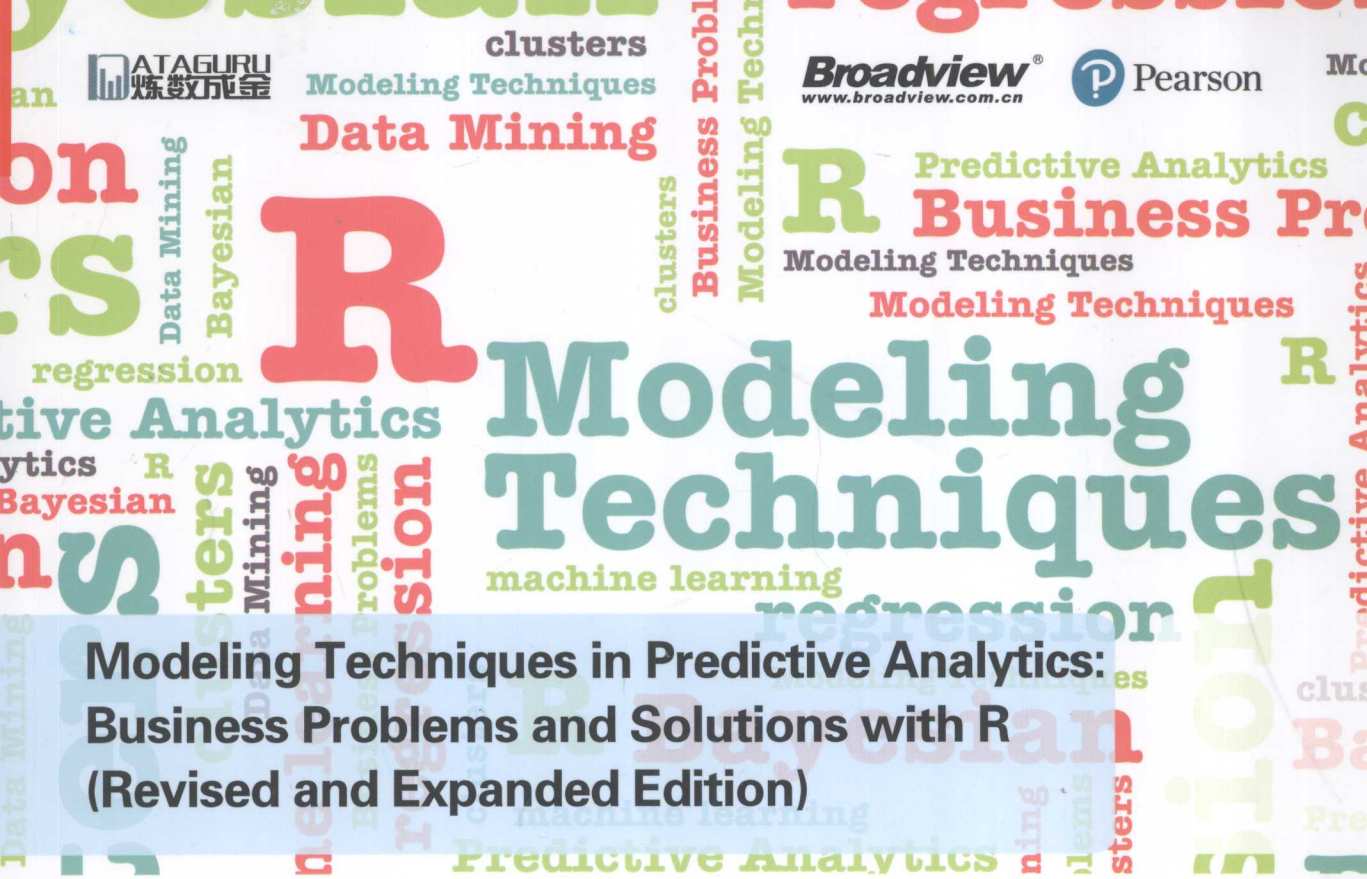

预测分析中的建模技术

商务问题与R语言解决方案

【美】Thomas W.Miller 著 美国西北大学预测分析项目主任教授
【美】陈宇红 译

- 借助可视化数据与易于学习的R语言代码揭示商务问题的解决方案
- 将战略与管理、方法与模型、信息技术与代码三者完美结合
- 通过对历史数据的分析创建模型，为预测分析打下可靠基础

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

Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R (Revised and Expanded Edition)

预测分析中的建模技术

商务问题与R语言解决方案

【美】Thomas W.Miller 著
【美】陈宇红 译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

这是一本商务智能方面的著作，旨在帮助读者解决真实工作中的商务问题，发现问题、定义数据、创建和优化模型，编写高效的代码，对结果进行分析，等等。

本书着眼于真实的案例和真实的数据。每章通过对一个实际问题的描述和讨论引出特定的预测分析模型，分析的结果通过可视化图表进行展示，章节末尾还提供了 R 语言编写的应用程序。通过对建模技术和编程工具的实际演示，把抽象化的概念转化为具体的例子，让这些可以成功运行的案例程序更易于理解。

