大数据分析与挖掘综合能力提升实战

【课程目标】

本课程为进阶课程，面向所有业务支撑部门及数据分析部门。

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。

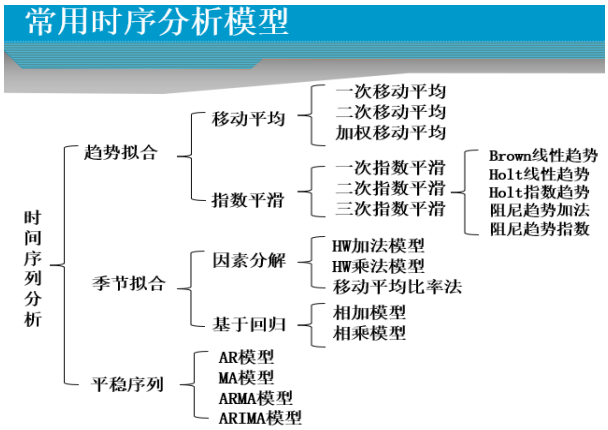
本课程具体内容包括：

1、 数据建模流程，特征工程处理

2、 时序预测模型，分类预测模型

3、 模型基本原理，模型含义解读

4、 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、 了解数据建模的标准过程

2、 明白时序预测的基本思想，熟悉常用的时序预测模型

3、 掌握常用的分类预测模型，理解模型基本原理

4、 学会解读分类预测模型的含义

5、 理解并掌握定性预测模型的质量评估指标

6、 了解分类预测模型的集成优化思想

【授课时间】

2 天时间（每天 6 个小时）

【授课对象】

产品销量部、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 v19 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—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

第二部分： 定量预测模型—时序预测篇

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

第三部分： 定性预测模型—分类预测篇

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

第四部分： 定性预测模型—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）

第五部分： 定性预测模型—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

结束：课程总结与问题答疑。