第 1 章 成功之路

“如果你不知道要去哪里，就只能随波逐流，无所适从。”

——刘易斯·卡罗尔

“如果你不能将要做的事情描述成一个流程，那么你就不知道自己在做什么。”

——爱德华兹·戴明

乍一看，这一章跟机器学习没有什么关系，但实际上本章内容对于机器学习非常重要（特别是对于机器学习的实施以及由此造成的改变）。不管我们如何定义成功，最聪明的人、最好的软件和最好的算法都不能确保其实现。

在大多数（即便不是全部）项目中，成功解决问题或改进决策的关键因素不是算法，而是沟通能力和影响力之类的非定量的软技能。很多人认为其中的问题在于，我们很难量化这些软技能的效果。一般来说，人们遇到不想做的事情都会止步不前。别忘了，爆红的电视喜剧《生活大爆炸》就是这么拍的。所以，本章目的是使你走向成功，意在提供一个流程，至少是一个灵活的流程，使你成为一位**变革推动者**：一个不靠位高权重以势压人，而是具有真知灼见并能付诸实施的人。我们将集中讨论**跨行业数据挖掘标准流程（Cross-Industry Standard Process for Data Mining，CRISP-DM）**，这可能是最著名也是最受重视的项目分析方法。即使你使用的是其他成熟方法或专有技术，也可以在本章有所收获。

我可以毫不犹豫地说，事情说起来容易，做起来难。对于本章内容中的错误和遗漏，我感觉非常内疚和遗憾。希望你能凭能力和运气避免我在过去12年中受到的各种身心上的伤害。

最后，我会介绍一个流程图（快速指南），你可以使用它判断应当使用何种方法解决手头的问题。

1.1 流程

CRISP-DM流程本来是专门为数据挖掘设计的，但它非常灵活和全面，完全可以应用于任何项目分析，无论是预测性分析、数据科学还是机器学习项目。不要被长长的任务列表吓倒，因为你可以在流程实施过程中根据自己的判断对流程进行调整以适应实际情况。图1-1给出了这个流

程的可视化表示，以及可以使流程非常灵活的反馈回路。

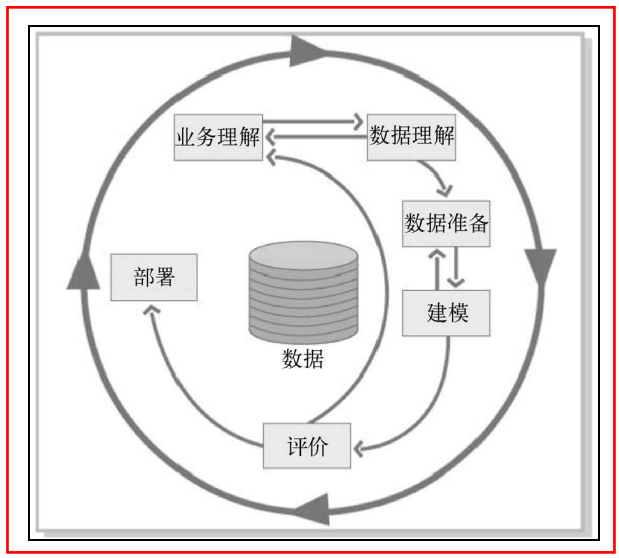

图1-1　CRISP-DM 1.0，循序渐进地进行数据挖掘

流程分以下6个阶段：

- 业务理解
- 数据理解
- 数据准备
- 建模
- 评价
- 部署

如果想查看包括所有任务和子任务的完整流程说明，请参考SPSS的文章“CRISP-DM 1.0, step-by-step data mining guide”（https://the-modeling-agency.com/crisp-dm.pdf）。

我会对流程中的每个步骤（包括其中的重要任务）进行说明，但这不是包含详细细节的说明书，而是更高层次的介绍。我不会跳过任何关键细节，但会重点介绍可以应用在任务上的技术。请记住，在后面的章节中我们将使用这些流程步骤，作为机器学习方法实际应用中的一个通用框架，特别是R语言实现。

