前言

要先提醒的是，學習和實踐資料科學非常困難。在希望成為優秀的程式設計師之前，不只要瞭解資料結構及其計算複雜性的細節，還要熟悉 Python 和 SQL。對您來說，統計學和最新的機器學習預測技術應該如同熟悉的第二種語言；想當然耳，您也要能夠應用這一切知識，來解決所有可能出現的實際業務問題。此外，這份工作的困難點也在於您必須是一位出色的溝通者，能夠向非技術人員講述引人入勝的故事，而這些人可能不習慣以資料為基礎做出決策。

所以，大方承認吧：資料科學的理論和實務幾乎是說不出口的困難。任何旨在涵蓋資料科學困難點（*hard part*）的書籍，要不像百科全書般詳盡，要不然就是會優先淘汰掉一些主題。

我也會在一開始就承認，這裡所選擇的主題，都是我認為資料科學的學習困難點，而這個標籤當然很主觀。如果想要讓它不那麼主觀，我會提出這樣的觀點：它們之所以難以學習，不是因為複雜，而是因為此時此刻，這個行業對學習這些主題的重視程度不高，以至於難以在資料科學職業生涯的入門階段，找到相對應的學習資源。因此在實務上，只是因為難以找到相關素材，才讓它們這麼不好學。

資料科學課程通常會強調的，是我稱之為大主題的程式設計和機器學習。但是，其他幾乎所有這份工作需要學到之事，只能祈禱您運氣好，在第一份或第二份工作時就遇到一位良師教導，這實屬大不幸。所以大型科技公司才會這麼美好，因為就算是有點見不得光，對許多從業者來說無法接觸而只能成為公司次文化的主題，它們也擁有同樣龐大的人才。

這本書探討的是能夠幫助您成為更高生產力的資料科學家技巧，分為兩部分：第一部分涵蓋資料分析和資料科學較為軟性的主題，第二部分則全與機器學習（ML）相關。

雖然這本書的各章節可以跳著閱讀，而不會影響學習，但某些章節仍會提到先前內容；大多數情況下，您可以跳過參考資料，內容仍然清楚且不需額外解釋；參考資料主要是在那些看似獨立的主題之間，提供統一說法。

第一部分涵蓋以下主題：

第 *1* 章：〈摘要重點！用資料科學創造價值〉

資料科學為組織創造價值的角色為何，又要如何衡量？

第 *2* 章：〈度量設計〉

我認為，資料科學家最適合改進可行動度量的設計，這章將說明這點。

第 *3* 章：〈增長分解：理解順風和逆風〉

理解業務狀況並提出引人入勝的敘事，是資料科學家的常見需求。本章介紹一些可用來自動化部分工作流程的增長分解（*growth decomposition*）方法。

第 *4* 章：〈*2*×*2* 設計〉

學習簡單的方法，才可以做更多複雜的事，而 *2*×*2* 設計（*2*×*2 design*）就可以幫助您達成這項目標，同時改善與利益相關者之間的溝通。

第 *5* 章：〈建構業務案例〉

在開始專案之前，您應該有一個業務案例（*business case*），此即本章內容。

第 *6* 章：〈提升的奧祕〉

提升（*lift*）很簡單，而且可以加速您也許想以機器學習進行的分析，本章將解釋提升的概念。

第 *7* 章：〈敘事〉

資料科學家需要提升自己講故事（storytelling），和構建引人入勝的敘事（*narrative*）能力，這正是這章的內容。

第 *8* 章：〈資料視覺化：選擇合適的圖表來傳達訊息〉

投入足夠的時間讓資料視覺化（*data visualization*），應該也會有助於構建敘述。本章會討論一些最佳實務。

第二部分與機器學習有關，涵蓋以下主題：

第 *9* 章：〈模擬和自助法〉

模擬（*simulation*）技術可以幫助您加強對不同預測演算法的理解，相關內容可見這一章，還會討論使用迴歸和分類技術時的一些注意事項。本章還會討論用於找到一些難以計算的估算值（estimand）信賴區間自助法（*bootstrapping*）。

第 *10* 章：〈線性迴歸：回歸到基礎〉

深入瞭解線性迴歸（*linear regression*），能夠理解更進階的相關主題，因此非常重要。這一章會回歸基礎知識，希望提供對於機器學習演算法更紮實的直觀基礎。

