第三章 数据预处理

- 1、数据源冲突
  - 冲突类型
    - 数据类型（数据格式）
    - 数据结构（相同关联的主键有不同的数据结构）
    - 记录粒度（记录根据不同特征的拆分）
    - 数据值域的定义（数据的值域）
    - 数据值（最重要和最关键的问题）
  - 冲突原因（举例）
    - 内部工具与第三方工具的数据冲突 主体：网站分析工具和代理商或广告媒体的数据
      - 对比指标不同（不同的指标计算逻辑不同）
      - 测量的数据不同
      - 网络丢包的问题
      - 去重机制的问题
      - 用户中途退出的问题
      - 页面跟踪加载的问题
      - 动机导致的数据夸大
      - 其他因素
    - 内部同一业务主体的同一类数据工具的数据测量冲突 主体：不同监测跟踪代码对同一个指标的监测数据
      - 指标定义不同
      - 采集逻辑不同（采集逻辑、触发机制、异常处理等）
      - 系统过滤规则不同（后端的规律规则不同必导致数据值的差异）
      - 更新时间不同
      - 监测位置不同
    - 内部同一业务主体的不同数据工具的数据测量冲突 主体：销售系统跟会员管理系统统计的订单数据
      - 订单来源差异（如线上线下）
      - 特殊商品订单跟踪（如实体货物与虚拟货币）
      - 订单状态差异
      - 数据同步问题
      - 内部系统拆单问题（子订单）
  - 处理冲突
    - 消除数据冲突并形成唯一一份数据
    - 不消除冲突也不做任何处理（冲突占比较小可忽略）
    - 不消除冲突但是使用全部冲突数据
    - 数据冲突必要关注
      - 差异性

- 稳定性

- 2、抽样或是全量数据
  - 抽要的原因
    - 数据计算资源不足
    - 数据采集限制
    - 时效性要求
    - 通过抽样实现快速的概率验证
    - 通过抽样解决样本不均衡问题（过抽样/欠抽样/组合&集成）
    - 无法实现对全部样本覆盖的数据化运营场景
    - 定性分析的工作需要
  - 如何抽样
    - 简单随机抽样
    - 等距抽样
    - 分层抽样
    - 整群抽样
  - 抽样需要注意的问题
    - 抽样要能反映运营背景（以下反例）
      - 数据时效性问题
      - 缺少关键因素数据
      - 不具备业务随机性
      - 没有考虑业务增长性
      - 没有考虑数据来源的多样性
      - 业务数据的可行性问题
    - 抽样要能满足数据分析和建模需求
      - 抽样样本量的问题
        - 月度销售预测
        - 预测（分类和回归）分析建模
        - 关联规则分析建模
        - 异常检测类分析建模
      - 抽样样本在不同类别中的分布问题 抽样样本能准确代表全部整体特征
        - 非数值型
          - 特征值域分布与总体一致
        - 数值型
          - 数据分布区间和各个统计量与整体数据分布区间一致
        - 特殊数据
          - 与整体数据分布区间一致
        - 异常检测类数据
          - 应用包含全部异常样本
            - 数据稀少优先包含异常数据
          - 需去除非业务因素的数据异常
            - 有类别特征

- 需与类别特征分布保持一致
- 没类别特征
  - 属于非监督式学习，需与整体分布一致

- 3、解决运营数据的共线性问题
  - 共线性（多重共线性）问题概述
    - 定义
      - 输入的自变量之间存在较高的线性相关度
    - 常见案例
      - 访问量和页面浏览量
      - 页面浏览量和访问时间
      - 订单量和销售额
      - 订单量和转化率
      - 促销费用和销售额
      - 网络展示广告费和访客数
    - 变量间共线性的原因
      - 数据样本不足
      - 多变量基于时间有相同或相反的演变趋势
      - 多变量间存在一定的推移关系，但总体变量间趋势一致，只是发生时间不一致
      - 多变量间存在近似线性关系
    - 影响
      - 导致回归模型的稳定性和准确性大大降低
      - 过多无关的维度参与计算浪费资源和时间
  - 检验共线性
    - 容忍度
      - 容忍度介于0到1，值越小，说明这个自变量与其它自变量间越可能存在共线性问题，
    - 方差膨胀因子VIF
      - 容忍度的倒数，值越大则共线性问题越明显，通常以10作为判断边界
        - VIF<10
          - 不存在多重共线性
        - 10≤VIF<100
          - 较强的多重共线性
        - VIF≥100
          - 严重多重共线性
    - 特征值
      - 对自变量进行主成分分析
        - 多个维度的特征值等于0，则可能有比较严重的共线性
  - 解决共线性问题 注：完全解决是不可能的
    - 增大样本量

