IBM SPSS Modeler 18.2.1 應用程式手冊

IBM

附註

使用本資訊及其所支援的產品之前，請先閱讀第 349 頁的『注意事項』中的資訊。

產品資訊

本版適用於 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

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專門知識快速建立預測性模型，並將其部署於企業運作，從而改進決策過程。IBM SPSS Modeler 參照線業標準 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、人工智慧以及統計量的各種建模方法。「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 是功能完整的產品版本，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。通過使用唯一視覺化介面，您可以輕鬆地視覺化資料採礦程序。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 使用主從式架構將資源密集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，在分散式分析模式下，該產品連同一個以上的 IBM SPSS Modeler 安裝在伺服器主機上持續執行。這種運行方式大大提高了 SPSS Modeler Server 對大型資料集的處理速度，因為在伺服器上可以運行耗用記憶體的作業，並且無需將資料下載到用戶端電腦上。IBM SPSS Modeler Server 還提供對 SQL 最佳化和資料庫內建模功能的支援，從而在效能和自動化方面帶來更多優勢。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是一個圖形使用者介面，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

雖然資料採礦通常是互動式程序，但也可以從指令行執行 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，或者從小型用戶端應用程式 IBM SPSS Modeler Advantage 進行存取。請將配接器安裝在管理儲存庫的系統上。

IBM SPSS Modeler 版本

SPSS Modeler 推出了下列版本。

SPSS Modeler Professional

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

SPSS Modeler Premium

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

IBM SPSS Modeler Text Analytics 採用了先進語言技術和自然語言處理 (NLP)，以快速處理大量非結構化文字資料，擷取和組織關鍵概念，以及將這些概念分組成各式各樣的種類。擷取的概念和種類可以和現有結構化資料（例如個人背景資訊）進行結合，並且可套用於使用 IBM SPSS Modeler 資料採礦工具完整套組來進行的建模，以作出更好更集中的決策。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription 會提供與傳統 IBM SPSS Modeler 用戶端完全相同的預測分析功能。使用 Subscription 版，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，包括如何建立資料串流、處理遺漏值、建立 CLEM 表示式、處理專案和報告，以及包裝串流以部署至 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。實際上這表示除建模節點以外的所有節點。
- **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、人工智慧以及統計量的各種建模方法。
- **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 開發者手冊。** 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 資料庫內採礦手冊。**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採礦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* 資料庫內採礦手冊 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

與應用程式範例搭配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 產品概觀

新手啟動使用

作為一種資料採礦應用程式，IBM SPSS Modeler 提供了用以尋找大資料集中有用關係的策略性方法。與更傳統的統計方法相比，您在開始時不必知道您要尋找什麼。您可以通過配適不同的模型和研究不同的關係來探索您的資料，直到發現有用的資訊。

啟動 IBM SPSS Modeler

若要啟動此應用程式，請按一下：

開始 > **[所有] 程式** > **IBM SPSS Modeler <版本>** > **IBM SPSS Modeler <版本>**

主視窗將在幾秒鐘後顯示。

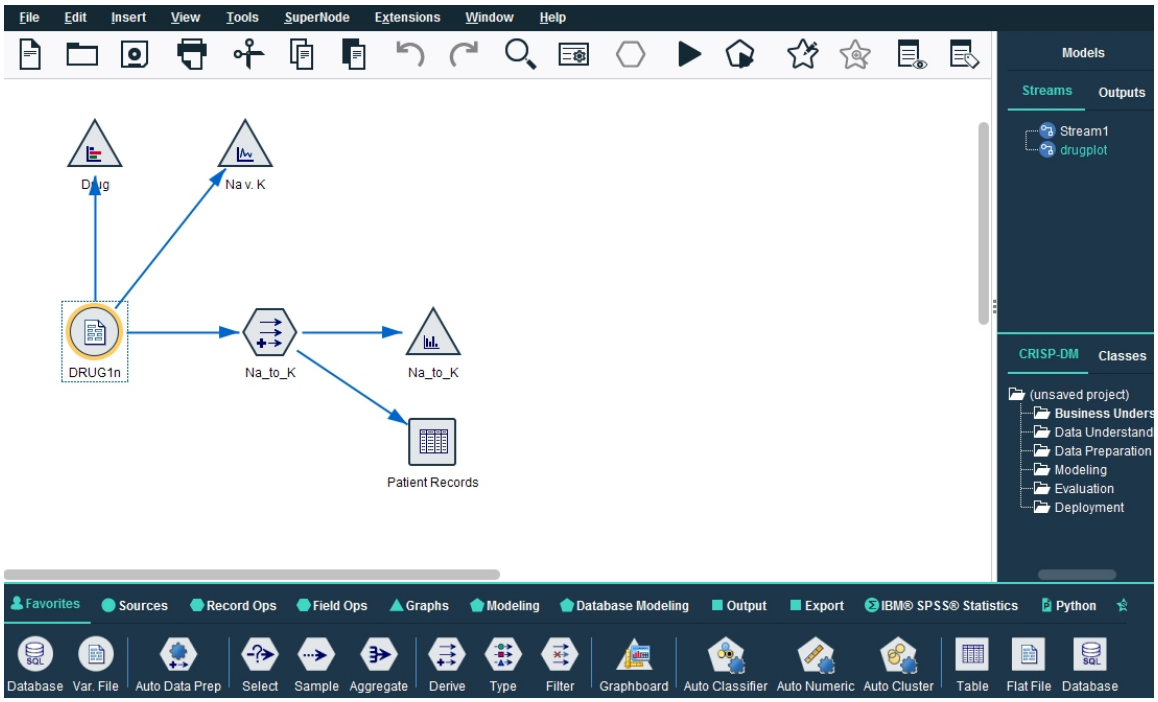


圖 *1. IBM SPSS Modeler* 主應用程式視窗

從指令行啟動

您可以使用作業系統的指令行來如下啟動 IBM SPSS Modeler：

1. 在安裝了 IBM SPSS Modeler 的電腦上，開啟 DOS 或命令提示字元視窗。
2. 要以互動模式啟動 IBM SPSS Modeler 介面，請輸入 `modelerclient` 指令，然後輸入所需的引數；例如：

```
modelerclient -stream report.str -execute
```

可用引數（旗標）容許您連接至伺服器、載入串流、執行 Script 或根據需要指定其他參數。

連線到 IBM SPSS Modeler Server

IBM SPSS Modeler 可作為單獨的應用程式執行，或作為直接連接至 IBM SPSS Modeler Server 的用戶端執行，或者作為通過處理程序協調器 (COP) 外掛程式從 IBM SPSS Collaboration and Deployment Services 連接至 IBM SPSS Modeler Server 或伺服器叢集的用戶端執行。現行連線狀態顯示在 IBM SPSS Modeler 視窗的左下角。

無論何時想連接至伺服器，都請手動輸入想要連接的伺服器的名稱或選取之前已定義的名稱。但是，如果您擁有 IBM SPSS Collaboration and Deployment Services，則可以從「伺服器登入」對話框搜尋伺服器清單或伺服器叢集清單。可以通過處理程序協調器執行瀏覽網路上執行的 Statistics 服務的功能。

連接至伺服器

1. 在「工具」功能表上，按一下**伺服器登入**。會開啟「伺服器登入」對話框。或者按兩下 IBM SPSS Modeler 視窗的連線狀態區域。
2. 使用該對話框指定要連接至本機伺服器電腦的選項或從表格中選取連線。
   - 按一下**新增**或**編輯**以新增或編輯連線。請參閱第 7 頁的『新增並編輯 IBM SPSS Modeler Server 連線』主題，以取得更多資訊。
   - 按一下**搜尋**以存取處理程序協調器中的伺服器或伺服器叢集。請參閱第 7 頁的『搜尋 IBM SPSS Collaboration and Deployment Services 中的伺服器』主題，以取得更多資訊。

   伺服器表格。此表格包含已定義的伺服器連線集。該表格顯示預設連線、伺服器名稱、說明和埠號。您可以手動新增新的連線，以及選取或搜尋現有連線。要將特定的伺服器設定為預設連線，請在表格中「預設」欄中為此連線選取勾選框。

   預設資料路徑。指定用於伺服器電腦上的資料的路徑。按一下省略號按鈕（**...**），以瀏覽至所需要的位置。

   設定認證。 不勾選此方框可啟用**單一登入**功能，該功能試圖使您使用本端電腦使用者名稱和密碼詳細資料登入伺服器。如果無法使用單一登入，或您勾選此方框以停用單一登入（例如，登入管理者帳戶），則啟用下列欄位讓您輸入您的認證。

   使用者 ID。輸入用於登入伺服器的使用者名稱。

   密碼。 輸入與指定使用者名稱相關聯的密碼。

   網域。 指定用於登入伺服器的網域。只有伺服器電腦與用戶端電腦處於不同的 Windows 網域時，才需要網域名稱。
3. 按一下**確定**以完成此連線。

切斷與伺服器的連接

1. 在「工具」功能表上，按一下**伺服器登入**。會開啟「伺服器登入」對話框。或者按兩下 IBM SPSS Modeler 視窗的連線狀態區域。
2. 在此對話框中，選取「本機伺服器」，然後按一下**確定**。

新增並編輯 IBM SPSS Modeler Server 連線

您可以在「伺服器登入」對話框中手動編輯或新增伺服器連線。按一下「新增」可以存取空白的「新增/編輯伺服器」對話框，在此對話框中可以輸入伺服器連線的詳細資料。在「伺服器登入」對話框中選取現有連線並按一下「編輯」，將開啟「新增/編輯伺服器」對話框，其中包含所選連線的詳細資料，以便可以進行任何變更。

註：不能編輯從 IBM SPSS Collaboration and Deployment Services 中新增的伺服器連線，因為名稱、埠及其他詳細資料已在 IBM SPSS Collaboration and Deployment Services 中做過定義。最佳實踐指出，應該使用相同的埠與 IBM SPSS Collaboration and Deployment Services 和 SPSS Modeler Client 進行通訊。這些埠可以設定為 options.cfg 檔案中的 `max_server_port` 和 `min_server_port`。

新增伺服器連線

1. 在「工具」功能表上，按一下**伺服器登入**。會開啟「伺服器登入」對話框。
2. 在此對話框中，按一下**新增**。即會開啟「伺服器登入新增/編輯伺服器」對話框。
3. 輸入伺服器連線的詳細資料，然後按一下**確定**儲存此連線並傳回「伺服器登入」對話框。

- **伺服器**。指定可用伺服器或從清單選取一個伺服器。伺服器電腦可以由英數名稱（例如 *myserver*）或指派給伺服器電腦的 IP 位址（例如，202.123.456.78）來識別。
- **埠**。 指定伺服器要接聽的埠號。如果預設值不可用，請向系統管理者索取正確的埠號。
- **說明**。 輸入此伺服器連線的選用說明。
- **確保連線安全(使用 SSL)**。指定是否應該使用 SSL (**Secure Sockets Layer**) 連線。SSL 是一個常用通訊協定，用於確保通過網路傳送的資料的安全。要使用此功能，必須在管理 IBM SPSS Modeler Server 的伺服器中啟用 SSL。必要時請與本端管理者聯絡，以瞭解詳細資料。

編輯伺服器連線

1. 在「工具」功能表上，按一下**伺服器登入**。會開啟「伺服器登入」對話框。
2. 在此對話框中，選取希望編輯的連線，然後按一下**編輯**。即會開啟「伺服器登入新增/編輯伺服器」對話框。
3. 變更伺服器連線詳細資料，然後按一下**確認**儲存變更內容並傳回至「伺服器登入」對話框。

搜尋 IBM SPSS Collaboration and Deployment Services 中的伺服器

在 IBM SPSS Collaboration and Deployment Services 中，可以使用處理程序協調器選取網路上可用的伺服器或伺服器叢集，從而代替手動輸入伺服器連線。伺服器叢集是一組伺服器，處理程序協調器從這群組伺服器中確定最適合對處理要求作出回應的伺服器。

雖然您可在「伺服器登入」對話框中手動新增伺服器，但搜尋可用伺服器能讓您不需要知道正確的伺服器名稱和埠號，即可連線至伺服器。此資訊會自動提供。但仍需輸入正確的登入資訊，如使用者名稱、網域和密碼。

附註：如果您無權存取處理程序協調器功能，那麼仍然可以手動輸入要連接的伺服器名稱或選取先前已定義的名稱。請參閱『新增並編輯 IBM SPSS Modeler Server 連線』主題，以取得更多資訊。

搜尋伺服器和伺服器叢集

1. 在「工具」功能表上，按一下**伺服器登入**。會開啟「伺服器登入」對話框。
2. 在此對話框中，按一下**搜尋**開啟「搜尋伺服器」對話框。如果在試圖瀏覽處理程序協調器時未登入到 IBM SPSS Collaboration and Deployment Services，則系統會提示您執行此項操作。
3. 從清單中選取伺服器或伺服器叢集。
4. 按一下**確定**以關閉對話框，然後將此連線新增到「伺服器登入」對話框的表格中。

連線到 Analytic Server

如果有多個 Analytic Server 可用，可以使用「Analytic Server 連線」對話框來定義多個伺服器以在 IBM SPSS Modeler 中使用。您的管理者可能已經在 <Modeler_install_path>/config/options.cfg 檔案中設定預設 Analytic Server。但在定義之後，也可以使用其他可用伺服器。例如，使用 Analytic Server「來源」及「匯出」節點時，您可能想要在串流的不同分支中使用不同的 Analytic Server 連線，因此每一個分支執行時，它會使用自己的 Analytic Server，且不會將任何資料拉取到 IBM SPSS Modeler Server。請注意，如果分支包含多個 Analytic Server 連線，則會從 Analytic Server 拉取到 IBM SPSS Modeler Server。如需相關資訊（包括限制），請參閱Analytic Server 串流內容。

要建立新的 Analytic Server 連線，轉至**工具** > **Analytic Server 連線**並在對話框的下列部分中提供所需資訊。

連線

URL。 以格式 https://hostname:port/contextroot 輸入 Analytic Server 的 URL，其中，hostname 是 Analytic Server 的 IP 位址或主機名稱，port 是其埠號，contextroot 是 Analytic Server 的環境定義根目錄。

租戶。 輸入 IBM SPSS Modeler Server 所屬的租戶的名稱。如果不知道租戶，請聯絡您的管理者。

鑑別

眾數。 從下列鑑別方式中進行選取。

- **使用者名稱和密碼**要求您輸入使用者名稱和密碼。
- **儲存的認證**要求您從 IBM SPSS Collaboration and Deployment Services 儲存庫 中選取認證。
- **Kerberos** 要求您輸入服務主體名稱和設定檔路徑。如果不知道該資訊，請聯絡您的管理者。

使用者名稱。輸入 Analytic Server 使用者名稱。

範圍。選取要用於 Analytic Server 連線的範圍。

密碼。 輸入 Analytic Server 密碼。

連接。按一下**連線**以測試新的連線。

連線

在指定上述資訊並按一下**連線**之後，連線將新增到該「連線」表格。如果需要移除連線，請將其選取並按一下**移除**。

如果管理者在 options.cfg 檔案中定義了預設 Analytic Server 連線，也可以按一下**新增預設連線**以將其新增到可用連線中。將提示您輸入使用者名稱和密碼。

變更 temp 目錄

IBM SPSS Modeler Server 執行的某些作業可能需要建立暫存檔。依預設，IBM SPSS Modeler 在系統暫存目錄下建立暫時檔案。可通過下列步驟更改暫存目錄的位置。

1. 建立新目錄 spss 及其子目錄 *servertemp*。
2. 編輯 *options.cfg*，該文件位於 IBM SPSS Modeler 安裝目錄的 */config* 目錄下。在此檔案中編輯 temp_directory 參數，將其更改為：temp_directory, "C:/spss/servertemp"。

3. 完成此操作後，必須重新啟動 IBM SPSS Modeler Server 服務。可通過按一下 Windows 控制台中的**服務**標籤進行此服務重啟操作。只需停止該服務然後將其重新啟動即可啟動所作的變更。重新啟動機器也會重新啟動該服務。

所有臨時檔案此時將寫入該新目錄。

註：

- 必須使用正斜線。
- 當通過 IBM SPSS Collaboration and Deployment Services 工作執行評估串流時，不適用 `temp_directory` 設定。當您執行這類工作時，會建立暫存檔。依預設，檔案會儲存至 IBM SPSS Modeler Server 的安裝目錄。當您在 IBM SPSS Modeler 中建立 IBM SPSS Modeler Server 連線時，可以變更儲存暫存檔的預設資料資料夾。

啟動多個 IBM SPSS Modeler 階段作業

如果需要同時啟動一個以上的 IBM SPSS Modeler 階段作業，則必須對 IBM SPSS Modeler 和 Windows 的設定做一些變更。例如，如果您有兩個獨立的伺服器授權，並且希望從同一台用戶端針對兩台不同的伺服器執行兩個串流，則需要對上述設置做一些更改。

要啟用多個 IBM SPSS Modeler 階段作業：

1. 按一下：

   開始 > **[所有] 程式** > **IBM SPSS Modeler**
2. 在 IBM SPSS Modeler 快速鍵（帶箭頭的圖示）上按一下滑鼠右鍵並選取**內容**。
3. 在**目標**文字框中，將 `-noshare` 新增到該字串的結尾。
4. 在 Windows 檔案總管中選取：

   工具 > **資料夾選項...**
5. 在「檔案類型」標籤上選取「IBM SPSS Modeler 串流」選項，然後按一下**進階**。
6. 在「編輯檔案類型」對話框中，選取「使用 IBM SPSS Modeler 開啟」，然後按一下**編輯**。
7. 在**用於執行動作的應用程式**文字框中，在 **-stream** 引數前新增 `-noshare`。

IBM SPSS Modeler 介面概覽

在資料採礦程序中的每個階段，易於使用的 IBM SPSS Modeler 介面都會邀請您的特定商業專門知識。建模演算法（如預測、分類、分區段和關聯偵測）可確保得到強大而準確的模型。模型結果可以方便地部署和讀取到資料庫、IBM SPSS Statistics 和各式各樣的其他應用程式中。

使用 IBM SPSS Modeler 是一個處理資料的三步驟程序。

- 首先，將資料讀入 IBM SPSS Modeler。
- 接著，通過一系列處理來執行資料。
- 最後，將資料傳送至目的地。

這一作業序列稱為**資料串流**，因為資料以逐條記錄形式流動，從來源開始，流經每個操作，最終到達目的地（模型或某種資料輸出）。

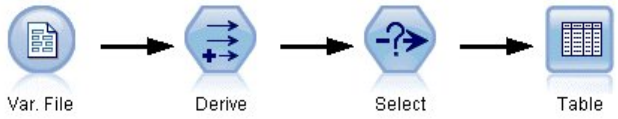


圖 2. 簡單串流

IBM SPSS Modeler 串流畫布

串流畫布是 IBM SPSS Modeler 視窗的最大區域，也是您建置和操作資料串流的位置。

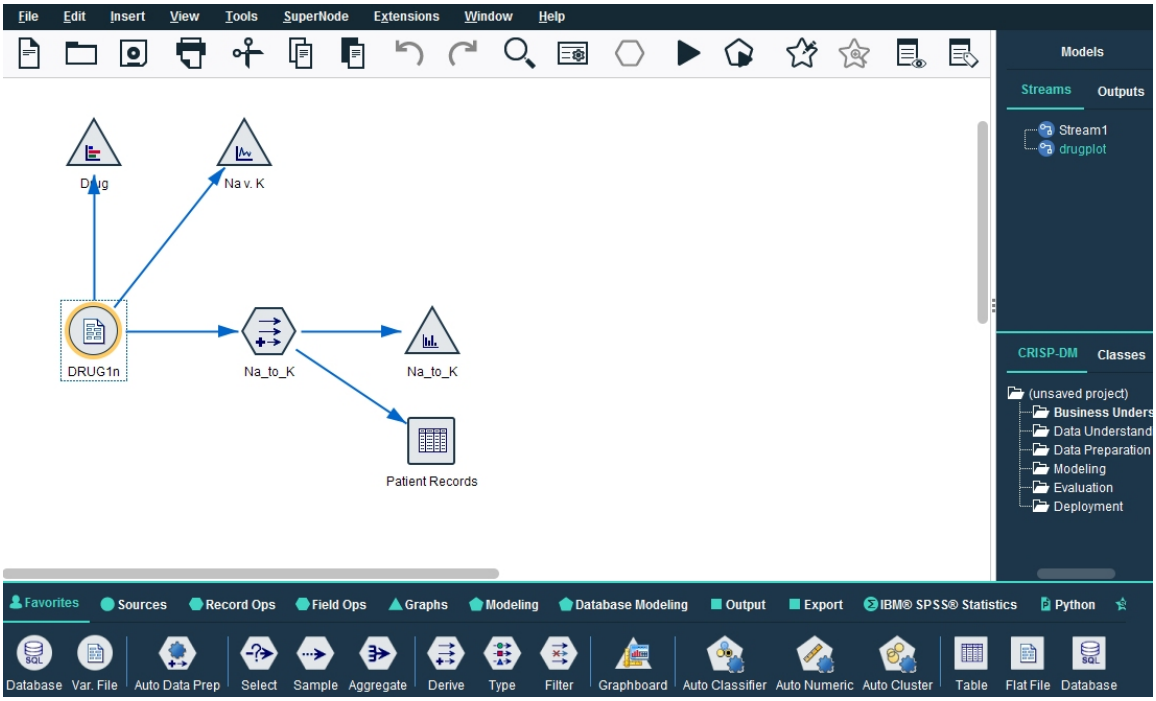


圖 3. *IBM SPSS Modeler* 工作區（預設視圖）

串流是在介面的主畫布中透過繪製與業務相關的資料作業圖來建立的。每個作業都用一個圖示或**節點**代表，這些節點在**串流**中鏈結在一起，串流代表資料在各個作業中的流動。

在 IBM SPSS Modeler 中，可以在同一串流畫布或透過開啟新的串流畫布來一次處理多個串流。階段作業期間，串流儲存在 IBM SPSS Modeler 視窗右上角的「串流」管理程式中。

註：如果使用啟用了內建觸控板的 **Force Click and haptic feedback** 設定的 MacBook，那麼從節點選用區將節點拖曳到串流畫布可能會導致向畫布新增複製節點。為避免發生此問題，建議停用 **Force Click and haptic feedback** 觸控板系統喜好設定。

節點選用區

SPSS Modeler 中的大部分資料和建模工具，都可從橫跨串流畫布下方視窗底部的節點選用區中獲取。

例如，**記錄作業**選用區標籤包含的節點，可用於對資料記錄執行作業（如選取、合併和附加）。

要向畫布新增節點，請按兩下節點選用區中的圖示或將節點拖至畫布上。隨後可連接各個圖示以建立一個代表資料流程的串流。

圖 *4.* 節點選用區中的「記錄作業」標籤

每個選用區標籤均包含一組不同的串流作業階段中使用的相關節點，如：

- **來源**節點用於將資料引入到 SPSS Modeler 中。
- **記錄作業**節點用於對資料記錄執行作業，例如選取、合併和附加。
- **欄位作業**節點用於對資料欄位執行作業，例如過濾、衍生新欄位和確定給定欄位的測量層次。
- **圖形**節點用於以圖形方式顯示建模前後的資料。圖形包含圖表、直方圖、Web 節點和評估表。
- **建模**節點使用 SPSS Modeler 中提供的建模演算法，例如神經網路、決策樹狀結構、叢集作業演算法和資料排序等。
- **資料庫建模**節點使用 Microsoft SQL Server、IBM Db2 以及 Oracle 和 Netezza 資料庫中提供的建模演算法。
- **輸出**節點用於為資料、圖表和模型結果生成可以在 SPSS Modeler 中檢視的各種輸出。
- **匯出**節點用於生成可以在外部應用程式（例如，IBM SPSS Data Collection 或 Excel）中檢視的各種輸出。
- **IBM SPSS Statistics** 節點從 IBM SPSS Statistics 匯入資料或向其匯出資料，以及執行 IBM SPSS Statistics 程序。
- **Python** 節點可用於執行 Python 演算法。
- **Spark** 節點可以用來執行 Spark 演算法。

隨著對 SPSS Modeler 的熟悉，您也可以自訂供自己使用的選用區內容。

在「節點選用區」的左端，您可以通過選取受監督、關聯或分區段對所顯示的節點進行過濾。

「節點選用區」下方是一個報告窗格，此窗格提供對各種作業進展的意見回饋，例如何時將資料讀入資料串流中。「節點選用區」下方還有一個狀態窗格，此窗格提供有關應用程式目前正在執行的操作的資訊以及何時需要使用者意見的指示資訊。

註：如果使用啟用了內建觸控板的 **Force Click and haptic feedback** 設定的 MacBook，那麼從節點選用區將節點拖曳到串流畫布可能會導致向畫布新增複製節點。為避免發生此問題，建議停用 **Force Click and haptic feedback** 觸控板系統喜好設定。

IBM SPSS Modeler 管理員

管理程式窗格位於視窗右上角。此窗格包含用於管理串流、輸出和模型的三個標籤。

可以使用「串流」標籤開啟、重新命名、儲存和刪除階段作業中建立的串流。

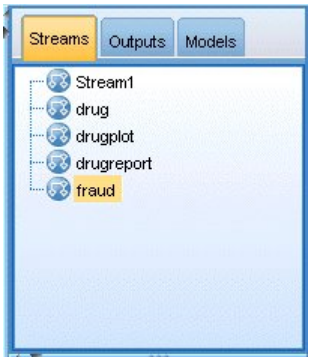


圖 *5.* 「串流」標籤

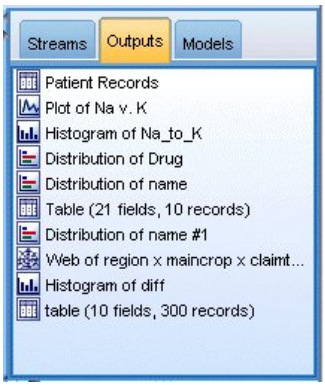


圖 *6.* 「輸出」標籤

「輸出」標籤中包含由 IBM SPSS Modeler 中的串流作業生成的各類檔案，如圖形和表格。您可以顯示、儲存、重新命名和關閉此標籤上列出的表格、圖形和報告。

圖 *7.* 包含模型區塊的「模型」標籤

在管理器標籤中，「模型」標籤具有最強大的功能。該標籤中包含所有模型**塊**，這些模型區塊包含針對現行階段作業在 IBM SPSS Modeler 中產生的模型。可以直接從「模型」標籤瀏覽這些模型或者將它們新增到畫布內的串流中。

IBM SPSS Modeler 專案

視窗右端底部是專案窗格，用於建立和管理資料採礦**專案**（與資料採礦作業相關的檔案群組）。可以通過兩種方法來檢視您在 IBM SPSS Modeler 中建立的專案：「類別」視圖和 CRISP-DM 視圖。

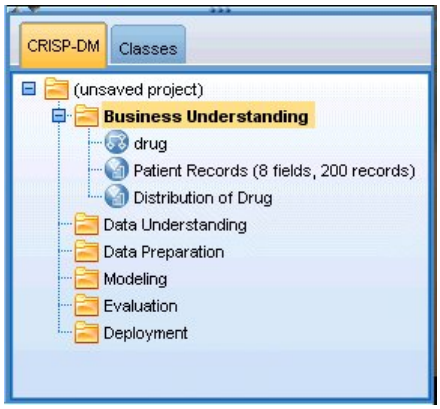


圖 *8. CRISP-DM* 視圖

依據業內認可的非專利方法「跨行業資料採礦標準程序」，CRISP-DM 標籤提供了一種專案組織方法。無論是有經驗的資料採礦人員還是新手，使用 CRISP-DM 工具都會使您事半功倍。

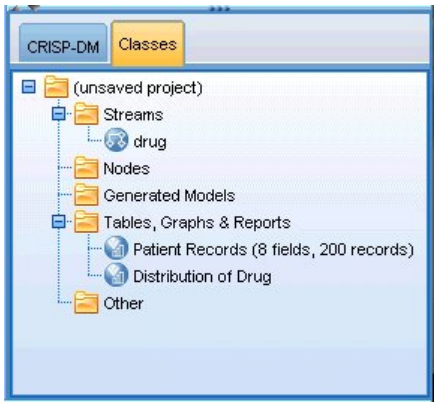


圖 *9.* 「類別」視圖

「類別」標籤提供了一種在 IBM SPSS Modeler 中按種類（即按照所建立物件的種類）組織您工作的方式。此視圖在盤點資料、串流和模型的庫存時非常有用。

IBM SPSS Modeler工具列

IBM SPSS Modeler 視窗頂部有一個圖示工具列，其中包含許多有用功能。下面是一些工具列按鈕及其功能。

建立新串流

開啟串流

儲存串流

列印目前串流

剪下並移到剪貼簿

複製到剪貼簿

貼上選擇範圍

還原上一個動作

重做

搜尋節點

編輯串流內容

預覽 SQL 產生

執行目前串流

執行串流選擇

停止串流（僅在串流處於執行狀態時可用）

新增 SuperNode

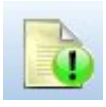
放大（僅限於 SuperNode）

縮小（僅限於 SuperNode）

串流中無標示

插入備註

隱藏串流標記（如果有）



顯示隱藏的串流標記

在 IBM SPSS Modeler Advantage 中開啟串流

串流標記由串流備註、模型鏈結和評分分支指示組成。

在《*IBM SPSS* 建模節點》手冊中介紹了模型鏈結。

自訂工具列

您可以變更工具列的各個方面，例如：

- 是否顯示
- 圖示是否有可用工具提示
- 使用大或小圖示

要開啟或關閉工具列顯示，請執行以下操作：

1. 在主功能表上，按一下：

   視圖 > **工具列** > **顯示**

要變更工具提示或圖示大小設定，請執行以下操作：

1. 在主功能表上，按一下：

   檢視 > **自訂** > **顯示**

根據需要按一下**顯示工具提示**或**大按鈕**。

自訂 IBM SPSS Modeler 視窗

使用 SPSS Modeler 介面各部分之間的分界線，可以調整工具的大小或關閉某些工具以符合個人偏好。例如，如果要處理大型串流，那麼可以使用每條分界線上的小箭頭來關閉節點選用區、管理程式窗格和專案窗格。這樣可以最大化串流畫布，從而為處理大型串流或多個串流提供足夠的畫布。

此外，從「檢視」功能表上，按一下**節點選用區**、**管理程式**或**專案**可開啟或關閉這些項目的顯示。

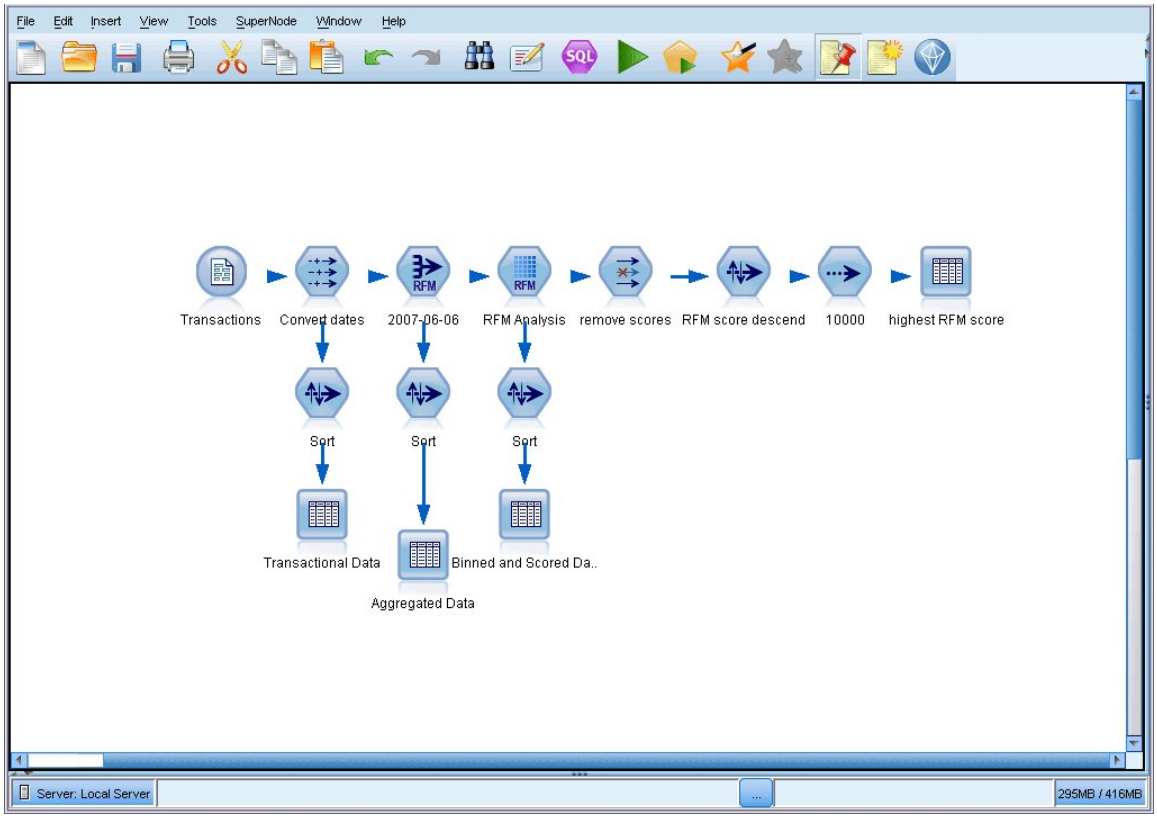


圖 *10.* 最大化的串流畫布

另外一種關閉節點選用區和管理程式以及專案窗格的方法是：垂直或水平移動 SPSS Modeler 視窗一端或底部的捲軸，將串流畫布當作捲動頁使用。

您也可以控制畫面標示的顯示，此標示由串流備註、模型鏈結和評分分支指示組成。要開啟或關閉此顯示，請按一下：

視圖 > **串流標記**

變更串流的圖示大小

您可以通過下列方式變更串流圖示的大小。

- 串流內容設定
- 串流中的蹦現功能表
- 使用鍵盤

您可以調整整個串流視圖的大小，將其調整為標準圖示大小的 8% 至 200% 之間的某個尺寸。

要調整整個串流的大小（串流內容方法）

1. 從主功能表中，選擇

   工具 > **串流內容** > **選項** > **佈置**。
2. 從「圖示大小」功能表中選擇所需大小。
3. 按一下**套用**以請參閱結果。
4. 按一下**確定**以儲存變更。

要調整整個串流的大小（功能表方法）

1. 用滑鼠右鍵按一下畫布上的串流背景。
2. 選擇**圖示大小**，並選擇所需的大小。

要調整整個串流的大小（鍵盤方法）

1. 同時按住主鍵盤上的 Ctrl + [-] 來縮小至下一個較小的尺寸。
2. 同時按住主鍵盤上的 Ctrl + Shift + [+] 來放大至下一個較大的尺寸。

獲取複合串流的整體視圖時，此功能尤其有用。您還可以使用此功能來最大程度地減少列印串流所需的頁數。

在 IBM SPSS Modeler 中使用滑鼠

IBM SPSS Modeler 中最常見的滑鼠用法如下所示：

- **按一下。**使用滑鼠右鍵或左鍵從功能表中選取選項、開啟蹦現功能表，以及存取其他標準控制項和選項。按一下並按住按鈕可移動和拖曳節點。
- **按兩下。**按兩下滑鼠左鍵可將節點放入於串流畫布中以及編輯現有節點。
- **按一下滑鼠中鍵。**按一下滑鼠中鍵並拖曳游標可連接串流畫布中的節點。按兩下滑鼠中鍵可切斷某個節點的連接。如果沒有三鍵滑鼠，可在按一下並拖曳滑鼠時通過按 Alt 鍵來模擬此功能。

使用快速鍵

IBM SPSS Modeler 中的多數可視化程式設計作業均有與之關聯的快速鍵。例如，可通過按一下某個節點並按鍵盤上的 Delete 鍵將此節點刪除。同樣地，可在按住 Ctrl 鍵的同時按 S 鍵來快速儲存某個串流。控制指令（例如此指令）由 Ctrl 和其他鍵的組合指定，例如 Ctrl+S。

標準 Windows 作業中使用了大量快速鍵，例如使用 Ctrl+X 來執行剪下作業。IBM SPSS Modeler 不僅受支援這些快速鍵，而且還受支援下列應用程式特定的快速鍵。

註：在某些情況下，IBM SPSS Modeler 中使用的舊快速鍵與標準 Windows 快速鍵相衝突。受支援將這些舊快速鍵與 Alt 鍵組合使用。例如，可以使用 Ctrl+Alt+C 來開啟或關閉快取。

表 *1.* 支援的快速鍵

快速鍵	功能
Ctrl+A	選擇全部
Ctrl+X	剪下
Ctrl+N	新串流
Ctrl+O	開啟串流
Ctrl+P	列印(P)
Ctrl+C	複製
Ctrl+V	貼上
Ctrl+Z	復原
Ctrl+Q	選取選定節點的所有下游節點
Ctrl+W	取消全選下游節點（使用 Ctrl+Q 進行切換）
Ctrl+E	從選定節點執行
Ctrl+S	儲存目前串流
Alt+方向鍵	向所使用的箭頭方向移動串流畫布上的選定節點
Shift+F10	開啟選定節點的蹦現功能表

表 2. 原有快速鍵支援的捷徑

快速鍵	功能
Ctrl+Alt+D	複製節點
Ctrl+Alt+L	載入節點
Ctrl+Alt+R	更名節點
Ctrl+Alt+U	建立使用者輸入節點
Ctrl+Alt+C	切換快取開啟關
Ctrl+Alt+F	清除快取
Ctrl+Alt+X	展開受
Ctrl+Alt+Z	放大/縮小
Delete	刪除節點或連線

正在列印

可在 IBM SPSS Modeler 中列印下列物件：

- 串流圖
- 圖形
- 表格(T)
- 報告（來自報告節點和專案報告）
- Script（來自「串流內容」、「獨立式 Script」或「SuperNode Script」對話框）
- 模型（模型瀏覽器、包含目前內容的對話框標籤、樹狀結構檢視器）
- 註解（使用輸出的「註解」標籤）

要列印物件：

- 要不預覽就列印，請按一下工具列上的「列印」按鈕。
- 要在列印前設定頁，請選取「檔案」功能表中的**版面設定**。
- 要在列印前預覽，請選取「檔案」功能表中的**預覽列印**。
- 要檢視標準列印對話框中用於選取印表機以及指定外觀的選項，請選取「檔案」功能表中的**列印**。

實現 IBM SPSS Modeler 的自動化

由於進階資料採礦往往是一個冗長的複合過程，因此 IBM SPSS Modeler 包含對幾種類型的編碼和自動處理的支援。

- **表示式操作控制語言** (CLEM) 是一種用於分析和操作在 IBM SPSS Modeler 串流中流動的資料的語言。資料採礦者在串流作業中大量使用 CLEM 來執行作業，例如根據成本衍生利潤這種簡單的作業或是將 Web 日誌資料轉換成一組含有用資訊的欄位和記錄。
- **Script 編寫**是用於在使用者介面上實現過程自動化的強大工具。Script 可以執行使用者使用滑鼠或鍵盤執行的同一類動作。還可以指定輸出並操作已產生的模型。

第 3 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。

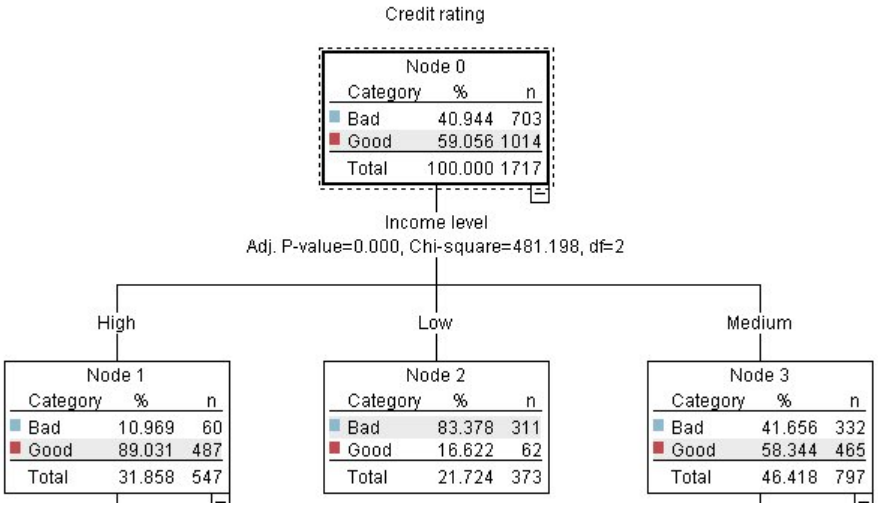


圖 *11.* 簡式決策樹狀結構模型

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。

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

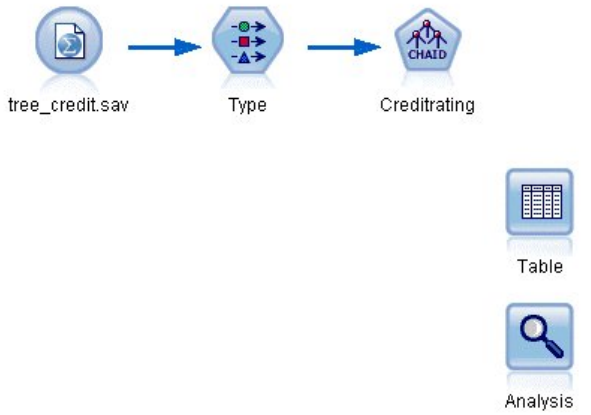


圖 *12.* 建模串流

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區塊。

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，在建立模型區塊並將其新增至串流後，將使用這些節點來檢視評分結果。

「統計量檔案」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 *13.* 使用統計量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 *Age* 欄位）包含連續數值，而**名義**欄位（如 *Credit rating* 欄位）有兩個以上的不同值，例如*Bad*、*Good* 或 *No credit history*。**序數**欄位（如 *Income level* 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 *Low*、*Medium* 和 *High*。

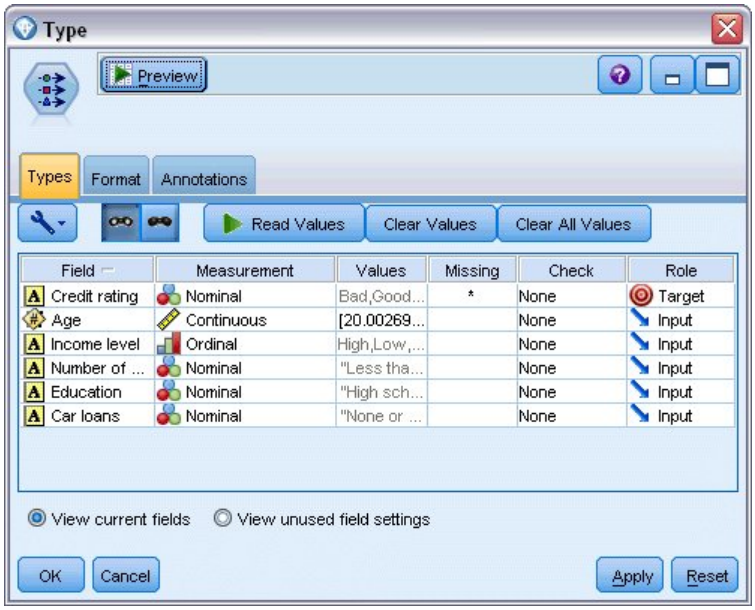


圖 *14.* 使用類型節點設定目標和輸入欄位

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 *Credit rating*，角色設為 *Target*，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 *Input*。輸入欄位有時稱為**預測工具**，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。

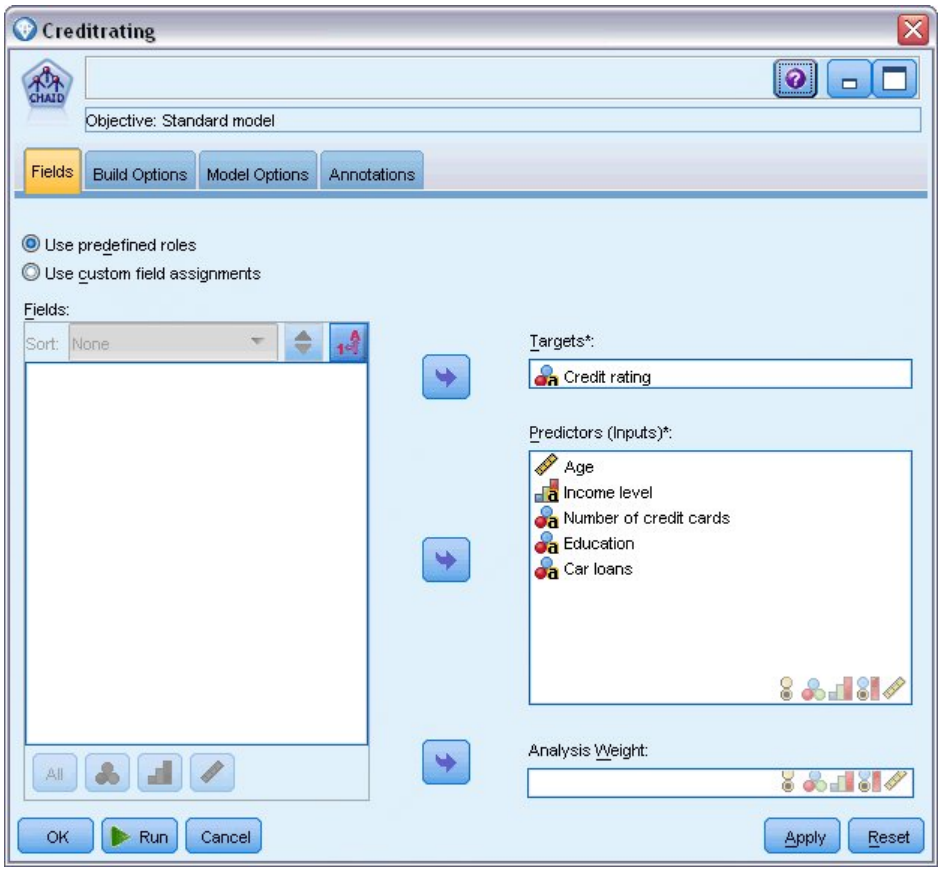


圖 *15. CHAID* 建模節點，欄位標籤

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。

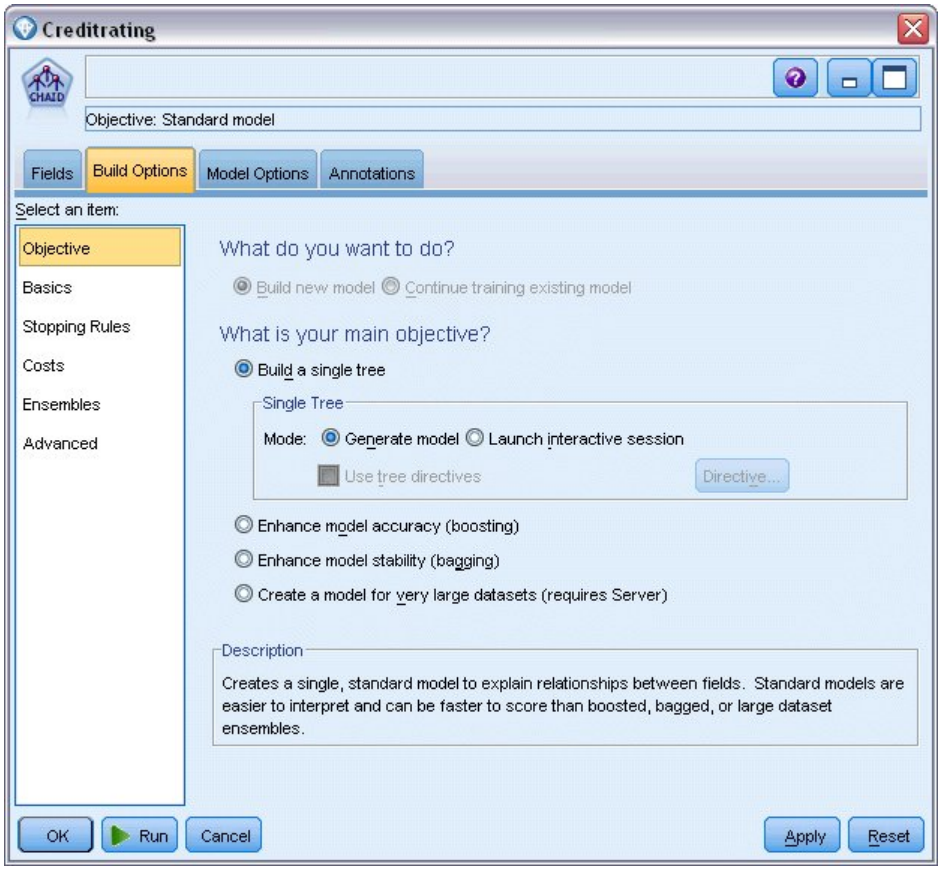


圖 *16. CHAID* 建模節點，建置選項標籤

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。

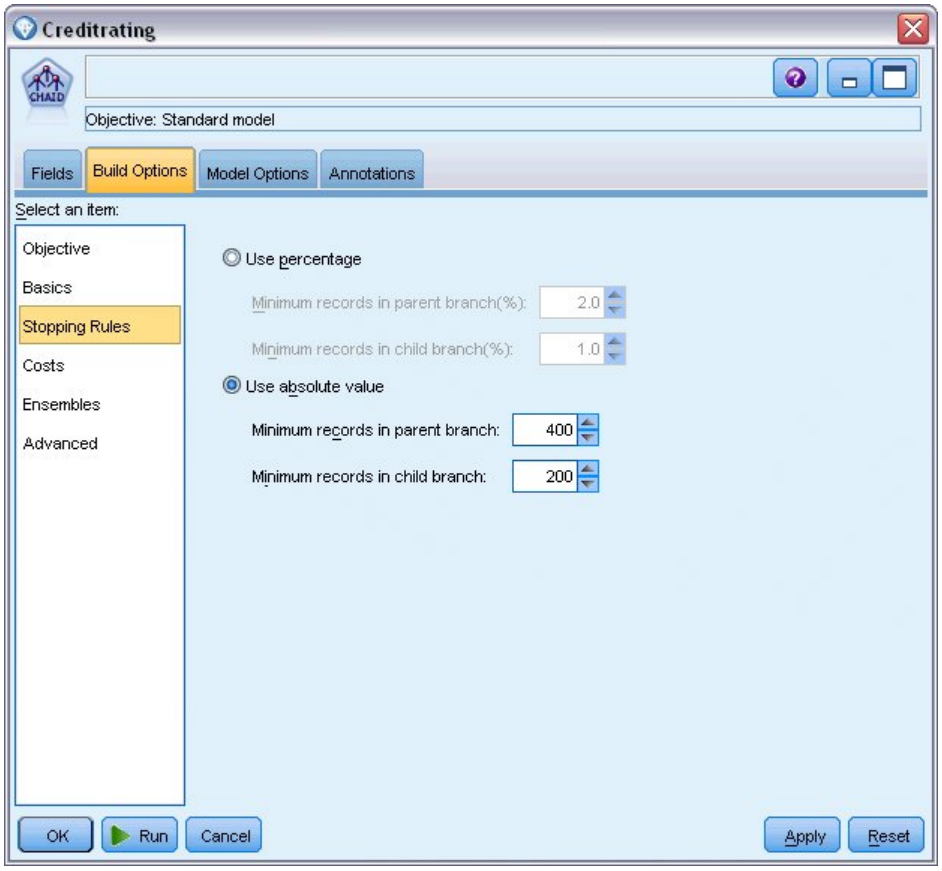


圖 *17.* 為決策樹狀結構建置設定停止準則

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，模型區塊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區塊，並選擇**瀏覽**（在模型選用區上）或**編輯**（在畫布上）。

