



CDA 数据分析师
CERTIFIED DATA ANALYST

CDA 企业内训课程库

备注：实际课程可选择已有模块+定制模块

一、业务分析师（一级）内训课纲（9 天）	2
二、Python 数据挖掘算法课纲（二级建模）（6 天）	7
三、机器学习课纲（18 天）	9
四、SPSS MODELER 数据挖掘课纲（4 天）	11
五、SAS 数据挖掘（6 天）	14
六、SAS 数据可视化（4 天）	21
七、SAS 评分模型数据分析（4 天）	22
八、大数据分析应用课纲（5 天）	23
九、Power BI 培训（2 天）	25
十、Python 基础及可视化（2 天）	26
十一、大数据认知与应用（4 天）	32
十二、大数据治理实战（4.5 天）	37
十三、如何撰写有价值的数据分析报告（1-2 天）	40
十四、SPSS 数据分析课程（5 天）	43
十五、R 语言数据挖掘培训（3 天）	48
十六、数据化运营课程培训（2 天）	50
十七、互联网数据分析思维培训（2 天）	53
十八、金融数字化转型人才训练营（10 天）	56
十九、金融投行业大数据应用	64

一、业务分析师（一级）内训课纲（9 天）

课程模块	内容设置	课时
概率论与 数理统计基础	part1: 数据分析方法概述 1、数据分析过程 2、数据分析的商业驱动 part2: 概率论基础 1、随机事件及其概率 2、概率的性质与运算法则 3、离散型随机变量及其分布 4、连续型随机变量的概率分布 part3: 描述性统计 1、集中趋势的度量 2、离散程度的度量 3、偏态与峰态的度量 part4: 常见分布族 1、分布概念 2、正态分布及其衍生重要分布 3、中心极限定理 4、样本比例抽样分布 part5: 多维随机变量 1、联合分布 2、协方差 3、相关系数 part6: 数据简化原理 1、似然函数 2、辅助函数 part7: 参数估计 1、点估计 2、区间估计	2 天

	3、样本量的确定 part8: 假设检验 1、单侧检验 2、总体参数检验 3、P 值 part9: 方差分析 1、单因素方差分析 2、双因素方差分析 part10: 一元线性回归 1、变量间关系的度量 2、一元线性回归方程 2、线性相关检验 3、残差分析	
Mysql 数据库管理	part0: MySQL 的安装与卸载 1、MySQL 安装 2、MySQL 的报错解决 3、MySQL 环境变量配置 part1: 数据库基本操作 1、MySQL 数据库的创建: 创建、查看、使用、删除数据库 part2: 数据表操作 1、数据表与数据库之间的关系 2、数据表的创建、查看、删除语句 part3: 数据库、数据表与字段操作 1、数据类型: 整数型、小数型、日期和时间、字符串 2、约束条件: 主键约束、非空约束、唯一性约束、默认约束、自增字段 part4: SQL 查询与函数 1、查询语句: 单表查询、多表查询、内连接查询、左连接、右连接、聚合查询等	2 天

	2、聚合函数：AVG、SUM、MAX、MIN、COUNT 3、查询操作符与子查询：and、or、in、like、is null、distinct 4、操作符与子查询的组合应用 5、as 重命名与 limit 限制查询结果行数 6、函数：常用数学函数、字符串函数、日期及时间函数、其他函数 7、为字段赋值 8、删除记录 part5: SQL 查询应用 1、电商数据表 ER 图 2、电商数据多表查询练习	
SPSS 数据分析实战 (SPSS 为实现工具， 主要讲解案例)	part0: SPSS 软件介绍 1、数据分析流程介绍：以员工短期绩效统计为例 2、SPSS 特点 3、SPSS 界面及操作介绍 part1: 数据的输入与保存 1、数据的获取与数据格式 2、数据变量详解 3、SPSS 访问外部数据 part2: 数据的预处理 1、数据清洗 2、变量衍生 3、描述性统计分析功能 4、使用 SPSS 绘制常用统计图形 part3: 一元线性回归 1、最小二乘估计 2、线性回归与相关 3、线性回归与方差 4、数据分析流程 part4: 多元线性回归	4 天

	1、正态多元线性回归的假设 2、正态分布的问题 3、异方差问题与处理 4、异常值问题与处理 5、共线性问题与处理 part5: Logistic 回归 1、logistic 回归与卡方：卡方分析原理、相关系数指标（kappa、风险、OR 等） 2、最大似估计 3、logistic 回归解析：模型与线性回归对比、两种 R2 对比、回归系数与 OR 值、RR 值、OR 值域 gamma 值 4、评分与预测：老样本预测与市场细分、数据的分箱化、数据离散化、评分卡模型中观点的意义 part6: 特征筛选、主成分分析 1、特征筛选：特征筛选流程、DB 特征筛选合理性、DB 特征筛选方法 2、主成分回归：主成分原理、主成分判断标准、主成分的应用场景、主成分的因子、主成分回归 part7: 市场细分-客户画像 1、用户画像应用：商业场景、模型算法应用 2、模型变量的确定：标签数、标签筛选 3、聚类算法：K-Means 聚类、两步聚类、Kohonen 聚类 4、市场细分：分组 5、标签制作原则 part8: 市场细分-客户价值模型 1、FRM 模型：商业应用、模型算法应用 2、因变量的确定：理论性及专家意见、潜变量及显变量评价 3、价值细分：分箱化、秩变换 4、价值细分应用：价值细分、价值评分 part9: 时间序列	
--	---	--

	1、时间序列原理介绍：AR 模型、MA 模型、ARIMA 模型 2、时间序列数据预处理：定义时间变量、时序数据的处理、序列图形 3、使劲序列的建模与预测：确定性分析、随机性分析 part10：不同市场订户信息的序列预测	
考点讲解及测试	Level 1 考点讲解 模拟测试	1 天

二、Python 数据挖掘算法课纲（二级建模）（6 天）

01 章 Python 数据挖掘基础理论与数据预处理技术

01-01 数据挖掘概要

01-02 数据挖掘的方法和原理

01-03 数据挖掘基础和进阶技术概述

01-04 数据预处理技术：字段选择-数据清洗-字段扩充-数据编码

01-05 人工特征工程：特征构造 - 特征抽取 - 特征选择

02 章 Python 数据挖掘之决策树算法

02-01 决策树建模思路

02-02 Quinlan 系列决策树 (ID3、C4.5、C8.0) 建模原理

02-03 CART 建模原理

02-04 决策树模型修剪

02-05 决策树模型效果评估

02-06 案例：使用决策树进行初始信用评级

03 章 Python 数据挖掘之神经网络和贝叶斯网络

03-01 了解神经网络的基本概念

03-02 明确人工神经网络结构

03-03 神经元模型

03-04 掌握 BP 神经网络学习算法

03-05 多层感知器的 scikit-learn 代码实现

03-06 贝叶斯公式与分类原理

03-07 朴素贝叶斯的参数估计

03-08 在 Python 中实现朴素贝叶斯

04 章 Python 数据挖掘之分类器与集成学习方法

04-09 KNN 算法原理与 Python 实现

04-01 线性可分与线性不可分

04-02 线性可分的支持向量机

04-03 线性支持向量机与软间隔最大化

04-04 非线性支持向量机与核函数

04-05 集成学习方法: Bagging、Boosting、随机森林

05 章 Python 数据挖掘之聚类分析

05-01 聚类算法的概述

05-02 聚类算法基本概念

05-03 聚类模型的评估

05-04 层次聚类原理与 Python 实现

05-05 基于划分的聚类 K-means 的原理及应用

05-06 详谈基于密度的聚类方法与 Python 实现

05-07 案例: 通信客户业务使用偏好聚类

06 章 Python 数据挖掘之关联规则与序列模式

06-01 关联规则的一些基本概念

06-02 关联规则 Ariori 算法的原理与 Python 实现

06-03 关联规则 FP-growth 算法

06-04 序列模式的简介与概念

06-05 序列模式 AprioriAll 算法与 Python 实现

06-06 基于用户和商品的协同过滤算法仅供参考

三、机器学习课纲（18 天）

大类	小类	培训项目	说明	时间 (天)
建模基础	数学基础	高等数学	高等数据基础与微积分	2
		概率论	概率基础、概率分布、最大似然估计等	
		统计学	描述统计、推断统计	
		线性代数	向量、行列式、矩阵、特征值与特征变量、奇异值分解等	
		最优化	梯度下降法、牛顿法等	
	PYTHON 基础	基础知识	安装配置、基础语法等	4
		常用库	Numpy、Pandas、Matplotlib、Ipython、	
		机器学习库	Scipy、Scikit-learn、TensorFlow	
		数据处理	数据清洗、数据标准化方法、缺失值处理等	
		模型评价	模型评价的方法，如准确率、召回率、KS、PSI 等	
机器学习算法	分类算法	逻辑回归算法	算法原理、Python 实现、示例讲解	6
		决策树算法	算法原理、Python 实现、示例讲解	
		K 近邻算法	算法原理、Python 实现、示例讲解	
		朴素贝叶斯算法	算法原理、Python 实现、示例讲解	
	回归算法	线性回归算法	算法原理、Python 实现、示例讲解	
		岭回归算法	算法原理、Python 实现、示例讲解	
		Lasso 回归算法	算法原理、Python 实现、示例讲解	
	聚类算法	K-Means 算法（基于划分）	算法原理、Python 实现、示例讲解	
		DBSCAN 算法（基于密度）	算法原理、Python 实现、示例讲解	
		层次聚类法	算法原理、Python 实现、示例讲解	
	降维算法	主成分分析算法（PCA）	算法原理、Python 实现、示例讲解	
		线性判别分析算法（LDA）	算法原理、Python 实现、示例讲解	
		LLE 局部线性嵌入算法（非线性降维）	算法原理、Python 实现、示例讲解	

