重点：

P10-11：EDA数据探索分析；

P12-15：预处理，通过可视化分析来找一场样本；

P16-17：特征工程；

P18：特征选择；

P21：强烈推荐的两个模型：XGBoost、LightGBM；

P25：baseline = 基本处理 + 原始特征 + 验证方法(时序验证或交叉验证) + LightGBM

如何进行一场数据挖掘算法竞赛

主讲：王贺 （鱼遇雨欲语与余）

内容合作者：张杰（杰少，老师）

2019/05/02

本次LIVE的内容

- 1. 为什么要参加数据挖掘竞赛？能带来什么？
- 2. 参加竞赛需要哪些基础知识和技能？
- 3. 如何选择适合自己的竞赛？
- 4. 竞赛中的几个主要模块
- 5. 竞赛过程中最重要的事情
- 6. 好的竞赛总结比竞赛过程更重要
- 7.案例分享（天池“全国城市计算AI挑战赛”）

1. 为什么要参加数据挖掘竞赛？能带来什么？

1.1 从理论知识到工程应用

1.2 求职加分，企业看重

1.3 奖金的激励（丰厚）

1.4 交友，学习，PK高手

oppo

- 精英奖：TOP20主要参赛选手可直接入围OPPO校招终面（即从“笔试+专业+部长+HR”简化为“部长+HR”）
- 极客奖：TOP100主要参赛选手可参与OPPO校招可免去笔试环节

YOUKU 复赛审核通过的排名前10队伍，可进入阿里（优酷）校招绿色通道。

腾讯广告 复赛最终成绩前 20 名，获得校园招聘（包括实习）免笔试绿色通道。

京东

奖项	数量	奖励（/支队伍）
大赛冠军	1支队伍	¥50万+Special offer+证书
大赛亚军	1支队伍	¥10万+Special offer+证书
大赛季军	1支队伍	¥5万+Special offer+证书
大赛第4-5名	2支队伍	¥2万+技术岗直通终面（1年内有效）+证书
大赛第6-7名	2支队伍	JDRead京东电子书阅读器+技术岗直通终面（1年内有效）
大赛第8-10名	3支队伍	京东叮咚智能音箱+技术岗直通终面（1年内有效）
大赛周冠军	线上赛期间每周A榜Top1团队	京东叮咚智能音箱
大赛入围第11-50名	线上赛结赛日B榜排行榜Top11-50团队	招聘免笔试绿色通道（1年内有效）

2. 参加竞赛需要哪些基础知识和技能?

很多人会问当理论学到什么程度的时候才能参加算法竞赛?

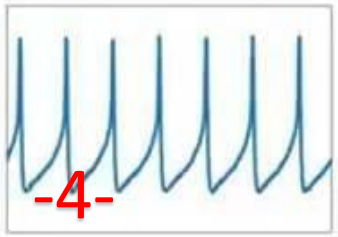
2.1 理论知识掌握：评价指标、数据分析、特征工程、常用模型

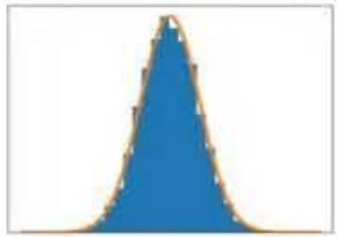
2.2 工具的掌握　语言的选择：Python

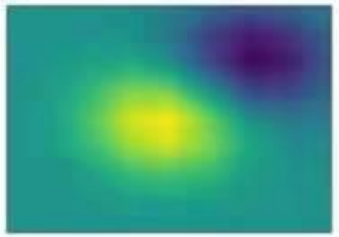
可视化工具：Matplotlib、Seaborn

数据处理工具：Pandas、NumPy

机器学习库：Sklearn、XGBoost、LightGBM







3. 如何选择适合自己的竞赛?