1.2　业务理解

我们绝对不能低估流程的第一阶段对于最后成功的重要性，这是最基础的一个阶段，它的成败在很大程度上将决定项目其余部分的成败。本阶段的目的是确定业务需求，进而转化为分析目

标。这一阶段有以下4个任务：

(1) 确定业务目标
(2) 现状评估
(3) 确定分析目标
(4) 建立项目计划

1.2.1 确定业务目标

这一任务的关键是确定组织的目标并且限定问题的范围。一个好问题是：我们要做什么改变？这看上去是一个老掉牙的问题，但它确实可以使人们从分析的视角去思索到底想要什么，并且认识到需要做出的决策的根本目的。这个问题还可以防止你在调研中走得太远，或做一些不必要的工作。所以，最关键的一点就是确定**决策**。对于管理团队来说，决策的一个范例就是，对是否进行资源投入做出明确的选择。当然，选择不做任何改变也是一种决策。

这并不意味着如果选择没有彻底明确，项目就不能开始。有些时候，问题不存在或不能清楚定义，正如美国前国防部长唐纳德·拉姆斯菲尔德所说，这是“已知的未知”。实际上，问题经常没有得到清楚定义，项目的主要目标就是更加深入地理解问题并提出假设。又如拉姆斯菲尔德部长所说，还有“未知的未知”，也就是说你根本不知道自己不知道什么。但是，在没有定义清楚的问题中，我们可以根据基于各种假设所产出结果的资源投入，来理解接下来将发生什么。

在这一任务中，还需要重视对期望的管理。完美的数据是不存在的——不管其深度与广度如何。现在要基于你的专业知识说明什么是可行的，而不是做出保证。

我建议完成这一任务之后，要得到两个成果。第一个就是任务说明。这可不是单位中的琐碎冗长的任务说明，而是你自己的，更进一步说，是经过项目负责人确认过的任务说明。我是从多年的军旅生涯中悟到这一点的，我可以长篇大论地说明为什么它非常有效，但那是以后的事。我们可以认为，当没有明确的方向和指导时，任务说明（或者随便你想叫它什么）是所有利益相关者的统一声明，可以防止需求范围蔓延。它包括以下几部分。

- **谁**：你、你的团队或者项目名称，所有人都喜欢一个酷酷的项目名，例如“蝰蛇行动”“融合”，等等。
- **什么**：你要进行的任务，例如，实施机器学习。
- **何时**：最后期限。
- **何地**：可以是地理意义上的，也可以是功能、部门、自发项目等。
- **为什么**：项目的目的，也就是业务目标。

第二个成果就是要尽可能明确成功的定义。达到哪些条件才算是成功？要帮助团队或负责人清晰描述出你所理解的成功，然后你的工作就是将其转化为建模需求。

1.2.2　现状评估

这项任务用于在项目计划中收集信息，包括可用资源、限制条件和假设；还要识别风险，并做出应急方案。多说一句，在这个阶段，还要确定关键的利益相关者，也就是要受到未来决策影响的人。

这里有几点需要注意。当检查可用资源时，不要忘了将过去和现在的项目记录都看一遍。要特别关注组织中曾经或正在处理相同问题的人，要将你的工作和他们的工作结合起来。不要忘了列举风险，时间、人员和费用都会产生风险。尽你的最大努力建立一个利益相关者列表，包括那些影响项目的人和被项目影响的人，确定这些人是谁，如何影响决策或被决策影响。这项工作完成之后，就要和项目负责人一起建立一个与这些利益相关者的沟通计划。

1.2.3　确定分析目标

在这一步，你需要将业务目标转化为技术需求。这需要将“确定业务目标”任务中的成功标准转化为技术上的成功标准，此时可能要引入均方根误差或预测准确度水平等标准。

1.2.4　建立项目计划

这一步的任务是基于目前收集到的所有信息建立一个有效的项目计划。不管你使用什么技术，甘特图或其他图表都可以，一定要使其成为沟通计划的一部分。要使大多数利益相关者了解这个计划，并根据实际情况定期更新。

