数据分析与决策技术丛书

Practical Data Analysis Cookbook

数据分析实战

[美] 托马兹·卓巴斯（Tomasz Drabas）著

刁寿钧 译

微软数据科学家系统讲解数据分析与建模技术，并提供Python源码示例和大量实战技巧

机械工业出版社
China Machine Press

需要全本请访问 www.ertongbook.com

数据分析与决策
技术丛书

Practical Data Analysis Cookbook

数据分析实战

[美] 托马兹·卓巴斯（Tomasz Drabas）著

刁寿钧 译

学院图书馆
藏书专用章

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据分析实战 /（美）托马兹·卓巴斯（Tomasz Drabas）著；刁寿钧译．—北京：机械工业出版社，2018.4
（数据分析与决策技术丛书）
书名原文：Practical Data Analysis Cookbook

ISBN 978-7-111-59779-7

I. 数… II. ① 托… ② 刁… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 085790 号

本书版权登记号：图字 01-2016-8655

Tomasz Drabas: *Practical Data Analysis Cookbook* (ISBN: 978-1-78355-166-8).

Copyright © 2016 Packt Publishing. First published in the English language under the title "Practical Data Analysis Cookbook".

All rights reserved.

Chinese simplified language edition published by China Machine Press.

Copyright © 2018 by China Machine Press.

本书中文简体字版由 Packt Publishing 授权机械工业出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

数据分析实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：张志铭 责任校对：殷 虹
印 刷：北京文昌阁彩色印刷有限责任公司 版 次：2018 年 6 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：18.25
书 号：ISBN 978-7-111-59779-7 定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

版权所有·侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

华章IT
HZBOOKS | Information Technology



Forword 推 荐 序

程序员、科学家、工程师之间有什么差别？

这个问题，我问过许多“软件工程师”，大家似乎都没法说得特别清楚，常见的答案是从名字上来区分：程序员只关心代码，工程师负责的是系统，科学家的理论知识非常深厚。它的潜台词是：你看，我是个工程师，我不只关心代码，我还为系统负责，但是，你没法对我的理论知识做太多要求。

看来，实际情况也确实如此。前些年，我在一支颇有效率的开发团队里，组织大家“补习”数据结构和算法。我惊奇地发现，不少主力开发人员做起业务功能来异常拿手，毫不畏惧，但面对简单的“理论问题”——比如如何从一个数组中找到和最大的两个元素——竟然束手无策。而且很多人都认为，这不是问题。

我想，这就是很大的问题。

“工程”这回事，是什么时候出现，并让大家意识到它不等于“手艺”的呢？具体时间或者不可考，但有个故事很能说明问题。

1742 年，教皇本尼迪克特十四世（Benedict XIV）需要派人诊断罗马圣彼得大教堂拱顶出现的裂纹。传统上，这种事情总是要找建造经验最丰富的工匠。但是这次不一样，教皇把任务指派给了三位数学家，其中一位还曾编辑和注释过艾萨克·牛顿的《自然哲学的数学原理》。在那个年代，他们的诊断方法和结论都引发了巨大的争议，因为其违背了无数工匠的经验和直觉。按照三位数学家的结论，拱顶的箍环承受不了水平的推力，必须新增三个带链条和铁钉的铁环，才能确保建筑的完整。

他们的建议被采纳了。今天，如果你去罗马，仍然可以看到完整的圣彼得大教堂。

土木工程师兼历史学家斯特劳布评论说：这份报告在土木工程史上有划时代的意义……重要性在于，与所有的传统和常规相反，对建筑结构的稳定性的勘测，不是建立在经验规则和静态感觉的基础之上，而是建立在科学的分析和研究之上。

从此大家逐渐相信，建筑不再是一门“手艺”，要想建造更复杂、更伟大的建筑，科学和研究是无论如何离不开的。今天，如果土木工程师在工作时不依照模型、理论、计算，

需要全本请访问 www.ertongbook.com

而是完全按照经验和直觉，哪怕他的经验再丰富，也不能称为“工程师”。

在我看来，软件开发，在某种程度上也处于相同的时间点。我承认，之前的大量开发工作，不需要太多理论和科学知识，单纯凭经验和直觉就可以完成。但是如今，我们已经无可避免地被卷入大数据的洪流之中——哪怕是“传统”的互联网开发，也已经大不相同了。我们写的每一个功能，都可能被成千上万人，在成千上万的场景下，成千上万次地使用。在整个过程中，成千上万的埋点、成千上万的传感器，会留下巨细靡遗的数据。如何还原场景、找到问题、做出改进，答案往往就藏在这些数据当中，谁看得懂、玩得转这些数据，谁就能找到答案。