附录比较系统地介绍了数据分析常用的统计学方法和测量的方法，以及为商务分析在 R 语言环境下特别扩展开发的程序代码。

本书不但适合计算机、统计等相关专业选作教材，还适合进行公司决策分析、大数据分析等的相关人员参考阅读。

图书在版编目（CIP）数据

预测分析中的建模技术：商务问题与 R 语言解决方案/（美）托马斯·W.米勒（Thomas W. Miller）著；陈宇红译. —北京：电子工业出版社，2016.7

书名原文：Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R，Revised and Expanded Edition

ISBN 978-7-121-29207-1

Ⅰ. ①预… Ⅱ. ①托… ②陈… Ⅲ. ①程序语言—程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字（2016）第 152238 号

策划编辑：张月萍
责任编辑：徐津平
特约编辑：顾慧芳
印　　刷：北京京科印刷有限公司
装　　订：北京京科印刷有限公司
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 × 1092　1/16　　印张：18.75　　字数：433 千字
版　　次：2016 年 7 月第 1 版
印　　次：2016 年 7 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。

译者序

大数据与商务智能

大数据（Big Data）是 2012 年开始炒作起来的一个新“词汇”。记得当时我有个在硅谷从事 IT 投资工作的朋友到纽约参加投资商会，她说，上一年度的会议，所有的话题都是社交网络（彼时 Facebook 如日中天），现在呢，人人说的都是大数据。

任何行业大致都差不多，隔上一段时间，总会有个新东西出来炒作一下，无论是 911 后的容灾系统、dot com，还是社交网络、大数据，到今天的“云”。这也是社会进步的一种方式吧。

在大数据这个概念出来之后，我曾经不止一次被问到大数据的问题。比方之前有一次被问到大数据的问题时，一开始我完全不知道对方问的是什么，几个回合下来，才了解到对方的问题完全是商务智能 Business Intelligence 方面的；还有一次和一些从业 IT 技术十几二十年的同行聊天，有人说，大数据把人忽悠得云里雾里，弄不清楚到底是什么东西。于是，许多人就很简单地把所有数据相关的东西，统统都说成是大数据，因此，商务智能也是大数据。

我们知道，传统的数据，多是指结构化的数据，如早期的 dBase、FoxPro，到现在普遍使用的关系型数据库 SQL Server、Oracle 或者 DB2，存储的都是结构化的数据。这些数据都可以用两维的行和列的表格形式表现出来。随着互联网技术的飞速发展，出现了很多非结构化的数据，比如音像数码文件、博客文章、网站搜索索引、社交网络的留言，对这些数据的收集和使用，是我所认知的“大数据”技术。人们对“大数据”有各种各样的定义，其中有一个定义是“大数据是不能用传统的数据库工具管理的所有数据—— big data is any data that can't be managed using conventional database tools”。我颇以为然。在当前的自然数据中，结构化的数据还不足两成，绝大部分的数据，都是非结构化的数据。

商务智能（Business Intelligence，BI）和商务分析（Business Analytics）却不是一个新兴的概念。根据相关资料的记述，商务智能这个词汇早在 1988 年就出现了，到了 20 世

纪 90 年代，关系型数据开始被广泛使用后，SQL 查询语言报表成为了常规，这便是最早的商务智能实例。商务智能在过去的这些年里发展迅速，如我们后来熟知的数据仓库（data warehouse）、数据集市（data mart）、建造数据仓库的抽取-转换-加载（ETL）技术、在线分析处理（OLTP）、数据可视化（data visualization）、信息中心化（dashboard）等，都可看作是商务智能的组成部分。

收集到了数据后，需要进行商务分析，回答商务问题，进行数据驱动的商务决策。按照商务问题的种类，数据分析划分为三种类型：第一种是描述分析（descriptive analytics），从历史数据中，总结过去的商务行为都发生了什么事情，是什么时候、什么原因，怎样发生的；第二种是预测分析（predictive analytics），是通过数据分析预测将来的商务行为中可能会发生什么样的事情；最后一种是规范分析（prescriptive analytics），是基于描述分析和预测分析的结果来推荐的未来的商务行为。有些数据科学家还提出，在描述分析和预测分析中间，应该加入一种新的分析类型——诊断分析（diagnostic analytics），通过对历史数据的分析创建模型，为预测分析打下基础。

对原始数据进行上述各种分析的过程，就是我们熟知的数据挖掘（data mining）。习惯上，我们把数据挖掘的过程分为四大类：分类（classes），类聚（clusters），关联（associations），序列模式 （sequential patterns）。在本书中，与这四种过程相关的技术，会使用经典统计（classical statistics）、贝叶斯方法（Bayesian Statistics）、回归（regression）、分类（classification）、机器学习（machine leaning），等等。

