HZ BOOKS
华章IT

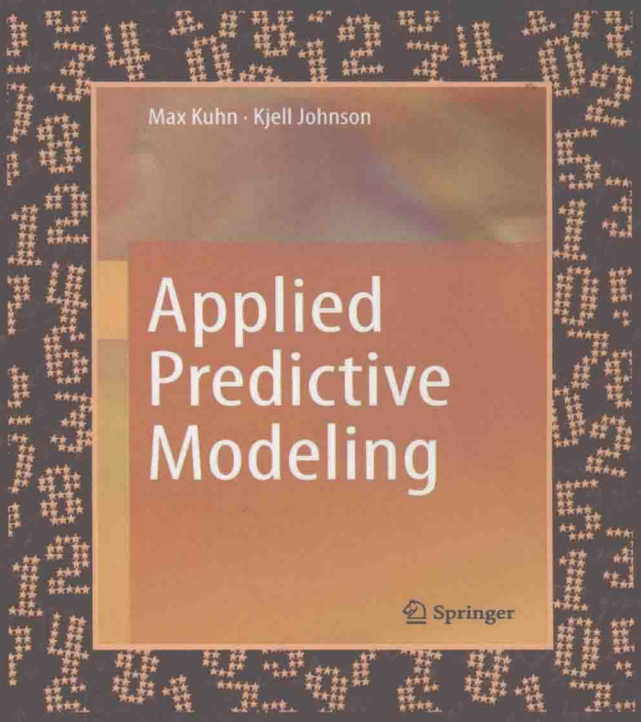
Springer

数据科学与工程技术丛书

应用预测建模

[美] 马克斯·库恩（Max Kuhn）
谢尔·约翰逊（Kjell Johnson） 著

林荟 邱怡轩 马恩驰 肖楠 张尚轩 译



APPLIED PREDICTIVE MODELING

机械工业出版社
China Machine Press

数据科学与工程技术丛书

APPLIED PREDICTIVE MODELING

应用预测建模

[美] 马克斯·库恩（Max Kuhn）
谢尔·约翰逊（Kjell Johnson） 著

林荟 邱怡轩 马恩驰 肖楠 张尚轩 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

应用预测建模 /（美）库恩（Kuhn，M.），（美）约翰逊（Johnson，K.）著；林荟等译 . —北京：机械工业出版社，2016.4
（数据科学与工程技术丛书）
书名原文：Applied Predictive Modeling

ISBN 978-7-111-53342-9

I. 应… II. ① 库… ② 约… ③ 林… III. 系统建模 IV. N945.12

中国版本图书馆 CIP 数据核字（2016）第 068979 号

本书版权登记号：图字：01-2014-0619

这是一本专注于预测建模的数据分析书，意在为实践者提供预测建模过程的指导，比如如何进行数据预处理、模型调优、预测变量重要性度量、变量选择等。读者可以从中学到许多建模方法以及提高对许多常用的、现代的有效模型的认识，如线性回归、非线性回归和分类模型，涉及树方法、支持向量机等。第 10 章和第 17 章分别研究混凝土混合物的抗压强度和作业调度两个案例。

作者重实际应用，轻数学理论，从实际数据出发，结合开源软件 R 语言来求解实际问题，详细给出 R 代码和处理的步骤。R 包 AppliedPredictiveModeling 包含书中使用的数据，以及可以用于重复书中每一章分析的 R 代码，让读者能在一定精度范围内重复本书的结果，并自然地将书中的预测建模方法应用到自己的数据上。章后附有习题，方便读者巩固所学。

这本业界互相推荐的好书，适合所有数据分析人员阅读。

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：明永玲　　责任校对：殷　虹
印　　刷：三河市宏图印务有限公司　　版　　次：2016 年 5 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：28.5（含 2.25 印张彩插）
书　　号：ISBN 978-7-111-53342-9　　定　　价：99.00 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）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

译 者 序

无需西装、领带、高脚杯和红唇，“数据科学家”本身就是21世纪“性感”的代名词！数据科学家在北美是高薪职业，相关人才成为各大科技公司争夺的对象。随着计算机技术的进步，数据科学成为热门话题，预测模型几乎能够用于你所能想到的任何一个领域。通过互联网上的海量数据加上如R、Python之类的开源工具，使得很多还是新手的数据分析从业者能够进行相对复杂的建模。数据建模分析竞赛平台Kaggle使企业和研究者可在其上发布数据，统计学者和数据挖掘专家可在其上进行竞赛以产生最好的模型，最优建模者可以获得企业提供的奖金或面试机会。

