HZ BOOKS
华章IT

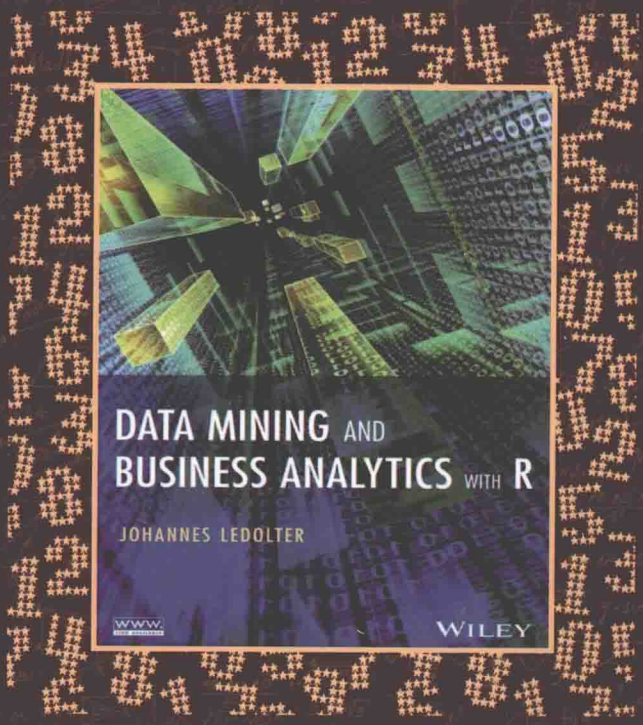
WILEY

数据科学与工程技术丛书

# 数据挖掘与商务分析
## R语言

[美] 约翰尼斯·莱道尔特（Johannes Ledolter）著

宋涛 王星 曹方 译




DATA MINING AND
BUSINESS ANALYTICS WITH R

机械工业出版社
China Machine Press

科学与工程技术丛书

# DATA MINING AND BUSINESS ANALYTICS WITH R

# 数据挖掘与商务分析

## R语言



[美] 约翰尼斯·莱道尔特（Johannes Ledolter）著
宋涛 王星 曹方 译

机械工业出版社
China Machine Press

**图书在版编目（CIP）数据**

数据挖掘与商务分析：R 语言 /（美）约翰尼斯·莱道尔特（Johannes Ledolter）著；宋涛，王星，曹方译 . —北京：机械工业出版社，2016.10
（数据科学与工程技术丛书）
书名原文：Data Mining and Business Analytics with R

ISBN 978-7-111-54940-6

I. 数… II. ① 约… ② 宋… ③ 王… ④ 曹… III. ① 数据采集 ② 程序语言－程序设计－应用－商务－经济分析 IV. ① TP274 ② F7-39

中国版本图书馆 CIP 数据核字（2016）第 231157 号

**本书版权登记号：图字：01-2013-7578**

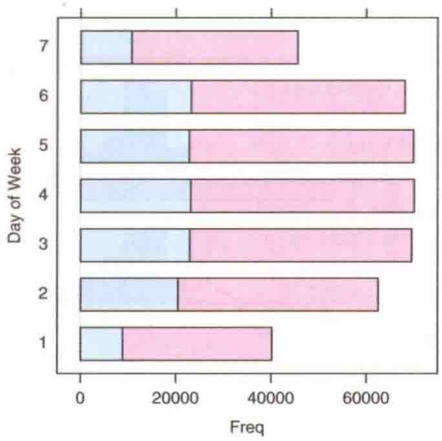


从海量的数据中收集、分析、提取有价值的信息需要功能强大的分析工具，本书结合 R 软件详细介绍了数据挖掘和数据分析的实用方法，主要内容包括处理信息和获取数据、标准线性回归、局部多项式回归、统计建模中简约的重要性、Logistic 回归、贝叶斯分析、多项式 Logistic 回归、决策树、聚类、购物篮分析、降维和网络数据等。书后配有练习并且书中所有例子涉及的数据集和 R 代码可以从本书配套网站获取。

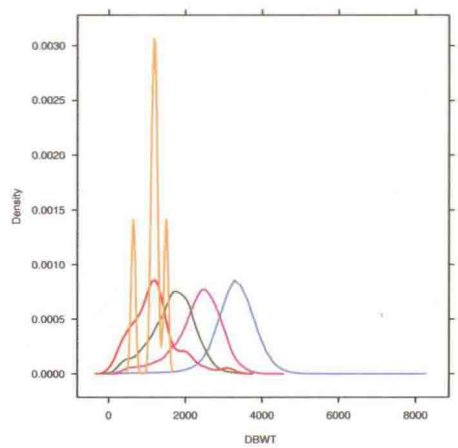
本书适用于数据分析相关专业学生和教师以及 R 语言使用者。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：刘诗灏　　责任校对：董纪丽
印　　刷：中国电影出版社印刷厂　　版　　次：2016 年 10 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：17.75（含 0.5 印张彩插）
书　　号：ISBN 978-7-111-54940-6　　定　　价：69.00 元

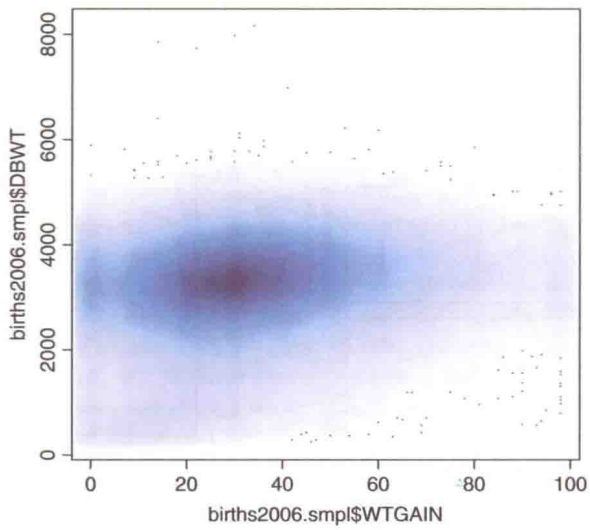
凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com



