ORACLE

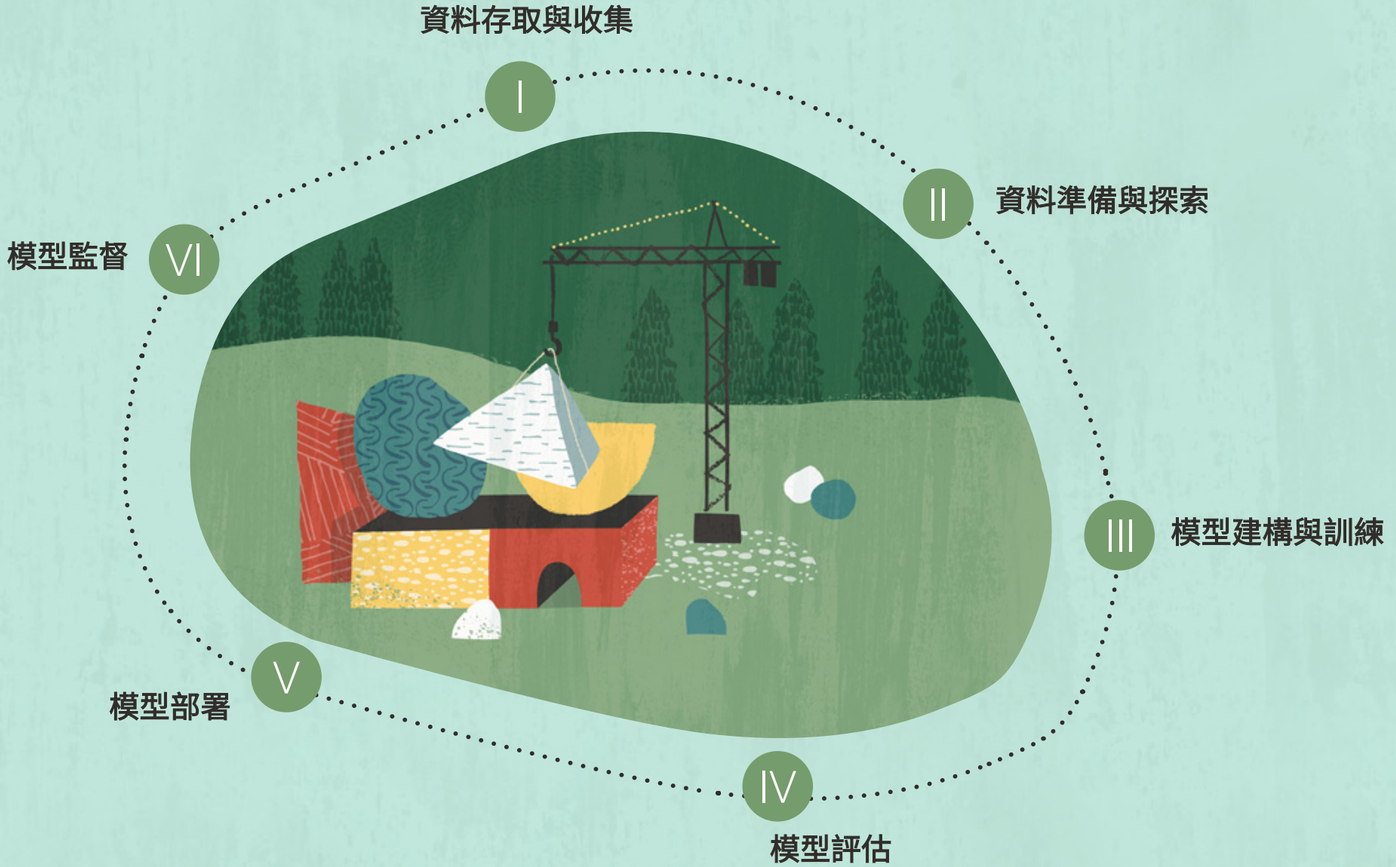
# 機器學習模型的生命週期

# 簡介

越來越多企業投資機器學習領域或視其為潛在成長領域。從能夠利用資料尋找有關客戶的深入解析，到提高流程效率，都是企業投資機器學習的原因。我們在本書將機器學習模型的建構方式分解為六個步驟：資料存取和收集、資料準備與探索、模型建構與訓練、模型評估、模型部署和模型監督。

