ctrip tech 携程技术中心

ctrip tech 携程技术中心 | IT大咖说 知识分享平台

携程技术沙龙

携程酒店浏览客户流失概率预测

分享人：陈无忌

姓名：陈无忌

- 中国科学技术大学-计算机学院-研二
- 研究方向：大数据、车联网
- 携程云海两次大数据竞赛第一名
- 阿里天池大数据竞赛多次排名前1%
- 曾在科大讯飞实习-研究智慧城市

团队组成

张随远：中国科学院
计算技术研究所

秦宇君　中国科学院
计算技术研究所

陈无忌：中国科学技术大学
计算机学院

初赛排行榜

排名	团队	最终分数	提交次数	最后提交时间
1	我中了你的贱	0.094532	18	2016-08-31 23:45:19
2	TOBE1	0.086309	10	2016-08-31 00:12:31
3	打酱油的	0.085893	24	2016-08-31 22:50:39
4	doubi	0.080010	19	2016-08-31 03:59:46
5	海淀吴彦祖	0.079532	25	2016-08-31 20:23:51
6	caesar	0.079122	10	2016-08-31 01:51:27
7	ISI	0.079122	8	2016-08-31 00:56:00
8	AK47	0.079110	19	2016-08-31 02:45:03
9	NULL	0.078520	10	2016-08-31 01:02:14
10	三人行	0.078478	12	2016-08-31 19:42:31

目录
CONTENTS

目录
CONTENTS

1 问题分析-问题描述

深入了解用户画像及行为偏好，找到最优算法，挖掘出
响用户流失的关键因素，从而更好地完善产品设计
用户体验！

字段	
sampleid	样
label	目标变量
d	访问日期
arrival	入住日期
iforderpv_24h	24小时内是否访问订单填写页
decisionhabit_user	决策习惯：以用户为单位观察决策习惯
historyvisit_7ordernum	近7天用户历史订单数
historyvisit_totalordernum	近1年用户历史订单数

1 问题分析-评价标准

$$F1=\frac{2\times Recall\times Precise}{Recall+Precise}$$

问题优化目标：

- 精确度$(Precision)=\frac{预测为流失且实际发生流失的样本数量}{预测为流失的样本数量}$

- 召回率$(Recall)=\frac{预测为流失且实际发生流失的样本数量}{实际流失的样本数量}$

$$F12=\frac{6\times Recall\times Precise}{5\times Recall+Precise}$$

- 第一步：先按prob从高到低排序
- 第二步：根据你的输出即n个概率值，将这些概率值分别作为阈值，依次计算precision和recall,分别得到长度为n的precision数组和recall数组
- 第三步：在precision>=0.97的recall中，选取max(recall)

1 问题分析-数据概况-摘要

一个用户访问一条酒店产生的记录，但是这个label跟用户相关

sampleid	样本id
label	目标变量
d	访问日期
arrival	入住日期
iforderpv_24h	24小时内是否访问订单填写页
decisionhabit_user	决策习惯：以用户为单位观察决策习惯
historyvisit_7ordernum	近7天用户历史订单数
historyvisit_totalordernum	近1年用户历史订单数
……	……

订单本身特征

用户特征

1 问题分析-数据概况-摘要

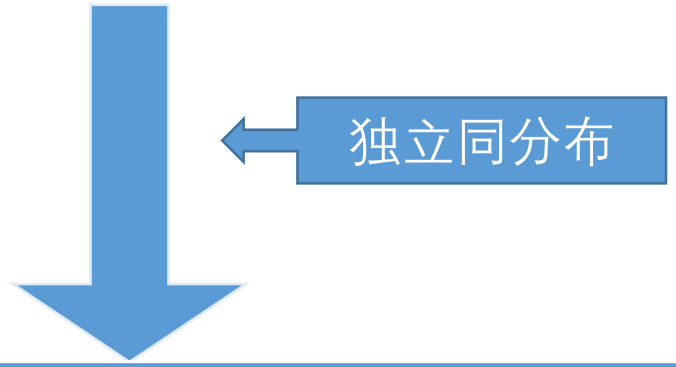
.....	.....
ordercanceledprecent	用户一年内取消订单率
landhalfhours	24小时内登陆时长
ordercannceledнum	用户一年内取消订单数
commentnums	当前酒店点评数
starprefer	星级偏好
novoters	当前酒店评分人数
consuming_capacity	消费能力指数

