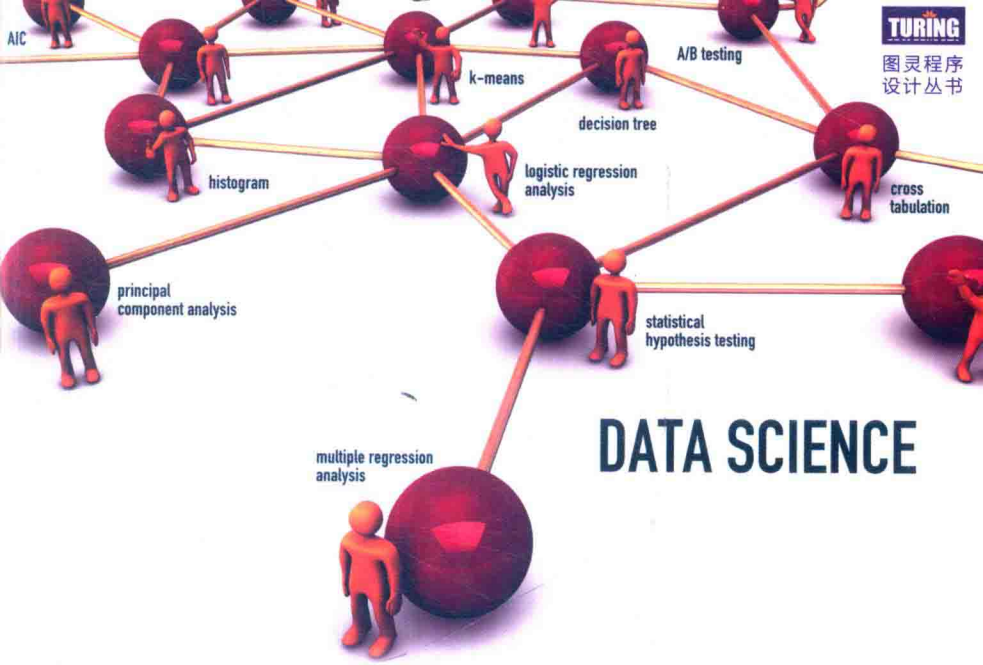


TURING
图灵程序设计丛书

数据分析实战

[日] 酒卷隆治 里洋平 / 著 肖峰 / 译

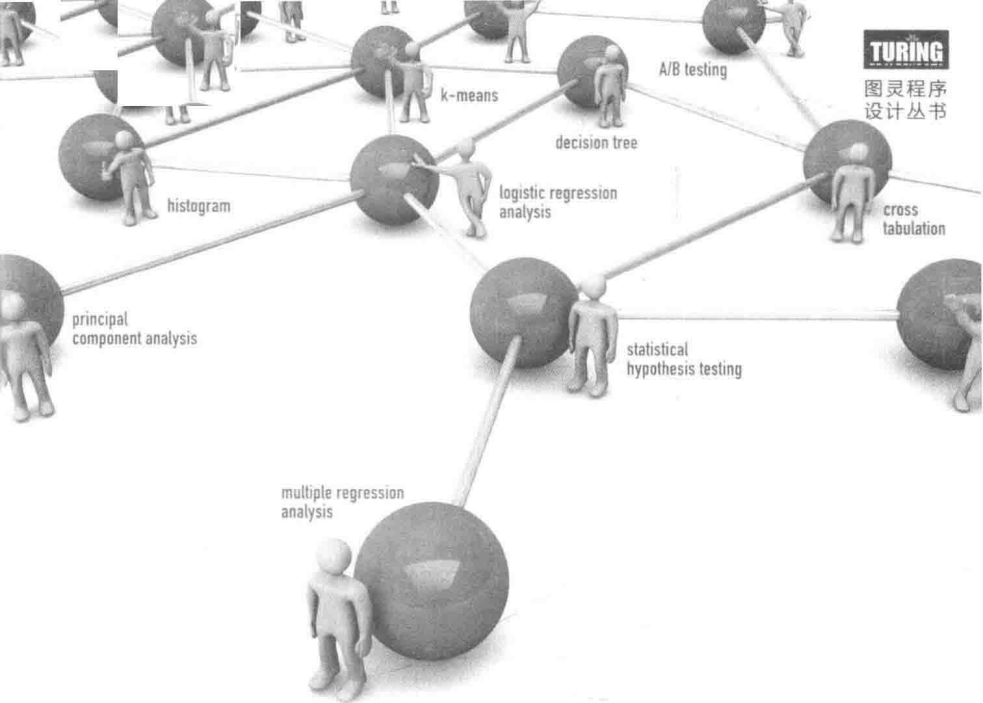
◆ 8个真实的商业案例
◆ 可下载的原始数据+R语言代码

玩转数据分析 告别纸上谈兵

让你学会用数据分析解决实际问题！

叉列表统计 / A/B测试 / 多元回归分析 / 逻辑回归分析
分分析 / 决策树分析 / 机器学习……

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



TURING
图灵程序设计丛书

数据分析实战

[日] 酒卷隆治 里洋平 / 著 肖峰 / 译

人民邮电出版社
北 京

图书在版编目（CIP）数据

数据分析实战 /（日）酒卷隆治，（日）里洋平著；肖峰译. -- 北京：人民邮电出版社，2017.6
（图灵程序设计丛书）
ISBN 978-7-115-45453-9

Ⅰ. ①数… Ⅱ. ①酒… ②里… ③肖… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2017）第083684号

内 容 提 要

本书由实战经验丰富的两位数据分析师执笔，首先介绍了商业领域里通用的数据分析框架，然后根据该框架，结合8个真实的案例，详细解说了通过数据分析解决各种商业问题的流程，让读者在解决问题的过程中学习各种数据分析方法，包括柱状图、交叉列表统计、A/B测试、多元回归分析、逻辑回归分析、主成分分析、聚类、决策树分析、机器学习等。特别是书中使用的数据都是未经清洗的原始数据，能够让读者了解真实的数据分析流程，避免纸上谈兵。

本书适合各大公司的算法工程师、数据分析师、数据挖掘工程师以及今后立志从事相关工作的高校学生阅读。

◆ 著　[日]酒卷隆治　里洋平
译　肖　峰
责任编辑　杜晓静
执行编辑　刘香娣
责任印制　彭志环

◆ 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
北京隆昌伟业印刷有限公司印刷

◆ 开本：880×1230　1/32
