IBM SPSS Modeler 18.2 資料庫內採礦手冊

IBM

請注意

使用本資訊及其所支援的產品之前，請先詳閱第 89 頁的『注意事項』中的資訊。

產品資訊

本版適用於 18.2.0 版 IBM SPSS Modeler 及所有後續版本與修訂版，除非新版中另有指示。

目錄

前言

IBM® SPSS® Modeler 是 IBM Corp. 企業能力資料採礦工作台。SPSS Modeler 可幫助組織透過深入瞭解資料來改善客戶和居民關係。組織利用從 SPSS Modeler 獲取的見解來留住有利的客戶、識別交叉銷售商機、吸引新客戶、偵測詐欺、降低風險，以及改進政府服務交付。

SPSS Modeler 的視覺化介面會邀請使用者應用他們特有的商業專門知識，帶來更強大的預測模型及縮短解決問題的時間。SPSS Modeler 提供許多建模技術，例如預測、分類、分割和關聯偵測演算法。建立模型後，IBM SPSS Modeler Solution Publisher 就能在企業層面交付給決策者或資料庫。

關於 IBM Business Analytics

IBM Business Analytics 軟體提供完整、一致且準確的資訊，決策者可信任此資訊，並藉以改善營運績效。商業智慧、預測分析、財務績效及策略管理，以及分析應用程式的綜合性產品組合會對現行績效提供清晰、即時而可行的洞察，且能夠預測未來結果。結合了豐富的業界解決方案、有效實證和專業服務，每種規模的組織都能引爆最高效能，確實自動化執行決策，並且交付更棒的成果。

作為此產品組合的一部分，IBM SPSS Predictive Analytics 軟體可協助組織預測未來事件，並根據促進較佳業務結果的洞察，主動採取行動。全球的商業、政府和學術客戶相當倚重 IBM SPSS 技術所帶來的競爭優勢，藉此做為吸引、保有和發展更多客戶，同時降低可能的不實詐欺風險。透過將 IBM SPSS 軟體引入其每天的作業，組織成為具有預測能力的企業，能夠直接或自動進行決策，以符合業務目標，並達成可測量的競爭優勢。如需更多資訊，或是聯絡代表人員，請造訪 http://www.ibm.com/spss。

技術支援

技術支援可用於維護客戶。客戶可能會聯絡技術支援，以取得使用 IBM Corp. 產品的協助，或其中一個受支援硬體環境的安裝協助。若要聯絡技術支援，請參閱 IBM Corp. 網站，網址為 http://www.ibm.com/support。請求協助時，請準備好識別您個人、組織和支援合約的相關資訊。

第 1 章 關於 IBM SPSS Modeler

IBM SPSS Modeler 是一組資料採礦工具，通過這些工具可以採用商業技術快速建立預測性模型，並將其應用於商業活動，從而改進決策過程。IBM SPSS Modeler 參照行業標準 CRISP-DM 模型設計而成，可支援從資料到更優商業成果的整個資料採礦過程。

IBM SPSS Modeler 提供擷取自機器學習人工智慧以及統計資料的各種建模方法。「建模」選用區上提供的方法可讓您根據資料衍生新資訊，以及開發預測模型。每種方法都具有特定的強度且最適合因應特定類型的問題。

SPSS Modeler 可以作為單獨產品購買，也可以作為用戶端與 SPSS Modeler Server 一起使用。同時提供了大量其他選項，下列各節將對這些選項進行概述。有關進一步資訊，請參閱https://www.ibm.com/analytics/us/en/technology/spss/。

IBM SPSS Modeler 產品

IBM SPSS Modeler 系列產品及關聯的軟體包括下列各項。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console（包含在 IBM SPSS Deployment Manager 中）
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 配接器

IBM SPSS Modeler

SPSS Modeler 是具有完整功能的產品，它安裝並執行於個人電腦上。您可以在本端方式作為單獨產品執行 SPSS Modeler，也可以在分佈方式下將其與 IBM SPSS Modeler Server 一起使用來提高大型資料集的效能。

借助 SPSS Modeler，您可以快速直接地建立準確的預測模型，而不進行程式設計。通過使用唯一可視介面，您可以輕鬆地查看資料採礦過程。借助該產品隨附的進階分析支援，您可以探索資料中先前隱藏的型樣和趨勢。您可以構建結果模型並瞭解影響結果的因素，從而利用業務機會並降低風險。

SPSS Modeler 推出了兩個版本：SPSS Modeler Professional 和 SPSS Modeler Premium。請參閱第 2 頁的『IBM SPSS Modeler 版本』主題，以取得更多資訊。

IBM SPSS Modeler Server

SPSS Modeler 使用用戶端/伺服器架構將資源集約型作業的要求分發給功能強大的伺服器軟體，因而使大資料集的傳輸速度大大加快。

SPSS Modeler Server 是一個個別授權的產品，在分佈分析方式下，該產品在安裝了一個或多個 IBM SPSS Modeler 的伺服器主機上持續執行。這種運行方式大大提高了 SPSS Modeler Server 對大型資料集的處理速度，因為在伺服器上可以運行耗用記憶體的作業，並且無需將資料下載到用戶端電腦上。IBM SPSS Modeler Server 還提供對 SQL 最佳化和資料庫內建模功能的支援，從而在效能和自動化方面帶來更多優勢。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是一個圖表使用者介面，用於管理多個 SPSS Modeler Server 配置選項，這些選項還可以通過選項檔案進行配置。主控台包含在 IBM SPSS Deployment Manager，可以用於監視和配置 SPSS Modeler Server 安裝，並且可供目前 SPSS Modeler Server 客戶免費使用。應用程式僅可以在 Windows 電腦上安裝；但它可以管理在任何受支援平台上安裝的伺服器。

IBM SPSS Modeler Batch

資料採礦通常是交互過程，因此，還可以從指令行執行 SPSS Modeler 而不需要圖形使用者介面。例如，您可能具有長時間執行或重複作業，並且希望在使用者不進行人為介入的情況下執行這些作業。SPSS Modeler Batch 是該產品的一個特殊版本，可提供對 SPSS Modeler 完整分析性能的支援，而無需存取一般的使用者介面。要使用 SPSS Modeler Batch，需要 SPSS Modeler Server。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher 是一個支持建立 SPSS Modeler 串流的打包版本的工具，該版本的串流可以由外部執行時期引擎執行或內含到外部應用程式中。通過這種方式，您可以發行和部署完整的 SPSS Modeler 串流以用於未安裝 SPSS Modeler 的環境。SPSS Modeler Solution Publisher 作為 IBM SPSS Collaboration and Deployment Services - 評分 服務的組成部分分發，需要個別的授權。通過此授權，您可以接收 SPSS Modeler Solution Publisher Runtime，它使您能夠執行已發佈的串流。

有關 SPSS Modeler Solution Publisher 的進一步資訊，請參閱 IBM SPSS Collaboration and Deployment Services 文件。IBM SPSS Collaboration and Deployment Services Knowledge Center 包含名為 "IBM SPSS Modeler Solution Publisher" 和 "IBM SPSS Analytics Toolkit" 的部分。

IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 配接器

IBM SPSS Collaboration and Deployment Services 的一些配接器使 SPSS Modeler 和 SPSS Modeler Server 能夠與 IBM SPSS Collaboration and Deployment Services 儲存庫進行交互。通過這種方式，部署到儲存庫的 SPSS Modeler 串流可以由多個使用者共用，或者從瘦用戶端應用程式 IBM SPSS Modeler Advantage 進行存取。請將配接器安裝在管理儲存庫的系統上。

IBM SPSS Modeler 版本

SPSS Modeler 推出了下列版本。

SPSS Modeler Professional

SPSS Modeler Professional 提供處理大多數類型的結構化資料所需要的所有工具，例如 CRM 系統中追蹤的行為和互動、個人背景資訊、採購行為和銷售資料。

SPSS Modeler Premium

SPSS Modeler Premium 是一個個別授權的產品，它對 SPSS Modeler Professional 進行了延伸，以便後者能夠處理專門的資料和非結構化文字資料。SPSS Modeler Premium 包含 IBM SPSS Modeler Text Analytics：

IBM SPSS Modeler Text Analytics 採用了先進語言技術和自然語言正在處理 (NLP)，以快速正在處理大量非結構化文字資料，擷取和群組織關鍵概念，以及將這些概念分為各式各樣的種類。擷取的概念和種類可以和現有結構化資料中進行已結合（例如人口統計學），並且可用於借助 IBM SPSS Modeler 的一整套資料採礦工具來進行建模，以此實現更好更集中的決策。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription 提供與傳統 IBM SPSS Modeler 用戶端完全相同的預測分析功能。使用 Subscription 版本時，您可以定期下載產品更新項目。

說明文件

文件可以從 SPSS Modeler 中的「說明」功能表獲取。這樣會開啟可在產品外部公開存取的 Knowledge Center。

作為產品下載的一部分，還會在個別的壓縮資料夾中以 PDF 格式提供每個產品的完整文件（包括安裝指示）。也可以從 Web 下載 PDF 文件，地址為：http://www.ibm.com/support/docview.wss?uid=swg27046871。

SPSS Modeler Professional 文件

SPSS Modeler Professional 文件套組（安裝指示除外）如下。

- **IBM SPSS Modeler 使用者手冊。**使用 SPSS Modeler 的一般簡介，包括如何建置資料串流、處理遺漏值、建置 CLEM 表示式，處理專案和報告，以及將用於部署的串流打包到 IBM SPSS Collaboration and Deployment Services 或 IBM SPSS Modeler Advantage。
- **IBM SPSS Modeler Source、Process 和 Output 節點。**說明用於以不同格式讀取、處理和輸出資料的所有節點。實際上這表示所有節點而非建模節點。
- **IBM SPSS Modeler Modeling 節點。**說明所有用於建立資料採礦模型的節點。IBM SPSS Modeler 提供擷取自機器學習人工智慧以及統計資料的各種建模方法。
- **IBM SPSS Modeler 應用程式手冊。**本手冊中的範例旨在為具體的建模方法和技術提供具有針對性的簡介。還可以在「說明」功能表中查閱本手冊的線上版本。請參閱第 4 頁的『應用程式範例』主題，以取得更多資訊。
- **IBM SPSS Modeler Python Scripting 和自動化。**通過編寫 Python Script 實現系統自動化的相關資訊，其中包含可以用於操作節點和串流的內容的資訊。
- **IBM SPSS Modeler 部署手冊。**有關在 IBM SPSS Deployment Manager 下以正在處理工作的步驟形式執行 IBM SPSS Modeler 串流的資訊。
- **IBM SPSS Modeler CLEF 開發人員手冊。** CLEF 提供了將第三方程式（例如，資料處理常式或建模演算法）作為節點整合到 IBM SPSS Modeler 的功能。
- **IBM SPSS Modeler 資料庫內挖掘手冊。**有關如何利用資料庫的功能通過第三方演算法來改進效能並增強分析功能的資訊。
- **IBM SPSS Modeler Server 管理和效能手冊。**提供有關如何配置和管理 IBM SPSS Modeler Server 的資訊。
- **IBM SPSS Deployment Manager 使用手冊。**有關使用 Deployment Manager 應用程式中包含的管理主控台使用者介面來監視和配置 IBM SPSS Modeler Server 的資訊。
- **IBM SPSS Modeler CRISP-DM 手冊。**借助 CRISP-DM 方法進行 SPSS Modeler 資料採礦的分步手冊。
- **IBM SPSS Modeler Batch 使用者手冊。**提供在批次模式下使用 IBM SPSS Modeler 的完整指導，包含批次模式執行和指令行引數的詳細資料。本手冊僅以 PDF 格式提供。

SPSS Modeler Premium 文件

SPSS Modeler Premium 文件套組（安裝指示除外）如下。

- **SPSS Modeler Text Analytics 使用者手冊。**提供有關將文字分析與 SPSS Modeler 配合使用的資訊，包括文字採集節點、互動式工作台、範本和其他資源。

應用程式範例

SPSS Modeler 中的資料採礦工具可以說明解決很多業務和組織問題，應用程式範例將提供有關特定建模方法和技術的簡要的針對性說明。此處使用的資料集比某些資料挖掘器管理的大量資料儲存庫小得多，但涉及的概念和方法可擴展到實際應用程式。

要存取範例，請在 SPSS Modeler 中按一下「說明」功能表中的**應用程式範例**。

資料檔案和樣本串流安裝在產品安裝目錄下的 Demos 資料夾中。如需相關資訊，請參閱『Demos 資料夾』。

資料庫建模範例。請參閱 *IBM SPSS Modeler* 資料庫內挖掘手冊 中的範例。

Scripting 範例。請參閱 *IBM SPSS Modeler Script* 編寫和自動化手冊 中的範例。

Demos 資料夾

與應用程式範例一起使用的資料檔案和樣本串流安裝在產品安裝目錄下的 Demos 資料夾中（例如：C:\Program Files\IBM\SPSS\Modeler\<version>\Demos）。可以從 Windows「開始」功能表上的 IBM SPSS Modeler 程式群組存取此資料夾，也可以通過按一下**檔案** > **開啟串流**對話框中最近的目錄的清單上的 Demos 來進行存取。

授權追蹤

當您使用 SPSS Modeler 時，系統會定期追蹤並記錄授權使用情況。所記錄的授權度量值為 *AUTHORIZED_USER* 和 *CONCURRENT_USER*，並且記錄的度量值類型取決於您針對 SPSS Modeler 具有的授權類型。

產生的日誌檔可由 IBM License Metric Tool 處理，通過該工具可產生授權使用情形報告。

授權日誌檔建立在記錄 SPSS Modeler 用戶端日誌檔的目錄（依預設為 %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log）中。

第 2 章 資料庫內採礦

資料庫建模概觀

IBM SPSS Modeler Server 支援與多家資料庫供應商的資料採礦和建模工具整合，這包含 IBM Netezza、Oracle Data Miner 和 Microsoft Analysis Services。您可以在資料庫內建置及儲存模型以及為模型評分 － 所有這些作業都是在 IBM SPSS Modeler 應用程式中進行。通過整合，可將 IBM SPSS Modeler 的分析功能和易用性將與資料庫的功能和效能相同分值合，同時還兼備資料庫供應商提供的資料庫自有演算法。模型在資料庫內創建，然後可以借助 IBM SPSS Modeler 介面以正常方式瀏覽模型並為之分數，必要時還可使用 IBM SPSS Modeler Solution Publisher 來對模型進行部署。在 IBM SPSS Modeler 的「資料庫建模」選用區中列出了受支援的演算法。

使用 IBM SPSS Modeler 存取資料庫自有演算法的若干優勢：

- 資料庫內的演算法常常與資料庫伺服器緊密整合，這可能有助於提高效能。
- 在「資料庫內」建立和儲存的模型不僅由可存取該資料庫的應用程式共用，且更易於在這些應用程式中部署。

SQL 產生。資料庫內建模與 SQL 產生（又稱為「SQL 回送」）存在明顯區別。使用此功能可以產生原生 IBM SPSS Modeler 作業的 SQL 陳述式，這些陳述式可以「回送」到資料庫（即，在其中執行）以提高效能。例如，「合併」、「聚合」和「選取」節點均可產生可以通過上述方式回送到資料庫的 SQL 代碼。將 SQL 產生與資料庫建模結合使用可以使串流自始至終在資料庫中執行，相比於在 IBM SPSS Modeler 中執行串流，前者具有極大的效能優勢。

註：資料庫建模和 SQL 最佳化需要在 IBM SPSS Modeler 電腦上啟用 IBM SPSS Modeler Server 連接。通過啟用此設定，您可以存取資料庫演算法，直接從 IBM SPSS Modeler 回送 SQL 以及存取 IBM SPSS Modeler Server。要驗證目前授權的狀態，請從 IBM SPSS Modeler 功能表中選擇下列項目。

說明 > **關於** > **其他詳細資訊**

如果啟用了連接，您可以在「授權狀態」標籤中看到選項**伺服器啟用**。

關於所受支援的演算法的更多資訊，請參閱針對指定供應商的後續章節。

需要項目

進行資料庫建模，需要進行下列設定：

- 在安裝了必要分析元件（Microsoft Analysis Services 或 Oracle Data Miner）的前提下，與相應資料庫建立 ODBC 連線).
- 在 IBM SPSS Modeler 中，必須在 Helper 應用程式對話框（**工具 > 說明應用程式**）中啟用資料庫建模。
- 應該啟用 IBM SPSS Modeler 以及 IBM SPSS Modeler Server（如果採用）中「使用者選項」對話框內的**產生 SQL** 和 **SQL 最佳化**設定。請注意，進行資料庫建模並非必須啟用 SQL 最佳化，但強烈建議您啟用此功能以提高效能。

註：資料庫建模和 SQL 最佳化需要在 IBM SPSS Modeler 電腦上啟用 IBM SPSS Modeler Server 連接。通過啟用此設定，您可以存取資料庫演算法，直接從 IBM SPSS Modeler 回送 SQL 以及存取 IBM SPSS Modeler Server。要驗證目前授權的狀態，請從 IBM SPSS Modeler 功能表中選擇下列項目。

說明 > **關於** > **其他詳細資訊**

如果啟用了連接，您可以在「授權狀態」標籤中看到選項**伺服器啟用**。

關於詳細資訊，請參閱針對指定供應商的後續章節。

模型建置

採用資料庫演算法建立模型和對模型評分的過程類似於 IBM SPSS Modeler 中其他類型的資料採礦。節點和建模「塊」的一般處理過程類似於 IBM SPSS Modeler 中其他的串流處理過程。唯一的區別是，實際正在處理和模型建置回送到資料庫中進行。

資料庫建模串流在概念上與 IBM SPSS Modeler 中的其他資料串流完全相同；但是，這個串流的所有操作均在資料庫中執行，例如，使用「Microsoft 決策樹」節點進行模型建立便是如此。執行串流時，IBM SPSS Modeler 會指示資料庫建立和儲存最終模型，而且詳細資料將下載到 IBM SPSS Modeler。資料庫中的執行由串流中使用的紫色陰影節點指示。

資料預備

無論是否使用了資料庫自有演算法，為了提高效能，應該盡可能將資料預備工作回送到資料庫完成。

- 如果原始資料儲存在資料庫中，那麼目標就是通過確保所有必要上游作業均可轉換為 SQL 使資料留在資料庫中。這樣可以避免將資料下載到 IBM SPSS Modeler，從而避免可能抵消增益的瓶頸，並容許在資料庫中執行整個串流。
- 如果原始資料 沒有儲存於資料庫，那麼仍可使用資料庫建模。此種情況下，將在 IBM SPSS Modeler 中進行資料預備，所準備的資料集將自動上傳到此資料庫並進行模型建置。

模型評分

採用資料庫中採礦從 IBM SPSS Modeler 中產生的模型與一般的 IBM SPSS Modeler 模型不同。雖然這些模型作為產生的模型「塊」顯示在模型管理器中，但實際上，它們是儲存在遠端資料挖掘或資料庫伺服器上的遠端模型。您在 IBM SPSS Modeler 中所看到的其實是對這些遠端模型的參照。換言之，您所看到的 IBM SPSS Modeler 模型是「空」模型，其中包含資料庫伺服器主機名、資料庫名和模型名等資訊。當對採用資料庫自有演算法建立的模型進行瀏覽和分數時，您應當清楚這個明顯差別。

建立模型後，您可以將其新增到串流並像其他所有在 IBM SPSS Modeler 中產生的模型一樣進行評分。所有評分均在資料庫中完成，即使上游作業並非如此。（可能的話，上游作業仍可能會被推回資料庫，以改善效能，但評分時並不一定要求這樣。）在大多數情況下，您還可以使用資料庫供應商提供的標準瀏覽器來瀏覽產生的模型。

為了同時進行瀏覽和評分，需要連線執行 Oracle Data Miner 或 Microsoft Analysis Services 的伺服器。

檢視結果和指定設定

要檢視結果以及指定評分設定，請在串流畫布中按兩下模型。您還可以選擇用滑鼠右鍵按一下此模型，然後選擇**瀏覽**或**編輯**。具體設定取決於模型的類型。

匯出和儲存資料庫模型

借助「檔案」功能表中的選項，可以從模型瀏覽器中匯出資料庫模型和摘要，就像匯出在 IBM SPSS Modeler 中建立的模型一樣。

1. 在模型瀏覽器的「檔案」功能表中，選擇下列某項：

- **匯出文字**將模型摘要匯出到文字檔

- **匯出 HTML** 將模型摘要匯出到 HTML 檔案
- **匯出 PMML**（僅受支援 IBM Db2 IM 模型）以預測模型標記語言 (PMML) 格式匯出模型，匯出的模型可以與其他 PMML 相容軟體配合使用。

註：還可以透過從「檔案」功能表中選擇**儲存節點**來儲存產生的模型。

模型一致性

對於產生的每個資料庫模型，IBM SPSS Modeler 會儲存一個模型結構說明，同時會以資料庫中的模型名稱來儲存一個模型參照。產生模型的「伺服器」標籤將顯示為此模型所產生的唯一索引鍵，此鍵與資料庫中的實際模型相符。

IBM SPSS Modeler 使用這個隨機產生的鍵來檢查模型是否仍然一致。這個鍵會在創建模型時儲存在模型說明中。最好在執行部署串流之前檢查鍵符合情況。

1. 要通過將資料庫中儲存的模型的說明與 IBM SPSS Modeler 儲存的隨機鍵進行比較來檢查該模型的一致性，請按一下**檢查**按鈕。如果未找到資料庫模型或索引鍵不相符，那麼系統將報錯。

檢視和匯出產生的 SQL

可以在執行前預覽所產生的 SQL 代碼，這可能有助於您進行除錯。

第 3 章 使用 Microsoft Analysis Services 進行資料庫建模

IBM SPSS Modeler 與 Microsoft Analysis Services

IBM SPSS Modeler支援與 Microsoft SQL Server Analysis Services 的整合。此功能作為 IBM SPSS Modeler 中的建模節點實現，並且可以從「資料庫建模」選用區上使用此功能。如果此選用區不可見，您可以通過啟用 MS Analysis Services 整合（位於 Helper 應用程式對話框的 Microsoft 標籤上）將其啟動。請參閱第 11 頁的『啟用與 Analysis Services 的整合』主題，以取得更多資訊。

IBM SPSS Modeler 支援整合下列 Analysis Services 演算法：

- 決策樹
- 叢集
- 關聯規則
- Naive Bayes
- 線性迴歸
- 神經網路
- 邏輯迴歸
- 時間序列
- 序列叢集作業

下圖說明從用戶端到伺服器的資料流程，其中資料庫內挖掘由 IBM SPSS Modeler Server 管理。模型建置使用 Analysis Services 進行。生成的模型由 Analysis Services 儲存。對此模型的參照在 IBM SPSS Modeler 串流中維護。然後，該模型從 Analysis Services 下載到 Microsoft SQL Server 或 IBM SPSS Modeler 中進行評分。

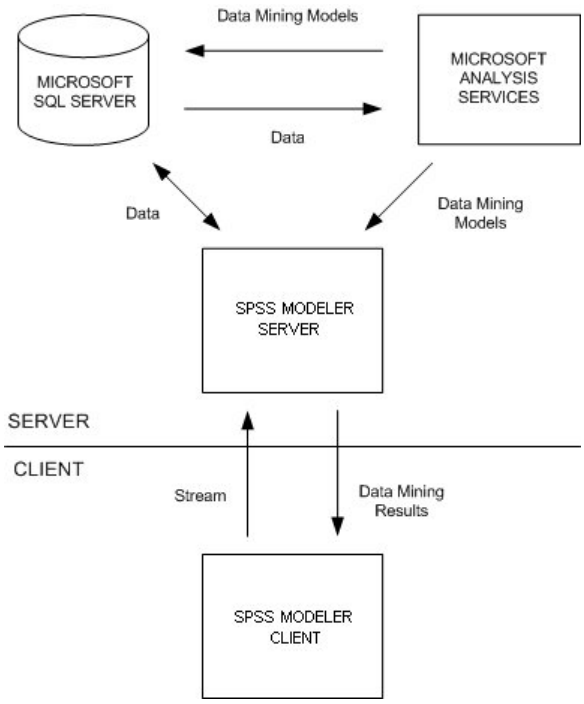

圖 *1.* 模型建置過程中，*IBM SPSS Modeler*、*Microsoft SQL Server* 與 *Microsoft Analysis Services* 之間的資料流程

注意：盡管可以使用 IBM SPSS Modeler Server，但它不是必要的。IBM SPSS Modeler 用戶端本身就能夠正在處理資料庫內挖掘計算。

與 Microsoft Analysis Services 進行整合的需求

下列是在 IBM SPSS Modeler 中使用 Analysis Services 演算法處理資料庫內建模的必備條件。您可能需要諮詢您的資料庫管理者以確保符合這些條件。

- 在 Windows 上安裝 IBM SPSS Modeler Server 後（分散式模式）執行 IBM SPSS Modeler。與 Analysis Services 的整合不支援 UNIX 平台。

  重要：IBM SPSS Modeler 使用者必須使用從『其他 IBM SPSS Modeler Server 需求』中列出的 URL 獲取的 Microsoft SQL Native Client 驅動程式來配置 ODBC 連線。建議您不要將 *IBM SPSS Data Access Pack* 隨附的驅動程式（一般推薦用於 *IBM SPSS Modeler* 的其他用途）用於此用途。驅動程式應配置為在啟用**與 Windows 鑑別整合**的條件下使用 SQL Server，因為 IBM SPSS Modeler 不支援 SQL Server 鑑別。如果您在建立或設定 ODBC 資料來源許可權方面存在問題，請聯絡您的資料庫管理者。
- 必須安裝 SQL Server，但不一定與 IBM SPSS Modeler 安裝在同一主機上。IBM SPSS Modeler 使用者必須具有足夠的權限以讀取寫入資料以及刪除和建立表格和視圖。

  註：建議您使用 SQL Server Enterprise Edition。Enterprise Edition 提供了用於調整演算法結果的進階參數，從而提供了更大的靈活性。Standard Edition 版本提供了相同的參數但不容許使用者編輯某些進階參數。
- Microsoft SQL Server Analysis Services 必須安裝在與 SQL Server 相同的主機上。

其他 IBM SPSS Modeler Server 需求

要在 IBM SPSS Modeler Server 中使用 Analysis Services 演算法，那麼必須在 IBM SPSS Modeler Server 主機上安裝下列元件。

註：如果 SQL Server 安裝在 IBM SPSS Modeler Server 所在的主機上，那麼這些元件已經可用。

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider（確保選取適合您作業系統的正確版本）
- Microsoft SQL Server Native Client（確保選取適合您操作系統的正確版本）
- 如果您使用的是 Microsoft SQL Server 2008 或 2012，那麼可能還需要安裝 Microsoft Core XML Services (MSXML) 6.0。

要下載這些元件，跳至 *www.microsoft.com/downloads*，搜尋 **.NET Framework** 或（對於所有其他元件）**SQL Server Feature Pack**，並選取您的 SQL Server 版本的最新的軟件包。

這些組件可能需要首先安裝其他套件，此類套件也可從 Microsoft 下載網站獲得。

其他 IBM SPSS Modeler 需求

要在 IBM SPSS Modeler 中使用 Analysis Services 演算法，必須安裝以上方元件，同時在用戶端新增下列元件：

- Microsoft SQL Server Datamining Viewer Controls（確保選取了適合您作業系統的正確版本）- 這還需要：
- Microsoft ADOMD.NET

要下載這些元件，跳至 *www.microsoft.com/downloads*，搜尋 **SQL Server Feature Pack**，並選取您的 SQL Server 版本的最新的軟件包。

註：資料庫建模和 SQL 最佳化需要在 IBM SPSS Modeler 電腦上啟用 IBM SPSS Modeler Server 連接。通過啟用此設定，您可以存取資料庫演算法，直接從 IBM SPSS Modeler 回送 SQL 以及存取 IBM SPSS Modeler Server。要驗證目前授權的狀態，請從 IBM SPSS Modeler 功能表中選擇下列項目。

說明 > **關於** > **其他詳細資訊**

如果啟用了連接，您可以在「授權狀態」標籤中看到選項**伺服器啟用**。

啟用與 Analysis Services 的整合

要啟用 IBM SPSS Modeler 與 Analysis Services 的整合，需要配置 SQL Server 和 Analysis Services，建立 ODBC 來源，在 IBM SPSS Modeler 的 Helper 應用程式對話框中啟用整合，並啟用 SQL 產生和最佳化。