酒店特征

为保护客户隐私，
不提供uid等信息

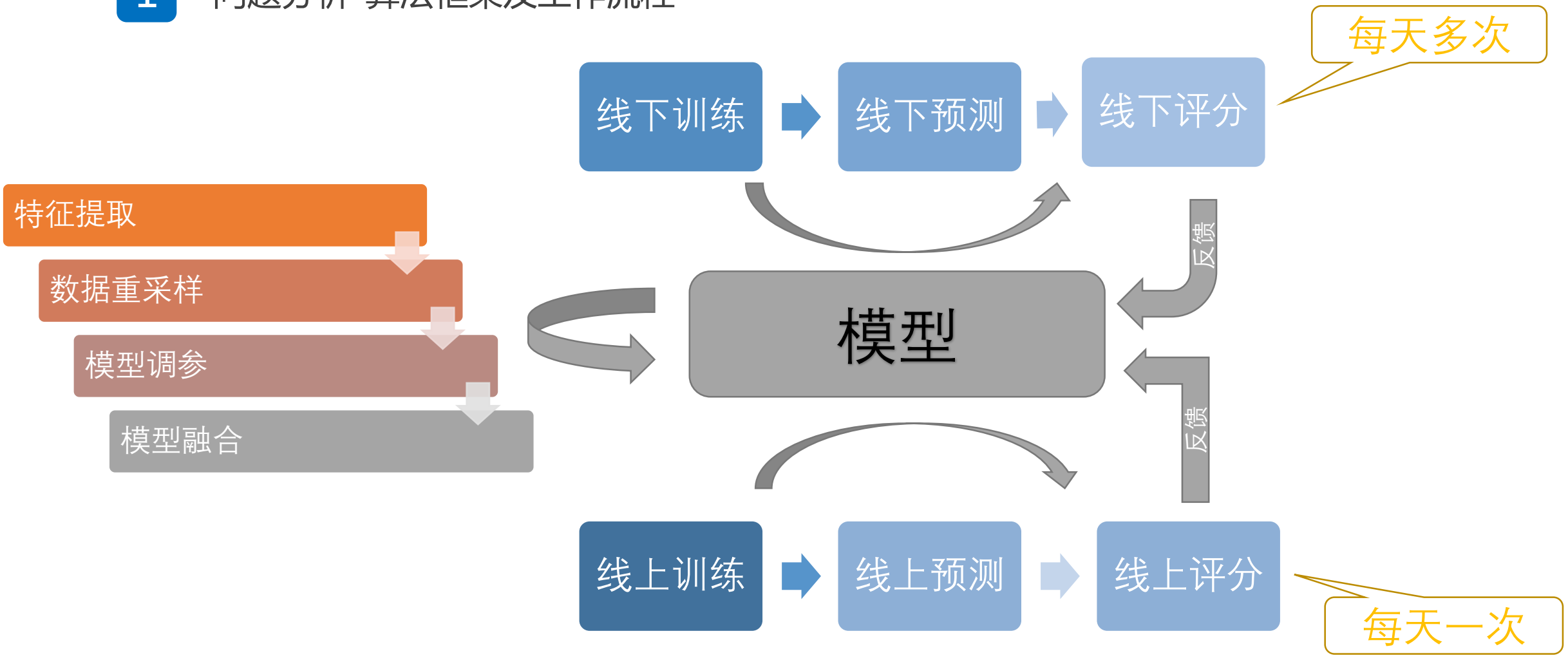
1 问题分析-数据集划分

线上训练集	线上测试集
2016.05.15-2016.05.21期间一周的访问数据	