第 *11* 章：〈資料洩漏〉

什麼是資料洩漏（*data leakage*）？要如何識別並預防發生？可見本章的相關介紹。

第 *12* 章：〈將模型投入生產〉

只有達到生產階段（*production stage*）的模型，才是好模型。幸運的是，這已經是廣為人所知並結構化的問題，這裡會介紹其中最關鍵的步驟。

第 *13* 章：〈機器學習講的故事〉

有一些很棒的技術可以幫助您打開黑箱，並在機器學習的故事講述方面表現出色。

第 *14* 章：〈從預測到決策〉

透過資料和機器學習驅動的過程，而提高決策能力以創造價值，本章會展示從預測轉向決策（*prediction to decision*）的方法。

第 *15* 章：〈增量性：資料科學的聖盃？〉

因果性（*causality*）在資料科學中已經取得了一些動力，但仍認定為相對專業的領域。這一章介紹基礎知識，並提供一些可以直接應用在貴組織的範例和程式碼。

第 *16* 章：〈*A/B* 測試〉

A/B 測試（*A/B test*）是估算替代行動方案增量性的典型範例，這個實驗需要一些統計（和業務）的強大背景知識。

最後一章（第 17 章）非常特別，因為我完全沒有介紹任何技術，而是推測生成式人工智慧（AI）出現後的資料科學未來。先說結論，我預期接下來的幾年內，這個職業的工作內容將會發生根本性的變化，而資料科學家應該為這種變化（革命）做好準備。

這本書適用於各種等級和資歷的資料科學家。而為了完整發揮這本書的價值，您最好具備一些中高階機器學習演算法的知識，因為我不會花時間介紹線性迴歸、分類和迴歸樹，或者像是隨機森林或梯度提升機這類的集成學習器。

本書編排慣例

本書使用以下排版慣例：

斜體字（*Italic*）

表示新的術語、URL、電子郵件地址、檔名和延伸檔名。

定寬字（`Constant width`）

用於程式列表，以及在段落中參照的程式元素，例如變數或函數名稱、資料庫、資料型別、環境變數、敘事和關鍵字。

此圖表示提示或建議。

此圖表示一般性注意事項。

此圖表示警告或警示事項。

摘要重點！用資料科學創造價值

資料科學（data science, DS）在過去的二十年中取得了令人矚目的成長，從最初只有矽谷頂尖科技公司才負擔得起的相對小眾領域，發展成為現今許多跨足各個行業和國家的組織都在參與的領域。然而，許多團隊仍然不知如何計量它所能對公司產生的價值。

因此，DS 對組織的價值為何？我發現各層級的資料科學家都在苦思這個問題，無怪乎組織本身也答不出來。我在第 1 章的目標，就是勾勒出一些創造 DS 價值的基本原則，相信理解和內化這些原則，可以幫助您成為更優秀的資料科學家。

價值何在？

公司的存在，是為了為股東、顧客和員工，希望也包括整個社會創造價值，自然而然，相對於其他替代方案，股東更期望他們的投資能獲得回報，顧客也想從產品的消費中獲得價值，並期望這價值至少大於他們支付的價格。

原則上，所有團隊和功能都應以某種可測量的方式，為創造價值的過程做出貢獻；但在許多情況下，要量化這一點並不容易，DS 就必須和這種不可測量性打交道。

我編寫的作品《Analytical Skills for AI and Data Science》（O'Reilly），提出使用資料來創造價值的一般方法（圖 1-1）。這個想法很簡單：資料本身並不會創造價值，它的價值來自於使用做出決策的品質。一開始，您描述公司目前和過去的狀態，這通常會使用傳統的商業智慧（business intelligence, BI）工具，例如儀表板（dashboard）和報告等完成；而使用機器學習（machine learning, ML），則可以預測（*prediction*）未來狀態，並

試圖避免決策過程的困難性和不確定性；如果能夠自動化和優化（*optimize*）決策過程的某些部分，則是再好不過。本書的主題，就是在是幫助實務工作者使用資料做出更好的決策，這裡先不多說。

