大数据时代的商业建模

范若愚　王金陵　赵丽丽　范承懿　编著

上海图书馆
上海科学技术文献出版社

大数据时代的商业建模

范若愚　王金陵　赵丽丽　范承懿　编著

上海图書館
上海科学技术文献出版社

图书在版编目（CIP）数据

大数据时代的商业建模 / 范若愚等编著．—上海：上海科学技术文献出版社，2013.7
ISBN 978-7-5439-5868-5

Ⅰ.① 大… Ⅱ.①范… Ⅲ.①数据采集—研究 Ⅳ.① TP274

中国版本图书馆 CIP 数据核字（2013）第 128003 号

责任编辑：应丽春
封面设计：许　菲

大数据时代的商业建模
范若愚　王金陵　赵丽丽　范承懿　编著
出版发行：上海科学技术文献出版社
地　　址：上海市长乐路 746 号
邮政编码：200040
经　　销：全国新华书店
印　　刷：常熟市人民印刷厂
开　　本：787×1092　1/16
印　　张：13.5
字　　数：219 000
版　　次：2013 年 7 月第 1 版　2013 年 7 月第 1 次印刷
书　　号：ISBN 978-7-5439-5868-5
定　　价：38.00 元
http://www.sstlp.com

序

近年来，在世界信息化浪潮、云计算技术的推动下，海量数据存储、处理的成本急剧下降，大数据时代也由此降临，大数据所蕴含的价值如今正在逐步释放，挖掘利用大数据对提升政府管理职能和企业的决策能力、创新发展模式都产生着深远的影响。发展我国的大数据产业对于推动经济由粗放型向集约型发展，加速经济发展转型会起到至关重要的作用。

我曾经在金融系统工作多年，金融服务从某种意义上来说是一场为客户提供优质服务的数字战役：金融机构通过多种渠道和产品与客户进行业务交流，掌握客户海量的数据，并以先进的数据挖掘技术加以分析，将其转化为极具价值的洞察力，在风险管理、精准营销、客户关系管理上正发挥着越来越大的作用。

2012年12月在北京召开的“中国数据分析业峰会”上发布了《中国企业数据化现状调查报告》。报告显示，超过17%的企业尚未进行数据化建设，将近70%的企业认为其数据化建设处于初级和中级阶段。这表明我国企业的数据化建设和应用均处于较低水平，企业在经营管理数据化的意识、数据化基础建设的投入、人才的培养和使用上都亟待加强。

我国数据化建设水平的落后与数据挖掘技术专家的缺乏非常有关系。在美国和欧洲的市场上，数据挖掘技术早已成熟，实施数据挖掘的各个环节分工明确，商业策略师和数据分析师彼此配合。而我国的数据挖掘还刚刚起步，专

业人士不多且非常缺乏经验，了解数据挖掘技术的商业策略人士更是稀少，这是很难适应形势发展要求的。与此同时，市场上完整的传授数据挖掘技术的书籍也不多见，这种状况急需改变。

本书的主要作者范若愚先生是我多年好友的儿子。他在美国学习、实践数据挖掘技术十多年，曾率领咨询团队负责为花旗银行的联名信用卡业务提供一切与数据分析有关的服务，是数据挖掘领域的顶尖专家。他回国后主导了大量数据挖掘项目，遍及银行、信用卡、电信、证券、保险、零售、互联网、新能源等领域，积累了丰富的实践经验，入选上海“千人计划”专家。在几年前的一次交谈中，我建议他将掌握的专业技术结合商业实践经验写成书，让更多的人能分享实践的成果。若愚采纳了我的意见，进行大量细致的整理、归纳、提炼，反复推敲，写成此书，我深感欣慰。

我不是数据挖掘的专家，但是这本书结合大量商业案例讲解技术流程，我想一定是很有实际意义的。若愚从待遇丰厚的风险投资行业回到国内从事数据挖掘工作所表现出的责任感也令我感动。这些也是我向读者推荐这本书的理由，希望它对数据挖掘技术在国内的发展起到一定的推动作用。

是为序。

原光大集团董事长、党委书记

王明权

前　言

现代社会，信息为王，银行、电信、零售、互联网等各行各业都对信息格外看重，收集了百万、千万乃至上亿客户的数据。可是信息数据如何转化为企业的知识和经济效益呢？这就需要数据挖掘技术了。庞大的客户信息、数据，单靠人力是无法来分析的，只能依靠统计软件、数据挖掘技术的支持。