2016.05.22-2016.05.28期间一周的访问数据



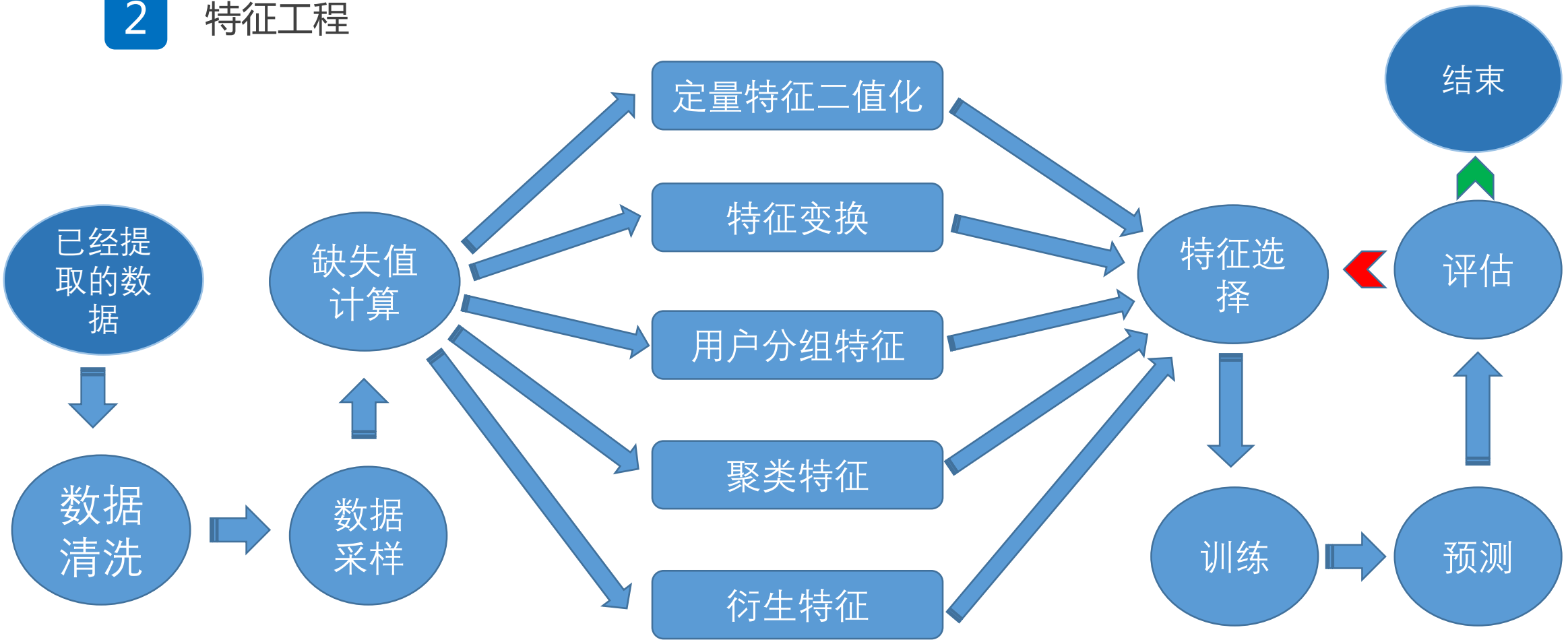
线下训练集	线下测试集
基于用户特征做数据集划分，2/3的数据作为本地训练数据	基于用户特征做数据集划分，2/3的数据作为本地训练数据

