大数据分析方法

用分析驱动商业价值

[美] 米歇尔·钱伯斯（Michele Chambers）
托马斯 W. 迪斯莫尔（Thomas W. Dinsmore）◎著

韩光辉 孙丽军◎等译

ADVANCED ANALYTICS METHODOLOGIES

DRIVING BUSINESS VALUE WITH ANALYTICS

机械工业出版社
China Machine Press

ADVANCED ANALYTICS METHODOLOGIES
DRIVING BUSINESS VALUE WITH ANALYTICS

大数据分析方法
用分析驱动商业价值

[美] 米歇尔·钱伯斯（Michele Chambers）
托马斯 W. 迪斯莫尔（Thomas W. Dinsmore）◎著

韩光辉 孙丽军◎等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据分析方法：用分析驱动商业价值 /（美）钱伯斯（Chambers, M.），（美）迪斯莫尔（Dinsmore, T. W.）著；韩光辉等译 . —北京：机械工业出版社，2016.6
书名原文：Advanced Analytics Methodologies: Driving Business Value with Analytics

ISBN 978-7-111-53731-1

I. 大… II. ① 钱… ② 迪… ③ 韩… III. 商业信息学 IV. F713.51

中国版本图书馆 CIP 数据核字（2016）第 096888 号

本书版权登记号：图字：01-2015-5214

本书全面介绍了针对大数据的分析方法。本书内容全面、前沿，可帮助读者针对当前的组织需求和分析能力找到合适的技术和方式来进行合理的分析。本书采用循序渐进的讲授方式，帮助读者制定能支持其企业需求、实现分析功能的路线图，同时兼顾企业文化及客户和企业相关利益群体的需求。

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：刘诗灏　　责任校对：殷　虹
印　　刷：北京瑞德印刷有限公司　　版　　次：2016 年 8 月第 1 版第 1 次印刷
开　　本：170mm × 242mm　1/16　　印　　张：17.5
书　　号：ISBN 978-7-111-53731-1　　定　　价：69.00 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

大数据可谓当今商业与技术领域最热的词之一，但是，经常见到大家对大数据的滥用和误解，以为做几个数据分析就是大数据了。翻译这本书起源于译者在从事大数据相关工作时，阅读了大量关于大数据方面的书籍和文献，发现有些书只重点讲大数据的商业案例和商业价值，没有涉及背后的技术，而有些书则一下深入到大数据艰深的技术层面，缺少对商业价值以及配套的战略、组织、人才的阐述。总之，很少有书籍和文献能够把大数据商业与技术两个不可或缺的方面结合在一起讲清楚。看到本书，我们感到非常惊喜，因为它是目前看到的唯一一本从大数据战略、组织、人才、技术、落地路径以及具体大数据技术，甚至市场上一些可用工具方面来讲述的书。像作者阐述的，大数据商业价值的实现依靠大数据的战略、人才、组织以及技术的完美结合，本书向读者勾画了一幅大数据全图，适合企业管理者、技术管理者以及技术“极客”们阅读。

翻译是一个“苦活”，译者作为大数据的从业者以及管理者，之所以能够在繁忙的工作中坚持翻译完此书，一方面是对大数据深深的热爱，另一方面是对本书独特价值的认可，因此，我们推荐你们通读此书，并在本书的指导下实践大数据的技术，从而实现大数据的商业价值。

推 荐 序

在大数据时代，客户越来越认可高级分析可以为其业务带来差异化竞争价值 。因此，预测模型可以转化为关键业务资产，这可以带来巨大收益，但同时也需要更加严格的流程来进行运营部署，才能产生这样的商业价值。

在此背景下我们惊讶地看到，只有一小部分预测模型实际部署了，并且部署往往需要几个月的时间。组织面临着大量的业务需求、众多 IT 解决方案和数据仓库平台以及迅速增长的一系列数据挖掘工具。对于组织来说，要真正有效利用高级分析方法所提供的机会，需要打破常规，比如打破往往受限于单一厂商提供的解决方案或人工流程，才能迈向一个现代分析基础架构。这就是为什么在最近的一系列报告中，Gartner 特意强调了给最终用户组织提供厂商中立的行业标准和开放的软件平台的益处——可以通过一系列硬件和软件快速部署和执行预测模型。