拿简单的“上课前给用户打电话通知”来说，它绝不再是“调用供应商接口发一个语音”那么简单。提前多久给用户打电话反馈最好？什么样的语音信息最容易接受？各地用户有什么特别偏好？不同年龄段的用户接受程度如何？……如果我们承认“用户体验”重要，那么搞清楚这些问题便也很重要。公司搞不清楚，就会被用户嫌弃。工程师搞不清楚，就只能被动接受产品经理的指挥。而这些问题的答案，只能来自对数据的积累和分析。

不要妄想“数据科学家”能帮忙解决这些问题，“数据科学家”太宝贵了，只能用在业务价值最关键的场合。更多的场合，工程师只能挽起袖子自己上场。然而，大部分工程师目前的数据处理能力还局限于极值、算数平均、方差等等少数几项。许多工程师也承认，数据分析能力很重要，也希望学习，可是打开数据分析的专门教材，一看到密集的公式，就已经打了退堂鼓。

怎么办？

我认为，阅读《数据分析实战》是个不错的出路。之所以这么说，有四个原因：第一，这本书的风格是“代码先行”，而且是可以运行的 Python 代码先行，对于要用到的各种工具包、类库的安装，都有详细的说明，这种风格对于工程师来说，天然有亲切感；第二，例子都是非常实际的，无论是银行电话营销数据分析，还是电网的发电量分析，以及影评文本分析，都很容易和生活经验结合起来，容易理解；第三，非常注重出图，全书用了相当多的篇幅讲述画图，而且是用 JavaScript 直接出图，图形非常有助于建立直观的认识，JavaScript 画图几乎可以保证“人人都可以完成”；第四，译者是工程师出身，在数据分析这个行业有足够丰富的经验，同时也足够谦虚，对于译文的质量来说，这都是很好的保证。

实际上在我看来，普通的技术人员，无论之前做的是什么工作，只要不是极端排斥数据分析，那么阅读这本书，都可以建立对数据分析的准确认知，并且还能真刀真枪地玩上几下：看，数据分析就是这样。一旦掌握了基础的知识和技能，无论把它们当作继续深入学习的基础，还是直接应用到现有工作当中去，都会有巨大的收获。

不要在 1742 年以后再当一个不懂数学的建筑工匠，不要在大数据时代当一个完全不懂数据分析的工程师。如果认同这个判断，欢迎你把本书当成高性价比的选择。

余　晟

The Translator's Words 译 者 序

作为译者，我其实不太敢说些什么。一是译者的话当然不如作者的话重要。作者尚未开口，译者先在读者面前高谈阔论，怕是难免佛头着粪引人讥笑。二是译者的话也可能会给读者带来一些令人啼笑皆非的影响。我还记得大二那年，因为恺蒂作了序，我在图书馆仔细阅读了“译林少儿文库”中的一本书——从序言开始仔细阅读。读完序言就已经了解了大致情节。等到最后二十页，作者要给我带来震撼时，我只好假装不知道真相，配合性地震撼了一下。这本书就是《少年 Pi 的奇幻漂流》。

但是，幸好摆在你面前的是一本技术书。而且是工具性质的技术书。这就好办了。不用担心译序的罅隙泄露了情节。而我毕竟译完了全书，对本书内容算是比较熟悉，也有义务介绍一下这本书，希望对你的使用有些助益。

本书大体上分三部分。第一部分是数据的基本处理，这部分内容读者可能最熟悉，用到的频率可能也最高；第二部分可以看成大学里人工智能概论课的补充材料，包括分类、聚类、降维、时间序列等主题；第三部分就是一些更深入的课题了，比如自然语言处理。

作为工具类技术书，本书讲的都是具体的实战技巧，非常实用。读者可以根据实际需要，挑选特定的主题阅读。不过我还是建议大家完整地读一遍。作者很贴心，大部分知识点都给出了可以深入了解的参考资料。不妨通过阅读本书，拓宽一下技术视野。

最近两年，DeepMind 公司不停地刷头条。2016 年年初 AlphaGo 战胜了李世石，之后又有 Master 缔造了连胜神迹，AlphaGo Zero 战胜了 Master。近期推出的 AlphaZero，经过很短时间的训练，就超越了最强的国际象棋软件、最强的将棋软件，以及最强的围棋软件——之前的 AlphaGo Zero。围棋界和技术界都在向阿老师及其缔造者学习。AI 与数据科学也变得更加广为人知。毕竟我们不仅要靠个人奋斗，还要考虑历史进程。