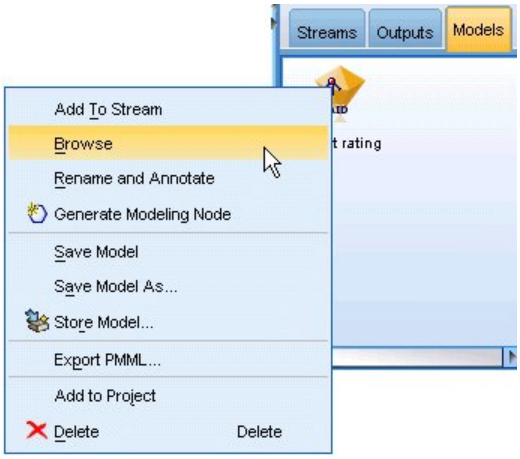


圖 *18.* 「模型」選用區

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。

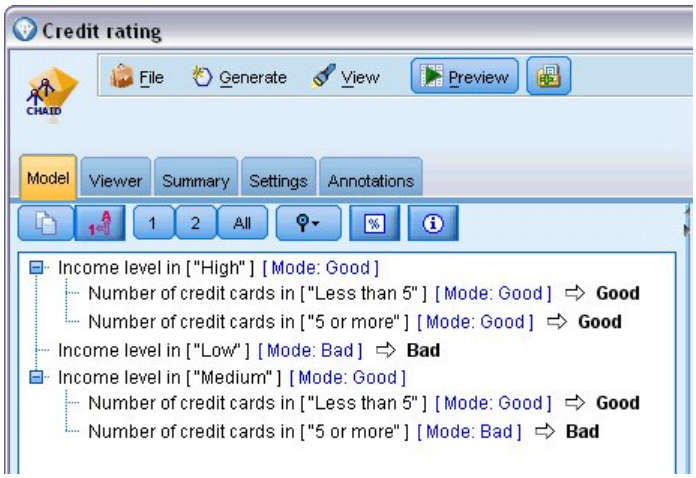


圖 *19. CHAID* 模型區塊與規則集

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「預測工具重要性」圖表，其顯示評估模型時每個預測工具的相對重要性。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。

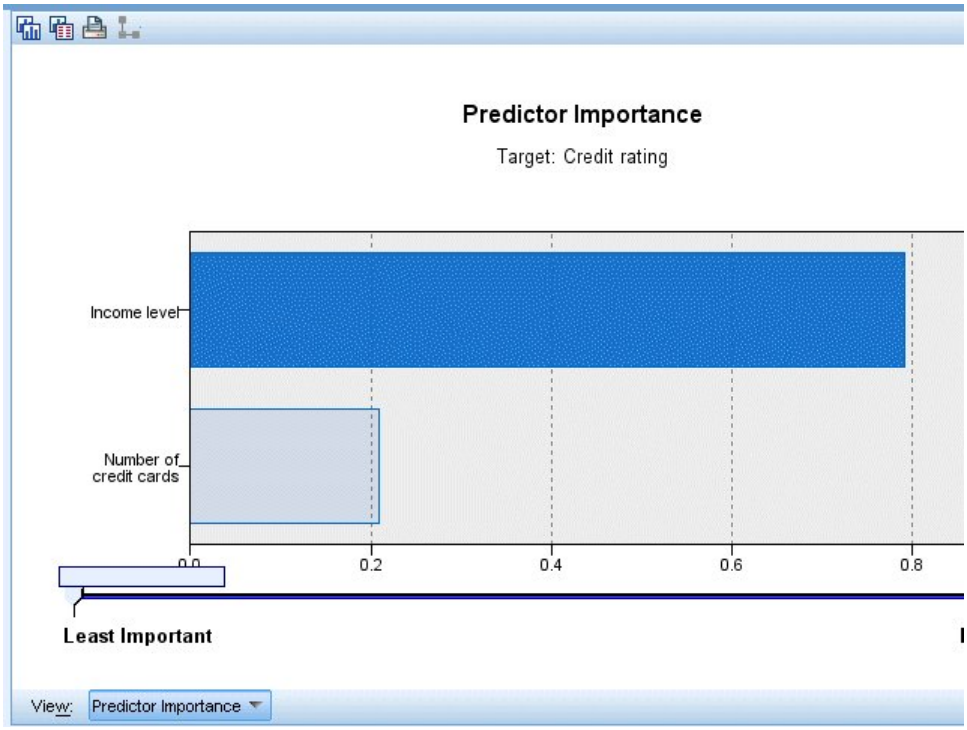


圖 *20.* 預測工具重要性圖表

模型區塊中的「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。

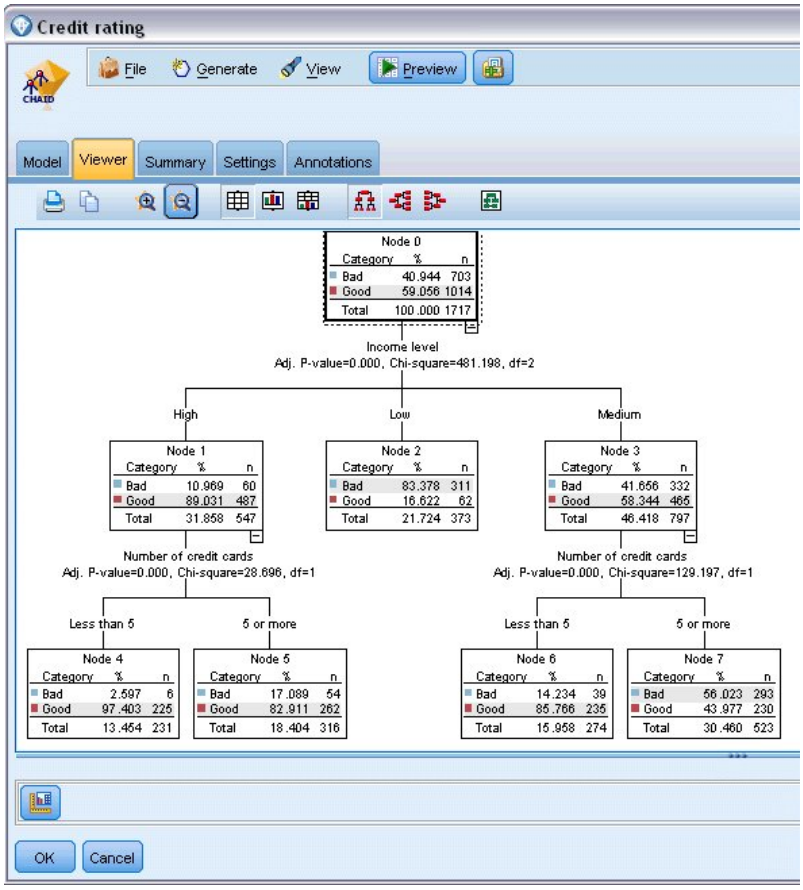


圖 *21.* 模型區塊中的檢視器標籤（已選取縮小）

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。

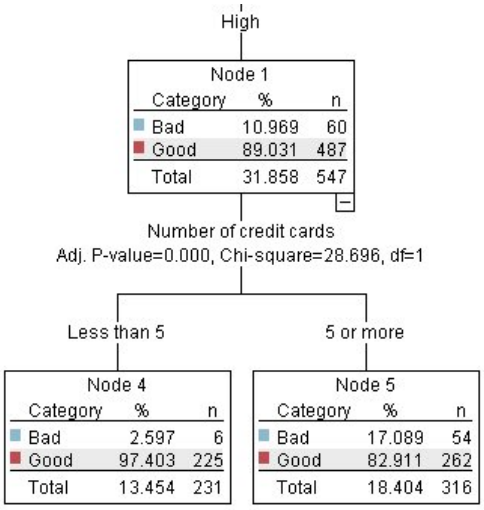


圖 22. 高收入客戶的樹狀結構視圖

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。

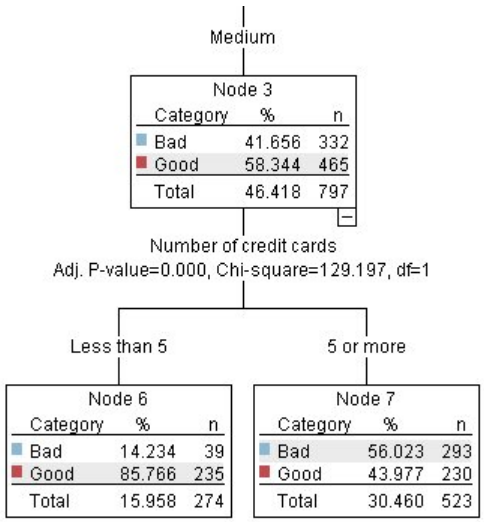


圖 23. 中等收入客戶的樹狀結構視圖

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。

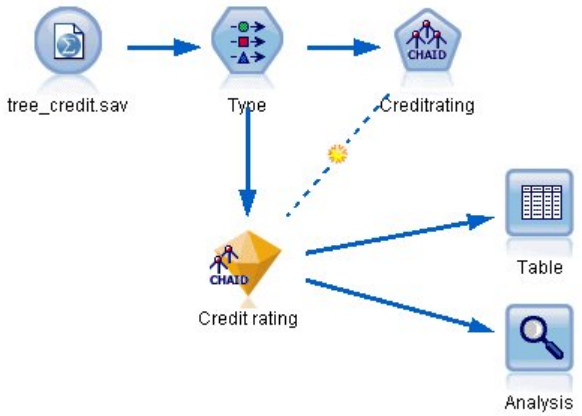
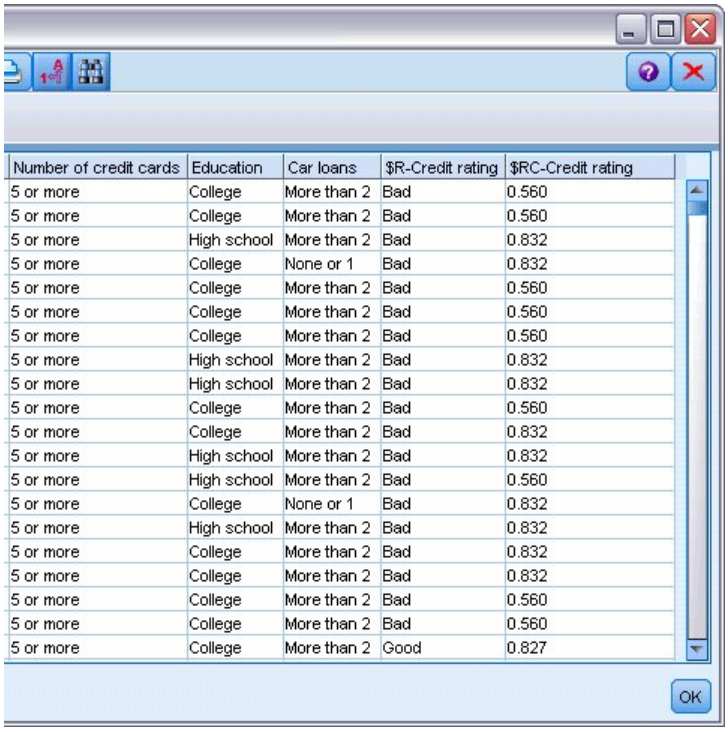


圖 *24.* 將模型區塊連接到輸出節點以評估模型

1. 若要查看評分或預測，請將「表格」節點連接到模型區塊，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 *Credit rating* 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 *25.* 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*Credit rating* 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. 將模型區塊連接至「分析」節點。
3. 按兩下「分析」節點，然後按一下**執行**。

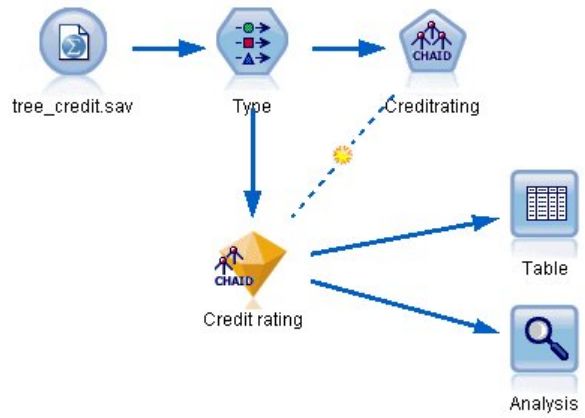


圖 *26.* 連接分析節點

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。

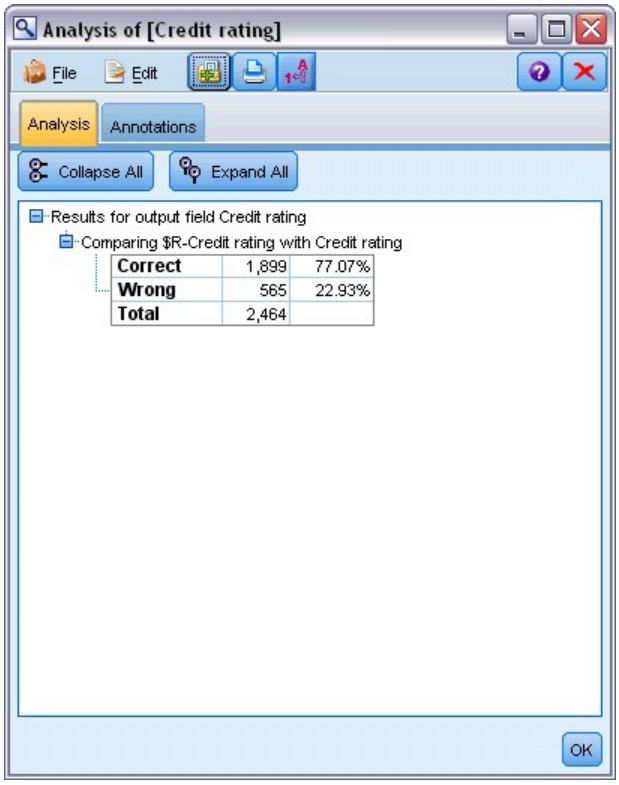


圖 *27.* 比較已觀察和預測回應值的分析結果

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。

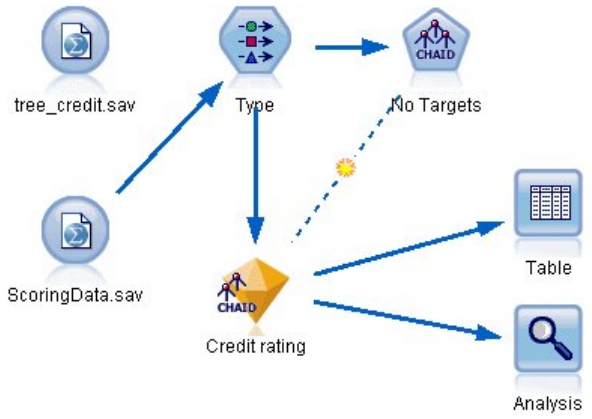


圖 *28.* 連接新資料以進行評分

您可以更新「統計量檔案」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（*Age*、*Income level*、*Education* 等），而不是目標欄位 *Credit rating*。

或者，您可以將模型區塊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。

您也可以將模型區塊儲存為個別檔案，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，並建立模型區塊。這有時稱為訓練模型。
- 模型區塊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 4 章 旗靶的自動建模

給客戶回應（自動分類器）建模

「自動分類器」節點可讓您自動建立並比較旗標（例如給定的客戶是否有可能拖欠貸款或對特定優惠作出回應）或名義（集）目標的數個不同模型。在此範例中，我們將搜尋旗標（yes 或 no）結果。在相對簡單的串流中，節點會產生一組候選模型並進行排名，選擇執行結果最好的模型，然後將它們結合到單個聚集（組合）模型中。這種方法透過結合多個模型將自動化的便利與好處結合在一起，這樣所獲得的預測會比通過任意一個模型獲得的預測更為準確。

本範例基於虛構公司，該公司希望通過向每個客戶提供合適的優惠以獲取更多利潤。

此方法強調了自動化的好處。對于使用連續（數值範圍）目標的類似範例，請參閱內容值（自動數值）。

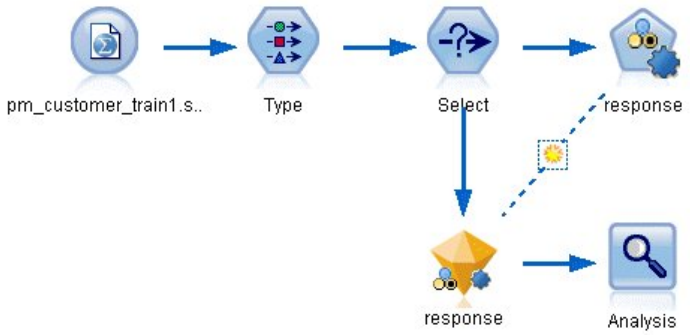


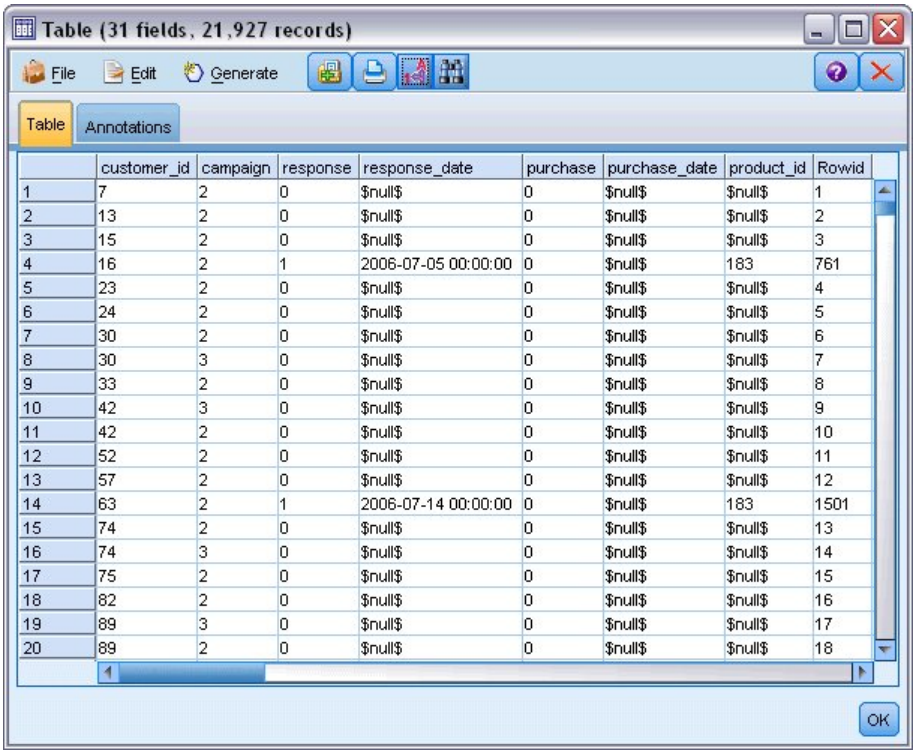
圖 *29.* 自動分類器串流範例

本範例使用安裝在 Demo 資料夾的 *streams* 之下的串流 *pm_binaryclassifier.str*。使用的資料檔案為 *pm_customer_train1.sav*。請參閱『歷程資料』主題，以取得更多資訊。

歷程資料

檔案 *pm_customer_train1.sav* 具有歷程資料，可用於追蹤過去行銷活動中針對特定客戶提供的優惠，如 *campaign* 欄位的值所指示。最大數目的記錄位於進階帳戶 行銷活動範圍之下。

campaign 欄位的值在資料中實際編碼為整數（例如 *2* = 進階帳戶）。之後，您將為這些值定義標籤，供您用來提供更有意義的輸出。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

圖 *30.* 先前促銷的相關資料

該檔案還包括回應 欄位，其中包括是否接受了優惠（0 = 否，而 1 = 是）。這將會是**目標欄位**或您要預測的值。包含每個客戶相關的個人背景資訊及財務資訊的數個欄位也包括在內。這些欄位可用來建置或「訓練」一個模型，以根據收入、年齡或每月交易數等性質來預測個人或群組的回應率。

建置串流

1. 新增指向 *pm_customer_train1.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的 Statistics「檔案」來源節點。（您可以在檔案路徑中指定 `$CLEO_DEMOS/` 作為參照此資料夾的捷徑。請注意，必須在路徑中使用正斜線而非反斜線，如範例所示）。

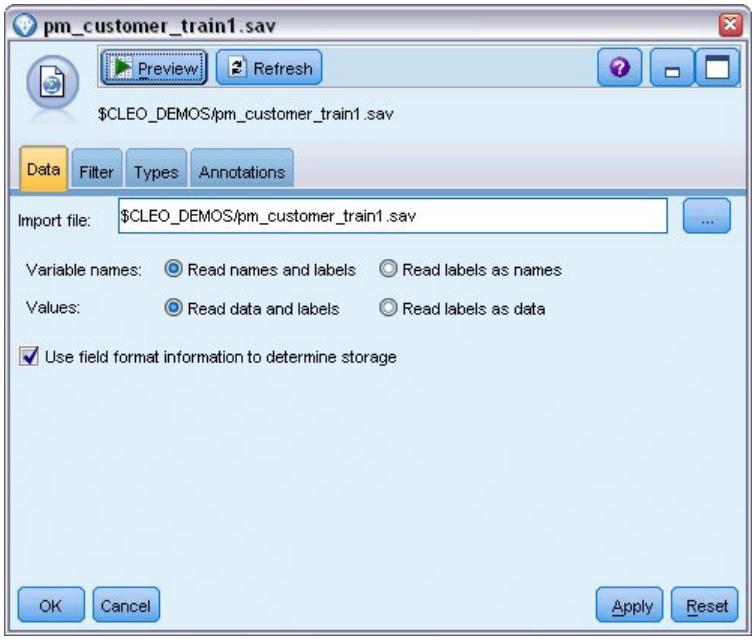


圖 *31.* 在資料中讀取

2. 新增「類型」節點，並選取 *response* 作為目標欄位（角色 = **目標**）。將此欄位的「測量」層次設為**旗標**。

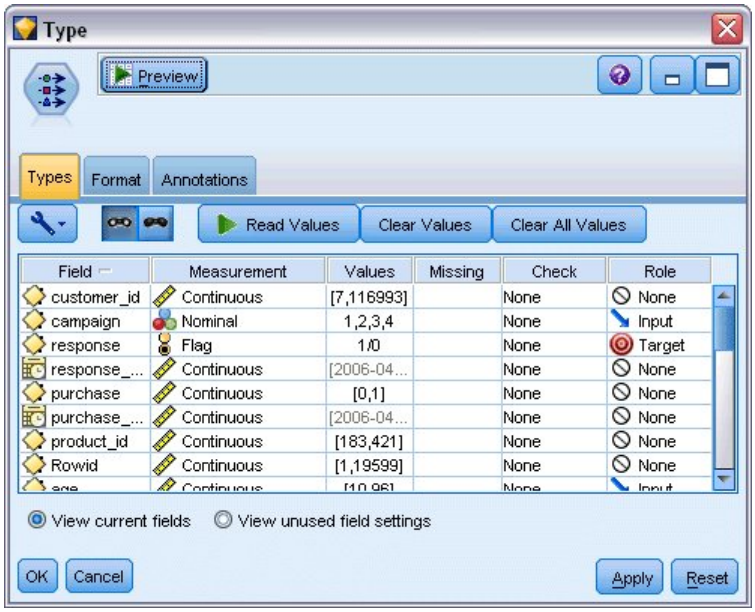


圖 *32.* 設定測量層次和角色

3. 針對下列欄位，將角色設為**無**：*customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid* 和 *X_random*。當您建置模型時，將會忽略這些欄位。
4. 按一下「類型」節點中的**讀取值**按鈕以確保已實例化值。

如先前所見，我們的來源資料包括四個不同行銷活動的相關資訊，每個行銷活動的目標設為不同類型的客戶帳號。這些行銷活動在資料中編碼成整數，因此，為了更容易記住每個整數代表的帳戶類型，讓我

們為每個帳戶定義標籤吧。

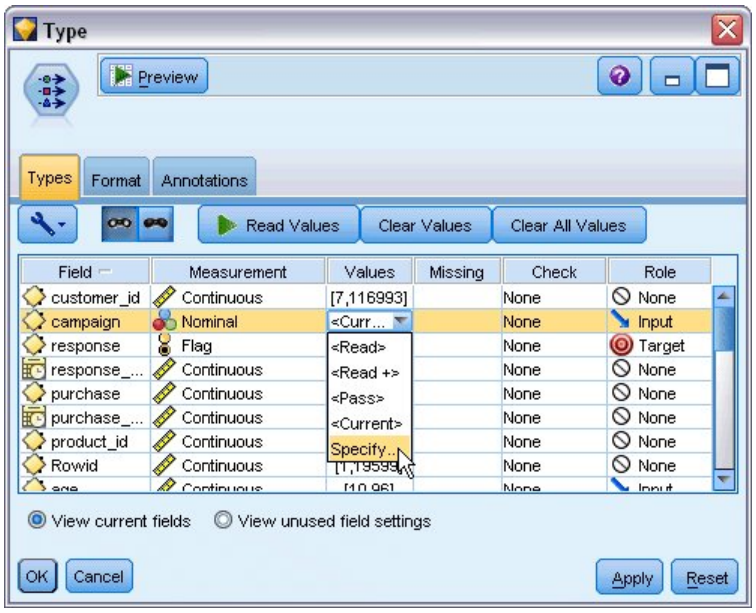


圖 *33.* 選擇以為欄位指定值

5. 在**行銷活動**欄位的列上，按一下**值**直欄中的項目。
6. 從下拉清單中選擇**指定**。

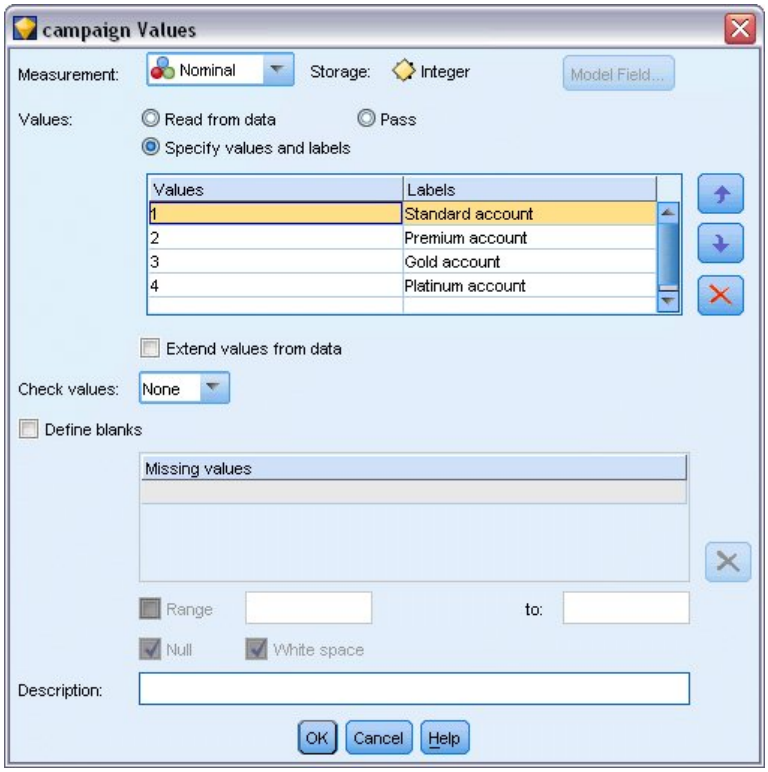
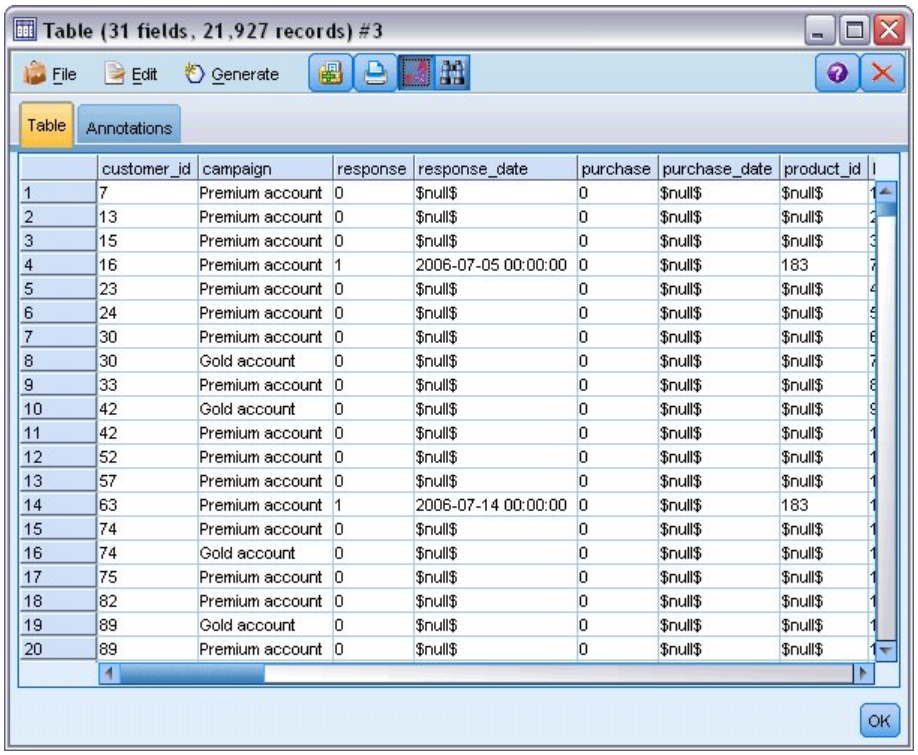


圖 *34.* 為欄位值定義標籤

7. 在**標籤**直欄中，為**行銷活動**欄位的四個值中的每個值輸入所顯示的標籤。

8. 按一下「**確定**」。

現在，您可以在輸出視窗中顯示標籤而非整數。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold account	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold account	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold account	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold account	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

圖 *35.* 顯示欄位值標籤

9. 將「表格」節點連接至「類型」節點。
10. 開啟「表格」節點並按一下**執行**。
11. 在輸出視窗上，按一下**顯示欄位和值標籤**工具列按鈕以顯示標籤。
12. 按一下**確定**以關閉輸出視窗。

雖然資料包括四個不同行銷活動的相關資訊，但您將著重於一次分析一個行銷活動。由於最大數量的記錄位於進階帳戶行銷活動範圍之下（在資料中編碼為 *campaign=2*），因此您可以使用「選取」節點以僅在串流中包括這些記錄。

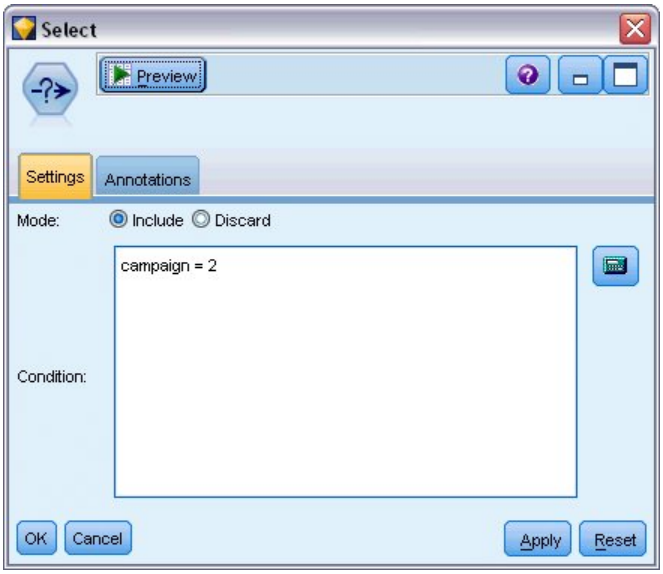


圖 *36.* 選取單一行銷活動的記錄

產生並比較模型

1. 連接「自動分類器」節點，然後選取**整體正確性**作為模型分級所用的度量值。
2. 將**要使用的模型數目**設為 3。這表示當您執行節點時，將建置三個最佳模型。

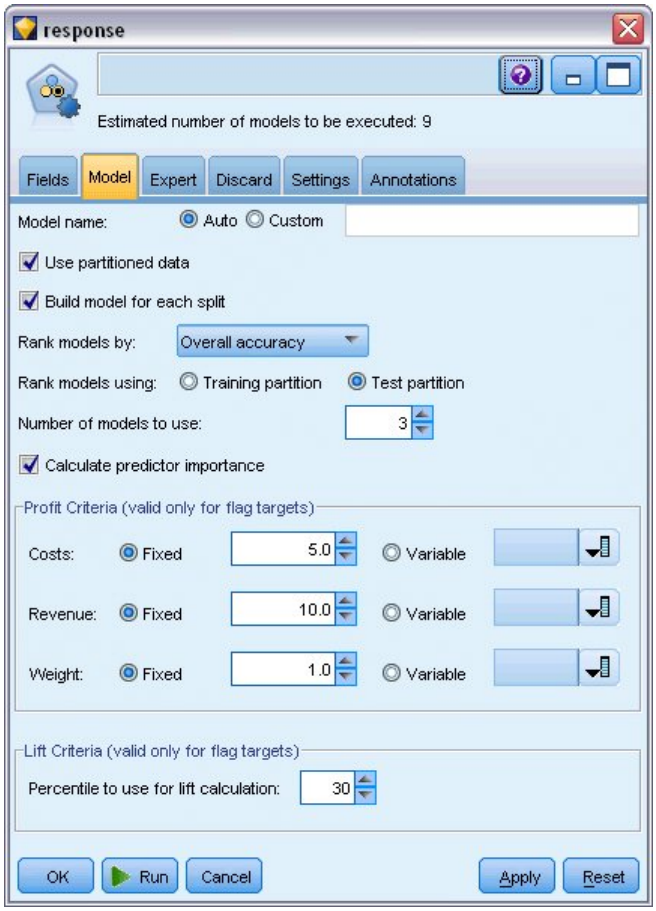


圖 *37.* 自動分類器節點的模型標籤

在「專家」標籤上，您可以從多達 11 個不同的模型演算法中進行選擇。

3. 取消選取**區別元件**和 **SVM** 模型類型。（這些模型需要更長時間來訓練這些資料，因此取消選取它們將加快範例的速度。如果您不介意等待，可以保留選取這些模型類型。）

由於您在「模型」標籤上將**要使用的模型數目**設為 3，因此節點將計算剩餘 9 個演算法的正確性，並建置一個包含三個最精確演算法的模型區塊。

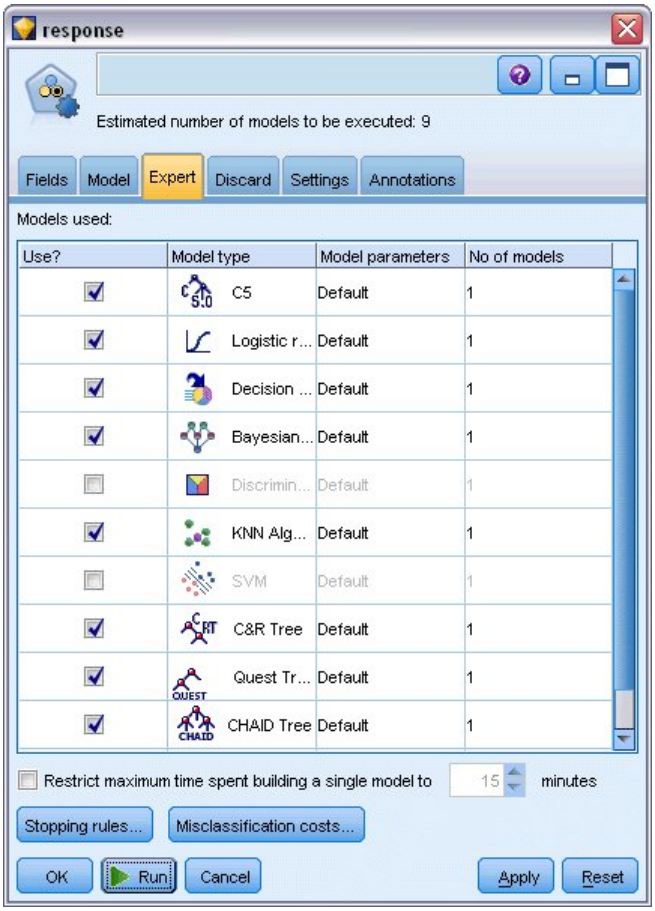


圖 *38.* 自動分類器節點的專家標籤

4. 在「設定」標籤上，針對組合方法選取**信賴度加權投票**。這可確定為每筆記錄產生單一聚集評分的方法。

   使用簡式投票，若三個模型中有兩個預測是 ，則是 以 2 比 1 的投票結果取勝。如果是使用信賴度加權投票，則是根據每個預測的信賴度值對投票進行加權。因此，如果預測否 的一個模型的信賴度高於兩個預測是 的模型組合，則否 勝出。

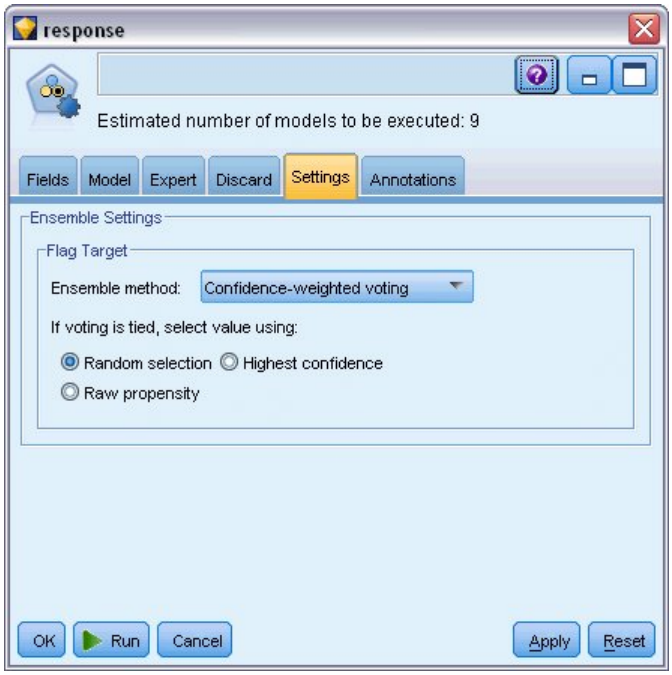


圖 *39.* 自動分類器節點：設定標籤

5. 按一下「**執行**」。

幾分鐘之後，便會建置產生的模型區塊並放置在畫布上，以及視窗右上角的「模型」選用區上。您可以瀏覽模型區塊，儲存它或使用數種其他方法進行部署。

開啟模型區塊；它列出在執行期間所建立的每個模型的相關詳細資料。（在實際情況下，可能在大型資料集上建立數百個模型，這將花費數個小時。） 請參閱第 33 頁的圖 29。

如果您要進一步探索任何個別的模型，則可以按兩下**模型**欄中的模型區塊，以往下探查並瀏覽個別的模型結果；您可以從中產生建模節點、模型區塊或評估圖表。在**圖形**欄中，您可以按兩下縮圖來產生最大的圖形。

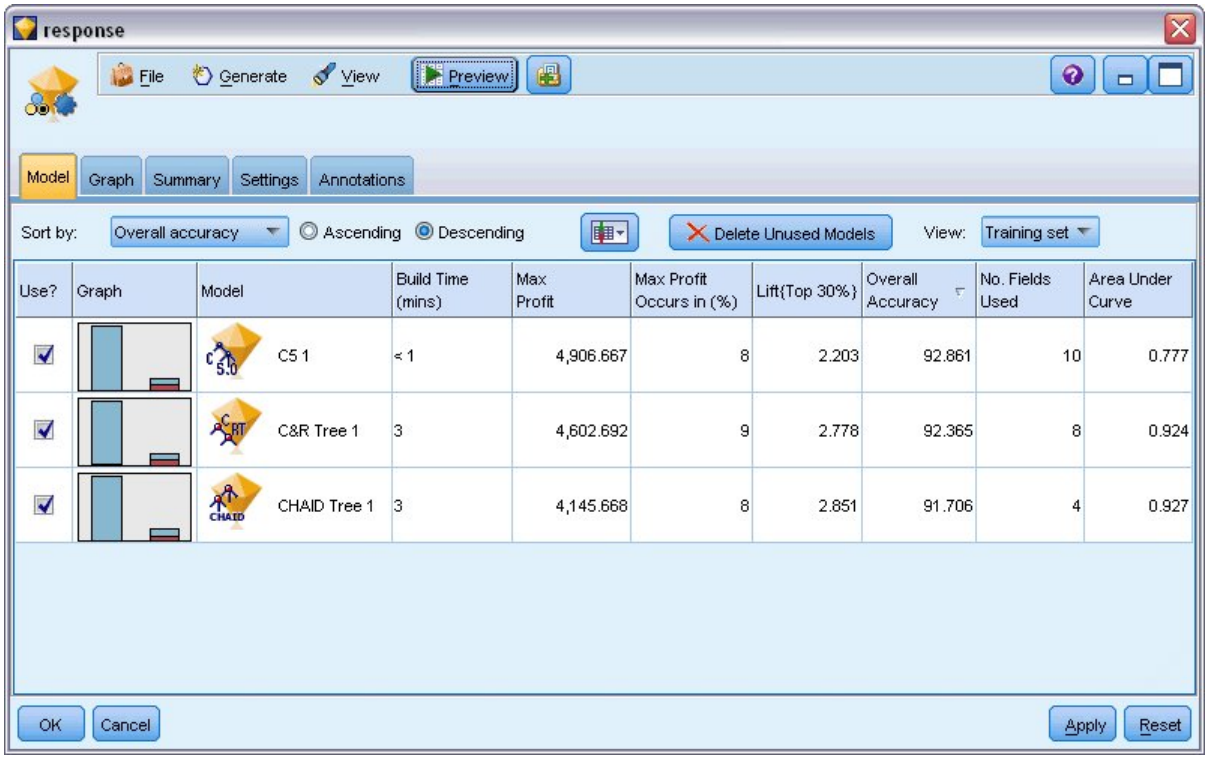


圖 *40.* 自動分類器結果

依預設，根據整體正確性對模型排序，因為這是您在「自動分類器」節點的「模型」標籤上選取的測量。C51 模型根據這個測量分級效果最佳，但 C&R Tree 和 CHAID 模型幾乎一樣精確。

您可以透過按一下不同欄的標頭來排序該欄，或者您可以從工具列上的**排序方式**下拉清單中選擇想要的測量。

根據這些結果，您決定使用這三種最精確的模型。通過結合多個模型的預測，可以避免單個模型的局限性，從而使整體正確性更高。

在**使用？**欄中，選取 C51、C&R Tree 和 CHAID 模型。

將「分析」節點（「輸出」選用區）附加在模型區塊之後。在「分析」節點上按一下滑鼠右鍵，然後選擇**執行**以執行串流。

由組合模型產生的聚集評分顯示在 *$XF-response* 欄位中。根據訓練資料測量時，預測值符合實際回應（如原始 *response* 欄位所記錄），且整體正確性為 92.82%。

在本案例中，如果正確性不及最佳的三個個別模型（C51 的正確性為 92.86%），則差異太小而沒有意義。一般來講，將組合模型套用至資料集而非訓練資料時，通常更有可能會良好地執行。

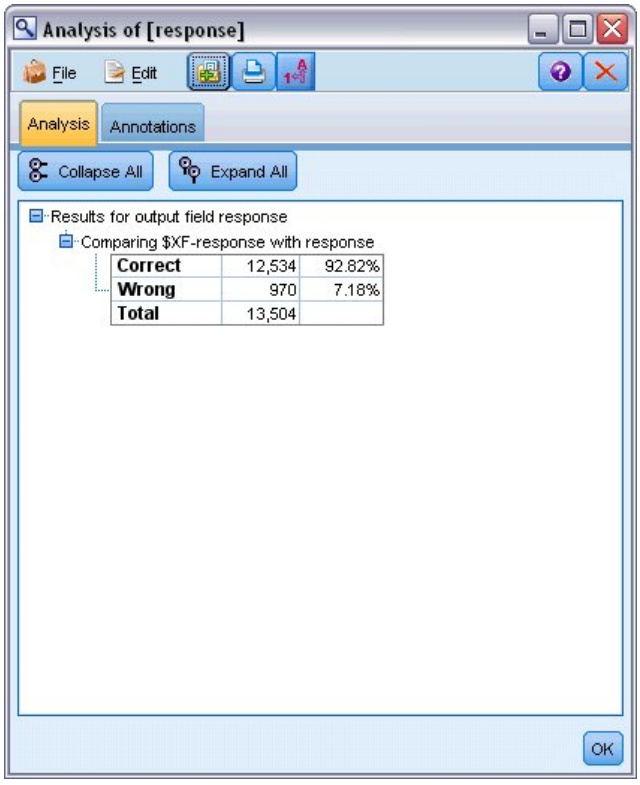


圖 *41.* 三個組合模型的分析

摘要

為了加總，您已使用「自動分類器」節點來比較數個不同的模型、已使用三個最準確的模型並將其新增至組合「自動分類器」模型區塊中的串流。

- 根據整體正確性，C51、C&R Tree 和 CHAID 模型在訓練資料上的執行效果最佳。
- 組合模型的執行效果幾乎與個別模型一樣好，當套用到其他資料集時執行效果可能會更佳。如果您的目標是盡可能多地自動化程序，則此方法可讓您在大部分情況下都能夠取得強大模型，而不必深入挖掘任何模型的特性。

第 5 章 連續目標的自動建模

內容值（自動數值）

「自動數值」節點可讓您自動建立及比較不同的模型以瞭解（數值）結果，例如預測財產的應納稅款值。使用單一節點，您可以預估並比較一組候選模型並產生一部分模型來進一步進行分析。該節點和「自動分類器」節點的工作方式相同，但針對的是連續目標而不是旗標或名義目標。

該節點將最佳的候選模型結合至單個聚集（組合）模型區塊中。這種方法透過結合多個模型將自動化的便利與好處結合在一起，這樣所獲得的預測會比通過任意一個模型獲得的預測更為準確。

此範例的焦點是負責調整和評量房地產稅的虛擬市政當局。若要更準確地執行此動作，他們將建置一個模型來根據建築物類型、街區、規模及其他已知因素來預測財產價值。

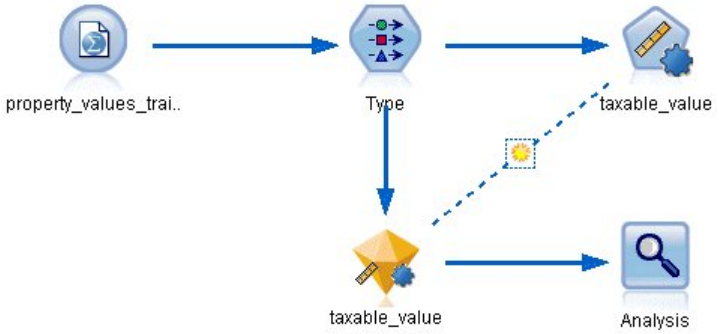


圖 *42.* 自動數值串流範例

本範例使用安裝在 Demos 資料夾的 `streams` 之下的串流 `property_values_numericpredictor.str`。使用的資料檔案為 `property_values_train.sav`。請參閱第 4 頁的『Demos 資料夾』主題，以取得更多資訊。

訓練資料

資料檔案包括名稱為 *taxable_value* 的欄位（它是**目標欄位**）或您要預測的值。其他欄位包含街區、建築物類型以及內部空間之類的資訊，可用來作為預測工具。

欄位名稱	標籤
property_id	Property ID
街區	城市內的區域
building_type	建築物類型
year_built	建築物年份
volume_interior	內部空間
volume_other	車庫和額外建築物的空間
lot_size	廣場大小
taxable_value	應納稅值

名為 *property_values_score.sav* 的評分資料檔案也包括在 Demos 資料夾中。它包含相同的欄位但不含 *taxable_value* 欄位。使用應納稅值已知的資料集來訓練模型之後，您可以對此值尚未知的記錄進行評分。

