大数据挖掘工具: SPSS Statistics 入门与提高

【课程目标】

随着大数据分析的需求越来越旺盛，大数据分析工具也越来越琳琅满目，然而，绝大多数的分析工具都只具有单一用途，无法满足企业的复杂的多样化的全面的业务分析需求，因此分析工具的选择成为了一个挑战。

一个良好的分析工具必须满足如下要求：

1） 易学易用易操作。

2） 分析效率要高。

3） 满足业务分析需求。

如果要说前两个要求，显然类似于 Excel/Power BI/Tableau 等工具都是满足要求的，但此类工具却无法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，这些需求用 Excel/PBI 等工具就难以胜任了，需要用到更高级的数据挖掘工具，比如 IBM SPSS 工具。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，能够解决的业务问题更丰富，提供

了更加强大的业务数据分析功能，并且它封装了具体的分析算法，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。

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、 掌握基本的统计分析，常用的影响因素分析。

3、 理解数据挖掘的常见模型，原理及适用场景。

4、 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

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	√	√
数值预测模型	√回归时序	√季节模型
回归模型优化		√
分类预测模型	√仅决策树	√ANN/SVM/…
市场客户划分		√
客户价值评估		√
产品推荐模型		√
假设检验		√
实战		√

【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、 数据集的基本知识

a) 存储类型

b) 统计类型

c) 角度

4、SPSS 工具简介

第二部分： 数据预处理过程

1、 数据预处理的基本步骤

- 数据读取、数据理解、数据处理、变量处理、探索分析

2、 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

3、 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）

4、 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

5、 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

6、 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

7、 数据降维

- 常用降维方法
- 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

8、 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： 数据可视化篇

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

问题：如何判断一个因素对另一个因素有影响？比如营销费用是否会影响销售额？产品价格是否会影响销量？产品的陈列位置是否会影响销量？风险控制的关键因素有哪些？如何判断？

1、影响因素分析的常见方法

2、相关分析（衡量变量间的的相关性）

问题：这两个属性是否会相互影响？影响程度大吗？营销费用会影响销售额吗？

- 什么是相关关系
- 相关系数：衡量相关程度的指标
- 相关系数的三个计算公式
- 相关分析的假设检验
- 相关分析的基本步骤
- 相关分析应用场景

演练：体重与腰围的关系

演练：营销费用会影响销售额吗

演练：哪些因素与汽车销量有相关性

演练：通信费用与开通月数的相关分析

案例：酒楼生意好坏与报纸销量的相关分析

- 偏相关分析
- 距离相关分析

3、方差分析

问题：哪些才是影响销量的关键因素？

- 方差分析解决什么问题
- 方差分析种类：单因素/双因素可重复/双因素无重复
- 方差分析的应用场景
- 方差分析的原理与步骤
- 如何解决方差分析结果

演练：终端摆放位置与终端销量有关吗？

演练：开通月数驿客户流失的影响分析

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、技能级别产品销量有影响吗？

案例：2015 年大学生工资与父母职业的关系

案例：医生洗手与婴儿存活率的关系

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

演练：饲料对生猪体重的影响分析（协方差分析）

4、列联分析（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

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

第五部分： 数据建模过程篇

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
- 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

第六部分： 数值预测模型篇

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、销量预测与市场预测——让你看得更远

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）
- 得到回归方程的几种常用方法
- 回归分析的五个步骤与结果解读
- 回归预测结果评估（如何评估预测质量，如何选择最佳回归模型）

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、时序预测

问题：随着时间变化，未来的销量变化趋势如何？

- 时序分析的应用场景（基于时间的变化规律）
- 移动平均 MA 的预测原理
- 指数平滑 ES 的预测原理
- 自回归移动平均 ARIMA 模型
- 如何评估预测值的准确性？

案例：销售额的时序预测及评估

演练：汽车销量预测及评估

演练：电视机销量预测分析

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

4、季节性预测模型

- 季节性回归模型的参数
- 常用季节性预测模型（相加、相乘）

案例：美国航空旅客里程的季节性趋势分析

案例：产品销售季节性趋势预测分析

5、新产品预测模型与 S 曲线

- 如何评估销量增长的拐点
- 珀尔曲线与龚铂兹曲线

案例：如何预测产品的销售增长拐点，以及销量上限

演戏：预测 IPad 产品的销量

6、自定义模型（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

第七部分： 回归模型优化篇

1、回归模型的基本原理

- 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？
- 拟合优度检验：回归模型的质量评估？

- 因素的显著性检验：自变量是否可用？
- 理解标准误差的含义：预测的准确性？

2、模型优化思路：寻找最佳回归拟合线

- 如何处理异常数据（残差与异常值排除）
- 如何剔除非显著因素（因素显著性检验）
- 如何进行非线性关系检验
- 如何进行相互作用检验
- 如何进行多重共线性检验
- 如何检验误差项
- 如何判断模型过拟合

案例：模型优化案例

第八部分： 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

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、逻辑回归模型（LR）

- 逻辑回归模型原理及适用场景
- 逻辑回归种类：二项/多项逻辑回归
- 如何解读逻辑回归方程

案例：如何评估用户是否会购买某产品（二项逻辑回归）

- 消费者品牌选择模型分析

案例：多品牌选择模型分析（多项逻辑回归）

5、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
- 如何评估分类性能？

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：电信运营商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

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、K 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

第九部分： 市场细分模型篇

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 PCA 分析

- 主成分分析原理
- 主成分分析基本步骤
- 主成分分析结果解读

演练：PCA 探索汽车购买者的细分市场

4、RFM 模型客户细分框架

第十部分： 客户价值评估

1、 客户价值评估与 RFM 模型

问题：如何评估客户的价值？如何针对不同客户采取不同的营销策略？

- RFM 模型，更深入了解你的客户价值
- RFM 的客户细分框架理解
- RFM 模型与市场策略
- RFM 模型与活跃度

演练：“双 11”淘宝商家如何选择客户进行促销

演练：结合响应模型，宜家 IKE 实现最大化营销利润

演练：重购用户特征分析

第十一部分： 假设检验篇

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)

第十二部分： 产品定价策略及最优定价

（根据需要讲解）

第十三部分： 实战-数据挖掘项目

实战 1：客户流失预警与客户挽留之真实数据分析实践

实战 2：银行信用风险分析

结束：课程总结与问题答疑。