aws

開發人員指南

Amazon Machine Learning

版本 Latest

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon Machine Learning: 開發人員指南

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商標和商業外觀不得用於任何非 Amazon 的產品或服務，也不能以任何可能造成客戶混淆、任何貶低或使 Amazon 名譽受損的方式使用 Amazon 的商標和商業外觀。所有其他非 Amazon 擁有的商標均為其各自擁有者的財產，這些擁有者可能附屬於 Amazon，或與 Amazon 有合作關係，亦或受到 Amazon 贊助。

Table of Contents

..... viii
什麼是 Amazon Machine Learning？ 1
Amazon Machine Learning 關鍵概念 1
資料來源 1
ML 模型 3
評估 4
批次預測 4
即時預測 5
存取 Amazon Machine Learning 5
區域與終端節點 6
Amazon ML 定價 6
估計批次預測成本 7
估計即時預測成本 8
機器學習概念 9
使用 Amazon Machine Learning 解決商務問題 9
Machine Learning 使用時機 10
建置機器學習應用程式 10
明確地描述問題 (建立問題的格式) 11
收集標記資料 11
分析您的資料 12
功能處理 12
將資料分割為訓練和評估資料 13
訓練模型 14
評估模型準確性 16
改進模型準確性 20
使用模型來進行預測 21
在新資料上重新訓練模型 21
Amazon Machine Learning 程序 22
設定 Amazon Machine Learning 24
註冊 AWS 帳戶 24
教學課程：使用 Amazon ML 預測對行銷優惠的回應 25
先決條件 25
步驟 25
步驟 1：準備資料 25

步驟 2：建立訓練資料來源 28
步驟 3：建立 ML 模型 32
步驟 4：檢閱 ML 模型的預測效能並設定分數閾值 34
步驟 5：使用 ML 模型產生預測 36
步驟 6：清除 44
建立和使用資料來源 46
了解 Amazon ML 的資料格式 46
屬性 47
輸入檔格式需求 47
使用多個檔案做為 Amazon ML 的資料輸入 48
CSV 格式的行尾字元 48
建立 Amazon ML 的資料結構描述 49
範例結構描述 49
使用 targetAttributeName 欄位 51
使用 rowID 欄位 52
使用 AttributeType 欄位 52
提供結構描述給 Amazon ML 54
分割您的資料 55
預先分割資料 55
序列分割資料 55
隨機分割資料 56
資料的深入解析 58
描述性統計資料 58
在 Amazon ML 主控台上存取 Data Insights 58
搭配使用 Amazon S3 與 Amazon ML 67
將您的資料上傳至 Amazon S3 67
許可 68
從 Amazon Redshift 中的資料建立 Amazon ML 資料來源 68
建立資料來源精靈的必要參數 69
使用 Amazon Redshift 資料建立資料來源 （主控台） 73
對 Amazon Redshift 問題進行故障診斷 76
使用 Amazon RDS 資料庫的資料來建立 Amazon ML 資料來源 81
RDS 資料庫執行個體識別符 82
MySQL 資料庫名稱 82
資料庫使用者登入資料 82
AWS Data Pipeline 安全資訊 82

Amazon RDS 安全資訊 83
MySQL SQL 查詢 83
S3 輸出位置 84
定型 ML 模型 85
ML 模型的類型 85
二元分類模型 85
多類別分類模型 86
回歸模型 86
訓練處理 86
培訓參數 87
最大模型大小 87
資料的最大通過數目 88
培訓資料的隨機播放類型 88
正規化類型和數量 89
培訓參數：類型和預設值 89
建立 ML 模型 90
先決條件 91
使用預設選項建立 ML 模型 91
使用自訂選項建立 ML 模型 92
機器學習的資料轉換 94
特徵轉型的重要性 94
使用資料配方轉換特徵 95
配方格式參考 95
群組 95
Assignments (指派) 96
輸出 96
完整配方範例 99
建議配方 100
資料轉換參考 100
N 元語法轉換 101
正交稀疏二元 (OSB) 轉換 102
小寫轉換 103
移除標點符號轉換 103
四分位數分箱轉換 103
標準化轉型 104
笛卡兒乘積轉換 104

資料重新安排 106
DataRearrangement 參數 106
評估 ML 模型 110
ML 模型深入分析 110
二元模型的深入解析 111
解譯預測 111
多類別模型深入分析 114
解譯預測 114
迴歸模型的深入解析 117
解譯預測 117
防止過度擬合 118
交叉驗證 119
調整您的模型 121
評估提醒 121
產生和解譯預測 123
建立批次預測 123
建立批次預測 (主控台) 123
建立批次預測 (API) 124
檢閱批次預測指標 125
檢閱批次預測指標 (主控台) 125
檢閱批次預測指標和詳細資訊 (API) 125
讀取批次預測輸出檔案 125
尋找批次預測資訊清單檔案 126
讀取資訊清單檔案 126
擷取批次預測輸出檔案 127
解譯二元分類 ML 模型的批次預測檔案內容 127
解譯二進位多級分類 ML 模型的批次預測檔案內容 128
解譯回歸 ML 模型的批次預測檔案內容 129
要求即時預測 129
嘗試即時預測 130
建立即時端點 132
找到即時預測端點 (主控台) 133
找到即時預測端點 (API) 134
建立即時預測要求 134
刪除即時端點 136
管理 Amazon ML 物件 138

列出物件 138
列出物件 (主控台) 138
列出物件 (API) 140
擷取物件描述 140
主控台中的詳細描述 141
API 中的詳細描述 141
更新物件 141
刪除物件 141
刪除物件 (主控台) 142
刪除物件 (API) 143
使用 Amazon CloudWatch 指標監控 Amazon ML 144
使用 記錄 Amazon ML API 呼叫 AWS CloudTrail 145
CloudTrail 中的 Amazon ML 資訊 145
範例：Amazon ML 日誌檔案項目 147
標記物件 150
標籤基本概念 150
標籤限制 151
標記 Amazon ML 物件 （主控台） 151
標記 Amazon ML 物件 (API) 153
Amazon Machine Learning 參考 154
授予 Amazon ML 許可從 Amazon S3 讀取您的資料 154
授予 Amazon ML 將預測輸出至 Amazon S3 的許可 156
控制 Amazon ML 資源的存取 - 使用 IAM 159
IAM 政策語法 159
指定 Amazon MLAmazon ML 的 IAM 政策動作 160
在 IAM 政策中指定 Amazon ML 資源ARNs 161
Amazon MLs的範例政策 162
預防跨服務混淆代理人 164
非同步操作的相依性管理 166
檢查要求狀態 167
系統限制 168
所有物件的名稱和 ID 169
物件生命週期 170
資源 171
文件歷史記錄 172

我們不再更新 Amazon Machine Learning 服務或接受新使用者。本文件可供現有使用者使用，但我們不再更新。如需詳細資訊，請參閱什麼是 Amazon Machine Learning。

本文為英文版的機器翻譯版本，如內容有任何歧義或不一致之處，概以英文版為準。

什麼是 Amazon Machine Learning？

我們不再更新 Amazon Machine Learning (Amazon ML) 服務或接受新使用者。本文件可供現有使用者使用，但我們不再更新。

AWS 現在提供強大的雲端型服務：Amazon SageMaker AI，讓所有技能水準的開發人員都可以使用機器學習技術。SageMaker AI 是一種全受管機器學習服務，可協助您建立強大的機器學習模型。透過 SageMaker AI，資料科學家和開發人員可以建置和訓練機器學習模型，然後將其直接部署到生產就緒的託管環境中。

如需詳細資訊，請參閱 SageMaker AI 文件。

主題

- Amazon Machine Learning 關鍵概念
- 存取 Amazon Machine Learning
- 區域與終端節點
- Amazon ML 定價

Amazon Machine Learning 關鍵概念

本節摘要說明下列重要概念，並更詳細地說明在 Amazon ML 中的使用方式：

- 資料來源 包含與 Amazon ML 資料輸入相關聯的中繼資料
- ML 模型使用從輸入資料擷取的模式產生預測結果
- 評估衡量 ML 模型的品質
- 批次預測「非同步」產生多個輸入資料觀察的預測結果
- 即時預測「同步」產生個別資料觀察的預測結果

資料來源

資料來源是包含輸入資料中繼資料的物件。Amazon ML 會讀取您的輸入資料、計算其屬性的描述性統計資料，並將統計資料與結構描述和其他資訊一起存放，做為資料來源物件的一部分。接下來，Amazon ML 會使用資料來源來訓練和評估 ML 模型，並產生批次預測。

> **Important**
>
> 資料來源不會存放輸入資料的副本。相反地，它會存放對於輸入資料所在的 Amazon S3 位置的參考。如果您移動或變更 Amazon S3 檔案，Amazon ML 無法存取或使用它來建立 ML 模型、產生評估或產生預測。

下表定義與資料來源相關的術語。

期間	定義
屬性	觀察內唯一具名的屬性。在表格格式資料中，例如試算表或逗號分隔值 (CSV) 檔案，欄標題代表屬性，而列則包含各個屬性的值。 同義詞：變數、變數名稱、欄位、欄
資料來源名稱	(選用) 可讓您為資料來源定義人類可讀取的名稱。這些名稱可讓您在 Amazon ML 主控台中尋找和管理資料來源。
輸入資料	資料來源參考的所有觀察的集體名稱。
位置	輸入資料的位置。目前，Amazon ML 可以使用存放在 Amazon S3 儲存貯體、Amazon Redshift 資料庫或 Amazon Relational Database Service (RDS) 中的 MySQL 資料庫內的資料。
觀察	單一輸入資料單位。例如，如果您建立 ML 模型來偵測詐騙交易，您的輸入資料會包含許多觀察，每個觀察各代表一個個別交易。 同義詞：記錄、範例、執行個體、資料列
列 ID	(選用) 旗標，若指定則可在輸入資料中識別要包含在預測輸出中的屬性。此屬性可讓您更輕鬆地將哪個預測與哪個觀察建立關聯。 同義詞：資料列識別符
結構描述	解譯輸入資料所需的資訊，包括屬性名稱及其指派資料類型，還有特殊屬性的名稱。
統計資料	輸入資料中每個屬性的摘要統計資料。這些統計資料有兩個用途：

期間	定義
	Amazon ML 主控台會以圖形顯示它們，協助您at-a-glance地了解您的資料，並識別異常或錯誤。 Amazon ML 會在訓練過程中使用這些項目來改善產生的 ML 模型品質。
Status	代表資料來源的目前狀態，例如，進行中、已完成或失敗。
目標屬性	在訓練 ML 模型的情況下，目標屬性會在包含「正確」答案的輸入資料中識別屬性的名稱。Amazon ML 使用此功能來探索輸入資料中的模式，並產生 ML 模型。在評估並產生預測的環境中，目標屬性是由受過訓練的 ML 模型預測其值的屬性。 同義詞：目標

ML 模型

ML 模型是一種數學模型，可透過尋找資料中的模式來產生預測。Amazon ML 支援三種類型的 ML 模型：二進位分類、多類別分類和迴歸。

下表定義與 ML 模型相關的術語。

期間	定義
迴歸	訓練回歸 ML 模型的目標是預測數值。
多類別	訓練多類別 ML 模型的目標是預測屬於一組有限、預先定義之允許值的值。
二進位	訓練二元 ML 模型的目標是預測只能兩種狀態其中之一 (例如 true 或 false) 的值。
模型大小	ML 模型會擷取和存放模式。ML 模型存放的模式越多，該模型就會越大。ML 模型大小是以 MB 為單位。
通過次數	當您訓練 ML 模型，您使用來自資料來源的資料。有時候在學習過程中多次使用每個資料記錄會有好處。您讓 Amazon ML 使用相同資料記錄的次數稱為傳遞次數。

期間	定義
正規化	正規化是一種機器學習技術，可用來取得更高品質的模型。Amazon ML 提供預設設定，適用於大多數情況。

評估

評估會測量您 ML 模型的品質，並判斷其是否執行效果良好。

下表定義與評估相關的術語。

期間	定義
模型深入分析	Amazon ML 為您提供指標和一些洞見，可用於評估模型的預測效能。
AUC	ROC 曲線下面積 (AUC) 會測量模型對陽性範例相較於陰性範例預測出較高分數的二元 ML 能力。
巨集平均 F1 分數	巨集平均 F1 分數是用來評估多類別 ML 模型的預測效能。
RMSE	均方根誤差 (RMSE) 是一種指標，用來評估回歸 ML 模型的預測效能。
截止值	ML 模型的運作方法是產生數值預測分數。透過套用截止值，系統可將這些分數轉換為 0 和 1 標籤。
準確性	準確性測量正確預測的百分比。
精確度	精確度顯示實際陽性執行個體 (而不是偽陽性) 在已擷取的這些執行個體 (已預測為陽性) 之間所佔的百分比。換言之，選取的項目是多少是陽性？
取回	取回會顯示真實正確占相關執行個體總數的百分比 (真實正確)。換言之，已選取多少陽性項目？

批次預測

批次預測適用於可以同時一起執行的觀察組。這很適合沒有即時需求的預測分析。

下表定義與批次預測相關的術語。

期間	定義
輸出位置	存放在 S3 儲存貯體輸出位置的批次預測結果。
資訊清單檔案	此檔案將每個輸入資料檔案，與其相關聯的批次預測結果建立關係。其存放在 S3 儲存貯體輸出位置。

即時預測

即時預測適用於具有低延遲要求的應用程式，例如互動式 Web、行動或桌面應用程式。使用低延遲即時預測 API 可以查詢任何 ML 模型的預測。

下表定義與即時預測相關的術語。

期間	定義
即時預測 API	即時預測 API 接受要求承載中的單一輸入觀察，並在回應中傳回預測。
即時預測端點	若要使用 ML 模型搭配即時預測 API，您需要建立即時預測端點。建立後，端點包含 URL，您可以用來請求即時預測。

存取 Amazon Machine Learning

您可以使用下列任一項來存取 Amazon ML：

Amazon ML 主控台

您可以登入 AWS 管理主控台，並在 https://console.aws.amazon.com/machinelearning/：// 開啟 Amazon ML 主控台，以存取 Amazon ML 主控台。

AWS CLI

如需有關如何安裝和設定 AWS CLI 的資訊，請參閱AWS Command Line Interface 《使用者指南》中的使用 AWS 命令列界面進行設定。

Amazon ML API

如需 Amazon ML API 的詳細資訊，請參閱Amazon ML API 參考。

AWS 開發套件

如需 AWS 開發套件的詳細資訊，請參閱 Amazon Web Services 適用工具。

區域與終端節點

Amazon Machine Learning (Amazon ML) 支援以下兩個區域中的即時預測端點：

區域名稱	區域	端點	通訊協定
美國東部 (維吉尼亞北部)	us-east-1	machinelearning.us-east-1.amazonaws.com	HTTPS
歐洲 (愛爾蘭)	eu-west-1	machinelearning.eu-west-1.amazonaws.com	HTTPS

您可以在任何區域中託管資料集、訓練和評估模型，及觸發預測。

建議您將所有資源保留在同一個區域。如果您的輸入資料與 Amazon ML 資源位於不同的區域，則會產生跨區域資料傳輸費用。您可以從任何區域呼叫即時預測端點，但從沒有要呼叫之端點的區域呼叫端點，會影響即時預測延遲。

Amazon ML 定價

使用 AWS 服務時，您只需支付使用量的費用。沒有最低費用，也沒有前期承諾。

Amazon Machine Learning (Amazon ML) 會針對用於運算資料統計資料和訓練和評估模型的運算時間收取每小時費率，然後您支付為應用程式產生的預測數量。對於即時預測，您還需根據模型的大小支付每小時預留容量費用。

Amazon ML 只會在 Amazon ML 主控台中預估預測的成本。

如需 Amazon ML 定價的詳細資訊，請參閱 Amazon Machine Learning 定價。

主題

- 估計批次預測成本
- 估計即時預測成本

估計批次預測成本

當您使用建立批次預測精靈從 Amazon ML 模型請求批次預測時，Amazon ML 會估計這些預測的成本。計算估計的方法會依可用的資料類型而有所不同。

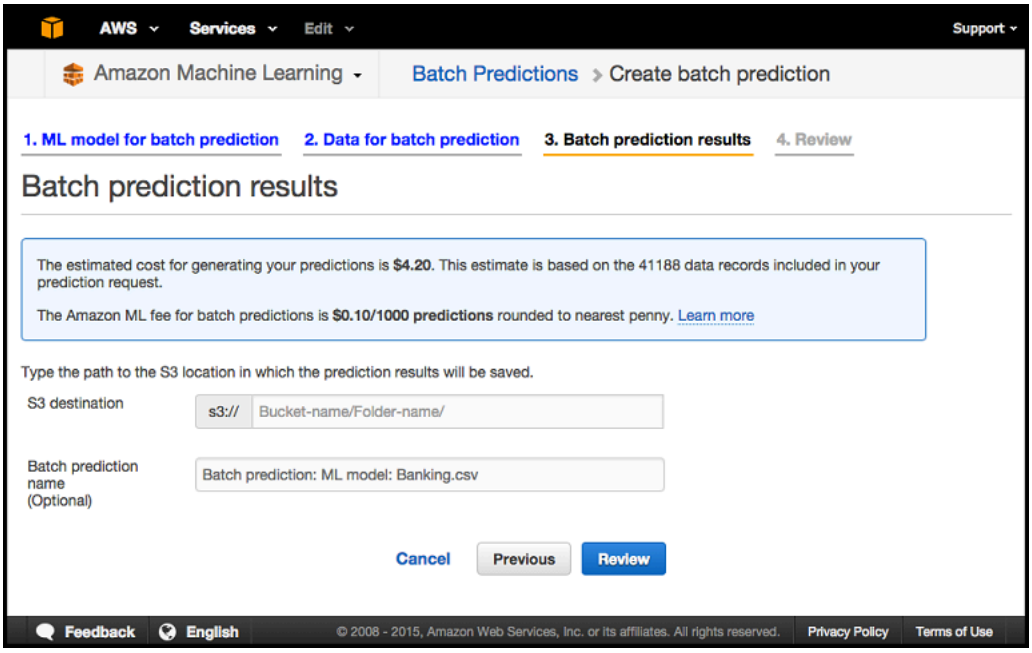
當資料統計資料可用時估計批次預測成本

當 Amazon ML 已計算用於請求預測之資料來源的摘要統計資料時，會取得最準確的成本估算。對於使用 Amazon ML 主控台建立的資料來源，一律會計算這些統計資料。API 使用者使用 CreateDataSourceFromS3、CreateDataSourceFromRedshift 或 CreateDataSourceFromRDS API 以程式設計方式建立資料來源`True`時，必須將 `ComputeStatistics`旗標設定為 。 APIs 資料來源必須在 `READY` 狀態，統計資料才可供使用。

Amazon ML 運算的其中一個統計資料是資料記錄的數量。當資料記錄的數量可用時，Amazon ML Create Batch Prediction 精靈會將資料記錄的數量乘以批次預測的費用，以估計預測的數量。

您的實際成本可能會與此預估不同，原因如下：

- 有些資料記錄可能無法處理。您不必支付來自失敗資料記錄的預測。
- 估計不會考慮既有的點數或 AWS 套用的其他調整。



只有資料大小可用時估計批次預測成本

當您請求批次預測，且請求資料來源的資料統計資料不可用時，Amazon ML 會根據下列項目估算成本：

- 在資料來源驗證期間計算和保留的總資料大小
- Amazon ML 透過讀取和剖析資料檔案的前 100 MB 來預估的平均資料記錄大小

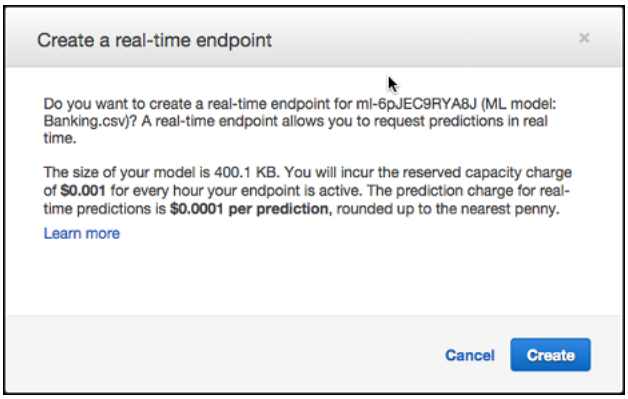
為了估計批次預測的成本，Amazon ML 會將總資料大小除以平均資料記錄大小。這種成本預測方法的精確度比資料記錄數量可用時來得低，因為您的資料檔案的第一筆記錄可能無法正確代表平均記錄大小。

當資料統計資料和資料大小都不可用時估計批次預測成本

當資料統計資料或資料大小都無法使用時，Amazon ML 無法估計批次預測的成本。當您用於請求批次預測的資料來源尚未經過 Amazon ML 驗證時，通常會發生這種情況。當您建立的資料來源是以 Amazon Redshift (Amazon Redshift) 或 Amazon Relational Database Service (Amazon RDS) 查詢為基礎，且資料傳輸尚未完成，或是資料來源建立排在帳戶中的其他操作後面時，就會發生這種情況。在此情況下，Amazon ML 主控台會通知您批次預測的費用。您可以選擇在沒有預估的情形下繼續執行批次預測請求，或取消精靈並在用於預測的資料來源位於 INPROGRESS 或 READY 狀態時再返回這裡。

估計即時預測成本

當您使用 Amazon ML 主控台建立即時預測端點時，將會顯示預估預留容量費用，這是保留端點以進行預測處理的持續費用。此費用根據模型的大小而異，如服務定價頁面所述。您也會收到標準 Amazon ML 即時預測費用的通知。



機器學習概念

機器學習 (ML) 可協助您使用歷史資料作出更好的商業決策。ML 演算法會找出資料中的模式，並利用這些發現結果建構數學模型。然後，您就可以利用這些模型對未來的資料作出預測。例如，機器學習模型的某個應用程式可能會根據客戶過去的行為，預測他們購買特定產品的可能性。

主題

- 使用 Amazon Machine Learning 解決商務問題
- Machine Learning 使用時機
- 建置機器學習應用程式
- Amazon Machine Learning 程序

使用 Amazon Machine Learning 解決商務問題

您可以使用 Amazon Machine Learning，將 Machine Learning 應用到您有現成實際答案範例的問題。例如，如果您想要使用 Amazon Machine Learning 預測電子郵件是否為垃圾郵件，您需要收集已正確標記為垃圾郵件或非垃圾郵件的電子郵件範例。然後，您可以使用 Machine Learning 從這些電子郵件範例一般化，來預測新電子郵件是否為垃圾郵件的可能性。從已標記實際答案的資料來學習的這種方法稱為受監督的 Machine Learning。

您可以針對這些特定 Machine Learning 任務使用受監督的 ML 方法：二元分類 (預測兩個可能的結果之一)、多類別分類 (預測兩個以上的結果之一) 與迴歸 (預測數值)。

二元分類問題範例：

- 客戶是否要購買這個產品？
- 這個電子郵件是否為垃圾郵件？
- 這個產品是書籍還是農畜？
- 這個評論是由客戶或機器人所撰寫？

多類別分類問題範例：

- 這個產品是書籍、電影還是衣物？
- 這個電影是浪漫喜劇片、紀錄片還是驚悚片？

- 這個客戶最感興趣的產品類別為何？

迴歸分類問題範例：

- 西雅圖明天的溫度為何？
- 這個產品會售出多少單位？
- 這個客戶過了多少天才停止使用應用程式？
- 這棟房屋的售價為何？

Machine Learning 使用時機

請務必記住，ML 不一定是每個問題類型的解決方案。有些案例可以開發健全的解決方案，不須使用 ML 技術。例如，如果您憑著簡單的規則、運算或預定步驟，不必經過任何資料導向學習的程式設計，就能判斷目標值，那就不需要 ML。

請在下列情況使用機器學習：

- 您無法為規則編寫程式碼：許多人力工作 (例如，識別電子郵件是否為垃圾郵件) 無法使用簡單 (決定性) 的規則式解決方案來妥善解決。許多因素都有可能影響答案。當規則取決於太多因素時，而且許多規則重疊或需要精細調整時，人類很快地就難以精準編碼規則。您可以使用 ML 來有效地解決這個問題。
- 無法擴展：您或許可以肉眼查出幾百封電子郵件是否為垃圾郵件。但是，若是數百萬封電子郵件，這個任務就會變得單調乏味。ML 解決方案能有效處理大規模問題。

建置機器學習應用程式

建置 ML 應用程式是一種反覆運算過程，包含一系列步驟。若要建置 ML 應用程式，一般步驟如下：

1. 根據想觀察的對象，以及您希望模型預測什麼答案，來建構核心 ML 問題。
2. 收集、清理和準備資料，讓資料適合由 ML 模型訓練演算法使用。視覺化和分析資料，執行例行性檢查以驗證資料品質並了解資料。
3. 通常，原始資料 (輸入變數) 和答案 (目標) 的呈現方式無法用來訓練高度預測模型。因此，通常您應該嘗試從原始變數建構更具預測性的輸入表示法或特徵。
4. 將產生的特徵饋送給學習演算法，來建置模型並評估對於從中提出模型建置之資料的模型品質。
5. 使用模型來針對新資料執行個體，產生目標答案預測。

明確地描述問題 (建立問題的格式)

機器學習的第一步，是決定您要預測什麼，這稱為標籤或目標答案。假設您想要製造產品，但您對於製造每個產品的決策，取決於其潛在銷售數量。在此案例中，您想要預測每個產品的購買次數 (預測銷售數量)。使用機器學習來定義此問題的方法有很多種。選擇如何定義問題，取決於您的使用案例或業務需求。

您想要預測客戶對於每個產品的購買數量 (在這種情況下，目標是數值而您需解決回歸問題)？ 還是想預測哪些產品會被購買超過 10 次 (在這種情況下，目標是二元而您需解決二元分類問題)？

請勿把問題過度複雜化，應建構可滿足您需求的最簡單解決方案。不過，也請避免遺漏資訊，尤其是歷史答案中的資訊。在這種情況下，將實際的過去銷售數量轉換成二元變數「超過 10」相較於「較少」會遺失寶貴的資訊。花點時間來決定哪個目標對您來說是最有意義的預測，可讓您省下時間以免建置無法回答您問題的模型。

收集標記資料

ML 問題從資料開始 - 最好是您已經知道資料目標答案的大量資料 (範例或觀察)。您已經知道資料目標答案的資料稱為「標記資料」。在受監督的 ML 中，演算法會教導自己從我們提供的標記範例去學習。

資料中的每個範例/觀察必須包含兩個元素：

- 目標 - 您要預測的答案。您提供標示為目標 (正確答案) 的資料供 ML 演算法從中學習。然後，使用受過訓練的 ML 模型，針對您不知道目標答案的資料來預測答案。
- 變數/特徵 - 這些是範例屬性，可用來識別模式以預測目標答案。

例如，對於電子郵件分類問題，目標是指出電子郵件是否為垃圾郵件的一個標籤。變數的範例是電子郵件的寄件者、電子郵件內文的文字、主旨行的文字、電子郵件的傳送時間，以及寄件者和接收者之間是否存在先前的通訊。

通常，資料不會以現成可用的標記形式提供。收集和準備變數與目標，通常是解決 ML 問題的最重要步驟。範例資料應該要能代表當您使用模型來進行預測時所擁有的資料。例如，如果您想要預測電子郵件是否為垃圾郵件，您必須收集陽性 (垃圾郵件電子郵件) 和陰性 (非垃圾郵件的電子郵件) 供機器學習演算法來找出模式，用以區分這兩種類型電子郵件。

擁有標記資料後，可能需要將資料轉換為演算法或軟體可接受的格式。例如，若要使用 Amazon ML，您需要將資料轉換為逗號分隔 (CSV) 格式，每個範例組成 CSV 檔案的一列、包含一個輸入變數的每一欄，以及包含目標答案的一欄。

分析您的資料

將標記資料饋送至 ML 演算法之前，最好先檢查您的資料以識別問題並獲得所用資料的深入分析。饋送的資料有多優良，模型的預測能力就有多優良。

分析資料時，應牢記以下幾點：

- 變數和目標資料摘要 - 很適合用來了解變數所具備的值，以及在資料中佔主導地位的值。可以由您想要解決之問題的主題專家來執行這些摘要。問問您自己或主題專家：資料是否符合您的期望？ 是否看起來像是您有資料收集問題？ 目標中是否某個類別比其他類別更頻繁出現？ 遺漏值或無效資料的數量是否超出您的預期？
- 變數-目標關聯 - 了解每個變數和目標類別之間的關聯非常有用，因為高度關聯表示變數和目標類別之間有關係。一般而言，您會納入具有高度關聯的變數，因為它們具有較高的預測能力 (信號)，並排除具有低度關聯的變數，因為它們可能無關。

在 Amazon ML 中，您可以透過建立資料來源和檢閱產生的資料報告來分析資料。

功能處理

透過資料摘要和視覺化效果了解您的資料之後，您可能會想進一步轉換變數，讓它們更具有意義。這就是所謂的「特徵處理」。例如，假設您有一個變數，用來擷取事件發生的日期和時間。這個日期和時間不會再次發生，因此不適合用來預測目標。不過，如果將此變數轉換成一天中小時、星期幾和月份的特徵，這些變數就能用來了解事件是否傾向於特定的小時、星期幾或月份發生。這種特徵處理可形成供學習的更一般化資料點，顯著改善預測模型。

其他常見特徵處理的範例包括：

- 使用更有意義的值來取代遺漏或無效的資料 (例如，如果您知道某個產品類型變數的遺漏值實際上代表書籍，您就可以將產品類型中的所有遺漏值替換為書籍值)。用來推算遺漏值的常用策略是將遺漏值替換為平均數或中位數。選擇替換遺漏值的策略之前，請務必先了解您的資料。
- 形成一個變數與另一個變數的笛卡兒乘積。例如，如果您有兩個變數，假設為人口密度 (urban 都市、suburban 郊區、rural 鄉村) 和州/省 (Washington 華盛頓州、Oregon 奧勒岡州、California 加州)，這兩個變數之笛卡兒乘積所產生特徵 (urban_Washington、suburban_Washington、rural_Washington、urban_Oregon、suburban_Oregon、rur 形成的特徵中，可能會有有用的資訊。
- 非線性轉換，例如將數值變數分箱轉換為類別。在許多情況下，變數特徵和目標之間的關係並非線性 (特徵值不會隨著目標單純地增加或減少)。在這種情況下，將數值特徵分箱轉換為代表不同數值特徵

範圍的類別，可能會很有用。接著，可再為每個類別特徵 (分箱) 建立模型，讓它們各自與目標產生線性關係。例如，假設您知道連續數值特徵年齡與購買書籍的可能性不存在線性關聯。您可將年齡分箱至不同的類別特徵，然後可以更精確擷取與目標的關係。數值變數的最佳分箱數取決於變數特徵以及其與目標的關係，而這最好透過實驗確定。Amazon ML 會根據建議配方中的資料統計資料，建議數值功能的最佳儲存貯體編號。如需「建議的配方」詳細資訊，請參閱《開發人員指南》。

- 領域特定的特徵 (例如，您有長度、寬度和高度的單獨變數；您可以根據這三個變數的乘積建立新的體積特徵)。
- 變數特定的特徵。有些變數類型 (例如文字特徵、擷取網頁結構或句子結構的特徵)，有一些通用的處理方法可以協助擷取結構和內容。例如，從文字「the fox jumped over the fence」(狐狸跳過籬笆) 形成的「n 元語法」可以使用「一元語法」來代表：the、fox、jumped、over、fence，也可使用「二元語法」來代表：the fox、fox jumped、jumped over、over the、the fence。

包含多個相關特徵，有助於改善預測能力。顯然，我們並非永遠能事先得知具有「信號」或預測性影響的特徵。因此，建議包含所有可能與目標標籤相關的特徵，並讓模型訓練演算法挑選具有最強關聯性的特徵。在 Amazon ML 中，可在建立模型時於配方中指定功能處理。如需可用特徵處理器的清單，請參閱《開發人員指南》。

將資料分割為訓練和評估資料

ML 的基本目標是將用於訓練模型的資料執行個體「一般化」。我們評估模型的目的，是要估計模型對於其尚未據以訓練之資料的模式一般化品質。不過，由於未來執行個體擁有不明的目標值，且我們無法現在檢查對於未來執行個體的預測準確性，因此我們需要使用一些現在已知其答案的資料，來做為未來資料的代理。使用用於訓練的相同資料來評估模型並不適合，因為這樣會獎勵能「記住」訓練資料的模型，而非從資料加以一般化的模型。

常見策略是採用所有可用的標記資料，並將其分割為訓練和評估子集，通常是 70-80% 的訓練資料、20-30% 的評估資料。ML 系統使用訓練資料來訓練模型，以查看模式並使用評估資料來評估訓練模型的預測品質。ML 系統透過使用各項指標來比較評估預測資料集的評估值與真正值 (稱為基本事實)，來評估預測效能。通常，您可以使用評估子集的「最佳」部分，來對您不知道目標答案的未來執行個體進行預測。

Amazon ML 會將透過 Amazon ML 主控台傳送用於訓練模型的資料分割為 70% 用於訓練，30% 用於評估。根據預設，Amazon ML 會依出現在訓練資料來源來源資料的順序使用前 70% 的輸入資料，其餘 30% 的資料則用於評估資料來源。Amazon ML 也可讓您隨機選取 70% 的來源資料進行訓練，而不是使用前 70%，並使用此隨機子集的補充進行評估。您可以使用 Amazon ML APIs 來指定自訂分割比率，並提供在 Amazon ML 外部分割的訓練和評估資料。Amazon ML 也提供分割資料的策略。如需分割策略的詳細資訊，請參閱分割您的資料。

訓練模型

您現在已準備好提供訓練資料給 ML 演算法 (也就是「學習演算法」)。演算法會學習到將變數對應到目標的訓練資料模式，並輸出擷取這些關係的模型。隨後可使用 ML 模型來預測您不知道目標答案的新資料。

線性模型

有大量的 ML 模型可供使用。Amazon ML 學習一種 ML 模型類型：線性模型。「線性模型」一詞表示模型指定為特徵的線性組合。根據訓練資料，學習過程會為每個特徵計算一個權重，以形成可預測或預估目標值的模型。例如，如果您的目標是顧客將購買的保險金額，您的變數是年齡和收入，簡單的線性模型將如下所示：

```
Estimated target = 0.2 + 5·age + 0.0003·income
```

學習演算法

學習演算法的任務是學習模型的權重。權重說明模型所學習的模式反映資料中實際關係的可能性。學習演算法由損失函數和最佳化技術組成。損失是當 ML 模型提供的預估目標不等於實際目標時，所產生的懲罰。損失函數將此懲罰量化為單一值。最佳化技術則尋求將損失降至最低。在 Amazon Machine Learning 中，我們使用三個損失函數，每個各對應到一種預測問題 (總共三種)。Amazon ML 中使用的最佳化技術為線上隨機梯度下降 (SGD)。SGD 會循序傳遞訓練資料，並在每個傳遞期間一次更新一個範例的特徵權重，旨在接近最小化損失的最佳權重。

Amazon ML 使用以下學習演算法：

- 對於二進位分類，Amazon ML 使用邏輯回歸 （ 邏輯損失函數 + SGD)。
- 對於多類別分類，Amazon ML 使用多節點邏輯迴歸 （ 多節點邏輯遺失 + SGD)。
- 對於迴歸，Amazon ML 使用線性迴歸 （ 平方損失函數 + SGD)。

培訓參數

Amazon ML 學習演算法接受稱為超參數或訓練參數的參數，可讓您控制所產生模型的品質。根據超參數，Amazon ML 會自動選取設定或為超參數提供靜態預設值。雖然預設超級參數設定通常可產生有用的模型，但您仍可以透過變更超級參數值來改善模型的預測效能。下列各節說明與線性模型學習演算法相關的常見超參數，例如 Amazon ML 所建立的超參數。

學習率

學習率是用於隨機漸層下降 (SGD) 演算法的常數值。學習率會影響演算法達到 (收斂至) 最佳權重的速度。SGD 演算法會針對它看到的每個資料範例，對線性模型的權重進行更新。這些更新的大小都由學習率控制。學習率太大，可能會阻止權重接近最佳解決方案。值太小又會導致演算法需要許多傳遞才能接近最佳權重。

在 Amazon ML 中，會根據您的資料自動選取學習率。

模型大小

如果您有許多輸入特徵，資料中的可能模式數會導致大型模型。大型模型擁有實際影響，例如在訓練和產生預測結果時需要更多的 RAM 來保有模型。在 Amazon ML 中，您可以使用 L1 正規化或透過指定大小上限來具體限制模型大小，以減少模型大小。請注意，如果您將模型大小降低太多，可能會降低模型的預測能力。

如需預設模型大小的詳細資訊，請參閱培訓參數：類型和預設值。如需正規化的詳細資訊，請參閱正規化。

通過次數

SGD 演算法會循序傳遞訓練資料。`Number of passes` 參數控制演算法傳遞訓練資料的通過次數。更多次傳遞可產生更符合資料的模型 (如果學習率不會太大)，但優點會隨著通過次數的增加而減少。對於較小的資料集，您可以大幅增加通過次數，如此可讓學習演算法有效地更密切符合資料。對於非常大的資料集，單次通過可能已足夠。

如需預設通過次數的詳細資訊，請參閱培訓參數：類型和預設值。

資料隨機播放

在 Amazon ML 中，您必須隨機播放資料，因為 SGD 演算法會受到訓練資料中資料列的順序影響。隨機播放訓練資料可產生較佳的 ML 模型，因為如此可協助 SGD 演算法避免產生最適合它看到的第一個資料類型、但不適合完整資料的解決方案。隨機播放會混合資料順序，讓 SGD 演算法不會連續遇到單一資料類型的太多觀察值。如果它只看到單一資料類型的許多連續權重更新，演算法可能無法更正新資料類型的模型權重，因為更新可能會太大。此外，未隨機呈現資料時，演算法將難以為所有資料類型快速找到最佳解決方案；在某些情況下，演算法可能永遠找不到最佳解決方案。隨機播放訓練資料有助於演算法更快收斂至最佳解決方案。

例如，假設您想要訓練 ML 模型預測產品類型，而您的訓練資料包含電影、玩具和電動遊戲產品類型。如果您在將資料上傳至 Amazon S3 之前，依產品類型欄排序資料，則演算法會依產品類型字母順序查

看資料。演算法會先看到所有的電影資料，因此您的 ML 模型開始學習電影的模式。接著，當您的模型遇到玩具的資料，演算法進行的每次更新都會讓演算法更符合玩具產品類型的模型，即使這些更新會降低符合電影的模式。這樣從電影突然切換到玩具類型，會產生不了解如何準確預測產品類型的模式。

如需預設隨機播放類型的詳細資訊，請參閱培訓參數：類型和預設值。

正規化

正規化會透過懲罰極端加權值，有助於避免線性模型過度擬合訓練資料範例 (也就是死記模式而非對其進行一般化)。L1 正規化可以將具有極小權重的特徵權重推送為零，以有效降低模型中使用的特徵數目。因此，L1 正規化會產生稀疏模型，並降低模型的雜訊量。L2 正規化可產生較小的整體加權值，並在輸入特徵之間有高相互關聯性時穩定加權。您可以使用 `Regularization type` 和 `Regularization amount` 參數控制套用的 L1 或 L2 正規化量。極大的正則化值可能導致所有特徵具有零加權，使得模型無法學習模式。

如需預設正則化值的詳細資訊，請參閱培訓參數：類型和預設值。

評估模型準確性

ML 模型的目標是學習將未知資料妥善一般化的模式，而非僅是死記它在訓練期間看到的資料。您擁有模型後，請務必檢查模型對於您未用於訓練模型的未知範例，是否執行效果良好。若要進行此操作，您可以使用模型來預測評估資料集 (留存資料) 的答案，然後將預測目標與實際答案 (基底事實) 進行比較。

ML 使用許多指標來測量模型的預測準確性。準確性指標的選擇，取決於 ML 任務。請務必檢閱這些指標，來判斷您的模型是否執行效果良好。

二元分類

許多二元分類演算法的實際輸出是一種預測「分數」。分數表示系統對於指定觀察屬於陽性類別的確定程度。身為此分數的取用者，若要決定觀察應該分類為陽性或陰性，您要選擇分類閾值 (截止值) 並以該值為準來比較分數，以解譯分數。對於分數高於閾值的任何觀察，將預測為陽性分類，而分數低於閾值者，則預測為陰性分類。

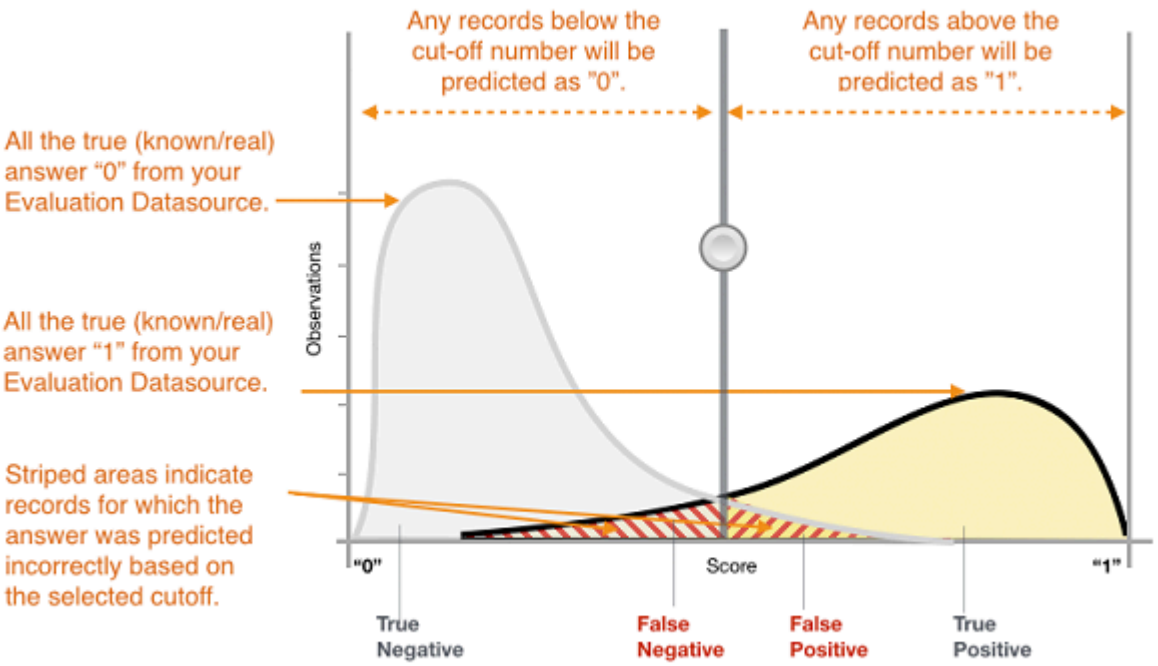


圖 1：二元分類模型的分數分佈

根據實際的已知答案和預測答案，預測結果現在分為四個群組：正確的陽性預測 (真陽性)、正確的陰性預測 (真陰性)、錯誤的陽性預測 (偽陽性) 和錯誤的陰性預測 (偽陰性)。

二元分類準確性指標會量化兩種正確預測類型和兩種錯誤類型。典型指標為「正確性」(ACC)、「精確度」、「取回」、「錯誤的正確率」、「F1 測量」。每個指標測量預測模型的不同面向。「正確性」(ACC) 會測量正確預測的分數。「精確度」會測量真實正確占這些預測為正確之範例的分數。「取回」會測量有多少真實正確被預測為正確。「F1 測量」是精確度和取回之間的調和平均數。

AUC 是不同的指標類型。它會測量模型對陽性範例相較於陰性範例預測出較高分數的能力。由於 AUC 與所選閾值無關，因此您不需要選擇閾值，就能從 AUC 指標得知模型的預測效能。

