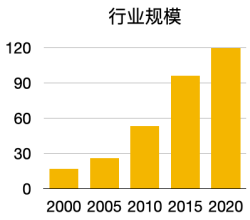
9.2 案例分析的一般流程

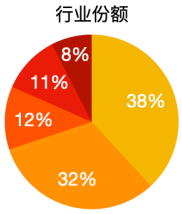
2023 年 9 月 18 日

9.1 案例分析的一般流程

1 选题与背景介绍

案例的背景介绍描述了案例所属行业的发展状况（市场规模）、行业的主要特点、行业的主要参与者（竞争格局）、商业的模式、消费者群体特点、消费场景、未来的发展趋势等。这部分撰写资料可参考商业咨询机构的调研报告。





2 研究的问题

（1）描述案例研究的商业问题，同时阐述研究目的是什么。

（2）明确案例所考察的核心变量，例如商业变量、经济水平变量、地理数据等，以及是否使用代理变量。

（3）针对变量的数据类型的不同，划分为字符型和数值型变量。

（4）同时需要考虑企业成本问题，以最小化企业的成本为导向。

变量考查

变量类型		变量名	详细说明	取值范围	备注
因变量Y					
自变量X	类别一				
	类别二				
	类别三				

3 数据采集与预处理

（1）数据的采集方面，介绍和使用网络数据采集软件，来抓取网站上的数据。市面上的网络数据采集软件很多，国内有八爪鱼，国外有 Uipath Studio 等。

（2）标注获取数据的来源、明确获取的数据内容、制定数据筛选的规则。

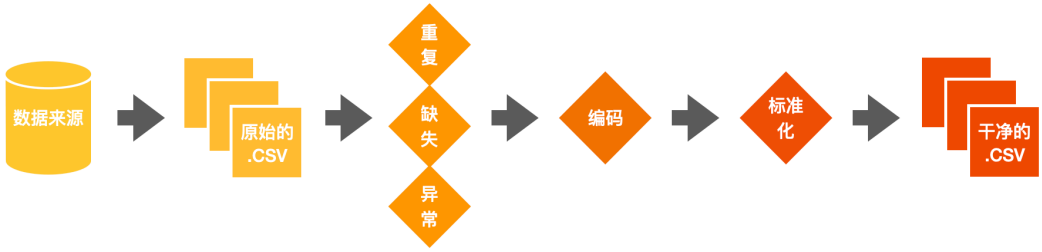
（3）对于结构化数据集的存储和读取，例如读取和存储 CSV 和 XLS 格式的文件。

（4）数据预处理，也称数据清洗，它是在我们开始分析数据和建模前，对获得数据中可能存在的问题进行排查和解决的过程。它主要包括对于数据中存在的重复问题，缺失问题，以及异常值（outliers）问题等进行剔除、填补和修正等方法。

（5）考虑数据的分布特性，如果数据样本的某一属性在之后模型使用中有正态分布要求，那么需要做对数处理。如果是时间序列可以做差分处理。

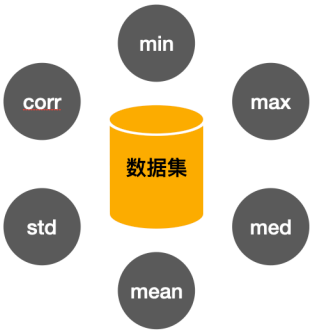
（6）编码变量也是一个必要的部分，数据集的部分特征（属性）往往属于字符型，就需要建立一个映射表，将字符型变量映射为数值型变量，例如，整数编码。

