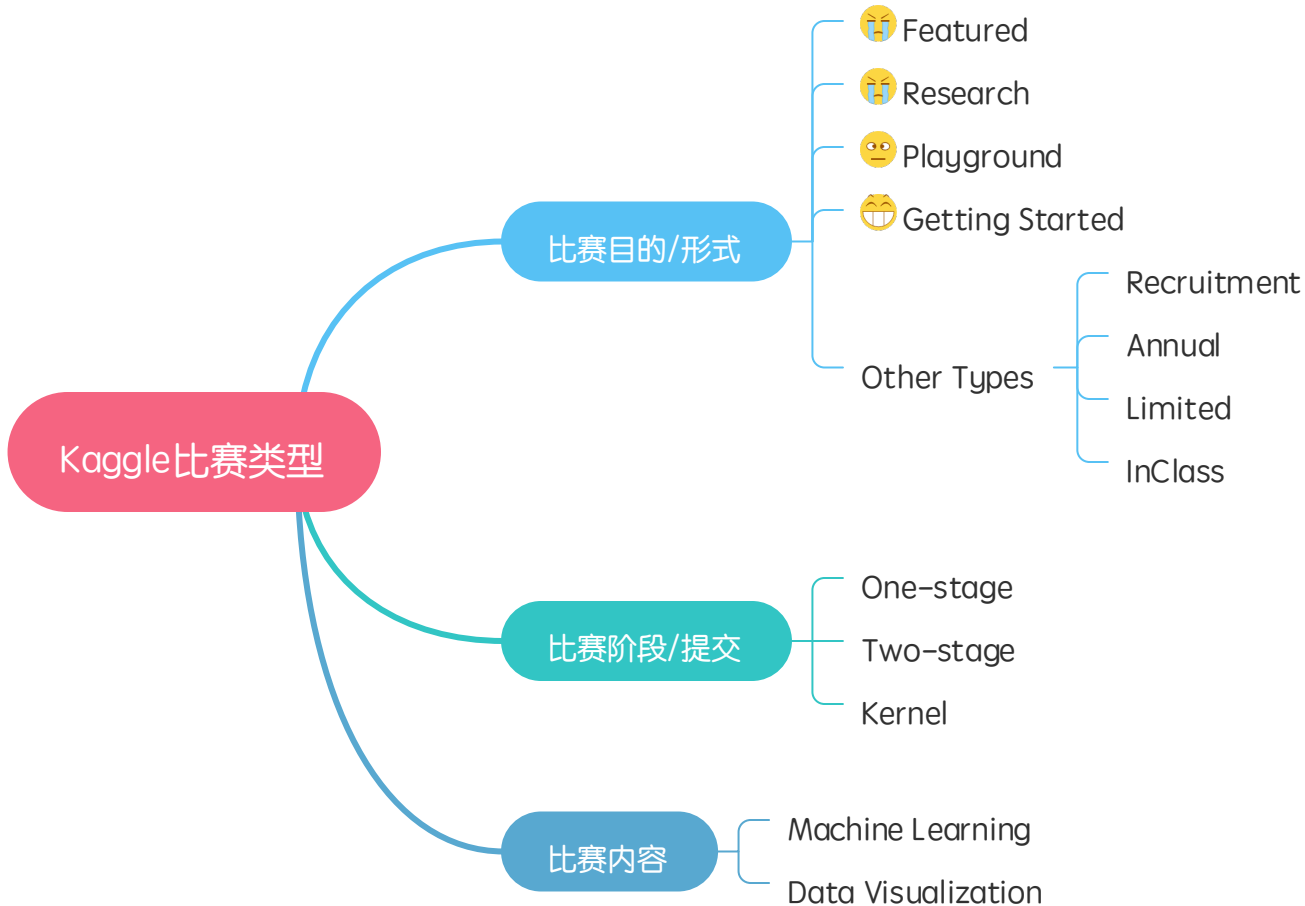
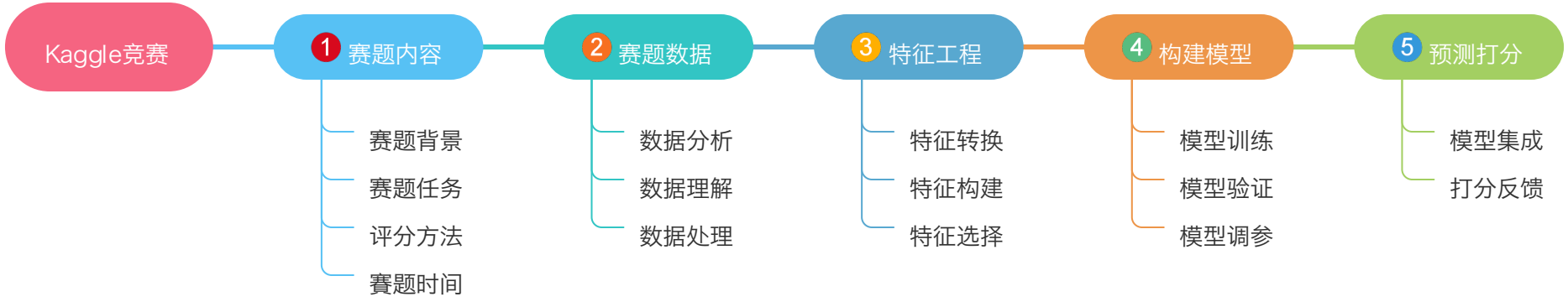