数据科学是很多不同学科的结合体（统计学、计算机科学、人工智能等，基于其应用的领域还要求特定的行业知识），从业者的背景跨度很大。相关书籍有些注重应用而没有提供足够的理论说明，有些又过于偏重理论而让读者不知如何有效应用。本书很好地平衡了两者，与其他书不同的是，本书对应有一个R包，其中包含许多代码示例，极大地方便了读者使用书中介绍的模型。

除了可重复性外，在我看来，本书的最大优点是介绍了从数据预处理到建模再到模型评估选择的整个过程，以及背后的**统计思想**。统计研究的不是**确定性**而是**不确定性**。统计学界泰斗George E. P. Box有这样一句名言：

“本质上讲，所有模型都是错的，但有一些是有用的。”

这短短的一句话体现了很高的**统计成熟度**。记得博士期间，讲《高级应用统计》的教授说过：“这门课的主要目的不是教统计知识，而是提高你们的统计成熟度。”该教授讲课天马行空，一学期下来让我觉得不着边际，但这五个字我牢牢地记住了，并在之后从业过程中不断隔空回响，成为我的职业箴言。阅读本书不仅可以学习统计知识，更重要的是可以提高统计成熟度。预测模型不是万能的，每一个预测都带有不确定性，建模者不是提供了预测值就万事大吉，更重要的是尝试尽可能多的模型，通过严格的训练测试探究模型的不确定性并且选出最优模型。在实际应用中，对不确定性的理解越深，越能在风险和收益之中做出权衡，预测模型产生的实际影响就越大。理论和应用之间还有相当长一段路要走，本书就是连接这两点的一条路。

在负责杜邦先锋北美市场预测建模两年多来，本书给我很大的帮助。我相信无论你是

数据分析的新手，还是数理统计的博士，本书都会让你受益匪浅。如果你打算从事预测建模的工作，本书绝对不容错过。

本书的翻译工作是由5人合作完成的。林荟翻译了书的第1～4，16，18、19章和第14章的后半部分。邱怡轩和肖楠共同翻译了第5～10章。马恩驰翻译了第11、15、17章，以及第14章的前半部分。张尚轩翻译了第12、13章。邱怡轩、肖楠和林荟负责审校。在翻译和校对过程中，我们对原书的一些明显错误做了修订，有的地方加上了译者注以帮助读者理解。机械工业出版社的明永玲编辑对该书的翻译工作给予了大力的支持和帮助。在此对所有为本书中文版问世做出努力的人表示感谢！

限于译者水平，书中难免有错误和不妥之处，恳请读者批评指正。

林　荟

前　言

这是一本关于**数据分析**的书，专注于**预测建模的实际应用**。“预测建模”一词可能让人联想起诸如机器学习、模式识别和数据挖掘。事实上，这样的联想是很自然的，这些专业名词指代的方法是预测建模整体过程的一部分。但是预测建模所涵盖的范围远大于发现数据模式的工具和技术。**应用预测建模**定义了这样一个建立模型的过程，我们能理解和量化模型对未来即将看到的数据的预测准确度。本书的核心内容就是其中的整个过程。

本书意在为实践者提供预测建模过程的指导，读者可以从阅读中学到许多（建模）方法以及提高对许多常用的、现代的有效模型的认识。我们会介绍许多统计和数学技术，但在任何情况下我们描述技术细节的动机都是帮助读者理解模型的优缺点，而非（单纯）数理统计知识。我们极力避免复杂的公式，但是有少数例外。关于预测模型的理论知识，推荐这两本书，即 Hastie 等（2008）和 Bishop（2006）。本书的读者需要有一些基本的统计学知识，包括方差、相关性、简单线性回归以及基本的统计假设检验（如 p 值和检验统计量）。

预测建模的过程本质上具有很强的应用实践性。但我们研究发现，很多文章、出版物不能让读者再现（他们的）建模结果，因为数据不公开，或读者无法使用相应软件，又或软件需付费。Buckheit 和 Donoho（1995）对传统学术界提出了相似的批评：

