狗熊会

潘蕊 等 著

数据思维实践

从零经验到数据英才

北京大学出版社
PEKING UNIVERSITY PRESS

数据思维实践

从零经验到数据英才

潘蕊 等 著

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

在大数据时代的背景下，商业分析能力显得尤为重要，具有商业分析能力的人才供不应求。不同于其他经典的统计学教科书，本书是一本非常实用的数据分析实战指导手册。

本书的灵感来源于狗熊会“人才计划”，全书框架也沿用人才计划，以一系列TASK的形式构建。全书涵盖数据分析的选题与背景、数据的获取与描述、模型的建立、表达与沟通和实战案例收录五大核心模块，具体内容为：第1章主要介绍数据分析中选题的确定方法，以及数据分析报告中背景介绍部分的撰写思路；第2章主要介绍数据的获取方式，以及数据介绍与描述分析部分的撰写、展示方法；第3章主要介绍数据建模的基本思路，以及常用模型方法；第4章主要介绍数据分析报告的撰写及展示分享时的表达与沟通技巧，以及代码规范的一系列问题；第5章主要分享一些优秀的数据分析报告案例，供读者学习参考。

本书适合数据分析入门者、对商业分析感兴趣的或正在从事相关工作的读者，可以帮助读者建立系统的数据分析框架，提高利用数据分析工具进行业务分析的能力，从而成为一位具有商业分析能力的数据科学人才。

图书在版编目(CIP)数据

数据思维实践 / 潘蕊等著. — 北京 : 北京大学出版社, 2018.8
ISBN 978-7-301-29614-1

Ⅰ. ①数… Ⅱ. ①潘… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第123482号

书　　名　数据思维实践
SHUJU SIWEI SHIJIAN
著作责任者　潘　蕊　等　著
责任编辑　吴晓月
标准书号　ISBN 978-7-301-29614-1
出版发行　北京大学出版社
地　　址　北京市海淀区成府路205 号　100871
网　　址　http://www. pup. cn　　新浪微博: @ 北京大学出版社
电子信箱　pup7@ pup. cn
电　　话　邮购部 62752015　发行部 62750672　编辑部 62570390
印 刷 者　北京大学印刷厂
经 销 者　新华书店
720毫米×1020毫米　16开本　16印张　256千字
2018年8月第1版　2018年9月第2次印刷
印　　数　8001-11000册
定　　价　79.00 元

序

FOREWORD

今天，我想跟大家隆重推荐《数据思维实践》这本书。这是继《数据思维：从数据分析到商业价值》后，狗熊会团队的又一心血之作。这本书凝聚了狗熊会“人才计划”部分创始团队成员的心血。他们分别是常象宇（政委）、陈昱（昱姐）、关蓉（关关）、刘婧媛（媛子）、潘蕊（水妈）、王菲菲（灰灰）及周静（静静）。尤其是水妈潘蕊，作为狗熊会“人才计划”的核心创始人，贡献巨大。

要想更好地了解这本书，您首先需要了解一下狗熊会对数据科学人才培养的一些基本看法、“人才计划”这个非常有趣的公益项目，以及狗熊会在人才培养方面的美好梦想。

狗熊会对数据科学人才培养的基本看法

狗熊会的使命由两句话构成：聚数据英才，助产业振兴！其中第一句“聚数据英才”关心的就是数据科学人才的培养。但是，具体到执行层面，它的内涵是什么，如何落地，并不十分明了。为此，狗熊会做了各种尝试，并渐渐形成以下基本看法。

第一，数据科学人才奇缺。狗熊会核心团队成员都是各个高校的老师。对此，我们有切身的体会。就我自己而言，经常有行业中的朋友请我推荐靠谱的学生。但遗憾的是，我却没有多少学生可推荐。而且，似乎无论推荐多少，大家都觉得不够。我甚至认为，随着国家人工智能、大数据战略的实施，数据产业的快速发展，数据科学行业人才的匮乏情况会更加严重。

