详解数据清洗、整理、探索、可视化、
建模和评估等流程

Numpy、Pandas、MatPlotLib、
Sklearn、Seaborn、Statsmodels和SciPy模块应用

十大常用数据挖掘算法及案例实战

从零开始学 Python 数据分析与挖掘

刘顺祥 著

数据源和源代码下载

清华大学出版社

从零开始学 Python

数据分析与挖掘

刘顺祥 著

清華大學出版社
北京

内 容 简 介

本书以 Python 3 版本作为数据分析与挖掘实战的应用工具，从 Pyhton 的基础语法开始，陆续介绍有关数值计算的 Numpy、数据处理的 Pandas、数据可视化的 Matplotlib 和数据挖掘的 Sklearn 等内容。全书共涵盖 15 种可视化图形以及 10 个常用的数据挖掘算法和实战项目，通过本书的学习，读者可以掌握数据分析与挖掘的理论知识和实战技能。

本书适于统计学、数学、经济学、金融学、管理学以及相关理工科专业的本科生、研究生使用，也能够提高从事数据咨询、研究或分析等人士的专业水平和技能。

图书在版编目（CIP）数据

从零开始学 Python 数据分析与挖掘 / 刘顺祥著. —北京：清华大学出版社，2018
ISBN 978-7-302-50987-5

Ⅰ. ①从… Ⅱ. ①刘… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字（2018）第 192204 号

责任编辑： 王金柱
封面设计： 王　翔
责任校对： 闫秀华
责任印制： 刘海龙

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 装 者： 三河市君旺印务有限公司
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 23.5　　**字　　数：** 602 千字
版　　次： 2018 年 10 月第 1 版　　**印　　次：** 2018 年 10 月第 1 次印刷
定　　价： 79.00 元

产品编号：079655-01

前　言

为什么写这本书

随着大数据时代的演进，越来越多的企业在搜集数据的同时，也开始关注并重视数据分析与挖掘的价值，因为他们正尝到这项技术所带来的甜头。例如，通过该技术可以帮助企业很好地认识其用户的画像特征，为用户提供个性化的优质服务，进而使用户的忠诚度不断提升；通过该技术提前识别出不利于企业健康发展的“毒瘤”用户（如黄牛群体、欺诈群体等），进而降低企业不必要的损失；通过该技术可以为企业实现某些核心指标的判断和预测，进而为企业高层的决策提供参考依据等。企业对数据分析与挖掘技术的重视就意味着对人才的重视，这就要求希望或正在从事数据相关岗位的人员具备该技术的理论知识和实战能力。

Python 作为大数据相关岗位的应用利器，具有开源、简洁易读、快速上手、多场景应用以及完善的生态和服务体系等优点，使其在数据分析与挖掘领域中的地位显得尤为突出。基于 Python 可以对各种常见的脏数据完成清洗、绘制各式各样的统计图形，并实现各种有监督、无监督和半监督的机器学习算法的落地，在数据面前做到游刃有余，所以说 Python 是数据分析与挖掘工作的不二之选。根据多家招聘网站的统计，几乎所有的数据分析或挖掘岗位都要求应聘者掌握至少一种编程语言，其中就包括 Python。

纵观国内的图书市场，关于 Python 的书籍还是非常多的，它们主要偏向于工具本身的用法，如关于 Python 的语法、参数、异常处理、调用以及开发类实例等。但是基于 Python 的数据分析与挖掘书籍并不是特别多，关于这方面技术的书籍更多的是基于 R 语言等工具。本书将通过具体的实例讲解数据的处理和可视化技术，同时也结合数据挖掘的理论知识和项目案例讲解 10 种常用的挖掘算法。

2015 年 9 月，笔者申请了微信公众号，取名为“数据分析 1480”，目前已经陆续更新了近 200 篇文章。一方面是为了将自己所学、所知记录下来，作为自己的知识沉淀；另一方面是希望尽自己的微薄之力，将记录下来的内容分享给更多热爱或从事数据分析与挖掘事业的朋友。但是公众号的内容并没有形成系统的知识框架，在王金柱老师的鼓励和支持下才开始了本书的写作，希望读者能够从中获得所需的知识点。