3.1 竞赛平台：kaggle TIANCHI DataFountain DC竞赛 www.dcjingsai.com

3.2 竞赛分类：按任务目标划分：分为分类问题和回归问题；

按领域归属划分：搜索推荐、时间序列（销量预测、股票预测）、自然语言处理（文本分类、情感分析）、计算机视觉（目标检测、图像分类）等

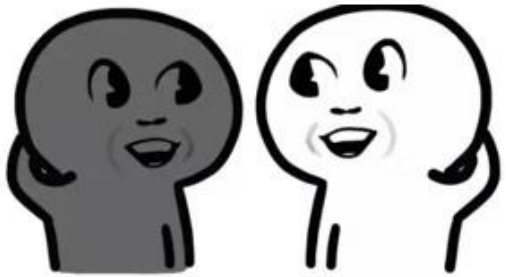
3.3 硬件要求：自己的机器能够支持并完成这场比赛，内存、显卡等，或者借助云服务器。根据比赛类型，比赛数据大小来确定。

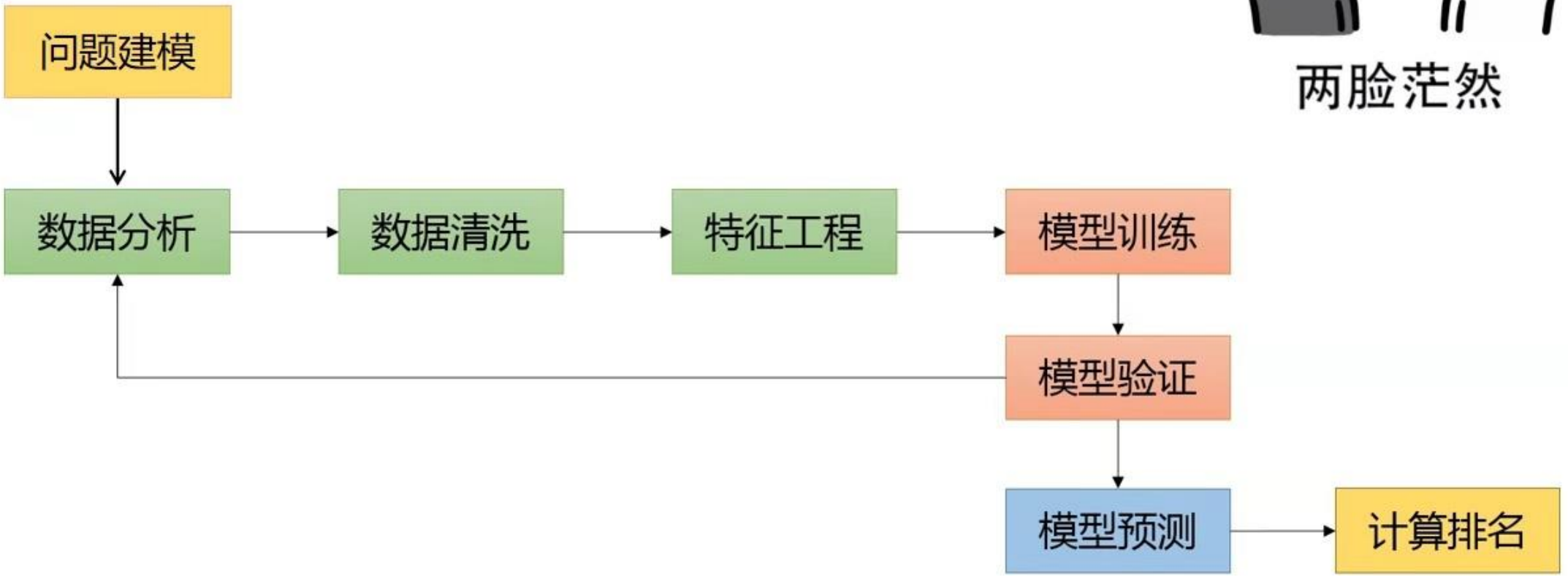
七、建议使用的计算资源

单机运行内存不超过 128G，CPU 不超过 24 核。

3.4 与自己专业的相关性：研究方向

4. 竞赛中的几个主要模块





4. 竞赛中的几个主要模块

4. 1 问题建模

4.1.1 赛题理解

- 业务理解
- 赛题数据
- 评价指标
  - 分类指标：精确率、召回率、AUC、logloss
  - 回归指标：MAE、MAPE、RMSE等

及对用户体验的控制策略。通常来说，基本竞争力可以用千次曝光收益 ecpm = 1000 * cpc_bid * pctr = 1000 * cpa_bid * pctr * pcvr （cpc, cpa 分别代表按点击付费模式和按转化付费模式)。综上，其中前者决定广告能参与竞争的次数以及竞争对象，后者决定在每次竞争中的胜出概率。二者最终决定广告每天的曝光量。

$$\mathbf{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(F_t + A_t)/2}$$

$$\text{score} = \frac{1}{n}\sum_{k=1}^{n}\frac{(imp_0 - imp_k)(bid_0 - bid_k)}{|(imp_0 - imp_k)(bid_0 - bid_k)|}$$