根據您的業務問題，您可能對特定指標子集執行效果良好的模型更感興趣。例如，兩個商務應用程式的 ML 模型可能會有非常不同的需求：

- 應用程式可能需要相當確定正確預測實際上為正確 (高精確度)，並能容忍將一些正確的範例分類為錯誤 (中度取回)。
- 而另一個應用程式可能只需要盡可能地正確預測正確的範例 (高度取回)，而且能夠接受將一些錯誤的範例不正確地分類為正確 (中精確度)。

在 Amazon ML 中，觀察獲得範圍 【0，1】 的預測分數。將範例分類為 0 或 1 的決策分數閾值預設為 0.5。Amazon ML 可讓您檢閱選擇不同分數閾值的影響，並可讓您挑選符合您業務需求的適當閾值。

多類別分類

與二元分類問題的程序不同，您不需要選擇分數閾值以進行預測。預測答案是具有最高預測分數的類別 (亦即，標籤)。在某些情況下，您可能會希望使用預估，只有預測回答高分。在這種情況下，您可以根據您是否接受預測答案，來選擇預測分數的閾值。

多類別中使用的典型指標與二元分類案例中使用的指標相同。在將所有其他類別分組為屬於第二個類別之後，透過將其視為二元分類問題，來計算每個類別的指標。然後，對所有類別的二元指標進行平均計算，以取得巨集平均 (平等對待每個類別) 或加權平均 (根據類別頻率加權) 指標。在 Amazon ML 中，巨集平均 F1-measure用於評估多類別分類器的預測成功。

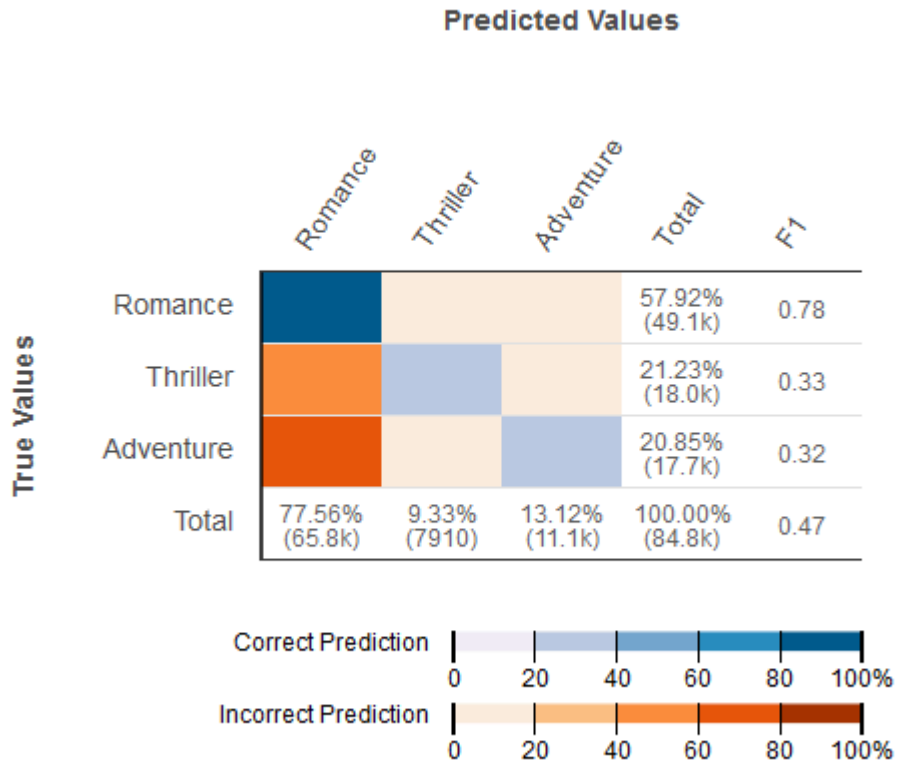


圖 2：多類別分類模型的混淆矩陣

這很適合用來檢閱多類別問題的「混淆矩陣」。每個混淆矩陣都是一個表格，顯示評估資料中的每個類別，以及正確預測和錯誤預測的數量或百分比。

迴歸

對於回歸任務，典型的準確性指標根是均方根誤差 (RMSE) 和平均絕對百分差 (MAPE)。這些指標測量預測數值目標與實際數值答案 (基本事實) 之間的差距。在 Amazon ML 中，RMSE 指標用於評估迴歸模型的預測準確性。

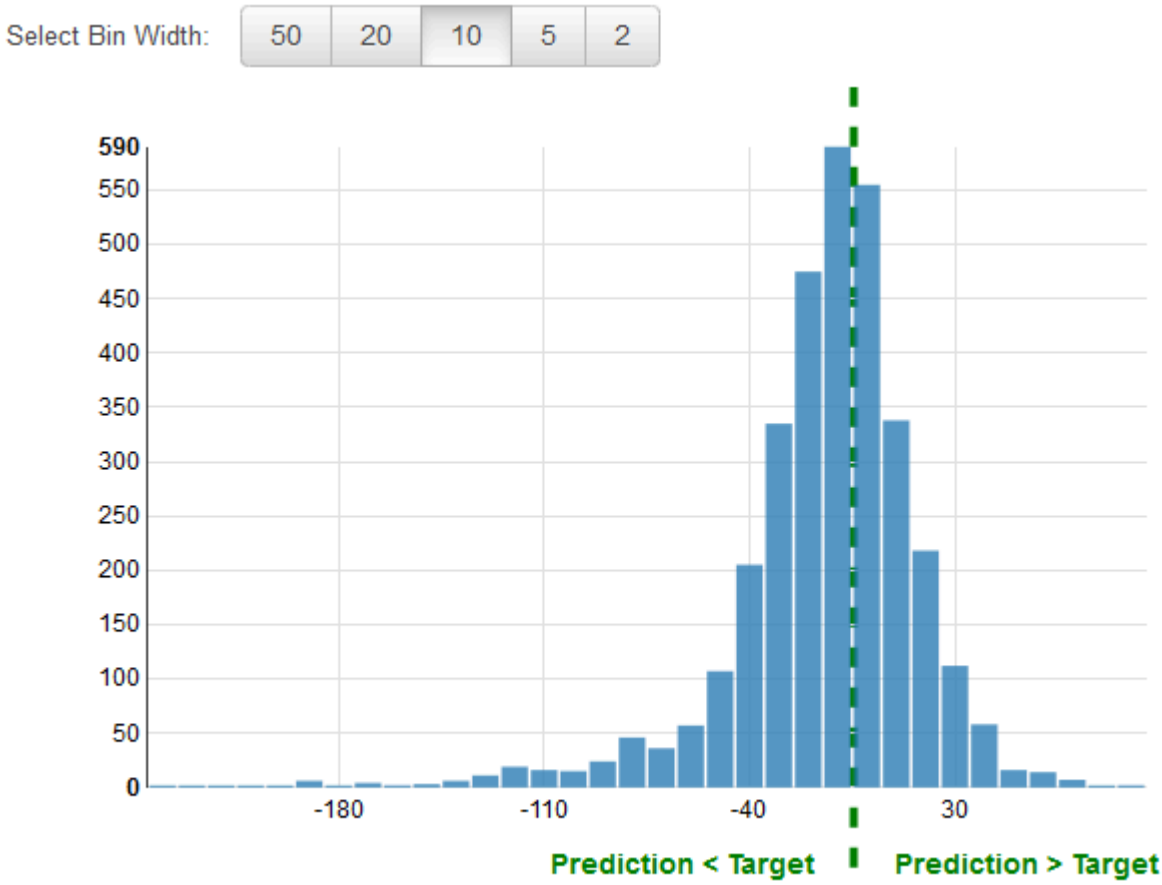


圖 3：回歸模型的殘差分佈

解決回歸問題的常見做法是檢閱「殘差」。評估資料中觀察的殘差是真正目標和預測目標之間的差距。殘差代表模型無法預測的目標部分。正殘差表示模型低估目標 (實際目標大於預測目標)。負殘差表示高估 (實際目標小於預測目標)。當評估資料的殘差長條圖呈鐘形分佈並以零為中心，表示模型以隨機的方式出錯，並未系統性過度預測或不足預測目標值的任何特定範圍。如果殘差未以零為中心呈鐘形分佈，模型的預測誤差會呈現某種結構。將更多變數新增至模型可能有助於模型擷取目前模型未擷取到的模式。

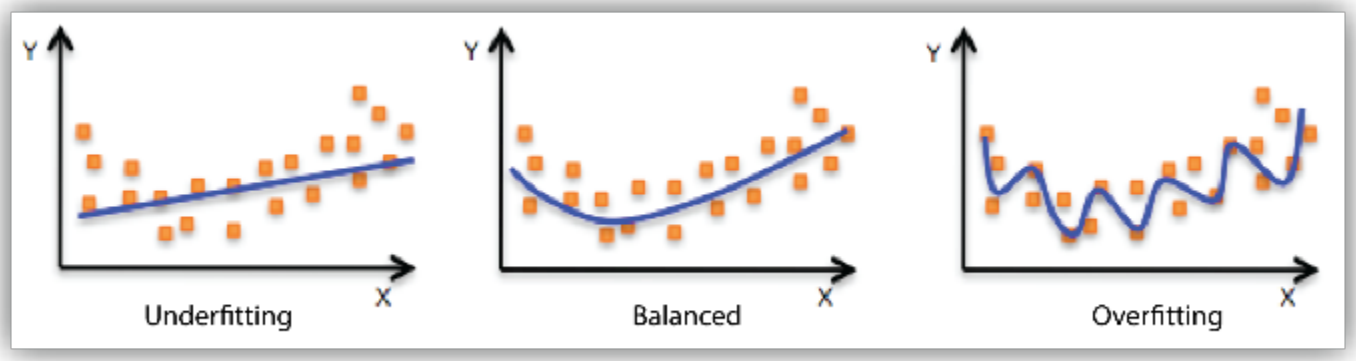
改進模型準確性

取得符合您需求的 ML 模型，通常涉及反覆運算此 ML 程序，以及嘗試一些變化。您可能無法在第一次反覆運算就取得預測性極高的模型，或者您可能想要改善模型以取得更好的預測結果。若要改進效能，您可以重複執行以下步驟：

1. 收集資料：增加訓練範例數
2. 特徵處理：增加更多變數和更佳的特徵處理
3. 模型參數調校：考慮對學習演算法使用的訓練參數採用替代值

模型擬合：低度擬合與過度擬合

了解模型擬合對於了解模型準確性不佳的根本原因相當重要。了解此項可引導您採取修正步驟。查看訓練資料和評估資的預測誤差，即可判斷預測模型是低度擬合還是過度擬合訓練資料。



當模型對訓練資料的執行效能不佳，您的模型是「低度擬合」訓練資料。這是因為模型無法擷取輸入範例 (通常稱為 X) 和目標值 (通常稱為 Y) 之間的關係。當您看到模型對訓練資料有很好的執行效果，但是對於評估資料無法執行得很好，您的模型是「過度擬合」訓練資料。這是因為模型是記憶它看到的資料，但無法一般化未知的範例。

對於訓練資料效能不佳，可能是因為模型太過簡單 (輸入特徵不夠豐富)，無法充分描述目標。您可以透過增加模型彈性來改進效能。若要提高模型彈性，請嘗試以下操作：

- 增加新的領域特定特徵和更多笛卡兒乘積特徵，以及變更特徵處理使用的類型 (例如，提高 n 元語法的大小)
- 減少使用的正規化數量

如果您的模型過度擬合訓練資料，採取降低模型彈性的措施是有道理的。若要降低模型彈性，請嘗試以下操作：

- 特徵選擇：考慮使用較少的特徵組合、減少 n 元語法的大小，以及減少數值屬性分箱數。
- 增加使用的正規化數量。

如果學習演算法沒有足夠的資料可供學習，訓練和測試資料的準確性可能不佳。您可以執行以下動作來提升效能：

- 增加訓練資料範例的數量。
- 增加現有訓練資料的通過次數。

使用模型來進行預測

現在您已擁有執行效能量好的 ML 模型，您將使用它來進行預測。在 Amazon Machine Learning 中，有兩種方式可以使用模型來進行預測：

批次預測

當您想同時為一組觀察產生預測，然後對一定比例或數量的觀察採取動作，批次預測會很有用。一般而言，您對於此類應用程式沒有低延遲要求。例如，當您想要決定鎖定哪些客戶做為產品廣告活動的對象，您需取得所有客戶的預測分數、排序模型的預測結果以識別哪些客戶最可能購買，然後鎖定最可能購買的前 5% 客戶。

線上預測

線上預測案例適用於當您想要在低延遲環境中，針對每個範例單獨產生預測結果。例如，您可以使用預測來立即決定特定交易是否可能是詐騙交易。

在新資料上重新訓練模型

模型要能夠準確預測，它據以進行預測的資料必須擁有與該模型據以訓練的資料類似的分佈。由於資料分佈預計會隨著時間而漂移，因此部署模型不是一次性的活動，而是持續不斷的程序。最好能夠持續監控傳入的資料，並在發現資料分佈已與原始的訓練資料分佈有顯著偏差時，在新的資料上重新訓練您的模型。如果監控資料以偵測資料分佈變更的開銷過高，則更簡單的策略是定期訓練模型，例如，每日、每週或每月。為了在 Amazon ML 中重新訓練模型，您需要根據您的新訓練資料建立新模型。

Amazon Machine Learning 程序

下表說明如何使用 Amazon ML 主控台來執行本文件中概述的 ML 程序。

ML 程序	Amazon ML 任務
分析您的資料	若要在 Amazon ML 中分析您的資料，請建立資料來源並檢閱資料洞見頁面。
將資料分割為定型與評估資料來源	Amazon ML 可以將資料來源分割為使用 70% 的資料進行模型訓練，並將 30% 的資料用於評估模型的預測效能。 當您使用建立 ML 模型精靈搭配預設設定時，Amazon ML 會為您分割資料。 如果您使用建立 ML 模型精靈搭配自訂設定，並選擇評估 ML 模型，您會看到允許 Amazon ML 為您分割資料並在 30% 的資料上執行評估的選項。
隨機播放您的定型資料	當您使用建立 ML 模型精靈搭配預設設定時，Amazon ML 會為您隨機切換資料。您也可以在將資料匯入 Amazon ML 之前隨機播放資料。
處理特徵	以理想格式將定型資料放在一起以供學習與一般化的程序，稱為特徵轉換。當您使用具有預設設定的建立 ML 模型精靈時，Amazon ML 建議為您的資料進行功能處理設定。 若要指定特徵處理設定，請使用 Create ML Model (建立 ML 模型) 精靈的 Custom (自訂) 選項，並提供特徵處理配方。
定型模型	當您使用建立 ML 模型精靈在 Amazon ML 中建立模型時，Amazon ML 會訓練您的模型。
選取模型參數	在 Amazon ML 中，您可以調校影響模型預測效能的四個參數：模型大小、通過次數、隨機分組類型和正規化。當您使用 Create ML Model (建立 ML 模型) 精靈建立 ML 模型時，可透過選擇 Custom (自訂) 選項設定這些參數。
評估模型效能	使用 Create Evaluation (建立評估) 精靈來評估模型的預測效能。

ML 程序	Amazon ML 任務
特徵選取	Amazon ML 學習演算法可能會捨棄對學習程序貢獻不大的功能。若要指出您想要捨棄這些特徵，請在建立 ML 模型時選擇 `L1 regularization` 參數。
設定分數閾值以取得預測準確度	檢閱評估報告中不同分數閾值的模型預測效能，然後根據您的商務應用程式設定分數閾值。分數閾值決定模型如何定義預測相符。調整此數值可控制錯誤肯定與錯誤否定。
使用模型	使用您的模型，透過 Create Batch Prediction (建立批次預測) 精靈來預測觀察批次。 或者，讓 ML 模型使用 `Predict` API 處理即時預測，以視需要來預測個別觀察。

設定 Amazon Machine Learning

註冊 AWS 帳戶

若要開始使用 AWS，您需要 AWS 帳戶。如需建立 的相關資訊 AWS 帳戶，請參閱《 AWS 帳戶管理參考指南》中的 入門 AWS 帳戶。

教學課程：使用 Amazon ML 預測對行銷優惠的回應

透過 Amazon Machine Learning (Amazon ML)，您可以建置和訓練預測模型，並將應用程式託管在可擴展的雲端解決方案中。在本教學課程中，我們會示範如何使用 Amazon ML 主控台來建立資料來源、建置機器學習 (ML) 模型，以及使用該模型來產生您可以在應用程式中使用的預測。

我們的練習範例會說明如何找出目標行銷活動的潛在客戶，但您可以套用該相同原則，來建立及使用各種 ML 模型。為了完成範例練習，您將使用 University of California at Irvine (UCI) Machine Learning Repository 公開提供的銀行和行銷資料集。這些資料集包含客戶的一般資訊，以及其之前對行銷聯絡人的回應。您將使用此資料找出哪些客戶最可能訂閱您的新產品：銀行定期存款，也稱為定期存單 (CD)。

> **Warning**
>
> AWS 免費方案不包含此教學課程。如需 Amazon ML 定價的詳細資訊，請參閱 Amazon Machine Learning 定價。

先決條件

若要執行教學課程，您需要具備 AWS 帳戶。如果您還沒有 AWS 帳戶，請參閱設定 Amazon Machine Learning。

步驟

- 步驟 1：準備資料
- 步驟 2：建立訓練資料來源
- 步驟 3：建立 ML 模型
- 步驟 4：檢閱 ML 模型的預測效能並設定分數閾值
- 步驟 5：使用 ML 模型產生預測
- 步驟 6：清除

步驟 1：準備資料

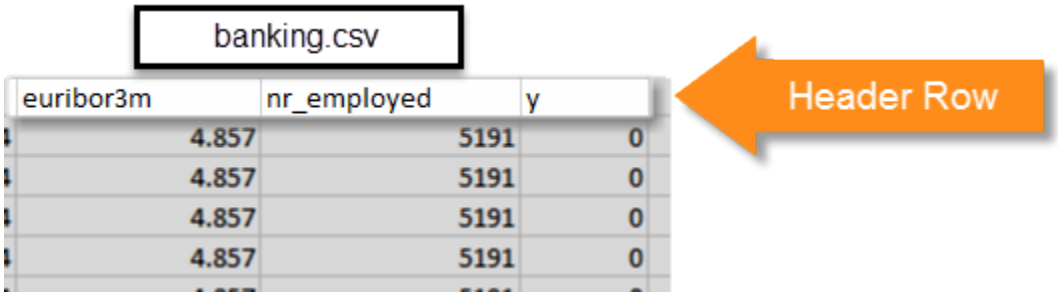
在機器學習中，您通常會取得資料，並先確保它的格式良好，再啟動培訓程序。基於本教學的目的，我們從 UCI Machine Learning Repository 取得範例資料集，格式化為符合 Amazon ML 準則，並供您下

載。從 Amazon Simple Storage Service (Amazon S3) 儲存位置下載資料集，並遵循本主題中的程序將其上傳至您自己的 S3 儲存貯體。

如需 Amazon ML 格式要求，請參閱 了解 Amazon ML 的資料格式。

下載資料集

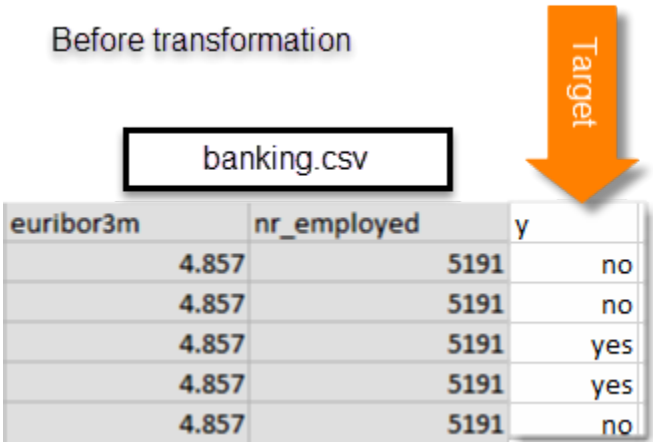
1. 按一下 banking.zip，下載包含已購買類似銀行定期存款產品之客戶歷史資料的檔案。解壓縮資料夾，並將 banking.csv 檔案儲存到您的電腦。
2. 按一下 banking-batch.zip，下載您將用來預測潛在客戶是否會回應您的報價的檔案。解壓縮資料夾，並將 banking-batch.csv 檔案儲存到您的電腦。
3. 打開 `banking.csv`。您將會看到資料的資料列和資料行。「標題列」包含每個資料行的屬性名稱。「屬性」(Attribute) 是唯一具名屬性 (Property)，說明每個客戶的特定特性，例如，nr_employed 指出客戶的雇用狀態。每個資料列都代表單一客戶的觀察集合。



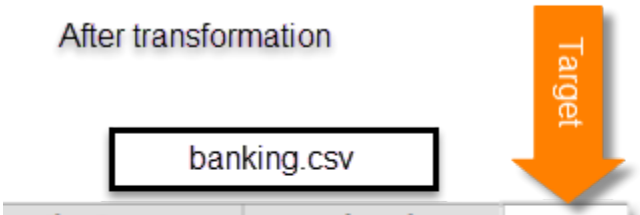
您希望 ML 模型回答「這位客戶將訂閱我的新產品嗎？」問題。在 `banking.csv` 資料集內，這個問題的答案是 y 屬性，其包含值 1 (表示「是」) 或 0 (表示「否」)。您希望 Amazon ML 了解如何預測的屬性稱為目標屬性。

> **Note**
>
> 屬性 y 是二元屬性。它可以只包含兩個值的其中一個值，在這種情況下為 0 或 1。在原始 UCI 資料集內，y 屬性為「是」或「否」。我們已為您編輯妥原始資料集。屬性 y 表示「是」的所有值現在是 1，而表示「否」的所有值現在是 0。如果您使用自己的資料，則可以使用其他二元屬性值。如需有效值的詳細資訊，請參閱使用 AttributeType 欄位。

以下範例顯示將 y 屬性的值變更為二元屬性 0 和 1 前後的資料。



euribor3m	nr_employed	y
4.857	5191	no
4.857	5191	no
4.857	5191	yes
4.857	5191	yes
4.857	5191	no



euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	1
4.857	5191	1
4.857	5191	0

`banking-batch.csv` 檔案未包含 y 屬性。在您建立 ML 模型之後，將會使用此模型來預測該檔案中每筆記錄的 y。

接著，將 `banking.csv` 和 `banking-batch.csv` 檔案上傳至 Amazon S3。

將檔案上傳至 Amazon S3 位置

1. 登入 AWS 管理主控台 ，並在 https://console.aws.amazon.com/s3/ : // 開啟 Amazon S3 主控台。
2. 在 All Buckets (所有儲存貯體) 清單中，建立儲存貯體或選擇您要上傳檔案的位置。
3. 在導覽列中，選擇 Upload (上傳)。
4. 選擇 Add Files (新增檔案)。
5. 在對話方塊中，導覽至您的桌面並選擇 `banking.csv` 和 `banking-batch.csv`，然後選擇 Open (開啟)。

您現在已準備好可建立培訓資料來源。

步驟 2：建立訓練資料來源

將`banking.csv`資料集上傳到 Amazon Simple Storage Service (Amazon S3) 位置後，您可以使用它來建立訓練資料來源。資料來源是 Amazon Machine Learning (Amazon ML) 物件，其中包含輸入資料的位置和輸入資料的重要中繼資料。Amazon ML 會將資料來源用於 ML 模型訓練和評估等操作。

若要建立資料來源，請提供下列項目：

- 資料的 Amazon S3 位置和存取資料的許可
- 結構描述，包含資料中的屬性名稱及各屬性的類型 (數字、文字、分類或二元)
- 屬性的名稱，其中包含您希望 Amazon ML 學習預測的答案，即目標屬性

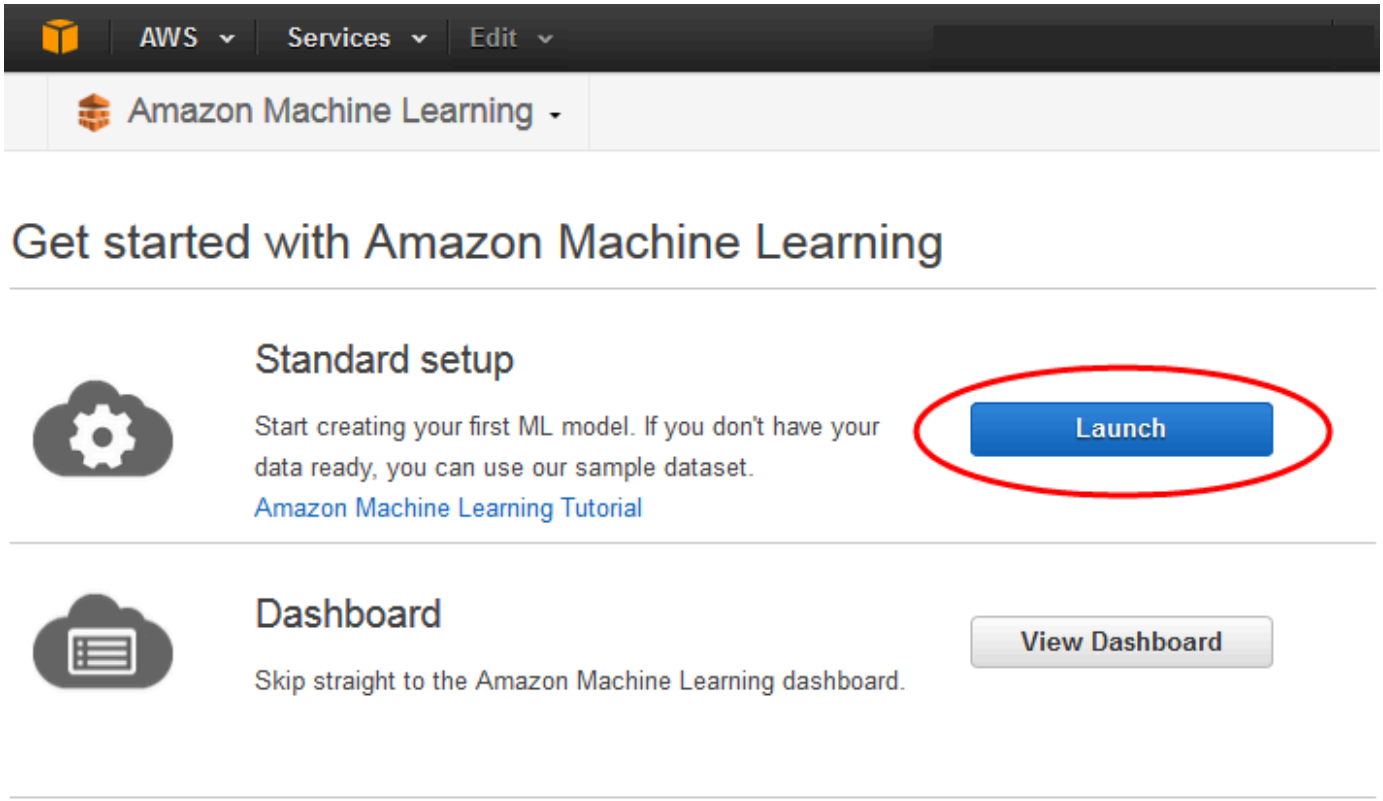
> **Note**
>
> 資料來源並不會實際地存放您的資料，只是參考該資料而已。避免移動或變更存放在 Amazon S3 中的檔案。如果您確實移動或變更它們，Amazon ML 無法存取它們來建立 ML 模型、產生評估或產生預測。

建立訓練資料來源

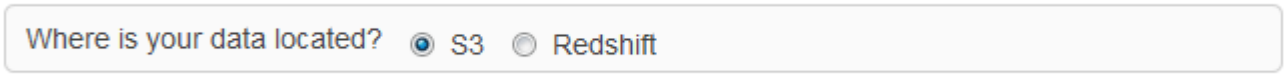
1. 開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 選擇開始使用。

   > **Note**
   >
   > 本教學假設這是您第一次使用 Amazon ML。如果您之前已使用 Amazon ML，則可以使用 Amazon ML 儀表板上的建立新的...下拉式清單來建立新的資料來源。

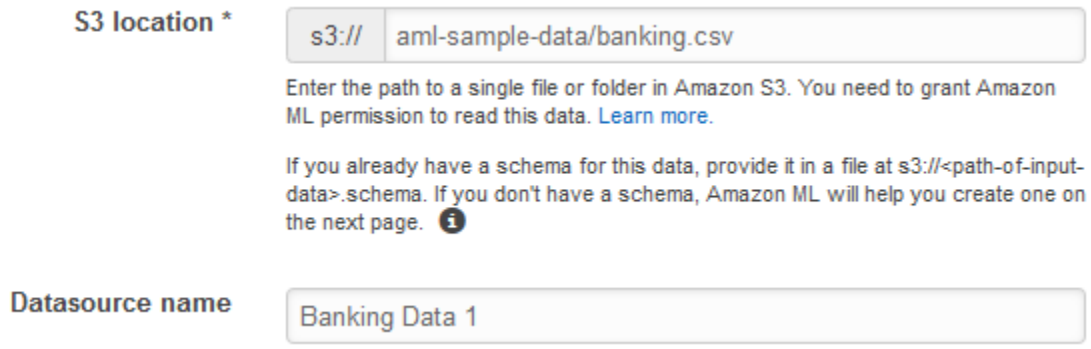
3. 在開始使用 Amazon Machine Learning 頁面上，選擇啟動。



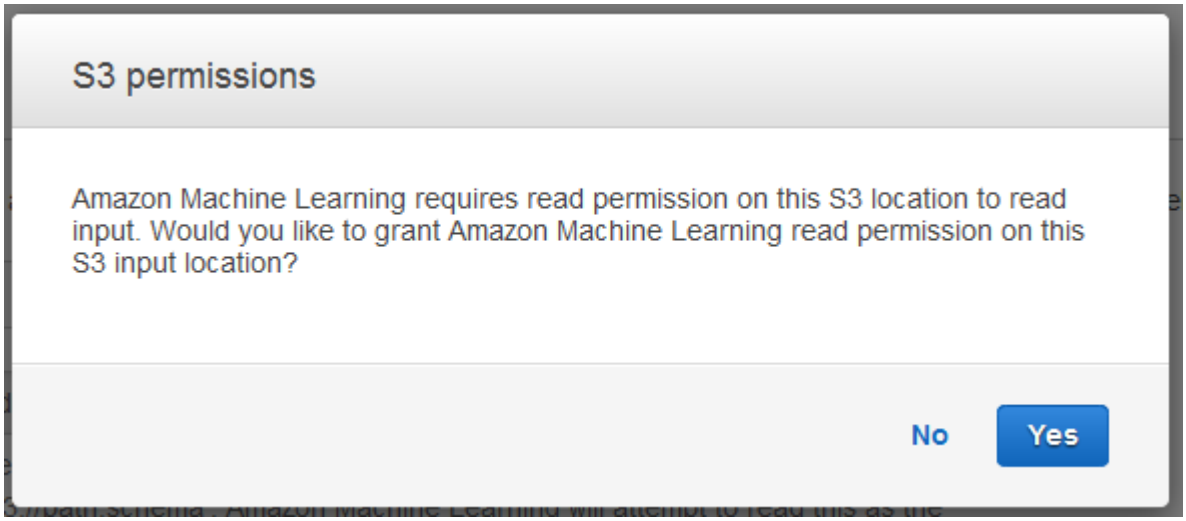
4. 在 Input Data (輸入資料) 頁面上，確定 Where is your data located? (您的資料在哪個位置？) 已選取 S3。



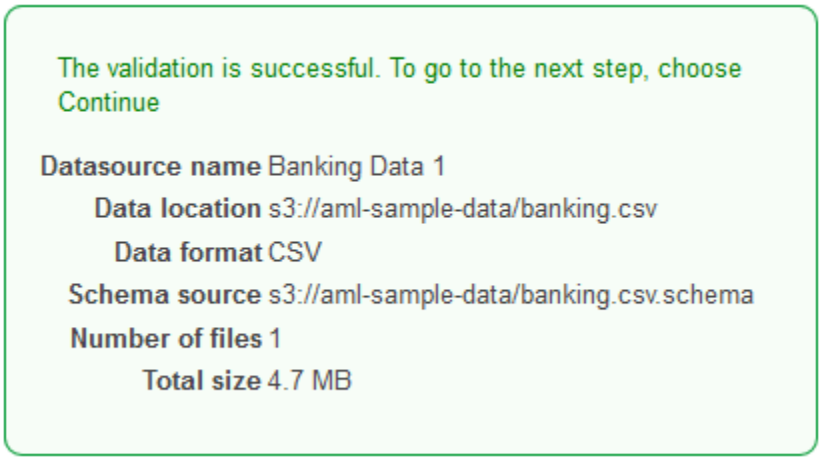
5. 為 S3 Location (S3 位置) 輸入「步驟 1：準備資料」中 `banking.csv` 檔案的完整位置。例如：*`your-bucket`***`/banking.csv`**。Amazon ML 會在儲存貯體名稱前面加上 s3：//。
6. 針對 Datasource name (資料來源名稱) 輸入 **`Banking Data 1`**。



7. 選擇 Verify (驗證)。
8. 在 S3 permissions (S3 許可) 對話方塊中，選擇 Yes (是)。



9. 如果 Amazon ML 可以存取和讀取 S3 位置的資料檔案，您會看到類似以下的頁面。檢閱屬性，然後選擇 Continue (繼續)。



接著，您要建立結構描述。結構描述是 Amazon ML 解譯 ML 模型的輸入資料所需的資訊，包括屬性名稱及其指派的資料類型，以及特殊屬性的名稱。提供 Amazon ML 結構描述的方式有兩種：

- 當您上傳 Amazon S3 資料時，請提供個別的結構描述檔案。
- 允許 Amazon ML 推斷屬性類型並為您建立結構描述。

在本教學課程中，我們會要求 Amazon ML 推斷結構描述。

如需建立獨立結構描述檔案的相關資訊，請參閱建立 Amazon ML 的資料結構描述。

允許 Amazon ML 推斷結構描述

1. 在結構描述頁面上，Amazon ML 會顯示它推斷的結構描述。檢閱 Amazon ML 針對屬性推斷的資料類型。務必為屬性指派正確的資料類型，以協助 Amazon ML 正確擷取資料，並對屬性啟用正確的特徵處理。

   - 若屬性只有兩種可能狀態 (例如，是或否)，應標示為 Binary (二元)。
   - 若屬性為用來表示分類的數字或字串，應標示為 Categorical (分類)。
   - 若屬性為順序有意義的數字量，應標示為 Numeric (數值)。
   - 若屬性為您想要視為以空格分隔之單詞的字串，應標示為 Text (文字)。

Name	Data Type	Sample Field Value 1
age	Numeric	56
campaign	Numeric	1
cons_conf_idx	Numeric	-36.4
cons_price_idx	Numeric	93.994
contact	Categorical	telephone
day_of_week	Categorical	mon
default	Categorical	no
duration	Numeric	261
education	Categorical	basic.4y
emp_var_rate	Numeric	1.1

2. 在本教學課程中，Amazon ML 已正確識別所有屬性的資料類型，因此請選擇繼續。

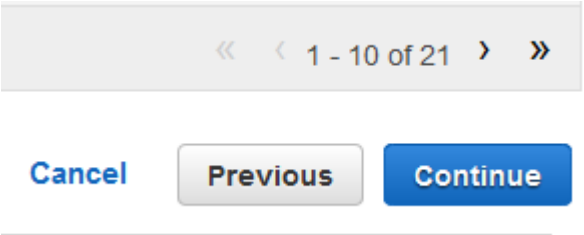
接著選取目標屬性。

請記住，目標是 ML 模型必須學習預測的屬性。屬性 y 會指出某人過去是否訂閱過活動：1 (是) 或 0 (否)。

> **Note**
>
> 只有當您要使用資料來源來訓練和評估 ML 模型時，才選擇目標屬性。

選取 y 做為目標屬性

1. 在表格右下方，選擇單箭頭前往表格的最後一頁，該頁會顯示名為 y 的屬性。



2. 在 Target (目標) 欄，選取 y。

   Amazon ML 會確認已選取 y 做為您的目標。
3. 選擇繼續。
4. 在 Row ID (列 ID) 頁面上，確定 Does your data contain an identifier? (您的資料包含識別符嗎？) 已選取預設值 No (否)。
5. 選擇 Review (檢閱)，然後選擇 Continue (繼續)。

既然您已具有訓練資料來源，就可以建立模型。

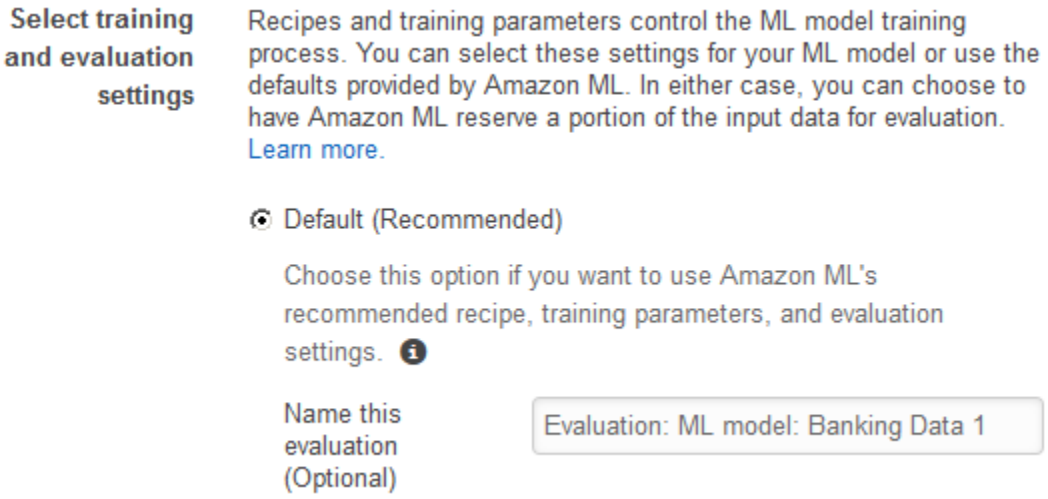
步驟 3：建立 ML 模型

建立好訓練資料來源後，可以使用它來建立 ML 模型、訓練模型，然後評估結果。ML 模型是 Amazon ML 在訓練期間在您的資料中找到的模式集合。您可以使用模型來建立預測。

建立 ML 模型

1. 由於入門精靈同時建立訓練資料來源和模型，Amazon Machine Learning (Amazon ML) 會自動使用您剛建立的訓練資料來源，並直接帶您前往 ML 模型設定頁面。在 ML model settings (ML 模型設定) 頁面上，確定 ML model name (ML 模型名稱) 已顯示預設選項 **`ML model: Banking Data 1`**。

   使用好記的名稱 (例如預設值)，可協助您輕鬆地識別和管理 ML 模型。
2. 對於 Training and evaluation settings (訓練與評估設定)，確定已選取 Default (預設)。



3. 對於 Name this evaluation (命名此評估)，接受預設值 **`Evaluation: ML model: Banking Data 1`**。
4. 選擇 Review (檢閱)、檢閱您的設定，然後選擇 Finish (完成)。

   選擇完成之後，Amazon ML 會將您的模型新增至處理佇列。Amazon ML 建立模型時，會套用預設值並執行下列動作：

   - 將訓練資料來源分割為兩個部分，其一包含 70% 的資料，另一部分包含其餘的 30%
   - 使用包含 70% 輸入資料的部分來訓練 ML 模型
   - 使用其餘 30% 輸入資料的部分來評估模型

   當您的模型在佇列中時，Amazon ML 會將狀態報告為待定。Amazon ML 建立模型時，會將狀態報告為進行中。待已完成所有動作後，則回報狀態為 Completed (已完成)。請等待評估完成，然後再繼續。

現在您已準備好要開始檢閱模型的效能和設定分界分數。

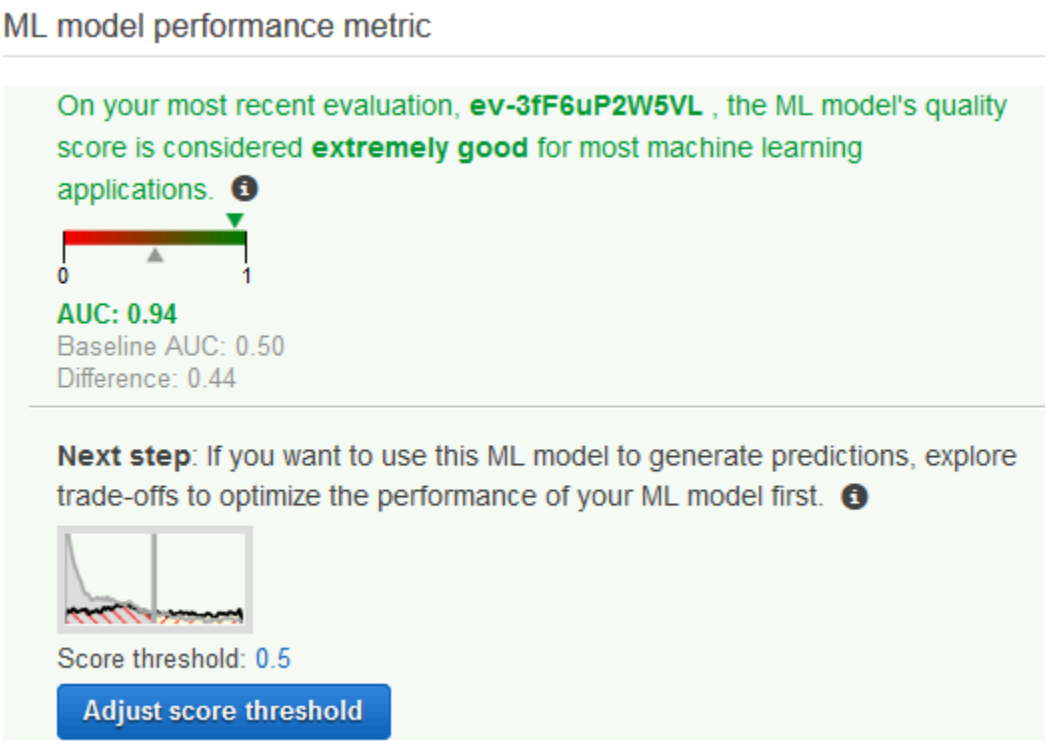
如需訓練和評估模型的詳細資訊，請參閱定型 ML 模型和evaluate an ML model。

步驟 4：檢閱 ML 模型的預測效能並設定分數閾值

現在您已建立 ML 模型，且 Amazon Machine Learning (Amazon ML) 已對其進行評估，讓我們來看看它是否足夠使用。在評估期間，Amazon ML 計算了一個產業標準品質指標，稱為曲線下面積 (AUC) 指標，可表達 ML 模型的效能品質。Amazon ML 也會解譯 AUC 指標，以告訴您 ML 模型的品質是否適用於大多數機器學習應用程式。(請至衡量 ML 模型準確性進一步了解 AUC)。讓我們檢閱 AUC 指標，然後調整分數閾值或分界值以最佳化模型的預測效能。