本书的作者 Michele Chambers 和 Thomas W. Dinsmore 勾画了开放性分析的新蓝图，而不是单一供应商专有的分析解决方案。我们会看到通过开放标准连接在一起的基于各种商业和开源工具的开放分析平台的崛起。要成为分析的高手，组织必须定义一个独特的架构和路线图，它们可以识别不同的应用程序、用例以及用户角色的复杂性。这种架构将包括许多厂商和项目，因为单个供应商不能够满足你所有

的需求。

本书提供了定义分析架构和路线图所需的基本背景、知识和工具。我鼓励你们通读此书，因为它会为从商业思维、人文因素、组织结构到分析应用洞察和预测的分析方法论各种主题提供有价值的指导。

——Michael Zeller

Zementis 公司 CEO

致　谢

出书的难度可想而知，翻翻这本书你就会感受到写书所需要完成的工作量。我们之所以承担这个工程，是出于对这个领域的热爱，并且想把我们的真知灼见分享和回馈给其他人。虽然关于技术的书永远是无法写完的，因为这个行业在不断地发展和演变，但现在本书终于截稿了。

一路走来，我们曾与许多思想领袖、相关领域专家进行合作，其中充满了乐趣。感谢他们给予的时间、支持和贡献。

感谢如下朋友的特殊贡献：

George Matthew——Alteryx

Greta Roberts——Talent Analytics

Les Sztandera——费城大学

Sujha Balaji——费城大学

感谢如下专家的经验分享：

Dean Abbott——Smarter Remarketer & Abbott Analytics

Thomas Baeck 博士——Divis Intelligent Solutions

Michael Forhez——CSC

Bob Gabruk——Cognizant

Rayid Ghani——EdgeFlip & 芝加哥大学

Kevin Kostuik——Charlotte Software Systems

Doug Laney——Gartner

Bob Muenchen——r4stats.org

Tess Nesbitt 博士——DataSong

Karl Rexer——Rexer Analytics

Greta Roberts——Talent Analytics

George Roumeliotis——Intuit

感谢如下朋友的大力支持：

埃森哲公司的 Jeffrey Brown 对本书提供的反馈。

Bill Jacobs、Lee Edlefson、Neera Talbert、Rich Kittler 和 Derek McCrae Norton 的宝贵评论和反馈。

目　录

第 1 章

现代分析基本原则

不久之前，曾经有一段时间企业的分析方法非常简单：他们从领先的开发商那里购买软件并将其安装在一个盒子中。当需求发生变化时，就从同一个开发商那里购买更多的软件，并且安装在更大的盒子里。那时的分析是特定专家的专业领域，他们使用的软件和研究生阶段的完全相同。他们相信一个简单的数据仓库就可以装下一切所需的有用信息。

回顾过去的商业节奏，可以用“悠闲”来描述：即使要花上两年的时间来实施一个预测模型，大家也都接受，这就是那时的工作效率。就在不久前，一家大型银行每年进行四次营销大战来促销它的信用卡，高管们认为这种成绩就算很好了。

好了，和过去的一切说再见吧。现在是数字媒体时代，也是 Web 2.0、移动、云和大数据的时代。数据的容量、速率、多样性都呈爆发式增长。各大企业已经放弃使用单一数据仓库的理念，因为数据复杂的多样性已经让单一数据库很难驾驭。我们已有过剩的原始资料，各类令人眼花缭乱的平台，以及无处不在的数据：本地的、第三方托管的、云端的。

数据的这种巨变给分析学领域带来了颠覆式的改变：新的业务问题、应用、用例、技术、工具和平台。当今被视为主流的技术五年之前并不

入流。原来一家软件开发商垄断了分析软件，而如今在 CrunchBase（一个关于创业公司信息的主要数据库）上就列有 851 家开发分析软件的初创公司。开源软件正在继续吞食软件世界：在 Gartner 公司最近的高级分析魔力象限的四个领导者中有两个是开源项目，调查中超过三分之二的分析师更加喜欢开源分析方式而不是流行的商业软件。