4. 竞赛中的几个主要模块

4. 1 问题建模

4.1.2 线下验证

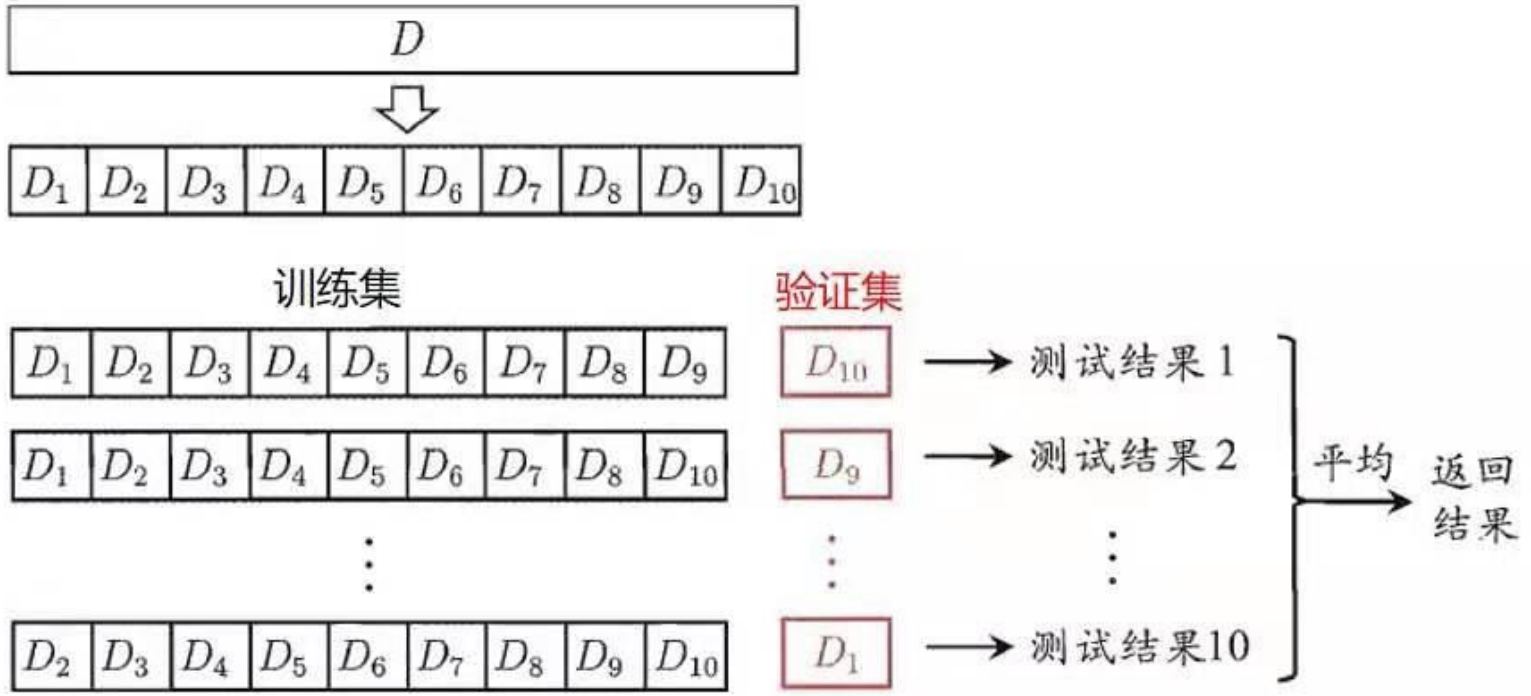
时序验证

4. 竞赛中的几个主要模块

4. 1 问题建模

4.1.2 线下验证

K折交叉验证



4. 竞赛中的几个主要模块

4. 2 数据探索性分析（EDA）

4.2.1 Why EDA?

- 了解数据
  - **数据类型大小**(需要什么配置,参赛代价大不大)...
  - **数据是否干净**(明显错误的数据,例如身高5m...)
  - **标签是什么类型的,是否需要格式转换?**... (DataFrame.info())
- 为建模做准备
  - **线下验证集的构建,是否可能会穿越?**
  - **是否存在某些奇异的现象? 为特征工程做准备:**
    - 例如时序的周期变化现象