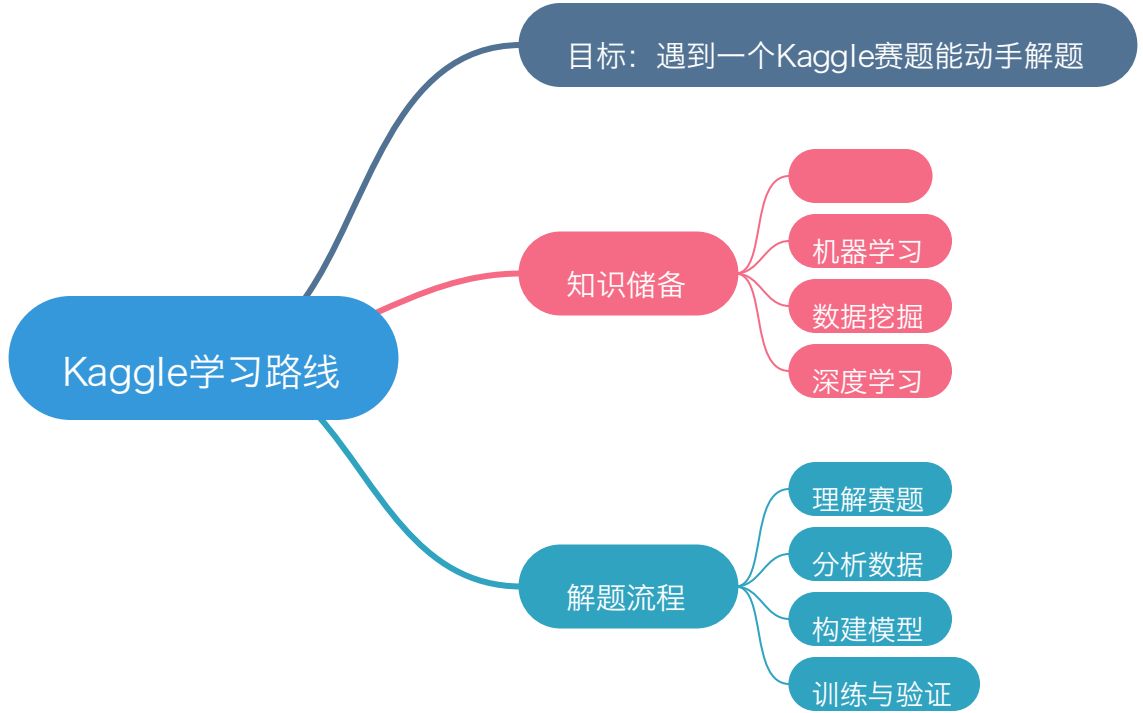

