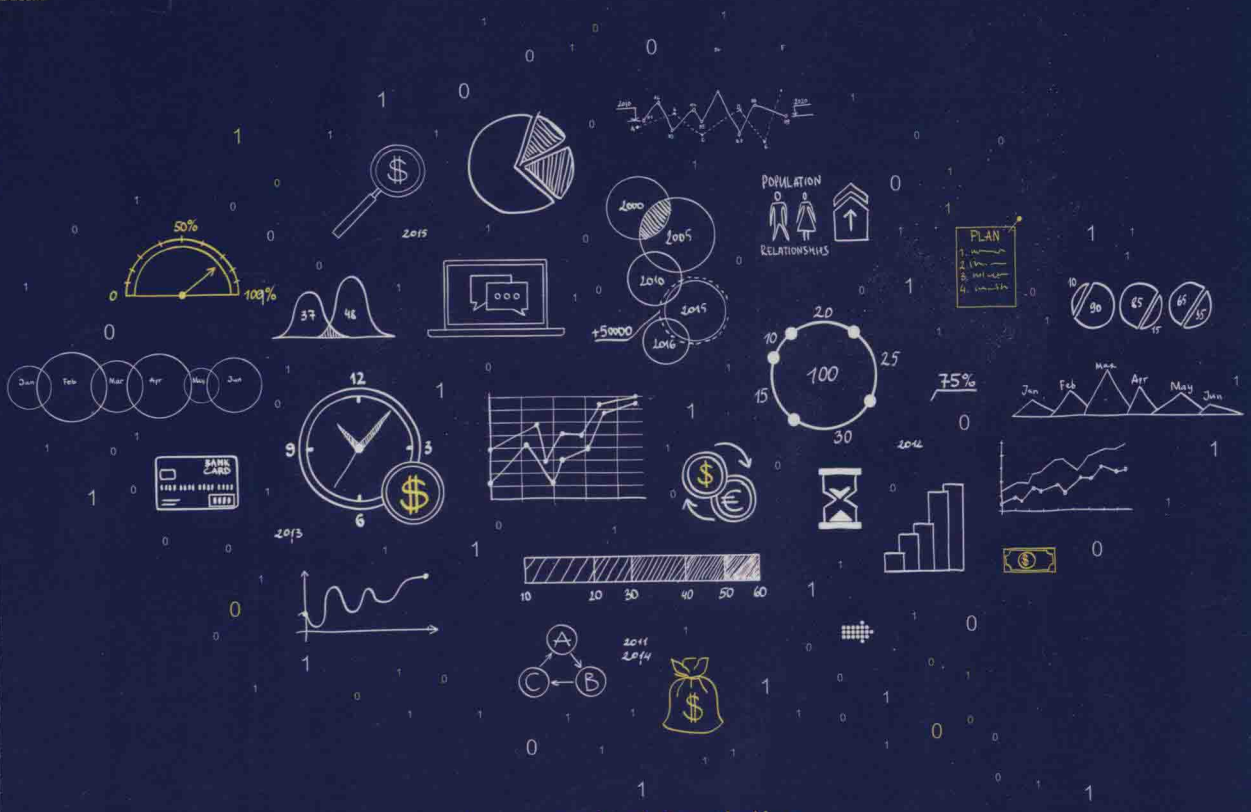
数据分析与决策技术丛书



Augmented Analytics

AI Driven Data Analytics, Business Decisions and Case Practice

增强型分析

AI驱动的数据分析、业务决策与案例实践

彭鸿涛 张宗耀 聂磊 ◎ 著

德勤首席数据科学家、前华为资深数据科学家、前 IBM 资深数据科学家联合撰写

从技术和业务双重维度讲解如何利用前沿的大数据和 AI 技术实现智能数据分析与业务决策

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

增强型分析

AI驱动的数据分析、业务决策与案例实践

彭鸿涛 张宗耀 聂磊 ◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

增强型分析：AI 驱动的数据分析、业务决策与案例实践 / 彭鸿涛，张宗耀，聂磊著 . —北京：机械工业出版社，2019.8
（数据分析与决策技术丛书）

ISBN 978-7-111-63416-4

I. 增… II. ①彭… ②张… ③聂… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2019）第 172986 号

增强型分析
AI 驱动的数据分析、业务决策与案例实践

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：张锡鹏　　责任校对：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2019 年 9 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：17.25
书　　号：ISBN 978-7-111-63416-4　　定　　价：89.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　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