第二，数据科学体系中，尤其缺乏的是商业分析（Business Analytics）人才。在我看来，数据科学相关领域大概可以被分为三大领域。第一个是工程领域。它关注大规模计算机集群的软硬件基础设施，如大规模集群建设（硬件）、分布式存储与计算系统开发（软件）。第二个是算法领域。它关注新的模型与算法开发，如传统的统计模型和机器学习模型（含深度学习）的研发。这两个领域（尤其是第二个）备受关注。但人们似乎忽视了第三个领域——商业分析。商业分析领域关注大数据软硬件环境及各种成熟模型与算法，不关心具体的建设与开发。商业分析的核心问题是如何洞察数据的商业价值，致力于如何把各种具体的商业问题转化为数据可分析问题。这是商业分析关注的重点。由此可见，商业分析承担着连接商业问题与数据模型的重要使命。因此，人才需求如井喷一样旺盛，而市场供给却严重不足。

第三，现有教育体系的人才供给严重不足。现有的教育体系主要分为两大块：一块是正规高校，另一块是各种培训班。正规高校的优点是知识体系完整、理论素养扎实、学生素质高；缺点是理论与实际脱节严重，经典的知识体系与校园外正在如火如荼发展的数据产业相比严重滞后。而各种培训班的优点是直接把企业专家请入课堂，把企业实际问题带入学习，因此非常接地气；缺点是短期性和功利性太强，基本上以传授一种“技能”为主，对于学生的终身学习与成长帮助非常有限。而且，市场混乱，鱼龙混杂。

狗熊会“人才计划”公益项目

基于以上几点共识，狗熊会开始思考一个问题：能否成规模地培养高质量的商业分析人才，弥补现有教育体系的不足？这里有两个关键点：一是成规模，二是高质量。我们认为这个目标与狗熊会“聚数据英才”的使命高度一致，是我们愿意为之努力拼搏的一项重要工作。带着对这个问题无限的使命感，狗熊会核心创作团队产生了创办狗熊会“人才计划”的想法，并且已经成功举办了两期。

狗熊会的“人才计划”是什么？简单地说，就是一个培训班。但是，它与普通的培训班又有很大的不同。

第一，免费公益。虽然狗熊会是一个 100% 的盈利机构，但是人才计划是一个 100% 的公益项目，狗熊会不向学生收取任何费用。也正因为不收取任何费用，所以狗熊会人才计划执行非常严格的纪律要求和淘汰制度，这样才能保证顺利毕业的同学符合高质量的标准。

第二，全球招生。“人才计划”为了能够帮助更多的学生成长，会在全国甚至全球范围内招生。事实上，第一次“人才计划”就吸引了来自海内外 300 多名同学报名，最后录取了 100 多名，有 65 名同学顺利结业。申请的同学有的来自国内 985 或 211 高校，有的来自民办大学，有的来自海外，如 MIT（麻省理工学院）、MSU（密歇根州立大学）等。

第三，全部线上。既然“人才计划”学员来自世界各地，那么相应的教学管理必须是 100% 在线上执行，线下教学活动不太好实行，这其实并不是狗熊会特意而为。“人才计划”创办之初，狗熊会设想从北京开始，这样可以安排一些线下学习的机会。但是，报名名单送上来的时候，我们自己也很惊讶：大量的同学来自全国各地，甚至海外。因此，我们意识到，“人才计划”不可能采用传统的线下教学方式，必须是 100% 线上执行。

第四，TASK 驱动。线上教学如何开展？是不是放个视频，做个直

播，发点讲义？这些都不是好的培养方式。因为在这样的培养方式下，学生是被动“教”出来的，而不是主动“学”出来的。因此，我们创造了TASK这样一种独特的人才培养方式。简单地说，TASK就是一个具体可被执行的任务。一个大的项目（如利用刷卡交易数据做征信），可以被切分成多个细小、可执行的任务。例如，一个任务可能就是在R中读入数据，并对各个指标做描述统计。这些TASK不见得都是与数据编程相关，但是都与业务目标相关。TASK布置给学生后，老师不承担“教”的任务，学生需要想办法去自学。在TASK的驱动下，学生开始自主思考探索，并通过各种手段去“自学”。因此，学生练就了非常强的自学能力，而知识的增长仅仅是一个结果而已。

狗熊会人才培养的美好梦想

你看，这就是狗熊会误打误撞、自己摸索出来的“人才计划”公益项目。到目前为止，已经成功举办了两期。有兴趣的朋友可以在狗熊会微信公众号（CluBear）中输入“人才”，即可看到非常详细的介绍及往期学员的作品。