註：Microsoft SQL Server 和 Microsoft Analysis Services 必須可用。請參閱第 10 頁的『與 Microsoft Analysis Services 進行整合的需求』主題，以取得更多資訊。

配置 SQL Server

配置 SQL Server 以便可以在資料庫內進行評分。

1. 在 SQL Server 主機上建立下列登錄鍵：

   ```
   HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP
   ```

2. 為該鍵新增如下 DWORD 值：

   ```
   AllowInProcess 1
   ```

3. 完成上述變更後，重新啟動 SQL Server。

配置 Analysis Services

必須首先在 Analysis Server「內容」對話框中手動配置兩項設定後，IBM SPSS Modeler 才能與 Analysis Services 進行通訊：

1. 通過 MS SQL Server Management Studio 登入到 Analysis Server。
2. 要存取「內容」對話框，請用滑鼠右鍵按一下伺服器名稱，然後選擇**內容**。
3. 選中**顯示進階（所有）內容**勾選框。
4. 變更下列內容：

- 將 `DataMining\AllowAdHocOpenRowsetQueries` 的值變更為 `True`（預設值為 `False`）。
- 將 `DataMining\AllowProvidersInOpenRowset` 的值變更為 `[all]`（無預設值）。

為 SQL Server 建立 ODBC DSN

若要讀取或寫入資料庫，您必須已針對相關資料庫，安裝並配置 ODBC 資料來源，並視需要具有讀取或寫入權。 Microsoft SQL Native Client ODBC 驅動程式是必要的，並且自動隨 SQL Server 一起安裝。建議您不要將 *IBM SPSS Data Access Pack* 提供的驅動程式（一般推薦用於 *IBM SPSS Modeler* 的其他用途）用於此用途。如果 IBM SPSS Modeler 和 SQL Server 駐留在不同的主機上，可以下載 Microsoft SQL Native Client ODBC 驅動程式。請參閱第 10 頁的『與 Microsoft Analysis Services 進行整合的需求』主題，以取得更多資訊。

如果您在建立或設定 ODBC 資料來源許可權方面存在問題，請聯絡您的資料庫管理者。

1. 通過使用 Microsoft SQL Native Client ODBC 驅動程式，建立一個 ODBC DSN，使其指向資料採礦過程中使用的 SQL Server 資料庫。餘下的驅動程式設定應使用預設設定。
2. 對於此 DSN，請確保已選取**使用整合的 Windows 鑑別**。

- 如果 IBM SPSS Modeler 和 IBM SPSS Modeler Server 執行在不同的主機上，請在每個主機上建立相同的 ODBC DSN。確保每台主機上使用的 DSN 名稱相同。

在 IBM SPSS Modeler 中啟用 Analysis Services 整合

要使 IBM SPSS Modeler 能夠使用 Analysis Services，首先必須在 Helper 應用程式對話框中提供伺服器指定資訊。

1. 從 IBM SPSS Modeler 功能表中選擇：

   工具 > **選項** > **Helper 應用程式**
2. 按一下 **Microsoft** 標籤。

- **啟用 Microsoft Analysis Services 整合**。啟用 IBM SPSS Modeler 視窗底部的「資料庫建模」選用區（如果尚未顯示）並為 Analysis Services 演算法新增節點。
- **Analysis Server 主機**。指定執行 Analysis Services 的機器的名稱。
- **Analysis Server 資料庫**。通過按一下省略號 (...) 按鈕開啟一個子對話框，在該對話框中，您可以從可用資料庫中選擇所需資料庫。清單中移入的資料庫都是可供指定 Analysis Server 使用的資料庫。由於 Microsoft Analysis Services 在指定資料庫中儲存資料採礦模型，因此應選取在其中儲存了由 IBM SPSS Modeler 建立的 Microsoft 模型的相應資料庫。
- **SQL 伺服器連線**。指定 DSN 資訊，SQL Server 資料庫使用此資訊來儲存要傳送到 Analysis Server 的資料。請選擇 ODBC 資料來源，以用來提供用於建立 Analysis Services 資料採礦模型的資料。如果您要根據純文字檔案或 ODBC 資料來源提供的資料建立 Analysis Services 模型，那麼此類資料將自動上傳到在此 ODBC 資料來源所指向的 SQL Server 資料庫中建立的暫時表格。
- **改寫資料採礦模型前發出警告**。選中此選項可以確保資料庫中儲存的模型不會在未經警告的情況下被 IBM SPSS Modeler 改寫。

註：可以在各個 Analysis Services 節點中覆蓋 Helper 應用程式對話框中所做的設定。

啟用 SQL 產生及最佳化

1. 從 IBM SPSS Modeler 功能表中選擇：

   工具 > **串流內容** > **選項**
2. 按一下導覽窗格中的**最佳化**選項。
3. 確認已啟用**產生 SQL** 選項。資料庫建模需要此設定才能夠運作。
4. 選中**最佳化 SQL 產生**和**最佳化其他執行**（非嚴格必要但強烈推薦使用，以使效能更優）。

使用 Analysis Services 建立模型

Analysis Services 模型建置要求訓練資料集位於 SQL Server 資料庫的表格或視圖中。如果資料不在 SQL Server 中，或者需要通過無法在 SQL Server 中執行的資料預備過程在 IBM SPSS Modeler 中進行處理，那麼此類資料將在模型建置前自動上傳到 SQL Server 中的暫時表格。

管理 Analysis Services 模型

通過 IBM SPSS Modeler 建立 Analysis Services 模型會在 IBM SPSS Modeler 中建立一個模型，然後在 SQL Server 資料庫中建立一個模型或取代其中一個模型。IBM SPSS Modeler 模型會參照儲存在資料庫伺服器中的資料庫模型的內容。 IBM SPSS Modeler 可通過將完全相同的產生模型鍵字串儲存在 IBM SPSS Modeler 模型和 SQL Server 模型中執行一致性檢查。

MS 決策樹建模節點可同時用於種類屬性和連續屬性的預測建模。對於種類屬性，此節點根據資料集中輸入欄之間的關係進行預測。例如，某案例要預測哪些顧客最有可能購買自行車，如果在年輕顧客中購買自行車的比例是十分之九，而在年紀較大的顧客中購買比例僅為十分之二，那麼該節點可推斷出年齡是有關自行車購買行為的良好預測值。決策樹可以根據此特定輸出結果的趨勢進行預測。對於連續屬性，此演算法將使用線性迴歸來確定決策樹分割位置。如果有一個以上的欄被設定為可預測的欄，或如果輸入資料包含一個被設定為可預測的巢套表格，那麼該節點可為每個可預測的欄建立個別的決策樹。

MS Clustering 建模節點使用疊代技術將某個資料集中的觀察值分組為包含類似性質的叢集。這些分組對於探索資料、識別資料異常和建立預測而言非常有用。叢集模型可以識別您無法通過表面觀測進行邏輯推導而獲得的資料集中的關係。例如，在邏輯上，您可以判斷騎自行車上下班的人的工作地點通常離家不遠。但是，此演算法可以找出騎自行車上下班的人員的其他不明顯特性。叢集節點與其他資料採礦節點的區別在於未指定目標欄位。叢集節點將通過資料中的關係和節點所識別的叢集對模型進行嚴格訓練。

MS 關聯規則建模節點對於推薦引擎十分有用。推薦引擎根據客戶已購買的項目或客戶表示有興趣的項目，向客戶推薦產品。關聯模型是以資料集為建置基礎，而資料集同時包含個別觀察值以及觀察值所含項目的 ID。觀察值中的項目群組稱為**項目集**。關聯模型包含一系列項目集，以及用來說明這些項目如何在觀察值中分組在一起的規則。演算法識別的規則可用來根據客戶購物車中已有的項目，來預測客戶未來可能購買的項目。

MS 貝式邏輯分類演算法 (Naive Bayes) 建模節點可計算目標欄位和預測值欄位之間的條件機率，並假定這些直欄是相互獨立的。此模型將所有建議預測變數視為彼此獨立，因此被稱為「樸素」。此方法比其他 Analysis Services 演算法的計算量小，因此對於在建模初期迅速探索關係非常有用。您可以套用此節點對資料執行初始探索，然後套用結果，以便建立含有其他計算時間可能更長但結果更為準確的節點的附加模型。

MS 線性迴歸建模節點是決策樹節點的變異，其中 MINIMUM_LEAF_CASES 參數被設定為大於或等於節點用來訓練挖掘模型的資料集中的案例總數。如果按上述方法設定參數，那麼該節點將永遠不會建立分割，因此可執行線性迴歸。

MS 神經網路建模節點類似於 MS 決策樹節點，即，當給定可預測屬性的每個狀態時，MS 神經網路節點會為輸入屬性的每個可能的狀態計算機率。之後，可以根據輸入屬性，使用這些機率對預測屬性的結果進行預測。

MS 邏輯迴歸建模節點是 MS 神經網路節點的變異，其中 HIDDEN_NODE_RATIO 參數設定為 0。此設定可建立不包含隱藏層的神經網路模型，因此相當於邏輯迴歸。

MS 時間序列建模節點提供的迴歸演算法對連續值（如產品銷售）在時間上的預測進行了最佳化。雖然其他 Microsoft 演算法（例如決策樹）需要更多的新資訊欄作為輸入才能預測趨勢，但「時間序列」模型卻非如此。「時間序列」模型可以僅根據用於建立模型的原始資料集來預測趨勢。您還可以在進行預測時向模型中新增新資料，並將新資料自動併入趨勢分析。請參閱第 16 頁的『MS 時間序列節點』主題，以取得更多資訊。

MS 序列叢集建模節點識別資料中的順序序列，並將此分析的結果與叢集技術結合以基於序列和其他屬性產生叢集。請參閱第 17 頁的『MS 序列叢集節點』主題，以取得更多資訊。

您可以從 IBM SPSS Modeler 視窗底部的「資料庫建模」選用區中存取每個節點。

對所有演算法節點通用的設定

下列設定通用於所有 Analysis Services 演算法。

伺服器選項

在「伺服器」標籤上，可以配置分析伺服器主機、資料庫和 SQL Server 資料來源。此處指定的選項將改寫 Helper 應用程式對話框的 Microsoft 標籤上指定的選項。請參閱第 11 頁的『啟用與 Analysis Services 的整合』主題，以取得更多資訊。

注意：對 Analysis Services 模型進行評分時，還可以使用此標籤的變體。請參閱第 18 頁的『Analysis Services 模型塊伺服器標籤』主題，以取得更多資訊。

模型選項

要建立最基本的模型，在進行處理前，需要在「模型」標籤上指定選項。評分方法和其他進階選項可在「專家」標籤上找到。

提供下列基本建模選項：

模型名稱。指定執行節點時指派給所建立模型的名稱。

- **自動填滿。** 根據目標或 ID 欄位名稱或模型類型名稱（如果未指定任何目標，例如叢集作業模型），來自動產生模型名稱。
- **自訂。** 讓您為所建立模型指定自訂名稱。

使用分割的資料。 將資料分割成多個不同的子集或樣本，以根據目前分割區欄位進行訓練、測試和驗證。通過使用一個樣本來建立模型並使用另一個樣本對模型進行測試，可以確定此模型適用於與現行資料類似的更大型資料集的程度。如果未在串流中指定分割區欄位，那麼將忽略此選項。

往下探查。如果顯示此選項，那麼您可以查詢模型以瞭解模型中所包含觀察值的詳細資料。

唯一欄位。 從下拉清單中，選取用於唯一地識別每個觀察值的欄位。通常，這個欄位為 ID 欄位，例如 **CustomerID**。

MS 決策樹專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 叢集專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 貝式邏輯分類演算法 (Naive Bayes)專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 線性迴歸專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 神經網路專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS邏輯迴歸專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 關聯規則節點

「MS 關聯規則」建模節點對於推薦引擎十分有用。推薦引擎根據客戶已購買的項目或客戶表示有興趣的項目，向客戶推薦產品。關聯模型是以資料集為建置基礎，而資料集同時包含個別觀察值以及觀察值所含項目的 ID。觀察值中的項目群組稱為**項目集**。

關聯模型包含一系列項目集，以及用來說明這些項目如何在觀察值中分組在一起的規則。演算法識別的規則可用來根據客戶購物車中已有的項目，來預測客戶未來可能購買的項目。

對於表格式資料，該演算法建立代表每個產生推薦 ($M-*field*) 的機率 ($MP-*field*) 的分數。對於交易處理格式資料，為支援 ($MS-*field*)、每個產生推薦 ($M-*field*) 的機率 ($MP-*field*) 和已調整機率 ($MAP-*field*) 建立分數。

需求

交易處理關聯模型的需求如下所示：

- **唯一欄位。** 關聯規則模型需要一個用於唯一地識別記錄的鍵。
- **ID 欄位。** 在建立具有交易處理格式資料的 MS 關聯規則模型時，用於識別每個交易的識別欄位為必填項。ID 欄位可以設定為與唯一欄位相同。
- **至少一個輸入欄位。** 相關規則演算法至少需要一個輸入欄位。
- **目標欄位。** 當建立具有交易資料的 MS 關聯模型時，目標欄位必須為交易欄位，例如使用者購買的產品。

MS 關聯規則專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

MS 時間序列節點

「MS 時間序列」建模節點支援兩種類型的預測：

- 未來
- 歷程記錄

未來預測評估在歷程資料結束之外若干指定時段的目標欄位值，並總是得到執行。**歷程預測**是在歷程資料中具有實際值的若干指定時段的評估目標欄位值。通過使用歷程預測，可以將實際歷程值與預測值進行比較，從而評估模型品質。預測起始點的值確定了是否執行歷程預測。

與 IBM SPSS Modeler 時間序列節點不同，MS 時間序列節點不需要提前的時間間隔節點。另一項區別是，依預設，僅針對預測的列生成分數，而不會針對時間序列資料中的所有歷程列生成分數。

需求

MS 時間序列模型的需求如下所示：

- **單個鍵時間欄位**。每個模型必須包含一個數值型或日期欄位，該欄位將用作觀察值數列並定義模型使用的時間塊。鍵時間欄位的資料類型可以是日期時間資料類型或數值資料類型。但是，此欄位必須包含連續的值，並且這些值對於每個系列必須唯一。
- **單個目標欄位**。在每個模型中，只能指定一個目標欄位。目標欄位的資料類型必須具有連續的值。例如，可以預測數值屬性（例如收入、銷售額或溫度）隨時間推移的變化情況。但是，無法使用包含種類值的欄位（例如採購狀態或教育程度）作為目標欄位。
- **至少一個輸入欄位。** MS 時間序列演算法需要至少一個輸入欄位。輸入欄位的資料類型必須具有連續的值。建立模型時，將忽略不連續的輸入欄位。
- **資料集必須已排序**。輸入資料集必須排序（在鍵時間欄位上），否則模型建置會因錯誤而中斷。

MS 時間序列模型選項

模型名稱。指定執行節點時指派給所建立模型的名稱。

- **自動填滿。** 根據目標或 ID 欄位名稱或模型類型名稱（如果未指定任何目標，例如叢集作業模型），來自動產生模型名稱。
- **自訂。** 讓您為所建立模型指定自訂名稱。

使用分割的資料。 如果定義了分割區欄位，那麼此選項可確保僅訓練分割的資料用於建立模型。

往下探查。如果顯示此選項，那麼您可以查詢模型以瞭解模型中所包含觀察值的詳細資料。

唯一欄位。 從下拉清單選取鍵時間欄位，該欄位用於建立時間序列模型。

MS 時間序列專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

如果要進行歷程預測，那麼可以併入在評分結果中的歷程步驟數由 (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP) 的值確定。依預設，此限制為 10，這表示只進行 10 項歷程預測。此時，例如當您在模型塊的「設定」標籤上為**歷程預測**輸入少於 -10 的值時，會發生錯誤（參見 第 20 頁的『MS 時間序列模型塊設定標籤』）。如果如需更多歷程預測，可以增大 HISTORIC_MODEL_COUNT 或 HISTORIC_MODEL_GAP 的值，但這將導致模型的建立時間延長。

MS 時間序列設定選項

啟動估計。 指定預測開始的時段。

- **啟動位置：新預測**。未來預測的開始時段，表示為相對於最後一個歷程資料時段的偏移值。例如，如果您的歷程資料在 12/99 結束，且您想在 01/00 開始預測，那麼應使用值 1；但如果您想在 03/00 開始預測，那麼應使用值 3。
- **啟動位置：歷程預測**。歷程預測的開始時段，表示為相對於最後一個歷程資料時段的負偏移值。例如，如果歷程資料在 12/99 結束，並且要對資料的最後一欄五個時段進行歷程預測，請使用值 -5。

結束估計。 指定預測停止的時段。

- **預測的結尾步驟**。預測的停止時段，表示為相對於最後一個歷程資料時段的偏移值。例如，如果歷程資料在 12/99 結束，並且您希望預測停止於 6/00，請在這裡使用值 6。對於未來預測，值必須總是大於或等於**啟動位置**值。

MS 序列叢集節點

MS 序列叢集節點使用一種序列分析演算法，該演算法探索包含可由下列路徑鏈結的事件的資料或序列。這方面的一些範例包括使用者在網站中進行導覽和瀏覽時建立的按一下路徑，或者顧客在網上零售店將商品新增到購物車的順序。演算法按照分組或叢集 找出最常見的序列和等同的序列。

需求

Microsoft 序列叢集模型的需求如下所示：

- **ID 欄位。** Microsoft 序列叢集演算法要求序列資訊以交易處理格式儲存。因此，用於識別每個交易的識別欄位為必填。
- **至少一個輸入欄位。** 此演算法至少需要一個輸入欄位。
- **序列欄位**。演算法還需要序列 ID 欄位，該欄位必須具有「連續」測量層次。例如，您可以使用網頁 ID、整數或字串，前提是此欄位按順序識別事件。每個序列僅容許一個序列 ID，每個模型中僅容許一種類型的序列。序列欄位不得與 ID 欄位和唯一欄位相同。
- **目標欄位。** 建立序列叢集模型時，目標欄位為必填。
- **唯一欄位。** 序列叢集模型需要一個用於唯一地識別記錄的鍵欄位。可以將唯一欄位設定為與 ID 欄位相同。

MS 序列叢集欄位選項

所有建模節點都有一個「欄位」標籤，您可以在其中指定要用於建立模型的欄位。

必須先指定要用作目標和輸入的欄位，然後才能建立序列叢集模型。請注意，對於「MS 序列叢集」節點，無法使用來自上游「類型」節點的欄位資訊；必須在此處指定欄位設定。

ID。 從清單中選取 ID 欄位。可以將數值或符號欄位用作 ID 欄位。此欄位的每一個唯一值都應指出一個特定的分析單位。例如，在購物籃應用程式中，每一個 ID 都可能代表一個客戶。對於 Web 日誌分析應用程式，每一個 ID 都可能代表一部電腦（依 IP 位址）或一位使用者（依登入資料）。

輸入。請為模型選取一個或多個輸入欄位。這些是包含序列建模所關注的事件的欄位。

序列。請從清單中選擇一個欄位用作序列 ID 欄位。例如，您可以使用網頁 ID、整數或字串，前提是此欄位按順序識別事件。每個序列僅容許一個序列 ID，每個模型中僅容許一種類型的序列。序列欄位不得與此標籤上指定的 ID 欄位以及「模型」標籤上指定的唯一欄位相同。

目標。 選擇一個欄位用作目標欄位，即您將基於序列資料嘗試預測其值的欄位。

MS 序列叢集專家選項

「專家」標籤上提供的選項根據所選串流的結構不同而有所變化。有關選定的 Analysis Services 模式節點的專家選項的詳細資料，請參閱使用者介面現場說明。

對 Analysis Services 模型評分

模型評分發生在 SQL Server 中，並由 Analysis Services 執行。如果資料源自 IBM SPSS Modeler 內或需要在 IBM SPSS Modeler 內準備，那麼可能需要將資料集上傳到暫時表格。您使用資料庫內挖掘從 IBM SPSS Modeler 建立的模型實際是儲存在遠端資料採礦或資料庫伺服器上的遠端模型。對使用 Microsoft Analysis Services 演算法建立的模型進行瀏覽和評分時，需要瞭解這項重要區別。

在 IBM SPSS Modeler 中，一般來講，只會遞送單一預測與關聯的機率或信賴度。

要獲取模型評分範例，請參閱第 20 頁的『Analysis Services 挖掘範例』。

對所有 Analysis Services 模型通用的設定

下列設定是所有 Analysis Services 模型的一般設定。

Analysis Services 模型塊伺服器標籤

「伺服器」標籤用於為資料庫內挖掘指定連線。此標籤還提供了唯一的模型鍵。當模型已建置並同時儲存在 IBM SPSS Modeler 中的模型內，以及「分析服務」資料庫內儲存之模型物件的說明內時，會隨機產生該索引鍵。

在「伺服器」標籤上，可以為評分作業配置分析伺服器主機和資料庫及 SQL Server 資料來源。在 IBM SPSS Modeler 中，此處指定的選項將改寫那些在 Helper 應用程式或「建立模型」對話框中指定的選項。請參閱第 11 頁的『啟用與 Analysis Services 的整合』主題，以取得更多資訊。

模型 GUID。模型鍵顯示在此處。當模型已建置並同時儲存在 IBM SPSS Modeler 中的模型內，以及「分析服務」資料庫內儲存之模型物件的說明內時，會隨機產生該索引鍵。

檢查。 按一下此按鈕將根據 Analysis Services 資料庫中儲存的模型中的鍵檢查模型鍵。此操作有助於驗證模型是否仍存在於分析伺服器中，並表示模型的結構未變更。

註：「檢查」按鈕僅適用於在準備評分時新增到串流畫布中的模型。如果檢查失敗，可調查此模型是否已被刪除或被伺服器上的其他模型取代。

檢視。 按一下此項可以開啟決策樹模型的圖表視圖。決策樹檢視器由 IBM SPSS Modeler 中的其他決策樹演算法所共用，且功能相同。

Analysis Services 模型塊彙總標籤

模型區塊的「摘要」標籤會顯示模型本身的相關資訊（分析），模型中使用的欄位（欄位），建置模型時使用的設定（建置設定）以及模型訓練（訓練摘要）。

當您第一次瀏覽節點時，「摘要」標籤結果會收合。若要查看相關結果，請使用項目左側的展開程式控制項來展開結果，或按一下**全部展開**按鈕以顯示所有結果。檢視完成後要隱藏結果時，請使用展開控制項來摺疊想要隱藏的具體結果，或者按一下**全部收合**按鈕來摺疊所有結果。

分析。 顯示特定模型的相關資訊。如果已執行附加到此模型塊的分析節點，那麼分析中的資訊也將顯示在此部分中。

欄位。 列出用來作為目標的欄位以及用來建置模型的輸入。

建置設定。 包含用來建置模型之設定的相關資訊。

訓練摘要。 顯示模型類型、用來建立模型的串流，模型建立者、模型建置時間以及建置模型的經歷時間。

MS 時間序列模型塊

MS 時間序列模型僅針對預測的時段生成分數，而不針對歷程資料生成分數。

下表格顯示新增到模型中的欄位。

表 *1.* 已新增至模型的欄位

欄位名稱	說明
$M-*field*	*field* 的預測值
$Var-*field*	*field* 的計算變異數
$Stdev-*field*	*field* 的標準差

MS 時間序列模型塊伺服器標籤

「伺服器」標籤用於為資料庫內挖掘指定連線。此標籤還提供了唯一的模型鍵。當模型已建置並同時儲存在 IBM SPSS Modeler 中的模型內，以及「分析服務」資料庫內儲存之模型物件的說明內時，會隨機產生該索引鍵。

在「伺服器」標籤上，可以為評分作業配置分析伺服器主機和資料庫及 SQL Server 資料來源。在 IBM SPSS Modeler 中，此處指定的選項將改寫那些在 Helper 應用程式或「建立模型」對話框中指定的選項。請參閱第 11 頁的『啟用與 Analysis Services 的整合』主題，以取得更多資訊。

模型 GUID。模型鍵顯示在此處。當模型已建置並同時儲存在 IBM SPSS Modeler 中的模型內，以及「分析服務」資料庫內儲存之模型物件的說明內時，會隨機產生該索引鍵。

檢查。 按一下此按鈕將根據 Analysis Services 資料庫中儲存的模型中的鍵檢查模型鍵。此操作有助於驗證模型是否仍存在於分析伺服器中，並表示模型的結構未變更。

註：「檢查」按鈕僅適用於在準備評分時新增到串流畫布中的模型。如果檢查失敗，可調查此模型是否已被刪除或被伺服器上的其他模型取代。

檢視。 按一下此項可以開啟「時間序列」模型的圖表視圖。Analysis Services 將整個模型顯示為樹狀結構。您還可以檢視圖，該圖將顯示目標欄位在一段時間內的歷程值以及預測的未來值。

有關進一步資訊，請參閱 MSDN 程式庫中對時間序列檢視器的說明，位置在 http://msdn.microsoft.com/en-us/library/ms175331.aspx。

MS 時間序列模型塊設定標籤

啟動估計。 指定預測開始的時段。

- **啟動位置：新預測**。未來預測的開始時段，表示為相對於最後一個歷程資料時段的偏移值。例如，如果您的歷程資料在 12/99 結束，且您想在 01/00 開始預測，那麼應使用值 1；但如果您想在 03/00 開始預測，那麼應使用值 3。
- **啟動位置：歷程預測**。歷程預測的開始時段，表示為相對於最後一個歷程資料時段的負偏移值。例如，如果歷程資料在 12/99 結束，並且要對資料的最後一欄五個時段進行歷程預測，請使用值 -5。

結束估計。 指定預測停止的時段。

- **預測的結尾步驟**。預測的停止時段，表示為相對於最後一個歷程資料時段的偏移值。例如，如果歷程資料在 12/99 結束，並且您希望預測停止於 6/00，請在這裡使用值 6。對於未來預測，值必須總是大於或等於**啟動位置**值。

MS 序列叢集模型塊

下表格顯示新增到 MS 序列叢集模型中的欄位（其中 *field* 是目標欄位的名稱）。

表 2. 已新增至模型的欄位

欄位名稱	說明
$MC-*field*	此序列所屬的叢集的預測。
$MCP-*field*	此序列的所預測叢集的機率。
$MS-*field*	*field* 的預測值
$MSP-*field*	$MS-*field* 值正確的機率。

匯出模型和產生節點

可以將模型摘要和結構匯出到文字格式的檔案和 HTML 格式的檔案。適當時，可以產生適當的「選取」和「過濾器」節點。

與 IBM SPSS Modeler 中的其他模型塊類似，Microsoft Analysis Services 模型塊支援直接產生記錄和欄位作業節點。使用模型塊的「產生」功能表選項，可以產生下列節點：

- 選取節點（僅當在「模型」標籤上已選取某項目時）
- 過濾器節點

Analysis Services 挖掘範例

其中包含多個樣本串流，這些樣本串流展示了如何搭配 IBM SPSS Modeler 使用 MS Analysis Services 資料採礦。這些串流位於 IBM SPSS Modeler 安裝資料夾中的以下位置：

*Demos**Database_Modelling**Microsoft*

附註：「展示」資料夾可從 Windows「啟動」功能表的 IBM SPSS Modeler 程式集存取。

範例串流：決策樹

下列串流按順序一起使用可作為使用由 MS Analysis Services 提供的決策樹演算法的資料庫挖掘過程的範例。

表 *3.* 決策樹 - 串流範例

串流	說明
1_upload_data.str	用來清除純文字檔中的資料以及將資料上傳至資料庫。
2_explore_data.str	提供使用 IBM SPSS Modeler 來探索資料的範例
3_build_model.str	使用資料庫的原生演算法來建置模型。
4_evaluate_model.str	用來作為使用 IBM SPSS Modeler 來評估模型的範例
5_deploy_model.str	部署用來在資料庫內進行評分的模型。

附註：若要執行該範例，必須依序執行串流。此外，還必須更新每個串流中的來源與建模節點，以參照您要使用之資料庫的有效資料來源。

串流範例中所使用的資料集涉及信用卡應用程式，並呈現類別預測值與連續預測值之混合發生的分類問題。如需此資料集的相關資訊，請參閱串流範例所在資料夾中的 *crx.names* 檔。