本书的内容

本书一共分为三大部分，系统地介绍数据分析与挖掘过程中所涉及的数据清洗与整理、数据可视化以及数据挖掘的落地。

第一部分（第 1~3 章）介绍有关数据分析与挖掘的概述以及 Python 的基础知识，并通过一个

有趣的案例引入本书内容的学习。本部分内容可以为初学 Python 的朋友奠定基础，进而为后续章节的学习做准备。

第二部分（第 4~6 章）涉及 numpy 模块的数值计算、Pandas 模块的数据清洗与整理以及 Matplotlib 模块的可视化技术。本部分内容可以为数据预处理过程中的清洗、整理以及探索性分析环节提供技术支撑。

第三部分（第 7~16 章）一共包含 10 种数据挖掘算法的应用，如线性回归、决策树、支持向量机、GBDT 等，使用通俗易懂的术语介绍每一个挖掘算法的理论知识，并借助于具体的数据项目完成算法的实战。本部分内容可以提高热爱或从事数据分析相关岗位朋友的水平和技能，也可以作为数据挖掘算法落地的模板。

本书每一章都有对应的数据源和完整代码，代码均包含具体的中文注释，读者可以在笔者的 github 网站 https://github.com/SnakeLiu/Python-Data-Aanalysis-and-Miner 下载，或者在百度网盘 https://pan.baidu.com/s/18REQ_J057i7KL7ivBCX-cw （密码：xt4i）下载。

勘误和支持

由于笔者水平有限，书中难免会出现不当的地方，欢迎专家和读者朋友给予批评和指正。可以通过下方的途径联系并反馈建议：

- 即时通信：添加个人微信（lsx19890717）或者 QQ（1029776077），及时反馈问题。
- 公众号：添加个人微信公众号“数据分析 1480”，可参与后台互动。
- 电子邮箱：发送邮件至 lsxxx2011@163.com。

致谢

特别感谢清华大学出版社的王金柱老师，感谢他的热情相邀和宝贵建议，是他促成了本书的完成，同时他专业而高效的审阅也使本书增色不少。感谢参与本书封面设计的王翔老师、责任校对闫秀华老师，以及其他背后默默支持的出版工作者，在他们的努力和付出下，保证了本书的顺利出版。

最后，感谢我的家人和朋友，尤其是我的妻子许欣女士，是她在我写书期间把家里的一切整理得有条不紊，对我的照顾更是无微不至，才使我能够聚精会神地完成本书全部内容的撰写。

刘顺祥（Sim Liu）
2018 年 8 月于上海

目　录

第1章

数据分析与挖掘概述

马云曾说“中国正迎来从IT时代到DT时代的变革”，DT就是大数据时代。随着移动互联网的发展，人们越来越感受到技术所带来的便捷，同时企业也将搜集到越来越多与用户相关的数据，包括用户的基本信息、交易记录、个人喜好、行为特征等。这些数据就相当于隐藏在地球深处的宝贵资源，企业都想从数据红利中分得一杯羹，进而推进企业重视并善加利用数据分析与挖掘相关的技术。

本章将以概述的形式介绍数据分析和挖掘相关的内容，通过本章的学习，你将了解如下几方面的知识点：

- 数据分析与挖掘的认识；
- 数据分析与挖掘的几个应用案例；
- 数据分析与挖掘的几方面区别；
- 数据分析与挖掘的具体操作流程；
- 数据分析与挖掘的常用工具。

1.1 什么是数据分析和挖掘

随着数据时代的蓬勃发展，越来越多的企事业单位开始认识到数据的重要性，并通过各种手段进行数据的搜集。例如，使用问卷调查法获取用户对产品的评价或改善意见；通过每一次的实验获得产品性能的改良状况；基于各种设备记录空气质量状况、人体健康状态、机器运行寿命等；通过网页或APP记录用户的每一次登录、浏览、交易、评论等操作；基于数据接口、网络爬虫等手段获取万维网中的公开数据；甚至是企业间的合作实现多方数据的共享。企事业单位花费人力、物力获取各种数据的主要目的就是通过数据分析和挖掘手段实现数据的变现，否则囤积的数据就是资源的浪费。

