大数据应用与技术丛书

R数据科学实战

(第 2 版)

[美] 尼娜·祖梅尔(Nina Zumel)
约翰·蒙特(John Mount) 著

张骏温 许向东 张博远 译

清华大学出版社

北 京

Nina Zumel, John Mount
Practical Data Science with R, Second Edition
EISBN: 978-1-61729-587-4

图书在版编目(CIP)数据

R 数据科学实战 / (美)尼娜 • 祖梅尔(Nina Zumel)，(美)约翰 • 蒙特 (John Mount) 著；张骏温，许向东，张博远译. —2 版. —北京：清华大学出版社，2022.1
(大数据应用与技术丛书)
书名原文：Practical Data Science with R, Second Edition
ISBN 978-7-302-59544-1

I. ①R… II. ①尼… ②约… ③张… ④许… ⑤张… III. ①程序语言－程序设计 IV. ①TP312

中国版本图书馆 CIP 数据核字(2021)第 231297 号

责任编辑：王 军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：朱雨萌

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：三河市金元印装有限公司
经 销：全国新华书店
开 本：170mm×240mm 印 张：36.75 字 数：740 千字
版 次：2022 年 1 月第 1 版 印 次：2022 年 1 月第 1 次印刷
定 价：139.00 元

产品编号：086798-01

本书(第1版)的赞誉

该书内容清晰简洁，针对快速变化的商业、统计学和机器学习之间的相互关系，提供了第一手丰富的实践指南。

—Dwight Barry,
Group Health Cooperative

我非常希望本书是我学习数据科学的起点。作者为我们绘制了一个全面的、结构合理的数据科学的知识体系，并且提供了快速掌握其原理的方法。本书也是集统计学、数据分析和计算机科学于一体的技术书籍。

—Justin Fister, AI 研究员,
PaperRater.com

这是我看到过的关于“数据科学与 R 语言”最全面的内容描述。

—Romit Singhai, SGI 公司

它涵盖了从数据探索到数据建模再到结果交付的一个端到端的流程。

—Nezih Yigitbasi, Intel 公司

即使对雄心勃勃或是经验丰富的数据科学家，本书也大有裨益。

—Fred Rahmanian,
Siemens Healthcare

采用了现实世界的案例对数据分析进行实验，强烈推荐此书。

—Kostas Passadis 博士, IPTO

在阅读本书的过程中，你会感觉到它是由知识渊博、经验丰富的专业人士编写的，并且他们对你毫无隐瞒。

—Amazon 读者

序　言

《R 数据科学实战》(第 2 版)是一本针对数据科学的实践指南，重点介绍了使用 R 语言和统计程序包处理结构化或表格数据的相关技术，也着重介绍了机器学习的技术。但它的独特之处在于专门讨论了数据科学家在项目中的角色、所管理的交付结果，甚至设计演示文稿等主题。本书不仅研究了如何编写模型，还讨论了如何与不同的团队协作，如何将业务目标转化为度量值，以及如何组织工作和编写报告等。如果你想学习如何使用 R 语言来从事数据科学家的工作，那么建议你阅读本书。

我们认识 Nina Zumel 和 John Mount 已经很多年了，曾经邀请他们到奇点大学(Singularity University)和我们一起教书，他们是我们所知道的最优秀的两位数据科学家。我们定期推荐他们关于交叉验证和影响编码(也称为目标编码)的原创性研究。实际上，他们在本书第 8 章中讲授了影响编码的相关理论，并通过自己的 R 程序包 vtreat 实现了其应用。

在《R 数据科学实战》(第 2 版)这本书中，作者用了一些篇幅描述了什么是数据科学、数据科学家是如何解决问题的，以及对他们工作的描述。其中，包括对经典监督学习方法(如线性回归和逻辑回归)的详细描述。我们喜欢本书的调研式风格，以及使用的大量的竞赛获奖方法和程序包的示例(如随机森林和 xgboost)。本书涵盖了非常有用的、可共享的经验和实践建议。我们注意到，在本书中甚至包括了我们自己使用过的一些技巧，例如使用随机森林变量重要性进行初始变量的筛选。

总体而言，这是一本很棒的图书，我们强烈推荐。

—Jeremy Howard 和 Rachel Thomas

前　言

本书是我们在自学时所希望拥有的书，它所汇集的主题和技能被称为数据科学。本书也是我们想分发给客户和同行的书。它的目的是解释统计学、计算机科学和机器学习等学科中对数据科学至关重要的内容。

数据科学利用了来自经验科学、统计学、报表技术、分析技术、可视化技术、商业智能、专家系统、机器学习、数据库、数据仓库、数据挖掘和大数据技术的各种工具。正是因为我们有太多的工具，所以需要一个涵盖所有工具的指导原则。数据科学本身与这些工具和技术的区别就在于数据科学的中心目标是将有效的决策模型部署到生产环境中。

我们的目标是从务实的、面向实践的角度来展示数据科学。我们通过聚焦在完全成功的真实数据上的示例来实现这一目标，本书展示了超过 10 个重要的数据集。我们认为这种方法能举例说明我们真正想要达到的教学目标，并能演示实际项目中所需要的各种准备步骤。

在本书中，我们讨论了实用的统计学和机器学习的概念，包括具体的代码示例，并探索了与非专业人员的合作和沟通方式。如果你觉得这些话题中没有新颖的主题，那么我们希望本书内容能为你最近没有想到的其他一两个话题提供一些启示。

致　谢

感谢我们的同事和其他阅读与评论本书各章节草稿的人。特别要感谢如下评审员：Charles C. Earl、Christopher Kardell、David Meza、Domingo Salazar、Doug Sparling、James Black、John MacKintosh、Owen Morris、Pascal Barbedo、Robert Samohyl 和 Taylor Dolezal。他们的意见、疑问和修订极大地改进了本书的质量。我们要特别感谢为本书第 1 版工作的策划编辑 Dustin Archibald 和 Cynthia Kane，感谢他们提供了想法和支持。同时，还要感谢 Nichole Beard、Benjamin Berg、Rachael Herbert、Katie Tennant、Lori Weidert、Cheryl Weisman 以及所有为使本书成为一本优秀的著作而努力工作的编辑们。

此外，还要感谢我们的同事 David Steier、加州大学伯克利分校信息科学学院的 Doug Tygar 教授、威斯康星大学白水分校生物科学和计算机科学系的 Robert K. Kuzoff 教授，以及所有用这本书作为教学范本的院系同事和教师。感谢 Jim Porzak、Joseph Rickert 和 Alan Miller 经常邀请我们在 R 用户组发表与本书相关的主题演讲。我们要特别感谢 Jim Porzak 为第 1 版撰写前言，并积极宣传我们的书。当我们疲惫不堪、灰心丧气，不知道为什么要把自己投入这项任务中时，他的热情帮助使我们意识到所分享的内容和所采用的方式是有大众需求的。如果没有他的鼓励，完成本书会变得困难许多。同时，我们也要感谢 Jeremy Howard 和 Rachel Thomas 撰写了新版的序言，并邀请我们演讲，为我们提供了大力支持。

关于本书

本书是关于数据科学的：一个使用统计学、机器学习以及计算机科学的结果来创建预测模型的领域。由于数据科学涉猎广泛，因此在本书中讨论并概述一些我们采用的方法是很重要的。

