IBM SPSS Modeler
CRISP-DM 指南

注释

在使用本信息及其支持的产品前，请阅读第 37 页的『声明』中的信息。

产品信息

本版本适用于 IBM(r) SPSS(r) Modeler V17.0.0 及所有后续发行版和修订版，直到在新版本中另有声明为止。

目录

前言

IBM® SPSS® Modeler 是 IBM Corp. 企业级数据挖掘工作平台。SPSS Modeler 通过深度的数据分析帮助组织改进与客户和市民的关系。组织通过借助源自 SPSS Modeler 的洞察力可以留住优质客户，识别交叉销售机遇，吸引新客户，检测欺诈，降低风险，促进政府服务交付。

SPSS Modeler 的可视化界面让用户可以应用他们自己的业务专长，这将生成更加强有力的预测模型，缩减实现解决方案所需时间。SPSS Modeler 提供了多种建模技术，例如预测、分类、分割和关联检测算法。模型创建成功后，通过 IBM SPSS Modeler Solution Publisher，在广泛的企业内交付给决策者，或通过数据库交付。

关于 IBM Business Analytics

IBM Business Analytics 软件提供完整、一致和正确的信息，决策人依据此信息来提高业务性能。企业智能、预测分析、财务业绩和战略管理的完整产品组合，和分析应用程序一起提供对当前业绩的清晰、直接和实用的洞察力，以及预测未来结果的能力。结合丰富的行业解决方案，久经证明的实践和专业服务，各种规模的组织都能够实现最高生产力、确信地自动作出决策以及获得更好的结果。

作为此产品服务组合的组成部分，IBM SPSS Predictive Analytics 软件可帮助组织预测未来事件，并在该洞察的基础上提前行动以实现更好的业务结果。商业、政府和学术客户依靠 IBM SPSS 技术吸引、挽留和增长客户，同时减少欺诈和降低风险。通过在日常活动中融入 IBM SPSS 软件，成为预测企业的组织可指引并实现决策的自动化，以满足企业目标并实现可衡量的竞争优势。有关详细信息或要联系一位代表，请访问 http://www.ibm.com/spss。

技术支持

技术支持可供维护客户使用。客户可就 IBM Corp. 产品使用问题或某一受支持硬件环境的安装帮助寻求技术支持。要获取技术支持，请访问 IBM Corp. Web 站点 http://www.ibm.com/support。请求帮助时，请准备好标识您自身、组织和支持协议。

第 1 章 CRISP-DM 简介

CRISP-DM 帮助概述

CRISP-DM（即“跨行业数据挖掘标准流程”的缩写）是一种业界认可的用于指导数据挖掘工作的方法。

- 作为一种 **方法** ，它包含项目中各个典型阶段的说明、每个阶段所包含的任务以及这些任务之间的关系的说明。
- 作为一种 **流程模型** ，CRISP-DM 概述了数据挖掘的生命周期。

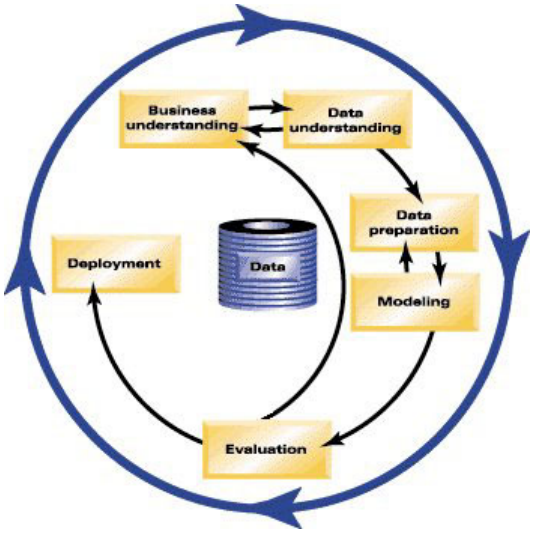

图 1. 数据挖掘生命周期。

生命周期模型中由六个阶段组成，其中的箭头表示这些阶段间最重要和最频繁使用的依赖关系。阶段之间并不一定要严格遵守顺序。实际上，大多数项目都会根据需要在这些阶段之间来回移动。

CRISP-DM 模型具备灵活性，可以轻松地自定义。例如，如果您的组织旨在检测洗钱行为，您很有可能将在不设定具体建模目标的情况下对大量数据进行筛选。此时，您的工作不是建模，而是以数据探索和数据展现为主，以便揭示可疑的财务数据模式。使用 CRISP-DM，您可以创建满足特定需求的数据挖掘模型。

在此情况下，与数据理解和准备阶段相比，建模、评估和部署阶段之间的关联性可能相对较小。但是，仍然需要考虑在这些后期阶段引发的某些问题，以便进行长期规划和制定未来的数据挖掘目标。

在 IBM SPSS Modeler 中的 CRISP-DM

IBM SPSS Modeler 采用两种方式合并 CRISP-DM 方法，从而为有效的数据挖掘提供独特的支持。

- CRISP-DM 项目工具可帮助您根据典型数据挖掘项目的各阶段组织项目流、输出和注解。您随时都可以在项目中基于流和各 CRISP-DM 阶段的注意事项生成报告。
- CRISP-DM 的帮助将指导您完成执行数据挖掘项目的整个过程。该帮助系统包括各步骤的任务列表以及 CRISP-DM 如何在实际应用中发挥作用的示例。您可以通过从主窗口的“帮助”菜单选择 **CRISP-DM 帮助**访问 CRISP-DM 帮助。

CRISP-DM 项目工具

CRISP-DM 项目工具提供了一种结构化的数据挖掘方法，有助于确保项目成功。它实际上是标准 IBM SPSS Modeler 项目工具的扩展。实际上，您可以在 CRISP-DM 视图和标准类视图之间切换以便查看按类型或 CRISP-DM 阶段组织的流和输出。

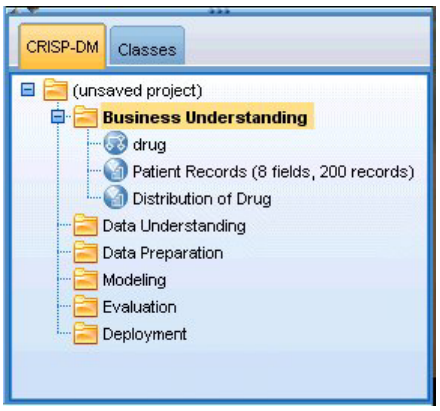

图 2. CRISP-DM 项目工具

通过使用项目工具的 CRISP-DM 视图，您可以:

- 根据各个数据挖掘阶段组织项目的流和输出。
- 记录您的组织在各个阶段的目标。
- 为每个阶段创建自定义工具提示。
- 记录根据特定图形或模型得出的结论。
- 生成 HTML 报告或更新以便发布给项目团队。

CRISP-DM 帮助。

IBM SPSS Modeler 对非专有 CRISP-DM 流程模型提供在线指南。该指南按项目阶段组织，可以提供下列支持:

- 每个 CRISP-DM 阶段的概述和任务列表
- 有关对各种重要事件生成相应报告的帮助
- 说明项目团队如何使用 CRISP-DM 获取数据挖掘帮助的实际应用示例
- 指向其他 CRISP-DM 相关资源的链接

您可以通过从主窗口的“帮助”菜单选择 **CRISP-DM 帮助**访问 CRISP-DM 帮助。

其他资源

除了由 IBM SPSS Modeler 对 CRISP-DM 提供的支持之外，您还可以通过多种其他方式增进您对数据挖掘流程的了解。

- 访问 CRISP-DM Consortium 的网站: www.crisp-dm.org
- 阅读由 CRISP-DM Consortium 制作并随此版本提供的《CRISP-DM 手册》。
- 阅读《*Data Mining with Confidence*》，版权 2002 归 SPSS Inc. 所有；ISBN 1-56827-287-1。

第 2 章 商业了解

业务理解概述

在使用 IBM SPSS Modeler 之前，您应该花些时间探讨一下您的组织期望通过数据挖掘获得什么。尽可能多地请一些重要人员参与此类讨论并将结果记录下来。这个 CRISP-DM 阶段的最后一个步骤是讨论如何使用在此阶段收集的信息生成一份项目计划。

尽管此项研究看起来似乎可有可无，但实际并非如此。了解进行数据挖掘的业务原因有助于确保在花费宝贵的资源之前所有人都达成一致意见。

确定业务目标

您的第一个任务就是尝试尽可能多地了解数据挖掘的业务目标。这可能并没有看起来这么简单，但通过详细说明问题、目标和资源，您可以将今后的风险降至最低。

CRISP-DM 方法为您提供了一种结构化的方法实现这一目的。

任务列表

- 开始收集有关当前业务情况的 背景信息。
- 记录由关键决策者决定的具体业务目标。
- 一致同意用于确定从业务角度判定 数据挖掘成功与否的标准。

电子商务零售业示例 - 寻找业务目标

使用 CRISP-DM 的 Web 挖掘方案

随着越来越多的公司通过 Web 开展销售业务，对于一家已创立的计算机/电子设备电子商务零售商来说，所面临的来自新网站的竞争日益加剧。此公司所面临的现实是网络商店如雨后春笋般快速（或飞速!）崛起，其速度远超过客户迁移到网上的速度，而尽管获取客户的成本不断增长，公司也必须想方设法保持盈利。一个提议的解决方案是培养现有客户关系，以便最大程度地发挥当前每个公司客户的价值。

