HZ BOOKS 华章IT
数据分析与决策技术丛书
[PACKT] PUBLISHING

Practical Predictive Analytics

实用预测分析

[美] 拉尔夫·温特斯（Ralph Winters）著
刘江一 陈瑶 刘旭斌 译

以真实案例驱动，详细讲解使用R、Spark等开源工具
进行预测分析的实用方法和技巧

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

Practical Predictive Analytics

实用预测分析

[美] 拉尔夫·温特斯（Ralph Winters）著

刘江一 陈瑶 刘旭斌 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

实用预测分析 /（美）拉尔夫·温特斯（Ralph Winters）著；刘江一，陈瑶，刘旭斌译．
—北京：机械工业出版社，2018.7
（数据分析与决策技术丛书）
书名原文：Practical Predictive Analytics

ISBN 978-7-111-60335-1

I. 实… II. ①拉… ②刘… ③陈… ④刘… III. 决策预测 IV. C934

中国版本图书馆 CIP 数据核字（2018）第 143529 号

本书版权登记号：图字 01-2017-7517

实用预测分析

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：唐晓琳 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2018 年 7 月第 1 版第 1 次印
开 本：186mm × 240mm 1/16 印 张：24.5
书 号：ISBN 978-7-111-60335-1 定 价：89.00 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

The Translator's Words 译 者 序

接触本书之前，我们刚刚完成了另外一本书《 Thoughtful Machine Learning with Python 》的翻译工作——那是一本非常适合机器学习入门的图书，也是该领域中的经典之一。而讯速决定开展对本书的翻译，自然也是因为对其喜爱有加：

第一，预测分析是机器学习中非常有应用价值的一个子领域；

第二，本书相当适合作为一本进阶的教材，能帮助读者对机器学习在真实世界的应用有直观的、详细的认识；

第三，可以借此机会熟悉一门在机器学习和统计学领域广受欢迎的编程语言：R 语言。

可能很多读者被本书吸引，也是出于类似的原因吧。

这里谈谈 R 语言。国内很多读者对 R 语言还不是很熟悉，但 R 语言在国外高校的统计系是一门必修的课程。R 语言在部分运行环境中是开源的，这使它具有很强的生命力，其功能也日益丰富、强大、稳定。安装 R 语言本身所使用的资源很少，而且对不同操作系统的兼容性令人满意。可以用它方便地对数据进行必要的处理，并绘制出漂亮的图形以供深入观察分析。在项目初期选用 R 语言作为建模语言，数据接口的兼容性较高，能够快速搭建模型，并且和传统的统计型语言相比，可移植性较高，对机器学习模型的可扩展支持 Package 的资源也非常丰富。值得注意的是，从语言开发产品的能力来看，C 语言和 Java 语言的商业可扩展性较高。例如商业化集成使用 R 语言进行大数据建模分析，主流服务器端的 R 语言环境多是基于 Microsoft R Server, 其他基于 Linux 服务器的 R 语言环境多由 R 语言 IDE 开发商来定制化支持。总结而言，R 语言能够快速探索、搭建初期的模型、原型，可以称其为学术派语言，值得期待的是，R 语言正在向商业化语言渐渐迈进。

有人说：“R 固然好用，但学起来却头疼无比！”放心，已经有人用 R 编写好了丰富的示例代码，并详加解释，让你知道为何要这么做、为何不选另一种方法，而你还有哪些其他选择等。没错，这些示例在本书中随处可见。而且作者还会贴心地反复提醒读者注意避免某些

错误，其重视程度，让人禁不住猜测，作者本人是否也是在各种错误中摸爬滚打，才练成了今天的段位……

还有人说："数据量一大，R 就慢得像爬行一样。"经验丰富的作者当然不会忘记为你提供趁手的解决方案，比如 SparkR、抽样等。在本书的多个示例中，数据量较小的示例用于演示算法的基本原理，使用基本 R 足够。数据量大的示例中会展示何时需要从基本 R 转换到 SparkR，高效地完成处理和抽样，再转换回基本 R，开始绘制图形等 R 擅长的任务。

本书对算法的解释简练而形象，但它本质上仍是一本偏重动手操作类的书籍。本书的目的是通过真实的数据绘制出各种对比图形，让你真真切切地感受到预测分析项目是如何实现的，并会指导人们做出判断和行动——有时会令人莫名激动，恨不得马上找到真实数据集来动手试一试，看自己能否利用强大的预测分析能力去解释世界、影响世界。

