Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition

数据挖掘技术(第3版)

——应用于市场营销、销售与客户关系管理

[美] Gordon S. Linoff
Michael J. A. Berry 著
巢文涵 张小明 王芳 译

清华大学出版社

数据挖掘技术(第 3 版)

——应用于市场营销、销售与客户关系管理

[美] Gordon S. Linoff
Michael J. A. Berry 著

巢文涵 张小明 王芳 译

清华大学出版社

北　京

Gordon S. Linoff，Michael J. A. Berry
Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition
EISBN：978-0-470-65093-6

图书在版编目(CIP)数据

数据挖掘技术(第 3 版)——应用于市场营销、销售与客户关系管理/ (美)林那夫(Linoff, G. S.)，(美)贝里(Berry, M.J.A.) 著；巢文涵，张小明，王芳 译. —北京：清华大学出版社，2013.3

书名原文：Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition

ISBN 978-7-302-31014-3

Ⅰ. ①数… Ⅱ. ①林… ②贝… ③巢… ④张… ⑤王… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2012)第 304161 号

责任编辑：王　军　刘伟琴
装帧设计：康　博
责任校对：蔡　娟
责任印制：

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 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：40.25　　**字　　数**：980 千字
版　　次：2013 年 3 月第 1 版　　**印　　次**：2013 年 3 月第 1 次印刷
印　　数：1~4000
定　　价：79.80 元

产品编号：

作 者 简 介

Gordon S. Linoff 和 Michael J. A. Berry 在数据挖掘领域的知名度众所周知。他们是 Data Miners 公司——一家从事数据挖掘的咨询公司——的创始人，而且他们已经共同撰写了一些在该领域有影响力和得到广泛阅读的书籍。他们共同撰写的第一本书是 *Data Mining Techniques* 的第一个版本，于 1997 年出版。自那时起，他们就一直积极地挖掘各种行业的数据。持续的实践分析工作使得两位作者能够紧跟数据挖掘、预测以及预测分析领域的快速发展。Gordon 和 Michael 严格地独立于供应商。通过其咨询工作，作者接触了所有主要软件供应商(以及一些小的供应商)的数据分析软件。他们相信好的结果不在于是采用专用的还是开源的软件，命令行的还是点击的软件，而是在于创新思维和健全的方法。

Gordon 和 Michael 专注于数据挖掘在营销和客户关系管理方面的应用——例如，为交叉销售和向上销售改进推荐，预测未来的用户级别，建模客户生存期价值，根据用户行为对客户进行划分，为访问网站的客户选择最佳登录页面，确定适合列入营销活动的候选者，以及预测哪些客户处于停止使用软件包、服务或药物治疗的风险中。Gordon 和 Michael 致力于分享他们的知识、技能以及对这个主题的热情。当他们自己不挖掘数据时，他们非常喜欢通过课程、讲座、文章、现场课堂，当然还有你要读的这本书来教其他人。经常可以发现他们在会议上发言和在课堂上授课。作者还在 blog.data-miners.com 维护了一个数据挖掘的博客。

Gordon 生活在曼哈顿。在本书之前，他最近的一本书是 *Data Analysis Using SQL and Excel*，已经由 Wiley 于 2008 年出版。

Michael 生活在马萨诸塞州剑桥市。他除了在 Data Miners 从事咨询工作之外，还在波士顿大学卡罗尔管理学院讲授市场营销分析(Marketing Analytics)课程。

致　　谢

令我们幸运的是，在我们周围到处都是一些极有才华的数据挖掘人员，所以首先要感谢我们在 Data Miners 公司的过去的以及现在的同事 Will Potts、Dorian Pyle 和 Brij Masand，从他们身上我们学会了很多。还有一些客户，由于我们与他们的合作非常密切，所以我们同样把他们看成是同事和朋友，Harrison Sohmer、Stuart E. Ward、III 和 Michael Benigno 就属于这一类。我们的编辑 Bob Elliott 使我们能够(或多或少)如期完工，并帮助我们保持一致的风格。

SAS 研究所和数据仓库研究所在过去 12 年为我们提供了无与伦比的教学机会。我们要特别感谢 Herb Edelstein(现已退休)、Herb Kirk、Anne Milley、Bob Lucas、Hillary Kokes、Karen Washburn，以及其他许多人，是他们使得这些课程成为可能。

在过去的一年中，当我们在撰写本书时，几位朋友和同事一直都非常支持我们。我们要感谢 Diane 以及 Savvas Mavridis、Steve Mullaney、Lounette Dyer、Maciej Zworski、John Wallace、Paul Rosenblum 和 Don Wedding。

我们还要感谢多年来所有与我们一起参与数据挖掘活动的人。我们从他们每个人身上都学到了许多东西。其中许多人这些年一直在帮助我们：

Alan Parker
Dave Waltz
Craig Stanfi ll
Dirk De Roos
Michael Alidio
Michael Cavaretta
Dave Duling
Jeff Hammerbacher
Andrew Gelman
Doug Newell
Ed Freeman
Erin McCarthy
Josh Goff
Karen Kennedy
Ronnie Rowton
Kurt Thearling
Mark Smith
Gary King
Tim Manns
Jeremy Pollock
Richard James
Georgia Tourasi
Avery Wang
Eric Jiang
Bruce Rylander
Daryl Berry
Adam Schwebber
Tiha Ghyczy
Usama Fayyad
Patrick Ott
John Muller
Frank Travisano
Jim Stagnito
Stephen Boyer

Nick Radcliffe
Patrick Surry
Ronny Kohavi
Terri Kowalchuk
Victor Lo
Yasmin Namini
Zai Ying Huang
Amber Batata
Yugo Kanazawa
Xu He
Kiran Nagarur
Ramana Thumu
Jacob Hauskens
Jeremy Pollock
Lutz Hamel

当然，我们依然要感谢所有我们在第一版中感谢过的人们:
Bob Flynn
Bryan McNeely
Claire Budden
David Isaac
David Waltz
Dena d'Ebin
Diana Lin
Don Peppers
Ed Horton
Edward Ewen
Fred Chapman
Gary Drescher
Gregory Lampshire
Janet Smith
Jerry Modes
Jim Flynn
Kamran Parsaye
Karen Stewart
Larry Bookman
Larry Scroggins
Lars Rohrberg
Lounette Dyer
Marc Goodman
Marc Reifeis
Marge Sherold
Mario Bourgoin
Prof. Michael Jordan
Patsy Campbell
Paul Becker
Paul Berry
Rakesh Agrawal
Ric Amari
Rich Cohen
Robert Groth
Robert Utzschnieder
Roland Pesch
Stephen Smith
Sue Osterfelt
Susan Buchanan
Syamala Srinivasan
Wei-Xing Ho
William Petefi sh
Yvonne McCollin

最后，我们要感谢家人和朋友，特别要感谢 Stephanie 和 Giuseppe，她们在我们撰写这本书期间默默地忍受和奉献。

前　言

15 年前，Michael 和我合写了这本书的第一版。那本书 400 页多一点，通过弥合技术和实践之间的差距，通过帮助商业人士了解数据挖掘技术以及帮助技术人员理解这些技术的商业应用，从而满足了我们调查数据挖掘领域的目标。当 Wiley 出版社的编辑 Bob Elliott 让我们撰写 *Data Mining Techniques* 的第 3 版时，我们欣然同意，浑然忘记了撰写一本书给我们的个人生 ai ok de 活所带来的牺牲。我们也知道新版本将会大幅改写以前的两个版本。

在过去的 15 年中，这个领域无论是在内涵上还是在字面上都已经得到了扩展，这本书中同样如此。2004 年出版了第 2 版，这一版本增加到了 600 页，并引入了两个新的章节，分别介绍了生存分析和统计算法这两种新的关键技术，它们对于数据挖掘人员而言已经变得(并依然)越来越重要。现在的这个版本将再度引入新的技术领域——尤其是文本挖掘和主成分分析，同时在所有章节中引入了丰富的新实例，并增强了技术描述。这些例子来自各行各业，其中包括金融服务、零售、电信、媒体、保险、保健和基于 Web 的服务。

作为该领域的从业人员，我们也一直在学习。我们现在大约已经有半个世纪的数据挖掘方面的经验。自 1999 年以来，Michael 和我一直在通过 SAS 研究所的业务知识系列(本系列与业务的软件方面分离，引入外部专家讲授非软件特定的课程)、数据仓库研究所以及许多不同企业的现场课程进行授课。我们在这些课程中的讲师角色使我们有机会接触成千上万各种行业中的不同业务人员。其中商业数据挖掘技术这门课程就是基于这本书的第二版。这些课程提供了大量有关数据挖掘主题的反馈，比如现实世界的人们正在做什么，以及如何以最佳方式来表示这些思想，从而使它们易于理解。大部分的反馈在这个新版本中都有所反映。我们从学生那里学到的东西看起来与学生从我们这里学到的一样多。

过去两年，Michael 也一直在波士顿大学的卡罗尔管理学院讲授市场营销分析课程。*Data Mining Techniques* 的前两个版本在许多学院和大学的课程中也广受欢迎，包括商业课程，以及越来越多的数据挖掘课程——在过去十年中其已在各大学中出现。虽然并不打算作为教科书，但是 *Data Mining Techniques* 为所有类型的学生提供了一个出色的概述。多年来，我们已经在我们的网站上提供了各种可用的数据集，讲师可以在课程中使用它们。

这本书分为 4 个部分。第一部分讨论数据挖掘的业务上下文。第 1 章对数据挖掘进行了概述，并给出了如何将其用于现实世界的例子。第 2 章解释了数据挖掘的良性循环，以及数据挖掘如何帮助理解客户。这一章有几个例子，显示了如何在整个客户生命周期中使用数据挖掘。第 3 章是数据挖掘方法的概述。第 5 章和第 12 章对整体方法进行了精化，分别对应于有指导和无指导数据挖掘。第 4 章涉及商业统计学知识，介绍了一些贯穿整本书其余部分的关键技术思想。这一章还扩展了 MyBuys 的案例研究，显示了用于分析 A/B 营销测试结果的不同方法的长处和短处。

早期版本把所有的数据挖掘技术都放在一个单一的部分。我们现在决定把这些技术划分为两个不同的类别，因此有指导和无指导技术分别拥有它们各自的章节。有指导数据挖掘部分首先在第 3 章针对有指导数据挖掘对数据挖掘方法进行了精化。后续章节则介绍各种有指导数据挖掘技术，其中包括统计技术、决策树、神经网络、基于记忆的推理、生存分析以及遗传算法。

在第 2 版中已经覆盖了所有的有指导数据挖掘技术。然而，我们在几个重要方面对它们进行了增强，特别是包含了更多在现实世界中使用它们的例子。第 7 章现在包括一个关于美国银行提升建模的案例研究，同时还介绍了支持向量机。第 8 章讨论了径向基函数神经网络。第 9 章现在有两个很有趣的案例研究，一个是关于 Shazam 如何识别歌曲，另一个使用 MBR 帮助放射学家确定 X 线检查是正常还是异常。第 10 章介绍生存分析，其中包括了一个针对客户价值的急需的讨论。第 11 章介绍了遗传算法，其中还包括群体智慧——另一个来自“计算生物学”世界的相关概念，其在数据挖掘领域具有广阔的应用前景。

第三部分专门讨论了无指导数据挖掘技术。第 12 章解释了四种不同类型的无指导数据挖掘。聚类算法分成两章。其中第 13 章重点介绍了最常见的聚类技术——K-均值聚类及其三个变体：K-中位数、K-中心点和 K-众数。同时它还扩展了关于群集解释的讨论，无论采用哪种技术来识别群集，解释群集都非常重要。第 14 章介绍了许多技术，包括层次聚类、分裂聚类、自组织网络和高斯混合模型(期望值最大化聚类)，它在此版本中是新的内容。第 15 章的购物篮分析在例子方面进行了加强，这些例子超越了关联规则，其中还包括一个关于种族营销的案例研究。第 16 章是无指导数据挖掘部分的最后一章，在 20 世纪 90 年代，当我们写这本书的第 1 版时，它几乎还处于外围。现在，它已经处于相当中央的位置，正如这一章的三个案例研究所示。

这本书的最后一部分专注于数据挖掘这一名称中的数据。第 17 章介绍支持数据的计算机体系结构，例如关系数据库、数据仓库和数据集市。同时，它还介绍了 Hadoop 和分析沙箱，它们都用于处理不适合关系数据库和传统数据挖掘工具的数据。两个早期的版本也有一章介绍数据挖掘的数据准备。由于这个问题如此重要，所以这个版本将该主题分成三章。第 18 章是关于如何在数据中发现客户和构建客户签名，这是一种许多数据挖掘算法所使用的数据结构。第 19 章涉及派生变量，以及如何定义变量以帮助模型表现更好的提示和技巧。第 20 章侧重于如何减少变量的数量，无论是针对诸如神经网络之类的喜欢较少变量的技术，还是出于数据可视化的目的。这一章的关键技术之一——主成分，在这个版本中是新的内容。

第 21 章涉及的主题本身也可以是一本书，这一主题就是文本挖掘。由于分析文本是构建在本书之前所介绍的许多思想之上，所以我们认为涉及文本挖掘的章节必须放在这本书的最后。其压轴出场凸显了文本挖掘是贯穿本书所覆盖主题的高潮部分。来自 DIRECTV 的最后一个案例研究，不仅是针对业务客户服务方面的一个有趣的文本挖掘应用，同时也是一个极佳的实践中的数据挖掘例子。

与前两个版本一样，这本书的读者对象也是当前的和未来的数据挖掘从业人员和他们的经理。它不适合寻找如何实现各种数据挖掘算法详细说明的软件开发人员，也不适合试图改进这些算法的研究人员，虽然这两组人都可以通过了解这种软件如何使用而受益。各种思想均是以非技术语言提出，其中尽量减少数学公式和神秘行话的使用。整本书的重点

既包括技术解释，也包括数据挖掘的实际应用，因此这些技术都包含了实际业务上下文的例子。

总之，我们试图写这样一本书：当我们开始自己的数据挖掘职业生涯时，也会想要阅读它。

——Gordon S. Linoff，2011 年 1 月于纽约

目　录

第 1 章　什么是数据挖掘以及为什么要进行数据挖掘 ……1
1.1　什么是数据挖掘 ……2
1.1.1　数据挖掘是一项业务流程……2
1.1.2　大量的数据……2
1.1.3　有意义的模式和规则……3
1.1.4　数据挖掘和客户关系管理……3
1.2　为什么是现在 ……4
1.2.1　数据正在产生……5
1.2.2　数据正存在于数据仓库中……5
1.2.3　计算能力能够承受……5
1.2.4　对客户关系管理的兴趣非常强烈……5
1.2.5　商业的数据挖掘软件产品变得可用……6
1.3　数据挖掘人员的技能 ……7
1.4　数据挖掘的良性循环 ……7
1.5　业务数据挖掘的案例研究 ……8
1.5.1　识别美国银行的业务挑战……9
1.5.2　应用数据挖掘……9
1.5.3　对结果采取行动……10
1.5.4　度量数据挖掘的影响……11
1.6　良性循环的步骤 ……11
1.6.1　识别业务机会……12
1.6.2　将数据转换为信息……13
1.6.3　根据信息采取行动……14
1.6.4　度量结果……15
1.7　良性循环上下文中的数据挖掘……17
1.8　经验教训 ……19

第 2 章　数据挖掘在营销和客户关系管理中的应用 ……21
2.1　两个客户生存周期 ……21
2.1.1　客户个人生存周期……21
2.1.2　客户关系生存周期……22
2.1.3　基于订阅的关系和基于事件的关系……23
2.2　围绕客户生存周期组织业务流程……25
2.2.1　客户获取……25
2.2.2　客户激活……27
2.2.3　客户关系管理……29
2.2.4　赢回……29
2.3　数据挖掘应用于客户获取……30
2.3.1　识别好的潜在客户……30
2.3.2　选择通信渠道……30
2.3.3　挑选适当的信息……31
2.4　数据挖掘示例：选择合适的地方做广告 ……31
2.4.1　谁符合剖析……31
2.4.2　度量读者群的适应度……33
2.5　数据挖掘改进直接营销活动……34
2.5.1　响应建模……35
2.5.2　优化固定预算的响应……35
2.5.3　优化活动收益率……37
2.5.4　抵达最受信息影响的人……40
2.6　通过当前客户了解潜在客户……41
2.6.1　在客户成为“客户”以前开始跟踪他们……41
2.6.2　收集新的客户信息……41

2.6.3 获取时间变量可以预测将来的结果……42
2.7 数据挖掘应用于客户关系管理……42
2.7.1 匹配客户的活动……42
2.7.2 减少信用风险……43
2.7.3 确定客户价值……44
2.7.4 交叉销售、追加销售和推荐……44
2.8 保留……45
2.8.1 识别流失……45
2.8.2 为什么流失是问题……46
2.8.3 不同类型的流失……46
2.8.4 不同种类的流失模型……47
2.9 超越客户生存周期……48
2.10 经验教训……48

第 3 章 数据挖掘过程……51
3.1 会出什么问题……51
3.1.1 学习的东西不真实……52
3.1.2 学习的东西真实但是无用……55
3.2 数据挖掘类型……56
3.2.1 假设检验……56
3.2.2 有指导数据挖掘……60
3.2.3 无指导数据挖掘……61
3.3 目标、任务和技术……61
3.3.1 数据挖掘业务目标……62
3.3.2 数据挖掘任务……62
3.3.3 数据挖掘技术……66
3.4 制定数据挖掘问题：从目标到任务再到技术……66
3.4.1 选择广告的最佳位置……66
3.4.2 确定向客户提供的最佳产品……67
3.4.3 发现分支或商店的最佳位置……68
3.4.4 根据未来利润划分客户……68
3.4.5 减少暴露于违约的风险……69
3.4.6 提高客户保留……69
3.4.7 检测欺诈性索赔……70
3.5 不同技术对应的任务……71
3.5.1 有一个或多个目标……72
3.5.2 目标数据是什么……72
3.5.3 输入数据是什么……72
3.5.4 易于使用的重要性……72
3.5.5 模型可解释性的重要性……72
3.6 经验教训……73

第 4 章 统计学入门：关于数据，你该了解些什么……75
4.1 奥卡姆(Occam)剃刀……76
4.1.1 怀疑论和辛普森悖论……77
4.1.2 零假设(Null Hypothesis)……77
4.1.3 p-值……78
4.2 观察和度量数据……79
4.2.1 类别值……79
4.2.2 数值变量……87
4.2.3 更多的统计思想……89
4.3 度量响应……90
4.3.1 比例标准误差……90
4.3.2 使用置信区间比较结果……91
4.3.3 利用比例差异比较结果……92
4.3.4 样本大小……93
4.3.5 置信区间的真正含义是什么……94
4.3.6 实验中检验和对照的大小……95
4.4 多重比较……96
4.4.1 多重比较的置信水平……96
4.4.2 Bonferroni 修正……96
4.5 卡方检验……97
4.5.1 期望值……97
4.5.2 卡方值……98
4.5.3 卡方值与比例差异的比较……100
4.6 示例：区域和开局卡方……101
4.7 案例研究：利用 A/B 检验比较两种推荐系统……103
4.7.1 第一个指标：参与会话……104
4.7.2 第二个指标：每个会话的日收益……104
4.7.3 第三个指标：每天谁取胜……106
4.7.4 第四个指标：每个会话的平均收益……106

4.7.5 第五个指标：每个客户的增量收益 107
4.8 数据挖掘与统计 107
4.8.1 基本数据中没有度量误差 108
4.8.2 大量的数据 108
4.8.3 无处不在的时间依赖性 109
4.8.4 实验非常困难 109
4.8.5 数据被删截 109
4.9 经验教训 110

第 5 章 描述和预测：剖析与预测建模 113
5.1 有指导数据挖掘模型 113
5.1.1 定义模型结构和目标 114
5.1.2 增量响应建模 115
5.1.3 模型稳定性 116
5.1.4 模型集中的时间帧 117
5.2 有指导数据挖掘方法 119
5.3 步骤 1：把业务问题转化为数据挖掘问题 120
5.3.1 如何使用结果 122
5.3.2 如何交付结果 122
5.3.3 领域专家和信息技术的角色 123
5.4 步骤 2：选择合适的数据 123
5.4.1 什么数据可用 124
5.4.2 多少数据才足够 125
5.4.3 需要多久的历史 125
5.4.4 多少变量 126
5.4.5 数据必须包含什么 126
5.5 步骤 3：认识数据 126
5.5.1 检查分布 127
5.5.2 值与描述的比较 127
5.5.3 验证假设 127
5.5.4 询问大量问题 128
5.6 步骤 4：创建模型集 128
5.6.1 聚合客户签名 128
5.6.2 创建一个平衡的样本 129
5.6.3 包括多个时间帧 130
5.6.4 创建一个预测模型集 130
5.6.5 创建一个剖析模型集 131
5.6.6 划分模型集 132
5.7 步骤 5：修复问题数据 132
5.7.1 分类变量的值太多 133
5.7.2 包含偏态分布和离群点的数值变量 133
5.7.3 缺失值 133
5.7.4 含义随时间而变化的值 134
5.7.5 不一致的数据编码 134
5.8 步骤 6：转换数据以揭露信息 134
5.9 步骤 7：构建模型 134
5.10 步骤 8：评估模型 135
5.10.1 评估二元响应模型和分类器 135
5.10.2 利用提升评估二元响应模型 136
5.10.3 利用提升图评估二元响应模型分数 137
5.10.4 利用剖析模型评估二元响应模型得分 139
5.10.5 使用 ROC 图表评估二元响应模型 139
5.10.6 评估估计模型 141
5.10.7 利用分数排名评估估计模型 141
5.11 步骤 9：部署模型 142
5.11.1 模型部署中的实际问题 142
5.11.2 优化模型以进行部署 143
5.12 步骤 10：评估结果 143
5.13 步骤 11：重新开始 144
5.14 经验教训 144

第 6 章 使用经典统计技术的数据挖掘 147
6.1 相似度模型 147
6.1.1 相似度和距离 148

6.1.2 示例：产品普及率的相似度模型 148
6.2 表查询模型 153
6.2.1 选择维度 153
6.2.2 维度的划分 154
6.2.3 从训练数据到得分 154
6.2.4 通过删除维度处理稀疏和缺失数据 155
6.3 RFM：一种广泛使用的查询模型 155
6.3.1 RFM 单元格迁移 156
6.3.2 RFM 与测试和度量(Test-and-Measure)方法论 156
6.3.3 RFM 和增量响应建模 157
6.4 朴素贝叶斯模型 158
6.4.1 概率论的一些思想 158
6.4.2 朴素贝叶斯计算 160
6.4.3 与表查询模型的比较 160
6.5 线性回归 161
6.5.1 最佳拟合曲线 162
6.5.2 拟合的优点 164
6.5.3 全局效应 166
6.6 多元回归 166
6.6.1 等式 166
6.6.2 目标变量的范围 166
6.6.3 解释线性回归方程的系数 167
6.6.4 用线性回归捕捉局部影响 168
6.6.5 使用多元回归的其他注意事项 169
6.6.6 多元回归的变量选择 170
6.7 逻辑回归分析 171
6.7.1 建模二元输出 171
6.7.2 逻辑函数 172
6.8 固定效应和分层效应 174
6.8.1 分层效应 175
6.8.2 内部效应与之间效应 175
6.8.3 固定效应 175
6.9 经验教训 177
第 7 章 决策树 179
7.1 决策树是什么以及如何使用 180
7.1.1 一棵典型的决策树 180
7.1.2 使用决策树学习客户流失 181
7.1.3 使用决策树来了解数据和选择变量 182
7.1.4 使用决策树生成排名 183
7.1.5 使用决策树估计类别概率 183
7.1.6 使用决策树分类记录 184
7.1.7 使用决策树估计数值 184
7.2 决策树是局部模型 184
7.3 决策树的生长 187
7.3.1 发现初始划分 187
7.3.2 生成整棵决策树 189
7.4 寻找最佳划分 190
7.4.1 Gini(总体多样性)作为划分标准 191
7.4.2 熵减少或信息增益作为划分标准 192
7.4.3 信息增益率 193
7.4.4 卡方检验作为划分标准 194
7.4.5 增量响应作为划分标准 195
7.4.6 减小方差作为数值型目标的划分标准 196
7.4.7 F 检验 198
7.5 剪枝 198
7.5.1 CART 剪枝算法 198
7.5.2 悲观修剪：C5.0 剪枝算法 202
7.5.3 基于稳定性的修剪 202
7.6 从决策树中提取规则 203
7.7 决策树变种 204
7.7.1 多路划分 204
7.7.2 一次在多个字段上进行划分 205
7.7.3 创建非矩形框 205
7.8 评估决策树的质量 209
7.9 什么时候使用决策树才合适 209
7.10 案例研究：咖啡烘焙厂的过程控制 210

7.10.1 模拟器的目标 …… 210
7.10.2 构建烘焙机模拟器 …… 210
7.10.3 评价烘焙机模拟器 ……211
7.11 经验教训……211

第 8 章 人工神经网络 ……213
8.1 历史回顾 ……214
8.2 生物学模型 ……215
8.2.1 生物神经元…… 216
8.2.2 生物输入层…… 217
8.2.3 生物输出层…… 217
8.2.4 神经网络与人工智能…… 217
8.3 人工神经网络 ……218
8.3.1 人工神经元…… 218
8.3.2 多层感知器…… 220
8.3.3 神经网络的一个例子…… 221
8.3.4 神经网络拓扑结构…… 223
8.4 应用实例：房地产估价 ……224
8.5 神经网络的训练 ……227
8.5.1 神经网络如何使用反向传播算法学习…… 227
8.5.2 神经网络的修剪…… 228
8.6 径向基函数网络 ……230
8.6.1 RBF 神经网络概述 …… 230
8.6.2 选择径向基函数的位置…… 231
8.6.3 万能逼近器…… 232
8.7 神经网络的应用 ……233
8.8 选择训练集 ……235
8.8.1 覆盖特征的所有值…… 235
8.8.2 特征数…… 235
8.8.3 训练集大小…… 235
8.8.4 输出的数目和值域…… 235
8.8.5 使用 MLP 的经验规则…… 235
8.9 数据准备 ……236
8.10 神经网络输出结果的解释 ….238
8.11 时间序列神经网络……239
8.11.1 时间序列建模 …… 239
8.11.2 时间序列神经网络的示例…… 240
8.12 神经网络模型是否能解释…..241
8.12.1 灵敏度分析 …… 241
8.12.2 使用规则来描述得分 …… 242
8.13 经验教训 ……242

第 9 章 最近邻方法：基于记忆的推理和协同过滤 ……245
9.1 基于记忆的推理 ……246
9.1.1 类众模型…… 247
9.1.2 实例：使用 MBR 估计纽约州 Tuxedo 镇的房租价格 …. 248
9.2 MBR 面临的挑战……250
9.2.1 选择一个平衡的历史记录集…… 250
9.2.2 训练数据表示…… 250
9.2.3 确定距离函数、组合函数和邻居数…… 253
9.3 案例研究：使用 MBR 分类乳房 X 线照片异常 ……253
9.3.1 业务问题：识别 X 射线异常…… 253
9.3.2 使用 MBR 应对这一问题 … 253
9.3.3 总体解决方案…… 255
9.4 距离和相似度计算 ……255
9.4.1 距离函数是什么…… 256
9.4.2 “一次一个字段”地建立距离函数 …… 257
9.4.3 其他数据类型的距离函数 … 259
9.4.4 当存在一个距离度量指标时…… 260
9.5 组合函数：向邻居征求建议 …260
9.5.1 最简单的方法：一个邻居 …. 260
9.5.2 针对类别目标的基本方法：民主…… 261
9.5.3 针对类别目标的加权投票 … 262
9.5.4 数值目标…… 262
9.6 案例研究：Shazam——发现音频文件的最近邻居 ……263
9.6.1 为何这一技能存在挑战 …… 264

9.6.2 音频签名…… 264
9.6.3 相似度计算…… 265
9.7 协同过滤：一种用于推荐的最近邻方法 ……267
9.7.1 构建个人信息…… 268
9.7.2 比较个人信息…… 268
9.7.3 预测…… 269
9.8 经验教训 ……270

第 10 章 了解何时应担忧：使用生存分析了解客户 ……271
10.1 客户生存 ……273
10.1.1 生存曲线揭示的含义…… 273
10.1.2 从生存曲线中寻找平均持续期…… 274
10.1.3 使用生存分析保留客户 …. 276
10.1.4 将生存视为衰变…… 277
10.2 风险概率 ……279
10.2.1 基本思想 …… 279
10.2.2 风险函数例子…… 280
10.2.3 删截 …… 282
10.2.4 风险计算 …… 283
10.2.5 其他类型的删截…… 284
10.3 从风险到生存 ……285
10.3.1 保留 …… 285
10.3.2 生存 …… 286
10.3.3 比较保留和生存…… 287
10.4 比例风险 ……288
10.4.1 比例风险的示例…… 288
10.4.2 分层：度量生存的初始影响…… 289
10.4.3 Cox 比例风险…… 290
10.5 生存分析实践 ……292
10.5.1 处理不同的客户流失类型 …… 292
10.5.2 客户何时还会返回…… 293
10.5.3 理解客户价值…… 295
10.5.4 预测 …… 297
10.5.5 风险随时间变化…… 298
10.6 经验教训 ……299

第 11 章 遗传算法与群体智能 ……301
11.1 优化……302
11.1.1 优化问题是什么…… 302
11.1.2 蚁群世界的优化问题…… 302
11.1.3 合众为一(E Pluribus Unum)…… 303
11.1.4 聪明的蚂蚁…… 304
11.2 遗传算法……306
11.2.1 一点历史…… 306
11.2.2 计算机中的遗传学…… 306
11.2.3 基因组的表示…… 312
11.2.4 模式：遗传算法的构造模块…… 313
11.2.5 超越简单算法…… 315
11.3 旅行商问题……316
11.3.1 穷举搜索…… 316
11.3.2 简单的贪婪算法…… 317
11.3.3 遗传算法的方法…… 317
11.3.4 群体智慧的方法…… 317
11.4 案例研究：使用遗传算法优化资源 ……319
11.5 案例研究：进化出分类投诉的解 ……320
11.5.1 业务上下文…… 320
11.5.2 数据…… 321
11.5.3 评论签名…… 321
11.5.4 基因组…… 322
11.5.5 适应度函数…… 323
11.5.6 结果…… 323
11.6 经验教训……323

第 12 章 一些新知识：模式识别与数据挖掘 ……325
12.1 无指导技术和无指导数据挖掘 ……326
12.1.1 无指导技术与有指导技术的对比…… 326

12.1.2 无指导数据挖掘与有指导数据挖掘的对比…… 327
12.1.3 案例研究：使用有指导技术的无指导数据挖掘… 327
12.2 什么是无指导数据挖掘 ……329
12.2.1 数据探索 …… 329
12.2.2 划分和聚类 …… 330
12.2.3 当目标不明确时目标变量的定义…… 332
12.2.4 模拟、预测和基于智能体的建模…… 335
12.3 无指导数据挖掘的方法论 …344
12.3.1 不存在方法论…… 345
12.3.2 需要谨记的事情…… 345
12.4 经验教训 ……345