Kaggle比赛类型
比赛目的/形式
Featured
Research
Playground
Getting Started
Other Types
Recruitment
Annual
Limited
InClass
比赛阶段/提交
One-stage
Two-stage
Kernel
比赛内容
Machine Learning
Data Visualization


Kaggle竞赛
1 赛题内容
赛题背景
赛题任务
评分方法
赛题时间
2 赛题数据
数据分析
数据理解
数据处理
3 特征工程
特征转换
特征构建
特征选择
4 构建模型
模型训练
模型验证
模型调参
5 预测打分
模型集成
打分反馈


Kaggle竞赛
Overview：比赛介绍
Data：比赛数据
Notebooks：比赛Notebook
Discussion：比赛讨论
Leaderboard：比赛排名
Rules：比赛规则
Team：比赛队伍


Kaggle竞赛
是否熟悉赛题数据
是：下载数据开始建模
否：看Notebooks和Discussion
是否了解类似任务
是：寻找类似解决方法
否：re- search


天池NLP赛题
比赛思路：文本分类
迭代模型1：TF-IDF提取特征，SVM进行分类
迭代模型2：FastText训练词向量，并进行分类
迭代模型3：Word2Vec训练词向量，TextCNN进行分类
迭代模型4：Bert词向量并进行分类
最终模型：使用Bert分类 + 统计特征的树模型


Kaggle学习路线
目标：遇到一个Kaggle赛题能动手解题
知识储备
机器学习
数据挖掘
深度学习
解题流程
理解赛题
分析数据
构建模型
训练与验证

- Kaggle知识点
  - EDA
    - 读取并分析数据质量
    - 探索性分析每个变量
      - 变量是什么类型
      - 变量是否有缺失值
      - 变量是否有异常值
      - 变量是否有重复值
      - 变量是否均匀
      - 变量是否需要转换
    - 探索性分析变量与target标签的关系
      - 变量与标签是否存在相关性
      - 变量与标签是否存在业务逻辑
    - 变量与标签是否存在业务逻辑
      - 连续型变量与连续型变量
        - 可视化：散点图、相关性热力图
        - 皮尔逊系数
        - 互信息
      - 离散变量与离散变量
        - 可视化：柱状图、饼图、分组表
        - 卡方检验
      - 检查变量之间的正态性
        - 直方图
        - 箱线图
        - Quantile–Quantile (QQ图)
    - EDA结论
      - 变量是否需要筛选、替换和清洗
      - 变量是否需要转换
      - 变量之间是否需要交叉
      - 变量是否需要采样
  - 特征工程
    - 特征清洗
      - 离群点检测
      - 不均衡采样
    - 特征预处理
      - 特征缩放
        - Min Max Scaler
        - Standard Scaler
        - Max Abs Scaler
        - Robust Scaler
        - Quantile Transformer Scaler
        - Power Transformer Scaler
        - Unit Vector Scaler
      - 缺失值填充 Imputation
        - Numerical Imputation
        - Categorical Imputation
      - 特征编码
        - 类别特征 Categorical
        - 数值特征
          - Counts
          - Binarization
          - Rounding
          - Interactions
          - Binning
          - Statistical Transformations
            - Log Transform
            - Box–Cox Transform
        - 日期特征
          - 时间特征
            - 连续时间
              - 持续时间
              - 间隔时间
            - 离散时间
              - 年、季度、季节、月、星期、日、时
              - 节假日、节点日间隔
              - 小时时间段（上午、下午、晚上）
              - 高峰时间段
          - 时序历史特征
            - 统计值
              - 四分位数
              - 中位数
              - 平均数
              - 偏度、峰度
              - 离散系数
            - 同期值
              - 与同期值相差
              - 与同期值相比
        - 文本特征
          - 文本清洗
            - Lowercasing
            - Unidecode
          - Tokenizing
            - Tokenize
            - N–Grams
          - Removing
            - Stopwords
            - Rare words
            - Common words
          - Enrich
            - Document features
            - Entity insertion
            - Parse Trees
          - Similarities
            - Token similarity
            - Compression distance
            - Levenshtein/Hamming/Jaccard Distance
            - Word2Vec / Glove
          - TF–IDF
            - Term Frequency
            - Inverse Document Frequency
            - TF–IDF
        - 图像特征
          - 底层特征
            - 颜色分布（颜色直方图）
            - 边缘信息
            - 关键点信息
          - 高层特征
            - CNN特征
            - 图像语义标签
          - 图像文件特征
            - 创建时间
            - 图片MD5
    - 特征筛选
      - Filter
      - Wrapper
      - Embedded