建置串流

1. 新增指向 *property_values_train.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的 Statistics「檔案」來源節點。（您可以在檔案路徑中指定 `$CLEO_DEMOS/` 作為參照此資料夾的捷徑。請注意，必須在路徑中使用正斜線而非反斜線，如範例所示）。

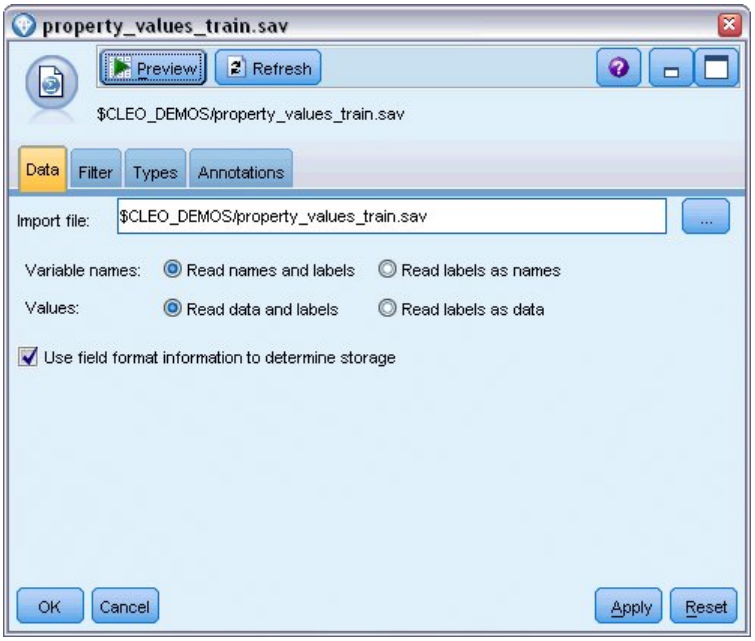


圖 *43.* 在資料中讀取

2. 新增「類型」節點，並選取 *taxable_value* 作為目標欄位（角色 = **目標**）。所有其他欄位的角色都應該設為**輸入**，表示它們將用來作為預測工具。

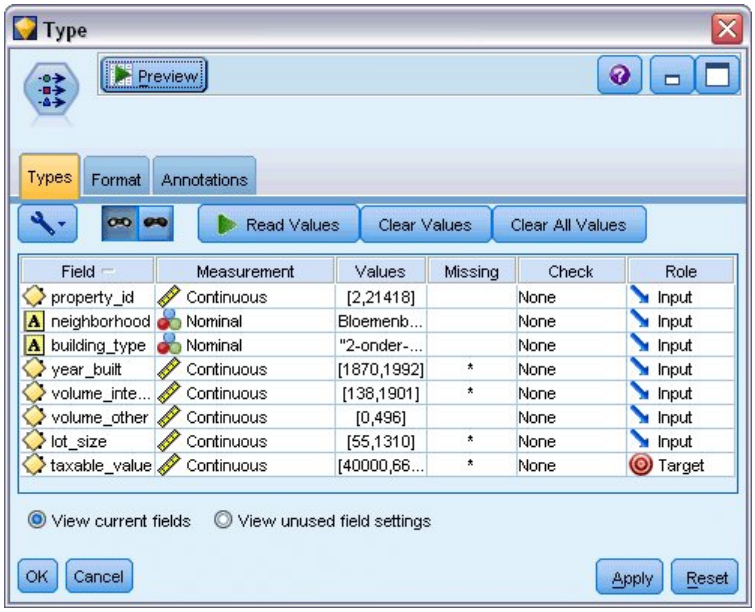


圖 *44.* 設定目標欄位

3. 連接「自動數值」節點，然後選取**相關性**作為模型分級所用的度量值。
4. 將**要使用的模型數目**設為 3。這表示當您執行節點時，將建置三個最佳模型。

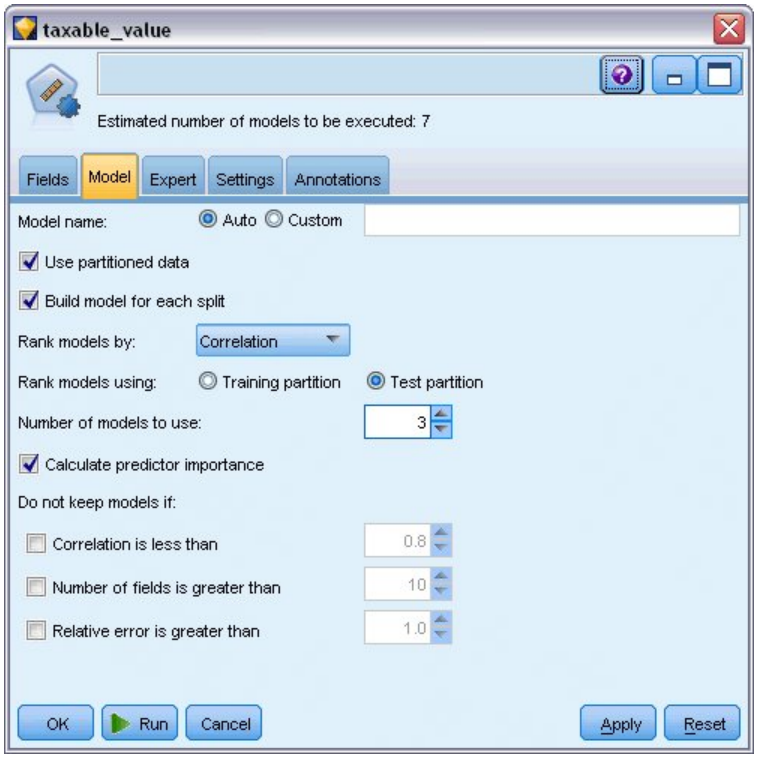


圖 *45.* 自動數值節點的模型標籤

5. 在「專家」標籤上，保留預設值不變；節點將評估每個演算法的單一模型（總共七個模型）。（或者，您可以修改這些設定來比較每個模型類型的多個變式。）

由於您在「模型」標籤上將**要使用的模型數目**設為 3，因此節點將計算 7 個演算法的正確性，並建置一個包含三個最精確演算法的模型區塊。

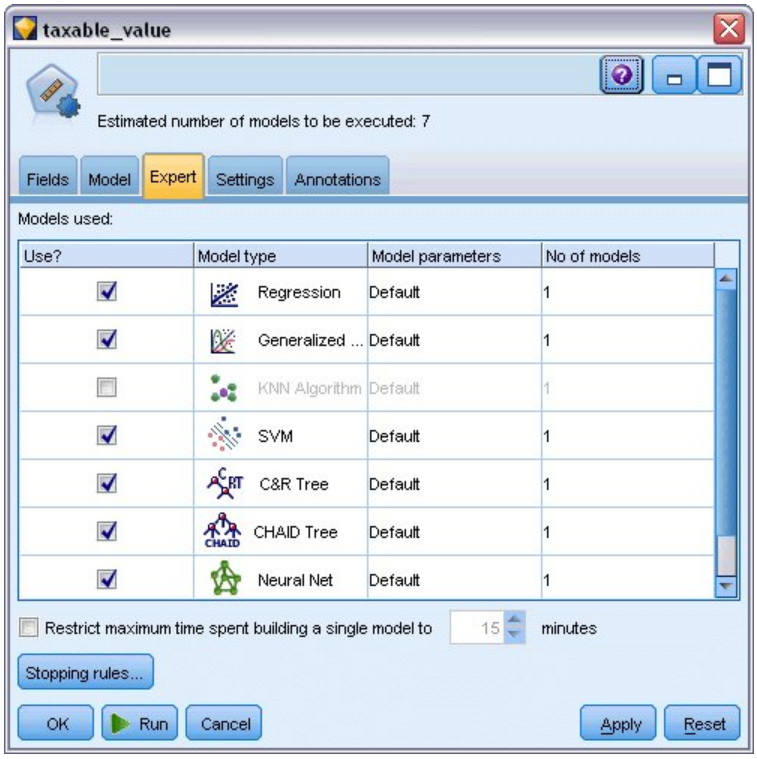


圖 *46.* 自動數值節點的專家標籤

6. 在「設定」標籤上，保留預設值不變。由於這是一個連續目標，因此會對個別模型評分求平均值來產生整體評分。

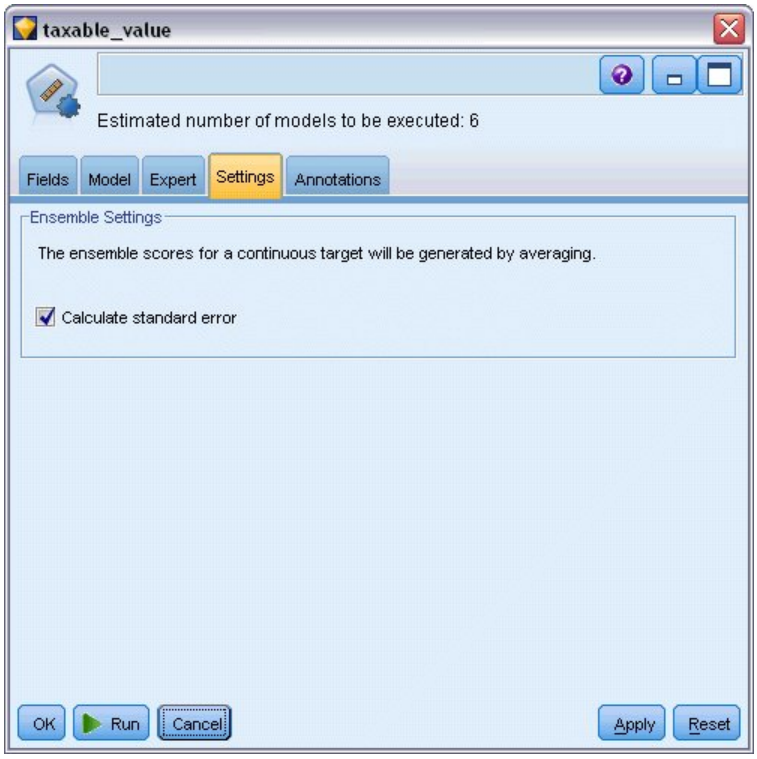


圖 *47.* 自動數值節點的設定標籤

比較模型

1. 按一下「執行」按鈕。

會建置模型區塊並放置在畫布上，以及視窗右上角的「模型」選用區上。您可以瀏覽該區塊，儲存它或使用數種其他方法進行部署。

開啟模型區塊；它列出在執行期間所建立的每個模型的相關詳細資料。（在實際情況下，會在大型資料集上預估數百個模型，這將花費數個小時。） 請參閱第 45 頁的圖 42。

如果您要進一步探索任何個別的模型，則可以按兩下**模型**欄中的模型區塊，以往下探查並瀏覽個別的模型結果；您可以從中產生建模節點、模型區塊或評估圖表。

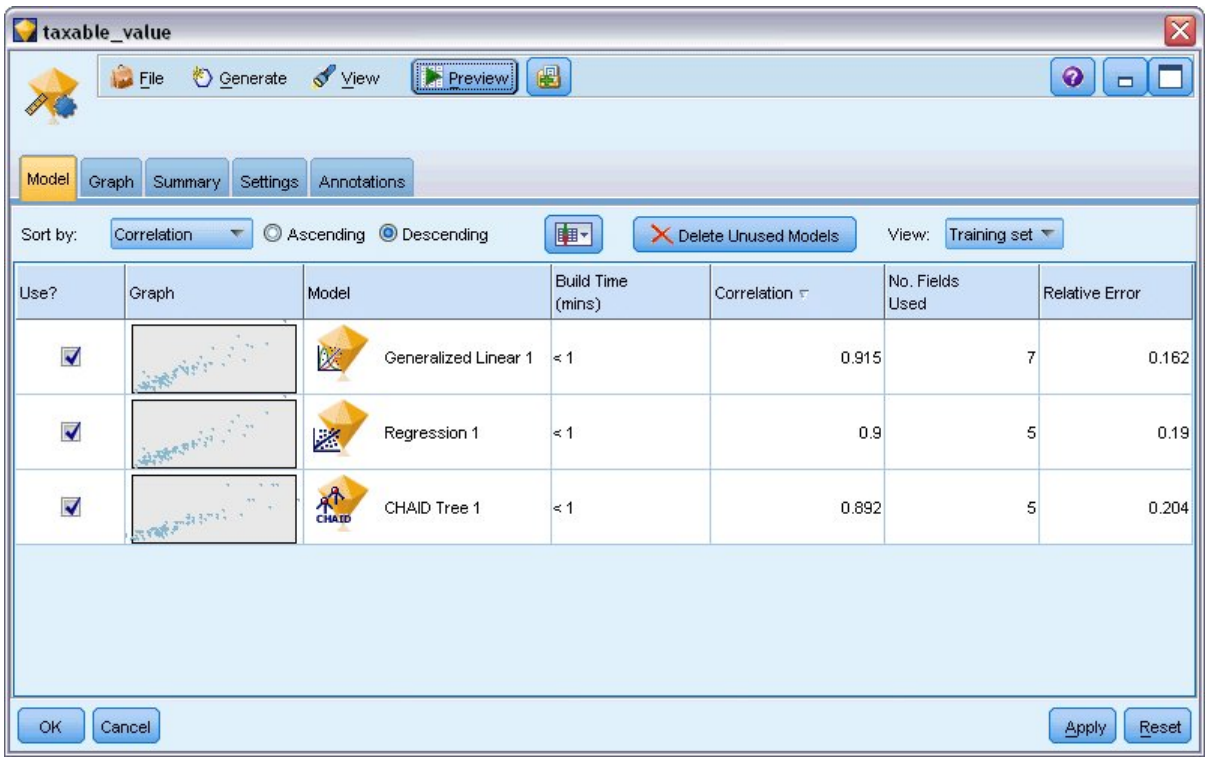


圖 *48.* 自動數值結果

依預設，將按照相關性來排序模型，因為這是您在「自動數值」欄位中選取的測量。若要分級，會使用相關性的絕對值，值越靠近 1 表示關係越強。「廣義線性」模型分級基於此測量，但數個其他模型幾乎是正確的。「廣義線性」模型也具有最低的相對錯誤。

您可以透過按一下不同欄的標頭來排序該欄，或者您可以從工具列上的**排序方式**清單中選擇想要的測量。

每個圖形將根據模型的預測值來顯示觀察值的圖，從而快速直觀地表示模型之間的相關性。對於好的模型，點應該沿著對角線形成叢集，在此範例中的所有模型都是這種情況。

在**圖形**欄中，您則可以按兩下縮圖來產生最大的圖形。

根據這些結果，您決定使用這三種最精確的模型。通過結合多個模型的預測，可以避免單個模型的局限性，從而使整體正確性更高。

在**使用？**直欄中，確保選取了所有模型（3 個）。

將「分析」節點（「輸出」選用區）附加在模型區塊之後。在「分析」節點上按一下滑鼠右鍵，然後選擇**執行**以執行串流。

由組合模型產生的平均評分會新增至欄位 *$XR-taxable_value*，且相關性為 0.922，此值高於三個個別模型的相關性。組合分數還顯示較低的平均絕對錯誤，套用到其他資料集時，其效能可能好於任何個別模型。

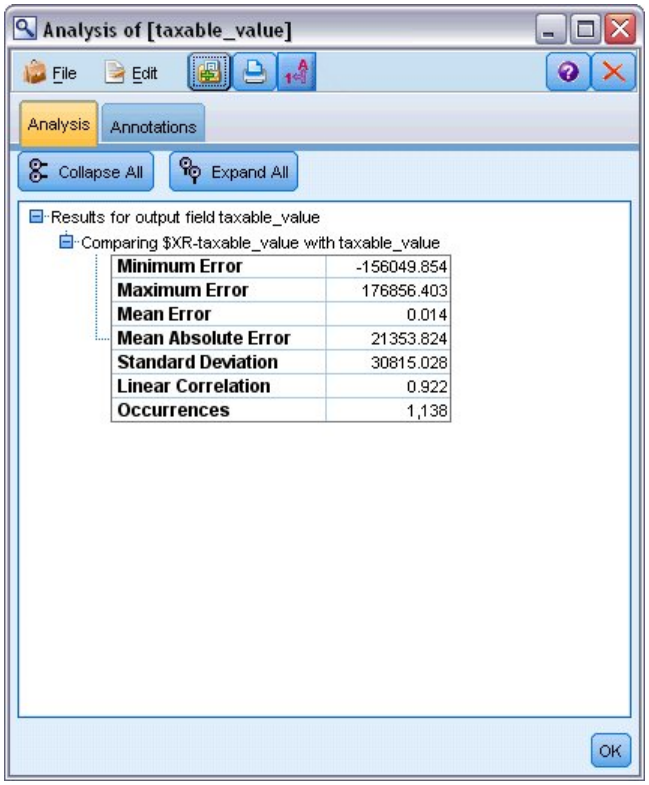


圖 *49.* 自動數值串流範例

摘要

為了加總，您已使用「自動數值」節點來比較數個不同的模型、已選取三個最準確的模型並將其新增至組合「自動數值」模型區塊中的串流。

- 根據整體正確性，廣義線性、迴歸和 CHAID 模型在訓練資料上的執行效果最佳。
- 組合模型顯示執行效果好於兩個個別模型，當套用到其他資料集時執行效果可能會更佳。如果您的目標是盡可能多地自動化程序，則此方法可讓您在大部分情況下都能夠取得強大模型，而不必深入挖掘任何模型的特性。

第 6 章 自動化資料準備 (ADP)

準備資料進行分析是任何資料採礦專案中的最重要步驟之一，並且在傳統上也是最耗時的作業之一。「自動化資料準備 (ADP)」節點會為您處理作業，分析您的資料並識別修正程式，篩選出存在問題或可能無用的欄位，並在適當的情況下衍生新的屬性，通過智能篩選技術改進效能。您可以完全自動化方式使用節點，容許節點選擇並套用修正程式，或者可在進行變更前預覽變更，並按照需要接受或拒絕變更。

使用 ADP 節點可讓您快速、輕鬆地準備資料以進行資料採礦，不需事先瞭解統計相關概念。如果您執行具有預設值的節點，則模型將更快速地建置和評分。

此範例使用參照資料檔案 *telco.sav* 的串流 *ADP_basic_demo.str* 來示範增加的正確性，可在建置模型時使用預設 ADP 節點設定來找到正確性。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*ADP_basic_demo.str* 檔位於 *streams* 目錄。

建置串流

1. 若要建置串流，請新增指向 `telco.sav`（位於 IBM SPSS Modeler 安裝的 `Demos` 目錄中）的「統計量檔案」來源節點。

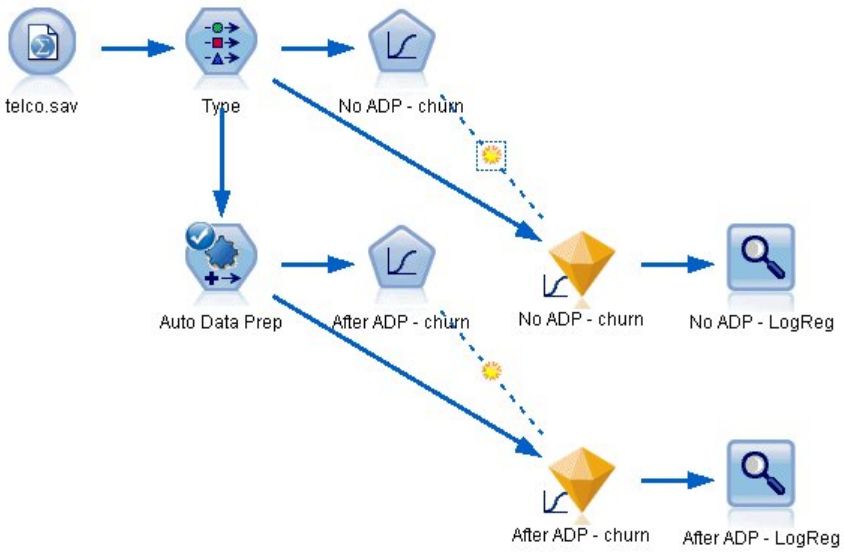


圖 *50.* 建置串流

2. 將「類型」節點連接至來源節點，將 *churn* 欄位的測量層次設為**旗標**，並將角色設為**目標**。所有其他欄位應該將其角色設為**輸入**。

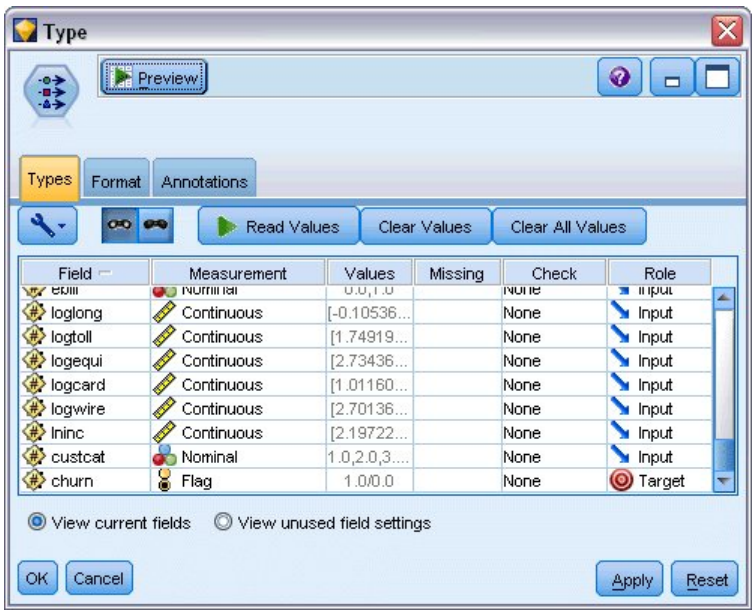


圖 *51.* 選取目標

3. 將「邏輯」節點連接至「類型」節點。
4. 在「邏輯」節點中，按一下「模型」標籤並選取**二項式**程序。在 *Model name* 欄位中，選取**自訂**並輸入 `No ADP - churn`。

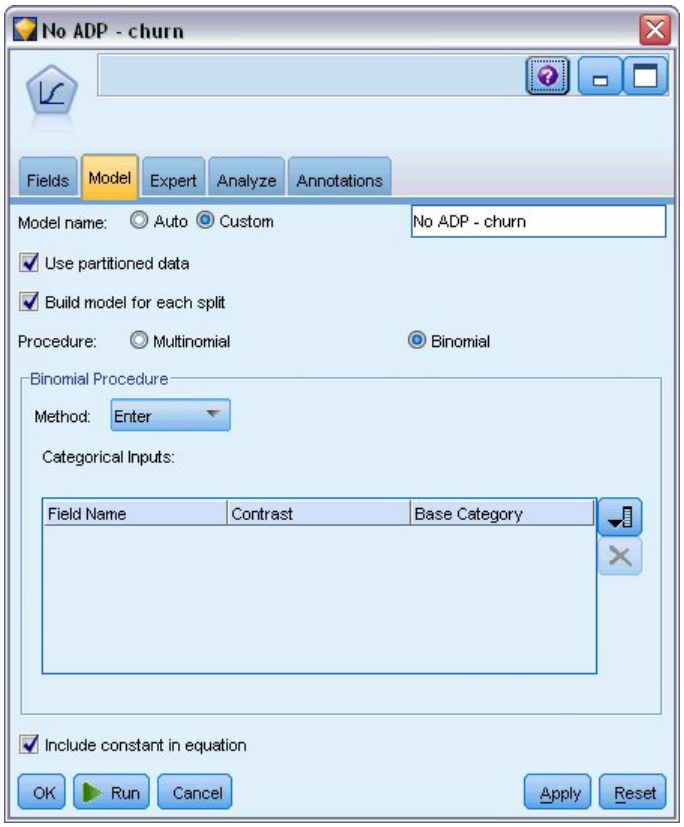


圖 *52.* 選擇模型選項

5. 將 ADP 節點連接至「類型」節點。在「目標」標籤上，原樣保留預設值以透過平衡速度和正確性來分析和準備資料。
6. 在「目標」標籤的頂端，按一下**分析資料**以分析和處理資料。

   ADP 節點上的其他選項可讓您指定您想要更專注於正確性和處理速度，或者細部調整多個資料準備處理步驟。

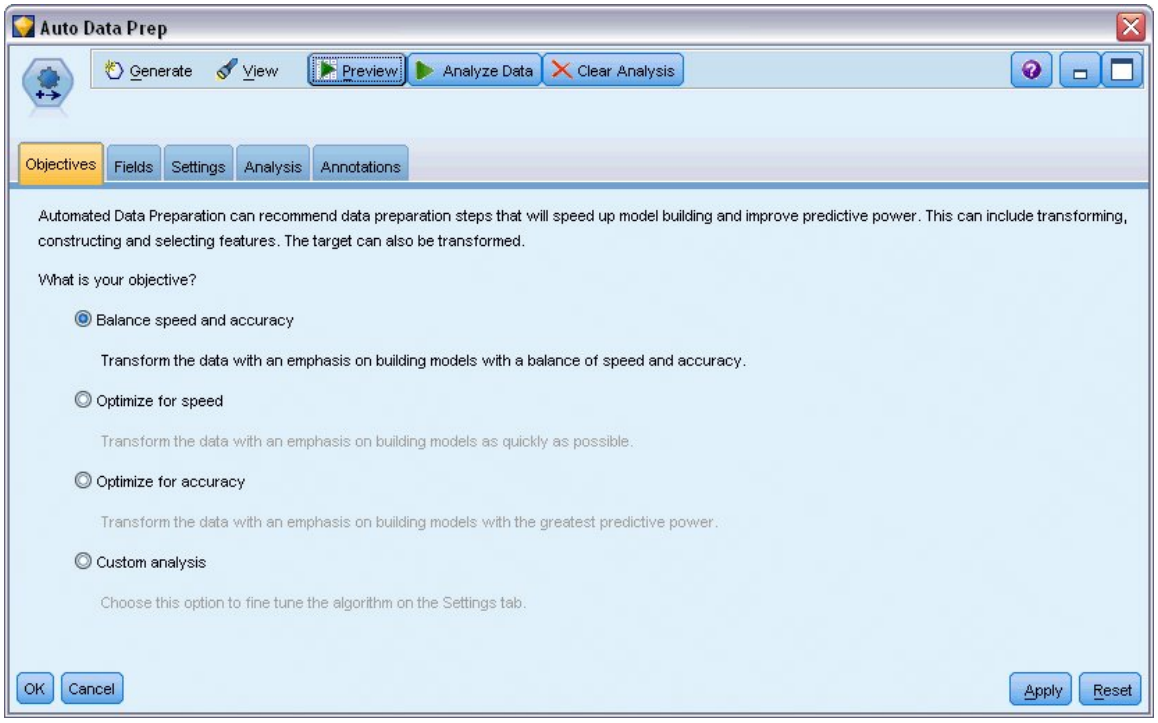


圖 *53. ADP* 預設目標

資料處理的結果顯示在「分析」標籤上。**欄位處理摘要**顯示有 41 個資料功能引入了 ADP 節點，已轉換 19 個以協助處理，3 個因未用而遭捨棄。

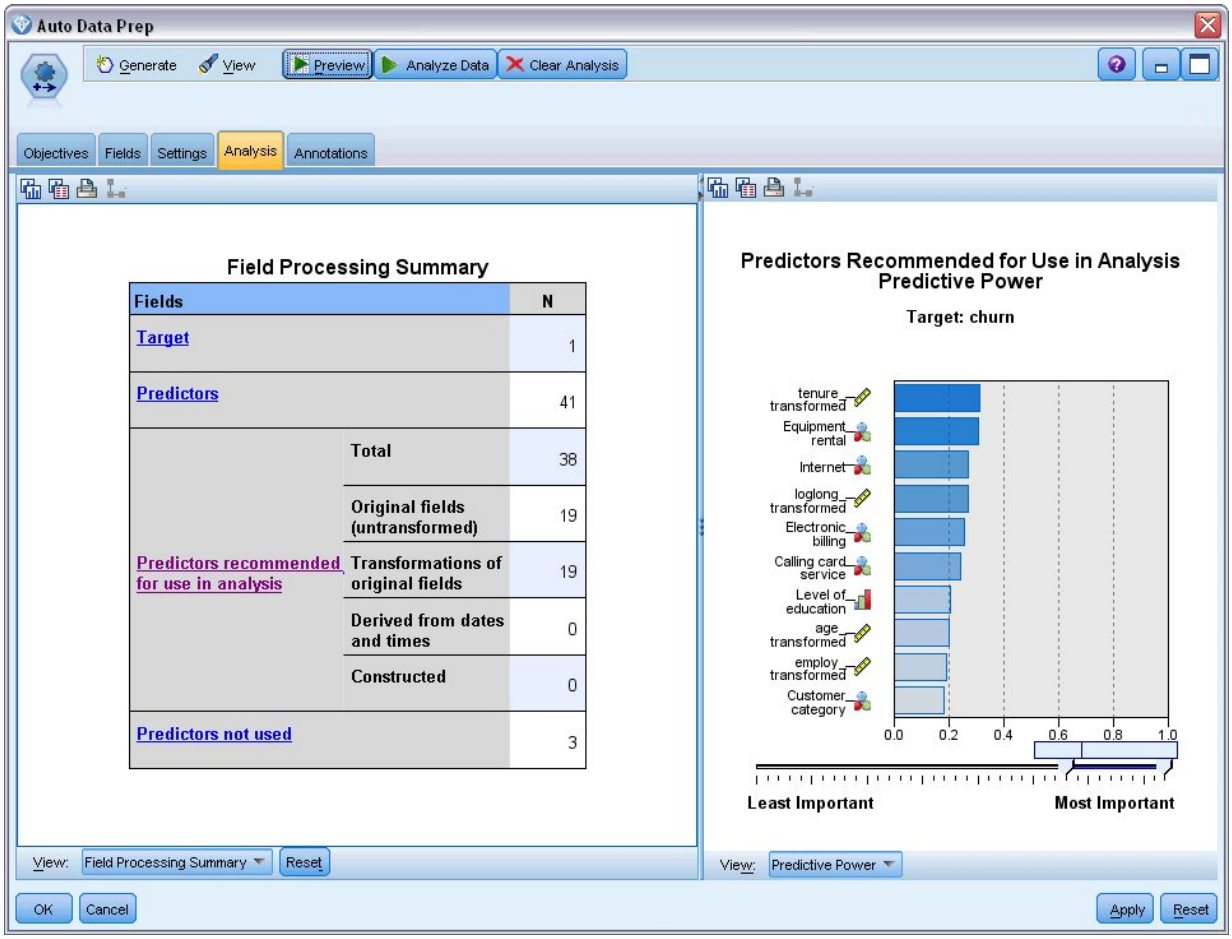


圖 *54.* 資料處理的摘要

7. 將「邏輯」節點連接至 ADP 節點。
8. 在「邏輯」節點中，按一下「模型」標籤並選取**二項式**程序。在 *Modeling name* 欄位中，選取**自訂**並輸入 `After ADP - churn`。

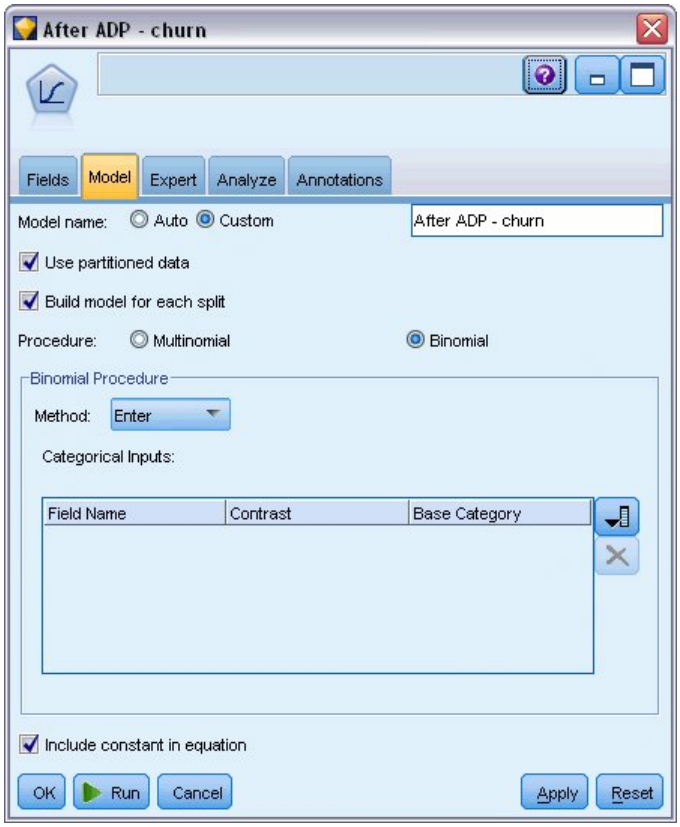


圖 *55.* 選擇模型選項

比較模型正確性

1. 執行兩個「邏輯」節點來建立模型區塊，模型區塊將會新增至串流以及右上角的「模型」選用區中。

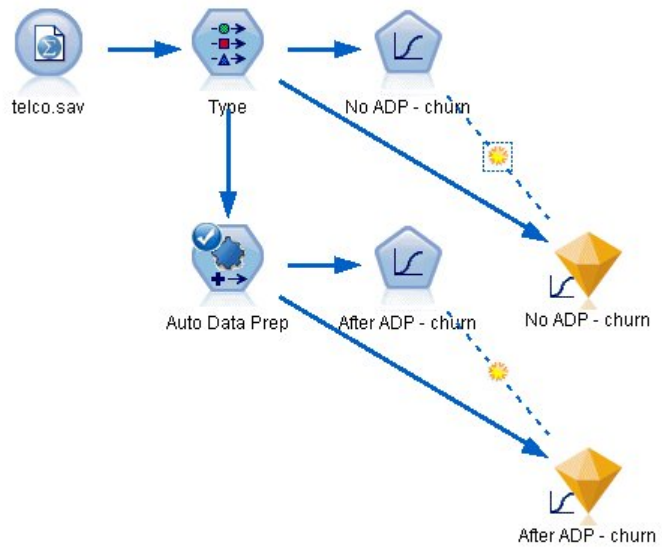


圖 *56.* 連接模型區塊

2. 將「分析」節點連接至模型區塊並使用其預設值來執行「分析」節點。

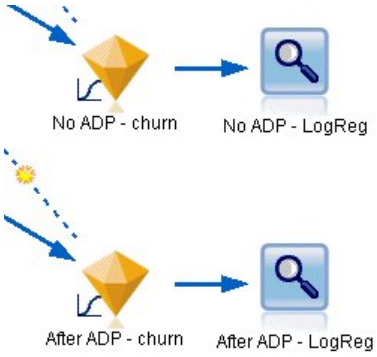


圖 *57.* 連接分析節點

分析非 ADP 衍生的模型會顯示，只透過「邏輯迴歸」節點並使用其預設值來執行資料會提供低正確性（只有 10.6%）的模型。

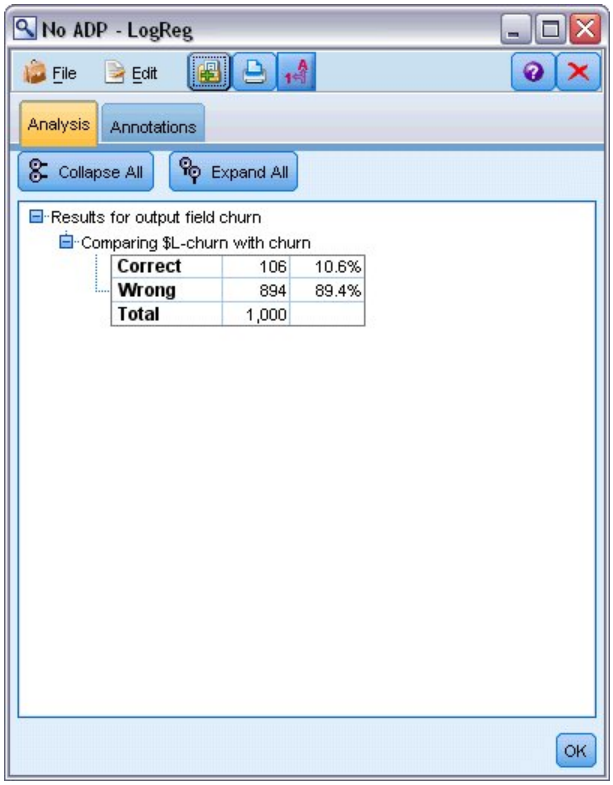


圖 *58.* 非 *ADP* 衍生的模型結果

分析 ADP 衍生的模型會顯示透過預設 ADP 設定執行資料，您建置的模型具有更高的正確性（正確性為 78.8%）。

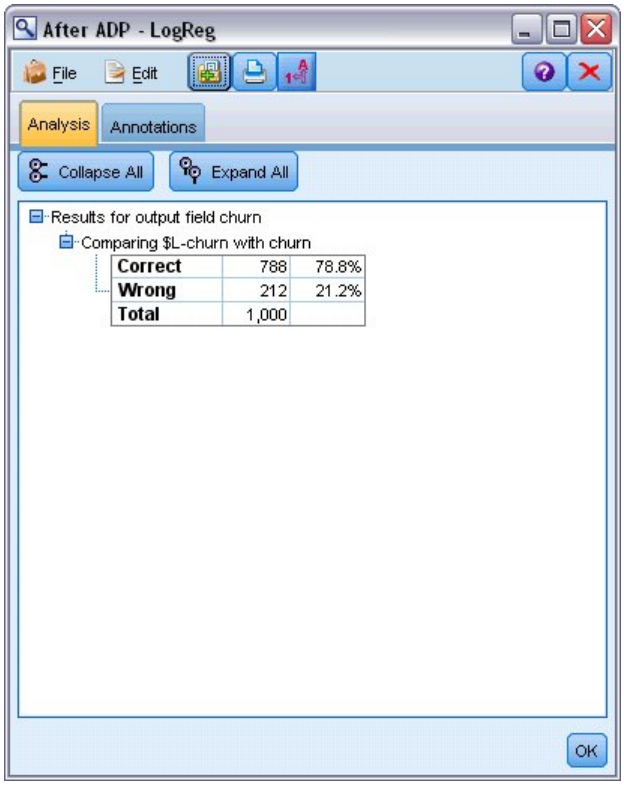


圖 *59. ADP* 衍生的模型結果

總而言之，僅透過執行 ADP 節點來細部調整資料的處理，您便能夠建置更精確的模型，無需直接操作資料。

顯然，如果您的興趣是證明或反證明某個理論，或想要建置特定的模型，您可能會發現直接使用模型設定有好處；但是，對於需要縮減時間量或需要準備大量資料的模型，ADP 節點可以提供優勢。

在《*IBM SPSS Modeler* 演算法手冊》中列出了在 IBM SPSS Modeler 中所使用建模方法的數學基礎說明，該手冊位於安裝磁碟的 *\Documentation* 目錄中。

請注意，本範例中的結果僅根據訓練資料得出。若要評量模型推廣到真實世界中的其他資料的程度，您可以使用「分割區」節點來送出一部分記錄用於測試和驗證。

第 7 章 準備用於分析的資料（資料審核）

透過「資料審核」節點，您可對帶入 IBM SPSS Modeler 的資料有個初步的全面瞭解。資料審核報告經常在起始資料探索期間使用，會顯示每一個資料欄位的彙總統計量以及直方圖和分佈圖，並容許您指定針對遺漏值、偏離值和極端值的處理方法。

此範例使用名為 *telco_dataaudit.str* 的串流，其參照的資料檔名為 *telco.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*telco_dataaudit.str* 檔案位於 *streams* 目錄中。

建置串流

1. 若要建置串流，請新增指向 *telco.sav*（位於 IBM SPSS Modeler 安裝的 *Demos* 目錄中）的「統計量檔案」來源節點。

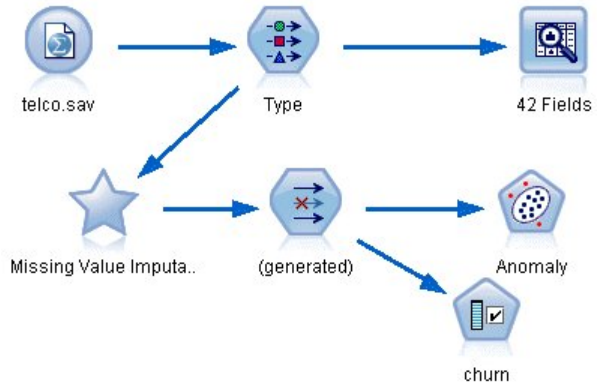


圖 *60.* 建置串流

2. 新增「類型」節點來定義欄位，並指定 *churn* 作為目標欄位（角色 = **目標**）。角色應針對所有其他欄位設定為**輸入**，這樣該欄位才會是唯一目標。

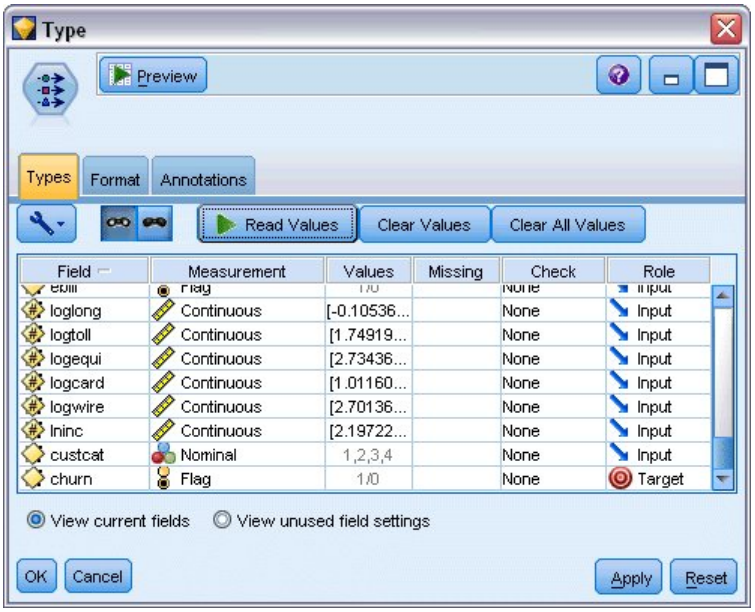


圖 *61.* 設定目標

3. 確認欄位測量層次定義正確。例如，包含值 0 及 1 的大部分欄位都可視為旗標，但特定欄位（例如，性別）會更精確地被視為包含兩個值的名義欄位。

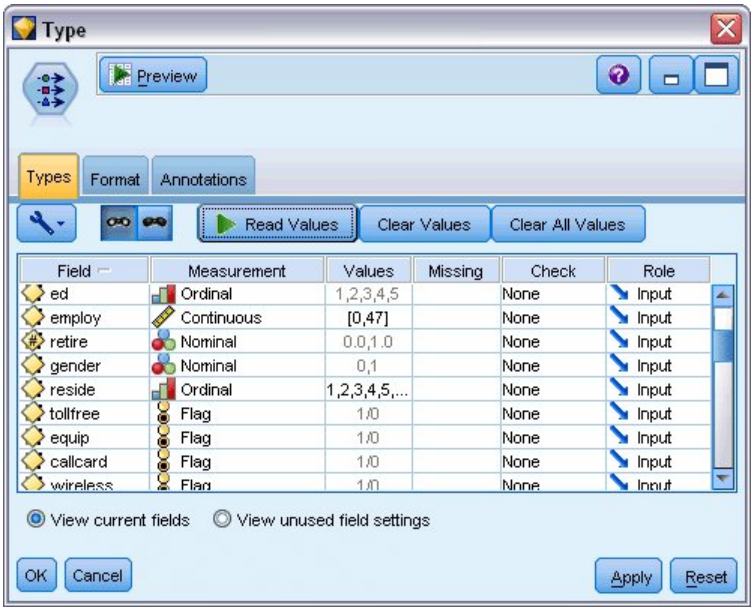


圖 *62.* 設定測量層次

提示：若要變更具有相似值（例如 0/1）的多個欄位的內容，請按一下值直欄標頭來依該直欄排序欄位，並使用 Shift 鍵選取要變更的所有欄位。然後您可以用滑鼠右鍵按一下選定內容來變更所有所選欄位的測量層次或其他屬性。

4. 將「資料審核」節點連接至串流。在「設定」標籤上，保留預設值不變以在報告中包含所有欄位。因為流失是「類型」節點中定義的唯一目標欄位，所以會將其自動用作套版。

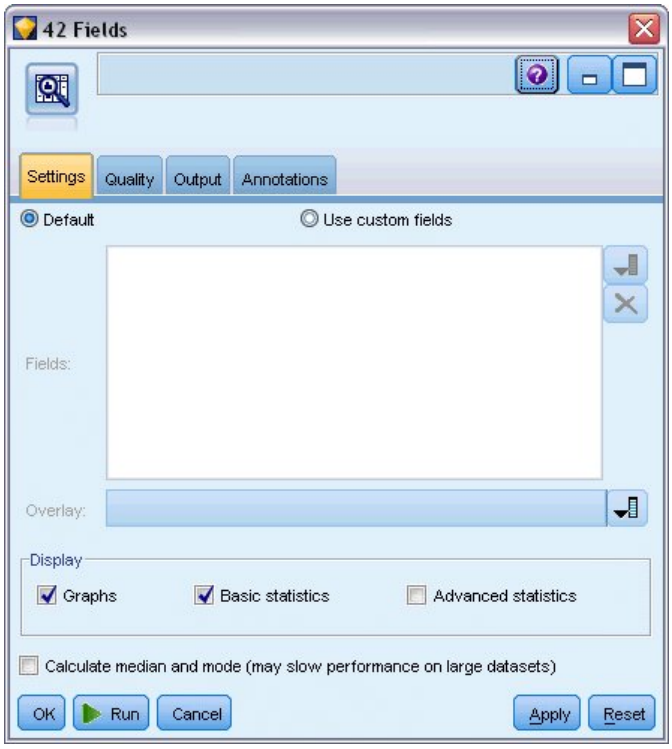


圖 *63.* 資料審核節點，設定標籤

在「品質」標籤上，保留用來偵測遺漏值、偏離值和極端值的預設值不變，並按一下**執行**。

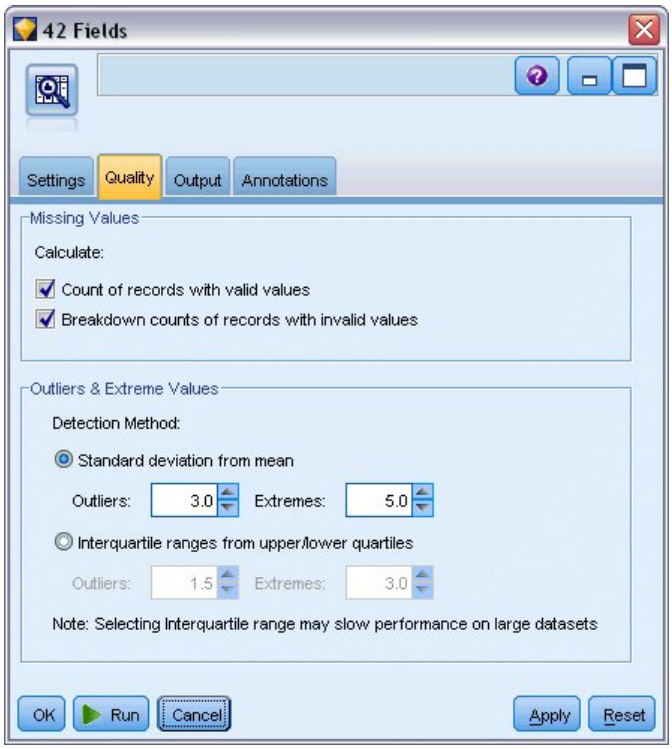


圖 *64.* 資料審核節點，品質標籤

瀏覽統計量及圖表

會顯示「資料審核」瀏覽器，其中包含每一個欄位的縮圖及描述性統計量。

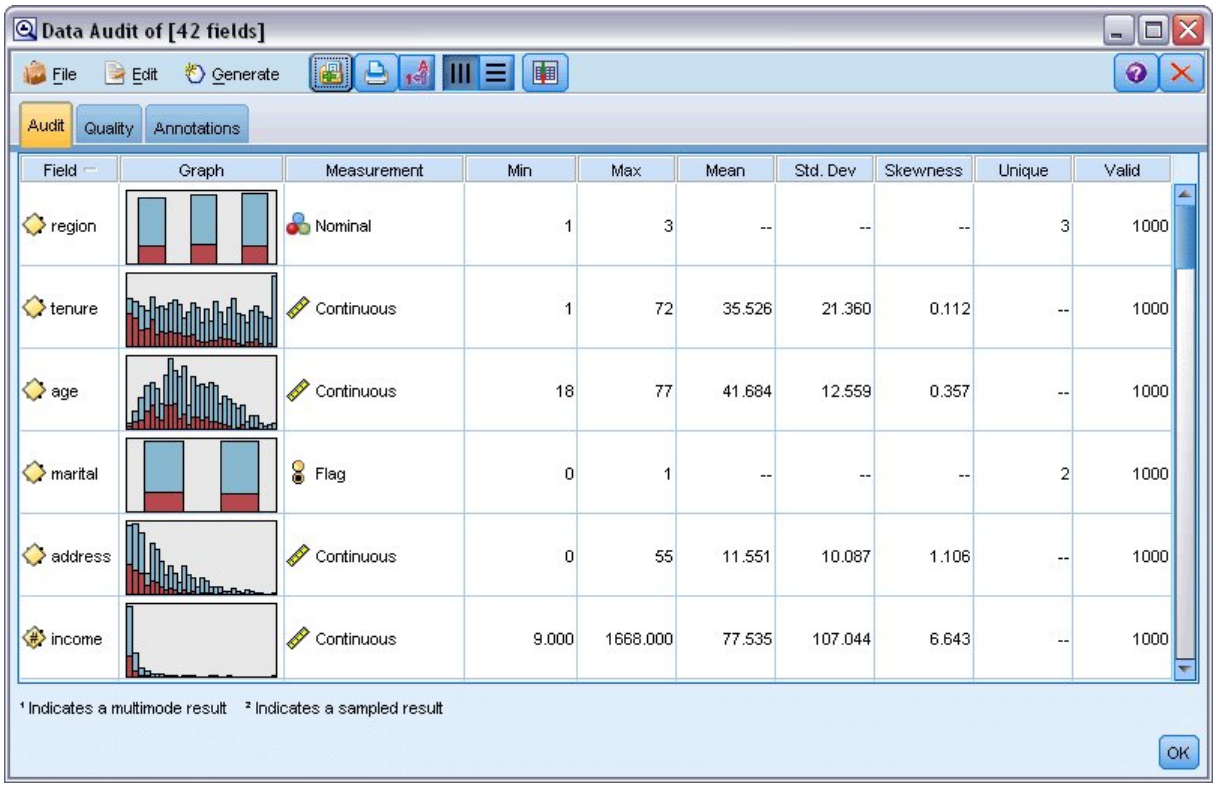


圖 *65.* 資料審核瀏覽器

使用工具列顯示欄位和值標籤，並將圖表的對齊方式從水平切換到垂直（僅針對種類欄位）。

1. 您也可以使用工具列或「編輯」功能表來選擇要顯示的統計量。

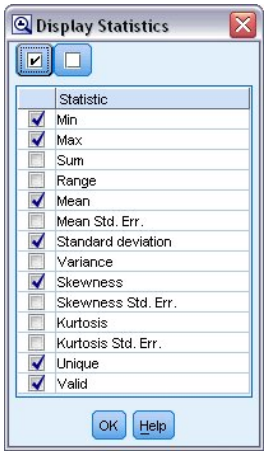


圖 *66.* 顯示統計量

按兩下審核報告中的任何縮圖可檢視該圖表的最大版本。因為流失是串流中唯一的目標欄位，所以會將其自動用作套版。您可以使用圖形視窗工具列切換欄位和值標籤的顯示，也可以按一下「編輯」模式按鈕來進一步自