4. 竞赛中的几个主要模块

4. 2 数据探索性分析（EDA）

4.2.2 What must see?

EDA必看

- **数据集大小,字段类型**：数据多大,每个字段是什么类型的
- **缺失值的情况**：缺失是否严重,是否缺失有特殊含义
- **特征之间是否冗余**：比如身高用cm表示和m表示就存在冗余
- **是否存在时间信息**：潜在的穿越问题
- **标签的分布**：是否类别分布不平衡等
- **训练集测试集的分布**：测试集中有的字段很多特征训练集没有

EDA主要是通过可视化、统计监测来进行，看一下均值情况或者是方差

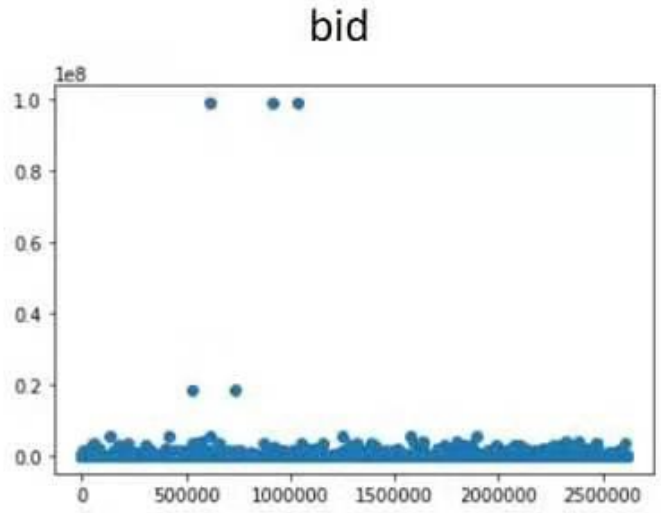
4. 竞赛中的几个主要模块

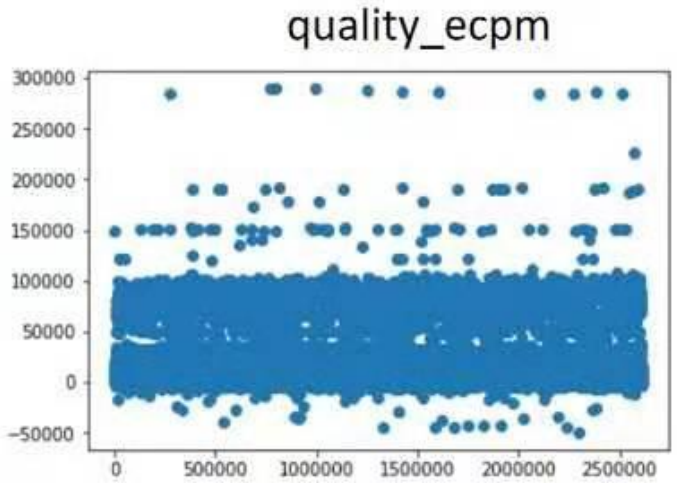
4. 3 特征工程

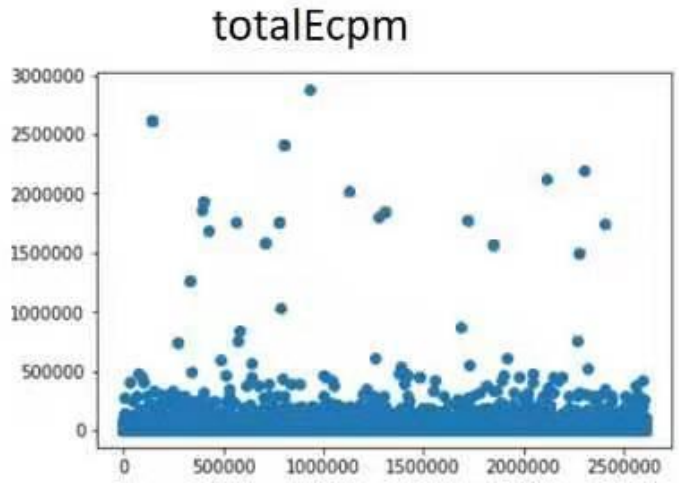
4.3.1 数据预处理

(1)离群点处理，通过可视化查看，对弈哪些离群点可以考虑删除

离群点处理







4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.1 数据预处理

缺失值处理

真正（意义）缺失?