因此，研究具有以下目标:

- 通过提供更好的推荐增加交叉销售的数量。
- 通过提供更个性化的服务提高客户的忠诚度。

如果达到以下目标，研究将暂时视为成功:

- 交叉销售增长 10%。
- 客户每次访问时在网站停留的时间更长，并且观看了更多的页面。
- 此项研究在不超出预算的情况下按时完成。

编译业务背景资料

理解组织的业务情况有助于您了解在以下这些方面需要解决什么问题:

- 可用资源（人力资源和物资）
- 问题

- 目标

您将需要对当前商业情况进行一些研究，以便找到对影响数据挖掘项目结果的那些问题的正确答案。

任务 1 - 确定组织结构

- 建立组织结构图来说明企业分公司、部门和项目团队的结构。请确保包含管理者的名字和职责。
- 识别组织中的关键个人。
- 识别将提供财务支持和/或领域专门知识的内部负责人。
- 确定是否存在指导委员会并制作一份成员列表。
- 识别将受到数据挖掘项目影响的业务单位。

任务 2 - 说明存在问题的领域

- 识别存在问题的领域，例如市场营销、客户服务或业务发展。
- 使用常规术语来描述问题。
- 阐明项目的先决条件。项目背后的动机是什么？企业是否已经在使用数据挖掘？
- 检查业务团队内数据挖掘项目的状态。努力是否得到认可，或者是否需要将数据挖掘作为业务团队的关键技术进行“通告”？
- 如果需要，请准备有关您的组织进行数据挖掘的信息演示文稿。

任务 3 - 说明当前的解决方案

- 说明当前用于解决业务问题的所有解决方案。
- 说明当前解决方案的优点和缺点。此外，指出这个解决方案在组织内的接受程度。

定义业务目标

本部分将介绍如何使目标具体化。作为您的研究和会议的结果，您应该拟定一个主要具体目标，并得到项目负责人和受结果影响的其他业务单位的一致同意。这个目标将最终从模糊的概念“减少客户流失”转变为可以指导您进行分析的具体数据挖掘目标。

任务列表

请确保记录以下几点以便以后写入项目计划。记住目标要符合实际。

- 说明您希望使用数据挖掘解决的问题。
- 尽可能准确地指出所有业务问题。
- 确定任何其他业务要求（例如在增加交叉销售机会的同时不丢失任何现有客户）。
- 使用业务术语指定预期收益（例如高价值客户流失减少 10%）。

业务成功标准

眼前的目标可能很清晰，但您能否在达到这一目标时知道已经达到该目标？在继续推进之前，定义数据挖掘项目的业务成功特征很重要。成功标准分为两类:

- **目标。**这些标准可能很简单，例如审核准确度或商定的流失率减少值具体提高了多少。
- **主观标准。**主观标准（如“发现一组有效解决方案”）比较难于确定，但你们可以商定由谁进行最终决策。

任务列表

- 尽可能准确地记录此项目的成功标准。
- 确保每个业务目标都有相关的成功标准。

- 调整决定者的主观成功衡量标准使其一致。如果可能，记录下他们的期望值。

评估情况

既然您已经有了一个明确定义的目标，那么现在应该评估您处于什么情况。这一步骤需要询问一些问题，例如:

- 什么种类的数据可供分析？
- 您是否具有完成此项目所需的人员？
- 所涉及的最大风险因素是什么？
- 对于这些风险，您是否具有相应的应急计划？

电子商务零售业示例 - 评估情况

使用 CRISP-DM 的 Web 挖掘方案

这是电子设备的电子商务零售商首次尝试 Web 挖掘，而且该公司已决定聘请一个数据挖掘专家担任顾问以帮助他们入门。这位顾问面临的首要任务之一是评估该公司的数据挖掘资源。

人员。明确的一点是公司内部有具有管理服务器日志和产品以及采购数据库的专家，但他们在用于分析的数据仓库和数据清理方面却没有什么经验。因此，还需要咨询一名数据库专家。由于公司希望研究的结果将成为连续 Web 挖掘过程的一部分，管理层还必须考虑当前工作中产生的任何职位是否为永久性职位。

数据。由于这是一家成立多年的公司，要提取的 Web 日志和采购信息数据量非常多。实际上，对于这个初始的研究，公司将限定只分析那些已经在网站“注册”的客户。如果成功了，再扩展项目。

风险。除了聘请顾问的开销和员工花费在研究上的时间之外，并没有很多与此事件相关的直接风险。但是，时间始终都是重要的，因此这个初期项目计划在一个财政季度内完成。

此外，目前公司的额外现金流并不多，因此此项研究一定不能超出预算。如果实现其中任何一个目标存在风险，业务经理就会建议缩小项目的范围。

资源清单

获取准确的资源清单是必不可少的步骤。通过实际查看硬件、数据源和人力资源问题，您可以节省很多时间以及避免很多头疼的问题。

任务 1 - 调查硬件资源

- 您需要支持什么硬件？

任务 2 - 识别数据源和知识存储

- 哪些数据源可用于数据挖掘？记录数据类型和数据格式。
- 采用什么方式存储数据？您是否可以对数据仓库或操作数据库进行实时访问？
- 您是否计划购买外部数据，例如人口统计信息？
- 是否存在任何让您无法访问所需数据的安全问题？

任务 3 - 识别人力资源

- 您是否能找到业务和数据专家？
- 您是否确定了数据库管理员以及可能会需要的其他技术支持人员？

一旦您询问了这些问题，请在阶段报告中包含一个联系人和资源列表。

要求、假设和约束

如果您真实地评估了项目的负载情况，您的努力获得回报的可能性就更大。尽可能清楚地阐明这些利害关系，这将有助于预防未来出现问题。

任务 1 - 确定要求

最基本的要求就是之前讨论过的业务目标，但需要考虑下面这些问题:

- 对于数据或项目结果，是否存在安全或法律方面的限制？
- 是否所有人都已对项目计划要求达成共识？
- 是否存在任何对结果部署的要求（例如，发布到网上或将评分读取到数据库中）？

任务 2 - 说明假设

- 是否存在可能影响项目的经济因素（例如，咨询费或竞争产品）？
- 是否存在对数据质量的假设？
- 项目负责人/管理团队期望采用什么方式查看结果？换句话说，他们是希望了解模型本身，还是只想看到结果？

任务 3 - 验证约束

- 您是否具有数据访问所需的所有密码？
- 您是否验证过所有对数据使用的法律约束？
- 所有财务约束是否都在项目的预算内？

风险和或有费用

考虑项目进行中可能会遇到的风险是一种明智的做法。风险的类型包括:

- 计划（如果项目花费的时间比预期时间长怎么办？）
- 财务（如果项目负责人遇到预算问题怎么办？）
- 数据（如果数据质量较差或者范围过窄怎么办？）
- 结果（如果初期结果达不到预期怎么办？）

当您考虑了各种风险之后，制定一个应急计划以帮助避免失败。

任务列表

- 记录下每种可能遇到的风险。
- 记录每种风险的相应应急计划。

术语

为了确保业务和数据挖掘团队“说同一种语言”，您应该考虑为技术术语和需要解释的专门用语编写一个词汇表。例如，如果“流失”对于您的业务具有特殊且独特的意思，就值得为了整个团队的利益对其进行明确说明。同样，团队还会受益于对收益图的使用说明。

任务列表

- 在表中记录术语或团队成员容易混淆的行话。包括业务和数据挖掘术语。
- 考虑在公司内部网或其他项目文档中发布此列表。

成本/收益分析

这一步请回答问题: **您的底线是什么** ？作为最终评估的一部分，将项目成本和潜在的成功收益进行比较非常重要。

任务列表

将下列估计成本包括在您的分析中:

- 数据收集和使用的任何外部数据
- 结果部署
- 运营成本

然后，考虑下列收益:

- 要达到的主要目标
- 其他通过数据探索获得的深入见解
- 因深刻理解数据而可能获得的收益

确定数据挖掘目标

既然已经明确了业务目标，现在应该将其转换为数据挖掘实体。例如，“减少流失”的业务目标可以转换为包含下列信息的数据挖掘目标:

- 基于最近的采购数据识别高价值客户
- 使用可用的客户数据构建一个模型，用于预测每个客户的流失可能性
- 基于流失倾向和客户价值为每个客户指定等级

这些数据挖掘目标（如果达到）可以随即被企业用于减少最有价值客户的流失。

正如您所见到的，业务和技术必须紧密配合才能获得有效的数据挖掘。请继续向下读，下面将向您提供如何确定数据挖掘目标的详细提示。

数据挖掘目标

当您与业务和数据分析师一起共同制定业务问题的技术解决方案时，请确保方案比较具体。

任务列表

- 描述数据挖掘问题的 **类型** ，如聚类、预测或分类。
- 使用具体的时间单位记录技术目标，例如预测在三个月内有效。
- 如果可能，为所需结果提供实际的数字，例如为 80% 的现有客户生成流失评分。

