大数据分析与挖掘综合能力提升实战

【课程目标】

本课程为大数据分析中级课程，需要在初级课程之后学习。面向所有应用型人员，包括业务部分，以及数据分析部门，系统开发人员也同样需要学习。

本课程核心内容为数据挖掘，预测模型，以及模型优化，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。

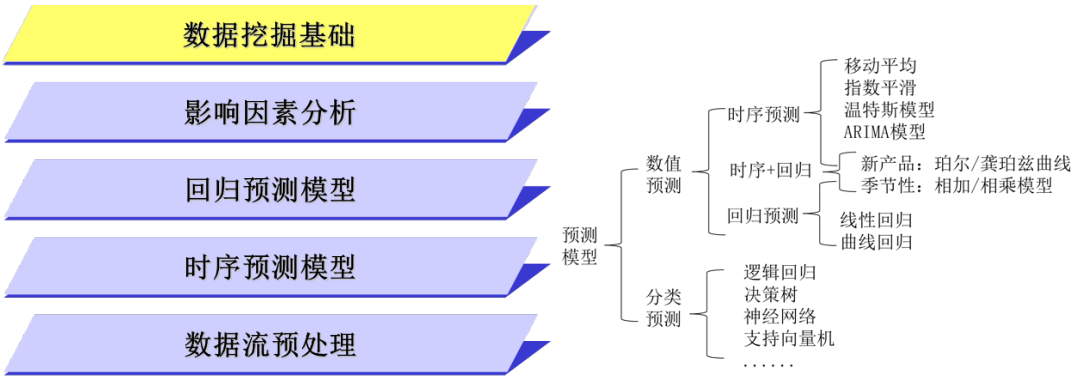
本课程覆盖了如下内容：

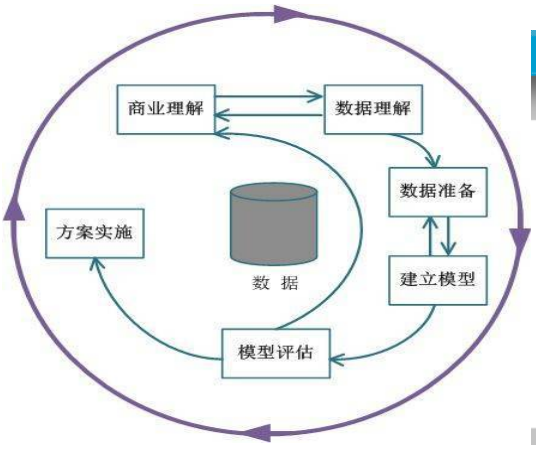
1、 数据挖掘基础知识。

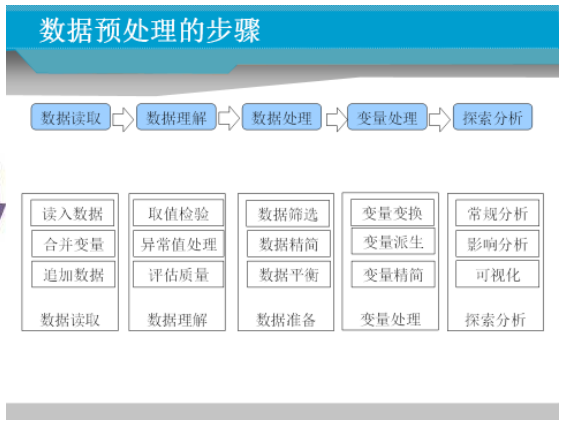
2、 常用数值预测模型。

3、 常用时序预测模型。

4、 数据预处理的基本过程。







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、 了解数据挖掘基础知识，以及数据挖掘标准过程。

2、 掌握建模前的影响因素分析，学会寻找影响业务的关键因素。

3、 熟练使用数值预测模型，掌握回归预测模型，学会解读模型中业务规律。

4、 学会自定义回归模型，能够对回归模型进行优化，并找到最优的回归模型。

5、 熟练掌握预处理的基本过程，并根据业务实际情况进行处理。

【授课时间】

2-3 天时间（每天 6 个小时）

【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 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： 数据挖掘基础

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、 变量的类型

- 存储类型
- 度量类型
- 角色

5、 SPSS 工具介绍

6、 数据挖掘常用模型

第二部分： 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：价格是否会影响产品销量？产品的陈列位置是否会影响销量？学历是否与客户流失有关系？影响风险的关键因素有哪些？

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）

问题：这两个属性是否会相互影响？影响程度大吗？

- 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：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：话费与网龄的相关分析

- 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）

问题：哪些才是影响销量的关键因素？

- 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
- 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：终端摆放位置与终端销量有关吗

演练：开通月数对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（多因素）

- 协方差分析原理
- 协方差分析的适用场景

演练：饲料对生猪体重的影响分析（协方差分析）

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
- 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

第三部分： 回归预测模型篇

问题：如何预测产品的销量/销售金额？

1、常用预测模型

- 数值预测：回归预测/时序预测
- 分类预测：逻辑回归、决策树、神经网络、…

2、回归分析/回归预测

问题：如何预测未来的销售量（定量分析）？

- 回归分析简介
- 回归分析的种类（一元/多元、线性/曲线）
- 得到回归方程的常用工具
  - 散点图+趋势线
  - 线性回归工具

- 规划求解工具

演练：散点图找营销费用与销售额的关系（一元回归）

- 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）

- 解读线性回归分析结果的技巧
  - 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
- 回归预测模型质量评估
  - 评估指标：判定系数 R^2、标准误差
  - 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）

- 预测值准确性评估
  - MAD、MSE/RMSE、MAPE 等
- 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、自动筛选不显著因素（自变量）

第四部分： 回归模型优化篇

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：是否可以做回归分析？
- 因素的显著性检验：自变量是否可用？
- 拟合优度检验：回归模型的质量评估？
- 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）
- 如何检验误差项（修改因变量）
- 如何判断模型过拟合（模型过拟合判断）

演练：模型优化案例

3、规划求解工具简介（自定义回归模型的工具）

4、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

5、好模型都是优化出来的

第五部分： 时序预测模型篇

问题：类似于 GDP 这种无法找到或找全影响因素，无法进行回归建模，怎么办？

1、时间序列简介

2、时间序列常用模型

3、评估预测值的准确度指标

- 平均绝对误差 MAE
- 均方差 MSE/RMSE
- 平均误差率 MAPE

4、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
  - 期数 N 的最佳选择方法
  - 最优权重系数的选取方法

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

5、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

6、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

7、回归季节预测模型

- 回归季节模型的参数
- 基于时期 t 的相加模型
- 基于时期 t 的相乘模型
- 怎样解读模型的含义

案例：美国航空旅客里程的季节性趋势分析

8、ARIMA 模型

- 适用场景及原理
- ARIMA 操作

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

9、 新产品销量预测模型

- 新产品累计销量的 S 曲线
- 如何评估销量增长的拐点及销量上限
- 珀尔曲线与龚铂兹曲线

演练：预测 IPad 产品的销量

演练：预测 Facebook 的用户增长情况

第六部分： 数据预处理篇（了解你的数据集）

1、 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

2、 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）

3、 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

4、 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

5、 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

6、 数据降维

- 常用降维的方法
- 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

7、 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）

8、 数据可视化

- 数据可视化：柱状图、条形图、饼图、折线图、箱图、散点图等
- 图形的表达及适用场景

演练：各种图形绘制

结束：课程总结与问题答疑。