建立機器學習模型是反覆的過程。構建機器學習模型所需的許多步驟可重複及修改，直到資料科學家對模型效能滿意為止。此過程需要大量的資料探索，視覺化和實驗，因為每個步驟必須獨立探索、修改及審核。

## 建構機器學習模型的步驟



# I. 資料存取與收集

機器學習問題的第一步是存取資料。通常，資料科學家會查詢公司儲存資料的資料庫，針對處理中的商業問題取得資料。此外，非結構化資料集也有許多價值，這些資料集不適合關聯式資料庫 (例如記錄、原始文字、圖像、影片等)。這些資料集經由資料工程師和資料科學家編寫的擷取、轉換、載入 (ETL) 管道進行大量處理。這些資料集可能位於資料湖或資料庫中 (無論是否關聯)。 當資料科學家缺乏解決問題所需的資料時，他們可從網站抓取資料，向資料提供者購買資料或從問卷、點選流資料、感測器、攝影機等等收集以便獲得資料。

# II. 資料準備與探索

在取得資料之後，資料科學家必須準備原始資料、執行資料探索、視覺化資料、轉換資料，並可能重複這些步驟，直到適合建立模型。資料準備是在分析之前清理並處理原始資料。在建構任何機器學習模型之前，資料科學家需要瞭解可用的資料。原始資料可能混亂、重複或不準確。資料科學家會探索他們可用的資料，透過取代或刪除損毀、不準確及不完整的資料加以清理。

此外，資料科學家需要確定資料是否具有標籤。例如，如果您有一系列圖像，並且想要開發偵測模型來判斷圖像中是否包含汽車，則需要有一組圖像標示其中是否包含汽車，而且很可能圖像中的汽車周圍需要邊界方框。如果圖像缺少標籤，資料科學家則必須加上標籤。有開源工具和商業供應商可提供資料標籤平台，並可供僱用人工來加上標籤。

在清理資料之後，資料科學家會探索資料集的特徵(或變數)，識別特徵轉換之間的任何關係。資料科學家可以使用各種工具在開放原始碼程式庫和分析/資料科學平台進行探索性資料分析。在此步驟中，可執行資料集統計分析並建立資料視覺效果來產生特徵繪圖的工具非常有用。

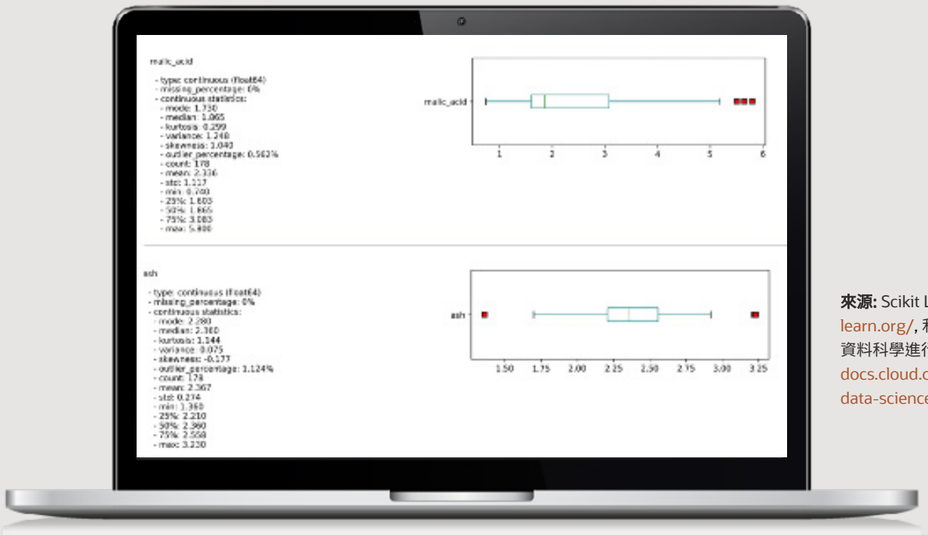
查看資料集包含哪些特徵類型很重要。特徵可以是數字，包含浮點數或整數。分類特徵包含有限數量的可能值，通常會將資料指派為群組。例如，如果您的客戶調查資料集，受訪者的性別 (男性或女性) 為分類特徵。序數特徵是包含既定順序或級數的分類特性。例如，客戶滿意度回應：非常滿意、滿意、無所謂、不滿意和非常不滿意都有既定順序。您可把順序轉換為整數的級數 (1->5)。在確定了有哪些特徵之後，接下來將獲得每個特徵所包含的值分佈以及摘要統計資訊。這樣有助於回答以下有關資料集的問題：