	集成学习算法	Bagging/Boosting	算法原理、Python 实现、示例讲解	2
		GBDT 算法	算法原理、Python 实现、示例讲解	
		随机森林算法	算法原理、Python 实现、示例讲解	
		Adaboost 算法	算法原理、Python 实现、示例讲解	
		Xgboost 算法	算法原理、Python 实现、示例讲解	
	深度学习算法	文本挖掘算法	算法原理、Python 实现、示例讲解	2
		关系图谱算法	算法原理、Python 实现、示例讲解	
	其他算法	神经网络算法	算法原理、Python 实现、示例讲解	
		关联规则算法	算法原理、Python 实现、示例讲解	
		SVM 支持向量机算法	算法原理、Python 实现、示例讲解	
实践项目	案例	客户生命周期案例	结合客户生命周期案例，讲解精准营销、银行反欺诈、信用评分等内容。	2

四、SPSS MODELER 数据挖掘课纲（4 天）

时 程	项 目	时数
第 1 天	第一部分：数据挖掘实务应用	1 小时
	涵盖内容： 1. 数据挖掘在银行业的应用 2. 数据挖掘在零售业的应用 3. 数据挖掘在保险业的应用 4. 数据挖掘在政府-财政部国税局、工商局的应用	
第 1 天	第二部分：数据挖掘在银行业中的应用	1 小时
	涵盖内容： 1. 数据挖掘在银行顾客关系管理中的应用 2. 数据挖掘在银行信用风险评等中的应用 IBM SPSS MODELER 实务操作： IBM SPSS MODELER 分析环境介绍 利用 IBM SPSS MODELER 建立项目 利用 IBM SPSS MODELER 实作一个简单案例	
第 1 天	第三部分：数据挖掘基础	4 小时
	涵盖内容： 1. 数据挖掘的起源、定义及目标 2. 数据挖掘的发展历程 3. 数据库中的知识发掘步骤 4. 数据挖掘的产业标准 5. 数据挖掘技术的功能分类 6. 数据挖掘相关网站介绍 7. 基本数据挖掘技术简介(查询工具、统计技术、可视化技术、K-最近邻技术、...) 8. 进阶数据挖掘技术简介(分类、预测、关联规则、序列型样、聚类、...) 9. 数据挖掘的绩效评估及顾客数优化	

	10. 问题导向式数据挖掘分析流程	
	IBM SPSS MODELER 实务操作： 利用 IBM SPSS MODELER 进行初步数据探索 利用 IBM SPSS MODELER 中的 K-最近邻技术进行案例分析 实作如何将数据挖掘的分析结果转为利润图(Profit Chart)以建立获利最高的数据挖掘模型	
第 2 天	第四部分：银行数据挖掘案例：银行信用风险评估之进件评分卡案例	3 小时
	IBM SPSS MODELER 实务操作： 利用 IBM SPSS MODELER 对信用评分卡数据进行数据前处理及发掘关键变量	
第 2 天	第五部分：银行数据挖掘案例：对的产品如何销售给对的人	3 小时
	涵盖内容： 1. 商业问题(Business Issue) 2. 数据集描述(Dataset Description) 3. 数据探索(Data Explore) 4. 数据清理(Data Cleansing) 5. 字段扩充 6. 数据编码 7. 数据挖掘 8. 挖掘结果解读 9. 营销活动设计 10. 预期收益	
	IBM SPSS MODELER 实务操作： 利用 IBM SPSS MODELER 对银行营销数据进行分析、建模并发展营销策略	
第 3 天	第六部分：银行数据挖掘案例：信用行为的因素筛选	6 小时
	涵盖内容： 1. 信用行为的标准是什么？ 2. 信用指标的信息是否充分 3. 区分好客户和坏客户的特征怎么选 4. 初步搭建预测模型与项目评估	

	5. 对老客户还是新客户感兴趣	
	6. 技术与业务评估	
	IBM SPSS MODELER 实务操作	
第 4 天	第七部分：银行数据挖掘案例：基于订购行为的营销优化	6 小时
	涵盖内容：	
	1. 数据分布信息	
	2. 数据是否真的准确	
	3. 经验与数据怎样结合起来	
	4. sas 代码的优化	
	5. 如何定义投资回报率	
	6. 客户行为特征与营销优化	
	7. 制定市场细分标签	

五、SAS 数据挖掘（6 天）

时 程	大纲内容
第 1 天	主题：数据挖掘基础
	企业使用之范围： 说明企业如何利用数据挖掘来进行营销活动及信用风险控管
	理论介绍： 数据挖掘在政府部门及各行业的应用 数据挖掘的起源、定义及目标 数据挖掘的发展历程 数据库中的知识发掘步骤 (KDD) 数据挖掘技术的产业标准 (CRISP DM) 数据挖掘技术的功能分类 数据挖掘相关网站介绍 (KDnuggets & Kaggle)
	实作，使用数据集：信用评等数据集、天气数据集、玻璃制品数据集
	主题：基础数据挖掘及数据前处理技术
	企业使用之范围： 说明企业如何利用数据前处理技术来进行数据的重整，建置数据仓库，以作为数据挖掘之输入
	理论介绍： 叙述性统计及可视化技术 案例为本的学习 (Case-based Learning)：KNN (K Nearest Neighbors)
第 1 天 第 2 天	数据前处理 (Data Preprocessing) 技术 字段选择 (Attribute Selection) *数据整合 (Data Integration) *数据探索 (Data Explore) *数据过滤 (Data Filtering) 数据清洗 (Data Cleansing)： *错误值 (Wrong Value) 的侦测及处理

	*离群值(Outlier)的侦测及处理 *缺失值(Missing Value)的侦测及处理 - 字段扩充(Attribute Enrichment): *内/外部数据的扩充方法 - 数据编码(Data Coding): *数据正规化(Data Normalization) *数据一般化(Data Generalization) *数据连续性指派(Data Arbitrarily Assignment) *数据离散化(Data Discretization) *数据精简(Record Reduction) *域值精简(Attribute Value Reduction) *字段精简(Attribute Reduction)
	实作, 使用数据集: 新车设计数据集、信用评等数据集、药物治疗数据集、电信客户流失数据集、鸢尾花数据集、糖尿病数据集、电离层雷达侦测数据集
第 2 天	主题: 进阶数据挖掘技术 1
	企业使用之范围: 说明企业如何利用关键变量发掘技术来发掘对项目目标有效之关键变量, 以做为数据挖掘之输入变量
	理论介绍: 训练数据与测试数据的产生方法(随机取样切割法&分层抽样切割法) 无效变量(不相关变量&多余变量) 统计方式的变量选择: *卡方检定(Chi-square Test) *t 检定及 ANOVA 检定(t Test & ANOVA Test) 模型方式的变量选择: *决策树(Decision Tree)选择法 *罗吉斯回归(Logistic Regression)选择法 *包装法(Wrapper Methods)
	实作, 使用数据集: 信用评等数据集、玻璃制品数据集

第 3 天	主题：进阶数据挖掘技术 2
	企业使用之范围： 说明企业如何利用分类技术之贝式网络及决策树来建立交叉销售 (Cross-Selling) 模型，以提升公司获利
	理论介绍： 分类之简单贝式网络 (Naive Bayes) 及贝式网络 (Bayes Net) 分类之决策树 (Decision Tree)：ID3, C5.0, CART & CHAID 分类模型的评估 *混乱矩阵 (正确率、响应率、捕捉率、F-指标) *Gain Chart, Lift Chart, Profit Chart
	实作，使用数据集：银行客户营销数据集、电信产品跨销售 (Cross-Selling) 数据集、疾病诊断数据集、寿险推销数据集、糖尿病预测数据集、乳癌数据集、天气数据集、影像数据集
第 3 天	主题：进阶数据挖掘技术 3
	企业使用之范围： 说明企业如何利用分类技术之神经网络、罗吉斯回归来建立信用评分 (Credit Scoring) 模型，以降低公司损失
	理论介绍： 分类之神经网络 (Neural Network) 分类之罗吉斯回归 (Logistic Regression) 分类模型的评估 *ROC Curve, AUC *GINI Index, KS
	实作，使用数据集：细胞样本数据集、零售促销数据集、电信客户分类 (不同套餐选择) 数据集、电信客户流失数据集、便利超商选点数据集、房价数据集、AISPACE 神经网络系统介绍、可视化工具 (Boundary Visualizer) 实作
第 4 天	主题：进阶数据挖掘技术 4
	企业使用之范围： 说明企业如何利用更进阶的分类技术来提升模型的分类效能，增加公司获利

	理论介绍: 分类之支持向量机 (Support Vector Machine) 分类之多模型整合 (Ensemble): 装袋 (Bagging)、增强 (Boosting) 学习 分类之多模型整合 (Ensemble): 随机森林 (Random Forest)
	实作, 使用数据集: 电信客户流失数据集、糖尿病数据集
第 4 天	主题: 进阶数据挖掘技术 5
	企业使用之范围: 说明企业如何利用预测 (Prediction) 技术之线性回归、回归树及类神经网络来建立数值预测模型-如预测客户之年收入, 以利公司设计营销活动
	理论介绍: 预测之简单线性回归 (Simple Linear Regression)、复回归 (Multiple Linear Regression) 及相关系数 (Correlation Coefficient) 预测之回归树 (Regression Tree) 预测之类神经网络 (Neural Network)-非线性回归 预测模型的效能评估 (MAE, MSE, RMSE, R2, Adjusted R2, AIC & BIC)
	实作, 使用数据集: 汽车油耗数据集、CPU 效能数据集
第 4 天	主题: 进阶数据挖掘技术 6
	企业使用之范围: 说明企业如何利用聚类技术来建立客群分析模型, 使企业针对不同客群, 采用不同之营销策略, 让银行获利最大化
	理论介绍: 相似性的衡量及距离的计算 阶层式聚类 (Hierarchical Clustering) *单一链结法 (Single Linkage Method) *完全链结法 (Complete Linkage Method) *平均链结法 (Average Linkage Method) *中心法 (Centroid Method) *Ward' s 法 (Ward' s Method) 分割式聚类 (Partitioning Clustering)