一篇发表于科学刊物上关于计算机科学的文章本身不是学术，仅是关于学术的广告。真正的学术是完整的软件开发环境和能够生成那些图的所有指令集。

因此，我们的目标是尽可能地具有实践应用性，让读者能在一定精度范围内重复本书的结果，且可以自然地将书中的预测建模方法应用到他们自己的数据上。再者，对于整个建模过程，我们使用 R 语言（Ihaka 和 Gentleman 1996；R Development Core Team 2010），这是一个用于数学和统计计算的免费软件。几乎所有例子中的数据集都可以在相应 R 包中找到。R 包 AppliedPredictiveModeling 包含了书中使用的很多数据，以及可以用于再现书中每一章分析结果的 R 代码。

我们选择 R 作为计算引擎有如下几个原因。首先 R 是免费的（虽然也有商业版

的 R），可以在不同的操作系统上使用。其次，它在**通用公共许可**（General Public License）下发行（免费软件基金 2007 年 6 月），该许可阐明程序再次发布的规则。在此构架下，任何人可以任意检查、修改源程序。由于开源特性，很多预测模型已经由 R 包可以实现。再者 R 有进行预测建模的大量强大的功能。不熟悉 R 的读者可以在网上找到大量的入门教程（见附录）。

由于篇幅所限，本书没有涵盖广义加性模型、模型集成、网络模型、时间序列等内容。

本书还有一个配套网站：

http://appliedpredictivemodeling.com/

其中含有一些相关内容。

没有如下这些人的指导和帮助不会有本书的问世：Walter H. Carter, Jim Garrett, Chris Gennings, Paul Harms, Chris Keefer, William Klinger, Daijin Ko, Rich Moore, David Neuhouser, David Potter, David Pyne, William Rayens, Arnold Stromberg 和 Thomas Vidmar。我们还要感谢 Ross Quinlan 对 Cubist 和 C5.0 部分的帮助，他们帮我们修正了这两部分的一些描述。我们还要感谢 Springer 出版社的 Marc Strauss 和 Hannah Bracken 以及审阅者 Vini Bonato、Thomas Miller、Ross Quinlan、Eric Siegel、Stan Young 和一位匿名审阅者。最后我们要感谢家人的支持：Miranda Kuhn, Stefan Kuhn, Bobby Kuhn, Robert Kuhn, Karen Kuhn 和 Mary Ann Kuhn; Warren 和 Kay Johnson, Valerie 和 Truman Johnson。

Max Kuhn

Kjell Johnson

目　录

第1章
导　论

人们每天都面临着如下问题："今天我该从哪条路上班？""是否要更换自己的移动电话运营商？""我该如何投资储蓄？"或者"我会得癌症吗？"这些问题说明我们渴望知道将来发生的事情，我们热切地想要为将来做出现在最好的决策。

我们通常是基于信息做出决策的。在某些情况下，我们有具体的、客观的数据，例如早晨的交通状况或者天气预报。其他时候我们用自己的直觉和经验做出判断，如"我今早不应该从那座桥通过，因为那座桥通常一到下雪天就无法通行"或者"我需要做一个PSA检查因为我的父亲得过前列腺癌"。在这两个例子里，我们根据当前掌握的信息和经验来预测将来的事件，进而根据自己对将来的预测做出决策。

由于通过互联网和媒体，信息的获取越来越容易，我们更加强烈地希望利用这些信息来帮助我们做出决策。虽然人类大脑能够有意识、无意识地收集大量的数据，但大脑能处理的信息量甚至无法超过与当下要解决问题相关的、极易获得的那部分信息。为了帮助决策，我们转而使用一些工具。比如用Google过滤几十亿的网页，从中找到我们搜寻的信息；WebMD可以通过我们提供的症状诊断疾病；还有E*TRADE能够浏览成千上万的股票信息并找出对我们而言最优的投资组合。

如同其他许多网站一样，这些网站通过工具读取我们目前所有的数据，从头到尾筛选、寻找与我们要解决的问题相关的信息模式并且反馈结果。开发这类工具的过程在许多领域中都在不断发展演变，如化学、计算机科学、物理学和统计学。这个过程被称为"机器学习"、"人工智能"、"模式识别"、"数据挖掘"、"预测分析"以及"知识探索"。尽管每个领域使用不同的工具从不同的角度解决问题，但是它们的最终目标是一致的：**给出精确的预测**。在本书中，我们用一个词统一代表这些不同的名称：**预测建模**。