1 问题分析-算法框架及工作流程

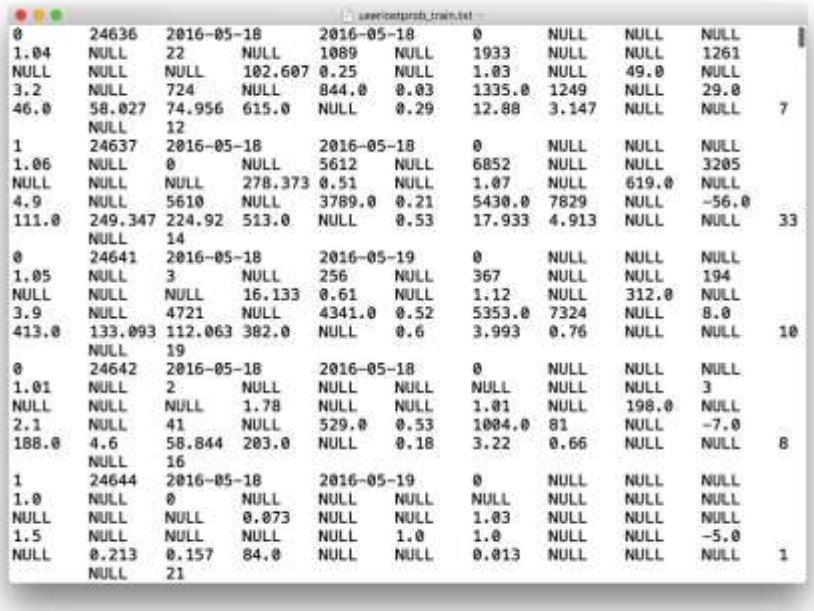


目录
CONTENTS

2 特征工程


已经提取的数据
数据清洗
数据采样
缺失值计算
定量特征二值化
特征变换
用户分组特征
聚类特征
衍生特征
特征选择
训练
预测
评估
结束

2 特征工程-缺失值处理



hoteluv	当前酒店历史uv
businessrate_pre	24小时历史浏览次数最多酒店商务属性指数
ordernum_oneyear	用户年订单数
cr_pre	24小时历史浏览次数最多酒店历史cr
avgprice	平均价格
lowestprice	当前酒店可定最低价

2 特征工程-二值化与特征变换

二值化

- one-hot编码
- sklearn-preprocessing-OneHotEncoder

特征变换

- 多项式变换
- Sklearn-preprocessing-PolynomialFeatures

2 特征工程-用户分组特征

用户特征
近7天用户历史订单数
近1年用户历史订单数
用户一年内取消订单率
用户一年内取消订单数
近3个月用户历史日均访问酒店数
消费能力指数
用户年订单数
……

基于
用户特征
分组

组内特征提取
每个用户组对应特征的最大值
每个用户组对应特征的最小值

2 特征工程-用户分组特征-Example

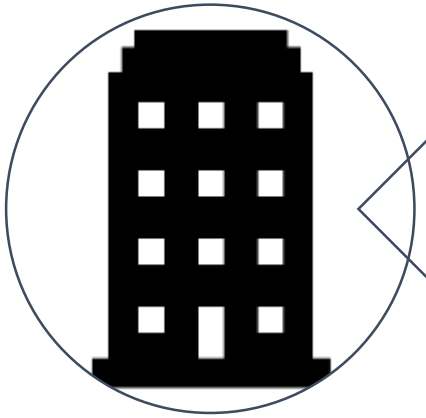
样本id	用户近一周订单数	用户近一年订单数	用户评级	酒店均价
1	12	132	5	214
2	2	53	2	332
3	12	132	5	432
4	12	132	5	142

样本id	用户近一周订单数	用户近一年订单数	用户评级	酒店均价	基于用户分组的酒店均价（MAX）	基于用户分组的酒店均价（MIN）
1	12	132	5	333	444	222
2	2	53	2	666	666	666
3	12	132	5	444	444	222
4	12	132	5	222	444	222

2 特征工程-聚类产生的特征

基于用户聚类

聚类属性
近7天用户历史订单数
近1年用户历史订单数
用户一年内取消订单率
用户一年内取消订单数
近3个月用户历史日均访问酒店数
用户偏好价格-24小时浏览最多酒店价格