	*K-中心 (K-Means) 法 *自我组织映像 (Self-Organizing Maps; SOM) 法 *两步 (Two-Steps) 法 群数的判断 (R^2, Semi-Partial R^2)
	实作, 使用数据集: 银行客户数据集、鸢尾花数据集
第 5 天	主题: 进阶数据挖掘技术 7
	企业使用之范围: 说明企业如何利用关联及序列分析技术来建立交叉销售 (Cross-Selling) 及提升销售 (Up-Selling) 模型, 以提升公司获利
	理论介绍: 关联规则的评估指针 (支持度、信赖度及提升度) 关联分析 (Association Analysis): Apriori 算法 关联规则的延伸 (虚拟商品、负向关联规则及相依性网络) 序列分析的评估指针 (支持度、信赖度) 序列分析 (Sequence Analysis): AprioriAll 算法 序列分析的延伸 (状态移转网络)
	实作, 使用数据集: 零售购物篮数据集
	主题: 项目实作一
	企业使用之范围: 电信商品营销项目
第 5 天	牵涉理论: 数据探索及数据质量报告 数据前处理 (空值、离群值处理) 分类技术 (贝式网络、决策树、神经网络、罗吉斯回归等) 进阶分类技术 (支持向量机、多模型整合之装袋、增强、随机森林等) 建模结果报告及模型解读
第 6 天	主题: 项目实作二
	企业使用之范围: 银行信用风险评分卡建置

	牵涉理论： 样本期间及好坏客户定义 数据探索及数据质量报告 数据前处理(区隔化模型建置、数值型属性趋势离散化) 单变量分析(母体稳定度指针、信息价值) 罗吉斯回归、WOE、Dummy Variable 罗吉斯回归转评分卡 建模结果报告(Gini 系数、KS 值、ROC 曲线)及模型解读
	主题：项目实作三
	企业使用之范围： 零售大数据促销项目
	牵涉理论： 数据探索及数据质量报告 3 亿 5 千万笔交易数据处理方法 数据前处理(空值、离群值处理、类别型属性趋势离散化、衍生字段处理) 分类技术(贝式网络、决策树、神经网络、罗吉斯回归等) 进阶分类技术(支持向量机、多模型整合之装袋、增强、随机森林等) 建模结果报告及模型解读

主要操作案例

1. 银行客户营销案例
2. 信用评等案例
3. 电信客户分类(不同套餐选择)案例
4. 电信客户流失案例
5. 新车设计案例

次要操作案例

1. 天气(Weather)案例
2. 玻璃制品(Glass)案例
3. 电信产品跨销售(Cross-Selling)案例
4. 药物治疗(Drug Diagnosis)案例

5. 糖尿病 (Diabetes) 案例
6. 乳癌 (Breast Cancer) 案例
7. 临床路径选择 (Clinical Path) 案例
8. 电离层雷达侦测 (Ionosphere) 案例
9. 寿险推销 (Insurance Promotion) 案例
10. 影像分类 (Image) 案例
11. 便利超商选点 (Convenient Store) 案例
12. 零售促销预测 (Retail Promotion) 案例
13. 房价 (Home Price) 预测案例
14. 汽车油耗 (MPG) 预测案例
15. CPU 效能 (CPU Performance) 预测案例
16. 银行客户购买金融商品 (Financial Product) 之关联分析 (Association Analysis) 案例
17. 文具 (Stationery) 及健康美容 (Health & Beauty) 用品之关联分析 (Association Analysis) 案例
18. 银行客户购买金融商品 (Financial Product) 之序列分析 (Sequential Analysis) 案例

课程收益

- (1) 了解什么是顾客关系管理;
- (2) 了解顾客关系管理系统的架构及其组成元素;
- (3) 了解如何利用顾客关系管理系统来进行营销活动;
- (4) 了解什么是数据挖掘 (Data Mining);
- (5) 掌握数据挖掘技术的功能分类;
- (6) 掌握数据挖掘技术的绩效增益;
- (7) 了解数据挖掘技术的产业标准;
- (8) 掌握如何利用数据挖掘技术来筛选关键变量 (Key Attribute);
- (9) 掌握如何利用数据挖掘技术来进行交叉销售 (Cross-Selling);
- (10) 掌握如何利用数据挖掘技术来评估客户的信用风险 (Credit Risk);
- (11) 了解如何利用数据挖掘技术来分析顾客行为、产生商业智慧并发展营销策略。

六、SAS 数据可视化（4 天）

课程名称	包含模块	天数	具体内容
报表与可视化体系	常规报表与可视化报告	1	1、描述性数据分析与 SAS EG 的实现 2、统计制图与 SAS EG 的实现 3、报表制作与 SAS EG 的实现 4、业务分析报告标准模板与 SAS EG 实现（使用房屋价格分析案例，实现报告的第一部分：描述）
影响因素归因分析（初级）	违约因素分析	1	1、总体（随机变量）分布与抽样 2、大数定理、中心极限定理、常见统计分布 3、假设检验基础：置信区间检验、两样本 T 检验 4、方差分析与 SAS EG 的实现 5、相关分析与 SAS EG 的实现 6、卡方检验与 SAS EG 的实现 7、业务分析报告标准模板与 SAS EG 实现（使用房屋价格分析案例，实现报告的第二部分：因素检验）
	客户流失归因分析		
	客户价值分析		
客群分析（初级）	信用卡持卡人客群分析	1	1、SQL 单表查询语句与 SAS EG 的实现 2、选择列、创建列 3、筛选样本 4、汇总函数 5、SQL 多表合并语句与 SAS EG 的实现
两阶段精准营销模型	价值预测模型	1	1、线性回归和逻辑回归讲解 2、模型参数估计、虚拟变量设置、显著性 3、模型拟合优度评估 4、模型预测、预测置信区间 5、模型修正：异方差、异常值、共线性 6、建立逻辑回归 7、分类模型的模型评估 8、业务分析报告标准模板与 SAS EG 实现（使用房屋价格分析案例，实现报告的第三部分：进行预测）
	客户获取模型		

七、SAS 评分模型数据分析（4 天）

课程名称	包含模块	天数	具体内容
构建行为 信用评分卡	业务理解与 特征体系设计	1	1、SEMMA 数据挖掘流程与数据挖掘模型开发报告模板 2、业务理解、数据采集步骤规划 3、使用 Proc SQL 建立宽表：基本信息、状态信息、业务信息、财务信息等 4、数据清洗步骤 5、建立评分卡模型：逻辑回归、预测与分量化 6、模型评估 7、基于人行征信标和第三方数据的风险数据集市与风险条线的数据治理 8、编程方面，需要学习 SQL 步、SAS 宏变量、宏等内容
	数据理解与 构建特征		
	数据清洗		
	建立模型与 模型评估		
产品与 客群分析	数据降维	1	1、信息冗余与信息压缩 2、线性代数知识点回归 3、特征根与特征向量 4、变量降维与主成分分析 5、因子分析与构建波士顿矩阵 6、层次聚类法与产品分类描述 7、K-means 聚类与客群细分 8、客户画像与标签体系
	客群细分		
构建申请 评分卡	业务分析与 数据提取规范	2	本课程演练使用 SASBASE 进行评分卡开发技术，要求学员熟练掌握 Data 步、SQL 步、宏和逻辑回归 1、申请信用评分的业务理解：建模的目标函数 2、数据提取：确定标签提取准则（滚动分析、Vintage 分析） 3、拒绝推断：聚类算法、KNN 算法、截断回归算法 4、使用 SAS 进行数据转换：概化、分箱、WoE 转换 5、使用 SAS 建立模型与尺度化 6、设定额度规则 7、模型评估与生命周期管理、模型管理平台讲解 8、广义可加模型与 SAS 实现
	数据补足与 清洗流程		
	模型开发与 落地部署		

八、大数据分析应用课纲（5 天）

时间	培训主题	培训内容
第一天上午	大数据前沿知识 及 Hadoop	1. 大数据前沿知识介绍 2. 课程介绍 3. Linux 及 Ubuntu 安装和使用 4. Linux/Ubuntu 文件系统操作 5. Hadoop 的单机、伪分布 【操作】 完成 Hadoop 伪分布搭建
第一天下午	Hadoop 集群搭建 及 HDFS 操作	1. Hadoop 完全分布式模式的安装配置 2. Hadoop2.x 体系结构 3. HDFS 分布式文件系统，HDFS Shell 操作 4. YARN 的基本构成和工作原理 【操作】 完成 Hadoop 集群搭建
第二天上午	Mapreduce 思想及案例	1. MapReduce 并行计算框架 2. 使用 MapReduce 编程框架实现数据去重 3. 使用 MapReduce 编程框架实现数据排序 3. Hadoop 集群上部署和执行 MR Job 【案例】 ：论坛日志分析
第二天下午	大数据导入与存储 Tableau 大数据分析	1. mysql 数据库基础知识 2. hive 的基本语法 3. hive 安装部署 4. Tableau 的维度与度量 5. 使用 Tableau 快速创建交互式绘图 6. 利用 Tableau 的 GUI 创建交互式仪表盘 7. 使用 Tableau 连接 mysql 数据源 【操作】 ：Tableau 实践
第三天上午	Hbase 理论及实战	1. Hbase 简介、安装及配置 2. Hbase 的数据存储