内容简介

增强型分析是数据科学的未来，本书讲解了如何通过前沿的大数据技术和AI技术实现智能的数据分析和业务决策，即增强型分析。

本书的三位作者是来自德勤、前华为和前IBM的资深数据科学家，在大数据和AI领域至少都有10年以上的工作经验，他们将各自多年来在“构建数据挖掘模型，解决实际业务问题”方面积累的经验全部总结于本书中。

全书的内容由两条主线贯穿。

技术主线：一方面讲解了预测模型、序列分析、预测分析、Prescriptive分析等前沿数据处理技术，一方面讲解了CNN、RNN和GNN等前沿的AI技术如何为数据分析赋能。

业务主线：在数字化转型的大时代背景下，如何通过数据分析实现智慧营销、智慧风险管控，实现由初级的“主动营销”到“被动营销”，再到“全渠道协同营销”等营销手段的升级应用。

本书的重点聚焦在本质内容上，即数据处理、算法及模型、“模型洞见到业务决策”的分析等。

全书共8章。

第1章：作者结合自己的从业经验介绍了数据科学家的职业生涯发展、工作模式和工作方法要点等内容，为有志成为数据科学家的读者指明了道路和方向；

第2章：从描述性分析的角度讲解了数据探索、数据预处理衍生指标加工方面的技巧；

第3章：介绍了预测类模型构建时的新方法、新思路、新工具；

第4章：讲解了序列分析的相关内容，包括序列模式、序列规则、序列预测等的挖掘与应用，用实例的方式说明了算法的原理、特点和使用技巧；

第5章：介绍了人工智能下一个阶段的重点领域，即如何应用数据分析做出最优决策；

第6~8章：通过与传统模型的对比，介绍了CNN、RNN、GNN等算法的原理，通过大量的实例说明了这些AI技术在数据分析与决策领域的用法和实际效果。

作者简介

彭鸿涛

德勤企业咨询总监兼首席数据科学家，德勤全球AI团队核心成员，德勤数字化转型、智慧营销、智慧风控、客户体验等核心咨询服务方案的资深顾问。

2008年加入SPSS并与跨国团队一起进行Analytical Decision Management决策自动化工具的开发，与国内外团队一起构建了SPSS在不同应用领域的解决方案，其中某些方案现已成长为IBM的知名解决方案；2014年加入IBM GBSC部门，领导数据分析团队，针对不同客户设计和实施数据分析的方案；2016年加入IBM GBS GBS Cognitive Business Decision Support担任CTO和首席数据科学家；2017年加入德勤企业咨询担任金融服务总监及首席数据科学家，领导团队开拓数字化转型背景下的新型咨询服务方案，期间高质量交付大型银行的数字化转型及实施相关项目并得到客户高度认可。

张宗耀

上海全应科技有限公司资深数据科学家，前华为企业智能部门资深数据科学家，前IBM SPSS 算法组件团队资深算法工程师。

2009年加入IBM SPSS算法组件团队，负责Statistic和Modeler产品的升级和维护；2012年开始大数据算法组件的设计和开发，为分布式分析引擎提供了核心计算单元，主导完成开发了分布式平台下的广义线性模型、自动建模算法、ADMM优化算法等，打造了分析引擎平台以及SPSS Modeler产品的最具竞争力算法模块；2015年开始投入Spark分布式框架的算法设计和开发，主导开发了生存分析算法、时间序列相关算法等。2016年加入华为，先后就职于华为的数据挖掘团队，以及企业智能部门的机器学习服务团队和工业解决方案团队，负责算法、机器学习、实时预测、数据分析，以及行业解决方案的设计、开发和部署相关的工作。

聂磊

陕西万禾数字科技有限公司CTO，前IBM SPSS 资深数据科学家，前IBM Watson Analytics数据分析引擎技术主管及架构师。

2008 年加入IBM Analytical Decision Management团队，主导开发了业务规则引擎和基于优化技术的预测性维护解决方案；2014年加入IBM Watson Analytics团队，担任技术主管兼架构师，主导了IBM Watson Analytics数据分析引擎基于Spark技术的转换，极大提高了平台的计算能力；2017年担任IBM Cognos Analytics团队架构师，主持了自动化技术的引入。