那么大数据和商务智能之间是否有关联呢？美国的一些数据科学家是这样说的，商务智能是帮你找到你想要知道的问题的答案，大数据是帮你发现那些你不知道要问的问题。这个答案也是蛮可爱的。商务智能分析的数据是结构化数据，大数据技术则需要分析所有的结构化，连同非结构化的数据。商务智能和大数据对数据的存储方式和对数据的分析手段的要求是不同的。但无论大数据也好，商务智能也好，数据存在的目的就是让我们通过分析，得到尽可能好的分析且结果为相关的商务服务。从这点上看，大数据和商务智能都有一个共同的目标，这大概就是大数据和商务智能常常被合二为一的原因吧。现在，有诸多数据专家致力于开发商务智能应用于大数据的数据分析技术，希望这一技术能早日成熟。

有人问过我，如果没有任何统计学基础，数学的根底也很有限，可以学习商务分析吗？诚如本书的作者米勒教授在前言中所说，在这本书里可以了解到，数据分析可以在哪些不同商务领域中解决什么样的问题。认识到哪些问题可以从现有的数据中找到答案，是利用商务智能的起点。如果你从事数据分析工作，或者是程序员，则可以通过本书的案例，认知到如何一步一步地分析问题、解决问题、找出问题的答案所在。

本书的所有案例都是在 R 语言环境下实现的。R 语言最初是为生物统计开发的一个开源软件。记得多年之前，我在纽约大学上生物统计的课程时，第一次接触到 R 语言。当时

我们的教授是这样描述 R 语言的：R 环境不依赖计算机操作系统，你可以在 UNIX、Linux、Windows 或者 Macintosh 甚至 OS X 系统下使用；R 语言很容易学习，即使没有任何编程基础的人，也可以掌握；在 R 环境下可以很轻松地进行数据分析，并绘制出可供图书出版级别的数据可视化图表。因为 R 语言的这些优势，R 的用户不断开发出各种增强功能的软件包，现在 R 语言已经被广泛用于经济计量、财经分析和商务智能等各个领域。

本书的作者米勒先生是美国西北大学的教授，他酷爱运动和电影。书中的案例包含了非常多的美国文化。例如在每个章节的开头，都以一段美国电影对白开始，如果了解这些电影，或者了解这段对白出现的场景，便会知道这段对白跟这个章节所讲述的内容之间的契合。在“文本分析”和“情绪分析”的章节，原始的数据都是非结构化的文本数据，需要先格式化处理，并使用“语料库”的技术进行分析，因为英文这种语言的特性，会让我们的读者很难理解将文本数据格式化所采用的方式。

非常感谢“炼数成金 dataguru.cn”社区的创始人黄志洪先生，在我翻译本书的过程中给予我诸多的建议和帮助。感谢“炼数成金”社区的何翠仪小姐、张晓仪小姐和吴仕灿先生的校阅。非常感谢我的家人，支持我在过去的这段日子仅因我个人的兴趣而花费的大量时间。感谢我的父母从小对我严谨的治学态度的培养，他们一向是我进步的榜样。也感谢我身边一直鼓励我的所有朋友。

作为这本书的译者，我尽力对书中的一些美国文化的背景做了一些注释。能使读者们最大程度地从这本译著中获益，是我的初衷。

陈宇红

2016 年 1 月于纽约

前言

“托托，我觉得我们已经不在堪萨斯城了哦。”

——陶乐思·高尔（茱蒂·格兰特饰演），美国电影《绿野仙踪》（The Wizard of Oz 1939）

数据和算法统治了当下。欢迎您来到这个崭新的商务世界，一个必须通过强大的分析能力和信息的交流，才能取得稍纵即逝的竞争优势的、快节奏的、数据密集的开源环境。

现有的许多论述预测分析或数据科学的书籍，谈论的是战略和管理；还有一部分着眼于方法和模型；其余的着重于信息技术和代码。本书少有的试图将上述三者结合起来，深受建模人员、程序员和商务经理的喜爱。

我们已经意识到了通过分析的手段来获得竞争优势的重要性。我们为研究人员和分析师提供一个现成的资源和建模技术参考指南；我们为程序员展示如何编写解决实际商务问题的基本代码；我们将模型运行的结果转化成管理人员可以理解的文字和图片；我们解释数据和模型的含义。

随着数据采集和储存数量的快速增长，随着各种可用于分析的数据的增长，随着每日数据的更新频率及需要分析的数据的增长，相较往日，数据分析变得至为重要。要获取竞争优势，就意味着必须实施新的信息管理和分析系统。这也同样意味着要改变经营的方式。

数据科学这个领域拥有巨大的文献资料，来自于诸多的学科和应用程序。相关的开源代码也在迅速增长。事实上，这是对我们撰写一本全面的预测分析和数据科学指南书籍的挑战。

我们着眼于真实的案例和真实的数据。我们提供一系列范例：在本书中的每一章，将针对一个特定的商务问题作出分析并附上应用程序。我们提供有意义的解决方案。通过对建模技术和编程工具的实际演示，把抽象化的概念转化为具体的例子，让完全可以成功运行的案例程序易于理解。

我们的目标是对预测分析和数据科学做一个概述，让大多数的读者能够读懂。本书没有很多数学理论，统计人员和建模人员可以从参考文献获取详细的推导方法。我们这里仅仅使用简单的文字和可视化的数据来显示商务问题的解决方案。