但是，这么好的一种人才培养方式是不是只能狗熊会自己独享？显然不是！狗熊会的力量太渺小，微不足道。如果想要为数据产业做出更大的贡献，需要更多志同道合的朋友加入进来，尤其是高校的优秀老师。我们希望TASK驱动的教学理念能够进入更多高校、培训班的课堂，并通过更多的教学实践不断改进。为此，一本高质量的教科书必不可少！

“水妈”带领狗熊会的核心创作团队，将过去一年关于“人才计划”的宝贵教学经验整理成册呈现给大家，希望能够帮助更多的老师、同学一起学习成长。大家从目录可以看出，这本书与所有的数据科学教材非常不一样，因为它承载着完全不同的教学理念。

本书一共5章。第1章介绍TASK的学习理念。TASK要求每个学生（或学习小组）确定一个自己感兴趣的研究题目，我们建议这个题目尽可能贴近生活、贴近真实的业务。例如，我们可以关心一下知乎上都在讨论什么；大学生活如何才能有一段美好的爱情；游戏达人怎样才能疯狂“吃鸡”。选题非常重要。一个好的题目，能够激发同学们的好奇心，并因此产生无穷探索的勇气。这是TASK学习的重要理念。

第2章学习数据的获取与描述。在实际工作中，很少有数据都整理好了就等着分析这样的场景，更多的时候需要自己去收集数据。而移动互联网时代赋予我们非常多的、非常便捷的数据采集手段。例如，使用“问卷星”可以通过微信发放问卷收集数据；又例如，使用“八爪鱼”可以非常便捷地采集网站数据等。在此基础上，如何对数据做最基本而有效的描述也是非常重要的内容。我们不盲目追逐各种“高大上”的可视化软件，而是希望所有数据描述都能够准确地瞄准业务需求。在这个前提下，越简单越好。

第3章系统学习数据建模。包括两大类最常见的无监督学习方法，即数据降维与聚类分析。还有两大类有监督学习方法，它们分别对应连续型因变量及离散型因变量。最后，还有一节关于文本分析的内容。这部分所涉及的技术细节是任何经典统计学或机器学习教科书中都可以找到的。本书无意呈现过多的技术细节，也无意覆盖很多的内容，但是希望学生能够在数据描述的基础上顺利过渡到数据建模的任务上来。整个学习的核心仍然是理解数据分析和业务问题的互动关系。

第4章是极具特色的一章，是绝大多数数据科学教材里不会涉及的。这一章的核心问题是表达与沟通，这里的表达与沟通不局限于口语，更多的是书面的表达与沟通。其中包括报告撰写、PPT制作及代码规范，这是实际工作中最常见的表达与沟通的手段。相关教学内容都围绕如何有效地呈现数据分析结果，而不至于让数据分析的辛苦努力白费。

第 5 章收录了一些具体的实战案例，供读者参考、学习。

最后要深深感谢本书的作者团队，他们也是“人才计划”的创始团队成员。他们的辛苦付出，尤其是水妈，成就了本书。

希望读者通过本书建立起自己的数据分析思维。

王汉生（熊大）

前言

PREFACE

2017 年 7 月 17 日，狗熊会“人才计划”第一期正式开始。“人才计划”旨在培养具有商业分析能力的人才，通过数据分析工具，解决实际业务问题。在培养过程中，注重锻炼学生快速学习和解决实际问题的能力。“人才计划”以 TASK 的形式推进，学生以自学为主，定期汇报成果。从“人才计划”的培养效果来看，学生们在自学能力和实战经验方面都有了很大的提高。于是，熊大（王汉生教授）鼓励我把“人才计划”的 TASK 内容梳理出来，与更多的学生或读者分享。这便是本书形成的初衷。

本书的目标主要有两个，一是培养读者的快速自学能力，二是培养读者的数据分析实战能力。

第一，快速自学能力。强调快速自学能力，缘于我这几年的教学经历。在学校，老师教什么，学生就学什么，一旦需要用新的知识解决问题，学生便束手无策。这与今后的工作需要极其不符。在工作中，学过的知识可能大都用不上，实际需要用到的知识都需要重新去学习。在校期间的时间非常宝贵，对于养成良好的学习习惯至关重要。因此，尽早掌握快速自学能力非常有益。