第 13 章 发现相似的岛屿：自动群集检测……347
13.1 搜索简化的岛屿 ……348
13.2 客户细分和聚类 ……349
13.2.1 相似性聚类 …… 350
13.2.2 基于群集划分的跟踪活动 …… 351
13.2.3 聚类揭示被忽视的细分市场 …… 352
13.2.4 适应军队需求…… 353
13.3 K-均值聚类算法 ……353
13.3.1 K-均值算法的两个步骤 .. 354
13.3.2 Voronoi 图和 K-均值群集 …… 355
13.3.3 选择群集种子点…… 357
13.3.4 选择 K 值…… 357
13.3.5 使用 K-均值检测离群点…… 358
13.3.6 半指导聚类 …… 359
13.4 解释群集 ……359
13.4.1 使用质心表征群集…… 359
13.4.2 使用群集之间的差异表征群集…… 360
13.4.3 使用决策树描述群集…… 361
13.5 评价聚类 ……362
13.5.1 群集的度量和术语…… 362
13.5.2 群集轮廓…… 363
13.5.3 为打分限制群集直径…… 365
13.6 案例研究：城镇聚类……366
13.6.1 创建城镇签名…… 366
13.6.2 创建群集…… 367
13.6.3 确定合适的群集数目…… 367
13.6.4 评价群集…… 368
13.6.5 使用人口统计学群集调整区域边界…… 370
13.6.6 商业成功…… 370
13.7 K-均值算法的变种算法 ……371
13.7.1 K-中位数、K-中心点和 K-众数 …… 371
13.7.2 K-均值的软层面…… 374
13.8 聚类的数据准备 ……375
13.8.1 一致性缩放…… 375
13.8.2 使用权重编码外部信息… 375
13.8.3 选择聚类变量…… 376
13.9 经验教训 ……376

第 14 章 其他的群集检测方法……379
14.1 K-均值聚类的缺点 ……379
14.1.1 合理性…… 380
14.1.2 一个直观的例子…… 380
14.1.3 通过改变度量范围来修正问题…… 382
14.1.4 这在实际中意味着什么… 383
14.2 混合高斯模型 ……383
14.2.1 把高斯过程引入 K-均值聚类…… 384
14.2.2 回到混合高斯模型…… 386
14.2.3 混合高斯模型的打分…… 388
14.2.4 混合高斯模型的应用…… 388
14.3 分裂聚类 ……389
14.3.1 一种类决策树的聚类算法…… 390

14.3.2 分裂聚类的打分…… 391
14.3.3 群集和树 …… 391
14.4 凝聚(层次化)聚类……392
14.4.1 凝聚聚类方法的综述…… 392
14.4.2 凝聚聚类算法…… 395
14.4.3 为凝聚群集打分…… 397
14.4.4 凝聚聚类的局限性…… 398
14.4.5 凝聚聚类的实际应用…… 399
14.5 自组织映射 ……400
14.5.1 什么是自组织映射…… 401
14.5.2 SOM 的训练…… 403
14.5.3 SOM 的打分…… 404
14.6 继续搜索简化的岛屿 ……404
14.7 经验教训 ……405
第 15 章 购物篮分析和关联规则 ……407
15.1 购物篮分析的定义 ……408
15.1.1 购物篮数据的四个级别 … 408
15.1.2 购物篮分析的基础：基本度量…… 409
15.1.3 订单特征 …… 410
15.1.4 项目(产品)人气……411
15.1.5 跟踪市场干预…… 412
15.2 案例研究：西班牙语或英语 ……413
15.2.1 业务问题 …… 413
15.2.2 数据 …… 414
15.2.3 “西班牙裔城市”偏好的定义…… 414
15.2.4 解决方案 …… 415
15.3 关联分析 ……416
15.3.1 规则不是万能的…… 416
15.3.2 关联规则中的项目集…… 418
15.3.3 关联规则的益处…… 420
15.4 构建关联规则 ……421
15.4.1 选择正确的项目集…… 422
15.4.2 从所有这些数据中生成规则 …… 426
15.4.3 克服实际限制…… 429
15.4.4 大数据问题…… 432
15.5 思想扩展 ……432
15.5.1 左右两侧包含不同的项目…… 432
15.5.2 利用关联规则比较商店… 433
15.6 关联规则和交叉销售……434
15.6.1 一个经典的交叉销售模型…… 435
15.6.2 更可信的倾向度产生方法…… 435
15.6.3 使用置信度所产生的结果…… 436
15.7 序列模式分析 ……436
15.7.1 序列的发现…… 436
15.7.2 序列关联规则…… 439
15.7.3 利用其他数据挖掘技术的序列分析…… 440
15.8 经验教训 ……440
第 16 章 链接分析 ……443
16.1 图论基础 ……444
16.1.1 图是什么…… 444
16.1.2 有向图…… 445
16.1.3 加权图…… 446
16.1.4 哥尼斯堡的七桥问题…… 447
16.1.5 图中的回路检测…… 449
16.1.6 旅行商问题的反思…… 449
16.2 社交网络分析 ……452
16.2.1 六度分割理论…… 453
16.2.2 你朋友说了关于你的什么事情…… 454
16.2.3 发现托儿福利欺诈…… 454
16.2.4 交友网站中谁响应了谁… 455
16.2.5 社会营销…… 456
16.3 呼叫图挖掘 ……456
16.4 案例研究：追踪领袖……458
16.4.1 业务目标…… 458
16.4.2 数据处理面临的挑战…… 459

16.4.3 发现呼叫数据中的社交网络 …… 459
16.4.4 这些结果如何用于营销 … 460
16.4.5 估计客户年龄 …… 460
16.5 案例研究：谁正在家里使用传真机 …… 460
16.5.1 寻找传真机为何有用 …… 461
16.5.2 传真机的行为如何 …… 461
16.5.3 图着色算法 …… 462
16.5.4 对图进行着色以识别传真机 …… 462
16.6 Google 如何成为世界的统治者 …… 463
16.6.1 中心和权威 …… 464
16.6.2 算法细节 …… 465
16.6.3 实践中的中心和权威 …… 466
16.7 经验教训 …… 466

第 17 章 数据仓库、OLAP、分析沙箱和数据挖掘 …… 469
17.1 数据体系结构 …… 470
17.1.1 事务数据：基础层 …… 471
17.1.2 操作汇总数据 …… 472
17.1.3 决策支持汇总数据 …… 472
17.1.4 数据库模式/数据模型 …… 473
17.1.5 元数据 …… 476
17.1.6 业务规则 …… 476
17.2 数据仓库的通用体系结构 …… 477
17.2.1 源系统 …… 477
17.2.2 提取、转换和加载 …… 479
17.2.3 中央存储库 …… 479
17.2.4 元数据存储库 …… 481
17.2.5 数据集市 …… 482
17.2.6 操作反馈 …… 482
17.2.7 用户和桌面工具 …… 482
17.3 分析沙箱 …… 484
17.3.1 为什么需要分析沙箱 …… 484
17.3.2 支持分析沙箱的技术 …… 486
17.4 OLAP 的适用时机 …… 488
17.4.1 立方体中是什么 …… 490
17.4.2 星型模式 …… 494
17.4.3 OLAP 和数据挖掘 …… 495
17.5 数据挖掘与数据仓库如何匹配 …… 496
17.5.1 大量的数据 …… 497
17.5.2 一致的、干净的数据 …… 497
17.5.3 假设检验和度量 …… 498
17.5.4 可扩展的硬件和 RDBMS 支持 …… 498
17.6 经验教训 …… 499

第 18 章 构建客户签名 …… 501
18.1 在数据中寻找客户 …… 502
18.1.1 客户是什么 …… 502
18.1.2 账户、客户与家庭 …… 503
18.1.3 匿名事务 …… 503
18.1.4 链接到卡的事务 …… 503
18.1.5 链接到 cookie 的事务 …… 504
18.1.6 链接到账户的事务 …… 504
18.1.7 链接到客户的事务 …… 505
18.2 设计签名 …… 505
18.2.1 客户签名是否有必要 …… 509
18.2.2 每一行代表什么 …… 509
18.2.3 签名对预测建模有用吗 … 512
18.2.4 目标已经被定义了吗 …… 513
18.2.5 是否应用了由特定的数据挖掘技术所强加的约束 … 513
18.2.6 将会引入哪些客户 …… 513
18.2.7 可能想了解客户的哪些情况 …… 514
18.3 签名看起来像什么 …… 514
18.4 创建签名的过程 …… 517
18.4.1 有些数据已处于正确的粒度 …… 517
18.4.2 旋转到规则时间序列 …… 517
18.4.3 聚集时间戳事务 …… 519
18.5 处理缺失值 …… 520
18.5.1 源数据中的缺失值 …… 520

18.5.2 未知或不存在…… 521
18.5.3 什么不该做 …… 521
18.5.4 需要考虑的事情…… 523
18.6 经验教训 ……524

第 19 章 派生变量：使数据的含义更丰富……527
19.1 基于手机流失率的流失预测 ……527
19.2 单变量转换 ……529
19.2.1 标准化数字变量…… 529
19.2.2 转换数值为百分位数…… 530
19.2.3 把数量转为比率…… 530
19.2.4 相对度量 …… 531
19.2.5 把类别变量替换为数值 … 532
19.3 变量组合 ……536
19.3.1 经典组合 …… 536
19.3.2 组合高度相关的变量…… 539
19.4 从时间序列中提取特征 ……545
19.4.1 趋势 …… 545
19.4.2 季节性 …… 546
19.5 从地理位置中提取特征 ……547
19.5.1 地理编码 …… 547
19.5.2 映射 …… 548
19.5.3 利用地理位置创建相对度量 …… 549
19.5.4 使用目标变量的历史值 … 549
19.6 使用模型分数作为输入 ……550
19.7 稀疏数据的处理 ……550
19.7.1 账户集模式 …… 550
19.7.2 分箱稀疏值 …… 551
19.8 从事务中捕获客户行为 ……551
19.8.1 拓宽窄数据 …… 552
19.8.2 影响范围作为良好客户的预测…… 552
19.8.3 示例：对评分者剖析的评分 …… 553
19.8.4 评分者签名中的样本字段 …… 553
19.8.5 评分签名与派生变量…… 555
19.9 经验教训 ……555

第 20 章 减少变量数量的技术……557
20.1 变量太多存在的问题……558
20.1.1 输入变量之间彼此相关的风险…… 558
20.1.2 过拟合风险…… 559
20.2 数据稀疏问题 ……560
20.2.1 稀疏性的可视化…… 560
20.2.2 独立性…… 561
20.2.3 穷举法特征选择…… 563
20.3 变量约简技术的类型……564
20.3.1 使用目标…… 564
20.3.2 原始变量与新变量…… 564
20.4 特征的顺序选择 ……565
20.4.1 传统的前向选择方法…… 565
20.4.2 使用验证集的前向选择… 566
20.4.3 逐步选择…… 567
20.4.4 使用非回归的前向选择技术…… 567
20.4.5 后向选择…… 567
20.4.6 无指导的前向选择…… 568
20.5 其他有指导的变量选择方法 ……568
20.5.1 利用决策树来选择变量… 568
20.5.2 使用神经网络来约简变量…… 571
20.6 主成分 ……571
20.6.1 主成分是什么…… 571
20.6.2 主成分分析的例子…… 575
20.6.3 主成分分析…… 578
20.6.4 因子分析…… 581
20.7 变量聚类 ……582
20.7.1 变量群集的例子…… 582
20.7.2 使用变量群集…… 583
20.7.3 层次变量聚类…… 583
20.7.4 分裂变量聚类…… 585
20.8 经验教训 ……586

第 21 章 仔细聆听客户所述：文本挖掘 ……587
21.1 什么是文本挖掘 ……588
21.1.1 文本挖掘用于派生列…… 588
21.1.2 派生特征之外…… 588
21.1.3 文本分析应用…… 589
21.2 处理文本数据 ……591
21.2.1 文本源 …… 591
21.2.2 语言影响 …… 592
21.2.3 表示文档的基本方法…… 593
21.2.4 实践中的文档表示…… 594
21.2.5 文档和语料库…… 595
21.3 案例研究：特设文本挖掘 ….595
21.3.1 抵制行动 …… 596
21.3.2 照常营业 …… 596
21.3.3 结合文本挖掘和假设检验 …… 596
21.3.4 结果 …… 597
21.4 使用 MBR 分类新闻报道……598
21.4.1 什么是编码 …… 598
21.4.2 应用 MBR…… 599
21.4.3 结果…… 601
21.5 从文本到数字 ……601
21.5.1 以“词袋”开始…… 602
21.5.2 词-文档矩阵 …… 603
21.5.3 语料库影响…… 604
21.5.4 奇异值分解(SVD) …… 604
21.6 文本挖掘和朴素贝叶斯模型 ……606
21.6.1 文本世界中的朴素贝叶斯…… 607
21.6.2 使用朴素贝叶斯识别垃圾邮件…… 607
21.6.3 情感分析…… 611
21.7 DIRECTV：客户服务案例研究 ……613
21.7.1 背景…… 613
21.7.2 应用文本挖掘…… 614
21.7.3 采取技术手段…… 616
21.7.4 持续受益…… 619
21.8 经验教训 ……620

第 1 章

什么是数据挖掘以及为什么要进行数据挖掘

本书第 1 版第 1 章的开场白如下:“马萨诸塞州萨默维尔市,本书作者之一的家乡.....”,接着分别介绍了那座小镇上的两家小商店，以及它们如何形成与客户的学习关系(learning relationship)。其中的一家商店——头发编织店，已经不再给那个小女孩编头发。第 1 版之后的这些年里，那个小女孩已经长大成人并搬离了小镇，而且她也不再梳着小辫了。而她的父亲作为作者之一，也搬到了附近的剑桥大学。但是，有一件事情并未改变。作者依然是 Wine Cask 商店的忠实顾客,店中的某些人在 1978 年首次向他介绍了便宜的阿尔及利亚红酒(Algerian reds)，其后是法国葡萄种植区，现在同样是这批人正在帮助他探讨意大利和德国的葡萄酒。

数十载之后，Wine Cask 依然还有一位忠实的客户。这种忠诚并非偶然。工作人员了解他们客户的口味以及他们能够接受的价格范围。当客户咨询时，他们不仅根据库存信息，而且还根据累积的客户口味和预算信息进行回答。

Wine Cask 的工作人员掌握了许多有关葡萄酒的知识。虽然这些知识是客户选择他们而不选择有大折扣的酒铺的原因之一，但是他们对每一位客户的深入了解才是获得回头客的主要原因。尽管可以在街道对面另开一家酒铺，同时雇佣一批品酒专家，但是要想获得有关客户的同样详尽知识，则需要耗费他们数月或数年时间。

经营有方的小商店会自然地形成与客户之间的学习关系。随着时间的推移，他们对客户的了解会越来越多，从而可以利用这些知识为他们提供更好的服务。结果则皆大欢喜，忠实的顾客和盈利的商店。

不过，拥有数十万或数百万客户的大公司，则不能奢望与每个客户形成密切的私人关系。因此大公司必须依赖于其他方式来形成与客户的学习关系。特别是，他们必须学会充分利用所拥有的大量信息——几乎是每次与客户交互所产生的数据。本书将要介绍的就是可用于将客户数据转换成客户知识的分析技术。

1.1 什么是数据挖掘

虽然有些数据挖掘技术非常新颖，但是数据挖掘本身并非一项新的技术：自从第一台计算机发明以来，人们就一直在计算机上分析数据——而且在此之前的数个世纪里，人们一直在没有计算机的情况下分析数据。多年来，数据挖掘有许多不同的名称，诸如知识发现(knowledge discovery)、商业智能(business intelligence)、预测建模(predictive modeling)以及预测分析(predictive analytics)，等等。本书作者所使用的数据挖掘定义是：

数据挖掘是一项探测大量数据以发现有意义的模式(pattern)和规则(rule)的业务流程(business process)。

这个定义包含了好几个组成部分，它们都非常重要。

1.1.1 数据挖掘是一项业务流程

数据挖掘是一项与其他业务流程交互的业务流程。特别是，它是一项没有开始和结束的流程：它一直在进行。数据挖掘以数据作为开始，然后通过分析来启动或激励行动，这些行动反过来又将创建更多需要数据挖掘的数据。

因此，对于那些希望充分利用它们的数据来改善业务的公司而言，不应仅仅把数据挖掘看作是细枝末节。相反，它们的业务策略必须包括收集数据、为长期利益分析数据，并针对分析结果做出行动。

同时，数据挖掘很容易适应为了解市场和客户的其他策略。市场调研、客户小组(customer panel)及其他技术，与数据挖掘和更为广泛的数据分析是兼容的。关键在于要认识到重心是客户，以及不同企业数据之间的共性。

1.1.2 大量的数据

本书作者之一在演讲时经常会询问听众：“多大才是大量的数据？”学生们会给出诸如“处理 1000 万客户的所有事务”或“TB 级的数据”之类的答案。他本人的回答则更为温和——“65 356 行”，尽管自从 2007 年之后微软已经在 Excel 电子表格中允许超过 100 万行的数据，但是这个回答依然获得了理解。

诸如 Excel 之类的工具，对于处理相对较小的数据量具有难以置信的通用性。它允许对每一行或列的值进行多种计算；而透视表(pivot table)对于理解数据和趋势也是惊人的实用；同时图表还提供了强大的数据可视化机制。

在数据挖掘初期(20 世纪 60 年代和 20 世纪 70 年代)，数据非常稀少。本书所描述的一些技术是在只包含几百个记录的数据集上发展而来的。那时，一个典型的数据集可能是关于蘑菇的一些属性，以及它们是否有毒或可食用。或者，可能是关于车的一些属性，而目标则是估计燃气里程。无论哪种特定的数据集，它们都证明了当时所开发的技术的能力；对于不再适合于电子表格的数据，这些技术依然适用。

由于计算能力易于获得，因此大量的数据不再是一个障碍，而是一个优势。本书所介

绍的许多技术在大数据集上比在小数据集上表现得更好——你可以用数据来替代智慧。换句话说，数据挖掘让计算机完成其最擅长的工作——从许许多多的数据中挖掘。反过来，这将让人们完成人类最擅长的工作：提出问题并理解结果。

另一方面，在本书的某些案例研究中仍然使用相对较小的数据规模。其中最小的也许是第 13 章聚类案例研究中所使用的数据。该案例研究在新英格兰(New England)的几百个城镇中发现人口学上类似的城镇。即使是强大的 Excel 也不具有实现类似“组合类似的城镇”这样的内置函数。

这就是数据挖掘的切入点。无论目标是发现类似的新英格兰城镇组，还是寻找客户减少的原因，或是整本书中无数的其他目标，数据挖掘技术都可以利用更为简单的桌面工具不再适用的数据。

1.1.3　有意义的模式和规则

在数据挖掘定义中最重要的也许是关于“有意义的模式”这一部分。虽然数据挖掘很有趣，但是帮助提高业务比让挖掘者觉得有趣更为重要。

在许多方面，发现数据的模式并不是非常困难。业务操作有时会生成数据，与此同时必然产生模式。然而，数据挖掘的目标——至少当作者使用该术语时——不是要找到数据的任何模式，而是要发现对业务有益的模式。

这意味着需要发现帮助日常业务操作的模式。考虑一个呼叫中心的应用程序，其目标是指定客户的颜色。“绿色”意味着非常好，因为呼叫者是一个有价值的客户，值得付出使其保持快乐；“黄色”则是给出警告，因为该客户也许是有价值的，但是存在一些风险迹象；而“红色”意味着不给客户任何特殊的待遇，因为该客户非常危险。发现模式也可能意味着把保留活动(retention campaign)的目标定位为最有可能流失的客户。这意味着优化客户获取(customer acquisition)，既考虑客户数量上的短期收益，同时也考虑客户价值的中期和长期收益。

逐渐地，公司将围绕数据挖掘发展业务模型——尽管他们可能不使用这个词。本书作者所供职的一家公司帮助零售商在 Web 上推荐商品；这家公司仅当 Web 购物者单击其推荐商品时获到报酬。这只是一个例子。有些公司聚合来自不同来源的数据，将数据汇合在一起以获得客户的更完整的形象。一些公司，如 LinkedIn，使用某些人提供的信息向其他人提供优质服务——当招聘负责人为其开放的工作岗位找到合适的候选人时，每个人都将受益。在所有这些情况下，目标都是向最有可能需要他们的人提供直接的产品和服务，使得每个人参与的购买和销售过程更为有效。

1.1.4　数据挖掘和客户关系管理

本书不是介绍通用的数据挖掘技术，而是特定于客户关系管理系统(customer relationship management)的数据挖掘。各种规模的公司都必须学会模仿小规模的、面向服务的公司所一直擅长的工作——与其客户创建一对一的关系。客户关系管理是许多文章、书籍以及会议所讨论的一个广泛主题。从引导-追踪(lead-tracking)软件，到活动管理软件(campaign management software)，再到呼叫中心软件等，都被标记为客户关系管理工具。

本书的重点则更窄一些——数据挖掘在通过提高公司与其客户形成学习关系的能力，从而改善客户关系管理时可以发挥的作用。

在各行各业中，高瞻远瞩的公司的目标都是理解每个客户，并通过利用这种理解，使得客户与他们(而不是与竞争对手)做生意更加容易，也更加有利可图。这些公司同样都学习分析每个客户的价值，从而清楚哪些客户值得投资和努力来保留，哪些则允许流失。重心从广泛的市场领域到个人客户的转变要求整个企业的变化，尤其在营销、销售和客户支持等方面更是如此。

围绕客户关系构建业务对大多数公司而言是一个革命性的变化。传统上，银行的重心是维持存钱和借钱的利差；电话公司聚焦于通过网络连接电话；而保险公司则集中在处理索赔要求、管理投资以及维持其赔付率(loss ratio)。把一个以产品为中心的企业转变成以客户为中心的企业的代价超过了数据挖掘。假设数据挖掘的结果建议为一个特定客户提供一个小装饰(widget)而不是一个小发明(gizmo)，但是如果经理的奖金取决于小发明本季度销售的数量而不是小装饰的数量(即便后者更为有利可图或者收获长期盈利更多的客户)，那么数据挖掘的结果也将会被忽略。

从狭义上讲，数据挖掘是工具和技术的集合。它是用来支持以客户为中心的企业的几种必要技术之一。从广义上讲，数据挖掘是一种态度，即业务行动应该基于学习、知情的决定比不知情的决定要好，以及度量结果对业务有益等。数据挖掘也是一个应用分析工具和技术的过程和方法论(methodology)。若要使数据挖掘发挥作用，则对分析型客户关系管理(CRM)的其他需求也必须到位。公司为了与其客户形成学习关系，必须能够：

- 注意客户正在做的事情。
- 记录它和客户随着时间推移所做的事情。
- 从记录的内容中学习。
- 根据学到的内容采取行动，从而使客户的收益更高。

虽然本书的重点是第 3 项——从以往发生的事情中学习，但是学习不能在真空中进行。必须存在事务处理系统捕获客户交互，数据仓库存储历史的客户行为信息，数据挖掘把历史数据转换成未来的行动计划，以及一个客户关系战略把这些计划付诸实施。

数据挖掘，重复一下之前的定义，是一项探测大量数据以发现有意义的模式和规则的业务流程。本书假设数据挖掘的目标是，使得一家公司能够通过更好地理解客户来改进其市场、销售和客户支持操作。然而，需要记住的是，本书所描述的数据挖掘技术和工具同样也适用于不同的领域，诸如执法、射电天文学、医学以及工业过程控制等。

1.2 为什么是现在

大多数的数据挖掘技术，至少作为学术算法，已经存在了数十年(最古老的生存分析实际上可追溯到数世纪之前)。数据挖掘已经形成一个大的流行趋势，自从 20 世纪 90 年代以来得到了急剧增长。这是由几个方面的因素所形成的：

- 数据正在产生。
- 数据正存储于数据仓库中。

- 计算能力能够承受。
- 对客户关系管理的兴趣非常强烈。
- 商业的数据挖掘软件产品已经形成。

这些因素的结合意味着数据挖掘越来越成为业务策略的基础。Google 不是第一个搜索引擎，但是它是第一个将复杂的搜索算法与基于点击收益(click-through revenue)最大化的业务模型相结合的搜索引擎。几乎在每个业务领域，公司都会发现它们拥有许多信息——订阅者、Web 访问者、托运人的信息以及支付模式、呼叫模式、朋友和邻居等信息。公司越来越多地转向数据分析以利用他们的信息。

1.2.1　数据正在产生

数据挖掘在大数据量可用时将体现其最大的价值。事实上，大多数数据挖掘算法在构建和训练模型时往往需要较大的数据量。

本书的潜在主题之一就是到处存在大量可用的数据。这对于拥有客户的公司(以及客户是其全部的公司)而言尤为正确。一个人浏览网站时会在一天内产生几十 KB 的数据。将其乘以百万的客户和潜在客户，数据量将快速超过单个电子表格的大小。

Web 不是产生大量数据的唯一制造者。电话公司和信用卡公司是第一批使用 TB 级大小的数据库的公司，在 20 世纪 90 年代末这是一个异常大的数据库规模。那个时代已经过去了。虽然有可用的、规模巨大的数据，但是如何使之产生价值呢？

1.2.2　数据正存在于数据仓库中

公司不仅正在产生大量的数据，而且其往往越来越多地从产生它们的业务账单、预约、索赔过程以及订单输入系统中被提取出来，然后输入到数据仓库中成为企业内存的一部分。

数据仓库是数据挖掘故事中如此重要的一部分，因此第 17 章将专门讨论这个话题。数据仓库以一个共同的格式汇集许多不同来源的数据，该格式具有一致的关键字和字段定义。业务系统旨在快速地向终端用户提供结果，他们可能是网站的客户或者是正在工作的员工。这些系统旨在完成手头的任务，而不是为了保持干净、一致的数据以进行分析。而另一方面，数据仓库是为决策支持而专门设计的，它将简化数据挖掘者的工作。

1.2.3　计算能力能够承受

数据挖掘算法通常需要多次遍历规模巨大的数据，同时许多算法也是计算密集型的。磁盘、内存、处理能力和网络带宽等价格的持续急剧下降，已把曾经昂贵的、仅在几个政府资助的实验室中使用的技术推向一般业务。

1.2.4　对客户关系管理的兴趣非常强烈

在各行各业中，公司已经认识到他们的客户是业务的中心，而客户信息则是他们的关键资产。

1. 每个业务都是服务业务

处于服务行业的公司，信息将赋予其竞争优势。这就是为什么连锁饭店会记录你首选无烟的房间，而租车公司会记录你喜欢的车的类型。此外，传统上认为自身不是服务提供者的公司也开始从不同的角度来思考。汽车经销商是出售汽车还是运输工具？如果是后者，那么每当你自己的车在商店里时，经销商就为你提供一辆替代车是合理的，许多经销商现在就是这么做的。

即使是日用商品也可以通过服务得到加强。一家家庭供热石油公司如果能够监视你的使用情况，并在你需要更多的石油时向你提供石油，那么相比一家公司期望你在油箱枯竭和管道冻结前记得打电话来安排你的订单，它销售的产品更好。对于信用卡公司、长途运输公司、航空公司以及所有类型的零售商而言，服务竞争通常会与价格竞争一样多或更多。

2. 信息即产品

许多公司发现他们拥有的客户信息不仅对自己有价值，而且对其他人同样有价值。一家具有忠诚卡方案的超市有一些消费者包装食品行业会喜欢的信息——关于谁在购买哪些产品的知识。信用卡公司有一些航空公司想要了解的信息——谁在买大量的机票。超市和信用卡公司都处于知识经纪人的位置。超市可以通过打印优惠券向消费者包装食品公司索取更高的收费，此时超市会承诺通过向适当的购物者打印适当的优惠券获得更高的回报率。信用卡公司可以向航空公司收费，其目标是为经常旅行、但乘坐其他航空公司航班的人提供频繁的飞行积分。

Google 了解人们正在 Web 上寻找什么。它在出售赞助商链接(以及其他事物)时利用这种知识。保险公司会为确保某人在搜索“汽车保险”时，为其提供它们站点的链接而支付相应的费用。金融企业将支付赞助商链接，从而当有人搜索诸如“抵押贷款再融资”之类的短语时显示其链接。

事实上，任何收集了宝贵数据的公司都能够成为一个信息代理。Cedar Rapids Gazette 利用其在爱荷华州东部 22 个县中的优势地位，为本地企业提供直接的市场服务。报纸使用它自身的讣告页面和结婚通告来保持其营销数据库的流通。

1.2.5 商业的数据挖掘软件产品变得可用

新的算法首先出现在学术刊物上并且在会议中引发讨论的时间，与在商业软件中纳入这些算法从而变得可用的时间之间总会存在时间差。在首批产品初始可用的时间以及它们获得广泛接受的时间之间，也存在另一个时间差。对于数据挖掘而言，具备广泛可用性和被广泛接受的时期已经来临。

本书所讨论的许多技术都是起源于统计、人工智能或机器学习等领域。在大学和政府实验室中经过数年之后，一项新的技术就会被商业部门中一些早期的采用者开始启用。在新技术演化之时，软件通常会向勇敢的用户提供源代码，他们愿意通过 FTP 来获取它、对它进行编译，并通过阅读作者的博士论文来了解如何使用它。只有在几个先驱成功地应用一项新技术之后，它才开始在真正的产品中出现，包括用户手册、帮助热线以及培训课程等。

现在，新技术正在开发之中；然而，其中许多工作也是用于扩展和改进现有的技术。本书讨论的所有技术在商业和开放源代码软件产品中均存在，尽管没有一款单一的产品把它们全部囊括。

1.3　数据挖掘人员的技能

谁能成为一名数据挖掘人员(data miner)？答案不是每个人，因为这需要一些特殊的技能。一个好的数据挖掘人员需要有数字技能，并且对统计有一些基本的了解(更强的统计知识总会有用)。第 4 章和第 6 章介绍了数据挖掘所需的许多关键统计概念。对 Excel 有很好的了解也非常有用，因为它是商业世界中主要的电子表格。诸如 Excel 之类的电子表格对分析少量的数据并且向广大读者展示结果非常有用。

当然，熟悉数据挖掘技术是成为数据挖掘人员的关键。本书的大部分内容是专门讨论各种技术。了解技术本身很重要；更重要的是了解它们何时有用以及如何有用。与技术细节同样重要的或许是数据挖掘技术的启蒙。尽管许多技术都很复杂，但是它们往往都是基于一个非常容易理解的基础。这些技术并非魔术。即便没有在数学或统计方面的博士学位，你不能准确地解释它们如何得到答案，但是理解它们还是可能的。这些技术比魔术更好，因为它们有用而且可以帮助解决实际问题。

对一名数据挖掘人员而言，另一项非常重要的技能实际上是一种态度：不畏惧为了得到结果可能需要处理的大数据量和复杂的过程。处理大型数据集、数据仓库以及分析沙箱是数据挖掘成功的关键。

