大数据建模大赛辅导实战

【课程目标】

本课程主要面向专业人士的大数据建模竞赛辅导需求（假定学员已经完成Python 建模及优化--回归篇/分类篇的学习）。

通过本课程的学习，达到如下目的：

1、 熟悉大赛常用集成模型

2、 掌握模型优化常用措施，掌握超参优化策略

3、 掌握特征工程处理，以及对模型质量的影响

4、 掌握建模工程管道类(Pipeline, ColumnTransformer)的使用

【授课时间】

2-3 天时间，大致内容安排（会根据需求和学员水平调整进度）

时间	主题	具体内容	目的
第一天 上午	建模流程	建模步骤 模型评估指标 模型基本原理	常用建模步骤，构建通用 common 模型，完成模型训练、评估等的封装
第一天 下午	数据清洗	数据清洗 缺失值填充	1、 理解异常数据对模型的影响 2、 缺失值常用的填充方式(固定值、分类填充、拉格朗日、预测填充) 3、 不同填充对模型的影响
第二天 上午	特征选择	特征选择模式(Filter/Wrapper/Embedded)	1、 特征选择的封装实现 2、 优缺点及应用场景(SelectKBest,REF,SelectFromModel)

第二天 下午	变量降维	因子分析 主成份分析 管道实现	1、 变量降维 PCA/FA 2、 掌握管道处理技能 (Pipeline, FeatureUnion, ColumnsTransformer 等)
第三天 上午	变量变换	变量派生 变量标准化 模型集成思想	1、 利用探索性分析，指导变量派生 2、 不同标准化对模型的影响 3、 特征处理的不同顺序对模型效果的影响
第 三 天 下午	超参优化	超参优化方法 欠拟合优化 过拟合优化 其 它 优 化 (性能、样本均衡处理)	1、 不同超参的作用(欠拟合/过拟合) 2、 超参优化方法 3、 超参优化策略 4、 Stacking 集成

【授课对象】

参加大数据建模大赛的 IT 专业人士。

要求精通 Python 语言，熟悉 sklearn 库的基本使用等。

【授课方式】

理论框架 + 落地措施 + 实战训练

【课程大纲】

第一部分： 常用集成模型

问题：数据建模的基本步骤是什么？每一步要重点考虑哪些知识和技能？

1、 数据建模六步法

- 选择模型：基于业务选择恰当的数据模型
- 特征工程：选择对目标变量有显著影响的属性来建模
- 训练模型：采用合适的算法，寻找到最合适的模型参数
- 评估模型：进行评估模型的质量，判断模型是否可用
- 优化模型：如果评估结果不理想，则需要对模型进行优化
- 应用模型：如果评估结果满足要求，则可应用模型于业务场景

2、模型集成思想

- Bagging
- Boosting
- Stacking

3、竞赛常用的集成模型

- RandomForest
- Adaboosting/GBDT/XGBoost

4、各模型的原理及适用场景

第二部分： 数据清洗技巧

1、数据清洗处理

- 重复值
- 错误值
- 离群值

- 缺失值

2、缺失值填充的常见方式

- 固定值填充
- 同类别均值填充
- 相邻值填充(向下/向上填充)
- 两点插值（相邻值均值填充）
- 拉格朗日插值
- 预测方法填充

3、不同填充方式对模型效果的影响

案例：泰坦尼克号沉船幸存者预测

第三部分： 特征选择模式

1、降维的两大方式：特征选择与因子合并

2、特征选择的三种模式

3、基于变量本身的重要性筛选

- 缺失值所占比例过大
- 标准差/变异系数过小（VarianceThreshold）

- 类别值比值失衡严重
- 类别值与样本量比例过大

4、Filter 式(特征选择与模型分离)

- 常用评估指标(相关系数/显著性/互信息等)
- f_regression, f_classif, chi2,
- mutual_info_regression, mutual_info_classif

案例：运营商流失预测的特征选择

5、Wrapper 式(利用模型结果进行特征选择)

- Sklearn 实现（RFE/RFECV-Recursive Feature Elimination)

6、Embedded 式(模型自带特征选择功能)

- L1 正则项(Lasso/ElasticNet)
- 信息增益(决策树)
- Sklearn 实现（SelectFromModel）

7、不同模式的优缺点及应用场景

8、特征选择的变量个数

第四部分： 特征合并方法

1、特征合并与特征选择

2、因子分析（FactorAnalysis）

- FA 原理及思想
- 载荷矩阵相关概念(变量共同度/方差贡献率)
- 如何确定降维的因子个数

3、主成份分析（Principal Component Analysis）

- PCA 原理
- PCA 的几何意义

案例：汽车油效预测

第五部分： 变量变换影响

1、为何需要变换变换

- 假设条件需求，可比性需要，同权重需要

2、因变量变换对模型质量的影响

案例：波士顿房价预测

3、特征标准化

- 标准化的作用: 缩小，消除/统一量纲
- 常用标准化方法:MinMaxScaler, StandardScaler,…

- 不同模型对标准化的要求
- 不同标准化对模型的影响

案例：医院肿瘤预测

4、其它变换：正态化、正则化

5、变量派生：多项式等

案例：用户收入预测

6、管道实现，简化代码

- 管道类 Pipeline
- 列转换类 ColumnTransformer
- 特征合并类 FeatureUnion

第六部分： XGBoost 模型详解及优化

1、基本参数配置

- 框架基本参数: n_estimators, objective
- 性能相关参数: learning_rate
- 模型复杂度参数:max_depth,min_child_weight,gamma
- 生长策略参数: grow_policy, tree_method, max_bin

- 随机性参数：subsample,colsample_bytree
- 正则项参数:reg_alpha,reg_lambda
- 样本不均衡参数: scale_pos_weight

2、早期停止与基类个数优化（n_estimators、early_stopping_rounds）

3、样本不平衡处理

- 欠抽样与过抽样
- scale_pos_weight= neg_num/pos_num

4、XGBoost 模型欠拟合优化措施

- 增维，派生新特征
    a) 非线性检验
    b) 相互作用检验
- 降噪，剔除噪声数据
    a) 剔除不显著影响因素
    b) 剔除预测离群值（仅回归）
    c) 多重共线性检验（仅回归）
- 变量变换

a) 自变量标准化

b) 残差项检验与因变量变换

- 增加树的深度与复杂度

a) 增大 max_depth

b) 减小 min_child_weight, gamma 等

- 禁止正则项生效

5、特征重要性评估与自动特征选择

6、超参优化策略：

- 分组调参：参数分组分别调优
- 分层调参：先粗调再细调

7、XGBoost 模型过拟合优化措施

- 降维，减少特征数量
- 限制树的深度和复杂度

a) 减小 max_depth

b) 增大 min_child_weight，gamma 等

- 采用 dart 模型来控制过拟合(引入 dropout 技术)
- 启用正则项惩罚:reg_alpha,reg_lambda 等

- 启用随机采样:subsample,colsample_bytree 等

8、Stacking 模式：XGBoost+LR、XGBoost+RF 等

9、XGBoost 的优化模型：LightGBM

第七部分： 实战训练篇

1、互联网广告判断模型

2、客户流失预测模型

3、直销响应模型

结束：课程总结与问题答疑。