Apress®

Business Analytics Using R
A Practical Approach

R语言商业分析实战

[美]Dr. Umesh R. Hodeghatta　Umesha Nayak | 著
王胜夏　杨莉灵　秦双夏 | 译

清华大学出版社

R 语言商业分析实战

[美] Dr. Umesh R. Hodeghatta，Umesha Nayak 著

王胜夏 杨莉灵 秦双夏 译

清华大学出版社

北 京

内 容 简 介

本书详细阐述了与R语言商业分析相关的基本解决方案，主要包括商业分析简介、R语言概述、R语言数据分析、描述性分析概述、商业分析过程与数据探索、机器学习、线性回归分析以及逻辑回归分析等内容。此外，本书还提供了相应的示例，以帮助读者进一步理解相关方案的实现过程。

本书适合作为高等院校计算机及相关专业的教材和教学参考书，也可作为相关开发人员的自学教材和参考手册。

Business Analytics Using R – A Practical Approach 1st Edition/by Dr. Umesh R. Hodeghatta,Umesha Nayak/
ISBN: 978-1-4842-2513-4

图书在版编目（CIP）数据

R 语言商业分析实战/（美）优曼许·R. 霍奇哈塔（Dr. Umesh R. Hodeghatta），（美）优曼许·纳亚克（Umesha Nayak）著；王胜夏，杨莉灵，秦双夏译. —北京：清华大学出版社，2018
（书名原文：Business Analytics Using R - A Practical Approach）
ISBN 978-7-302-48966-5

Ⅰ. ①R… Ⅱ. ①优… ②优… ③王… ④杨… ⑤秦… Ⅲ. ①程序语言-应用-商业管理 Ⅳ. ①F712-39

中国版本图书馆 CIP 数据核字（2017）第 292751 号

责任编辑：贾小红
封面设计：刘　超
版式设计：魏　远
责任校对：马子杰
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张**：15.25　　**字　　数**：306 千字
版　　次：2018 年 2 月第 1 版　　**印　　次**：2018 年 2 月第 1 次印刷
印　　数：1～3000
定　　价：79.00 元

产品编号：074949-01

译 者 序

近几年来，随着计算机和新一代信息技术的蓬勃发展，商业大数据也呈爆炸性增长。在商业领域，各个行业、企业或组织都遇到了前所未有的全球化、区域化或细分市场多元化的挑战和机遇，他们在激烈竞争中对生存和成长的需求推动了对大数据发展的探索和研究。有效地处理、分析和应用这些大数据解决商业问题，已成为当今各个商业领域的迫切需求，相应地也使商业分析师变得倍受欢迎。

有志于涉足商业分析的读者，将会从本书中获益匪浅。本书最显著的特点是以丰富的案例阐述、解释和演示商业分析涉及的理论和技术，虽然涉及 R 语言、数学和统计学理论，但并不晦涩难懂。本书列举了数据分析在许多应用场景和商业分析中的案例，探讨了商业分析师所需的技能和商业分析项目的过程，并且以 R 语言作为演示本书当中概念的一种统计工具，通过详实的案例讲述了如何编写 R 语言程序处理数据、分析模型的构建以及从中吸取有益的推论。同时，阐述了数据挖掘及其相关技术，也揭示了机器学习的基本概念和生成预测建模，最后还讲述了如何定义商业问题及提出相关的解决方案。

本书将有助于读者了解当前行业、企业或组织中所面临的具体商业问题，以及学习采取哪些相应的措施去分析和解决读者所研究的商业问题。书中还包含了一些重要的商业分析技术的学习案例，如分类、关联、聚类以及回归分析等。读者可以根据商业需求选择不同的方法和技术去解决实际商业问题。如果读者对商业分析感兴趣，这就是一本具有实战指导意义的书籍。

除王胜夏、杨莉灵和秦双夏之外，参与本书翻译的还有吴骅、罗平章、李远明、王学昌、周娟、刘红军、王玲、郑正正、莫鸿强等人，感谢这些同行对本书的辛勤付出。由于水平有限，译文中的不当之处在所难免，恳请各位同行和读者朋友不吝赐教。

译 者

作 者 简 介

Umesh Rao Hodeghatta 博士在机器学习、NLP 和商业分析领域中是一名备受赞誉的专家。他拥有俄克拉荷马州立大学电气工程硕士学位、印度理工学院（IIT）卡哈拉格普尔（分院）的博士学位，专业（方向）为机器学习和 NLP。Hodeghatta 博士在 Wipro Technologies、McAfee、Cisco Systems 和 AT&T Bell 实验室担任技术和高级管理职务。而且，他在国际期刊和会议论文集中发表了许多期刊文章。此外，他也是《信息安全导论指南》（*The InfoSec Handbook: An Introduction to Information Security*）的合著者。Hodeghatta 博士为许多专业组织和监管机构做出了不少贡献，其中包括 IEEE 计算机学会（印度）、美国的信息系统审计和控制协会（ISACA）、（印度）奥迪萨邦政府、印度的国际神经网络学会（INNS），以及商业智能与知识管理专门小组。同时，他也是 IEEE 的资深成员。如需了解 Hodeghatta 博士的更多详细信息，请访问 www.mytechnospeak.com。可以通过 Email：umesh_hr@yahoo.com 与他取得联系。