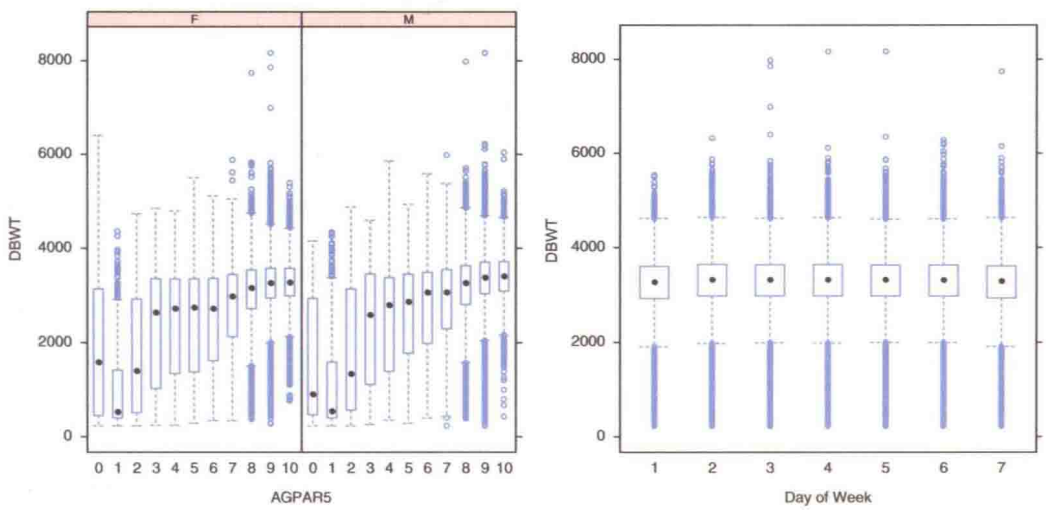

彩图 1



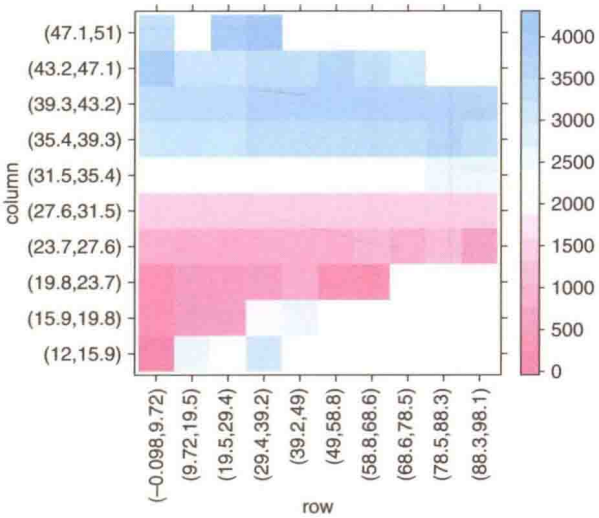

彩图 2



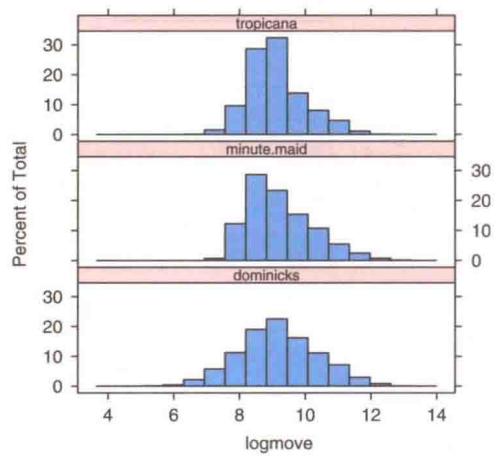

彩图 3



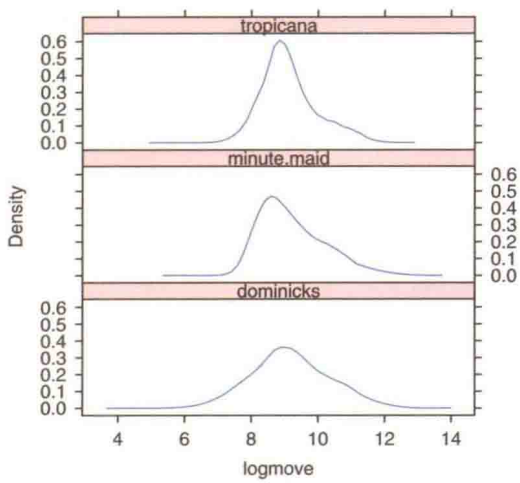

彩图 4

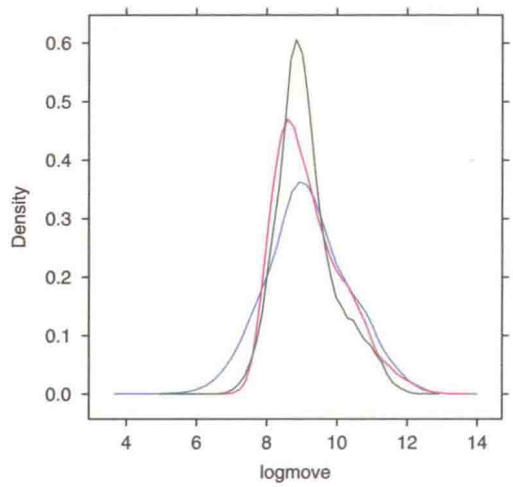



