IBM SPSS Modeler 18.2.1 建模節點

IBM

附註

在使用本資訊及其支援的產品之前，請先閱讀第 331 頁的『注意事項』中的資訊。

產品資訊

此版本適用於 IBM SPSS Modeler 18.2.0 版及所有後續版本和修正，直到新版中另有指示。

目錄

前言

IBM® SPSS® Modeler 是 IBM Corp. 企業級資料採礦工作台。SPSS Modeler 協助組織透過深入瞭解資料來改進客戶與居民關係。組織使用從 SPSS Modeler 得出的見解保留盈利性客戶，識別交叉銷售機會，吸引新客戶，偵測缺陷，降低風險，並改進政府服務交付。

SPSS Modeler 的視覺化介面可讓使用者發揮其特定的商業專門知識，這樣可形成更強大的預測模型並縮短得出解決方案的時間。SPSS Modeler 提供了很多建模技術，如預測、分類、分區段和關聯偵測演算法。建立模型後，IBM SPSS Modeler Solution Publisher 可讓他們向決策制訂者或資料庫進行企業層面的交付。

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。包括商業智慧、預測分析、財務績效和策略管理，以及分析應用程式的整合型產品組合，為目前績效提供了清晰、即時且具行動性的前瞻眼界，以及預測未來成果的能力。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。

在這項產品組合中，IBM SPSS Predictive Analytics 軟體有助於組織預測未來事件，並且針對前瞻概念提前行動，創造更棒的營運成果。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。藉由將 IBM SPSS 軟體併入每天作業，這些組織成為預測型企業 – 足以駕馭決策並使決策自動化處理，以符合營運目標，並且達到可測知的競爭優勢。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

技術支援人員可提供客戶維護的服務。客戶可以聯絡技術支援人員，尋求 IBM Corp. 產品使用協助，或尋求其中一個受支援硬體環境的安裝協助。若要聯絡技術支援人員，請參閱 IBM Corp. 網站，網址：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。

© Copyright IBM Corp. 1994, 2019

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 採用先進的語言技術和自然語言處理 (NLP)，可快速處理各種各樣的非結構化文字資料，擷取並組織關鍵概念，以及將這些概念分類。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 用戶端相同的所有預測分析功能。使用訂閱版，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。這樣會開啟可一律在產品外部存取的線上 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 下載最新的 PDF 文件，網址為：http://www.ibm.com/support/docview.wss?uid=ibm10874788。

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 建模簡介

模型是一組規則、公式或方程式，可用於根據輸入欄位或變數集預測結果。例如，金融機構可以使用模型，根據過去已知的申請人來預測貸款申請人有可能造成較低還是較高風險。

預測結果的能力是預測性分析的核心目標，而瞭解建模過程是使用 IBM SPSS Modeler 的關鍵所在。

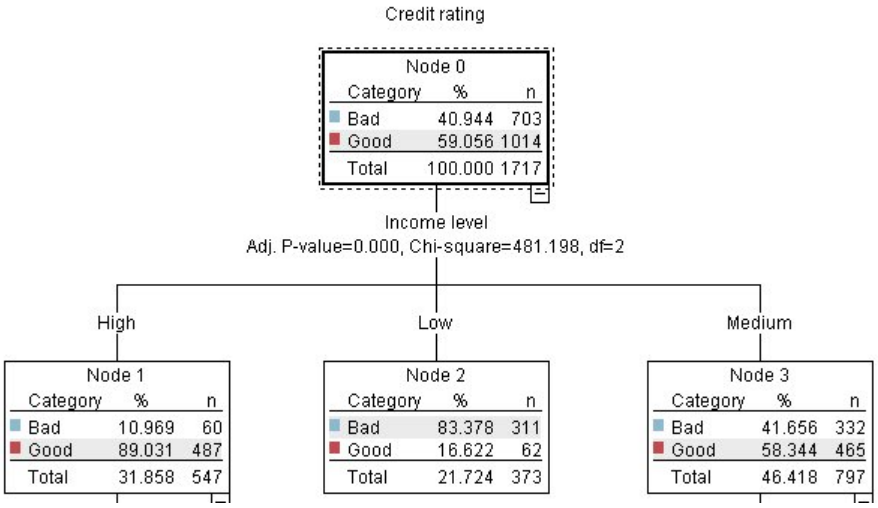


圖 *1.* 簡式決策樹狀結構模型

此範例使用**決策樹狀結構**模型，其使用一系列決策規則來將記錄分類，例如：

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

雖然此範例使用 CHAID（卡方自動互動偵測）模型，但其只是作為一般簡介，而大部分概念廣泛適用於 IBM SPSS Modeler 中的其他建模。

若要瞭解任何模型，首先需要瞭解放入其中的資料。此範例中的資料包含銀行客戶的相關資訊。使用了下列欄位：

欄位名稱(F)	**說明**
信用評級	信用評級：0=差，1=佳，9=遺漏值
年齡	年齡（年為單位）
收入	收入層級：1=低，2=中等，3=高
信用卡	持有的信用卡數目：1=少於五張，2=五張以上
教育	教育程度：1=高中，2=大學
汽車貸款	汽車貸款數目：1=無或一輛，2=兩輛以上

銀行保留客戶向銀行貸款之歷程資訊的資料庫，包括他們是否償還貸款（信用評級 = 佳）或拖欠（信用評級 = 差）。使用此現有資料，銀行想要建置一個模型，可讓他們預測未來貸款申請人拖欠貸款的可能性。

使用決策樹模型，您可以分析兩組客戶的特性，並預測拖欠貸款的可能性。

© Copyright IBM Corp. 1994, 2019

此範例使用名為 *modelingintro.str* 的串流，位於 *Demos* 資料夾下的 *streams* 子資料夾中。資料檔案為 *tree_credit.sav*。請參閱第 4 頁的『Demos 資料夾』主題，以取得更多資訊。

讓我們來看看串流。

1. 從主功能表中選擇下列項目：

   檔案 > **開啟串流**

2. 按一下「開啟」對話框之工具列上的金色片段圖示，並選擇 Demos 資料夾。
3. 按兩下 *streams* 資料夾。
4. 按兩下名為 *modelingintro.str* 的檔案。

建置串流

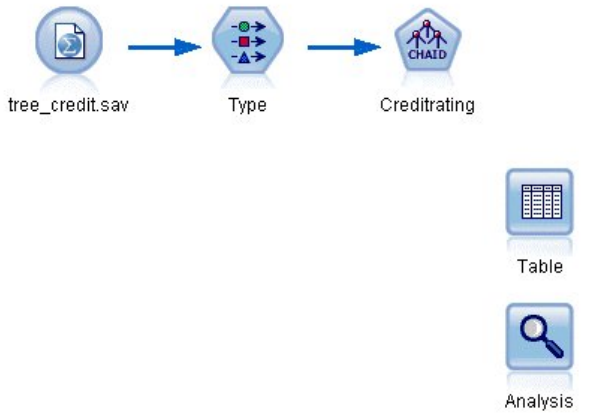


圖 2. 建模串流

若要建置用於建立模型的串流，我們需要至少三個元素：

- 來源節點，用於讀取外部來源的資料，在此情況下為 IBM SPSS Statistics 資料檔案。
- 來源或「類型」節點，用於指定欄位內容（如測量層級，即欄位所包含資料的類型）以及每個欄位在建模中的角色是目標還是輸入。
- 建模節點，用於在執行串流時產生模型片段。

在此範例中，我們使用的是 CHAID 建模節點。CHAID（或卡方自動互動偵測）是一種分類方法，可透過使用稱為卡方統計學的特定統計學類型來找出決策樹中進行分割的最佳位置。

如果來源節點中指定了測量層級，則可以刪除個別類型節點。從功能而言，結果都相同。

此串流還具有「表格」及「分析」節點，在建立模型片段並將其新增至串流後，將使用這些節點來檢視評分結果。

「統計資料檔案」來源節點可讀取 *tree_credit.sav* 資料檔案中以 IBM SPSS Statistics 格式表示的資料，該資料檔案安裝於 *Demos* 資料夾中。（名為 *$CLEO_DEMOS* 的特殊變數用於參照現行 IBM SPSS Modeler 安裝下的此資料夾。如此一來，無論現行安裝資料夾或版本為何，該路徑都將有效。

圖 *3.* 使用統計資料檔案來源節點讀取資料

「類型」節點指定每一個欄位的**測量層級**。測量層級是用於指出欄位中資料類型的種類。我們的來源資料檔案使用三種不同的測量層級。

連續欄位（如年齡欄位）包含連續數值，而**名義**欄位（如信用評級欄位）有兩個以上的不同值，例如差、佳或無信用記錄。**序數**欄位（如收入層級欄位）使用多個具有固有順序的不同值來說明資料 - 在此情況下為低、中等及高。

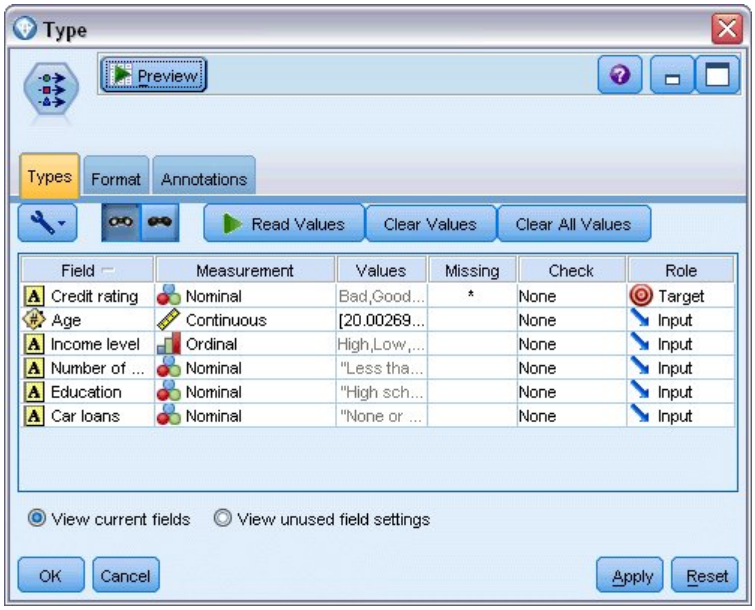


圖 *4.* 使用類型節點設定目標和輸入欄位

針對每一個欄位，「類型」節點也指定**角色**，以指出每一個欄位在建模中所扮演的部分。對於欄位信用評級，角色設為目標，該欄位指出給定客戶是否拖欠貸款。這是**目標**欄位，即我們想要預測值的欄位。

對於其他欄位，角色設為輸入。輸入欄位有時稱為**預測值**，建模演算法會使用這些欄位的值來預測目標欄位的值。

CHAID 建模節點會產生模型。

在建模節點的「欄位」標籤上，已選取選項**使用預先定義的角色**，表示目標和輸入將使用「類型」節點中指定的項目。此時我們可變更欄位角色，但在本範例中我們就按原樣使用。

1. 按一下「建置選項」標籤。

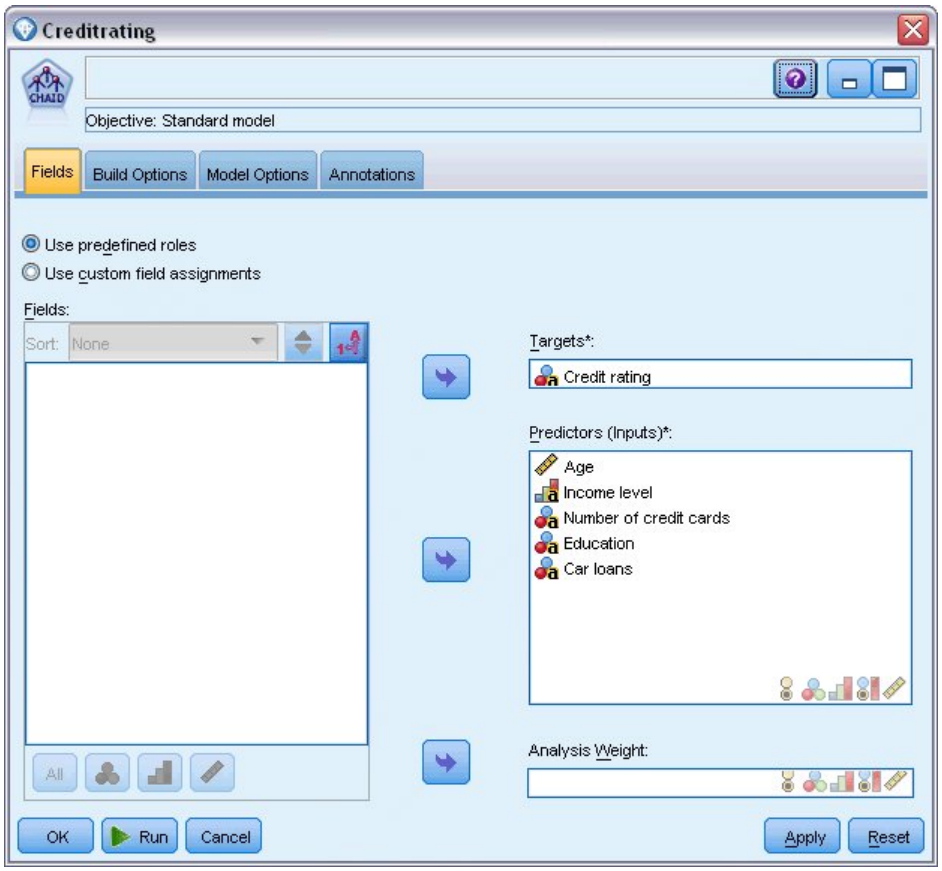


圖 *5. CHAID* 建模節點，欄位標籤

這裡有數個選項，我們可在其中指定要建置的模型類型。

我們想要全新的模型，因此將使用預設選項**建置新模型**。

此外，我們只需要單一標準決策樹狀結構，而無需任何加強功能，因此我們也保留預設的目標選項**建置單一樹狀結構**。

我們可以選擇性地啟動互動式建模階段作業，以容許細部調整模型，此範例只是簡單使用預設模式設定**產生模型**來產生了一個模型。

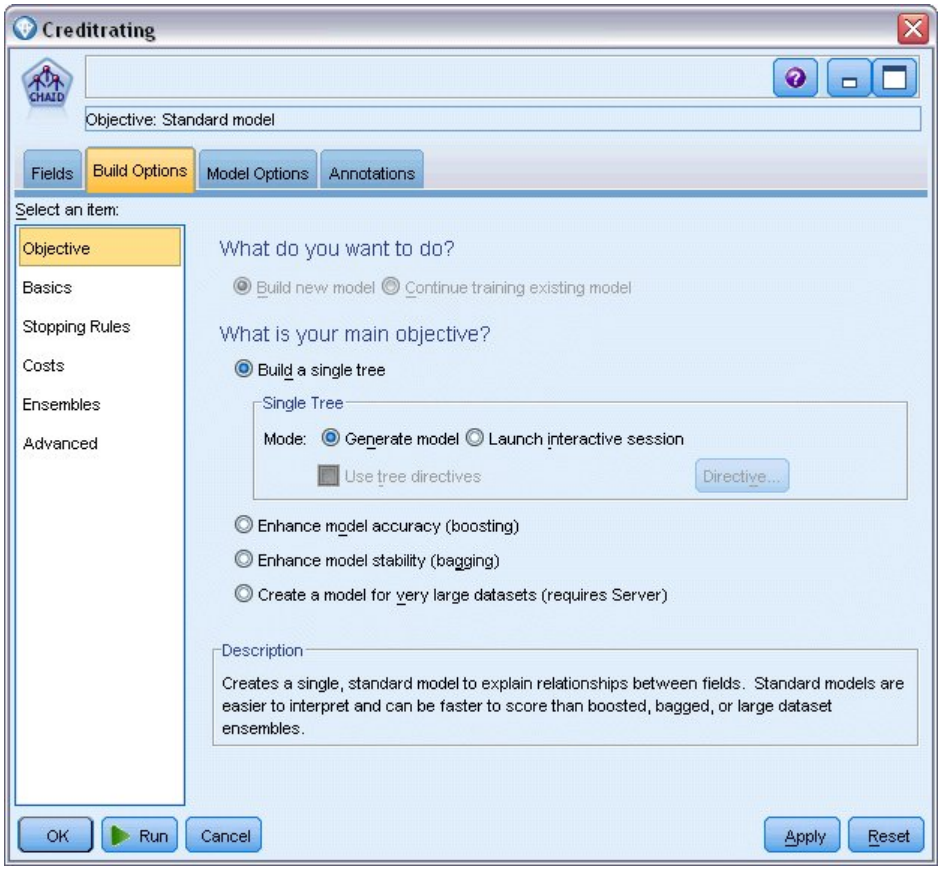


圖 *6. CHAID* 建模節點，建置選項標籤

就此例而言，我們要保持樹狀結構十分的簡單，所以我們要增加上層節點和子節點觀察值的最小值，來限制樹狀結構的成長。

2. 在「建置選項」標籤上，從左側導覽窗格中選取**停止規則**。
3. 選取**使用絕對值**選項。
4. 將**上層分支中的最小記錄**設為 400。
5. 將**子分支中的最小記錄**設為 200。

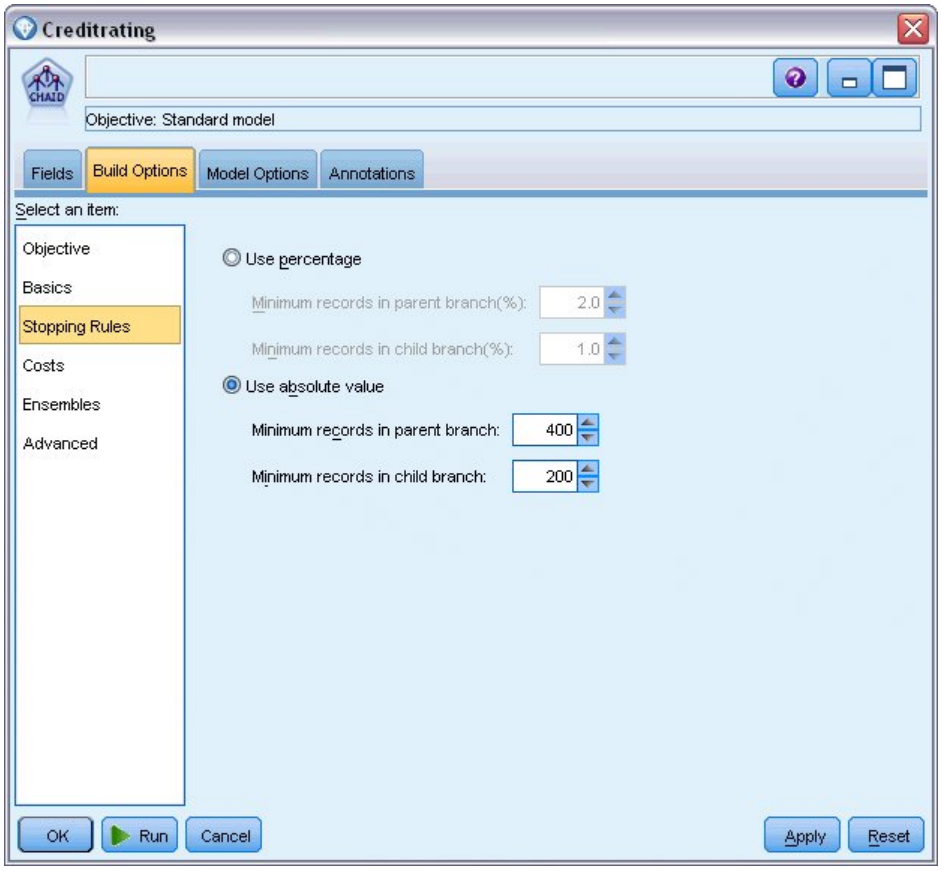


圖 7. 為決策樹狀結構建置設定停止準則

在此範例中，我們可以使用所有其他預設選項，因此按一下**執行**來建立模型。（或者，用滑鼠右鍵按一下節點，並從快速功能表中選擇**執行**，或選取節點並從工具功能表中選擇**執行**。）

瀏覽模型

在執行完成後，模型片段會新增至應用程式視窗右上角的「模型」選用區中，同時也會置於串流畫布上，其中包含建立它的建模節點的鏈結。若要檢視模型詳細資料，請用滑鼠右鍵按一下模型片段，並選擇**瀏覽**（在模型選用區上）或**編輯**（在畫布上）。

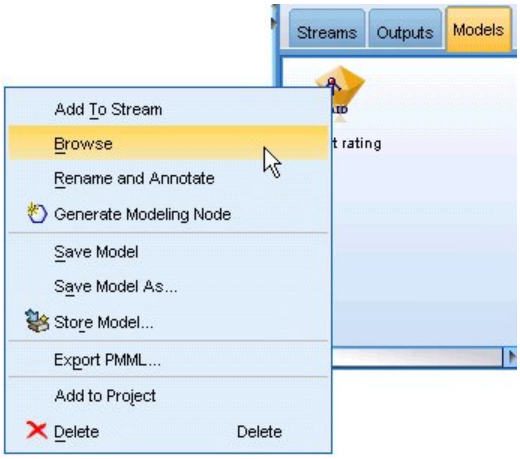


圖 *8.* 「模型」選用區

如果是 CHAID 片段，「模型」標籤會以規則集的形式顯示詳細資料 - 本質上是一系列規則，可基於不同輸入欄位的值將單個記錄分配給子節點。

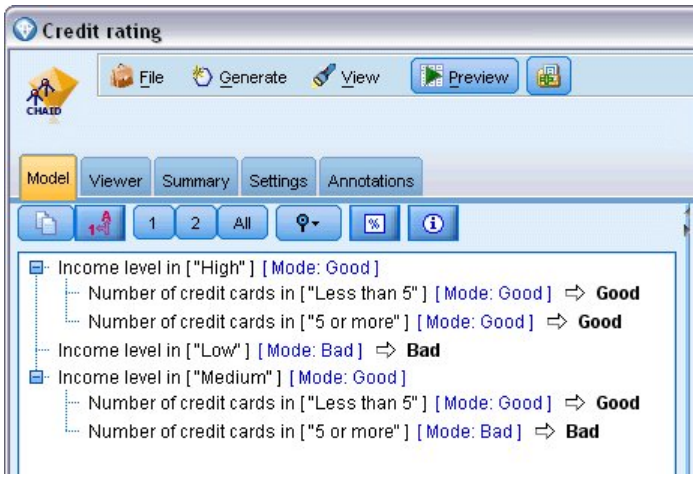


圖 *9. CHAID* 模型片段與規則集

對於每一個決策樹狀結構終端節點（表示將來不再分割的那些樹狀結構節點）- 將傳回佳或差。無論是其中哪一種情況，預測都由落入該節點內的記錄的**眾數**決定，即最常見的回應。

在規則集的右側，「模型」標籤會顯示「預測值重要性」圖表，其顯示評估模型時每個預測值的相對重要性。從中我們可以發現在此情況下，收入層次很顯然是最重要的，而僅剩的另一個重要因素是信用卡數目。

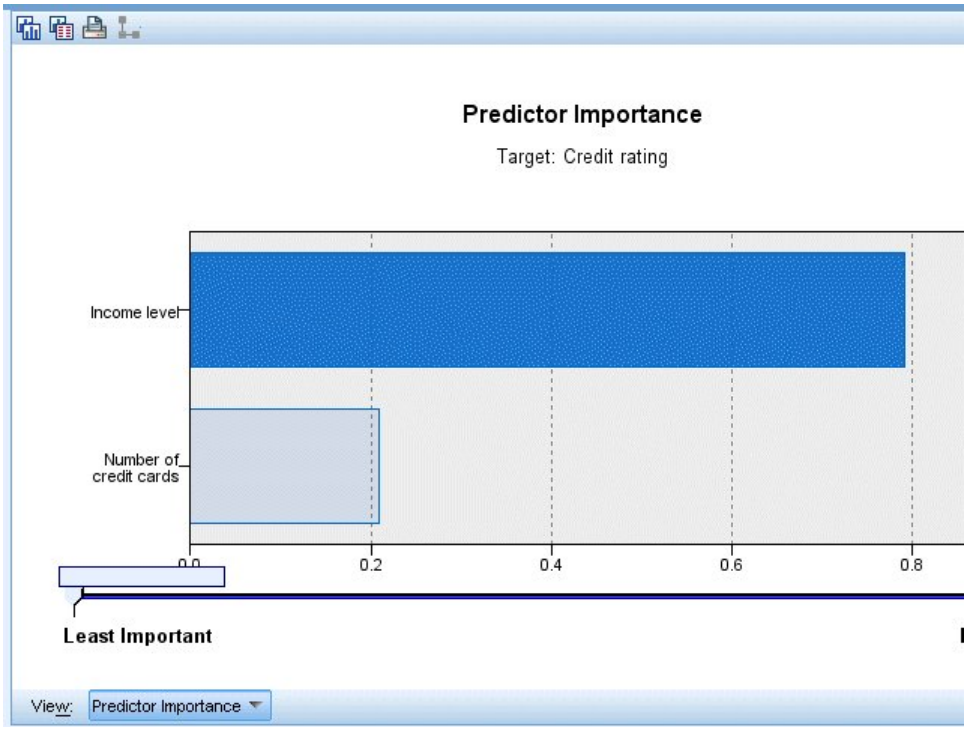


圖 *10.* 預測值重要性圖表

模型片段中的「檢視器」標籤以樹狀結構形式顯示與每一個決策點的節點相同的節點。使用工具列上的「縮放」控制項，可放大或縮小特定節點，以查看樹狀結構的更多內容。

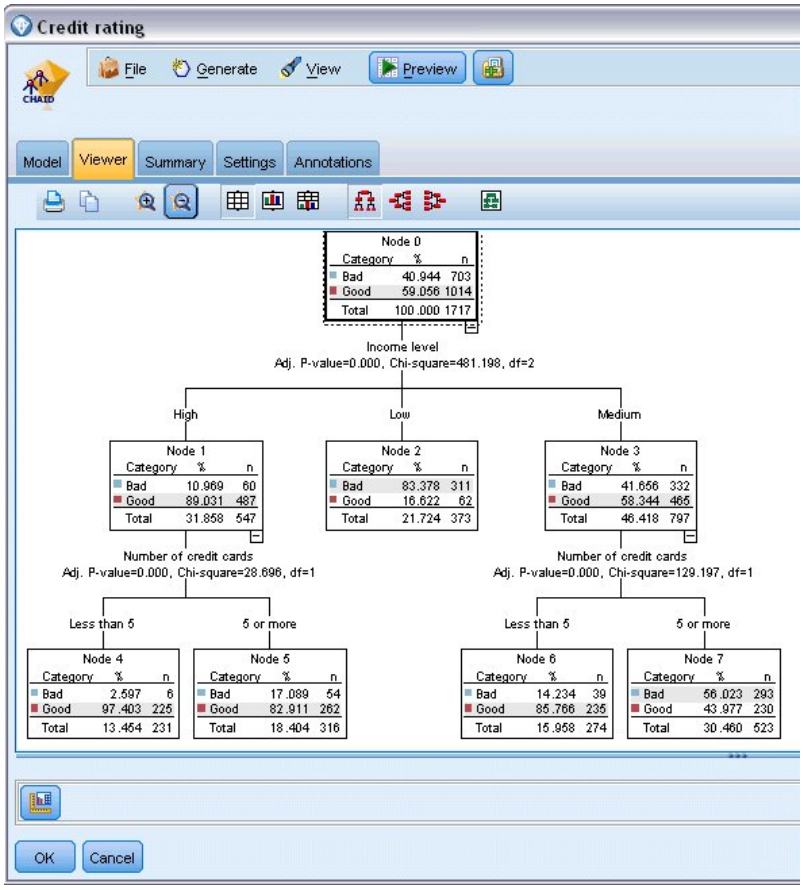


圖 *11.* 模型片段中的檢視器標籤（已選取縮小）

查看樹狀結構上方，第一個節點（節點 0）提供了資料集內所有記錄的摘要。資料集內剛好有 40% 多一點的項目被分類為風險很大。這是非常高的比例，因此我們看看樹狀結構是否提供了造成此結果之因素的任何線索。

我們可以看到，首先由收入層級進行了分割。收入層級處於低種類的記錄會指派給「節點 2」，毫無疑問，此種類包含最高百分比的貸款拖欠者。很明顯，對此種類中客戶借款的風險很高。

但是，此種類中有 16% 的客戶實際並未拖欠，因此，預測也不是一律都正確。沒有任何模型可切實預測每一個回應，但好的模型應能讓我們根據可用的資料預測每一個記錄最可能做出的回應。

同樣，如果我們查看高收入客戶（節點 1），我們發現絕大部分 (89%) 風險很低。但這些客戶超過 10% 的人也會拖欠。我們能否修正借貸準則來將這裡的風險降到最低？

注意查看該模型是如何根據持有的信用卡數目來將這些客戶分割為兩個子種類（節點 4 及 5）。對於高收入客戶，如果我們僅借貸給少於 5 張信用卡的客戶，則可將成功率從 89% 提高到 97% - 更令人滿意的結果。

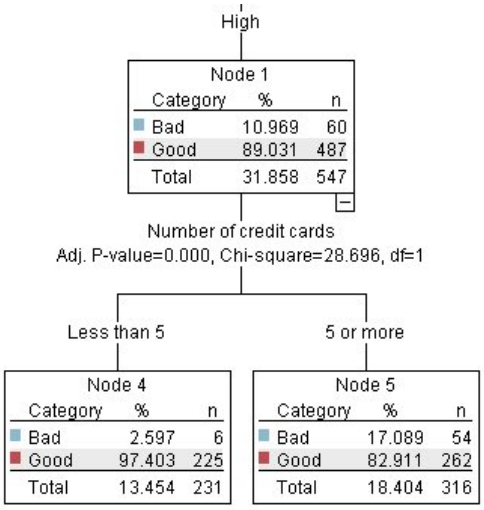


圖 *12.* 高收入客戶的樹狀結構視圖

但中等收入種類（節點 3）的這些客戶又如何呢？他們差不多平分為「佳」和「差」等級。

同樣，子種類（此情況下為節點 6 及 7）可協助我們。此處，僅借貸給少於 5 張信用卡的這些中等收入客戶可將「佳」等級從 58% 提高到 85%，從而得以大幅提高。

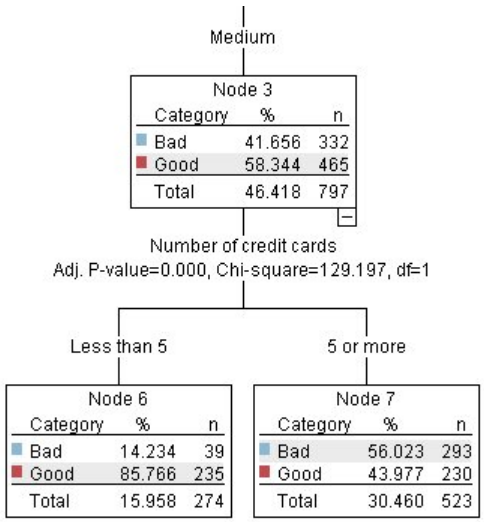


圖 *13.* 中等收入客戶的樹狀結構視圖

因此，我們已瞭解到輸入到此模型的每個記錄都將指派給特定節點，並根據該節點最常見的回應指派佳或差回應。

這種對個別記錄指派預測的過程稱為**評分**。透過對用於評估模型的相同記錄進行評分，我們可評估其在訓練資料（從中得出結果的資料）上執行時的準確程度。我們看一下如何執行此操作。

評估模型

我們已瀏覽模型而瞭解評分的運作方式。但若要評估其運作的準確性，我們需要對部分記錄評分，並將模型預測的回應與實際結果進行比較。我們將對用於評估模型的相同記錄進行評分，從而對觀察的回應及預測回應進行比較。

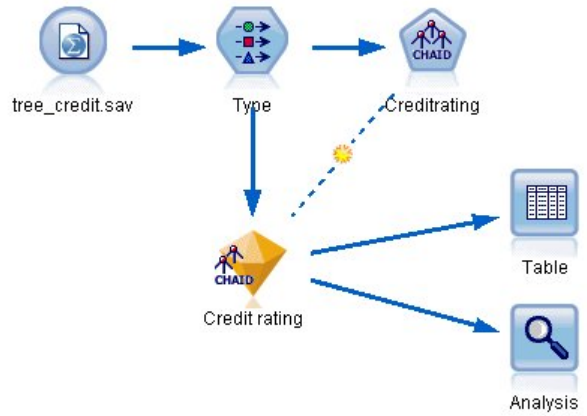
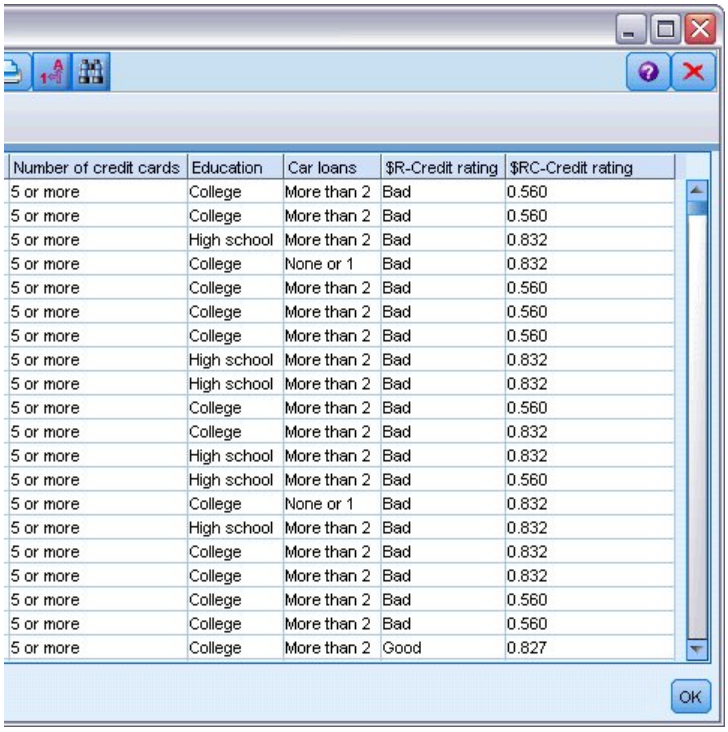


圖 *14.* 將模型片段連接到輸出節點以評估模型

1. 若要查看評分或預測，請將「表格」節點連接到模型片段，按兩下「表格」節點，然後按一下**執行**。

   表格會在名為 *$R-Credit rating* 的欄位中顯示預測的評分，該欄位由模型建立。我們可以將這些值與包含實際回應的原始信用評級欄位進行比較。

   按照慣例，在評分期間產生的欄位名稱基於目標欄位，但具有標準字首。字首 *$G* 及 *$GE* 由「一般線性模型」產生，*$R* 在此情況下是用於 CHAID 模型所產生預測的字首，*$RC* 表示信賴值，*$X* 一般使用集合產生，而 *$XR*、*$XS* 和 *$XF* 分別在目標欄位是「連續」、「種類」、「集」或「旗標」欄位的情況下用作字首。不同的模型類型使用不同的字首集。**信賴度值**是模型本身對每一個預測值之準確性的估計值，範圍從 0.0 到 1.0。



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

圖 *15.* 此表格顯示產生的分數和信賴度值

依預期，預測值與許多記錄的實際回應相符，但並非全部。原因在於每一個 CHAID 終端節點都會有混合回應。預測會符合最常見的那一個，但對於該節點中的其他所有項目都將是錯誤的。（請注意，有 16% 少數未拖欠的低收入客戶。）

若要避免此現象，我們可將樹狀結構繼續分割成越來越小的分支，直到每一個節點完全為 100% - 全部為佳或差，而不含混合回應。但這種模型非常複雜，並且可能無法對其他資料集一般化。

若要找出到底有多少預測是正確的，我們要查看整個表格，並計算預測欄位 *$R-Credit rating* 值符合信用評級值的記錄數。所幸我們可以使用更簡單的方法，即「分析」節點，來自動執行此動作。

2. 將模型片段連接至「分析」節點。
3. 按兩下「分析」節點，然後按一下**執行**。

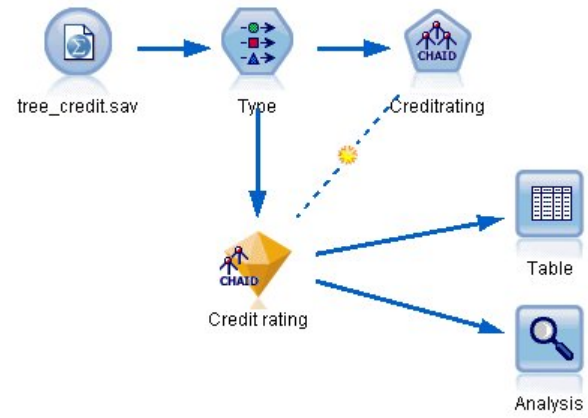


圖 *16.* 連接分析節點

分析顯示，在模型對值進行預測的 2464 條記錄中，有 1899 條記錄（超過 77%）符合實際回應。

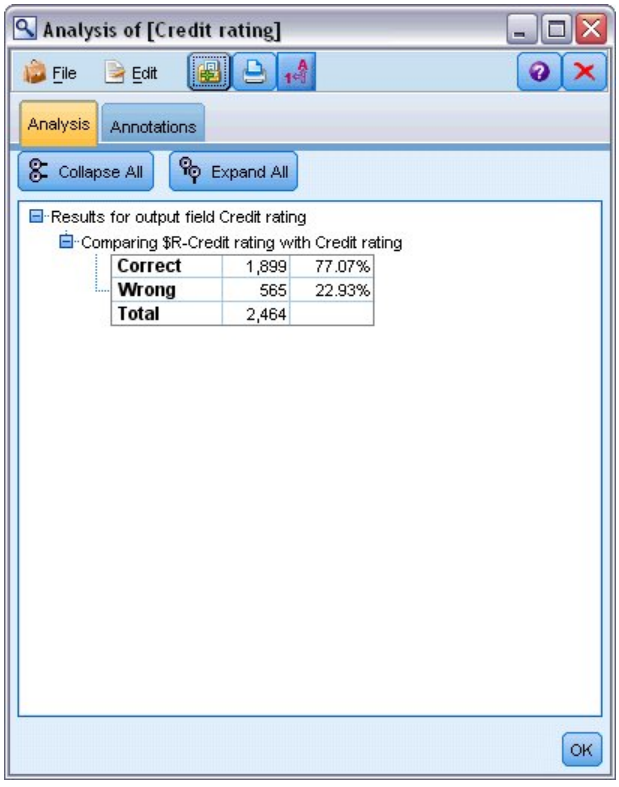


圖 *17.* 比較已觀察和預測回應值的分析結果

因為用於評分的記錄與用於評估模型的記錄相同，所以此結果也有所限制。在實際狀況中，您可以使用「分割」節點來將資料分割為個別樣本，以進行訓練和評估。

通過用某個樣本分割產生模型並用另一個樣本對模型進行測試，可以預判其對其他資料集的擬合優劣。

「分析」節點可讓我們根據已知實際結果的記錄來測試模型。下一階段說明如何使用模型對不知道結果的記錄進行評分。例如，這可能包括目前尚未成為銀行客戶但卻是促銷郵寄潛在目標的人員。

評分記錄

之前，我們對用於評估模型的相同記錄進行評分，以評估模型的準確性。現在，我們將查看如何對用於建立模型的不同記錄集進行評分。使用目標欄位建模的目標如下：研究已知結果的記錄以識別型樣，從而可讓您預測尚不知道的結果。

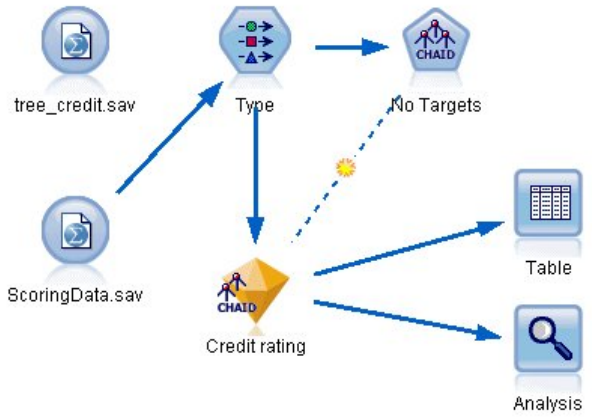


圖 *18.* 連接新資料以進行評分

您可以更新「統計資料檔案」來源節點以指向不同的資料檔案，或者您可以新增來源節點以讀取要進行評分的資料。無論採用何種方式，新資料集必須包含模型使用的相同輸入欄位（年齡、收入層級、教育等），而不是目標欄位信用評級。

或者，您可以將模型片段新增至包含預期輸入欄位的任何串流。無論是從檔案還是資料庫讀取，只要欄位名稱和類型符合模型使用的項目，則來源類型不受影響。

您也可以將模型片段儲存為個別檔案，或以 PMML 格式匯出模型以用於支援此格式的其他應用程式，或者將模型儲存於 IBM SPSS Collaboration and Deployment Services 儲存庫，其為模型提供企業層面的部署、評分及管理。

不論使用何種基礎架構，模型本身都會以相同的方式運作。

摘要

本範例示範建立、評估及對模型進行評分的基本步驟。

- 建模節點透過研究已知結果的記錄來評估模型，並建立模型片段。這有時稱為訓練模型。
- 模型片段可新增至具有預期欄位的任何串流，以對記錄進行評分。透過對已知結果（如現有客戶）的記錄進行評分，可評估執行的效果。
- 當模型執行得足夠良好，讓您感到滿意之後，您可以對新資料（如潛在客戶）進行評分，以預測他們將如何回應。
- 用於訓練或評估模型的資料可能稱為分析或歷程資料；評分資料也可能稱為作業資料。

第 3 章 建模概觀

建模節點概觀

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。

IBM SPSS Modeler 應用程式手冊 為上述多種方法提供了範例以及建模過程的一般簡介。本手冊作為線上指導教學提供，也有 PDF 格式。如需相關資訊，請參閱主題第 4 頁的『應用程式範例』。

建模方法分為以下種類：

- 受監督
- 關聯
- 分區段

受監督模型

受監督模型使用一個或多個**輸入**欄位的值來預測一個或多個輸出（或**目標**）欄位的值。這些技術的一些範例包括：決策樹（C&R 樹狀結構、QUEST、CHAID 和 C5.0 演算法）、迴歸方法（線性、logistic、通用性線性和 Cox 迴歸演算法）、神經網路、支援向量機器和貝葉斯網路。

「受監督」模型可說明組織預測已知的結果，例如顧客是否購買、流失或某交易是否符合某種已知的犯罪型樣。其建模技術包含機器學習、規則歸納、子群組識別、統計技術和多模型產生。

受監督節點

「自動分類器」節點用於建立和對比二元結果（是或否，流失或不流失等）的若干不同模型，使用戶可以選擇給定分析的最佳處理方法。由於受支援 多種建模演算法，因此可以對用戶希望使用的方法、每種方法的特定選項以及對比結果的準則進行選取。節點根據指定的選項產生一組模型並根據用戶指定的準則排列最佳候選項的順等級。

自動數值節點使用多種不同方法估計和對比模型的連續數值範圍結果。此節點和自動分類器節點的工作方式相同，因此可以選擇要使用和要在單個建模傳送中使用多個選項組合進行測試的演算法。受支援的演算法包含神經網路、C&R 樹狀結構、CHAID、線性迴歸、通用性線性迴歸以及受支援向量機器 (SVM)。可基於相關係數度、相對錯誤或已用變數數目對模型進行對比。

分類和迴歸方法 (C&R) 樹狀結構節點產生可用於預測或分類未來觀察的決策樹。該方法通過在每一個步驟最大限度降低不純潔度，使用遞歸分區來將訓練記錄分割為組。如果樹狀結構中某個節點中 100% 的觀察值都的目標欄位的一個特定種類，那麼該節點將被認定為「純潔」。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

QUEST 節點可提供用於建立決策樹的二元分類法，此方法的設計目的是減少大型 C&R 樹狀結構分析所需的處理時間，同時也減少在分類樹狀結構方法中發現的趨勢以便偏愛容許有多個分割的輸入。輸入欄位可以是數值範圍（連續），但目標欄位必須是種類。所有分割都是二元的。

CHAID 使用卡方統計資料來產生決策樹，以確定最佳的分割。與 C&R Tree 和 QUEST 節點不同，CHAID 可產生非二元樹狀結構，表示部分分割有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

C5.0 節點建立決策樹或規則 集。該模型的工作原理是根據在每個層次提供上限資訊收穫的欄位分割樣本。目標欄位必須為種類欄位。容許進行多次多於兩個子群組的分割。

決策清單節點可識別子群組或區段，顯示與整體相關的給定二元結果的概似度的高低。例如，您或許在尋找那些最不可能流失的客戶或最有可能對某個商業活動作出積極回應的客戶。通過自訂區段和並排預覽備選模型來比較結果，您可以將自己的業務知識體現在模型中。決策清單模型由一組規則構成，其中每個規則具備一個條件和一個結果。規則依順序套用，相符的第一個規則將決定結果。

線性迴歸模型基於目標和一個或多個預測值之間的線性關係預測連續目標。

PCA/因素節點提供強大的資料減少技術來減少資料的複雜性。主成份分析（PCA）可找出輸入欄位的線性組合，該組合最好地擷取了整個欄位集合中的變異數，且組合中的各個成分相互正交（相互垂直）。因數分析則試圖識別底層因素，這些因素說明觀測的欄位集合內的相關性型樣。對於這兩種方法，其共同的目標是找到可對原始欄位集合中的資訊進行有效總結的少量衍生欄位。

功能選擇節點會根據某組準則（例如遺漏百分比）篩選可移除的輸入欄位，對於保留的輸入，隨後將對其相對於指定目標的重要性進行排等級。例如，假如某個給定資料集有上千個潛在輸入，那麼哪些輸入最有可能用於對病患結果進行建模呢？

判別分析所做的假設比邏輯迴歸方法的假設更嚴格，但在符合這些假設時，判別分析可以作為邏輯迴歸方法分析的有用替代項或補充。

邏輯迴歸是一種統計技術，它可根據輸入欄位的值對記錄進行分類。它類似於線性迴歸方法，但採用的是種類目標欄位而非數值範圍。

「通用性線性」模型對一般線性模型進行了擴展，這樣依變數通過指定的鏈結函數與因子和共變數線性相關。此外，此模式允許」變數具有非常態分配。它包括統計模型大部分的功能，其中包括線性迴歸、邏輯迴歸方法、用於計數資料的對數線性模型以及區間刪失生存分析模型。

概化線性混合模型 (GLMM) 延伸了線性模型，使得目標可以有非常態分佈，通過指定的連接函數與因子和共變數線性相關，並且觀察可能相關。通用性線性混合模型涵蓋多種模式，從非常態縱向資料的簡單線性迴歸，到複雜的多層級模式。

使用 Cox 迴歸節點，您可以在已有的檢查記錄中建立時間事件的生存分析模型。該模型會生成一個生存分析函數，該函數可預測在給定時間 (*t*) 內對於所給定的輸入變數值相關事件的發生機率。

使用支援向量機器 (SVM) 節點，可以將資料分為兩群組，而無需過度配適。SVM 可以與大量資料集配合使用，例如那些含有大量輸入欄位的資料集。

通過貝葉斯網路節點，你可以利用對真實世界認知的判斷力並結合所觀察和記錄的證據來建立機率模型。該節點重點應用了樹狀結構擴展簡單貝葉斯 (TAN) 和馬爾可夫覆蓋網路，這些算法主要用於分類問題。

自習回應模型（SLRM）節點可用於建立一個包含單個新觀察值或少量新觀察值的模型，通過此模型，無需使用全部資料對模型進行重新訓練即可對模型進行重新評估。

「時間序列」節點會為時間序列資料評估指數平滑化、單變量自身迴歸整合移動平均 (ARIMA)，以及多變量 ARIMA（或稱轉換函數）模型，然後產生未來效能的預測。此「時間序列」節點類似於 SPSS Modeler 第 18 版中不推薦使用的先前「時間序列」節點。但是，此較新「時間序列」節點旨在利用 IBM SPSS Analytic Server 的能力來處理大資料，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。

The *k*-最近相鄰元素 (KNN) 節點將新的觀察值關聯到預測值空間中與其最鄰近的 *k* 個物件的種類或值（其中 *k* 為整數）。相似的觀察值會彼此相鄰，相異的觀察值則會彼此相隔。

空間-時間預測 (STP) 節點使用包含位置資料、預測輸入欄位（預測值）、時間欄位和目標欄位的資料。每個位置有數個資料列，用來代表在每次測量時每個預測工具的值。分析資料之後，它可用來在分析中所使用形狀資料內的任何位置，預測目標值。

關聯模型

關聯模型尋找您資料中的型樣，其中一個或多個實體（如事件、購買或屬性）與一個或多個其他實體相關聯。這些模型構建定義這些關係的規則集。資料中的欄位可以作為輸入和目標。您可以手動尋找這些關聯，但關聯規則演算法可以更快速地完成，並能探索更多複合的型樣。Apriori 和 Carma 模型是使用此類演算法的範例。另一種類型的關聯模型是序列偵測模型，後者可以在按時間建立結構的資料中尋找順序型樣。

關聯模型在預測多個結果時非常有用，例如，購買了產品 X 的顧客也購買了產品 Y 和 Z。關聯模型可以將特定結論（如購買某些產品的決策）與一組條件關聯起來。關聯規則演算法相對於更標準的決策樹演算法（C5.0 和 C&RT）的優勢在於，它可以找到任何屬性間存在的關聯。決策樹演算法將建置只有一個結果的規則，而關聯演算法會嘗試尋找許多規則，每個規則可能具有不同的結果。

關聯節點

「事前」節點從資料擷取一組規則，即擷取資訊內容最多的規則。Apriori 節點提供五種選取規則的方法並使用複雜的編製索引模式來高效地處理大資料集。對於較大的問題，Apriori 訓練的速度通常較快較快；它對可保留的規則數目量沒有任何限制，而且可處理最多帶有 32 個前提條件的規則。「事前」要求輸入和輸出欄位均為種類型欄位，但因為它專為處理此類型資料而進行最佳化，因而處理速度快得多。

CARMA 模型會從資料中擷取一組規則，而不需要您指定輸入或目標欄位。與 Apriori 不同，CARMA 節點提供建立規則設定支援（前提條件和結果支援），而不僅僅是前提條件支援。這就意味著產生的規則可以用於更多應用程式，例如用於尋找產品或服務（前提條件）的清單，這些產品或服務的結果為想在節日期間促銷的商品。

序列節點可探索循序資料或與時間有關的資料中的相關規則。序列是一系列可能會以可預測順序發生的項目集合。例如，一個購買了剃刀和須後水的顧客可能在下次購物時購買剃須膏。「序列」節點基於 CARMA 關聯規則演算法，使用有效的兩段式方法來尋找序列。

「相關規則」節點與 Apriori 節點類似；但是，與 Apriori 不同，「相關規則」節點能夠處理清單資料。另外，「關聯規則」節點可以與 IBM SPSS Analytic Server 配合使用，以正在處理大型資料以及利用更快的平行處理功能。

分區段模型

分區段模型將資料劃分為具有類似輸入欄位型樣的記錄區段或叢集。分區段模型只對輸入欄位感興趣，沒有輸出或目標欄位的概念。分區段模型的範例為 Kohonen 網路、K-Means 叢集、二階叢集和異常偵測等。

在不知道特定結果的情況下（例如，需要識別新犯罪型樣或在客戶群中識別利益群體時），分區段模型（也稱為「叢集模型」）非常有用。叢集作業模型著重於識別相似記錄的群組，以及根據它們所隸屬的群組來標示記錄。此技術的優點在於，不用提前瞭解這些群組及其特性就可以使用，它使叢集模型（其中沒有需要模型預測的預先定義輸出或目標欄位）區別於其他的建模技術。對於這些模型來說，沒有正確或錯誤的結果之分。其值根據它們的以下能力來決定：它們能夠擷取資料中的相關分組並為這些分組提供有用說明。叢集模型通常用於建立在後續分析中用作輸入的叢集或區段（例如，將潛在用戶分成幾個相似的子群組）。

分區段節點

自動叢集節點估計和比較識別具有類似特性記錄群組的叢集模型。節點工作方式與其他自動建模節點相同，使您在一次建模運行中即可試驗多個選項組合。可以使用基本測量比較模型，透過測量嘗試過濾並分級叢集模型的實用性，並基於特定欄位的重要性提供測量。

K-Means 節點將資料集叢集到不同群組（或叢集）。此方法將定義固定的叢集數目量，將記錄迭代分配給叢集，以及調整叢集中心，直到進一步優化無法再精確模型。*k*-means 節點作為一種非監督學習機制，它並不試圖預測結果，而是揭示隱含在輸入欄位集中的型樣。

Kohonen 節點會產生一種類神經網路，此類神經網路可用於將資料集叢集到各個差異群組。此網路訓練完成後，相似的記錄應在輸出對映中緊密地聚集，差異大的記錄則應彼此遠離。您可以通過查看模型塊 中每個單位所擷取觀察的數量來找出規模較大的單元。這將讓您對叢集的相應數量有所估計。

TwoStep 節點使用二階叢集方法。第一步完成簡單資料製作，以便將原始輸入資料壓縮為可管理的子叢集集合。第二步使用層級叢集方法將子叢集一步一步合併為更大的叢集。TwoStep 具有一個優點，就是能夠為訓練資料自動估計最佳叢集數目。它可以高效處理混合的欄位類型和大型的資料集。

Anomaly Detection 節點確定不符合「正常」資料格式的異常觀察值（離群值）。即使離群值不匹配任何已知格式或用戶不清楚自己的查找目標，也可以使用此節點來確定離群值。

資料庫內資料採礦模型

IBM SPSS Modeler 支援與多家資料庫供應商的資料挖掘和建模工具整合，這包含 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。有關完整的詳細資料，請參閱《IBM SPSS Modeler 資料庫內資料採礦手冊》。

IBM SPSS Statistics 模型

如果您在電腦上擁有 IBM SPSS Statistics 安裝和軟體使用權的一個副本，您可以從 IBM SPSS Modeler 存取和執行某些 IBM SPSS Statistics 常式以建立模型和給模型分數。

建立分割模型

通過分割建模，可以使用單個串流來為旗標、列名或連續輸入欄位的每個可能值建立個別的模型，並且生成的模型全部都可從單個模型塊進行存取。輸入欄位的可能值可能對模型具有非常不同的效應。使用分割建模，您可以容易地在串流的一次執行中為每個可能的欄位值建立最佳配適模型。

請注意，交互建模階段作業不能使用分割。您通過互動建模個別指定每個模型，而使用分割會自動建立多個模型，所以使用分割沒有優勢。

分割建模會指定某個輸入欄位為分割欄位。您可以通過在「類型」規格中將欄位角色設定為**分割**來執行此操作。

只能將測量層次為**旗標**、**列名**、**序數**或**連續**的欄位指定為分割欄位。

您可以將多個輸入欄位分配為分割欄位。但是這種情況下，所建立模型數量可能大增。給所選分割欄位值的每個可能組合建立一個模型。例如，如果三個輸入欄位指定為分割欄位，每個欄位具有三個可能值，那麼結果會建立 27 個不同模型。

即使在您將一個或多個欄位指定為分割欄位後，您仍然可以通過建模節點對話框上的勾選框設定來選擇建立多個分割模型還是單個模型。

如果定義了分割欄位但未選取勾選框，那麼只產生一個模型。同樣，如果選取了勾選框但未定義分割欄位，那麼分割被忽略，產生一個模型。

當您執行串流時，在背景為分割欄位的每個可能值建立個別的模型，但只有一個模型塊置於模型選用區和串流畫布中。分割模型塊由分割符號指示；這是疊加在模型塊影像上的兩個灰色矩形。

瀏覽分割模型塊時，您會看到包含已建立的所有個別模型的清單。

您可以通過在檢視器中按兩下塊從清單中檢視單個模型。這樣開啟單個模型的標準瀏覽器視窗。當塊位於畫布中時，按兩下縮圖開啟完整大小的圖表。請參閱第 40 頁的『分割模型檢視器』主題，以取得更多資訊。

一旦將模型建立為分割模型之後，就不能刪除其分割正在處理，也不能從分割建模節點或模型塊下游復原分割。

範例。 某個國內零售商希望按產品種類估計其國內每家店鋪的銷售情況。則其通過使用分割建模，將其輸入資料的「店鋪」欄位指定為分割欄位，這樣能在一次作業中為每個店鋪的每個分類建立個別的模型。其然後可以使用所得資訊比只使用一個模型更加準確地控制庫存層次。

分割和分區

分割與分區共有某些特徵，但其使用方式截然不同。

分區將資料集隨機分為兩部分或三部分：訓練、測試和（選用）驗證，並用於測試單個模型的效能。

分割將按分割欄位的可能值的數目劃分資料集，並用於建立多個模型。

分區和分割工作方式彼此完全不同。您可以在建模節點中選擇一個、兩個或一個也不選。

支援分割模型的建模節點

大量建模節點可建立分割模型。例外狀況的情況是自動叢集、PCA/因子、功能選擇、SLRM、隨機樹狀結構、樹狀結構 AS、線性 AS、LSVM、關聯模型（Apriori、Carma 和序列）、叢集作業模型（K-Means、Kohonen、二階和例外狀況）、Statistics 模型以及用於資料庫內建模的節點。

支援分割建模的建模節點是：


C&R 樹狀結構


Bayes 網路


線性


QUEST


GenLin


GLMM


CHAID


KNN


STP


C5.0


Cox


一類 SVM


神經網路


自動分類器


XGBoost 樹狀結構


決策清單


自動數值


XGBoost 線性


迴歸方法(R)


邏輯


HDBSCAN


區別


SVM


時間序列

受分割影響的特徵

使用分割模型以各種方式影響大量 IBM SPSS Modeler 特徵。本部分提供有關在串流中將分割模型與其他節點配合使用的指引。

記錄處理節點

當在包含年「樣本」節點的串流中使用分割模型時，按分割欄位對記錄進行分層，以實現記錄的平均採樣。當選擇複合作為樣本方法時，此選項可用。

如果串流包含「平衡」節點，那麼平衡適用於輸入記錄的整體集合，而非分割內的記錄子集合。

當通過「聚合」節點來聚合記錄時，如果要計算每個分割的聚合，請將分割欄位設定為索引鍵欄位。

欄位作業節點

通過「類型」節點，可以指定將哪個或哪些欄位用作分割欄位。

註：盡管「總體」節點用於組合兩個或多個模型塊，但其無法用於置換分割動作，因為分割模型包含在單個模型塊內。

建模節點

分割模型不支援預測值重要性（估計模型時預測值輸入欄位的相對重要性）計算。建置分割模型時，會忽略預測值重要性設定。

註：使用分割模型時，會忽略調整傾向分數設定。

KNN（最近相鄰元素）節點只有在設定為預測目標欄位時才支援分割模型。其他設定（只識別最近鄰接項）不建立模型。如果選取選項**自動選取 k**，那麼每個分割模型可能具有不同數量的最近鄰接項。因此，整體模型產生的欄數等於所有分割模型找到的最近鄰接項的最大數。對於最近鄰接項數少於此最大值的分割模型，存在對應數量的已填入 $null 值的欄。請參閱第 293 頁的『KNN 節點』主題，以取得更多資訊。

資料庫建模節點

資料庫內建模節點不支援分割模型。

模型區塊

不可能從分割模型塊**匯出到 PMML**，因為塊包含多個模型，而 PMML 不支援這種包裝。可以匯出到文字或 HTML。

建模節點欄位選項

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。

您必須先指定要用做目標與輸入的欄位，然後才能建置模型。在某些例外狀況下，所有建模節點會使用來自上游「類型」節點的欄位資訊。若您使用「類型」節點來選取輸入和目標欄位，則無須變更此標籤的任何內容。（特殊情況包含序列節點和文字擷取節點，這兩個節點需要在建模節點中指定欄位設定。）

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，請視需要於下方指定欄位。

註：並非所有欄位都對所有節點顯示。

- **使用交易格式（僅限 Apriori、CARMA、MS 關聯規則和 Oracle Apriori 節點）**。如果來源資料為**交易處理格式**，那麼選中此勾選框。此格式的記錄具有兩個欄位，一個為 ID 欄位，一個為內容欄位。每條記錄代形式個交易或單個項目，關聯的項目通過相同的 ID 得以鏈結。如果資料為**表格式**，請取消勾選此方框，表格式中項目由獨立旗標代表，其中每個旗標欄位代表某個特定項目是否出現，且每個記錄代表關聯的項目的完整集合。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。
  - **ID**。 對於交易處理資料，請從清單中選取 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）。
  - **ID 是連續的**。 （僅限 Apriori 和 CARMA 節點）如果您的資料進行了預先排序，以便所有 ID 相同的記錄在資料串流中群組在一起，那麼選取此選項可以加快正在處理速度。如果您的資料未經預先排序（或者您不確定），請將此選項保持未選取狀態，那麼該節點將自動對資料進行排序。

    註：如果您的資料未經過排序而您選取了此選項，那麼可能會在模型中得到無效結果。
  - **內容**。 指定模型的內容欄位。這些欄位包含與關聯建模有關的項目。您可以指定多個旗標欄位（如果資料為表格式）或者一個列名欄位（如果資料為交易格式）。
- **目標**。 針對需要使用一或多個目標欄位的模式，請選取一個或多個目標欄位。這與在「類型」節點中將欄位角色設為目標類似。
- **評估**。 （僅適合自動叢集模型。）不為叢集模型指定目標，但可選取一個評估欄位以確定其重要性等級。此外，還可評估叢集區分此欄位值的程度，從而指示是否可使用叢集來預測此欄位。附註：評估欄位必須是具有多個值的字串。
  - **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。
  - **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。透過使用一個樣本來產生模型，並使用另一個樣本來測試模型，您可以很好地指出模型將概化為與現行資料相似的更大型資料集的程度。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。）
- **分割**。 針對分割模型，選取一或多個分割欄位。這與在「類型」節點中將欄位角色設為分割類似。您可以只將測量層次為**旗標**、**標稱**、**序數**或**連續**的欄位指定為分割欄位。選擇作為分割欄位的欄位無法用作目標、輸入、分割區、頻率或加權欄位。請參閱第 24 頁的『建立分割模型』主題，以取得更多資訊。
- **使用頻率欄位**。 此選項可讓您選取某個欄位作為一個頻率加權。如果訓練資料中的每條記錄代表多個單元（例如，您正在使用聚合的資料），那麼可採用此項。欄位值應該為每筆記錄代表的單位數。請參閱第 28 頁的『使用頻率和加權欄位』主題，以取得更多資訊。

註：如果您看到錯誤訊息**meta 資料（在輸入/輸出欄位上）無效**，請確保已指定所有必填欄位，例如「頻率」欄位。

- **使用加權欄位**。 此選項可讓您選取某個欄位作為一個觀察值加權。觀察值加權是用來說明輸出欄位不同等級間的變異數差異。請參閱第 28 頁的『使用頻率和加權欄位』主題，以取得更多資訊。
- **後繼**。 對於規則歸納節點 (Apriori)，請選取在生成的規則集中用作結果的欄位。（這對應於「類型」節點中角色為目標或兩者的欄位。）
- **先行**。 對於規則歸納節點 (Apriori)，請選取在生成的規則集中用作前提條件的欄位。（這對應於「類型」節點中角色為輸入或兩者的欄位。）

某些模型的「欄位」標籤與本節所述「欄位」標籤不同。

- 請參閱第 231 頁的『序列節點欄位選項』主題，以取得更多資訊。
- 請參閱第 221 頁的『CARMA 節點欄位選項』主題，以取得更多資訊。

使用頻率和加權欄位

頻率和加權欄位用於賦予某些記錄高於其他記錄的附加重要性，例如，因為您知道一部分人未在訓練資料（加權）中代表出來，或者因為一個記錄代表多個相同觀察值（頻率）。

- 頻率欄位的值應為正整數。次數加權小於或等於零的記錄將排除在分析之外。非整數次數加權將四捨五入為最近的整數。
- 觀察值加權應為正數但不一定是整數值。觀察值加權小於或等於零的記錄將排除在分析之外。

評分頻率和加權欄位

頻率和加權欄位用於訓練模型，但不用於評分，因為每條記錄的分數基於該記錄的特性，而與它代表的觀察值個數無關。例如，假設您有下表格中的資料。

表 *1.* 資料範例

已婚	**已回應**
是	是
是	是
是	是
是	否
否	是
否	否
否	否

基於上表，可以得出這樣的結論：四分之三的已婚者對促銷作出回應；而三分之二的未婚者對此未作出回應。因此，您將相應地對任何新記錄進行評分，如下表格所示。

表 *2.* 已評分記錄範例

已婚	**$-已回應**	**$RP-已回應**
是	是	0.75 (3/4)
否	否	0.67 (2/3)

另外，您可以使用頻率欄位更簡潔地儲存訓練資料，如下表格所示。

表 *3.* 已評分記錄替代範例

已婚	**已回應**	**次數**
是	是	3
是	否	1
否	是	1
否	否	2

因為此表完全代表同一資料集，因此可以建立相同的模型並僅根據婚姻狀況預測回應率。如果評分資料中有十位已婚者的記錄，那麼無論這十個人是代表十條獨立的記錄，還是頻率為 10 的一個人，都可預測他們每位的回答均為是。雖然通常情況下加權不是整數，但可以認為它近似表示記錄的重要性。這就是對記錄進行評分時不使用頻率和加權欄位的原因。

評估和比較模型

某些模型類型可支援頻率欄位，某些可支援加權欄位，還有一些可同時支援這兩種欄位。但在套用這兩種字段的所有情況中，它們僅用於建立模型，在套用「評估」節點或「分析」節點對模型進行評估時，或者在套用受「自動分類器」節點和「自動數值」節點受支援的大部分方法進行模型分級時，均不考慮套用這兩種字段。

- 例如，在使用評估表比較模型時將忽略頻率和加權值。這將在使用頻率和加權欄位的模型與不使用這些欄位的模型之間進行層次比較，但同時意味著，必須使用不依賴頻率或加權欄位並且可以準確代表總體的資料集才能獲得準確的評估。在實際套用中，要執行此操作，就要確保套用頻率欄位值或加權欄位值永遠無效或 1 的測試樣本評估模型。（這種限制僅適用於評估模型；如果訓練樣本和測試樣本的頻率值或加權值始終為 1，那麼首次不必套用這兩種欄位。）
- 如果使用「自動分類器」基於「利潤」對模型進行分級，那麼可考慮頻率，在這種情況下推薦使用此方法。
- 如果有必要，可以使用分割區節點，將資料分割為訓練樣本和測試樣本。

建模節點分析選項

多數建模節點都包含「分析」標籤，您可以通過該標籤獲取預測值重要性資訊以及原始傾向分數和已調整的傾向分數。

模型評估

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，對於部分模型，需要較長時間來計算預測值重要性，尤其是處理大型資料集時，結果便是依預設會關閉部分模型的預測值重要性。預測值重要性對於決策清單模型無法使用。如需相關資訊，請參閱第 37 頁的『預測值重要性』。

傾向分數

可以在建模節點中和模型塊的「設定」標籤上啟用傾向分數。唯有當選取的目標是旗標欄位時此功能才可用。請參閱第 30 頁的『傾向分數』主題，以取得更多資訊。

計算原始傾向評分。 原始傾向分數僅衍生自基於訓練資料的模型。如果模型預測值為 *true*（將回應），那麼傾向與 P 相同，其中 P 為預測的可能性。如果模型預測的值為假，那麼計算出的傾向為 (1 – P)。

- 如果建立模型時選擇了此選項，那麼依預設將在模型塊中啟用傾向分數。不過，無論是否在建模節點中選擇了原始傾向分數，都可以始終在模型塊中選擇啟用原始傾向分數。
- 對模型進行評分時，原始傾向評分將被新增到將 *RP* 字母附加到標準字首的欄位中。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RRP-churn*。

計算調整傾向評分。 原始傾向僅基於由可能過度擬合的模型指定的估計，這將導致過於樂觀地估計傾向。已調整的傾向試圖通過查看模型在測試或驗證分割區的性能或通過調整傾向來彌補，以相應地給作出更好的估計。

- 此設定要求串流中出現有效的分割區欄位。
- 與原始信賴度分數不同，已調整的傾向評分必須在建立模型時計算；否則，對模型塊進行評分時該分數將不存在。

- 對模型進行評分時，在將 *AP* 字母附加到標準字首的欄位中新增已調整的傾向評分。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RAP-churn*。已調整的傾向分數不適用於邏輯迴歸模型。
- 在計算已調整的傾向分數時，必須尚未已平衡用於計算的測試或驗證分割區。為避免這一點，請確保在任何上游平衡節點中已選取**僅平衡訓練資料**選項。此外，如果已在上游獲取了複合樣本，那麼這將導致已調整的傾向分數無效。
- 已調整的傾向分數不適用於「增強型」樹狀結構和規則集模型。請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。

依據。對於要進行計算的已調整的傾向分數，串流中必須出現一個分割區欄位。可以指定是使用測試分割區還是驗證分割區進行此計算。為獲取最佳結果，測試或驗證分割區包含的記錄數量應至少與用於訓練原始模型的分割區所包含的記錄數相同。

傾向分數

對於傳回預測為 是或 否的模型，您除了可以要求標準預測和信賴度值以外，還可要求傾向分數。傾向分數指示特定結果或回應的可能性。下表格提供了一個範例。

表 *4.* 傾向評分

客戶	**要回應的傾向**
Joe Smith	35%
Jane Smith	15%

傾向分數僅適用於有旗標目標的模型，並且指示為欄位定義的值為 *true*的可能性，如在來源節點或類型節點中指定的那樣。

傾向分數與信賴度分數

傾向分數不同於信賴度分數，後者適用於目前預測，值為是或否。例如，如果預測為否，那麼高信賴度實際表示不作出回應的可能性較大。傾向分數可以迴避此限制，從而輕鬆比較所有記錄。例如，信賴度為 *0.85* 的 否 預測將轉換為 *0.15*（或 *1* 減 *0.85*）的原始傾向。

表 *5.* 信任評分

客戶	**預測**	**信賴度**
Joe Smith	會回應	.35
Jane Smith	不會回應	.85

獲得傾向分數

- 可以在建模節點中的「分析」標籤或模型塊中的「設定」標籤上啟用傾向分數。唯有當選取的目標是旗標欄位時此功能才可用。請參閱第 29 頁的『建模節點分析選項』主題，以取得更多資訊。
- 也可以通過總體節點計算傾向分數，具體取決於所用的總體方法。

計算已調整的傾向分數

計算已調整的傾向分數將作為建立模型過程的一部分，否則沒有可用的已調整的傾向分數。構建模型後，則可使用測試或驗證分割區中的資料對模型進行評分，同時通過在該分割區上分析原始模型的效能，構建一個提供已調整的傾向分數的新模型。根據模型的類型，可以使用兩種方法之一來計算已調整的傾向分數。

- 對於規則集模型和樹狀結構模型，要產生已調整的傾向分數，可通過重新計算每個樹狀結構節點上每個種類的頻率（適用於樹狀結構模型）或重新計算每個規則的支援和信賴度（適用於規則集模型）。這樣一來，已要求已調整的傾向分數時將使用與原始模型一起儲存的新規則集模型或樹狀結構模型。每次將原始模型套用到新資料時，都會隨之將新模型套用到原始傾向分數以產生已調整的分數。
- 對於其他模型，通過對測試或驗證分割區上的原始模型進行評分而生成的記錄將按其原始傾向評分進行分級。接著，對定義非線性函數的神經網路模型進行訓練，該函數從每個分級的平均原始傾向中對映到相同分級的平均觀測傾向中。正如之前對樹狀結構模型的說明，得出的類神經網路模型將與原始模型一起儲存，並且在已要求已調整的傾向分數時套用到原始傾向分數。

關於測試分割區中遺漏值的警告說明。測試/驗證分割區中遺漏輸入值的處理方法隨模型不同而有所差異（請參閱各個模型評分演算法以獲取詳細資料）。有遺漏輸入值時，C5 模式無法計算調整傾向。

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。

註：僅「決策樹」模型容許在建立時指定成本。

模型塊

圖 *19.* 模型區塊

模型塊是模型的儲存器，其中包含一組規則、方程式或方程式，它們代表在 SPSS Modeler 中模型建置作業的結果。模型塊的主要用途是對資料進行評分以產生預測，或者實現對模型內容進行進一步分析。在畫面上開啟模型塊後，可以請參閱有關模型的各類詳細資料，例如，在模型建立中輸入欄位的相對重要性。要檢視預測，則需要進一步添加並執行處理或輸出節點。請參閱第 41 頁的『使用串流中的模型塊』主題，以取得更多資訊。

圖 *20.* 從建模節點指向模型區塊的模型鏈結

在成功地執行建模節點後，會在串流畫布上放置對應的模型塊，並以金色鑽石形圖示代表（因此稱之為「塊」）。在串流畫布上顯示的模型塊，帶有到位於建模節點之前的最近合適節點的連線（實行），以及到建模節點本身的鏈結（虛行）。

此外，模型塊也放置在位於 IBM SPSS Modeler 視窗右上角落的「模型」選用區中。從任一位置均可已選取模型塊，並瀏覽模型的詳細資料。

在建模節點成功執行後，模型塊始終位於「模型」選用區中。可以設定使用者選項來控制是否也將模型塊置於串流畫布上。

下列主題提供了使用 IBM SPSS Modeler 中模型塊的相關資訊。要深入瞭解所使用的演算法，請參閱產品下載過程中以 PDF 檔案形式提供的《*IBM SPSS Modeler* 演算法手冊》。

模型鏈結

依預設，在畫布上顯示的模型塊帶有指向建立它的建模節點的鏈結。這在具有多個模型塊的複合串流中特別有用，它使您能夠識別將被每個建模節點更新的模型塊。每個連結包含一個符號，指出在執行建模節點時是否取代模型。請參閱第 34 頁的『取代模型』主題，以取得更多資訊。

定義和刪除模型鏈結

您可以在畫布上手動定義和刪除模型鏈結。在定義新的鏈結後，游標將變成鏈結游標。

圖 *21.* 鏈結游標

定義新鏈結（快速功能表）

1. 用滑鼠右鍵按一下要作為鏈結起點的建模節點。
2. 從環境定義功能表中選擇**定義模型鏈結**。
3. 按一下想要鏈結在其中結束的區塊。

定義新鏈結（主功能表）

1. 按一下要作為鏈結起點的建模節點。
2. 從主功能表中，選擇：

   編輯 > **節點** > **定義模型鏈結**
3. 按一下想要鏈結在其中結束的區塊。

刪除現有鏈結（快速功能表）

1. 用滑鼠右鍵按一下位於鏈結終點的模型塊。
2. 從環境定義功能表中選擇**刪除模型鏈結**。

或者：

1. 用滑鼠右鍵按一下位於鏈結中部的符號。
2. 從快速功能表中選擇**刪除鏈結**。

刪除現有鏈結（主功能表）

1. 按一下要刪除其鏈結的建模節點或模型塊。
2. 從主功能表中，選擇：

   編輯 > **節點** > **移除模型鏈結**

複製和貼上模型鏈結

如果複製了帶鏈結的模型塊，但未包括其建模節點，那麼當將其貼上到同一串流中時，貼上後的模型塊將具有到建模節點的鏈結。新鏈結具有與原始鏈結相同的模型取代狀態（請參閱第 34 頁的『取代模型』）。

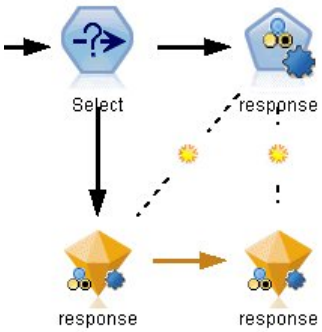


圖 22. 複製並貼上鏈結的區塊

如果將模型塊連同其鏈結的建模節點一起複製並貼上，那麼無論物件貼上到同一串流還是新串流中，鏈結都將保留。

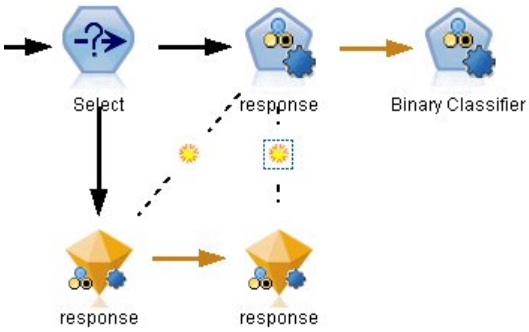


圖 23. 複製並貼上鏈結的區塊

註：如果複製了帶鏈結的模型塊，但未包括其建模節點，那麼將其貼上到新串流（或不包含建模節點的 SuperNode）中時，鏈結將中斷，並且將只貼上模型塊。

模型鏈結和 SuperNode

如果定義 SuperNode 以包含鏈結模型的建模節點或模型塊（但未同時包含），鏈結會中斷。展開 SuperNode 不會還原鏈結，只能通過復原建立 SuperNode 來完成此操作。

取代模型

您可以選擇在重新執行建立模型塊的建模節點時是否取代（即更新）現有模型塊。如果關閉取代選項，那麼重新執行建模節點時將建立新的模型塊。

每個從建模節點指向模型塊的鏈結都包含一個符號，指出在重新執行建模節點時是否取代模型。

圖 *24.* 模型取代處於開啟狀態的模型鏈結

初始顯示鏈結時，模型取代處於開啟，並通過鏈結中的小旭日形符號指示。在此狀態下，重新執行位於鏈結一端的建模節點就會更新另一端的模型塊。

圖 *25.* 模型取代處於關閉狀態的模型鏈結

如果模型取代處於關閉，那麼鏈結符號取代為灰色點。在此狀態下，重新執行位於鏈結一端的建模節點會在畫布上新增一個更新後的模型塊。

在任一情況下，在「模型」選用區中是更新現有模型塊還是新增模型塊，取決於**取代原有模型**系統選項的設定。

執行順序

當執行具有包含模型塊的多個分支的串流時，首先對串流進行評估，以確保先執行模型取代處於開啟的分支，然後再執行使用結果模型塊的任何分支。

如果您的需求更為複合，那麼可通過 Script 手動設定執行順序。

變更模型取代設定

1. 用滑鼠右鍵按一下鏈結上的符號。
2. 根據情況選擇**開啟（關閉）模型取代**。

註：模型鏈結上的模型取代設定將置換「使用者選項」對話框（工具 > 選項 > 使用者選項）的「通知」標籤上的設定。

模型選用區

通過模型選用區（位於管理員視窗的「模型」標籤中），您可以按各種方式使用、檢查和修改模型塊。

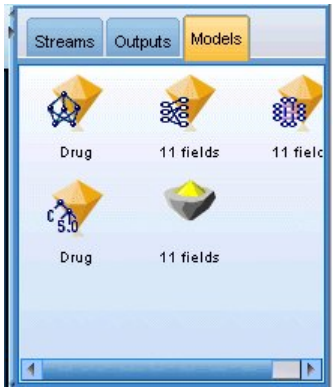


圖 *26.* 「模型」選用區

用滑鼠右鍵按一下模型選用區中的模型塊，開啟帶有下列選項的環境定義功能表：

- **新增至串流中。**將模型塊新增到目前處於作用中狀態的串流中。如果串流中存在選定節點，當可以連線時，模型塊將連接至選定節點，否則鏈接到最近的可能節點。如果建立模型的建模節點仍然在串流中，那麼顯示的模型塊將帶有到建模節點的鏈結。
- **瀏覽。** 開啟塊的模型瀏覽器。
- **更名並進行註釋。**通過此選項，您可以更名模型塊和/或修改模型塊的註解。
- **產生建模節點。** 如果要修改或更新某個模型塊，但無法使用用於建立該模型的串流，那麼可以使用此選項與建立原始模型相同的選項來重新生成一個建模節點。
- **儲存模型，儲存模型為。**將此模型塊儲存到外部產生模型 (.gm) 二進位檔。
- **儲存模型。**在 IBM SPSS Collaboration and Deployment Services 儲存庫 中儲存模型塊。
- **匯出 PMML。** 以預測模型標記語言 (PMML) 格式匯出模型塊，其可用於 IBM SPSS Modeler 之外的新資料評分。**匯出 PMML** 可用於所有產生的模式節點。
- **新增到專案中。**儲存模型區塊並將其新增至現行專案。在「類別」標籤上，該片段將新增至「產生的模型」資料夾。在 CRISP-DM 標籤上，它將新增至預設專案階段。
- **刪除。**從選用區中刪除模型塊。

用滑鼠右鍵按一下模型選用區中的未佔用區域，開啟帶有下列選項的快速功能表：

- **開啟模型。**載入先前在 IBM SPSS Modeler 中建立的模型塊。
- **擷取模型。**從 IBM SPSS Collaboration and Deployment Services 儲存庫擷取儲存的模型。
- **載入選用區。**從外部檔案載入儲存的模型選用區。
- **擷取選用區。**從 IBM SPSS Collaboration and Deployment Services 儲存庫擷取儲存的模型選用區。
- **儲存選用區。**將模型選用區的所有內容儲存到外部產生模型選用區 (.gen) 檔案。
- **儲存選用區。**將模型選用區的所有內容儲存到 IBM SPSS Collaboration and Deployment Services 儲存庫中。
- **清空選用區。**從選用區中刪除所有模型塊。
- **將選用區新增到專案中。**儲存模型選用區並將其新增到目前專案。在「類別」標籤上，該片段將新增至「產生的模型」資料夾。在 CRISP-DM 標籤上，它將新增至預設專案階段。
- **匯入 PMML。**從外部檔案載入模型。可以開啟、瀏覽由 IBM SPSS Statistics 或其他支援此格式的應用程式所建立的 PMML 模型並對其進行評分。請參閱第 41 頁的『將模型匯入和匯出為 PMML』主題，以取得更多資訊。

瀏覽模型塊

通過模型塊瀏覽器，您可以檢查和使用模型的結果。在瀏覽器中，您可以儲存、列印或匯出產生模型，檢查模型摘要，檢視或編輯模型註釋。對於某些類型的模型塊，還可以產生新的節點，例如「過濾器」節點或「規則集」節點。對於某些模型，您還可以檢視模式參數，如規則或叢集中心。對於某些類型的模型（基於樹狀結構的模型和叢集模型），您可以檢視其模型結構的圖表顯示。使用模型塊瀏覽器的控制項如下方所述。

功能表

「檔案」功能表。所有模型塊均有一個「檔案」功能表，其中包括下列選項的子集合：

- **儲存節點。**將模型塊儲存到某個節點 (.nod) 檔案。
- **儲存節點。**在 IBM SPSS Collaboration and Deployment Services 儲存庫中儲存模型塊。
- **頁首和頁尾。**通過此選項，您可以編輯頁的頁首和頁尾，以便從模型塊進行列印。
- **版面設定。**通過此選項，您可以變更版面設定，以便於從模型塊進行列印。
- **預覽列印。**顯示模型塊的列印預覽。從子功能表中選取要預覽的資訊。
- **列印。** 列印模型塊的內容。從子功能表中選取要列印的資訊。
- **列印檢視。**列印現行視圖或全部視圖。
- **匯出文字。**將模型塊內容匯出到某個文字檔中。從子功能表中選取您要匯出的資訊。
- **匯出 HTML。**將模型塊內容匯出到 HTML 檔案中。從子功能表中選取您要匯出的資訊。
- **匯出 PMML。** 以預測模型標記語言 (PMML) 格式匯出模型，匯出的文件可在其他 PMML 相容軟體中使用。請參閱第 41 頁的『將模型匯入和匯出為 PMML』主題，以取得更多資訊。
- **匯出 SQL。**以結構化查詢語言 (SQL) (SQL) 匯出模型，可以通過其他資料庫來編輯和使用匯出的 SQL。

  註：僅在下列模型中提供了 SQL 匯出：C5、C&RT、CHAID、QUEST、線性迴歸、邏輯迴歸、類神經網路、PCA／因子以及決策清單模型。
- **針對伺服器評分配接器發佈。**將模型發行到安裝有評分配接器的資料庫中，可在資料庫中進行模型評分。請參閱第 43 頁的『為評分配接器發行模型』主題，以取得更多資訊。

「產生」功能表 多數模型塊還具有「產生」功能表，通過此功能表可以根據模型塊產生新節點。此功能表中的可用選項取決於您所瀏覽模型的類型。請參閱特定的模型區塊類型以瞭解您可以從特定模型產生之項目的相關詳細資料。

「檢視」功能表。在模型塊的「模型」標籤上，此功能表允許您顯示或隱藏在目前模式下可用的各類直觀表示工具列。要使全部工具列可用，可從「一般」工具列中選取「編輯模式」（畫筆圖示）。

「預覽」按鈕。某些模型塊具有「預覽」按鈕，允許您檢視模型資料的樣本，包含由建模過程建立的額外欄位。顯示的預設列數為 10；但是，您可以在串流內容中變更此值。

「新增到目前專案」按鈕。儲存模型區塊並將其新增至現行專案。在「類別」標籤上，該片段將新增至「產生的模型」資料夾。在 CRISP-DM 標籤上，它將新增至預設專案階段。

模型塊概要/資訊

模型塊的「概要」標籤或「資訊」視圖顯示了關於欄位、建構設定值和模型估計過程的資訊。結果呈現在樹狀結構視圖中，可以按一下特定項目來展開或收合該視圖。

分析。 顯示模型的相關資訊。具體詳細資料因模型類型而異，這些資訊可在每種模型塊的相應章節中找到。另外，如果執行了附加到此建模節點的「分析」節點，那麼還會在此部分顯示該分析中的資訊。

欄位。 列出建立模型時用作目標和輸入的欄位。對於分割模型，也列出確定分割的欄位。

註：在具有增強或組裝模式的神經網路模型、線性模型和其他模型的資訊視圖中，所顯示的圖示相同（列名圖示），而與類型為旗標、列名還是序數無關。

建構設定值/選項。包含建立模型時使用的設定的相關資訊。

訓練摘要。 顯示模型類型、用於建立模型的串流、建立模型的使用者、模型建立時間和建立模型所用時間。請注意，只有「摘要」標籤上提供建立模型所耗用的時間，「資訊」視圖中不提供此時間。

預測值重要性

一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

預測值重要性對於可生成相應重要性統計標準的模型可用，包含神經網路模型、決策樹（C&R 樹狀結構、C5.0、CHAID 和 QUEST）、貝葉斯網路模型、判別分析模型、SVM 和 SLRM 模型、線性和邏輯迴歸模型、通用性線性模型以及最近相鄰元素 (KNN) 模型。對於這些模型中的大部分而言，可以在建模節點的「分析」標籤上啟用預測值重要性。請參閱第 29 頁的『建模節點分析選項』主題，以取得更多資訊。有關 KNN 模型，請參閱第 295 頁的『鄰接項』。

註：對於分割模型，預測值重要性不受支援。建置分割模型時，會忽略預測值重要性設定。請參閱第 24 頁的『建立分割模型』主題，以取得更多資訊。

計算預測值重要性所用的時間遠遠大於建立模型的用時，特別當使用大的資料集時。對於 SVM 和邏輯迴歸模型，計算變量重要性的用時比對其他模型執行此操作的用時都要長，所以依預設這兩種模型均停用此功能。使用一個包含許多預測值的資料集時，使用「功能選擇」節點進行初始篩選可以較快地生成結果（請參閱以下方內容）。

- 如果適用，可以從測試分割區計算出預測值重要性。否則，就使用訓練資料。
- 預測值重要性也適用於 SLRM 模型，但需要使用 SLRM 演算法進行計算。請參閱第 283 頁的『SLRM 模型塊』主題，以取得更多資訊。
- 可以使用 IBM SPSS Modeler 的圖表工具進行交互、編輯，並儲存圖表。
- 您可以選擇性地根據預測值重要性圖表中的資訊，來產生「過濾器」節點。請參閱第 38 頁的『基於重要性過濾器變數』主題，以取得更多資訊。

預測值重要性和功能選擇

在某些情況下，模型塊中顯示的預測值重要性圖表可能似乎給出與「功能選擇」節點相似的結果。當功能選擇基於每個輸入欄位與特定目標（與其他輸入無關）的關係強度對輸入欄位進行排等級時，預測值重要性圖表將顯示此特定模型中各個輸入的相對重要性。因此，在篩選輸入時使用功能選擇可能較為保守。例如，如果工作職務 和工作種類 與薪資之間有強烈的關係，特性選取就會指示這兩者都很重要。但在建模時，還需考慮互動性和相關性。這樣，當兩個輸入的大部分資訊都相同時，您可能會發現僅使用了兩個輸入之一。在實際應用中，功能選擇對預篩選最有用，特別是處理包含大量變數的較大資料集時，而預測值重要性在微調模型時更為有用。

單一模型與自動化建模節點之間的預測值重要性差異

根據您要從個別節點建立單一模型還是使用自動化建模節點來生成結果，您可能會看到預測值重要性的細微差異。這種實現上的差異是由一些專案限制所致。

例如，借助 CHAID 之類的單一分類器，此算法在計算重要性值時應用停止規則並使用機率值。相反，自動分類器不使用中止規則，而是直接在計算中使用預測的標籤。這些差異可能意味著，如果您使用自動分類器來生成單一模型，那麼與針對單一分類器計算出的值相比，重要性值可以被認為是粗略估計值。要獲取最準確的預測值重要性值，我們建議使用單一節點來取代自動化建模節點。

基於重要性過濾器變數

您可以選擇性地根據預測值重要性圖表中的資訊，來產生「過濾器」節點。

標示要包含在圖表上的預測值（如果適用），然後從功能表中選擇：

產生 > **「過濾器」節點（預測值重要性）**

OR

> **欄位選擇（預測值重要性）**

變數數上限。 包含或排除到達指定數的最重要預測值。

重要性大於。 包含或排除相對重要性大於指定值的所有預測值。

集合檢視器

集合的模型

集合模式提供關於集合中的成分模型與集合效能的整體資訊。

您可以使用主要（獨立檢視）工具列來選擇是否使用集合或參考模型來評分。若您使用集合來評分，則也可以選取合併規則。上述變更不要求重新執行模式；不過，系統會將這些選擇儲存至模式 (項目)，以作為評分和/或下游模式評估用途。其也會影響從集合檢視器匯出的 PMML。

合併規則。系統執行集合評分時，可使用此規則來合併基底模型的預測值，以計算集合分數值。

- 您可以使用投票、最高機率或最高平均數機率來合併**類別**目標的集合預測值。**投票**會選取所有基底模式中最常擁有最高機率的類別。**最高機率**會在所有基底模式中選取達到單一最高機率的類別。**最高平均數機率**會在平均計算所有基底模型的類別機率時，選取具有最高值的類別。
- 系統會使用基底模型的平均數或中位數，來合併**連續**目標的集合預測值。

模型建置期間會從規格當中擷取預設值。變更合併規則時會重新計算模式準確性，並且更新所有模式準確性的檢視。此外也會更新「預測值重要性」圖表。若已選取用來評分的參考模式，則會停用此控制項。

顯示所有合併規則。 若選取此項目，則會在模式品質圖表中顯示所有可用合併規則的結果。此外也會更新「成分模型準確性」圖表，以顯示每個投票方法的參考行。

模型摘要： 「模型摘要」視圖是一種Snapshot，集合品質和差異的一覽摘要。

品質。此圖表會顯示與參考模型及單純模型比較過後的最終模型準確性。準確性越大，則模式越佳；「最佳」模式會擁有最高準確性。對於類別目標而言，準確性等於預測值與觀察值的相符記錄百分比。對於連續目標而言，準確性等於將 1 減去預測平均絕對誤差（預測值減去觀察值所得出的平均絕對值）與預測值範圍（最大預測值減去最小預測值）的比率。

對於 Bagging 集合而言，參考模式是內建於整體訓練區隔的標準模式。對於 Boosted 集合而言，參考模式是第一個成分模型。

單純模式代表未建立任何模式時的準確性，並將所有記錄指定至典型類別。系統不會計算連續目標的單純模式。

差異。此圖表會顯示用於建立集合之成分模型之間的「意見差異」，以越大代表差異越大的格式表示。這是一種所有基底模式之間預測差異的測量。Boosted 集合模式無法使用差異，此外也不會顯示連續目標的差異。

預測值重要性： 一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

所有集合模式均無法使用預測值重要性。所有成分模型的預測值集可能會有所差異，但系統可以針對至少一個用於成分模型的預測值來計算其重要性。

預測值次數： 由於選擇的建模方法或預測值選擇不相同，因此所有成分模型的預測值集也會有所差異。「預測值次數」圖是一種顯示集合中所有成分模型預測值分配狀況的點形圖。每個點代表一或多個內含預測值的成分模型。預測值會畫在 y 軸上，並依遞減順序排列次數；因此位於最頂端的是最多成分模式使用的預測值，而最底端的是最少成分模型使用的預測值。系統會顯示前 10 個預測值。

一般而言，最常出現的預測值代表其重要性最高。對於無法在所有成分模型中採用不同預測值集的方法，則不適用於此圖。

成分模型準確性： 此圖表是一種成分模型預測準確性的點形圖。每個點代表一或多個在 y 軸上繪有準確度等級的成分模型。將滑鼠游標移到任一點上，即可取得個別對應成分模型的相關資訊。

參考線。圖上會顯示經過彩色編碼的集合線條，以及參照模型和單純模型。在用於評分之模型的對應線旁，會顯示核取記號。

互動性。如果變更合併規則，便會更新此圖表。

Boosted 集合。系統顯示的 Boosted 集合折線圖。

成分模型詳細資料： 此表格會顯示成分模型的資訊（依列排列）。依預設，系統會以遞增的模式編號順序來排序成分模型。您可以依照任何欄值，採用遞增或遞減順序來排序列。

模式。一個代表成分模型建立順序的數字。

準確性。採用百分比格式的整體準確性。

方法。建模方法。

預測值。用於成分模型的預測值數量。

模式大小。模型大小視建模方法而定：對於樹狀結構而言，模型大小是樹狀結構中的節點數；對線性模型而言，模型大小是係數個數；對於神經網路而言，模型大小是接合處的數目。

記錄。訓練樣本中的輸入記錄加權數。

自動資料準備：

此檢視會顯示排除的欄位，以及在自動式資料準備 (ADP) 步驟中衍生轉換欄位方式的相關資訊。針對每個已轉換或排除的欄位，此表格會列出欄位名稱、分析當中的欄位角色，以及 ADP 步驟所採取的動作。系統會依照欄位名稱的字母順序以遞增方式來排序欄位。

若顯示「**刪除離群值**」動作，則表示已將超過分割值 (平均數的 3 個標準差) 的連續預測值設為分割值。

分割模型的模型塊

分割模型的模型塊可以存取分割建立的所有個別模型。

分割模型塊包含：

- 建立的所有分割模型清單，連同每個模型的統計資料集合
- 有關整體模型的資訊

從分割模型清單中，您可以開啟單個模型以進一步檢查。

分割模型檢視器

「模型」標籤列出塊中包含的所有模型，以各種形式提供有關分割模型的統計資料。它有兩種一般形式，具體取決於建模節點。

排序依據。使用此清單選擇列出模型的順序。您可以根據任何顯示直欄的值將清單按遞增或遞減排序。或者，按一下欄標題，按該欄將清單排序。預設是總精確性的遞減。

顯示/隱藏欄功能表。按一下此按鈕，以顯示功能表，以便選擇單個欄以顯示或隱藏。

檢視。如果您正在使用分區，您可以選擇檢視訓練資料或檢定資料的結果。

對於每個分割，詳細資料顯示如下：

圖表。指示此模型資料分佈的縮圖。當塊位於畫布中時，按兩下縮圖開啟完整大小的圖表。

模型。模型類型圖示。按兩下圖示開啟此特定分割的模型塊。

分割欄位。建模節點中指定為分割欄位的欄位及其各個可能值。

分割中記錄數。此特定分割中涉及的記錄數。

使用的欄位數。根據所使用的輸入欄位數對分割模型評級。

總準確度 (%)。分割模型正確預測的記錄數佔該分割中記錄總數的百分比。

分割。直欄標題會顯示用於建立分割的欄位，而儲存格即為分割值。按兩下任何分割，即會開啟「模型檢視器」以執行該分割的模型建置。

準確性。採用百分比格式的整體準確性。

模式大小。模型大小視建模方法而定：對於樹狀結構而言，模型大小是樹狀結構中的節點數；對線性模型而言，模型大小是係數個數；對於神經網路而言，模型大小是接合處的數目。

記錄。訓練樣本中的輸入記錄加權數。

使用串流中的模型塊

模型塊置於串流中，允許您對新資料進行評分並產生新節點。通過對資料**進行評分**，您可以使用通過模型建置獲得的資訊來為新記錄建立預測。如需評分結果，需要為模型塊添加終端節點（即正在處理或輸出節點）並執行終端節點。

對於某些模型而言，還可從模型塊中獲得有關預測品質的其他資訊，例如信賴度值或到叢集中心的距離。通過產生新節點，您可以輕鬆地根據產生的模型的結構來建立新節點。例如，您可以根據執行輸入欄位選擇的多數模型產生「過濾器」節點，此節點僅傳送模型 ID 為「重要」的輸入欄位。

註：在 IBM SPSS Modeler 的不同版本中執行時，給定模型為給定觀察值指定的分數可能會有細小差別。這通常是由於各個版本之間的軟體增強所致。

使用模型塊對資料進行評分

1. 將模型塊連接到向其傳送資料的資料來源或串流。
2. 將一個或多個正在處理或輸出節點（如表格或分析節點）新增或連接到模型塊。
3. 執行模型塊中的某個下游節點。

註：您無法使用「未優化規則」節點對資料進行評分。要根據關聯規則模型對資料進行評分，請使用「未優化規則」節點產生「規則集」模型塊，然後使用「規則集」模型塊進行評分。請參閱第 227 頁的『從關聯模型塊產生規則集』主題，以取得更多資訊。

使用模型塊產生正在處理節點

1. 在此選用區中瀏覽模型，或者在串流畫布中編輯模型。
2. 在「模型塊瀏覽器」視窗的「產生」功能表中選取所需節點類型。可用選項將因模型塊類型的不同而有所不同。請參閱特定的模型區塊類型以瞭解您可以從特定模型產生之項目的相關詳細資料。

重新產生建模節點

如果要修改或更新某個模型塊，但無法使用用於建立該模型的串流，那麼可以使用與建立原始模型相同的選項來重新產生一個建模節點。

要重新建立模型，用滑鼠右鍵按一下模型選用區中的模型，然後選擇**產生建模節點**。

此外，當瀏覽模型時，請選擇「產生」功能表中的**產生建模節點**。

多數情況下，重新產生的建模節點應與建立原始模型的建模節點在功能上一致。

- 對決策樹模型而言，還可以將互動式階段作業過程中的其他設定儲存到節點，重新產生建模節點的過程中將啟用**使用樹狀結構型指引**選項。
- 對於決策清單模型而言，將啟用**使用儲存的交互階段作業資訊**選項。請參閱第 130 頁的『決策清單模型選項』主題，以取得更多資訊。
- 對於「時間序列」模型，將啟用**使用現有模型繼續估計**選項，通過該選項您可以使用現行資料重新產生先前的模型。請參閱「時間序列」模型選項主題，以取得更多資訊。

將模型匯入和匯出為 PMML

PMML（也稱為預測模型標記語言）是一種 XML 格式，用於描述資料採礦和統計模型，包含模型的輸入、用於為資料採礦準備資料的轉換，以及定義模型自身的參數。IBM SPSS Modeler 可匯入並匯出 PMML，這使得其能夠與其他支援此格式的應用程式（如 IBM SPSS Statistics）共用模型。

有關 PMML 的詳細資訊，請參閱資料採礦群組網站 (*http://www.dmg.org*)。

匯出模型

PMML 匯出受支援大多數模型類型，這些模型類型產生在 IBM SPSS Modeler 中。請參閱『支援 PMML 的模型類型』主題，以取得更多資訊。

1. 用滑鼠右鍵按一下模型選用區上的模型塊。（或者，按兩下畫布上的模型塊並選取「檔案」功能表。）
2. 在功能表上，按一下**匯出 PMML**。
3. 在「匯出」（或「儲存」）對話框中，指定此模型的目標目錄及唯一名稱。

註：

您可在「使用者選項」對話框中變更 PMML 匯出的選項。在主功能表上，按一下：

工具 > **選項** > **使用者選項**

然後按一下 PMML 標籤。

匯入儲存為 PMML 的模型

以 PMML 格式從 IBM SPSS Modeler 或其他應用程式中匯出的模型可以匯入到模型選用區中。請參閱『支援 PMML 的模型類型』主題，以取得更多資訊。

1. 在模型選用區上，用滑鼠右鍵按一下選用區並從功能表中選取**匯入 PMML**。
2. 選取要匯入的檔案並根據需要為變數標籤指定選項。
3. 按一下「**開啟**」。

使用變數標籤（如果模型中出現這些標籤）。PMML 可為資料字典中的變數同時指定變數名稱和變數標籤（例如，Referrer ID，簡稱 *RefID*）。如果在最初匯出的 PMML 中出現變數標籤，則選中此選項可以使用這些變數標籤。

如果已選取變數標籤選項但在 PMML 中沒有變數標籤，則按常態使用變數名稱。

支援 PMML 的模型類型

PMML 匯出

IBM SPSS Modeler 模型。在 IBM SPSS Modeler 中建立的下列模型都可匯出為 PMML 4.0 格式：

- C&R 樹狀結構
- QUEST
- CHAID
- 神經網路
- C5.0
- 邏輯迴歸
- Genlin
- SVM
- Apriori
- Carma
- K-Means
- Kohonen
- TwoStep

- TwoStep-AS
- GLMM（針對所有 GMML 模型匯出 PLMM，但 PMML 僅具有固定效應）
- 決策清單
- Cox
- 序列（不支援序列 PMML 模型評分）
- 隨機樹狀結構
- Tree-AS
- 線性
- Linear-AS
- 迴歸方法(R)
- 羅吉斯
- GLE
- LSVM
- 異常偵測
- KNN
- 關聯規則

資料庫原生模型。對於使用資料庫原生演算法產生的模型，PMML 匯出不可用. 無法匯出使用 Microsoft 的 Analysis Services 或 Oracle Data Miner 建立的模型。

PMML 匯入

IBM SPSS Modeler 可以匯入並分數由所有 IBM SPSS Statistics 產品的目前版本產生的 PMML 模型，包括從 IBM SPSS Modeler 匯出的模型和由 IBM SPSS Statistics 17.0 或以後版本產生的模型或變換 PMML。這實質上意味著評分引擎可評分的任何 PMML，下列除外：

- 無法匯入 Apriori、CARMA、異常偵測、序列和關聯規則模型。
- 將 PMML 模型匯入到 IBM SPSS Modeler 中後，雖然可以對其進行評分，但不能進行瀏覽。（注意，其中包含最初從 IBM SPSS Modeler 中匯出的模型。為避免此限制，可將模型按產生的模型檔 (`*.gm`) 匯出而不是按 PMML 匯出。）
- 在匯入時會執行有限的驗證，但在試圖對模型分數時會執行全面驗證。因此有可能匯入成功，但評分卻失敗或產生不正確的結果。

註：對於匯入到 IBM SPSS Modeler 中的第三方 PMML，IBM SPSS Modeler 將試圖對可以識別並進行評分的有效 PMML 進行評分。但是，無法保證將對所有 PMML 進行評分，也無法保證以應用程式產生 PMML 的方式對 PMML 進行評分。

為評分配接器發行模型

您可以將模型發行到安裝有評分配接器的資料庫伺服器。評分配接器可通過使用資料庫的使用者定義函數 (UDF) 功能在資料庫中進行模型評分。在資料庫中進行評分可免除在評分之前擷取資料。發行到評分配接器也將產生一些範例 SQL 以執行 UDF。

發行評分配接器

1. 按兩下模型塊將其開啟。
2. 從模型塊功能表中選擇：

檔案 > **為伺服器評分配接器發行**

3. 填寫對話框中的相關欄位，然後按一下**確定**。

資料庫連線。要為模型使用的資料庫的連線細節。

發行 ID。（僅限 Db2 for z/OS 資料庫）模型的 ID。如果您重新建立同一模型並使用相同發行 ID，那麼產生的 SQL 也保持不變，所以無需變更使用之前產生的 SQL 的應用程式即可重新建立模型。（對於其他資料庫，產生的 SQL 對模型則是唯一。）

產生範例 SQL。如果選取此項，將在**檔案**欄位中指定的檔案中產生範例 SQL。

未優化模型

未優化模型包含從資料中擷取的資訊，但並不用於直接產生預測。即這些模型不能新增至串流。未優化模型在「產生的模型」選用區上顯示為「未打磨的鑽石」。

圖 *27.* 未優化模型的圖示

如需未優化規則模型的詳細資訊，用滑鼠右鍵按一下模型，然後選擇快速功能表中的 **瀏覽** 。像其他在 IBM SPSS Modeler 中產生的模型一樣，各種標籤將提供所建立模型的相關概要和規則資訊。

產生節點。 「產生」功能表允許您基於規則建立新節點。

- **選取節點。** 產生「選取」節點以選取要對其套用目前選定規則的記錄。如果未選取任何規則，則停用此選項。
- **規則集。**產生「規則集」節點以預測單個目標欄位的值。請參閱第 227 頁的『從關聯模型塊產生規則集』主題，以取得更多資訊。

第 4 章 篩選模型

篩選欄位和記錄

分析的預備階段中可以使用多個建模節點來查找對建模最有用的欄位和記錄。可使用功能選擇節點來按照重要性篩選欄位並為之排等級，以及使用異常偵測節點來查找不符合「正常」資料已知型樣的不正常記錄。

功能選擇節點會根據某組準則（例如遺漏百分比）篩選可移除的輸入欄位，對於保留的輸入，隨後將對其相對於指定目標的重要性進行排等級。例如，假如某個給定資料集有上千個潛在輸入，那麼哪些輸入最有可能用於對病患結果進行建模呢？

Anomaly Detection 節點確定不符合「正常」資料格式的異常觀察值（離群值）。即使離群值不匹配任何已知格式或用戶不清楚自己的查找目標，也可以使用此節點來確定離群值。

請注意，異常偵測會透過叢集分析，根據模型中選取的欄位集來識別不尋常的記錄或觀察值，而不會考慮任何特定的目標（相依）欄位，也不管那些欄位與您嘗試預測的型樣是否相關。鑑於此，您可能想要結合使用異常偵測與功能選擇或用來對欄位進行篩選和分級的其他技術。例如，您可以使用功能選擇來識別相對於特定目標的最重要欄位，然後使用異常偵測來尋找對那些欄位而言最不尋常的記錄。（另一種方法是建置決策樹模型，然後檢查任何錯誤分類的記錄作為潛在的異常。但是，使用此方法難以大規模進行抄寫或自動化。）

功能選擇節點

資料採礦問題可能包括成百甚至上千個可用作輸入的備選欄位。從而花費大量的時間和精力來檢查模型究竟應該包含哪些欄位或變數。為了縮小選擇範圍，可以使用功能選擇演算法來識別對某給定分析最為重要的欄位。例如，如果你試著根據多種因素來預測病患結果，那麼哪些因素最為重要呢？

功能選擇由以下三個步驟組成：

- **篩選。**刪除不重要或有問題的輸入、記錄或觀察值（例如輸入欄位含有過多遺漏值，或者輸入欄位的變異太大或太少而變得無用）。
- **分等級。** 對剩餘輸入進行排序並根據重要性進行分級。
- **選取。**確定要在後續模型中使用的功能子集合，例如通過僅保留最重要的輸入以及過濾或排除所有其他項目輸入來進行確定。

當下，多數組織的資料均已超載，因此簡化和加快建模過程是功能選擇的根本優勢。通過將注意力迅速集中到最重要的欄位上，可以降低所需的計算數量，並且可以方便地找到因某種原因被忽略的小而重要的關係，最終獲得更簡單、精確和易於解釋的模型。通過減少模型中的欄位個數數量，可以減少評分時間以及未來疊代中所收集的資料數量。

範例。 某電話公司有一個資料倉儲，其中包含 5000 名公司客戶對某次特別促銷活動的回應的相關資訊。資料包含有客戶年齡、職業、收入、電話使用情況的統計資料資料等大量資料。三個目標欄位表示客戶是否對三個報價做出回應。該公司希望使用這些資料來說明預測哪些客戶最有可能在將來對類似報價做出回應。

需求。 單個目標欄位（其角色設定為目標），以及要根據目標進行篩選或排等級的多個輸入欄位。目標和輸入欄位均具有連續（數值型範圍）或種類的測量層次。

功能選項模式設定

「模型」標籤上的設定包含標準模型選項以及用於對輸入欄位篩選準則進行微調的設定。

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。

篩選輸入欄位

篩選就是剔除不提供關於輸入/目標關係的任何有用資訊的輸入或觀察值。篩選選項只依據在問題中使用欄位的屬性，而不遵循該欄位針對於選定目標欄位的預測能力。被篩選出來的欄位將不參與有關輸入排等級的計算，同時還選用擇將這些欄位過濾掉，或是從用於建模的資料中刪除。

可根據下列準則篩選欄位：

- **遺漏值的上限百分比。**篩選具有過多遺漏值的欄位，以佔記錄總數的百分比表示。遺漏值百分比大的欄位幾乎不提供任何預測資訊。
- **單個種類中的記錄的上限百分比。**篩選相對於記錄總數而言同一種類中具有過多記錄的欄位。例如，如果資料庫中 95% 的客戶開啟同一類型的車，那麼此資訊無助於識別客戶。任何超過指定最大值的欄位都將被篩選掉。此選項僅適用於類別欄位。
- **以記錄的百分比表示的種類數目上限。**篩選相對於記錄總數而言具有過多種類的欄位。如果很高百分比個種類只含有一個觀察值，那麼該欄位用處有限。例如，如果每名客戶都戴不同的帽子，那麼此資訊在建立行為型樣模型時就不太可能有用。此選項僅適用於類別欄位。
- **下限變異係數。**篩選變異係數小於或等於指定最小值的欄位。此測量值是輸入欄位標準離差與輸入欄位平均數之間的比值。如果此值接近 0，那麼變數值的變異性就不高。此選項僅適用於連續（數值範圍）欄位。
- **最小標準偏差。**篩選標準差小於或等於指定最小值的欄位。此選項僅適用於連續（數值範圍）欄位。

包含遺漏資料的記錄。目標欄位具有遺漏值或所有輸入都具有遺漏值的記錄或觀察值將被從用於分等級的計算式中排除。

功能選擇選項

「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。

請參閱第 47 頁的『功能選項模式結果』主題，以取得更多資訊。

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。

重要性分等級選項

所有種類。當所有輸入和目標均為種類字段時，可以根據以下任何一個測量對重要性進行排等級：

- **Pearson 卡方。**無需現有關係的強度或方向即可測試目標和輸入的獨立性。
- **概似比卡方。**與 Pearson 卡方類似，也用於測試目標 - 輸入的獨立性。
- **Cramer's V。**基於 Pearson 卡方測試統計資料的關聯的測量。值範圍為 0 到 1，0 表示無關聯，1 表示完全關聯。
- **Lambda (λ)。**這是反映變數用於預測目標值時誤降低比例的關聯的測量。值為 1 表示輸入欄位完美地預測了目標，值為 0 則表示輸入未提供目標的任何有用資訊。

部分種類。當部分但並非所有輸入為種類字段且目標也為種類字段時，可以根據 Pearson 或概似比卡方對重要性進行排等級。（除非所有輸入均為種類變量，否則 Cramer's *V* 和 lambda 均無法使用。）

種類與連續。針對連續目標來為種類輸入分等級或與之相反的情形時（即其中之一為種類字段，但不能兩者均為種類字段），則使用 *F* 統計資料。

兩者均為連續字段。針對連續目標來為連續輸入分等級時，將使用基於相關係數的 *t* 統計資料。

功能選項模式塊

「功能選擇」模型塊顯示每個輸入相對於選定目標的重要性（遵循「功能選擇」節點的排等級）。分等級前已篩選掉的所有欄位也將被列出。請參閱第 45 頁的『功能選擇節點』主題，以取得更多資訊。