Umesha Nayak 先生是 MUSA 软件工程总监兼首席顾问，主要负责系统、流程和管理咨询。他拥有 35 年的工作经验，在此期间从事 IT/制造和全球其他组织的咨询工作，时间长达 14 年。此外，Nayak 先生拥有软件系统硕士学位和经济硕士学位。他获得的认证包括印度银行家协会认证专员（CAIIB）、国际信息系统审计师（CISA）和 ISACA 的风险及信息系统控制认证（CRISC）、财务管理研究生（PGDFM）、多项标准的总审计师，以及认证培训师等。他的工作范围广泛，包括银行、软件开发、产品设计与开发、项目管理、计划管理、信息技术审计、信息应用审计、质量保证、培训、产品可靠性、人力资源管理、商业分析和咨询等。在现任职务之前，他曾在印度金奈的北极星软件实验室担任副总裁兼公司执行理事会成员。1981 年，他开始涉足计算机行业，从 ICL 大型机开始，然后是迷你计算机和个人电脑。Nayak 先生也是印度银行业信息系统审计的创始成员之一。依靠成功的 ISO 9001、ISO 27001、CMMI 和其他认证，以及流程/产品改进和商业分析，Nayak 先生给许多组织机构提供了卓有成效的指导。他还和 Hodeghatta 博士共同撰写了《信息安全导论指南》。他的联系方式为 Email：aum136@rediffmail.com。

技 术 审 校

乔乔·穆拉伊尔（Jojo Moolayil）是一名数据科学家，著有《智慧决策——物联网和决策科学的交叉点》（*Smarter Decisions—The Intersection of Internet of Things and Decision Science*）。他在数据科学、决策科学和物联网领域拥有 4 年以上的工业经验，并与工业领先企业进行跨多个垂直方向的合作，合作的项目都是一些具有重大影响的关键项目。目前，他和工业物联网数据科学的先锋和领先者通用电气合作，现居住在班加罗尔——印度的硅谷。

穆拉伊尔出生和成长在印度的浦那，毕业于浦那大学，主修信息技术工程学。之后在世界最大的纯游戏分析提供商 Mu Sigma 开始他的职业生涯，和众多财富 50 强客户的领导一起工作。作为最早进入物联网分析行业的冒险者之一，他现今专注于解决工业物联网应用案例的决策科学问题。而他在 GE 工作的职责之一，是为工业物联网开发数据科学和决策科学相关的产品和平台。此外，穆拉伊尔还是一名活跃的数据科学导师，在 www.jojomoolayil.com/web/blog/上发布并维护博客。

目　录

第 1 章　商业分析简介

知识是当今世界的基石。人们早期通过观察积累知识。后来人们不仅在观察中积累知识，也通过实际行为去确认，然后又通过进一步的实验去证实。人们将积累的知识应用到实践领域，以此类推也应用到其他领域。现在，人们借助各种计算机应用程序、网站和更多的工具，通过分析数据或深入研究数据积累和应用知识。而计算机的出现补充了统计学、数学和编程这方面的知识，尤其是云的巨大存储和扩展计算能力，从大量数据中快速提炼出知识，也用于进一步的预防或生产。本章将介绍商业分析的应用场景和应用方式的基础知识。

想象下列情景：

- 当您到达瑞士的一家酒店，酒店以您最喜欢的饮料和菜肴欢迎您。此刻您感到多么高兴！
- 当您去喜欢的地方旅行，您最喜爱的酒店给您提供了非常优惠的住宿价格。
- 您接收到预警提示，您极其可能会患上糖尿病，您也认为这个预警事出有因，所以您采取正确的做法，避免患上糖尿病。
- 当您计划旅行时，收到了目的地可能会发生骚乱的预警信息，您根据预警取消了行程。随后您从新闻报道中得知那个地方确实发生了骚乱！
- 您正在考虑是否和另外一人结为终身伴侣，但根据两人的性格特征，数据预测您和此人并非天作之合。您因此避免了将来可能发生的离婚！
- 当您走进杂货店，发现商家已把您每月定期购买的商品选置妥当。而您唯一需要做的是决定全部购买这些商品或者只购买其中一部分。此时此刻您是多么开心！
- 在您假期旅行前，您喜欢的航空公司为您提前预订了机票，而且机票价格远比其他市售机票优惠得多。
- 您计划旅行时，收到了目的地可能会发生飓风的预警信息，您因而推迟了行程。后来您得知当地飓风肆虐，而您却幸免于此。