訂圖表。

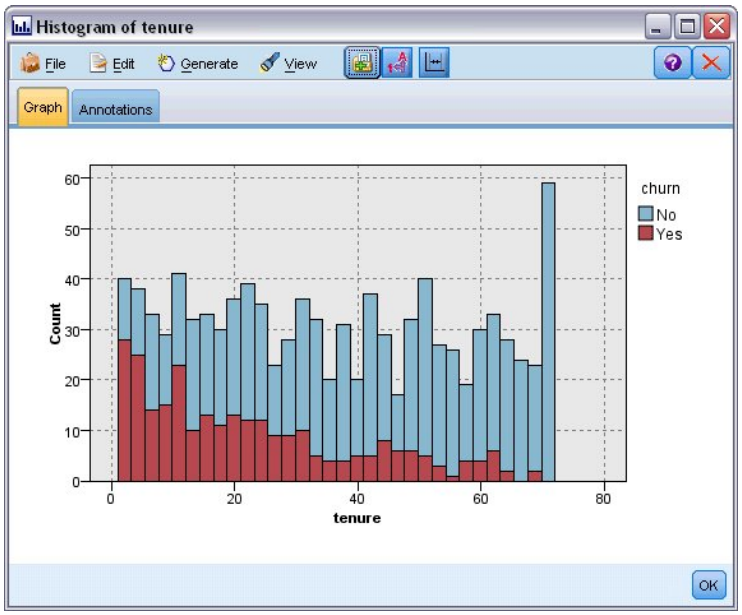


圖 *67.* 保有期直方圖

或者，您可以選取一個以上縮圖並針對每一個產生「圖形」節點。產生的節點放在串流畫布上，並可以新增至串流來重建該特定圖形。

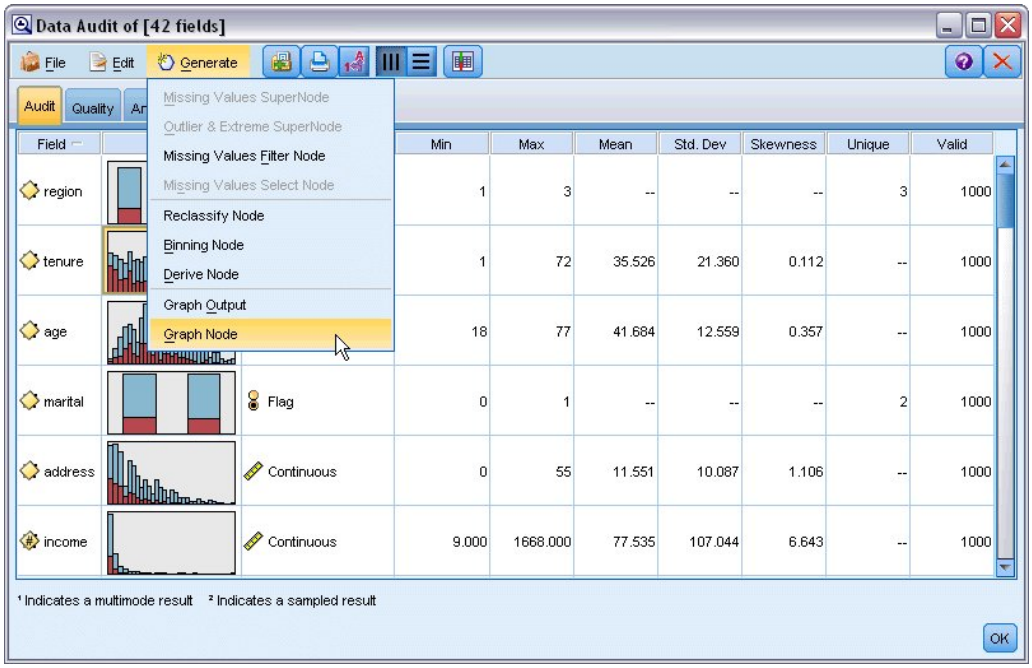


圖 *68.* 產生圖形節點

處理偏離值與遺漏值

審核報告中的「品質」標籤顯示偏離值、極端值和遺漏值的相關資訊。

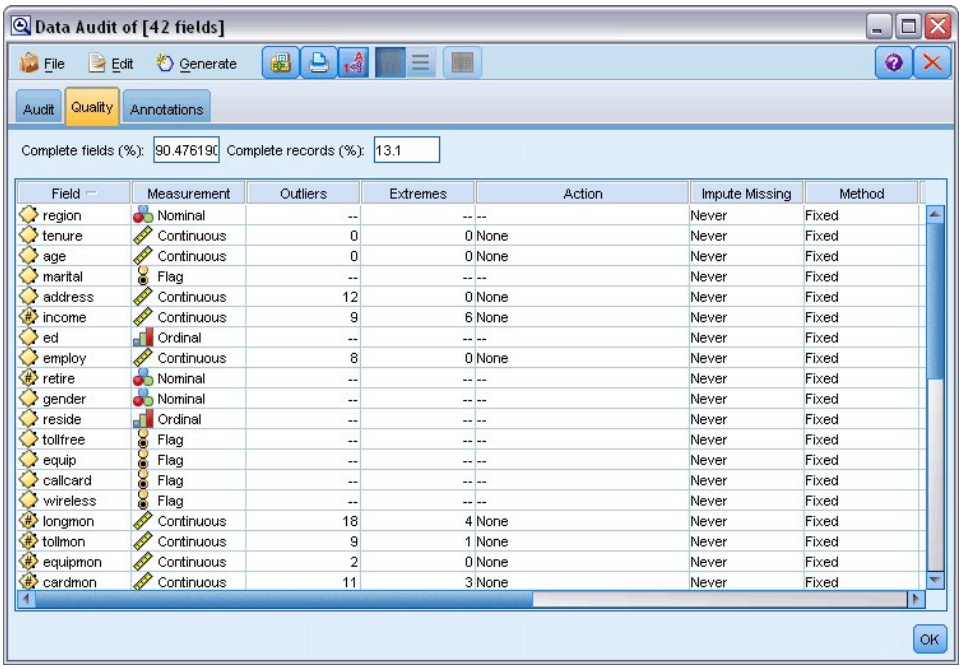


圖 *69.* 資料審核瀏覽器，品質標籤

您還可以指定用來處理這些值的方法，並產生 SuperNode 來自動套用轉換。例如，您可以選取一個以上欄位並選擇使用大量方法（包括 C&RT 演算法）插補或取代這些欄位的遺漏值。

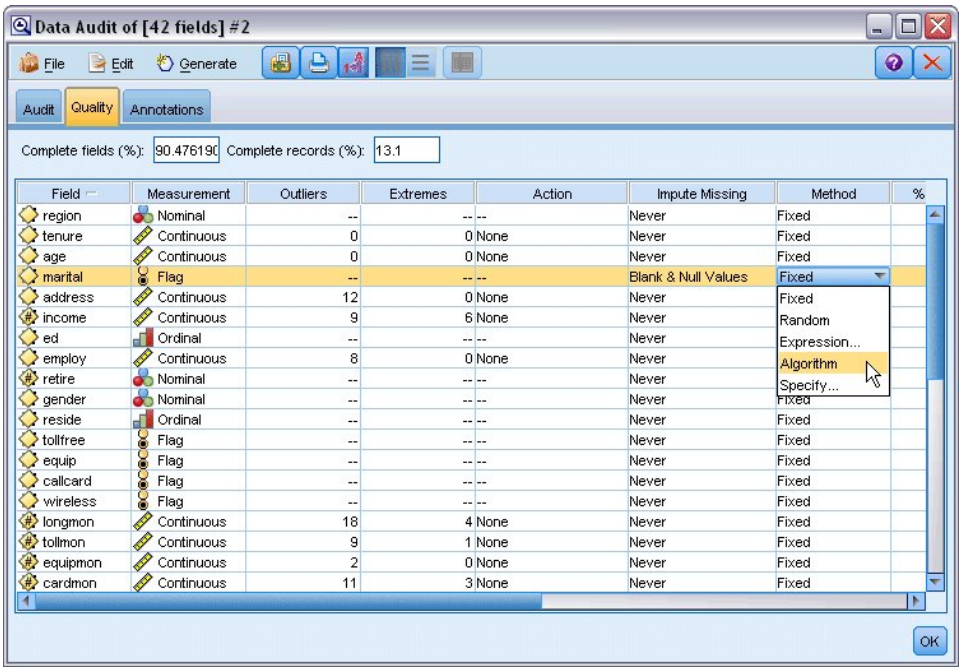


圖 *70.* 選擇插補方法

為一個以上欄位指定插補方法之後，若要產生「遺漏值 SuperNode」，請從功能表中選擇：

產生 > 遺漏值 SuperNode

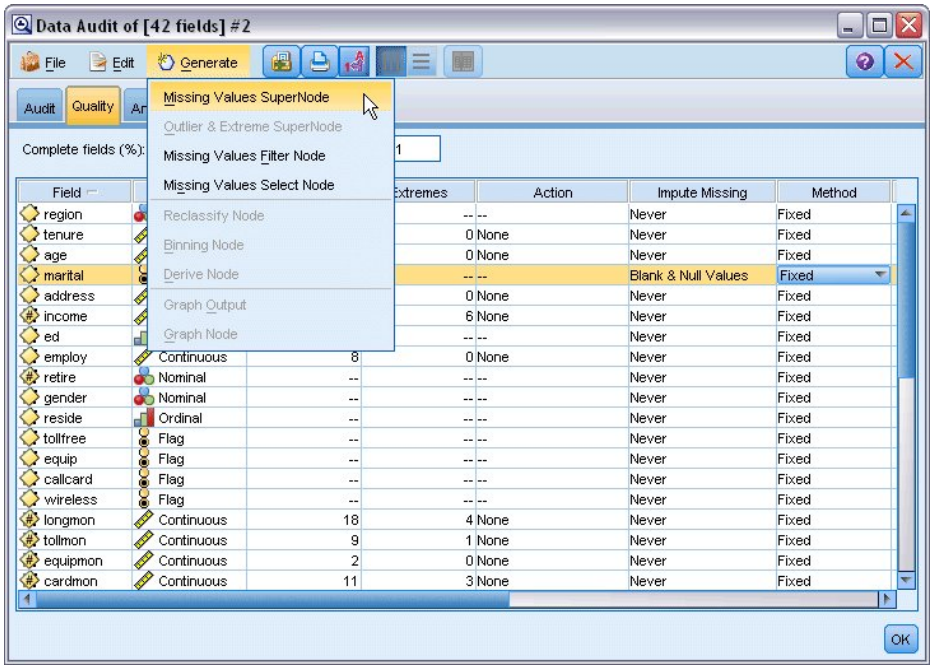


圖 *71.* 產生 *SuperNode*

產生的 SuperNode 會新增至串流畫布，其中您可以將其連接至串流以套用轉換。

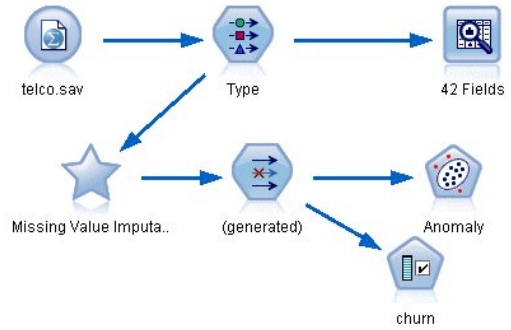


圖 *72.* 包含遺漏值 *SuperNode* 的串流

SuperNode 實際包含一系列執行所要求轉換的節點。若要瞭解如何運作，您可以編輯 SuperNode 並按一下**放大**。

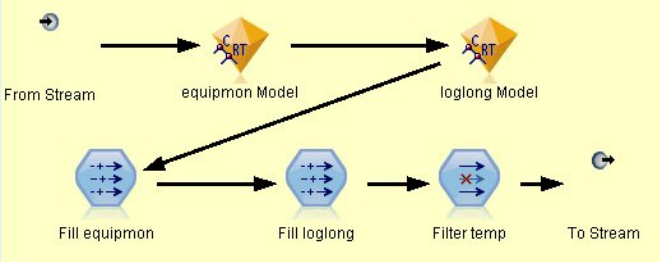


圖 *73.* 放大 *SuperNode*

對於使用演算法插補的每一個欄位，例如，都會有一個個別的 C&RT 模型，還有一個「填入器」節點會使用該模型預測的值取代空白值和空值。您可以新增、編輯或移除 SuperNode 內的特定節點來進一步自訂行為。

或者，您可以產生「選取」或「過濾器」節點來移除具有遺漏值的欄位或記錄。例如，您可以過濾品質百分比低於指定臨界值的任何欄位。

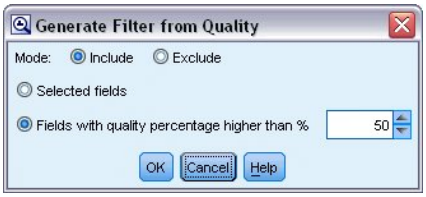


圖 *74.* 產生過濾器節點

偏離值和極端值可透過相似的方式進行處理。指定您要針對每一個欄位採取的動作 — 強制轉型、捨棄或設為空值 — 並產生 SuperNode 以套用轉換。

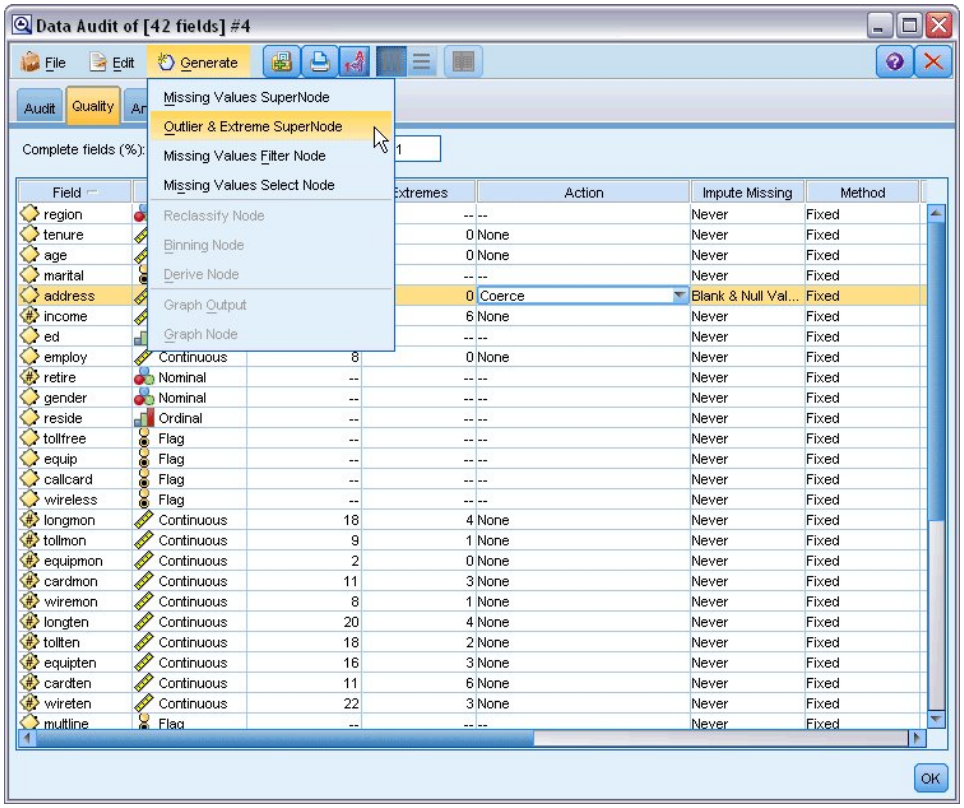


圖 *75.* 產生過濾器節點

完成審核並將產生的節點新增至串流之後，您可以繼續分析。此外，您可能想要使用「異常偵測」、「功能選擇」或多種其他方法進一步篩選資料。

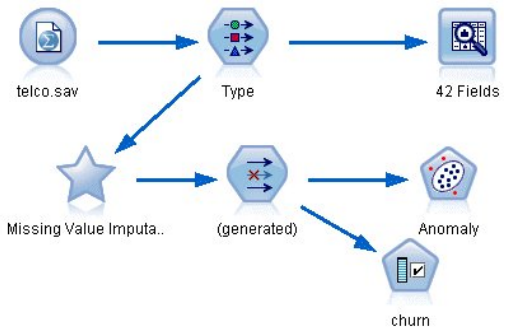


圖 *76.* 包含遺漏值 *SuperNode* 的串流

第 8 章 藥品治療（指數圖形/C5.0）

在這個小節中，假設您是編譯資料進行研究的醫療研究人員。您收集了一組罹患同一種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。您在工作過程中使用資料採礦來找出哪種藥品可能適合未來罹患同一種病的病患。

此範例使用參照資料檔 *DRUG1n* 的串流 *druglearn.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*druglearn.str* 檔位於 *streams* 目錄。

用於展示的資料欄位如下：

資料欄位	說明
年齡	（數字）
性別	*M* 或 *F*
BP	血壓：*HIGH*、*NORMAL* 或 *LOW*
膽固醇	血液膽固醇：*NORMAL* 或 *HIGH*
Na	血鈉濃度
K	血鉀濃度
藥物	病患產生反應的處方藥

在文字資料中讀取

圖 *77.* 新增變數檔案節點

您可以使用**變數檔案節點**在定界文字資料中讀取。您可以從選用區新增「變數檔案」節點--按一下**來源**標籤來尋找節點或使用**我的最愛**標籤（依預設，其中包括此節點）。接下來按兩下新放置的節點以開啟其對話框。

只需按一下標示了省略符號 (...) 的檔案方框右側的按鈕來瀏覽至您系統上安裝了 IBM SPSS Modeler 的目錄。開啟 *Demos* 目錄並選取檔案 *DRUG1n*。

確保選取了**從檔案中讀取欄位名稱**，請注意剛載入到對話框中的欄位和值。

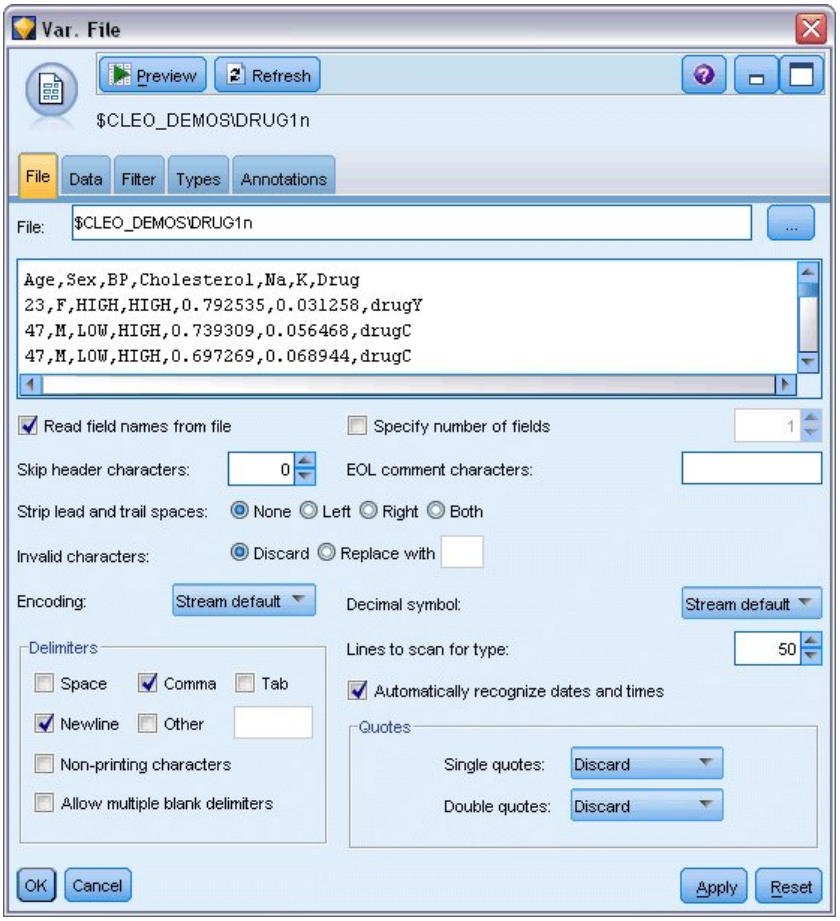


圖 *78.* 變數檔案對話框

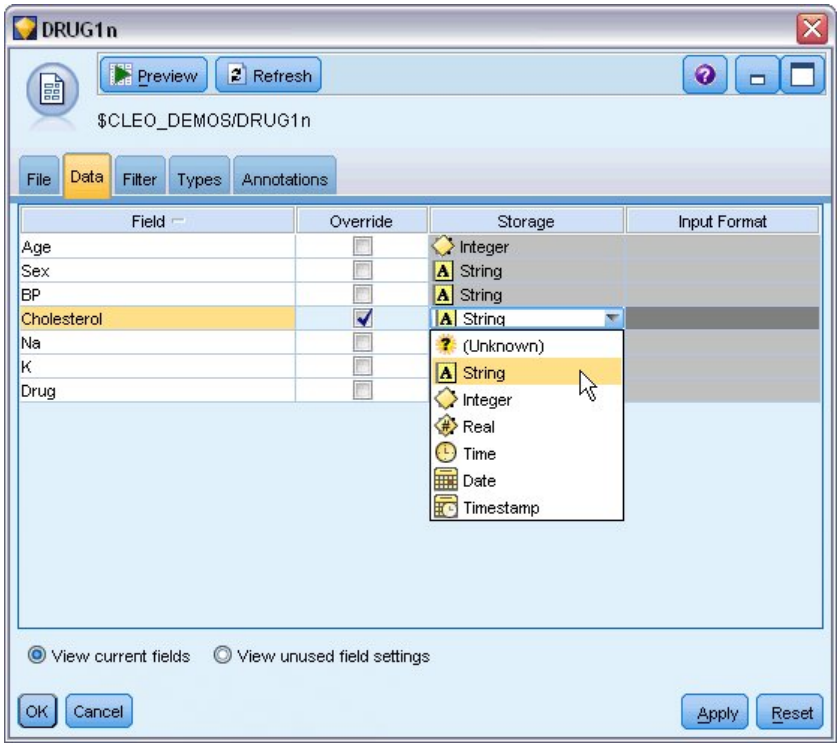


圖 *79.* 變更欄位的儲存體類型

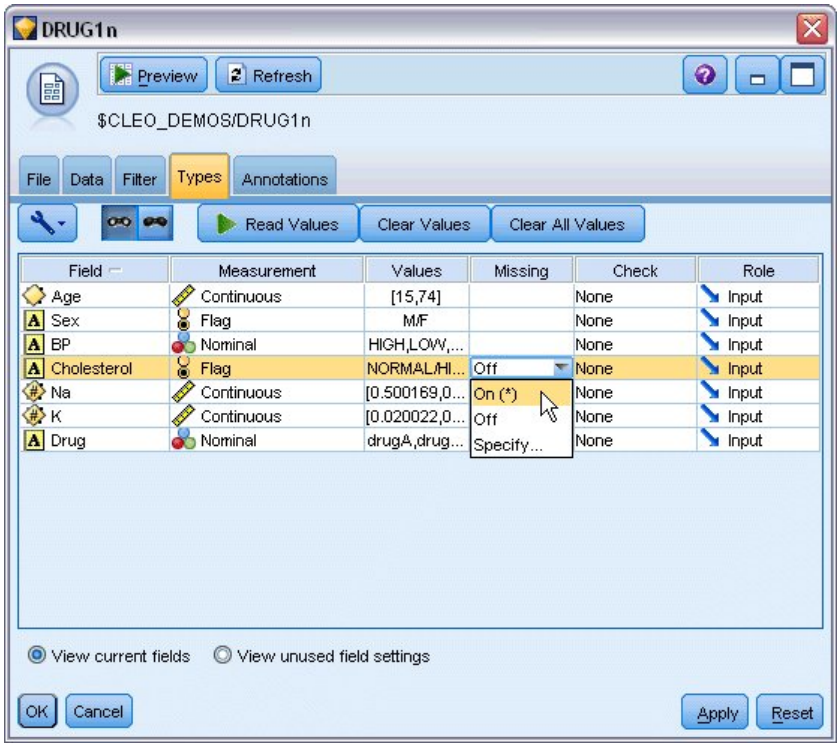


圖 *80.* 在類型標籤上選取值選項

按一下**資料**標籤以置換和變更欄位的**儲存體**。附註，儲存體不同于**測量**，亦即資料欄位的測量層次（或用量類型）。**類型**標籤可協助您進一步瞭解資料中的欄位類型。您也可以選擇**讀取值**以根據您從值 直欄進行的選擇來檢視每個欄位的實際值。此程序稱為**實例化**。

新增表格

現在您已載入資料檔案，您可能想要大致瞭解部分記錄的值。執行此動作的一個方法是建置包括「表格」節點的串流。若要在串流中放置「表格」節點，請按兩下選用區中的圖示或將圖示拖放至畫布。

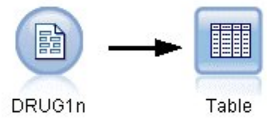


圖 *81.* 連接至資料來源的表格節點

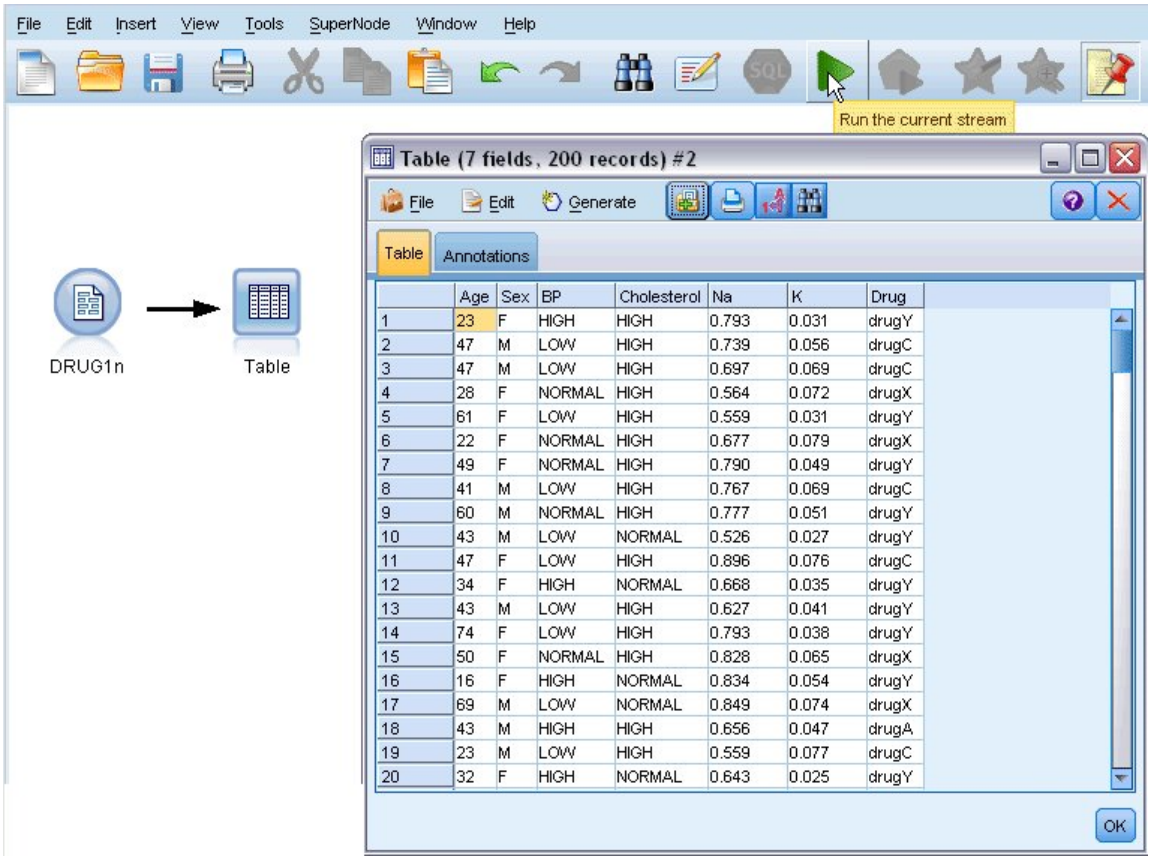


圖 *82.* 從工具列執行串流

按兩下選用區中的節點將自動將它連接至串流畫布中的所選節點。或者，如果節點尚未連接，您可以使用中間滑鼠按鈕將「來源」節點連接至「表格」節點。若要模擬中間滑鼠按鈕，請在使用滑鼠時按住 Alt 鍵。若要檢視表格，請按一下工具列上的綠色方向鈕來執行串流，或在「表格」節點上按一下滑鼠右鍵然後選擇**執行**。

建立分佈圖形

在資料採礦期間，透過建立視覺化摘要來探索資料通常很有用。IBM SPSS Modeler 提供數種類型的圖形供您選擇，視您要彙總的資料種類而定。例如，若要找出對每種藥品產生反應的病患比例，可使用「分佈」節點。

將「分佈」節點新增至串流並將它連接至「來源」節點，然後按兩下該節點以編輯要顯示的選項。

選取藥品作為您想要顯示其分佈的目標欄位。然後從對話框中按一下**執行**。

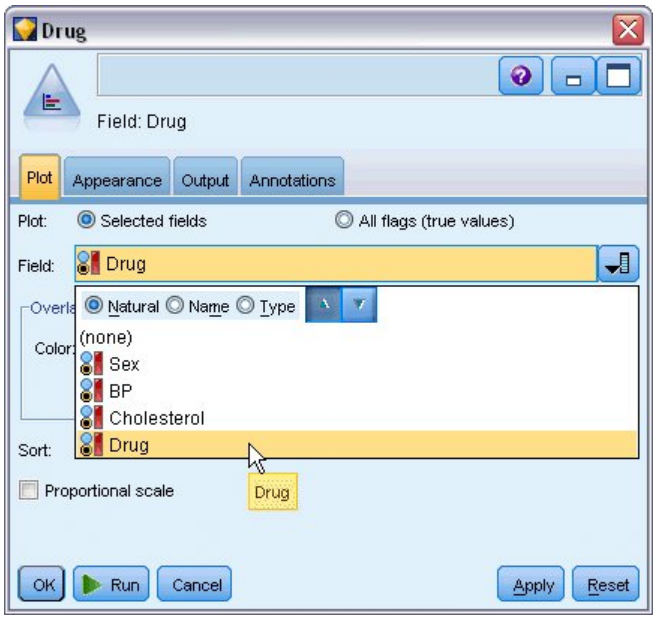


圖 *83.* 選取藥品作為目標欄位

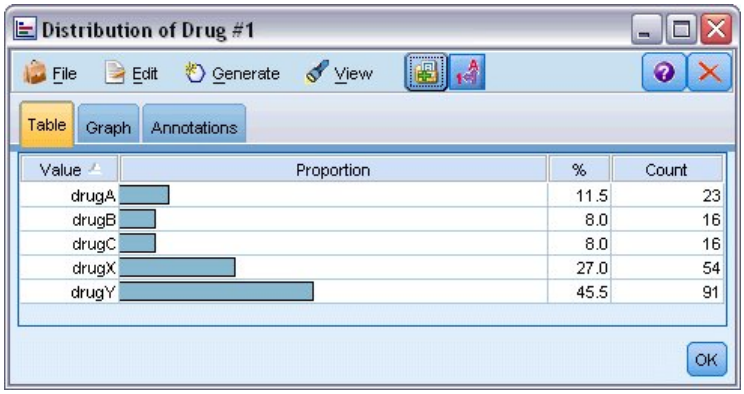


圖 *84.* 對藥品類型產生反應的分佈

產生的圖形協助您查看資料的「形狀」。它顯示病患對藥品 *Y* 經常產生反應，而對藥品 *B* 和 *C* 產生反應的次數最少。

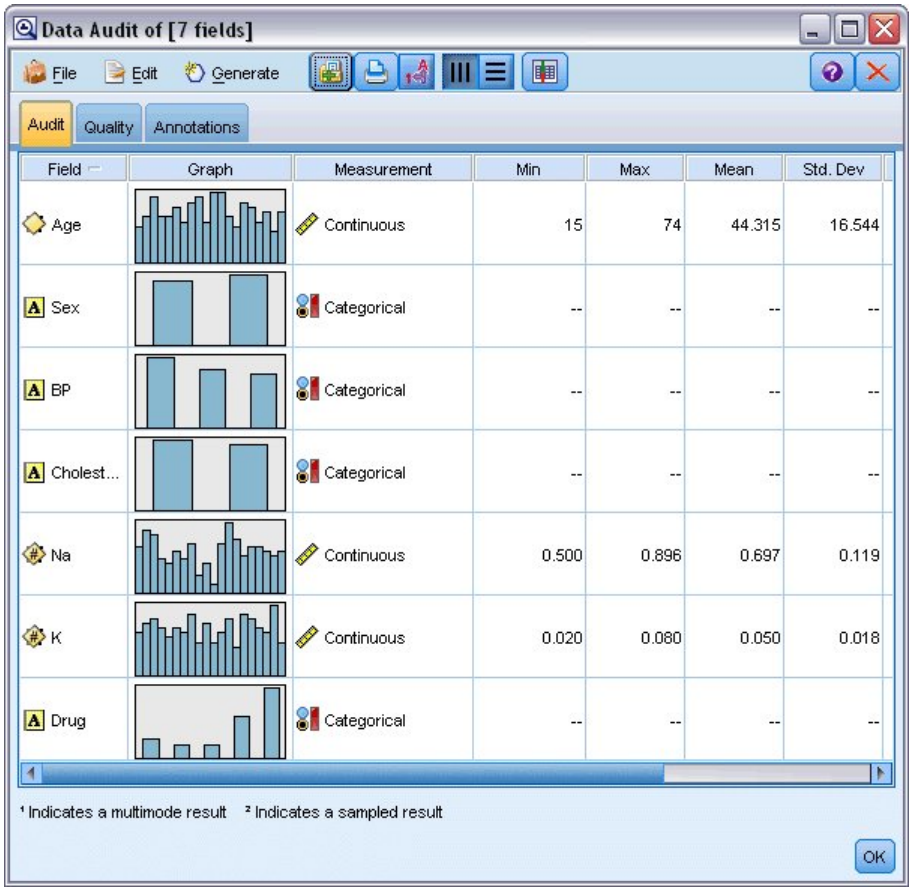


圖 *85.* 資料審核的結果

或者，您可以連接並執行「資料審核」節點以立即快速查看所有欄位的分佈和直方圖。「資料審核」節點在「輸出」標籤上有提供。

建立散佈平面圖

現在，讓我們看看哪些因素可能會影響藥品，亦即目標變數。作為研究人員，您知道血液中鉀和鈉的濃度是重要的因素。由於這兩個都是數值，因此您可以建立鉀和鈉的散佈平面圖，並使用藥品種類作為顏色套版。

在工作區中放置一個「圖形」節點，然後將它連接至「來源」節點並按兩下以編輯該節點。

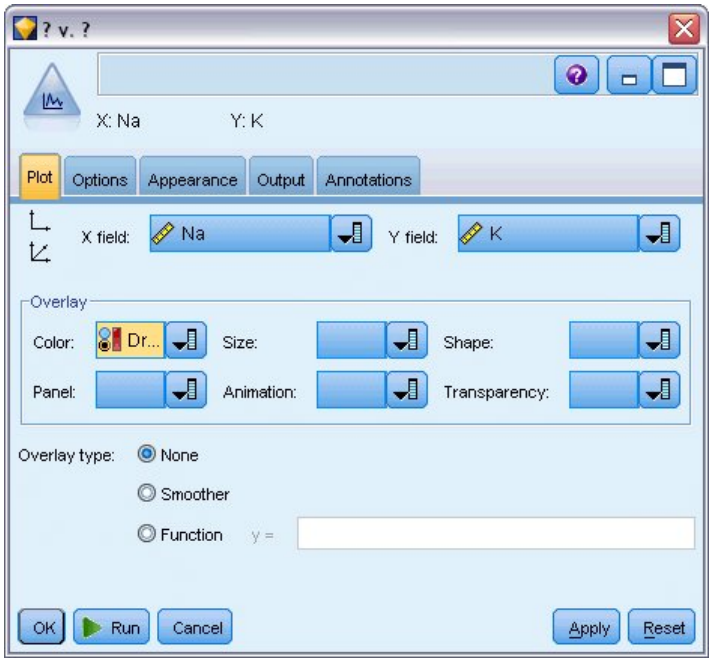


圖 *86.* 建立散佈平面圖

在「圖形」標籤上，選取 *Na* 作為 X 欄位，選取 *K* 作為 Y 欄位，選取 *Drug* 作為套版欄位。然後按一下**執行**。

圖形清楚地顯示在臨界值上方正確的藥品一律為藥品 *Y*，而在臨界值之下正確的藥品永不為藥品 *Y*。此臨界值是一個比例，亦即鈉 (*Na*) 與鉀 (*K*) 的比例。

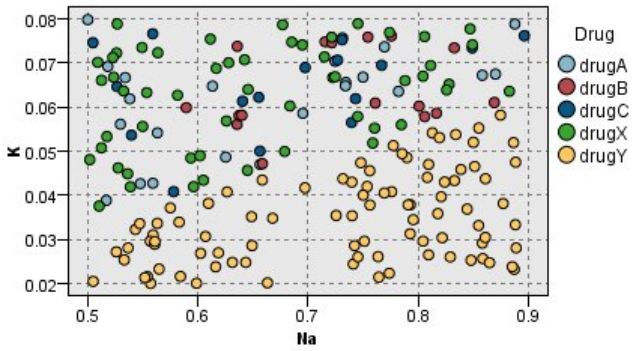


圖 *87.* 藥品分佈的散佈平面圖

建立 Web 圖形

由於許多資料欄位是種類欄位，因此您也可以嘗試繪制 Web 圖形，以對映不同種類之間的關聯。開始時請在工作區中將 Web 節點連接至來源節點。在「Web 節點」對話框中，選取 *BP*（表示血壓）和 *Drug*。然後按一下**執行**。

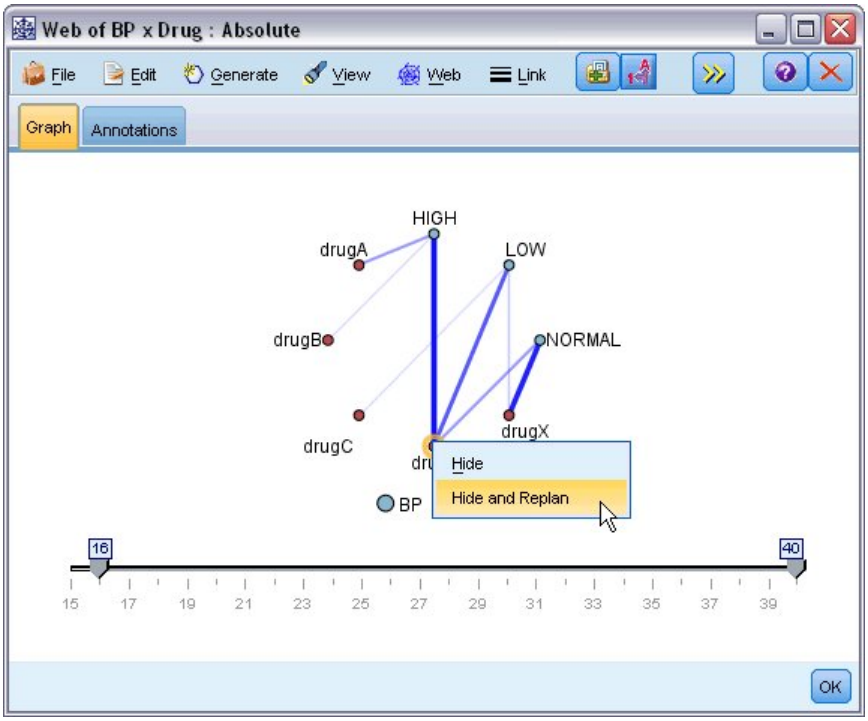


圖 *88.* 藥品和血壓的 *Web* 圖形

圖中顯示，藥品 *Y* 與所有三個血壓層次相關聯。這並不驚訝--您已確定這種情況下藥品 *Y* 為最佳。若要關注其他藥品，您可以隱藏藥品 *Y*。在**視圖**功能表上，選擇**編輯模式**，然後在藥品 *Y* 點上按一下滑鼠右鍵並選擇**隱藏並重新規劃**。

在簡化的圖形中，會隱藏藥品 *Y* 及其所有鏈結。現在，您可以清楚地看到只有藥品 *A* 和 *B* 與高血壓相關聯。只有藥品 *C* 和 *X* 與低血壓相關聯。正常血壓僅與藥品 *X* 相關聯。但是，在這種情況下，您仍然不知道對於給定的病患，如何在藥品 *A* 和 *B* 或藥品 *C* 和 *X* 之間選擇。這種情況下可以使用建模來協助您。

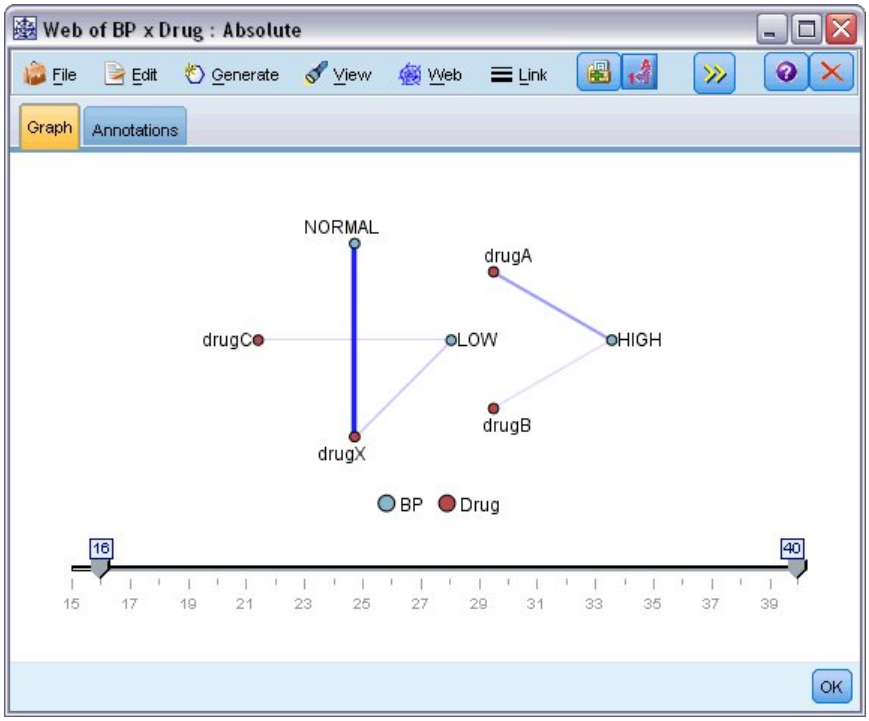


圖 *89.* 隱藏了藥品 *Y* 及其鏈結的 *Web* 圖形

衍生新欄位

由于在使用藥品 *Y* 時似乎可以預測納與鉀的比例，您可以衍生一個欄位來包含每筆記錄的這個比例值。當您建置模型以預測何時使用這五種藥品中的每一種時，此欄位可能會很有用。若要簡化串流佈置，請從刪除 DRUG1n 來源節點以外的所有節點開始。將「衍生」節點（欄位作業標籤）連接至 DRUG1n，然後按兩下「衍生」節點以編輯它。

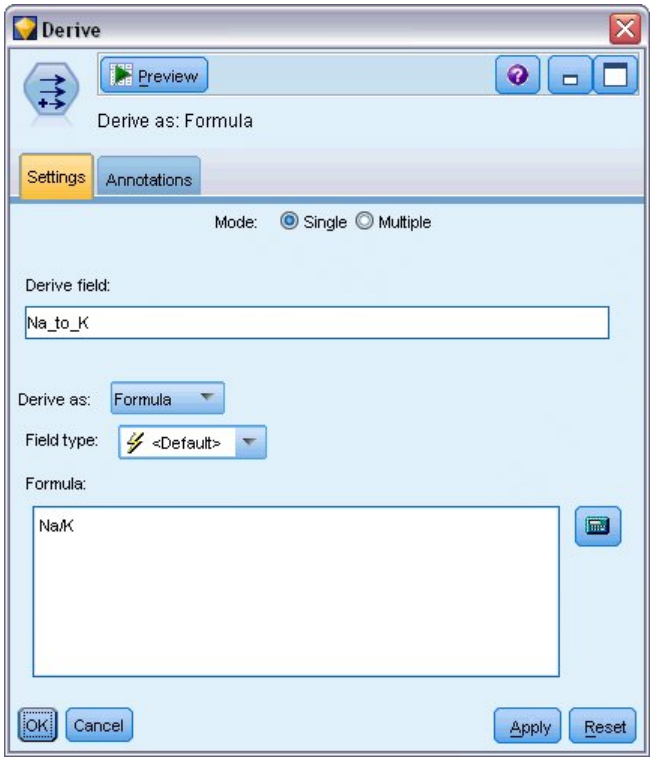


圖 *90.* 編輯衍生節點

將新欄位命名為 *Na_to_K*。由于您是將納值除以鉀值來取得新欄位，因此請為公式輸入 `Na/K`。您也可以透過按一下欄位右側的圖示來建立公式。這時會開啟「表示式建置器」，這是一種使用內建的函數、運算元、欄位及其值清單以互動方式建立表示式的方式。

您可以透過將「直方圖」節點連接至「衍生」節點來檢查新欄位的分佈。在「直方圖」節點對話框中，指定 *Na_to_K* 作為要繪制的欄位，指定 *Drug* 作為套版欄位。

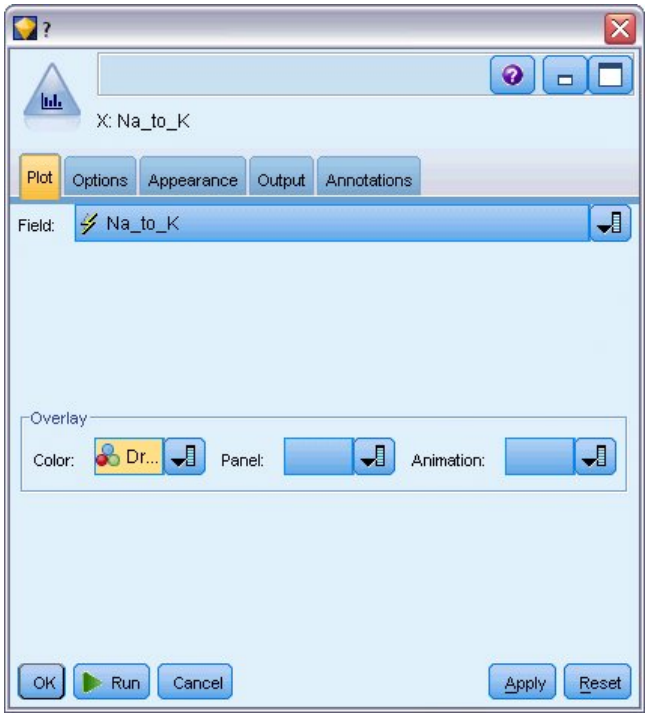


圖 *91.* 編輯直方圖節點

執行串流時，可取得這裡所顯示的圖形。根據顯示，您可以得出當 *Na_to_K* 值為大約 15 或以上時，藥品 *Y* 是要選擇的藥品。

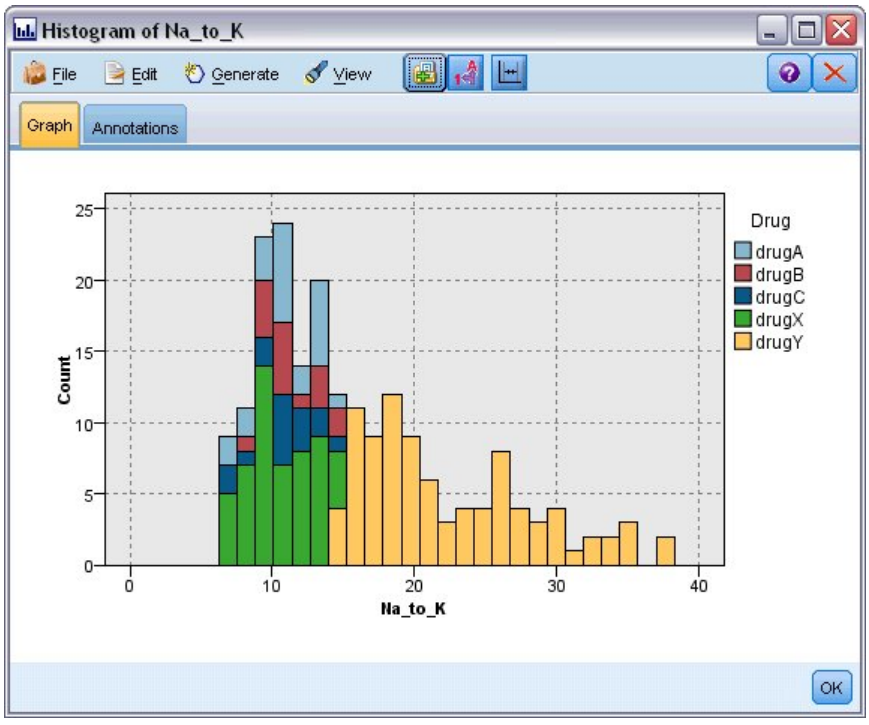


圖 *92.* 直方圖顯示

建置模型

透過探索和操作資料，您已經能夠形成部分假設。血液中的納與鉀的比例似乎會影響藥品的選擇，血壓也一樣。但是您尚且無法完整說明所有關係。這種情況下採用建模可能會提供一些答案。在這個案例中，您將嘗試使用規則建置模型 C5.0 來適合資料。

