第 5 卷◆第 4 期◆版本 1.0◆2022 年
文章类型：论文|刊号（ISSN）：2630-4686 /（中图刊号）：380GL020

“新商科”下商业数据分析课程的教学案例设计

曹香汉
上海外国语大学贤达经济人文学院 商学院
DOI:10.12238/er.v5i4.4608

[摘　要] 数字经济对“新商科”人才的技能和素质提出了更高的要求。商业数据分析课程作为“新商科”背景下高校的热门课程,融合了多门学科(商科、统计学和计算机学),着重培养学生数据分析的思维与实践能力。本文探讨了该课程案例教学中案例的选择标准,以及筛选案例背后数据集所需要考虑的因素。同时,提供了该课程案例设计的一般化流程,供授课教师参考。
[关键词] “新商科”；商业数据分析；教学案例
中图分类号：G424.1　文献标识码：A

Teaching Cases Design of Business Data Analysis Course under "New Business"

Xianghan Cao
School of Business, Xianda College of Economics & Humanities, Shanghai International Studies University
[Abstract] Digital economy has put forward higher requirements on the skills and qualities of "New Business" talents. As a popular course in colleges and universities under the background of "New Business", the course of business data analysis integrates multiple disciplines (business, statistics and computer science) and focuses on cultivating students' thinking and practical ability of data analysis. This paper discusses the selection criteria of cases in case teaching of this course and the factors that need to be considered in selecting data sets behind cases. Meanwhile, the general process of case design of the course is provided for teachers' reference.
[Key words] "New Business"; business data analysis; teaching case

引言

在数字经济和人工智能的背景下,“新商科”教学理念逐渐渗透到高校商科类人才培育的方案中。越来越多的高校商学院开始建设商业数据分析类的课程,培养相应的人才,来应对社会经济的变革。案例教学作为商科重要的教学手段,也是商务数据分析课程授课中不可或缺的部分。传统的商业案例库虽然丰富,但多数并不具有数据分析的基础,比如数据的样本少、变量少以及分析任务不明确等问题。同时此课程案例设计的一般化流程是怎样的,这些问题亟待探讨和梳理。

1 “新商科”下商业数据分析课程

1.1“新商科”的人才培养。2019年“新商科：概念、内涵与实现路径”研讨会暨教育部高等学校工商管理类专业教学指导委员会首次系统地对比了传统商科与新商科的不同点[1]。不同于传统商科的服务形态偏向于工业经济,“新商科”服务的经济形态是数字经济。作为全球经济增长日益重要的驱动力,数字经济呈现出四个主要特征,即规模经济、技术驱动、生态化发展、动态治理[2]。数字经济对于人才的技能和素质要求,相对工业经济更高。所以,“新商科”人才培养应在传统商科培养的知识和能力的基础上,增加新技术和新思维方面的知识和能力,从而构建“商科+技术+思维”的知识与能力结构体系[3]。所以,高校教学需要打破学科边界、融合人工智能和数据科学的优势、加强案例教学占比,加大校企协同,提升学生的实践创新能力,培养出符合时代发展需要、满足行业需求的复合型商业人才[4]。

1.2商业数据分析课程。“新商科”下的商业数据分析类的课程属于多学科交叉课程(商科、统计学和计算机学三个学科),授课内容涵盖了编程语言、数据处理、建模以及商业案例写作等知识,着重培养学生数据分析的思维与实践能力。这门课程宜设置为一门平台课,授课对象适用于商学院各系。

随着数字经济、大数据和人工智能时代对于数据分析的迫切需求,这门课在高校越来越受欢迎。通过这门课程的接触,商科学生可以扩大就业面,满足企业对于商业分析师(Business Analyst)、数据分析师(Data Analyst)、行业研究员(Industry Researcher)等职位的需求。

2 课程案例的选择

2.1课程案例的选择标准。1921年哈佛大学商学院成立商业研究处,雇佣了一批学者进入商业实践领域收集和写作工商管理案例[5]。这不仅保证了哈佛大学商业教育拥有充足的案例来源,也巩固了其作为案例教学倡导者的地位。经过数十年的大力