執行含有特「功能選擇」型塊的串流時，模型行為將如同過濾器，僅保留「模型」標籤上目前已選取的輸入。例如，可以選取評定為「重要」的所有欄位（預設選項之一）或在「模型」標籤上手動選取一個欄位子集合。不管如何選擇，目標欄位總是被保留。所有其他欄位將被排除。

過濾僅基於欄位名稱；例如，如果選取年齡和收入，那麼相符其中一個名稱的任何欄位都將被保留。該模型不是基於新資料更新欄位分等級，而只是根據選定的名稱來過濾欄位。所以，將模型套用到新的或更新過的資料時應多加注意。存有疑問時，最好重新產生模型。

功能選項模式結果

「功能選擇」模型塊的「模型」標籤在頂部窗格中顯示所有輸入的排等級和重要性，並且使您可以通過左側欄中的勾選框來選擇用於過濾的欄位。執行串流時，將只保留選定的欄位；其他欄位將被捨棄。預設選擇是基於模型建立節點中指定的選項，但可以根據需要選擇或棄選其他欄位。

底部窗格列出依據遺漏值百分比或建模節點中指定的其他準則而從分等級中排除的輸入。與其他排等級欄位一樣，可以通過左欄勾選框來選擇包含或捨棄這些欄位。請參閱第 46 頁的『功能選項模式設定』主題，以取得更多資訊。

- 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。如果要使用工具列，那麼可以從「排序依據」清單選取需要的項目，並使用「向上」和「向下」箭頭來變更排序方向。
- 您可以使用工具列來選中或取消勾選所有欄位以及存取「檢查欄位」對話框，您可以通過該對話框根據排等級或重要性來選取欄位。也可以按住 Shift 和 Ctrl 鍵並按一下欄位，以選擇更多的欄位，並使用空格鍵來切換選定的欄位群組。請參閱第 48 頁的『按照重要性選取欄位』主題，以取得更多資訊。
- 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。請參閱第 46 頁的『功能選擇選項』主題，以取得更多資訊。

按照重要性選取欄位

使用功能選項模式塊對資料進行評分時，由排等級或篩選欄位已選取的所有欄位都將被保留，如左欄勾選框所示。其他欄位將被捨棄。要變更選擇，您可以使用工具列存取「檢查欄位」對話框，該對話框使您可以根據排等級或重要性來選取欄位。

標示的所有欄位。 選取標示為重要、一般或不重要的所有欄位。

欄位數目上限。 可讓您根據重要性選取前 *n* 個欄位。

重要性大於。 選取重要性大於指定臨界值的所有欄位。

從功能選項模式中產生過濾器

根據「功能選擇」模型的結果，您可以使用「根據特徵產生過濾」對話框來產生一個或多個「過濾器」節點，該節點根據相對於指定目標的重要性包含或排除欄位子集合。雖然模型塊也可以用於過濾，但使用此方法可以在不複製或不修改模型的情況下自由地嘗試不同的欄位子集合。不管是選取包含還是選取排除，過濾時將總是保留目標欄位。

包含/排除。您可以選擇包含或排除欄位，例如包含前 10 個欄位或排除所有標示為「不重要」的欄位。

選取的欄位。 包含或排除表格中目前選定的所有欄位。

標示的所有欄位。 選取標示為重要、一般或不重要的所有欄位。

欄位數目上限。 可讓您根據重要性選取前 *n* 個欄位。

重要性大於。 選取重要性大於指定臨界值的所有欄位。

異常偵測節點

異常偵測模型用於識別資料中的離群值或異常觀察值。與儲存有關異常觀察值的規則的其他建模方法不同，異常偵測模型儲存有關正常行為的資訊。因此即使在離群值不符合任何已知型樣的情況下，異常偵測模型也使識別離群值成為可能，在新型樣可能不斷湧現的應用（如缺陷偵測）中，該模型可能尤其有用。異常偵測是一種不受監督的方法，這就意味著它不需要包含已知缺陷觀察值的訓練資料集作為起點。

識別離群值的傳統方法通常是一次檢查一個或兩個變數，而異常偵測可以檢查大量欄位以識別相似記錄所屬的叢集或對等群組。然後，可將每條記錄與其對等群組中的其他記錄進行比較，以識別出可能的異常值。觀察值與正常中心值離得越遠，它越有可能是異常觀察值。例如，該演算法可能會將記錄聚合為三個不同的叢集，並對離任何一個叢集的中心值較遠的那些記錄進行旗標。

每條記錄都指定了一個異常指數，該指數是群組離差指數與該觀察值所屬叢集中平均值的比。此指數的值越大，觀察值與平均值的離差就越大。通常情況下，異常指數值少於 1 甚至少於 1.5 的觀察值都不會被視為異常值，因為該離差與平均值相同或者只是大一點。但是，指數值大於 2 的觀察值有可能是異常觀察值，因為該離差至少是平均值的兩倍。

異常偵測是一種探索性方法，它是為對應該進行進一步分析的可能異常觀察值或記錄進行快速偵測而設計的。這些觀測值應視為 疑似 異常值，在進行進一步檢查後，可以證明它們是或不是真正的異常值。您可能會發現某個記錄完全有效，但無法選擇從資料中將其篩選出來用於模型建置。另外，如果演算法重複檢測出虛假異常值，那麼可能表示資料收集過程中存在錯誤或假象。

請注意，異常偵測會透過叢集分析，根據模型中選取的欄位集來識別不尋常的記錄或觀察值，而不會考慮任何特定的目標（相依）欄位，也不管那些欄位與您嘗試預測的型樣是否相關。鑑於此，您可能想要結合使用異常

偵測與功能選擇或用來對欄位進行篩選和分級的其他技術。例如，您可以使用功能選擇來識別相對於特定目標的最重要欄位，然後使用異常偵測來尋找對那些欄位而言最不尋常的記錄。（另一種方法是建置決策樹模型，然後檢查任何錯誤分類的記錄作為潛在的異常。但是，使用此方法難以大規模進行抄寫或自動化。）

範例。 對農業發展補貼進行審查以確定是否可能存在內部欺詐觀察值時，異常偵測可用於探索有悖於標準值的離差，並強調顯示值得進一步調查的異常記錄。特別值得關注的是那些相對農場類型和規模而言似乎申請了過多（或過少）補助金的補貼申請。

需求。 一個或多個輸入欄位。請注意，只有其角色使用來源節點或「類型」節點設定為**輸入**的欄位才能用作輸入。目標欄位（角色設定為**目標**或**兩者**）將被忽略。

強度。 通過旗標不符合已知規則集（而不是符合已知規則集）的觀察值，異常偵測模型可以識別異常觀察值，即使它們未遵循先前已知的型樣也是如此。與功能選擇結合使用時，異常偵測可用於篩選大數量資料，以便相對較快地識別最受關注的記錄。

異常偵測模型選項

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。

確定截斷值以用於異常標記的根據。指定用於確定截斷值以旗標異常的方法。您可以使用的選項如下：

- **最低異常指數層次。**指定用來標示例外狀況的最小截斷值。達到或超過此臨界值的記錄將進行旗標。
- **訓練資料中最異常記錄所佔百分比。**自動設定某一層次的臨界值，以標示訓練資料中記錄的指定百分比。所生成的分割值作為參數包含在模型中。請注意，此選項確定了截斷值的設定方式，而並非確定評分期間要旗標的記錄的實際百分比。實際評分結果可能隨資料而異。
- **訓練資料中最異常記錄的號碼。**自動設定某一層次的臨界值，以標示訓練資料中的指定記錄數。所生成的臨界值作為參數包含在模型中。請注意，此選項確定了截斷值的設定方式，而並非確定評分期間要旗標的記錄的具體號碼。實際評分結果可能隨資料而異。

註：無論如何確定截斷值，這都不會影響對每條記錄報告的潛在異常指數值。它僅在對模型進行估算和評分時指定用於將記錄旗標為異常的臨界值。如果您稍後要檢查更多或更少記錄，那麼可以使用「選取」節點來根據異常指數值 (`$O-AnomalyIndex > X`) 識別記錄子集。

要報告的異常欄位個數。指定要報告的欄位個數，用於指示將特定記錄標示為異常的原因。會報告大部分異常欄位，定義為顯示為其指派記錄之叢集的最大欄位標準偏差的欄位。

異常偵測專家選項

要指定遺漏值和其他設定的選項，請在「專家」標籤上將模式設定為**專家**。

調整係數。用於平衡計算距離時指定給連續（數值範圍）和種類欄位的相對加權的值。值越大，對連續欄位的影響也越大。它必須為非零值。

自動計算對等群組數目。異常偵測可用於快速分析大量可行的解決方案，以選擇訓練資料的最佳對等群組數目。可通過設定對等節點群組的數目下限和上限來擴大或縮小範圍。較大值將使系統可以探索更大範圍的可行解決方案，但相應代價是處理時間增加。

指定對等群組數。如果您知道要在模型中包含的叢集數，請選取此選項並輸入對等群組數目。通常，選取此選項可提高效能。

雜訊等級和比例。這些設定用於確定二階叢集期間離群值的處理方式。在第一階段中，使用叢集特性 (CF) 樹狀結構將資料從大量個別記錄壓縮成可管理數量的叢集。該樹狀結構基於親緣性測量構建，並且當樹狀結構的

某個節點中記錄過多時，它會分割為子節點。在第二階段中，將從 CF 樹狀結構的終端節點開始創建階層式叢集。雜訊處理在第一次資料傳送時開啟，並在第二次資料傳送時關閉。第一次資料傳送時，雜訊叢集中的觀察值將分配給第二次資料傳送中的一般叢集。

- **雜訊等級。**指定介於 0 到 0.5 之間的值。只有在下列情況下此設定才相關：CF 樹狀結構在成長階段填滿，即該樹狀結構不再接收葉節點中的觀察值並且葉節點無法分割。

  如果 CF 樹狀結構進行填充並且雜訊等級設定為 0，那麼臨界值將增大並且 CF 樹狀結構將使用所有觀察值重新增長。在最終叢集之後，無法指定到叢集的值就會標示為偏離值。將對離群值叢集指定識別號 -1。離群值叢集不併入在叢集數目的計數中；即，如果您指定 *n* 個叢集和雜訊處理，那麼演算法將輸出 *n* 個叢集和一個雜訊叢集。實際上，增大此值可使演算法在將異常記錄納入樹狀結構中時有更大餘地，而不是將它們分配給個別的離群值叢集。

  如果 CF 樹狀結構進行填充並且雜訊等級大於 0，那麼在將稀疏葉節點中的任何資料放入其自身的雜訊葉節點中之後，該 CF 樹狀結構會重新增長。如果稀疏葉節點中的觀察值數與最大葉節點中的觀察值數的比例少於雜訊等級，那麼認為該葉節點是稀疏葉節點。在樹狀結構增長完成後，系統會在可能的情況下將離群值放入 CF 樹狀結構中。如果未放入 CF 樹中，那麼對於第二階段叢集，將捨棄離群值。
- **噪音比率。**指定為元件配置的應用於噪音緩衝的記憶體部分。此值必須介於 0.0 到 0.5 之間。如果將特定觀察值插入樹狀結構的葉節點中之後，所產生的緊性少於臨界值，那麼葉節點將不再分割。如果緊性超過臨界值，那麼葉節點將進行分割，同時將另一個小叢集新增至 CF 樹狀結構。實際上，增大此設定可能會導致演算法更快速地向較簡單的樹狀結構傾斜。

插補遺漏值。 對於連續欄位，請用欄位平均數替換遺漏值。對於種類欄位，多個遺漏值種類將進行已結合併被視為一個有效種類。如果取消選取此選項，那麼將從分析中排除任何具有遺漏值的記錄。

異常偵測模型塊

異常偵測模型塊包含異常偵測模型所擷取的所有資訊以及有關訓練資料和估計過程的資訊。

執行包含異常偵測模型塊的串流時，多個新欄位將按照模型塊中「設定」標籤上的選擇新增至串流。請參閱第 51 頁的『異常偵測模型設定』主題，以取得更多資訊。新的欄位名稱基於模型名稱，並帶有前綴 *$O*，下表格對這些名稱進行了概述。

表 *6.* 新欄位名稱產生

欄位名稱	說明
$O-Anomaly	指示記錄是否異常的旗標欄位。
$O-AnomalyIndex	記錄的異常指數值。
$O-PeerGroup	指定記錄分配給哪個對等群組。
$O-Field-n	與叢集標準值離差相關的第 *n* 個最異常欄位的名稱。
$O-FieldImpact-n	欄位的變數離差指數。此值用於測量與記錄分配到的叢集欄位標準值的離差。

也可以選擇抑止非異常記錄的分數，以使結果更易於讀取。請參閱第 51 頁的『異常偵測模型設定』主題，以取得更多資訊。

異常偵測模式詳細資料

所產生的異常偵測模型的「模型」標籤顯示模型中對等群組的相關資訊。

請注意，所報告的對等群組大小和統計資料是基於訓練資料的估計值，並且可能與實際評分結果略有不同，即使對相同資料執行也是如此。

異常偵測模型摘要

異常偵測模型塊的「摘要」標籤顯示欄位、建構設定值和估計過程的相關資訊。另外，還顯示了對等群組數目以及用於將記錄旗標為異常的截斷值。

異常偵測模型設定

使用「設定」標籤可以指定用於對模型塊進行評分的選項。

異常記錄指示方法 指定在輸出中處理異常記錄的方式。

- **旗標和指標** 建立旗標欄位，對於模型中包含的所有超過截斷值的記錄，此欄位設定為 *True*。另外，將報告另一個欄位中每條記錄的異常指數。請參閱第 49 頁的『異常偵測模型選項』主題，以取得更多資訊。
- **僅旗標** 建立旗標欄位，但不報告每條記錄的異常指數。
- **僅指標** 報告異常指數但不建立旗標欄位。

要報告的異常欄位數量：指定要報告的欄位個數，用於指示將特定記錄標示為異常的原因。會報告大部分異常欄位，定義為顯示為其指派記錄之叢集的最大欄位標準偏差的欄位。

捨棄記錄 選取此選項可捨棄串流中的所有**非異常**記錄，從而更容易地專註於任何下游節點中的潛在異常。另外，您也可以捨棄所有**異常**記錄，以便將後續分析限制為那些未根據模型旗標為潛在異常的記錄。

註：由於取整造成的細微差異，評分期間旗標的實際記錄數可能與訓練模型時旗標的記錄數不同，即使對相同資料執行也是如此。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

第 5 章 自動建模節點

自動建模節點對多種不同的建模方法進行估計和比較，這使您可以在一次建模執行中嘗試多種方法。您可以選取所使用的建模演算法，以及每個建模演算法的具體選項，包含可能互斥的組合。例如，您無需為類神經網路選擇快速、動態或刪改之中的某個方式，完全可以全部嘗試。該節點會探索每個可能的選項組合，根據您指定的測量為每個候選模型評級，並儲存最佳實務以用於評分蘗或未來分析。

您可以根據分析需要從三個自動建模節點中進行選擇：

「自動分類器」節點用於建立和對比二元結果（是或否，流失或不流失等）的若干不同模型，使用戶可以選擇給定分析的最佳處理方法。由於受支援 多種建模演算法，因此可以對用戶希望使用的方法、每種方法的特定選項以及對比結果的準則進行選取。節點根據指定的選項產生一組模型並根據用戶指定的準則排列最佳候選項的順等級。

自動數值節點使用多種不同方法估計和對比模型的連續數值範圍結果。此節點和自動分類器節點的工作方式相同，因此可以選擇要使用和要在單個建模傳送中使用多個選項組合進行測試的演算法。受支援的演算法包含神經網路、C&R 樹狀結構、CHAID、線性迴歸、通用性線性迴歸以及受支援向量機器 (SVM)。可基於相關係數度、相對錯誤或已用變數數目對模型進行對比。

自動叢集節點估計和比較識別具有類似特性記錄群組的叢集模型。節點工作方式與其他自動建模節點相同，使您在一次建模運行中即可試驗多個選項組合。可以使用基本測量比較模型，透過測量嘗試過濾並分級叢集模型的實用性，並基於特定欄位的重要性提供測量。

最佳模型儲存在一個複合模型塊中，可對其進行瀏覽和比較，並選擇評分中使用的模型。

- 只有對於二元、列名和數值目標，您才可以選取多個評分模型，並將評分組合在一個模型總體中。透過結合來自多個模型的預測，可避免個別模型中存在的限制，通常會導致整體精確度高於可從任何一個模型獲取的精確度。
- 您還可以選擇往下探查結果，並為要使用或進一步探索的所有個別模型產生建模節點或模型塊。

模型和執行時間

根據資料集和模型的數量，自動建模節點執行時間可能為數小時或甚至更長。在選取選項時，請注意正在生成的模型個數。如果現實條件允許，您可能希望將建模執行的時間安排在夜晚或週末，因為此時對系統資源的需求可能比較小。

- 必要的話，可以使用分割區節點或樣本節點減少併入在初始訓練傳送中的記錄數。一旦將選擇限制在幾個生成的候選模型內，就可以還原全部資料集。
- 要減少輸入欄位數，請使用功能選擇。請參閱第 45 頁的『功能選擇節點』主題，以取得更多資訊。另外，您可以使用初始建模執行來識別需要進一步探索的欄位和選項。例如，如果性能最佳的模型似乎都使用了相同的三個欄位，那麼有力地說明這些欄位值得保留。
- 您還可以限制評估任一模型所需的時間並且指定用於過濾和排等級模型的測量。

自動建模節點演算法設定

對於每個模型類型，可以使用預設值，或為每個模型類型選擇選項。這些特定選項類似於獨立建模節點中可用的選項，不同之處在於並非只能選擇一種設定而是大多數情況下可以根據套用需要選擇多種。例如，如果對比類神經網路模型，可以選擇幾種不同的訓練方法，並且在使用隨機種子和不使用隨機種子的情況下嘗試每種方法。選定選項的所有可能組合都將使用，從而使得在單次傳遞中產生多數不同模型變得更容易。但是，使用時要小心，因為選擇多個設定會引起模式個數非常快速地增加。

要為每個模型類型選擇選項：

1. 在自動建模節點上，選取**專家**標籤。
2. 按一下模型類型的**模型參數**欄。
3. 從下拉功能表中，選擇**指定**。
4. 在**演算法設定**對話框上，從**選項**欄中選取選項。

註：**演算法設定**對話框的「專家」標籤上提供了更多選項。

自動建模節點停止規則

為自動建模節點指定的停止規則與整體節點執行有關，與停止節點建置的個別模型無關。

限制整體執行時間。（僅限類神經網路、K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和 C&R 模型）在指定小時數後停止執行。所有在該時間點之前（併入該點）產生的模型都將併入在模型塊中，但這之後不會再產生模型。

生成有效模型後立即停止。當模型傳送了所有在「捨棄」標籤（自動分類器或自動叢集節點的）和「模型」標籤（自動數值節點的）上指定的準則時將停止執行。請參閱第 59 頁的『自動分類器節點捨棄選項』主題，以取得更多資訊。請參閱第 66 頁的『自動叢集節點捨棄選項』主題，以取得更多資訊。

自動分類器節點

「自動分類器」節點使用多種不同的方法來估計和比較名義（集合）或二元（是/否）目標的模型，這使您可以在一次建模執行中嘗試多種方法。您可以選取要使用的演算法，且可以試用多個選項的組合。例如，您無需在徑向基底函數、多項式、sigmoid 或線性方法中選擇一種來用於 SVM，您可以全部都嘗試一下。該節點將探究每種可能的選項組合，並根據您指定的測量對每個候選模型進行排等級，然後儲存最佳模型以用於評分或進行進一步分析。如需相關資訊，請參閱第 53 頁的第 5 章, 『自動建模節點』。

範例 某零售公司具有歷程資料，可用於追蹤過去行銷活動中針對特定客戶的報價。公司現在希望通過向每個客戶提供合適的報價來獲取更多的利潤。

需求 一個測量層次為名義或旗標的目標欄位（角色設定為**目標**）和至少一個輸入欄位（角色設定為**輸入**）。對於「旗標」欄位，假定為目標欄位定義的*true*值代表計算利潤、提升和相關統計資料時的命中數。輸入欄位的測量層次可以是連續或種類，但具有限制，即某些輸入可能不適合一些模型類型。例如，在 C&R 樹狀結構、CHAID 和 QUEST 模型中用作輸入的序數欄位必須是數值儲存類型（而不是字串），如果指定了其他類型，將被這些模型忽略。類似地，在某些情況下可對連續輸入欄位進行分級。這和使用單個建模節點時的要求一樣；例如，不管是從貝葉斯網絡節點還是自動分類器節點產生，貝葉斯網絡模型都以同樣的方式工作。

頻率和加權欄位

頻率與加權用來向部分記錄提供高於其他記錄的額外重要性，例如，使用者知道建置資料集未充分代表一部分母體（加權），或者因為一筆記錄代表數個相同的觀察值（頻率）。如果指定了頻率欄位，那麼 C&R 樹狀結構、CHAID、QUEST、決策清單和貝葉斯網絡模型可以使用此欄位。

C&RT、CHAID 和 C5.0 模型可以使用加權欄位。其他模型類型將忽略這些欄位並且無論如何都會建置模型。頻率和加權欄位僅用於模型建置，並且在評估和評分模型時不予以考慮。如需相關資訊，請參閱第 28 頁的『使用頻率和加權欄位』。

字首 如果您將表格節點附加到自動分類器節點塊，那麼表格中存在多個名稱以字首 $ 開頭的新變數。

在評分期間產生的欄位名稱基於目標欄位，但具有標準字首。不同的模型類型使用不同的字首集。

例如，字首 $G、$R、$C 分別用作「一般線性」模型、CHAID 模型及 C5.0 模型所產生預測的字首。$X 通是利用集合來產生，而 $XR、$XS 和 $XF 分別在目標欄位是「連續」、「種類」或「旗標」欄位的情況下用作字首。

$..C 字首用於預測種類或旗標目標的信賴度；例如，$XFC 用作總體旗標預測信賴度的字首。$RC 和 $CC 分別是 CHAID 模型和 C5.0 模型的單一預測信賴度的字首。

支援的模型類型

受支援的模型類型包含類神經網路、C&R 樹狀結構、QUEST、CHAID、C5.0、邏輯迴歸、決策清單、貝葉斯網絡、判別、最近鄰接項、SVM、XGBoost 樹狀結構及 XGBoost-AS。請參閱第 56 頁的『自動分類器節點專家選項』主題，以取得更多資訊。

自動分類器節點模型選項

通過「自動分類器」節點的「模型」標籤，您可以指定要建立的模式個數以及用於比較模型的準則。

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

模型的評級依據。 指定用來比較及分級模型的準則。選項包含整體精確度、ROC 曲線下的區域、利潤、提升和欄位的數量。請注意，無論在此處選定哪些尺度，所有這些尺度都能在彙總報告中使用。

註：對於名義（集合）目標，排等級限制為**整體精確度**或**欄位數**。

計算利潤、提升和相關統計資料時，將假定為目標欄位定義的 *true* 值代表一個命中。

- **整體精確度** 模型正確預測的記錄相對於記錄總數的百分比。
- **ROC 曲線下方區域** ROC 曲線提供模型的效能指標。曲線位置距參照線越遠，則測試準確度越高。
- **利潤（累加）** 根據指定的成本、營收和加權準則計算出的各個累加百分位數（按預測的信賴度排序）的利潤總和。通常，頂部百分位數的初始利潤接近於零，然後逐步增加，最後減少。對於構建完好的模型，利潤將顯示一個正確定義的尖峰，隨尖峰出現所在的百分位數一起報告。對於不包含任何資訊的模型，利潤曲線將相對較直，可能顯示為遞增、遞減或水平，具體取決於所採用的成本/營收結構。
- **提昇（累加）** 相對於整個樣本（其中分位數按預測的信賴度排序）的累加分位數匹配率。例如，頂部分位數提升值 3 表示其匹配率是整個樣本的三倍。對於構建完好的模型，頂部分位數提升起點應該遠高於 1.0，然後針對低分位數徑直向1.0 跌落。對於不包含任何資訊的模型，提升將在 1.0 上下徘徊。
- **欄位個數** 根據所使用的輸入欄位數對模型進行排等級。

模型評級使用的方法。 如果分割區在使用中，您可以指定等級是基於訓練資料集還是測試集。對於大型資料集，使用分割區來初步顯示模型可能會極大地改善效能。

要使用的模型數。 指定要在節點生成的模型塊中列出的上限模式個數。會根據指定的分級準則列出等級最高的模型。請注意，增加此限制可能會降低效能。容許的上限值是 100。

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，預測值重要性可能會延長計算某些模型所需的時間，如果您只需要在數個不同的模型之間進行廣泛比較，則不建議使用它。當您將分析縮至您要更詳細地探索的少量模型時，此函數更為有用。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

利潤準則。附註。僅適用於旗標目標。利潤等於每條記錄的營收減去該記錄的成本。分位數的利潤僅僅是分位數中所有記錄的利潤總和。這裡假定利潤僅套用至命中數，但成本可套用至所有的記錄。

- **成本。**指定與每個記錄關聯的成本。您可以選取**固定**或**變數**成本。. 對於固定成本，請指定成本值。對於變動成本，請按一下「欄位選擇器」按鈕，將某個欄位選擇為成本欄位。（**成本**不適用於 ROC 圖表。）
- **營收。**指定與代表命中數的每個記錄關聯的營收。您可以選取**固定**或**變數**成本。. 對於固定收益，請指定營收值。對於可變營收，請按一下「欄位選擇器」按鈕，將某個欄位選擇為營收欄位。（**營收**不適用於 ROC 圖表。）
- **加權。** 如果資料中的記錄代表多個單元，那麼可以使用頻率加權來調整結果。使用**固定**或**可變**加權，指定與每個記錄關聯的加權。對於固定加權，請指定加權值（每個記錄的單元數）。對於變動加權，請按一下「欄位選擇器」按鈕，將某個欄位選擇為加權欄位。（**加權**不適用於 ROC 圖表。）

提昇準則。附註。僅適用於旗標目標。指定提升計算使用的百分位數。注意，在比較結果時也可以變更此值。請參閱第 66 頁的『自動模型塊』主題，以取得更多資訊。

自動分類器節點專家選項

通過「自動分類器」節點的「專家」標籤，您可以套用分割區（如果可用），選取要套用的演算法以及指定停止規則。

選取模型。 依預設，將選取所有模型以進行建置；但是，如果您具有 Analytic Server，則可以將模型限制為可在 Analytic Server 上執行的那些模型，並預設模型以讓它們建置分割模型或準備處理非常大的資料集。

註：不支援在「自動分類器」節點內對 Analytic Server 模型進行本端建立.

已使用的模型。使用左側欄中的勾選框選取要在比較中包含的模型類型（演算法）。選取的類型越多，建立的模型就會越多，且正在處理的時間就會越長。

模型類型 列出可用的演算法（請參閱下方的內容）。

模型參數。對於每個模型類型，可以使用預設值，或選擇**指定**為每個模型類型選擇選項。這些特定選項類似於獨立建模節點中可用的選項，不同之處在於可以選取多個選項或組合。例如，比較類神經網路模型時，與其選擇六種訓練方法之一，還不如一次選中全部六種方法以在一次傳送中訓練六種模型。

模式個數。列出根據目前設定為每個演算法生成的模型的號碼。當組合選項時，模式個數會激增，因此強烈建議密切關注該模式個數，尤其在使用大型資料集時。

限制建立單個模型所花費的最長時間。（僅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和決策清單模型）為任意一個模型設定最長時間限制。例如，如果由於某些複合的互動效應，某個特定模型所需的訓練時間長得出乎意料，那麼您大概不希望它使得整個的建模執行停滯。

註：如果目標為名義（集合）欄位，那麼「決策清單」選項不提供。

支援的演算法

使用支援向量機器 (SVM) 節點，可以將資料分為兩群組，而無需過度配適。SVM 可以與大量資料集配合使用，例如那些含有大量輸入欄位的資料集。

The *k*-最近相鄰元素 (KNN) 節點將新的觀察值關聯到預測值空間中與其最鄰近的 *k* 個物件的種類或值（其中 *k* 為整數）。相似的觀察值會彼此相鄰，相異的觀察值則會彼此相隔。

判別分析所做的假設比邏輯迴歸方法的假設更嚴格，但在符合這些假設時，判別分析可以作為邏輯迴歸方法分析的有用替代項或補充。

通過貝葉斯網路節點，你可以利用對真實世界認知的判斷力並結合所觀察和記錄的證據來建立機率模型。該節點重點應用了樹狀結構擴展簡單貝葉斯 (TAN) 和馬爾可夫覆蓋網路，這些算法主要用於分類問題。

決策清單節點可識別子群組或區段，顯示與整體相關的給定二元結果的概似度的高低。例如，您或許在尋找那些最不可能流失的客戶或最有可能對某個商業活動作出積極回應的客戶。通過自訂區段和並排預覽備選模型來比較結果，您可以將自己的業務知識體現在模型中。決策清單模型由一組規則構成，其中每個規則具備一個條件和一個結果。規則依順序套用，相符的第一個規則將決定結果。

邏輯迴歸是一種統計技術，它可根據輸入欄位的值對記錄進行分類。它類似於線性迴歸方法，但採用的是種類目標欄位而非數值範圍。

CHAID 使用卡方統計資料來產生決策樹，以確定最佳的分割。與 C&R Tree 和 QUEST 節點不同，CHAID 可產生非二元樹狀結構，表示部分分割有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

QUEST 節點可提供用於建立決策樹的二元分類法，此方法的設計目的是減少大型 C&R 樹狀結構分析所需的處理時間，同時也減少在分類樹狀結構方法中發現的趨勢以便偏愛容許有多個分割的輸入。輸入欄位可以是數值範圍（連續），但目標欄位必須是種類。所有分割都是二元的。

分類和迴歸方法 (C&R) 樹狀結構節點產生可用於預測或分類未來觀察的決策樹。該方法通過在每一個步驟最大限度降低不純潔度，使用遞歸分區來將訓練記錄分割為組。如果樹狀結構中某個節點中 100% 的觀察值都的目標欄位的一個特定種類，那麼該節點將被認定為「純潔」。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

C5.0 節點建立決策樹或規則 集。該模型的工作原理是根據在每個層次提供上限資訊收穫的欄位分割樣本。目標欄位必須為種類欄位。容許進行多次多於兩個子群組的分割。

類神經網路節點使用的模型是對人類大腦處理資訊的方式簡化了的模型。此模型通過模擬大量類似於神經元的抽象形式的互連簡單處理裝置而運行。神經網路是功能強大的一般函數估計器，只需要最少的統計或數學知識就可以對其進行訓練或套用。

線性迴歸模型基於目標和一個或多個預測值之間的線性關係預測連續目標。

通過線性支援向量機器 (LSVM) 節點，您可以將資料分為兩群組，而無需過度配適。LSVM 是線性的，並且可以與大量資料集配合使用，例如包含大量記錄的資料集。

「隨機樹狀結構」節點類似於現有 C&RT 節點；但是，「隨機樹狀結構」節點旨在處理大資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。「隨機樹狀結構」節點將產生您可以對未來觀察進行預測或分類的決策樹。通過在每一個步驟最大限度降低不純潔度，此方法使用遞歸分區將訓練記錄分割為多個區段。如果樹狀結構中某個節點的全部觀察值都的目標欄位的一個特定種類，那麼系統會將該節點視為純潔。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

樹狀結構-AS 節點類似於現有的 CHAID 節點；但是，「樹狀結構-AS」節點旨在處理大量資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。此節點通過使用卡方統計量 (CHAID) 來識別最佳化分割，從而產生決策樹。 對 CHAID 的這一使用可產生非二元樹狀結構，意味著某些分割將具有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

XGBoost Tree© 是將樹狀結構模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost Tree 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost Tree 節點僅顯示了核心功能和一般參數。該節點是以 Python 來實作的。

XGBoost© 是梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost-AS 節點僅顯示了核心功能和一般參數。XGBoost-AS 節點是以 Spark 來實作的。

註： 如果選取「樹狀結構-AS」以在 Analytic Server 上執行，當存在「分割區」節點上游時，它將無法建立模型。在此情況下，為使「自動分類器」能夠與 Analytic Server 上的其他建模節點一起工作，請取消選取「樹狀結構-AS」模型類型。

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。

自動分類器節點捨棄選項

通過「自動分類器」節點的「捨棄」標籤，您可以自動捨棄不符合特定準則的模型。這些模型將不會列在彙總報告中。

可以為總精確度指定下限臨界值，為模型中使用的變數數目指定上限臨界值。此外，對於旗標目標，可以為提升、利潤和曲線下方區域指定下限臨界值，提升和利潤由在「模型」標籤上指定的內容所確定。請參閱第 55 頁的『自動分類器節點模型選項』主題，以取得更多資訊。

您可以選擇性地配置節點，以在第一次產生符合所有指定準則的模型時停止執行節點。請參閱第 54 頁的『自動建模節點停止規則』主題，以取得更多資訊。

自動分類器節點設定選項

通過「自動分類器」節點的「設定」標籤，您可以預配置塊上可用的分數時間選項。

過濾掉組合模型所產生的欄位。 從輸出中移除由個別模型（為組合節點提供資訊來源）產生的所有其他欄位。如果您僅對所有輸入模型中的結合評分感興趣，請選取這個勾選框。確保在下列情況下取消選取此選項：例如，如果您要使用「分析」節點或「評估」節點來比較結合評分的準確性與每一個個別輸入模型的評分準確性。

自動數值節點

「自動數值」節點使用多種不同方法來估計和比較模型以得出連續數值型範圍結果，這使您可以在一次建模執行中嘗試多種方法。您可以選取要使用的演算法，且可以試用多個選項的組合。例如，您可以使用神經網絡、線性迴歸、C&RT 和 CHAID 模型預測住房價值，以確定哪種模型的性能最好，並且可以嘗試逐步、向前和向後迴歸方法的不同組合。該節點會探索每個可能的選項組合，根據您指定的測量為每個候選模型評級，並儲存最佳實務以用於評分蘗或未來分析。請參閱第 53 頁的第 5 章, 『自動建模節點』主題，以取得更多資訊。

範例 市政當局需要更準確地估計房地產稅以及無需檢查財產就可以按需要調整特定財產的價值。通過使用「自動數值」節點，分析師可以產生並對比許多模型，這些模型根據建立類型、芳鄰、大小和其他已知因子來預測內容值。

需求 一個目標欄位（角色設定為**目標**）和至少一個輸入欄位（角色設定為**輸入**）。目標必須為連續（數值型範圍）欄位，如年齡或收入。輸入欄位可以是連續或種類，但具有限制，即某些輸入可能不適合一些模型類型。例如，C&R 樹狀結構模型能將種類字串欄位作為輸入使用，而線性迴歸模型不能使用這些欄位並將在指定這些欄位後省略它們。這和使用個別建模節點時的要求相同。例如，不管 CHAID 模型是在 CHAID 節點中還是在自動數值節點中產生，其工作方式都相同。

頻率和加權欄位

頻率與加權用來向部分記錄提供高於其他記錄的額外重要性，例如，使用者知道建置資料集未充分代表一部分母體（加權），或者因為一筆記錄代表數個相同的觀察值（頻率）。如果指定頻率欄位，那麼 C&R 樹狀結構和 CHAID 演算法可以使用該欄位。C&RT、CHAID 迴歸方法和 GenLin 演算法可以使用加權欄位。其他模型類型將忽略這些欄位並且無論如何都會建置模型。頻率和加權欄位僅用於模型建置，並且在評估和評分模型時不予以考慮。請參閱第 28 頁的『使用頻率和加權欄位』主題，以取得更多資訊。

字首 如果您將表格節點附加到自動數值節點塊，那麼表格中存在多個名稱以字首 $ 開頭的新變數。

在評分期間產生的欄位名稱基於目標欄位，但具有標準字首。不同的模型類型使用不同的字首集。

例如，字首 $G、$R、$C 分別用作「一般線性」模型、CHAID 模型及 C5.0 模型所產生預測的字首。$X 通是利用集合來產生，而 $XR、$XS 和 $XF 分別在目標欄位是「連續」、「種類」或「旗標」欄位的情況下用作字首。

$..E 字首用於連續目標的預測信賴度；例如，$XRE 用作總體連續預測信賴度的字首。$GE 是廣義線性模型的單個預測信賴度的字首。

支援的模型類型

受支援的模型類型包含類神經網路、C&R 樹狀結構、CHAID、迴歸方法、GenLin、最近相鄰元素、SVM、XGBoost Linear、GLE 及 XGBoost-AS。如需相關資訊，請參閱第 61 頁的『自動數值節點專家選項』。

自動數值節點模型選項

通過「自動數值」節點的「模型」標籤，您可以指定要儲存的模式個數以及用於比較模型的準則。

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

模型的評級依據。 指定用於比較模型的準則。

- **相關性。** 這是每條記錄的觀察值和模型預測的值之間的 Pearson 相關性。相關性是兩種變數之間的線性關聯尺度，值越接近 1 說明變數之間的關係越強。（相關性的值在 –1 和 +1 之間，–1 代表完全負關係，+1 代表完全正關係。值為 0 表示無線性關係，但具有負相關性的模型將排在最後。）
- **欄位個數。**模型中用作預測值的欄位的號碼。在某些情況下，選擇使用較少欄位的模型可簡化資料預備過程並提高效能。
- **相對錯誤。**相對錯誤是模型觀察值相對於預測值的變異數與觀察值相對於平均數的變異數的比例。在實際應用的角度，它對比模型相對於**無效**或**截距**模型（僅傳回目標欄位的平均數作為預測值）的性能。對於好的模型，此值應少於 1，說明此模型比無效模型更精確。相對錯誤大於 1 的模型不如無效模型精確，因此這樣的模型沒有意義。對於線性迴歸模型，相對錯誤等同於相關性的平方並且未新增任何新的資訊。對於非線性模型，相對錯誤與相關性無關並且為評估模型效能提供了附加尺度。

模型評級使用的方法。 如果正在使用分割區，那麼可以指定根據訓練分割區還是測試分割區進行排等級。對於大型資料集，使用分割區來初步顯示模型可能會極大地改善效能。

要使用的模型數。 指定要在節點生成的模型塊中顯示的上限模式個數。會根據指定的分級準則列出等級最高的模型。通過增加此限制，您可以對比更多模型的結果，但這可能會降低效能。容許的上限值是 100。

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，預測值重要性可能會延長計算某些模型所需的時間，如果您只需要在數個不同的模型之間進行廣泛比較，則不建議使用它。當您將分析縮至您要更詳細地探索的少量模型時，此函數更為有用。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

在下列情況下不儲存模型。指定相關性、相對誤和所用欄位個數的臨界值。無法符合這些準則中的任意一個的模型將被捨棄，並且不會在彙總報告中列出。

- **相關性少於。**這是要包含在彙總報告中的模型的下限相關性（以絕對值表示）。
- **使用的欄位個數超過。**這是要包含的任意模型將使用的上限欄位個數。
- **相對錯誤大於。**這是要包含的任意模型的上限相對錯誤。

您可以選擇性地配置節點，以在第一次產生符合所有指定準則的模型時停止執行節點。請參閱第 54 頁的『自動建模節點停止規則』主題，以取得更多資訊。

自動數值節點專家選項

通過「自動數值」節點的「專家」標籤，您可以選取要使用的演算法和選項並指定中止規則。

選取模型。 依預設，將選取所有模型以進行建置；但是，如果您具有 Analytic Server，則可以將模型限制為可在 Analytic Server 上執行的那些模型，並預設模型以讓它們建置分割模型或準備處理非常大的資料集。

註：不支援在「自動數值」節點內對 Analytic Server 模型進行本端建立.

已使用的模型。使用左側欄中的勾選框選取要在比較中包含的模型類型（演算法）。選取的類型越多，建立的模型就會越多，且正在處理的時間就會越長。

模型類型 列出可用的演算法（請參閱下方的內容）。

模型參數。對於每個模型類型，可以使用預設值，或選擇**指定**為每個模型類型選擇選項。這些特定選項類似於獨立建模節點中可用的選項，不同之處在於可以選取多個選項或組合。例如，比較類神經網路模型時，與其選擇六種訓練方法之一，還不如一次選中全部六種方法以在一次傳送中訓練六種模型。

模式個數。列出根據目前設定為每個演算法生成的模型的號碼。當組合選項時，模式個數會激增，因此強烈建議密切關注該模式個數，尤其在使用大型資料集時。

限制建立單個模型所花費的最長時間。（僅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和決策清單模型）為任意一個模型設定最長時間限制。例如，如果由於某些複合的互動效應，某個特定模型所需的訓練時間長得出乎意料，那麼您大概不希望它使得整個的建模執行停滯。

支援的演算法

類神經網路節點使用的模型是對人類大腦處理資訊的方式簡化了的模型。此模型通過模擬大量類似於神經元的抽象形式的互連簡單處理裝置而運行。神經網路是功能強大的一般函數估計器，只需要最少的統計或數學知識就可以對其進行訓練或套用。

分類和迴歸方法 (C&R) 樹狀結構節點產生可用於預測或分類未來觀察的決策樹。該方法通過在每一個步驟最大限度降低不純潔度，使用遞歸分區來將訓練記錄分割為組。如果樹狀結構中某個節點中 100% 的觀察值都的目標欄位的一個特定種類，那麼該節點將被認定為「純潔」。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

CHAID 使用卡方統計資料來產生決策樹，以確定最佳的分割。與 C&R Tree 和 QUEST 節點不同，CHAID 可產生非二元樹狀結構，表示部分分割有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

線性迴歸是一種通過配適直線或平面以實現彙總資料和預測的普通統計技術，它可使預測值和實際輸出值之間的差異最小化。

「通用性線性」模型對一般線性模型進行了擴展，這樣依變數通過指定的鏈結函數與因子和共變數線性相關。此外，此模式允許」變數具有非常態分配。它包括統計模型大部分的功能，其中包括線性迴歸、邏輯迴歸方法、用於計數資料的對數線性模型以及區間刪失生存分析模型。

The *k*-最近相鄰元素 (KNN) 節點將新的觀察值關聯到預測值空間中與其最鄰近的 *k* 個物件的種類或值（其中 *k* 為整數）。相似的觀察值會彼此相鄰，相異的觀察值則會彼此相隔。

使用支援向量機器 (SVM) 節點，可以將資料分為兩群組，而無需過度配適。SVM 可以與大量資料集配合使用，例如那些含有大量輸入欄位的資料集。

線性迴歸模型基於目標和一個或多個預測值之間的線性關係預測連續目標。

通過線性支援向量機器 (LSVM) 節點，您可以將資料分為兩群組，而無需過度配適。LSVM 是線性的，並且可以與大量資料集配合使用，例如包含大量記錄的資料集。

「隨機樹狀結構」節點類似於現有 C&RT 節點；但是，「隨機樹狀結構」節點旨在處理大資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。「隨機樹狀結構」節點將產生您可以對未來觀察進行預測或分類的決策樹。通過在每一個步驟最大限度降低不純潔度，此方法使用遞歸分區將訓練記錄分割為多個區段。如果樹狀結構中某個節點的全部觀察值都的目標欄位的一個特定種類，那麼系統會將該節點視為純潔。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

樹狀結構-AS 節點類似於現有的 CHAID 節點；但是，「樹狀結構-AS」節點旨在處理大量資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。此節點通過使用卡方統計量 (CHAID) 來識別最佳化分割，從而產生決策樹。 對 CHAID 的這一使用可產生非二元樹狀結構，意味著某些分割將具有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

XGBoost Linear© 是將線性模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 SPSS Modeler 中的 XGBoost Linear 節點使用 Python 進行實現。

GLE 延伸了線性模型，以便目標可以有非常態分佈，通過指定的連接函數與因子和共變數線性相關，並且觀察可能相關。通用性線性混合模型涵蓋多種模式，從非常態縱向資料的簡單線性迴歸，到複雜的多層級模式。

XGBoost© 是梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost-AS 節點僅顯示了核心功能和一般參數。XGBoost-AS 節點是以 Spark 來實作的。

自動數值節點設定選項

通過「自動數值」節點的「設定」標籤，您可以預先配置塊上可用的分數時間選項。

過濾掉組合模型所產生的欄位。 從輸出中移除由個別模型（為組合節點提供資訊來源）產生的所有其他欄位。如果您僅對所有輸入模型中的結合評分感興趣，請選取這個勾選框。確保在下列情況下取消選取此選項：例如，如果您要使用「分析」節點或「評估」節點來比較結合評分的準確性與每一個個別輸入模型的評分準確性。

計算標準誤差。對於連續（數值型範圍）目標，依預設會執行標準誤差計算以計算測量或估計值與 true 值之間的差分；並顯示這些估計值的相近符合程度。

自動叢集節點

自動叢集節點估計和比較識別具有類似特性記錄群組的叢集模型。節點的工作方式與其他自動建模節點相同，這使您可以在一次建模運行中試驗多個選項組合。可以使用基本測量比較模型，透過測量嘗試過濾並分級叢集模型的實用性，並基於特定欄位的重要性提供測量。

叢集模型常常用於識別在後續分析中可用作輸入的群組。例如，您可能希望基於如收入的統計特性來針對客戶群，或基於客戶過去購買的服務而針對客戶群。可以在不瞭解客戶群及其特性的情況下進行此操作 -- 您可能不知道要尋找多少個客戶群，或該用什麼特性去定義客戶群。叢集模型常稱作不受監督的學習模型，因為其不使用目標欄位，且不傳回可評估為 true 或 false 的特殊預測。叢集模型的值由模型擷取資料中感興趣的分組並提供這些分組的有用說明資訊的能力來確定。如需相關資訊，請參閱第 197 頁的第 11 章, 『叢集模型』。

需求。 這是用於定義興趣特性的一個或多個欄位。叢集模型使用目標欄位的方式與其他模型不同，因為其不作出能被評估為 true 或假的特定預測。相反，其用於識別可能相關的觀察值群組。例如，您無法使用預測給定客戶會流失還是對預訂作出積極回應的叢集模型。但您可以使用叢集模型，根據客戶對做此類事項的傾向，將客戶指派給群組。不會使用加權和頻率欄位。

評估欄位。 雖然不使用目標，但是您可以選擇性地指定要在比較模型中使用的一個或多個評估欄位。可通過衡量叢集是否能有效區分這些欄位，評估叢集模型的效果。

支援的模型類型

受支援的模型類型包括 TwoStep、K-Means、Kohonen、一類 SVM 及 K-Means-AS。

自動叢集節點模型選項

使用「自動叢集」節點的「模型」標籤可以指定要儲存的模式個數，以及用於比較模型的準則。

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。

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。

模型的評級依據。 指定用來比較及分級模型的準則。

- **Silhouette。**這是用於衡量叢集結合與分離特性的指數。有關詳細資訊，請參閱下方的 *Silhouette* 分等級測量。
- **叢集數目。** 模型中的叢集數目。
- **最小叢集的大小。**最小叢集的大小。
- **最大叢集的大小。**最大叢集的大小。
- **最小/最大叢集。**最小叢集與最大叢集的大小比例。
- **重要性。** **欄位**標籤上的**評估**欄位的重要性。注意只有在**評估**欄位已指定時，才能計算。

模型評級使用的方法。 如果分割區在使用中，您可以指定等級是基於訓練資料集還是測試集。對於大型資料集，使用分割區來初步顯示模型可能會極大地改善效能。

要保留的模式個數。指定要在節點生成的塊中列出的上限模式個數。會根據指定的分級準則列出等級最高的模型。請注意，增加此限制可能會降低效能。容許的上限值是 100。

Silhouette 分等級測量

預設分等級測量，Sil預設排名測量 Silhouette 的預設值為 0，這是因為少於 0 的值（即負值）表示觀察值和其指派叢集中的點之間平均距離大於觀察值與另一個叢集中點的下限平均距離。因此，具有負 Silhouette 值的模型可以安全地捨棄。

分等級測量實際上為修改的 silhouette 係數，它結合了叢集結合（偏向包含緊密結合叢集的模型）和叢集分離（偏向包含高度分離叢集的模型）的概念。平均 Silhouette 係數是在所有觀察值上的簡單平均值，每個個別觀察值應用下列計算：

$(B - A) / \max(A, B)$

其中 A 為從觀察值到其所屬叢集的矩心的距離，B 為從觀察值到每個其他叢集矩心的最小距離。

Silhouette 係數（及其平均值）大小在 -1（表示極低劣的模型）與 1（表示極好的模型）之間。可以在總體觀察值層次上求平均值（得到總體 Silhouette），也可在叢集層次上求平均值（得到叢集 Silhouette）。距離可以使用 Euclidean 距離進行計算。

自動叢集節點專家選項

通過「自動叢集」節點的「專家」標籤，您可以套用分割區（如果可用），選取要套用的演算法以及指定停止規則。

選取模型。 依預設，將選取所有模型以進行建置；但是，如果您具有 Analytic Server，則可以將模型限制為可在 Analytic Server 上執行的那些模型，並預設模型以讓它們建置分割模型或準備處理非常大的資料集。

註：不支援在「自動叢集」節點內對 Analytic Server 模型進行本端建置。

已使用的模型。使用左側欄中的勾選框選取要在比較中包含的模型類型（演算法）。選取的類型越多，建立的模型就會越多，且正在處理的時間就會越長。

模型類型 列出可用的演算法（請參閱下方的內容）。

模型參數。對於每個模型類型，可以使用預設值，或選擇**指定**為每個模型類型選擇選項。這些特定選項類似於獨立建模節點中可用的選項，不同之處在於可以選取多個選項或組合。例如，比較類神經網路模型時，與其選擇六種訓練方法之一，還不如一次選中全部六種方法以在一次傳送中訓練六種模型。

模式個數。列出根據目前設定為每個演算法生成的模型的號碼。當組合選項時，模式個數會激增，因此強烈建議密切關注該模式個數，尤其在使用大型資料集時。

限制建立單個模型所花費的最長時間。（僅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和決策清單模型）為任意一個模型設定最長時間限制。例如，如果由於某些複合的互動效應，某個特定模型所需的訓練時間長得出乎意料，那麼您大概不希望它使得整個的建模執行停滯。

支援的演算法

K-Means 節點將資料集叢集到不同群組（或叢集）。此方法將定義固定的叢集數目量，將記錄迭代分配給叢集，以及調整叢集中心，直到進一步優化無法再精確模型。*k*-means 節點作為一種非監督學習機制，它並不試圖預測結果，而是揭示隱含在輸入欄位集中的型樣。

Kohonen 節點會產生一種類神經網路，此類神經網路可用於將資料集叢集到各個差異群組。此網路訓練完成後，相似的記錄應在輸出對映中緊密地聚集，差異大的記錄則應彼此遠離。您可以通過查看模型塊 中每個單位所擷取觀察的數量來找出規模較大的單元。這將讓您對叢集的相應數量有所估計。

TwoStep 節點使用二階叢集方法。第一步完成簡單資料製作，以便將原始輸入資料壓縮為可管理的子叢集集合。第二步使用層級叢集方法將子叢集一步一步合併為更大的叢集。TwoStep 具有一個優點，就是能夠為訓練資料自動估計最佳叢集數目。它可以高效處理混合的欄位類型和大型的資料集。

「一類 SVM」節點使用未受監督的學習演算法。該節點可用來偵測新事件。它將偵測給定樣本集的軟性界限，然後將新的點分類成是否的該集合。此一級 SVM 建模節點在 SPSS Modeler 中使用 Python 進行實現並且需要 `scikit-learn`© Python 程式庫。

K-Means 是其中一個最常用的叢集演算法。它將資料點叢集化至預先定義的叢集數。SPSS Modeler 中的 K-Means-AS 節點是在 Spark 中進行實作。如需 K-Means 演算法的詳細資料，請參閱 https://spark.apache.org/docs/2.2.0/ml-clustering.html。請注意，K-Means-AS 節點自動針對類別變數執行 one-hot 編碼。

自動叢集節點捨棄選項

使用「自動叢集」節點的「捨棄」標籤可以自動捨棄不符合特定準則的模型。這些模型將不會列在模型塊中。

您可以指定下限 silhouette 值、叢集數、叢集大小和模型中所用評估欄位的重要性。Silhouette 以及叢集的數量和大小根據建模節點中指定的值確定。請參閱第 64 頁的『自動叢集節點模型選項』主題，以取得更多資訊。

您可以選擇性地配置節點，以在第一次產生符合所有指定準則的模型時停止執行節點。請參閱第 54 頁的『自動建模節點停止規則』主題，以取得更多資訊。

自動模型塊

執行自動建模節點時，節點評估每個可能選項組合的候選模型，基於您指定的測量為每個候選模型排等級，並將最佳模型儲存在複合自動模型塊中。此模型塊實際上包含該節點產生的一個或多個模型的集合，其中模型可個別被瀏覽或已選取用於評分。每個模型列有模型類型和建立時間，以及適合該模型類型的多個其他測量。可以按照這些欄中的任意一欄對表格進行排序，以便快速確定最關注的模型。

- 要瀏覽任何一個個別的模型塊，請按兩下模型塊圖示。然後，可以從這裡產生該模型的建模節點到串流畫布，或產生模型塊副本到模型選用區。
- 使用縮圖圖形可以快速而直觀地評量每個模型類型，總結如下方。可以按兩下縮圖產生完整大小的圖表。標準大小的統計圖可以最多顯示 1000 個點並且會在資料集包含更多點時基於樣本。（僅對於散佈圖，圖表每顯示一次就重新產生一次，所以上游資料中的任意變更（例如在未選取**設定隨機種子**時更新隨機樣本或分割區）在每次重新繪製散佈圖時都會反映出來。
- 使用工具列在「模型」標籤上顯示或隱藏特定的欄或變更用於對表格排序的欄。（也可以通過按一下欄標題變更順序。）
- 使用「刪除」按鈕以永久地刪除任何未用的模型。
- 要重新為欄排序，請按一下欄標題並將該欄拖放到所需位置。
- 如果正在使用分割區，那麼可選擇檢視可應用的訓練分割區或測試分割區的結果。

特定的欄取決於要對比的模型的類型，如下所述。

二元目標

- 對於二元模型，縮圖圖表顯示實際值的分佈，上面覆蓋預測值，來快速直觀地表示每個種類中正確預測的記錄數。
- 分級準則符合「自動分類器」建模節點中的選項。請參閱第 55 頁的『自動分類器節點模型選項』主題，以取得更多資訊。
- 對於最大值利潤，還會報告產生的數目上限的百分位數。
- 對於累加提升，可以使用工具列變更選定的百分位數。

列名目標

- 對於列名（集合）模型，縮圖圖表顯示實際值的分佈，上面覆蓋預測值，來快速直觀地表示每個種類中正確預測的記錄數。
- 分級準則符合「自動分類器」建模節點中的選項。請參閱第 55 頁的『自動分類器節點模型選項』主題，以取得更多資訊。

連續目標

- 對於連續（數值型範圍）模型，將根據每個模型的觀察值預測圖表散點，從而快速直觀地表示模型之間的相關性。對於好的模型，點應趨向於聚集在對角線周圍，而不是在整個圖表中隨機分佈。
- 分等級準則與「自動數值」建模節點中的選項比對。請參閱第 60 頁的『自動數值節點模型選項』主題，以取得更多資訊。

叢集目標

- 對於叢集模型，將根據每個模型的叢集計算圖表散點，從而快速直觀地表示叢集分佈。
- 分等級準則與「自動叢集」建模節點中的選項比對。請參閱第 64 頁的『自動叢集節點模型選項』主題，以取得更多資訊。

選取評分模型

使用？欄可選取評分中使用的模型。

- 對於二元、列名和數值目標，您可以選取多個評分模型，並將評分組合在一個總體模型塊中。透過結合來自多個模型的預測，可避免個別模型中存在的限制，通常會導致整體精確度高於可從任何一個模型獲取的精確度。
- 對於叢集模型，一次只能選取一個評分模型。依預設，首先選取頂級模型。

產生節點和模型

可以從複合自動模型塊的建立位置產生其副本，或自動建模節點。例如，當您沒有從中建立自動模型塊的原始串流時，這可能非常有用。此外，還可以為自動模型塊中列出的任何個別模型產生模型塊或建模節點。

自動建模塊

從「產生」功能表中，選取**模型至選用區**將自動模型塊新增到模型選用區上。可對產生的模型進行儲存，或者在不重新執行串流的情況下使用它。

或者，可以從「產生」功能表中選取**產生建模節點**以便將建模節點新增到串流畫布。可以不用重複完整的建模執行，而使用此節點重新估計選定的模型。

個別模型塊

1. 在**模型**功能表中，按兩下所需的個別模型塊。塊副本在新的對話框中開啟。
2. 從新對話框中的「產生」功能表中，選取**模型至選用區**將個別建模塊新增到模型選用區上。
3. 或者，可以從新對話框中的「產生」功能表中選取**產生建模節點**以便將個別建模節點新增到串流畫布。

產生評估表

對於二元模型，可以產生評估表以直觀評價和對比每個模型的效能。評估表不適用於自動數值型或自動叢集節點產生的模型。

1. 在自動分類器自動模型塊的使用？欄下，選取要評估的模型。
2. 從「產生」功能表中，選擇**評估表**。這將顯示「評估表」對話框。
3. 選取圖表類別和其他需要的選項。

評估圖

在自動模型塊的「模型」標籤上，可以向下瀏覽以顯示所示每個模型的個別圖形。對於自動分類器和自動數值塊，「圖表」標籤同時顯示圖表和預測值重要性以反映所有結合模型的結果。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

對於自動分類器，則顯示分佈圖形；而對於自動數值則顯示多重繪圖（也稱為「散佈圖」）。

第 6 章 決策樹

決策樹模型

決策樹模型可用於開發分類系統，此分類系統可以基於一組決策規則來預測或分類未來的觀察。 如果已將資料分成您感興趣的類別（例如，高風險和低風險貸款、訂閱者和非訂閱者、投票人和非投票人或細菌類型），那麼您可以使用自己的資料來建立用於對具有最高精確度的舊觀察值或新觀察值進行分類的規則。例如，可以基於年齡和其他因素建立對信貸風險或購買目的進行分類的樹狀結構。

此方法（有時稱為規則歸納）有多個優點。首先，瀏覽樹狀結構的同時可以明顯地看出模型背後的推論過程。這與其他「黑箱」建模技術不同的地方，在其他「黑箱」建模技術中，您很難瞭解其內部邏輯。

其次，此過程只會將真正影響決策的屬性自動包含在其規則中。不會提高樹狀結構的精確度的屬性將被忽略。此技術可獲得非常有用的資料資訊，並且可用於在訓練其他學習技術（如類神經網路）之前將資料縮減到相關欄位。

決策樹模塊可轉換成 if-then 規則的集合（規則集），在多數情況下此規則集以更為複雜的形式顯示資訊。樹狀結構表示法可以讓您知道資料屬性是如何將總體分割或分割區成與問題相關的子集。樹狀結構-AS 節點輸出不同於其他決策樹節點，因為它在塊中直接包含規則清單，無需建立規則集。 規則集表示法可以讓您知道特定項目群組與具體結論是如何關聯的。例如，下列規則就提供了關於值得購買的一群組汽車的分析概要：

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

樹狀結構建置演算法

有多種演算法可用於執行分類和分區段分析。這些演算法執行的操作基本相同，檢查資料集中的所有欄位，通過將資料分割為多個子群組來找到能夠實現最佳分類或預測的欄位。 此過程將重複套用以將子群組分割成越來越小的單位，直到樹狀結構結束生長（由特定的停止準則所定義）。建立樹狀結構的過程中所用的目標和輸入欄位可以是連續（數值範圍）或種類（這取決於所採用的演算法）。如果使用的是連續目標，那麼產生迴歸方法樹狀結構；如果使用的是種類目標，那麼產生分類樹狀結構。

分類和迴歸方法 (C&R) 樹狀結構節點產生可用於預測或分類未來觀察的決策樹。該方法通過在每一個步驟最大限度降低不純潔度，使用遞歸分區來將訓練記錄分割為組。如果樹狀結構中某個節點中 100% 的觀察值都的目標欄位的一個特定種類，那麼該節點將被認定為「純潔」。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

CHAID 使用卡方統計資料來產生決策樹，以確定最佳的分割。與 C&R Tree 和 QUEST 節點不同，CHAID 可產生非二元樹狀結構，表示部分分割有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

© Copyright IBM Corp. 1994, 2019

QUEST 節點可提供用於建立決策樹的二元分類法，此方法的設計目的是減少大型 C&R 樹狀結構分析所需的處理時間，同時也減少在分類樹狀結構方法中發現的趨勢以便偏愛容許有多個分割的輸入。輸入欄位可以是數值範圍（連續），但目標欄位必須是種類。所有分割都是二元的。

C5.0 節點建立決策樹或規則 集。該模型的工作原理是根據在每個層次提供上限資訊收穫的欄位分割樣本。目標欄位必須為種類欄位。容許進行多次多於兩個子群組的分割。

樹狀結構-AS 節點類似於現有的 CHAID 節點；但是，「樹狀結構-AS」節點旨在處理大量資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。此節點通過使用卡方統計量 (CHAID) 來識別最佳化分割，從而產生決策樹。 對 CHAID 的這一使用可產生非二元樹狀結構，意味著某些分割將具有兩個以上的分支。目標欄位和輸入欄位可以是數值範圍（連續）或類別欄位。「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查所有可能的分割，但計算時間較長。

「隨機樹狀結構」節點類似於現有 C&RT 節點；但是，「隨機樹狀結構」節點旨在處理大資料以建立單一樹狀結構，並在 SPSS Modeler 第 17 版中新增的輸出檢視器中顯示生成的模型。「隨機樹狀結構」節點將產生您可以對未來觀察進行預測或分類的決策樹。通過在每一個步驟最大限度降低不純潔度，此方法使用遞歸分區將訓練記錄分割為多個區段。如果樹狀結構中某個節點的全部觀察值都的目標欄位的一個特定種類，那麼系統會將該節點視為純潔。目標欄位和輸入欄位可以是數值範圍或類別（標稱、序數或旗標）欄位；所有分割都是二元的（只有兩個子群組）。

基於樹狀結構的分析的一般用法

下列為一些基於樹狀結構的分析的多個用法：

分區段：確定可能成為特定類別成員的人員。

分層：將觀察值指派至數個種類之一，如高風險群組、中等風險群組和低風險群組。

預測：建立規則並使用這些規則來預測未來事件。預測還可能意味著試圖將預測屬性與連續變數值相關聯。

資料縮減和變數篩選：從大型變數集中選取有用的預測值子集，以用於建置正式的參數模型。

互動識別：識別僅與特定子群組有關的關係，並在正式的參數模型中指定這些關係。

種類合併和帶狀化連續變數：以最小的資訊損失對群組預測值種類和連續變數進行重新編碼。

交互樹狀結構建置器

可以自動產生樹狀結構模型，由運用演算法在其中決定每一級的最佳分割，也可以套用交互樹狀結構建置器來控制模型的產生，並在儲存模型塊之前運用專業知識精練或簡化樹狀結構。

1. 建立串流並新增以下任一決策樹節點：C&R 樹狀結構、CHAID 或 QUEST。

   註：樹狀結構-AS 或 C5.0 樹狀結構都不支援互動式樹狀結構建立。

2. 開啟節點，並在「欄位」標籤上選取至欄位和預測值欄位，然後在視需要指定其他模型選項。如需特定指示，請參閱每一個樹狀結構建置節點的說明文件。
3. 在「建置選項」標籤的「目標」畫面上，選取**啟動互動式工作階段**。
4. 按一下**執行**以啟動樹狀結構建置器。

其中顯示了從根節點開始的目前樹狀結構。可以逐層次編輯和刪改樹狀結構，並在產生一個或多個模型之前存取增益、風險和相關的資訊。

備註(O)

- 使用 C&R 樹狀結構、CHAID 和 QUEST 節點時，模型中使用的所有序數欄位的儲存類型都必須是數值（而非字串）。必要的話，可以使用「重新分類」節點來對其進行轉換。
- 還可以選擇使用分割區欄位將資料分隔到訓練樣本和測試樣本中。
- 作為使用樹狀結構建置器的另一種替代方案，也可以直接從建模節點中產生樹狀結構模型或其他 IBM SPSS Modeler 模型。請參閱第 80 頁的『直接建立樹狀結構模型』主題，以取得更多資訊。

生成和刪改樹狀結構

使用樹狀結構建置器的「檢視器」標籤可以檢視從根節點開始的目前樹狀結構。

1. 要生成樹狀結構，請從功能表中選擇：

   樹狀結構 > **增長樹狀結構**

   系統將通過遞歸分割每個分支直到符合一個或多個停止準則來建立樹狀結構。然後，可根據使用的建模方法在每個分割處自動選取最合適的預測值。
2. 也可以選取**生成樹狀結構的第一層次**新增一個層次。
3. 要在一個特定節點下方新增分支，可選取該節點，然後選取**生成分支**。
4. 要選擇某個分割所使用的預測值，請選擇所需的節點，然後選擇**使用自訂分割生成分支**。請參閱第 72 頁的『定義自訂分割』主題，以取得更多資訊。
5. 要刪改分支，可選取某個節點，然後選取**移除分支**以清除所選取的節點。
6. 要移除樹狀結構的最底層次，可選取**移除第一層次**。
7. 僅對於 C&R 樹狀結構和 QUEST 樹狀結構，可選取**生成樹狀結構和刪改**根據成本複雜性演算法（此演算法可根據終端節點數目調整風險評估）進行刪改，通常會生成一個較簡單的樹狀結構。請參閱第 82 頁的『C&R 樹狀結構節點』主題，以取得更多資訊。

在檢視器標籤上讀取分割規則

在「檢視器」標籤上檢視分割規則時，方括弧表示相鄰值包括在範圍內，而括弧指出相鄰值從範圍內排除。因此，表示式 (23,37] 表示從 23（不含）至 37（含）；即，從 23 以上至 37。在「模型」標籤上，相同的條件會顯示為：

年齡 > 23 和年齡 <= 37

中斷樹狀結構的生成。要中斷樹狀結構生成作業（例如，如果此作業所用的時間比預期的長），可按一下工具列上的「停止於執行」按鈕。

圖 *28.* 「停止執行」按鈕

此按鈕僅在樹狀結構生成期間啟用。此按鈕將使目前的生成作業在其目前點停止，並保留所有已新增的節點，但不儲存所作的變更，也不關閉該視窗。樹狀結構建置器將保持開啟狀態，以便您根據需要產生模型、更新指引，或以適當的格式匯出輸出。

定義自訂分割

通過「定義分割」對話框，您可以選取預測值並為每個分割指定條件。

1. 在樹狀結構建置器的「檢視器」標籤上選擇一個節點，然後從功能表中選擇：

   樹狀結構 > **通過自訂分割增長分支**
2. 從下拉清單中選取所需的預測值，或按一下**預測值**按鈕，以檢視每個預測值的詳細資料。請參閱『檢視預測值詳細資料』主題，以取得更多資訊。
3. 可接受為每個分割選取的預設條件，或選取**自訂**為分割指定適當的條件。
   - 對於連續（數值型範圍）的預測值，可以使用**編輯範圍值**欄位以指定落在每個新節點中的值的範圍。
   - 對於種類預測值，可使用**編輯集合值**或**編輯次序值**欄位，以指定對映至每個新節點的特定值（如果是序數預測值，那麼指定值的範圍）。
4. 選取**生成**，使用選定的預測值重新生成分支。

在不考慮停止規則的情況下，通常可使用任何預測值分割樹狀結構。唯一的例外狀況情況是當節點是純節點（即所有觀察值都落在相同的目標類別中，從而沒有可分割的觀察值），或所選擇的預測值是常數（即沒有可分割的預測值）時無法分割樹。

遺漏值。僅對於 CHAID 樹狀結構，如果給定的預測值中有遺漏值，那麼可以在定義自訂分割時選擇將這些遺漏值分配給特定的子節點。（對於 C&R 樹狀結構和 QUEST，可使用代理項按演算法中定義的方式處理遺漏值。請參閱『分割的詳細資料和替代』主題，以取得更多資訊。)

檢視預測值詳細資料

「選取預測值」對話框中顯示了可用於目前分割的預測值（有時稱為「代替預測值」）的統計資料。

- 對於 CHAID 和 Exhaustive CHAID，列出了每個種類預測值的卡方測試統計資料；如果預測值是數值範圍類型，那麼顯示 *F* 統計資料。卡方測試統計資料可用來測量目標欄位與分割欄位的不相關程度。較高的卡方測試統計資料通常與較低的機率有關，這意味著兩個欄位間不相關的機率較低 - 表示此分割情況良好。這裡也將自由度併入在內，因為自由度考慮了以下事實，即與雙向分割相比，三向分割更易具有較高的統計資料和較低的機率。
- 對於 C&R 樹狀結構和 QUEST，顯示了每個預測值的改善值。如果使用此預測值，那麼改善值越大，母節點和子節點間的純度差異越大。（純節點指其中所有的觀察值都落在一個目標種類中的節點；樹狀結構中的雜質越少，此模型配適度資料的效果就越好。）換句話說，較高的改善值通常表示對此類型的樹狀結構進行了有用的分割。所使用的雜質測量在樹狀結構建立節點中指定。

分割的詳細資料和替代

可在「檢視器」標籤中選取任意節點，然後選取位於工具列右端的分割資訊按鈕檢視有關該節點的分割詳細資料。此時將顯示所使用的分割規則及相關的統計資料。對於 C&R 樹狀結構種類樹狀結構，將顯示改善值和關聯值。關聯值可用於測量替代與原始分割欄位間的一致性，其中「最佳」替代通常是對分割欄位模擬得最像的欄位。對於 C&R 樹狀結構和 QUEST，還將列出所有用於代替主要預測值的代理項。

要編輯選定節點的分割，可按一下位於替代窗格左端的圖示以開啟「定義分割」對話框。（作為快速鍵，可以在按一下圖示選取替代作為原始分割欄位之前，從清單中選取此替代。）

代理項。如果適用，那麼會針對所選節點顯示主要分割欄位的所有替代。替代是在給定記錄的主要預測值遺漏時使用的代理欄位。給定分割容許的上限替代數在樹狀結構建立節點中指定，但實際數量取決於訓練資料。一般來講，遺漏資料越多，可能使用的替代越多。對於其他決策樹模型，此標籤為空白。

註：要在模型中包含代理項，必須在訓練階段對其進行 ID。 如果訓練樣本沒有遺漏值，那麼不會 ID 任何替代；在正在測試或評分過程中遇到的具有遺漏值的所有記錄將自動落入記錄數最大的子節點。如果在正在測試或評分過程中預期出現遺漏值，請確保值在訓練樣本中也處於遺漏狀態。替代對於 CHAID 樹狀結構無法使用。

雖然 CHAID 樹狀結構中不使用替代，但當定義自訂分割時，仍可選擇將這些替代分配給特定的子節點。請參閱第 72 頁的『定義自訂分割』主題，以取得更多資訊。

自訂樹狀結構形視圖

在樹狀結構建置器的「檢視器」標籤中顯示目前的樹狀結構。預設情況下，將展開樹狀結構中所有的分支，但也可以按照需要展開和摺疊分支並自訂其他設定。

- 按一下母節點右下角落的減號 (**–**) 可以隱藏其所有子節點。按一下母節點右下角落的加號 (**+**) 顯示其子節點。
- 使用「檢視」功能表或工具列變更樹狀結構的方向（從上至下、從左至右或從右至左）。
- 按一下主工具列上的「顯示欄位與值標籤」按鈕以顯示或隱藏欄位和值標籤。
- 使用放大鏡按鈕放大或縮小視圖，或按一下工具列右端的樹狀對映按鈕以檢視完整樹狀結構的圖。
- 如果正在使用分割區欄位，那麼可在樹狀結構形視圖的訓練分割區和測試分割區之間進行交換（選擇 **視圖 > 分割區** ）。顯示正在測試樣本時，可以檢視但不能編輯樹狀結構。（將在視窗右下角落的狀態列中顯示目前分割區。）
- 按一下分割資訊按鈕（工具列最右側的 "i" 按鈕）以檢視目前分割的詳細資料。請參閱第 72 頁的『分割的詳細資料和替代』主題，以取得更多資訊。
- 將在每個節點中顯示統計資料、圖形或同時顯示兩者（請參見下方文）。

顯示統計資料和圖形

節點統計資料。對於種類目標欄位，每個節點中的表格顯示每個種類中的記錄數和百分比以及該節點代表的整個樣本的百分比。對於連續（數值型範圍）目標欄位，該表格顯示目標欄位的平均數、標準差、記錄數和預測值。

節點圖形。對於種類目標欄位，圖表為目標欄位的每個種類中的百分比長條圖。表格中每列的前面是一個顏色樣本，其對應的顏色代表該節點圖形中的每個目標欄位種類。對於連續（數值範圍）目標欄位，該圖表顯示節點中記錄的目標欄位的直方圖。

增益

「增益」標籤可顯示樹狀結構中所有終端節點的統計資料。增益可用於測量給定節點上的平均數或比例與整體平均值之間的差異大小。一般來說，此差異越大，作為決策工具的樹狀結構就越有效。例如，某個節點的指數或「提升」值為 148% 表示，該節點中的記錄落在目標種類中的可能性大概是其作為一個整體用於資料集的可能性的 1.5 倍。

針對指定了過適預防集的 C&R 樹狀結構及 QUEST 節點，會顯示兩組統計資料：

- 樹狀結構成長集 - 移除了過適預防集的訓練樣本
- 過適預防集

針對其他 C&R 樹狀結構及 QUEST 互動式樹狀結構，以及針對所有 CHAID 互動式樹狀結構，僅會顯示樹狀結構成長集統計資料。

通過「增益」標籤，您可以執行下列操作：

- 顯示每個節點統計資料、累加數統計資料或分位數統計資料。
- 顯示增益或利潤。
- 將視圖在表格和圖表間進行交換。
- 選取目標種類（僅種類目標）。
- 根據指數百分比對表格按遞增或遞減排序。如果顯示的是多個分割區的統計資料，那麼一律將排序套用於訓練樣本而不是正在測試樣本。

一般來說，在增益表格中選定的內容也會在樹狀結構形視圖中得到更新，反之亦然。例如，如果在表格中選取某個列，那麼也會在樹狀結構中已選取對應的節點。

分類增益

對於分類樹狀結構（指使用種類目標變數的樹狀結構），從增益指數百分比可看出每個節點上給定目標種類的比例與總比例間的差異有多大。

依次顯示節點統計資料

在此視圖的表格中，將為每個終端節點顯示一列。例如，如果直郵活動的總回應是 10%，但有 20% 的記錄落在節點 X 內並且做出積極的回應，那麼該節點的指數百分比應為 200%，表示該群組中的受訪者進行購買的可能性大概是總人數的兩倍。

針對指定了過適預防集的 C&R 樹狀結構及 QUEST 節點，會顯示兩組統計資料：

- 樹狀結構成長集 - 移除了過適預防集的訓練樣本
- 過適預防集

針對其他 C&R 樹狀結構及 QUEST 互動式樹狀結構，以及針對所有 CHAID 互動式樹狀結構，僅會顯示樹狀結構成長集統計資料。

節點。 目前節點的 ID（顯示在「檢視器」標籤上）。

節點：n。該節點中的總記錄數。

節點 (%)。資料集中所有落在此節點上的記錄的百分比。

增益：n。落在此節點上並且具有選定目標種類的記錄數。換句話說，在資料集的所有落在目標種類的記錄中，有多少記錄落在該節點？

增益 (%)。在整個資料集的目標種類中，所有落在此節點上的記錄的百分比。

回應 (%)。目前節點中的目標種類的記錄的百分比。該環境定義中的回應有時也稱為「命中數」。

指數 (%)。目前節點的回應百分比，以回應百分比相對於整個資料集的百分比表示。例如，指數值為 300% 表示該節點中的記錄落在目標種類中的可能性大概是其作為一個整體用於資料集的可能性的三倍。

累加統計資料

在累加視圖中，表格的每列顯示一個節點，但統計資料是累加的，並按指數百分比以遞增或遞減順序排序。例如，如果按遞減排序，那麼首先列出指數百分比最高的節點，並且接下來的列中的統計資料是對該列及上方的列的累加數。

隨著所新增節點的回應百分比越來越低，累加指數百分比將逐列降低。最後一列的累加指數一律是 100%，因為此時將併入整個資料集。

分位數

在此視圖中，表格中的每一列都代表一個分位數而不是節點。分位數可以是四分位數 (4)、五分位數 (5)、十分位數 (10)、二十分位數 (20) 或百分位數 (100)。如果需要多個節點以補足此百分比（例如，如果顯示四分位數時，而前兩個節點包含的觀察值不到所有觀察值的 50%），那麼可在一個分位數中列出多個節點。可以對表格的其餘部分進行累加，且與累加視圖的解釋方式相同。

分類利潤和投資回報率

對於分類樹狀結構，增益統計資料也可按利潤和投資返回率顯示。通過「定義利潤」對話框，您可以為每個種類指定營收和支出。

1. 在「增益」標籤上，按一下工具列上的「利潤」按鈕（標註為 $/$）以存取該對話框。
2. 輸入目標欄位的每個種類的營收和支出值。

例如，如果為每個客戶郵寄報價的成本是 $0.48，而從接受三個月預訂的積極回應中獲得的營收是 $9.95，那麼每個 *no* 回應將花費 $0.48，而每個 *yes* 回應將賺取 $9.47（按 9.95–0.48 計算）。

在增益表格中，**利潤**的計算方式為終端節點的每條記錄中的總營收減去支出。**ROI** 是某個節點的總利潤除以總支出得到的值。

備註(O)

- 利潤值僅影響在增益表格中顯示的平均利潤和投資回報率，可以明確檢視統計資料，尤其適合檢視利潤。但是不會影響基本樹狀結構的模式結構。不應將利潤與誤分類成本相混淆，誤分類成本在樹狀結構建立節點中指定，且可化為模型中的側邊欄（作為避免高成本錯誤的一種方式）。
- 在兩個交互樹狀結構建立階段作業之間不會保留利潤說明。

迴歸方法增益

對於迴歸方法樹狀結構，可以選擇依次顯示節點視圖、累加節點視圖和分位數視圖。表格中可顯示平均值。只有在分位數視圖中才可使用圖表。

增益圖表

在「增益」標籤上，圖表可作為表格的替代項顯示。

1. 在「增益」標籤上，選取「分位數」圖示（工具列從左數第三個圖示）。（對於依次顯示節點統計資料或累加統計資料，不可使用圖表。）
2. 選取「圖表」圖示。
3. 按照需要從下拉清單中選取所顯示的單位（百分位數、十分位數等等）。
4. 選取**提昇**、**回應**或**提升**變更所顯示的測量。

增益圖表

收益圖表繪製的是表格中 增益 *(%)* 欄值的統計圖。增益定義為每個增量中的命中數相對於樹狀結構中總命中數的比例，使用的方程式如下：

(增量中的命中數 / 總命中數) x 100%

該圖有效協助您需要撒出多大範圍的網絡，才能獲取樹狀結構中所有命中數的給定百分比。對角線繪製的是整個樣本的預期回應（如果未使用模型的話）。在此情況下，回應率將不變，因為某個人員的回應可能與另一個

人員一樣。若要將收益翻倍，您需要詢問兩倍的人員。曲線指出在您僅併入根據收益排名在較高百分位數的人員時，您的回應可以提升的幅度。例如，併入前 50% 可能讓您得到 70% 以上的正向回應。曲線越陡峭，增益越高。

提升圖表

提升圖表對表格中 指數 *(%)* 欄中的值進行了繪製。此圖表將每個增量中具有積極響應的記錄的百分比與訓練資料集中具有積極響應的記錄的總百分比作了比較，其方程式為：

（增量中的命中數/增量中的記錄數）/（總命中數/總記錄數）

回應圖表

回應圖表表格對表格中 回應 *(%)* 欄中的值進行了繪製。回應是增量中具有積極回應的記錄的百分比，其方程式為：

（增量中具有積極回應的記錄/增量中的記錄） x 100%

基於增益的選擇

通過「基於增益的選擇」對話框，您可以根據指定的規則或臨界值來自動選擇具有最佳（或最差）增益的終端節點。然後可以根據該選擇產生一個選取節點。

1. 在「增益」標籤上，選擇依次顯示節點視圖或累加視圖，然後選擇該選擇所基於的目標種類。（該選擇基於目前的表格顯示，無法使用於分位數視圖。）
2. 從「增益」標籤的功能表中選擇以下項：

   編輯 > **選擇終端節點** > **基於增益的選擇**

   僅「選取」。可以選取相符的節點或不相符的節點 - 例如，選取前 100 條記錄以外的所有記錄。

   增益資訊比對。根據目前目標種類的增益統計資料來相符節點，包含：
   - 其提昇、回應或提升（指數）與指定的臨界值相符的節點，例如，回應大於或等於 50%。
   - 基於目標種類的增益的頂部 *n* 個節點。
   - 上限為指定記錄數的頂部節點。
   - 上限為指定訓練資料百分比的頂部節點。
3. 按一下**確定**更新「檢視器」標籤上的選擇。
4. 要根據「檢視器」標籤上的目前選擇新建「選取」節點，請從「產生」功能表中選擇**選取節點**。請參閱第 79 頁的『產生過濾節點和選取節點』主題，以取得更多資訊。

註：由於實際上選擇的是節點而不是記錄或百分比，因此無法始終獲取與選取準則完全相符的結果。系統選取上限為指定等級的完整節點。例如，如果選取頂部 12 個觀察值，而第一個節點中有 10 個觀察值，第二個節點中有 2 個觀察值，那麼將只選取第一個節點。

風險

風險指任意類別上誤分類的機率。「風險」標籤可顯示某點的風險評估和（分類輸出的）誤分類表。

- 對於數值預測，風險是每個終端節點上的合併變異數評估。
- 對於分類預測，風險是錯誤分類觀察值的比例，可根據任意先驗分佈或誤分類成本進行調整。

儲存樹狀結構模型和結果

可以用以下多種方式儲存或匯出交互樹狀結構建立階段作業的結果：

- 基於目前樹狀結構產生模型（ **產生 > 產生模型** ）。
- 儲存用於生成目前樹狀結構的指引。下次執行樹狀結構建立節點時，將自動重新生成目前樹狀結構（包含已定義的任何自訂分割）。
- 匯出模型、增益和風險資訊。請參閱第 79 頁的『匯出模型、增益和風險資訊』主題，以取得更多資訊。

通過樹狀結構建置器或樹狀結構模型塊，可以執行下列操作：

- 根據目前的樹狀結構產生過濾節點或選取節點。如需相關資訊，請參閱第 79 頁的『產生過濾節點和選取節點』。
- 產生一個規則集塊，該節點將樹狀結構結構代表成一組定義了樹狀結構的終端機分支的規則。如需相關資訊，請參閱第 80 頁的『從決策樹中產生規則集』。
- 此外，還可以按 PMML 格式匯出模型（僅限樹狀結構模型塊）。如需相關資訊，請參閱第 34 頁的『模型選用區』。如果模型包含任何自訂分割，那麼不會在匯出的 PMML 中保留此資訊。（保留分割，但不保留它是自訂分割而不是通過演算法選擇的分割這一事實。）
- 基於目前樹狀結構的所選部分產生圖表。請注意，此操作僅對附加到串流中其他節點的區塊有效。如需相關資訊，請參閱第 105 頁的『產生圖形』。

註：無法儲存交互樹狀結構自身。為了避免丟失所執行的操作，請在關閉樹狀結構建置器視窗之前產生模型和/或更新樹狀結構指引。

從樹狀結構建置器產生模型

要基於目前樹狀結構產生模型，可從樹狀結構建置器功能表中選擇以下項：

產生 > **模型**

在「產生新模型」對話框中，您可以從下列選項中進行選擇：

模型名稱。可以指定自訂名稱或根據建模節點的名稱自動產生模型名稱。

建立節點位置。 可以在**畫布**、**GM 選用區**或**兩者**上新增節點。

包含樹狀指令。要在產生模型中包含來自目前樹狀結構的指引，選取此選項。通過此選項，您可以根據需要重新產生樹狀結構。請參閱『樹狀結構增長指引』主題，以取得更多資訊。

樹狀結構增長指引

對於 C&R 樹狀結構、CHAID 和 QUEST 模型，樹狀結構指參照於指定生成樹狀結構（一次一級）的條件。每當從節點中啟動交互樹狀結構建置器時，都會套用指引。

- 指引可作為一種最安全的方法用來重新產生在以前的交互階段作業中建立的樹狀結構。請參閱第 79 頁的『更新樹狀結構指引』主題，以取得更多資訊。也可以手動編輯指引，但操作時需要格外小心。
- 指引與其所說明的樹狀結構結構高度相關。因此，對原始資料或建模選項的任何變更都可能會導致以前有效的一組指引失效。例如，如果 CHAID 演算法基於更新的資料將雙向分割變更為三向分割，那麼基於以前的雙向分割的所有指引都將失效。

註：如果選擇直接產生模型（不使用樹狀結構建置器），那麼將忽略所有的樹狀結構指引。

編輯指引

1. 要檢視或編輯已儲存的指引，請開啟樹狀結構建立節點並選取「建置選項」標籤的「目標」畫面。

2. 選取**啟動互動式工作階段**以啟用控制項，選中**使用樹狀結構指引**，然後按一下**指引**。

指令語法

指引可指定從根節點開始生成樹狀結構的條件。例如，生成樹狀結構的第一層次：

```
成長節點索引 0 子項 1 2
```

由於未指定任何預測值，演算法將選擇最佳分割。

注意，一律必須在根節點 (Index 0) 上進行第一次分割，且必須指定兩個子節點的索引值（在本例中為 1 和 2）。除非首先生成已建立節點 2 的根節點，否則指定 Grow Node Index 2 Children 3 4 是無效的。

要生成樹狀結構，請使用：

```
增長樹狀結構(T)
```

要生成並刪改樹狀結構（僅限 C&R 樹狀結構），請使用：

```
Grow_And_Prune Tree
```

要為連續預測值指定自訂分割，請使用：

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

要分割具有兩個值的列名預測值，請使用：

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

要分割具有多個值的列名預測值，請使用：

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

要分割序數預測值，請使用：

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

註：指定自訂分割時，欄位名稱和值（EDUCATE、GENDER 和 CHILDS 等）區分觀察值。

CHAID 樹狀結構的指引

CHAID 樹狀結構的指引對資料或模型中的變更非常敏感，因為這些指引與 C&R 樹狀結構和 QUEST 中的不同，它們並非只能使用二元分割。例如，下面的語法看起來很有效，但如果演算法將根節點分割為兩個以上的子節點時，這些語法將失效：

```
成長節點索引 0 子項 1 2
Grow Node Index 1 Children 3 4
```

對於 CHAID，節點 0 可能具有 3 個或 4 個子節點，這種情況將使上述第二行語法失效。

在 Script 中使用指引

也可使用三重引號將指引內含到 Script 中。

更新樹狀結構指引

要保留在交互樹狀結構建立階段作業中執行的操作，可以儲存用於產生目前樹狀結構的指引。與儲存無法進一步進行編輯的模型塊不同的是，您可以通過儲存指令來按樹狀結構的現行狀態重新產生該樹狀結構以進行進一步編輯。

要更新指引，請從樹狀結構建置器功能表中選擇以下項：

檔案 > **更新指引**

指引儲存在用於建立樹狀結構（C&R 樹狀結構、QUEST 或 CHAID）的建模節點中，並可用於重新產生目前樹狀結構。請參閱第 77 頁的『樹狀結構增長指引』主題，以取得更多資訊。

匯出模型、增益和風險資訊

可以從樹狀結構建置器中根據需要以文字、HTML 或影像格式匯出模型、增益和風險統計資料。

1. 在樹狀結構建置器視窗中，選取要匯出的標籤或視圖。
2. 在功能表上，選擇：

   檔案 > **匯出**
3. 根據需要選取**文字**、**HTML** 或**圖表**，並從子功能表中選取要匯出的特定項目。

在適用的情況下，匯出基於目前的選擇。

匯出文字或 HTML 格式。您可以為訓練分割區或測試分割區（如果已定義）匯出增益統計資料或風險統計資料。匯出基於「增益」標籤上的目前的選擇 - 例如，可以選擇依次顯示節點統計資料、累加統計資料或分位數統計資料。

匯出圖表。可以匯出在「檢視器」標籤上顯示的目前樹狀結構，或為訓練分割區或測試分割區（如果已定義）匯出收益圖表。可用的格式包含 *.JPEG*、 *.PNG* 和 *.BMP*。對於增益，匯出基於「增益」標籤上的目前的選擇（僅當顯示圖表時可用）。

產生過濾節點和選取節點

在樹狀結構建置器視窗中，或在瀏覽決策樹模型塊時，從功能表中選擇以下項：

產生 > **「過濾器」節點**

或

> **「選取」節點**

過濾節點。 產生用於過濾目前樹狀結構不使用的任何欄位的節點。此方法可以快速削減資料集，使其僅包含那些演算法選取為重要欄位的欄位。如果此決策樹節點的上游存在「類型」節點，那麼「過濾」模型塊將傳送所有角色為目標的欄位。

選取節點。 產生用於選取所有落在目前節點中的記錄的節點。此選項需要在「檢視器」標籤中選取一個或多個樹狀結構分支。

該模型塊位於串流畫布中。

從決策樹中產生規則集

產生的規則集模型塊可作為定義樹狀結構的終端機分支的一組規則來代表樹狀結構的結構。 規則集通常可以保留來自某個完整決策樹（但具有的模型不太複雜）的大部分重新資訊。最重要的區別是，套用規則集時，可以為任意特定記錄套用多個規則，也可以不套用任何規則。 例如，可以看到所有預測結果為否的規則，緊隨其後是所有預測為是的規則。 如果套用多個規則，那麼每個規則將根據與此規則關聯的信賴度獲得一個加權「投票」，並通過組合套用到所討論記錄的所有規則的加權投票來確定最終的預測。 如果沒有套用任何規則，則會將預設預測指派給記錄。

註：對規則集進行評分時，您可能會注意到此評分相比於針對樹狀結構的評分存在差異；這是由於樹狀結構種每個終端機分支都是獨立評分的。 如果在資料中存在遺漏值，那麼您可以明顯注意到此差異。

僅可從具有種類目標欄位的樹狀結構（不是迴歸方法樹狀結構）中產生規則集。

在樹狀結構建置器視窗中，或在瀏覽決策樹模型塊時，從功能表中選擇以下項：

產生 > **規則集**

規則集名稱：指定新的規則集模型塊的名稱。

節點的建立位置：控制項新的規則集模型塊的位置。選取**畫布**、**GM 選用區**或**兩者**。

最少實例數：指定在規則集模型塊中保留的最小實例（已應用規則的記錄數）。 支援小於指定值的規則將不會包含在新的規則集中。

最小信賴度：指定規則集模型塊中要保留規則的最小信賴度。 信賴度少於指定值的規則將不會包含在新的規則集中。

直接建立樹狀結構模型

作為使用互動式樹狀結構建置器的替代方案，您可以在執行串流時直接從節點建立決策樹模型。這與大多數其他模型建立節點相一致。對於交互樹狀結構建置器所不支援的 C5.0 樹狀結構模型和樹狀結構-AS 模型來說，這是唯一可以使用的方法。

1. 建立串流並新增其中一個決策樹節點 - C&R 樹狀結構、CHAID、QUEST、C5.0 或樹狀結構-AS。
2. 對於 C&R 樹狀結構、QUEST 或 CHAID，在「建置選項」標籤的「目標」畫面上，選擇一個主目標。如果您選擇**建立單一樹狀結構**，請確保將**模式**設為**產生模型**。

   對於 C5.0，在「模型」標籤上，將**輸出類型**設為**決策樹**。

   對於樹狀結構-AS，在「建置選項」標籤的「基本」窗格上，選取**樹狀結構增長演算法**類型。
3. 選取至欄位和預測值欄位，並在視需要指定其他模型選項。如需特定指示，請參閱每一個樹狀結構建置節點的說明文件。
4. 執行串流以產生模型。

關於建立樹狀結構的備註

- 使用此方法產生樹狀結構時，會忽略樹狀結構增長指引。
- 無論使用交互模式還是直接模式，這兩種創建決策樹的方法最終都會產生相似的模型。只需考慮希望在此過程中執行多大程度的控制。

決策樹節點

IBM SPSS Modeler 中的決策樹節點提供對下列樹狀結構建立演算法的存取：

- C&R 樹狀結構
- QUEST
- CHAID
- C5.0
- Tree-AS
- 隨機樹狀結構

請參閱第 69 頁的『決策樹模型』主題，以取得更多資訊。

這些演算法有一點很相似：它們可以通過將資料遞歸分割為越來越小的子群組來構造樹狀結構。但是，存在一些重要的差異。

輸入欄位。輸入欄位（預測值）可以是下列任何類型（測量層次）：連續、種類、旗標、列名或序數。

目標欄位。僅可指定一個目標欄位。對於 C&R 樹狀結構、CHAID、樹狀結構-AS 和隨機樹狀結構，目標可以為連續、種類、旗標、列名或序數。對於 QUEST，目標字段可以是種類、旗標或列名。對於 C5.0，目標字段可以是旗標、列名或序數。

分割類型。C&R 樹狀結構、QUEST 和隨機樹狀結構僅支援二元分割（即，每個樹狀結構節點不能分割成兩個以上的分支）。相比之下，CHAID、C5.0 和樹狀結構-AS 支援一次分割為兩個以上的分支。

用於分割的方法。不同演算法在用於確定分割的準則上有所不同。C&R 樹狀結構在預測種類輸出時使用離差測量（預設為 Gini 係數，不過您可以進行變更）。對於連續目標，將使用最小平方離差法。CHAID 和樹狀結構-AS 使用卡方測試；QUEST 將卡方測試用於種類預測值並將變異數分析用於連續輸入。 對於 C5.0，將使用資訊論測量，即資訊增益率。

遺漏值處理。所有演算法均容許預測值欄位遺漏值，但它們使用不同的遺漏值處理方法。C&R 樹狀結構和 QUEST 根據需要使用替代預測欄位，以確保具有遺漏值的記錄在訓練期間通過樹狀結構。CHAID 將遺漏值分為個別的種類，並使它們可以用於樹狀結構建立。C5.0 使用分離方法，該方法將記錄的分離部分從節點（此節點中的分割取決於具有遺漏值的欄位）向下傳送到樹狀結構的各個分支。

刪改。C&R 樹狀結構、QUEST 和 C5.0 提供的選項允許完全生成樹狀結構，然後刪除對於樹狀結構的精確性沒有顯著影響的底層次分割以進行刪改。但是，所有決策樹演算法都容許控制下限子群組大小，這有助於避免出現資料記錄較少的分支。

交互樹狀結構建立。C&R 樹狀結構、QUEST 和 CHAID 提供了啟動互動式階段作業的選項。通過此選項，您可以建立樹狀結構（一次一級），編輯分割並在建立模型之前對樹狀結構進行刪改。C5.0、Tree-AS 和隨機樹狀結構沒有交互選項。

事前機率。C&R 樹狀結構和 QUEST 支援在預測種類目標欄位時為種類指定事前機率。事前機率是對整體（從中抽取訓練資料）中每個目標分類的總相對次數的估計。換而言之，事前機率是對預測值有任何瞭解之前對每個可能的目標值的機率估計。CHAID、C5.0、樹狀結構-AS 和隨機樹狀結構不支援指定事前機率。

過長。不適用於樹狀結構-AS 或隨機樹狀結構。對於具有種類目標欄位的模型，決策樹節點提供了以過長形式建立模型的選項，這有時比複合決策樹更容易解釋。對 C&R 樹狀結構、QUEST 和 CHAID，您可以從互動式階段作業中產生過長；對於 C5.0，可以在建模節點上指定此選項。另外，所有決策樹模型都支持根據模型塊產生過長。請參閱第 80 頁的『從決策樹中產生規則集』主題，以取得更多資訊。

C&R 樹狀結構節點

分類和迴歸方法 (C&R) 樹狀結構節點是一種基於樹狀結構的分類和預測方法。與 C5.0 類似，此方法可使用遞歸分區將訓練記錄分割為具有相似輸出欄位值的區段。首先，「C&R 樹狀結構」節點通過檢查輸入欄位來尋找最佳分割（以分割所引起的雜質指標下降情況進行測量）。分割可定義兩個子群組，其中每個子群組隨後又被分割為兩個子群組，依此類推，直到觸發其中一個停止準則為止。所有分割都是二元的（只有兩個子群組）。

刪改

C&R 樹狀結構允許您先生成樹狀結構，然後根據成本複雜性演算法（該演算法可根據終端節點數目調整風險評估）刪改此樹狀結構。通過此方法（此方法可以使樹狀結構在長大後再根據更複合的準則進行刪改）可生成交叉驗證內容較好的小型樹狀結構。增加終端節點數目通常會降低目前（訓練）資料的風險，但當模型擴展為適用不可見資料時，實際的風險可能會更大。假設在一種極端的情況下，訓練集合中的每條記錄都有一個個別的終端節點。此時的風險評估可能是 0%，因為每條記錄都落在了它自己的節點內，但對於不可見的（正在測試）資料，誤分類的風險幾乎肯定大於 0。成本複雜性測量將試圖彌補這種風險。

範例。 某有線電視公司委託進行行銷研究，以確定有意預訂有線電視互動新聞服務的用戶。使用研究中得來的資料可建立串流，其中的目標欄位為有意預訂有線電視服務，預測值欄位則包含年齡、性別、教育、收入種類、每天看電視的時間和子女數。通過將 C&R 樹狀結構節點套用到串流，您可以預測回應並對回應進行分類以取得行銷活動的最高回應率。

需求。 要訓練 C&R 樹狀結構模型，需要一個或多個輸入欄位和唯一一個目標欄位。目標欄位和輸入欄位可以是連續（數值範圍）或類別欄位。會忽略設為兩者 或無 的欄位。對於模型中使用的欄位，必須將它們的類型完全實例化，並且模型中使用的所有序數（排定次序的集）欄位的儲存類型必須是數值（而不是字串）。必要的話，可以使用「重新分類」節點來對其進行轉換。

強度。 遇到缺少資料及大量欄位等問題時，C&R 樹狀結構模型的表現十分穩健。它們通常不需要很長的訓練時間來進行估計。另外，C&R 樹狀結構模型似乎比某些其他模型類型更易於理解 - 衍生自模型的規則解譯起來更簡明易懂。與 C5.0 不同的是，C&R 樹狀結構可同時相容連續欄位和種類輸出欄位。

CHAID 節點

CHAID（或卡方自動互動偵測）是一種分類方法，可透過使用卡方統計學來識別最佳分割以建置決策樹。

CHAID 首先會檢查每個輸入欄位與結果之間的交叉表，然後使用卡方獨立性測試來測試顯著性。如果這些關係中有多個關係在統計上顯著，則 CHAID 將選取最顯著（*p* 值最小）的輸入欄位。如果一個輸入有兩個以上的種類，則會比較這些種類，且結果無差異的種類會收合在一起。這是透過順利結合顯示最不明顯之差異的種類配對來完成。如果在指定的測試層次所有剩餘的種類都不同，則這個種類合併程序會停止。若為標稱輸入欄位，可以合併任何種類；若為序數集，則只能合併連續的種類。

「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查每個預測值所有可能的分割，但計算時間較長。

需求。 目標及輸入欄位可以為連續欄位或類別欄位；在每一個層次上，節點可分割成兩個以上子群組。模型中使用的任何序數欄位都必須具有數值儲存空間（而不是字串）。必要的話，可以使用「重新分類」節點來對其進行轉換。

強度。 與 C&R Tree 和 QUEST 節點不同，CHAID 可產生非二元樹狀結構，表示部分分割有兩個以上的分支。因此，相較於二元增長方法而言，它將建立更寬的樹狀結構。 CHAID 適用於所有類型的輸入，並且它同時接受觀察值加權與頻率變數。

QUEST 節點

QUEST，或稱快速、無偏倚、高效率統計樹狀結構，是一種用於建立決策樹的二元分類法。開發此方法的一個主要原因是減少包含很多變數或觀察值的大型 C&R 樹狀結構分析所需的處理時間。QUEST 的第二個目的是減少在分類樹狀結構方法中發現的趨勢以便偏愛容許有多個分割的輸入，即連續（數值型範圍）輸入或具有多個種類的輸入。

- 根據顯著性測試，QUEST 使用一系列規則來評估節點上的輸入欄位。為了進行選擇，可能需要對節點的每個輸入執行一次測試。與 C&R 樹狀結構不同，所有的分割都不用檢查，而與 C&R 樹狀結構和 CHAID 都不同的是，在評估輸入欄位以供選擇時不會測試種類組合。因此可加快分析的速度。
- 通過使用由目標種類形成的群組中選定的輸入來執行二次判別分析可以確定分割。而且，與窮舉搜尋（C&R 樹狀結構）相比，此方法確定最佳化分割的速度得到了改善。

需求。 輸入欄位可以是連續（數值型範圍）的，但目標欄位必須是種類的。所有分割都是二元的。不能使用加權欄位。模型中使用的所有序數（排定次序的集）欄位的儲存類型都必須是數值型類型（不是字串）。必要的話，可以使用「重新分類」節點來對其進行轉換。

強度。 與 CHAID 相似但與 C&R 樹狀結構不同的是，QUEST 可使用統計測試確定是否使用輸入欄位。QUEST 還可以將輸入的選擇與分割問題分開啟，分別為其套用不同的準則。不過在 CHAID 中，確定變數選擇的統計測試結果還可生成分割。同樣，C&R 樹狀結構也可採用雜質變更測量在選取輸入欄位的同時確定分割。

決策樹節點欄位選項

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。

預測值（輸入）。 選擇一或多個欄位作為預測的輸入。

分析加權。（僅限CHAID、C&RT 和樹狀結構-AS）要使用欄位作為觀察值加權，在此處指定。 觀察值加權是用來說明輸出欄位不同等級間的變異數差異。請參閱第 28 頁的『使用頻率和加權欄位』主題，以取得更多資訊。

決策樹節點建置選項

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。

該標籤包含數個不同窗格，您可以在其中設定特定於您模型的自訂作業。

「決策樹」節點 - 目標

對於「建置選項」標籤上的「目標」畫面中的 C&R 樹狀結構節點、QUEST 節點和 CHAID 節點，您可以選擇是建立新模型還是更新現有模型。 還可以設定節點的主要目標：建立標準模型、建立精確度或穩定性增強的模型，或者建立用於大型資料集的模型。

您現在要進行什麼作業？

建立新模型。（預設）每次執行包含此建模結點的串流時，就會建立一個全新模型。

繼續訓練現有模型。 依預設，每次執行建模節點時，將建立一個全新的模型。如果已選取該選項，那麼會繼續訓練該節點成功生成的最後一個模型。這樣就可以在無需存取原始資料的情況下更新或重新整理現有的模型，並可能會顯著提升效能，這是因為只有新的或更新後的記錄被反饋到串流中。系統會使用建模節點來儲存上一個模型的詳細資料，即使資料流或「模型」色板中已不再提供上個模型塊，也能使用此選項。

註：只有在您選取**建立單個樹狀結構**（對於 C&R 樹狀結構、CHAID 和 QUEST）、**建立標準模型**（對於類神經網路和線性）或**為超大型資料集建立模型**作為目標時，此選項才會被啟動。

您的主要目標是什麼？

- **建立單一樹狀結構。**建立單個標準決策樹模型。通常，與使用其他目標選項構建的模型相比，標準模型更易於說明並可以更快速地進行分數。

  註：對於分割模型，要將此選項與**繼續訓練現有模型**配合使用，您必須連接至 Analytic Server。

  眾數。 指定用來建置模型的方法。**產生模型**可在執行串流時自動建立模型。**啟動互動式階段作業**將開啟樹狀結構建置器，您可以通過該建置器在建立模型塊之前建立樹狀結構（一次一級）、編輯分割並根據需要進行刪改。

  使用樹狀結構指引。選中此選項可以指定從節點中產生交互樹狀結構時套用的指引。例如，可以指定第一級分割和第二級分割，當啟動樹狀結構建置器時會自動套用這些分割。還可以儲存交互樹狀結構建立階段作業中的指引，以便將來重新建立樹狀結構時使用。請參閱第 79 頁的『更新樹狀結構指引』主題，以取得更多資訊。
- **強化模式準確性 (Boosting)。** 如果您要使用一種名為**增強**的特殊方法來提高模型準確率，請選擇此項。增強的工作原理是在序列中建立多個模型。第一個模型以平常方式建置。然後，會建置第二個模型以便讓它著重於由第一個模型分類錯誤的記錄。然後會建置第三個模型以著重於第二個模型的錯誤，依此類推。最後會對觀察值分類，方法對其套用整個模型集，使用加權投票程序將個別的預測結合至一個整體預測。增強方法可以顯著提高決策樹模型的精確度，但也需要更長的訓練時間。
- **強化模式穩定性 (Bagging)。** 如果您要使用一種名為 **組裝**（Bootstrap 彙總）的特殊方法來提高模型穩定性並避免過度配適，請選擇此項。此選項將建立多個模型並將其進行組合，以獲取更加可靠的預測。與標準模型相比，使用此選項獲取的模型建立和分數所花費的時間更長。
- **為大型資料集建立模型。**如果您的資料集過大，而無法使用任何上述目標選項建立模型，請選擇此項。此選項用於將資料劃分為更小的資料區塊，並對每個區塊建立一個模型。然後，將自動選取最準確的模型並將它們合併到單一模型塊中。如果您在此畫面上選取**繼續訓練現有模型**選項，可以執行增量式模型更新。

  註：此選項適合大型資料集，需要連接至 IBM SPSS Modeler Server。

「決策樹」節點 - 基本

指定關於要如何建置決策樹的基本選項。

樹狀結構增長演算法（僅限 CHAID 和樹狀結構-AS）選擇您要使用的 **CHAID** 演算法類型。 **詳盡的 CHAID** 是對 CHAID 的修改，它會徹底檢查每個預測值所有可能的分割，但計算時間較長。

樹狀結構深度上限指定根節點以下方的上限級數（對樣本進行遞歸分割的次數）。 預設值是 5；選擇**自訂**，並輸入值以指定其他級數。

刪改（僅限 C&RT 和 QUEST）

對樹狀結構進行刪改以避免過度配適刪改包括刪除對於樹狀結構的精確度沒有顯著影響的底層次分割。 刪改有助於簡化樹狀結構，使樹狀結構更容易被理解，在某些情況下還可提高廣義性。如果需要未刪改的完整樹狀結構，請取消選取此選項。

- **設定風險上限差分（在標準誤範圍內）**通過此選項，您可以指定更自由的刪改規則。 標準誤差規則使演算法可以選取最簡單的樹狀結構，該樹狀結構的風險評估接近於（但也可能大於）風險最小的子樹狀結構的風險評估。該值表示已刪改樹和風險最小的樹狀結構之間所容許的風險評估差異大小。例如，如果指定 2，那麼將選取其風險評估（2 × 標準誤差）大於完整樹狀結構的風險評估的樹狀結構。

代理數上限。 代理項是用於處理遺漏值的方法。對於樹狀結構中的每個分割，演算法都會對與選定的分割欄位最相似的輸入欄位進行識別。這些被識別的欄位就是該分割的替代。 當必須對某個記錄進行分類，但此記錄中的分割欄位中具有遺漏值時，可以使用替代欄位的值填補此分割。增加此設定將可以更加靈活地處理遺漏值，但也會導致記憶體使用量和訓練時間增加。

「決策樹」節點 - 中止規則

這些選項可控制樹狀結構的構建方式。停止規則可確定何時停止分割樹狀結構的特定分支。設定下限分支大小可阻止通過分割建立非常小的子群組。如果節點（父級）中要分割的記錄數少於指定值，那麼**上層分支中的最小記錄**將阻止進行分割。 如果由分割建立的任何分支（子級）中的記錄數少於指定值，那麼**子分支中的最小記錄**將阻止進行分割。

- **使用百分比**：按總訓練資料的百分比指定大小...。
- **使用絕對值**：按絕對記錄數指定大小...。

「決策樹」節點 - 總體

系統在「目標」中要求 boosting、bagging 或極大資料集時，這些設定會決定所發生的集合行為。系統會忽略無法套用至所選目標的選項。

Bagging 與極大資料集。 系統執行集合評分時，可使用此規則來合併基底模型的預測值，以運算集合分數值。

- **種類目標的預設組合規則**。可以通過投票、最高機率或最高平均值機率來對種類目標的總體預測值進行已結合。**投票**會選取所有基底模式中最常擁有最高機率的類別。**最高機率**會在所有基底模式中選取達到單一最高機率的類別。**最高平均數機率**會在平均計算所有基底模型的類別機率時，選取具有最高值的類別。
- **連續目標的預設合併規則**。 系統會使用基底模型預測值的平均數或中位數，來合併連續目標的集合預測值。

請注意，若目標是用於強化模式準確性，則系統會忽略合併規則選擇。Boosting 會一律使用大部分的加權投票來為類別目標評分，並且使用加權中位數來為連續目標評分。

Boosting 與 Bagging。 當目標用於強化模式準確性或穩定性時，可指定欲建立的基底模式數目；若為 bagging，則此為重複取樣範例的數目。其應為正整數。

C&R 樹狀結構和 QUEST 節點 - 成本和先驗

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。

先驗

通過這些選項可以在預測種類目標欄位時為種類指定事前機率。**事前機率**是對整體（從中抽取訓練資料）中每個目標分類的總相對次數的估計。換句話說，事前機率是對預測值有任何瞭解 之前對每個可能的目標值的機率估計。有三種方法用來設定先驗概率：

- **基於訓練資料。** 此為預設值。事前機率基於訓練資料中分類的相對次數。
- **對所有類別相等。** 所有種類的事前機率都定義為 $1/k$，其中 k 是目標分類數。
- **自訂。** 您可以自行指定事前機率。對於所有類別，都將事前機率的初值設定為相等。可以將單個分類的機率調整為使用者定義的值。要調整特定分類的機率，可在表格中對應於所需分類的機率 Cell 中，先清除其內容，然後輸入所需的值。

所有分類的事前機率之和應為 1.0（ **機率限制** ）。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。此自動調整操作可在強制執行機率限制時保留分類中的比例。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。

使用錯誤分類成本調整先驗概率。通過此選項可以根據錯誤分類成本（在「成本」標籤中指定）調整先驗概率。從而可為使用兩分雜質測量的樹狀結構將損失資訊直接合併到樹狀結構生成過程中。（未選取此選項時，損失資訊僅用於為基於兩分測量的樹狀結構分類記錄和計算風險評估。）

CHAID 節點 - 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。

C&R 樹狀結構節點 - 進階

進階選項可讓您細部調整樹狀結構建置程序。

下限雜質改變。指定雜質中的變更下限以便在樹狀結構中建立新的分割。**雜質**是指由樹狀結構定義的子群組在每個群組中所具有的輸出欄位值的廣度。對於種類目標，如果節點中 100% 的觀察值都落在目標欄位的特定種類中，那麼該節點被認為是「純節點」。樹狀結構建立的目的是建立具有相似輸出值的子群組 - 換句話說，是為了減少每個節點中的雜質。如果某個分支的最佳分割按少於指定值的數量減少雜質，那麼不會進行此分割。

種類目標的雜質測量。對於種類目標欄位，指定用於測量樹狀結構的雜質的方法。（對於連續目標，將忽略此選項，而一律會使用**最小平方差**雜質測量。）

- **吉尼**是基於分支的種類會員資格機率的一般雜質測量。
- **兩分**是強調二元分割並更有可能導致從分割中生成大小近似相同的分支的雜質測量。
- **依序**新增了額外的限制，即只有相鄰的目標類別才可以群組成一群組，此選項僅適用於依序目標。如果對於列名目標已選取此選項，將預設使用標準的兩分測量。

過適預防集。 演算法會在內部將記錄分割為模型建置集和過適預防集，後者是一組獨立的資料記錄，用來追蹤訓練期間的錯誤以避免方法對資料中的機會變異進行建模。指定記錄百分比。預設值是 30。

複製結果。 設定隨機種子可讓您抄寫分析。請指定一個整數，或是按一下「**產生**」以建立介於 1 和 2147483647 之間（含）的虛擬亂數整數。

QUEST 節點 - 進階

進階選項可讓您細部調整樹狀結構建置程序。

分割的顯著性層次。指定用於分割節點的顯著性水準 (Alpha 值)。該值必須位於 0 和 1 之間。值越小，生成的樹狀結構的節點也會越少。

過適預防集。 演算法會在內部將記錄分割為模型建置集和過適預防集，後者是一組獨立的資料記錄，用來追蹤訓練期間的錯誤以避免方法對資料中的機會變異進行建模。指定記錄百分比。預設值是 30。