		3. Hbase Shell Hbase API 4. Hbase 数据备份与恢复方法 【案例】: 针对 XX 高校学生管理信息系统中的学生选课
第三天下午	Spark 配置及使用场景	1. Scala 基本语法 2. Spark 介绍及发展历史 3. Spark standalone 模式部署 4. Spark RDD 详解 5. Spark 案例分析 【操作】: Spark1.4 Standalone 分布式集群
第四天全天	Spark 大数据分析原理	1. Spark 作业调度流程与策略 2. 集群模式运行、监控 Spark 应用程序的步骤与方法 3. Hive、Shark 简介与对比分析 4. Spark MLlib 体系结构简介 5. 机器学习简介及经典算法案例详解 6. Spark SQL 体系简介及语句流程 7. DataFrame 模型的定义、功能 【案例】: Sogou 搜索日志数据之用户行为进行分析
第五天上午	Hadoop+Spark 智慧高速大数据分析	1. 复习 Hive、Sqoop、Spark、Mysql 的配置 2. Sqoop 和 Mysql 数据库进行交互 3. Spark SQL 语句操作大数据平台上的数据 4. Tableau 进行高速大数据的漏斗分析
第五天下午	Hadoop+Spark 互联网金融监管大数据分析	1. 使用爬虫进行文本数据采集 2. 使用 spark 实现文本分词 3. 互联网金融公司提取 4. 公司名词判别 5. 使用 SPARK 程序实现情感语义分析 6. 结果展示

九、Power BI 培训（2 天）

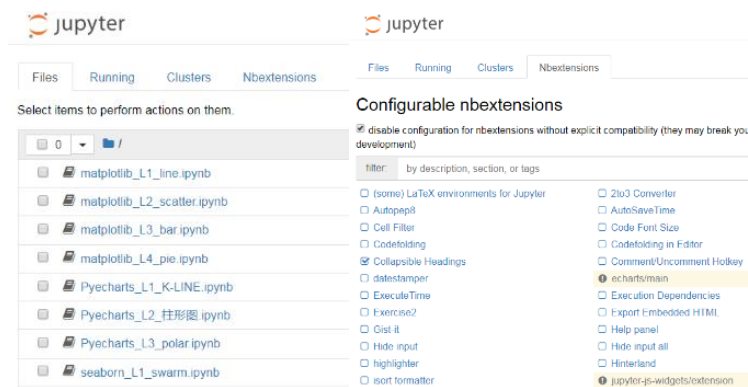
模块	内容	时长
Excel 进阶 Power BI 商业智能	1. Power BI 简介与数据可视化 2. Power Query 基本功能：数据导入、数据转换、数据组合、加载数据模式、合并查询等 3. Power Query 常用移除文本操作 4. Power Pivot 数据导入功能 5. Power Pivot 搭建多维数据集与多维数据透视 6. Power Pivot 基本功能：数据加载、计算字段、KPI、层次结构等 Power Pivot 综合应用案例：商机相关企业信息	6h
客户数据分析案例	1. 经典客户分析方法介绍 2. RFM 分析：“最近 12 个月”动态客户 RFM 分析 3. 帕累托分析：多种方法构造的交互式帕累托分析 4. 客户价值分层分析 5. 客户 ABC 分析方法	3h
销售管理数据分析 案例	销售部业务分析仪表盘报告 1. 销售管理分析-销售漏斗分析方法 2. 销售运营业务图 3. 销售管理分析 - 搭建多维数据模型：客户维度、产品维度、商机维度、地域维度、渠道维度、销售维度 4. 销售管理分析-洞察风险 5. 销售管理分析设计思路 6. Power Query 进行数据加工 7. Power Pivot 进行数据加工 8. 仪表盘展现总体趋势问题 9. 瀑布图展现各阶段占比情况 10. 了解各个阶段商机构成 11. 销售管理分析仪呈现与细节问题	3h

十、Python 基础及可视化（2 天）

第一天

（1）厉兵秣马——Python 简介与 Jupyter 配置

样例：



内容简介：学习基本的环境设置，为后期的代码编辑做准备。

（2）雾里看花——可视化基本逻辑与商业应用

内容简介：

- 数据可视化是什么？
- 数据可视化如何在不同的商业场景中进行应用？
- 当下热门的“一图看天下”是什么？
- 数据可视化的误区有哪些？

（3）无中生有——变量的理解与随机数据生成

样例：

构造样本数据1

```
In [2]: num=100
np.random.seed(1)
capital=np.random.randn(num)*10+100+np.random.randn(num)*20
np.random.seed(2)
sale=1.2*capital+20+np.random.randn(num)*20
np.random.seed(3)
cost=-0.8*capital+200+np.random.randn(num)*5

In [3]: data=pd.DataFrame([capital,sale,cost]).T
data.columns=['capital','sale','cost']
print(data.head())
```

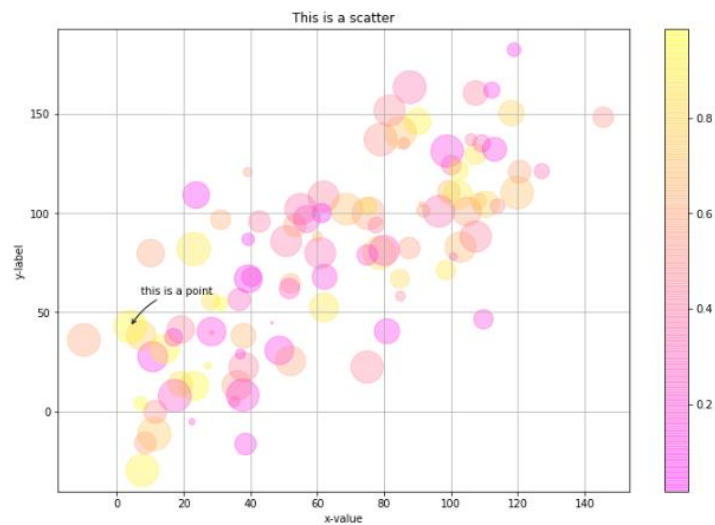
	capital	sale	cost
0	132.385558	170.527512	103.034696
1	101.639649	140.842242	120.870830
2	111.989120	111.663022	110.891192
3	104.212747	177.860712	107.312339
4	119.778881	127.865945	102.789954

内容简介:

- 产生模拟的数据
- 对数据进行拼接
- 对数据进行分割
- 数据的筛选与补充

(4) 循序渐进——散点图与条形图的进化过程

样例:

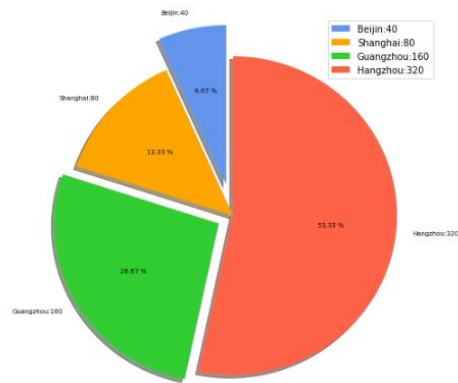


内容简介:

- 散点图和条形图的适用数据类型
- 散点图的主要参数
- 条形图的主要参数
- 散点图与条形图的实现
- 可视化过程中的数据降维问题

(5) 渐入佳境——用饼图与环形图呈现数据比例

样例:

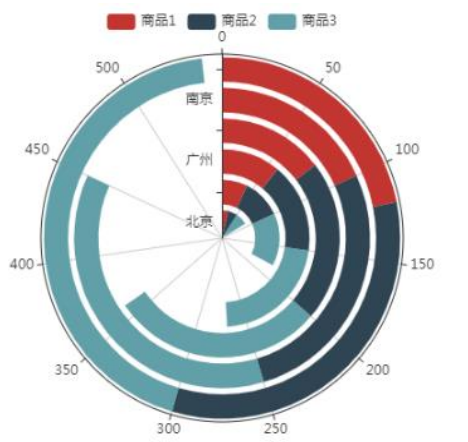


内容简介:

- 饼图与环形图的适用场景
- 饼图与环形图的关联与区别
- 饼图与环形图的重要参数
- 饼图与环形图的代码实现

(6) 秘境寻踪——南丁格尔的往事与玫瑰花图

样例:



内容简介:

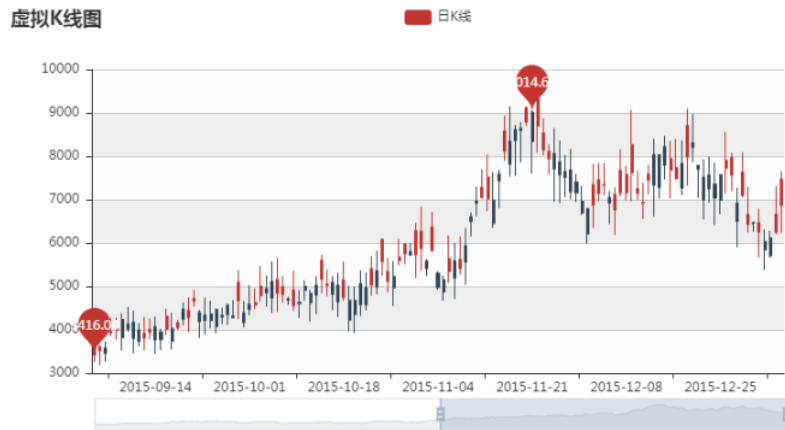
- 南丁格尔与玫瑰花图的起源
- 玫瑰花图的基本原理
- 玫瑰花图的衍生
- 相关图像的代码实现

第一天学员收获: 学员掌握 Python 基本的数据赋值, 数据分解与合并, 数据构造, 变量转换, 条件语句, 循环语句等。同时能够运用 Python 进行可视化编程。