看过了这本书的主题之后，可能会有人想知道我到底是经典统计的拥护者还是贝叶斯阵营的。在美国明尼苏达大学统计学院时，我创立一个对经典统计及贝叶斯理论都予以尊重的观点。无论是采用经验贝叶斯方法，还是从经典统计学习的方法入手，都会存在一个结合机器学习和经典统计学的领域，这个观点我深以为然。当涉及建模和推理这样的问题时，我是一个实用主义者。我希望大家能够理解我所做的工作，以及我所表达的不确定性。

在世界各地成千上万的专家的帮助下，让我们能够出版这本书。他们对开源环境贡献了时间和想法。开源环境的增长及易于发展的特点，确保了已开发出的解决方案将会成为未来许多年的中心所在。阿拉丁神灯里面的精灵已经跳出油灯获得自由，在帷幕的后面施展魔术——高深的科学不再神奇，秘密正在显露。这本书正是这个进程中的一部分。

本书采用的大多数数据是从公共数据资源中取得的。美国职棒大联盟的促销和上座率数据来自埃里卡·科斯特洛（Erica Costello）先生。莎朗·张伯伦（Sharon Chamberlain）女士对计算机选择研究数据方面的工作给予了有力的支持。阿维·曼德尔鲍姆（AviMandelbaum）和宜兰·高迪（IlanGuedj）提供了“匿名银行”呼叫中心的数据。每章开头的电影对白，承蒙互联网电影数据库的许可使用。斯坦福大学的安德鲁·L·摩斯（Andrew L. Mass）和他的同事承担了获取互联网电影数据库（IMDb）电影评论数据许可的工作。本书的某些范例的灵感来自于同下面公司的合作项目：佛罗里达 ToutBay of Tampa 公司、NCR Comten、联合惠普（HP）公司、纽约网站分析公司、威斯康星州麦迪逊市的 Sunseed Research LLC，以及麦迪逊市联合出租车公司 Union Cab Cooperative of Madison。

我们都工作在开源的社区，彼此分享程序资源。我们所做的真实工作以我们编写的程序代码呈现。所有的人都可以看，如果你们愿意，还可以调试这些程序代码。为了便于学生的学习，每段程序代码均包含逐句注释和建议，可供更深入的研究。本书所涉及的所有数据集和计算机程序代码可以从本书的网站 http://www.ftpress.com/miller/ 上下载。

在过去的很多年，有许多人对我的人生发展影响颇大，他们都是优秀的思想家和善良的人，是我的老师和导师，我永远铭谢。难过的是，伍尔西斯学院（Ursinus College）哲学系的杰拉尔德·哈恩·欣克尔（Gerald Hahn Hinkle）、语言学系的 艾伦·雷克·莱斯（Allan Lake Rice）、明尼苏达大学哲学系的赫伯特·费格尔（Herbert Feigl）已经离开了我们。我亦非常感谢明尼苏达大学心理测量学系的戴维·J·魏斯（David J. Weiss），以及经济学系的凯利·埃金（Kelly Eakin），她之前就职于俄勒冈大学。好老师是这样的伟大，使我受益终生。

感谢迈克尔·L·洛希尔（Michael L. Rothschild）、尼尔·M·福特（Meal M. Ford）、彼得·R·迪克森（Peter R. Dickson）和珍妮特·克里（Janet Christopher），在我们共同为威斯康星-麦迪逊大学，以及 A. C. 尼尔森市场研究中心工作期间提供的宝贵支持。

我居住在美国加利福尼亚州，位于道奇体育场北面四英里处。我在伊利诺伊州埃文斯顿市的西北大学任教，并且兼任佛罗里达州坦帕市 ToutBay 公司的产品开发总监，ToutBay 是一家数据科学公司。这所有的一切都受益于高速的互联网络。

我很幸运地参与了美国西北大学进修学院的研究生远程教育项目。感谢格伦·佛格提（Glen Fogerty）给了我在西北大学预测分析专业任教并担任领导者的机会。感谢我的同事和职员们，和我一起承担这个卓越的研究生专业的工作。感谢我的学生和资深教师，他们令我受教颇多。

ToutBay 是一家数据科学领域的新兴公司。我期许这家公司在其联合创始人 Greg Blence 先生的领导下，在今后的几年中大展宏图。非常感谢 Greg 邀请我加入他们，一起为公司的发展而努力，这也令我能够在实际的商务需求中进行演练，让学术研究和数据科学模型得以深远的发展。我们最终所期待的作为，就是要实施我们的理念和模型，让所有需要的人彼此分享。

感谢德科纳米公司（Texnology Inc.）的艾米·亨德里克森（Amy Hendrickson），她的美工使文字、表格、图表印刷得十分精美，这是另一种开源的成功。感谢高德纳（Donald Knuth）及 TEX / LATEX 社区贡献的这个美妙排版和出版系统。