彩图 5

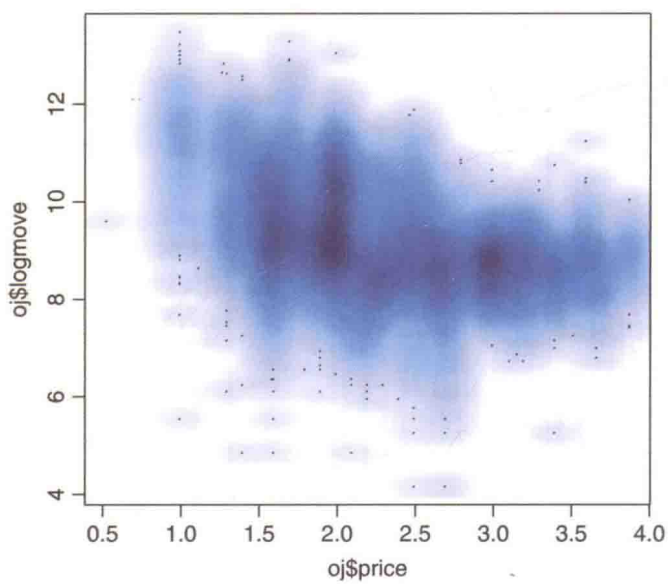



彩图 6

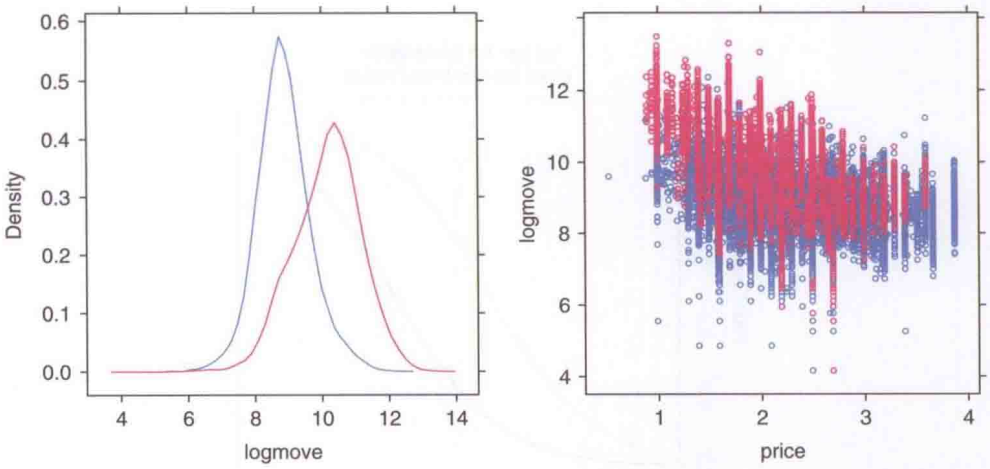


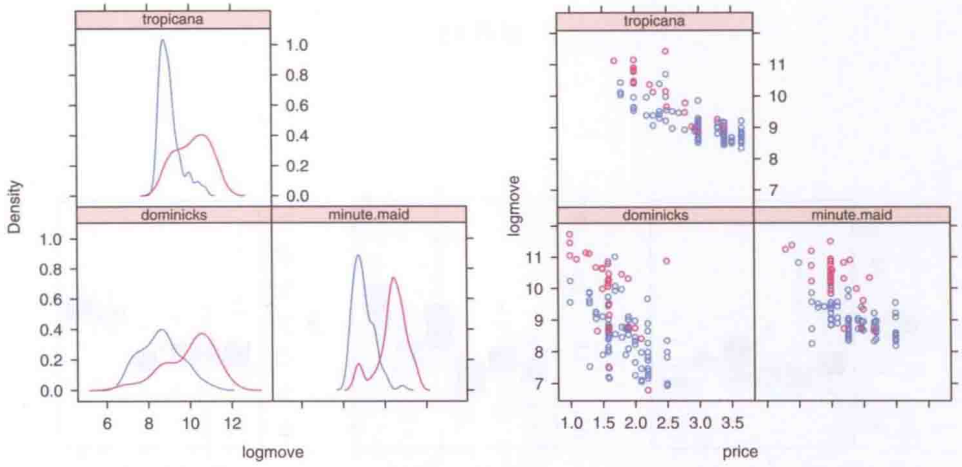

彩图 7



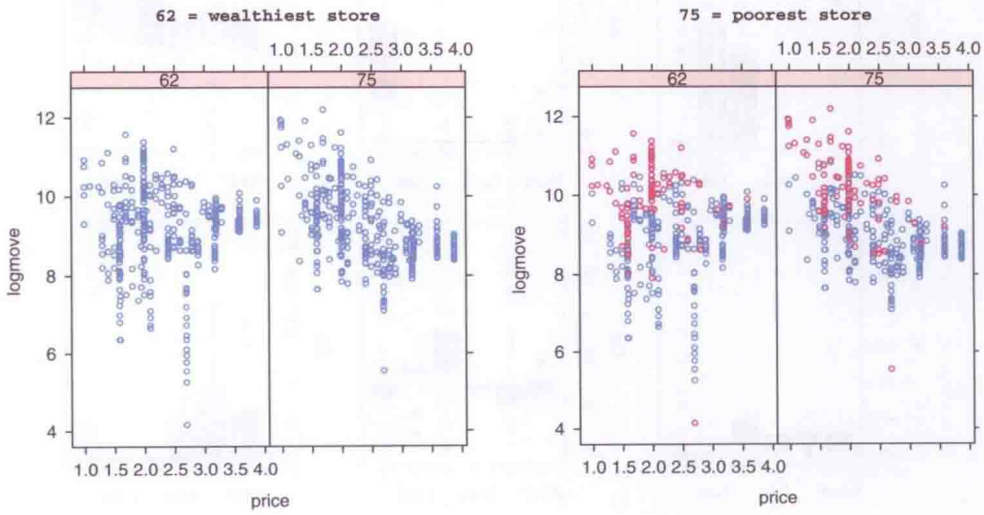

彩图 8



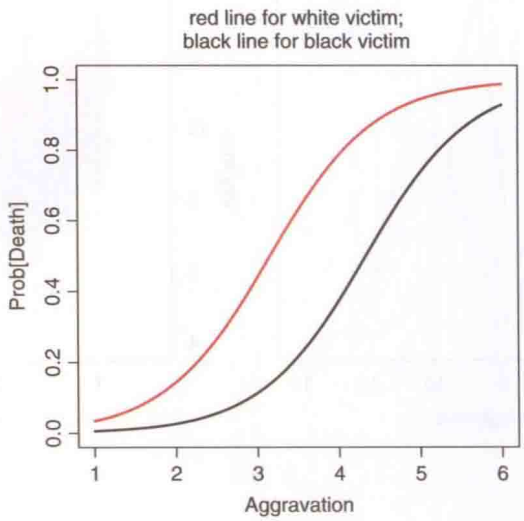

