金融行业风险预测模型实战

【课程目标】

本课程专注于金融行业的风控模型，面向数据分析部等专门负责数据分析与建模的人士。

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、 掌握数据建模前的属性筛选的系统方法，为建模打下基础。

3、 掌握常用的分类预测模型，包括逻辑回归、决策树、神经网络、判别分析等等，以及分类模型的优化。

4、 掌握金融行业信用评分卡模型，构建信用评分模型。

主要内容包括数据建模的过程和步骤，以及建模涉及到的分析方法、分析模型，以及模型优化等。

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。

【授课时间】

2-3 天时间

【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】

基础知识精讲 + 案例演练 + 实际业务问题分析 + SPSS 实际操作

【课程大纲】

第一部分： 数据建模基本过程

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
- 属性筛选：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、 数据挖掘常用的模型

- 数值预测模型：回归预测、时序预测等
- 分类预测模型：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、 属性筛选/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

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

6、模型实现算法（暂略）

7、好模型是优化出来的

案例：通信客户流失分析及预警模型

第二部分： 属性筛选方法

问题：如何选择合适的属性来进行建模预测？
比如：价格是否可用于产品销量的预测？套餐的合理性是否会影响客户流失？在欺诈风险中有哪些数据会有异常表现？

1、属性筛选/变量降维的常用方法

- 基于变量本身特征来选择属性
- 基于数据间的相关性来选择属性
- 基于因子合并（如 PCA 分析）实现变量的合并
- 利用 IV 值筛选
- 基于信息增益来选择属性

2、相关分析（衡量变量间的线性相关性）

问题：这两个属性是否会相互影响？影响程度大吗？

- 相关分析简介
- 相关分析的三个种类
  - 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：年龄和收入的相关分析

演练：营销费用会影响销售额吗

演练：工作时间与收入有相关性吗

演练：话费与网龄的相关分析

- 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

3、方差分析(衡量类别变量与数据变量的相关性)

问题：哪些才是影响销量的关键因素？

- 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
- 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：用户收入对银行欠贷的影响分析

演练：家庭人数对银行欠贷的影响分析

演练：年龄大小对欠贷有影响吗

演练：寻找影响贷款风险的关键因素

- 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（多因素）

- 协方差分析原理
- 协方差分析的适用场景

  演练：饲料对生猪体重的影响分析（协方差分析）

4、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
- 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

演练：不同的信用卡类型会有不同欠贷风险吗

演练：有无住房对欠贷的影响分析

案例：行业/规模对风控的影响分析

5、相关性分析各种方法的适用场景

6、主成份分析（PCA）

- 因子分析的原理
- 因子个数如何选择
- 如何解读因子含义

案例：提取影响电信客户流失的主成分分析

第三部分： 回归预测模型篇

问题：如何预测产品的销量/销售金额？如果产品跟随季节性变动，该如何预测？

新产品上市，如果评估销量上限及销售增速？

1、常用的数值预测模型

- 回归预测
- 时序预测

2、回归预测/回归分析

问题：如何预测未来的销售量（定量分析）？

- 回归分析的基本原理和应用场景
- 回归分析的种类（一元/多元、线性/曲线）
- 得到回归方程的四种常用方法
  - Excel 函数

- 散点图+趋势线
- 线性回归工具
- 规范求解

- 线性回归分析的五个步骤
- 回归方程结果的解读要点
- 评估回归模型质量的常用指标
- 评估预测值的准确度的常用指标

演练：散点图找推广费用与销售额的关系（一元线性回归）

演练：推广费用、办公费用与销售额的关系（多元线性回归）

演练：让你的营销费用预算更准确

演练：如何选择最佳的回归预测模型（曲线回归）

- 带分类变量的回归预测

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配置（营业厅）

3、自动筛选不显著自变量

第四部分： 回归预测模型优化篇

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：是否可以做回归分析？
- 因素的显著性检验：自变量是否可用？
- 拟合优度检验：回归模型的质量评估？
- 理解标准误差的含义：预测的准确性？

2、 回归模型优化思路：寻找最佳回归拟合线

- 如何处理预测离群值（剔除离群值）
- 如何剔除非显著因素（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）
- 如何检验误差项（修改因变量）
- 如何判断模型过拟合（模型过拟合判断）