- 資料集是否偏向某個值範圍或類別子集？
- 特徵的最小值、最大值、平均值、中間值和眾數值是多少？
- 是否包含遺漏值或無效值 (例如空值)？如果是，有多少個？
- 資料集是否包含異常值？

在資料探索步驟中，繪製特徵甚至交互繪製有助識別資料集的模式。這有助於判斷資料轉換的必要性。您需要回答的問題包含：

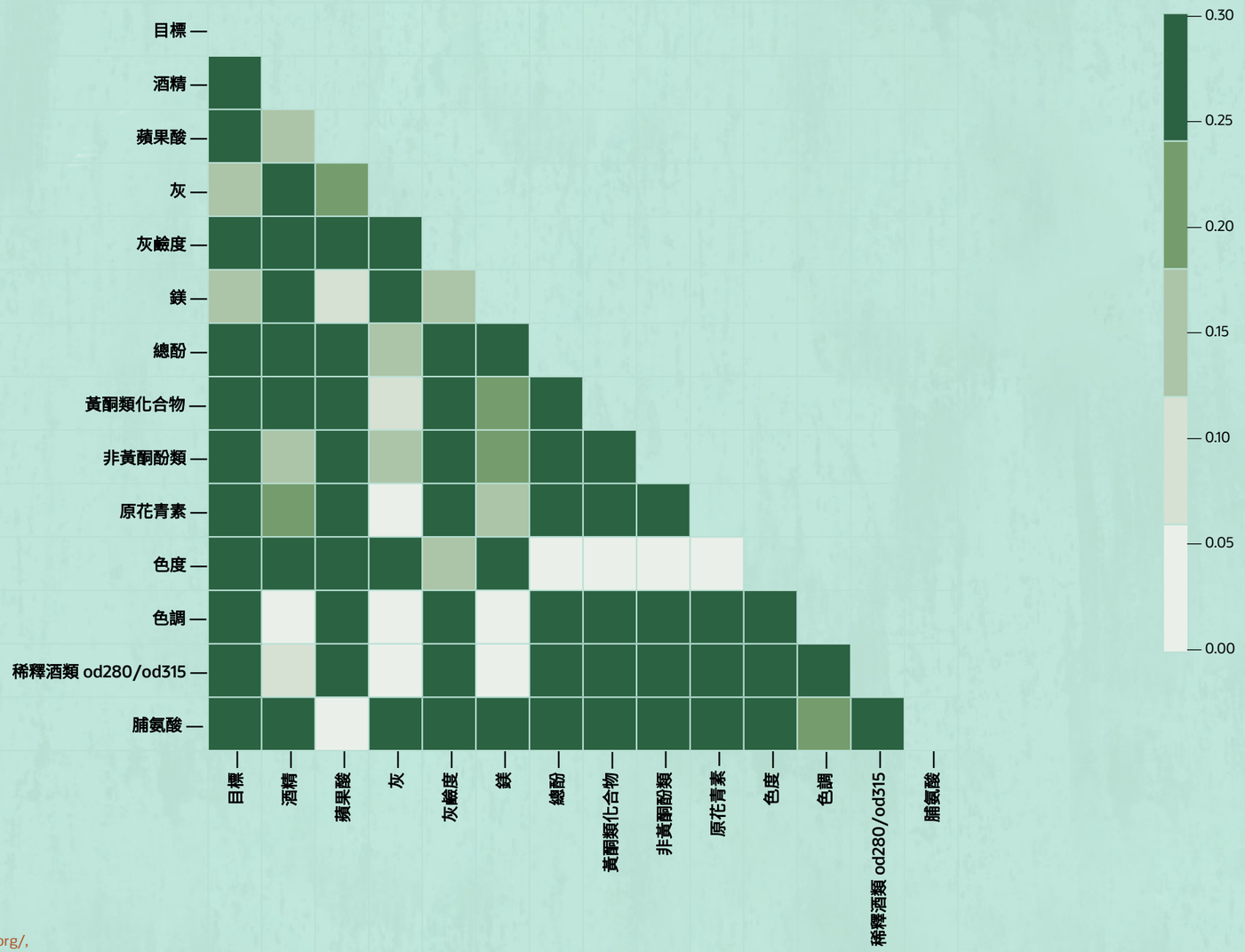
- 如何處理遺漏值？您是否要填入這些值？如果是，您打算採用什麼方法來填入遺漏值？方法包含取平均值、中位數、眾數、附近項目值和附近項目的平均值。
- 如何處理異常值？
- 您的部分特徵是否相互關聯？
- 您是否需要對資料集進行歸一化或執行其他轉換以重新調整資料 (例如記錄轉換)？
- 分類值長尾的處理方法是什麼？您是否按原樣使用，以某種有意義的方式進行分組，還是完全忽略其子集？