檢閱 ML 模型的 AUC 指標

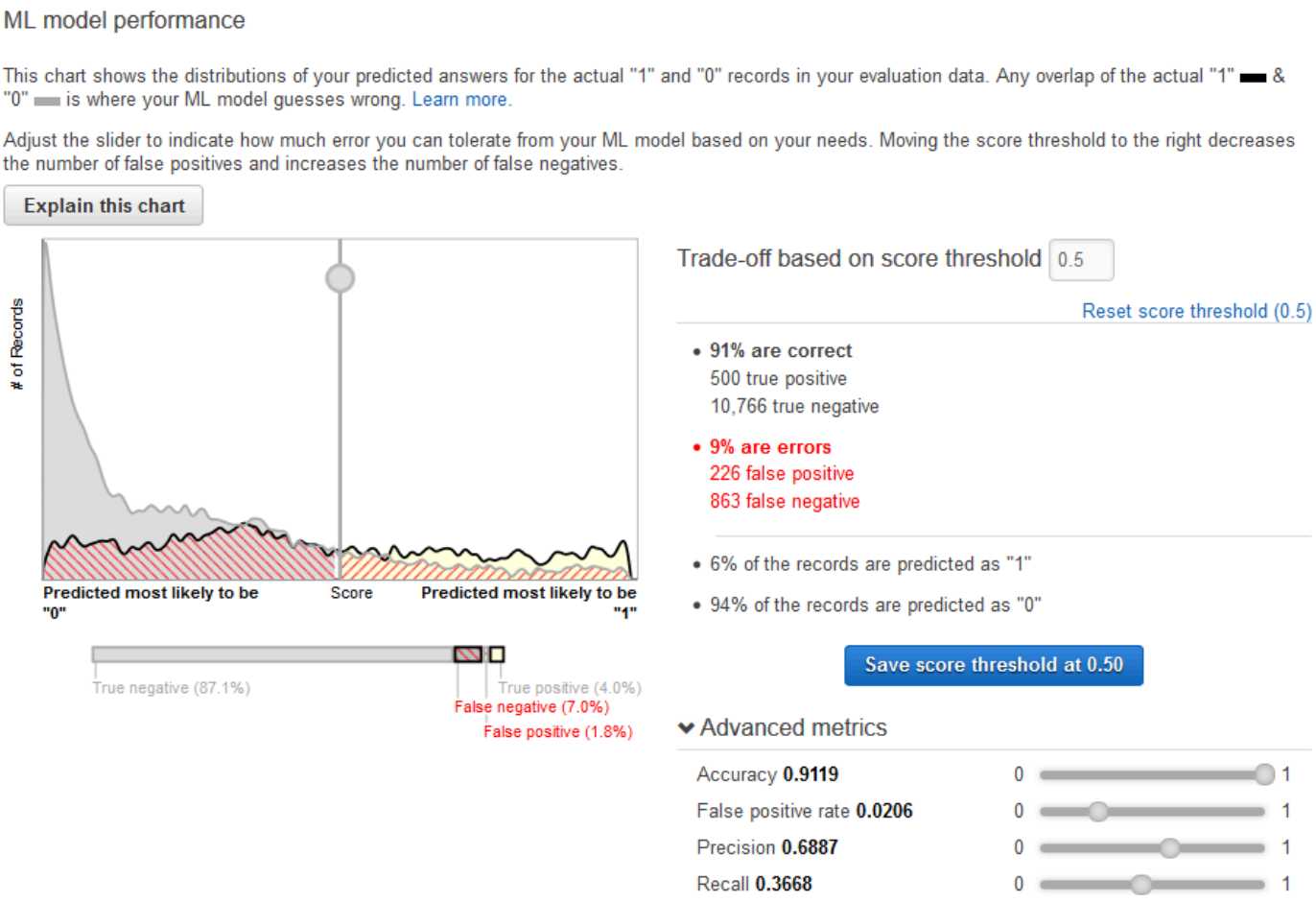
1. 在 ML model summary (ML 模型摘要) 頁面的 ML model report (ML 模型報告) 導覽窗格中，選擇 Evaluations (評估)、選擇 Evaluation: ML model: Banking model 1 (評估：ML 模型：Banking model 1)，然後選擇 Summary (摘要)。
2. 在 Evaluation summary (評估摘要) 頁面上，檢閱評估摘要，包括模型的 AUC 效能指標。



ML 模型會為預測資料來源中的每個記錄各產生數值預測分數，然後套用閾值以將這些分數轉換為 0 (否) 或 1 (是) 的二元標籤。透過變更 score threshold (分數閾值)，您可以調整 ML 模型指派這些標籤的方式。現在，設定分數閾值。

為 ML 模型設定分數閾值

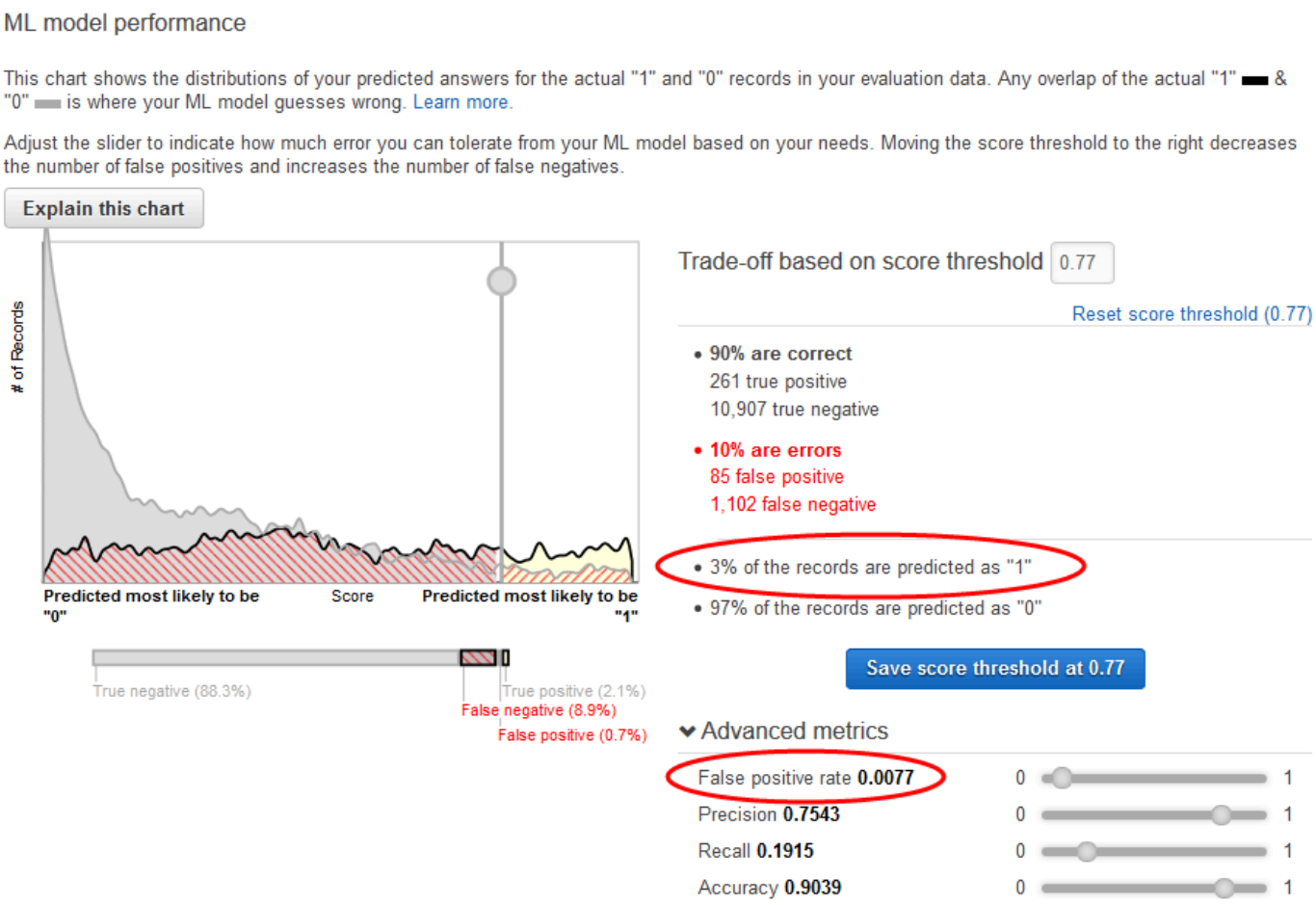
1. 在 Evaluation Summary (評估摘要) 頁面上，選擇 Adjust Score Threshold (調整分數閾值)。



   您可以透過調整分數閾值來微調 ML 模型的效能指標。調整這個值會改變模型在預測中必須具備的可信度等級，用以將某個預測視為陽性。也會改變您願意在預測中容忍多少偽陽性和偽陰性結果。

   您可以透過增加分數閾值來控制模型會將哪些考慮為陽性預測的分界值，直到模型只將具有最高真陽性可能性的預測看作是陽性的。您也可以降低分數閾值，直到您不再有任何偽陽性。選擇您的分界值，以反映您的業務需求。在本教學課程中，每個偽陽性都會花費行銷活動資金，所以我們希望真陽性的比例高於偽陽性。

2. 舉例來說，您想要鎖定會訂閱產品的前 3% 客戶。滑動垂直選取器，將分數閾值設定至對應到 3% of the records are predicted as "1" (3% 的記錄會預測為「1」) 的值。



請注意，此分數閾值對 ML 模型效能的影響：偽陽性率為 0.007。假設此偽陽性率是可接受的值。

3. 選擇 Save score threshold at 0.77 (將分數閾值儲存在 0.77)。

每當您使用此 ML 模型來進行預測，它會將分數超過 0.77 的記錄預測為「1」，其餘記錄預測為「0」。

若要進一步了解分數閾值，請參閱二元分類。

現在您已準備好要開始使用您的模型建立預測。

步驟 5：使用 ML 模型產生預測

Amazon Machine Learning (Amazon ML) 可以產生兩種類型的預測：批次和即時。

即時預測是 Amazon ML 隨需產生的單一觀察的預測。即時預測適用於行動應用程式、網站和其他需要以互動方式使用結果的應用程式。

批次預測是一組觀察的一組預測。Amazon ML 會一起處理批次預測中的記錄，因此處理可能需要一些時間。批次預測適用於需要一組觀察的預測或不以互動方式使用結果的預測之應用程式。

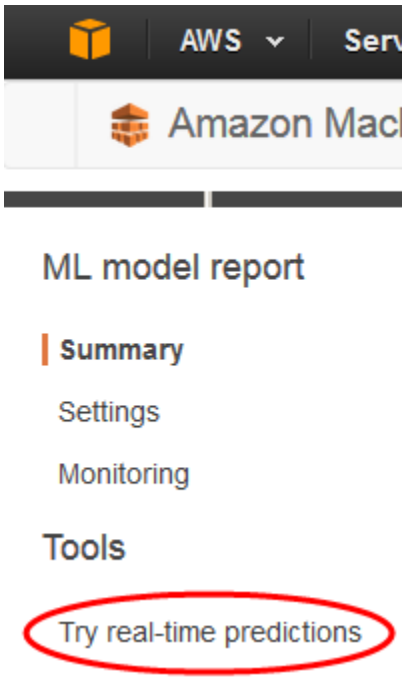
在本教學課程中，您會產生即時預測，預測某位潛在客戶是否會訂閱新產品。您也會產生大批次潛在客戶的預測。在批次預測方面，您將會使用您在「banking-batch.csv」中上傳的 步驟 1：準備資料 檔案。

那麼就先從即時預測開始。

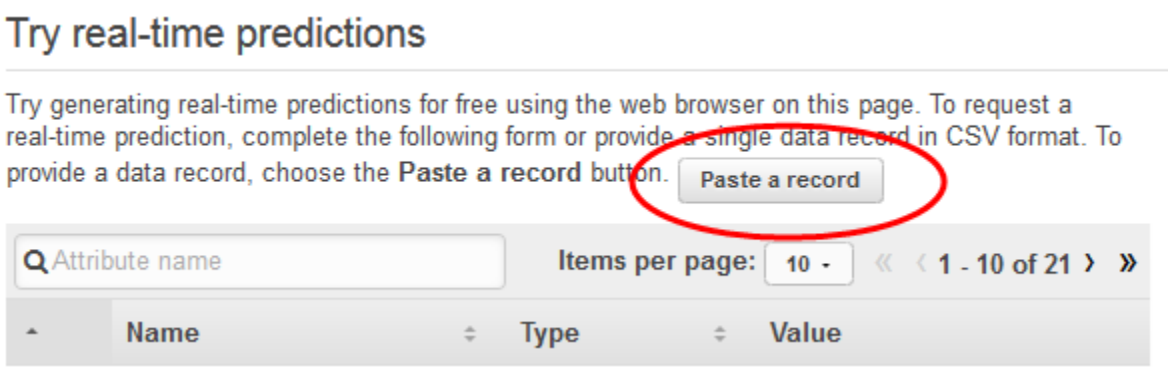
> **Note**
>
> 對於需要即時預測的應用程式，您必須建立 ML 模型的即時端點。當即時端點可使用時，會產生費用。在您承諾使用即時預測並開始產生與其相關的費用前，可以先在 Web 瀏覽器中試用即時預測功能，而不用建立即時端點。這就是在本教學課程中將要操作的內容。

試用即時預測

1. 在 ML model report (ML 模型報告) 導覽窗格中，選擇 Try real-time predictions (試用即時預測)。



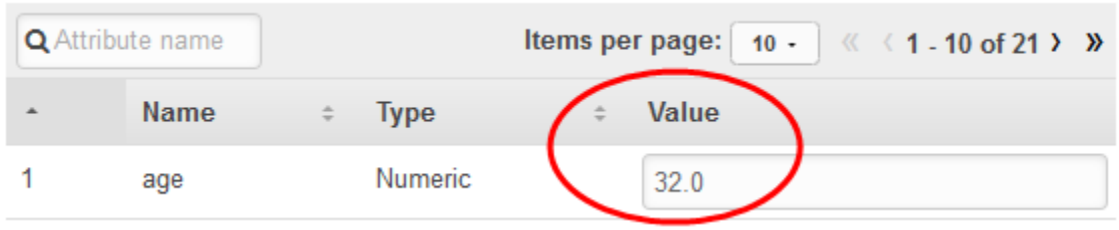
2. 選擇 Paste a record (貼上記錄)。



3. 在 Paste a record (貼上記錄) 對話方塊中，貼上以下觀察：

```
32,services,divorced,basic.9y,no,unknown,yes,cellular,dec,mon,110,1,11,0,nonexistent,-1.8,9
```

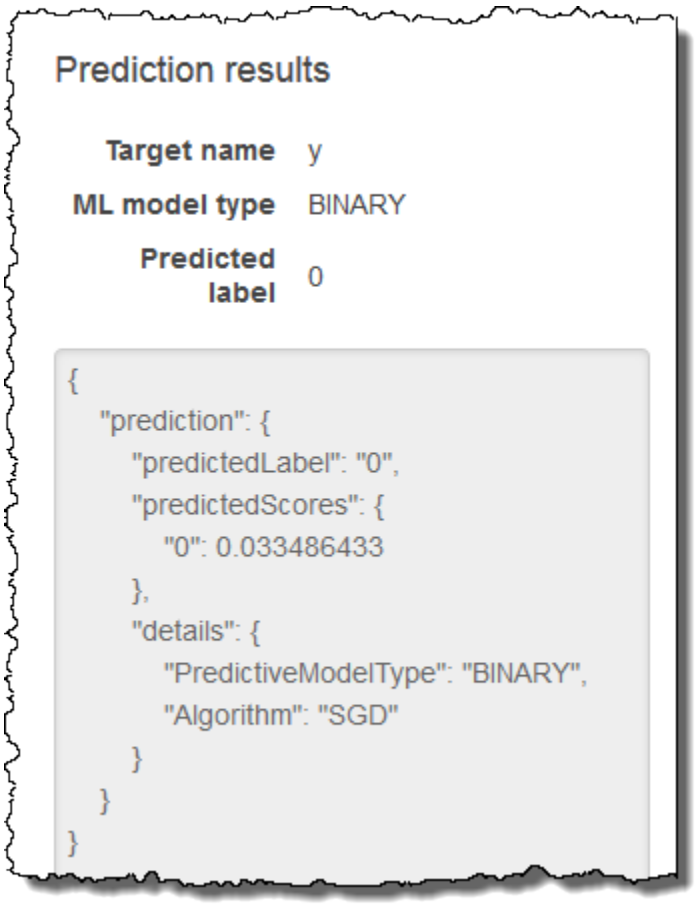
4. 在貼上記錄對話方塊中，選擇提交以確認您想要為此觀察產生預測。Amazon ML 會以即時預測形式填入值。



> **Note**
>
> 您也可以輸入個別的值，藉此填入 Value (值) 欄位。無論您選擇何種方法，都應提供不是用於訓練模型的觀察。

5. 在頁面底部，選擇 Create prediction (建立預測)。

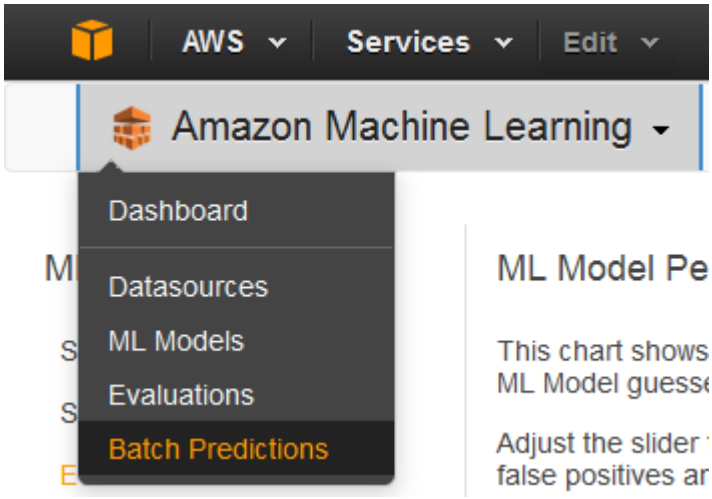
   預測會顯示在右側的 Prediction results (預測結果) 窗格中。此預測的 Predicted label (預測標籤) 為 `0`，表示此潛在客戶不可能回應活動。Predicted label (預測標籤) 為 `1` 則表示客戶可能回應活動。



現在來建立批次預測。您將向 Amazon ML 提供您正在使用的 ML 模型名稱、您要產生預測之輸入資料的 Amazon Simple Storage Service (Amazon S3) 位置 (Amazon ML 會從此資料建立批次預測資料來源），以及用於儲存結果的 Amazon S3 位置。

建立批次預測

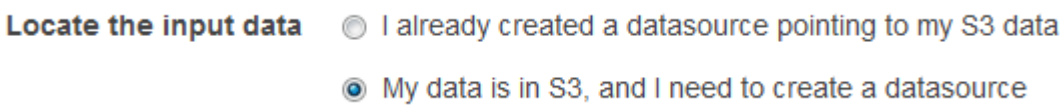
1. 選擇 Amazon Machine Learning，然後選擇 Batch Predictions (批次預測)。



2. 選擇 Create new batch prediction (建立新的批次預測)。
3. 在 ML model for batch predictions (批次預測的 ML 模型) 頁面上，選擇 ML model: Banking Data 1 (ML 模型：Banking Data 1)。

   Amazon ML 會顯示 ML 模型名稱、ID、建立時間和相關聯的資料來源 ID。
4. 選擇繼續。
5. 若要產生預測，您需要向 Amazon ML 提供您需要預測的資料。這稱為「輸入資料」。首先，將輸入資料放入資料來源，以便 Amazon ML 可以存取它。

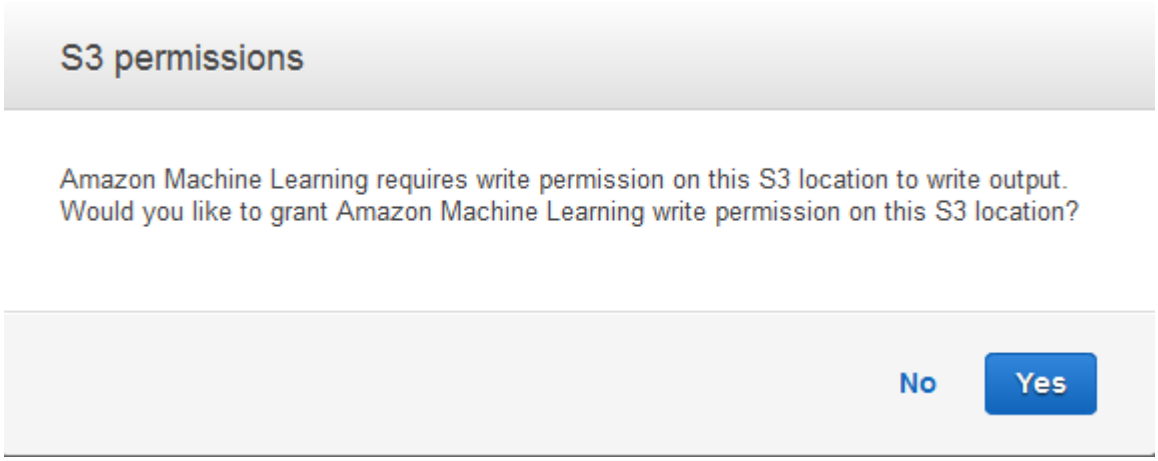
   Locate the input data (尋找輸入資料) 選擇 My data is in S3, and I need to create a datasource (我的資料存放在 S3 中，而且我需要建立資料來源)。

6. 針對 Datasource name (資料來源名稱) 輸入 **`Banking Data 2`**。
7. S3 Location (S3 位置) 輸入 `banking-batch.csv` 檔案的完整位置：**`######/banking-batch.csv`**。
8. Does the first line in your CSV contain the column names? (CSV 的第一行是否包含欄名？) 選擇 Yes (是)。
9. 選擇 Verify (驗證)。

   Amazon ML 會驗證資料的位置。
10. 選擇繼續。

11. 針對 S3 目的地，輸入您在步驟 1：準備資料中上傳檔案的 Amazon S3 位置名稱。Amazon ML 會在那裡上傳預測結果。
12. 對於批次預測名稱，接受預設值 **`Batch prediction: ML model: Banking Data 1`**。Amazon ML 會根據用於建立預測的模型來選擇預設名稱。在本教學課程中，模型和預測結果都會以訓練資料來源 `Banking Data 1` 命名。
13. 選擇檢閱。
14. 在 S3 permissions (S3 許可) 對話方塊中，選擇 Yes (是)。

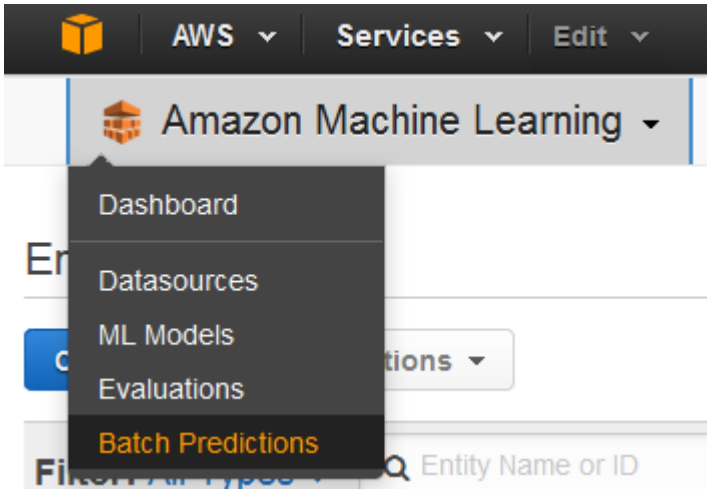


15. 在 Review (檢閱) 頁面上，選擇 Finish (完成)。

    批次預測請求會傳送至 Amazon ML 並輸入佇列。Amazon ML 處理批次預測所需的時間取決於資料來源的大小和 ML 模型的複雜性。當 Amazon ML 處理請求時，它會報告進行中的狀態。在批次預測完成後，請求的狀態會變更為 Completed (已完成)。您現在可以檢視結果。

檢視預測

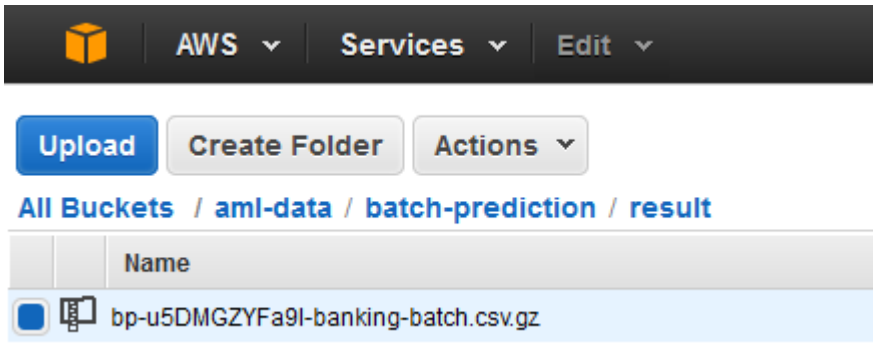
1. 選擇 Amazon Machine Learning，然後選擇 Batch Predictions (批次預測)。



2. 在預測清單中，選擇 Batch prediction: ML model: Banking Data 1 (批次預測：ML 模型：Banking Data 1)。Batch prediction info (批次預測資訊) 頁面隨即出現。

Name	Subscription propensity Predictions
ID	bp-u5DMGZYFa9l
Creation Time	Mar 5, 2015 3:28:33 PM
Status	Completed
Log	Download Log
Datasource ID	ds-33Rqgz9w3ee
ML Model ID	ml-u7ljoShX2kX
Input S3 URL	s3://aml-data/banking-batch.csv
Output S3 URL	s3://aml-data/

3. 若要檢視批次預測的結果，請前往 Amazon S3 主控台，網址為 https://console.aws.amazon.com/s3/：//Amazon S3S3。從該位置再導覽至結果資料夾，其名稱類似於 `s3://aml-data/batch-prediction/result`。



預測會存放在壓縮的 .gzip 檔案中，副檔名為 .gz。

4. 將預測檔案下載到您的桌面，解壓縮後開啟它。

bestAnswer	score
0	0.06046
0	0.00507
0	0.01410
0	0.00170
0	0.00184
0	0.07133
0	0.30811

該檔案會有兩欄：bestAnswer (最佳答案) 和 score (分數)，同時您的資料來源中的每項觀察各占一列。bestAnswer (最佳答案) 欄內的結果是以您在步驟 4：檢閱 ML 模型的預測效能並設定分數閾值時設定的分數閾值 0.77 為依據。score (分數) 大於 0.77 會得到 bestAnswer (最佳答案) 為 1 表示正回應或預測，而 score (分數) 小於 0.77 會得到 bestAnswer (最佳答案) 為 0 表示負回應或預測。

下列範例顯示以 0.77 為分數閾值的正負預測。

正預測：

bestAnswer	score
1	0.8228876

在此範例中，bestAnswer (最佳答案) 的值為 1，而 score (分數) 的值為 0.8228876。bestAnswer (最佳答案) 的值為 1 是因為 score (分數) 大於分數閾值 0.77。bestAnswer (最佳答案) 為 1 表示客戶可能會購買您的產品，因而視為正預測。

負預測：

bestAnswer	score
0	0.7695356

在此範例中，bestAnswer (最佳答案) 的值為 0 是因為 score (分數) 值 0.7695356 小於分數閾值 0.77。bestAnswer (最佳答案) 為 0 表示客戶不太可能購買您的產品，因而視為負預測。

批次結果的每列均會對應到您批次輸入中的一列 (您資料來源中的觀察)。

分析了預測結果後，您就可以執行您的目標行銷活動，例如向所有預測分數為 1 的人發送傳單。

既然您已建立、檢閱並使用了模型，請清理您建立的資料和 AWS 資源，以免產生不必要的費用，並保持工作空間整齊。

步驟 6：清除

若要避免產生額外的 Amazon Simple Storage Service (Amazon S3) 費用，請刪除存放在 Amazon S3 中的資料。您不需要支付其他未使用 Amazon ML 資源的費用，但我們建議您刪除這些資源，以保持工作區的乾淨。

刪除存放在 Amazon S3 中的輸入資料

1. 開啟位於 https://console.aws.amazon.com/s3/ 的 Amazon S3 主控台。
2. 導覽至您存放 `banking.csv`和 `banking-batch.csv` 檔案的 Amazon S3 位置。
3. 選取 `banking.csv`、`banking-batch.csv` 和 `.writePermissionCheck.tmp` 檔案。
4. 選擇動作，然後選擇刪除。
5. 出現確認提示時，請選擇 OK (確定)。

雖然您不需要付費保留 Amazon ML 執行的批次預測記錄，或是您在教學課程中建立的資料來源、模型和評估記錄，但建議您刪除這些預測，以避免工作區的雜亂。

刪除批次預測

1. 導覽至您存放批次預測輸出的 Amazon S3 位置。
2. 選擇 `batch-prediction` 資料夾。
3. 選擇動作，然後選擇刪除。
4. 出現確認提示時，請選擇 OK (確定)。

刪除 Amazon ML 資源

1. 在 Amazon ML 儀表板上，選取下列資源。
   - `Banking Data 1` 資料來源
   - `Banking Data 1_[percentBegin=0, percentEnd=70, strategy=sequential]` 資料來源
   - `Banking Data 1_[percentBegin=70, percentEnd=100, strategy=sequential]` 資料來源

- Banking Data 2 資料來源
- ML model: Banking Data 1 ML 模型
- Evaluation: ML model: Banking Data 1 評估

2. 選擇動作，然後選擇刪除。
3. 在對話方塊中選擇 Delete (刪除)，刪除所有選取的資源。

您現在已成功完成指導教學。若要繼續使用主控台建立資料來源、模型和預測，請參閱 Amazon Machine Learning 開發人員指南。若要了解如何使用 API，請參閱《Amazon Machine Learning API 參考》。

建立和使用資料來源

您可以使用 Amazon ML 資料來源來訓練 ML 模型、評估 ML 模型，並使用 ML 模型產生批次預測。資料來源物件包含輸入資料的相關中繼資料。當您建立資料來源時，Amazon ML 會讀取輸入資料、計算其屬性的描述性統計資料，並將統計資料、結構描述和其他資訊儲存為資料來源物件的一部分。建立資料來源之後，您可以使用 Amazon ML 資料洞見來探索輸入資料的統計屬性，也可以使用資料來源來訓練 ML 模型。

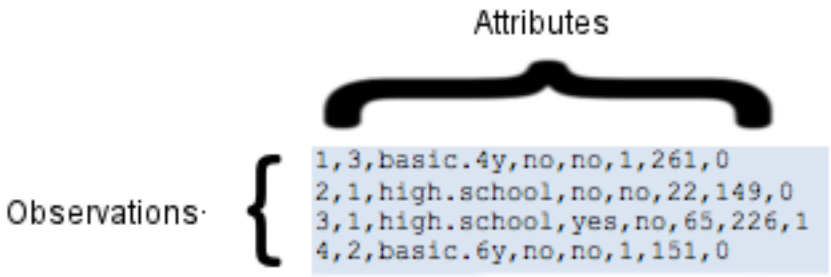
> **Note**
>
> 本節假設您熟悉 Amazon Machine Learning 概念。

主題

- 了解 Amazon ML 的資料格式
- 建立 Amazon ML 的資料結構描述
- 分割您的資料
- 資料的深入解析
- 搭配使用 Amazon S3 與 Amazon ML
- 從 Amazon Redshift 中的資料建立 Amazon ML 資料來源
- 使用 Amazon RDS 資料庫的資料來建立 Amazon ML 資料來源

了解 Amazon ML 的資料格式

輸入資料是用來建立資料來源的資料。您必須以逗號分隔值 (.csv) 格式儲存輸入資料。.csv 檔案中的每個資料列都是單一資料記錄或觀察。.csv 檔案中的每個資料行都會包含觀察的屬性。例如，下圖顯示 .csv 檔案的內容，而此檔案包含四個觀察，且各有自己的資料列。每個觀察都會包含八個以逗號分隔的屬性。這些屬性代表觀察所代表之每個個人的下列資訊：customerId,jobId,education,housing,loan,campaign,duration,willRespondToCampaign。



屬性

Amazon ML 需要每個屬性的名稱。您可以透過下列方式指定屬性名稱：

- .csv 檔案第一行 (也稱為標頭行) 包含用作您輸入資料的屬性名稱
- 在個別結構描述檔中包含屬性名稱，而結構描述檔位在與輸入資料相同的 S3 儲存貯體中

如需使用結構描述檔的詳細資訊，請參閱建立資料結構描述。

下列 .csv 檔案範例將屬性名稱包含在標頭行中。

```
customerId,jobId,education,housing,loan,campaign,duration,willRespondToCampaign
1,3,basic.4y,no,no,1,261,0
2,1,high.school,no,no,22,149,0
3,1,high.school,yes,no,65,226,1
4,2,basic.6y,no,no,1,151,0
```

輸入檔格式需求

包含您輸入資料的 .csv 檔案必須符合下列需求：

- 必須為使用 ASCII、Unicode 或 EBCDIC 這類字元集的純文字。
- 由觀察組成，一行一個觀察。
- 對於每個觀察，必須以逗號分隔屬性值。
- 如果屬性值包含逗號 (分隔符號)，則必須用雙引號括住整個屬性值。
- 每個觀察的結尾都必須是行尾字元，此字元是指出行尾的特殊字元或一系列字元。

- 屬性值不可以包含行尾字元，即使使用雙引號括住屬性值也是一樣。
- 每個觀察都必須有相同數目的屬性和一系列的屬性。
- 每個觀察值不得大於 100 KB。Amazon ML 會在處理期間拒絕任何大於 100 KB 的觀察。如果 Amazon ML 拒絕超過 10，000 個觀察，則會拒絕整個 .csv 檔案。

使用多個檔案做為 Amazon ML 的資料輸入

您可以將輸入做為單一檔案或檔案集合提供給 Amazon ML。集合必須滿足這些條件：

- 所有檔案都必須具有相同的資料結構描述。
- 所有檔案都必須位於相同的 Amazon Simple Storage Service (Amazon S3) 字首，而您為集合提供的路徑必須以斜線 ('/') 字元結尾。

例如，如果您的資料檔案命名為 input1.csv、input2.csv 和 input3.csv，而 S3 儲存貯體名稱為 s3://examplebucket，則您的檔案路徑可能如下所示：

s3://examplebucket/path/to/data/input1.csv

s3://examplebucket/path/to/data/input2.csv

s3://examplebucket/path/to/data/input3.csv

您會提供下列 S3 位置做為 Amazon ML 的輸入：

's3://examplebucket/path/to/data/'

CSV 格式的行尾字元

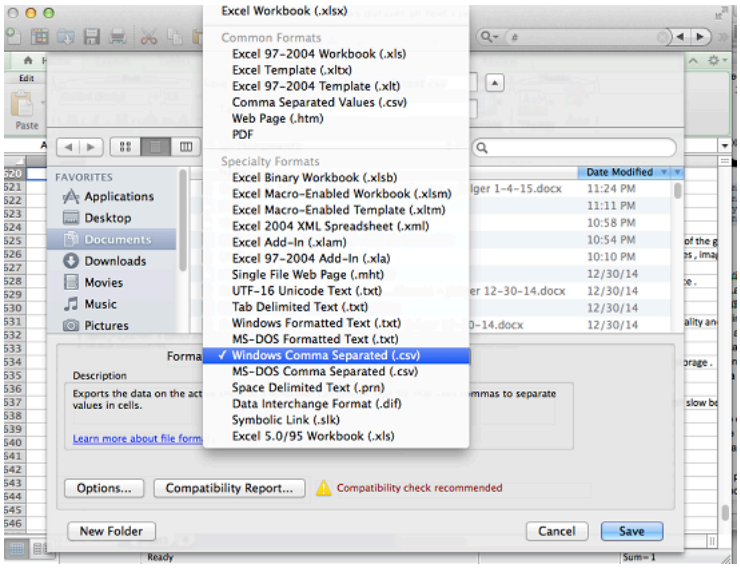
當您建立 .csv 檔案時，每個觀察的結尾都是特殊行尾字元。當您按 Enter 或 Return 鍵時，此字元不會顯示，但會自動包含在每個觀察的結尾。代表行尾的特殊字元會根據作業系統而不同。Linux 或 OS X 這類 Unix 系統使用「換行」字元，其以 "\n" (十進位 ASCII 代碼 10，或十六進位的 0x0a) 表示。Microsoft Windows 使用「歸位字元」和「換行字元」這兩個字元，其以 "\r\n" (十進位 ASCII 代碼 13 和 10，或十六進位的 0x0d 和 0x0a) 表示。

如果您想要使用 OS X 和 Microsoft Excel 建立 .csv 檔案，請執行下列程序。請務必選擇正確的格式。

在使用 OS X 和 Excel 時儲存 .csv 檔案

1. 儲存 .csv 檔案時，請選擇 Format (格式)，然後選擇 Windows Comma Separated (.csv) (Windows 逗號分隔)。

2. 選擇儲存。



> **Important**
>
> 請不要使用逗號分隔值 (.csv) 或 MS-DOS 逗號分隔 (.csv) 格式來儲存 .csv 檔案，因為 Amazon ML 無法讀取它們。

建立 Amazon ML 的資料結構描述

「結構描述」包含輸入資料中的所有屬性和其對應資料類型。它可讓 Amazon ML 了解資料來源中的資料。Amazon ML 會使用結構描述中的資訊來讀取和解譯輸入資料、運算統計資料、套用正確的屬性轉換，以及微調其學習演算法。如果您不提供結構描述，Amazon ML 會從資料推斷一個結構描述。

範例結構描述

若要讓 Amazon ML 正確讀取輸入資料並產生準確的預測，必須為每個屬性指派正確的資料類型。讓我們演練範例以查看如何將資料類型指派給屬性，以及如何在結構描述中包含屬性和資料類型。我們將範例稱為 "Customer Campaign" (客戶行銷活動)，因為我們想要預測哪些客戶將會回應我們的電子郵件行銷活動。我們的輸入檔是一個含 9 個資料行的 .csv 檔案：

```
1,3,web developer,basic.4y,no,no,1,261,0

2,1,car repair,high.school,no,no,22,149,0
```

```
3,1,car mechanic,high.school,yes,no,65,226,1

4,2,software developer,basic.6y,no,no,1,151,0
```

此資料的這個結構描述：

```
{
    "version": "1.0",
    "rowId": "customerId",
    "targetAttributeName": "willRespondToCampaign",
    "dataFormat": "CSV",
    "dataFileContainsHeader": false,
    "attributes": [
        {
            "attributeName": "customerId",
            "attributeType": "CATEGORICAL"
        },
        {
            "attributeName": "jobId",
            "attributeType": "CATEGORICAL"
        },
        {
            "attributeName": "jobDescription",
            "attributeType": "TEXT"
        },
        {
            "attributeName": "education",
            "attributeType": "CATEGORICAL"
        },
        {
            "attributeName": "housing",
            "attributeType": "CATEGORICAL"

        },
        {
            "attributeName": "loan",
            "attributeType": "CATEGORICAL"

        },
        {
            "attributeName": "campaign",
            "attributeType": "NUMERIC"
        },
```

```
        {
            "attributeName": "duration",
            "attributeType": "NUMERIC"
        },

        {
            "attributeName": "willRespondToCampaign",
            "attributeType": "BINARY"
        }

    ]
}
```

在這個範例的結構描述檔案中，rowId 的值為 customerId：

```
"rowId": "customerId",
```

willRespondToCampaign 屬性會定義為目標屬性：

```
"targetAttributeName": "willRespondToCampaign ",
```

customerId 屬性和 CATEGORICAL 資料類型是與第一個資料行建立關聯、jobId 屬性和 CATEGORICAL 資料類型是與第二個資料行建立關聯、jobDescription 屬性和 TEXT 資料類型是與第三個資料行建立關聯、education 屬性和 CATEGORICAL 資料類型是與第四個資料行建立關聯，以此類推。第 9 個資料行是與具有 willRespondToCampaign 資料類型的 BINARY 屬性建立關聯，而且此屬性也定義為目標屬性。

使用 targetAttributeName 欄位

targetAttributeName 值是您想要預測的屬性名稱。建立或評估模型時，您必須指派 targetAttributeName。

當您訓練或評估 ML 模型時， 會targetAttributeName識別輸入資料中的屬性名稱，其中包含目標屬性的「正確」答案。Amazon ML 使用目標，其中包含正確答案，以探索模式並產生 ML 模型。

當您評估模型時，Amazon ML 會使用目標來檢查預測的準確性。建立和評估 ML 模型之後，即可搭配使用資料與未指派的 targetAttributeName，以使用 ML 模型來產生預測。

您可以在建立資料來源或在結構描述檔案中定義 Amazon ML 主控台中的目標屬性。如果您建立自己的結構描述檔案，請使用下列語法來定義目標屬性：

```
"targetAttributeName": "exampleAttributeTarget",
```

在這個範例中，exampleAttributeTarget 是輸入檔中為目標屬性的屬性名稱。

使用 rowID 欄位

row ID 是與輸入資料中屬性建立關聯的選用旗標。如果指定，標示為 row ID 的屬性會包含在預測輸出中。此屬性可讓您更輕鬆地將哪個預測與哪個觀察建立關聯。不錯的 row ID 範例是客戶識別符或類似的唯一屬性。

> **Note**
>
> 資料列 ID 僅供您參考。Amazon ML 在訓練 ML 模型時不會使用它。選取屬性作為資料列識別符，不要將它用於培訓 ML 模型。

您可以在建立資料來源或在結構描述檔案中定義 row IDAmazon ML 主控台中的 。如果您要建立自己的結構描述檔案，請使用下列語法來定義 row ID：

```
"rowId": "exampleRow",
```

在前述範例中，exampleRow 是輸入檔中定義為資料列識別符的屬性名稱。

產生批次預測時，您可能會收到下列輸出：

```
tag,bestAnswer,score
55,0,0.46317
102,1,0.89625
```

在這個範例中，RowID 代表 customerId 屬性。例如，customerId 55 預期會回應低可信度 (0.46317) 的電子郵件行銷活動，而 customerId 102 預期會回應高可信度 (0.89625) 的電子郵件行銷活動。

使用 AttributeType 欄位

在 Amazon ML 中，屬性有四種資料類型：

：二進位

針對只有兩種可能狀態 (例如 BINARY 或 yes) 的屬性，選擇 no。

例如，追蹤人員是否為新客戶的 isNew 屬性會有 true 值指出個人是新客戶，而 false 值指出其非新客戶。

有效負值是 0、n、no、f 和 false。

有效正值是 1、y、yes、t 和 true。

Amazon ML 會忽略二進位輸入的情況，並去除周圍的空格。例如，" FaLSe " 是有效的二元值。您可以混合在相同資料來源中使用的二進位值，例如使用 true、 no和 1。Amazon ML 只會1針對二進位屬性輸出 0和 。

分類

針對採用有限數目之唯一字串值的屬性，選擇 CATEGORICAL。例如，使用者識別符、月分和郵遞區號是分類值。分類屬性視為單一字串，而且不再進一步進行字符化。

數值

針對採用數量作為值的屬性，選擇 NUMERIC。

例如，溫度、重量和點擊速率是數值。

並非所有保留數字的屬性都是數值。分類屬性 (例如該月天數和識別符) 通常呈現為數字。若要視為數值，某個數字必須相當於另一個數字。例如，客戶識別符 664727 不會告訴您有關客戶識別符 124552 的任何資訊，但重量 10 告訴您該屬性比重量 5 的屬性還要重。該月天數不是數值，因為某個月的第一天可能發生在另一個月的第二天之前或之後。

> **Note**
>
> 當您使用 Amazon ML 建立結構描述時，它會將Numeric資料類型指派給使用數字的所有屬性。如果 Amazon ML 建立您的結構描述，請檢查指派是否不正確，並將這些屬性設定為 CATEGORICAL。

Text (文字)

針對為文字字串的屬性，選擇 TEXT。在文字屬性中讀取時，Amazon ML 會將它們轉換為字符，以空格分隔。

例如，email subject 會成為 email 和 subject，而 email-subject here 會成為 email-subject 和 here。

如果訓練結構描述中變數的資料類型不符合評估結構描述中該變數的資料類型，Amazon ML 會變更評估資料類型以符合訓練資料類型。例如，如果訓練資料結構描述將 的資料類型指派給TEXT變數 age，但評估結構描述將 的資料類型指派給 NUMERIC age，則 Amazon ML 會將評估資料中的存留期視為TEXT變數，而不是 NUMERIC。