複製結果。 設定隨機種子可讓您抄寫分析。請指定一個整數，或是按一下「**產生**」以建立介於 1 和 2147483647 之間（含）的虛擬亂數整數。

CHAID 節點 - 進階

進階選項可讓您細部調整樹狀結構建置程序。

分割的顯著性層次。指定用於分割節點的顯著性水準 (Alpha 值)。該值必須位於 0 和 1 之間。值越小，生成的樹狀結構的節點也會越少。

合併的顯著性層級。指定用於合併種類的顯著性層次 (Alpha 值)。該值必須大於 0 並小於或等於 1。要阻止任何種類合併，可以將值指定為 1。對於連續目標，這意味著最終樹狀結構中變數個種類數與指定的時間間隔數相符。耗盡的 CHAID 無法使用此選項。

使用 Bonferroni 方法調整顯著值。 在測試預測值的各種種類組合時調整顯著性值。根據與預測值的種類數及測量層級直接相關的測試數調整值。這通常是理想的選項，因為它可更好地控制誤判的誤差率。取消此選項將增加您的分析能力以找到 true 差分，但以增加假陽性率為代價。特別是，可以建議針對較小的樣本停用此選項。

容許重新分割節點內的已合併種類。CHAID 演算法試圖合併種類以生成用於說明模型的最簡單的樹狀結構。如果已選取，為了更好地解決問題，則可使用此選項重新分割合併的種類。

種類目標的卡方值。對於種類目標，您可以指定用於計算卡方測試統計資料的方法。

- **Pearson。** 這個方法會提供更快速的計算，但是用在小型樣本時則必須小心。
- **概似比。** 與 Pearson 方法相比，此方法更加穩健，但計算時間更長。對於小型的樣本，這是較佳的方法。針對連續目標，一律使用此方法。

儲存格期望次數中的最少變更。 （為列名模型和列作用順序模型）估計 Cell 頻率時，迭代程序 (epsilon) 用於對最佳化估計（在特定分割的卡方測試中使用）進行收斂。Epsilon 可決定必須發生多少變更，疊代才能繼續；如果前次疊代中的變更小於指定的值，則疊代會停止。如果遇到演算法不收斂的問題，則增加此值或增加疊代次數上限，直到發生收斂。

聚合的疊代數上限。 指定停止前的上限疊代數，而不考慮是否已進行收斂。

過適預防集。 （只有在使用交互樹狀結構建置器時，此選項才可用。）演算法會在內部將記錄分割為模型建置集和過適預防集，後者是一組獨立的資料記錄，用來追蹤訓練期間的錯誤以避免方法對資料中的機會變異進行建模。指定記錄百分比。預設值是 30。

複製結果。 設定隨機種子可讓您抄寫分析。請指定一個整數，或是按一下**「產生」**以建立介於 1 和 2147483647 之間（含）的虛擬亂數整數。

決策樹節點模型選項

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。還可以選擇取得預測值重要性資訊，以及旗標目標的原始傾向分數和已調整的傾向分數。

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。

模型評估

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，對於部分模型，需要較長時間來計算預測值重要性，尤其是處理大型資料集時，結果便是依預設會關閉部分模型的預測值重要性。預測值重要性對於決策清單模型無法使用。如需相關資訊，請參閱第 37 頁的『預測值重要性』。

傾向分數

可以在建模節點中和模型塊的「設定」標籤上啟用傾向分數。唯有當選取的目標是旗標欄位時此功能才可用。請參閱第 30 頁的『傾向分數』主題，以取得更多資訊。

計算原始傾向評分。 原始傾向分數僅衍生自基於訓練資料的模型。如果模型預測值為 *true*（將回應），那麼傾向與 P 相同，其中 P 為預測的可能性。如果模型預測的值為假，那麼計算出的傾向為 (1 – P)。

- 如果建立模型時選擇了此選項，那麼依預設將在模型塊中啟用傾向分數。不過，無論是否在建模節點中選擇了原始傾向分數，都可以始終在模型塊中選擇啟用原始傾向分數。
- 對模型進行評分時，原始傾向評分將被新增到將 *RP* 字母附加到標準字首的欄位中。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RRP-churn*。

計算調整傾向評分。 原始傾向僅基於由可能過度擬合的模型指定的估計，這將導致過於樂觀地估計傾向。已調整的傾向試圖通過查看模型在測試或驗證分割區的性能或通過調整傾向來彌補，以相應地給作出更好的估計。

- 此設定要求串流中出現有效的分割區欄位。
- 與原始信賴度分數不同，已調整的傾向評分必須在建立模型時計算；否則，對模型塊進行評分時該分數將不存在。
- 對模型進行評分時，在將 *AP* 字母附加到標準字首的欄位中新增已調整的傾向評分。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RAP-churn*。已調整的傾向分數不適用於邏輯迴歸模型。
- 在計算已調整的傾向分數時，必須尚未已平衡用於計算的測試或驗證分割區。為避免這一點，請確保在任何上游平衡節點中已選取**僅平衡訓練資料**選項。此外，如果已在上游獲取了複合樣本，那麼這將導致已調整的傾向分數無效。
- 已調整的傾向分數不適用於「增強型」樹狀結構和規則集模型。請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。

依據。對於要進行計算的已調整的傾向分數，串流中必須出現一個分割區欄位。可以指定是使用測試分割區還是驗證分割區進行此計算。為獲取最佳結果，測試或驗證分割區包含的記錄數量應至少與用於訓練原始模型的分割區所包含的記錄數相同。

C5.0 節點

SPSS Modeler Professional 和 SPSS Modeler Premium 中提供了此功能。

該節點使用 C5.0 演算法建立 **決策樹** 或 **規則集** 。C5.0 模型的工作原理是根據提供上限 **資訊增益** 的欄位分割樣本。然後通常會根據不同的欄位再次分割由第一次分割定義的每個子樣本，且此過程會重複下去直到無法繼續分割子樣本。最後，將重新檢查最底層次分割，並刪除或 **刪改** 對模型值沒有顯著影響的分割。

註：C5.0 節點只能預測種類目標。分析包含種類（名義或序數）欄位的資料時，與 11.0 版以前的 C5.0 版本相比將種類群組合在一起的可能性更大。

C5.0 可以生成兩種模型。**決策樹** 是對由演算法建立的分割的簡單說明。每個終端機（或「葉節點」）節點可說明訓練資料的特定子集合，而訓練資料中的每個觀察值都完全的樹狀結構中的某個終端節點。換句話說，對於在決策樹中顯示的任何特定資料記錄，僅可能有一個預測。

反過來， **規則集** 則是嘗試對單個記錄進行預測的一組規則。規則集源自決策樹，並且在某種程度上代表在決策樹中建立的經簡化或提取的資訊版本。規則集通常可以保留來自某個完整決策樹（但具有的模型不太複雜）的大部分重新資訊。由於規則集的這種工作方式，其內容與決策樹的內容不同。最重要的區別是，套用規則集時，可以為任意特定記錄套用多個規則，也可以不套用任何規則。如果套用多個規則，則每個規則將根據與此規則關聯的信賴度獲得一個加權「投票」，並通過組合套用到所討論記錄的所有規則的加權投票來確定最終的預測。如果沒有套用任何規則，則會將預設預測指派給記錄。

範例。 醫學研究員已收集一組患有相同疾病的病患的相關資料。在治療過程中，每位病患均對五種藥物中的一種有明顯反應。您可以將 C5.0 模型與其他節點結合使用，以說明找出可能適用於今後患有相同疾病的病患的藥物。

需求。 要訓練 C5.0 模型，必須有一個種類（即名義或序數）目標欄位和一個或多個任意類型的輸入欄位。會忽略設為兩者 或無 的欄位。模型中所用的欄位必須已完全實例化其類型。還可以指定加權欄位。

強度。 遇到缺少資料及存在大量輸入欄位等問題時，C5.0 模型的表現十分穩健。它們通常不需要很長的訓練時間來進行估計。此外，C5.0 模型與某些其他模型類型相比似乎更容易理解，因為源自模型的規則解譯起來更簡明易懂。C5.0 還提供功能強大的 **增強** 方法來增加分類的精確度。

註：啟用平行處理可以有助於提高 C5.0 模型建置速度。

C5.0 節點模型選項

模型名稱。指定要生成的模型的名稱。

- **自動填滿。** 在已選取此選項的情況下，將根據目標欄位名稱自動產生模型名稱。此為預設值。
- **自訂。** 選中此選項可以為此節點將建立的模型塊指定專屬名稱。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

輸出類型。在此處指定您希望生成的模型塊是**決策樹**還是**規則集**。

群組符號。如果已選取了此選項，那麼 C5.0 將試圖對輸出欄位具有相似型樣的符號值進行組合。如果未選取此選項，C5.0 將為用於分割母節點的符號欄位的每個值建立一個子節點。例如，如果 C5.0 分割的是 顏色欄位（其值為 紅色、 綠色和 藍色），則它將預設建立一個三向分割。但是，如果已選取此選項，且 顏色 = 紅色的記錄與 顏色 = 藍色的記錄非常相似，則 C5.0 將建立一個雙向分割，其中所有 綠色記錄在一個群組中，而所有 藍色記錄連同所有 紅色記錄在另一個群組中。

使用增強。C5.0 演算法包含一個用於提高其準確率的特殊方法，稱為**增強**。它的工作原理是在序列中建立多個模型。第一個模型以平常方式建置。然後，會建置第二個模型以便讓它著重於由第一個模型分類錯誤的記錄。然後會建置第三個模型以著重於第二個模型的錯誤，依此類推。最後會對觀察值分類，方法對其套用整個模型集，使用加權投票程序將個別的預測結合至一個整體預測。增強方法可以顯著提高 C5.0 模型的精確度，但也需要更長的訓練時間。通過**嘗試次數**選項，您可以控制用於增強型模型的模型數。此功能基於 Freund 和 Schapire 的研究，通過一些專有改善更好地處理雜訊資料。

交互驗證。如果已選取此選項，那麼 C5.0 將使用一組根據訓練資料的子集合構建的模型來估計根據整個資料集構建的模型的精確度。如果資料集太小以致於無法將其分割為傳統的訓練集合和測試集，此選項非常有用。在計算精確度評估後，交叉驗證模型將被捨棄。可以指定用於交叉驗證的**摺疊次數**或模式個數。注意，在 IBM SPSS Modeler 以前的版本中，建立模型和交叉驗證模型是兩個個別的作業。在目前的版本中，則無需執行個別的模型建立步驟。模型建置和交叉驗證將同時執行。

眾數。 對於**簡單**訓練，大多數 C5.0 參數是自動設定的。**專家**訓練容許更直接地控制訓練參數。

簡單模式選項

偏向。依預設，C5.0 將嘗試盡可能生成最準確的樹狀結構。在某些情況下，此操作可能會導致過度配適，從而在將此模型套用至新資料時導致效能偏低。選取**通用性**以使用受此問題影響較小的演算法設定。

註：不保證在已選取**通用性**選項的情況下建立的模型的適用性優於其他模型。當通用性問題比較嚴重時，一律可使用保留測試樣本驗證模型。

預期的雜訊 (%)。指定訓練集合中雜訊資料或錯誤資料所佔的預期比例。

專家模式選項

刪改重要性。確定決策樹或規則集的刪改程度。增加該值可獲得一個更簡潔的小型樹狀結構。減小該值可獲得一個更精確的樹狀結構。此設定僅影響本端刪改（請參見下方的「廣域刪改」）。

每個子分支的記錄數下限。可以使用子群組的大小來限制樹狀結構的任何分支中的分割數。僅當兩個或多個生成的子分支中至少包含從訓練集合得到的這一最小記錄數時，才可分割樹狀結構的分支。預設值為 2。增加該值有助於防止使用雜訊資料進行 **過度訓練** 。

使用廣域刪改。樹狀結構的刪改分為兩個階段：第一個階段是本端刪改，將檢查子樹狀結構並摺疊分支以增加模型的精確度。第二個階段是刪改，在此階段中將把樹狀結構視作一個整體並摺疊虛弱的子樹狀結構。依預設將執行刪改。要忽略刪改階段，請取消選中此選項。

精選屬性。如果已選取此選項，那麼 C5.0 在開始建立模型前檢查預測值的有效性。如果發現不相關的預測值，則會將其從模型建立過程中排除。此選項對於具有多數預測值欄位的模型非常有用，並且有助於防止過度配適。

註：啟用平行處理可以有助於提高 C5.0 模型建置速度。

Tree-AS 節點

「樹狀結構-AS」節點可以與分散式環境中的資料配合使用。在此節點中，您可以使用 CHAID 或 Exhaustive CHAID 模型來建立決策樹。

CHAID（或卡方自動互動偵測）是一種分類方法，可透過使用卡方統計學來識別最佳分割以建置決策樹。

CHAID 首先會檢查每個輸入欄位與結果之間的交叉表，然後使用卡方獨立性測試來測試顯著性。如果這些關係中有多個關係在統計上顯著，則 CHAID 將選取最顯著（*p* 值最小）的輸入欄位。如果一個輸入有兩個以上的種類，則會比較這些種類，且結果無差異的種類會收合在一起。這是透過順利結合顯示最不明顯之差異的種類配對來完成。如果在指定的測試層次所有剩餘的種類都不同，則這個種類合併程序會停止。若為標稱輸入欄位，可以合併任何種類；若為序數集，則只能合併連續的種類。

「詳盡的 CHAID」是對 CHAID 的修改，它會徹底檢查每個預測值所有可能的分割，但計算時間較長。

需求。 目標及輸入欄位可以為連續欄位或類別欄位；在每一個層次上，節點可分割成兩個以上子群組。模型中使用的任何序數欄位都必須具有數值儲存空間（而不是字串）。必要的話，可以使用再分類節點來對其進行轉換。

強度。 CHAID 可以產生非二元樹狀結構，這意味著有些分割將有多於兩個的分支。 因此，相較於二元增長方法而言，它將建立更寬的樹狀結構。 CHAID 適用於所有類型的輸入，並且它同時接受觀察值加權與頻率變數。

樹狀結構-AS 節點欄位選項

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。

預測值 選取一或多個欄位作為預測的輸入。

分析加權 若要將欄位用作觀察值加權，請在這裡指定欄位。觀察值加權是用來說明輸出欄位不同等級間的變異數差異。如需相關資訊，請參閱第 28 頁的『使用頻率和加權欄位』。

樹狀結構-AS 節點建置選項

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。

該標籤包含數個不同窗格，您可以在其中設定特定於您模型的自訂作業。

樹狀結構-AS 節點 - 基礎

指定關於要如何建置決策樹的基本選項。

樹狀結構增長演算法：選取您要使用的 **CHAID** 演算法類型。 **詳盡的 CHAID** 是對 CHAID 的修改，它會徹底檢查每個預測值所有可能的分割，但計算時間較長。

樹狀結構深度上限：指定根節點以下方的上限級數（對樣本進行遞歸分割的次數）；預設值是 5。上限層次數（也稱為節點）為 50,000。

分級：如果您使用連續資料，您必須對輸入進行分級。 您可以在前一個節點中執行此操作；但是，樹狀結構-AS 節點會對任何連續輸入進行自動 Bin。 如果您使用樹狀結構-AS 節點來自動對資料進行 Bin，請選取要對輸入進行分割的**Bin 個數**。 資料將按等頻率 Bin；可用選項為 2、4、5、10、20、25、50 或 100。

樹狀結構-AS 節點 - 增長

使用增長選項來對樹狀結構建置程序進行微調。

從 p 值切換至作用大小的記錄臨界值：指定在建立樹狀結構時，模型將從使用 **p值設定**切換至**反映大小設定**的記錄數。 預設值為 1,000,000。

分割的顯著性水準：指定用於用於分割節點的顯著性水準 (Alpha 值)。 值必須介於 0.01 到 0.99 之間。較低的值會產生具有較少節點的樹狀結構。

合併的顯著性層級：指定用於合併種類的顯著性水準 (Alpha 值)。 值必須介於 0.01 到 0.99 之間。耗盡的 CHAID 無法使用此選項。

使用 Bonferroni 法調整顯著性值：正在測試預測值的各種種類組合時，調整顯著性值。 根據與預測值的種類數及測量層級直接相關的測試數調整值。這通常是理想的選項，因為它可更好地控制誤判的誤差率。取消此選項將增加您的分析能力以找到 true 差分，但以增加假陽性率為代價。 特別是，可以建議針對較小的樣本停用此選項。

反映大小臨界值（僅限連續目標）：設定使用連續目標時，分割節點和合併種類時要使用的作用大小臨界值。值必須介於 0.01 到 0.99 之間。

反映大小臨界值（僅限類別目標）：設定使用種類目標時，分割節點和合併種類時要使用的作用大小臨界值。值必須介於 0.01 到 0.99 之間。

容許重新分割節點內的已合併種類：CHAID 演算法試圖合併種類以生成用於說明模型的最簡單的樹狀結構。 如果已選取，為了更好地解決問題，則可使用此選項重新分割合併的種類。

葉節點分組的顯著性水準：指定確定如何形成葉節點分組或者如何識別不正常的葉節點的顯著性水準。

種類目標的卡方：對於種類目標，您可以指定用於計算卡方測試統計資料的方法。

- **Pearson**：此方法提供更快的計算，但是對於小型樣本應該謹慎使用它。
- **概似比率檢驗**：與 Pearson 方法相比，此方法更加穩健，但計算時間更長。 對於小型的樣本，這是較佳的方法。針對連續目標，一律使用此方法。

樹狀結構-AS 節點 - 停止規則

這些選項可控制樹狀結構的構建方式。停止規則可確定何時停止分割樹狀結構的特定分支。設定下限分支大小可阻止通過分割建立非常小的子群組。如果節點（父級）中要分割的記錄數少於指定值，那麼**上層分支中的最小記錄**將阻止進行分割。 如果由分割建立的任何分支（子級）中的記錄數少於指定值，那麼**子分支中的最小記錄**將阻止進行分割。

- **使用百分比**：按總訓練資料的百分比指定大小...。
- **使用絕對值**：按絕對記錄數指定大小...。

預期儲存格頻率的變更下限值：（為列名模型和列作用順序模型）估計 Cell 頻率時，迭代程序 (epsilon) 用於對最佳化估計（在特定分割的卡方測試中使用）進行收斂。 Epsilon 可決定必須發生多少變更，疊代才能繼續；如果前次疊代中的變更小於指定的值，則疊代會停止。如果遇到演算法不收斂的問題，則增加此值或增加疊代次數上限，直到發生收斂。

收斂的最大疊代：指定停止前的疊代次數上限，而不考慮是否已進行收斂。

樹狀結構-AS 節點 - 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）。

併入成本的模型產生的誤差不能比不併入成本的模型少，且在整體準確性方面的等級不會更高，但實際效能可能更好，因為它的內建偏差偏好成本較低的誤差。

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。

僅針對序數目標，您可以選取**序數目標的預設成本增加**並在成本矩陣中設定預設值。可用的選項在下列清單中進行說明。

- **無增加** - 針對每個正確的預測使用預設值 1.0。
- **線性** - 每一個連續的不正確預測會將成本增加 1。
- **平方** - 每一個連續的不正確預測是線性值的平方。在這種情況下，值可能為：1、4、9 等。
- **自訂** - 如果您手動編輯表格中的任何值，下拉選項會自動變更為**自訂**。如果您將下拉選項變更為任何其他選項，則您編輯的值會取代為所選取選項的值。

樹狀結構-AS 節點模型選項

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。您也可以選擇計算信賴度值，並在對模型評分期間新增 ID。

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。

計算信賴度：要在對模型進行分數時新增信賴度欄位，請選中此勾選框。

規則 ID：要在對模型進行分數時新增包含記錄分配到的葉節點 ID 的欄位，請選中該欄位。

樹狀結構-AS 模型塊

樹狀結構-AS 模型塊輸出

在建立樹狀結構-AS 模型後，在輸出檢視器中提供了下列資訊。

模型資訊表格

「模型資訊」表格提供模型的關鍵資訊。該表格識別一些高階模型設定，例如：

- 使用的演算法類型；CHAID 或 Exhaustive CHAID。
- 「類型」節點或樹狀結構-AS 節點「欄位」標籤中已選取的目標欄位的名稱。
- 在「類型」節點或樹狀結構-AS 節點「欄位」標籤中，選取作為預測值的欄位名稱。
- 資料中的記錄數。如果使用次數加權建置模型，則此值是代表樹狀結構所基於的記錄數的有效加權計數。
- 產生的樹狀結構中葉節點的數量。
- 樹狀結構種的層次數：即，樹狀結構深度。

預測值重要性

「預測值重要性」圖形以長條圖形式顯示模型中前 10 個輸入（預測值）的重要性。

如果圖表中具有 10 個以上欄位，則您可以使用圖表下的調節器來變更圖表中包含的預測值選擇。調節器上的指示標示是固定寬度，並且調節器上的每個標示都代表 10 個欄位。您可以沿著調節器來移動指示標示以顯示後面或前面的 10 個欄位，依預測值重要性排序。

您可以按兩下圖表以開啟個別對話框來編輯圖形設定。例如，您可以修正一些項目，例如圖形大小，以及所用字型的大小和顏色。關閉這個個別的編輯對話框時，變更會套用至「輸出」標籤中顯示的圖表。

熱門決策規則表格

依預設，此互動式表格會基於葉節點中包含的合計記錄的百分比，顯示輸出中前五個葉節點的規則的統計資料。

您可以按兩下表格來開啟個別的對話框，您可以在其中編輯表格中顯示的規則資訊。顯示的資訊及對話框中可用的選項視目標的資料類型而定：例如，種類或連續。

下列規則資訊會在表格中顯示：

- 規則 ID
- 如何套用及草擬規則的詳細資料
- 每項規則的記錄計數。如果使用次數加權建置模型，則此值是代表樹狀結構所基於的記錄數的有效加權計數。
- 每條規則的記錄百分比

此外，對於連續目標，表格中包含一個額外的欄，其中顯示每條規則的**平均**值。

您可以使用下列**表格內容**選項來更改規則表格佈置：

- **主要決策規則**：按葉節點中包含的合計記錄百分比排序的前五條決策規則。
- **所有規則**：該表格包含模型生成的所有葉節點，但每頁僅顯示 20 條規則。 選取此佈置時，您可以使用其他選項**依 ID 尋找規則**和**頁面**搜尋規則。

此外，對於種類目標，您可以通過使用**按種類種類的主要規則**選項來更改規則表格的佈置。 前五個決策規則依您選取的**目標種類**的記錄總數百分比排序。

如果您變更規則表格的佈置，那麼可以通過按一下位於對話框左上角的「複製到檢視器」按鈕來講修改後的規則表格複製回「輸出檢視器」。

樹狀結構-AS 模型塊設定

在樹狀結構-AS 模型塊的「設定」標籤上，可以在模型評分期間指定用於信賴度及 SQL 產生的選項。 僅當模型片段已新增至串流之後，此標籤才可用。

計算信賴度 要在評分作業中包括信賴度，請選取此勾選框。在資料庫中對模型評分時，排除信賴度表示您可以產生更有效率的 SQL。對於迴歸樹狀結構，不會指派信賴度。

規則 ID：要在評分輸出中新增一個欄位，表示每個記錄分配到的終端節點的 ID，請選中此勾選框。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何產生 SQL：

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

「隨機樹狀結構」節點

「隨機樹狀結構」節點可以與分散式環境中的資料配合使用。此節點中，您可以建立包含多個決策樹的總體模型。

「隨機樹狀結構」節點是一種基於樹狀結構的分類和預測方法，此方法根據分類和迴歸方法方法建立。與 C&R 樹狀結構類似，此預測方法使用遞歸分區將訓練記錄分割為具有相似輸出欄位值的區段。首先，此節點通過檢查可供其使用的輸入欄位來尋找最佳分割（以分割所引起的雜質指標下降情況進行測量）。分割可定義兩個子群組，其中每個子群組隨後又分割為兩個子群組，依此類推，直到觸發其中一項停止準則為止。所有分割都是二元的（只有兩個子群組）。

「隨機樹狀結構」節點使用利用取代的引導取樣，以產生取樣資料。取樣資料用於建立樹狀結構模型。在建立樹狀結構期間，「隨機樹狀結構」不會再次對資料進行取樣。它會隨機選取部分預測值，並使用最佳預測值來分割樹狀結構節點。分割每一個樹狀結構節點時，會重複此處理程序。這是在隨機森林中建立樹狀結構的基本構想。

「隨機樹狀結構」使用類似於 C&R Tree 的樹狀結構。由於此類樹狀結構為二進位，因此用於分割的每一個欄位都會產生兩個分支。對於具有多個種類的種類欄位，種類會根據內部分割準則分組為兩個群組。每一個樹狀結構都會成長為最大的延伸可能（無刪改）。評分時，「隨機樹狀結構」依大多數投票（適用於分類）或平均值（適用於迴歸）結合個別樹狀結構評分。

「隨機樹狀結構」與 C&R Tree 不同，如下所示：

- 「隨機樹狀結構」節點隨機選取指定數目的預測值，並使用所選項目中最佳的預測值以分割節點。與此相反，C&R Tree 從所有預測值中尋找最佳預測值。
- 「隨機樹狀結構」中的每一個樹狀結構都完整成長，通常直到每一個葉節點都包含一筆單一記錄。因此樹狀結構深度可能非常大。但是標準 C&R Tree 將不同的停止規則用於樹狀結構成長，這通常導致樹狀結構淺得多。

與 C&R 樹狀結構相比，隨機樹狀結構將新增兩項功能：

- 第一項目功能是組裝，其中訓練資料集的抄本是通過對原始資料集進行放回取樣來建立的。此動作將大小與原始資料集相等的 Bootstrap 樣本，在此動作執行後將根據每個抄本建立成份模型。與這些元件模型一起組成複合模型。
- 第二項功能是，在樹狀結構的每個分割處僅考慮將輸入欄位採樣進行雜質測量。

需求。 要訓練「隨機樹狀結構」模型，您需要一個或多個輸入欄位以及一個目標欄位。目標欄位和輸入欄位可以是連續（數值範圍）或類別欄位。將忽略設定為兩者或無的欄位。對於模型中使用的欄位，必須將它們的類型完全實例化，並且模型中使用的任何序數（排定次序的集）欄位的儲存類型必須是數值類型（而不是字串）。必要的話，可以使用「重新分類」節點來對其進行轉換。

強度。 處理大型資料集和許多欄位時，「隨機樹狀結構」模型是穩健的模型。由於使用組裝和欄位採樣，因此它們更不容易過度配適，並且正在測試中看到的結果更可能在您使用新資料時重複。

「隨機樹狀結構」節點欄位選項

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。

預測值 選取一或多個欄位作為預測的輸入。

分析加權 若要將欄位用作觀察值加權，請在這裡指定欄位。觀察值加權是用來說明輸出欄位不同等級間的變異數差異。如需相關資訊，請參閱第 28 頁的『使用頻率和加權欄位』。

「隨機樹狀結構」節點建置選項

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。

該標籤包含數個不同窗格，您可以在其中設定特定於您模型的自訂作業。

「隨機樹狀結構」節點 - 基本

指定如何建置決策樹狀結構的基本選項。

要建置的模型數目。指定節點可以建置的樹狀結構數目上限。

樣本大小。依預設，Bootstrap 樣本的大小等於原始訓練資料。處理大型資料集時，減少樣本大小可以增加效能。它是從 0 到 1 的比例。例如，將樣本大小設為 0.6，以將它降低為原始訓練資料大小的 60%。

處理不平衡的資料。如果模型的目標是旗標結果（例如，購買或不購買）並且所需結果與非所需結果的比例很小，那麼資料是不已平衡資料並且模型所處理的 Bootstrap 採樣可能會影響模型精確性。要提高精確度，請選中此勾選框；模型隨後會擷取所需結果中的更大比例部分並產生更好的模型。

使用加權採樣選擇變數。依預設，每個葉節點的變數是使用同一機率隨機選擇的。要將加權用於變數並改進選擇過程，請選中此勾選框。加權是由「隨機樹狀結構」節點自行計算。更重要的欄位（具有更高加權）更可能選取作為預測值。

節點數目上限。指定個別樹狀結構中容許的葉節點數目上限。如果下一次分割時將超過此數字，那麼樹狀結構成長將在進行分割之前停止。

樹狀結構深度上限。指定根節點下方的上限葉節點層次數；即，樣本進行遞歸分割的次數。

子節點大小下限。指定分割母節點之後必須包含在子節點中的下限記錄數。如果子節點包含的記錄數少於您輸入的數目，那麼不會分割母節點。

指定要用於分割的預測值數目。如果是建立分割模型，請設定要用於建立每個分割的下限預測值數目。這防止分割建立過小的子群組。如果您未選取此選項，則預設值為 *sqrt*(M)（適用於分類）及 M/3（適用於迴歸），其中 M 是預測值變數的數目總計。如果選取此選項，則將使用指定數目的預測值。

註：用於分割的預測值數目不能大於資料中的預測值總數。

當精確度無法再提高時停止建構。要改進模型建置時間，請選取此選項，以在結果的精確度無法提高時停止模型建置過程。特別是，如果現行集合精確度的改良小於指定的臨界值，則將停止新增樹狀結構。這可能導致模型的樹狀結構少於您為**要建置的模型數目**選項指定的值。

「隨機樹狀結構」節點 - 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）。

併入成本的模型產生的誤差不能比不併入成本的模型少，且在整體準確性方面的等級不會更高，但實際效能可能更好，因為它的內建偏差偏好成本較低的誤差。

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。

僅針對序數目標，您可以選取**序數目標的預設成本增加**並在成本矩陣中設定預設值。可用的選項在下列清單中進行說明。

- **無增加** - 對於每一個不正確的預測，預設值為 1.0。
- **線性** - 每一個連續的不正確預測會將成本增加 1。
- **平方** - 每一個連續的不正確預測是線性值的平方。在這種情況下，值可能為：1、4、9 等。
- **自訂** - 如果您手動編輯表格中的任何值，下拉選項會自動變更為**自訂**。如果您將下拉選項變更為任何其他選項，則您編輯的值會取代為所選取選項的值。

「隨機樹狀結構」節點 - 進階

指定如何建置決策樹狀結構的進階選項。

遺漏值的最大百分比。指定任何輸入中容許的遺漏值的上限百分比。如果該百分比超過了此數字，那麼將從模型建置中排除此輸出。

排除單個種類多數超過以下值的欄位。指定單個種類可以在某個欄位中具有的上限記錄百分比。如果任何種類值代表的記錄百分比高於指定值，那麼將從模型建立中排除整個欄位。

上限欄位種類數。指定欄位中可以包含的上限種類數。如果種類數超過了此數字，那麼將從模型建置中排除此欄位。

欄位變異下限。如果某個連續欄位的變異係數小於您在此處指定的值（換言之，該欄位接近常數），那麼將從模型建置中排除此欄位。

分組數目。 請指定要用於連續輸入的均等頻率 Bin 數。可用選項包括：2、4、5、10、20、25、50 或 100。

要報告的相關規則數。指定要報告的規則數目量（最小值為 1，最大值為 1000，預設值為 50）。

「隨機樹狀結構」節點模型選項

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。對模型進行評分的過程中，您還可以選擇計算預測值的重要性。

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。

隨機樹狀結構模型塊

隨機樹狀結構模型塊輸出

建立隨機樹狀結構模型之後，輸出檢視器中提供了下列資訊：

模型資訊表格

模型資訊表格提供關於模型的關鍵資訊。該表格一律包括下列高階模型設定：

- 在「類型」節點或「隨機樹狀結構」節點欄位標籤中選取的目標欄位的名稱。
- 模型建置方法 - 隨機樹狀結構。
- 輸入到模型中的預測值數。

表格中顯示的其他詳細資訊取決於您建立的是分類模型還是迴歸模型以及建立模型是否旨在處理不已平衡資料：

- 分類模型（預設值）
  - 模型精確性
  - 誤分類規則
- 分類模型（已選取**處理不已平衡資料**）
  - Gmean
  - true 陽性率（細分為多個類別）。
- 迴歸模型
  - 均方根誤差
  - 相對錯誤
  - 解釋的變異數

記錄摘要

此摘要顯示用於擬合模型的記錄數以及排除的記錄數。將顯示這兩種記錄數以及整數所佔百分比。如果建立模型旨在併入次數加權，那麼還將顯示併入和排除的記錄的未加權號碼。

預測值重要性

「預測值重要性」圖形以長條圖形式顯示模型中前 10 個輸入（預測值）的重要性。

如果圖表中具有 10 個以上欄位，則您可以使用圖表下的調節器來變更圖表中包含的預測值選擇。調節器上的指示標示是固定寬度，並且調節器上的每個標示都代表 10 個欄位。您可以沿著調節器來移動指示標示以顯示後面或前面的 10 個欄位，依預測值重要性排序。

您可以按兩下圖表以開啟個別的對話框，您可以在其中編輯圖表大小。關閉這個個別的編輯對話框時，變更會套用至「輸出」標籤中顯示的圖表。

熱門決策規則表格

依預設，此交互表格顯示按相關度排序的主要規則的統計資料。

您可以按兩下表格來開啟個別的對話框，您可以在其中編輯表格中顯示的規則資訊。顯示的資訊及對話框中可用的選項視目標的資料類型而定：例如，種類或連續。

下列規則資訊會在表格中顯示：

- 如何套用及草擬規則的詳細資料
- 如果結果的最常用的種類
- 規則精確度
- 樹狀結構精確度
- 相關度指標

相關度指標將使用下列公式進行計算：

$$I_{index}(t) = P(A(t)) * P(B(t)) * (P(B(t)|A(t)) + P(\bar{B}(t)|\bar{A}(t)))$$

在此公式中：

- P(A(t)) 是樹狀結構精確度
- P(B(t)) 是規則精確度
- P(B(t) | A(t)) 代表樹狀結構和節點由樹狀結構和節點進行的正確預測
- 此公式的其餘部分代表由樹狀結構和節點進行的不正確預測

您可以使用下列**表格內容**選項來變更規則表格：

- **主要決策規則** 按相關度指標排序的前五條主要決策規則。
- **所有規則** 該表格包含由模型生成的所有規則，但每頁僅顯示 20 條規則。選取此佈置時，您可以使用其他選項**依 ID 尋找規則**和**頁面**搜尋規則。

另外，對於種類目標，您可以使用**按種類列出的主要規則**選項來變更規則表格佈置。前五個決策規則依您選取的**目標種類**的記錄總數百分比排序。

註：對於種類目標，僅當未在建置選項的「基本」標籤中選取**處理不已平衡資料**時，此表格才可用。

如果變更了規則表格的佈置，那麼通過按一下對話框左上角的「複製到檢視器」按鈕，您可以將修改後的規則表格複製回輸出檢視器。

混淆矩陣

對於分類模型，混淆矩陣顯示預測結果數與實際觀測結果數，包含正確預測所佔的比例。

註：混淆矩陣不適用於迴歸模型，並且在建置選項的「基本」標籤上選取了**處理不已平衡資料**時，也無法使用混淆矩陣。

隨機樹狀結構模型塊設定

在隨機樹狀結構模型塊的「設定」標籤上，您可以指定模型評分期間用於信賴度和 SQL 產生的選項。僅當模型片段已新增至串流之後，此標籤才可用。

計算信賴度 要在評分作業中包括信賴度，請選取此勾選框。在資料庫中對模型評分時，排除信賴度表示您可以產生更有效率的 SQL。對於迴歸樹狀結構，不會指派信賴度。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何產生 SQL：

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

C&R 樹狀結構、CHAID、QUEST 和 C5.0 決策樹模型塊

決策樹模型塊代表用於預測其中一個決策樹建模節點（C&R 樹狀結構、CHAID、QUEST 或 C5.0）所探索的特定輸出欄位的樹狀結構結構。樹狀結構模型可以直接從樹狀結構建立節點中產生，也可以從互動式樹狀結構建置器中間接產生。請參閱第 70 頁的『交互樹狀結構建置器』主題，以取得更多資訊。

評分樹狀結構模型

執行包含樹狀結構模型塊的串流時，特定的結果取決於樹狀結構的類型。

- 對於分類樹狀結構（種類目標），會將兩個新欄位（其中分別包含每條記錄的預測值和信賴度）新增到資料中。預測取決於為其分配記錄的終端節點的使用最頻繁的種類；如果在給定節點中大多數回應為 是，那麼對分配到該節點的所有記錄的預測也為是。
- 對於迴歸方法樹狀結構，僅產生預測值；而不會分配信賴度。
- 另外，對於 CHAID、QUEST 和 C&R 樹狀結構模型，也可以新增表示節點 ID 的附加欄位，每條記錄都將分配到此節點中。

新欄位名稱透過新增字首從模型名稱衍生。對於 C&R 樹狀結構、CHAID 和 QUEST，預測欄位的字首是 *$R-*，信賴度欄位的字首是 *$RC-*，而節點 ID 欄位的字首是 *$RI-*。對於 C5.0 樹狀結構，預測欄位的字首是 *$C-*，而信賴度欄位的字首是 *$CC-*。如果出現多個樹狀結構模型節點，那麼必要時新的欄位名稱的字首中將包含數字以進行識別 - 例如，*$R1-*、*$RC1-* 和 *$R2-*。

使用樹狀結構模型塊

可以多種方式儲存或匯出與模型相關的資訊。

註：樹狀結構建置器視窗中也提供了其中的多數選項。

通過樹狀結構建置器或樹狀結構模型塊，可以執行下列操作：

- 根據目前的樹狀結構產生過濾節點或選取節點。如需相關資訊，請參閱第 79 頁的『產生過濾節點和選取節點』。
- 產生一個規則集塊，該節點將樹狀結構結構代表成一組定義了樹狀結構的終端機分支的規則。如需相關資訊，請參閱第 80 頁的『從決策樹中產生規則集』。
- 此外，還可以按 PMML 格式匯出模型（僅限樹狀結構模型塊）。如需相關資訊，請參閱第 34 頁的『模型選用區』。如果模型包含任何自訂分割，那麼不會在匯出的 PMML 中保留此資訊。（保留分割，但不保留它是自訂分割而不是通過演算法選擇的分割這一事實。）
- 基於目前樹狀結構的所選部分產生圖表。請注意，此操作僅對附加到串流中其他節點的區塊有效。如需相關資訊，請參閱第 105 頁的『產生圖形』。
- 僅在增強型 C5.0 模型中，可以選擇**單一決策樹（畫布）**或**單一決策樹（GM 選用區）**以根據目前選定的規則建立一個新的過長。請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。

註：雖然 C&R 樹狀結構節點已替代「建立規則」節點，但現有串流中最初使用「建立規則」節點建立的決策樹節點仍然正常工作。

單一樹狀結構模型塊

如果在建模節點上選取**建立單一樹狀結構**作為主目標，那麼結果模型塊包含下列標籤。

表 7. 單一樹狀結構塊上的標籤

Tab	說明	進一步資訊
模型	顯示用於定義模型的規則。	請參閱『決策樹模型規則』主題，以取得更多資訊。
檢視器	顯示模型的樹狀結構形視圖。	請參閱第 104 頁的『決策樹模型檢視器』主題，以取得更多資訊。
摘要	顯示欄位、建置設定及模型估計程序的相關資訊。	請參閱第 36 頁的『模型塊概要/資訊』主題，以取得更多資訊。
設定	使您可以在模型評分期間為信賴度及 SQL 產生指定選項。	請參閱第 104 頁的『決策樹/規則集模型塊設定』主題，以取得更多資訊。
註釋	可讓您新增敘述註釋、指定自訂名稱、新增工具提示文字並指定模型的搜尋關鍵字。	

決策樹模型規則

決策樹模型塊的「模型」標籤顯示定義該模型的規則。此外，還可以顯示預測值重要性的圖形和包含有關歷史、頻率和替代資訊的第三個窗格。

註：如果您在 CHAID 節點的「建置選項」標籤（「目標」畫面）上選中**為大型資料集建立模型**選項，那麼「模型」標籤只顯示樹狀結構規則詳細資料。

樹狀結構規則

左側窗格顯示了條件清單，這些條件定義演算法探索的資料的分區 - 本質上是一系列規則，可基於不同預測值的值將單個記錄分配給子節點。

決策樹基於輸入欄位值的對資料進行遞歸分區。資料分割區稱為分支。 初始分支（有時稱為根）包含所有資料記錄。根將根據特定輸入欄位的值被分成多個子集或子分支。每個子分支可以進一步分割成次級子分支，次級子分支還可進一步分割，如此類推。不再分割的分支是樹狀結構的最底層次分支。這樣的分支稱為終端機分支（或葉節點）。

樹狀結構規則詳細資料

規則瀏覽器顯示了輸入值，輸入值定義了每個分割區或分支以及這些分割中記錄的輸出欄位值概要。如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』。

對於基於數值型型欄位的分割，分支將以下行所示的形式顯示：

```
fieldname relation value [summary]
```

這裡的 *relation* 是數值型型關係。例如，由 *revenue* 欄位大於 100 的值所定義的分支將顯示為如下形式：

```
revenue  > 100 [summary]
```

對於基於符號型欄位的分割，分支將以下行所示的形式顯示：

```
fieldname = value [summary] or fieldname in [values] [summary]
```

這裡的 *values* 代表定義分支的欄位值。例如，包含 *region* 字段值為 *North*、*West* 或 *South* 的記錄的分支將以如下形式代表：

```
region in ["North" "West" "South"] [summary]
```

終端機分支也將進行預測，同時會在規則條件的尾部新增箭頭和預測值。例如，定義時 *revenue* > 100 且預測輸出欄位值為 *high* 的葉節點將顯示如下：

```
revenue > 100 [Mode: high] → high
```

數值型型和符號型輸出欄位的分支概要定義有所不同。對於含有數值型型輸出欄位的樹狀結構，分支的平均值值便是概要，分支的作用便是分支平均值與其父分支平均值的差。對於含有符號型輸出欄位的樹狀結構，分支中記錄的中位數（或出現頻率最高的值）便是概要。

要完全說明分支，需要包含定義分支的條件以及定義樹狀結構中更深層分割的條件。例如，在樹狀結構中：

```
revenue  > 100  region = "North"  region in ["South" "East" "West"]    revenue <= 200
```

第二行所代表的分支由條件 *revenue > 100* 和 *region = "North"* 進行定義。

如果按一下工具列上的 **顯示實例/信賴度**，那麼每條規則還將顯示其所適用的記錄數（實例數）和規則為 true 的記錄所佔的比例（信賴度）。

預測值重要性

選擇性地，指出評估模型時每個預測值的相對重要性的圖表，可能也會顯示在「模型」標籤上。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。

註：只有在產生模型之前已選取「分析」標籤上的**計算預測值重要性**，才可以使用此圖表。 請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

其他模型資訊

如果按一下工具列中的 **顯示其他資訊窗格** ，您將在視窗底部看到顯示選定規則詳細資訊的窗格。資訊窗格包含三個標籤。

歷史。此標籤追蹤從根節點至選定節點的分割準則。從而給出了一個條件清單，據此可以判斷出記錄何時分配給了選定節點。所有條件均為 true 的記錄將分配給此節點。

頻率。對於含符號型目標欄位的模型而言，此標籤（為每個可能的目標值）顯示了分配給包含目標值（訓練資料中）節點的記錄的數量。還將顯示頻率圖（顯示為最多三位小數的百分比）。對於含數值型型目標的模型，此標籤為空白。

代理項。如果適用，那麼會針對所選節點顯示主要分割欄位的所有替代。替代是在給定記錄的主要預測值遺漏時使用的代理欄位。給定分割容許的上限替代數在樹狀結構建立節點中指定，但實際數量取決於訓練資料。一般來講，遺漏資料越多，可能使用的替代越多。對於其他決策樹模型，此標籤為空白。

註：要在模型中包含代理項，必須在訓練階段對其進行 ID。 如果訓練樣本沒有遺漏值，那麼不會 ID 任何替代；在正在測試或評分過程中遇到的具有遺漏值的所有記錄將自動落入記錄數最大的子節點。如果在正在測試或評分過程中預期出現遺漏值，請確保值在訓練樣本中也處於遺漏狀態。替代對於 CHAID 樹狀結構無法使用。

效果

節點的效果是平均值（與母節點相比較的預測值）的增加或減少。例如，如果節點的平均值是 0.2，其母項的平均值是 0.6，則節點的效果是 0.2-0.6=-0.4。此統計資料僅適用於連續目標。

決策樹模型檢視器

決策樹模型塊的「檢視器」標籤類似於樹狀結構建置器中的顯示。主要的區別是當瀏覽模型塊時，無法生成或修改樹狀結構。兩個元件中用於檢視和自訂顯示的其他選項都類似。請參閱第 73 頁的『自訂樹狀結構形視圖』主題，以取得更多資訊。

註：如果您在「建置選項」標籤的「目標」畫面上選中了**為大型資料集建立模型**選項，那麼對於已建立的 CHAID 模型塊，將不會顯示「檢視器」標籤。

在「檢視器」標籤上檢視分割規則時，方括弧表示相鄰值包括在範圍內，而括弧指出相鄰值從範圍內排除。因此，表示式 (23,37] 表示從 23（不含）至 37（含），即，從 23 以上至 37。在「模型」標籤上，相同的條件會顯示為：

年齡 > 23 和年齡 <= 37

決策樹/規則集模型塊設定

通過決策樹或規則集模型塊的「設定」標籤，您可以在模型評分期間指定用於信賴度及 SQL 產生的選項。只有將模型塊新增到串流之後，此標籤才可用。

計算信賴度：選中此選項以便在評分作業中包含信賴度。在資料庫中評分模型時，排除信賴度有助於產生更有效的 SQL。對於迴歸樹狀結構，不會指派信賴度。

註：如果您在 CHAID 模型的「方法」畫面上的「建置選項」標籤中選中**為大型資料集建立模型**選項，那麼此勾選框僅在列名或旗標種類目標的模型塊中可用。

計算原始傾向分數：對於含旗標目標（傳回「是」或「否」預測）的模型，您可以要求傾向分數，這些分數指示為目標欄位指定結果為 true 的可能性。 這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

註：如果您在 CHAID 模型的「方法」畫面上的「建置選項」標籤中選中**為大型資料集建立模型**選項，那麼此勾選框僅在旗標種類目標的模型塊中可用。

計算已調整的傾向分數：原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

註：已調整的傾向分數不適用於增強型樹狀結構和規則集模型。 請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。

規則 ID：對於 CHAID、QUEST 和 C&R 樹狀結構模型，此選項將在評分輸出中新增指示終端節點的 ID 的欄位，每條記錄都將分配到此終端節點中。

註：已選取此選項後，SQL 產生將無法使用。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **通過轉換至無遺漏值支援的原生 SQL 來進行評分：**如果選取此項，將產生原生 SQL 在資料庫中對模型進行評分，而沒有處理遺漏值的開啟銷。 如果在評分觀察值時遇到遺漏值，那麼此選項會將預測設定為空值 ($null$)。

註：此選項對於 CHAID 模型不適用。 對於其他模型類型，此選項僅適用於決策樹（而非規則集）。

- **通過轉換至含遺漏值支援的原生 SQL 來進行評分**對於 CHAID、QUEST 和 C&R 樹狀結構模型，您可以產生原生 SQL 以在資料庫中對模型進行評分，此資料庫具有完整的遺漏值支援。 因此，產生 SQL 意味著已按模型中指定的方式處理遺漏值。例如，C&R 樹狀結構使用代理規則和最大子退回。

  註：對於 C5.0 模型，此選項僅可用於規則集（而非樹狀結構）。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

增強型 C5.0 模型

SPSS Modeler Professional 和 SPSS Modeler Premium 中提供了此功能。

建立增強型 C5.0 模型（規則集或樹狀結構）時，實際上建立了一組相關模型。增強型 C5.0 模型的模型規則瀏覽器顯示位於階層頂層的模型的清單，以及每個模型的估計精確度和增強型模型的總體的精確度。要檢查特定模型的規則或分割，可選取並根據在單模型中擴展規則或分支的方式擴展該模型。

也可以從增強型模型集中擷取特定的模型並建立恰好包含此模型的新規則集模型塊。要從增強型 C5.0 模型中建立新的規則集，可選擇所需規則集或樹狀結構，並從「產生」功能表中選擇**單一決策樹（GM 選用區）**或**單一決策樹（畫布）**。

產生圖形

「樹狀結構」節點提供了大量資訊；但是，對於業務使用者，此資訊可能並非始終採用可輕鬆存取的格式。若要提供資料以便可以將資料輕易納入商業報告、簡報等，您可以產生所選資料的圖形。例如，您可以從模型塊的「模型」或「檢視器」標籤，或者從交互樹狀結構的「檢視器」標籤為樹狀結構的選定部分產生圖表，從而僅為選定的樹狀結構或分支節點中的觀察值建立圖表。

註：僅當塊附加到串流中的其他節點時，您才能根據該塊產生圖表。

產生圖形

第一步是選取要在圖表上顯示的資訊：

- 在塊的「模型」標籤上，展開左側窗格中的條件和規則清單，然後選取感興趣的項。
- 在塊的「檢視器」標籤上，展開分支清單並選取感興趣的節點。
- 在交互樹狀結構的「檢視器」標籤上，展開分支清單並選取感興趣的節點。

  註：您無法在上述「檢視器」標籤中選取頂級節點。

建立圖表的方式相同，而與選取顯示資料的方式無關：

1. 從「產生」功能表選擇**圖表（從選擇）**；或者在「檢視器」標籤上按一下左下角落處的**圖表（從選擇）**按鈕。即會顯示「圖表板基本」標籤。

   附註：以此方式顯示圖形板時，只有「基本」及「詳細」標籤可用。
2. 使用「基本」或「詳細」標籤設定來指定要顯示在圖形上的詳細資料。
3. 按一下「確定」以產生圖形。

圖表標題識別已選擇要併入的節點或規則。

用於增強、組裝和超大型資料集的模型塊

如果在建模節點上選取**提高模型精確度（增強）**、**提高模型穩定性（組裝）**或**為大型資料集建立模型**作為主目標，那麼 IBM SPSS Modeler 會建立多個模型的總體。請參閱第 38 頁的『集合的模型』主題，以取得更多資訊。

生成的模型塊包括下列標籤。「模型」標籤提供多個不同的模型視圖。

表 *8.* 模型塊中可用的標籤

Tab	視圖	說明	進一步資訊
模型	模型摘要	顯示總體品質和（增強型模型和連續目標除外）差異性的摘要，以及有關預測變量在不同模型中的變化程度的測量。	請參閱第 38 頁的『模型摘要』主題，以取得更多資訊。
	預測值重要性	顯示了指示估計模型時每個預測值（輸入欄位）的相對重要性的圖表。	請參閱第 39 頁的『預測值重要性』主題，以取得更多資訊。
	預測值頻率	顯示了表示每個預測值在模型集中的相對使用頻率的圖表。	請參閱第 39 頁的『預測值次數』主題，以取得更多資訊。
	成分模型準確度	繪製關於總體中每個不同模型的預測精確度的圖表。	
	成分模式詳細資料	顯示總體中每個不同模型的資訊。	請參閱第 39 頁的『成分模型詳細資料』主題，以取得更多資訊。
	資訊	顯示欄位、建置設定及模型估計程序的相關資訊。	請參閱第 36 頁的『模型塊概要/資訊』主題，以取得更多資訊。
設定		使您可以在評分作業中包含信賴度。	請參閱第 104 頁的『決策樹/規則集模型塊設定』主題，以取得更多資訊。
註釋		可讓您新增敘述註釋、指定自訂名稱、新增工具提示文字並指定模型的搜尋關鍵字。	

C&R 樹狀結構、CHAID、QUEST、C5.0 和 Apriori 規則集模型塊

對於關聯規則建模節點 (Apriori) 或某個樹狀結構建立節點（C&R 樹狀結構、CHAID、QUEST 或 C5.0）所探索的特定輸出欄位，規則集模型塊代表用於預測此欄位的規則。對於相關規則，必須從未優化規則塊中產生規則集。對於樹狀結構，可以從互動式樹狀結構建置器、C5.0 模型建立節點或任何樹狀結構模型塊中產生規則集。與未優化規則塊不同，可將規則集塊放置在串流中以產生預測。

執行包含規則集塊的串流時，會將兩個新欄位（分別包含每條記錄對資料的預測值和信賴度）新增到串流中。新欄位名稱透過新增字首從模型名稱衍生。對於關聯規則集，預測欄位的字首是 *$A-*，而信賴度欄位的字首是 *$AC-*。對於 C5.0 規則集，預測欄位的字首是 *$C-*，而信賴度欄位的字首是 *$CC-*。對於 C&R 樹狀結構規則集，預測欄位的字首是 *$R-*，而信賴度欄位的字首是 *$RC-*。在一個數列（可預測相同的輸出欄位）中具有多個規則集塊的串流中，新的欄位名稱將在字首中包含數字，以便彼此區分開來。串流中的第一個關聯規則集塊將使用常用名稱，第二個節點將使用以 *$A1-* 和 *$AC1-* 開頭的名稱，第三個節點將使用以 *$A2-* 和 *$AC2-* 開頭的名稱，依此類推。

規則是如何套用的。從相關規則中產生的規則集與其他模型塊不同，因為對於任何特定記錄，都可以產生多個預測並且這些預測可能並非都同意。可使用兩種方法從規則集中產生預測。

註：不論使用哪種方法，從決策樹中產生的規則集都會傳回相同的結果，因為從決策樹衍生的規則是互斥的。

- **投票。**此方法試圖組合對記錄套用的所有規則的預測。對於每條記錄，會檢查所有的規則，並使用套用至該記錄的每個規則產生一個預測和一個關聯信賴度。計算每個輸出值的信任值總和，具有最大信賴度總和的值將被選作最終預測。最終預測的信賴度是該值的信賴度總和除以對該記錄套用的規則數。
- **第一個符合項。**此方法僅僅是依順序測試規則，並且對記錄應用的第一項規則即為用於產生預測的規則。

可在串流選項中控制所使用的方法。

產生節點。 「產生」功能表使您可以根據規則集建立新節點。

- **「過濾器」節點**建立新的「過濾器」節點以過濾規則集中的規則不使用的欄位。
- **「選取」節點**建立新的「選取」節點以選取對其套用選定規則的記錄。 產生的節點將選取所應用規則的所有條件均為 true 的記錄。此選項需要選定一個規則。
- **規則追蹤節點**建立將計算欄位（用於表示對每條記錄進行預測時所使用的規則）的新 SuperNode。 當使用第一次命中方法評估規則集時，僅用一個表明將發動第一個規則的符號來表示。當使用投票方法評估規則集時，則用一個顯示投票機制的輸入的複合字串來表示。
- **單一決策樹**（畫布）/**單一決策樹**（GM 選用區）。 建立從目前選定的規則中衍生的單個新規則集塊。僅適用於 **增強型** C5.0 模型。請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。
- **從模型到選用區**將模型傳回到模型選用區。 當同事寄給您一個包含模型的資料流，而非模式本身時，這個功能很有用。

註：規則集塊中的「設定」和「摘要」標籤與決策樹模型完全相同。

規則集模型標籤

規則集塊的「模型」標籤中顯示由演算法擷取自資料的規則清單。

規則按結果（預測種類）劃分，並按下列格式顯示：

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

其中 consequent 和 antecedent_1 直到 antecedent_n 都是條件。該規則可解釋為「對於其中 antecedent_1 直到 antecedent_n 都為 true 的記錄，consequent 也可能為 true。」如果按一下工具列上的**顯示實例/信賴度**按鈕，則每個規則還將顯示有關套用該規則的記錄（即前提條件為 true 的記錄）數資訊（**實例數**），及整個規則為 true 的記錄的比例資訊（**信賴度**）。

注意，對於 C5.0 規則集，信賴度的計算方式有些不同。C5.0 使用下列公式計算規則的信賴度：

```
(1 + 規則正確的記錄數)
/
(2 + 規則的前提條件為 true 的記錄數)
```

這一信賴度估計計算方式可調整從決策樹中生成規則（即 C5.0 建立規則集時所執行的操作）的過程。

從 AnswerTree 3.0 中匯入專案

IBM SPSS Modeler 可使用標準的「檔案」>「開啟」對話框匯入在 AnswerTree 3.0 或 3.1 中儲存的專案，示例如下：

1. 從 IBM SPSS Modeler 功能表中選擇：

   檔案 > **開啟串流**

2. 從檔案類型下拉清單中選取 **AT 專案檔案 (*.atp、*.ats)**。

   使用下列節點將匯入的每個專案轉換到 IBM SPSS Modeler 串流中：

   - 一個來源節點，它可定義所使用的資料來源（例如，IBM SPSS Statistics 資料檔案或資料庫來源）。
   - 對於專案中的每個樹狀結構（可能有多個樹狀結構），將建立一個「類型」節點，該節點可為每個欄位（變數）定義內容，包括類型、角色（輸入或者預測工具欄位與輸出或者預測欄位）、遺漏值及其他選項。
   - 對於專案中的每個樹狀結構，將建立一個「分割」節點，該節點可分割訓練或測試樣本的資料，還將建立一個樹狀結構建立節點，該節點可定義用於生成樹狀結構（C&R 樹狀結構、QUEST 或 CHAID 節點）的參數。

3. 要檢視產生的樹狀結構，請執行該串流。

備註

- 不能將在 IBM SPSS Modeler 中產生的決策樹匯出到 AnswerTree 中；從 AnswerTree 匯入 IBM SPSS Modeler 是一個單向過程。
- 將專案匯入到 IBM SPSS Modeler 時，無法保留在 AnswerTree 中定義的利潤。

第 7 章 貝式網路模型

貝葉斯網路節點

貝葉斯網路節點可讓您透過以下方式建立機率模型：結合觀察並記錄的證據與真實世界常識，使用看似不相關屬性以建立發生事件的可能性。該節點側重於樹狀結構擴展素樸貝葉斯 (TAN) 網路和馬爾可夫覆蓋網路，這些網路主要用於分類。

貝葉斯網路可用於在多數不同的狀況下進行預測，範例如下：

- 選取違約風險較低的貸款時機。
- 根據感應器輸入和現有記錄，估計設備何時需要維修、增加零件或更換。
- 借助線上疑難排解工具解決客戶問題。。
- 即時診斷並排除移動電話網路故障。
- 評估研發專案的潛在風險和回報，以在最佳時機集中資源。

貝葉斯網路是一種圖形模型，可顯示資料集中的變數（通常稱之為**節點**）以及這些變數之間的機率或條件獨立性。貝葉斯網路可呈現節點之間的因果關係；但是，網路中的鏈結（也稱為**弧**）不一定呈現直接因果關係。例如，如果圖表中所顯示的症狀和疾病之間的機率性的獨立性成立，貝葉斯網路可根據特定症狀和其他相關資料是否存在，計算病患患有某種特殊疾病的幾率。這種網路非常穩健，即使在資訊遺漏時，也可以利用現有的任何資訊作出最佳預測。

標準的基礎貝葉斯網路範例由 Lauritzen 和 Spiegelhalter 於 1988 年建立。該網路示例是一種簡化的網路版本，通常稱作 "Asia" 模型，醫生可用它來診斷新病患的病情，所有鏈結的方向可大體指示因果關係。每個節點代表與病患狀況相關的一個方面，例如「吸煙」代表這些病患確為吸煙者，而 "VisitAsia" 代表他們最近是否去過亞洲。機率關係由所有節點之間的鏈結指示，例如，吸煙會增大病患患有支氣管炎和肺癌的幾率，而年齡僅與肺癌的患病率相關。同樣地，肺部 x 光檢查異常可能由肺結核或肺癌引起。同時，如果病患本身患有支氣管炎或肺癌，那麼他們更有可能出現呼吸急促（呼吸困難）症狀。

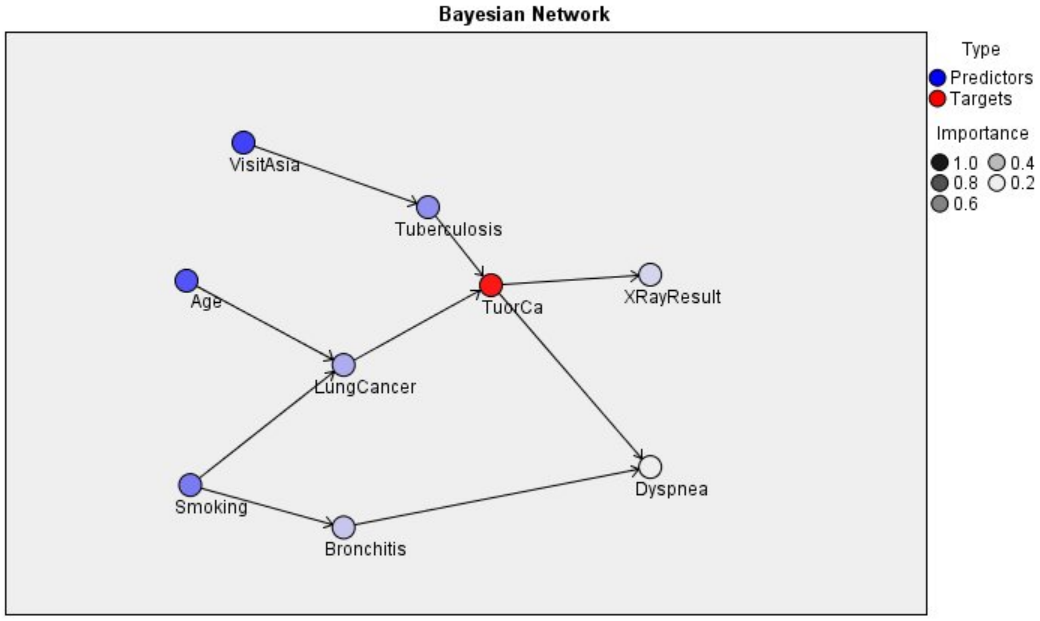


圖 *29. Lauritzen* 和 *Spegelhalter* 的 *Asia* 網路範例

以下是您有可能決定使用貝葉斯網路的幾點原因：

- 它可協助您瞭解因果關係。由此，您可以瞭解出現問題的地方並可預測任何干涉可能引發的後果。
- 該網路可提供避免資料過度配適的有效方法。
- 可以輕鬆地觀測到所涉及關係的清晰視圖。

需求。目標欄位必須為種類且測量層次為名義、序數或旗標。輸入內容可以為任何類型的欄位。連續（數值型範圍）輸入欄位將自動分級；但是，如果分佈出現不對稱，則可使用貝葉斯網路節點之前的分組節點對欄位進行手動分級，從而獲得較好的效果。例如，在**管理員欄位**與貝葉斯網路節點**目標**欄位相同的位置處，使用最佳化分級。

範例。 銀行分析師希望可以預測可能拖欠償還貸款的客戶或潛在客戶。您可使用貝式網路模型來識別最有可能拖欠還款的客戶的特性，並建立幾種不同類型的模型，以確定哪種模型可以最好地預測潛在的貸款拖欠者。

範例。 一位電信運營商希望減少中斷服務（又稱為「流失」）的客戶數量，並使用上一個月的資料對模型每月進行更新。您可以使用貝式網路模型確定最有可能流失的客戶的特性，然後每月使用新資料繼續訓練該模型。

貝葉斯網路節點模型選項

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。

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。

建置每個分割的模型。 針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。請參閱第 24 頁的『建立分割模型』主題，以取得更多資訊。

分割區。通過此欄位，您可以指定用於針對模型建置中的訓練、測試和驗證階段將資料劃分為不同樣本的欄位。透過使用一個樣本來產生模型，並使用另一個樣本來測試模型，您可以很好地指出模型將概化為與現行資料相似的更大型資料集的程度。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。）

分割。 針對分割模型，選取一或多個分割欄位。這與在「類型」節點中將欄位角色設為分割類似。您可以只將測量層次為**旗標**、**標稱**、**序數**或**連續**的欄位指定為分割欄位。選擇作為分割欄位的欄位無法用作目標、輸入、分割區、頻率或加權欄位。請參閱第 24 頁的『建立分割模型』主題，以取得更多資訊。

繼續訓練現有模型。 如果選取此選項，則在模型塊「模型」標籤上顯示的結果，將在每次執行模型時重新產生和更新。例如，如果已為現有模型新增新的或更新的資料來源，則需要執行此操作。

註：此操作只能更新現有網路；它無法新增或者移除節點或連線。每次重新訓練模型時，網路的形狀都將保持不變，只會變更條件機率和預測值重要性。如果新資料與舊資料大致相似也無妨，因為您所預期的是關注相同的內容；但是，如果您希望檢查或更新 重要的內容（針對其重要程度），則需要建立新模型，即建立新網路。

結構類型。選取建立貝葉斯網路時要使用的結構：

- **TAN。**樹狀結構擴展素樸貝葉斯模型 (TAN) 用於建立簡單的貝葉斯網路模型，後者是對標準素樸貝葉斯模型的改善。這是由於該模型容許每一個預測值除了依賴於目標變數之外，還依賴於其他預測值，由此增加了分類的準確度。
- **馬爾可夫覆蓋。**此結構用於選取資料集中的節點的集合，這些節點包含目標變數的父項、其子項以及子項的父項。馬爾可夫覆蓋基本可以確定網路中預測目標變數的所需的所有變數。用戶認為這種建立網路的方法更為準確；但是，當正在處理大型資料集時，由於所包含的變數數目較多，所以可能會消耗許多正在處理時間。要減少正在處理工作數量，可以使用「專家」標籤上的**功能選擇**選項，選擇與目標變數有重大相關性的變數。

包含功能選擇前置處理步驟。選擇該框，您可以使用「專家」標籤上的**功能選擇**選項。

參數學習方法。貝葉斯網路參數是指給定每個節點的父項值時，該節點具有的條件機率。有兩種可能的選擇，您可以用來控制估算節點（此處父項值已知）間條件機率表格這一作業。

- **最大概似法。**使用大型資料集時，請選中此框。這是預設選項。
- **對小儲存格計數的貝葉斯調整。**對於較小的資料集，存在模型過度配適的風險以及出現大量零計數的可能性。選中此選項可通過套用平滑來減少任何零計數以及不可靠的估計結果帶來的效應，從而解決這些問題。

貝葉斯網路節點專家選項

使用節點專家選項可微調模型建立過程。若要存取專家選項，請在「專家」標籤上將「模式」設為**專家**。

遺漏值。 依預設，IBM SPSS Modeler 只會將具有有效值的記錄用於模型中所使用的所有欄位。（這有時稱為遺漏值的**整批 (Listwise) 刪除**。）如果您有很多遺漏資料，則可能會發現此方法會刪除太多記錄，從而留給您的資料不足以產生良好的模型。在這種情況下，可以取消選中**僅使用完整記錄**選項。IBM SPSS Modeler 隨後會嘗試使用盡可能多的資訊來估計模型，其中包括部分欄位具有遺漏值的記錄。（有時候，這稱為成對刪除遺漏值。）但是，在一些情況下，以這種方式使用不完整記錄可能會導致評估模型時發生計算問題。

附加所有機率。 指定是否將每個種類的輸出欄位的機率新增至節點所處理的每筆記錄。如果未選取此選項，則只新增預測種類的機率。

獨立性測試。獨立性測試會評量兩個變數的配對觀察是否彼此獨立。請從以下可用選項中選取要使用的測試類型：

- **概似比。** 通過計算兩種不同假設下結果的上限機率之間的比例來測試目標與預測值之間的獨立性。

- **Pearson 卡方**。通過使用虛無假設（所觀察事件的相對出現頻率遵循指定的頻率分佈）來測試目標與預測值之間的獨立性。

貝式網路模型可處理在測試配對以外使用了附加變數的條件獨立性測試。此外，模型不僅可以研究目標和預測值之間的關係，還可研究預測值自身之間的關係。

註：只有在「模型」標籤上選中馬爾可夫覆蓋的**包含功能選擇前置處理步驟**或**結構類型**時，獨立性測試選項才可用。

顯著性層級。 與獨立性測試設定結合使用，可讓您設定要在處理測試時使用的截斷值。該值越小，網路中的鏈結就越少；預設層次值為 0.01。

註：唯有當您選取**包括功能選擇預先處理步驟**或「模型」標籤上 Markov Blanket 的**結構類型**，才能使用此選項。

上限條件集大小。該演算法用於建立馬爾可夫覆蓋結構，它使用大小不斷增加的條件集來執行獨立性測試並從網路中移除不需要的鏈結。由於包含大量條件變數的測試需要更多的時間和記憶體進行處理，因此您可以限制要併入的變數數目。尤其是在處理眾多變數間具有較強相依關係的資料時，這非常有用。但請注意，最終形成的網路可能包含一些多餘鏈結。

指定執行獨立性測試時要使用的條件變數的上限號碼。預設設定為 5。

註：唯有當您選取**包括功能選擇預先處理步驟**或「模型」標籤上 Markov Blanket 的**結構類型**，才能使用此選項。

功能選擇。通過這些選項，您可以限制在正在處理模型時所使用的輸入數，以便加快模型建置過程。由於在建立馬爾可夫覆蓋結構時存在大量的潛在輸入，因此該操作特別有用；通過此項操作，您可以選取與目標變數有重大關聯的輸入。

註：只有在「模型」標籤上選中**包含功能選擇前置處理步驟**時，功能選擇選項才可用。

- **始終選取輸入**。通過使用「欄位選擇器」（文字欄位右側的按鈕），從資料集中選擇建立貝式網路模型時始終使用的欄位。一律會選取目標欄位。請注意，如果其他測試認為某些項目不重要，那麼在模型建置過程中，貝葉斯網路可能仍然會從此清單中刪除這些項目。因此，該選項僅確保清單中的項目用在模型建置程序中，而不確保它們絕對顯示在生成的貝葉斯模型中。
- **上限輸入數。** 指定建立貝式網路模型時要使用的來自資料集中的總輸入數。您可以輸入的最大號碼為資料集中的總輸入量。

  註：如果在**始終選取輸入**中選取的欄位個數超過**上限輸入數**的值，那麼將顯示一條錯誤訊息。

貝式網路模型塊

註：如果在建模節點的「模型」標籤中選中了**繼續訓練現有參數**，那麼將在每次重新產生模型時更新模型塊的「模型」標籤上顯示的資訊。

模型塊「模型」標籤分為兩個窗格：

左側窗格

基本：此視圖包含節點網路圖表，此圖表顯示目標與其最重要的預測值之間的關係，以及各預測值之間的關係。 各預測值的重要性可通過其顏色的深淺顯示；顏色越深表示預測值越重要，反之亦然。

當您將滑鼠指標懸停在節點上時，工具提示中會顯示代表範圍的節點的 Bin 值。

可以使用 IBM SPSS Modeler 中的圖表工具進行交互、編輯，並儲存圖表。例如，可以在其他應用程式（如 MS Word）中使用圖表。

提示：如果網路包含大量節點，那麼可以按一下以選中某個節點，然後拖曳它以使圖表更加清晰。

分佈：此視圖將以微型圖表顯示網路中每個節點的條件機率。 將滑鼠懸停在圖表上方，可在工具提示中顯示條件機率值。

右窗格

預測值重要性：這將顯示一個圖表，以指示在估計模型時所使用的各個預測值的相對重要性。 如需相關資訊，請參閱第 37 頁的『預測值重要性』。

條件機率：當在左窗格中選取了某個節點或微型分佈圖時，右窗格則會顯示相關的條件機率表格。 該表格包含各個節點值的條件機率值，以及各節點的母節點中的值組合。此外，該表還包含為每個記錄值和母節點中各個值組合所觀測的記錄數量。

貝式網路模型設定

在貝式網路模型塊的「設定」標籤中可指定選項以修改已建立的模型。例如，可以通過貝葉斯網路節點使用相同的資料和設定建立幾個不同的模型，然後使用每個模型中的此標籤對設定稍做修改以查看其對結果的影響。

註：僅當模型片段已新增至串流之後，此標籤才可用。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

附加所有機率 指定是否將每個種類的輸出欄位的機率新增至節點所處理的每筆記錄。如果未選取此選項，則只新增預測種類的機率。

此勾選框的預設設定由建模節點的「專家」標籤上的相應勾選框確定。請參閱第 111 頁的『貝葉斯網路節點專家選項』主題，以取得更多資訊。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

貝式網路模型摘要

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。當結束對項目的檢視時，為了隱藏結果，可使用展開控制項摺疊要隱藏的特定結果，或按一下**全部收合**按鈕摺疊所有結果。

分析。 顯示特定模型的相關資訊。

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。

第 8 章 神經網路

神經網路可以近似多種預測模型，而對模型結構和假設只有最小需求。關係的形式是在學習過程中確定的。如果目標與預測值之間的線性關係合適，那麼類神經網路的結果應該與傳統線性模型的結果非常相似。 如果非線性關係更為恰當，則神經網路會自動近似「正確的」模型結構。

此靈活性的缺點是，不容易對類神經網路進行解釋。如果要嘗試解釋在目標與預測值之間生成關係的底層過程，那麼最好使用更為傳統的統計模型。但是，如果模型可解釋性並不重要，那麼使用類神經網路可以獲得良好的預測。

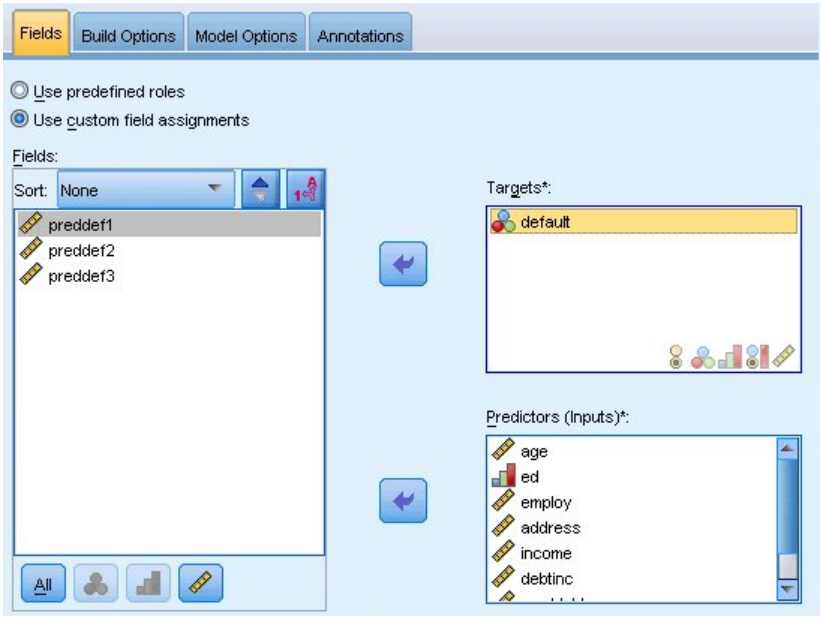


圖 *30.* 欄位標籤

欄位要求。必須至少有一個目標字段和一個輸入字段。將忽略設定為「兩者」或「無」的欄位。對於目標或預測值（輸入），沒有測量層次限制。如需相關資訊，請參閱第 26 頁的『建模節點欄位選項』。

在模型建置期間指派給神經網路的起始加權，以及因此產生的最終模型，視資料中的欄位順序而定。SPSS Modeler 會先按欄位名稱自動排序資料，然後再將其提供給神經網路進行訓練。這表示在模型建置器中設定隨機種子時，明確變更資料上游中欄位的順序不會影響產生的神經網路模型。但以變更排序順序的方式變更輸入欄位名稱會產生不同的神經網路模型，即使在模型建置器中設定了隨機種子也是如此。使用不同的欄位名稱排序順序不會顯著影響模型品質。

神經網路模型

神經網路是神經系統運轉方式的簡式模型。其基本單元是**神經元**，通常將其組織到**層**中，如下面的圖所示。

© Copyright IBM Corp. 1994, 2019

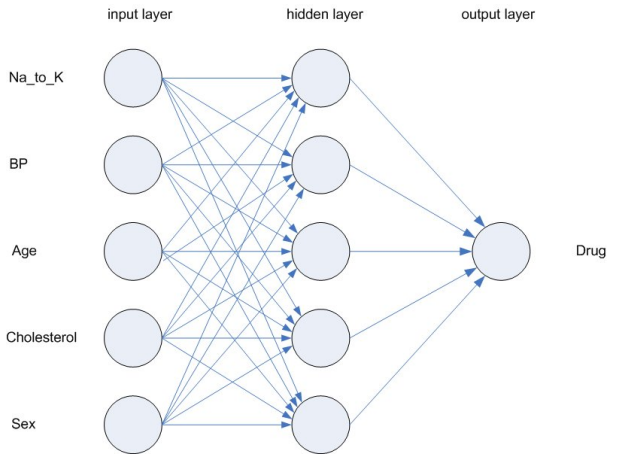


圖 *31.* 類神經網路的結構

類神經網路是模擬人類大腦處理資訊方式的簡化模型。此模型的工作方式為模擬大量類似於神經元的抽象形式的互連處理裝置。

這些正在處理單元都位於層中。類神經網路通常包含三個部分：**輸入層**，其中的單元代表輸入欄位；一個或多個**隱藏層**；一個**輸出層**，帶有一個或多個代表目標欄位的單元。這些單元通過可變的連線強度（或**加權**）連線。輸入資料 顯示在第一層，其值從每個神經元傳播到下一層的每個神經元。最終從輸出層中輸出結果。

該網路可通過以下過程進行學習，即檢查單個記錄，然後為每條記錄產生預測，並且當產生的預測不正確時，對加權進行調整。在符合一個或多個停止準則之前，此過程會不斷重複，而網路會持續提高其預測準確度。

最初，所有的加權都是隨機生成的，並且從網絡輸出的結果很可能沒有意義的。網路可通過**訓練**來學習。向該網路重複應用已知道結果的範例，並將網路給出的結果與已知的結果進行比較。從此比較中得出的資訊會傳送回網路，並逐漸改變加權。隨著訓練的進行，該網路對已知結果的抄寫會變得越來越準確。一旦訓練完畢，就可以將網路套用到未知結果的未來案例中。

對舊式串流使用神經網路

IBM SPSS Modeler 版本 14 引入了新的類神經網路節點，支援增強和組裝技術，並可針對大型資料集進行最佳化。在較新的發行版中，包含舊節點的現有串流可能仍會建置模型並進行評分。但是，未來的發行版將移除此支援，因此我們建議您使用新版本。

從第 13 版以後，帶有未知值的欄位（即，值在訓練資料中不出現）不再自動按照遺漏值進行處理，而是使用 `$null$` 值進行評分。因此，如果要在第 13 版或更高版本中使用第 13 版以前的舊類神經網路模型將具有未知值的欄位分數為非無效，那麼應該將未知值標示為遺漏值（例如，通過使用「類型」節點完成此任務）。

請注意，為了實現相容性，任何仍然包含舊節點的遺存串流仍然可以套用**工具 > 串流內容 > 選項**中的限制集合大小選項；此選項僅適用於第 14 版以後的 Kohonen 網路和 *K* 「平均數」節點。

目標

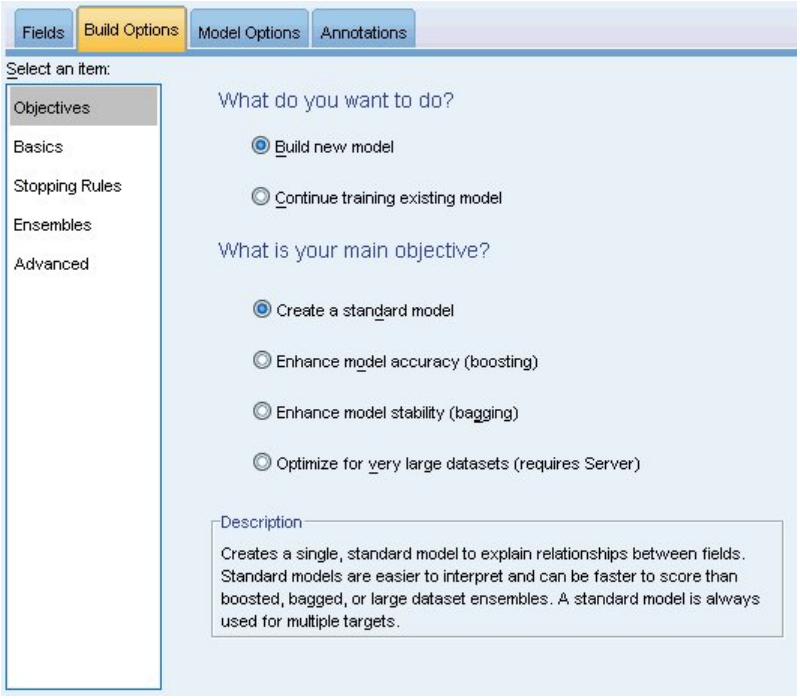


圖 32. 目標設定

您現在要進行什麼作業？

- **建立新模式。** 建立全新的模式。此為一般的節點作業。
- **繼續執行現有模型的訓練。** 系統會繼續使用節點上次成功產生的模式來執行訓練。這可讓您更新或重新整理現有模式，而無需存取原始資料，此外由於資料流中僅會容納全新或已更新的記錄，因此會使效能速度顯著提升。系統會使用建模節點來儲存上一個模型的詳細資料，即使資料流或「模型」色板中已不再提供上個模型塊，也能使用此選項。

  註：若啟用此選項，則系統會停用「欄位」和「建置選項」標籤上的其他所有控制項。

您的主要目標是什麼？ 選取適當的目標。

- **建立標準模式。** 此方法會建立單一模式，以預測使用預測值的目標。一般而言，標準模式在解讀上較為容易，且評分速度比 Boosted、Bagged 或大型資料集集合更快。

  註：對於分割模型，若要搭配使用此選項與**繼續訓練現有模型**，您必須連接到 Analytic Server。
- **強化模式準確性 (Boosting)。** 此方法會使用 Boosting 來建立集合模式，其會產生一系列的模式，以取得更為準確的預測。集合的建立和評分時間均長於標準模式。

  Boosting 會產生一系列的「元件模型」，其中每一個都建置於整個資料集。在建置每一個系列元件模型之前，會根據前一個元件模型的殘差加權記錄。對於具有較大殘差的觀察值一般會給定較高的分析加權，因此下一個元件模型將更加注重於預測這些記錄。與這些元件模型一起組成複合模型。複合模型會使用結合規則對新記錄進行評分；可用的規則取決於目標的測量水準。
- **強化模式穩定性 (Bagging)。** 此方法會使用 Bagging (重複取樣整合) 來建立集合模式，其會產生多個模式，以取得更為可靠的預測結果。集合的建立和評分時間均長於標準模式。

引導聚集 (bagging) 會透過取樣產生訓練資料集的複本，以替換原始資料。這樣做可建立與原始資料集大小相同的引導樣本。然後，在每一個複本上建置「元件模型」。與這些元件模型一起組成複合模型。複合模型會使用結合規則對新記錄進行評分；可用的規則取決於目標的測量水準。

- **為大型資料集建立模型。**此方法會將資料集分割為個別的資料區塊，以建立集合模式。若資料集過大而無法建置上述任何模型，或是想要建立增量模型，請選擇此選項。此選項的建立時間較短，但評分時間會長於標準模式。

存在多個目標時，此方法將僅建立標準模型，而不考慮選定的目標。

基本

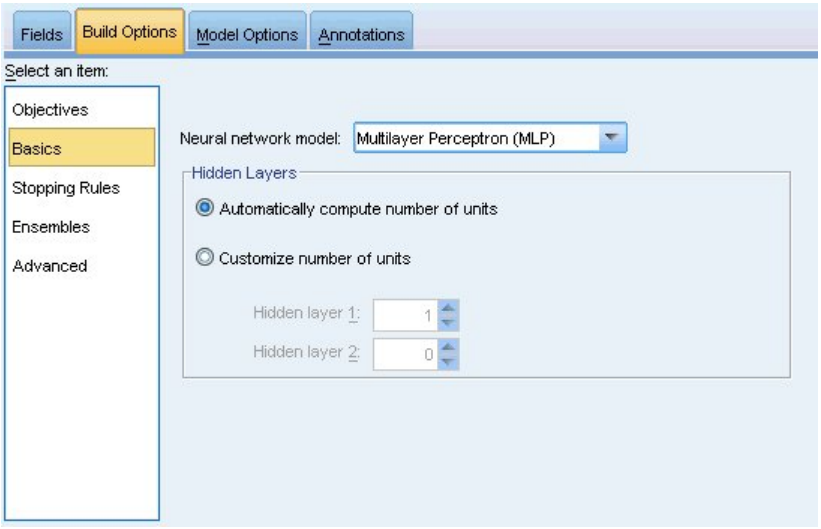


圖 *33.* 基本設定

神經網路模型。此類模型用於確定神經網路如何通過隱藏層將預測值連接到目標。**多層感知器(MLP)**容許構建較為複合的關係，但代價是更長的訓練與評分時間。**徑向基底函數(RBF)**可以縮短訓練與評分時間，但與 MLP 相比其預測能力要差些。

隱藏層。類神經網路的隱藏層包含無法觀察到的單元。每個隱藏單元的值都是預測值的某個函數；此函數的準確形式部分取決於網路類型。多層感知器可能有一個或兩個隱藏層；徑向基底函數網路可以有一個隱藏層。

- **自動計算單位數。**此選項建立具有單個隱藏層的網路，並計算隱藏層中的「最佳」單元數。
- **自訂單位數。**此選項容許您指定每個隱藏層中的單位數。第一個隱藏層必須至少具有一個單元。對第二個隱藏層指定 0 個單位將建立具有單個隱藏層的多層感知器。

  註：在選擇值時，應確保節點數目不超過連續預測值數加上所有種類（旗標、名義和序數）預測值中的種類總數之和。

停止規則

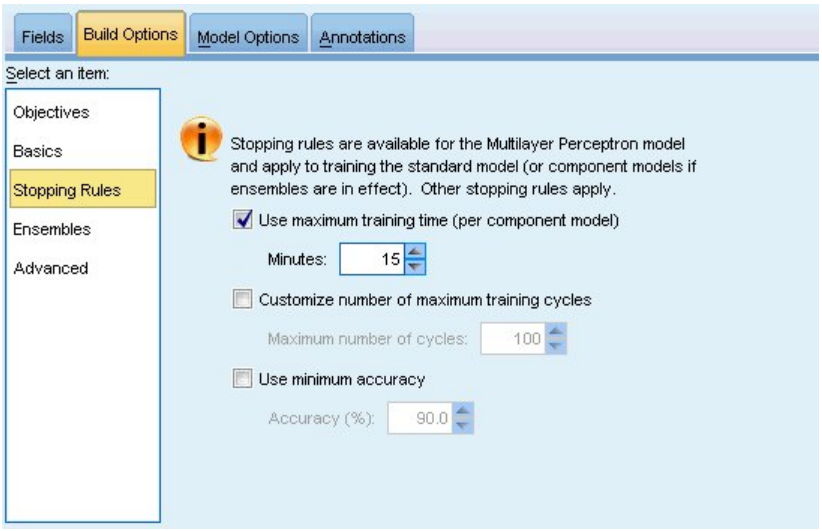


圖 *34.* 「中止規則」設定

這些規則用於確定何時停止訓練多層感知器網路；使用徑向基底函數演算法時，將忽略這些設定。訓練將至少進行一個循環（資料傳遞），然後可以根據下列準則中止訓練。

使用訓練時間上限（每個成份模型）。選擇是否要指定演算法執行的最大分鐘數。請指定一個大於 0 的數字。構建總體模型時，這是總體的每個成份模型所容許的訓練時間。請注意，訓練時間可能會超出指定的時間限制一點點才能完成現行週期。

自訂訓練週期數目上限。容許的上限訓練循環數。如果超過週期數目上限，那麼訓練將中止。請指定一個大於 0 的整數。

使用精確度下限。使用此選項，訓練將繼續進行，直至獲得指定的精確度為止。這種情況可能永遠不會發生，但您可以隨時岔斷訓練，並儲存目前為止達到的最精確的淨值。

如果防止過度擬合集合中的誤並未在每個循環後減小，訓練誤中的相對變化較小，或者目前訓練誤與初始誤相比較小，那麼訓練演算法也將中止。

總體

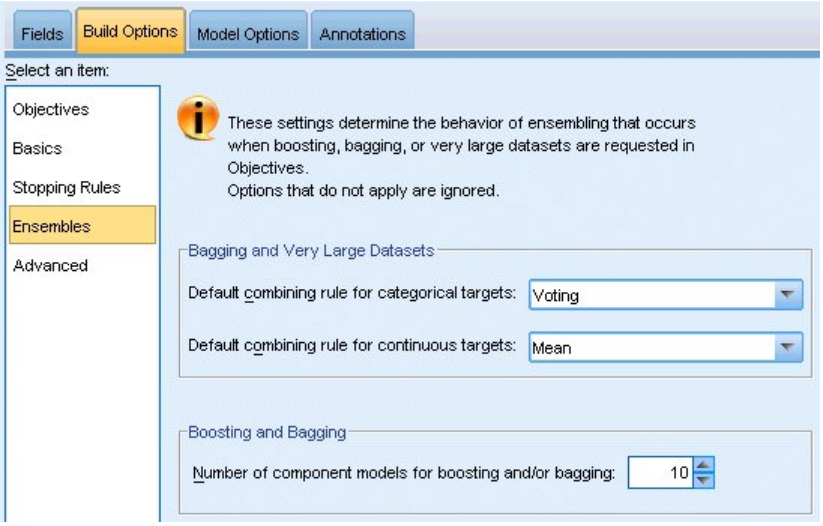


圖 *35.* 總體設定

系統在「目標」中要求 boosting、bagging 或極大資料集時，這些設定會決定所發生的集合行為。系統會忽略無法套用至所選目標的選項。

Bagging 與極大資料集。 系統執行集合評分時，可使用此規則來合併基底模型的預測值，以運算集合分數值。

- **種類目標的預設組合規則。**可以通過投票、最高機率或最高平均值機率來對種類目標的總體預測值進行已結合。**投票**會選取所有基底模式中最常擁有最高機率的類別。**最高機率**會在所有基底模式中選取達到單一最高機率的類別。**最高平均數機率**會在平均計算所有基底模型的類別機率時，選取具有最高值的類別。
- **連續目標的預設合併規則。** 系統會使用基底模型預測值的平均數或中位數，來合併連續目標的集合預測值。

請注意，若目標是用於強化模式準確性，則系統會忽略合併規則選擇。Boosting 會一律使用大部分的加權投票來為類別目標評分，並且使用加權中位數來為連續目標評分。

Boosting 與 Bagging。 當目標用於強化模式準確性或穩定性時，可指定欲建立的基底模式數目；若為 bagging，則此為重複取樣範例的數目。其應為正整數。

進階

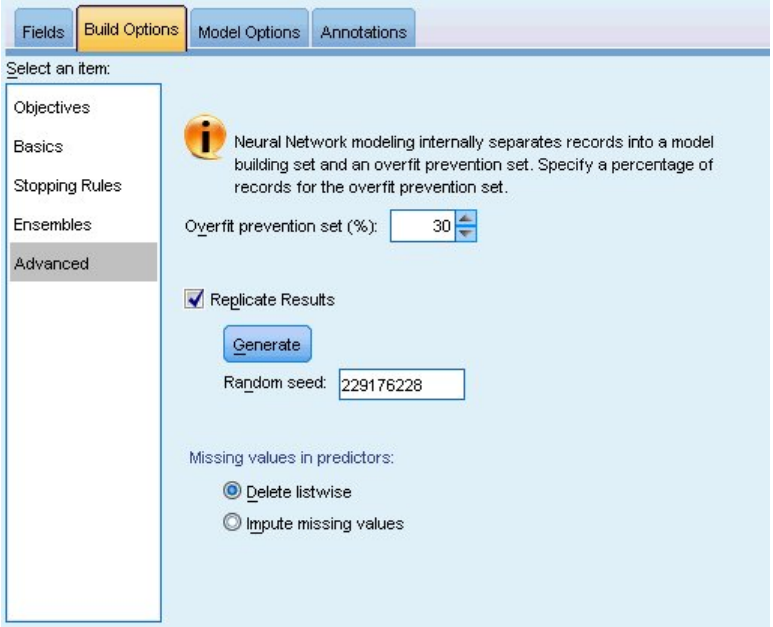


圖 *36.* 進階設定

進階設定提供對無法很好地歸入其他設定群組的選項的控制。

過適預防集。 類神經網路方法在內部將記錄劃分為模型建置集和防止過度擬合集合，後者作為不相依的資料記錄集，用於追蹤訓練過程中的錯誤，以防止該方法對資料中的幾率變異進行建模。指定記錄百分比。預設值是 30。

複製結果。 設定亂數種子以供您複製分析。請指定一個整數，或是按一下「**產生**」以建立介於 1 和 2147483647 之間（含）的虛擬亂數整數。依預設，這些分析以種子值 229176228 進行抄寫。

預測值中遺漏值。這將指定如何處理遺漏值。**整批刪除**將在預測值上存在遺漏值的記錄從模型建置中排除。**轉嫁遺漏值**將取代預測值中遺漏值，並在分析中使用這些記錄。連續欄位插補下限和上限觀察值的平均值；種類欄位插補最經常出現的種類。請注意，將始終從模型建置中移除「欄位」標籤上指定的任何其他欄位中包含遺漏值的記錄。

模型選項

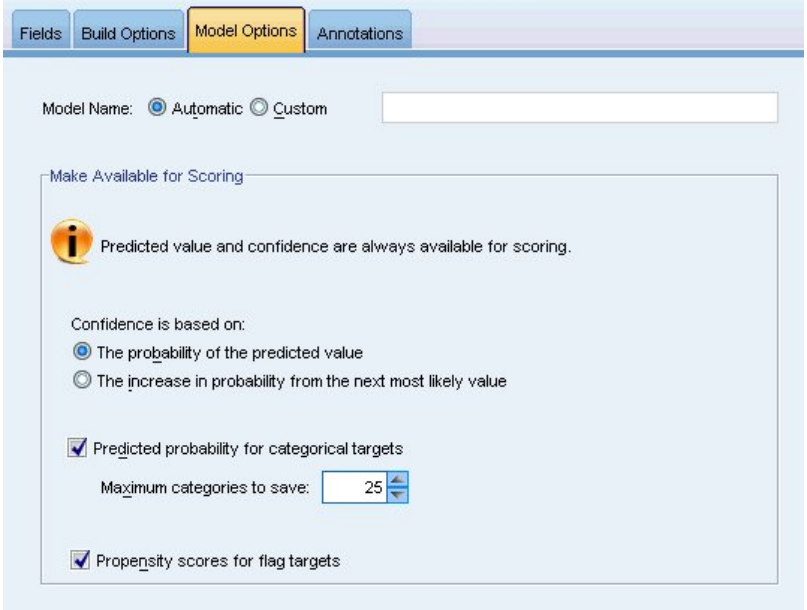


圖 *37.* 模型選項標籤

模型名稱。您可以根據目標欄位來自動產生模型名稱，或是指定自訂名稱。自動產生的名稱為目標欄位名稱。如果存在多個目標，則模型名稱是依序排列的欄位名稱，名稱之間以 '&' 符號連接。例如，如果目標為 *field1*、*field2* 和 *field3*，那麼模型名稱為 *field1 & field2 & field3*。

可用於評分。對模型進行評分時，應生成此群組中的選定項目。對模型評分時，一律會計算所有目標的預測值以及類別目標的信賴度。所計算的信賴可以基於預測值的機率（最高預測機率），或是基於最高預測機率與次高預測機率之間的差異。

- **類別目標的預測機率。** 這會產生類別目標的預測機率。會為每一個類別建立一個欄位。
- **旗標目標的傾向評分。** 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。該模型會產生原始傾向評分；如果分割區有效，則該模型也會根據測試分割區產生調整傾向評分。

模型摘要

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

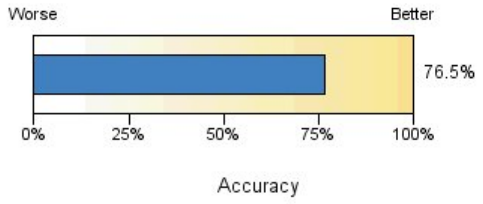


圖 *38.* 神經網路模型摘要視圖

「模型摘要」視圖是一個 Snapshot，即類神經網路預測或分類精確度的概覽摘要。

模型摘要。 此表格識別目標、已訓練的類神經網路類型、中止訓練的中止規則（已訓練多層感知器網路時顯示），以及網路的每個隱藏層中的神經元數。

神經網路品質。此圖表顯示最終模型的精確度，數值越大越好。對於種類目標，此指數只是預測值與觀察值相符的記錄所佔的百分比。對於連續目標，精確度指定為 R^2 值。

多個目標。如果有多個目標，那麼每個目標都將顯示在表格的**目標**列中。圖表中顯示的精確度是各個目標精確度的平均值。

預測值重要性

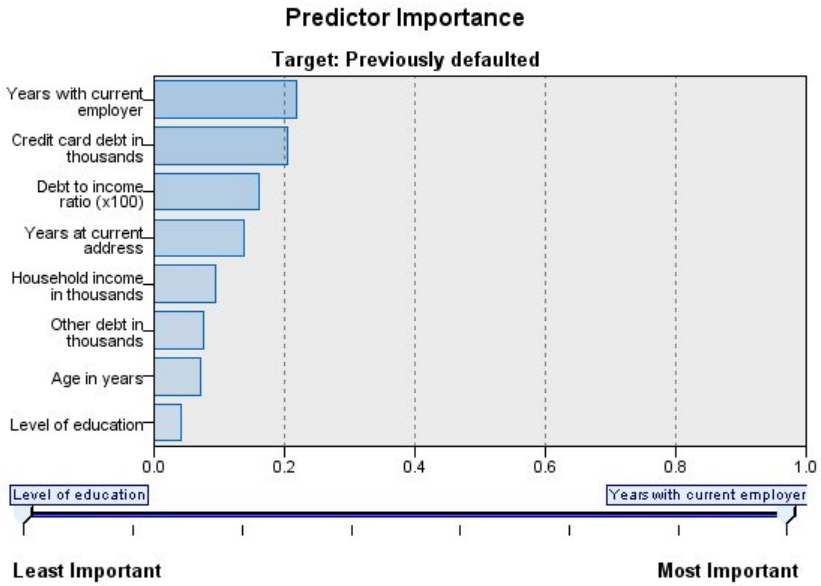


圖 *39.* 「預測值重要性」視圖

一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

多個目標。如果存在多個目標，則每一個目標都會顯示在個別圖表中，並會提供**目標**下拉清單，用於控制要顯示的目標。

按已觀測進行預測

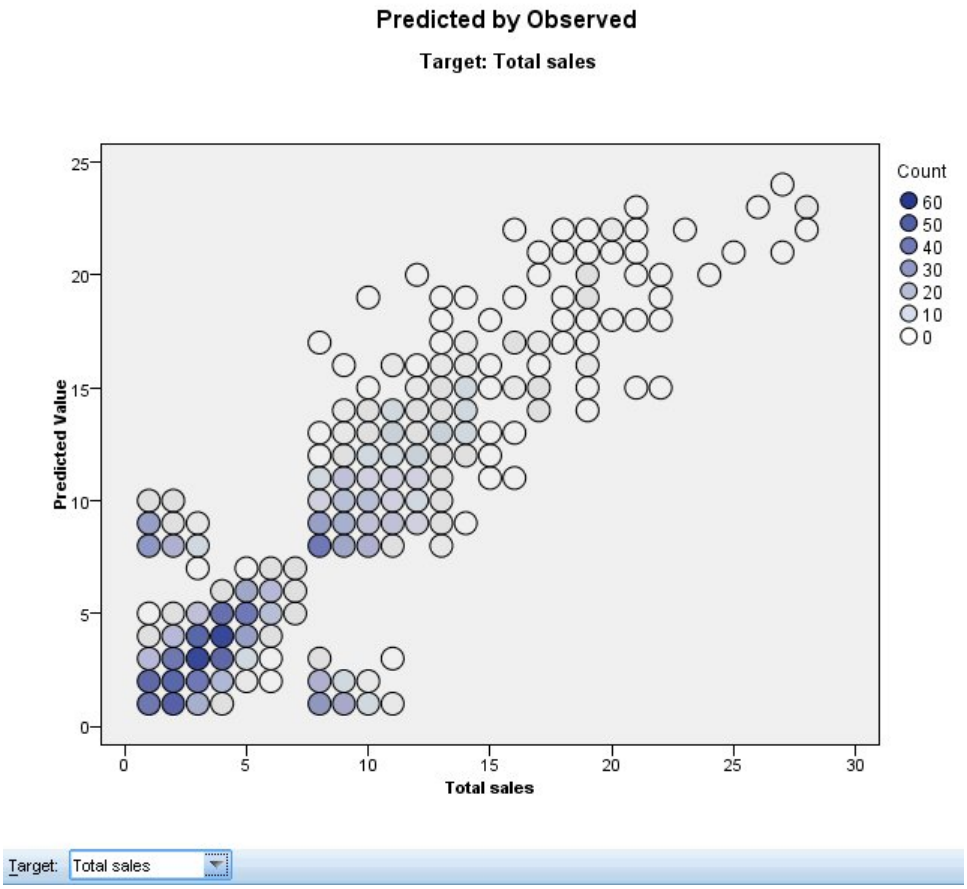


圖 *40.* 「預測值與觀察值」視圖

對於連續目標，這將顯示預測值位於垂直軸上，而觀察值位於水平軸上的離散化散佈圖。

多個目標。如果存在多個連續目標，則每一個目標都會顯示在個別圖表中，並會提供**目標**下拉清單，用於控制要顯示的目標。

分類

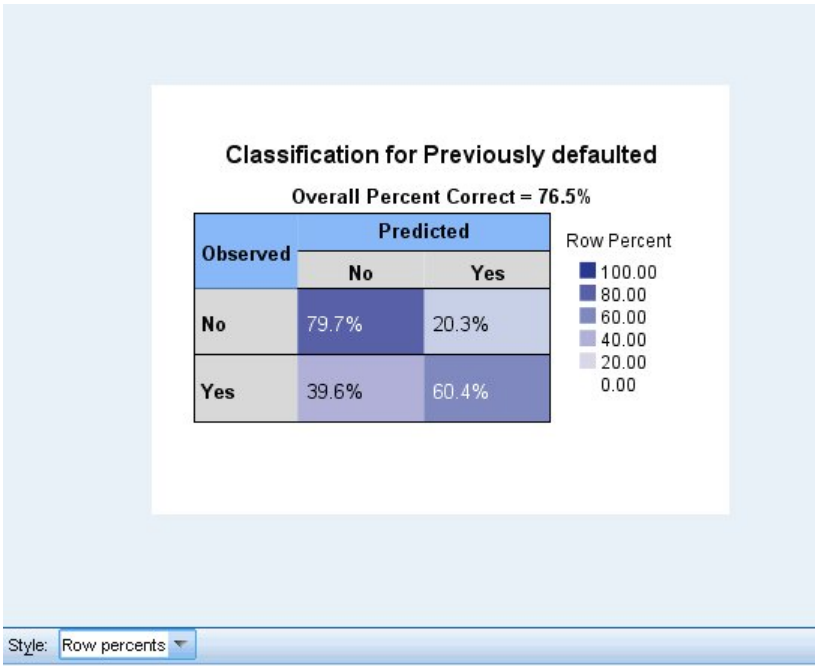


圖 *41.* 分類視圖，列百分比樣式

對於類別目標，這會顯示熱圖中觀察值相對於預測值的交叉分類，以及整體正確百分比。

表格樣式。提供幾種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **列百分比。**這會顯示儲存格中的列百分比（以列總和的百分比表示的儲存格個數）。此為預設值。
- **儲存格個數。**這會顯示儲存格中的儲存格個數。熱圖的陰影仍會以列百分比為基礎。
- **熱圖。**這不在儲存格中顯示值，只會顯示陰影。
- **壓縮。**這不會在儲存格中顯示列標題、直欄標題或值。這在目標具有大量類別時很有用。

遺漏值。如果目標中有任何記錄具有遺漏值，則這些記錄會顯示於所有有效列下方的（**遺漏值**）列中。具有遺漏值的記錄不會納入整體百分比修正中。

多個目標。如果有多個類別目標，則每個目標都會顯示在個別的表格中，並有**目標**下拉清單可控制要顯示的目標。

大型表格。如果顯示的目標具有 100 個以上類別，則不會顯示表格。

網路

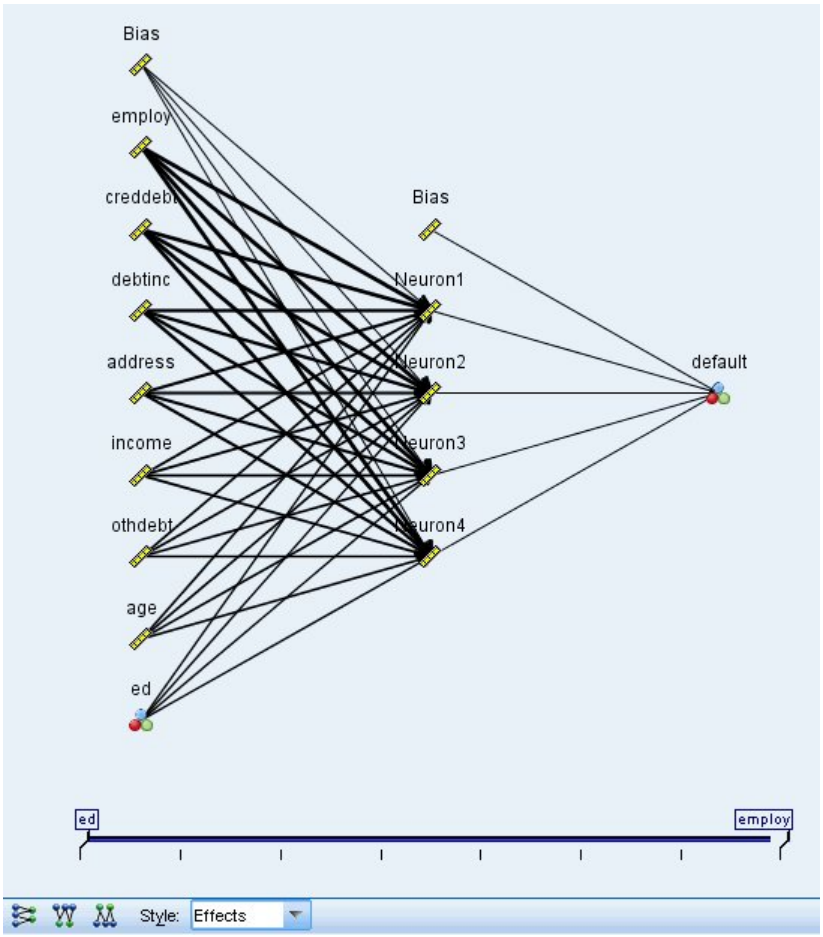


圖 *42.* 網路視圖，左側的輸入，作用樣式

這將顯示類神經網路的圖形表示法。

圖表樣式。有兩種不同的顯示樣式，可以從**樣式**下拉清單中進行存取。

- **效應。**這會在圖中將每個預測值與目標顯示為單個節點，而不遵循測量尺度是連續還是種類。此為預設值。
- **係數。**這將為種類預測值與目標顯示多個指示節點。係數樣式圖中的連接線根據突觸加權的估計值進行著色。

圖方向。依預設，輸入位於網路圖的左側，而目標位於右側。通過使用工具列控制項，您可以變更方向，以使輸入位於頂部而目標位於底部，或者輸入位於底部而目標位於頂部。

預測值重要性。在圖中，連接線條根據預測值的重要性進行加權，粗線條表示重要性較高。工具列中有一個「預測值重要性」調節器，用於控制項網路圖中顯示的預測值。這不會變更模式，僅會讓您更能著重於最重要的預測值。

多個目標。如果存在多個目標，那麼所有目標都將顯示在圖表中。

設定

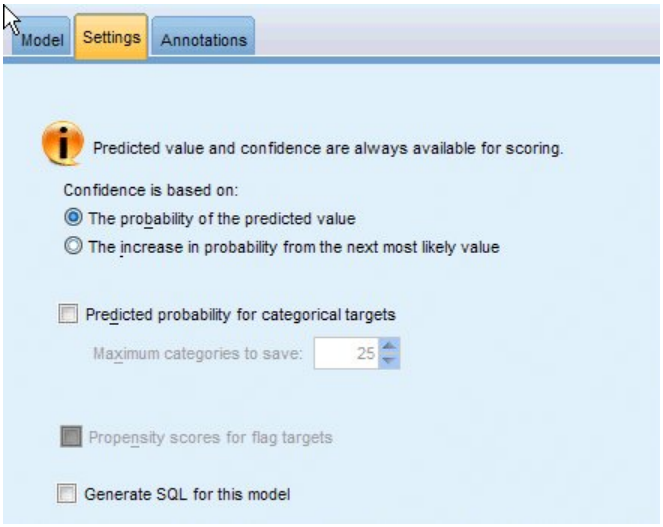


圖 *43.* 「設定」標籤

在對模型分數時，應生成此標籤中的選定項目。對模型評分時，一律會計算所有目標的預測值以及類別目標的信賴度。所計算的信賴可以基於預測值的機率（最高預測機率），或是基於最高預測機率與次高預測機率之間的差異。

- **類別目標的預測機率。** 這會產生類別目標的預測機率。會為每一個類別建立一個欄位。
- **旗標目標的傾向評分。** 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。該模型會產生原始傾向評分；如果分割區有效，則該模型也會根據測試分割區產生調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

透過轉換為原生 SQL 進行評分 如果選取此項，則會產生原生 SQL 來在資料庫內對模型進行評分。

註：雖然這個選項可以更快地提供結果，但隨著模型複雜性的增加，原生 SQL 的大小和複雜性也會增加。

在資料庫外部評分 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

第 9 章 決策清單

決策清單 模型識別了子群組或**區段**，即，顯示了與整體樣本相關的二值（yes 或 no）結果的概似度的高低。例如，您或許在尋找那些最不可能流失的客戶或最有可能對某個商業活動作出積極響應的客戶。通過 決策清單檢視器 可以實現對模型的完全控制，它允許您編輯區段、新增自己的商業規則、指定每個區段的分數方式，以及採用其他多種方式自訂模型從而對所有區段的匹配比例進行最佳化。因此，它尤其適用於產生郵件清單，或確定作為特定活動目標的記錄。此外，還可以使用多個**挖掘作業**對不同建模方法進行組合，例如，確定同一模型中性能較高和較低的區段，並根據需要在評分階區段包含或排除每個區段。

區段、規則和條件

模型由區段清單組成，每個區段由選取相符記錄的規則進行定義。給定的規則可以有多個條件，例如：

```
RFM_SCORE  > 10 andMONTHS_CURRENT <= 9
```

規則的清單順序即為套用順序，第一個相符規則將決定給定記錄的輸出結果。如果單獨採用，那麼規則或條件可能會發生重疊，但規則的順序排除了二義性。如果規則不相符，那麼記錄將會分配給其餘規則。

完全控制評分

通過 決策清單檢視器，您可以檢視、修改和重組區段，並且可以出於評分目的來選擇包含或排除哪些區段。例如，您可以選擇在未來報價中排除某群組客戶和包含其他客戶，並且可以立即請參閱這對於整體匹配率的影響。對於已併入的區段和所有其他區段（已併入剩餘區段），決策清單 模型分別傳回分數 *Yes* 和 *$null$*。對評分的這種直接控制使得 決策清單 模型成為產生郵件發送清單的理想工具，而這些模型被廣泛應用於客戶關係管理中，包含呼叫中心或市場應用方面。

採礦作業、測量和選擇

建模過程由**採礦作業**實現。每項採礦作業可以有效地起始一次新的建模，並且會傳回一組新的備選模型。預設作業基於 決策清單 節點的初始規格，您可以定義任意數量的自訂作業。您還可以重複套用作業，例如您可以在整個訓練集合中執行高機率搜尋，然後在剩餘集合中執行低機率搜尋以除去性能較低的區段。

資料選擇

可以定義資料選擇和自訂模型測量以進行模型建置和評估。例如，可以在採礦作業中指定資料選擇以裁剪模型，使之符合具體區域的要求，並且可以建立自訂測量以評估其就整個國家範圍而言的性能優劣。不同於採礦作業的是，測量並不改變底層模型而是以其他視角對其性能進行評估。