1.3　数据理解

在经过了虽然痛苦但却至关重要的第一阶段之后，你可以着手于数据工作了。这个阶段的任务如下：

(1) 数据收集
(2) 数据描述
(3) 数据探索
(4) 数据质量校验

这一步是**ETL**（**数据抽取、转化和加载，Extrack, Transform, Load**）的典型案例，也有几个需要注意的地方。你需要做初步判定，确认目前可用的数据能够满足你的分析要求。当你使用可视化或其他方法进行数据探索时，需要确定变量是否稀疏以及数据缺失程度，从而确定要使用的机器学习方法，并确认对缺失数据的填补是否必要和可行。

数据质量的校验是非常重要的。需要花费一些时间来搞清楚数据是由谁收集的，如何收集的，

甚至还要搞清为什么要收集数据。你经常会遇到数据收集不完整的情况，意外的IT问题会导致数据错误，业务规则也会按计划进行调整。这在时间序列分析中非常重要，决定数据分类的业务规则会经常随时间变化。最后，从这个阶段开始对程序代码进行归档是个好主意。作为归档过程的一部分，如果数据字典不可用，那么为了防止潜在的严重问题，请一定要做一个数据字典。

1.4 数据准备

差不多了！这个阶段包括下面5个任务：

(1) 数据选择
(2) 数据清洗
(3) 数据构建
(4) 数据整合
(5) 数据格式化

这些任务不需太多解释，其目的就是准备好数据以输入算法，包括数据合并、特征工程、数据转换等。如果需要填补缺失数据，也要在这个阶段完成。尤其是使用R，则要注意输出结果需要如何标记。如果输出变量（响应变量）是Yes/No的形式，那么在有些程序包中是不被支持的，需要进行数据转换，否则就没有1/0变量。在这个阶段中，如果需要，还要将数据分成不同的集合：训练集、测试集或验证集。这个阶段的工作极其繁重，但是很多过来人会告诉你，这就是你脱颖而出的机会。做好这些工作，我们就可以得到丰厚的回报。

1.5 建模

在这个阶段，你之前所做的一切工作都该有个结果了——或者挥拳庆祝，或者抱头痛哭。别在意，如果这个工作那么简单，那不是谁都能做了吗？本阶段的任务包括：

(1) 选择建模技术
(2) 设计检验方法
(3) 建立模型
(4) 评估模型

奇怪的是，这个流程阶段中需要注意的事情都是你已经考虑过的和准备好的。在第一阶段，你对如何进行建模总会有一点概念。请记住，这是一个灵活的、可迭代的流程，而不是像机组人员备忘录那样严格的线性流图。

本章后面的快速指南可以帮助你正确选择建模技术。检验设计指的是如何建立你的测试数据集和训练数据集，以及如何使用交叉验证。在数据准备阶段就需要考虑这些事情了。

模型评估需要将模型与在业务理解阶段建立的成功标准进行对比，比如均方根误差、提升度、ROC曲线等。

1.6　评价

在评价阶段，主要目的是确认已经完成的工作和选择的模型是否符合业务目标。问一下自己和他人，我们达到项目成功的要求了吗？看看Netflix这一反面教材。我相信你一定知道，Netflix悬赏100万美元寻找团队，开发最好的、具有最低均方根误差的推荐算法。但是，Netflix并没有实施这个算法，因为算法获得的精确度提升根本配不上实施成本。请永远记住奥卡姆剃刀原理。评价阶段至少包括以下任务：