可從 *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* 的 UCI 機器學習儲存庫中取得此資料集。

串流範例：上傳資料

第 1 個範例串流 *1_upload_data.str* 用於清理純文字檔中的資料並將其上傳到 SQL Server。

由於 Analysis Services 資料採礦需要鍵欄位，因而這個初始串流通過 IBM SPSS Modeler 的 `@INDEX` 函數，使用「衍生」節點將名為 *KEY* 的新欄位新增到資料集中，其唯一值為 *1*、*2* 和 *3*。

隨後的填入節點用於遺漏值處理，並將從文字檔 *crx.data* 中讀取的空白欄位取代為空白值。

串流範例：探索資料

第二個串流範例 *2_explore_data.str* 用來示範如何使用「資料審核」節點來取得資料的一般概觀，其中包括摘要統計資料與圖形。

按兩下「資料審核報告」中的圖形可產生更詳細的圖形，以便更深入地探索給定的欄位。

串流範例：建置模型

第三個串流範例 *3_build_model.str* 說明在 IBM SPSS Modeler 中建置模型。可將資料庫模型附加到串流並通過按兩下指定建構設定值。

在此對話框的「模型」標籤上，可以指定下列設置：

1. 選取 **Key** 欄位作為唯一 ID 欄位。

在「專家」標籤上，可以微調設定以建立模型。

在執行之前，請確保指定正確的資料庫用於模型建置。使用「伺服器」標籤調整設定。

串流範例：評估模型

第四個串流範例 *4_evaluate_model.str* 說明使用 IBM SPSS Modeler 在資料庫內建模的優點。執行該模型之後，可以將其新增回資料串流，並使用 IBM SPSS Modeler 中提供的幾個工具評估該模型。

檢視建模結果

您可以按兩下模型塊以探索結果。「摘要」標籤提供了結果的規則樹狀結構視圖。還可以按一下**視圖**按鈕（位於「伺服器」標籤上）來檢視決策樹模型的圖表視圖。

評估模型結果

樣本串流中的「分析」節點建立一個重合矩陣，以顯示每個預測欄位與其目標欄位之間的相符型樣。然後，執行「分析」節點以檢視結果。

樣本串流中的「評估」節點可以建立一個收益圖表，用於顯示模型對精確性的提高。然後，執行「評估」節點以檢視結果。

串流範例：部署模型

您對模型的精確度感到滿意以後，便可以部署模型以用於外部應用程式或發佈回資料庫。在最後一個範例串流 *5_deploy_model.str* 中，將從表格 CREDIT 中讀取資料，然後進行評分，再使用資料庫匯出節點將資料發佈到表格 CREDITSCORES。

執行這個串流將產生下列 SQL：

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
 T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
 T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
 T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
 SELECT CONVERT(NVARCHAR,[TA].[field1])  AS C0, CONVERT(NVARCHAR,[TA].[field2])  AS C1,
  [TA].[
field3] AS C2, CONVERT(NVARCHAR,[TA].[field4])  AS C3,
  CONVERT(NVARCHAR,[TA].[
field5])  AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
  CONVERT(NVARCHAR,[TA].[
field7])  AS C6, [TA].[field8] AS C7,
  CONVERT(NVARCHAR,[TA].[
field9])  AS C8, CONVERT(NVARCHAR,[TA].[field10])  AS C9,
  [TA].[
field11] AS C10, CONVERT(NVARCHAR,[TA].[field12])  AS C11,
  CONVERT(NVARCHAR,[TA].[
field13])  AS C12, [TA].[field14] AS C13,
  [TA].[
field15] AS C14, CONVERT(NVARCHAR,[TA].[field16])  AS C15,
  [TA].[
KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16])  AS C17,
  [TA].[$MC-field16] AS C18
 FROM openrowset('MSOLAP',
  'Datasource=localhost;Initial catalog=FoodMart 2000',
  'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
   [T].[
C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
   [T].[
C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
   [T].[
C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
   [T].[
C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
   [T].[
C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
   PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
  FROM [CREDIT1] PREDICTION JOIN
   openrowset(''MSDASQL'',
   ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
   T0."
field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
   T0."
field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
   T0."
```

```
field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
   T0."
field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
   T0."
KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
  ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
   及 [T].[
C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
   及 [T].[
C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
   及 [T].[
C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

第 4 章 使用 Oracle Data Mining 構建資料庫模型

關於 Oracle Data Mining

IBM SPSS Modeler 支援與 Oracle Data Mining (ODM) 的整合，ODM 提供了緊密內嵌於 Oracle RDBMS 中的一系列資料採礦演算法。這些功能可透過 IBM SPSS Modeler 圖形使用者介面和工作流程導向的開發環境進行存取，使您可以直接在 IBM Netezza 環境中執行資料採礦演算法。

IBM SPSS Modeler 支援整合 Oracle Data Mining 的下列演算法：

- Naive Bayes
- Adaptive Bayes
- 支援向量機器 (SVM)
- 通用性線性模型 (GLM)*
- 決策樹
- O-叢集
- k-Means
- 非負矩陣分解 (NMF)
- Apriori
- 下限描述子長度 (MDL)
- 屬性重要性 (AI)

* 僅限於 11*g* R1

與 Oracle 進行整合的需求

下列是使用 Oracle Data Mining 處理資料庫內建模的必備項目。您可能需要諮詢您的資料庫管理者以確保符合這些條件。

- 以本機模式或在 Windows 或 UNIX 上安裝 IBM SPSS Modeler Server 後執行 IBM SPSS Modeler。
- 帶有 Oracle Data Mining 選項的 Oracle 10 *g* R2 或 11 *g* R1（10.2 或更高版本的資料庫）。

註：10*g*R2 支援除通用性線性模型（需要 11*g*R1）以外的所有資料庫建模演算法。

- 連接 Oracle（ 如下方所述 ）的 ODBC 資料來源。

註：資料庫建模和 SQL 最佳化需要在 IBM SPSS Modeler 電腦上啟用 IBM SPSS Modeler Server 連接。通過啟用此設定，您可以存取資料庫演算法，直接從 IBM SPSS Modeler 回送 SQL 以及存取 IBM SPSS Modeler Server。要驗證目前授權的狀態，請從 IBM SPSS Modeler 功能表中選擇下列項目。

說明 > **關於** > **其他詳細資訊**

如果啟用了連接，您可以在「授權狀態」標籤中看到選項**伺服器啟用**。

啟用與 Oracle 的整合

要啟用 IBM SPSS Modeler 與 Oracle Data Mining 的整合，需要配置 Oracle，建立 ODBC 來源，在 IBM SPSS Modeler 的 Helper 應用程式對話框中啟用整合，並啟用 SQL 產生和最佳化。

配置 Oracle

要安裝和配置 Oracle Data Mining，請參閱 Oracle 文件（特別是 *Oracle Administrator's Guide*）以獲得更多詳細資料。

為 Oracle 建立 ODBC 來源

要啟用 Oracle 和 IBM SPSS Modeler 之間的連線，您需要建立 ODBC 系統資料來源名稱 (DSN)。

在建立 DSN 之前，您應該基本瞭解 ODBC 資料來源與驅動程式以及 IBM SPSS Modeler 中的資料庫支援。

如果您針對 IBM SPSS Modeler Server 以分散式模式執行，請在伺服器電腦上建立 DSN。如果您以本端（用戶端）模式執行，請在用戶端電腦上建立 DSN。

1. 安裝 ODBC 驅動程式。您可在此版本隨附的 IBM SPSS Data Access Pack 安裝盤上找到這些驅動程序。執行 *setup.exe* 檔案以啟動安裝程式，並選取所有相關的驅動程式。請按畫面上的指示執行操作，以安裝驅動程式。
   a. 建立 DSN。

      註：功能表序列隨 Windows 版本不同而有所變化。

      - **Windows XP**。 從「開始」功能表中選擇**控制台**。按兩下**系統管理工具**，然後按兩下**資料來源 (ODBC)**。
      - **Windows Vista**。從「開始」功能表中選擇**控制台**，然後選擇**系統維護**。按兩下**系統管理工具**，選取**資料來源 (ODBC)**，然後按一下**開啟**。
      - **Windows 7**。從「開始」功能表中選擇**控制台**，選擇**系統和安全**，然後選擇**系統管理工具**。選取**資料來源 (ODBC)**，然後按一下**開啟**。

   b. 跳至**系統 DSN** 標籤，然後按一下**新增**。
2. 選取 **SPSS OEM 6.0 Oracle Wire Protocol** 驅動程式。
3. 按一下**完成**。
4. 在「ODBC Oracle Wire Protocol 驅動程式安裝」畫面中，輸入選擇的資料來源名稱、Oracle 伺服器的主機名稱、連線埠號及使用的 Oracle 實例的 SID。

   如果已使用 *tnsnames.ora* 檔案實現了 TNS，那麼可以從伺服器上的 *tnsnames.ora* 檔案獲取主機名稱、埠和 SID。要獲取進一步資訊，請與 Oracle 管理者聯絡。
5. 請按一下**測試**按鈕以測試連線。

在 IBM SPSS Modeler 中啟用 Oracle Data Mining 整合

1. 從 IBM SPSS Modeler 功能表中選擇：

   工具 > **選項** > **Helper 應用程式**
2. 按一下 **Oracle** 標籤。

啟用 Oracle Data Mining 整合。啟用 IBM SPSS Modeler 視窗底部的「資料庫建模」選用區（如果尚未顯示）並為 Oracle Data Mining 演算法新增節點。

Oracle 連線。指定用於建立和儲存模型的預設 Oracle ODBC 資料來源以及有效的使用者名稱和密碼。可在各個建模節點和模型塊上併入。

註：用於建模的資料庫連線可以與用於存取資料的連線相同，也可以不相同。例如，您可能有一個串流是從某個 Oracle 資料庫存取資料，將資料下載至 IBM SPSS Modeler 以進行清除或其他操作，然後將資料上傳至其他 Oracle 資料庫進行建模。另外，原始資料也可以位於平面檔案或其他（非 Oracle）來源中，但在這種情況下，需要將資料上傳到 Oracle 才能進行建模。所有情況下資料都將自動上傳到用於建模的資料庫中建立的暫時表格。

改寫 Oracle Data Mining 模型前發出警告。選中此選項可以確保資料庫中儲存的模型不會在未經警告的情況下被 IBM SPSS Modeler 覆蓋。

列出 Oracle Data Mining 模型。顯示可用的資料採礦模型。

允許啟動 Oracle Data Miner。**（選用）**啟用該選項後，IBM SPSS Modeler 便可以啟動 Oracle Data Miner 應用程式。請參閱 第 41 頁的『Oracle 資料採礦程式』 以瞭解進一步資訊。

Oracle Data Miner 執行檔的路徑。**（選用）**用於指定 Oracle Data Miner for Windows 執行檔的實體位置（例如 *C:\odm\bin\odminerw.exe*）。Oracle Data Miner 不會隨著 IBM SPSS Modeler 一起安裝，必須從 Oracle 網站 (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) 下載正確的版本並在用戶端進行安裝。

啟用 SQL 產生及最佳化

1. 從 IBM SPSS Modeler 功能表中選擇：

   工具 > **串流內容** > **選項**
2. 按一下導覽窗格中的**最佳化**選項。
3. 確認已啟用**產生 SQL** 選項。資料庫建模需要此設定才能夠運作。
4. 選中**最佳化 SQL 產生**和**最佳化其他執行**（非嚴格必要但強烈推薦使用，以使效能更優）。

使用 Oracle Data Mining 建立模型

Oracle 建模節點的工作方式與 IBM SPSS Modeler 中其他建模節點的一樣，不過也有幾個例外狀況。可通過橫向顯示在 IBM SPSS Modeler 視窗底部的資料庫建模選用區來存取這些節點。

資料考量

Oracle 要求以字串格式（字元 或 VARCHAR2）儲存種類資料。因此，IBM SPSS Modeler 不容許將測量層次為旗標或名義（種類）的數值儲存欄位指定為 ODM 模型的輸入。如有必要，可在 IBM SPSS Modeler 中使用「再分類」節點將數字轉換為字串。

目標欄位。 只能選取一個欄位作為 ODM 分類模型的輸出（目標）欄位。

模型名稱。從 Oracle 11*g*R1 開始，名稱 `unique` 已成為關鍵字，不能用作自訂模型名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

一般評論

- 對於 Oracle Data Mining 建立的模型，IBM SPSS Modeler 不提供 PMML 匯出/匯入功能。
- 模型評分始終在 ODM 中進行。如果資料來自於 IBM SPSS Modeler 或需要在其中準備資料，那麼需要將資料集上傳到暫時表格。
- 在 IBM SPSS Modeler 中，一般來講，只會遞送單一預測與關聯的機率或信賴度。
- IBM SPSS Modeler 將可以用於模型建置和評分的欄位個數限制為 1000。
- IBM SPSS Modeler 可以從使用 IBM SPSS Modeler Solution Publisher 發行執行的串流中對 ODM 模型進行評分。

Oracle 模型伺服器選項

指定用於上傳建模資料的 Oracle 連線。如果需要，您可以在「伺服器」標籤上為每個建模節點都選取一個連線，以置換在 Helper 應用程式對話框中指定的預設 Oracle 連線。 請參閱第 26 頁的『啟用與 Oracle 的整合』主題，以取得更多資訊。

備註(O)

- 用於建模的連線可以與串流的來源節點中使用的連線相同，也可以不相同。例如，您可能有一個串流是從某個 Oracle 資料庫存取資料，將資料下載至 IBM SPSS Modeler 以進行清除或其他操作，然後將資料上傳至其他 Oracle 資料庫進行建模。
- ODBC 資料來源名稱有效地內嵌在每一個 IBM SPSS Modeler 串流中。如果在某個主機上建立的串流在不同主機上執行，則資料來源的名稱在每一個主機上必須相同。或者，可以在每一個來源或建模節點中的「伺服器」標籤上選取不同的資料來源。

錯誤分類成本

在某些環境定義中，特定錯誤類別的成本高於其他錯誤的成本。例如，將高風險信貸申請人分類為低風險申請人（一種錯誤類別）的成本高於將低風險申請人分類為高風險申請人（另一種錯誤類別）的成本。使用錯誤分類成本可指定不同類別的預測誤的相對重要性。

錯誤分類成本在本質上指應用於特定結果的加權。這些加權可化為模型中的因素，並可能在實際上變更預測（作為避免高成本錯誤的一種方式）。

除 C5.0 模型之外，在對模型進行評分時，錯誤分類成本是不適用的；在套用自動分類器節點、評估表或分析節點對模型進行分類別或比較時，錯誤分類成本也不予以考慮。將成本計算在內的模型不比不將成本計算在內的模型產生的誤小，這樣的模型不會也不可能按照整體精確度排等級到任何更高的級別，但是在實際應用中，這樣的模型執行的結果可能更好，因為它有一個內建的偏移，從而有利於將錯誤的 成本降低。

成本矩陣顯示了預測種類和實際種類的每個可能的組合的成本。預設情況下，所有錯誤分類成本都設定為 1.0。要輸入自訂成本值，可選取**使用誤分類成本**並將自訂值輸入到成本矩陣中。

要變更誤分類成本，可選取與所需的預測值和實際值的組合對應的 Cell，清除此 Cell 內現有的內容，然後為其輸入所需的成本。成本不會自動均攤。例如，如果將 *A* 誤分類為 *B* 的成本設定為 2.0，那麼將 *B* 誤分類為 *A* 的成本將仍是預設值 1.0，除非也明確地對它進行變更。

注意：僅容許在建置「決策樹」模型時指定成本。

Oracle Naive Bayes

Naive Bayes 是用來解決分類問題的著名演算法。此模型將所有建議預測變數視為彼此獨立，因此被稱為樸素。Naive Bayes 是一個快速可調式演算法，能夠計算屬性及目標屬性組合的條件式機率。從訓練資料中，會建立獨立機率。鑒于每個輸入變數中出現的每個值種類，此機率會提供每個目標類別的可能性。

- 交叉驗證用於檢定建立模型所使用之相同資料的模型精確度。如果可用於建立模型的觀察值的數量很小，那麼該交叉驗證特別有用。
- 模型輸出可用矩陣格式瀏覽。矩陣中的數字為條件式機率，與預測的類別（欄）和預測變數/值的組合（列）相關聯。

貝式邏輯分類演算法模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

使用分割的資料。 如果定義了分割區欄位，那麼此選項可確保僅訓練分割的資料用於建立模型。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

貝式邏輯分類演算法專家選項

除非給定的值或值成對在訓練資料中具有足夠高的發生率，否則在模型構建後，單個預測值屬性值或值成對將被忽略。用於忽略值的臨界值將根據訓練資料中的記錄數指定為分數值。調整此臨界值可減少雜訊並改進模型擬合其他資料集的能力。

- **單臨界值**。指定給定的預測值屬性值的臨界值。給定值的出現次數必須等於或大於指定的分數，否則該值將被忽略。
- **雙臨界值**。指定給定屬性和預測值對的臨界值。給定值對的出現次數必須等於或大於指定的分數，否則該值對將被忽略。

預測機率。容許模型包含目標欄位可能結果的正確預測機率。要啟用此功能，選擇**選擇**，按一下**指定**按鈕，選擇一個可能結果，然後按一下**插入**。

使用預測集。對於目標欄位的所有可能輸出結果，產生所有可能結果的表格。

Oracle 調適性 Bayes

Adaptive Bayes Network (ABN) 使用最小說明長度 (MDL) 和自動功能選項來構造 Bayesian Network 分類器。盡管 ABN 的執行速度慢些，但在貝式邏輯分類演算法表現糟糕的某些狀況中它仍有良好表現，而在其他大多數狀況下也至少不比貝式邏輯分類演算法低劣。ABN 演算法能夠用於建立三種進階的、基於 Bayesian 的模型，包含簡化的樹狀結構（單功能）、刪改的貝式邏輯分類演算法和增強型多功能模型。

註：Oracle 12C 中已丟棄 Oracle Adaptive Bayes 演算法，而且使用 Oracle 12C 時此演算法在 IBM SPSS Modeler 中不受支援。請參閱http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726。

已產生的模型

在單功能建立模式中，ABN 可根據一組人員可讀取的規則生成一個簡化的決策樹，使業務使用者或分析師可以瞭解模型預測的基準並據此向其他人演示或解說。相比於貝式邏輯分類演算法和多功能模型，這是一個突出的優勢。這些規則可以像 IBM SPSS Modeler 中的標準規則集一樣進行瀏覽。如下所示的是一個簡單的規則集：

```
IF MARITAL_STATUS = "Married"AND EDUCATION_NUM = "13-16"THEN CHURN= "TRUE"Confidence = .78, Support = 570 cases
```

刪改的貝式邏輯分類演算法和多功能模型無法在 IBM SPSS Modeler 中瀏覽。

Adaptive Bayes 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

使用分割的資料。 如果定義了分割區欄位，那麼此選項可確保僅訓練分割的資料用於建立模型。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

模型類型

建立模型時有三種不同模式可供選擇。

- **多重功能**。建立和對比若干個模型，包含 NB（素樸貝葉斯）模型、單功能產品機率模型和多功能產品機率模型。這是最詳盡的模式，但通常所需的計算時間也最長。只有單功能模型勝出而成為最佳模型時，才會產生規則。如果選擇了多功能模型或 NB 模型，那麼不會生成任何規則。
- **單一功能**。根據規則集建立簡化決策樹。每個規則均含有一個條件以及與每個結果關聯的機率。各規則互相排斥且其為人員可讀取格式，這可能是相比於貝式邏輯分類演算法和多功能模型的重要優點。
- **貝式邏輯分類演算法**。建立單一 NB 模型並將它與廣域樣本事前分佈進行對比（廣域樣本中目標值的分配）。只有 NB 模型勝出而成為比廣域事前分佈更好的目標值預測值時，才產生 NB 模型作為輸出。否則，將不會輸出任何模型。

Adaptive Bayes 專家選項

限制執行時間。請選取此選項來指定以分鐘表示的最長建立時間。此選項可用於縮短模型生成時間，不過這樣一來，所生成的模型準確性較差。該演算法將在建模過程中的每個重要步驟檢驗是否能夠在指定的時間內完成下一個重要步驟，然後再繼續下一步，並在達到限制時傳回可用的最佳模型。

預測值的數量上限。此選項可用於通過限制使用的預測值的數量，來限制模型的複雜性和提高執行速度。預測值將根據預測值與目標相關性的 MDL 測量值來進行排等級，此排等級測量了預測值包含在模型中的可能性。

貝式邏輯分類演算法預測值的數量上限。此選項指定貝式邏輯分類演算法模型中使用的預測值的最大數目。

Oracle 支援向量機器 (SVM)

支援向量機器 (SVM) 是一種分類和迴歸方法演算法，它使用機器學習理論在不過度配適資料的同時，最大限度地提高預測精確度。SVM 使用訓練資料的選用非線性轉換，接著在轉換後的資料中搜尋迴歸方程式以分隔類別（對於種類目標）或擬合目標（對於連續目標）。Oracle 上配置了 SVM 後，就可以使用這兩個可用核心函數（線性和高斯）中的其中一個來建立模型。線性核心函數完全忽略了非線性變換，使得生成的模型本質上為迴歸模型。

如需相關資訊，請參閱《*Oracle* 資料採礦應用程式開發者手冊》和《*Oracle* 資料採礦概念》。

Oracle SVM 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

主動學習。提供處理大型建模數據集的方法。演算法可套用主動學習，根據小樣本建立一個初始模型，隨後將初始模型套用到完整的訓練資料集中，再根據結果遞增地更新樣本和模型。更新週期將不斷重複，直到模型在訓練資料上收斂，或支援向量的數量達到了容許的最大值。

核心功能。選擇**線性**或**高斯**，或保留預設的**系統已確定**容許系統選擇最適合的核心。高斯核心函數模擬更複雜的關係，但一般來說，耗費的計算時間更長。可先使用線性核心函數，然後如果線性核心函數未能找到合適的擬合，再嘗試使用高斯核心函數。這種方法在迴歸模型中更常用，因為迴歸模型中核心函數的選取更重要。同時請注意，用高斯核心函數建立的 SVM 模型在 IBM SPSS Modeler 中無法瀏覽。用線性核心函數建立的模型則可以像瀏覽標準迴歸模型一樣在 IBM SPSS Modeler 中進行瀏覽。

常態化方法。 指定用於連續輸入欄位和目標欄位的正規化方法。您可以選擇 **Z 評分**、**最小-最大**或**無**。如果已選取**自動資料預備**複選框，Oracle 將自動執行標準。取消勾選此複選框以選取手動常態化方法。

Oracle SVM 專家選項

核心快取大小。指定以位元組表示的快取大小，該快取用於儲存建立作業期間計算的核心函數。如所預期，較大的快取通常建立速度更快。預設值是 50 MB。

收斂容差。指定模型建立終止前容許的允差值。該值必須處於 0 到 1 之間，預設值是 0.001。值較大，建立速度也較快，但模型準確率較低。

指定標準差。指定高斯核心函數使用的標準差參數。此參數影響著模型的複合度和拓展到其他資料集的能力（即資料的過度配適和失度配適）之間的平衡。標準差值越高，越容易傾向於失度配適。此參數值預設通過訓練資料估計得出。

指定ε(Epsilon)。僅適用於迴歸模型，用於指定建立對 epsilon 不敏感的模型時可容許錯誤的區間的值。換言之，它用於識別小錯誤（忽略）與大錯誤（不可忽略）。該值必須處於 0 到 1 之間。依預設，該值將通過訓練資料估計得出。

指定複雜性因子r。指定複雜性因子，複雜性因子用於平衡模型錯誤（通過訓練資料測量出）和模型複合度，以防止資料的過度配適和失度配適。該值越高則對錯誤的罰分就越高，資料過度配適的風險也越高；值越低則對錯誤的罰分就越低，也就越容易導致資料失度配適。

指定離群值比率。指定訓練資料中期望的離群值比率。只對一級 SVM 模型有效。不能與**指定複雜性因子**設定一起使用。

預測機率。容許模型包含目標欄位可能結果的正確預測機率。要啟用此功能，選擇**選擇**，按一下**指定**按鈕，選擇一個可能結果，然後按一下**插入**。

使用預測集。對於目標欄位的所有可能輸出結果，產生所有可能結果的表格。

Oracle SVM 加權選項

在分類模型中，通過使用加權，可以指定各個可能的目標值的相對重要性。這樣做可能很有用，例如，如果您訓練資料中的資料點實際上未分佈在種類之間。加權可讓您調整模型，以便您可以對未充分呈現在資料中的那些種類進行補償。增加某個目標值的加權應該增加該類型的正確預測百分比。

有三種方法可用來設定加權：

- **基於訓練資料。** 此為預設值。加權以訓練資料中種類的相對次數為基礎。
- **對所有類別相類別。** 所有種類的加權都定義為 $1/k$，其中 k 是目標種類數。
- **自訂。** 您可以指定自己的加權。所有的類別的加權起始值設定為相類別。您可以將各個種類的加權調整為使用者定義的值。要調整特定分類的加權，可在表格中對應於所需種類的加權 Cell 中，先清除其內容，然後輸入所需的值。

所有種類的加權之和應為 1.0。如果它們的總和不是 1.0，則會顯示一則警告，提供一個用來自動正規化值的選項。此項自動調整操作可以保留各種類的比例，同時實施加權限制。您可以隨時按一下**正規化**按鈕來執行此調整。若要重設表格以讓所有種類的值相等，請按一下**均分**按鈕。

Oracle 通用性線性模型 (GLM)

（僅限於 11g）「通用性線性模型」放寬了線性模型所作的限制假設。例如，這包含假設情況目標變數具有常態分佈，以及假設情況預測值對目標變數的作用在本質上是線性作用。通用性線性模型適合於目標分佈可能是非常態分佈的預測，例如多項式分佈或 Poisson 分佈。同樣，通用性線性模型在預測值與目標之間的關係或鏈結有可能是非線性關係或鏈結的情況下非常有用。

如需相關資訊，請參閱《*Oracle* 資料採礦應用程式開發者手冊》和《*Oracle* 資料採礦概念》。

Oracle GLM 模式選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

常態化方法。 指定用於連續輸入欄位和目標欄位的正規化方法。您可以選擇 **Z 評分**、**最小-最大**或**無**。如果已選取**自動資料預備**複選框，Oracle 將自動執行標準。取消勾選此複選框以選取手動常態化方法。

遺漏值處理。指定如何處理輸入資料中的遺漏值：

- **取代為平均數或眾數**將數值型屬性的遺漏值取代為平均數，並將種類屬性的遺漏值取代為眾數。
- **只使用完整的記錄**忽略帶有遺漏值的記錄。

Oracle GLM 專家選項

使用列的權重。勾選此方框以啟動相鄰下拉清單，從中可以為列選取包含加權因子的欄。

將列診斷儲存至表格。勾選此方框以啟動相鄰文字欄位，在此可以輸入表格名稱以包含列層次診斷。

係數信賴度層次。目標的預測值在模型計算的信賴度區間內的確定性程度，從 0.0 到 1.0。信賴度範圍隨係數統計資料一起傳回。

目標的參照種類。選擇**自訂**為用作參照種類的目標欄位選擇值或保留預設值**自動**。

脊迴歸。脊迴歸是一種補償在變數中有太高相關性程度的狀況的技術。您可以使用**自動** 選項，容許演算法控制此技術的使用，或者也可通過**取消** 和**啟用** 選項手動控制。如果您選擇手動啟用脊迴歸，那麼可以通過在相鄰欄位中輸入值來置換脊參數的系統預設值。

生成脊迴歸的 VIF。如果您想當脊正在用於線性迴歸時生成變異數膨脹因子 (VIF) 統計資料，勾選此方框。

預測機率。容許模型包含目標欄位可能結果的正確預測機率。要啟用此功能，選擇**選擇**，按一下**指定**按鈕，選擇一個可能結果，然後按一下**插入**。

使用預測集。對於目標欄位的所有可能輸出結果，產生所有可能結果的表格。

Oracle GLM 加權選項

在分類模型中，通過使用加權，可以指定各個可能的目標值的相對重要性。這樣做可能很有用，例如，如果您訓練資料中的資料點實際上未分佈在種類之間。加權可讓您調整模型，以便您可以對未充分呈現在資料中的那些種類進行補償。增加某個目標值的加權應該增加該類型的正確預測百分比。

有三種方法可用來設定加權：

- **基於訓練資料。** 此為預設值。加權以訓練資料中種類的相對次數為基礎。
- **對所有類別相類別。** 所有種類的加權都定義為 $1/k$，其中 k 是目標種類數。
- **自訂。** 您可以指定自己的加權。所有的類別的加權起始值設定為相類別。您可以將各個種類的加權調整為使用者定義的值。要調整特定分類的加權，可在表格中對應於所需種類的加權 Cell 中，先清除其內容，然後輸入所需的值。