不过不管行业风口如何变幻，希望我们都能立足技术，从实际业务出发，潜心学习。老子说千里之行始于足下，庄子也讲庖丁解牛丈人承蜩。我们不是一个简简单单写代码的人，我们是程序员，是要做数据科学家的。与大家共勉。

本书得以出版，首先要感谢机械工业出版社的王春华老师与张志铭老师，是他们的建议使我开始本书的翻译，也感谢他们在翻译过程中对我的信任、宽容与帮助；感谢姜承尧老师、余晟老师在技术和专业翻译方面给我的指导，让我获益良多；感谢陈文备、杨天宇、束文奂、康墨、胡林军和刘祯，他们参与了全书译稿的审校；还要感谢我的好妻子翁联吉，感谢她对我的理解与支持。

如果你对本书内容有任何疑问，欢迎与我联系。电子邮件地址：shoujun.diao@gmail.com。GitHub 账号：diaosj。祝各位阅读愉快。

刁寿钧

Preface 前　言

数据分析与数据科学已经成功引起了各行各业的注意。当下产生的数据总量已让人惊叹，并且这个数据量每天仍在增长；随着手机使用量的激增，人们对 Facebook、Youtube、Netflix 或其他 4K 视频提供方的访问将越发地倚重云计算，这是我们可以预见的必然趋势。

数据科学家的工作内容包括但远不限于清理数据、转换数据和分析数据，为客户提供业务洞察力，监控公司服务的健康情况，并且自动呈现推荐以促成交叉销售。

在本书中，你将学到如何读取、写入、清理和转换数据——这些工作最为耗时，但也最为关键。接着，会提供相当广泛的工具与技巧——可以说是数据科学家行走江湖的必备技能——内容涉及分类、聚类与回归，图论与时间序列分析，以及离散选择模型与模拟。每一章会给出一系列 Python 代码示例，帮你演练那些今后作为数据科学家会遇到的任务。

本书内容

第 1 章讲解了利用多种数据格式与数据库来读取与写入数据的过程，以及使用 OpenRefine 与 Python 对数据进行清理。

第 2 章描述了用于理解数据的多种技巧。我们会了解如何计算变量的分布与相关性，并生成多种图表。

第 3 章介绍了处理分类问题的种种技巧，从朴素贝叶斯分类器到复杂的神经网络和随机树森林。

第 4 章解释了多种聚类模型；从最常见的 k 均值算法开始，一直到高级的 BIRCH 算法和 DBSCAN 算法。

第 5 章展示了很多降维的技巧，从最知名的主成分分析出发，经由其核版本与随机化版本，一直讲到线性判别分析。

第 6 章涵盖了许多回归模型，有线性的，也有非线性的。我们还会复习随机森林和支持向量机，它们可用来解决分类或回归问题。

第 7 章探索了如何处理和理解时间序列数据，并建立 ARMA 模型以及 ARIMA 模型。

第 8 章介绍了如何使用 NetworkX 和 Gephi 来对图数据进行处理、理解、可视化和分析。

第 9 章描述了多种与分析文本信息流相关的技巧：词性标注、主题抽取以及对文本数据的分类。

第 10 章解释了选择模型理论以及一些流行的模型：多项式 Logit 模型、嵌套 Logit 模型以及混合 Logit 模型。

第 11 章涵盖了代理人基的模拟；我们模拟的场景有：加油站的加油过程，电动车耗尽电量以及狼—羊的掠食。

阅读准备

阅读本书，你需要一台个人计算机（Windows、Mac 或 Linux 机器均可），配置好 Python 3.5 环境；我们使用的是 Anaconda 版本，可从这里下载：https://www.continuum.io/downloads。

书中使用了多种 Python 模块：pandas、NumPy/SciPy、Scikit-learn、MLPY、StatsModels、PyBrain、NLTK、BeautifulSoup、Optunity、Matplotlib、Seaborn、Bokeh、PyLab、OpenPyXI、PyMongo、SQLAlchemy、NetworkX 和 SimPy。大部分已经预装在 Anaconda 版本中了，也有一些需要通过 conda 安装器或 python setup.py install 命令进行安装。如果你机器上缺一些模块，也没有关系，我们会指导你完成安装。

我们也使用了一些 Python 之外的工具：用于数据清理和分析的 OpenRefine，用于可视化数据的 D3.js，用于存储数据的 Postgres 和 MongoDB 数据库，用于可视化图的 Gephi，以及预测离散选择模型的 PythonBiogeme。我们也会提供详细的安装指导。