第二天

(1) 有条不紊——折线图的设计与多线条呈现

样例：

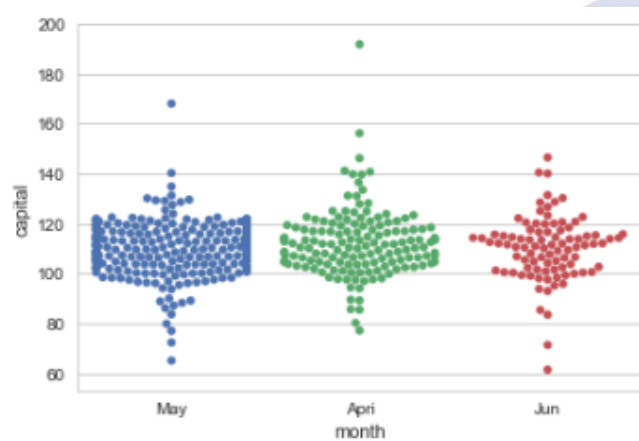


内容简介：

- 折线图的应用范围
- 折线图的主要参数
- 折线图的代码实现
- 多折线的呈现技巧

(2) 火眼金睛——分布与箱线图的异常值发现

样例：

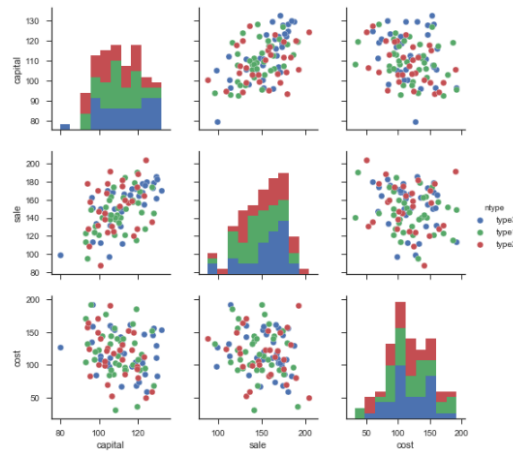


内容简介：

- 什么是数据分布
- 数据分布与异常值
- 分布与箱线图的主要参数
- 分布与箱线图的代码实现

（3）抽丝剥茧——关联与配对图的相关性识别

样例：

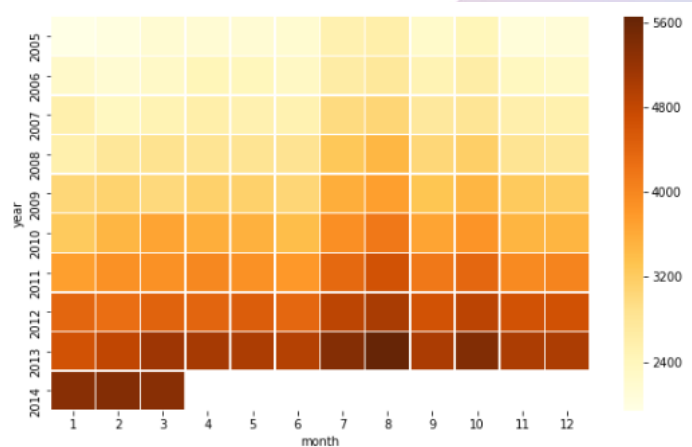


内容简介：

- 关联图与配对图的关系与区别
- 什么是相关性
- 关联图与配对图的主要参数
- 关联图与配对图的代码实现
- 图像内涵的信息解释

（4）冰火相融——日历热力图中的周期性探索

样例：



内容简介：

- 热图的使用场景
- 热图的数据要求

- 热图的核心参数
- 热图的代码实现
- 如何解读热图

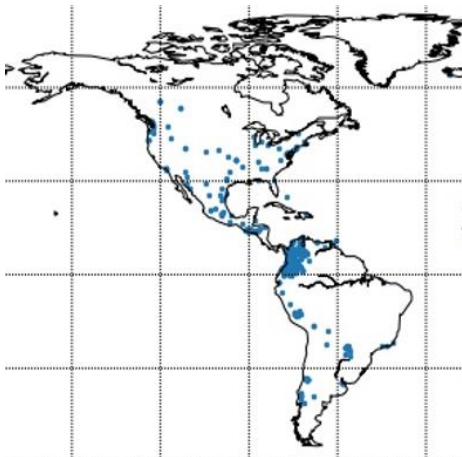
(5) 妙笔生花——可视化数据的动画设计基础

内容简介：

- 动画设计的基本原理
- 动画设计的误区
- 动画设计如何结合具体问题
- 动画的代码实现

(6) 图看天下——基于地理信息的数据可视化

样例：



内容简介：

- 什么是地理信息可视化
- 地理信息可视化的注意要点
- 地理信息可视化的数据要求
- 地理信息可视化的参数
- 地理信息可视化的代码实现与解读

第二天学员收获：学员较为熟练运用 Python 中的 Numpy, Pandas 等常用库。能够在不同的具体场景选择不同的可视化方案并实现。

十一、大数据认知与应用（4 天）

Part 1 数据化思维导论

1 走近数据化管理

1.1 什么是数据化管理（30 min）

1.1.1 数据化管理的四个层次

1.1.2 驱动业务的三驾马车

1.1.3 经验与创新

1.2 数据思维（30min）

1.2.1 立体化：点、线、面

1.2.2 元素化：时间、对象、事件

1.2.3 多角度：广度、宽度、深度

1.2.4 常见误区解析

1.3 增长的秘密（60 min）

1.3.1 从 0 到 1：新业务线的爆发式推进

1.3.2 从 1 到 N：消费升级与品牌价值

1.3.3 新零售：融合流量与场景的生态圈

1.4 数据驱动（60 min）

（企业维度）

1.4.1 品牌价值评估

1.4.2 业务线分析

1.4.3 供应链管理

（客户维度）

1.4.4 客户生命周期管理

1.4.5 大数据营销

1.4.6 客户行为分析

1.5 常见问题与困难（60min）

1.5.1 如何将问题抽象成数据的语言：概率与分布

1.5.2 如何选择恰当的维度：方差分析与群体切分

1.5.3 如何设计实验方案：对照组与实验组的选择

1.5.4 如何评估最终效果：常用指标设定与调整

1.5.5 缺失数据怎么处理：常用填补方法

1.5.6 结果不符合预期怎么处理：寻因法与纠错法

1.5.7 陌生问题没有思路怎么处理：矩阵式分析与图形化思考

1.5.8 考虑的方面过多怎么处理：降维与独立变量法

2 走近客户：想说懂你不容易

2.1 了解客户：特征画像（60min）

2.1.1 为什么我们需要做客户画像

2.1.2 客户画像的常见方法

2.1.3 数据从哪儿来

2.1.4 应用场景

2.2 需求满足（30min）

2.2.1 主动性需求：客户想要什么

2.2.2 被动性需求：我们希望给客户什么

2.3 制造惊喜（30min）

2.3.1 大数据不止可以“杀熟”：抓住客户的潜在需求

2.3.2 “快了一点”与“快乐一点”：2B 与 2C

Part 2 大数据营销与数据化客户管理(CRM)

1 大数据营销的常见应用（30min）

1.1 高效获客

1.2 精准营销

1.3 流失预警

1.4 产品线管理

1.5 客户生命周期维护

2 高效获客（60min）

2.1 常用获客指标

2.2 低端客户过滤器：识别“羊毛党”的常用方法(含案例剖析)

2.3 CRM 系统与新客获取(含案例剖析)

3 精准营销（60min）

- 3.1 常用营销指标
- 3.2 大数据活动方案设计与效果预测
- 3.3 CRM 客户标签完善路径与组合使用(含案例剖析)
- 3.4 利用 CRM 系统进行客户 A/B 试验(含案例剖析)

4 流失预警（60min）

- 4.1 常用流失预警指标
- 4.2 客户流失潜伏期识别(含案例剖析)
- 4.3 CRM 系统与流失挽回的应用(含案例剖析)

5 产品线管理（60min）

- 5.1 常用产品销售指标
- 5.2 客户覆盖率及产品线竞争力分析(含案例剖析)
- 5.3 结合 CRM 系统，产品上下架业务影响预测(含案例剖析)

6 客户生命周期维护（60min）

- 6.1 客户的两个生命周期
- 6.2 大数据在生命周期不同阶段的应用案例(含案例剖析)
- 6.3 CRM 系统与客户生命周期管理(含案例剖析)

7 CRM 系统的应用实例（30min）

- 7.1 闭环 CRM 系统案例(含案例剖析)
- 7.2 “云系统”案例(含案例剖析)

Part 3 金融领域的数字化应用

一 客户篇（3 h）

1. 整体了解：KYC 模型原理解析（60 min）

- 1.1 主观因子

- 1.2 客观因子
- 1.3 评分原理
- 1.4 应用举例
- 2. 新客：潜力筛选（60 min）
 - 2.1 渠道分析
 - 2.2 羊毛党分析
 - 2.3 客户价值评分与成本因子
- 3. 老客：价值提升（60 min）
 - 3.1 数据化客户档案
 - 3.2 客户成长路径
 - 3.3 客户社交网络分析
- 二 产品篇（3 h）
 - 1. 产品体系规划（60 min）
 - 1.1 产品分类
 - 1.2 产品网络分析
 - 1.3 引流产品、中途岛产品与粘性产品
 - 3. 组合推荐（60 min）
 - 3.1 常用推荐模型
 - 3.2 推荐效果评估
 - 3.3 结合客户标签的推荐
 - 3.4 结合行为数据的客户 A/B 试验
 - 3. 风险管理（60 min）
 - 3.1 异常行为识别
 - 3.2 潜在风险敞口管理
 - 3.3 平衡型风险管理

Part 4 应用方法与工具

由于时间限制，这部分以介绍每个工具的基本功能和应用场景为主，包括简单的操作展示和基础类的应用。后续建议根据实际需要重点选择学习某一工具。

一 分析类工具（3 h）

1. Excel (60 min)
 - 1.1 Excel 日常数据处理流程
 - 1.2 常用技巧分享
 - 1.3 案例：自动化数据处理模板设计
2. R (60 min)
 - 2.1 R 的介绍：主要特点与功能
 - 2.2 常用处理流程和场景
 - 2.3 案例：R 语言做客户分类
3. Python (60 min)
 - 3.1 Python 的介绍：主要特点与功能
 - 3.2 常用处理流程和场景
 - 3.3 案例：智能作诗

二 展示类工具 (3 h)

1. Excel (60 min)
 - 1.1 Excel 报表的设计思路
 - 1.2 案例：监控周报设计
2. PPT (60 min)
 - 2.1 PPT 报表的设计思路
 - 2.2 案例：分析型月报设计
3. Tableau (60 min)
 - 3.1 Tableau 介绍：主要特点与功能
 - 3.2 案例：交互式报表设计