Geisser（1993）将预测建模定义为"建立或者选择能最好预测将来事件发生概率的模型的过程"。在本书中，我们对这个定义进行了一些微调：

预测建模：开发能够给出精确预测的数学工具或者模型的过程。

《Wired》杂志的Steve Levy针对预测模型的普及写道（Levy 2010），"[人工智能的]例子随处可见：Google全球的机器用它解密人工搜索；信用卡公司用它追踪信用欺诈；Netflix用它向订阅者推荐电影；财务系统用它处理数十亿的交易信息（系统偶尔会崩溃）。"如下列举了我们想要预测的一些问题：

- 这本书的销量会有多少？
- 这个客户会不会离开我们去别的公司？

- 在当前市场下我的房子的售价将会是多少？
- 某个病人是否患上了某种疾病？
- 基于某人之前的浏览记录，他会对哪些电影感兴趣？
- 我应该卖掉这只股票吗？
- 我们的在线婚介服务该把什么样的人配在一起？
- 这是一封垃圾邮件吗？
- 该患者会对这种治疗有反应吗？

另外一个例子是保险公司需要对汽车、健康和生命保险的投保人进行风险预测。这些信息将被用于决定投保人是否能买某种保险，如果可以，保险费是多少。和保险公司类似，政府需要进行风险预测，但目的是为了保护公民。最近关于政府预测模型的例子包括使用计量生物学模型鉴定恐怖主义嫌疑人、诈骗监测模型（Westphal 2008）以及对动荡混乱的局势建模预测（Shachtman 2011）等。即使是去杂货铺或加油站［我们在日常活动场所的消费信息会被收集分析，（商家）试图明白他们的消费者都是什么样的人，想要什么样的产品（Duhigg 2012）］这样的事情也能涉及预测建模，而我们很多时候甚至没有觉察到。预测模型在我们的生活中无孔不入。

虽然预测模型指引我们制造更令人满意的产品、发现更好的医疗手段以及进行更有回报的投资，但它也时常会给我们不准确的预测结果和错误的答案。例如，我们中大多数人都有因为预测模型错误地屏蔽垃圾邮件而错过一封重要邮件的经历（邮件过滤器）。类似的例子还包括，医学预测模型错误地诊断疾病，金融预测模型错误地提示人买入卖出股票而造成损失。最后这个金融预测模型出错的例子在 2010 年影响了很多股民。热衷于股票的人应该熟悉 2010 年 5 月 6 日的“闪电崩盘”。那时股市急速下降 600 点，然后急速回升至原来的点数。经过几个月的调查，日用品期货贸易委员会和证券交易管理委员会将此次崩盘的原因归结为算法模型的错误（U. S. Commodity Futures Trading Commission and U. S. Securities & Exchange Commission 2010）。

基于这次的闪电崩盘以及其他的一些预测模型失误的案例，Rodriguez（2011）写道，“预测建模，即建立或者选择能最好预测将来事件发生概率的模型这一过程，已经丧失其可信度了。”他认为预测模型时常失效是由于它们没有捕捉到一些复杂的变量，例如人类行为。事实上，我们预测或者做决定的能力受限于我们当前和过往的知识，同时也受我们没能考虑到的一些因素影响。这些现实情况是所有模型的限制，但是这并不能让我们停止优化过程以及建立更好的模型。

接下来的章节讨论一系列导致预测模型失效的公认因素。这些公认因素包括：（1）没有对数据进行充分的预处理，（2）没有充分验证模型，（3）未被证实的推论（例如，将模型应用到其已知范围以外的数据上），或者，更重要地，（4）过度拟合模型。此外，建模者在寻找变量间预测关系的时候通常只探索了少量的模型。这通常由于建模者对小部分专业技术的偏好，或者缺乏能让他们尝试更多不同技术的软件。