如需與每種資料類型建立關聯之統計資料的資訊，請參閱描述性統計資料。

提供結構描述給 Amazon ML

每個資料來源都需要結構描述。您可以選擇兩種方式為 Amazon ML 提供結構描述：

- 允許 Amazon ML 推斷輸入資料檔案中每個屬性的資料類型，並自動為您建立結構描述。
- 當您上傳 Amazon Simple Storage Service (Amazon S3) 資料時，請提供結構描述檔案。

允許 Amazon ML 建立您的結構描述

當您使用 Amazon ML 主控台建立資料來源時，Amazon ML 會根據變數的值使用簡單的規則來建立結構描述。強烈建議您檢閱 Amazon ML 建立的結構描述，並在資料類型不正確時加以更正。

提供結構描述

建立結構描述檔案後，您需要將其提供給 Amazon ML。您有兩種選擇：

1. 使用 Amazon ML 主控台提供結構描述。

   使用主控台建立資料來源，以及將 .schema 副檔名附加到輸入資料檔案的檔案名稱後面，來包含結構描述檔案。例如，如果輸入資料的 Amazon Simple Storage Service (Amazon S3) URI 為 s3：//my-bucket-name/data/input.csv，結構描述的 URI 將為 s3：//my-bucket-name/data/input.csv.schema。Amazon ML 會自動尋找您提供的結構描述檔案，而不是嘗試從資料推斷結構描述。

   若要使用檔案目錄做為 Amazon ML 的資料輸入，請將 .schema 延伸模組附加到您的目錄路徑。例如，如果您的資料檔案位於位置 s3://examplebucket/path/to/data/，則結構描述的 URI 會是 s3://examplebucket/path/to/data/.schema。
2. 使用 Amazon ML API 提供結構描述。

   如果您打算呼叫 Amazon ML API 來建立資料來源，您可以將結構描述檔案上傳至 Amazon S3，然後在 CreateDataSourceFromS3 API DataSchemaLocationS3 的 屬性中提供該檔案的 URI。如需詳細資訊，請參閱 CreateDataSourceFromS3。

您可以直接在 CreateDataSource* 承載中提供結構描述，APIs而不是先將其儲存至 Amazon S3。做法是將完整結構描述字串放在 DataSchema、CreateDataSourceFromS3 或 CreateDataSourceFromRDS API 的 CreateDataSourceFromRedshift 屬性中。如需詳細資訊，請參閱《Amazon Machine Learning API 參考》。

分割您的資料

ML 模型的基本目標是對於用於訓練模型之外的未來資料執行個體，可進行準確的預測。在使用 ML 模型來進行預測之前，我們需要評估模型的預測效能。為了估計 ML 模型對於未知資料的預測品質，我們可以針對現在已知其答案的資料，保留或分割其一部分來做為未來資料的代理，並評估 ML 模型對於該資料預測正確答案的準確程度。您將資料來源分割成一部分做為訓練資料來源，另一部分做為評估資料來源。

Amazon ML 提供三種分割資料的選項：

- 預先分割資料 - 您可以將資料分割成兩個資料輸入位置，然後再將其上傳至 Amazon Simple Storage Service (Amazon S3)，並建立兩個單獨的資料來源。
- Amazon ML 循序分割 - 您可以指示 Amazon ML 在建立訓練和評估資料來源時循序分割資料。
- Amazon ML 隨機分割 - 您可以在建立訓練和評估資料來源時，指示 Amazon ML 使用種子隨機方法分割資料。

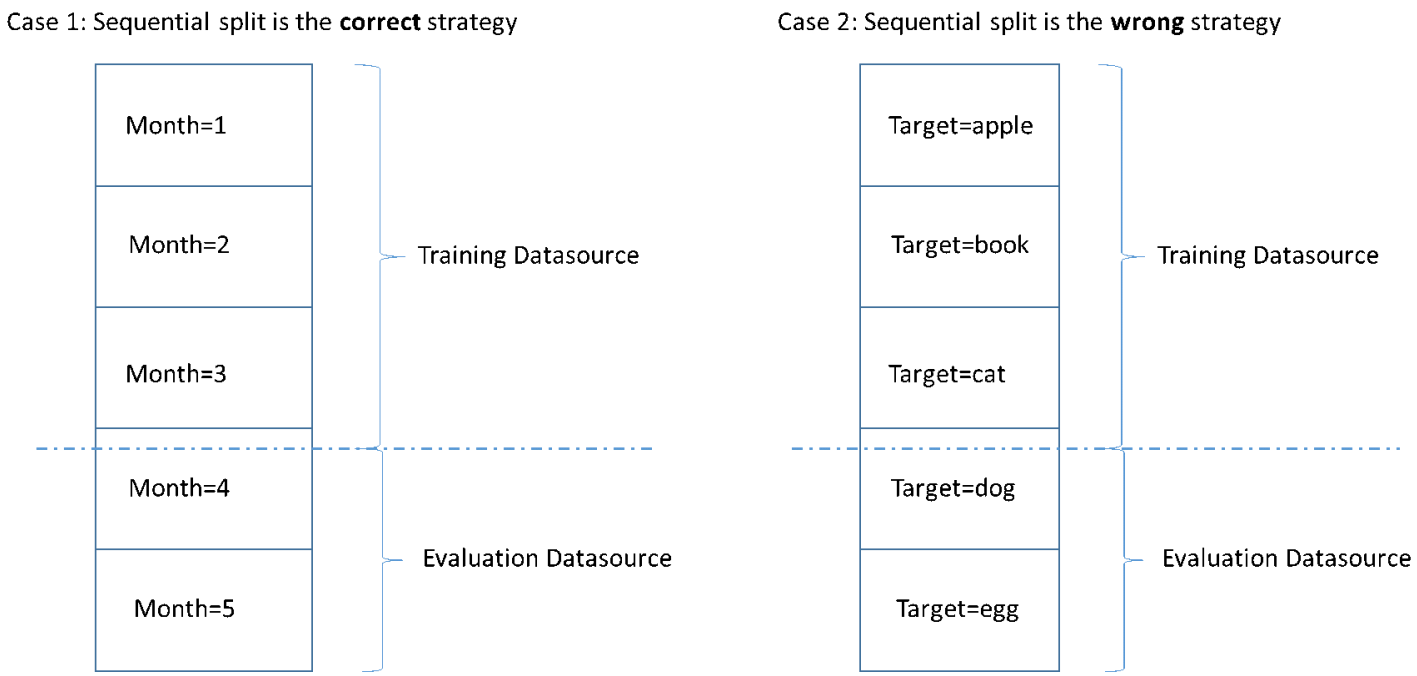
預先分割資料

如果您想明確控制訓練和評估資料來源中的資料，請將您的資料分割為不同的資料位置，並建立輸入和評估位置的不同資料來源。

序列分割資料

分割輸入資料用於訓練和評估的簡單方式，就是選取不重疊的資料子集，同時保留資料記錄的順序。如果您想要評估特定日期或特定時間範圍內的 ML 模型，這個方法非常有用。例如，假設您有過去五個月的客戶互動資料，而且您想要使用此歷史資料來預測下個月的客戶互動。將範圍開頭的資料用於訓練，範圍結束的資料用於評估，可能會比使用從整個資料範圍抽取的資料記錄，能產生更準確的預估值資料。

下圖顯示何時應使用序列分割策略，以及何時應使用隨機策略的範例。

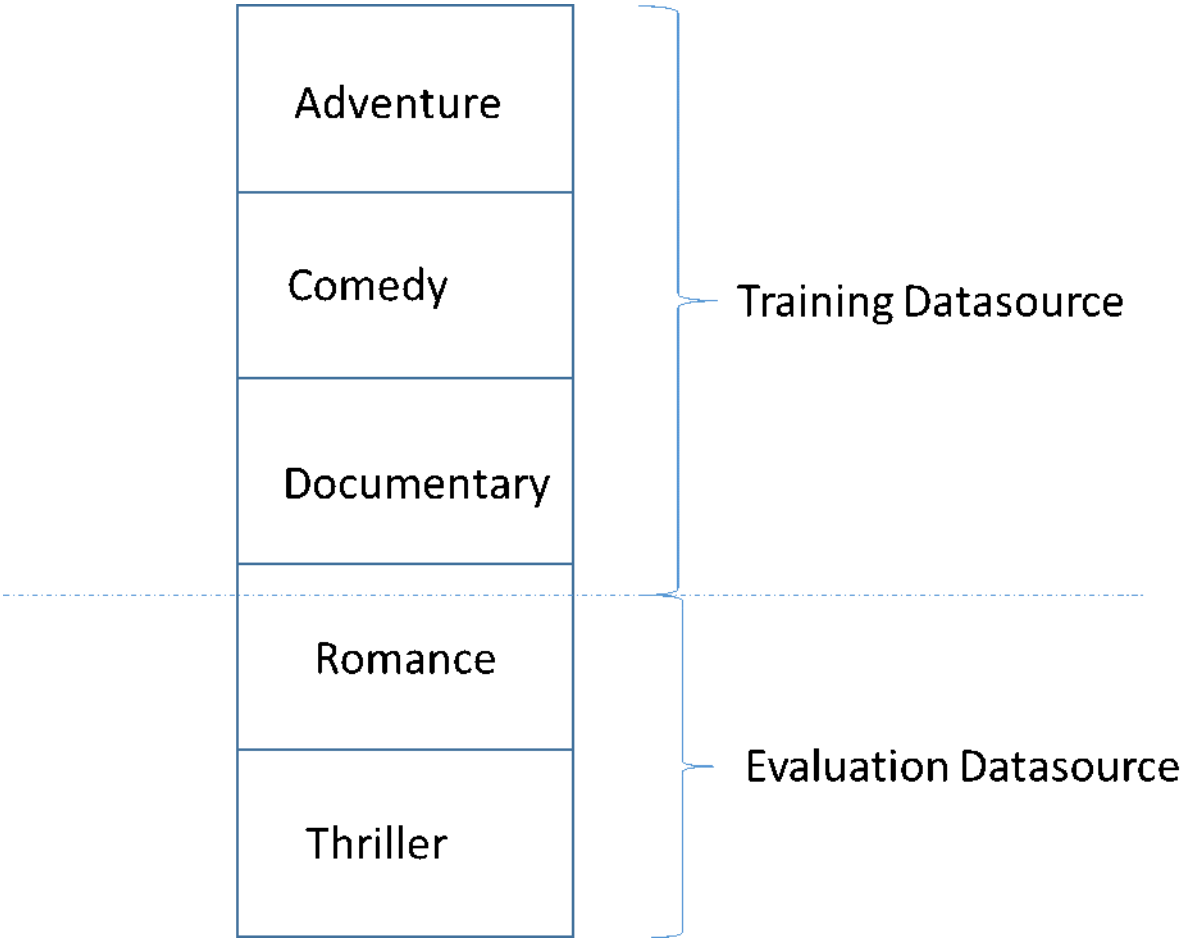


建立資料來源時，您可以選擇依序分割資料來源，Amazon ML 會使用前 70% 的資料進行訓練，其餘 30% 的資料用於評估。這是您使用 Amazon ML 主控台分割資料的預設方法。

隨機分割資料

隨機將輸入資料分割為訓練和評估資料來源，可確保訓練和評估資料來源中的資料分佈類似。當您不需要保留輸入資料的順序，請選擇此選項。

Amazon ML 使用種子虛擬隨機數字產生方法來分割您的資料。種子是部分根據輸入字串值，部分根據資料本身的內容。根據預設，Amazon ML 主控台會使用輸入資料的 S3 位置做為字串。API 使用者可以提供自訂的字串。這表示指定相同的 S3 儲存貯體和資料，Amazon ML 每次都會以相同的方式分割資料。若要變更 Amazon ML 分割資料的方式，您可以使用 `CreateDatasourceFromS3`、 或 `CreateDatasourceFromRDS` API`CreateDatasourceFromRedshift`，並提供種子字串的值。使用這些 API 來建立用於訓練和評估的個別資料來源時，請務必對這兩個資料來源使用相同的種子字串值並對一個資料來源使用補充旗標，以確保訓練和評估資料之間沒有重疊。



開發高品質 ML 模型中常見的陷阱，是在與用於訓練之資料不類似的資料上評估 ML 模型。例如，假設您使用 ML 來預測電影類型，而您的訓練資料包含冒險片、喜劇片以及紀錄片類型的電影。不過，您的評估資料只包含愛情片和驚悚片類型的資料。在這種情況下，ML 模型並未學習到愛情片和驚悚片類型的任何資訊，評估程序也無法評估模型從冒險片、喜劇片以及紀錄片類型的學習程度。因此，類型資訊無用，對於所有類型的 ML 模型預測品質受到損害。模型和評估太過不同 (有非常不同的描述統計資料)，因此無用。這可能發生在輸入資料依資料集的某一欄排序，然後依序分割。

如果您的訓練和評估資料來源有不同的資料分佈，您會在模型評估中看到評估提醒。如需評估提醒的詳細資訊，請參閱評估提醒。

如果您已將輸入資料隨機化，例如，在 Amazon S3 中隨機隨機播放輸入資料，或在建立資料來源時使用 Amazon Redshift SQL 查詢的 `random()`函數或 MySQL SQL 查詢的 `rand()`函數，則不需要在 Amazon ML 中使用隨機分割。在這些情況下，您可以倚賴序列分割選項來建立具有類似分佈的訓練和評估資料來源。

資料的深入解析

Amazon ML 會計算輸入資料的描述性統計資料，供您用來了解資料。

描述性統計資料

Amazon ML 會針對不同的屬性類型計算下列描述性統計資料：

數值：

- 色階分佈圖
- 無效值的數量
- 最小值、中間值、平均值與最大值

二元與分類：

- (每個類別的相異值) 計數
- 數值色階分佈圖
- 最常出現的值
- 不重複的值計數
- true 值的百分比 (僅限二元)
- 最重要的單字
- 最常出現的單字

文字：

- 屬性的名稱
- 與目標的相互關聯性 (如有設定目標)
- 總字數
- 不重複的文字
- 單一資料列中的字數範圍
- 單字長度範圍
- 最重要的單字

在 Amazon ML 主控台上存取 Data Insights

在 Amazon ML 主控台上，您可以選擇任何資料來源的名稱或 ID，以檢視其 Data Insights 頁面。此頁面提供指標與視覺效果，可讓您了解資料來源相關聯的輸入資料，包括下列資訊：

- 資料摘要
- 目標分佈
- 缺少值
- 無效值
- 變數的摘要統計資料 (依資料類型)
- 變數的分佈 (依資料類型)

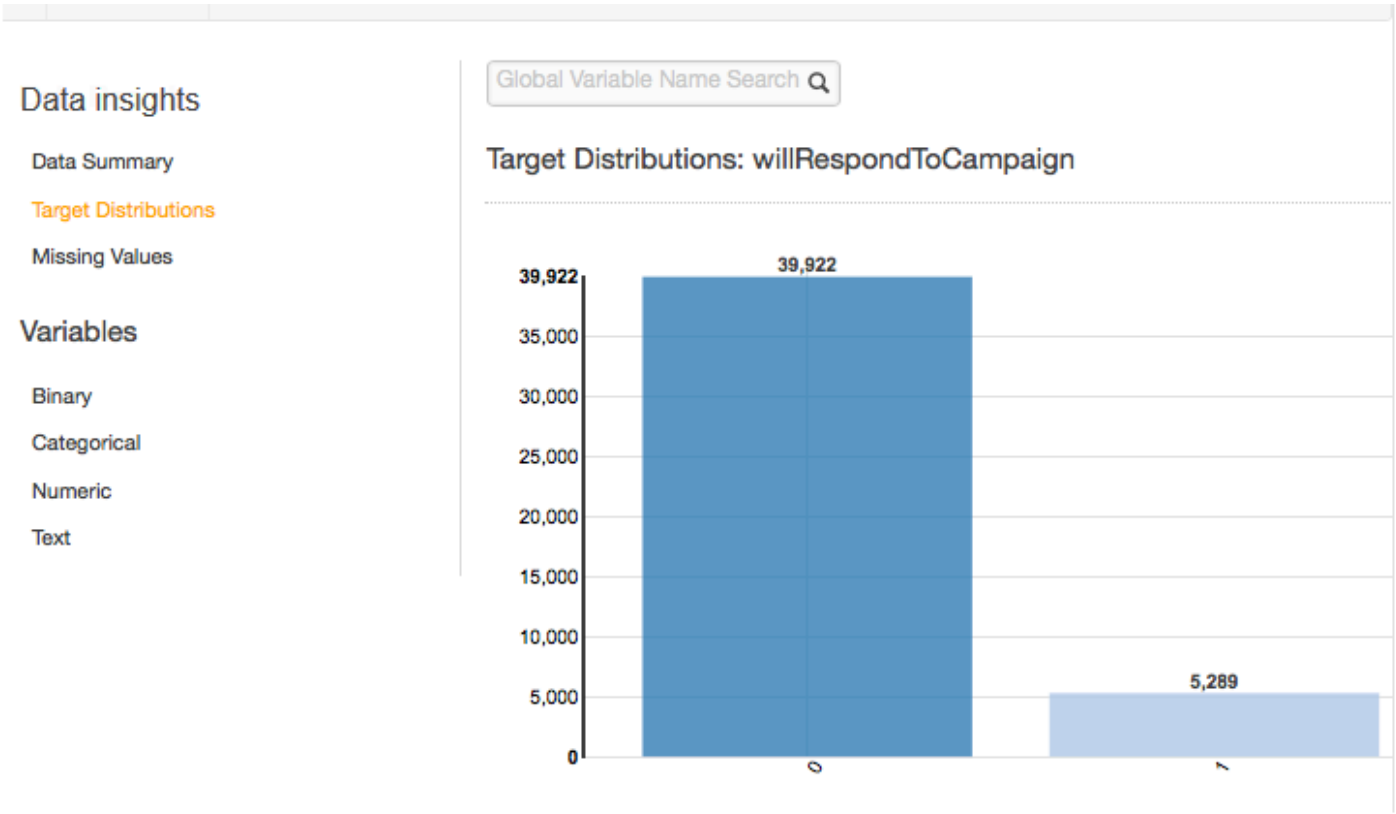
下列各節將詳細說明指標與視覺效果。

資料摘要

資料來源的資料摘要報告顯示摘要資訊，包括資料來源 ID、名稱、完成位置、目前的狀態、目標屬性、輸入資料資訊 (S3 儲存貯體位置、資料格式、已處理的記錄數與處理期間發生的錯誤記錄數)，以及變數的數量 (依資料類型)。

目標分佈

目標分佈報告顯示資料來源的目標屬性分佈。下列範例有 39,922 個觀察，其中的 willRespondToCampaign 目標屬性等於 0。這是未回應電子郵件行銷活動的客戶人數。總共有 5,289 個觀察，其中 willRespondToCampaign 等於 1。這是已回應電子郵件行銷活動的客戶人數。



缺少的值

缺少的值報告會列出輸入資料中缺少值的屬性。只有資料類型為數值的屬性才會缺少值。因為缺少值可能會影響 ML 模型的訓練品質，所以建議您盡可能提供所缺少的值。

在 ML 模型訓練期間，如果遺失目標屬性，Amazon ML 會拒絕對應的記錄。如果目標屬性存在於記錄中，但缺少另一個數值屬性的值，則 Amazon ML 會忽略缺少的值。在此情況下，Amazon ML 會建立替代屬性，並將其設定為 1，以表示缺少此屬性。這可讓 Amazon ML 從遺失值的出現中學習模式。

無效值

只有數值與二元資料類型才會出現無效值。您可以檢視資料類型報告中的變數摘要統計資料來尋找無效值。在下列範例中，持續時間的數值屬性有一個無效值，二元資料類型有兩個無效值 (分別在房屋屬性與貸款屬性中)。

Numeric Variables

Variables	Correlations to Target	Missing Values	Invalid Values	Range	Mean	Median	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

Binary Variables

Variables	Correlations to Target	Percent True	Invalid Values	Preview
campaign	NA	100%	27667 (61%)	
housing	0.01842	56%	1 (0%)	
loan	0.00656	16%	1 (0%)	
willRespondToCampaign	NA	12%	0 (0%)	

變數與目標的相互關聯性

建立資料來源之後，Amazon ML 可以評估資料來源，並識別變數與目標之間的相互關聯或影響。例如，產品價格對於能否成為暢銷商品的影響可能很大，但產品大小對於預測的影響可能就很小。

一般會建議在訓練資料中加入愈多的變數愈好。但加入許多對於預測影響很小之變數所帶來的干擾，可能會對您的 ML 模型品質與正確性造成負面影響。

您可以移除訓練模型時影響力很小的變數，藉此改善模型的預測效能。您可以在配方中定義哪些變數可供機器學習程序使用，這是 Amazon ML 的轉換機制。若要進一步了解配方，請參閱機器學習的資料轉換。

屬性的摘要統計資料 (依資料類型)

在資料深入解析報告中，您可以依下列資料類型檢視屬性摘要統計資料：

- 二進位
- 分類
- 數值
- 文字

二元資料類型的摘要統計資料會顯示所有的二元屬性。Correlations to target (與目標的相互關聯性) 資料行會顯示目標資料行與屬性資料行中相同的資訊。Percent true (true 的百分比) 資料行顯示觀察值為 1 的百分比。Invalid values (無效值) 資料行顯示無效值的數目，以及每個屬性的無效值所占百分比。Preview (預覽) 資料行提供每個屬性圖形化分佈的連結。

Binary Variables

Variables	Correlations to Target	Percent True	Invalid Values	Preview
campaign	NA	100%	27667 (61%)	
housing	0.01842	56%	1 (0%)	
loan	0.00656	16%	1 (0%)	
willRespondToCampaign	NA	12%	0 (0%)	

分類資料類型的摘要統計資料顯示所有分類屬性，以及不重複的值、最常出現的值與最少出現的值的數量。Preview (預覽) 資料行提供每個屬性圖形化分佈的連結。

Categorical Variables

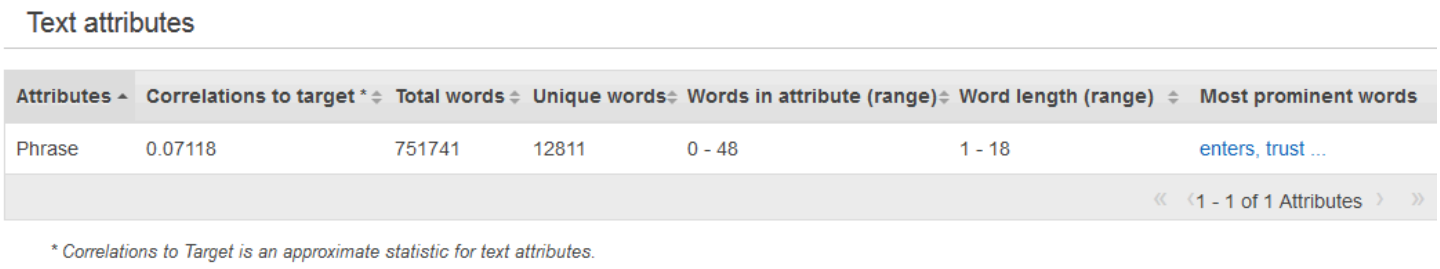
Variables	Correlations to Target	Unique Values	Most Frequent	Least Frequent	Preview
campaign	0.00433	49	1	39	
customerId	NA	45211	45211	1	
education	0.00355	5	secondary		
housing	0.01846	4	1		
jobId	0.00671	13	blue-collar		
willRespondToCampaign	NA	3	0		

數值資料類型的摘要統計資料顯示所有數值屬性，以及缺少的值、無效值、值的範圍、平均值與中間值的數量。Preview (預覽) 資料行提供每個屬性圖形化分佈的連結。

Numeric Variables

Variables	Correlations to Target	Missing Values	Invalid Values	Range	Mean	Median	Preview
duration	0.05165	2 (0%)	1 (0%)	0 - 4918	258.1618	180	

文字資料類型的摘要統計資料顯示所有的文字屬性、該屬性的總字數、該屬性中不重複的單字數、屬性的字數範圍、單字長度的範圍與最重要的單字。Preview (預覽) 資料行提供每個屬性圖形化分佈的連結。



下一個範例顯示文字變數 "review" 的文字資料類型統計資料，其中包含四筆記錄。

```
1. The fox jumped over the fence.
2. This movie is intriguing.
3.
4. Fascinating movie.
```

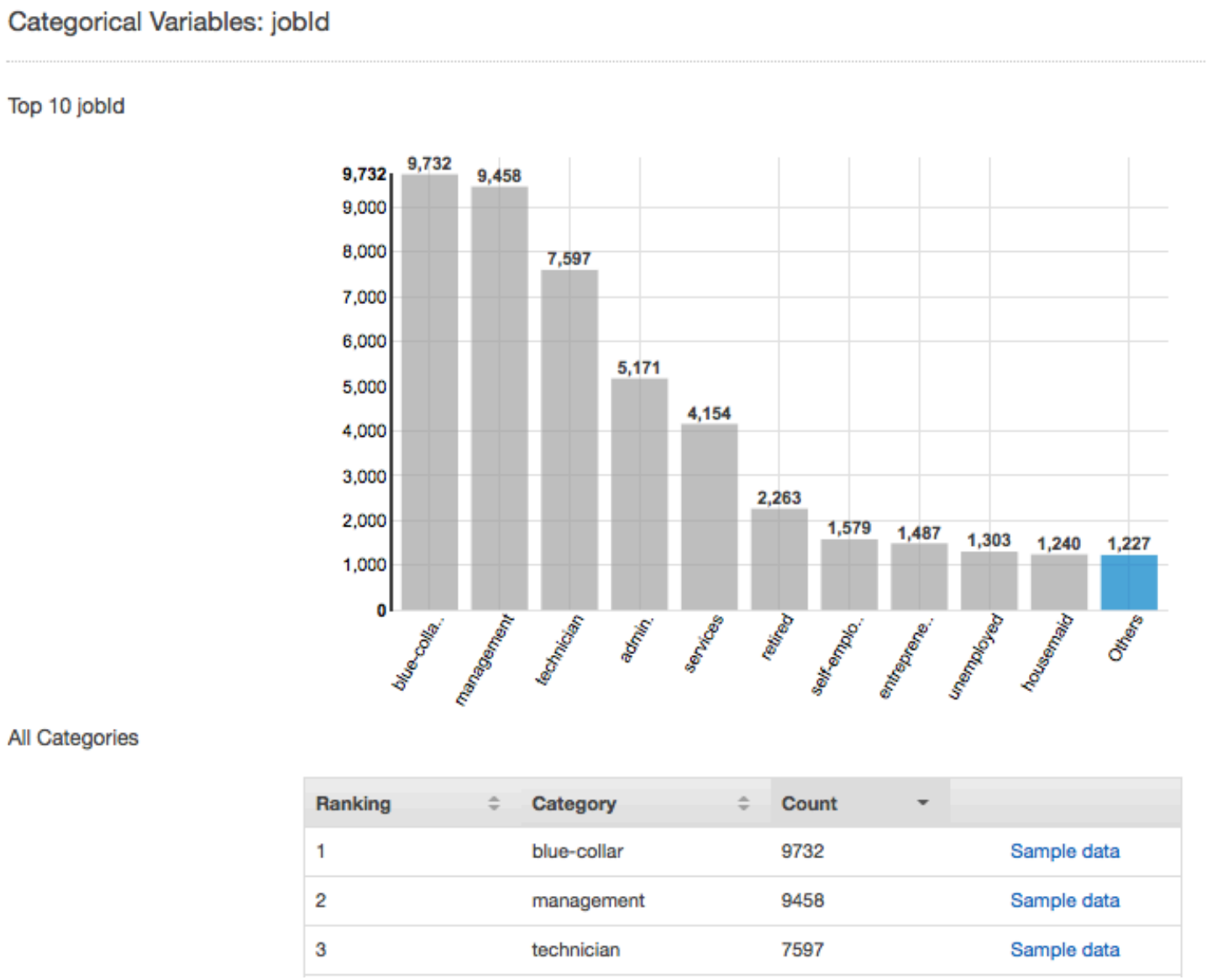
此範例的資料行顯示下列資訊。

- Attributes (屬性) 資料行顯示變數的名稱。在此範例中，此資料行會顯示 "review"。
- 若已指定目標，才會有 Correlations to target (與目標的相互關聯性) 資料行。相互關聯性測量此屬性所提供與目標相關的資訊量。相互關聯性愈高，此屬性所提供與目標相關的資訊量愈多。相互關聯性在測量文字屬性的簡單表示與目標之間共有的資訊。
- Total words (總字數) 資料行顯示字符化每筆記錄後所產生的字數，並會以空格分隔每個單字。在此範例中，此資料行會顯示 "12"。
- Unique words (不重複的單字) 資料行顯示屬性中不重複的單字數。在此範例中，此資料行會顯示 "10"。
- Words in attribute (range) (屬性中的字數 (範圍)) 資料行顯示屬性之單一資料列中的字數。在此範例中，此資料行會顯示 "0-6"。
- Word length (range) (單字長度 (範圍)) 資料行顯示單字中的字元數範圍。在此範例中，此資料行會顯示 "2-11"。
- Most prominent words (最重要的單字) 資料行顯示屬性中出現之單字的排名清單。如有目標屬性，所有單字會依其與目標的相互關聯性排名，亦即，相互關聯性最高的單字會最先列出。若資料中沒有目標，則所有單字會依其熵排名。

了解分類與二元屬性的分佈

您可以按一下與分類或二元屬性相關聯的 Preview (預覽) 連結檢視該屬性的分佈，以及屬性的每一個分類值輸入檔案中的範例資料。

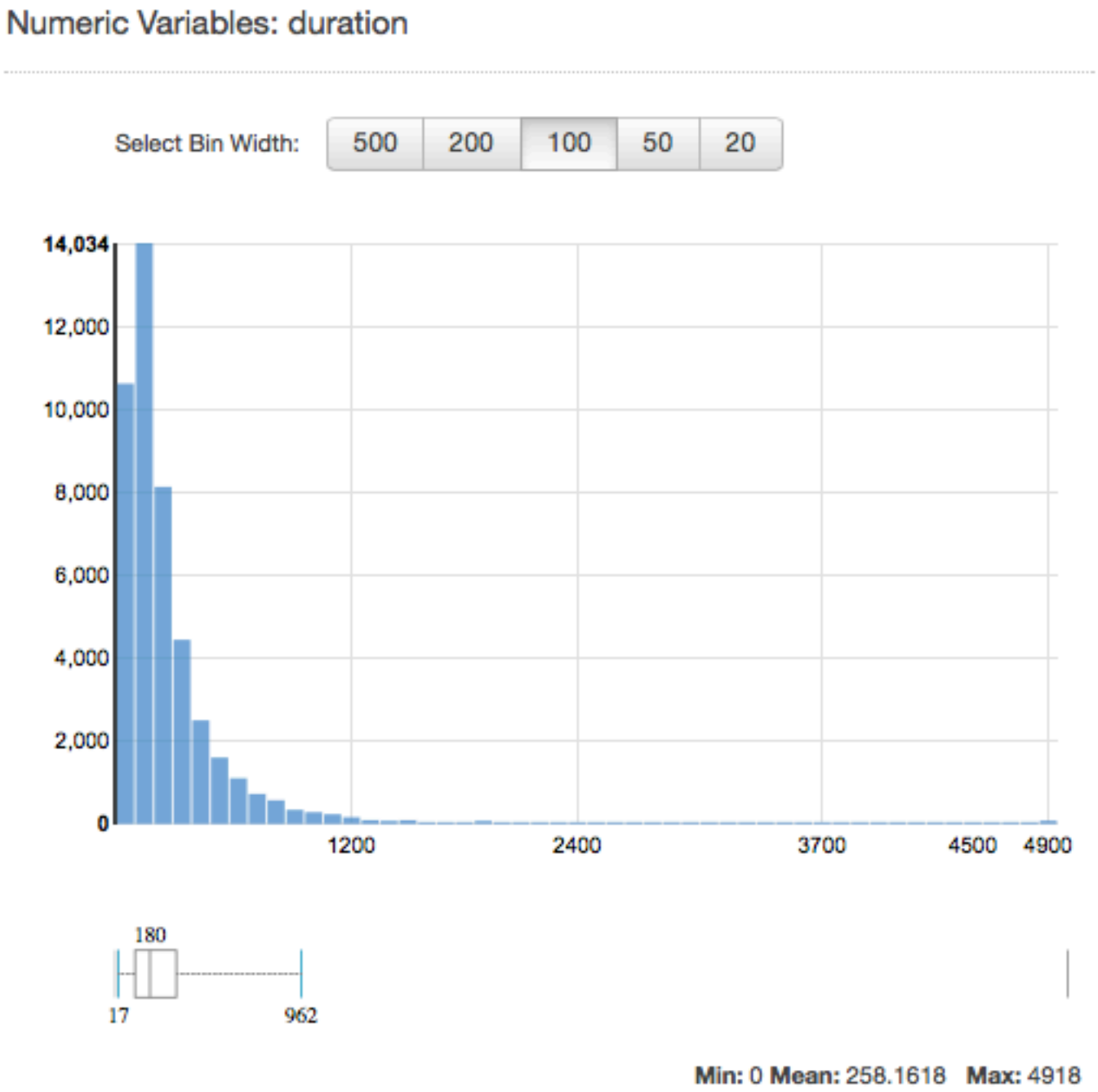
例如，下列螢幕擷取畫面顯示分類屬性 jobId 的分佈。此分佈顯示前 10 個分類值，以及分組為「其他」的所有其他值。其會排名前 10 個分類值，並提供輸入檔案中包含該值的觀察數，以及可檢視輸入資料檔案中範例觀察的連結。



了解數值屬性的分佈

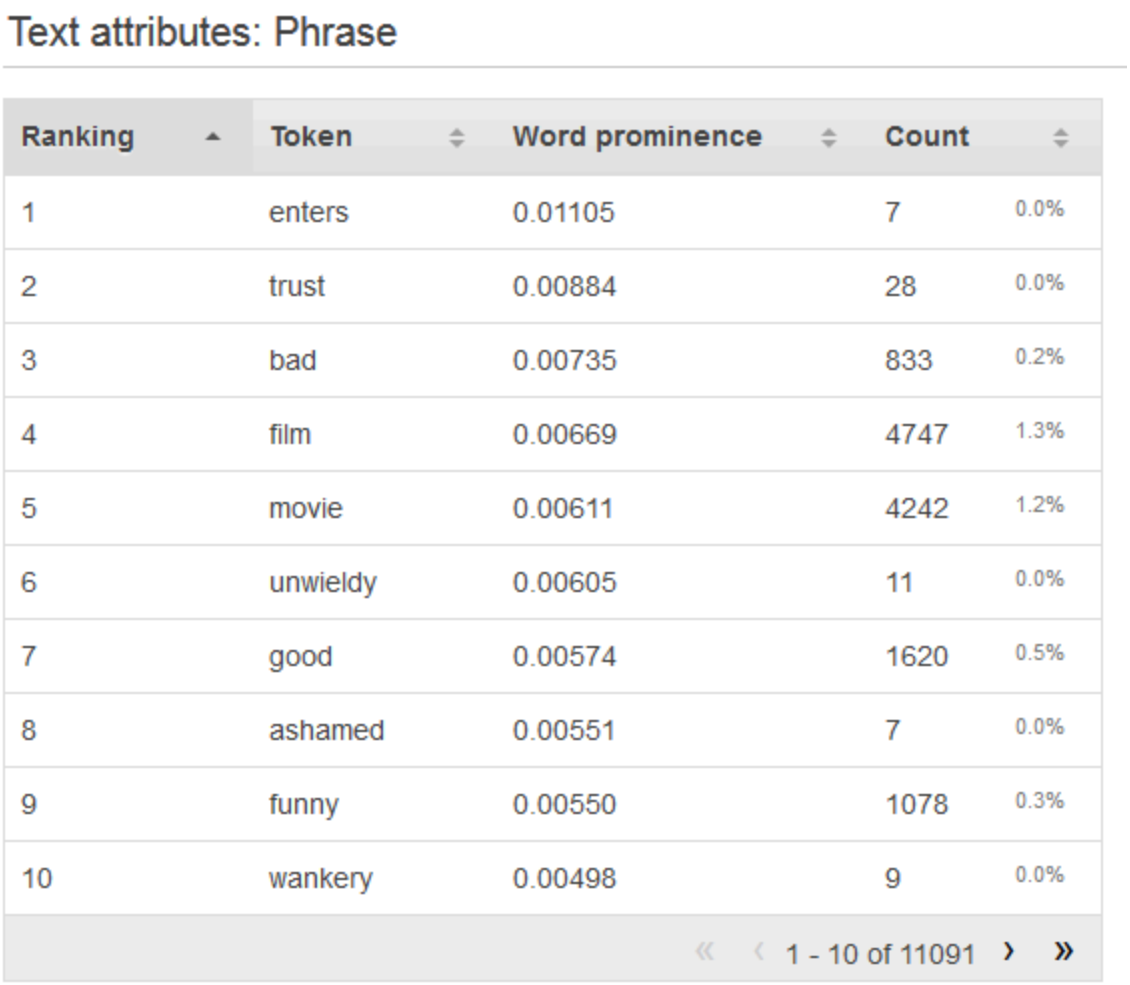
若要檢視數值屬性的分佈，可按一下該屬性的 Preview (預覽) 連結。檢視數值屬性的分佈時，可以選擇量化大小 500、200、100、50 或 20。量化大小愈大，顯示的長條圖數值愈小。此外，量化大小很大的分佈解析度會比較粗糙。反之，若將儲存貯體大小設定為 20，顯示的分佈解析度會相對提升。

此外也會顯示最小值、平均值與最大值，如下列螢幕擷取畫面所示。



了解文字屬性的分佈

若要檢視文字屬性的分佈，可按一下該屬性的 Preview (預覽) 連結。檢視文字屬性的分佈時，會看到下列資訊。



Ranking (排名)

文字字符會依其傳達的資訊量排名，從最多到最少。

Token (字符)

Token (字符) 顯示輸入文字中與統計資料列相關的單文。

Word prominence (單字重要性)

如有目標屬性，文字會依其與目標的相互關聯性排名；因此，相互關聯性最高的文字會最先列出。若資料中沒有目標，則文字會依其熵排名，亦即其可傳達的資訊量。

Count (計數)

Count (計數) 顯示包含此字符之輸入記錄的數量。

Count percentage (計數百分比)

Count Percentage (計數百分比) 顯示字符所在之輸入資料列的百分比。

搭配使用 Amazon S3 與 Amazon ML

Amazon Simple Storage Service (Amazon S3) 是網際網路儲存服務。您可以使用 Amazon S3 隨時從 Web 任何地方存放和擷取任意資料量。Amazon ML 使用 Amazon S3 做為下列任務的主要資料儲存庫：

- 存取輸入檔案以建立資料來源物件，藉此訓練和評估 ML 模型。
- 存取輸入檔來產生批次預測。
- 使用 ML 模型產生批次預測時，將所指定的 S3 儲存貯體輸出到預測檔案。
- 若要將儲存在 Amazon Redshift 或 Amazon Relational Database Service (Amazon RDS) 中的資料複製到 .csv 檔案，並將其上傳至 Amazon S3。

若要讓 Amazon ML 執行這些任務，您必須授予 Amazon ML 存取 Amazon S3 資料的許可。

> **Note**
>
> 您不能將批次預測的檔案輸出到僅接受伺服器端加密的 S3 儲存貯體。在請求中沒有 Deny 標題的情況下，請確定儲存貯體政策中沒有 `s3:PutObject` 動作的 `s3:x-amz-server-side-encryption` 效果，就能確認該政策允許上傳未加密的檔案。如需 S3 伺服器端加密儲存貯體政策的詳細資訊，請參閱《Amazon Simple Storage Service 使用者指南》中的使用伺服器端加密保護資料。

將您的資料上傳至 Amazon S3

您必須將輸入資料上傳至 Amazon Simple Storage Service (Amazon S3)，因為 Amazon ML 從 Amazon S3 位置讀取資料。您可以直接將資料上傳至 Amazon S3 （例如，從您的電腦），或者 Amazon ML 可以將存放在 Amazon Redshift 或 Amazon Relational Database Service (RDS) 中的資料複製到 .csv 檔案中，並將其上傳至 Amazon S3。

如需從 Amazon Redshift 或 Amazon RDS 複製資料的詳細資訊，請分別參閱搭配 Amazon ML 使用 Amazon Redshift 或搭配 Amazon ML 使用 Amazon RDS。

本節的其餘部分說明如何將輸入資料直接從您的電腦上傳至 Amazon S3。在開始閱讀本節程序之前，您必須將資料轉成 .csv 檔案。如需有關如何正確格式化 .csv 檔案以便 Amazon ML 可以使用它的資訊，請參閱了解 Amazon ML 的資料格式。

將您的資料從電腦上傳至 Amazon S3

1. 登入 AWS 管理主控台，然後前往 https://console.aws.amazon.com/s3 開啟 Amazon S3 主控台。
2. 建立儲存貯體或選擇現有的儲存貯體。
    a. 若要建立儲存貯體，請選擇 Create Bucket (建立儲存貯體)。為儲存貯體命名，選擇區域 (您可以選擇任何可用區域)，然後選擇 Create (建立)。如需詳細資訊，請參閱 Amazon 簡易儲存入門指南中的建立儲存貯體相關文章。
    b. 若要使用現有的儲存貯體，請從 All Buckets (所有儲存貯體) 清單中選擇儲存貯體，搜尋該儲存貯體。出現該儲存貯體的名稱後，選取其名稱，然後選擇 Upload (上傳)。
3. 在 Upload (上傳) 對話方塊中，選擇 Add Files (新增檔案)。
4. 導覽到其中包含輸入資料 .csv 檔案的資料夾，然後選擇 Open (開啟)。

許可

若要授予 Amazon ML 存取其中一個 S3 儲存貯體的許可，您必須編輯儲存貯體政策。

如需有關授予 Amazon ML 許可以從 Amazon S3 中的儲存貯體讀取資料的資訊，請參閱授予 Amazon ML 許可以從 Amazon S3 讀取您的資料。

如需授予 Amazon ML 許可以將批次預測結果輸出至 Amazon S3 中儲存貯體的詳細資訊，請參閱授予 Amazon ML 許可以將預測輸出至 Amazon S3。

如需有關管理 Amazon S3 資源存取許可的資訊，請參閱 Amazon S3 開發人員指南。

從 Amazon Redshift 中的資料建立 Amazon ML 資料來源

如果您的資料存放在 Amazon Redshift 中，您可以使用 Amazon Machine Learning (Amazon ML) 主控台中的建立資料來源精靈來建立資料來源物件。當您從 Amazon Redshift 資料建立資料來源時，您可以指定包含資料的叢集，以及擷取資料的 SQL 查詢。Amazon ML 會透過在叢集上叫用

Amazon Redshift `Unload`命令來執行查詢。Amazon ML 會將結果存放在您選擇的 Amazon Simple Storage Service (Amazon S3) 位置，然後使用存放在 Amazon S3 中的資料來建立資料來源。資料來源、Amazon Redshift 叢集和 S3 儲存貯體都必須位於相同的區域。

> **Note**
>
> Amazon ML 不支援從私有 VPCs 中的 Amazon Redshift 叢集建立資料來源。叢集必須有公有 IP 地址。

主題

- 建立資料來源精靈的必要參數
- 使用 Amazon Redshift 資料建立資料來源 （主控台）
- 對 Amazon Redshift 問題進行故障診斷

建立資料來源精靈的必要參數

若要允許 Amazon ML 連線至您的 Amazon Redshift 資料庫並代表您讀取資料，您必須提供下列項目：

- Amazon Redshift `ClusterIdentifier`
- Amazon Redshift 資料庫名稱
- Amazon Redshift 資料庫登入資料 （使用者名稱和密碼）
- Amazon ML Amazon Redshift AWS Identity and Access Management (IAM) 角色
- Amazon Redshift SQL 查詢
- （選用）Amazon ML 結構描述的位置
- Amazon S3 預備位置 (Amazon ML 在建立資料來源之前放置資料的位置）