否

有特定的业务含义!
- **填充max(fea)+1/min(fea)-1**

是

填充?
- **各种填充方案**
- **不填充,设为np.nan**
- **对比效果选择**

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.1 数据预处理

错误值处理

假值（明暗）处理

明

明显错误!
- **血压999999**
- **体重800**

暗

Santander Product Recommendation

Can you pair products with people?

匿名比赛
- **出现-1和999,表示了缺失值,替换np.nan**

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.1 数据预处理

假标签处理

常见两种假标签

错误标签

标签错误

- **血压999999等**

标签和评估指标不一致

CAT

Caterpillar Tube Pricing

Model quoted prices for industrial tube assemblies

$30,000 · 1,323 teams · 4 years ago

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

- **标签log1p化,再用mse进行优化学习**

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.2 特征提取

类别特征

- 自然数编码
- 独热编码
- 计数统计（异常值敏感）
- 计数排名（异常值不敏感）
- 目标编码
- 交叉组合（类别-类别、类别-数值）

类别特征的构造思路有旁边几种

例如 腾讯比赛原始类别特征

广告id
广告账户 id
商品 id
商品类型
广告行业 id

＋

定向人群
投放时间

- 独热编码one-hot
- 计数统计count
- 目标编码/mean
- 交叉组合

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.2 特征提取

数值特征

- 分桶（转化为离散特征）
- 特征交叉（加减乘除，根据业务）
- 交叉组合（类别-数值）

数值特征的构造方法少一些

腾讯比赛原始数值特征

bid
pctr
quality_ecpm
totalEcpm

粒度可以变细，如某一类别的平均出价*点击率

特征交叉

交叉组合

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.2 特征提取

时间特征

- 日期变量（年、月、周、日、小时、分钟）
- 时序相关特征（历史平移，滑窗统计）

时间特征是一个重要的特征，离得时间越近越能代表最近的行为

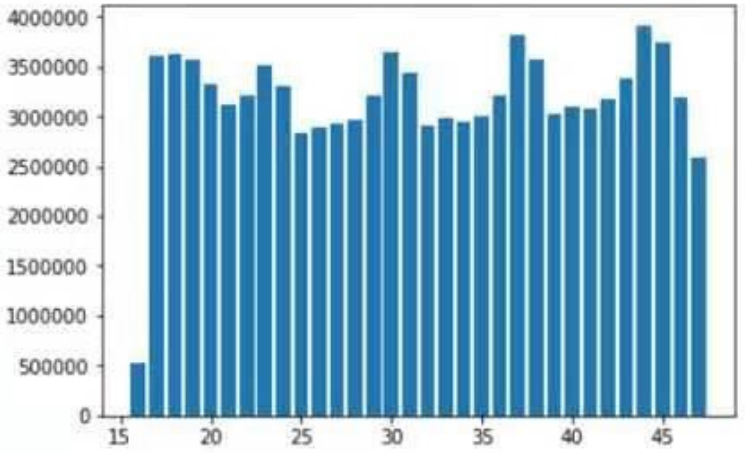
后面案例详细介绍

腾讯比赛原始时间特征

请求时间
创建时间

构造月、周、日特征

根据日进行历史平移



4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.3 特征选择

过滤法 — 相关系数 + 卡方检验 + 互信息

可以通过旁边的三类方法进行特征选择

封装法 — 前向搜索 + 后向搜索

嵌入法 — 基于学习模型的特征排序

如树模型可以直接返回特征的重要性

4. 竞赛中的几个主要模块

4. 3 特征工程

4.3.3 特征选择

数据（1416维）

字段	解释
user_tags	用户标签信息,以逗号分隔

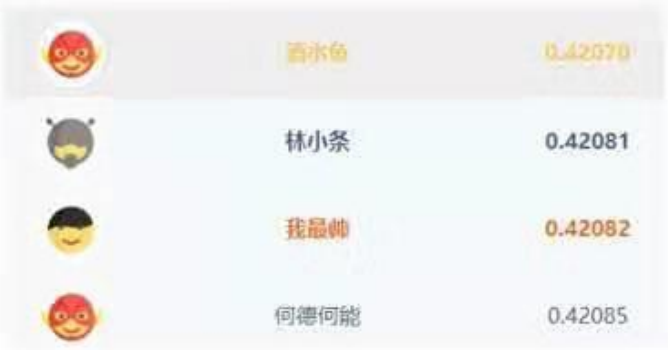
例子：当时数据完全展开有1416维

2018科大讯飞AI营销算法大赛 算法竞赛

科大讯飞

本次大赛提供了讯飞AI营销云的海量广告投放数据，参赛选手
的广告点击概率

商业 分类 时间：2018/09/01-201



可能包含重要信息，也可能包含噪音，给出三种做法

完全展开 → 1416维，分数一般

完全展开 + 卡方检验 → 100维，分数较好

完全展开 + 重要性得分 → 100维，分数很好

4. 竞赛中的几个主要模块

4.4 必备模型

XGBoost

对特征处理要求低

对类别和连续特征友好

LightGBM

缺失值不需要填充

(冠军) LightGBM

(亚军) LightGBM

(季军) LightGBM

4. 竞赛中的几个主要模块

4.5 模型融合

5. 竞赛过程中最重要的事情

5.1 海量数据分析（对于数据的理解，业务的分析能力提升）

5.2 不断尝试新的idea（相关论文或自己的想法）

5.3 多向优秀的选手学习提问

6. 好的竞赛总结比竞赛过程更重要

6.1 赛后及时总结：自己的整体思路、关键代码、自己的不足、还需要做哪些尝试。

6.2 学习优秀方案：不仅局限于自己的思维方式，其他人是如何思考的，哪里是可以借鉴的，进行对比发现自己的不足。

7. 案例分享（全球城市计算AI挑战赛）

基本处理+原始特征+验证方法+LightGBM=Baseline

各站点每十分钟的进出流量预测

列名	类型	说明	示例
time	String	刷卡发生时间	2019-02-01 00:30:53
lineID	String	地铁线路 ID	C
stationID	int	地铁站 ID	15
deviceID	int	刷卡设备编号 ID	2992
status	int	进出站状态，0 为出站，1 为进站	1
userID	String	用户身份 ID	Ad53ce59370e8b141dbc99c03d2158fe4
payType	int	用户刷卡类型	0