数据科学概述

统计学家 William S. Cleveland 将数据科学定义为一个比统计学本身更大的跨学科领域。我们将数据科学定义为可以将假设和数据转化为可操作的预测的过程。典型的预测分析目标包括预测谁将赢得选举、哪些产品放在一起可以畅销、哪些贷款将违约，以及哪些广告将被点击。数据科学家负责获取和管理数据，选择建模技术，编写代码并验证结果。

因为数据科学涉及很多学科，所以它常常是一种“第二种选择”。我们遇到的许多最好的数据科学家都是从程序员、统计学家、商业智能分析师或科学家开始的。通过在他们已掌握的技能上增加一些技术，他们成为了优秀的数据科学家。正是这一观察推动了本书的编写：我们通过在真实数据上实际操作所有常见项目的具体步骤来介绍数据科学家所需要的实用技能。对读者而言，有些步骤你会比我们更了解，有些你会很快学会，而有些你可能需要进一步研究。

数据科学的许多理论基础都来自统计学。但我们所熟知的数据科学因受到技术和软件工程方法论的巨大影响，很大程度上是在计算机科学和信息技术驱动的群体中发展起来的。我们可以列举出数据科学的一些工程学特性：

- 亚马逊的产品推荐系统
- 谷歌的广告评估系统
- 领英的联系人推荐系统
- 推特的热门话题
- 沃尔玛的消费者需求预测系统

以上这些系统有很多共同的特点：

- 所有的这些系统都建立在海量的数据集上。这并不是说它们都属于大数据的领域。但如果它们仅仅使用小的数据集，那么相信没有一个系统会成功。为了管理数据，这些系统需要来自计算机科学的理论：数据库理论、并行编程理论、流数据技术和数据仓库系统。
- 这些系统大部分都是在线或者实时的。不同于制作一份报告或分析，数据科学团队会部署一个决策程序或评分程序来直接做决策或直接向大量终端用户展示结果。生产部署是解决问题的最后机会，因为数据科学家不会随时都来解释这些缺陷。
- 所有这些系统都可能会出现某些无法预知的错误。
- 这些系统中没有一个与原因有关。它们的成功在于找到了有用的关联性，并且它们不被用来区分正确的因果关系。

本书所教授的内容是建立这些系统所需要的原理和工具，涉及成功交付此类项目的常见任务、步骤和工具。我们的重点是全过程项目管理、与他人合作，并向非专业人士展示结果。

本书的路线图

本书包括以下内容：

- 数据科学管理的全过程。数据科学家必须具备衡量和检测他们自身项目的能力。
- 许多数据科学项目中使用的最强大的统计和机器学习的技术。本书包含众多使用 R 编程语言执行实际数据科学工作的实际操作。
- 为所有利益相关者(管理者、用户、部署团队等)准备演示文稿。要知道，你必须能用具体的术语向不同的受众解释你的工作，而不是坚持使用某一特定领域的技术词汇。当然，也不能随便把数据科学项目的结果扔到一边，置之不理。

我们把书中的主题按照能增进对数据科学理解的顺序进行了排列。所有材料的组织如下：

第 I 部分描述了数据科学过程的基本目标和技术，主要强调协作和数据。第 1 章讨论了数据科学家是如何工作的。第 2 章展示了如何将数据加载到 R 中，以及如何开始使用 R。

第 3 章讲授了如何在数据中找寻所需的信息，以及描述和理解数据的几个重要步骤。数据必须为分析做好准备，并且数据问题需要被修正。第 4 章展示了如何修正第

3 章发现的问题。

第 5 章介绍了数据准备的一个步骤：基本数据整理。数据并不总是以最适合分析的形式或“形态”提供给数据科学家。R 提供了许多工具来管理和重构数据以获得正确的结构。本章涵盖了这些内容。

第 II 部分从描述和准备数据转向建立有效的预测模型。第 6 章提供了从业务需求到技术评估和建模技术的映射关系。它涵盖了用于评估模型性能的标准指标和程序，以及一项专用技术——LIME。LIME 用来解释由一个模型做出的具体预测。

第 7 章介绍了基本的线性模型：线性回归、逻辑回归和正则线性模型。线性模型是许多分析任务要用到的重要工具，并且对于识别关键变量和深入了解问题的内部结构有极大的帮助。深入了解这些模型对数据科学家来说极其有价值。

第 8 章暂时脱离了建模任务，涵盖了更高级的数据处理的内容：如何为建模步骤规整杂乱的真实世界数据。因为要理解这些数据处理方法是如何工作的，需要对线性模型和模型评估指标有一定的理解，所以这个话题被放到第 II 部分。

第 9 章介绍了无监督模型的方法：一种不使用带标签的训练数据的建模方法。第 10 章涵盖了提高预测性能、修复特定建模问题的更高级的建模方法。涉及的主题包括基于决策树的集成方法、广义相加模型和支持向量机。

第III部分从建模回到了流程，展示了如何交付结果。第 11 章演示了如何管理、记录和部署自己的模型。第 12 章介绍如何为不同的听众创建有效的演示幻灯片。

附录包括了有关 R、统计学和其他可用工具的技术细节。附录 A 展示了如何安装 R，如何启动工作，以及如何使用其他工具(如 SQL)。附录 B 是对一些关键统计学概念的复习。

这些内容是按照目标和任务来组织的，用到的工具会一并介绍。每章的主题都涉及一个有相关数据集的项目。在学习本书的过程中，你将完成许多实质性的项目。本书提到的所有数据集都存储在本书的 GitHub 资料库中：https://github.com/WinVector/PDSwR2。你可以将整个资料库作为单个 zip 文件(GitHub 的服务之一)下载，或者将资料库复制到你自己的机器上，或者根据需要复制单个文件。

本书的读者对象

为了使用书中示例，首先需要对 R 和统计学有一些了解。我们建议你准备一些优秀的入门书籍。在开始阅读本书时不需要精通 R，但需要对它有所了解。如果你想从 R 学起，我们推荐你阅读 Jonathan Carroll(Manning, 20108)的 *Beyond Spreadsheets with R* 或者 Robert Kabacoff 的 *R in Action*(现在已经发行了第 2 版：http://www.manning.com/kabacoff2/)，

以及本书的相关网站，*Quick-R*(http://www.statmethods.net)。对于统计学，我们推荐你阅读 David Freedman、Robert Pisani 和 Roger Purves 合作编写的 *Statistics*(第四版) (W.W. Norton & Company, 2007)。

概括而言，我们希望你：

- 对工作示例感兴趣。通过研究这些示例，你至少学会一种方法来实现一个完整项目的所有步骤。你必须愿意尝试简单的脚本和编程以便获取本书的全部价值。对于我们提到的每一个例子，你应该尝试一些变化并预期有一些变化会失败(当你的变化不起作用时)而有一些变化会成功(当你的变化优于我们的示例分析时)。
- 对 R 统计系统比较熟悉，并愿意使用 R 语言编写简短的脚本和程序。除了 Kabacoff 外，我们在附录 C 中也列出了几本好书。我们会在 R 中处理具体的问题。你需要运行示例并参考附加文档以了解我们没有演示的命令的变化。
- 对基本的统计学概念比较熟悉，比如概率、平均值、标准差和显著性。我们将根据需要介绍这些概念，但在运行示例时，你要根据自己的情况阅读额外的参考资料以更好地理解示例。我们将定义一些术语并且参考一些主题资料和有用的博客，但我们也要求你自行在网上搜索某些主题。
- 准备一台计算机(macOS、Linux 或 Windows)来安装 R 和其他工具，以及用来下载工具和数据集的互联网连接。强烈建议你亲自运行那些示例，在不同方法上使用 R 的 help ()来查看相应的帮助文档，并阅读一些附加的参考文献。