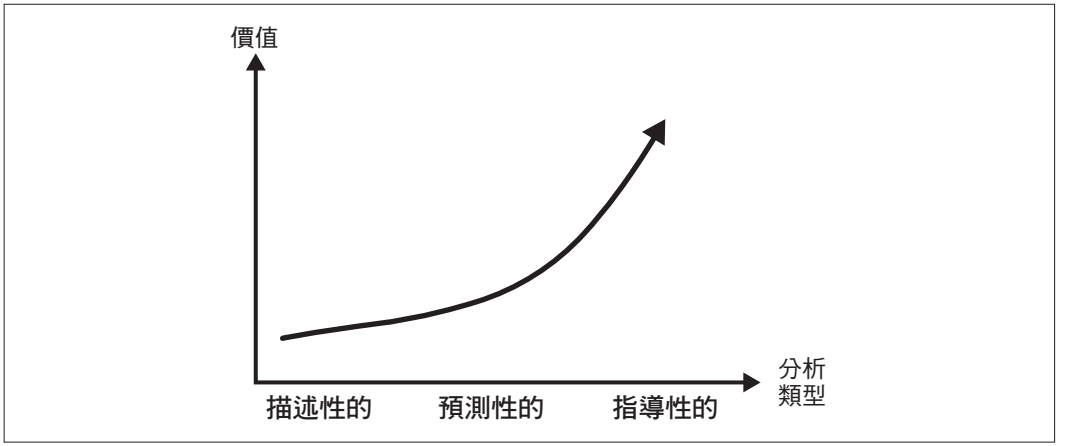

圖 1-1　使用資料建立價值

儘管這可能很直觀，但我發現這種描繪過於一般性和抽象，難以讓資料科學家在實務中使用，因此隨著時間過去，我將其轉化為一個框架，在介紹敘事主題時（第 7 章）也會派上用場。

大原則基本上不變：透過提升組織的決策能力，增加價值。為此，您真的需要瞭解手上的業務問題：什麼！（*what*），深入思考影響這些問題的手段：摘要重點！（*so what*），並對此採取積極主動的態度：現在該怎麼辦？（*now what*）。

什麼！瞭解業務

我常說，資料科學家應該和他們的利益相關者一樣瞭解業務。而在這裡，業務的範疇無所不包，從營運的事務，如瞭解和提議新的度量（第 2 章），以及他們的利益相關者可以操縱的手段，即拉動的「槓桿」（lever），到支撐業務的基礎經濟和心理因素，例如，驅使消費者購買產品的動機等。

對於資料科學家來說，要學習這些內容似乎很多，尤其是因為您需要不斷更新那些一直在發展的技術工具箱相關知識。但真的需要這樣嗎？難道不能只專注於演算法、技術堆疊和資料的有趣技術，而讓利益相關者專注於他們不那麼有趣的工作嗎？

我要先聲明的是，業務也很有趣！就算您不覺得這很令人振奮，如果資料科學家想要讓實際決策者聽到他們的聲音，絕對有必要贏得利益相關者的尊重。

在繼續之前，讓我強調一點，資料科學家很少是業務戰略和戰術的實際決策者：那實際上就是利益相關者，市場行銷、財務、產品還是銷售也好，也有可能是公司中的任何其他團隊。

要如何達成這一點？以下是我認為有用的一些建議：

參加非技術性會議

沒有教科書會教您業務的具體內容；您必須真的參與其中，並從組織中的集體知識中學習。

與決策者共事

確保參加會做出決策的會議。我在組織中最常為團隊用的理由就是，如果他們在場，對每個人來說都會是最有利；舉例來說，不瞭解業務細節的話，怎麼可能為模型策劃出色的特徵呢？

學習關鍵績效指標（*KPI*）

相對於組織的其他部分，資料科學家有一個優勢：他們擁有資料，並且經常收到計算和呈現團隊關鍵度量的要求，因此必須學習這些關鍵度量。這聽起來很理所當然，但許多資料科學家認為這很無聊，而且由於他們與度量方法無關，也就是說，他們很可能不用負責達到目標，所以很開心地將這個工作丟給利益相關者。但實際上，資料科學家應該要是度量設計的專家（第 2 章）。

保持好奇心並持開放態度

資料科學家應該永保好奇心。我指的是不要羞於提問問題，挑戰組織中的既定事實。有趣的是，我發現許多資料科學家缺乏這種整體好奇心，好在這是可以學習的，本章最後就會分享一些資源。

建立分散式結構

這可能無法取決於您，或您的上司或您上司的上司，但在公司這樣的團隊使用資料科學，將有助於達成業務專業化，甚至帶來信任等其他明顯正面性質。分散式的資料科學結構組織，會有各種不同背景的團隊成員，如資料科學家、業務分析師、工程師或產品等，並且能夠讓每個人都成為自己領域的專家；相反的，有一組「專家」讓整個公司組織的顧問中央集權化也具有優勢，但在獲得必要業務專業知識這方面就差了點。