所有種類的加權之和應為 1.0。如果它們的總和不是 1.0，則會顯示一則警告，提供一個用來自動正規化值的選項。此項自動調整操作可以保留各種類的比例，同時實施加權限制。您可以隨時按一下**正規化**按鈕來執行此調整。若要重設表格以讓所有種類的值相等，請按一下**均分**按鈕。

Oracle 決策樹

Oracle Data Mining 根據常用的分類和迴歸方法樹狀結構演算法，提供了一種經典的決策樹功能。 ODM 決策樹模型含有每個節點的完整資訊，包括信賴度、支援和分割準則。 可以顯示每個節點的完整規則，而且還提供每個節點的代理屬性，該代理屬性用於在將模型套用到具有遺漏值的觀測數據時作為代理。

決策樹的廣泛套用是因為它適用性廣、便於套用及易於理解。 樹狀結構將對所有可能的輸入屬性進行篩選，以查找最佳「分割器」，即屬性切割點（例如，AGE > 55），以便將下游資料記錄分割成若干更均質的總體。每次分割決策後，ODM 將重複長出整個樹狀結構和建立終端機「葉子」的過程，該葉子代表具有類似記錄、項目或人員的總體。 從樹狀結構節點的根部往下看（例如，總人口），決策樹提供人員可讀取的 IF A, then B 陳述式規則。這些決策樹規則還提供每個樹狀結構節點的支援和信賴度。

Adaptive Bayes Network 也可以提供用於解釋每項預測的簡式規則，但每個分割決策的 Oracle Data Mining 完整規則是由決策樹提供。 決策樹還可以為最佳客戶、已恢復健康的病人以及與欺騙關聯的因子等建立詳細的設定檔。

決策樹模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

雜質矩陣。指定尋求分割每個節點資料的最佳測試問題時使用的度量值。最佳分割器和分隔值是那些能最大限度增加節點中各實體的目標值均一性的分割器和分隔值。均一性通過一個測量值來衡量。受支援的度量值為 **吉尼** 和 **熵**。

決策樹專家選項

最大深度。設定要建立的樹狀結構模型的最大深度。

節點中記錄的下限百分比。設定節點中記錄的下限百分比。

進行分割的記錄下限百分比。設定母節點中記錄數目的下限，該數目下限以用於訓練模型的記錄總數的百分比表示。如果記錄數目小於此百分比，那麼不會試圖進行任何分割。

節點中的記錄數目下限。設定傳回記錄數目的下限。

用於分割的記錄數目的下限。設定母節點中記錄數目的下限，該數目下限以數字表示。如果記錄數目小於此值，那麼不會試圖進行任何分割。

規則 ID。 如果已勾選，模型中會包含一個字串以在已進行特定分割的樹狀結構中識別節點。

預測機率。容許模型包含目標欄位可能結果的正確預測機率。要啟用此功能，選擇**選擇**，按一下**指定**按鈕，選擇一個可能結果，然後按一下**插入**。

使用預測集。對於目標欄位的所有可能輸出結果，產生所有可能結果的表格。

Oracle O-叢集

Oracle O-Cluster 演算法確定資料中自然發生的分組。 正交分區叢集 (O-Cluster) 是 Oracle 專有的叢集演算法，它建立基於階層式網格的叢集模型，也就是說，它在輸入屬性空間中建立軸平行（正交）分區。該演算法遞歸式地運行。所產生的階層式結構為一個不規則的網格，該網格將屬性空間分割成各個叢集。

O-Cluster 演算法可處理數值屬性和種類屬性，且 ODM 將自動選取最佳的叢集定義。ODM 會提供叢集詳細資訊、叢集規則、叢集重心值，並可以用來對叢整合員資格上的母體進行評分。

O-Cluster 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

上限叢集數。設定產生的叢集的最大數量。

O-Cluster 專家選項

最大緩衝區。設定最大緩衝區大小。

敏感度。設定一個分數，該分數指定分割新叢集所要求的最高密度。該分數與廣域均勻分配密度相關聯。

Oracle K-Means

Oracle k-Means 演算法確定資料中自然發生的分組。k-Means 演算法是基於距離的叢集演算法，該演算法將資料分割為預定數量的叢集（條件是存在足夠的不同觀察值）。 基於距離的演算法根據距離測量（函數）來衡量資料點之間的親緣性。 根據所使用的距離度量值，資料點被指派到與之距離最近的叢集。ODM 提供增強版的 k-Means。

k-Means 演算法支援階層式叢集，處理數值和種類屬性並將總體分割為使用者指定數量的叢集。 ODM 會提供叢集詳細資訊、叢集規則、叢集重心值，並可以用來對叢整合員資格上的母體進行評分。

k-Means 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

叢集數目。 設定產生叢集的數量。

距離函數。指定 k-Means 叢集使用的距離函數。

分割準則。指定 k-Means 叢集使用的分割準則。

常態化方法。 指定用於連續輸入欄位和目標欄位的正規化方法。您可以選擇 **Z 評分**、**最小-最大**或**無**。

k-Means 專家選項

疊代。 設定 k-Means 演算法的疊代次數。

收斂容錯。 設定 k-Means 演算法的收斂容差。

分組數目。 指定 k-Means 生成的屬性直方圖中的 Bin 的數目。每個屬性的 bin 範圍都是通過對整個訓練資料集進行廣域計算得到的。分級方法為等寬法。具有單一值的屬性只有一個分類，除此以外，其他所有屬性均具有同樣數量的分級。

區塊成長。設定配置用於容納叢集資料的記憶體的成長因素。

下限百分比屬性支援。設定屬性值分數，該屬性值必須為非無效，才能使該屬性包含在叢集的規則說明中。如果參數值在具有遺漏值的資料中設定得過高，那麼可能導致規則過短，或甚至為空白。

Oracle 非負矩陣分解 (NMF)

非負矩陣分解 (NMF) 用於將大資料集簡化為若干具有代表性的屬性。它與主成分分析 (PCA) 的原理類似，但可以處理更大數量的屬性，在可加性代表模型中，NMF 是功能強大的先進資料採礦演算法，而且用途廣泛。

NMF 可以用於將大數量資料（比如文字資料）簡化為小的、稀疏得多的代表，NMF 降低了資料的維度，即用少得多的變數儲存了等數量的資訊。NMF 模型的輸出可用有監督的學習技術（比如 SVM）或沒有監督的學習技術（比如 叢集）來進行分析。Oracle Data Mining 用 NMF 和 SVM 演算法來挖掘尚未結構化的文字資料。

NMF 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

常態化方法。 指定用於連續輸入欄位和目標欄位的正規化方法。您可以選擇 **Z 評分**、**最小-最大**或**無**。如果已選取**自動資料預備**複選框，Oracle 將自動執行標準。取消勾選此複選框以選取手動常態化方法。

NMF 專家選項

指定功能個數。指定要擷取的特徵的數量。

隨機種子。 設定 NMF 演算法的隨機種子。

疊代次數。 設定 NMF 演算法的疊代數。

收斂容錯。 設定 NMF 演算法的收斂容差。

顯示所有功能。 顯示所有功能的特徵 ID 和信賴度，而不是僅顯示最佳特徵的特徵 ID 和信賴度。

Oracle Apriori

Apriori 演算法會探索資料中的相關規則。例如，「如果客戶購買剃須刀和須後產品，那麼該客戶還會購買剃須膏，並且信賴度為 80%。」關聯挖掘問題可以拆解為兩個子問題：

- 找到所有稱為頻繁項目集合的項目組合，即支援大於最小支援的項目組合。
- 使用頻繁項目集合來產生所需要的規則。 舉例說明規則的生成原理，例如，ABC 和 BC 為頻繁項，如果 support(ABC) 與 support(BC) 的比例大於等於最小信賴度時，那麼可使用「從規則 A 推衍生 BC」。注意：如果 ABCD 為頻繁項，該規則將具有最小支援。ODM 關聯僅支援單一結果規則（從 ABC 推衍生 D）。

頻繁項目集合的數量取決於最小支援參數。產生規則的數量取決於頻繁項目集合的數量和信賴度參數。如果信賴度參數設得過高，那麼關聯模型中可能存在頻繁項目集合，但不存在規則。

ODM 將基於 SQL 來執行 Apriori 演算法。候選產生和支援計數步驟使用 SQL 查詢來執行。 不使用專門的記憶體內資料結構。SQL 查詢將使用各種提示進行優化，以便能在資料庫伺服器中高效執行。

Apriori 欄位選項

所有建模節點都有一個「欄位」標籤，您可以在其中指定要用於建置模型的欄位。

在建立 Apriori 模型之前，需要指定要將哪些欄位用作與關聯建模有關的項目。

使用類型節點設定。此選項會告知節點使用來自上游「類型」節點的欄位資訊。此為預設值。

使用自訂設定。此選項會告知節點使用此處指定的欄位資訊，而不使用任何上游「類型」節點的指定欄位資訊。選取此選項後，根據是否正在使用交易格式來指定對話框中的剩餘欄位。

如果沒有 使用格式，請指定：

- **輸入**。選取輸入欄位。這類似於在「類型」節點中將欄位角色設定為輸入。
- **分割區**。此欄位容許您指定一個欄位，以用來將資料分割為不同樣本以用於模型建置的訓練、測試及驗證階段。

如果 正在 使用交易處理格式，請指定：

使用交易式格式。 如果希望將每個項目列中的資料轉換為每個觀察值列中的資料，請使用此選項。

選取此選項會變更該對話框下半部分中的欄位控制項：

對於交易處理格式，請指定：

- **ID**。 從清單中選取 ID 欄位。可以將數值或符號欄位用作 ID 欄位。此欄位的每一個唯一值都應指出一個特定的分析單位。例如，在購物籃應用程式中，每一個 ID 都可能代表一個客戶。對於 Web 日誌分析應用程式，每一個 ID 都可能代表一部電腦（依 IP 位址）或一位使用者（依登入資料）。
- **內容**。 指定模型的內容欄位。該欄位包含與關聯建模有關的項目。
- **分割區**。此欄位容許您指定一個欄位，以用來將資料分割為不同樣本以用於模型建置的訓練、測試及驗證階段。通過用某個樣本建立模型並用另一個樣本對模型進行測試，您可以預判出此模型對類似於目前資料的大型資料集的擬合優劣。如果已使用「類型」或「分割區」節點來定義多個分割區欄位，則必須在使用分割的每一個建模節點中的「欄位」標籤上選取單一分割區欄位。（如果僅存在一個分割區，則每當啟用分隔時，都會自動使用該分割區。）另請注意，若要套用您分析中選取的分割區，則還必須在節點的「模型選項」標籤中啟用分割。（取消選取此選項可能會停用分割而不變更欄位設定。）

Apriori 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

規則最大長度長度。為任何規則設定前置條件數上限，該值為從 2 到 20 的整數。這是限制規則複雜性的一種方法。如果規則過於複合或過於具體，或者訓練規則集所需的時間過長，請嘗試減小此設定。

最小信賴度。 設定最小信賴度層次，該值介於 0 和 1 之間。信任層次低於指定準則的規則將被放棄。

最小支援。 設定最小支援臨界值，該值介於 0 和 1 之間。「事前」探索頻率高於最小支援臨界值的型樣。

Oracle 說明長度下限 (MDL)

Oracle 最小說明長度 (MDL) 演算法用於確定對目標屬性具有最大影響的屬性。通常情況下，知道哪個是最有影響的屬性可以更好地瞭解和管理業務並且有助於簡化建模操作。另外，這些屬性可以指示為擴大模型而希望新增的資料的類型。例如，MDL 可用於找到與以下預測內容最相關的屬性：製造的零件的品質、與流失相關聯的因素以及最有可能用於治療特定疾病的基因等等。

Oracle MDL 將捨棄它認為對於預測目標而言不重要的輸入欄位。然後，它使用餘下的輸入欄位建立與 Oracle Data Miner 中顯示的 Oracle 模型相關聯的未優化模型塊。在 Oracle Data Miner 中瀏覽模型將顯示一個圖表，其中顯示了餘下的輸入欄位，按照它們在預測目標方面的重要性順序排名。

負分等級指示雜訊。排名為零或更小值的輸入欄位不影響預測，應從資料中移除。

要顯示圖表

1. 在「模型」選用區中用滑鼠右鍵按一下非優化模型塊並選擇**瀏覽**。
2. 在模型視窗中，按一下按鈕以啟動 Oracle Data Miner。
3. 連接至 Oracle Data Miner。請參閱第 41 頁的『Oracle 資料採礦程式』主題，以取得更多資訊。
4. 在 Oracle Data Miner 導覽器畫面中，展開**模型**，然後展開**屬性重要性**。
5. 選取相關的 Oracle 模型（其名稱與您在 IBM SPSS Modeler 中指定的目標欄位名稱相同）。如果您不確定哪個正確，請選取「屬性重要性」資料夾並按建立日期查找模型。

MDL 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

唯一欄位。 指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。 IBM SPSS Modeler 限制這個鍵欄位必須為數值。

註：此欄位對於除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 節點外的所有 Oracle 節點都是選用的。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

Oracle 屬性重要性 (AI)

屬性重要性的目標是找出資料集中的哪些屬性與結果相關，以及它們影響最終結果的程度。「Oracle 屬性重要性」節點將分析資料、尋找型樣並預測具有相關聯信賴度的結果。

AI 模型選項

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

使用分割的資料。 如果定義了分割區欄位，那麼此選項可確保僅訓練分割的資料用於建立模型。

自動資料預備。（僅 11*g*）啟用（預設）或取消 Oracle Data Mining 的自動資料預備模式。如果已勾選此框，那麼 ODM 將自動執行演算法所需的資料變換。有關詳細資訊，請參閱 *Oracle Data Mining* 概念。

AI 選取選項

「選項」標籤用於指定在模型塊中選取或排除輸入欄位的預設值。然後可以將模型新增到串流，以選取用於後續模型建立的欄位子集合。或者，也可以通過在產生模型後在模型瀏覽器中選取或棄選其他欄位，以置換這些設定。但是，預設值下，無需更多修改即可套用模型塊，這點在 Script 編寫方面特別有用。

您可以使用的選項如下：

所有已排等級的欄位。根據欄位的重要、邊際或不重要分等級等級來選取欄位。可編輯每項排等級的標籤及用於指派記錄的分等級等級的截斷值。

欄位數目上限。 根據重要性選取前 *n* 個欄位。

重要性大於。 選取重要性大於指定值的所有欄位。

不管如何選擇，目標欄位總是被保留。

AI 模型塊模型標籤

針對 Oracle AI 模型塊的「模型」標籤顯示所有輸入的排名和重要性，並容許您使用左側欄中的勾選框來選取用於進行過濾的欄位。執行這個串流時，將只保留已勾選的欄位以及目標預測。其他輸入欄位將被捨棄。 預設選擇基於建模節點中指定的選項，但您可以根據需要選擇或取消選擇其他欄位。

- 若要依等級、欄位名稱、重要性或任何其他顯示的直欄來排序清單，請按一下直欄標頭。另外，可以從「排序依據」按鈕旁的清單中選取期望的項目，並使用向上和向下箭頭來變更排序方向。
- 可使用工具列來選中或棄選所有欄位和存取「檢查欄位」對話框，可在該對話框上根據排等級或重要性來選取欄位。另外，還可以在按住 Shift 或 Ctrl 鍵的情況下按一下各欄位以延伸選擇。
- 用來將輸入分級成重要、一般或不重要的臨界值會顯示在表格下面的圖註中。這些值在建模節點中指定。

管理 Oracle 模型

Oracle 模型新增到模型選用區的方式與其他 IBM SPSS Modeler 模型的新增方式一樣，而且使用方法也大致相同。但是，也有幾點重大差異，比如 IBM SPSS Modeler 中生成的每個 Oracle 模型實際參照的是儲存在資料庫伺服器上的模型。

Oracle 模型塊伺服器標籤

通過 IBM SPSS Modeler 建立 ODM 模型即可在 IBM SPSS Modeler 中建立一個模型，並建立或替代 Oracle 資料庫中的一個模型。這種 IBM SPSS Modeler 模型將參照資料庫伺服器上儲存的資料庫模型的內容。IBM SPSS Modeler 可以通過將完全相同的產生**模型鍵**字串儲存在 IBM SPSS Modeler 模型和 Oracle 模型中執行一致性檢查。

每個 Oracle 模型的鍵字串顯示在「列出模型」對話框中的模型資訊欄下。IBM SPSS Modeler 模型的鍵字串在 IBM SPSS Modeler 模型的「伺服器」標籤上顯示為**模型鍵**（放置在串流中時）。

模型塊「伺服器」標籤上的「檢驗」按鈕，可用於檢驗 IBM SPSS Modeler 模型中的模型鍵和 Oracle 模型是否相符。如果 Oracle 中無法找到名稱相同的模型，或者模型鍵不相符，那麼 Oracle 模型已被刪除或在 IBM SPSS Modeler 模型建立後重新建立。

Oracle 模型塊彙總標籤

模型區塊的「摘要」標籤會顯示模型本身的相關資訊（分析），模型中使用的欄位（欄位），建置模型時使用的設定（建置設定）以及模型訓練（訓練摘要）。

當您第一次瀏覽節點時，「摘要」標籤結果會收合。若要查看相關結果，請使用項目左側的展開程式控制項來展開結果，或按一下**全部展開**按鈕以顯示所有結果。檢視完成後要隱藏結果時，請使用展開控制項來摺疊想要隱藏的具體結果，或者按一下**全部收合**按鈕來摺疊所有結果。

分析。 顯示特定模型的相關資訊。如果已執行附加到此模型塊的分析節點，那麼分析中的資訊也將顯示在此部分中。

欄位。 列出用來作為目標的欄位以及用來建置模型的輸入。

建置設定。 包含用來建置模型之設定的相關資訊。

訓練摘要。 顯示模型類型、用來建立模型的串流，模型建立者、模型建置時間以及建置模型的經歷時間。

Oracle 模型塊設定標籤

模型塊的「設定」標籤容許您置換建模節點上某些選項的設定，以便進行評分。

Oracle 決策樹

使用錯誤分類成本。確定是否在 Oracle 決策樹模型中使用錯誤分類成本。請參閱第 28 頁的『錯誤分類成本』主題，以取得更多資訊。

規則 ID。 如果選取（已選取），將規則 ID 欄新增到 Oracle 決策樹模型中。規則 ID 用於識別樹狀結構中進行特定分割的節點。

Oracle NMF

顯示所有功能。 如果選取（已選取），顯示所有特徵的特徵 ID 和信賴度，而不是僅在 Oracle NMF 模式中顯示最佳特徵的特徵 ID 和信賴度。

列出 Oracle 模型

「列出 Oracle Data Mining 模型」按鈕用於啟動一個對話框，該對話框列出現有資料庫模型並容許刪除模型。此對話框可以從 Helper 應用程式對話框中啟動，也可以通過 ODM 相關節點的建立、瀏覽和套用對話框啟動。

針對每個模型顯示下列資訊：

- **模型名稱。** 模型的名稱，用於對清單進行排序
- **模型資訊**。模型鍵資訊，由建立日期/時間和目標欄名組成
- **模式類型。** 建立此模型的演算法的名稱

Oracle 資料採礦程式

Oracle Data Miner 是 Oracle Data Mining (ODM) 的使用者介面，並替代以前 IBM SPSS Modeler 的 ODM 使用者介面。Oracle Data Miner 旨在增加分析師在使用 ODM 演算法方面的成功率。該目標通過以下方式來實現：

- 使用者在套用能同時處理資料預備和演算法選擇的方法學方面需要更多協助。Oracle Data Miner 通過提供資料採礦活動來逐步引導使用者使用正確的方法，來滿足此需求。
- Oracle Data Miner 為模型建置提供改進的和擴展的試探法，為指定模型和變換設定提供可降低錯誤幾率的變換精靈。

定義 Oracle Data Miner 連線

1. Oracle Data Miner 可通過任何版本的 Oracle 進行啟動，可通過**啟動 Oracle Data Miner** 按鈕套用節點和輸出對話框。

圖 2. 啟動 *Oracle Data Miner* 按鈕

2. 如果正確設置了 Helper 應用程式選項，那麼 Oracle Data Miner 的**編輯連線**對話框將在 Oracle Data Miner 外部應用程式啟動之前顯示在使用者面前。

   註：此對話框僅在不存在已定義連線名稱時顯示。

   - 提供一個 Data Miner 連線名稱並輸入對應的 Oracle 10gR1 或 10gR2 伺服器資訊。Oracle 伺服器應與 IBM SPSS Modeler 中指定的伺服器一樣。
3. Oracle Data Miner 的**選擇連線**對話框提供用於指定使用哪個（以上方步驟中定義的）連線名稱的選項。

關於 Oracle Data Miner 需求、安裝和使用的詳細資訊，請參閱 Oracle 網站上的 Oracle Data Miner。

準備資料

使用 Oracle Data Mining 演算法的貝式邏輯分類演算法、Adaptive Bayes 和支援向量機器來建模時，可以使用兩種類型的資料預備：

- **bin**，即，對於無法接受連續資料的演算法，將連續數值範圍欄位轉換為種類。
- **標準**，即套用至數值範圍的轉換，以使這些數值範圍具有類似的平均值和標準差。

分組

IBM SPSS Modeler 的「 bin 」節點提供了許多執行 bin 作業的技術。定義了可以套用至一個或多個欄位的 bin 作業。如在資料集上執行 bin 作業，那麼將建立臨界值並容許建立 IBM SPSS Modeler 的「衍生」節點。「衍生」作業可轉換為 SQL 並模型建置和評分前被套用。此方法將在模型與執行 bin 的「衍生」節點之間建立相依關係，但容許 bin 規格由多個建模作業重複使用。

標準

用作支援向量機器模型的輸入的連續（數值範圍）欄位應該先進行正規化，然後再用於模型建置。對於迴歸模型，還必須反轉標準，以根據模型輸出重新構建分數。SVM 模型設定用於選擇 **Z 分數**、**最值法** 或**無**。通過 Oracle 建立標準係數是模型建置程序中的一個步驟，這些係數將被上傳到 IBM SPSS Modeler 並儲存在模型中。套用時，這些係數將被轉換為 IBM SPSS Modeler 衍生表示式，並用於準備（評分時套用的）資料，然後再將資料傳輸到模型。此情況中，標準與建模作業緊密關聯。

Oracle Data Mining 範例

提供若干樣本串流，以展示如何在 IBM SPSS Modeler 中使用 ODM。這些串流位於 IBM SPSS Modeler 安裝資料夾中的 *\Demos\Database_Modelling\Oracle Data Mining* 目錄下。

附註：「展示」資料夾可從 Windows「啟動」功能表的 IBM SPSS Modeler 程式集存取。

下表格中的串流是資料庫挖掘過程的範例，通過使用 Oracle Data Mining 提供的支援向量機器 (SVM) 演算法，依次使用這些樣本串流。

表 *4.* 資料庫採礦 - 範例串流

串流	說明
1_upload_data.str	用來清除純文字檔中的資料以及將資料上傳至資料庫。
2_explore_data.str	提供使用 IBM SPSS Modeler 來探索資料的範例
3_build_model.str	使用資料庫的原生演算法來建置模型。
4_evaluate_model.str	用來作為使用 IBM SPSS Modeler 來評估模型的範例
5_deploy_model.str	部署用來在資料庫內進行評分的模型。

附註：若要執行該範例，必須依序執行串流。此外，還必須更新每個串流中的來源與建模節點，以參照您要使用之資料庫的有效資料來源。

串流範例中所使用的資料集涉及信用卡應用程式，並呈現類別預測值與連續預測值之混合發生的分類問題。如需此資料集的相關資訊，請參閱串流範例所在資料夾中的 *crx.names* 檔。

可從 *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* 的 UCI 機器學習儲存庫中取得此資料集。

串流範例：上傳資料

第一個範例串流 *1_upload_data.str* 用於清理純文字檔案中的資料並將其上傳到 Oracle。

由於 Oracle Data Mining 需要唯一的 ID 欄位，因而這個初始串流通過 IBM SPSS Modeler 的 `@INDEX` 函數，使用「衍生」節點將名為*ID*的新欄位新增到資料集中，其唯一值為 1、2 和 3。

「填入器」節點用於處理遺漏值，並將從文字檔 *crx.data* 讀取的空白欄位取代為 空白 值。

串流範例：探索資料

第二個串流範例 *2_explore_data.str* 用來示範如何使用「資料審核」節點來取得資料的一般概觀，其中包括摘要統計資料與圖形。

按兩下「資料審核報告」中的圖形可產生更詳細的圖形，以便更深入地探索給定的欄位。

串流範例：建置模型

第三個串流範例 *3_build_model.str* 說明在 IBM SPSS Modeler 中建置模型。按兩下「資料庫來源」節點（標註為 CREDIT）以指定資料來源。要指定建構設定值，請按兩下建立節點（最初標註為 CLASS，指定資料來源後將變更為 FIELD16）。

在對話框的「模型」標籤上：

1. 確保選取**ID**作為唯一欄位。
2. 確保選取**線性**作為核心功能，選取 **z 分數**作為常態化方法。

串流範例：評估模型

第四個串流範例 *4_evaluate_model.str* 說明使用 IBM SPSS Modeler 在資料庫內建模的優點。一旦執行完模型，即可將它新增回資料串流中並使用 IBM SPSS Modeler 提供的多種工具來評估模型。

檢視建模結果

將一個「表格」節點附加到模型塊以探索結果。**$O-field16** 欄位顯示每個觀察值中 *field16* 的預測值，而 **$OC-field16** 欄位顯示該預測的信賴度值。

評估模型結果

您可以使用「分析」節點建立重合矩陣，以顯示每個預測欄位與其目標欄位之間的相符型樣。然後，執行「分析」節點以請參閱結果。

您可以使用「評估」節點建立收益圖表，用於顯示模型對精確性的提高。然後，執行「評估」節點以請參閱結果。

串流範例：部署模型

您對模型的精確度感到滿意以後，便可以部署模型以用於外部應用程式或發佈回資料庫。最後一個範例串流，即 *5_deploy_model.str* 中，資料從表格 CREDITDATA 讀取，然後進行評分並使用名稱為部署解決方案的 Publisher 節點將資料發行到表格 CREDITSCORES。

第 5 章 使用 IBM Data Warehouse 和 IBM Netezza Analytics 進行資料庫建模

SPSS Modeler with IBM Data Warehouse 和 IBM Netezza Analytics

IBM SPSS Modeler 支援與 IBM Data Warehouse 和 IBM Netezza® Analytics 整合，以提供在這些 IBM 伺服器上執行進階分析的能力。這些功能可透過 IBM SPSS Modeler 圖形使用者介面和工作流程導向的開發環境進行存取，使您可以直接在 IBM Netezza 或 IBM Data Warehouse 環境中執行資料採礦演算法。

SPSS Modeler 支援整合來自 **IBM Netezza Analytics** 的下列演算法：

- 決策樹
- K-Means
- TwoStep
- Bayes 網路
- Naive Bayes
- KNN
- 區分叢集
- PCA
- 迴歸樹狀結構
- 線性迴歸
- 時間序列
- 廣義線性

如需這些演算法的相關資訊，請參閱 *IBM Netezza Analytics* 開發人員手冊和 *IBM Netezza Analytics* 參考手冊。

SPSS Modeler 支援整合來自 **IBM Data Warehouse** 的下列演算法（不支援貝氏網路、分割叢集法和時間序列）：

- 決策樹
- K-Means
- TwoStep
- Naive Bayes
- KNN
- PCA
- 迴歸樹狀結構
- 線性迴歸
- 廣義線性

整合需求

下列是使用 IBM Netezza Analytics 或 IBM Data Warehouse 處理資料庫內建模的必備條件。您可能需要諮詢您的資料庫管理者以確保符合這些條件。

- 對在 Windows 或 UNIX（不包括 zLinux，未提供可用的 IBM Netezza ODBC 驅動程式）上安裝 IBM SPSS Modeler Server 執行的 IBM SPSS Modeler。
- 執行 IBM Netezza Analytics 套件的 IBM Netezza Performance Server。

  註：所需的最低 Netezza Performance Server (NPS) 版本取決於所需的 INZA 版本，如下所示：
  - NPS 6.0.0 P8 以上的所有版本都支援 2.0 以前的 INZA 版本。
  - 要使用 INZA 2.0 或更高版本，需要 NPS 6.0.5 P5 或更高版本。