本书中未包含的内容

- 本书不是一本 R 手册。我们使用 R 来具体说明数据科学项目的重要步骤。我们会讲授大量关于 R 的知识来帮助你完成书中的示例，但不熟悉 R 的话就需要参考附录 A，以及许多优秀的 R 书籍和可用的教学视频。
- 本书不是一系列的案例研究。我们强调方法理论和技术。书中给出的示例数据和代码只是为了确保我们给出的是具体且可用的建议。
- 本书不是一本大数据的书。我们认为最有意义的数据科学是发生在可管理的数据库或者文件规模(通常比内存大些，但仍然易于管理)。要产生那些能将度量的各种条件映射到相关结果的有价值的数据往往需要昂贵的成本，而这往往限制了数据规模。但对于某些报表生成、数据挖掘和自然语言处理，你需要进一步探索大数据的领域。

- 本书不是一本理论书籍。我们不聚焦在任何一种技术的细致理论描述上。数据科学讲究灵活性，有许多可用的好技术。如果某种技术看起来能解决你手头的问题，我们会深入探讨它。即使是在我们的文本中，相比于漂亮的排版公式，我们也更喜欢使用R代码符号，因为R代码可以直接使用。
- 本书不是一本机器学习的技术书。我们只关注已经在 R 中实现的方法。对于每种方法，我们会践行操作理论并展示其优点。我们通常不讨论如何实现它们(即使实现起来很容易)，因为优秀的R实现已经是可用的。

代码约定和下载

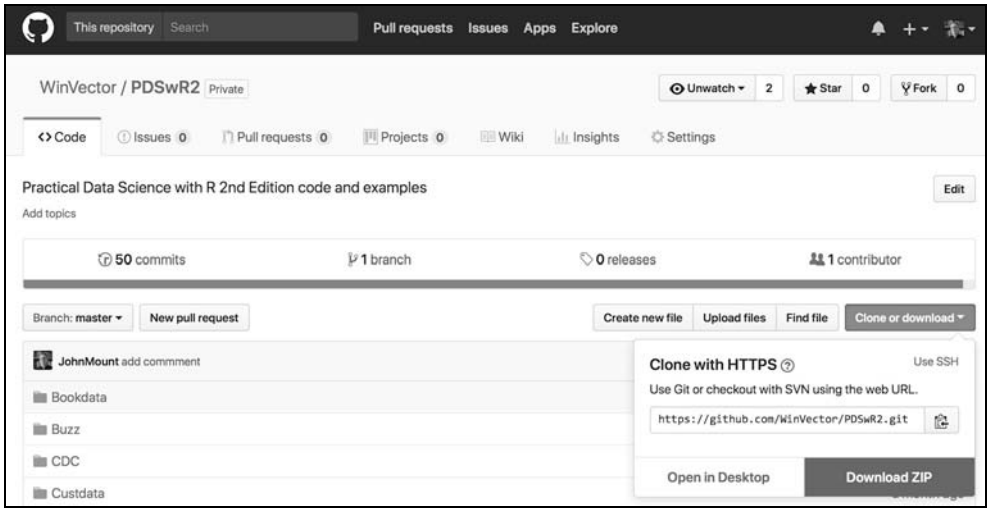
本书是以示例为导向的。我们在 GitHub 资料库里提供了准备好的示例数据(https://github.com/WinVector/PDSwR2)，同时附有 R 代码并且有原始的链接。你可以在线浏览此资料库，也可以将其复制到自己的计算机上。我们也以 zip 文件的方式提供了产生所有结果的代码和几乎所有能在本书中找到的图表(https://github.com/WinVector/PDSwR2/raw/master/CodeExamples.zip)，因为从 zip 文件里复制代码比从书上复制和粘贴要容易得多。关于下载、安装和使用所有推荐的工具和示例数据的指令可以在附录A中的A.1节中找到。我们鼓励你在阅读文本时尝试运行R代码的示例。即使我们讨论的是数据科学中比较抽象的内容，也会用具体的数据和代码来举例说明。每个章节都包含引用特定数据集的链接。R 代码的编写不需要任何命令行提示符，例如>(它通常在运行R代码时出现，但不能作为新的R代码输入)。内联结果是以R的注释字符#作为前缀。在大多数情况下，初始源代码已经被重新格式化；我们添加了换行符并重新编写行首缩进以便适应书中可用的页面空间。在极少数情况下，代码清单中还包括了行连接标记(➥)。另外，当在文本中描述代码时，会将源代码中的注释删除。而许多代码清单中包含了代码注释，用于强调重要的概念。

如何使用本书

我们建议最好在阅读本书的同时至少运行一些示例。为此，建议你安装 R、RStudio，以及一些书中常用的程序包。附录A的A.1节中分享了关于如何执行此操作的说明。我们也建议你通过 https://github.com/WinVector/PDSwR2 的 GitHub 资料库或本书封底上的二维码下载所有的示例，包括代码和数据。

下载本书的辅助资料/资料库

可以使用“download as zip”GitHub 这一特性，将资料库的内容以 zip 文件的格式下载下来，如下图所示，下载来自 GitHub 网址 https://github.com/WinVector/PDSwR2 的内容。



GitHub 下载示例

点击 Download ZIP 链接应该会下载程序包的压缩版本(或者你可以尝试直接访问 ZIP 资料的链接：https://github.com/WinVector/PDSwR2/archive/master.zip)。或者，如果你熟悉在命令行中运行 Git 资源管理系统，可以使用 Bash shell 命令(不是 R 命令)：

```
git clone https://github.com/WinVector/PDSwR2.git
```

在本书所有示例中，我们假设读者已经复制了资料库或下载并解压了其中的内容。这会生成一个名为 PDSwR2 的目录。我们讨论的路径将会从这个目录开始。比如，如果我们提到使用 PDSwR2/UCICar 目录，则意味着无论你在哪里解压了 PDSwR2 目录，我们都会对 UCICar 子目录下的内容进行操作。你可以通过 setwd()命令来更改 R 的工作目录(请在 R 控制台中输入 help(setwd)来获得更多帮助信息)。当然，如果你使用的是 RStudio 工具，那么也能通过文件浏览界面的 gear/more 菜单选项来设置工作目录。本书中的所有代码示例都包含在 PDSwR2/CodeExamples 目录中，因此不需要输入代码(尽管你需要在自己定义的数据目录中运行它们——而不是在下载代码的目录中执行它们)。

本书并没有提供完整的练习，只是提供了一些示例。我们建议读者学习示例并尝试修改示例。例如，在 2.3.1 节中，我们展示了如何预测收入与受教育程度和性别的关系，尝试将收入与就业状况和年龄联系起来也是有意义的。实际上，数据科学要求你对编程、函数、数据、变量和关系有一定的好奇心，越早在自己的数据中发现不同，就越容易对它们进行处理。

关 于 作 者

Nina Zumel 曾在一家独立的、非营利性研究机构 SRI International 担任科学家。她曾在一家价格优化公司担任首席科学家，并创办了一家合同研究公司。Nina 现在是 Win-Vector LLC 的首席顾问。读者可以通过 nzumel@win-vector.com 联系她。

John Mount 曾是生物科技领域的计算科学家和股票交易算法的设计师，并且为 Shopping.com 管理过一个研究团队。他现在是 Win-Vector LLC 的首席顾问。读者可以通过 jmount@win-vector.com 联系他。