数据分析和挖掘都是基于搜集来的数据，应用数学、统计、计算机等技术抽取出数据中的有

用信息，进而为决策提供依据和指导方向。例如，应用漏斗分析法挖掘出用户体验过程中的不足之处，从而进一步改善产品的用户流程；利用 AB 测试法检验网页布局的变动对交易转化率的影响，从而确定这种变动是否有利；基于 RFM 模型实现用户的价值分析，进而针对不同价值等级的用户采用各自的营销方案，实现精准触达；运用预测分析法对历史的交通数据进行建模，预测城市各路线的车流量，进而改善交通的拥堵状况；采用分类手段，对患者的体检指标进行挖掘，判断其所属的病情状况；利用聚类分析法对交易的商品进行归类，可以实现商品的捆绑销售、推荐销售等营销手段。应用数据分析和挖掘方法，让数据产生价值的案例还有很多，这里就不一一枚举了，所以只有很好地利用数据，它才能产生价值，毫不夸张地说，大部分功劳都要归功于数据分析和挖掘。

1.2 数据分析与挖掘的应用领域

也许读者也曾自我发问——学会了数据分析和挖掘技术，可以从事哪些行业的相关工作呢？在笔者看来，有数据的地方就有用武之地。现在的数据充斥在各个领域，如庞大的互联网行业，包含各种电商平台、游戏平台、社交平台、中介类平台等；金融行业，包含银行、P2P、互联网金融等；影响国计民生的教育、医疗行业；各类乙方数据服务行业；传统行业，如房地产、餐饮、美容等。这些行业都需要借助数据分析和挖掘技术来指导下一步的决策方向，以下仅举 3 个行业应用的例子，进一步说明数据分析和挖掘的用武之地。

1.2.1 电商领域——发现破坏规则的“害群之马”

移动互联网时代下，电商平台之间的竞争都特别激烈，为了获得更多的新用户，往往会针对新用户发放一些诱人的福利，如红包券、满减券、折扣券、限时抢购优惠券等，当用户产生交易时，就能够使用这些券减免一部分交易金额。电商平台通过类似的营销手段一方面可以促进新用户的获取，增添新鲜血液；另一方面也可以刺激商城的交易，增加用户的活跃度，可谓各取所需的双赢效果。

然而，某些心念不正的用户为了从中牟取利益，破坏大环境下的游戏规则。某电商数据分析人员在一次促销活动的复盘过程中发现交易记录存在异常，于是就对这批异常交易作更深层次的分析和挖掘。最终发现这批异常交易都有两个共同特点，那就是一张银行卡对应数百个甚至上千个用户 id，同时，这些 id 自始至终就发生一笔交易。暗示了什么问题？这说明用户很可能通过廉价的方式获得多个手机号，利用这些手机号去注册 APP 成为享受福利的多个新用户，然后利用低价优势买入这些商品，最后再以更高的价格卖出这些商品，这种用户我们一般称为“黄牛”。

这些“害群之马”的行为至少给电商平台造成两方面的影响，一是导致真正想买商品的新用户买不到，因为有限的福利或商品都被这些用户抢走了；二是虚增了很多“薅羊毛”的假用户，因为他们很可能利用完新用户的福利资格后就不会再交易了。如果没有数据分析与挖掘技术在互联网行业的应用，就很难发现这些“害群之马”，企业针对“害群之马”对游戏规则做了相应的调整，从而减少了不必要的损失，同时也挽回了真实用户的利益。