基于酒店聚类

聚类属性
当前酒店历史cr
当前酒店点评数
当前酒店评分人数
当前酒店历史取消率
当前酒店历史uv

2 特征工程-衍生特征

字段	解释
sampleid	样本id
d	访问日期
arrival	入住日期

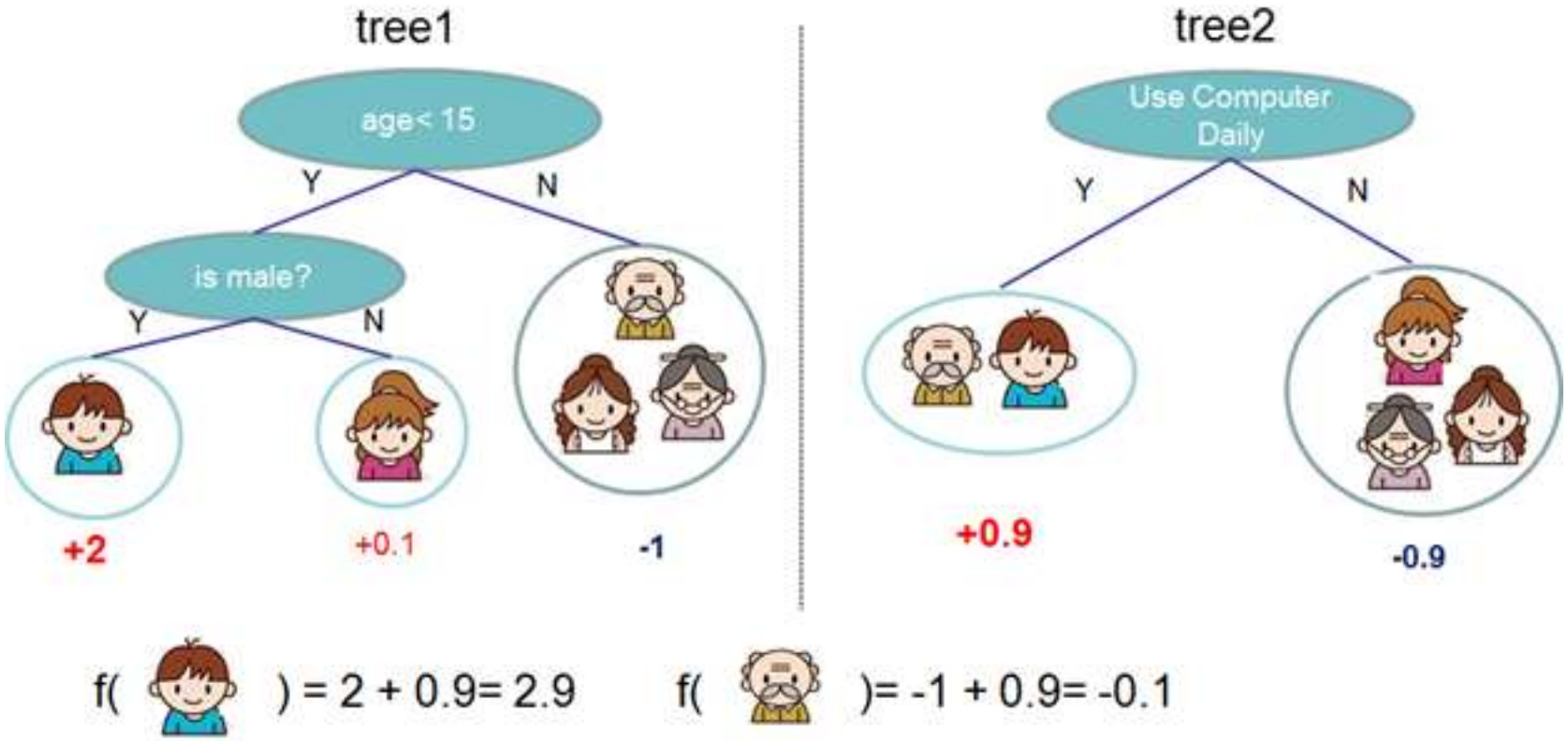
字段	解释
sampleid	样本id
intervals	访问日期和入住日期的差值
D_ifweekend	访问日期是否是周末
arrival_ifweekend	入住日期是否是周末