  「Netezza 廣義線性」和「Netezza 時間序列」需要 INZA 2.0 及更高版本才能正常運行。所有其他 Netezza 資料庫內節點都需要 INZA 1.1 或更高版本。
- 連接至 IBM Netezza 資料庫所需的 ODBC 資料來源。請參閱『啟用整合』主題，以取得更多資訊。
- 連接至 IBM Data Warehouse 資料庫所需的 ODBC 資料來源。
- IBM SPSS Modeler 中啟用的 SQL 產生和最佳化。請參閱『啟用整合』主題，以取得更多資訊。

註：資料庫建模和 SQL 最佳化需要在 IBM SPSS Modeler 電腦上啟用 IBM SPSS Modeler Server 連接。通過啟用此設定，您可以存取資料庫演算法，直接從 IBM SPSS Modeler 回送 SQL 以及存取 IBM SPSS Modeler Server。要驗證目前授權的狀態，請從 IBM SPSS Modeler 功能表中選擇下列項目。

說明 > **關於** > **其他詳細資訊**

如果啟用了連接，您可以在「授權狀態」標籤中看到選項**伺服器啟用**。

啟用整合

啟用與 IBM Netezza Analytics 或 IBM Data Warehouse 的整合包含下列步驟。

- 配置 IBM Netezza Analytics 或 IBM Data Warehouse
- 建立 ODBC 來源
- 在 IBM SPSS Modeler 中啟用整合
- 在 IBM SPSS Modeler 中啟用 SQL 產生和最佳化

在以下部分中將介紹這些內容。

配置 IBM Netezza Analytics 或 IBM Data Warehouse

若要安裝或配置 IBM Netezza Analytics 或 IBM Data Warehouse，請參閱適當的 IBM 文件。例如，若為 IBM Netezza Analytics，請參閱該產品隨附的 *IBM Netezza Analytics* 安裝手冊。該手冊中的設定資料庫權限小節包含容許 IBM SPSS Modeler 串流寫入資料庫所需執行的 Script 詳細資料。

註：如果您要使用依賴於矩陣計算的節點，那麼必須透過執行 `CALL NZM..INITIALIZE();` 來起始設定矩陣引擎，否則執行儲存程序將失敗。對於每個資料庫，該起始設定為一次性設定步驟。

為 IBM Netezza Analytics 建立 ODBC 來源

要啟用 IBM Netezza 資料庫和 IBM SPSS Modeler 之間的連線，您需要建立 ODBC 系統資料來源名稱 (DSN)。

在建立 DSN 之前，您應該基本瞭解 ODBC 資料來源與驅動程式以及 IBM SPSS Modeler 中的資料庫支援。

如果您針對 IBM SPSS Modeler Server 以分散式模式執行，請在伺服器電腦上建立 DSN。如果您以本端（用戶端）模式執行，請在用戶端電腦上建立 DSN。

Windows 用戶端

1. 從您的 *Netezza Client* CD 上，執行 *nzodbcsetup.exe* 檔案以啟動安裝程式。請按畫面上的指示執行操作，以安裝驅動程式。有關詳細說明，請參閱《IBM Netezza ODBC、JDBC 和 OLE DB 安裝與配置手冊》。
   a. 建立 DSN。

      註：功能表序列隨 Windows 版本不同而有所變化。

      - **Windows XP**。 從「開始」功能表中選擇**控制台**。按兩下**系統管理工具**，然後按兩下**資料來源 (ODBC)**。
      - **Windows Vista**。從「開始」功能表中選擇**控制台**，然後選擇**系統維護**。按兩下**系統管理工具**，選取**資料來源 (ODBC)**，然後按一下**開啟**。
      - **Windows 7**。從「開始」功能表中選擇**控制台**，選擇**系統和安全**，然後選擇**系統管理工具**。選取**資料來源 (ODBC)**，然後按一下**開啟**。

   b. 跳至**系統 DSN** 標籤，然後按一下**新增**。
2. 從清單中選取 **NetezzaSQL**，然後按一下**完成**。
3. 在 Netezza ODBC 驅動程式設定畫面的 **DSN 選項**標籤上，鍵入選擇的資料來源名稱、IBM Netezza 伺服器的主機名稱或 IP 位址、連線埠號、使用的 Netezza 實例的資料庫，以及用於資料庫連線的使用者名稱和密碼資訊。按一下**說明**按鈕獲得欄位說明。
4. 按一下**測試連線**按鈕並確保您連接至資料庫。
5. 在成功連線後，重複按一下**確定**以結束 ODBC 資料來源管理器畫面。

Windows 伺服器

對於 Windows Server，該程序與 Windows XP 用戶端的程序相同。

UNIX 或 Linux 伺服器

下列程序適用於 UNIX 或 Linux 伺服器（不包括 zLinux，未提供適用的 IBM Netezza ODBC 驅動程式）。

1. 從您的 Netezza Client CD/DVD 上，將對應的 `<platform>cli.package.tar.gz` 檔案複製到伺服器上的暫時位置。
2. 通過 **gunzip** 和 **untar** 指令，解壓縮存檔內容。
3. 為解壓縮的 *unpack* Script 新增執行權限。
4. 執行 Script，並在畫面提示時給出回答。
5. 編輯 `modelersrv.sh` 檔案以包含下列行。

   ```
   . <SDAP Install Path>/odbc.sh
   LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
   NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
   ```

   例如：

   ```
   . /usr/IBM/SPSS/SDAP/odbc.sh
   LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
   NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
   ```

6. 找到檔案 `/usr/local/nz/lib64/odbc.ini` 並將其內容複製到隨 SDAP 安裝的 `odbc.ini` 檔案（由環境變數 $ODBCINI 定義）中。

注意：對於 64 位元 Linux 系統，**Driver** 參數錯誤地參照了 32 位元驅動程式。當您在上一步驟中複製 odbc.ini 內容時，應相應地編輯該參數中的路徑，例如：

```
/usr/local/nz/lib64/libnzodbc.so
```

7. 編輯 Netezza DSN 定義中的參數，以反映要使用的資料庫。
8. 重新啟動 IBM SPSS Modeler Server，並在用戶端上測試使用 Netezza 資料庫內挖掘節點。

在 SPSS Modeler 中啟用整合

1. 在 IBM SPSS Modeler 主功能表中，選擇

   工具 > **選項** > **說明應用程式**。
2. 按一下 **IBM Data Warehouse** 標籤。

啟用 IBM Data Warehouse 分析整合。啟用 IBM SPSS Modeler 視窗底部的「資料庫建模」選用區（如果尚未顯示）並為 IBM Data Warehouse 和 Netezza 資料採礦演算法新增節點。

IBM Data Warehouse 連線。按一下**編輯**按鈕，並選擇在建立 ODBC 來源時設定的 IBM Data Warehouse 連線字串。如需相關資訊，請參閱 IBM Data Warehouse 管理主控台。

啟用 SQL 產生及最佳化

由於使用超大型資料集的可能性，出於效能的原因，您應在 IBM SPSS Modeler 中啟用 SQL 產生和最佳化選項。

1. 從 IBM SPSS Modeler 功能表中選擇：

   工具 > **串流內容** > **選項**
2. 按一下導覽窗格中的**最佳化**選項。
3. 確認已啟用**產生 SQL** 選項。資料庫建模需要此設定才能夠運作。
4. 選中**最佳化 SQL 產生**和**最佳化其他執行**（非嚴格必要但強烈推薦使用，以使效能更優）。

使用 IBM Netezza Analytics 和 IBM Data Warehouse 來建置模型

每個受支援的演算法都有對應的建模節點。您可以從節點選用區上的**資料庫建模**標籤中存取 IBM Data Warehouse 和 IBM Netezza 建模節點。

資料考量

資料來源中的欄位可包含各種資料類型的變數，視建模節點而定。在 IBM SPSS Modeler 中，資料類型稱為測量層次。建模節點的「欄位」標籤使用圖示來指出其輸入與目標欄位允許的測量層次類型。

目標欄位 目標欄位是您嘗試預測其值的欄位。在可指定目標的位置，只能選取其中一個來源資料欄位作為目標欄位。

記錄識別欄位 指定用於唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。如果來源資料不包含 ID 欄位，您可以按照「衍生」節點的方法建立此欄位，如下列程序所示。

1. 選取來源節點。
2. 從節點選用區上的「欄位作業」標籤中，按兩下「衍生」節點。
3. 在畫布上按兩下「衍生」節點圖示來開啟「衍生」節點。
4. 例如，在**衍生**欄位中輸入 ID。

5. 在**公式**欄位中，輸入 `@INDEX` 並按一下**確定**。
6. 將「衍生」節點連接至串流的剩餘部分。

註：如果您使用 `NUMERIC(18,0)` 資料類型從 Netezza 資料庫擷取長數值資料，在匯入期間 SPSS Modeler 有時會向上捨入資料。為避免此問題，請使用 `BIGINT` 或 `NUMERIC(36,0)` 資料類型儲存資料。

註：由於存在針對可以使用的欄位類型的限制，因此具有無類型測量層次和記錄 ID 角色的欄位不會顯示在 Netezza 資料庫內建模節點（例如，K-Means）中。

處理空值

如果輸入資料包含空值，那麼使用某些 Netezza 節點可能會導致產生錯誤訊息或者長時間執行的串流，因此我們建議移除包含空值的記錄。使用下列方法。

1. 將「選取」節點附加至來源節點。
2. 將「選取」節點的**模式**選項設為**捨棄**。
3. 在**條件**欄位中輸入下列內容：

   `@NULL(`*`field1`*`) [or @NULL(`*`field2`*`)[... or @NULL(`*`fieldN`*`]])`

   請務必包括每一個輸入欄位。
4. 將「選取」節點連接至串流的剩餘部分。

模型輸出

包含 Data Warehouse 或 Netezza 建模節點的串流有可能每次執行都產生略微不同的結果。這是因為在模型建置之前，將資料讀取至暫時表格時，節點讀取來源資料的順序不一定相同。但是，此效果產生的差異可以忽略不計。

一般註解

- 在 IBM SPSS Collaboration and Deployment Services 中，不能使用包含 IBM Data Warehouse 或 IBM Netezza 資料庫建模節點的串流來建立評分配置。
- Data Warehouse 或 Netezza 節點構建的模型無法進行 PMML 匯出或匯入。

欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

目標。 選擇一個欄位作為預測的目標。若為「一般線性」模型，另請參閱此畫面上的**試用**欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

伺服器選項

在「伺服器」標籤上，指定將在其中建置模型的 IBM Data Warehouse 資料庫。

IBM Data Warehouse 伺服器詳細資料。在這裡，可以指定要用於模型的資料庫的連線細節。

- **使用上游連線**。（預設值）使用上游節點（例如資料庫來源節點）中指定的連線詳細資料。僅當所有上游節點都能夠使用 SQL 回送功能時，此選項才有效。在此情況下，無需將資料移出資料庫，因為 SQL 完全實現所有的上游節點。
- **將資料移至資料庫**。將資料移至您在這裡指定的資料庫。這樣，即使資料位於另一個 IBM Data Warehouse 資料庫或者另一供應商的資料庫中，甚至位於純文字檔案中，也仍然可以進行建模。此外，如果因節點未執行 SQL 推回而擷取了資料，則資料會移回至在這裡指定的資料庫。按一下**編輯**按鈕以瀏覽並選取連線。

  注意：
  IBM Netezza Analytics 和 IBM Data Warehouse 通常與非常大型的資料集配合使用。在資料庫之間傳輸大數量資料，或者從資料庫中取出或存入大數量資料，可能非常耗時，應盡可能避免。

註：ODBC 資料來源名稱有效地內嵌在每一個 IBM SPSS Modeler 串流中。如果在某個主機上建立的串流在不同主機上執行，則資料來源的名稱在每一個主機上必須相同。或者，可以在每一個來源或建模節點中的「伺服器」標籤上選取不同的資料來源。

模型選項

在「模型選項」標籤上，您可以選擇是否指定模型的名稱，還是自動產生名稱。您還可以設定評分選項的預設值。

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

名稱已使用時取代現有項。如果您選取此勾選框，則會改寫相同名稱的任何現有模型。

使其可用於評分。您可以在這裡為模型塊對話框上顯示的評分選項設定預設值。有關這些選項的詳細資料，請參閱特定模型塊的「設定」標籤的說明主題。

管理模型

透過 SPSS Modeler 建置 IBM Netezza 或 IBM Data Warehouse 模型將在 SPSS Modeler 中建立一個模型，並在 IBM Data Warehouse 資料庫中建立或取代一個模型。這種 SPSS Modeler 模型將參照資料庫伺服器上儲存的資料庫模型的內容。SPSS Modeler 可以通過將完全相同的產生模型鍵字串儲存在 SPSS Modeler 模型和 Netezza 或 Data Warehouse 模型中來執行一致性檢查。

每個 Netezza 或 Data Warehouse 模型的模型名稱都顯示在「列出資料庫模型」對話框的模型資訊欄下面。SPSS Modeler 模型的模型名稱在 SPSS Modeler 模型的「伺服器」標籤上顯示為「模型鍵」（放置在串流中時）。

「檢查」按鈕可用於檢查 SPSS Modeler 模型和 Netezza 或 Data Warehouse 模型中的模型鍵是否相符。如果在 Netezza 或 Data Warehouse 中找不到具有相同名稱的模型，或模型鍵不相符，那麼說明在建立 SPSS Modeler 模型之後刪除或重新建立了該 Netezza 或 Data Warehouse 模型。

列出資料庫模型

SPSS Modeler 提供了一個用於列出 IBM Data Warehouse 中儲存的模型的對話框，並允許刪除模型。此對話框可以從「IBM Helper 應用程式」對話框以及 IBM Data Warehouse 和 IBM Netezza 資料採礦相關節點的建立、瀏覽和套用對話框中進行存取。針對每個模型顯示下列資訊：

- 模型名稱（模型的名稱，用於對清單進行排序）。
- 擁有人名稱。
- 模型中使用的演算法。
- 模型的現行狀態；例如「已完成」。
- 模型的建立日期。

IBM Data WH 迴歸樹

迴歸樹狀結構是基於樹狀結構的演算法，反覆地分割觀察值的樣本以基於數值目標欄位衍生相同種類的子集。與決策樹一樣，迴歸樹狀結構將資料拆解為子集，其中樹狀結構的葉節點對應於足夠小或足夠統一的子集。選取分割來減少目標屬性值的離散，因此能夠透過葉節點上的平均值合理地進行預測。

IBM Data WH 迴歸樹建置選項 - 樹狀結構成長

您可以針對樹狀結構成長及樹狀結構刪改設定建置選項。

下列建置選項可用於樹狀結構成長：

樹狀結構深度上限。 樹狀結構在根節點下可以成長到的層次數目上限，亦即可以遞迴地分割樣本的次數。預設值是 62，這是用於建模的樹狀結構深度上限。

註：如果模型塊中的檢視器顯示模型的文字呈現，則最多可以顯示 12 個層次的樹狀結構。

分割準則。 這些選項控制何時停止分割樹狀結構。如果您不想使用預設值，請按一下**自訂**並變更這些值。

- **分割評估測量。** 此類別評估測量會評估分割樹狀結構的最佳位置。

  註：目前變異數是唯一可能的選項。
- **分割下限改善。** 必須減少的雜質數量下限，之後才能在樹狀結構中建立新分割。樹狀結構建置的目標是建立具有相似輸出值的子群組以最小化每個節點內的雜質。如果分支的最佳分割會將雜質減少到小於分割準則所指定的數量，則不會分割分支。
- **分割的實例數下限。** 可以分割的記錄數目下限。當剩餘的未分割記錄少於此數目時，不會執行進一步分割。您可以使用此欄位來阻止在樹狀結構中建立小型子群組。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

IBM Data WH 樹狀結構建置選項 - 樹狀結構刪改

您可以使用刪改選項來為迴歸樹狀結構指定刪改準則。刪改的目的是要減少因為移除不會改進新資料預期精確度的過度成長子群組而導致過適的風險。

刪改測量。 刪改測量可確保從樹狀結構移除葉節點之後，預估的模型準確性仍處於可接受的限制內。您可以選取下列其中一個測量。

- **mse。** 均方誤差 - （預設值）測量適合行離資料點有多近。
- **r2。** R 平方 - 測量迴歸模型說明的應變數中變異的比例。
- **Pearson。** Pearson 的相關係數 - 測量正常分配的線性應變數之間的關係強度。
- **Spearman。** Spearman 的相關係數 - 偵測根據 Pearson 的相關性顯示為弱但實際可能很強的非線性關係。

用於刪改的資料。您可以使用部分或全部訓練資料來預估新資料的預期準確性。或者，您可以將指定表格中的個別刪改資料集用於此目的。

- **使用所有訓練資料。**此選項（預設值）會使用所有訓練資料來預估模型準確性。
- **使用特定百分比的訓練資料來進行刪改。**使用此選項以將資料分割為兩個集合，一個用於訓練，另一個用於刪改，並將這裡指定的百分比用於刪改資料。

  如果要指定隨機種子以確保每次執行串流時都以相同的方式來分割資料，請選取**抄寫結果**。您可以在**用於刪改的種子**欄位中指定一個整數，或是按一下**產生**以建立虛擬亂數整數。
- **使用現有表格中的資料。**指定個別刪改資料集的表格名稱以估計模型精確度。這樣做比使用訓練資料更為可靠。不過，此選項可能導致從移除訓練集中去除較大的資料子集，因而會降低樹狀結構的品質。

Netezza 分割叢集

分裂式叢集是一種叢集分析方法，它通過重複執行演算法，使叢集分裂為子叢集，直至達到規定的停止點。

叢集的構造以包含全部訓練實例（記錄）的單個叢集開始。此演算法的第一次疊代將資料集分為兩個子叢集，後續疊代將這些子叢集劃分為進一步的子叢集。停止準則指定為疊代數目上限、資料集細分為的上限層次數以及進行進一步分區所需的下限實例數。

產生的層次叢集樹狀結構可以用於將實例從根叢集向下傳播，以便對它們進行分類，如下例所示。

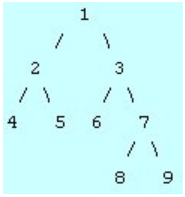

圖 3. 分裂式叢集樹狀結構範例

在每一層次，根據實例與子叢集中心之間的距離來選擇最符合的子叢集。

在套用了階層層次 -1（預設值）的情況下對實例進行評分時，此評分將只傳回一個葉節點叢集，這是因為葉節點由負數指定。在本例中，這將是叢集 4、5、6、8 或 9 中之一。不過，如果將階層層次設定為 2，那麼評分會傳回根叢集下方第二層次上的叢集（4、5、6 或 7）之一。

Netezza 分裂式叢集分析欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

Netezza 分裂式叢集建置選項

可以在「建置選項」標籤中設定用於建置模型的所有選項。您只需按一下**執行**按鈕即可使用所有預設選項來建置模型，但通常您想要自訂建置以用於您的專屬用途。

距離測量。這是用於測量資料點之間的距離的方法；距離越大，表示相異性越大。選項是：

- **歐式距離**。（預設）通過將兩個點用一條直線結合起來計算得出的兩點之間的距離。
- **曼哈頓距離**。兩點之間的距離計算為其坐標之間的絕對差總和。
- **堪培拉距離**。類似於曼哈頓距離，但對更加靠近原點的資料點更加敏感。
- **最大值**。兩點之間的距離計算為任何座標尺寸之差的最大值。

疊代數目上限。此演算法通過執行同一過程的多次疊代來完成操作。此選項可讓您在指定的疊代次數之後停止模型訓練。

叢集樹狀結構的最大深度。這是資料集可以細分為的上限層次數。

重複結果。如果您要設定隨機種子，請選中此勾選框，這將允許您重複進行分析。可指定一個整數或按一下**產生**來建立偽隨機整數。

用於分割的下限實例數。可以分割的記錄數目下限。如果剩餘的未分割記錄數小於此數目，那麼將不執行進一步分割。您可以使用此欄位來防止在叢集樹狀結構中建立非常小的子群組。

IBM Data WH 廣義線性

線性迴歸是一種廣為接受的統計技術，用於根據數值輸入欄位的值對記錄進行分類。線性回歸適合用來最小化預測輸出值與實際輸出值之間差異的直線或平面。線性模型由於訓練簡單且模型應用方便，在構建各種真實世界現象的模型方面用途甚廣。然而，線性模型假設依變數（目標）呈常態分佈，且自變數（預測值）對應變數的影響是線性的。

線性迴歸在多數狀況下非常有用，但是上述假設並不適用。例如，對顧客從給定數量的商品中進行選取的行為建模時，依變數可能呈多項式分佈。同樣，對年齡與收入的關係建模時，收入通常隨年齡增長而增加，但這二者的關聯卻不像一條直線那麼簡單。

對於這些狀況，可以使用廣義線性模型。廣義線性模型擴展了線性迴歸模型，使應變數與預測值之間通過特定的鏈結函數建立關聯，由解釋功能選擇合適的函數。另外，此模型容許依變數呈非正常分佈，例如 Poisson 分佈。

此演算法以疊代方式（次數可達指定的疊代數）求出配適度最佳的模型。在計算最佳擬合時，誤由依變數的預測值和實際值之間的差異的平方和來代表。

IBM Data WH 廣義線性模型欄位選項

在「欄位」標籤上，您可選擇是想要使用已在上游節點中定義的欄位角色設定，還是手動指派欄位。

使用預先定義的角色。 此選項使用上游「類型」節點或者上游來源節點的「類型」標籤中的角色設定（例如目標或預測值）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值以及其他角色，則選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

目標。 選擇一個欄位作為預測的目標。

記錄 ID。 要用作唯一記錄 ID 的欄位。此欄位的值對於每個記錄而言必須是唯一的（例如，客戶 ID 號）。

實例加權。指定欄位以使用實例加權。實例加權是每列輸入資料的加權。依預設，假定所有輸入記錄具有相同的相對重要性。可以通過向輸入記錄分配各項加權來變更重要性。指定的欄位必須包含每列輸入資料的數值加權。

預測值（輸入）。 選取一個或多個輸入欄位。此動作與在「類型」節點中將欄位的角色設定為輸入類似。

IBM Data WH 廣義線性模型選項 - 一般

在「模型選項」標籤上，您可以選擇是否指定模型的名稱，還是自動產生名稱。您可以進行有關模型的多項設定，像是鏈結函數、輸入欄位的互動（如果有的話）以及設定評分選項的預設值。

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

欄位選項。可以指定用於建立模型的輸入欄位的角色。

一般設定。這些設定關係到演算法的停止準則。

- **最大疊代次數。** 演算法最多進行疊代的次數；最小值為 1 次，預設為 20 次。
- **最大錯誤 (1e)。**最大錯誤值（以科學記號表示法），達到此值後，此演算法應停止尋找最佳擬合模型。最小值為 0，預設值為 -3，表示 1E-3 或 0.001。
- **不顯著誤值的臨界值 (1e)。**這是一個值（以科學記號表示法表示），所有小於此值的錯誤均被視為具有零值。最小值為 -1，預設值為 -7，表示誤值若低於 1E-7 （或 0.0000001），那麼被視為不顯著。

分佈設定。這些設定與應變數（目標變數）的分佈相關。

- **回應變數的分配。**分佈類型；這是下列其中一項：**白努利**（預設），**高斯**、**Poisson**、**二項式**、**負二項**、**Wald**（反向高斯）和**Gamma**。
- **參數。** （僅限 Poisson 或二項式分佈）您必須在**指定參數**欄位中指定下列其中一個選項：
  - 要自動從資料估計參數，請選取**預設值**。
  - 要容許對分佈準概似進行最佳化，請選取**準**。
  - 要明確指定參數值，請選取**明確地**。

  （僅限二項式分佈）您必須按二項式分佈的要求指定將用作試驗欄位的輸入表格欄。此欄包含二項式分佈的試驗數。

  （僅限負二項式分佈）您可以使用預設值 -1 或指定不同的參數值。

鏈結函數設定。這些設定與關聯函數相關，後者用於使依變數與預測值相關。

- **鏈結函數。** 要使用的函數，這是下列其中一項：**Identity**、**Inverse**、**Invnegative**、**Invsquare**、**Sqrt**、**Power**、**Oddspower**、**Log**、**Clog**、**Loglog**、**Cloglog**、**Logit**（預設）、**Probit**、**Gaussit**、**Cauchit**、**Canbinom**、**Cangeom** 和 **Cannegbinom**。
- **參數。** （僅限於 Power 或 Oddspower 鏈結函數），如果鏈結函數為 **Power** 或 **Oddspower**，您可以指定其參數值。請選擇是指定一個值，還是使用預設值 1。

IBM Data WH 廣義線性模型選項 - 互動

「互動」控制台包含了選項，可以指定互動行為（即，輸入欄位間的product作用）。

欄互動。選取此勾選框來指定輸入欄位之間的互動。若無互動行為，請將選項框留空。

通過從來源清單中選取一個或多個欄位，並將其拖曳到互動清單，在模型中輸入互動。所建立的互動類型取決於將選項拖放到何種焦點資訊值。

- **主要。** 拖入的欄位作為個別的主要互動，顯示在互動清單的底部。
- **雙向。** 所有可能配對的拖入欄位作為雙向互動，顯示在互動清單的底部。
- **三向。** 所有可能配成三項值的拖入欄位，均作為三向互動，顯示在互動清單的底部。
- ***.** 放入的全部欄位的組合作為單一互動顯示在互動清單底部。

包含截距。模型中通常會包含截距，如果可以假設資料穿過原點，就可以將截距排除在外。

對話框按鈕

顯示在右側的按鈕允許您對模型中使用的項目進行變更。

圖 *4.* 「刪除」按鈕

透過選取您想要刪除的項目，並按一下刪除按鈕，從模型中刪除項目。

圖 *5.* 「重新排序」按鈕

透過選取您想要重新排序的項目，並按一下上移鍵或下移鍵，對模型內的項目進行重新排序。

圖 *6.* 「自訂互動」按鈕

新增自訂項目

您可以採用 *n1*x1*x1*x1..* 格式來指定自訂互動。請從**欄位**清單中選取一個欄位，按一下右箭頭按鈕將該欄位新增到**自訂項目**，按一下**按 ***，選取下一個欄位，再按一下右箭頭按鈕，依此類推。當您已完成構建自訂互動，可按一下**新增項目**將其傳回「互動」畫面。

IBM Data WH 廣義線性模型選項 - 評分選項

使其可用於評分。您可以在這裡為模型塊對話框上顯示的評分選項設定預設值。請參閱第 76 頁的『IBM Data WH 廣義線性模型塊 - 設定標籤』主題，以取得更多資訊。

- **包含輸入欄位**。如果您需要將該輸入欄位連同預測值一同顯示在模型輸出中，請選取該選框。

IBM Data WH 決策樹

決策樹是代表分類模型的階層式結構。使用決策樹模型，您可以開發分類系統來預測或分類一組訓練資料中的未來觀察。分類採取樹狀結構的形式，其中分支代表分類中的分割點。分割會將資料遞迴地分為子群組，直到達到停止點為止。停止點處的樹狀結構節點稱為**葉節點**。每片樹葉節點分配一個標籤（稱為 **類別標籤**）給其子群組或類別會員。

實例加權和類別加權

依預設，假定所有輸入記錄和類別具有相同的相對重要性。您可以透過對兩個項目或其中之一的成員指派個別加權，對此進行變更。這樣做可能很有用，例如，如果您訓練資料中的資料點實際上未分佈在種類之間。加權可讓您調整模型，以便您可以對未充分呈現在資料中的那些種類進行補償。增加某個目標值的加權應該增加該類型的正確預測百分比。

在「決策樹」建模節點中，可以指定兩種類型的加權。**實例加權**分配一個加權給每一列輸入資料。對於大部分觀察值，加權通常指定為 1.0，同時僅對那些比大部分觀察值更加重要或更加不重要的觀察值指定較大或較小的值，如下表格所示。

表 *5.* 實例加權範例

記錄 ID	目標	實例加權
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

類別加權對目標欄位的每個種類分配一個加權，如下表格所示。

表 *6.* 類別加權範例

類別	類別加權
drugA	1.0
drugB	1.5

可以同時使用這兩種類型的加權，在這種情況下，它們將相乘並用作實例加權。因此，如果將之前的兩個範例一起使用，那麼此演算法將使用下表格所示的實例加權。

表 *7.* 實例加權計算範例

記錄 ID	計算	實例加權
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Netezza 決策樹欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色：此選項使用上游類型節點（或上游來源節點的「類型」標籤）的角色設定（目標、預測值等等）。