需要全本请访问 www.ertongbook.com



华章圖書

一本打开的书，一扇开启的门，
通向科学殿堂的阶梯，托起一流人才的基石。

Forword 推荐序一

近年来，人工智能技术在应用领域已经有了比较大的发展，它正在逐步改变我们的生活，同时也在促进和推动企业的变革。以我熟悉的金融行业为例，人工智能预计会改变过去金融企业成功的基本要素，对于金融企业运营架构重塑、产品定制化、预测和决策，以及金融行业格局都有深远的影响。金融企业的高管们已经慢慢领略到了数据的威力和价值，很多银行家已开始将企业数字化转型提升到了企业战略的高度。

运用人工智能技术，可以使人类社会变得更美好。人们总是期待产品更适合、服务更贴心、生活更便利。在实践中，技术给企业赋能，企业通过优质的产品和服务满足社会，提升人类福祉。很多金融企业已经开始尝试向潜在客户推送更加精准的产品信息，通过智能投顾及产品交叉销售挖掘来满足客户多样化的潜在需求，开发各种人工智能助手协助客户获得更便利的服务体验。高德纳（Gartner）公司提出了客户体验的金字塔模型，如图 1 所示，根据客户是否需要、客户是否知道、产品服务触达情况细分了六个层次。我相信，在满足客户体验方面，还存在广泛的技术应用空间。

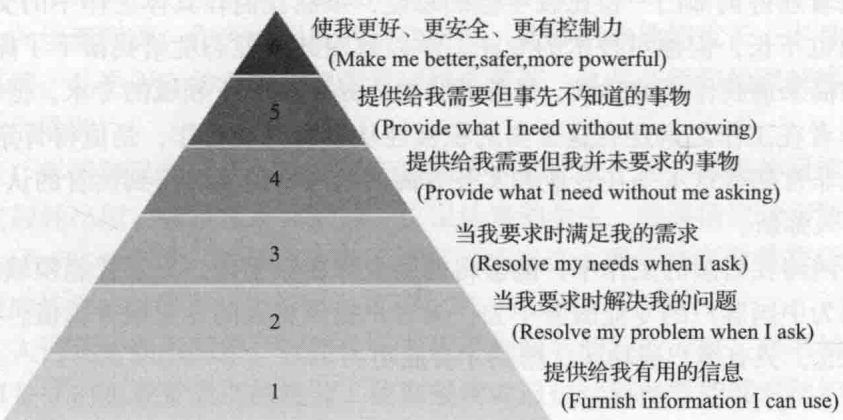


图 1　Gartner 客户体验金字塔模型[⊖]

最近一两年，中国人工智能技术应用市场逐渐走向成熟，企业已经认识到大数据平

⊖ https://www.gartner.com/en/newsroom/press-releases/2018-07-30-gartner-says-customer-experience-pyramid-drives-loyalty-satisfaction-and-advocacy。

台、人工智能平台等 IT 投入不能直接解决业务问题，还需要咨询及业务管理的广泛投入才能取得优化的管理和业务成果。作为一家具有百年历史的会计师事务所，德勤自身的变革从未中断。德勤紧盯市场需求，在人工智能领域投入非常大，而且取得了很大的成效。在各个业务条线，提供“业务咨询 + 大数据及人工智能技术的应用”已经逐渐成为德勤咨询新业务模式的常态。业务咨询与大数据及人工智能技术的结合，既解决了客户业务发展方向及策略问题，又通过具体的数据分析及智能应用让客户看到了实际的业务成果。

鸿涛凭借在数据分析领域十多年的工作经历，以及多年的咨询项目经验，坚持技术应聚焦于解决实际问题的理念，在实践中综合考虑技术、业务模式、方法路径和策略等核心要素，追求产生最佳的应用效果。他和其他两位作者在书中围绕业务分析的三个层次，即描述性分析、预测性分析及规范性分析，对数据科学家的成长路径、大数据探索和预测、预测模型信息技术、序列分析、决策分析以及三种人工智能技术等方面进行了深入阐述，相对全面地介绍了所涉及的典型算法、工具、业务问题的解决案例等。

