普华经管

写给经理人的数据挖掘书

运用大数据解决商业难题

【美】理查德·博伊尔（Richard Boire）◎著
杨华勇　王若琼◎译

拥有30多年行业经验的理查德·博伊尔教你运用数据挖掘获得具有可行性的观点和方案
数据分析和挖掘的方法、工具、技巧、陷阱，以及极具启发性的真实案例

四步数据挖掘法：锁定商业问题或挑战，
建立分析文件，运用正确的工具和技术，以及执行与跟踪

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

写给经理人的数据挖掘书

运用大数据解决商业难题

Data Mining For Managers

How to Use Data（Big and Small）to Solve Business Challenges

【美】理查德·博伊尔（Richard Boire） 著

杨华勇 王若琼 译

人民邮电出版社

北 京

图书在版编目（CIP）数据

写给经理人的数据挖掘书 : 运用大数据解决商业难题 / (美) 博伊尔 (Boire,R.) 著 ; 杨华勇, 王若琼译. -- 北京 : 人民邮电出版社, 2016.6
ISBN 978-7-115-42400-6

Ⅰ. ①写… Ⅱ. ①博… ②杨… ③王… Ⅲ. ①商业信息－数据采集 Ⅳ. ①F713.51

中国版本图书馆CIP数据核字(2016)第095331号

内 容 提 要

海量数据正在向我们袭来，你所在的企业准备好迎接这一挑战了吗？如何找出优质客户名单？如何防止特定的客户群体流失？针对哪些客户开展促销活动能够获得更大的回报？如何利用大数据和小数据，以商业人士和数据专家都能理解和认同的方式提高企业经营绩效？

理查德·博伊尔拥有近 30 年的数据分析和挖掘经验，他将数据挖掘的过程归纳为四个清晰的步骤，分别是锁定商业问题或挑战、建立分析文件、运用正确的工具和技术，以及执行与跟踪。本书围绕这四个步骤，介绍了相关的数据分析和挖掘的方法、工具、技巧、陷阱以及颇有启发性的真实案例。这些知识和经验可以帮助各类读者优化数据挖掘方案，提高投资回报率，并获得实实在在的市场竞争优势。

本书适合各类组织中的管理者阅读，也适合企业中的数据挖掘师、分析师等 IT 人员，以及营销决策者、业务负责人阅读，还可作为高校相关专业的师生的参考读物。

◆ 著 【美】理查德·博伊尔（Richard Boire）
译 杨华勇 王若琼
责任编辑 庞卫军
执行编辑 陈 宏
责任印制 焦志炜
◆人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷
◆开本：700×1000 1/16
印张：17 2016 年 6 月第 1 版
字数：150 千字 2016 年 6 月河北第 1 次印刷
著作权合同登记号 图字：01-2015-2960 号

定 价：59.00 元

读者服务热线：（010） 81055656 印装质量热线：（010） 81055316
反盗版热线：（010） 81055315
广告经营许可证：京东工商广字第 8052 号

Data Mining
For Managers

How to Use Data to Solve Business Challenges

推荐序

大数据、数据挖掘以及预测分析是当今工作中经常会碰到的术语和话题，但是人们真的理解这些概念吗？分析学是一门新兴学科，它现在正处于飞速发展阶段，然而分析对我们来说并不是一个新名词。借助计算机技术捕获、存储、管理以及处理海量数据，我们通过自身能力在很大程度上促进了这门学科的发展。数据挖掘和分析存在于各个行业以及当今社会的方方面面。无论是分析消费者及其购买行为，监测恐怖分子的活动，预测某一个人是否具备好员工的潜质，还是判断一位棒球手是否能够击中迎面飞来的球，都涉及数据挖掘和分析。无论你是否已经意识到这一点，数据挖掘和分析早已经无处不在。

