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。

本课程的主要目的是，培养学员的大数据意识和大数据思维，掌握常用的数据分析方法和数据分析模型，并能够用于对客户行为作分析和预测，提升学员的数据分析综合能力。

通过本课程的学习，达到如下目的：

1、 掌握数据分析和数据建模的基本过程和步骤

2、 掌握客户行为分析中常用的分析方法

3、 掌握业务的影响因素分析常用的方法

4、 掌握常用客户行为预测模型，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

5、 掌握金融行业信用评分卡模型，构建信用评分模型

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（每天 6 个小时）

【授课对象】

风险控制部、业务支撑、网络中心、IT 系统部、数据分析部等对数据建模有较高要求的相关领域人员。

【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

【课程大纲】

第一部分： 数据核心理念—数据思维篇

问题：什么是数据思维？大数据决策的底层逻辑以及决策依据是什么？

1、数字化五大技术战略：ABCDI 战略

- A：人工智能，目的是用机器模拟人类行为
- B：区块链，构建不可篡改的分布记账系统
- C：云计算，搭建按需分配的计算资源平台
- D：大数据，实现智能化的判断和决策机制

- I：物联网，实现万物互联通信的基础架构

2、大数据的本质

- 数据，是事物发展和变化过程中留下的痕迹
- 大数据不在于量大，而在于全（多维性）
- 业务导向还是技术导向

3、大数据决策的底层逻辑（即四大核心价值）

- 探索业务规律，按规律来管理决策

  案例：客流规律与排班及最佳营销时机

  案例：致命交通事故发生的时间规律

- 发现运营变化，定短板来运营决策

  案例：考核周期导致的员工月初懈怠

  案例：工序信号异常监测设备故障

- 理清要素关系，找影响因素来决策

  案例：情绪对于股市涨跌的影响

  案例：为何升职反而会增加离职风险？

- 预测未来趋势，通过预判进行决策

  案例：惠普预测员工离职风险及挽留

案例：保险公司的车险预测与个性化保费定价

4、 大数据决策的三个关键环节

- 业务数据化：将业务问题转化为数据问题
- 数据信息化：提取数据中的业务规律信息
- 信息策略化：基于规律形成业务应对策略

案例：用数据来识别喜欢赚“差价”的营业员

第二部分： 数据分析基础—流程步骤篇

1、 数据分析的六步曲

2、 步骤 1：明确目的，确定分析思路

- 确定分析目的：要解决什么样的业务问题
- 确定分析思路：分解业务问题，构建分析框架

3、 步骤 2：收集数据，寻找分析素材

- 明确数据范围
- 确定收集来源
- 确定收集方法

4、 步骤 3：整理数据，确保数据质量

- 数据质量评估
- 数据清洗、数据处理和变量处理
- 探索性分析

5、步骤 4：分析数据，寻找业务答案

- 选择合适的分析方法
- 构建合适的分析模型
- 选择合适的分析工具

6、步骤 5：呈现数，解读业务规律

- 选择恰当的图表
- 选择合适的可视化工具
- 提炼业务含义

7、步骤 6：撰写报告，形成业务策略

- 选择报告种类
- 完整的报告结构

演练：产品精准营销案例分析

- 如何搭建精准营销分析框架
- 精准营销分析的过程和步骤

第三部分： 用户行为分析—统计方法篇

问题：数据分析方法的种类？分析方法的不同应用场景？

1、业务分析的三个阶段

- 现状分析：通过企业运营指标来发现规律及短板
- 原因分析：查找数据相关性，探寻目标影响因素
- 预测分析：合理配置资源，预判业务未来的趋势

2、常用的数据分析方法种类

- 描述性分析法（对比/分组/结构/趋势/交叉…）
- 相关性分析法（相关/方差/卡方…）
- 预测性分析法（回归/时序/决策树/神经网络…）
- 专题性分析法（聚类/关联/RFM 模型/...）

3、统计分析基础

- 统计分析两大关键要素（类别、指标）
- 统计分析的操作模式（类别□指标）
- 统计分析三个操作步骤（统计、画图、解读）
- 透视表的三个组成部分

4、常用的描述性指标

- 集中程度：均值、中位数、众数
- 离散程度：极差、方差/标准差、IQR
- 分布形态：偏度、峰度

5、基本分析方法及其适用场景

- 对比分析（查看数据差距，发现事物变化）

演练：寻找用户的地域分布特征

演练：分析产品受欢迎情况及贡献大小

演练：用数据来探索增量不增收困境的解决方案

- 分布分析（查看数据分布，探索业务层次）

演练：银行用户的消费水平和消费层次分析

演练：客户年龄分布/收入分布分析

案例：通信运营商的流量套餐划分合理性的评估

演练：呼叫中心接听电话效率分析（呼叫中心）

- 结构分析（查看指标构成，评估结构合理性）

案例：增值业务收入结构分析（通信）

案例：物流费用成本结构分析（物流）

案例：中移动用户群动态结构分析

演练：财务领域的结构瀑布图、财务收支的变化瀑布图

- 趋势分析（发现事物随时间的变化规律）

案例：破解零售店销售规律

案例：手机销量的淡旺季分析

案例：微信用户的活跃时间规律

演练：发现客流量的时间规律

- 交叉分析（从多个维度的数据指标分析）

演练：用户性别+地域分布分析

演练：不同客户的产品偏好分析

演练：不同学历用户的套餐偏好分析

演练：银行用户的违约影响因素分析

第四部分： 用户行为分析—分析框架篇

问题：如何才能全面/系统地分析而不遗漏？如何分解和细化业务问题？

1、业务分析思路和分析框架来源于业务模型

2、常用的业务模型

- 外部环境分析：PEST

- 业务专题分析：5W2H
- 竞品/竞争分析：SWOT、波特五力
- 营销市场专题分析：4P/4C 等

3、用户行为分析（5W2H 分析思路和框架）

- WHY：原因（用户需求、产品亮点、竞品优劣势）
- WHAT：产品（产品喜好、产品贡献、产品功能、产品结构）
- WHO：客户（基本特征、消费能力、产品偏好）
- WHEN：时间（淡旺季、活跃时间、重购周期）
- WHERE：区域/渠道（区域喜好、渠道偏好）
- HOW：支付/促销（支付方式、促销方式有效性评估等）
- HOW MUCH：价格（费用、成本、利润、收入结构、价格偏好等）

案例讨论：结合公司情况，搭建用户消费习惯的分析框架（5W2H）

第五部分：　数据建模基础—流程步骤篇

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

第六部分：　影响因素分析—根因分析篇

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

第七部分： 客户行为预测—分类模型篇

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

第八部分： 客户行为预测—模型评估篇

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）

第九部分： 客户行为预测—集成优化篇

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

第十部分： 银行客户信用卡模型

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

第十一部分： 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。