十二、大数据治理实战（4.5 天）

一、培训收益

通过此次课程培训，可使学习者获得如下收益：

1. 掌握数据治理、DAMA DMBOK 数据管理知识体系核心内容
2. 理解数据架构设计与数据标准、元数据管理等的基本知识；
3. 掌握企业级数据架构下的数据建模方法；
4. 理解数据质量、数据能力成熟度 DMM、主数据管理、数据安全等内容；
5. 理解数据治理落地实施最佳实践，各数据管理核心领域的重点、难点的深度解析和行业最佳实践案例；
7. 培养数据治理、数据管理的问题思考和方案设计、实施执行能力。

二、培训对象

学习者的职业定位：

1. CIO 首席信息官
2. CDO 首席数据官
3. 业务部门的信息化负责人与业务骨干
4. 数据管理团队、数据服务及数据运营团队
5. 数据资产管理产品经理及数据管理专家

三、培训特色

1. 理论与实践相结合、案例分析与理论穿插进行；
2. 专家精彩内容解析、学员专题讨论、分组研究；
3. 通过全面知识理解、专题技能掌握和安全实践增强的授课方式。

四、日程安排

第一天 上午 大数据治理目标与定位

1. 什么是数据治理？
2. 数据治理的基本理论
3. 数据管理的基本概念
4. 国内外数据治理发展及趋势

第一天 下午 数据架构管理

1. 企业架构基本理论
2. TOGAF 企业架构概述

3. 什么是数据架构?
4. 数据架构的基本概念及内容组成
5. 数据架构的建设方法
6. 数据架构与数据治理之间的关系
7. 数据架构的相关实践案例

第二天 上午 数据标准管理

1. 数据标准的基本概念
2. 数据标准的分类
3. 业务术语标准化
4. 参考数据和主数据标准化
5. 数据元标准化
6. 指标数据标准化
7. 数据标准的建设方法
8. 数据标准与数据治理之间的关系
9. 数据标准化的案例

第二天 下午 元数据管理

1. 元数据的基本概念
2. 元数据的分类
3. 元数据管理
4. 元数据管理工具
5. 元数据管理案例, 构建企业数据资源目录

第三天 上午 数据质量管理

1. 数据质量基本概念
2. 全面数据质量管理及通用方法
3. 数据质量评估维度
4. 数据质量管理手段及工具
5. 数据质量管理案例

第三天 下午 数据能力成熟度 DMM

1. DMM 数据能力成熟度评估模型
2. DMM 数据能力成熟度评估方法及过程

3. DMM 数据能力成熟度评估案例

第四天 上午 主数据管理

1. 参考数据和主数据的基本概念
2. 主数据管理的主要内容和实施难点
3. 主数据管理工具
4. 主数据管理方法及实施方案
5. 主数据管理案例

第四天 下午 数据安全治理

1. 数据安全治理基本概念及内容框架
2. 数据安全分级
3. 数据安全策略
4. 数据安全治理
5. 数据安全审计
6. 数据安全治理案例

第五天 上午 行业数据治理实践

1. 能源行业数据治理实践分享
2. 银行、金融行业数据治理实践分享
3. 通信行业数据治理实践分享

十三、如何撰写有价值的数据分析报告（1-2 天）

篇	章	内容
Part 1 理论篇	数据报告的核心	- 目标 - 原则 - 效果
	数据报告的基本步骤	- 数据收集：内部数据与外部数据 - 数据处理：清洗与筛选 - 确定方法：分析与模型、实验与分组 - 总结结论：解释型、结论型与预测型 - 结果呈现：时效性、互动性与准确性
Part 2 思维篇	数据分析的本质： 解决问题	- 问题的三种形式 - 原因寻找型问题 - 风险防范型问题 - 追求理想型问题
	数据分析的路径： 从业务到抽象，再从抽象回归业务	- 战略规划：市场与竞对分析 - 客户管理：客户特征画像 - 产品设计：品牌价值主张与产品线规划 - 活动营销：精准营销与增长黑客 - 业务运营：客群维护与日常监控
Part 3 技巧篇	图表选择	- 单维度图表 - 双维度图表 - 多维度图表 - 综合信息图表
	常见问题	- 如何将问题抽象成数据的语言 - 如何选择恰当的分析维度 - 如何设计实验方案：对照组与实验组的选择 - 如何评估最终效果：常用指标设定与调整 - 缺失数据怎么处理：常用填补方法

		- 结果不符合预期怎么处理：寻因法与纠错法 - 陌生问题没有思路怎么处理：矩阵式分析与图形化思考 - 考虑的方面过多怎么处理：降维与独立变量法
Part 4 场景篇	方案介绍型报告	- 战略规划 - 活动营销 - 产品设计
	结果展示型报告	- 业绩展示 - 活动效果 - 市场分析
	原因剖析型报告	- 业绩问题 - 预警评估 - 落差归因
	常规周期型报告	- 汇报类 - 监控类 - 高管类
	创意分享型报告	- 团队内部 - 公司内部 - 市场客户
Part 5 工具篇	数据类报表：Excel	- 报表设计 - 常用公式 - 案例
	数据类报告：PPT	- 报告设计 - 结构与内容 - 案例
	交互式报告：Tableau	- 设计思路：数据来源、交互环节、逻辑关联 - 基本技巧 - 案例
Part 6	基于客户相关行业设计，	- 分析情景

综合案例篇	一般包括如下部分	- 需求确定 - 报告类型 - 工具选择 - 结构拆分 - 分析过程（节选展示） - 结果呈现（节选展示）
-------	----------	---

十四、SPSS 数据分析课程（5 天）

第一天 SPSS 软件初探

1. 数据分析全貌

- *业务与项目理解：阅读项目简报与报告
- *统计角色：因变量测量
- *统计角色：自变量的选择
- *数据描述：分类与连续变量及可视化
- *数据预分析：异常值、缺省值、共线性处理等
- *构建统计模型：线性回归模型
- *模型假设与修正：假设条件与处理方案
- *模型应用：结构问题与预测问题
- *模型评价：信度与业务

2. 软件丛中的 SPSS

- *软件：SAS、R、PYTHON 等
- *运行速度测试
- *统计功能与数据挖掘功能评析
- *优势互补：如何学习软件

3. 功能布局与特征

- *软件功能分布与学习
- *算法集成与 3.0 版本
- *算法功能解析：开发与使用

4. 数据分析师需要如何看待数据

- *行的信息：数据质量与统计方法
- *列的信息：变量、测量级别、角色
- *单元格信息：文本编码、缺失填补
- *数据分析师为谁而工作：数据分析目标

SPSS stat 实作：短期绩效评估

第二天 数据源在哪里

1. 数据录入与数据模板

- *手工录入技巧：快速录制
- *数据模板：值标签、测量级别、类型、缺失的定义

2. 访问本地文件

- *excel 文件导入：表与单元格的导入
- *文本文件导入：文本特性、数据分割、抽样
- *本地文件优势评析：不同格式对比

3. 数据库数据

- *访问数据库：软件与数据库连接、mysql
- *数据质量评估：数据描述与质量评估

4. 问卷编制与数据处理

- *问卷与量表的设计思路对比
- *问卷设计架构：理论与业务理解
- *问卷题目设计：题目形式、措辞、测量、精确度
- *判断问卷质量的优劣：量表与问卷的信效度
- *问卷设计注意事项：规则与流程

5. 抽样技术

- *为什么需要抽样：样本与误差
- *抽样样本确定：小数据与大数据抽样场景
- *抽样质量评估：指标评价法
- *软件实现：简单随机抽样和分层抽样

SPSS stat 实作：（1）短期绩效评估，（2）问卷设计纲要

第三天 数据描述：统计量与可视化

1. 分类变量描述

- *分类占比与众数、中位数
- *分类比例平衡与异常值处理
- *变量可视化：波动风险、累计比例、趋势评估

2. 连续变量描述

- *大众群体与均值、中位数
- *离散指标与方差、分位距
- *变量可视化：分布信息、拐点意义、极值的新意义

3. 绘制统计图形的注意事项

- *条形图与折线图的整合
- *饼图占比与条形图
- *分布与蝴蝶图
- *箱图与高低波动
- *散点图与气泡图
- *分组与嵌板的图形绘制

4. 分析报告撰写

- *报告中图表布局
- *图形元素注意事项
- *图形设计与模板
- *表格格式设计与模型

SPSS stat 实作：客户违约信息评估与预测

第四天 假设检验

1. 假设检验原理

- *假设检验流程：假设、小概率、证伪
- *原假设探析：猜测原假设
- *备择假设与归因性分析
- *P 值在哪些场景中应用
- *小数据方法论：大数据方法论的对比

2. t 检验

- *t 检验原理：差异与误差
- *单样本 t 检验：正态分布与理论值确定
- *独立样本 t 检验：分组、异方差问题

*t 检验与算法家族：一般线性模型

3. 方差分析模型

*方差分析原理：差异与误差

*方差分析：均值稳定性、正态变换

*方差分析：方差的稳定性、非参检验

*单因素方差分析：差异分析与趋势追踪

*方差分析：残差图诊断

4. 相关分析

*相关分析原理：方差的延伸

*相关分析种类与使用

*相关应用：信度与风向标

SPSS stat 实作：（1）不同技术水平的间产品差异分析（2）工作能力差异判断

第五天 相关与回归

1. 回归分析基础

*回归分析：相关性的趋势

*回归系数与信度解释

*回归预测与归因探析

*中介效应分解

*调节效应分析

SPSS stat 实作：（1）产品元素的影响因素及预测分析（2）管理特征的中介效应分解

2. logistic 回归分析

*卡方分析

*卡方分析原理：期望与频数

*OR 效应与 RR 效应的解释

*模型修正技术：工具变量

*市场细分：评分卡制作

SPSS stat 实作：欺诈与影响因素间的关系

3. 主成分模型

*主成分模型与回归：算法、操作

*主成分优劣的评估标准：第一主成分特征

*主成分的综合应用：降维与主成分回归

SPSS stat 实作：电商用户潜在购买动机评估

十五、R 语言数据挖掘培训（3 天）

第一天：基础

第一部分：数据挖掘基础

- 1、数据挖掘在政府部门及各行业的应用
- 2、数据挖掘的起源、定义及目标
- 3、数据挖掘的发展历程
- 4、数据挖掘技术的功能分类

第二部分：R 语言基础

- 1、R 与 RSTUDIO 的下载和安装
- 2、R 的基本对象
- 3、R 的基本数据类型
- 4、R 的程序控制
- 5、R 函数与包介绍
- 6、R 的日期与时间的数据类型
- 7、在 R 中读写数据