金融行业的数字化转型浪潮衍生出了大量的技术应用场景。技术赋能的方法和经验是在实践中不断积累下来的宝贵财富。纯粹的技术型书籍往往侧重于原理及工具的介绍，较少包含“如何应用这些技术解决实际问题”的内容，而本书则围绕“技术介绍”与“业务应用”两条主线展开，还融合了作者在过往金融业服务项目中的各类“业务咨询 + 大数据及人工智能技术的应用”方案的心得与总结。相信这对于有志于开发应用大数据及人工智能技术并解决实际问题的读者会有非常大的参考价值。

我和彭鸿涛都在德勤中国工作，并且都专注于为金融行业客户提供专业服务。鸿涛是德勤风险管理咨询部门一位比较年轻的总监，虽然我们在具体工作中的交集并不多，认识的时间也不长，但他对技术的专注、学以致用的研究态度给我留下了深刻的印象。当他拿着书稿来请我作序的时候，虽然我自己不是数据分析领域的专家，但我想鸿涛和其他两位作者在工作之余还有这么高的积极性从事研究和写作，是值得肯定和鼓励的。同时，我也非常期待这本书在专业和实务方面的参考价值能够得到读者的认可并在时间的长河中历久弥新。

我期待鸿涛在后续的工作中，能够将德勤全球在数字化、人工智能领域的领先成果实际运用到为中国客户的专业服务中去，为客户提供更大的专业服务价值。德勤本身没有成功的概念，只有客户成功了，德勤才会成功。

是为序。

吴卫军

德勤中国副主席

2019 年于北京

Forword 推荐序二

从研究生毕业到现在超过十年的时间中，彭鸿涛和我联系并不多。但是我知道他一直在不断努力上进。研究生刚毕业时他在西安的富士通工作，主要做对日的外包开发工作。一年多后他去了知名的数据分析软件公司 SPSS，从事数据分析的软件开发以及应用数据挖掘技术解决实际问题等工作。IBM 收购了 SPSS 后，他一直在 IBM 工作，直到两年前又去了德勤。

彭鸿涛 2005 年在交大软件学院上学，当时学院刚刚成立。那时对他印象较深的事情有两个。一个是在上学期间如果要找他，去机房肯定能找到；还有一个就是他总能较快地完成复杂程序，譬如他可以很快完成实现视频目标自动跟踪的小软件。

在 2014 年他出版了第一本书《发现数据之美》。图书出版后，他专门来学校与我面谈。他当时强调，写书既是总结归纳的学习过程，也是期望能够借此突破工作内容，做一些更能解决实际问题的事情。我欣喜于他能不断前行，也对他在工作之余能够完成图书写作而感到高兴。

学校只是学生人生旅程中的一站。如同这美丽的校园中的植物，百花在四季中绽放，然后又是下一个轮回。一批批朝气蓬勃的菁华学生，在这里相聚、学习、成长，然后又奔赴下一个人生里程碑。作为校园里的常住居民，能够陪伴他们成长，本身就是一件非常快乐的事情。几乎所有的老师都期待看到桃李芳菲，这是老师们的成就感所在，也是老师们的人生乐事。

前几天，彭鸿涛又告诉我，他和张宗耀师弟一起完成了第二本书的写作，很快将由机械工业出版社出版。张宗耀是我的另一位很认真的学生，做事情比较细致。当年他一毕业就加入了 SPSS 公司，与彭鸿涛成了同事。经过这么多年在专业数据分析公司的淬炼，他已经成长为一位资深的机器学习算法专家。

他们二人请我给新书写序。我建议他们找更为知名的行业专家，对书的推广可能更有帮助，但是他们更看重老师的推荐，我期望读者通过这篇推荐能够对他们的背景多一点了解。

人工智能将深刻地改变我们的生活

象牙塔内教授和研究的内容偏向基础，象牙塔外的应用更看重实际价值。人工智能

从 20 世纪到现在已经经历了几轮研究、投资的起伏，最近几年人们的热情又空前高涨。这是因为从数据、计算能力、算法及研究、软件、人才等各个方面，都可以极大地促进人工智能的大发展、大应用，这当然也会极大地改变我们的生活。