电子商务零售业示例 - 数据挖掘目标

使用 CRISP-DM 的 Web 挖掘方案

在数据挖掘顾问的帮助下，该电子商务零售商已经能够将公司的业务目标转换为数据挖掘术语。初期研究要在本季度内完成的目标是:

- 使用先前的历史记录采购信息生成一个包含“相关”链接项目的模型。当用户查看某个项目说明时，为其提供指向相关组中其他项目的链接（ **市场购物篮分析** ）。
- 使用 Web 日志确定不同的客户分别尝试查找哪些项目，然后对网站进行重新设计，以突出显示这些项目。每个不同的“类型”的客户将看到不同的网站主页（ **客户分类** ）。

- 通过给出他或她来自哪里以及去过网站中的什么地方（ **序列分析** ），使用 Web 日志来尝试预测客户接下来要去哪里。

数据挖掘成功标准

此外，还必须使用技术术语来定义成功，以便随时了解数据挖掘工作的进度。使用之前确定的数据挖掘目标来明确说明成功的基准。IBM SPSS Modeler 提供多种工具（如评估图表节点和分析节点）来帮助您分析结果是否正确和有效。

任务列表

- 描述模型评估方法（例如，准确度、性能等）。
- 定义评估成功的基准。提供具体的数字。
- 尽可能详细地定义主观衡量标准，并确定成功的决定者。
- 考虑成功部署模型结果是否算是数据挖掘成功的一部分。立即开始对部署进行计划。

制定项目计划

现在，您可以制定数据挖掘项目的计划了。您之前询问的那些问题以及详细制定的业务和数据挖掘目标将作为这个路线图的基础。

编写项目计划

项目计划是适用于所有数据挖掘工作的主要文档。如果计划制定得好，它可以为每个项目相关人员提供各个数据挖掘阶段的目标、资源、风险以及计划等信息。您可能希望在公司内部网中发布此计划，同时发布这个阶段收集到的所有文档。

任务列表

创建计划时，请确保您已经回答了下面这些问题:

- 您是否已经和所涉及的每个人讨论了项目任务和提议的计划？
- 是否所有阶段或任务都包含估计的时间？
- 您是否包含了部署结果或业务解决方案所需的工作量和资源？
- 计划中是否突出显示了决策点和审核请求？
- 您是否已经标记出通常会发生多个迭代的阶段，例如建模阶段？

项目计划样本

研究的总体计划如下表所示:

表 1. 样本项目计划概述

Phase	**时间**	**Resources**	**风险**
业务理解	1 周	所有分析师	经济环境变化
数据理解	3 周	所有分析师	数据问题，技术问题
数据准备	5 周	数据挖掘顾问，数据库分析师的一些时间	数据问题，技术问题
建模	2 周	数据挖掘顾问，数据库分析师的一些时间	技术问题，无法找到合适的模型

表 1. 样本项目计划概述 (续)

Phase	时间	Resources	风险
评估	1 周	所有分析师	经济环境变化，无法实施结果
部署	1 周	数据挖掘顾问，数据库分析师的一些时间	经济环境变化，无法实施结果

评估工具和技术

由于您已经选择使用 IBM SPSS Modeler 作为您的数据挖掘成功的工具，您可以使用这个步骤来调查哪些数据挖掘技术最能满足您的业务需要。 IBM SPSS Modeler 为每个数据挖掘阶段都提供了全面的工具系列。要决定什么时候使用哪一种技术，请参阅联机帮助的建模部分。

准备好进入下一个步骤了吗?

在探索数据并开始使用 IBM SPSS Modeler 之前，请确保您已经回答了下面这些问题。

从业务角度:

- 您的企业希望从此项目中获得什么？
- 您将如何定义我们是否成功完成了工作？
- 您是否具有实现我们的目标所需的预算和资源？
- 您是否可以访问此项目所需的所有数据？
- 您和您的团队是否讨论过与此项目相关的风险和或有费用？
- 您的成本/收益分析的结果是否表明这个项目值得尝试？

当您回答了上述问题之后，您是否将这些问题的答案转换为数据挖掘目标？

从数据挖掘角度:

- 具体地说，数据挖掘如何帮助您达到业务目标？
- 您对哪种数据挖掘技术可能获得最好的结果是怎么看的？
- 您怎样才能知道您的结果是否已达到期望的准确度或有效性？（*我们是否设定了衡量数据挖掘成功的标准？*）
- 将如何部署建模结果？您是否考虑了项目计划中的部署？
- 项目计划是否包含所有 CRISP-DM 阶段？
- 计划中是否说明了风险和依赖关系？

如果对于上述问题您都可以回答“是”，则表示您已经作好深入了解数据的准备。

第 3 章 数据了解

数据理解概述

CRISP-DM 的数据理解阶段包含深入了解可用于挖掘的数据。此步骤是在下一个阶段（数据准备）中避免意外问题发生的关键，这个后续阶段通常是项目中耗时最长的部分。

数据理解包含使用可以在 IBM SPSS Modeler 中通过 CRISP-DM 项目工具组织的表格和图形访问数据以及探索数据。在这一阶段中，您可以确定数据的质量并在项目文档中描述这些步骤的结果。

收集初始数据

此时在 CRISP-DM 中，您已作好访问数据并将其带入 IBM SPSS Modeler 中的准备。数据来自各种不同的数据源，例如:

- **现有数据。**这包括各种不同的数据，例如事务处理数据、调查数据、Web 日志等。请考虑现有数据是否足以满足您的需要。
- **购买的数据。**您的组织是否使用补充性数据（例如人口统计数据）？如果没有，请考虑是否需要使用此类数据。
- **其他数据。**如果上面的数据源并不能满足您的需求，您可能需要开展调查或开始进行其他跟踪以便补充现有的数据存储。

任务列表

查看 IBM SPSS Modeler 中的数据然后考虑以下问题。请确保记录下您发现的问题。请参阅主题第 12 页的『编写数据收集报告』，了解更多信息。

- 数据库中的哪些属性（列）看起来最有用？
- 哪些属性看起来并不相关，可以排除在外？
- 要想得出概括的结论或者做出准确的预测，现有数据是否足够？
- 您所选的建模方法是否存在过多属性？
- 您是否要合并不同的数据源？如果要合并，是否存在合并时会引发问题的区域？
- 您是否考虑过如何处理各个数据源中的缺失值？

电子商务零售业示例 - 初始数据收集

使用 CRISP-DM 的 Web 挖掘方案

这个示例中的电子商务零售商使用多个重要的数据源，包括:

Web 日志。原始访问日志包含与客户如何浏览 Web 站点相关的所有信息。作为数据准备过程的一部分，将需要移除对 Web 日志中的图像文件以及其他非信息条目的引用。

采购数据。当客户提交订单时，将会保存与该订单相关的所有信息。采购信息数据库中的订单需要映射到 Web 日志的相应会话中。

产品数据库。在确定“相关”产品时，产品属性将会很有用。产品信息需要映射到相应的订单中。

客户数据库。此数据库包含收集自注册客户的附加信息。这些记录绝不会是完整的，因为许多客户都没有填写调查问卷。客户信息需要映射到 Web 日志的相应购买信息和会话中。

目前，公司没有购买外部数据库或花钱开展调查的计划，因为分析师们正在忙于处理他们当前所具有的数据。但是，在某些时候，他们可能想要考虑扩大数据挖掘结果的部署规模，在此情况下购买其他对未注册客户的人口统计数据相当有用。此外，使用人口统计信息来了解此电子商务零售商的客户群与普通网络购物者之间的区别也是很有用的。

编写数据收集报告

使用上述步骤中收集的材料，您可以开始编写数据收集报告。一旦完成，可将此报告添加到项目 Web 站点或向项目团队发布。它也可以与后续步骤中准备的报告组合在一起，如数据说明、探索和质量验证。这些报告将在整个数据准备阶段指导您的工作。

描述数据

您可以采用多种方式对数据进行描述，但是大多数描述都将重点放在数据的数量和质量上，即可提供多少数据以及这些数据的具体情况。以下列出了描述数据时需要用到的一些关键特征。

- **数据量。**对于大多数建模技术，数据大小都具有相关的协定。大型数据集可以生成更准确的模型，但它们也会增加处理时间。考虑是否可以使用数据的一个子集。当为最终报告记录信息时，请确保包括所有数据集的大小统计数据量，并且记住在描述数据时考虑记录和字段（属性）的数量。
- **值类型。**数据可以采用多种格式，例如**数字**、**分类**（字符串）或**布尔值** (true/false)。注意值类型可以防止在后面的建模阶段出现问题。
- **编码方案。**数据库中的值常用于表示特征，如性别或产品类型。例如，一个数据集可以使用 *M* 和 *F* 来表示 *男性* 和 *女性* ，此外也可以使用数字值 *1* 和 *2* 表示。请注意数据报告中的那些冲突的方案。

掌握了这项知识，您现在可以编写 数据说明报告并且与广大读者共享您发现的问题。

电子商务零售业示例 - 描述数据

使用 CRISP-DM 的 Web 挖掘方案

在 Web 挖掘应用的过程中，需要处理许多记录和属性。尽管电子商务零售商执行此数据挖掘项目时将初始研究限定为大约 30,000 名已经在站点中注册的客户，Web 日志中仍有数百万条记录。