1.2.2　交通出行领域——为打车平台进行私人订制

打车工具的出现，改变了人们的出行习惯，也改善了乘车的便捷性，以前都是通过路边招手才能搭乘出租车，现在坐在家里就可以完成一对一的打车服务。起初滴滴、快滴、优步、易到等打车平台，为了抢占市场份额，不惜花费巨资补贴给司机端和乘客端，在一定程度上获得了用户的青睐，甚至导致用户在短途出行中都依赖上了这些打车工具。然而随着时间的推移，打车市场的格局基本定型，企业为了自身的利益和长远的发展，不再进行这种粗放式的“烧钱”运营手段。

当司机端和乘客端不再享受以前的福利待遇时，在一定程度上影响了乘客端的乘车频率和司机端的接单积极性。为了弥补这方面的影响，某打车平台利用用户的历史交易数据，为司机端和乘客端的定价进行私人订制。

例如，针对乘客端，通过各种广告渠道将折扣券送到用户手中，一方面可以唤醒部分沉默用户（此时的折扣力度会相对比较高），让他们再次回到应用中产生交易，另一方面继续刺激活跃用户的使用频率（此时的折扣力度会相对比较低），进而提高用户的忠诚度。针对司机端，根据司机在平台的历史数据，将其接单习惯、路线熟悉度、路线拥堵状况、距离乘客远近、天气变化、乘客乘坐距离等信息输入到逻辑模型中，可以预测出司机接单的概率大小。这里的概率在一定程度上可以理解为司机接单的意愿，概率越高，说明司机接单的意愿越强，否则意愿就越弱。当模型发现司机接单的意愿比较低时，就会发放较高的补贴给司机端，否则司机就会获得较少的补贴甚至没有补贴。如果不将数据分析与挖掘手段应用于大数据的交通领域，就无法刺激司机端和乘客端的更多交易，同时，也会浪费更多的资金，造成运营成本居高不下，影响企业的发展和股东的利益。

1.2.3　医疗健康领域——找到最佳医疗方案

众所周知，癌症的产生是由于体内某些细胞的 DNA 或 RNA 发生了病变，这种病变会导致癌细胞不断地繁殖，进而扩散至全身，最终形成可怕的肿瘤。早在 2003 年，乔布斯在一次身体检查时发现胰腺处有一块阴影，医生怀疑是一块肿瘤，建议乔布斯马上进行手术，但乔布斯选择了药物治疗。遗憾的是，一年后，医生从乔布斯的身体检查中发现可怕的癌细胞已经扩散到了全身，认为乔布斯的生命即将走到人生的终点。

乐观的乔布斯认为还可以有治疗的希望，于是花费几十万美元，让专业的医疗团队将自己体内的 DNA 与历史肿瘤 DNA 样本进行比对，目的就是找到符合肿瘤病变的 DNA。这样，对于乔布斯体内的 DNA 来说就有了病变与正常的标签，然后基于这个标签构建分类算法。当正常 DNA 出现病变特征时，该算法就能够准确地找出即将病变的 DNA，从而指导医生及时地改变医疗方案和寻找有效的药物。最终，使得原本即将走到终点的生命，延续了八年时间，正是这短短的八年，让乔布斯一次次地创造了苹果的辉煌。如果没有数据分析与挖掘在医疗行业的应用，也许就没有现在的苹果。

1.3 数据分析与挖掘的区别

从广义的角度来说，数据分析的范畴会更大一些，涵盖了数据分析和数据挖掘两个部分。数据分析就是针对搜集来的数据运用基础探索、统计分析、深层挖掘等方法，发现数据中有用的信息和未知的规律与模式，进而为下一步的业务决策提供理论与实践依据。所以广义的数据分析就包含了数据挖掘的部分，正如读者在各招聘网站中所看见的，对于数据分析师的任职资格中常常需要应聘者熟练使用数据挖掘技术解决工作中的问题。从狭义的角度来说，两者存在一些不同之处，主要体现在两者的定义说明、侧重点、技能要求和最终的输出形式。接下来阐述这几个方面的差异。