关于序言作者

Jeremy Howard 是一位企业家、商业战略家、开发人员和教育家。他是 fast.ai 的创始研究员，fast.ai 是一家致力于让深度学习变得更加容易访问的研究机构。他也是旧金山大学的教员，并且是 doc.ai 和 platform.ai 的首席科学家。

在此之前，Jeremy 是 Enlitic 的创始 CEO(首席执行官)，这是第一家将深度学习应用于医学的公司，并被《麻省理工科技评论》连续两年评选为世界最聪明的 50 家公司之一。他曾经是数据科学平台 Kaggle 的会长和首席科学家，在 Kaggle 他连续两年参加国际机器学习大赛并获顶级排名。

Rachel Thomas 是 USF Center for Applied Data Ethics 的董事以及 fast.ai 的联合创始人，在《经济学人》《麻省理工科技评论》和《福克斯》中都被提到过。她被福布斯选为人工智能领域中 20 位不可思议的女性之一，她在杜克大学获得了数学博士学位，并且曾是 Uber 的早期工程师。Rachel 是一位受欢迎的作家和主题演讲人。在她的 TEDx 演讲中，她分享了人工智能让她害怕的地方以及为什么我们需要拥有不同背景的人来参与人工智能。

关于封面插图

本书封面上插图的标题为 Habit of a Lady of China in 1703。该插图是取自 Thomas Jefferys 的 *A Collection of the Dressesof Different Nations, Ancient and Modern* (4 卷本)。该图集于 1757 年到 1772 年期间在伦敦出版。标题页上注明这些图画是手工着色的铜板雕刻，并用阿拉伯树胶增色。Thomas Jefferys (1719—1771)被称为“国王乔治三世时期的地理学家”。他曾是一位英国制图师，是当时主要的地图供应商。他为政府和其他官员刻印地图并制作了各种各样的商业地图和地图集，尤其是北美地区的地图册。作为一名制图员，他的工作令他对其所勘测和绘制的这些国家的当地服饰和习俗产生了兴趣。

对遥远土地的迷恋和休闲旅行在 18 世纪是相当新奇的现象，像这样的收藏品很受欢迎，向游客和神游旅行者介绍其他国家的风土人情也是很受欢迎的。Jefferys 作品中绘画的多样性生动地说明了一个世纪前世界各国的独特性和个性。从那时起，人们的着装方式发生了变化，当时如此丰富的地区差异也逐渐消失。现在很难区分不同大陆的居民，更不用说不同的城镇、地区或国家了。也许我们已经用文化多样性换取了更为多样化的个人生活，当然也换成了更加多样化和快节奏的科技生活。

在一个很难区分一本书和另一本书的时代，Manning 出版社以三个世纪前丰富多样的服饰为基础来制作图书封面(由 Jefferys 提供的生动图片)，从而展现出计算机行业的差异化和创新性。

目　录

第Ⅰ部分

数据科学引论

在本书的第Ⅰ部分，我们会重点介绍数据科学中最基本的任务：如何与你的合作伙伴一起工作，如何定义问题，以及如何检查数据。

第 1 章介绍了一个典型的数据科学项目的生命周期，包括项目团队成员的不同角色和职责，一个典型项目的不同阶段，以及如何定义目标和设定项目预期。该章作为本书其余部分的概览，内容也按照我们讨论的主题顺序进行组织。

第 2 章深入到具体细节，介绍了如何将数据从各种外部格式加载到 R 系统中，以及如何将数据转换为适合分析的格式，还讨论了对于数据科学家来讲最重要的 R 语言数据结构：数据框(data frame)。有关 R 编程语言的更多详细信息将在附录 A 中提供。

第 3 章和第 4 章介绍了建模阶段之前应做的数据探索和处理。在第 3 章，我们讨论了在数据处理中将遇到的一些典型问题和难点，以及如何使用概要统计信息和可视化功能来检测这些问题。在第 4 章，讨论了数据处理方法，以帮助你解决数据处理中的问题和难点。我们也推荐了一些规则和方法，以帮助你在项目的不同阶段更好地管理数据。

第 5 章介绍了如何将数据整理成易于分析的格式。

在完成第Ⅰ部分之后，你将学会如何定义数据科学项目，如何将数据加载到 R 系统中，以及如何为建模和分析准备数据。

第1章

数据科学处理过程

本章内容：

- 数据科学的定义
- 数据科学项目角色的定义
- 了解数据科学项目的各个阶段
- 为一个新的数据科学项目设定预期

数据科学是一门跨学科的实践，它综合了数据工程、描述统计学、数据挖掘、机器学习和预测分析等理论和方法。与运筹学一样，数据科学也侧重于实施以数据为依据的决策和管理。在本书中，我们将集中讨论使用这些技术将数据科学应用于商业和科学问题的方法。

数据科学家负责从头到尾地指导一个数据科学项目。数据科学项目的成功不是由于使用了任何一种外来工具，而是由于有可量化的目标、良好的方法、跨学科的交互和可重复的工作流程。

本章将介绍一个典型的数据科学项目，具体分析你遇到的问题类型、你应该拥有的目标类型、你可能要处理的任务以及预期的结果类型。

我们将使用一个具体的、真实的例子来开始本章的讨论[1]。

示例 假设你在一家德国银行工作。这家银行认为，由于不良贷款而损失了太多钱，因此希望减少损失。为此，他们需要一种工具来帮助信贷员更准确地发现风险贷款。

这就是你的数据科学团队的切入点。

1 在本章，我们将使用综合生物学教授汉斯·霍夫曼博士于 1994 年捐赠给 UCI 机器学习知识库的一个信用数据集。为了清楚起见，我们简化了一些列名。原始数据集可以在 http://archive.ics.uci.edu/ml/datasets/Statlog 上通过查找(German+Credit+Data)关键字找到。我们将在第 2 章演示如何加载这些数据并为分析做好数据准备。请注意，在收集数据时，德国货币单位为德国马克(DM)。

1.1　数据科学项目中的角色

一个数据科学项目并非凭空进行的。它需要众多的角色、技能和工具协同工作。在讨论项目本身之前，先来看看在一个成功的项目中必须拥有的角色。长期以来，项目管理一直是软件工程的核心问题，因此我们可以从中寻求指导。在定义角色时，我们从 Fredrick Brooks 的“外科手术团队”的观点中借鉴了一些软件开发的想法，如他在 *The Mythical Man-Month：Essays on Software Engineering*(Addison-Wesley, 1995)中所述。我们还借鉴了敏捷软件开发范式的思想。

项目角色

下面列出了数据科学项目中始终存在的几个角色，如表 1.1 所示。

表 1.1　数据科学项目中的角色和职责

角　　色	职　　责
项目出资方	代表商业利益，提供项目支持
客户	代表最终用户的利益，领域专家
数据科学家	设定和执行分析战略，与出资方和客户沟通
数据架构师	管理数据和数据存储，有时要管理数据的收集
运营工程师	管理基础设施，部署最终项目成果

有时这些角色可能会重叠。有些角色，特别是客户、数据架构师和运营工程师，通常并不是数据科学项目团队中的人员，但他们却是关键的合作者。

1. 项目出资方

数据科学项目中最重要的角色是项目出资方。出资方是想获得数据科学成果的人。通常，他们代表了商业利益。在贷款申请示例中，出资方可能是银行的消费信贷主管。出资方负责认定项目是成功还是失败。如果数据科学家认为自己了解并可以代表业务需求，则可以担任自己项目的出资方角色，但这不是最佳安排。理想的出资方应满足以下条件：如果他们对项目结果感到满意，那么该项目可定义为成功。让出资方签收成为一个数据科学项目的重要的组织目标。