最后，数据挖掘不仅仅产生技术结果。例如，除了在计算机中移位之外，数据挖掘模型不会真正做任何其他事情。结果必须用来帮助人们(或者帮助越来越自动化的流程)做出更明智的决定。产生技术结果只是数据挖掘过程第一步的结束。能够与其他人一起工作、交流结果，并认识到真正的需求是作为一名好的数据挖掘人员至关重要的技能。贯穿本书的是许多商业上下文中的数据挖掘实例，它们分布在下两章和专门讨论每项技术的章节中。数据挖掘是一个基于数据的学习过程，如下一节所描述的，一位好的数据挖掘人员必须对新思想持开放态度。

1.4　数据挖掘的良性循环

在 19 世纪初期，纺织厂是工业成功故事。这些纺织厂在英格兰和新英格兰境内的河流沿岸不断增长的村庄和城市中涌现，以便充分地利用水电。水在水轮上运转，从而驱动纺纱、针织和编织机器。一个世纪以来，工业革命的象征是水倒在为纺织机器提供动力的轮子上。

商业世界已经发生了改变。老式的磨房小镇现在已成为独特的历史文物。河流沿岸长长的工厂建筑现在是仓库、购物中心、艺术家工作室以及其他各式各样的商店。甚至制造公司在服务上产生的价值通常也比在商品上更多。作者曾被领先的国际水泥制造商——Cemex——的一次广告宣传活动所吸引，它把混凝土作为服务来展示。该广告的重心不在

于水泥质量、价格或可用性，而是在一条河流上画了一座桥，并出售“水泥”是一项服务的想法，该服务连接着桥梁之间的人。混凝土作为服务？欢迎来到 21 世纪。

世界已经改变。获得电气和机械的力量不再是商业成功的标准。对于大众市场的产品而言，客户交互的数据是新的水力；知识驱动着服务经济以及许多制造业经济的涡轮，因为服务与制造之间的界限越来越模糊。来自数据的信息使得能够通过以客户为目标集中销售和营销方面的努力，通过解决客户的实际需要改进产品设计，通过了解和预测客户喜好增强资源分配。

数据是许多核心业务流程的心脏。它是由业务系统中的事务所产生，无论何种工业——例如零售、电信、制造、保健、实用工具、交通运输、保险、信用卡和金融服务等。大量的内部数据拥有许多外部来源，包括零售客户的人口、生活方式以及信用信息；业务客户的信贷、金融和市场信息；以及各种规模社区的人口统计信息。数据挖掘承诺找到隐藏在磁盘或计算机内存中所有这些数十亿或数万亿比特数据中的有趣模式。仅仅发现模式是不够的。你必须响应这些模式，针对它们执行相应的动作，最终把数据转化为信息，信息转化为行动，行动转化为价值。简而言之，这就是数据挖掘的良性循环。

为了实现这一承诺，需要把数据挖掘变成一个基本的业务流程，以纳入到其他的流程，包括营销、销售、客户支持、产品设计和存货控制等。良性循环把数据挖掘放在更大的业务范围内，重心从发现机制转移到基于发现的行动上。本书强调数据挖掘的可操作结果(此处所指的“可操作(actionable)”不应与它在法律领域的定义所混淆，在法律领域中，它意味着某项行动具有合法行动的基础)。

营销方面的文献使数据挖掘看上去很容易。仅仅应用由学术界成熟理论创建的自动算法，如神经网络、决策树与遗传算法，你就会收获极大的成功。虽然算法很重要，但是数据挖掘解决方案不仅仅只是一组强大的技术和数据结构。这些技术必须应用于正确的问题，针对正确的数据。数据挖掘的良性循环是一个迭代的学习过程，其构建在随时间推移产生的结果之上。成功使用数据将把企业从反应类型(reactive)转化为主动类型(proactive)。这就是数据挖掘的良性循环，作者使用它从本书稍后所描述的技术中获取最大收益。在解释数据挖掘的良性循环之前，先看一个数据挖掘在实践中的案例研究。

1.5 业务数据挖掘的案例研究

从前，有一家银行存在业务问题：一项特定的业务——房屋净值信贷额度，未能吸引到足够好的客户。银行可以选择几种不同的方式来解决这个问题。

例如，该银行可以降低房屋净值贷款的利率。这将招揽更多的顾客，并在利润率较低的费用下增加市场份额。现有的客户可能会转移到更低的利率，从而进一步压缩利润。更糟糕的是，假设初始利率具有合理的竞争力，降低利率可能会带来最坏的客户——不忠诚。竞争对手可以以稍微好一点的条件轻松地引诱他们离开。补充内容“赚钱或亏损”讨论了保持忠实客户的问题。

赚钱或亏损

房屋净值贷款会通过支付贷款的利息为银行创造收益，但有时公司需要应付快速亏损服务。

举一个例子，富达投资(Fidelity Investments)曾经一度停止了其账单支付服务，因为该项服务一直在亏损。但是最后的分析使之得以保存，因为分析发现 Fidelity 最忠实和最有利可图的客户会使用该项服务。虽然该项服务亏损，但是 Fidelity 在这些客户的其他账户上赚得更多的钱。毕竟，信任金融机构并愿意支付账单的客户对该机构的信任级别也非常高。削减这种增值服务可能会无意中导致最好的客户到别处寻求更好的服务，从而无意中加剧收益率问题。

即使是如房屋净值贷款之类的产品也给某些银行出了一个难题。拥有自己的房子并且有一大笔信用卡债务的客户，往往是适合房屋净值信用额度的好的候选者。这对客户很合适，因为贷款额度通常比原来的信用卡具有更低的利率。银行应该鼓励客户把他们的债务从信用卡转换为家庭净值贷款吗？

答案似乎比看上去更为复杂。在短期内，这样的转换对客户而言很合适，但对银行而言很糟糕：客户支付更少的利息意味着银行获得更少的收益。在银行内部，这样的转换也会引起问题。信用卡组可能已经很努力地工作来获得会每个月支付利息的客户。这个组并不想失去自己的优质客户。

另一方面而言，转换用户可能会构建一个终生的关系，包括许多汽车贷款、抵押贷款与投资产品等。当重点是客户时，长期的观点有时更为重要，而且它可能会与短期目标冲突。

在这个特定实例中的银行是美国银行(Bank of America，BofA)，经过几次直接邮寄活动产生失望的结果之后，它非常渴望扩展其家庭净值贷款的投资组合。美国全国消费者资产组(National Consumer Assets Group，NCAG)决定使用数据挖掘来解决问题，因此引入数据挖掘的良性循环。(作者要感谢研究这个问题的 Lounette Dyer、Larry Flynn 和 Jerry Modes，以及允许我们使用美国银行案例研究材料的 Larry Scroggins)。

1.5.1　识别美国银行的业务挑战

为了向客户营销家庭净值贷款，美国银行需要更加努力。基于常识和商务顾问的信息，得出了以下的一些结论：

- 有上大学年龄孩子的人会想要利用家庭净值贷款来支付学费。
- 收入高但不稳定的人想要使用家庭净值来平滑他们收入中的波峰和波谷。

这些见解可能为真或者为假。不过，家庭净值贷款额度产品的营销行业反映了这种关于潜在客户的观点，正如电话营销所拟定的名单一样。这些见解导致了前面所提及的令人失望的结果。

1.5.2　应用数据挖掘

美国银行与 Hyperparallel(一个数据挖掘工具供应商，后来被 Yahoo!收购)的数据挖掘

顾问一道，利用一系列的数据挖掘技术来解决这个问题。它并不缺乏数据。多年以来，美国银行已经在一个大型关系数据库中存储了其数百万零售客户的数据，该数据库位于 Teradata 公司一个强大的并行计算机之上。对来自 42 个记录系统的数据进行清理、转换、对齐，然后输入企业数据仓库。通过该系统，美国银行可以看到每个客户与银行保持的所有关系。

这一历史数据库是真正的名副其实——一些记录甚至可追溯到 1914 年！最近的客户记录中有大约 250 个字段，其中包括人口统计字段，如收入、小孩数量和家庭类型，以及内部数据等。将这些客户属性合并为一个客户签名，然后使用 Hyperparallel 的数据挖掘工具对其进行分析。

决策树(Decision Tree，第 7 章讨论的一种技术)派生出的规则将现有的银行客户分类成可能或不太可能响应家庭净值贷款的客户。决策树，在数以万计获得以及未获得该产品的客户实例上进行训练，最终的学习规则将给出他们之间的区别。发现这些规则之后，结果模型将用来向每个待预测记录添加另一个属性。这个属性，作为“家庭净值贷款额度标志的预测”标志，由数据挖掘模型生成。

接下来，序列化的模式发现技术(例如第 15 章在介绍购物篮分析和序列化模式分析时所描述的技术)可用来确定客户何时最有可能想要一笔这种类型的贷款。这种分析的目标是要发现过去在成功的贷款之前经常出现的一系列事件。

最后，聚类技术(见第 13 章)用来自动地把具有相似属性的客户分组。此时，该工具会找到 14 个客户群集(cluster)，其中许多群集似乎并不令人特别感兴趣。不过，这 14 个群集中的其中一个群集具有两个有趣的属性：

- 该客户群集中 39%的人同时有企业账户和个人账户。
- 这个群集中超过 25%的客户被决策树分类为家庭净值贷款的可能响应者。

此结果提示好奇的数据挖掘人员，人们可能是使用家庭净值贷款来创业。

1.5.3 对结果采取行动

根据这个新的结论，家庭净值贷款额度的业务单位(NCAG)与零售银行业务部(Retail Banking Division)一起，执行银行在这种情况下该做的事情：他们提交与客户交谈的市场调研提案。美国银行将发布一项银行分支机构的调查，以找出基层实际正在发生的事情，这种发布每年会有四次。利用数据挖掘获得的知识，该银行需要把一个问题添加到该列表中：“贷款的收益是否用于启动公司？”数据挖掘研究的结果是针对内部调查的一个问题。

调查结果证实了由数据挖掘所引发的怀疑。最后，NCAG 修改了活动的广告语，从“利用房屋价值送你的孩子上大学”到更具煽动性的“现在房子是空的，使用你的净资产做你一直想做的事”。

顺便提一下，市场调研和数据挖掘常常用于类似的结果——更好地理解客户。虽然市场调研很强大，但是它也有一些缺点：

- 响应者可能不是整体人口的代表。也就是说，响应者集合可能带有偏向性，特别是会偏向于作为过去营销努力的目标的组(从而形成所谓的机会主义样例(opportunistic sample))。

- 客户(特别是不满意的客户和前客户)没什么理由是有益的或者诚实的。
- 任何给定的操作可能是各种原因积累的结果。银行客户可能会因为各种原因而流失，比如一个银行分支被关闭，银行退还了一次支票，以及他们必须在自动取款机上等太久等。虽然序列往往更为重要，但是市场调研也许只能发现近似的原因。

尽管有这些缺点，但是与客户和前客户交谈提供了任何其他方式不能提供的见解。美国银行的这个例子表明这两种方法是兼容的。

提示：当针对现有客户进行市场调研时，使用数据挖掘来考虑对他们已有的了解是一个好主意。

1.5.4　度量数据挖掘的影响

作为营销活动采用更好广告语的结果，家庭净值贷款活动的响应率从 0.7%增加至 7%。该集团副总裁 Dave McDonald 说，数据挖掘对银行零售环节的战略影响是不折不扣地将银行从大规模营销机构转换为学习机构。“我们想要达到不断执行营销方案的目的——不只是季度邮件，而是在统一基础之上的方案。”他预见到一个闭环的营销过程：业务数据作为快速分析过程的数据源，该分析过程将创建执行和测试的方案，这又将产生更多的数据以更新该过程。总之，这将形成数据挖掘的良性循环。

1.6　良性循环的步骤

美国银行的例子显示了实践中的数据挖掘良性循环。图 1-1 显示了 4 个阶段：

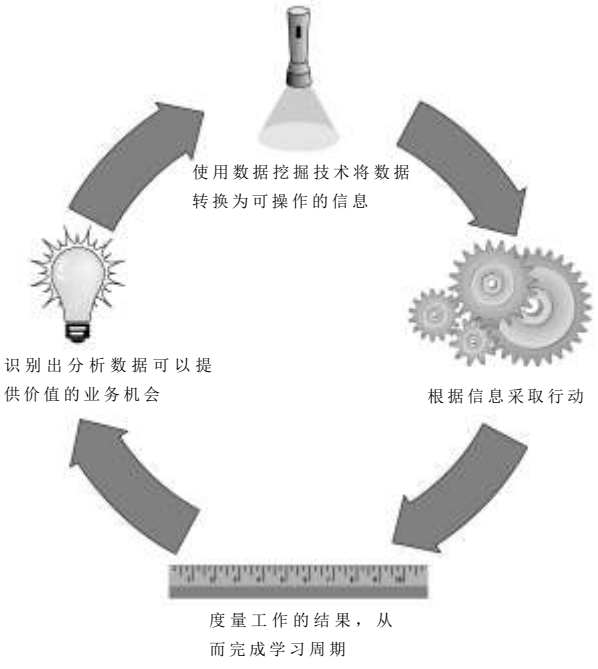


图 1-1　数据挖掘良性循环的重心在于业务的结果，而不只是利用先进的技术

(1) 识别业务机会。
(2) 挖掘数据将其转换为可操作的信息。
(3) 根据信息采取行动。
(4) 度量结果。

根据这些步骤的提示，成功的关键是把数据挖掘合并到业务流程，并能够促进数据挖掘人员和使用结果的业务用户之间的通信。

1.6.1 识别业务机会

数据挖掘的良性循环首先识别合适的业务机会。遗憾的是，有许多很好的统计师和主管分析师的工作基本上都是浪费，因为他们解决的是对业务没有帮助的问题。优秀的数据挖掘人员希望避免这种情况。

为避免浪费分析工作，首先应愿意针对结果采取行动。许多正常的业务流程都很适合数据挖掘：

- 规划新的产品介绍
- 计划直接营销活动
- 理解客户的流失/波动
- 评价营销测试的结果
- 分配营销预算以吸引最有利可图的客户

这些都是数据挖掘可以提高现有业务工作的实例，允许业务经理做出更明智的决定——如定位不同的目标群，更改广告语等。

为了避免浪费分析工作，度量所采取的行动影响，从而判断数据挖掘工作本身的价值也非常重要。正如 George Santayana 所说(在他所有的引述中，通常只有其中的最后一句被记住)：

> 进步，远不在于改变，而在于记忆力。即便改变是绝对的，但是仍有未改进之处以及没有可能改进的方向：如果经验不会保留，则像野蛮人一样，永远保持幼年。那些不吸取教训的人，注定要重蹈覆辙。

在数据挖掘方面，这也同样适用：如果不能度量挖掘数据的结果，那么不能在以往的工作中学习，从而不存在良性循环。

度量过去的努力和有关业务的特设问题同时启示了数据挖掘的机会：

- 什么类型的客户会响应过去的活动？
- 最好的客户住在哪里？
- 在自动柜员机前漫长的等待是客户流失的原因吗？
- 有利可图的客户会使用客户支持吗？
- 哪些产品应该通过 Clorox 漂白来升级？

与业务专家面谈是另一个好的开始方式。因为在业务方面的人员可能不熟悉数据挖掘，因此他们可能不了解如何针对结果采取行动。通过解释数据挖掘对企业的价值，这种面谈为双向沟通提供了一个讨论会。

本书的作者之一曾参加了在电信公司的一系列会议，讨论分析呼叫详细记录(每一个客户完成的呼叫记录)的价值。在其中一次会议中，与会者缓慢地了解到这可能有用。然后，一位同事指出隐藏在数据里面的信息是客户在家中使用传真机(在第 16 章介绍链接分析时详细讨论该结果项目)。这种看法引发了与会者的思考。传真机的使用将会是人是否在家中工作的一个好的指标。对于在家工作的人群，该公司已经根据他们的需要量身定制了产品。然而，如果没有理解数据和技术的人的激励，那么这个营销团队永远不会考虑通过搜索数据来发现在家工作的人群。连接技术和业务突出了一个非常宝贵的机会。

提示：当与业务用户讨论数据挖掘的机会时，确保重心在业务问题而不是技术和算法。让我们的技术专家专注于技术，同时让业务专家专注于业务。

1.6.2　将数据转换为信息

数据挖掘——本书的重点——是将数据转换成可操作的结果。成功的关键在于使得数据在业务中具有意义，而不是使用特定的算法或工具。许多缺陷降低了使用数据挖掘结果的能力：

- 坏的数据格式，例如在客户地址中不包括邮政编码。
- 混乱的数据字段，例如交货日期在一个系统称为“计划交付日期”，而在另一个系统中称为“实际交付日期”。
- 功能缺乏，例如呼叫中心应用程序不允许基于每个客户进行注解。
- 法律影响，如拒绝贷款时必须提供法律依据(而且“我的神经网络告诉我如此”是不可接受的)。
- 组织因素，因为一些业务组不愿改变他们的行动，特别是在没有奖励的情况下。
- 不及时，因为结果可能来得太晚而不再适合采取行动。

不同形式、不同格式以及来自多个系统的数据，如图 1-2 所示。确定适当的数据源并把它们整合在一起是成功的关键因素。每个数据挖掘项目都有数据的问题：不一致的系统、跨数据库的表的键不匹配、记录每隔几个月就会覆盖等。抱怨数据是不做任何事的首要借口。第 17、18、19 章讨论了有关数据的各种问题，首先是数据仓库，然后介绍如何转换成适合数据挖掘的格式。真正的问题是“现有数据能做什么？”这就是本书后续所描述的各项技术的落脚点。

在已经获得了一个强大的服务器和一个数据挖掘的软件包之后，一家无线通信公司曾想过整合一个数据挖掘组。在后期阶段，该公司联系作者来帮助其研究数据挖掘的机会。其中一个机会非常明显。客户流失的一个关键因素是过度呼叫(overcall)：新的客户在第一个月使用的分钟数超出了他们的费用计划。当第一次的账单送达时——有时会在第二个月的中间，客户会了解过度使用。到那时候，与第一个月一样，客户已经在第二个月产生了一个很大的账单，从而更加不快乐。遗憾的是，客户服务组还要等待相同的账单周期之后才能检测出过度使用，没有时间来主动反应。

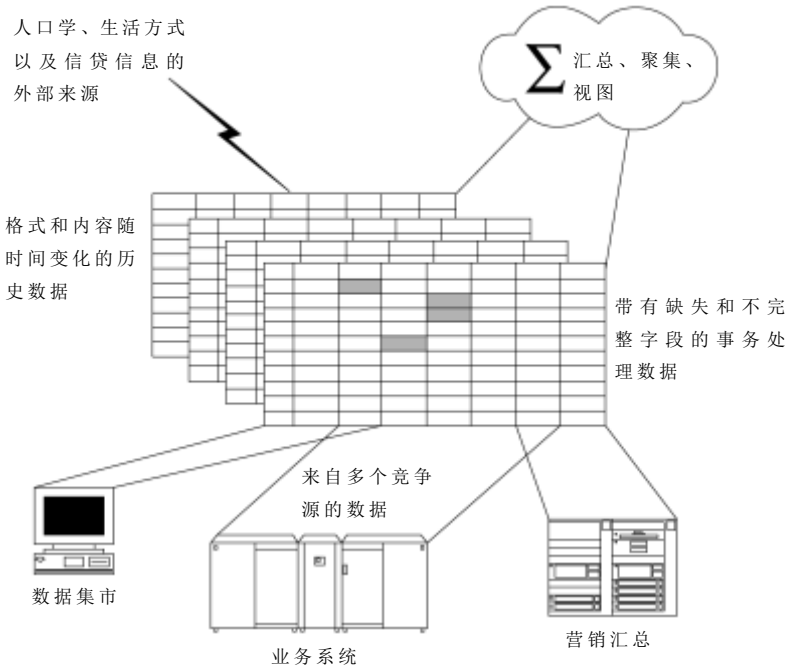


图 1-2　数据从来就不干净，它来自内部和外部的不同形式、不同来源

然而，新生的数据挖掘组拥有资源，而且已经鉴别和调查了适当的数据源。采用一些相当简单的程序，该小组能够在这些客户第一次过度呼叫时把他们标识出来。使用这个信息，客户服务中心能够联系处于风险中的客户，并在第一个账单失效之前把他们移到适当的账单计划中。这个简单的系统获得了重大胜利，展示了数据挖掘的优点。只需拥有一个数据挖掘组——具有技能、硬件、软件以及访问能力——就是把合适的触发器结合起来以保存处于风险中的客户的有利因素。

1.6.3　根据信息采取行动

采取行动是数据挖掘良性循环的目的。正如之前已经提到的，可以采取多种形式的行动。数据挖掘将使得业务决策更为明智。随着时间的推进，更明智的决定会导致更好的结果。

有时，“行动”只不过是做曾经做过的事情——但是此时行动将具有更大(或更少)的信心。即便如此，它也是数据挖掘的成功，因为减少担忧的程度也是一件好事。

更典型地，采取的行动一般会基于正在进行的业务：

- 当客户在线出现时，把结果合并到自动推荐系统。
- 通过直接邮寄、电子邮件、电话推销等向客户和潜在客户发送信息；使用数据挖掘，不同的信息可能会发送给不同的人。
- 优先客户服务。
- 调整库存水平。

数据挖掘的结果必须提交给可以接触客户和影响客户关系的业务流程。

1.6.4　度量结果

我们已经强调过了度量结果的重要性，虽然这个阶段在良性循环中最有可能被忽视。虽然度量并持续改进的价值得到了广泛承认，但是依然没有得到足够的注意，因为它不会立即产生投资回报。没有人回头去检查实际结果与计划相匹配的程度的业务案例会有多少？个人通过比较和学习，通过问为什么计划会匹配或不匹配实际的结果，以及通过愿意学习早期的假设何时以及如何发生错误，从而改进自身的工作。对个人有效同样也适用于企业。

通常，需要基于金融度量来评价营销工作——这些度量非常重要。然而，还需要评价建模工作。考虑一个大型加拿大银行曾经发生的事情，该银行计划向其客户交叉销售投资账户。这个营销信息遍布整个银行：在电视和收音机的广告里、在各分行的海报上、打印在 ATM 收据背后的信息中，在为客户提供客户服务时的信息里，等等。客户不会错过这些信息。

不过，这个故事是关于一个不同的渠道——直接邮寄。数据挖掘的一项工作是识别最有可能对一项投资活动做出响应的客户。营销活动旨在针对可能会做出响应的客户。不过，在当前情况下，该银行包括一个特别的对照组：本组预计会响应良好，但未收到直接邮寄。(补充内容“数据挖掘和营销测试”更详细地讨论了这一思想)。对于直接邮寄的经理而言，把潜在的响应者晾在一边的行动具有相当大的争议。数据挖掘人员的解释是：“我们认为这一组将会做出响应，但是不与他们任何人联系，以便留出一部分让我们可以从这个测试中学习”。

相对于不与一些好的客户联系的代价，学到的知识是相当值得的。对于在投资账户提议中得分较高的客户，无论他们是否接受提议，都有相同比例的客户会开设账户。事实上，该模型确实会找到将开设账户的客户。然而，营销测试同时还发现营销传播是多余的。给定所有其他的营销工作，并不需要这个特别的直接邮递活动。

当开始识别业务问题时，是开始考虑度量的时机。如何才能度量结果？一家公司在发出优惠券鼓励销售其产品时，无疑将度量优惠券的赎回率(redemption rate)。但是无论如何，优惠券的买回者可能已经购买了产品。另一个适当的度量是考虑在特定的商店或者地区的销售增长，其中增长是与特定的营销工作相关联。这些度量可能很难完成，因为它们需要更详细的销售信息。然而，如果目标是要增加销售，那么需要一种直接或间接的方式对其进行度量。否则，营销工作可能都是“慷慨激昂，却毫无意义”。

数据挖掘和营销测试

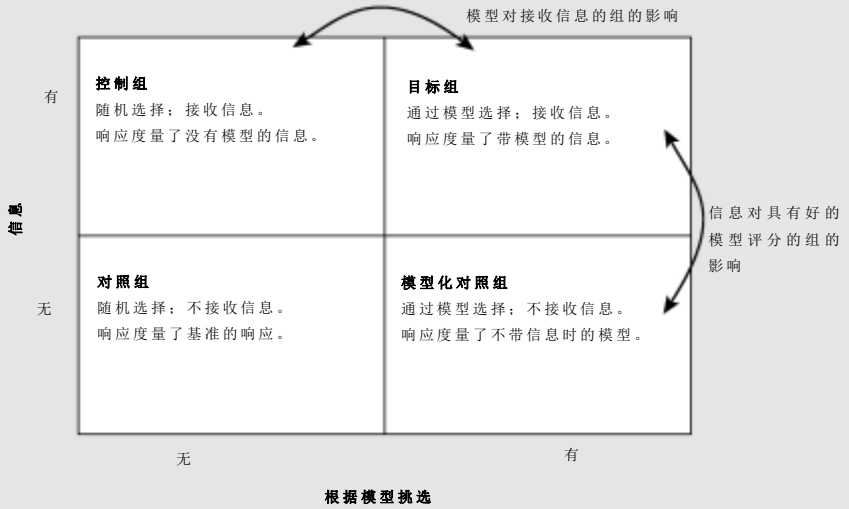
营销测试是分析营销的一个重要部分，数据挖掘同样如此。这两者常常相互补充，营销测试是了解数据挖掘工作是否有效的重要部分。当使用数据挖掘进行营销处置时，通常应该测试两件事情。第一，营销信息是否有效？第二，数据挖掘模型是否有效？

其中的关键在于聪明地使用对照组来分析这两个因素。在实践中，存在 4 个潜在的组：

- 目标组(Target Group)：接受处置，且具有指示响应的模型评分。
- 控制组(Control Group)：接受处置，且随机或基于较低的模型评分进行选择。
- 对照组(Holdout Group)：不接受处置，且随机或基于较低的模型评分进行选择。

- **模型化对照组(Modeled Holdout Group)**：不接受处置，且具有指示响应的模型评分。

这四个组如下图所示：

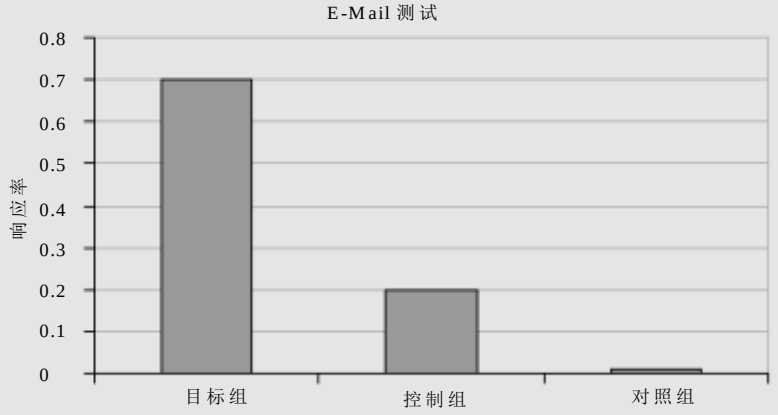


这四个组可用于度量信息和模型工作的有效性

这四个组的响应将会提供有用的信息。使用这些组建模称为增量响应建模(incremental response modeling)，第 5 章将对其进行更详细的讨论。

在加拿大银行学习到直接邮递工作没有必要的示例中，模型化对照组的响应率与目标组的一样。这表明该项工作不会产生效果。目标组和控制组之间的差异可度量建模是否有效。

下面的图表是另一个银行的例子，其使用这些图表来度量活动的有效性。下面这张图是来自银行的实际图：



这张图清晰地显示了响应的区别，以确定处置以及建模是否起作用

头两栏显示目标组比控制组具有更高的响应率，表明建模正在起作用。第二个两栏显示控制组比对照组具有更高的响应率，表明营销处理在起作用。

仅仅度量这四个组只是度量数据挖掘有效性的开始。例如，模型评分往往会分成等分。在这种情况下，为了确定该模型正在起作用，包含来自活动中所有等分的一个样本非常重

要。当然，工作中将包含顶部等分中的每个样本(因为这会实现业务目标)。对于更低层的等分，只需包含一个样本。样本应该足够大，以确定等分是否真的起作用——当使用模型时这相当重要。第 4 章解释了用于确定这些测试的合适大小所需的相关统计背景知识。

可能在营销干预发生数周或数月之后生成的标准报告中会包含摘要。即使拥有信息，营销经理可能也不具备从这种报告中收集重要发现的技术技能。了解保留客户的影响意味着需要在更长的时间里跟踪已有的营销工作。精心设计的报告应用程序对营销组和营销分析师的帮助会很大。然而，对某些问题而言需要更多的详细信息。

将每个营销工作看作一个小的业务案例是一个好主意。比较期望与实际的结果使得能够识别出在下一轮良性循环中可利用的数据挖掘机会。你常常会太忙于投入精力处理下一个问题而不能度量当前工作的有效性。这是个错误。不管成功与否，每次数据挖掘工作都会带来可应用到未来工作的经验。问题是度量什么以及如何应用该度量方法，使之为将来使用提供最好的输入。

作为示例，让我们以有针对性地获取活动需要度量什么作为开始。标准化的度量是响应率：本活动所针对的人群中有多少人会真正响应？这使得在数据表中会存在很多信息。对于使用模型评分(其中高得分意味着响应的可能性更高)的客户获取工作，一些有未来价值的问题示例包括：

- 该活动是否抵达(reach)并带来有利可图的客户？
- 得分更高的模型评分会表明更高的响应率吗？
- 这些客户是被保留还是可预期？
- 本活动所抵达的最忠实客户的特征是什么？
- 新获得的顾客会购买额外的产品吗？
- 一些信息或优惠会比其他的更有效吗？
- 活动所抵达的客户可通过其他备用渠道抵达吗？

所有这些度量会为在未来做出更明智的决定提供信息。通过学习，数据挖掘将连接过去与未来的行动。

较为特殊的是度量客户生存期价值(lifetime customer value)。顾名思义，这是对客户在他的整个关系过程中(或者在未来的某个固定的时期，如接下来的两年)所体现价值的一种估计。在某些行业，为了估计客户生存期价值，已经发展出非常复杂的模型。即使没有复杂的模型，短期的估计(如一个月后、六个月和一年之后的价值)也可证明是相当有用的。下一章将更详细地讨论客户价值。

1.7　良性循环上下文中的数据挖掘

考虑美国的一家大型电信公司。该公司有数百万客户，它拥有位于中央办公室的数百或数千个交换机，其通常是位于多个时区中多个不同的州。每个交换机可以同时处理数千起电话——包括诸如呼叫等待、会议呼叫、呼叫转移、声音邮件与数字服务等高级要求。作为曾经开发过的最复杂的计算设备，交换机可以由少数几个制造商来提供。一个典型的电话公司会从每个供应商处获得几个交换机的多个版本。其中，每个交换机会在每次呼叫

和尝试呼叫中提供大量的、具有自己格式的数据卷——每天会产生 10 GB 以上的卷。此外，每个州都有自己影响行业的规定，更不用说联邦法律和条例的改变会更为频繁。为了添加困惑度，该公司向其客户提供了成千上万的不同账单计划，其客户包括从偶尔的住宅用户到《财富》前 100 的公司。