摘要重點！在資料科學中創造價值的精髓

您的專案對公司的重要性為何？為什麼大家要在乎您的分析或模型？更重要的是，這會帶動什麼行動？這是本章討論的核心問題，順帶一提，這也是資料科學用來界定資深程度的屬性之一。在面試應徵者時，除了必要的技術篩選問題外，我都會馬上進入摘要重點！（*so what*）層面。

以下錯誤層出不窮：資料科學家花很多時間執行他們的模型或分析，但在展示時，只會顯示漂亮的圖表和資料視覺化，真的。

不要誤會，解釋手上資料非常重要，因為利益相關者通常不太瞭解資料，或資料的視覺化，尤其是對於較具技術性的東西，但他們絕對可以理解報告中的圓餅圖。不過您不應該止於此，第 7 章將介紹講述故事的實際情況，讓我提供一些發展這種技能的一般性指南：

從一開始就思考「摘要重點！」的重要性

每當決定啟動一個新專案時，我總是反向解決問題：決策者會如何使用我的分析或模型結果？他們手上有哪些方法？好不好用？這些問題都有答案之後，才能開始。

記錄下來

一旦您弄清楚了摘要重點！把它寫下來絕對是一個好習慣，不要讓它在許多技術層面淪為次要角色。很多時候，我們都會深陷技術細節而迷失了方向，如果能把它寫下來，這些摘要重點！將在絕望時成為您的指路北極星。

理解手段

摘要重點！的核心要能行動。您關心的 KPI 通常無法直接行動，因此您或公司中的某人需要發揮手段，嘗試影響諸如價格、行銷活動或銷售獎勵等度量。您需要深入思考可能的行動方案。同時，請隨時跳脫框架思考。

考慮您的受眾

他們在乎的是您在預測模型中使用的那些花俏新模型，還是在乎如何使用您的模型來改善他們的度量？我猜是後者：如果您能幫助他們成功，您也將成功。

第二章

度量設計

我有一個論點，出色的資料科學家，同時也會擅長度量（metric）設計。什麼是度量設計？簡單說，這是一門找出具有良好特性度量標準的藝術和科學。我會很快說一下其中的理想特性，但首先要說說資料科學家應該擅長這個領域的原因。

用最簡單的說法就是：如果不是我們，還有誰可以做這件事？理想情況下，組織中的每個人都應該擅長度量設計。但是，資料從業者最適合這項任務，因為資料科學家一直在使用度量：包括計算、報告、分析，並希望能夠優化它們。以 *A/B* 測試為例：每個好測試的起點都來自正確的輸出度量；機器學習（ML）也有類似的基本原理：獲取正確的結果度量至關重要。

度量應該具備的特性

公司為什麼需要度量？如同第 1 章所論述的，好的度量可以帶動行動。有了這個成功準則，讓我們逆向思考這個問題，並確定成功的必要條件。

可測量

度量在定義上就是可測量的。不幸的是，許多度量並不完美，學會認出它們的缺點將讓您受益匪淺。所謂的代理（*proxy*）度量，或直接稱代理（*proxy*），通常與期望的結果相關，需要瞭解使用它們的利弊[1]。

1 例如，在線性迴歸中，特徵上的測量誤差會使參數估計產生統計偏差。

有個簡單的例子是意圖性（*intentionality*）。假設您想瞭解早期流失（*early churn*），也就是新使用者的流失原因，因為有些人實際上並未打算使用產品，只是試用而已；如此，測量意圖性將能有效改善您的預測模型。但意圖性實際上並不容易測量，因此需要尋找代理，例如瞭解應用程式的用法之後，和開始使用它之間的時間間隔。我認為，開始使用應用程式的速度越快，就代表有越強烈的意圖。

另一個例子是增長從業者使用的習慣（*habit*）概念。應用程式的使用者通常會完成引導過程、嘗試產品（發出「啊哈！」的時刻）、並希望能藉此養成習慣。哪些證據表明使用者達到了這個階段？這裡常見的代理是使用者第一次嘗試使用後 *X* 天間的互動次數。對我來說，「習慣」完全與重複（*recurrence*）有關，不論這對不同使用者來說代表什麼意思。以此來看，代理最多只是重複的一個早期指標。

可行動性

為了推動決策，度量必須具有可行動性（actionable）；不幸的是，許多最高層級（top-line）度量不具行動性。例如收入（*revenue*）：它取決於使用者購買產品，而這是無法強迫的。但是，如果您將度量分解為子度量，可能會出現一些有效的手段，正如我等一下展示的範例。