彩图 9



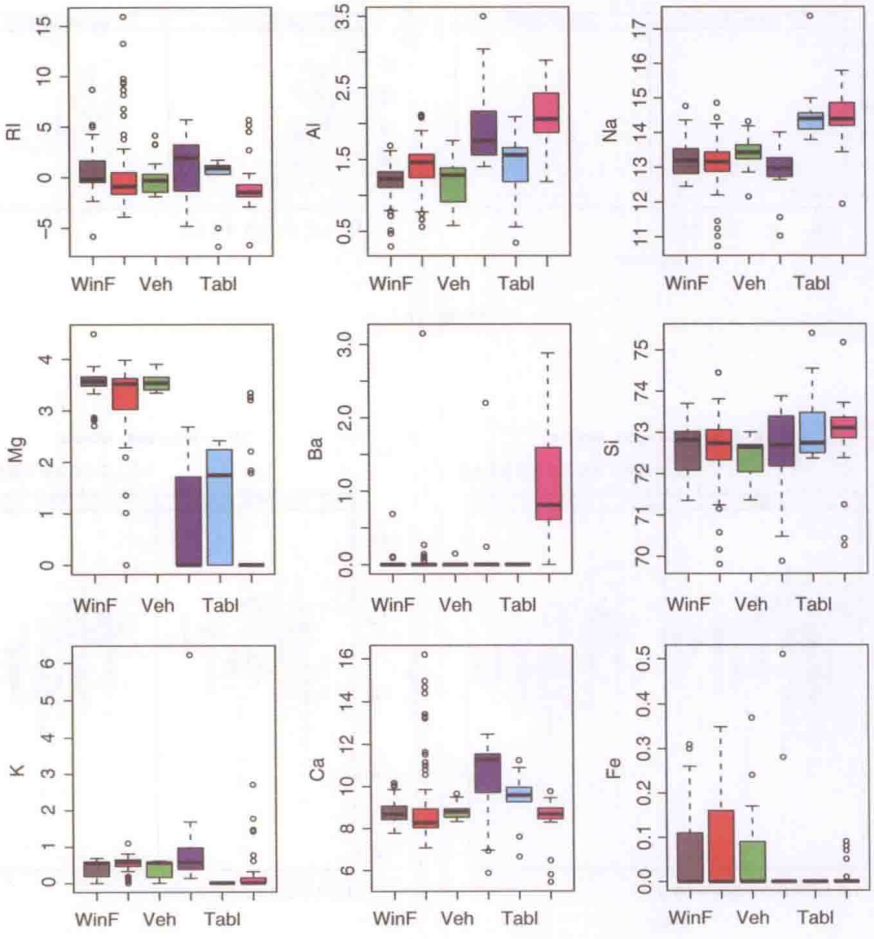

彩图 10



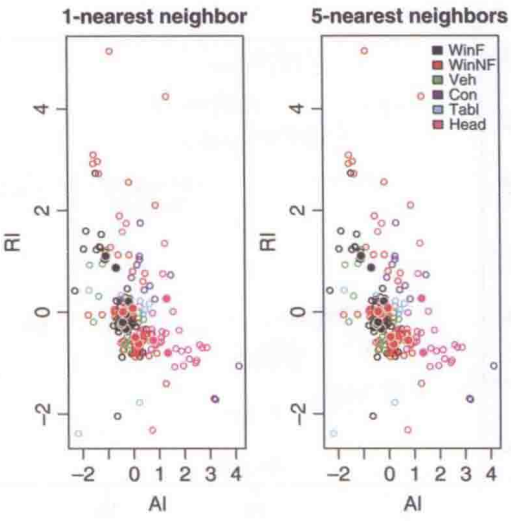

彩图 11


red line for white victim;
black line for black victim
Prob[Death]
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Aggravation