意识到这一点，便能更好地去理解这本书，以及这本书的基本构成单元——TASK。TASK 就是一个个的小任务，每个 TASK 有一个明确的任务主题、与主题相关的内容讲解、参考资料及任务作业。读者需要在阅读 TASK 后，花时间找到更多的学习资源进行学习和思考，才有可能完成任务

作业。在这个过程中，读者会在寻找学习资源的能力、提炼和归纳相关知识点的能力及以报告撰写为主的表达沟通能力等方面得到锻炼。

第二，数据分析实战能力。在教学过程中，我所看到的另一个普遍现象是学生们学习了很多理论知识和统计学模型，但拿到一个实际数据却不会做分析，更不会形成一份规范的报告。数据分析实战能力的匮乏，令人感到遗憾。因此，TASK 的设计的重点在于提高数据分析实战能力，没有过多的理论细节。各种统计模型（如回归分析）的理论基础已经有太多经典教材做过详细的讲解，因而不是这本书的侧重点。

为了支持更多的教师授课，本书配备了精美的 PPT 课件，教师不必再自己费心制作课件。此课件连同书中所涉及的数据、代码和案例的 PPT，都可以在狗熊会的官方网站或狗熊会公众号（CluBear）上下载，这些资料将免费提供给读者使用。左侧是狗熊会公众号的二维码。

本书能够出版，要感谢我的导师熊大和狗熊会 CEO 李广雨先生，他们的不断鼓励让我坚信这是一件非常有意义的事情。感谢狗熊会的核心创作团队成员：常象宇（政委）、陈昱（昱姐）、关蓉（关关）、刘婧媛（媛子）、王菲菲（灰灰）及周静（静静）。没有他们的大力支持，本书不会如此顺利地完成。未来，狗熊会还将继续并肩作战。感谢中央财经大学的樊津畅、高天辰、王晶冰、王蕾、张宇轩和翟晋同学，书中的图表、代码及 PPT 大多出自他们之手。

“聚数据英才，助产业振兴”，狗熊会一直在路上！

水妈

目录

CONTENTS

第 1 章　选题与背景　// 1

1.1　TASK 概述　// 1

1.2　TASK 1 确定选题　// 3

1.2.1　选题的思考路径　// 3

1.2.2　可能的选题方向　// 3

1.2.3　补充材料　// 7

1.2.4　课后作业　// 7

1.3　TASK 2 学写背景介绍　// 8

1.3.1　如何写背景介绍　// 8

1.3.2　背景介绍经常出现的问题　// 9

1.3.3　课后作业　// 9

1.4　范例与点评　// 10

1.4.1　范例一　// 10

1.4.2　范例二　// 12

1.4.3　范例三　// 15

第 2 章　数据的获取与描述　// 19

2.1　TASK 3 数据的获取　// 19

2.1.1　搭建框架　// 22

2.1.2　确定问题形式　// 24

2.1.3 选措辞、排结构 // 25
2.1.4 评估、预测试 // 27
2.1.5 课后作业 // 28
2.2 TASK 4 数据介绍与说明 // 32
2.2.1 数据变量说明表 // 32
2.2.2 用 PPT 介绍数据 // 34
2.2.3 常见的问题 // 35
2.2.4 课后作业 // 36
2.3 TASK 5 数据的描述——外表美 // 37
2.3.1 描述分析简介 // 37
2.3.2 描述分析的整体规范 // 37
2.3.3 统计图的规范 // 39
2.3.4 课后作业 // 40
2.4 TASK 6 数据的描述——内在美 // 41
2.4.1 准确使用统计图 // 41
2.4.2 写好描述性文字 // 42
2.4.3 扩展阅读材料 // 46
2.4.4 课后作业 // 47
2.5 范例与点评 // 48
2.5.1 范例一 // 48
2.5.2 范例二 // 50
2.5.3 范例三 // 54