新增您的業務知識

通過微調或延伸由演算法識別的區段，決策清單檢視器 使您可以將業務知識併入模型。您可以編輯模型所產生的區段或新增基於指定規則的其他區段。然後可以套用變更並預覽結果。

為了進行深入瞭解，Excel 動態鏈結使您可以將資料匯出到 Excel，這些資料可用於在 Excel 中建立呈現圖表和計算自訂測量（例如綜合利潤和 ROI），您可在建立模型的同時在 決策清單檢視器 中檢視這些自訂測量。

範例。 某金融機構的市場行銷部門希望通過向每個客戶提供合適的報價來在未來的行銷活動中獲取更多利潤。您可以使用決策清單模型來根據以前的促銷活動確定最有可能做出積極回應的客戶所具備的特性，並根據結果產生郵寄清單。

© Copyright IBM Corp. 1994, 2019

需求。 一個表示要預測的二元結果（是/否）且測量層次為旗標或名義的種類目標欄位和至少一個輸入欄位。當目標欄位類型為名義時，必須手動選擇一個值作為**匹配**或**回應**；所有其他值集中在一起作為**不匹配**。還可以指定一個選用的頻率欄位。連續日期/時間欄位將被忽略。使用在建模節點的「專家」標籤上指定的演算法對連續數值範圍的輸入自動分級。為了更好地控制分級，可新增上游分組節點並使用已分級的欄位作為測量層次為序數的輸入。

決策清單模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

眾數。 指定用來建置模型的方法。

- **產生模型。**執行節點時自動在模型選用區上產生模型。可將生成的模型新增至串流中以便評分，但是此模型無法繼續編輯。
- **啟動互動式階段作業。**開啟 決策清單檢視器 互動式建模（輸出）視窗，在此視窗中您可以從多個替代項中進行選擇並重複套用具有不同設定的演算法以逐步生成或修改模型。請參閱第 132 頁的『決策清單檢視器』主題，以取得更多資訊。
- **使用已儲存的互動式階段作業資訊。**使用先前儲存的設定來啟動互動式階段作業。可以使用 決策清單檢視器中的「產生」功能表（用於建立模型或建模節點）或「檔案」功能表（用於更新從中啟動階段作業的節點）儲存交互設定。

目標值。 指定目標欄位的值，以指出您要對其建模的結果。例如，如果目標欄位變換的編碼為：`0 = no` 與 `1 = yes`，則指定 1 以識別指出可能要變換哪些記錄的規則。

尋找具有以下特徵的區段。表示搜尋目標變數是否應該查找出現的**高機率**或**低機率**。尋找和排除這些段可能對於改善您的模型非常有幫助，當剩下的段為低機率段時尤其有用。

上限區段數。指定要傳回的上限區段數。建立頂部的 *N* 個區段，其中最好的區段是機率最高的區段，如果多個模型具有相同的機率，那麼為涵蓋面最高的區段。容許的下限設定是 1；沒有上限設定。

下限區段大小。下方的兩項設定指定了下限區段大小。兩個值中的較大者優先。例如，如果百分比值等於比絕對值高的數字，那麼百分比設定優先。

- **以上一個區段的百分比表示 (%)。**以記錄的百分比指定下限群組大小。容許的下限設定為 0；容許的上限設定為 99.9。
- **以絕對值表示 (N)。**以記錄的絕對數量指定下限群組大小。容許的下限設定是 1；沒有上限設定。

區段規則。

上限屬性數。指定每個區段規則的上限條件數。容許的下限設定是 1；沒有上限設定。

- **容許複用屬性。**如果啟用，那麼每個循環可以使用所有屬性，即使以前的循環已使用過這些屬性。區段的條件是在循環內建立的，每個循環都會增加一個新條件。循環數使用**上限屬性數**設定定義。

新條件的信賴區間 (%)。指定用於測試區段顯著性的信賴等級。此設定在傳回的區段數（如果存在）以及每個區段規則的條件數中具有非常重要的作用。值越高，傳回的結果集越小。容許的下限設定為 50；容許的上限設定為 99.9。

決策清單節點專家選項

通過「專家」選項，您可以對模型建置程序進行微調。

分組方法。 用來對連續欄位（計數相等或寬度相等）執行 Binning 的方法。

分組數目。 針對連續欄位建立的 Bin 數目。容許的下限設定是 2；沒有上限設定。

模型搜尋寬度。 每個週期中可用於下一個週期的模型結果數目上限。容許的下限設定是 1；沒有上限設定。

規則搜尋寬度。 每個週期中可用於下一個週期的規則結果數目上限。容許的下限設定是 1；沒有上限設定。

Bin 合併因素。 當某個區段與其相鄰區段合併時，該區段必須增長的最小數量。容許的下限設定是 1.01；沒有上限設定。

- **在條件中容許遺漏值。** True 表示在規則中容許 IS MISSING 測試。
- **捨棄中間結果。** 若為 True，則只會傳回搜尋程序的最終結果。最終結果是搜尋程序中不會進一步精簡的結果。若為 False，還會傳回中間結果。

  最大替代項目數。 指定執行採礦作業時可以傳回的上限替代項數。容許的下限設定是 1；沒有上限設定。

  注意，採礦作業將只傳回替代值的實際數量，上限為指定的數量上限。例如，如果數量上限設為 100，但只找到 3 個替代值，那麼只顯示這 3 個替代值。

決策清單模型塊

模型包括一個**區段**清單，每個區段都由**規則**進行定義，從而可以選取相符的記錄。在產生模型前可輕鬆檢視或修改這些區段，並選擇包含哪些區段或不包含哪些區段。用於評分時，決策清單模型對於包含的區段傳回是，對於所有其他區段（包括剩餘區段）傳回 *$null$*。對評分的這種直接控制使得決策清單模型成為產生郵件發送清單的理想工具，而這些模型被廣泛應用於客戶關係管理中，包含呼叫中心或市場應用方面。

執行包含決策清單模型的串流時，節點將新增三個新欄位，其中包括分數（對於併入的欄位為 *1*，表示是，對於排除的欄位為 *$null*）、記錄所在區段的機率（命中率）以及區段的 ID 號碼。新欄位的名稱衍生自要預測的輸出欄位的名稱，並帶有表示分數的字首 *$D-*、表示機率的字首 *$DP-* 或表示區段 ID 的字首 *$DI-*。

按照建立模型時指定的目標值對模型進行評分。可以手動去除某些區段以便使它們的分數為 *$null$*。例如，如果執行低機率搜尋以尋找低於平均值匹配率的區段，那麼這些「低匹配率」區段的分數將為是，除非您手動將這些區段排除。如果必要，可以使用衍生節點或充填節點將空值重新撰寫為否。

PMML

使用「第一次命中」選取準則可將決策清單模型儲存為 PMML RuleSetModel。但是，希望所有的規則具有相同的分數。為容許對目標欄位或目標值進行變更，可將多個規則集模型儲存到一個檔案中依順序進行套用，無法與第一個模型符合的案例將傳送到第二個模型，依此類推。演算法名稱 *DecisionList* 用於表示此非標準的行為，且僅具有該名稱的規則集模型可被識別為決策清單模型並如上所述進行評分。

決策清單模型塊設定

通過決策清單模型塊的「設定」標籤，您可以取得傾向分數，還可以啟用或取消 SQL 最佳化。只有將模型塊新增到串流之後，才可以使用此標籤。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **透過轉換為原生 SQL 進行評分** 如果選取此項，則會產生原生 SQL 來在資料庫內對模型進行評分。

  註：雖然這個選項可以更快地提供結果，但隨著模型複雜性的增加，原生 SQL 的大小和複雜性也會增加。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

決策清單檢視器

基於作業的 決策清單檢視器 圖表介面簡單易用，它消除了模型建置過程的複雜性，使您可以擺脫資料採礦技術的低層次詳細資料而將全部精力投入需要使用者參與的分析內容上，如設定目標、選取目標群組、分析結果，以及選取最佳化模型。

工作模型窗格

工作模型窗格將顯示目前模型，包含挖掘作業和適用於該工作模型的其他動作。

ID。 識別循序區段順序。模型區段根據其 ID 號按順序進行計算。

區段規則。提供區段名稱和已定義的區段條件。依預設，區段名稱是欄位名稱或條件中使用的連接欄位名稱（以逗點為分隔字元）。

分數。代表要預測的欄位，假定其值與其他欄位（預測值）的值相關。

註：下列選項可切換為通過第 140 頁的『組織模型測量』對話框顯示。

封面。 該圓餅圖直觀地識別出每個區段的收訊涵蓋範圍與整個收訊涵蓋範圍的對比情況。

範圍 (n)。列出每個區段相對於整個收訊涵蓋範圍的收訊涵蓋範圍量。

次數。 列出接收到的相對於覆蓋範圍的命中數數。例如，如果涉及範圍為 79，頻率為 50，那麼表示在 79 個之中有 50 個對所選區段進行了回應。

機率。指示區段機率。例如，如果涉及範圍為 79，頻率為 50，那麼表示該區段的機率為 63.29%（50 除以 79）。

錯誤。指示區段錯誤。

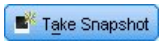
窗格底部的資訊顯示整個模型的涉及範圍、頻率和機率。

工作模型工具列

工作模型窗格的工具列提供了下列功能。

註：也可以通過用滑鼠右鍵按一下模型區段來存取某些功能。

表 9. 工作中模型工具列按鈕

工具列按鈕	說明
	啟動產生新模型對話框，該對話框提供用於建立新模型塊的選項。
	儲存交互階段作業的現行狀態。這會將「決策清單」建模節點更新為目前設定，包括採礦作業、模型 Snapshot資料選擇和自訂測量。要將階段作業還原至此狀態，選中建模節點的「模型」標籤中的**使用儲存的階段作業資訊**勾選框，然後按一下**執行**。
	顯示「組織模型測量」對話框。請參閱第 140 頁的『組織模型測量』主題，以取得更多資訊。
	顯示「組織資料選擇」對話框。請參閱第 137 頁的『組織資料選擇』主題，以取得更多資訊。
	顯示 Snapshot 標籤。請參閱第 134 頁的『Snapshot 標籤』主題，以取得更多資訊。
	顯示「替代」標籤。請參閱第 134 頁的『「替代」標籤』主題，以取得更多資訊。
	獲取目前模型結構的 Snapshot。Snapshot 顯示在 Snapshot 標籤中，一般用於模型比較。
	啟動插入區段對話框，該對話框提供用於建立新模型區段的選項。
	啟動編輯區段規則對話框，該對話框提供的選項可用於將條件新增到模型區段，或變更先前定義的模型區段條件。
	在模型層次中將所選區段上移。
	在模型層次中將所選區段下移。
	刪除所選區段。
	在模型中併入/排除所選區段的情況之間進行切換。排除時，區段結果將計入餘數。不同於刪除區段的是，排除區段允許您選擇重新啟動區段。

「替代」標籤

按一下**尋找區段**產生「替代」標籤，該標籤將針對工作模型窗格中的選定模型或區段列出所有替代挖掘結果。

要將替代模型升級為工作模型，強調顯示所需替代模型並按一下**載入**；則替代模型顯示在工作模型窗格中。

註：只有當您已在決策清單建模節點的「專家」標籤上設定了**上限替代項數**時，才會顯示「替代」標籤以建立多個替代項。

每個產生的模型替代項會顯示特定的模型資訊：

名稱。 每個替代模型都有順序編號。第一個替代項通常包含最佳結果。

目標。 指出目標值。例如：1 相當於 "true"。

區段數目。 替代模型中所使用的區段規則數。

封面。 替代模型的涉及範圍。

頻率。 與覆蓋相關的命中數。

機率。 指明替代模型的機率百分比。

註：替代結果不會隨模型儲存；結果僅在作用中階段作業期間有效。

Snapshot 標籤

Snapshot 是模型在特定復原點的視圖。例如，如果您需要將另一個替代模型載入工作模型窗格、但不希望失去目前模型的相關工作，那麼可以獲取模型 Snapshot。Snapshot 標籤將列出在任意數量的工作模型狀態下手動獲取的所有模型Snapshot。

註：Snapshot 將隨模型儲存。我們建議在您載入首個模型時建立 Snapshot。該 Snapshot 用於儲存原始模型結構，從而確保您可隨時傳回原始模型狀態。產生的 Snapshot 名稱顯示為時間戳記，指示其產生時間。

建立模型 Snapshot

1. 選取要在工作模型窗格中顯示的適當的模型/替代項。
2. 對該工作模型進行必要的變更。
3. 按一下**建立 Snapshot**。此時將在 Snapshot 標籤中顯示一個新Snapshot。

   名稱。 Snapshot 名稱。您可以按兩下 Snapshot 名稱對其進行變更。

   目標。 指出目標值。例如：1 相當於 "true"。

   區段數目。 模型中所使用的區段規則數。

   封面。 模型的涉及範圍。

   頻率。 與覆蓋相關的命中數。

   機率。 指明模型的機率百分比。
4. 要將 Snapshot 升級為工作模型，強調顯示所需 Snapshot 並按一下**載入**；則 Snapshot 模型顯示在工作模型窗格中。
5. 可通過以下方法刪除 Snapshot：按一下**刪除**，或用滑鼠右鍵按一下 Snapshot，然後在功能表中選擇**刪除**。

使用 決策清單檢視器

將以最佳方式預測客戶回應和行為的模型是通過多個階段進行構建的。啟動 決策清單檢視器 時，工作模型將填入已定義的模型區段和測量，並且準備就緒，等待您啟動採礦作業、根據需要修改區段/測量，並產生新的模型或建模節點。

您可新增一個或多個區段規則，直到獲得滿意的模型。可以通過執行採礦作業或使用**編輯區段規則**功能為模型新增區段規則。

在模型建置過程中，您可以對模型的效能進行評估，方法是根據測量資料驗證模型、在圖表中對模型進行可視化處理，或產生自訂 Excel 測量。

肯定模型的品質後，您可以產生新模型並將其置於 IBM SPSS Modeler 畫布或模型選用區中。

採礦作業

採礦作業是確定新規則產生方式的參數的收集。其中某些參數是可以選取的，以便為您提供使模型適應新狀況的靈活性。作業由作業範本（類型）、目標和建立選擇（挖掘資料集）組成。

下列各部分詳細介紹各種採礦工作作業：

- 『執行採礦作業』
- 『建立和編輯採礦作業』
- 第 137 頁的『組織資料選擇』

執行採礦作業： 通過 決策清單檢視器，您可以執行採礦作業或在模型之間複製和貼上區段規則以手動向模型新增區段規則。採礦作業包含有關如何產生新區段規則的資訊（資料採礦參數設定，如搜尋策略、來源屬性、搜尋寬度、信賴等級等）、待預測的客戶行為，以及要調查的資料。採礦作業的目標是搜尋可能的最佳區段規則。

要通過執行採礦作業產生模型區段規則，請執行下列操作：

1. 按一下**剩餘項**列。如果工作模型窗格中已有顯示的區段，您也可以選取其中某一個，根據所選區段尋找其他規則。選取剩餘項或區段之後，可採用下列方法之一產生模型或替代模型：

- 從「工具」功能表選擇**尋找區段**。
- 用滑鼠右鍵按一下**剩餘項**列/區段，然後選擇**尋找區段**。
- 按一下工作模型窗格上的**尋找區段**按鈕。

在作業正在處理過程中，進度將在畫布底部顯示，並在作業完成時提示您。作業完成所用的時間完全取決於採礦作業的複雜性以及資料集的大小。如果結果中只有一個模型，那麼作業完成後它將立即顯示在工作模型窗格上；但是，如果結果包含多個模型，那麼模型顯示在「替代」標籤上。

註：作業結果將為：完成並更新模型、完成但不更新模型或失敗。

可以重複尋找新區段規則的過程，直到不再有新規則新增到模型中。這表示已找到所有有意義的客戶群組。

可以對任何現有的模型區段執行採礦作業。如果對作業的結果不滿意，您可以選擇對同一模型區段啟動另一個採礦作業。此操作將基於所選區段提供找到的其他規則。位於所選區段「下方」的區段（即，在所選區段之後新增到模型的區段）將被新區段替代，因為每個區段都取決於其前提條件。

建立和編輯採礦作業： 採礦作業是搜尋組成資料模型的規則收集的機制。除所選範本中定義的搜尋準則外，作業還會定義目標（激發分析的實際問題，如有多少客戶可能對郵件做出回應），並識別要使用的資料集。採礦作業的目標是搜尋可能的最佳模型。

建立採礦作業

要建立採礦作業，請執行下列操作：

1. 選取要在其中挖掘其他區段條件的區段。
2. 按一下**設定**。此時將開啟「建立/編輯挖掘作業」對話框。該對話框提供用於定義挖掘作業的選項。
3. 進行必要的變更並按一下**確定**傳回到工作模型窗格。決策清單檢視器 使用這些設定作為預設值以針對每個作業執行，直到選取了替代作業或設定。
4. 按一下**尋找區段**以啟動選定區段上的挖掘作業。

編輯採礦作業

「建立/編輯採礦作業」對話框提供的選項可用於定義新的採礦作業或編輯現有採礦作業。

可用於採礦作業的大部分參數與決策清單節點中提供的參數類似。例外在下方顯示。請參閱第 130 頁的『決策清單模型選項』主題，以取得更多資訊。

載入設定：建立多個挖掘作業後，請選取所需作業。

新建... 按一下以根據目前顯示的作業的設定建立新採礦作業。

目標

目標欄位：代表要預測的欄位，假定其值與其他欄位（預測值）的值相關。

目標值。 指定目標欄位的值，以指出您要對其建模的結果。例如，如果目標欄位變換的編碼為：0 = no 與 1 = yes，則指定 1 以識別指出可能要變換哪些記錄的規則。

簡式設定

最大替代項目數。 指定執行採礦作業後將顯示的替代項數。容許的下限設定是 1；沒有上限設定。

專家設定

編輯... 用於開啟**編輯進階參數**對話框，您可在其中定義進階設定。請參閱『編輯進階參數』主題，以取得更多資訊。

資料

建立選擇。提供的選項用於指定 決策清單檢視器 應對其進行分析以尋找新規則的評估測量方式。列出的測量在「組織資料選擇」對話框中進行建立/編輯。

可用的欄位。 提供用於顯示所有欄位或手動選取要顯示的欄位的選項。

編輯... 如果選取了**自訂**選項，那麼這將開啟**自訂可用欄位**對話框，您可以在其中選取可用作通過採礦作業找到的區段屬性的欄位。請參閱第 137 頁的『自訂可用欄位』主題，以取得更多資訊。

編輯進階參數： 「編輯進階參數」對話框提供下列配置選項。

分組方法。 用來對連續欄位（計數相等或寬度相等）執行 Binning 的方法。

分組數目。 針對連續欄位建立的 Bin 數目。容許的下限設定是 2；沒有上限設定。

模型搜尋寬度。 每個週期中可用於下一個週期的模型結果數目上限。容許的下限設定是 1；沒有上限設定。

規則搜尋寬度。 每個週期中可用於下一個週期的規則結果數目上限。容許的下限設定是 1；沒有上限設定。

Bin 合併因素。 當某個區段與其相鄰區段合併時，該區段必須增長的最小數量。容許的下限設定是 1.01；沒有上限設定。

- **在條件中容許遺漏值。** True 表示在規則中容許 IS MISSING 測試。
- **捨棄中間結果。** 若為 True，則只會傳回搜尋程序的最終結果。最終結果是搜尋程序中不會進一步精簡的結果。若為 False，還會傳回中間結果。

自訂可用欄位： 通過「自訂可用欄位」對話框，您可以選取可用作通過採礦作業找到的區段屬性的欄位。

可用。列出目前可用作區段屬性的欄位。要從清單中刪除欄位，請選取適當的欄位，然後按一下 **刪除 >>**。此時所選欄位將從「可用」清單移至「無法使用」清單。

無法使用。列出無法使用作區段屬性的欄位。要將欄位包含在「可用」清單中，請選取適當的欄位，然後按一下**<< 新增**。此時所選欄位將從「無法使用」清單移至「可用」清單。

組織資料選擇： 通過組織資料選擇（挖掘資料集），可以指定 決策清單檢視器應對哪些測量進行分析以尋找新規則，並選擇要用作尺度基準的資料選擇。

要組織資料選擇，請執行下列操作：

1. 從「工具」功能表中選擇**組織資料選擇**，或用滑鼠右鍵按一下某個區段並選擇該選項。此時將開啟「組織資料選擇」對話框。

   註：通過「組織資料選擇」對話框，您也可以編輯或刪除現有資料選擇。
2. 按一下**新增新的資料選擇**按鈕。此時會將一個新的資料選擇登錄新增到現有的表格中。
3. 按一下**名稱**並輸入適當的選擇名稱。
4. 按一下**分割區**並選取適當的分割區類型。
5. 按一下**條件**並選取適當的條件選項。如果選擇**指定**，那麼會開啟「指定選擇條件」對話框，其中包含定義特定欄位條件的選項。
6. 定義適當的條件，然後按一下**確定**。

通過「建立/編輯採礦作業」對話框中的「建立選擇」下拉清單可存取這些資料選擇。通過該清單，您可以選取用於特定採礦作業的評估測量方式。

叢集規則

通過執行基於作業範本的採礦作業，可以尋找模型區段規則。您可以使用「插入區段」或「編輯區段規則」功能手動為模型新增區段規則。

如果選擇挖掘新的區段規則，結果（如果有）將在「互動式清單」對話框的「檢視器」標籤中顯示。通過從「模型相簿」對話框中選取替代結果，並按一下**載入**，可以快速精練您的模型。這樣，您可以嘗試不同結果，直到準備好建立出準確說明最佳目標群組的模型。

插入區段： 您可以使用「插入區段」功能手動為模型新增區段規則。

要將區段規則條件新增到模型，請執行下列操作：

1. 在互動式清單對話框中，選取您要新增新區段的位置。新區段將直接插在所選區段的上方。
2. 在「編輯」功能表中，選擇**插入區段**或通過用滑鼠右鍵按一下區段存取此選項。

   這將開啟「插入區段」對話框，您可以在其中插入新的區段規則條件。

3. 按一下**插入**。這將開啟「插入條件」對話框，您可以在其中定義新規則條件的屬性。
4. 從下拉清單中選取欄位和運算子。

   註：如果選取 **Not in** 運算子，那麼所選條件將用作排除條件，並且在「插入規則」對話框中顯示為紅色。例如，當條件 `region = 'TOWN'` 顯示為紅色時，表示將 TOWN 從結果集中排除。
5. 輸入一個或多個值，或者按一下**插入值**圖示，以顯示「插入值」對話框。該對話框可讓您選擇定義給所選欄位的值。例如，欄位 **married** 將提供值 **yes** 和 **no**。
6. 按一下**確定**傳回「插入區段」對話框。再次按一下**確定**將所建立的區段新增到模型中。

此時該新區段將顯示在指定的模型位置。

編輯區段規則： 通過「編輯區段規則」功能，您可以新增、變更或刪除區段規則條件。

要變更區段規則條件，請執行下列操作：

1. 選取要編輯的模型區段。
2. 從「編輯」功能表選擇**編輯區段規則**，或用滑鼠右鍵按一下規則以存取此選項。

   此時將開啟「編輯區段規則」對話框。
3. 選取適當的條件，然後按一下**編輯**。

   這將開啟「編輯條件」對話框，您可以在其中定義所選規則條件的屬性。
4. 從下拉清單中選取欄位和運算子。

   註：如果選取 **Not in** 運算子，那麼所選條件將用作排除條件，並且在「編輯區段規則」對話框中顯示為紅色。例如，當條件 `region = 'TOWN'` 顯示為紅色時，表示將 TOWN 從結果集中排除。
5. 輸入一個或多個值，或按一下**插入值**按鈕以顯示「插入值」對話框。該對話框可讓您選擇定義給所選欄位的值。例如，欄位 **married** 將提供值 **yes** 和 **no**。
6. 按一下**確定**傳回到「編輯區段規則」對話框。再次按一下**確定**傳回工作模型。

此時所選取的區段將與更新的規則條件一起顯示。

刪除區段規則條件： 要刪除區段規則條件，請執行下列操作：

1. 選取包含要刪除的規則條件的模型區段。
2. 從「編輯」功能表中選擇**編輯區段規則**，或用滑鼠右鍵按一下區段以存取此選項。

   這將開啟「編輯區段規則」對話框，您可在其中刪除一項或多項區段規則條件。
3. 選取適當的規則條件，然後按一下**刪除**。
4. 按一下**確定**。

刪除一個或多個區段規則條件將使工作模型窗格重新整理其測量的度量值。

複製區段： 決策清單檢視器 為您提供了一種複製模型區段的簡便方法。如果要將一個模型中的區段套用於另一個模型時，只需將該區段從一個模型複製（或剪下）並貼上到另一個模型中即可。此外，您還可以從「替代預覽」窗格中顯示的模型複製區段並將其貼上到工作模型窗格中顯示的模型中。這些剪下、複製和貼上功能使用系統剪貼簿儲存或擷取暫時資料。這意味著將在剪貼簿中複製條件和目標。剪貼簿內容不僅僅保留用於 決策清單檢視器，也可以貼上在其他應用程式中。例如，在文字編輯器中貼上剪貼簿內容時，會以 XML 格式貼上條件和目標。

要複製或剪下模型區段，請執行下列操作：

1. 選取要在其他模型中使用的模型區段。
2. 從「編輯」功能表中選擇**複製**（或**剪下**），或用滑鼠右鍵按一下模型區段並選擇**複製**或**剪下**。
3. 開啟適當的模型（將在其中貼上模型區段的模型）。
4. 選取某個模型區段，然後按一下**貼上**。

註：您也可以使用以下組合鍵，來代替**剪下**、**複製**和**貼上**指令：**Ctrl+X**、**Ctrl+C** 和 **Ctrl+V**。

複製（剪下）的區段將插入先前選取的模型區段上方。下方貼上的一或多個區段的測量將重新計算。

註：此程序中的兩個模型必須基於同一基礎模型範本，並包含同一目標，否則將顯示錯誤訊息。

替代模型： 當有多個結果時，「替代」標籤顯示每個採礦作業的結果。每個結果包含所選資料中與目標最接近符合的條件，以及所有「相當符合」的替代項。顯示的替代項總數取決於分析過程中採用的搜尋準則。

要檢視替代模型，請執行下列操作：

1. 按一下「替代」標籤上的替代模型。在「替代預覽」窗格中，替代模型區段顯示或替代目前模型區段。
2. 要在工作模型窗格中使用替代模型，在「替代預覽」窗格中選取模式並按一下**載入**，或在「替代」標籤上用滑鼠右鍵按一下替代模型名稱並選擇**載入**。

註：產生新模型時，不會儲存替代模型。

自訂模型

資料不是靜態的。客戶會遷移、結婚和更換工作。產品會隨之失去市場對焦點並作廢。

決策清單檢視器 為業務使用者提供了使模型方便迅速地適應新狀況的靈活性。您可通過編輯、設定優先順序、刪除或停用特定模型區段來變更模型。

為區段設定優先順序： 您可選擇任意順序，對模型規則進行排列。依預設，模型區段按優先順序顯示，第一個區段具有最高優先順序。當您為一個或多個區段指定不同的優先順序時，模型會發生相應的變更。您可以根據需要通過將區段移至較高或較低的優先順序位置來更改模型。

要為模型區段設定優先順序，請執行下列操作：

1. 選取要為其指定不同優先順序的模型區段。
2. 按一下工作模型窗格工具列中的兩個箭頭按鈕之一，將所選模型區段在清單中上移或下移。

設定優先順序後，會重新計算先前的所有評量結果，並顯示新值。

刪除區段： 要刪除一個或多個區段，請執行下列操作：

1. 選取模型區段。
2. 從「編輯」功能表中選擇**刪除區段**，或在工作模型窗格的工具列中按一下刪除按鈕。

測量將針對修改後的模型重新計算，模型也會發生相應的變更。

排除區段： 在搜尋特定群組時，您可能會將一部分模型區段作為商業動作的基礎。部署模型時，您可能會選擇排除模型中的某些區段。排除的區段作為空值進行評分。排除某個區段並不代表不使用該區段，而是從郵件清單中排除與該規則相符的所有記錄。該規則仍在套用，但方式不同。

要排除特定的模型區段，請執行下列操作：

1. 在工作模型窗格中選取一個區段。
2. 在工作模型窗格的工具列中按一下**切換區段排除**按鈕。此時將在所選區段的所選「目標」欄中顯示**排除**。

註：與刪除的區段不同，排除的區段在最終模型中仍然可供重複使用。排除的區段仍將影響圖表結果。

變更目標值： 通過「變更目標值」對話框，您可以變更目前目標欄位的目標值。

與工作模型具有不同目標值的 Snapshot 和階段作業結果會通過將該列的表格背景變為黃色進行 ID。這表示該 Snapshot/階段作業結果已過時。

建立/編輯採礦作業對話框將顯示目前工作模型的目標值。該目標值不會隨採礦作業儲存，而是取自工作模型的值。

當您將某個與目前工作模型具有不同目標值的已儲存模型升級為工作模型（例如，通過編輯替代結果或編輯 Snapshot 副本）時，已儲存模型的目標值將變更為工作模型的目標值（工作模型窗格中顯示的目標值不會變更）。模型度量值將使用新目標重新評估。

產生新的模式

「產生新模型」對話框提供的選項可用於命名模型並選取建立新節點的位置。

模型名稱。選取**自訂**可調整自動產生的名稱，或建立在串流畫布中顯示的唯一節點名稱。

建立節點位置。 選取**畫布**會將新模型置於畫布中；選取 **GM 選用區**會將新模型置於「模型」選用區中；選取**兩者**會將新模型同時置於畫布和「模型」選用區中。

包含互動式階段作業狀態。啟用此選項後，互動式階段作業狀態將保留在產生的模型中。稍後從模型產生建模節點時，該狀態將繼續傳遞並用於起始設定交互階段作業。無論是否選取此選項，模型本身對新資料的分數方式都是相同的。如果未選取此選項，模型仍然可以建立建置節點，但該節點將更為一般化，它會啟動新的交互階段作業而不是從原有階段作業停止的位置繼續前進。如果變更節點設定但以儲存的某種狀態執行，那麼會忽略已變更的設定以採用儲存狀態的設定。

註：標準度量值是的模型的唯一度量值。其他度量值將保留在交互狀態。產生的模型不會顯示已儲存的交互採礦作業狀態。啟動 決策清單檢視器時，它會顯示通過檢視器所做的初始設定。

請參閱第 41 頁的『重新產生建模節點』主題，以取得更多資訊。

模型評量

成功的建模需要在正式作業環境中執行實作之前進行謹慎的模型評量。決策清單檢視器提供了可用於評量模型實際應用效果的多種統計測量和商業測量。其中包含收益圖表和與 Excel 的全面互操作，從而實現成本/增益案例的模擬，以便評估部署的作用。

您可採用下列方式評估自己的模型：

- 使用 決策清單檢視器 中提供的預先定義的統計測量和商業模型測量（機率、頻率）。
- 評估從 Microsoft Excel 中匯入的測量。
- 使用收益圖表對模型進行視覺化處理。

組織模型測量： 決策清單檢視器 提供了用於定義按欄計算並顯示的測量的選項。每個區段可包含預設的涉及範圍、頻率、機率和錯誤等測量，按欄顯示。此外，您也可以建立將按欄顯示的新測量。

定義模型測量

要為模型新增測量或定義現有的測量，請執行下列操作：

1. 從「工具」功能表中選擇**組織模型測量**，或用滑鼠右鍵按一下模型以選擇此選項。此時將開啟「組織模型測量」對話框。

2. 按一下**新增新的模型測量**按鈕（位於「顯示」欄右側）。此時將在表格中顯示一個新的測量。
3. 提供測量名稱，並選擇適當的類型、顯示選項和選擇。「顯示」欄指示是否為工作模型顯示測量。定義現有測量時，請選擇適當的測量和選擇，並指定該測量是否將在工作模型中顯示。
4. 按一下**確定**傳回決策清單檢視器畫布。如果已勾選新測量的「顯示」欄，那麼會為工作模型顯示該新測量。

Excel 中的自訂度量值

請參閱『Excel 中的評量』主題，以取得更多資訊。

重新整理測量： 在某些特定情況下，可能需要重新計算模型測量，例如對一組新客戶套用現有模型時。

要重新計算（重新整理）模型測量，請執行下列操作：

在「編輯」功能表中選擇**重新整理所有測量**。

或

按 F5。

此時將重新計算所有測量，並針對工作模型顯示新值。

Excel 中的評量： 決策清單檢視器 可以與 Microsoft Excel 進行整合，使您可以在模型建置過程中直接使用自己的值計算和利潤公式來模擬成本/收益案例。通過與 Excel 的鏈結，您可以將資料匯出至 Excel（資料在其中可用於建立呈現圖表）、計算自訂測量（如複合利潤和 ROI 測量），並且可以在建立模型時通過 決策清單檢視器 檢視這些測量。

註：要使用 Excel 試算表，CRM 分析專家必須針對 決策清單檢視器 與 Microsoft Excel 的同步化定義配置資訊。該配置包含於 Excel 試算表檔案中，用於指明 決策清單檢視器 與 Excel 之間相互傳輸的資訊。

下列步驟僅在已安裝 MS Excel 的情況下有效。如果未安裝 Excel，那麼不會顯示使模型與 Excel 同步的選項。

要使模型與 MS Excel 同步，請執行下列操作：

1. 開啟模型，執行交互階段作業，並從「工具」功能表中選擇**組織模型測量**。
2. 為**計算 Excel 中的自訂測量**選項選取**是**。這將啟動**活頁簿**欄位，使您可以選取預先配置的 Excel 活頁簿範本。
3. 按一下**連接至 Excel** 按鈕。這將開啟「開啟」對話框，使您可以導覽至本端或網路檔案系統中預先配置的範本所在的位置。
4. 選取適當的 Excel 範本，然後按一下**開啟**。此時將啟動所選的 Excel 範本；使用 Windows 工作列（或按 Alt+Tab）返回到「選擇自訂測量的輸入」對話框。
5. 在 Excel 範本中定義的度量值名稱與模型度量值名稱之間選取適當的對映，然後按一下**確定**。

建立鏈結後，Excel 將立即採用預先配置的 Excel 範本啟動，該範本以試算表顯示模型規則。Excel 中的計算結果在 決策清單檢視器中顯示為新欄。

註：儲存模型時，不會保留 Excel 度量值；度量值僅在作用中階段作業期間有效。但是，您可以建立包含 Excel 度量值的Snapshot。在 Snapshot 視圖中儲存的 Excel 度量值僅適用於歷程記錄比較，在重新開啟時不會重新整理。請參閱第 134 頁的『Snapshot 標籤』主題，以取得更多資訊。重新建立與 Excel 範本的連線前，Excel 度量值將不會顯示在 Snapshot 中。

***MS Excel* 整合設定：** 決策清單檢視器 與 Microsoft Excel 的整合是通過使用預先配置的 Excel 試算表範本實現的。該範本由以下三個工作表組成：

模型測量。顯示匯入的決策清單檢視器測量、自訂 Excel 測量，以及計算總計（在「設定」工作表中定義）。

設定。提供用於根據已匯入的 決策清單檢視器 測量和自訂 Excel 測量產生計算的變數。

配置。 提供用於指定從 決策清單檢視器 匯入哪些測量以及用於定義自訂 Excel 測量的選項。

警告：「配置」工作表的結構已嚴格定義。請**勿**編輯綠色陰影區域中的任何蜂巢 Cell。

- **來自模型的度量值。**指示在計算中使用哪些 決策清單檢視器 度量值。
- **要返回到模型的度量值。**指示 Excel 產生的哪些度量值將傳回到 決策清單檢視器。Excel 產生的測量在 決策清單檢視器中顯示為新的測量欄。

註：產生新模型時，Excel 度量值不會隨模型一起保留；這些度量值僅在作用中階段作業期間有效。

變更模型測量： 下列範例演示如何通過多種方法變更模型測量：

- 變更現有測量。
- 從模型匯入其他標準測量。
- 將其他自訂測量匯出到模型。

變更現有測量

1. 開啟範本並選取「配置」工作表。
2. 通過強調顯示並重寫名稱或說明來編輯任何**名稱**或**說明**。

   請注意，如果要變更測量（例如，為了提示使用者機率而非頻率），只需變更**來自模型的測量**中的名稱和說明，該名稱和說明隨後將顯示在模型中並且使用者可以選擇要對映的恰當測量。

從模型匯入其他標準測量

1. 開啟範本並選取「配置」工作表。
2. 在功能表上，選擇：

   工具 > **保護** > **撤消工作表保護**
3. 選取 A5 Cell，該 Cell 有黃色陰影且包含**結束**單字。
4. 在功能表上，選擇：

   插入 > **列**
5. 在新測量的**名稱**和**說明**中輸入值。例如，**錯誤**和**區段的相關錯誤**。
6. 在 C5 Cell 中輸入公式 **=COLUMN('Model Measures'!N3)**。
7. 在 D5 Cell 中輸入公式 **=ROW('Model Measures'!N3)+1**。

   這些公式會使新的測量顯示在模型測量工作表的 N 欄中，此欄目前為空白。
8. 在功能表上，選擇：

   工具 > **保護** > **保護工作表**
9. 按一下**確定**。
10. 在模型測量工作表中，確保 N3 Cell 已將**錯誤**作為新欄的標題。
11. 選取整個 N 欄。

12. 在功能表上，選擇：

    格式 > 單元

13. 依預設，所有蜂巢 Cell 均有一個**一般**數字種類。按一下**百分比**可變更數字顯示的方式。這可協助您檢查 Excel 中的數字；此外，也可讓您能夠以其他方式利用資料，例如作為圖形的輸出。
14. 按一下**確定**。
15. 將試算表儲存成 Excel 2003 範本，使用副檔名為 *.xlt* 的唯一名稱。為了輕鬆找到新範本，建議您將它儲存在您本端或網路檔案系統上的預先配置範本位置。

將其他自訂測量匯出到模型

1. 開啟之前範例中已新增「錯誤」欄的範本；選取「配置」工作表。
2. 在功能表上，選擇：

    工具 > 保護 > 撤消工作表保護

3. 選取 Cell A14，該 Cell 有黃色陰影且包含**結束**單字。
4. 在功能表上，選擇：

    插入 > 列

5. 在新測量的**名稱**和**說明**中輸入值。例如，**調整比例錯誤**和**套用至 Excel 錯誤的調整比例**。
6. 在 Cell C14 中輸入公式 **=COLUMN('Model Measures'!O3)**。
7. 在 Cell D14 中輸入公式 **=ROW('Model Measures'!O3)+1**。

    這些公式指定 O 欄將提供模型的新測量。

8. 選取「設定」工作表。
9. 在 Cell A17 中輸入說明 **'- 定比變化錯誤**。
10. 在 Cell B17 中輸入調整比例係數 **10**。
11. 在「模型測量」工作表中，在 O3 Cell 中輸入說明**調整比例錯誤**作為新欄的標題。
12. 在 Cell O4 中輸入公式 **=N4*Settings!B17**。
13. 選取 Cell O4 的右下角落並將其向下拖曳到 Cell O22，以將公式複製到每一個 Cell 中。
14. 在功能表上，選擇：

    工具 > 保護 > 保護工作表

15. 按一下**確定**。
16. 將試算表儲存成 Excel 2003 範本，使用副檔名為 *.xlt* 的唯一名稱。為了輕鬆找到新範本，建議您將它儲存在您本端或網路檔案系統上的預先配置範本位置。

當使用該範本連接 Excel 時，「錯誤」值可用作新的自訂測量。

視覺化模式

瞭解模型作用的最佳方式是對其進行視覺化處理。使用收益圖表，您可以透過即時研究多個替代模型的效果，獲得對模型商業和技術效益的有價值日常見解。『增益圖表』部分顯示了某個模型在隨機決策過程中的增益，並且可以在存在替代模型時實現對多個圖表的直接比較。

增益圖表： 收益圖表繪制的是表格增益 % 直欄中的值。增益定義為每個增量中的命中數相對於樹狀結構中總命中數的比例，使用的方程式如下：

(增量中的命中數 / 總命中數) x 100%

收益圖表有效地為您說明需要怎樣的撒網廣度才能擷取樹狀結構中所有命中數的給定百分比。斜行繪製整個樣本在未使用模型的情況下的預期回應。在此情況下，回應率將不變，因為某個人員的回應可能與另一個人員一樣。若要將收益翻倍，您需要詢問兩倍的人員。曲線指出在您僅併入根據收益排名在較高百分位數的人員時，您的回應可以提升的幅度。例如，併入前 50% 可能讓您得到 70% 以上的正向回應。曲線越陡峭，增益越高。

要檢視收益圖表，請執行下列操作：

1. 開啟包含決策清單節點的串流，並從該節點啟動一個交互階段作業。
2. 按一下**增益**標籤。根據指定的分割區，您會看到一個或兩個圖表（例如，如果同時為模型測量定義了訓練分割區和測試分割區，那麼會顯示兩個圖表）。

依預設，圖表會顯示為區段。您可以將圖表切換為分位數顯示，方法是選取**分位數**，然後在下拉功能表中選取適當的分位數方法。

圖表選項： 「圖表選項」功能提供的選項可用於選取以圖表顯示哪些模型和 Snapshot繪製哪些分割區，以及是否顯示區段標籤。

要繪製的模型

目前模型。使您可以選取要繪製的模型。您可以選取工作模型或任何已建立的 Snapshot 模型。

要繪製的分割區

左側圖表的分割區。該下拉清單提供用於顯示所有已定義分割區或所有資料的選項。

右側圖表的分割區。該下拉清單提供用於顯示所有已定義分割區、所有資料或僅顯示左側圖表的選項。如果選取**只繪製左側圖**，那麼僅顯示左側圖表。

顯示區段標籤。啟用此選項後，將在圖表中顯示全部區段標籤。

第 10 章 統計模型

統計模型使用數學方程式對擷取自資料的資訊進行編碼。在某些情況下，統計建模技術可以非常快速地給出合適的模型。甚至對於那些只有更加靈活的機器學習技術（例如神經網路）才能最終給出更好結果的問題，仍然可以將某些統計模型用作基準線預測模型以判斷更先進技術的效能。

提供了下列統計建模節點。

線性迴歸模型基於目標和一個或多個預測值之間的線性關係預測連續目標。

邏輯迴歸是一種統計技術，它可根據輸入欄位的值對記錄進行分類。它類似於線性迴歸方法，但採用的是種類目標欄位而非數值範圍。

PCA/因素節點提供強大的資料減少技術來減少資料的複雜性。主成份分析（PCA）可找出輸入欄位的線性組合，該組合最好地擷取了整個欄位集合中的變異數，且組合中的各個成分相互正交（相互垂直）。因數分析則試圖識別底層因素，這些因素說明觀測的欄位集合內的相關性型樣。對於這兩種方法，其共同的目標是找到可對原始欄位集合中的資訊進行有效總結的少量衍生欄位。

判別分析所做的假設比邏輯迴歸方法的假設更嚴格，但在符合這些假設時，判別分析可以作為邏輯迴歸方法分析的有用替代項或補充。

「通用性線性」模型對一般線性模型進行了擴展，這樣依變數通過指定的鏈結函數與因子和共變數線性相關。此外，此模式允許」變數具有非常態分配。它包括統計模型大部分的功能，其中包括線性迴歸、邏輯迴歸方法、用於計數資料的對數線性模型以及區間刪失生存分析模型。

概化線性混合模型 (GLMM) 延伸了線性模型，使得目標可以有非常態分佈，通過指定的連接函數與因子和共變數線性相關，並且觀察可能相關。通用性線性混合模型涵蓋多種模式，從非常態縱向資料的簡單線性迴歸，到複雜的多層級模式。

使用 Cox 迴歸節點，您可以在已有的檢查記錄中建立時間事件的生存分析模型。該模型會生成一個生存分析函數，該函數可預測在給定時間 (*t*) 內對於所給定的輸入變數值相關事件的發生機率。

© Copyright IBM Corp. 1994, 2019

線性節點

線性回歸是一個一般統計技術，用來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。

需求。 在線性迴歸模型中只能使用數值型欄位。您必須只有一個目標欄位（角色設為**目標**）以及一個以上預測值（角色設為**輸入**）。角色為**兩者**或**無**的欄位會予以忽略，因為它們是非數值欄位。（必要的話，可使用「衍生」節點對非數值欄位重新編碼。）

強度。 線性迴歸模型相對簡單，提供易於解譯的數學公式來產生預測內容。由於線性迴歸是一種由來已久的統計方法，因此這些模型的內容已廣為人所熟知。通常，線性模型的訓練速度也非常快。「線性」節點提供了自動欄位選擇方法，以排除方程式中不重要的輸入欄位。

註：如果目標欄位為種類（例如**是/否**或**流失/未流失**）而非連續範圍，那麼可以將邏輯迴歸用作替代項。 邏輯迴歸也提供對非數值輸入的支援，不需要對這些欄位重新編碼。請參閱第 154 頁的『Logistic 節點』主題，以取得更多資訊。

線性模型

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。

線性模型相當簡便，且提供了簡易的評分數學公式。與同一個資料集上的其他模式類型 (如神經網路或決策樹) 相較，這些模型的內容比較容易瞭解，通常可以快速建立。

範例。 某資源有限的保險公司，打算調查屋主的保險理賠，希望建立估計理賠成本的模式。透過在服務中心部署這個模式，代表就能在和客戶通電話的同時輸入資訊，並根據過去的資料立即取得「預期的」理賠成本。

欄位需求。必須具有「目標」以及至少一個的「輸入」。依預設，不會使用預先定義角色為「兩者」或「無」的欄位。目標必須是連續的 (尺度)。對預測值（輸入）沒有測量層次限制。種類（旗標、列名和序數）欄位用作模型中的因子，同時連續欄位用作共變數。

目標

您現在要進行什麼作業？

- **建立新模式。** 建立全新的模式。此為一般的節點作業。
- **繼續執行現有模型的訓練。** 系統會繼續使用節點上次成功產生的模式來執行訓練。這可讓您更新或重新整理現有模式，而無需存取原始資料，此外由於資料流中僅會容納全新或已更新的記錄，因此會使效能速度顯著提升。系統會使用建模節點來儲存上一個模型的詳細資料，即使資料流或「模型」色板中已不再提供上個模型塊，也能使用此選項。

  註：若啟用此選項，則系統會停用「欄位」和「建置選項」標籤上的其他所有控制項。

您的主要目標是什麼？ 選取適當的目標。

- **建立標準模式。** 此方法會建立單一模式，以預測使用預測值的目標。一般而言，標準模式在解讀上較為容易，且評分速度比 Boosted、Bagged 或大型資料集集合更快。

  註：對於分割模型，若要搭配使用此選項與**繼續訓練現有模型**，您必須連接到 Analytic Server。
- **強化模式準確性 (Boosting)。** 此方法會使用 Boosting 來建立集合模式，其會產生一系列的模式，以取得更為準確的預測。集合的建立和評分時間均長於標準模式。

Boosting 會產生一系列的「元件模型」，其中每一個都建置於整個資料集。在建置每一個系列元件模型之前，會根據前一個元件模型的殘差加權記錄。對於具有較大殘差的觀察值一般會給定較高的分析加權，因此下一個元件模型將更加注重於預測這些記錄。與這些元件模型一起組成複合模型。複合模型會使用結合規則對新記錄進行評分；可用的規則取決於目標的測量水準。

- **強化模式穩定性 (Bagging)。** 此方法會使用 Bagging (重複取樣整合) 來建立集合模式，其會產生多個模式，以取得更為可靠的預測結果。集合的建立和評分時間均長於標準模式。

  引導聚集 (bagging) 會透過取樣產生訓練資料集的複本，以替換原始資料。這樣做可建立與原始資料集大小相同的引導樣本。然後，在每一個複本上建置「元件模型」。與這些元件模型一起組成複合模型。複合模型會使用結合規則對新記錄進行評分；可用的規則取決於目標的測量水準。
- **為大型資料集建立模型。**此方法會將資料集分割為個別的資料區塊，以建立集合模式。若資料集過大而無法建置上述任何模型，或是想要建立增量模型，請選擇此選項。此選項的建立時間較短，但評分時間會長於標準模式。

請參閱第 148 頁的『集合』，以取得 boosting、bagging 和極大資料集的相關設定。

基本

自動準備資料。此選項可讓程序在內部轉換目標和預測值，以發揮最高的模型預測能力；所有轉換都會和模式一起儲存，並套用至新資料以進行評分。系統會將已轉換欄位的原始版本自模式中排除。依預設，系統會執行下列的自動資料準備作業。

- **日期與時間處理。**每個日期預測值皆會轉換至新的連續預測值，其中內含參考日期 (1970-01-01) 之後的經過時間。每個時間預測值皆會轉換為新的連續預測值，其中包含參考時間 (00:00:00) 之後的經過時間。
- **調整測量層級。**系統會將不同值數目小於 5 的連續預測值重新分配為次序預測值。系統會將不同值數目大於 10 的次序預測值重新分配為連續預測值。
- **偏離值處理。**系統會將超出分割值 (與平均值相距 3 個標準差) 的連續預測值設為分割值。
- **遺漏值處理。**系統會以訓練區隔的眾數來取代名義預測變數的遺漏值。系統會以訓練區隔的中位數來取代次序預測值的遺漏值。系統會以訓練區隔的平均數來取代連續預測值的遺漏值。
- **受監督合併。**此項目會透過減少要處理的目標相關欄位數目，建立較為精簡的模式。相同的類別是根據輸入和目標之間的關係來識別。系統會合併沒有顯著差異 (即 p 值大於 0.1) 的類別。若將所有類別合而為一，則會從模式中排除欄位的原始和衍生版本，這是因為它們沒有可作為預測值的數值。

信賴等級。 此信賴等級用於在係數檢視中計算模型係數的區間估計值。指定一個大於 0 且小於 100 的值。預設值是 95。

模型選擇

模型選擇方法。選擇其中一種模式選擇方法 (詳細資訊如下) 或「**包含所有預測值**」，以僅將所有可用的預測值輸入為主效應模式項目。依預設，系統會使用「**向前逐步**」。

轉遞逐步選擇。此方法一啟動並不會對模式產生任何效果，但會根據逐步條件一步一步新增或移除效果，直至無法再新增或移除任何效果為止。

- **進入/移除準則。**此統計量決定是否應在模式中新增或移除效果。**資訊準則 (AICC)** 是以指定模式訓練集的概似為基礎，且會經過調整以懲罰過於複雜的模式。**F 統計量**是以模式錯誤改善的統計測試為基礎。**已調整 R 平方**是以訓練集的配適度為基礎，且會經過調整以懲罰過於複雜的模式。**過適預防準則 (ASE)** 過適預防準則 (平均平方誤，或 ASE) 為基礎。過適預防集是原始資料集 30% 左右的隨機子樣本，不用來訓練模式。

  若選擇任何非「**F 統計量**」的條件，則系統會將每個步驟中對應至最大正向增加條件的效果新增至模式。系統會移除模式中任何對應至減少條件的效果。

若選擇「**F 統計量**」作為條件，則系統會在每個步驟中，將最小 *p* 值小於指定臨界值（「**包含 p 值小於此值的效果**」）的效果新增至模式。預設值為 0.05。系統會移除任何 *p* 值大於指定臨界值（「**移除 p 值大於此值的效果**」）之模式中的效果。預設值是 0.10。

- **自訂最終模型中的效果最大數目。**依預設，系統會將所有可用的效果輸入至模式。或者，若逐步演算法以指定的最大效果數目來結束步驟，則演算法在停止時會保有目前的效果集。
- **自訂步驟的最大數目。** 逐步演算法在經過特定步驟數目後即會停止。依預設，此為可用效果數目的 3 倍。或者，請指定正整數的步驟最大數目。

最佳子集選擇。這會檢查「所有可能」模式或是大於向前逐步的可能模式子集，以根據最佳子集條件來選擇最佳子集。**資訊準則 (AICC)** 是以指定模式訓練集的概似為基礎，且會經過調整以懲罰過於複雜的模式。**已調整 R 平方**是以訓練集的配適度為基礎，且會經過調整以懲罰過於複雜的模式。**過適預防準則 (ASE)** 過適預防準則 (平均平方誤，或 ASE) 為基礎。過適預防集是原始資料集 30% 左右的隨機子樣本，不用來訓練模式。

系統會將具有最大條件值的模式選作最佳模式。

註：最佳子集選擇較向前逐步選擇更需要大量計算。搭配 boosting、bagging 或極大資料集執行最佳子集時，其建立時間會長於使用向前逐步選擇建立的標準模式。

集合

系統在「目標」中要求 boosting、bagging 或極大資料集時，這些設定會決定所發生的集合行為。系統會忽略無法套用至所選目標的選項。

Bagging 與極大資料集。 系統執行集合評分時，可使用此規則來合併基底模型的預測值，以運算集合分數值。

- **連續目標的預設合併規則。** 系統會使用基底模型預測值的平均數或中位數，來合併連續目標的集合預測值。

請注意，若目標是用於強化模式準確性，則系統會忽略合併規則選擇。Boosting 會一律使用大部分的加權投票來為類別目標評分，並且使用加權中位數來為連續目標評分。

Boosting 與 Bagging。 當目標用於強化模式準確性或穩定性時，可指定欲建立的基底模式數目；若為 bagging，則此為重複取樣範例的數目。其應為正整數。

進階

複製結果。 設定亂數種子以供您複製分析。系統會使用亂數產生器來選擇過適預防集中的記錄。請指定一個整數，或是按一下「**產生**」以建立介於 1 和 2147483647 之間（含）的虛擬亂數整數。預設值為 54752075。

模型選項

模型名稱。您可以根據目標欄位來自動產生模型名稱，或是指定自訂名稱。自動產生的名稱為目標欄位名稱。

請注意，系統為模式評分時會一律計算預測值。新欄位的名稱即為目標欄位的名稱，且以 *$L-* 為字首。例如，若目標欄位名稱為 *sales*，則新欄位的名稱便是 *$L-sales*。

模型摘要

「模型摘要」視圖是一種 Snapshot模型及其配適的一覽摘要。

表格 該表格會識別一些高階模型設定，包括：

- 指定於欄位標籤的目標名稱，
- 自動資料準備是否如基本設定之指定執行，
- 模型選擇設定中指定的模型選擇方法和選擇準則。同時也會顯示最終模型的選擇準則值，以越小越佳的格式表示。

圖表　此圖表顯示最終模型的精確度，數值越大越好。此值為 100（最終模型的已調整 R^2）。

自動式資料準備

此檢視會顯示排除的欄位，以及在自動式資料準備 (ADP) 步驟中衍生轉換欄位方式的相關資訊。針對每個已轉換或排除的欄位，此表格會列出欄位名稱、分析當中的欄位角色，以及 ADP 步驟所採取的動作。系統會依照欄位名稱的字母順序以遞增方式來排序欄位。可能為每個欄位採取的行動包括：

- **衍生期間：月份**計算從包含日期的欄位值到目前系統日期的經過時間 (以月為單位)。
- **衍生期間：小時**計算從包含時間的欄位值到目前系統時間的經過時間 (以小時為單位)。
- **將測量層級從連續變更為次序**將唯一值少於 5 個的連續欄位重新分配為序數欄位。
- **將測量層級從序數變更為連續**將唯一值多於 10 個的序數欄位重新分配為連續欄位。
- **修整偏離值**會將超出分割值 (與平均值相距 3 個標準差) 的連續預測值設為分割值。
- **置換遺漏值**以模式置換名義欄位的遺漏值，以中位數置換序數欄位，以平均數置換連續欄位。
- **合併類別以最大化和目標的關聯**根據輸入和目標之間的關係來識別「類似的」預測值類別。系統會合併沒有顯著差異 (即 p 值大於 0.05) 的類別。
- **排除常數預測值 / 偏離值處理之後 / 合併類別之後**移除具有單一值的預測值，可能在採取其他 ADP 行動之後。

預測值重要性

一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

依觀察預測

這會根據水平軸上的觀察值，來顯示垂直軸上預測值的 Bin 散佈圖。理想的狀況下，點應排列在 45 度的線上；此檢視可以告訴您模式是否有預測結果特別差的記錄。

殘差

這會顯示模型殘差的診斷圖表。

圖表樣式。有不同的顯示樣式，可以從**樣式**下拉清單中進行存取。

- **直方圖。**此為經過 bin 處理的 studentized 殘差直方圖，其中含有常態分配的重疊圖。線性模型會假定殘差具有常態分配，因此直方圖應會十分貼近平滑線條。
- **P-P 圖。**此為經過 bin 處理的機率-機率圖，其會將 studentized 殘差與常態分配加以比較。若圖中各點所構成的線條斜率小於一般線條，則殘差所顯示的變異性會高於常態分配；若斜率越大，則殘差所顯示的變異性就會越低於常態分配。若繪製的點構成 S 形曲線，則殘差分配會呈現偏斜。

偏離值

此表格會列出對模式產生不當影響的記錄，並會顯示記錄 ID（若已於「欄位」標籤上指定）、目標值和 Cook's 距離。Cook's 距離是一種測量方式，若自模型係數計算中排除特定記錄，則其會測量所有記錄的殘差變更程度。若 Cook's 距離較大，則代表排除記錄已足以造成係數變更，因此應將其視為具有影響力。

您應該仔細檢查具有影響力的記錄，以決定是否要在估計模式時給予較少的加權，或是要將偏離值截斷至某些可接受的臨界值，或是完整移除具有影響力的記錄。

效果

此檢視會顯示模式中每個效果的大小。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **圖。**此圖表會依遞減的預測值重要性，來從高至低進行效果排序。系統會根據效果顯著性來加權處理圖中的連接線條，線條寬度越大代表符合越多的顯著效果 (較小的 p 值)。游標停留於連接線上時，會以工具提示的方式顯示 p 值和效果的重要性。此為預設值。
- **表格。** 此為整體模型與個別模型效應的 ANOVA 表格。系統會依遞減的預測值重要性，來從高至低進行個別效果排序。請注意，依預測表格會收合起來，僅顯示整體模型的結果。若要觀看個別模型效應的結果，請按一下表格中的「**已修正模型**」儲存格。

預測值重要性。具有「預測值重要性」滑塊，可控制視圖中所顯示的預測值。這不會變更模式，僅會讓您更能著重於最重要的預測值。依預設，系統會顯示前 10 個效果。

顯著性。「顯著」滑塊除了根據預測值重要性所顯示的效果之外，還可進一步控制視圖中可顯示的效果。系統會隱藏顯著值大於滑塊值的效果。這不會變更模式，僅會讓您更能著重於最重要的效果。預設值為 1.00，因此系統不會根據顯著性來過濾任何效果。

係數

此檢視會顯示模式中每個係數的值。請注意，模式當中的各項因素（類別預測值）皆已經過指標編碼，因此一般來說內含因素的**效果**會具有多個相關**係數**；除了對應於冗餘（參考）參數的類別外，每個類別會具有一個係數。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **圖。**此圖表會先顯示截距，然後再依遞減的預測值重要性，來從高至低進行效果排序。在內含因素的效果當中，系統會依升冪的資料值順序進行係數排序。圖表中的連接線會根據係數符號 (請參閱圖表鍵) 以彩色顯示和根據係數顯著性加權處理，線條寬度越大代表符合越多的顯著係數 (較小的 p 值)。游標停留在連接線上時，會以工具提示的方式顯示係數值、其 p 值，以及與參數相關之效果的重要性。此為預設樣式。
- **表格。** 此表格會顯示個別模型係數的值、顯著性測試和信賴區間。在截距之後，系統會依遞減的預測值重要性來從高至低進行效果排序。在內含因素的效果當中，系統會依升冪的資料值順序進行係數排序。請注意，依預設表格會收合起來，僅顯示係數、顯著性和每個模型參數的重要性。若要觀看標準誤、t 統計量和信賴區間，請按一下表格中的「**係數**」儲存格。游標停留在表格中的模型參數名稱上時，會以工具提示的方式顯示參數名稱、與參數相關的效果，以及（在類別預測值方面）與模型參數相關的數值標籤。在自動資料準備合併類別預測值的類似類別時，這對於觀看所建立的新類別建立特別有用。

預測值重要性。具有「預測值重要性」滑塊，可控制視圖中所顯示的預測值。這不會變更模式，僅會讓您更能著重於最重要的預測值。依預設，系統會顯示前 10 個效果。

顯著性。具有「顯著」滑塊，其除了根據預測值重要性所顯示的係數之外，還可進一步控制視圖中顯示的係數。系統會隱藏顯著值大於滑塊值的係數。這不會變更模式，僅會讓您更能著重於最重要的係數。預設值為 1.00，因此系統不會根據顯著性來過濾任何係數。

估計平均數

此為顯示顯著預測值的圖表。此圖表會針對水平軸上的每個預測值顯示垂直軸上的目標模型估計值，並且保留其他所有的預測值常數。其針對目標中每個預測值係數的效果提供實用的視覺化內容。

註：若無任何顯著預測值，則不會產生任何估計平均數。

模型建置摘要

若在「模型選擇」設定中選擇有別於「**無**」的其他模式選擇演算法，則此項目會提供有關模型建置程序的部分詳細資料。

向前逐步。若採用向前逐步作為選擇演算法，則表格會顯示逐步演算法中的最後 10 個步驟。系統會針對每個步驟，顯示步驟中模型的選擇條件值和效果。這有助於使您瞭解每個步驟對於模型的貢獻度。您可以在每個行中排序列，以便更加輕鬆地查看指定步驟中的模型效應。

最佳子集。若採用最佳子集作為選擇演算法，則表格會顯示前 10 個模式。系統會針對每個模式，顯示模式中的選擇條件值和效果。這有助於使您瞭解頂端模型的穩定性；若這些模式當中的類似效果差異不大，則您便可安心信賴這些「頂端」模式；若這些模式當中的效果差異過大，則表示某些效果過於近似而應加以合併 (或移除)。您可以在每個行中排序列，以便更加輕鬆地查看指定步驟中的模型效應。

設定

請注意，系統為模式評分時會一律計算預測值。新欄位的名稱即為目標欄位的名稱，且以 *$L-* 為字首。例如，若目標欄位名稱為 *sales*，則新欄位的名稱便是 *$L-sales*。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **透過轉換為原生 SQL 進行評分** 如果選取此項，則會產生原生 SQL 來在資料庫內對模型進行評分。

  註：雖然這個選項可以更快地提供結果，但隨著模型複雜性的增加，原生 SQL 的大小和複雜性也會增加。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

線性-AS 節點

IBM SPSS Modeler 有兩個不同版本的線性節點：

- **線性**是在 IBM SPSS Modeler Server 上執行的傳統節點。
- 連接至 IBM SPSS Analytic Server 時，可以執行**線性 AS**。

線性回歸是一個一般統計技術，用來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。

需求。 在線性迴歸模型中只能使用數值型欄位和種類預測值。 您必須只有一個目標欄位（角色設為**目標**）以及一個以上預測值（角色設為**輸入**）。角色為**兩者**或**無**的欄位會予以忽略，因為它們是非數值欄位。（必要的話，可使用「衍生」節點對非數值欄位重新編碼。）

強度。 線性迴歸模型相對簡單，提供易於解譯的數學公式來產生預測內容。由於線性迴歸是一種由來已久的統計方法，因此這些模型的內容已廣為人所熟知。通常，線性模型的訓練速度也非常快。「線性」節點提供了自動欄位選擇方法，以排除方程式中不重要的輸入欄位。

註：如果目標欄位為種類（例如**是/否**或**流失/未流失**）而非連續範圍，那麼可以將邏輯迴歸用作替代項。 邏輯迴歸也提供對非數值輸入的支援，不需要對這些欄位重新編碼。請參閱第 154 頁的『Logistic 節點』主題，以取得更多資訊。

線性-AS 模型

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。

線性模型相當簡便，且提供了簡易的評分數學公式。與同一個資料集上的其他模式類型 (如神經網路或決策樹) 相較，這些模型的內容比較容易瞭解，通常可以快速建立。

範例。 某資源有限的保險公司，打算調查屋主的保險理賠，希望建立估計理賠成本的模式。透過在服務中心部署這個模式，代表就能在和客戶通電話的同時輸入資訊，並根據過去的資料立即取得「預期的」理賠成本。

欄位需求。必須具有「目標」以及至少一個的「輸入」。依預設，不會使用預先定義角色為「兩者」或「無」的欄位。目標必須是連續的 (尺度)。對預測值（輸入）沒有測量層次限制。種類（旗標、列名和序數）欄位用作模型中的因子，同時連續欄位用作共變數。

基本

包括截距。 當 x 軸為 0 時，該選項在 y 軸上包含偏移量。截距通常包含在模型中。 但是如果可以假設資料會穿過原點，則可以將截距排除在外。

考慮雙向互動。 該選項會告訴模型比較每個可能的輸入對，以瞭解各輸入對的趨勢之間是否會互相影響。 如果會互相影響，那麼這些輸入更有可能包含在設計矩陣中。

係數預估值的信賴區間 (%)。 此為用於係數視圖中計算模型係數的估計值的信賴度區間。 指定一個大於 0 且小於 100 的值。預設值是 95。

類別預測的排序。 這些控制項用於確定因子（種類輸入）種類的順序，以確定「最後一個」種類。 如果輸入不是種類目標或者指定了自訂參照種類，那麼將忽略排序設定。

模型選擇

模型選擇方法。選擇其中一種模式選擇方法 (詳細資訊如下) 或「**包含所有預測值**」，以僅將所有可用的預測值輸入為主效應模式項目。依預設，系統會使用「**向前逐步**」。

轉遞逐步選擇。此方法一啟動並不會對模式產生任何效果，但會根據逐步條件一步一步新增或移除效果，直至無法再新增或移除任何效果為止。

- **進入/移除準則。**此統計量決定是否應在模式中新增或移除效果。**資訊準則 (AICC)** 是以指定模式訓練集的概似為基礎，且會經過調整以懲罰過於複雜的模式。**F 統計量**是以模式錯誤改善的統計測試為基礎。**已調整 R 平方**是以訓練集的配適度為基礎，且會經過調整以懲罰過於複雜的模式。**過適預防準則 (ASE)** 過適預防準則 (平均平方誤，或 ASE) 為基礎。過適預防集是原始資料集 30% 左右的隨機子樣本，不用來訓練模式。

  若選擇任何非「**F 統計量**」的條件，則系統會將每個步驟中對應至最大正向增加條件的效果新增至模式。系統會移除模式中任何對應至減少條件的效果。

  若選擇「**F 統計量**」作為條件，則系統會在每個步驟中，將最小 p 值小於指定臨界值（「**包含 p 值小於此值的效果**」）的效果新增至模式。預設值為 0.05。系統會移除任何 p 值大於指定臨界值（「**移除 p 值大於此值的效果**」）之模式中的效果。預設值是 0.10。
- **自訂最終模型中的效果最大數目。**依預設，系統會將所有可用的效果輸入至模式。或者，若逐步演算法以指定的最大效果數目來結束步驟，則演算法在停止時會保有目前的效果集。
- **自訂步驟的最大數目。** 逐步演算法在經過特定步驟數目後即會停止。依預設，此為可用效果數目的 3 倍。或者，請指定正整數的步驟最大數目。

最佳子集選擇。這會檢查「所有可能」模式或是大於向前逐步的可能模式子集，以根據最佳子集條件來選擇最佳子集。**資訊準則 (AICC)** 是以指定模式訓練集的概似為基礎，且會經過調整以懲罰過於複雜的模式。**已調整 R 平方**是以訓練集的配適度為基礎，且會經過調整以懲罰過於複雜的模式。**過適預防準則 (ASE)** 過適預防準則 (平均平方誤，或 ASE) 為基礎。過適預防集是原始資料集 30% 左右的隨機子樣本，不用來訓練模式。

系統會將具有最大條件值的模式選作最佳模式。

註：最佳子集選擇較向前逐步選擇更需要大量計算。搭配 boosting、bagging 或極大資料集執行最佳子集時，其建立時間會長於使用向前逐步選擇建立的標準模式。

模型選項

模型名稱。您可以根據目標欄位來自動產生模型名稱，或是指定自訂名稱。自動產生的名稱為目標欄位名稱。

請注意，系統為模式評分時會一律計算預測值。新欄位的名稱即為目標欄位的名稱，且以 *$L-* 為字首。例如，若目標欄位名稱為 *sales*，則新欄位的名稱便是 *$L-sales*。

互動式輸出

執行線性-AS 模型後，下列輸出可用。

模型資訊

「模型資訊」視圖提供有關模型的關鍵資訊。該表格識別一些高階模型設定，例如：

- 欄位標籤上指定的目標名稱
- 迴歸方法加權欄位
- 模型選擇設定上指定的模型建置方法
- 預測值輸入數目
- 最終模型中預測值的數目
- 經過糾正的 Akaike 資訊準則 (AICC)。 AICC 是一種用於基於 -2（受限）對數概似選取和比較混合模型的測量方法。 數值越小代表模式越佳。AICC 會「修正」較小 AIC 的樣本大小。當樣本大小增加時，AICC 會收斂至 AIC。
- R 平方。 這是線性模型的適合度測量，有時稱為決定係數。 由迴歸模型所解釋的因變數變異的比例。它的範圍介於 0 到 1 之間。數值越小代表模式越不適合資料。
- 已調整的 R 平方

記錄摘要

「記錄摘要」視圖提供模型中所包括以及排除的記錄（觀察值）的數目與百分比。

預測值重要性

一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

已依觀察預測

這會根據水平軸上的觀察值，來顯示垂直軸上預測值的 Bin 散佈圖。理想的狀況下，點應排列在 45 度的線上；此檢視可以告訴您模式是否有預測結果特別差的記錄。

設定

請注意，系統為模式評分時會一律計算預測值。新欄位的名稱即為目標欄位的名稱，且以 *$L-* 為字首。例如，若目標欄位名稱為 *sales*，則新欄位的名稱便是 *$L-sales*。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：在程序中使用「伺服器評分配接卡」（如有安裝的話）來評分。** 如果連接至安裝了評分配接卡的資料庫，則使用評分配接卡和關聯使用者定義的函數 (UDF) 來產生 SQL，並對資料庫中的模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫之外評分。** 此選項會從資料庫提取回您的資料，並在 SPSS Modeler 中對其進行評分。

Logistic 節點

邏輯迴歸（也稱為**名義迴歸**）是一種用於依據輸入欄位的值對記錄進行分類的統計技術。這種技術與線性迴歸類似，但用種類目標欄位代替了數值型欄位。同時受支援二項式模型（用於具有兩種離散種類的目標）和多項式模型（用於具有兩種以上種類的目標）。

邏輯迴歸的工作原理是建立一組方程式，使輸入欄位值與每個輸入欄位種類所關聯的機率相關。產生模型後，便可以用它來估計新資料的機率。對於每一筆記錄，會為每個可能的輸出種類計算成員資格的機率。會將具有最高機率的目標種類指派為該記錄的預測輸出值。

二項式模型範例。某個電信服務提供商關心流失到競爭對手那裡的客戶數。使用服務利用率資料，可以建立二項式模型以預測哪些客戶有可能轉向其他提供商，並自訂服務以保留盡可能多的客戶。由於目標具有兩個截然不同的種類（可能流失或不流失），因此將使用二項式模型。

註：字串欄位的長度限制為 8 個字元（僅適用於二項式模型）。如有必要，可以使用「再分類」節點或使用「匿名化」節點對較長的字串進行重新編碼。

多項式模型範例。某電信公司根據服務使用方式來切割客戶數量，並將客戶分成四個組別。通過使用人口統計資料預測群組成員資格，您可以建立多項式模型來將潛在客戶分成多個群組，然後針對各個客戶自訂報價。

需求。 一個或多個輸入欄位和唯一一個具有兩個或多個種類的種類目標欄位。對於二項式模型，目標必須具有旗標測量層次。於多項式模型，目標可以具有旗標，或列名的測量層次，以及兩個或多個種類。會忽略設為兩者 或無 的欄位。模型中所用的欄位必須已完全實例化其類型。

強度。 通常，邏輯迴歸模型非常準確。它們可處理符號和數值類型的輸入欄位。它們可以給出所有目標種類的預測機率，從而能夠輕鬆識別出次佳推測值。當群組成員資格是真正種類欄位時，Logistic 模型最為有效；如果群組成員資格基於連續範圍欄位的值（例如，高 IQ 與低 IQ），那麼應考慮使用線性迴歸，以利用整個範圍的值所提供的更豐富的資訊。Logistic 模型還可以執行自動欄位選擇，但其他方法（例如樹狀結構模型或功能選擇）在對大型資料集執行此操作時可能速度更快。最後，由於 Logistic 模型被很多分析師和資料挖掘人員所熟知，因此他們可能會將其用作比較其他建模技術的基礎。

正在處理大型資料集時，可以取消進階輸出選項概似比測試，從而顯著提高效能。請參閱第 158 頁的『邏輯迴歸進階輸出』主題，以取得更多資訊。

重要：如果暫時磁碟空間較少，二項式邏輯迴歸可能無法建立，並會顯示錯誤。當根據大型資料集（10GB 或更多）進行建立時，需要相同的可用的磁碟空間數量。您可以使用環境變數 SPSSTMPDIR 來設定暫存目錄的位置。

Logistic 節點模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

程序。指定將建立二項式模型還是多項式模型。對話框中提供的選項會因所選建模程序的類型而異。

- **二項式。** 當目標欄位是具有兩個離散（二分）值（如是/否、啟動/關閉或男/女）的旗標或列名欄位時使用。
- **多項式。** 當目標欄位是具有兩個以上值的列名欄位時，應使用此選項。可以指定**主效應**、**全析因**或**自訂**。

方程式中含有常數項。 此選項用於確定生成的方程式中是否將包含常數項目。在大部分狀況下，您應該選取此選項。

二項式模型

對於二項式模型，可用的方法和選項如下：

方法。指定要用來建置邏輯迴歸模型的方法。

- **輸入。** 這是預設方法，其直接在方程式中輸入所有項。在建置模型時並未執行欄位選擇。
- **逐步向前法。**顧名思義，欄位選擇逐步向前法用於分步建立方程式。起始模型可能是最簡單的模型，其方程式中沒有任何模型項（除了常數）。在每一個步驟中，會對尚未新增至模型的項目進行評估，並且如果這些項目的最適性會顯著性地新增至模型的預測能力，則會予以新增。此外，會對目前位於模型中的項進行重新評估，以判定是否可以移除其中任一項而不會顯著地減損模型。如果是的話，則會將其移除。程序重複執行，新增及/或移除其他項。當無法新增更多項以改善模型，以及無法移除更多項而不減損模型時，便會產生最終模型。
- **向後逐步。** 逐步向後法與向前逐步方法在本質上是相反的。使用此方法時，起始模型包含所有項作為預測值。在每一個步驟中，會對模型中的項進行評估，並且會移除可以移除而不會顯著地減損模型的任何項。此外，會對先前移除的項進行重新評估，以判定這些項的最適性是否會顯著性地新增至模型的預測能力。若是如此，則會將其新增回模型。當無法移除更多項而不會顯著地減損模型，以及無法新增更多項以改善模型時，便會產生最終模型。

種類輸入。列出識別為種類欄位的欄位，即具有旗標、列名或序數的測量層次。可以為每個種類欄位指定對照和基本種類。

- **欄位名稱。** 此欄包含種類輸入的欄位名稱。要向此欄中新增連續輸入欄位或數值型輸入欄位，請按一下清單右邊的「新增欄位」圖示，然後選取所需輸入欄位。
- **對比** 種類欄位的迴歸係數的解譯取決於使用的對照。對照會決定如何設定假設測試，以比較估計平均數。例如，如果已知某個種類欄位具有隱含順序（如型樣或分組），那麼可以使用對照為該順序建模。可用的對照如下：

  指標。 對比指出該類別成員是否存在。此為預設的方法。

  簡單。 將預測值欄位的每個種類（參照種類除外）與參照種類進行比較。

  差異。 將預測值欄位的每個種類（第一個類別除外）與先前種類的平均值效果進行比較。這種對比也叫作反「赫爾莫特 (Helmert) 對比」。

Helmert。 將預測值欄位的每個種類（最後一個種類除外）與隨後的一個種類的平均效果進行比較。

重複。 將預測值欄位的每個種類（第一個類別除外）與前一個種類進行比較。

多項式。 正交多項式對比。它假設類別間距都是相等的。多項式對照僅適用於數值型欄位。

離差。 將預測值欄位的每個種類（參照種類除外）與整體效果進行比較。

- **基本種類。**指定如何針對選定的對照類型確定參照種類。選取**第一個**以使用輸入欄位的第一個類別（按字母排列），或選取**最後一個**以使用最後一個種類。預設基底種類適用於**種類輸入**區域中列出的變數。

  註：如果對照設定為「差分」、Helmert、「重複」或「多項式」，那麼此欄位不提供。

每個欄位對整體回應效果的估計，可以計算為其他各個種類相對於參照種類的概似增量或減量。這可協助您識別更有可能提供特定回應的欄位和值。

基本種類在輸出中顯示為 0.0。這是因為將它與自身相比較產生的結果為空。所有其他種類顯示為與基本種類相關的對等項目。請參閱第 160 頁的『Logistic 模型塊詳細資料』主題，以取得更多資訊。

多項式模型

對於多項式模型，可用的方法和選項如下：

方法。指定要用來建置邏輯迴歸模型的方法。

- **輸入。** 這是預設方法，其直接在方程式中輸入所有項。在建置模型時並未執行欄位選擇。
- **逐步。** 用於欄位選擇的「逐步」方法會逐步建置方程式，如名稱所示。起始模型可能是最簡單的模型，其方程式中沒有任何模型項（除了常數）。在每一個步驟中，會對尚未新增至模型的項目進行評估，並且如果這些項目的最適性會顯著性地新增至模型的預測能力，則會予以新增。此外，會對目前位於模型中的項進行重新評估，以判定是否可以移除其中任一項而不會顯著地減損模型。如果是的話，則會將其移除。程序重複執行，新增及/或移除其他項。當無法新增更多項以改善模型，以及無法移除更多項而不減損模型時，便會產生最終模型。
- **向前。** 欄位選擇向前法與分步建立模型的逐步方法類似。但採用這種方法時，初始模型是最簡單的模型，只能向模型中新增常數和項目。每個步驟會對尚未納入到模型中的項目進行測試，看它們對模型的改進起多大作用，然後將其中的最佳項目新增至模型中。當無法再新增任何項目、或最佳備選項目無法對模型產生足夠的改善時，便會產生最終模型。
- **往回。** 向後法與向前法在本質上是相反的。但採用這種方法時，初始模型包含作為預測值的所有項目，只能從模型中刪除項目。對模型影響較小的模型項目將被逐一刪除，直到無法再刪除任何項目而不對模型功能造成重大損害，從而生產最終模型。
- **向後逐步。** 「逐步往回」方法實質上與「逐步」方法相反。使用此方法時，起始模型包含所有項作為預測值。在每一個步驟中，會對模型中的項進行評估，並且會移除可以移除而不會顯著地減損模型的任何項。此外，會對先前移除的項進行重新評估，以判定這些項的最適性是否會顯著性地新增至模型的預測能力。若是如此，則會將其新增回模型。當無法移除更多項而不會顯著地減損模型，以及無法新增更多項以改善模型時，便會產生最終模型。

註：自動方法（包含逐步、逐步向前和逐步向後）是適應性強的學習方法，並且特別容易過度擬合訓練資料。使用這些方法時，請務必使用新資料或以「分割區」節點建立的留出法測試範例，來驗證所產生模型的有效性。

目標的基本種類。指定如何決定參照種類。這將用作對目標中所有其他種類的迴歸方程式進行估計的基礎。選取**第一個**以使用目前目標欄位的第一個類別（按字母排列），或選取**最後一個**以使用最後一個種類。或者，可以選擇**指定**以選擇特定種類，並從清單中選擇所需的值。可以在類型節點中為每個欄位定義可用的值。

通常應將關注程度最低的種類指定為基本種類，例如低價促銷產品。然後再以相對方式將其他種類與該基本種類相關，從而確定什麼使它們更有可能自成種類。這可協助您識別更有可能提供特定回應的欄位和值。

基本種類在輸出中顯示為 0.0。這是因為將它與自身相比較產生的結果為空。所有其他種類顯示為與基本種類相關的對等項目。請參閱第 160 頁的『Logistic 模型塊詳細資料』主題，以取得更多資訊。

模型類型 有三個選項用於定義模型中的項目。**主效應**模型僅包含各個輸入欄位，而不測試輸入欄位之間的互動（相乘性作用）。**全因子**模型包含所有互動以及輸入欄位主效應。全因子模型擷取複合關係的能力較強，但也比較難以解釋，而且更有可能出現過度配適情況。由於有可能出現大量可能組合，因此對於全因子模型，自動欄位選擇方法（輸入以外的方法）處於停用狀態。**自訂**模型僅包括您指定的項目（主效應與互動）。選取此選項時，請使用「模型塊」清單在模型中新增或移除項目。

模型塊。 建置自訂模型時，您必須在模型中明確指定項目。該清單顯示模型目前的項目集。「模型塊」清單右邊的按鈕用於新增和移除模型塊。

- 若要將項目新增至模型，請按一下新增模型塊 按鈕。
- 若要刪除項目，請選取所需的項目，然後按一下刪除選定的模型塊 按鈕。

將項目新增到邏輯迴歸模型

要求自訂邏輯迴歸模型時，可以通過按一下「邏輯迴歸模型」標籤中的新增新的模型塊按鈕將項目新增到模型中。這將開啟「新建項目」對話框，您可在其中指定項目。

要新增的項目類型。 可根據在可用欄位清單中選取的輸入欄位，使用數種方法將項目新增至模型。

- **單一互動。** 插入代表所有所選欄位之互動的項目。
- **主效應。** 為每個選定的輸入欄位插入一個主效應項目（欄位本身）。
- **所有雙向互動。** 針對所選輸入欄位的每個可能的配對，插入一個雙向互動項目（輸入欄位的product）。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B* 和 *C*，則此方法將插入項目 *A * B*、*A * C* 和 *B * C* 中。
- **所有 3 向互動。** 針對所選輸入欄位的每個可能的組合，插入一個 3 向互動項目（輸入欄位的product），一次性採用 3 個。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B*、*C* 和 *D*，則此方法將插入項目 *A * B * C*、*A * B * D*、*A * C * D* 和 *B * C * D* 中。
- **所有 4 向互動。** 針對所選輸入欄位的每個可能的組合，插入一個 4 向互動項目（輸入欄位的product），一次性採用 4 個。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B*、*C*、*D* 和 *E*，則此方法將插入項目 *A * B * C * D*、*A * B * C * E*、*A * B * D * E*、*A * C * D * E* 和 *B * C * D * E* 中。

可用的欄位。 列出要用來建構模型塊的可用輸入欄位。

預覽。 根據上述選取的欄位和項目類型，顯示按一下**插入**時將新增到模型中的項目。

插入。 根據目前選取的欄位與項目類型將項目插入模型，然後關閉對話框。

Logistic 節點專家選項

如果您對邏輯迴歸方法有詳細瞭解，那麼可以通過專家選項對訓練過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

尺度（僅限多項式模型）。您可以指定將用於更正參數共變數矩陣的估計值的離差尺度值。**Pearson** 使用 Pearson 卡方測試統計資料來估計尺度值。**離差**使用離差函數（概似比卡方）統計資料來估計尺度值。您也可以指定自己的使用者定義尺度值。但其必須為正數值。

附加所有機率。 如果已選取此選項，那麼會將輸出欄位的每個種類的機率新增到節點所處理的每條記錄中。如果未選取此選項，則只新增預測種類的機率。

例如，包含具有三個種類的多項式模型結果的表格將包含五個新欄。一個欄將列出預測正確的結果的機率，第二個欄將顯示該預測準確或失誤的機率，第三個欄將顯示每個種類的預測失誤或準確的機率。請參閱第 160 頁的『Logistic 模型塊』主題，以取得更多資訊。

附註：對於二項式模型，此選項始終處於已選取狀態。

奇異性容忍值。 指定檢查特異值時使用的允差。

聚合。 通過這些選項，您可以控制用於模型收斂的參數。執行模型時，聚合設定會控制重複執行不同參數以查看其適合度的次數。嘗試參數的頻率更高，結果越接近（亦即，結果將聚合）。請參閱『邏輯迴歸收斂選項』主題，以取得更多資訊。

輸出。這些選項可讓您要求將顯示在節點所建置之模型區塊進階輸出中的其他統計資料。請參閱『邏輯迴歸進階輸出』主題，以取得更多資訊。

逐步。 這些選項可讓您使用「逐步」、「向前」、「往回」、「逐步往回」等估計方法來控制新增及移除欄位的準則。（如果選取了「輸入」方法，則會停用該按鈕。）請參閱第 159 頁的『邏輯迴歸執行步驟選項』主題，以取得更多資訊。

邏輯迴歸收斂選項

您可設定用於邏輯迴歸模型估計的收斂參數。

最大疊代。 指定用來估計模型的疊代數目上限。

最大的半階次數。 逐步二分法是邏輯迴歸用於處理估計過程的複雜性的技術。在通常情況下，應使用預設設定。

對數概似收斂。 如果對數概似中的相對變更小於此值，則會停止疊代。如果值為 0 便不會使用這個條件。

參數收斂條件。 如果參數估計值中的絕對變更或相對變更小於此值，則會停止疊代。如果值為 0 便不會使用這個條件。

Delta（僅限多項式模型）。您可以指定要新增到每個空白 Cell（輸入欄位和輸出欄位值的組合）中的值，該值介於 0 和 1 之間。當相對於資料中的記錄數有多數可能的欄位值組合時，此值有助於估計演算法處理資料。預設值為 0。

邏輯迴歸進階輸出

選取要在迴歸模型塊的進階輸出中顯示的選用輸出。若要檢視進階輸出，請瀏覽模型片段並按一下**進階**標籤。請參閱第 162 頁的『Logistic 模型塊進階輸出』主題，以取得更多資訊。

二項選項

選取要為模型產生的輸出類型。請參閱第 162 頁的『Logistic 模型塊進階輸出』主題，以取得更多資訊。

顯示。 選取是在每一步驟顯示結果還是等到所有步驟已完成時再顯示結果。

exp(B) 的 CI。選取表示式中每個係數（顯示為 Beta）的信賴區間。指定信賴區間的等級（預設值為 95%）。

殘差診斷。要求殘差的「全部觀察值診斷」表格。

- **離群值極限（標準差）。.** 僅列出這樣的殘差觀察值：所列變數的絕對標準化值至少與您指定的值一樣大。預設值為 2。
- **全部觀察值。** 在殘差的「全部觀察值診斷」表格中包含全部觀察值。

  註：由於此選項將列出每條輸入記錄，因此可能在報告中產生極大的表格的表格，其中每條記錄佔一行。

分類分割值。 此選項可用於確定對觀察值進行分類的分割點。超過分類分割的預測觀察值會分類為正向，而小於分割的預測觀察值會分類為負向。若要改變預設值，請輸入介於 0.01 和 0.99 之間的數值。

多項式選項

選取要為模型產生的輸出類型。請參閱第 162 頁的『Logistic 模型塊進階輸出』主題，以取得更多資訊。

註：選取**概似比測試**選項將極大地增加建立邏輯迴歸模型所需的處理時間。如果模型建立時間過長，可以考慮取消此選項，或利用 Wald 統計資料和分數統計資料。請參閱『邏輯迴歸執行步驟選項』主題，以取得更多資訊。

疊代過程間隔。選取在進階輸出中列印疊代狀態的分步間隔。

信賴區間。 方程式中係數的信賴區間。指定信賴區間的等級（預設值為 95%）。

邏輯迴歸執行步驟選項

這些選項可讓您使用「逐步」、「向前」、「往回」、「逐步往回」等估計方法來控制新增及移除欄位的準則。

模型中的項目數（僅限多項式模型）。您可以指定模型中的下限項目數（針對向後法和逐步向後法模型）和上限項目數（針對向前法和逐步模型）。如果指定大於 0 的最小值，模型將包含該數量的項目，即使根據統計準則應將其中某些項目刪除也是如此。對於前進法、逐步法和輸入模型，將忽略最小值設定。如果指定最大值，可能會省略模型中的某些項目，即使根據統計準則應將其已選取也是如此。對於向後法、逐步向後法和輸入模型，將忽略**指定最大值**設定。

輸入準則（僅限多項式模型）。選取**分數**可以最大化處理速度。**概似比率**選項可能會稍微多提供一些有力的估計值，但所需的計算時間較長。預設設定是使用分數統計資料。

移除準則。 為更強健的模型選取**概似比**。若要縮短建置模型的時間，您可以嘗試選取 **Wald**。但是，如果資料中有完全或半完全分隔（可使用模型塊的「進階」標籤確定），Wald 統計資料將變得極不可靠，不應採用。預設設定是使用概似比統計資料。對於二項式模型，還有其他選項**條件式**。此選項提供以基於條件參數估計值的概似比統計資料的機率為依據的移除測試。

準則的顯著性臨界值。 通過此選項，您可以根據與每個欄位相關聯的統計機率（*p* 值）來指定選擇準則。唯有當關聯的 *p* 值小於**輸入**值時才會將欄位新增至模型，而唯有當 *p* 值大於**移除**值時才會移除欄位。**輸入**值必須小於**移除**值。

輸入或移除的要求（僅限多項式模型）。對於某些應用程式，除非模型也包含互動項目所涉及欄位的低階數項目，否則將互動項目新增到模型中在數學上沒有意義。例如，除非 *A* 和 *B* 也納入到模型中，否則將 *A * B* 納入到模型中沒有意義。使用這些選項，可以確定如何在逐步模型項目選擇過程中處理這些相依關係。

- **用於離散作用的階層。**僅當相關欄位的所有低階數作用（涉及較少欄位的主效應或互動）均位於模型中時，高階數作用（涉及較多欄位的互動）才會進入模型，並且如果涉及相同欄位的高階數作用位於模型中，那麼將不會刪除低階數作用。此選項僅適用於種類欄位。
- **所有作用的階層。**此選項的工作原理與上一選項相同，但它適用於所有輸入欄位。

- **包含所有作用。**僅當作用中包含的所有作用也納入到模型中時，該作用才能納入到模型中。此選項與**用於所有效果的層次**選項類似，只是連續欄位的處理方式略有不同。要讓一個作用包含另一個作用，被包含（低階數）作用必須包含（高階數）作用中涉及的所有連續欄位，且被包含作用的種類欄位必須是包含作用中種類欄位的子集合。例如，如果 *A* 和 *B* 是種類欄位，*X* 是連續欄位，那麼項目 *A * B * X* 將包含項目 *A * X* 和 *B * X*。
- **無。** 不會強制執行任何關係；模型中項目的新增和刪除是獨立的。

Logistic 模型塊

Logistic 模型塊代表由 Logistic 節點估計的方程式。其中包含邏輯迴歸模型擷取的所有資訊，以及有關模型結構和效能的資訊。這種類型的方程式也可以通過其他模型（如 Oracle SVM）產生。

執行包含 Logistic 模型塊的串流時，該節點將新增兩個包含模型預測和相關機率的新欄位。新欄位的名稱衍生自所預測的輸出欄位的名稱，並帶有表示預測種類的字首 *$L-* 或表示相關機率的字首 *$LP-*。例如，對於名為 *colorpref* 的輸出欄位，新欄位將命名為 *$L-colorpref* 和 *$LP-colorpref*。此外，如果在 Logistic 節點中已選取了**附加所有可能性**選項，那麼會針對輸出欄位的每個種類新增一個附加欄位，其中包含的每條記錄對應種類的機率。這些附加欄位的名稱基於輸出欄位的值，並帶有字首 *$LP-*。例如，如果 *colorpref* 的合法值為 *Red*、*Green* 和 *Blue*，那麼將新增以下三個新欄位：*$LP-Red*、*$LP-Green* 和 *$LP-Blue*。

產生過濾器節點。 「產生」功能表可讓您建立新的「過濾器」節點以根據模型結果來傳遞輸入欄位。因多重共線性而從模型中刪除的欄位以及模型中未使用的欄位將被產生的節點過濾。

Logistic 模型塊詳細資料

對於多項式模型，Logistic 模型塊的「模型」標籤採用分屏顯示，模型方程式顯示在左側窗格中，而預測值重要性顯示在右側窗格中。對於二項式模型，此標籤只顯示預測值重要性。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

模型方程式

對於多項式模型，左窗格顯示為邏輯迴歸模型估計的實際方程式。在目標欄位中，除基本種類之外，每種種類均有一個方程式。這些方程式以樹狀結構格式顯示。這種類型的方程式也可以通過某些其他模型（如 Oracle SVM）產生。

使用的方程式。顯示用於在給定一組預測值的情況下推衍生目標種類機率的迴歸方程式。目標欄位的最後一個種類將被視為**基本種類**；顯示的方程式將針對一組特定預測值給出其他種類相對於基本種類的對數機率。給定預測值型樣的每個種類的預測機率根據這些對數機率值推導得出。

如何計算機率

每個方程式會計算一個特定目標種類相對於基本種類的對數機率。**對數機率**（也稱為**羅吉特機率**）是指定目標種類相對於基本種類的機率比，並在結果中套用對數函數。對於基本種類，種類相對於自身的優勢比為 1.0，因此其對數機率為 0。可以將這種情況視為基本種類的隱含方程式，其中所有係數均為 0。

要根據特定目標種類的對數機率推衍生機率，需要取該種類的方程式計算的羅吉特機率值，並套用下列方程式：

$$P(group_i) = \exp(g_i) / \Sigma_k \exp(g_k)$$

其中 g 是計算的對數機率， i 是種類參考號， k 為 1 至目標種類數之間的數字。

預測值重要性

選擇性地，指出評估模型時每個預測值的相對重要性的圖表，可能也會顯示在「模型」標籤上。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。注意，只有在產生模型之前已選取「分析」標籤上的**計算預測值重要性**，才可以使用此圖表。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

附註：與其他類型的模型相比，計算邏輯迴歸的預測值重要性可能需要更長時間，因此依預設在「分析」標籤中未選取預測值重要性。選取此選項可能會降低效能，尤其是使用大型資料集時。

Logistic 模型塊概要

邏輯迴歸模型的概要顯示用於產生該模型的欄位和設定。此外，如果您已執行連接至此建模節點的分析節點，則也會在此小節中顯示來自該分析的資訊。如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』。

Logistic 模型塊設定

Logistic 模型塊中的「設定」標籤用於指定模型評分過程中的信賴度、機率、傾向評分和 SQL 產生選項。該標籤僅在已將模型塊新增到串流中之後才可用，而且可以根據模型和目標的類型顯示不同選項。

多項式模型

對於多項式模型，可用的選項如下。

計算信賴度：指定是否在評分期間計算信賴度。

計算原始傾向分數（僅限旗標目標）：（僅限於具有旗標目標的模型）您可以要求生成原始傾向分數，這些分數指示對目標欄位指定的 true 結果的概似值。 這些是標準預測及信賴度值的附加項目。無法使用調整傾向評分。請參閱第 29 頁的『建模節點分析選項』主題，以取得更多資訊。

附加所有機率 指定是否將每個種類的輸出欄位的機率新增至節點所處理的每筆記錄。如果未選取此選項，則只新增預測種類的機率。例如，對於具有三種種類的列名目標，評分輸出針對三種種類的每一種都包含一欄，並包含第四欄指示預測任一種類的機率。例如，如果種類紅色、綠色和藍色的機率分別是 0.6、0.3 和 0.1，那麼預測種類將為紅色，其中機率為 0.6。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **透過轉換為原生 SQL 進行評分** 如果選取此項，則會產生原生 SQL 來在資料庫內對模型進行評分。

  註：雖然這個選項可以更快地提供結果，但隨著模型複雜性的增加，原生 SQL 的大小和複雜性也會增加。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

註：對於多項式模型，如果已選取**附加所有機率**，那麼 SQL 產生功能無法使用；或者，對於具有標準目標的模型，如果已選取**計算信賴度**，那麼 SQL 產生功能無法使用。僅僅對具有旗標目標的多項式模型，支援具有信賴度計算的 SQL 產生功能。SQL 產生功能無法用於二項式模型。

二項式模型

對於二項式模型，信賴度和機率始終處於啟用狀態，並且用於取消這些選項的設定無法使用。SQL 產生功能無法用於二項式模型。對於二項式模型，唯一可以變更的設定是計算原始傾向分數的功能。正如以上針對多項式模型的說明，此內容適用於只具有旗標目標的模型。請參閱第 29 頁的『建模節點分析選項』主題，以取得更多資訊。

Logistic 模型塊進階輸出

邏輯迴歸（也稱為**名義迴歸**）的進階輸出將提供有關估計模型及其效能的詳細資訊。進階輸出包含的大部分資訊技術含量很高，需要具備邏輯迴歸分析方面的廣泛知識才能正確理解該輸出。

警告。 指出結果的任何警告或潛在問題。

觀察值處理摘要。 列出由模型中的每個符號欄位處理和細分的記錄數。

步驟摘要（選用）。列出使用自動欄位選擇時在模型建立過程的每個步驟中新增或移除的作用。

註：僅針對逐步、向前法、向後法或逐步向後法顯示此選項。

疊代過程（選用）。顯示從起始估計值開始的每 *n* 次疊代的參數估計值的疊代過程，其中 *n* 是列印間隔值。預設設置是列印每一個疊代 (n=1)。

模型配適度資訊（多項式模型）。顯示根據其中所有參數係數均為 0（僅有截距）的模型對模型（最終模型）進行的概似比測試。

分類（選用）。顯示輸出欄位預測值和實際值的百分比矩陣。

擬合度卡方統計資料（選用）。顯示 Pearson 和概似比卡方統計資料。這些統計資料可測試模型對訓練資料的整體擬合度。

Hosmer-Lemeshow 擬合度（選用）。顯示將觀察值分組為風險的十分位數並對每個十分位數中的觀測機率與預期機率進行比較的結果。此適合度統計值量比多項式模型中採用的傳統適合度統計值量更為穩健，尤其適用於具有連續共變數的模型和小樣本的研究。

假 R-平方（選用）。顯示模型適合度的 Cox 和 Snell、Nagelkerke 以及 McFadden *R* 平方測量。這些統計資料在某些方面與線性迴歸中的 *R* 平方統計資料類似。

單調性測量（選用）。顯示資料中一致成對、不一致成對和約束成對的號碼，以及每類佔總成對數的百分比。Somers′ D、Goodman 與 Kruskal's Gamma、Kendall's tau-a 以及和諧指數 C 也會呈現在本表。

資訊準則（選用）。顯示 Akaike 的資訊準則（AIC）和 Schwarz 的貝葉斯資訊準則（BIC）。

概似比率測試（選用）。顯示模型效應係數是否在統計上不等於 0 的統計量測試。有意義的輸入欄位是輸出中顯著性層次很低（標註為 *Sig.*）的輸入欄位。.