由於您使用衍生欄位 *Na_to_K*，因此您可以濾出原始欄位 *Na* 和 *K*，使其不要在建模演算法中使用兩次。您可以使用「過濾器」節點來執行此動作。

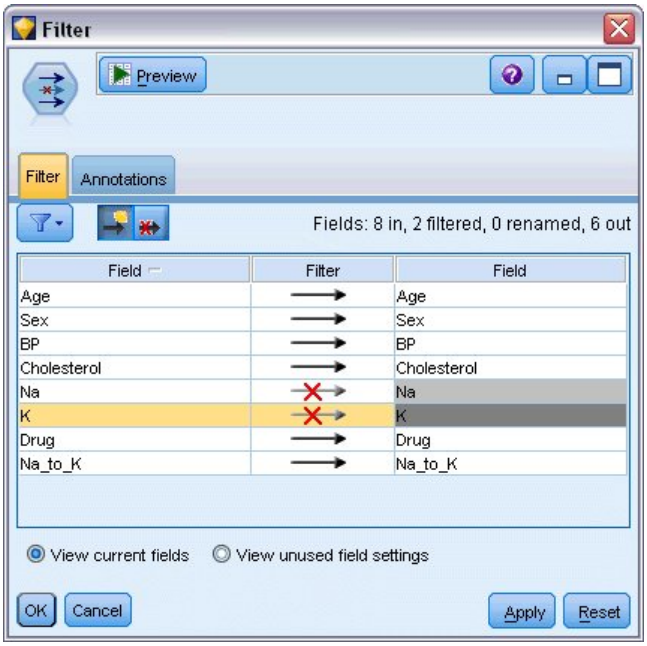


圖 *93.* 編輯過濾器節點

在「過濾器」標籤上，按一下 *Na* 和 *K* 旁邊的箭頭。紅色 Xs 出現在箭頭上方，表示現在已濾出欄位。

接下來，將「類型」節點連接至「過濾器」節點。「類型」節點可讓您指出所使用的欄位類型，以及如何使用這些欄位來預測結果。

在「類型」標籤上，將 *Drug* 欄位的角色設為**目標**，表示 *Drug* 是您要預測的欄位。保留其他欄位的角色設為**輸入**，以便將它們用作預測工具。

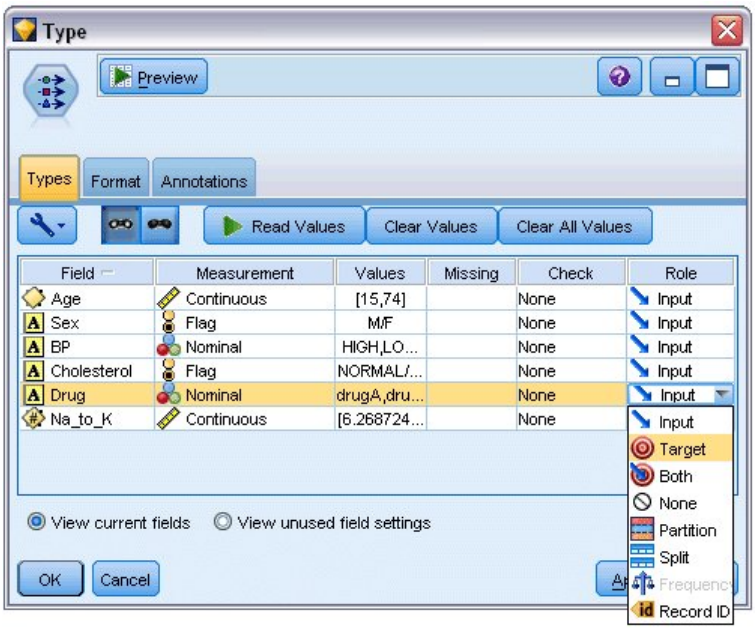


圖 *94.* 編輯類型節點

若要估計模型，請在工作區中放置 C5.0 節點並將它連接至所顯示的串流結尾。然後，按一下綠色**執行**工具列按鈕來執行串流。

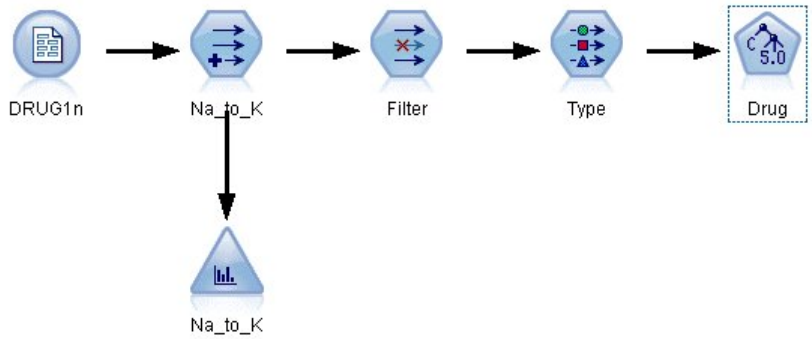


圖 *95.* 新增 *C5.0* 節點

瀏覽模型

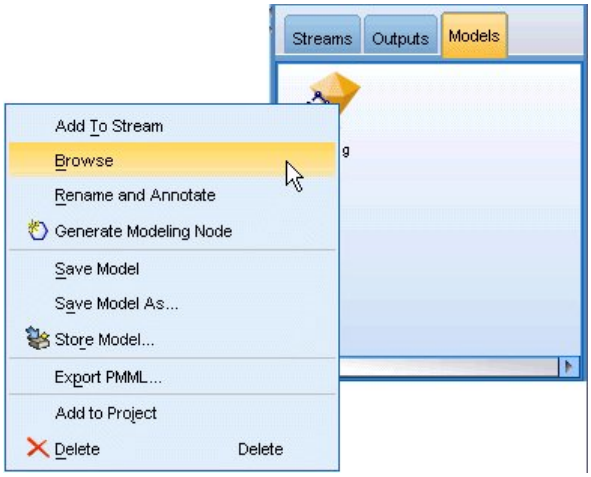


圖 *96.* 瀏覽模型

執行 C5.0 節點時，模型區塊即會新增至串流，也會新增至視窗右上角的「模型」選用區。若要瀏覽模型，請在圖示上按一下滑鼠右鍵，然後從快速功能表中選擇**編輯**或**瀏覽**。

「規則」瀏覽器會顯示採用決策樹狀結構格式之 C5.0 節點產生的規則集。起初，樹狀結構是收合的。若要展開它，請按一下**全部**按鈕以顯示所有層次。

圖 *97.* 規則瀏覽器

現在，您可以看到拼圖的遺漏部分。對於 *Na/K* 比例小於 14.64 且血壓高的人員，年齡決定藥品的選擇。對於低血壓的人員，膽固醇層次似乎是最佳預測工具。

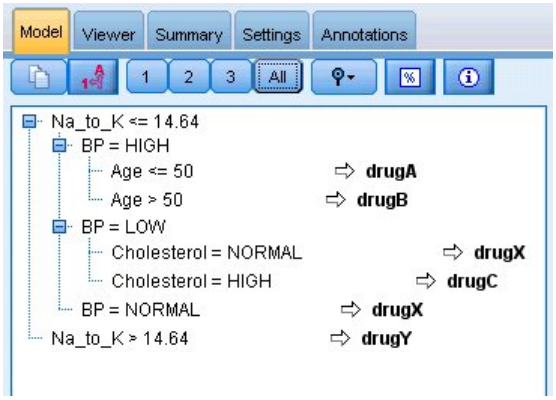


圖 *98.* 完全展開的規則瀏覽器

可透過按一下**檢視器**標籤以更準確的圖形格式來檢視相同的決策樹狀結構。在這裡，您可以更輕鬆地查看每個血壓種類的觀察值數目，以及觀察值的百分比。

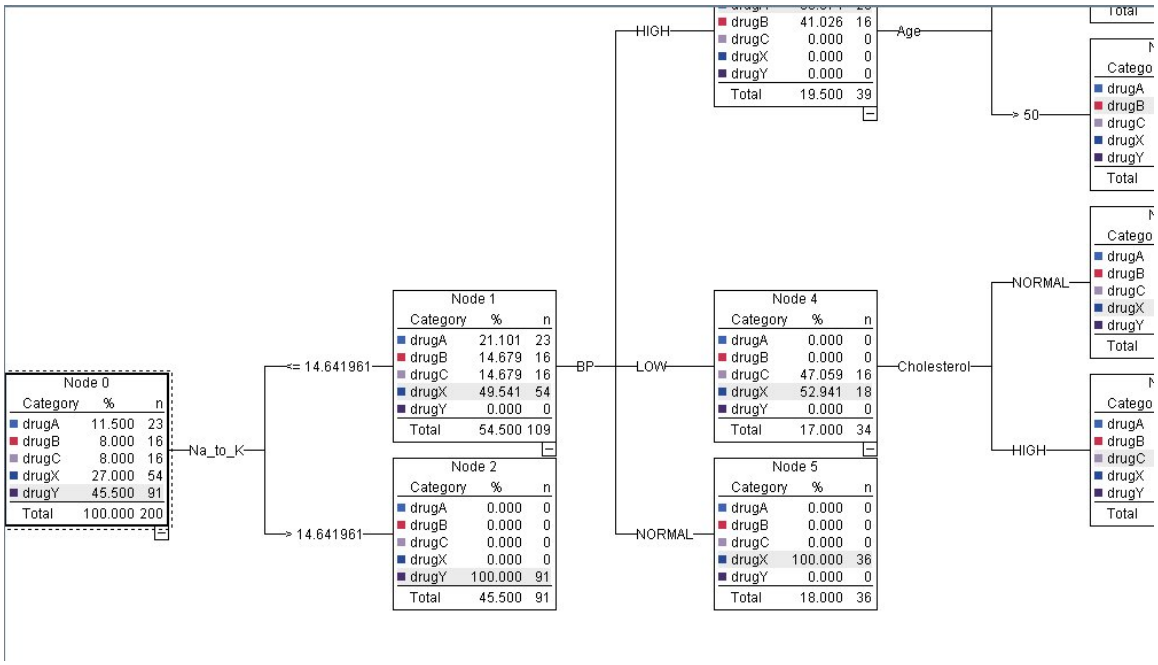


圖 *99.* 採用圖形格式的決策樹狀結構

使用分析節點

可使用分析節點來評量模型的正確性。將「分析」節點（從「輸出」節點選用區）連接至模型區塊，開啟「分析」節點並按一下**執行**。

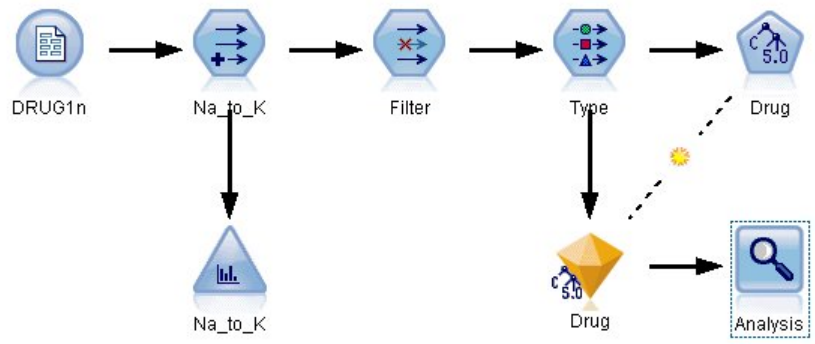


圖 *100.* 新增分析節點

「分析」節點輸出顯示使用這個人工資料集，模型會正確地預測針對資料集中的每筆記錄選擇的藥品。使用真實資料集，您不可能看到 100% 的正確性，但是您可以使用「分析」節點來協助判定模型的正確性對於您的特定應用程式而言是否可以接受。

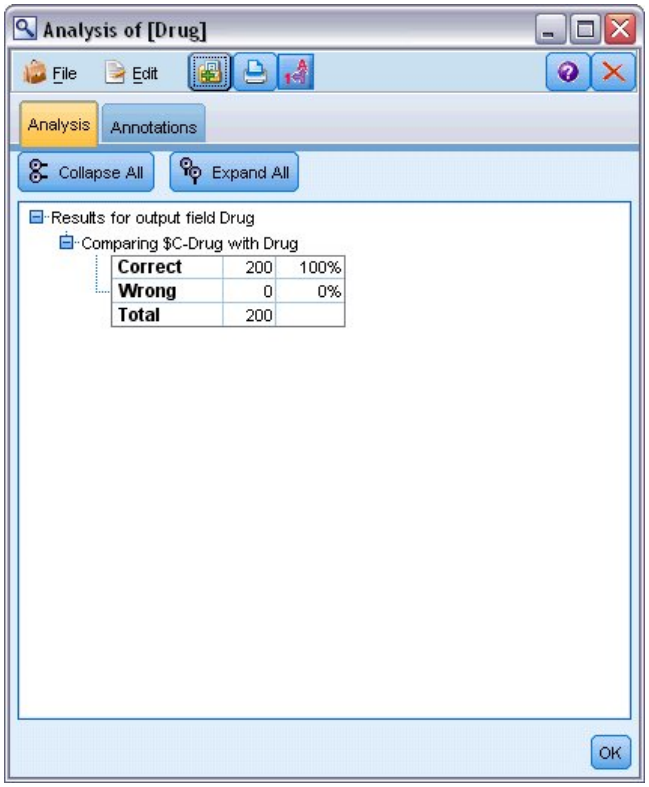


圖 *101.* 分析節點輸出

第 9 章 篩選預測值（功能選擇）

「功能選擇」節點可協助您識別對於預測某個結果極其重要的欄位。「功能選擇」節點可從一組成百甚至上千個預測值中篩選極其重要的預測值並對其進行等級排列和選取。最終，您可能會得到更快速更高效的模型 — 該模型使用的預測值更少，執行速度更快，可能更容易理解。

此範例中使用的資料代表假設的電話公司的資料倉儲，其中包含 5000 名公司客戶對某次特價促銷活動回應的相關資訊。該資料包括大量欄位，其中包括客戶年齡、職業、收入及電話使用情況統計量。三個「目標」欄位會顯示客戶是否對三項優惠的每一項做出回應。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。

此範例使用名為 *featureselection.str* 的串流，其參照的資料檔名為 *customer_dbase.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*featureselection.str* 檔案位於 *streams* 目錄中。

此範例的重點僅集中在其中一項優惠上作為目標。它使用 CHAID 樹狀結構建置節點來開發模型以說明哪些客戶最有可能對促銷做出回應。它會將兩種方法進行對比：

- 不使用功能選擇。資料集中的所有預測值欄位都用作 CHAID 樹狀結構的輸入。
- 使用功能選擇。使用「功能選擇」節點來選取前 10 個預測值。然後將這些值輸入 CHAID 樹狀結構。

透過比較兩個產生的樹狀結構模型，我們可以看到功能選擇如何產生有效的結果。

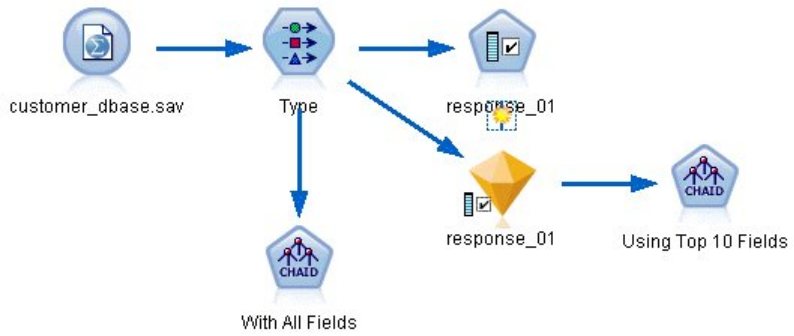


圖 *102.* 功能選擇範例串流

1. 將「統計量檔案」來源節點放到空白串流畫布上。將此節點指向位於 IBM SPSS Modeler 安裝下 *Demos* 目錄中的範例資料檔 *customer_dbase.sav*。（或者開啟 *streams* 目錄中的範例串流檔 *featureselection.str*。）
2. 新增「類型」節點。在「類型」標籤上，向下捲動至底端並將 *response_01* 的角色變更為目標。將其他回應欄位（*response_02* 和 *response_03*）以及清單頂部的客戶 ID (*custid*) 的角色變更為無。將所有其他欄位的角色設定為輸入，按一下**讀取值**按鈕，然後按一下**確定**。

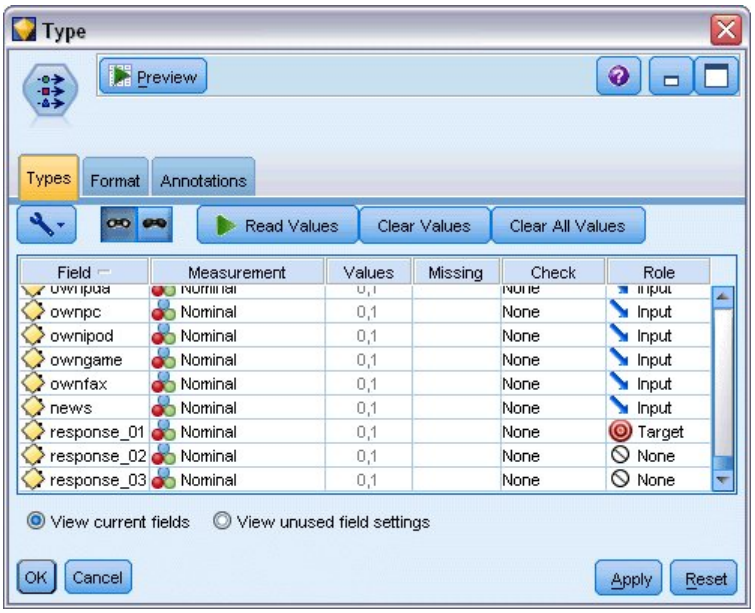


圖 *103.* 新增類型節點

3. 將「功能選擇」建模節點新增至串流。在此節點上，您可以指定規則及準則來篩選欄位或取消欄位資格。
4. 執行串流以建立「功能選擇」模型區塊。
5. 用滑鼠右鍵按一下串流上或「模型」選用區中的模型區塊，然後選擇**編輯**或**瀏覽**來查看結果。

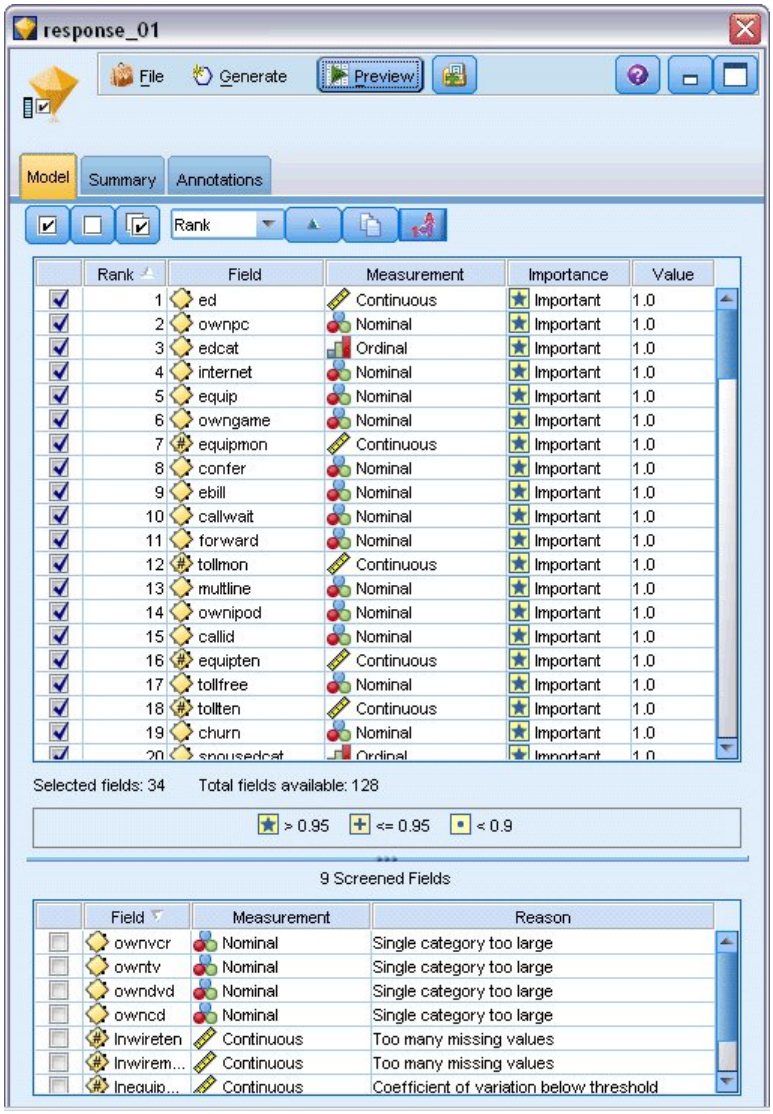


圖 *104.* 功能選擇模型區塊中的模型標籤

頂部畫面顯示對預測有用的欄位。這些欄位根據重要性排列等級。底部畫面顯示從分析中篩選了哪些欄位及其原因。透過檢查頂部畫面中的欄位，您可以決定在後續建模階段作業中使用哪些欄位。

6. 現在我們可以選取下游要使用的欄位。盡管原本有 34 個欄位識別為重要，但我們想要更進一步縮減預測值集。
7. 在第一個直欄中使用勾號僅選取前 10 個預測值來取消選取不需要的預測值。（按一下列 11 中的勾號，按住 Shift 鍵並按一下列 34 中的勾號。）關閉模型區塊。
8. 若要比較不使用功能選擇的結果，則必須將兩個 CHAID 建模節點新增至串流：一個使用功能選擇，一個不使用。
9. 將一個 CHAID 節點連接至「類型」節點，將另一個連接至「功能選擇」模型區塊。
10. 開啟每一個 CHAID 節點，選取「建置選項」標籤並確保「目標」窗格中已選取選項**建置新模型**、**建置單一樹狀結構**及**啟動互動式階段作業**。

在「基本」窗格上，請確定**最大樹狀結構深度**設為 5。

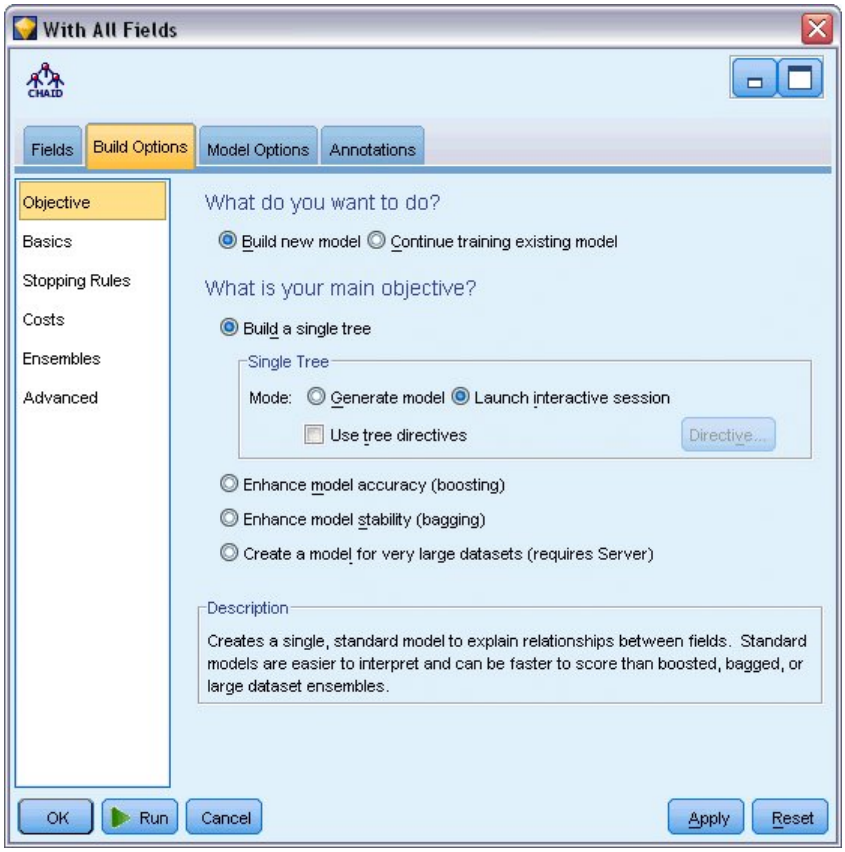


圖 *105.* 所有預測值欄位之 *CHAID* 建模節點的目標設定

建置模型

1. 執行使用資料集中所有預測值的 CHAID 節點（連接至「類型」節點的節點）。在它執行時，注意它執行所花費的時長。結果視窗會顯示一個表格。
2. 從功能表中，選擇**樹狀結構 > 增長樹狀結構**以增長並顯示展開的樹狀結構。

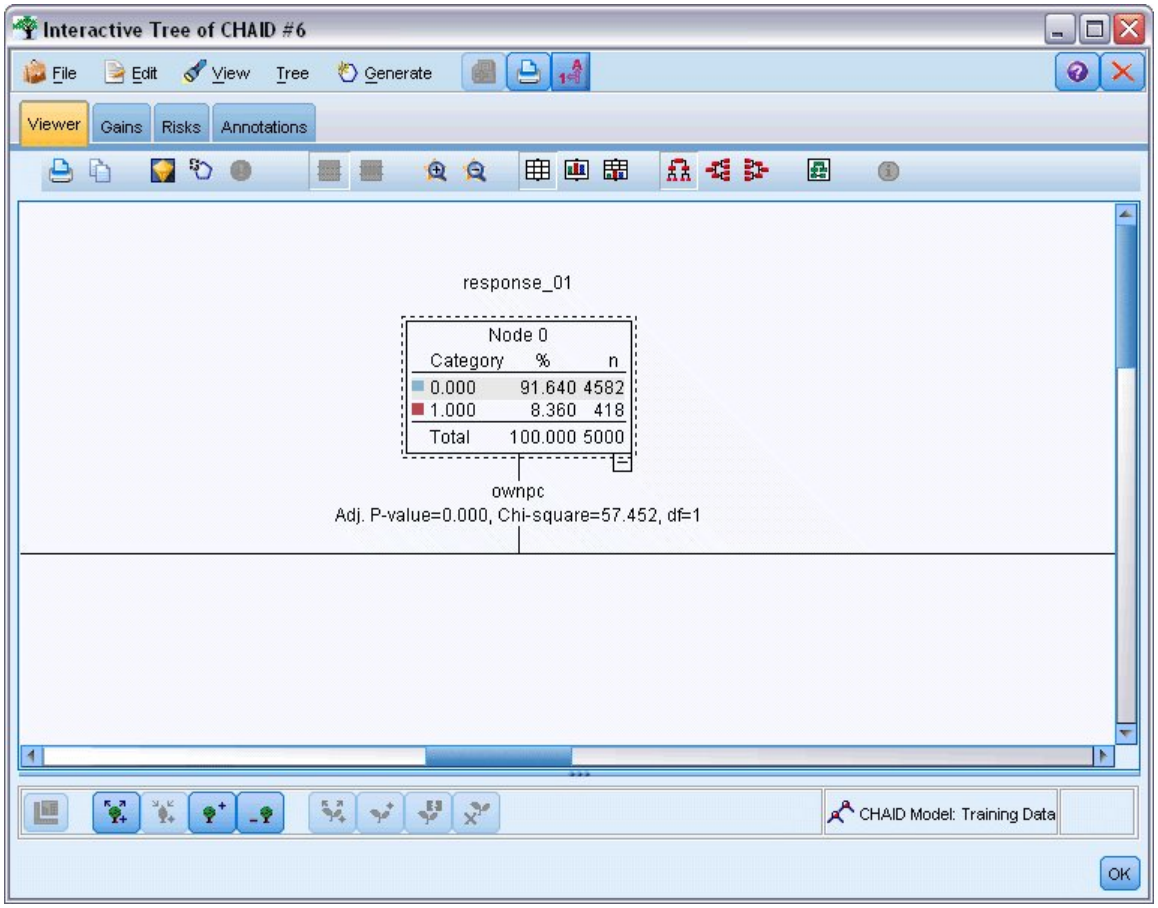


圖 *106.* 在樹狀結構建置器中增長樹狀結構

3. 現在針對僅使用 10 個預測值的另一個 CHAID 節點執行相同的動作。當樹狀結構建置器開啟時再次增長樹狀結構。

第二個模型執行速度應比第一個更快。因為此資料集相當小，執行時間的差異可能為幾秒鐘；但對於更大的真實資料集來說，差異可能會非常明顯 — 幾分鐘甚至幾個小時。使用功能選擇可以大大加快處理時間。

第二個樹狀結構包含的樹狀結構節點也比第一個少。它更容易理解。但在決定使用它之前，需要搞清楚它是否有效以及與使用所有預測值的模型相比結果如何。

比較結果

若要比較兩個結果，我們需要測量有效性。為此，我們會使用樹狀結構建置器中的「增益」標籤。我們會查看**提升**，它測量的是節點中的記錄處於目標種類的可能性比資料集中的所有記錄增加了多少。例如，提升值為 148% 表示節點中的記錄處於目標種類中的可能性是資料集中所有記錄的 1.48 倍。提升在「增益」標籤上的指數直欄中指出。

1. 在整個預測值集的樹狀結構建置器中，按一下「增益」標籤。將目標種類變更為 1.0。透過先按一下「四分位數」工具列按鈕將顯示變更為四分位數。然後從此按鈕右側的下拉清單中選取**四分位數**。
2. 在樹狀結構建置器中針對包含 10 個預測值的集合重複此程序，這樣您就有兩個類似的「增益」表格進行比較，如下圖中所示。

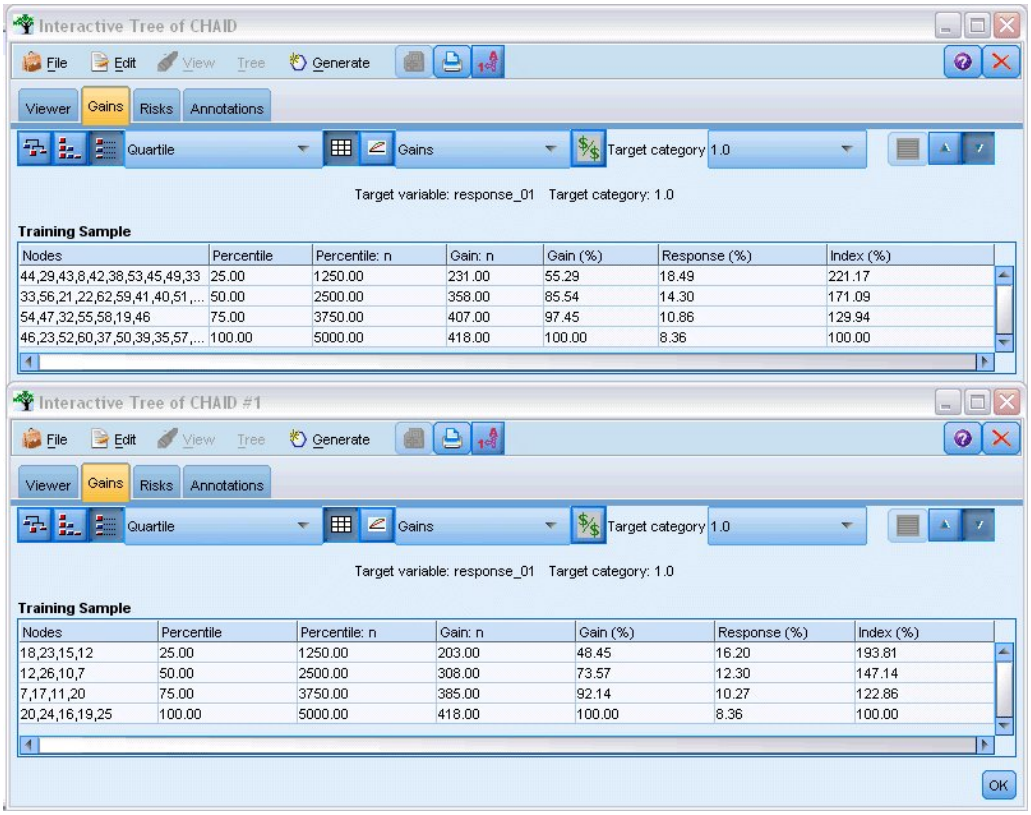


圖 *107.* 兩個 *CHAID* 模型的增益圖

每一個增益表將其樹狀結構的終端節點分組成四分位數。若要比較兩個模型的有效性，請查看每一個表格中頂部四分位數的提升（指數值）。

包括所有預測值後，模型顯示提升了 221%。即這些節點中具有特性的觀察值回應目標促銷的可能性是 2.2 倍。若要瞭解那些特性是什麼，請按一下以選取頂端列。然後切換至「檢視器」標籤，其中對應節點的外框現在為黑色。沿著樹狀結構追蹤到每一個強調顯示的終端節點來查看如何分割預測值。頂部四分位數單獨包括 10 個節點。轉換為真實評分模型後，10 個不同的客戶設定檔可能很難管理。

僅包括前 10 個預測值時（由功能選擇識別）時，提升接近 194%。盡管此模型不像使用所有預測值的模型那麼好，但也確實有用。在這裡，頂部四分位數僅包括四個節點，更為簡單。因此，我們可以判定功能選擇模型優於包含所有預測值的模型。

摘要

我們來回顧一下功能選擇的優點。使用更少的預測值，成本更低。這表示您要收集、處理及為模型提供的資料更少。計算時間得到改進。在此範例中，即使額外增加功能選擇步驟，較小預測值集的模型建置也會明顯加快。對於更大的真實資料集，節省的時間應該會大大增加。

使用更少的預測值，使得評分更為簡單。正如範例所示，您只能識別四個可能回應促銷的客戶的設定檔。請注意，預測值數目越大，您承擔模型過適的風險越大。模型越簡單，對其他資料集的廣義化程度越好（但您需要進行測試才能確定）。

您本可以使用樹狀結構建置演算法來進行功能選擇，讓樹狀結構能夠識別對您來說最重要的預測值。實際上，CHAID 演算法通常會用於此目的，甚至可以逐個層次增長樹狀結構來控制其深度及複雜性。但「功能選擇」節

點使用起來更加輕鬆快捷。它會在一個快速步驟中將所有預測值排列等級，讓您能夠快速識別最重要的欄位。它還容許您改變要包括的預測值數。您可以使用前 15 或 20 個預測值而非前 10 個再次輕鬆地執行此範例，從而比較結果來判定最佳模型。

第 10 章 減少輸入資料字串長度（重新分類節點）

減少輸入資料字串長度（重新分類）

對於二項式邏輯迴歸以及包括二項式邏輯迴歸模型的自動分類器模型，字串欄位限制為最多 8 個字元。如果字串多於 8 個字元，則可以使用「重新分類」節點來記錄。

此範例使用參照資料檔 *drug_long_name* 的串流 *reclassify_strings.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*reclassify_strings.str* 檔位於 *streams* 目錄。

此範例著重於一小部分串流以顯示可能使用超長字串產生的錯誤種類，並說明如何使用「重新分類」節點將字串詳細資料變更為可接受的長度。雖然該範例使用二項式「邏輯迴歸」節點，但在使用「自動分類器」節點來產生二項式「邏輯迴歸」模型時也同樣適用。

重新分類資料

1. 使用「變數檔案」來源節點，連接至 *Demos* 資料夾中的 *drug_long_name*。

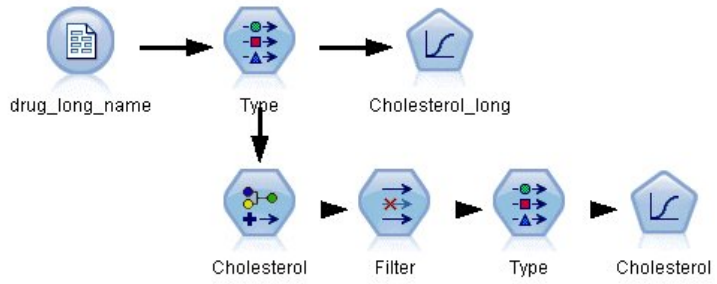


圖 *108.* 顯示二項式邏輯迴歸字串重新分類的串流範例

2. 將「類型」節點新增至「來源」節點並選取 **Cholesterol_long** 作為目標。
3. 將「邏輯迴歸」節點新增至「類型」節點。
4. 在「邏輯迴歸」節點中，按一下「模型」標籤並選取**二項式**程序。

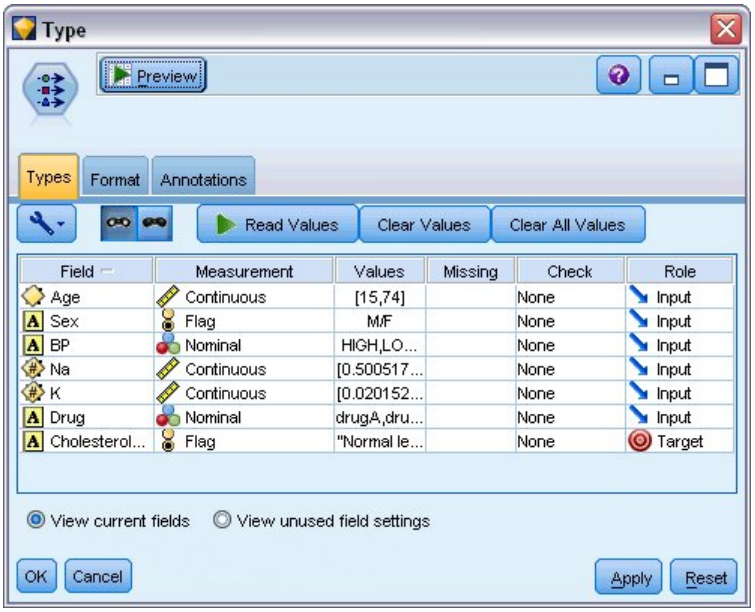


圖 *109. "Cholesterol_long"* 欄位中的長字串詳細資料

5. 當您執行 *reclassify_strings.str* 中的「邏輯迴歸」節點時，會顯示一則錯誤訊息，警告 **Cholesterol_long** 字串值太長。

   如果您遇到這種類型的錯誤訊息，請遵循此範例中剩餘內容所說明的程序來修改資料。

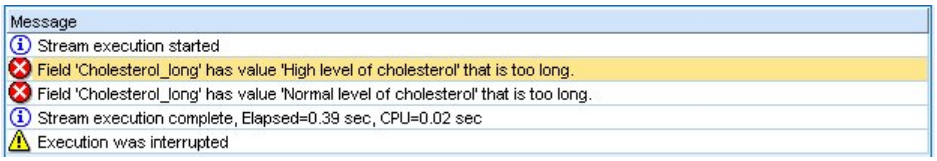


圖 *110.* 執行二項式邏輯迴歸節點時顯示的錯誤訊息

6. 將「重新分類」節點新增至「類型」節點。
7. 在「重新分類」欄位中，選取 **Cholesterol_long**。
8. 輸入**膽固醇**作為新欄位名稱。
9. 按一下**取得**按鈕以將 **Cholesterol_long** 值新增至原始值直欄。
10. 在新值直欄中，在**高水平膽固醇**的原始值旁邊輸入**高**，並在**正常水平膽固醇**的原始值旁邊輸入**正常**。

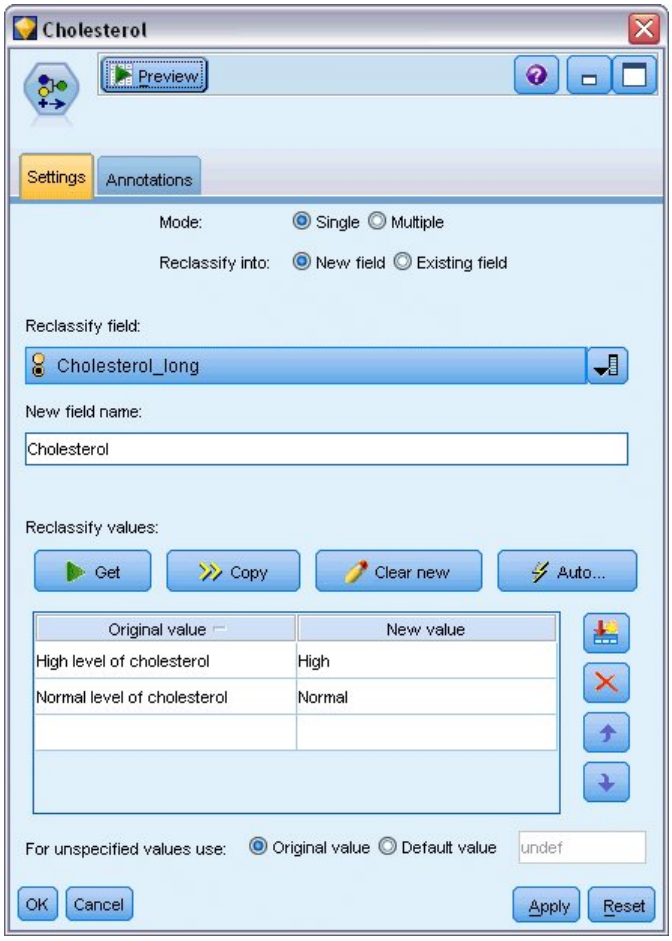


圖 *111.* 重新分類長字串

11. 將「過濾器」節點新增至「重新分類」節點。
12. 在「過濾器」直欄中，按一下以移除 **Cholesterol_long**。

圖 *112.* 從資料中過濾 *"Cholesterol_long"* 欄位

13. 將「類型」節點新增至「過濾器」節點並選取**膽固醇**作為目標。

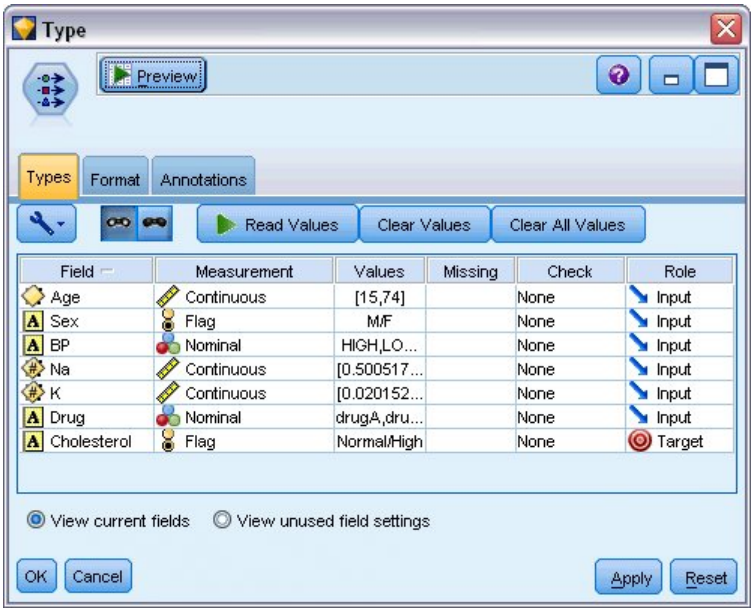


圖 *113.* 「膽固醇」欄位中的簡短字串詳細資料

14. 將「邏輯」節點新增至「類型」節點。
15. 在「邏輯」節點中，按一下「模型」標籤並選取**二項式**程序。
16. 您現在可以執行「二項式邏輯」節點並產生模型而不顯示錯誤訊息。

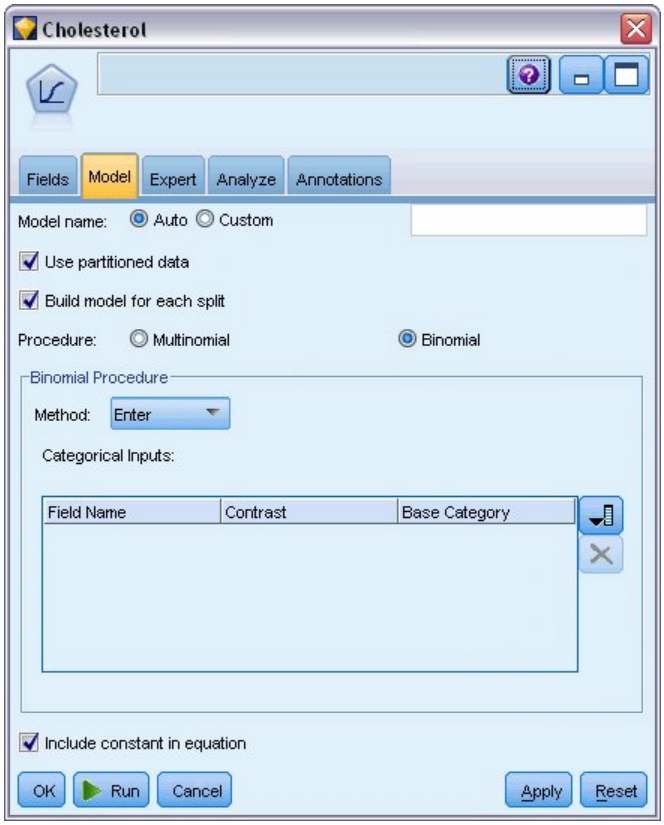


圖 *114.* 選擇二項式作為程序

此範例僅顯示串流的一部分。如果您需要串流類型（您可能需要在其中重新分類長字串）的進一步相關資訊，提供了下列範例：

- 自動分類器節點。請參閱第 33 頁的『給客戶回應（自動分類器）建模』主題，以取得更多資訊。
- 二項式邏輯迴歸節點。請參閱第 133 頁的第 13 章, 『電信客戶流失（二項式邏輯迴歸）』主題，以取得更多資訊。

有關如何使用 IBM SPSS Modeler 的更多資訊（例如使用手冊、節點參照以及演算法手冊）可從安裝磁碟的 *Documentation* 目錄取得。

第 11 章 給客戶回應（決策清單）建模

「決策清單」演算法產生規則以指出給定二元（yes 或 no）結果的更高或更低可能性。「決策清單」模型廣泛用於客戶關係管理，例如呼叫中心或行銷應用程式。

本範例基於虛構公司，該公司希望通過向每個客戶提供合適的優惠以在未來的市場行銷活動中獲取更多利潤。尤其是，該範例使用「決策清單」模型來根據以前的促銷活動識別最有可能做出積極回應的客戶的特性，並根據結果產生郵寄清單。

「決策清單」模型尤其適合互動式建模，讓您調整模型中的參數並立即查看結果。對於可讓您自動建立數個不同模型並將結果排名的不同方法，可以改為使用「自動分類器」節點。

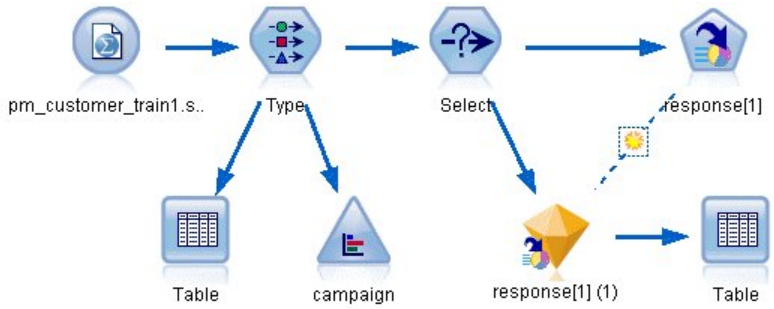
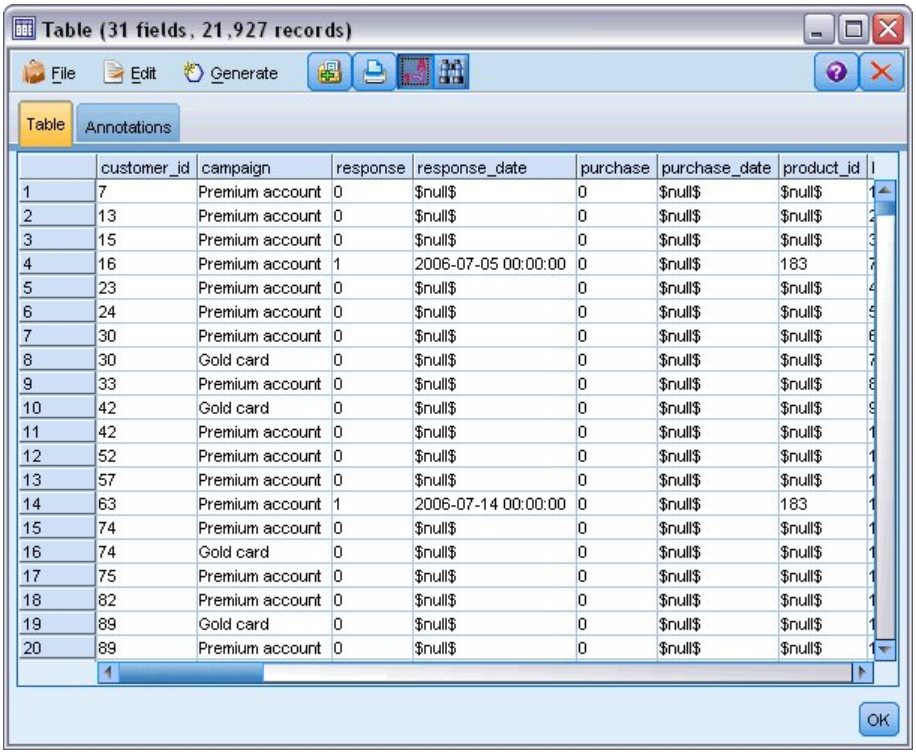