**三種葡萄酒資料集特徵的摘要統計和視覺化，以及每種葡萄酒的特徵。**




**來源:** Scikit Learn https:// scikit-learn.org/, 利用 Oracle 雲端基礎結構資料科學進行視覺化 https://docs.cloud.oracle.com/en-us/iaas/data-science/using/data-science.htm

**從包含三種葡萄酒和每一種葡萄酒之特色的資料集為基礎，顯示這些特色如何相互關聯的熱圖。**




**來源 :** Scikit Learn https://scikit-learn.org/, 利用 Oracle 雲端基礎結構資料科學進行視覺化 https://docs.cloud.oracle.com/en-us/iaas/data-science/using/data-science.htm

在資料探索步驟中，您可以識別資料集的圖樣，瞭解如何開發更能代表資料集的新特徵。這稱為特徵工程。例如，如果您有個交通資料集有關每小時通過主要十字路口的車輛數量，您希望建立新特徵，將小時分類為一天的不同部分，例如清晨、中午、下午、傍晚和夜間。

對於分類特徵，通常需要針對特徵進行熱編碼。熱編碼代表將分類特徵轉換為二進制特徵，每個類別一個。例如，假設您有個客戶資料集，同時包含客戶來自哪個州的特徵：華盛頓、俄勒岡和加利福尼亞。一個熱編碼會產生兩個二進制特徵，其中一個特徵是客戶是否來自華盛頓，第二個特徵是客戶是否來自俄勒岡。據推測，如果客戶並非來自華盛頓或俄勒岡州，他/她將來自加利福尼亞，因此無須第三個特徵。

# III. 模型建構與訓練

模型建構包含選取正確的機器學習模型來解決問題以及納入模型的特徵。在模型建構的第一步，資料科學家需要決定適當的機器學習模型可能是什麼，以解決問題。機器學習模型包含兩種主要類型：受監督和無監督。監督學習牽涉為一組輸入資料建立模型來產生輸出或標籤。分類及迴歸屬於監督學習的問題。無監督學習牽涉為一組沒有標籤的輸入資料建立模型。例如，客戶群屬於無監督的學習問題。您事先不知道客戶屬於哪個客戶群。客戶群由模型指派。

