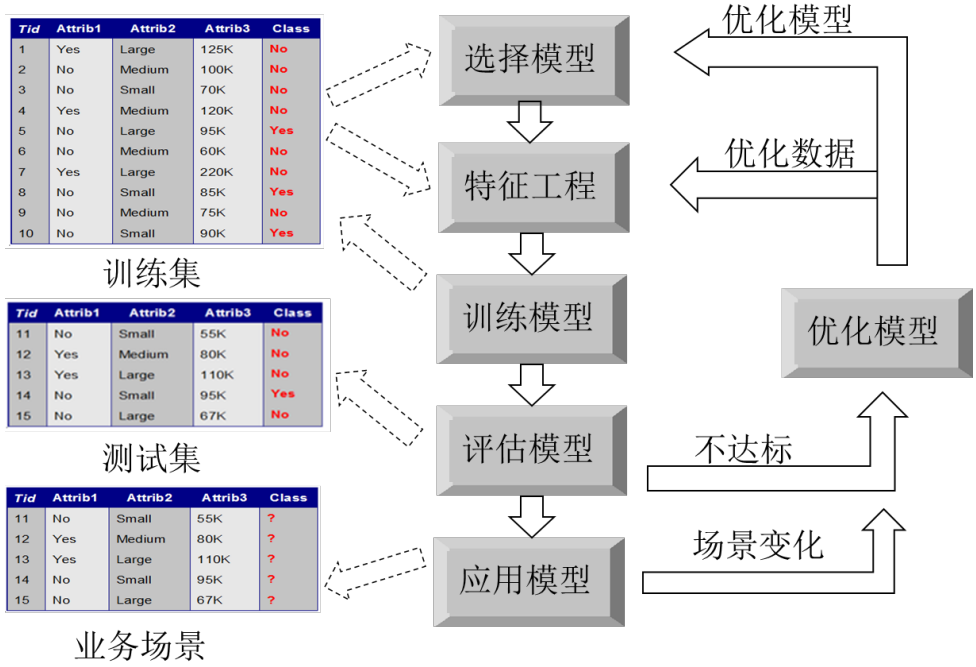
Python 数据建模及模型优化（回归篇）

【课程目标】

本课程主要讲解如何利用 Python 进行数据建模，建立数学模型，来拟合业务的各个要素之间的关系，来模拟业务的未来发展和变化。

基于真实的业务问题，在数据建模的标准过程指导下，从模型选择到特征工程，从训练模型到算法实现，从模型评估到模型优化，再到模型解读及模型应用，带领大家一步步实现一个回归预测模型。



通过本课程的学习，达到如下目的：

1、 掌握数据建模的标准流程。

2、掌握数据预处理常用的方法，包括特征筛选、变量合并等。

3、掌握回归模型的原理，以及算法实现。

4、熟练使用模型的评估指标，评估方法，以及过拟合的评估。

5、掌握模型优化的基本措施，学会欠拟合的解决方法。

6、学会过拟合评估，学会使用有正则项来解决过拟合问题。

7、熟练使用 sklearn 库的常用回归类。

8、学会超参优化的常用方法，能够设置最优超参。

【授课时间】

2-3 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、数据分析部、系统设计部、系统开发部、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】

建模流程+ 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： 预测建模基础

1、 数据建模六步法

- 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，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、训练模型及实现算法