相關性

度量對於手上的問題是否具有資訊性？可稱之為相關性（*relevance*），因為它強調度量只對特定業務問題具有價值。也可以使用資訊性（*informative*），但是所有度量都對某事具有資訊性，而相關性則是指擁有正確度量，以正確解決問題的特性。

及時性

好的度量在您需要它們時會推動行動。如果我發現自己已是癌症末期，醫生也沒辦法做太多事情；但是如果我定期檢查，他們可能會發現早期症狀，從而為我提供多種治療方案。

顧客流失是另一個例子。通常會使用長度為一個月的不活動視窗以測量和報告，上個月活躍的使用者在下個月轉為不活躍的百分比。不幸的是，這個度量可能會產生偽陽性：有些使用者只是休息一下，並未真正流失。

獲得更強固度量的方法之一，是將不活躍視窗從一個月增加到三個月，時間視窗越長，使用者只是休息一下的可能性就越小。但是新的度量在及時性（timeliness）方面已經降低：您現在必須等待三個月才能標記一個已流失的顧客，而這樣可能會來不及發起留客活動。

度量分解

透過分解度量，可能可以改善這些屬性中的任何一個，以下將詳細介紹一些技巧，會幫助您達成這一目標。

漏斗分析

漏斗（funnel）是一系列依序進行的動作。例如，在前面的習慣範例中，使用者首先需要設定他們的帳戶、嘗試產品、然後定期使用；若是擁有漏斗，就可以使用簡單技巧來找到子度量。我將先以抽象方式展示這個技巧，再提供一些簡單範例。

圖 2-1 為典型漏斗：從入口點 E 到輸出 M 的階段序列（此處的這些符號同時也都代表相應度量）。我的目標是改進 M，內部階段表示為 s_1、s_2、s_3，每個都提供一個具有相應索引的度量。

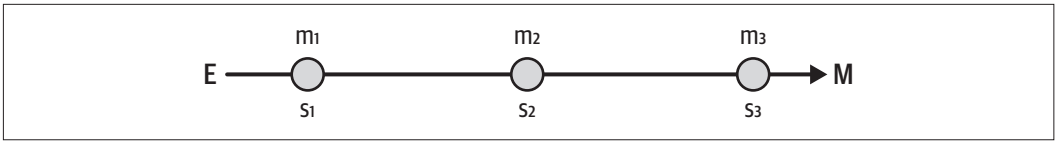

圖 2-1　典型漏斗

分解方法如下：從右到左移動、乘以當前的子度量、然後除以前一個子度量。為了確保您永遠不會失去相等性，請在結束時乘以漏斗開始處的度量（E）。請注意，在取消公共項之後，最終結果是 $M = M$，確保它的確分解自原始度量。

$$M = \frac{M}{m_3} \times \frac{m_3}{m_2} \times \frac{m_2}{m_1} \times \frac{m_1}{E} \times E$$

每個分數都可以解釋為轉換率，即在前一階段可用的單元中，有多少百分比達到了目前階段。通常，這些子度量中的一個或全部都具有比原始度量 M 更好的性質。瞭解這個技巧後，就可以將其付諸實踐了。

典型的銷售漏斗會如下，我的目標是增加銷售，但這需要幾個步驟，這裡會簡化漏斗：

- 潛在顧客產生（L：潛在顧客數量）
- 第一次聯繫（C_1：第一次聯繫的次數）
- 第二次聯繫（C_2：第二次聯繫的次數）
- 提出報價（O：提出的報價次數）
- 成交銷售（S：銷售數量）

分解如下：

$$S = \frac{S}{O} \times \frac{O}{C_2} \times \frac{C_2}{C_1} \times \frac{C_1}{L} \times L$$

要增加銷售次數，可以增加潛在顧客數量，或增加各階段之間的轉換率。這之中的某些行動會與資料科學家有關，例如提高潛在顧客的品質；而其他行動則可能與銷售團隊有關，例如，是否有足夠的第一次聯繫；沒有的話，公司可能需要增加銷售人員的規模或聘請不同人員。也許他們應該改變談判或價格策略，以提高報價成交率，甚至可以針對產品改善！也是有可能已擁有最好潛在顧客或最好銷售團隊，但仍然缺乏產品 - 市場契合度。