地铁站之间的连接关系表

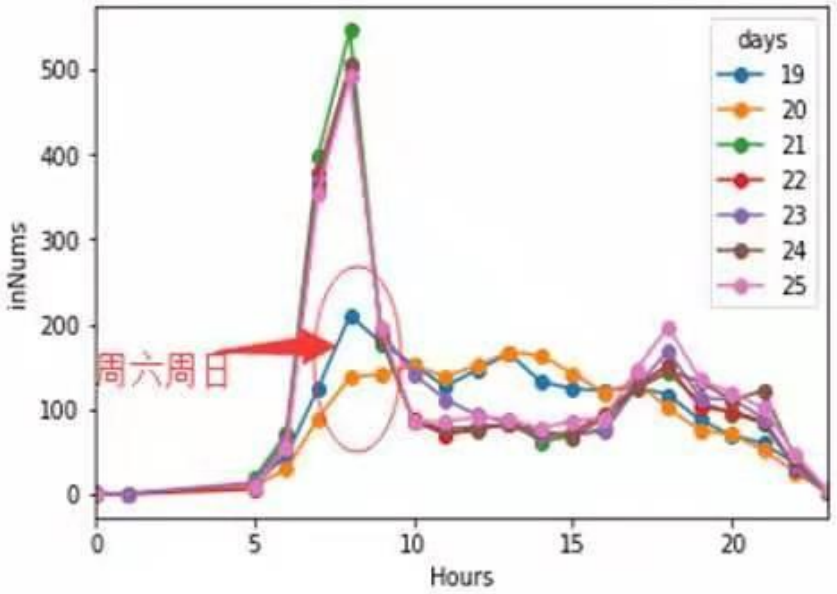
7. 案例分享（全球城市计算AI挑战赛）

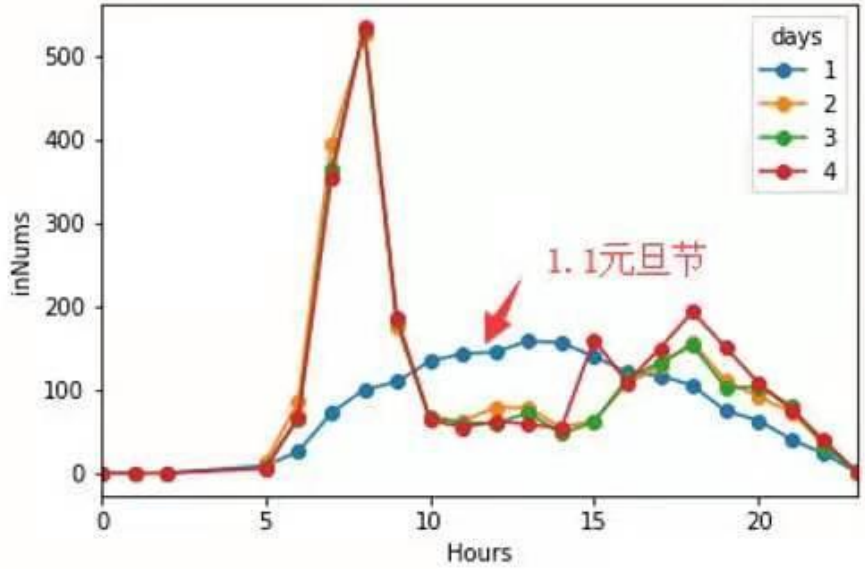
特征工程

离群点处理

缺失值处理