圖 *115.* 決策清單串流範例

此範例使用串流 *pm_decisionlist.str*，它參照資料檔案 *pm_customer_train1.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*pm_decisionlist.str* 檔案位於 *streams* 目錄。

歷程資料

檔案 *pm_customer_train1.sav* 具有歷程資料，可用於追蹤過去行銷活動中針對特定客戶提供的優惠，如 *campaign* 欄位的值所指示。最大數目的記錄位於進階帳戶 行銷活動範圍之下。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold card	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold card	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold card	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold card	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

圖 *116.* 先前促銷的相關資料

campaign 欄位的值在資料中實際編碼為整數，並在「類型」節點中定義了標籤（例如 *2* = 進階帳戶）。您可以使用工具列切換值標籤在表格中的顯示。

該檔案還包括數個欄位，其中包含每個客戶個人背景資訊與財務相關資訊，可利用這些資訊來建置或「訓練」模型以根據特定性質來預測不同群組的回應率。

建置串流

1. 新增指向 *pm_customer_train1.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「統計量檔案」節點。（您可以在檔案路徑中指定 `$CLEO_DEMOS/` 作為參照此資料夾的捷徑。）

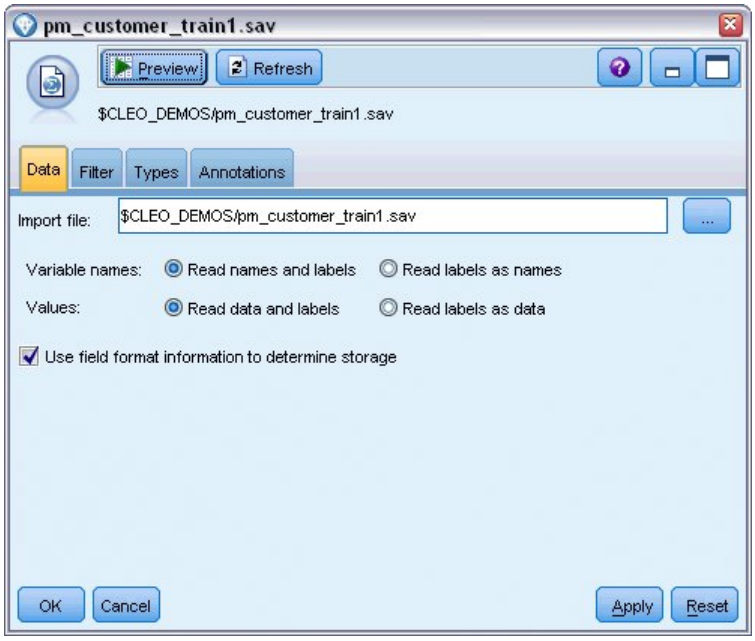


圖 *117.* 在資料中讀取

2. 新增「類型」節點，並選取 *response* 作為目標欄位（角色 = **目標**）。將此欄位的「測量」層次設為**旗標**。

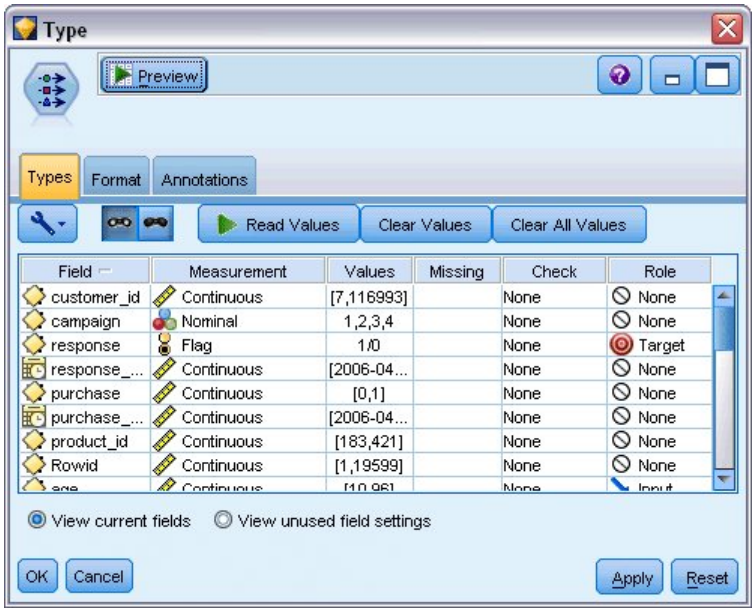


圖 *118.* 設定測量層次和角色

3. 針對下列欄位，將角色設為**無**：*customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid* 和 *X_random*。這些欄位全部用於資料，但將不會用於建置實際模型。
4. 按一下「類型」節點中的**讀取值**按鈕以確保已實例化值。

雖然資料包括四個不同行銷活動的相關資訊，但您將著重於一次分析一個行銷活動。由於最大數量的記錄位於進階行銷活動範圍之下（在資料中編碼為 *campaign=2*），因此您可以使用「選取」節點以僅在串流中包括這些

記錄。

圖 *119.* 選取單一行銷活動的記錄

建立模型

1. 將「決策清單」節點新增至串流。在「模型」標籤上，將**目標值**設為 1 以指出您要搜尋的結果。在本案例中，您要尋找向前一個優惠回應了是 的客戶。

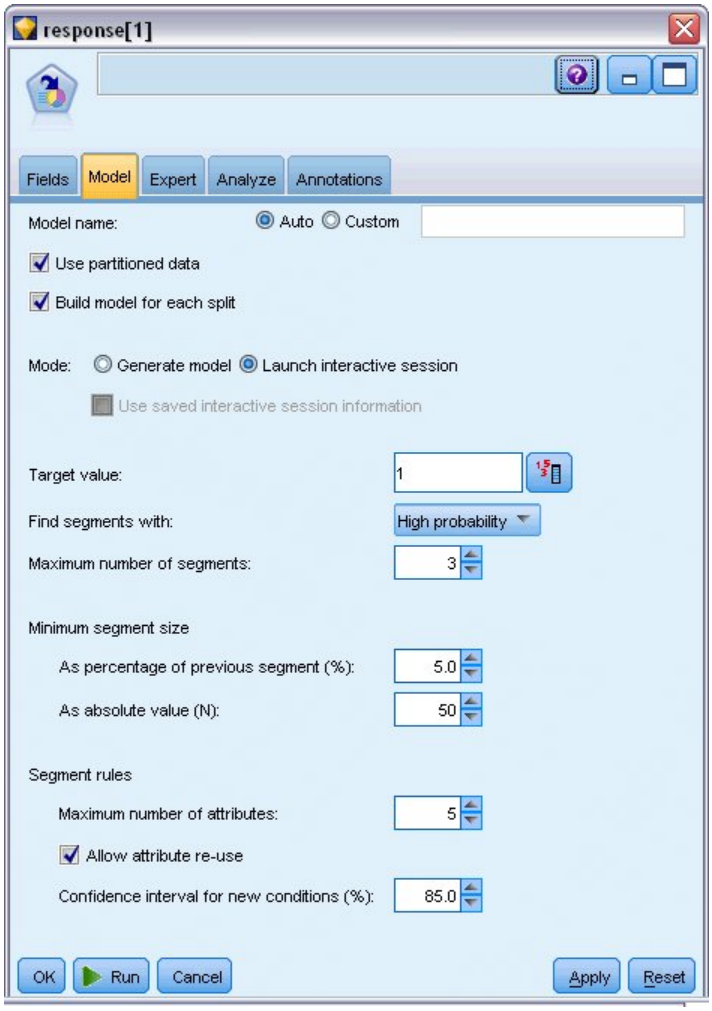


圖 *120.* 決策清單節點的模型標籤

2. 選取**啟動互動式階段作業**。
3. 若要將模型保持為簡式模型以用於本範例，請將最大區段數目設為 3。
4. 將新條件的信賴區間設為 85%。
5. 在「專家」標籤上將**模式**設為**專家**。

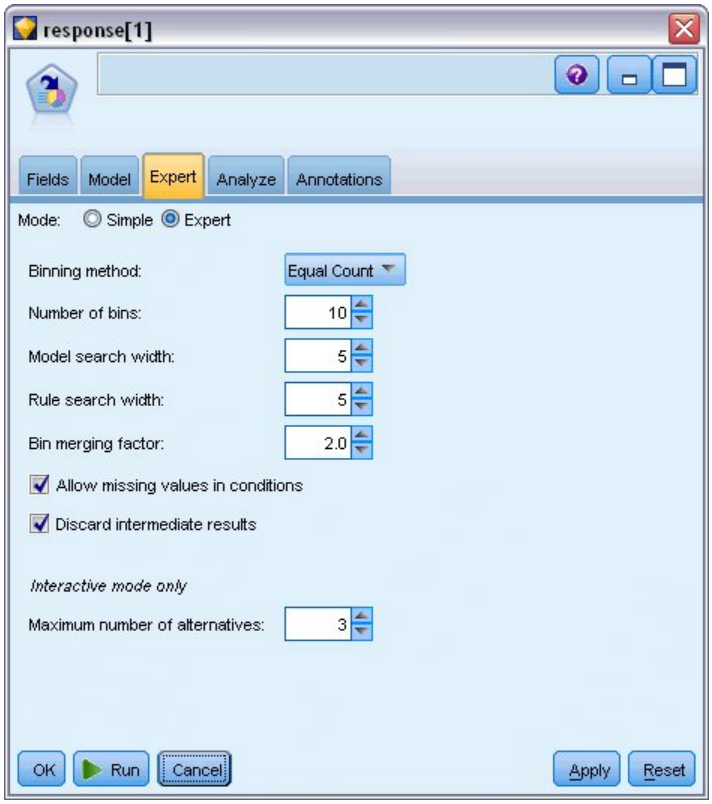


圖 *121.* 決策清單節點的專家標籤

6. 將**替代模型數目上限**增加到 3。此選項與您在「模型」標籤上選取的**啟動互動式階段作業**設定一起使用。
7. 按一下**執行**以顯示「互動式清單」檢視器。

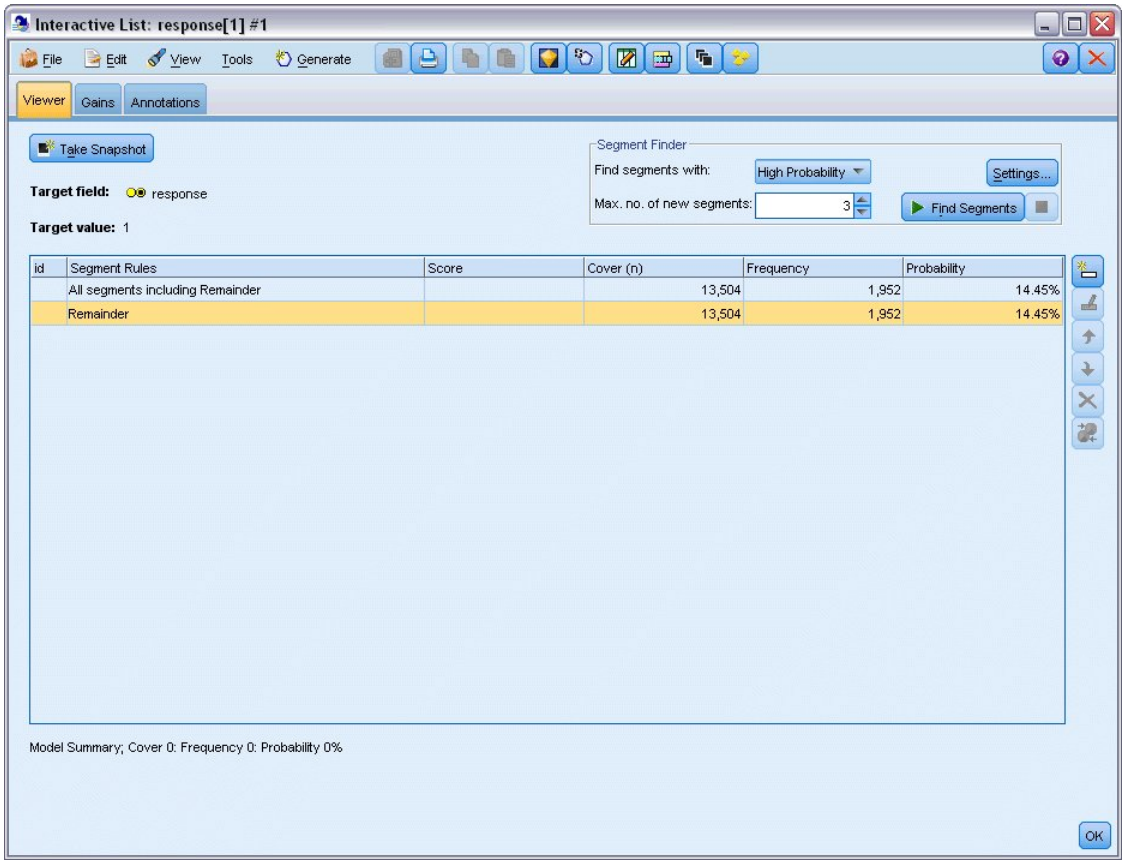


圖 *122.* 互動式清單檢視器

由於尚未定義任何區段，因此所有記錄都歸入剩餘項目。在範例中的 13,504 筆記錄中，有 1,952 表示是，整體命中率是 14.45%。您想要識別更易或更不易提供良好回應之客戶的區段，來提升此比率。

8. 在「互動式清單」檢視器中，從功能表選擇：

工具 > **尋找區段**

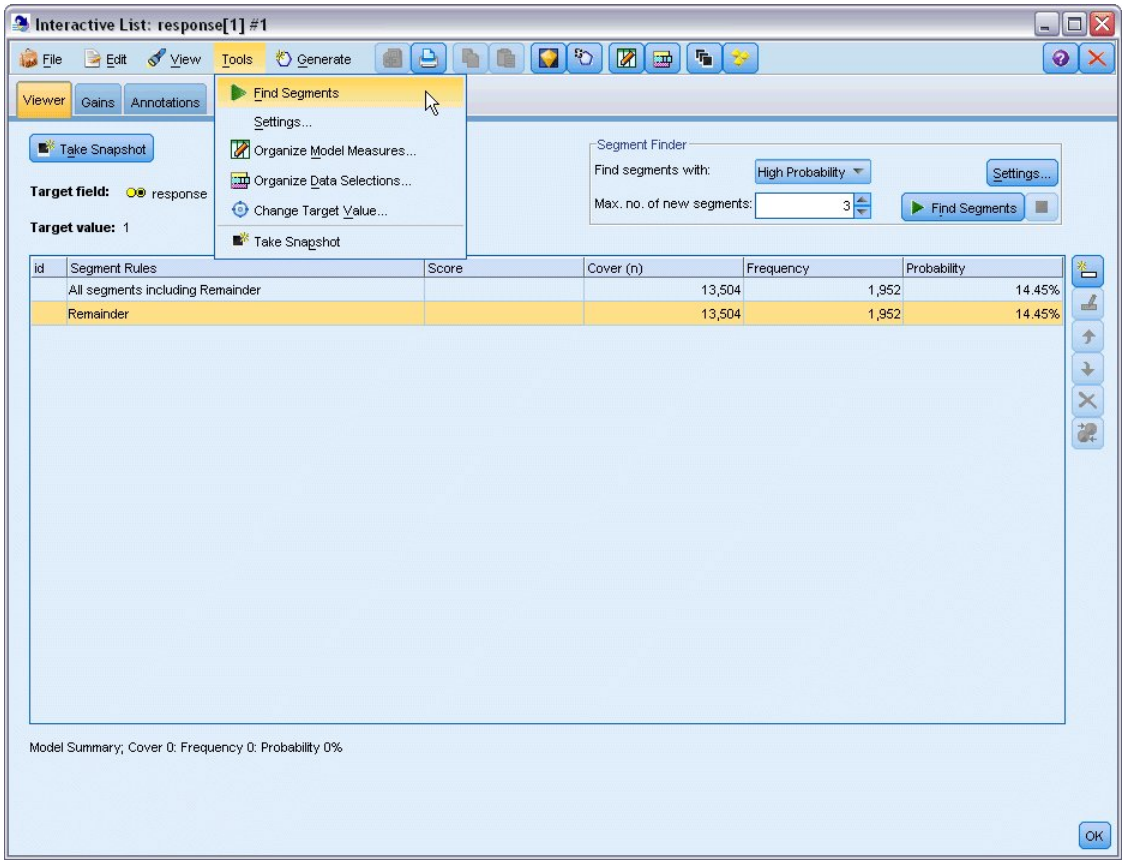


圖 *123.* 互動式清單檢視器

這會根據您在「決策清單」節點中指定的設定來執行預設採礦作業。已完成作業傳回三個互動式模型，這些模型列在「模型相簿」對話框的「替代模型」標籤中。

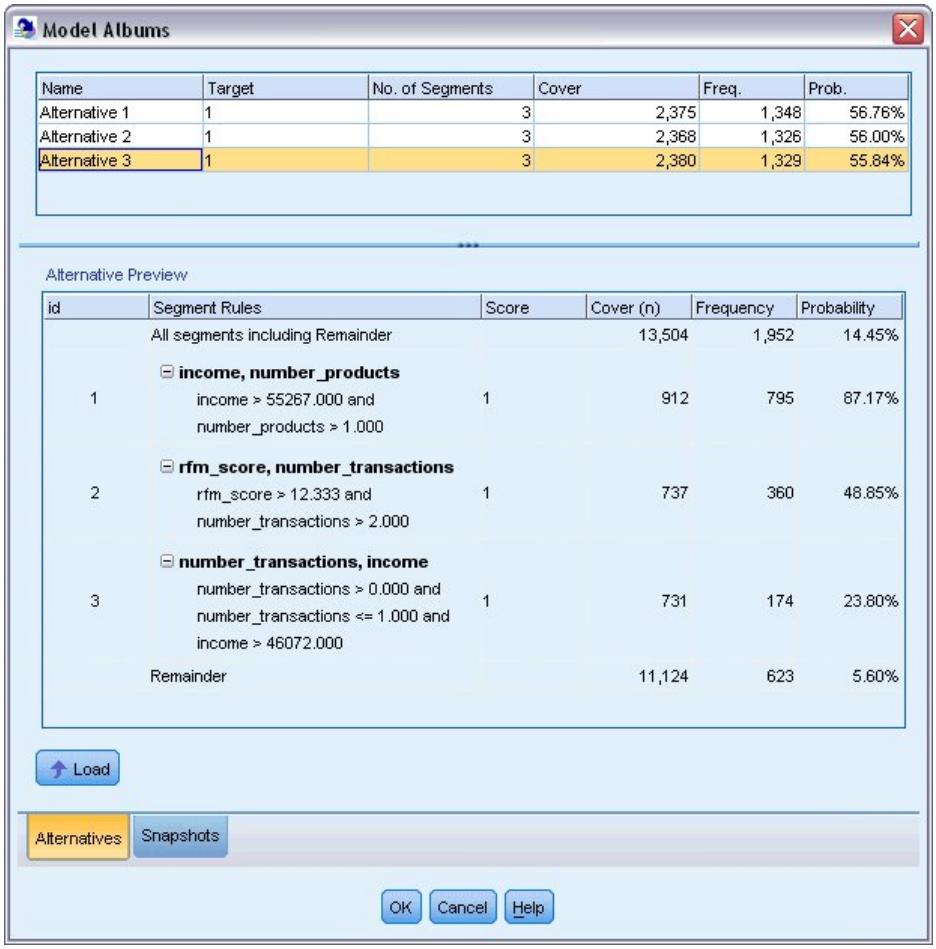


圖 *124.* 可用的替代模型

9. 從清單中選取第一個替代模型；其詳細資料顯示在「替代模型預覽」畫面中。

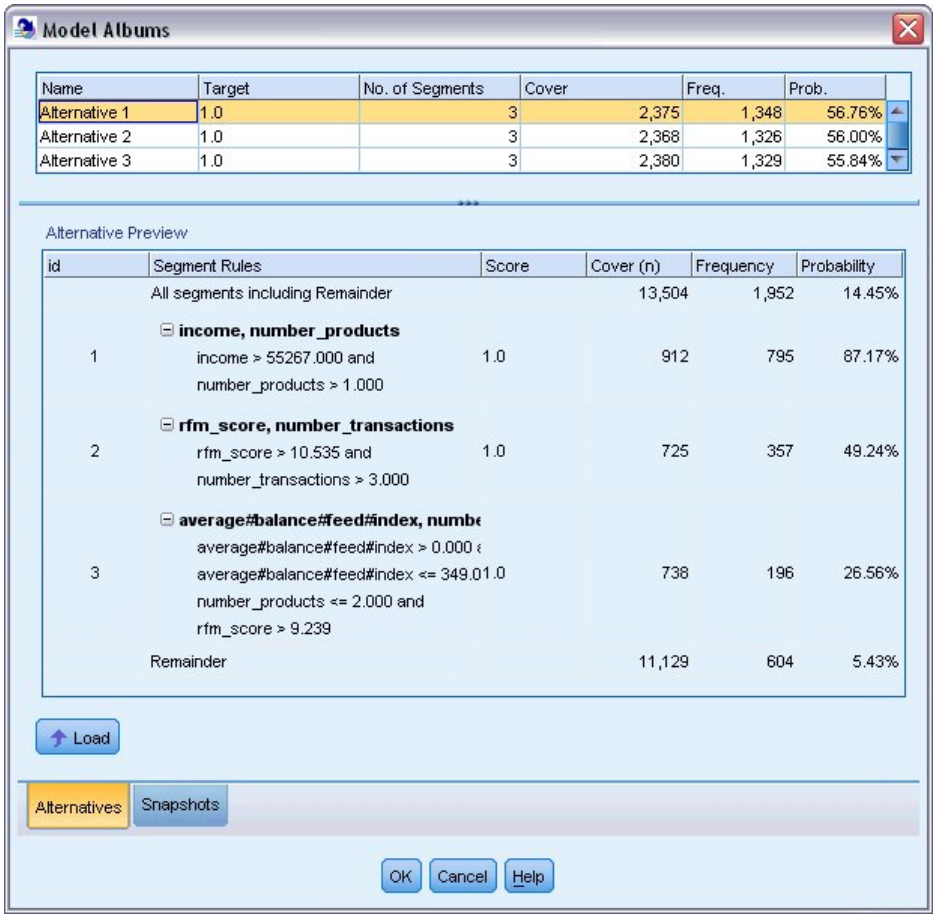


圖 *125.* 選取的替代模型

「替代模型預覽」畫面可讓您快速瀏覽任意數目的替代模型，而無需變更工作中模型，從而輕鬆體驗不同的方法。

附註：若要更好地查看模型，您可能希望在對話框中最大化「替代模型預覽」畫面，如這裡所示。可以拖曳畫面邊框來執行此動作。

根據預測工具（例如收入、每個月的交易數和 RFM 評分）來使用規則，模型會識別回應率高於範例整體的回應率的區段。結合區段時，此模型會建議您可將命中率提升到 56.76%。但是，該模型只涵蓋了整體範例中的一小部分，留下 11,000 筆記錄（其中有數百個命中）以歸入剩餘項目。您想要一個模型，它將擷取更多命中，同時仍然排除低效能區段。

10. 若要嘗試不同的建模方法，請從功能表中選擇：

工具 > **設定**

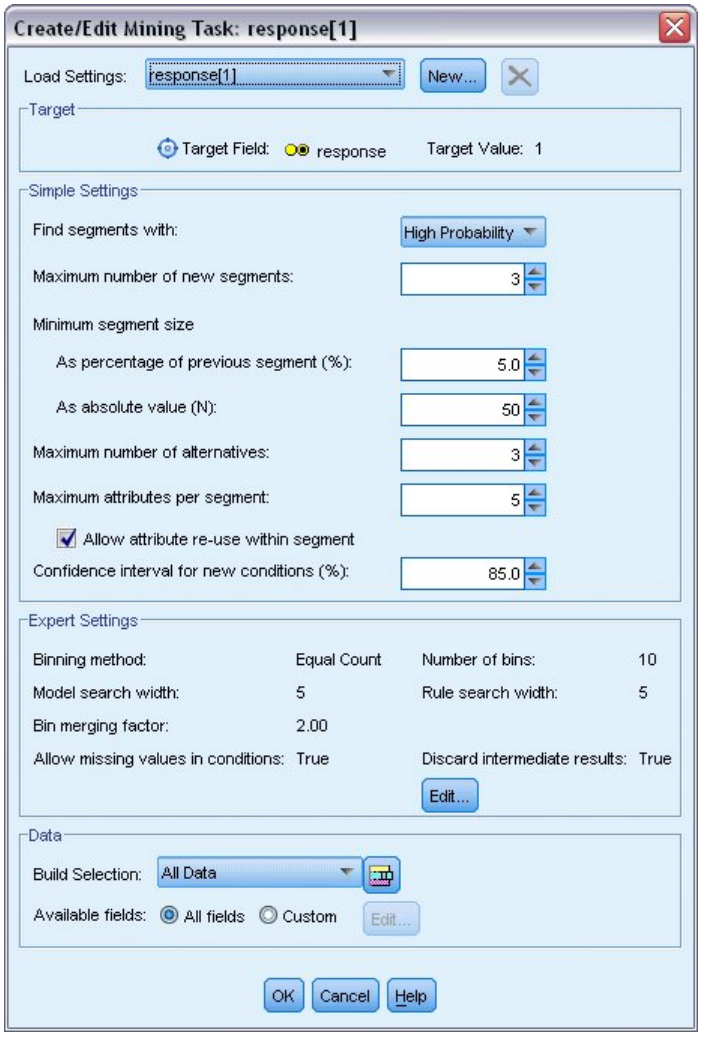


圖 *126.* 建立/編輯採礦作業對話框

11. 按一下**新建**按鈕（右上角）以另外建立一個採礦作業，並在「新建設定」對話框中指定向下搜尋 作為作業名稱。

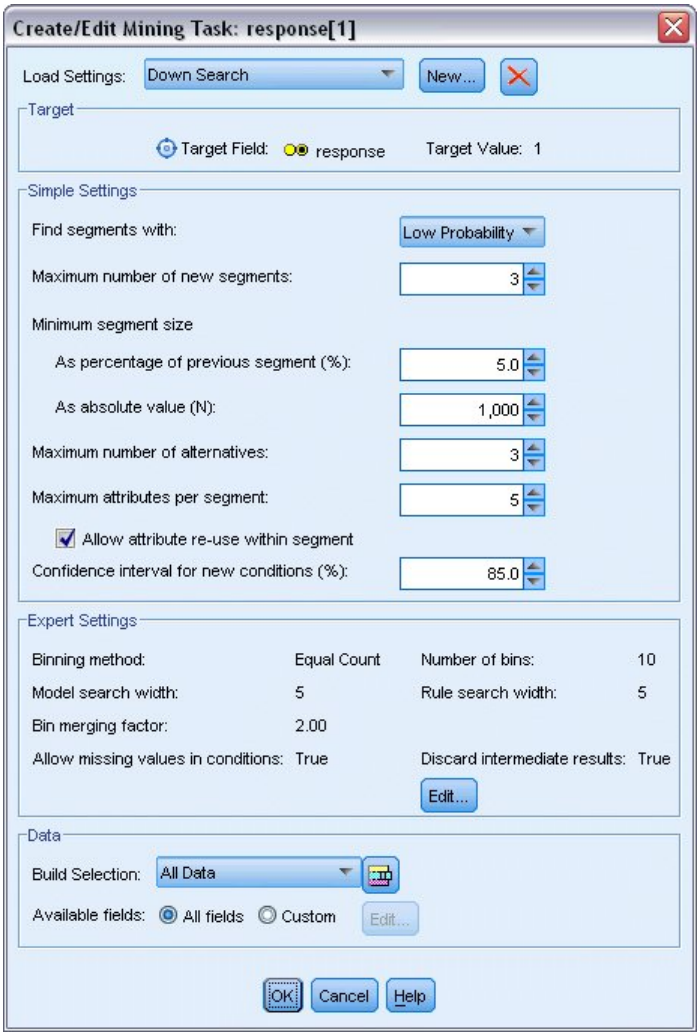


圖 *127.* 建立/編輯採礦作業對話框

12. 將作業的搜尋方向變更為**低可能性**。這樣做會導致演算法搜尋最低（而非最高）回應率的區段。
13. 將最小區段大小增加到 1,000。按一下**確定**以回到「互動式清單」檢視器。
14. 在「互動式清單」檢視器中，確保區段搜尋器 畫面顯示機關報的作業詳細資料，然後按一下**尋找區段**。

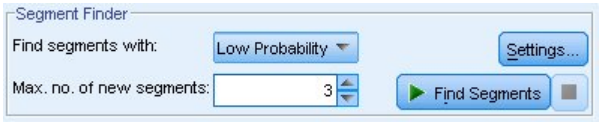


圖 *128.* 在新的採礦作業中尋找區段

該作業會傳回一組新的替代模型，這些模型顯示在「模型相簿」對話框的「替代模型」標籤中，並且可採用與預覽結果相同的方式來預覽這些模型。

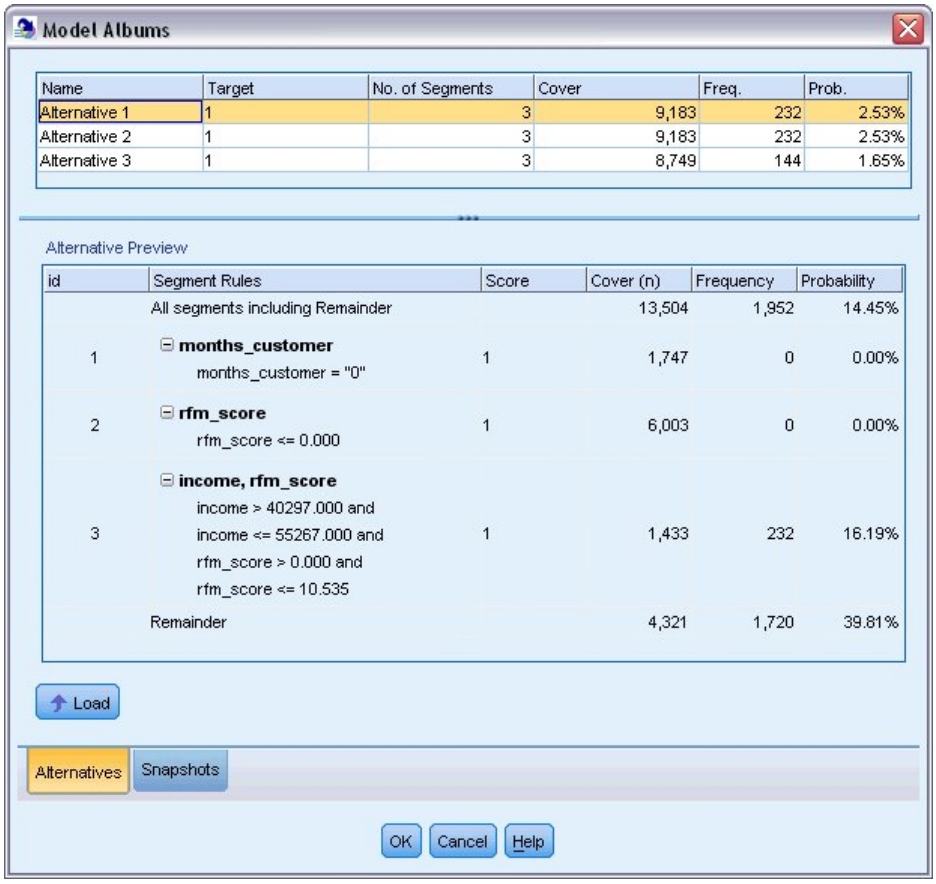


圖 *129.* 向下搜尋模型結果

此時，每一個模型都會識別具有低回應可能性而非高回應可能性的區段。查看第一個替代模型，僅排除這些區段會將剩餘項目的命中率增加到 39.81%。此模型低於您之前查看的模型，但涵蓋範圍更高（表示命中總數更多）。

透過結合兩種方法—使用「低可能性」搜尋來清除不感興趣的記錄，然後再使用「高可能性」搜尋—您可以改善此結果。

15. 按一下**載入**以將此模型（即第一個「向下搜尋」替代模型）設為工作中模型，然後按一下**確定**以關閉「模型相簿」對話框。

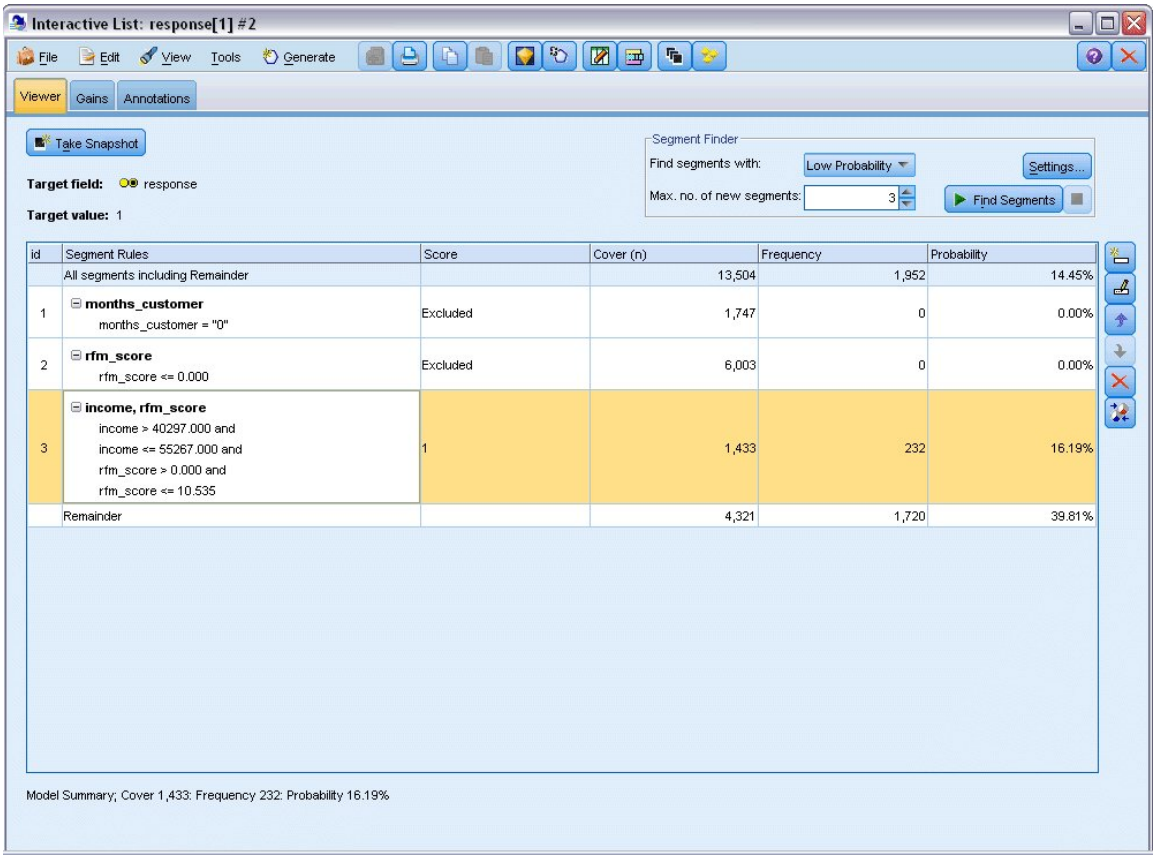


圖 *130.* 排除區段

16. 在前兩個區段中的每一個區段上按一下滑鼠右鍵，然後選取**排除區段**。這些區段一起將擷取幾乎 8,000 筆記錄，這些記錄之間的命中為零，因此將其從未來優惠中排除是有意義的。（已排除區段將評分為空值以指出這一點）。
17. 用滑鼠右鍵按一下第三個區段，然後選取**刪除區段**。在 16.19% 時，此區段的命中率與基準線比率 14.45% 沒有什麼區別，因此它不會新增足夠的資訊來合理地保持它。

    附註：刪除區段與排除區段不同。排除區段只會變更它的評分方式，而刪除它會完整將其從模型中移除。

    排除最低效能區段之後，您現在可以在剩餘項目中搜尋高效能區段。
18. 按一下表格中的剩餘項目列以選取它，因此下一個採礦作業將僅套用於剩餘項目。

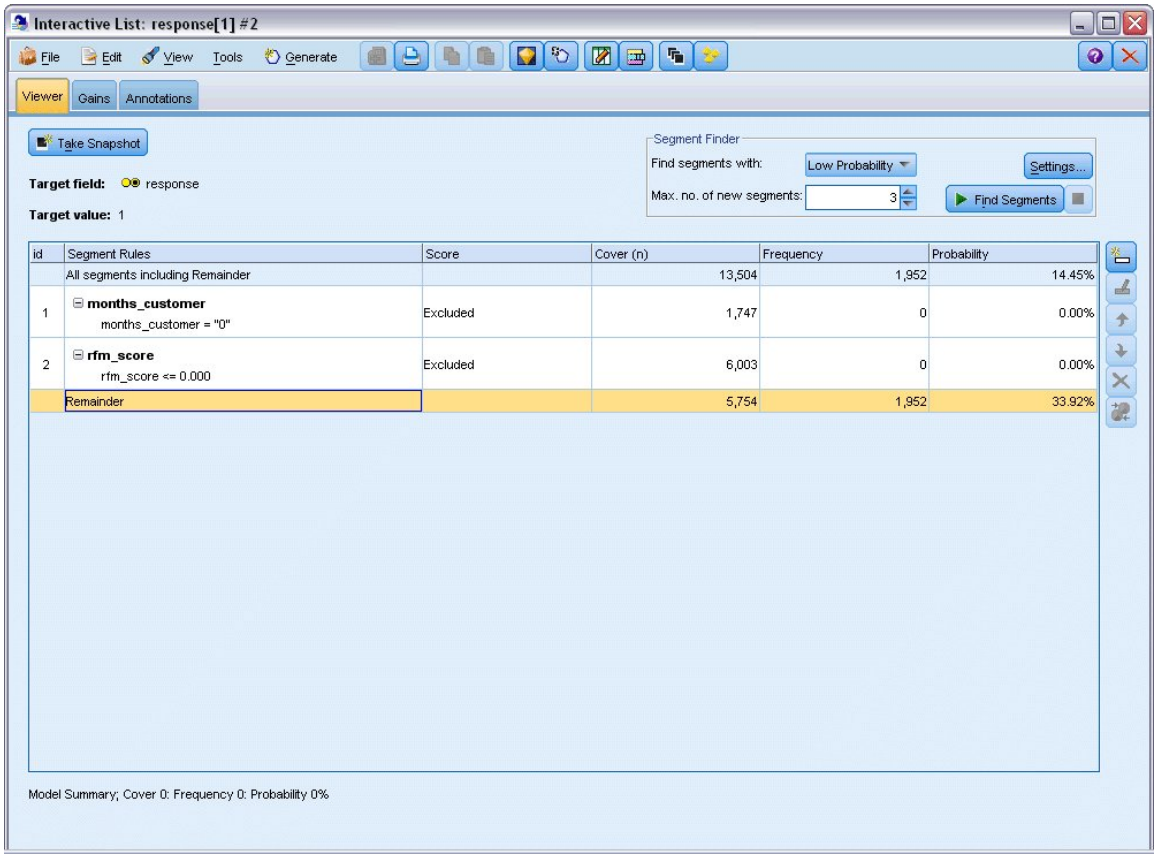


圖 *131.* 選取區段

19. 選取剩餘項目之後，按一下**設定**以重新開啟「建立/編輯採礦作業」對話框。
20. 在頂端的**載入設定**中，選取預設採礦作業：**response[1]**。
21. 編輯**簡式設定**以將新區段數目增加到 5，將最低區段大小增加到 500。
22. 按一下**確定**以回到「互動式清單」檢視器。

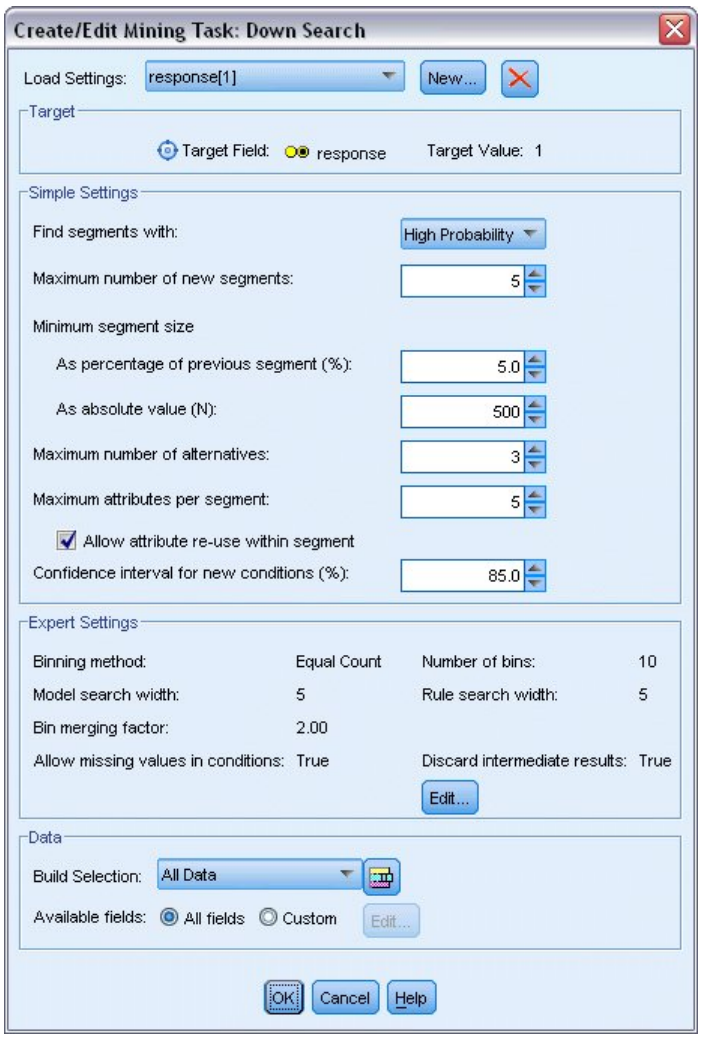


圖 *132.* 選取預設採礦作業

23. 按一下**尋找區段**。

這會顯示另一組替代模型。透過將一個採礦作業的結果送入另一個採礦作業，這些最新模型混合了高效能區段和低效能區段。將會排除低回應率的區段，這表示將它們評分為空值，而併入的區段將評分為 1。整體統計量會反應這些排除項目，第一個替代模型顯示命中率為 45.63%，其涵蓋範圍（3,456 筆記錄中有 1,577 次命中）高於任何先前的模型。

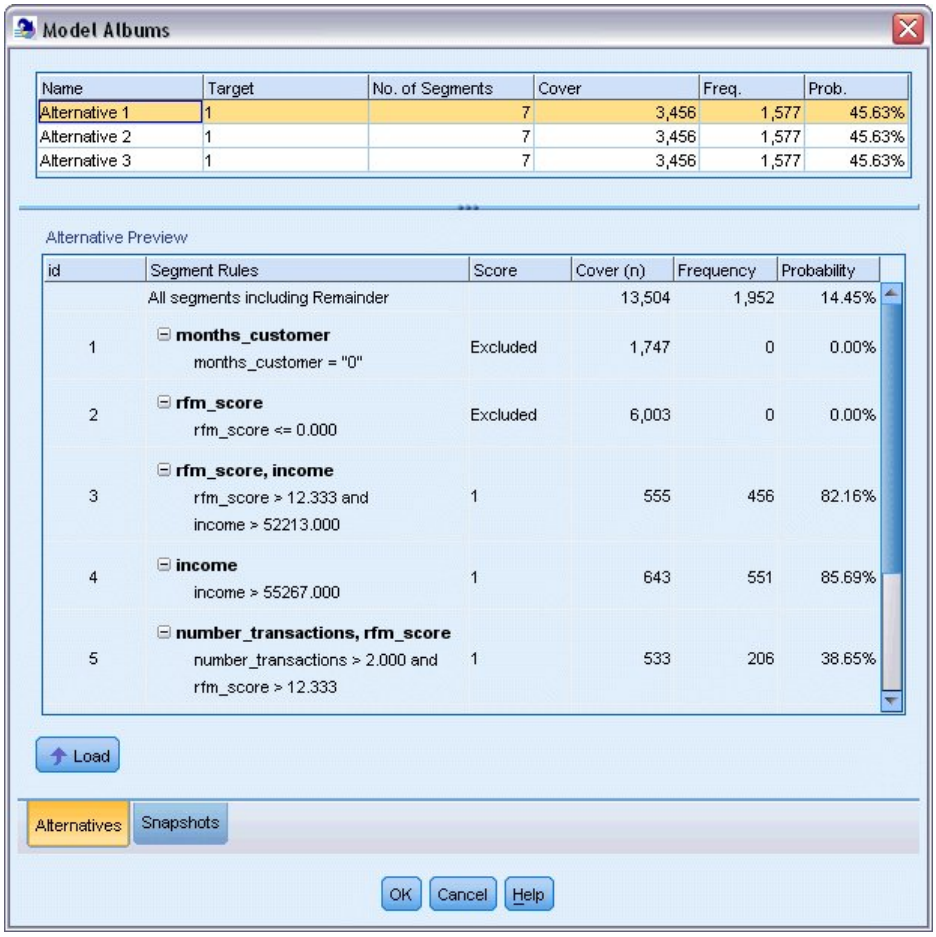


圖 *133.* 組合模型的替代模型

24. 預覽第一個替代模型，然後按一下**載入**以將它設為工作中模型。

使用 Excel 計算自訂測量

1. 若要進一步瞭解模型在實際狀況中如何執行，請從「工具」功能表選擇**組織模型測量**。

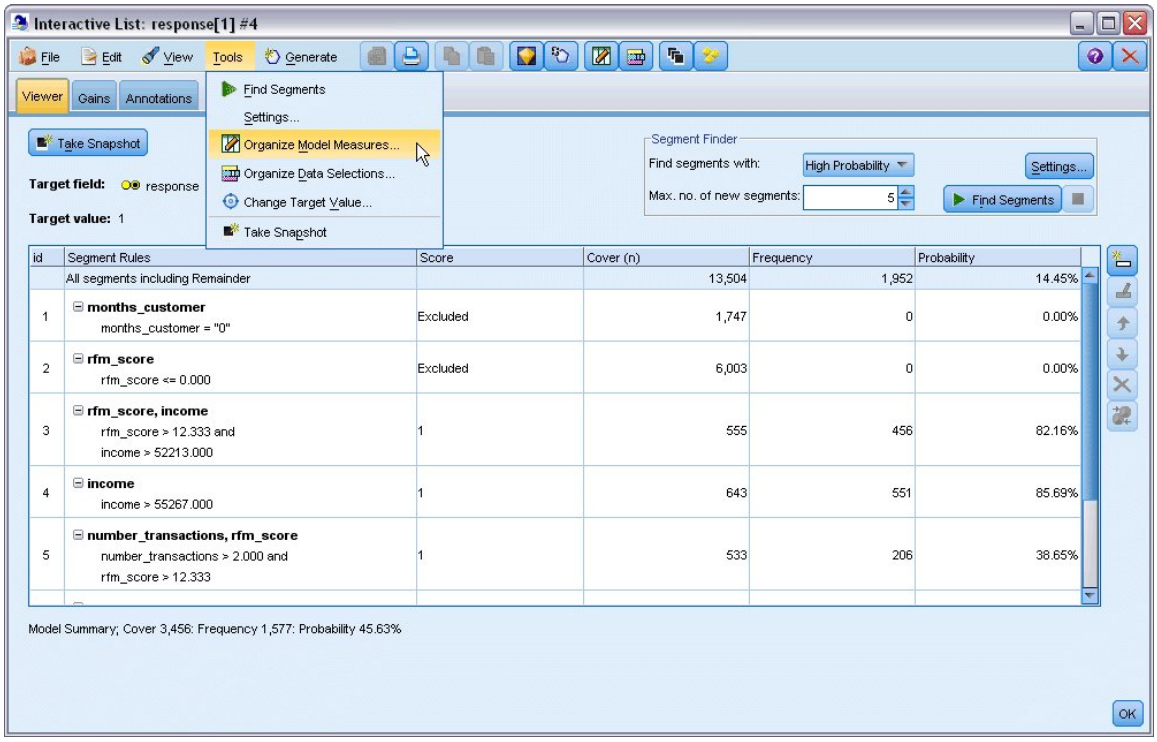


圖 *134.* 組織模型測量

「組織模型測量」對話框可讓您選擇要顯示在「互動式清單」檢視器中的測量（或直欄）。您也可以指定是根據所有記錄還是一個選取的子集來計算測量，並且可選擇顯示一個圓餅圖而非數字（適用的話）。