本书致力于帮助建模者开发可靠、可信的预测模型，循序渐进地引导建立模型的整个过程，对林林总总的常用模型进行直观的讲解。本书意在介绍：

- 建立预测模型的基本原则
- 直观讲解各种常用的解决分类以及回归问题的预测模型方法
- 验证预测模型的方法和步骤

- 实施建模和验证预测模型的关键步骤的程序代码

为了阐明这些原则和方法，本书将会列举许多不同的实际案例，涵盖范围从金融到医药，具体内容将在 1.4 节里进行介绍。在描述实际数据之前，让我们先了解一个阻碍所有预测建模技术的现实状况：预测和解释之间的取舍。

1.1 预测与解释

对于之前提到的例子，似乎存在历史数据可以被用来建立数学工具以预测未知的将来。此外，这些例子最重要的目的不是了解为什么某事情会发生或者不发生。我们的首要兴趣在于准确预测某事发生或者不发生的概率。注意，这类模型着眼于优化预测的准确度。例如，我们并不关心邮件过滤器为什么认为某封邮件是垃圾邮件。相反，我们只在意它是不是能准确过滤掉垃圾邮件，同时让我们想要看到的邮件顺利地到达收件箱里。又例如，如果我要卖一所房子，我首要关注的不是房产交易网站（例如 zillow.com）如何对我的房子估价。相反，我对 zillow.com 是否能给我的房产正确地估价更感兴趣。过低的估价会导致低的竞标和销售价格，而过高的估值可能让潜在的买家望而却步。

预测和解释之间的微妙平衡同样体现在医药领域。例如，一个癌症患者和其医生正考虑改变治疗的方式。医生和病人需要考虑许多因素，如服药的时间安排，可能的副作用以及存活率等。然而，如果之前有足够的病人已经采用过这个候选的疗法，那么我们可以收集关于这些病人的疾病、疗程以及人口构成的数据。同时，病人的基因或者其他生物学方面的数据（如蛋白质测量）可以通过实验室检测得到。基于这些病人的状况，我们可以建模预测这种候选治疗方式的结果。对于医生和病人来说，关键的问题是要知道更换治疗方式会对病人有**怎样**的影响。最重要的是，这个预测需要相当准确。如果建模用于这样的预测，它不应该被模型的可解释性所限制。或许有强大的舆论认为这可能不道德，但只要模型预测的结果能被恰当地验证，模型本身是个黑匣子还是一个容易解释的模型无关紧要。

预测模型的首要目的在于得到精确的预测，其次是解释模型并且明白为什么模型管用。但是，当我们努力建立更加精确的模型时，模型通常会变得更加复杂和难以解释。当预测准确率是我们的首要目标时，几乎总要面对这样的取舍。

1.2 预测模型的关键部分

目前为止人们谈论的许多实例表明，为了解决研究中的问题，收集数据（事实上非常大的数据）是相对容易的。此外，免费或者相对廉价的建模软件，如 JMP、WEKA 以及众多 R 包，加上功能越来越强的个人电脑，将使得预测建模的门槛降低。任何人只要懂一些计算技术，即可开始建立预测模型。但与此同时，如 Rodriguez（2011）精准指出的那样，模型的可信度被削弱了，特别是当获取数据和分析工具的途径大大拓宽时。

我们之后将会在书中反复谈到，如果预测信号存在于一个数据集当中，那么对于很多模型，无论它们依赖于什么样的技术或者假设，都或多或少可以捕捉到一些信号。不假思索地套用模型在某种程度上也能产生效果，俗语说："瞎猫碰到死耗子。"但能否建立最优的预测模型根本上是受建模者的专业知识以及要解决问题所处的实际情况制约的。专业知识首先应该用于获取与研究问题**相关的**数据上。尽管大量的数据库信息可以作为预测模型

建立的基础，但其中无关的信息（噪声）会减弱许多模型的预测能力。有针对性的知识能够帮助将可能有意义的信息从无关信息中分离出来，即在去除有害的噪声的同时加强潜在有用的信号。我们不想要的混淆信号也可能存在于数据中，并且这些信号只有运用专业知识才能分离出来。如下是一个关于混淆信号且需要专业知识去处理的极端例子。美国食品药品监管局不良反应报告系统数据库中存储了上百万种药品以及相应副作用的信息。这个数据集普遍存在明显的偏差。例如，当搜索治疗恶心的药物时，可能会发现大部分使用过该药物的患者患有白血病。想当然的分析或许会表明白血病是该药物的潜在副作用，但更可能的解释是，患者之所以服用治疗恶心的药物，是因为他们想缓解癌症治疗过程中产生的副作用。这或许显而易见，但重点在于，大量数据的可获取性并不能防止数据的滥用。