案例：模型优化案例

3、 规划求解工具简介

4、 自定义回归模型（如何利用规划求解进行自定义模型）

案例：如何对餐厅客流量进行建模及模型优化

5、 好模型都是优化出来的

第五部分： 分类预测模型

问题：如何评估客户购买产品的可能性？或者说，影响客户购买意向的产品关键特性是什么？

1、 分类预测模型概述

2、 常见分类预测模型

3、 评估分类模型的常用指标

- 正确率、查全率/查准率、特异性等

4、 逻辑回归分析模型（LR）

问题：如果评估用户是否购买产品的概率？

- 逻辑回归模型原理及适用场景
- 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

案例：如何评估用户是否会有违约风险（二项逻辑回归）

案例：多品牌选择模型分析（多项逻辑回归）

5、决策树分类（DT）

问题：如何提取客户流失者、拖欠货款者的特征？如何预测其流失的概率？

- 决策树分类的原理
- 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
  - 如何修剪决策树
- 选择最优属性
  - 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（最优划分点）

- 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：识别银行欠货风险，提取欠货者的特征

案例：客户流失预警与客户挽留模型

6、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

7、判别分析（DA）

- 判别分析原理
- 距离判别法
- 典型判别法
- 贝叶斯判别法

案例：MBA 学生录取判别分析

案例：上市公司类别评估

8、最近邻分类（KNN）

- 基本原理
- 关键问题

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
- 贝叶斯网络种类：TAN/马尔科夫毯

- 预测分类概率（计算概率）

案例：评估银行用户拖欠货款的概率

10、 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
- 维空难与核函数

第六部分： 分类模型优化篇（集成方法）

1、分类模型的优化思路：利用弱分类器构建强分类模型

2、集成方法的基本原理

- 选取多个数据集，构建多个弱分类器
- 多个弱分类器投票决定

3、集成方法/元算法的种类

- Bagging 算法
- Boosting 算法

4、Bagging 原理

- 如何选择数据集
- 如何进行投票
- 随机森林

5、 Boosting 的原理

- AdaBoost 算法流程
- 样本选择权重计算公式
- 分类器投票权重计算公式

第七部分： 银行信用评分卡模型

1、信用评分卡模型简介

2、评分卡的关键问题

3、信用评分卡建立过程

- 筛选重要属性
- 数据集转化
- 建立分类模型
- 计算属性分值
- 确定审批阈值

4、筛选重要属性

- 属性分段
- 基本概念：WOE、IV
- 属性重要性评估

5、数据集转化

- 连续属性最优分段
- 计算属性取值的 WOE

6、建立分类模型

- 训练逻辑回归模型
- 评估模型
- 得到字段系数

7、计算属性分值

- 计算补偿与刻度值
- 计算各字段得分
- 生成评分卡

8、确定审批阈值

- 画 K-S 曲线

- 计算 K-S 值
- 获取最优阈值

案例：构建银行小额贷款的用户信用模型

第八部分： 数据预处理篇（了解你的数据集）

1、 数据预处理的主要任务

- 数据集成：多个数据集的合并
- 数据清理：异常值的处理
- 数据处理：数据筛选、数据精简、数据平衡
- 变量处理：变量变换、变量派生、变量精简
- 数据归约：实现降维，避免维灾难

2、 数据集成

- 外部数据读入：Txt/Excel/SPSS/Database
- 数据追加（添加数据）
- 变量合并（添加变量）

3、 数据理解（异常数据处理）

- 取值范围限定

- 重复值处理
- 无效值/错误值处理
- 缺失值处理
- 离群值/极端值处理
- 数据质量评估

4、 数据准备：数据处理

- 数据筛选：数据抽样/选择（减少样本数量）
- 数据精简：数据分段/离散化（减少变量的取值个数）
- 数据平衡：正反样本比例均衡

5、 数据准备：变量处理

- 变量变换：原变量取值更新，比如标准化
- 变量派生：根据旧变量生成新的变量
- 变量精简：降维，减少变量个数

6、 数据降维

- 常用降维的方法
- 如何确定变量个数
- 特征选择：选择重要变量，剔除不重要的变量

- 从变量本身考虑
- 从输入变量与目标变量的相关性考虑
- 对输入变量进行合并

- 因子分析（主成分分析）
  - 因子分析的原理
  - 因子个数如何选择
  - 如何解读因子含义

案例：提取影响电信客户流失的主成分分析

7、 数据探索性分析

- 常用统计指标分析
- 单变量：数值变量/分类变量
- 双变量：交叉分析/相关性分析
- 多变量：特征选择、因子分析

演练：描述性分析（频数、描述、探索、分类汇总）

8、 数据可视化

- 数据可视化：柱状图、条形图、饼图、折线图、箱图、散点图等
- 图形的表达及适用场景

演练：各种图形绘制

第九部分： 数据建模实战篇

1、 电信业客户流失预警和客户挽留模型实战

2、 银行欠贷风险预测模型实战

3、 银行信用卡评分模型实战

结束：课程总结与问题答疑。