数据分析题框架

- 特征工程
  - 数据清洗
    - 低方差滤波器：删除方差为0的特征
    - 缺失值滤除：设定阈值，缺失值超出该阈值则剔除该变量
  - 数据处理
    - 归一化
    - 标准化
  - 特征选择
    - 目的：实现代表性与独立性
    - 加权集成筛选法（代表性）
      - 稳定性（基于统计）
        - 灰色关联分析
        - 最大信息系数（2021-D-02)
          - 检验变量间的线性与非线性相关性
        - 距离相关系数(2021-D-176)
          - 检验变量间的非线性相关性
      - 非稳定性（基于机器学习）
        - 随机森林
        - 弹性网络（2021-D-02)
        - 平均精度损失(2021-D-176)
    - 相关性分析（独立性)
      - 变量间线性相关分析
        - pearson系数
      - 变量间非线性相关分析
        - 最大信息系数
        - 距离相关系数
- 预测与分类
  - 数据准备
    - 特征变量的选择
      - 若本文需要的特征前面未进行特征工程，则此问构建两阶段模型XGBoost+预测模型
      - 若前面进行过特征工程，则简要阐明
    - 数据预处理

        - 归一化
        - 标准化
    - 数据集的划分
        - 是否存在不均衡问题
            - 过采样与欠采样
        - 简要阐明划分比例的原因
- 变量的独立性
    - 变量间相关性检验
- 模型选择与超参的调整
    - 选择多种机器学习方法，说明选择的原因
    - 列出多个模型训练的参数设置
    - 给出参数优化方法（网格搜索等），以及最终模型参数
- 模型的评价与分析
    - 模型性能分析
        - 建立评价指标，阐述模型表现
        - 预测与分类的评价指标不同
            - 预测：MAE/RMSE 等
            - 分类：ACC、召回率、AUC等
    - 模型鲁棒性（稳定性）分析
        - 重复多次实验，结合图标，阐述稳定性
    - 过拟合分析
        - 列出评价指标中训练集与测试集的表现，几句话表明过拟合问题不存在即可
- 模型预测结果
    - 根据题目要求，对指定数据进行预测或分类
- 优化问题
    - 建立优化目标函数，描述约束条件
        - 约束条件或目标函数可能是前面的预测函数，需结合起来
    - 求解
        - 智能优化算法
        - 图标展示模型优化过程与参数设置
    - 对结果进行分析
        - 最好利用相关背景知识
    - 展示结果