(1) 评价结果
(2) 回顾过程
(3) 确定下一步

回顾过程时，非常有必要使工作结果获得管理层的认可，并与其他利益相关者进行沟通，以取得他们的支持。当然，在流程的前几个阶段也是如此。如果你想成为变革推动者，下一个步骤就是确保你已经回答利益相关者头脑中的这几个问题：**是什么**？**要什么**？**现在要做什么**？如果你能用前面阶段中产生的决策来说服他们现在要做什么，那你就成功了。

1.7　部署

如果直到现在所有事情都按照计划进行，那么只需按动一下开关，你的模型就会运转。假设情况并非如此简单，那么这个阶段要进行以下任务：

(1) 按计划部署
(2) 监测与维护
(3) 完成总结报告
(4) 项目回顾

在部署、监测和维护工作展开之后，对于你和你的团队来说，最重要的事情就是完成一份完整详尽的总结报告。报告应该包括一份白皮书和一份幻灯片简介。我要解释一下，我没有按照自己的意愿将工作成果放在白皮书中，因为我和军方签有合同，军方强烈要求使用PowerPoint幻灯片。但是，幻灯片很可能是某些人出于自己的目的来应付你的，华而不实，空话连篇。相信我，白皮书不会有这个问题，因为它是你的工作成果和信念的延续。如果你所在的组织坚持使用PowerPoint，那么你可以使用它向利益相关者做简单的介绍，但要使用白皮书作为文档记录和预读资料。在R中，生成白皮书的标准过程是使用knitr和LaTex包。

现在，我们已经讨论了所有重要阶段，你可以大展身手了。但是不管你以正式方式还是非正式方式推进流程，其中总要包括以下几点：

- 计划做什么？
- 实际做了什么？
- 为什么做这些工作，或为什么不做？
- 以后的项目中还需要做什么支持工作？
- 以后的项目中要进行哪些改进？
- 确定行动计划，保证支持和改进工作顺利进行。

综上所述，我们完成了对CRISP-DM流程的介绍。这个流程提供了一个综合而又灵活的框架，以保证项目成功实施，并使你成为那个推进变革的人。

1.8 算法流程图

本节的目的是建立一个工具，它不但可以帮助你选择合适的建模技术，而且可以帮助你更加深入地思考问题，你也可以用它和项目团队或项目负责人一起做出问题的框架。这种使用流程图的技术并不复杂，足以让你开始工作了。它还包括了一些本书没有涉及的技术。