使用自訂欄位指派。 要手動分配目標、預測值和其他角色，請選中此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

若要選取清單中的全部欄位，請按一下**全部**按鈕，或按一下個別測量層次按鈕來選取具有該測量層次的所有欄位。

目標。 選取一個欄位作為預測目標。

記錄 ID。 要用作唯一記錄 ID 的欄位。此欄位的值針對每個記錄必須是唯一的（例如，客戶身分證號碼）。

實例加權。在這裡指定欄位會讓您使用實例加權（每個輸入資料列一個加權），而不是預設值類別加權（每個目標欄位的種類一個加權）。您在這裡指定的欄位必須包含每一列輸入資料的數字加權。請參閱第 56 頁的『實例加權和類別加權』主題，以取得更多資訊。

預測值（輸入）。 選取一個或多個輸入欄位。這與在「類型」節點中將欄位角色設為輸入類似。

IBM Data WH 決策樹建置選項

下列建置選項可用於樹狀結構成長：

成長測量。 這些選項可控制測量樹狀結構成長的方式。

- **雜質測量。** 此測量會評估分割樹狀結構的最佳位置。這是測量資料子群組或區段中的變異性。較低的雜質測量指出群組中大部分成員的準則或目標欄位值類似。

  支援的測量為**熵** 和 **Gini**。這些測量以分支的種類成員資格的機率為基礎。
- **樹狀結構深度上限。** 樹狀結構在根節點下可以成長到的層次數目上限，亦即可以遞迴地分割樣本的次數。此內容的預設值是 10，您可以為此內容設定的最大值為 62。

  註：如果模型塊中的檢視器顯示模型的文字呈現，則最多可以顯示 12 個層次的樹狀結構。

分割準則。 這些選項控制何時停止分割樹狀結構。

- **分割下限改善。** 必須減少的雜質數量下限，之後才能在樹狀結構中建立新分割。樹狀結構建置的目標是建立具有相似輸出值的子群組以最小化每個節點內的雜質。如果分支的最佳分割會將雜質減少到小於分割準則所指定的數量，則不會分割分支。
- **分割的實例數下限。** 可以分割的記錄數目下限。當剩餘的未分割記錄少於此數目時，不會執行進一步分割。您可以使用此欄位來阻止在樹狀結構中建立小型子群組。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

IBM Data WH 決策樹節點 - 類別加權

在這裡您可以為個別類別指派加權。預設是指派值 1 給所有類別，使它們的加權相類別。透過為不同類別標籤指定不同分值加權，您指示演算法相應地對特定類別的訓練集進行加權。

若要變更加權，請在**加權**直欄中按兩下加權，然後進行所需的變更。

值。類別標籤的集，衍生自目標欄位的可能值。

加權。要指派給特定類別的加權。將較高的加權指派給某個類別會讓模型對該類別相較於其他類別而言更為敏感。

您可以結合使用類別加權與實例加權。請參閱第 56 頁的『實例加權和類別加權』主題，以取得更多資訊。

IBM Data WH 決策樹節點 - 樹狀結構刪改

您可以使用刪改選項來為決策樹指定刪改準則。刪改的目的是要減少因為移除不會改進新資料預期精確度的過度成長子群組而導致過適的風險。

刪改測量。預設刪改測量**準確性**可確保從樹狀結構移除葉節點之後，預估的模型準確性仍處於可接受的限制內。如果要在套用刪改時將類別加權納入考量，請使用替代測量**加權的準確性**。

用於刪改的資料。您可以使用部分或全部訓練資料來預估新資料的預期準確性。或者，您可以將指定表格中的個別刪改資料集用於此目的。

- **使用所有訓練資料**。此選項（預設值）會使用所有訓練資料來預估模型準確性。
- **使用特定百分比的訓練資料來進行刪改**。使用此選項以將資料分割為兩個集合，一個用於訓練，另一個用於刪改，並將這裡指定的百分比用於刪改資料。

  如果要指定隨機種子以確保每次執行串流時都以相同的方式來分割資料，請選取**抄寫結果**。您可以在**用於刪改的種子**欄位中指定一個整數，或是按一下**產生**以建立虛擬亂數整數。
- **使用現有表格中的資料**。指定個別刪改資料集的表格名稱以估計模型精確度。這樣做比使用訓練資料更為可靠。不過，此選項可能導致從移除訓練集中去除較大的資料子集，因而會降低樹狀結構的品質。

IBM Data WH 線性回歸

線性模型會根據目標與一或多個預測值之間的線性關係預測連續目標。線性迴歸模型僅限於直接建模線性關係，但它相對簡單，用於評分的數學公式也易於解釋。與其他更優化的迴歸方法演算法產生的模型相比，線性模型快速、高效，並且簡單易用，但其應用範圍有限。

IBM Data WH 線性回歸建置選項

可以在「建置選項」標籤中設定用於建置模型的所有選項。您只需按一下**執行**按鈕即可使用所有預設選項來建置模型，但通常您想要自訂建置以用於您的專屬用途。

使用特異值分解來求解方程式。使用特異值分解矩陣而不是原始矩陣，不但能夠更有效地應對數值誤，並且可以加快計算過程。

在模型中包括截距。包含截距可以增加解的整體精確度。

計算模型診斷。此選項將導致對模型計算大量診斷資訊。這些結果將儲存在矩陣或表格中，以供稍後檢視。診斷選項包含 r 平方、殘差平方和、估計變異數、標準差、p 值和 t 值。

這些診斷資訊與模型的有效性和可用性相關。您應當針對底層資料執行其他診斷，以確保其符合線性假設。

IBM Data WH KNN

最近鄰法分析是以和其他觀察值的親緣性為基礎來分類觀察值的方法。在機器學習中，這是辨認資料形式的方法，完全不需要確切符合任何已儲存的形式或觀察值。相似的觀察值會彼此相鄰，相異的觀察值則會彼此相隔。因此，兩個觀察值相距的距離可用來判斷彼此的相異性。

彼此接近的觀察值稱為「鄰接項」。新的觀察值 (保留) 存在時，會計算模式中各觀察值的距離。計算最相似觀察值的分類（最近鄰法），新觀察值會放在包含最近鄰法中個數最多的類別。

您可以指定要檢查的最近鄰接項數目；此值稱為 *k*。圖片顯示如何使用兩個不同的 *k* 值來分類新的觀察值。當 *k* = 5 時，新的觀察值會放置在種類 *1* 中，因為大部分最近鄰接項都屬於種類 *1*。當 *k* = 9，則新的觀察值會放置在種類 *0* 中，因為大部分最近鄰接項都屬於種類 *0*。

最近鄰法分析也可以用來計算連續目標的數值。在此狀況下，會使用最近鄰的平均數或中位數目標值來取得新觀察值的預測值。

IBM Data WH KNN 模型選項 - 一般

在「模型選項 - 一般」標籤上，您可以選擇是指定模型名稱，還是自動產生名稱。您還可以設定那些控制如何計算最近鄰接項數量的選項，並設定相關選項以獲得增強的模型效能和準確度。

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

鄰接項

距離測量。這是用於測量資料點之間的距離的方法；距離越大，表示相異性越大。選項是：

- **歐式距離**。（預設）通過將兩個點用一條直線結合起來計算得出的兩點之間的距離。
- **曼哈頓距離**。兩點之間的距離計算為其坐標之間的絕對差總和。
- **堪培拉距離**。類似於曼哈頓距離，但對更加靠近原點的資料點更加敏感。
- **最大值**。兩點之間的距離計算為任何座標尺寸之差的最大值。

最近鄰數目 (k)。 特定觀察值的最近鄰接項數量。請注意，使用較大相鄰數目未必可得出較精確的模式。

通過選取 *k*，您可以控制項在防止過度配適（這可能很重要，尤其對於「雜訊」資料）和求解（針對類似實例產生不同預測結果）之間的平衡。您通常需要針對每個資料集來調整 *k* 值，其一般值在 1 至幾十之間。

增強效能和準確度

在計算距離前規範測量結果。如果已選取，該選項將標準化連續輸入欄位的測量結果，然後再計算距離值。

對大型資料集使用核心集以增加效能。如果已選取，該選項將針對大型資料集採用核心集取樣以加快計算過程。

IBM Data WH KNN 模型選項 - 評分選項

在「模型選項 - 評分選項」標籤上，您可以設定評分選項的預設值，並為個別類別指定相對加權。

可用於評分

包含輸入欄位。指定依預設是否將輸入欄位併入在評分中。

類別加權

如果您要變更個別類別在建立模型中的相對重要性，請使用此選項。

注意：僅當您使用 KNN 進行分類時，此選項才處於啟用狀態。如果您要執行迴歸方法（即，目標欄位類型為連續），此選項將被停用。

預設是指派值 1 給所有類別，使它們的加權相類別。透過為不同類別標籤指定不同分值加權，您指示演算法相應地對特定類別的訓練集進行加權。

若要變更加權，請在**加權**直欄中按兩下加權，然後進行所需的變更。

值。類別標籤的集，衍生自目標欄位的可能值。

加權。要指派給特定類別的加權。將較高的加權指派給某個類別會讓模型對該類別相較於其他類別而言更為敏感。

IBM Data WH K-Means

K-Means 節點實作提供叢集分析方法的 *k* 平均數演算法。您可以使用此節點將資料集叢集到不同的群組。

演算法是基於距離的叢集演算法，根據距離度量（函數）來測量資料點之間的親緣性。資料點根據使用的距離度量指派給最近的叢集。

演算法透過對相同的基本處理程序執行數次疊代進行運算，程序中的每一個訓練實例都指派給最近的叢集（就指定的距離函數而言，則套用至距離及叢集中心）。然後會重新計算所有叢集中心作為指派給特定叢集的實例的平均屬性值向量。

IBM Data WH K-Means 欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Data WH K-Means 建置選項標籤

透過設定建置選項，您可以自訂用於您專屬用途的模型建置。

如果要使用預設選項來建置模型，請按一下**執行**。

距離測量 此參數定義用於測量資料點之間距離的方法。距離越大指出相異性越大。請選取下列其中一個選項：

- **歐基里得。** 歐基里得測量是兩個資料點之間的直線距離。

- **正規化歐基里得。** 常態化歐基里得測量類似於歐基里得測量，但前者以標準偏差平方進行常態化。與歐基里得測量不同，常態化歐基里得測量也是比例不變。
- **馬氏距離。**馬氏距離測量是考慮輸入資料的相關性的通用性歐式距離測量。與正規化歐式距離測量一樣，馬氏距離測量具有尺度不變性。
- **曼哈頓距離。**曼哈頓距離測量是計算為其座標的絕對差總和的兩個資料點之間的距離。
- **堪培拉距離。**堪培拉距離測量類似於曼哈頓距離測量，但對距離原點越近的資料點越敏感。
- **最大值。** 最大值測量是計算為任何座標尺寸之差的最大值的兩個資料點之間的距離。

叢集數目。 此參數定義要建立的叢集數目。

最大疊代次數。 演算法對相同的處理程序執行數次疊代。此參數定義模型訓練停止之前的疊代次數。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

複製結果。 如果要設定隨機種子以抄寫分析，請選取這個勾選框。您可以指定一個整數，也可以透過按一下**產生**來建立虛擬隨機整數。

IBM Data WH Naive Bayes

Naive Bayes 是用來解決分類問題的著名演算法。此模型將所有建議預測變數視為彼此獨立，因此被稱為樸素。Naive Bayes 是一個快速可調式演算法，能夠計算屬性及目標屬性組合的條件式機率。從訓練資料中，會建立獨立機率。鑒于每個輸入變數中出現的每個值種類，此機率會提供每個目標類別的可能性。

Netezza Bayes Net

貝葉斯網路是一個模型，它顯示資料集中的變數，以及這些變數之間隨機的或條件式的獨立性。使用貝葉斯網絡節點，可以透過將觀察到並記錄下的證據與實際常識結合起來建立機率模型，以透過使用表面看上去不相關的屬性確定發生的可能性。

Netezza 貝葉斯網絡欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

對於此節點，只有評分才需要目標欄位，所以此欄位未顯示在此標籤上。您可以在「類型」節點、此節點的「模型選項」標籤或模型塊的「設定」標籤上設定或變更目標。請參閱第 71 頁的『Netezza 貝葉斯網絡塊 - 「設定」標籤』主題，以取得更多資訊。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

Netezza 貝葉斯網絡建置選項

可以在「建置選項」標籤中設定用於建置模型的所有選項。您只需按一下**執行**按鈕即可使用所有預設選項來建置模型，但通常您想要自訂建置以用於您的專屬用途。

基底指標。為第一個屬性（輸入欄位）分配的數值 ID，以方便內部管理。

樣本大小。 當屬性數量過多並可能導致正在處理時間過長 時，要採用的樣本大小。

在執行期間顯示更多資訊。如果已勾選此框（預設情況），那麼將在訊息對話框中顯示附加的進度資訊。

Netezza 時間序列

時間序列 是一個數值型序列，以時間上前後接續的（但不必是規律的）點計量 - 例如，每日股票價格或每周銷售資料。分析此類資料有時會很有用，例如，用於突顯某些行為，像是趨勢或週期性變動（一項重複性的型樣），或是通過過去的事件預測未來的行為的時候。

Netezza 時間序列支援下列時間序列演算法。

- 光譜分析
- 指數平滑化
- 自身迴歸整合移動均數 (ARIMA)
- 週期性趨勢分解

這些演算法將時間序列分解成一個趨勢和一個週期性成分。再對這些成分進行分析，以建立出一個可用於預測的模型。

光譜分析用於識別時間序列中的週期性行為。對於包含多個底層週期性的時間序列，或者在資料中出現大數量隨機雜訊時，光譜分析提供了最為明確的方法來識別週期性成分。此方法通過將數列從時間網域轉換為一系列頻率網域，偵測週期性行為的頻率。

指數平滑化是使用之前數列觀察的加權值預測未來值的預測方法。採用指數平滑化，觀測所造成的影響隨時間推移而以指數級減少。該方法一次預測一個點，當有新資料進入時再對預測作出調整，對資料的加法入、趨勢以及週期性變化作出整體性考慮。

ARIMA 模型提供了比指數平滑化模型更複雜的方法進行趨勢建模和週期性成分建模。此方法涉及明確指定自動迴歸階數和移動平均數階數以及差異分析度。

附註：在實際情況中，如果您要包括可以協助說明預測數列行為的預測值，例如，郵寄的型錄數或公司網頁的點閱數，則 ARIMA 模型會非常有用。而指數平滑化模型在說明時間序列的行為時，並不試圖解釋其行為原因。

週期性趨勢分解先將週期性行為從時間序列中刪除，以便進行趨勢分析，之後再為趨勢選取一個基本形狀，例如一個二次函數。這些基本形狀帶有數個已確定值的參數，以將殘差的平均方差（即時間序列擬合值與觀察值之間的差異）減到最小。

Netezza 時間序列值的插補

插補是估計時間序列中遺漏的資料並插補一個值的過程。

如果時間序列的時間間隔有規律，但某些值不出現，那麼可以使用線性插補來估計這些遺漏值。考慮如下示例數列中某機場航廈每月的乘客抵達人數。

表 *8.* 客運航站樓的每月抵達數

月	乘客
3	3,500,000
4	3,900,000
5	-
6	3,400,000
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

在此例中，通過線性插補可以估計出第 5 個月的遺漏值為 3,650,000 （第 4 個月與第 6 個月的中間點）。

不規律間隔的處理方式有所不同。考慮如下數列中的溫度讀取數。

表 *9.* 溫度讀數

日期	時間(M)	溫度
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

這裡，我們有 3 天內從 3 個點所取得的一系列讀取數，但除了少數讀取數外，大部分讀取數的獲取時間並不相同。另外，其中只有 2 天是連續的。

這種狀況可以通過下列兩種方法中的一種來處理：計算聚合，或者確定步長大小。

聚合可能是根據對資料語義的瞭解，使用公式計算得出的每日聚合。執行這一步會得到如下的資料集。

表 *10.* 溫度讀數（已聚集）

日期	時間(M)	溫度
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	Null
2011-07-27	24:00	72

另外，此演算法可以將該數列視為差異數列以確定適當的步驟大小。在此例中，演算法所確定的步驟大小可能是 8 個小時，這樣會得到如下結果。

表 *11.* 已計算步驟大小的溫度讀數

日期	時間(M)	溫度
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

在這裡，只有 4 個讀取數與原始測量值對應，但借著原始系列中的其他已知值，遺漏的值可再次通過插補計算出來。

Netezza 時間序列欄位選項

在「欄位」標籤上，指定來源資料輸入欄位的角色。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

目標。 選擇一個欄位作為預測的目標。這必須是測量層次為「連續」的欄位。

（預測值）時間點。（必填）這是包含時間序列的日期或時間值的輸入欄位。這必須是測量層次為「連續」或「種類」且資料儲存類型為日期、時間、時間戳記或數值的欄位。此處指定的欄位的資料儲存類型同時也定義了此建模節點的其他標籤上某些欄位的輸入類型。

（預測值）時間序列 ID（按）。包含時間序列 ID 的欄位。如果輸入項包含一個以上的時間序列，那麼使用這個欄位。

Netezza 時間序列建置選項

建置選項分兩個層次：

- 基本 - 演算法選項、插補以及所要採用的時間範圍的設定。
- 進階 - 設定預測

本節說明基本選項。

可以在「建置選項」標籤中設定用於建置模型的所有選項。您只需按一下**執行**按鈕即可使用所有預設選項來建置模型，但通常您想要自訂建置以用於您的專屬用途。

演算法

這些是有關所要採用的時間序列演算法的設定。

演算法名稱。選擇要使用的時間序列演算法。可選的演算法包括**光譜分析**、**指數平滑化**（預設）、**ARIMA** 或**週期性趨勢分解**。請參閱第 62 頁的『Netezza 時間序列』主題，以取得更多資訊。

趨勢。（僅限於指數平滑化）如果時間序列呈現出一種趨勢，那麼簡單指數平滑化效果不佳。若有趨勢，使用該欄位來指定它，以使演算法可將它納入考量。

- **系統已決定。** （預設值）系統嘗試為此參數尋找最佳值。
- **無(N)。** 時間序列未呈現趨勢。
- **可加性(A)。** 隨著時間推移而穩定增加的趨勢。
- **減幅可加性 (DA)**。隨著時間推移最終會消失的可加性趨勢。
- **相乘性(M)。** 該趨勢也是隨時間而增加，但速度通常比穩定可加性趨勢快。
- **減幅相乘性 (DM)**。隨著時間推移最終會消失的增幅趨勢。

週期性。（僅限於指數平滑法）使用該欄位指定時間序列中的資料是否呈現週期性特徵。

- **系統已決定。** （預設值）系統嘗試為此參數尋找最佳值。
- **無(N)。** 時間序列未呈現週期性型樣。
- **可加性(A)。** 週期性浮動型樣呈現隨時間推移穩定上行的趨勢。
- **相乘性(M)。** 具有與遞增的週期性相同的特點，但除此之外，其週期性浮動的電流幅度（高低點間的距離）圍繞著浮動的整體上行趨勢而增加。

對 ARIMA 使用系統確定的設定。（僅限於 ARIMA）如果您希望由系統來確定 ARIMA 演算法的設定，請選擇此選項。

指定。 （僅限於 ARIMA）選擇此選項並按一下按鈕可以手動指定 ARIMA 設定。

插補法

時間序列來源資料包含遺漏值時，選擇一種方法來插入估計值以填補資料中的間隙。請參閱第 62 頁的『Netezza 時間序列值的插補』主題，以取得更多資訊。

- **線性。** 時間序列的間隔有規律，僅僅是某些值缺失時，請選擇此方法。
- **指數自由曲線**。把資料值以高速增加或減少的已知點擬合成一條平滑曲線。
- **立體模式自由曲線**。將已知資料點擬合成一條平滑曲線來估計遺漏的值。

時間範圍

在此可選擇是否使用全範圍的時間序列資料，或時間序列資料的一個連續的子集合來建立模型。這些欄位的有效輸入由針對「欄位」標籤上「時間點」指定之欄位的資料儲存類型定義。請參閱第 64 頁的『Netezza 時間序列欄位選項』主題，以取得更多資訊。

- **使用資料中的最早和最晚時間**。如果您想要使用全範圍的時間序列資料，請選擇此選項。
- **指定時間範圍**。如果您希望只使用時間序列的一部分，請選擇此選項。使用**最早時間（自）**與**最晚時間（至）**欄位來界定範圍。

ARIMA 結構

指定 ARIMA 模型中各種非週期性成分及週期性成分的值。在每一種情況下，均先將運算子設定為 **=**（等於）或 **<=**（小於或等於），然後指定相鄰欄位的值（小於或等於），然後指定相鄰欄位的值。指定度數的所有值都必須為非負整數。

非週期性。模型中各非週期性成分的值。

- **自相關係數度 (p)**。模式中自我迴歸階數的個數。自我迴歸階數指定要從數列中取用哪個先前值來預測目前值。例如，自我迴歸階數 2 指定要使用過去兩段時間的序列值來預測目前值。
- **衍生 (d)**。指定在估計模式之前套用至數列的差分階數。當趨勢存在時（包含趨勢的數列一般都是非平穩性數列，但 ARIMA 模式中假定數列為穩定性），就必須對數列進行差分，以移除趨勢的影響。差分的階數對應於數列趨勢的程度，第一階差分代表線性趨勢，第二階差分代表二次趨勢，依此類推。
- **移動均數 (q)**。模式中移動平均階數的個數。移動平均階數指定如何使用先前數值的數列平均數離差來預測目前值。例如，移動平均階數 1 和 2 指定在預測數列的目前值時，要考慮最後兩段時間中各個數列平均值的離差。

週期性。 週期性自相關係數 (SP)、衍生 (SD) 以及移動均數 (SQ) 成分扮演與其非週期性對應成分相同的角色。但對週期性階數來說，目前序列值是受由一個或多個週期性期間分隔的先前序列值影響。例如，以每月資料（週期性期間為 12）來說，週期性階數 1 代表目前序列值受 12 個週期之前的序列值影響。如此對每月資料來說，週期性階數 1 就與指定非週期性階數 12 相同。

僅當在資料中偵測到週期性趨勢時，或您從「進階」標籤中指定了「期間設定」時，才需用到週期性設定。

Netezza 時間序列建置選項 - 進階

您可以使用進階設定來指定預測選項。

對模型建置選項使用系統確定的設定。如果您希望由系統確定進階設定，請選擇此選項。

指定。 如果您希望手動指定進階選項，請選擇此選項。（演算法為光譜分析時，該選項不可選。）

- **期間/期間單位**。這是一個時間期間，在此之後，時間序列的一些特性行為不斷重複。例如，對於一個包含每週銷售數字的時間序列，您可以指定期間為 1，單位為 星期。**期間**必須為非負整數；**期間單位**可以是**毫秒**、**秒**、**分**、**小時**、**天**、**星期**、**季**或者**年**之一。如果未設定**期間**，或時間類型不為數值，請勿設定**期間單位**。但是，如果您指定**期間**，您必須也指定**期間單位**。

預測設定。您可以選擇在特定復原點之前進行預測，或者在特定復原點進行預測。這些欄位的有效輸入由針對「欄位」標籤上「時間點」指定之欄位的資料儲存類型定義。請參閱第 64 頁的『Netezza 時間序列欄位選項』主題，以取得更多資訊。

- **預測範圍**。僅當您只想指定預測結束點時，才應選擇此選項。預測將到此復原點為止。
- **預測時間**。選擇此選項可以指定一個或多個復原點，作為預測復原點。按一下**新增**在時間點的表格中增加一列。要刪除一列，請選定該列，再按一下**刪除**。

Netezza 「時間序列」模型選項

在「模型選項」標籤上，您可以選擇是否指定模型的名稱，還是自動產生名稱。您還可以設定模型輸出選項的預設值。

模型名稱 您可以根據目標或 ID 欄位（如果未指定此類欄位，則根據模型類型）自動產生模型名稱，或者指定自訂名稱。

使其可用於評分。您可以在這裡為模型片段對話框上顯示的評分選項設定預設值。

- **在輸出結果中包含歷程值**。按照預設，模型輸出不包含資料的歷程資料值（之前用來進行預測的值）。選取此勾選框以包含這些值。
- **在輸出結果中包含插入值**。如果您選擇在輸出中包含歷程值，且希望同時包含插入值（如果有），請選中此框。請注意，插補僅對歷程資料起作用，所以如果未選取**在輸出中包含歷程記錄值**，那麼此框不提供。請參閱第 62 頁的『Netezza 時間序列值的插補』主題，以取得更多資訊。

IBM Data WH TwoStep

TwoStep 節點實作 TwoStep 演算法，可提供叢集資料除以大資料集的方法。

您可以在考量可用資源（例如，記憶體和時間限制）時使用此節點來叢集資料。

TwoStep 演算法是透過下列方式叢集資料的資料庫採礦演算法：

1. 叢集特性 (CF) 樹狀結構已建立。這一高度平衡的樹狀結構儲存階層式叢集的叢集特性，其中類似的輸入記錄會變成相同樹狀結構節點的一部分。
2. CF 樹狀結構的葉節點是階層式地叢集到記憶體內，以產生最終的叢集結果。最佳叢集數會自動決定。如果您指定叢集數上限，則會決定指定限制內的最佳叢集數。
3. 叢集結果在另一個步驟中精簡，其中資料會套用與 K-Means 演算法類似的演算法。

IBM Data WH TwoStep 欄位選項

透過設定欄位選項，您可以指定使用在上游節點中定義的欄位角色設定。您還可以手動指派欄位。

選取項目。 選擇此選項可使用上游「類型」節點或上游來源節點的「類型」標籤中的角色設定。例如，角色設定是目標及預測值。

使用自訂欄位指派。 如果您要手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭可將此清單中的項目手動指派給右側的角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Data WH TwoStep 建置選項

透過設定建置選項，您可以自訂用於您專屬用途的模型建置。

如果要使用預設選項來建置模型，請按一下**執行**。

距離測量 此參數定義用於測量資料點之間距離的方法。距離越大指出相異性越大。選項是：

- **對數概似。** 概似量數會對變數進行機率分配。連續變數假設為常態分配，而類別變數則假設為多項式分配。所有變數皆假設為自變數。
- **歐基里得。** 歐基里得測量是兩個資料點之間的直線距離。
- **正規化歐基里得。** 常態化歐基里得測量類似於歐基里得測量，但前者以標準偏差平方進行常態化。與歐基里得測量不同，常態化歐基里得測量也是比例不變。

叢集數目。 此參數定義要建立的叢集數目。選項是：

- **自動計算叢集數目。** 叢集數會自動計算。您可以在**上限**欄位中指定叢集數上限。
- **指定叢集數目。** 指定應建立多少叢集。

統計量。 此參數定義模型中包含的統計量數量。選項是：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量

- **無。** 僅包含對模型進行評分所需的統計量。

複製結果。 如果要設定隨機種子以抄寫分析，請選取這個勾選框。您可以指定一個整數，也可以透過按一下**產生**來建立虛擬隨機整數。

IBM Data WH PCA

主成份分析 (PCA) 是一種強大的資料削減技術，用於降低資料複雜性。PCA 可以找出輸入欄位的線性組合，這些組合能夠最好地捕獲整個欄位集合中的變異數，且組合中的各個成分相互正交（不相關）。其目標在於找到有效概括原始輸入欄位集中的資訊的少量衍生欄位（主成分）。

IBM Data WH PCA 欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Data WH PCA 建置選項

可以在「建置選項」標籤中設定用於建置模型的所有選項。您只需按一下**執行**按鈕即可使用所有預設選項來建置模型，但通常您想要自訂建置以用於您的專屬用途。

在計算 PCA 之前集中資料。如果已勾選此選項（預設情況），那麼將在進行分析前集中資料（也稱為「平均數消去法」）。為了確保第一個主成分說明上限變異數的方向，需要集中資料，否則該成分可能更接近於資料的平均數。如果已採用這種方式來準備資料，您通常可以取消勾選此選項以提升效能。

在計算 PCA 之前執行資料換算。該選項將在分析之前執行資料換算。這樣做可以在以不同的單位測量不同的變數時使分析不那麼任意。如果資料採用最簡單的形式，可以通過將每個變數除以其標準差來完成資料定標。