感谢读者和审稿人提供的诸多帮助，他们是苏珊娜·卡伦德（Suzanne Callender）、菲利普·M·戈德费德（Philip M. Goldfeder）、梅尔文·奥特（Melvin Ott）和托马斯·P·瑞恩（Thomas P. Ryan）。在编写这一修订版时，洛雷娜·马丁（Lorena Martin）为本书的改进提供了诸多的反馈和建议。康迪斯·布拉德利（Candice Bradley）承担了审稿和文字编辑的双重工作。罗伊·L·桑福德（Roy L. Sanford）给予了有关统计模型和方案的专业建议。我非常感谢他们的反馈和鼓励。感谢我的编辑，珍妮·格拉瑟·莱文（Jeanne Glasser Levine），出版商培生教育出版社，使这本书能够出版。当然了，本书的任何文字问题、任何错误，或尚未完成的商务方案，由我个人负责。

我的好朋友布兰妮和她的女儿杰尼娅在时间允许时一直陪伴我。我的儿子丹尼尔不计我脾气的好坏，容忍并照料我的生活。他们的信任是我最大的负疚。

托马斯·米勒（Thomas W. Miller）
于加利福尼亚州的格伦代尔

目录

第 1 章　分析和数据科学 1

第 2 章　广告和促销 11

第 3 章　偏好与选择 25

第 4 章　购物篮分析 33

第 5 章　经济数据分析 44

第 6 章　运营管理 56

第 7 章　文本分析 71

第 8 章　情绪分析 93

第 9 章　体育分析 129

第 10 章　空间数据分析 148

第 11 章　品牌和定价 167

第 12 章　大数据的小游戏 200

附录 A　数据科学的方法 203

A.1　数据库和数据准备 204

A.2　经典统计与贝叶斯统计 206

A.3　回归与分类 208

A.4　机器学习 212

A.5　互联网和社交网络分析 213

A.6　推荐系统 215

A.7　产品定位 216

A.8 市场细分 218
A.9 选址 219
A.10 金融数据科学 220
附录 B 测量 222
附录 C 个案分析 232
C.1 回到我们的“摇头娃娃”个案 232
C.2 DriveTime 公司的轿车销售 233
C.3 钻石价更高 237
C.4 威斯康星 Dells 度假中心 240
C.5 个人电脑选择研究 244
附录 D 代码和实用程序 248

1 第1章 分析和数据科学

马奎尔先生："我想跟你讲一个词，就一个词。"

本："请讲。"

马奎尔先生："你真的在听么？"

本："是啊。"

马奎尔先生："塑胶。"

——马奎尔先生（华特·不莱克饰演）与本杰明布拉达克（达斯汀·霍夫曼饰演）的对白，美国电影《毕业生》（The Graduate 1967）[1]

取得一个哲学专业的学位职业前景并不怎么被看好，除非你想谋得一个教职，但哲学老师的职位却是很少的。尽管如此，我仍然非常看重我在文学院做哲学系学生的那些日子。我的学士学位的毕业论文是关于伯特兰·罗素（Bertrand Russell）的，这篇论文获得了荣誉论文的奖励。在明尼苏达大学的研究生院，我从师于一个真正伟大的哲学家赫伯特·费格尔（Herbert Feigl）。我阅读过有关科学和如何发现真理的文章，反之而行的理论被称为认识论。我最喜爱的哲学是逻辑经验主义。

虽然我的那些费格尔哲学"思考再思考"的日子早已离我远去，但早年的这些学术训练，使我对区分什么是真理，什么仅仅是泛泛而谈十分的有感觉。

"模型"是一个事物的表现，是对现实的渲染或描述。在数据科学中，一个典型的模型是一种将一组变量与其他变量联系起来的企图。虽然"模型"这个表述是有限的，是不精确的，但它是可用的，"模型"可以帮助我们了解这个世界。由于模型是基于数据而建

[1] 在20世纪60年代，塑胶工业在美国方兴未艾，是一个极具前景的新兴行业。作者在这里以此比喻数据分析也是一个极具前景的新兴领域。

立的，它所要表达的远不止于这里讨论的。

预测分析将管理、信息技术和建模连接起来。它在当前这个数据爆炸的时代应运而生。预测分析是数据科学，是成功企业、非营利机构及政府机构必不可少的多学科技能组合。无论是预测销售还是市场份额，寻找一个好的零售地点或是一个投资机会，确定消费群体以及目标市场，挖掘新产品的潜力以及这会对现有产品带来的风险评估，预测分析的建模方法都能找出问题的关键所在。

那些在预测分析领域工作的数据科学家，讲的是商务语言——会计、金融、市场和管理。他们不但精通诸如数据结构、算法、面向对象编程等信息技术，还熟练运用统计建模、机器学习和数学规划。数据科学家是方法论的折中者，借鉴多个科学学科的专业知识，把经验研究的结果转化为管理人员能够理解的图片和文字。

预测分析，与统计学有很多相像之处，都是在诸多的变量中寻找有意义的关系，并将其在模型中呈现出来。那些称之为响应变量的，就是试图预测的事情；而称之为解释变量或预测因子的，是那些通过观察、调控及控制等手段，能够产生响应的事情。