20世纪90年代中期，随着计算机运算速度的飞速提升，数据挖掘技术得到蓬勃发展，成为美国社会进步的一个重要因素。近几年数据挖掘在国内也逐渐兴起。国内外多个媒体把数据分析师捧为21世纪最赚钱的技术人才。谷歌的首席经济学家哈尔·范里安(Hal Varian)有段著名的言论："数据非常之多而且具有战略重要性，但是真正缺少的是从数据中提取价值的能力。这也就是为什么统计学家(数据分析师)才是真正了不起的人。"

如何成为一个称职的、优秀的数据分析师呢？仅仅依靠学校里教授的统计知识是远远不够的，还必须掌握必要的数据挖掘流程和方法。然而目前国内外涉及数据挖掘的书籍，大部分都是泛泛地谈及数据挖掘的理论或方法，很少触碰数据挖掘、数据建模的细节和流程，使得国内有志于学习数据挖掘的人士无从着手。本书作者希望结合自己的实际经验，系统具体地介绍数据挖掘尤其是数据建模的流程和方法，并给出大量实用SAS程序。

SAS是全球最广泛使用的统计软件，本书也以它作为数据挖掘、数据建模的工具。SAS软件很容易上手，读者如果对它不熟悉甚至以前没有接触过，也

不构成学习数据挖掘技术的障碍。

本书理想的读者是对数据挖掘有兴趣，且有一定统计基础、熟悉统计软件SAS的人士，因为本书跳过了一些初级统计知识、基本SAS程序的传授。对数据挖掘领域的初涉者，阅读本书可能要花更多的时间和精力，最好能同时补充些统计和SAS软件的基础知识。我们希望各个层级的数据挖掘爱好者都能从本书中获益。

本书前两章简单介绍数据挖掘的常识。第三章介绍如何把商业问题转化为数学问题。第四章是建模前的数据处理。第五到第九章介绍具体建模流程。第十、十一章介绍模型的评估、验证、使用和维护。第十二章通过一个实际商业案例重温整个建模流程。

本书大量列举了数据挖掘技术在各行各业的应用案例，因此商业策划、市场策划专业人士也能从中获益。你们对数据挖掘技术了解越多，数据挖掘在国内的应用发展就会越快。

我们在本书中反复强调数据挖掘离不开商业，成功的模型不仅取决于建模处理，更需要对数据、市场、商业目标的准确理解。我们希望本书的读者能够体会这一点。

数据挖掘也是一门艺术，有较强的主观性，我们也希望本书给读者留下这个印象。

越来越多的人被数据挖掘的远大前景所吸引而投入这个行业，却发现行业中的专业书籍既缺乏又不具体。我们最大的希望是本书能填补这个空白，起到抛砖引玉的作用。

本书在写作过程中，杨宇新(英国)、何锦阳、张皓菀给予了大量校对、修改的帮助，并提出了很多建设性意见，在此深表谢意！

本书作者的邮箱地址是fanr55@126.com，欢迎读者与我们联系。

目　录

第三篇　实际商业案例

第一篇　基本知识

第1章 简 介

1.1 数据挖掘简介

数据挖掘技术以现代数理统计学为基础，以计算机技术的发展为支撑，使用SAS等专业工具，对海量数据进行分析和信息挖掘，从而实现对客户的需求和行为进行准确地预测。

数据挖掘技术的雏形诞生于20世纪60年代。70年代，数据模型中的一类重要应用——信用评分模型开始发展，商业银行可以量化地管理风险。90年代初，随着计算机技术的发展与普及、现代数理统计学的进步和数据库技术的突破，海量数据的计算、分析成为现实，数据挖掘技术迅速成为美国等发达国家金融机构进行市场营销、风险控制、客户关系管理的主要手段。以Capital One[①]、First USA[②]为代表的信用卡公司，率先利用数据挖掘技术开发出能有效预测消费者未来行为的模型，从而获取了巨大的商业利润。数据挖掘战略的成功也使得Capital One等公司在短时间内成为行业中的领军人物。90年代末，数据建模技术逐渐被各大商业银行掌握并运用，进而扩展到商业银行的传统业务，成为推动美国金融业蓬勃发