使用不太準確但速度較快的方法來計算 PCA。該選項將導致演算法使用低準確度但快速的方法 (forceEigensolve) 來尋找主成分。

管理 IBM Data WH 和 Netezza 模型

IBM Data Warehouse 和 IBM Netezza Analytics 模型可以使用與其他 IBM SPSS Modeler 模型相同的方式新增到畫布和模型選用區中，並以幾乎相同的方式來使用。但是，也有幾點重大差異，比如 IBM SPSS Modeler 中建立的每個 IBM Data Warehouse 或 IBM Netezza Analytics 模型實際參照的是儲存在資料庫伺服器上的模型。因此，要使串流正常工作，必須將其連接至建立模型所在的資料庫，並且模型表格未被外部處理程序修改。

對 IBM Data Warehouse 和 IBM Netezza Analytics 模型評分

模型在畫布上由金色的模型塊圖示代表。區塊的主要目的是對資料評分以產生預測，或進一步分析模型內容。評分以一或多個額外資料欄位的形式來新增，可透過將「表格」節點附加至區塊並執行串流分支設為可見，本節稍後會加以說明。部分區塊對話框（例如「決策樹」或「回歸樹」的區塊對話框）額外會有一個「模型」標籤，用來提供模型的視覺化表示法。

額外欄位由目標欄位名稱中新增的字首 $<id>- 加以區分，其中 <id> 取決於模型，用來識別所新增資訊的類型。在每個模型塊的主題中說明了不同的 ID。

若要檢視評分，請完成下列步驟：

1. 將「表格」節點連接至模型塊。
2. 開啟「表格」節點。
3. 按一下「**執行**」。
4. 捲動至表格輸出視窗的右邊，以檢視額外的欄位及其評分。

IBM Data WH 和 Netezza 模型塊 - 伺服器標籤

在「伺服器」標籤上，可以設定模型評分的伺服器選項。您可以繼續使用在上游指定的伺服器連線，也可將資料移動到在此指定的其他資料庫。

IBM Data Warehouse 伺服器詳細資料。在這裡，可以指定要用於模型的資料庫的連線細節。

- **使用上游連線**。（預設值）使用上游節點（例如資料庫來源節點）中指定的連線詳細資料。僅當所有上游節點都能夠使用 SQL 回送功能時，此選項才有效。在此情況下，無需將資料移出資料庫，因為 SQL 完全實現所有的上游節點。
- **將資料移至資料庫**。將資料移至您在這裡指定的資料庫。這樣，即使資料位於另一個 IBM Data Warehouse 資料庫或者另一供應商的資料庫中，甚至位於純文字檔案中，也仍然可以進行建模。此外，如果因節點未執行 SQL 推回而擷取了資料，則資料會移回至在這裡指定的資料庫。按一下**編輯**按鈕以瀏覽並選取連線。

  注意：
  IBM Netezza Analytics 和 IBM Data Warehouse 通常與非常大型的資料集配合使用。在資料庫之間傳輸大數量資料，或者從資料庫中取出或存入大數量資料，可能非常耗時，應盡可能避免。

模型名稱。模型的名稱。該名稱的顯示僅供參考；無法在此對其進行變更。

IBM Data WH 決策樹模型塊

決策樹模型塊顯示建模作業的輸出，還允許您設定一些選項來為模型評分。

在您執行包含決策樹模型塊的串流時，該節點會預設新增一個新的欄位，其名稱將從目標名稱衍生。

表 *12.* 決策樹的模型評分欄位

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。

如果您在建模節點或模型塊上選取選項**計算所分配類別用於記錄評分的機率**，並執行串流，那麼會再新增一個欄位。

表 *13.* 決策樹的模型評分欄位 - 其他

新增欄位的名稱	意義
$IP-*target_name*	預測的信賴度值（從 0.0 到 1.0）。

IBM Data WH 決策樹區塊 - 模型標籤

模型標籤以圖形格式顯示決策樹模型的「預測值重要性」。條欄長度代表預測值的重要性。

註：使用 IBM Netezza Analytics 2.x 版或更早版本時，決策樹模型的內容僅以文字格式顯示。

對於這些版本，會顯示下列資訊：

- 每一行的文字對應於一個節點或葉節點。
- 縮排可反映樹狀結構層次。
- 針對節點，會顯示分割準則。
- 針對葉節點，會顯示指派的類別標籤。

IBM Data WH 決策樹區塊 - 設定標籤

「設定」標籤可讓您為模型評分設定部分選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

計算所分配類別用於記錄評分的機率。（僅限於決策樹和貝式邏輯分類演算法）如果已選取此選項，那麼表示附加的模型欄位包含信賴度（即，機率）欄位和預測欄位。如果您取消選中該勾選框，將只生成預測欄位。

使用確定性輸入資料。如果已選取此選項，那麼將確保任何對同一個視圖多次執行傳遞的 Netezza 演算法都將使用同一組資料進行傳遞。如果清除此方框以表明正在使用非確定性資料，那麼將建立一個暫時表格以存放要處理的資料輸出，例如由分割區節點產生的輸出；建立模型後，將刪除這個表格。

IBM Data WH 決策樹區塊 - 檢視器標籤

檢視器標籤會顯示樹狀結構模型的樹狀結構呈現，與 SPSS Modeler 針對其決策樹模型顯示的方式相同。

註：如果使用 IBM Netezza Analytics 2.x 版或更舊版本建置模型，則**檢視器**標籤為空。

IBM Data WH K-Means 模型塊

K-Means 模型包含叢集模型所擷取的所有資訊，以及訓練資料和預估程序的相關資訊。

當執行包含 K-Means 模型塊的串流時，該節點將新增兩個新欄位，這兩個欄位包含叢集成員資格以及與該記錄所分配到的叢集中心的距離。名為 $KM-K-Means 的新欄位用於叢集成員資格資訊，而名為 $KMD-K-Means 的新欄位用於與叢集中心的距離。

IBM Data WH K-Means 區塊 - 模型標籤

模型標籤包含各種圖形視圖，以顯示叢集欄位的摘要統計量及分佈。您可以從模型匯出資料，也可以將視圖作為圖形匯出。

使用 IBM Netezza Analytics 2.x 版或更早版本時，或者使用馬氏距離作為距離測量來建立模型時，K-Means 模型的內容僅以文字格式顯示。

對於這些版本，會顯示下列資訊：

- **摘要統計量。** 對於最小叢集和最大叢集，彙總統計資料顯示記錄數量。另外，彙總統計資料還顯示這些叢集所擁有的資料集的百分比。該清單還顯示了最大叢集與最小叢集的比值。
- **叢集彙總。**叢集彙總列出了演算法所建立的叢集。對於每個叢集，該表格顯示了該叢集中的記錄數量，以及這些記錄離叢集中心的平均距離。

IBM Data WH K-Means 區塊 - 設定標籤

「設定」標籤可讓您為模型評分設定部分選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

距離測量。這是用於測量資料點之間的距離的方法；距離越大，表示相異性越大。選項是：

- **歐式距離**。（預設）通過將兩個點用一條直線結合起來計算得出的兩點之間的距離。
- **曼哈頓距離**。兩點之間的距離計算為其坐標之間的絕對差總和。
- **堪培拉距離**。類似於曼哈頓距離，但對更加靠近原點的資料點更加敏感。
- **最大值**。兩點之間的距離計算為任何座標尺寸之差的最大值。

Netezza 貝葉斯網絡模型塊

貝葉斯網絡模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含貝葉斯網絡模型塊的串流時，該節點會新增一個新的欄位，其名稱將從目標名稱衍生。

表 *14. Bayes Net* 的模型評分欄位

新增欄位的名稱	意義
$BN-*target_name*	現行記錄的預測值。

您可以透過將「表格」節點連接到模型塊並執行「表格」節點來檢視額外欄位。

Netezza 貝葉斯網絡塊 -「設定」標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

目標。 如果您要對不同於目前目標的目標欄位分數，在此選擇新的目標。

記錄 ID。 如果未指定記錄 ID 欄位，在此選擇要使用的欄位。

預測類型。您要使用的預測演算法的變異：

- **最佳（最相關相鄰元素）**。（預設）使用最相關的相鄰元素節點。
- **鄰接項（鄰接項的加權預測）**。使用所有相鄰元素節點的加權預測。
- **NN 鄰接項（非無效鄰接項）**。與上一選項相同，不同之處在於它將忽略具有空值的節點（即，對於要計算預測結果的實例，該節點對應的屬性存在遺漏值）。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

IBM Data WH Naive Bayes 模型塊

貝式邏輯分類演算法模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含貝式邏輯分類演算法模型塊的串流時，該節點會預設新增一個新的欄位，其名稱將從目標名稱衍生。

表 *15. Naive Bayes* 的模型評分欄位 - 預設值

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。

如果您在建模節點或模型塊上選取選項**計算所分配類別用於記錄評分的機率**，並執行此串流，那麼會再新增兩個欄位。

表 *16. Naive Bayes* 的模型評分欄位 - 其他

新增欄位的名稱	意義
$IP-*target_name*	實例類別的 Bayesian 分子（即，事前類別機率與條件實例屬性值機率的product）。
$ILP-*target_name*	後者的自然對數。

您可以透過將「表格」節點連接至模型塊並執行「表格」節點來檢視額外欄位。

IBM Data WH Naive Bayes 區塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

計算所分配類別用於記錄評分的機率。（僅限於決策樹和貝式邏輯分類演算法）如果已選取此選項，那麼表示附加的模型欄位包含信賴度（即，機率）欄位和預測欄位。如果您取消選中該勾選框，將只生成預測欄位。

針對較小或嚴重失衡的資料集提高機率準確度。在計算機率時，該選項將呼叫 *m* 估計技術，以避免在估計期間出現零機率。這種類型的機率估計可能速度較慢，但可針對較小或嚴重失衡資料集提供更好的結果。

IBM Data WH KNN 模型塊

KNN 模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含 KNN 模型塊的串流時，該節點會新增一個新的欄位，其名稱將從目標名稱衍生。

表 *17. KNN* 的模型評分欄位

新增欄位的名稱	意義
$KNN-*target_name*	現行記錄的預測值。

您可以透過將「表格」節點連接到模型塊並執行「表格」節點來檢視額外欄位。

IBM Data WH KNN 區塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

距離測量。這是用於測量資料點之間的距離的方法；距離越大，表示相異性越大。選項是：

- **歐式距離**。（預設）通過將兩個點用一條直線結合起來計算得出的兩點之間的距離。
- **曼哈頓距離**。兩點之間的距離計算為其坐標之間的絕對差總和。
- **堪培拉距離**。類似於曼哈頓距離，但對更加靠近原點的資料點更加敏感。

- **最大值**。兩點之間的距離計算為任何座標尺寸之差的最大值。

最近鄰數目 (k)。 特定觀察值的最近鄰接項數量。請注意，使用較大相鄰數目未必可得出較精確的模式。

通過選取 *k*，您可以控制項在防止過度配適（這可能很重要，尤其對於「雜訊」資料）和求解（針對類似實例產生不同預測結果）之間的平衡。您通常需要針對每個資料集來調整 *k* 值，其一般值在 1 至幾十之間。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

在計算距離前規範測量結果。如果已選取，該選項將標準化連續輸入欄位的測量結果，然後再計算距離值。

對大型資料集使用核心集以增加效能。如果已選取，該選項將針對大型資料集採用核心集取樣以加快計算過程。

Netezza 分裂式叢集模型塊

分裂式叢集模型塊提供了一種方法來設定選項以為模型評分。

執行包含分裂式叢集模型塊的串流時，該節點將新增兩個新欄位，這兩個新欄位的名稱從目標名稱衍生。

表 *18.* 分割叢集的模型評分欄位

新增欄位的名稱	意義
$DC-*target_name*	現行記錄所分配到的子叢集的 ID。
$DCD-*target_name*	與現行記錄的子叢集中心的距離。

您可以透過將「表格」節點連接至模型片段並執行「表格」節點來檢視額外欄位。

Netezza 分裂式叢集塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

距離測量。這是用於測量資料點之間的距離的方法；距離越大，表示相異性越大。選項是：

- **歐式距離**。（預設）通過將兩個點用一條直線結合起來計算得出的兩點之間的距離。
- **曼哈頓距離**。兩點之間的距離計算為其坐標之間的絕對差總和。
- **堪培拉距離**。類似於曼哈頓距離，但對更加靠近原點的資料點更加敏感。
- **最大值**。兩點之間的距離計算為任何座標尺寸之差的最大值。

套用的階層層次。套用至資料的階層層次。

IBM Data WH PCA 模型塊

PCA 模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含 PCA 模型塊的串流時，該節點會預設新增一個新的欄位，其名稱將從目標名稱衍生。

表 *19. PCA* 的模型評分欄位

新增欄位的名稱	意義
$F-*target_name*	現行記錄的預測值。

如果您在建模節點或模型塊上的**主成分個數 ...**欄位中指定大於 1 的值，並執行此串流，該節點將為每個成分新增一個新的欄位。在此情況下，欄位名稱帶有字尾 *-n*，其中 *n* 是成分的編號。例如，如果模型名為 *pca* 且包含三個成分，那麼新欄位將命名為 *$F-pca-1*、*$F-pca-2* 和 *$F-pca-3*。

您可以透過將「表格」節點連接至模型塊並執行「表格」節點來檢視額外欄位。

IBM Data WH PCA 區塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

要在投射中使用的主成分號碼。您要用來減小資料集的主成分個數。該值不得超過屬性（輸入欄位）數量。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

Netezza 迴歸方法樹狀結構模型塊

迴歸方法樹狀結構模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含迴歸方法樹狀結構模型塊的串流時，該節點會預設新增一個新的欄位，其名稱將從目標名稱衍生。

表 *20.* 迴歸樹狀結構的模型評分欄位

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。

如果您在建模節點或模型塊上選取選項**計算估計變異數**，並執行串流，那麼會再新增一個欄位。

表 *21.* 迴歸樹狀結構的模型評分欄位 - 其他

新增欄位的名稱	意義
$IV-*target_name*	所預測的值的估計變異數。

您可以透過將「表格」節點連接至模型片段並執行「表格」節點來檢視額外欄位。

Netezza 迴歸方法樹狀結構塊 - 模型標籤

模型標籤以圖形格式顯示迴歸樹狀結構模型的「預測值重要性」。條欄長度代表預測值的重要性。

註：使用 IBM Netezza Analytics 2.x 版或更早版本時，迴歸方法樹狀結構模型的內容僅以文字格式顯示。

對於這些版本，會顯示下列資訊：

- 每一行的文字對應於一個節點或葉節點。
- 縮排可反映樹狀結構層次。
- 針對節點，會顯示分割準則。
- 針對葉節點，會顯示指派的類別標籤。

Netezza 迴歸方法樹狀結構塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

計算估計變異數。表示是否應在輸出中包含所分配類別的變異數。

Netezza 迴歸方法樹狀結構塊 -「檢視器」標籤

檢視器標籤會顯示樹狀結構模型的樹狀結構呈現，與 SPSS Modeler 針對其迴歸樹狀結構模型顯示的方式相同。

註：如果使用 IBM Netezza Analytics 2.x 版或更舊版本建置模型，則**檢視器**標籤為空。

IBM Data WH 線性迴歸模型塊

線性迴歸模型塊提供了一種方法來設定選項以為模型評分。

在您執行包含線性迴歸模型塊的串流時，該節點會新增一個新的欄位，其名稱將從目標名稱衍生。

表 22. 線性迴歸的模型評分欄位

新增欄位的名稱	意義
$LR-*target_name*	現行記錄的預測值。

IBM Data WH 線性回歸區塊 - 設定標籤

在「設定」標籤上，您可以設定用於對模型進行評分的選項。

包含輸入欄位。選取之後，此選項會向下傳遞所有原始輸入欄位，將額外的一或多個建模欄位附加至每列資料。如果您清除這個勾選框，則只會傳送記錄 ID 欄位和額外建模欄位，因此串流的執行速度會更快。

Netezza 「時間序列」模型塊

此模型塊使您能夠存取時間序列建模作業的輸出。輸出由下列欄位組成。

表 23. 時間序列模型輸出欄位

欄位	說明
TSID	時間序列的 ID；這是在建模節點的「欄位」標籤上對「時間序列 ID」指定的欄位中的內容。請參閱第 64 頁的『Netezza 時間序列欄位選項』主題，以取得更多資訊。
TIME	目前時間序列內的時間期間。
HISTORY	歷程資料值（曾用於預測）。僅當模型塊的「設定」標籤中的**在輸出中包含歷程記錄值**選項被選取時，該欄位才被包含在內。
$TS-INTERPOLATED	插入值（如果使用）。僅當模型塊的「設定」標籤中的**在輸出中包含插補的值**選項被選取時，該欄位才被包含在內。插補是建模節點的「建置選項」標籤上的一個選項。
$TS-FORECAST	時間序列的預測值。

要檢視模型輸出，請從節點選用區的「輸出」標籤將一個「表格」節點附加到模型塊，並執行這個「表格」節點。

Netezza 時間序列塊 - 設定標籤

在「設定」標籤中，您可以指定選項來自訂模型輸出。

模型名稱。 模型名稱，在建模節點的「模型選項」標籤上定義。

其他選項與建模節點的「建模選項」標籤上的選項相同。

IBM Data WH 廣義線性模型塊

此模型塊使您能夠存取建模作業的輸出。

執行包含廣義線性模型塊的串流時，該節點將新增一個新的欄位，其名稱是從目標名稱衍生。

表 *24.* 廣義線性模型的模型評分欄位

新增欄位的名稱	意義
$GLM-*target_name*	現行記錄的預測值。

「模型」標籤顯示各種與模型有關的統計資料。

輸出由下列欄位組成。

表 *25.* 廣義線性模型中的輸出欄位

輸出欄位	說明
參數	模型使用的參數（即，預測值）。這些是數值型和列名欄，以及截距（迴歸模型中的常數項目）。
Beta	相關係數（即，模型的線性成分）。
標準誤	Beta 的標準差。
測試	用於評估參數有效性的測試統計資料。
p-value	假定參數為顯著參數時的錯誤機率。
殘差彙總	
殘差類型	顯示彙總值的預測殘差類型。
RSS	殘差的值。
df	殘差的自由度。
p-value	誤的機率。高值表示擬合度欠佳的模型；低值表示擬合度良好。

IBM Data WH 廣義線性模型塊 - 設定標籤

在「設定」標籤中，您可以自訂模型的輸出。

此選項與建模節點的評分選項中顯示的選項相同。請參閱第 56 頁的『IBM Data WH 廣義線性模型選項 - 評分選項』主題，以取得更多資訊。

IBM Data WH TwoStep 模型塊

執行包含 TwoStep 模型塊的串流時，節點會新增兩個新欄位，包含叢整合員資格和距離該記錄之已指派叢集中心的距離。名為 $TS-Twostep 的新欄位用於叢集成員資格資訊，而名為 $TSP-Twostep 的新欄位用於與叢集中心的距離。

IBM Data WH TwoStep 區塊 - 模型標籤

模型標籤包含各種圖形視圖，以顯示叢集欄位的摘要統計量及分佈。您可以從模型匯出資料，也可以將視圖作為圖形匯出。

第 6 章 使用 IBM Db2 for z/OS 進行資料庫建模

IBM SPSS Modeler 和 IBM Db2 for z/OS

SPSS Modeler 支援與 Db2 for z/OS 進行整合，這提供了在 Db2 for z/OS 伺服器上執行進階分析的能力。您可以通過 SPSS Modeler 圖表使用者介面以及面向工作流程的開發環境存取這些功能。這樣，就可以直接在 Db2 for z/OS 環境中執行資料採礦演算法，從而利用 IBM Db2 Analytics Accelerator。

SPSS Modeler 支援與 Db2 for z/OS 中的下列演算法整合。

- 決策樹
- K-Means
- Naive Bayes
- 迴歸樹狀結構
- TwoStep

與 IBM Db2 for z/OS 進行整合的需求

下列是使用 Db2® for z/OS® 和 IBM Db2 Analytics Accelerator for z/OS 處理資料庫內建模的必備項目。為了確保滿足這些條件，您可能需要咨詢資料庫管理者。如需詳細需求，包括支援的版本，請參閱軟體產品相容性報告。

- 以本端方式執行或者針對 Windows 或 UNIX 上的 SPSS Modeler Server 安裝來執行的 IBM SPSS Modeler
- Db2 for z/OS 以及 Db2 Analytics Accelerator for z/OS
- IBM SPSS Data Access Pack
- 在執行 SPSS Modeler Server 的伺服器上，下列其中一個系統：
  - IBM Db2 Data Server Driver for ODBC and CLI
  - 帶有配置用於 DB2 for z/OS 的 ODBC 資料來源的任何 Db2 for Linux、UNIX 和 Windows 版本
- Db2 Connect for System z® 授權
- 在 SPSS Modeler 中啟用 SQL 產生和最佳化
- Db2 z/OS 資料庫內資料採礦需要僅加速器表格 (AOT) 或加速表格以及 INZA 支援。IDAA 5.1 中引入了 IDAA INZA。這意味著 Db2 z/OS 資料庫內資料採礦節點不適用於舊版 IDAA。

  如果在 Modeler 中使用啟用 IDAA 的 DSN，在使用 DSN 的「資料庫來源」節點中傳回的表格清單中，僅顯示 AOT 或加速表格。

啟用與 IBM Db2 Analytics Accelerator for z/OS 整合

啟用與 Db2 Analytics Accelerator for z/OS 整合的過程由下列步驟組成：

- 配置 Db2 for z/OS 和 Db2 Analytics Accelerator for z/OS
- 建立 ODBC 來源
- 在 IBM SPSS Modeler 中，啟用 IBM Db2 for z/OS 的整合
- 在 SPSS Modeler 中啟用 SQL 產生和最佳化

- 啟用 IBM SPSS Modeler Server Scoring Adapter for Db2 for z/OS
- 在 IBM SPSS Modeler 中，使用 IBM Db2 Client 來配置 DSN

配置 IBM Db2 for z/OS 和 IBM Analytics Accelerator for z/OS

下列網站上的內容說明了配置 Db2 for z/OS 和 Analytics Accelerator for z/OS 的方法：

Db2 Analytics Accelerator for z/OS。

為 IBM Db2 for z/OS 和 IBM Db2 Analytics Accelerator 建立 ODBC 來源

有關如何在 Db2 for z/OS 與 IBM Db2 Analytics Accelerator 之間啟用連線的資訊，請參閱下列網站：

- 對於 V4：Db2 Analytics Accelerator for z/OS 4.1.0
- 對於 V3：Db2 Analytics Accelerator for z/OS 3.1.0
- 在不修改應用程式下使用 IBM Db2 Analytics Accelerator for ODBC 和 JDBC 應用程式啟用查詢加速
- 在 Db2 Analytics Accelerator for z/OS 中執行查詢時 ODBC 驅動程式發生 SQL 錯誤

在 IBM SPSS Modeler 中，啟用 IBM Db2 for z/OS 的整合

要在 SPSS Modeler 中啟用 Db2 for z/OS 整合，請執行下列步驟：

1. 從 SPSS Modeler `config` 目錄，開啟 `odbc-db2-accelerator-names.cfg` 檔案。

   如果此檔案不存在，那麼必須進行建立。
2. 新增所有資料來源的名稱和所有加速器的名稱。例如：

   ```
   dsn1, acceleratorname1
   dsn2, acceleratorname2
   ```
3. 用於僅加速器表格 (AOT) 的預設 CCSID 為 Unicode；要置換此設置，向加速器名稱中新增編碼字串來修改條目。例如：

   ```
   dsn1, acceleratorname1, EBCDIC
   dsn2, acceleratorname2, UNICODE
   ```
4. 儲存並關閉 `odbc-db2-accelerator-names.cfg` 檔案，然後開啟同一目錄中的 `odbc-db2-custom-properties.cfg` 檔案。
5. SPSS Modeler 使用 SQL 設定 IDAA 登錄。如果需要，可以將 SQL 變更為所需值來置換這些項目。例如：

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```
6. 依預設， SPSS Modeler 使用 SQL 為資料庫快取建立暫時表格。如果需要，可以指定預期資料庫名稱來置換此設置。例如：

   ```
   [OSZ]
   table_create_temp_sql_acc, 'CREATE TABLE <table-name> <(table-columns)> IN DATABASE NAME_OF_DATABASE_FOR_AOT'
   ```
7. 依預設，SPSS Modeler 認為 ODBC 來源節點中編寫的 SQL 查詢是不可重播的，意味著在多次執行時，會認為查詢傳回不同的結果。但是在某些情況下，這可能會阻止 Modeler 為下游節點產生 SQL，可以通過將相關值變更為 `Y` 置換此設置。例如：

   ```
   assume_custom_sql_replayable, Y
   ```
8. 在 SPSS Modeler 主功能表中，按一下**工具** > **選項** > **說明應用程式**。
9. 按一下 **IBM Db2 for z/OS** 標籤。
10. 選取**啟用 IBM Db2 for z/OS 資料採礦整合**，然後按一下**確定**。

註：您無法在 Modeler 中同時檢視 IDAA 和非 IDAA 表格。

啟用 SQL 產生及最佳化

由於使用超大型資料集的可能性，出於效能的原因，您應在 IBM SPSS Modeler 中啟用 SQL 產生和最佳化選項。

要配置 SPSS Modeler，請完成下列步驟：

1. 從 IBM SPSS Modeler 功能表中選擇**工具** > **串流內容** > **選項**。
2. 按一下導覽窗格中的**最佳化**選項。
3. 確認已啟用**產生 SQL** 選項。資料庫建模需要此設定才能夠運作。
4. 選中**最佳化 SQL 產生和最佳化其他執行**（非嚴格必要但強烈推薦使用，以使效能更優）。

在 IBM SPSS Modeler 中，使用 IBM Db2 Client 來配置 DSN

如果需要在 SPSS Modeler 中使用 Db2 Client for Db2 配置資料來源名稱 (DSN)，請完成下列步驟：

1. 如果尚未安裝，請在安裝了 Modeler Server 的作業系統上安裝 Db2 用戶端。
2. 通過使用 **db2 catalog** 指令，對資料庫編目，並將新資料來源新增到 DB2 用戶端中的 `db2cli.ini` 檔案。確保指向定義的資料庫別名。
3. 配置資料存取權；Modeler 文件中提供了詳細步驟。

   有關進一步資訊，請參閱《Modeler Server 管理和效能文檔》(ModelerServerAdminPerformance.pdf) 中的主題**架構和硬體建議** > **資料存取權**。
4. 通過參照步驟 2 中定義的資料庫別名在 `odbc.ini` 中建立新的 ODBC 資料來源。
5. 對於 Linux 或 UNIX 使用者：
   a. 確保使用驅動程式程式庫 `libdb2o.so`（代替 `libdb2.so`），並確保為新資料來源定義 `'DriverUnicodeType=1'`。
   b. 在 IBM SPSS 資料存取包安裝中，確保將 Db2 用戶端的程式庫路徑新增到 `odbc.sh`。
   c. 確保 Modeler Server 使用具有 UTF-16 編碼的 ODBC 驅動程式封套程式庫（這稱為 `'libspssodbc_datadirect_utf16.so'`）。
6. 確保連接至 Db2 的使用者具有執行下列查詢的必需專用權：

   ```
   SELECT ACCELERATORNAME FROM SYSACCEL.SYSACCELERATORS
   ```

使用 IBM Db2 for z/OS 來建置模型

每個受支援的演算法都有對應的建模節點。您可以從節點選用區上的「資料庫建模」標籤中存取 Db2 for z/OS 建模節點。

資料考量

資料來源中的欄位可包含各種資料類型的變數，視建模節點而定。在 SPSS Modeler 中，資料類型稱為測量層次。建模節點的「欄位」標籤使用圖示來指出其輸入與目標欄位允許的測量層次類型。

目標欄位。目標欄位是您嘗試預測值的欄位。在可指定目標的位置，只能選取其中一個來源資料欄位作為目標欄位。

記錄 ID 欄位。指定用來唯一識別每個觀察值的欄位。例如，這可能是一個 ID 欄位，如 *CustomerID*。如果來源資料不包含 ID 欄位，您可以按照「衍生」節點的方法建立此欄位，如下列程序所示。