保持出资方知情并介入　保持出资方知情并参与项目至关重要。要按照他们能够理解的程度，向他们展示项目的计划、进度以及阶段性的成功或失败等情况。反之，让出资方两眼一抹黑必然导致项目失败。

为确保出资方签收，你必须通过与出资方的直接交谈获得清晰的目标。你应该力求用量化的语言得到出资方所表述的目标。这是一个目标示例：“在出现第一笔逾期还款的至少 2 个月之前，标识出 90%的拖欠账目，且假阳性率不高于 25%”。这个精确的目标表述使你能够同时检查该目标满足时是否能够达到商业意图，以及你是否拥有高质量的数据和工具去达到这个目标。

2. 客户

出资方是代表商业利益的角色，而客户是代表模型中最终用户的利益的角色。有时，出资方和客户的角色由同一个人担任。如果数据科学家能权衡商业利弊，也就能承担客户角色，但这不是理想的情况。

客户比出资方更有实践经验。在构建一个好的模型所需的技术细节与部署该模型所需的日常工作流程之间，客户起到了接口作用。他们不需要精通数学或统计学，但需要熟悉相关业务流程，可担当团队的领域专家。在本章后面将讨论的贷款应用示例中，客户可以是信贷员，或是代表信贷员利益的人。

作为出资方，应保证客户能知情和介入项目。理想的情况是，你可以与他们一起定期召开会议，使你的工作与最终用户的需求一致。通常，客户会隶属于一个机构中的不同部门，有该项目之外的其他职责。所以，必须保持会议聚焦，以客户容易理解的方式展现成果和进展，并将客户的批评牢记于心。如果最终用户不能或者不想使用你的模型，则项目最终是不成功的。

3. 数据科学家

数据科学项目中的下一个角色是数据科学家，负责执行保障项目成功的所有必要步骤，包括设定项目战略和保证客户知悉情况。他们设计项目步骤，选取数据源，选取使用的工具。由于他们要选择各种尝试使用的技术，因此必须精通统计学和机器学习。他们也负责项目计划和跟踪，有时会和他们的项目管理伙伴一起来做这项工作。

在更多的技术层面上，数据科学家也要检查数据，进行统计检验和处理，应用机器学习模型和评价结果，这些是数据科学项目中的科学部分。

领域理解能力

要求数据科学家成为领域专家往往太过分了。但在实际项目中，数据科学家必须培养自己强大的领域理解能力，以帮助正确地定义和解决问题。

4. 数据架构师

数据架构师负责所有的数据及其存储。这个角色常常是由数据科学团队之外的人来担当，例如数据库管理员或架构师。数据架构师经常忙于为不同的项目管理数据仓

库，因此，只有在需要紧急咨询时才可寻求他们的帮助。

5. 运营工程师

运营工程师在获取数据和提交最终结果过程中都是关键的项目角色。担任该角色的人通常在数据科学团队之外承担运维职责。例如，如果要部署一个数据科学结果，该结果会影响在线商店网站中的商品排序，那么负责运营该网站的人在决定如何进行部署上具有较大的发言权。此人会给出响应时间、编程语言和数据大小等方面在部署时需要考虑的约束。担任运营工程师角色的人可能一直在支持你的出资方和客户，所以他们很容易被找到(尽管他们已经按要求花费了很多时间)。

1.2　数据科学项目的阶段

理想的数据科学环境是鼓励数据科学家与所有其他利益方之间进行反馈和反复沟通的。这在数据科学项目的生命周期中得到充分反映。尽管本书像其他关于数据科学处理过程的讨论一样，将其生命周期分成不同阶段，但现实中各阶段之间的边界并不是固定的，一个阶段的活动经常与另一个阶段的活动重叠[1]。在整个处理过程向前推进时，你经常需要在两个或更多的阶段之间往复循环，如图 1.1 所示。

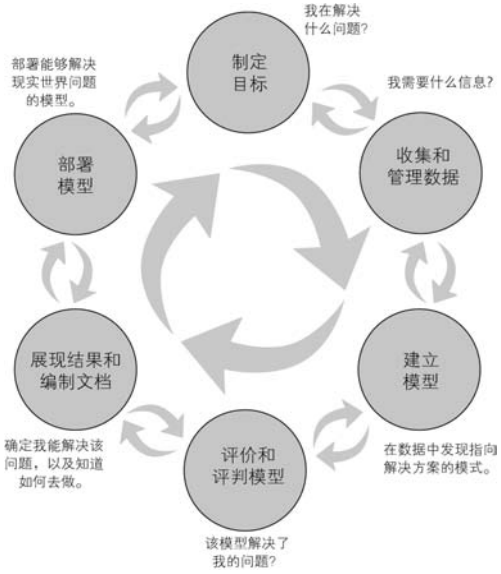


图 1.1　一个数据科学项目的生命周期：嵌套循环过程

1　机器学习过程的一个常见模型是针对数据挖掘的跨行业标准处理过程(CRISP-DM)(https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining)。与我们将在此处讨论的模型是类似的，但我们强调在处理过程的任何阶段都可以循环往复。

即使在你完成一个项目并部署一个模型之后，从模型运行过程中也能发现新的问题和可疑点。一个项目的结束可能会导致一个后续项目的开始。

下面来看图 1.1 中给出的各个阶段。

1.2.1 制定目标

数据科学项目的第一个任务就是制定一个可衡量和可量化的目标。在该阶段，要尽你所能了解该项目的背景信息：

- 为什么出资方需要设立该项目？他们缺少什么以及需要什么？
- 出资方正在做什么事情来解决问题？为什么还不够好？
- 你需要什么资源：什么种类的数据以及需要多少？你是否需要有一起合作的领域专家？计算资源是什么？
- 项目出资方计划如何部署你的结果？为了成功地实施部署，必须满足哪些约束条件？

下面回到贷款应用示例，其最终的商业目标是减少银行因不良贷款导致的损失。项目出资方希望有一种帮助信贷员的工具，可以更精确地为贷款申请人打分，从而减少产生不良贷款的数量。同时，让信贷员觉得他们对于批准贷款具有最终的决断能力，这也是很重要的。

一旦你和项目出资方及其他利益方对这些疑问确立了基本答案，你就可以和他们一起开始制定项目的精确目标了。该目标应该是具体的和可衡量的，不是“我们想要更好地发现不良贷款”，而是“我们要利用一个模型去预测哪些贷款申请人可能会拖欠贷款，从而将放款坏账率降低至少 10%”。

通过项目的具体目标可得出该项目具体的结束条件和接受条件。目标越不具体，项目就越可能没有界限，因为没有任何结果可以是“足够好的”。如果你不知道想达到什么目标，就不知道何时停止尝试，甚至不知去尝试什么。当项目由于时间到期或者资源耗尽而不得不结束时，没有人会对结果满意。

当然，这并不意味着不需要更具探索性的项目。例如，哪些数据与高拖欠率相关？或者是否应该减少发放贷款的种类？应该去掉哪种贷款？对于这些情况，要仔细检查项目的具体结束条件，如时间限制。例如，你可能决定花两周而不是更多的时间浏览数据，以期得出候选的目标假设。然后，这些假设能够变成针对整个建模项目的具体问题或目标。