第二天：实际案例操作一

说明：如何利用聚类技术来建立客群分析模型，使企业针对不同客群，采用不同之营销策略，让公司获利最大化。

- 1、客户画像和客户分群
- 2、客户特征提取与 FRM 方法
- 3、相似性的衡量及距离的计算
- 4、阶层式聚类(Hierarchical Clustering)
 - *单一链结法(Single Linkage Method)
 - *完全链结法(Complete Linkage Method)
 - *平均链结法(Average Linkage Method)
 - *中心法(Centroid Method)
 - *Ward' s 法(Ward' s Method)
- 5、分割式聚类(Partitioning Clustering)
 - *K-中心(K-Means)法

*自我组织映像(Self-Organizing Maps;SOM)法

*两步(Two-Steps)法

6、群数的判断(R^2 , Semi-Partial R^2)

7、联规则的评估指针(支持度、信赖度及提升度)

8、关联分析(Association Analysis): Apriori 算法

9、关联规则的延伸(虚拟商品、负向关联规则及相依性网络)

10、序列分析的评估指针(支持度、信赖度)

11、序列分析(Sequence Analysis): AprioriAll 算法

12、序列分析的延伸(状态移转网络)

第三天上午：实际案例操作二

说明：企业如何利用分类技术之神经网络、罗吉斯回归来建立信用评分(Credit Scoring)模型，以降低公司损失。

1、分类之神经网络(Neural Network)

2、分类之罗吉斯回归(Logistic Regression)

3、分类模型的评估

*ROC Curve, AUC

*GINI Index, KS

4、类之支持向量机(Support Vector Machine)

5、分类之多模型整合(Ensemble): 装袋(Bagging)、增强(Boosting)学习

6、分类之多模型整合(Ensemble): 随机森林(Random Forest)

7、分类之深度学习(Deep Learning)

第三天下午：考试辅导

主要以模考和讲解的方式进行

十六、数据化运营课程培训（2 天）

第一部 数据化管理导论

1 什么是数据化管理

1.1 四个主要的方向（30min）

- 1.1.1 战略规划
- 1.1.2 策略分解
- 1.1.3 市场营销
- 1.1.4 业务运营

1.2 数据化管理流程图（30min）

- 1.2.1 明确需求
- 1.2.2 设定指标
- 1.2.3 收集数据
- 1.2.4 整理维度
- 1.2.5 分析建模
- 1.2.6 结果展示
- 1.2.7 应用方案
- 1.2.8 迭代优化

1.3 行业应用展示（30min）

- 1.3.1 物流行业应用：微观层（配送动态规划）、过渡层（供应链数据监控）、宏观层（数字化商品管理）（含案例介绍）
- 1.3.2 电商行业应用：动态客户管理与长线维护（含案例介绍）
- 1.3.3 金融行业应用：精准营销与产品线交叉运营（含案例介绍）

2 数据化管理的理论基础

2.1 知识图谱：常见问题与困难（60min）

- 2.1.1 如何将问题抽象成数据的语言：概率与分布
- 2.1.2 如何选择恰当的维度：方差分析与群体切分
- 2.1.3 如何设计实验方案：对照组与实验组的选择
- 2.1.4 如何评估最终效果：常用指标设定与调整

- 2.1.5 缺失数据怎么处理：常用填补方法
- 2.1.6 结果不符合预期怎么处理：寻因法与纠错法
- 2.1.7 陌生问题没有思路怎么处理：矩阵式分析与图形化思考
- 2.1.8 考虑维度过多怎么处理：降维与独立变量法
- 2.2 分析方法（60min）
 - 2.3.1 立体化：点、线、面
 - 2.3.2 元素化：时间、对象、事件
 - 2.3.3 多角度：广度、宽度、深度
 - 2.3.4 常见误区解析
- 2.3 常用模型（60min）
 - 2.3.1 未来数据预测：回归与时间序列分析
 - 2.3.2 目标客户筛选：逻辑回归分析
 - 2.3.3 营销方案设计：聚类分析与决策树分析
 - 2.3.4 客户管理方法：RFM 模型及定制化拓展
- 2.4 数据展示（60min）
 - 2.4.1 流量渠道与获客分析：PV/UV 和客户漏斗图
 - 2.4.2 客户成长周期与长尾效应：帕累托图与关键节点识别
 - 2.4.3 产品线交叉分析：客户覆盖率与新产品定位
 - 2.4.4 动态客户档案：雷达图与散点图
- 2.5 数据项目管理（30min）
 - 2.5.1 团队技能搭建
 - 2.5.2 进阶式能力提升方案
 - 2.5.3 常用数据工具介绍
 - 2.5.4 数据项目流程与管理

第二部 大数据营销与实践

- 1 大数据营销的常见应用（30min）
 - 1.1 高效获客
 - 1.2 精准营销
 - 1.3 流失预警
 - 1.4 产品线管理

1.5 客户生命周期维护

2 高效获客 (60min)

2.1 常用获客指标

2.2 低端客户过滤器：识别“羊毛党”的常用方法(含案例剖析)

2.3 CRM 系统与新客获取(含案例剖析)

3 精准营销 (60min)

3.1 常用营销指标

3.2 大数据活动方案设计与效果预测

3.3 CRM 客户标签完善路径与组合使用(含案例剖析)

3.4 利用 CRM 系统进行客户 A/B 试验(含案例剖析)

4 流失预警 (60min)

4.1 常用流失预警指标

4.2 客户流失潜伏期识别(含案例剖析)

4.3 CRM 系统与流失挽回的应用(含案例剖析)

5 产品线管理 (60min)

5.1 常用产品销售指标

5.2 客户覆盖率及产品线竞争力分析(含案例剖析)

5.3 结合 CRM 系统，产品上下架业务影响预测(含案例剖析)

6 客户生命周期维护 (60min)

6.1 客户的两个生命周期

6.2 大数据在生命周期不同阶段的应用案例(含案例剖析)

6.3 CRM 系统与客户生命周期管理(含案例剖析)

7 CRM 系统的应用实例 (30min)

7.1 闭环 CRM 系统案例(含案例剖析)

7.2 “云系统”案例(含案例剖析)

十七、互联网数据分析思维培训（2 天）

课程大纲：洞察：提升你的数据思维力

Part 1 总述

- 1 数据：既不遥远，也不陌生（30 min）
 - 1.1 如期而至：大数据时代的拐点
 - 1.2 移动互联网时代的客户关系：场景化到多元化
 - 1.3 触不到的客户：中间商困局（线上与线下）
 - 1.4 大而全 vs 小而精：数量与质量的平衡
 - 1.5 2C 数据分析：冷启动、轻成本、高效率
- 2 从“有数据”到“用数据”（30 min）
 - 2.1 数据化管理的四个层次
 - 2.2 驱动业务的三驾马车
 - 2.3 经验与创新
- 3 增长的秘密（30 min）
 - 3.1 从 0 到 N：新业务线的爆发式推进
 - 3.2 从 1 到 N：消费升级与品牌价值
 - 3.3 新客户关系：流量 plus 场景

Part 2 思维篇

- 1 分析方法（60 min）
 - 1.1 立体化：点、线、面
 - 1.2 元素化：时间、对象、事件
 - 1.3 多角度：广度、宽度、深度
 - 1.4 常见误区解析
- 2 通用流程（60 min）
 - 2.1 框架：常见步骤
 - 2.2 分解：找到问题的关键

2.3 交叉：原因的锁定

2.4 对策：制定解决方案

3 常用模型（60 min）

3.1 未来数据预测：回归与时间序列分析

3.2 目标客户筛选：逻辑回归分析

3.3 营销方案设计：聚类分析与决策树分析

3.4 客户管理方法：RFM 模型及定制化拓展

4 数据展示（60 min）

4.1 流量渠道与获客分析：PV/UV 和客户漏斗图

4.2 客户成长周期与长尾效应：帕累托图与关键节点识别

4.3 产品线交叉分析：客户覆盖率与新产品定位

4.4 动态客户档案：雷达图与散点图

5 常见问题与困难（60 min）

5.1 如何将问题抽象成数据的语言：概率与分布

5.2 如何选择恰当的维度：方差分析与群体切分

5.3 如何设计实验方案：对照组与实验组的选择

5.4 如何评估最终效果：常用指标设定与调整

5.5 缺失数据怎么处理：常用填补方法

5.6 结果不符合预期怎么处理：寻因法与纠错法

5.7 陌生问题没有思路怎么处理：矩阵式分析与图形化思考

5.8 考虑维度过多怎么处理：降维与独立变量法

6 常用工具介绍（30 min）

6.1 分析类工具

6.2 挖掘类工具

6.3 可视化工具

Part 3 应用篇

1 业绩知多少：战略与战术（60 min）

1.1 现在怎么样：描述性分析

1.2 未来会如何：推断性分析

1.3 原因是什么：挖掘性分析

2 产品怎么管：数据化的产品线运营（60 min）

2.1 市场导向：品牌价值与竞对分析

2.2 客户导向：客户价值主张

2.3 二八法则：产品线的广度、宽度与深度

3 客户怎么想：用数据走进客户（60 min）

3.1 了解客户：读者知多少

3.2 满足客户：主动性需求与被动性需求

3.3 制造惊喜：从“焦虑”到“治愈”