（7）数据的标准化对于某些模型，比如神经网络这类对输入敏感的模型是必须的。标准化处理包含归一化，最小最大值标准化，均值标准差标准化等。



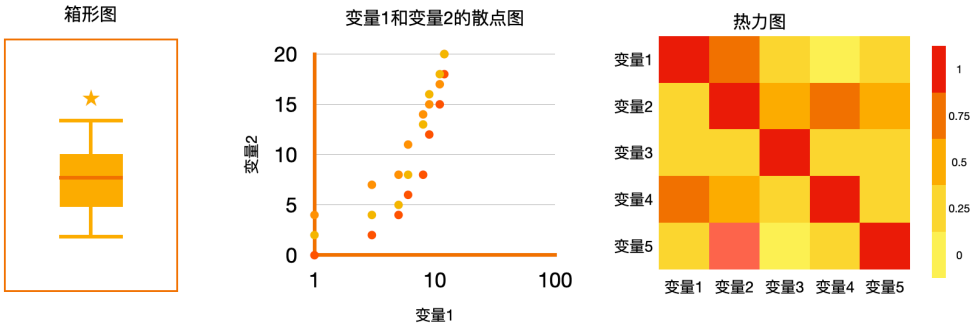
4 描述性统计

常用的描述性统计方法有最小值、最大值、均值、中位数、方差、标准差、协方差和相关系数。



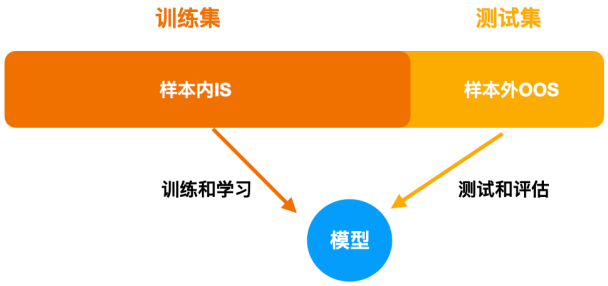
	样本数 num	均值 mean	标准差 std	最小值 min	25分位 25%	中位数 med	75分位 75%	最大值 max
属征/特征1	-	-	-	-	-	-	-	-
属征/特征2	-	-	-	-	-	-	-	-
属征/特征3	-	-	-	-	-	-	-	-

通过绘制解释变量和被解释变量之间的箱线图、散点图以及热力图，来观察变量之间是否存在一定的相关关系（正相关、负相关还是不明显相关）。



5 数据集的划分

随机对照试验（randomized controlled experiment）是统计学里一个很重要的方法。当考虑因果效应时，对于实验样本设置处理组和对照组是非常有必要的，它们主要用来消除因为抽样的非随机性造成研究过程中的偏差。在数据科学领域中，一般将数据集划分为训练集（training sets）和测试集（test sets），前者负责模型的训练任务，后者用来评价模型的表现。两者的划分比例按照经验，可以设置为



0.8:0.2。

6 模型的建立

商业数据分析的模型主要使用数据挖掘、机器学习、计量等应用统计学科的模型，负责完成回归、预测、分类、聚类等任务。

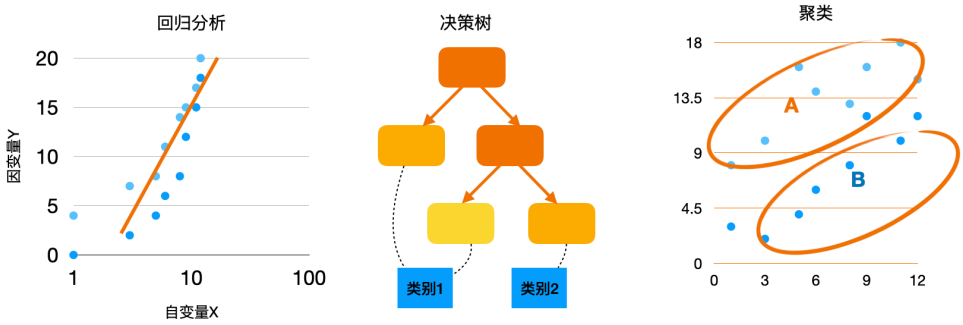
（1）回归分析（Regression Analysis）包括一元和多元线性回归，它考察了解释变量和被解释变量之间的统计相关关系。因为拥有良好的解释性，它被广泛应用在各个社会学科领域，包括经济学、管理学、心理学等领域需要定量分析的任务中。通过统计学的假设检验，考察回归解释变量的 P 值是否显著，将不显著的变量剔除出回归方程。