这个公司——或任何拥有大量数据和大量客户的类似公司——会如何管理其负责收益的账单流程？答案很简单：要非常仔细！公司已经开发了处理标准操作的详细过程；他们拥有政策和程序。这些过程很可靠。即使业务重组、数据库管理员在度假、计算机临时崩溃，甚至即使法律法规发生了变化、交换机已经升级或者飓风袭击，账单都将会发给客户。如果企业可以管理的流程与每个月针对数以百万计的普通客户、企业客户和政府客户的精确账单一样复杂，那么显然把数据挖掘与决策过程相结合会非常容易。确实如此吗？

在设计和实现业务相关的关键任务应用程序方面，大公司拥有数十年的经验。然而，数据挖掘不同于典型的业务系统(如表 1-1 所示)。在运行成功的业务系统时所需的技能并不一定能保证成功的数据挖掘。

表 1-1　数据挖掘不同于典型的操作业务流程

典型的业务系统	数据挖掘系统
在历史数据上操作和报告	对历史数据的分析通常应用于大多数当前数据以确定未来的行动
可预测的和周期性的工作流，通常与日历关联	不可预测的工作流，其取决于企业和市场的需求
专注于单个项目(item)，每次一个(干草堆里的一根针)	每次专注于一个更大的组，试图了解整个干草堆
限制使用企业范围的数据	数据越多，结果越好(通常而言)
重点是经营范围(如账户、区域、产品代码以及使用时间等)，而非客户	重点是可操作的实体、产品、客户与销售区域
响应时间通常以秒/毫秒(对于交互式系统而言)为单位度量，但是等待报告需要数周/数月	迭代过程的响应时间通常以分钟或小时为单位度量
数据记录系统	数据的拷贝
叙述性和重复性	创造性

通过数据挖掘解决的问题不同于业务问题——数据挖掘系统不寻求完全复制以前的结果。事实上，复制以前的工作会导致灾难性的结果。这可能会导致反复向相同的目标客户进行营销活动。你不想通过分析数据了解到一大群客户符合之前活动中联系的客户的剖析(profile)。数据挖掘过程需要考虑这些问题，与典型的业务系统想要反复复制相同的结果不同——比如是否完成了一个电话、寄汇票、授权信用购买、跟踪库存或其他无数的日常业务。

数据挖掘是创造性的过程。数据包含许多明显的相关性，它们毫无用处或者仅仅表示当前的业务策略。例如，分析一个大型零售商的数据表明购买维修合同的人也很可能会购买大型家用电器。除非零售商要分析电器维修合同的销售有效性，否则这种信息比无用更

糟，因为考虑的维修合同仅仅与大型设备一起销售。花费数百万美元的硬件、软件以及数据挖掘人员成本，最后只发现这样的结果实在是在浪费资源，若把它们应用于业务的其他地方或许会更好。分析师必须理解业务的价值是什么，以及如何合理安排数据以挖掘出金块(nugget)。

数据挖掘的结果会随着时间推移而变化。随着时间的流逝，模型将会过期而渐渐降低其用处。原因之一是数据在快速地衰老。同样，市场和客户也在快速地变化。

数据挖掘为其他可能需要更改的过程提供反馈。在商业世界中所做出的决定经常影响当前过程以及与客户的交互。通常，观察数据可以发现业务系统的缺陷，以及应该加以修正以便提高未来客户理解的不完善之处。

1.8　经验教训

数据挖掘是客户关系管理系统的重要组成部分。客户关系管理系统的目标是要尽可能地重新构建与客户的密切学习关系，它们往往为经营有方的小企业所享有。公司与客户的交互会产生大量的数据。数据最初是由事务处理系统捕获，如自动柜员机、电话交换机的记录以及超市的扫描仪文件等。然后可以对数据收集、清洗，并对其进行汇总以包含在一个客户数据仓库中。一个精心设计的客户数据仓库将包含客户交互的历史记录，它们会成为公司的记忆。数据挖掘工具可应用于该历史记录以学习客户的信息，从而使得公司未来向客户提供更好的服务。本章给出了几个数据挖掘商业应用程序的示例，诸如更好的优惠券定位、推荐、交叉销售、客户保留以及信贷风险降低等。

数据挖掘本身就是在大量的数据中找到有用的模式和规则的过程。为了获得成功，数据挖掘必须成为大的业务流程的一个组成部分，即数据挖掘的良性循环。

数据挖掘的良性循环将利用数据的力量，并把它转化为可执行的业务结果。正如水曾经推动轮子转动以驱动整个工厂中的机器一样，必须将数据收集起来，并在整个企业中传播以产生价值。如果数据是这个比喻中的水，那么数据挖掘就是那个轮子，而良性循环将把数据力量传播到所有的业务流程。

数据挖掘的良性循环是一个基于客户数据的学习过程。它首先识别适合数据挖掘的业务机会。最好的业务机会是那些将要执行的行动。如果没有执行，则从了解客户中获得的收益会很少或没有价值。度量行动的结果也非常重要。这样就完成了整个良性循环的循环，并且经常会给出进一步的数据挖掘机会。

下一章将考虑在客户上下文中的数据挖掘，首先介绍客户生存周期，接下来是良性循环在执行中的几个例子。

第 2 章

数据挖掘在营销和客户关系管理中的应用

数据挖掘技术并非存在于真空之中，它们与业务上下文息息相关。尽管这些技术自身都很有趣，但是它们终究只是一种工具。本章将介绍业务上下文。

本章首先描述客户生存周期(customer lifecycle)，以及与每一阶段相关联的业务流程。正如贯穿本章始终所描述的，客户生存周期的每个阶段都为客户关系管理和数据挖掘提供了机会。客户生存周期是中心主题，因为数据挖掘所支持的业务流程都围绕该生存周期来组织。

本章解决的业务主题所涉及的客户关系复杂度大致会逐渐升高：以潜在客户开始，接着是已建立的客户关系，最后以保留(retention)和赢回(winback)结束。在讨论业务应用的过程中，本章会介绍相关的技术资料，但特定数据挖掘技术的详细信息将留待后续章节介绍。

2.1 两个客户生存周期

术语“客户生存周期”有两种不同的意思——客户的个人生存周期，或者客户关系的生存周期。从数据挖掘的观点来看，后者通常更为重要。

2.1.1 客户个人生存周期

客户，无论他们是个人、家庭或者企业，都会随着时间推移而发生变化。自从创业者创建公司之后，有些会成为收购的目标；有些会独立地持续增长。其中大部分公司最终会以失败而告终。个人生存周期是以生活事件而标记，例如高中毕业、有了孩子以及找工作等等。

这些不同的生活阶段对于市场和客户关系管理而言非常重要。例如，搬家是一项重大事件。当人们搬家时，他们可能会购买新家具、订阅当地报纸和开设一个新的银行账户等。

了解到谁正在搬家对于把他们定为营销目标非常有用，其中对于家具商、报纸、有线电视公司和银行而言尤为重要，特别是在搬家之后的几天或几周之内。对于其他生活事件而言同样如此，从要高中毕业和大学毕业，到要结婚、有孩子，换工作和退休等。了解这些生活阶段将使公司能够定义与特定人群产生共鸣的产品和信息。

一些企业正是围绕特定的生活阶段而构建的。一家婚礼商店专注于结婚礼服，这种业务要获得增长不是依靠女性更加频繁的结婚，而是通过声誉和推荐来实现。类似地，搬家公司不必鼓励他们的近期客户搬迁，他们需要的是招揽新客户。

对于大多数企业而言，客户的个人生存周期相对不那么重要。在任何情况下，基于生活阶段管理客户关系都非常困难，因为：

- 及时地识别事件是一项挑战。
- 许多事件都是一次性的或者非常罕见的。
- 生活阶段事件通常不可预测或超出控制能力。

无论如何，这些缺点并不会致使这些生活阶段无用，因为它们对于理解客户可能的需求非常关键。然而，大多数业务过程都是围绕着另一个不同的生存周期——客户关系的生存周期——而组织。

2.1.2 客户关系生存周期

与客户的业务关系会随着时间演变。虽然每个业务各自不同，但是客户关系都将客户分为 5 个主要阶段，如图 2-1 所示：

- 目标市场的潜在客户，但还不是客户。
- 响应者是那些表现出一定兴趣的潜在客户——例如，通过填写申请单或者在网站上注册表示兴趣。
- 新客户是已经作出承诺的响应者，通常是一项支付协议，例如已经完成第一次购买，已经签订合同，或者已经在网站中注册一些个人信息。
- 已建立的客户是返回的新客户，他们是关系有望扩大或加深的对象。
- 前客户是那些已经离开的客户，他们或者是自愿流失(因为他们流失到竞争对手那里或不再能发现产品的价值)，强制流失(因为他们没有支付账单)，或者是预期流失(因为他们不再在目标市场；例如，因为他们已经搬家了)。

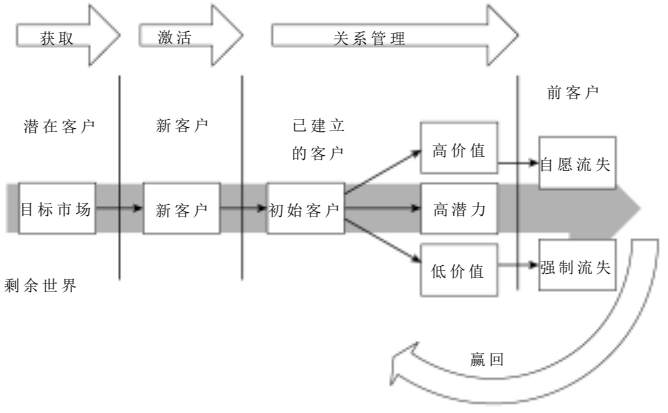


图 2-1　客户生存周期在不同阶段的进展

这些阶段的精确定义都取决于业务。对于一个 e-media(电子媒体)网站而言，潜在客户可能是 Web 上的每个人；响应者是访问该网站的人；新客户是已经注册的人；而已建立的客户则是重复的访问者。前客户是那些在一定时间内未返回的人，该时间取决于站点的性质。对于其他业务而言，这些定义可能会完全不同。例如，人寿保险公司有一个目标市场。响应者是那些填写了一种申请单的人——然后往往对他们会抽血进行血液检查。新客户是那些被接受的申请者，而已建立的客户是那些为保险付款支付保险费的人。

2.1.3　基于订阅的关系和基于事件的关系

客户生存周期关系的另一个维度是在每次交互中保持不变的承诺。考虑下列成为电话公司客户的方式：

- 在公用电话上(如果您还能找到的话！)打电话。
- 购买一张可打一定分钟数的预付费电话卡。
- 购买一部预付费移动电话。
- 购买一部带不定期合同的延后支付移动电话。
- 购买一部带合同的移动电话。

前三个示例是基于事件的关系。后两个示例是基于订阅的关系。下面将更详细地探讨这些关系的特点。

提示：不间断的账单关系是不间断订阅关系的一个好迹象。这种不间断的客户关系提供了在商业活动过程中与客户进行对话的机会。

1. 基于事件的关系

基于事件的关系是基于事务的。客户可能返回，也可能不返回；随着时间的推移，追踪客户可能会很困难或者根本不可能。在之前的示例中，电话公司对客户可能根本就没有太多的信息，当客户采用现金支付时尤为如此。匿名事务仍然包含信息，但是，显然没有什么机会直接向未提供任何联系信息的客户发送信息。

当基于事件的关系占主导时，公司通常通过广播信息与潜在客户进行通信(例如，广告、网站广告、病毒式营销，等等)，而不是把特定的个人作为目标发送信息。在这些情况下，分析工作非常集中于产品、地理和时间，因为这些是在客户事务中所能了解的信息。

广播广告不是抵达潜在客户的唯一方式。通过邮件或在 Web 上分发优惠券是另一种方式。美国的制药公司已经非常善于鼓励潜在客户访问它们的网站以获得更多信息——与此同时，公司将会收集他们的一些信息。同时，许多公司采用 Web 和社交网络与其他匿名客户通信。

有时，基于事件的关系意味着具有中间人的企业对企业关系。制药公司再次提供了一个示例，因为它们的许多营销预算是花在医生，而不是购买药物的病人之上。

2. 基于订阅的关系

基于订阅的关系为了解客户提供了更为自然的机会。在早期所给出的列表中，最后两个示例都具有不间断的账单关系，其中客户已经同意为一段时间的服务支付费用。订阅关

系为未来的现金流(未来的客户付款流)提供了机会，并且为与每个客户的交互提供了许多机会。

基于订阅的关系可能采取账单关系的形式，但是它们也可能采取零售亲和卡(retailing affinity card)或在网站注册的形式。在某些情况下，账单关系是某种类型的订阅，其没有为追加销售(up-sell)或交叉销售(cross-sell)提供空间。已经订阅了杂志的客户几乎可能没有机会扩展关系。其实还有某个机会。杂志客户可以购买一个礼物订阅或者购买名牌产品。然而，未来的现金流非常取决于目前的产品结构。

在其他情况下，不间断的关系只是一个开始。信用卡公司会每个月邮寄一张账单；然而，没有费用则没有亏欠。长距离的提供商可能会每个月向客户收费，但它可能只是每月的最低限额。一个编目员(cataloger)向客户发送目录，但是大部分的客户不会购买。在这些示例中，使用激励(usage stimulation)是关系的一个重要部分。

关系的开始和结束是定义基于订阅关系的两个关键事件。当这些事件明确定义时，生存分析(见第 10 章)是一种了解关系持续期的好的候选方式。但是，有时定义关系的结束会很难：

警告：定义一个客户关系的结束可能会很困难。不同的定义会产生不同的模型，而且有时候会导致不同的结论。在定义目标变量之前，在关系何时被认为结束方面达成共识。

- 信用卡关系可能当一个客户收支不平衡，并且在指定的时限(如 3 个月或 6 个月)内没有产生事务时结束。
- 目录关系可能当一个客户在指定时限(如 18 个月)内没有从目录中购买时结束。
- 亲和力卡关系可能当一个客户未在指定的时限(如 12 个月)内使用该卡时结束。

即使充分理解了关系，可能还会出现一些棘手的局面。关系的结束日期应该是客户打电话取消的日期还是账户关闭的日期？对于未成功支付其最后账单的客户，在自动请求终止服务之后，是否应该与因为滞纳而被停止的客户等同视之？

这些情况被作为理解客户关系的指导方针。图 2-2 针对报纸订阅客户的简单案例，制定了客户体验的不同阶段。基本上，这些客户具有以下类型的交互行为：

- 通过一些渠道开始订阅
- 更改产品(工作日更改为 7 天、周末更改为 7 天、7 天更改为工作日、7 天更改为周末)
- 暂停交付(通常是因为休假)
- 抱怨
- 停止订阅(无论是自愿或非自愿)

在基于订阅的关系中，通过收集所有这些不同类型的事件，从而形成客户关系的一张图，使得随着时间的推移，理解客户成为可能。

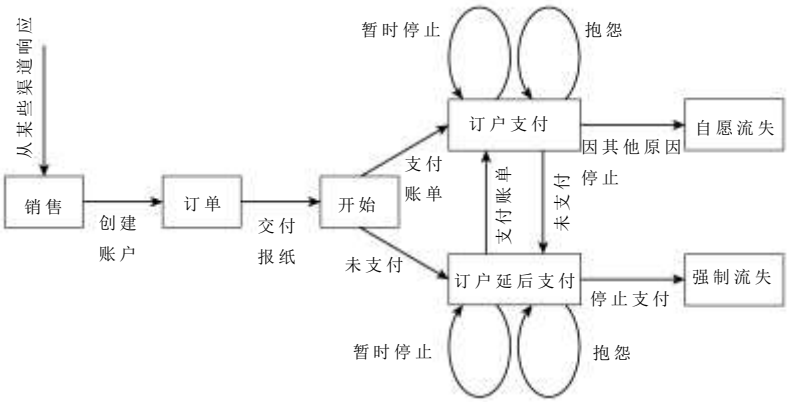


图 2-2　报纸订户的(简化)客户体验，包括几种不同类型的交互

2.2　围绕客户生存周期组织业务流程

业务流程将客户从客户生存周期的一个阶段转移到另一个阶段。这些业务流程非常重要，因为它们将使得客户随着时间推移而更具价值。本节将介绍这些不同的流程以及数据挖掘在其中所起的作用。

2.2.1　客户获取

客户获取(Customer Acquisition)是吸引潜在客户，并把他们转化为客户的过程。这通常是通过广告、口碑以及有针对性的营销来完成的。数据挖掘能够并且确实在获取过程中发挥了重要作用。

关于获取有三个重要问题：潜在客户是谁？何时获取一个客户？数据挖掘的作用是什么？

1. 潜在客户是谁

了解潜在客户非常重要，因为信息应该针对适当的受众。使用历史数据的挑战之一是潜在客户可能会随着时间而变化：

- 地理上的扩张会带来潜在客户，他们可能与也可能不与原始地区的客户类似。
- 更改产品、服务以及价格可能带来不同的目标受众。
- 竞争可能会改变潜在客户的结构。

这些类型的状况可能会带来以下问题：过去是对未来的一个好的预测器(predictor)吗？在大多数情况下，答案是“是的”，但是必须对过去明智地加以利用。

下面的故事举例说明了一个必须注意的情况。在纽约地区，一家公司在曼哈顿已经拥有大量的客户基础，并希望把它扩大到郊区。它已经针对曼哈顿完成了直接邮寄活动，并根据这些活动的响应者建立了模型集(model set)。这个故事的重要之处在于曼哈顿的富裕居民社区具有较高的浓度(concentration)，因此该模型集会向富人偏置。正如预期，响应者比周围地区的潜在客户更加富有。然而，非响应者同样更加富有。

当模型扩展到曼哈顿以外的地区时，该模型会选择什么地区？它选择一些最富有的社区，因为这些地区的人在人口统计上看起来与在曼哈顿的响应者类似。尽管在这些地区有良好的潜在客户，该模型错过了许多其他的潜在客户，这一点可以通过使用邮件中的对照组(本质上，它们是名字的随机抽样)来发现。具有较高响应率的地区为富有的地区，但是并非与用于构建模型的曼哈顿社区一样富有。

警告：当把响应模型从一个地理区域扩展到另一个区域时要小心。结果可能会给出更多类似的人口信息，而非响应信息。

2. 何时获取客户

通常获取客户有一个基本过程，该过程的细节取决于特定行业，但一些一般性的步骤如下：

- 客户以某种方式在某个日期响应。这是“销售”日期。
- 在一个基于账号的关系中，创建该账户。这是“账户开设日期”。
- 以某种方式使用该账户，这是“激活日期”或“第一次购买日期”。

有时，所有这些事情会在同一时间发生。混乱也屡屡发生——错误的信用卡号、拼写错误的地址以及客户反悔等。结果是好几个日期都可能作为获取日期。

假设所有这些有关的日期都可用，那么最好使用哪一个呢？这取决于业务需求。在直接邮寄空投或电子邮件轰炸之后，检查响应曲线以了解响应预计何时会发生将很有趣，如图 2-3 所示。针对此目的，销售日期是最重要的日期，因为它表明了客户行为，而问题正是针对客户行为。

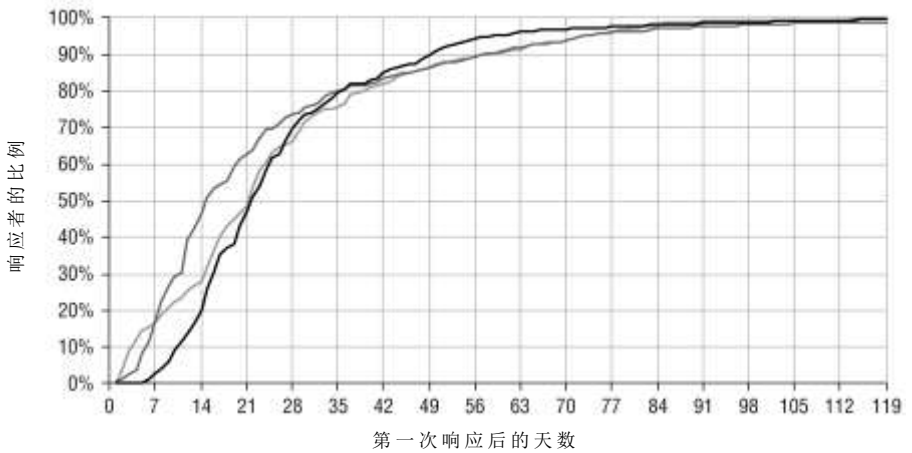


图 2-3　三次直接邮寄活动对应的响应曲线表明，80%的响应来自五六个星期之内

不同的问题可能会有不同的答案。例如，为了比较不同组的响应率，账户设立日期可能更为重要。那些注册了一次“销售”，但是从未开设账户的潜在客户应从这种分析中排除。

3. 数据挖掘的作用是什么

可用的数据限制了数据挖掘可以发挥的作用。响应模型用于诸如直接邮寄和电话销售之类的渠道，其中联系的成本相对较高。因此，目标是将联系限制为更有可能做出响应并

成为优质客户的潜在客户。可用于这一努力的数据分为三类：

- 潜在客户源
- 额外的个人或家庭数据
- 附加的地理级别的人口统计数据(典型的有人口普查或人口普查组)

这里的目的是要从数据挖掘的角度讨论潜在客户。

概述一个典型的获取策略将是一个好的开始。使用电子邮件、直接邮寄或出站(outbound)电话销售的公司会购买清单(list)。一些清单在以往表现很好，所以会完全使用它们。对于其他清单，可能使用建模来确定联系哪些潜在客户。当人口统计信息在家庭级可用时，可能会基于这些附加人口统计信息建模。当这些人口统计信息不可用时，可能使用社区人口统计信息，而不是在不同的模型集合中选择。

在构建用于客户获取的数据挖掘模型时，回声效应(echo effect)(也称为光环效应)是一个挑战。可能通过某一种渠道抵达潜在客户，但是其通过另一个渠道做出响应。公司向一组潜在客户发送电子邮件，其中有些人可能不会单击该电子邮件中的链接，而是通过电话响应。潜在客户可能收到广告信息或直接邮件，但是通过网站进行响应。或者，广告宣传活动可能会鼓励在相同时间内通过几种不同的渠道进行响应。图 2-4 显示了一个回声效应示例，展示了两个渠道——入站电话(inbound call)和直接邮寄之间的相关性。

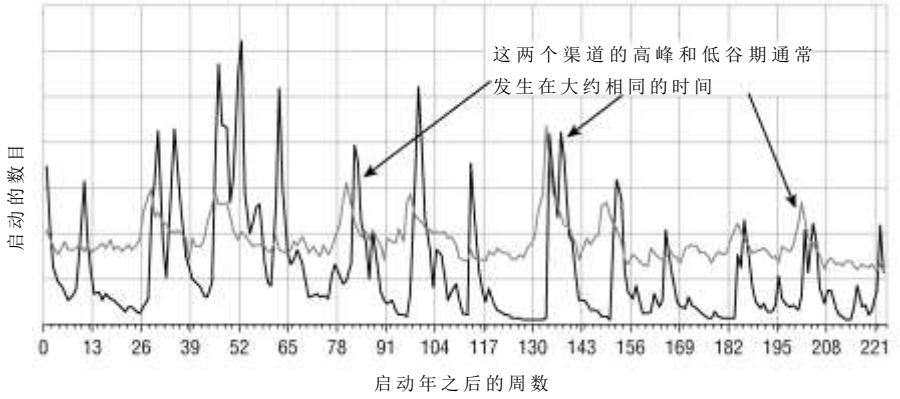


图 2-4　回声效应可能导致人为地低估或高估渠道的性能，因为通过一个渠道激发的客户可能归功于另一个渠道

2.2.2　客户激活

激活过程可能像客户在网站上填写注册表一样简单。也可能会涉及更为冗长的审批过程，如信用检查。甚至可能会更加繁杂，例如在寿险公司的示例中，其在设置比率之前需要进行一次保险业测验。一般而言，激活是一种业务流程，更多的是关注业务需求而非分析需要。

作为一种业务流程，客户激活可能看起来与数据挖掘无关。激活提供了一个新客户在其启动时的视图。这种视角非常重要，而且作为一种数据源，需要对其进行维护。初始条件和随后的变化都很有趣。

提示：客户激活提供客户关系的初始条件。这样的初始条件对于预测客户的长期行为通常会很有用。

激活过程常常被描绘成一个漏斗，如图 2-5 所示，尽管一个过滤器堆栈可能是一个更贴切的比喻。从漏斗顶部倒入的一切最终都将从底部流出。对于潜在客户而言并非如此。

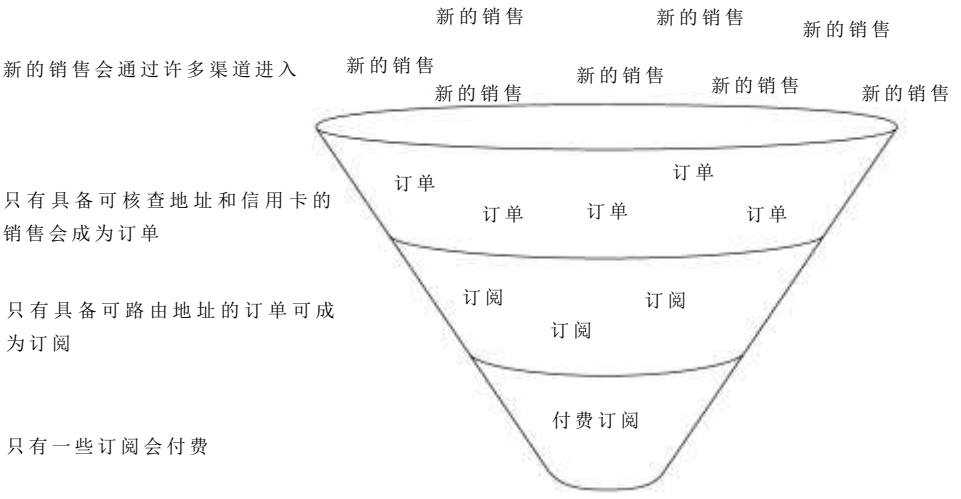


图 2-5　客户激活过程漏斗会在激活过程中的每个步骤消减响应者

该图阐明了一家报纸送货上门订户的激活过程。它具有下列步骤：

(1) 销售。一个潜在用户表示有兴趣获得订阅，通过 Web、电话或者邮件中的响应卡提供地址和支付信息。

(2) 订单。创建一个账户，其中包括对地址和支付信息的初步核查。

(3) 订阅。报纸实际上是物理交付，需要进一步核实地址和特别的交货说明。

(4) 付费订阅。客户支付该份报纸。

每个步骤都将失去一些客户，也许仅有几个百分点，也许会更多。例如，信用卡可能会无效、失效日期不合适，或者提供的地址不匹配。客户可能生活在交付区域以外。交付员可能不了解特别的交货说明。地址可能是在一个不允许访问的公寓楼内。其中大多数都是基于操作考虑(而例外是客户是否会支付)，它们说明了客户激活所涉及的不同类型的操作关注以及过程。

数据挖掘可以在理解客户是否按照其合理的方式移动中发挥作用——或者哪些特征会导致在客户激活阶段失败。这些结果可以帮助改进业务流程。通过强调带来销售但还没有转换为付费订阅的策略，它们还可以为客户获取提供指导。

对于与 Web 相关的业务，客户激活通常是——虽然并非总是——花费少量时间的自动处理过程。当它正常运转时，不会存在问题。尽管激活阶段的时间很短，但是它是客户获取流程的关键部分。当其失效时，则会失去潜在的有价值客户。

2.2.3　客户关系管理

客户关系管理系统的主要目标是提高客户价值。这通常包括以下活动：

- **追加销售**(up-selling)。使客户购买高端产品和服务。
- **交叉销售**(cross-selling)。拓宽客户关系，例如使客户除书籍之外，还购买 CD、机票和汽车等。
- **使用激励**(usage stimulation)。确保客户再次消费；例如，通过确保客户看到更多的广告或使用信用卡更多地消费。
- **客户价值计算**。为每个客户指定一个预期价值。

对于提供许多产品和服务的公司而言，一种危险是不能获得正确的信息。客户并不一定想要选择，他们可能只想要简单。让客户在接二连三的营销沟通中发现他们感兴趣的内容，说明在信息传递方面做得很糟糕。因此，向每个客户重点介绍其可能会感兴趣的少数几个产品的信息将会非常有用。当然，每个客户都具有不同的潜在设置。数据挖掘在发现这些关系方面发挥了关键的作用。

数据挖掘在了解业务的操作方面也可以发挥作用。第 21 章包含一个案例研究，讨论一个大型的卫星电视提供商如何结合结构化和非结构化的呼叫中心数据来发现一个与其业务系统相关的性能问题。正如挖掘评论文本之后所确定的，特定主题的呼叫花费了太长时间而不能解决。这个问题与服务代表无关，而是因为用来解决特定问题类型的系统反映迟钝而导致的。

客户关系管理系统中一个最重要的部分也许就是保留客户。这是预测模型最常应用的领域之一。保留客户有两种方法。第一种是比较很短时间就离开的客户与保持时间很长的客户。第二种方法是生存分析(见第 10 章)，其直接建模客户的保留时间。

2.2.4　赢回

即使客户已经离开，仍有可能将他们吸引回来。赢回(winback)就是试图这么做，具体操作方法是向有价值的前客户提供激励措施、产品以及价格促销等。

赢回往往比数据分析更加依赖于业务策略，但是数据挖掘可能在确定客户为什么会离开方面发挥作用，尤其是当客户服务投诉和其他行为的数据可以纳入模型集时。第 21 章介绍一个案例研究，其识别由于抵制而离开一家媒体公司的客户。显然，对于这些客户的赢回策略不同于其他客户。

有些公司具有专门的“拯救团队(save team)”。客户在离开之前，将必须与一个接受过留住客户培训的人沟通。除了努力留住客户之外，拯救团队还将完成另一项跟踪客户离开原因的工作——这些信息对于未来的客户保留工作非常有价值。

试图把不满的客户吸引回来相当困难。因此，更重要的工作是一开始就向他们提供有竞争力的产品、有吸引力和有用的服务，以尽量留住他们。

2.3 数据挖掘应用于客户获取

对于大多数企业而言，地球上 70 亿左右的人中只有很少一部分是真正的潜在用户，大部分将根据地理、年龄、偿还能力、语言，或者产品或服务的需要而排除在外。对于提供房屋净值信贷额度的银行，它们会自然地把该服务限定为在银行所注册运行的辖区内的房屋所有者。一家出售后院秋千装置的公司会希望得到有孩子且很可能有后院的家庭的信息。一本杂志的目标将是会阅读适当的语言，并且其广告商会感兴趣的人。

