大数据分析与挖掘综合能力提升实战

【课程目标】

本课程为高级课程，培训的内容是继中级课程之后学习的，同时提供了更复杂的数据模型来解决实际工作中的商业决策问题。

本课程面向高级数据分析人员，以及系统开发人员。

本课程核心内容为数据挖掘，分类预测模型，以及专题模型分析，帮助学员构建系统全面的业务分析思维，提升学员的数据分析综合能力。

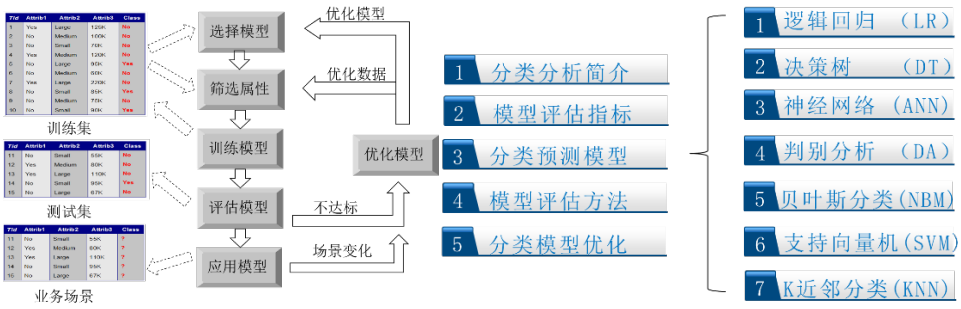
本课程覆盖了如下内容：

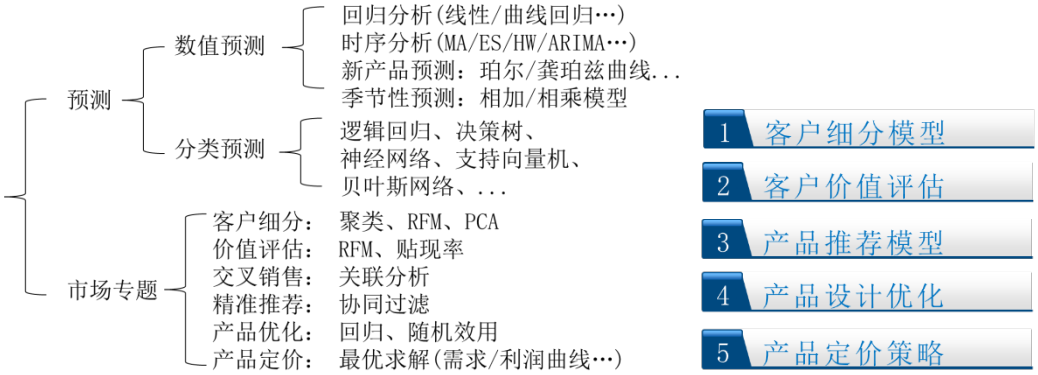
1、 数据建模过程

2、 分类预测模型

3、 分类模型优化思路

4、 市场专题分析模型





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，对数据分析及数据挖掘技术进行了全面的介绍（从数据收集与处理，到数据分析与挖掘，再到数据可视化和报告撰写），通过大量的操作演练，帮助学员掌握数据分析和数据挖掘的思路、方法、表达、工具，从大量的企业经营数据中进行分析，挖掘客户行为特点，帮助运营团队深入理解业务运作，以达到提升学员的数据综合分析能力，支撑运营决策的目的。

通过本课程的学习，达到如下目的：

1、 熟悉建模的一般过程，能够独立完成整个预测建模项目的实现。

2、 熟练使用各种分类预测模型，以及其应用场景。

3、 熟悉模型质量评估的关键指标，掌握模型优化的整体思路。

4、 熟练掌握常用市场专题分析模型：

a) 学会做市场客户细分，划分客户群

b) 学会实现客户价值评估

c) 学会产品功能设计与优化

d) 掌握产品精准推荐模型，学会推荐产品

e) 熟悉产品定价策略，寻找产品最优定价

【授课时间】

2-3 天时间

【授课对象】

业务支撑部、运营分析部、数据分析部、大数据系统开发部等对业务数据分析有较高要求的相关人员。

【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 Microsoft Office Excel 2013 版本及以上。

3、 便携机中事先安装好 IBM SPSS Statistics v24 版本及以上。

注：讲师可以提供试用版本软件及分析数据源。

【授课方式】

数据分析基础 + 方法讲解 + 实际业务问题分析 + 工具实践操作

采用互动式教学，围绕业务问题，展开数据分析过程，全过程演练操作，让学员在分析、分享、讲授、总结、自我实践过程中获得能力提升。

【课程大纲】

第一部分： 数据建模过程

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： 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