- 模型原理
- 算法实现

5、模型评估

- 评估指标
- 评估方法
- 过拟合评估

6、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

7、模型应用

- 模型解读
- 模型部署
- 模型应用

8、好模型是优化出来的

第二部分： 回归模型评估

1、三个基本概念：SST、SSR、SSE

2、三个方面评估：指标、方法、过拟合

3、拟合程度指标

- 简单判定系数： R^2
- 调整判定系数： $\acute{R}^2$

4、预测值误差指标

- 平均绝对误差：MAE
- 根均方差：RMSE
- 平均绝对误差率：MAPE

5、信息损失准则指标

- 赤池信息准则：AIC
- 贝叶斯信息准则：BIC
- HQ 信息准则：HQIC

6、评估方法

- 原始评估法
- 留出法（Hold-Out）
- 交叉验证法（k-fold cross validation）
- 自助采样法（Bootstrapping）

7、其它评估

- 过拟合评估：学习曲线
- 残差评估：白噪声评估

第三部分： 影响因素分析

问题：如何选择合适的属性来进行建模预测？如何做特征选择/特征降维？

1、 属性筛选/变量降维的常用方法

2、 影响因素分析常用方法

- 相关分析
- 方差分析
- 卡方检验

3、 相关分析（衡量变量间的线性相关性）

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
- 相关分析的三个种类
  - 简单相关分析
  - 偏相关分析
- 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：网龄与消费水平的关系

- 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(衡量类别变量与数据变量的相关性)

问题：哪些才是影响销量的关键因素？主要因素是哪些？次要因素是哪些？

- 方差分析的应用场景
- 方差分析原理

- 方差分析前提：齐性检验
- 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
- 方差分析的四个步骤
- 分析结果解读要点

演练：终端摆放位置与终端销量有关吗

演练：客户学历对消费水平的影响分析

演练：广告形式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：排除收入后，网龄对消费水平的影响大小分析

5、列联分析/卡方检验（两类别变量的相关性分析）

- 卡方检验应用场景
- 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：银行用户违约的影响因素分析

6、属性重要程度排序/筛选

第四部分： 线性回归模型

问题：如何预测产品的销量/销售金额？

1、常用数值预测的模型

- 通用预测模型：回归模型

2、线性回归应用场景

3、线性回归模型种类

- 一元线性回归
- 多元线性回归

4、线性回归建模过程

5、带分类变量的回归建模

6、回归模型的质量评估

7、回归方程的解读

第五部分： 回归算法实现

1、基本概念

- 损失函数

2、普通最小二乘法 OLS

- 数学推导
- OLS 存在的问题

3、梯度下降算法

- 梯度概念
- 梯度下降/上升算法
- 批量梯度/随机梯度/小批量梯度
- 学习率的影响
- 早期停止法

4、牛顿法/拟牛顿法

- 泰勒公式(Taylor)
- 牛顿法(Newton)
- 拟牛顿法(Quasi-Newton)的优化
  - DFP/BFGS/L-BFGS

5、算法比较-优缺点

第六部分： 回归模型优化

6、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？

7、欠拟合解决：多项式回归

- 剔除离群值
- 剔除非显著因素
- 非线性关系检验
- 相互作用检验
- 共线性检验
- 检验误差项

案例：销量预测模型优化示例

8、过拟合解决：正则项

- 岭回归（Ridge）
- 套索回归（Lasso）
- 弹性网络回归（ElasticNet）

9、超参优化

- 手工遍历 cross_val_score
- 交叉验证 RidgeCV/LassCV/ElasticNetCV

- 网格搜索 GridSearchCV
- 随机搜索 RandomizedSearchCV

第七部分： 自定义回归模型

1、自定义回归模型

2、模型参数最优法方法

- 全局优化/暴力破解 brute
- 局部优化 fmin
- 有约束优化 minimize

3、好模型都是优化出来的

案例：餐厅客流量进行建模及模型优化

4、基于回归季节模型

- 季节性回归模型的参数
- 相加模型
- 相乘模型
- 模型解读/模型含义

案例：美国航空旅客里程的季节性趋势分析

5、新产品预测与 S 曲线

- 产品累计销量的 S 曲线模型
- 如何评估销量增长的上限以及拐点
- 珀尔曲线
- 龚铂兹曲线

案例：预测 IPAD 的销售增长拐点，以及销量上限

第八部分： 案例实战

1、客户消费金额预测模型

2、房价预测模型及优化

结束：课程总结与问题答疑。