最重要的是，企业运转的节奏呈指数级加速。昨天，我们每年开展四项市场活动；今天，我们可以每小时进行四项市场活动。我们已无法容忍花两年时间来实施一个预测模型。如果还是这个节奏的话，我们就会被市场淘汰。

我们再也负担不起蓝筹股、单一厂商专有的分析架构的奢侈花销。取而代之的是，我们看到企业搭建起来的通过开放标准连接在一起的基于各种商业和开源工具的开放分析平台。在这个新的世界里，每个组织必须定义一个独特的分析架构和路线图，能够支持现代组织和经营战略的复杂性。这种架构将包括众多厂商和开源项目，因为没有任何一家厂商能够满足所有的需求。

在这本书中，我们提出了一个基于 9 项核心原则的方法：

- **实现商业价值和影响**——构建并持续改进分析方法以实现高价值业务影响力。
- **专注于最后一英里**——将分析部署到生产中，从而实现可复制的、持续的商业价值。
- **持续改善**——从小处开始进而走向成功。
- **加速学习能力和执行力**——行动、学习、适应、重复。
- **差异化分析**——探索你的分析方法从而产生新的结果。
- **嵌入分析**——将分析嵌入业务流程。
- **建立现代分析架构**——利用通用硬件和下一代技术来降低成本。
- **构建人力因素**——培养并充分发挥人才潜力。

❑ **利用消费化趋势**——利用不同的选择进行创新。

接下来，我们将充分介绍这些原则，因为它们是建立现代分析方法的基础。

1.1　实现商业价值和影响

本书后面将会描述如何创造一个独特的分析路线图，以及如何对不同项目进行优先级排序。现在，简单地说，现代分析方法的原则之一就是聚焦分析那些具有潜在的改变组织游戏规则的价值的项目。要保证组织能够实现价值，你需要评估目前的状态来确定基线，并设定初始的、可以量化的业务目标和持续的业务目标。例如，目前的收入是每年 1 亿美元，复合增长率是 4%。初步设定实现 15% 的新增收入，并且希望未来每年贡献 10% 的新增业务收入。

这样的指标可以很容易识别和衡量，而其他指标在识别和衡量上有一定难度。为了发现这些潜在的指标，需要确定商业决策通常是由哪些因素决定的。首先要衡量这些因素的影响，然后有目的地建立对业务有直接影响的指标。过去，公司通常情况下只是想有一个收益指标或者是一个运营成本指标，而不是两者兼顾。而如今，成熟的分析型组织通常建立起兼顾资产负债表两头的衡量标准。这向团队发出一个非常明确的信号——实现收益增长的同时必须有效控制成本。

精明的企业可以通过逆向思维找到潜在的分析机遇。通常情况下，在一个行业或公司内最难以解决的、根深蒂固的问题长时间存在，员工开始把这些问题看作工作中最难改变的限制条件。然而，现实是往往那些在过去看来不可能解决的问题的壁垒其实早已不复存在。从瓶颈中释放出来之后，通常会创造出大量的商业价值。分析驱动型的组织敢于打破条条框框，并在他们所面临的行业或企业中寻找出最具挑战性的问题。当做到这一点之后，他们将开始确定如何通过创新数据、技术手段来解

决或减少这类问题。通常是通过头脑风暴来寻找问题的答案，并假设解决问题所需的各种资源都可以获得。在想法确定之后，团队通常会进行另一轮头脑风暴，以确定如何获得他们所需要的但还没有得到的资源。这个团队会寻找潜在的新资源——数据、共生关系的合作者或技术来帮助他们实现业务目标，而不是使用样本或回溯测试[⊖]来找到解决方案。