庫存 - 流量分解

想知道累積度量時，*庫存 - 流量分解*（*stock-flow decomposition*）就很有用。首先要先定義以下概念：*庫存*（*stock*）變數是會累積並在特定時間點測量的變數；*流量*（*flow*）變數則不會累積，而且是在一段時間內測量。最好的例子就是以浴缸說明：時間 t 的水量等於時間 $t-1$ 的水量、加上這兩個時間點之間開水龍頭而灌入的水量、再減去流失水量。

想瞭解月活躍使用者（Monthly Active User, *MAU*）時最常用到這個方法。以下先寫出分解方式再來評論：

$$MAU_t = MAU_{t-1} + \text{新增使用者}_t - \text{流失使用者}_t$$

如果目標是增加公司的 *MAU*，方法有獲取更多顧客，或減少顧客流失量。新增使用者（*Incoming Users*）可能分為新使用者（*New Users*）和回頭使用者（*Resurrected Users*），這樣提供至少一種新的控制手段。

第十四章

從預測到決策

據麥肯錫（McKinsey）的一項調查（*https://oreil.ly/Kl_7y*）顯示，他們的受訪組織中，有 50% 在 2022 年採用了人工智慧（AI）或機器學習（ML），相對於 2017 年大幅增加 2.5 倍，但仍低於 2019 年的 58%。如果 AI 是新的電力（*https://oreil.ly/O_tsb*），資料是新的石油（*https://oreil.ly/bU0xd*），為什麼在大型語言模型（large language model, LLM）如 ChatGPT 和 Bard 出現之前，它們的採用率會陷入停滯呢[1]？

儘管有各式各樣的根本原因，但最直接的原因是，大多數組織尚未找到正面的投資報酬（return on investment, ROI：*https://oreil.ly/Stpro*）。在〈Expanding AI's Impact With Organizational Learning〉（*https://oreil.ly/izJb7*）一文中，Sam Ransbotham 及其合作者認為，只有「10% 的公司，會從人工智慧技術中獲得明顯的經濟利益」。

這個 ROI 來自哪裡？在其核心，由於機器學習演算法是預測性程序，因此合理地期望大多數價值是透過改善決策能力而建立。本章將探討預測改進決策的一些方式，在此過程中，我將介紹一些實用的方法，幫助您從預測轉向改進決策。

剖析決策

預測演算法試圖繞過不確定性，這樣做對提高我們的決策能力非常重要。例如，我可以純粹為了高興而試圖預測家鄉明天的天氣，但是預測本身可以促使並改善我們在面對這種不確定性時，做出更好決策的能力。很容易就可以找到各種不同的人和組織，都願意為這些資訊付費的使用案例，如農民、派對策劃者、電信行業或 NASA 等政府機構等。

1 有人可能會想問，大型語言模型（LLM）是否真的會以明顯方式改變採用趨勢。我認為基本立論至今還未真正改變，至少在機器達到人工通用智能（artificial general intelligence, AGI）之前。第 17 章也會討論這個主題。

圖 14-1 以圖表方式顯示不確定性在決策中的角色。從右側開始，一旦解決不確定性，就會有一個能影響您所關心的某個度量結果，這個結果取決於您掌握的槓桿（行動）集合，以及它們和底層不確定性的交互作用。例如，您不知道今天是否會下雨（不確定性），但您想保持舒適和乾爽（結果），您可以決定要不要帶傘（槓桿）。當然，如果下雨，帶傘是比較好的選擇，因為能讓您保持乾爽；但如果不下雨，最好的決策就是不帶傘，您會因無須攜帶它而感到更舒適。