数据挖掘可以在发现潜在客户方面扮演多种角色。其中最重要的是：

- 识别好的潜在客户。
- 为抵达潜在客户选择通信渠道。
- 对不同的潜在客户群挑选合适的信息。

虽然所有这些都很重要，但是第一个——即识别好的潜在客户——得到了最广泛的实施。

2.3.1 识别好的潜在客户

好的潜在客户的最简单定义是那些至少表示有兴趣成为客户的人，这一定义被许多公司所采用。更复杂的定义要求会更高。真正好的潜在客户不仅有兴趣成为客户而且他们有条件成为客户，他们将会是有价值的客户，他们不太可能会欺诈公司，并且可能会支付账单，如果处理得好，他们将会是忠实的客户并会推荐其他客户。无论潜在客户的定义多么简单或复杂，第一个任务就是发现他们。

无论是通过广告或是通过诸如邮件、电话或电子邮件等更直接的渠道发送信息，目标对象很重要。在某种程度上，甚至广告牌上的信息也是定向的；航空公司和租车公司的广告牌往往会出现在指向机场的高速公路上，使用这些服务的人很可能会是驶向机场的客户。

应用数据挖掘，首先定义什么是好的潜在客户，然后寻找规则把满足这些特征的人群作为营销目标。对于许多公司而言，使用数据挖掘识别好的潜在客户的第一步是构建一个响应模型。本章稍后会深入讨论响应模型，它们的各种使用方式，以及它们能做什么和不能做什么。

2.3.2 选择通信渠道

潜在客户需要通信。广义地说，公司会有意以几种不同的方式与潜在客户进行通信。一种方法是通过公共关系，即鼓励媒体介绍公司的故事并通过口碑传播正面的信息。虽然这种方法对有些公司高度有效(如 Facebook、Google 和 eBay)，但是公共关系并非直接的营销信息。正如本书第 21.6.3 一节中所介绍的，即使是在这里，数据挖掘也可以提供帮助。

从数据挖掘的观点来看，更有趣的是广告和直接营销。广告可以采用任何形式，从火柴盒封面到广告词，从商业网站的赞助商链接到重大体育赛事期间的电视节目以及电影中的产品放置等。在这方面，广告基于共同特点定位目标人群；然而，许多广告媒介不能对个体定制信息。

2.3.3　挑选适当的信息

即使是售卖同样的基础产品或服务，不同的信息也只适合不同的人。一个典型的例子是权衡价格和便利程度。有些人对价格很敏感，他们愿意在仓库购物，在深夜打电话，以及不断更改飞机以获得更好的交易。而有些人则愿意支付额外的费用以获得最便捷的服务。基于价格的信息不仅不能激发寻求便利者，而且还有把他们引向获利较少的产品的风险，即使他们乐意支付更多。

2.4　数据挖掘示例：选择合适的地方做广告

定位潜在客户的方法之一是寻找类似于当前客户的人。通过调查，一个全国性的出版物明确其读者具有以下特征：

- 59%的读者受过大学教育。
- 46%具有专业或行政职务。
- 21%的家庭收入超过 75 000 美元/年。
- 7%的家庭收入超过 10 万美元/年。

了解该剖析将在两个方面帮助出版社：首先，通过把符合该剖析的人作为潜在客户，可以提高自身宣传工作的响应率。第二，受过良好教育、高收入的读者可以用来向想要抵达这类客户的公司销售杂志上的广告空间。

由于本节的主题是目标潜在客户，让我们看看该杂志如何使用剖析来加强其定位潜在客户的工作。基本思想非常简单：当该杂志想要在电台中做广告时，它应该寻找听众匹配该剖析的电台；当它想要放置布告牌时，它应该在匹配该剖析的社区这么做；当它想要进行呼出电话营销时，它应该呼叫匹配该剖析的人。数据挖掘的挑战在于对“匹配剖析”是指什么做一个好的定义。

2.4.1　谁符合剖析

确定客户是否符合剖析的方法之一是度量客户和剖析之间的相似度——也称作距离。数据中包括了表示订阅者在特定时间的一个快照的调查结果。什么样的度量适合这种数据呢？剖析是以百分比的形式表示(58%是大学教育；7%超过 10 万美元)，但是对于个人而言，要么是大学教育，要么不是大学教育；收入要么超过 10 万美元，要么不超过 10 万美元，对于这种情况应如何处理呢？

考虑两个受访者。艾米受过大学教育，赚 80 000 美元/年，并且是一个专业人士。鲍勃是一个高中毕业生，挣 50 000 美元/年。哪个人更匹配读者剖析？答案取决于如何作比较。表 2-1 显示了一种用来打分的方法，其仅仅使用剖析和一个简单的距离度量。

表 2-1 通过比较个人与每个人口统计学度量计算个人的匹配值

	读 者	“是”得分	“否”得分	艾 米	鲍 勃	艾米得分	鲍勃得分
大学教育	58%	0.58	0.42	是	否	0.58	0.42
专业或行政	46%	0.46	0.54	是	否	0.46	0.54
收入>$75K	21%	0.21	0.79	是	否	0.21	0.79
收入>$100K	7%	0.07	0.93	否	否	0.93	0.93
总计						2.18	2.68

这个表基于读者符合每个特征的比例计算分数。例如，因为 58%的读者是大学教育，所以艾米这一特性得分 0.58。而鲍勃没有从大学毕业，获得分数为 0.42，因为其他 42%的读者没有从大学毕业。继续对每个特征如此处理，并且把这些分数累加在一起。艾米最后得分为 2.18，而鲍勃得分更高一点为 2.68。他的得分更高反映出，相比艾米而言，他与目前读者的剖析更相似。

这种方法的问题在于，虽然鲍勃看起来比艾美与这个剖析更相似，但是艾米看上去更像是该杂志的真正目标读者——即受过大学教育、个人收入较高的人。这一定位的成功显然是比较了读者剖析与美国整个人口的特征。

与整体人口相比，读者受过更好的教育、更专业，获得更高的薪资。在表 2-2 中，“指数(Index)”列比较了读者特征与整个人口特征，它是将具有特定属性的读者比例除以整个人口中具有该特征的比例。读者中受过大学教育的比例大约是整个人口中该比例的三倍。类似地，他们没有接受大学教育的比例大约只有整个人口中该比例的一半。通过使用指数作为每个特征的得分，艾米得分 8.42 (2.86 + 2.40 + 2.21 + 0.95)，而鲍勃得分只有 3.02 (0.53 + 0.67 + 0.87 + 0.95)。基于指数评分更好地反应了该杂志的目标受众。新的得分更有意义，因为它们现在结合了额外的信息，即目标受众与美国整个人口的不同。

表 2-2 通过考虑在整个人口中的比例计算分数

	是			否		
	读 者	美 国 人 口	指 数	读 者	美 国 人 口	指 数
大学教育	58%	20.3%	2.86	42%	79.7%	0.53
专业或行政	46%	19.2%	2.40	54%	80.8/%	0.67
收入>$75 K	21%	9.5%	2.21	79%	90.5%	0.87
收入>$100K	7%	2.4%	2.92	93%	97.6%	0.95

人口普查域数据

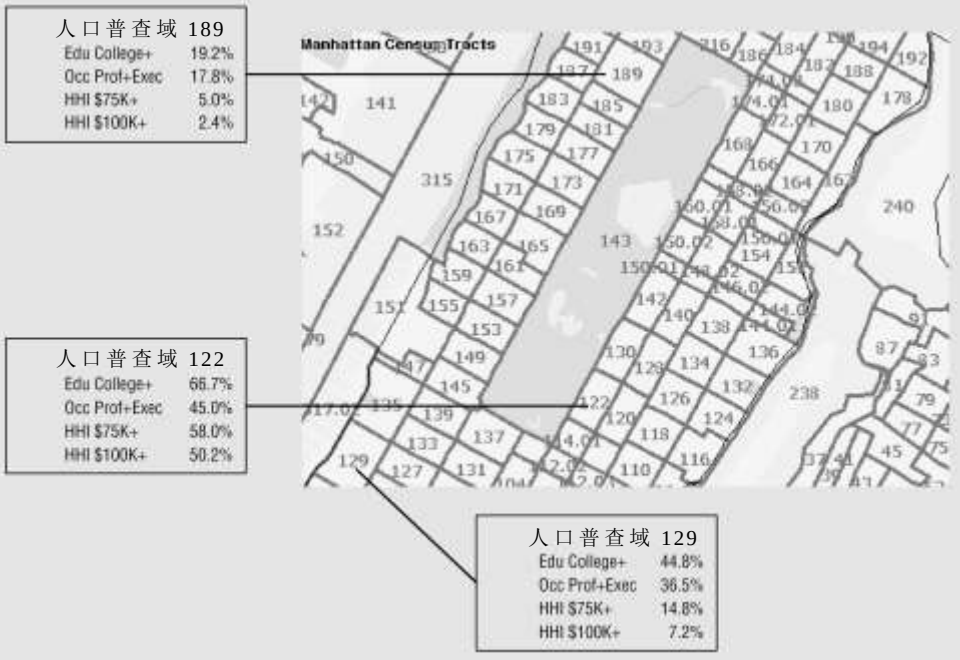
美国政府宪法规定每 10 年执行一次人口普查。人口普查的主要目的是为了分配每个州在众议院的席位。在满足这一任务的过程中，对美国人口的普查还提供了丰富的信息。

即使在非普查年，美国人口普查局(www.census.gov)也通过问卷的方式调查了美国人口，其中包含了详细的问题，诸如收入、职务、上下班的习惯、消费模式和其他更多的信

息等。对这些调查表的响应为人口的剖析提供了基础。

人口普查局不发布个人信息，而是对小的地理区域的信息进行聚合。最常用的是人口普查域(census tract)，平均约包含 4000 位个人。虽然人口普查域大小不一，但是它们在人口上比其他的地理单元(如县和邮政编码)更一致。

人口普查有更小的地理单元、块(block)和块组(block group)；为了保护居民的隐私，不提供人口普查域级别之下的一些数据。利用这些单元，可以分别根据国家、州、都市统计区(metropolitan statistical area，MSA)、立法区等聚合信息。下图显示了一些曼哈顿中心的人口普查域：



营销的哲学之一是基于一个古老的谚语“有羽毛的鸟会聚在一起(物以类聚)”。具有类似兴趣和品味的人会生活在类似的区域(无论是自愿或是因为历史的歧视模式)。根据这一思想，向您已有客户所在的地区以及类似的地区进行营销是一个好主意。人口普查资料对于了解客户浓度(concentrations)所在位置以及确定相似地区的剖析都非常有价值。

提示：比较客户剖析时，需要铭记整体人口的剖析非常重要。因此，使用指数往往比使用原始值表现更好。

2.4.2　度量读者群的适应度

基于指数打分所蕴含的思想可以扩展到更大的人群。这很重要，因为用于度量人口的特征可能不是对每个客户或潜在客户有效。幸运的是，前述特征都是可以通过美国人口普查获得的人口统计特征，并且可以通过地理区来度量，如人口普查域、邮政编码、县和州等(参阅补充内容“人口普查域数据”)。

目标是根据每个普查域与该杂志的适应度对其进行打分。例如，如果有一个普查域的

成年人口 58%受过大学教育，那么其中的每个人都将因为这一特性得到 1 分。如果 100%是大学教育，那么比分仍然是 1——这是您能做的最完美适应度。但是，如果只有 5.8%从大学毕业，那么这一特征的适应度得分降为 0.1。整体的适应度得分是个体得分的平均。

图 2-6 阐明了补充内容所提及的三个曼哈顿的人口普查域。每个域都有四个正在考虑的特征的不同比例，它们将结合起来得到每个域的整体适应度得分。得分表示那一个域符合剖析的人口比例。

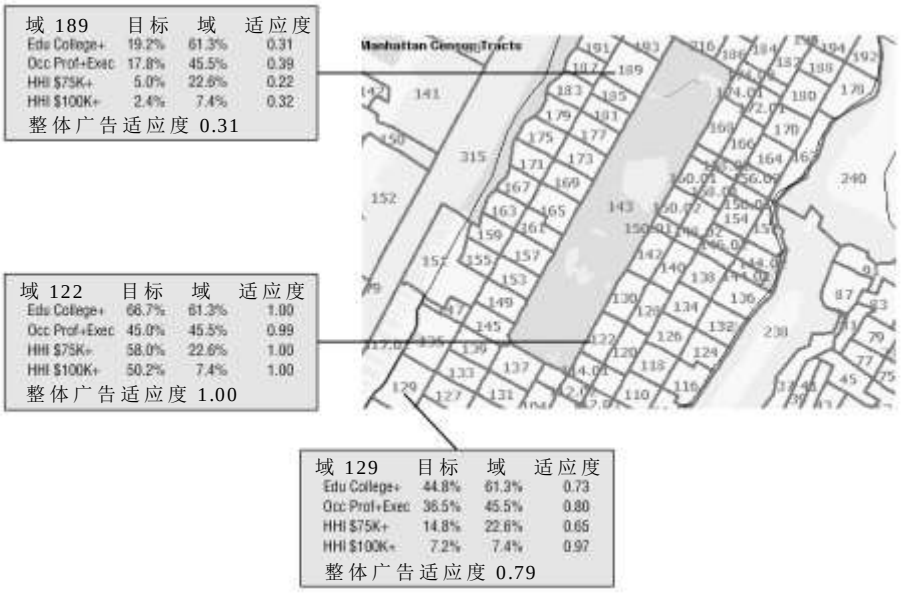


图 2-6　计算三个曼哈顿人口普查域的读者适应度示例

2.5　数据挖掘改进直接营销活动

广告可以用来抵达对其个人信息一无所知的潜在客户。直接营销需要至少有一小部分额外的信息，例如名称和地址、电话号码或电子邮件地址。信息越多，则数据挖掘的机会就越多。最起码，数据挖掘可以通过选择联系人来改进目标市场选择(targeting)。

第一级目标市场选择并不需要数据挖掘，而只需要数据。在美国，存在相当多的人口数据。在许多国家，各种公司会编译和出售家庭级别的数据，包括各种各样的内容，诸如收入、孩子数量、教育程度，甚至业余爱好等。其中一些数据是从公共记录中收集的。房屋购置、婚姻、生育和死亡都属于公共记录的范畴，它们可以从县法院和证书注册表中收集。其他的数据可以从产品注册表单中收集。其中有些数据是使用模型估算得出的。

该数据用于商业目的的规则会因国家的不同而发生变化。在某些国家，可以出售带地址的数据，但不能带姓名。在其他国家，数据可能只能用于某些批准的用途。在一些国家，数据使用的限制较少，但只覆盖了有限数量的家庭。在美国，有些数据(如医疗记录)是完全不可用的。而有些数据，如信用记录，只能用于某些批准的用途。其余许多数据则不受限制。

警告：在商业上可用的家庭数据的范围以及对使用它们有相对较少的限制方面，美国都是不同寻常的。尽管许多国家提供了家庭数据，但是使用不同的规则。对于跨界转移个人数据有特别严格的规则。在计划使用家庭数据进行营销之前，需要熟悉使用它们的法律限制。

基于诸如收入、汽车的所有权或有孩子之类的信息，可以首先直接使用家庭级别的数据对群组进行粗粒度的划分。问题在于，即使应用了明显的筛选器，相对于可能会做出响应的潜在客户数量，剩余池可能依然非常大。因此，数据挖掘的一个主要应用是目标市场选择——发现最有可能实际响应的潜在客户。

2.5.1　响应建模

通常，直接营销活动的响应率只有很低的个位数。通过识别更有可能对直接征求进行响应的潜在客户，响应模型可用于提高响应率。最有用的响应模型会提供一个实际的响应似然(likelihood of response)估计，但并非严格要求如此。任何模型只要能够根据响应似然对潜在客户排序就够了。给出一个排序列表，直接市场营销人员可以通过邮寄或呼叫列表顶部附近的人，提高活动所能抵达的响应者比例。

以下部分描述了该模型得分可以用来改进直接营销的几种方法。这些讨论独立于用来生成得分的数据挖掘技术。本书中的许多数据挖掘技术都可应用于响应建模。

根据直接营销协会(Direct Marketing Association，一个行业协会)，通常一次 10 万封邮寄活动的成本约为 10 万美元，虽然价格会根据邮寄的复杂度而变化很大。其中一些费用，如开发创造性的内容、准备艺术品以及初始打印设置等，独立于该次邮寄的规模。其余的费用会直接根据邮寄的邮件数量而变化。已知邮件订单响应者或积极的杂志订阅者的邮件列表可以在“每千个姓名的价格”的基础上购买。邮件的车间生产成本和邮费都可以在类似的基础上收费。邮寄规模越大，固定成本将变得越不重要。为便于计算，本章的示例假设直接邮寄活动抵达一个人的费用是一美元。这并非一种不合理的估计，虽然简单邮件的成本较低而独特邮件的成本较高。

2.5.2　优化固定预算的响应

利用模型得分的最简单方法是使用它们来指定排名。根据响应倾向评分对潜在客户排名之后，可以对潜在客户列表进行排序，从而使那些最可能响应的潜在客户位于该列表的顶部，而那些最不可能响应的位于列表底部。许多建模技术可用于生成响应评分，其中包括回归模型、决策树和神经网络。

每当没有足够的时间和预算达到所有的潜在客户时，对潜在客户列表进行排序就有意义。如果必须排除某些人，那么应该留下最有可能响应的那些人。并非所有的业务都必须排除潜在客户。一个当地的有线电视公司可以考虑城里的每个家庭都是一个潜在客户，而且它有能力一年内联系所有的家庭好几次。当营销计划要求针对每个潜在客户而努力时，则不太需要响应模型！然而，数据挖掘仍然可能用于选择适当的信息和预测潜在客户可能的行为方式。

更可能的场景是，营销预算不允许对每个潜在客户采取相同级别的花费。考虑一家公司 Simplifying Assumptions Corporation(SAC)的潜在客户列表上有 100 万个姓名，其在某次营销活动中的预算是 30 万美元，其中每次联系的成本是 1 美元。这家公司可以通过响应模型对潜在客户列表打分，并向顶端 30 万个得分所对应的潜在客户发送优惠品，以最大化 30 万美元支出所能获得的响应数目，如图 2-7 所示。

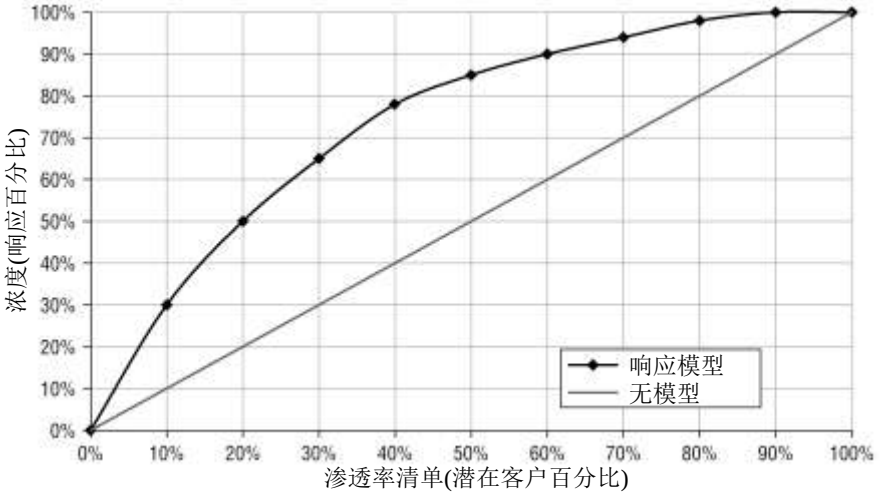


图 2-7　累积收益或浓度图显示了使用模型的收益

上层的曲线画出了浓度，即随着包括在活动中的潜在客户越来越多而获得的所有响应者的比例。其中直对角线是用于比较。它表示没有模型时会发生什么，此时浓度不是根据渗透率(penetration，联系的潜在客户的百分比)变化的函数。邮寄到随机选择的 30%的潜在客户将会发现 30%的响应者。利用该模型，邮寄到顶部 30%的潜在客户将会发现 65%的响应者。浓度对渗透率的比称为提升(lift)。这两条线之间的差异就是收益(benefit)。收益和提升将在补充内容中讨论。

此处画出的模型在第三个十分位值(decile)处提升了 2.17。利用模型，SAC 在支出 30 万美元时获取的响应者将两倍于随机联系 30%的潜在客户。

收益和提升

累积收益图表(参见图 2-7)通常是用来讨论提升的。提升度量了浓度与渗透率的关系。如果人口中的响应者是 10%，但是根据模型挑选的一组中响应者为 20%，则提升为 2。提升是在给定潜在客户列表的深度时比较两种模型性能的一种有用方法。然而，它未捕获另一个概念，在观察图时其直观上看起来很重要——即两行相距多远，以及在哪个渗透率它们分离得最远？

统计人员 Will Potts 将浓度和渗透率之间的差异命名为收益。根据他的术语，差异最大的点称为最大收益点。注意最大收益点并不对应最高提升点。通常，提升在浓度图的左边获得最大化，此时浓度最高且曲线的斜率最陡峭。

最大收益与每类累积概率分布函数之间的最大距离成正比

在渗透率最大的位置分割潜在客户列表的模型得分，同样也是最大化 Kolmogorov-

Smirnov(KS)统计的得分。KS 测试在一些统计员中颇受欢迎，尤其是在金融服务行业。它用于测试两个分布是否不同。在最大化收益的点处划分潜在客户列表将形成“好的列表”和“坏的列表”，其中响应者的分布最大限度地彼此分离。在这种情况下，“好的列表”中响应者的比例最大，而“坏的列表”中比例最小。

浓度曲线上最大收益点反映了相应的 ROC 曲线和无模型线之间的最大垂直距离

ROC 曲线(详见第 5 章的描述)，类似于更为熟悉的浓度或累积收益图表，因此它们之间有关系并不令人惊讶。ROC 曲线显示了在两种类型的误分类(misclassification)错误之间的权衡。在累计收益图中的最大收益点对应了 ROC 曲线上类之间分离最大的点。

最大收益点反映了最大化敏感性(sensitivity)和特异性(specificity)无加权平均值的决策规则

如同医疗世界所使用的，敏感性度量了基于测试的诊断正确的似然性。它是在测试中获得阳性结果的人中是真阳性的比例。换句话说，它是真阳性除以真阳性和假阳性的和。特异性是在测试中获得阴性结果的人中是真阴性的比例。一个好的测试应该既是敏感的又是特异的。最大收益点是最大化这两种度量平均值的割点(cutoff)。

假设误分类成本与目标类的发生率成反比，最大收益点反映了一个最小化预期损失的决策规则

评价分类规则的方法之一是为每种类型的误分类指定成本，并基于该成本比较规则。无论他们是否代表响应者、不遵守规则者、黑客或具有某种特殊疾病的人，罕见的案例通常是最令人感兴趣的，因此错过其中一个案例比误分类一个常见案例的代价更高。根据这种假设，最大化收益会挑选一个好的分类规则。

2.5.3　优化活动收益率

毫无疑问，活动的响应率翻倍是一个理想的结果，但是它真正的价值是多少呢？甚至活动是否有盈利？虽然提升是比较模型的一种有用方法，但是它并没有回答这些重要问题。为了处理收益率，需要更多的信息。特别是，关于收益率需要收入以及成本信息。让我们对 SAC 示例添加一些更详细的信息。

SAC 公司以单一价格出售一款单一产品。该产品的价格是 100 美元。SAC 制造、库存以及分发产品的费用总额是 55 美元。前面已经提到，它要抵达潜在客户的成本是 1 美元。现在有足够的信息来计算响应的价值。每个响应的毛收入是 100 美元。每个响应的净值在考虑与响应相关的费用(55 美元货物成本，1 美元的联系成本)之后，获得每个响应的净收益为 44 美元。信息概述如表 2-3 所示。

表 2-3　SAC 公司的损益矩阵

邮　寄	响　应	
	是	否
是	$44	-$1
否	$0	$0

此表说明，若联系上一个潜在客户并且其响应，则该公司赚 44 美元。如果联系上一个潜在客户，但是没有响应，则该公司损失 1 美元。在这个简化的示例中，没有在选择不联系潜在客户时的成本和收益。一个更复杂的分析可能会考虑到这样的事实：存在不联系潜在客户的机会成本，其可能已经响应过，甚至非响应者也可能因为联系，而通过提高品牌意识成为更好的潜在客户，并且这种响应者可能比单次购买所表明的客户具有更高的生存周期值。

这个简单的损益矩阵可用来将活动的响应转换成利润图。忽略活动开销的固定成本，如果每 44 个潜在客户未响应之后会有 1 个响应，则该活动就会保本。如果响应率比它要高，则该活动有利可图。

警告：如果把失败的联系成本设置得太低，则损益矩阵会建议联系每个人。对于其他理由而言，这可能不是一个好主意。它可能会导致接二连三地向潜在客户提供不适当的优惠。

一个更复杂的活动收益率分析会考虑该活动的启动成本、整个人口中的基本响应率，以及所联系人群的截止渗透率(cutoff penetration)。回顾一下，SAC 的预算为 300 000 美元。假设整个人口中基本的响应率为 1%。预算足以联系 30 万潜在客户，或者潜在客户池中的 30%。在 30%的深度时，该模型会提供的提升大约为 2 ，所以 SAC 可以预期响应者的数量为它没有使用模型时的两倍。在这种情况下，两倍意味着 2%而不是 1%，获得 6000 个(2% ×300 000)响应者，其中每个人值净收入 44 美元。根据上述假设，SAC 从响应者处获得 600 000 美元的毛额以及 264 000 美元的净额。与此同时，98%的潜在客户或者 294 000 个潜在客户没有响应。其中每个人的成本为 1 美元，因此 SAC 在活动中亏损 30 000 美元。

表 2-4 显示了用于生成图 2-7 中累积收益图表的数据。这表明该活动可能有利可图，通过花较少的钱来联系少量、但是响应率更好的潜在客户。仅向 1 万个潜在客户，或者潜在客户列表顶部的 10%邮寄，则会获得的提升为 3。它将把基本的响应率从 1%提到 3%。在当前场景下，有 3 000 人会作出反应，产生 132 000 美元的收入。现在有 97 000 人没有响应，且他们每人的成本为一美元。因此，结果会产生 35 000 美元的利润。更好的消息是，SAC 将留下 200 000 美元的营销预算，可将其用于另一个营销活动或者在此活动中提高优惠，也许会进一步提高响应。

表 2-4 每个十分位值获得的提升和累积收益

渗 透 率	收 益	累 积 收 益	提 升
0%	0%	0%	0.000
10%	30%	30%	3.000
20%	20%	50%	2.500
30%	15%	65%	2.167
40%	13%	78%	1.950
50%	7%	85%	1.700
60%	5%	90%	1.500

(续表)

渗　透　率	收　　益	累 积 收 益	提　　升
70%	4%	94%	1.343
80%	4%	96%	1.225
90%	2%	100%	1.111
100%	0%	100%	1.000

一个较小、目标市场选择更好的活动可能比更大、更贵的活动更为有利可图。提升会随着列表变小而增加，那么更小会一直表现更好吗？答案是“否”，因为绝对收入会随着响应者数目的下降而减少。举一个极端的例子，假设模型可以通过发现一组响应率为 100%(此时基本响应率为 1%)的潜在客户而产生 100 的提升。这听上去很神奇，但是如果该组中只有 10 个人，那么它仍然仅仅值 440 美元。同时，更切合实际的例子中会包括一些预先的固定成本。图 2-8 给出了假定活动除了 1 美元的联系成本之外，还有固定成本 20000 美元时会发生什么，其中每个响应的收益为 44 美元，以及基本响应率为 1%。该活动只在大约 10%的小范围文件渗透率时才有利可图。

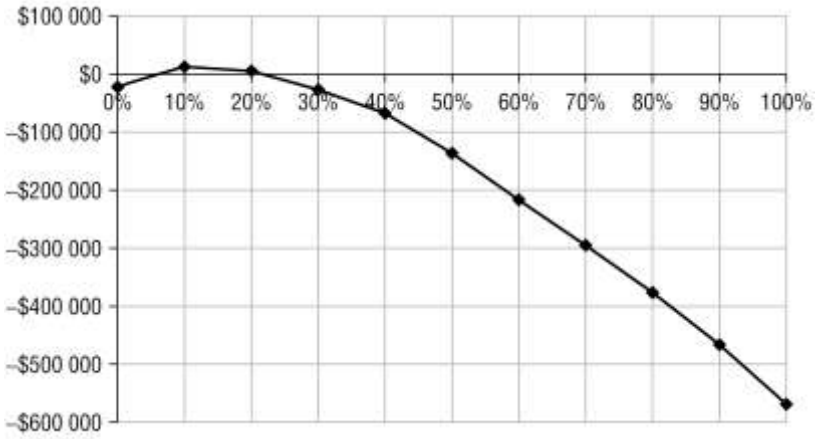


图 2-8　活动收益率是渗透率的函数

利用模型来优化活动的收益率，似乎比简单地用于挑选将谁放入一个预先确定大小的邮件或者电话清单中更具有吸引力，但是该方法同样存在缺陷。首先，结果取决于该活动的成本、响应率以及每个响应者的收益，但是这些都是在活动运行之前所不知道的。在现实生活中，对这些只能进行估计。其中任何一项发生小变化都能把之前示例中的活动变得完全无利可图，或者使其在更大范围的十分位值处有利可图。

图 2-9 显示了如果成本假设、响应率和收益都降低 20%时，活动将会怎样变化。在悲观情况下，可以获得的最好结果是亏损 20 000 美元。在乐观情况下，这个活动在 40%的渗透率处会获得最大 161 696 美元的利润。成本估计往往相当准确，因为它们都是基于邮费、打印费和其他可以事先定好的因素。响应率和收益的估计通常只能是猜想。

优化活动的收益率听上去很诱人，但是如果没有在实际的测试活动中实施，则不可能

有用。预先建模活动的收益率主要是假设分析，以确定基于各种假设所能获得的收益率边界。在运行之后计算活动的收益率更为有用。为了有效地做到这一点，具有各种响应得分的客户都必须包括在活动中——甚至是来自较低的十分位值的客户。