在当今这个大数据时代，短时间内处理

海量数据这件事已经变得稀松平常，但是对大多数企业来说，如何利用大数据仍然是一个巨大的挑战。只有具备丰富经验、知识渊博的人才能在大数据迷宫中找到方向，并将原始数据转化为有意义、有价值的资源。在如今这个技术资本优于人力资本的时代，本书将加快你的学习步调。

理查德·博伊尔的这本书能帮助你从分析学领域的领导者身上汲取经验。早在出现“大数据”这样的概念之前，理查德就已经是数据挖掘领域的佼佼者了。通过阅读本书，你将从他长达 25 年的职业生涯的经验和教训中获益。

理查德解决商业问题时采用的兼具实用性和可操作性的方法贯穿于整本书之中。通过真实的商业案例，你将学会：评估或者定义自己所面临的问题；运用内部或外部数据解决问题；建立、创造以及筛选出最佳的工具或者解决方案；以及正确运用解决方案并评测这种方案所带来的影响。

作者作为导师和教师所获得的经验也将使你获益。你将了解到作者对于利用分析获得成功这方面的洞见。高效的数据挖掘者（或称数据科学家）无法独立工作。因为成功的数据挖掘需要团队合作以及不同学科之间的配合，例如数据库管理员（或称数据技术人员或者 IT 人士）和终端商务用户，这些人一般被称为“主题专家”（Subject Matter Expert，SME）或价值架构师。而最成功的解决方案往往需要人文与科学的结合。

我很荣幸在过去 25 年间以同事、合作伙伴以及朋友的身份与理查德·博伊尔共事。我在数据挖掘以及预测分析领域的知识都来自于他。这本书对读者来说是掌握这些知识的好机会，也是从业于数据挖掘与分析领域的人都应当阅读的一本好书。

——拉里·菲尔勒（Larry Filler）

博伊尔菲尔勒集团（Boire Filler Group）合伙人

Data Mining
For Managers

How to Use Data to Solve Business Challenges

目录

Data Mining For Managers

How to Use Data to Solve Business Challenges

第 1 章

导论

创作一本数据挖掘图书的挑战在于如何避免拾人牙慧，与其他同类图书、论文以及研讨会的内容有所区别。随着数据量的激增，人们谈论这一话题的热情也空前高涨，然而真正为人所熟知的内容仅仅是冰山一角。本书的关注点在于数据本身，以及数据如何撬动商业杠杆，并最终提高投资回报率。数据挖掘之所以流行，是因为大多数商业人士已经意识到了信息的价值，并利用信息作决策，以获得更多利润。对现今的大多数企业来说，数据挖掘已经成为或即将成为商业中极为重要的一部分。因为数据挖掘还是一个相对较新的概念，所以了解相关知识对成功利用这一技术来说至关重要。诚然，这一领域的咨询业务正在逐渐兴起，也有很多人自称是数据挖掘方面的专家。但是企业必须明白一点，数据挖掘仍然属于新兴领域，很少有人在这一领域拥有丰富的经验。同其他学科一样，学习相关知识并掌握相关技能的关键就在于不断实践，并观察实践结果。尽管大多数数据挖掘的经验都来自于直销（Direct Marketing）领域，但是在其他领域数据挖掘的方法和过程都是相类似的。精通数据挖掘的关键在于如何处理数据。数据挖掘研究的是数据，以及我们如何从信息中提炼数据。这是数据挖掘的重点，也是本书将要重点阐述的内容。

数字影响力

经验丰富的从业者表示，随着新型软件和技术的爆炸式增长，他们如今所面临的最大挑战在于优化数据挖掘过程。而另一个棘手问题则是如何在网络环境中运用数据挖掘。例如，在市场营销部门中，从营销活动开始的那一

刻，就可以进行数据挖掘分析。随着分析此类信息成为可能，人们的需求也日渐增长，他们开始追求加快此类分析数量和速度的工具及技术。过去，提炼有用信息大概需要 6~8 周的时间。而有了网络和电子数据之后，我们就能够即时整合信息。正因为如此，渴望掌握这一技术的人才越来越多，企业则希望更快速地获取深刻洞见，并提升自己的决策能力。