- 从定义说明出发：数据分析采用适当的统计学方法，对搜集来的数据进行描述性分析和探索性分析，并从描述和探索的结果中发现数据背后存在的价值信息，用以评估现状和修正当前的不足；数据挖掘则广泛交叉数据库知识、统计学、机器学习、人工智能等方法，对搜集来的数据进行“采矿”，发现其中未知的规律和有用的知识，进一步应用于数据化运营，让数据产生更大的价值。
- 从侧重点出发：数据分析更侧重于实际的业务知识，如果将数据和业务分开，往往会导致数据的输出不是业务所需，业务的需求无法通过数据体现，故数据分析需要两者的紧密结合，实现功效的最大化；数据挖掘更侧重于技术的实现，对业务知识的熟练度并没有很高的要求，如何从海量的数据中发现未知的模式和规律，是数据挖掘的目的所在，只有技术过硬，才能实现挖掘项目的落地。
- 从掌握的技能出发：数据分析一般要求具备基本的统计学知识、数据库操作技能、Excel 报表开发和常用可视化图表展现的能力，就可以解决工作中的分析任务；数据挖掘对数学功底和编程能力有较高的要求，数学功底是数据挖掘、机器学习、人工智能等方面的基础，没有好的数学功底，在数据挖掘领域是走不远的，编程能力是从数据中发现未知模式和规律途径，没有编程技能，就无法实现算法的落地。
- 从输出的结果出发：数据分析更多的是统计描述结果的呈现，如平均水平、总体趋势、差异对比、数据转化等，这些结果都必须结合业务知识进行解读，否则一组数据是没有任何实际意义的；数据挖掘更多的是模型或规则的输出，通过模型或规则可对未知标签的数据进行预测，如预测交通的畅通度（预测模型）、判别用户是否响应某种营销活动（分类算法）；通过模型或规则实现智能的商业决策，如推荐用户可能购买的商品（推荐算法）、划分产品所属的群类（聚类算法）等。

为了读者更容易理解和区分两者之间的差异，这里将上面描述的四方面内容做一个简短的对比和总结，如表 1-1 所示。

表 1-1 数据分析与挖掘对比

差异角度	数据分析	数据挖掘
定义	描述和探索性分析，评估现状和修正不足	技术性的“采矿”过程，发现未知的模式和规律
侧重点	实际的业务知识	挖掘技术的落地，完成“采矿”过程

（续表）

差异角度	数据分析	数据挖掘
技能	统计学、数据库、Excel、可视化等	过硬的数学功底和编程技术
结果	需结合业务知识解读统计结果	模型或规则

1.4　数据挖掘的流程

本书将安排 10 个章节的内容来讲解具体的数据挖掘算法和应用案例，故需要对数据挖掘的具体流程做一个详细的说明。这里的流程可以理解为数据挖掘过程中的规范，只有熟悉了这些具体的规范，才可以在数据挖掘过程中做到游刃有余。首先通过图 1-1 中的金字塔了解数据挖掘中具体的操作步骤。