以上只是我们觉得本书对读者帮助较大的地方，本书当然不止这一两项优点，它还有很多精彩等待你去发现。

在本书的翻译过程中，陈瑶翻译了第 1 章（部分）、第 4 章、第 7 章和第 12 章，刘旭斌翻译了第 2 章、第 5 章、第 8 章和第 11 章，刘江一翻译了前言、第 1 章（部分）、第 3 章、第 6 章、第 9 章和第 10 章。

感谢诸位译者在百忙之中挤出时间完成了这项有趣的工程！

感谢机械工业出版社华章公司的编辑在翻译过程中给予的悉心帮助和指导！

刘江一

参与本书翻译的初衷，是因为当时负责的有关性别预测分析 (Gender Analysis) 和情感倾向分析（Sentiment Analysis）的项目，在初期选用了 R 语言作为建模语言，数据接口的兼容性较高，能够快速搭建模型，并且和传统的统计型语言相比，可移植性较高，对机器学习模型的可扩展支持 Package 的资源也非常丰富。值得注意的是，从语言开发产品的能力来看，C 语言和 Java 语言的商业可扩展性较高。例如商业化集成使用 R 语言进行大数据建模分析，主流服务器端的 R 语言环境，多是基于 Microsoft R Server, 其他基于 Linux 服务器的 R 语言环境多由 R 语言 IDE 开发商来定制化支持。

伴随着项目的进行，翻译完本书，总结而言，R 语言能够快速探索、搭建初期的模型、原型，可以称其为学术派语言，值得期待的是，R 语言正在向商业化语言渐渐迈进。

陈　瑶

About the Auther 关于作者

Ralph Winters 的职业生涯始于在一个音乐表演权利组织担任数据库研究人员（他甚至会作曲），继而延伸到医疗调查研究，最后落脚于分析和信息技术领域。他已经给很多名列世界500强的大企业提供过自己在统计和分析方面的经验，包括金融、直销、保险、医疗和制药领域的企业。他的工作涉及很多不同类型的预测分析项目，包括客户保留、反洗钱、客户之声文本挖掘分析，以及医疗风险和客户选择模型。

他如今在一家医疗服务公司担任数据架构师，在数据和高级分析组工作。他很喜欢与一个拥有业务分析师、技术专家、保险精算师及其他数据科学家的智囊团协同合作。

Ralph 认为自己是个务实的人。除了为 Packt 出版社写作了《Practical Predictive Analytics》之外，他还参与写作了另外两本著作，即 2014 年 9 月 Elsevier 出版的《 Practical Predictive Analytics and Decisioning Systems for Medicine》(Miner 等人著)，以及 2013 年在马萨诸塞州剑桥第 11 届年度文本和社会分析峰会上发表的《 Practical Text Mining with SQL using Relational Databases 》。

Ralph 和他挚爱的妻子 Katherine、迷人的女儿 Clair 与 Anna 居住在新泽西州，Ralph 的个人网站是 ralphwinters.com。

关于审校者 *About the Reviewers*

Armando Fandango 在 REAL 公司担任首席技术官，开发基于 AI 的产品和平台，用于在品牌、代理、出版商和读者之间生成智能的连接。Armando 创立了 NeuraSights，目标是使用神经网络和机器学习从大数据和小数据中发掘洞见。在此之前，他还担任过 Epic 工程咨询集团有限公司的首席数据科学家和首席技术官，曾经与政府部门和大型个人组织合作开发智能产品，涉及机器学习、大数据工程、企业数据仓库和企业仪表板。Armando 曾经在 Sonobi 公司担任数据主管，领导若干个数据科学与工程团队，为 Sonobi 的 AdTech 平台 JetStream 推动大数据和预测分析技术及策略。Armando 曾经在中佛罗里达大学的高级计算研究中心管理高性能计算（HPC）的咨询和基础建设。Armando 还曾经为高科技初创公司 QuantFarm、Cortxia Foundation 和 Studyrite 做过顾问团成员及 AI 专家。Armando 的著作包括一本名为《Python Data Analysis》(第 2 版）的书，以及在国际期刊和会议上发表的研究论文。

Alberto Boschetti 是一位数据科学家，在信号处理和统计学方面有丰富的经验。他拥有电信工程博士学位，现在居住于伦敦。在他工作的项目中，日常面对的挑战涉及自然语言处理（NLP）、机器学习以及分布式处理。他对工作极具热忱，持续跟进数据科学技术的最新进展，参加小组讨论、会议以及其他活动。他的著作有《Python Data Science Essentials》《Regression Analysis with Python》和《Large Scale Machine Learning with Python》，全部由 Packt 出版。