服务

数据挖掘领域仍然处于萌芽阶段，观察它将会给商业带来什么样的影响将是一件十分有趣的事情。数据挖掘对技术的依赖表明，这一行业将以工具和软件的形式飞速发展。而这一发展过程也将大幅增加对该领域从业者的需求。现在对数据挖掘从业者的需求正在不断增长，而目前的市场无法满足这一需求，并且大多数高校对此也无能为力。

人文与科学

尽管技术正在朝日趋复杂的方向发展——未来它有可能具备帮助我们发现目标市场的功能，但在任何时候，利用数据挖掘寻找解决方案都离不开人文因素。如果不能运用专业的商业思维去分析数据，那么很有可能得出错误的结论或者建议。

举个例子，某位零售商在数据挖掘中发现这样一个现象：啤酒和纸尿裤的销量成正比关系，并且相互之间影响很大。这是否意味着那些购买啤酒的消费者往往也愿意购买纸尿裤呢？深入调查之后发现，在无意

识的情况下，某家商店的店员恰巧把啤酒与纸尿裤摆在了一起，而那些深夜来买纸尿裤的年轻爸爸们，往往会顺手再买几瓶啤酒。这件事告诉我们，深入调查就能发现原始数据中真正的商机。在这个例子中，由于商店中物品的摆放位置影响了数据挖掘的结果，因此它并不能真正反映消费者的购物意向。

再举一个例子，在某个时间点，新、老客户突然都出现了购买同一家企业产品的倾向。深入调查后发现，这家企业刚刚并购了另外一家企业，于是他们开始针对被并购企业的消费者开展促销活动。根据数据库中的记录，这些消费者购买被并购企业产品的习惯均维持了一年左右。因此，将这些消费者视为新客户显然是不妥的。

数据挖掘中的人文因素能够帮助分析师更好地理解结论。只有充分理解商业运作，才能真正懂得为什么会出现特定的结果。至少，分析师能够更充分地调查商业的不同侧面，从而找出暗藏在结果之下的真正原因。尽管数据挖掘是一种注重数据的技术，但是数据背后隐藏的商业知识才是帮助业内人士得出最佳结论的利器。

数据挖掘并不仅仅关乎技术，它更大程度上是利用技术找出商业解决方案，从而提高投资回报率（Return On Investment，ROI）的手段。这就要求企业在注重投资技术（例如购买软件和新的数据库系统）的同时，也注重投资智慧资本。事实上，比起投资前沿技术，成功的企业往往更加注重投资人才。数据挖掘获得成功的关键就在于找出解决方案，在最小粒度上提高投资回报率。大多数情况下，这个最小粒度指的是消费者个人层面。

数据挖掘既是一门科学，也是一门行业学科，它一直处于发展变化之中。但是，对一位成功的数据挖掘师来说，必须注重数据挖掘中的人文因素。与此同时，还有很多人认为数据挖掘能够为社会带来有益影响。例如，

如果能够为任意一款商品在最恰当的时机找到最合适的消费者，将大大缩减行销成本，从而有效增加边际利润。这种方式往往能够 100% 甚至更高幅度地提高投资回报率。最终，高投资回报率会进一步降低成本，增加收入。在信用风险或者信用欺诈领域（这一领域通常被视为数据挖掘的发源地，我们会在后面的章节中详细论述），分析师需要辨别哪些人有可能无力偿还贷款，或者从欺诈的角度说，哪些人的消费模式有些异常。如果数据挖掘能够降低哪怕 1% 的信用风险或者预防 1% 的欺诈行为，就能够挽回上百万的损失，这些最终都将对盈亏平衡造成直接影响。