① 美国第一资本金融公司，20世纪90年代初期还是美国弗吉尼亚州一城市银行(Signet Bank)的信用卡部门。它基于信息、数据挖掘的战略获得巨大成功后，跻身全美前十大银行。

② 后被美国第一银行(Bank One)收购。

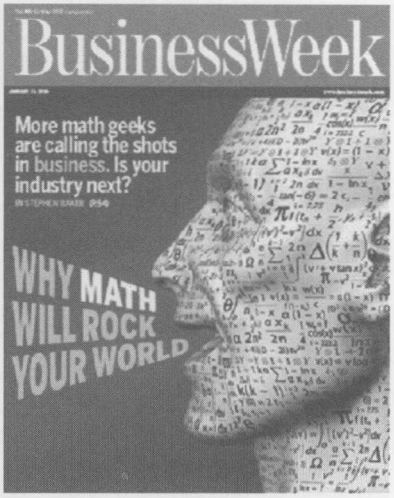

图 1-1 2006 年 1 月 23 日美国《商业周刊》封面

展的核心技术。

美国著名的咨询公司 Gartner Group 在 2000 年把数据挖掘列为未来 5 年改变人类社会的五大技术之一；美国麻省理工学院 2002 年 1 月发表的《技术评论》(《Technology Review》)也把数据挖掘列为改变未来的十大技术之一；美国著名的《商业周刊》在 2006 年 1 月 23 日的封面文章中介绍了数据挖掘技术，并称数据挖掘技术在各行各业越来越得到广泛应用，将震撼整个世界；2011 年 5 月，“麦肯锡全球研究所”(McKinsey Global Institute)发布的一份咨询报告中称数据分析为“下一个创新、竞争和生产力的前沿”①。

数据挖掘的重要意义可见一斑。

我国的数据挖掘尚处在初级阶段，面对庞大的国内市场，数据挖掘将有很大的发展空间。

数据挖掘离不开信息，数据挖掘的效果很大程度取决于信息的质量，如果信息数据缺失、杂乱、毫无特征可循，任何数据挖掘的手段都会黯然失色。我们举一个简单的例子：如图 1-2 所示，两个具有相同基本信息的客户，由于所处人生阶段不同，她们对于金融产品的需求存在明显的差异：左边的客户育有两个孩子，希望进行稳妥型投资，而右边的客户没有子女，风险承受能力较大，可以进行风险性稍高而收益较大的投资。两者的基本信息完全一样，但是需要的理财产品却是截然不同的。如果没有客户子女方面的信息，任何数据挖掘的手段可能都无法区分这两个客户。当然如果数据挖掘不够深入的话，我们也很可能忽视区分这两个客户的重要特征，所以说信息质量和数据挖掘技术相辅相成。

我们举几个数据挖掘在不同行业、不同领域的应用实例，让读者对数据挖掘有更直观的认识。

---

① 报告英文名为《Big data: The next frontier for innovation, competition, and productivity》。

图 1-2 基本信息相同的客户可能有不同的需求

实例一：精准营销，零售

美国大型运动鞋厂商 R 公司于 1999 年希望发展邮购业务，一方面促进销售，另一方面通过邮购建立会员用户网络，因此 R 公司列出其用户的主要特征（如年龄、性别、收入情况等），并通过美国信用局按照这些特征筛选出 1 万个潜在用户，向他们邮寄公司产品的宣传资料，并提供比较有吸引力的价格。1 万个潜在用户中有 113 人响应了邮购销售，购买了产品，他们也成为 R 公司的首批会员。接着 R 公司委托数据挖掘公司，对上述营销数据进行分析，建立了邮购营销响应模型。R 公司使用模型给在美国信用局有信息数据的个人打分，预测每个人购买 R 公司产品的可能性。根据模型的评分，R 公司从美国信用局筛选出 10 万个最可能购买其产品的潜在用户（模型评分最高），进行了又一轮的邮购营销，结果 10 万个潜在用户中有 9 710 人购买了 R 公司的产品，响应率从第一次的 1.13%提高到数据建模以后的 9.71%，充分说明了数据挖掘的重要性。

实例二：风险管理，电信

某电信公司希望向预付款客户提供一定的信用额度，使他们在使用完预付款后不必马上充值，仍然能继续通话。但是如果电信公司不分青红皂白向所有的预

付款客户提供此项服务，一定会出现大量的坏账损失。为此，电信公司聘请了数据挖掘专业公司为其建立风险模型。数据挖掘通过对历史坏账客户的特征挖掘，建立了预测每个电信客户欠账不还可能性的数据模型，电信公司只对风险预测值较低的（即坏账可能性较小）客户提供信用额度。一年以后，预付款客户使用的信用额度总量十分可观，而信用额度这项服务带来的坏账率却几乎可以忽略不计，说明模型的风险预测起到了很好的效果。