Preface 前　言

这是另一类关于预测分析的书。我写这本书的初衷是为传统分析人员介绍一些使用开放源码工具的预测分析技术。

不过，我很快意识到，传统分析工具的某些特性可以使新一代数据科学家受益。我曾经在企业数据解决方案方面做了大量工作，我很有兴趣撰写一些不同类型的主题，如分析方法、敏捷、元数据、SQL 分析和可重复的研究，这些研究在一些数据科学 / 预测分析书中经常被忽略，但对分析项目的成功是至关重要的。

我还想写一些很少被提及的分析技术，这些技术超出了标准回归和分类任务的范围，例如使用生存分析来预测客户流失，使用购物篮分析作为推荐引擎。

由于基于云计算的解决方案已经有了很大的进展，我认为增加一些关于云分析（大数据）的内容很重要，所以我加入了一些在 Spark 环境中开发预测分析解决方案的章节。

本书的重点之一是触类旁通，我希望无论你的技术方向是什么，也无论你如何理解数据科学、预测分析、大数据，甚至是诸如预测这样的术语，都可以在这里找到适合自己需求的内容。

此外，作为数据科学团队的一部分，我要向领域专家们致敬。通常情况下，这些精通领域业务知识的分析师没有耀眼的头衔，但他们对于分析项目的成功至关重要。希望我讨论的一些话题能打动他们的心弦，让他们对预测分析的一些技术概念更感兴趣。

当 Packt 邀请我写一本关于预测分析的书时，我首先想到的是寻找一种优秀的开源语言，来弥合传统分析与当今数据科学家之间的鸿沟。我认真地考虑过这个问题，是因为每种语言在如何表达问题的解决方案方面都有细微的差别。然而，我决定最终不在意那些细节，因为预测分析这个概念不是依赖于任何一种编程语言的，而且编程语言的选择通常由个人偏好以及你所在的公司决定。

我最终选择了 R 语言，因为我的专业背景是统计学，我觉得 R 语言具有良好的统计学

严谨性，现在它不但已经和 SAS 等适合的软件做了合理的整合，而且还与关系数据库系统以及 Web 协议有很好的整合。它还具有出色的绘图和可视化系统，以及用户贡献的许多好用的软件包，涵盖了大部分的统计和预测分析功能。

关于统计数据，我建议你尽可能多地学习相关知识。了解统计数据可以帮助你区分优良的模型与糟糕的模型，并通过了解基本概念——如中心倾向度量（平均值、中位数、众数）、假设检验、p 值和效应大小——来帮助你识别不良数据中的许多问题。如果你了解数据统计，将不再仅仅以自动的方式运行封装好的软件，而是可以多少了解一些底层的运行机制。

R 语言的一个缺点是它在内存中处理数据，因此在单个 PC 上使用时，软件会限制数据集的大小，使之处理不了更大的数据集。对于本书中使用的数据集，在单个 PC 上运行 R 程序来处理应该没有问题。 如果你有兴趣分析大数据，本书将用几章的篇幅讨论在云环境中的 R 和 Spark，你可以在这些章中看到如何处理分布在许多不同计算机上的大型数据集。

谈到本书中使用的数据集，我不想使用那些你经常看到的、被人们反复分析的数据集。其中一些数据集的确非常适合用来演示技术，但我想要一些新的东西。然而，我没有看到多少我认为对本书有用的数据。有些数据来源不明，有些需要正式的使用许可，有些缺少好的数据字典。所以，在许多章节中，我最终使用 R 中的模拟技术生成自己的数据。我觉得这是一个不错的选择，因为借此机会我能够介绍一些可以在工作中使用的数据生成技术。

我使用的数据涵盖了广泛的范围，包括市场营销、零售和医疗保健应用。我本来希望能增加一些财务方面的预测分析用例，但时间不够用了。也许我会把这方面的内容留到另一本书中去讲！

本书主要内容

第 1 章从介绍预测分析的发展历史开始，然后讨论预测分析从业人员的一些不同角色，并描述他们从事的行业。接下来讨论在 PC 上组织预测分析项目的方法，介绍 R 语言，并以简短的预测模型为例结束该章。

第 2 章讨论如何将预测模型的开发过程组织成几个阶段，每个阶段都有不同的目标，如探索和问题定义，最后是预测模型的实际开发。该章讨论两种重要的分析方法：CRISP-DM 和 SEMMA。在该章中贯穿了一些示例代码，以展示一些方法的核心思想，希望你不会感到枯燥。