移动互联网的出现已经深刻地影响了我们生活的方方面面，在 5G 和人工智能时代，会出现什么样的变化，让我们拭目以待。从技术应用的趋势来看，人工智能的应用场景势必会深刻切入到人们生活的方方面面。以智能家居为例，人工智能技术可以根据喜好、行为以及环境变化在我们无感的情况下调节好温度、湿度、灯光、音乐等环境氛围要素。

我们期待人工智能技术应用的涌现，也期望技术能够提升人们的生活品质。在 5G 时代，能够互联的信息会爆炸式地增长，新技术更多地应该帮助人们利用这些数据解决问题，避免沉迷于特定内容。

积极投身于这次人工智能的浪潮

人工智能已经是一个国家战略，其实施需要配套的人、财、物的不断投入。作为教书匠，我很高兴地看到，不论是国家层面，还是具体的细分行业，都已经形成了人工智能应用的浪潮。

不论是在校的学生还是从业者，在做人生规划或实践自身的人生道路时，都可以考虑下是否愿意在这次技术应用的大爆炸中发挥一些作用，做一点贡献。

国家的发展、社会的进步，教育是核心的基础工作，接受教育成为有用之人也是学生需要努力去做的事情。如果你是一位学生，在学习成长过程中，除了学习课程外，还需要“抬头看、向远处看”，问一下自己的内心，想成为一个什么样的人，然后努力去实践。

期望作者们能走得更远

我深信本书的三位作者付出了巨大的努力才使得本书能够出版。引用交通大学的校训，“精勤求学、敦笃励志、果毅力行、忠恕任事”，可以认为他们在成长中是不断践行着这些原则。

“路漫漫其修远兮”，我期望他们能够在工作中继续前行，努力创造出新的天地，到达新的高度！

朱利

西安交通大学软件学院副院长、教授

Preface 前　言

人工智能技术由于数据、算法、硬件支撑的计算能力等核心要素的共同发展，进入了广泛的、实质性的应用阶段。在不远的将来，我们肯定能看到人工智能及其相关技术在不同行业发挥巨大的价值。

增强型分析将会长足发展

多年以来，人们在构建模型时总是要花费大量的时间和精力在准备数据、数据预处理、多次尝试构建模型、模型验证等过程上。在工业发展的历程中，纯手工打造的时代势必要被标准化流水线的工厂取代，因为工序分解后可以按照统一的模式来处理。构造模型的过程也可以从纯手工打造时代发展为一个更加智能化的时代。笔者十年前在 SPSS 任职时，就深度参与了自动化建模相关组件的开发，即同一个模型可以按照不同的算法来实现并通过同一个评价指标筛选出最优模型。这样的功能在现在的开源算法库（如 sklearn）中已经非常常见。最近 AutoML、H2O 等知名开源平台使得自动化建模又有了长足的发展。然而建模自动化并不是终点。

增强型分析（Augmented Analytics）于 Gartner 在 2017 年 7 月发表的《增强型分析是数据及分析的未来》[⊖]报告中首次进入人们的视野。其核心的概念包括[⊜]：

- 智慧数据洞察（Smart Data Discovery）。应用相关的工具能够比较智能和自动化地实现数据收集、准备、集成、分析、建模，能够输出各种洞察，可以为人们在战略方向、对应具体范围的战术活动（如针对某市场机会发起营销）、执行（具体执行营销策略）等不同层面的活动提供指导，包括相关关系的发现、模式识别、趋势判断与预测、决策建议等。
- 增强型数据准备（Augmented Data Preparation）。提供智能化的工具使得业务人员能够快速、轻松地访问数据，并连接各种数据源通过统一的、标准化的、可交互的视图展现内容、数据间的关系等。同时提供丰富的工具进行自动数据归约、清