行业观察

医疗

通常情况下，在医疗以及健康领域，医生通过分析大量的数据和信息来确认病人的病情。如果能够查看病人病史，并参考其他病人的病史，即使遇到相对罕见的病症，医疗专家也能够及时确诊。病人病史中包含了大量数据，检索这些数据能够帮助医生准确地为病人制定治疗方案。多亏了数据挖掘技术，当今的医疗专家能够分析病人数据，从而迅速为病人制定最佳的治疗方案。

政府与执法

政府部门通常将数据挖掘当成执法工具。如果把数据挖掘看成是一种学习知识的手段，那么，掌握知识的关键就在于能够在数据中发现独特的模式。我们都了解执法中的“侦查过程”，这是指侦查人员利用自己的知识进

行分析。这些知识包括了侦查过程中的线索、过去的经验，以及侦查人员的判断。掌握这种知识需要投入大量的时间和精力。只有掌握了这种知识，侦查人员才能找到独特的模式，从而确认嫌疑人。而有了数据挖掘技术，这种模式就可以在更大程度上实现自动化，我们可以获取大量侦查人员在成千上万的案件中获取的数据和信息。将这些数据和信息汇总在一个分析文件或者数据库中，对它们进行计算和分析，就能够发现特殊的行为模式和事件，从而解决案件。

除了帮助解决案件，这些数据还能用于确认哪些城市容易发生哪种类型的犯罪。在数据挖掘的过程中，通过寻找特定案件的相关信息，我们就可以预测未来事件。有了这些数据，警察就能快速锁定目标。通过优化城市资源配置，我们就可以根据某个城市特定区域的犯罪类型和数量指派相应数量的负责人去解决。例如，在市长鲁道夫·朱利亚尼（Rudolph Giuliani）的领导下，纽约市就在利用这一技术减少重大犯罪的犯罪率，尤其是谋杀。自 20 世纪 70 年代末到 80 年代初，犯罪率高已经成为纽约市的一大特征。现在，纽约市的谋杀犯罪率不到 20 世纪 70 年代末谋杀犯罪率的三分之一。数据挖掘技术的应用是犯罪率下降的主要原因之一。

“9·11事件”之后，将数据挖掘作为执法工具的概念就变得更加明晰。“9·11 事件”之后，美国政府成立了一个部门，专门监测不同政府部门搜集来的个人数据，并利用这些数据对抗恐怖主义。这个庞大的数据库中包含了医疗、保险以及银行等领域的数据。实际上，政府能够收集到一个人一生中所有活动的资料。除了对于个人隐私的担忧，反对这一做法的其他理由是为了阻止恐怖袭击开发相关工具和解决方案需要进行海量的监测。对“9·11 事件”以及预防未来恐怖主义袭击，争论的焦点在于无法搜集到足够的数据来鉴别基地组织恐怖分子的特征，因为我们拥有北美地区大量的人

口统计信息，但我们仅仅有 19 个数据点（即 19 名恐怖分子）。

各种辩论和争议都聚焦于如何利用数据确认潜在的恐怖分子。在这些讨论中衍生出了一个“种族形象定性”的概念，和这种行为可能带来的结果。通过数据挖掘，执法部门能够锁定某一类高危人群（即进行种族形象定性）。但是像种族形象定性这种敏感的处理方式，需要我们在保障公共安全和防止执法歧视之间寻求一种平衡。许多重要的利益相关者都需要参与到这种具有启发性的讨论中来。我希望政府最终能够制定出一种政策，为执法过程中的种族形象定性这种行为提供指导和规范。

数据挖掘实战经验

1983 年获得 MBA 学位之后，我开启了自己的直销生涯。幸运的是，我能够把自己学到的知识应用到统计学中。就像大多数大学生一样，我自信能够在企业中有所作为。然而，尽管我确实在自己的岗位上为公司作出了贡献，但是公司对我的帮助更大，它让我能够在商业中应用统计学，从而开辟自己的事业。在这个过程中我发现，我们需要遵循统计学中那些晦涩难懂的规则，但也需要灵活变通。例如，在大多数情况下，我们都不会按照传统的假设对样本群进行统计分析，因为商业环境中只有非常态样本才需要这样做。常态样本是指反映样本某种特征（如年龄、收入等）的数据中，一半低于平均值，另一半高于平均值，数据呈钟形曲线分布。尽管如此，商业领域还是一直在应用统计分析的方法，这纯粹是因为数学家的恐怖能力——他们能够计算出令人满意的数据，让人看到极大的增量利润。关于直销领域的统计学应用，我最想强调的也许就是不必完全遵循统计学规则。在学术界，最