目录
CONTENTS

3 模型原理及调参-尝试的模型

模型	速度	效果	特点
Logistic Regression	快	一般	速度快，易于解释。特征空间大或缺失值过多的时候表现欠佳
Random Forest	一般	一般	采用bagging的思想对特征进行采样，抗噪能力强
GBDT	一般	较好	采用boosting思想，对残差迭代，模型拟合效果好
XGBoost 	快	好	针对gbdt算法改进，增加二阶导及并行化支持

3 模型原理及调参-gbdt原理

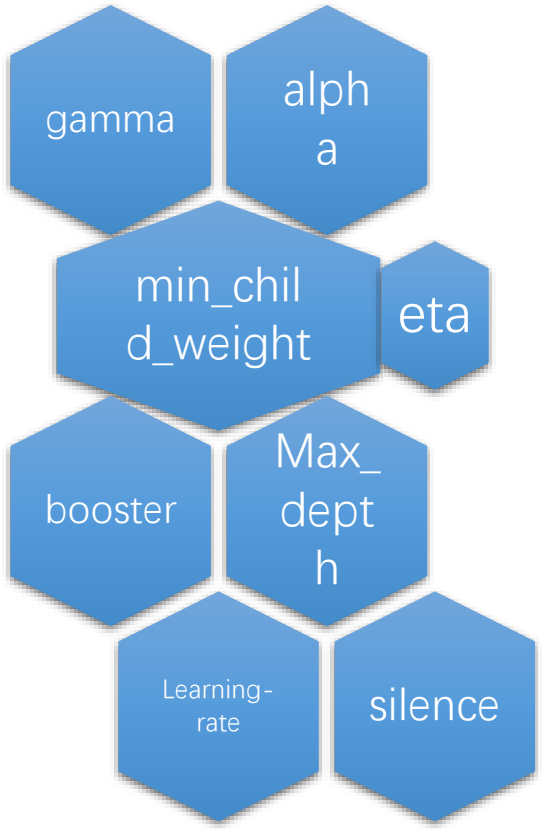


3 模型原理及调参-Xgboost原理

- 目标 $Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant$
- 用泰勒展开来近似我们原来的目标
  - 泰勒展开：$f(x+\Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2$
  - 定义：$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}),\quad h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + constant$$