⊖ https://www.gartner.com/doc/3773164/augmented-analytics-future-data-analytics。

⊜ https://www.dataversity.net/augmented-analytics-matter/。

需要全本请访问 www.ertongbook.com

洗、智能化分箱、降噪等功能。增强型数据要能够在原数据和经过数据治理后的数据间灵活处理，尽量避免因为数据治理而丢失信息，同时也避免在大量原数据间进行无序的探索。

从上述的定义中可以看出，增强型分析的特点是其可以智能和自动地完成数据准备和数据分析的工作。对于增强型分析的一个美好的预期就是“交给机器大量的原数据，机器直接针对特定场景给出决策建议”。要实现这个愿景需要人们至少完成如下的几个要点。

（1）大数据存储与访问

基于大数据平台的存储、计算的相关技术发展很快，目前已经比较成熟，能够高效地处理大量数据。

（2）数据分析流程的组件化、标准化改造

数据分析过程中关键步骤如数据收集、准备、集成、分析、建模等过程，需要细分为不同的子任务，并通过子任务间的灵活搭配构成数据分析的流程。流程的自动化运行以及对应的有价值的结果输出已经有了较好的组件，如 H2O 等。

（3）提供大量的算法支持数据处理、模型构建

算法既可以用来构建业务模型，也可以用来分析数据间的关系、进行变量聚类等工作。

（4）将“模型洞见到业务决策”纳入分析范围

模型输出洞见，如模型输出每一个客户的购买可能性，还需要配套如“当购买可能性大于 90% 时再根据时机因素进行推荐”的业务决策，才能在实际营销活动中实施。这是一个“洞见—决策—行动”的过程。

实现增强型分析所需的技术势必是庞杂的，本书的重点涵盖范围是数据处理、算法及模型、“模型洞见到业务决策”的分析等内容。这些内容既是我们日常建模时要用到的技术，也是增强型分析中必不可少的内容。虽然增强型分析的表现形式是追求智能化、自动化等功能，但是增强型分析的终极目标还是通过数据分析发挥数据价值。目前增强型分析还处于概念在逐步清晰但需要不断发展的阶段，所以本书的重点是聚焦在其本质内容，即数据处理、算法及模型、“模型洞见到业务决策”的分析等内容。

本书特点

应用机器学习、人工智能技术不仅需要理解算法原理，还需要对算法参数调优、算法使用时的数据要求、算法输出结果，以及如何在具体业务场景使用数据挖掘模型等方面都有所了解，这样才能真正发挥数据价值，产生实际的业务效果。

本书作者结合多年来给不同的大型机构“构建数据挖掘模型、解决实际业务问题”的实践，总结归纳技术、应用等方面的经验，以“介绍较新机器学习及人工智能技术”和“如何应用这些技术解决实际问题”两个方面作为本书的整体选题思路。总体来讲，

本书具有如下两个主要特点。

（1）介绍较新的技术

有监督学习的建模技术早已不是只懂得算法就可以了。目前基于集成学习、Grid Search、交叉验证等自动化建模技术方兴未艾，这些技术在专门的章节作了重点介绍；基于序列模式挖掘、序列规则、序列预测等进入公众视野还较新的技术在实际业务中有巨大的价值，这些也是本书介绍的重点；对于目前比较火热的深度学习、对抗学习等内容，本书也有专门的章节进行介绍。从这些技术的特点来看，已经具备了增强型分析的部分特点，如集成学习的技术就是旨在将多个模型结合起来，达到相对于单独采用一个模型而明显改善的效果。

（2）兼顾原理与大量实例

按照深入浅出的方式介绍算法原理、参数调优及使用方法等信息，并结合实际例子展示如何使用以及使用时的思路。笔者采用“深入浅出的原理介绍 + 实际使用的案例”的内容安排，期望能够让读者真正了解机器学习及人工智能的技术原理、特点与使用方法，并能直接在实践中起到指导作用。

除此之外，在本书中涉及汉语直译不能达意的词汇时都是采用英语原词，方便读者能够与科技类的英文材料对应，尽量避免生硬翻译带来的疑惑。在本书的大量实例中，代码注释基本上都是英文的，这与笔者多年的编码习惯有关。

读者范围

本书的目标读者是实际解决业务问题的数据分析建模人员。目前各个企业在应用机器学习及人工智能方面，不断在人才、技术、平台方面进行投入，特别是不断招聘了大量的数理统计、机器学习方面的人才。但是能够实际解决业务问题的数据分析建模人员，除了对算法原理要了解外，还需要对业务有一定了解，同时需要打开眼界快速了解不同的建模方法能够解决什么问题；除此之外，还要具备较高的实践能力，能够灵活应用不同的技术工具来快速完成任务。