第 3 章介绍可以将自己的输入数据引入到 R 程序中的各种方法。该章还讨论使用标准 SQL 函数和 R dplyr 包的各种数据预处理方法。没有输入数据？没问题。该章将展示如何使用 R 语言的 wakefield 包生成你自己的模拟数据。

第 4 章从对有监督算法和无监督算法的讨论开始。该章的其余部分集中在回归算法，它是一种代表性的有监督算法。你将了解如何解释回归算法的输出，如模型系数和残差图。该章甚至提供一个交互式游戏，利用交互测试，看看你是否能够辨别一系列的残差是不是随机的。

第 5 章重点讨论另外三种广泛使用的核心预测算法，而且把它们与回归结合起来，可用于解决许多（可能是大部分）预测分析问题。该章讨论的最后一个算法（支持向量机（SVM））通常用于诸如非结构化文本之类的高维数据，因此示例代码将附带使用一些客户投诉评论的文本挖掘技术。

第 6 章讨论一种称为生存分析的具体建模技术，并展示一个假设的客户营销满意度和保留示例。我们还将深入研究利用 R 中的抽样功能模拟客户选择的方法。

第 7 章介绍关联规则和购物篮分析的概念，并介绍一些可以根据在线零售商店销售的各种数据组合来预测未来销售情况的技术。该章还会引入一些文本分析技术和一些聚类分析技术，用来将各种客户分为不同的分组。除此之外，你将学到一些数据清理技术，还可以知道如何生成一些有趣的关联图。

第 8 章介绍时间序列分析。首先探讨 CMS 网站的医疗保健注册资料。接下来定义一些基本的时间序列概念，如简单和指数移动平均线。最后，在示例代码中使用 R 的预测软件包，顾名思义，它可以帮助你执行一些时间序列预测。

第 9 章介绍 SparkR，它是使用 R 访问大型 Spark 聚类的环境，不需要安装本地版本的 R。该章还会引入 Databricks，一种用来针对基于 Spark 的大数据运行 R（以及 Python、SQL 等）的基于云的环境。该章还介绍使用 Pima Indians 糖尿病数据库作为参考，将小型数据集转换为更大的 Spark 聚类的技术。

第 10 章展示如何利用 SparkR 和 Spark SQL 的组合，使用加载到 Spark 中的 Pima Indians 糖尿病数据，执行一些探索性数据分析。我们将使用一些特定的 Spark 命令来了解 Spark 数据的基础知识，这些命令允许我们过滤、分组和汇总，并将 Spark 数据可视化。

第 11 章先介绍一个使用 Spark 聚类构建的逻辑回归模型，进而对机器学习进行阐述。我们将学习如何将 Spark 数据分解为训练数据和测试数据，运行逻辑回归模型，然后评估其性能。

第 12 章教你如何使用 Stop 和 Frisk 数据集在 Spark 中运行决策树模型。你将学习如何将一些聚类样本提取到本地计算机中，然后运行一些你已经熟悉的非 Spark 算法来克服 Spark MLlib 环境的一些算法限制。该章还将介绍一种新的基于规则的算法 OneR，并演示如何在 Spark 中混用不同的语言，例如在同一个笔记本中使用 % 魔法指令将 R、SQL，甚至 Python

代码混合在一起。

阅读本书你需要什么知识

这不是一本预测分析的入门书，也不是学习 R 或 Spark 的入门书。我们希望读者有一些基础的 R 数据操作技术的知识。预先获取一些预测分析的知识也是有用的。如前所述，了解假设检验、相关性、平均值、标准偏差和 p 值等基本统计概念也有助于你阅读本书。

本书的读者对象

本书适用于已经接触过 R，并且正在寻求学习如何开发企业预测分析解决方案的读者。此外，如果传统的业务分析师和经理希望扩展一些使用开源 R 程序进行预测分析的技能，可能会发现这本书很有用。了解其他编程语言、目前正在进行预测分析实践，或希望使用 Spark 了解预测分析的读者，也将会从有关 Spark 和 R 的章节中获益。

下载示例代码和彩图

在 GitHub 上提供了本书的代码，网址是：https://github.com/PacktPublishing/Practical-Predictive-Analytics。

我们还为读者提供了一个 PDF 文件，其中包含本书中使用的截图 / 图表的彩图。彩图将帮助你更好地了解输出数据中的变化。可以从以下网址下载此文件：https://www.packtpub.com/sites/default/files/downloads/PracticalPredictiveAnalytics_ColorImages.pdf。

Contents 目　录