大数据建模与模型优化实战培训

【课程目标】

本课程为建模课程，面向数据分析部等专门负责数据分析与建模的人员。

本课程的主要目的是，帮助学员掌握大数据建模基础知识，帮助学员构建系统全面的预测建模思维，提升学员的数据建模综合能力。

本课程具体内容包括：

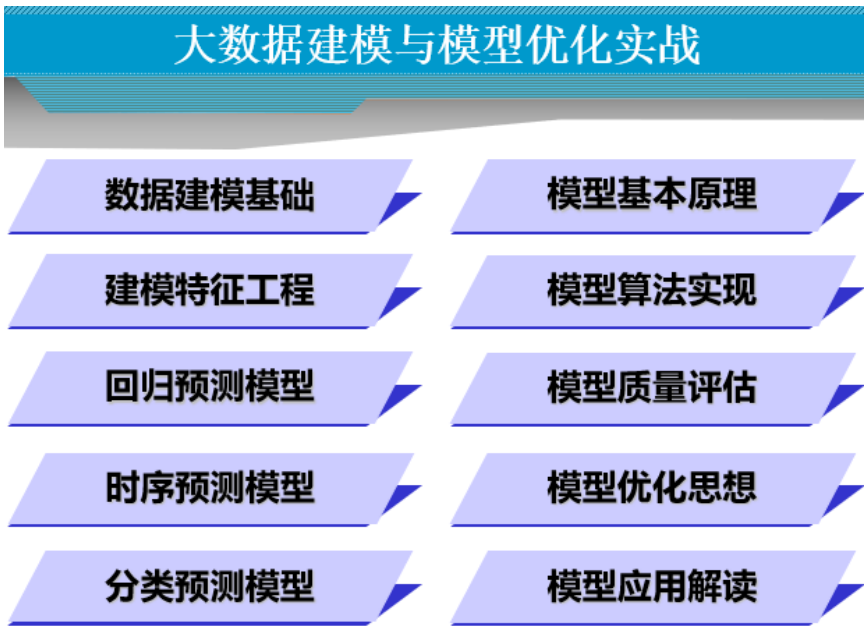
1、 数据建模流程，特征工程处理

2、 回归预测模型，时序预测模型

3、 分类预测模型，模型含义解读

4、 模型基本原理，模型算法实现

5、 模型质量评估，模型优化措施



本系列课程从实际的业务需求出发，结合行业的典型应用特点，围绕实际的商业问题，对数据预测建模的过程进行了全面的介绍（从模型选择，到特征选择，再到训练模型，评估模型，以及优化模型和模型解读），通过大量的操作演练，帮助学员掌握数据建模的思路、方法、技巧，以提升学员的数据建模的能力，支撑运营决策的目的。

通过本课程的学习，达到如下目的：

1、 掌握数据建模的基本过程和步骤

2、 掌握数据建模前的特征选择的系统方法，学会寻找影响业务的关键要素

3、 掌握回归预测模型基本原理，学会解读回归方程的含义

4、 掌握常用的时序预测模型，以及各模型的适用场景

5、 掌握常用的分类预测模型，以及分类模型的优化

【授课时间】

2-4 天时间（每天 6 个小时）

【授课对象】

业务支撑、网络中心、IT 系统部、数据分析部等对业务数据分析有较高要求的相关专业人员。

【学员要求】

1、 每个学员自备一台便携机(必须)。

2、 便携机中事先安装好 Office Excel 2013 版本及以上。

3、 便携机中事先安装好 IBM SPSS Statistics v24 版本以上软件。

注：讲师可以提供试用版本软件及分析数据源。

【授课方式】

理论精讲 + 案例演练 + 实际业务问题分析 + SPSS 实际操作

本课程突出数据挖掘的实际应用，结合行业的典型应用特点，从实际问题

入手，引出相关知识，进行大数据的收集与处理；探索数据之间的规律及关联性，帮助学员掌握系统的数据预处理方法；介绍常用的模型，训练模型，并优化模型，以达到最优分析结果。

【课程大纲】

第一部分：　　数据建模流程

1、预测建模六步法

- 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：采用合适的算法对模型进行训练，寻找到最优参数
- 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、数据挖掘常用的模型