印张：8.375
字数：258千字　2017年6月第1版
印数：1－3 000册　2017年6月北京第1次印刷

著作权合同登记号　图字：01-2015-6982号

定价：45.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

译者序

伴随着互联网的高速发展和以云计算为代表的技术创新，过去难以收集和存储的大量数据得以集中管理和使用，大数据时代已经来临。然而，如何发掘大数据这座金矿，使之在商业领域体现其最终价值，这属于数据科学（Data Science）的范畴，也是数据科学家（Data Scientist）的工作职责所在。

要想成为一名合格的数据科学家，不仅需要拥有深厚的统计学等理论基础，更需要有较强的业务能力、对数据的敏感性以及处理实际商业数据的经验。对于这一点我深有体会。在日本留学期间我有幸进入乐天技术研究所（Rakuten Institute of Technology），这也是我第一次接触到生产环境下的数据。然而在面对公司实际问题时，我拿着海量的各种日志数据却经常有种不知从何下手的感觉。

这是因为现实中我们需要面对的不再是抽象的理论，而是真实的生产环境下的问题，并且生产环境下的数据噪声远比实验室中使用的数据要高。面对这些问题，作者在本书中给出了很好的答案，在理论与实际之间搭建了一座桥梁。本书没有教条似的说教，而是使用某个游戏公司实际工作中所遇到的问题作为案例，让读者在解决具体问题的过程中理解数据分析的整个流程。特别是当涉及逻辑回归、聚类、主成分分析等理论性较强的内容时，本书没有使用大篇幅的理论解释和数学推导，而是在实践过程中对各类机器学习算法进行详尽而又通俗易懂的介绍。更难能可贵的是，为了使得本书的内容更贴近生产实践，书中使用的各种案例的数据均为未经清洗的原始日志数据，这使得读者可以接触到实际生产环境下的数据，从而避免了纸上谈兵。

本书适合各大公司的数据分析师、算法工程师、数据挖掘工程师以及今后立志从事数据分析和数据挖掘工作的高校学生阅读。

在翻译本书的过程中，非常感谢图灵公司的各位编辑所给予的帮助。同时也感谢我的妻子李春姬在怀孕期间对我工作的默默支持，感谢我的儿子明诚一直很乖地陪伴。希望本书能够对读者在实际数据分析工作中有所帮助。