这些数据源中的大多数为符号类型的值，它们分别是日期和时间、访问过的网页或者是注册调查问卷中的多选问题的答案。这些变量中有部分变量将用于创建新的数字变量，例如浏览过的网页数量以及花费在网站中的时间。数据源中很少的现有数字变量包含每个产品订购的数量、一次采购所花费的金额以及来自产品数据库的产品重量和尺寸规格。

不同数据源的编码方案几乎没有重叠，因为这些数据源均包含非常不同的属性。唯一重叠的变量是“关键字”，例如客户标识和产品代码。不同的数据源中的这些变量必须具有完全相同的编码方案；否则，将无法合并这些数据源。还将需要一些数据准备工作，用于记录这些要合并的关键字段。

编写数据说明报告

要有效地推进您的数据挖掘项目，请考虑使用下列度量标准生成准确数据说明报告的值:

数据数量

- 数据的格式是什么？
- 指定用于捕获数据的方法，例如，ODBC。

- 数据库有多大（使用行数和列数描述）？

数据质量

- 数据是否包含与业务问题相关的特征？
- 所呈现的是什么数据类型（符号、数字等）？
- 您是否为关键属性计算了基本统计数据？这些数据为业务问题提供了哪些深入的见解？
- 您是否能够为相关的属性设置优先级？如果不能，业务分析师是否可以提供进一步的见解？

探索数据

使用这个 CRISP-DM 阶段通过 IBM SPSS Modeler 中提供的表格、图表和其他可视化工具来探索数据。此类分析可以帮助解决在 业务理解阶段构建的数据挖掘目标。它们还可以帮助用于设定假设以及制定将在数据准备阶段进行的数据转换任务。

电子商务零售业示例 - 探索数据

使用 CRISP-DM 的 Web 挖掘方案

尽管 CRISP-DM 建议在此时执行一次初始探索，但正如我们的电子商务零售商发现的那样，在原始 Web 日志上进行数据探索是一件艰难的工作（如果不是不可能）。通常，Web 日志数据必须先在数据准备阶段进行处理，以便生成具有实际探索意义的数据。这个与 CRISP-DM 不符的操作强调处理过程可以而且应该自定义以便满足您的特定数据挖掘需要。CRISP-DM 具有循环性，数据挖掘工作人员通常会在各个阶段之间来回穿梭。

尽管 Web 日志必须在进行探索之前进行处理，可供此电子商务零售商使用的其他数据源则更服从于探索。使用采购信息数据库进行探索可以揭示有趣的客户汇总信息，例如他们消费了多少钱、他们每次购物时购买了多少项目以及他们来自什么地方。客户数据库的汇总信息将显示对注册调查问卷项目的回答分布情况。

探索对于查找数据中的错误同样有用。尽管大多数数据源都是自动生成的，但产品数据库中的信息却是手动输入的。一些列表产品尺寸的快速摘要将有助于您发现打字错误，例如“119-inch”（而不是“19-inch”）监视器。

编写数据探索报告

当您创建图形并对可用数据进行统计时，应该开始设定数据如何才能解决技术和业务目标的假设。

任务列表

记录您发现的问题以便将其包含在数据探索报告中。请确保回答以下问题:

- 您对数据设定了什么类型的假设？
- 哪些属性看起来对于进一步的分析有用？
- 您的探索是否揭示了新的数据特征？
- 这些探索怎样改变了您的初始假设？
- 您是否能标识特定的数据子集以供过后使用？
- 再次查看一下您的数据挖掘计划。此次探索是否更改了目标？

验证数据质量

数据几乎没有完美的。事实上，大多数数据都包含代码错误、缺失值或其他类型的不一致现象，这往往让分析很棘手。一种可避免可能出现缺陷的方法是在建模前对可用数据进行全面的质量分析。

IBM SPSS Modeler 中的报告工具（例如，数据审核、表格与其他输出节点） 可以帮助您查找以下这些类型的问题:

- **缺失数据**包括空白值或编码为无应答的值（例如 *$null$*、*?* 或 *999*）。
- **数据错误**通常是在输入数据时造成的拼写错误。
- **度量标准误差**包括正确输入但基于不正确的度量方案的数据。
- **编码不一致**通常包括非标准度量单位或不一致的值，例如同时使用 *M* 和 *male* 表示性别。
- **元数据不正确**包括字段的表面意思和字段名称或定义中陈述的意思不匹配。

请务必记录此类质量问题。请参阅主题『编写数据质量报告』，了解更多信息。

电子商务零售业示例 - 验证数据质量

使用 CRISP-DM 的 Web 挖掘方案

验证数据质量通常在说明和探索数据的处理过程中完成。电子商务零售商可能会遇到的一些问题包括:

缺失数据。已知的缺失数据包括部分注册用户没有回答的调查问卷。不具有调查问卷提供的这些额外信息，这些客户可能必须排除在部分后续模型之外。

数据错误。大多数数据源都是自动生成的，因此这并不需要太过担心。产品数据库中的拼写错误可以在数据探索过程中发现。

度量标准错误。最有可能存在度量标准误差的数据源是调查问卷。如果这些项目中有任何项目是未经仔细考虑或用词不当的，它们可能无法提供电子商务零售商希望获得的信息。而且，在探索过程中，特别关注那些答案分布异常的项目也很重要。

编写数据质量报告

基于您对数据质量的探索和验证，您现在可以开始准备将用于指导下一个 CRISP-DM 阶段的报告。请参阅主题第 13 页的『验证数据质量』，了解更多信息。

任务列表

正如之前讨论的，存在 多种类型的数据质量问题。在进入下一个步骤之前，请考虑下列质量问题并规划解决方案。将所有答复记录在数据质量报告中。

- 您有没有找到任何缺失属性和空白字段？如果找到了，此类缺失值是否暗含什么意思？
- 是否存在可能会在后面的合并或转换的过程中导致问题的拼写前后不一致的情况？
- 您是否探索了偏差值以确定它们是“无效数据”还是值得进一步分析的现象？
- 您是否对值执行了真实性检查？记录下所有明显的冲突（例如青少年具有高收入）。
- 您是否考虑过将那些对您的假设没有任何影响的数据排除在外？
- 数据是否存储在平面文件中？如果是，这些文件中的定界符是否一致？每条记录是否都包含相同数量的字段？

准备好进入下一个步骤了吗?

在为在 IBM SPSS Modeler 中建模而准备数据之前，请考虑下面这几点问题:

您对数据了解到什么程度？

- 是否所有数据源都已清楚标识和访问？您是否知道存在的任何问题或限制？
- 您是否已经在可用数据中标识了关键属性？
- 这些属性是否有助于您设定假设？
- 您是否记录下了所有数据源的大小？
- 您是否可以在适当的地方使用数据的子集？
- 您是否为所感兴趣的每个属性计算了基本统计数据？是否有具有意义的信息显现？
- 您是否使用探索图形来获取关键属性的深入见解？这些深入的见解是否使您对自己的所有假设进行了重新设定？
- 这个项目的数据质量问题是什么？您是否有解决这些问题的计划？
- 数据准备步骤是否明确？例如，您是否知道要合并哪些数据源以及要过滤或选择哪些属性？

既然您已经在业务和数据理解阶段都作好了准备，现在该轮到使用 IBM SPSS Modeler 来准备建模使用的数据了。

第 4 章 数据准备

数据准备概述

数据准备是数据挖掘最重要的阶段之一，通常需要花费大量的时间。据估计，实际的数据准备工作通常占 50-70% 的项目时间和工作量。在前期的 业务理解和 数据理解阶段投入足够的精力可以将对这一阶段的投入降至最低，但您仍需花费大量的精力为挖掘准备和打包数据。

取决于您的组织及组织目标，数据准备通常包含以下任务:

- 合并数据集和/或记录
- 选择数据子集样本
- 汇总记录
- 导出新的属性
- 排序数据以便建模
- 删除或替换空白值或缺失值
- 分为训练数据集和测试数据集

选择数据

基于在前面的 CRISP-DM 阶段执行的 初始数据收集，您可以开始选择与您的数据挖掘目标相关的数据。通常，有以下两种选择数据的方式:

- **选择项（行）**包括各种决策的制定，例如要包含哪些帐户、产品或客户。
- **选择属性或特征（列）**包括有关使用哪些特征的决策制定，例如交易金额或家庭收入。

电子商务零售业示例 - 选择数据

使用 CRISP-DM 的 Web 挖掘方案

电子商务零售商的许多关于选择什么数据的决策在数据挖掘流程的早期阶段早已确定。

选择项。初始研究的对象将限于已经在站点中注册的（大约）30,000 名客户，因此需要设置一些过滤器，以便将非注册客户的购买日志和 Web 日志排除在外。还应建立一些其他过滤器，用于移除对 Web 日志中的图像文件以及其他非信息条目的调用。