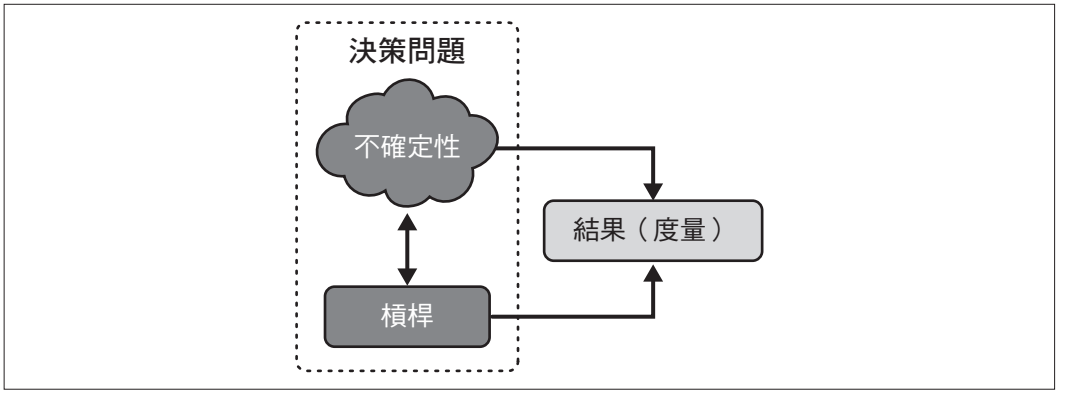

圖 14-1　不確定情況下的決策

表 14-1 彙總了一些 ML 常見使用案例，強調決策和不確定性的角色，以及一些可能的結果。讓我們來看一下第一列，即健康保險理賠處理的案例。如果有一份新的理賠，您必須決定手動審查還是批准支付，因為理賠可能是違規的，而違規的理賠會不必要地增加保險公司的成本，但審核過程通常相當複雜，需要大量的時間和精力。如果您能夠正確預測，就可以降低預測誤差和成本，同時又提高顧客滿意度。

表 14-1　ML 使用案例範例

類別	使用案例	決策	不確定性	結果
服務營運	理賠處理	自動支付與審核	是否違規	降低手動流程（成本），提高顧客滿意度，減少詐欺
服務營運	人員配置	雇用或重新安置	員工規模取決於需求	提高顧客滿意度，減少未使用的資源（成本）
服務營運	主動顧客支援	要不要打給顧客	顧客是否有我可以解決的問題	提高滿意度，減少流失
供應鏈優化	需求預測	管理庫存	庫存取決於需求	提高銷售，降低折舊成本

類別	使用案例	決策	不確定性	結果
詐欺偵測	退款預防	批准或拒絕交易	是否違規	降低與詐欺相關的成本，提高顧客滿意度
行銷	潛在顧客產生	要不要打給潛在顧客	他們會不會購買	提高銷售效率
基於 ML 的產品	推薦系統	推薦 A 或 B	他們會不會購買	提高參與度，減少流失

首先考慮決策和結果，然後再考慮 ML 應用，這樣可以讓您在組織內發展強大的資料科學實務方面走得更遠。

在職場尋找新的 ML 使用案例時，思考決策和槓桿是一種很好的方法。該過程包括：

1. 確定利益相關者所做的關鍵決策，以及相關度量和槓桿。
2. 瞭解不確定性的作用。
3. 為建構 ML 解決方案提出商業案例。

智慧型閾值的簡單決策規則

與迴歸不同，簡單的決策規則在分類模型中，會以閾值化（*thresholding*）的形式自然產生。我將描述二項模型，即兩個結果的情況，但這個原則可以調整為更一般的多項模型。典型的情境可能如下：

$$做(\tau) = \begin{cases} 若\ \hat{p}_i \geq \tau\ 則\ A \\ 若\ \hat{p}_i < \tau\ 則\ B \end{cases}$$

在這裡，$\hat{p}_i$ 是單元 i 的預測機率分數，τ 是您選擇的閾值，該規則在分數夠高時，會激發行動 A；在分數不夠高時會激發行動 B。請注意，如果您用預測的連續型結果來替換預測機率，則適用類似原理。然而，分類設定固有的簡化結構，讓您可以在深思熟慮後納入不同預測錯誤的成本。

簡而言之，一切都歸結於更深入的理解偽陽性和偽陰性。在二項模型中，結果通常標記為正（1）或負（0），一旦有了預測的機率分數和閾值，具有較高機率的單元會預測為正，反之為負。請參見表 14-2 中的混淆矩陣（confusion matrix）。

表 14-2　典型混淆矩陣

實際 / 預測	$\widehat{N}(\tau)$	$\hat{P}(\tau)$
N	TN	FP
P	FN	TP

混淆矩陣中的列和行分別表示實際標籤和預測標籤。如前所述，預測結果取決於所選閾值（τ），因此，根據預測標籤是否與實際標籤匹配，您可以將樣本中的每個實例分類為真陰性（true negative, *TN*）、真陽性（true positive, *TP*）、偽陰性（false negative, *FN*）或偽陽性（false positive, *FP*）。矩陣中的單元格表示每個類別的案例數量。

