用商业案例学
R语言数据挖掘

经管之家 主编　常国珍 曾珂 朱江 编著

一本面向商业数据分析初学者的教材，从具体的商业数据分析案例入手，使读者掌握数据挖掘的目的、理念、思路与分析步骤。

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

CDA数据分析师系列丛书

用商业案例学
R语言数据挖掘

经管之家 主编　常国珍 曾珂 朱江 编著

Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

商业智能时代已经全面到来，分析型人才的岗位数量在就业市场中呈现井喷式增长。无论是从事产品研发的工程师，还是从事产品推广的市场人员、人力资源和财务会计人员，都需要掌握数据分析技术，否则很有可能被人工智能替代。

本书包括 18 章，涉及使用 R 语言做数据分析和数据挖掘的主要分析方法。其中，第 1、2 章为数据分析方法概述，第 3 章为 R 语言编程基础，第 4 章到第 8 章为统计学习方法，第 9 章到第 16 章为数据挖掘方法，第 17 章为特征工程，第 18 章为 R 文本挖掘。每章都根据所涉及的知识点的不同，选取了实用的案例，并为读者准备了相应的练习题。

本书作为 CDA 数据分析师系列丛书中《如虎添翼！数据处理的 SPSS 和 SAS EG 实现（第 2 版）》和《胸有成竹!数据分析的 SPSS 和 SAS EG 进阶（第 2 版）》的姊妹篇，将前两本书的内容进行整合并做了重大拓展，而且秉承了该系列丛书的特点：内容精练、重点突出、示例丰富、语言通俗。可以作为广大从业人员自学商业数据分析的读物，适合大中专院校师生学习和阅读，同时也可以作为高等院校商科、社会科学及相关培训机构的教材。

图书在版编目（CIP）数据

用商业案例学 R 语言数据挖掘 / 经管之家主编；常国珍，曾珂，朱江编著. —北京：电子工业出版社，2017.9
（CDA 数据分析师系列丛书）
ISBN 978-7-121-31958-7

Ⅰ. ①用… Ⅱ. ①经… ②常… ③曾… ④朱… Ⅲ.①程序语言－程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字(2017)第 139693 号

策划编辑：张慧敏
责任编辑：石　倩
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：28.75　字数：736 千字
版　　次：2017 年 9 月第 1 版
印　　次：2017 年 9 月第 1 次印刷
定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

作者简介

常国珍，北京大学会计学博士，中国大数据产业生态联盟专家委员会专家委员。主要从事金融、电信行业客户画像，信用与操作风险识别与防范，客户终生价值预测与价值提升等工作。

曾珂，华中师范大学管理科学工程硕士，现为第一车贷金融产品部产品经理，精通Python与R语言数据挖掘。曾为华为、国家电网等公司提供培训。以金融信用与欺诈风险建模、文本分析、数据可视化等为主要研究方向。

朱江，挪威科技大学工学硕士，现为CDA数据分析研究院课程开发副总监，CDA数据挖掘竞赛评委。精通R与SAS语言数据挖掘，从事电商与互联网数据分析的教学工作。研究方向为电商推荐系统开发、数据可视化、客户特征提取和客户行为模式发现。

CDA数据分析师

CDA（Certified Data Analyst），亦称“CDA数据分析师”，指在互联网、零售、金融、电信、医学、旅游等行业专门从事数据的采集、清洗、处理、分析并能制作业务报告、提供决策的新型数据分析人才。CDA以总结凝练最先进的商业数据分析实践为使命，明晰各类数据分析从业者的知识体系为职责，旨在加强全球范围内正规化、科学化、专业化的大数据及数据分析人才队伍建设，进一步提升数据分析师的职业素养与能力水平，促进数据分析行业的高质量持续快速发展。