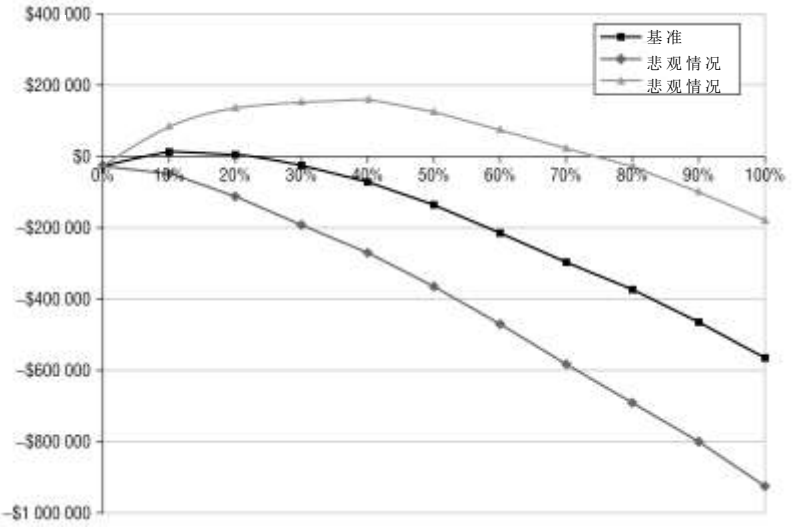


图 2-9　在响应率、成本以及每个响应者的收益方面发生 20%的变化对活动的收益率有很大的影响

警告：活动的收益率取决于许多只能预先估计的因素，因此唯一可靠的方法是使用实际的市场测试。

2.5.4　抵达最受信息影响的人

一个微妙的简化假设是营销活动会激励响应。但是，存在另一种可能：模型仅仅可以识别在有活动或者没有活动时那些可能会购买这种产品的人。

提示：为了测试模型以及它支持的活动是否有效，可以跟踪响应率与模型得分的关系，考虑那些不是活动一部分的支持组中的潜在客户，以及那些包括在活动中的潜在客户。

营销活动的目标是改变行为。在这方面，抵达无论如何都会购买的潜在客户，不会比抵达尽管已接受优惠但不会购买的潜在客户更有用。标识为可能响应者的组也可能不太受营销消息的影响。他们在目标组的身份意味着他们过去很可能已经从竞争对手获得了类似的信息。他们可能已经拥有了产品或类似替代品，或者会坚持拒绝购买它。营销信息对于之前从未听过这一切的人而言可能会有更大差异。得分最高的分组无论如何都可能会有响应，即使没有营销投资。这导致了几乎自相矛盾的结论：在响应模型中得分最高的分组可能不会提供最大的营销投资回报。

走出这种困境的方法是直接建模活动的实际目标，它不仅仅是抵达后续会购买的潜在客户。目标应该是抵达那些因为联系过而更有可能购买的潜在客户。这被称为增量响应建模(incremental response modeling)，是第 5 章和第 7 章讨论的主题。

2.6　通过当前客户了解潜在客户

找到好的潜在客户的一种好办法是查看目前最好的客户来自哪里。这意味着使用某种方法来确定谁是当前最好的客户。这也意味着需要记录当前客户是如何获取的，以及在获取时他们看起来如何。

依赖当前客户学习在哪里寻找潜在客户的危险在于当前客户反映了过去的营销决策。研究当前的客户不会提出在尚未尝试的位置寻找新的潜在客户。不过，当前客户的性能是评估现有获取渠道的好方法。

为了潜在客户的目的，了解当前客户在他们自身还是潜在客户时的样子很重要。理想情况下您应该：

- 在客户成为客户以前开始跟踪他们。
- 在获取时收集新客户的信息。
- 建模获取时的数据和未来感兴趣的结果之间的关系。

以下几小节对它们进行更详细的阐述。

2.6.1　在客户成为“客户”以前开始跟踪他们

甚至可以在潜在客户成为客户之前开始记录他们的信息。网站会在第一次看到访问者时发出一个 cookie，开始一个匿名的剖析以记录访问者在该网站上所做的事情。当访问者返回(在同一计算机上使用相同的浏览器)时，该 cookie 会被识别，同时剖析将会更新。当访问者最终成为一个客户或者注册用户时，导致这种转变的活动将成为客户记录的一部分。

在脱机世界跟踪响应和响应者同样是好的做法。第一个需要记录的关键信息是潜在客户响应或者没有响应的事实。描述谁响应了、谁没有响应的数据是未来响应模型的一个要素。只要有可能，响应数据还应该包括刺激响应的营销行为、捕获响应的渠道，营销信息的时间选择以及响应进来的时间。

确定许多营销信息中的哪些信息刺激了响应需要技巧。在某些情况下，它甚至是不可能的。为了使工作变得更加轻松，响应表单和目录中应包括标识代码(identifying code)。网站访问会获取引用链接。甚至连广告宣传活动也可以通过使用不同的电话号码、邮政信箱、Web 地址，以及最后手段——询问响应者来区分。

2.6.2　收集新的客户信息

当潜在客户开始成为客户时，存在一个收集更多信息的黄金机会。在从潜在客户转换为客户之前，关于潜在客户的任何数据往往都是地理和人口统计数据。购买列表中除了姓名、联系信息以及列表源之外不可能提供任何其他信息。使用地址可以根据所在社区的特征推断出潜在客户的其他事情。姓名和地址在一起可用于从营销数据提供商购买潜在客户家庭有关的信息。这类数据可用于较为广泛的目标，例如“年轻母亲”或“城市青少年”等一般性分组，但是它们不足以详细到形成个性化的客户关系。

提示：在地理级别(邮政编码、人口普查域等等)的人口统计信息非常强大。然而，这些信息不提供个人客户或家庭的信息；它提供了他所在的社区信息。

其中收集的对未来数据挖掘最有用的字段是初始购买日期、初始获取渠道、响应的优惠、初始产品、初始信用评分、响应时间和地理位置。作者发现这些字段可用于预测大量的有趣结果，如预期的关系持续期、坏账以及额外购买等。应保持这些初始值，而不是随着客户关系的发展用新值来覆盖它们。

2.6.3 获取时间变量可以预测将来的结果

通过在获取时记录客户的一切信息，然后在随后的时间跟踪客户，企业就可以使用数据挖掘将获取时间变量与将来的结果相关联，诸如客户关系的寿命、客户价值和默认的风险等。然后，这些信息就可用来指导营销工作，使之集中于可产生最佳结果的渠道和信息。例如，您可以使用第 10 章描述的生存分析技术建立每个渠道的平均客户周期。通常，有些渠道所产生客户的生存周期会是从其他渠道所产生客户的两倍。假设可以估计客户每个月的价值，则可以将它翻译成一个实际的美元图，其可用于比较典型渠道 A 的客户和典型渠道 B 的客户的价值——该图和常常用来估计渠道的“每个响应的成本”度量一样有价值。

2.7 数据挖掘应用于客户关系管理

客户关系管理的重点自然是已建立的客户。幸运的是，已建立的客户对挖掘而言是最丰富的数据源。最重要的是，通过已建立的客户生成的数据反映了实际的个人行为。客户是否准时付账？是支票、信用卡或 PayPal？最后购买是什么时候？所购买的产品是什么？花了多少钱？客户呼叫了多少次客户服务？联系了客户多少次？客户最常使用什么运送方法？客户退货多少次？这类行为数据可用来评估客户的潜在价值，评估他们将结束关系的风险，评估他们将停止支付账单的风险，以及预见他们未来的需要。

2.7.1 匹配客户的活动

对于向现有客户定制混合营销信息而言，响应模型评分对现有客户比潜在客户更有用。营销不会在一旦获取客户之后停止。有交叉销售活动、追加销售活动、使用激励活动、忠诚方案、保留活动等等。您可以将这些活动看作是在争夺客户的访问。

当单独考虑每个活动，并且对所有客户给定在每个活动中的响应评分时，通常会出现一个相似的客户组会在许多活动中得到高分。一些客户就是比其他客户具有更多的响应，这一事实反映在模型的评分上。这种方法会导致客户关系管理质量较差。高分组会因为收到接二连三的信息而生气和不做出响应。与此同时，其他客户从未收到过公司的信息，因此无法激发他们扩大关系。

一种替代方法是限制发送到每个客户的信息数量，使用评分来决定最适合每个人的信息。即使客户对每种优惠都评分很低，但是他们对某些优惠的评分也可能比其他人高。*Mastering Data Mining*(Wiley，1999 年)这本书描述了该系统如何用于定制一家银行网站，

根据客户的银行行为突出显示他们最有可能感兴趣的产品和服务。图 2-10 显示了其工作原理。

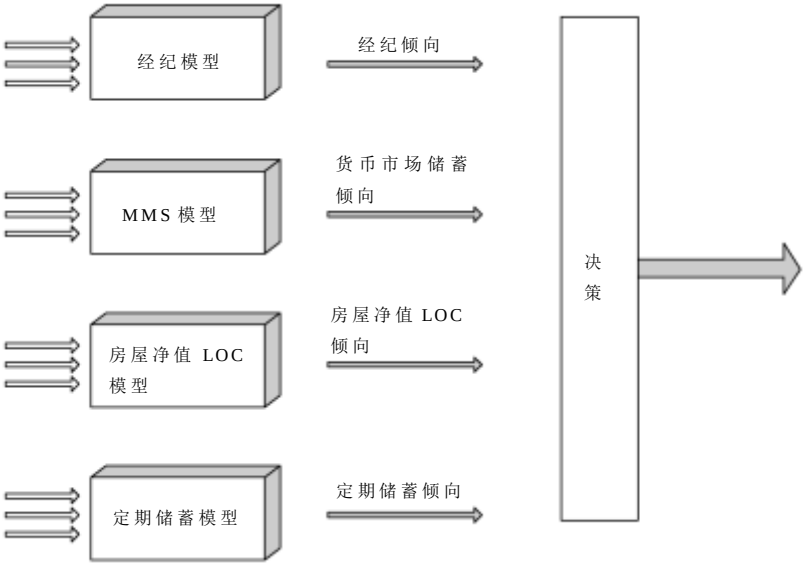


图 2-10　比较多个模型的评分以决定向客户提供哪些优惠

对每个客户给予每个产品的倾向得分。倾向得分是估计客户将响应特定产品优惠的概率。得分逻辑之一是那些已经有产品的客户的倾向得分为 0。在决策框中，倾向得分将乘以与每个产品相关联的第一年利润的平均值，以获得预期的美元价值。最后将向客户提供具有最高期望值的产品。

eBay 提供了另一个示例。在线商城使用一个决策引擎来决定向通过 Google 搜索到达该站点的人所显示的内容。基于用户在搜索引擎输入的搜索字符串，以及在 eBay 系统存储的用户剖析中的搜索字符串，动态地创建一个登录页。登录页通常会包含一系列的链接组合，包括指向 eBay 公司自己卖方的链接，以及由外部广告服务放置在页面上的链接。

当访问客户点击一个广告链接时，eBay 会获得一个小额报酬，而当他们真正从一个商城卖家购买商品时，eBay 将会获得一个更大额的报酬。经常实际购买的客户只给他们显示指向 eBay 卖家的链接。经常浏览而不购买的客户会看到更多的广告链接。对于这些客户，广告点击的期望值会高于链接向卖方的期望值。

2.7.2　减少信用风险

学会避免不良客户(并注意到好的客户将要变坏)与抓住好的客户同样重要。由于暴露在消费者信用风险之下，因此大多数公司会把消费者信用筛选(credit screening)作为获取过程的一部分，但风险建模在获取客户之后不会结束。

1. 预测谁将违约

评估现有客户的信用风险，对于任何向客户提供延后支付的业务而言都是一个问题。总有一些客户在接受服务之后，不对其进行支付。不偿还债务是一个明显的例子。报纸订阅、电话服务、电和煤气以及有线服务是众多通常只在使用之后支付的服务中的一些例子。

当然，若客户在足够长的时间内不支付，则最后会被终止服务。到那时候，他们可能会欠下一大笔钱，而这笔钱必须得注销。根据预测模型的早期预警，公司可以采取措施来保护自己。这些措施可能包括限制获得服务的机会，或者减少延迟还款和中止服务之间的时间长度。

有时会把未付款而导致服务终止称为非自愿流失，我们可以采取多种方法对其进行建模。通常，非自愿流失被视为一个二分结果，诸如逻辑回归分析和决策树之类的技术很适合这种情况。第 10 章将说明这个问题也可以被视为一个生存分析问题，实际上，把问题“客户下个月是否会不支付？”转换成“因为非自愿流失而损失一半客户需要多长时间？”

自愿流失和非自愿流失的很大区别之一在于非自愿流失往往涉及复杂的业务流程，例如账单因为经过不同阶段而延迟(这些通常被称为邓宁水平(Dunning Level)，以纪念 IBM 的一位研究人员，其最早开发出处理延迟支付客户的自动化技术)。最好的方法通常是建模业务流程中的每个步骤。

2. 改进催缴

当客户已经停止支付，数据挖掘可以辅助催缴。模型用于预测可以催缴的数量，以及在某些情况下帮助选择催缴策略。催缴基本上是一种销售。公司尝试向拖欠债务的客户出售其应支付的账单，而不是其他一些账单。与任何其他的销售活动一样，有些潜在的支付者会倾向于接受某种类型的信息，而其他一些倾向于另一种类型的信息。

2.7.3 确定客户价值

在客户价值计算方面，数据挖掘发挥了重要作用，尽管这种计算还需要获得正确的财务定义。客户价值的一个简单定义是在一段时间内该客户的总收益减去维护客户的总成本。但多少收益应归功于一个客户？是他到某个时间点的总开支吗？他这个月的开支是多少或预计在下一年的开支是多少？间接收益如广告收入和名单租赁等应该如何分配给客户？

成本甚至会产生更多问题。业务具有各种各样的成本，它们可能会以独特的方式分配给客户。把客户无法控制的成本归咎于他们是否公平？两个 Web 客户订购完全相同的商品，并且都是保证免费送货。其中生活得离仓库更远的那一个客户可能要花更多的运输成本，但是他真的是价值更低的客户吗？如果下一个订单是从不同的位置发货呢？随着无处不在的全国统一利率计划，移动电话服务提供商正面临着同样的问题。当提供商并不拥有整个网络时，他们的成本远不能统一。其中一些呼叫是经过该公司自己的网络，而其他一些呼叫会在收费利率更高的竞争对手的网络上漫游。通过设法劝阻客户在访问提供商成本较高的州时呼叫，该公司可以提高客户价值吗？

在整理出所有这些问题之后，而且一家公司已认可可追溯性的客户价值定义时，为了估计潜在客户的价值，数据挖掘可以发挥作用。其实际上是估计单位时间内客户带来的收益，然后估计客户的剩余生存周期。其中的第二个问题是第 10 章的主题。

2.7.4 交叉销售、追加销售和推荐

对于现有的客户，客户关系管理系统的一个主要关注点是通过交叉销售和追加销售提

高客户的收益率。数据挖掘用于找出对谁提供什么，以及什么时候提供。

1. 找到正确的提供时间

Charles Schwab(一家投资公司)发现客户通常以数千美元开立账户，即使他们有更大一部分用于储蓄和投资账户。当然，Schwab 希望吸引其余的钱。通过分析历史数据，分析师发现：把大笔盈余转入投资账户的客户，这么做通常是在开设他们第一个账户之后的头几个月内。几个月后，试图让客户移入大笔盈余的可能性很小。窗口已经关闭了。因此，Schwab 改变了策略，不再在客户的整个生存周期不断地发送信息，而是把精力集中在最初几个月。

2. 推荐

一种交叉销售的方法是使用关联规则，这是第 15 章的主题。关联规则用于查找产品组，它们通常可以一起出售，或者同一个人会在一段时间内购买它们。若客户购买了其中的一些产品，但并非该组中的所有元素，则对于缺失元素而言，他是好的潜在客户。这种方法对于零售产品有效，在那里可以找到许多这样的群集。相同思想在其他领域的巧妙应用也很有效，例如产品更少的金融服务领域。

2.8　保留

对于任何公司而言，客户流失都是一个重要问题；而且对于成熟的行业尤为重要，因为其初期阶段的指数性增长已经过去。毫不奇怪，流失(或者从积极的角度来看，是保留)是数据挖掘的一个主要应用。

2.8.1　识别流失

建模流失(attrition)的挑战之一是明确它是什么，并了解它何时会发生。有些行业相比其他行业会更难。一个极端的例子是处理匿名现金交易的业务。当一个曾经忠诚的客户放弃他经常喝咖啡的酒吧，去街区南边的另一家酒吧时，记牢客户订单的咖啡师可能会注意到，但是这个事实不会被记录到任何公司的数据库中。即使在按名称标识客户的情况下，区分已经流失的客户和从未动摇过的客户之间的差异也可能很困难。如果一个忠诚的福特公司(Ford)客户每五年会购买一辆新的 F150 小货车，但是他在第六年没有买，那么该客户是否已经流失到另一个品牌了？

当存在月度账单关系(如信用卡)时，流失会更容易被发现。即便如此，流失也可能会悄无声息。客户可能停止使用信用卡，但并未取消。流失在基于订阅的业务中很容易定义，部分因为这个原因，这些业务中的流失模型最受欢迎。长途公司、移动电话服务提供商、保险公司、有线电视公司、金融服务公司，Internet 服务提供商、报纸、杂志以及某些零售商都共享一个订阅模型，其中客户有一个正式的、必须显式结束的合同关系。

2.8.2 为什么流失是问题

损失的客户必须用新的客户来替代，而获取新的客户很昂贵。通常，新的客户在近期内所产生的收益比已建立的客户要少。对于市场相当饱和的成熟行业而言尤其如此——其中，想要拥有产品或服务的人可能都已经拥有了它，因此新客户的主要来源是离开竞争对手的人。

图 2-11 说明了随着市场逐渐饱和，以及获取活动的响应率逐渐下降，获得新客户的成本将上升。该图显示了每个新的客户在直接邮寄获取活动中的成本，假定邮寄成本为 1 美元，而且它还包括某种形式的 20 美元优惠，比如一张优惠券或降低信用卡的利率。若获取活动的响应率高，如 5%，那么一个新客户的成本是 40 美元。(送抵 100 人的成本为 100 美元，其中 5 位响应的成本是每人 20 美元。因此，获取 5 个新客户的成本为 200 美元)。随着响应率下降，成本会迅速地增加。当响应率下降到 1%时，每个新客户的成本是 120 美元。在某些时候，花那些钱来保留现有客户比用来吸引新的客户更有意义。

保留活动有效，但也很昂贵。移动电话公司可能向延长合同的客户提供一款昂贵的新电话。信用卡公司可能会降低利率。这些优惠的问题在于得到优惠的客户都将接受它。谁不想要一个免费电话或较低的利率？许多接受该优惠的人无论如何都将一直保持为客户。构建流失模型的动机是要找出谁最有可能流失，从而向高价值的客户提供优惠，因为如果没有额外的奖励他们可能会流失。

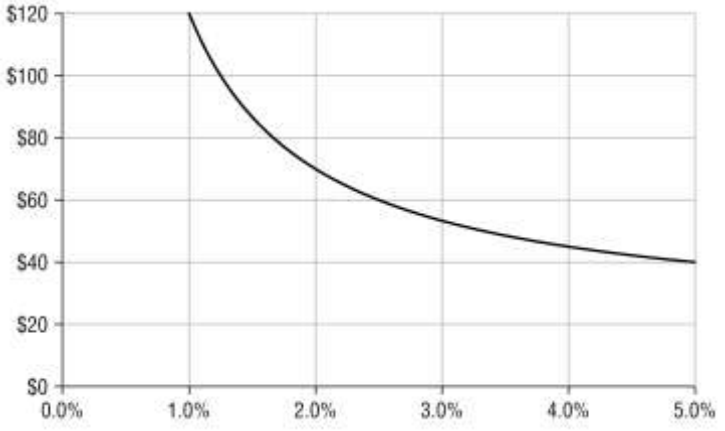


图 2-11 随着获取活动的响应率下降，获取每个客户的成本在上升

2.8.3 不同类型的流失

前面讨论了为什么流失问题会集中在自愿类型。客户根据他们自己的自由意志，决定把他们的业务放在其他地方。这种类型的流失，称为自愿流失，实际上是三种可能性之一。其他两种分别是非自愿流失和预期流失(expected attrition)。

非自愿流失，也称为强制流失(forced attrition)，发生在公司而非客户终止关系时——最常是因为未支付账单。预期流失在客户不再是产品的目标市场时发生。婴儿长牙齿后不再需要婴儿食品。家庭搬家后更换有线电视提供商。

不混淆不同类型的流失很重要，但也很容易做到。考虑两个相同财务状况下的移动电话客户。由于一些不幸，两个客户都不能继续接受移动电话服务。两个电话都被取消了。其中一个通知了客户服务代理，其被记录为自愿离开。另一个在等待十分钟之后挂掉电话，并在不支付账单的情况下继续使用这个电话。第二个客户是被迫离开。两个客户的根本问题——缺乏资金——都是相同的，所以他们很可能将得到类似的分数。模型不能预测两个用户所体验的持有时间方面的差别。

公司把强迫流失误判断为自愿流失将导致两次损失——一次是当他们花钱试图保留后来变坏的客户，另一次是增加了注销代价。

对强制流失不准确的预测也可能会很危险。因为对不可能支付账单的客户的处理往往令人不快——停止电话服务、增加滞纳金，并更快地发送催缴信。这些补救办法可能会疏远其他的好客户，增加他们自愿离开的机会。

从数据挖掘的角度看，同时解决自愿和非自愿流失会更好，因为所有的客户都不同程度地存在这两种风险。

提示：当对流失建模时，对所有类型的流失进行建模是一个好主意。在他们彻底倒向其中一种或另一种流失之前，用户都处于非自愿和自愿流失的风险。对某一种风险得分较高的客户可能(或可能不)对其他的风险同样具有较高的分数。

2.8.4　不同种类的流失模型

对流失建模有两种基本方法。第一种是把流失看成是二分结果，其中客户将离开或者留下。第二种试图估计客户的剩余生存周期。

1. 预测谁会离开

为了把流失建模成二分结果，必须选定某个时限。如果问题是：“明天谁会离开？”，答案是几乎没有人。但是如果这个问题是：“谁将在下一个百年里离开？”，那么在大多数企业答案几乎是每个人。二分结果流失模型通常有一个较短的时限，例如 60 天或 90 天，或者一年。当然，时限不能太短，否则将没有时间实施模型预测。

可以使用任何通常用于分类的工具来建立这种模型，包括逻辑回归分析、决策树和神经网络等。描述某一刻客户人口的历史数据将带有一个标志，以显示客户在某个后续时刻是否依然是活跃的。建模任务是区分哪些客户会离开以及哪些会留下。

这种模型通常会根据客户离开的可能性对他们打分并排序。最自然的得分是简单地使用模型，得出客户在某个时期内离开的可能性。那些自愿流失得分超过某个阈值的客户将被包含在一个保留方案中。那些非自愿流失得分超过某个阈值的客户将放在一个观察列表中。

通常，流失预测器是基于一组客户的混合信息，包括在获取时所了解的客户信息，如获取渠道和初始信用类别；在客户关系期间发生的事情，例如服务相关的问题、延迟还款和意外的高账单或低账单；以及客户的人口统计信息。第一类流失驱动给出了如何通过少获取易于流失的客户来降低未来的客户流失的信息。第二类流失驱动提供了如何减少已存在客户的风险的洞察力。

2. 预计客户将保留多长时间

流失建模的第二种方法是生存分析，详见第 10 章。其基本思想是计算每个客户(或者每组客户，他们具有相同的模型输入变量的值，例如地理、信用等级以及获取渠道等)到目前为止将在明天之前离开的可能性。对于任何阶段，这种所谓的灾难可能性(hazard probability)都相当小，但是某些阶段的可能性会高于其他阶段。通过干预灾难，能够估算出客户能够生存到某个更远的未来日期的机会。

2.9 超越客户生存周期

数据挖掘很自然地适合于客户生存周期。然而，并非所有的数据挖掘应用都直接与生存周期相关联。例如，预测通常是一个关键的业务流程。您可使用数据挖掘预测客户的数量和未来的流失率，同样可以使用它对客户进行划分，甚至发现意外的客户类型或行为。

第 11 章给出了一个案例研究，分析如何区分客户投诉与其他类型的客户评论。这是一个将文本挖掘应用到客户关系管理的示例。并非所有的数据挖掘应用都处理客户数据。第 21 章包含了文本挖掘的示例，其中一个示例是为新闻故事指定关键词；关键词可以帮助读者发现需要的故事。第 12 章描述了 Boston Globe 的一个数据挖掘项目，其关注的是整个城镇而非个别订户。根据人口的相似性对城镇进行聚类。然后，将这些人口簇与地理邻接性相结合以创建不同的区域，针对它们定制报纸的版本。

数据挖掘有许多应用。因为客户对于所有的业务都是相同的，因此以客户为中心的应用也最常见。本书所介绍的技术已经被用于客户关系管理系统以及其他系统。

2.10 经验教训

在本书的大部分内容中，以客户为中心的应用程序作为重点是隐含在用于说明技术的示例选择上。但是，在本章该重点更为显式化。客户关系遵循一种自然的生存周期，开始于潜在客户和客户的获取，接着是激活，然后继续一段扩展周期，管理与已建立的客户的关系。客户关系管理系统的部分工作是，努力保持这些已建立的客户，同时设法赢回已流失的客户。

客户关系所有阶段所生成的数据都可以用于挖掘。在获取阶段，数据挖掘同时支持广告和直接营销以识别正确的受众，选择最佳的通信渠道，以及挑选最适当的信息。潜在的客户可以与预期受众的剖析进行比较，并给出一个适应度得分。如果潜在客户的个人信息不可用，那么您可以运用相同的方法对地理上的社区指定适应度得分，利用可以从美国人口普查局、加拿大统计机构以及许多国家类似的官方来源等得到的类型的数据。

数据挖掘在直接建模中的一个常见应用是响应建模。响应模型对潜在客户响应直接营销活动的可能性进行了打分。该信息可以用于提高活动的响应率，但是它自身不足以决定活动的收益率。估计活动的收益率需要依赖于对未来活动基本响应率的估计，与响应相关联的平均订单大小估计，以及落实和活动本身的成本估计。对响应得分的进一步以客户为

中心的使用，是为每个客户从一些相互竞争的活动中选择最佳活动。这种方法避免了通常在独立的、基于得分的活动中会出现的问题，它们倾向于每次选择相同的人。

在一个模型识别对产品或服务感兴趣的人的能力，以及该模型识别因为特定活动或优惠而发生购买行为的人的能力之间进行区分非常重要。增量响应分析提供了一种方法，以识别活动将在其中产生最大影响的市场分块。

通过从当前客户在成为客户之前的已知信息中发现期望结果预测器，公司可以使用当前的客户信息来确定可能的潜在客户。这种分析对选择获取渠道和联系策略，以及筛选潜在客户列表都非常有价值。通过从客户第一次响应(甚至在他们成为客户之前)时，开始跟踪他们，同时当获取客户时收集和存储额外信息，企业可以提高客户数据的价值。

在已经获取客户之后，重点将转移到客户关系管理系统。可用于活跃客户的数据比可用于潜在客户的数据更为丰富，因为它本质上是行为数据而不仅仅是地理和人口数据，它将更具预测性。数据挖掘能够基于客户当前的使用模式，确定应该向客户提供的额外产品和服务。它还可以建议进行交叉销售或追加销售的最佳时间。

客户关系管理方案的目标之一是保留有价值的客户。数据挖掘可以帮助识别哪些客户是最具价值的，同时评估与每个客户相关联的自愿或非自愿流失风险。基于这一信息，公司可以把保留工作的目标定位为既有价值又有风险的客户，并采取步骤保护自己远离可能违约的客户。当把流失模型构造成“客户什么时候会离开？”时，该模型可用于估计客户价值。

数据挖掘是一种贯穿整个客户生存周期的有效工具。下一章将从数据挖掘如何辅助业务转移到在业务环境中实现数据挖掘的挑战上。

第 3 章

数据挖掘过程

第 1 章将数据挖掘的良性循环描述为一个业务流程，其中把数据挖掘划分为 4 个阶段：

(1) 识别问题

(2) 将数据转换为信息

(3) 采取行动

(4) 度量结果

本章的重点转向把数据挖掘作为技术过程，把识别业务问题转变为将业务问题转化为数据挖掘问题。同时，第二个阶段——把数据转换为信息，将扩展到几个主题，包括假设检验(hypothesis testing)、模型构建(model building)和模式发现(pattern discovery)。本章所介绍的思想和最佳实践将在本书的剩余部分进一步阐明。本章的目的是为了在一个地方把不同类型的数据挖掘放在一起介绍。

避免打破数据挖掘良性循环的最好方式是了解它可能会失败的方式，并采取预防措施。多年来，本书作者曾遇到过数据挖掘项目发生错误的许多方式。本章一开始讨论其中的一些陷阱，剩余部分介绍数据挖掘的过程。后续章节涵盖了数据挖掘方法的各个方面，它们特定于不同类型的数据挖掘——有指导数据挖掘(directed data mining)和无指导数据挖掘(undirected data mining)。本章重点介绍这些方法的共同之处。

介绍了三个主要的数据挖掘类型，首先是简单的方法——通常通过使用特设查询检验假设，接着介绍更加复杂的活动，例如可用于评分的模型构建，以及使用无指导数据挖掘技术的模式发现。这一章的主题是从业务目标的明确说明，转移到为实现目标所需的对数据挖掘任务的明确理解，以及适合该任务的数据挖掘技术。

3.1 会出什么问题

数据挖掘是一种从过往中学习以便在将来更好地做决定的方法。本章中描述的最佳实践旨在避免两个不可取的学习过程结果：

- 学习的东西不真实。

- 学习的东西为真，但是无用。

古代水手学会了如何避免为保护西西里和意大利大陆之间狭窄海峡的锡拉岩礁(Scylla)岩石和卡律布迪斯(Charybdis)漩涡。像学会避免这些威胁的古代水手们一样，数据挖掘人员也需要知道如何避免常见的危险。

3.1.1 学习的东西不真实

学习的东西不真实比学习的东西无用更危险，因为可能会基于不正确的信息做出重要的业务决策。数据挖掘结果经常看上去是可靠的，因为它们是基于实际的数据，以看似科学的态度进行处理。这种可靠性看上去非常具有欺骗性。数据可能不正确或者与手头的问题无关。发现的模式可能反映了过去的业务决策或者什么都不是。诸如汇总之类的数据转换可能会破坏或隐藏重要信息。下面将讨论一些会导致虚假结论的更常见的问题。

警告：当数据不正确或根本不相关时，即使采用最先进的技术，最认真细致的分析也会产生错误的结果。在信息技术的圈子里，有一个流行的格言：“无用输入，无用输出(garbage in，garbage out)”。

1. 模式可能不代表任何基本规则

俗话说：数字不会说谎，但说谎者会玩弄数字。当在数据中发现模式时，数字不会为了提议不真实的东西而说谎。可用来构造模式的方法如此众多，从而给定任何随机的数据点集合，只要检查足够长的数据，总能发现一个模式。人类严重依赖于我们生活中的模式，即使它们不存在我们也会倾向于看到它们。我们仰望夜空，映入眼帘的并非星星的随意排列，而是北斗七星、南十字星座或猎户座腰带。有人甚至看到可用来预测未来的占星模式和预兆。普遍接受的古怪阴谋论是人类需要发现模式的进一步证据。

人类演化成对模式如此亲切的原因可能在于模式通常反映了世界运转的一些基本事实。月相、季节变幻、昼夜不断交替，甚至最喜爱的电视节目定期在一周同一天的同一时间播放也是有用的，因为它们稳定从而可预测。我们可以使用这些模式来决定什么时候种植番茄是安全的，什么时候吃早餐，以及如何安排 DVR。其他模式显然没有任何预测能力。如果一枚硬币连续出现 5 次头向上，那么在第 6 次它仍有 50%的机会掷出尾朝上。

数据挖掘人员的挑战在于要找出哪些模式有益，哪些无益。考虑以下模式，所有这些模式都曾在大众媒体文章中被引用过，就像它们有预测价值：

- 在野党(总统竞选失败的政党)在非大选年选举中会获得国会席位。
- 当美国职业棒球联盟赢得世界职业棒球大赛时，共和党会继续把持着白宫。
- 华盛顿红人队赢得最后一场主场比赛时，执政党继续执掌白宫。
- 在美国总统竞选中，高个子的男人常常会赢。