參數估計值（選用）。顯示方程式係數的估計值、這些係數的測試、衍生自標註為 *Exp(B)* 的係數的幾率比及其置信區間。

漸近共變異數/相關性矩陣（選用）。顯示係數預估值的漸近共變異數和/或相關性。

觀測頻率和預測頻率（選用）。對於每個共變數型樣，顯示每個輸出欄位值的觀測頻率和預測頻率。此表格可能很大，對於具有數值輸入欄位的模型來說尤其如此。如果結果表格過大而無法應用，那麼將省略該表格，並顯示一條警告。

PCA/因子節點

PCA/因素節點提供強大的資料減少技術來減少資料的複雜性。該技術提供以下兩種相似但不同的方法。

- **主成分分析 (PCA)** 可以找出輸入欄位的線性組合，這些組合能夠出色地擷取整個欄位集合中的變異數，且組合中的各個成分相互正交（相互垂直）。主成分分析集中關注所有變異數，包含共用變異數和獨有變異數。
- **因數分析**試圖確定可以解釋一組觀測欄位中的相關性型樣的基本概念（即**因子**）。因數分析只集中關注共用變異數。估計模型時不考慮特定欄位獨有的變異數。因子/主成分分析節點提供幾種因數分析方法。

這兩種方式的目標都是找到有效概括原始欄位集中的資訊的少量衍生欄位。

需求。 主成分分析因子模型中只能使用數值型欄位。要估計因數分析或主成分分析，需要一個或多個角色設定為輸入欄位的欄位。角色設定為目標、兩者或無的欄位將被忽略，就像對待非數值型欄位一樣。

強度。 因數分析和 PCA 可以在不犧牲太多資訊內容的情況下有效地降低資料的複雜性。這些技術可協助您建立更穩健的模型，並實現比原始輸入欄位更高的執行速度。

PCA/因子節點模型選項

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。

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。

擷取方法。指定要用於資料縮減的方法。

- **主成分。**這是預設方法，它將使用 PCA 來尋找對輸入欄位進行彙總的成分。
- **未加權最小平方。**此因數分析方法的工作原理是找出最能重新產生輸入欄位之間的關係（相關性）型樣的因子集合。
- **通用性最小二乘法。**此因數分析方法與未加權最小平方類似，區別在於它利用加權降低具有大量獨有（非共用）變異數的欄位的重要程度。
- **最大概似法。**根據關於這些關係的形式的假設情況，此因數分析方法將產生最有可能生成輸入欄位中觀測到的關係（相關性）型樣的因子方程式。 特別是，該方法假定訓練資料服從多元常態分佈。
- **主軸分解。**此因數分析方法與主成分法十分類似，區別在於它僅側重於共用變異數。
- **Alpha分解。**此因數分析方法將分析中的欄位視為潛在輸入欄位範圍內的樣本。它會將因子的統計可靠性最大化。
- **映像因子法。**此因數分析方法使用資料估計來隔離共用變異，並尋找說明該變異數的因子。

主成份分析（PCA）/因子節點專家選項

如果您對因數分析和 PCA 有詳細瞭解，那麼可以通過專家選項對訓練過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

遺漏值。 依預設，IBM SPSS Modeler 只會將具有有效值的記錄用於模型中所使用的所有欄位。（這有時稱為遺漏值的**整批 (Listwise) 刪除**。）如果您有很多遺漏資料，則可能會發現此方法會刪除太多記錄，從而留給您的資料不足以產生良好的模型。在此類情況下，您可以取消選取**僅使用完整記錄**選項。 IBM SPSS Modeler 隨後會嘗試使用盡可能多的資訊來估計模型，其中包括部分欄位具有遺漏值的記錄。（有時候，這稱為**成對刪除**遺漏值。）但是，在一些情況下，以這種方式使用不完整記錄可能會導致評估模型時發生計算問題。

欄位。 指定估計模型時是使用輸入欄位的相關性矩陣（預設設置）還是使用其共變異數矩陣。

聚合的疊代數上限。 指定用來估計模型的疊代數目上限。

擷取因子。選取要從輸入欄位中擷取的因子數目的方法有兩種。

- **大於以下值的固有值。**此選項將保留固有值大於指定準則的所有因子或成分。**固有值**用於測量每個因子或成分對輸入欄位集合中的變異數進行彙總的能力。使用相關係數矩陣時，模型將保留固有值大於指定值的所有因子或成分。使用共變數矩陣時，準則是指定的乘以平均固有值。該尺度變換使此選項對於兩種類型的矩陣具有類似的意義。
- **最大數目。**此選項將保留指定號碼的因子或成分，按固有值的遞減排列。換言之，將保留 *n* 個最高固有值所對應的因子或成分，其中 *n* 為指定準則。預設擷取準則為五個因子/成分。

成分/因子矩陣格式。這些選項用於控制因子矩陣（或 PCA 模型的成分矩陣）的格式。

- **對值進行排序。**如果已選取此選項，那麼將按數值順序對模型輸出中的因子載入進行排序。
- **隱藏值低於以下方值的值。**如果已選取此選項，那麼將在矩陣中隱藏小於指定臨界值的分數，以便於請參閱矩陣中的型樣。

旋轉。通過這些選項，您可以控制模型的旋轉方法。請參閱『主成分分析（PCA）/因子節點旋轉選項』主題，以取得更多資訊。

主成分分析（PCA）/因子節點旋轉選項

多數情況下，對保留的因子集合進行數學旋轉可增加其實用性，尤其可以降低其解釋難度。選取一種旋轉方法：

- **沒有旋轉。**預設選項。不使用旋轉。
- **上限方差。**這是可以將每個因子上負荷較高的欄位的號碼降至最低的正交旋轉法。它簡化了因子的解譯過程。
- **直接斜交旋轉。**斜交（非正交）旋轉法。當 **Delta** 等於 0（預設值）時，解將採用斜交法。若 delta 越趨近負數，則因素就越不會趨向斜交。要置換預設的 Delta 值 0，請輸入小於或等於 0.8 的數字。
- **上限四次方值。**這是可以將解釋每個欄位所需的因子的數量降至最低的正交旋轉法。它簡化了被觀測欄位的解譯過程。
- **上限平衡值。**此旋轉法結合了 Varimax 法與 Quartimax 法，前者用於簡化因子，後者用於簡化欄位。可將某個因子上載荷較高的欄位數量和解釋某個欄位所需的因子數目量降至最低。
- **最優斜交。**這是實現了因子關聯的斜交旋轉法。它計算起來比斜交旋轉更快，因此適用於大型資料集。**卡帕(Kappa) 統計量數** 用於控制項解的傾斜度（因子相關的程度）。

主成分/因子模型塊

主成分/因子模型塊代表由 PCA/因子節點建立的因子分析和主成份分析 (PCA) 模型。其中包含被訓練模型擷取的所有資訊，以及有關模型效能和特性的資訊。

當您執行包含因子方程式模型的串流時，節點會為模型中的每個因子或成分新增一個新欄位。新的欄位名稱衍生自模型名稱並帶有字首和字尾，字首為 *$F-*，而字尾為 *-n*，其中 *n* 是因子或成分的編號。例如，如果模型名為 *Factor* 且包含三個因子，新欄位將命名為 *$F-Factor-1*、 *$F-Factor-2* 和 *$F-Factor-3*。

為更好地瞭解因子模型的編碼內容，可以進一步執行一些下游分析。檢視因子模型結果的一種實用方法是使用統計資料節點檢視因子與輸入欄位之間的相關性。這種方法可顯示哪些輸入欄位對哪些因子的載荷較重，並協助您探索因子是否具有潛在的意義或解譯。

您還可以使用進階輸出中提供的資訊對因子模型進行評估。若要檢視進階輸出，請按一下模型區塊瀏覽器的**高階**標籤。進階輸出包含大量詳細資訊，適合於在因數分析或主成分分析方面具有廣泛知識的使用者。請參閱『主成分/因子模型塊進階輸出』主題，以取得更多資訊。

主成分/因子模型塊方程式

因子模型快的「模型」標籤顯示每個因子的因子分數方程式。因子或成分的分數是通過將每個輸入欄位值相乘以其係數並將結果相加計算得出的。

主成分/因子模型塊概要

因子模型的「概要」標籤顯示因子/主成分分析模型中保留的因子數目，以及有關用於產生模型的欄位和設定的其他資訊。請參閱第 36 頁的『瀏覽模型塊』主題，以取得更多資訊。

主成分/因子模型塊進階輸出

因數分析的進階輸出提供有關所估計模型及其效能的詳細資訊。進階輸出中包含的大部分資訊技術含量很高，需要具備因數分析方面的廣泛知識才能正確理解該輸出。

警告。 指出結果的任何警告或潛在問題。

共同性。顯示因子或成分佔每個欄位的變異數的比例。初始指定具有整個因子集合（最初，模型的因子數與輸入欄位數相同）的初始共同性，擷取指定基於保留因子集合的共同性。

解釋的總變異數。顯示由模型中的因子解釋的總變異數。初始固有值顯示由整個初始因子集合解釋的變異數。擷取平方和載入顯示由模型中保留的因子解釋的變異數。旋轉平方和載入顯示由旋轉因子解釋的變異數。請注意，對於斜交旋轉法， 旋轉載入平方和僅顯示載入平方和，而不顯示變異數百分比。

因子（或成分）矩陣。顯示輸入欄位與未旋轉因子之間的相關性。

旋轉因子（或成分）矩陣。顯示輸入欄位與正交旋轉的旋轉因子之間的相關性。

型樣矩陣。顯示輸入欄位與斜交旋轉法的旋轉因子之間的偏相關。

結構矩陣。顯示輸入欄位與斜交旋轉法的旋轉因子之間的簡單相關性。

因子相關性矩陣。顯示斜交旋轉法的因子之間的相關性。

區別節點

區別元件分析會為群組成員資格建置預測模型。此模型由區別函數組成，以預測變數的線性組合為基礎 (如果有兩個群組以上，將產生一組區別函數)，而該預測變數，必須能提供群組之間最佳判斷依據。上述函數從觀察值樣本中產生，而且觀察值的組別成員是已知的，這些函數隨後便可以套用到新的觀察值上，新觀察值會有預測變數的測量，但是，組別成員是未知的。

範例。 根據使用情況資料，電信公司可以使用判別分析來對用戶進行群組。此操作使電信公司可以對潛在的用戶進行評分，並將最有可能的最有價值的群組的客戶作為目標。

需求。 您需要一個或多個輸入欄位，且只需要一個目標欄位。目標必須為帶有字串或整數儲存的種類欄位（測量層次為旗標或列名）。（必要的話，可使用「填充值」或「衍生」節點來轉換儲存體。）會忽略設為兩者 或無 的欄位。模型中所用的欄位必須已完全實例化其類型。

強度。 判別分析和邏輯迴歸都是合適的分類模型。但是，判別分析會對輸入欄位進行更多的假設，例如，假設這些欄位為正常分佈且連續，當滿足這些要求時它們能提供更好的結果，在樣本量比較小時尤其如此。

判別節點模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

方法。下列選項可用於將預測值輸入模型：

- **輸入。** 這是預設方法，其直接在方程式中輸入所有項。不能顯著增加模型預測能力的項目將不被新增。
- **逐步。** 起始模型可能是最簡單的模型，其方程式中沒有任何模型項（除了常數）。在每一個步驟中，會對尚未新增至模型的項目進行評估，並且如果這些項目的最適性會顯著性地新增至模型的預測能力，則會予以新增。

註：逐步方法特別容易過度擬合訓練資料。使用這些方法時，請務必使用留出法測試範例或新資料來驗證所產生模型的有效性。

判別節點專家選項

如果對判別分析有詳盡瞭解，可用專家選項調整訓練過程。若要存取專家選項，請在「專家」標籤上將**模式**設為**專家**。

事前機率。 此選項可判定是否針對群組成員資格的事前瞭解調整分類係數。

- **所有群組大小均等。** 假設所有群組事前機率均等；對係數沒有任何效果。
- **依據群組大小計算。** 樣本中的觀察群組大小可判定群組成員資格的事前機率。例如，如果分析中有 50% 的觀察值介於第一個群組內，25% 介於第二個群組內，而另外 25% 介於第三個群組內，則分類係數將調整為增加第一組中的成員資格與其他兩組的相關可能性。

使用共變異數矩陣。 您可以選擇使用組內共變異數矩陣、或各組共變異數矩陣，來將觀察值分類。

- 在組別內。合併的群組內共變異數矩陣用於分類觀察值。
- 各組。各組共變異數矩陣在分類時使用。因為分類是以區別函數為依據（而不是以原始變數為依據），因此此選項並不是一律相當於二次區別。

輸出。這些選項可讓您要求將顯示在節點所建置之模型區塊進階輸出中的其他統計資料。請參閱『判別節點輸出選項』主題，以取得更多資訊。

逐步。 這些選項可讓您使用「逐步」估計方法來控制新增及移除欄位的準則。（如果選取了「輸入」方法，則會停用該按鈕。）請參閱第 167 頁的『判別節點執行步驟選項』主題，以取得更多資訊。

判別節點輸出選項

選取要在邏輯迴歸模型塊的進階輸出中顯示的選用輸出。若要檢視進階輸出，請瀏覽模型片段並按一下**進階**標籤。請參閱第 168 頁的『判別分析模型塊進階輸出』主題，以取得更多資訊。

描述性統計量。 可用的選項包括：平均數 (包括標準差)、單變量 ANOVA、Box's *M* 測試。

- 平均數. 顯示總和與群組平均數，以及自變數的標準差。
- 單變量 *ANOVA(A)*. 針對每個自變數的群組平均數之相等性執行單向變異數分析測試。
- *Box's M*。對群組共變異數矩陣相等性進行的測試。若樣本夠大，則非顯著 p 值代表沒有充足證據顯示矩陣有所不同。此測試對多變量常態的偏差很敏感。

函數係數。 可用的選項包括：Fisher 分類係數、未標準化係數。

- *Fisher's* 線性區別函數係數*(F)*. 顯示可直接供分類使用的 Fisher 分類函數係數。將會針對每個群組取得個別的分類函數係數集，並將觀察值指派給擁有最高區別分數（分類函數值）的群組。
- 未標準化*(U)*. 顯示未標準化的區別函數係數。

矩陣。 可用的自變數係數矩陣包括：組內相關性矩陣、組內共變異數矩陣、各組共變異數矩陣、總和共變異數矩陣。

- 組別內相關。顯示合併的群組內相關性矩陣，此矩陣是將所有群組的個別共變異數矩陣平均，再計算相關性而來。
- 組別內共變異。顯示合併的群組內共變異數矩陣，此矩陣可能與共變異數矩陣總計不同。透過將所有群組的個別共變異數矩陣平均來取得矩陣。
- 各組共變異。顯示各組不同的共變異數矩陣。
- 共變異總計。顯示所有觀察值中的共變異數矩陣，如同它們來自單一樣本一樣。

分類。 下列輸出與分類結果有關。

- 逐觀察值結果。顯示每個觀察值之實際群組、預測群組、事後機率及區別分數的代碼。
- 摘要表格。根據區別分析，正確及錯誤地指派給每個群組的觀察值數目。有時亦稱為「混淆矩陣」。
- 留一分類。分析中的每個觀察值皆由該觀察值除外之所有觀察值的衍生函數來分類。其亦稱作「U 方法」。
- 地域圖。以函數值為基礎，用於將觀察值分類至群組的邊界圖。將觀察值分類至對應之組別的數目。每個組別的平均數是用其邊界中的星號來表示。如果只有一個區別函數，不會顯示地圖。
- 組合組別。建立前兩個區別函數值的所有群組散佈圖。如果僅有一個函數，則會改為顯示直方圖。
- 各組。建立前兩個區別函數值的各組散佈圖。在只有一個函數的情況下，則會改為顯示直方圖。

逐步。 **步驟摘要**顯示了執行每個步驟後所有變數的統計資料；**配對距離的 F 值**顯示了每個群組對的配對 *F* 比矩陣。*F* 比可用於群組之間 Mahalanobis 距離的顯著性測試。

判別節點執行步驟選項

方法。選取統計量，以輸入或刪除新變數。您可以使用的項目包括：Wilks′ Lambda (λ) 值、無法解釋的變異數、馬氏 (Mahalanobis) 距離、最小 *F* 比例和 Rao′s *V* 量數。利用 Rao′s *V* 量數，您可以指定要輸入的變數 *V* 最小增量。

- *Wilks′ lambda*。一種進行逐步迴歸分析區別分析的功能選擇方法，它會根據變數將 Wilks Lambda 降低的程度來為輸入方程式的項目選擇變數。每一個步驟都要輸入將整體 Wilks Lambda 最小化的變數。
- 未說明的變異數。在每個步驟中，輸入會將群組之間未說明的變異總和最小化的變數。
- *Mahalanobis* 距離。測量自變數之觀察值的值與整體觀察值平均數的變異程度。較大 Mahalanobis 距離會將觀察值識別為在一個以上自變數中具有極端值。
- 最小 *F* 比例。一種逐步分析的功能選擇方法，其所根據的作法，是將從群組間 Mahalanobis 距離計算而來的 F 比例最大化。
- *Rao's V*。群組平均數之間差異的量數。也稱為 Lawley-Hotelling 跡。每一個步驟，都要輸入將 Rao′s V 中的增加最大化的變數。選取此選項後，輸入變數必須輸入分析的最小值。

條件。 可用的替代方案包括：**使用 F 值**和**使用 F 機率**。輸入值表示輸入和移除變數。

- 使用 *F* 值。變數的 F 值大於「輸入」值時，系統會將該變數輸入模型，而當 F 值小於「移除」值時，系統會將變數移除。「輸入」必須大於「移除」，而且兩個數值都必須是正數。若要將更多變數輸入模式，請調低「輸入」值。若要從模型中移除更多變數，請調高「移除」值。

- 使用 *F* 機率。變數的 F 值顯著性層次小於「輸入」值時，系統會將變數輸入模型，而顯著性層次大於「移除」值時，系統會將變數移除。「輸入」必須小於「移除」，而且兩個數值都必須是正數。若要將更多變數輸入模式，請調高「輸入」值。若要從模型中移除更多變數，請調低「移除」值。

判別分析模型塊

判別分析模型塊代表由判別節點估計的方程式。這些方程式包含由判別分析模型所擷取的所有資訊及有關模型結構和效能的資訊。

當執行包含判別分析模型塊的串流時，該節點可新增包含模型預測和相關機率的兩個新欄位。新欄位的名稱衍生自要預測的輸出欄位的名稱，並帶有表示預測種類的字首 *$D-* 或表示相關機率的字首 *$DP-*。例如，對於名稱為 *colorpref* 的輸出欄位，新欄位的名稱應是 *$D-colorpref* 和 *$DP-colorpref*。

產生過濾器節點。 「產生」功能表可讓您建立一個新的「過濾器」節點，以根據模型的結果來傳遞輸入欄位。

預測值重要性

選擇性地，指出評估模型時每個預測值的相對重要性的圖表，可能也會顯示在「模型」標籤上。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。注意，只有在產生模型之前已選取「分析」標籤上的**計算預測值重要性**，才可以使用此圖表。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

判別分析模型塊進階輸出

判別分析的進階輸出給出了有關估計模型及其效能的詳細資訊。在進階輸出中包含的多數資訊具有很強的技術性，需要具有廣泛的判別分析方面的知識才能夠對此輸出作出正確地解釋。請參閱第 166 頁的『判別節點輸出選項』主題，以取得更多資訊。

判別分析模型塊設定

通過判別分析模型塊中的「設定」標籤，您可以在對模型進行評分時取得傾向評分。此標籤適用於只包含旗標目標的模型，且只有在模型區塊已新增至串流之後才可用。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

判別分析模型塊彙總

判別分析模型塊的「摘要」標籤顯示了用於產生模型的欄位和設定。此外，如果您已執行連接至此建模節點的分析節點，則也會在此小節中顯示來自該分析的資訊。如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』。

GenLin 節點

廣義線性模型會延伸一般線性模型，因此應變數可透過指定的鏈結函數與因數和共變數成線性相關。此外，此模式允許」變數具有非常態分配。它包含廣泛使用的統計模型，例如一般分佈回應的線性迴歸、二進位資料的邏輯模型、計數資料的對數線性模型、區間受限存活資料的互補對數存活函數的對數模型，以及透過其極其常見的模型規劃的許多其他統計模型。

範例。運輸公司可以使用通用性線性模型對不同期間建造的多種類型的船隻的損壞計數應用 Poisson 迴歸方法，生成的模型可說明確定哪些船隻類型最容易損壞。

車輛保險公司可以使用通用性線性模型對車輛的損壞索賠應用 Gamma 迴歸，生成的模型可說明確定對索賠金額影響最大的因素。

醫學研究者可以使用通用性線性模型對區間型刪失的生存分析資料應用互補對數存活函數的對數迴歸方法，從而預測病理狀況再發生的時間。

通用性線性模型的工作原理是建立一個方程式，從而使輸入欄位值與輸出欄位值關聯起來。一旦產生了模型，就可以用來估計新資料的值。對於每一筆記錄，會為每個可能的輸出種類計算成員資格的機率。會將具有最高機率的目標種類指派為該記錄的預測輸出值。

需求。 您需要一個或多個輸入欄位，同時有且僅有一個具有兩個或多個種類的目標欄位（其測量層次可以為連續或旗標）。模型中所用的欄位必須已完全實例化其類型。

強度。 廣義線性模型極為靈活，但選擇模型結構的過程並未自動化，因此您需要對資料有一定的瞭解（這在「黑盒」演算法中是不需要的）。

GenLin 節點欄位選項

除建模節點的「欄位」標籤通常提供的目標、輸入和分割區自訂選項外（請參閱第 26 頁的『建模節點欄位選項』），GenLin 節點還提供下列附加功能。

使用加權欄位。 尺度參數是與回應的變異數有關的估計模型參數。尺度加權是可能依據不同觀察值而有所不同的已知值。若指定尺度加權變數，則會針對每個觀察值將尺度參數 (與回應的變異數有關) 除以尺度加權變數。分析中不會使用尺度加權值少於等於 0 或遺漏的記錄。

目標欄位代表一組試驗中發生的事件數目。如果回應是一組試驗中發生的事件數目，那麼目標欄位將包含該事件數目，並且您可選取包含試驗次數的附加變數。或者，若所有受試者的試驗數都相同，則或許可使用固定值來指定試驗。試驗數應大於或等於每個記錄的事件數。事件必須是非負值的整數，而試驗必須是正整數。

GenLin 節點模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

模型類型 有兩個選項用於要構造的模型類型。**僅主效應** 使模型僅分別包含各個輸入欄位，而不測試輸入欄位之間的互動作用（相乘性作用）。**主效應和所有雙向互動**包含所有雙向互動以及輸入欄位主效應。

偏移。 偏移項是「結構」預測值。它的係數無法由模式估計，但系統假設其具有數值 1；因此偏移的呼叫會新增至目標的線性預測值。這在卜瓦松 (Poisson) 迴歸模型中特別有用，因為每個觀察值暴露於所需事件的層級可能不同。

例如，在建立個別駕駛員的意外事件比率模式時，具有三年經驗並曾於某一意外事件中出差錯的駕駛員，與具有 25 年經驗並曾於某一意外事件中出差錯的駕駛員間有重要的差異。如果駕駛員的經驗是以偏移項的方式納入，則可將意外事件數量模型化為含有對數鏈結的 Poisson 回應值或負二項式回應。

其他分配組合和鏈結類型可能需要偏移變數的其他轉換。

註：如果使用了變數位移欄位，那麼不應同時將指定欄位用作輸入。如果需要，可在上游來源節點或「類型」節點中將位移欄位的角色設定為**無**。

旗標目標的基本種類。

對於二進位回應，您可以選擇應變數的參照種類。這會影響特定輸出，如參數估計值與儲存值，但不會變更模式適合度。例如，如果二進位回應採用值 0 與 1：

- 依預設，程序會使最後一個（最高值）種類或 1 作為參照種類。在這種情況下，模式儲存的機率會預估給定觀察值採取數值 0 的機會，而參數預估應該解譯為與種類 0 的概似相關。
- 如果您指定第一個（最低值）種類或 0 作為參照種類，則模型儲存的機率會預估給定觀察值採用值 1 的機會。
- 如果您指定自訂種類，且變數已定義標籤，則可以透過從清單中選擇值來設定參照種類。當您在指定模型的過程中，不記得特定變數的確切編碼方式時，這會是相當方便的功能。

模式中包括截距。 模式中通常會包括截距，但是如果假設資料會穿過原點的話，就可以將截距排除在外。

GenLin 節點專家選項

如果具備通用性線性模型的深入知識，那麼可以使用專家選項對訓練過程進行微調。若要存取專家選項，請在「專家」標籤上將**模式**設為**專家**。

目標欄位分佈與鏈結函數

分佈。

本節說明因變數的分配。指定非常態分配與非識別鏈結函數的能力，對於在一般線性模型改善廣義線性模型而言是必備的。可能的分配鏈結函數組合有很多，且其中有好幾個都適用於指定的任何資料集，因此您的選擇可遵循先期提出的理論考量，或看起來最適合的組合。

- **二項式。** 此分配唯有變數代表二元回應或事件個數時才合適。
- **伽瑪參數。** 此分配適用於具有正值尺度的變數且偏向較大正數值的變數。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。
- **反向高斯 (Gaussian)。** 此分配適用於具有正值尺度的變數且偏向較大正數值的變數。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。
- **負二項式。** 此分配可視為觀察 *k* 成功所需的試驗次數，且適用於具有非負整數值的變數。若資料值非整數、小於零或遺漏，則不會在分析中使用對應觀察值。負二項式分佈的輔助參數的固定值可以為大於或等於 0 的任意數字。輔助參數設為 0 時，使用此分配等同於使用卜瓦松 (Poisson) 分配。
- **常態。** 此分配適用於值呈對稱、約於中央 (平均數) 值呈鐘型分佈的尺度變數。因變數必須為數值。
- **Poisson。** 此分配可視為在固定時間內所需事件的發生次數，且適用於具有非負整數值的變數。若資料值非整數、小於零或遺漏，則不會在分析中使用對應觀察值。

- **Tweedie。** 此分配適用於可以伽瑪分配的 Poisson 混合表示的變數；此分配「混合」的意思是說，其結合了連續（如非負實值）與離散分配（單一值上的正機率量，0）的特性。因變數必須為數值，且資料值大於或等於零。若資料值小於零或遺漏，則不會在分析中使用對應觀察值。Tweedie 分配的固定值可以是任何大於 1 且小於 2 的數字。
- **多項式。** 此分配適用於表示序數回應值的變數。因變數可以是數值或字串，且必須至少具備兩個相異的有效資料值。

鏈結函數。

鏈結函數是允許模式估計的因變數轉換。您可以使用的函數如下：

- **單位。** $f(x)=x$。因變數不會進行轉換。此鏈結可以和任何分配一起使用。
- **互補對數存活函數的對數。** $f(x)=\log(-\log(1-x))$。這僅適用於二項式分配。
- **累積 Cauchit。** $f(x) = \tan(\pi\ (x - 0.5))$，套用至每一種回應的累積機率。這僅適用於多項式分配。
- **累積互補對數存活函數的對數。** $f(x)=\ln(-\ln(1-x))$，套用至每一種回應的累積機率。這僅適用於多項式分配。
- **累積 Logit。** $f(x)=\ln(x\ /\ (1-x))$，套用至每一種回應的累積機率。這僅適用於多項式分配。
- **累積負對數存活函數的對數。** $f(x)=-\ln(-\ln(x))$，套用至每一種回應的累積機率。這僅適用於多項式分配。
- **累積機率。** $f(x)=\Phi^{-1}(x)$，套用到每一種回應的累積機率，其中 Φ^{-1} 是反向標準正態累積分佈函數。這僅適用於多項式分配。
- **對數。** $f(x)=\log(x)$。此鏈結可以和任何分配一起使用。
- **對數互補。** $f(x)=\log(1-x)$。這僅適用於二項式分配。
- **Logit 分析。** $f(x)=\log(x\ /\ (1-x))$。這僅適用於二項式分配。
- **負二項式。** $f(x)=\log(x\ /\ (x+k^{-1}))$，其中 k 是負二項式分佈的輔助參數。這僅適用於負值二項式分配。
- **負對數存活函數的對數。** $f(x)=-\log(-\log(x))$。這僅適用於二項式分配。
- **勝算冪次。** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, if $\alpha \neq 0$。$f(x)=\log(x)$，如果 $\alpha=0$。α 是必要的數值規格且必須為實數。這僅適用於二項式分配。
- **機率值。** $f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是反向標準正態累積分佈函數。這僅適用於二項式分配。
- **次方。** $f(x)=x^{\alpha}$, if $\alpha \neq 0$。$f(x)=\log(x)$，如果 $\alpha=0$。 α 是必要的數值規格且必須是實數。此鏈結可以和任何分配一起使用。

參數。 通過此群組中的控制項，可以在選中某些分佈選項時指定參數值。

- **用於負二項的參數。**對於負二項式分佈，選擇以指定一個值或容許系統提供估計值。
- **用於 Tweedie 的參數。**對於 Tweedie 分佈，給固定值指定在 1.0 與 2.0 之間的一個數字。

  參數估計。 這個群組中的控制項可用來指定估計方法，並為參數估計提供初始值。

  – **方法。** 您可以選取參數預估方法。在 Newton-Raphson、Fisher 分數或混合方法（會在切換至 Newton-Raphson 方法前，先執行 Fisher 分數疊代）之間選擇。如果在混合方法的費雪 (Fisher) 評分階段期間，尚未到達 Fisher 疊代的最大數量就已達到收斂，則演算法會繼續進行 Newton-Raphson 方法。

  – **尺度參數方法。**您可以選取尺度參數估計方法。最大概似會以模型效果共同預估尺度參數；請注意，如果回應具有負二項式、Poisson 或二項式分配，則此選項無效。離差和 Pearson 卡方選項會從那些統計量的值預估尺度參數。或者，您可以為尺度參數指定固定值。
- **共變異數矩陣。** 以模式為基礎的估計值是赫氏 (Hessian) 矩陣通用性反向的負數。穩健 (也稱為 Huber/White/sandwich) 估計值是「修正」過後以模式為基礎的估計值，可提供一致的共變異數估計值，即使變異數規格和鏈結函數不正確時也是如此。

疊代。 這些選項可讓您控制模型聚合的參數。請參閱『通用性線性模型疊代』主題，以取得更多資訊。

輸出。這些選項可讓您要求將顯示在節點所建置之模型區塊進階輸出中的其他統計資料。請參閱『通用性線性模型進階輸出』主題，以取得更多資訊。

奇異性容忍值。 奇異（或不可翻轉的）矩陣具有線性相關直欄，這可能導致預估演算法發生嚴重問題。即使接近奇異的矩陣也可能導致較差的結果，因此程序會將行列式小於容錯的矩陣視為奇異。指定一個正值。

通用性線性模型疊代

您可設定用於對通用性線性模型進行估計的收斂參數。

疊代。 您可以使用的選項如下：

- **最大疊代。** 將執行運算的疊代最大值。指定一個非負整數。
- **最大的半階次數。** 每次疊代時，步驟大小會因乘以因素 0.5 而減少，直到對數概似增加或到達最大半階次數。指定一個正整數。
- **檢查資料點的分隔。**選取的話，演算法會執行測試，以確定參數估計值具有唯一值。當程序能夠產生可正確分組各觀察值的模式，就會啟動分組。可以從二進位格式的二項式回應取得此選項。

收斂準則。 下列選項可用

- **參數收斂條件。** 選取的話，演算法會在參數估計值之絕對或相對變更小於所指定數值的疊代之後停止，該數值必須為正數。
- **對數概似收斂。** 選取的話，演算法會在對數概似函數之絕對或相對變更小於所指定數值的疊代之後停止，該數值必須為正數。
- **Hessian 收斂。** 針對「絕對」指定，如果基於 Hessian 收斂的統計量小於指定的正數值，則假設收斂。針對「相對」指定，如果統計量小於指定的正數值與對數概似絕對值的product，則假設收斂。

通用性線性模型進階輸出

選取要在廣義線性模型塊的進階輸出中顯示的選用輸出。若要檢視進階輸出，請瀏覽模型片段並按一下**進階**標籤。請參閱第 173 頁的『GenLin 模型塊進階輸出』主題，以取得更多資訊。

下列是可用的輸出：

- **觀察值處理摘要。** 顯示分析和「相關資料摘要」表格內含和排除的觀察值個數與百分比。
- **描述性統計量。** 顯示因變數、共變異數和因素的描述性統計量和摘要資訊。
- **模式資訊。** 顯示資料集名稱、因變數或事件和試驗變數、偏移變數、尺度加權變數、機率分配和鏈結函數。
- **適合度統計量。** 顯示離差和調整後的離差、皮爾森 (Pearson) 卡方和調整後的皮爾森 (Pearson) 卡方、對數概似、Akaike 資訊準則 (AIC)、最終樣本修正 AIC (AICC)、Bayesian 訊息準則 (BIC) 和一致的 AIC (CAIC)。
- **模型摘要統計量。** 顯示模式適合度的測試，包括模式適合度綜合測試的概似比統計量，以及每個效應項的類型 I 或 III 對比的統計量。
- **參數估計值。** 顯示參數估計值和對應的測試統計量與信賴區間。除了原始參數估計值以外，您也可以選擇顯示指數化參數估計值。
- **參數估計值的共變異數矩陣。** 顯示估計的參數共變異數矩陣。
- **參數估計值的相關性矩陣。** 顯示估計的參數相關性矩陣。
- **對比係數 (L) 矩陣。** 顯示預設效應項和估計的邊際平均數 (若在「EM 平均數」標籤中要求) 的對比係數。
- **一般可估計函數。** 顯示產生對比係數 (L) 矩陣的矩陣。

- **疊代歷程。** 顯示參數估計值和對數概似的疊代歷程，並列印梯度向量和赫氏 (Hessian) 矩陣的最後評估。疊代歷程表格每 n^{th} 個疊代會顯示一次參數估計值 (從第 0^{th} 個疊代 (初始估計值) 啟動)，其中 n 是列印間隔值。若要求疊代歷程，則永遠會顯示最後一個疊代，無論 n 的值為何。
- **拉氏 (Lagrange) 乘數測試。** 顯示 Lagrange 乘數測試統計資料，用於為正常、Gamma 和逆模型高斯分佈評估使用離差或 Pearson 卡方計算或者設定為固定值的尺度參數的有效性。對於負值二項式分配，此測試會固定輔助參數。

模型效應。 您可以使用的選項如下：

- **分析類型。** 指定要產生的分析類型。當模式中的訂購預測值是您的首要原因時，通常適用類型 I 分析，而類型 III 則是較普遍適用的。Wald 或概似比統計量是依據在「卡方統計量」群組中的選擇來計算。
- **信賴區間。** 指定大於 50 或小於 100 的信賴等級。Wald 區間的基礎是假設參數有標準常態分配； 組合概似區間更為精確，但需要進行大量計算。組合概似區間的容差層級是一項準則，用來停止用於計算區間的疊代演算法。
- **對數概似函數。** 這會控制對數概似函數的顯示格式。完整函數包含一個額外的項目，是與參數估計值有關的常數；其對參數估計沒有影響，因而在某些軟體產品中不會顯示。

GenLin 模型塊

GenLin 模型塊代表由 GenLin 節點估計的方程式。它們包含模型擷取的全部資訊，以及模型結構與效能的相關資訊。

當您執行包含 GenLin 模型塊的串流時，該節點會新增一些新欄位，這些欄位的內容取決於目標欄位的本質：

- **旗標目標。**新增包含預測種類和相關機率的欄位，並為每個種類新增機率。前兩個新欄位的名稱衍生自要預測的輸出欄位的名稱，並帶有表示預測種類的字首 *$G-* 或表示相關機率的字首 *$GP-*。例如，對於名為 *default* 的輸出欄位，新欄位將命名為 *$G-default* 和 *$GP-default*。後兩個附加欄位的名稱基於輸出欄位的值，並帶有字首 *$GP-*。例如，如果 *default* 的有效值為 *Yes* 和 *No*，那麼新欄位會以 *$GP-Yes* 和 *$GP-No* 命名。
- **連續目標。**新增包含預測平均值以及標準錯誤的欄位。
- **連續目標，代表一系列試驗中發生的事件數目。**新增包含預測平均值以及標準錯誤的欄位。
- **序數目標。**為依序集合的每個值新增包含預測種類和相關機率的欄位。欄位的名稱衍生自要預測的依序集合的值，並帶有表示預測種類的字首 *$G-* 或表示相關機率的字首 *$GP-*。

產生過濾器節點。 「產生」功能表可讓您建立一個新的「過濾器」節點，以根據模型的結果來傳遞輸入欄位。

預測值重要性

選擇性地，指出評估模型時每個預測值的相對重要性的圖表，可能也會顯示在「模型」標籤上。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。注意，只有在產生模型之前已選取「分析」標籤上的**計算預測值重要性**，才可以使用此圖表。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

GenLin 模型塊進階輸出

廣義線性模型的進階輸出可提供有關估計模型及其效能的詳細資訊。進階輸出中包含的大部分資訊的技術性含量都很高，需要進行此類分析所需的豐富知識才能夠對此輸出作出正確解釋。請參閱第 172 頁的『通用性線性模型進階輸出』主題，以取得更多資訊。

GenLin 模型塊設定

使用 GenLin 模型塊的「設定」標籤，您可以在對模型進行評分時取得傾向評分，並且也適用於對模型進行評分期間產生 SQL。 此標籤適用於只包含旗標目標的模型，且只有在模型區塊已新增至串流之後才可用。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

GenLin 模型塊彙總

GenLin 模型塊的「摘要」標籤顯示了用於產生模型的欄位和設定。此外，如果您已執行連接至此建模節點的分析節點，則也會在此小節中顯示來自該分析的資訊。如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』。

通用性線性混合模型

GLMM 節點

使用此節點可以建立概化線性混合模型 (GLMM)。

通用性線性混合模型

通用性線性混合模型會延伸線性模型，這麼一來：

- 目標即可透過指定的鏈結函數與因素和共變量成線性相關。
- 目標可具有非常態分佈。
- 觀察值可以具有相關性。

通用性線性混合模型涵蓋多種模型，從非常態縱向資料的簡單線性迴歸，到複雜的多層級模型。

範例。地區教育局可使用通用性線性混合模型以判斷實驗性教學方式是否能有效提升數學分數。相同教室的學生應該具有相關性，因為他們都是由同一位教師授課，且相同學校中的教室也具有相關性，所以我們可以在學校和教室層級包含隨機效應以說明變異性的不同來源。

醫療研究員可以使用通用性線性混合模型，以判斷新的抗癲癇藥物是否能減少病患癲癇發作的機率。相同病患的重複測量結果通常會是正相關，所以加入一些隨機效應的混合模型應較為合適。目標欄位和發作次數採用正整數值，因此使用卜瓦松 (Poisson) 分佈和對數鏈結的通用性線性混合模型可能較為合適。

電視、電話及網際網路服務之纜線提供者的高階監督者可以使用通用性線性混合模型，來進一步了解潛在客戶。因為可能的答案具有名義測量層級，所以公司分析師會透過使用隨機截距的概化 logit 混合模型，以擷取特定問卷回答者答案中跨服務類型（電視、電話、網際網路）之服務使用問題答案間的相關性。

「資料結構」標籤可以讓您在觀察值具有相關性時，指定資料集中記錄之間的結構關係。如果資料集中的記錄代表獨立的觀察值，則無需在此標籤中指定任何值。

受試者。 特定類別欄位的值組合應該唯一定義資料集內的受試者。例如，單一的病患 *ID* 欄位必須足以定義單一醫院中的受試者，但如果病患的 ID 號碼不是所有醫院中的唯一 ID，就可能會需要醫院 *ID* 和病患 *ID* 的組合。在重複測量設定中，會為每個受試者記錄多個觀察值，因此每個受試者可能會佔用資料集內的多個記錄。

受試者是一個觀察單位，並且可以視為與其他受試者無關。例如，某個醫療研究中的病患，其血壓讀數可以視為與其他病患的血壓讀數無關。當每個受試者有重複測量，並且您想要為這些觀察值之間的相關性建模時，定義受試者便變得非常重要。例如，您可能會預期某個病患在連續看病期間的血壓讀數是相關的。

在「資料結構」標籤上指定為**受試者**的所有欄位，都會用來定義殘差共變異結構的受試者，並提供可能的欄位清單來定義隨機效應區塊上隨機效應共變異結構的受試者。

重複測量。 在此指定的欄位會用來識別重複觀察值。例如，單一變數週可識別醫療研究中 10 週的觀察值，或可以合併使用月和天來識別一年中某個時期每天的觀察值。

定義共變異數群組依據。在此指定的類別欄位定義獨立的重複效應共變異數參數集；每一個欄位適用於一種由分組欄位交叉分類定義的類別。所有受試者都具有相同的共變異數類型；同一個共變異數分組中的受試者將具有相同的參數值。

空間共變異數座標。選取其中一個空間共變異數類型作為重複的共變異數類型時，此清單中的變數指定重複觀察值的座標。

重複共變異數類型。這可指定殘差的共變異數結構。可用的結構包括：

- 第一階自身迴歸 (AR1)
- 自身迴歸移動平均 (1,1) (ARMA11)
- 複合對稱
- 對角線
- 尺度單位
- 空間：冪次
- 空間：指數
- 空間：高斯
- 空間：線性
- 空間：線性對數
- 空間：球面
- Toeplitz
- 非結構化
- 變異數成分

目標： 這些設定透過鏈結函數，來定義目標、其分佈及其與預測值的關係。

目標。 目標是必要的。它可以具有任何測量層級，而目標的測量層級會限制哪些分佈和鏈結函數是適當的。

- **將試驗數當成分母。** 當目標回應值是在一組試驗中發生的事件數時，目標欄位則包含事件數，而您可以選取包含試驗數的其他欄位。例如，在試驗新的殺蟲劑時，您可能會讓螞蟻樣本暴露在不同濃度的殺蟲劑之下，然後記錄殺死的螞蟻數量及每個樣本的螞蟻數。在這種情況下，用來記錄殺死的螞蟻數量的欄位應指定為目標（事件）欄位，用來記錄每個樣本中螞蟻數量的欄位應指定為試驗欄位。如果每個樣本的螞蟻數量均相同，則可使用固定值來指定試驗數。

試驗數應大於或等於每個記錄的事件數。事件必須是非負值的整數，而試驗必須是正整數。

- **自訂參考類別。** 您可以針對類別目標選擇參考類別。這會影響特定輸出，如參數估計值，但不會變更模型適合度。例如，如果您的目標採用數值 0、1 和 2，則依預設，程序會使最後一個（最高值）類別或 2 成為參考類別。在此情況下，參數估計值應該解譯為與類別 0 或 1 的概似相關（相對於類別 2 的概似）。如果您指定自訂類別，且目標已定義標籤，則可以透過從清單中選擇值來設定參考類別。當您在指定模型的過程中，如果不記得特定欄位的確切編碼方式，這會是相當方便的功能。

目標分佈以及與線性模型的關係（鏈結）。 給定預測值之後，模型可預期目標值的分佈遵循指定形狀，而目標值要透過指定的鏈結函數與預測值呈線性關係。提供了數種常見模型的捷徑，如果您想要配合捷徑清單上沒有的特定分佈和鏈結函數組合，可以選擇**自訂**設定。

- **線性模型。** 以單位鏈結指定常態分佈，這在可以使用線性迴歸或 ANOVA 模型來預測目標時很有用。
- **伽瑪迴歸。** 以對數鏈結指定伽瑪分佈，這應在目標包含所有正數值並朝較大值偏斜時使用。
- **對數線性。** 以對數鏈結指定卜瓦松 (Poisson) 分佈，這應在目標代表固定時段中的出現次數時使用。
- **負二項式迴歸。** 以對數鏈結指定負二項式分佈，這應在目標和分母代表觀察第 *k* 次成功時所需之試驗數時使用。
- **多項式邏輯迴歸。** 指定多項式分佈，這應在目標為多類別回應時使用。多項式分佈會使用累積 logit 鏈結（序數結果）或概化 logit 鏈結（多類別名義回應）。
- **二元邏輯迴歸。** 以 logit 鏈結指定二項式分佈，這應在目標為邏輯迴歸模型所預測的二元回應時使用。
- **二元機率值。** 以機率值鏈結指定二項式分佈，這應在目標為具有基礎常態分佈的二元回應時使用。
- **區間受限存活。** 以互補對數存活函數的對數鏈結指定二項式分佈，當部分觀察沒有終止事件時，這在存活分析中是很有用的。

分佈 本選項指定目標的分佈。指定非常態分佈與非單位鏈結函數的能力，對於在線性混合模型基礎上改善通用性線性混合模型而言是必備項目。可能的分佈-鏈結函數組合有很多，且其中有好幾個都適用於給定的任何資料集，因此您的選擇可遵循先期提出的理論考量，或看起來最適合的組合。

二項式

此分佈僅適用於代表二元回應或事件數的目標。

Gamma

此分佈適用於具有正值尺度且偏向較大正數值的目標。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。

反向高斯

此分佈適用於具有正值尺度且偏向較大正數值的目標。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。

多項式

此分佈適用於代表多種類回應的目標。模型形式取決於目標的測量層級。

名義目標會導致產生名義多項式模型，其中會針對目標的每個類別（參考類別除外）估計一組個別的模型參數。給定預測值的參數估計值會顯示該預測值與目標之每個類別的概似之間的關係（相對於參考類別）。

序數目標會導致產生序數多項式模型，其中傳統截距項目將由一組**臨界值**參數取代，這些參數與目標類別的累積機率相關。

負二項式

負二項式迴歸搭配對數鏈結使用負二項式分佈，這應在目標代表具有高變異數的出現次數時使用。

常態 這適用於值呈對稱、約於中央（平均數）值呈鐘型分佈的連續目標。

Poisson

此分佈可視為在固定時段內相關事件的出現次數，且適用於具有非負整數值的變數。若資料值為非整數、小於零或遺漏，則不會在分析中使用對應觀察值。

鏈結函數

鏈結函數是允許模型估計的目標轉換。您可以使用的函數如下：

恆等式

$f(x)=x$。目標並未轉換。此鏈結可以和多項式以外的任何分佈搭配使用。

互補對數存活函數的對數

$f(x)=\log(-\log(1-x))$。這僅適用於二項式或多項式分佈。

Cauchit

$f(x) = \tan(\pi\ (x - 0.5))$. 這僅適用於二項式或多項式分佈。

對數 $f(x)=\log(x)$。此鏈結可以和多項式以外的任何分佈搭配使用。

對數互補

$f(x)=\log(1-x)$。這僅適用於二項式分配。

Logit $f(x)=\log(x\ /\ (1-x))$。這僅適用於二項式或多項式分佈。

負對數存活函數的對數

$f(x)=-\log(-\log(x))$。這僅適用於二項式或多項式分佈。

機率值

$f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是反向標準正態累積分佈函數。這僅適用於二項式或多項式分佈。

冪次 $f(x)=x^{\alpha}$，如果 $\alpha \neq 0$。$f(x)=\log(x)$，如果 $\alpha=0$。α 是必須指定的數值，且必須是實數。此鏈結可以和多項式以外的任何分佈搭配使用。

固定效應： 固定效應因素通常被認為是相關值都呈現於資料集中、並能用於評分的欄位。依預設，具有預先定義輸入角色（未在對話框的任何地方指定）的欄位，會輸入於模型的固定效應部分中。類別（旗標、名義和序數）欄位在模型中會被當成因素使用，而連續欄位則會當成共變數使用。

在來源清單中選取一或多個欄位並拖曳至效應清單，以便將效應輸入到模型中。所建立的效應類型視您將選項置於哪個熱點而定。

- **主要。** 已放置欄位會在效應清單底端，以個別的主效應呈現。
- **雙向。** 已放置欄位所有可能的組合（兩個一組）都在效應清單底端，以雙向互動呈現。
- **三向。** 已放置欄位所有可能的組合（三個一組）都在效應清單底端，以三向互動呈現。
- ***.** 所有已放置欄位的組合都在效應清單底端，以單一互動呈現。

「效應建置器」右側的按鈕可讓您執行各種動作。

表 *10.* 效應建置器按鈕說明

圖示	說明
	透過選取您想要刪除的項目，並按一下刪除按鈕，從固定效應模型中刪除項目。

表 *10.* 效應建置器按鈕說明 *(*繼續*)*

圖示	說明
	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，對固定效應模型內的項目進行重新排序。
	使用『新增自訂項目』對話框，將巢狀項目新增至模型，方法是按一下「新增自訂項目」按鈕。

包含截距。模型中通常會包含截距，但是如果可以假設資料會穿過原點，則可以將截距排除在外。

新增自訂項目： 您可以在這個程序中，為您的模式建立巢狀項目。通常巢狀項目在建立因素或共變量效果項的模式時非常有用，但因素或共變量的值不可以與其他因素層級互動。例如，連鎖雜貨店可能會追蹤他們客戶在數個商店位置的消費習慣。因為每個客戶通常只在其中一個地點消費，因此您可以說客戶效果項是**巢狀**於商店位置效果項內。

此外，您可以包含互動項（例如與相同的共變量有關的多項式項目）或新增多層巢狀結構到巢狀項目中。

限制：巢狀項目有下列限制：

- 互動內的所有因素都必須是唯一的。因此，如果 *A* 是因素，那麼指定 *A*A* 是無效的。
- 巢狀效果項中的所有因素都必須是唯一的。因此，如果 *A* 是因素，那麼指定 *A(A)* 是無效的。
- 共變量內不可巢套效應項。因此，如果 *A* 是因素，而 *X* 是共變量，那麼指定 *A(X)* 是無效的。

建構巢狀項目

1. 選取巢套於其他因素中的因素或共變量，並按一下箭頭按鈕。
2. 按一下**（之內）**。
3. 選取前一個因素或共變量巢套於其中的因素，並按一下箭頭按鈕。
4. 按一下**新增項目**。

您可以選擇性地包含互動項，或新增多層巢狀結構至巢狀項目。

隨機效應： 隨機效應因素為具有以下特色的欄位：其在資料檔中的值可視為從較大數值母群取得的隨機樣本。它們對於解釋目標中的過量變異性很有用。依預設，如果您在「資料結構」標籤中選取了多個受試者，則會針對最內層受試者之外的每個受試者建立「隨機效應」區塊。例如，如果您在「資料結構」標籤中選取了「學校」、「課程」和「學生」作為受試者，則會自動建立下列隨機效應區塊：

- 隨機效應 1：受試者是學校（沒有效應，只有截距）
- 隨機效應 2：受試者是學校 * 課程（沒有效應，只有截距）

隨機效應區塊的使用方法如下所示：

1. 若要新增區塊，請按一下**新增區塊...**。這會開啟第 179 頁的『隨機效應區塊』對話框。
2. 若要編輯現有區塊，請選取想要編輯的區塊，並按一下**編輯區塊...**。這會開啟第 179 頁的『隨機效應區塊』對話框。
3. 若要刪除一或多個區塊，請選取要刪除的區塊，並按一下「刪除」按鈕。

隨機效應區塊： 在來源清單中選取一或多個欄位並拖曳至效應清單，以便將效應輸入到模型中。所建立的效應類型視您將選項置於哪個熱點而定。類別（旗標、名義和序數）欄位在模型中會被當成因素使用，而連續欄位則會當成共變數使用。

- **主要。** 已放置欄位會在效應清單底端，以個別的主效應呈現。
- **雙向。** 已放置欄位所有可能的組合（兩個一組）都在效應清單底端，以雙向互動呈現。
- **三向。** 已放置欄位所有可能的組合（三個一組）都在效應清單底端，以三向互動呈現。
- ***.** 所有已放置欄位的組合都在效應清單底端，以單一互動呈現。

「效應建置器」右側的按鈕可讓您執行各種動作。

表 *11.* 效應建置器按鈕說明

圖示	說明
	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。
	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。
	使用第 178 頁的『新增自訂項目』對話框，將巢狀項目新增至模型，方法是按一下「新增自訂項目」按鈕。

包含截距。依預設，截距並未包含在隨機效應模型中。但是如果可以假設資料會穿過原點，則可以將截距排除在外。

顯示此區塊的參數預測。 指定此項可顯示隨機效應參數估計值。

定義共變異數群組依據。在此指定的類別欄位會定義獨立的隨機效應共變異數參數集；每一個欄位適用於一種由分組欄位交叉分類定義的種類。可以為每個隨機效應區塊指定不同的分組欄位集。所有受試者都具有相同的共變異數類型；同一個共變異數分組中的受試者將具有相同的參數值。

受試者組合。這可讓您在「資料結構」標籤中，從受試者的預設組合中指定隨機效應受試者。例如，如果學校、課程以及學生依序被定義為「資料結構」標籤中的受試者，那麼依照該順序，「受試者組合」下拉清單將會有**無**、**學校**、**學校 * 課程**以及**學校 * 課程 * 學生**這些選項。

隨機效應共變異數類型。這可指定殘差的共變異數結構。可用的結構包括：

- 第一階自身迴歸 (AR1)
- 自動迴歸異質 (ARH1)
- 自身迴歸移動平均 (1,1) (ARMA11)
- 複合對稱
- 非對稱性異質複合 (CSH)
- 對角線
- 尺度單位
- Toeplitz

- 非結構化
- 變異數成分

加權與偏移： 分析加權。尺度參數是與回應的變異數相關的估計模型參數。分析加權是可能依據不同觀察值而有所不同的已知值。若指定分析加權欄位，則會將尺度參數（與回應的變異數相關）除以每個觀察值的分析加權值。若記錄的分析加權值小於或等於零或遺漏，則不會用於分析中。

偏移。 偏移項是「結構」預測值。它的係數無法由模式估計，但系統假設其具有數值 1；因此偏移的呼叫會新增至目標的線性預測值。這在卜瓦松 (Poisson) 迴歸模型中特別有用，因為每個觀察值暴露於所需事件的層級可能不同。

例如，在建立個別駕駛員的意外事件比率模式時，具有三年經驗並曾於某一意外事件中出差錯的駕駛員，與具有 25 年經驗並曾於某一意外事件中出差錯的駕駛員間有重要的差異。如果駕駛員的經驗是以偏移項的方式納入，則可將意外事件數量模型化為含有對數鏈結的 Poisson 回應值或負二項式回應。

其他分配組合和鏈結類型可能需要偏移變數的其他轉換。

一般建置選項： 這些選項指定用來建置模型的更為進階的部分準則。

排序順序

這些控制項會決定目標和因素（類別輸入）的類別順序，以判斷「最後的」類別。如果目標並非類別目標，或者在第 175 頁的『目標』設定上指定自訂參照類別，則會忽略目標排序設定。

停止規則

您可以指定演算法將執行的疊代最大數目。演算法使用包含一個內部迴圈與一個外部迴圈的雙重疊代處理程序。為疊代最大數目指定的值會套用至這兩個迴圈。指定一個非負的整數。預設值是 100。

估計後設定

這些設定會判斷部分模型輸出要如何計算以供檢視。

信賴等級 (%)

此信賴等級用於計算模型係數的區間估計值。指定一個大於 0 且小於 100 的值。預設值是 95。

自由度

指定顯著性測試的自由度計算方式。如果您的樣本大小足夠大，或是資料已平衡，或是模型使用較簡單的共變異數類型（例如，尺度單位或對角線），則請選擇**殘差方法**。這是預設值。如果您的樣本大小較小，或是資料尚未平衡，或是模型使用複雜的共變異數類型（例如無結構），則請選擇**沙特斯懷特 (Satterthwaite) 近似法**。如果您的樣本大小較小，並且您具有「受限最大概似 (REML)」模型，請選擇 **Kenward-Roger 近似法**。

固定效應和係數的測試

這是計算參數估計值共變異數矩陣的方法。如果您擔心會違反模型假設，請選擇穩健性估計值。

估計： 模型建置演算法使用包含一個內部迴圈與一個外部迴圈的雙重疊代處理程序。下列設定適用於內部迴圈。

參數收斂。

如果參數估計值中的最大絕對變更或最大相對變更小於指定值 (必須為非負數值)，便假設收斂。如果指定的值等於 0，則不會使用準則。

對數概似收斂。

如果對數概似函數中的絕對變更或相對變更小於指定值 (必須為非負數值)，便假設收斂。如果指定的值等於 0，則不會使用準則。

Hessian 收斂。

當指定**絕對**時，如果以 Hessian 為主的統計量小於指定值，便假設收斂。當指定**相對**時，如果統計量小於指定值與對數概似絕對值的product，便假設收斂。如果指定的值等於 0，則不會使用準則。

Fisher 評分步驟數目上限。

指定一個非負的整數。值為 0 表示 Newton-Raphson 方法。值大於 0 表示先使用 Fisher 評分演算法，並且疊代次數最大為 *n*（其中 *n* 是指定的整數），然後再使用 Newton-Raphson。

奇異性容錯。

此值用作檢查奇異性時的容錯。請指定正數值。

註：依預設，使用「參數收斂」，其中，會勾選 1E-6 容錯的最大**絕對**變更。此設定產生的結果可能與第 22 版之前版本中取得的結果不同。若要從第 22 版之前的版本中重新產生結果，請針對「參數收斂」準則使用**相對**，並保持預設容錯值 1E-6。

一般： 模型名稱。您可以根據目標欄位來自動產生模型名稱，或是指定自訂名稱。自動產生的名稱為目標欄位名稱。如果存在多個目標，則模型名稱是依序排列的欄位名稱，名稱之間以 '&' 符號連接。例如，如果目標為 *field1*、*field2* 和 *field3*，那麼模型名稱為 *field1 & field2 & field3*。

可用於評分。對模型進行評分時，應生成此群組中的選定項目。對模型評分時，一律會計算所有目標的預測值以及類別目標的信賴度。所計算的信賴可以基於預測值的機率（最高預測機率），或是基於最高預測機率與次高預測機率之間的差異。

- **類別目標的預測機率。** 這會產生類別目標的預測機率。會為每一個類別建立一個欄位。
- **旗標目標的傾向評分。** 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。該模型會產生原始傾向評分；如果分割區有效，則該模型也會根據測試分割區產生調整傾向評分。

平均數估計值： 這個標籤可讓您顯示因素層次和因素互動的邊際平均數估計值。邊際平均數估計值不適用於多項式模型。

項目 「固定效應」中完全由類別欄位組成的模型塊列於此。檢查您要模型產生邊際平均數估計值的每個項目。

對比類型

這會指定要用於對比欄位層級的對比類型。

無 不會產生對比。

成對 產生指定因素所有層次組合的成對比較。這是因素互動唯一可用的對比。

離差 對比會比較每個因素層次與總平均數。

簡單 對比會比較每個因素層次（除了最後一個）與最後一個層次。「最後一個」層次由「建置選項」上指定的因素排序順序所決定。請注意，所有這些對比類型都不是正交。

對比欄位

這會指定使用選取的對比類型來比較其層級的因素。若選取**無**作為對比類型，則無法選取（或不需要選取）任何對比欄位。

連續欄位

所列的連續欄位是從「固定效應」中使用連續欄位的項目擷取而來。在計算邊際平均數估計值時，共變量會固定在指定值。選取平均數或指定自訂值。

為使用多重比較進行調整

使用多重對比執行假設測試時，可從所包含對比的顯著性層級調整整體顯著性層級。這可讓您選擇調整方法。

最小顯著差異

這個方法無法控制以下假設之整體可能性，此假設為某些線性的對比不同於虛無假設值，。

循序 *Bonferroni* 法*(Q)*

這是循序逐步拒絕的 Bonferroni 程序；就拒絕個別假設而言，此程序做法相當不保守，但整體顯著性層次仍維持相同。

循序 *Sidak* 法*(S)*

這是循序逐步拒絕的 Sidak 程序；就拒絕個別假設而言，此程序做法相當不保守，但整體顯著性層次仍維持相同。

序列 Sidak 方法比最小顯著差異方法保守，而序列 Bonferroni 又比序列 Sidak 方法保守；也就是說，最小顯著差異會拒絕至少和序列 Sidak 一樣多的個別假設，而序列 Sidak 會拒絕至少和序列 Bonferroni 一樣多的個別假設。

顯示平均數估計值依據

這會指定是根據目標的原始尺度還是鏈結函數轉換來計算邊際平均數估計值。

原始目標尺度

計算目標的邊際平均數估計值。請注意，當目標是以事件/試驗選項指定時，這會產生事件/試驗比例的邊際平均數估計值，而非事件數的邊際平均數估計值。

鏈結函數轉換

計算線性預測值的邊際平均數估計值。

模型視圖： 依預設，會顯示「模型摘要」視圖。若要查看其他模型視圖，請在視圖縮圖中選取。

模型摘要： 此視圖是一種Snapshot，是模型及其配適的一覽摘要。

表格。表格會指出在目標設定上指定的目標、機率分佈和鏈結函數。如果目標是由事件和試驗定義，則儲存格會分割以顯示事件欄位和試驗欄位或固定的試驗次數。另外，也會顯示以有限樣本修正的 Akaike 資訊準則 (AICC) 和 Bayesian 資訊準則 (BIC)。

- *Akaike* 修正。用於根據 -2（受限）對數概似值來選取及比較混合模型的量數。數值越小代表模式越佳。AICC 會「修正」較小 AIC 的樣本大小。當樣本大小增加時，AICC 會收斂至 AIC。
- *Bayesian*。用於根據 -2 對數概似值來選取及比較模型的量數。數值越小代表模式越佳。BIC 也會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型），但比 AIC 更嚴格。

圖表。如果目標為類別目標，則圖表會顯示最終模型的準確性，即正確分類的百分比。

資料結構： 此視圖提供您所指定的資料結構摘要，並協助您檢查是否已正確指定受試者和重複測量。第一個受試者的觀察資訊會顯示於每個受試者欄位和重複測量欄位以及目標中。另外，也會顯示每個受試者欄位和重複測量欄位的層級數量。

依觀察預測： 對於連續目標，包括指定為事件/試驗的目標，這會根據水平軸上的觀察值，來顯示垂直軸上預測值的 Bin 散佈平面圖。理想的狀況下，點應排列在 45 度的線上；此視圖可以告訴您模型是否有預測結果特別差的記錄。

分類： 對於類別目標，這會顯示熱圖中觀察值相對於預測值的交叉分類，以及整體正確百分比。

表格樣式。提供幾種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **列百分比。**這會顯示儲存格中的列百分比（以列總和的百分比表示的儲存格個數）。此為預設值。

- **儲存格個數。**這會顯示儲存格中的儲存格個數。熱圖的陰影仍會以列百分比為基礎。
- **熱圖。**這不在儲存格中顯示值，只會顯示陰影。
- **壓縮。**這不會在儲存格中顯示列標題、直欄標題或值。這在目標具有大量類別時很有用。

遺漏值。如果目標中有任何記錄具有遺漏值，則這些記錄會顯示於所有有效列下方的**（遺漏值）**列中。具有遺漏值的記錄不會納入整體百分比修正中。

多個目標。如果有多個類別目標，則每個目標都會顯示在個別的表格中，並有**目標**下拉清單可控制要顯示的目標。

大型表格。如果顯示的目標具有 100 個以上類別，則不會顯示表格。

固定效應： 此視圖會顯示模型中每個固定效應的大小。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **圖。**在此圖表中，效應是依照其在「固定效應」設定中的指定順序，從上到下排序。系統會根據效應顯著性來加權處理圖中的連接線，線條寬度越大代表效應顯著性越大（較小的 p 值）。此為預設值。
- **表格。** 此為整體模型與個別模型效應的 ANOVA 表格。個別效應是依照其在「固定效應」設定中的指定順序，從上到下排序。

顯著性。有「顯著性」滑塊，可控制視圖中所顯示的效應。系統會隱藏顯著性值大於滑塊值的效應。這不會變更模型，僅會讓您著重於最重要的效應。預設值為 1.00，因此系統不會根據顯著性來過濾任何效應。

固定係數： 此檢視會顯示模型中每個固定係數的值。請注意，模型當中的各項因素（類別預測值）皆已經過指標編碼，因此一般來說內含因素的**效應**會具有多個相關**係數**；除了對應於冗餘係數的類別外，每個類別會具有一個係數。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **圖。**此圖表先顯示截距，然後依照效應在「固定效應」設定中的指定順序從上到下排序。在內含因素的效應當中，系統會依資料值的升冪對係數進行排序。系統會根據係數顯著性來加權處理圖表中的連接線並以彩色顯示，線條寬度越大代表係數顯著性越大（較小的 p 值）。此為預設樣式。
- **表格。** 此表格會顯示個別模型係數的值、顯著性測試和信賴區間。在截距之後，效應是依照其在「固定效應」設定中的指定順序，從上到下排序。在內含因素的效應當中，系統會依資料值的升冪對係數進行排序。

多項式。如果多項式分佈有作用，則「多項式」下拉清單會控制要顯示的目標類別。清單中值的排序是由「建置選項」設定上的指定項目所決定。

指數。這會顯示特定模型類型的指數係數估計值和信賴區間，包括二元邏輯迴歸（二項式分佈和 logit 鏈結）、名義邏輯迴歸（多項式分佈和 logit 鏈結）、負二項式迴歸（負二項式分佈和對數鏈結）以及對數線性模型（卜瓦松 (Poisson) 分佈和對數鏈結）。

顯著性。有「顯著性」滑塊，可控制視圖中所顯示的係數。系統會隱藏顯著性值大於滑塊值的係數。這不會變更模型，僅會讓您著重於最重要的係數。預設值為 1.00，因此系統不會根據顯著性來過濾任何係數。

隨機效應共變異數： 此視圖會顯示隨機效應共變異數矩陣 (**G**)。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **共變異數值。**這是共變異數矩陣的熱圖，其中效應是依照其在「固定效應」設定中的指定順序，從上到下排序。相關圖中的色彩對應至儲存格值，如該鍵所示。此為預設值。

- **相關圖。** 這是共變異數矩陣的熱圖。
- **壓縮。**這是共變異數矩陣的熱圖，沒有列標題和直欄標題。

區塊。如果有多個隨機效應區塊，則會有「區塊」下拉清單，以供選取要顯示的區塊。

群組。 如果隨機效應區塊有群組指定項目，則會有「群組」下拉清單，以供選取要顯示的群組層級。

多項式。如果多項式分佈有作用，則「多項式」下拉清單會控制要顯示的目標類別。清單中值的排序是由「建置選項」設定上的指定項目所決定。

共變異數參數： 此檢視會顯示殘差和隨機效應的共變異數參數估計值和相關統計量。這些是進階、也是基本的結果，可提供共變異數結構是否適合的資訊。

摘要表。這是殘差 (**R**) 和隨機效應 (**G**) 共變異數矩陣中的參數數量、固定效應 (**X**) 和隨機效應 (**Z**) 設計矩陣中的等級（欄數），以及由用來定義資料結構之受試者欄位所定義的受試者數量的快速參照。

共變異數參數表格。對於選取的效應，會顯示每個共變異數參數的估計值、標準誤和信賴區間。所顯示的參數數量視效應的共變異數結構而定，對於隨機效應區塊，則由區塊中的效應數量而定。若您發現對角線外的參數並不顯著，您可以使用較為簡單的共變異數結構。

效應。如果有隨機效應區塊，則會有「效應」下拉清單，以供選取要顯示的殘差或隨機效應區塊。殘差效應一律可供使用。

群組。 如果殘差或隨機效應區塊有群組指定項目，則會有「群組」下拉清單，以供選取要顯示的群組層級。

多項式。如果多項式分佈有作用，則「多項式」下拉清單會控制要顯示的目標類別。清單中值的排序是由「建置選項」設定上的指定項目所決定。

估計平均數：顯著性效應： 這些是針對 10 個「最顯著」固定全因素效應所顯示的圖表，從三向互動啟動，然後是雙向互動，最後是主效應。此圖表會針對水平軸上主效應（或互動中第一個列出的效應）的每個值顯示垂直軸上的目標模型估計值；為互動中第二個列出之效應的每個值產生個別線條；為三向交互作中第三個列出之效應的每個值產生個別圖表；其他所有預測值則維持不變。其針對目標中每個預測值係數的效應提供實用的視覺化內容。請注意，若無任何顯著預測值，則不會產生任何估計平均數。

信賴度。這會使用指定為「建置選項」一部分的信賴等級，來顯示邊際平均數的信賴上限和下限。

估計平均值：自訂效應： 這些是適用於使用者所要求的固定全因素效應的表格和圖表。

樣式。提供各種不同的顯示樣式，您可以從**樣式**下拉清單中存取這些樣式。

- **圖。**此樣式會針對水平軸上主效應（或互動中第一個列出的效應）的每個值顯示垂直軸上的目標模型估計值折線圖；為互動中第二個列出之效應的每個值產生個別線條；為三向交互作中第三個列出之效應的每個值產生個別圖表；其他所有預測值則維持不變。

  如果要求了對比，則會顯示另一個圖表以比較對比欄位的層級；在互動方面，會為對比欄位以外的每個效應層級組合顯示圖表。在**成對**對比中，距離網路圖是以圖形表示的比較表，其中網路中節點之間的距離對應至樣本之間的差異。黃線對應至統計上的顯著差異；黑線對應至非顯著差異。游標停留於網路中的線上時，會以工具提示的方式顯示線所連接節點之間差異的調整顯著性。

  在**偏差**對比方面會顯示一個長條圖，其垂直軸上的是目標模型估計值，水平軸上的是對比欄位的值；在互動方面，則會為對比欄位以外的每個效應層級組合顯示圖表。長條顯示每個對比欄位層級與總平均數之間的差異，由黑色水平線條表示。

在**簡單**對比方面會顯示一個長條圖，其垂直軸上的是目標模型估計值，水平軸上的是對比欄位的值；在互動方面，則會為對比欄位以外的每個效應層級組合顯示圖表。長條顯示每個對比欄位層級（除了最後一個）與最後一個層級之間的差異，由黑色水平線條表示。

- **表格。**此樣式顯示一個表格，其中包含效應中每個欄位層級組合的目標模型估計值、標準誤、以及信賴區間；其他所有預測值則維持不變。

  如果要求了對比，則會顯示另一個表格，其中包含每個對比的估計值、標準誤、顯著性測試和信賴區間；在互動方面，除對比欄位以外，每個效應層級組合都會有個別的列集合。另外，還會顯示含整體測試結果的表格；在互動方面，除對比欄位以外，每個效應層級組合都會有個別的整體測試。

信賴度。這會使用指定為「建置選項」一部分的信賴等級，來切換顯示邊際平均數的信賴上限和下限。

佈置。這會切換成對對比圖的佈置。與網路佈置相比較，圓形佈置較無法顯示對比，但避免了重疊的線條。

設定： 在對模型分數時，應生成此標籤中的選定項目。對模型評分時，一律會計算所有目標的預測值以及類別目標的信賴度。所計算的信賴可以基於預測值的機率（最高預測機率），或是基於最高預測機率與次高預測機率之間的差異。

- **類別目標的預測機率。** 這會產生類別目標的預測機率。會為每一個類別建立一個欄位。
- **旗標目標的傾向評分。** 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。該模型會產生原始傾向評分；如果分割區有效，則該模型也會根據測試分割區產生調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

GLE 節點

GLE 模型通過指定的鏈結函數識別與因子和共變數線性相關的依變數。此外，此模式允許」變數具有非常態分配。它涵蓋了廣泛使用的統計模型，如用於正常分佈回應的線性迴歸、用於二進位資料的 logistic 模型、用於計數資料的對數線性模型、用於區間刪失生存分析資料的互補對數存活函數的對數模型以及使用其非常通用的模型公式的其他多數統計模型。

範例。航運公司可以使用廣義線性模型將卜瓦松 (Poisson) 迴歸調整到符合數種在不同時期建造的船隻損壞個數，而結果模式可以協助判斷何種船隻類型最容易損壞。

汽車保險公司可以使用廣義線性模型將 gamma 迴歸調整到符合汽車損壞理賠，而結果模式可以協助判斷影響理賠金額最鉅的因素。

醫療研究員可以使用廣義線性模型將「互補對數存活函數的對數」迴歸調整到符合區間受限存活資料，以預測某種疾病復發的時間。

GLE 模型的工作原理是建立一個方程式，從而使輸入欄位值與輸出欄位值進行關聯。一旦產生了模型，就可以用來估計新資料的值。

對於種類目標、每條記錄，將計算每個可能的輸出種類的會員資格機率。會將具有最高機率的目標種類指派為該記錄的預測輸出值。

需求。 您需要一個或多個輸入欄位，同時有且僅有一個具有兩個或多個種類的目標欄位（其測量層次可以為連續、種類或旗標）。模型中所用的欄位必須已完全實例化其類型。

目標

這些設定透過鏈結函數，來定義目標、其分佈及其與預測值的關係。

目標 目標為必要設置。目標可以具有任何測量層次，並且目標的測量層次會影響適合的分佈與鏈結函數。

- **使用預先定義目標** 要使用上游「類型」節點 （或上游來源節點的「類型」標籤）中的目標設定，請選取此選項。
- **使用自訂目標** 要手動分配目標，請選取此選項。
- **使用試用數作為分母** 如果目標回應是一組試驗中發生的事件數目量，目標欄位將包含該事件數目量，您可選取包含試驗次數的附加欄位。例如，在試驗新的殺蟲劑時，您可能會讓螞蟻樣本暴露在不同濃度的殺蟲劑之下，然後記錄殺死的螞蟻數量及每個樣本的螞蟻數。在這種情況下，用來記錄殺死的螞蟻數量的欄位應指定為目標（事件）欄位，用來記錄每個樣本中螞蟻數量的欄位應指定為試驗欄位。如果每個樣本的螞蟻數量均相同，則可使用固定值來指定試驗數。

  試驗數應大於或等於每個記錄的事件數。事件必須是非負值的整數，而試驗必須是正整數。
- **自訂參照種類**。您可以針對類別目標選擇參考類別。這會影響特定輸出，如參數估計值，但不會變更模型適合度。例如，如果您的目標採用數值 0、1 和 2，則依預設，程序會使最後一個（最高值）類別或 2 成為參考類別。這種狀況下，參數估計值應解釋為與種類 0 的概似度有關，或 1 相對於種類 2 的概似度。如果指定一個自訂種類，且目標具有定義標籤，可通過從清單中選一個值來設定參照種類。當您在指定模型的過程中，如果不記得特定欄位的確切編碼方式，這會是相當方便的功能。

目標分佈以及與線性模型的關係（鏈結） 給定預測值的值，模型的預期為目標值按照指定的形狀分佈，並在指定的鏈結函數中與預測值呈線性相關。提供了數種常見模型的捷徑，如果您想要配合捷徑清單上沒有的特定分佈和鏈結函數組合，可以選擇**自訂**設定。

- **線性模型** 使用身分鏈結函數指定常態分佈，在目標可用線性迴歸或 ANOVA 模型來預測時特別有用。
- **Gamma 迴歸** 使用對數鏈結函數指定 Gamma 分佈，在目標包含所有正值並向更大值偏斜時使用。
- **對數線性** 使用對數鏈結函數指定 Poisson 分佈，在目標代表某個固定時段內事件發生的次數時使用。
- **負二項式迴歸方法** 使用對數鏈結函數指定負二項式分佈，在目標和分母代表觀察到第 *k* 次成功所需的試驗次數時使用。
- **Tweedie 迴歸方法** 指定包含身分式、對數或冪次鏈結函數的 Tweedie 分佈，用於混合有零及正實數值的建模回應。這些分佈也稱為複合 *Poisson*、複合 *Gamma*及*Poisson Gamma*分佈。
- **多項式邏輯迴歸** 指定多項式分佈，在目標為多種類回應時使用。多項式分佈會使用累積 logit 鏈結（序數結果）或概化 logit 鏈結（多類別名義回應）。
- **二進位邏輯迴歸** 使用分對數鏈結函數指定二項式分佈，在目標為邏輯迴歸模型預測的二元回應時使用。
- **二元機率值** 使用機率值鏈結函數指定二項式分佈，在目標為具基本常態分佈的二元回應時使用。
- **區間刪失生存分析** 使用互補對數存活函數的對數鏈結函數指定二項式分佈，在存活分析的觀察沒有終止事件時特別有用。
- **自訂** 指定您自己的分佈和鏈結函數組合。

分佈

此選項指定目標的**分佈**。能夠指定非常態分佈和非身分鏈結函數是廣義線性模型相對於線性模型的重大改善。可能的分佈-鏈結函數組合有很多，且其中有好幾個都適用於給定的任何資料集，因此您的選擇可遵循先期提出的理論考量，或看起來最適合的組合。

- **自動** 如果您不確定要套用哪個分佈，請選取此選項；節點將分析您的資料以估計並套用最佳的分佈方法。
- **二項** 此分佈僅適用於代表二元回應或事件數目的目標。
- **Gamma** 此分佈適用於具有向更大正值偏斜的正尺度值的目標。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。
- **反向高斯** 此分佈適用於具有向更大正值偏斜的正尺度值的目標。若資料值小於或等於零或遺漏，則不會在分析中使用對應觀察值。
- **多項式** 此分佈適用於代表多種類回應的目標。模型形式取決於目標的測量層級。

  名義目標會導致產生名義多項式模型，其中會針對目標的每個類別（參考類別除外）估計一組個別的模型參數。給定預測值的參數估計值會顯示該預測值與目標之每個類別的概似之間的關係（相對於參考類別）。

  序數目標會導致產生序數多項式模型，其中傳統截距項目將由一組**臨界值**參數取代，這些參數與目標類別的累積機率相關。
- **負二項式** 負二項式迴歸方法使用帶對數關聯的負二項式分佈，它在目標代表具有較高變異數的發生計次時使用。
- **正太** 此分佈適用於其值圍繞中心值（平均數）呈對稱鐘形分佈的連續目標。
- **Poisson** 該分佈可視為被觀察事件在固定時間段內發生的次數，適合具有非負整數值的變數。若資料值為非整數、小於零或遺漏，則不會在分析中使用對應觀察值。
- **Tweedie** 此分佈適用於可以由 Poisson 分佈和 Gamma 分佈混合代表的變數；從某種意義上說，此分佈的「混合型」分佈，因為該分佈同時具有連續分佈（採用非負實數值）和離散分佈（正機率群位於單一值 0）的內容。因變數必須為數值，且資料值大於或等於零。若資料值小於零或遺漏，則不會在分析中使用對應觀察值。Tweedie 分配的固定值可以是任何大於 1 且小於 2 的數字。

鏈結函數

鏈結函數是容許進行模型估計的目標轉換。可以使用的函數如下：

- **自動** 如果您不確定要套用哪個鏈結函數，請選取此選項；節點將分析您的資料以估計並套用最佳的鏈結函數。
- **身分** $f(x)=x$。目標並未轉換。此鏈結可以和多項式以外的任何分佈搭配使用。
- **互補對數存活函數的對數** $f(x)=\log(-\log(1-x))$。這僅適用於二項式或多項式分佈。
- **Cauchit** $f(x) = \tan(\pi\ (x - 0.5))$。這僅適用於二項式或多項式分佈。
- **對數** $f(x)=\log(x)$。此鏈結可以和多項式以外的任何分佈搭配使用。
- **對數互補** $f(x)=\log(1-x)$。這僅適用於二項式分配。
- **Logit** $f(x)=\log(x\ /\ (1-x))$。這僅適用於二項式或多項式分佈。
- **負重對數** $f(x)=-\log(-\log(x))$。這僅適用於二項式或多項式分佈。
- **機率值** $f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是標準正太累積分佈的反函數。這僅適用於二項式或多項式分佈。
- **冪次** $f(x)=x^{\alpha}$（如果 $\alpha \neq 0$）。$f(x)=\log(x)$，如果 $\alpha=0$。α 是必須指定的數值，且必須是實數。此鏈結可以和多項式以外的任何分佈搭配使用。

Tweedie 參數 僅當您已選取 **Tweedie 迴歸方法**圓鈕或選取 Tweedie 作為**分佈**方法時才可用。選取介於 1 與 2 之間的某個值。

模型效應

固定效應因素通常被認為是相關值都呈現於資料集中、並能用於評分的欄位。依預設，具有預先定義輸入角色（未在對話框的任何地方指定）的欄位，會輸入於模型的固定效應部分中。種類（旗標、列名和序數）欄位可用作模型中的因子，連續欄位可用作共變數。

在來源清單中選取一或多個欄位並拖曳至效應清單，以便將效應輸入到模型中。所建立的效應類型視您將選項置於哪個熱點而定。

- **主要** 拖入的欄位顯示為獨立主效應，列在作用清單的底部。
- **雙向** 所有可能的拖入欄位配對顯示為雙向互動作用，列在作用清單的底部。
- **三向** 所有可能的三組拖入欄位顯示為三向互動作用，列在作用清單的底部。
- ***** 所拖動的全部欄位組合起來，作為單一的互動顯示在作用清單的底部。

「效應建置器」右側的按鈕可讓您執行各種動作。

表 *12.* 效應建置器按鈕說明

圖示	說明
	透過選取您想要刪除的項目，並按一下刪除按鈕，從固定效應模型中刪除項目。
	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，對固定效應模型內的項目進行重新排序。
	通過按一下「新增自訂項目」按鈕，使用「新增自訂項目」對話框向模型中新增巢套的項目。

包含截距 模型中通常包含截距。但是如果假設資料會穿過原點的話，就可以將截距排除在外。

新增自訂項目

您可以在這個程序中，為您的模式建立巢狀項目。通常巢狀項目在建立因素或共變量效果項的模式時非常有用，但因素或共變量的值不可以與其他因素層級互動。例如，連鎖雜貨店可能會追蹤他們客戶在數個商店位置的消費習慣。由於每個顧客只經常光顧其中一個店址，因此「顧客」作用可視為嵌入在「店址」作用中。

此外，您可以包含互動項（例如與相同的共變量有關的多項式項目）或新增多層巢狀結構到巢狀項目中。

限制。 巢狀項目有下列限制：

- 互動內的所有因素都必須是唯一的。因此，如果 *A* 是因素，那麼指定 *A*A* 是無效的。
- 巢狀效果項中的所有因素都必須是唯一的。因此，如果 *A* 是因素，那麼指定 *A(A)* 是無效的。
- 共變量內不可巢套效應項。因此，如果 *A* 是因素，而 *X* 是共變量，那麼指定 *A(X)* 是無效的。

建構巢狀項目

1. 選取巢套於其他因素中的因素或共變量，並按一下箭頭按鈕。
2. 按一下**（之內）**。
3. 選取前一個因素或共變量巢套於其中的因素，並按一下箭頭按鈕。
4. 按一下**新增項目**。

您可以隨意包含互動項，或新增多層巢狀結構至巢狀項目。

加權和偏移量

分析加權 尺度參數是與回應變異數相關的估計模型參數。分析加權是可能依據不同觀察值而有所不同的已知值。如果指定了**分析加權**欄位，那麼對每個觀察，都會用與回應變異數相關的尺度參數除以該分析加權值。分析中不會使用分析加權值少於等於 0 或遺漏的記錄。

偏移量 偏移量項目是一個結構預測值。它的係數無法由模式估計，但系統假設其具有數值 1；因此偏移的呼叫會新增至目標的線性預測值。這在卜瓦松 (Poisson) 迴歸模型中特別有用，因為每個觀察值暴露於所需事件的層級可能不同。

例如，為各個駕駛員的事故率建模時，有三年駕駛經驗的駕駛員在一次事故中的過錯率與有 25 年駕駛經驗的駕駛員在一次事故中的過錯率存在重大差別。如果駕駛員的經驗是以偏移項的方式納入，則可將意外事件數量模型化為含有對數鏈結的 Poisson 回應值或負二項式回應。

其他分配組合和鏈結類型可能需要偏移變數的其他轉換。

建置選項

這些選項指定用來建置模型的更為進階的部分準則。

排序 這些控制項用於確定目標和因子（種類輸入）種類的順序，以確定「最後一個」種類。如果目標並非類別目標，或者在第 186 頁的『目標』設定上指定自訂參照類別，則會忽略目標排序設定。

後置估計設定 這些設定可確定計算某些模型輸出供檢視的方式。

- **信賴等級 %** 這是用於計算模型係數的區間估計值的信賴等級。指定一個大於 0 且小於 100 的值。預設值是 95。
- **自由度** 用來指定如何為顯著性測試計算自由度。如果您的樣本大小足夠大，或是資料已平衡，或是模型使用較簡單的共變異數類型；例如，尺度單位或對角線，則請選擇**所有測試都為固定（殘差方法）**。此為預設值。如果您的樣本大小較小，或是資料尚未平衡，或是模型使用複雜的共變異數類型；例如無結構，則請選擇**測試各自不同（沙特斯懷特 (Satterthwaite) 近似法）**。
- **固定效果和係數的測試**。這是計算參數估計值共變異數矩陣的方法。如果您擔心會違反模型假設，請選擇穩健性估計值。

偵測影響離群值 對於除多項式分佈之外的所有分佈，請選取此選項以確定影響離群值。

處理趨勢分析 對於散佈平面圖，選取此選項可以處理趨勢分析。

估計

方法 選取要使用的極大概似估計方法；可用選項包括：

- 費雪 (Fisher) 評分(F)
- Newton-Raphson
- Hybrid

上限 Fisher 疊代次數 請指定一個非負整數。值為 0 表示 Newton-Raphson 方法。值大於 0 表示先使用 Fisher 評分演算法，並且疊代次數最大為 *n*（其中 *n* 是指定的整數），然後再使用 Newton-Raphson。

尺度參數方法 選取估計尺度參數的方法；可用選項包括：

- 最大概似估計
- 固定值。您也可以設定要使用的**值**。
- 離差
- Pearson 卡方

負二項式方法 選取估計負二項式輔助參數的方法；可用選項包括：

- 最大概似估計
- 固定值。您也可以設定要使用的**值**。

參數收斂 如果參數估計值中的最大絕對變更或最大相對變更小於指定值 (必須為非負數值)，便假設收斂。如果指定的值等於 0，則不會使用準則。

對數概似收斂 如果對數概似函數中的絕對變更或相對變更小於指定值 (必須為非負數值)，便假設收斂。如果指定的值等於 0，則不會使用準則。

Hessian 收斂 當指定**絕對**時，如果以 Hessian 為主的統計量小於指定值，便假設收斂。當指定**相對**時，如果統計量小於指定值與對數概似絕對值的product，便假設收斂。如果指定的值等於 0，則不會使用準則。

疊代數目上限 您可指定演算法執行的疊代數目上限。演算法使用包含一個內部迴圈與一個外部迴圈的雙重疊代處理程序。為疊代最大數目指定的值會套用至這兩個迴圈。指定一個非負的整數。預設值是 100。

單一容差 此值用作檢查奇異性時的容錯。請指定正數值。

註：依預設使用**參數收斂**，在此設置中，將檢查允差為 1E-6 的上限**絕對**變更。此設定可能會生成與第 17 版之前的版本中獲取的結果不同的結果。要重新產生第 17 版之前的版本中的結果，請對「參數收斂」準則使用**相對**，並保留預設允差值 1E-6。

模型選擇

使用模型選擇或正規化 要啟動此窗格中的控制項，請選中此勾選框。

方法 選取模型選擇方法，或者選擇要使用的正規化（如果使用**嶺**）。您可以從下列選項中進行選擇：

- **Lasso** 也稱為 L1 正規化，如果有許多預測值，那麼此方法的速度比向前逐步快。此方法將通過減小（即，施加懲罰值）參數來避免過度配適。它可以將某些參數減少為零，從而執行變數選擇 Lasso。
- **嶺** 也稱為 L2 正規化，此方法將通過減小（即，施加懲罰值）參數來避免過度配適。它會按相同比例縮小所有參數，但不會清除任何參數，並且不是變數選擇方法。
- **彈性網絡** 也稱為 L1 + L2 正規化，此方法將通過減小（即，施加懲罰值）參數來避免過度配適。它可以將某些參數減少為零，從而執行變數選擇。
- **逐步向前** 此方法開始時在模型中沒有任何作用，然後每次在一個步驟中新增或刪除作用，直到根據逐步準則不能再新增或移除作用為止。

自動偵測雙向互動 要自動偵測雙向互動，請選取此選項。

懲罰值參數

只有您選取了 Lasso 或彈性網絡**方法**時，這些選項才可用。

自動選取懲罰值參數 如果您不確定要設定何種參數懲罰值，請選中此勾選框，節點將識別並套用懲罰值。

Lasso 懲罰值參數 請輸入要由 Lasso 模型選擇**方法**使用的懲罰值參數。

彈性網絡懲罰值參數 1 請輸入要由彈性網絡模型選擇**方法**使用的 L1 懲罰值參數。

彈性網絡懲罰值參數 2 請輸入要由彈性網絡模型選擇**方法**使用的 L2 懲罰值參數。

逐步向前

只有您選取了逐步向前**方法**時，這些選項才可用。

包含 p 值不少於以下值的作用 指定作用必須具有的、要包含在計算中的下限機率值。

移除 p 值大於以下值的作用 指定作用必須具有的、要包含在計算中的上限機率值。

自訂最終模型中的效果數目上限 要啟動**效果數目上限**選項，請選中此選項。

效果數目上限 指定使用逐步向前建立方法時的效果數目上限。

自訂上限步進數 要啟動**上限步進數**選項，請選中此選項。

上限逐步數 指定使用逐步向前建立方法時的上限逐步數。

模型選項

模型名稱 您可以自動根據目標欄位產生模型名稱，也可以指定**自訂**名稱。自動產生的名稱為目標欄位名稱。如果存在多個目標，則模型名稱是依序排列的欄位名稱，名稱之間以 '&' 符號連接。例如，如果目標為 field1、field2 和 field3，那麼模型名稱為 *field1 & field2 & field3*。

計算預測值重要性 對於生成相應重要性測量的模型，可以顯示一個圖表來說明評估模型中每個預測值的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，對於部分模型，需要較長時間來計算預測值重要性，尤其是處理大型資料集時，結果便是依預設會關閉部分模型的預測值重要性。

如需相關資訊，請參閱第 37 頁的『預測值重要性』。

GLE 模型塊

GLE 模型塊輸出

建立 GLE 模型後，輸出中會提供下列資訊。

模型資訊表格

「模型資訊」表格提供模型的關鍵資訊。該表格識別一些高階模型設定，例如：

- 在「類型」節點或 GLE 節點欄位標籤中選取的目標欄位的名稱。
- 已建模和參照目標種類百分比。
- 機率分配和關聯的鏈結函數。
- 所使用的模型建置方法。
- 最終模型中輸入的預測值數目和號碼。
- 分類精確度百分比。
- 模型類型。
- 模型的百分比精確度（如果目標是連續目標）。

記錄摘要

摘要表顯示用於擬合模型的記錄數以及排除的記錄數。顯示的詳細資料包含所併入和排除的記錄數和所佔百分比以及未加權號碼（如果您使用了頻率加權）。

預測值重要性

「預測值重要性」圖形以長條圖形式顯示模型中前 10 個輸入（預測值）的重要性。

如果圖表中具有 10 個以上欄位，則您可以使用圖表下的調節器來變更圖表中包含的預測值選擇。調節器上的指示標示是固定寬度，並且調節器上的每個標示都代表 10 個欄位。您可以沿著調節器來移動指示標示以顯示後面或前面的 10 個欄位，依預測值重要性排序。

您可以按兩下圖表以開啟個別對話框來編輯圖形設定。例如，您可以修正一些項目，例如圖形大小，以及所用字型的大小和顏色。關閉這個個別的編輯對話框時，變更會套用至「輸出」標籤中顯示的圖表。

殘差（按預測圖形列出）

您可以使用此圖形來識別離群值，也可以使用它來診斷非線性或非恆定錯誤變異。理想圖形將顯示隨機分佈在基準行四周的點。

預期的型樣為標準化偏差殘差在線性預測值的預測值之間的分配具有平均數零和恆定範圍。預期的型樣是穿過零的水平行。

GLE 模型塊設定

在 GLE 模型塊的「設定」標籤上，您可以指定模型評分期間用於原始傾向的選項和用於 SQL 產生的選項。僅當模型片段已新增至串流之後，此標籤才可用。

計算原始傾向評分 對於只含有旗標目標的模型，您可以要求原始傾向評分來指出為目標欄位指定之真實結果的可能性。這些是標準預測及信賴度值的附加項目。無法使用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何產生 SQL：

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

Cox 節點

「Cox 迴歸」為時間對事件資料建置預測模型。模型會產生一個存活函數，預測針對預測值變數的給定值，所需事件在給定時間 *t* 的發生機率。生存函數的形狀及預測值的迴歸係數是從所觀察受試者預估；然後，模型可以套用至具有預測值變數測量的新觀察值。請注意，受限受試者（即在觀察時間期間未經歷所需事件的那些受試者）對於預估模型非常有用。

範例。 作為其減少客戶流失所做工作的一部分，電信公司對「流失時間」很感興趣，借此他們可以確定哪些因素導致客戶在很短的時間內更換使用其他電信服務。為此，選取了一個隨機的客戶樣本，並且從資料庫中抽取了他們作為客戶的時間（無論他們是否仍為活躍客戶）以及各種人口統計欄位。

需求。 您需要一個或多個輸入欄位，只需一個目標欄位，且必須在 Cox 節點中指定存活時間欄位。應對目標欄位進行編碼，使得 "false" 值表示存活時間，"true" 值表示所關注事件已發生；目標欄位的測量層次必須為旗標，且帶有字串或整數儲存。（必要的話，可使用「填充值」或「衍生」節點來轉換儲存體。）會忽略設為兩者 或無 的欄位。模型中所用的欄位必須已完全實例化其類型。存活時間可以是任意數值欄位。

註：在對「Cox 迴歸」模型進行評分時，如果種類變數中的空字串用作模型建置的輸入，那麼將報告錯誤。請避免使用空字串作為輸入。

日期和時間。「日期和時間」欄位不能直接用於定義存活時間；如果有「日期和時間」欄位，那麼應根據輸入研究的日期和觀測日期之間的差分，使用這些欄位建立包含存活時間的欄位。

Kaplan-Meier 分析。 可以在沒有輸入欄位的情況下執行 Cox 迴歸。這等效於 Kaplan-Meier 分析。

Cox 節點欄位選項

存活時間。選擇數值型欄位（測量層次為連續的欄位）以使節點可執行。存活時間表示所預測記錄的有效期限。例如，對流失的客戶時間進行建模時，它可能是記錄客戶在組織內的時間長度的欄位。客戶加入或流失的日期不會影響該模型；只有客戶保有期的持續時間與其相關。

存活時間為無單位的持續時間。您必須確保輸入欄位與存活時間相符。例如，在按月測量流失的研究中，您可將月銷售量而非年銷售量用作輸入。如果您的資料具有開始日期和結束日期而不是持續時間，您必須在 Cox 代碼上游將這些日期重新編碼為持續時間。

此對話框中的剩餘欄位是在整個 IBM SPSS Modeler 中使用的標準欄位。請參閱第 26 頁的『建模節點欄位選項』主題，以取得更多資訊。

Cox 節點模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

方法。下列選項可用於將預測值輸入模型：

- **輸入。** 這是預設方法，用於將所有項目直接輸入模型中。在建置模型時並未執行欄位選擇。
- **逐步。** 顧名思義，欄位選擇逐步方法用於分步建立模型。初始模型可能是最簡單的模型，其模型中不含任何模型塊（除常數外）。在每一個步驟中，會對尚未新增至模型的項目進行評估，並且如果這些項目的最適性會顯著性地新增至模型的預測能力，則會予以新增。此外，會對目前位於模型中的項進行重新評估，以判定是否可以移除其中任一項而不會顯著地減損模型。如果是的話，則會將其移除。程序重複執行，新增及/或移除其他項。當無法新增更多項以改善模型，以及無法移除更多項而不減損模型時，便會產生最終模型。
- **向後逐步。** 「逐步往回」方法實質上與「逐步」方法相反。使用此方法時，起始模型包含所有項作為預測值。在每一個步驟中，會對模型中的項進行評估，並且會移除可以移除而不會顯著地減損模型的任何項。此外，會對先前移除的項進行重新評估，以判定這些項的最適性是否會顯著性地新增至模型的預測能力。若是如此，則會將其新增回模型。當無法移除更多項而不會顯著地減損模型，以及無法新增更多項以改善模型時，便會產生最終模型。

註：自動方法（包含逐步和向後逐步）是適應性強的學習方法，並且特別容易過度擬合訓練資料。使用這些方法時，請務必使用新資料或以「分割區」節點建立的留出法測試範例，來驗證所產生模型的有效性。

群組。 指定群組欄位會導致節點為該欄位的每個種類計算個別的模型。該欄位可以是儲存類型為字串或整數的種類欄位（旗標或名義）。

模型類型 有兩個選項用於定義模型中的項目。**主效應**模型僅包含各個輸入欄位，而不測試輸入欄位之間的互動（相乘性作用）。**自訂**模型僅包括您指定的項目（主效應與互動）。選取此選項時，請使用「模型塊」清單在模型中新增或移除項目。

模型塊。 建置自訂模型時，您必須在模型中明確指定項目。該清單顯示模型目前的項目集。「模型塊」清單右邊的按鈕可用於新增或刪除模型塊。

- 若要將項目新增至模型，請按一下新增模型塊 按鈕。
- 若要刪除項目，請選取所需的項目，然後按一下刪除選定的模型塊 按鈕。

將項目新增到 Cox 迴歸模型

在要求自訂模型時，可以通過按一下「模型」標籤中的新增新的模型塊按鈕將各項目新增到模型中。此時將開啟一個新的對話框，您可在其中指定項目。

要新增的項目類型。 可根據在可用欄位清單中選取的輸入欄位，使用數種方法將項目新增至模型。

- **單一互動。** 插入代表所有所選欄位之互動的項目。
- **主效應。** 為每個選定的輸入欄位插入一個主效應項目（欄位本身）。
- **所有雙向互動。** 針對所選輸入欄位的每個可能的配對，插入一個雙向互動項目（輸入欄位的product）。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B* 和 *C*，則此方法將插入項目 *A * B*、*A * C* 和 *B * C* 中。
- **所有 3 向互動。** 針對所選輸入欄位的每個可能的組合，插入一個 3 向互動項目（輸入欄位的product），一次性採用 3 個。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B*、*C* 和 *D*，則此方法將插入項目 *A * B * C*、*A * B * D*、*A * C * D* 和 *B * C * D* 中。
- **所有 4 向互動。** 針對所選輸入欄位的每個可能的組合，插入一個 4 向互動項目（輸入欄位的product），一次性採用 4 個。例如，如果您已選取可用欄位清單中的輸入欄位 *A*、*B*、*C*、*D* 和 *E*，則此方法將插入項目 *A * B * C * D*、*A * B * C * E*、*A * B * D * E*、*A * C * D * E* 和 *B * C * D * E* 中。

可用的欄位。 列出要用來建構模型塊的可用輸入欄位。請注意，清單中可能包含非法輸入欄位，因此務必確保所有的模型塊都只包含輸入欄位。

預覽。 根據上述選取的欄位和項目類型，顯示按一下**插入**時將新增到模型中的項目。

插入。 根據目前選取的欄位與項目類型將項目插入模型，然後關閉對話框。

Cox 節點專家選項

聚合。 這些選項可讓您控制模型聚合的參數。執行模型時，聚合設定會控制重複執行不同參數以查看其適合度的次數。嘗試參數的頻率更高，結果越接近（亦即，結果將聚合）。請參閱第 195 頁的『Cox 節點收斂準則』主題，以取得更多資訊。

輸出。通過這些選項，可以要求將顯示在由節點建立的產生模型的進階輸出中的附加統計資料和統計圖（包括生存分析曲線）。請參閱第 195 頁的『Cox 節點進階輸出選項』主題，以取得更多資訊。

逐步。 這些選項可讓您使用「逐步」估計方法來控制新增及移除欄位的準則。（如果選取了「輸入」方法，則會停用該按鈕。）請參閱第 195 頁的『Cox 節點執行步驟準則』主題，以取得更多資訊。

Cox 節點收斂準則

最大疊代。 讓您指定模型的最大疊代，這個選項會控制程序尋找解決方案所需之時間。

對數概似收斂。 如果對數概似中的相對變更小於此值，則會停止疊代。如果值為 0 便不會使用這個條件。

參數收斂條件。 如果參數估計值中的絕對變更或相對變更小於此值，則會停止疊代。如果值為 0 便不會使用這個條件。

Cox 節點進階輸出選項

統計量。 您可以取得模型參數的統計量，包括 exp(*B*) 的信賴區間和估計相關。您可以在每個步驟都要求算出這些統計量，也可以最後一個步驟才要求。

顯示基準線函數。 讓您利用共變量的平均數來顯示基準線風險函數和累積存活函數。

圖形(L)

圖形可以幫您評估所用之估計模式，並且還會解釋結果。您可以繪製存活、風險、負對數存活函數的對數，以及被 1 減後的存活函數。

- 存活。可在線性尺度上顯示累積存活函數。
- 風險。可在線性尺度上顯示累積風險函數。
- **負對數存活函數的對數。** 顯示對估計運用了 ln(-ln) 變換之後的累加剩餘估計。
- 壹減存活。繪製線性尺度上被 1 減後的存活機率函數。

為每個值繪製個別的行。此選項僅可用於種類欄位。

要用於圖形的值。由於這些函數都依賴於預測值的值，因此您必須使用預測值的常數值來繪製函數隨時間推移的變化情況。預設情況下，將使用各個預測值的平均數作為常數值，但您可以使用網格為統計圖輸入自己的值。對於種類輸入，使用指示符編碼，因此每個種類都具有迴歸方法係數（最後一個種類除外）。因此，種類輸入具有每個指示符對照的平均數，等於種類中對應於指示符對照的觀察值比例。

Cox 節點執行步驟準則

移除準則。 為更強健的模型選取**概似比**。若要縮短建置模型的時間，您可以嘗試選取 **Wald**。還有附加選項 **條件** ，此選項提供以基於條件參數估計值的概似比統計資料的機率為依據的移除測試。

準則的顯著性臨界值。 通過使用此選項，您可以根據與每個欄位關聯的統計機率（*p* 值）來指定選擇準則。唯有當關聯的 *p* 值小於**輸入**值時才會將欄位新增至模型，而唯有當 *p* 值大於**移除**值時才會移除欄位。**輸入**值必須小於**移除**值。

Cox 節點設定選項

預測未來時間的生存分析情況。指定一個或多個未來時間。即在未發生終端機事件的情況下，無論每個觀察值是否可能至少在此時間長度（從現在開始）內生存分析，都將在每個時間值為每條記錄預測存活時間，一個時間值對應一個預測值。請注意，存活時間為目標欄位的 "false" 值。

- **定期。**存活時間值是根據指定的**時間間隔**和**要分數的時段數**產生的。例如，如果已要求 3 個時段，時間間隔為 2，那麼對未來時間的存活時間將為 2、4 和 6。以相同時間值評估每條記錄。
- **時間欄位。**在所選的時間欄位中為每條記錄提供存活時間（產生一個預測欄位），因此可以在不同的時間評估各條記錄。

過去存活時間。將迄今為止的記錄的存活時間指定為一個欄位，例如將現有客戶的保有期指定為一個欄位。在未來時間對生存分析的概似進行評分取決於過去存活時間。

註：未來和過去存活時間的值必須在用於訓練模型的資料的存活時間範圍內。時間外此範圍的記錄將標記為無效。

附加所有機率。 指定是否將每個種類的輸出欄位的機率新增至節點所處理的每筆記錄。如果未選取此選項，則只新增預測種類的機率。為每個未來時間計算機率。

計算累加故障函數。指定是否將累加故障的值新增到每條記錄中。為每個未來時間計算累積風險。

Cox 模型塊

Cox 迴歸模型代表由 Cox 節點所估計的方程式。它們包含模型擷取的全部資訊，以及模型結構與效能的相關資訊。

執行包含產生的 Cox 迴歸模型的串流時，該節點可新增包含模型預測和相關機率在內的兩個新欄位。新欄位的名稱衍生自要預測的輸出欄位的名稱並帶有字首和字尾，字首為表示預測種類的 *$C-* 或表示相關機率的 *$CP-*，而字尾為未來時間間隔的號碼或用於定義時間間隔的時間欄位的名稱。例如，對於名為 *churn* 的輸出欄位，以及定期定義的兩個未來時間間隔，新欄位命名為 *$C-churn-1*、*$CP-churn-1*、*$C-churn-2* 和 *$CP-churn-2*。如果使用時間欄位 *tenure* 定義未來時間，那麼新欄位為 *$C-churn_tenure* 和 *$CP-churn_tenure*。

如果在 Cox 節點中已選取了 **附加所有機率** 設定選項，那麼會針對每個未來時間新增兩個附加欄位，其中包含每條記錄生存分析和失敗的機率。這些附加欄位是根據輸出欄位的名稱進行命名的並帶有字首和字尾，字首為表示生存分析機率的 *$CP-<false value>-* 或表示事件已發生機率的 *$CP-<true value>-*，而字尾為未來時間間隔的號碼。例如，對於 "false" 值為 0，"true" 值為 1 的輸出欄位和定期定義的兩個未來時間間隔，新欄位命名為 *$CP-0-1*、*$CP-1-1*、*$CP-0-2* 和 *$CP-1-2*。如果使用單個時間欄位 *tenure* 定義未來時間，由於存在單個的未來區間，那麼新欄位為 *$CP-0-1* 和 *$CP-1-1*

如果在 Cox 節點中已選取了**計算累積風險函數**設定選項，那麼會針對每個未來時間新增附加欄位，其中包含每條記錄的累積風險函數。這些附加欄位是根據輸出欄位的名稱進行命名的並帶有字首和字尾，字首為 *$CH-*，而字尾為未來時間間隔的號碼或用於定義時間間隔的時間欄位的名稱。例如，對於名為 *churn* 的輸出欄位，以及定期定義的兩個未來時間間隔，新欄位命名為 *$CH-churn-1* 和 *$CH-churn-2*。如果使用時間欄位 *tenure* 定義未來時間，那麼新欄位為 *$CH-churn-1*。

Cox 迴歸輸出設定

除產生 SQL 外，塊的「設定」標籤與模式節點的「設定」標籤包含相同的控制項。 塊控制項的預設值由模式節點中設定的值決定。請參閱第 195 頁的『Cox 節點設定選項』主題，以取得更多資訊。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

Cox 迴歸方法進階輸出

Cox 迴歸的進階輸出可提供有關所估計模型及其效能的詳細資訊，其中包含生存分析曲線。進階輸出中包含的大部分資訊的技術含量都很高，需要具備 Cox 迴歸方面的廣泛知識才能正確理解該輸出。

第 11 章 叢集模型

叢集作業模型著重於識別相似記錄的群組，以及根據它們所隸屬的群組來標示記錄。不需事先瞭解群組資訊及群組特性即可完成該操作。事實上，甚至無法確切知道要查找多少個群組。這點將叢集模型與其他機器學習技術區別開來，即不存在供模型預測的預先定義輸出或目標欄位。由於不存在用於判斷模型的分類效果的外部標準，因而這些模型通常被稱作 **不受監督學習** 模型。對於這些模型而言，不存在 對或 錯的回答。其值根據它們的以下能力來決定：它們能夠擷取資料中的相關分組並為這些分組提供有用說明。

叢集方法基於對記錄間距離和叢集間距離的測量。將記錄指派給叢集時將盡量縮短的同一個叢集的記錄之間的距離。

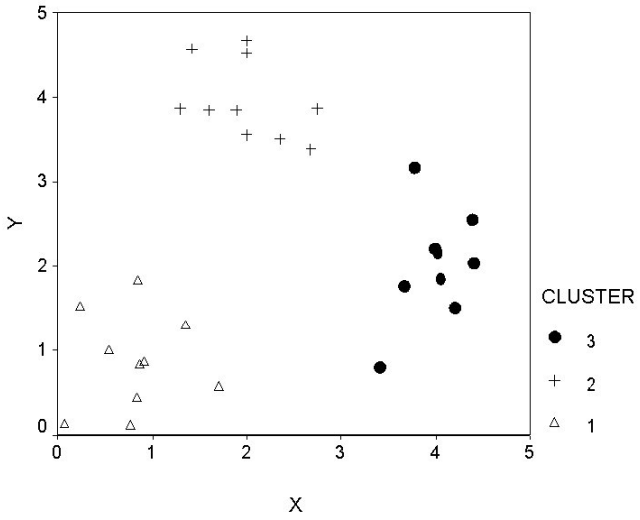


圖 *44.* 簡單叢集模型

提供了下列叢集方法：

K-Means 節點將資料集叢集到不同群組（或叢集）。此方法將定義固定的叢集數目量，將記錄迭代分配給叢集，以及調整叢集中心，直到進一步優化無法再精確模型。*k*-means 節點作為一種非監督學習機制，它並不試圖預測結果，而是揭示隱含在輸入欄位集中的型樣。

TwoStep 節點使用二階叢集方法。第一步完成簡單資料製作，以便將原始輸入資料壓縮為可管理的子叢集集合。第二步使用層級叢集方法將子叢集一步一步合併為更大的叢集。TwoStep 具有一個優點，就是能夠為訓練資料自動估計最佳叢集數目。它可以高效處理混合的欄位類型和大型的資料集。

Kohonen 節點會產生一種類神經網路，此類神經網路可用於將資料集叢集到各個差異群組。此網路訓練完成後，相似的記錄應在輸出對映中緊密地聚集，差異大的記錄則應彼此遠離。您可以通過查看模型塊 中每個單位所擷取觀察的數量來找出規模較大的單元。這將讓您對叢集的相應數量有所估計。

© Copyright IBM Corp. 1994, 2019

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非監督式學習來尋找資料集的叢集或密集區域。SPSS Modeler 中的 HDBSCAN 節點顯示了 HDBSCAN 程式庫的核心功能及常用參數。該節點在 Python 中實作，當您一開始不瞭解那是些什麼群組時，您可以使用它來將資料集叢集至不同的群組。

通常使用叢集模型來建立叢集或區段，然後將叢集或區段用作後續分析的輸入。常見例子如營銷人員常使用市場分區段來將整個市場劃分為多個類似的子群組。每個市場分區段都有自己的特性，該特性將影響到針對該分段的行銷努力是否能取得成功。如果您使用資料採礦來最佳化行銷戰略，通常可以通過識別合適的市場分區段和在預測模型中使用分區段資訊來顯著改進模型。

Kohonen 節點

Kohonen 網路是一種執行叢集的類神經網路類型，也稱為 **knet** 或 **自組織對映** 。如果在開始時沒有群組的相關資訊，那麼可使用此類型的網路將資料集叢集到有明顯區別的不同群組。對記錄進行群組，以便群組或叢集中的記錄趨於相似，而不同群組中的記錄則有所差異。

基本單元為**神經元**，並且它們分為兩層：**輸入層**和**輸出層**（也稱為**輸出對映**）。所有輸入神經元都和所有輸出神經元相連線，這些連線有與其相關的**強度**或**加權**。訓練過程中，每個單元會與所有其他單元進行競爭以「贏得」每條記錄。

輸出圖是神經元的二維網格（單元之間沒有連線）。

輸入資料會顯示在輸入層，相應值將傳播到輸出層。回應最強的輸出神經元將稱為**勝利者**並且會成為輸入的結果。

最初的加權隨機產生。如果某個單位贏得一條記錄，那麼其加權（與其附近單元的加權一起統稱為**芳鄰**）將作調整以盡可能地與此條記錄的預測值的型樣相符。顯示所有輸入記錄，並且加權將相應更新。將重複此過程，直到變化非常小為止。當進行訓練時，網格單元的加權將作調整從而形成叢集的一個二維度「對映」（所以會有項目**自組織對映**）。

此網路訓練完成後，相似的記錄應在輸出對映中緊密地聚集，差異很大的記錄則應彼此遠離。

與 IBM SPSS Modeler 中的大多數學習方法不同的是，Kohonen 網路 不 使用目標欄位。這種類型的學習（沒有目標欄位）稱為**未受監督的學習**。 Kohonen 網路試圖揭示輸入欄位集中的型樣而不是預測結果。通常，Kohonen 網路最終會形成幾個彙總多數觀測數據的單元（**強**單元），以及幾個實際不對應任何觀測數據的單元（**弱**單元）。強單元（有時也包括網格中與其鄰近的其他單元）代表可能的叢集中心。

Kohonen 網路的另一種用途是**維度縮減**。二維度網格的空間特性可提供從 *k* 個原始預測值到保留了原始預測值中親緣性關係的兩個衍生特性的對映。在某些情況下，此方法的作用與因子分析或主成分分析的作用相同。

請注意，計算輸出網格預設大小的方法與 IBM SPSS Modeler 以前的版本相比已發生了變化。通常，新方法將生成更小的輸出層，這些輸出層訓練起來更快且通用性更強。如果您發現使用預設大小得到的結果不理想，可以嘗試在「專家」標籤上增加輸出網格的大小。請參閱第 199 頁的『Kohonen 節點專家選項』主題，以取得更多資訊。

需求。 要訓練 Kohonen 網路，您需要一個或多個角色設定為輸入的欄位。角色設為目標、兩者或無的欄位會被忽略。

強度。 您不需要關於群組成員資格的資料即可建立 Kohonen 網路模型。您甚至不需要知道要尋找的群組的個數。Kohonen 網路剛開始會有大量的單元，隨著訓練的進行，這些單元會向資料中的自然叢集集中。可通過查看模型塊中每個單位擷取的觀察數來識別強單元，進而瞭解適當的叢集數目。

Kohonen 節點模型選項

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。

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。

繼續訓練現有模型。 依預設，每次執行 Kohonen 節點時，都會建立一個全新的網路。如果選中此選項，那麼會繼續訓練該節點成功生成的最後一個網絡。

顯示回饋意見圖形。 如果已選取此選項，那麼將在訓練期間顯示二維度陣列的直觀代表。每個節點的強度用顏色代表。紅色表示聚集了多數記錄的單元（**強**單元），白色表示聚集的記錄較少或沒有記錄的單元（**弱**單元）。如果建立模型所花費的時間相對較短，可能不會顯示意見。注意，此功能會減慢訓練進度。要加快訓練進度，請取消選中此選項。

停止時間。 預設停止準則將根據內部參數中止訓練。也可以指定時間作為停止準則。以分鐘為單位輸入網路訓練的時間。

設定隨機種子。 如果未設定隨機種子，那麼每一個執行節點時，用於起始設定網路加權的隨機值的序列都是不同的。這將導致即使節點設定和資料值都完全相同，節點也會在不同的執行中建立不同的模型。通過選取該選項，可以將隨機種子設定為特定值，從而使結果模型具有精確的可再現性。特定的隨機種子一律會產生相同的隨機值序列，在這種情況下執行節點一律會產生相同的產生模型。

註：對從資料庫中讀取的記錄使用**設定隨機種子**選項時，可能需要在取樣前使用「排序」節點以確保每次執行節點時都獲得相同的結果。 這是因為隨機種子取決於記錄的順序，而並不能保證順序在關聯式資料庫中保持相同。

附註：如果要在模型中包含標準（集合）欄位，但在建置模型時遇到記憶體問題，或者建置模型所需的時間過長，那麼可以考慮對大型集合欄位進行重新編碼以減少值的數量，或者考慮使用包含較少值的其他欄位作為該大型集合的 Proxy。例如，如果包含個別產品值的 *product_id* 欄位存在問題，您可以考量將其從模型中移除，然後改為新增不太詳細的 *product_category* 欄位。

最佳化。 選取旨在基於特定需要而在模型建置期間提高效能的選項。

- 選取**速度**以指示演算法永不為了增進效能而使用磁碟溢出。
- 選取**記憶體**來指示演算法在適當時以犧牲速度為代價使用磁碟溢出。依預設會選取這個選項。

  註：以分散式方式執行時期，`options.cfg` 中指定的管理者選項可能會置換此設定。

附加叢集標記。預設對新模型已選取此選項，但對從較早版本的 IBM SPSS Modeler 載入的模型取消選取。該選項會建立一個由 K-Means 和 TwoStep 節點共同建立的相同類型的種類分數欄位。在計算不同模型類型的分等級測量時，該字串欄位用於「自動叢集」節點。請參閱第 64 頁的『自動叢集節點』主題，以取得更多資訊。

Kohonen 節點專家選項

對於對 Kohonen 網路有詳盡瞭解的用戶，可使用專家選項對訓練過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

寬度和長度。將二維度輸出圖的大小（寬度和長度）指定為每個維度上的輸出單元數。

學習速率衰減。選取線性或指數學習速率衰減。**學習速率**是隨時間遞減的加權因素，使得網路可以從資料的大尺度特徵開始進行編碼，然後逐漸集中於更細微的資料資訊。

階段 1 和階段 2。Kohonen 網路訓練分為兩個階段。階段 1 是粗略估計階段，用於擷取資料中的大致型樣。階段 2 是調整階段，用於調整圖以便為資料更精細的特徵建模。每個階段都有以下三個參數：

- **芳鄰。**設定芳鄰的起始大小（半徑）。此參數確定在訓練期間與贏得單元一起被更新的「鄰近」單元數。在階段 1，芳鄰大小以 階段 *1* 芳鄰為起始值，然後減少到（ 階段 *2* 芳鄰 + 1）。在階段 2，芳鄰大小起始為 階段 *2* 芳鄰，然後減少到 1.0。階段 *1* 芳鄰應該大於階段 *2* 芳鄰。
- **起始 Eta。**設定學習速率 **eta** 的起始值。在階段 1，eta 起始於階段 *1* 起始 *Eta*，然後減少到階段 *2* 起始 *Eta*。在階段 2，eta 起始於階段 *2* 起始 *Eta*，然後減少到 0。階段 *1* 起始 *Eta* 應大於階段 *2* 起始 *Eta*。
- **週期。** 為訓練的每個階段設定循環數。每個階段均會進行指定次數的資料處理。

Kohonen 模型塊

Kohonen 模型塊包含由經過訓練的 Kohonen 網路擷取的所有資訊，還包含有關網路架構的資訊。

當執行包含 Kohonen 模型塊的串流時，節點將新增兩個新欄位，這兩個欄位包含 Kohonen 輸出網格中對該記錄反應最強烈的單元的 *X* 座標和 *Y* 座標。新的欄位名稱衍生自模型名稱，並帶有字首 *$KX-* 或 *$KY-*。例如，如果模型名稱為 *Kohonen*，那麼新欄位的名稱應是 *$KX-Kohonen* 和 *$KY-Kohonen*。

為了更好地瞭解 Kohonen 網路編碼的內容，可按一下模型塊瀏覽器上的「模型」標籤。此時會顯示叢集檢視器，提供叢集、欄位和重要性等級的圖形表示法。請參閱第 210 頁的『叢集檢視器 - 「模型」標籤』主題，以取得更多資訊。

如果要將叢集視覺化為網格，那麼可以通過使用「繪圖」節點繪製 *$KX-* 和 *$KY-* 欄位來檢視 Kohonen 網路的結果。（應在「繪圖」節點中選取 **X-Agitation** 和 **Y-Agitation** 以防止每個單元的記錄彼此重疊。）在圖形中，也可以重疊符號欄位以調查 Kohonen 網路如何將資料加入叢集。

另一個深入瞭解 Kohonen 網路的有力技術是，使用規則歸納找到可以對通過該網路探索的各個叢集加以識別的特性。如需相關資訊，請參閱主題第 89 頁的『C5.0 節點』。

如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』

Kohonen 模型彙總

Kohonen 模型塊的「摘要」標籤顯示網路的架構或拓蹼的相關資訊。二維度 Kohonen 特徵圖（輸出層）的長度和寬度顯示為 **$KX-** *model_name* 和 **$KY-** *model_name*。對於輸入層和輸出層，將列出該層的單位數。

K-Means 節點

K-Means 節點提供一種進行**叢集分析**的方法。當您一啟動不瞭解那是些什麼群組時，它可用來將資料集分組至不同的群組。與 IBM SPSS Modeler 中的大多數學習方法不同的是，K-Means 模型 不 使用目標欄位。這種類型的學習（沒有目標欄位）稱為**未受監督的學習**。 K-Means 模型試圖揭示輸入欄位集的型樣而不是預測結果。記錄會進行分組，因此某個群組或叢集內的記錄彼此會相似，但不同群組中的記錄並不同。

K-Means 的工作原理是根據資料定義一組起始叢集中心。然後根據記錄的輸入欄位值，將每條記錄分配到與其最相似的叢集中。在分配完所有記錄後，更新叢集中心以反映分配到每個叢集的新記錄集。然後再次檢查記錄，以確定是否應將這些記錄重新指派到不同的叢集中，這個記錄分配/叢集疊代過程將一直持續，直到達到疊代數目上限或一次疊代與下次疊代之間的改變不超過指定臨界值為止。

註：生成的模型在一定程度上取決於訓練資料的順序。對資料重新排序以及重建模型可能會得出不同的最終叢集模型。

需求。 要訓練 K-Means 模型，您需要一個或多個角色設定為輸入的欄位。角色設定為輸出、兩者或無的欄位將被忽略。

強度。 您不需要關於群組成員資格的資料即可建立 K-Means 模型。通常，K-Means 模型是進行大型資料集叢集的最快方法。

K-Means 節點模型選項

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。

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。

指定的叢集數。指定要產生的叢集數目。預設值為 5。

產生距離欄位。如果已選取此選項，那麼模型塊將包含一個欄位，該欄位包含每條記錄與所分配到的叢集的中心之間的距離。

叢集標籤。 為產生的叢集成員資格欄位的值指定格式。叢集成員資格可表示為具有指定**標籤字首**的**字串**（例如，「叢集 1」、「叢集 2」等等），也可以表示為**數字**。

附註：如果要在模型中包含標準（集合）欄位，但在建置模型時遇到記憶體問題，或者建置模型所需的時間過長，那麼可以考慮對大型集合欄位進行重新編碼以減少值的數量，或者考慮使用包含較少值的其他欄位作為該大型集合的 Proxy。例如，如果包含個別產品值的 *product_id* 欄位存在問題，您可以考量將其從模型中移除，然後改為新增不太詳細的 *product_category* 欄位。

最佳化。 選取旨在基於特定需要而在模型建置期間提高效能的選項。

- 選取**速度**以指示演算法永不為了增進效能而使用磁碟溢出。
- 選取**記憶體**來指示演算法在適當時以犧牲速度為代價使用磁碟溢出。依預設會選取這個選項。

  註：以分散式方式執行時期，`options.cfg` 中指定的管理者選項可能會置換此設定。

K-Means 節點專家選項

對於對 *k*-means 叢集有詳盡瞭解的用戶，可使用專家選項對訓練過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

停止時間。 指定訓練模型時要使用的停止準則。**預設**停止於準則為 20 次疊代或差異 < 0.000001，以先滿足的準則為準。選中**自訂**可指定自己的停止準則。

- **最大疊代。** 此選項可讓您在指定的疊代次數之後停止模型訓練。
- **差異允差。**通過此選項，您可以在某次疊代的叢集中心中的最大變更少於指定的層次時停止模型訓練。

集合的編碼值。指定 0 到 1.0 之間的值，以用於將集合欄位重新編碼為數字欄位群組。預設值是 0.5 的平方根（近似為 0.707107），它可為重新編碼的旗標欄位提供適當的加權。值越接近 1.0，對集合欄位的加權就越高於對數值型欄位的加權。

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 建模節點的串流時，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。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*。

深入瞭解 K-Means 模型的一種有力技術是，使用規則歸納法找到可以對通過該模型探索的各個叢集加以識別的特性。請參閱第 89 頁的『C5.0 節點』主題，以取得更多資訊。 您也可以按一下模型區塊瀏覽器上的「模型」標籤以顯示「叢集檢視器」，提供叢集、欄位以及重要性層次的圖形表示法。請參閱第 210 頁的『叢集檢視器 - 「模型」標籤』主題，以取得更多資訊。

如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』

K-Means 模型摘要

K-Means 模型塊的「摘要」標籤包含有關訓練資料、估計過程和由模型定義的叢集的資訊。顯示的資訊有叢集數目，還有疊代過程。如果您執行了附加至此建模節點的「分析」節點，則該分析中的資訊也將顯示在此區段中。

TwoStep 叢集節點

「TwoStep 叢集」節點提供一種形式的**叢集分析**。當您一啟動不瞭解那是些什麼群組時，它可用來將資料集分組至不同的群組。與 Kohonen 節點和 K-Means 節點一樣，「TwoStep 叢集」模型也不使用目標欄位。TwoStep 叢集模型試圖揭示輸入欄位集的型樣而不是預測結果。記錄會進行分組，因此某個群組或叢集內的記錄彼此會相似，但不同群組中的記錄並不同。

TwoStep 叢集是一種兩步驟叢集方法。第一步進行一次資料透通，這個過程將原始輸入資料壓縮為可管理的一組子叢集。第二步，採用階層式叢集方法逐漸將這些子叢集合併成越來越大的叢集，不需要再次進行資料透通。階層式叢集的優點在於不需要事先選取叢集數目。多數階層式叢集方法剛開始都將單個記錄作為最初的叢集，然後遞歸合併這些記錄以不斷生成更大的叢集。雖然此類方法常因資料數量巨大而失敗，但 TwoStep 的初始預叢集使得階層式叢集即使對於大型資料集也非常快。

註：得到的模型一定程度上取決於訓練資料的順序。對資料重新排序以及重建模型可能會得出不同的最終叢集模型。

需求。 要訓練「TwoStep 叢集」模型，您需要一個或多個角色設定為輸入的欄位。角色設為目標、兩者或無的欄位會被忽略。 TwoStep 叢集演算法不處理遺漏值。建立模型時將忽略任意輸入欄位包含空白的記錄。

強度。 「TwoStep 叢集」可以處理混合欄位類型並能有效處理大型資料集。它還能測試多種叢集解決方案並選擇其中最有效的一種，因此不必知道開始時應有多少個叢集。可將「二階叢集」設定為自動排除**離群值**或能對結果造成損害的極其異常情況。