图1-2启动了一个流程，通过这个流程你可以选择可能使用的建模技术。你只需回答问题，它会指引你进入图1-3~图1-6。

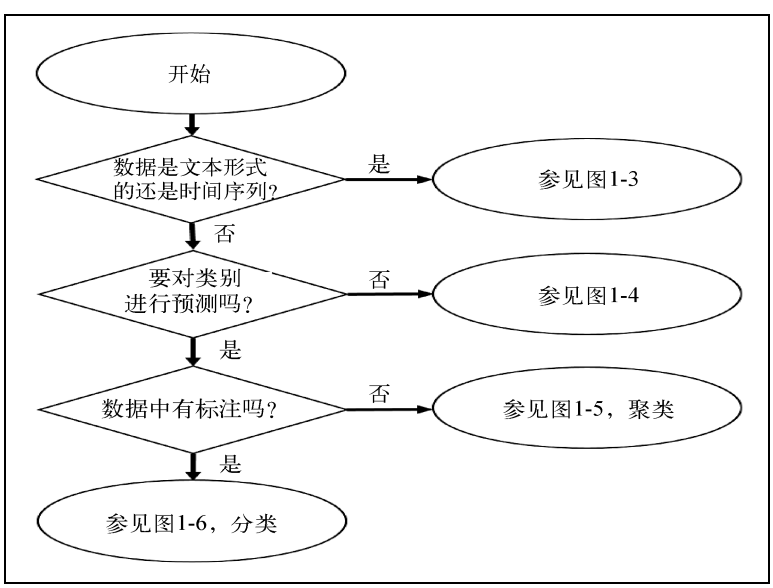

图1-2

如果数据是文本形式的或时间序列形式的，那么你应该采用图1-3中的流程。

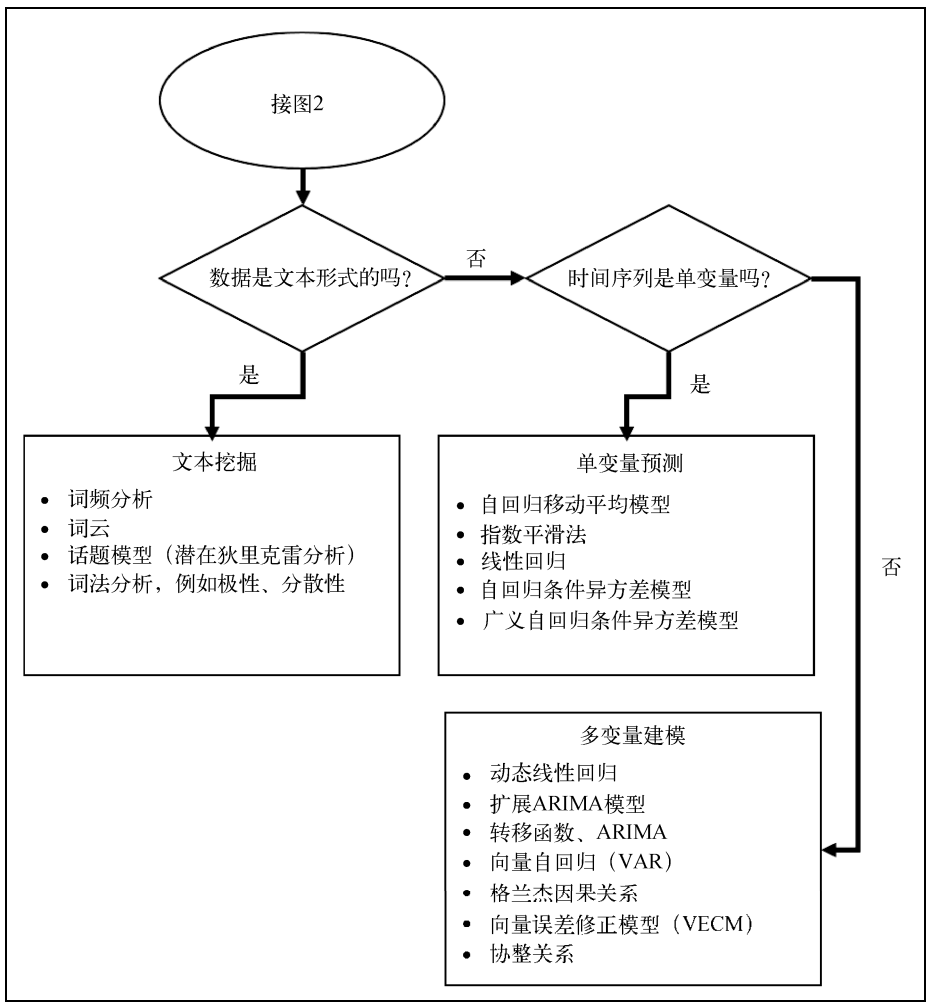

图1-3

图1-4所示的这个算法分支中，你不需要文本数据和时间序列数据，也不需要预测分类。所以，你的目的应该是做出推荐、理解关联规则或预测出一个数值量。

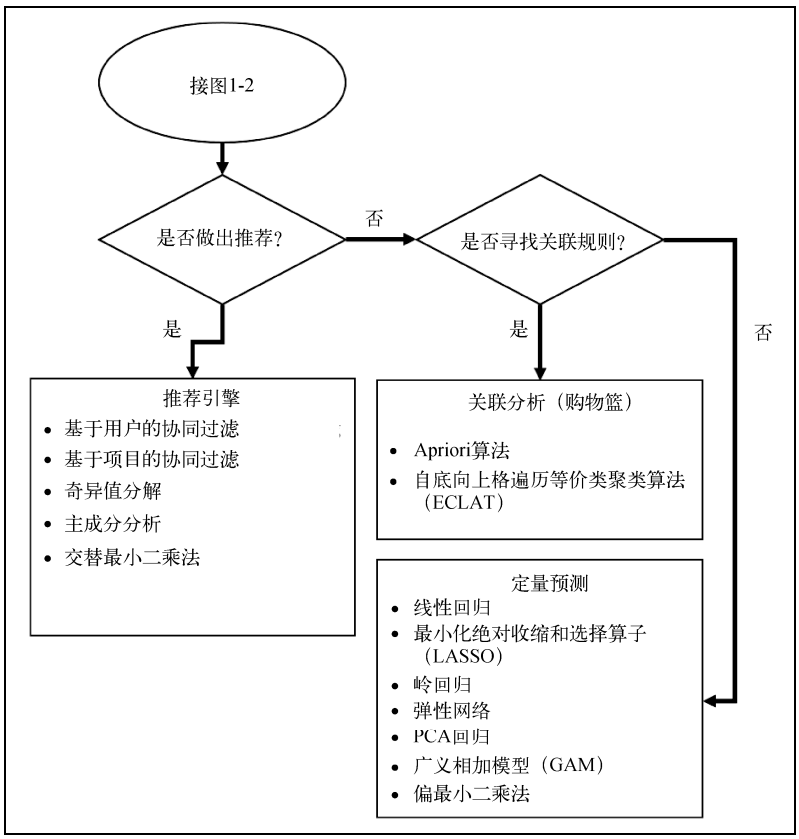

图1-4

要进入图1-5，你的数据不能是文本形式的或时间序列。你的目的是对数据进行分类，但分类结果不用标记，这就要使用聚类方法。