实例三：客户关系管理，信用卡

美国C信用卡公司是世界上最著名的依靠“信息化决策”的金融机构之一。因为美国人力成本较高，C公司希望能够提高客户服务中心的工作效率。它的客户打入电话后，会被要求输入信用卡号码和密码，系统核对后连接该客户的实时信息数据，预测客户今天来电的目的，然后自动转接到相关处理部门。这个预测的准确率达到了惊人的75%，为公司省下了巨额的电话转接成本，也为客户节省了大量时间。

为什么预测客户来电原因的准确率会如此之高？这是因为信用卡公司掌握了客户大量的基本信息和交易信息，然后通过数据挖掘找出了一系列规律：如果持卡人在还款日过后的几天内打来电话，那很可能是他忘了还款；如果持卡人在账单日以后的几天内打来电话，那很可能是持卡人收到了账单而对账单有异议；如果持卡人本月的消费金额已经接近了他的信用额度，那么持卡人来电目的可能是申请增加信用额度等。在这些规律之外，还有预测模型的帮助：如果流失模型预测持卡人很可能流失，那么电话就会被转接到处理客户流失的部门。正是完整、实时的客户信息和通过数据挖掘技术建立的一系列规则和模型，帮助C公司建立了智能电话接入系统，节省了数亿美元的人力物力成本，也大大提升了客户的满意度。

1.2 数据建模简介

数据挖掘有许多种方法，例如关联分析、聚类分析、决策树分析、神经网络、数据建模等等。本书仅仅针对数据建模展开探讨研究。数据建模是最高级的数据挖掘方法之一，也是商业应用最广泛的。

数据模型是对实际商业问题的一种数学表述，是为了一个特定目标，对特定的对象，根据特有的内在规律，做出一些必要的简化假设后，得到的一个数学结构。数据模型也是一种数学的思考方法，把一些复杂的商业问题转化成数学问题，然后运用统计学的方法得到数据模型，即数学意义上的答案。

数据建模具有很多优越性：

● 数据模型可以把预测量化，使决策简单有效。

解释一件事情最有效的方法可能就是用数字来表达。数据模型通过数据挖掘，根据每个个体的基本信息、交易信息、行为信息等计算出该个体发生某行为的可能性，将研究目标量化，从而有利于个体之间的比较，帮助管理人员快速有效地进行决策。

● 数据模型可以综合考虑多种因素，避免了条件筛选中“一刀切”的现象。

普通的条件筛选往往是根据某几个变量截取客户群，例如在数据库中找到年龄为35岁以上、性别为女性、收入为20万元以上、职业为中层管理人员以上、学历为大学以上的客户作为目标人群。这样做往往会遗漏一些虽然不能满足某个条件，但在另外一个甚至几个条件上表现十分优秀的客户。再者，各个筛选条件对选择目标客户的重要程度是不一样的，比如收入的重要性可能远大于年龄。如果按照上述的筛选条件，可能忽略了年龄稍稍低于35岁但收入远高于20万的优质客户，造成“捡了芝麻丢了西瓜”的后果。一般的条件筛选只能机械地“一刀切”，而数据模型能够同时考虑多种因素以及各因素的重要程度，覆盖了不同条件此消彼长的情况，故而能够更合理地选择目标客户。

对于数据建模有两个极端的认识。一种认为数据建模很简单，理由是一些高级的统计软件都有自动建模功能，所以建模无须任何分析技能。这个认识是非常错误的。数据建模有三个环节：把商业问题转化为数学问题，解答数学问题，把数学答案转化为商业解决方案。软件自动建模最多只针对第二个环节，而且解答数学问题时也不考虑任何商业含义。因此统计软件自动建立的模型一般只能参考，没有多大的实际意义。

另外一种极端的认识：数据建模是最高级的数据挖掘技术，只要有机会就应该建模。其实许多商业上的数据挖掘不需要建模那么复杂、费时，一些简单的方法就可以取得很好的效果。例如网上商城向客户推荐商品就不需要太复杂的规则，