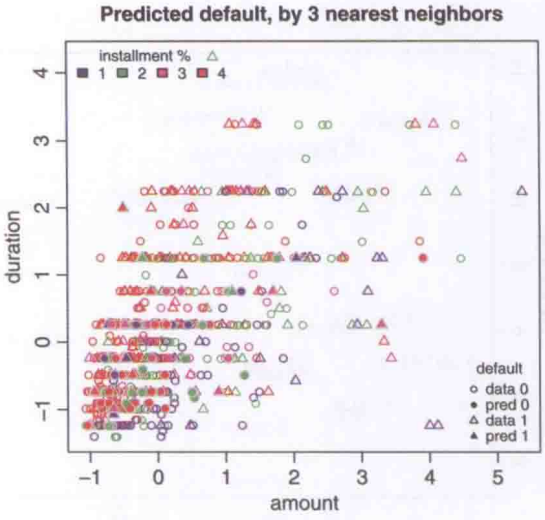
彩图 12


RI
Al
Na
Mg
Ba
Si
K
Ca
Fe
WinF
Veh
Tabl

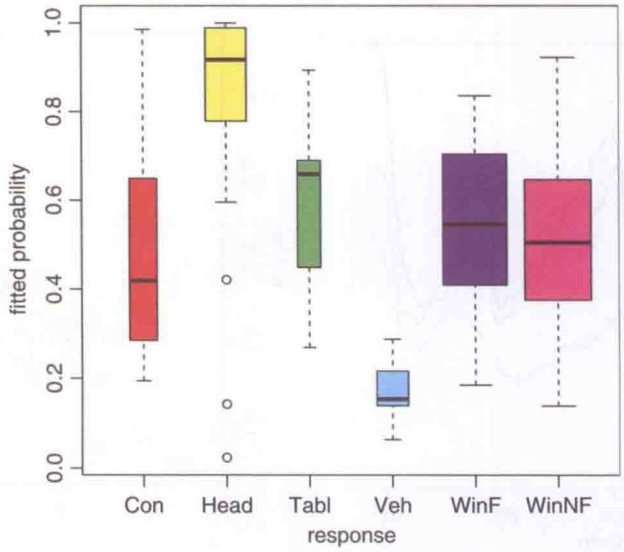
彩图 13



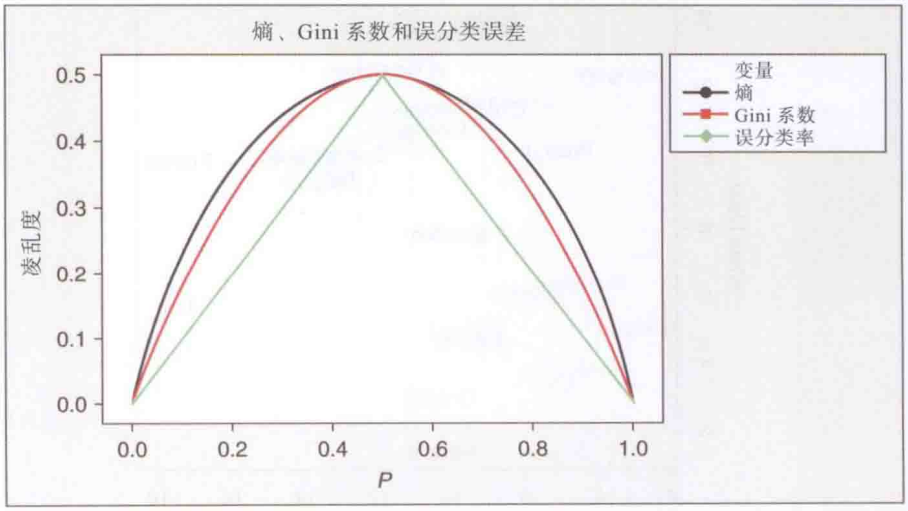

彩图 14



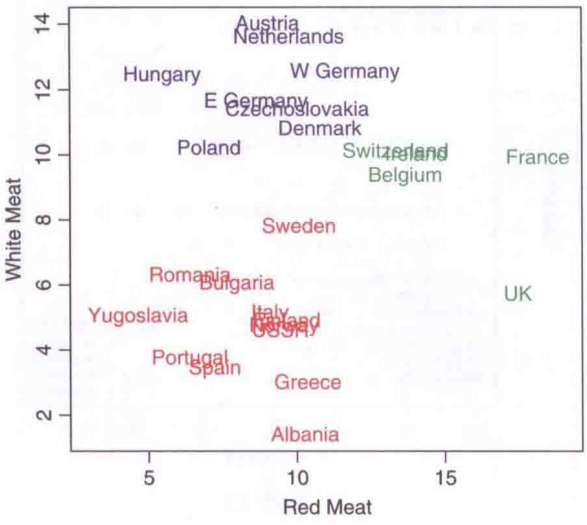

彩图 15



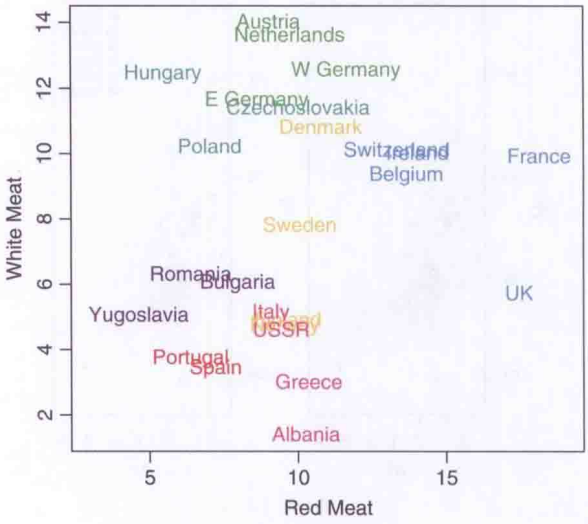

彩图 16



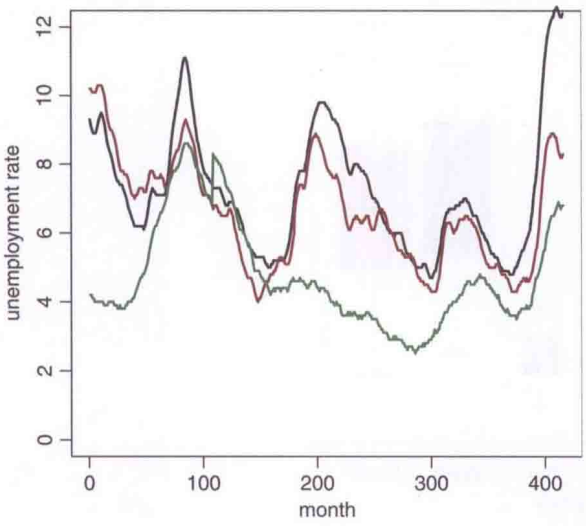

彩图 17



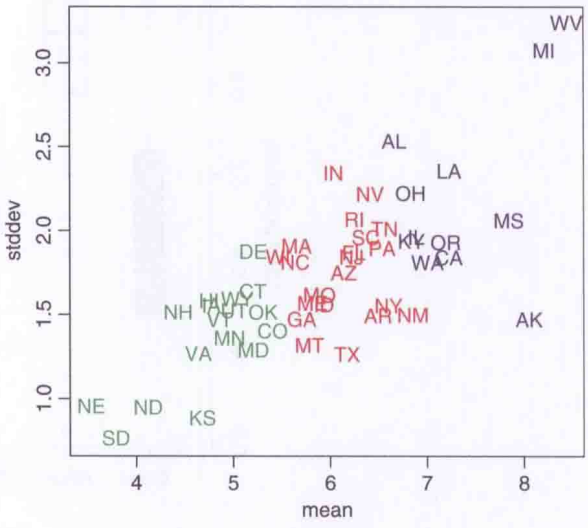

彩图 18



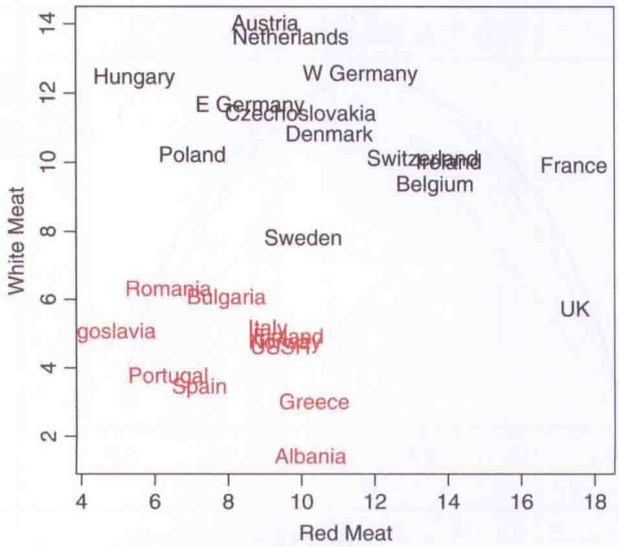

彩图 19



