Python 数据挖掘专题实战培训

【课程目标】

本课程主要讲解如何利用 Python 进行时间序列的数据建模。

通过本课程的学习，达到如下目的：

1、全面掌握 Python 语言以及其编程思想。

2、掌握常用扩展库的使用，特别是数据挖掘相关库的使用。

3、学会使用 Python 完成数据挖掘项目整个过程。

4、掌握利用 Python 实现可视化呈现。

5、掌握数据挖掘常见算法在 Python 中的实现。

【授课时间】

2-5 天时间

（要根据学员的实际情况调整重点内容及时间）

【授课对象】

业务支持部、IT 系统部、大数据系统开发部、大数据分析中心、网络运维部等相关技术人员。

【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 Python 3.9 版本及以上。

3、 安装好 Numpy,Pandas,statsmodels,sklearn,scipy 等常用库。

注：讲师现场提供分析的数据源。

【授课方式】

语言基础 + 挖掘模型 + 案例演练 + 开发实践 + 可视化呈现

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、 数据挖掘常用模型

第二部分： 数据预处理篇

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

- 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

第三部分： 用户专题分析

1、 用户专题分析的主要任务

2、 客户群细分与聚类分析

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？

- 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

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

- 最优 K 值选择
  - Elbow 手肘法
  - Silhouette Coefficient 轮廓系数
  - Calinski-Harabasz Index 准则
- 双聚类 bicluster 及评估
- 谱聚类联合
  - 联合谱聚类 SpectralCoclustering
  - 双向谱聚类 SpectralBiclustering
- DBSCAN 邻近聚类

3、客户喜好评估与主成分分析 PCA

营销问题：如何汇聚大众的共同喜好？

- 主成分分析方法介绍
- 主成分分析基本思想
- 主成分分析步骤

案例：如何评估汽车购买者的客户细分市场

4、客户价值评估与 RFM 模型

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？

- 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

案例：淘宝客户价值评估与促销名单

案例：重购用户特征分析

第四部分： 产品专题分析

1、产品专题分析主要任务

- 产品设计分析
- 市场占有分析
- 累计销量分析
- 定价策略分析

2、产品设计优化（联合分析法）

问题：如何设计最优的功能特征？

- 评估功能特征的重要性

- 评估功能特征的价值

案例：产品开发与设计分析

3、产品评估模型（随机效用理论）

- 属性重要性评估
- 市场占有率评估
- 产品价格弹性评估
- 评估产品的品牌价值
- 动态调价（纳会均衡价格）

案例：品牌价值与价格敏感度分析

案例：纳什均衡价格

第五部分： 产品定价策略

营销问题：产品如何实现最优定价？套餐价格如何确定？采用哪些定价策略可达到利润最大化？

1、常见的定价方法

2、产品定价的理论依据

- 需求曲线与利润最大化

- 如何求解最优定价

案例：产品最优定价求解

3、如何评估需求曲线

- 价格弹性
- 曲线方程（线性、乘幂）

4、如何做产品组合定价

5、如何做产品捆绑/套餐定价

- 最大收益定价（演进规划求解）
- 避免价格反转的套餐定价

案例：电信公司的宽带、IPTV、移动电话套餐定价

6、非线性定价原理

- 要理解支付意愿曲线
- 支付意愿曲线与需求曲线的异同

案例：双重收费如何定价（如会费+按次计费）

7、阶梯定价策略

案例：电力公司如何做阶梯定价

8、数量折扣定价策略

案例：如何通过折扣来实现薄利多销

9、定价策略的评估与选择

案例：零售公司如何选择最优定价策略

10、 航空公司的收益管理

- 收益管理介绍
- 如何确定机票预订限制
- 如何确定机票超售数量
- 如何评估模型的收益

案例：FBN 航空公司如何实现收益管理（预订/超售）

第六部分： 产品推荐与协同过滤

问题：购买 A 产品的顾客还常常要购买其他什么产品？应该给客户推荐什么产品最有可能被接受？

1、从搜索引擎到推荐引擎

2、常用产品推荐模型及算法

3、基于流行度的推荐

- 基于排行榜的推荐，适用于刚注册的用户

- 优化思路：分群推荐

4、基于内容的推荐 CBR

- 关键问题：如何计算物品的相似度
- 优缺点
- 优化：Rocchio 算法、基于标签的推荐、基于兴趣度的推荐

5、基于用户的推荐

- 关键问题：如何对用户分类/计算用户的相似度
- 算法：按属性分类、按偏好分类、按地理位置

6、协同过滤的推荐

- 基于用户的协同过滤
- 基于物品的协同过滤
- 冷启动的问题

案例：计算用户相似度、计算物品相似度

7、基于分类模型的推荐

8、其它推荐算法

- LFM 基于隐语义模型
- 按社交关系

- 基于时间上下文

9、多推荐引擎的协同工作

第七部分： 信用评分卡模型

1、信用评分卡模型简介

2、评分卡的关键问题

3、信用评分卡建立过程

- 筛选重要属性
- 数据集转化
- 建立分类模型
- 计算属性分值
- 确定审批阈值

4、筛选重要属性

- 属性分段
- 基本概念：WOE、IV
- 属性重要性评估

5、数据集转化

- 连续属性最优分段
- 计算属性取值的 WOE

6、建立分类模型

- 训练逻辑回归模型
- 评估模型
- 得到字段系数

7、计算属性分值

- 计算补偿与刻度值
- 计算各字段得分
- 生成评分卡

8、确定审批阈值

- 画 K-S 曲线
- 计算 K-S 值
- 获取最优阈值

第八部分： 交叉销售与关联规则

1、 关联规则概述

2、 常用关联规则算法

- Apriori 算法
  - 发现频繁集
  - 生成关联规则
- FP-Growth 算法
  - 构建 FP 树
  - 提取规则

案例：使用 apriori 实现关联分析

10、 基于关联分析的推荐

- 如何制定套餐，实现交叉/捆绑销售

案例：啤酒与尿布、飓风与蛋挞

- 关联分析模型原理（Association）
- 关联规则的两个关键参数
  - 支持度
  - 置信度

- 关联分析的适用场景

案例：购物篮分析与产品捆绑销售/布局优化

案例：通信产品的交叉销售与产品推荐

结束：课程总结与问题答疑。