回归的方法帮助我们预测一个响应具有意义的幅度，例如：销售数量、股票价格或投资回报。分类的方法则是帮助我们预测响应的类别，例如：哪种品牌会畅销？消费者是否会买这种产品？客户是会偿清还是拖延贷款？某银行的交易是真实的还是欺诈？

预测问题定义为数据集里面潜在预测因子的宽度或数量，以及观察对象的深度或数量。诸多存在于商务、市场及投资分析中的潜在预测因子使分析变得非常困难，而那些成千上万的潜在预测因子与响应之间可能只有微弱关系。在计算机的帮助下，可以使用成百上千的模型去拟合数据子集，并使用另一部分数据子集进行测试，对每个预测因子进行评估。预测建模需要找出良好的预测因子子集。模型拟合数据良好优于模型拟合数据恶劣。简单的模型优于复杂的模型。

预测分析有三种普遍的研究和建模方法：传统方法、自适应数据法和依赖模型法。请参阅图 1.1。研究、统计推断和建模的传统方法从理论与模型自身着手，采用经典或者贝叶斯统计推断方法。传统方法，诸如线性回归和 Logistic 回归，需要预估线性预测因子参数。建模过程涉及数据对模型的拟合和模型的诊断检验。在使用传统模型做预测之前，我们必须先进行模型验证。

当采用自适应数据法时，要从数据和研究着手，通过这些数据找出有用的预测因子。在进行分析之前，不会考虑太多理论或假设。这是机器学习的领域，有时候也会被称为统计学习或数据挖掘。自适应数据法顺应现有的数据，代表了非线性关系和变量之间的相互作用。数据确定模型。数据自适应法是数据驱动的。同传统模型一样，在使用数据自适应模型做预测之前，也需要验证自适应模型。

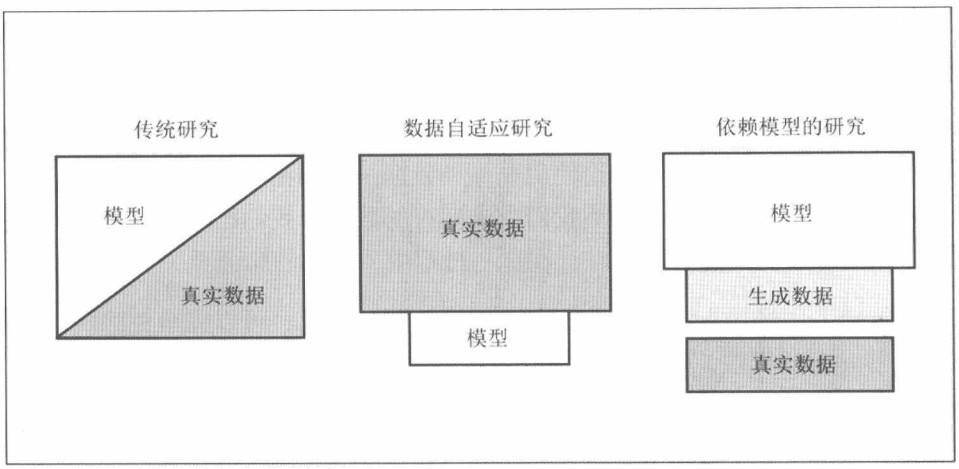

图 1.1　研究的数据和模型

第三种方法是依赖模型法。它从模型规范开始，并使用该模型生成数据、预测或建议。模拟和数学规划法，以及运筹学的主要工具等，都是依赖模型研究的范例。当采用依赖模型或模拟法时，我们会通过对生成数据与真实数据的比较，来优化模型。我们要求模型模拟消费者、企业和市场行为，如同真正的消费者、企业和市场行为。与真实数据相比较，是验证依赖模型的方法。

实际上常常是将模型和方法相结合才能得到最好效果。例如：在金融研究领域的一个应用程序，互惠基金经理正在寻找更多的股票作为基金的投资目标。金融工程师采用自适应数据模型（也许是一个神经网络）来搜索成千上万的绩效指标和股票，生成了一个可被进一步分析的股票子集。然后，金融工程师采用理论基础法（CAPM 资本资产定价模型）分析这个股票子集，确定一个较小的股票列表推荐给基金经理。最后的这个步骤，正是使用了依靠模型的研究（数学规划），工程师确定了投资目标中最低风险资本投资的个股。

数据可依照观测单元、时间和空间进行整理。观测单元或横断面单元可能是一个消费者个体，或商务，或任何其他收集和分组的基础数据。按照时间整理的数据，可采用秒、分钟、小时、天等为单位。按照空间或位置整理的数据常常以经度和纬度为单位。

举例来说，我们需要观测有多少顾客在星期一进入了位于美国加利福尼亚州格伦代尔镇上的杂货店。这个观测的分析单位是进入杂货店的顾客数目，时间单元是星期一。忽略杂货店的空间位置，这些都是横断面数据。假设我们为其中的一间商店工作，分析在六个月内顾客在每周的每一日的客流量数据，这些就是时间序列数据。然后我们来看看六个月内顾客进入格伦代尔镇所有杂货店的客流量，这些就是纵向或面板数据。为了完善我们的研究，我们还需要确定这些商店的经度和纬度，于是我们也拥有了空间或时空数据。其他的这些数据结构，都是在收集进入商店的顾客数量的同时，需要额外考虑测量的。我们观察商店销售、消费者或附近居民人口统计，格伦代尔街道交通状况，等等，这个案例就成