此外，您需要確保建立 Amazon Redshift 資料來源的 IAM 使用者或角色 （無論是透過主控台或使用`CreateDatasourceFromRedshift`動作） 具有 `iam:PassRole`許可。

Amazon Redshift **`ClusterIdentifier`**

使用此區分大小寫參數，讓 Amazon ML 能夠尋找並連線至您的叢集。您可以從 Amazon Redshift 主控台取得叢集識別符 （名稱）。如需叢集的詳細資訊，請參閱 Amazon Redshift 叢集。

Amazon Redshift 資料庫名稱

使用此參數來告知 Amazon ML Amazon Redshift 叢集中的哪個資料庫包含您要用作資料來源的資料。

Amazon Redshift 資料庫登入資料

使用這些參數來指定將在其中執行安全性查詢的 Amazon Redshift 資料庫使用者的使用者名稱和密碼。

> **Note**
>
> Amazon ML 需要 Amazon Redshift 使用者名稱和密碼才能連線至您的 Amazon Redshift 資料庫。將資料卸載至 Amazon S3 之後，Amazon ML 永遠不會重複使用您的密碼，也不會儲存密碼。

Amazon ML Amazon Redshift 角色

使用此參數指定 Amazon ML 應該用來設定 Amazon Redshift 叢集安全群組的 IAM 角色名稱，以及 Amazon S3 預備位置的儲存貯體政策。

如果您沒有可存取 Amazon Redshift 的 IAM 角色，Amazon ML 可以為您建立角色。當 Amazon ML 建立角色時，它會建立客戶受管政策並將其連接至 IAM 角色。Amazon ML 建立的政策會授予 Amazon ML 僅存取您指定叢集的許可。

如果您已有存取 Amazon Redshift 的 IAM 角色，您可以輸入角色的 ARN，或從下拉式清單中選擇角色。具有 Amazon Redshift 存取權的 IAM 角色會列在下拉式清單的頂端。

IAM 角色必須具有下列內容：

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": {
            "Service": "machinelearning.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
```

```
            "StringEquals": { "aws:SourceAccount": "123456789012" },
           "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:datasource/*" }
        }
    }]
}
```

如需客戶受管政策的詳細資訊，請參閱《IAM 使用者指南》中的客戶受管政策。

Amazon Redshift SQL 查詢

使用此參數指定 Amazon ML 在 Amazon Redshift 資料庫上執行的 SQL SELECT 查詢，以選取您的資料。Amazon ML 使用 Amazon Redshift UNLOAD 動作，將查詢結果安全地複製到 Amazon S3 位置。

> **Note**
>
> 當輸入記錄以隨機順序 （隨機播放） 顯示時，Amazon ML 效果最佳。您可以使用 Amazon Redshift random() 函數，輕鬆隨機播放 Amazon Redshift SQL 查詢的結果。例如，假設這是原始查詢：
>
> ```
> "SELECT col1, col2, … FROM training_table"
> ```
>
> 您可以更新下列這類查詢來內嵌隨機播放：
>
> ```
> "SELECT col1, col2, … FROM training_table ORDER BY random()"
> ```

結構描述位置 (選用)

使用此參數為 Amazon ML 將匯出的 Amazon Redshift 資料指定結構描述的 Amazon S3 路徑。

如果您未提供資料來源的結構描述，Amazon ML 主控台會根據 Amazon Redshift SQL 查詢的資料結構描述自動建立 Amazon ML 結構描述。Amazon ML 結構描述的資料類型少於 Amazon Redshift 結構描述，因此不是one-to-one轉換。Amazon ML 主控台會使用下列轉換機制，將 Amazon Redshift 資料類型轉換為 Amazon ML 資料類型。

Amazon Redshift 資料類型	Amazon Redshift 別名	Amazon ML 資料類型
SMALLINT	INT2	NUMERIC

Amazon Redshift 資料類型	Amazon Redshift 別名	Amazon ML 資料類型
INTEGER	INT、INT4	NUMERIC
BIGINT	INT8	NUMERIC
DECIMAL	NUMERIC	NUMERIC
REAL	FLOAT4	NUMERIC
DOUBLE PRECISION	FLOAT8、FLOAT	NUMERIC
BOOLEAN	BOOL	BINARY
CHAR	CHARACTER、NCHAR、BPCHAR	CATEGORICAL
VARCHAR	CHARACTER VARYING、NVARCHAR、TEXT	TEXT
DATE		TEXT
TIMESTAMP	TIMESTAMP WITHOUT TIME ZONE	TEXT

若要轉換為 Amazon ML `Binary`資料類型，資料中的 Amazon Redshift 布林值必須支援 Amazon ML 二進位值。如果您的布林值資料類型具有不支援的值，Amazon ML 會將它們轉換為最具體的資料類型。例如，如果 Amazon Redshift 布林值為 `0`、 `1`和 `2`，Amazon ML 會將布林值轉換為`Numeric`資料類型。如需所支援二元值的詳細資訊，請參閱使用 AttributeType 欄位。

如果 Amazon ML 無法找出資料類型，則預設為 `Text`。

Amazon ML 轉換結構描述後，您可以在建立資料來源精靈中檢閱和更正指派的 Amazon ML 資料類型，並在 Amazon ML 建立資料來源之前修改結構描述。

Amazon S3 預備位置

使用此參數指定 Amazon ML 存放 Amazon Redshift SQL 查詢結果的 Amazon S3 預備位置名稱。建立資料來源之後，Amazon ML 會使用預備位置中的資料，而不是返回 Amazon Redshift。

> **Note**
>
> 由於 Amazon ML 擔任 Amazon ML Amazon Redshift 角色定義的 IAM 角色，因此 Amazon ML 具有存取指定 Amazon S3 預備位置中任何物件的許可。因此，建議您只將不包含敏感資訊的檔案存放在 Amazon S3 預備位置。例如，如果您的根儲存貯體是 `s3://mybucket/`，我們建議您建立一個位置來僅存放您希望 Amazon ML 存取的檔案，例如 `s3://mybucket/AmazonMLInput/`。

使用 Amazon Redshift 資料建立資料來源 （主控台）

Amazon ML 主控台提供兩種使用 Amazon Redshift 資料建立資料來源的方式。您可以透過完成建立資料來源精靈來建立資料來源，或者，如果您已經從 Amazon Redshift 資料建立資料來源，則可以複製原始資料來源並修改其設定。複製資料來源可讓您輕鬆地建立多個類似的資料來源。

如需使用 API 建立資料來源的資訊，請參閱 CreateDataSourceFromRedshift。

如需下列程序中參數的詳細資訊，請參閱建立資料來源精靈的必要參數。

主題

- 建立資料來源 (主控台)
- 複製資料來源 (主控台)

建立資料來源 (主控台)

若要將資料從 Amazon Redshift 卸載至 Amazon ML 資料來源，請使用建立資料來源精靈。

從 Amazon Redshift 中的資料建立資料來源

1. 開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 在 Amazon ML 儀表板的實體下，選擇建立新...，然後選擇資料來源。
3. 在輸入資料頁面上，選擇 Amazon Redshift。
4. 在 Create Datasource (建立資料來源) 精靈中，Cluster identifier (叢集識別符) 輸入叢集的名稱。
5. 針對資料庫名稱，輸入 Amazon Redshift 資料庫的名稱。
6. Database user name (資料庫使用者名稱) 輸入您的資料庫使用者名稱。
7. Database password (資料庫密碼) 輸入您的資料庫密碼。

8. 針對 IAM role (IAM 角色)，選擇您的 IAM 角色。如果您還沒有，請選擇建立新角色。Amazon ML 會為您建立 IAM Amazon Redshift 角色。
9. 若要測試 Amazon Redshift 設定，請選擇測試存取 (IAM 角色旁）。如果 Amazon ML 無法使用提供的設定連線至 Amazon Redshift，您就無法繼續建立資料來源。如需故障診斷協助，請參閱對錯誤進行故障診斷。
10. 針對 SQL query (SQL 查詢)，輸入您的 SQL 查詢。
11. 針對結構描述位置，選擇您是否希望 Amazon ML 為您建立結構描述。如果您已自行建立結構描述，請輸入結構描述檔案的 Amazon S3 路徑。
12. 對於 Amazon S3 預備位置，輸入您希望 Amazon ML 從 Amazon Redshift 卸載資料的儲存貯體的 Amazon S3 路徑。
13. (選用) 針對 Datasource name (資料來源名稱)，輸入資料來源的名稱。
14. 選擇 Verify (驗證)。Amazon ML 會驗證它是否可以連接到您的 Amazon Redshift 資料庫。
15. 在 Schema (結構描述) 頁面上，檢閱所有屬性的資料類型，並視需要進行更正。
16. 選擇繼續。
17. 若您想要使用此資料來源建立或評估 ML 模型，則針對 Do you plan to use this dataset to create or evaluate an ML model? (您要使用此資料集建立或評估 ML 模型嗎？) 選擇 Yes (是)。如果您選擇 Yes (是)，請選擇目標資料列。如需目標的資訊，請參閱使用 targetAttributeName 欄位。

    若您想要使用此資料來源與已建立的模型來建立預測，請選擇 No (否)。
18. 選擇繼續。
19. 如果您的資料未包含資料列識別符，針對 Does your data contain an identifier? (您的資料包含識別符嗎？) 請選擇 No (否)。

    若您的資料包含資料列識別符，則選擇 Yes (是)。如需資料列識別符的資訊，請參閱使用 rowID 欄位。
20. 選擇檢閱。
21. 在 Review (檢閱) 頁面上檢閱設定，然後選擇 Finish (完成)。

建立資料來源之後，即可使用它來create an ML model。如果您已建立模型，則可以使用資料來源evaluate an ML model或generate predictions。

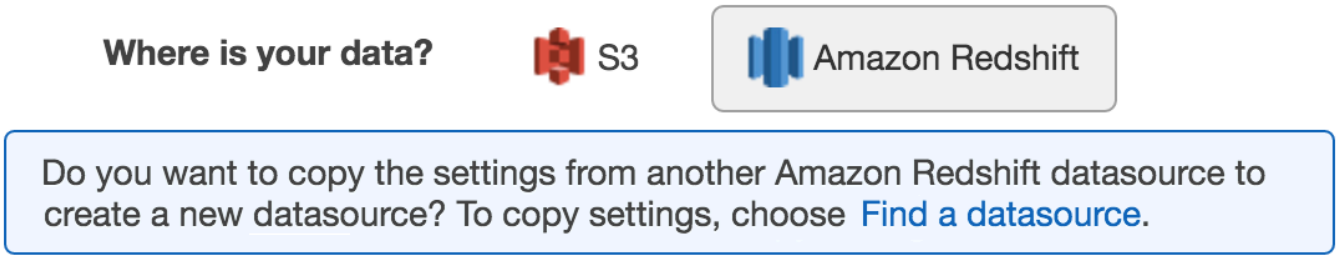
複製資料來源 (主控台)

當您想要建立與現有資料來源類似的資料來源時，您可以使用 Amazon ML 主控台來複製原始資料來源並修改其設定。例如，您可以選擇從現有的資料來源開始，然後修改資料結構描述以更接近您的資

料；變更用於從 Amazon Redshift 卸載資料的 SQL 查詢；或指定不同的 AWS Identity and Access Management (IAM) 使用者來存取 Amazon Redshift 叢集。

複製和修改 Amazon Redshift 資料來源

1. 開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 在 Amazon ML 儀表板的實體下，選擇建立新...，然後選擇資料來源。
3. 在輸入資料頁面上，針對您的資料在哪裡？，選擇 Amazon Redshift。如果您已從 Amazon Redshift 資料建立資料來源，您可以選擇從另一個資料來源複製設定。



如果您尚未從 Amazon Redshift 資料建立資料來源，則不會顯示此選項。

4. 選擇 Find a datasource (尋找資料來源)。
5. 選取您要複製的資料來源，然後選擇複製設定。Amazon ML 會使用原始資料來源的設定自動填入大部分的資料來源設定。它不會複製原始資料來源中的資料庫密碼、結構描述位置或資料來源名稱。
6. 修改您要變更之任何自動填入的設定。例如，如果您想要變更 Amazon ML 從 Amazon Redshift 卸載的資料，請變更 SQL 查詢。
7. Database password (資料庫密碼) 輸入您的資料庫密碼。Amazon ML 不會儲存或重複使用您的密碼，因此您必須一律提供密碼。
8. （選用）對於結構描述位置，Amazon ML 會預先選取我想要 Amazon ML 為您產生建議的結構描述。如果您已建立結構描述，請選擇我想要使用在 Amazon S3 中建立和存放的結構描述，然後在 Amazon S3 中輸入結構描述檔案的路徑。
9. (選用) 針對 Datasource name (資料來源名稱)，輸入資料來源的名稱。否則，Amazon ML 會為您產生新的資料來源名稱。
10. 選擇 Verify (驗證)。Amazon ML 會驗證它是否可以連接到您的 Amazon Redshift 資料庫。
11. （選用）如果 Amazon ML 為您推斷結構描述，請在結構描述頁面上檢閱所有屬性的資料類型，並視需要加以更正。

12. 選擇繼續。
13. 若您想要使用此資料來源建立或評估 ML 模型，則針對 Do you plan to use this dataset to create or evaluate an ML model? (您要使用此資料集建立或評估 ML 模型嗎？) 選擇 Yes (是)。如果您選擇 Yes (是)，請選擇目標資料列。如需目標的資訊，請參閱使用 targetAttributeName 欄位。

    若您想要使用此資料來源與已建立的模型來建立預測，請選擇 No (否)。
14. 選擇繼續。
15. 如果您的資料未包含資料列識別符，針對 Does your data contain an identifier? (您的資料包含識別符嗎？) 請選擇 No (否)。

    如果您的資料包含資料列識別符，請選擇 Yes (是)，然後選取您想要用作識別符的資料列。如需資料列識別符的資訊，請參閱使用 rowID 欄位。
16. 選擇檢閱。
17. 檢閱您的設定，然後選擇 Finish (完成)。

建立資料來源之後，即可使用它來create an ML model。如果您已建立模型，則可以使用資料來源evaluate an ML model或generate predictions。

對 Amazon Redshift 問題進行故障診斷

當您建立 Amazon Redshift 資料來源、ML 模型和評估時，Amazon Machine Learning (Amazon ML) 會在 Amazon ML 主控台中報告 Amazon ML 物件的狀態。如果 Amazon ML 傳回錯誤訊息，請使用下列資訊和資源來疑難排解問題。

如需 Amazon ML 一般問題的答案，請參閱 Amazon Machine Learning FAQs。您也可以在 Amazon Machine Learning 論壇中搜尋答案和張貼問題。

主題

- 對錯誤進行故障診斷
- 聯絡 AWS Support

對錯誤進行故障診斷

角色的格式無效。請提供有效的 IAM 角色。例如，arn:aws:iam::YourAccountID:role/YourRedshiftRole。

原因

IAM 角色的 Amazon Resource Name (ARN) 格式不正確。

解決方案

在 Create Datasource (建立資料來源) 精靈中，更正您角色的 ARN。如需格式化角色 ARNs 的資訊，請參閱《IAM 使用者指南》中的 IAM ARNs。 IAM 角色 ARNs。

角色無效。Amazon ML 無法擔任 <role ARN> IAM 角色。提供有效的 IAM 角色，並使其可供 Amazon ML 存取。

原因

您的角色未設定為允許 Amazon ML 擔任該角色。

解決方案

在 IAM 主控台中，編輯您的角色，使其具有信任政策，允許 Amazon ML 擔任與其連接的角色。

這個 <使用者 ARN > 使用者無權傳遞 <角色 ARN > IAM 角色。

原因

您的 IAM 使用者沒有允許其將角色傳遞至 Amazon ML 的許可政策。

解決方案

將許可政策連接至 IAM 使用者，可讓您將角色傳遞至 Amazon ML。您可以在 IAM 主控台將許可政策附加至您的 IAM 使用者。

不允許跨帳戶傳遞 IAM 角色。IAM 角色必須屬於此帳戶。

原因

您無法傳遞屬於另一個 IAM 帳戶的角色。

解決方案

登入您用來建立角色的 AWS 帳戶。您可以在 IAM 主控台看到您的 IAM 角色。

指定的角色沒有執行操作的許可。提供具有為 Amazon ML 提供必要許可之政策的角色。

原因

您的 IAM 角色沒有執行所請求操作的許可。

解決方案

在 IAM 主控台編輯附加至您角色的許可政策，以提供必要的許可。

Amazon ML 無法使用指定的 IAM 角色在該 Amazon Redshift 叢集上設定安全群組。

原因

您的 IAM 角色沒有設定 Amazon Redshift 安全叢集所需的許可。

解決方案

在 IAM 主控台編輯附加至您角色的許可政策，以提供必要的許可。

當 Amazon ML 嘗試在您的叢集上設定安全群組時發生錯誤。請稍後再試。

原因

當 Amazon ML 嘗試連線到您的 Amazon Redshift 叢集時，遇到問題。

解決方案

確定您在 Create Datasource (建立資料來源) 精靈中提供的 IAM 角色具有所有必要的許可。

叢集 ID 的格式無效。叢集 ID 的開頭必須是字母，且必須僅包含英數字元以及連字號。不能包含兩個連續連字號或以連字號結尾。

原因

您的 Amazon Redshift 叢集 ID 格式不正確。

解決方案

在 Create Datasource (建立資料來源) 精靈中，更正您的叢集 ID，讓它只包含英數字元和連字號，且不包含兩個連續連字號或以連字號結尾。

沒有 <Amazon Redshift 叢集名稱> 叢集，或叢集與 Amazon ML 服務不在同一區域。指定與此 Amazon ML 位於相同區域的叢集。

原因

Amazon ML 找不到您的 Amazon Redshift 叢集，因為它不在您要建立 Amazon ML 資料來源的區域中。

解決方案

確認您的叢集存在於 Amazon Redshift 主控台叢集頁面上、您在 Amazon Redshift 叢集所在的相同區域中建立資料來源，以及建立資料來源精靈中指定的叢集 ID 是否正確。

Amazon ML 無法讀取 Amazon Redshift 叢集中的資料。提供正確的 Amazon Redshift 叢集 ID。

原因

Amazon ML 無法讀取您指定的 Amazon Redshift 叢集中的資料。

解決方案

在建立資料來源精靈中，指定正確的 Amazon Redshift 叢集 ID、確認您正在具有 Amazon Redshift 叢集的相同區域中建立資料來源，以及您的叢集已列在 Amazon Redshift 叢集頁面上。

<Amazon Redshift 叢集名稱> 叢集無法公開存取。

原因

Amazon ML 無法存取您的叢集，因為叢集無法公開存取，也沒有公有 IP 地址。

解決方案

將叢集設為可公開存取，並讓它擁有公有 IP 地址。如需公開存取叢集的資訊，請參閱《Amazon Redshift 管理指南》中的修改叢集。

Amazon ML 無法使用 <Redshift> 叢集狀態。使用 Amazon Redshift 主控台來檢視和解決此叢集狀態問題。叢集狀態必須為「可用」。

原因

Amazon ML 看不到叢集狀態。

解決方案

確定您的叢集可供使用。如需檢查叢集狀態的資訊，請參閱《Amazon Redshift 管理指南》中的取得叢集狀態概觀。如需重新啟動叢集使其可用的詳細資訊，請參閱《Amazon Redshift 管理指南》中的重新啟動叢集。

這個叢集中沒有 <資料庫名稱> 資料庫。請確認資料庫名稱正確或指定另一個叢集和資料庫。

原因

Amazon ML 在指定的叢集中找不到指定的資料庫。

解決方案

確定在 Create Datasource (建立資料來源) 精靈中輸入的資料庫名稱正確，或指定正確的叢集和資料庫名稱。

Amazon ML 無法存取您的資料庫。請提供資料庫使用者 <使用者名稱> 的有效密碼。

原因

您在建立資料來源精靈中提供以允許 Amazon ML 存取 Amazon Redshift 資料庫的密碼不正確。

解決方案

為您的 Amazon Redshift 資料庫使用者提供正確的密碼。

當 Amazon ML 嘗試驗證查詢時發生錯誤。

原因

您的 SQL 查詢有問題。

解決方案

確認您的查詢是有效的 SQL。

執行您的 SQL 查詢時發生錯誤。請確認資料庫名稱和提供的查詢。根本原因：{serverMessage}。

原因

Amazon Redshift 無法執行您的查詢。

解決方案

確認您在 Create Datasource (建立資料來源) 精靈中指定正確的資料庫名稱，而且您的查詢是有效的 SQL。

執行您的 SQL 查詢時發生錯誤。根本原因：{serverMessage}。

原因

Amazon Redshift 找不到指定的資料表。

解決方案

請確認您在建立資料來源精靈中指定的資料表存在於 Amazon Redshift 叢集資料庫中，而且您輸入的叢集 ID、資料庫名稱和 SQL 查詢正確無誤。

聯絡 AWS Support

如果您有 AWS Premium Support，您可在 AWS Support 中心建立技術支援案例。

使用 Amazon RDS 資料庫的資料來建立 Amazon ML 資料來源

Amazon ML 可讓您從存放在 Amazon Relational Database Service (Amazon RDS) 中 MySQL 資料庫的資料建立資料來源物件。當您執行此動作時，Amazon ML 會建立 AWS Data Pipeline 物件，執行您指定的 SQL 查詢，並將輸出放入您選擇的 S3 儲存貯體。Amazon ML 使用該資料來建立資料來源。

> **Note**
>
> Amazon ML 僅支援 VPCs中的 MySQL 資料庫。

您必須先將資料匯出至 Amazon Simple Storage Service (Amazon S3)，Amazon ML 才能讀取輸入資料。您可以使用 API 設定 Amazon ML 來為您執行匯出。(RDS 僅限於 API，不可從主控台使用。)

為了讓 Amazon ML 連線到 Amazon RDS 中的 MySQL 資料庫並代表您讀取資料，您需要提供下列項目：

- RDS 資料庫執行個體識別符
- MySQL 資料庫名稱
- 用來建立、啟用和執行資料管道的 AWS Identity and Access Management (IAM) 角色
- 資料庫使用者登入資料：
  - 使用者名稱
  - 密碼
- AWS Data Pipeline 安全資訊：
  - IAM 資源角色
  - IAM 服務角色
- Amazon RDS 安全資訊：
  - 子網路 ID
  - 安全群組 ID

- SQL 查詢，指定您想要用來建立資料來源的資料
- 用於存放查詢結果的 S3 輸出位置 (儲存貯體)
- (選用) 資料結構描述檔案的位置

此外，您需要確保使用 CreateDataSourceFromRDS 操作建立 Amazon RDS 資料來源的 IAM 使用者或角色具有 `iam:PassRole`許可。如需詳細資訊，請參閱控制 Amazon ML 資源的存取 - 使用 IAM。

主題

- RDS 資料庫執行個體識別符
- MySQL 資料庫名稱
- 資料庫使用者登入資料
- AWS Data Pipeline 安全資訊
- Amazon RDS 安全資訊
- MySQL SQL 查詢
- S3 輸出位置

RDS 資料庫執行個體識別符

RDS 資料庫執行個體識別符是您提供的唯一名稱，可識別 Amazon ML 與 Amazon RDS 互動時應使用的資料庫執行個體。您可以在 Amazon RDS 主控台中找到 RDS 資料庫執行個體識別符。

MySQL 資料庫名稱

MySQL 資料庫名稱指定 RDS 資料庫執行個體中的 MySQL 資料庫名稱。

資料庫使用者登入資料

若要連接到 RDS 資料庫執行個體，您必須提供具有足夠許可之資料庫使用者的使用者名稱和密碼，才能執行您提供的 SQL 查詢。

AWS Data Pipeline 安全資訊

若要啟用安全的 AWS Data Pipeline 存取，您必須提供 IAM 資源角色和 IAM 服務角色的名稱。

EC2 執行個體會擔任資源角色，將資料從 Amazon RDS 複製到 Amazon S3。建立此資源角色最簡單的方式是使用 `DataPipelineDefaultResourceRole` 範本，並將 **`machinelearning.aws.com`**

列為信任的服務。如需範本的詳細資訊，請參閱 AWS Data Pipeline 開發人員指南中的設定 IAM 角色。

如果您建立自己的角色，它必須具有下列內容：

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
    {
        "Effect": "Allow",
        "Principal": {
            "Service": "machinelearning.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "StringEquals": { "aws:SourceAccount": "123456789012" },
            "ArnLike": { "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:datasource/*" }
        }
    }]
}
```

AWS Data Pipeline 擔任服務角色，以監控將資料從 Amazon RDS 複製到 Amazon S3 的進度。建立此資源角色最簡單的方式是使用 DataPipelineDefaultRole 範本，並將 machinelearning.aws.com 列為信任的服務。如需範本的詳細資訊，請參閱 AWS Data Pipeline 開發人員指南中的設定 IAM 角色。

Amazon RDS 安全資訊

若要啟用安全的 Amazon RDS 存取，您需要提供 VPC Subnet ID和 RDS Security Group IDs。您也需要設定由 Subnet ID 參數指向的 VPC 子網路適當輸入規則，並提供具有此許可的安全群組 ID。

MySQL SQL 查詢

MySQL SQL Query 參數指定您想在 MySQL 資料庫上執行的 SQL SELECT 查詢。查詢的結果會複製到您指定的 S3 輸出位置 (儲存貯體)。

> **Note**
>
> 當輸入記錄以隨機順序 (隨機播放) 呈現時，機器學習記數的運作最為良好。您可以使用 `rand()` 函數，輕鬆地隨機播放 MySQL SQL 查詢的結果。例如，假設這是原始查詢：
> "SELECT col1, col2, … FROM training_table"
> 您可以更新查詢來增加隨機播放，如下所示：
> "SELECT col1, col2, … FROM training_table ORDER BY rand()"

S3 輸出位置

`S3 Output Location` 參數指定輸出 MySQL SQL 查詢結果的「預備」Amazon S3 位置的名稱。

> **Note**
>
> 從 Amazon RDS 匯出資料後，您需要確保 Amazon ML 具有從此位置讀取資料的許可。如需設定這些許可的詳細資訊，請參閱「授予 Amazon ML 許可從 Amazon S3 讀取您的資料」。

定型 ML 模型

定型 ML 模型的程序需要提供 ML 演算法 (也就是「學習演算法」)，以及要從中學習的定型資料。「ML 模型」一詞是指由定型程序所建立的模型成品。

定型資料必須包含正確答案，也稱為「目標」或「目標屬性」。學習演算法會在定型資料中尋找將輸入資料屬性對應至目標的模式 (您想要預測的答案)，並輸出擷取這些模式的 ML 模型。

您可以使用 ML 模型，來預測您不知道目標的新資料。例如，假設您想要定型 ML 模型來預測電子郵件是否為垃圾郵件。您會向 Amazon ML 提供訓練資料，其中包含您知道目標的電子郵件 （即指示電子郵件是否為垃圾郵件的標籤）。Amazon ML 會使用此資料來訓練 ML 模型，導致模型嘗試預測新電子郵件是否是垃圾郵件。

如需 ML 模型與 ML 演算法的一般資訊，請參閱機器學習概念。

主題

- ML 模型的類型
- 訓練處理
- 培訓參數
- 建立 ML 模型

ML 模型的類型

Amazon ML 支援三種類型的 ML 模型：二進位分類、多類別分類和迴歸。您應選擇的模型類型，取決於您想要預測的目標類型。

二元分類模型

二元分類問題的 ML 模型預測二元結果 (兩個可能類別其中之一)。為了訓練二進位分類模型，Amazon ML 使用稱為邏輯回歸的產業標準學習演算法。

二元分類問題範例

- 「這個電子郵件是否為垃圾郵件？」
- 「客戶會不會購買此產品？」
- 「這個產品是書籍還是農畜品？」

- 「這個評論是由客戶還是機器人所撰寫？」

多類別分類模型

多類別分類問題的 ML 模型可讓您為多類別產生預測 (預測兩個以上結果的其中一個)。對於訓練多類別模型，Amazon ML 使用稱為多節點羅吉斯迴歸的產業標準學習演算法。

多類別問題範例

- 「這個產品是書籍、電影還是衣服？」
- 「這個電影是浪漫喜劇片、紀錄片還是驚悚片？」
- 「這個客戶最感興趣的產品類別為何？」

回歸模型

回歸問題的 ML 模型預測數值。對於訓練迴歸模型，Amazon ML 使用稱為線性迴歸的產業標準學習演算法。

回歸問題範例

- 「西雅圖明天的溫度為何？」
- 「這個產品會售出多少單位？」
- 「這棟房屋的售價為何？」

訓練處理

若要訓練 ML 模型，您需要指定以下資訊：

- 輸入訓練資料來源
- 包含要預測之目標的資料屬性名稱
- 必要的資料轉換指示
- 用以控制學習演算法的訓練參數

在訓練過程中，Amazon ML 會根據您在訓練資料來源中指定的目標類型，自動為您選擇正確的學習演算法。

培訓參數

一般而言，機器學習演算法接受參數，而參數可以用來控制培訓程序的特定屬性和產生的 ML 模型。在 Amazon Machine Learning 中，這些稱為訓練參數。您可以使用 Amazon ML 主控台、API 或命令列界面 (CLI) 來設定這些參數。如果您未設定任何參數，Amazon ML 將使用已知適用於大量機器學習任務的預設值。

您可以指定下列培訓參數的值：

- 最大模型大小
- 培訓資料的最大通過數目
- 隨機播放類型
- 正規化類型
- 正規化數量

在 Amazon ML 主控台中，預設會設定訓練參數。預設設定可適用於大部分 ML 問題，但您可以選擇其他值來微調效能。某些其他培訓參數 (例如學習速率) 是根據您的資料進行設定。

下列各節提供培訓參數的詳細資訊。

最大模型大小

模型大小上限是 Amazon ML 在 ML 模型訓練期間建立之模式的總大小，以位元組為單位。

根據預設，Amazon ML 會建立 100 MB 模型。您可以指定不同的大小，指示 Amazon ML 建立較小或較大的模型。對於各種可用的大小，請參閱 ML 模型的類型

如果 Amazon ML 找不到足夠的模式來填滿模型大小，則會建立較小的模型。例如，如果您指定 100 MB 的模型大小上限，但 Amazon ML 找到總計只有 50 MB 的模式，則產生的模型將為 50 MB。如果 Amazon ML 找到的模式比符合指定大小的模式更多，它會修剪對學習模型品質影響最小的模式，以強制執行最大截止值。

選擇模型大小可讓您控制模型預測品質與使用成本之間的取捨。較小的模型可能會導致 Amazon ML 移除許多模式，使其符合最大大小限制，進而影響預測品質。另一方面，較大模型的成本高於查詢即時預測。

> **Note**
>
> 如果您使用 ML 模型來產生即時預測，則會引起透過模型大小所決定的少量容量保留費用。如需詳細資訊，請參閱Amazon ML 定價。

較大輸入資料集不一定會產生較大的模型，因為模型存放模式，而不是輸入資料；如果模式少而簡單，產生的模型就會小。具有大量原始屬性（輸入資料欄）或衍生功能 (Amazon ML 資料轉換的輸出）的輸入資料，可能會在訓練過程中找到和儲存更多模式。只要幾次實驗，就能選擇您資料和問題的正確模型大小。Amazon ML 模型訓練日誌（您可以從主控台或透過 API 下載）包含訓練程序期間發生多少模型修剪（如果有的話）的訊息，可讓您預估潛在的hit-to-prediction品質。

資料的最大通過數目

為了獲得最佳結果，Amazon ML 可能需要對資料進行多次傳遞，以探索模式。根據預設，Amazon ML 會進行 10 次傳遞，但您可以透過將數字設定為 100 來變更預設值。Amazon ML 會追蹤模式的品質（模型收斂），並在沒有更多資料點或模式需要探索時自動停止訓練。例如，如果您將傳遞次數設定為 20，但 Amazon ML 發現在 15 次傳遞結束時找不到新模式，則會停止 15 次傳遞的訓練。

一般而言，只有幾個觀察的資料集通常需要資料的更多通過以取得較高的模型品質。較大資料集通常包含許多類似的資料點，因此不需要大量通過。選擇資料的更多資料通過的影響是兩倍：模型培訓需要較長的時間，而且成本較高。

培訓資料的隨機播放類型

在 Amazon ML 中，您必須隨機播放訓練資料。隨機播放會混合資料順序，讓 SGD 演算法不會連續遇到單一資料類型的太多觀察值。例如，如果您要培訓 ML 模型預測產品類型，而且您的培訓資料包含電影、玩具和影片遊戲產品類型，如果您在上傳資料之前依產品類型資料行排序資料，則演算法會依產品類型字母順序查看資料。演算法會先看到所有的電影資料，因此您的 ML 模型開始學習電影的模式。接著，當您的模型遇到玩具的資料，演算法進行的每次更新都會讓演算法更符合玩具產品類型的模型，即使這些更新會降低符合電影的模式。這樣從電影突然切換到玩具類型，會產生不了解如何準確預測產品類型的模式。

您必須隨機播放培訓資料，即使您在將輸入資料來源分割為培訓和評估部分時選擇隨機分割選項也是一樣。隨機分割策略會選擇每個資料來源之資料的隨機子集，但不會變更資料來源中的資料列順序。如需分割資料的詳細資訊，請參閱分割您的資料。

當您使用主控台建立 ML 模型時，Amazon ML 預設會使用虛擬隨機隨機隨機隨機隨機隨機播放技術來隨機播放資料。無論請求的通過次數為何，Amazon ML 在訓練 ML 模型之前只會隨機播放資料一次。如果

您在將資料提供給 Amazon ML 之前隨機播放資料，且不希望 Amazon ML 再次隨機播放資料，您可以將隨機播放類型設定為 none。例如，如果您在將 .csv 檔案中的記錄上傳到 Amazon S3 之前隨機隨機隨機隨機播放，在從 Amazon RDS 建立資料來源時使用 MySQL SQL 查詢中的 `rand()` 函數，或在從 Amazon Redshift 建立資料來源時使用 Amazon Redshift SQL 查詢中的 `random()`函數，請將隨機播放類型設定為 none 不會影響 ML 模型的預測準確性。隨機播放資料一次可減少建立 ML 模型的執行時間和成本。

> **Important**
>
> 當您使用 Amazon ML API 建立 ML 模型時，Amazon ML 預設不會隨機播放您的資料。如果您使用 API 建立 ML 模型，而非主控台，則強烈建議您將 `sgd.shuffleType` 參數設定為 `auto` 來隨機播放資料。

正規化類型和數量

資料包含太多模式時，複雜 ML 模型的預測效能 (有許多輸入屬性的模型) 會降低。隨著模式數量的增加，模型學習意外資料成品的可能性也會增加，而不是真正資料模式。在這種情況下，模型可以很好地學習培訓資料，但無法適當地歸納新資料。這個現象稱為「過度擬合」培訓資料。

正規化會加上極大加權值，協助避免線性模型過度擬合培訓資料範例。L1 正規化將具有極小加權的功能加權推送為零，以減少模型中使用的功能數目。L1 正規化會產生稀疏模型，並降低模型的雜訊量。L2 正規化可產生較小的整體加權值，以在功能之間有高相互關聯性時穩定加權。您可以使用 `Regularization amount` 參數控制 L1 或 L2 正規化量。指定極大 `Regularization amount` 值，可能導致所有功能都具有零加權。

選取並調整最佳正規化值是機器學習研究的主旨。您可能會受益於選擇中等數量的 L2 正規化，這是 Amazon ML 主控台中的預設值。進階使用者可以選擇三種類型的正規化 (none、L1 或 L2) 和數量。如需正規化的詳細資訊，請前往正規化 (數學)。

培訓參數：類型和預設值

下表列出 Amazon ML 訓練參數，以及每個參數的預設值和允許範圍。

培訓參數	類型	預設值	Description
maxMLModelSizeInBytes	Integer	100,000,000 位元組 (100 MiB)	允許範圍：100,000 (100 KiB) 到 2,147,483,648 (2 GiB)

培訓參數	類型	預設值	Description
			根據輸入資料，模型大小可能會影響效能。
sgd.maxPasses	Integer	10	允許範圍：1-100
sgd.shuffleType	String	auto	允許值：`auto` 或 `none`
sgd.l1RegularizationAmount	Double	0 (預設不會使用 L1)	允許範圍：0 到 MAX_DOUBLE 發現 1E-4 與 1E-8 之間的 L1 值可以產生不錯的結果。較大值可能會產生不太有用的模型。 您無法設定 L1 和 L2。您必須選擇其中一項。
sgd.l2RegularizationAmount	Double	1E-6 (L2 預設會與這個數量的正規化搭配使用)	允許範圍：0 到 MAX_DOUBLE 發現 1E-2 與 1E-6 之間的 L2 值可以產生不錯的結果。較大值可能會產生不太有用的模型。 您無法設定 L1 和 L2。您必須選擇其中一項。

建立 ML 模型

建立資料來源之後，您可以開始建立 ML 模型。如果您使用 Amazon Machine Learning 主控台建立模型，您可以選擇使用預設設定，或套用自訂選項來自訂模型。

自訂選項包括：

- 評估設定：您可以選擇讓 Amazon ML 保留部分輸入資料，以評估 ML 模型的預測品質。如需評估的資訊，請參閱評估 ML 模型。
- 配方：配方會告知 Amazon ML 哪些屬性和屬性轉換可用於模型訓練。如需 Amazon ML 配方的資訊，請參閱使用資料配方進行特徵轉換。

- 定型參數：這些參數可控制定型程序與所產生之 ML 模型的特定屬性。如需定型參數的詳細資訊，請參閱定型參數。

若要選取或指定這些設定的值，請在使用 Create ML Model (建立 ML 模型) 精靈時選擇 Custom (自訂) 選項。如果您希望 Amazon ML 套用預設設定，請選擇預設值。

當您建立 ML 模型時，Amazon ML 會根據目標屬性的屬性類型，選取將使用的學習演算法類型。(目標屬性是包含「正確」答案的屬性)。如果您的目標屬性是二進位，Amazon ML 會建立使用邏輯迴歸演算法的二進位分類模型。如果您的目標屬性是分類屬性，Amazon ML 會建立多類別模型，該模型使用多項式邏輯回歸演算法。如果您的目標屬性是數值，Amazon ML 會建立使用線性迴歸演算法的迴歸模型。

主題

- 先決條件
- 使用預設選項建立 ML 模型
- 使用自訂選項建立 ML 模型

先決條件

使用 Amazon ML 主控台建立 ML 模型之前，您需要建立兩個資料來源，一個用於訓練模型，另一個用於評估模型。如果您尚未建立兩個資料來源，請參閱本教學課程中的步驟 2：建立訓練資料來源。

使用預設選項建立 ML 模型

如果您希望 Amazon ML 執行下列動作，請選擇預設選項：

- 分割輸入資料，將前 70% 用於定型，並將剩餘的 30% 用於評估
- 以定型資料來源上所收集的統計資料為基礎的建議配方，其為輸入資料來源的 70%
- 選擇預設定型參數

選擇預設選項

1. 在 Amazon ML 主控台中，選擇 Amazon Machine Learning，然後選擇 ML 模型。
2. 在 ML models (ML 模型) 摘要頁面上，選擇 Create a new ML model (建立新的 ML 模型)。
3. 在 Input data (輸入資料) 頁面上，確定已選取 I already created a datasource pointing to my S3 data (我已建立指向我的 S3 資料的資料來源)。

4. 在資料表中，選擇您的資料來源，然後選擇 Continue (繼續)。
5. 在 ML model settings (ML 模型設定) 頁面上，於 ML model name (ML 模型名稱) 輸入您的 ML 模型的名稱。
6. 針對 Training and evaluation settings (定型與評估設定)，確定已選取 Default (預設)。
7. 針對此評估的名稱，輸入評估的名稱，然後選擇檢閱。Amazon ML 會略過精靈的其餘部分，並帶您前往檢閱頁面。
8. 檢閱您的資料，刪除您從資料來源複製但不想要套用至模型與評估的任何標籤，然後選擇 Finish (完成)。

使用自訂選項建立 ML 模型

自訂 ML 模型可讓您：

- 提供您自己的配方。如需如何提供您自己的配方的資訊，請參閱配方格式參考。
- 選擇定型參數。如需定型參數的詳細資訊，請參閱定型參數。
- 選擇預設 70/30 比例以外的定型/評估分割比例，或提供另一個已準備好評估的資料來源。如需分割策略的資訊，請參閱分割您的資料。

您也可以針對任何這些設定選擇預設值。

如果您已使用預設選項建立模型，並想要改善模型的預測效能，請使用 Custom (自訂) 選項建立具有一些自訂設定的新模型。例如，您可以將更多特徵轉換新增至配方，以增加定型參數中的傳遞數目。

使用自訂選項建立模型

1. 在 Amazon ML 主控台中，選擇 Amazon Machine Learning，然後選擇 ML 模型。
2. 在 ML models (ML 模型) 摘要頁面上，選擇 Create a new ML model (建立新的 ML 模型)。
3. 如果您已建立資料來源，請在 Input data (輸入資料) 頁面上，選擇 I already created a datasource pointing to my S3 data (我已建立指向我的 S3 資料的資料來源)。在資料表中，選擇您的資料來源，然後選擇 Continue (繼續)。

   如果您需要建立資料來源，請選擇 My data is in S3, and I need to create a datasource (我的資料在 S3 中，而且我需要建立資料來源)，然後選擇 Continue (繼續)。系統會將您重新導向至 Create a Datasource (建立資料來源) 精靈。指定您的資料是在 S3 或 Redshift 中，然後選擇 Verify (驗證)。完成建立資料來源的程序。

建立資料來源之後，系統會將您重新導向至 Create ML Model (建立 ML 模型) 精靈的下一個步驟。

4. 在 ML model settings (ML 模型設定) 頁面上，於 ML model name (ML 模型名稱) 輸入您的 ML 模型的名稱。
5. 在 Select training and evaluation settings (選取定型與評估設定) 中選擇 Custom (自訂)，然後選擇 Continue (繼續)。
6. 在 Recipe (配方) 頁面上，您可以customize a recipe。如果您不想自訂配方，Amazon ML 會為您建議配方。選擇繼續。
7. 在 Advanced settings (進階設定) 頁面上，指定 Maximum ML model Size (最大 ML 模型大小)、Maximum number of data passes (最大資料傳遞數目)、Shuffle type for training data (培訓資料的隨機播放類型)、Regularization type (正規化類型) 與 Regularization amount (正規化數量)。如果您未指定這些參數，Amazon ML 會使用預設訓練參數。

   如需這些參數與其預設值的詳細資訊，請參閱培訓參數。

   選擇繼續。
8. 在 Evaluation (評估) 頁面上，指定您是否要立即評估 ML 模型。如果您不想要立即評估 ML 模型，請選擇 Review (檢閱)。

   如果您想要立即評估 ML 模型：

   a. 針對 Name this evaluation (命名此評估) 輸入評估的名稱。
   b. 對於選取評估資料，選擇您是否希望 Amazon ML 保留部分輸入資料以供評估，如果要保留，則選擇如何分割資料來源，或選擇提供不同的資料來源以供評估。
   c. 選擇檢閱。
9. 在 Review (檢閱) 頁面上，編輯您的選擇，刪除您從資料來源複製但不想要套用至模型與評估的任何標籤，然後選擇 Finish (完成)。

建立模型之後，請參閱步驟 4：檢閱 ML 模型的預測效能並設定分數閾值。

機器學習的資料轉換

機器學習模型的良好程度取決於用來訓練模型的資料。良好訓練資料的關鍵特性在於，該資料的提供方式已針對學習與一般化進行最佳化。這項將資料以此最佳格式放在一起的程序，業界稱為「特徵轉換」。

主題

- 特徵轉型的重要性
- 使用資料配方轉換特徵
- 配方格式參考
- 建議配方
- 資料轉換參考
- 資料重新安排

特徵轉型的重要性

假設某個機器學習模型的任務是確定信用卡交易是否為詐騙行為。根據您的應用程式背景知識和資料分析，您可以決定輸入資料應該要包含哪些重要的資料欄位 (或特徵)。例如，交易金額、商家名稱、地址和信用卡擁有者的地址，都是提供給學習程序的重要內容。另一方面，隨機產生的交易 ID 並沒有任何資訊 (若真的是隨機)，而且也沒有用。