目标读者

本书的目标读者是，所有想进入数据科学领域、通过解决业界实战问题来练出一身本领的同学。另外，有经验的从业人员也能从本书高阶主题中的一些示例里找到一些乐趣。

板块

这些板块会在本书中频繁出现：“准备”“怎么做”“原理”“更多”以及“参考”。

为了提供完整而清晰的指导，我们将按下面的方式使用这些标题：

准备

这部分介绍技巧的目标，并描述相关软件的安装与设定。

怎么做

这部分包含技巧的步骤。

原理

这部分通常是前一部分的详细解释。

更多

这部分会附加一些信息，帮助读者更了解对应技巧。

参考

这部分会列出一些有助于了解技巧其他信息的链接。

本书约定



表示警告或重要的提示。



表示建议与技巧。

下载示例代码及彩色图像

本书的样例源码，以及所有的截图和样图，可以从 http://www.packtub.com 通过个人账号下载，也可以访问华章图书官网 http://www.hzbook.com，通过注册并登录个人账号下载。

需要全本请访问 www.ertongbook.com

致　谢 *Acknowledgements*

首先，我要感谢我的妻子 Rachel 与女儿 Skye；感谢她们鼓励我接受挑战，感谢她们容忍我整日整夜地编程与写作。她们最棒了，而我对她们的爱意无边无际！

在学习编程时，Tomasz Bednarz 是我的良师益友——感谢你！我也要感谢我的现任与前任经理 Mike Stephenson 和 Rory Carter，以及各位鼓励我完成本书的同事和朋友。

特别感谢我的两位导师 Richard Cheng-Lung Wu 博士与 Tomasz Jablonski 博士。与 Tomasz 合作的项目引发了我对神经网络的兴趣——这是我永远不会忘记的。没有 Richard 的帮助，我可能都拿不到博士学位，感谢他的帮助、引导和友情。

About the Author 关于作者

托马兹·卓巴斯（Tomasz Drabas）是微软的数据科学家，目前工作于西雅图。他拥有超过 13 年的数据分析经验，行业领域覆盖高新技术、航空、电信、金融以及咨询。

2003 年，Tomasz 获得战略管理的硕士学位后，从位于波兰华沙的 LOT 波兰航空公司开启了他的职业生涯。2007 年，他前往悉尼，在新南威尔士大学航空学院攻读运筹学博士学位；他的研究结合了离散选择模型和航空作业。在悉尼的日子里，他曾担任过 Beyond Analysis Australia 公司的数据分析师，沃达丰和记澳大利亚公司的高级数据分析师 / 数据科学家，以及其他职位。他也发表过学术论文，参加过国际会议，并且担任过学术期刊的审稿人。

2015 年，他搬到西雅图，开始在微软工作。在这里他致力于解决高维特征空间的问题。

关于审稿人 *About the Reviewers*

Brett Bloomquist 拥有数学学士学位和计算机科学硕士学位，精通是计算机辅助几何设计。他在软件行业拥有 26 年工作经验，专注于几何建模算法和计算机图形学。最近，Brett 正在以数学和图形学背景申请首席数据科学家。

Khaled Tannir 拥有 20 多年大数据技术和机器学习经验，以及 8 年左右的数据挖掘经验，是一位富有远见的方案架构师。

他在前述技术领域被认为专家，还拥有电子学学士学位与系统信息架构硕士学位。目前正在攻读博士学位。

Khaled 拥有超过 15 项证书（R 编程、大数据等），是微软认证方案开发人员（Microsoft Certified Solution Developer，MCSD），也是一位技术达人。

他在许多法国公司工作过（最近在加拿大工作），主导软件方案的开发与实现，并发表技术演讲。

他是《RavenDB 2.x Beginner's Guide》和《Optimizing Hadoop MapReduce》的作者，两本书都由 Packt 出版公司出版（并有简体中文版）；也是 Packt 出版公司《Pentaho Analytics for MongoDB》《MongoDB High Availability》和《Learning Predictive Analytics with R》这些书的审稿人。

他喜欢拍摄风景，旅行，打视频游戏，以及用 Arduino、Raspberry Pi 和 .Net Gadgeteer 创建有趣的电子装置，当然还喜欢与家人共度时光。