为具有多重时间序列和多元变量方法的案例。我们收集到的数据的整理方法，会影响我们采用的模型结构。

本书中所讨论的商务问题，让我们能够接触到很多类型的模型，例如横断面模型、时间序列模型、空间数据模型等。无论使用哪一种数据结构和相关模型，预测都是唯一的主题。使用我们已经拥有的数据来预测我们尚未拥有的数据，我们已经意识到预测的不稳定性。这就是推断和预测的过程。并且，在这一过程中，模型的验证至关重要。

我们可能会采用经典方法或贝叶斯方法来做预测。或者，可以完全免除传统的统计，仅仅依靠机器学习的算法。我们只使用有效果的方法[2]。我们的预测分析方法是基于以下一个简单的前提：

一个模型的价值在于其预测的质量。

从统计学中可以得知：必须量化不确定性。一方面，采用经典方法，使用置信区间、点估计与其相关的标准差，还有显著性检验以及 p 值。另一方面，沿袭贝叶斯统计的路径，使用后验概率分布、概率区间、预测区间、贝叶斯因子、主观（可能是离散的）先验。在对两个模型选择比较时，赤池信息量准则（AIC）指数或贝叶斯信息标准（BIC）指数可以帮助我们在拟合优度与简约之间取得平衡。

回到*训练并测试方案*的方法上来。我们将样本数据划分为训练集和测试集两类。我们在训练集上建模再以测试集对模型进行评估。简单的数据两区分法和数据三区分法，如图 1.2 所示。

将样本数据随机地分割为训练集和测试集，可能会极具偶然性，尤其是在小量样本数据集的情形下，所以有时我们会在统计试验中，进行若干次的随机分割，以便分割出的测试集具有平均性能指标。这是训练并测试方案的扩展和变化。

多重交叉检验是训练并测试方案的一种变化，如图 1.3 所示。将数据样本近似等分成 M 段，并进行一系列的测试。如图 1.3 中所示的五等分交叉验证，先以 B 到 E 子集做训练集，A 子集做测试集；之后，将 B 子集做测试集，其他的 A 子集，C 到 E 子集做训练集，诸如这般，直到每个子集都做过一次测试集。性能评估的结果是这五个测试集观测值的平均数。留一交叉评估法，是多重交叉检验的逻辑极端，在样本数据中有多少个测试集就会有多少个观测对象。

---

[2] 在统计文献中，Seymour Geisser（1929—2004）在 1993 年介绍了一种最佳描述贝叶斯预测推断的方法(Geisser 1993)。贝叶斯统计在创建人 Thomas Bayes 牧师（1706—1761）死后被命名为贝叶斯定理（贝氏定理）。在强调预测的成功性时，我们倾向于 Geisser 的理论。然而，我们的方法是纯粹经验主义的，无法依赖于经典或贝叶斯理论。