精確度和召回率

在分類問題中，兩個常見的效能度量是精確度（precision）和召回率（recall）：

$$\text{精確度} = \frac{TP}{TP + FP}$$

$$\text{召回率} = \frac{TP}{TP + FN}$$

這兩個度量都可以視為真陽性率，但每個考慮情況各自不同。[2] 精確度回答了這個問題：*在所有我說的正面案例中，實際上有多少百分比是正面的？*另一方面，召回率回答的是這個問題：*在所有實際上為正面的案例中，我預測正確的百分比是多少？*當您將精確度作為考慮因素時，實際上在考慮偽陽性的成本；而對於召回率，重要的是偽陰性的成本。

圖 14-2 顯示 3 個不同模型的精確度和召回率曲線，這些模型來自一個產出平衡結果的模擬潛在變數線性模型上的訓練，第一行顯示一個分類器，在單元間隔中抽取隨機均勻數來分配機率分數；這個隨機分類器將作為基線。中間一行繪製從羅吉斯迴歸獲得的精確度和召回率，最後一行刻意地切換預測類別，以建立一個反機率分數，其中較高的分數與較低的發生率相關聯。

您可以很容易地看到幾種模式：精確度始終以樣本中正向案例的一部分開始，它可以是相對直的線（隨機分類器）、增加或減少。大多數情況下，您會得到一個增加的精確度，因為大多數模型傾向於優於隨機分類器，並且至少在某種程度上對您想要預測的結果具有一定資訊性。儘管這在理論上是可能的，但精確度不太可能出現負斜率。

2　注意，在機器學習文獻中，召回率通常會視為真陽性率（*true positive rate*）。

精確度的表現較好，它會從一開始然後下降到 0，只有曲率發生變化。一個良好的凹函數（中間的案例）通常是可以預期的，這也與在良好的分類模型中，分數能夠描述發生機率的事實相關聯。

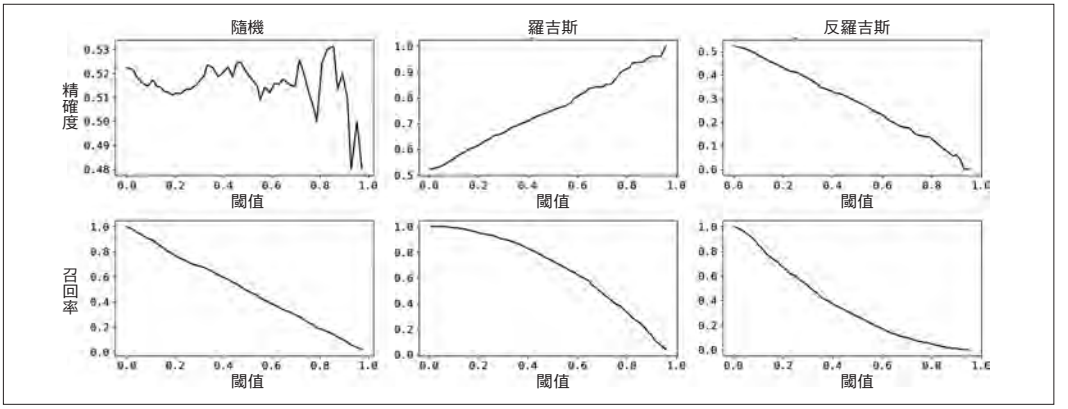

圖 14-2　不同模型的精確度和召回率

範例：潛在顧客產生

以潛在顧客產生（lead generation）活動為例，您對潛在顧客評分，以預測最終哪些顧客會購買。您的資料包括以前由電話銷售團隊使用的潛在顧客樣本中，成功銷售和失敗未銷售的聯繫紀錄。

考慮一個簡單的決策規則，即在預測的機率高於某一閾值時再與顧客聯繫。FN 是一個本應交給行銷團隊的潛在顧客，但由於沒有交給他們，未能完成銷售；FP 是錯送給行銷團隊的潛在顧客，因此最終未轉化為銷售。偽陰性的成本是由於未能銷售而產生的收入損失，而偽陽性的成本則是用於處理潛在顧客的任何資源，例如，一名電話業務代表的每小時薪資為 *$X*，為每個潛在顧客的處理時間為 *k* 分鐘，則每個偽陽性的成本為 *$kX/60*。

簡單的處理量（*volume*）閾值規則運作方式如下：銷售團隊告訴您每個週期，如一天或一週他們可以處理的銷售量（*V*），然後您根據估計的機率分數，向他們發送前 *V* 名潛在顧客。顯然，透過固定處理量，您也內隱式地設定您的決策規則閾值。