“CDA数据分析师认证”是一套专业化、科学化、国际化、系统化的人才考核标准，分为CDA LEVEL Ⅰ、LEVEL Ⅱ、LEVEL Ⅲ，涉及金融、电商、医疗、互联网、电信等行业大数据及数据分析从业者所需要具备的技能，符合当今全球大数据及数据分析技术潮流，为各界企业、机构提供数据分析人才参照标准。经管之家为中国区CDA数据分析师认证考试唯一主办机构，于每年6月与12月底在全国范围举办线下数据分析师考试，通过考试者可获得CDA数据分析师认证证书。

“CDA数据分析师培训”是根据CDA数据分析师认证体系标准而设立的一套专业化、科学化、系统化的学习方案。培训内容不仅包含认证标准中的技能知识要求，还包含企业环境中的真实项目和案例，能满足不同层次的学员需求，使学员能学到真本事技能并能够落地运用，实现商业价值。

品读经典，分享精华
我们期待您的加入

投稿邮箱：
zhanghm@phei.com.cn

交流学习：

序言：数据分析是当代商业的主旋律

CDA 数据分析研究院历经多年研发，最终呈现给大家这一系列教材。“CDA”是注册数据分析师的英文缩写。CDA 行业有当前的发展，主要是时势使然。遥想成立之时，金融海啸正逐步向实体经济蔓延，国际大型跨国企业由于经营业绩下滑，纷纷裁员。例如国际制药企业默沙东全球裁员 5000 余人，但奇特的是其不仅没有裁减亚太研究中心的数据分析人员，反而还在各大高校积极招聘。出现这种怪现象主要是因为数据分析职业是逆经济周期发展的。商业发展前景越悲观、行业竞争越激烈，企业对数据分析人员的需求就越旺盛。这和在经济低迷时，化妆品和电信公司收入反而提高是一个道理。我国大数据的元年为 2013 年，与金融海啸相隔 5 年，在这 5 年里，金融的风险向实体经济逐渐释放。自 2008 年后，国际贸易逐渐走弱，代工类企业的收入明显下降。很多外向型企业逐渐瞄准国内市场，但是这谈何容易。这类企业对国内市场很陌生，市场推介主要是依靠各类展销，一年的生产目标仅靠几个大订单就能确定。企业过去的商业模式基本上就是一个成本中心，只要控制好成本，就算万事大吉。如今一旦进入国内散客市场，创建自主品牌，商业模式便会完全改变。企业要进行客户分析、了解市场结构与客户偏好，并投入研发、宣传、开拓市场等工作中。之前接触过一些转型中的企业，它们一开始都不知道客户在哪里，产品需求分析与趋势预测更是无从谈起。管理学大师德鲁克曾指出，“在未来的社会中，不能正确预测趋势将导致企业 100%的失败”。这从侧面反映了当时企业的尴尬境地。企业在困境中一方面要进一步控制成本，对内通过数据治理实现效率的提高；另一方面，要积极获取外部数据用于市场分析、客户研究，从而指导产品研发和市场策略，这就是大数据相关行业火热起来的根本原因。

大环境利好数据分析，但是企业在实施数据分析项目时却步履维艰，这主要是由于专业人才的匮乏。麦肯锡公司的一份研究预测称，到 2018 年，在“具有深入分析能力的人才”方面，美国可能面临着 14 万到 19 万人的缺口，而“可以利用大数据分析来做出有效决策的经理和分析师”的缺口则会达到 150 万人。谷歌首席科学家范里安直接指出“数据非常之多而且具有战略重要性，但是真正缺少的是从数据中提取价值的能力”。我国是人口大国，却是人力资源弱国。据艾瑞的研究报告，未来与数据分析相关的就业岗位会在 1000 万个左右，而目前国内合格的数据分析师不足 5 万人，建立一个科学有效的数据分析师认证与培训体系迫在眉睫。北京国富如荷网络科技有限公司应时代需求，依托经管之家（原人大经济论坛）十几年来在商业、金融、管理等方向的数据分析教学领域的基奠，联合中华资料采矿协会及数据分析领域专家、学者于 2013 年发起成立“CDA 数据分析师”职业认证，积极推动商业数据分析知识体系建设和认证标准制定等工作的开展。国富如荷在 CDA 数