推广,案例教学在世界范围内产生了深远而广泛的影响。

作为案例教学的支撑材料,案例(Cases)是为了满足教学目的,围绕着一个或几个问题,以事实为素材而编写的对于客观请假的描述[6]。考虑案例具有的一般特点,商业数据分析课程的案例选择与设计应突出以下：

(1)真实性。案例取材于商业或企业遇到的真实问题,不是授课教师凭个人知识和经验杜撰出来的问题。案例要尽可能得包含背景、问题以及相关的企业和行业诉求。同时案例需要附带一定的名词解释和相关的学术资料,来弥补学生的知识不足的情况。(2)典型性。案例符合常见的商业类别,具有一定的商业理论支撑。传统商业案例被划分成更细的类别。例如,哈佛商学院的案例库[7]的类别：会计、商务和政府关系、商业道德、经济学、企业与创业、金融与财务、综合管理、人力资源管理、信息科技、国际商务、市场营销、谈判、运营管理、组织行为学、销售等。在设计商务数据分析课程的案例时,也需要考虑类别,避免属于边缘、小众的商业问题。(3)数据可得性。案例需附有一定的数据集,包括数据集描述、数据集的授权使用证书。如果没有数据集或者数据集不全,就会造成数据分析任务无法开展,更不会有后续的工作。同时,需要注意数据集获取手段是否符合当地的法律,以及数据集的使用证书是否对数据集的用途有所限制。

2.2案例数据集的选择标准。数据集对于案例设计非常重要,获取数据集的方式和来源也很多。部分高校、开源数据科学社区、商业咨询机构等提供了较多开源易用的数据集,例如：

(1)Kaggle是全球最大的数据科学社区,提供了数据科学相关的竞赛和数据集[8]。很多著名企业,例如,在Kaggle平台上发布企业的数据和商业需求,并附带奖金。众多数据科学爱好者在网站上提交自己的解决方案,已赢取排名和奖金。此外,个人用户也可以上传和公开自己的数据集,开放给其他用户使用。这种竞争、开放和分享的氛围,让Kaggle在数据科学领域具有良好的口碑。(2)UCI Machine Learning Repository 是加州大学欧文分校机器学习与智能系统中心提供的一个开放数据集[9]。目前维护了600个数据集,覆盖了生命科学、心理学、计算机/软件工程、社会学、商业等领域。(3)DataShop@CMU是CMU提供的专门为学生和教育软件提供交互的数据集存储库[10]。(4)data.world提供了现代数据堆栈的企业数据目录[11]。它提供云原生SaaS平台,让数据分析师利用知识图谱使数据发现、治理和分析变得简单。

在判断数据集背后的领域是否属于商业领域后,需要进一步对得到的数据集进行筛选。筛选所考虑的因素有：

(1)任务类型：数据分析的任务类型,从数据科学的模型角度来看,主要分为有监督任务和无监督任务,有监督任务主要分为分类和回归,而无监督任务主要为聚类任务。(2)数据集大小：数据集的存储大小往往代表了计算资源的要求,分析任务的难度。在目前常规的数据分析任务中,往往需要把数据集一次性加载到内存中。如果数据集过大,就需要利用大数据技术例如Hadoop和各类数据库软件。而这个已经超越了商务数据分析这门课程的教学范围。基于主流个人电脑内存大小,将数据集的存储大小分为3类：在<10MB下为小数据集；10MB-1GB为中型数据集；在1GB以上为大型数据集。(3)样本量：指所收集的样本数量,它决定了数据集是否足够包含所研究对象的集合,应符合统计学对于样本显著性的最低要求,同时也要考虑到样本是否能够划分可能(例如划分为实验集和对照集),所以暂不考虑样本量在100以下的数据集。(4)特征数量：特征也称研究对象的属性。特征数量的多少和数据分析任务的难度有很强的关联性。当特征在10个以内时候,学生可以逐个考察特征的统计特性(例如均值、标准差和正态性)。当超过10个特征时候,就需要批量的去处理特征,并且在模型的解释性上面临挑战。