要在最初和一段较长时间内实现商业价值，你需要将分析应用到生产。在任何分析展开之前，需要验证分析模型结果的准确性。今天，这经常发生在一个“沙箱”中——使用原始数据的一个有限子集，在一个人工的、非生产的环境中进行演练。有一个十分普遍的现象：“沙箱”分析模型能够满足甚至超过各项业务测试指标，但是却在实际生产环境中表现得不尽如人意。所以，你要记得在实际实施的环境中对分析模型进行评估，而不是在理想的环境中评估。在上线前，应该在一个模拟的生产环境中部署模型进行全面测试，以获得是否可以实现业务目标的实际评估。过去，部署通常在模型建立之后，而现在部署是全生命周期分析过程的一部分。一旦所有的潜在技术部署障碍确定之后，在上线投入生产前要获得法律批准或程序确认。分析模型部署后，评估最初的业务影响并确定快速方法以便不断改善结果。

1.2 专注于最后一英里

今天，很少有团队实现了将分析部署到生产环境和承诺的为组织改变游戏规则的商业价值。为了实现这个终极目标，我们从结局开始进行反向逆推。通过与一线工人交流从战略到执行的每一个细节，了解组织中每一层级、每一天面临的挑战。这些领域的专家能敏锐地意识到制约他们成功的问题。清楚地认识取得成功的代价，而不是如何取得成功。有了这样的认识后，为你的分析方法建立起量化的、远大的目标。例如：

⊖ 通常也叫回溯（backtesting）。

- 要获得的企业目标价值是多少？

  收益提高 3%？

  库存每年节省 1000 万美元？

  部署的第一年总费用节省 1 亿美元？

- 业务预期的服务水平协议（SLA）是什么？

  隔夜重新评估信用等级？

  5 分钟内完成投资组合评价？

- 运作模式是什么？

  如何将模型运用到生产？

  这个分析模型需要与其他业务系统结合吗？如果需要，操作流程和决策如何改变？

  分析模型是否由其他商业系统引发？

  这个分析模型部署在一个地点还是多个地点？

  是否有跨国或本地的要求？

  模型更新的频率是多少？

- 什么是衡量商业影响的关键成功因素？

  如何衡量成功？

  什么是失败？

  团队要经历多长时间才能取得成功？

- 什么是模型的准确性？

  模型准确性是否“足够好”可以马上实现商业价值？

  模型需要多少改进以及在什么时间改进？

传统上，一个团队的定量分析师、统计人员或数据挖掘人员负责模型的创建，而第二个团队通常是信息技术团队负责生产部署。因为这往往会跨越组织边界，所以有可能在模型创建和模型部署或评分之间存在较长时间的滞后和割裂。这两个团队必须像一个团队一样发挥作用，即

使组织边界存在且会持续下去。完整的生命周期方法可以使这两个团队进入合作状态，要求分析方法并不仅仅是创建和评估初始分析模型，还要涵盖分析模型的实际生产部署和为了实现企业的经营目标而持续地重新评估。

运用现代分析方法，团队专注于提供快速的结果，而不是等待打造出“完美的”分析模型。他们通常以概念性验证（POC）或是原型开始，虽然项目范围有局限，但是可以帮助团队加快实现商业价值。他们迅速完善并改进 POC 或原型，使其可以进行生产部署，取得系统性的收益。

1.3 持续改善

持续改善，即在生产活动中不断提高，已经被许多不同的学科借鉴，包括分析学。持续改善的核心是：

- ❑ 从小处入手
- ❑ 去除过于复杂的工作
- ❑ 进行实验以确定和消除无用之处

重点在于快速实现价值而不在于完美。测试和学习可以带来许多小的改进，并通向最终目标。

这与花费较长的开发周期建设“完美”模型的现状形成鲜明的对比。日前构建和部署分析方法是十分复杂的定制项目，涉及多个不同的功能领域。在这个新的时代，现代分析团队要去除象牙塔的学术束缚，从传统分析方法转变到消除项目周期中不必要的耗时步骤。这有助于提高在将商业反馈纳入流程的过程中的灵活性和响应能力，从而改善结果。

有持续改善作为指导原则，现代分析团队可以立即构建、部署模型，然后在很短的周期里提高模型在分析和信息技术方面的应用，从而创建一个无摩擦环境，不断地提供商业价值。由此，分析团队经常使用混合型敏捷或快速应用开发方法来缩短周期，降低与跨部门团队合作的障碍。