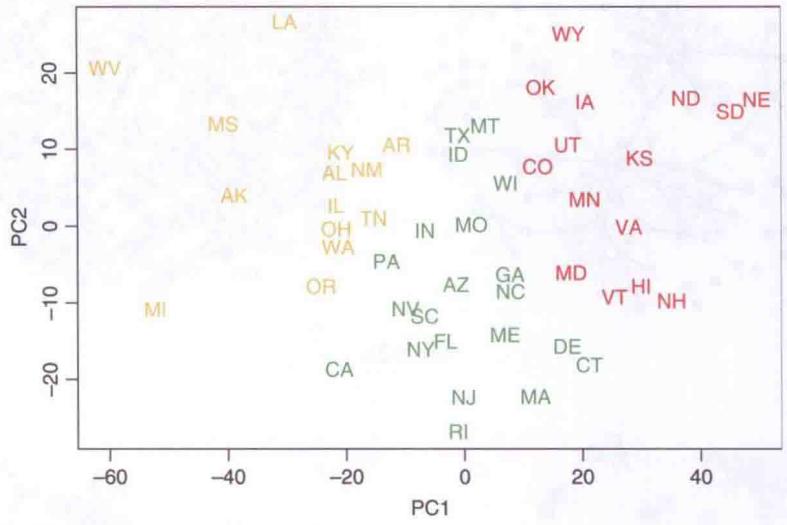

彩图 20



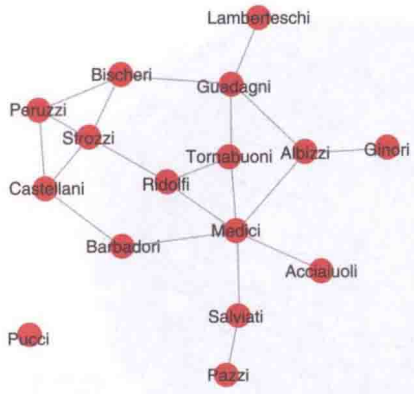

彩图 21



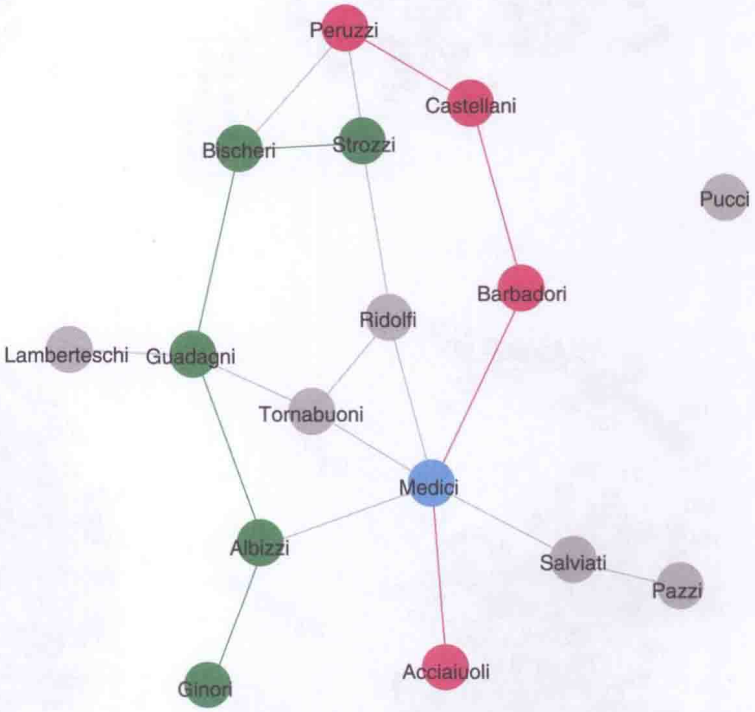

彩图 22



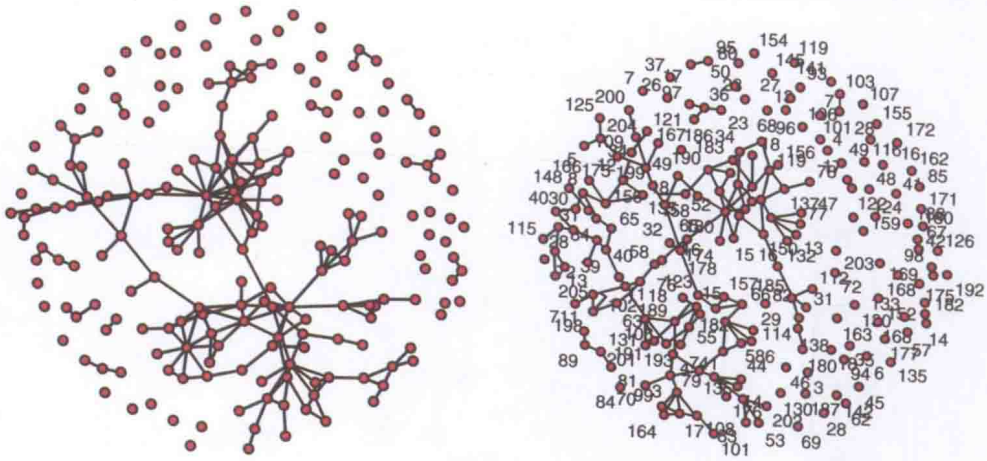

彩图 23



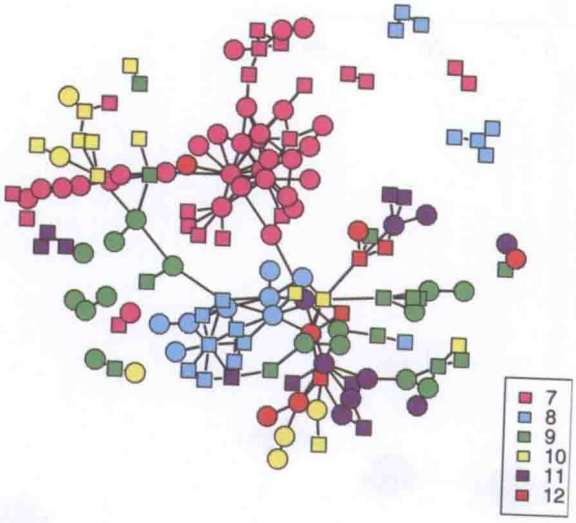

彩图 24



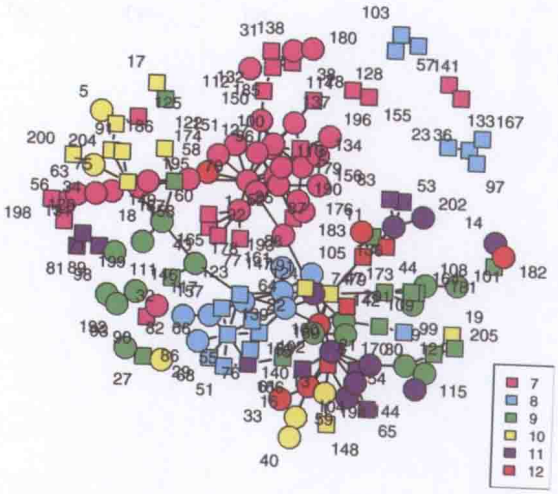

彩图 25



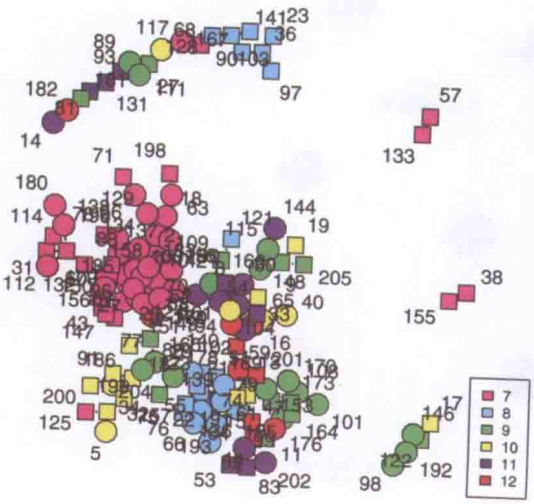
彩图 26



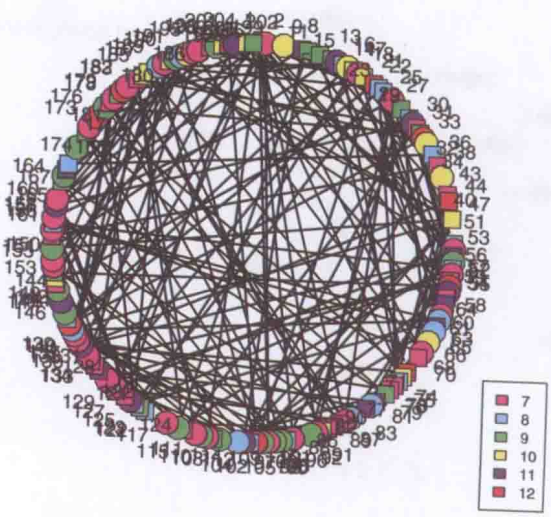

彩图 27




彩图 28




彩图 29




彩图 30

# 译 者 序

本书英文版自出版后就在 Amazon 上得到了极高的评价，曾经是 Amazon 网站上最畅销的数据挖掘类书籍之一。

本书的作者 Johannes Ledolter 是世界顶尖商学院——美国艾奥瓦大学 Tippie 商学院管理科学系的一位数据挖掘专家，同时也是一位 R 资深开发者。本书包括多达 19 个数据挖掘的翔实案例，内容十分丰富，涉及医疗、慈善、汽车、二手市场等行业领域。书中案例从数据量、分析目标、数据类型等方面提出了各种具有挑战性的问题，并给出了克服这些挑战的方法和技巧。本书专注于数据挖掘的建模，以实际问题、解决方案以及探讨解决方案为主线组织内容。读者需要具备一定的数据挖掘基础知识，同时对 R 有一定的了解。但本书也对 R 计算进行了详尽完整的说明，对于零基础的读者来说，还可以通过直接复制书中提供的 R 程序来学习相应的数据挖掘算法。本书定位于面向定量方法的 MBA 学生，同时也适用于大数据分析的本科生及研究生，适合作为数据挖掘的教材或学习指南。