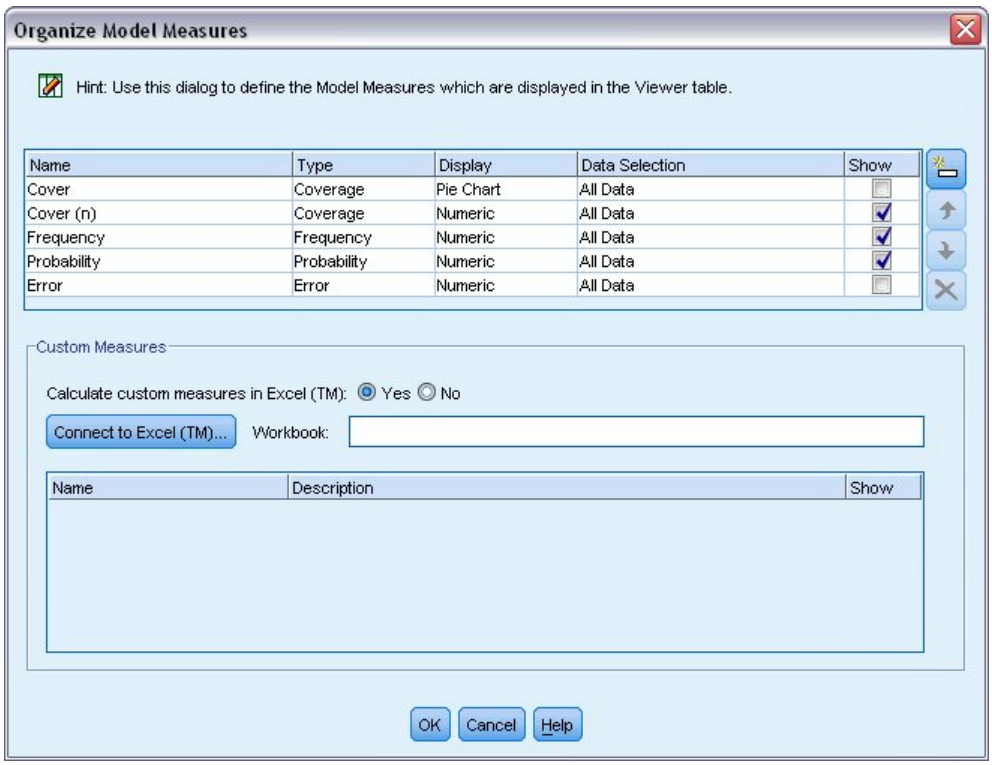


圖 *135.* 組織模型測量對話框

此外，如果您安裝了 Microsoft Excel，則可以鏈結至將會計算自訂測量的 Excel 範本，並將其新增至互動式顯示畫面。

2. 在「組織模型測量」對話框中，將**在 Excel (TM) 中計算自訂測量** 設為**是**。
3. 按一下**連接至 Excel (TM)**
4. 選取 *template_profit.xlt* 活頁簿（其位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾的 *streams* 之下），然後按一下**開啟**以啟動試算表。

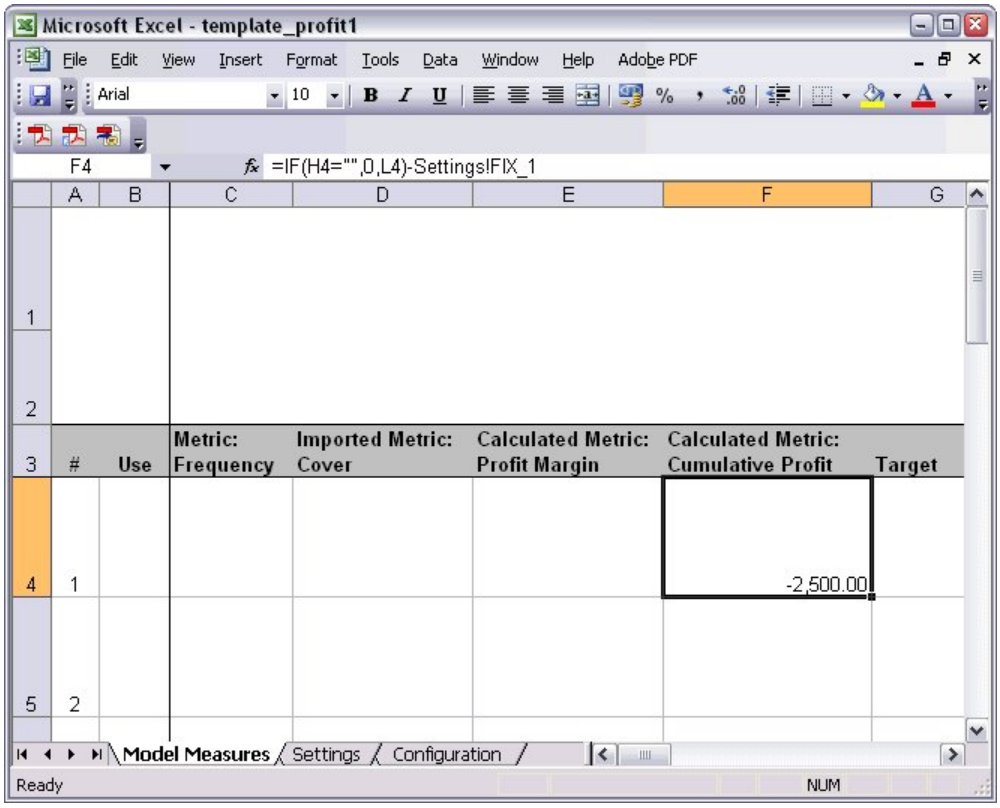


圖 *136. Excel* 模型測量工作表

Excel 範本包含三個工作表：

- **模型測量**顯示從模型中匯入的模型測量，並計算自訂測量以匯回到模型。
- **設定**包含要用來計算自訂測量的參數。
- **配置**定義要匯入的測量以及要匯出至模型的測量。

匯回至模型的測量包括：

- **利潤**。區段中的淨收益
- **累積利潤**。行銷活動中的總利潤

如下列公式所定義：

利潤 = 頻率 * 每個回應者的收益 - 涵蓋範圍 * 可變成本

累積利潤 = 總利潤 - 固定成本

請注意，「頻率」和「涵蓋範圍」是從模型中匯入的。

成本和收益參數由使用者在「設定」工作表上指定。

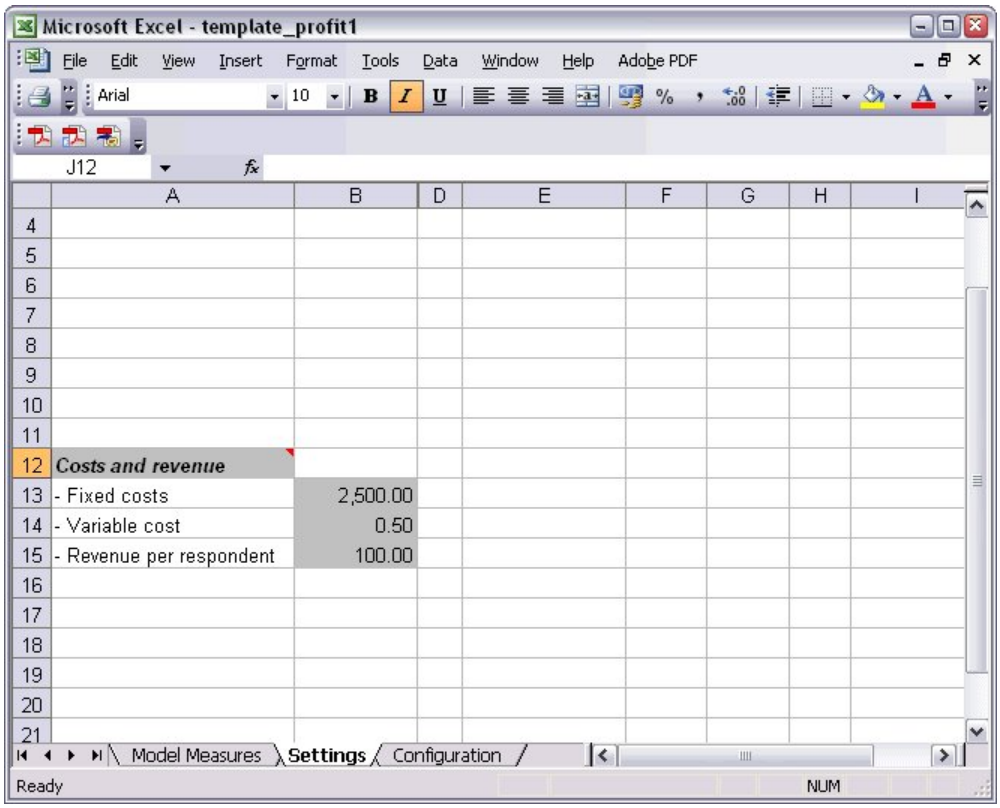


圖 *137. Excel* 設定工作表

固定成本是行銷活動的設定成本，例如設計和規劃。

可變成本是將提議延伸到每個客戶的成本，例如封套和戳記。

每個回應者的收益是來自回應提議之客戶的淨收益。

5. 若要完成鏈結回模型，請使用 Windows 工作列（或按 Alt+Tab）以導覽回到「互動式清單」檢視器。

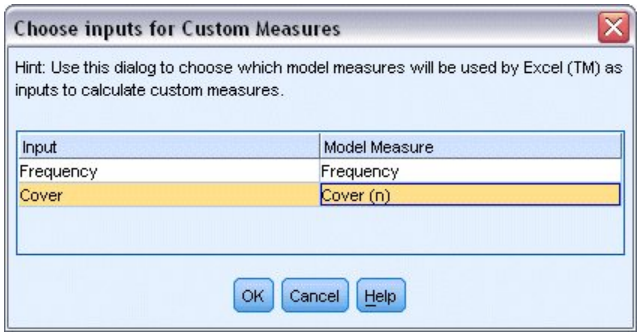


圖 *138.* 選擇自訂測量的輸入

這時會顯示「選取自訂測量的輸入」對話框，可讓您將模型中的輸入對映至範本中所定義的特定參數。左欄列出可用的測量，而右欄將這些測量對映至在「配置」工作表中定義的試算表參數。

6. 在**模型測量**直欄中，針對相應的輸入選取**頻率**和**涵蓋範圍 (n)**，然後按一下**確定**。

在本案例中，範本中的參數名稱（頻率和涵蓋範圍 (n)）剛好符合輸入，但也可以使用其他名稱。

7. 在「組織模型測量」對話框中按一下**確定**以更新「互動式清單」檢視器。

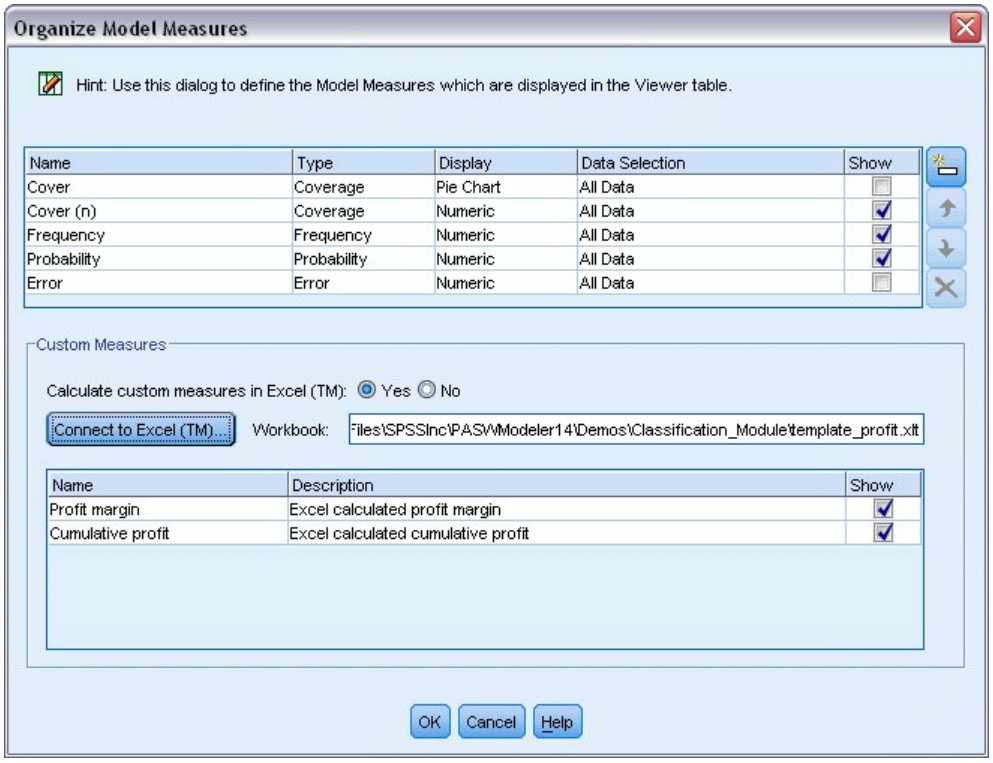


圖 *139.* 「組織模型測量」對話框顯示來自 *Excel* 的自訂測量

現在，新測量已新增為視窗中的新直欄，並且將會每次更新模型時重新計算。

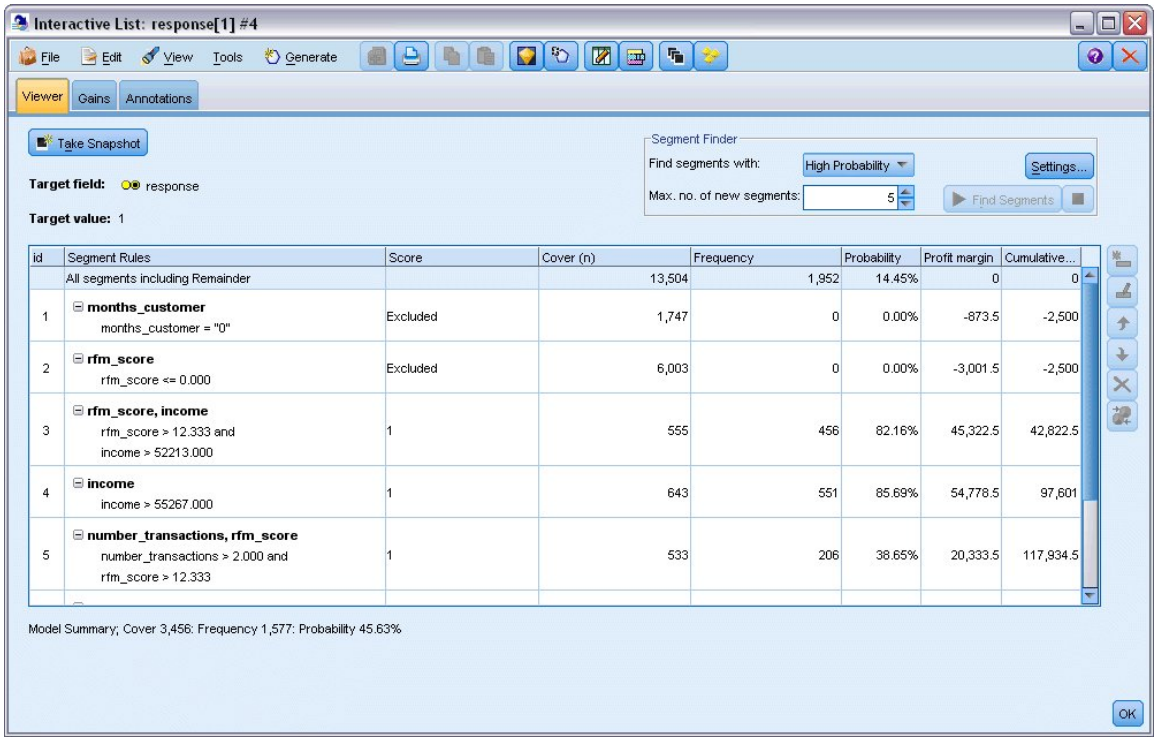


圖 *140.* 來自 *Excel* 的自訂測量顯示在互動式清單檢視器中

透過編輯 Excel 範本，可以建立任意數量的自訂測量。

修改 Excel 範本

雖然 IBM SPSS Modeler 隨預設 Excel 範本一起提供以與「互動式清單」檢視器一起使用，但您可能希望變更設定或新增自己的設定。例如，範本中的成本對您的組織而言可能不正確，需要修正。

附註：如果確實修改現有範本或建立自己的範本，請記得使用 Excel 2003 *.xlt* 字尾來儲存檔案。

若要以新的成本和收益詳細資料來修改預設範本，並以新圖來更新「互動式清單」檢視器，請執行下列動作：

1. 在「互動式清單」檢視器中，從「工具」功能表選擇**組織模型測量**。
2. 在「組織模型測量」對話框中，按一下**連接至 Excel™**。
3. 選取 *template_profit.xlt* 活頁簿，按一下**開啟**以啟動試算表。
4. 選取「設定」工作表。
5. 編輯**固定成本**以變更為 3,250.00，將**每個應答者的收益**變更為 150.00。

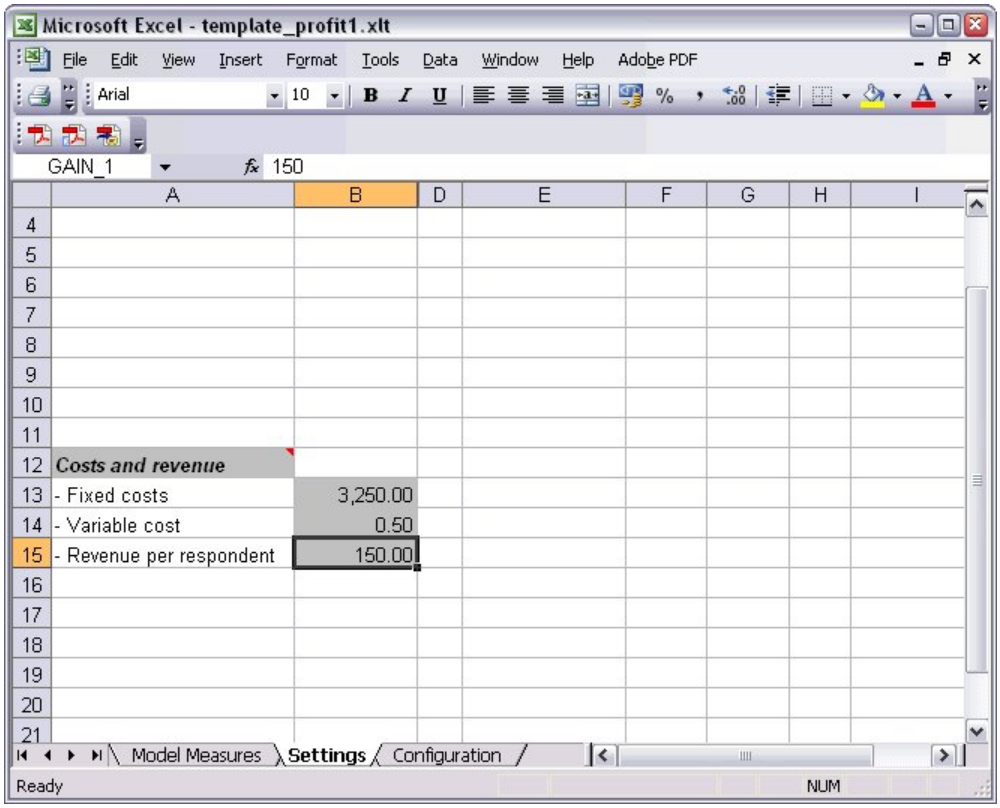


圖 *141.* 在 *Excel* 設定工作表上已修改的值

6. 使用唯一的相關檔名來儲存已修改的範本。確保它的副檔名為 Excel 2003 *.xlt*。

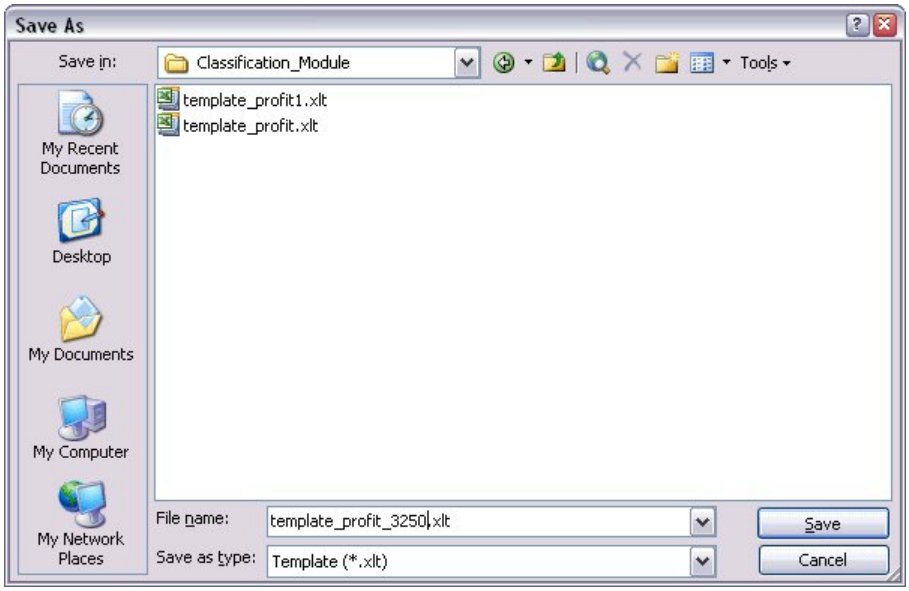


圖 *142.* 儲存已修改的 *Excel* 範本

7. 使用 Windows 工作列（或按 Alt+Tab）以導覽回到「互動式清單」檢視器。

在「選擇自訂測量的輸入」對話框中，選取您要顯示的測量，然後按一下**確定**。

8. 在「組織模型測量」對話框中，按一下**確定**以更新「互動式清單」檢視器。

很明顯，本範例僅顯示了一種簡單的方式來修改 Excel 範本；您可以進一步變更以從「互動式清單」檢視器提取資料並傳遞資料至其中，或在 Excel 中工作來產生其他輸出，例如圖形。

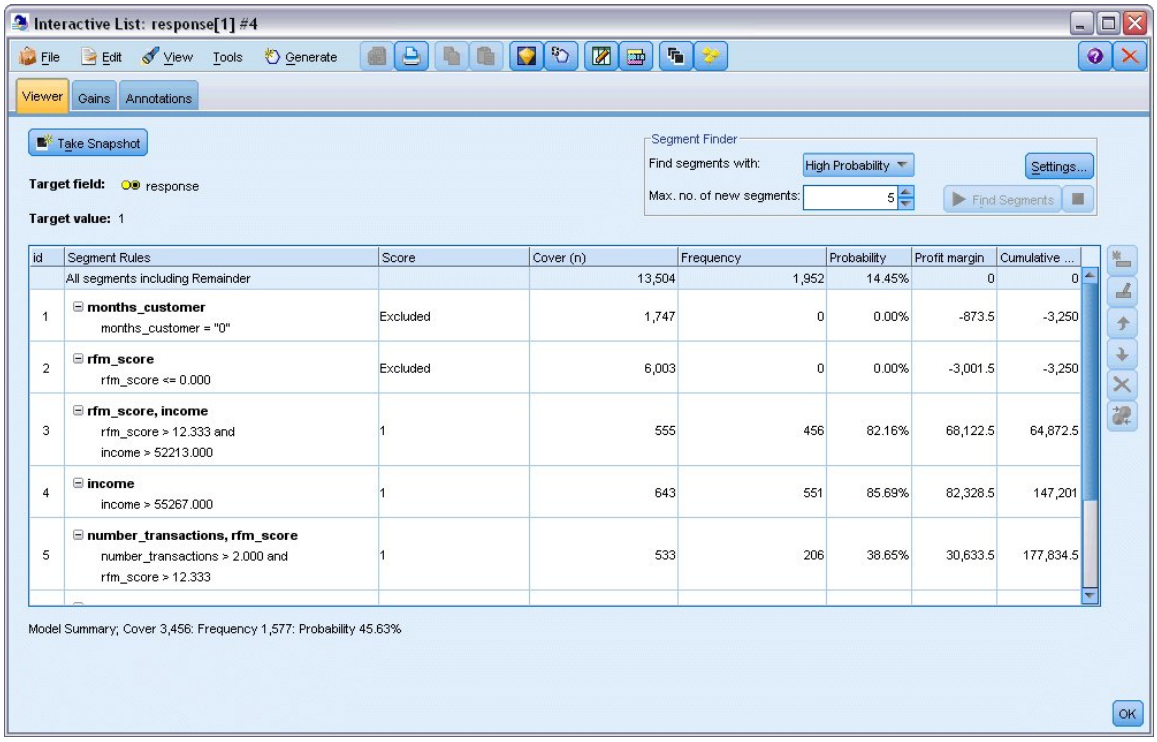


圖 *143.* 來自 *Excel* 的已修改測量顯示在互動式清單檢視器中

儲存結果

若要儲存模型以供稍後用於互動式階段作業，您可以擷取模型的 Snapshot，它將在 Snapshot 標籤上列出。您可以在互動式階段作業期間隨時返回到任何儲存的 Snapshot。

使用這種方式繼續，您可以實驗其他採礦作業來搜尋其他區段。您也可以編輯現有區段、根據自己的商業規則插入自訂區段、建立資料選項以最佳化特定群組的模型，以及採用數種其他方式來自訂模型。最終，您可以適當地明確併入或排除每個區段以指定每個區段如何評分。

如果對結果滿意，則可以使用「產生」功能表來產生可以新增至串流或部署進行評分的模型。

此外，若要儲存互動式階段作業的現行狀態以便在其他日期使用，請從「檔案」功能表中選擇**更新建模節點**。這將會以目前設定來更新「決策清單」建模節點，包括採礦作業、模型 Snapshot資料選擇和自訂測量。下次執行串流時，只需確保在「決策清單」建模節點中選取了**使用儲存的階段作業資訊**以將階段作業還原為其現行狀態。

第 12 章 分類電信客戶（多項式邏輯迴歸）

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。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。

例如，假設某電信公司根據服務使用型樣來切割客戶群，將客戶分成四組。如果可以使用人口資料來預測群組成員資格，則可以針對個別潛在客戶自訂報價。

此範例使用名為 *telco_custcat.str* 的串流，其參照的資料檔名為 *telco.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*telco_custcat.str* 檔案位於 *streams* 目錄中。

該範例的重點集中在使用人口資料來預測使用型樣。目標欄位 *custcat* 有四個可能的值，對應於四組客戶，如下所示：

數值	標籤
1	基本服務
2	電子服務
3	附加服務
4	總服務

因為目標有多個種類，所以會使用多項式模型。如果目標有兩個相異的種類，例如，是/否、true/false 或流失/不流失，則可以改為建立二項式模型。請參閱第 133 頁的第 13 章, 『電信客戶流失（二項式邏輯迴歸）』主題，以取得更多資訊。

建置串流

1. 新增指向 *Demos* 資料夾中 *telco.sav* 的「統計量檔案」來源節點。

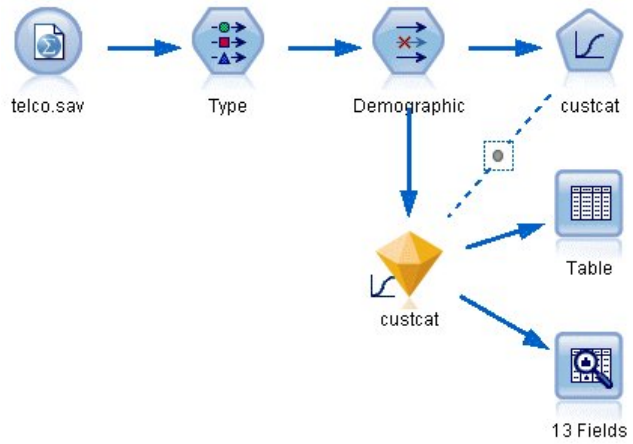


圖 *144.* 用來使用多項式邏輯迴歸分類客戶的樣本串流

a. 新增「類型」節點並按一下**讀取值**，確保已正確設定所有測量層次。例如，值為 0 和 1 的大部分欄位可以視為旗標。

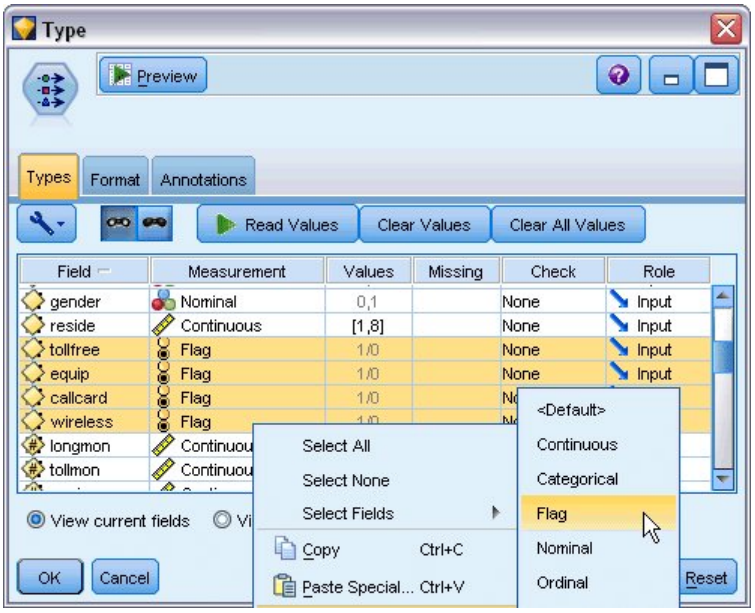


圖 *145.* 設定多個欄位的測量層次

提示：若要變更具有類似值（例如 0/1）的多個欄位的內容，請按一下值 直欄標頭以依值排序欄位，然後在使用滑鼠或方向鍵的同時按住 shift 鍵，以選取您要變更的所有欄位。然後，您可以在選項上按一下滑鼠右鍵以變更所選欄位的測量層次或其他屬性。

請注意，將 *gender* 視為具有一組值（包含兩個值）的欄位而不是一個旗標更為正確，因此請將其測量值保留為**名義**。

b. 將 *custcat* 欄位的角色設為**目標**。所有其他欄位應該將其角色設為**輸入**。

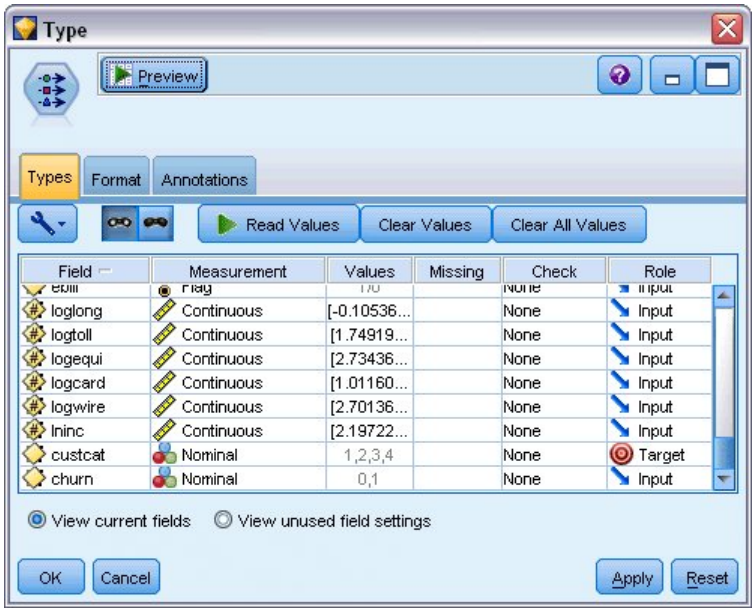


圖 *146.* 設定欄位角色

由於此範例的焦點是個人背景資訊，因此使用「過濾器」節點可以只包括相關欄位（*region*、*age*、*marital*、*address*、*income*、*ed*、*employ*、*retire*、*gender*、*reside* 和 *custcat*）。可以排除其他欄位以便進行此分析。

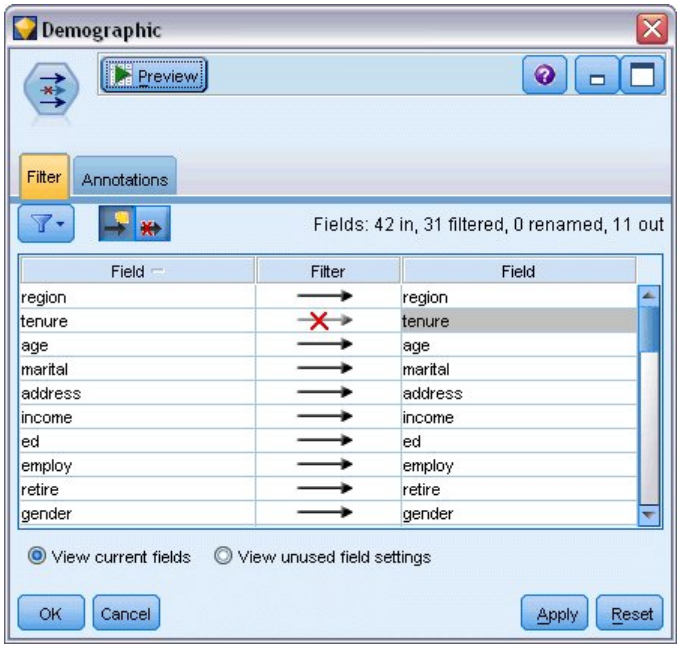


圖 *147.* 根據個人背景資訊欄位過濾

（此外，您可以針對這些欄位將角色變更為**無**而非排除他們，或是選取您要用在建模節點中的欄位。）

2. 在「邏輯」節點中，按一下**模型**標籤並選取**逐步迴歸分析**法。選取**多項式**、**主效應**及**在方程式中包含常數**。

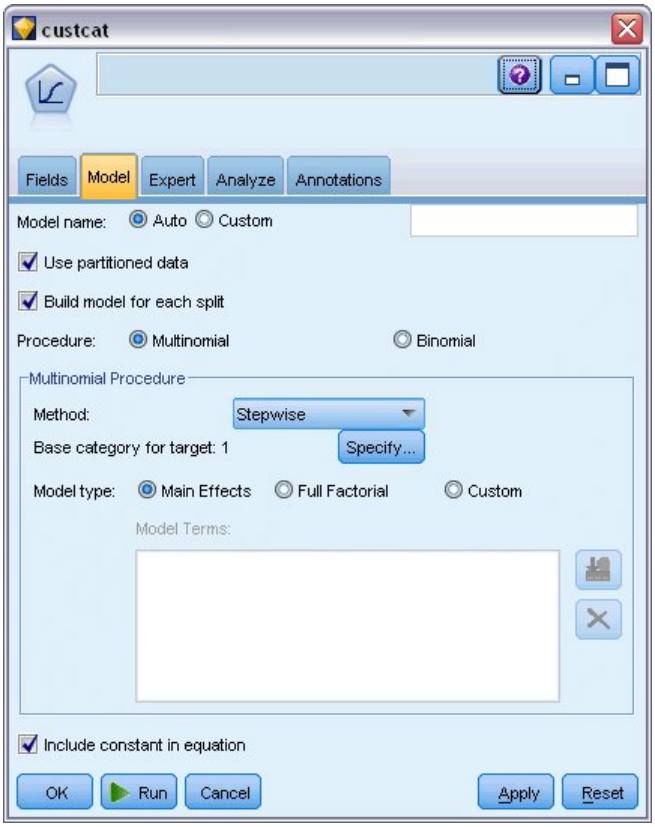


圖 *148.* 選擇模型選項

將「目標的基本種類」保留為 1。該模型會將其他客戶與訂閱「基本服務」的那些客戶進行比較。

3. 在「專家」標籤上，選取**專家**模式，選取**輸出**，並在「進階輸出」對話框中選取**分類表**。

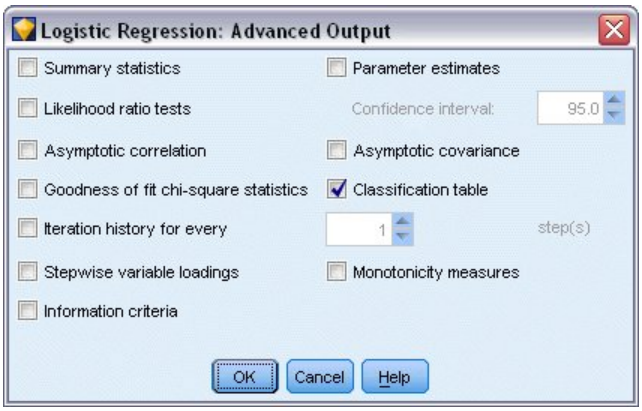


圖 *149.* 選擇輸出選項

瀏覽模型

1. 執行節點以產生模型，該模型會新增到右上角的「模型」選用區。若要檢視其詳細資料，請用滑鼠右鍵按一下產生的模型節點，並選擇**瀏覽**。

模型標籤會顯示用來將記錄指派給目標欄位每一個種類的方程式。有四種可能的種類，其中一種是基本種類，不會顯示該種類的方程式詳細資料。其餘三個方程式的詳細資料會顯示，其中種類 3 代表「附加服務」，依此類推。

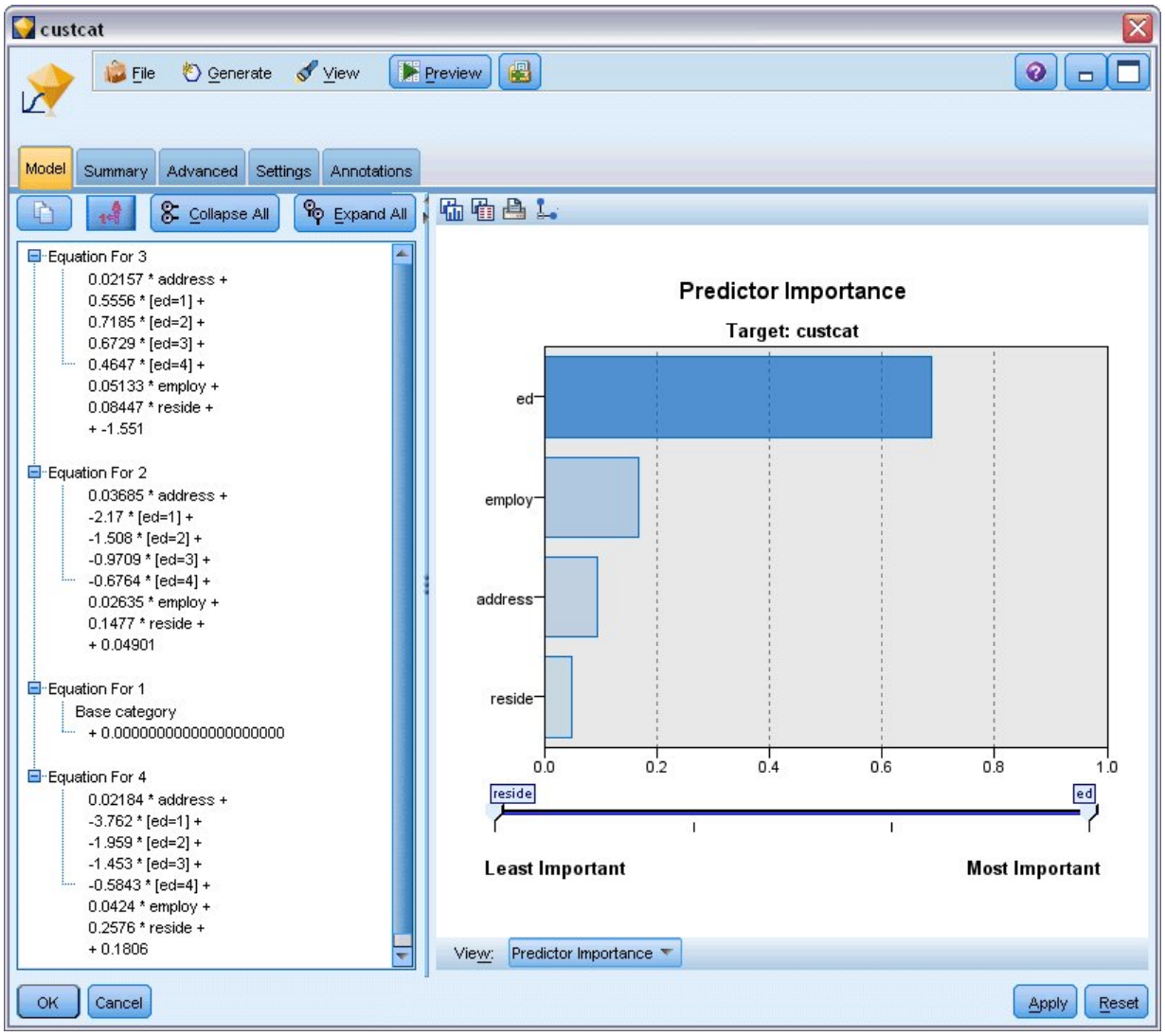


圖 *150.* 瀏覽模型結果

「摘要」標籤會顯示（除其他事物以外）模型使用的目標及輸入（預測值欄位）。請注意，這些是基於逐步迴歸分析法實際選擇的欄位，而非提交以供考量的完整清單。

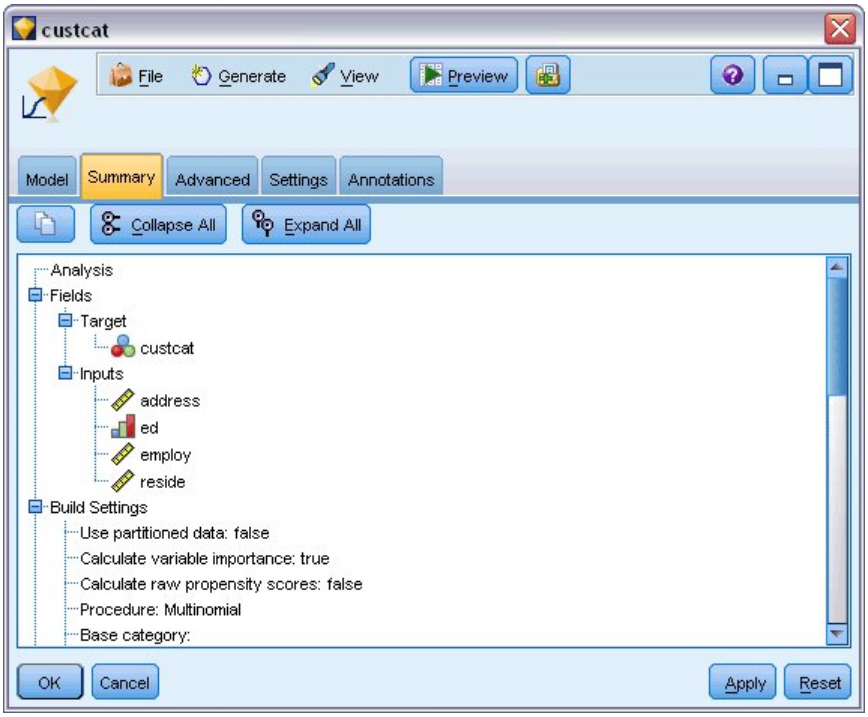


圖 *151.* 顯示目標及及輸入欄位的模型摘要

在「進階」標籤上顯示的項目視在建模節點中「進階輸出」對話框上選取的選項而定。

一律顯示的一個項目是「觀察值處理摘要」，其會顯示處於目標欄位每一個種類的記錄百分比。這會為您提供虛無模型以用作比較的基準。

如果不建置使用預測值的模型，您的最佳猜測會是將所有客戶全都指派給最常見的群組，即用於「附加服務」的群組。

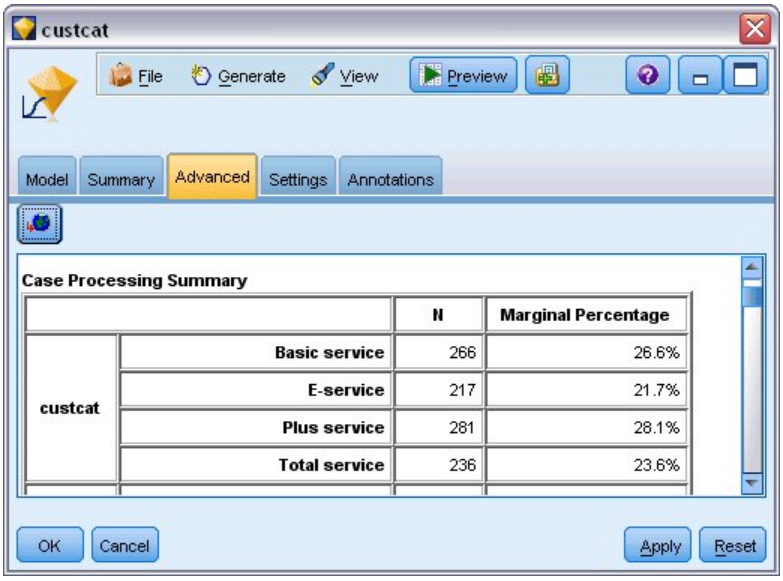


圖 *152.* 觀察值處理摘要

根據訓練資料，如果將所有客戶都指派給虛無模型，則您的正確率為 281/1000 = 28.1%。「進階」標籤包含進一步的資訊，您可以使用這些資訊檢查模型的預測。然後您可以將預測與虛無模型的結果進行比較，以查看模型與資料的適合度。

在「進階」標籤底部，分類表會顯示模型的結果，即正確率 39.9%。

特別是，您的模型的優點在於可以識別出「總服務」客戶（種類 4），但在識別「電子服務」客戶（種類 2）方面卻做得非常差。如果您想要針對種類 2 中的客戶提高精確度，則可能需要找到另一個預測值來識別他們。

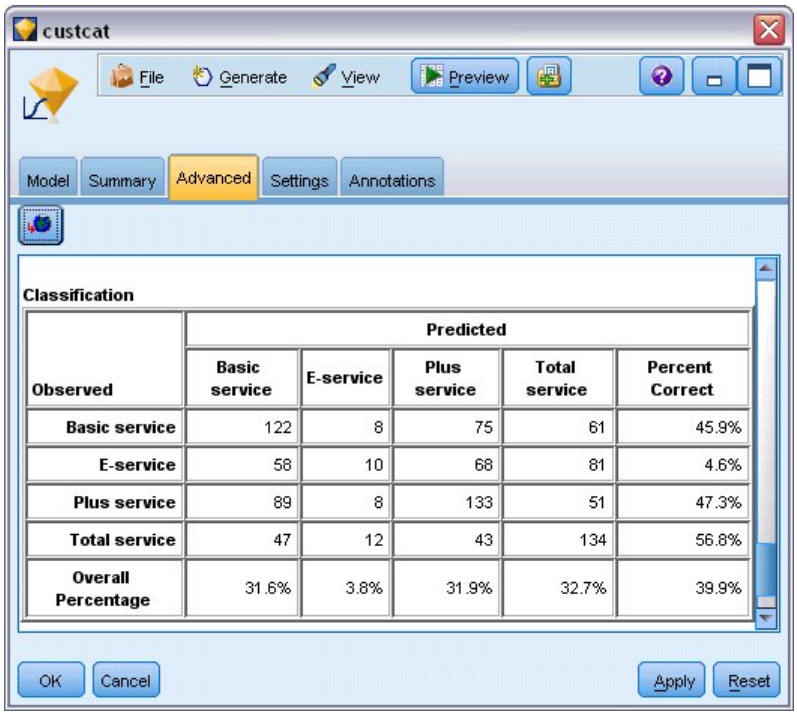


圖 *153.* 分類表

視您想要預測的內容而定，該模型可能會完全滿足您的需要。例如，如果您對識別種類 2 中的客戶並不關心，則該模型對您來說是足夠精確的。在這種情況下，電子服務可能是低價促銷，幾乎不會帶來利潤。

如果，例如，您的最高投資報酬率來自處於種類 3 或 4 的客戶，則該模型可能會為您提供需要的資訊。

為了評量模型對資料的實際適合度，建置模型時會在「進階輸出」對話框中提供多個診斷程式。在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出，該手冊可從安裝磁碟的 *Documentation* 目錄取得。

另請注意，這些結果僅基於訓練資料。若要評量現實世界中模型對其他資料的廣義化程度，您可以使用「分割區」節點來保留一部分記錄以便進行測試及驗證。

第 13 章 電信客戶流失（二項式邏輯迴歸）

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。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。

此範例使用名為 *telco_churn.str* 的串流，其參照的資料檔名為 *telco.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*telco_churn.str* 檔案位於 *streams* 目錄中。

例如，假設某個電信服務提供商關心流失到競爭對手那裡的客戶數。如果可以使用服務使用情況資料來預測哪些客戶有可能轉向其他提供商，則可以自訂報價來留住盡可能多的客戶。

此範例的重點集中在利用使用情況資料來預測客戶的流失。因為目標有兩個相異的種類，所以會使用二項式模型。如果目標有多個種類，則可以改為建立多項式模型。請參閱第 125 頁的第 12 章, 『分類電信客戶（多項式邏輯迴歸）』主題，以取得更多資訊。