（2）决策树 (Decision Tree) 作为一种常见的分类模型，用来解决目标变量是非连续型变量。在构建决策树的过程中，通常采用信息熵来作为决策规则。

（3）聚类（Clustering）是一种无监督的学习模型，它将相似的对象归到同一个簇中。簇内的对象越相似，聚类的效果越好。聚类有时候也被称为无监督分类（unsupervised classification）。

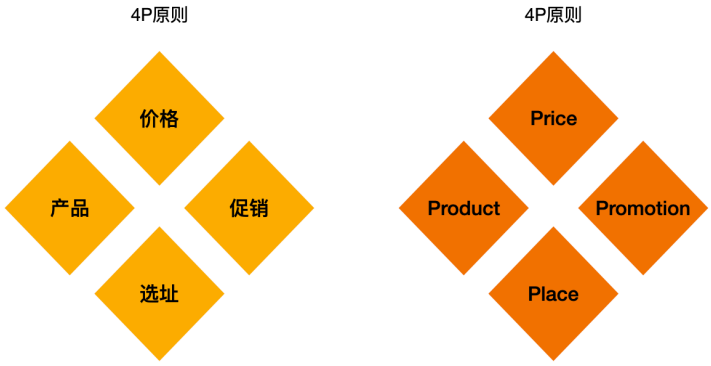
（4）人工神经网络（Artificial Neural Network）又称多层感知机，通过输入数据集的训练样本，训练得到一个映射函数，其中是输入自变量的维度，是输出因变量的维度。该近似函数可以是非线性的，既可以用于回归也可以用于分类。神经网络模型的优点：1. 能够学习数据集中的非线性关系；2. 能够进行增量学习，即基于部分新的样本，在原模型的基础上进行学习，而不需要基于全部的数据集。

（5）时间序列分析（Time Series Analysis）模型主要针对时间序列的一些特性，例如趋势、季节性周期和随机性，进行回归和预测的任务。它包含一些系列模型，例如自回归模型（AR）、滑动平均模型（MA）、自回归滑动模型（ARMA）、条件异方差模型（ARCH、GARCH）等。



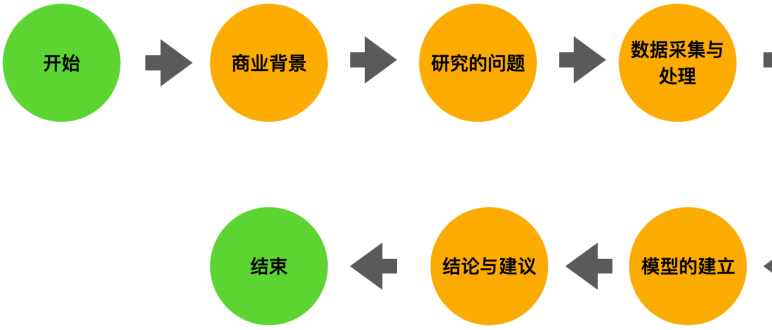
7 结论与建议

在案例撰写的结论部分，一般的需要给出建议方案，同时突出设计方案的优点，以及不足与待改进之处。很多商业的建议方案可以围绕着 4 个方面，简称 4P 原则：(1) 产品 (2) 价格 (3) 位置或选址 (4) 促销或营销



如果因为商业机密，无法获取部分企业的真实数据，模型简化了一部分处理，或者是存在一定的假设条件，一定要在结论的位置阐述清楚。

8 总结



最后，我们总结下案例分析的一般化流程，如下：

实际撰写报告的时候，可以按照以上流程按部就班，当然也可以简化流程，突出重点。

9. 案例分析报告撰写常见问题

案例分析报告撰写常见问题如下：

- 背景介绍不够详细
- 没说明数据来源、怎样获取数据的
- 数据集包含的变量、单位、取值范围
- 数据集存在哪些问题
- 背景介绍令人担忧的抄袭问题
- 需要使用 for 循环从字符串中提取信息