不同類別的機器學習模型用於解決無監督和有監督的學習問題。通常，資料科學家會嘗試多個模型和演算法，並產生多個候選模型。資料科學家事先不知道哪種模型在資料集上表現最好，因此他們會對其中幾個進行試驗。在模型訓練期間，資料科學家可能會進行特徵選擇，這是僅選擇特徵子集作為機器學習模型輸入的過程。降低輸入變數量的好處在於降低模型訓練的運算成本，使模型更加普遍化，並可能改善模型效能。

在模型訓練期間，資料集會分割成訓練集和測試集。訓練資料集是用來訓練模型，而測試資料集則是用來針對未見過的資料查看模型執行效能。模型評估將在下面詳細討論。

模型超參數調整是模型訓練過程的主要任務。模型是演算法，而資料科學家可利用超參數調整改善模型效能。例如，決策樹的深度為超參數。

您可選擇極深或極淺的決策樹。這將影響模型的偏差和變異。偏差是由於學習不足或未擷取特性與輸出之間的關係所造成的錯誤。變異是模型在訓練資料集中表現良好但對於未見過的資料表現不佳的過度學習錯誤。調整模型的超參數可以部分自動化，儘管資料科學家應該參與該過程。

資料科學家也必須決定訓練模型所需的運算資源類型。您可在電腦上準備資料並在本機訓練模型。然而，視訓練模型需要準備多少資料而定，您的電腦可能無法勝任。您可能必須將工作負載轉移到雲端，您可在此選擇存取包含 GPU 在內的更多運算資源。

部分模型在專用硬體可更快速進行訓練 (例如，在 GPU 訓練感知/深度神經網路模型)。您也可探索能夠加速程序的分散式訓練環境，尤其是當資料量無法放入機器最大可用記憶體時，將資料分割並分配到多台機器，或者當您想要同時在單獨的機器訓練多個候選模型時。

# AutoML

AutoML 在過去幾年獲得了相當多的關注，因為其有望讓機器學習容易為更多受眾所接受。AutoML 代表自動化機器學習。它自動化了特徵選擇、模型/演算法選擇及超參數調整的過程。這是所有主要資料科學平台都會包含的功能。使用者可將資料集提供給 AutoML，訓練多個機器學習模型、調整這些模型的超參數，彼此評估其效能。

AutoML 可透過自動化訓練過程提高資料科學家的生產力。它讓資料分析師和開發人員無需透過數據科學專業知識來調整模型訓練過程的各個層面，便能建構機器學習模型。大多數 AutoML 功能都支援分類及迴歸問題的表格式資料，而其他功能則包含支援圖像及文字資料，以及時間序預測的較進階產品。

AutoML 或任何複雜模型的缺點是它看起來像是黑箱解決方案，使用者難以瞭解模型如何獲得預測結果。使用者可查看 AutoML 系統提供的模型解釋能力，瞭解有哪些功能可幫助使用者解釋模型，並瞭解選定模型如何得出預測結果。

模型解釋通常分為全域解釋和區域解釋。全域解釋是瞭解整個機器學習模型的一般行為。這包含解釋每個特徵對於模型預測的重要性。區域解釋提供了機器學習模型如何針對資料樣本進行特定預測的說明。例如，為什麼詐騙偵測演算法會將特定交易預測為詐騙？