Ayres（2007）着力研究了专家意见和经验的相互作用，基于数据的模型给出两个结论，它们肯定了我们对有针对性专业知识的需求。第一，

"最终［预测建模］不能够代替直觉，更确切地说是一个补充。"

简单地只是选取数据模型或专家意见这两者中的一个，都不如把这两者结合在一起更好。第二，

"传统的专家在得知统计预测结果之后可以做更好的决定。那些依附传统行业专家权威的人倾向于支持综合两种'知识'的观点，即向行业专家提供'统计支持'……人们在知晓了统计分析预测结果之后通常可以给出更好的预测。"

在某些情况下，如垃圾邮件检测，让计算机做大部分的"思考"工作是可以接受的。当事情可能的后果更加严重时，如预测患者的反应，综合性的方法常常会有更好的效果。

总而言之，有效的预测模型与**直觉**和**对要解决问题所处的实际情况的深刻理解**是分不开的。这些方面的结合对于进一步改进模型极其关键。这个过程从获取**相关的**数据开始，这又是一个关键点。第三个要点在于有一个**多功能**的计算工具箱，它包含了预处理数据和数据可视化的各种工具，此外还有一套针对不同情形的建模工具。表 1-1 列举了一些情形。

1.3 专业术语

如前所述，有很多名词用来指代从数据中挖掘出变量相互关系信息并预测某个因变量的过程，而"预测建模"只是诸多称呼中的一个。由于多个科学领域都对该主题（预测建模）有贡献，因此下面列举了不同科学领域用来描述同一个概念的同义词：

- **样本**、**数据点**、**观测值**或者**实例**，这些术语指代一个单独的独立数据单元，如一个消费者，一位患者，或者一个群体。**样本**这个词同时也可指代一个数据点集合的子集，如训练集样本。词语的具体意义可以通过上下文推断得到。
- **训练集**是用于建立模型的数据，而**测试集**（或**验证集**）则用来评估最终模型的效果。
- **预测变量**、**自变量**、**属性**或**描述量**是预测公式中输入的变量。
- **结果变量**、**因变量**、**目标变量**、**类**或**响应变量**，用来指代事件的结果或我们要预测的那个量。
- **连续**数据具有天然的小数位数。血压、某物品的价格或者浴室的数目都是连续的。

在最后的这个例子中，虽然计数不能是小数，但我们仍然把其当作连续数据。

- **分类数据**，也叫**定类数据**、**属性数据**或**离散数据**，只能取一些特定的值，且不具有小数位数。信用等级（“优”或“差”）和颜色（“红”、“蓝”等）都是分类数据。
- **模型建立**、**模型训练**以及**参数估计**都指代通过数据得到模型方程的过程。

1.4 实例数据集和典型数据场景

在之后的章节里，我们将用案例分析阐明书中介绍的不同建模方法。在进入正题之前，首先简要考察一些预测建模的例子及其使用的数据集。本小节着眼于模型问题和相应数据类型的多样性。一些实例数据源于机器学习竞赛，这些比赛会提供一个现实需要解决的问题，并且对提出最优解决方案的参赛者进行奖励（常是金钱奖励）。这样的比赛在预测建模领域中有很长的历史，并极大地促进了这一领域的发展。

音乐流派

这是一个发布在 TunedIT 网站（http://tunedit.org/challenge/music-retrieval/genres）上的竞赛数据集。此次比赛的题目是建立预测模型将不同流派的音乐分为六大类。数据包括 12 495个音乐样本，对应 191 个特征变量。响应变量的类型分布并不平衡（图 1-1），其中最小的类是重金属音乐（占 7%），最大的类是古典音乐（占 28%）。所有的预测变量都是连续的：许多变量之间高度相关，并且它们具有不同的测量标度。整个数据集来自 60 个演奏者，每个演奏者提供 15～20 首乐曲。之后从每首乐曲中摘取 20 个片段进行参数化，从而得到最终的数据集。[⊖]因此，这些样本本身就彼此不独立。[⊜]

