大数据挖掘工具: SPSS Statistics 入门与提高

【课程目标】

本课程为数据分析和挖掘的工具篇，本课程面向数据分析部等专门负责数据分析与挖掘的人士，专注大数据挖掘工具 SPSS Statistics 的培训。

IBM SPSS 工具是面向非专业人士的高级的分析工具（挖掘工具），它提供大量的分析方法和分析模型，能够解决更复杂的业务问题，比如影响因素分析、客户行为预测/精准营销、客户群划分、产品交叉销售、产品销量预测等等。工具它封装了复杂难懂的算法实现，即使你没有深厚的技能能力，也能够胜任复杂的数据分析和挖掘。

本课程从实际的业务需求出发，对数据分析及数据挖掘技术进行了全面的

介绍，将数据挖掘标准流程、分析思路、分析方法、分析模型，全部落地在SPSS 工具中，通过大量的工具操作和演练，帮助学员熟练掌握 SPSS 工具的使用，并能够将 SPSS 工具在实际的业务数据分析中满地，实现“知行合一”。

通过本课程的学习，达到如下目的：

1、 了解大数据挖掘的标准过程和挖掘步骤

2、 掌握常用的统计分析方法，以及可视化

3、 掌握常用的影响因素分析方法，学会根因分析

4、 理解数据挖掘的常见模型，原理及适用场景。

5、 熟练掌握 SPSS 基本操作，能利用 SPSS 解决实际的商业问题。

【授课时间】

2~4 天时间，或根据客户需求选择（每天 6 个小时）

知识点	2 天	4 天
数据挖掘标准流程	√	√
数据流预处理	√	√
数据可视化	√	√
影响因素分析	√	√
回归预测模型	√	√

时序预测模型	√（部分）	√
回归模型优化		√
分类预测模型		√
市场客户划分		√
客户价值评估		√
假设检验		√

【授课对象】

市场部、业务支撑部、数据分析部、运营分析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 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】

基础知识精讲 + 案例演练 + 实际业务问题分析 + 工具实际操作

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；引导学员思考，构建分析模型，进行数据分析与挖掘，以及数据呈现与解读，全过程演练操作，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。

【课程大纲】

第一部分： 数据挖掘标准流程

1、 数据挖掘概述

2、 数据挖掘的标准流程（CRISP-DM）

- 商业理解
- 数据准备
- 数据理解
- 模型建立
- 模型评估
- 模型应用

案例：客户流失预测及客户挽留

3、 数据集概述

4、 SPSS 工具介绍

5、 数据挖掘常用模型

第二部分： 数据预处理

如何整理数据，了解数据，对数据进行预处理？

1、 数据预处理的四大任务

- 数据集成：多个数据集合并
- 数据清洗：异常值的处理
- 样本处理：样本筛选、样本抽样、样本平衡
- 变量处理：变量变换、变量派生、变量精简

2、 数据集成（数据集合并）

- 样本追加（添加数据行）：横向合并
- 变量合并（添加变量列）：纵向合并

3、 数据清洗（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、 样本处理：行处理

- 样本筛选：指定条件筛选指定样本集（减少样本数量）
- 样本抽样：随机抽取部分样本集（减少样本数量）
- 样本平衡：正反样本比例均衡

5、 变量处理：列处理

- 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：变量删除/降维，减少变量个数
- 类型转换：数据类型的相互转换

6、 变量精简/变量降维常用方法

- 常用降维方法
- 如何确定降维后变量个数
- 特征选择：选择重要变量，剔除不重要变量
    - 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
- 利用 IV 值筛选
- 基于信息增益来选择属性

- 因子合并：将多个变量进行合并
  - PCA 主成分分析
  - 判别分析

7、 类型转换

8、 因子合并/主成分分析

- 因子分析的原因
- 因子个数选择原则
- 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

9、 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）

第三部分： 数据可视化

1、 数据可视化的原则

2、 常用可视化工具

3、 常用可视化图形

- 柱状图、条形图、饼图、折线图、箱图、散点图等

4、 图形的表达及适用场景

演练：各种图形绘制

第四部分： 影响因素分析篇

营销问题：哪些因素是影响业务目标的关键要素？比如，产品在货架上的位置是否对销量有影响？价格和广告开销是如何影响销量的？影响风控的关键因素有哪些？如何判断？

1、影响因素分析的常见方法

2、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

3、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？

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：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

5、相关性分析方法总结

第五部分： 回归预测模型

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

第六部分：　　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

第七部分：　　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

第八部分： 回归模型质量评估

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

第九部分： 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

第十部分： 分类预测模型篇

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

第十一部分： 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
  - 层次聚类
  - 两步聚类
- 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：如何自动评选优秀员工？

  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、客户细分与 PCA 分析法

- PCA 主成分分析的原理
- PCA 分析法的适用场景

演练：利用 PCA 对汽车客户群进行细分

演练：如何针对汽车客户群设计汽车

第十二部分： 客户价值评估

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

演练：“双 11”淘宝商家如何选择价值客户进行促销

演练：结合响应模型，宜家 IKE 实现最大化营销利润

案例：重购用户特征分析

第十三部分： 假设检验

1、参数检验分析（样本均值检验）

问题：如何验证营销效果的有效性？

- 假设检验概述
  - 单样本 T 检验
  - 两独立样本 T 检验
  - 两配对样本 T 检验

- 假设检验适用场景

电信行业

案例：电信运营商 ARPU 值评估分析（单样本）

案例：营销活动前后分析（两配对样本）

金融行业

案例：信用卡消费金额评估分析（单样本）

医疗行业

案例：吸烟与胆固醇升高的分析（两独立样本）

案例：减肥效果评估（两配对样本）

2、非参数检验分析（样本分布检验）

问题：这些属性数据的分布情况如何？如何从数据分布中看出问题？

- 非参数检验概述
  - 单样本检验
  - 两独立样本检验
  - 两相关样本检验

- 两配对样本检验

- 非参数检验适用场景

案例：产品合格率检验（单样本-二项分布）

案例：训练新方法有效性检验（两配对样本-符号/秩检验）

案例：促销方式效果检验(多相关样本-Friedman 检验)

案例：客户满意度差异检验(多相关样本-Cochran Q 检验)

结束：课程总结与问题答疑。