表1 部分案例数据集的选择标准参考

序号	中文标题	来源	任务类型(分类、预测、聚类、其他)	数据集大小	样本量	特征数量	主要变量类型
1	波士顿房屋价格预测	Kaggle	回归	100 kB	<1K	14	数值型
2	埃姆斯市房屋价格预测	Kaggle	回归	1MB	3K	79	数值型+字符型
3	Elo 商户类别推荐	Kaggle	回归	12MB	>300K	3	数值型
4	银行客户签约预测	Kaggle	分类	5.8MB	<50K	20	数值型+字符型
5	奥托集团产品分类挑战	Kaggle	分类	40MB	>100K	94	数值型
6	Quora 不真诚的问题分类	Kaggle	分类	159MB	>1000K	1	字符型
7	房屋信用违约风险	Kaggle	分类	192MB	>300K	122	数值型+字符型
8	桑坦德银行客户交易预测	Kaggle	分类	606MB	>200K	201	数值型
10	预测零售商店产品的销售额	Kaggle	回归	125MB	>1M	5	数值型+字符型
11	泰国化妆品零售商脸书评论数据	UCI Machine Learning Repository	聚类	366KB	>7K	11	数值型

根据以上因素,本文提供了部分案例数据集作为参考。

3 案例的设计流程

3.1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、消费场景等。设计这个部分过程中可参考商业咨询机构的调研报告。

3.2研究问题。(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。

3.3数据采集与处理。(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)在案例教学中,要求学生掌握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。

3.4描述性统计方法。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。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)。

3.5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。

3.6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)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)。(2)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。(3)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。(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$f(\cdot): R^m \to R^o$,其中m是输入自变量的维度,O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：第一,能够学习数据集中的非线性关系；第二,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)。

3.7报告的撰写。在商业数据分析的报告撰写,往往并不是单纯按照上述流程进行撰写。案例的受众往往更关住问题的描述、过程的分析和结论。所以,需要重点描述以下部分,见图1。

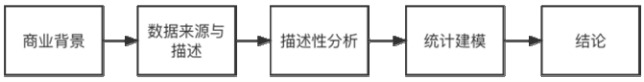

图1 案例报告撰写的主要流程

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。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。

4 结论

商业数据分析课程作为“新商科”背景下高校的热门课程,融合了多门学科的知识,对于学生的要求偏向实践性和应用性。本文通过探讨该课程案例的选择标准和数据集的筛选考虑因素帮助授课教师筛选课程案例和相关数据集,同时,提供此课程案例设计的一般化流程,供授课教师参考。

[参考文献]

[1]齐佳音,张国锋,吴联仁.人工智能背景下的商科教育变革[J].中国大学教学,2019(Z1):58-62.

[2]陈玲,孙君,李鑫.评估数字经济:理论视角与框架构建[J].电子政务,2022(03):40-53.

[3]张国平.新商科人才培养模式与实现路径[J].中国高等教育,2021(02):43-44+50.

[4]陈晓芳,夏文蕾,张逸石,等.新时代新商科的内涵及“多维度协同”培养体系改革[J].财会月刊,2021(05):107-113.

[5]杨光富,张宏菊.案例教学:从哈佛走向世界——案例教学发展历史研究[J].外国中小学教育,2008(06):1-5.

[6]张家军,靳玉乐.论案例教学的本质与特点[J].中国教育学刊,2004(01):51-53+65.

[7]哈佛大学商学院案例库:https://hbsp.harvard.edu/cases/.

[8]Kaggle数据科学社区:https://kaggle.com.

[9]UCI Machine Learning Repository: https://archive.-beta.ics.uci.edu.

[10]DataShop@CMU: https://pslcdatashop.web.cmu.edu.

[11]data.world:https://data.world.

作者简介：

曹香汉(1989--),男,汉族,上海人,硕士,上海外国语大学贤达经济人文学院,助教,研究方向：系统性金融风险、量化投资。