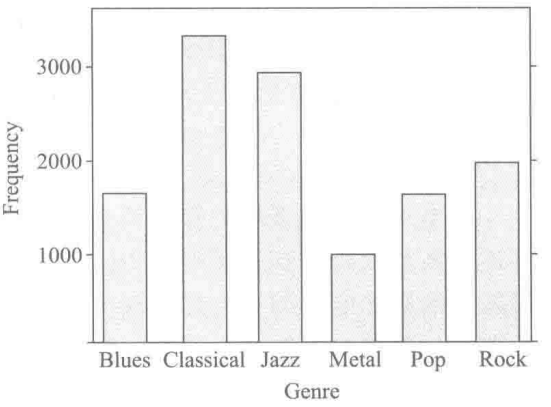


图 1-1 音乐流派频数分布

经费申请

这是发布在 Kaggle 网站（http://www.kaggle.com）上的竞赛数据集，比赛题目是对于经费申请成功率建立预测模型。数据库中包含了墨尔本大学 2009 年和 2010 年的 8707 条经费申请记录，共 249 个预测变量。经费申请状况（“成功”或者“失败”）是响应变量，且分布较为均衡（46%的成功比例）。网站注明目前澳大利亚经费申请成功率是小于 25%的，因此该历史数据并不代表澳大利亚的整体情形。预测变量中有测量值也有分类变量，如赞助方 ID、经费类别、经费额度范围、科研领域和院系等，这其中包括了连续变量、计数变量以及分类变量。关于这个数据集另外需要注意的是预测变量中有很多缺失值

⊖ 这里参数化指的是对于每个音乐片段添加相应的特征变量列。——译者注

⊜ 因为有些音乐片段来自相同的演奏者。——译者注

(比例为83%)。此外，样本彼此之间并不独立，因为一个申请人可能在不同时间多次申请。这些数据会在书中反复用来阐明不同的分类模型。

在第12章和第15章中，我们会大量地使用这个数据集。12.1节会对这个数据集进行更多细节上的梳理汇总。

肝损伤

这是来自制药行业的数据，建模的目标是预测不同化合物导致肝损伤的概率。数据集包括281种化合物，每种化合物对应376个预测变量。响应变量是分类变量（“不会导致损伤”、“轻微损伤”或者“严重损伤”），并且分布极度不均衡（图1-2）。这种类型的响应变量常出现于制药行业中，因为制药企业会避免合成那些有不利于健康的化合物。因此，表现良好的化合物的数目通常远远超过那些不良的化合物。376个预测变量中包括了184种生物扫描测量值和192个化学特征预测变量。生物学预测变量表示每次扫描得到的活跃度观测，观测值区间从0到10，众数为4。化学特征预测变量表示重要的亚结构的数目，以及被认为和肝损伤相关的生理特征测量。第5章里有对数据更详细的介绍。

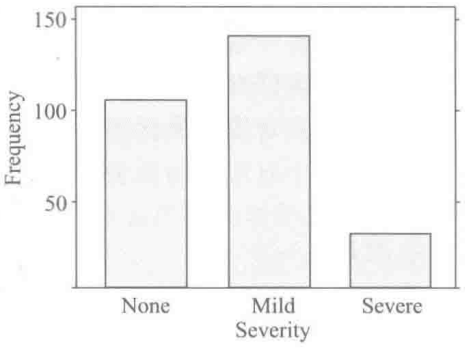


图1-2 肝损伤类型分布

渗透性

这个来自制药领域的数据集是用来建立模型以预测不同化合物渗透性的。简单地说，渗透性衡量化合物穿透膜的能力。例如在人体中，躯干和大脑之间有显著的膜，称做血脑屏障，而肠和躯干间也有膜阻隔。这些膜帮助身体关键部分抵御有害的物质。然而，口服的药物如果想对大脑产生效果，就首先得穿过肠壁，然后通过血脑屏障才能够到达神经目标。因此，研究化合物渗透生物膜的能力在药物研发的初期阶段是极其关键的。那些在科研遴选实验中表现出对治疗某疾病有效的化合物，如果其渗透性很差，则还需要改良提高渗透性，以使其能够到达目标部位。对渗透性识别问题的研究有助于指引化学家发现更好的分子结构。