- 定量预测模型：回归预测、时序预测等
- 定性预测模型：逻辑回归、决策树、神经网络、支持向量机等
- 市场细分：聚类、RFM、PCA 等

- 产品推荐：关联分析、协同过滤等
- 产品优化：回归、随机效用等
- 产品定价：定价策略/最优定价等

3、特征工程/特征选择/变量降维

- 基于变量本身特征
- 基于相关性判断
- 因子合并（PCA 等）
- IV 值筛选（评分卡使用）
- 基于信息增益判断（决策树使用）

4、模型评估

- 模型质量评估指标：R^2、正确率/查全率/查准率/特异性等
- 预测值评估指标：MAD、MSE/RMSE、MAPE、概率等
- 模型评估方法：留出法、K 拆交叉验证、自助法等
- 其它评估：过拟合评估、残差检验

5、模型优化

- 优化模型：选择新模型/修改模型
- 优化数据：新增显著自变量

- 优化公式：采用新的计算公式
- 集成思想：Bagging/Boosting/Stacking

6、常用预测模型介绍

- 时序预测模型
- 回归预测模型
- 分类预测模型

第二部分：　建模特征工程

问题：如何选择合适的属性/特征来建模呢？选择的依据是什么？比如价格是否可用于产品销量预测？

1、数据预处理 vs 特征工程

2、特征工程处理内容

- 变量变换
- 变量派生
- 变量精简（特征选择、因子合并）
- 类型转换

3、特征选择常用方法

- 相关分析、方差分析、卡方检验

4、相关分析（衡量两数据型变量的线性相关性）

- 相关分析简介
- 相关分析的应用场景
- 相关分析的种类
  - 简单相关分析
  - 偏相关分析
  - 距离相关分析
- 相关系数的三种计算公式
  - Pearson 相关系数
  - Spearman 相关系数
  - Kendall 相关系数
- 相关分析的假设检验
- 相关分析的四个基本步骤

演练：营销费用会影响销售额吗？影响程度如何量化？

演练：哪些因素与汽车销量有相关性

演练：影响用户消费水平的因素会有哪些

- 偏相关分析
  - 偏相关原理：排除不可控因素后的两变量的相关性
  - 偏相关系数的计算公式
  - 偏相关分析的适用场景
- 距离相关分析

5、方差分析（衡量类别变量与数值变量间的相关性）

- 方差分析的应用场景
- 方差分析的三个种类
  - 单因素方差分析
  - 多因素方差分析
  - 协方差分析
- 单因素方差分析的原理
- 方差分析的四个步骤
- 解读方差分析结果的两个要点