重要：

IBM SPSS Modeler 有兩個不同版本的 TwoStep 叢集節點：

- **TwoStep 叢集**是在 IBM SPSS Modeler Server 上執行的傳統節點。
- 在連接至 IBM SPSS Analytic Server 時，可以執行**二階 AS 叢集**。

TwoStep 叢集節點模型選項

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。

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。

標準化數值欄位。依預設，TwoStep 會對所有數值型輸入欄位進行標準化，使它們具有相同的尺度，即平均數為 0 且變異數為 1。要保留數值型欄位的原始尺度，可取消選中此選項。符號欄位不受影響。

不含離群值。如果選中此選項，那麼那些與主要叢集似乎格格不入的記錄將自動排除在分析之外。這樣可以防止此類情況歪曲結果。

離群值偵測在預叢集步驟進行。已選取此選項時，會將相對於其他子叢集具有較少記錄的子叢集視為潛在離群值，且重新建立不包括這些記錄的子叢集樹狀結構。子叢集被視為包含潛在離群值的下方限大小由**百分比**選項控制。如果其中某些潛在離群值記錄與任何新子叢集配置足夠相似，那麼可將其新增到重新構建的子叢集中。將其餘無法合併的潛在離群值視為離群值新增到「雜訊」叢集中並排除在階層式叢集步驟之外。

使用經過離群值處理的 TwoStep 模型對資料進行評分時，會將與最近主要叢集的距離大於特定臨界值距離（基於對數概似值）的新觀察值視為離群值分配到「雜訊」叢集中，名稱為 -1。

叢集標籤。 為產生的叢集成員資格欄位指定格式。叢集會員資格可表示為具有指定**標籤字首**的**字串**（例如，"Cluster 1"、"Cluster 2" 等），也可以表示為**數字**。

自動計算叢集數目。 「TwoStep 叢集」可以非常迅速地對大量叢集解決方案進行分析並為訓練資料選擇最佳叢集數目。通過設定**上限**和**下限**叢集數目來指定要嘗試的解決方案的範圍。「二階叢集」通過一個兩階段過程確定最佳叢集數目。在第一個階段，隨著所新增叢集的增多，可基於貝葉斯資訊準則 (BIC) 中的差異選取模型中叢集數目的上限。在第二個階段，為叢集數比下限 BIC 解決方案還少的所有模型找出叢集間下限距離的差異。距離的最大差異用於識別最終叢集模型。

指定叢集數目。 如果知道模型中要包含的叢集的號碼，請選中此選項並輸入叢集數目。

距離測量 此選項可決定兩個叢集間計算的相似琵度。

- **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。** 歐基里得量數是兩個叢集間的「直線」距離。它只能在所有變數皆為連續時使用。

叢集準則。 此選項可決定自動叢集演算法決定叢集數目的方式。您可以指定「Bayesian 資訊準則」(Bayesian Information Criterion，BIC)，或指定「Akaike 資訊準則」(Akaike Information Criterion，AIC)。

TwoStep 叢集模型塊

TwoStep 叢集模型塊包含由叢集模型擷取的所有資訊，還包含有關訓練資料和估計過程的資訊。

當執行包含「TwoStep 叢集」模型塊的串流時，節點將為該記錄新增包含叢集成員資格的新欄位。新欄位名稱衍生自模型名稱，以 *$T-* 為字首。例如，如果模型名稱為 *TwoStep*，那麼新欄位的名稱應是 *$T-TwoStep*。

深入瞭解 TwoStep 模型的一種有力技術是，使用規則歸納找到可以對通過該模型探索的各個叢集加以識別的特性。請參閱第 89 頁的『C5.0 節點』主題，以取得更多資訊。 您也可以按一下模型區塊瀏覽器上的「模型」標籤以顯示「叢集檢視器」，提供叢集、欄位以及重要性層次的圖形表示法。請參閱第 210 頁的『叢集檢視器 - 「模型」標籤』主題，以取得更多資訊。

如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』

TwoStep 模型彙總

TwoStep 叢集模型塊的「摘要」標籤顯示找出的叢集數目以及有關訓練資料、估計過程和所使用的建構設定值的資訊。

請參閱第 36 頁的『瀏覽模型塊』主題，以取得更多資訊。

TwoStep-AS 叢集節點

IBM SPSS Modeler 有兩個不同版本的 TwoStep 叢集節點：

- **TwoStep 叢集**是在 IBM SPSS Modeler Server 上執行的傳統節點。
- 在連接至 IBM SPSS Analytic Server 時，可以執行**二階 AS 叢集**。

Twostep-AS 叢集分析

「TwoStep 叢集」是設計用來顯示資料集中自然分組 (或叢集) 的探索工具 (原本不會加以顯示)。此程序使用的演算法有多個不錯的特徵使其有別於傳統叢集技術：

- **處理類別和連續變數。** 藉由假設變數為自變數，結合的多項式-常態分配就可以放置在類別和連續變數上。
- **自動選擇叢集數目。** 藉由比較不同叢集解之間的模型-選項準則的值，此程序可自動決定最適叢集數目。
- **擴展性** 通過構造對記錄進行彙總的叢集特徵 (CF) 樹狀結構，二階演算法能夠分析大型資料檔案。

例如，零售和消費者產品公司定期地對說明客戶的購買習慣、性別、年齡、收入層次和其他屬性的資訊套用叢集技術。這些公司針對每個消費者群體定制其行銷和產品開發戰略，以增加銷售量並建立品牌忠誠度。

欄位標籤

「欄位」標籤指定要在分析中使用的欄位。

使用預先定義的角色。已選取所有具有已定義的「輸入」角色的欄位。

使用自訂欄位指定。新增和移除欄位，而不考慮對其定義的角色指派。您可以選取具有任意角色的欄位，並將其移入或移出**預測值（輸入）**清單。

基本

叢集數目

自動決定

此程序確定指定範圍內的最佳叢集數目。**最小值**必須大於 1。這是預設選項。

指定固定

此程序產生指定的叢集數目。**號碼**必須大於 1。

叢集準則

此選項控制項自動叢集演算法如何確定叢集數。

BIC 準則

用於根據 -2 對數概似值來選取及比較模型的量數。數值越小代表模式越佳。BIC 也會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型），但比 AIC 更嚴格。

AIC 資訊準則 (AIC)

用於根據 -2 對數概似值來選取及比較模型的量數。數值越小代表模式越佳。AIC 會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型）。

自動叢集方法

如果您選擇了**自動決定**，請從下面用於自動決定叢集數目的叢集方法中選擇：

使用叢集準則設定

資訊條件收斂是對應於兩個目前叢集解的資訊條件與第一個叢集解的比例。所使用的準則是在「叢集準則」群組中選取的準則。

距離跳轉

距離跳轉是與兩個連續叢集解相對應的距離的關係。

上限 對資訊條件收斂法的結果和距離跳躍法的結果進行組合，以生成與第二次跳躍相對應的叢集數目。

下限 對資訊條件收斂法的結果和距離跳躍法的結果進行組合，以生成與第一次跳躍相對應的叢集數目。

特徵重要性方法

特徵重要性方法確定特徵（欄位）在叢集解中的重要性。輸出包含有關整體特徵重要性和每個叢集中的每個特徵欄位的重要性的資訊。將排除不符合下限臨界值的特徵。

使用叢集準則設定

這是預設方法，此方法基於在「叢集準則」群組中選取的準則。

作用大小

特徵重要性基於作用大小而不是顯著性值。

特徵樹狀結構準則

這些設定確定如何建立叢集特徵樹狀結構。通過建立叢集特徵樹狀結構並對記錄進行彙總，二階演算法能夠分析大型資料檔案。換而言之，TwoStep 叢集使用叢集特徵樹狀結構來建置叢集，從而使其能夠處理多數觀察值。

距離測量

此選項可決定兩個叢集間計算的相似琵度。

對數概似值

概似測量假設欄位服從某種機率分配。假設連續欄位呈正常分佈，而假設種類欄位呈多項式分佈。所有欄位皆假設為不相依。

歐幾里德

歐基里得量數是兩個叢集間的「直線」距離。使用平方歐幾里德距離測量和 Ward 法來計算叢集之間的親緣性。僅當所有欄位都是連續欄位時，才能使用此測量。

離群值叢集

包含離群值叢集

包含作為一般叢集離群值的觀察值的叢集。如果未選取此選項，那麼所有觀察值都將併入在一般叢集中。

特徵樹狀結構葉節點中的觀察值數少於

如果特徵樹狀結構葉節點中的觀察值數少於指定的值，那麼將該葉節點視為離群值。此值必須是大於 1 的整數。如果您變更了此值，那麼較大的值可能會產生較多的離群值叢集。

離群值的最高百分比

構建叢集模型時，離群值將按離群值強度進行排名。進入離群值主要百分比所需的離群值強度作為臨界值，用於確定是否將觀察值分類為離群值。較大的值意味著將較多的觀察值分類為離群值。此值必須介於 1 與 100 之間。

其他設定

初始距離變動臨界值

這是用於使叢集特徵樹狀結構增長的初始臨界值。如果將葉節點插入項目到樹狀結構中的葉節點後，所產生的緊性少於此臨界值，那麼該葉節點將不再分割。如果緊性超過此臨界值，那麼該葉節點將進行分割。

葉節點上限分支數

分葉節點所能擁有的子節點最大數目。

非葉節點上限分支數

非葉節點可以具有的上限子節點數。

樹狀結構深度上限

叢集樹狀結構可以具有的上限層次數。

測量層次的調整加權

通過增加連續欄位的加權降低種類欄位的影響。此值代表用於降低種類欄位加權的分母。例如，預設值 6 使種類字段的加權為 1/6。

記憶體配置

叢集演算法使用的上限記憶體數量，以兆位元組 (MB) 計。如果此程序使用的空間量超過此最大值，那麼將使用磁碟來儲存記憶體中放不下的資訊。

延遲分割

延遲叢集特徵樹狀結構的重建。叢集演算法在評估新觀察值時，將多次重建叢集特徵樹狀結構。此選項將延遲該作業並減少重建該樹狀結構的次數，從而提高效能。

標準化

叢集演算法處理已標準化的連續欄位。依預設，所有連續欄位都已標準化。為了節省部分時間和計算工作，您可以將已標準化的連續欄位移到**不標準化**清單。

功能選擇

在「功能選擇」畫面上，可以設定規則以確定何時排除欄位。例如，可以排除包含許多遺漏值的欄位。

用於排除欄位的規則

遺漏值百分比大於

在分析中，將排除遺漏值百分比大於指定值的欄位。此值必須是大於零且少於 100 的正數。

種類欄位個種類數大於

在分析中，將排除種類數大於指定號碼的種類欄位。此值必須是大於 1 的正整數。

趨向於單一值的欄位

連續欄位的變異係數少於

在分析中，將排除變異係數少於指定值的連續欄位。變異係數是標準差與平均數之比。較小的值常常表示這些值的變異程度較小。此值必須在 0 到 1 之間。

種類欄位的單個種類中的觀察值百分比大於

在分析中，將排除單個種類中的觀察值百分比大於指定值的種類欄位。數值必須大於 0 且小於 100。

自適應功能選擇

此選項將執行額外的資料傳遞，以尋找並移除最不重要的欄位。

模型輸出

模型建置摘要

模型規格

模型規格數、最終模型中的叢集數目以及最終模型中包含的輸入數（欄位數）的摘要。

記錄摘要

模型中併入和排除的記錄（觀察值）的號碼和百分比。

已排除輸入

對於任何不包含在最終模型中的欄位，顯示欄位被排除的原因。

評估

模型品質

這個表格顯示每個叢集的優度和重要性以及整體模型擬合度。

特徵重要性長條圖

這個長條圖顯示特徵（欄位）在所有叢集中的重要性。在長條中，長條圖較長的特徵（欄位）比長條圖較短的特徵（欄位）更為重要。特徵（字段）還按重要性以遞減排序（最前面的條形最重要）。

特徵重要性單字雲

這個單字雲顯示特徵（欄位）在所有叢集中的重要性。文字較大的特徵（欄位）比文字較小的特徵（欄位）更為重要。

離群值叢集

如果您選擇不包含離群值，那麼這些選項將處於停用狀態。

互動式表格和圖表

這個表格和圖表顯示離群值強度以及離群值叢集與一般叢集的相對親緣性。在表格中選取不同的列，將會在圖表中顯示不同離群值叢集的資訊。

樞紐表

這個表格顯示離群值強度以及離群值叢集與一般叢集的相對親緣性。這個表格包含的資訊與互動式顯示畫面的資訊相同。這個表格支援所有用來將表格置於樞軸上並進行編輯的標準功能。

最大數目

輸出中要顯示的上限離群值數目。如果有超過 20 個離群值叢集，將會改為顯示樞紐表。

解譯

叢集間的特徵重要性剖面圖

互動式表格和圖表。

這些表格和圖表顯示叢集解中使用的每個輸入（欄位）的特徵重要性和叢集中心。在表格中選取不同的列將會顯示一個不同的圖表。對於種類欄位，顯示長條圖。對於連續欄位，將顯示平均數和標準離差的圖表。

樞紐表。

這個表格顯示每個輸入（欄位）的特徵重要性和叢集中心。這個表格包含的資訊與互動式顯示畫面的資訊相同。這個表格支援所有用來將表格置於樞軸上並進行編輯的標準功能。

叢集中的特徵重要性

對於每個叢集，顯示每個輸入（欄位）的叢集中心和特徵重要性。每個叢集都有一個個別的表格。

叢集距離

這個窗格圖表顯示叢集之間的距離。每個叢集都有一個個別的窗格。

叢集標籤

文字 每個叢集的標籤由為**字首**指定的值和後跟的序列號組成。

數字 每個叢集的標籤是一個序列號。

模型選項

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。

兩階 AS 叢集模型塊

TwoStep-AS 模型區塊會在「輸出檢視器」的「模型」標籤中顯示模型的詳細資料。如需使用檢視器的相關資訊，請參閱《Modeler 使用手冊》(ModelerUsersGuide.pdf) 中標題為「使用輸出」的小節。

兩階 AS 叢集模型塊包含由叢集模型擷取的所有資訊，以及有關訓練資料和估計過程的資訊。

執行包含兩階 AS 叢集模型塊的串流時，節點將為該記錄新增包含叢集成員資格的新欄位。新欄位的名稱從模型名稱衍生，並以 ***$AS-*** 作為字首。例如，如果模型名為 TwoStep，那麼新欄位將名為 ***$AS-TwoStep***。

深入瞭解兩階 AS 模型的一種有力技術是，使用規則歸納法找到可以對該模型所探索的各個叢集加以識別的特性。如需相關資訊，請參閱主題第 89 頁的『C5.0 節點』。

如需使用模型瀏覽器的一般資訊，請參閱第 36 頁的『瀏覽模型塊』

二階-AS 叢集模型塊設定

「設定」標籤為二階-AS 模型塊提供了額外的選項。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **透過轉換為原生 SQL 進行評分** 如果選取此項，則會產生原生 SQL 來在資料庫內對模型進行評分。

  註：雖然這個選項可以更快地提供結果，但隨著模型複雜性的增加，原生 SQL 的大小和複雜性也會增加。
- **在資料庫外進行評分**此選項會從資料庫提取回您的資料，並在 SPSS Modeler 中對其進行評分。

K-Means-AS 節點

K-Means 是其中一個最常用的叢集演算法。它將資料點叢集化至預先定義的叢集數。[1] SPSS Modeler 中的 K-Means-AS 節點是在 Spark 中實作。

如需 K-Means 演算法的詳細資料，請參閱https://spark.apache.org/docs/2.2.0/ml-clustering.html。

請注意，K-Means-AS 節點自動針對類別變數執行 one-hot 編碼。

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-Means-AS 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

使用預先定義的角色。 此選項會告知節點使用來自上游「類型」節點的欄位資訊。依預設選取該項。

使用自訂欄位指派。 如果您想要手動指派輸入欄位，請選取此選項，然後選取一或多個輸入欄位。使用此選項與在「類型」節點中將欄位角色設為**輸入**類似。

K-Means-AS 節點建置選項

使用「建置選項」標籤可以指定 K-Means-AS 節點的建置選項，包括用於模型建置的一般選項、用於起始設定叢集中心的起始設定選項，以及用於計算反覆運算及隨機種子的進階選項。如需相關資訊，請參閱 SparkML 上 K-Means 的 JavaDoc。[1]

一般

模型名稱。對特定叢集評分之後產生的欄位名稱。選取**自動**（預設值），或者選取**自訂**，然後鍵入一個名稱。

叢集數目。指定要產生的叢集數目。預設值為 **5**，且最小值為 **2**。

起始設定

起始設定模式。指定起始設定叢集中心的方法。**K-Means\|\|** 是預設值。如需有關這兩個方法的詳細資料，請參閱可調式 K-Means++。[2]

起始設定步驟。如果已選取 **K-Means\|\|** 起始設定模式，請指定起始設定步驟數。**2** 是預設值。

進階

進階設定。如果您要如下所示設定進階選項，請選取此選項。

最大疊代。指定搜尋叢集中心時執行的疊代數目上限。**20** 是預設值。

容錯。指定疊代演算法的聚合容錯。**1.0E-4** 是預設值。

設定隨機種子。選取此資訊並按一下**產生**可以產生由亂數字產生器使用的種子。

顯示

顯示圖形。如果您想要在輸出中包括圖形，請選取此選項。

下表顯示了 SPSS Modeler K-Means-AS 節點中的設定與 K-Means Spark 參數之間的關係。

表 *13.* 對映至 *Spark* 參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	K-Means SparkML 參數
輸入欄位	features	
叢集數目	clustersNum	k
起始設定模式	initMode	initMode
起始設定步驟	initSteps	initSteps
最大疊代	maxIter	maxIter
容錯	toleration	tol
隨機種子	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

叢集檢視器

叢集模型通常用於根據檢驗的變數尋找類似記錄的群組（或叢集），其中相同群組中成員之間的親緣性高，不同群組中成員之間的親緣性低。尋找的結果可用於找出不明顯的關聯性。例如，透過叢集分析客戶喜好、收入層級與購買習慣，可找出較可能回應特定市場行銷活動的客戶類型。

有兩種方法可解讀叢集顯示的結果：

- 檢驗叢集，以判定該叢集獨一無二的特性。一個叢集是否包含所有高收入借款人？此叢集包含的記錄是否比其他叢集的多？
- 檢驗各叢集的欄位，以判定數值在叢集之間如何分佈。一個人的教育程度是否決定了在叢集中的成員資格？高信用評分是否區分不同叢集中的成員資格？

您可以在「叢集檢視器」中使用主要視圖與各種鏈結的視圖來深入瞭解，以協助您回答這些問題。

下列叢集模型塊會在 IBM SPSS Modeler 中產生：

- Kohonen 淨模型塊
- K-Means 模型塊
- TwoStep 叢集模型塊

若要檢視叢集模型塊的相關資訊，請在模型節點上用滑鼠右鍵按一下，並選擇快顯功能表中的「**瀏覽**」（若是資料流中的節點，則選擇「**編輯**」）。或者，如果您使用的是「自動叢集」建模節點，則在「自動叢集」模型塊內連按兩下所需的叢集項目。請參閱第 64 頁的『自動叢集節點』主題，以取得更多資訊。

叢集檢視器 - 「模型」標籤

叢集模型的「模型」標籤會以圖形顯示摘要統計量，與叢集之間的欄位分佈，這也是所謂的「**叢集檢視器**」。

附註：以 IBM SPSS Modeler 13 版之前版建立的模式沒有「模型」標籤。

「叢集檢視器」由兩個畫面組成，主要視圖位於左邊，鏈結或輔助視圖位於右邊。主要視圖有兩種：

- 模型摘要（預設值）。請參閱第 211 頁的『模型摘要視圖』主題，以取得更多資訊。
- 叢集。請參閱第 211 頁的『叢集檢視』主題，以取得更多資訊。

鏈結/輔助視圖有四種：

- 預測值重要性。請參閱第 212 頁的『叢集預測值重要性視圖』主題，以取得更多資訊。
- 叢集大小（預設值）。請參閱第 212 頁的『叢集大小視圖』主題，以取得更多資訊。
- 儲存格分配。請參閱第 213 頁的『儲存格分配檢視』主題，以取得更多資訊。
- 叢集比較。請參閱第 213 頁的『叢集比較檢視』主題，以取得更多資訊。

模型摘要視圖

「模型摘要」視圖會顯示叢集模型的Snapshot，也就是摘要，包括叢集結合與分組的 Silhouette 量數，此量數會以不同顏色來表示差、可或佳的結果。此 Snapshot 可讓您快速檢查品質是否為差，如果為差，您可以決定回到建模節點修正叢集模型設定，以產生更好的結果。

差、可與佳的結果是依據 Kaufman 與 Rousseeuw (1990) 關於叢集結構解讀的著作而決定。在「模型摘要」視圖中，結果「佳」則表示到達 Kaufman 與 Rousseeuw 的資料為叢集結構合理或強力證據等級，結果「可」則表示到達薄弱證據等級，結果「差」則表示到達非顯著證明的等級。

所有記錄的 Silhouette 量數平均為 (B A) / max(A,B)，其中 A 為記錄距離其叢集中心的距離，B 為距離最近的非其所屬叢集中心的距離。Silhouette 係數 1 是指所有的觀察值直接位於其叢集中心。值 1 表示所有觀察值位於另一個叢集的叢集中心。平均而言，平均數值為 0 表示觀察值距離其叢集中心與另一個最近叢集中心的距離是相等的。

摘要會提供一張表格，其中包含下列資訊：

- **演算法。** 所使用的叢集演算法，例如 "TwoStep"。
- **輸入功能。** 欄位數目，也就是所謂的**輸入**或**預測值**。
- **叢集。** 解中的叢集數目。

叢集檢視

「叢集」視圖包含叢集對特徵網格中，網格包含叢集名稱、大小和每個叢集的分析概要。

網格中的直欄包含下列資訊：

- **叢集。** 演算法建立的叢集數目。
- **標籤。** 任何套用到每個叢集上的標記（預設為空白）。在儲存格內部連按兩下以輸入可說明叢集內容的標記，例如 Luxury 汽車買主。
- **說明。** 任何叢集內容的說明（預設為空白）。在儲存格內連按兩下以輸入叢集的說明，例如「55 歲以上，專業人員，收入超過 100,000 美元」。
- **大小。** 每個叢集大小，以佔整體叢集樣本的百分比表示。網格內每個大小儲存格會顯示一個垂直列，會顯示在叢集內的大小百分比、以數值格示表示的大小百分比，以及叢集觀察值個數。
- **功能。** 個別的輸入值或預測值，依照預設會按照整體重要性排序。如果有任何直欄的大小相同，則會以叢集數目的遞增排序順序顯示。

  整體特徵重要性會以儲存格背景顏色來表示，最重要的特徵顏色最深，最不重要的特徵則沒有顏色。上表的網格指出每個特徵儲存格色彩的重要性。

將滑鼠停留在儲存格上方，會顯示特徵的完整名稱/標記與儲存格的重要性值。是否顯示進一步的資訊，取決於檢視與特徵的類型。在「叢集中心」視圖中，這包含儲存格統計資料與儲存格值；例如："平均數：4.32"。若是類別特徵，則儲存格會顯示最多（典型）的類別及其百分比。

在「叢集」視圖內，您可以選取各種方式來顯示叢集資訊：

- 轉置叢集與特徵。請參閱『轉置叢集與特徵。』主題，以取得更多資訊。
- 排序特徵。請參閱『排序特徵』主題，以取得更多資訊。
- 排序叢集。請參閱『排序叢集』主題，以取得更多資訊。
- 選取儲存格內容。請參閱『儲存格內容』主題，以取得更多資訊。

轉置叢集與特徵。： 依預設，叢集顯示為欄，功能顯示為列。若要將此顯示方式顛倒，請按一下「**排序特徵依據**」按鈕左側的「**轉置叢集與特徵**」按鈕。例如，當您有許多叢集要顯示時，可能需要這麼做，如此不需要太多水平捲動就能看到資料。

排序特徵： 「**排序特徵依據**」按鈕可讓您選取顯示特徵儲存格的方式：

- **整體重要性。** 這是預設的排序順序。特徵是按照整體重要性遞減排序，各叢集之間的排序順序皆相同。如果有任何特徵的重要性值同分，則會以特徵名稱的遞增排序順序列出同分的特徵。
- **叢集內重要性。** 系統會根據特徵對每個叢集的重要性來排序特徵。如果有任何特徵的重要性值同分，則會以特徵名稱的遞增排序順序列出同分的特徵。選擇此選項後，排序順序通常會隨著叢集改變。
- **名稱。** 特徵是依照名稱的字母順序排序。
- **資料順序。** 特徵是按照其資料集順序排序。

排序叢集： 依照預設，叢集會依照大小的遞減順序排序。「**排序叢集依據**」按鈕可讓您按照名稱字母順序排序叢集，如果您已為叢集建立唯一的標記，則改為以標記字母順序排序。

標記相同的特徵會依照叢集名稱排序。如果叢集是按照標記排序的，當您編輯了叢集的標記後，排序順序會自動更新。

儲存格內容： 「**儲存格**」按鈕可讓您變更特徵的儲存格內容與評估欄位的顯示方式。

- **叢集中心。** 依預設，儲存格會顯示特徵名稱/標記，以及每個叢集/特徵組合的集中趨勢。系統會針對連續欄位與眾數（最常出現的類別）顯示平均數，以及類別欄位的欄表百分比。
- **絕對分配。** 顯示每個叢集內的特徵名稱/標記，以及特徵的絕對分配。若是類別特徵，顯示畫面會出現與類別重疊的長條圖，這些類別按照資料值的遞增順序排序。若是連續特徵，顯示畫面會顯示平滑密度圖，此圖會為每個叢集使用相同的端點與間隔。

  著上實心紅色的顯示畫面會顯示叢集分配，較淡色的顯示畫面則表示整體資料。
- **相對分配。** 在儲存格中顯示特徵名稱/標記與相對分配。一般而言，顯示畫面與顯示絕對分配的分畫面類似，不同處為顯示的是相對分配。

  著上實心紅色的顯示畫面會顯示叢集分配，較淡色的顯示畫面則表示整體資料。
- **基本檢視。** 當叢集很多時，如果沒有捲動畫面，很難看見所有的詳細資訊。若要減少捲動的次數，請選取此檢視將顯示畫面變更為表格的精簡版。

叢集預測值重要性視圖

「預測值重要性」視圖會顯示每個欄位在估計模型時的相對重要性。

叢集大小視圖

「叢集大小」視圖會顯示圓餅圖，其中包含每個叢集。每個圖塊上會顯示每個叢集的百分比大小，將滑鼠停在每個圖塊上方，圖塊中會顯示個數。

在圖表下方的表格會列出下列大小資訊：

- 最小叢集大小（個數與佔整體的百分比）。
- 最大叢集大小（個數與佔整體的百分比）。

- 最大叢集對最小叢集的大小比例。

儲存格分配檢視

「儲存格分配」視圖會為您在「叢集」主畫面的表格中選取的任何特徵儲存格，以展開的方式顯示更詳細的資料分配圖。

叢集比較檢視

「叢集比較」視圖包含網格模式佈置，此佈置會在列中顯示特徵，在直欄中顯示選取的叢集。此檢視可幫助您更瞭解構成叢集的因素，它也可以讓您看見叢集之間的差異，這些差異不只是叢集與整體資料的比較，也會有叢集與叢集間的比較。

若要選取要顯示的叢集，請按一下「叢集」主畫面上叢集直欄的頂端。使用 Ctrl-按一下或 Shift-按一下來選取或取消選取多個要進行比較的叢集。

附註：您最多可以選擇顯示五個叢集。

系統會以選取叢集的順序來顯示叢集，欄位順序則是由「**排序特徵依據**」選項所決定。選取「**叢集內重要性**」時，一律依據整體重要性排序欄位。

背景圖會顯示每個特徵的整體分配：

- 類別特徵會顯示為點形圖，點的大小代表每個叢集最多/最常用的類別（按照特徵）。
- 連續特徵會顯示為盒形圖，圖中會顯示整體中位數與四分位距。

所選取叢集的盒形圖會在這些背景檢視上重疊：

- 若是連續特徵，方形點標記與水平線會表示每個叢集的中位數與四分位距。
- 每個叢集都以不同的色彩表示，並在檢視頂端顯示。

瀏覽叢集檢視器

「叢集檢視器」是互動式的顯示畫面。您可以：

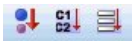
- 選取欄位或叢集可檢視更多詳細資訊。
- 比較叢集以選取感興趣的項目。
- 改變顯示畫面。
- 轉置軸。
- 使用「產生」功能表產生「衍生」、「過濾」與「選取」節點。

使用工具列

您可以使用工具列選項來控制在左窗格與右窗格中顯示的資訊。您可以使用工具列控制項來變更顯示畫面的方向（上到下、左到右或右到左）。此外，您也可以將檢視器重設為預設設定，並開啟對話框來指定主畫面中「叢集」視圖的內容。

「**排序特徵依據**」、「**排序叢集依據**」、「**儲存格**」與「**顯示**」選項只有在您選取了主畫面中的「**叢集**」視圖後才能使用。請參閱第 211 頁的『叢集檢視』主題，以取得更多資訊。

表 *14.* 工具列圖示

圖示	主題
	請參閱轉置叢集與特性
	請參閱特性排序依據
	請參閱叢集排序依據
	請參閱儲存格

從叢集模型產生節點

「產生」功能表可讓您根據叢集模型建立新節點。此選項位於所產生模型的「模型」標籤中，可讓您根據目前的顯示畫面或選項（亦即所有可見的叢集或所有選取的節點）產生節點。例如，您可以選取單一特徵，然後產生「過濾器」節點來捨棄所有其他（不可見）的特徵。產生的節點會不連接地放置在構圖區中。此外，您可以在模型色板上產生模型塊的複本。記得先將節點連接起來，並進行任何所需的編輯後再執行。

- **產生建模節點。** 在資料流構圖區上產生建模節點。例如，您有一個資料流，而您要在這個資料流中使用這些模式設定，但您卻沒有過去用來產生這些設定時使用的建模節點時，這個功能便很有用。
- **模式至色板。** 在「模型」色板上建立項目。當同事寄給您一個包含模型的資料流，而非模式本身時，這個功能很有用。
- **過濾節點。** 建立新「過濾器」節點以過濾出叢集模型未使用的欄位，和/或目前「叢集檢視器」顯示畫面中看不見的欄位。如果「叢集」節點的上游處有「類型」節點，則產生的「過濾器」節點會捨棄所有角色為「目標」的欄位。
- **過濾節點（從選擇內容）。** 建立新「過濾器」節點，以根據「叢集檢視器」中的選擇內容來過濾欄位。使用 Ctrl-按一下方法可選取多個欄位。在「叢集檢視器」中選取的欄位會在下游捨棄，但您在執行之前可透過編輯「過濾器」節點來變更這個行為。
- **選取節點。** 建立新的「選取」節點，以根據記錄在目前「叢集檢視器」中任何可見叢集內的成員資格來選取記錄。系統會自動產生選取條件。
- **選取節點（從選擇內容）。** 建立新的「選取」節點，以根據在「叢集檢視器」中所選取叢集內的成員資格來選取記錄。使用 Ctrl-按一下方法可選取多個叢集。
- **衍生節點。** 建立新的「衍生」節點，此節點可根據在「叢集檢視器」中所有可見叢集內的成員資格來衍生旗標欄位，此欄位可將 *True* 或 *False* 值指定給記錄。系統會自動產生衍生條件。
- **衍生節點（從選擇內容）。** 建立新的「衍生」節點，以根據在「叢集檢視器」中所選取叢集內的成員資格來衍生旗標欄位。使用 Ctrl-按一下方法可選取多個叢集。

除了產生節點外，您也可以使用「產生」功能表建立圖形。請參閱第 215 頁的『從叢集模型產生圖形』主題，以取得更多資訊。

控制叢集視圖顯示畫面

若要控制主畫面「叢集」視圖中顯示的內容，請按一下「**顯示**」按鈕；「顯示」對話框隨即開啟。

功能。 預設為選取。若要隱藏所有輸入特徵，請取消選取此勾選框。

評估欄位。 選擇要顯示的評估欄位（建立叢集模型時未使用，但已傳送到模型檢視器以評估叢集的欄位）；預設不會顯示任何欄位。附註：評估欄位必須是具有多個值的字串。如果沒有任何的評估欄位，則無法使用此勾選框。

叢集說明。 預設為選取。若要隱藏所有叢集說明儲存格，請取消選取此勾選框。

叢集大小。 預設為選取。若要隱藏所有叢集大小儲存格，請取消選取此勾選框。

最大類別個數。 指定在類別特徵的圖表中顯示的類別數目上限，預設值為 20。

從叢集模型產生圖形

叢集模型提供許多資訊，但其格式有時不便於業務使用者存取。若要提供資料以便可以將資料輕易納入商業報告、簡報等，您可以產生所選資料的圖形。例如，可從「叢集檢視器」產生所選叢集的圖表，這樣可以只建立該叢集中觀察值的圖表。

註：僅當模型塊附加到串流中的其他節點時，您才能從叢集檢視器中產生圖表。

產生圖形

1. 開啟包含「叢集檢視器」的模型塊。
2. 在「模型」標籤上，從**視圖**下拉清單選取叢集。
3. 在主面板上，選取您要為其生成圖表的一個或多個叢集。
4. 從「產生」功能表，選擇**圖表（從選擇創建）**；顯示「圖版基本」標籤。

   附註：以此方式顯示圖形板時，只有「基本」及「詳細」標籤可用。
5. 使用「基本」或「詳細」標籤設定來指定要顯示在圖形上的詳細資料。
6. 按一下「確定」以產生圖形。

圖表標題識別模型類型和選擇併入在內的一個或多個叢集。

第 12 章 關聯規則

相關規則將特定結論（例如，購買特定產品）與一組條件（例如，購買多個其他產品）關聯起來。例如，規則

```
啤酒 <= 罐裝蔬菜 & 冷凍食品 (173, 17.0%, 0.84)
```

表述的是：啤酒經常與罐裝蔬菜和 冷凍食品一起成對出現。該規則可靠率為 84% 並適用於 17% 的資料或 173 條記錄。關聯規則演算法自動找到可使用可視技術（比如 Web 節點）手動找到的關聯。

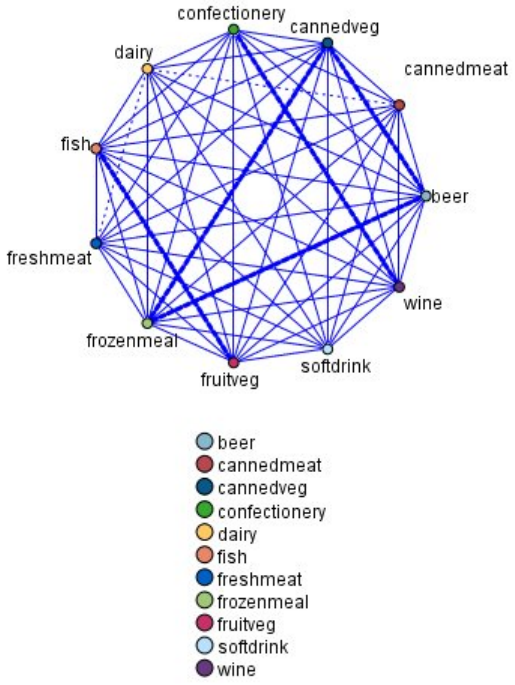


圖 *45.* 顯示市場購物籃商品之間的關聯的網絡節點

與標準的決策樹演算法（C5.0 和 C &R 樹狀結構）相比，關聯規則演算法的優點是 任何屬性之間都可以存在關聯。決策樹演算法將建置只有一個結果的規則，而關聯演算法會嘗試尋找許多規則，每個規則可能具有不同的結果。

關聯演算法的缺點是試圖在可能非常大的搜尋空間中尋找規則，因而執行時期間比樹狀結構演算法長得多。關聯演算法使用 **產生與測試** 方法來尋找規則（簡式規則將初始產生）並對照資料集來驗證這些規則。將儲存符合條件的規則，然後規範所有受各種限制限制的規則。 **規範** 是將條件新增到規則的過程。然後這些新規則將對照資料進行驗證，並且驗證過程中將迭代儲存最符合條件和最有用的規則。使用者通常會對容許進入規則的前提條件的可能的數量給出一定限制，並根據資訊理論和高效編製索引方式使用各種技術來縮小原來可能很大的搜尋空間。

正在處理結束後，將給出最符合條件的規則的列表格。此組相關規則不能直接用於做出預測，這點與標準的模型（比如決策樹或類神經網路）不同。這是由於規則可能有多數不同的結論。需要將相關規則變換為分類規則集的另外一層次變換。因此，關聯演算法生成的相關規則被稱作 **未優化模型** 。雖然使用者可以瀏覽這些未優化模型，但除非使用者指令系統從未優化模型產生分類模型，否則無法明確地將這些模型用作分類模型。用戶可通過瀏覽器的「產生」功能表選項來完成這種轉換。

© Copyright IBM Corp. 1994, 2019

受支援兩種關聯規則演算法：

「事前」節點從資料擷取一組規則，即擷取資訊內容最多的規則。Apriori 節點提供五種選取規則的方法並使用複雜的編製索引模式來高效地處理大資料集。對於較大的問題，Apriori 訓練的速度通常較快較快；它對可保留的規則數目量沒有任何限制，而且可處理最多帶有 32 個前提條件的規則。「事前」要求輸入和輸出欄位均為種類型欄位，但因為它專為處理此類型資料而進行最佳化，因而處理速度快得多。

序列節點可探索循序資料或與時間有關的資料中的相關規則。序列是一系列可能會以可預測順序發生的項目集合。例如，一個購買了剃刀和須後水的顧客可能在下次購物時購買剃須膏。「序列」節點基於 CARMA 關聯規則演算法，使用有效的兩段式方法來尋找序列。

表格資料與交易資料

關聯規則模型使用的資料可能是交易處理格式，也可能表格式，如下方所述。下面的內容是一般說明；具體的要求可能有所不同，請參見每種模型類型文件中的討論。請注意，對模型進行評分時，要評分的資料必須反映用於建立該模型的資料格式。使用表格資料建立的模型只能用於對表格資料進行評分；使用交易資料建立的模型只能對交易資料進行評分。

交易處理格式

交易資料對於每個交易或項目具有一個個別的記錄。比方說，如果某位客戶購買多次，則每一次的購買都是獨立的記錄，並以客戶 ID 鏈結相關聯項目。有時這又稱為**收銀機捲紙 (till-roll)** 格式。

客戶	購買
1	jam
2	牛奶
3	jam
3	bread
4	jam
4	bread
4	牛奶

Apriori、CARMA 和序列節點都可使用交易資料。

表格資料

表格資料（也稱為 **籃子** 資料或 **真值表格** 資料）由個別的旗標代表項目，其中每個旗標欄位代表一個特定項目的出現或不出現。每條記錄代表一個相關項目的完整集合。旗標欄位可以是種類的也可以是數值的，但某些模型具有更具體的要求。

客戶	Jam	麵包	Milk
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Apriori、CARMA、GSAR 和序列節點都可以使用表格資料。

Apriori 節點

Apriori 節點還會探索資料中的相關規則。Apriori 提供了五種不同的規則選取方法，並使用一種複雜的編製索引編制方案來高效處理大型資料集。

需求。 要建立 Apriori 規則集，您需要一個或多個輸入欄位和一個或多個目標欄位。輸入欄位和輸出欄位（角色為輸入、目標或兩者的欄位）必須是符號型欄位。會忽略角色為無的欄位。執行節點之前欄位類型必須完全實例化。資料可以是表格式，也可以是交易格式。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。

強度。 對於較大的問題，Apriori 訓練的速度通常較快. 它對於可以保留的規則數目也沒有任何限制，可以處理最多帶有 32 個前置條件的規則。Apriori 提供了五種不同的訓練方法，因此將資料挖掘方法與當前問題相符時可以實現更強的靈活性。

Apriori 節點模型選項

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。

最低前提條件支援。 您可以指定針對在規則集中保留規則的支援準則。**支援**指的是訓練資料中條件（規則中的 "if" 部分）為 true 的記錄的百分比。（請注意，此支援定義與 CARMA 和序列節點中使用的定義不同。請參閱第 232 頁的『序列節點模型選項』主題，以取得更多資訊。）如果您希望獲得套用至很小部分資料的規則，請嘗試增大此設定。

註：Apriori 的支援定義基於帶有前提條件的記錄的號碼。這與 CARMA 和序列演算法不同，對於這兩種演算法，支援定義基於具有規則中所有項目（即條件和結果）的記錄的數量。關聯模型的結果顯示（條件）支援和規則支援度兩個測量。

最低規則信賴度。您還可以指定信賴度準則。**信賴度**基於其規則的前提條件為 true 的記錄，並且是其結果也為 true 的記錄的百分比。換句話說，置信度是基於規則的正確預測的百分比。會捨棄信賴度低於指定準則的規則。如果您獲得的規則太多，請嘗試增加此設定。如果您獲得的規則太少（甚至根本無法獲得規則），請嘗試降低此設定。

註：必要的話，您可以在您的專屬值中強調顯示值和類型。請注意，如果您將信賴度值降低到 1.0 以下，則除了處理程序需要大量可用記憶體之外，您還可能發現需要花費極長時間才能建置規則。

上限前提條件數。 您可以為任何規則指定上限前置條件數。這是限制規則複雜性的一種方法。如果規則太複合或者太具體，請嘗試降低此設定。此設定對於訓練時間也具有很大的影響。如果規則集訓練所需的時間過長，請嘗試降低此設定。

旗標只有真值。如果對於表格（真值表格）格式的資料選取了此選項，則在生成的規則中只會併入 true 值。這樣可能有助於使得規則更容易理解。選項不套用至交易式格式的資料。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。

註：如果欄位類型為旗標，而 Apriori 模型建置節點包括空記錄，則 CARMA 模型建置節點會在建置模型時忽略空記錄。空記錄是模型建置中使用的所有欄位都具有 false 值的記錄。

最佳化。 選取旨在基於特定需要而在模型建置期間提高效能的選項。

- 選取**速度**以指示演算法永不為了增進效能而使用磁碟溢出。

- 選取**記憶體**來指示演算法在適當時以犧牲速度為代價使用磁碟溢出。依預設會選取這個選項。

  註：以分散式方式執行時期，*options.cfg* 檔案中指定的管理者選項可能會置換此設定。 有關進一步資訊，請參閱《*IBM SPSS Modeler Server* 管理者手冊》。

Apriori 節點專家選項

對於那些詳細瞭解 CARMA 節點作業的人員來說，通過下列專家選項可以對歸納過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

評估測量。Apriori 支援五種評估潛在規則的方法。

- **規則信賴度。**此預設方法使用規則的信賴度（或精確度）來評估規則。對於此測量，**測量下限**為停用狀態，因為此選項對於「模型」標籤上的**規則信賴度最小值**選項來說是多餘的。請參閱第 219 頁的『Apriori 節點模型選項』主題，以取得更多資訊。
- **信賴度差異。**（也稱為**與事前相比的絕對信賴度差異**。）此測量是規則的信賴度與其事前信賴度之間的絕對差。此選項會防止出現偏移，即結果分佈不均勻。這有助於防止遵守「明顯的」規則。例如，可能會有 80% 的客戶購買您最受歡迎的產品。某項以 85% 的精確度預測購買該受歡迎產品的規則不會使您的瞭解加深，盡管 85% 的精確度對於絕對尺度來說似乎已經相當不錯。請將該評估測量下限設定為您希望保留的規則的信賴度下限差。
- **信賴度比率。**（也稱為**信賴度商數與 1 之間的差**。）此測量為 1 減去規則信賴度與事前信賴度之間的比（如果該比例大於一，則減去其倒數）。與信賴度差異相似，此方法會考慮不均勻分佈。此方法尤其適用於尋找預測小概率事件的規則。例如，假設有一種罕見的病理狀況只在 1% 的病人中出現。如果某個規則有 10% 的幾率預測出這種病理狀況，那麼它與隨機猜測相比是一種很大的提高，盡管從絕對尺度角度來看 10% 的精確度好像非常不起眼。請將該評估測量下限設定為您希望保留的規則的最小差。
- **資訊差異。**（也稱為**與事前的資訊差異**。）此測量基於 **資訊增益** 測量。如果某個特定結果的機率被視為一個邏輯值（一個**位元** ），則資訊增益為基於條件可以確定的該位元的比例。資訊差異是給定條件的情況下資訊增益與只給定了結果的事前信賴度的情況下資訊增益之間的差。此方法的一個重要特徵在於，它考慮了支援，因此對於給定層次的信賴度，它傾向於覆蓋更多記錄的規則。將評估測量下界設為想讓規則儲存的資訊差異。

  附註：此測量的尺度不像其他尺度那麼直觀，因此您可能需要試驗不同的下限才能取得滿意的規則集。
- **正規化的卡方。**（也稱為**正規化的卡方測量方式**。）此測量是條件與結果之間關聯的一個統計學指數。此測量進行了正規化，採用 0 和 1 之間的值。此測量尺度甚至比資訊差異測量更依賴於支援。將評估測量下界設為想讓規則儲存的資訊差異。

  附註：與資訊差異測量相同，此測量的尺度不像其他尺度那麼直觀，因此您可能需要試驗不同的下限才能取得滿意的規則集。

容許沒有先行的規則。 選取以容許僅包括後繼（項目或項目集）的規則。當您對決定一般項目或項目集感興趣時，這會非常有用。例如，cannedveg 是沒有先行的單項規則，指出購買 *cannedveg* 一般會在資料中出現。在某些情況下，如果您只是對大部分確信的預測感興趣，則可能想要包括這類規則。依預設，此選項處於關閉狀態。按照慣例，沒有條件的規則的先例支援表示為 100%，規則支援度與信賴度相同。

CARMA 節點

CARMA 節點使用相關規則探索演算法來探索資料中的相關規則。相關規則是下列形式的陳述式：

if *antecedent(s)* **then** *consequent(s)*

例如，如果某個 Web 客戶購買了無線網卡和高端無線路由器，那麼該客戶還可能購買無線音樂伺服器（如果提供該產品的話）。CARMA 模型會從資料中擷取一組規則，而不需要您指定輸入或目標欄位。這就意味著產生的規則可用於很多種應用程式。例如，您可以使用此節點產生的規則來尋找一系列產品或服務（前提條件），其結果是您要在此假期內進行促銷的商品。使用 IBM SPSS Modeler，您可以確定哪些用戶購買了這些條件產品，然後舉辦一個旨在促銷這些結果產品的營銷活動。

需求。 與 Apriori 不同，CARMA 節點不需要輸入欄位或目標欄位。這是該演算法工作方式的重要組成部分，相當於在將所有欄位設定為兩者的情況下建立 Apriori 模型。您可以通過在建立模型後對該模型進行過濾來限制僅作為前提條件或結果列出的項目。例如，您可以使用模型瀏覽器來尋找一系列產品或服務（條件），其結果是您要在此假期內進行促銷的項目。

要建立 CARMA 規則集，您需要指定一個 ID 欄位以及一個或多個內容欄位。ID 欄位可以具有任何角色或測量層次。會忽略角色為無的欄位。執行節點之前欄位類型必須完全實例化。與 Apriori 相似，資料可以是表格式，也可以是交易格式。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。

強度。 CARMA 節點基於 CARMA 關聯規則演算法。與 Apriori 相比，CARMA 節點為規則支援度（對前提條件和結果的支援）提供建立設定，而不是為前提條件支援提供建立設定。CARMA 還容許帶有多個結果的規則。與 Apriori 相似，CARMA 節點產生的模型可以插入到資料串流中用來建立預測。請參閱第 31 頁的『模型塊』主題，以取得更多資訊。

CARMA 節點欄位選項

執行 CARMA 節點之前，必須在 CARMA 節點的「欄位」標籤上指定輸入欄位。雖然大多數建模節點的欄位標籤選項都相同，但 CARMA 節點有幾個獨特的選項。所有選項均在下方討論。

使用「類型」節點設定。此選項用於告知節點使用上游類型節點中的欄位資訊。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。選取了此選項之後，請根據您要讀取交易格式的資料還是表格式的資料來指定下方的欄位。

使用交易式格式。 此選項將根據您的資料是交易處理格式還是表格式來變更此對話框中的其他欄位控制項。如果您使用多個具有交易式資料的欄位，則在某個特定記錄的這些欄位中指定的項目，會被視為代表在單一交易（具有單一時間戳記）中找到的項目。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。

表格資料

如果未選取**使用交易格式**，則顯示下列欄位。

- **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。
- **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。透過使用一個樣本來產生模型，並使用另一個樣本來測試模型，您可以很好地指出模型將概化為與現行資料相似的更大型資料集的程度。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。）

交易資料

如果選中了**使用交易格式**，則顯示下列欄位。

- **ID。** 對於交易處理資料，請從清單中選取 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）。

- **ID 是連續的。** （僅限 Apriori 和 CARMA 節點）如果您的資料進行了預先排序，以便所有 ID 相同的記錄在資料串流中群組在一起，那麼選取此選項可以加快正在處理速度。如果您的資料未經預先排序（或者您不確定），請將此選項保持未選取狀態，那麼該節點將自動對資料進行排序。

  註：如果您的資料未經過排序而您選取了此選項，那麼可能會在模型中得到無效結果。
- **內容。** 指定模型的內容欄位。這些欄位包含與關聯建模有關的項目。您可以指定多個旗標欄位（如果資料為表格式）或者一個列名欄位（如果資料為交易格式）。

CARMA 節點模型選項

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。

最低規則支援度 (%)。您可以指定支援準則。**規則支援度**指的是訓練資料中包含整個規則的 ID 所佔的比例。（請注意，此支援定義不同於 Apriori 節點中使用的前提條件支援。）如果您要關注更常見的規則，請增大此設定的值。

最低規則信賴度 (%)。您可以指定針對在規則集中保留規則的信賴度準則。**信賴度**是指預測正確的 ID 在所有由規則進行了預測的 ID 中所佔的百分比。基於訓練資料，該百分比的計算如下：包含整個規則的 ID 數量除以其中包含條件的 ID 數量。會捨棄信賴度低於指定準則的規則。如果您獲得的規則無關或者太多，請嘗試增加此設定。如果您獲得的規則太少，請嘗試降低此設定。

註：必要的話，您可以在您的專屬值中強調顯示值和類型。請注意，如果您將信賴度值降低到 1.0 以下，則除了處理程序需要大量可用記憶體之外，您還可能發現需要花費極長時間才能建置規則。

最大規則尺寸。您可以設定規則中不同項目集（而不是項目）的最大數量。如果相關規則相對較短，那麼可以降低此設定，以加快規則集建立速度。

註：如果欄位類型為旗標，而 Apriori 模型建置節點包括空記錄，則 CARMA 模型建置節點會在建置模型時忽略空記錄。空記錄是模型建置中使用的所有欄位都具有 false 值的記錄。

CARMA 節點專家選項

對於那些詳細瞭解 CARMA 節點作業的人員來說，通過下列專家選項可以對建模過程進行微調。要存取專家選項，請將「專家」標籤上的「模式」設定為**專家**。

排除具有多個結果的規則。選取此選項可以排除「雙頭」結果，即包含兩個項目的結果。例如，規則 `bread & cheese & fish -> wine &fruit` 包含一個雙頭結果 `wine &fruit` 。預設情況下，這樣的規則併入在內。

設定刪改值。 為了節省記憶體，所使用的 CARMA 演算法在正在處理期間會定期從其潛在項目集的清單中移除（**刪改**）不常用的項目集。選取此選項可調整刪改頻率，您指定的數字將決定刪改頻率。輸入較小的值來減少演算法的記憶體需求（但可能會增加所需的訓練時間），或輸入較大值來加快訓練速度（但可能會增加記憶體需求）。預設值是 500。

變換支援度。選取該選項目會排除因為納入不平均而好像表現為非常頻繁的不頻繁項目集合，從而增加效率。這是通過這樣的方式實現的：首先從較高的支援層次開始，然後逐漸下降到「模型」標籤上指定的層次。對於**交易的估計數量** 輸入一個值可指定支援層次應採用的下降速度。

容許沒有先行的規則。 選取以容許僅包括後繼（項目或項目集）的規則。當您對決定一般項目或項目集感興趣時，這會非常有用。例如，`cannedveg` 是沒有先行的單項規則，指出購買 *cannedveg* 一般會在資料中出現。在某些情況下，如果您只是對大部分確信的預測感興趣，則可能想要包括這類規則。預設不選取此選項。

關聯規則模型塊

關聯規則模型塊代表由下列關聯規則建模節點之一所探索的規則：

- Apriori
- CARMA

模型塊包含建模期間擷取自資料的規則的相關資訊。

註：如果未按 ID 對交易資料進行排序，那麼關聯規則塊評分可能不正確。

檢視結果

您可以使用該對話框上的「模型」標籤來瀏覽關聯模型（Apriori 和 CARMA）以及序列模型產生的規則。在生成新節點或對模型評分之前瀏覽模型塊會使您看到規則的相關資訊，還會提供用於過濾結果和對結果進行排序的選項。

為模型評分

精煉模型塊（Apriori、CARMA 和序列）可以新增到串流中，用於進行評分。請參閱第 41 頁的『使用串流中的模型塊』主題，以取得更多資訊。用來評分的模型塊在其各自的對話框上包括額外的「設定」標籤。請參閱第 226 頁的『關聯規則模型塊設定』主題，以取得更多資訊。

無法以其原始格式將未優化模型塊用於評分。而您可以產生一個規則集，並將該規則集用於評分。請參閱第 227 頁的『從關聯模型塊產生規則集』主題，以取得更多資訊。

關聯規則模型塊詳細資料

在關聯規則模型塊的「模型」標籤上，您可以看到一個表格，其中包含了該演算法所擷取的規則。表格中的每列都代表一個規則。第一欄代表結果（規則的 "then" 部分），而下一欄代表條件（規則的 "if" 部分）。後面的欄包含規則資訊，如信賴度、支援和提升。

相關規則通常下列表格中的格式顯示。

表 *15.* 關聯規則範例

後果	前提條件
Drug = drugY	Sex = F BP = HIGH

該範例規則的解釋為：如果 *Sex = "F"* 且 *BP = "HIGH"*，則 *Drug* 很可能為 *drugY*；或者以另一種方式解釋：對於 *Sex = "F"* 且 *BP = "HIGH"* 的記錄，*Drug* 很可能為 *drugY*。使用對話框工具列，可以選擇顯示其他資訊，如信賴度、支援和實例數。

排序功能表。 工具列上的「排序」功能表按鈕控制規則的排序。排序方向（遞增或遞減）可以使用排序方向按鈕（向上或向下箭頭）變更。

您可以依下列條件排序規則：

- 支援
- 信賴度
- 規則支援
- 後果

- 評估
- 增譯
- 可部署性

「顯示/隱藏」功能表。「顯示/隱藏」功能表（準則工具列按鈕）用於控制規則的顯示選項。

圖 *46.* 「顯示/隱藏」按鈕

下列顯示選項可用：

- **規則 ID**，顯示模型建置期間分配的規則 ID。通過規則 ID，可以識別哪些規則要套用於某個給定的預測。通過規則 ID，還可以在以後合併附加的規則資訊，如可部署性、產品資訊或條件。
- **實例數**，顯示規則所適用的唯一 ID 數（即，前提條件為 true 的 ID）的相關資訊。例如，假設規則為 bread -> cheese，訓練資料中包含條件 *bread* 的記錄數量稱為**實例數**。
- **支援**，顯示前提條件支援，即其前提條件為 true 的 ID 在訓練資料中所佔的比例。例如，如果 50% 的訓練資料包含 bread（面包）的購買，那麼規則 bread > cheese 的先例支援為 50%。註：此處定義的支援與實例數相同，但以百分比的形式代表。
- **信賴度**，顯示規則支援度與前提條件支援的比例。此值指出帶有指定條件、並且其結果也為 true 的 ID 的比例。例如，如果 50% 的訓練資料包含 bread（面包）（表明條件支援），但只有 20% 既包含 bread（面包）又包含 cheese（奶酪）（表明規則支援度），則規則 bread - > cheese 的信賴度為 規則支援度/先例支援 ，在這裡為 40%。
- **規則支援度**，顯示其整個規則、前提條件和結果均為 true 的 ID 所佔的比例。例如，如果 20% 的訓練資料既包含 bread（面包）又包含 cheese（奶酪），那麼規則 bread - > cheese 的規則支援為 20%。
- **評估**，如果選取其中一個專家關聯規則準則（信賴度差異、信賴度比率、資訊差異或正規化的卡方），則包含此選項。這些專家準則測量會與使用者設定的**評估測量下限**數進行比較（且僅在選取專家準則規則時適用）。「評估」統計量對於每一個專家關聯規則準則具有下列意義：
  - 信賴度差異：事後信賴度 - 事前信賴度
  - 信賴度比率：（事後信賴度 - 事前信賴度）/事後信賴度
  - 資訊差異：資訊增益測量
  - 正規化的卡方：正規化的卡方統計量

  上述每一個統計量都會與使用者設定的**評估測量下限**數進行比較，如果統計量超過此數，則會選取一個規則。
- **提昇**，顯示規則信賴度與具有結果的事前機率的比例。例如，如果整個人口統計中 10% 購買了 bread（面包），那麼預測人們是否購買 bread（面包）、信賴度為 20% 的規則具有的提升將為 20/10 = 2。如果另一個規則告訴您人們將購買 bread（面包），並且信賴度為 11%，則該規則的提升接近 1，這就意味著具有條件對於具有結果的機率不會造成太大的影響。總之，提升度不為 1 的規則比提升度接近 1 的規則的相關性更強。
- **可部署性**，這是對訓練資料中滿足前提條件但不滿足結果的部分所佔百分比的測量。在產品購買領域，它的意思大致為：總的客戶群中有多少百分比擁有了（或已經購買了）條件，但尚未購買結果。可部署性統計資料定義為（（以記錄數表示的先例支援 - 以記錄數表示的規則支援度）/記錄數）*100，其中先例支援表示其條件為 true 的記錄數，規則支援度表示條件和結果都為 true 的記錄數。

過濾器按鈕。 功能表上的「過濾器」按鈕（漏斗形圖示）會展開對話框的底端，以顯示其中顯示作用中規則過濾器的畫面。使用過濾器來減少「模型」標籤上顯示的規則數。

圖 *47.* 過濾器按鈕

若要建立過濾器，請按一下已展開畫面右側的「過濾器」圖示。這會開啟個別的對話框，您可以在其中指定顯示規則的限制。請注意，「過濾器」按鈕通常與「產生」功能表一起使用，先過濾規則，然後產生包含該規則子集的模型。如需相關資訊，請參閱下方的『為規則指定過濾器』。

「尋找規則」按鈕。「尋找規則」按鈕（望遠鏡圖示）使您可以搜尋對指定規則 ID 顯示的規則。相鄰的顯示框指示可用規則數目中目前顯示的規則數目。規則 ID 由模型按照探索時間的順序指定，並且會在評分期間新增到資料中。

圖 *48.*「尋找規則」按鈕

要對規則 ID 重新排序：

1. 您可以在 IBM SPSS Modeler 中對規則 ID 進行重新排序，方法是，首先根據所需的測量標準（如信賴度或提升）對規則顯示表格進行排序。
2. 然後使用「產生」功能表中的選項，建立一個過濾的模型。
3. 在「已過濾的模型」對話框中，選取**對規則重新進行連續編號的起始號碼**，然後指定一個開始號碼。

如需相關資訊，請參閱第 228 頁的『產生已過濾的模型』。

為規則指定過濾器

依預設，規則演算法（如 Apriori、CARMA 和序列）可能會產生非常大量的規則。為了在瀏覽時增強明確度，或者為了簡化規則評分，您應該考慮過濾規則，以便更加顯著地顯示相關的結果和條件。使用規則瀏覽器「模型」標籤上的過濾選項，可以開啟一個用於指定過濾條件的對話框。

後繼。 選取**啟用過濾器**可啟動根據併入或排除指定結果來對規則進行過濾的選項。選取**包含任意**可建立一個過濾器，該過濾器中的規則至少包含一個指定結果。另外，選取**排除**可建立一個排除指定結果的過濾器。您可以使用清單框右側的選取器圖示選取結果。這樣將開啟一個對話框，其中列出產生的規則中包含的所有結果。

註：結果可能包含多個項目。過濾器只會檢查結果是否包含一個指定項目。

先行。 選取**啟用過濾器**可啟動根據併入或排除指定前提條件來對規則進行過濾的選項。您可以使用清單框右側的選取器圖示選取項目。這樣將開啟一個對話框，其中列出產生的規則中包含的所有條件。

- 選取**併入所有**可將過濾器設定為一個併入過濾器，其中的規則必須併入指定的所有條件。
- 選取**包含任意**可建立一個過濾器，該過濾器中的規則至少包含一個指定條件。
- 選取**排除**可建立一個排除包含指定條件的規則的過濾器。

信賴度。 選取**啟用過濾器**可啟動根據規則的信賴度層次來對規則進行過濾的選項。您可以使用 **下限** 和 **上限** 控制項來指定信賴度範圍。當您瀏覽已產生的模型時，信賴度將以百分比的形式列出。當您對輸出評分時，信賴度則表示為一個介於 0 和 1 之間的數字。

前提條件支援。選取**啟用過濾器**可啟動根據規則的前提條件支援層次來對規則進行過濾的選項。前提條件支援指示包含與目前規則相同前提條件的訓練資料比例，因此與普及性指數有點類似。您可以使用**下限**和**上限**控制項，根據支援層次來指定用於過濾規則的範圍。

提昇。選取**啟用過濾器**可啟動根據規則的提昇測量來對規則進行過濾的選項。註：提昇過濾僅可用於發行版 8.5 之後建立的關聯模型，或包含提昇測量的先前版本的模型。序列模型不包含此選項。

按一下**確定**可套用已在此對話框中啟用的所有過濾器。

為規則產生圖形

關聯節點提供了大量資訊，但對業務使用者來說，它可能並不始終是一種方便存取的格式。若要提供資料以便可以將資料輕易納入商業報告、簡報等，您可以產生所選資料的圖形。從「模型」標籤上，可以為選定規則產生圖表，從而只為該規則中的觀察值建立圖表。

1. 在「模型」標籤上，選取感興趣的規則。
2. 從「產生」功能表中，選擇**圖表（從選定內容）**。即會顯示「圖表板基本」標籤。

   附註：以此方式顯示圖形板時，只有「基本」及「詳細」標籤可用。
3. 使用「基本」或「詳細」標籤設定來指定要顯示在圖形上的詳細資料。
4. 按一下「確定」以產生圖形。

圖形標題識別選擇要併入的規則和前提條件詳細資料。

關聯規則模型塊設定

此「設定」標籤用於為關聯模型（Apriori 和 CARMA）指定評分選項。此標籤僅在模型塊新增到用於評分的串流後才可用。

註：用於瀏覽非精簡 模型的對話框不包含「設定」標籤，因為無法對此模型進行評分。 要對 「非精簡」模型進行評分，您必須先產生規則集。 請參閱第 227 頁的『從關聯模型塊產生規則集』主題，以取得更多資訊。

預測的最大數量：指定為每個購物籃項目集合併入的預測的最大數量。 此選項與下方的「規則準則」一起使用可生成「最佳」預測，其中 最佳指的是信賴度、支援、提升等的最高層次，如下方的內容所述。

規則準則：選取用於確定規則強度的測量。規則按照此處選取的準則強度進行排序，以便傳回項目集合的最佳預測。下列清單中顯示了可用準則。

- 信賴度
- 支援
- 規則支援度（支援 * 信賴度）
- 增譯
- 可部署性

容許重複預測：選取此選項可在評分時包含具有相同結果的多項規則。 例如，選取此選項可容許對下列規則進行評分：

```
麵包 & 芝士 -> 紅酒
芝士 & 水果 -> 紅酒
```

關閉此選項可在評分時排除重複的預測。

註：僅在所有後繼 (`wine & pate`) 之前都已預測的情況下，將包含多重後繼的規則 (`bread & cheese & fruit -> wine & pate`) 視為重複預測。

忽略不符合的購物籃項目：選取此選項可忽略項目集中附加項目的存在。 例如，如果對於包含 [tent & sleeping bag & kettle] 的購物籃選取了此選項，規則 tent & sleeping bag - > gas_stove 套用時則會忽略該購物籃中出現的額外項目 (kettle)。

可能存在一些情況應該排除額外的項目。例如，購買了 tent（帳篷）、sleeping bag（睡袋）和 kettle（水壺）的某人可能已經擁有了 gas stove（燃氣爐），這可由存在水壺指出。換句話說，gas stove（燃氣爐）可能不是最佳預測。這種情況下，您應該取消選取**忽略不符合的購物籃項目**以確保規則條件與購物籃內容精確符合。依預設，不相符的項目將被忽略。

核實預測不存在於購物籃中。選取此選項可確保結果也不出現於購物籃中。例如，如果進行評分的目的是為了進行一項傢具產品推薦，那麼已經包含餐桌的購物籃可能不會購買另一個這樣的傢具。這種情況下，您應該選取此選項。另一方面，如果是易變質或一次性產品（如奶酪、嬰兒配方奶粉或者衛生紙），那麼結果已出現於購物籃的規則可能有些價值。在後一種情況下，最有用的選項可能是下方的 **不檢查購物籃中是否存在預測值** 。

檢查購物籃中是否出現預測值：選取此選項可確保結果也出現於購物籃中。 當您嘗試深入瞭解現有客戶或交易時，此方法很有用。例如，您可能希望確定提升最高的規則，然後探索哪些客戶符合這些規則。

不檢查購物籃中是否存在預測值：選取此選項可在評分時包含所有規則，而與購物籃中是否存在結果無關。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

關聯規則模型塊概要

關聯規則模型塊的「概要」標籤顯示探索的規則數目量，以及規則集中規則的上限和下限支援、提升值、信賴度和可部署性。

從關聯模型塊產生規則集

關聯模型塊（如 Apriori 和 CARMA）可用於直接對資料分數，您也可以首先產生一個規則子集合，稱為 **規則集** 。當您使用無法直接用於評分的未精簡模型時，規則集特別有用。請參閱第 44 頁的『未優化模型』主題，以取得更多資訊。

要產生規則集，請從模型塊瀏覽器的「產生」功能表中選擇**規則集**。您可以指定下列選項，將規則轉換為規則集：

規則集名稱。通過此選項，您可以指定新產生的「規則集」節點的名稱。

建立節點位置。 控制項新產生的「規則集」節點的位置。選取**畫布**、**GM 選用區**或**兩者**。

目標欄位。 確定哪個輸出欄位將用於產生的「規則集」節點。從清單中選取一個輸出欄位。

最小支援。 指定要在產生的規則集中保留規則的最小支援。支援少於指定值的規則將不會包含在新的規則集中。

最小信賴度。 指定要在產生的規則集中保留規則的最小信賴度。信賴度少於指定值的規則將不會包含在新的規則集中。

預設值。 通過此選項，您可以為分配到不會發動任何規則的已評分記錄的目標欄位指定預設值。

產生已過濾的模型

要從關聯模型塊（如 Apriori、CARMA 或序列規則集節點）產生已過濾的模型，請從模型塊瀏覽器的「產生」功能表中選擇**已過濾的模型**。這樣將建立一個子集模型，其中只包含瀏覽器中目前顯示的那些規則。註：無法對未優化模型產生已過濾的模型。

您可以指定下列用於過濾器規則的選項：

新模型的名稱。通過此選項，您可以指定新的「已過濾模型」節點的名稱。

建立節點位置。 控制項新的「已過濾模型」節點的位置。選取**畫布**、**GM 選用區**或**兩者**。

規則編號。指定規則 ID 在已過濾模型所包含的規則子集合中的編號方式。

- **保留原始規則 ID 編號。**選取此選項可以保持原始的規則編號。依預設，會為規則提供一個與演算法探索它們的順序相對應的 ID。該順序可能會因所採用演算法的不同而有所差別。
- **對規則重新進行連續編號的起始號碼。**選取此選項可以為已過濾規則指定新的規則 ID。新的 ID 將根據「模型」標籤上規則瀏覽器表格中顯示的排序進行指定，從您在此處指定的數字開始。您可以使用右側的箭頭指定 ID 的開始號碼。

相關規則評分

通過關聯規則模型塊執行新資料生成的分數會傳回到不同的欄位中。對於每個預測會新增三個新欄位，其中 *P* 代表預測， *C* 代表信賴度， *I* 代表規則 ID。這些輸出欄位的排列取決於輸入資料是交易格式還是表格式。請參閱第 218 頁的『表格資料與交易資料』以獲取有關這些格式的概觀。

例如，假設您要使用一個基於下面三個規則產生預測的模型對購物籃資料進行評分：

```
Rule_15  bread&wine -> meat (confidence 54%)
Rule_22  cheese -> fruit (confidence 43%)
Rule_5  bread&cheese -> frozveg (confidence 24%)
```

表格資料。對於表格資料，這三個預測（3 為預設值）將在單一記錄中傳回。

表 *16.* 表格式的評分

ID	麵包	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	肉類	0.54	15	水果	0.43	22	frozveg	.24	5

交易資料。對於交易資料，將對每個預測產生一個個別的記錄。預測仍然會新增到個別的欄中，但分數在計算時傳回。這樣會生成帶有不完整預測的記錄，如下方的示例輸出所示。第二個和第三個預測（P2 和 P3）以及相關聯的信賴度和規則 ID 在第一個記錄中空白。但傳回分數時，最後一個記錄將包含所有三個預測。

表 *17.* 交易式格式的評分

ID	項目	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	肉類	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	芝士	肉類	0.54	14	水果	0.43	22	$null$	$null$	$null$
Fred	葡萄酒	肉類	0.54	14	水果	0.43	22	frozveg	0.24	5

要只包含用於報告或部署目的的完整預測，請使用選取節點選取完整的記錄。

註：為了清楚起見，這些範例中使用的欄位名稱都是縮寫。在實際應用中，關聯模型的結果欄位將按下表格所示進行命名。

表 *18.* 關聯模型的結果欄位的名稱

新欄位	欄位名稱範例
預測	*$A-TRANSACTION_NUMBER-1*
信賴度（或其他準則）	*$AC-TRANSACTION_NUMBER-1*
規則 ID	*$A-Rule_ID-1*

帶有多個結果的規則

CARMA 演算法容許規則帶有多個結果，例如：

```
bread -> wine&cheese
```

對這樣的「雙頭」規則進行評分時，預測將下列表格中顯示的格式傳回。

表 *19.* 包括具有多重後繼的預測的評分結果

ID	麵包	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat&veg	0.54	16	水果	0.43	22	frozveg	.24	5

在某些情況下，您可能需要在部署之前分割這樣的分數。要分割帶有多個結果的預測，您需要使用 CLEM 字串函數剖析該欄位。

部署關聯模型

對關聯模型進行評分時，預測和信賴度將輸出到個別的欄中（其中 *P* 代表預測， *C* 代表信賴度， *I* 代表規則 ID）。這種情況要區分輸入資料是表格式還是交易格式。請參閱第 228 頁的『相關規則評分』主題，以取得更多資訊。

準備分數進行部署時，您可能會發現您的應用程式需要將輸出資料轉換為預測位於列中的格式，而不是位於欄中的格式（每列一個預測，有時稱為「列窮盡」格式）。

轉置表格分數

您可以使用 IBM SPSS Modeler 中的一些步驟將表格分數從欄轉置為列，如下面的步驟所示。

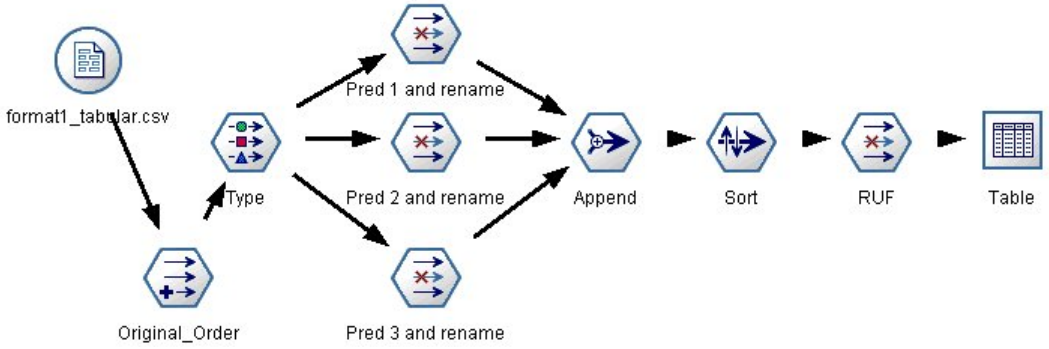


圖 *49.* 用於將表格資料變換為行窮盡格式的範例串流

1. 在衍生節點中使用 @INDEX 函數可確定預測的目前順序，並將此指示儲存在一個新欄位中，如 *Original_order*。
2. 新增一個類型節點，確保所有欄位均實例化。
3. 使用過濾器節點將預設的預測、信賴度和 ID 欄位（ *P1*、*C1*、*I1*）更名為普通欄位，如 *Pred*、 *Crit* 和 *Rule_ID*，這些欄位將用於在以後附加記錄。對於每個產生的預測都需要一個過濾節點。

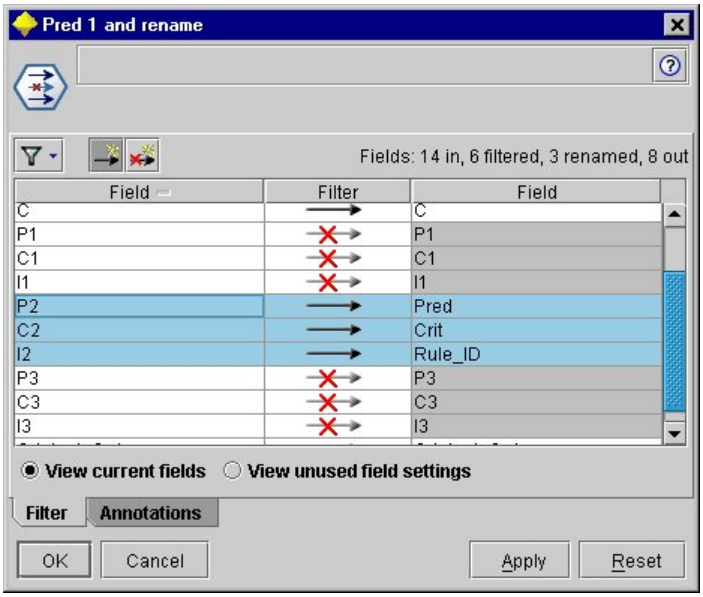


圖 *50.* 更名預測 *2* 的欄位時過濾預測 *1* 和預測 *3* 的欄位。

4. 使用附加節點附加共用 *Pred*、 *Crit* 和 *Rule_ID* 的值。
5. 連接一個排序節點，以便按照欄位 *Original_order* 的升冪對記錄進行排序，按照 *Crit* 的遞減對記錄進行排序，後面一個欄位是用於按準則（如信賴度、提升和支援）對預測進行排序的欄位。
6. 使用另一個過濾器節點將欄位 *Original_order* 從輸出中過濾掉。

此時，資料就可以進行部署了。

轉置交易分數

轉置交易分數的過程與上面的過程相似。例如，下方顯示的串流會根據部署需要，將分數轉置為每列一個預測的格式。

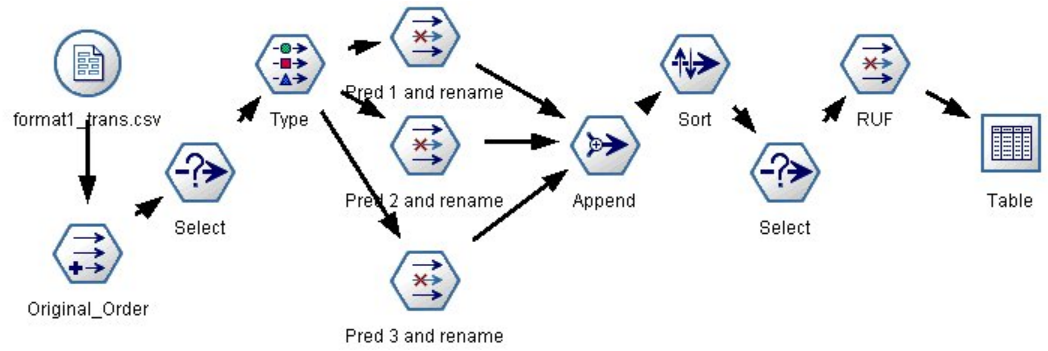


圖 *51.* 用於將交易資料變換為行窮盡格式的範例串流

除了新增兩個「選取」節點之外，該過程與先前對表格資料說明的過程完全相同。

- 第一個選取節點用於對相鄰記錄的規則 ID 進行比較，並且只包含唯一的或未定義的記錄。此「選取」節點使用以下 CLEM 表示式來選取記錄：`ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1)= undef`。
- 第二個選取節點用於放棄多餘的規則，或者 `Rule_ID` 為空值的規則。此「選取」節點使用下列 CLEM 表示式來捨棄記錄：`not(@NULL(Rule_ID))`。

有關轉置分數進行部署的詳細資訊，請聯絡技術支援部門。

順序節點

序列節點會探索循序資料或面向時間的資料中的型樣，其格式為 `bread - > cheese` 。序列的元素為組成一個交易的**項目集合**。例如，如果某人進入商店，購買了面包和牛奶，幾天之後傳回了該商店，購買了一些奶酪，那麼這個人的購買活動可以代表為兩個項目集合。第一個項目集合包含面包和牛奶，第二個包含奶酪。**序列**是一系列可能會以可預測順序發生的項目集合。序列節點會偵測頻繁出現的序列，並建立一個可用於產生預測的產生模型節點。

需求。 要建立序列規則集，您需要指定一個 ID 欄位、一個選用的時間欄位，以及一個或多個內容欄位。請注意，這些設定必須在建模節點的「欄位」標籤上進行；不能從上游「類型」節點中讀取。ID 欄位可以具有任何角色或測量層次。如果指定時間欄位，那麼該欄位可以是任意角色，但其儲存必須是數值、日期、時間或時間戳記。如果不指定時間欄位，序列節點則會使用隱含的時間戳記，實際上是使用列號作為時間值。內容欄位可具有任意測量層次和角色，但所有內容欄位的類型必須相同。如果這些字段是數值型的，那麼必須為整數範圍（不是實數範圍）。

強度。 「序列」節點基於 CARMA 關聯規則演算法，使用有效的兩段式方法來尋找序列。另外，序列節點建立的產生的模式節點可以插入到資料串流中來建立預測。產生的模式節點還可產生 SuperNode。用於偵測或計數特定的序列，以及基於特定的序列作出預測。

序列節點欄位選項

執行序列節點之前，必須在序列節點的「欄位」標籤上指定 ID 欄位和內容欄位。如果您要使用時間欄位，也需要在此處指定。

ID 欄位。 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）。

- **ID 是連續的。** 如果您的資料進行了預先排序，以便所有 ID 相同的記錄在資料串流中群組在一起，那麼選取此選項可以加快正在處理速度。如果您的資料未經預先排序（或者您不確定），請將此選項保持不已選取狀態，序列節點將自動對該資料進行排序。

註：如果您的資料未經過排序而您選取了此選項，那麼可能會在序列模型中得到無效結果。

時間欄位。如果您要在資料中使用欄位來指示事件時間，請選取**使用時間欄位**並指定要使用的欄位。時間欄位必須是數值、日期、時間或時間戳記。如果不指定時間欄位，那麼假設記錄按照從資料來源出發的順序到達，記錄編號將用作時間值（第一個記錄發生在時間 "1"；第二個記錄發生在時間 "2"；依此類推。）

內容欄位。指定模型的內容欄位。這些欄位包含與序列建模有關的事件。

序列節點可以處理表格式的資料，也可以處理交易格式的資料。如果您使用多個具有交易式資料的欄位，則在某個特定記錄的這些欄位中指定的項目，會被視為代表在單一交易（具有單一時間戳記）中找到的項目。請參閱第 218 頁的『表格資料與交易資料』主題，以取得更多資訊。

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。透過使用一個樣本來產生模型，並使用另一個樣本來測試模型，您可以很好地指出模型將概化為與現行資料相似的更大型資料集的程度。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。）

序列節點模型選項

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。

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。

最低規則支援度 (%) 您可以指定支援準則。規則支援指的是訓練資料中包含整個序列的 ID 所佔的比例。如果您要關注更常見的序列，請增加此設定。

最低規則信賴度 (%) 您可以指定針對在序列集中保留序列的信賴度準則。信賴度是指預測正確的 ID 在所有由規則進行了預測的 ID 中所佔的百分比。它會根據訓練資料，按為其尋找整個序列的 ID 數除以為其尋找先行的 ID 數的方式來計算。信賴度低於指定準則的序列將被放棄。如果您獲得的序列太多或者不是非常相關，請嘗試增加此設定。如果您獲得的序列太少，請嘗試降低此設定。

註：必要的話，您可以在您的專屬值中強調顯示值和類型。請注意，如果您將信賴度值降低到 1.0 以下，則除了處理程序需要大量可用記憶體之外，您還可能發現需要花費極長時間才能建置規則。

最大序列大小 您可以設定序列中不同值的最大數量。如果相關序列相對較短，那麼可以降低此設定，以加快序列集建立速度。

要新增至串流中的預測數 指定最終產生的「模型」節點要新增至串流中的預測的號碼。如需相關資訊，請參閱第 234 頁的『序列模型塊』。

序列節點專家選項

對於那些詳細瞭解序列節點作業的人員來說，通過下列專家選項可以對建模過程進行微調。若要存取專家選項，請在「專家」標籤上將「模式」設定為**專家**。

設定最大持續期。如果已選取了此選項，那麼會將序列限制為持續時間（第一個項目集與最後一個項目集之間的時間）少於或等於指定值的序列。如果沒有指定時間欄位，該持續時間則以原始資料中的列數（記錄數）表示。如果使用的時間欄位為時間、日期或時間戳記欄位，該持續時間則表示為秒數。對於數字欄位，持續時間則使用與欄位相同的單位數表示。

設定刪改值。 為了節省記憶體，「序列」節點中使用的 CARMA 演算法在正在處理期間會定期從其潛在項目集的清單中移除（**刪改**）不常用的項目集。選取此選項可調整刪改的頻率。指定的數字決定了刪改頻率。輸入較小的值來減少演算法的記憶體需求（但可能會增加所需的訓練時間），或輸入較大值來加快訓練速度（但可能會增加記憶體需求）。

設定記憶體中的最大序列。如果已選取了此選項，那麼 CARMA 演算法會將建模期間候選序列的記憶體儲存限制為指定的序列數。如果 IBM SPSS Modeler 在序列建模期間使用的記憶體過多，請選取此選項。請注意，您在此處指定的上限序列值指的是在建立模型期間進行內部追蹤的備選序列數。此數字應該比最終模型中預期的序列數大很多。

項目集之間的約束間隙。通過此選項，您可以指定對分隔項目集的時間間隙的限制。如果選取了此選項，那麼不會考慮時間間距小於您所指定的**下限間距**或大於**上限間距**的項目集作為序列的組成部分。使用此選項可避免考慮包含較長時間間隔或者在很短的時間跨距內發生的那些序列。

註：如果使用的時間欄位為時間、日期或時間戳記欄位，那麼時間間隙以秒為單位。對於數值型欄位，時間間距則使用與時間欄位相同的單位數表示。

例如，請考慮下面的交易清單。

表 *20.* 交易的範例清單

ID	時間(M)	內容
1001	1	apples
1001	2	bread
1001	5	芝士
1001	6	dressing

如果您針對這些資料建模時指定的下限間距為 2，那麼會得到下列序列：

```
apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing
```

您不會看到像 `apples - > bread` 這樣的序列，因為 `apples` 和 `bread` 之間的間距小於下限間距。同樣地，請考慮下列替代資料。

表 *21.* 交易的範例清單

ID	時間(M)	內容
1001	1	apples
1001	2	bread
1001	5	芝士
1001	20	dressing

如果上限間隙設定為 10，那麼您將不會看到任何帶有 dressing 的序列，因為 cheese 與 dressing 之間的間隙過大，而無法將它們視為同一序列的組成部分。

序列模型塊

「序列」模型塊代表「序列」節點針對某個特定輸出欄位探索的序列，可以新增至串流中以產生預測。

當您執行包含「序列」節點的串流時，「序列」節點會將包含預測的一對欄位，以及序列模型中每個預測的相關信賴度值新增到資料中。預設情況下，會新增包含三個最佳預測的三成對欄位（以及它們相關聯的信賴度值）。您既可以通過在建立時設定序列節點模型選項變更建立模型時產生的預測數，也可以在將模型塊新增到串流之後在「設定」標籤上變更此數量。請參閱第 236 頁的『序列模型塊設定』主題，以取得更多資訊。

新的欄位名稱衍生自模型名稱。預測欄位的欄位名稱為 *$S-sequence-n*（其中 *n* 表示第 *n* 個預測）信賴度欄位的欄位名稱為 *$SC-sequence-n*。在一個序列中具有多個序列規則節點的串流中，新的欄位名稱將在字首中包含數字，以便彼此區分開來。串流中的第一個「序列」節點將使用常用名稱，第二個節點將使用以 *$S1-* 和 *$SC1-* 開頭的名稱，第三個節點將使用以 *$S2-* 和 *$SC2-* 開頭的名稱，依此類推。預測按照信賴度的順序顯示，因此 *$S-sequence-1* 所包含預測的信賴度最高，*$S-sequence-2* 所包含預測的信賴度次高，依此類推。對於其中可用預測數量小於所已要求預測數量的記錄，剩餘的預測包含值 $null$。例如，如果對於某個特定的記錄只能進行兩個預測，那麼 *$S-sequence-3* 和 *$SC-sequence-3* 的值將為 $null$。

對於每條記錄，會將模型中的規則與目前對於目前 ID 已經處理的交易集合（包含現行記錄和具有相同 ID 和較早時間戳記的所有以前記錄）進行比較。將套用適用於此交易集合的、信賴度值最高的 *k* 個規則為該記錄產生 *k* 個預測，其中 *k* 為模型新增到串流之後在「設定」標籤上指定的預測數。（如果多個規則對於該交易集合預測了相同的結果，那麼只使用信賴度最高的規則。）請參閱第 236 頁的『序列模型塊設定』主題，以取得更多資訊。

與其他類型的關聯規則模型相同，資料格式必須與建立序列模型時使用的格式相符。例如，使用表格資料建立的模型只能用於對表格資料進行評分。請參閱第 228 頁的『相關規則評分』主題，以取得更多資訊。

註：在串流中使用產生的「序列集」節點對資料進行評分時，您在建立模型時選取的任何允差或間隙設定都將被忽略，不會用於評分目的。

根據序列規則進行的預測

該節點以與時間相關（如果在建立模型時未使用時間戳記欄位的話，那麼與順序相關）的方式處理記錄。記錄應該按照 ID 欄位和時間戳記欄位（如果出現的話）排序。但是，預測與新增到其中的記錄的時間戳記沒有關係。它們只是在給出到現行記錄為止目前 ID 的交易歷史的情況下，指出最可能在將來的某個時間出現的項目。

請注意，每條記錄的預測不一定與該記錄的交易相關。如果現行記錄的交易不觸發、觸發程式某個特定的規則，那麼會根據目前 ID 的以前交易選取規則。換句話說，如果現行記錄不向序列新增任何有用的預測資訊，那麼會將此 ID 的最後一個有用交易中的預測轉到現行記錄。

例如，假設您擁有的序列模型具有一個規則

```
Jam -> Bread (0.66)
```

然後您將其傳送到下列記錄。

表 *22.* 記錄範例

ID	購買	預測
001	jam	bread
001	牛奶	bread

請注意，與您的預期相同，第一個記錄產生了預測 *bread*。第二個記錄也包含 *bread* 預測，因為 *milk* 前面的 *jam* 沒有規則；因此， *milk* 交易不會增加任何有用資訊，所以規則 `Jam -> Bread` 仍然適用。

產生新節點

通過「產生」功能表可以基於序列模型建立新的 SuperNode。

- **規則 SuperNode。**建立一個可以偵測和計算已評分資料中序列發生次數的 SuperNode。如果未選取任何規則，則停用此選項。請參閱第 237 頁的『從序列模型塊產生規則 SuperNode』主題，以取得更多資訊。
- **模式至色板。** 將模型傳回到模型選用區。當同事寄給您一個包含模型的資料流，而非模式本身時，這個功能很有用。

序列模型塊詳細資料

序列模型塊的「模型」標籤顯示演算法擷取的規則。表格中的每列都代表一個規則，其中條件（規則的 "if" 部分）位於第一欄，結果（規則的 "then" 部分）位於後面的第二欄。

每項規則都下列列格式顯示。

表 *23.* 規則格式

先行	後果
`beer and cannedveg`	`beer`
`fish` `fish`	`fish`

第一個規則範例解釋為：對於在同一交易中具有 *"beer"* 和 *"cannedveg"* 的 *ID*，後面可能會出現 *"beer"*。第二個規則範例解釋為：對於在一個交易中具有 *"fish"*，而在另一個交易中也具有 *"fish"* 的 *ID*，後面可能會出現 *"fish"*。請注意在第一個規則中， *beer* 和 *cannedveg* 是同時購買的；在第二個規則中，*fish* 是在兩個不同的交易中購買的。

排序功能表。 工具列上的「排序」功能表按鈕控制規則的排序。排序方向（遞增或遞減）可以使用排序方向按鈕（向上或向下箭頭）變更。

您可以依下列條件排序規則：

- 支援度
- 信賴度
- 規則支援百分比
- 後果
- 第一個前提條件
- 最後一個前提條件
- 項目數目（前提條件）

例如，下表格按照項目數目，以遞減進行排序。條件集合中具有多個項目的規則排在條件集合中項目數較少的規則前面。

表 *24.* 依項目數排序的規則

先行	後果
`beer and cannedveg and frozenmeal`	`frozenmeal`
`beer and cannedveg`	`beer`

表 24. 依項目數排序的規則 (繼續)

先行	後果
fish fish	fish
softdrink	softdrink

顯示/隱藏準則功能表。顯示/隱藏準則功能表按鈕（網格圖示）控制項著規則的顯示選項。下列顯示選項可用：

- **實例數**顯示出現完整序列（同時包含前提條件和結果）的唯一 ID 數的相關資訊。（請注意，此內容與關聯模型不同，後者的實例數指的是其中 僅 條件適用的 ID 數。例如，假設規則為 bread - > cheese ，訓練資料中同時包含 *bread* 和 *cheese* 的記錄數稱為 **實例數** 。
- **支援**顯示訓練資料中前提條件為 true 的 ID 所佔的比例。例如，如果訓練資料中 50% 包含條件 *bread*，那麼規則 bread - > cheese 的支援為 50%。（與關聯模型不同，支援不基於實例數，如前面所述）。
- **信賴度**顯示預測正確的 ID 在所有由規則進行了預測的 ID 中所佔的百分比。它會根據訓練資料，按為其尋找整個序列的 ID 數除以為其尋找先行的 ID 數的方式來計算。例如，如果 50% 的訓練資料包含 cannedveg （表明條件支援），但只有 20% 既包含 cannedveg 又包含 frozenmeal , ，那麼規則 cannedveg - > frozenmeal 的信賴度為 規則支援度/先例支援 ，在這裡為 40%。
- 序列模型的**規則支援度**基於實例數，並顯示整個規則、前提條件和結果都為 true 的訓練記錄的所佔比例。例如，如果 20% 的訓練資料既包含 *bread* 也包含 *cheese*，那麼規則 bread - > cheese 的規則支援度為 20%。

請注意，這些比例基於有效交易（至少具有一個觀測項目或 true 值的交易），而不基於總的交易。在這些計算中不會考慮無效交易（沒有項目或 true 值的交易）。

過濾器按鈕。 功能表上的「過濾器」按鈕（漏斗形圖示）會展開對話框的底端，以顯示其中顯示作用中規則過濾器的畫面。使用過濾器來減少「模型」標籤上顯示的規則數。

圖 52. 過濾器按鈕

若要建立過濾器，請按一下已展開畫面右側的「過濾器」圖示。這會開啟個別的對話框，您可以在其中指定顯示規則的限制。請注意，「過濾器」按鈕通常與「產生」功能表一起使用，先過濾規則，然後產生包含該規則子集的模型。如需相關資訊，請參閱下方的第 225 頁的『為規則指定過濾器』。

序列模型塊設定

序列模型塊的「設定」標籤顯示模型的評分選項。此標籤僅在模型新增到串流畫布用於評分之後可用。

上限預測數。指定每個購物籃項目集合併入的預測的最大數量。套用至這個交易集且具有最高信賴度值的規則，用來為記錄（最多到指定的限制）產生預測。

序列模型塊概要

序列規則模型塊的「概要」標籤顯示探索的規則數量，以及規則的上限和下限支援和信賴度。如果您執行了附加至此建模節點的「分析」節點，則該分析中的資訊也將顯示在此區段中。

請參閱第 36 頁的『瀏覽模型塊』主題，以取得更多資訊。

從序列模型塊產生規則 SuperNode

要基於序列規則產生規則 SuperNode：

1. 在序列規則模型塊的「模型」標籤上，按一下表格中的某列以選取所需的規則。
2. 從規則瀏覽器功能表中選擇：

產生 > **規則 SuperNode**

重要：要使用產生的 SuperNode，您必須在將資料傳入 SuperNode 之前按 ID 欄位（和時間欄位，如果有的話）對資料進行排序。SuperNode 無法在未排序的資料中正確偵測序列。

您可以指定下列用於產生規則 SuperNode 的選項：

偵測。指定如何為傳入 SuperNode 的資料定義相符項。

- **僅前提條件。**每當 SuperNode 在具有同一識別的一組記錄中發現選定規則的前提條件的順序正確時，它都會確定一個相符項，而不考慮是否還找到了結果。請注意，此選項不考慮原始序列建模節點中的時間戳記允差或項目間距限制設定。在串流中偵測到最後一個條件項目集合（所有其他條件均以正確順序發現）最後一個，具有目前 ID 的所有最後一個續記錄都將包含下方選取的概要。
- **整個序列。**每當 SuperNode 在具有同一識別的一組記錄中發現選定規則的前提條件和結果的順序正確時，它都會確定一個相符項。此選項不考慮原始序列建模節點中的時間印記容忍度或項目間隙限制設定。在串流中偵測到最後一個結果（所有條件均以正確順序發現）後，現行記錄和具有目前 ID 的所有後續記錄都將包含下方選取的概要。

顯示。 控制項如何將相符項摘要新增到規則 SuperNode 輸出中的資料內。

- **首次出現的結果值。**新增到資料中的值是根據相符項的首次出現預測的結果值。這些值將作為一個名為 *rule_n_consequent* 的新欄位進行新增，其中 *n* 為規則編號（基於串流中規則 SuperNode 的建立順序）。
- **首次出現的 true 值。**如果對於該 ID 至少存在一個相符項，那麼新增到資料中的值為 true；如果沒有任何相符項，那麼新增的值為 false。這些值將作為一個名為 *rule_n_flag* 的新欄位新增。
- **發生計次。**新增到資料中的值為該 ID 的相符項數。值將新增為新欄位 *rule_n_count*。
- **規則編號。**新增的值為選定規則的規則編號。**規則編號**是根據 SuperNode 新增到串流中的順序指定的。例如，第一個規則 SuperNode 被視為 規則 *1*，第二個規則 SuperNode 被視為 規則 *2*，依此類推。當您要在串流中包含多個規則 SuperNode 時，此選項最有用。這些值將作為一個名為 *rule_n_number* 的新欄位新增。
- **納入信賴度指數。**如果已選取此選項，那麼會將規則信賴度以及選定概要新增到資料串流中。這些值將作為一個名為 *rule_n_confidence* 的新欄位新增。

關聯規則節點

相關規則是下列格式的陳述式。

例如，「如果顧客購買了剃須刀和須後水，那麼該顧客還會購買剃須膏，並且信賴度為 80%」。「相關規則」節點從資料中擷取一組規則，抽出的規則具有出現頻率最高的資訊內容。「相關規則」節點與 Apriori 節點非常類似，但是，存在一些明顯的差異：

- 「相關規則」節點無法處理交易性資料。
- 「相關規則」節點能夠處理儲存類型為「清單」且測量層次為「收集」的資料。
- 「相關規則」節點可以與 IBM SPSS Analytic Server 配合使用。這提供了可調整性，並且意味著您可以正在處理大型資料並利用速度更快的平行正在處理。
- 「相關規則」節點提供了更多設定，例如能夠限制產生的規則數目，從而增加正在處理速度。

- 模型塊的輸出將顯示在輸出檢視器中。

註：「相關規則」節點不支援 IBM SPSS Collaboration and Deployment Services 中的「模型評估」步驟或「冠軍挑戰者」步驟。

註：如果欄位類型為旗標，「相關規則」節點在建立模型時會忽略空白的記錄。 空記錄是模型建置中使用的所有欄位都具有 false 值的記錄。

在 IBM SPSS Modeler 安裝的 Demos 目錄中，提供了名為 geospatial_association.str 的串流，這個串流顯示了有關如何使用「相關規則」的有效範例，並參照資料檔案 InsuranceData.sav、CountyData.sav 和 ChicagoAreaCounties.shp。您可從 Windows「啟動」功能表的 IBM SPSS Modeler 程式集存取 Demos目錄。geospatial_association.str 檔案位於 streams 目錄中。

相關規則 - 欄位選項

在**欄位**標籤上，您可以選擇是使用上游節點（例如上一個「類型」節點）中已定義的欄位角色設定，還是手動進行欄位分配。

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（例如目標或預測值）。具有輸入角色的欄位被視為條件，具有目標角色的欄位被視為預測，而那些同時用作輸入和目標的欄位被視為具有這兩種角色。

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 如果您已選取了**使用自訂欄位指定**，那麼使用箭頭按鈕可以將此清單中的項目手動分配到畫面右側的框。圖示指出每個欄位的有效測量層次。

兩者（條件或預測）

新增到此清單中的欄位可以在模型所產生的規則中充當條件或預測角色。這基於每條規則，因此，某個欄位可能是一條規則中的條件，並且是另一條規則中的預測。

僅限預測

新增到此清單中的欄位只能顯示為規則的預測（也稱為「結果」）。欄位出現在此清單中並不表示在任何規則中使用了該欄位，只要使用了該欄位，那麼它只能是預測。

僅限條件

新增到此清單中的欄位只能顯示為規則的條件（也稱為「前提條件」）。欄位出現在此清單中並不表示在任何規則中使用了該欄位，只要使用了該欄位，那麼它只能是條件。

相關規則 - 規則建置

每規則項目

使用這些選項可以指定每條規則中可以使用的項目或值的數目。

註：這兩個欄位的總和不得超過 10。

條件數目上限

選取單條規則中可以包含的條件數目上限。

預測數目上限

選取單條規則中可以包含的預測的數目上限。

規則建置

使用這些選項可以指定要建立的規則的號碼和類型。

規則數目上限
指定在為模型建立規則時可以考慮使用的規則數目上限。

針對前 N 個的規則準則
選取用於建立前 N 條規則的準則，其中 N 是在**規則數目上限**欄位中已輸入的值。您可以從下列準則中選擇。

- **信賴度**
- **規則支援**
- **條件支援度**
- **增譯**
- **可部署性**

僅包含旗標變量的 true 值
當資料採用表格式時，選中此選項只會將旗標欄位的 true 值包含在生成的規則中。選取 true 值可能有助於使規則更容易理解。選項不套用至交易式格式的資料。如需相關資訊，請參閱第 218 頁的『表格資料與交易資料』。

規則準則

如果您選中**啟用規則準則**，那麼可以使用這些選項來選取最小強度，只有那些符合此強度的規則才能在模型中使用。

- **信任層次** 指定模型所生成的規則的信賴等級的下限百分比值。如果模型產生的規則層次小於此數量，則會捨棄該規則。
- **規則支援度** 指定模型所生成的規則的規則支援度層次的下限百分比值。如果模型產生的規則層次小於此數量，則會捨棄該規則。
- **條件支援度** 指定模型所生成的規則的條件支援度層次的下限百分比值。如果模型所生成的規則的條件支持度層次少於指定的數量，那麼該規則將被廢棄。
- **提昇** 指定模型所生成的規則容許的下限提昇值。如果模型所生成的規則的值少於指定的數量，那麼該規則將被廢棄。

排除規則

在某些情況下，兩個或兩個以上欄位之間的關聯已知或者不證自明，在這種情況下，您可以排除其中的欄位預測彼此的規則。通過排除包含這兩個值的規則，您可以減少不相關的輸入並增加找到有用發現項目的機會。

欄位 選取不想在規則建置中同時使用的關聯欄位。例如，關聯欄位可能是「製造商」和「汽車型號」，或者是「學年」和「學生時代」。當模型建立規則時，如果規則至少包含在規則的任一端（條件或預測）選取的其中一個欄位，那麼該規則將被廢棄。

相關規則 - 變換
分組

使用這些選項可以指定如何對連續（數值範圍）欄位進行 bin。

bin 數目
所有設定為自動 Bin 的連續欄位都劃分為您指定的間距相等的箱數。您可以選取 2 - 10 範圍內的任意數字。

清單欄位

最大清單長度

要在清單欄位的長度未知時限制要併入在模型中的項目數，請輸入上限清單長度。您可以選取 1 到 100 範圍內的任何數字。如果清單長度超過輸入的數字，那麼模型仍將使用此欄位，但僅包含截至此數字為止的值；此欄位中的所有其他值都將被忽略。

相關規則 - 輸出

使用此窗格中的選項可以控制在建立模型時所產生的輸出。

規則表格

使用這些選項可以建立一種或多種表格類型，這些表格類型針對每個選定的準則顯示最佳規則數目（基於您指定的號碼）。

信賴度

信賴度是規則支援度與條件支援度的比例。在具有列出的條件值的項目中，具有預測結果值的項目所佔的百分比。將建立一個包含最佳 N 條相關規則的表格，這些規則基於輸出中要併入的信賴度（其中 N 是**要顯示的規則數**值）。

規則支援

整個規則、條件和預測均為 true 的項目所佔的比例。對於資料集中所有的項目，規則所正確解釋並預測的百分比。此測量給出規則的整體重要性。將建立一個包含最佳 N 條相關規則的表格，這些規則基於輸出中要併入的規則支援度（其中 N 是**要顯示的規則數**值）。

增譯 規則信賴度與具有預測的事前機率的比例。規則的信賴度值與結果值出現在整體中的百分比的比例。此比例測量規則對機會的改進程度。將建立一個包含最佳 N 條相關規則的表格，這些規則基於輸出中要併入的提昇（其中 N 是**要顯示的規則數**值）。

條件支援度

條件為 true 的項目所佔的比例。將建立一個包含最佳 N 條相關規則的表格，這些規則基於輸出中要併入的前提條件支援（其中 N 是**要顯示的規則數**值）。

可部署性

用於測量訓練資料中滿足條件但不滿足預測的部分所佔的百分比。此測量顯示規則未命中的頻率。它實際上與信賴度是相反的。將建立一個包含最佳 N 條相關規則的表格，這些規則基於輸出中要併入的可部署性（N 是**要顯示的規則數**值）。

要顯示的規則數

設定表格中要顯示的最大規則數目。

模型資訊表格

使用這些選項中的一個或多個選項可以選取輸出中要包含的模型表格。

- **欄位轉換**
- **記錄摘要**
- **規則統計資料**
- **最頻率的值**
- **最頻率的欄位**

規則的可排序的單字雲。

使用這些選項可以建立顯示規則輸出的單字雲。單字以不斷遞增的文字大小顯示，以表明其重要性。

建立可排序的單字雲。

選中此框將在輸出中建立可排序的單字雲。

預設排序

選取最初建立單字雲時要使用的排序類型。單字雲是互動式的，您可以在模型檢視器中變更準則以檢視不同的規則和排序。可以從下列排序選項中選擇：

- 信賴度。
- 規則支援
- 增譯
- 條件支援。
- 可部署性

要顯示的規則數上限

設定單字雲中要顯示的規則數目；可以選擇的最大值為 20。

相關規則 - 模型選項

使用此標籤上的設定可以指定「相關規則」模型的評分選項。

模型名稱 可以根據目標欄位自動產生模型名稱（未指定此類欄位時，將根據模型類型產生模型名稱），也可以指定自訂名稱。

預測的最大數量 指定可以併入在分數結果中的預測的最大數量。將此選項與**規則準則**條目配合使用可生成「最佳」預測，其中「最佳」表示最高層次的信賴度、支援、提昇等等。

規則準則 選取用於確定規則強度的測量。規則依這裡選取的準則強度排序，以便傳回項目集的前幾個預測。您可以從 5 個不同的準則中選擇。

- **信賴度** 信賴度是規則支援度與條件支援度的比例。在具有列出的條件值的項目中，具有預測結果值的項目所佔的百分比。
- **條件支援度** 條件為 true 的項目所佔的比例。
- **規則支援度** 整個規則、條件和預測均為 true 的項目所佔的比例。用**條件支援度**值相乘以**信賴度**值計算得出。
- **提昇** 規則信賴度與具有預測的事前機率的比例。
- **可部署性** 用於測量訓練資料中滿足條件但不滿足預測的部分所佔的百分比。

容許重複預測 要在評分期間包含多條具有相同預測的規則，請選中此勾選框。例如，選中此框將允許對下列規則進行評分。

```
麵包 & 芝士 -> 紅酒
芝士 & 水果 -> 紅酒
```

註：只有在之前已預測過所有預測 (`wine & pate`) 的情況下，才會將具有多個預測的規則 (`bread & cheese & fruit -> wine & pate`) 視為重複預測。

只有在預測未出現在輸入中時才對規則進行評分 要確保預測不會也出現在輸入中，請選中此選項。例如，如果進行評分的目的是為了進行一項傢具產品推薦，那麼已經包含餐桌的輸入可能不會購買另一個這樣的傢具。在這種情況下，選取此選項。但是，如果產品易腐爛或者是一次性的（如奶酪、嬰兒代乳品或者衛生紙），那麼結果已出現於輸入中的規則可能有些價值。如果是後面這種情況，則最有用的選項可能是**對所有規則評分**。

只有在預測出現在輸入中時才對規則進行評分 要確保預測也出現在輸入中，請選中此選項。當您嘗試深入瞭解現有客戶或交易時，此方法很有用。例如，您可能想要識別提升最高的規則，然後探索哪些客戶適合這些規則。

對所有規則進行評分 要在評分期間包含所有規則，而無論是否存在預測，請選中此選項。

「相關規則」模型塊

模型塊包含模型建立期間擷取自資料的規則的相關資訊。

檢視結果

您可以使用此對話框的「模型」標籤來瀏覽「相關規則」模型所產生的規則。在產生新節點或者對模型進行評分前瀏覽模型塊將顯示有關規則的資訊。

為模型評分

經過優化的模型塊可以新增到串流中，用於進行評分。請參閱第 41 頁的『使用串流中的模型塊』主題，以取得更多資訊。用來評分的模型塊在其各自的對話框上包括額外的「設定」標籤。請參閱『關聯規則模型塊設定』主題，以取得更多資訊。

關聯規則模型塊詳細資料

「關聯規則」模型區塊會在「輸出檢視器」的「模型」標籤中顯示模型的詳細資料。如需使用檢視器的相關資訊，請參閱《Modeler 使用手冊》(ModelerUsersGuide.pdf) 中標題為「使用輸出」的小節。

GSAR 建模作業將建立多個具有字首 $A 的新欄位，如下表格所示。

表 *25.* 由「關聯規則」建模作業建立的新欄位

欄位名稱	說明
$A-<prediction>#	此欄位包含模型對已評分記錄的預測。 <prediction> 是包括在模型中「預測」角色內的欄位名稱，而 # 是輸出規則的一系列數目（例如，如果分數設為包括 3 個規則，則數目順序將從 1 到 3）。
$AC-<prediction>#	此欄位包含預測中的信賴度。 <prediction> 是包括在模型中「預測」角色內的欄位名稱，而 # 是輸出規則的一系列數目（例如，如果分數設為包括 3 個規則，則數目順序將從 1 到 3）。
$A-Rule_ID#	此字段包含針對已評分資料集中每個記錄進行預測的規則的 ID。 # 是輸出規則的編號序列（例如，如果將分數設定為包含 3 條規則，那麼編號序列將從 1 到 3）。

關聯規則模型塊設定

「相關規則」模型塊的「設定」標籤顯示模型的評分選項。只有在將模型新增到串流畫布中用於評分之後，此標籤才可用。

預測的最大數量 指定為每個項目集合併入的預測的最大數量。套用至這個交易集且具有最高信賴度值的規則，用來為記錄（最多到指定的限制）產生預測。將此選項與**規則準則**選項配合使用可生成「最佳」預測，其中最佳指的是最高層次的信賴度、支援、提昇等等。

規則準則 選取用於確定規則強度的測量。規則依這裡選取的準則強度排序，以便傳回項目集的前幾個預測。您可以從下列準則中選擇。

- **信賴度**
- **規則支援**
- **增譯**

- **條件支援度**
- **可部署性**

容許重複預測 要在評分時包含多條具有相同結果的規則，請選中此勾選框。例如，選中此選項表示可以對下列規則進行評分：

```
麵包 & 芝士 -> 紅酒
芝士 & 水果 -> 紅酒
```

要在評分時排除重複預測，請取消選中此勾選框。

註：僅在所有後繼 (wine & pate) 之前都已預測的情況下，將包含多重後繼的規則 (bread & cheese & fruit -> wine & pate) 視為重複預測。

只有在預測未出現在輸入中時才對規則進行評分 選中此框可確保結果不會也出現在輸入中。例如，如果進行評分的目的是為了進行一項傢具產品推薦，那麼已經包含餐桌的輸入可能不會購買另一個這樣的傢具。在這種情況下，選取此選項。另一方面，如果產品易腐爛或者是一次性的（如奶酪、嬰兒代乳品或者衛生紙），那麼結果已出現於輸入中的規則可能有些價值。如果是後面這種情況，則最有用的選項可能是**對所有規則評分**。

只有在預測出現在輸入中時才對規則進行評分 選中此框可確保結果也出現在輸入中。當您嘗試深入瞭解現有客戶或交易時，此方法很有用。例如，您可能想要識別提升最高的規則，然後探索哪些客戶適合這些規則。

對所有規則進行評分 要在評分時包含所有規則，而無論結果是否出現在輸入中，請選中此選項。

第 13 章 「時間序列」模型

為什麼要進行預測？

預測的意思就是對一個或多個數列在一定時間內的值進行預言。例如，您可能希望預測某個系列產品或服務的預期需求，以便配置資源進行製造或配送。因為計劃決策的實施需要時間，所以預測在很多計劃過程中都是一個必不可少的工具。

時間序列建模方法假定歷史總會自我重演，即使不是完全一樣也會非常接近，足以通過研究過去對將來作出更好的決策。例如，為了預測下一年的銷售量，您可能得從分析今年的銷售量開始，看看近年來都有哪些發展趨勢或型樣（如果存在的話）。但型樣可能很難測量。例如，如果您的銷售量在幾周之內連續上升，那麼這是週期性原因呢還是一種長期趨勢的開始？

使用統計建模技術，可以分析過去資料中存在的型樣並加以預測，以確定該數列的未來值可能的的範圍。其結果是您的決策所依據的預測更為準確。

時間數列資料

時間序列是以規律的時間間隔採集的測量值的依序收集，例如，每日的股票價格或每周的銷售資料。測量值可以是您感興趣的任何內容，每個數列通常可以歸為下列類別之一：

- **應變數**。要預測的數列。
- **預測值**。 這是可能有助於解釋目標的數列，例如使用廣告預算來預測銷售量。預測值只能用於 ARIMA 模型。
- **事件**。一種特殊的預測值系列，用於說明可預測的重複發生發生事件，例如促銷活動。
- **人為介入**。一種特殊的預測值系列，用於說明一次性發生事件，例如停電或員工罷工。

時間間隔可以代表任何時間單位，但所有測量值的時間間隔必須相同。而且，沒有測量值的任何時間間隔必須設定為遺漏值。因此，有測量值的時間間隔數（包括測量值為遺漏值的時間間隔）定義資料歷程記錄展開範圍的時間長度。