常见的回归分析数据就是 R^2 的值大于或等于 0.7（我们会在后续章节中详细讨论这个话题）。而在直销领域，数据通常显示 R^2 的值小于或等于 0.05，而这种情况仍然属于正常范围。学术界中的非常态数据放在商业活动中就属于正常数据。出现这种差异的原因在于，不同领域对特定解决方案的看法完全不同。关于这个问题，我们也会在后续章节中详细讨论。

正如任何一本意义重大的图书一样，这本书与其他作品有所区别。如果你希望看到一本论述数据挖掘领域数学与技术之间细微差别的作品，那么请放下这本书。但如果你希望寻找一本从商业实践角度解读数据挖掘的读物，那么本书绝对不会让你失望。我将从实践角度聚焦于数据挖掘，讲述自己 30 多年来在商业领域应用这些工具和技巧的经验，让你有所收获。但是我绝不是要用这本书把读者塑造成一位数据挖掘专家。如果你在数据挖掘领域已经有一定的经验，那么这本书能够帮助你从另一个角度理解数据挖掘。还要明确的一点是，这本书讲述的是加拿大数据挖掘领域中的经验。现在市面上的数据挖掘图书大多都是美国人基于美国市场而得出的经验，现在换一种角度，相信能够帮读者从整体角度理解该领域。如果这本书对你来说具有参考价值，能够帮助你理解特定的数据挖掘战略和这一技术对某个商业问题的影响，那么我就实现了创作这本书的目的。

衷心希望你们喜欢这本书。

Data Mining For Managers

How to Use Data to Solve Business Challenges

第 2 章

从历史角度解读数据挖掘的发展

创作任何一个学科的书，都需要先了解其历史。数据挖掘经历过怎样的挑战和发展，才成为了今天商业领域中的前沿技术？为了了解数据挖掘的历史，我们要先从早期运用这一技术的实践者们谈起。他们就是大型目录直销公司和出版社的直销人员。

1982 年，刚刚获得 MBA 学位的我有幸成为这些先锋直销企业中的一员，我在《读者文摘》（*Reader's Digest*）获得了一份工作。那时，我在列表选择部门做回归分析。尽管我们为直邮开发出了回归预测模型，但是在涉及单个客户数据时，我们仍然需要做分析和市场细分的工作。我和部门同事们是《读者文摘》的客户信息专家，并且亲眼见证了商业文化带来的巨大投资回报率。自然而然的，我认为这是大多数企业都有的一种典型业务流程。但是，通过跟那些 20 世纪 80 年代初在各个行业龙头企业中打拼的同事们交流，我发现事实并非如此。当时，我们这些人没有那些了不起的头衔，像业务分析专家、新概念分析师、数据科学家或者客户关系管理（Customer Relationship Management，CRM）分析师这些职位根本就不存在，因为当时还处在数据挖掘和预测分析的萌芽阶段。

我意识到了数据挖掘技术在提高整体投资回报率方面的巨大商业潜能，并了解到《读者文摘》只是少数应用这一技术的企业之一。我看到了数据挖掘领域巨大的创业机遇。但是，这一领域的一大阻碍就在于技术。在当时的直销市场中，可应用统计技术的计算设备成本高昂，因为那是 1982 年，个人计算机还没有问世，我们所有的工作都是在主机[①] 上完成的。但在当时，《读者文摘》的分析环境非常稳健、高效。我们针对每一位客户都建立了广

① 主机是指可同时供多人使用的大型计算机。——译者注