一旦您決定要包含哪些欄位，就能改變特徵結構，有利於學習程序。轉型就是要為輸入資料新增背景經驗，讓機器學習模型能從經驗中取經。例如，以下商家地址以字串來表示：

「123 Main Street, Seattle, WA 98101」(華盛頓州 98101 西雅圖市 Main Street 123 號)

地址本身的表示能力有限，只有在與該確切地址有所關聯的學習模式中才有用。不過將地址分為多個組成部分，就能建立像是「地址」(123 Main Street)、「城市」(西雅圖)、「州」(華盛頓州) 和「郵遞區號」(98101) 的額外特徵。現在，學習演算法可以將更多不同的交易分門別類，並探索更廣泛的模式，也許還能找到相較其他郵遞區號，遇到較多詐騙活動的商業郵遞區號。

如需特徵轉換方法和程序的詳細資訊，請參閱機器學習概念。

使用資料配方轉換特徵

使用 Amazon ML 建立 ML 模型之前，有兩種方法可以轉換特徵：您可以在 Amazon ML 中顯示您的輸入資料之前直接轉換這些輸入資料，或者您可以使用 Amazon ML 內建的資料轉換。您可以使用 Amazon ML 配方，這是常見轉換的預先格式化指示。搭配配方，您可執行以下操作：

- 從內建常見機器學習轉換清單中選擇，並將這些轉換套用至個別變數或變數群組
- 選擇要將哪些輸入變數和轉換提供給機器學習過程

使用 Amazon ML 配方可提供多種好處。Amazon ML 會為您執行資料轉換，因此您不需自行實作。此外，這些程序很快速，因為 Amazon ML 會在讀取輸入資料的同時就套用轉換，並提供結果給學習過程，無須儲存結果至磁碟的中繼步驟。

配方格式參考

Amazon ML 配方包含在機器學習程序中轉換資料的指示。配方是使用 JSON 類似的語法所定義，但有一般 JSON 限制以外的其他限制。配方具有下列各區段，必須以這裡顯示的順序出現：

- Groups (群組) 可將多個變數分組，以簡化套用轉換。例如，您可以建立一組所有必須處理網頁任意文字部分 (標題、內文) 的變數，然後一次執行所有這些部分的轉換。
- Assignments (指派) 能夠建立可重複用於處理的中繼具名變數。
- Outputs (輸出) 定義哪些變數將用於學習過程，以及套用至該等變數的轉換 (若有)。

群組

您可以定義一組變數，統一轉換群組內的所有變數，或使用這些變數進行機器學習，而不進行轉換。根據預設，Amazon ML 會為您建立下列群組：

ALL_TEXT、ALL_NUMERIC、ALL_CATEGORICAL、ALL_BINARY – 根據資料來源結構描述中所定義的變數而形成的類型特定群組。

> **Note**
>
> 您無法使用 `ALL_INPUTS` 建立群組。

這些變數不需要定義，即可用於配方的 outputs 區段。您也可以新增或扣除現有群組的變數，或直接新增或扣除變數集合的變數，來建立自訂群組。在下列範例中，我們示範所有這三種方法，以及群組指派語法：

```
"groups": {

"Custom_Group": "group(var1, var2)",
"All_Categorical_plus_one_other": "group(ALL_CATEGORICAL, var2)"

}
```

群組名稱的開頭必須是字母字元，而且長度可以介於 1 到 64 個字元。如果群組名稱的開頭不是字母字元或包含特殊字元 (, ' " \t \r \n ( ) \)，則必須括住名稱，才能包含在配方中。

Assignments (指派)

為求便利性和可讀性，您可以將一或多個轉換指派給中繼變數。例如，如果您有一個名為 email_subject 的文字變數，並且對其套用小寫轉換，則可以將產生的變數命名為 email_subject_lowercase，以輕鬆在配方中的其他位置追蹤到它。指派也可以進行鏈結，讓您依指定的順序套用多次轉換。下列範例使用配方語法來示範單一和鏈結指派：

```
"assignments": {

"email_subject_lowercase": "lowercase(email_subject)",

"email_subject_lowercase_ngram":"ngram(lowercase(email_subject), 2)"

}
```

中繼變數名稱的開頭必須是字母字元，而且長度可以介於 1 到 64 個字元。如果名稱的開頭不是字母字元或包含特殊字元 (, ' " \t \r \n ( ) \)，則必須括住名稱，才能包含在配方中。

輸出

outputs 區段控制將用於學習過程的輸入變數，以及其所套用的轉換。空或不存在的 outputs 區段都是錯誤，因為不會將資料傳遞給學習過程。

最簡單的 outputs 區段只包含預先定義的 ALL_INPUTS 群組，指示 Amazon ML 使用資料來源中定義的所有變數進行學習：

```
"outputs": [
"ALL_INPUTS"
]
```

outputs 區段也可以參照其他預先定義的群組，指示 Amazon ML 使用這些群組中的所有變數：

```
"outputs": [
"ALL_NUMERIC",
"ALL_CATEGORICAL"
]
```

outputs 區段也可以參照自訂群組。在下列範例中，只會將前述範例之群組指派區段中所定義的一個自訂群組用於機器學習。所有其他變數都會予以捨棄：

```
"outputs": [
"All_Categorical_plus_one_other"
]
```

outputs 區段也可以參照指派區段中所定義的變數指派：

```
"outputs": [
"email_subject_lowercase"
]
```

此外，您可以直接在 outputs 區段中定義輸入變數或轉換：

```
"outputs": [
```

```
"var1",

"lowercase(var2)"

]
```

輸出需要明確地指定預期可用於學習過程的所有變數和轉換變數。例如，假設您要在輸出中包含 Cartesian 產品 var1 和 var2。如果您想要同時包含原始變數 var1 和 var2，則需要在 outputs 區段中新增原始變數：

```
"outputs": [

"cartesian(var1,var2)",

"var1",

"var2"

]
```

輸出可以新增註解文字和變數，以包含註解的可讀性：

```
"outputs": [

"quantile_bin(age, 10) //quantile bin age",

"age // explicitly include the original numeric variable along with the
binned version"

]
```

您可以在 outputs 區段內混合使用並比對所有這些方法。

> **Note**
>
> 新增配方時，Amazon ML 主控台不允許評論。

完整配方範例

下列範例參照前述範例中引進的數個內建資料處理器：

```
{

"groups": {

"LONGTEXT": "group_remove(ALL_TEXT, title, subject)",

"SPECIALTEXT": "group(title, subject)",

"BINCAT": "group(ALL_CATEGORICAL, ALL_BINARY)"

},

"assignments": {

"binned_age" : "quantile_bin(age,30)",

"country_gender_interaction" : "cartesian(country, gender)"

},

"outputs": [

"lowercase(no_punct(LONGTEXT))",

"ngram(lowercase(no_punct(SPECIALTEXT)),3)",

"quantile_bin(hours-per-week, 10)",

"hours-per-week // explicitly include the original numeric variable
along with the binned version",

"cartesian(binned_age, quantile_bin(hours-per-week,10)) // this one is
critical",

"country_gender_interaction",

"BINCAT"
```

```
]

}
```

建議配方

當您在 Amazon ML 中建立新的資料來源，並對該資料來源計算統計資料時，Amazon ML 也會建立建議的配方，該配方可用來從資料來源建立新的 ML 模型。建議的資料來源以資料和存在於資料中的目標屬性為基礎，為您建立和微調 ML 模型提供有用的起點。

若要在 Amazon ML 主控台上使用建議的配方，請從建立新下拉式清單中選擇資料來源或資料來源和 ML 模型。對於 ML 模型設定，您可以在建立 ML 模型精靈的 ML 模型設定步驟中選擇預設或自訂訓練和評估設定。如果您選取 Default (預設) 選項，Amazon ML 會自動使用建議的配方。如果您選取 Custom (自訂) 選項，下一個步驟中的配方編輯器會顯示建議的配方，您就可以視需要加以確認或修改。

> **Note**
>
> Amazon ML 可讓您在統計資料運算完成前，先建立資料來源，然後立即用其來建立 ML 模型。在這種情況下，您在 Custom (自訂) 選項中將看不到建議的配方，但仍可繼續通過該步驟，讓 Amazon ML 使用預設配方訓練模型。

若要以 Amazon ML API 使用建議的配方，您可以在 Recipe 和 RecipeUri API 兩個參數中傳送空白字串。您無法使用 Amazon ML API 擷取建議的配方。

資料轉換參考

主題

- N 元語法轉換
- 正交稀疏二元 (OSB) 轉換
- 小寫轉換
- 移除標點符號轉換
- 四分位數分箱轉換
- 標準化轉型

- 笛卡兒乘積轉換

N 元語法轉換

N 元語法轉換採用文字變數做為輸入，並產生當滑動 n 個單詞 (使用者可設定的) 視窗時的對應字串，以便在過程中產生輸出。例如，假設有一個文字字串："I really enjoyed reading this book" (我真的很喜歡閱讀這本書)。

指定 n 元語法轉換並使用視窗大小 = 1，只會提供該字串中的所有個別單詞：

```
{"I", "really", "enjoyed", "reading", "this", "book"}
```

將 n 元語法轉換指定為視窗大小 = 2，則會提供所有兩個字組合以及一個字組合：

```
{"I really", "really enjoyed", "enjoyed reading", "reading this", "this
book", "I", "really", "enjoyed", "reading", "this", "book"}
```

將 n 元語法轉換指定為視窗大小 = 3 則會增加三個字組合，產生下列項目：

```
{"I really enjoyed", "really enjoyed reading", "enjoyed reading this",
"reading this book", "I really", "really enjoyed", "enjoyed reading",
"reading this", "this book", "I", "really", "enjoyed", "reading",
"this", "book"}
```

您可以請求 2-10 個單詞大小範圍的 n 元語法。所有在資料結構描述中標示為文字類型的輸入，都會隱含產生大小為 1 的 n 元語法，因此您不需要特定要求。最後，請記住，n 元語法是根據空白字元來中斷輸入資料而產生。這表示，舉例來說，標點符號字元會被視為單詞符記的一部分：為字串 "red, green, blue" (「紅色、綠色、藍色」) 產生視窗為 2 的 n 元語法會得到：{"red,", "green,", "blue,", "red, green", "green, blue"}。如果這不是您要的結果，您可以使用標點符號移除處理器 (本文稍後說明) 來移除標點符號。

若要對變數 var1 計算視窗大小為 3 的 n 元語法：

```
"ngram(var1, 3)"
```

正交稀疏二元 (OSB) 轉換

OSB 轉換旨在協助文字字串分析，且是二元語法 (視窗大小為 2 的 n 元語法) 的替代方法。OSB 透過在文字上滑動視窗大小 n 來產生，然後輸出每對字詞，其中包含視窗中的第一個字。

若要建置每個 OSB，其構成單字使用「_」(底線) 字元連結，並在 OSB 中新增其他底線來指出每個略過的符記。因此，OSB 不只編碼視窗中看到的符記，也會指出相同視窗中略過的符記數。

舉例來說，假設有個字串 "The quick brown fox jumps over the lazy dog" (敏捷的棕色狐狸跳過了一隻懶狗)，且 OSB 的大小為 4。以下範例中顯示六個四字視窗，以及從字串尾端產生的最後兩個較短視窗，以及為每個視窗產生的 OSB：

視窗, {產生的 OSB}

```
"The quick brown fox", {The_quick, The__brown, The___fox}

"quick brown fox jumps", {quick_brown, quick__fox, quick___jumps}

"brown fox jumps over", {brown_fox, brown__jumps, brown___over}

"fox jumps over the", {fox_jumps, fox__over, fox___the}

"jumps over the lazy", {jumps_over, jumps__the, jumps___lazy}

"over the lazy dog", {over_the, over__lazy, over___dog}

"the lazy dog", {the_lazy, the__dog}

"lazy dog", {lazy_dog}
```

正交稀疏二元語法在某些情況下，可能是比 n 元語法運作更良好的替代選擇。如果您的資料擁有大型文字欄位 (10 個或更多單詞)，請實驗看看哪個的運作方式更佳。請注意，大型文字欄位的內容可能因情況而異。不過，對於較大的文字欄位，會憑經驗顯示 OSB，因為特殊「略過」符號 (底線) 而唯一代表該文字。

您可以在輸入文字變數上，要求視窗大小為 2 到 10 個單字的 OSB 轉換。

若要針對變數 var1 計算 OSB，並使用視窗大小 5：

"osb(var1, 5)"

小寫轉換

小寫轉換處理器會將輸入文字轉換為小寫。例如，提供輸入 "The Quick Brown Fox Jumps Over the Lazy Dog"，處理器將輸出 "the quick brown fox jumps over the lazy dog"。

若要套用小寫轉換到變數 var1：

"lowercase(var1)"

移除標點符號轉換

Amazon ML 會在資料結構描述中根據空格隱含分割標記為文字的輸入。因此，字串中的標點符號會變成鄰接文字符記，或變成單獨的福記，這取決其周圍的空格。如果您不想要這種結果，可使用標點符號移除器轉換來從產生的特徵中移除標點符號。例如，提供字串 "Welcome to AML - please fasten your seat-belts!" (歡迎來到 AML - 請繫好安全帶！)，會隱含產生下列符記組：

```
{"Welcome", "to", "Amazon", "ML", "-", "please", "fasten", "your", "seat-belts!"}
```

套用標點符號移除處理器到這個字串，會產生此組結果：

```
{"Welcome", "to", "Amazon", "ML", "please", "fasten", "your", "seat-belts"}
```

請注意，只會移除字首和尾碼標點符號。出現在符記中間的標點符號，例如 "seat-belts" 中的連字號，並不會移除。

若要套用標點符號移除到變數 var1：

"no_punct(var1)"

四分位數分箱轉換

四分位數分箱處理器採用兩個輸入：一個數值變數和一個稱為「分箱數」的參數，然後輸出類別變數。其目的是將觀察值分組在一起，來探索變數分佈中的非線性狀況。

在許多情況下，數值變數與目標之間的關係並非線性 (數值變數值不會隨著目標單純地增加或減少)。在這種情況下，將數值特徵分箱至可代表不同數值特徵範圍的類別特徵，可能會很有用。接著可以為每個類別特徵值 (分箱) 建立模型，分別擁有自己與目標的線性關係。例如，假設您知道連續數值特徵 account_age 與購買書籍的可能性沒有線性相關。您可將年齡分箱至不同的類別特徵，然後可以更精確擷取與目標的關係。

四分位數分箱處理器可用來指示 Amazon ML 根據年齡變數的所有可用輸入值分佈來建立 n 個相等大小的分箱，然後將每個數字替換成包含分箱的文字符記。數值變數的最佳分箱數取決於變數特徵以及其與目標的關係，而這最好透過實驗確定。Amazon ML 會根據建議的配方中的統計資料，對數值特徵建議最佳的分箱數。

您可以請求 5 到 1000 個四分位數分箱，用來計算任何數值輸入變數。

以下範例說明如何運算和使用 50 個分箱來代替數值變數 var1：

"quantile_bin(var1, 50)"

標準化轉型

標準化轉型會將數值變數標準化，使其平均值為零、而方差為一。如果數值變數之間有非常大的差異範圍，由於無論特徵對於目標是否資訊量足夠，具有最高量值的變數都會主導 ML 模型，因此將數值變數標準化有助於學習過程。

若要將此轉換套用到數值變數 var1，請新增此行至配方：

normalize(var1)

這個轉換器也可以採用使用者定義的數值變數群組或預先定義的所有數值變數群組 (ALL_NUMERIC) 做為輸入：

normalize(ALL_NUMERIC)

注意

您「不一定」要使用標準化處理器來處理數值變數。

笛卡兒乘積轉換

笛卡兒轉換會產生兩個或多個文字或類別輸入變數的置換。此轉換是用於猜測變數之間應有相互影響時。例如，假設在「教學課程：使用 Amazon ML 預測對行銷優惠的回應」中使用銀行行銷資料集。使用此資料集，我們想要根據經濟和人口統計資訊，預測人們是否會主動回應銀行促銷。我們猜想人們的工作類型應該有點重要 (例如在特定領域受雇，和有餘錢可使用之間應有關聯)，另外最高教育程度也很重要。我們也可能直覺這兩個變數的相互影響中具備強烈信號，例如，促銷特別針對大學畢業的企業家客戶而打造。

笛卡兒乘積轉換採用類別變數或文字做為輸入，並產生新特徵，以擷取這些輸入變數之間的相互影響。值得注意的是，它將為每個訓練範例建立特徵組合，並將其新增做為獨立的特徵。例如，假設我們的簡化輸入資料列如下所示：

目標, 教育程度, 工作

0, 大學學位, 技術人員

0, 高中, 服務業

1, 大學學位, 行政管理

如果我們指定將笛卡兒轉換套用至類別變數 [教育程度] 和 [工作] 欄位，產生的特徵 [教育程度_工作_相互影響] 看起來會像這樣：

目標, 教育程度_工作_相互影響

0, 大學學位_技術人員

0, 高中_服務業

1, 大學學位_行政管理

如果將笛卡兒轉換運作在符記序列甚至會更加強大，因為其中一個引數是會隱含或明確分割成符記的文字變數。例如，假設有一項任務是將書籍分類為是教科書和不是教科書。直覺上，我們可能認為書本的標題可以說明是否為教科書 (教科書標題中會更頻繁出現特定單詞)，另外我們也可能認為書本的裝訂版能用來進行預測 (教科書更可能是精裝書)，但標題中某些單詞和裝訂版的組合才是最具預測性的。對於真實世界的範例，下表顯示將笛卡兒處理器套用至輸入變數 [裝訂版] 和 [標題] 的結果：

教科書	Title	裝訂版	no_punct(Title) 和裝訂版的笛卡兒乘積
1	Economics : Principles, Problems, Policies (經濟學：原則、問題、政策)	精裝	{"Economics_Hardcover", "Principles_Hardcover", "Problems_Hardcover", "Policies_Hardcover"}
0	愛上經濟：一個談經濟學的愛情故事 (The Invisible Heart: An Economic Romance)	平裝	{"The_Softcover", "Invisible_Softcover", "Heart_Softcover", "An_Softcover", "Economics_Softcover", "Romance_Softcover"}

教科書	Title	裝訂版	no_punct(Title) 和裝訂版的笛卡兒乘積
0	Fun With Problems (問題的樂趣)	平裝	{"Fun_Softcover", "With_Softcover", "Problems_Softcover"}

以下範例說明如何套用笛卡兒轉換器至 var1 和 var2：

cartesian(var1, var2)

資料重新安排

資料重新安排功能可讓您建立只根據所指向輸入資料一部分的資料來源。例如，當您使用 Amazon ML 主控台中的建立 ML 模型精靈建立 ML 模型，並選擇預設評估選項時，Amazon ML 會自動保留 30% 的資料用於 ML 模型評估，並使用其他 70% 進行訓練。此功能由 Amazon ML 的資料重新排列功能啟用。

如果您使用 Amazon ML API 來建立資料來源，您可以指定新資料來源將以哪個部分為基礎的輸入資料。做法是將 `DataRearrangement` 參數中的指示傳遞給 `CreateDataSourceFromS3`、`CreateDataSourceFromRedshift` 或 `CreateDataSourceFromRDS` API。DataRearrangement 字串的內容是包含資料開始和結束位置的 JSON 字串 (以百分比表示)、補充旗標和分割策略。例如，下列 DataRearrangement 字串指定資料的前 70% 將用來建立資料來源：

```
{
    "splitting": {
        "percentBegin": 0,
        "percentEnd": 70,
        "complement": false,
        "strategy": "sequential"
    }
}
```

DataRearrangement 參數

若要變更 Amazon ML 建立資料來源的方式，請使用下列參數。

PercentBegin (選用)

使用 percentBegin 指出資料來源的資料開始位置。如果您未包含 percentBegin和 percentEnd，Amazon ML 會在建立資料來源時包含所有資料。

有效值為 0 到 100 (含)。

PercentEnd (選用)

使用 percentEnd 指出資料來源的資料結束位置。如果您未包含 percentBegin和 percentEnd，Amazon ML 會在建立資料來源時包含所有資料。

有效值為 0 到 100 (含)。

Complement (選用)

complement 參數會指示 Amazon ML 使用 到 範圍中未包含的資料percentBeginpercentEnd來建立資料來源。如果您需要建立補充資料來源來進行培訓和評估，則 complement 參數十分有用。若要建立補充資料來源，請使用相同的 percentBegin 和 percentEnd 值，以及 complement 參數。

例如，下列兩個資料來源不共用任何資料，而且可以用來培訓和評估模型。第一個資料來源有 25% 的資料，而第二個資料來源有 75% 的資料。

評估的資料來源：

```
{
    "splitting":{
        "percentBegin":0,
        "percentEnd":25
    }
}
```

培訓的資料來源：

```
{
    "splitting":{
        "percentBegin":0,
        "percentEnd":25,
        "complement":"true"
    }
}
```

有效值為 `true` 和 `false`。

Strategy (選用)

若要變更 Amazon ML 分割資料來源資料的方式，請使用 `strategy` 參數。

`strategy` 參數的預設值為 `sequential`，這表示 Amazon ML 會依記錄出現在輸入資料中的順序，取得資料來源 `percentBegin`和 `percentEnd` 參數之間的所有資料記錄

下列兩行 `DataRearrangement` 是循序排序培訓和評估資料來源範例：

評估的資料來源：`{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential"}}`

培訓的資料來源：`{"splitting":{"percentBegin":70, "percentEnd":100, "strategy":"sequential", "complement":"true"}}`

若要透過隨機選取資料來建立資料來源，請將 `strategy` 參數設定為 `random`，並提供一個字串，用作進行隨機資料分割的種子值 (例如，您可以使用資料的 S3 路徑作為隨機種子字串)。如果您選擇隨機分割策略，Amazon ML 會為每個資料列指派虛擬隨機號碼，然後選取指派號碼介於 `percentBegin`和 之間的資料列`percentEnd`。虛擬亂數是使用位元組位移作為種子進行指派，因此變更資料會導致不同的分割。保留任何現有排序。隨機分割策略確保以類似的方式分佈培訓和評估資料中的變數。它適用於輸入資料可能有隱含排序順序時，這可能會導致包含非類似資料記錄的培訓和評估資料來源。

下列兩行 `DataRearrangement` 是非循序排序培訓和評估資料來源範例：

評估的資料來源：

```
{
    "splitting":{
        "percentBegin":70,
        "percentEnd":100,
        "strategy":"random",
        "strategyParams": {
            "randomSeed":"RANDOMSEED"
        }
    }
}
```

培訓的資料來源：

```
{
    "splitting":{
        "percentBegin":70,
        "percentEnd":100,
        "strategy":"random",
        "strategyParams": {
            "randomSeed":"RANDOMSEED"
        }
        "complement":"true"
    }
}
```

有效值為 `sequential` 和 `random`。

(選用) Strategy:RandomSeed

Amazon ML 使用 randomSeed 來分割資料。API 的預設種子是空字串。若要指定隨機分割策略的種子，請傳入字串。如需隨機種子的詳細資訊，請參閱《Amazon Machine Learning 開發人員指南隨機分割資料》中的 。

如需示範如何搭配 Amazon ML 使用跨驗證的範例程式碼，請前往 Github Machine Learning Samples。

評估 ML 模型

您應該持續「評估模型」以判斷其能否勝任預測新資料和未來資料的預測任務。由於未來的執行個體有不明目標值，您需要檢查 ML 模型對於您已知目標答案之資料的準確性指標，並使用此評估做為預測未來資料準確性的代理。

若要正確地評估模型，請從訓練資料來源取出已標示為目標 (基本事實) 的資料樣本。使用用於訓練的相同資料來評估 ML 模型的預測準確性並不適合，因為這樣是獎勵能「死記」訓練資料的模型，而非能從資料加以一般化的模型。您完成訓練 ML 模型後，您傳送已知目標值的保留觀察給模型。然後，比較 ML 模型傳回的預測結果和已知目標數值。最後，您運算摘要指標，告訴您預測和真實值的相符程度。

在 Amazon ML 中，您可以透過建立評估來評估 ML 模型。若要建立 ML 模型的評估，您需要一個您想要評估的 ML 模型，您也需要未使用於訓練的標記資料。首先，使用保留資料建立 Amazon ML 資料來源來建立資料來源以供評估。用於評估的資料必須和用於訓練的資料具備相同的結構描述，並包含目標變數的實際值。

如果您的所有資料都位於單一檔案或目錄中，您可以使用 Amazon ML 主控台來分割資料。Create ML model (建立 ML 模型) 精靈中的預設路徑會分割輸入資料來源，並使用前 70% 做為訓練資料來源，其餘 30% 做為評估資料來源。Create ML model (建立 ML 模型) 精靈中的 Custom (自訂) 選項也可供您自訂分割比，您可以在此處隨機選取 70% 的樣本用於訓練，並將其餘 30% 用於評估。為了進一步指定自訂分割比，請使用建立資料來源 API 中的資料重新安排字串。擁有評估資料來源和 ML 模型後，您可以建立評估並檢閱評估的結果。

主題

- ML 模型深入分析
- 二元模型的深入解析
- 多類別模型深入分析
- 迴歸模型的深入解析
- 防止過度擬合
- 交叉驗證
- 評估提醒

ML 模型深入分析

當您評估 ML 模型時，Amazon ML 會提供產業標準指標和許多深入分析，用以檢閱您模型的預測準確性。在 Amazon ML 中，評估結果包含下列項目：

- 預測準確性指標，用以報告模型的整體成功情況
- 視覺化，用以協助探索預測準確性指標外的模型準確性
- 能夠檢閱設定分數閾值的影響 (僅適用於二元分類)
- 條件的警示，用以檢查評估的有效性

指標和視覺化的選擇取決於您評估的 ML 模型類型。請務必檢閱這些視覺化，以決定您模型的效能是否足以滿足您的業務需求。

二元模型的深入解析

解譯預測

許多二元分類演算法的實際輸出是一種預測「分數」。此分數指出系統確定指定的觀察屬於正確類別 (真實目標值為 1)。Amazon ML 中的二進位分類模型會輸出 0 到 1 的分數。此分數的取用者可以決定應將觀察分類為 1 或 0。您可以挑選分類閾值或「分界值」做為分數的比較依據，從而解譯分數。所有分數高於此分界值的觀察，都會將其目標預測為 1；所有分數低於此分界值的觀察，都會將其目標預測為 0。

在 Amazon ML 中，預設分數截止值為 0.5。您可以依據您的業務需求，選擇更新此分界值。您可以利用主控台中的視覺效果，了解分界值選擇對於您應用程式的影響。

衡量 ML 模型準確性

Amazon ML 為稱為「（接收者操作特性）曲線下「面積」(AUC) 的二進位分類模型提供業界標準的準確性指標。AUC 會測量模型在預測較高分數之正確範例與錯誤範例上的能力，並將兩者相比較。因為這無關乎分界分數，所以您無須選取閾值，就能從 AUC 指標得知模型的預測正確性。

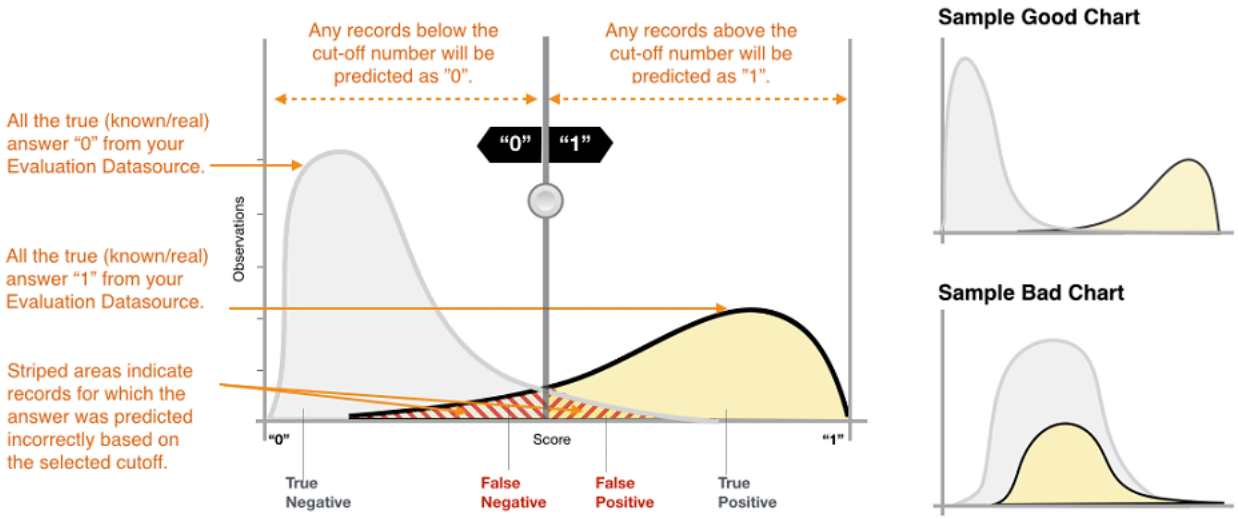
AUC 指標會傳回介於 0 至 1 的小數值。接近 1 的 AUC 值代表 ML 模型準確性很高。值接近 0.5 表示 ML 模型與隨機猜測差異不大。接近 0 的值並不常見，通常表示資料有問題。基本上，若 AUC 接近 0，表示 ML 模型已學會正確模式，但使用這些模式進行的預測會與現實相反 (將 '0' 預測為 '1'，反之亦然)。如需 AUC 的詳細資訊，請參閱 Wikipedia 上的接收者操作特性頁面。

二元模型的基準 AUC 指標為 0.5，這是 ML 假設模型的值，會隨機預測 1 或 0 的答案。您二元 ML 模型的執行效果應優於此值，此模型才有價值。

使用效能視覺化

若要探索 ML 模型的準確性，您可以在 Amazon ML 主控台的評估頁面上檢閱圖形。此頁面顯示兩個色階分佈圖：a) 評估資料中真實正確 (目標為 1) 之分數的色階分佈圖，以及 b) 評估資料中真實錯誤 (目標為 0) 之分數的色階分佈圖。

具備良好正確性預測的 ML 模型，會將高分預測真實的 1，並將低分預測為真實的 0。完美的模型在 X 軸兩端各有一個色階分佈圖，分別顯示所有得到高分的真實正確，以及所有得到低分的真實錯誤。但 ML 模型會犯錯，而且常見圖表的這兩個色階分佈圖會在特定分數重疊。效能極差的模型無法區分正確與錯誤的類別，而且這兩個類別的色階分佈圖大部分會重疊。



透過視覺效果，您可以得出落入兩種正確預測類型與兩種錯誤預測類型的預測數量。

正確預測

- 真陽性 (TP)：Amazon ML 預測值為 1，而真值為 1。
- True negative (TN)：Amazon ML 預測值為 0，而 true 值為 0。

錯誤預測

- 偽陽性 (FP)：Amazon ML 預測值為 1，但真實值為 0。
- False negative (FN)：Amazon ML 預測值為 0，但 true 值為 1。

> **Note**
>
> TP、TN、FP 與 FN 的數量取決於選取的分數閾值，而最佳化其中任何一個數量意味著其他數量也會受到影響。TP 數量高通常會導致 FP 的數量高及 TN 數量低。

調整分界分數

ML 模型的運作方式是先產生數值預測分數，然後再套用分界值，將這些分數轉換成二元的 0/1 標籤。只要變更分界分數，就能在模型犯錯時調整其行為。在 Amazon ML 主控台的評估頁面上，您可以檢閱各種分數分界點的影響，也可以儲存您想要用於模型的分數分界點。

當您調整分界分數的閾值時，請觀察這兩種誤差類型之間的交互影響。將分界值向左移會得到比較多真正的正確，但代價是錯誤的錯誤數量會增加。將此值向右移會得到比較少錯誤的錯誤，但代價是會漏失一些真正的正確。您可以為您自己的預測應用程式選取適當的分界分數，決定比較能容忍的誤差種類。

檢閱進階指標

Amazon ML 提供下列其他指標來測量 ML 模型的預測準確性：準確性、精確度、召回和誤報率。

準確性

「正確性」(ACC) 會測量正確預測的分數。範圍介於 0 至 1 之間。值越大，表示預測準確性越高：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

精確度

「精確度」會測量真實正確占這些預測為正確之範例的分數。範圍介於 0 至 1 之間。值越大，表示預測準確性越高：

$$Precision = \frac{TP}{TP + FP}$$

取回

「取回」會測量被預測為正確的真實正確分數。範圍介於 0 至 1 之間。值越大，表示預測準確性越高：

$$Recall = \frac{TP}{TP + FN}$$

錯誤的正確率

「錯誤的正確率」(FPR) 會測量被預測為正確的誤報率或真實錯誤分數。範圍介於 0 至 1 之間。值愈小表示預測正確性愈佳：

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

根據您的業務問題，您可能對特定指標子集執行效果良好的模型更感興趣。舉例來說，兩個商務應用程式的 ML 模型在需求上可能截然不同：

- 其中一個應用程式可能需要相當確定正確預測實際上為正確 (高精確度)，並能容忍將一些正確的範例分類為錯誤 (中度取回)。
- 另一個應用程式可能只需要盡可能地正確預測正確的範例 (高度取回)，而且能夠接受將一些錯誤的範例不正確地分類為正確 (中精確度)。

Amazon ML 可讓您選擇與上述任何進階指標的特定值對應的分數分界。此外，它也會顯示最佳化任何一個指標所帶來的相互影響。例如，若您選取的分界值對應到了高精確度，通常帶來的相互影響就是較低的取回數量。

> **Note**
>
> 您必須儲存截止分數，才能有效地分類 ML 模型未來所做的任何預測。

多類別模型深入分析

解譯預測

多類別分類演算法的實際輸出是一組預測「分數」。分數指出模型對於特定觀察屬於每個類別的確定程度。與二元分類問題不同的是，您不需要選擇分數截止值來進行預測。預測答案是具有最高預測分數的類別 (例如，標籤)。

衡量 ML 模型準確性

多類別中使用的典型指標在平均所有類別後所使用的指標和二元分類案例中使用的指標相同。在 Amazon ML 中，巨集平均 F1 分數用於評估多類別指標的預測準確性。

巨集平均 F1 分數

F1 分數是一個二元分類指標，同時參考二元指標精確度和取回。這是精確度和取回之間的調和平均數。範圍介於 0 至 1 之間。值越大，表示預測準確性越高：

$$F1\ score = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

巨集平均 F1 分數是多類別案例中所有類別的 F1 分數未加權平均。它不將類別在評估資料集中的出現頻率列入考慮。值越大，表示預測準確性越高。以下範例顯示評估資料來源中的 K 類別：

$$Macro\ average\ F1\ score = \frac{1}{K}\sum_{k=1}^{K} F1\ score\ for\ class\ k$$

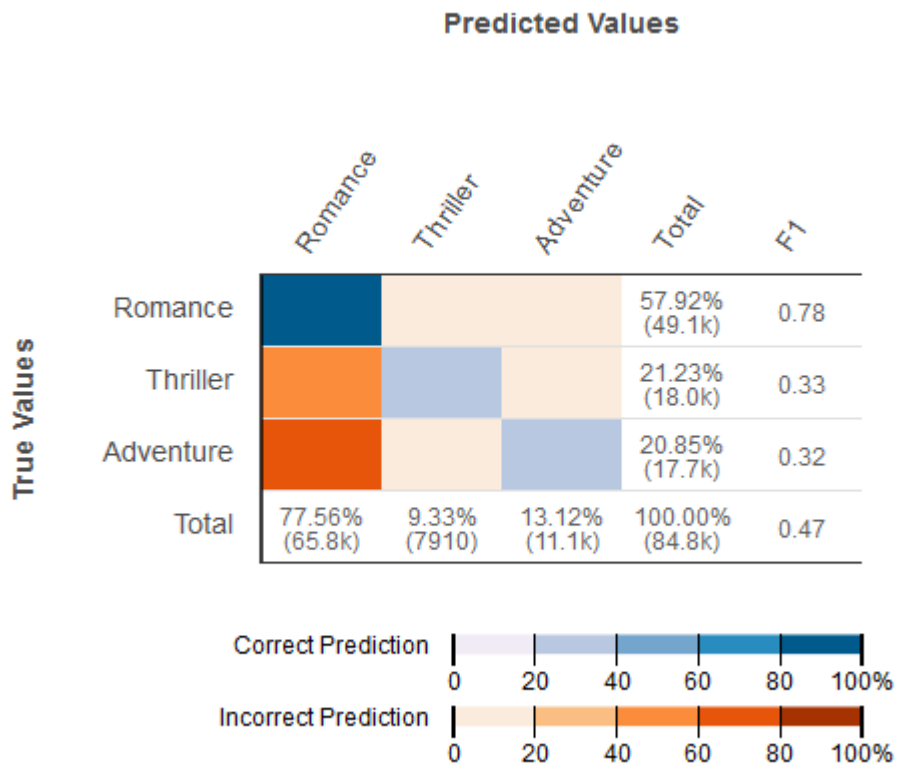
基準巨集平均 F1 分數

Amazon ML 為多類別模型提供基準指標。這是假設多類別模型的巨集平均 F1 分數，一律預測最頻繁的類別做為答案。例如，如果您要預測電影類型而您的訓練資料中最常見的類型是愛情片，則基準模型會一律將類型預測為愛情片。您應根據此基準來比較您的 ML 模型，以驗證您的 ML 模型是否比預測此固定答案的 ML 模型更佳。

使用效能視覺化

Amazon ML 提供混淆矩陣，做為視覺化多類別分類預測模型準確性的方法。混淆矩陣透過比較觀察的預估類別及其真正類別，以表格說明每個類別的正確和錯誤預測的數量或百分比。

例如，如果您嘗試將一部電影分類為類型，預測模型可能會將其類型 (類別) 預測為愛情片。不過，其真正類型實際上可能是驚悚片。當您評估多類別分類 ML 模型的準確性時，Amazon ML 會識別這些錯誤分類，並在混淆矩陣中顯示結果，如下圖所示。



以下資訊會顯示在混淆矩陣中：

- 每個類別的正確和錯誤預測數：混淆矩陣中的每一列都對應至其中一個真正類別的指標。例如，第一列顯示實際是愛情片類型的電影，多類別 ML 模型取得超過 80% 的正確預測。它將類型錯誤預測為驚悚片不到 20%，冒險片也是少於 20%。
- 智慧類別 F1 分數：最後一欄顯示每個類別的 F1 分數。
- 評估資料中的真正類別頻率：倒數第二欄顯示在評估資料集內，評估資料中 57.92% 的觀察是愛情片、21.23% 是驚悚片、20.85% 是冒險片。
- 評估資料的預測類別頻率：最後一列顯示預測中每個類別的頻率。 77.56% 的觀察結果預測為浪漫，9.33% 預測為 Thriller，13.12% 預測為冒險。

Amazon ML 主控台提供視覺化顯示，最多可容納混淆矩陣中的 10 個類別，以評估資料中最常到最不頻繁類別的順序列出。如果您的評估資料有 10 個以上的類別，您會在混淆矩陣中看到前 9 個最常發生的類別，所有其他類別則會摺疊成名為「其他」的類別。Amazon ML 也提供透過多類別視覺化頁面上的連結下載完整混淆矩陣的功能。

迴歸模型的深入解析

解譯預測

ML 迴歸模型的輸出為數值，是此模型對於目標的預測。例如您若要預測房價，模型就可能會預測出 254,013 一類的值。

> **Note**
>
> 預測範圍可能會與訓練資料中的目標範圍不同。例如，假設您要預測房價，而且訓練資料中目標的值範圍介於 0 到 450,000 之間。預測的目標無須在同一個範圍，而且可以接受任何正確值 (大於 450,000) 或錯誤值 (小於零)。請務必規劃如何處理落在您應用程式容許範圍以外的預測值。

衡量 ML 模型準確性

對於迴歸工作，Amazon ML 使用業界標準的均方根誤差 (RMSE) 指標。這是預測的數值目標到真實的數值答案 (真實數值) 之間的差距。RMSE 的值愈小，模型的預測正確性愈佳。模型的預測若是十分正確，其 RMSE 會是 0。下列範例顯示包含了 N 筆記錄的評估資料：

$$RMSE = \sqrt{1/N \sum_{i=1}^{N} (actual\ target - predicted\ target)^2}$$

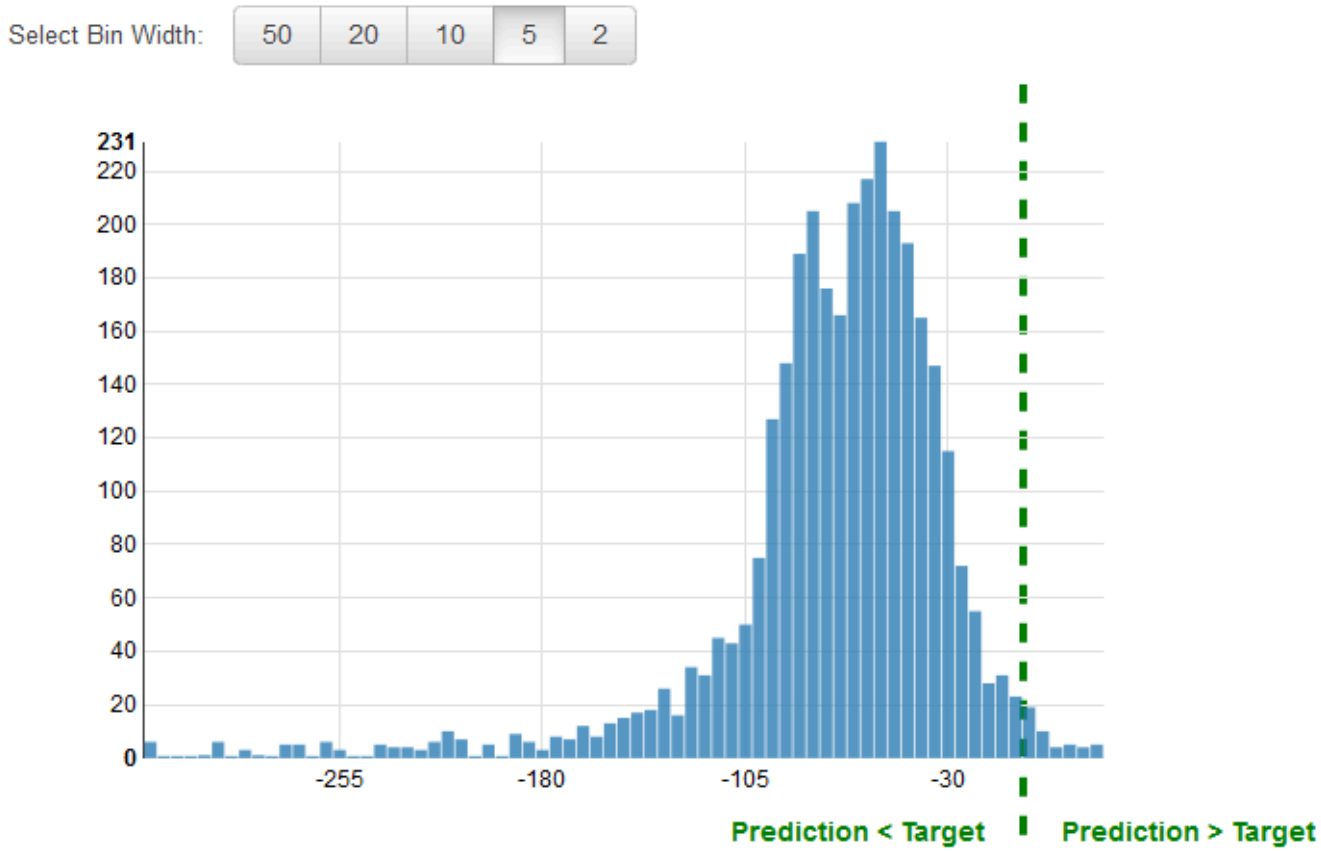
基準 RMSE

Amazon ML 提供迴歸模型的基準指標。這是假設迴歸模型的 RMSE，預策的答案一律是目標的平均值。例如，若您要預測購屋者的年齡，而訓練資料中觀察到的平均年齡為 35，則基準模型的預測答案一律是 35。您依據此基準來比較您的 ML 模型，從而確認您的 ML 模型是否優於預測此固定答案的 ML 模型。