一旦有了关于项目目标的好想法，就能集中精力收集数据以满足那些目标。

1.2.2　收集和管理数据

这个阶段包含识别你需要的数据并对数据进行探索，使其满足分析的条件。通常，这个阶段是处理过程中最耗时的一个步骤，也是最重要的阶段之一：

- 什么数据可以供我用？
- 这些数据是否有助于我解决问题？
- 这些数据是否足够多？
- 数据的质量是否足够好？

假设在上述贷款申请问题中，你已经收集到最近十年的代表性贷款的样本数据。一些贷款已经拖欠，但大多数(大约 70%)没有拖欠。你还收集到了贷款申请的各种属性，如表 1.2 所示。

表 1.2　贷款数据属性

Status_of_existing_checking_account(状态为申请中)
Duration_in_month (按月计算的贷款时间)
Credit_history(信用历史)
Purpose (汽车贷款、学费贷款等用途)
Credit_amount(贷款数量)
Savings_Account_or_bonds(储蓄账号或债券的结余和数量)
Present_employment_since(当前职业)
Installment_rate_in_percentage_of_disposable_income(可支配收入的分期还款百分比)
Personal_status_and_sex(个人状态和性别)
Cosigners(联合签署人)
Present_residence_since(当前住址)
Collateral (汽车、财产等担保)
Age_in_years(年龄)
Other_installment_plans (其他分期还款计划，其他贷款、贷款种类)
Housing(自有的、租住的住宅等)
Number_of_existing_credits_at_this_bank(在本银行的信贷数量)
Job (职业类型)
Number_of_dependents(赡养者人数)
Telephone (电话号码，如果有的话)
Loan_status (贷款的优劣，因变量)

在你的数据中，Loan_status 有两个取值：GoodLoan 和 BadLoan。为便于讨论，假定 GoodLoan 表示还清贷款，BadLoan 表示拖欠贷款。

尽可能地使用可直接度量的信息 要尽可能多地使用可以直接度量的信息，而不是使用从另一个度量中推导出的信息。例如，你也许可以将收入用作一个变量，从而推测低收入者可能会更难付清贷款。但是，如果考虑与借贷者可支配收入相比的还贷数额大小，这个值可以更加直接地度量还清贷款的能力。这个信息比单独的收入数据更有用。你可以通过变量 Installment_rate_in_percentage_of_disposable_income 得到这个数据。

在这个阶段，你将对数据进行初始的探索和可视化处理，也将清洗数据，包括根据需要修复数据错误、转换变量。在探索和清洗数据的过程中，你可能会发现这些数据不适合，或者还需要其他类型的信息。也许你会在数据中发现一些端倪，引出新的问题，而这些问题比你原计划要解决的问题更加重要。例如，图 1.2 中的数据似乎看起来是不合常理的。

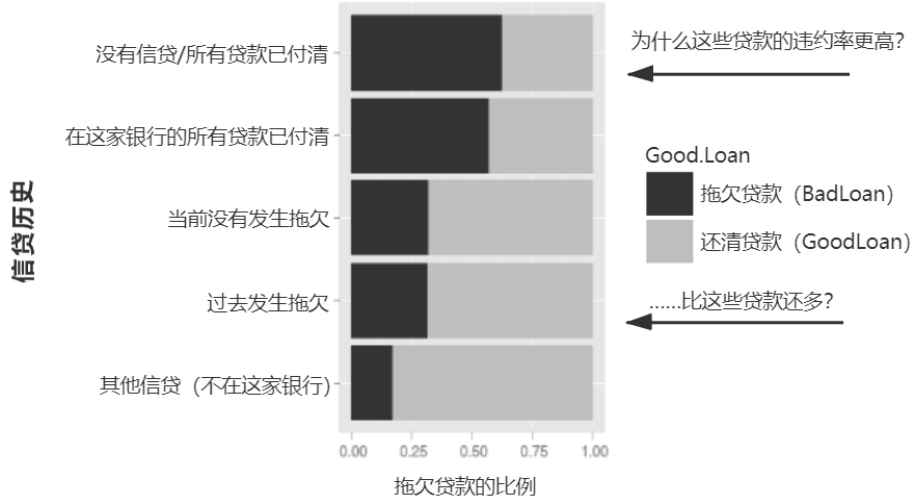


图 1.2 按照信贷历史的类别显示拖欠贷款所占比例
(条形的深色部分代表在该类别中拖欠贷款所占比例)

为什么看起来安全的申请者(已向银行还清了所有信贷)比看起来有风险的人(在过去拖欠过贷款)具有更高的拖欠贷款比例？当你更仔细地检查数据，并与其他利益方和领域专家就该问题进行讨论后，你会认识到这个样本数据存在固有的偏差：你只有实际已发生的贷款数据(已被批准的)。一个真实没有偏见的贷款申请样本应包括已接受的和被拒绝的贷款申请。总的来说，因为你的样本只包括已接受的贷款，因此在数据中看起来有风险的贷款比看起来安全的贷款要少得多。可能的情况是，看起来

有风险的贷款经过非常严格的审查过程才能被批准，而看起来安全的贷款申请也许绕过了这些审查。这说明，如果你的模型用于当前申请批准流程的下游，则信贷历史不再是有用的变量。也说明对于那些即使看起来安全的贷款申请也应该更加仔细地严格审查。

类似这样的发现将引导你和其他利益方去修改或者改善项目目标。在这个示例中，你将决定专门关注那些看起来安全的贷款申请。当你在数据中发现有用的信息后，在这个阶段和前一阶段之间，以及在这个阶段和建模阶段之间，进行反复循环处理是常见的情况。我们将在第 3 章和第 4 章深入介绍数据的探索和管理。

1.2.3　建立模型

在模型建立或分析阶段，最终要用到统计学和机器学习。因此，你需要从数据中获取有用的见解，以达到目标。由于许多建模过程需要关于数据分布和关联关系的具体假设，因此，为了找到最好的数据表达方法和最好的数据建模方式，在建模阶段和数据清洗阶段之间会有重叠和反复。

最常见的数据科学建模任务有：

- 分类——决定某个特性属于哪个类别。
- 打分——预测或者预估一个数值，如定价或概率。
- 排名——按照偏好对条目(item)进行排序。
- 聚类——将条目分到最相似的组。
- 发现关系——在数据中找出相关性或潜在原因。
- 特征化——从数据中生成通用的绘图或报表。

对于每个任务，都有多种可用的方法。在本书中，我们将针对每种任务介绍一些最常用的方法。

上述贷款申请问题是个分类问题：识别出可能拖欠的贷款申请人。常用的方法是逻辑回归法和基于树的方法(将在第 7 章和第 10 章深入介绍)。当你与信贷员以及将要实际使用模型的人进行交谈后，了解到他们想要知道在该模型的分类功能背后的推理链，以及该模型所做出决策的置信度指标：该申请人很可能拖欠，或者只是有些可能？为了解决这个问题，你认为决策树是最适合的方法，但就目前而言，我们只能查看决策树模型的结果，因为我们将在第 10 章更详细地讨论决策树[1]。

1　本章为了演示方便，仅选用了一棵分支分层少的“小”树。

假设你使用图 1.3 所示的模型。让我们遍历一下树的示例路径。假设有一笔 10 000 马克(德国马克，当时研究用的货币)的一年期贷款申请。在树的顶部(图 1.3 中的节点 1)该模型检查贷款是否超过 34 个月。答案是“否”，因此该模型沿树的右分支向下移动。该分支从节点 1 开始显示为一条颜色加重的线。下一个问题(节点 3)是贷款是否大于 11 000 德国马克。同样，答案是“否”，因此模型沿着右边(从节点 3 开始显示为一条颜色加重的线)的分支向下移动并到达叶节点 3。