可以通过 LinkedIn 或 contact@khaledtannir.net 联系他。

Contents 目　录

推荐序
译者序
前言
致谢
关于作者
关于审稿人

第 1 章　准备数据 ····· 1

1.1　导论 ····· 1
1.2　使用 Python 读写 CSV/TSV 文件 ····· 2
1.3　使用 Python 读写 JSON 文件 ····· 6
1.4　使用 Python 读写 Excel 文件 ····· 7
1.5　使用 Python 读写 XML 文件 ····· 10
1.6　使用 pandas 检索 HTML 页面 ····· 13
1.7　存储并检索关系数据库 ····· 15
1.8　存储并检索 MongoDB ····· 18
1.9　使用 OpenRefine 打开并转换数据 ····· 20
1.10　使用 OpenRefine 探索数据 ····· 23
1.11　排重 ····· 25
1.12　使用正则表达式与 GREL 清理数据 ····· 27
1.13　插补缺失值 ····· 28
1.14　将特征规范化、标准化 ····· 29
1.15　分级数据 ····· 30
1.16　编码分类变量 ····· 32

第 2 章　探索数据 ····· 34

2.1　导论 ····· 34
2.2　生成描述性的统计数据 ····· 34
2.3　探索特征之间的相关性 ····· 37
2.4　可视化特征之间的相互作用 ····· 38
2.5　生成直方图 ····· 43
2.6　创建多变量的图表 ····· 46
2.7　数据取样 ····· 49
2.8　将数据集拆分成训练集、交叉验证集和测试集 ····· 51

第 3 章　分类技巧 ····· 53

3.1　导论 ····· 53
3.2　测试并比较模型 ····· 53
3.3　朴素贝叶斯分类器 ····· 56
3.4　将逻辑回归作为通用分类器使用 ····· 58
3.5　将支持向量机用作分类引擎 ····· 61
3.6　使用决策树进行分类 ····· 65

3.7 使用随机森林预测订阅者……69
3.8 使用神经网络对呼叫进行分类……72

第 4 章 聚类技巧……79

4.1 导论……79
4.2 评估聚类方法的表现……79
4.3 用 *k* 均值算法聚类数据……82
4.4 为 *k* 均值算法找到最优的聚类数……84
4.5 使用 mean shift 聚类模型发现聚类…90
4.6 使用 *c* 均值构建模糊聚类模型……91
4.7 使用层次模型聚类数据……93
4.8 使用 DBSCAN 和 BIRCH 算法发现潜在的订阅者……96

第 5 章 降维……99

5.1 导论……99
5.2 创建三维散点图，显示主成分……99
5.3 使用核 PCA 降维……102
5.4 用主成分分析找到关键因素……105
5.5 使用随机 PCA 在数据中寻找主成分……109
5.6 使用线性判别分析提取有用的维度…114
5.7 用 kNN 分类模型给电话分类时使用多种降维技巧……117

第 6 章 回归模型……122

6.1 导论……122
6.2 识别并解决数据中的多重共线性…124
6.3 构建线性回归模型……128
6.4 使用 OLS 预测生产的电量……134
6.5 使用 CART 估算发电厂生产的电量……138
6.6 将 kNN 模型用于回归问题……141
6.7 将随机森林模型用于回归分析……143
6.8 使用 SVM 预测发电厂生产的电量…145
6.9 训练神经网络，预测发电厂生产的电量……151

第 7 章 时间序列技术……154

7.1 导论……154
7.2 在 Python 中如何处理日期对象……155
7.3 理解时间序列数据……159
7.4 平滑并转换观测值……163
7.5 过滤时间序列数据……166
7.6 移除趋势和季节性……169
7.7 使用 ARMA 和 ARIMA 模型预测未来……173

第 8 章 图……181

8.1 导论……181
8.2 使用 NetworkX 在 Python 中处理图对象……182
8.3 使用 Gephi 将图可视化……190
8.4 识别信用卡信息被盗的用户……200
8.5 识别谁盗窃了信用卡……204

第 9 章 自然语言处理……207

9.1 导论……207
9.2 从网络读入原始文本……208
9.3 标记化和标准化……212
9.4 识别词类，处理 *n*-gram，识别命名实体……218
9.5 识别文章主题……224
9.6 识别句子结构……226
9.7 根据评论给影片归类……229

第 10 章　离散选择模型……233

10.1　导论……233
10.2　准备数据集以估算离散选择模型…235
10.3　估算知名的多项 Logit 模型……239
10.4　测试来自无关选项的独立性冲突…244
10.5　用巢式 Logit 模型处理 IIA 冲突…249
10.6　用混合 Logit 模型处理复杂的替代模式……251

第 11 章　模拟……254

11.1　导论……254
11.2　使用 SimPy 模拟加油站的加油过程……255
11.3　模拟电动车耗尽电量的场景……264
11.4　判断羊群面对群狼时是否有团灭的风险……269

需要全本请访问 www.ertongbook.com