第一个模式(涉及非选举年选举的那一个)在纯粹的政治术语下是可解释的。每隔四年，略高于半数的美国选民都很兴奋地投票选举总统候选人。几个月后，候选人接管而失望就此开始——政客们根本不能兑现选民所期望的所有承诺。两年后的国会选举会出现反弹，通常是由感到失望的支持者不投票导致的结果。因为该模式有一个基本的解释，似乎很可能会持续到将来，表明它有预测价值。

而后两个所谓的预言——涉及体育赛事的那两个预言，看上去明显没有预测价值。无论共和党人和美国联盟在过去可能已经共享过多少次胜利(作者没有研究过这一点)，但是没有理由可以预测在将来会继续关联。

那候选人的高度呢？自 1948 年杜鲁门(其身材矮小，但比杜威高)当选之后，只有卡特击败福特和布什打败克里是较为矮小的候选人赢得普选的两次选举。在 2000 年的选举中，如果我们假设模式是与赢得普选而非选举人票相关，那么戈尔的 6 英尺 1 英寸对布什总统的 6 英尺还是符合该模式。在 2008 年，打篮球的奥巴马击败了较为矮小的麦凯恩。高度看上去与当总统这份工作毫不相关。然而，我们的语言展示了“身高歧视”：我们把仰视看成是表示尊敬的姿态，而俯视表示蔑视。身高与更好的童年营养相关，其反过来会提高智商以及其他社会成功的指标。如本章所述，决定规则是否稳定和具有预测性的正确方式是比较它在多个从同一总体中随机选择的样本中的性能。关于总统高度的案例，作者把它留给读者作为练习。通常情况下，任务中最难的部分是收集数据——在 Google 年代之前，确定过去几个世纪以来不成功总统候选人的高度很不容易。

发现的模式无法泛化的术语是过拟合(overfitting)。过拟合会导致模型不稳定，它可能今天工作但是明天不工作；或者在一个数据集上工作，但在另一个数据集上不工作。建立稳定的模型是数据挖掘方法的主要目标。

2. 模型集可能不反映相关的总体

模型集(model set)是用来创建数据挖掘模型的数据，它必须描述过去发生了什么。模型的质量只能与用来创建它的数据一样。为了使推理有效，模型集必须反映模型用来描述、分类或者评分的总体。若对没有正确地反映总体的样本进行评分，则会导致总体有偏置(biased)。

有偏置的模型集会导致学习的东西不真实。除非考虑了偏置，否则结果模型也会有偏置。偏置很难避免。考虑：

- 客户不像潜在客户。
- 调研响应者不像非响应者。
- 阅读电子邮件的人不像那些不阅读电子邮件的人。
- 在网站上注册的人不像那些不注册的人。
- 获取之后，来自获取公司的客户不一定像获取需求方的客户。
- 没有缺失值的记录与具有缺失值的记录反映了不同的总体。

考虑第一点。客户不像潜在客户，因为他们代表了那些对任何过去用来吸引客户的信息、优惠以及促销都积极响应的人。对当前客户的研究也可能会给出相同的提议。如果过去的活动都是针对富有的城市消费者，那么当前客户与一般人群的任何比较都很可能会显示客户往往是富有的城市人群。这样的模型可能会错过在中等收入郊区的机会。

提示：仔细选择和采样数据集对于数据挖掘的成功至关重要。

使用有偏置的样本比只是错失营销机会的后果更为糟糕。在美国，有一个“歧视”的历史，在某些社区(通常是低收入社区或少数群体社区)存在拒绝写贷款或保险政策的非法做法。对于一家有歧视历史的公司，如果从它的历史数据中搜索模式，那么将显示某些社

区的人根本不可能是客户。如果未来的营销工作是基于这样的发现，那么数据挖掘将导致非法和不道德做法延续下去。

3. 数据的详细程度有误

在多个行业中，作者曾经被告知在客户离开之前的那个月内使用量通常会下降。经过仔细观察之后，这可能是一个学习的东西不真实的例子。图 3-1 显示了一组移动电话订户每月使用的分钟数，这些用户被记录为在 9 月份停止使用。前 7 个月，订户每个月平均使用约 100 分钟。在第 8 个月时，其使用量下降为约一半左右。再下一个月，则根本不存在使用，因为订户已经停止了。这说明由使用下降触发的营销努力也许能够保存这些客户。

这些订户似乎符合在放弃服务之前一个月减少使用的模式。表象往往具有欺骗性。这些客户在第 9 个月没有使用是因为实际的停止日期是在第 8 个月。平均而言，停止日期是这个月的中间。在停止之前，这些客户继续以恒定的速率使用服务，大概是因为在那一天，客户开始使用与其竞争的服务。假定的使用下降区间实际上并不存在，因此当然不会为保留客户提供一个机会之窗。这似乎是一个领先指标，但实际上却是一个滞后指标。

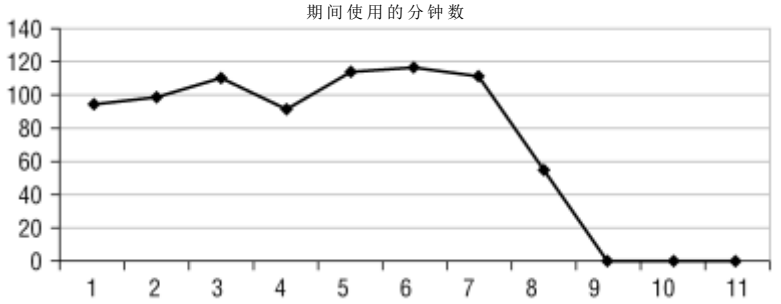


图 3-1　第 8 个月的使用下降是否能够预测在第 9 个月的流失

图 3-2 显示了另一个由聚合引起混乱的例子。与 8 月份和 9 月份相比，10 月份的销售似乎下降了。图片来自一家只有当金融市场开放的日子才有销售活动的企业。因为在 2003 年周末和假日的安排方式，10 月的贸易日比 8 月和 9 月的要少。该事实仅仅考虑了销售的整体下降。

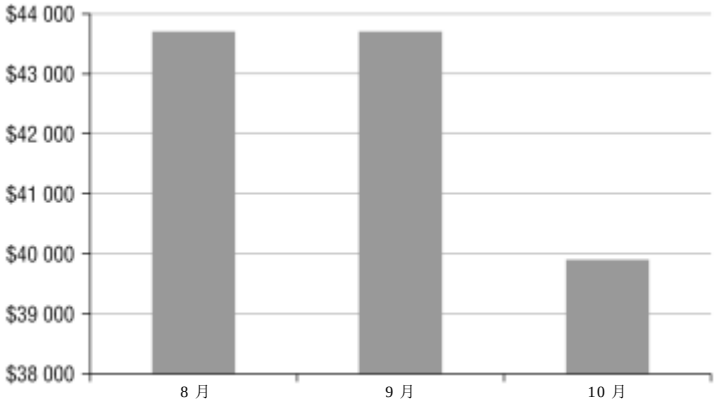


图 3-2　10 月份的销售是否真的下跌了

在前面的示例中，聚合导致了混淆。没有聚合到适当的级别也可能会导致混淆。一个家庭成员可能会有一个余额较低和无活动的支票账户，而相同家庭的另一个成员有多个大账户。把小账户持有者看做是低价值的客户可能会导致与整个家庭的关系处于危险境地。在这种情况下，总的余额数可能比其中任何一个账户的余额更为重要。

提示：当汇总数据时，选择合适的聚合级别，使得不会隐藏在一段时期内的重要模式。一个每周具有很强变化的企业不应报告每月汇总的活动。

3.1.2　学习的东西真实但是无用

尽管不如学习的东西不真实那么危险，但是学习的东西无用更为常见。它可能会以几种不同的方式发生。

1. 学习的东西众所周知(或者应该知道)

数据挖掘应该提供新的信息。在数据中许多最强的模式都表示了众所周知的事情。过了退休年龄的人往往不会响应退休储蓄计划的优惠。生活在送货区之外的人们不会成为报纸订户。尽管他们可能响应订阅优惠，但是永远不会启动服务。自己不拥有车的人不会购买汽车保险。

数据挖掘还可以发现那些应该已经知道为真的模式。在一个有趣的例子中，本书作者们正致力于一个分析杂货店数据中的购买模式的项目。当第一组数据到达时，我们首先发现会被同时购买的产品。首批组合包括“鸡蛋和肉”、“鸡蛋和牛奶”以及“鸡蛋和苏打水”。后续的规则类似——鸡蛋忙于与杂货店里的每款产品同时下架。首先，这似乎是数据中的一个潜在问题。其次，我们的同事之一注意到该数据来自复活节前一周。事实上，当人们在复活节之前去杂货店购物时，他们常常会为复活节假期购买鸡蛋进行涂色或者隐藏。

最强的模式通常会反映业务规则。如果数据挖掘“发现”具有匿名呼叫阻止服务的人也有来电显示，那么原因可能是因为匿名呼叫阻止只是作为包括来电显示在内的捆绑服务的一部分。如果数据挖掘“发现”维修协议与大型设备一起出售(如 Sears 曾经发现的一样)，那么这是因为维修协议几乎总是在设备之后出售。这些模式不仅无趣，而且它们的强度可能会阻碍不太明显但更具可操作性的模式。

不过，学习的东西众所周知确实会起到一个有益的作用。它表明，在技术层面上数据挖掘技术有效，而且数据相当准确。即使没有帮助，这也多少会令人有所安慰。当数据挖掘技术足以发现众所周知是真实的事情时，那么有理由相信它们也能发现更有用的模式。

2. 学习的东西不可用

数据挖掘能够发现真实且先前未知的关系，但是仍然难以使用。有时这个问题是可控的。客户的无线呼叫模式可能与某些固定长途包类似，但是由于法律限制，提供这两种服务的公司可能不允许利用这个事实。类似地，客户的信用记录对于未来的保险索赔可能具有预测性，但是监管机构可能禁止基于这种信息做出承保的决定。或者，在一个更普遍的例子中，一个人的遗传物质可能会表明对某些疾病的倾向性——在美国和大多数欧洲国家的保险公司都禁止使用这一特征。

其他时候，数据挖掘发现的重要结果超出了公司的控制范围。相比其他气候，产品可能更加适合于某些气候，但是天气很难改变。由于地形的缘故，移动电话服务在一些地区可能会更糟，但是这也很难改变。

提示：客户流失的研究表明，客户离开的一个重要预测点是获取他们的方式。对于现有客户而言，这个发现已经太迟而不能倒回去改变，但是这并非无用的信息。通过改变混合获取渠道为能带来长期持续客户的渠道，可以减少未来的流失。

数据挖掘人员必须注意避开锡拉岩礁，以免学习的东西不真实，以及避开卡律布迪斯，以免学习任何无用的东西。第 5 章和第 12 章介绍的方法旨在确保数据挖掘的努力会导致稳定的模型，从而能够成功地解决业务问题。

3.2 数据挖掘类型

第 1 章定义数据挖掘为“探索和分析大量数据以产生有意义的结果”。这是一个广泛的定义，足以包括许多不同的方法。主要有三种类型：

- 假设检验(Hypothesis testing)
- 有指导数据挖掘(Directed data mining)
- 无指导数据挖掘(Undirected data mining)

假设检验的目标是使用数据来回答问题或掌握知识。有指导数据挖掘的目标是构建一个模型，其能够解释或预测一个或多个特定的目标变量。无指导数据挖掘的目标是找到全部的模式，而不是绑定到一个特定的目标。在一个数据挖掘项目过程中，你可以采用这些类型中的任何一种类型或者所有类型，这取决于问题的性质和你对数据的熟悉程度。

虽然这三种类型的数据挖掘有一些技术性的差异，但是它们也有很多共同之处。在第 5 章有指导数据挖掘上下文下讨论的许多主题对于假设检验和发现模式也非常重要。事实上，有指导数据挖掘方法的前三步——把业务问题转化成数据挖掘问题，选择适当的数据，以及了解数据——也同样可以在本章介绍。

3.2.1 假设检验

假设检验几乎是所有数据挖掘努力的一部分。数据挖掘人员通常在以下两个方法之间来回处理，首先构造出对观察行为的可能解释(通常借助业务专家的帮助)，同时让那些假设来决定待分析的数据，然后让数据提出新的假设来检验。

假设是提出的一个解释，可以通过分析数据检验其有效性。这些数据可以简单地通过观测收集或者通过试验(如一个营销活动测试)来生成。假设检验有时会发现指导公司行动的假设不正确。例如，一家公司的广告是基于对产品或服务目标市场以及响应性质的一些假设。这些假设能否通过实际的响应来证明是值得检验的。

取决于假设的不同，这可能意味着解释从一个简单查询返回的单一值、费力地阅读通

过购物篮分析生成的关联规则集合、决定回归模型发现的相关的显著性，或者设计对照实验。在所有情况下，有必要仔细思考以确保结果不会以意外的方式发生偏置。对数据挖掘结果的适当评价需要分析知识和业务知识。由于这些评价不是同一个人所能给出的，因此好好利用新的信息需要跨职能的合作。

本质上假设检验是特设的，但是这个过程有一些可辨认的步骤，第一步也是最重要的一步在于产生好的假设来检验。接下来是查找或生成数据以证实或推翻假设。

1. 生成假设

生成假设的关键在于从整个机构获得多样性的输入，并且在适当的时候，也可以从机构外部获得数据。局外人可能会质疑内部人员认为理所当然的事——因此可能提供宝贵的见解。通常，启动思想交流所需的一切就是对问题本身的一个明确说明——如果它在以前不被认为是一个问题的话，尤其需要如此。

超出人们想象的是，问题不被重视的原因通常在于它们没有被用来评估性能的度量指标所覆盖。如果公司一直根据每个月新完成的销售量来考察它的销售人员，那么推销员可能从来就不会太多地思考这样的问题：新客户会在多长时间内保持活跃，或者他们会在维持关系过程中花多少钱。但是，当被问到正确的问题时，销售人员可能会洞察营销(由于其与客户的更大距离)所漏掉的客户行为。

目标是提出既可检验又可操作的想法。考虑以下假设：

- 大多数接受保留优惠的客户无论如何将保持。
- 拥有高中阶段孩子的家庭更可能比其他家庭对家庭净值贷款提议作出响应。
- 购买更多不同产品类型的客户有较高的总体开支。

所有这些主张可能是或可能不是真实的，并且在每种情况下，知道答案会暗示一些具体的行动。如果第一个假设是真的，那么停止花钱来吸引没有离开风险的客户，或者发现一种更好的方式，把保留优惠定位到真正打算离开的客户。如果第二个假设是真的，那么继续围绕这一群体开展目前的营销。如果第三个假设是正确的，那么鼓励销售人员进行更多的交叉销售。

2. 已有数据检验假设

常常可以通过在已有的历史数据中寻找证据来检验新的假设。例如，把医疗设备卖给医院的制造商会有如下假设：购买多种类型产品的客户往往会花更多的钱。首先，他们查看具有不同产品数量的平均销售，生成如图 3-3 所示的图表。

该图清楚地表明购买多种产品的客户会生成更大的平均客户收益，但是它没有显示交叉销售会在多大程度上驱动额外收益。较大的机构自然会花更多的钱，也许他们还更有可能需要多个类别的产品。或许高收益和多个产品类别都是由客户大小所决定的——这不是公司的控制范围。同样有一个可检验的假设：根据大小和类型对客户分组，并寻找每组客户中不同产品和收益之间的关系。

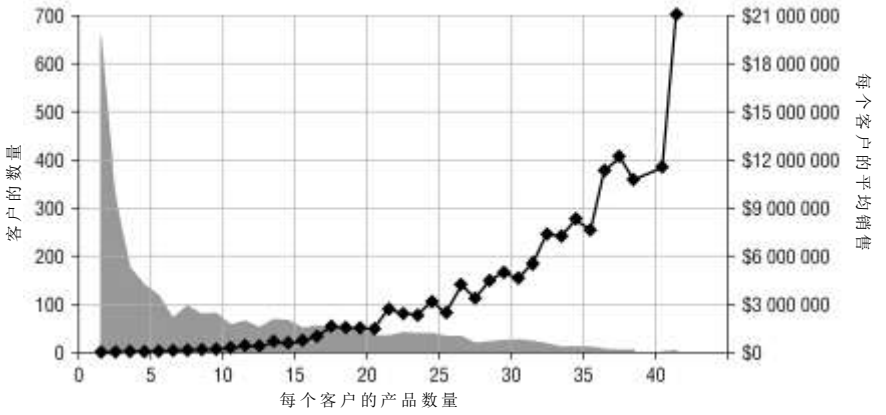


图 3-3　购买更多产品类型的客户会花更多的钱

检验长期持有的信念可能会更难，因为历史数据反映了过去所给出的任何假设。如果拥有高中孩子的家庭一直是某一特定产品的目标，那么这一事实将在这些家庭较高的采用率上得到反映。这并不能证明他们是最可能响应的群体；其他一些团体，如小生意人，甚至会产生更多的响应。在这种情况下，进行对照实验是可取的。

在收集什么以及如何收集数据方面的细小变化能够极大地增加其分析的价值。例如，使用不同的 Web 地址或在不同广告中的呼入数，以及跟踪每个响应的渠道。

提示：每次当公司请求其客户的响应时，无论是通过广告或更直接的通信形式，都有机会收集信息。在通信设计方面的细微变化，例如包含一种方法来识别当潜在客户作出响应时的渠道，都可以极大地提高收集到的数据价值。

3. 假设检验和实验

虽然许多假设可以根据历史数据进行检验，但是还有许多假设不能如此检验。例如，我们采取如下假设：接受了一个保留优惠的客户将在无论是否有额外诱惑时都将保持。历史数据给出了谁收到了优惠、谁接受了优惠，以及谁最终离开了，但是除非活动设计为一个带有对照组的适当实验，否则它不会回答如果不提供优惠会发生什么的问题。这个问题不能通过比较收到和没有收到优惠的客户的保留情况来回答，因为这两个组几乎肯定会存在系统方法上的不同。

如果向那些被认为具有高流失风险的客户提供优惠，那么即使优惠确实保住了一大批客户，但是没有得到优惠的人可能依然会有更好的保留。另一方面，如果向被认为特别有价值的客户提供优惠，那么他们可能会比因为某些原因而对优惠不做任何事的未接收者具有更好的保留。除了测试的事情之外，该方案有效性的合理测试需要比较两组在各方面都类似的客户。那样的数据可能不会自然产生，所以你必须设计实验来生成它。实验设计和分析是一个广泛的统计领域。本小节介绍常见营销测试细节中的一些关键点。

1) 检验和控制

最基本的实验设计需要创建两个组。其中一组称为检验组(test group)或处置组

(treatment group)，其会接受某种处置，如电子邮件或电话等。另一组称为控制组(control group)，不接受处置。选择的两个组尽可能相似——相同的平均年龄、相同的平均收入、相同的男女分布以及相同的客户持续期分布等。这听起来可能会很困难，但其实不然。基本上，选择一个总体组，然后随机分为检验组和控制组。只要检验组和控制组足够大，概率法则会确保这两组彼此相似(且与总体相似)。如果你想要确保这些组代表一些关键特征(比如，性别和持续期)，那么根据这些字段对总体排序，然后每次把第 n 个记录放入控制组。

实验后，两个组之间的任何显著性区别都可以自信地归因于处置。第 4 章说明统计意义的显著性概念以及如何测试它。

2) A/B 检验

A/B 检验比较两个(或可能更多的)处置。客户随机分配到 A 组或 B 组。两个组接受不同的处置，如不同的广告信息、Web 页面布局、价格或者支付选项。面向分析的公司经常运行 A/B 检验以确定甚至看起来微小变化的影响，因为小的改变可以起到巨大和意料之外的作用。

一家网上零售公司发现添加一个客户可以输入优惠券代码的框，导致了实施购买的客户比例显著(6.5%)减少。大多数客户没有优惠券，显然邀请提供折扣代码会造成人们认为他们得到了一个坏的对待。也许这种购物者会被激励在 Google 上搜索一张优惠券，在这个过程中可能找到一个更好的价格。

A/B 检验通常与直接营销和基于 Web 的零售相关，因为在这些环境中控制哪些客户得到哪些信息相对简单。A/B 检验对于非直接类型的广告也有用，比如广告牌、广播、电视等。其中的技巧是在相似的市场运行不同的活动。这种试验称为配对测试(paired test)，因为它们依赖于为测试目的而尽可能相似的不同市场(或商店位置或任何其他事物)对。这一对中的一半得到处理，而另一半处于控制之下。第 9 章对配对测试进行了更详细的讨论。

3) 冠军/挑战者检验

A/B 检验的常见形式是比较新的处置(挑战者)和已有的处置(冠军)。这种思想常常是应用于对客户打分的数据挖掘模型。只有当新模型表明其比旧模型更好时才会被采纳。

Amazon.com 特别擅长这种形式的 A/B 检验。在其网站上的一切——从产品评论和产品说明的位置到用户评论的数量和关键字——都已经针对“冠军”最佳布局进行了测试。在 Amazon 的活动环境中，随机选择网站访问者作为测试组来查看一个修改布局。几个小时或几天后，收集的资料足以说明布局的测试修改是否会产生比冠军布局更高或更低的销售。如果改进非常显著，那么测试将成为新的冠军。

4. 假设检验中的案例研究：度量错误的事情

这是一家公司的故事，其为零售网站开发推荐软件。它的委托人，即零售商，在某些特定的 Web 页面上留下空白区域，如产品页面、购物车，以及结账页面等。当客户在网站购物时，该推荐软件会提供产品推荐来填补空白。当客户购买推荐的物品时，该软件公司获得佣金。当然，目标是提高网站的整体销售，这将有利于零售商，从而鼓励他们继续使用该推荐软件。

该软件公司遇到了一个难题：根据其所有的度量指标，它的推荐在年复一年地改进。更多的客户点击和购买了推荐的物品。然而，零售商抱怨收益的增长不如预期。在一些一对一测试(head-to-head test)中，这款复杂的推荐软件甚至不如委托人所开发的简单的通用规则。

这不是一个有指导数据挖掘的好问题。目标变量是什么？它也不是无指导模式发现的好对象，模式实在太明显了。它对于假设检验而言最适合不过。数据挖掘人员的工作是思考可能有什么问题，然后测试结果假设。

该软件公司请求数据挖掘人员(作者成立的咨询公司)来帮助搞清楚这个难题。我们获得了 A/B 检验的数据，其产生了令人失望的结果。在 A/B 检验中，随机选择一半消费者从该公司收到建议，而另一半收到来自零售商的竞争建议。该数据包括一个定单表，其中包含了每种物品的详细信息，如价格、产品类别，而且在消费者点击产品推荐时，还包括一个点击 ID。对于每个点击，点击表显示了是几个推荐算法中的哪一个生成了该推荐，以及当作出推荐时，消费者正在看什么物品。

使用简单的 SQL 查询，我们发现测试中我们委托人一方的客户的确点击了更多的推荐，而且，在双方的测试中，点击客户更有可能购买。更多的采购意味着赚更多的钱。更多的钱意味着零售商很高兴。

给定这些度量指标，A 方——我们的委托人一方——怎么会输呢？第一个线索是 A 方点击物品的平均价格比 B 方低。我们的第一个假设是 A 推荐的是与 B 不同的产品混合，但是这很容易被推翻。我们不断地尝试其他假说，直到最终找到了两个假设，它们在一起能够解释发生了什么：

- A 方的推荐产生了更多的替代，且交叉销售更少。
- 许多 A 方的推荐是向下销售。

交叉销售是指除了已经在考虑的产品之外，消费者还会购买推荐产品，从而产生更大的购买总额。替代是指消费者购买推荐产品以代替原来考虑的产品。交叉销售对零售商更有价值，因为它会增加客户的支出总额。然而，我们的委托方仅仅是基于最终的消费者是否购买其推荐产品。零售商设计的推荐是为了产生交叉销售。而且，当它确实在推荐替代时，其推荐的产品几乎都是更贵的东西——向上销售。通过比较，我们委托方的推荐平均来说是向下销售。

我们的结论是委托方度量了错误的事情。随着时间的推移，它的推荐在吸引更多的点击方面得到了改进，但是点击自身是无用的。为了吸引点击，最简单的方法是向消费者展示比他们正在看的物品更便宜的替代。这种行为为我们的委托方生成佣金，但是(无意中)导致零售商终止销售更为便宜的物品，以及终止为此优惠支付佣金。我们建议软件公司改变佣金结构，从而它可以根据增量收益而非点击获得回报，这是一个使用假设检验类型数据挖掘的宝贵结果。

3.2.2 有指导数据挖掘

有指导数据挖掘是另一种类型的数据挖掘。有指导数据挖掘的重点是一个或多个目标变量，而且历史数据中的实例包含了所有这些目标值。换句话说，有指导数据挖掘不仅寻

找数据中的任何模式，而且还会寻找能够解释目标值的模式。一个非常典型的例子是保留模型。历史数据包含活跃客户实例以及其他已经停止的客户实例。有指导数据挖掘的目标是找到一些模式，能够区分导致客户离开和客户留下来的因素。

在统计中，术语“预测建模(predictive modeling)”经常用于有指导数据挖掘。依作者看来，这多少有一点用词不当，因为虽然预测建模绝对是有指导数据挖掘的一个方面，但是它同样有其他的方面。基于目标变量和输入之间的时间关系，第 5 章区分了预测建模和剖析建模(profile modeling)。预测建模特指目标来自一个晚于输入的时间；剖析建模特指目标和输入来自相同的时间。

3.2.3　无指导数据挖掘

无指导数据挖掘是一种不使用目标变量(至少不明确使用)的数据挖掘。在有指导数据挖掘中，不同的变量发挥不同的角色。目标变量是研究对象；其余的变量用来解释或预测目标值。在无指导数据挖掘中，没有特殊的角色。目标是找到全体模式。在检测到模式之后，由人来负责解释和决定它们是否有用。

术语“无指导”实际上有一点误导。虽然没有使用任何目标变量，但是仍然必须解决业务目标。由无指导数据挖掘解决的业务目标可能听起来与其他任何目标一样直接，“发现欺诈的例子”是一个业务目标的例子，它可能需要有指导或无指导数据挖掘，这取决于训练数据是否包含明确的欺诈交易。有指导方法将寻找与已知欺诈案例类似的新记录。无指导方法是寻找异常的新记录。

业务目标的另一个例子是增加平均订单大小，这可以由无指导数据挖掘来解决。关联规则是一种无指导数据挖掘技术，可以发现哪些相关物品会经常在一起出售的模式。此信息可用来通过改进交叉销售来提高订单大小。

有时，业务目标自身可能有一点模糊，而数据挖掘的努力是一种精炼它们的方法。例如，一家公司可能会有为不同客户群开发专门服务的目标，但是没有明确了解应该如何划分客户。聚类是一种无指导数据挖掘技术，可以用来发现客户的划分。研究划分可能会深入了解组内成员的共同之处，这反过来会提出新产品可能解决的共同需要。

3.3　目标、任务和技术

作者认识的一位数据挖掘顾问说，他生活在恐惧当中，害怕客户会阅读一篇提到一些特定数据挖掘技术名称的杂志文章。当营销副总裁开始询问神经网络与支持向量机时，可能就到了重置会话的时间。数据挖掘总是开始于一个业务目标，而数据挖掘人员的首要工作就是很好地理解这一目标。这一步骤需要在设定目标的上层管理人员以及负责将这些目标纳入数据挖掘任务的分析师之间进行良好的沟通。下一步工作是根据数据挖掘任务重申业务目标，直到此时才会选择特定的数据挖掘技术。

3.3.1 数据挖掘业务目标

第 2 章的数据挖掘应用提供了几个业务目标的好例子：

- 选择广告的最佳位置。
- 为分支或商店寻找最佳位置。
- 获取更多有利可图的客户。
- 降低暴露于违约的风险。
- 改进客户保留。
- 检测欺诈性索赔要求。

本书的其余部分还包含了许多数据挖掘的示例，它们用于解决现实世界中的实际问题。并非所有的业务目标本身都可直接作为数据挖掘的业务目标；有时候必须把它们转变成数据挖掘的业务目标。数据挖掘要想成功，业务目标应该明确，同时指向适于使用现有数据进行分析的特定努力。数据挖掘的业务目标通常可以表示成可测量的事物，如增量收益、响应率、订单大小或者等待时间等。

当然，实现这些目标需要的不只是数据挖掘，但是数据挖掘可以发挥重要的作用。第一步是为问题设计一个高级方法。为了获得更多有利可图的客户，你可以首先学习驱动已有客户的盈利因素，然后获取具有合适特征的新客户。降低信贷风险可能意味着预测哪些目前信誉良好的客户可能会变质，同时提前减少他们的信用额度。提高客户保留可能会聚焦于改进现有客户的体验，或者获取持续期预期会很长的新客户。高级方法提出特定的建模任务。

3.3.2 数据挖掘任务

数据挖掘任务是可以独立于任何特定业务目标进行描述的技术活动。如果一个业务目标适合于数据挖掘，那么它通常可以表示成下列任务：

- 为挖掘准备数据
- 探索性数据分析
- 二元响应建模(也称为二元分类)
- 离散值分类和预测
- 数值估计
- 发现群集和关联
- 将模型应用于新的数据

数据挖掘项目通常会涉及若干个这样的任务。我们来看这样一个例子：在直接营销活动中决定覆盖哪些客户。探索性数据分析表明哪些变量对于描述客户响应的特征来说非常重要。这些变量可以用来发现类似客户的群集。客户的群集标识可能是二元响应模型中的一个重要的解释变量。当然，创建模型的目标是为了将其应用到表示潜在客户的新数据，对他们响应活动的倾向进行打分。

1. 为挖掘准备数据

为挖掘准备数据是从第 18 章到第 20 章的主题。需要的工作量取决于数据源的性质和特定数据挖掘技术的要求。一些数据准备工作几乎总是必需的，而且数据准备通常是数据挖掘项目中最费时的一部分。需要一些数据准备来修复源数据的问题，但是它主要还是旨在加强数据的信息内容。更好的数据意味着更好的模型。

通常，多来源的数据必须结合成一个客户签名，其中每个客户一条记录，同时有大量的字段用来获取对他们感兴趣的一切。因为源数据通常不是客户级别的，因此构建客户签名需要多次转换。事务必须以有用的方式汇总。可以捕获时间序列中的趋势作为斜率或差异。对于只能在数字上工作的数据挖掘技术，类别数据必须表示成数值。有些数据挖掘技术不能处理缺失值，因此必须以某种方式处理缺失值；同样，离群点(outlier)也需要处理。当某些结果非常罕见时，有必要使用分层抽样来平衡数据。当变量采用不同的尺度度量时，也有必要对它们进行标准化。