选择属性。采购数据库将包含有关电子商务零售商的敏感客户信息，因此过滤掉类似客户姓名、地址、电话号码和信用卡号码等属性信息非常重要。

包括或排除数据

在您决定要包括或排除哪些数据子集的时候，请确保记录下做出这些决定的根本原因。

要考虑的问题

- 某个给定的属性是否与您的数据挖掘目标相关？
- 某个特定数据集或属性的质量是否会导致您的结果无效？
- 您是否能对此类数据进行数据挽救？

- 对于使用某些特定字段，如 *性别* 或 *种族* 是否存在任何限制？

您在此阶段所作的决定是否与您在数据理解阶段所作的假设不同？如果不同，请确保在项目报告中记录下您的原因。

清理数据

清理数据包括深入了解您选择包含在分析中的数据存在的问题。您可以通过多种方式使用 IBM SPSS Modeler 的记录和字段操作节点来清理数据。

表 2. 清理数据

数据问题	可能的解决方案
缺失数据	排除行或特征。或者，使用估计值填充空白值。
数据错误	通过逻辑关系手动发现错误并进行替换。或者，排除特征。
编码不一致	决定使用其中一种编码方案，然后转换及替换相应的值。
缺失或无效的元数据	手动检测可疑字段并追踪其正确的意思。

在数据理解阶段准备的 数据质量报告包含您的数据的特定问题类型的详细信息。您可以使用它作为在 IBM SPSS Modeler 中进行数据操作的起始点。

电子商务零售业示例 - 清理数据

使用 CRISP-DM 的 Web 挖掘方案

此电子商务零售商使用数据清理过程来解决数据质量报告中指出的问题。

缺少数据。在以后创建的部分模型中，可能必须将没有完成在线调查问卷的客户排除在外。电子商务零售商可以让这些客户再次填写此调查问卷，但这样做所花费的时间和费用此零售商均负担不起。此电子商务零售商所能做的是构建回答和不回答此调查问卷的客户的购买行为差异模型。如果这两组客户具有相似的购买习惯，则无需担心缺失的调查问卷。

数据错误。在探索过程中发现的错误可以在此阶段更正。尽管多数情况下，在客户将页面提交到后端的数据库中之前，网站均强制要求输入正确的数据。

测量错误。调查问卷中存在用词不当的项目会极大地影响数据的质量。与缺失的调查问卷一样，这也是一个难以解决的问题，因为没有更多的时间和金钱可用于收集重新更换过问题的答案。对于存在问题的项目，最佳的解决方案将是返回选择过程并从后期的分析中过滤掉这些项目。

编写数据清理报告

报告您的数据清理成果对于跟踪数据的更改是必不可少的步骤。轻松掌握工作的详细信息将有助于将来的数据挖掘项目。

任务列表

编写报告时考虑以下问题是不错的选择:

- 数据中产生了哪些类型的无用数据？
- 您使用什么方法删除这些无用数据？哪些技术获得了成功？
- 是否存在无法挽救的情况或属性？请确保记录因无用数据而排除的数据。

构建新数据

您经常会遇到需要构建新数据的情况。例如，创建一个新列用于标记每项交易是否购买了延长保修是非常有用的。这个新字段，*purchased_warranty*，可以通过 IBM SPSS Modeler 中的“设为标志”节点轻松生成。

有以下两种构建新数据的方式:

- 导出属性（列或特征）
- 生成记录（行）

IBM SPSS Modeler 使用其记录和字段操作节点提供多种构建数据的方式。

电子商务零售业示例 - 构建数据

使用 CRISP-DM 的 Web 挖掘方案

处理 Web 日志时会生成许多新的属性。对于日志中记录的事件，电子商务零售商将希望创建时间戳，用于标识访问者和会话，并记录所访问的页面和事件所代表的活动类型。这些变量中的一部分将用于创建更多的属性，例如会话中事件之间的时间。

合并或其他数据重新结构化操作也会创建更多的属性。例如，当对每行一个事件的 Web 日志进行“累计”以便每行表示一个会话时，将会创建一些新的属性用于记录操作的总数、花费的时间总量以及会话期间的总购买量。当 Web 日志与客户数据库合并以便每行表示一个客户的数据时，将会创建一些新的属性用于记录会话数、操作总数、花费的时间总量以及每个客户的总购买量。

构建新数据之后，电子商务零售商执行了一次探索过程以确保正确执行了数据创建过程。

导出属性

在 IBM SPSS Modeler 中，您可以使用以下字段操作节点导出新属性:

- 使用 **导出节点** 创建源自现有字段的新字段。
- 使用 **设为标志节点** 创建标志字段。

任务列表

- 考虑导出属性时建模的数据要求。建模算法是否具有预期的特定数据类型，例如数字？如果有，请执行必要的转换。
- 建模之前是否需要对数据进行标准化？
- 缺失的属性是否可以使用汇总、求平均值或归纳来构建？
- 基于您的背景知识，是否能够从现有字段生成任何重要的事实（例如在 Web 站点中花费的时间长度）？

集成数据

同一组业务问题具有多个数据源的情况很多见。例如，您可以访问同一组客户的抵押贷款数据以及购买的人口统计数据。如果这些数据集包含相同的唯一标识（如社会保险号），您可以在 IBM SPSS Modeler 中使用这个关键字段将它们合并在一起。

合并数据的基本方法有以下两种:

- **合并**数据，涉及合并两个具有相似记录但不同属性的数据集。这些数据通过各记录的相同关键标识（例如客户标识）合并。生成的数据将会增加一些列或特征。

- **追加**数据，涉及集成两个或多个具有相似属性但不同记录的数据集。数据基于相似字段（例如产品名称或合同时长）集成。

电子商务零售业示例 - 集成数据

使用 CRISP-DM 的 Web 挖掘方案

具有多个数据源时，电子商务零售商可使用多种方法集成数据:

- **在事件数据中添加客户和产品属性。**为了使用其他数据库的属性创建 Web 日志事件模型，必须正确标识与每个事件相关的任意客户标识、产品编号和采购单编号，而且相应的属性也必须合并到经过处理的 Web 日志中。请注意，合并的文件会在每次客户或产品与事件关联时复制客户和产品信息。
- **在客户数据中添加购买信息和 Web 日志信息。**为了构建客户价值模型，必须从相应的数据库中挑选出他们的采购和会话信息，在合计之后与客户数据库相合并。这包括之前在构建数据过程中讨论过的创建新属性操作。

集成数据库之后，电子商务零售商执行了一次探索过程以确保正确执行了数据合并过程。

集成任务

如果您没有花费足够的时间开发和理解您的数据，集成数据将会变得很复杂。更多地思考一下那些看上去与数据挖掘目标关系最大的项目和属性，然后开始集成您的数据。

任务列表

- 使用 IBM SPSS Modeler 中的合并或追加节点，集成那些您认为对于建模有用的数据集。
- 考虑在建模之前保存生成的输出。
- 合并之后，可以通过 **汇总** 值简化数据。汇总表示通过总结多条记录和/或表中的信息计算出新值。
- 此外，您也可能需要生成一些新记录（例如多年联合退税的平均减免额）。

格式化数据

作为建模前的最后一个步骤，检查某些特定技术是否需要数据具有特定格式或顺序很有用。例如，某种序列算法要求数据在运行模型前预先排序的情况很常见。即使模型可以执行排序操作，但是在建模前使用排序节点可以节省处理时间。

任务列表

格式化数据时请考虑下列问题:

- 您计划使用哪些模型？
- 这些模型是否需要特定的数据格式或顺序？

如果建议进行更改， IBM SPSS Modeler 的处理工具可以帮助您应用必要的数据操作。

作好建模的准备了吗?

在 IBM SPSS Modeler 中构建模型之前，请确保您已经回答了以下这些问题。

- 是否所有数据都可以从 IBM SPSS Modeler 中访问？
- 基于您对数据的初始探索和理解，是否能够选择相关的数据子集？
- 您是否已经进行了有效的数据清理或者是否已经删除了无法挽救的项目？在最终报告中记录所有决定。
- 多个数据集是否已经正确集成？还有没有应该记录在案的合并问题？

- 您是否仔细研究过您希望使用的建模工具的要求？
- 在建模之前，是否还有任何可以解决的格式化问题？这既包括对需要进行格式化的担心，也包括可能减少建模时间的任务。

如果您可以回答上述问题，则您已经准备好可以进入数据挖掘的核心阶段 - 建模。

第 5 章 建模

建模概述

这是您的努力工作开始有所回报的阶段。您花费时间准备的数据将导入到 IBM SPSS Modeler 的分析工具中，此时这些结果开始表现在 业务理解阶段呈现的业务问题。

建模时通常会执行多次迭代。通常，数据挖掘人员会使用缺省参数运行多个模型，然后再对这些参数进行微调或回到数据准备阶段以便执行所选模型所需的操作。仅使用一个模型且仅执行一次就能圆满地解答组织的数据挖掘问题，这样的情况几乎不存在。这就是数据挖掘如此有趣的原因，您可以使用多种方法来考虑某个已知的问题，而 IBM SPSS Modeler 为您提供了助您实现此目的的多种工具。

选择建模技术

尽管您可能已经知道哪种类型的建模方式最能满足您组织的需要，但现在应该做出有关使用哪些建模方式的正式决定。通常，将会基于下列因素确定最适用的模型:

- **可供挖掘的数据类型。**例如，感兴趣的字段是否为分类（符号型）？
- **您的数据挖掘目标。**您是否只想获取有关事务处理数据存储的深入见解并挖掘出令客户感兴趣的购买模式？或者您是否需要生成一个评分，例如用于表明拖欠学生贷款的倾向？
- **具体的建模要求。**模型是否要求使用特定的数据大小或类型？您是否需要一个具有易于演示的结果的模型？

有关 IBM SPSS Modeler 中的模型类型及其要求的更多信息，请参阅 IBM SPSS Modeler 文档或联机帮助。

电子商务零售业示例 - 建模技术

此电子商务零售商使用的建模技术由公司的数据挖掘目标决定:

改进建议。简单来说，这包括将采购单聚类以确定哪些是最经常同时购买的产品。客户数据，甚至是访问记录，都可以添加到其中以便获得更丰富的结果。两步聚类或 Kohonen 网络聚类技术适用于此类建模方式。随后，可以使用 C5.0 规则集描述聚类，以确定客户在访问过程中随时得到最适用的建议。

提高站点导航能力。对于现在而言，此电子商务零售商将重点放在标识那些经常使用但用户需要多次点击才能找到的页面。这就必须对 Web 日志应用排序算法以便生成客户在 Web 站点中浏览时的“独有路径”，然后特别查找具有大量页面访问但却没有采取任何操作（或采取操作之前）的会话。过后，在更深入的分析中，聚类技术可用于标识不同“类型”的访问和访问者，然后就可以根据类型整理和展示站点内容。

选择正确的建模技术

在 IBM SPSS Modeler 中提供了许多建模技术。通常，数据挖掘人员使用多种技术从多个不同方向处理问题。

任务列表

当决定要使用哪种（些）模型之后，请考虑以下的问题是否会影响您的选择:

- 此模型是否需要将数据分为测试集和训练集？
- 您是否具有足够的数据为给定的模型生成可靠的结果？
- 此模型是否需要特定的数据质量级别？您的当前数据是否达到这一级别？

- 您的数据是不是适用于此特定模型的恰当类型？如果不是，您是否可以使用数据操控类节点进行必要的转换？

有关 IBM SPSS Modeler 中的模型类型及其要求的更多信息，请参阅 IBM SPSS Modeler 文档或联机帮助。

建模假设

当您开始缩小建模工具的选择范围时，请记录下决策制定过程。记录下所有为了达到模型的要求而设定的数据假设以及为此而执行的数据操作。

例如，Logistic 回归和神经网络节点都要求其数据类型在执行前经过完全**实例化**（数据类型已知）。这就意味着您将需要在流中添加一个类型节点并执行该节点以便在构建和运行模型前全面运行数据。与之类似，预测模型（例如 C5.0）可以在为少见的事件预测规则时，从重新平衡数据中获益。当进行此类预测时，通过在流中插入一个平衡节点并在模型中增加平衡性更强的子集通常可以获得更好的结果。

请确保记录下这些决定的类型。

生成测试设计

作为实际构建模型之前的最后一个步骤，您应该花些时间再次考虑要采用什么方式对模型的结果进行测试。生成一个全面的测试设计操作包含两个部分:

- 描述模型的“优异性”标准
- 定义将要对其测试这些标准的数据

模型的 **优异性** 可以通过多种方法度量。对于监督式模型（例如 C5.0、和 C&R 树）优异性的度量方法通常估计特定模型的错误率。对于非监督式模型，例如 Kohonen 聚类网络，度量方法可以包括易于解释、部署或所需处理时间等标准。

记住，模型构建操作是一个迭代的过程。这意味着您通常需要测试多个模型的结果才能决定使用和部署哪些模型。

编写测试设计

测试设计就是您将用于测试生成的模型的步骤说明。因为建模是一个迭代过程，因此知道何时应该停止调整参数以及尝试另一种方法或模型非常重要。

任务列表

当创建测试设计时，请考虑以下问题:

- 将使用什么数据测试模型？您是否已将数据分为训练/测试集？（这是在建模时常会使用的方法。）
- 您如何度量监督式模型（例如 C5.0）是否成功？
- 您要如何度量非监督式模型是否成功（例如 Kohonen 聚类网络）？
- 您愿意在尝试另一种模型类型前使用调整的设置重新运行多少次模型？

电子商务零售业示例 - 测试设计

使用 CRISP-DM 的 Web 挖掘方案

评估模型的标准取决于正在考虑的模型和数据挖掘目标:

改进建议。在将改善的建议展现给真实的客户观看之前，没有纯粹客观地评估它们的方式。但是，电子商务零售商可能会要求生成建议的规则非常简单（从业务角度看）。同样，这些规则也应该足够复杂，可以为不同的客户和会话生成不同的建议。

提高站点导航能力。如果有客户在 Web 站点上访问哪些页面的证据，此电子商务零售商可以根据是否能够轻松访问重要页面从客观的角度评估更新后的站点设计。但是，对于推荐，很难进一步在事前评估客户对于重新组织后的站点的接受程度。如果时间和经济许可，应该定制一些使用性能测试。

构建模型

此时，您应该已经作好准备，以构建那些经过长时间思考的模型的。在做出最终决定之前，给自己充分的时间和空间去体验不同的模型。大多数数据挖掘人员通常都会在部署或集成模型之前构建多个模型，然后再比较它们的结果。

为了跟踪您处理多个模型的过程，请确保记录下每个模型所使用的设置和数据。这可以在您与其他人讨论这些结果时提供帮助，并且还可以在需要时重新跟踪您的步骤。在模型构建过程的最后阶段，您将获得三类将在数据挖掘决策时使用的信息:

- **参数设置**包含您记录的生成最佳结果的参数。
- 生成的实际 **模型** 。
- **模型结果描述**包含在执行模型并探索其结果时发生的性能和数据问题。

电子商务零售业示例 - 模型构建

使用 CRISP-DM 的 Web 挖掘方案

改进建议。为不同级别的数据集成生成了不同的聚类，由采购信息数据库开始，然后包括相关的客户和会话信息。对于每个集成级别，聚类均在两步算法和 Kohonen 网络算法的不同参数设置下生成。对于这些聚类中的每一个，通过不同的参数设置生成了少数 C5.0 规则集。（对于这些聚类中的每一个，都生成了具有不同参数设置的一些 C5.0 规则集。）

提高站点导航能力。序列建模节点均用于生成客户路径。该算法允许最小支持度标准规范，这在重点处理最常见的客户路径时很有用。这些参数的各种设置都已经过尝试。

参数设置

大多数建模技术都具有大量参数或设置，对这些参数和设置进行调整即可控制建模过程。例如，可用通过调整决策树的深度、分割和一些其他设置对它进行控制。通常情况下，大多数人都会先使用缺省选项构建一个模型，然后再在后续的会话中改进参数。

一旦您确定了可生成最准确结果的参数，请确保保存流和生成的模型节点。此外，记录下最佳设置也可以在您决定使用新数据自动构建或重新构建模型时提供帮助。

运行模型

在 IBM SPSS Modeler 中，运行模型是一项简单的任务。一旦您将模型节点插入流中并编辑任意参数之后，只需执行模型即可生成可查看的结果。结果将显示在工作区右侧的“生成的模型”导航器中。您可以右键单击模型浏览其结果。对于大多数模型，您可以将生成的模型插入流中以便进一步评估和部署结果。模型也可以保存在 IBM SPSS Modeler 中以便轻松重新使用。

模型说明

当检验模型的结果时，请确保记录下您的建模经验。您可以使用节点注解对话框或项目工具将这些记录存储在模型自身中。

任务列表

对于每个模型，记录以下信息:

- 您是否能从此模型得出有意义的结论？
- 此模型是否揭示了新的深入见解或不寻常的模式？
- 模型是否存在执行问题？执行时间是否合理？
- 此模型是否存在数据质量难题，例如具有大量缺失值？
- 有没有应该记录的计算不一致问题？

评估模型

既然您已经具有一组初始模型，请深入了解它们以确定哪些模型既准确又有效，足以成为最终的模型。最终包含多层含义，例如“可以部署”或“展现了用户感兴趣的模式”。参考您之前创建的测试计划可能有助于您从组织的观点出发进行此评估。

综合模型评估

对于每个正在考虑的模型，最好基于测试计划中生成的标准进行一次系统评估。在这里，您可以将生成的模型添加到流中并使用评估图表或分析节点分析结果是否有效。您还应该考虑结果从逻辑上看是否合理或者它们是否对于您的业务目标来说太过简单（例如，所揭示的采购顺序为酒 > 酒 > 酒）。

一旦进行了评估，请基于客观（模型正确性）和主观（易于使用或结果无需解释）标准对模型进行排序。

任务列表

- 使用 IBM SPSS Modeler 的数据挖掘工具，例如评估图表、分析节点或交叉验证图表来评估您的模型结果。
- 基于您对业务问题的理解对结果进行审核。咨询对某个特定结果的相关性具有深入了解的数据分析师或其他专家。
- 考虑某个模型的结果是否易于部署。您的组织是要求将该结果部署到 Web 上还是发送回数据仓库中？
- 分析结果对您的成功标准的影响。它们是否达到您在业务理解阶段建立的目标？