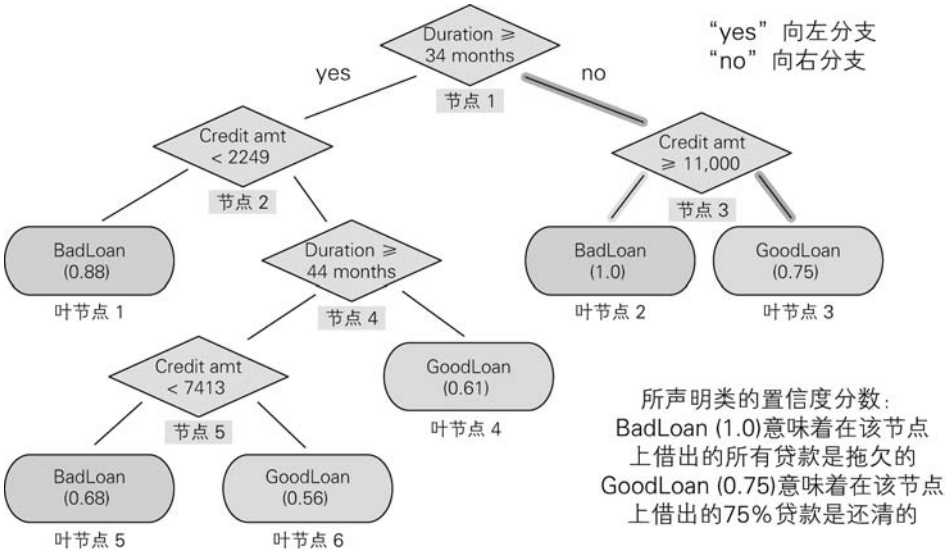


图 1.3 决策树模型，用于找出不良贷款申请，且带有置信度分数

从历史数据看，到达此叶节点的贷款中有 75%是可以还清的贷款，因此模型建议你批准这笔贷款，因为这笔贷款被偿还的可能性很高。

此外，假设有一笔 15 000 德国马克的一年期贷款申请。在这种情况下，该模型将首先在节点 1 上沿着右边向下分支，然后在节点 3 上向左分支，到达叶节点 2。从历史数据看，到达叶节点 2 的所有贷款都已发生拖欠，因此模型建议你拒绝此贷款申请。

我们将在第 6 章讨论一般的建模策略，并在第 II 部分详细介绍特定的建模算法。

1.2.4 评价和评判模型

一旦有了模型，就需要确定它是否满足目标：

- 对你的需求来说是否足够准确？它能否很好地概括需求？
- 它是否比“直观猜测”表现得更好？比你当前使用的任何估计都表现得更好？
- 模型结果(系数、聚簇、规则、置信区间、重要性和诊断)在问题领域的情景中是否有意义？

如果任何一个提问的答案为“否”，则需要重新循环建模步骤，或者，可推断出这些数据不支持你要达到的目标。虽然没有人愿意得到否定的结果，但如果能提前了解到不能用现有资源满足你的成功标准，就可以省去你无谓的努力。这样，你可以把更多的精力花在如何构思成功上，即制定更现实的目标，收集其他额外的数据，或者收集为了达到原始目标所需的其他资源。

回到贷款申请示例，首先要检查的是模型发现的规则是否有意义。从图 1.3 可以看出，你没有观察到任何明显异常的规则，接下来就可以评估模型的精确度。混淆矩阵可用于总结分类器的精确度，它将实际分类结果与预测分类结果用表格形式列出来[1]。

在代码清单 1.1 中，将创建一个混淆矩阵，矩阵的行表示实际贷款状态，而列表示预测的贷款状态。为了提高易读性，代码以矩阵元素的名称命名而不是按索引值命名。例如，conf_mat ["GoodLoan", "BadLoan"]表示 conf_mat [2, 1]元素。矩阵的对角项(diagonal entries)表示正确的预测。

代码清单 1.1　计算混淆矩阵

可以通过网址 https://github.com/WinVector/PDSwR2/blob/master/packages.R 找到运行本书示例所需的所有软件包

该文件可以通过网址 https://github.com/WinVector/PDSwR2/tree/master/Statlog 找到

```
library("rpart")
load("loan_model_example.RData")
conf_mat <-
      table(actual = d$Loan_status, pred = predict(model, type = 'class'))
##            pred
## actual      BadLoan GoodLoan
##    BadLoan        41      259
##    GoodLoan       13      687

(accuracy <- sum(diag(conf_mat)) / sum(conf_mat))
## [1] 0.728
```

创建混淆矩阵

模型总的精确度：73%的预测是正确的

1　正常情况下，我们会使用测试集(不用于建模的数据)评估模型。在本示例中，为简单起见，将用训练数据(用于建模的数据)评估模型。另外请注意，在绘图时我们遵循一个约定：x 轴表示预测，对于表格方式，预测用列名表示。请注意，混淆矩阵还有其他的约定规则。

```
(precision <- conf_mat["BadLoan", "BadLoan"] / sum(conf_mat[, "BadLoan"])
## [1] 0.7592593
(recall <- conf_mat["BadLoan", "BadLoan"] / sum(conf_mat["BadLoan", ]))
## [1] 0.1366667

(fpr <- conf_mat["GoodLoan","BadLoan"] / sum(conf_mat["GoodLoan", ]))
## [1] 0.01857143
```

模型精确度：预测为不良的申请人中 76%的申请人确实拖欠贷款

假阳性率：2%的优质申请人被误判为不良申请人

模型召回率：模型发现了14%的申请人拖欠贷款

该模型正确地预测了 73%的贷款状态，比机会概率(50%)好。在原始数据集中，有 30%的不良贷款，剩余均猜测为 GoodLoan 也能达到 70%的精度(虽然不是很有用)。因此，该模型明显优于随机预测，稍微优于直观猜测。

该模型总的精度不够好，你想知道究竟出了什么错误：是弄错了太多的不良贷款，还是把太多的优质贷款识别为了不良贷款？召回率用于衡量该模型可实际发现多少不良贷款；精确度用于衡量有多少识别出的不良贷款确实属实；假阳性率衡量有多少优质贷款被错误地识别为不良贷款。理想情况下，希望召回率和精确度要高，而假阳性率要低。而什么是“足够高”和“足够低”则由你和其他利益方一起决定。做到恰当的平衡经常要求在召回率和精确度之间进行折中。

此外，也有其他衡量模型的精确度和质量的方法。我们将在第 6 章讨论模型评估内容。

1.2.5 展现结果和编制文档

一旦有了满足成功标准的模型，你会把结果展现给项目出资方和其他利益方。在部署模型之后，还必须给那些负责使用、运行和维护模型的机构编写模型文档。

不同的受众需要不同种类的信息。业务受众需要根据业务标准来理解你的发现所产生的影响。在贷款例子中，在向业务受众展现结果时，最重要的一点是解释你的贷款申请模型是如何减少损失的(银行在不良贷款里损失的钱)。假如该模型识别出的一组不良贷款占违约损失总额的 22%，那么，在你的展现结果或者执行摘要中，应强调该模型可能会将银行的损失减少到该数额，如图 1.4 所示。