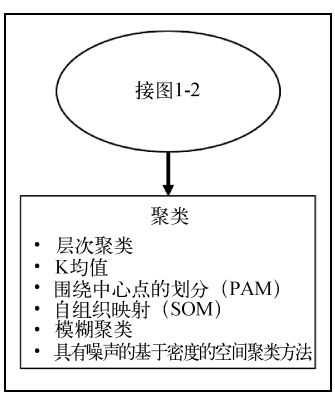

图1-5

如果想对数据分类并进行标记，则使用分类技术，如图1-6所示。

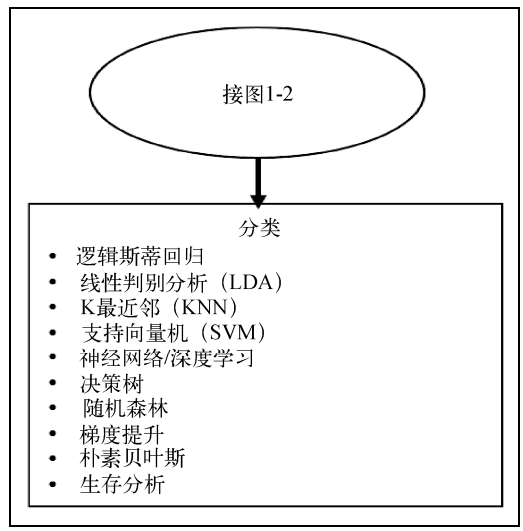

图1-6

1.9　小结

本章介绍如何使你自己和你的团队在承担的项目中取得成功。我们介绍了CRISP-DM流程，这是一种灵活而又全面的框架，其目的是提高沟通能力和影响力等软技能。我们列举了流程的每个阶段及其任务，还详细介绍了一些技术和注意事项，以保证流程顺利执行。如果你能认真遵循流程，那么在任何组织中，你都可以成为积极的变革推动者。

本章的另一部分内容是关于算法流程图的，这是一份快速指南，可以帮助你确定合适的技术，以解决业务问题。基础已经打好了，我们的下一步工作就是使用机器学习技术解决实际问题。