数据准备可能涉及以创造性的方式结合现有的变量来创建新的变量。它也可能涉及使用主成分和其他技术来缩减变量的数量。

2. 探索性数据分析

探索性数据分析不是本书的重点，但是这并非因为我们认为它不重要。事实上，作者之一(Gordon)曾经写过一本书 *Data Analysis Using SQL and Excel*，重点介绍了这种数据挖掘任务。探索性数据分析的结果可能是一份报告或一系列的图，它们描述了感兴趣的东西。探索性数据分析也可用于在数据中添加新的度量和变量。

剖析是一种熟悉的方法，可用于许多问题，它根本不需要涉及任何复杂的数据挖掘算法。剖析常常基于人口统计变量，例如地理位置、性别、年龄等。因为广告是根据这些相同的变量来销售的，因此人口统计剖析可以直接转化为媒体策略。简单的剖析可用于设置保险费。一名 17 岁的男性会比一个 60 岁的女性支付更多的汽车保险。同样，一个简单的人寿保险合同申请表会询问年龄、性别、是否吸烟等——而不是其他更多的信息。

虽然剖析很强大，但是它存在严重的局限性。其中之一是不能区分原因和结果。只要剖析是基于常见的人口统计变量，这一限制就不明显。如果男性比女性购买更多的啤酒，你不必怀疑喝啤酒是否就是客户为男性的原因。我们可以假设从男人到啤酒之间存在链路是合理的，但反之则不然。

对于行为数据而言，因果关系的方向并不总是那么清晰。考虑一些实际的数据挖掘项目例子：

- 购买了存款证(certificate of deposit，CD)的人在他们的储蓄账户中钱很少或没有钱。
- 使用语音信箱的客户经常拨打自己的电话号码。

不把钱存到储蓄账户中是 CD 持有人的一个共同行为，就像男性是啤酒消费者的一个共同特征。啤酒公司会寻找男性来推销他们的产品，那么银行是否应该找出没有存款余额的人，以便向他们出售存款证明？可能不会！据推测，CD 持有人的储蓄账户没有钱，是因为他们用这笔钱购买了 CD。储蓄账户中没有钱的一个更常见原因是他们没有任何钱，而没钱的人不是投资账户的好的潜在客户。类似地，语音信箱用户如此频繁地呼叫他们自己

的号码只是因为在这个特定的系统中，这是一种检查语音信箱的方式。该模式对寻找潜在客户没用。

3. 二元响应建模(二元分类)

许多业务目标归结为彼此分为两类——把好坏分开，把绵羊和山羊分开，或者(存在性别和年龄歧视的风险)把男人和男孩分开。在直接营销活动中，好的会响应，而坏的不会。当信贷扩展时，好的会支付欠款，而坏的会违约。当提出索赔要求时，好的是有效的索赔，而坏的则是欺诈。有些技术，如逻辑回归，是专门针对这类是或否的模型。

取决于不同的应用，响应模型评分可以是类标签本身，也可以是属于感兴趣类的可能性的估计。信用卡公司想要向滑雪靴制造商出售其在账单信封空白处的广告空间，它可以构建一个分类模型将其所有持卡人分为两类，滑雪者或非滑雪者。更典型的是，它将对每个持卡人指定一个滑雪倾向评分。任何评分大于或等于某个阈值的持卡人都被归类为滑雪者，而得分较低的人则被认为不是一个滑雪者。

这种估计方法有一个很大的优点，就是可以根据该估计值对个人记录进行排序。为了看看这么做的重要性，想象一下这个滑雪靴公司为 50 万份邮件制定了预算。如果使用该分类方法确定了 150 万名滑雪者，那么它只需简单地把广告投放在此集合中随机选择的 50 万人的账单上。另一方面，如果每个持卡人有一个滑雪倾向得分，那么它可以联系 50 万最有可能的候选者。

4. 分类

分类(Classification)作为最常见的数据挖掘任务之一，似乎是人类必不可少的。为了理解世界并与世界交流，我们不断地进行划分、归类和分级。我们把生物按门类(phyla)、种类(specie)和种属(genera)分类；把事物按元素分类；把狗按品种分类；把人按种族分类；把牛排和枫树糖浆按照美国农业部(USDA)的等级分类。

分类涉及为新给定的对象指定预定义类集合中的某一类。分类任务由明确定义的类来描绘，同时模型集由划分好的实例组成。其任务是建立某种类型的模型，可应用于对未分类的数据进行归类。

使用本书所描述的技术已经解决的分类任务例子包括：

- 把信贷申请者分类为低、中或高风险。
- 选择 Web 页面上要显示的内容。
- 确定哪些电话号码对应传真机，哪些对应声音线路以及哪些是两者共享。
- 识别欺诈性的保险索赔。
- 基于自由文本(free-text)的描述指定行业代码和职务名称。

在所有这些例子中，类的数量有限，而任务就是把任何记录指定到它们中的一个或另一个。

5. 评估

分类处理离散输出：是或否；例如是否为麻疹、风疹或水痘。评估处理连续值输出。给定一些输入数据，评估生成一些未知的连续变量的值，如收入、订单的大小，或信用卡

余额。

评估任务的示例包括：

- 估计家庭的收入总额。
- 估计客户的生存周期值。
- 估计客户违约的风险。
- 估计某人响应余额转账请求的概率。
- 估计需要转账的余额大小。

最后两个估计的乘积即是余额转账营销的预期价值。如果预期价值低于营销的成本，那么就不应该发起请求。

6. 发现群集、关联及相关组

无指导数据挖掘的两个例子有：决定超市购物车中什么东西会放在一起，以及发现具有类似购买习惯的客户组。倾向于一起出售的产品称为相关组(affinity group)，具有类似行为的客户形成了市场划分。零售商可以使用相关组来计划商品在货架上或目录中的安排，从而使常常被一起购买的物品会同时看到。营销人员可以为特定的分组设计产品和服务。

相关组合是一种从数据中生成规则的简单方法。如果两个物品，比如猫食和猫窝，同现次数足够频繁，那么你可以想想如何在营销活动中使用此信息。它还带来了另一个问题：客户应该买而没买的东西是什么？买了许多猫窝的客户还应该买猫食——他们在哪可找到它？

聚类是把异构的总体数据划分成一些更均匀的小组或群集的任务。聚类与分类的区别在于聚类不依赖预定义的类。在分类中，通过在预分类的实例上训练得到模型，然后基于该模型对每条记录分配一个预定义的类。

在聚类中，没有预定义的类，也没有实例。基于记录的自相似性把它们分组在一起。它由用户决定把什么意义(如果有的话)附加到结果群集上。症状群集可以表示不同的疾病。客户属性的群集可能表示不同的市场划分。

聚类通常是某些其他形式的数据挖掘或模型的前奏。例如，聚类可能是市场划分工作的第一步：相对于试图对“客户对哪种类型的促销响应最好”获得一个通用规则，其首先把客户库划分成具有类似消费习惯的群集或人群，然后问什么样的促销最适合每个群集。第 13 章和第 14 章详细介绍了群集检测技术。

7. 将模型应用于新的数据

前面列出的许多任务通常会涉及把模型应用于新的数据。探索性数据分析并非如此，对于聚类而言它可能为真也可能不为真，但是对于二元响应模型、分类和评估，用来创建模型的数据中包含了目标变量的已知值。将模型应用于目标值已知的数据的理由之一是为了评价模型。配置模型之后，其目的是为了对新的数据打分，其中响应概率、类别或者要估计的值未知。

将模型应用于新的数据称为打分(scoring)。要打分的数据必须包含模型所需的所有输入变量，同时每行有一个唯一的标识符。打分的结果是一张新表，其中至少有两列——标识符和分数。

3.3.3 数据挖掘技术

本书的大多数章节描述了数据挖掘技术的某种独特技术。

在许多情况下，数据挖掘是通过构建模型来完成的。从某种词义来看，模型是对事物如何工作的一种解释或说明，其足以反映现实，从而可以用来对现实世界进行推理。虽然没有意识到，但是人类无时无刻不在使用模型。当你看到两个餐厅，并判断在每张桌子上有白桌布和真花的其中一家餐厅，会比另一家有胶木桌和塑料花的餐厅更贵时，你实际上是根据头脑中构建(基于过去经验)的模型进行了推理。当你走进其中一家餐厅时，你再次调用了一个心理模型。

从更为技术化的词义来看，模型是指某种东西，可使用数据对事物进行分类、预测、估计值，或者生成其他一些有用的结果。如图 3-4 所示，有相当多的东西可应用于数据并产生某种得分，它们都符合模型的定义。

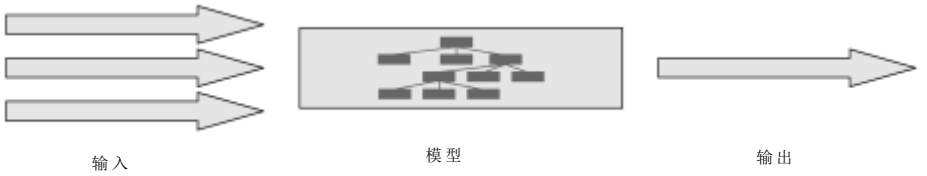


图 3-4 模型接受一个输入并产生输出

数据挖掘模型可以满足两个目的。第一个目的是产生得分，其可用来指导决策。第二个目的是为了洞察用于构建模型和目标的解释变量之间的关系。取决于不同的应用，这些目的中的一个可能会比另一个更重要。

数据挖掘技术分为两类：它们可以是有指导或无指导的，分别意味着技术本身是否需要或不需要目标变量。有指导和无指导技术不应与有指导和无指导数据挖掘混淆，因为这两类技术均可用于两类数据挖掘。

3.4 制定数据挖掘问题：从目标到任务再到技术

业务目标、数据挖掘任务和数据挖掘技术形成了一个阶梯，分别为从一般到具体和从非技术性到技术性。制定数据挖掘问题涉及从该阶梯上下降，每次一步；先从业务目标到数据挖掘任务，然后从数据挖掘任务到数据挖掘技术。通常，每个步骤都需要具有不同技能集的不同工作人员参与。设置目标及其优先次序是上层管理人员的责任。把这些目标转换成数据挖掘任务，并使用数据挖掘技术来完成它们是数据挖掘人员的责任。收集必要数据，并把它转化为合适的形式常常需要数据库管理员与信息技术组其他成员的合作。

3.4.1 选择广告的最佳位置

一家公司正试图得到新的有利可图客户。它应该去哪里做广告？Google AdWords？现实的电视烹调节目？杂志？如果是杂志，应选择哪一份？Architectural Digest？People en Español？还是 Rolling Stone？

许多因素会影响决策，包括总体成本，每个效果的费用，以及每次转换的费用。通过匹配广告媒介的人口统计信息与最佳客户的人口统计信息，数据挖掘可以为决策提供输入。有利可图客户的行为数据没有作用，因为广告只是基于人口统计数据。

一种可能的方法是：

(1) 使用人口统计和地理特征剖析现有的有利可图客户，诸如年龄、性别、职业、婚姻状况和社区特征等。使用这个剖析定义有利可图客户的原型。

(2) 使用用于剖析有利可图客户的相同变量，定义每个潜在广告媒介的受众。

(3) 估计每个广告渠道到有利可图客户原型的距离。这个距离是广告渠道的相似得分，正如打高尔夫一样，越小会越好。

(4) 在得分最低的场地做广告。

这是一个相似度模型(similarity model)的示例，将在第 6 章对其进行介绍。

3.4.2　确定向客户提供的最佳产品

下一个向客户提供的最佳优惠是什么？这个问题是在许多行业都会发生的交叉销售的一个例子。

针对这个问题有几种可能的办法，在许多因素中它们主要取决于可供选择的产品数量。如果可管理的产品数量比较少，那么一个好办法是为每款产品建立一个单独的模型，从而可以对每个客户给定每款产品的得分，如图 3-5 所示。客户的最佳优惠是他具有最高分数的产品(可能已经排除了客户已有的产品)。

(1) 对于每款产品，构建一个二元响应模型来估计客户对该产品的倾向。

(2) 对于已经有一个产品的客户，设置其倾向得分为 0。

(3) 使用这些倾向得分，设计为每个客户指定最佳产品的决策过程，该过程将基于类似最高倾向或最高期望利润之类的度量。

步骤(1)中的自然选择包括决策树、人工神经网络和逻辑回归。

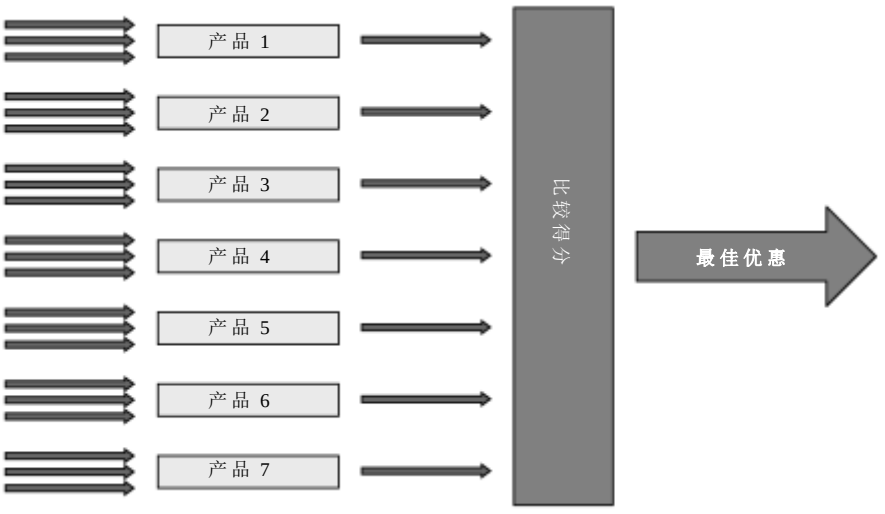


图 3-5　对比每款产品的个人得分倾向以确定最佳优惠

二元响应模型并非计算倾向得分的唯一方法。另一种方法是使用输入变量对数据进行聚类，查看每个群集中哪些产品占主导地位。可以把一款给定产品在群集中的比例指定为倾向的得分。该方法可以使用 K-Means 聚类或另一种聚类技术。

3.4.3 发现分支或商店的最佳位置

新商店的最佳位置应该在哪里？在这个场景中，现有商店的绩效数据与服务区(每个商店服务客户的自然市场区域)的数据相连。其思想是要找到一组解释变量能够预测一家商店的良好绩效。

以下建模任务是解决该问题的一种方法：

(1) 建立一个模型来估计某个商店绩效指标，该指标是基于该服务区的可用解释变量。

(2) 把该模型应用到候选位置，从而可以选择得分最高的位置。

这基本上是一种估计模型，可以使用各种技术，如神经网络、回归或 MBR 等。

一种替代方法是把商店分类为好或者坏，然后构建模型来预测这些组。通常，这么做的一种好办法是采取排除中间的做法：每个商店的利润分成三个级别——高、中、低。去掉位于“中间的”商店，并构建一个模型来区分高和低(第 15 章的一个案例研究采用这种方法来发现能够区分拉美裔区域与非拉美裔区域的商店的因素)：

(1) 把现有的商店分类为好或坏，同时建立一个能够区分这两个类的模型。

(2) 把该模型应用到候选位置，从而可以选择好的候选者。

可能的解释变量包括驾驶距离内的人口、驾驶距离内竞争对手的数量以及人口因素。这是一个剖析模型，因为目标是把当前绩效与当前环境相连。建模的技术是那些用于分类的技术，如逻辑回归分析、决策树以及 MBR 等。

3.4.4 根据未来利润划分客户

我们已经建立了一种定义利润的方法，如客户在一年中产生的总收益或净收益。现在的目标是基于客户在下一年的预期盈利能力对他们进行划分。

有许多方法可用来计算盈利能力。这种方法省略了一些更为困难的主题，如预测一个客户在多长时间内仍将是客户(并因此决定将来的折扣率)，以及如何定义客户的网络效应(network effect)。

对于这种方法，把时钟往回拨一年，为每个当时活跃的客户生成一个快照。然后，度量下一年期间的总收益。以下就是该模型：

(1) 为建模准备数据，把时钟往回拨一年，为每个当时活跃的客户生成一个快照。然后，度量下一年期间的总收益。这将创建一个预测模型集。

(2) 使用此模式集估计某个人在下一年将会产生多少收益。

(3) 将预期收入分为三个级别，分别是得到高、中、低的预期收益。

第(2)步需要建立一种估算模型，利用诸如神经网络、MBR 或回归等技术。

对这种方法进行轻微改变，将模型集中的客户分类为在来年中的高、中或低收益产生者。这将使用一个分类模型，它可能是决策树(具有三个目标)或三个逻辑回归模型(每个分组一个模型)。

3.4.5　减少暴露于违约的风险

此业务问题的目标是当仍有时间采取措施降低暴露时，检测出违约的警告信号。一种检测方法是使用二元响应模型，其以“违约”为目标。该模型集是在给定时间点(例如，第一年)所有客户的快照，同时也是一个标志，表明他们在快照日期之后的三个月是否违约。可以用二元响应模型对新客户打分，预测其违约的概率。也许，对于具有高违约水平的客户应该降低他们的信贷额度。

可以使用各种技术构建这样一个二元响应模型，如逻辑回归、决策树或神经网络等。甚至可以使用无指导技术，例如聚类。在输入变量上建立群集，然后度量群集的能力以分离目标值。这是一个将无指导技术用于有指导模型的例子。

另一种方法是将违约的概率与违约的数量相结合。这个两阶段模型会估计一个客户在违约后的欠款数额。为此，该模型集只包含已经违约的客户，其目标是欠款数额。该模型将用于计算预期的亏损值，即违约的概率乘以估计的欠款数额。对欠款数额的估计可以使用 MBR、神经网络、回归或可能使用决策树来构建。

然而，另一种方法是把它当作实时事件(time-to-event)问题，当一个客户可能会违约时对其进行估计。在这种情况下，该模型集包含所有的客户，包括他们的开始日期、结束日期以及该客户是否违约。该模型将估计客户违约的时间。当对新客户打分时，如果违约的估计时间是在不久之后，那么将采取行动来降低违约风险。这种类型的模型通常会使用生存分析来构建。

3.4.6　提高客户保留

有许多不同的方式可用来提高客户保留：

- 发现离开风险最高的客户，并鼓励他们留下来。
- 量化改进操作的价值，从而使得客户将继续保留。
- 确定哪些获取客户的方法会带来更好的客户。
- 确定哪些客户无益，并让他们离开。

本节只讨论其中的第一个方式。

确定谁将留下的任务列表与任何二元响应模型的任务列表类似。建立一个模型集，其中包含留下和离开的客户，同时构建模型以发现区分他们的模式。这会提供一个模型评分，你可以将它用于保留工作。

这种类型的二元响应模型可以使用许多诸如决策树、神经网络、逻辑回归与 MBR 之类的技术来构建。一种估计客户剩余持续期的替代方法是使用生存分析，并把保留信息应用到不久之后最有可能离开的客户。

有时，一个模型最重要的输出不是其产生的分数，而是通过检查该模型本身所产生的知识。该模型能够解释客户是否主要是由于服务中断、价格敏感性或者其他原因而离开。然而，这需要使用一种能够解释其结果的技术。决策树和逻辑回归是最具可解释性的突出代表。

3.4.7 检测欺诈性索赔

把这一目标转换成模型任务取决于是否存在已知欺诈的例子。如果是，那么这是一个有指导数据挖掘任务：

(1) 构建一个能够从合法索赔中区分欺诈性索赔的剖析模型。

(2) 利用该模型对所有进来的索赔评分。标注得分高于某个阈值的索赔，从而在批准之前对其进行额外的审查。

决策树和逻辑回归是可用于在步骤(1)中构建剖析模型的技术。

有时，虽然涉嫌欺诈，但是并不清楚哪些事务是欺诈性的。这种情况要求无指导数据挖掘：

(1) 形成类似索赔的群集。大多数的索赔可能会落到少数几个大的群集中，它们代表了不同类型的合法索赔。

(2) 审查较小的群集，以了解它们如此特殊的原因。

在较小群集中的索赔也可以是完全合法的。聚类操作所能表明的仅仅是它们不同寻常。因为一些不寻常的索赔被证明是欺诈，所以所有这些都值得进一步审查。

一个目标，两个任务：赢得数据挖掘的比赛

每年，学术界和工业界的竞争者们都会在一项与年度 KDD(知识发现和数据挖掘)会议同时举行的竞赛中测试他们的数据挖掘技能。某一年，非常清楚的是优胜者与失败者的区别不在于他们使用的算法或软件，而是他们如何将业务问题转化为数据挖掘任务。

业务问题是要最大化一个非营利慈善团体的捐款。数据是一个关于捐献的历史数据库。

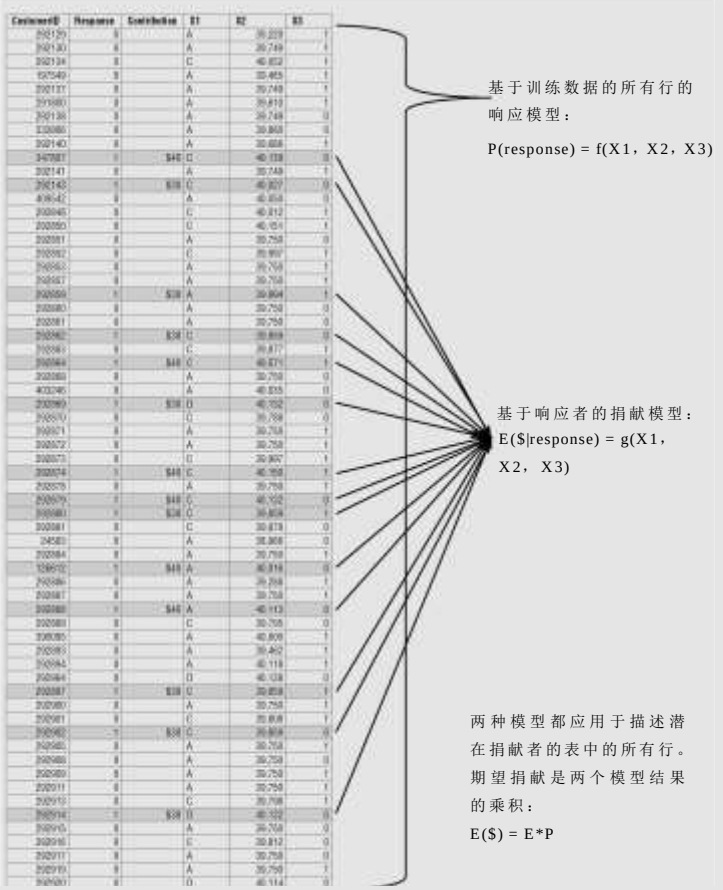
探索数据之后发现了第一个结论：某些人捐献的次数越多，则他们每次捐献的钱就越少。最好的捐助者是那些最频繁的响应者，这样的预期非常合理。

不过，在这种情况下，人们似乎是以年为基准计划他们的慈善捐款。他们可能会一次性捐赠，或者随着时间推移进行间隔捐献。更多的检查并不总是意味着更多的钱。这表明决定捐赠与捐赠多少是分离的。这两项决定很有可能受到不同因素的影响。也许，无论是哪个收入水平的人，如果他们自身在军队服过役，那么他们都更有可能捐赠给一个老兵组织。在他们已决定作出捐献之后，收入水平可能会影响捐赠多少。

这些结论可以导致赢的方法，即分别对响应和捐献的大小进行建模。响应模型是在一个同时包含捐献者和非捐献者的训练集上构建。这是一个二元结果分类任务。

捐献大小模型是在一个仅包含捐献者的训练集上构建。这是一种评估任务。下图显示了这两个模型，以及如何结合它们的结果以产生每个潜在客户的期望响应值。

三个优胜组都是采用这种模型相结合的方法。另一方面，大多数竞争者建立了一个以捐献数额为目标的单一模型。这些模型把整个问题看作是一个评估任务，其中未响应表示为一个零美元的捐献。



期望捐献值的两阶段模型

3.5　不同技术对应的任务

你可以以创造性的方式使用本书描述的所有数据挖掘技术，将它们应用于通常最相关的应用之外的应用。每个主要的技术系列都对应一章(甚至不止一章)。各个技术章节包括了如何将这些技术应用于各种目的的示例。不过，仍然有一些技术更适合于某些任务。当选择一种技术时，询问自己这些问题：

- 有一个或多个目标吗？
- 目标数据是什么？
- 输入数据是什么？
- 易于使用有多重要？
- 可解释性有多重要？

这些问题的答案会缩小技术的选择范围。

3.5.1 有一个或多个目标

所有的有指导数据挖掘技术，包括回归、决策树和神经网络，都需要用已知的目标变量值进行训练。当数据没有包含这样的目标时，需要诸如聚类或探索性数据分析之类的无指导技术。

3.5.2 目标数据是什么

当目标是数值并且值域很宽泛时，适合采取产生连续值的技术。线性回归模型可以产生从负无穷到正无穷之间的任何值，神经网络同样如此。当任务是估计一个连续目标的值时，这些是自然的选择。回归树模型和表查询模型也可以用来估计数字值，但它们会产生数量相对较小的离散值。基于记忆的推理是数值目标的另一个选择，它可以产生一个值域比较大的值，但从未超出原始数据的范围。

当目标是一个二元响应或者类别变量时，将采用产生属于每个类的概率的技术。决策树非常适合这类问题，逻辑回归和神经网络同样如此。取决于问题的其他方面以及输入的性质，其他的技术如相似度模型、基于记忆的推理以及朴素贝叶斯模型可能也是不错的选择。

3.5.3 输入数据是什么

回归模型、神经网络和许多其他技术对输入值进行数学运算，因此不能处理类别数据或缺失值。当然可以重新编码类别数据，或者以数字字段替换表示重要分类特征的类别字段。也可以输入缺失值。然而，这些操作可能会既费时又不准确。随着类别字段的数量以及带有缺失值字段的数量不断增长，决策树、表查询模型以及朴素贝叶斯模型的吸引力也在不断上升，所有这些模型都可以轻松地处理类别字段和缺失值。当输入是数字且不包含缺失值时，回归模型和神经网络能够使用数据中的更多信息。

3.5.4 易于使用的重要性

一些技术比其他技术需要更多的数据准备。例如，神经网络需要所有的输入都是数字且在一个小的值域范围内。同时，它们对噪声也很敏感，且不能处理缺失值。其他技术，如决策树，容错性更好且所需的数据准备较少，但是可能效果会较差。通常需要在能力、准确性以及易于使用之间权衡。作为一种极端的例子，遗传算法需要数据挖掘人员做太多的工作，以至于如果有替代方法的话，则很少会使用它们。

自从本书第一版在 20 世纪 90 年代出版之后，数据挖掘软件工具已在易用性领域取得了重大进展。其中最好的工具提供了支持最佳实践的用户界面，甚至使得诸如神经网络之类的复杂技术也相对用户友好。

3.5.5 模型可解释性的重要性

对有些问题而言，尽快获得正确的答案至关重要。一个现代化的、无需封皮

(no-envelope-required)的自动柜员机必须能够准确地识别手写金额以接受存款的支票。虽然了解算法如何区别美国的“7s”和欧洲的“1s”肯定会引人入胜，但是并不迫切需要这么做。在刷信用卡并传输批准代码的短暂间隔内，将根据欺诈的似然性对事务进行打分。获得正确的决策非常重要。批准一个欺诈事务会带来直接和明显的成本；拒绝合法交易会令一个有价值的客户心生厌烦。在这两个例子中，获得正确的答案显然比明确解释如何做决定更为重要。

而另一个极端是，有些决定——例如，是否授予或拒绝贷款——可能易受管理评论的影响。把申请者因为有太多的开场白，而且债务收入比太高作为贷款被拒绝的理由进行解释会比较好。“这个模型确定申请者为高风险，但是我们不知道为什么”的解释是不可接受的。

不同的技术提供了在准确性和可解释性之间的不同权衡。决策树可以说提供了最佳的解释，因为每片叶子都有一个规则形式的准确描述。虽然这意味着对于任何给定记录的得分都可以解释，但是这并不意味着一个大型的、复杂的树作为一个整体易于理解。权衡之处在于，决策树可能不是与其他技术一样使用变量的许多固有信息，那些技术会直接使用变量的值，而不是仅仅与分裂值(splitting value)相比较。

稍微注意一下数据准备，回归模型也会发现有助于得分的因素。若已经对解释变量进行了标准化，则模型系数的相对大小会表明每个变量对得分的贡献大小。在回归时，解释变量值的每次细小变化都对得分有影响。在这种意义上，回归模型比决策树更充分地利用了解释变量所提供的信息。

神经网络相当灵活，并且能够非常精确地建模相当复杂的功能，但基本上不可解释。这些技术中的每一种技术都提供了在最好得分和最好解释性之间的不同权衡。了解了这些技术的优缺点，你必须决定哪些技术最适合你的应用程序。

表 3-1 显示了哪些技术通常用于哪些任务。如表中所明确给出的，相当多的有指导技术可用于分类、预测和估计问题。最后的选择除了产生分数、待分析的数据特征，还取决于模型能解释的程度。

表 3-1　各种技术分别对应哪些任务

任　务	最　适　合	也　考　虑
分类和预测	决策树、逻辑回归、神经网络	相似度模型、表查询模型、最近邻模型、朴素贝叶斯模型
评估	线性回归、神经网络	回归树、最近邻模型
二元响应	逻辑回归、决策树	相似度模型、表查询模型、最近邻模型、朴素贝叶斯模型
发现群集和模式	任何聚类算法	关联规则

3.6　经验教训

数据挖掘过程可能会因为多种原因导致失败。失败的形式多种多样，包括仅仅是不能回答你提出的问题，以及“发现”你已经知道了的事情。一种特别有害的失败类型是学习

的东西不真实。其发生的原因多种多样：当用于挖掘的数据不具代表性时；或者当它包含不能泛化的意外模式时；或者当它以破坏信息的方式进行了汇总时；或者当它把本应保持独立的不同时期的信息混合在一起时。

有三种类型的数据挖掘。探索性数据挖掘产生见解或回答问题，而不是生成用于评分的模型。探索性数据挖掘通常涉及构造出使用数据能够证明或不能证明的假设。探索性数据挖掘非常重要；然而，它不是本书高级技术的主题。

当历史数据包含正在寻找的实例时，使用有指导数据挖掘。对于流失模型，其假设历史数据中包含了已经留下或离开的客户实例。对于客户价值模型，其假设可以使用历史数据来估计客户价值。该模型的目标就是这些变量。该模型的“解释”变量是输入。

无指导数据挖掘不使用目标变量。这就像是把数据扔进计算机，然后看它在哪着陆。为了使无指导数据挖掘有意义，需要对结果进行理解和解释。由于没有目标，计算机无法判断结果是好还是坏。

你可以单独使用这三种数据挖掘，或者组合它们来完成一个范围广泛的业务目标。数据挖掘过程以业务目标作为开始。数据挖掘过程涉及将业务目标转化为一个或多个数据挖掘任务。在明确定义这些任务之后，任务的性质、可用的数据类型、提交结果的方式，以及模型的准确性和模型可解释性之间的折衷，都会影响数据挖掘技术的选择。

无论你选择哪种技术，以及无论采取什么数据挖掘类型，有效地使用数据挖掘都需要一些统计学知识，这是下一章的主题。