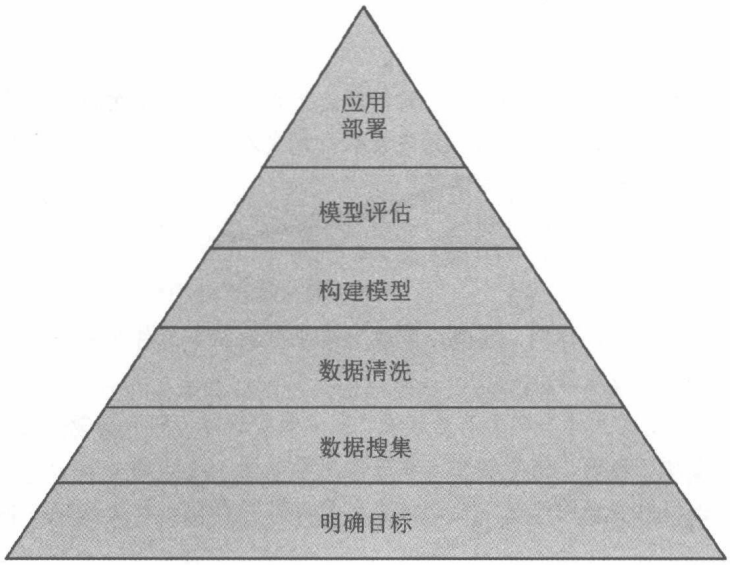

图 1-1　数据挖掘步骤

1.4.1　明确目标

前面讲了几个有关数据分析和数据挖掘在电商行业、交通领域和医疗健康方面的案例，体现了数据分析与挖掘的重要性。你可能非常期待数据分析与挖掘在工作中的应用，先别急，在实施数据挖掘之前必须明确自己需要解决的问题是什么，然后才可以有的放矢。

这里通过三个实际的案例来加以说明数据挖掘流程中的第一步，即明确目标：

- 在餐饮行业，可能都会存在这方面的痛点，即如何调整中餐或晚餐的当班人数，以及为下一餐准备多少食材比较合理。如果解决了这个问题，那么对于餐厅来说既可以降低人工成本，又可以避免食材的浪费。
- 当前互联网经济下的消费信贷和现金信贷都非常流行，对于企业来说可以达到“以钱赚钱”

的功效，对于用户来说短期内可以在一定程度上减轻经济压力，从而实现两端的双赢。但是企业会面临给什么样的用户放发信贷的选择，如果选择正确了，可以赚取用户的利息，如果选择错误了，就得赔上本金。所以风险控制（简称“风控”）尤其重要，如果风控做得好，就能够降低损失，否则就会导致大批“坏账”甚至是面临倒闭。

- 对于任何一个企业来说，用户的价值高低决定了企业可从用户身上获得的利润空间。用户越忠诚、价值越高，企业从用户身上获取的利润就越多，反之利润就越少。所以摆在企业眼前的重大问题就是如何提升用户的生命价值。

1.4.2 数据搜集

当读者明确企业面临的痛点或工作中需要处理的问题后，下一步就得规划哪些数据可能会影响到这些问题的答案，这一步就称为数据的搜集过程。数据搜集过程显得尤为重要，其决定了后续工作进展的顺利程度。接下来继续第一步中的例子，说明这三个案例中都需要搜集哪些相关的数据。

1. 餐饮相关

- 食材数据：食材名称、食材品类、采购时间、采购数量、采购金额、当天剩余量等。
- 经营数据：经营时间、预定时间、预定台数、预定人数、上座台数、上座人数、上菜名称、上菜价格、上菜数量、特价菜信息等。
- 其他数据：天气状况、交通便捷性、竞争对手动向、是否为节假日、用户口碑等。

2. 金融授信

- 用户基本数据：姓名、性别、年龄、受教育水平、职业、工作年限、收入状况、婚姻状态、借贷情况、房产、汽车等。
- 刷卡数据：是否有信用卡、刷卡消费频次、刷卡缴费规律、刷卡金额、是否分期、是否逾期、逾期天数、未偿还金额、信用额度、额度使用率等。
- 其他数据：信用报告查询记录、电话核查记录、银行存款、社交人脉、其他 APP 数据等。

3. 影响用户价值高低

- 会员数据：性别、年龄、教育水平、会员等级、会员积分、收入状况等。
- 交易数据：用户浏览记录、交易商品、交易数量、交易频次、交易金额、客单价、最后交易时间、偏好、下单与结账时差等。
- 促销数据：用户活动参与度、优惠券领取率、优惠券使用率、购买数量、购买金额等。
- 客服数据：实时沟通渠道数量、用户沟通次数、用户疑问响应速度、疑问解答率、客户服务满意度等。

1.4.3 数据清洗

为解决企业痛点或面临的问题，需要搜集相关的数据。即使数据搜集上来，也必须保证数据“干净”，因为数据质量的高低将影响最终结果的准确性。通常都有哪些“不干净”的数据会影响后面的建模呢？针对这些数据都有哪些解决方案呢？这里不妨做一个简要的概述。