        - 岭回归法
            - 存在较强共线性的回归应用中较为常用
        - 逐步回归法
        - 主成分回归
        - 人工去除
- 4、有相关性分析的混沌
    - 概述
        - 多个具备相关关系的变量进行分析以衡量变量间相关程度
        - R（相关系数） ∈[–1,1] \in [-1,1]∈[–1,1]
    - 相关和因果
        - 相关
            - x1 x_{1} x1 和 x2 x_{2} x2 是逻辑上的并列相关关系
        - 因果
            - 因为 x1 x_{1} x1 所以 x2 x_{2} x2
    - 相关系数低并不是不相关
        - 负相关只是意味着两个变量的增长趋势相反
        - R的绝对值来判断相关性的强弱
        - 变量间除了线性关系外，还包括指数关系、多项式关系、幂关系
- 5、标准化
    - Z-Score（标准差标准化）

$$x^* = \frac{x - \bar{x}}{\sigma}$$

        - 实现中心化和正态分布，改变原有数据分布
        - 不适合用于对稀疏数据做处理
    - Max-Min

$$x^* = \frac{x - min}{max - min}$$

        - 实现归一化，离差标准化，将数值映射到[0,1]
    - MaxAbs
        - 用于稀疏数据
        - x‘=x ∣ max ∣ ,x∈[0,1]x`= \frac{x}{ \left| max \right| },x \in [0,1]x‘= ∣ max ∣ x ,x∈[-1,1]
    - RobustScaler
        - 针对离群点
- 6、离散化
    - 概述
        - 把无限空间中有限的个体映射到有限的空间中
        - 大多针对连续数据

- 必要性
  - 节约资源，提高效率
  - 计算的需要
  - 提高模型的稳定性和准确度
  - 特定数据处理和分析必要步骤，在图像处理应用广泛
  - 模型结果应该和部署的需要
- 不同数据类型离散化
  - 时间数据
    - 切分
      - 一天时间的切分
      - 日为粒度的切分
    - 数据类型
      - 分类数据（上、下午）
      - 顺序数据（周一、周二、周三...）
      - 数值型数据（一年有52个周，周数）
  - 多值离散数据
    - 合并/重新划分
  - 连续数据
    - 分位数法
      - 四分位、五分位、十分位等分位数
    - 距离区间法
      - 等距区间或自定义区间
    - 频率区间法
      - 不同数据的频率分布进行排序
    - 聚类法
      - K均值将样本集分为多个离散化的簇
    - 卡方
      - 找出数据的最佳临近区间并合并，形成较大的区间
  - 连续数据的二值化
    - 与阈值比较大小得某固定住1/0，便得到两个值域的数据集
- 7、考虑的运营业务因素
  - 考虑固定和突发运营周期
    - 有计划的周期性
      - 对于时间序列特征明显的建模影响很大，包括时间序列、时序关联、隐马尔可夫模型等
    - 临时或突发周期
    - 不同周期下所产生的数据可能有差异
  - 考虑运营需求的有效性
    - 缺少数据
    - 需求不合理
    - 条件限制

- 资源限制
- 低价值需求

- 考虑交付时要贴合运营落地场景
  - 维持原有指标
  - 更容易理解的算法模型
  - 数据生产和应用环境
- 不要忽视专家经验
  - 数据工作方向与逻辑
- 考虑业务需求的变动因素