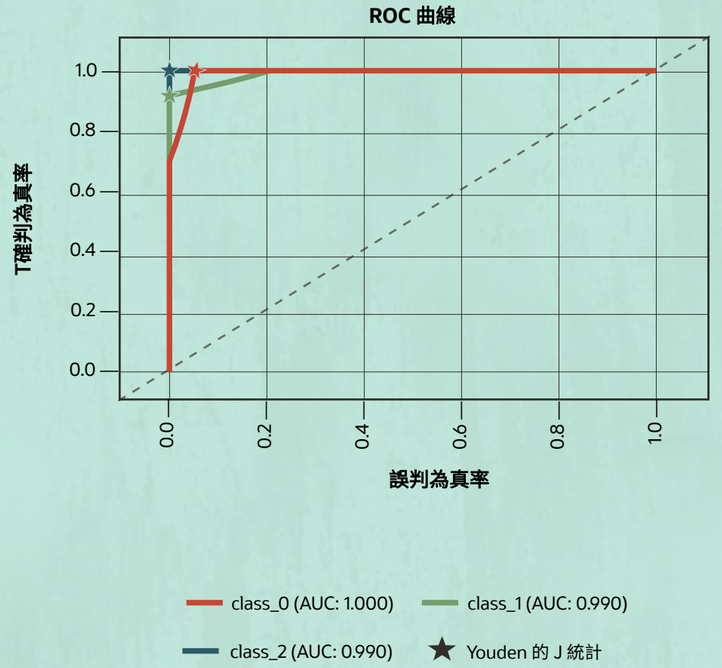
# IV. 模型評估

有許多開放原始碼工具可協助資料科學家來計算指標以及視覺化指標 (例如 AUC-ROC 曲線、增益和提升圖) 以便評估機器學習模型。在評估機器學習模型時，資料科學家需要決定哪些指標對於他們試圖解決的商業問題很重要。

對於分類問題，可利用精確度進行模型評估，但有時可能並非最佳衡量標準。如果問題牽涉偵測某人是否患有罕見疾病，則更好的衡量標準可能是準確診斷出的罹病人數除以所有罹病人數。在這種情況下，查看顯示真陽性、真陰性、偽陽性和偽陰性數量的混淆矩陣，計算精度並重新叫用會更有效。對於迴歸問題，您可利用諸如均方根誤差、平均絕對誤差等指標或計算決定係數 r2。對於無監督的問題，在叢集內及之間具有高度一致性的一組叢集較為理想。這可透過輪廓得分和 Calinski-Harabasz 係數等指標來衡量。

**多類分類結果的混淆矩陣，根據葡萄酒特徵預測葡萄酒類型的隨機森林模型，來自包含三種葡萄酒類型和每種葡萄酒特徵的資料集。**


**多類分類 ROC 曲線係為隨機森林模型的結果而建構，模型從包含三種類型的葡萄酒及每種葡萄酒的特徵資料集來預測葡萄酒類型。**




**來源：** Scikit Learn Library https://scikit-learn.org/, 利用 Oracle 雲端基礎結構資料科學進行視覺化 https://docs.cloud.oracle.com/en-us/iaas/data-science/using/data-science.htm

# V. 模型部署

在完成模型訓練和評估流程之後，會儲存最佳候選模型。模型通常以 Pickle、ONNX 及 PMML 格式儲存。視目標而定，資料科學家可能會針對機器學習問題進行概念證明、實驗或將其部署到生產環境。模型部署係以某種方式利用機器學習模型所做的預測。極有可能，也必須部署資料轉換的管道。通常，資料科學家會搭配工程師進行模型部署。

根據您打算預測的方式，您可以部署為批次用途或即時用途。對於批次用途，可安排預測時間 (例如，每小時、每天)。然後可將預測結果儲存在資料庫，並由其他應用程式運用。通常，您處理的資料量大於即時預測。例如，如果您經營電子商務網站，並且您希望每週向客戶傳送電子郵件，根據過去的購買情況向他們推薦產品。可以排定提前執行機器學習模型。

對於即時用途，將起始處理程序，使用常駐模型來提供預測。例如，在開始付款時決定交易是否為欺詐，需要即時預測。您必須考慮提供預測服務的速度（毫秒、秒？）、服務需求量，以及執行預測的資料大小。盡可能減少預測服務的延遲相當重要。您可使用尺寸更小的模型，使用 GPU 等加速器，改進即時預測檢索實體相關特徵的方式來改善服務延遲（例如，如果您在使用者瀏覽網站時向使用者推薦產品，改進如何獲取使用者過去的購買資訊可以改善延遲。）

不同工具和雲端平台產品可用於模型部署，例如，功能即服務 (FaaS) 平台、針對 HTTP 端點模型完全受管理的部署、在容器協調器平台 (例如 k8 及 docker swarm 等等) 透過 flask 或 Django 自助部署。