2、常见分类预测模型

3、逻辑回归模型

- 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
- 多元逻辑回归

案例：如何评估用户是否会购买某产品（二元逻辑回归）

案例：多品牌选择模型分析（多元逻辑回归）

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

- 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
  - 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：商场酸奶购买用户特征提取

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

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
- 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

7、最近邻分类（KNN）

- 基本原理
- 关键问题

8、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

9、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

第三部分： 分类模型优化

1、集成方法的基本原理：利用弱分类器构建强分类模型

- 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

2、集成方法/元算法的种类

- Bagging 算法
- Boosting 算法

3、Bagging 原理

- 如何选择数据集
- 如何进行投票
- 随机森林

4、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

第四部分： 市场细分模型

问题：我们的客户有几类？各类特征是什么？如何实现客户细分，开发符合细分市场的新产品？如何提取客户特征，从而对产品进行市场定位？

1、 市场细分的常用方法

- 有指导细分
- 无指导细分

2、 聚类分析

- 如何更好的了解客户群体和市场细分？
- 如何识别客户群体特征？
- 如何确定客户要分成多少适当的类别？
- 聚类方法原理介绍
- 聚类方法作用及其适用场景
- 聚类分析的种类
- K均值聚类（快速聚类）

  案例：移动三大品牌细分市场合适吗？

演练：宝洁公司如何选择新产品试销区域？

演练：如何评选优秀员工？

演练：中国各省份发达程度分析，让数据自动聚类

- 层次聚类（系统聚类）：发现多个类别
- R 型聚类与 Q 型聚类的区别

案例：中移动如何实现客户细分及营销策略

演练：中国省市经济发展情况分析（Q 型聚类）

演练：裁判评分的标准衡量，避免“黑哨”（R 型聚类）

- 两步聚类

3、主成分分析

- 主成分分析方法介绍
- 主成分分析基本思想
- 主成分分析步骤

案例：如何评估汽车购买者的客户细分市场

第五部分： 客户价值分析

营销问题：如何评估客户的价值？不同的价值客户有何区别对待？

1、如何评价客户生命周期的价值

- 贴现率与留存率
- 评估客户的真实价值
- 使用双向表衡量属性敏感度
- 变化的边际利润

案例：评估营销行为的合理性

2、RFM 模型（客户价值评估）

- RFM 模型，更深入了解你的客户价值
- RFM 模型与市场策略
- RFM 模型与活跃度分析

案例：淘宝客户价值评估与促销名单

案例：重购用户特征分析

第六部分： 产品推荐模型

问题：购买 A 产品的顾客还常常要购买其他什么产品？应该给客户推荐什么产品最有可能被接受？

1、常用产品推荐模型

2、 关联分析

- 如何制定套餐，实现交叉/捆绑销售

案例：啤酒与尿布、飓风与蛋挞

- 关联分析模型原理（Association）
- 关联规则的两个关键参数
  - 支持度
  - 置信度
- 关联分析的适用场景

案例：购物篮分析与产品捆绑销售/布局优化

案例：通信产品的交叉销售与产品推荐

3、 协同过滤

第七部分： 产品设计优化

1、联合分析法

2、 离散选择模型

- 如何评估客户购买产品的概率
- 如何指导产品开发？如何确定产品的重要特性

- 竞争下的产品动态调价
- 如何评估产品的价格弹性

案例：产品开发与设计分析

案例：品牌价值与价格敏感度分析

案例：纳什均衡价格

3、品牌价值评估

4、新产品市场占有率评估

第八部分： 产品定价策略及产品最优定价

营销问题：产品如何实现最估定价？套餐价格如何确定？采用哪些定价策略可达到利润最大化？

1、常见的定价方法

2、产品定价的理论依据

- 需求曲线与利润最大化
- 如何求解最优定价

案例：产品最优定价求解

3、如何评估需求曲线

- 价格弹性
- 曲线方程（线性、乘幂）

4、如何做产品组合定价

5、如何做产品捆绑/套餐定价

- 最大收益定价（演进规划求解）
- 避免价格反转的套餐定价

案例：电信公司的宽带、IPTV、移动电话套餐定价

6、非线性定价原理

- 要理解支付意愿曲线
- 支付意愿曲线与需求曲线的异同

案例：双重收费如何定价（如会费+按次计费）

7、阶梯定价策略

案例：电力公司如何做阶梯定价

8、数量折扣定价策略

案例：如何通过折扣来实现薄利多销

9、定价策略的评估与选择

案例：零售公司如何选择最优定价策略

10、 航空公司的收益管理

- 收益管理介绍
- 如何确定机票预订限制
- 如何确定机票超售数量
- 如何评估模型的收益

案例：FBN 航空公司如何实现收益管理（预订/超售）

第九部分： 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

第十部分： 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。