3 模型原理及调参-Xgboost调参



再基于用户分组将训练集划分出1/3数据作为验证集

↓

通过GridSearch的方法尝试**max_depth、learning_rate、n_estimators**三个参数值的组合

↓

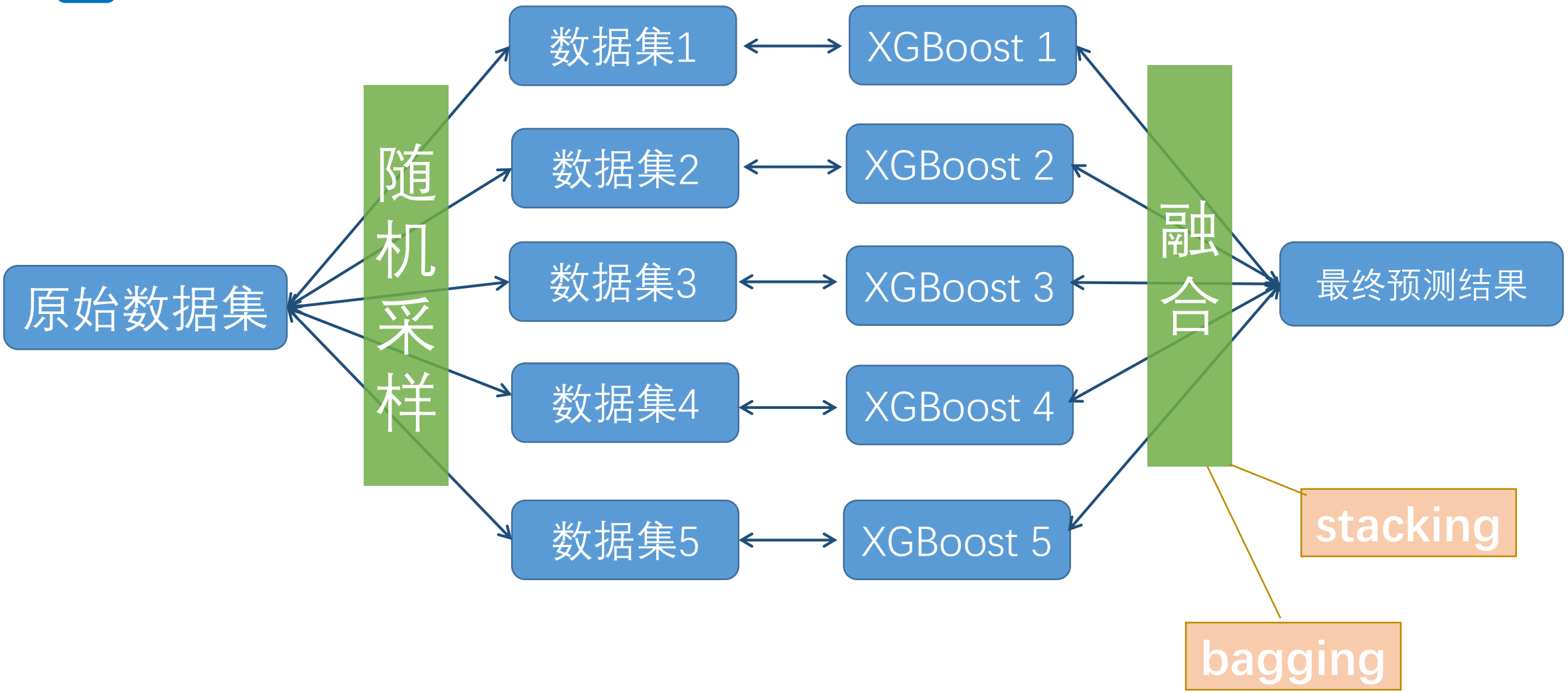
选出较优的几组参数做ensemble learning

Tips：

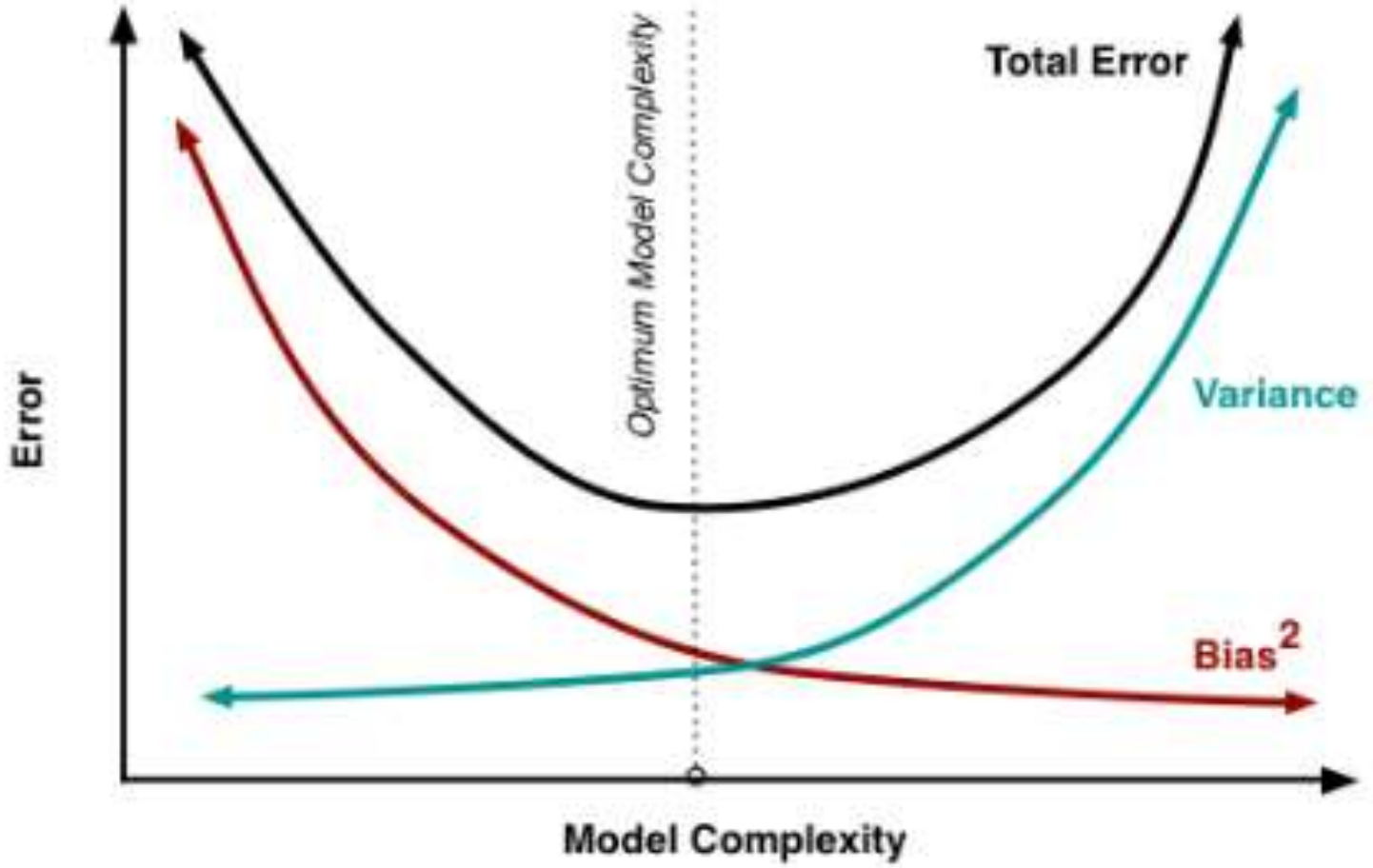
- 线下做gridSearch的时候，不用太过于追求调参的效果，适度即可。
- 因为做ensemble learning，所以挑选最优参数组合的时候尽量挑选参数差别大的组合。
- 更多技巧参见：Complete Guide to Parameter Tuning in XGBoost

目录
CONTENTS

4 模型融合-框架


原始数据集
随机采样
数据集1
数据集2
数据集3
数据集4
数据集5
XGBoost 1
XGBoost 2
XGBoost 3
XGBoost 4
XGBoost 5
融合
最终预测结果
stacking
bagging

4 模型提升-提升的原理


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

目录
CONTENTS

5 总结-经验分析

数据分析及预处理是首要

- 不要上来就抽特征跑模型

特征很重要

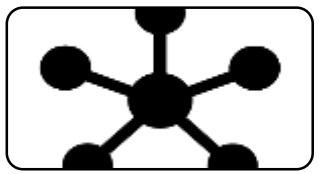
- 特征决定模型的上限

调参是精益求精

- 先调好特征最后调模型，尽量线下调参

别忘了关注评价标准

- 及时修改模型的loss



模型融合是杀手锏

- 模型融合勿急躁，调好单模型是关键

ctrip tech 携程技术中心

ctrip tech 携程技术中心 | IT大咖说 知识分享平台

THANK YOU!

Q&A