使用效能視覺化

解決回歸問題的常見做法是檢閱「殘差」。評估資料中觀察的殘差是真正目標和預測目標之間的差距。殘差代表模型無法預測的目標部分。正殘差表示模型低估目標 (實際目標大於預測目標)。負殘差表示高估 (實際目標小於預測目標)。當評估資料的殘差長條圖呈鐘形分佈並以零為中心，表示模型以隨機的方式出錯，並未系統性過度預測或不足預測目標值的任何特定範圍。若餘數未形成中心為零的鐘形，模型

的預測誤差必有其特定的結構。將更多變數新增至模型可能有助於模型擷取目前模型未擷取到的模式。下圖顯示中心不是零的餘數。



防止過度擬合

建立和訓練 ML 模型時，目標是選擇可進行最佳預測的模型，這表示選擇具有最佳設定 (ML 模型設定或超級參數) 的模式。在 Amazon Machine Learning 中，您可以設定四個超參數：通過次數、正規化、模型大小和隨機類型。不過，如果您選擇會對評估資料產生「最佳」預測效能的模型參數設定，您可能會過度擬合模型。當模型記住訓練和評估資料來源中發生的模式，但無法一般化資料中的模式，就會發生過度擬合。它通常發生在訓練資料包含用於評估的所有資料。過度擬合的模型在評估期間表現良好，但無法對未知資料進行準確的預測。

為了避免選取過度擬合的模型做為最佳模型，您可以保留額外的資料來驗證 ML 模型的效能。例如，您可以將您的資料分為 60% 用於訓練、20% 用於評估，其他 20% 用於驗證。在選擇很適合執行評估資料的模型參數後，您須使用驗證資料執行第二個評估，以查看 ML 模型對於驗證資料的執行效能。如果模型在驗證資料上符合您的期望，就表示模型未過度擬合資料。

使用第三組資料進行驗證，可協助您選擇適當的 ML 模型參數以防止過度擬合。不過，從訓練程序提取用於評估和驗證的資料，會讓可用於訓練的資料變得更少。這是小型資料集要特別留意的問題，因為能

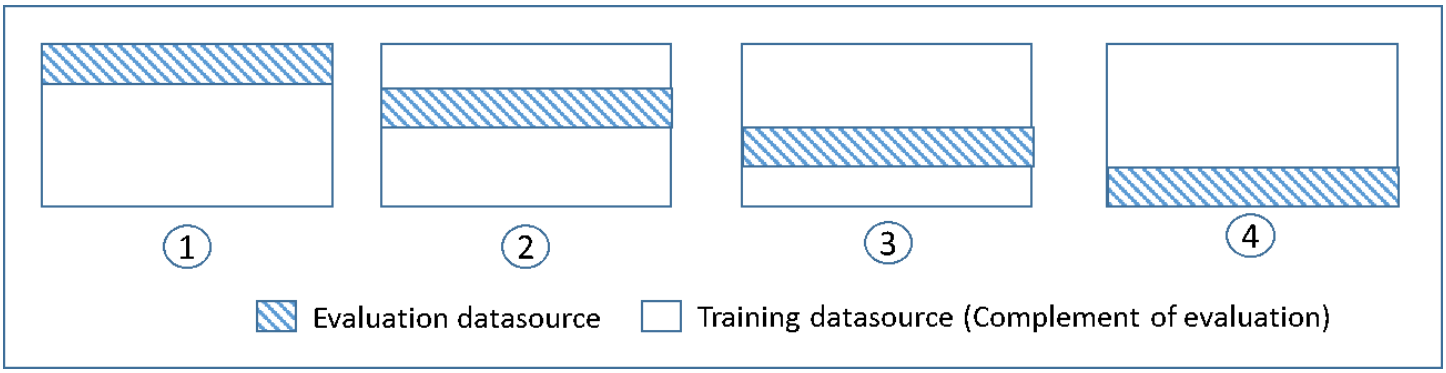
用於訓練的資料總是越多越好。若要解決這個問題，您可以執行交叉驗證。如需交叉驗證的詳細資訊，請參閱交叉驗證。

交叉驗證

交叉驗證是一種評估 ML 模型的技術，採用的方法是使用可用輸入資料的子集來訓練數個 ML 模型並根據資料的互補子集來評估這些模型。使用交叉驗證可以偵測過度擬合，亦即無法一般化模式。

在 Amazon ML 中，您可以使用 k 倍交叉驗證方法來執行交叉驗證。在 k 倍交叉驗證中，您可以將輸入資料分割為 k 個資料子集 （ 也稱為折疊 ）。 您可以在除了一個 (k-1) 以外的所有子集上訓練 ML 模型，然後在未用於訓練的子集上評估模型。此程序會重複 k 次，每次保留不同的子集用於評估 (以及排除不用於訓練)。

下圖顯示在 4 倍交叉驗證期間建立和訓練的四個模型，為每個模型所產生的訓練子集和互補評估子集範例。模型一將前 25% 的資料用於評估，其餘 75% 用於訓練。模型二將第二個 25% 子集 (25% 到 50%) 用於評估，其餘三個子集 的資料用於訓練，依此類推。



每個模型使用互補資料來源來訓練和評估 - 評估資料來源中的資料包含並受限於不屬於訓練資料來源的所有資料。您使用 DataRearrangement、createDatasourceFromS3 和 createDatasourceFromRedShift API 中的 createDatasourceFromRDS 參數來建立這些子集的資料來源。在 DataRearrangement 參數中，指定每個區段的開始和結束位置，來指定資料來源要包含哪些資料子集。若要建立 4k 倍交叉驗證所需的互補資料來源，請指定 DataRearrangement 參數，如以下範例所示：

模型一：

評估的資料來源：

```
{"splitting":{"percentBegin":0, "percentEnd":25}}
```

培訓的資料來源：

```
{"splitting":{"percentBegin":0, "percentEnd":25, "complement":"true"}}
```

模型二：

評估的資料來源：

```
{"splitting":{"percentBegin":25, "percentEnd":50}}
```

培訓的資料來源：

```
{"splitting":{"percentBegin":25, "percentEnd":50, "complement":"true"}}
```

模型三：

評估的資料來源：

```
{"splitting":{"percentBegin":50, "percentEnd":75}}
```

培訓的資料來源：

```
{"splitting":{"percentBegin":50, "percentEnd":75, "complement":"true"}}
```

模型四：

評估的資料來源：

```
{"splitting":{"percentBegin":75, "percentEnd":100}}
```

培訓的資料來源：

```
{"splitting":{"percentBegin":75, "percentEnd":100, "complement":"true"}}
```

執行 4 倍交叉驗證會產生四個模型、四個資料來源來訓練模型、四個資料來源來評估模型，以及四個評估，每個模型各一個。Amazon ML 會為每個評估產生模型效能指標。例如，在二元分類問題的 4 倍交叉驗證中，每個評估報告一個曲線下的區域 (AUC) 指標。您可以透過計算這四個 AUC 指標的平均，取得整體效能測量。如需 AUC 指標的相關資訊，請參閱衡量 ML 模型準確性。

如需示範如何建立跨驗證和平均模型分數的範例程式碼，請參閱 Amazon ML 範例程式碼。

調整您的模型

交叉驗證模型之後，如果您的模型效能不符合你的標準，您可為下一個模型調整設定。如需過度擬合的詳細資訊，請參閱模型擬合：低度擬合與過度擬合。如需正規化的詳細資訊，請參閱正規化。如需變更正規化設定的詳細資訊，請參閱使用自訂選項建立 ML 模型。

評估提醒

Amazon ML 提供洞見，協助您驗證是否正確評估模型。如果評估不符合任何驗證條件，Amazon ML 主控台會顯示已違反的驗證條件來提醒您，如下所示。

- ML 模型是使用留存資料進行評估

  如果您使用相同的資料來源進行訓練和評估，Amazon ML 會提醒您。如果您使用 Amazon ML 分割資料，您將符合此有效性條件。如果您不使用 Amazon ML 分割資料，請務必使用訓練資料來源以外的資料來源來評估您的 ML 模型。

- 足夠的資料用於評估預測模型

  如果評估資料中的觀察/記錄數量少於您在訓練資料來源中觀察的 10%，Amazon ML 會提醒您。若要正確地評估模型，請務必提供足夠的大型資料範例。此條件會提供檢查，讓您知道是否使用太少的資料。評估 ML 模型所需的資料量是主觀的。此處選擇 10% 作為停止差距，因為沒有更好的措施。

- 符合的結構描述

  如果訓練和評估資料來源的結構描述不同，Amazon ML 會提醒您。如果您有某些屬性不存在於評估資料來源中，或有其他屬性，Amazon ML 會顯示此提醒。

- 評估檔案中的所有記錄都用於預測模型效能評估

  請務必了解提供用於評估的所有記錄是否實際用於評估模型。如果評估資料來源中的某些記錄無效，且未包含在準確性指標運算中，Amazon ML 會提醒您。例如，如果評估資料來源中的某些觀察缺少目標變數，Amazon ML 無法檢查 ML 模型對這些觀察的預測是否正確。在這種情況下，會將具有遺漏目標值的記錄視為無效。

- 目標變數的分佈

  Amazon ML 會顯示訓練和評估資料來源中目標屬性的分佈，以便您可以檢閱目標是否在兩個資料來源中以類似方式分佈。如果根據目標分佈與評估資料上目標分佈不同的培訓資料來培訓模型，則評估品質可能不佳，因為會根據具有極不同統計資料的資料來計算它。最好以類似的方式將資料分佈到培訓和評估資料，並讓這些資料集盡可能模仿模型在預測時將遇到的資料。

如果觸發此提醒，請嘗試使用隨機分割策略，將資料分割為培訓和評估資料來源。在極少數情況下，即使您隨機分割資料，此提醒仍可能會錯誤地警告您目標分佈差異。Amazon ML 使用近似資料統計資料來評估資料分佈，偶爾會錯誤觸發此提醒。

產生和解譯預測

Amazon ML 提供兩種產生預測的機制：非同步 （ 批次型 ） 和同步 one-at-a-time)。

若您有多個觀察且希望同時獲得所有觀察的預測結果，請使用非同步預測，或稱「批次預測」。程序使用資料來源做為輸入，並將預測輸出到存放在您所選 S3 儲存貯體的 .csv 檔案中。您需要等待批次預測程序完成後，才能存取預測結果。Amazon ML 可在批次檔案中處理的資料來源大小上限為 1 TB （約 1 億筆記錄）。如果您的資料來源大於 1 TB，您的任務將失敗，Amazon ML 將傳回錯誤碼。若要避免發生這種狀況，請將您的資料分為多個批次。如果您的記錄通常較長，在處理 1 億筆記錄之前，您就會到達 1 TB 的限制。在這種情況下，建議您聯絡 AWS Support 以提高您批次預測的任務大小。

若您希望以低延遲獲得預測結果，請使用同步預測，或稱「即時預測」。即時預測 API 接受序列化為 JSON 字串的一個輸入觀察，並在 API 回應中同步傳回預測和相關中繼資料。您可以同時叫用 API 多次以取得平行同步預測結果。如需即時預測 API 輸送量限制的詳細資訊，請參閱《Amazon ML API 參考》中的即時預測限制。

主題

- 建立批次預測
- 檢閱批次預測指標
- 讀取批次預測輸出檔案
- 要求即時預測

建立批次預測

若要建立批次預測，您可以使用 Amazon Machine Learning (Amazon ML) 主控台或 API 建立`BatchPrediction`物件。`BatchPrediction` 物件描述了 Amazon ML 使用您的 ML 模型和一組輸入觀察所產生的一組預測。當您建立`BatchPrediction`物件時，Amazon ML 會啟動非同步工作流程來計算預測。

對於您用來取得批次預測的資料來源以及您用於訓練 ML 模型以用於查詢預測的資料來源，都必須使用相同的結構描述。唯一的例外是，批次預測的資料來源不需要包含目標屬性，因為 Amazon ML 會預測目標。如果您提供目標屬性，Amazon ML 會忽略其值。

建立批次預測 (主控台)

若要使用 Amazon ML 主控台建立批次預測，請使用建立批次預測精靈。

建立批次預測 (主控台)

1. 登入 AWS 管理主控台 並開啟 Amazon Machine Learning 主控台，網址為 https://console.aws.amazon.com/machinelearning/：//。
2. 在 Amazon ML 儀表板的物件下，選擇建立新...，然後選擇批次預測。
3. 選擇您要用來建立批次預測的 Amazon ML 模型。
4. 若要確認您要使用此模型，請選擇 Continue (繼續)。
5. 選擇您要為其建立預測的資料來源。資料來源必須擁有跟模型相同的結構描述，但不需要包含目標屬性。
6. 選擇繼續。
7. 對於 S3 destination (S3 目的地)，輸入 S3 儲存貯體的名稱。
8. 選擇檢閱。
9. 檢閱您的設定，然後選擇 Create batch prediction (建立批次預測)。

建立批次預測 (API)

若要使用 Amazon ML API 建立`BatchPrediction`物件，您必須提供下列參數：

資料來源 ID

指向您想要預測之觀察的資料來源 ID。例如，如果您想要 `s3://examplebucket/input.csv` 檔案中資料的預測結果，您可以建立一個資料來源物件，指向該資料檔案，然後使用此參數傳遞該資料來源的 ID。

BatchPrediction ID

要指派給批次預測的 ID。

ML 模型 ID

Amazon ML 應查詢預測的 ML 模型 ID。

輸出 Uri

要存放預測輸出的 S3 儲存貯體 URI。Amazon ML 必須具有將資料寫入此儲存貯體的許可。

`OutputUri` 參數必須參考結尾為正斜線 ('/') 字元的 S3 路徑，如下所示：

s3://examplebucket/examplepath/

如需設定 S3 許可的詳細資訊，請參閱授予 Amazon ML 將預測輸出至 Amazon S3 的許可。

(選用) BatchPrediction 名稱

(選用) 批次預測的人類可讀取名稱。

檢閱批次預測指標

Amazon Machine Learning (Amazon ML) 建立批次預測後，會提供兩個指標： `Records seen`和`Records failed to process`。 `Records seen`會告訴您 Amazon ML 在執行批次預測時查看了多少筆記錄。 `Records failed to process`會告訴您 Amazon ML 無法處理多少筆記錄。

若要允許 Amazon ML 處理失敗的記錄，請檢查用於建立資料來源的資料中的記錄格式，並確認所有必要屬性都存在，且所有資料皆正確。修正您的資料後，您可以重新建立您的批次預測，或以失敗的記錄建立新的資料來源，然後使用新的資料來源建立新的批次預測。

檢閱批次預測指標 (主控台)

若要查看 Amazon ML 主控台中的指標，請開啟批次預測摘要頁面，並查看已處理的資訊區段。

檢閱批次預測指標和詳細資訊 (API)

您可以使用 Amazon ML APIs 擷取`BatchPrediction`物件的詳細資訊，包括記錄指標。Amazon ML 提供下列批次預測 API 呼叫：

- `CreateBatchPrediction`
- `UpdateBatchPrediction`
- `DeleteBatchPrediction`
- `GetBatchPrediction`
- `DescribeBatchPredictions`

如需詳細資訊，請參閱 Amazon ML API 參考。

讀取批次預測輸出檔案

執行以下步驟，擷取批次預測輸出檔案：

1. 尋找批次預測資訊清單檔案。
2. 讀取資訊清單檔案，判斷輸出檔案的位置。

3. 擷取包含預測的輸出檔案。
4. 解譯輸出檔案的內容。目錄會根據用來產生預測的 ML 模型類型而有所不同。

以下章節會更詳細地說明操作步驟。

尋找批次預測資訊清單檔案

批次預測資訊清單檔案中包含的資訊，可將輸入檔案對應到預測輸出檔案。

若要尋找資訊清單檔案，請從建立批次預測物件時所指定的輸出位置開始。您可以使用 Amazon ML API 或 https：//https://console.aws.amazon.com/machinelearning/ 查詢已完成的批次預測物件，以擷取此檔案的 S3 位置。

資訊清單檔案所在的輸出位置路徑，包含附加到輸出位置的靜態字串 `/batch-prediction/`，和資訊清單檔案的名稱，也就是批次預測的 ID，再加上 `.manifest`。

例如，如果您建立 ID 為`bp-example` 的批次預測物件，並指定了 S3 位置 `s3://examplebucket/output/` 做為輸出位置，將會在以下位置找到資訊清單檔案：

```
s3://examplebucket/output/batch-prediction/bp-example.manifest
```

讀取資訊清單檔案

.manifest 檔案的內容會編碼為 JSON 對應，其中的金鑰是一組 S3 輸入資料檔案名稱的字串，而值是一組相關聯批次預測結果檔案的字串。每對輸入/輸出檔案都有一行映射內容。繼續舉例而言，如果 `BatchPrediction` 物件的建立輸入包含名為 data.csv 的單一檔案，而此檔案位於 `s3://examplebucket/input/`，您可能就會看到與下文類似的映射字串：

```
{"s3://examplebucket/input/data.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data.csv.gz"}
```

如果 `BatchPrediction` 物件的建立輸入包含名為 data1.csv、data2.csv 及 data3.csv 三個檔案，而這些檔案全都儲存在 S3 位置 `s3://examplebucket/input/`，您可能就會看到與下文類似的映射字串：

```
{"s3://examplebucket/input/data1.csv":"s3://examplebucket/output/batch-prediction/
result/bp-example-data1.csv.gz",

"s3://examplebucket/input/data2.csv":"
```

```
s3://examplebucket/output/batch-prediction/result/bp-example-data2.csv.gz",

"s3://examplebucket/input/data3.csv":"
s3://examplebucket/output/batch-prediction/result/bp-example-data3.csv.gz"}
```

擷取批次預測輸出檔案

您可以從資訊清單映射下載各個批次預測檔案，然後直接在本機處理。檔案格式是使用 gzip 演算法壓縮的 CSV。在該檔案內，對應輸入檔案中的每個輸入觀察皆各有一行。

若要讓預測與批次預測的輸入檔案合併在一起，您可以對兩個檔案執行簡單的依記錄合併動作。批次預測的輸出檔案一律包含與預測輸入檔案相同的記錄數量，而且順序相同。如果輸入觀察處理失敗，就不會產生任何預測，那麼批次預測的輸出檔案將會在對應位置上出現一行空白。

解譯二元分類 ML 模型的批次預測檔案內容

二元分類模型的批次預測檔案欄名為 bestAnswer (最佳答案) 和 score (分數)。

bestAnswer (最佳答案) 欄包含的預測標籤 (「1」或「0」) 是評估了預測分數相較於分界分數而得出的結果。如需分界分數的詳細資訊，請參閱調整分界分數。您可以使用 Amazon ML API 或 Amazon ML 主控台上的模型評估功能，設定 ML 模型的分界分數。如果您未設定截止分數，Amazon ML 會使用預設值 0.5。

分數欄包含 ML 模型針對此預測指派的原始預測分數。Amazon ML 使用邏輯回歸模型，因此此分數會嘗試將對應至 true ("1") 值的觀察機率建立模型。請注意，score (分數) 是以科學記號標記法回報，因而下例第一列中的值 8.7642E-3 也就等於 0.0087642。

例如，若 ML 模型的分界分數是 0.75，二元分類模型的批次預測輸出檔案內容可能如下所示：

```
bestAnswer,score

0,8.7642E-3

1,7.899012E-1


0,6.323061E-3

0,2.143189E-2
```

```
1,8.944209E-1
```

輸入檔案中第二個和第五個觀察皆有高於 0.75 的預測分數，因此這些觀察的 bestAnswer 欄會出現「1」的值，而其他觀察則出現「0」的值。

解譯二進位多級分類 ML 模型的批次預測檔案內容

在多等級模型的批次預測檔案中，將會包含一個用於訓練資料中各個等級的欄位。欄名稱會出現在批次預測檔案的標題列。

當您從多類別模型請求預測時，Amazon ML 會針對輸入檔案中的每個觀察運算數個預測分數，而輸入資料集中定義的每個類別各一個。這等同於詢問「相對於其他等級，此觀察歸為此等級的機率為何？(以 0 和 1 來評估)」 每個分數可以轉譯為「該觀察屬於此等級的機率。」 由於預測分數會模擬觀察屬於一種等級或其他等級的基礎機率，因此一列中所有預測分數的總和為 1。您必須挑選一個等級做為模型的預測等級。通常，您可以挑選最有可能是最佳答案的等級。

例如，試想一下在試著預測客戶對產品的評價時，會以 1 到 5 顆星來評比。如果等級名為 `1_star`、`2_stars`、`3_stars`、`4_stars` 和 `5_stars`，那麼多級預測輸出檔案可能如下所示：

```
1_star, 2_stars, 3_stars, 4_stars, 5_stars

8.7642E-3, 2.7195E-1, 4.77781E-1, 1.75411E-1, 6.6094E-2

5.59931E-1, 3.10E-4, 2.48E-4, 1.99871E-1, 2.39640E-1

7.19022E-1, 7.366E-3, 1.95411E-1, 8.78E-4, 7.7323E-2

1.89813E-1, 2.18956E-1, 2.48910E-1, 2.26103E-1, 1.16218E-1

3.129E-3, 8.944209E-1, 3.902E-3, 7.2191E-2, 2.6357E-2
```

在此範例中，第一個觀察的預測分數最高，為 `3_stars` 級 (預測分數 = 4.77781E-1)，因此您可以將結果解譯為，表示 `3_stars` 級是此觀察的最佳答案。請注意，預測分數會以科學記號標記法來報告，所以預測分數 4.77781E-1 也就等於 0.477781。

有時您也許並不想要選擇最高機率的等級，例如，您可能會想要建立最低閾值，若低於此值，即使某個等級有最高預測分數也不會將其視為最佳答案。假設您正在為電影分類，而且希望預測分數至少要有 5E-1，才能表示該分類是最佳答案。喜劇片取得的預測分數是 3E-1、劇情片是 2.5E-1、紀錄片是 2.5E-1、動作片是 2E-1。在此情況下，ML 模型預測您最可能會選擇喜劇片，但您決定不依該最佳答案做出選擇。這是因為所有預測分數都不超過您的基準預測分數 5E-1，所以您判斷這些預測不足以準確

預估電影分類，而決定選擇其他方式。之後您的應用程式可能會將這部電影的「分類」欄位標為「不明」。

解譯回歸 ML 模型的批次預測檔案內容

回歸模型的批次預測檔案中有一個欄，名為 score (分數)。此欄包含輸入資料中每個觀察的原始數字預測。這些數值以科學記號標記法回報，因而下例第一列中的 score (分數) 值 -1.526385E1 也就等於 -15.26835。

此範例顯示在回歸模型上執行的批次預測輸出檔案：

```
score

-1.526385E1

-6.188034E0

-1.271108E1

-2.200578E1

8.359159E0
```

要求即時預測

即時預測是對 Amazon Machine Learning (Amazon ML) 的同步呼叫。當 Amazon ML 取得請求時，就會進行預測，並立即傳回回應。即時預測常用來在互動式 Web、行動或桌面應用程式內啟用預測功能。您可以使用低延遲 `Predict` API 即時查詢使用 Amazon ML 建立的 ML 模型以進行預測。`Predict` 操作會接受要求承載中的單一輸入觀察，並在回應中同步傳回預測。這與批次預測 API 不同，該 API 會使用指向輸入觀察位置的 Amazon ML 資料來源物件 ID 叫用，並以非同步方式將 URI 傳回至包含所有這些觀察預測的檔案。Amazon ML 會在 100 毫秒內回應大多數即時預測請求。

您可以在 Amazon ML 主控台中嘗試即時預測，而不會產生費用。如果您接著決定使用即時預測，則必須先建立即時預測產生的端點。您可以在 Amazon ML 主控台或使用 `CreateRealtimeEndpoint` API 執行此操作。在您有端點之後，請使用即時預測 API 來產生即時預測。

> **Note**
>
> 在您建立模型的即時端點之後，即會根據模型大小來開始產生容量保留費用。如需詳細資訊，請參閱 定價。如果您在主控台中建立即時端點，則主控台會顯示端點將持續增加的

預估費用明細。若要在不再需要取得該模型的即時預測時停止產生費用，請使用主控台或 `DeleteRealtimeEndpoint` 操作來移除即時端點。

如需`Predict`請求和回應的範例，請參閱《Amazon Machine Learning API 參考》中的預測。若要查看使用您模型的確切回應格式範例，請參閱嘗試即時預測。

主題

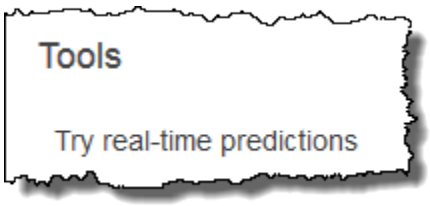
- 嘗試即時預測
- 建立即時端點
- 找到即時預測端點 (主控台)
- 找到即時預測端點 (API)
- 建立即時預測要求
- 刪除即時端點

嘗試即時預測

為了協助您決定是否啟用即時預測，Amazon ML 可讓您嘗試在單一資料記錄上產生預測，而不會產生與設定即時預測端點相關聯的額外費用。若要嘗試即時預測，您必須具有 ML 模型。若要大規模建立即時預測，請使用 Amazon Machine Learning API 參考中的預測 API。 Amazon Machine Learning

嘗試即時預測

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 從導覽列的 Amazon Machine Learning 下拉式選單中選擇 ML models (ML 模型)。
3. 選擇您想要用來嘗試即時預測的模型，例如教學課程中的 `Subscription propensity model`。
4. 在 ML 模型報告頁面上，於 Predictions (預測) 下選擇 Summary (摘要)，然後選擇 Try real-time predictions (嘗試即時預測)。



Amazon ML 會顯示構成 Amazon ML 用來訓練模型之資料記錄的變數清單。

5. 在表單的每個欄位中輸入資料，或以 CSV 格式將單一資料記錄貼入文字方塊，即可繼續進行。

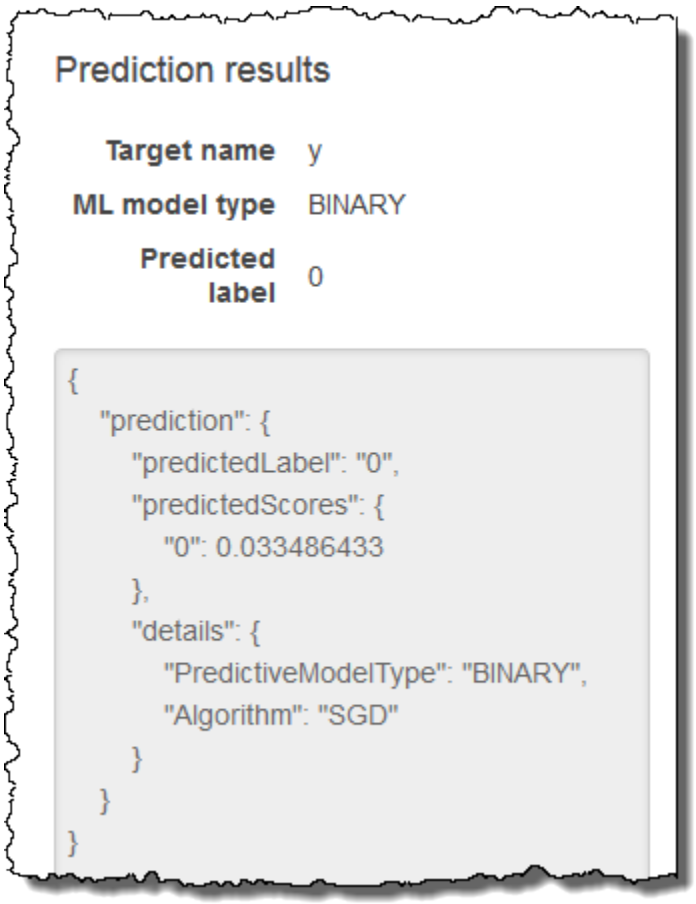
   若要使用表單，請在各個 Value (值) 欄位內輸入您想要用來測試即時預測的資料。如果您輸入的資料記錄未包含一或多個資料屬性的值，請將輸入欄位空白。

   若要提供資料記錄，請選擇 Paste a record (貼上記錄)。將單一 CSV 格式的資料列貼到文字欄位中，然後選擇提交。Amazon ML 會自動為您填入值欄位。

   > **Note**
   >
   > 資料記錄中的資料必須具有與培訓資料相同數目的資料行，並依相同的順序排列。唯一的例外是您應該省略目標值。如果您包含目標值，Amazon ML 會忽略它。

6. 在頁面底部，選擇 Create prediction (建立預測)。Amazon ML 會立即傳回預測。

   在 Prediction results (預測結果) 窗格中，您會看到 `Predict` API 呼叫所傳回的預測物件，以及 ML 模型類型、目標變數的名稱和預測類別或值。如需解譯結果的資訊，請參閱解譯二元分類 ML 模型的批次預測檔案內容。



建立即時端點

若要產生即時預測，您需要建立即時端點。若要建立即時端點，您必須已有要產生即時預測的 ML 模型。您可以使用 Amazon ML 主控台或呼叫 `CreateRealtimeEndpoint` API 來建立即時端點。如需使用 `CreateRealtimeEndpoint` API 的詳細資訊，請參閱《Amazon Machine Learning API 參考https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_CreateRealtimeEndpoint.html》中的 。

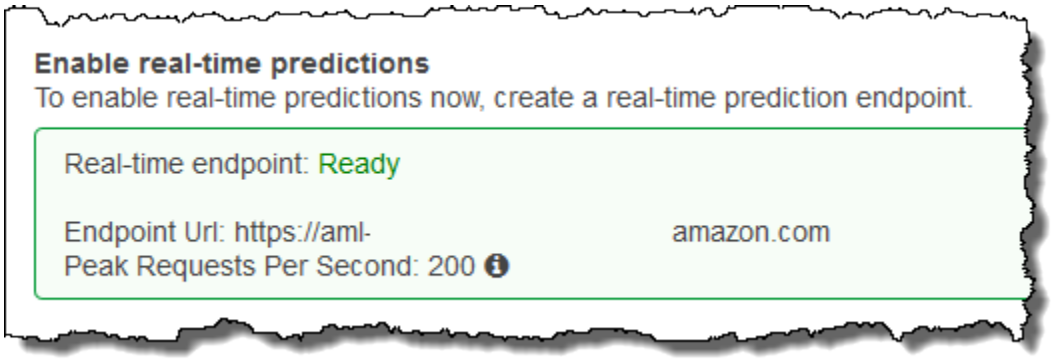
建立即時端點

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 從導覽列的 Amazon Machine Learning 下拉式選單中選擇 ML models (ML 模型)。
3. 選擇您要產生即時預測的模型。
4. 在 ML model summary (ML 模型摘要) 頁面上，於 Predictions (預測) 下選擇 Create real-time endpoint (建立即時端點)。

即會顯示說明即時預測定價方式的對話方塊。

5. 選擇建立。即時端點請求會傳送至 Amazon ML 並進入佇列。即時端點的狀態是 Updating (正在更新)。

6. 當即時端點就緒時，狀態會變更為就緒，Amazon ML 會顯示端點 URL。透過 `Predict` API，使用端點 URL 建立即時預測要求。如需使用 `Predict` API 的詳細資訊，請參閱《Amazon Machine Learning API 參考https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html》中的 。



找到即時預測端點 (主控台)

若要使用 Amazon ML 主控台尋找 ML 模型的端點 URL，請導覽至模型的 ML 模型摘要頁面。

找到即時端點 URL

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 從導覽列的 Amazon Machine Learning 下拉式選單中選擇 ML models (ML 模型)。
3. 選擇您要產生即時預測的模型。
4. 在 ML model summary (ML 模型摘要) 頁面上，向下捲動以查看 Predictions (預測) 區段。
5. 模型的端點 URL 會列在 Real-time prediction (即時預測) 中。使用此 URL 做為即時預測呼叫的 Endpoint Url (端點 URL)。如需如何使用端點產生預測的資訊，請參閱《Amazon Machine

Learning API 參考https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_Predict.html》中的 。

找到即時預測端點 (API)

當您使用 CreateRealtimeEndpoint 操作建立即時端點時，會在回應中將端點的 URL 和狀態傳回給您。如果您已使用主控台建立即時端點，或想要稍後擷取您所建立端點的 URL 和狀態，請呼叫具有您要查詢即時預測之模型識別符的 GetMLModel 操作。端點資訊包含在回應的 EndpointInfo 區段中。針對具有相關聯即時端點的模型，EndpointInfo 可能會如下所示：

```
"EndpointInfo":{
    "CreatedAt": 1427864874.227,
    "EndpointStatus": "READY",
    "EndpointUrl": "https://endpointUrl",
    "PeakRequestsPerSecond": 200
}
```

沒有即時端點的模型會傳回下列各項：

```
EndpointInfo":{
    "EndpointStatus": "NONE",
    "PeakRequestsPerSecond": 0
}
```

建立即時預測要求

範例 Predict 要求承載可能如下所示：

```
{
    "MLModelId": "model-id",
    "Record":{
        "key1": "value1",
        "key2": "value2"
    },
    "PredictEndpoint": "https://endpointUrl"
}
```

PredictEndpoint 欄位必須對應至 EndpointInfo 結構的 EndpointUrl 欄位。Amazon ML 使用此欄位將請求路由到即時預測機群中的適當伺服器。

MLModelId 是具有即時端點之先前培訓過模型的識別符。

Record 是變數名稱與變數值的對應。每個配對都代表一個觀察。Record 映射包含 Amazon ML 模型的輸入。它類似培訓資料集中的單一資料列，而沒有目標變數。不論培訓資料中的值類型為何，Record 都會包含字串對字串對應。

> **Note**
>
> 您可以省略沒有值的變數，但這可能會降低您預測的準確性。您可以包含的變數愈多，模型會更準確。

Predict 要求所傳回的回應格式取決於用於查詢預測的模型類型。在所有情況下，details 欄位都會包含預測要求的資訊，尤其是包含具有模型類型的 PredictiveModelType 欄位。

下列範例示範二元模型的回應：

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "BINARY"
        },
        "predictedLabel": "0",
        "predictedScores":{
            "0": 0.47380468249320984
        }
    }
}
```

請注意包含預測標籤predictedLabel的欄位，在此案例中為 0。Amazon ML 透過比較預測分數與分類截止值來計算預測標籤：

- 您可以檢查 GetMLModel 操作回應中的 ScoreThreshold 欄位，或在 Amazon ML 主控台中檢視模型資訊，以取得目前與 ML 模型相關聯的分類截止。如果您未設定分數閾值，Amazon ML 會使用預設值 0.5。
- 您可以檢查 predictedScores 對應，以取得二元分類模型的確切預測分數。在這個對應內，預測標籤會與確切的預測分數搭配使用。

如需二元預測的詳細資訊，請參閱解譯預測。

下列範例示範回歸模型的回應。請注意，預測數值位於 predictedValue 欄位中：

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "REGRESSION"
        },
        "predictedValue": 15.508452415466309
    }
}
```

下列範例示範多類別模型的回應：

```
{
    "Prediction":{
        "details":{
            "PredictiveModelType": "MULTICLASS"
        },
        "predictedLabel": "red",
        "predictedScores":{
            "red": 0.12923571467399597,
            "green": 0.08416014909744263,
            "orange": 0.22713537514209747,
            "blue": 0.1438363939523697,
            "pink": 0.184102863073349,
            "violet": 0.12816807627677917,
            "brown": 0.10336143523454666
        }
    }
}
```

與二元分類模型類似，預測標籤/類別位於 predictedLabel 欄位中。您可以查看 predictedScores 對應，以進一步了解預測與每個類別的緊密相關程度。此對應內的類別分數愈高，預測與類別的相關性愈強，而最高值最後會選取為 predictedLabel。

如需多類別預測的詳細資訊，請參閱多類別模型深入分析。

刪除即時端點

當您完成即時預測時，請刪除即時端點，以避免產生額外費用。只要刪除端點，就會立即停止產生費用。

刪除即時端點

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 從導覽列的 Amazon Machine Learning 下拉式選單中選擇 ML models (ML 模型)。
3. 選擇不再需要即時預測的模型。
4. 在 ML 模型報告頁面上，於 Predictions (預測) 下選擇 Summary (摘要)。
5. 選擇 Delete real-time endpoint (刪除即時端點)。
6. 在 Delete real-time endpoint (刪除即時端點) 對話方塊中，選擇 Delete (刪除)。

管理 Amazon ML 物件

Amazon ML 提供四個物件，您可以透過 Amazon ML 主控台或 Amazon ML API 進行管理：

- 資料來源
- ML 模型
- 評估
- 批次預測

每個物件在建立機器學習應用程式的生命週期中都有不同的用途，而且每個物件都有只適用於該物件的特定屬性和功能。雖有這些差異，但物件的管理方式都很類似。例如，您用來列出物件、擷取其描述及更新或刪除物件的程序幾乎相同。

下列各節會說明所有四項物件常見的管理操作，並在任何不同處加上備註。

主題

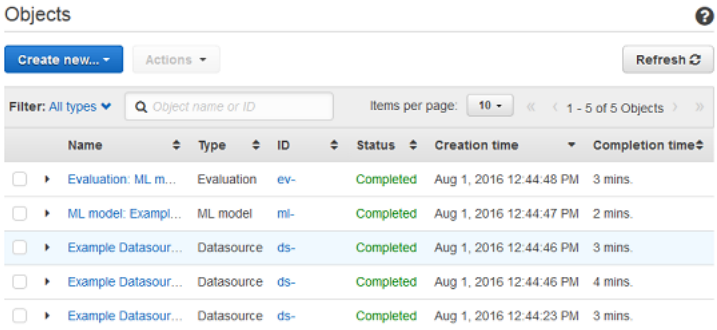
- 列出物件
- 擷取物件描述
- 更新物件
- 刪除物件

列出物件

如需有關 Amazon Machine Learning (Amazon ML) 資料來源、ML 模型、評估和批次預測的深入資訊，請列出這些資訊。對於每個物件，您會看到其名稱、類型、ID、狀態碼和建立時間。您也可以查看特定物件類型的專屬詳細資訊。例如，您可以查看資料來源的資料深入分析。

列出物件 (主控台)

若要查看您建立的最後 1，000 個物件清單，請在 Amazon ML 主控台中開啟物件儀表板。若要顯示物件儀表板，請登入 Amazon ML 主控台。



若要查看物件的詳細資訊，包括物件類型專屬的詳細資訊，選擇物件的名稱或 ID。例如，若要查看資料來源的 Data insights (資料深入分析)，請選擇資料來源名稱。

Objects (物件) 儀表板中的欄會顯示每個物件的下列資訊。

名稱

物件的名稱。

類型

物件的類型。有效值包括 Datasource (資料來源)、ML model (機器學習模型)、Evaluation (評估) 和 Batch prediction (批次預測)。

> **Note**
>
> 若要查看模型是否設定為支援即時預測，請選擇名稱或模型 ID，前往 ML model summary (機器學習模型摘要) 頁面。

ID

物件的 ID。

狀態

物件的狀態。值包括 Pending (待定)、In Progress (進行中)、Completed (已完成) 和 Failed (失敗)。如果狀態為 Failed (失敗)，請檢查您的資料後再重試。

建立時間

Amazon ML 完成建立此物件的日期和時間。

完成時間

Amazon ML 建立此物件所花費的時間長度。您可以使用模型的完成時間來預估新模型的訓練時間。

資料來源 ID

對於使用資料來源建立的物件，例如模型和評估，這是資料來源的 ID。如果刪除資料來源，您再也無法使用以該資料來源建立的 ML 模型來建立預測。

選擇欄標頭旁的雙三角形圖示，即可依據任何欄排序。

列出物件 (API)

在 Amazon ML API 中，您可以使用下列操作，依類型列出物件：

- `DescribeDataSources`
- `DescribeMLModels`
- `DescribeEvaluations`
- `DescribeBatchPredictions`

每個操作包含參數，用以篩選、排序和分頁很長的物件清單。您可以透過 API 存取任意數量的物件，沒有上限。若要限制清單的大小，請使用 `Limit` 參數，最大值為 100。

API 對 `Describe*` 命令的回應包含分頁符記 (`nextPageToken`) (若適用)，以及每個物件的簡要說明。物件描述包含主控台中顯示每個物件類型的相同資訊，包括物件類型專屬的詳細資訊。

> **Note**
>
> 即使回應中包含少於所指定限制的物件，它仍可能包含 `nextPageToken`，表示還有更多結果。即使包含 0 個項目的回應，也可能包含 `nextPageToken`。

如需詳細資訊，請參閱 Amazon ML API 參考。

擷取物件描述

您可以透過主控台或 API 檢視任何物件的詳細描述。

主控台中的詳細描述

若要在主控台查看描述，請導覽至特定物件類型的清單 (資料來源、ML 模型、評估或批次預測)。接著，透過瀏覽整個清單或者搜尋其名稱或 ID，找出表格中對應到物件的列。

API 中的詳細描述

每個物件類型都有一個擷取 Amazon ML 物件之完整詳細資訊的操作：

- GetDataSource
- GetMLModel
- GetEvaluation
- GetBatchPrediction

每項操作都需要兩個參數：物件 ID 和稱為 Verbose (詳細資訊) 的布林值旗標。Calls with Verbose (具有詳細資訊的呼叫) 設為 true 會包含物件的額外詳細資訊，這會造成延遲提高和回應增加。若要了解設定 Verbose (詳細資訊) 旗標會包含哪些欄位，請參閱《Amazon ML API 參考》。

更新物件

每個物件類型都有 操作，可更新 Amazon ML 物件的詳細資訊 （ 請參閱 Amazon ML API 參考) ：

- UpdateDataSource
- UpdateMLModel
- UpdateEvaluation
- UpdateBatchPrediction

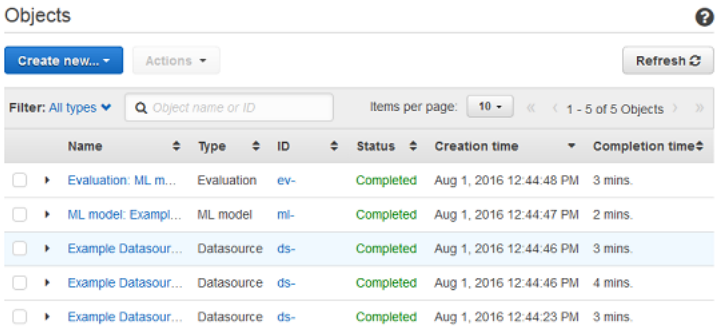
每項操作都需要物件的 ID，才能指定所要更新的物件。您可以更新所有物件的名稱。您無法更新資料來源、評估與批次預測之物件的任何其他屬性。對於 ML 模型，您可以更新 ScoreThreshold 欄位，但前提是 ML 模型沒有相關聯的即時預測端點。

刪除物件

當您不再需要資料來源、ML 模型、評估和批次預測時，可以將它們刪除。雖然在您完成批次預測後，保留 Amazon ML 物件以外的 Amazon ML 物件不會產生額外的成本，但刪除物件可讓您的工作區保持

井然有序且易於管理。您可以使用 Amazon Machine Learning (Amazon ML) 主控台或 API 刪除一或多個物件。

> **Warning**
>
> 當您刪除 Amazon ML 物件時，效果會立即、永久且不可復原。



刪除物件 (主控台)

您可以使用 Amazon ML 主控台刪除物件，包括模型。您用來刪除模型的程序，取決於您是否使用模型來產生即時預測。若要刪除用於產生即時預測的模型，請先刪除即時端點。