肖峰

2016 年 12 月于北京

前　言

如果你每天都在和数据分析业务打交道，那么很可能经常听到下述说法。

A. 在数据收集和分析上投入了巨大的成本，然而效果和预期相距甚远。
B. 从数据分析部门拿到了详细的报告，然而内容却非常晦涩难懂。
C. 数据大致都有了，然而因为工作很忙，没能很好地进行分析利用。
D. 数据全部都保存了下来，但是却不知道如何才能很好地利用这些数据。
E. 虽然每天都会检查和核对重要的数据，但对于如何利用这些数据来指导手头的工作，却还不是很清楚。
F. 我们靠的是相关负责人的业务经验，这个比数据更加可靠。
G. 数据？分析？还是算了吧，先把自己能力范围内的事情做好再说。

数据分析在很多情况下对提升工作效率很有帮助。根据我们的亲身经验，在各行各业的不同领域，尽管数据分析所起的效果大小可能有差异，但确实有着非常多的成功案例。然而，正如上述各种说法所反映的那样，实际上在企业内部的各个部门里，还有很多数据分析的问题没有解决。

站在分析者的角度来看，关于上述各种说法所产生的背景，可以大体总结出以下几点。

A′. 数据分析，特别是机器学习等，被误认为是一种普通人无法理解的魔法一样的东西。
B′. 数据分析只是在利用复杂的数值分析来对实际现象做出解释。
C′. 数据是保存了下来，但数据分析还需要时间和人力等。
D′. 数据是保存了下来，但是却不知如何来做数据分析。
E′. 虽然通过营业额等重要的数据能够把握经营的现状，然而这些

数据对于今后的经营策略有何指导作用，却不甚明了。

F′. 不知道数据分析和业务经验其实可以优势互补，并产生协同效应。

G′. 即便是在那些只需要一味蛮干的工作中，如果同时使用数据分析的手法，也可以有效地提高工作质量。然而很多人却不知道这一点。

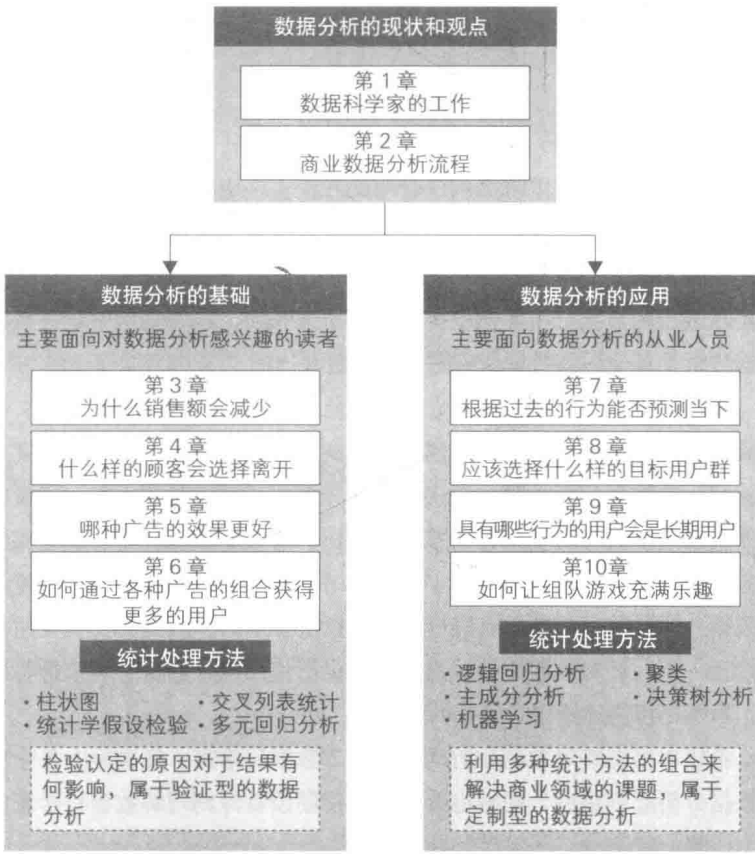
一般来说，为了成功地完成某件事，我们需要经历“知道、理解、掌握、精通”这几个阶段。要想从一个阶段迈入下一个阶段，就必须跨越阶段间的巨大障碍。从上述各种问题来看，虽然“不知道、无法理解、无法掌握、无法精通”这些阶段各有不同，但问题都可以归结到同一点，那就是不了解数据分析在商业领域是完全可以获得成功的。

