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、 线性回归模型，模型基本原理

3、 模型质量评估，模型优化措施

4、 回归方程解读，自定义回归模型

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、 掌握数据建模的标准过程和步骤

2、 掌握建模前的特征选择常用方法，学会寻找影响业务的关键要素

3、 掌握回归预测模型基本原理，学会解读回归方程的含义

4、 理解并掌握定量预测模型的评估指标的含义

5、 学会利用规划求解实现自定义回归模型（非线性回归模型）

6、 掌握常用的回归模型优化措施

7、 熟练掌握数据预处理的基本任务，并根据业务实际情况进行处理

【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】

理论精讲 + 模型原理 + 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—建模步骤篇

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： 影响因素分析—特征工程篇

问题：如何选择合适的属性/特征来建模呢？选择的依据是什么？比如价格是否

可用于产品销量预测？

1、数据预处理 vs 特征工程

2、特征选择常用方法

- 相关分析、方差分析、卡方检验

3、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

4、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？

5、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

第三部分： 定量预测模型—回归模型篇

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

8、自动筛选不显著因素（自变量）

第四部分： 定量预测模型—回归优化篇

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：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

第五部分： 定量预测模型—自定义回归篇

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

第六部分： 定量预测模型—模型评估篇

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

结束：课程总结与问题答疑。