刪除 Amazon ML 物件（主控台）

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 選取您要刪除的 Amazon ML 物件。若要選取多個物件，請使用 SHIFT 鍵。若要取消選取所有已選的物件，請使用

   或

   按鈕。
3. 對於 Actions (動作)，請選擇 Delete (刪除)。
4. 在對話方塊中選擇 Delete (刪除)，刪除模型。

刪除具有即時端點的 Amazon ML 模型 （主控台）

1. 登入 AWS 管理主控台 並開啟位於 https：//https://console.aws.amazon.com/machinelearning/ 的 Amazon Machine Learning 主控台。
2. 選取您要刪除的模型。
3. 對於 Actions (動作)，選擇 Delete real-time endpoint (刪除即時端點)。
4. 選擇 Delete (刪除)，刪除端點。
5. 再選取一次模型。
6. 對於 Actions (動作)，請選擇 Delete (刪除)。
7. 選擇刪除，刪除模型。

刪除物件 (API)

您可以使用下列 API 呼叫來刪除 Amazon ML 物件：

- `DeleteDataSource` - 取得參數 `DataSourceId`。
- `DeleteMLModel` - 取得參數 `MLModelId`。
- `DeleteEvaluation` - 取得參數 `EvaluationId`。
- `DeleteBatchPrediction` - 取得參數 `BatchPredictionId`。

如需詳細資訊，請參閱《Amazon Machine Learning API 參考》。

使用 Amazon CloudWatch 指標監控 Amazon ML

Amazon ML 會將指標自動傳送至 Amazon CloudWatch，讓您可以收集與分析 ML 模型的使用統計資料。例如，若要追蹤批次與即時預測，您可以根據 RequestMode 維度監控 PredictCount 指標。這些指標會每五分鐘自動收集並傳送至 Amazon CloudWatch。您可以使用 Amazon CloudWatch 主控台、AWS CLI 或 AWS 開發套件來監控這些指標。

Amazon ML 指標是免費的，並會透過 CloudWatch 回報。如果您設定指標的警示，則會按標準 CloudWatch 費率向您收費。

如需詳細資訊，請參閱《Amazon CloudWatch 開發人員指南》中 Amazon CloudWatch 命名空間、維度與指標參考一節的 Amazon ML 指標清單。

使用 記錄 Amazon ML API 呼叫 AWS CloudTrail

Amazon Machine Learning (Amazon ML) 已與 整合 AWS CloudTrail，此服務提供使用者、角色或 Amazon ML 中 AWS 服務所採取動作的記錄。CloudTrail 會將 Amazon ML 的所有 API 呼叫擷取為事件。擷取的呼叫包括從 Amazon ML 主控台呼叫，以及對 Amazon ML API 操作的程式碼呼叫。如果您建立線索，您可以啟用 CloudTrail 事件持續交付至 Amazon S3 儲存貯體，包括 Amazon ML 的事件。即使您未設定追蹤，依然可以透過 CloudTrail 主控台中的事件歷史記錄檢視最新事件。使用 CloudTrail 收集的資訊，您可以判斷對 Amazon ML 提出的請求、提出請求的 IP 地址、提出請求的人員、提出請求的時間，以及其他詳細資訊。

若要進一步了解 CloudTrail，包括如何設定及啟用，請參閱《AWS CloudTrail 使用者指南》。

CloudTrail 中的 Amazon ML 資訊

建立 AWS 帳戶時，會在您的帳戶上啟用 CloudTrail。當 Amazon ML 中發生支援的事件活動時，該活動會記錄於 CloudTrail 事件，以及事件歷史記錄中的其他服務 AWS 事件。您可以在 AWS 帳戶中檢視、搜尋和下載最近的事件。如需詳細資訊，請參閱《使用 CloudTrail 事件歷史記錄檢視事件》https://docs.aws.amazon.com/awscloudtrail/latest/userguide/view-cloudtrail-events.html。

若要持續記錄您 AWS 帳戶中的事件，包括 Amazon ML 的事件，請建立追蹤。線索能讓 CloudTrail 將日誌檔案交付至 Amazon S3 儲存貯體。根據預設，當您在主控台建立追蹤記錄時，追蹤記錄會套用到所有 AWS 區域。追蹤會記錄 AWS 分割區中所有 區域的事件，並將日誌檔案交付至您指定的 Amazon S3 儲存貯體。此外，您可以設定其他 AWS 服務，以進一步分析 CloudTrail 日誌中收集的事件資料並對其採取行動。如需詳細資訊，請參閱下列內容：

- 建立追蹤的概觀
- CloudTrail 支援的服務和整合
- 設定 CloudTrail 的 Amazon SNS 通知
- 從多個區域接收 CloudTrail 日誌檔案，以及從多個帳戶接收 CloudTrail 日誌檔案

Amazon ML 支援將下列動作記錄為 CloudTrail 日誌檔案中的事件：

- AddTags
- CreateBatchPrediction
- CreateDataSourceFromRDS
- CreateDataSourceFromRedshift

- CreateDataSourceFromS3
- CreateEvaluation
- CreateMLModel
- CreateRealtimeEndpoint
- DeleteBatchPrediction
- DeleteDataSource
- DeleteEvaluation
- DeleteMLModel
- DeleteRealtimeEndpoint
- DeleteTags
- DescribeTags
- UpdateBatchPrediction
- UpdateDataSource
- UpdateEvaluation
- UpdateMLModel

下列 Amazon ML 操作使用包含登入資料的請求參數。將這些請求傳送至 CloudTrail 之前，登入資料會替換為三個星號 ("***")：

- CreateDataSourceFromRDS
- CreateDataSourceFromRedshift

使用 Amazon ML 主控台執行下列 Amazon ML 操作時，CloudTrail 日誌的 `RequestParameters`元件`ComputeStatistics`中不包含 屬性：

- CreateDataSourceFromRedshift
- CreateDataSourceFromS3

每一筆事件或日誌專案都會包含產生請求者的資訊。身分資訊可協助您判斷下列事項：

- 請求是使用根還是 AWS Identity and Access Management (IAM) 使用者登入資料提出。
- 提出該請求時，是否使用了特定角色或聯合身分使用者的暫時安全憑證。
- 請求是否由其他 AWS 服務提出。

如需詳細資訊，請參閱 CloudTrail userIdentity 元素。

範例：Amazon ML 日誌檔案項目

追蹤是一種組態，能讓事件以日誌檔案的形式交付到您指定的 Amazon S3 儲存貯體。CloudTrail 日誌檔案包含一或多個日誌專案。一個事件為任何來源提出的單一請求，並包含請求動作、請求的日期和時間、請求參數等資訊。CloudTrail 日誌檔並非依公有 API 呼叫的堆疊追蹤排序，因此不會以任何特定順序出現。

以下範例顯示的是展示 動作的 CloudTrail 日誌項目。

```
{
    "Records": [
        {
            "eventVersion": "1.03",
            "userIdentity": {
                "type": "IAMUser",
                "principalId": "EX_PRINCIPAL_ID",
                "arn": "arn:aws:iam::012345678910:user/Alice",
                "accountId": "012345678910",
                "accessKeyId": "EXAMPLE_KEY_ID",
                "userName": "Alice"
            },
            "eventTime": "2015-11-12T15:04:02Z",
            "eventSource": "machinelearning.amazonaws.com",
            "eventName": "CreateDataSourceFromS3",
            "awsRegion": "us-east-1",
            "sourceIPAddress": "127.0.0.1",
            "userAgent": "console.amazonaws.com",
            "requestParameters": {
                "data": {
                    "dataLocationS3": "s3://aml-sample-data/banking-batch.csv",
                    "dataSchema": "{\"version\":\"1.0\",\"rowId\":null,\"rowWeight
\":null,
                        \"targetAttributeName\":null,\"dataFormat\":\"CSV\",
                        \"dataFileContainsHeader\":false,\"attributes\":[
                          {\"attributeName\":\"age\",\"attributeType\":\"NUMERIC\"},
                          {\"attributeName\":\"job\",\"attributeType\":\"CATEGORICAL
\"},
                          {\"attributeName\":\"marital\",\"attributeType\":
\"CATEGORICAL\"},
```

```
                            {\"attributeName\":\"education\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"default\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"housing\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"loan\",\"attributeType\":\"CATEGORICAL
\"},
                            {\"attributeName\":\"contact\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"month\",\"attributeType\":\"CATEGORICAL
\"},
                            {\"attributeName\":\"day_of_week\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"duration\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"campaign\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"pdays\",\"attributeType\":\"NUMERIC\"},
                            {\"attributeName\":\"previous\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"poutcome\",\"attributeType\":
\"CATEGORICAL\"},
                            {\"attributeName\":\"emp_var_rate\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"cons_price_idx\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"cons_conf_idx\",\"attributeType\":
\"NUMERIC\"},
                            {\"attributeName\":\"euribor3m\",\"attributeType\":\"NUMERIC
\"},
                            {\"attributeName\":\"nr_employed\",\"attributeType\":
\"NUMERIC\"}
                        ],\"excludedAttributeNames\":[]}"
                },
                "dataSourceId": "exampleDataSourceId",
                "dataSourceName": "Banking sample for batch prediction"
            },
            "responseElements": {
                "dataSourceId": "exampleDataSourceId"
            },
            "requestID": "9b14bc94-894e-11e5-a84d-2d2deb28fdec",
            "eventID": "f1d47f93-c708-495b-bff1-cb935a6064b2",
            "eventType": "AwsApiCall",
```

```
            "recipientAccountId": "012345678910"
        },
        {
            "eventVersion": "1.03",
            "userIdentity": {
                "type": "IAMUser",
                "principalId": "EX_PRINCIPAL_ID",
                "arn": "arn:aws:iam::012345678910:user/Alice",
                "accountId": "012345678910",
                "accessKeyId": "EXAMPLE_KEY_ID",
                "userName": "Alice"
            },
            "eventTime": "2015-11-11T15:24:05Z",
            "eventSource": "machinelearning.amazonaws.com",
            "eventName": "CreateBatchPrediction",
            "awsRegion": "us-east-1",
            "sourceIPAddress": "127.0.0.1",
            "userAgent": "console.amazonaws.com",
            "requestParameters": {
                "batchPredictionName": "Batch prediction: ML model: Banking sample",
                "batchPredictionId": "exampleBatchPredictionId",
                "batchPredictionDataSourceId": "exampleDataSourceId",
                "outputUri": "s3://EXAMPLE_BUCKET/BatchPredictionOutput/",
                "mLModelId": "exampleModelId"
            },
            "responseElements": {
                "batchPredictionId": "exampleBatchPredictionId"
            },
            "requestID": "3e18f252-8888-11e5-b6ca-c9da3c0f3955",
            "eventID": "db27a771-7a2e-4e9d-bfa0-59deee9d936d",
            "eventType": "AwsApiCall",
            "recipientAccountId": "012345678910"
        }
    ]
}
```

標記您的 Amazon ML 物件

透過使用標籤將中繼資料指派給 Amazon Machine Learning (Amazon ML) 物件，來組織和管理這些物件。「標籤」是您針對物件所定義的索引鍵值組。

除了使用標籤來組織和管理 Amazon ML 物件之外，您還可以使用它們來分類和追蹤您的 AWS 成本。當您將標籤套用至 AWS 物件 (包含 ML 模型) 時，AWS 成本分配報告會包含依標籤彙總的使用情況和成本。套用代表商業類別的標籤 (例如成本中心、應用程式名稱或擁有者)，來整理多個服務中的成本。如需詳細資訊，請參閱《AWS Billing 使用者指南》中的將成本分配標籤用於自訂帳單報告。

目錄

- 標籤基本概念
- 標籤限制
- 標記 Amazon ML 物件 （ 主控台 ）
- 標記 Amazon ML 物件 (API)

標籤基本概念

使用標籤分類物件，讓您更輕鬆地進行管理。例如，您可以依用途、擁有者或環境來分類物件。然後，您可以定義一組標籤，協助您依擁有者和相關聯應用程式來追蹤模型。以下是數個範例：

- 專案：專案名稱
- 擁有者：名稱
- 用途：行銷預測
- 應用程式：應用程式名稱
- 環境：生產

您可以使用 Amazon ML 主控台或 API 來完成下列任務：

- 為物件新增標籤
- 檢視物件的標籤
- 編輯物件的標籤
- 刪除物件的標籤

根據預設，套用至 Amazon ML 物件的標籤會複製到使用該物件建立的物件。例如，如果 Amazon Simple Storage Service (Amazon S3) 資料來源具有「行銷成本：目標行銷活動」標籤，則使用該資料來源建立的模型也會具有「行銷成本：目標行銷活動」標籤，如同模型的評估。這可讓您使用標籤來追蹤相關物件，例如用於行銷活動的所有物件。如果標籤來源之間發生衝突，例如標籤為「行銷成本：目標行銷活動」的模型，以及標籤為「行銷成本：目標行銷客戶」的資料來源，Amazon ML 會套用模型中的標籤。

標籤限制

下列限制適用於標籤。

基本限制：

- 每個物件的標籤數量上限為 50。
- 標籤鍵與值皆區分大小寫。
- 您無法變更或編輯已刪除物件的標籤。

標籤鍵限制：

- 每個標籤鍵都必須是唯一的。如果您新增具有已使用鍵的標籤，則新的標籤會覆寫該物件的現有鍵值組。
- 您無法使用 `aws:` 來啟動標籤鍵，因為此字首保留供 AWS 使用。AWS 會代表您建立開頭為此字首的標籤，但您無法編輯或刪除它們。
- 標籤鍵的長度必須介於 1 到 128 個 Unicode 字元之間。
- 標籤鍵必須包含下列字元：Unicode 字母、數字、空格以及下列特殊字元：`_ . / = + - @`。

標籤值限制：

- 標籤值的長度必須介於 0 到 255 個 Unicode 字元之間。
- 標籤值可以空白。否則，它們必須包含下列字元：Unicode 字母、數字、空格以及下列任何特殊字元：`_ . / = + - @`。

標記 Amazon ML 物件（主控台）

您可以使用 Amazon ML 主控台檢視、新增、編輯和刪除標籤。

檢視物件的標籤 (主控台)

1. 登入 AWS 管理主控台 並開啟 Amazon Machine Learning 主控台，網址為 https://console.aws.amazon.com/machinelearning/：//。
2. 在導覽列中，展開區域選取器，然後選擇區域。
3. 在 Objects (物件) 頁面上，選擇任一物件。
4. 捲動至所選擇物件的 Tags (標籤) 區段。該物件的標籤會列在區段底部。

將標籤新增至物件 (主控台)

1. 登入 AWS 管理主控台 並開啟 Amazon Machine Learning 主控台，網址為 https://console.aws.amazon.com/machinelearning/：//。
2. 在導覽列中，展開區域選取器，然後選擇區域。
3. 在 Objects (物件) 頁面上，選擇任一物件。
4. 捲動至所選擇物件的 Tags (標籤) 區段。該物件的標籤會列在區段底部。
5. 選擇 Add or edit tags (新增或編輯標籤)。
6. 在 Add Tag (新增標籤) 的 Key (索引鍵) 欄位內指定標籤鍵，(選用) 在 Value (值) 欄位內指定標籤值，然後選擇 Apply changes (套用變更)。

   如果 Apply changes (套用變更) 按鈕未啟用，表示您所指定的標籤鍵或標籤值不符合標籤限制。如需詳細資訊，請參閱標籤限制。
7. 重新整理頁面，即可在 Tags (標籤) 區段的清單中看到新標籤。

編輯標籤 (主控台)

1. 登入 AWS 管理主控台 並開啟 Amazon Machine Learning 主控台，網址為 https://console.aws.amazon.com/machinelearning/：//。
2. 在導覽列中，展開區域選取器，然後選取區域。
3. 在 Objects (物件) 頁面上，選擇任一物件。
4. 捲動至所選擇物件的 Tags (標籤) 區段。該物件的標籤會列在區段底部。
5. 選擇 Add or edit tags (新增或編輯標籤)。
6. 在 Applied tags (已套用標籤) 下編輯 Value (值) 欄位內的標籤值，然後選擇 Apply changes (套用變更)。

如果 Apply changes (套用變更) 按鈕未啟用，表示您所指定的標籤值不符合標籤限制。如需詳細資訊，請參閱標籤限制。

7. 重新整理頁面，即可在 Tags (標籤) 區段的清單中看到已更新的標籤。

刪除物件中的標籤 (主控台)

1. 登入 AWS 管理主控台 並開啟 Amazon Machine Learning 主控台，網址為 https://console.aws.amazon.com/machinelearning/：//。
2. 在導覽列中，展開區域選取器，然後選擇區域。
3. 在 Objects (物件) 頁面上，選擇任一物件。
4. 捲動至所選擇物件的 Tags (標籤) 區段。該物件的標籤會列在區段底部。
5. 選擇 Add or edit tags (新增或編輯標籤)。
6. 在 Applied tags (已套用標籤) 下選擇您想要刪除的標籤，然後選擇 Apply changes (套用變更)。

標記 Amazon ML 物件 (API)

您可以使用 Amazon ML API 新增、列出和刪除標籤。如需範例，請參閱下列文件：

AddTags

新增或編輯所指定物件的標籤。

DescribeTags

列出所指定物件的標籤。

DeleteTags

刪除所指定物件中的標籤。

Amazon Machine Learning 參考

主題

- 授予 Amazon ML 許可從 Amazon S3 讀取您的資料
- 授予 Amazon ML 將預測輸出至 Amazon S3 的許可
- 控制 Amazon ML 資源的存取 - 使用 IAM
- 預防跨服務混淆代理人
- 非同步操作的相依性管理
- 檢查要求狀態
- 系統限制
- 所有物件的名稱和 ID
- 物件生命週期

授予 Amazon ML 許可從 Amazon S3 讀取您的資料

若要從您在 Amazon S3 中的輸入資料建立資料來源物件，必須將下列您存放輸入資料之 S3 位置的許可授予 Amazon ML：

- S3 儲存貯體上的 GetObject 許可與字首。
- S3 儲存貯體上的 ListBucket 許可。不同於其他動作，必須將儲存貯體範圍的許可 (而不是字首) 授予 ListBucket。不過，您可以使用 Condition 子句，將許可範圍限制為特定字首。

如果您使用 Amazon ML 主控台來建立資料來源，則可為您將這些許可新增至儲存貯體。在您完成精靈的各步驟時，系統會提示您確認是否要新增許可。以下政策範例示範如何授予許可，讓 Amazon ML 從範例位置 s3://*examplebucket*/*exampleprefix* 讀取資料，同時限制 ListBucket 許可的範圍只在 *exampleprefix* 輸入路徑。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
```

```
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        },
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:ListBucket",
            "Resource": "arn:aws:s3:::examplebucket",
            "Condition": {
                "StringLike": {
                    "s3:prefix": "exampleprefix/*"
                },
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        }
    ]
}
```

若要將此政策套用到您的資料，您必須編輯與您存放資料之 S3 儲存貯體相關聯的政策陳述式。

編輯 S3 儲存貯體的許可政策 (使用舊版主控台)

1. 登入 AWS 管理主控台 ，並在 https://console.aws.amazon.com/s3/ ：// 開啟 Amazon S3 主控台。
2. 選取您資料所在的儲存貯體名稱。
3. 選擇 Properties (屬性)。
4. 選擇 Edit bucket policy (編輯儲存貯體政策)。
5. 輸入上述政策，並依照您的需求加以自訂，然後選擇 Save (儲存)。
6. 選擇儲存。

編輯 S3 儲存貯體的許可政策 (使用新版主控台)

1. 登入 AWS 管理主控台 ，並在 https://console.aws.amazon.com/s3/ ：// 開啟 Amazon S3 主控台。
2. 選擇儲存貯體名稱，然後選擇 Permissions (許可)。
3. 選擇 Bucket Policy (儲存貯體政策)。
4. 輸入上述政策，並加以自訂以符合您的需求。
5. 選擇儲存。

授予 Amazon ML 將預測輸出至 Amazon S3 的許可

若要將批次預測操作的結果輸出至 Amazon S3，必須就「建立批次預測」操作時提供做為輸入的輸出位置授予 Amazon ML 以下許可：

- 您的 S3 儲存貯體與字首的 GetObject 許可。
- 您的 S3 儲存貯體與字首的 PutObject 許可。
- 您的 S3 儲存貯體與字首的 PutObjectAcl 許可。
  - Amazon ML 需要此許可，以確保在建立物件之後可將標準 ACL 儲存貯體擁有者完整控制許可授予您的 AWS 帳戶。
- S3 儲存貯體上的 ListBucket 許可。不同於其他動作，必須將儲存貯體範圍的許可 (而不是字首) 授予 ListBucket。不過，您可以使用 Condition 子句，將許可範圍限定為特定字首。

若使用 Amazon ML 主控台建立批次預測請求，可以將這些許可新增到您的儲存貯體。當您完成精靈中的步驟時，會提示您確認是否要新增這些許可。

以下範例政策示範如何授予許可，讓 Amazon ML 可將資料寫入範例位置 s3://examplebucket/exampleprefix，並將 ListBucket 許可範圍侷限在 exampleprefix 輸入路徑，以及授予 Amazon ML 對輸出字首設定放置物件 ACL 的許可：

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": [
                "s3:GetObject",
                "s3:PutObject"
            ],
            "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        },
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:PutObjectAcl",
            "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
```

```
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        },
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:ListBucket",
            "Resource": "arn:aws:s3:::examplebucket",
            "Condition": {
                "StringLike": {
                    "s3:prefix": "exampleprefix/*"
                },
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        }
    ]
}
```

若要將此政策套用到您的資料，您必須編輯與您存放資料之 S3 儲存貯體相關聯的政策陳述式。

編輯 S3 儲存貯體的許可政策 (使用舊版主控台)

1. 登入 AWS 管理主控台 ，並在 https://console.aws.amazon.com/s3/ ：// 開啟 Amazon S3 主控台。
2. 選取您資料所在的儲存貯體名稱。
3. 選擇 Properties (屬性)。
4. 選擇 Edit bucket policy (編輯儲存貯體政策)。
5. 輸入上述政策，並依照您的需求加以自訂，然後選擇 Save (儲存)。
6. 選擇儲存。

編輯 S3 儲存貯體的許可政策 (使用新版主控台)

1. 登入 AWS 管理主控台 ，並在 https://console.aws.amazon.com/s3/：// 開啟 Amazon S3 主控台。
2. 選擇儲存貯體名稱，然後選擇 Permissions (許可)。
3. 選擇 Bucket Policy (儲存貯體政策)。
4. 輸入上述政策，並加以自訂以符合您的需求。
5. 選擇儲存。

控制 Amazon ML 資源的存取 - 使用 IAM

AWS Identity and Access Management (IAM) 可讓您安全地控制使用者對 AWS 服務和資源的存取。使用 IAM，您可以建立和管理 AWS 使用者、群組和角色，並使用許可來允許和拒絕使用 IAM 搭配 Amazon Machine Learning (Amazon ML) 存取 AWS resources.By，您可以控制組織中的使用者是否可以使用特定 AWS 資源，以及是否可以使用特定 Amazon ML API 動作來執行任務。

IAM 可讓您：

- 在 AWS 帳戶底下建立使用者與群組。
- 將唯一安全登入資料指派給 AWS 帳戶下的每位使用者
- 控制每個使用者使用 AWS 資源執行任務的許可
- 與 AWS 帳戶中的使用者輕鬆共用您的 AWS 資源
- 為您的 AWS 帳戶建立角色和管理許可，以定義可擔任這些角色的使用者或服務
- 您可以在 IAM 中建立角色並管理許可，控制擔任該角色的實體或 AWS 服務可執行哪些操作。您也可以定義允許擔任該角色的實體。

如果您的組織已經有 IAM 身分，您可以使用它們來授予使用 AWS 資源執行任務的許可。

如需 IAM 的詳細資訊，請參閱《IAM 使用者指南》。

IAM 政策語法

IAM 政策為包含一或多個陳述式的 JSON 文件。每個陳述式結構如下：

```
{
    "Statement":[{
```

```
        "Effect":"effect",
        "Action":"action",
        "Resource":"arn",
        "Condition":{
            "condition operator":{
                "key":"value"
            }
        }
    }]
}
```

政策陳述式包含下列元素：

- 效果：控制使用資源和 API 動作 (您稍後將於陳述式中指定) 的許可。有效值為 Allow 和 Deny。根據預設，IAM 使用者沒有使用資源和 API 動作的許可，因此所有請求均會遭到拒絕。明確 Allow 會覆寫預設值。明確 Deny會覆寫任何 Allows。
- 動作：您授予或拒絕許可的特定 API 動作。
- Resource (資源)：受動作影響的資源。若要在陳述式中指定資源，您可以使用它的 Amazon Resource Name (ARN)。
- 條件 (選用)：控制政策生效的時機。

若要簡化建立和管理 IAM 政策，您可以使用 AWS 政策產生器 和 IAM 政策模擬器。

指定 Amazon MLAmazon ML 的 IAM 政策動作

在 IAM 政策陳述式中，您可以為支援 IAM 的任何服務指定 API 動作。當您為 Amazon ML API 動作建立政策陳述式時，請在 API 動作machinelearning:的名稱前面加上 ，如下列範例所示：

- machinelearning:CreateDataSourceFromS3
- machinelearning:DescribeDataSources
- machinelearning:DeleteDataSource
- machinelearning:GetDataSource

若要在單一陳述式中指定多個動作，請用逗號分隔：

```
"Action": ["machinelearning:action1", "machinelearning:action2"]
```

您也可以使用萬用字元指定多個動作。例如，您可以指定名稱開頭有 "Get" 文字的所有動作：

```
"Action": "machinelearning:Get*"
```

若要指定所有 Amazon ML 動作，請使用 * 萬用字元：

```
"Action": "machinelearning:*"
```

如需 Amazon ML API 動作的完整清單，請參閱 Amazon Machine Learning API 參考。

在 IAM 政策中指定 Amazon ML 資源ARNs

IAM 政策陳述式適用於一或多個資源。您可以根據資源的 ARN 來為政策指定資源。

若要指定 Amazon ML 資源ARNs，請使用下列格式：

"Resource": `arn:aws:machinelearning:region:account:resource-type/identifier`

以下範例說明如何指定常見的 ARN。

資料來源 ID：`my-s3-datasource-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:datasource/my-s3-datasource-id
```

ML 模型 ID：`my-ml-model-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/my-ml-model-id
```

批次預測 ID：`my-batchprediction-id`

```
"Resource":
arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/my-batchprediction-
id
```

評估 ID：`my-evaluation-id`

```
"Resource": arn:aws:machinelearning:<region>:<your-account-id>:evaluation/my-
evaluation-id
```

Amazon MLs的範例政策

範例 1：允許使用者讀取機器學習資源中繼資料

以下政策允許使用者或群組透過對指定的資源執行 DescribeDataSources、DescribeMLModels、DescribeBatchPredictions、DescribeEvaluations、GetDataSou 和 GetEvaluation 動作，以讀取資料來源、ML 模型、批次預測和評估的中繼資料。Describe* 操作許可不能限制在特定的資源上。

JSON

```
{ "Version":"2012-10-17",        "Statement": [ { "Effect": "Allow", "Action": [
    "machinelearning:Get*" ], "Resource": [
        "arn:aws:machinelearning:us-east-1:123456789012:datasource/S3-DS-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:datasource/REDSHIFT-DS-
ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:mlmodel/ML-MODEL-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:batchprediction/BP-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:evaluation/EV-ID1"
    ] }, { "Effect": "Allow", "Action": [ "machinelearning:Describe*" ],
 "Resource": [ "*" ] } ]
    }
```

範例 2：允許使用者建立機器學習資源

以下政策允許使用者或群組透過執行 CreateDataSourceFromS3、CreateDataSourceFromRedshift、CreateDataSourceFromRDS、Cr 和 CreateEvaluation 動作，來建立機器學習資料來源、ML 模型、批次預測和評估。您無法將這些動作的許可限制在特定資源上。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
```

```
        {
            "Effect": "Allow",
            "Action": [
                "machinelearning:CreateDataSourceFrom*",
                "machinelearning:CreateMLModel",
                "machinelearning:CreateBatchPrediction",
                "machinelearning:CreateEvaluation"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

範例 3：允許使用者在 ML 模型上建立和刪除即時端點，以及執行即時預測

以下政策允許使用者或群組透過在模型上執行 `CreateRealtimeEndpoint`、`DeleteRealtimeEndpoint` 和 `Predict` 動作，來為特定 ML 模型建立和刪除即時端點，以及執行即時預測。

JSON

```
{ "Version":"2012-10-17",       "Statement": [ { "Effect": "Allow", "Action": [
    "machinelearning:CreateRealtimeEndpoint",
 "machinelearning:DeleteRealtimeEndpoint",
    "machinelearning:Predict" ], "Resource": [
        "arn:aws:machinelearning:us-east-1:123456789012:mlmodel/ML-MODEL"
    ] } ] }
```

範例 4：允許使用者更新和刪除特定資源

以下政策允許透過提供許可給使用者或群組，讓他們在您帳戶的這些資源上執行 `UpdateDataSource`、`UpdateMLModel`、`UpdateBatchPrediction`、`UpdateEvaluation`、`DeleteDa` 和 `DeleteEvaluation` 動作，更新和刪除您 AWS 帳戶中的特定資源。

JSON

```
{ "Version":"2012-10-17",       "Statement": [ { "Effect": "Allow", "Action": [
```

```
    "machinelearning:Update*", "machinelearning:DeleteDataSource",
 "machinelearning:DeleteMLModel",
    "machinelearning:DeleteBatchPrediction",
 "machinelearning:DeleteEvaluation" ], "Resource": [
        "arn:aws:machinelearning:us-east-1:123456789012:datasource/S3-DS-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:datasource/REDSHIFT-DS-
ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:mlmodel/ML-MODEL-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:batchprediction/BP-ID1",
        "arn:aws:machinelearning:us-east-1:123456789012:evaluation/EV-ID1"
    ] } ] }
```

範例 5：允許任何 Amazon MLaction

下列政策允許使用者或群組使用任何 Amazon ML 動作。由於此政策會授予您所有機器學習資源的完整存取權，因此僅限管理員使用。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "machinelearning:*"
            ],
            "Resource": [
                "*"
            ]
        }
    ]
}
```

預防跨服務混淆代理人

混淆代理人問題屬於安全性問題，其中沒有執行動作許可的實體可以強制具有更多許可的實體執行該動作。在 中 AWS，跨服務模擬可能會導致混淆代理人問題。在某個服務 (呼叫服務) 呼叫另一個服務 (被呼叫服務) 時，可能會發生跨服務模擬。可以操縱呼叫服務來使用其許可，以其不應有存取許可的方式

對其他客戶的資源採取動作。為了預防這種情況， AWS 提供的工具可協助您保護所有服務的資料，而這些服務主體已獲得您帳戶中資源的存取權。

我們建議在資源政策中使用 aws:SourceArn和 aws:SourceAccount全域條件內容索引鍵，以限制 Amazon Machine Learning 為資源提供其他服務的許可。如果 aws:SourceArn 值不包含帳戶 ID (例如 Amazon S3 儲存貯體 ARN)，您必須使用這兩個全域條件內容金鑰來限制許可。如果同時使用這兩個全域條件內容金鑰，且 aws:SourceArn 值包含帳戶 ID，則在相同政策陳述式中使用 aws:SourceAccount 值和 aws:SourceArn 值中的帳戶時，必須使用相同的帳戶 ID。如果您想要僅允許一個資源與跨服務存取相關聯，則請使用 aws:SourceArn。如果您想要允許該帳戶中的任何資源與跨服務使用相關聯，請使用 aws:SourceAccount。

防範混淆代理人問題最有效的方法，是使用 aws:SourceArn 全域條件內容金鑰，以及資源的完整 ARN。如果不知道資源的完整 ARN，或者如果您指定了多個資源，請使用 aws:SourceArn 全域條件內容金鑰，同時使用萬用字元 (*) 表示 ARN 的未知部分。例如 arn:aws:*servicename*:*:*123456789012*:*。

下列範例顯示如何在 Amazon ML 中使用 aws:SourceArn和 aws:SourceAccount全域條件內容金鑰，以防止從 Amazon S3 儲存貯體讀取資料時出現混淆代理人問題。

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::examplebucket/exampleprefix/*",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        },
```

```
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "machinelearning.amazonaws.com"
            },
            "Action": "s3:ListBucket",
            "Resource": "arn:aws:s3:::examplebucket",
            "Condition": {
                "StringLike": {
                    "s3:prefix": "exampleprefix/*"
                },
                "StringEquals": {
                    "aws:SourceAccount": "123456789012"
                },
                "ArnLike": {
                    "aws:SourceArn": "arn:aws:machinelearning:us-
east-1:123456789012:*"
                }
            }
        }
    ]
}
```

非同步操作的相依性管理

Amazon ML 中的批次操作需依賴其他操作，才能成功完成。為了管理這些相依性，Amazon ML 會識別具有相依性的請求，並確認操作已完成。如果操作尚未完成，Amazon ML 會將初始請求放在一旁，直到其依賴的操作已完成。

批次操作之間有一些相依性。例如，在您可以建立 ML 模型之前，您必須先建立資料來源，以便用來訓練 ML 模型。如果沒有可用的資料來源，Amazon ML 無法訓練 ML 模型。

不過，Amazon ML 支援非同步操作的相依性管理。例如，您不需要等待資料統計資料已計算，才能傳送根據資料來源訓練 ML 模型的請求。反之，只要資料來源建立好，您就可以傳送使用資料來源訓練 ML 模型的請求。在資料來源統計資料已計算完成前，Amazon ML 不會實際開始訓練操作。createMLModel 請求會放入佇列中，直到統計資料計算完成；一旦完成後，Amazon ML 就會立即嘗試執行 createMLModel 操作。同樣地，您可以傳送尚未完成訓練的 ML 模型的批次預測和評估請求。

下表顯示繼續不同 AmazonML 動作的要求

為了...	您必須有...
建立 ML 模型 (createMLModel)	已計算資料統計資料的資料來源
建立批次預測 (createBatchPrediction)	資料來源 機器學習 (ML) 模型
建立批次評估 (createBatchEvaluation)	資料來源 機器學習 (ML) 模型

檢查要求狀態

當您提交請求時，您可以使用 Amazon Machine Learning (Amazon ML) API 檢查其狀態。例如，如果您提交`createMLModel`請求，您可以使用 `describeMLModel`呼叫來檢查其狀態。Amazon ML 會以下列其中一種狀態回應。

狀態	定義
待定	Amazon ML 正在驗證要求。 或 Amazon ML 正在等候運算資源成為可用，再執行要求。當您的帳戶超過並行執行之批次操作要求的最大數目時，就可能會發生這種情況。如果是這種情況，當其他執行中的請求已完成或被取消時，狀態就會轉換成 InProgress (進行中)。 或 Amazon ML 正在等候您的要求相依的批次操作完成。
INPROGRESS (進行中)	您的要求仍在執行中。
COMPLETED (已完成)	要求已完成，而且物件已備妥可供使用 (ML 模型與資料來源) 或檢視 (批次預測與評估)。

狀態	定義
失敗	您提供的資料有誤，或者您已取消操作。例如，如果您嘗試計算資料來源中無法完成的資料統計資料，您可能會收到 Invalid (無效) 或 Failed (失敗) 狀態訊息。此錯誤訊息說明操作未成功完成的原因。
DELETED	已刪除物件。

Amazon ML 也提供物件的相關資訊，例如 Amazon ML 何時完成建立該物件。如需詳細資訊，請參閱列出物件。

系統限制

為了提供穩定且可靠的服務，Amazon ML 對您向系統提出的請求施加一定的限制。大多數 ML 問題能夠輕鬆符合這些限制。不過，若您發現您對 Amazon ML 的使用受到這些限制所侷限，則可聯絡 AWS 客戶服務以申請提高限制。例如，對於可以同時執行的任務，可能有 5 個的限制。如果您發現由於此項限制，您的任務經常被排入佇列需等待資源，則提高您的帳戶的限制便可能有道理。

下表顯示 Amazon ML 中每個帳戶的預設限制。並非所有限制都可由 AWS 客戶服務提高。

限制類型	系統限制
每個觀察的大小	100 KB
訓練資料的大小 *	100 GB
批次預測輸入的大小	1 TB
批次預測輸出的大小 (記錄數量)	1 億
資料檔案 (結構描述) 中的變數數量	1,000
配方複雜性 (處理輸出變數的數量)	10,000
每個即時預測端點的 TPS 數	200
所有即時預測端點的 TPS 總數	10,000
所有即時預測端點的 RAM 總數	10 GB

限制類型	系統限制
同時任務的數量	25
任何任務的最長執行時間	7 天
多類別 ML 模型的類別數量	100
ML 模型大小	下限為 1 MB，上限為 2 GB
每個物件的標籤數量	50

- 限制資料檔案大小是為了確保可及時完成工作。已執行超過七天的工作會自動終止，產生 FAILED 狀態。

所有物件的名稱和 ID

Amazon ML 中的每個物件都必須有一個識別符，或 ID。Amazon ML 主控台會為您產生 ID 值，但如果您使用的是 API，就必須自己產生 ID 值。在您的 AWS 帳戶中，每個 ID 必須是同一類型所有 Amazon ML 物件的唯一 ID。也就是說，同一個 ID 不能有兩個評估。同一個 ID 可能會有一個評估和一個資料來源，但是不建議這麼做。

建議您使用物件隨機產生的識別符，前面加上較短的字串來識別其類型。例如，當 Amazon ML 主控台產生資料來源時，便會為此資料來源指派一組隨機、唯一的 ID，如「ds-zScWIuWiOxF」。這組 ID 隨機性相當足夠，能避免與任何單一使用者造成砥觸，同時也相當簡潔可讀。「ds-」字首只是為了便利性和明確性，並非必要不可。如果您不確定要使用什麼當做 ID 字串，建議您使用十六進位的 UUID 值 (如 28b1e915-57e5-4e6c-a7bd-6fb4e729cb23)，這種值在任何現代程式設計環境中都隨時可用。

ID 字串可以包含 ASCII 字母、數字、連字號和底線，最多可達 64 個字元。或許為了便利性考量，將中繼資料編碼至 ID 字串也是可以的。但是不建議您這麼做，因為一旦物件建立完成後，就無法改變其 ID。

物件名稱提供了簡單的方式，能讓您將使用者易用的中繼資料與每個物件建立關聯。您可以在建立物件之後更新名稱。這樣物件名稱就能反映 ML 工作流程的某些部分。例如，您一開始可將 ML 模型命名為「實驗 3 號」，之後將模型重新命名為「最終生產模型」。名稱可以是您想要的任何字串，最多 1,024 個字元。

物件生命週期

您使用 Amazon ML 建立的任何資料來源、ML 模型、評估或批次預測物件，都可在建立後讓您至少使用兩年。Amazon ML 可能會自動移除超過兩年未存取或使用的物件。

資源

以下相關資源可協助您使用此服務。

- Amazon ML 產品資訊 – 在中央位置擷取有關 Amazon ML 的所有相關產品資訊。
- Amazon ML FAQs – 涵蓋開發人員針對此產品提出的熱門問題。
- Amazon ML 範例程式碼 – 使用 Amazon ML 的範例應用程式。您可以使用範本程式碼作為建立您自己 ML 應用程式的起點。
- Amazon ML API 參考 – 詳細說明 Amazon ML 的所有 API 操作。它還為支援的 Web 服務協定提供了要求與回應的範例。
- AWS 開發人員資源中心 – 提供中央起點，以尋找文件、程式碼範例、版本備註和其他資訊，協助您使用 AWS 建置創新應用程式。
- AWS 訓練和課程 – 角色型和特殊課程和自定進度實驗室的連結，可協助您強化 AWS 技能並取得實際體驗。
- AWS 開發人員工具 – 開發人員工具和資源的連結，提供文件、程式碼範例、版本備註和其他資訊，協助您使用 AWS 建置創新應用程式。
- AWS Support Center – 建立和管理 AWS 支援案例的中樞。這也包含與其他實用資源的連結，例如論壇、常見技術問答集、服務運作狀態，以及 AWS Trusted Advisor。
- AWS Support – AWS Support 相關資訊的主要網頁，這是one-on-one、快速回應的支援管道，可協助您在雲端中建置和執行應用程式。
- 聯絡我們 – 有關 AWS 帳單、您的帳戶、事件、濫用和其他問題的查詢的中央聯絡窗口。
- AWS 網站條款 – 有關我們的著作權和商標、您的帳戶、授權和網站存取，以及其他主題的詳細資訊。

文件歷史記錄

下表說明此版本 Amazon Machine Learning (Amazon ML) 中文件的重要變更。

- API 版本：2015-04-09
- 文件最近更新日期：2016 年 8 月 2 日

變更	描述	變更日期
新增指標	此版本的 Amazon ML 新增了 Amazon ML 物件的新指標。 如需詳細資訊，請參閱列出物件。	2016 年 8 月 2 日
刪除多個物件	此版本的 Amazon ML 新增刪除多個 Amazon ML 物件的功能。 如需詳細資訊，請參閱刪除物件。	2016 年 7 月 20 日
新增標記	此 Amazon ML 版本新增了將標籤套用至 Amazon ML 物件的功能。 如需詳細資訊，請參閱標記您的 Amazon ML 物件。	2016 年 6 月 23 日
複製 Amazon Redshift 資料來源	此版本的 Amazon ML 會將 Amazon Redshift 資料來源設定複製到新的 Amazon Redshift 資料來源。 如需複製 Amazon Redshift 資料來源設定的詳細資訊，請參閱複製資料來源 (主控台)。	2016 年 4 月 11 日
新增隨機播放功能	此 Amazon ML 版本新增了隨機播放輸入資料的功能。 如需使用 Shuffle type (隨機播放類型) 參數的詳細資訊，請參閱培訓資料的隨機播放類型。	2016 年 4 月 5 日
使用 Amazon Redshift 改善資料來源建立	此版本的 Amazon ML 新增了在主控台中建立 Amazon ML 資料來源時測試 Amazon Redshift 設定的功能，以驗證連線是否有效。如需詳細資訊，請參閱使用 Amazon Redshift 資料建立資料來源 （ 主控台 ）。	2016 年 3 月 21 日

變更	描述	變更日期
改善 Amazon Redshift 資料結構描述轉換	此版本的 Amazon ML 可改善 Amazon Redshift (Amazon Redshift) 資料結構描述轉換為 Amazon ML 資料結構描述。 如需搭配 Amazon ML 使用 Amazon Redshift 的詳細資訊，請參閱 從 Amazon Redshift 中的資料建立 Amazon ML 資料來源。	2016 年 2 月 9 日
已新增 CloudTrail 記錄	此 Amazon ML 版本新增了使用 AWS CloudTrail (CloudTrail) 記錄請求的功能。 如需使用 CloudTrail 記錄的詳細資訊，請參閱 使用 記錄 Amazon ML API 呼叫 AWS CloudTrail。	2015 年 12 月 10 日
新增其他 DataRearrangement 選項	此 Amazon ML 版本新增了隨機分割輸入資料和建立補充資料來源的功能。 如需使用 `DataRearrangement` 參數的詳細資訊，請參閱 資料重新安排 。如需如何使用交叉驗證新選項的相關資訊，請參閱交叉驗證。	2015 年 12 月 3 日
嘗試即時預測	此版本的 Amazon ML 新增了在服務主控台中嘗試即時預測的能力。 如需嘗試即時預測的詳細資訊，請參閱《Amazon Machine Learning 開發人員指南要求即時預測》中的 。	2015 年 11 月 19 日
新 區域	此 Amazon ML 版本新增了對歐洲 （愛爾蘭） 區域的支援。 如需歐洲 （愛爾蘭） 區域 Amazon ML 的詳細資訊，請參閱《Amazon Machine Learning 開發人員指南區域與終端節點》中的 。	2015 年 8 月 20 日
初始版本	這是 Amazon ML 開發人員指南的第一個版本。	2015 年 4 月 9 日