4 联结：运营与营销（60 min）

4.1 流行与个性：三步法完成精准营销

4.2 忠诚度陷阱：会员体系的防坑秘籍

4.3 价值提升计划：客户生命周期管理

4.4 爆点：活动策划中的数据驱动

十八、金融数字化转型人才训练营（10 天）

篇	章	内容
第一篇： 数据资产规划和管理	一、数据产品规划	1.1 数据应用为导向的数据产品组合设计 （1）业务战略与绩效评估体系介绍 • 企业愿景与业务战略 • 绩效评估体系的构建 （2）基于业务绩效评估体系的数据应用规划 • BI 应用规划 • 数据挖掘应用规划 1.2 基于企业 IT 架构的数据资源分析 （1）业务系统架构分析 （2）数据资产盘点
	二、敏捷数据产品开发	2.1 工具和度量 2.2 市场研究和调研方法 2.3 数据产品流程 （1）编写工作计划 （2）产品愿景对齐 （3）用户画像 （4）需求分析工具 （5）原型设计 （6）技术方案设计 （7）交付计划 （8）最终报告展示 2.4 数据产品生命周期管理
	三、数据分析常用方法	3.1 商业数据分析步骤 3.2 商业分析思维框架 3.2 业务分析模版 3.4 数据挖掘模版

	四、数据资产管理	4.1 数据资产管理体系 4.2 数据治理
	五、数据工程和数据平台	5.1 构建数据模型 5.2 数据平台架构 5.3 数据中台架构
第二篇： 智能客群运营	一、理论篇： 智能客群运营理论	1.1 数字化运营理论：发现问题 1.2 数字化营销理论：解决问题 1.3 数字化运营和数字化营销一体化
	二、实现篇： 智能客群运营实现体系	2.1 基于 NES 的客群运营监控 2.2 数字化的营销体系（模型库、标签库和 CRM 系统等） 2.3 数字化的营销闭环
	三、工具篇： 聚类分析及其在银行中的应用	3.1 聚类分析 3.2 K-means 聚类 3.3 层级聚类 3.4 谱聚类 3.5 聚类分析及其在银行中的实践
	四、工具篇： 推荐算法及其在银行中的应用	4.1 关联规则算法 4.2 协同过滤算法 4.3 产品推荐应用 4.4 银行产品推荐架构
	五、工具篇： 社区发现算法在银行中的应用	5.1 图论基本概念和实现 5.2 社区发现算法 5.3 银行交易圈及其营销应用

第三篇： 智能信用风控	一、综述部分	1.1 消费信贷全生命周期风险管理 1.2 消费信贷常见产品及基本要素 （1）消费信贷概念 （2）消费信贷参与主体 （3）常见消费信贷产品 （4）产品风险点 1.3 ABC 卡介绍 （1）ABC 评分卡介绍与特点 （2）在消费信贷风险管理中的应用
	二、贷前、贷中 风控模型	2.1 自动化信贷审批 （1）自动化贷款审批基本框架 2.2 贷款人识别 （1）生物识别与身份验证技术 （2）在审批流程中的应用 2.3 信贷准入 （1）监管性准入 （2）政策性准入 （3）黑名单性准入 2.4 信贷规则 （1）组合策略 2.5 申请信用评分卡 （1）业务理解 （2）数据获取 （3）构建模型 2.6 授信定价 （1）基于风险的差异化定价 （2）定价策略 2.7 申请阶段的数据监控 （1）申请信息监控

		(2) 策略监控 (3) 模型监控 (4) 贷款质量监控
	三、贷后风控模型	3.1 行为评分卡 (1) 业务理解 (2) 数据获取 (3) 构建模型 3.2 额度管理 (1) 风险差异化的额度调整策略 (2) 续贷客户额度策略 3.3 还款预警 (1) 还款预警
	四、催收策略模型	4.1 催收评分卡 (1) 业务理解 (2) 数据获取 (3) 构建模型 4.2 催收策略 (1) 基于催收评分卡的不同阶段的催收策略
第四篇： 智能操作风控	一、操作风险概念、 体系和特点	1.1 操作风险概念 (1) 新巴塞尔协议相关解析 (2) 操作风险场景 • 反欺诈 • 反洗钱 • 反舞弊 1.2 操作风险防控体系 (1) 流程制度 (2) 人员体系 (3) 数据体系 (4) 架构设计

		(5) 技术手段 1.3 操作风险建模特点 (1) 正样本积累定义难 (PU 问题) (2) 样本不平衡问题 (3) 高度非线性模型 (4) 特殊的评价指标
	二、信用卡盗刷反欺诈场景建模	2.1 任务和数据介绍 2.2 数据质量检查 2.3 模型搭建 (1) PU 问题 (2) 不平衡问题 (3) 无监督算法 (4) 有监督算法 2.4 代码演示讲解
	三、反洗钱建模	3.1 任务和数据介绍 3.2 数据质量检查 3.3 常见建模框架 (1) 基于规则引擎案例 (2) 基于原始交易流水 3.4 业务理解和特征工程 3.5 模型构建与评价 (1) 模型搭建 (2) 特征重要性 with 业务解释 3.6 图谱展示与案例分析 3.7 代码演示讲解
	四、反薅羊毛建模	4.1 任务和数据介绍 4.2 数据质量检查 4.3 特征工程

		4.4 模型搭建与融合 4.5 代码演示与讲解
第五篇： 数据中台和 AI 中台	一、数据集成平台	1.1 数据集成 （1）数据集成平台架构 （2）数据集成工具 • ETL • 调度平台 • 数据资源目录 （3）数据模型 • 三范式数据建模 • 维度数据建模 （4）数据集成形式 • 数据仓库 • 数据迁移 • 数据合并 • 数据同步 • 数据交换 1.2 数据平台 （1）数据平台演变路线 • 各阶段数据平台架构与特点 • 不同数据平台适合的应用 （2）数据平台建设 • 基于传统数据库的数据平台建设 • 基于大数据平台的数据平台建设 1.3 数据分析 （1）一个经营分析的数据模型示例 （2）数据探索区建设 （3）BI 平台建设

	二、数据治理与数据资产管理	2.1 数据管理 (1) 数据管理体系介绍 (2) 元数据管理 • 如何元数据管理体系 • 元数据管理建设重点与难点 • 元数据管理的应用 (3) 数据质量管理 • 如何量化数据质量 • 数据质量建设的重点与难点 • 数据质量管理的应用 (4) 数据标准管理 • 建立数据标准体系 • 数据标准建设的重点难点 • 数据标准管理的应用 (5) 主数据管理 • 如何建设主数据管理 • 主数据建设的重点与难点 (6) 数据管理其他领域介绍 2.2 数据管理平台 (1) 平台建设核心技术介绍 (2) 平台建设核心功能介绍
	三、数据治理与数据资产管理	3.1 银行数据治理体系建立 (1) 数据战略制定 (2) 数据治理组织架构 (3) 数据治理流程制度 (4) 数据治理管理办法 (5) 数据治理成功保障机制 3.2 数据治理与资产管理关系 3.3 数据资产管理建设

		(1) 数据资产管理体系 (2) 数据价值评估模型 (3) 数据生命周期模型 (4) 数据资产目录 (5) 数据共享与应用管理
	四、数据中台建设与 Hadoop 生态系统	4.1 什么是数据中台？ (1) 数据中台是如何产生的 (2) 数据中台能解决什么问题 (3) 数据中台的意义 4.2 如何用 Hadoop 生态组件搭建数据中台 (1) 数据中台架构分析 (2) 数据中台能力建设 (3) 数据中台核心功能

十九、金融投行业大数据应用

第一部分 基金业绩归因分析理论

一、以 Jensen- α 模型为基础的基金业绩归因模型

- 1、基础的 Jensen- α 模型
- 2、在 Jensen 模型基础上分离择时能力的 T-M 模型
- 3、在 Jensen 模型基础上分离择时能力的 H-M 模型
- 4、在 APT 定价理论上加入市场择时的 C-L 模型
- 5、F-S 条件性模型
- 6、针对 Jensen- α 模型中数据频度进行改进的 GII 模型
- 7、Fama 提出的超额收益分解模型
- 8、波动市场选择能力方法

二、摒弃传统 Jensen- α 模型的 BHB 模型

三、可随意添加多种因素的多因素 CAPM 归因模型

四、非参数评估方法

五、基金业绩归因分析总结

第二部分 基金业绩归因分析实证研究

一、实证研究

- 1、BHB 业绩归因分析理论
- 2、证券选择效应——由选股带来的超额收益
- 3、交互效应——由资产配置和选股综合效应带来的超额收益

二、实证分析

- 1、指标确定
- 2、样本选取

三、研究方法与分析

- 1、基金层面归因分析
 - (1) 横向比较分析
 - (2) 纵向比较分析
- 2、股票组合层面归因分析
 - (1) 行业配置能力分析

(2) 仓位控制能力分析

(3) 股票量价操作能力分析

第三部分 大数据在二级市场中的应用

一、可视化分析

1、主流的可视化工具

2、Python 中的可视化实现

二、金融市场大数据的应用分类

1、证券关注度

2、新新闻

3、主题聚类

4、行业聚合

5、舆情监控

6、事件影响预测

三、金融市场大数据的建模方法

1、文本数据的获取

2、文本数据的结构化

3、结构化数据建模

四、主流券商的大数据应用案例分析

1、新闻选股

2、大数据舆情分析策略

五、主流的机器/深度学习量化策略简介

1、支持向量机算法 （包括 OneClass SVM 算法）

2、XGBoost 算法 （包括 GBRT 算法）

3、K-Means 算法

4、强化学习算法

5、神经网络算法

六、基于 Hadoop、Spark 等的大数据风控架构

七、金融市场大数据的未来