本书的翻译工作由宋涛、王星和曹方共同完成。在本书的翻译过程中，原作者 Johannes 博士多次就译者提出的问题进行了耐心而细致的解答。这里对他的帮助表示由衷的谢意。由于水平所限，书中可能会有翻译不当之处，希望读者多加指正。

必须说明的是，本项工作是集体努力的结果。其中，王星老师在翻译和统稿过程付出了大量心血，她的坚持使我打消了放弃此项目的想法。此外，余阿炎、曹家铭、温丽、丁虹元、俞良、金璐等人也参与了本书的翻译。感谢王宝东、宋辰玉、宋燕、仓猛、刘宇等完成了清样的校对和通读。还有许多其他同学和同事在不同阶段参与了本项工作，在此不再一一列出。

为进一步探讨、解析和扩展本书中的案例，译者团队将在“数据科学家”公众号中免费为各位读者奉献更多更翔实的 R 案例内容。可扫描以下二维码，关注“数据科学家”微信公众号，获得更多有关数据科学和 R 应用的最新知识。

宋涛

2016 年 9 月

# 前　言

这是一本有关数据挖掘和商务分析的实用方法的图书，适用于迫切需要使用这些方法来了解运营状况并解决经营问题的读者。写作本书的目的是对获得公众口碑的数据挖掘工具进行全面讨论，而不仅仅局限于传统的黑箱式描述，展现这些方法的工作机理。

数据挖掘需要一套功能强大、计算精准、兼容良好的计算工具，在这方面微软的Excel难以胜任。尽管我们也多次获得许多供应商专门提供的卓越的数据挖掘商务软件，但通常来说这些软件价格昂贵。书中我们所使用的R统计软件功能强大而且免费。不过要想正常使用R需要一些学习代价，它需要用户写指令，而大多数电子表格用户对程序指令的编写并不熟悉，这也是我在书中和与本书相关的网页上提供R示例代码的原因。这些示例代码应该可以顺利地迁移到当下通用的、强大的计算机环境中，并有助于最小化R的学习成本。

本书采用了将软件与数据挖掘的统计基础相融合的写作风格，同时也推广了工具的应用。虽然市面上不乏深入阐述这些方法的教材，也不缺乏对R计算的详尽完整的说明手册。但是本书力图权衡理论与实践，定位于对定量方法感兴趣的MBA学生的认知层次。本书适用于MBA的数据挖掘课程，以及高年级本科生和研究生的分析与解释大数据集的课程。从事商学、社会学、自然科学、医学以及工科的学生都可以从本书受益。本书所涉大部分主题可以安排在一个学期的课程中，但是包括的主题并不适用于每一个读者。可能有些读者会认为其中一些主题内容太深或者太浅。建议主讲老师略去或适当扩展某些主题。从这个角度来看，本书可以适用于很多不同的读者。

数据挖掘的应用常常需要花大力气收集相关信息。在这种情况下，数据的准备工作比最终建立模型需要花费更多的时间。在另外一些应用中，数据收集的工作量并非大问题，工作的重点是大容量信息的存取（即数据仓库）。尽管如何获取、存储、合并和整理信息在数据分析全过程来说必不可少，但书中对这些技术细节并未做深入探讨，本书重点介绍数据挖掘的建模。

本书所述全部例子的数据集和R代码都可以在配套网页（http://www.biz.uiowa.edu/faculty/jledolter/DataMining）上找到。也可以通过在booksupport.wiley.com上输入ISBN 9781118447147获取本书的附加材料。读者可以将书中的代码复制粘贴到自己的R会话中，从而得到分析结果。也可以在软件中修改或添加一些代码来做数据实验，以及用我们给的R模板程序对自己的数据集进行分析。附录给出了练习和几个大的练习数据集。练习有助于老师布置课后作业，也为读者提供了一个实践书中所讨论技巧的机会。如何

使用这些数据集的相关说明请参见附录 A。

这是本书第 1 版，尽管在表述和例证数据集的分析上我们很小心谨慎，但不得不承认其中有很多地方还值得推敲。如果在阅读本书的过程中有任何反馈，我们将不胜感激，期待你将你的建议通过 johannes-ledolter@ uiowa. edu 邮箱写信给我。相关的勘误和评论我将在本书的网页上随时更新。

# 致　　谢

2011 年我访问芝加哥大学布斯商学院时，忽然为一篇 MBA 方面有关数据挖掘的文章中的素材产生了兴趣。芝加哥大学著名教授 Matt Taddy 的数据挖掘（BUS41201）课件为本书的撰写提供了灵感，在表述上我同样受到 Taddy 教授课件中的案例和 R 模板的影响。第 19 章中关于文本数据的分析也大量引用了他近期的研究成果，由衷感谢 Taddy 教授对本书的贡献。

著书是一项耗时的工作。如果没有妻子 Lea Vandervelde 的持续支持和鼓励，无法想象我的这项工作可以画上句号。她是艾奥瓦大学从事密苏里州奴隶自由史研究的教授，同时她的亲身体验告诉我，从文本数据的挖掘中构建数据集是一项多么重要和艰难的工作。

# 目　录