也就是说，我们或许可以将上述这些问题的原因归结为信息匮乏。

正因如此，我们希望通过本书向读者展示商业活动中数据分析的一些经典案例，通过这些案例来向读者揭示数据分析的作用。本书的读者对象主要有：

- **关注商业数据分析的人**
- **在实际工作中从事商业数据分析的人**

所谓关注商业数据分析的人，是指将来打算从事数据分析工作的学生，或者工作没多久的商业人士，以及想把数据分析作为自身技能之一的商业精英，还有正在筹划设立数据分析部门的公司经营管理层。



本书的结构和目标读者

第 1 章主要介绍从事商业数据分析的数据科学家的现实情况。第 2 章主要介绍在商业领域里通用的数据分析框架。

从第 3 章开始，基本上就是根据第 2 章提到的数据分析框架来介绍具体的数据分析案例。第 3 章到第 6 章主要介绍数据分析的基础。因为这部分是基础的东西，所以在商业环境下很少有机会用到，但本书会尽可能地将基础知识和实际业务相结合，对商业数据分析相关的观点和实际的业务进行充分说明。

从第 7 章到第 10 章的后半部分是具体案例分析，这部分内容是笔

者参考了实际从事商业数据分析的同行的工作而写成的。在应用篇中，我们将介绍更加高级的方法，比如将多种方法组合，或者比较几种方法中哪种更有效等。此处介绍的数据分析的应用案例主要针对在社交游戏产业或 IT 以及其他行业从事数据分析业务的人员，这些行业的数据分析人员很多时候都可以将案例中介绍的方法直接应用在工作中。对于其他行业的分析人员，也希望他们能够参考这些应用案例，灵活应用在自己的工作环境中，并提出新的分析观点。另外，本书的创新之处在于，从实际进行的数据分析案例中挑选了 4 个来介绍，这些案例中所进行的数据分析是其他同类数据分析图书中不曾有过的。

本书提供了各章中使用的数据，以及加工和分析数据时使用的 R 语言脚本代码。读者可以一边阅读各种案例，一边从数据分析师的角度来体验实际的数据分析流程。在讲解数据分析的其他同类图书中，经常使用那些和书中内容高度相符的数据来分析，但大多数读者会发现在实际业务中使用书中的方法却很困难。而本书的各个案例提供的都是最原始的数据，这些数据都是没有经过处理、杂乱无章的，因此需要在使用前先处理一下。针对这些原始数据，如何灵活使用统计分析工具来进行处理，本书尽可能地给出了详细的解释。