据分析研究方面，设有 CDA 数据分析师培训中心、CDA 数据分析师考试中心和大数据及数据分析教研部，分别负责知识体系构建、认证题库建设和商业数据分析教学研发工作。经过 4 年的发展，成果喜然，参与培训和认证的人数每年均以 50%以上的速度增长，成功见证了数千名数据分析师的成长。未来，我们将继续提供高水平、多层次的数据分析培训和认证服务，以在行业积累多年的影响力，吸引更好、更多的优秀师资，瞄准行业内重要的数据分析问题和难点，不断突破，建立更加规范的行业培训体系，引领数据分析培训行业向规范化、有效化和前瞻化方向发展，为数据分析的商业运用做出应有的贡献。

常国珍

2017 年 1 月 1 日

前 言

本书有别于其他数据挖掘书籍最大的特点在于参与写作的主要作者均为非理工科背景并具有数据挖掘岗位数年的实际工作经验，且从事 3 年以上的培训工作。这使得本书更贴近实际运用的同时，紧抓初学者的痛点，语言更浅显易懂，操作性更强。当然，这也使得本书在前沿方法的讲解上略显不足。因为一个算法要在商业数据挖掘中得到运用需要大致 3 ~ 5 年的时间。所以本书仅适合数据挖掘入门人员使用。而且本系列教材强调追求浅显易懂，只注重运用中是否够用，不关心算法知识的全面性，因此在算法推导过程中降低了难度，不涉及非关键且不易理解的部分。当读者从事数据挖掘 2~3 年后，本书的知识就不能满足其更高的需求了，需要参考内容更深入的书籍，比如更专业的《统计学习方法》、《机器学习》等。

本书按照数据挖掘工程师规范化学习体系而定，对于一名初学者，应该先掌握必要的编程工具、统计理论基础、数据挖掘算法等内容。进而，数据挖掘需要根据业务问题选择合适的方法，按照标准流程，即数据的获取、储存、整理、清洗、归约等一系列数据处理技术，并最终得出结果，绘制图表并解读数据，这些内容在本书中进行了详细的讲解和操作分析。

本书整体风格是“理论>技术>应用”的一个学习过程，最终目的在于商业业务应用，为欲从事数据挖掘的各界人士提供一个规范化的数据分析师学习体系。

读者对象

本书是一本面向商业数据分析初学者的教材，从具体的商业数据分析案例入手，使读者掌握数据挖掘的目的、理念、思路与分析步骤。本书力图淡化技术，对于方法的介绍也尽量避免涉及过多的数学内容，和高等数学相关的内容只在线形回归和主成分分析这两节中涉及，而且都辅以图形做形象的展现。因此本书的读者只需要具有高中水平的数学基础即可。但是本书强调每种方法的假设、适用条件都与商业数据分析的主题匹配。在教学实践中，我们发现业务经验丰富和有较好商业模式理解的学员，在学习数据挖掘时有更好的效果，主要原因可能是这类学员有较强的思辨能力、分析能力、学习目的性和质量意识，而不是简单地模仿和套用数学公式。

工具介绍

当前，R 和 Python 等开源软件方兴未艾，但是这类软件学习曲线缓慢，使很多初学者的热情在进入数据分析的核心领域之前就消逝殆尽。商业数据分析的真正目的是为了解决业务的分析需求，

构造稳健的数据挖掘模型。数据挖掘产品的质量是通过对分析流程的严格掌控而得以保障的。本书注重实用，直指数据挖掘实施的要点，精选业界使用最广泛的实施方案，为读者节约宝贵的时间。