1. 選取來源節點。
2. 從節點選用區上的「欄位作業」標籤中，按兩下「衍生」節點。
3. 在畫布上按兩下「衍生」節點圖示來開啟「衍生」節點。
4. 例如，在**衍生**欄位中輸入 ID。
5. 在**公式**欄位中，輸入 @INDEX 並按一下**確定**。
6. 將「衍生」節點連接至串流的剩餘部分。

處理空值

如果輸入資料包含空值，那麼使用某些 Db2 for z/OS 節點可能會導致產生錯誤訊息或者長時間執行的串流，因此我們建議移除包含空值的記錄。使用下列方法。

1. 將「選取」節點附加至來源節點。
2. 將「選取」節點的**模式**選項設為**捨棄**。
3. 在**條件**欄位中輸入下列內容：

   ```
   @NULL(field1) [or @NULL(field2)[... or @NULL(fieldN]])
   ```

   請務必包括每一個輸入欄位。
4. 將「選取」節點連接至串流的剩餘部分。

模型輸出

包含 Db2 for z/OS 建模節點的串流有可能每次執行都產生略微不同的結果。這是因為在模型建置之前，將資料讀取至暫時表格時，節點讀取來源資料的順序不一定相同。但是，此效果產生的差異可以忽略不計。

一般註解

- 在 SPSS Collaboration and Deployment Services 中，不能使用包含 Db2 for z/OS 建模節點的串流來建立評分配置。
- Db2 for z/OS 節點所建立的模型無法進行 PMML 匯出或匯入。

IBM Db2 for z/OS 模型 -「欄位」選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

目標。 選擇一個欄位作為預測的目標。若為「一般線性」模型，另請參閱此畫面上的**試用**欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Db2 for z/OS 模型 - 伺服器選項

在「伺服器」標籤上，指定要在其中建立模型的 Db2 for z/OS 系統。

- **使用上游連線**。（預設值）使用上游節點（例如資料庫來源節點）中指定的連線詳細資料。註：僅當所有上游節點都能夠使用 SQL 回送功能時，此選項才有效。在這種情況下，由於 SQL 完全實現了所有的上游節點，因此無需將資料移出資料庫。
- **將資料移至資料庫**。將資料移至您在這裡指定的資料庫。這樣，即使資料位於另一個 IBM 資料庫或者另一供應商的資料庫中，甚至位於純文字檔案中，也仍然可以進行建模。此外，如果因節點未執行 SQL 推回而擷取了資料，則資料會移回至在這裡指定的資料庫。按一下**編輯**按鈕以瀏覽並選取連線。

註：實際上，ODBC 資料來源名稱內含在每個 SPSS Modeler 串流中。如果在某個主機上建立的串流在不同主機上執行，則資料來源的名稱在每一個主機上必須相同。或者，可以在每一個來源或建模節點中的「伺服器」標籤上選取不同的資料來源。

IBM Db2 for z/OS 模型 -「模型」選項

在「模型選項」標籤上，您可以選擇是否指定模型的名稱，還是自動產生名稱。

模型名稱。您可以根據目標或 ID 欄位（或者模型類型，如果未指定此類欄位）自動產生模型名稱，或者指定自訂名稱。

名稱已使用時取代現有項。如果您選取此勾選框，則會改寫相同名稱的任何現有模型。

IBM Db2 for z/OS 模型 - K-Means

K-Means 節點實作提供叢集分析方法的 *k* 平均數演算法。您可以使用此節點將資料集叢集到不同的群組。

演算法是基於距離的叢集演算法，根據距離度量（函數）來測量資料點之間的親緣性。資料點根據使用的距離度量指派給最近的叢集。

演算法透過對相同的基本處理程序執行數次疊代進行運算，程序中的每一個訓練實例都指派給最近的叢集（就指定的距離函數而言，則套用至距離及叢集中心）。然後會重新計算所有叢集中心作為指派給特定叢集的實例的平均屬性值向量。

IBM Db2 for z/OS 模型 - K-Means 欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

記錄 ID。 要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Db2 for z/OS 模型 - K-Means 建置選項

透過設定建置選項，您可以自訂用於您專屬用途的模型建置。

如果要使用預設選項來建置模型，請按一下**執行**。

距離測量 此參數定義用於測量資料點之間距離的方法。距離越大指出相異性越大。請選取下列其中一個選項：

- **歐基里得。** 歐基里得測量是兩個資料點之間的直線距離。
- **正規化歐基里得。** 常態化歐基里得測量類似於歐基里得測量，但前者以標準偏差平方進行常態化。與歐基里得測量不同，常態化歐基里得測量也是比例不變。

叢集數目。 此參數定義要建立的叢集數目。

最大疊代次數。 演算法對相同的處理程序執行數次疊代。此參數定義模型訓練停止之前的疊代次數。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

複製結果。 如果要設定隨機種子以抄寫分析，請選取這個勾選框。您可以指定一個整數，也可以透過按一下**產生**來建立虛擬隨機整數。

IBM Db2 for z/OS 模型 - Naive Bayes

Naive Bayes 是用來解決分類問題的著名演算法。此模型將所有建議預測變數視為彼此獨立，因此被稱為「樸素」。Naive Bayes 是一個快速可調式演算法，能夠計算屬性及目標屬性組合的條件式機率。從訓練資料中，會建立獨立機率。鑒于每個輸入變數中出現的每個值種類，此機率會提供每個目標類別的可能性。

IBM Db2 for z/OS 模型 - 決策樹

決策樹是代表分類模型的階層式結構。使用決策樹模型，您可以開發分類系統來預測或分類一組訓練資料中的未來觀察。分類採取樹狀結構的形式，其中分支代表分類中的分割點。分割會將資料遞迴地分為子群組，直到達到停止點為止。停止點處的樹狀結構節點稱為葉節點。每片樹葉節點分配一個標籤（稱為 類別標籤）給其子群組或類別會員。

IBM Db2 for z/OS 模型 - 決策樹欄位選項

在「欄位」標籤上，您可以選擇是否要使用已在上游節點中定義的欄位角色設定，還是手動進行欄位指派。

使用預先定義的角色。 此選項使用上游「類型」節點（或上游來源節點的「類型」標籤）中的角色設定（目標、預測值等）。

使用自訂欄位指派。 如果您要在此畫面上手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭按鈕，將此清單中的項目手動指派給畫面右側上的各個角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

按一下**全部**按鈕以選取清單中的所有欄位，或按一下個別測量層次按鈕以選取該測量層次中的所有欄位。

目標。 選擇一個欄位作為預測的目標。

記錄 ID。 要用作唯一記錄 ID 的欄位。此欄位的值針對每個記錄必須是唯一的（例如，客戶身分證號碼）。

實例加權。在這裡指定欄位會讓您使用實例加權（每個輸入資料列一個加權），而不是預設值類別加權（每個目標欄位的種類一個加權）。您在這裡指定的欄位必須包含每一列輸入資料的數字加權。

預測值（輸入）。 選取一個或多個輸入欄位。這與在「類型」節點中將欄位角色設為輸入類似。

IBM Db2 for z/OS 模型 - 決策樹建置選項

下列建置選項可用於樹狀結構成長：

成長測量。 這些選項可控制測量樹狀結構成長的方式。

- **雜質測量。** 此測量會評估分割樹狀結構的最佳位置。這是測量資料子群組或區段中的變異性。較低的雜質測量指出群組中大部分成員的準則或目標欄位值類似。

  支援的測量為**熵** 和 **Gini**。這些測量以分支的種類成員資格的機率為基礎。
- **樹狀結構深度上限。** 樹狀結構在根節點下可以成長到的層次數目上限，亦即可以遞迴地分割樣本的次數。此內容的預設值是 10，您可以為此內容設定的最大值為 62。

  註：如果模型片段中的檢視器顯示模型的文字呈現，則最多可以顯示 12 個層次的樹狀結構。

分割準則。 這些選項控制何時停止分割樹狀結構。

- **分割下限改善。** 必須減少的雜質數量下限，之後才能在樹狀結構中建立新分割。樹狀結構建置的目標是建立具有相似輸出值的子群組以最小化每個節點內的雜質。如果分支的最佳分割會將雜質減少到小於分割準則所指定的數量，則不會分割分支。
- **分割的實例數下限。** 可以分割的記錄數目下限。當剩餘的未分割記錄少於此數目時，不會執行進一步分割。您可以使用此欄位來阻止在樹狀結構中建立小型子群組。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

IBM Db2 for z/OS 模型 - 決策樹節點 - 類別加權

在這裡您可以為個別類別指派加權。預設是指派值 1 給所有類別，使它們的加權相等。透過為不同類別標籤指定不同分值加權，您指示演算法相應地對特定類別的訓練集進行加權。

若要變更加權，請在**加權**直欄中按兩下加權，然後進行所需的變更。

值。 類別標籤的集，衍生自目標欄位的可能值。

加權。 要指派給特定類別的加權。將較高的加權指派給某個類別會讓模型對該類別相較於其他類別而言更為敏感。

您可以結合使用類別加權與實例加權。

IBM Db2 for z/OS 模型 - 決策樹節點 - 樹狀結構刪改

您可以使用刪改選項來為決策樹指定刪改準則。刪改的目的是要減少因為移除不會改進新資料預期精確度的過度成長子群組而導致過適的風險。

刪改測量。預設刪改測量**準確性**可確保從樹狀結構移除葉節點之後，預估的模型準確性仍處於可接受的限制內。如果要在套用刪改時將類別加權納入考量，請使用替代測量**加權的準確性**。

要刪改的資料。 您可以使用部分或全部訓練資料來預估新資料的預期準確性。或者，您可以將指定表格中的個別刪改資料集用於此目的。

- **使用所有訓練資料**。 此選項（預設值）會使用所有訓練資料來預估模型準確性。
- **用於刪改的訓練資料的使用百分比**。 使用此選項以將資料分割為兩個集合，一個用於訓練，另一個用於刪改，並將這裡指定的百分比用於刪改資料。
- 如果要指定隨機種子以確保每次執行串流時都以相同的方式來分割資料，請選取**抄寫結果**。您可以在**用於刪改的種子**欄位中指定一個整數，或是按一下**產生**以建立虛擬亂數整數。
- **使用現有表格中的資料**。 指定用於估計模型精確性的獨立刪改資料集的表格名稱。這樣做比使用訓練資料更為可靠。

IBM Db2 for z/OS 模型 - 迴歸樹狀結構

迴歸樹狀結構是基於樹狀結構的演算法，反覆地分割觀察值的樣本以基於數值目標欄位衍生相同種類的子集。與決策樹一樣，迴歸樹狀結構將資料拆解為子集，其中樹狀結構的葉節點對應於足夠小或足夠統一的子集。選取分割來減少目標屬性值的離散，因此能夠透過葉節點上的平均值合理地進行預測。

IBM Db2 for z/OS 模型 - 迴歸方法樹狀結構建置選項 - 樹狀結構成長

您可以針對樹狀結構成長及樹狀結構刪改設定建置選項。

下列建置選項可用於樹狀結構成長：

樹狀結構深度上限。 樹狀結構在根節點下可以成長到的層次數目上限，亦即可以遞迴地分割樣本的次數。預設值是 62，這是用於建模的樹狀結構深度上限。

註：如果模型片段中的檢視器顯示模型的文字呈現，則最多可以顯示 12 個層次的樹狀結構。

分割準則。 這些選項控制何時停止分割樹狀結構。

- **分割評估測量**。 此類別評估測量會評估分割樹狀結構的最佳位置。

  註：目前變異數是唯一可能的選項。
- **分割下限改善**。 必須減少的雜質數量下限，之後才能在樹狀結構中建立新分割。樹狀結構建置的目標是建立具有相似輸出值的子群組以最小化每個節點內的雜質。如果分支的最佳分割會將雜質減少到小於分割準則所指定的數量，則不會分割分支。
- **分割的實例數下限**。 可以分割的記錄數目下限。當剩餘的未分割記錄少於此數目時，不會執行進一步分割。您可以使用此欄位來阻止在樹狀結構中建立小型子群組。

統計量。 此參數定義模型中包含的統計量數量。請選取下列其中一個選項：

- **全部**。 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄**。 包含與直欄相關的統計量

- **無。** 僅包含對模型進行評分所需的統計量。

IBM Db2 for z/OS 模型 - 迴歸方法樹狀結構建置選項 - 樹狀結構刪改

您可以使用刪改選項來為迴歸樹狀結構指定刪改準則。刪改的目的是要減少因為移除不會改進新資料預期精確度的過度成長子群組而導致過適的風險。

刪改測量。 刪改測量可確保從樹狀結構移除葉節點之後，預估的模型準確性仍處於可接受的限制內。您可以選取下列其中一個測量。

- **mse。** 均方誤差 - （預設值）測量適合行離資料點有多近。
- **r2。** R 平方 - 測量迴歸模型說明的應變數中變異的比例。
- **Pearson。** Pearson 的相關係數 - 測量正常分配的線性應變數之間的關係強度。
- **Spearman。** Spearman 的相關係數 - 偵測根據 Pearson 的相關性顯示為弱但實際可能很強的非線性關係。

要刪改的資料。 您可以使用部分或全部訓練資料來預估新資料的預期準確性。或者，您可以將指定表格中的個別刪改資料集用於此目的。

- **使用所有訓練資料。** 此選項（預設值）會使用所有訓練資料來預估模型準確性。
- **用於刪改的訓練資料的使用百分比。** 使用此選項以將資料分割為兩個集合，一個用於訓練，另一個用於刪改，並將這裡指定的百分比用於刪改資料。

  如果要指定隨機種子以確保每次執行串流時都以相同的方式來分割資料，請選取**抄寫結果**。您可以在**用於刪改的種子**欄位中指定一個整數，或是按一下**產生**以建立虛擬亂數整數。
- **使用現有表格中的資料。** 指定個別刪改資料集的表格名稱以估計模型精確度。這樣做比使用訓練資料更為可靠。

IBM Db2 for z/OS 模型 - 二階

TwoStep 節點實作 TwoStep 演算法，可提供叢集資料除以大資料集的方法。

您可以在考量可用資源（例如，記憶體和時間限制）時使用此節點來叢集資料。

TwoStep 演算法是透過下列方式叢集資料的資料庫採礦演算法：

1. 叢集特性 (CF) 樹狀結構已建立。這一高度平衡的樹狀結構儲存階層式叢集的叢集特性，其中類似的輸入記錄會變成相同樹狀結構節點的一部分。
2. CF 樹狀結構的葉節點是階層式地叢集到記憶體內，以產生最終的叢集結果。最佳叢集數會自動決定。如果您指定叢集數上限，則會決定指定限制內的最佳叢集數。
3. 叢集結果在另一個步驟中精簡，其中資料會套用與 K-Means 演算法類似的演算法。

IBM Db2 for z/OS 模型 - TwoStep 欄位選項

透過設定欄位選項，您可以指定使用在上游節點中定義的欄位角色設定。您還可以手動指派欄位。

選取項目。 選擇此選項可使用上游「類型」節點或上游來源節點的「類型」標籤中的角色設定。例如，角色設定是目標及預測值。

使用自訂欄位指派。 如果您要手動指派目標、預測值及其他角色，請選擇此選項。

欄位。 使用箭頭可將此清單中的項目手動指派給右側的角色欄位。這些圖示指出每一個角色欄位的有效測量層次。

記錄 ID。要用作唯一記錄 ID 的欄位。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

IBM Db2 for z/OS 模型 - 二階建置選項

透過設定建置選項，您可以自訂用於您專屬用途的模型建置。

如果要使用預設選項來建置模型，請按一下**執行**。

距離測量 此參數定義用於測量資料點之間距離的方法。距離越大指出相異性越大。選項為：

- **對數概似。** 概似量數會對變數進行機率分配。連續變數假設為常態分配，而類別變數則假設為多項式分配。所有變數皆假設為不相依。

叢集數目。 此參數定義要建立的叢集數目。選項是：

- **自動計算叢集數目。** 叢集數會自動計算。您可以在**上限**欄位中指定叢集數上限。
- **指定叢集數目。** 指定應建立多少叢集。

統計量。 此參數定義模型中包含的統計量數量。選項是：

- **全部。** 包含所有與直欄相關的統計量以及所有與值相關的統計量。

  註：此參數包含統計量數量上限，因此可能會影響系統效能。如果不想以圖形格式檢視模型，請指定**無**。
- **直欄。** 包含與直欄相關的統計量
- **無。** 僅包含對模型進行評分所需的統計量。

複製結果。 如果要設定隨機種子以抄寫分析，請選取這個勾選框。您可以指定一個整數，也可以透過按一下**產生**來建立虛擬隨機整數。

IBM Db2 for z/OS 模型 - TwoStep 塊 -「模型」標籤

模型標籤包含各種圖形視圖，以顯示叢集欄位的摘要統計量及分佈。您可以從模型匯出資料，也可以將視圖作為圖形匯出。

管理 IBM Db2 for z/OS 模型

Db2 for z/OS 模型可以使用與其他 IBM SPSS Modeler 模型相同的方式新增到畫布和模型選用區中，並以幾乎相同的方式來使用。

要直接在 Db2 for z/OS 中對資料進行評分，請完成下列步驟：

1. 在資料所在的 Db2 for z/OS 資料庫中安裝 SPSS Scoring Adapter。
2. 確保串流連接至資料所在的 Db2 for z/OS 資料庫。

對 IBM Db2 for z/OS 模型進行評分

模型在畫布上由金色的模型區塊圖示代表。區塊的主要目的是對資料評分以產生預測，或進一步分析模型內容。評分以一或多個額外資料欄位的形式來新增，可透過將「表格」節點附加至區塊並執行串流分支設為可見，本節稍後會加以說明。部分區塊對話框（例如「決策樹」或「回歸樹」的區塊對話框）額外會有一個「模型」標籤，用來提供模型的視覺化表示法。

額外欄位由目標欄位名稱中新增的字首 `$<id>-` 加以區分，其中 `<id>` 取決於模型，用來識別所新增資訊的類型。在每個模型區塊的主題中說明了不同的 ID。

若要檢視評分，請完成下列步驟：

1. 將「表格」節點連接至模型片段。
2. 開啟「表格」節點。
3. 按一下「**執行**」。
4. 捲動至表格輸出視窗的右邊，以檢視額外的欄位及其評分。

註：評分過程不是在加速器中執行，而是在 Db2 中執行，因此要求用於評分的輸入表格必須實體上位於 Db2 中。因此，作為評分輸入，只能使用基於 DB2 的表格或加速表格。如果串流使用僅加速器表格，將發生下列錯誤："THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR."

IBM Db2 for z/OS 決策樹模型塊

決策樹模型塊顯示建模作業的輸出，還允許您設定一些選項來為模型評分。

執行包含決策樹模型塊的串流時，該節點將新增兩個新欄位，這兩個新欄位的名稱從目標衍生。

表 *26.* 決策樹的模型評分欄位.

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。
$IP-*target_name*	預測的信賴度值（從 0.0 到 1.0）。

註：由於 Db2 for z/OS 的限制，可能會截斷直欄名稱。

IBM Db2 for z/OS 決策樹塊 -「模型」標籤

模型標籤以圖形格式顯示決策樹模型的「預測值重要性」。條欄長度代表預測值的重要性。

IBM Db2 for z/OS 決策樹塊 -「檢視器」標籤

檢視器標籤會顯示樹狀結構模型的樹狀結構呈現，與 SPSS Modeler 針對其決策樹模型顯示的方式相同。

IBM Db2 for z/OS K-Means 模型塊

K-Means 模型包含叢集模型所擷取的所有資訊，以及訓練資料和預估程序的相關資訊。

執行包含 K-Means 模型塊的串流時，該節點將新增兩個新欄位，這兩個欄位包含叢集成員資格資訊以及與該記錄所分配到的叢集中心的距離。新的欄位名稱得自模型名稱，即為叢集成員資格加上 $KM- 字首，為與叢集中心的距離加上 $KMD- 字首。例如，如果模型名稱為 Kmeans，那麼新欄位的名稱應是 $KM-Kmeans 和 $KMD-Kmeans。

註：由於 Db2 for z/OS 的限制，可能會截斷直欄名稱。

IBM Db2 for z/OS K-Means 塊 - 「模型」標籤

模型標籤包含各種圖形視圖，以顯示叢集欄位的摘要統計量及分佈。您可以從模型匯出資料，也可以將視圖作為圖形匯出。

IBM Db2 for z/OS Naive Bayes 模型塊

執行包含貝式邏輯分類演算法模型塊的串流時，該節點將新增兩個新欄位，這兩個新欄位的名稱從目標名稱衍生。

表 27. *Naive Bayes* 的模型評分欄位.

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。
$IP-*target_name*	預測的信賴度值（從 0.0 到 1.0）。

註：由於 Db2 for z/OS 的限制，可能會截斷直欄名稱。

您可以透過將「表格」節點連接至模型片段並執行「表格」節點來檢視額外欄位。

IBM Db2 for z/OS 迴歸方法樹狀結構模型塊

執行包含迴歸方法樹狀結構模型塊的串流時，該節點將新增兩個新欄位，這兩個新欄位的名稱從目標名稱衍生。

表 28. 迴歸樹狀結構的模型評分欄位.

新增欄位的名稱	意義
$I-*target_name*	現行記錄的預測值。
$IS-*target_name*	所預測的值的估計標準差。

註：由於 Db2 for z/OS 的限制，可能會截斷直欄名稱。

您可以透過將「表格」節點連接至模型片段並執行「表格」節點來檢視額外欄位。

IBM Db2 for z/OS 迴歸方法樹狀結構塊 -「模型」標籤

模型標籤以圖形格式顯示迴歸樹狀結構模型的「預測值重要性」。條欄長度代表預測值的重要性。

IBM Db2 for z/OS 迴歸方法樹狀結構塊 -「檢視器」標籤

檢視器標籤會顯示樹狀結構模型的樹狀結構呈現，與 SPSS Modeler 針對其迴歸樹狀結構模型顯示的方式相同。

IBM Db2 for z/OS TwoStep 模型塊

執行包含 TwoStep 模型塊的串流時，節點會新增兩個新欄位，包含叢整合員資格和距離該記錄之已指派叢集中心的距離。新的欄位名稱衍生自模型名稱，即，添加 $KM- 字首（對於用於叢集成員資格資訊的欄位）和 $TSD-（對於用於與叢集中心的距離的欄位）。例如，如果模型名稱為 MDL，那麼新欄位的名稱將是 $TS-MDL 和 $TSD-MDL。

注意事項

本資訊係針對 IBM 在美國所提供之產品與服務所開發。IBM 可能會以其他語言提供本資料。不過，您可以要求擁有一份該語言的產品或產品版本副本以取用它。

在其他國家，IBM 不見得有提供本文件所提及之各項產品、服務或功能。請洽詢當地的 IBM 業務代表，以取得當地目前提供的產品和服務之相關資訊。本文件在提及 IBM 的產品、程式或服務時，不表示或暗示只能使用 IBM 的產品、程式或服務。只要未侵犯 IBM 之智慧財產權，任何功能相當之產品、程式或服務皆可取代 IBM 之產品、程式或服務。不過，任何非 IBM 之產品、程式或服務，使用者必須自行負責作業之評估和驗證責任。

本文件所說明之主題內容，IBM 可能擁有其專利或專利申請案。提供本文件不代表提供這些專利的授權。您可以書面提出授權查詢，來函請寄到：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

如果是有關雙位元組 (DBCS) 資訊的授權查詢，請洽詢所在國的 IBM 智慧財產部門，或書面提出授權查詢，來函請寄到：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM 僅以「現狀」提供本書，而不提供任何明示或默示之保證（包括但不限於可售性或符合特定效用的保證）。某些特定之法定管轄區在特定交易上，不允許排除明示或暗示的保證，因此，該項聲明不一定適合您。

本資訊中可能會有技術上或排版印刷上的訛誤。因此，IBM 會定期修訂；並將修訂後的內容納入新版中。IBM 隨時會改進及/或變更本出版品所提及的產品及/或程式，不另行通知。

本資訊中任何對非 IBM 網站的敘述僅供參考，IBM 對該網站並不提供任何保證。這些網站所提供的資料不是 IBM 本產品的資料內容，如果要使用這些網站的資料，您必須自行承擔風險。

IBM 得以各種 IBM 認為適當的方式使用或散布貴客戶提供的任何資訊，而無需對貴客戶負責。

如果本程式之獲授權人為了 (i) 在個別建立的程式和其他程式（包括本程式）之間交換資訊，以及 (ii) 相互使用所交換的資訊，因而需要相關的資訊，請洽詢：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

上述資料之取得有其特殊要件，在某些情況下必須付費方得使用。

IBM 基於 IBM 客戶合約、IBM 國際程式授權合約或雙方之任何同等合約的條款，提供本文件所提及的授權程式與其所有適用的授權資料。

所引用的效能資料和客戶範例僅為說明用途呈現。實際效能結果會隨著特定的配置和作業條件而不同。

本書所提及之非 IBM 產品資訊，係一由產品的供應商，或其出版的聲明或其他公開管道取得。IBM 並未測試過這些產品，也無法確認這些非 IBM 產品的執行效能、相容性或任何對產品的其他主張是否完全無誤。如果您對非 IBM 產品的性能有任何的疑問，請逕向該產品的供應商查詢。

關於 IBM 未來方針或目的之聲明，隨時可能更改或撤銷，不必另行通知，且僅代表目標與主旨。

此資訊包含日常企業運作所使用的資料和報告的範例。為了盡可能地加以完整說明，範例中含有個人、公司、品牌及產品的名稱。所有這些名稱全為虛構，任何與實際個人或商場企業類似之處，純屬巧合。

商標

IBM、IBM 標誌及 ibm.com 是 International Business Machines Corp. 在世界許多管轄區註冊的商標或註冊商標。其他產品及服務名稱可能是 IBM 或其他公司的商標。IBM 商標的最新清單可在 Web 的 "Copyright and trademark information" 中找到，網址為 www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 標誌、PostScript 及 PostScript 標誌是 Adobe Systems Incorporated 在美國及（或）其他國家或地區的註冊商標或商標。

Intel、Intel 標誌、Intel Inside、Intel Inside 標誌、Intel Centrino、Intel Centrino 標誌、Celeron、Intel Xeon、Intel SpeedStep、Itanium 及 Pentium 是 Intel Corporation 或其子公司在美國及其他國家或地區的商標或註冊商標。

Linux 是 Linus Torvalds 在美國及（或）其他國家或地區的註冊商標。

Microsoft、Windows、Windows NT 及 Windows 標誌是 Microsoft Corporation 在美國及/或其他國家或地區的商標。

UNIX 是 The Open Group 在美國及其他國家或地區的註冊商標。

Java 和所有以 Java 為基礎的商標及標誌是 Oracle 及（或）其子公司的商標或註冊商標。

產品說明文件條款

這些出版品的使用許可權，係遵循下列條款而授予。

適用性

這些條款附加於 IBM 網站所適用的任何條款。

個人用途

貴客戶可以為了非商務性的私人用途而複製這些出版品，但必須保留所有專利注意事項。如果沒有 IBM 明文同意，貴客戶不能散佈、顯示或衍生這些出版品或其中的任何部分。

商業用途

貴客戶可以在企業內複製、散佈和展示這些出版品，但必須保留所有專利注意事項。如果沒有 IBM 的明文同意， 貴客戶不得在企業外衍生這些出版品，或複製、散布或顯示這些出版品或其中的任何部分。

權限

除了本項許可權所明確授予者之外，並未明示或暗示授予出版品或任何資訊、資料、軟體或其中的其他智慧財產的任何其他許可權、授權或權利。

IBM 保留在判定出版品的使用將損害其利益或判定未適當遵守上述指示時，撤銷此處所授予之許可權的權利。

除非完全符合所有適當的法律和規章，其中包括所有美國輸出法律和規章，否則，貴客戶不能下載、輸出或再輸出本項資訊。

IBM 對於該等出版品之內容不為任何保證。這些出版品是依「現狀」提供，不含任何明示或默示之保證（包括但不限於可售性、未涉侵權及符合特定效用的保證）。

索引

索引順序以中文字，英文字，及特殊符號之次序排列。

〔四劃〕

〔五劃〕

〔六劃〕

〔七劃〕

〔八劃〕

〔九劃〕

〔十劃〕

〔十一劃〕

〔十二劃〕

〔十三劃〕

〔十四劃〕

〔十五劃〕

〔十六劃〕

〔十七劃〕

〔十八劃〕

〔十九劃〕

〔二十一劃〕

〔二十三劃〕

A

D

E

G

I

K

M

N

O

P

S

T

Z

Printed in Taiwan