希望本书能够对数据分析从业人员（数据科学家）在实际业务中应用数据分析起到帮助，同时也希望能够帮助读者深刻理解数据科学家的工作内容，并和他们一起更好地完成工作。

笔者

2014 年 5 月

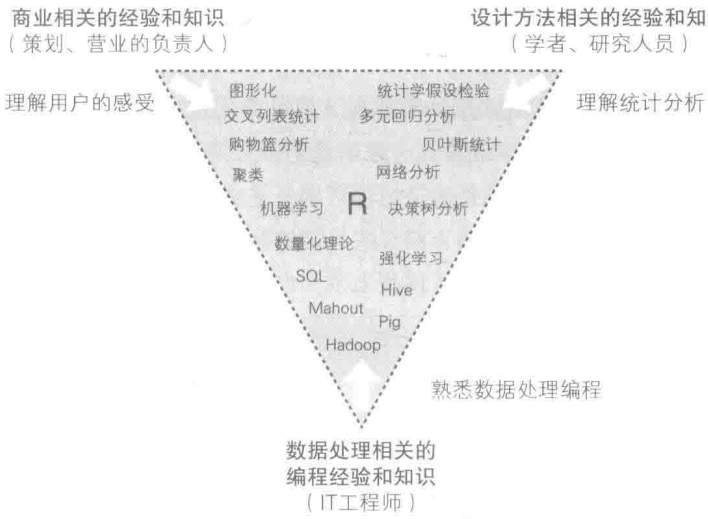
目录

[分析应用]篇

第1章

数据科学家的工作

市场营销的主要从业者并不是技术人员，所以很难立即找到合适的从业者。现在市场营销的工作主要由3种职业的人员相互协作来完成。



1.1 什么是数据科学家

什么是数据

首先我们需要回答的问题是：什么是数据？自古以来，人们通过观察客观现象，并对观测的数据进行分析，从而发现了各种各样的规律和法则。比如，开普勒根据天体观测的数据，发现行星是以太阳为中心沿着椭圆形轨道运行的。在统计学领域，棣莫弗通过对多种游戏的观测，提出了中心极限定理，而这个定理正是正态分布理论的核心。也就是说，如果我们把过去的某些事情记录下来，并由此推导出一些可能的规律，而这些规律又能够解释当前某些事情的前因后果，那么基于这样的过程，就可以根据现在的事情预测未来。也正是各领域内这样的规律不断被发现、考证和研究，才推动了科学的发展。而数据则是发现和验证这些规律的关键，是十分宝贵的材料。

数据在商业中的应用

近年来，随着网络或 POS（Point Of Sale，销售点）系统的快速发展，人的行为可以简单地作为数据存储起来，尤其是涉及购买行为的数据存储了特别多。如果我们能够从这些存储的数据中推导出与用户购买行为相关联的规则，那么这将会颠覆商业世界里一直以来所遵循的某类经验，而在这种科学的数据分析的指引下，将会打开新的商业局面。

话说回来，其实在日常生活中我们也经常使用数据分析。比如，大

多数人都会定期去称体重。称重的时候，体重数值本身并没有多大意义，我们并不能说“50 kg 比较好”或者“体重到了 51 kg 就糟糕了”。也就是说，50 kg、51 kg 之类的体重数值并没有什么绝对的意义，而仅仅是身体重量的观测数据而已。然而根据这些观测数据，我们可以完成以下事情。

- ★根据自己的性别、年龄、身高等其他数据，推断出理想的健康体重，并将其设定为目标值
- ★通过长期跟踪测量体重，得到体重随时间变化的观测数据，并将过去暴饮暴食等行为与体重的变化联系起来，从而反省自己的行为
- ★通过收集拥有理想身材的人的运动和饮食生活等方面的数据，效仿他们健康的生活方式

很多人为了达到控制体重的目的，都会根据自己的体重观测数据，选择采取上述某种具体的行为。

在商业领域里也是一样。人们通常会通过观测数据来推测出某种因果关系，再用这种因果关系预测未来，或者控制原因以达到预期的结果。最近，越来越多的企业开始为这种工作增设一个专门的职位，招募被称为数据科学家的人才加入。

为什么需要数据科学家

在商业领域，从观测数据中推导出因果关系曾经是市场营销部的工作内容，并由专门的市场营销人员来承担此项工作。市场营销部的主要目的是“理解用户的需求，迎合用户的口味开展商业活动”。

具体来说，市场营销部的主要工作有分析各种销售额数据，分析对于广告和新产品认知度以及正在销售商品的满意度问卷调查，分析售后部门收到的用户咨询等，并以“理解用户的需求，迎合用户的口味”为原则，开展企业的市场营销活动。市场营销活动在各个领域都取得了一定的成果，市场营销部才有继续存在的价值。

然而，随着信息技术的发展，商业环境也在发生着变化，企业可以

保存大量的商业日志。在商业活动中，通过尽早地对自身的细节问题进行反复修正来顺应用户需求的服务一直存在，于是就有人提出能否将这大量的日志应用于一直以来的市场营销活动中。也就是说，在分析过程中不仅要考虑到已有的市场营销数据，而且还要针对实时性较高的大量日志数据做出快速分析。

为了满足这种需求，就需要能够直接分析日志数据的人，也就是“会写代码的市场营销人员”。这类会写代码的市场营销人员现如今也被称为数据科学家。也就是说，为了应对近些年来商业环境的变化，在过去不曾有过的领域里产生了极大的可能性和值得关注的地方。然而，过去市场营销的主要从业者并不是技术人员，所以很难立即找到合适的从业者，现在这个工作主要由3种职业的人员相互协作来完成。