時間數列的性質

研究數列過去的行為有助於辨別其中的型樣從而作出更好的預測。將其繪製成圖形時，多數時間序列就會表現出下列一種或多種特徵：

- 趨勢
- 季節循環和非季節循環
- 脈衝和步進
- 離群值

趨勢

趨勢是指數列層次的逐漸上升或下降或數列值隨時間的推移而增大或減小的趨勢。

© Copyright IBM Corp. 1994, 2019

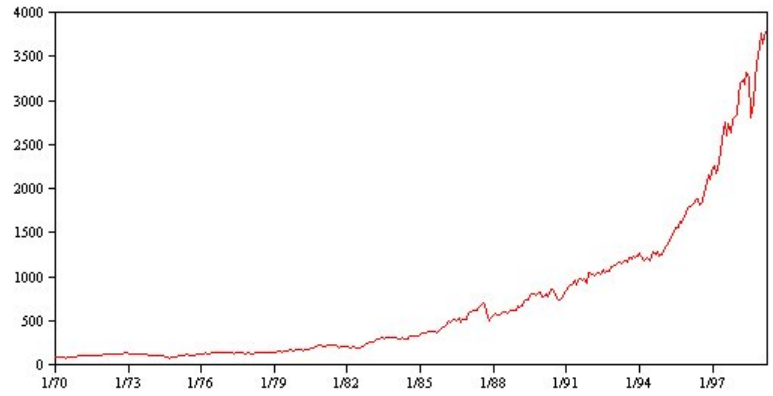


圖 *53.* 趨勢

趨勢既可以是**區域**的，也可以是**廣域**的，而一個數列可以同時體現這兩種趨勢。從歷程記錄來看，股票市場指數的數列圖形總的趨勢是上升的。經濟蕭條時期所表現出的是區域下降趨勢，而經濟繁榮時期表現出的是區域上升趨勢。

趨勢既可以是**線性**的，也可以是**非線性**的。線性趨勢是指數列層次表現為正增加或負增加，就和本金以單利計息差不多。非線性趨勢通常表現為倍增，即相對於以前的數列值成比例地增長。

廣域線性趨勢可通過指數平滑化模型和 ARIMA 模型很好地擬合和預測。在建立 ARIMA 模型的過程中，通常會對表現出趨勢的數列進行區分，以消除趨勢的效果。

週期性循環

週期性循環是數列值中可預測的重複型樣。

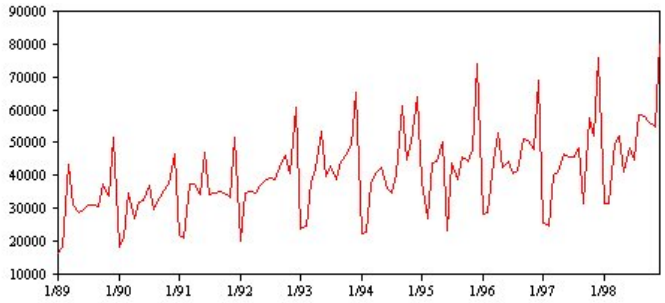


圖 *54.* 週期性循環

週期性循環與數列的時間間隔相聯系。例如，月度資料通常會隨季度和年度而週期。月度數列可能會表現出第一個季度較低的明顯季度循環或每年十二月都出現尖峰的年度循環。表現出週期性循環的數列稱之為具有**週期性**。

季節型樣對於獲取良好的擬合和預測非常有用，用來擷取週期性的有指數平滑化模型和 ARIMA 模型。

非週期性循環

非週期性循環是數列值中可能無法預測的重複型樣。

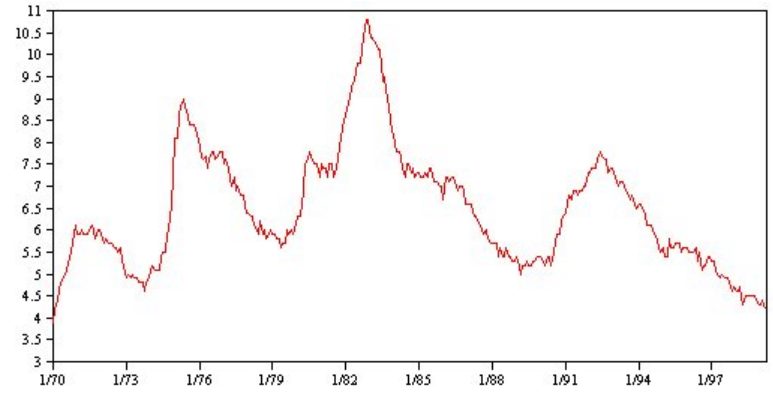


圖 *55.* 非週期性循環

某些數列（如失業率）明顯地表現出週期性性行為；但這種週期性性的週期性會隨時間而變化，因此很難預測何時高何時低。其他數列可能具有可預測的循環，但可能與陽曆並不完全吻合，或者其循環比一年長。例如，潮汐遵循陰曆，與奧林匹克運動會相關的國際旅遊和貿易每隔四年膨脹一次，還有多數宗教節日，其陽曆日期每年都會變化。

非季節週期型樣很難建模，通常會增加預測的不確定性。例如，股票市場的許多數列實例就常使預測者的努力無功而返。即便如此，當存在非季節型樣時，還是有必要加以說明。在多數情況下，您仍然可以找出與歷程資料擬合得很好的模型，從而最大限度地減小預測中的不確定性。

脈衝和步進

多數數列都會出現層次突變。它們通常分為兩種類型：

- 數列層次突然、 暫時性的變動，或稱**脈衝**
- 數列層次突然、 永久的性的變動，或稱**步進**

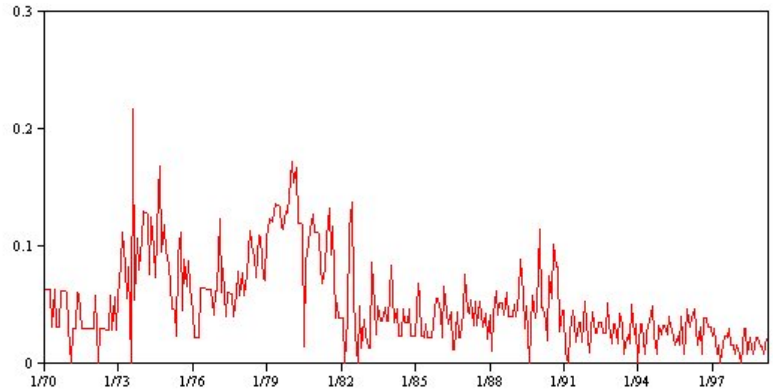


圖 *56.* 脈衝數列

觀測到步進或脈衝時，找到一種貌似合理的解釋很重要。「時間序列」模型是用來說明漸變而非突變的。因此，它們往往低估脈衝並為步進所瓦解，導致模型配適度低劣強人意，增加預測的不確定性。（某些週期性實例可能表現為突然的層次變化，但該層次在不同的季節期間之間則保持穩定。）

如果擾動是可以解釋的，那麼可以使用**人為介入**或**事件**為其建模。例如，1973 年 8 月，石油輸出國組織 (OPEC) 頒佈的石油禁運導致了通貨膨脹率的急劇變化，經過數月之後才恢復到正常層次。通過為該禁運月指定一個**點人為介入**，可以改善模型的擬合度，因此可以間接提高預測的準確性。例如，某個零售商店可能會發

現，所有商品均標示降價 50% 的當天銷售量比平時高出很多。通過將降價 50% 的促銷指定為一個定期的**事件**，可以改善模型的擬合度，估計將來重複該項促銷措施的效果。

離群值

時間序列層次中無法解釋的變動稱為**離群值**。這些觀察與數列中的其他值不一致，可能會顯著影響分析，從而影響「時間序列」模型的預測能力。

下圖顯示了時間序列中常見的幾種離群值。藍行代表沒有離群值的數列。紅行表示如果數列包含離群值情況下可能出現的型樣。這些離群值全部歸為**確定性**離群值，因為它們只影響數列的平均數層次。

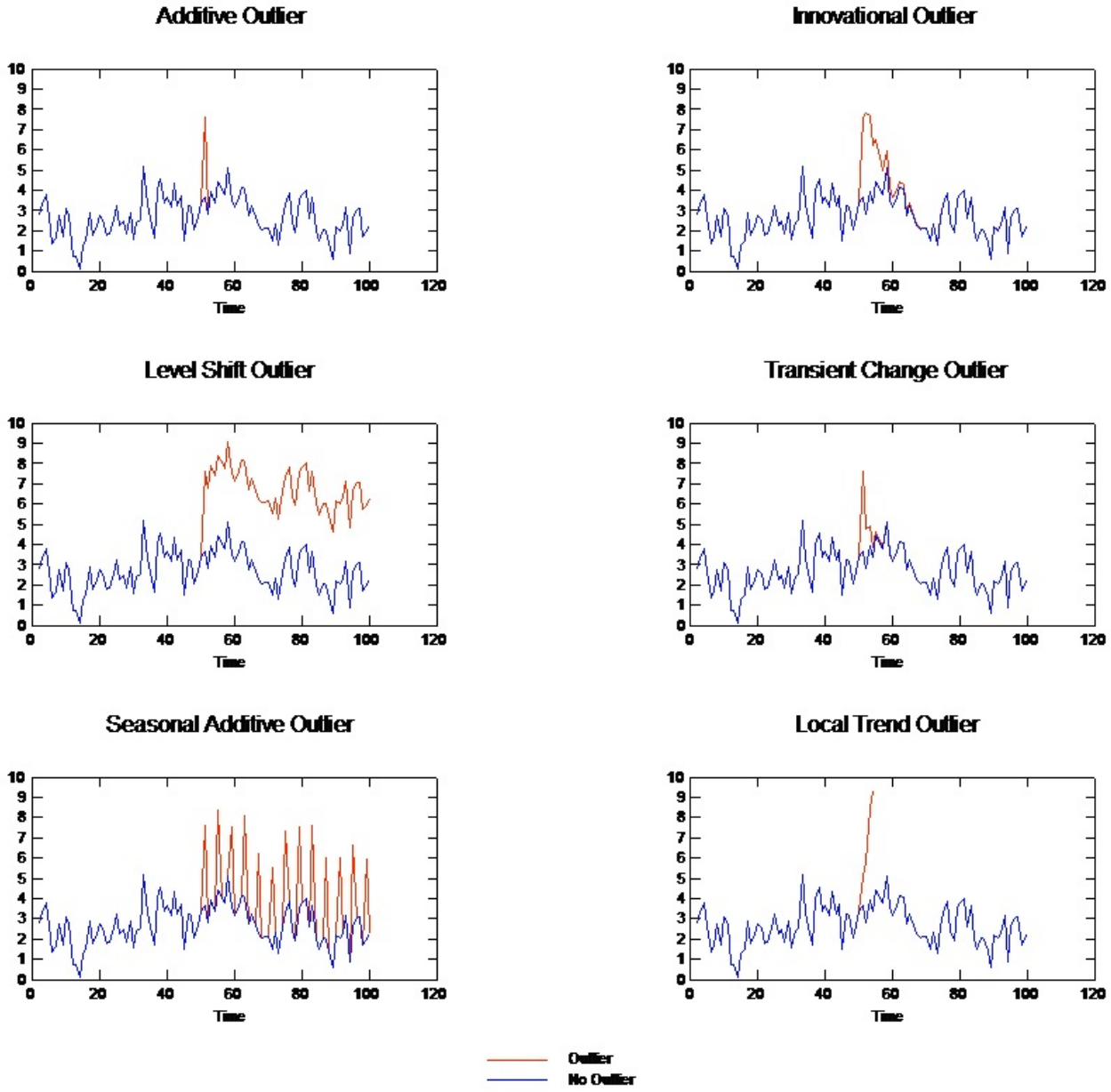


圖 *57.* 離群值類型

- **可加性離群值。**可加性離群值表現為一次觀測中出現的異常大或異常小的值。後續觀測不受可加性離群值的影響。連續的可加性離群值通常稱為**可加性離群值修補**。
- **創新離群值。**創新離群值的特徵為初始效應一直對後續觀測產生作用。這些離群值的影響可能會隨著時間的推移而不斷增強。
- **水平偏移離群值。**對於水平偏移，離群值之後出現的所有觀察均移動到新層次。與可加性離群值相反，水平偏移離群值會效果多數觀察，並且具有永久的性效果。
- **瞬時變化離群值。**瞬時變化離群值類似水平偏移離群值，只是這種離群值對後續觀測的效果呈指數遞減。最終，該數列會恢復到正常層次。
- **週期性可加性離群值。**週期性可加性離群值表現為以固定時間間隔重複出現的異常大或異常小的值。
- **局部趨勢離群值。**局部趨勢離群值會在出現初始離群值之後，在數列中產生一個由離群值中的型樣所導致的整體漂移。

時間序列中的離群值偵測包括確定出現的任何離群值的位置、類型和大小。Tsay (1988) 提出了一個用於偵測平均數層次變化以識別出確定性離群值的迭代程序。此過程是將一個假設不出現離群值的時間序列模型與另一個具有離群值的模型進行比較。從兩個模型之間的差異產生將任何給定點視為離群值的效果的估計。

自相關函數和偏自相關函數

自相關係數和局部自相關係數用於測量目前數列值和過去數列值之間的相關性，並指示預測將來值時最有用的過去數列值。瞭解了此內容，您就可以確定 ARIMA 模型中過程的順序。更具體來說，

- **自相關係數函數 (ACF)。**延遲為 *k* 時，這是相距 *k* 個時間間隔的數列值之間的相關性。
- **局部自相關係數函數 (PACF)。**延遲為 *k* 時，這是相距 *k* 個時間間隔的數列值之間的相關性，同時考慮了間隔之間的值。

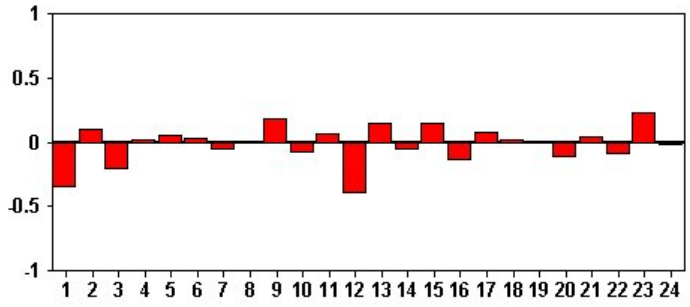


圖 *58.* 數列的 *ACF* 圖形

ACF 圖形的 *x* 軸表示計算自相關係數處的延遲； *y* 軸表示相關係數值（介於 –1 和 1 之間）。例如，ACF 圖形中延遲 1 處的峰值表示每個數列值與前面的值強相關係數，延遲 2 處的峰值表示每個值與以前兩個點之間的值強相關係數，依此類推。

- 正相關係數表示較大的目前的值與指定延遲處較大的值相對應；負相關係數表示較大的目前的值與指定延遲處較小的值相對應。
- 相關係數的絕對值是關聯強度的測量，絕對值越大表明關係越強。

數列轉換

轉換對在模型估計之前穩定數列常常有用。這對 ARIMA 模型尤其重要，因為估計這類模型之前需要數列保持**穩定**。如果在整個數列中，廣域層次（平均數）以及與該層次的平均值離差（變異數）保持不變，那麼該數列是穩定的。

盡管多數令人感興趣的數列都不穩定，但只要能夠通過套用轉換（如，自然對數、差異分析或季節差異分析）使數列保持穩定，那麼 ARIMA 就是有效的。

變異數穩定轉換。變異數隨時間變化的數列通常可以使用自然對數轉換或平方根轉換來保持穩定。這些轉換也稱為函數轉換。

- **自然對數。** 對數列值取自然對數。
- **平方根。** 對數列值套用平方根函數。

自然對數轉換和平方根轉換不能用於具有負值的數列。

層次穩定轉換。ACF 中值的緩慢下降表示每個數列值都與上一個值具有很強的相關性。通過分析數列值的變化，您可以獲得一個穩定層次。

- **簡單差異分析。**計算數列中每個值與上一個值之間的差，數列中最舊的值除外。這意味著經過差分的數列將比原始數列少一個值。
- **週期性差異分析。**除計算每個值與上一個季節值之間的差分外，其他均與簡單差異分析相同。

將簡單差異分析或季節差異分析同時用於對數轉換或平方根轉換時，總是先套用變異數穩定轉換。同時套用簡單差異分析和週期性差異分時，無論首先套用簡單差異分析還是週期性差異分，得到的數列值均相同。

預測值數列

預測值系列包含可能有助於解釋要預測數列的行為的相關資料。例如，一個網上零售商或型錄零售商可能會根據郵寄的型錄數量、開啟通的電話數量或公司網頁的點擊次數來預測銷售量。

任何數列都可以作為預測值，條件是該數列須延伸到要預測的將來時間，並且具有不存在遺漏值的完整資料。

向模型中新增預測值時以慎重為宜。新增大量預測值會增加估計模型所需的時間。雖然新增預測值可以提高模型擬合歷程資料的能力，但並不意味著該模型就一定能產生更好的預測結果，因為增加的複合怕有可能及不上所造成的麻煩。理想的目標是，找出的模型既是最簡單的，同時又能作出很好的預測。

一般而言，建議預測值的數量應少於樣本大小除以 15（即最多每 15 個觀察值一個預測值）。

包含遺漏資料的預測值。不能在預測中使用包含不完整資料或遺漏資料的預測值。這適用於歷程資料和將來值。在某些情況下，可通過設定模型的估計展開範圍以便在估計模型時排除最舊資料來避免上述限制。

空間-時間預測建模節點

空間-時間預測 (STP) 有多數潛在的應用，例如緊急管理建築物或設施、對機械服務專案師進行績效分析和預測或者進行公用交通規劃。在這些應用程式中，通常要對空間和時間進行測量（如能耗）。可能與記錄這些測量值相關的問題包含哪些因子效果未來的觀察、如何實現所需的變化或者如何更好地管理系統？為了回答這些問題，您可以在不同位置使用能夠預測未來值的統計技術，並可以明確地對可調因子進行建模以執行假設情況分析。

STP 分析使用包含位置資料、預測輸入欄位（預測值）、時間欄位和目標欄位的資料。每個位置有數個資料列，用來代表在每次測量時每個預測工具的值。分析資料之後，它可用來在分析中所使用形狀資料內的任何位置，預測目標值。並且，還可以預測何時能夠獲知未來復原點的輸入資料。

註：STP 節點不支援 IBM SPSS Collaboration and Deployment Services 中的「模型評估」步驟或「冠軍挑戰者」步驟。

在 IBM SPSS Modeler 安裝的 Demos 目錄中，提供了名為 stp_server_demo.str 的串流，這個串流顯示了有關如何使用 STP 的有效範例，並參照了資料檔 room_data.csv 和 score_data.csv。您可從 Windows「啟動」功能表的 IBM SPSS Modeler 程式集存取 Demos目錄。 stp_server_demo.str 檔案位於 streams 目錄中。

空間-時間預測 - 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（或上游來源節點的「類型」標籤）中的角色設定（僅限目標和預測值）。

使用自訂欄位指派

要在此畫面上手動分配目標、預測值和其他角色，請選中此選項。

欄位 顯示資料中所有可以選取的欄位。使用箭頭按鈕可以將此清單中的項目手動分配到畫面右側的各種框。圖示指出每個欄位的有效測量層次。

註：對於每個位置每個時間間隔，STP 都需要 1 條記錄才能正常運行；因此，這些是必需欄位。

在**欄位**窗格底部，按一下**全部**按鈕將選中所有欄位（這與測量層次無關），按一下個別的測量層次按鈕將選中所有具有該測量層次的欄位。

目標 選取一個欄位作為預測目標。

註：您只能選取測量層次為「連續」的欄位。

位置 選取要在模型中使用的位置類型。

註：您只能選取測量層次為「地理空間」的欄位。

位置標籤

形狀資料通常包含一個表明層特徵的名稱的欄位，例如，這可能是省/自治區/直轄市或者國家或地區的名稱。使用此欄位可以將名稱或標籤與位置相關聯，方法是選取一個種類欄位來標註輸出中的所選**位置**欄位。

時間欄位

選取要在預測中使用的時間欄位。

註：您只能選取測量層次為「連續」且儲存類型為時間、日期、時間戳記或整數的欄位。

預測值（輸入）

選擇一或多個欄位作為預測的輸入。

註：您只能選取測量層次為「連續」的欄位。

空間-時間預測 - 時間間隔

在「時間間隔」窗格中，您可以選取用於設定時間間隔和隨時間推移進行的任何必要聚集的選項。

在您可以建立 STP 模型之前，需要進行資料預備以便將時間欄位轉換為指標；要使得能夠進行這種轉換，時間欄位中的記錄之間必須有固定的區間。如果資料尚未包含此資訊，請使用此窗格中的選項來設定此區間，然後才能使用建模節點。

時間間隔 請選取要將資料集轉換為的時間間隔。可用的選項取決於在「欄位」標籤上選擇作為模型的**時間欄位**的欄位儲存類型。

- **期間** 僅適用於整數時間欄位；這是一系列與任何其他可用時間間隔均不相符的時間間隔，並且每項測量之間的間距一致。
- **年** 僅可用於「日期」或「時間戳記」時間欄位。
- **季度** 僅可用於「日期」或「時間戳記」時間欄位。如果您選擇了此選項，那麼系統將提示您選擇第一個季度的**開始月**。
- **月** 僅可用於「日期」或「時間戳記」時間欄位。
- **周** 僅可用於「日期」或「時間戳記」時間欄位。
- **天** 僅可用於「日期」或「時間戳記」時間欄位。
- **小時** 僅可用於「時間」或「時間戳記」時間欄位。
- **分鐘** 僅可用於「時間」或「時間戳記」時間欄位。
- **秒** 僅可用於「時間」或「時間戳記」時間欄位。

當您選取了**時間間隔**時，系統將提示您填寫更多欄位。可用的欄位同時取決於時間間隔和儲存類型。下列清單顯示了可能會顯示的欄位。

- **每週的天數**
- **一天的時數**
- **一周的開始日期** 一周的第一天
- **一天的開始時間** 新的一天的開始時間。
- **時間間隔值** 您可以選擇下列其中一個選項：1、2、3、4、5、6、10、12、15、20 或 30。
- **開始月** 財年的開始月。
- **開始期間** 如果使用了**期間**，請選取開始期間。

資料與指定的時間間隔設定相符 如果資料已包含正確的時間間隔資訊，並且不需要進行轉換，請選中此勾選框。選中此框後，**聚集**區域中的欄位將不提供。

聚集

只有在您取消選中**資料與指定的時間間隔設定相符**勾選框時才可用；請指定用於聚合欄位以便與指定區間相符的選項。例如，如果您具有每週及每月資料的混合，則可以聚集或「彙總」每週值來達到均勻的每月間隔。選取要用於不同欄位類型聚集的預設值，建立要用於任何特定欄位的任何自訂設定。

- **連續** 設定要套用至未逐個指定的所有連續欄位的預設聚集方法。可以從數個方法中選擇：
  - 總和
  - 平均數
  - 下限
  - 上限
  - 中位數
  - 第一個四分位數
  - 第三個四分位數

指定欄位的自訂設定 要將特定聚集函數套用於個別欄位，請在此表格中選擇欄位並選擇聚集方法。

- **欄位** 使用**新增欄位**按鈕以顯示「選取欄位」對話框並選擇所需欄位。所選擇的欄位將顯示在此欄中。
- **聚集函數** 從下拉清單中，選取用於將欄位轉換為指定時間間隔的聚集函數。

空間-時間預測 - 基本建置選項

使用此對話框中的設定可以設定基本模型建置選項。

模型設定

包含截距(E)

包括截取（模型中的固定項目）可增加解的整體精確度。如果可以假設資料穿過原點，就可以將截距排除在外。

上限自身迴歸階數

自身迴歸階數指定使用哪些先前值來預測目前的值。使用此選項可以指定用於計算新值的先前記錄數。您可以選擇介於 1 與 5 之間的任何整數。

空間共變量

估計方法

選擇要使用的估計方法；您可以選擇**參數**或**無母數**。對於**參數**方法，您可以從三種**模型**類型中進行選擇：

- **Gaussian**
- **指數**
- **冪指數** 如果選取此選項，那麼還必須指定要使用的**冪次**層次。此層次可以是介於 1 與 2 之間的任意值，並以 0.1 為增量變動。

空間-時間預測 - 進階建置選項

熟悉 STP 的使用者可以使用下列選項對模型建置程序進行微調。

遺漏值的最大百分比

指定模型中可以併入的包含遺漏值的記錄所佔的上限百分比。

模型建立中用於假設測試的顯著性水準

指定用於 STP 模型估計的所有測試（包括兩項適合度測試、作用 F 測試和係數 T 測試）的顯著性水準值。此層次可以是 0 與 1 之間的任何值，並以 0.01 為增量變動。

空間-時間預測 - 輸出

在建立模型之前，請使用此窗格中的選項來選取要包含在輸出檢視器中的輸出。

模型資訊

模型規格

選中此選項表示將模型規格資訊包含在模型輸出中。

時間資訊摘要

選中此選項表示將時間資訊摘要包含在模型輸出中。

評估

模型品質

選中此選項表示將模型品質包含在模型輸出中。

平均數結構模型中的作用測試

選中此選項表示將作用測試資訊包含在模型輸出中。

解譯

模型係數的平均數結構

選中此選項表示將平均數結構模型係數資訊包含在模型輸出中。

自身迴歸係數

選中此選項表示將自身迴歸係數資訊包含在模型輸出中。

空間衰變測試

選中此選項表示將空間共變數（即空間衰變）測試資訊包含在模型輸出中。

參數空間共變異模型參數圖形

選中此選項表示將參數空間共變異模型參數圖形資訊包含在模型輸出中。

註：僅當您在「基本」標籤上選取了**參數**估計方法時，此選項才可用。

相關性熱圖

選中此選項表示將目標值的圖包含在模型輸出中。

註：如果模型中的位置超過 500 個，則不會建立地圖輸出。

相關性圖

選中此選項表示將相關性的圖包含在模型輸出中。

註：如果模型中的位置超過 500 個，則不會建立地圖輸出。

位置叢集

選中此選項表示將位置叢集輸出包含在模型輸出中。只有不需要存取圖資料的輸出才會併入在叢集輸出中。

註：只能為非參數空間共變異模型建立此輸出。

如果選擇了此選項，那麼可以設定下列各項：

- **相似性臨界值** 請選取臨界值，達到此臨界值的輸出叢集將被視為足夠相似，從而合併到單個叢集中。
- **要顯示的最大叢集數目** 請設定模型輸出中可以併入的叢集數目的上限。

空間-時間預測 - 模型選項

模型名稱 可以根據目標欄位自動產生模型名稱，或者指定自訂名稱。自動產生的名稱為目標欄位名稱。

不確定性因數 (%) 不確定因數是用來在預測未來時表示成長不確定性的百分比值。每向未來邁進一步，預測不確定性的上限與下限就會增加此百分比。設定要套用至模型輸出的不確定因素；這會設定預測值的上限和下界。

空間-時間預測模型塊

時空性預測 (STP) 模型塊會在「輸出檢視器」的「模型」標籤中顯示模型的詳細資料。如需使用檢視器的相關資訊，請參閱《Modeler 使用手冊》(ModelerUsersGuide.pdf) 中標題為「使用輸出」的小節。

空間-時間預測 (STP) 建模作業將建立多個具有 $STP- 字首的新欄位，如下表格所示。

表 *26. STP* 建模作業建立的新欄位

欄位名稱	說明

表 *26. STP* 建模作業建立的新欄位 *(*繼續*)*

$STP-<Time>	在模型建立過程中建立的時間欄位。「建置選項」標籤的「時間間隔」窗格中的設定確定了建立此欄位的方式。 <Time> 是「欄位」標籤上選取作為**時間欄位**的欄位的原始名稱。 **註：**只有在模型建立過程中轉換了原始**時間欄位**的情況下，才會建立此欄位。
$STP-<Target>	此欄位包含目標值的預測。 <Target> 是模型原始**目標**欄位的名稱
$STPVAR-<Target>	此欄位包含 VarianceOfPointPrediction 值。 <Target> 是模型原始**目標**欄位的名稱
$STPLCI-<Target>	此欄位包含 LowerOfPredictionInterval 值（即信賴度下限）。 <Target> 是模型原始**目標**欄位的名稱
$STPUCI-<Target>	此欄位包含 UpperOfPredictionInterval 值（即信賴度上限）。 <Target> 是模型原始**目標**欄位的名稱

空間/時間預測模型設定

使用「設定」標籤可以控制您認為建模作業中可接受的不確定性層次。

不確定性因數 (%) 不確定因數是用來在預測未來時表示成長不確定性的百分比值。每向未來邁進一步，預測不確定性的上限與下限就會增加此百分比。設定要套用至模型輸出的不確定因素；這會設定預測值的上限和下界。

TCM 節點

使用此節點可以建立時間因果模型 (TCM)。

時間原因模型

時間原因建模嘗試在時間序列資料中探索主要原因關係。在時間原因建模中，您可以指定一組目標序列以及這些目標的一組候選輸入。隨後，該程序會為每個目標建置自動迴歸時間序列，並且只包括與目標之間具有原因關係的那些輸入。此方法不同於傳統時間序列建模，在後者中，您必須明確指定目標序列的預測因素。由於時間原因建模通常包括多個相關時間序列的建置模型，因此結果稱為模型系統。

在時間原因建模的環境定義中，原因一詞是指「葛蘭哲因果」。如果依照 X 和 Y 的過去值進行 Y 迴歸產生的 Y 的模型，優於只依照 Y 的過去值進行迴歸產生的模型，則時間序列 X 被稱為「葛蘭哲因果」的另一個時間序列 Y。

註：時間原因建模節點不支援IBM SPSS Collaboration and Deployment Services中的「模型評估」或「冠軍挑戰者」步驟。

範例

商業決策制訂者可使用時間原因建模，來揭露說明商業的眾多時間型度量值內的原因關係。該分析可能會顯示一些可控制的輸入，這對關鍵績效指標 (KPI) 有最大的影響。

大型 IT 系統的管理者可使用時間原因建模，在眾多相互關聯的作業度量值中偵測例外狀況。然後，原因模型容許超越例外狀況偵測並探索造成例外狀況最可能的主要原因。

欄位需求

必須至少存在一個目標。依預設，不會使用預先定義角色為「無」的欄位。

資料結構

時間原因建模支援兩種類型的資料結構。

直欄型資料

若為直欄型資料，每個時間序列欄位都包含單一時間序列的資料。此結構是時間序列資料的傳統結構，由「時間序列模型器」使用。

多維度資料

若為多維度資料，每個時間序列欄位都包含多個時間序列的資料。特定欄位中個別的時間序列隨後會被稱為維度欄位的種類欄位的一組值進行識別。 例如，兩個不同銷售通道（零售和網路）的銷售資料可能會儲存在單一 *sales* 欄位中。名為 *channel* 且值為「零售」和「網路」的維度欄位可識別與兩個銷售通道之一相關聯的記錄。

註：若要建置時間原因模型，您需要足夠的資料點。本產品使用如下限制：

```
m>(L + KL + 1)
```

其中，m 是資料點數目，L 是落差數目，而 K 是預測值數目。請確定您的資料集足夠大，以使資料點數目 (m) 滿足條件。

要建模的時間序列

在「欄位」標籤上，使用**時間序列**設定來指定要併入模型系統中的數列。

為適用於您資料的資料結構選取選項。若為多維度資料，請按一下**選取維度**以指定維度欄位。指定的維度欄位順序可定義維度欄位在所有後續對話框及輸出中的出現順序。使用「選取維度」子對話框上的向上和向下箭頭按鈕，可對維度欄位進行重新排序。

若為直欄型資料，術語數列與術語欄位的意義相同。若為多維度資料，包含時間序列的欄位是指度量值欄位。多維度資料的時間序列由度量值欄位和每個維度欄位的值來定義。下列考量同時適用於直欄型資料和多維度資料。

- 指定為候選輸入或目標及輸入的數列可考慮併入每一個目標的模型中。每個目標的模型一律包括目標本身的落階值。
- 指定為強制輸入的數列一律併入每個目標的模型中。
- 至少一個數列必須指定為目標或目標和輸入。
- 如果選取了**使用預先定義角色**，則角色為「輸入」的欄位會設為候選輸入。預先定義的角色未對映為強制輸入。

多維度資料

若為多維度資料，可以在網格中指定度量值欄位與相關聯的角色，其中，網格中的每一列指定單一度量值和角色。依預設，模型系統包括網格中每一列的維度欄位所有組合的數列。例如，如果存在 *region* 和 *brand* 的維度，則依預設，將度量值 *sales* 指定為目標表示每個 *region* 和 *brand* 的組合都有個別的銷售目標數列。

針對網格中的每一列，您可以按一下維度的省略符號按鈕，來自訂任何維度欄位的值集。執行此動作會開啟「選取維度值」子對話框。您還可以新增、刪除或複製網格列。

數列計數直欄會顯示目前為相關聯度量值指定的維度值集數。顯示的值可以大於數列的實際數目（每個集一個數列）。如果指定的部分維度值組合未對應於相關聯度量值包含的數列，則會發生這種情況。

選取維度值： 若為多維度資料，您可以指定將套用至具有特定角色的特定度量值欄位的維度值，以自訂分析。例如，如果 *sales* 為度量值欄位，而 *channel* 是值為「零售」和「網路」的維度，則您可以指定「網路」銷售是輸入，而「零售」銷售是目標。您也可以指定套用至分析中所使用之所有度量值欄位的維度子集。例如，如果 *region* 是指示地理空間區域的維度欄位，則您可以將分析限制為特定區域。

所有值

指定包括現行維度欄位的所有值。此選項是預設值。

選取要包括或排除的值

使用此選項來指定現行維度欄位的值集。如果針對**眾數**選取了**包括**，則只會包括**選取的值**清單中指定的值。如果針對**眾數**選取了**排除**，則會包括除**選取的值**清單中指定的值以外的所有值。

您可以過濾可從中選擇的值集。符合過濾條件的值會出現在**未選取的值**清單的**相符**標籤中，而不符合過濾條件的值則會出現在**不符合**標籤中。無論任何過濾條件，**所有**標籤都會列出所有未選取的值。

- 指定過濾器時，您可以使用星號 (*) 來表示萬用字元。
- 若要清除現行過濾器，請在「過濾顯示的值」對話框上，為搜尋詞彙指定空白值。

觀察

在「欄位」標籤上，使用**觀察**設定來指定用於定義觀察的欄位。

由日期/時間定義的觀察

您可以指定觀察由日期、時間或時間戳記欄位定義。除了定義觀察的欄位以外，還可以選取說明觀察的適當時間間隔。根據指定的時間間隔，您還可以指定其他設定，例如觀察間隔（增量）或每週的天數。下列考量僅適用於時間間隔：

- 如果觀察的時間（例如處理銷售單的時間）間隔無規律，請使用值**無規律**。如果選取**無規律**，您必須從「資料規格」標籤上的**時間間隔**設定中，指定用於分析的時間間隔。
- 如果觀察代表日期和時間，而時間間隔是小時數、分鐘數或秒數，則使用**每天小時數**、**每天分鐘數**或**每天秒數**。如果觀察代表無日期參照的時間（期間），且時間間隔為小時數、分鐘數或秒數，則使用**小時數（非週期）**、**分鐘數（非週期）**或**秒數（非週期）**。
- 根據選取的時間間隔，程序可以偵測遺漏的觀察。必須偵測遺漏觀察，因為該程序假設所有觀察的時間間隔相等，且沒有任何遺漏觀察。例如，如果時間間隔為「日」，而日期 2014-10-27 後接 2014-10-29，則遺漏 2014-10-28 的觀察。會針對任何遺漏觀察插補值。可從「資料規格」標籤指定用來處理遺漏值的設定。
- 指定的時間間隔可讓程序偵測相同時間間隔中需要總計在一起的多個觀察，並依據間隔界限（例如月份的第一天）排列觀察，以確保觀察間隔相等。例如，如果時間間隔為「月」，則會將相同月份內的多個日期總計在一起。這種類型的總計稱為分組。依預設，在分組觀察時會加總觀察。您可以從「資料規格」標籤上的**總計及分佈**設定中指定不同的分組方法，如觀察平均數。
- 對於部分時間間隔，其他設定可定義一般相等時間間隔中的岔斷。例如，如果時間間隔為「日」，但只有工作日有效，則您可以指定一週內有五天，一週從星期一啟動。

由期間或循環期間定義的觀察

觀察可由一或多個代表期間或重複循環期間（多達任意數目的循環水準）的整數欄位定義。您可以使用此結構來說明不適合其中一個標準時間間隔的觀察數列。例如，可使用代表年份的循環欄位以及代表月份的期間欄位來說明只有 10 個月的財政年度，其中一個循環的長度為 10。

指定循環期間的欄位可定義週期性水準的階層，其中最低水準由**期間**欄位定義。下一個最高水準由水準為 1 的循環欄位指定，接著由水準為 2 的循環欄位指定，依此類推。每個水準（除最高水準以外）

的欄位值相對於下一個最高水準必須是週期性的。最高水準的值不能為週期性值。例如，在財政年度為 10 個月的情況下，月份在年份中是週期性的，但年份不是週期性的。

- 特定水準上的循環長度是下一個最低水準的週期。以財政年度為例，只有一個循環水準且循環長度為 10，因為下一個最低水準代表月份，而指定的財政年度有 10 個月。
- 針對不從 1 啟動的任何週期性欄位指定起始值。偵測遺漏值時需要此設定。例如，如果週期性欄位從 2 啟動，但起始值指定為 1，則程序會假設該欄位每個循環中的第一個期間都存在遺漏值。

分析的時間間隔

用於分析的時間間隔可能與觀察的時間間隔不同。例如，如果觀察的時間間隔為「日」，您可以為分析的時間間隔選擇「月」。然後，在模型建置之前，將資料總計從每日改為每月。您也可以選擇將資料分佈從較長時間間隔改成較短時間間隔。例如，如果觀察是按季度發生的，則您可以將資料從按季度分佈改成按月分佈。

執行分析的時間間隔的可用選項，視觀察的定義方式以及這些觀察的時間間隔而定。特別地，如果觀察是由循環期間定義，則只支援總計。在此情況下，分析的時間間隔必須大於或等於觀察的時間間隔。

從「資料規格」標籤上的**時間間隔**設定中，指定用於分析的時間間隔。您可以從「資料規格」標籤上的**總計及分佈**設定指定總計或分佈資料所採用的方式。

總計和分佈

聚集函數

當用於分析的時間間隔長於觀察的時間間隔時，會總計輸入資料。例如，如果觀察時間間隔是「日」，而分析的時間間隔是「月」，則會執行總計。提供了下列聚集函數：平均數、加總、眾數、最小值或最大值。

分佈函數

當用於分析的時間間隔短於觀察的時間間隔時，會分佈輸入資料。例如，如果觀察時間間隔是「季度」，而分析的時間間隔是「月」，則會執行分佈。提供了下列分佈函數：平均數或加總。

分組函數

如果觀察由日期/時間定義且在相同的時間間隔發生多個觀察，則可套用分組。例如，如果觀察的時間間隔是「月」，則相同月份中的多個日期會分組在一起，並與其發生所在的月份相關聯。您可以使用的分組函數如下：mean、sum、mode、min 或 max。如果觀察由日期/時間定義，且觀察的時間間隔指定為「無規律」，則一律會執行分組。

註：雖然分組是一種總計形式，但它是在處理所有遺漏值之前執行，而正式的總計是在處理完所有遺漏值之後執行。當觀察的時間間隔指定為「無規律」時，只能使用分組函數來執行總計。

總計跨日期觀察與前一天的值

指定是否將在跨日期界限的時間內的觀察與前一天的值總計在一起。例如，針對一天 8 小時制（在 20:00 啟動）的每小時觀察，這個設定會指定在 00:00 和 04:00 之間的觀察是否併入前一天的總計結果。只有在觀察時間間隔是「每天小時數」、「每天分鐘數」或「每天秒數」，且分析的時間間隔為「日」時，此設定才適用。

所指定欄位的自訂設定

您可以逐個欄位指定聚集函數、分佈函數及分組函數。這些設定會置換聚集函數、分佈函數及分組函數的預設值。

遺漏值

輸入資料中的遺漏值會取代為插補值。提供了下列取代方法：

線性插補

使用線性插補來取代遺漏值。該選項會用遺漏值出現之前的最後一個有效值，還有遺漏值出現之後的第一個有效值當作內插。如果數列中的第一個或最後一個觀察具有遺漏值，則會使用數列開頭或結尾的兩個最近非遺漏值。

數列平均數

以整個數列的平均數置換遺漏值。

鄰近點平均數

以有效周圍值的平均數置換遺漏值。附近點的指距是指用來計算平均數的遺漏值，其之前及之後的有效值個數。

鄰近點中位數

以有效周圍值的中位數置換遺漏值。附近點的指距是指用來計算中位數的遺漏值，其之前及之後的有效值個數。

線性趨勢

此選項會使用數列中的所有非遺漏觀察來擬合簡式線性迴歸模型，然後使用該模型來插補遺漏值。

其他設定：

遺漏值百分比上限 (%)

指定任何數列可接受的遺漏值百分比上限。其遺漏值多於所指定上限的數列將會從分析中排除。

一般資料選項

每個維度欄位的最大相異值數目

此設定適用於多維度資料，並指定任何一個維度欄位可接受的最大相異值數目。依預設，此限制設為10000，但它可以增加到任意大數目。

一般建置選項

信賴區間寬度 (%)

此設定可控制預測及模型參數兩者的信賴區間。您可以指定小於 100 的任何正數值。依預設，會使用 95% 信賴區間。

每個目標的最大輸入數

此設定指定每個目標模型可接受的最大輸入數目。您可以指定 1 - 20 範圍內的整數。每個目標的模型一律包括本身的落階值，因此，將此值設為 1 表示唯一的輸入是目標本身。

模型容差

此設定可控制用來判定每個目標最佳輸入集的疊代處理程序。您可以指定大於零的任何值。預設值為 0.001。模型容差是指選取預測值的停止準則。其可影響最終模型中包含的預測值數目。但如果目標可良好預測本身，則最終模型中可能不會包含其他預測值B。可能需要進行一些試驗並包含錯誤（例如，如果已將此值設為較高值，您可以嘗試將其設為較小值，以查看是否會包含其他預測值）。

離群值臨界值 (%)

如果觀察為離群值的機率（從模型計算得出）超出此臨界值，則會將觀察標示為離群值。您可以指定 50 - 100 範圍內的值。

每個輸入的落階數

此設定指定每個目標模型中每個輸入的落階項目數。依預設，會透過用於分析的時間間隔自動判定落階項目數。例如，如果時間間隔為月數（增量為一個月），則落階數為 12。您可以選擇性地明確指定落階數。指定的值必須是 1-20 範圍內的整數。

使用現有模型繼續估計

如果您產生了時間原因模型，選取此選項可重複使用指定給該模型的準則設定，而不必建置新的模型。使用這種方式，您可以根據與之前相同的模型設定，但使用更新的資料，來重新估計並產生新預測，從而節省時間。

要顯示的數列

這些選項指定要顯示其輸出的數列（目標或輸入）。所指定數列的輸出內容由**輸出選項**設定來決定。

顯示與最適模型相關聯的目標

依預設，會顯示與 10 個最適模型（由均方根百分比誤差所判定）相關聯之目標的輸出。您可以指定最適模型的不同固定數目，也可以指定最適模型的百分比。您也可以從下列適合度量數中選擇：

R 平方(R)

線性模型的適合度量數，有時也稱為判斷係數。由模型所解釋的目標變數中變異的比例。它的範圍值介於 0 到 1 之間。值越小，表示模型越不適合資料。

均方根百分比誤差

測量模型預測值與數列觀察值差異的程度。其與使用的單元無關，因此可用來比較具有不同單元的數列。

均方根誤差(Q)

平均均方根誤差。是相依系列與其模型預測層級差異程度的量數，用與相依系列相同的單位來表示。

BIC Bayesian 資訊準則。用於根據 -2 減少對數概似值來選取及比較模型的量數。數值越小代表模式越佳。BIC 也會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型），但比 AIC 更嚴格。

AIC Akaike 資訊準則。用於根據 -2 減少對數概似值來選取及比較模型的量數。數值越小代表模式越佳。AIC 會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型）。

指定個別系列。

您可以指定要對其進行輸出的個別數列。

- 若為直欄型資料，您可以指定包含您想要之數列的欄位。所指定欄位的順序可定義欄位在輸出中的出現順序。
- 若為多維度資料，您可以將項目新增至包含數列之度量值欄位的網格中，來指定特定數列。然後，可指定用來定義數列的維度欄位值。
  - 您可以將每個維度欄位的值直接輸入到網格中，也可以從可用維度值清單中選取。若要從可用維度值清單中選取，請在您想要之維度的資料格中，按一下省略符號按鈕。執行此動作會開啟「選取維度值」子對話框。
  - 您可以在「選取維度值」子對話框中，按一下雙目望遠鏡圖示並指定搜尋詞彙，來搜尋維度值清單。空格會被視為搜尋詞彙的一部分。搜尋詞彙中的星號 (*) 不表示萬用字元。
  - 網格中的數列順序可定義數列在輸出中的出現順序。

若同時為直欄型資料與多維度資料，則輸出限制為 30 個數列。此限制包括您指定的個別數列（輸入或目標）以及與最適模型相關聯的目標。個別指定的數列優先於與最適模型相關聯的目標。

輸出選項

這些選項可指定輸出內容。**目標的輸出**群組中的選項會針對與**要顯示的數列**設定上的最適模型相關聯的目標，產生輸出。**數列的輸出**群組中的選項會針對**要顯示的數列**設定上指定的個別數列，產生輸出。

整體模型系統

顯示模型系統中數列之間原因關係的圖形表示法。所顯示目標的模型適合度統計量與離群值的表格都會併入為輸出項目的一部分。在**數列的輸出**群組中選取此選項時，會針對**要顯示的數列**設定上指定的每個個別數列建立個別輸出項目。

數列之間的原因關係具有關聯的顯著性水準，其中顯著性水準越小，表示連線越顯著。您可以選擇隱藏其顯著性水準大於指定值的關係。

模型適合度統計量與離群值

選取用來顯示之目標數列的模型適合度統計量與離群值的表格。這些表格包含的資訊與「整體模型系統」視覺化中的表格資訊相同。這些表格支援所有用來旋轉及編輯表格的標準功能。

模型效應與模型參數

選取用來顯示之目標數列的模型效應測試與模型參數的表格。模型效應測試包括模型中所包含每個輸入的 F 統計量及相關聯的顯著性值。

影響圖

顯示相關數列與它影響或影響它的其他數列之間原因關係的圖形表示法。影響相關數列的數列稱為原因。選取**效應**會產生起始設定為顯示效應的影響圖。選取**原因**會產生起始設定為顯示原因的影響圖。選取**原因與效應**會產生兩個個別影響圖：起始設定為原因的圖與起始設定為效應的圖。您能以互動方式在顯示影響圖的輸出項目中的原因與效應之間切換。

您可以指定要顯示的原因或效應水準數目，其中第一個水準只是相關數列。每個其他水準都會顯示相關數列更間接的原因或效應。例如，效應顯示中的第三個水準由數列組成，這些數列包含第二個水準中的數列作為直接輸入。然後，第三個水準中的數列會間接受到相關數列的影響，因為相關數列是第二個水準中數列的直接輸入。

數列圖

選取用來顯示之目標數列的觀察值與預測值圖表。如果要求預測，則該圖也會顯示預測的預測值與信賴區間。

殘差圖

選取用來顯示之目標數列的模型殘差圖。

最上層輸入

在一段時間內所顯示的每個目標以及目標前 3 個輸入的圖形。最上層輸入是顯著性值最低的輸入。若要容納輸入與目標的不同尺度，Y 軸代表每個數列的 z 評分。

預測表格

選取用來顯示之目標數列的那些預測的預測值與信賴區間表格。

離群值主要原因分析

判定哪些數列最有可能是相關數列中每個離群值的原因。會針對**要顯示的數列**設定上個別數列清單中包含的每個目標數列，執行離群值主要原因分析。

輸出

互動式離群值表格與圖表

每個相關數列的離群值及這些離群值主要原因的表格與圖表。表格包含每個離群值的單一列。圖表是影響圖。在表格中選取一列會強調顯示影響圖中的路徑，該路徑從相關數列指向最可能造成相關聯離群值的數列。

離群值的樞紐表

每個相關數列的離群值及這些離群值主要原因的表格。這個表格包含的資訊與互動式顯示畫面中的表格資訊相同。這個表格支援所有用來將表格置於樞軸上並進行編輯的標準功能。

原因水準

您可以指定要包含在主要原因搜尋中的水準數目。這裡使用的水準概念與針對影響圖說明的概念相同。

所有模型中的模型適合度

所有模型的模型適合度及所選取適合度統計量直方圖。提供了下列適合度統計量：

R 平方(R)

線性模型的適合度量數，有時也稱為判斷係數。由模型所解釋的目標變數中變異的比例。它的範圍值介於 0 到 1 之間。值越小，表示模型越不適合資料。

均方根百分比誤差

測量模型預測值與數列觀察值差異的程度。其與使用的單元無關，因此可用來比較具有不同單元的數列。

均方根誤差(Q)

平均均方根誤差。是相依系列與其模型預測層級差異程度的量數，用與相依系列相同的單位來表示。

BIC Bayesian 資訊準則。用於根據 -2 減少對數概似值來選取及比較模型的量數。數值越小代表模式越佳。BIC 也會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型），但比 AIC 更嚴格。

AIC Akaike 資訊準則。用於根據 -2 減少對數概似值來選取及比較模型的量數。數值越小代表模式越佳。AIC 會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型）。

一段時間內的離群值

在估計期間的每一個時間間隔，所有目標中的離群值數目的長條圖。

數列轉換

已套用至模型系統中數列的任何轉換的表格。可能的轉換是遺漏值插補、總計與分佈。

估計期間

依預設，估計期間啟動於所有數列中最早觀察的時間，在所有數列中最晚觀察的時間結束。

依啟動與結束時間

您可以同時指定估計期間的啟動與結束時間，也可以只指定啟動時間或結束時間。如果您省略估計期間的啟動或結束時間，則會使用預設值。

- 如果觀察是由日期/時間欄位來定義，請以日期/時間欄位使用的相同格式輸入啟動時間與結束時間的值。
- 如果觀察是由循環期間定義，請為每個循環期間欄位指定一個值。每個欄位都會顯示在個別的直欄中。

依最晚或最早時間間隔

將估計期間定義為資料中從最早時間間隔開始或在最晚時間間隔的指定時間間隔數（可包含選用偏移）結束。在此環境定義中，時間間隔是指分析的時間間隔。例如，假定觀察是按月份，而分析的時間間隔是按季度。指定**最晚**並為**時間間隔數**指定值 24 表示最晚 24 個季度。

您可以選擇性地排除指定的時間間隔數。例如，指定最晚 24 個時間間隔並指定 1 代表要排除的數目，表示估計期間由最後一個間隔之前的 24 個間隔組成。

模型選項

模型名稱

您可以對模型指定自訂名稱，也可以接受自動產生的名稱，即 *TCM*。

預測　用於**將記錄延伸至未來**的選項可設定要在估計期間結束以外預測的時間間隔數目。在這種情況下，時間間隔是分析的時間間隔，它在「資料規格」標籤上指定。要求預測時，會針對也不是目標的任何輸入數列自動建置自迴歸模型。然後，在預測期間使用這些模型產生那些輸入數列的值。此設定沒有上限限制。

互動式輸出

時間原因建模的輸出包括數個互動式輸出物件。透過啟動（按兩下）「輸出檢視器」中的物件，可提供互動式特性。

整體模型系統

顯示模型系統中數列之間的原因關係。將特定目標連接至輸入的所有行都具有相同的顏色。行的厚度指示原因連線的顯著性，其中更厚的行代表更顯著的連線。也不是目標的輸入以黑色方形表示。

- 您可以顯示最上層模型、指定數列、所有數列或沒有輸入之模型的關係。最上層模型是符合為**要顯示的數列**設定上最適模型指定之準則的模型。
- 可透過選取圖表中的數列名稱，用滑鼠右鍵按一下，然後從快速功能表中選擇**建立影響圖**，來產生一個以上數列的影響圖。
- 您可以選擇隱藏其顯著性水準大於指定值的原因關係。顯著性水準越小，表示原因關係越顯著。
- 您可以透過選取圖表中的數列名稱，用滑鼠右鍵按一下，然後從快速功能表中選擇**強調顯示數列關係**，來顯示特定數列的關係。

影響圖

顯示相關數列與它影響或影響它的其他數列之間原因關係的圖形表示法。影響相關數列的數列稱為原因。

- 您可以指定想要的數列名稱來變更相關數列。按兩下影響圖中的任何節點，可將相關數列變更為與該節點相關聯的數列。
- 您可以在原因與效應之間切換顯示，您也可以變更要顯示的原因或效應的水準數目。
- 只需按一下任意節點，就能開啟與該節點相關聯之數列的詳細序列圖。

離群值主要原因分析

判定哪些數列最有可能是相關數列中每個離群值的原因。

- 您可以在「離群值」表格中選取任何離群值的列，來顯示離群值的主要原因。您也可以在序列圖中按一下離群值的圖示來顯示主要原因。
- 只需按一下任意節點，就能開啟與該節點相關聯之數列的詳細序列圖。

整體模型品質

所有模型及特定適合度統計量的模型適合度直方圖。按一下長條圖中的長條可過濾點狀圖，以便它僅顯示與所選長條相關聯的模型。您可以指定特定目標數列的名稱來尋找點狀圖中該數列的模型。

離群值分佈

在估計期間的每一個時間間隔，所有目標中的離群值數目的長條圖。按一下長條圖中的長條可過濾點狀圖，以便它僅顯示與所選長條相關聯的離群值。

TCM 模型塊

TCM 建模作業將建立多個具有字首 $TCM 的新欄位，如下表格所示。

表 *27.* 由 *TCM* 建模作業建立的新欄位

欄位名稱	說明
$TCM-*colname*	模型針對每一個目標數列預測的值。
$TCMLCI-*colname*	每個已預測數列的信賴區間下限值。

表 27. 由 *TCM* 建模作業建立的新欄位 (繼續)

$TSUCI-*colname*	每個已預測數列的信賴區間上限值。
$TCMResidual-*colname*	每欄產生的模型資料中的雜訊殘差值。

TCM 模型塊設定

「設定」標籤為 TCM 模型塊提供了額外的選項。

預測

用於**將記錄延伸至未來**的選項可設定要在估計期間結束以外預測的時間間隔數目。在這種情況下，時間間隔為 TCM 節點的「資料規格」標籤上指定的分析時間間隔。要求預測時，會針對任何不是目標的輸入序列自動建置自動迴歸模型。然後，在預測期間使用這些模型產生那些輸入數列的值。

可用於評分

為要評分的每一個模型建立新欄位。 可讓您指定要為將進行評分之每個模型所建立的新欄位。

- **噪音殘差。** 如果已選取了此選項，那麼對於每個目標欄位，將為模型殘差建立新欄位（帶有預設字首 $TCM-），並同時建立這些值的總計。
- **信賴區間上限及下限。** 如果已選取了此選項，那麼對於每個目標欄位，將分別為信賴區間上限和下限建立新欄位（帶有預設字首 $TCM-），並同時建立這些值的總計。

包括的用來評分的目標。 選取可用的目標以包括在模型評分中。

時間原因模型實務

「時間原因模型實務」程序會使用作用中資料集中的資料，來執行時間原因模型系統的使用者定義實務。實務由稱為根序列的時間序列以及該序列在指定時間範圍內的一組使用者定義值來定義。隨後，指定的值用來為根序列影響的時間序列產生預測。該程序需要由「時間原因建模」程序建立的模型系統檔案。假設作用中資料集的資料與用來建立模型系統檔案的資料相同。

範例

商業決策制訂者使用「時間原因建模」程序探索到的關鍵度量值，會影響數個重要的效能指標。度量值是可控制的，因此決策制訂者想要調查度量值在下個季度中各種值集的效果。可以透過將模型系統載入「時間原因模型實務」程序並指定關鍵度量值的值集，來輕鬆執行該調查。

定義實務期間

實務期間是您指定用來執行實務的值所在的期間。它可以在估計期間結束之前或之後啟動。您可以選擇性地指定在實務期間結束範圍之外進行預測。依預設，會透過實務期間結束來產生預測。所有實務都是使用相同的實務期間和規格來瞭解預測範圍。

註：預測啟動於實務期間啟動之後的第一個時段。例如，如果實務期間啟動於 2014-11-01，且時間間隔是月份，則第一個預測是 2014-12-01。

依啟動時間、結束時間和預測經歷時間來指定

- 如果觀察是由日期/時間欄位來定義，請以日期/時間欄位使用的相同格式輸入啟動時間、結束時間以及預測經歷時間值。日期/時間欄位的值與相關聯時間間隔的開頭對齊。例如，如果分析的時間間隔是月份，則值 10/10/2014 與月份開頭 10/01/2014 對齊。
- 如果觀察是由循環期間定義，請為每個循環期間欄位指定一個值。每個欄位都會顯示在個別的直欄中。

依相對於估計期間結束的時間間隔來指定

根據相對於估計期間結束的時間間隔來定義啟動與結束時間，其中時間間隔是分析時間間隔。估計期間結束定義為時間間隔 0。估計期間結束之前的時間間隔值為負數，估計期間結束之後的間隔值為正數。您也可以指定要在實務期間結束範圍之外進行預測的間隔數。預設值為 0。

例如，假設分析時間間隔是月份，而您指定 1 表示啟動間隔，3 表示結束間隔，以及 1 表示在該範圍之外要預測的範圍。然後，實務期間為估計期間結束之後的 3 個月。會針對實務期間的第二個與第三個月，以及在實務期間結束之後的 1 個月產生預測。

新增實務及實務群組

「實務」標籤指定要執行的實務。若要定義實務，您必須先按一下**定義實務期間**來定義實務期間。透過按一下相關聯的**新增實務**或**新增實務群組**按鈕，來建立實務和實務群組（僅適用於多維度資料）。您可以透過選取相關聯網格中的特定實務或實務群組，來進行編輯、複製或刪除。

直欄型資料

網格中的**根欄位**直欄可指定時間序列欄位，其值會取代為實務值。**實務值**直欄以最早到最晚的順序來顯示指定的實務值。如果實務值由表示式來定義，則該直欄會顯示表示式。

多維度資料

個別實務

「個別實務」網格中的每一列皆可指定時間序列，其值會取代為指定的實務值。該數列是由**根度量值**直欄中指定的欄位與指定給每個維度欄位的值這兩者的組合來定義的。**實務值**直欄的內容與直欄型資料的內容相同。

實務群組

實務群組根據單一根度量值欄位及多個維度值集來定義一組實務。指定的度量值欄位的每組維度值（每個維度欄位一個值）定義一個時間序列。隨後會對每個這類時間序列產生一個個別實務，然後其值會取代為實務值。實務群組的實務值由表示式指定，該表示式隨後會套用到群組中的每個時間序列。

數列計數直欄會顯示與實務群組相關聯的維度值集數。顯示的值可以大於與實務群組相關聯的時間序列的實際數目（每集一個數列）。如果部分指定的維度值組合未對應於群組根度量值包含的數列，則會發生這種情況。

作為實務群組範例，請考量度量值欄位 *advertising* 和兩個維度欄位 *region* 和 *brand*。您可以定義一個基於將 *advertising* 作為根度量值並包括 *region* 和 *brand* 的所有組合的實務群組。然後，您可以將 `advertising*1.2` 指定為表示式，以調查對與 *advertising* 欄位相關聯的每個時間序列增加 20% *advertising* 造成的影響。如果 *region* 有 4 個值，而 *brand* 有 2 個值，則有 8 個此類時間序列，因此群組會定義 8 個實務。

實務定義： 用來定義實務的設定視您的資料是直欄型資料還是多維度資料而定。

根數列

指定實務的根數列。每個實務都是基於單一根數列。若為直欄型資料，請選取定義根數列的欄位。若為多維度資料，請將項目新增至包含數列之度量值欄位的網格中，來指定根數列。然後指定定義根數列的維度欄位值。下列內容適用於指定維度值：

- 您可以將每個維度欄位的值直接輸入到網格中，也可以從可用維度值清單中選取。若要從可用維度值清單中選取，請在您想要之維度的資料格中，按一下省略符號按鈕。執行此動作會開啟「選取維度值」子對話框。

- 您可以在「選取維度值」子對話框中，按一下雙目望遠鏡圖示並指定搜尋詞彙，來搜尋維度值清單。空格會被視為搜尋詞彙的一部分。搜尋詞彙中的星號 (*) 不表示萬用字元。

指定受影響目標

當您知道根數列影響的特定目標，且僅想調查對這些目標產生的影響時，使用這個選項。依預設，會自動判定受根數列影響的目標。您可以使用「選項」標籤上的設定，來指定受實務影響的數列幅度。

若為直欄型資料，請選取想要的目標。若為多維度資料，請將項目新增至包含數列之目標度量值欄位的網格中，來指定目標數列。依預設，會包括指定度量值欄位中包含的所有數列。您可以自訂一或多個維度欄位所包括的值，來自訂併入的數列集。若要自訂包括的維度值，請按一下想要之維度的省略符號按鈕。執行此動作會開啟「選取維度值」對話框。

數列計數直欄（針對多維度資料）會顯示目前為相關聯目標度量值指定的維度值集數。顯示的值可以大於受影響目標數列的實際數目（每集一個數列）。如果指定的部分維度值組合未對應於相關聯目標度量值包含的數列，則會發生這種情況。

實務 ID

每個實務都必須具有唯一 ID。該 ID 顯示在與實務相關聯的輸出中。ID 值除了唯一性以外，沒有任何其他限制。

為根數列指定實務值

使用這個選項，在實務期間為根數列指定明確值。您必須為網格中所列的每個時間間隔指定一個數值。您可以按一下**讀取**、**預測**或**讀取@3977測**，在實務期間取得每個間隔的根數列值（實際值或預測值）。

為根數列的實務值指定表示式

您可以定義表示式，以在實務期間計算根數列的值。您可以直接輸入表示式，或按一下計算機按鈕，從「實務值表示式建置器」建立表示式。

- 表示式可以包含模型系統中的任何目標或輸入。
- 如果實務期間超出現有資料，則表示式會套用至表示式中欄位的預測值。
- 對於多維度資料，表示式中的每一個欄位都指定由此欄位及為根度量值指定的維度值定義的時間序列。時間序列是用來對表示式求值的那些時間序列。

例如，假設根欄位為 *advertising* 而表示式為 `advertising*1.2`。則實務中所使用 *advertising* 的值代表對現有值增加 20%。

註：可透過按一下「實務」標籤上的**新增實務**來建立實務。

選取維度值： 若為多維度資料，您可以自訂維度值來定義受實務或實務群組影響的目標。您也可以自訂維度值來為實務群組定義根數列集。

所有值

指定包括現行維度欄位的所有值。此選項是預設值。

選取值

使用此選項來指定現行維度欄位的值集。您可以過濾可從中選擇的值集。符合過濾條件的值會出現在**未選取的值**清單的**相符**標籤中，而不符合過濾條件的值則會出現在**不符合**標籤中。無論任何過濾條件，**所有**標籤都會列出所有未選取的值。

- 指定過濾器時，您可以使用星號 (*) 來表示萬用字元。
- 若要清除現行過濾器，請在「過濾顯示的值」對話框上，為搜尋詞彙指定空白值。

若要自訂受影響目標的維度值，請執行下列動作：

1. 從「實務定義」或「實務群組定義」對話框中，選取您要為其自訂維度值的目標度量值。

2. 在您要自訂之維度的直欄中，按一下省略符號按鈕。

若要為實務群組的根數列自訂維度值，請執行下列動作：

1. 從「實務群組定義」對話框中，按一下您要自訂之維度的省略符號按鈕（位於根數列網格中）。

實務群組定義：

根數列

指定實務群組的根數列集。會針對集中的每個時間序列產生個別實務。將項目新增至包含所要數列之度量值欄位的網格中，來指定根數列。然後，可指定用來定義該集的維度欄位值。依預設，包括所指定根度量值欄位中包含的所有數列。您可以自訂一或多個維度欄位所包括的值，來自訂併入的數列集。若要自訂包括的維度值，請按一下維度的省略符號按鈕。執行此動作會開啟「選取維度值」對話框。

數列計數直欄會顯示相關聯根度量值目前包括的維度值集數。顯示的值可以大於實務群組的根數列的實際數目（每集一個數列）。如果指定的部分維度值組合未對應於根度量值包含的數列，則會發生這種情況。

指定受影響目標數列

當您知道根數列集影響的特定目標，且僅想要調查對這些目標產生的影響時，使用這個選項。依預設，會自動判定受每個根數列影響的目標。您可以使用「選項」標籤上的設定，來指定受每個個別實務影響的數列幅度。

將項目新增至包含數列之度量值欄位的網格中，來指定目標數列。依預設，會包括指定度量值欄位中包含的所有數列。您可以自訂一或多個維度欄位所包括的值，來自訂併入的數列集。若要自訂包括的維度值，請按一下想要之維度的省略符號按鈕。執行此動作會開啟「選取維度值」對話框。

數列計數直欄會顯示目前為相關聯目標度量值指定的維度值集數。顯示的值可以大於受影響目標數列的實際數目（每集一個數列）。如果指定的部分維度值組合未對應於相關聯目標度量值包含的數列，則會發生這種情況。

實務 ID 字首

每個實務群組都必須具有唯一字首。該字首用來建構顯示在與實務群組中每個個別實務相關聯之輸出中的 ID。個別實務的識別為字首後接底線，再接用來識別根數列的每個維度欄位值。維度值以底線分隔。字首值除了唯一性以外，沒有任何其他限制。

根數列的實務值表示式

用於實務範例群組的實務範例值由表示式指定，該表示式然後用於計算群組中每一個根數列的值。您可以直接輸入表示式，或按一下計算機按鈕，從「實務值表示式建置器」建立表示式。

- 表示式可以包含模型系統中的任何目標或輸入。
- 如果實務期間超出現有資料，則表示式會套用至表示式中欄位的預測值。
- 針對群組中的每個根數列，表示式中的欄位指定這些欄位所定義的時間序列，以及定義根數列的維度值。時間序列是用來對表示式求值的那些時間序列。例如，如果根數列由 `region='north'` 和 `brand='X'` 定義，則表示式中所用的時間序列是由這些相同的維度值所定義。

例如，假設根度量值欄位是 *advertising* 且存在兩個維度欄位 *region* 和 *brand*。還假設實務群組包括維度欄位值的所有組合。然後，您可以將 `advertising*1.2` 指定為表示式，以調查針對與 *advertising* 欄位相關聯的每個時間序列增加 20% *advertising* 所造成的影響。

註：實務群組僅適用於多維度資料，且可透過按一下「實務」標籤上的**新增實務群組**來建立。

選項

受影響目標的水準上限

指定受影響目標的水準數目上限。每個連續水準（最多 5 個）都包括根數列較間接影響的目標。具體而言，第一個水準包括將根數列作為直接輸入的目標。第二個水準中的目標將第一個水準中的目標作為直接輸入，依此類推。增加此設定的值會增加計算的複雜性，因此可能會影響效能。

自動偵測到的目標數上限

指定為每個根數列自動偵測到的受影響目標數目上限。增加此設定的值會增加計算的複雜性，因此可能會影響效能。

影響圖

顯示每個實務的根數列與它影響的目標數列之間原因關係的圖形表示法。受影響目標的實務值與預測值兩者的表格都會併入作為輸出項目的一部分。該圖形包括受影響目標的預測值圖。在影響圖中按一下任意節點，就會開啟與該節點相關聯之數列的詳細序列圖。系統會為每個實務產生個別影響圖。

數列圖

為每個實務中的每個受影響目標，產生預測值的數列圖。

預測與實務表格

每個實務的預測值與實務值的表格。這些表格包含的資訊與影響圖中的表格資訊相同。這些表格支援所有用來旋轉及編輯表格的標準功能。

在圖形與表格中併入信賴區間

指定是否同時在圖表和表格輸出中併入實務預測的信賴區間。

信賴區間寬度 (%)

此設定可控制實務預測的信賴區間。您可以指定小於 100 的任何正數值。依預設，將使用 95% 信賴區間。

「時間序列」節點

「時間序列」節點可以在本端或分散式環境中與資料配合使用；在分散式環境中，可以利用 IBM SPSS Analytic Server 的能力。通過此節點，可以選擇對時間序列的指數平滑化模型、單變異數自身迴歸整合移動均數 (ARIMA) 及多變異數 ARIMA（或轉換函數）模型進行估計和建立，並根據時間序列資料產生預測。

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。照這樣，指數平滑化並非基於對資料的理論了解。它一次預測一個點，隨著新資料的進入對預測進行調整。技術用於預測展示趨勢、週期性或兩者的數列。您可從以不同方式處理趨勢和週期性的各種指數平滑化模型中進行選擇。

與指數平滑化模型相比，**ARIMA** 模型在對趨勢和季節元件建模方面提供更成熟的方法，特別是，增加了可在模型中包含自變數（預測值）的優勢。這項作業涉及指定自我迴歸和移動平均階數，以及差分次數。您可以包括預測值變數並為任何或全部變數定義轉換函數，以及指定自動偵測偏離值或明確的一組偏離值。

註：在實際應用中，如果要包含預測值（這些預測值可能有助於說明正在預測的數列的行為，例如郵寄的型錄的號碼或某公司網頁的點擊數），那麼 ARIMA 模型最有用。指數平滑化模型說明時間序列的行為，而不嘗試理解它為何如此行事。例如，過去每 12 個月達到最大值的數列可能繼續保持該行為，即使您不瞭解其原因也是如此。

還提供了**專家建模器**選項，此選項將試圖自動識別和估計對一個或多個目標變數配適度最高的 ARIMA 模型或指數平滑化模型，從而無需通過試錯來識別適當的模型。如果您有任何疑問，請使用「專家建模器」選項。

如果指定了預測值，那麼專家建模器會選取將那些在統計意義上與相依數列具有顯著關係的變數包括在 ARIMA 模型中。若適當，可利用差分及/或平方根或自然對數進行模型變數轉換。依預設，Expert Modeler 會考量所

有指數平滑化模型及所有 ARIMA 模型，並為每一個目標欄位挑選其中的最佳模型。但您可以限制 Expert Modeler 僅挑選最佳的指數平滑化模型，或僅挑選最佳的 ARIMA 模型。您也可以指定自動偵測偏離值。

「時間序列」節點 - 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。 選取一個或多個欄位作為預測目標。

候選輸入。選取一或多個欄位作為預測的輸入。

事件及人為介入。 使用該區域來指定某些輸入欄位以作為事件或人為介入欄位。此指定可將欄位視為包含可受事件（可預期的重複出現的狀況，例如，促銷）或人為介入（一次性事件，例如，電源中斷或員工罷工）影響的時間序列資料。您選取的欄位必須為具有整數儲存的旗標。

「時間序列」節點 - 資料規格選項

通過「資料規格」標籤，您可以設定用於將資料包含在模型中的所有選項。只要同時指定**日期/時間欄位**和**時間間隔**，便可以按一下**執行**按鈕來建立包含所有預設選項的模型，但通常您會想要根據自己的用途自訂建立。

該標籤包含數個不同窗格，您可以在其中設定特定於您模型的自訂作業。

「時間序列」節點 - 觀察值

使用此窗格中的設定可以指定用於定義觀察的欄位。

由日期/時間欄位指定的觀察

您可以指定觀察由日期、時間或時間戳記欄位定義。除了定義觀察的欄位以外，還可以選取說明觀察的適當時間間隔。根據指定的時間間隔，您還可以指定其他設定，例如觀察間隔（增量）或每週的天數。下列考量僅適用於時間間隔：

- 如果觀察的時間（例如處理銷售單的時間）間隔無規律，請使用值**無規律**。如果選取**無規律**，您必須從「資料規格」標籤上的**時間間隔**設定中，指定用於分析的時間間隔。
- 如果觀察代表日期和時間，而時間間隔是小時數、分鐘數或秒數，則使用**每天小時數**、**每天分鐘數**或**每天秒數**。如果觀察代表無日期參照的時間（期間），且時間間隔為小時數、分鐘數或秒數，則使用**小時數（非週期）**、**分鐘數（非週期）**或**秒數（非週期）**。
- 根據選取的時間間隔，程序可以偵測遺漏的觀察。由於此程序假定所有觀察之間的時間間距相等，並假定未遺漏觀察，因此有必要偵測遺漏的觀察。例如，如果時間間隔為「天」，並且日期 2015-10-27 後跟 2015-10-29，那麼表示遺漏 2015-10-28 的觀察。對於任何遺漏觀察，將插補值；使用「資料規格」標籤的**遺漏值處理**區域可以指定用於處理遺漏值的設定。

- 指定的時間間隔可讓程序偵測相同時間間隔中需要總計在一起的多個觀察，並依據間隔界限（例如月份的第一天）排列觀察，以確保觀察間隔相等。例如，如果時間間隔為「月」，則會將相同月份內的多個日期總計在一起。這種類型的總計稱為分組。依預設，在分組觀察時會加總觀察。您可以從「資料規格」標籤上的**總計及分佈**設定中指定不同的分組方法，如觀察平均數。
- 對於部分時間間隔，其他設定可定義一般相等時間間隔中的岔斷。例如，如果時間間隔為「日」，但只有工作日有效，則您可以指定一週內有五天，一週從星期一啟動。

定義為期間或週期性時期的觀察

觀察可由一或多個代表期間或重複循環期間（多達任意數目的循環水準）的整數欄位定義。借助此結構，您可以說明任何標準時間間隔都無法支持的觀察數列。例如，可使用代表年份的循環欄位以及代表月份的期間欄位來說明只有 10 個月的財政年度，其中一個循環的長度為 10。

指定循環期間的欄位可定義週期性水準的階層，其中最低水準由**期間**欄位定義。下一個最高水準由水準為 1 的循環欄位指定，接著由水準為 2 的循環欄位指定，依此類推。除最高層次以外，每個層次的欄位值對於次高層次都必須具有週期性。最高水準的值不能為週期性值。例如，對於由 10 個月組成的財年，月在年中具有週期性，而年不具有週期性。

- 特定水準上的循環長度是下一個最低水準的週期。以財政年度為例，只有一個循環水準且循環長度為 10，因為下一個最低水準代表月份，而指定的財政年度有 10 個月。
- 請指定任何並非起始於 1 的週期性欄位的起始值。要偵測遺漏值，有必要進行此設定。例如，如果週期性欄位從 2 啟動，但起始值指定為 1，則程序會假設該欄位每個循環中的第一個期間都存在遺漏值。

「時間序列」節點 - 分析時間間隔

用於分析的時間間隔可以與觀察的時間間隔不同。例如，觀察的時間間隔為「天」時，您可以選擇「月」用作進行分析的時間間隔。系統先將每日資料聚合為每月資料，然後再建立模型。您也可以選擇將資料分佈從較長時間間隔改成較短時間間隔。例如，如果觀察是按季度發生的，則您可以將資料從按季度分佈改成按月分佈。

使用此窗格中的設定指定用於分析的時間間隔。聚集或分佈資料的方法是在「資料規格」標籤上的**聚集和分佈**設定中指定的。

執行分析的時間間隔的可用選項，視觀察的定義方式以及這些觀察的時間間隔而定。特別是，如果觀察由週期性時期定義，那麼僅受支援聚集。在此情況下，分析的時間間隔必須大於或等於觀察的時間間隔。

「時間序列」節點 - 聚集和分佈選項

使用此窗格中的設定可以指定用於對觀察的時間間隔的相關輸入資料進行聚合或分佈的設定。

聚集函數

當用於分析的時間間隔長於觀察的時間間隔時，會總計輸入資料。例如，如果觀察時間間隔是「日」，而分析的時間間隔是「月」，則會執行總計。提供了下列聚集函數：平均數、加總、眾數、最小值或最大值。

分佈函數

當用於分析的時間間隔短於觀察的時間間隔時，會分佈輸入資料。例如，如果觀察時間間隔是「季度」，而分析的時間間隔是「月」，則會執行分佈。提供了下列分佈函數：平均數或加總。

分組函數

如果觀察由日期/時間定義且在相同的時間間隔發生多個觀察，則可套用分組。例如，如果觀察的時間間隔是「月」，則相同月份中的多個日期會分組在一起，並與其發生所在的月份相關聯。您可以使用的分組函數如下：mean、sum、mode、min 或 max。如果觀察由日期/時間定義，且觀察的時間間隔指定為「無規律」，則一律會執行分組。

註：雖然分組是一種總計形式，但它是在處理所有遺漏值之前執行，而正式的總計是在處理完所有遺漏值之後執行。當觀察的時間間隔指定為「無規律」時，只能使用分組函數來執行總計。

總計跨日期觀察與前一天的值

指定是否將時間跨天界限的觀察值聚合到前一天的值。例如，對於在 20:00 開始的 8 小時一天的每小時觀察值，此設定指定是否將介於 00:00 與 04:00 之間的觀察值包含在前一天的聚合結果中。只有在觀察時間間隔是「每天小時數」、「每天分鐘數」或「每天秒數」，且分析的時間間隔為「日」時，此設定才適用。

所指定欄位的自訂設定

您可以逐個欄位指定聚集函數、分佈函數及分組函數。這些設定會置換聚集函數、分佈函數及分組函數的預設值。

「時間序列」節點 - 遺漏值選項

使用此窗格中的設定可以指定輸入資料中要取代為插補值的遺漏值數。提供了下列取代方法：

線性插補

使用線性插補取代遺漏值。該選項會用遺漏值出現之前的最後一個有效值，還有遺漏值出現之後的第一個有效值當作內插。如果數列中的第一個或最後一個觀察具有遺漏值，則會使用數列開頭或結尾的兩個最近非遺漏值。

數列平均數

以整個數列的平均數置換遺漏值。

鄰近點平均數

以有效周圍值的平均數置換遺漏值。附近點的指距是指用來計算平均數的遺漏值，其之前及之後的有效值個數。

鄰近點中位數

以有效周圍值的中位數置換遺漏值。附近點的指距是指用來計算中位數的遺漏值，其之前及之後的有效值個數。

線性趨勢

此選項使用數列中的所有非遺漏觀察來擬合簡單線性迴歸模型，該模型隨後用於插補遺漏值。

其他設定：

最低資料品質分數 (%)

針對時間變數以及對應於每個時間序列的輸入資料計算資料品質測量。如果資料品質分數低於此臨界值，那麼將捨棄對應的時間序列。

「時間序列」節點 - 估計期間

在「估計期間」窗格中，您可以指定要在模型估計中使用的記錄的範圍。依預設，估計期間啟動於所有數列中最早觀察的時間，在所有數列中最晚觀察的時間結束。

依啟動與結束時間

您可以同時指定估計期間的啟動與結束時間，也可以只指定啟動時間或結束時間。如果您省略估計期間的啟動或結束時間，則會使用預設值。

- 如果觀察值由日期/時間欄位定義，請以用於該日期/時間欄位的格式輸入開始時間值和結束時間值。
- 如果觀察是由循環期間定義，請為每個循環期間欄位指定一個值。每個欄位都會顯示在個別的直欄中。

依最晚或最早時間間隔

將估計期間定義為指定號碼的時間間隔，這些時間間隔從資料中的最早時間間隔開始或以最晚時間間隔結束，並具有選用偏移量。在此環境定義中，時間間隔是指分析時間間隔。例如，假定觀察是按月份，而分析的時間間隔是按季度。指定**最晚**並為**時間間隔數**指定值 24 表示最晚 24 個季度。

您可以選擇性地排除指定的時間間隔數。例如，指定最晚 24 個時間間隔並指定 1 代表要排除的數目，表示估計期間由最後一個間隔之前的 24 個間隔組成。

「時間序列」節點 - 建置選項

通過「數據規範」標籤，您可以設定用於建立模型的所有選項。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。

此標籤包含兩種不同的窗格，您可以在這些窗格中設定特定於模型的自訂作業內容。

「時間序列」節點 - 一般建置選項

此窗格中的可用選項取決於您從**方法**清單中選擇下列三項設定中的哪一項：

- **Expert Modeler**。 選擇此選項以使用 Expert Modeler，此組件將自動為每個相依數列尋找配適度最高的模型。
- **指數平滑化**。 使用此選項，可指定自訂指數平滑化模式。
- **ARIMA 程序** 使用此選項，可指定自訂 ARIMA 模式。

Expert Modeler

在**模型類型**下，選取您要建立的模型的類型：

- **所有模式**。 Expert Modeler 會同時考量 ARIMA 和指數平滑化模式。
- **僅指數平滑化模式**。Expert Modeler 僅考慮指數平滑化模型。
- **僅 ARIMA 模式**。Expert Modeler 僅考慮 ARIMA 模型。

Expert Modeler 會考量週期性模型。 只有在為作用中資料集定義了週期性性時才啟用此選項。選取此選項時，Expert Modeler 會同時考量週期性和非週期性模型。如果未選取此選項，那麼 Expert Modeler 將僅考慮非週期性模型。

Expert Modeler 考慮複雜指數平滑化模型。選取了此選項時，Expert Modeler 將搜尋所有 13 個指數平滑化模型（其中 7 個存在於原始時間序列節點中，而剩下 6 個是18.1 版 中新增的節點）。如果未選取此選項，那麼 Expert Modeler 將搜尋原始的 7 個指數平滑化模型。

在**離群值**下，從下列選項中進行選取

自動偵測偏離值。 依預設，不會執行自動偵測偏離值。選取此選項以執行自動偵測離群值，然後選取所期望的離群值類型：

輸入欄位必須具有旗標、列名或序數測量層次，並且必須是數值（例如，對於旗標欄位，必須為 1/0，而非 True/False），才能包含在此清單中。

對於在**欄位**標籤上 ID 為事件欄位或人為介入欄位的輸入，Expert Modeler 僅考慮簡單迴歸方法而不是任意變換函數。

指數平滑化

模式類型。 指數平滑化模型分類為週期性模型或非週期性模型。[1]只有使用「資料規格」標籤上的「時間間隔」窗格定義的週期性為週期性週期性時，才可以使用週期性模型。週期為：循環期間、年、季度、月、每週天數、每日小時數、每日分鐘數及每日秒數。提供了下列模型類型：

- **簡單。** 此模型適用於不具有趨勢或週期性的數列。層級是它唯一的相關平滑化參數。簡式指數平滑化與 ARIMA 最為相似，都具有零個自身迴歸階數、一個差分階數、一個移動平均數階數，且沒有常數。
- **Holt 線性趨勢。** 此模型適用於具有線性趨勢且沒有週期性的序列。其相關平滑化參數為層級與趨勢，且在此模型中不受彼此的數值所限制。 Holt's 模式較 Brown's 模式更為普遍，但對於大型數列的預估值計算時間會較長。 Holt 指數平滑化與 ARIMA 最為相似，都具有零個自身迴歸階數、兩個差分階數，以及兩個移動平均階數。
- **減幅趨勢。** 此模型適用於具有線性趨勢且具有漸失線性趨勢但沒有週期性的數列。其相關平滑化參數為層級、趨勢和減幅趨勢。減幅指數平滑化與 ARIMA 最為相似，都具有一個自身迴歸階數、一個差分階數，以及兩個移動平均數階數。
- **相乘性趨勢。**該模型適合於具有一種隨數列長度而變的趨勢且沒有週期性的數列。其相關的平滑參數是層次和趨勢。相乘性趨勢指數平滑化與任何 ARIMA 模型都不相似。
- **Brown 線性趨勢。** 此模型適用於具有線性趨勢且沒有週期性的序列。其相關平滑化參數為層級與趨勢，且在此模型中皆已假定為相等。因此，Brown's 模式為 Holt's 模型的特殊觀察值。Brown 指數平滑化與 ARIMA 最為相似，都具有零個自身迴歸階數、兩個差分階數以及兩個移動平均數階數，且移動平均數的第二階係數與第一階的二分之一係數平方相等。
- **簡單週期性。** 此模型適用於沒有趨勢的數列，以及在一段時間保持不變的週期效果。其相關平滑化參數為層級和週期。週期可加性指數平滑化與 ARIMA 最為相似，都具有零個自身迴歸階數、一個差分階數、一個週期差分階數及一個、p 個及 $p + 1$ 個移動平均數階數，其中 p 為週期區間內的期間個數。對於每月資料，p=12。
- **Winters 可加性。** 此模型適用於具有線性趨勢的數列，以及在一段時間保持不變的週期效果。其相關平滑化參數可為層級、趨勢和週期。 Winters 可加性指數平滑化與 ARIMA 最為相似，都具有零個自身迴歸階數、一個差分階數、一個週期差分階數及 $p + 1$ 個移動平均數階數，其中 p 為週期區間內的期間個數。對於每月資料，p=12。
- **具有可加性週期性的減幅趨勢。**此模型適合於具有逐漸消退的線性趨勢且季節作用不隨時間變化的數列。它的相對平滑化參數為層級、趨勢、減幅趨勢及週期。阻尼趨勢和可加性週期性指數平滑化與任何 ARIMA 模型都不相似。
- **具有可加性週期性的相乘性趨勢。**該模型適合於具有隨數列長度而變的趨勢且季節作用不隨時間變化的數列。其相關平滑化參數可為層級、趨勢和週期。相乘性趨勢和可加性週期性指數平滑化與任何 ARIMA 模型都不相似。
- **相乘性週期性。**此模型適合於不具有趨勢且季節作用隨數列長度而變的數列。其相關平滑化參數為層級和週期。相乘性週期性指數平滑化與任何 ARIMA 模型都不相似。
- **Winters 相乘性。** 此模型適用於具有線性趨勢的數列，以及隨著數列長度而變更的週期效果。其相關平滑化參數可為層級、趨勢和週期。 Winters 的可乘指數平滑法與任何 ARIMA 模型都不相似。
- **具有相乘性週期性的減幅趨勢。**此模型適合於具有逐漸消退的線性趨勢且季節作用隨數列的大小而變化的數列。它的相對平滑化參數為層級、趨勢、減幅趨勢及週期。減幅趨勢和相乘性週期性指數平滑化與任何 ARIMA 模型都不相似。

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

- **具有相乘性週期性的相乘性趨勢。**此模型適合於具有隨數列的長度發生變化的趨勢和季節作用的數列。其相關平滑化參數可為層級、趨勢和週期。相乘性趨勢和相乘性週期性指數平滑化與任何 ARIMA 模型都不相似。

目標轉換。 您可以指定先在各個因變數上執行轉換，然後再進行模式化作業。

- **無。** 沒有執行任何轉換。
- **平方根。** 執行平方根轉換。
- **自然對數。** 執行自然對數轉換。

ARIMA

指定自訂 ARIMA 模型的結構。

ARIMA 階數。 在網格的相應單元中，輸入模型的各個 ARIMA 成分的值。所有數值都必須為非負的整數。對自我迴歸和移動平均成份來說，該值表示最大階數。所有正較低階數都包括在模型中。例如，如果指定 2，那麼模型將包含階數數 2 和 1。只有在為作用中資料集定義了週期性性的情況下，才會啟用「週期性」欄中的單元。

- **自我迴歸** 。 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，自身迴歸階數數 2 指定數列中過去兩個時間段的值用於預測目前的值。
- **差分** 。 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。差異分析階數數對應於數列趨勢的程度 ; 第一階數差異分析表示線性趨勢，第二階數差異分析表示二次趨勢，依此類推。
- **移動平均數 (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。 週期性自我迴歸、移動平均數，和差分成份所扮演角色與其非週期性對等項目相同。但對週期性階數來說，目前序列值是受由一或多個週期性期間分隔的先前數列值影響。例如，對於月度資料（週期性期間為 12），週期性階數 1 表示目前數列值將受到目前期間之前 12 個期間的數列值的影響。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。

自動偵測偏離值。 選中此選項可以對離群值執行自動偵測，並選取可用的一個或多個離群值類型。

要偵測的偏離值的類型。 選取要偵測的離群值類型。支援的類型為：

- 可加性（預設值）
- 層級變遷（預設值）
- 創新
- 短暫
- 週期性相加(S)
- 局部趨勢
- 可加性補充(V)

轉換函數階數數和轉換。要指定轉換並為 ARIMA 模型中的任何或所有輸入欄位定義轉換函數，請按一下**設定**；這將顯示另一個對話框，您可以在其中輸入變換和轉換詳細資料。

常數項納入模式中。 納入常數是標準作業，除非您確定總平均序列值是 0；套用差分時，則建議排除常數。

其他詳細資料

- 有關離群值類型的進一步資訊，請參閱第 248 頁的『離群值』。
- 有關傳輸和轉換函數的進一步資訊，請參閱『轉換函數』。

轉換函數： 使用「轉換函數順序和轉換」對話框可以指定轉換以及為 ARIMA 模型中的任何或所有輸入欄位定義轉換函數。

目標轉換。在此窗格中，您可以指定對每個目標變數進行建模之前要對其執行的轉換。

- **無。** 沒有執行任何轉換。
- **平方根。** 執行平方根轉換。
- **自然對數。** 執行自然對數轉換。

候選輸入轉換函數及轉換。通過使用轉換函數，您可指定以何種方式使用輸入欄位的過去值來預測目標序列的未來值。左端窗格的清單中顯示了所有的輸入欄位。此窗格中的其餘資訊特定於您選取的輸入欄位。

轉換函數階數。 在**結構**網格的相應單元中，輸入轉換函數的各個成分的值。所有數值都必須為非負的整數。對分子和分母成份來說，該值表示最大階數。所有正較低階數都包括在模型中。此外，一定會包含階數 0 供分子成份使用。例如，如果將分子指定為 2，那麼模型將包含階數數 2、1 和 0。如果將分子指定為 3，那麼模型將包含階數數 3、2 和 1。只有在為作用中資料集定義了週期性性的情況下，才會啟用季節欄中的單元。

分子。 轉換函數的分子階數指定所選取不相依系列（預測值）中，哪些先前值用來預測不相依系列的現行值。例如，分子階數數 1 指定使用過去某個時段的相依系列的值（以及相依系列的目前的值）來預測各不相依系列的目前的值。

分母。 轉換函數的分母階數指定要如何使用所選取不相依系列（預測值）先前值的數列平均數離差來預測相依系列的現行值。例如，分母階數數 1 指定在預測每個相依系列的目前的值時，需要考慮與過去一個時間期間的不相依系列的平均數離差。

差異。 指定在估計模型之前套用至所選擇不相依系列（預測值）的差分階數。當趨勢存在時，就必須對數列進行差分，以移除趨勢的影響。

週期性。 週期性分子、分母，和差分成份所扮演角色與其非週期性對等項目相同。但對週期性階數來說，目前序列值是受由一或多個週期性期間分隔的先前數列值影響。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。

延遲。 設定延遲會造成輸入欄位的影響依指定區間數延遲。例如，如果延遲設定為 5，在時間 t 的輸入欄位值不會影響預測，要一直到經過五個週期 $(t + 5)$ 之後才會有影響。

轉換。 一組自變數的轉換函數規格，同時也包括要在這些變數上執行的選擇性轉換。

- **無。** 沒有執行任何轉換。
- **平方根。** 執行平方根轉換。
- **自然對數。** 執行自然對數轉換。

「時間序列」節點 - 建立輸出選項

ACF 和 PACF 輸出中的落階數目上限量。自相關係數 (ACF) 和局部自相關係數 (PACF) 用於測量目前數列值和過去數列值之間的相關性，並指示預測將來值時最有用的過去數列值。您可以設定自相關係數和局部自相關係數表格及圖形中顯示的落階數目上限。

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。通常，您希望將建模工作的主要精力放在最重要的預測值上，並考慮刪除或忽略那些最不重要的預測值。對於某些模型，計算預測值重要性（特別在處理大型資料集時）可能需要花費較長時間，因此依預設，預測值重要性對某些模型處於關閉狀態。

「時間序列」節點 - 模型選項

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。

信賴限制寬度 (%)。 信賴區間是供模型預測和殘差自我相關計算使用。您可以指定小於 100 的任何正數值。依預設，將使用 95% 信賴區間。

使用現有模式繼續估計。如果已產生一個「時間序列」模型，那麼選取此選項可以重新使用為該模型指定的準則設定，並在模型選用區中產生一個新的模式節點，而不必從頭建立一個新模型。這樣，您可以基於先前的模型設定但使用較新的資料來重新估算並生成新預測，從而節省時間。因此，例如，如果特定時間序列的原始模型為 Holt 線性趨勢，則可以使用相同類型的模型針對該資料進行重新估計及預測。系統不會重新試圖尋找最適合新資料的模型類型。

僅建立評分模式。要減少模型中儲存的資料數量，請選中此框。使用此選項可以在使用許多時間序列（數萬個）建立模型時提高效能。您仍可以按照常規方法對資料分數。

將記錄延伸到未來。啟用下列**要在預測中使用的未來值**部分，在該部分中可以設定估計期間結束後要預測的時間間隔數目量。在這種情況下，時間間隔是您在「資料規格」標籤上指定的分析的時間間隔。此設定沒有上限限制。通過使用下列選項，您可以自動計算輸入的未來值，或者手動為一個或多個預測值指定預測值。

要在用於預測的未來值

- **計算輸入的未來值** 如果選取此選項，那麼會自動計算預測值、雜訊預測、差異估計和未來時間值的預測值。要求預測時，會針對任何不是目標的輸入序列自動建置自動迴歸模型。然後，在預測期間使用這些模型產生那些輸入數列的值。
- **選取要將其值新增到資料中的欄位。**對於要預測的每條記錄（holdout 除外），如果您使用預測值欄位（角色設定為輸入），那麼可以為每個預測值的預測期間指定估計值。您可以手動指定值，也可以從清單中選擇。
  - **欄位。** 按一下欄位選取器按鈕，並選擇可以用作預測值的任何欄位。請注意，在這裡選取的欄位在建模時可以使用，也可能無法使用；若要實際將欄位用作預測值，則必須在下游建模節點中選取它。此對話框會為您直接提供方便的地方來指定未來值，因此可由多個下游建模節點共用而無需在每一個節點中單獨指定它們。另請注意，可用欄位；的清單可能會受到「建置選項」標籤中選項的約束。

    請注意，如果為串流中不再可用（因為已將其刪除或由於「建置選項」標籤中的選項更新）的欄位指定了未來值，那麼此欄位將顯示為紅色。
  - **值。** 針對每一個欄位，您可以從函數清單中選擇，或按一下**指定**以手動輸入值或從預定值清單中選擇。如果預測值欄位與您可以控制或可以提前知道的項目相關，則應手動輸入值。例如，如果您要根據房間預約數預測旅館下個月的收入，則可以指定該期間實際擁有的預約數。相反地，如果預測值欄位與您無法控制的某個項目相關，例如股價，則可以使用最近的值或最近點的平均數等函數。

    可用的函數視欄位的測量層級而定。

表 *28.* 可用於測量層次的函數

測量層級	函數
連續或名義欄位	空白 最近點的平均數 最新值 指定
旗標欄位	空白 最新值 TrueFalse指定

最近點的均數根據最後一欄三個資料點的均數計算未來值。

最近值將未來值設定為最近資料點的值。

true/假將旗標欄位的未來值設定為指定的 true 值或 false。

指定開啟一個對話框，用於手動指定未來值或從預先定義清單中選擇未來值。

可用於評分

您可以在此設定模型塊的對話框中顯示的評分選項的預設值。

- **計算置信限制的上限和下限。**如果選取了此選項，那麼對於每個目標欄位，將為信賴區間上限和下限建立新欄位（帶有預設字首 $TSLCI- 和 $TSUCI-）。
- **計算雜訊殘差。**如果已選取了此選項，那麼對於每個目標欄位，此選項將為模型殘差建立新欄位（帶有預設字首 $TSResidual-），並同時建立這些值的總計。

模型設定

要在輸出中顯示的上限模式個數。指定您要包含在輸出中的上限模式個數。請注意，如果建立的模式個數超過了此臨界值，那麼模型不會顯示在輸出中，但它們仍可用於評分。預設值是 10。顯示大量模型可能會導致效能不佳或不穩定。

時間序列模型片段

「時間序列」模型塊輸出

建立「時間序列」模型後，輸出檢視器中會提供下列資訊。請注意，「時間序列」模型的「輸出」檢視器中可以顯示的模型數量限制為 10 個。

時間資訊摘要

此摘要顯示下列資訊：

- 時間欄位
- 增量
- 起始點和結束點
- 唯一點的號碼

此摘要適用於所有目標。

模型資訊表格

（對於每個目標重複）模型資訊表格提供關於模型的關鍵資訊。該表格一律包括下列高階模型設定：

- 在「類型」節點或「時間序列」節點欄位標籤中選取的目標欄位的名稱。
- 模型建置方法 - 例如，指數平滑化或 ARIMA。
- 輸入到模型中的預測值數。
- 用於擬合模型類型的記錄數。不同類型的模型的範例可能包含：RMSE、MAE、AIC、BIC 和 R 平方。

另外，如果資料符合所需的條件，那麼還可能顯示 Ljunq-Box Q 統計資料。在下列情況下，此統計資料**不**可用：

- 非遺漏資料點的號碼小於或等於所需的總和項目的號碼（固定值 18）。
- 參數數目大於或等於所需的總和項目的號碼。
- 所計算的總和項目的號碼少於可接受的最小 k 值（固定值 7）。
- 對於每個目標，表格重複顯示。

預測值重要性

（對於每個目標重複）預測值重要性圖表以長條圖的形式顯示模型中前 10 個輸入（預測值）的重要性。

如果圖表中具有 10 個以上欄位，則您可以使用圖表下的調節器來變更圖表中包含的預測值選擇。調節器上的指示標示是固定寬度，並且調節器上的每個標示都代表 10 個欄位。您可以沿著調節器來移動指示標示以顯示後面或前面的 10 個欄位，依預測值重要性排序。

您可以按兩下圖表以開啟個別對話框來編輯圖形設定。例如，您可以修正一些項目，例如圖形大小，以及所用字型的大小和顏色。關閉這個個別的編輯對話框時，變更會套用至「輸出」標籤中顯示的圖表。

相關圖

將對每個目標顯示相關圖（即，自相關圖），並且該圖形顯示了殘值（期望與實際值之間的差分）與時間延遲的自相關係數函數 (ACF) 或局部自相關係數函數 (PACF)。信賴區間在整個圖表中強調顯示。

參數估計值

對於每個目標，參數估計值重複顯示（適用時），其中包含下列詳細資料：

- 目標名稱
- 所套用的轉換
- 對模型 (ARIMA) 中此參數使用的延遲
- 係數值
- 參數估計值的標準誤差
- 參數估計值除以標準誤差後的值
- 參數估計的顯著性層次。

「時間序列」模型塊設定

「設定」標籤為「時間序列」模型塊提供其他選項。

預測

用於**將記錄延伸到未來**。的選項設定在估計期間結束之後要預測的時間間隔數量。在這種情況下，時間間隔是在「時間序列」節點的「資料規格」標籤上指定的分析時間間隔。要求預測時，會針對任何不是目標的輸入序列自動建置自動迴歸模型。然後，在預測期間使用這些模型產生那些輸入數列的值。

計算輸入的未來值。如果選取此選項，那麼會計算預測值、雜訊預測、差異估計和未來時間值的預測值。

要在用於預測的未來值

- **計算輸入的未來值** 如果選取此選項，那麼會自動計算預測值、雜訊預測、差異估計和未來時間值的預測值。要求預測時，會針對任何不是目標的輸入序列自動建置自動迴歸模型。然後，在預測期間使用這些模型產生那些輸入數列的值。
- **選取要將其值新增到資料中的欄位**。對於要預測的每條記錄（holdout 除外），如果您使用預測值欄位（角色設定為輸入），那麼可以為每個預測值的預測期間指定估計值。您可以手動指定值，也可以從清單中選擇。
  - **欄位**。 按一下欄位選取器按鈕，並選擇可以用作預測值的任何欄位。請注意，在這裡選取的欄位在建模時可以使用，也可能無法使用；若要實際將欄位用作預測值，則必須在下游建模節點中選取它。此對話框會為您直接提供方便的地方來指定未來值，因此可由多個下游建模節點共用而無需在每一個節點中單獨指定它們。另請注意，可用欄位；的清單可能會受到「建置選項」標籤中選項的約束。

    請注意，如果為串流中不再可用（因為已將其刪除或由於「建置選項」標籤中的選項更新）的欄位指定了未來值，那麼此欄位將顯示為紅色。
  - **值**。 針對每一個欄位，您可以從函數清單中選擇，或按一下**指定**以手動輸入值或從預定值清單中選擇。如果預測值欄位與您可以控制或可以提前知道的項目相關，則應手動輸入值。例如，如果您要根據房間預約數預測旅館下個月的收入，則可以指定該期間實際擁有的預約數。相反地，如果預測值欄位與您無法控制的某個項目相關，例如股價，則可以使用最近的值或最近點的平均數等函數。

    可用的函數視欄位的測量層級而定。

表 *29.* 可用於測量層次的函數

測量層級	函數
連續或名義欄位	空白 最近點的平均數 最新值 指定
旗標欄位	空白 最新值 TrueFalse指定

最近點的均數根據最後一欄三個資料點的均數計算未來值。

最近值將未來值設定為最近資料點的值。

true/假將旗標欄位的未來值設定為指定的 true 值或 false。

指定開啟一個對話框，用於手動指定未來值或從預先定義清單中選擇未來值。

可用於評分

為要評分的每一個模型建立新欄位。 可讓您指定要為將進行評分之每個模型所建立的新欄位。

- **噪音殘差。** 如果已選取了此選項，那麼對於每個目標欄位，將為模型殘差建立新欄位（帶有預設字首 $TSResidual-），並同時建立這些值的總計。
- **信賴區間上限及下限。** 如果已選取了此選項，那麼對於每個目標欄位，此選項將分別為信賴區間上限和下限建立新欄位（帶有預設字首 $TSLCI- 和 $TSUCI-），並同時建立這些值的總計。

包括的用來評分的目標。 選取可用的目標以包括在模型評分中。

第 14 章 自習回應節點模型

SLRM 節點

使用**自習回應模型** (SLRM) 節點，可以建立這樣的模型：隨著資料集的增長，可以不斷對其進行更新或重新估計，而不必每一個使用整個資料集重新建立該模型。例如，如果有多個產品，而您希望確定某位客戶獲得報價後最有可能購買的產品，那麼這種模型將十分有用。此模型可用於預測最適合客戶的報價，以及該報價被接受的機率。

最初建立模型時，可以使用較小的資料集，其中的報價和對這些報價的回應可以隨機選擇。隨著資料集的增長，模型可得到更新，從而越發能夠根據其他輸入欄位（如年齡、性別、職業和收入）預測最適合客戶的報價以及這些客戶接受報價的機率。可以通過在節點對話框中新增或刪除這些可用報價對其進行變更，而不必變更資料集的目標欄位。

如果與 IBM SPSS Collaboration and Deployment Services 一起使用，那麼可以為模型設立自動定期更新。該過程不需要人工監督或動作就可以為不可能或沒必要由資料挖掘者自訂人為介入的組織和應用程式提供靈活且成本低的解決方案。

範例。 某金融機構希望通過向每個客戶提供最有可能接受的報價來獲取更多的利潤。您可以使用自習模型來根據先前的促銷活動確定最有可能對活動作出積極回應的客戶的特性，並根據最近的客戶回應即時更新該模型。

SLRM 節點欄位選項

執行 SLRM 節點之前，必須在節點的「欄位」標籤上同時指定目標欄位和目標回應欄位。

目標欄位。 從清單中選定目標欄位；例如，包含要為客戶提供的不同產品的列名（集合）欄位。

註：目標欄位的儲存格式必須為字串而不是數值。

目標回應欄位。從清單中選取目標回應欄位。例如，接受或拒絕。

註：此欄位必須是旗標欄位。旗標的 true 值表示報價接受，false 表示報價拒絕。

此對話框中的剩餘欄位是在整個 IBM SPSS Modeler 中使用的標準欄位。請參閱第 26 頁的『建模節點欄位選項』主題，以取得更多資訊。

註：如果來源資料包含要用作連續（數值型範圍）輸入欄位的範圍，那麼您必須確保 meta 資料包含每個範圍的最小值和最大值詳細資料。

SLRM 節點模型選項

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。

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。

繼續訓練現有模型。 依預設，每次執行建模節點時，將建立一個全新的模型。如果已選取該選項，那麼會繼續訓練該節點成功生成的最後一個模型。這樣就可以在無需存取原始資料的情況下更新或重新整理現有的模型，並可能會顯著提升效能，這是因為只有新的或更新後的記錄被反饋到串流中。系統會使用建模節點來儲存上一個模型的詳細資料，即使資料流或「模型」色板中已不再提供上個模型塊，也能使用此選項。

© Copyright IBM Corp. 1994, 2019

目標欄位值。預設情況下，此選項設定為**使用全部**，表示將建立其中包含與選定目標欄位值相關聯的每個報價的模型。如果希望產生僅包含目標欄位的某些報價的模型，請按一下**指定**，並使用**新增**、**編輯**和**刪除**按鈕輸入或修正要為其建立模型的報價的名稱。例如，如果選擇的目標是列出提供的所有產品，那麼可以使用此欄位將提供報價的產品限制為在此輸入的產品。

模型評量。此畫面中的欄位與模型無關，因為這些欄位不會影響評分。不過，這些字段有助於形成一個直觀代表，顯示模型預測結果的準確程度。

註：要在模型塊中顯示模型評估結果，您還必須選中**顯示模型評估**複選框。

- **包含模型評量**。已選取此複選框可以建立針對每項選定報價顯示模型的預測精確度的圖形。
- **設定隨機種子**。 當根據隨機百分比來估計模型的精確度時，此選項可讓您將相同的結果複製到其他階段作業中。透過指定亂數產生器所用的起始值，您可以確保每次執行節點時都指派相同的記錄。輸入期望的種子值。如果未選取此選項，則每次執行節點時，都將產生不同樣本。
- **模擬樣本大小**。指定評估模型時要在樣本中使用的記錄數。預設值是 100。
- **疊代次數**。 通過此選項，您可以在疊代次數達到指定值後停止建構模型評量。指定疊代數目上限；預設值是 20。

  註：請記住，如果樣本大小較大並且疊代數較多，那麼這將增加建立模型所用的時間。

顯示模型評估。選中此選項將在模型塊中顯示結果的圖形表示法。

SLRM 節點設定選項

使用節點設定選項可微調模型建立過程。

每秒的預測數上限。透過此選項，您可以限制對資料集中每條記錄進行的預測數。預設是 3。

例如，您有六項報價（如儲蓄、抵押、汽車貸款、退休金、信貸卡片和保險），但只想瞭解最適於推薦的兩項；這時應將此欄位設為 2。當您建立模型並將其附加到表格中時，會看到每條記錄有兩個預測欄（以及接受的報價的相關信賴度機率）。預測可以由六種可能報價中的任意報價組成。

隨機化層次。為了避免出現任何偏移（例如，在較小或不完整的資料集中）並平等對待所有可能的報價，您可以對報價的選擇及其成為推薦報價的機率新增隨機化層次。隨機化表示為百分比，以 0.0（無隨機化）與 1.0（完全隨機化）之間小數值的形式顯示。預設值為 0.0。

設定隨機種子。 向報價的選擇新增隨機化層次時，您可以透過此選項在另一個階段作業中複製相同結果。透過指定亂數產生器所用的起始值，您可以確保每次執行節點時都指派相同的記錄。輸入期望的種子值。如果未選取此選項，則每次執行節點時，都將產生不同樣本。

註：對從資料庫中讀取的記錄使用**設定隨機種子**選項時，可能需要在取樣前使用「排序」節點以確保每次執行節點時都獲得相同的結果。 這是因為隨機種子取決於記錄的順序，而並不能保證順序在關聯式資料庫中保持相同。

排序順序。選取報價在已建置模型中的顯示順序：

- **遞減**。模型首先顯示分數最高的報價。這些報價被接受的機率最高。
- **遞增**。模型首先顯示分數最低的報價。這些報價被拒絕的機率最高。例如，在決定要從某種特定報價的營銷活動中刪除哪些客戶時，這種順序相當實用。

目標欄位的喜好設定。建置模型時，您可能希望主動升級或移除資料的特定方面。例如，如果建立用於選取為某個客戶推薦的最佳財務報價的模型，您可能需要確保始終包含一種特定報價（無論其對於每個客戶的分數如何）。

要在此畫面中包含某項報價並編輯其喜好設定，請按一下**新增**，鍵入報價的名稱（例如，「儲蓄」或「抵押」），然後按一下**確定**。

- **值。** 此選項將顯示您新增的報價的名稱。
- **喜好設定。**指定要對報價套用的喜好設定層次。首選度表示為百分比，以 0.0（非首選）與 1.0（最首選）之間小數值的形式顯示。預設值為 0.0。
- **始終包含。**要確保某項特定報價始終包含在預測中，請選中此框。

註：如果**首選率**設定為 0.0，那麼將忽略**始終包含**設定。

考慮模型可靠性。與包含少量資料的全新模型相比，已透過多次重新產生來進行微調的結構良好、資料豐富的模型應當始終產生更準確的結果。要利用較成熟模型具有的較高可靠性，請選中此框。

SLRM 模型塊

註：如果在「模型選項」標籤上同時選中**包含模型評估**和**顯示模型評估**，那麼結果只會顯示在此標籤。

在執行包含 SLRM 模型的串流時，該節點會估計每個目標欄位值（報價）的預測精確度，以及所用的每個預測值的重要性。

註：如果在建模節點的「模型」標籤上已選取了**繼續訓練現有模型**，那麼將在每次重新產生模型時更新模型塊上顯示的資訊。

對於使用 IBM SPSS Modeler 12.0 或更高版本建立的模型，模型塊的「模型」標籤分為兩欄：

左欄。

- **檢視。** 有多項報價時，請選取要顯示其結果的一項報價。
- **模型效能。**此部分顯示每項報價的估計模型精確度。測試集通過模擬產生。

右欄。

- **檢視。** 選取要顯示**與回應的關聯**還是**變數重要性**詳細資料。
- **與回應的關聯。**顯示每個預測值與目標變數之間的關聯（相關性）。
- **預測值重要性。**表示在估計模型過程中每個預測值的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。雖然在使用 SLRM 的情況下，圖表是由 SLRM 演算法模擬產生的，但該圖表可用解釋其他顯示預測值重要性的模型的方式進行解釋。方法是：依次從模型中刪除每個預測值，然後請參閱此操作對模型精確度的影響如何。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

SLRM 模型設定

在 SLRM 模型塊的「設定」標籤中可指定選項以修改已建立的模型。例如，可以通過 SLRM 節點使用相同的資料和設定建立幾個不同的模型，然後使用每個模型中的此標籤對設定稍做修改以查看其對結果的影響。

註：僅當模型片段已新增至串流之後，此標籤才可用。

每秒的預測數上限。透過此選項，您可以限制對資料集中每條記錄進行的預測數。預設是 3。

例如，您有六項報價（如儲蓄、抵押、汽車貸款、退休金、信貸卡片和保險），但只想瞭解最適於推薦的兩項；這時應將此欄位設為 2。當您建立模型並將其附加到表格中時，會看到每條記錄有兩個預測欄（以及接受的報價的相關信賴度機率）。預測可以由六種可能報價中的任意報價組成。

隨機化層次。為了避免出現任何偏移（例如，在較小或不完整的資料集中）並平等對待所有可能的報價，您可以對報價的選擇及其成為推薦報價的機率新增隨機化層次。隨機化表示為百分比，以 0.0（無隨機化）與 1.0（完全隨機化）之間小數值的形式顯示。預設值為 0.0。

設定隨機種子。 向報價的選擇新增隨機化層次時，您可以透過此選項在另一個階段作業中複製相同結果。透過指定亂數產生器所用的起始值，您可以確保每次執行節點時都指派相同的記錄。輸入期望的種子值。如果未選取此選項，則每次執行節點時，都將產生不同樣本。

註：對從資料庫中讀取的記錄使用**設定隨機種子**選項時，可能需要在取樣前使用「排序」節點以確保每次執行節點時都獲得相同的結果。 這是因為隨機種子取決於記錄的順序，而並不能保證順序在關聯式資料庫中保持相同。