如果您能够成功解决上述问题并相信当前模型达到了您的目标，现在可以开始进一步执行更全面的模型评估并进行最终部署。否则，根据您所学到的知识使用经过调整的参数设置重新运行模型。

电子商务零售业示例 - 模型评估

使用 CRISP-DM 的 Web 挖掘方案

改进建议。Kohonen 网络之一和两步聚类方法均各自给出了合理的结果，而此电子商务零售商发现很难在它们之间作出选择。最后，公司希望两者都使用，接受这两种技术都认可的那些建议并仔细研究它们的不同之处。通过少许努力并应用相应的业务知识，此电子商务零售商可以开发出进一步的规则用于解决这两种技术的不同之处。

此电子商务零售商还发现包含会话信息的结果惊人的好。有证据表明可以将建议与站点导航绑定在一起。当客户浏览时，可以实时使用一个用于定义客户的下一可能目标位置的规则集来直接影响站点的内容。

提高站点导航能力。序列模型为电子商务零售商提供可以预测特定客户路径的高度置信度，生成的结果会建议对站点设计进行可控数量的更改。

跟踪已修正的参数

基于您在模型评估过程中了解的信息，现在应该再次查看一下模型。您此时具有两个选项:

- 调整现有模型的参数。
- 选择另一个模型来解决您的数据挖掘问题。

在这两种情况下，您都将返回构建模型任务并重复执行该任务直至结果成功。不要担心重复执行这一步骤。在找到满足需要的模型之前，数据挖掘人员多次评估和重新运行模型是非常常见的。这是一个在调整多个模型的参数之前用于同时构建这些模型并比较结果的实用参数。

准备好进入下一个步骤了吗?

在进入模型的最终评估阶段之前，请考虑您的初始评估够不够全面。

任务列表

- 您是否能够理解模型的结果？
- 单纯从逻辑角度考虑，这些模型结果是否逻辑通顺？是否存在需要进一步探索的明显不一致情况？
- 从您的第一眼看来，这些结果是否能解决您组织的业务问题？
- 您是否曾经使用分析节点以及提升图或收益图来比较和评估模型的准确度？
- 您是否探索了多种类型的模型并比较了它们的结果？
- 您的模型结果是否可以部署？

如果您的数据建模结果看上去准确并且相互关联，现在应该在最终部署之前执行更全面的评估。

第 6 章 评估

评估概述

现在，您的数据挖掘项目已经完成了一大半。而且，根据之前定义的 **数据挖掘成功标准** ，您还确定在建模阶段构建的模型从技术上说是正确而且有效的。

但是，在继续之前，您应该使用在项目开始时设立的 **业务成功标准** 评估您的努力结果。这是确保您的组织可以利用您所获得的结果的关键环节。数据挖掘可以生成两种类型的结果:

- 在 CRISP-DM 的前期阶段中选定的最终 **模型** 。
- 从模型本身以及数据挖掘过程中得出的任意结论或推论。这些均称为 **发现的问题** 。

评估结果

在这个阶段，您将对项目结果是否达到业务成功标准的评估进行规范。此步骤要求对声明的业务目标有清晰地了解，因此请确保在项目评估时包含关键决策制定者。

任务列表

首先，您需要将您对数据挖掘结果是否达到业务成功标准的评估记录在案。在报告中考虑以下问题:

- 您的结果是否明确声明并且采用可以轻松展示的格式？
- 是否存在应该突出强调的特别故事或独特的发现问题？
- 您是否能够按照模型和发现的问题对于业务目标的适用顺序对他们进行排序？
- 总的来说，这些结果能在多大程度上满足您组织的业务目标？
- 您的结果还引发了哪些其他问题？您将如何使用商业术语表述这些问题？

评估完结果后，编辑汇总一个已批准的模型列表以包含在最终报告中。此列表应该包含同时满足您组织的数据挖掘目标以及业务目标的那些模型。

电子商务零售业示例 - 评估结果

使用 CRISP-DM 的 Web 挖掘方案

此电子商务零售商首次体验数据挖掘的总体结果从业务角度来说很容易说明白: 研究表明什么有希望成为更好的产品推荐并且改善了站点的设计。站点设计基于客户的浏览顺序进行了改善，它展现了客户希望但却需要多个步骤才能达到的站点功能。因为决策规则可能很复杂，所以很难证明产品推荐更佳。为了生成最终报告，分析师将尽力从规则集中找到一些可以更容易解释的一般趋势。

模型评级。因为有多个初始模型看上去具有商业意义，所以将他们在组内基于统计标准和多样性进行排秩，这很容易解释。因此，该模型针对不同情况给出不同的推荐。

新问题。通过研究发现的最重要的问题是，这个电子商务零售商如何才能找到有关他或她的客户的更多信息？客户数据库中的信息在形成推荐聚类时起着很重要的作用。尽管在对那些信息缺失的客户进行推荐时可以使用一些特殊的规则，但从本质上来看，这些推荐与那些可以对注册客户进行的推荐相比，更具普遍性。

审核过程

有效的方法通常包含用于反映刚完成的过程的成功之处和不足之处的时间。数据挖掘也不例外。部分 CRISP-DM 会吸取您的经验，因此未来的数据挖掘项目将会变得更为有效。

任务列表

首先，您应该汇总每个阶段的活动和决定，包括数据准备步骤、模型构建等。然后对于每个阶段，考虑以下问题并提出改进建议:

- 这一阶段是否对最终结果的值有所贡献？
- 有没有方法可以简化或改善这一特定阶段或操作？
- 这一阶段的失败之处和失误分别是什么？下一次应该如何避免这些问题？
- 是否存在死端，例如某些已验证无效的特殊模型？是否有办法预测此类死端，以便可以更有效地开展工作？
- 这一阶段是否存在任何令人惊喜的结果（不论好或坏）？事后看来，是否有明显的办法可以预测此类事件的发生？
- 是否有其他的备选决定或策略可以在某个给定的阶段使用？请在将来的数据挖掘项目中注意此类替代选项。

电子商务零售业示例 - 审核报告

使用 CRISP-DM 的 Web 挖掘方案

作为审核最初的数据挖掘项目过程的结果，该电子商务零售商对于流程中各步骤之间的相关性表示极为赞赏。该电子商务零售商开始时并不愿意“回溯”CRISP-DM 过程，但现在发现过程的循环可以增强其功效。过程审核还让该电子商务零售商了解了:

- 当 CRISP-DM 过程中的另一个阶段发生了不正常的情况时，返回到探索过程始终是合理的。
- 数据准备，特别是 Web 日志的准备，需要耐心，因为这往往需要花费很长的时间。
- 将注意力集中在现有业务问题上很重要，这是因为一旦作好分析数据的准备，非常容易不考虑全局就开始构建模型。
- 一旦建模阶段结束，对于决定如何实施结果和确定还需要进行哪些深入研究来说业务理解甚至起着更为重要的作用。

确定后续步骤

至今为止，您已经生成了结果并且评估了您的数据挖掘经历，此时应该考虑，**接下来该做些什么？**这一阶段将帮助您根据数据挖掘业务目标回答上述问题。实际上，您此时具有两个选择:

- **继续进入部署阶段。**下一个阶段将帮助您将模型结果并入您的业务过程中，从而生成最终的报告。即使您的数据挖掘工作没有成功，您也应该使用 CRISP-DM 的部署阶段来创建最终报告以便将其发送给项目负责人。
- **返回到前面的步骤改进或替换您的模型。**如果您发现您的结果几乎可以算是（但并非）最佳结果，可以考虑另一轮建模。您可以将在此阶段中了解到的信息用于改进模型并生成更好的结果。

此时，您的决定关系到建模结果的准确度和相关性。如果结果实现了您的数据挖掘目标和业务目标，则您已经可以进入部署阶段。不论您作出什么决定，都请确保记录下完整的评估过程。

电子商务零售业示例 - 后续步骤

使用 CRISP-DM 的 Web 挖掘方案

此电子商务零售商对于项目结果的准确度和相关性相当自信，因此要继续进入部署阶段。

同时，项目团队也作好了返回前期步骤并增加一些模型以包含预测技术的准备。此时，他们正在等待递交最终报告以获得决策者的批准。

第 7 章 部署

部署概述

部署就是使用您的新的深入见解在组织内部进行改善的过程。这可以表示正式的集成，例如实施一个用于生成随后要读入数据仓库中的流失评分的 IBM SPSS Modeler 模型。此外，部署还意味着您可以使用从数据挖掘中获得的深入见解改善您的组织。例如，也许您会发现您数据中的报警模式指明年龄超过 30 岁的客户行为会发生改变。这些结果可能不会正式集成到您的信息系统中，但它们无疑对于计划和制定营销决策非常有用。

通常，CRISP-DM 的部署阶段包含两种类型的活动:

- 计划和监视结果的部署
- 完成包尾的任务，例如生成最终报告和执行项目审核

取决于您组织的要求，您可能需要完成上述步骤之一或全都完成。

制定部署计划

尽管您可能非常希望共享您数据挖掘工作的成果，但是请花时间计划顺利而全面的结果部署。

任务列表

- 第一个步骤是汇总您的结果，包含模型和发现的问题。这个步骤可以帮助您确定哪些模型可以集成到您的数据库系统中，哪些发现的问题应该向您的同事展示。
- 对于每种可部署的模型，创建一个分步骤执行的计划以便部署和集成到您的系统中。请注意任何技术细节问题，例如模型输出的数据库要求。例如，也许您的系统要求建模输出使用制表符分隔格式部署。
- 对于包含的每个发现问题，创建将此信息传递给策略制定者的计划。
- 对于两种值得说明的结果类型是否有备择部署计划？
- 考虑如何监控部署。例如，如何更新使用 IBM SPSS Modeler Solution Publisher 部署的模型？当模型不再适用时您将作何决定？
- 找出所有部署问题并准备应急预案。例如，决策制定者可能需要更多有关建模结果的信息，而且可能会要求您提供进一步的技术细节信息。

电子商务零售业示例 - 部署计划

使用 CRISP-DM 的 Web 挖掘方案

要成功部署电子商务零售商的数据挖掘结果要求将正确的信息传递给正确的人。

决策者。需要告知决策者要对站点进行哪些推荐和建议的改进，并简单地向他们解释这些改进能起到什么作用。假设他们接受研究的结果，则需要通知将要实施这些更改的人员。

Web 开发人员。维护 Web 站点的人必须将新的推荐和站点内容组织合并。通知他们因为将来的研究 *可能* 会发生的更改，以便他们现在就开始开展基础工作。让团队作好准备以便应对基于实时序列分析的动态站点构建，这可能会对后面的工作很有帮助。

数据库专家。应该保证通知维护客户、采购信息和产品数据库的人员数据库中的信息是如何使用的，以及在未来的项目中哪些属性可能会添加到数据库中。

总的来说，项目团队需要与这些团队全都分别保持联系，以便调整结果的部署方案并为将来的项目制定计划。

计划监视和维护

在一个成熟的建模结果部署和集成中，可能需要同时开展数据挖掘工作。例如，如果部署了一个模型用于预测电子购物篮的采购顺序，很有可能需要对此模型进行定期评估以确保其有效性，并且不断地对其进行改进。与之相似，对于为提高高价值客户的客户保持度而部署的模型而言，一旦达到特定的客户保持度级别后，将有可能需要对此模型进行调整。可能随即会对此模型进行修改，然后重新使用它来保留处于较低级别但仍处于价值金字塔盈利级别的客户。

任务列表

记录下列问题并确保将它们包含在最终报告中。

- 对于每个模型或发现的问题，需要跟踪哪些因素或影响（例如市场价值或季节性差异）。
- 如何度量和监视各个模型的有效性和准确度？
- 您将如何确定某个模型是否已经“过期”？给出有关数据中的准确度阈值或预期更改的具体信息
- 当模型过期时将会发生什么情况？您是否只需使用较新的数据重建模型，或者只是进行微调即可？或者所作的更改是否覆盖面很广，以至于需要一个新的数据挖掘项目？
- 一旦某个模型过期后，是否能对相似的业务问题使用该模型？在此情况下，好的文档对于评估每个数据挖掘项目的业务目的变得很重要。

电子商务零售业示例 - 监视和维护

使用 CRISP-DM 的 Web 挖掘方案

要监控的直接任务是确定新站点的组织和改进建议是否真的有效。也就是说，用户是否能够通过更多路径直接进入他们所要查找的页面？推荐项目的交叉销售量是否有所增加？经过几周的监视，此电子商务零售商将能够确定此次研究成功与否。

可以自动处理的是包含新的注册用户。当客户在站点中注册时，将在他们的信息中应用当前规则集，从而确定应该向他们提出什么建议。

决定何时更新确定建议的规则集是一件较为棘手的任务。因为创建聚类需要手动输入有关给定聚类解决方案准确性的内容，所以更新规则集并非自动执行的过程。

由于未来的项目会生成更多复杂的模型，对于监视的需求和数量都将毫无疑问地大大增加。当可能时，大量的监视工作应该使用可供审核的定期计划报告自动执行。此外，创建可提供动态预测的模型可能会是公司期望努力的方向。这要求团队的技巧比第一个数据挖掘项目时更熟练。

生成最终报告

编写最终报告不仅可以将早期文档中的零碎信息联系起来，而且还可以用于传达您的结果。尽管这看起来很简单，但重要的是将您的结果演示给各种与结果的相关人员。这可以包括将负责实施结果建模的技术管理员以及将基于您的结果制定决策的营销负责人和管理负责人。

任务列表

首先，考虑什么人观看您的报告。他们是技术开发人员还是只关注市场的管理人员？如果这些观众的需求完全不同，您可能需要为每一类观众创建单独的报告。不论哪种情况，您的报告都应包含下面这些点中的大多数:

- 原始业务问题的全面说明

- 用于执行数据挖掘的流程
- 项目的成本
- 与原始项目计划之间的偏差的记录
- 数据挖掘结果的汇总信息，包含模型和发现的问题
- 所提议的部署计划概况
- 数据挖掘工作的进一步建议，包括在数据探索和建模阶段发现的令人感兴趣的线索

准备最终演示

除了项目报告之外，您可能还需要向一组负责人或相关部门演示发现的项目问题。如果遇到这种情况，您可以使用与报告中的信息大致相同的信息进行演示，但需要从更宽的角度讲解。IBM SPSS Modeler 的图标和图形可以轻松导出供此类演示使用。

电子商务零售业示例 - 最终报告

使用 CRISP-DM 的 Web 挖掘方案

与原始项目计划之间的巨大差异也是进行进一步数据挖掘工作的令人感兴趣的线索。原始计划需要找出让客户每次访问站点时花更多的时间并且查看更多的网页的方法。

结果显示，拥有一个快乐的客户并不仅仅是使他们保持在线状态这么简单。每个会话持续时间的频率分布（对会话是否源自采购表现出不一致），显示大多数源自采购的会话的会话次数介于两种非采购会话的会话次数之间。

既然已经知道了这一情况，那么问题便在于找出那些在站点中花费很长时间但却不购买商品的客户只是在浏览还是只是找不到他们想要的商品。下一个步骤便是找出如何提供他们想要的商品以便鼓励购买。

执行最终项目审核

这是 CRISP-DM 方法的最后一个步骤，使用它，您可以明确说明您的最终印象以及对数据挖掘过程中吸取的教训进行比较。

任务列表

您应该对那些数据挖掘过程所涉及的重要人员进行简短的采访。要在这些采访中考虑的问题包括以下内容:

- 您对此项目的总体印象如何？
- 您在此过程中学到了有关一般数据挖掘和可用数据的哪些知识。
- 哪些部分的项目进展顺利？哪些部分是困难所在？有没有信息可以帮助减轻混淆的情况？

在部署数据挖掘结果之后，您也可以采访那些受到结果影响的人，例如客户或商业合作伙伴。您此时的目标应该是确定项目是否值得实施以及是否提供了创建此项项目时设定的收益。

这些采访的结果可以与您自己对项目的印象汇总在一起并记录在最终报告中，此报告应该将重点放在从挖掘您的数据存储的经验中吸取的教训。

电子商务零售业示例 - 最终审核

使用 CRISP-DM 的 Web 挖掘方案

项目成员采访。电子商务零售商发现，那些从始至终与研究关系最紧密的项目成员是对结果最热心并且盼望开展未来项目的人。数据库团队看上去持谨慎的乐观态度；尽管他们赞赏研究是有用的，他们仍指出这对数据库资源增加了负担。在研究的过程中一直都有一个顾问参与，但要进一步扩展项目范围，另一个专门维护数据库的员工将是必不可少的。

客户采访。至今为止，客户的反馈大多数都是正面的。一个之前没有经过深思熟虑的问题是站点设计的更改对现有客户的影响。经过这些年，注册客户对站点的组织方式已经产生了某种特定的预期。注册用户的反馈与来自非注册客户的反馈相比，表示认可的要少一些，而且还有一些用户表示非常不喜欢这样的更改。此电子商务零售商需要清楚知道这些问题，并且需要仔细考虑这样的更改是否能够在冒险失去现有客户的同时带来足够的新客户。

声明

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务，则由用户自行负责。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并未授予用户使用这些专利的任何许可。您可以用书面方式将许可查询寄往:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

本条款不适用英国或任何这样的条款与当地法律不一致的国家或地区: International Business Machines Corporation“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本信息中可能包含技术方面不够准确的地方或印刷错误。此处的信息将定期更改；这些更改将编入本资料的新版本中。IBM 可以随时对本出版物中描述的产品进行改进和/或更改，而不另行通知。

本信息中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的许可证持有者如果要了解有关程序的信息以达到如下目的: (i) 允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及 (ii) 允许对已经交换的信息进行相互使用，请与下列地址联系:

IBM Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

此处包含的任何性能数据都是在受控环境中测得的。因此，在其他操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。此外，有些测量是通过推算而估计的，实际结果可能会有差异。本文档的用户应当验证其特定环境的适用数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

如果您正在查看本信息的软拷贝，图片和彩色图例可能无法显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp. 在全球许多行政管辖地区的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。Web 页面“Copyright and trademark information” (www.ibm.com/legal/copytrade.shtml) 提供了 IBM 商标的最新列表。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和/或其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 和/或其子公司的商标或注册商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

索引

Printed in China