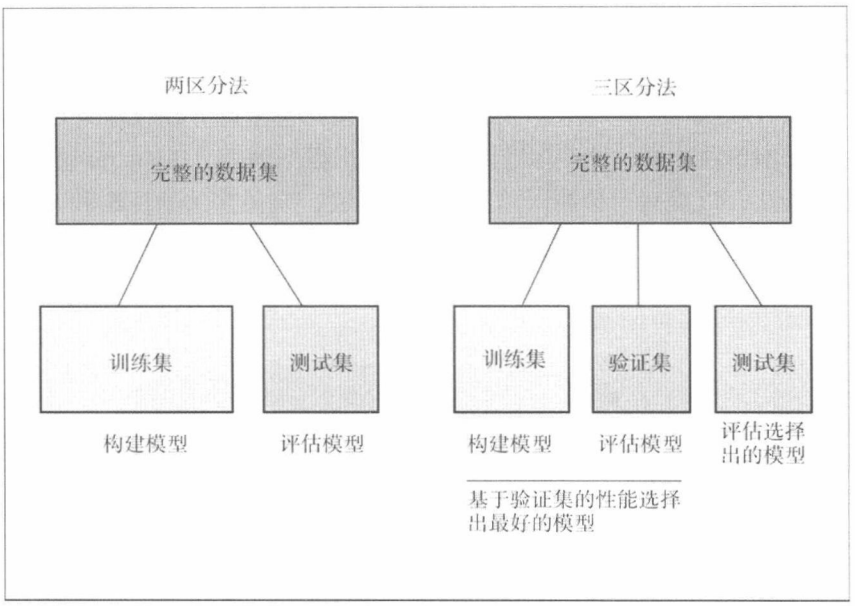

图 1.2　模型评估的训练并测试的方案

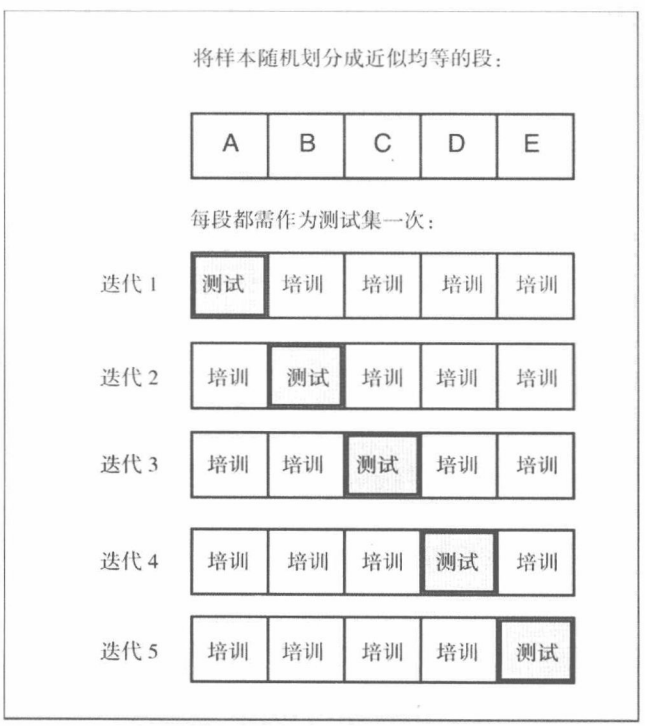

图 1.3　多重交叉验证法的训练并测试的方案

另一种训练并测试的方案是 Bootstrap 重采样分类法。如果样本接近采样的总体，那么从样本中抽取样本（所谓的重采样），也是接近被采样的总体的。如图 1.4 所示，引导

的过程中，采用反复重抽样来更迭样本。也就是说，通过多次随机抽样来更迭样本，对于每一次的重抽样，都计算利益统计。Bootstrap 重采样法分布的统计结果与样本分布的统计结果相似。那么 Bootstrap 重采样法有怎样的价值呢？它让我不必对总体分布做出假设，而可以对每个样本单独分析、估计标准误差，并做出概率报表。Bootstrap 重采样法也可用于留一交叉检验法，以提高对预测误差的估算准确度。有关交叉验证法和 Bootstrap 重采样法的文献，参见 Davison 和 Hinkley（1997），Efron 和 Tibshirani（1993），Hastie、Tibshirani 和 Friedman（2009）。

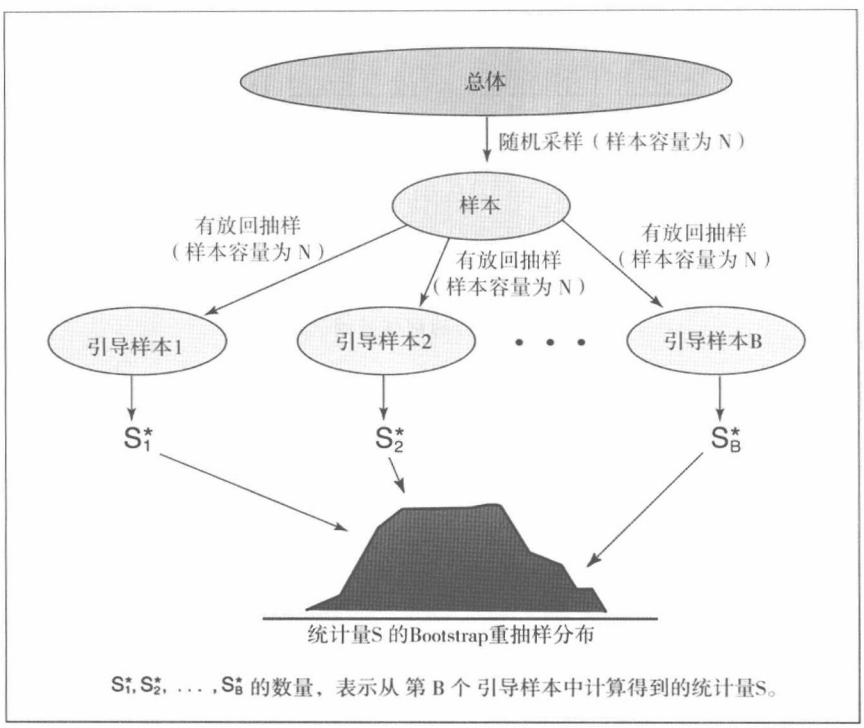

图 1.4 引导重采样的训练并测试的方案

数据可视化对数据科学工作至关重要。在本书的范例中演示了数据可视化在发现、诊断和设计中的重要性。我们将采用探索数据分析（发现）工具和统计建模（诊断）工具。在为管理层展示分析结果时，将使用图形来演绎（设计）。

在展示统计图表和数据可视化的重要性方面。没有什么能比称之为“安斯库姆四组 Anscombe Quartet”的数据更有说服力的例子了。安斯库姆于 1973 年发表了如表 1.1 中所示的数据集。观察表格中的数据，一般来说，读者都会注意到第四组数据集，明显与别的组不同。那么前三组数据集呢？在 x 与 y 之间的关系模式是否也存在明显的差异？