通过分析您和您的行为数据，我们可以想象出许多类似的场景。在您进行以下各种

U.R.霍德噶塔和U.纳亚克，《R语言商业分析实战》，DOI 10.1007/978-1-4842-2514-1_1

活动时，我们可以采集到关于您的行为数据：如 Google 搜索、访问各种网站、对社交媒体网站的评论、使用不同计算机应用的行为以及更多活动等。在这些情景中，我们主要从您个人的角度去进行数据分析应用。

现在，让我们从商业角度来观察一些应用场景。想象以下情形：

- 在酒店业务中，您既可以提供潜在客户具有竞争力的价格，同时也能保证您有利可赚。而且，您还可以通过提供附加优惠，确保酒店全部住满，这些附加优惠包括给其他从事当地旅行和观光旅游的中介提供一些折扣。
- 在出租车业务中，您根据乘客早期的历史行程和乘客对出租车类型和司机的偏好，再三地吸引相同的乘客。
- 在快餐业中，您在生意清淡时提供优惠价格吸引顾客。利用这些优惠价格，确保生意清淡时店内也座无虚席。
- 您在某组织的人力资源（HR）部门任职，却深陷人才高流失率的困境中。但根据表现良好、忠诚且愿意效力于组织的人才性格特征，您现在已经很清楚应该招聘什么类型的人才了。
- 您从事航空公司工作时，根据收集的发动机系统数据，发现了未来三个月内可能发生的潜在发动机故障的预警信息，您随即主动地采取必要的改进措施。
- 您从事设计、制造或销售医疗设备的业务时，通过分析设备日志中捕获到的错误或警告，在设备发生故障前能够清楚地了解故障发生的可能性。

通过分析企业和其他人从不同来源收集到的数据，可以实现所有这些场景。这种可能会出现的场景也不胜枚举。商业领域的数据分析应用称为商业分析。

您可能也观察到以下情况：

- 您过去几天在 Google 一直搜索一些可以去冒险的地方，也尝试搜索各种可用的旅行箱包。后来，您在访问其他网站或进行其他搜索时，突然发现页面在向您展示特定广告，广告的内容和您之前搜索的信息相关，而且提供了折扣价格。
- 您一直在 Amazon（或任何其他网站）上寻购一些特定商品。但在浏览其他网站时，您忽然发现和您曾搜索商品相关的广告，或者发现邮箱收到了相关的定制邮件，也都给您提供了折扣价格以及您可能感兴趣的其他商品。
- 您也看到过 Amazon 给您的推荐清单，这些清单都是根据您早期的搜索项目、愿望清单或以往订单进行推荐的。很多时候，您可能也会依据 Amazon 的现有数据去观察他们提供的折扣或促销产品。

由于企业使用特定的数据分析，如今上述可能发生的场景都一一变成了现实。我们随后将向您深入介绍商业分析，将您引入商业分析的领域，这的确令人感到兴奋。

1.1　本书目的

许多专业人士对学习商业分析越来越感兴趣，但他们并非全都具有丰富的统计学或数学背景。本书正是为这些人编写的，也是他们开始商业分析学习的最佳选择。首先，本书通过上机实验向初学者讲解 R 语言，接着介绍预测模型和大数据，这部分是商业分析的关键部分。所以，本书是一本使用 R 语言进行商业分析的入门指导书籍。

本书拥有以下优点：

- 本书涵盖了 R 语言编程和许多现实案例分析。
- 本书将理论和实践完美结合。使用商业场景或在需要时使用案例研究来解释概念。
- 本书由相关行业的专业人士撰写，这些专业人士目前正在为付费客户分析他们的现实问题。

本书涵盖了以下内容：

- 本书利用从政府或其他来源收集、整理、购买或免费提供的数据，和您分享对这些数据应用的独到见解。借助计算机编程、统计学和数学知识以及相关领域的专业知识，不仅可以让您对数据应用形成独到的见解，也可以让您理解数据，获得一定的预测能力。
- 有效使用和商业分析有关的各种技术。
- 如何有效使用 R 语言编程平台进行商业分析。
- 提供实践案例和示例，将本书中学习到的知识学以致用。
- 商业分析的注意事项。

本书不讲述以下内容：

- 对与分析有关的各种术语进行释义，因为过多释义可能会令人混淆。
- 详细阐述任何统计学或数学方法或超过一定限度的特定算法背后的基本原理。
- 提供在商业分析领域中使用的全部技术或算法的数据库（但本书确实运用了不少数据库）。

1.2　容易混淆的术语

本书讨论时会使用到许多术语，例如，数据分析、商业分析、数据科学和大数据分