第 3 章 模型的建立 // 62

3.1 TASK 7 建模的流程 // 62
3.1.1 建模前的准备 // 62
3.1.2 模型的选择与建立 // 63
3.1.3 模型的解读与评价 // 64
3.1.4 课后作业 // 65
3.2 TASK 8 无监督学习：数据降维 // 66

3.2.1 主成分分析 // 66
3.2.2 因子分析 // 71
3.2.3 课后作业 // 75
3.3 TASK 9 无监督学习：聚类分析 // 76
3.3.1 聚类分析概述 // 76
3.3.2 层次聚类法 // 78
3.3.3 K 均值聚类法 // 79
3.3.4 课后作业 // 82
3.4 TASK 10 有监督的学习：连续型因变量 // 83
3.4.1 模型的建立与估计 // 84
3.4.2 结果的整理与解读 // 86
3.4.3 模型诊断与改进技巧 // 89
3.4.4 模型选择：准则和步骤 // 95
3.4.5 课后作业 // 99
3.5 TASK 11 有监督的学习：离散型因变量 // 100
3.5.1 逻辑回归模型 // 101
3.5.2 模型的评价 // 105
3.5.3 决策树 // 108
3.5.4 课后作业 // 112
3.6 TASK 12 文本分析 // 113
3.6.1 文本分析可以干什么 // 113
3.6.2 文本分析的主要内容 // 113
3.6.3 文本分析基本流程 // 114
3.6.4 文本分析示例 // 119
3.6.5 课后作业 // 126

第 4 章 表达与沟通 // 127

4.1 TASK 13 报告的撰写 // 127
4.1.1 报告概述 // 127
4.1.2 报告的核心要素 // 128

4.1.3 如何撰写优秀的报告 // 128
4.1.4 课后作业 // 133
4.2 TASK14 PPT 的制作 // 134
4.2.1 PPT 的特点 // 134
4.2.2 制作 PPT 的步骤 // 134
4.2.3 示范与点评 // 141
4.2.4 课后作业 // 145
4.3 TASK15 以 PPT 为核心的表达与沟通 // 146
4.3.1 从“表达与沟通”的角度看 PPT 制作的问题 // 146
4.3.2 表达与沟通的注意事项 // 148
4.3.3 课后作业 // 151
4.4 TASK16 代码规范 // 152
4.4.1 代码注释 // 152
4.4.2 代码命名规则 // 153
4.4.3 代码模块化 // 154
4.4.4 代码调试 // 156
4.4.5 代码效率优化 // 158
4.4.6 课后作业 // 161

第 5 章 实战案例 // 162

5.1 案例一 // 162
5.2 案例二 // 173
5.3 案例三 // 186
5.4 案例四 // 199
5.5 案例五 // 211
5.6 案例六 // 229

参考文献 // 242

第 1 章
选题与背景

本章介绍如何确定选题及怎样撰写选题背景，并通过 3 个范例与相应点评，帮助读者快速熟悉选题思路及撰写选题背景的注意事项。

1.1 TASK 概述

TASK，即“任务”，是狗熊会人才培养的核心。狗熊会试图通过一系列 TASK，帮助数据分析爱好者完成一个完整的数据分析作品。本书分为如下几个 TASK 模块。

（1）模块一：选题和背景（2 个 TASK）。

（2）模块二：数据的获取与描述（4 个 TASK）。

（3）模块三：模型的建立（6 个 TASK）。

（4）模块四：表达与沟通（4 个 TASK）。

具体来说，每一个 TASK 包括如下内容。

（1）一个明确的任务主题。例如，确定选题、描述分析、数据降维等。通过这个任务主题，能够了解 TASK 的主要任务方向。

（2）关于任务主题的讲解。这是 TASK 的主体，详细陈述了相关任务主题的重要性、主要内容和学习重点等。

（3）与任务主题相关的学习材料。这是 TASK 的补充内容，为读者

提供可能的学习素材和资料来源，方便读者课后进行自学。

（4）课后作业。每一个 TASK 都会有相应的作业，读者需要认真完成并反复练习，才能有所提高。

在 TASK 的学习过程中，最主要的是锻炼快速自学的能力。TASK 的内容偏向于数据分析实战，并不会涉及太多的理论细节。希望读者通过本书的系列 TASK 的学习，享受数据分析所带来的快乐！