排序順序。選取報價在已建置模型中的顯示順序：

- **遞減。**模型首先顯示分數最高的報價。這些報價被接受的機率最高。
- **遞增。**模型首先顯示分數最低的報價。這些報價被拒絕的機率最高。例如，在決定要從某種特定報價的營銷活動中刪除哪些客戶時，這種順序相當實用。

目標欄位的喜好設定。建置模型時，您可能希望主動升級或移除資料的特定方面。例如，如果建立用於選取為某個客戶推薦的最佳財務報價的模型，您可能需要確保始終包含一種特定報價（無論其對於每個客戶的分數如何）。

要在此畫面中包含某項報價並編輯其喜好設定，請按一下**新增**，鍵入報價的名稱（例如，「儲蓄」或「抵押」），然後按一下**確定**。

- **值。** 此選項將顯示您新增的報價的名稱。
- **喜好設定。**指定要對報價套用的喜好設定層次。首選度表示為百分比，以 0.0（非首選）與 1.0（最首選）之間小數值的形式顯示。預設值為 0.0。
- **始終包含。**要確保某項特定報價始終包含在預測中，請選中此框。

  註：如果**首選率**設定為 0.0，那麼將忽略**始終包含**設定。

考慮模型可靠性。與包含少量資料的全新模型相比，已透過多次重新產生來進行微調的結構良好、資料豐富的模型應當始終產生更準確的結果。要利用較成熟模型具有的較高可靠性，請選中此框。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

第 15 章 支援向量機器模型

關於 SVM

支援向量機器 (SVM) 是一項功能穩健的分類和迴歸方法技術，可最大化模型的預測準確度，而不會過度配適訓練資料。SVM 特別適用於分析預測值欄位非常多（如數千個）的資料。

SVM 適用於多個學科，例如客戶關係管理 (CRM)、面部影像和其他影像識別、生物資訊學、文字採集概念擷取、入侵偵測、蛋白質結構預測以及語音識別。

SVM 如何運行

SVM 的工作原理是將資料對映到高維度特徵空間，這樣即使資料不是線性可分，也可以對該資料點進行分類。找到種類之間的分隔字元，然後以將分隔字元繪製成超平面的方式轉換資料。之後，可用新資料的特性預測新記錄所屬的群組。

例如，請考慮下圖，圖中的資料點落在兩個不同的種類中。

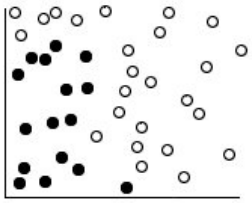


圖 *59.* 原始資料集

可以使用一條曲線分隔這兩個種類，如下圖所示。

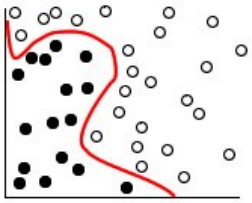


圖 *60.* 新增了分隔字元的資料

變換後，可以使用超平面定義這兩個種類之間的界限，如下圖所示。

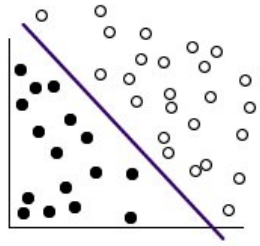


圖 *61.* 轉換後的資料

用於轉換的數學函數稱為**核**函數。IBM SPSS Modeler 中的 SVM 支援下列核類型：

- 線性
- 多項式
- 徑向基底函數 (RBF)
- Sigmoid

如果資料的線性分隔比較簡單，那麼建議使用線性核心功能。在其他情況下，應當使用其他函數。在所有情況下，您都需要嘗試使用不同的函數才能獲得最佳模型，因為每一個函數均使用不同的演算法和參數。

調整 SVM 模型

除了種類之間的分隔行，分類 SVM 模型還會尋找用於定義兩個種類之間的空間的邊際行。

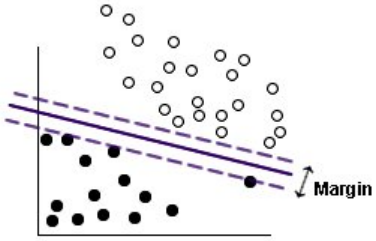


圖 *62.* 使用初步模型的資料

位於邊距上的資料點稱為**支援向量**。

兩個種類之間的邊距越寬，模型在預測新記錄所屬個種類方面性能越佳。在上一個範例中，邊距不是很寬，因此稱該模型**過度擬合**。為了增加邊界的寬度，可以接受少數量的誤分類；下圖中顯示了一個這樣的範例。

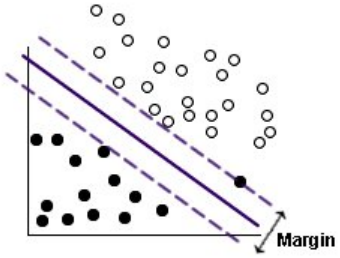


圖 *63.* 使用改進模型的資料

在某些情況下，線性分隔難度較大；下圖中顯示了一個這樣的實例。

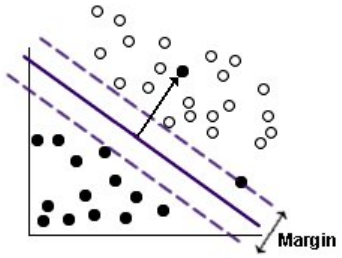


圖 *64.* 線性分隔存在的問題

在類別似這種情況中，目標是找到寬邊距和少量誤分類別資料點之間的最佳平衡。核心功能有一個**正規化參數**（稱為 C），該參數控制項這兩個值之間的平衡。如果要獲得最佳模型，您可能需要對該參數和其他核參數嘗試使用不同的值。

SVM 節點

通過 SVM 節點，可以使用支援向量機器對資料進行分類。SVM 特別適合於大型資料集，即具有大量預測值欄位的資料集。可以對節點使用預設值以相對較快地生成基本模型，也可以使用「專家」設定以嘗試使用不同類型的 SVM 模型。

如果已建置模型，您可以：

- 瀏覽模型區塊以顯示建置模型中輸入欄位的相對重要性。
- 將「表格」節點附加至模型區塊以檢視模型輸出。

範例。 醫學研究員已取得一個資料集，其中包含擷取自被認為有患癌風險之病人的數個人類細胞樣本的性質。分析原始資料表明良性與惡性樣本之間的許多性質存在顯著差異。該研究人員希望開發一個 SVM 模型，該模型可以使用其他病患的樣本中相似細胞特性的值，以盡早發現他們的樣本是良性還是惡性。

SVM 節點模型選項

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

SVM 節點專家選項

如果您對支援向量機器具有深入瞭解，那麼可以使用專家選項對訓練過程進行調整。若要存取專家選項，請在「專家」標籤上將「模式」設為**專家**。

附加所有可能性（僅對種類目標有效）。如果已選取了該選項，那麼將指定針對節點所處理的每條記錄顯示列名或旗標目標欄位的每個可能值的機率。如果未選取該選項，那麼僅為列名或旗標目標欄位顯示預測值的機率。該勾選框的設定將決定模型塊上的相應勾選框的預設狀態。

停止準則。決定停止最佳化演算法的時間。值的範圍從 1.0E–1 到 1.0E–6；預設值為 1.0E–3。減小該值會生成更準確的模型，但模型的訓練時間也要相應增加。

正規化參數(C)。控制項最大化邊距和最小化訓練錯誤項目之間的平衡。通常情況下，值應當介於 1 和 10（含本數）之間；預設值為 10。增加該值會改善訓練資料的分類準確性（或減少迴歸錯誤），但這樣也可能會導致過適。

迴歸方法精準度 (epsilon)。唯有當目標欄位的測量層次為 *Continuous* 時才使用。導致接受錯誤的原因是錯誤小於在這裡指定的值。增加值可能會加速建模，但以精確度為代價。

核心類型。確定用於轉換的核心功能的類型。核類型不同，計算分隔字元的方法也將不同，因此建議嘗試使用不同的選項。預設值為 **RBF**（徑向基底函數）。

RBF Gamma。僅在核類型設定為 **RBF** 時才啟用。通常情況下，值應當介於 $3/k$ 和 $6/k$ 之間，其中 k 為輸入欄位的數量。例如，如果有 12 個輸入欄位，那麼應當嘗試使用介於 0.25 和 0.5 之間的值。增加該值會改善訓練資料的分類準確性（或減少迴歸錯誤），但這樣也可能會導致過適。

Gamma。僅在核心類型設為**多項式**或 **Sigmoid** 的情況下啟用。增加該值會改善訓練資料的分類準確性（或減少迴歸錯誤），但這樣也可能會導致過適。

偏移。僅在核心類型設為**多項式**或 **Sigmoid** 的情況下啟用。在核心功能中設定 `coef0` 值。大多數情況下可以使用預設值 0。

度。僅在核類型設定為**多項式**時才啟用。控制項對映空間的複雜性（維度）。通常情況下，不使用大於 10 的值。

SVM 模型塊

SVM 模型會建立許多新欄位。其中最重要的是 **$S-fieldname** 欄位，該欄位顯示由模型預測的目標欄位值。

模型建立的新欄位的數量和名稱取決於目標欄位的測量層次（此欄位在下表格中由欄位名指示）。

如需這些欄位及其值，請將表格節點新增到 SVM 模型塊中，然後執行表格節點。

表 *30.* 目標欄位測量層級為「名義」或「旗標」

新欄位名稱	說明
$S-*fieldname*	目標欄位的預測值。
$SP-*fieldname*	預測值的機率。
$SP-*value*	列名或旗標的各個可能值的機率（僅在已勾選模型塊中「設定」標籤上的**附加所有可能性**時才顯示）。
$SRP-*value*	（僅適用於旗標目標）原始 (SRP) 和已調整的 (SAP) 傾向分數，表示目標欄位結果為 "true" 的可能性。僅當在產生模型之前已選取 SVM 建模節點的「分析」標籤上的相應勾選框之後，才顯示這些分數。請參閱第 29 頁的『建模節點分析選項』主題，以取得更多資訊。
$SAP-*value*	

表 *31.* 目標欄位測量層級為「連續」

新欄位名稱	說明
$S-*fieldname*	目標欄位的預測值。

預測值重要性

選擇性地，指出評估模型時每個預測值的相對重要性的圖表，可能也會顯示在「模型」標籤上。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。注意，只有在產生模型之前已選取「分析」標籤上的**計算預測值重要性**，才可以使用此圖表。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

註：與其他類型的模型相比，計算 SVM 的預測值重要性可能需要更長時間，因此依預設在「分析」標籤中未選取預測值重要性。選取此選項可能會降低效能，尤其是使用大型資料集時。

SVM 模型設定

通過「設定」標籤可以指定在檢視結果時顯示的附加欄位（例如，通過執行表格節點附加到塊）。通過選取這些選項可以檢視每個選項的效果,並且按一下「預覽」按鈕（捲動至「預覽」輸出右側）可以檢視附加欄位。

附加所有可能性（僅對種類目標有效）。如果勾選了此選項，則會針對該節點所處理的每一筆記錄，顯示標稱或旗標目標欄位的每個可能值的機率。如果未勾選此選項，則針對標稱或旗標目標欄位只會顯示預測值及其機率。

此勾選框的預設設定由建模節點的相應勾選框確定。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項來指定如何執行 SQL 產生。

- **預設值：使用伺服器評分配接器（如果已安裝）進行評分，否則在處理程序中評分** 如果連接至已安裝評分配接器的資料庫，則使用評分配接器及相關使用者定義函數 (UDF) 來產生 SQL，並在資料庫內對模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫外部評分** 如果選取此項，則此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。

LSVM 節點

通過 LSVM 節點，您可以使用線性支援向量機器對資料進行分類。LSVM 特別適用於大型資料集，即具有大量預測值欄位的資料集。可以對節點使用預設值以便相對較快地生成基本模型，也可以使用建置選項來試用不同的設定。

LSVM 節點類似於 SVM 節點，但它是線性的，更擅長處理大量記錄。

如果已建置模型，您可以：

- 瀏覽模型區塊以顯示建置模型中輸入欄位的相對重要性。
- 將「表格」節點附加至模型區塊以檢視模型輸出。

範例。 醫學研究員已取得一個資料集，其中包含擷取自被認為有患癌風險之病人的數個人類細胞樣本的性質。分析原始資料表明良性與惡性樣本之間的許多性質存在顯著差異。該研究人員希望開發一種 LSVM 模型，該模型可以使用其他病患的樣本中相似細胞特性的值，以盡早發現他們的樣本是良性還是惡性。

LSVM 節點模型選項

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。

計算預測值重要性。 對於可產生重要性適當測量的模型，您可以顯示一個圖表來指出每個預測值對於評估模型的相對重要性。一般而言，您會想要將建模焦點著重在最重要的預測值，並考慮捨棄或忽略最不重要的預測值。請注意，對於部分模型，需要較長時間來計算預測值重要性，尤其是處理大型資料集時，結果便是依預設會關閉部分模型的預測值重要性。預測值重要性對於決策清單模型無法使用。如需相關資訊，請參閱第 37 頁的『預測值重要性』。

LSVM 建置選項

模型設定

包括截距。包括截取（模型中的固定項目）可增加解的整體精確度。如果可以假設資料穿過原點，就可以將截距排除在外。

種類目標的排序。為分類目標指定排序。對於連續目標，將忽略此設定。

迴歸方法精準度 (epsilon)。唯有當目標欄位的測量層次為 *Continuous* 時才使用。導致接受錯誤的原因是錯誤小於在這裡指定的值。增加值可能會加速建模，但以精確度為代價。

排除具有任何遺漏值的記錄。設定為 **True** 時，如果任何單一值遺漏，那麼將排除記錄。

懲罰值設定

處罰函數。指定用於降低過度適合可能性的處罰函數類型。選項為 **L1** 或 **L2**。

L1 和 **L2** 通過增加係數罰分來降低過度配適的幾率。二者的區別是當有大量特徵時，在模型建置期間，**L1** 通過將某些係數設定為 0 來使用功能選擇。**L2** 不具備此能力，因此在有大量特徵時不應使用該選項。

懲罰值參數 (lambda)。指定懲罰值（正規化）參數。如果設定了**處罰函數**，那麼將啟用此設定。

LSVM 模型塊（互動式輸出）

執行 LSVM 模型後，下列輸出可用。

模型資訊

「模型資訊」視圖提供有關模型的關鍵資訊。該表格識別一些高階模型設定，例如：

- 在「欄位」標籤中指定的目標的名稱
- 模型選擇設定上指定的模型建置方法
- 預測值輸入數目
- 最終模型中預測值的數目
- 正規化類型（L1 或 L2）
- 懲罰值參數 (lambda)。這是正規化參數。
- 迴歸方法精準度 (epsilon)。如果錯誤少於此值，那麼將接受這些錯誤。更高的值可能會加快建模速度，但要以精確度為代價。僅當目標欄位的測量層次為連續時，此項才可用。
- 分類精確度百分比。這僅適用於分類。
- 平均值平方誤。這僅適用於迴歸方法。

記錄摘要

「記錄摘要」視圖提供模型中所包括以及排除的記錄（觀察值）的數目與百分比。

預測值重要性

一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

已依觀察預測

這會根據水平軸上的觀察值，來顯示垂直軸上預測值的 Bin 散佈圖。理想的狀況下，點應排列在 45 度的線上；此檢視可以告訴您模式是否有預測結果特別差的記錄。

註：與其他類型的模型相比，計算 LSVM 和 SVM 的預測值重要性可能需要更長時間。選取此選項可能會降低效能，尤其是使用大型資料集時。

混淆矩陣

混淆矩陣（有時也稱為摘要表）顯示了根據 LSVM 分析正確和不正確分配給每個群組的觀察值數。

LSVM 模型設定

在 SVLM 模型塊的「設定」標籤上，您可以指定模型評分期間用於原始傾向的選項和用於 SQL 產生的選項。僅當模型片段已新增至串流之後，此標籤才可用。

計算原始傾向評分 對於只含有旗標目標的模型，您可以要求原始傾向評分來指出為目標欄位指定之真實結果的可能性。這些是標準預測及信賴度值的附加項目。無法使用調整傾向評分。

產生此模式的 SQL：使用資料庫中的資料時，可以將 SQL 代碼推回到資料庫中以進行執行，這可以極大地提高多數作業的效能。

選取下列其中一個選項以指定 SQL 的產生方式。

- **預設值：在程序中使用「伺服器評分配接卡」（如有安裝的話）來評分。** 如果連接至安裝了評分配接卡的資料庫，則使用評分配接卡和關聯使用者定義的函數 (UDF) 來產生 SQL，並對資料庫中的模型進行評分。沒有可用的評分配接器時，此選項會從資料庫提取資料並在 SPSS Modeler 中對資料進行評分。
- **在資料庫之外評分。** 選取之後，此選項會從資料庫提取回您的資料，並在 SPSS Modeler 中對其進行評分。

第 16 章 最近相鄰元素模型

KNN 節點

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* 值對新觀察值進行分類。當 *k* = 5 時，新觀察值放置在種類 *1* 中，因為大部分最近的鄰接項都的種類 *1*。但是，當 *k* = 9 時，新觀察值會放置在種類 *0* 中，因為大部分最近的鄰接項都的種類 *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。

KNN 節點目標選項

您可以在「目標」標籤中，選擇在輸入資料中根據最近鄰接項的值建置預測目標欄位值的模型，或者只是尋找特定感興趣觀察值的最近鄰接項。

您要執行那種類型的分析？

預測目標欄位。如果您想根據最近鄰接項的值預測目標欄位的值，請選擇此選項。

僅識別最近鄰接項。如果您只希望看到特定輸入欄位的最近鄰接項，請選擇此選項。

如果您選擇只識別最近鄰接項，在此標籤上與精確度和速度相關的剩餘選項將被停用，因為其只與預測目標相關。

您的目標是什麼?

預測目標欄位時，您可以通過這組選項來決定速度、精確度或這二者的組合是否為最重要的因素。或者您可以選擇自己自訂設定。

如果您選擇平衡、速度或精確度選項，那麼演算法預先選擇該選項的最合適設定組合。進階使用者可能希望置換這些選擇；可在「設定」標籤上的各個畫面上進行此操作。

權衡速度與準確度。選取小範圍內鄰接項的最佳數量。

速度。找出固定數量的相鄰值。

精確度。選取較大範圍內的鄰接項的最佳數量，並在計算距離時使用預測值重要性。

自訂分析。 選擇該選項以微調「設定」標籤上的演算法。

註：與大多數其他模型不同的是，生成的 KNN 模型的大小隨著訓練資料量的增大呈線性增加。如果在嘗試建置 KNN 模型時看到報告「記憶體不足」錯誤，那麼嘗試增加 IBM SPSS Modeler 所使用的系統記憶體上限。要進行此操作，請選擇

工具 > **選項** > **系統選項**

並在**最大記憶體**欄位中輸入新大小。「系統選項」對話框中所作的變更要在重新啟動 IBM SPSS Modeler 之後才能生效。

KNN 節點設定

在「設定」標籤上您可以指定最近相鄰元素分析特有的選項。畫面左端的側邊列出了用於指定選項的畫面。

模型

「模型」窗格提供控制如何構建模型的選項，例如是否使用分區或分割模型、是否轉換數值型輸入欄位以使其落入相同範圍內和如何管理感興趣觀察值。您也可以給模型選擇一個自訂名稱。

註：使用分割的資料和**使用觀察值標籤**不能使用同一欄位。

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。

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。

建立分割模型。針對指定為分割欄位的輸入欄位的每個可能的值，建置個別的模型。如需相關資訊，請參閱第 24 頁的『建立分割模型』。

要手動選取欄位... 依預設，節點使用來自「類型」節點的分割區與分割欄位設定（如果存在），但您可以在此處置換這些設定。要啟動**分割區**與**分割**欄位，請選擇**欄位**標籤，並選擇**使用自訂設定**，然後傳回此處。

- **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。透過使用一個樣本來產生模型，並使用另一個樣本來測試模型，您可以很好地指出模型將概化為與現行資料相似的更大型資料集的程度。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。）
- **分割。** 針對分割模型，選取一或多個分割欄位。這與在「類型」節點中將欄位角色設為分割類似。您可以僅將類型為**旗標**、**列名**或**序數**的欄位指定為分割欄位。選擇作為分割欄位的欄位無法用作目標、輸入、分割區、頻率或加權欄位。請參閱第 24 頁的『建立分割模型』主題，以取得更多資訊。

使範圍輸入正規化。勾選此方框為連續輸入欄位正規化值。經過常態化的功能具有相同的值範圍，可改善估計演算法的效能。使用調整後常態化 [2*(x–min)/(max–min)]–1,調整後常態化的值介於 -1 和 1 之間。

使用觀察值標籤。勾選此方框以啟用下拉清單，從這裡您可以選擇欄位並將其值用作標籤，以在「模型檢視器」中識別在預測值空間圖表、對等項圖表和象限圖中所需的觀察值。您可以選擇測量層次為列名、序數或旗標的任何欄位用作標籤欄位。如果您不在這裡選擇欄位，則記錄會顯示在「模型檢視器」圖表中，且最近鄰接項依來源資料中的列號識別。如果您在建立模型之後要操作資料，可使用觀察值標籤，以避免每次需要參考來源資料在顯示中識別觀察值。

ID 焦點記錄。勾選此方框啟用下拉清單，容許您標示感興趣的輸入欄位（僅針對旗標欄位）。如果在此處指定了一個欄位，當建立模型時，首先會在模型檢視器中選取代表該欄位的點。在此處選取焦點記錄是選用的；任何點都可以暫時成為焦點記錄，只要在「模型檢視器」中手動已選取它。

鄰接項

「鄰接項」畫面具有一組控制如何計算最近鄰接項數量的選項。

最近鄰數目 (k)。 指定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。

如果目的是為了預測目標，則您具有兩個選擇：

- **指定固定 k。**如果您要指定尋找固定數目的最近鄰接項，請使用此選項。
- **自動選擇 k。**您也可以使用**最小值**和**最大值**欄位指定一個值範圍，並容許程序選擇該範圍內鄰接項的「最佳」數量。確定最近鄰接項數目的方法取決於「特性選取」畫面上是否要求特性選取。

  如果功能選擇有效，那麼針對已要求範圍中每個 *k* 值執行功能選擇，並選擇具有最低錯誤率（如果目標為連續，那麼為最低平方和誤差）的 *k* 值和特徵集。

  如果功能選擇無效，則會使用 *V* 折疊交叉驗證，以選擇相鄰的「最佳」數目。請參閱「交叉驗證」畫面以瞭解如何控制褶疊的指定。

距離計算。 這是指定距離矩陣所用的矩陣，可用來測量觀察值的親緣性。

- **歐基里得矩陣。** 兩個觀察值 x 和 y 之間的距離就是觀察值之間平方差的所有維度總和平方根。
- **城市街區度量值。**兩個觀察值之間的距離就是觀察值數值之間差異的所有維度總和。也稱為 Manhattan 距離。

或者，如果目的是為了預測目標，則您可以選擇在計算距離時按照其正規化重要性計算特徵加權。預測值的特徵重要性的計算方法為：不含預測值的模型的錯誤率或平方和誤差與完整模型的誤率或平方和誤差之比。經過常態化的重要性，是以重新加權功能重要性值進行計算，因此總和為 1。

計算距離時按照重要性計算特徵加權。（唯有當目的為預測目標時才顯示。）勾選此方框，當計算鄰接項之間距離時，使用預測值重要性。預測值重要性將在模型塊中顯示，並用於預測（因此影響評分）。請參閱第 37 頁的『預測值重要性』主題，以取得更多資訊。

對範圍目標的預測。（唯有當目的為預測目標時才顯示。）如果指定了連續（數值型範圍）目標，這可指定預測值是基於最近鄰接項的平均數還是中值來計算的。

功能選擇

唯有當目的為預測目標時才會啟動此畫面。使您能夠為功能選擇要求和指定選項。依預設，會針對功能選擇考量所有功能，但是您可以選擇強制用於模式中的功能子集。

執行功能選擇。選中此勾選框啟用功能選擇選項。

- **強制輸入。**按一下此框旁的欄位選擇器按鈕並選擇一個或多個特徵以強制輸入模型。

停止準則。 在每一步中，如果在模型中新增特性可以將錯誤減至最小（計算為種類目標的錯誤率和連續目標的平方和錯誤），那麼考慮將其納入模型集中。向前選取法會持續進行，直到符合指定的條件為止。

- **在選取了指定號碼的特徵時中止。**在強制用於模型的功能之外，此演算法會加上固定數目的功能。指定一個正整數。降低數目的值會建立較精簡的模式，但是有可能遺漏重要功能。增加數目的值將擷取所有的重要功能，但是有可能加入會實際造成模式錯誤的功能。
- **在絕對誤差比例的變化小於或等於最小值時中止。**當絕對錯誤比例中的變更指示加入更多功能也無法進一步改善模式，則此演算法會停止。指定正數。減少下限變化值將傾向於包含更多特徵，但存在包含對模型價值不大的特徵的風險。提高最小變更的值會排除較多功能，但是有可能遺漏對於模式很重要的功能。最小變更的「最佳」值需視您的資料和應用方式而定。請參閱輸出的「功能選擇錯誤記錄」，也協助您評估那些功能最重要。請參閱第 299 頁的『預測值選擇錯誤日誌』主題，以取得更多資訊。

交叉驗證

唯有當目的為預測目標時才會啟動此畫面。該窗格上的選項控制計算最近鄰接項時是否使用交叉驗證。

交叉驗證將樣本劃分為許多子樣本，或**褶疊**。然後會產生最近鄰法模式，並且從每個子樣本中排除資料。第一個模式是以第一個樣本折疊中以外的所有觀察值為基礎，第二個模式是以第二個範例折疊中以外的所有觀察值為基礎，依此類推。對於各個模型，都會將模式套用至在產生時被排除的子樣本，以評估錯誤。最近鄰法的「最佳」數目是在折疊產生最低錯誤的數目。

交叉驗證折疊。 *V* 折疊交叉驗證可用來判斷相鄰的「最佳」數目。基於效能考量，這無法搭配功能選擇使用。

- **將觀察值隨機指派給折疊。** 指定應該用於交叉驗證的折疊數目。此程序會將觀察值指派給折疊，數目介於 1 到 *V* (折疊的數目)。
- **設定隨機種子。** 當根據隨機百分比來估計模型的精確度時，此選項可讓您將相同的結果複製到其他階段作業中。透過指定亂數產生器所用的起始值，您可以確保每次執行節點時都指派相同的記錄。輸入期望的種子值。如果未選取此選項，則每次執行節點時，都將產生不同樣本。
- **使用欄位指定觀察值。**指定一個將作用中資料集中的每個觀察值分配到摺疊中的數值型欄位。此欄位必須是數值，並且接受從 1 到 *V* 的值。如果此範圍中的任何值遺漏，並且任何分隔欄位上的分隔模型都有效，則此值將導致錯誤。

分析

只有在目的是為了預測目標時才啟動「分析」畫面。您可以使用它指定模型是否要納入附加變數以包含：

- 每個可能目標欄位值的機率
- 觀察值和最近鄰接項之間的距離
- 原始和已調整的傾向分數（僅適用於旗標目標）。

附加所有機率。 如果勾選了此選項，則會針對該節點所處理的每一筆記錄，顯示標稱或旗標目標欄位的每個可能值的機率。如果未勾選此選項，則針對標稱或旗標目標欄位只會顯示預測值及其機率。

儲存觀察值和 k 個最近鄰接項之間的距離。對於每個焦點記錄，將為其 *k* 個最近鄰接項（來自訓練樣本）和對應的 *k* 個最近距離建立個別的變數。

傾向分數

可以在建模節點中和模型塊的「設定」標籤上啟用傾向分數。唯有當選取的目標是旗標欄位時此功能才可用。請參閱第 30 頁的『傾向分數』主題，以取得更多資訊。

計算原始傾向評分。 原始傾向分數僅衍生自基於訓練資料的模型。如果模型預測值為 *true*（將回應），那麼傾向與 P 相同，其中 P 為預測的可能性。如果模型預測的值為假，那麼計算出的傾向為 (1 – P)。

- 如果建立模型時選擇了此選項，那麼依預設將在模型塊中啟用傾向分數。不過，無論是否在建模節點中選擇了原始傾向分數，都可以始終在模型塊中選擇啟用原始傾向分數。
- 對模型進行評分時，原始傾向評分將被新增到將 *RP* 字母附加到標準字首的欄位中。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RRP-churn*。

計算調整傾向評分。 原始傾向僅基於由可能過度擬合的模型指定的估計，這將導致過於樂觀地估計傾向。已調整的傾向試圖通過查看模型在測試或驗證分割區的性能或通過調整傾向來彌補，以相應地給作出更好的估計。

- 此設定要求串流中出現有效的分割區欄位。
- 與原始信賴度分數不同，已調整的傾向評分必須在建立模型時計算；否則，對模型塊進行評分時該分數將不存在。

- 對模型進行評分時，在將 *AP* 字母附加到標準字首的欄位中新增已調整的傾向評分。例如，如果預測位於名為 *$R-churn* 的欄位中，那麼傾向分數欄位的名稱將是 *$RAP-churn*。已調整的傾向分數不適用於邏輯迴歸模型。
- 在計算已調整的傾向分數時，必須尚未已平衡用於計算的測試或驗證分割區。為避免這一點，請確保在任何上游平衡節點中已選取**僅平衡訓練資料**選項。此外，如果已在上游獲取了複合樣本，那麼這將導致已調整的傾向分數無效。
- 已調整的傾向分數不適用於「增強型」樹狀結構和規則集模型。請參閱第 105 頁的『增強型 C5.0 模型』主題，以取得更多資訊。

KNN 模型塊

KNN 模型會建立許多新欄位，如下表格所示。要檢視這些欄位及其值，請將表格節點新增到 KNN 模型塊中，然後執行表格節點，或按一下模型塊上的「預覽」按鈕。

表 *32. KNN* 模型欄位

新欄位名稱	說明
$KNN-*fieldname*	目標欄位的預測值。
$KNNP-*fieldname*	預測值的機率。
$KNNP-*value*	列名或旗標欄位的每個可能值的機率。只有在模型塊的「設定」標籤上已勾選了**附加所有可能性**才會被納入。
$KNN-neighbor-*n*	焦點記錄的第 *n* 個最近相鄰元素名稱。唯有當模型區塊的「設定」標籤上的 **顯示最近**設為非零值時才包括此項。
$KNN-distance-*n*	焦點記錄第 *n* 個最近相鄰元素到焦點記錄的相對距離。唯有當模型區塊的「設定」標籤上的 **顯示最近**設為非零值時才包括此項。

最近相鄰元素模型視圖

模型視圖

模型視圖有 2 個畫面視窗：

- 第一個畫面會顯示模型的概述，稱為主要視圖。
- 第二個窗格會顯示兩種檢視類型的其中一種：

  輔助模型視圖會顯示模型的詳細資訊，但是焦點不著重於模型本身。

  鏈結檢視會顯示使用者探索主要顯示各部分時模式功能的詳細資訊。

依預設，第一個窗格顯示預測值空間，第二個窗格顯示預測值重要性圖表。如果預測值重要性圖表無法使用；即如果未在「設定」標籤的「鄰接項」窗格上已選取**按照重要性計算特徵加權**，那麼將顯示「視圖」下拉清單格中的第一個可用視圖。

如果視圖不具有可用資訊，它將從「視圖」下拉清單中省略。

預測值空間： 預測值空間圖表是有關預測值空間（如果存在 3 個以上預測值，那麼為子空間）的互動式圖形形。每條軸代表模型中的某個預測值，圖表中的點位置顯示觀察值這些預測值在訓練和 holdout 分割區中的值。

鍵。 除了預測值外，圖形中的點還傳遞其他資訊。

- 形狀表示點所屬的訓練或保留區隔。

- 點的色彩/陰影表示該觀察值的目標值，其中不同色彩值等於類別目標的類別，而形狀表示連續目標的值範圍。訓練區隔的指示值是觀察值；對於保留區隔，這是預測值。如果未指定任何目標，則不會顯示任何鍵。
- 較粗的外框表示觀察值為焦點。顯示的焦點記錄鏈結至它們的 *k* 個最近鄰居。

控制和互動性。 使用圖表中的一些控制項可以探索預測值空間。

- 可以選擇在圖表中顯示哪個預測值子集合，還可變更在維度上代表哪些預測值。
- 「焦點記錄」僅僅是在「預測值空間」圖表中選定的點。如果指定了焦點記錄變數，那麼初始情況下會已選取代表焦點記錄的點。不過，如果選中了任何點，那麼它都可以暫時成為焦點記錄。其中適用點選擇的「一般」控制；按一下點會選取該點，並取消選取其他所有點；按住 Ctrl 並按一下點會將該點加入所選取的多個點中。鏈結的視圖，如對等項圖表，將根據在預測值空間中選取的觀察值自動更新。
- 您可以變更為焦點記錄顯示的最近鄰居號碼 (*k*)。
- 游標停留於圖表中的點時，會顯示個案標籤值的工具提示（如果未定義個案標籤，則顯示個案編號），以及觀察和預測目標值。
- 通過「重設」按鈕，您可以將「預測值空間」恢復到其原始狀態。

變更預測值空間圖表上的軸： 您可以控制在預測值空間圖表的軸上顯示的特徵。

要變更軸設定：

1. 按一下左側窗格上的「編輯模式」按鈕（畫筆圖示），為預測值空間選取編輯模式。
2. 在右側窗格中變更視圖。在兩個主要窗格之間出現**顯示區域**畫面。
3. 按一下**顯示區域**勾選框。
4. 按一下預測值空間中的任何資料點。
5. 要使用具有相同資料類型的預測值取代某個軸：
   - 將新預測值拖到您要取代的預測值的區域標籤（帶有小 X 按鈕）上。
6. 要使用具有不同資料類型的預測值取代某個軸：
   - 在您要取代的預測值的區域標籤上，按一下小 X 按鈕。預測值空間變為二維度視圖。
   - 將新預測值拖到**新增維度**區域標籤上。
7. 按一下左側窗格上的「探索模式」按鈕（箭頭的方向圖示），結束編輯模式。

預測值重要性： 一般而言，您會想要將焦點著重在建模過程中最重要的預測值欄位，並考慮捨棄或忽略最不重要的預測值欄位。預測值重要性圖可協助您指出評估模式時各預測值的相對重要性，以達成此目標。由於其中的值都是相對值，因此顯示中所有預測值的值總和為 1.0。預測值重要性與模式準確性無關。這只涉及進行預測時各預測值的重要性，而不涉及預測是否正確。

最近鄰接項距離： 該表格只顯示焦點記錄的 *k* 個最近鄰居與距離。如果焦點記錄 ID 指定在 建模節點上，那麼它為可用，且只顯示此變數 ID 的焦點記錄。

各列的：

- **焦點記錄**欄包含焦點記錄的觀察值標籤變數值；如果未定義觀察值標籤，那麼此欄包含焦點記錄的觀察值編號。
- 在**最近鄰接項**群組下的第 *i* 欄包含焦點記錄的第 *i* 個最近鄰接項的觀察值標籤變數值；如果未定義觀察值標籤，那麼此欄包含焦點記錄第 *i* 個最近鄰接項的觀察值數目。
- 在**最近距離**群組下的第 *i* 欄包含第 *i* 個最近相鄰元素與焦點記錄的距離。

對等項： 該圖表顯示焦點觀察值及其在每個預測值和目標上 k 個最近鄰接項。它僅在預測值空間圖表中選取了焦點觀察值時可用。

對等項圖表以兩種方式鏈結至預測值空間。

- 在預測值空間中所選的觀察值（焦點觀察值）顯示在對等項圖表中，也包括其 k 個最近鄰居。
- 在對等項圖表中使用在預測值空間中所選的 k 值。

選取預測值。 使您可選取在對等項圖表中顯示的預測值。

象限圖： 該圖表顯示焦點觀察值及其在散佈圖（或點圖，取決於目標的測量層次）上 k 個最近鄰接項。目標在 y 軸上，尺度預測值在 x 軸上，按預測值劃分窗格。它僅當存在目標，且在預測值空間圖表中選取了焦點觀察值時可用。

- 對於連續變數會在訓練區隔的變數平均數繪製參考線。

選取預測值。 使您可選取在象限圖中顯示的預測值。

預測值選擇錯誤日誌： 對於該圖表上的點，其 y 軸值為模型的誤（錯誤率或平方和誤差，取決於目標的測量層次），x 軸上列出模型的預測值（加號上 x 軸左側的所有功能）。如果目標和功能選擇有效，則可使用此圖表。

分類表： 此表按照區隔顯示目標觀察值與預測值的交叉分類。它僅當存在種類目標（旗標、列名或序數）時可用。

- 「保留」區隔的 **(遺漏)** 列包含具有目標遺漏值的保留觀察值。這些觀察值會構成保留樣本：整體百分比，而非正確百分比值。

誤摘要： 如果有目標變數，則可使用此表。此表會顯示模型相關的錯誤，以及類別目標的連續目標與誤差率（100% 正確的整體百分比）平方和。

KNN 模型設定

通過「設定」標籤可以指定在檢視結果時顯示的附加欄位（例如，通過執行表格節點附加到塊）。通過選取這些選項可以檢視每個選項的效果,並且按一下「預覽」按鈕（捲動至「預覽」輸出右側）可以檢視附加欄位。

附加所有可能性（僅對種類目標有效）。如果勾選了此選項，則會針對該節點所處理的每一筆記錄，顯示標稱或旗標目標欄位的每個可能值的機率。如果未勾選此選項，則針對標稱或旗標目標欄位只會顯示預測值及其機率。

此勾選框的預設設定由建模節點的相應勾選框確定。

計算原始傾向評分。 對於具有旗標目標的模型（傳回 yes 或 no 預測），您可以要求傾向評分以指出針對目標欄位指定的 true 結果的概似性。這些值是除了其他預測與信賴值以外，在評分期間可能產生的值。

計算調整傾向評分。 原始傾向分數僅依賴於訓練資料，並且由於多數模型過度擬合此資料的傾向，該分數可能會過度優化。調整傾向會嘗試透過向測試或驗證分割區評估模型效能來進行補償。此選項要求在產生模型之前，在串流中定義分割區欄位並且在建模節點中啟用調整傾向評分。

顯示最近。如果您將此值設為 n，其中 n 是非零正整數，那麼焦點記錄的第 n 個最近鄰接項與其到焦點記錄的相對距離一起納入在模型中。

第 17 章 Python 節點

SPSS Modeler 提供了用於使用 Python 原生演算法的節點。 節點選用區上的 **Python** 標籤包含您可用於執行 Python 演算法的下列節點。這些節點在 Windows 64、Linux64 和 Mac 上受支援。

Synthetic Minority Over-sampling Technique (SMOTE) 節點提供一個過度取樣演算法來處理不平衡的資料集。它提供進階方法來平衡資料。 SMOTE 處理節點在 SPSS Modeler 中使用 Python 進行實現並且需要 `imbalanced-learn`© Python 程式庫。

XGBoost Linear© 是將線性模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 SPSS Modeler 中的 XGBoost Linear 節點使用 Python 進行實現。

XGBoost Tree© 是將樹狀結構模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost Tree 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost Tree 節點僅顯示了核心功能和一般參數。該節點是以 Python 來實作的。

t-Distributed Stochastic Neighbor Embedding (t-SNE) 是一套用於視覺化高維度資料的工具。它會將資料點的親緣性轉換為可能性。此 t-SNE 節點在 SPSS Modeler 中使用 Python 進行實現並且需要 `scikit-learn`© Python 程式庫。

Gaussian Mixture© 模型是一種概率模型，它假設所有資料點都從使用未知參數進行無限次數混合產生的。人們可以將混合模型視為產生 k-means 叢集以納入有關資料的協方差結構以及潛在 Gaussians 中心的資訊。SPSS Modeler 中的 Gaussian Mixture 節點顯示了 Gaussian Mixture 程式庫的核心功能及常用參數。該節點是以 Python 來實作的。

Kernel Density Estimation (KDE)© 使用 Ball Tree 或 KD Tree 演算法來執行高效率查詢，並結合非意外學習、功能工程和資料建模的概念。基於鄰接項的方法（例如 KDE）是最熱門且最有用的密度估計技術。SPSS Modeler 中的 KDE 建模和 KDE 模擬節點揭示了 KDE 程式庫的核心功能和常用參數。這些節點是以 Python 來實作的。

隨機森林節點使用將樹狀結構模型用作基底模型的 bagging 演算法的進階實現。此隨機森林建模節點在 SPSS Modeler 中使用 Python 進行實現並且需要 `scikit-learn`© Python 程式庫。

© Copyright IBM Corp. 1994, 2019

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非監督式學習來尋找資料集的叢集或密集區域。SPSS Modeler 中的 HDBSCAN 節點顯示了 HDBSCAN 程式庫的核心功能及常用參數。該節點在 Python 中實作，當您一開始不瞭解那是些什麼群組時，您可以使用它來將資料集叢集至不同的群組。

「一類 SVM」節點使用未受監督的學習演算法。該節點可用來偵測新事件。它將偵測給定樣本集的軟性界限，然後將新的點分類成是否的該集合。此一級 SVM 建模節點在 SPSS Modeler 中使用 Python 進行實現並且需要 `scikit-learn`© Python 程式庫。

SMOTE 節點

Synthetic Minority Over-sampling Technique (SMOTE) 節點提供一個過度取樣演算法來處理不平衡的資料集。它提供進階方法來平衡資料。 SMOTE 處理節點使用 Python 進行實現並且需要 `imbalanced-learn`© Python 程式庫。有關 imbalanced-learn 程式庫的詳細資料，請參閱 http://contrib.scikit-learn.org/imbalanced-learn/about.html[1]。

節點選用區上的 Python 標籤包含 SMOTE 節點和其他 Python 節點。

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

SMOTE 節點設定

在 SMOTE 節點的**設定**標籤上定義下列設定。

目標設定

目標欄位。選取目標欄位。支援所有旗標、名義、序數及離散測量類型。如果在「分割區」部分中選取了**使用分割的資料**選項，那麼將對訓練資料進行過採樣。

超額取樣比例

選取**自動**以自動選取過採樣比例，或者選取**設定比例（少數相較多數）**以設定自訂比例值。比例是少數類別中的樣本數除以多數類別中的樣本數。此值必須大於 **0** 並小於或等於 **1**。

隨機種子

設定隨機種子。 選取此資訊並按一下**產生**可以產生由亂數字產生器使用的種子。

方法

演算法類型。選取您要使用的 SMOTE 演算法的類型。

樣本規則

K 鄰近值。指定要用於構建合成樣本的最近鄰居的數量

M 鄰近值。指定要用於確定是否少數樣本處於危險狀態的最近鄰居的數量。僅當選取了 **Borderline1** 或 **Borderline2** SMOTE 演算法類型時，才會使用此選項。

分割區

使用分割的資料。 如果您僅希望對訓練資料進行過採樣，請選取此選項。

此 SMOTE 節點需要 imbalanced-learn© Python 程式庫。下表顯示 SPSS Modeler SMOTE 節點對話框中的設定與 Python 演算法之間的關係。

表 *33.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	Python API 參數名稱
超額取樣比例（數字輸入控制）	sample_ratio_value	ratio
隨機種子	random_seed	random_state
K 鄰居	k_neighbours	k
M 鄰居	m_neighbours	m
演算法類型	algorithm_kind	kind

XGBoost 線性節點

XGBoost Linear© 是將線性模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 SPSS Modeler 中的 XGBoost Linear 節點使用 Python 進行實現。

有關提升演算法的進一步資訊，請參閱 XGBoost 指導教學，網址為 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。[1]

請注意，SPSS Modeler 中不支援 XGBoost 交叉驗證功能。您可以將 SPSS Modeler 分割區節點用於此功能。另外，請注意，XGBoost 在 SPSS Modeler 中用於自動對種類變數執行 One-Hot 編碼。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost Linear 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

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。 要手動分配目標和預測值，請選取此選項。

欄位。 使用方向鈕可以將項目從清單中手動分配給畫面右側的「目標」和「預測值角色」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。 選取要用作預測的目標的欄位。

預測值。選取一或多個欄位作為預測的輸入。

XGBoost Linear 節點的「建置選項」標籤

使用「建置選項」標籤可以指定 XGBoost Linear 節點的建置選項，包括線性提升參數和模型建置之類的**基本選項**以及用於目標的**學習作業選項**。如需這些選項的相關資訊，請參閱下列線上資源：

- XGBoost Parameter Reference[1]
- XGBoost Python API[2]

- XGBoost 首頁[3]

基本

超參數最佳化（基於 Rbfopt）。選取此選項以基於 Rbfopt 啟用超參數最佳化，其會自動探索參數最佳組合，以便模型可對樣本達到預期或較小錯誤率。如需 Rbfopt 的詳細資料，請參閱 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

Alpha 值。這是有關加權的 L1 正規化項目。增大此值將使模型更保守。

Lambda。 這是有關加權的 L2 正規化項目。增大此值將使模型更保守。

Lambda (λ) 偏差。這是有關基本選項目的 L2 正規化項目。（沒有關於偏移的 L1 正規化項目，因為它不重要。）

提高循環數。這是提升疊代的次數。

學習作業

目標。從下列學習作業目標類型中進行選取：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic** 或 **multi**。

隨機種子。您可以按一下**產生**來產生亂數字產生器所使用的種子。

下表格顯示了 SPSS Modeler XGBoost Linear 節點對話框中的設定與 Python XGBoost 程式庫參數之間的關係。

表 *34.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	XGBoost 參數
目標	`TargetField`	
預測	`InputFields`	
Lambda (λ) 值	`Lambda (λ) 值`	`Lambda (λ) 值`
Alpha	`alpha`	`alpha`
Lambda (λ) 偏差	`lambdaBias`	`lambda_bias`
提高循環數	`numBoostRound`	`num_boost_round`
目標	`objectiveType`	`objective`
隨機種子	`random_seed`	`seed`

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost Linear 節點模型選項

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。

XGBoost 樹狀結構節點

XGBoost Tree© 是將樹狀結構模型用作基底模型的梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost Tree 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost Tree 節點僅顯示了核心功能和一般參數。該節點是以 Python 來實作的。

有關提升演算法的進一步資訊，請參閱 XGBoost 指導教學，網址為 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。[1]

請注意，SPSS Modeler 中不支援 XGBoost 交叉驗證功能。您可以將 SPSS Modeler 分割區節點用於此功能。另外，請注意，XGBoost 在 SPSS Modeler 中用於自動對種類變數執行 One-Hot 編碼。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost Tree 節點的「欄位」標籤

「欄位」標籤指定要在分析中使用的欄位。

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。 要手動分配目標和預測值，請選取此選項。

欄位。 使用方向鈕可以將項目從清單中手動分配給畫面右側的「目標」和「預測值角色」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。 選取要用作預測的目標的欄位。

預測值。選取一或多個欄位作為預測的輸入。

XGBoost Tree 節點的「建置選項」標籤

使用「建置選項」標籤可以指定 XGBoost Tree 節點的建置選項，包括用於模型建置和樹狀結構成長的**基本選項**、用於目標的**學習作業選項**以及用於控制不已平衡資料集的過度配適及處理的**進階選項**。如需這些選項的相關資訊，請參閱下列線上資源：

- XGBoost Parameter Reference[1]
- XGBoost Python API[2]
- XGBoost 首頁[3]

基本

超參數最佳化（基於 Rbfopt）。選取此選項以基於 Rbfopt 啟用超參數最佳化，其會自動探索參數最佳組合，以便模型可對樣本達到預期或較小錯誤率。如需 Rbfopt 的詳細資料，請參閱 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

樹狀結構方法。選取要使用的 XGBoost Tree 構建演算法。

提高循環數。 指定提升疊代的次數。

深度上限。指定樹狀結構的最大深度。增大此值將導致模型更複合，並且很可能出現過度配適。

最小子項加權。指定子代中需要的實例加權 (hessian) 的下限總和。如果樹狀結構分區步驟生成實例加權總和少於此**最小子項加權**的葉節點，那麼建置程序將停止於進行進一步分區。在線性迴歸模式下，此項簡單地對應於每個節點中所需的下限實例數。加權越大，演算法越保守。

最大差異步驟。指定容許用於每個樹狀結構的加權估計的最大差異步驟。如果設定為 **0**，那麼沒有限制。如果設定為正值，那麼它可以使更新步驟更為保守。通常不需要此參數，但是在某個類別極度不已平衡的情況下，它可以用在邏輯迴歸中。

學習作業

目標。從下列學習作業目標類型中進行選取：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic** 或 **multi**。

提前停止。如果您想要使用提前停止功能，請選取此選項。對於**停止回合**，驗證錯誤必須至少在每個提前停止回合減少，才能繼續訓練。**評估資料比例**是用於驗證錯誤的輸入資料比例。

隨機種子。您可以按一下**產生**來產生亂數字產生器所使用的種子。

進階

子樣本。子樣本是訓練實例的比例。例如，如果您將此項設定為 **0.5**，那麼 XGBoost 將隨機收集一半的資料實例以生成樹狀結構，並且這將防止過度配適。

Eta。這是更新步驟期間用於防止過度配適的步驟大小收縮。在每個提升步驟後，可以直接獲取新功能的加權。Eta 也會縮小功能加權，以使提升過程更保守。

伽瑪參數。 這是對樹狀結構的某個葉節點進行進一步分割區所需的下限損失減小。Gamma 設定越大，演算法越保守。

Colsample（依樹狀結構）。這是構建每個樹狀結構時欄的子樣本比例。

Colsample（依層次）。這是在每個層次每個分割的欄的子樣本比例。

Lambda。 這是有關加權的 L2 正規化項目。增大此值將使模型更保守。

Alpha 值。這是有關加權的 L1 正規化項目。增大此值將使模型更保守。

比例 POS 加權。用於控制正加權和負加權的平衡。這對於不平衡類別非常有用。

下表格顯示了 SPSS Modeler XGBoost Tree 節點對話框中的設定與 Python XGBoost 程式庫參數之間的關係。

表 *35.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	XGBoost 參數
目標	`TargetField`	
預測	`InputFields`	
樹狀結構方法	`treeMethod`	`tree_method`
提高循環數	`numBoostRound`	`num_boost_round`
深度上限	`maxDepth`	`max_depth`
最小子項加權	`minChildWeight`	`min_child_weight`
最大差異步驟	`maxDeltaStep`	`max_delta_step`

表 35. 對映至 *Python* 程式庫參數的節點內容 (繼續)

SPSS Modeler 設定	Script 名稱（內容名稱）	XGBoost 參數
目標	objectiveType	objective
提前停止	earlyStopping	early_stopping_rounds
停止回合	stoppingRounds	
評估資料比例	evaluationDataRatio	
隨機種子	random_seed	seed
子樣本	sampleSize	subsample
Eta	Eta 值	Eta 值
Gamma	伽瑪分配	伽瑪分配
Colsample（依樹狀結構）	colsSampleRatio	colsample_bytree
Colsample（依層次）	colsSampleLevel	colsample_bylevel
Lambda (λ) 值	Lambda (λ) 值	Lambda (λ) 值
Alpha	alpha	alpha
比例 POS 加權	scalePosWeight	scale_pos_weight

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost Tree 節點的「構建選項」標籤

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。

t-SNE 節點

t-Distributed Stochastic Neighbor Embedding (t-SNE)© 是一套用於視覺化高維度資料的工具。它會將資料點的親緣性轉換為可能性。原始空間中的親緣性由 Gaussian 聯合機率代表，而內嵌空間中的親緣性則是由「學生 t 分布 (Student's t-distributions)」代表。這可讓 t-SNE 對於本端結構，並且擁有幾個超越現有技術的優點：[1]

- 在單一地圖上以許多比例顯示結構
- 顯示位於多個、不同、各種、叢集的資料
- 降低人群一起指向中心的傾向

此 t-SNE 節點在 SPSS Modeler 中使用 Python 進行實現並且需要 scikit-learn© Python 程式庫。有關 t-SNE 和 scikit-learn 程式庫的詳細資料，請參閱：

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

節點選用區上的 Python 標籤包含此節點和其他 Python 節點。「圖表」標籤上也可以使用 t-SNE 節點。

[1] 參考書目：

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE 節點專家選項

選擇**簡式**模式或**專家**模式，視您要對 t-SNE 節點設定的選項而定。

視覺化類型。 選取 **2D** 或 **3D**，指定要以二維還是三維來繪製圖表。

方法。選取 **Barnes Hut** 或 **Exact**。依預設，梯度計算演算法會使用 Barnes-Hut 近似值，其執行速度遠比 Exact 方法快。Barnes-Hut 近似值可讓 t-SNE 技術能套用到大型的實際資料集。Exact 演算法在避免最近鄰接項錯誤方面的運作較好。

起始設定。選取**隨機**或 **PCA** 作為內嵌的起始設定。

目標欄位。選取要顯示為輸出圖表上的彩色圖的目標欄位。如果未在這裡指定目標欄位，圖表將會使用一個顏色。

最佳化

複雜度。 複雜度與其他各種學習演算法中所用的最近鄰接項數目相關。較大的資料集通常需要較大的複雜度。請考量選取介於 **5** 與 **50** 之間的值。預設值是 **30**，且範圍是 **2 - 9999999**。

提早誇大。 這項設定控制原始空間中的自然叢集在內嵌空間中將會有多緊，以及它們之間將會有多少空間。預設值是 **12**，且範圍是 **2 - 9999999**。

學習率。 如果學習率太高，則資料很可能類似於「球形」，每個點都近乎與最近的鄰接項等距。如果學習率太低，則大部分點可能看起來壓縮在密雲中，且極少數點是極端值。如果成本停滯在較差的區域最小值，則提高學習率可能有所幫助。預設值是 **200**，且範圍是 **0 - 9999999**。

最大疊代數。 最佳化的疊代數目上限。預設值是 **1000**，且範圍是 **250 - 9999999**。

角距大小。 從一個點所測量的遠距節點的角距大小。輸入介於 **0** 與 **1** 之間的值。預設值是 **0.5**。

隨機種子

設定隨機種子。 選取此資訊並按一下**產生**可以產生由亂數字產生器使用的種子。

最佳化停止條件

沒有進度的最大疊代數。 在停滯最佳化之前沒有執行進度的最大疊代數，用在最初的 250 個疊代之後，且具有早期誇張。請注意，僅每 50 個疊代檢查進度一次，因此這個值捨入為下一個 50 的倍數。預設值是 **300**，且範圍是 **0 - 9999999**。

最小梯度規範。 如果梯度規範低於此最小臨界值，最佳化將會停止。預設值是 **1.0E-7**。

度量值。 計算特性陣列中實例之間的距離時使用的度量值。如果度量值是字串，則必須是 scipy.spatial.distance.pdist 針對其度量參數容許的其中一個選項，或者 pairwise.PAIRWISE_DISTANCE_FUNCTIONS 中列出的度量值。選取其中一個可用的度量值類型。預設值是 **euclidean**。

當記錄數大於。指定繪製大型資料集的方法。您可以指定資料集大小上限，或者使用預設 2,000 點。當您選取 **Bin** 或**範例**選項時，會加強大型資料集的效能。或者，您可以選擇透過選取**使用所有資料**繪製所有資料點，但是您應該注意到這可能自動降低軟體的效能。

- **Bin。**選取以在資料集包含的記錄數超過指定數目時啟用 Binning。Binning 在實際繪製之前將圖形分為較小的網格，並計數將在每一個網格資料格中出現的連線數目。在最終圖形中，在 Bin centroid，每個資料格使用一個連線（Bin 中所有連線點的平均值）。
- **取樣。**選取以隨機取樣包含指定記錄數的資料。

下表格顯示了 SPSS Modeler t-SNE 節點對話框的「專家」標籤上的設定與 Python t-SNE 程式庫參數之間的關係。

表 *36.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	Python t-SNE 參數
模式	mode_type	
視覺化類型	n_components	n_components
方法	method	method
內嵌起始設定	init	init
目標	target_field	target_field
複雜度	perplexity	perplexity
提早誇大	early_exaggeration	early_exaggeration
學習率	learning_rate	learning_rate
最大疊代數	n_iter	n_iter
角距大小	angle	angle
設定隨機種子	enable_random_seed	
隨機種子	random_seed	random_state
沒有進度的最大疊代數	n_iter_without_progress	n_iter_without_progress
最小梯度規範	min_grad_norm	min_grad_norm
使用多重複雜度來執行 t-SNE	isGridSearch	

t-SNE 節點輸出選項

在**輸出**標籤上指定 t-SNE 節點輸出的選項。

輸出名稱。指定節點執行時產生的輸出名稱。如果您選取**自動**，會自動設定輸出的名稱。

輸出至畫面。選取此選項可在新視窗中產生並顯示輸出。輸出也會新增至「輸出」管理程式。

輸出至檔案。選取此選項可將輸出儲存至檔案。這麼做會啟用**檔名**和**檔案類型**欄位。如果您想要使用其他欄位建立繪圖以進行比較，或者將其輸出作為分類或迴歸模型中的預測值，則 t-SNE 節點需要存取此輸出檔。t-SNE 模型會建立包含 x、y（及 z）座標欄位的結果檔，可輕鬆使用「固定檔案」來源節點進行存取。

t-SNE 模型塊

t-SNE 模型塊包含 t-SNE 模型所擷取的所有資訊。可用的標籤如下。

圖表

圖表標籤顯示 t-SNE 節點的圖表輸出。pyplot 散佈圖顯示低維度結果。如果您未選取 t-SNE 節點的專家標籤上的**使用多重複雜度來執行 t-SNE** 選項，則只會包含一個圖表而不是六個具有不同複雜度的圖表。

文字輸出

文字輸出標籤顯示 t-SNE 演算法的結果。如果您在 t-SNE 節點的專家標籤上選擇 **2D** 視覺化類型，則這裡的結果為二維點值。如果您選擇 **3D**，則結果為三維點值。

Gaussian Mixture 節點

Gaussian Mixture© 模型是一種概率模型，它假設所有資料點都從使用未知參數進行無限次數混合產生的。人們可以將混合模型視為產生 k-means 叢集以納入有關資料的協方差結構以及潛在 Gaussians 中心的資訊。[1]

SPSS Modeler 中的 Gaussian Mixture 節點顯示了 Gaussian Mixture 程式庫的核心功能及常用參數。該節點是以 Python 來實作的。

如需 Gaussian Mixture 建模演算法和參數的相關資訊，請參閱 Gaussian Mixture 說明文件，網址為：http://scikit-learn.org/stable/modules/mixture.html 和 https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html。[2]

[1] "User Guide." *Gaussian mixture models*. Web. © 2007 - 2017. scikit-learn developers.

[2] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Gaussian Mixture 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

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）中的輸入設定。

使用自訂欄位指派。 若要手動指派輸入，請選取此選項。

欄位。 使用箭頭按鈕可以將此清單中的項目手動指派給畫面右側的「預測值」清單。圖示指出每個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。

預測值。選取一個或多個欄位作為預測值。

Gaussian Mixture 節點建置選項

使用「建置選項」標籤可以指定 Gaussian Mixture 節點的建置選項，包括**基本選項**及**進階選項**。如需在此章節中未涵蓋的這些選項的相關資訊，請參閱下列線上資源：

- Gaussian mixture 參數參照[1]
- Gaussian mixture 節點使用手冊[2]

基本

協方差類型。 請選取下列一個協方差矩陣：

- **完全**。每個元件都會有它自己的一般協方差矩陣。
- **並列**。所有元件共用相同的一般協方差矩陣。
- **對角線**。每個元件都會有它自己的對角線協方差矩陣。
- **球面**。每個元件都會有它自己的單一變異。

元件數目。指定要在建置模型時使用的 mixture 元件數目。

叢集標籤。指定叢集標籤是數字還是字串。如果您選擇**字串**，請為叢集標籤指定字首（例如，預設字首為 cluster，這會產生 **cluster-1**、**cluster-2** 等叢集標籤）。.

隨機種子。選取此資訊並按一下**產生**可以產生由亂數字產生器使用的種子。

進階

容錯。指定聚合臨界值。預設值是 **0.001**。

疊代次數。指定要執行的疊代次數上限。預設值是 **100**。

起始設定參數。選取起始設定參數 **Kmeans**（使用 k-means 起始設定的責任）或 **Random**（隨機起始設定的責任）。

熱啟動。如果您選取 **True**，將會使用前次適合的解決方案作為下一次適合的起始設定。這可在針對類似的問題多次呼叫適合時，加快聚合的速度。

下表顯示了 SPSS Modeler Gaussian Mixture 節點對話框中的設定與 Python Gaussian 程式庫參數之間的關係。

表 *37.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	Gaussian Mixture 參數
使用預先定義的角色/使用自訂欄位指派	role_use	
輸入	預測值	
使用分割的資料	use_partition	
共變異類型	covariance_type	covariance_type
元件數目	number_component	n_components
叢集標籤	component_lable	
標籤字首	label_prefix	
設定隨機種子	enable_random_seed	
隨機種子	random_seed	random_state
容錯	tol	tol
疊代次數	max_iter	max_iter
起始設定參數	init_params	init_params
熱啟動	warm_start	warm_start

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] "User Guide." *Gaussian mixture models*. Web. © 2007 - 2017. scikit-learn developers.

Gaussian Mixture 節點模型選項

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。

KDE 節點

Kernel Density Estimation (KDE)© 使用 Ball Tree 或 KD Tree 演算法來執行高效率查詢，並連結非意外學習、功能工程和資料建模。基於鄰接項的方法（例如 KDE）是最熱門且最有用的密度估計技術。KDE 可在任何維度數目中執行，但在實踐中發現，高維度可能導致效能退化。SPSS Modeler 中的 KDE 建模和 KDE 模擬節點揭示了 KDE 程式庫的核心功能和常用參數。這些節點是以 Python 來實作的。[1]

若要使用 KDE 節點，您必須設定上游「類型」節點。KDE 節點將從「類型」節點（或上游來源節點的「類型」標籤）讀取輸入值。

KDE 建模節點在 SPSS Modeler 的「建模」標籤和 Python 標籤上可用。「KDE 建模」節點會產生模型片段，並且片段的得分值是來自輸入資料的核心密度值。

KDE 模擬節點在「輸出」標籤和 Python 標籤上可用。「KDE 模擬」節點會產生 KDE Gen 來源節點，後者可以建立分佈與輸入資料相同的一些記錄。KDE Gen 節點包含「設定」標籤，可在其中指定節點將產生多少個記錄（預設值為 1）並產生隨機種子。

如需 KDE 的相關資訊，請參閱 KDE 說明文件，網址為：http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation。[1]

[1] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

KDE 建模節點和 KDE 模擬節點欄位

「欄位」標籤指定要在分析中使用的欄位。

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）中的輸入設定。

使用自訂欄位指派。 若要手動指派輸入，請選取此選項。

欄位。 使用箭頭按鈕可以將此清單中的項目手動指派給畫面右側的「輸入」清單。圖示指出每個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。

輸入。選取一個或多個欄位作為叢集作業的輸入。KDE 只能處理連續欄位。

KDE 節點建置選項

使用「建置選項」標籤可以指定 KDE 節點的建置選項，包括用於核心密度參數和叢集標籤的**基本選項**，以及用於容錯、葉節點大小和是否使用寬度優先方法等的**進階選項**。如需這些選項的相關資訊，請參閱下列線上資源：

- Kernel Density Estimation Python API 參數參照[1]
- Kernel Density Estimation 使用手冊[2]

基本

頻寬。指定核心的頻寬。

核心。選取要使用的核心。「KDE 核心」節點的可用核心有 **Gaussian**、**Tophat**、**Epanechnikov**、**指數**、**線性**或**餘弦**。「KDE 模擬」節點的可用核心有 **Gaussian** 或 **Tophat**。如需可用核心的詳細資料，請參閱 Kernel Density Estimation 使用手冊。[2]

演算法。對於要使用的樹狀結構演算法，選取**自動**、**Ball Tree** 或 **KD Tree**。如需相關資訊，請參閱 Ball Tree[3] 和 KD Tree。[4]

度量值。 選取距離度量值。可用的度量值有 **Euclidean**、**Braycurtis**、**Chebyshev**、 **Canberra**、**Cityblock**、**Dice**、**Hamming**、**Infinity**、 **Jaccard**、**L1**、**L2**、 **Matching**、**Manhattan**、**P**、 **Rogerstanimoto**、**Russellrao**、 **Sokalmichener**、**Sokalsneath**、 **Kulsinski** 或 **Minkowski**。如果您選取 **Minkowski**，請根據需要設定 **P值**。

在此下拉清單中提供的度量值會因您選擇的演算法而異。另請注意，僅對於 Euclidean 距離度量值，密度輸出正規化才是正確的。

進階

絕對容錯。指定結果的絕對容錯。較大的容錯往往執行時期較快。預設值為 **0.0**。

相對容錯。指定結果的所需相對容錯。較大的容錯往往執行時期較快。預設值是 **1E-8**。

葉節點大小。指定基礎樹狀結構的葉節點大小。預設值為 **40**。變更葉節點大小可能會嚴重影響效能並需要記憶體。如需 Ball Tree 和 KD Tree 演算法的相關資訊，請參閱 Ball Tree[3] 和 KD Tree。[4]

寬度優先。如果您要使用寬度優先方法，請選取 **True**，或者選取 **False** 以使用深度優先方法。

下表顯示了 SPSS Modeler KDE 節點對話框中的設定與 Python KDE 程式庫參數之間的關係。

表 *38.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	KDE 參數
輸入	inputs	
頻寬	bandwidth	bandwidth
核心	kernel	kernel
演算法	algorithm	algorithm
度量值	metric	metric
P 值	pValue	pValue
絕對容錯	atol	atol
相對容錯	rtol	Rtol
葉節點大小	leafSize	leafSize
寬度優先	breadthFirst	breadthFirst

[1] "API Reference." *sklearn.neighbors.KernelDensity*. Web. © 2007-2018, scikit-learn developers.

[2] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

[3] "Ball Tree." *Five balltree construction algorithms*. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

[4] "K-D Tree." *Multidimensional binary search trees used for associative searching*. © 1975, Bentley, J.L., Communications of the ACM.

KDE 建模節點和 KDE 模擬節點模型選項

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。

隨機森林節點

Random Forest© 是使用將樹狀結構模型用作基底模型的 bagging 演算法的進階實現。在隨機森林中，集合中的每一個樹狀結構都根據從訓練集的取代繪製的樣本（例如，重複取樣樣本）建置而來。在建構樹狀結構期間分割節點時，選擇的分割不再是所有功能的最佳分割。而挑選的分割才是隨機功能子集中的最佳分割。由於此隨機性，通常森林偏差會稍有增加（就單一非隨機樹狀結構的偏差而言），但由於平均化的存在，其變異也會減少，通常足以補償偏差增長，因此產生整體而言更好的模型。[1]

SPSS Modeler 中的「隨機森林」節點使用 Python 進行實現。節點選用區上的 Python 標籤包含此節點和其他 Python 節點。

如需隨機森林演算法的相關資訊，請參閱 https://scikit-learn.org/stable/modules/ensemble.html#forest。

[1]L. Breiman，"Random Forests"，Machine Learning，45(1)，5-32，2001。

隨機森林節點欄位

「欄位」標籤指定要在分析中使用的欄位。

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。 要手動分配目標和預測值，請選取此選項。

欄位。 使用方向鈕可以將項目從清單中手動分配給畫面右側的「目標」和「預測值角色」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。 選取要用作預測的目標的欄位。

預測值。選取一或多個欄位作為預測的輸入。

隨機森林節點建置選項

使用「建置選項」標籤可以指定「隨機森林」節點的建置選項，包括**基本選項**及**進階選項**。如需這些選項的相關資訊，請參閱 https://scikit-learn.org/stable/modules/ensemble.html#forest

基本

要建置的樹狀結構數目。選取樹系中的樹狀結構數目。

指定深度上限。如果未選取，則會展開節點，直至顯示所有葉節點，或直至所有葉節點包含少於 `min_samples_split` 個樣本。

深度上限。樹狀結構的深度上限。

葉節點大小下限。成為葉節點所需的樣本數目下限。

要用於分割的功能數目。尋找最佳分割時要考量的功能數目：

- 如果為 auto，則 max_features=sqrt(n_features)（適用於分類器），且 max_features=sqrt(n_features)（適用於迴歸）。
- 如果為 sqrt，則 max_features=sqrt(n_features)。
- 如果為 log2，則 max_features=log2(n_features)。

進階

建置樹狀結構時使用重複取樣樣本。如果選取，則在建置樹狀結構時使用重複取樣樣本。

使用開袋即用樣本來評估一般化準確度。如果選取，則使用開袋即用樣本來評估一般化準確度。

使用極度隨機的樹狀結構。如果選取，則使用極度隨機的樹狀結構，而非一般隨機森林。在極度隨機的樹狀結構中，隨機性計算分割的方法更進一步。在隨機森林中，將使用候選功能的隨機子集，但並非尋找最具區分性的臨界值，而是為每一個候選功能隨機繪製臨界值，並且挑選這些隨機所產生臨界值中最好的一個用作分割規則。這樣便可在輕微增加偏差的情況下，進一步減少模型變異。[1]

複製結果。 如果選取，將抄寫模型建置程序，以達到相同的評分結果。

隨機種子。 您可以按一下**產生**來產生亂數字產生器所使用的種子。

超參數最佳化（基於 Rbfopt）。選取此選項以基於 Rbfopt 啟用超參數最佳化，其會自動探索參數最佳組合，以便模型可對樣本達到預期或較小錯誤率。如需 Rbfopt 的詳細資料，請參閱 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

目標。 您想要達到的目標函數值（模型在樣本上的錯誤率），亦即，不明最佳值。將可接受的值設為諸如 0.01。

最大疊代。要在模型上嘗試的疊代數目上限。預設值為 1000。

最大評估數。以準確模式嘗試執行模型的功能評估最大數目。預設值為 300。

下表顯示了 SPSS Modeler「隨機森林」節點對話框中的設定與 Python 隨機森林程式庫參數之間的關係。

表 *39.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	隨機森林參數
目標	target	
預測	inputs	
要建置的樹狀結構數目	n_estimators	n_estimators
指定深度上限	specify_max_depth	specify_max_depth
深度上限	max_depth	max_depth
葉節點大小下限	min_samples_leaf	min_samples_leaf
要用於分割的功能數目	max_features	max_features
建置樹狀結構時使用重複取樣樣本	bootstrap	bootstrap
使用開袋即用樣本來評估一般化準確度	oob_score	oob_score
使用極度隨機的樹狀結構	extreme	

表 39. 對映至 *Python* 程式庫參數的節點內容 *(繼續)*

SPSS Modeler 設定	Script 名稱（內容名稱）	隨機森林參數
抄寫結果	use_random_seed	
隨機種子	random_seed	random_seed
超參數最佳化（基於 Rbfopt）	enable_hpo	
目標（針對 HPO）	target_objval	
疊代上限（針對 HPO）	max_iterations	
Max evaluations (for HPO)	max_evaluations	

[1]L. Breiman，"Random Forests"，Machine Learning，45(1)，5-32，2001。

隨機森林節點模型選項

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。

隨機森林模型片段

「隨機森林」模型片段包含隨機森林模型所擷取的所有資訊。有下列區段可用。

模型資訊

這個視圖提供有關模型的關鍵資訊，其中包括輸入欄位、One-Hot 編碼值和模型參數。

預測值重要性

這個視圖將顯示一個圖表，以指示在估計模型時所使用的各個預測值的相對重要性。如需相關資訊，請參閱第 37 頁的『預測值重要性』。

HDBSCAN 節點

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非監督式學習來尋找資料集的叢集或密集區域。SPSS Modeler 中的 HDBSCAN 節點顯示了 HDBSCAN 程式庫的核心功能及常用參數。該節點在 Python 中實作，當您一開始不瞭解那是些什麼群組時，您可以使用它來將資料集叢集至不同的群組。與 SPSS Modeler 中的大多數學習方法不同的是，HDBSCAN 模型不使用目標欄位。這種類型的學習（沒有目標欄位）稱為未受監督的學習。 HDBSCAN 模型試圖揭示輸入欄位集中的型樣而不是預測結果。記錄會進行分組，因此某個群組或叢集內的記錄彼此會相似，但不同群組中的記錄並不同。 HDBSCAN 演算法將叢集視為與低密度區域分開的高密度區域。由於這種相當普遍的觀點，HDBSCAN 找到的叢集可能會是任何形狀，與假設叢集是凸形的 k-means 相反。低密度區域中單獨存在的離群值點也會標示出來。HDBSCAN 還支援對新樣本評分。[1]

若要使用 HDBSCAN 節點，您必須設定上游「類型」節點。HDBSCAN 節點將從「類型」節點（或上游來源節點的「類型」標籤）讀取輸入值。

如需 HDBSCAN 叢集演算法的相關資訊，請參閱 HDBSCAN 說明文件，網址為：http://hdbscan.readthedocs.io/en/latest/。[1]

[1] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

重要：若要訓練 HDBSCAN 模型，您必須使用一個或多個角色設定為**輸入**的欄位。角色設定為**輸出**、**兩者**或**無**的欄位將被忽略。

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）中的輸入設定。

使用自訂欄位指派。 若要手動指派輸入，請選取此選項。

欄位。 使用箭頭按鈕可以將此清單中的項目手動指派給畫面右側的「輸入」清單。圖示指出每個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。

輸入。選取一個或多個欄位作為叢集作業的輸入。

HDBSCAN 節點建置選項

使用「建置選項」標籤可以指定 HDBSCAN 節點的建置選項，包括用於叢集參數和叢集標籤的**基本選項**，以及用於進階參數和圖表輸出的**進階選項**。如需這些選項的相關資訊，請參閱下列線上資源：

- HDBSCAN Python API 參數參照[1]
- HDBSCAN 首頁[2]

基本

超參數最佳化（基於 Rbfopt）。選取此選項以基於 Rbfopt 啟用超參數最佳化，其會自動探索參數最佳組合，以便模型可對樣本達到預期或較小錯誤率。如需 Rbfopt 的詳細資料，請參閱 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

叢集大小下限。指定叢集的大小下限。如果單一鏈結分割包含的點少於在這裡指定的值，則會將其視為「脫離」叢集的點，而非分為兩個新叢集的叢集。

最小樣本數。 指定某個點要被視為核心點，其芳鄰中的最小樣本數。如果設為 **0**，則預設值為叢集大小值下限。

演算法。 選取要使用的演算法。HDBSCAN 具有專門用於不同資料性質的變式。依預設使用 **BEST** - 這會在給定資料本質的情況下自動選擇最佳演算法。如需這些演算法類型的相關詳細資料，請參閱 HDBSCAN 說明文件。[1] 請注意，您所選擇的演算法會影響效能。例如，針對大型資料，我們建議嘗試使用 Boruvka KDTree 或 Boruvka BallTree。

距離的度量值。 選取計算功能陣列中實例之間距離時要使用的度量值。

叢集標籤。指定叢集標籤是數字還是字串。如果您選擇**字串**，請為叢集標籤指定字首（例如，預設字首為 `cluster`，這會產生 **cluster-1**、**cluster-2** 等叢集標籤）。.

進階

近似最小跨距樹狀結構。如果您要接受近似最小跨距樹狀結構，請選取 **True**。對於部分演算法來說，這可以改進效能，但產生的叢集品質可能較為低劣。如果您願意為了正確性而犧牲速度，則可能應該嘗試 **False** 選項。在大部分觀察值中，建議使用 **True**。

用來選取叢集的方法。選取用來從壓縮樹狀結構中選取叢集的方法。HDBSCAN 的標準方法會使用 Excess of Mass (**EOM**) 演算法來尋找最具持續性的叢集。您也可以在樹狀結構的葉節點處選取叢集，樹狀結構會提供最精細的同質叢集。

接受單一叢集。僅在單一叢集是資料集有效結果的情況下，將此設定變更為 **True** 以僅容許該結果。

P 值。如果針對距離使用 Minkowski 度量值（在**基本**建置選項下面），則可以根據需要變更這個 p 值。

葉節點大小。如果使用空間樹狀結構演算法（Boruvka KDTree 或 Boruvka BallTree），則為樹狀結構葉節點中的點數。此設定不會變更產生的叢集，但可能會影響演算法的執行時期。

有效性指標。選取此選項以在模型區塊輸出中包括「有效性指標」圖表。

壓縮樹狀結構。選取此選項以在模型區塊輸出中包括「壓縮樹狀結構」圖表。

單一鏈結樹狀結構。選取此選項以在模型區塊輸出中包括「單一鏈結樹狀結構」圖表。

最小跨距樹狀結構。選取此選項以在模型區塊輸出中包括「最小跨距樹狀結構」圖表。

下表顯示了 SPSS Modeler HDBSCAN 節點對話框中的設定與 Python HDBSCAN 程式庫參數之間的關係。

表 *40.* 對映至 *Python* 程式庫參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	HDBSCAN 參數
輸入	inputs	inputs
超參數最佳化	useHPO	
叢集大小下限	min_cluster_size	min_cluster_size
最小樣本數	min_samples	min_samples
演算法	algorithm	algorithm
距離的度量值	metric	metric
叢集標籤	useStringLabel	
標籤字首	stringLabelPrefix	
近似最小跨距樹狀結構	approx_min_span_tree	approx_min_span_tree
用來選取叢集的方法	cluster_selection_method	cluster_selection_method
接受單一叢集	allow_single_cluster	allow_single_cluster
P 值	p_value	p_value
葉節點大小	leaf_size	leaf_size
有效性指標	outputValidity	
壓縮樹狀結構	outputCondensed	
單一鏈結樹狀結構	outputSingleLinkage	
最小跨距樹狀結構	outputMinSpan	

[1] "API Reference." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

[2] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN 節點模型選項

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。

一級 SVM 節點

「一類 SVM©」節點使用未受監督的學習演算法。該節點可用來偵測新事件。它將偵測給定樣本集的軟性界限，然後將新的點分類成是否的該集合。此一級 SVM 建模節點使用 Python 進行實現並且需要 scikit-learn© Python 程式庫。有關 scikit-learn 程式庫的詳細資料，請參閱 http://contrib.scikit-learn.org/imbalanced-learn/about.html[1]。

節點選用區上的 Python 標籤包含一級 SVM 節點和其他 Python 節點。

註：一級 SVM 用於無監督的離群值和新內容偵測。在大多數情況下，我們建議使用已知的「正常」資料集來建立模型，以使演算法可以為指定樣本設定正確的界限。模型的參數（例如，nu、Gamma 和核心）會顯著影響結果。因此，您可能需要試驗這些選項，直到找到適合於您的狀況的最佳設定為止。

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression."*Statistics and Computing Archive*, vol. 14, no. 3, August 2004, pp. 199-222. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

一級 SVM 節點的「欄位」標籤

「欄位」標籤指定要在分析中使用的欄位。

使用預先定義的角色。 選取此選項以選取具有已定義的「輸入」角色的所有欄位。

使用自訂欄位指派。 要手動選取欄位，請選擇此選項並選擇輸入欄位和分割欄位：

輸入。選取要在分析中使用的輸入欄位。除了無類型或未知以外，受支援所有儲存類型和測量類型。如果某個欄位具有「字串」儲存類型，那麼將通過 One-Hot 編碼演算法以單個對全部的方式對此欄位的值進行二進制化。

分割。選取要用作分割欄位的一個或多個欄位。支援所有旗標、名義、序數及離散測量類型。

使用分割的資料 如果定義了分割區欄位，那麼此選項用於確保僅來自訓練分割的資料用於建立模型。

一級 SVM 節點的「專家」標籤

在一級 SVM 節點的「專家」標籤上，您可以從**簡單**模式或**專家**模式中進行選擇。如果您選擇**簡單**，那麼將使用預設值設定所有參數，如下方所示。如果您選取**專家**，那麼可以為這些參數指定自訂值。有關這些選項的詳細資訊，請參閱 http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM。

停止準則。 指定停止準則的允差。預設值為 **1.0E-3** (0.001)。

迴歸方法精準度 (nu)。訓練錯誤及支援向量分數的範圍。預設值是 **0.1**。

核心類型。要在演算法中使用的核心類型。選項包含 **RBF**、**多項式**、**Sigmoid**、**線性**或**預算**。預設值為 **RBF**。

指定 Gamma。選取此選項以指定 Gamma。否則，將套用自動 gamma。

伽瑪參數。 Gamma 設定僅可用於 RBF、多項式和 Sigmoid 核心類型。

Coef0。Coef0 僅可用於多項式和 Sigmoid 核心類型。

次數。 次數僅可用於多項式核心類型。

使用縮小啟發式。選取此選項以使用縮小啟發式。依預設會選取這個選項。

設定隨機種子。 選取此選項以設定對資料進行排列來估計可能性時要使用的亂數初始值。依預設會選取這個選項。

指定核心快取的大小 (MB)。選取此選項以指定核心快取的大小。依預設會選取這個選項。已選取此選項時，預設值為 **200** MB。

超參數最佳化（基於 Rbfopt）。選取此選項以基於 Rbfopt 啟用超參數最佳化，其會自動探索參數最佳組合，以便模型可對樣本達到預期或較小錯誤率。如需 Rbfopt 的詳細資料，請參閱 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

目標。 我們想要達到的目標函數值（模型在樣本上的錯誤率），例如，不明最佳值。將可接受的值設為諸如 **0.01**。

最大疊代。要在模型上嘗試的疊代數目上限。預設值為 **1000**。

最大評估數。嘗試執行模型的功能評估最大數目，關注點在於基於相同速度的準確度。預設值為 **300**。

一級 SVM 節點需要 scikit-learn© Python 程式庫。下表顯示 SPSS Modeler SMOTE 節點對話框中的設定與 Python 演算法之間的關係。

表 *41.* 對映至 *Python* 程式庫參數的節點內容

參數名稱	Script 名稱（內容名稱）	Python API 參數名稱
停止準則	stopping_criteria	tol
迴歸方法精準度	precision	nu
核心類型	kernel	kernel
Gamma	伽瑪分配	伽瑪分配
Coef0	coef0	coef0
碩士	程度	程度
使用縮小啟發式	shrinking	shrinking
指定核心快取的大小（數字輸入框）	cache_size	cache_size
隨機種子	random_seed	random_state

一級 SVM 節點選項

在一級 SVM 節點的標籤「選項」上，您可以設定下列選項。

「平行座標類型」圖表。SPSS Modeler 可以繪製平行坐標形來表示已建立的模型。有時，顯示一些資料欄/功能的值時它們遠大於其他值，這會導致難以看到圖表的一些其他部分。對於這種情況，您可以選擇**獨立的垂直軸**選項為所有縱軸提供獨立的軸尺度，也可以選擇**一般垂直軸**強制所有縱軸共用同一軸尺度。

圖表上的上限行數。指定要在圖形輸出中顯示的資料行的最大數量。預設值是 100。出於效能原因，最多將顯示 20 個欄位。

繪製圖形上的所有輸入欄位。選取此選項可以在圖形輸出中顯示所有輸入欄位。依預設，會將每個資料欄位繪製為一條縱軸。為了提高效能，最多將顯示 30 個欄位。

要在圖表上繪製的自訂欄位。您可以選擇此選項並選擇要顯示的一部分欄位，而不是在圖形輸出中顯示所有輸入欄位。這可以提高效能。出於效能原因，最多將顯示 20 個欄位。

第 18 章 Spark 節點

SPSS Modeler 提供了用於使用 Spark 原生演算法的節點。 節點選用區上的 **Spark** 標籤包含您可用於執行 Spark 演算法的下列節點。這些節點在 Windows 64、Mac 64 和 Linux 64 上受支援。請注意，這些節點不支援指定整數/雙欄作為旗標/名義來建置模型。若要這麼做，您必須將直欄值轉換為 0/1 或 0,1,2,3,4...

Isotonic 迴歸演算法屬於迴歸演算法系列。SPSS Modeler 中的 Isotonic-AS 節點是在 Spark 中進行實作。如需有關 Isotonic Regression 演算法的詳細資料，請參閱https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html。

XGBoost© 是梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost-AS 節點僅顯示了核心功能和一般參數。XGBoost-AS 節點是以 Spark 來實作的。

K-Means 是其中一個最常用的叢集演算法。它將資料點叢集化至預先定義的叢集數。SPSS Modeler 中的 K-Means-AS 節點是在 Spark 中進行實作。如需 K-Means 演算法的詳細資料，請參閱 https://spark.apache.org/docs/2.2.0/ml-clustering.html。請注意，K-Means-AS 節點自動針對類別變數執行 one-hot 編碼。

多層感知器是基於前饋神經網的分類器，它包含多個層。每一層完全連接到神經網中的下一層。SPSS Modeler 中的 MultiLayerPerceptron-AS 節點是在 Spark 中進行實作。如需多層感知器分類器 (MLPC) 的詳細資料，請參閱 https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier。

Isotonic-AS 節點

Isotonic 迴歸演算法屬於迴歸演算法系列。SPSS Modeler 中的 Isotonic-AS 節點是在 Spark 中進行實作。

如需有關 Isotonic Regression 演算法的詳細資料，請參閱https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html。[1]

[1] "Regression - RDD-based API." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

Isotonic-AS 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

© Copyright IBM Corp. 1994, 2019

欄位。 列出資料來源中的所有欄位。使用方向鈕可以將項目從此清單中手動分配給畫面右側的「目標」、「輸入」和「加權」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。 選取要用作目標的欄位。

輸入。選取一個或多個輸入欄位。

加權。 選取指數加權的加權欄位。如果未設定，則將使用預設加權值 **1**。

Isotonic-AS 節點建置選項

使用「建置選項」標籤可以指定 Isotonic-AS 節點的建置選項，其中包括特性索引及 Isotonic 類型。如需相關資訊，請參閱http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.html。[1]

輸入欄位索引。指定輸入欄位的索引。預設值是 **0**。

Isotonic 類型。此設定會判定輸出順序應該是保序/遞增還是反序/遞減。預設值是 **Isotonic**。

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

Isotonic-AS 模型塊

Isotonic-AS 模型塊包含 isotonic 迴歸模型所擷取的所有資訊。有下列區段可用。

模型摘要

這個視圖提供有關模型的關鍵資訊，其中包括輸入欄位、目標欄位和模型建置選項。

模型圖表

此視圖顯示散佈圖表。

XGBoost-AS 節點

XGBoost© 是梯度提升演算法的進階實現。Boosting 演算法會反覆學習弱分類器，然後將其新增至最終的強分類器。 XGBoost 具有很高的靈活性，並提供了很多對於大多數使用者來說過於複雜的參數，因此 SPSS Modeler 中的 XGBoost-AS 節點僅顯示了核心功能和一般參數。XGBoost-AS 節點是以 Spark 來實作的。

有關提升演算法的進一步資訊，請參閱 XGBoost 指導教學，網址為 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。[1]

請注意，SPSS Modeler 中不支援 XGBoost 交叉驗證功能。您可以將 SPSS Modeler 分割區節點用於此功能。另外，請注意，XGBoost 在 SPSS Modeler 中用於自動對種類變數執行 One-Hot 編碼。

註：在 Mac 上，建置 XGBoost-AS 模型需要 10.12.3 版或更高版本。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost-AS 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

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。 要手動分配目標和預測值，請選取此選項。

欄位。 使用方向鈕可以將項目從清單中手動分配給畫面右側的「目標」和「預測值角色」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。 選取要用作預測的目標的欄位。

預測值。選取一或多個欄位作為預測的輸入。

XGBoost-AS 節點建置選項

使用「建置選項」標籤可以指定 XGBoost-AS 節點的建置選項，包括用於模型建置和處理不平衡資料集的**一般選項**、用於目標和評估度量值的**學習作業選項**以及用於特定 booster 的 **booster 參數**。如需這些選項的相關資訊，請參閱下列線上資源：

- XGBoost 首頁[1]
- XGBoost Parameter Reference[2]
- XGBoost Spark API[3]

一般

工作程式數目。用於訓練 XGBoost 模型的工作程式數目。

執行緒數目。每個工作程式使用的執行緒數目。

使用外部記憶體。是否將外部記憶體用作快取。

Booster 類型。要使用的 booster（**gbtree**、**gblinear** 或 **dart**）。

Booster 循環數目。boosting 的循環數目。

比例 POS 加權。此設定用於控制正加權和負加權的平衡，對於不平衡的類別很有用。

隨機種子。按一下**產生**以產生亂數產生器所使用的種子。

學習作業

目標。從下列學習作業目標類型中選取：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**rank:pairwise**、**binary:logistic** 或 **multi**。

評估度量值。驗證資料的評估度量值。將根據目標（針對迴歸的 **rmse**，分類的 **error** 或分級的 **mean average precision**）指派預設度量值。可用的選項為 **rmse**、**mae**、**logloss**、**error**、**merror**、**mlogloss**、**uac**、**ndcg**、**map** 或 **gamma-deviance**（預設值為 **rmse**）。

Booster 參數

Lambda。 這是有關加權的 L2 正規化項目。增大此值將使模型更保守。

Alpha 值。這是有關加權的 L1 正規化項目。增大此值將使模型更保守。

Lambda (λ) 偏差。這是有關基本選項目的 L2 正規化項目。（沒有關於偏移的 L1 正規化項目，因為它不重要。）

樹狀結構方法。選取要使用的 XGBoost Tree 構建演算法。

深度上限。指定樹狀結構的最大深度。增大此值將導致模型更複合，並且很可能出現過度配適。

最小子項加權。指定子代中需要的實例加權 (hessian) 的下限總和。如果樹狀結構分區步驟生成實例加權總和少於此**最小子項加權**的葉節點，那麼建置程序將停止於進行進一步分區。在線性迴歸模式下，此項簡單地對應於每個節點中所需的下限實例數。加權越大，演算法越保守。

最大差異步驟。指定容許用於每個樹狀結構的加權估計的最大差異步驟。如果設定為 **0**，那麼沒有限制。如果設定為正值，那麼它可以使更新步驟更為保守。通常不需要此參數，但是在某個類別極度不已平衡的情況下，它可以用在邏輯迴歸中。

子樣本。子樣本是訓練實例的比例。例如，如果您將此項設定為 **0.5**，那麼 XGBoost 將隨機收集一半的資料實例以生成樹狀結構，並且這將防止過度配適。

Eta。這是更新步驟期間用於防止過度配適的步驟大小收縮。在每個提升步驟後，可以直接獲取新功能的加權。Eta 也會縮小功能加權，以使提升過程更保守。

伽瑪參數。 這是對樹狀結構的某個葉節點進行進一步分割區所需的下限損失減小。Gamma 設定越大，演算法越保守。

Colsample（依樹狀結構）。這是構建每個樹狀結構時欄的子樣本比例。

Colsample（依層次）。這是在每個層次每個分割的欄的子樣本比例。

常態化演算法。選取「一般」選項下的 dart booster 類型時要使用的常態化演算法。可用的選項為**樹狀結構**或**樹系**（預設值為**樹狀結構**）。

取樣演算法。選取「一般」選項下的 dart booster 類型時要使用的取樣演算法。**均勻**演算法均勻選取放置的樹狀結構。**加權**演算法以加權比例選取放置的樹狀結構。預設值為**均勻**。

捨棄比率。選取「一般」選項下的 dart booster 類型時要使用的捨棄比率。

跳過捨棄的機率。選取「一般」選項下的 dart booster 類型時要使用的跳過捨棄機率。如果跳過捨棄，則將以 **gbtree** 的相同方式新增樹狀結構。

下表顯示了 SPSS Modeler XGBoost-AS 節點對話框中的設定與 XGBoost Spark 參數之間的關係。

表 *42.* 對映至 *Spark* 參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	XGBoost Spark 參數
目標	target_fields	
預測	input_fields	
Lambda (λ) 值	Lambda (λ) 值	Lambda (λ) 值
工作程式數目	nWorkers	nWorkers
執行緒數目	numThreadPerTask	numThreadPerTask
使用外部記憶體	useExternalMemory	useExternalMemory
Booster 類型	boosterType	boosterType
Boosting 循環數目	numBoostRound	round

表 42. 對映至 *Spark* 參數的節點內容 *(*繼續*)*

SPSS Modeler 設定	Script 名稱（內容名稱）	XGBoost Spark 參數
比例 POS 加權	scalePosWeight	scalePosWeight
目標	objectiveType	objective
評估度量值	evalMetric	evalMetric
Lambda (λ) 值	Lambda (λ) 值	Lambda (λ) 值
Alpha	alpha	alpha
Lambda (λ) 偏差	lambdaBias	lambdaBias
樹狀結構方法	treeMethod	treeMethod
深度上限	maxDepth	maxDepth
最小子項加權	minChildWeight	minChildWeight
最大差異步驟	maxDeltaStep	maxDeltaStep
子樣本	sampleSize	sampleSize
Eta	Eta 值	Eta 值
Gamma	伽瑪分配	伽瑪分配
Colsample（依樹狀結構）	colsSampleRation	colSampleByTree
Colsample（依層次）	colsSampleLevel	colsSampleLevel
常態化演算法	normalizeType	normalizeType
取樣演算法	sampleType	sampleType
捨棄比率	rateDrop	rateDrop
跳過捨棄的機率	skipDrop	skipDrop

[1] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

[2] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "ml.dmlc.xgboost4j.scala.spark Params." *DMLC for Scalable and Reliable Machine Learning*. Web. 3 Oct 2017.

XGBoost-AS 節點模型選項

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。

K-Means-AS 節點

K-Means 是其中一個最常用的叢集演算法。它將資料點叢集化至預先定義的叢集數。[1] SPSS Modeler 中的 K-Means-AS 節點是在 Spark 中實作。

如需 K-Means 演算法的詳細資料，請參閱https://spark.apache.org/docs/2.2.0/ml-clustering.html。

請注意，K-Means-AS 節點自動針對類別變數執行 one-hot 編碼。

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-Means-AS 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

使用預先定義的角色。 此選項會告知節點使用來自上游「類型」節點的欄位資訊。依預設選取該項。

使用自訂欄位指派。 如果您想要手動指派輸入欄位，請選取此選項，然後選取一或多個輸入欄位。使用此選項與在「類型」節點中將欄位角色設為**輸入**類似。

K-Means-AS 節點建置選項

使用「建置選項」標籤可以指定 K-Means-AS 節點的建置選項，包括用於模型建置的一般選項、用於起始設定叢集中心的起始設定選項，以及用於計算反覆運算及隨機種子的進階選項。如需相關資訊，請參閱 SparkML 上 K-Means 的 JavaDoc。[1]

一般

模型名稱。對特定叢集評分之後產生的欄位名稱。選取**自動**（預設值），或者選取**自訂**，然後鍵入一個名稱。

叢集數目。指定要產生的叢集數目。預設值為 **5**，且最小值為 **2**。

起始設定

起始設定模式。指定起始設定叢集中心的方法。**K-Means\|\|** 是預設值。如需有關這兩個方法的詳細資料，請參閱可調式 K-Means++。[2]

起始設定步驟。如果已選取 **K-Means\|\|** 起始設定模式，請指定起始設定步驟數。**2** 是預設值。

進階

進階設定。如果您要如下所示設定進階選項，請選取此選項。

最大疊代。指定搜尋叢集中心時執行的疊代數目上限。**20** 是預設值。

容錯。指定疊代演算法的聚合容錯。**1.0E-4** 是預設值。

設定隨機種子。選取此資訊並按一下**產生**可以產生由亂數字產生器使用的種子。

顯示

顯示圖形。如果您想要在輸出中包括圖形，請選取此選項。

下表顯示了 SPSS Modeler K-Means-AS 節點中的設定與 K-Means Spark 參數之間的關係。

表 *43.* 對映至 *Spark* 參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	K-Means SparkML 參數
輸入欄位	features	
叢集數目	clustersNum	k
起始設定模式	initMode	initMode
起始設定步驟	initSteps	initSteps
最大疊代	maxIter	maxIter
容錯	toleration	tol
隨機種子	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

MultiLayerPerceptron-AS 節點

多層感知器是基於前饋神經網的分類器，它包含多個層。每一層完全連接到神經網中的下一層。如需多層感知器分類器 (MLPC) 的詳細資料，請參閱 https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier。[1]

SPSS Modeler 中的 MultiLayerPerceptron-AS 節點是在 Spark 中進行實作。若要使用此節點，您必須設定上游「類型」節點。MultiLayerPerceptron-AS 節點將從「類型」節點（或上游來源節點的「類型」標籤）讀取輸入值。

[1] "Multilayer perceptron classifier." *Apache Spark*. MLlib: Main Guide. Web. 5 Oct 2018.

MultiLayerPerceptron-AS 節點欄位

「欄位」標籤指定要在分析中使用的欄位。

使用預先定義的角色。 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。

使用自訂欄位指派。 要手動分配目標和預測值，請選取此選項。

目標。 選取要用作預測的目標的欄位。

預測值。選取一或多個欄位以使用作為預測的輸入。

MultiLayerPerceptron-AS 節點建置選項

使用「建置選項」標籤可以指定 MultiLayerPerceptron-AS 節點的建置選項，其中包括效能、建模建置及專家級選項。如需這些選項的相關資訊，請參閱 http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/classification/MultilayerPerceptronClassifier.html。[1]

效能

感知器層。使用此設定來定義要包括的感知器層數目。此值必須大於感知器欄位的數目。預設值為 **1**。

隱藏層。指定隱藏層的數目。在多個隱藏層之間使用逗點。預設值為 **1**。

輸出層。指定輸出層的數目。預設值為 **1**。

隨機種子。如果您要產生亂數產生器所使用的種子，請按一下**產生**。

模型建置

最大疊代。指定要執行的疊代次數上限。預設值為 **10**。

僅限專家

區塊大小。如果您要指定矩陣中堆疊輸入資料的區塊大小，請在「模型建置」部分中選取**專家模式**選項。這樣可加快計算速度。預設區塊大小為 **128**。

下表顯示了 SPSS Modeler MultiLayerPerceptron-AS 節點對話框中的設定與 Spark KDE 程式庫參數之間的關係。

表 *44.* 對映至 *Spark* 參數的節點內容

SPSS Modeler 設定	Script 名稱（內容名稱）	Spark 參數
預測	features	
目標	label	
感知器層	layers[0]	layers[0]
隱藏層	layers[1...<latest-1>]	layers[1...<latest-1>]
輸出層	layers[<latest>]	layers[<latest>]
隨機種子	seed	seed
最大疊代數	maxiter	maxiter

[1] "Class MultilayerPerceptronClassifier." *Apache Spark*. JavaDoc. Web. 5 Oct 2018.

MultiLayerPerceptron 節點模型選項

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。

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 可能提供此材料的其他語言版本。不過，您可能需要擁有該語言的產品副本或產品版本，才能對它進行存取。

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 CorporationNorth Castle Drive, MD-NC119
Armonk, NY 10504-1785US

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）。有些轄區不允許放棄在特定交易中的明示或默示保證，因此，這項聲明對您可能不適用。

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。

本書對於非 IBM 網站的援引只是為了方便而提供，並不對這些網站作任何認可。該些網站上的內容並非本 IBM 產品內容的一部分，用戶使用該網站時應自行承擔風險。

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 CorporationNorth Castle Drive, MD-NC119
Armonk, NY 10504-1785US

這些資訊可能可以使用，但必須遵循適當的條款，在某些情況中需要付費。

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。

所引用的客戶範例為說明用途。實際的績效會因不同的配置與作業狀況而異。

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。有關非 IBM 產品的性能問題應直接洽詢該產品供應商。

IBM 不須通知即可變更或撤銷與 IBM 未來方向或目的相關的陳述，亦僅代表其目標及方針。

本資訊中含有日常商業活動所用的資料及報告範例。為了盡可能完整地說明，範例中包括了個人、公司行號、品牌以及產品等的名稱。 所有這些名稱都是虛構的，實際個人或商業企業的任何類似項目都純屬巧合。

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 商標最新清單可於下列網站之「著作權與商標資訊」("Copyright and trademark information") 網頁上取得，網址如下：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。

產品說明文件的條款

這些出版品的使用許可權係遵循下列條款而授與。

適用性

這些條款係附加於 IBM 網站的任何使用條款上。

個人使用

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。未經 IBM 明示同意，您不得散佈、展示或改作該等「出版品」或其任何部分。

商業使用

貴客戶只能在您的企業內重製、散布和顯示這些出版品，但必須保留所有的所有權聲明。未經 IBM 明示同意，您不得改作該等「出版品」，也不得於企業外複製、散佈或展示該等「出版品」或其任何部分。

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。

您不可下載、出口或再出口本資訊，除非完全遵守所有適用的法律及規定，包括所有的美國出口法律及規定。

IBM 對於該等出版品之內容不為任何保證。這些出版品係「依現狀」提供，無任何形式（明示或暗示）的擔保，包括但不限於對適售性、無侵權、符合特定使用目的的暗示保證。

詞彙

A

AICC ： 用於根據 -2（受限）對數概似值來選取及比較混合模型的量數。數值越小代表模式越佳。AICC 會「修正」較小 AIC 的樣本大小。當樣本大小增加時，AICC 會收斂至 AIC。

B

Bayesian 資訊準則 *(BIC)* ： 用於根據 -2 對數概似值來選取及比較模型的量數。數值越小代表模式越佳。BIC 也會「懲罰」過度參數化模型（例如，包含大量輸入的複雜模型），但比 AIC 更嚴格。

Box's M 測試 ： 對群組共變異數矩陣相等性進行的測試。若樣本夠大，則非顯著 p 值代表沒有充足證據顯示矩陣有所不同。此測試對多變量常態的偏差很敏感。

C

個案 ： 顯示每個觀察值之實際群組、預測群組、事後機率及區別分數的代碼。

分類結果 ： 根據區別分析，正確及錯誤地指派給每個群組的觀察值數目。有時亦稱為「混淆矩陣」。

組合的群組圖 *(Combined-Groups Plots)* ： 建立前兩個區別函數值的所有群組散佈圖。如果僅有一個函數，則會改為顯示直方圖。

共變異數 *(Covariance)* ： 兩個變數之間關聯的非標準化量數，它等於交叉product離差除以 N-1。

F

Fisher's 線性區別函數係數*(F)* ： 顯示可直接供分類使用的 Fisher 分類函數係數。將會針對每個群組取得個別的分類函數係數集，並將觀察值指派給擁有最高區別分數（分類函數值）的群組。

H

風險圖 *(Hazard Plot)* ： 可在線性尺度上顯示累積風險函數。

K

峰度 ： 偏離程度的測量。對常態分配而言，峰度統計量數值為零。正峰度值指出資料的極端偏離值超出常態分佈。負峰度值指出資料的極端偏離值小於常態分佈。

A

留一分類 *(Leave-one-out Classification)* ： 分析中的每個觀察值皆由該觀察值除外之所有觀察值的衍生函數來分類。其亦稱作「U 方法」。

M

MAE ： 平均絕對誤差。測量數列與模型預測層級之間的變異程度。在原始數列單元中會報告 MAE。

Mahalanobis 距離 *(Mahalanobis Distance)* ： 測量自變數之觀察值的值與整體觀察值平均數的變異程度。較大 Mahalanobis 距離會將觀察值識別為在一個以上自變數中具有極端值。

MAPE ： 平均絕對百分比誤差。測量相依數列與其模型預測層級之間的變異程度。其與使用的單元無關，因此可用來比較具有不同單元的數列。

MaxAE ： 最大絕對誤差。最大預測誤差，其會以作為相依系列的相同單元來表示。其與 MaxAPE 類似，對於想像您預測值的最糟情況非常有用。最大絕對誤差與最大絕對百分比誤差可能發生於不同的數列點--例如當大型序列值之絕對誤差僅略微大於小型序列值之絕對誤差時。在此情況下，最大絕對誤差會發生於較大的數列值，而最大絕對百分比誤差則會發生於較小的數列值。

MaxAPE ： 平均絕對百分比誤差。其為以百分比表示的最大預測誤差。此量數對於想像您預測值的最糟情境非常有用。

將最小 *F* 比例輸入方法最大化 *(Maximizing the Smallest F Ratio Method of Entry)* ： 一種逐步分析的功能選擇方法，其所根據的作法，是將從群組間 Mahalanobis 距離計算而來的 F 比例最大化。

最大值 ： 數值變數的最大值。

平均數*(M)* ： 集中趨勢的測量。算術平均數，總和除以觀察值數。

平均數 ： 顯示總和與群組平均數，以及自變數的標準差。

中位數 ： 半數觀察值落點上下的值，第 50 個百分位數。如果觀察值個數是偶數的話，中位數是中間兩個觀察值的平均值，此處的兩個中間是指，當觀察值按照遞增或遞減順序排列時，位於最中間的兩個值。中位數為集中趨勢的量數，其不會感應到偏離值（相反地，平均數會受到幾個高低極端值所影響）。

最小化 *Wilks Lambda (Minimize Wilks' Lambda)* ： 一種進行逐步迴歸分析區別分析的功能選擇方法，它會根據變數將 Wilks Lambda 降低的程度來為輸入方程式的項目選擇變數。每一個步驟都要輸入將整體 Wilks Lambda 最小化的變數。

最小值 ： 數值變數的最小值。

眾數*(O)* ： 最常出現的值。如果幾個值的最大出現次數相同，則每一個都是眾數。

N

正規化 *BIC* ： 正規化 Bayesian 資訊準則。它是一種模式整體適合度的一般測量，嘗試說明模型的複雜性。它是依據平均平方誤得到的分數，且包含對模型參數個數與數列長度設下的懲罰。該懲罰會移除參數過多的模型優點，讓統計量較能輕易地比較同一數列的不同模型。

O

壹減存活機率 *(One Minus Survival)* ： 繪製線性尺度上被 1 減後的存活機率函數。

R

範圍 ： 值變數最大與最小值之間的差異，也就是最大值減去最小值。

Rao's V（區別分析）*(Rao's V (Discriminant Analysis))* ： 群組平均數之間差異的量數。也稱為 Lawley-Hotelling 跡。每一個步驟，都要輸入將 Rao's V 中的增加最大化的變數。選取此選項後，輸入變數必須輸入分析的最小值。

RMSE ： 均方根誤差。平均均方根誤差。是應變數列與其模型預測層次差異程度的量數，用與應變數列相同的單位來表示。

R 平方 *(R-Squared)* ： 線性模型的適合度量數，有時也稱為判斷係數。由迴歸模型所解釋的因變數變異的比例。它的範圍值介於 0 到 1 之間。值越小，表示模型越不適合資料。

S

各組 *(Separate-Groups)* ： 各組共變異數矩陣在分類時使用。因為分類是以區別函數為依據（而不是以原始變數為依據），因此此選項並不是一律相當於二次區別。

各組共變異數 *(Separate-Groups Covariance)* ： 顯示各組不同的共變異數矩陣。

各組的圖 *(Separate-Groups Plots)* ： 建立前兩個區別函數值的各組散佈圖。在只有一個函數的情況下，則會改為顯示直方圖。

循序 *Bonferroni* 法*(Q)* ： 這是循序逐步拒絕的 Bonferroni 程序；就拒絕個別假設而言，此程序做法相當不保守，但整體顯著性層次仍維持相同。

循序 *Sidak* 法*(S)* ： 這是循序逐步拒絕的 Sidak 程序；就拒絕個別假設而言，此程序做法相當不保守，但整體顯著性層次仍維持相同。

偏斜度 ： 分配不對稱性的量數。常態分佈是一種對稱性分佈，其偏斜度值為 0。具有顯著性正偏斜度的分佈右側尾部較長。有顯著負偏斜度值的分配會有一個長的偏左尾部。偏斜度值有如指標，若大於它的兩倍標準誤，則表示背離對稱。

標準差 ： 在平均數四周離散的量數，等於變異數的平方根。標準差所用的測量單位與原始觀察值相同。

標準差 ： 測量平均數四周的離散情形。在常態分配中，68% 的觀察值會落在平均數的一個標準差內，95% 的觀察值會落在兩個標準差內。例如，如果平均年齡為 45 歲，標準差是 10 的話，在常態分配中 95% 的觀察值會介於 25 到 65 歲之間。

標準誤 ： 不同樣本間測試統計量值的變化大小量數。這是一項統計量取樣分配的標準差。例如，平均數的標準誤就是樣本平均數的標準差。

峰度的標準誤 *(Standard Error of Kurtosis)* ： 峰度對其標準誤的比例可用於測試常態性（也就是說，如果比例小於 -2 或大於 +2，則您可以拒絕常態性）。峰度若為大的正值，表示該分配的尾端比常態分配的尾端更長；峰度若為負值，表示尾端較短（變成類似箱型均勻分配的尾端）。

平均數的標準誤 *(Standard Error of Mean)* ： 測量從同一個分配取出來的不同樣本間平均數的變化大小。它可以用來大略地將觀察平均數與假設值相比較（也就是如果標準誤與差異的比值小於 -2 大於 +2 的話，您就可以下結論說兩個值不同）。

偏斜度的標準誤 *(Standard Error of Skewness)* ： 偏斜度對其標準誤的比例可用於測試常態性（也就是說，如果比例小於 -2 或大於 +2，則您可以拒絕常態性）。偏斜度若為大的正值，表示有長的偏右尾端：極端負值則表示有長的偏左尾端。

恆定 *R* 平方 *(Stationary R-squared)* ： 將模型中恆定的部分比作簡式平均數模型的測量法。有趨勢或週期性樣式時，此測量法比普通 R 平方更好。恆定 R 平方可以是從負無限大到 1 範圍內的負數。負值表示考量的模型比基準線模型更差。正值表示考量的模型比基準線模型更佳。

總和*(U)* ： 遍及所有包含遺漏值之觀察值的值總和或總數。

存活圖 *(Survival Plot)* ： 可在線性尺度上顯示累積存活函數。

T

地域圖 *(Territorial Map)* ： 以函數值為基礎，用於將觀察值分類至群組的邊界圖。將觀察值分類至對應之組別的數目。每個組別的平均數是用其邊界中的星號來表示。如果只有一個區別函數，不會顯示地圖。

共變異數總計 *(Total Covariance)* ： 顯示所有觀察值中的共變異數矩陣，如同它們來自單一樣本一樣。

U

未說明的變異數 *(Unexplained Variance)* ： 在每個步驟中，輸入會將群組之間未說明的變異總和最小化的變數。

唯一 *(Unique)* ： 針對任何類型的所有其他效果，調整每個效果來同時評估所有效果。

單變量 *ANOVA(A)* ： 針對每個自變數的群組平均數之相等性執行單向變異數分析測試。

未標準化*(U)* ： 顯示未標準化的區別函數係數。

使用 *F* 值 *(Use F Value)* ： 變數的 F 值大於「輸入」值時，系統會將該變數輸入模型，而當 F 值小於「移除」值時，系統會將變數移除。「輸入」必須大於「移除」，而且兩個數值都必須是正數。若要將更多變數輸入模式，請調低「輸入」值。若要從模型中移除更多變數，請調高「移除」值。

使用 *F* 機率值 *(Use Probability of F)* ： 變數的 F 值顯著性層次小於「輸入」值時，系統會將變數輸入模型，而顯著性層次大於「移除」值時，系統會將變數移除。「輸入」必須小於「移除」，而且兩個數值都必須是正數。若要將更多變數輸入模式，請調高「輸入」值。若要從模型中移除更多變數，請調低「移除」值。

V

有效 ： 沒有系統遺漏值，也沒有定義為使用者遺漏值的有效觀察值。

變異數 ： 測量平均數四周的離散情形，它等於平均數的平方離差總和除以觀察值數目減一。變異數的測量單位是變數本身的平方。

W

群組內 *(Within-Groups)* ： 合併的群組內共變異數矩陣用於分類觀察值。

>群組內相關性 *(Within-Groups Correlation)* ： 顯示合併的群組內相關性矩陣，此矩陣是將所有群組的個別共變異數矩陣平均，再計算相關性而來。

群組內共變異數 *(Within-Groups Covariance)* ： 顯示合併的群組內共變異數矩陣，此矩陣可能與共變異數矩陣總計不同。透過將所有群組的個別共變異數矩陣平均來取得矩陣。

索引

索引順序以中文字，英文字，及特殊符號之次序排列。

〔一劃〕

〔二劃〕

〔三劃〕

〔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〕

〔二十二劃〕

〔二十三劃〕

A

B

C

D

E

F

G

H

I

K

L

M

N

P

Q

R

S

T

W

X

Printed in Taiwan