本书“深入浅出的原理介绍 + 实际使用的案例”的内容安排能够使得数据分析建模人员从算法原理、数据挖掘知识结构、业务应用方法等方面得到提升，帮助数据分析建模人员开阔眼界、优化知识结构、提升实践技能。

从整体来说，本书适用于中、高级的数据分析建模人员，但是初学者也能从实例中得到重要的参考。

章节概要说明

在本书的内容安排中，保持业务和技术两个主线：业务主线是数字化转型背景下的

智慧营销、智慧风险管控如何通过数据分析完成具体工作，实现由初级的“主动营销”到“被动营销”，再到“全渠道协同营销”等营销手段的升级应用；技术主线是从常见算法的较新发展到深度学习及对抗学习的“复杂度由低向高”“分析技术由预测性分析到Prescriptive 分析”进行介绍。具体对应于大纲的内容如表 1 所示。

表 1　本书章节大纲概览

章节	技　术		业　务	
	内容概要	技术分类	内容概要	业务分类
第 1 章	鼓励数据科学家加入数字化变革的进程，与业务深度结合			
第 2 章	数据处理技巧、数据可视化等	描述性分析		
第 3 章	介绍预测类模型构建时的新方法、新思路、新工具	预测性分析（输出洞见）	通过一个具体案例，利用看重客户需求而从众多产品中寻找最可能的推荐。较之前单个产品响应预测，是从“以客户为中心的视角”来产生推荐，以解决多产品排序的问题，实际效果有较大提升	主动营销
第 4 章	介绍序列分析的相关技术，应用较新的算法以实例的方式说明算法原理、特点、注意事项等		客户行为是不是存在一些共有模式？客户下一个行为会是什么？这些都是具体营销和分享管控领域的实际问题，对营销和风险的具体决策具有非常大的影响。仔细挖掘，善于应用，往往能取得非常好的效果	事件式营销（被动营销）
第 5 章	介绍 Prescriptive Analytics 的相关技术，这方面的技术注定会成为数据分析不断深入应用时要用到的重点技术	Prescriptive 分析（输出决策）	因为传统模型大多只输出名单，而 Prescriptive Analytics 模型要输出的是“名单 + 决策”，实现真正的智能决策	全渠道协同营销（考虑成本、收益等诸多限制因素）
第 6 章	通过与传统模型的对比，介绍 CNN 算法的原理，通过大量实例说明其特点、用法、实际效果等			

（续）

章节	技　术		业　务	
	内容概要	技术分类	内容概要	业务分类
第 7 章	通过介绍 RNN 算法的原理、特点，以大量实例的方式说明其用法		通过 LSTM 算法研究客户行为预测，掌握精准的营销时机	基于客户行为事件式营销
第 8 章	通过介绍 Generative Adversarial Network 算法的原理、特点，以实例的方式说明其用法			

总体来讲，本书是一本既能扩展读者视野又具有实际参考价值，能够紧贴实际业务的关于大数据与人工智能的书籍。

在上述章节中，笔者完成了大部分工作，另外两位作者协助笔者做了一些内容补充，这些内容包括：张宗耀完成了 2.1 节、2.2 节、3.6 节、5.4 节、7.3 节；聂磊完成了第 2 章的大幅修改、5.5 节、5.7 节、6.2 节。在整个写作过程中，大家经常一起讨论、相互学习，这个过程很愉悦！

为什么写这本书

笔者自 2008 年加入 IBM SPSS，从一个单纯的软件开发者变身为数据分析行业的参与者至今已经快 11 年了。在这段时间，数据分析行业发生了巨大变化，作为行业的参与者，笔者自身从业经历也在不停地发生变化。总结下来，笔者遵从“数据分析驱动业务”的主线，按照“软件开发人员——数据挖掘工具开发者和团队管理者——资深数据科学家——深入理解业务的资深数据科学家——深刻理解数字化变革的高级咨询顾问和管理者”的职业路径，在数据分析行业的浩瀚波澜中前行。这些年的从业经历，笔者有如下几点感触。