渗透性分析如PAMPA和Caco-2可被用来测量化合物的渗透性（Kansy等1998）。这些分析能够有效地量化化合物的渗透性，但是实验的人力成本昂贵。在有大量的已检测化合物样本的条件下，我们可以建立渗透性的预测模型，希许能降低对实验的需求。这个项目包括了165种不同的化合物，每种化合物有1107个分子指纹。分子指纹指的是一系列二进制数值，用来表示该分子中是否存在某特定的分子亚结构。响应变量是高度偏态的（图1-3），同时预测变量非常稀疏（15%的稠密度），而且很多变量之间高度相关。

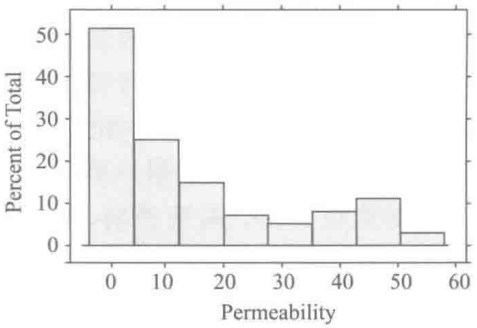


图1-3 渗透性分布

化学药品生产过程

该数据集包含了一项化学药品生产过程的信息，其目标是理解生产过程和最终生产率之间的关系。该生产过程的原材料将通过27个步骤得到最后的医药产品。最初的原料从生物个体中获取，数据中有关于其质量与特征的描述。项目的目标是预测生产过程的生产率。数据集由177个生物材料样本组成，每个样本有57个观测的特征。在这些特征中，12个来自最初的生物原材料，45个是在生产过程中测量的。生产过程中的这部分变量包括了温度、干燥时间、清洗时间以及某些步骤中副产品的浓度。其中一些步骤中的观测量能够被很好地控制，而另外一些则不行。预测变量包含了连续变量、计数变量和分类变量，其中一些是彼此相关的，一些有缺失值。样本之间并不独立，因为部分样本来自相同的初始生物原材料。

财务舞弊

Fanning和Cogger（1998）描述了一个用数据预测上市公司管理层财务舞弊的例子。利用公开的数据资源（如美国证监会的资料档案），作者检测到了102起财务舞弊案例。由于财务舞弊在整体数据中占比重较小，他们随机抽取了相同数量的没有财务舞弊的公司，[⊖]这些样本的抽取基于对一些重要因素的控制（如公司规模和行业性质）。这组数据中，150个样本点被用来当作训练集，剩下的54个作为测试集用来评估模型。

作者的分析从一些数目不确定的预测变量着手。这些变量从一些关键的领域获得，如行政主管的流失率、诉讼以及债务结构等。最终，他们的模型包含20个预测变量，诸如应收账款和销售额的比值，存货和销售额的比值，以及不同年份间的毛利变化。其中很多比值形式的预测变量有相同的分母（如应收账款和销售额的比值，存货和销售额的比值）。尽管真实的数据点没有公开，但这些预测变量之间很可能是高度相关的。

从建模的角度看，这个例子令人感兴趣的原因有如下几个方面。首先，现实中两个类（有无财务舞弊）的分布极不均衡，因此建模数据中两个类的频数分布[⊜]与实际要预测的目标群体的频数分布情况非常不同。这是一种用来减小数据不均衡影响的常见策略，有时称为“降采样”。其次，和样本量相比，预测变量的数目很多。这种情况下，预测变量选择是一项非常精细的任务，因为只有很少的样本，却需要利用它们来选择变量、建立模型以及评估模型。之后的章节会讨论过度拟合的问题，在这些问题中训练样本中的趋势在同样群体的其他样本中并不存在。当自变量多、样本量小的时候，可能存在的一种风险是从现有数据中找到的一个有效的预测变量却在其他数据中不可再现。

数据集比较

这些例子展示的数据特征普遍存在于在很多其他现实数据集中。首先，响应变量可能是连续的或是分类的，而分类变量可能有超过两个类别。对于连续型响应变量，分布可能是对称的（如化学药品制造的例子），也可能是有偏的（如渗透性的例子）；对于分类型响应变量，分布可能均衡（如经费申请的例子）也可能不均衡（如音乐流派和肝损伤的例

⊖ 这种抽样方式与医学领域中的病例对照研究极为相似。

⊜ 建模时进行了随机抽样，使两个类别数目相等。——译者注