演练：摆放位置与销量有关吗

演练：客户学历对消费水平的影响分析

演练：广告和价格是影响终端销量的关键因素吗

演练：营业员的性别、技能级别对产品销量有影响吗

演练：寻找影响产品销量的关键因素

- 多因素方差分析原理
- 多因素方差分析的作用
- 多因素方差结果的解读

演练：广告形式、地区对销量的影响因素分析

- 协方差分析原理
- 协方差分析的适用场景

演练：排除产品价格，收入对销量有影响吗？

6、列联分析/卡方检验（两类别变量的相关性分析）

- 交叉表与列联表：计数值与期望值
- 卡方检验的原理
- 卡方检验的几个计算公式
- 列联表分析的适用场景

案例：套餐类型对客户流失的影响分析

案例：学历对业务套餐偏好的影响分析

案例：行业/规模对风控的影响分析

第三部分： 线性回归模型

营销问题：如何预测未来的产品销量/销售额？如果产品跟随季节性变动，该如何预测？

1、回归分析简介和原理

2、回归分析的种类

- 一元回归/多元回归
- 线性回归/非线性回归

3、常用回归分析方法

- 散点图+趋势线（一元）
- 线性回归工具（多元线性）
- 规划求解工具（非线性回归）

演练：散点图找营销费用与销售额的关系

4、线性回归分析的五个步骤

演练：营销费用、办公费用与销售额的关系（线性回归）

5、线性回归方程的解读技巧

- 定性描述：正相关/负相关
- 定量描述：自变量变化导致因变量的变化程度

6、回归预测模型评估

- 质量评估指标：判定系数 R^2
- 如何选择最佳回归模型

演练：如何选择最佳的回归预测模型（一元曲线回归）

7、带分类自变量的回归预测

演练：汽车季度销量预测

演练：工龄、性别与终端销量的关系

演练：如何评估销售目标与资源最佳配置

8、自动筛选不显著因素（自变量）

第四部分： 回归模型优化

1、回归分析的基本原理

- 三个基本概念：总变差、回归变差、剩余变差
- 方程的显著性检验：方程可用性
- 因素的显著性检验：因素可用性
- 方程拟合优度检验：质量好坏程度
- 理解标准误差含义：预测准确性？

2、回归模型优化措施：寻找最佳回归拟合线

- 如何处理预测离群值（剔除离群值）
- 如何剔除不显著因素（剔除不显著因素）
- 如何进行非线性关系检验（增加非线性自变量）
- 如何进行相互作用检验（增加相互作用自变量）
- 如何进行多重共线性检验（剔除共线性自变量）

演练：模型优化演示

3、好模型都是优化出来的

第五部分： 自定义回归模型

1、回归建模的本质

2、规划求解工具简介

3、自定义回归模型

案例：如何对客流量进行建模预测及模型优化

4、回归季节预测模型模型

- 回归季节模型的原理及应用场景
- 加法季节模型

- 乘法季节模型
- 模型解读

案例：美国航空旅客里程的季节性趋势分析

5、新产品累计销量的 S 曲线

- S 曲线模型的应用场景（最大累计销量及销量增长的拐点）
- 珀尔曲线
- 龚铂兹曲线

案例：如何预测产品的销售增长拐点，以及销量上限

演练：预测 IPad 产品的销量

第六部分： 定量模型评估

1、 定量预测模型的评估

- 方程显著性评估
- 因素显著性评估
- 拟合优度的评估
- 估计标准误差评估
- 预测值准确度评估

2、 模型拟合度评估

- 判定系数： R^2
- 调整判定系数： $\acute{R}^2$

3、 预测值准确度评估

- 平均绝对误差：MAE
- 根均方差：RMSE
- 平均误差率：MAPE

4、 其它评估：残差检验、过拟合检验

第七部分： 时序预测模型

营销问题：像利率/CPI/GDP 等按时序变化的指标如何预测？当销量随季节周期变动时该如何预测？

1、回归预测 vs 时序预测

2、因素分解思想

3、时序预测常用模型

- 趋势拟合
- 季节拟合

- 平均序列拟合

4、评估预测值的准确度指标：MAD、RMSE、MAPE

5、移动平均（MA）

- 应用场景及原理
- 移动平均种类
  - 一次移动平均
  - 二次移动平均
  - 加权移动平均
  - 移动平均比率法
- 移动平均关键问题
  - 如何选取最优参数 N
  - 如何确定最优权重系数

演练：平板电脑销量预测及评估

演练：快销产品季节销量预测及评估

6、指数平滑（ES）

- 应用场景及原理
- 最优平滑系数的选取原则

- 指数平滑种类
  - 一次指数平滑
  - 二次指数平滑（Brown 线性、Holt 线性、Holt 指数、阻尼线性、阻尼指数）
  - 三次指数平滑

演练：煤炭产量预测

演练：航空旅客量预测及评估

7、温特斯季节预测模型

- 适用场景及原理
- Holt-Winters 加法模型
- Holt-Winters 乘法模型

演练：汽车销量预测及评估

8、平稳序列模型（ARIMA）

- 序列的平稳性检验
- 平稳序列的拟合模型
  - AR(p)自回归模型
  - MA(q)移动模型

- ARMA(p,q)自回归移动模型

- 模型的识别与定阶
  - ACF 图/PACF 图
  - 最小信息准则
- 序列平稳化处理
  - 变量变换
  - k 次差分
  - d 阶差分
- ARIMA(p,d,q)模型