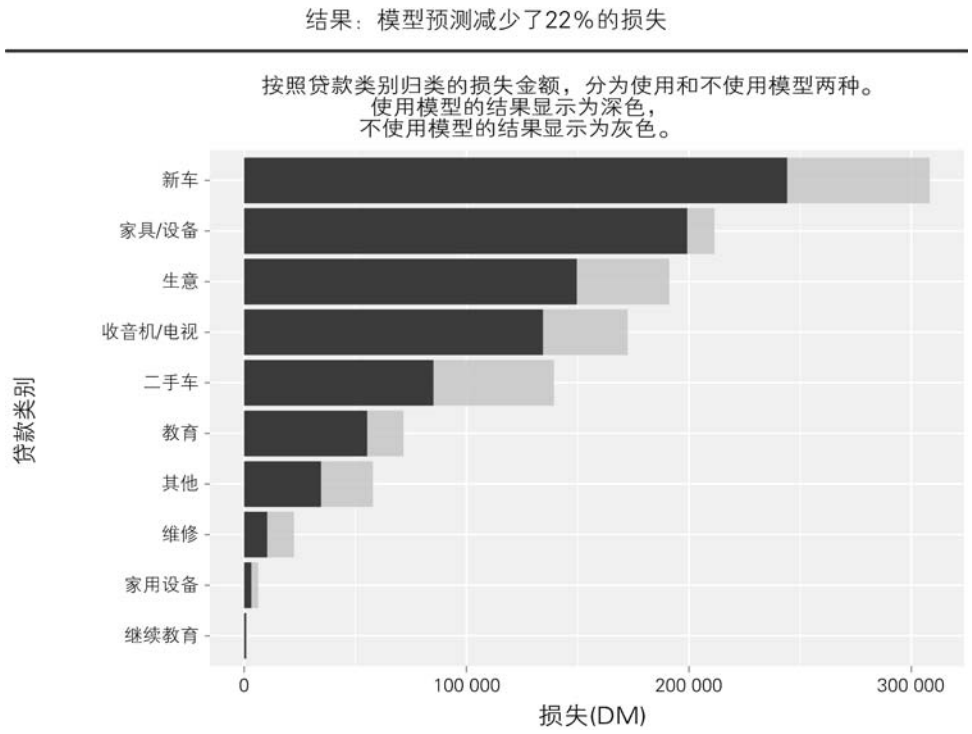


图 1.4　展现给执行主管所用的幻灯片示例

你也许想给这些受众最有意义的发现或推荐，例如，新车贷款比二手车贷款的风险更高，或者，最大的损失与不良汽车贷款和不良设备贷款有关(假设受众不知道这些事实)。对于受众来说，他们不会对模型的技术细节感兴趣，你应该略过这些技术细节或者只给他们提供高层次的展现。

为模型的最终用户(信贷员)做的展现要强调该模型如何帮助他们把工作做得更好：

- 他们应该如何解释该模型？
- 该模型的输出是什么？
- 如果该模型提供决策树规则的执行轨迹，应如何解读它？
- 如果该模型提供了分类的置信度分数，应该如何使用这个置信度分数？
- 他们何时可能会否决该模型？

为运营人员所做的展现或编制的文档应该强调你的模型对他们所负责内容的影响。我们将在本书的第III部分介绍为各种受众所做的展现和文档内容。

1.2.6 部署模型

最后，该模型要投入运行。在许多机构中，这意味着数据科学家不再主要负责该模型的日常操作。但你仍应该确保该模型平滑运行，不会产生灾难性的、无人监督的决策。你也要保证该模型可以随着环境的变化而更新。在很多情形下，宜将模型先部署为一个小型试点项目。在测试过程中可能会出现未预料到的问题，你必须随之调整模型。我们将在第 11 章讨论模型的部署。

在部署模型时，你会发现信贷员在某些情形下会频繁地推翻你的模型，因为该模型与他们的直觉相抵触。是他们的直觉有误？还是该模型不完备？另一方面，在得到充分肯定的情况下，你的模型可能运行得非常成功，银行会想进一步把它再扩展到住房贷款申请中。

在后续的章节中，我们将更深入地介绍数据科学生命周期的各个阶段。在此之前，我们先看看项目初始设计阶段的一个重要方面：设定预期。

1.3 设定预期

设定预期是制定项目目标和成功标准的关键。或许你团队中业务方面的成员(特别是项目出资方)已有一个关于如何满足商业目标所要求性能的想法，例如，银行希望至少减少 10%因不良贷款所造成的损失。在你深入参与一个项目之前，应该确定你所拥有的资源足以满足业务目标的需要。

举一个动态调整项目生命周期阶段的例子。在探索和清洗阶段，你对数据的了解变得更清晰了。在你对数据有概念后，就能够感知到数据是否足以满足预期的性能阈值。如果不能，就必须重新经历项目设计和目标设定阶段。

确定模型性能的下限

在定义验收标准时，了解模型应如何达到可接受的性能是非常重要的。

空值模型代表了你要努力获取的模型性能的下限。你可以将空值模型视为“直观的猜测”，你的模型必须做得比它好。如果你想改进某个工作模型或已有的解决方案，那么空值模型就是现成的方案。在没有现成的模型或解决方案的情况下，空值模型是最简单的可用模型：例如，猜测都为 GoodLoan，或者当你尝试要预测一个数值时，总是预测其输出的平均值。

在贷款申请示例中，数据集里 70%的贷款申请其实是优质贷款。若一个模型将所

有贷款标记为 GoodLoan(实际上，这仅使用现有的过程对贷款进行分类)，那么它能达到 70%的正确率。由此可知，建立在数据上的任何实际模型的精确度都应该高于 70%才有用——如果精确度是你的唯一标准。因为这是最简单的可用模型，所以它的出错率被称为基准错误率。

应该比 70%好多少？在统计学上，有个称为假设检验或显著性检验的过程，它检验模型是否等价于空值模型(这种情况下，就是检验一个新模型是否与猜测都为 GoodLoan 的精度一样)。你希望你的模型精度要“显著地优于”(按照统计学术语)70%。我们将在第 6 章详细介绍显著性检验。

精度不是仅有的(或者最好的)性能度量。正如之前所见，召回率是衡量模型识别出实际的不良贷款的比例。在我们的例子中，猜测都为 GoodLoan 的空值模型在识别不良贷款时的召回率为 0，这显然不是你想要的。一般地，如果实际中已经有模型或者处理过程，那么你可能已定义了精确度、召回率、假阳性率等值。改进其中的一项指标总是比只考虑精确度这一项更为重要。如果项目的目标是要改进现有的处理过程，那么说明当前模型的这些指标中至少有一个度量是不令人满意的。知晓了现有处理过程的局限，可以帮你确定所需性能的有用下限。

1.4　小结

数据科学处理过程存在着大量的反复——在数据科学家和其他项目利益方之间反复，或在处理过程的不同阶段之间反复。在处理过程中，你将遇到一些意外的事情和障碍。本书将教给你克服其中某些障碍的方法，从而确保所有利益方知情和参与，这很重要。这样，当项目完成时，与其有关的任何人都不会对最终结果感到意外。

在第 2 章，我们将跟随项目设计来介绍下一个阶段：加载、探索和管理数据。第 2 章涵盖了将数据加载到 R 系统的几种基本方法，采用的是一种便于分析的数据格式。

在本章中，你已学习了

- 一个成功的数据科学项目远不只是统计学，也要求有代表业务和客户利益的各种角色，以及运营中的关注点。
- 确保具有清楚的、可验证且可量化的目标。
- 确定为所有利益方设定现实的预期。