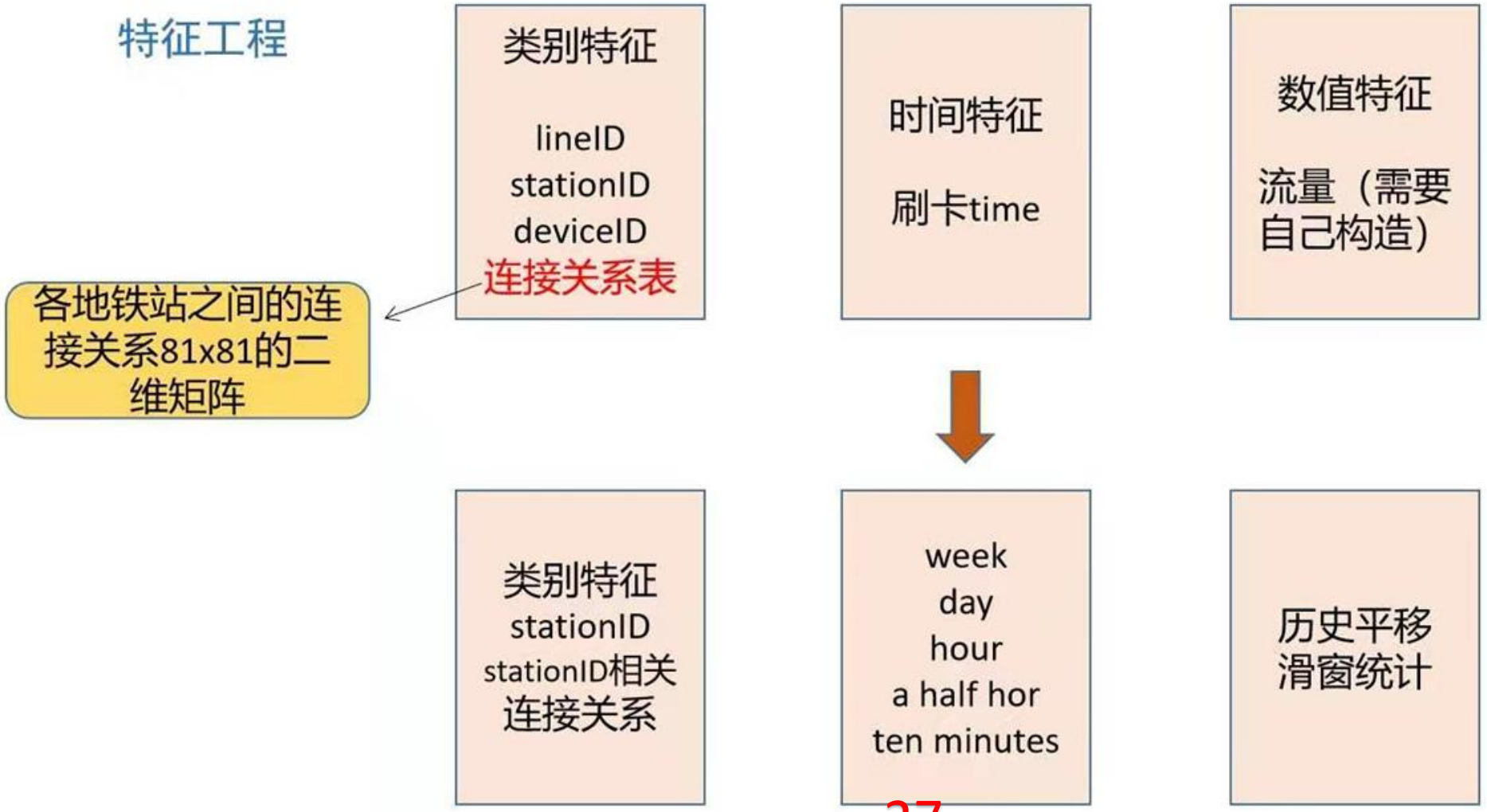
通过可视化分析来找异常样本





7. 案例分享（全球城市计算AI挑战赛）

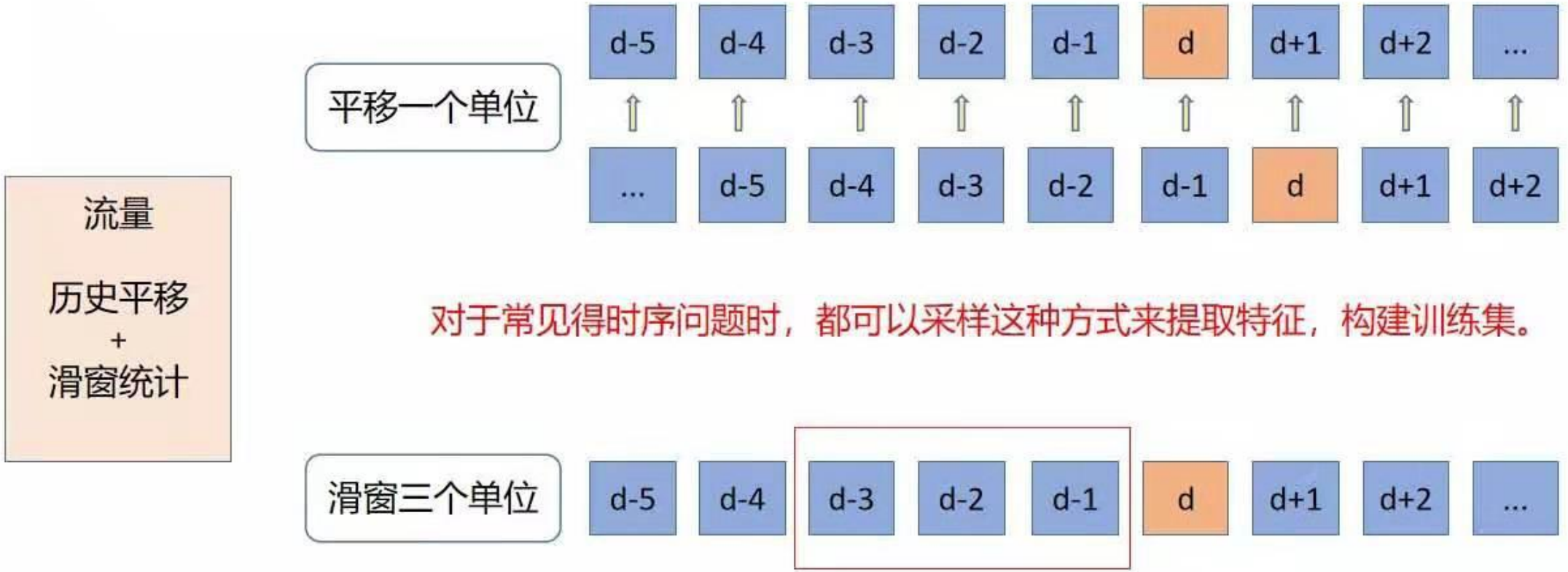
特征工程



7. 案例分享（全球城市计算AI挑战赛）

特征工程

以天为单位，每个方格分别代表
发生在每天对应10分钟的流量



对于常见得时序问题时，都可以采样这种方式来提取特征，构建训练集。

构造统计特征，如均值，最大值，
最小值来反应前三天的情况

4. 竞赛中的几个主要模块

4. 1 问题建模

4.1.2 线下验证

时序验证