# VI. 模型監督

模型監督是個具有挑戰性的步驟，有時，缺乏成熟機器學習及資料科學計畫的機構，會遺漏了這項步驟。模型的重新訓練及重新部署需要資料科學與工程團隊，以及運算資源的時間。模型監督可協助團隊決定需要重新訓練模型並重新部署的必要性與時間。模型監督可分為兩個部分：模型效能的飄移/統計監督及運轉監督。

在部署模型之後，衡量及訓練模型的指標會在生產中下滑。這是由於資料並非靜止不動。非穩定性經由數個方面表現：生產資料的特徵可能出現訓練資料集範圍外的值；值的分佈可能會出現緩慢漂移等等。

由於模型退化的緣故，需要監督模型品質，決定是否以及何時重新訓練並重新部署模型。有時無法立即進入生產系統獲得即時資料的預測精度。例如，您可能需要一些時間才能確定客戶流失預測模型或欺詐偵測模型是否提供了準確的預測。然而，可以將訓練資料的統計結果及分佈跟即時資料進行比較，並將模型預測的分佈與訓練跟即時資料進行比較。例如，如果您正在使用客戶流失模型，您可以將用於訓練模型的客戶特徵跟生產系統的客戶特徵進行比較。此外，您還可查看在訓練樣本中預計流失的客戶相較於現場生產的百分比。

機器學習系統的運轉監督需要資料科學家及工程團隊之間的合作。要監督的內容包括服務延遲、記憶體/CPU 使用率、輸送量和系統可靠性。需要設定記錄與指標來進行追蹤及監督。記錄包含事件的記錄，以及發生的時間。它們可用於調查特定事件並找出事件的原因。Kibana 是個用於搜尋並查看記錄的開放原始碼工具。指標衡量機器學習系統的使用與行為。Prometheus 和 Grafana 為適合監督指標的工具。

# 結論

我們希望這是實用的指南，說明建立機器學習模型.所需要的步驟。請務必記住，機器學習是非常反覆的過程，需要多次重覆並改善本書所述的步驟。

有許多資源可深入探討本書涵蓋的每個步驟，您可以在針對相關企業資料科學策略制定決策時加以深入瞭解。如果您已準備好，Oracle 會提供實作實驗室，讓您嘗試建構自己的資料科學模型。


## Oracle corporation

全球總部
500 Oracle Parkway, Redwood Shores, CA 94065, USA

全球諮詢
電話 + 1.650.506.7000 + 1.800.ORACLE1
傳真 + 1.650.506.7200

oracle.com


## 聯絡我們

撥打 +1.800.ORACLE1 或造訪 oracle.com. 在北美以外的地區，請前往 oracle.com/contact. 尋找您當地的辦事處。

blogs.oracle.com/oracle

facebook.com/oracle

twitter.com/oracle

## 作者


Wendy Yip, 資料科學家。



版權所有 © 2020, Oracle 和 (或) 其關係企業。保留一切權利。本文件僅作為資訊用途，如有變更恕不另行通知。本文件不保證無錯誤，也不受任何其他口頭表達或法律默示之擔保或條款約束，包括默示擔保與適售性或特定用途適用性條款。我們特別聲明，不擔負因本文件所生之損害賠償責任，並且也不會透過本文件形成直接或間接之契約義務。未經我們事先書面許可，不得出於任何目的以任何形式或任何方式 (電子或機械) 複製或傳播本文件。

Oracle 和 Java 是 Oracle 和 (或) 其關係企業的註冊商標。其他名稱為各商標持有人所擁有之商標。

Intel 和 Intel Xeon 是 Intel Corporation 的商標或註冊商標。所有 SPARC 商標的使用皆經過授權，且是 SPARC International, Inc. 的商標或註冊商標。AMD、Opteron、AMD 標誌與 AMD Opteron 標誌是 Advanced Micro Devices 的商標或註冊商標。UNIX 是 The Open Group 05.10.19 的註冊商標。