相对于 Python，R 偏向于统计分析、计量经济学和统计内容。R 不仅在学术研究中拥有广泛的用户基础，而且和 Oracle、SQL Sever 等数据库软件的结合使其不再受内存的限制，从而在商业上有了一定的用武之地。而且 R 和 Hadoop、Spark 等大数据分析平台也可以自由连接。

阅读指南

本书包括 18 章，内容涉及使用 R 做数据挖掘的主要分析方法。其中，第 1、2 章为数据分析方法概述，第 3 章为 R 语言编程基础，第 4 章至第 8 章为统计学习方法，第 9 章至第 16 章为数据挖掘方法，第 17 章为特征工程，第 18 章为 R 文本挖掘。每章都根据涉及的知识点的不同，选取了实用的案例，并为读者准备了相应的思考和练习题。

为方便读者学习，本书提供书中案例的源文件下载，请读者进入 CDA 官网（http://cda.cn/view/22045.html）的相应专栏下载数据和源代码。

本书特点

本书作为 CDA 第一本数据挖掘教材，和其他统计软件图书有很大的不同，文体结构新颖，案例贴近实际，讲解深入透彻。这些特点主要表现在以下几方面。

场景式设置

本书对互联网、电商、电信、银行等商业案例进行精心归纳，提炼出各类数据分析的运用场景，方便读者查找与实际工作相似的问题。

开创式结构

本书案例中的“解决方案”环节是对问题的解决思路的解说，结合“操作方法”环节中的步骤让读者更容易理解。“原理分析”环节则主要解释所使用代码的工作原理或者详细解释思路。“知识扩展”环节是对与案例相关的知识点的补充，既能拓展读者的视野，同时也有利于理解案例本身的解决思路。

启发式描述

本书注重培养读者解决问题的思路，以最朴实的思维方式结合启发式的描述，帮助读者发现、总结和运用规律，从而启发读者快速地找出解决问题的方法。

学习方法

俗话说，“打把势全凭架势，像不像，三分样”。只有熟悉数据挖掘的流程，才能实现从模仿到灵活运用的提升。在产品质量管理方面，对流程的掌控是成功的关键，在数据挖掘过程中，流程同

样是重中之重。数据挖掘是一个先后衔接的过程，一个步骤的失误会带来完全错误的结果。一个数据挖掘的流程大致包括抽样、数据清洗、数据转换、建模和模型评估这几个步骤。如果抽样中的取数逻辑不正确，就有可能使因果关系倒置，因而得到完全相反的结论。如果数据转换方法选择不正确，模型就难以得到预期的结果。而且，数据分析是一个反复试错的过程，每一步都要求有详细的记录和操作说明，否则数据挖掘人员很可能迷失方向。

学习数据挖掘最好的方法就是动手做一遍，本书语言通俗但高度凝炼，很少有公式，以避免读者麻痹大意。本书按照相关商业数据分析主题提供了相应的练习数据，同时提供相关方面的参考资料，供读者学习。

致谢

本书由经营之家主编，CDA 数据分析研究院策划，常国珍、曾珂、朱江负责编写和完成统稿。

丛书从策划到出版，倾注了电子工业出版社张慧敏、石倩、王静、杨嘉媛等多位编辑的心血，特在此表示衷心的感谢！

为保证丛书的质量，使其更贴近读者，我们组织了著名学者和工作在数据挖掘一线的工程师参与了本书的预读工作，他们是李御玺教授、瞿辉工程师。感谢两位预读员的辛勤、耐心与细致，使得本丛书能以更加完善的面目与各位读者见面。

尽管作者们对书中的案例精益求精，但疏漏仍然在所难免，如果您发现书中的错误或认为某个案例有更好的解决方案，敬请登录社区网站向作者反馈，我们将尽快在社区中给出回复，且在本书再次印刷时做出修正。

再次感谢您的支持！

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **下载资源**：本书如提供示例代码及资源文件，均可在 下载资源 处下载。
- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。
- 页面入口：http://www.broadview.com.cn/31958

反侵权盗版声明

目　录