（1）从事数据分析行业的人是需要不断充电的

用“日新月异”来形容数据分析的发展是最为确切的了，新技术、新论文不断涌现。大量书籍上描述的是一些基本的算法，对于新技术、新算法，我们应该永远保持不断学习的态度，才能在日常数据分析实践中不断发挥作用。书中并没有讲大家在很多书籍上能看到的传统算法，而是重点讲一些大多数书籍还未涉及的内容。

（2）真正发挥数据价值需要融会贯通数据与业务

在很多情况下，当数据科学家花费大量时间和精力构建出模型后，兴高采烈地试图交给业务人员使用时，往往会遇到一个有趣的情况：业务人员听不懂你对高深算法的解

释，甚至不在乎你对数据的各种费心处理，他们只关心实际的问题，如模型到底效果如何。所以在本书中穿插了大量与业务相关的例子。

（3）数字化变革的浪潮与数据分析的广泛应用密不可分

数字化变革是目前几乎所有企业都无法回避的任务。企业由于所处行业、自身特点等原因，需要量身定制数字化转型的战略。大型企业需要选择发展重点作为突破方向，在转型过程中既要做好技术基础，也需要大力推行敏捷的方法，同时要对人们的观念、组织内的流程等方面做出更新。数据分析的广泛应用在数字化变革中势必要发挥巨大作用。笔者认为数据分析者要“抬头看”，深刻地参与到数字化变革的浪潮中。

本书的写作历时近一年，笔者在做好本职工作的同时花费了巨大的精力总结归纳过往项目经验、学习研究新技术。这个过程既是一个自我充电的过程，也是一个不断总结归纳的过程。笔者试图尽力做到将自己走过的路按照深入浅出的方式讲出来，期望提供一定的参考价值。这也是笔者写这本书的目的。

笔者相信书中难免有一些疏漏，非常希望能够得到阅读反馈。读者可以通过 yfc@hzbook.com 联系到笔者。

感谢

笔者年近不惑，能够有大量时间花费在写书上，是因为笔者的父母、爱人、孩子给笔者铸就了一个坚强的后方。“风暖春日雪，化作涓涓流”，这是爱人、孩子和笔者在一次春游时看到终南山中的雪即兴而作的。其实这也能对应到现实中，家人的爱和关心让笔者在前行时如沐春风，遇到困难时他们就是笔者的动力！同时也感谢笔者的三个姐姐对笔者的关心和鼓励。

感谢另外两位作者张宗耀和聂磊，一位是我的师弟，另一位是与我完成过第一本书《发现数据之美——数据分析原理与实践》的合著者。兄弟之情已经在聚会、讨论、相互学习、写作中镌刻在我们各自的人生轨迹中！

感谢德勒中国副主席、金融服务业领导合伙人吴卫军能够在百忙之中给本书作序并给出非常积极的评价。吴总在写序过程中，严谨的工作态度给我留下深刻印象，这使我觉得他的序言非常重要。感谢笔者的研究生导师——西安交通大学朱利教授的鼓励与肯定，并欣然接受给本书作序的请求。青春挥洒的校园生活是笔者不能忘记的，特别感谢老师在笔者上学期间的关心与培养。

感谢笔者的老板吴颖兰（德勤全球主管合伙人）在笔者写作过程中的鼓励；感谢上海依图网络科技有限公司 COO 张小平在笔者写作过程中给予的鼓励；感谢美丽聪慧的同事崔璨、罗瑞丽能够在笔者写作过程中不断给予鼓励，并提出非常有价值的意见；感谢同事李敬军、曹文俊、刘田林、刘婷婷、仇敏讷、李宸豪、马克、母丹、张宇姮，在一

起做项目的过程中，我们相互学习、相互成长。

感谢我们的客户，在项目中我们能够相互学习、相互提高。可以非常肯定地说，客户的很多痛点是笔者不断学习的动力所在。

感谢机械工业出版社杨福川编辑对本书的肯定，他的专业性和工作效率让笔者惊叹。感谢机械工业出版社常晓敏老师在“鲜读”渠道对本书的大力推广，也万分感谢“鲜读”渠道上热心读者给本书内容提出的各种意见和建议。

彭鸿涛

需要全本请访问 www.ertongbook.com