  演练：上海证券交易所综合指数收益率序列分析

  演练：服装销售数据季节性趋势预测分析
- 平稳序列的建模流程

第八部分： 分类预测模型

问题：如何评估客户购买产品的可能性？如何预测客户的购买行为？如何提取某类客户的典型特征？如何向客户精准推荐产品或业务？

1、分类模型概述及其应用场景

2、常见分类预测模型

3、逻辑回归（LR）

- 逻辑回归的适用场景
- 逻辑回归的模型原理
- 逻辑回归分类的几何意义
- 逻辑回归的种类
  - 二项逻辑回归
  - 多项逻辑回归
- 如何解读逻辑回归方程
- 带分类自变量的逻辑回归分析
- 多项逻辑回归/多分类逻辑回归

案例：如何评估用户是否会购买某产品（二项逻辑回归）

案例：多品牌选择模型分析（多项逻辑回归）

4、分类决策树（DT）

问题：如何预测客户行为？如何识别潜在客户？

风控：如何识别欠贷者的特征，以及预测欠贷概率？

客户保有：如何识别流失客户特征，以及预测客户流失概率？

- 决策树分类简介

案例：美国零售商（Target）如何预测少女怀孕

演练：识别银行欠货风险，提取欠贷者的特征

- 决策树分类的几何意义
- 构建决策树的三个关键问题
  - 如何选择最佳属性来构建节点
  - 如何分裂变量
  - 修剪决策树
- 选择最优属性生长
  - 熵、基尼索引、分类错误
  - 属性划分增益
- 如何分裂变量
  - 多元划分与二元划分
  - 连续变量离散化（最优分割点）
- 修剪决策树
  - 剪枝原则
  - 预剪枝与后剪枝
- 构建决策树的四个算法

- C5.0、CHAID、CART、QUEST
- 各种算法的比较

- 如何选择最优分类模型？

  案例：商场用户的典型特征提取

  案例：客户流失预警与客户挽留

  案例：识别拖欠银行货款者的特征，避免不良货款

  案例：识别电信诈骗者嘴脸，让通信更安全

- 多分类决策树

  案例：不同套餐用户的典型特征

- 决策树模型的保存与应用

5、人工神经网络（ANN）

- 神经网络概述
- 神经网络基本原理
- 神经网络的结构
- 神经网络分类的几何意义
- 神经网络的建立步骤
- 神经网络的关键问题

- BP 反向传播网络（MLP）
- 径向基网络（RBF）

案例：评估银行用户拖欠货款的概率

6、判别分析（DA）

- 判别分析原理
- 判别分析种类
- Fisher 线性判别分析

案例：MBA 学生录取判别分析

案例：上市公司类别评估

7、最近邻分类（KNN）

- KNN 模型的基本原理
- KNN 分类的几何意义
- K 近邻的关键问题

8、支持向量机（SVM）

- SVM 基本原理
- 线性可分问题：最大边界超平面
- 线性不可分问题：特征空间的转换

- 维灾难与核函数

9、贝叶斯分类（NBN）

- 贝叶斯分类原理
- 计算类别属性的条件概率
- 估计连续属性的条件概率
- 预测分类概率（计算概率）
- 拉普拉斯修正

案例：评估银行用户拖欠货款的概率

第九部分： 定性模型评估

1、 模型的评估指标

- 两大矩阵：混淆矩阵，代价矩阵
- 六大指标：Acc,P,R,Spec,F1,lift
- 三条曲线：
  - ROC 曲线和 AUC
  - PR 曲线和 BEP
  - KS 曲线和 KS 值

2、 模型的评估方法

- 原始评估法
- 留出法（Hold-Out）
- 交叉验证法（k-fold cross validation）
- 自助采样法（Bootstrapping）

第十部分： 模型集成优化

1、 模型的优化思路

2、 集成算法基本原理

- 单独构建多个弱分类器
- 多个弱分类器组合投票，决定预测结果

3、 集成方法的种类

- Bagging
- Boosting
- Stacking

4、 Bagging 集成

- 数据/属性重抽样
- 决策依据：少数服从多数
- 典型模型：随机森林 RF

5、 Boosting 集成

- 基于误分数据建模
- 样本选择权重更新公式
- 决策依据：加权投票
- 典型模型：AdaBoost 模型

6、 其它高级集成算法：GBDT，XGBoost 等

结束：课程总结与问题答疑。