建置串流

1. 新增指向 *Demos* 資料夾中 *telco.sav* 的「統計量檔案」來源節點。

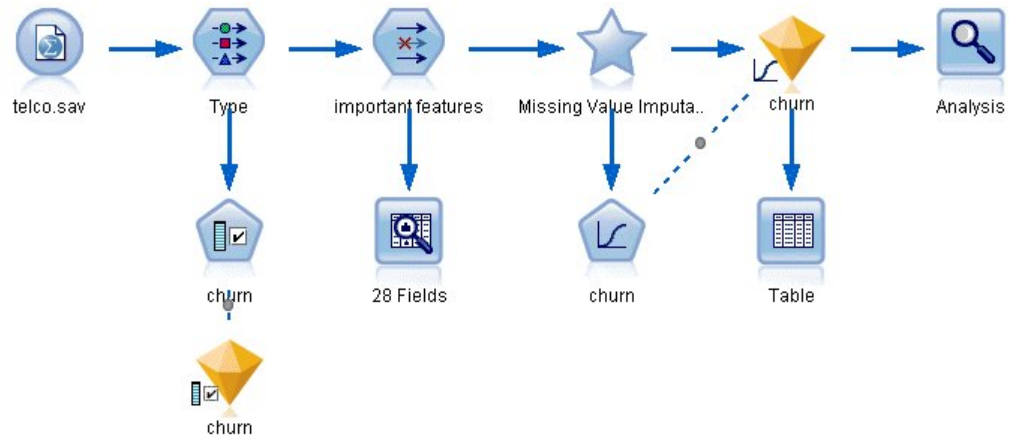


圖 *154.* 用來使用二項式邏輯迴歸分類客戶的樣本串流

2. 新增「類型」節點來定義欄位，從而確保所有測量層次都正確設定。例如，包含值 0 及 1 的大部分欄位都可視為旗標，但特定欄位（例如，性別）會更精確地被視為包含兩個值的名義欄位。

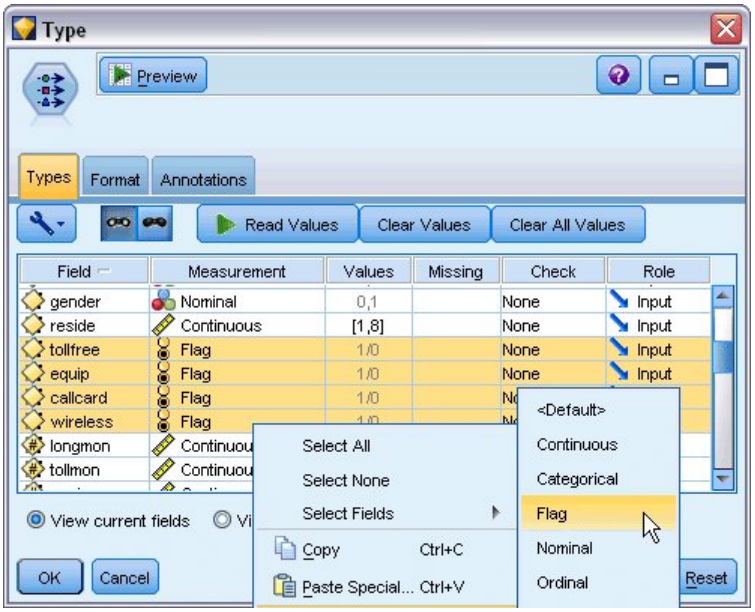


圖 *155.* 設定多個欄位的測量層次

提示：若要變更具有相似值（例如 0/1）的多個欄位的內容，請按一下值直欄標頭來依值排序欄位，然後按住 Shift 鍵，同時使用滑鼠或方向鍵來選取要變更的所有欄位。然後您可以用滑鼠右鍵按一下選定內容來變更所選欄位的測量層次或其他屬性。

3. 將流失欄位的測量層次設定為**旗標**，並將角色設定為**目標**。所有其他欄位都應該將其角色設定為**輸入**。

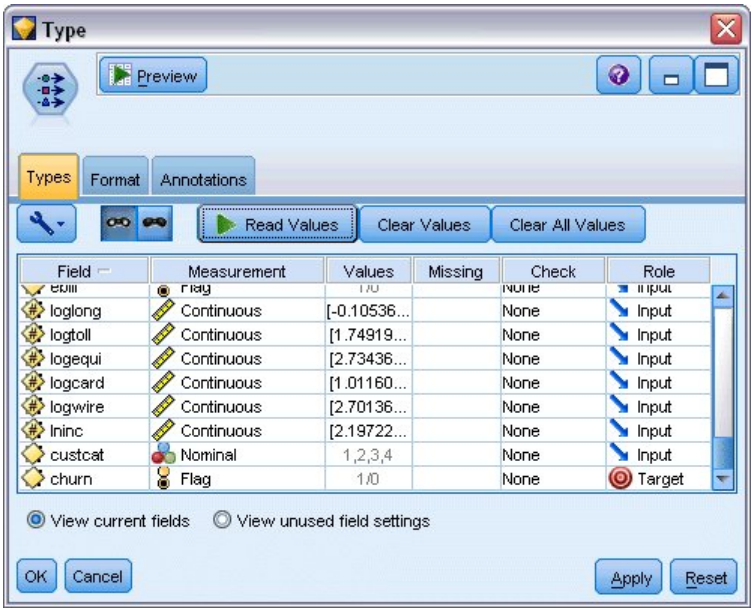


圖 *156.* 設定流失欄位的測量層次及角色

4. 將「功能選擇」建模節點新增至「類型」節點。

使用「功能選擇」節點可讓您移除不提供關於預測值/目標關係的任何有用資訊的預測值或資料。

5. 執行串流。

6. 開啟產生的模型區塊，然後從**產生**功能表中，選擇**過濾器**來建立「過濾器」節點。

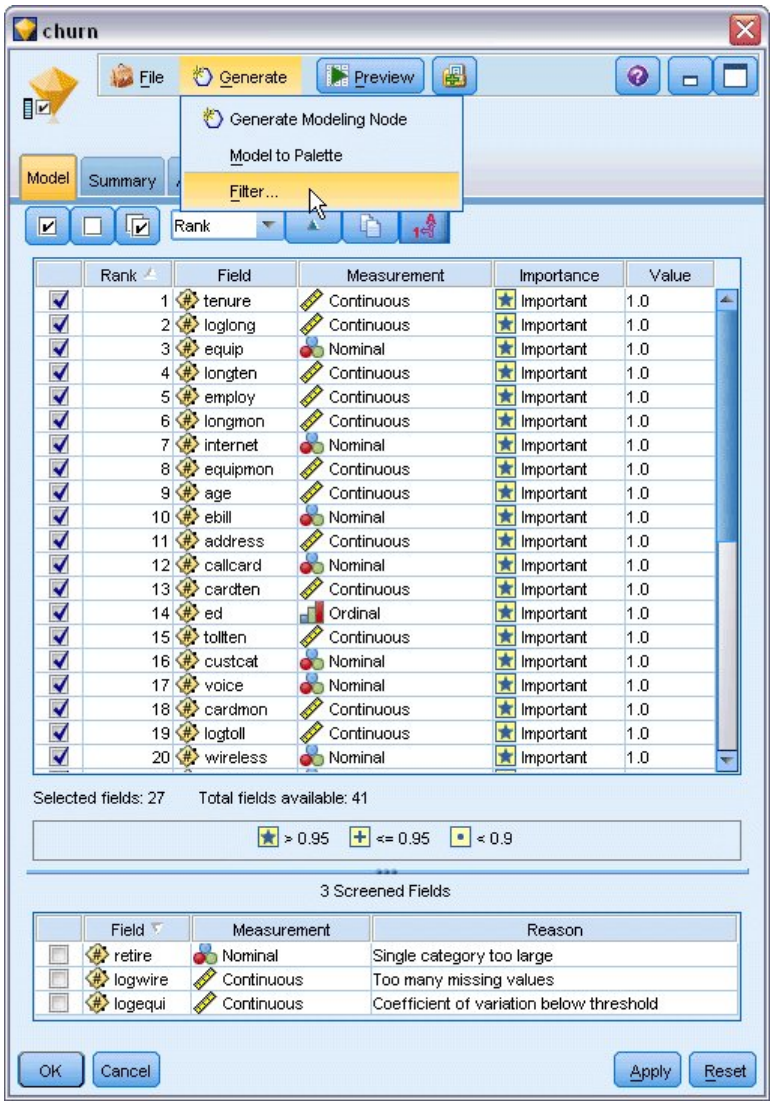


圖 *157.* 從功能選擇節點中產生過濾器節點

並非 *telco.sav* 檔案中的所有資料都對流失預測有用。您可以使用過濾器來僅選取被視為具有重要用途的資料作為預測值。

7. 在「產生過濾器」對話框中，選取**標示為重要的所有欄位**，然後按一下**確定**。
8. 將產生的「過濾器」節點連接至「來源」節點。

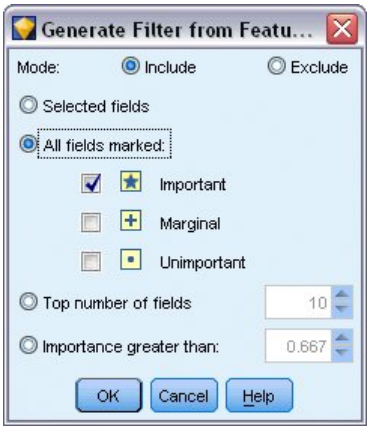


圖 *158.* 選取重要欄位

9. 將「資料審核」節點連接至產生的「過濾器」節點。

   開啟「資料審核」節點，然後按一下**執行**。

10. 在「資料審核」瀏覽器的「品質」標籤上，按一下完成 % 直欄來按數值遞增順序排序直欄。這可讓您識別出具有大量遺漏資料的任何欄位；在這種情況下，您需要修正的唯一欄位是 *logtoll*，其完成百分比不到 50%。
11. 在 *logtoll* 的插補遺漏直欄中，按一下**指定**。

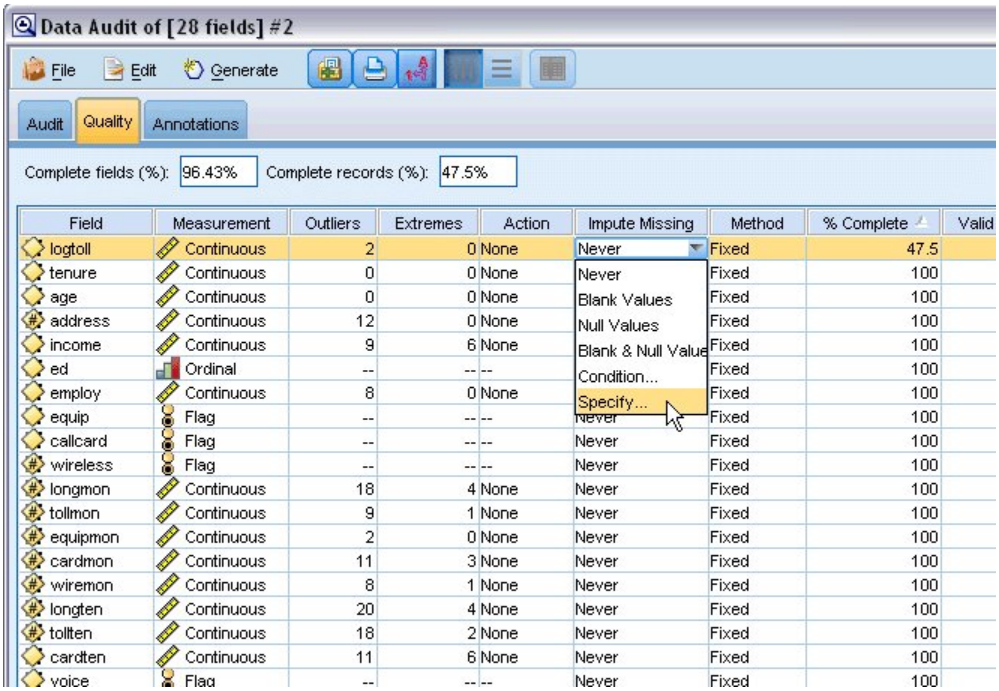


圖 *159.* 插補 *logtoll* 的遺漏值

12. 針對**插補條件**，選取**空白值與空值**。針對**固定為**，選取**平均數**，然後按一下**確定**。

    選取**平均數**可確保插補的值不會對整體資料中所有值的平均數產生不利影響。

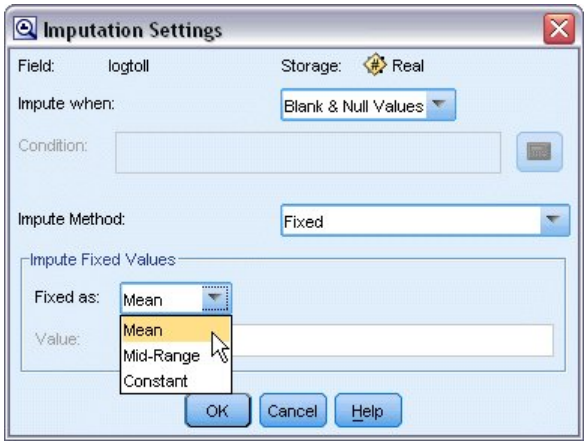


圖 *160.* 選取插補設定

13. 在「資料審核」瀏覽器的「品質」標籤上，產生「遺漏值 SuperNode」。若要進行此功能，在功能表中選擇：

    產生 > 遺漏值 SuperNode

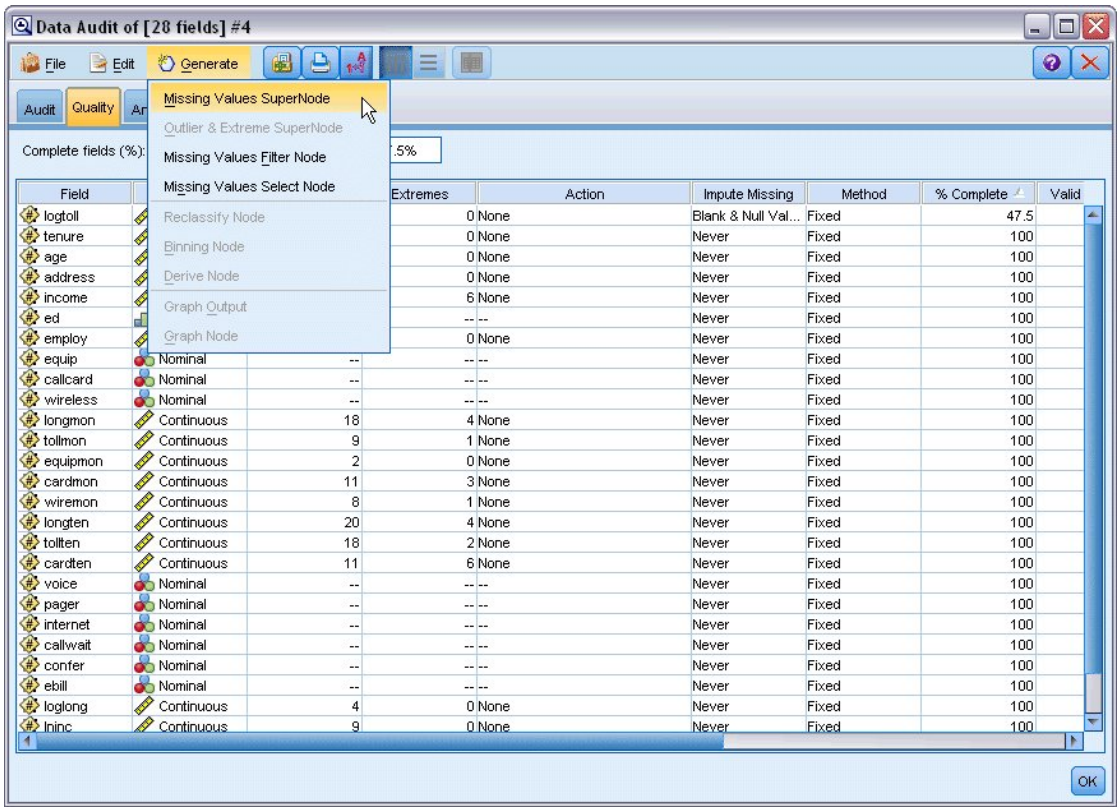


圖 *161.* 產生遺漏值 *SuperNode*

在「遺漏值 SuperNode」對話框中，將**樣本大小**增加到 50%，然後按一下**確定**。

SuperNode 即會在串流畫布上顯示，標題為：遺漏值插補。

14. 將 SuperNode 連接至「過濾器」節點。

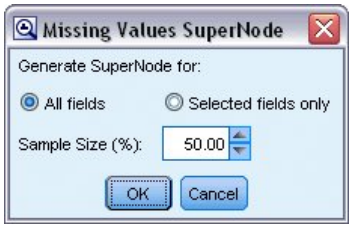


圖 *162.* 指定樣本大小

15. 將「邏輯」節點新增至 SuperNode。
16. 在「邏輯」節點中，按一下「模型」標籤並選取**二項式**程序。在二項式程序區域中，選取**向前**法。

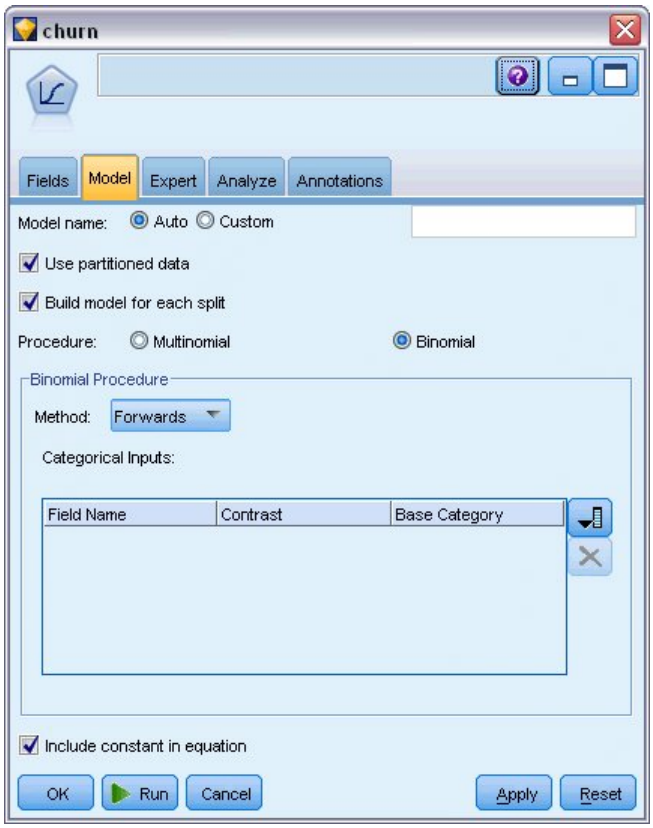


圖 *163.* 選擇模型選項

17. 在「專家」標籤上，選取**專家**模式，然後按一下 **輸出**。即會顯示「進階輸出」對話框。
18. 在「進階輸出」對話框中，選取**在每一個步驟**作為顯示類型。選取**疊代歷程**及**參數估計值**，然後按一下**確定**。

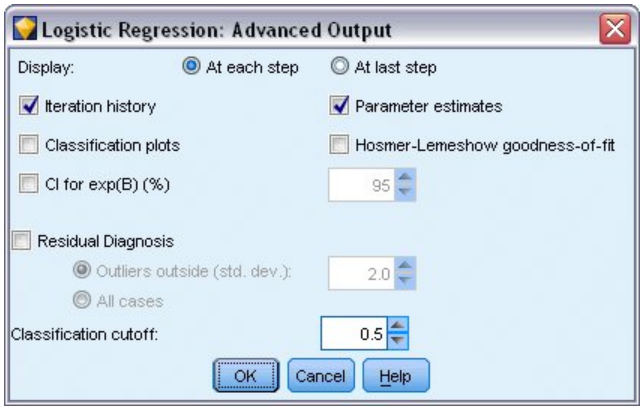


圖 *164.* 選擇輸出選項

瀏覽模型

1. 在「邏輯」節點上，按一下**執行**來建立模型。

   模型區塊會新增至串流畫布，還會新增至右上角的「模型」選用區。若要檢視其詳細資料，請用滑鼠右鍵按一下模型區塊，並選取**編輯**或**瀏覽**。

   「摘要」標籤會顯示（除其他事物以外）模型使用的目標及輸入（預測值欄位）。請注意，這些是基於向前法實際選擇的欄位，而非提交以供考量的完整清單。

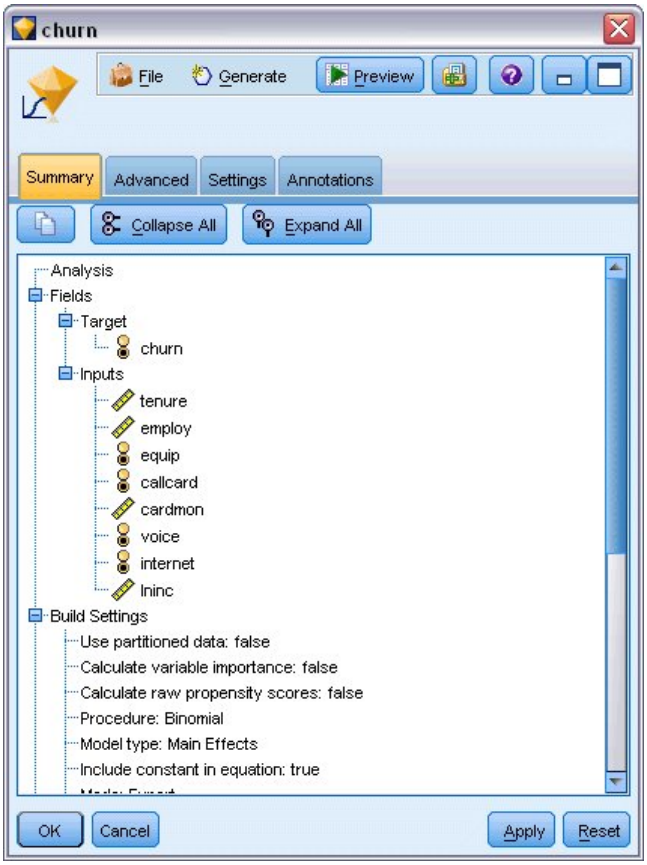


圖 *165.* 顯示目標及及輸入欄位的模型摘要

在「進階」標籤上顯示的項目視在「邏輯」節點中「進階輸出」對話框上選取的選項而定。一律顯示的一個項目是「觀察值處理摘要」，其會顯示分析中包括的記錄數及百分比。此外，它還會列出其中一個以上輸入欄位無法使用的遺漏觀察值數（如果有），以及任何未選取的觀察值數。

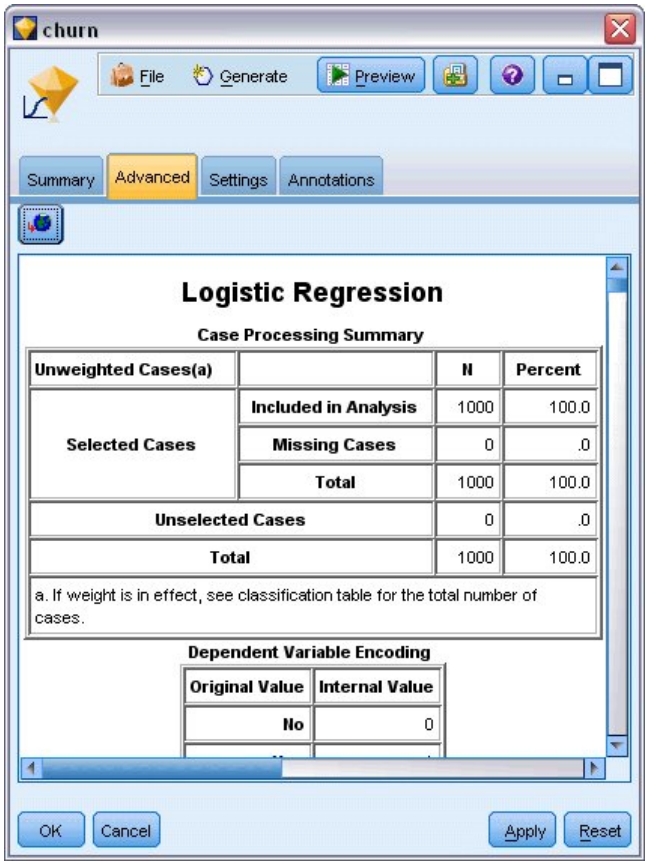


圖 *166.* 觀察值處理摘要

2. 從「觀察值處理摘要」向下捲動來顯示「區塊 0：開始區塊」下方的分類表。

   逐步向前法從虛無模型 — 即沒有預測值的模型 — 開始，該模型可作為基準與最終建置的模型進行比較。按照慣例，虛無模型會將一切預測為 0，因此虛無模型的精確度為 72.6%，只是因為正確預測了 726 位客戶沒有流失。但根本沒有正確預測出流失的客戶。

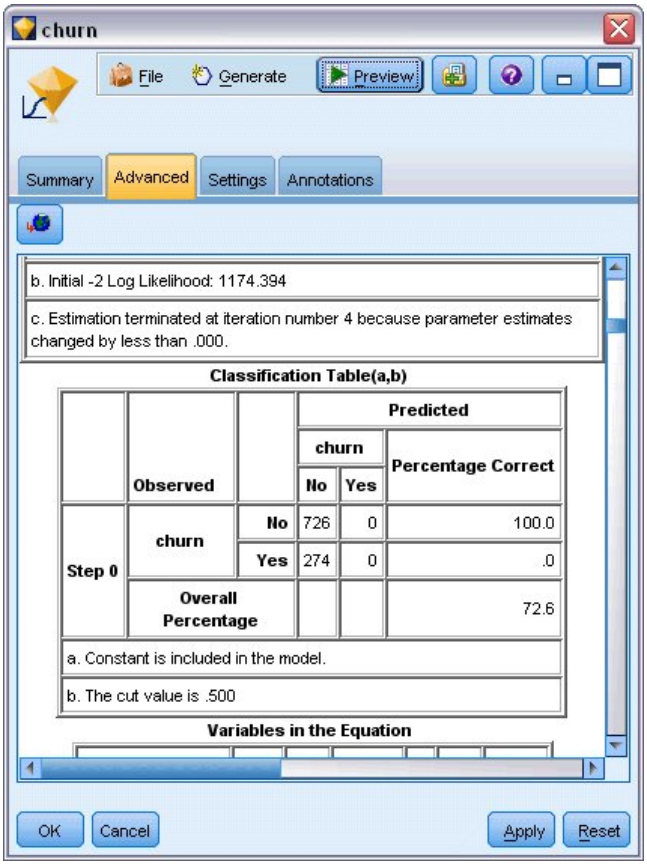


圖 *167.* 開始分類表 - 區塊 *0*

3. 現在向下捲動來顯示「區塊 1：方法 = 逐步向前」下方的分類表。

   此分類表會顯示模型的結果，因為在每一步中都新增了預測值。在第一步中 - 僅在使用了一個預測值之後 - 模型就已將流失預測精確度從 0.0% 增加到了 29.9%

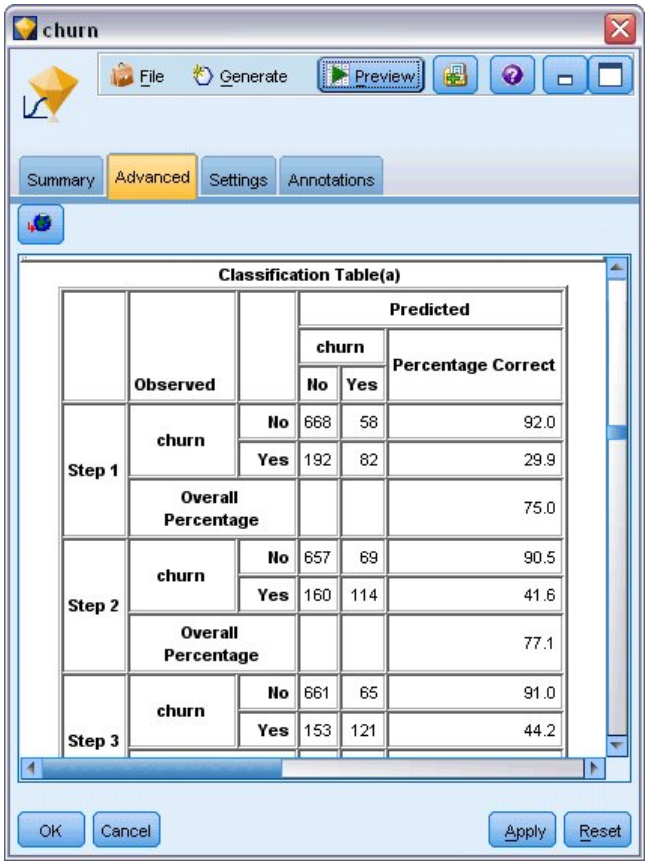


圖 *168.* 分類表 - 區塊 *1*

4. 向下捲動至此分類表的底部。

分類表顯示最後一個步驟是第 8 步。在這一階段，演算法已決定不需要再將更多的預測值新增至模型中。盡管非流失客戶的精確度略微下降到了 91.2%，但流失客戶的預測精確度卻從原始的 0% 上升到了47.1%。這對於不使用任何預測值的原始虛無模型來說是一種顯著的改進。

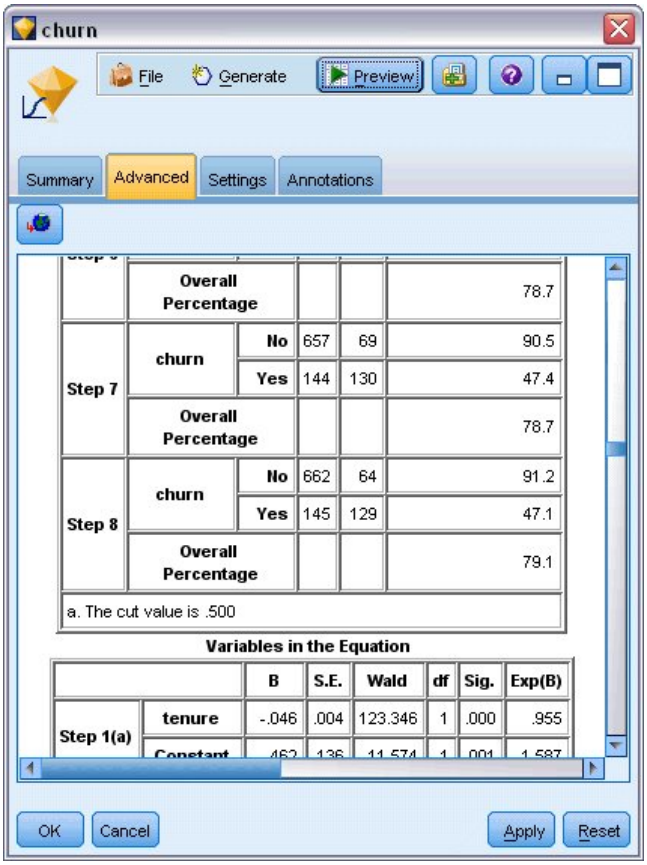


圖 *169.* 分類表 - 區塊 *1*

對於想要減少流失的客戶來說，能夠將流失減少將近一半在保護其收入流方面起到了舉足輕重的作用。

附註：此範例還會顯示將「整體百分比」作為模型精確度的指引在某些情況下如何會造成誤導。原始虛無模型的整體精確度為 72.6%，而最終預測模型的整體精確度為 79.1%；但正如我們所看到的，實際個別種類預測的精確度相差甚遠。

為了評量模型對資料的實際適合度，建置模型時會在「進階輸出」對話框中提供多個診斷程式。在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出，該手冊可從安裝磁碟的 *\Documentation* 目錄取得。

另請注意，這些結果僅基於訓練資料。為了評量現實世界中模型對其他資料的廣義化程度，您會使用「分割區」節點來保留一部分記錄以便進行測試及驗證。

第 14 章 預測頻寬使用率（時間序列）

使用時間序列節點來預測

全國性寬頻服務提供業者的分析師必須提出用戶訂購量的預測，以預測頻寬的使用量。分析師需要對各地市場進行預測，才能得出全國註冊用戶數量。您將使用時間序列建模對各地市場未來三個月註冊用戶數量進行預測。第二個範例顯示如何將輸入的不正確格式來源資料轉換成「時間序列」節點。

這些範例使用參照資料檔案 *broadband_1.sav* 的串流 *broadband_create_models.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*broadband_create_models.str* 檔案位于 *streams* 資料夾中。

最後一個範例示範如何將儲存的模型套用至更新的資料集以便將預測延期三個月。

在 IBM SPSS Modeler 中，您可以在單一作業中產生多個時間序列模型。您將要使用的原始檔擁有 85 個不同市場的時間序列資料，但為了簡化，您只會針對這些市場中的五個市場進行建模，外加針對所有市場的總計進行建模。

broadband_1.sav 資料檔案具有本地 85 個市場中每個市場的每月使用資料。基於此範例的用途，只會使用前五個序列；將針對這五個序列中的每一個以及針對總計建立個別的模型。

該檔案還包括日期欄位，指出每筆記錄的月份和年份。此欄位將用來標記記錄。日期欄位作為字串讀入 IBM SPSS Modeler，但是為了使用 IBM SPSS Modeler 中的欄位，您將使用「過濾器」節點將儲存體類型轉換成數值日期格式。

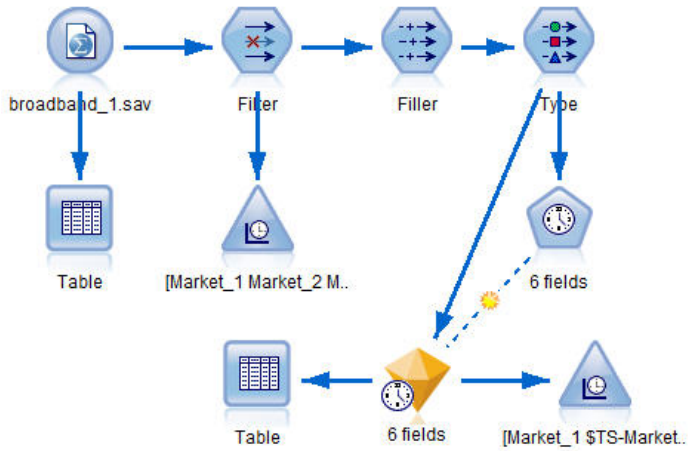
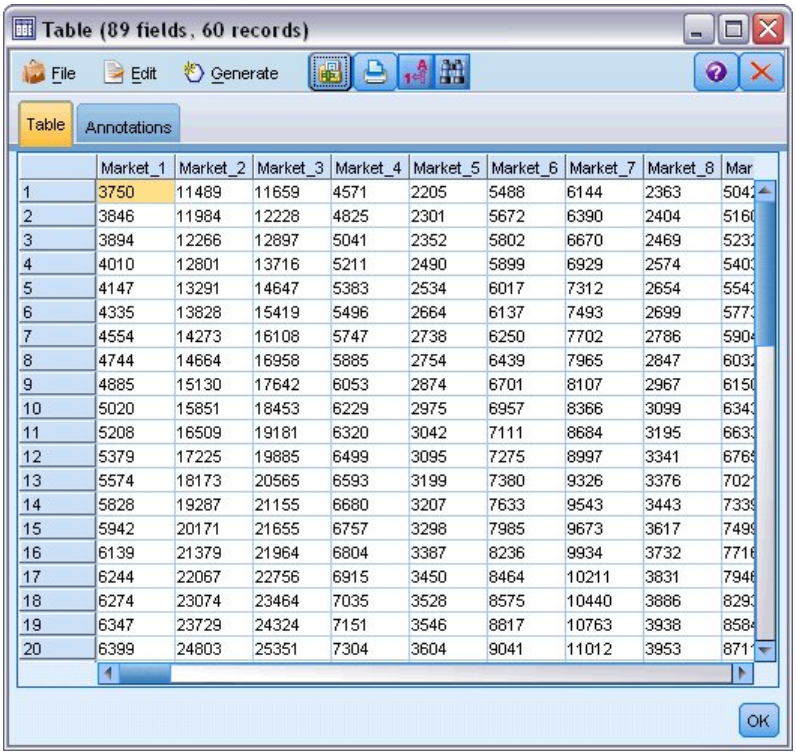


圖 *170.* 顯示時間序列建模的串流範例

「時間序列」節點要求每個序列位於個別的直欄，每個間隔各佔一列。必要的話，IBM SPSS Modeler 會提供方法來轉換資料以符合此格式。



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	5042
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5232
4	4010	12801	13716	5211	2490	5899	6929	2574	5403
5	4147	13291	14647	5383	2534	6017	7312	2654	5543
6	4335	13828	15419	5496	2664	6137	7493	2699	5773
7	4554	14273	16108	5747	2738	6250	7702	2786	5904
8	4744	14664	16958	5885	2754	6439	7965	2847	6032
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6343
11	5208	16509	19181	6320	3042	7111	8684	3195	6633
12	5379	17225	19885	6499	3095	7275	8997	3341	6765
13	5574	18173	20565	6593	3199	7380	9326	3376	7021
14	5828	19287	21155	6680	3207	7633	9543	3443	7339
15	5942	20171	21655	6757	3298	7985	9673	3617	7495
16	6139	21379	21964	6804	3387	8236	9934	3732	7716
17	6244	22067	22756	6915	3450	8464	10211	3831	7948
18	6274	23074	23464	7035	3528	8575	10440	3886	8293
19	6347	23729	24324	7151	3546	8817	10763	3938	8584
20	6399	24803	25351	7304	3604	9041	11012	3953	8711

圖 *171.* 寬頻本地市場的每月訂閱資料

建立串流

1. 建立新串流並新增指向 *broadband_1.sav* 的「統計量檔案」來源節點。
2. 使用「過濾器」節點來濾出 *Market_6* 到 *Market_85* 欄位以及 *MONTH_* 和 *YEAR_* 欄位以簡化模型。

提示：若要在單一作業中選取多個相鄰欄位，請按一下 *Market_6*，按住左滑鼠按鈕並將滑鼠向下拖曳至 *Market_85* 欄位。所選取的欄位會以藍色強調顯示。若要新增其他欄位，請按住 Ctrl 鍵並按一下 *MONTH_* 和 *YEAR_* 欄位。

圖 *172.* 簡化模型

檢驗資料

建置模型前，最好先瞭解您資料的性質。資料是否顯示週期性變動？雖然 Expert Modeler 可以自動尋找每個序列的最佳週期性或非週期性模型，但您通常可以在資料中不存在週期性時透過將搜尋限制為非週期性模型，更快地取得結果。如果不檢查每個本地市場的資料，則我們透過繪製所有五個市場的訂閱者總數的圖形，可以大致瞭解週期性的存在或缺少。

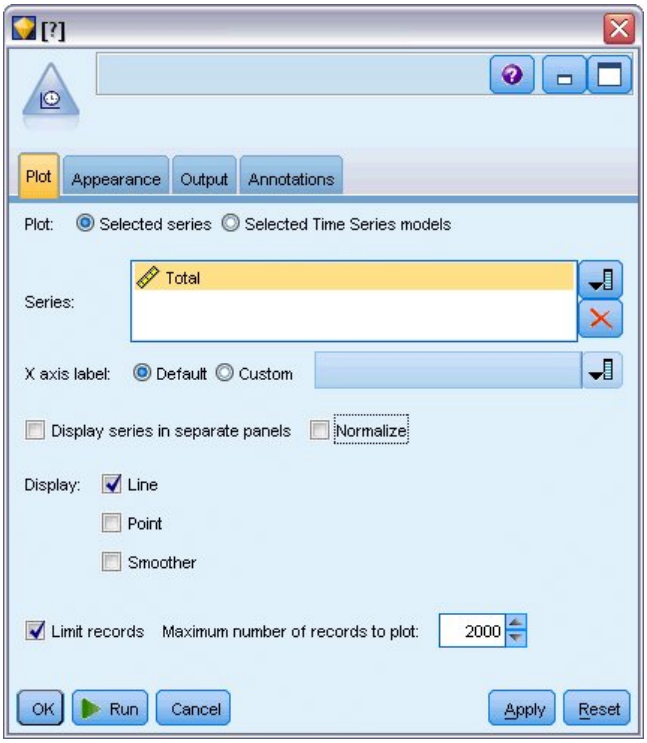


圖 *173.* 繪製訂閱者總數的圖形

1. 從「圖形」選用區，將「時間圖」節點連接至「過濾器」節點。
2. 將 *Total* 欄位新增至「序列」清單。
3. 取消選取**在個別畫面中顯示序列**和**正規化**勾選框。
4. 按一下「**執行**」。

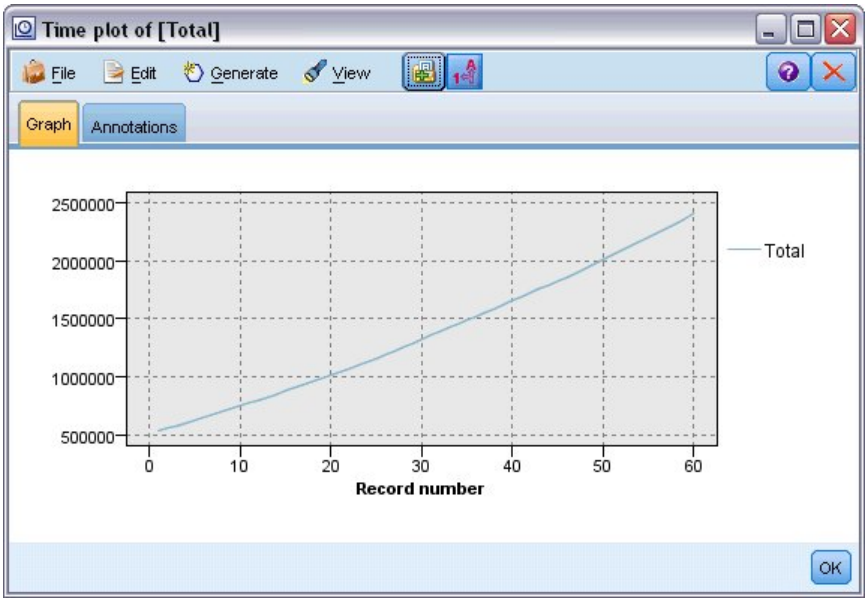


圖 *174. Total* 欄位的時間圖

數列展示了非常平滑的上升趨勢，且未顯示任何週期性變異。個別數列可能有週期性，但週期性一般而言似乎不是資料的顯著特色。

當然，您也應該在排除週期性模型前，檢閱每個數列。然後您可以分離出具備週期性的數列並分別建立他們的模式。

IBM SPSS Modeler 可讓您輕鬆一起繪製多個序列的圖形。

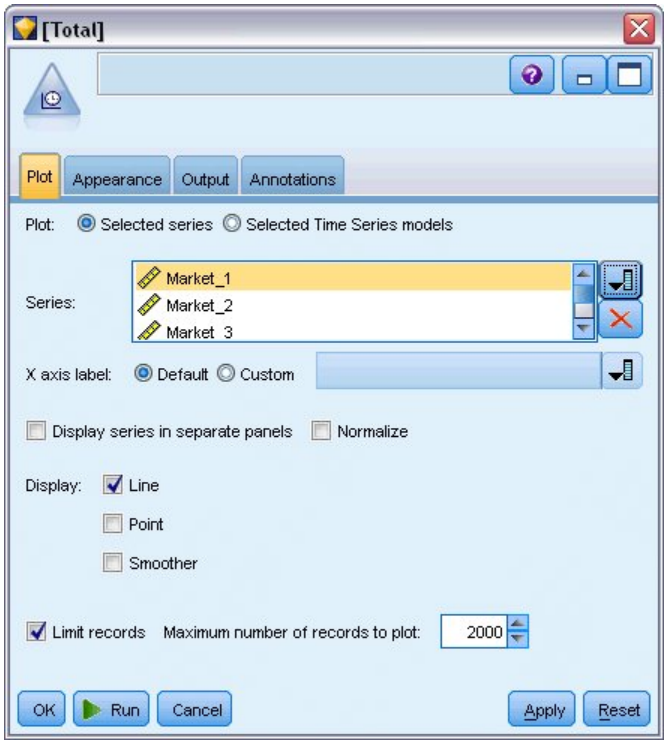


圖 *175.* 繪製多個時間序列的圖形

5. 重新開啟「時間圖」節點。
6. 將 *Total* 欄位從「序列」清單中移除（選取它並按一下紅色的 X 按鈕）。
7. 將 *Market_1* 到 *Market_5* 個欄位新增至清單。
8. 按一下「**執行**」。

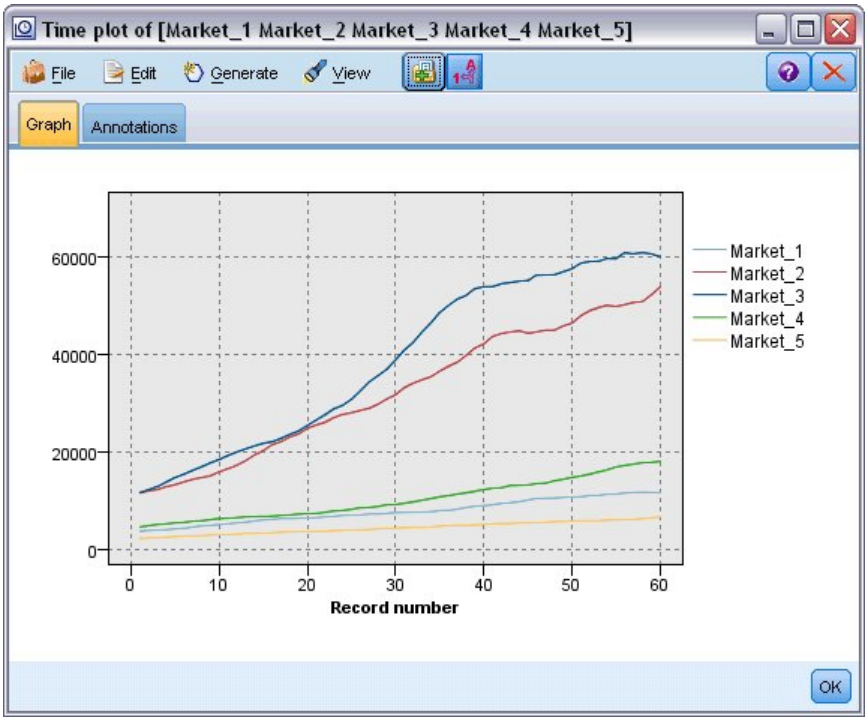


圖 *176.* 多個欄位的時間圖

檢驗每一個市場會顯示每個案例中的穩定上升趨勢。雖然部分市場較其他市場而言有點不穩定，但沒有證據表明有週期性跡象。

定義日期

現在，您需要將 *DATE_* 欄位的儲存體類型變更為日期格式。

1. 將「填充值」節點連接至「過濾器」節點。
2. 開啟「填充值」節點並按一下欄位選取器按鈕。
3. 選取 **DATE_** 以將它新增至**填寫欄位**。
4. 將**取代**條件設為**一律**。
5. 將**取代為**的值設為 **to_date(DATE_)**。

圖 *177.* 設定日期儲存體類型

變更預設日期格式以符合「日期」欄位的格式。這是讓「日期」欄位轉換如預期工作的必要動作。

6. 在功能表上，選擇**工具 > 串流內容 > 選項**以顯示「串流選項」對話框。
7. 選取**日期/時間**窗格，並將預設**日期格式**設為 **MON YYYY**。

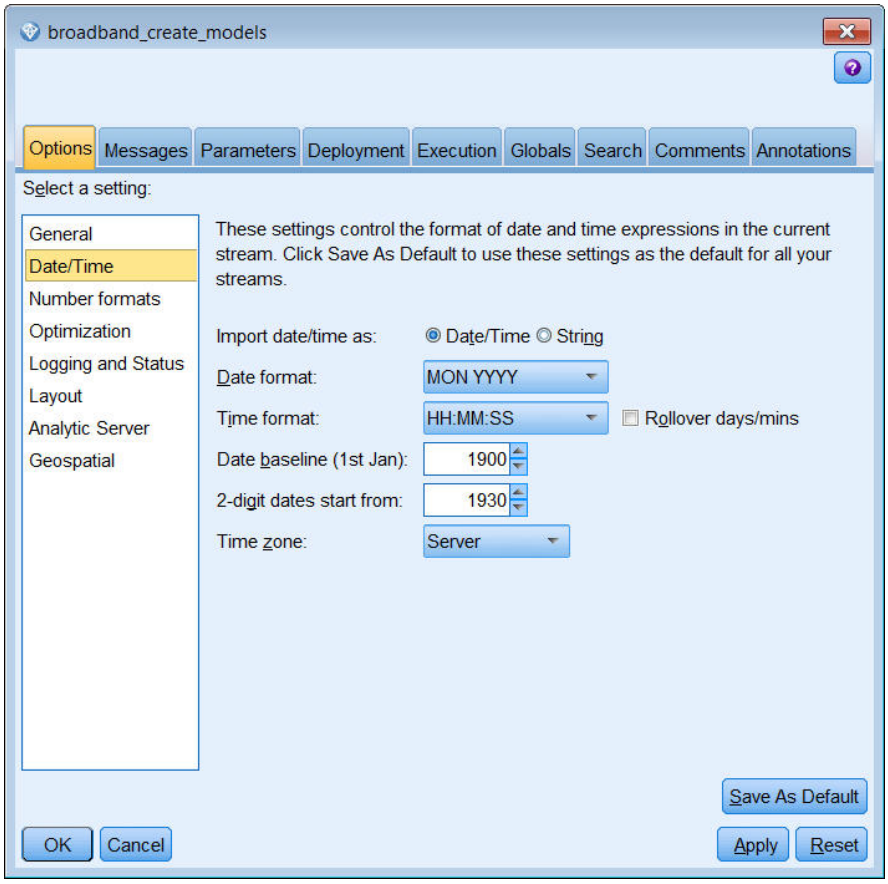


圖 *178.* 設定日期格式

定義目標

1. 新增「類型」節點並將欄位 *DATE_* 的角色設為**無**。將所有其他欄位（*Market_n* 欄位加上 *Total* 欄位）的角色設為**目標**。
2. 按一下**讀取值**按鈕以將值移入「值」直欄中。

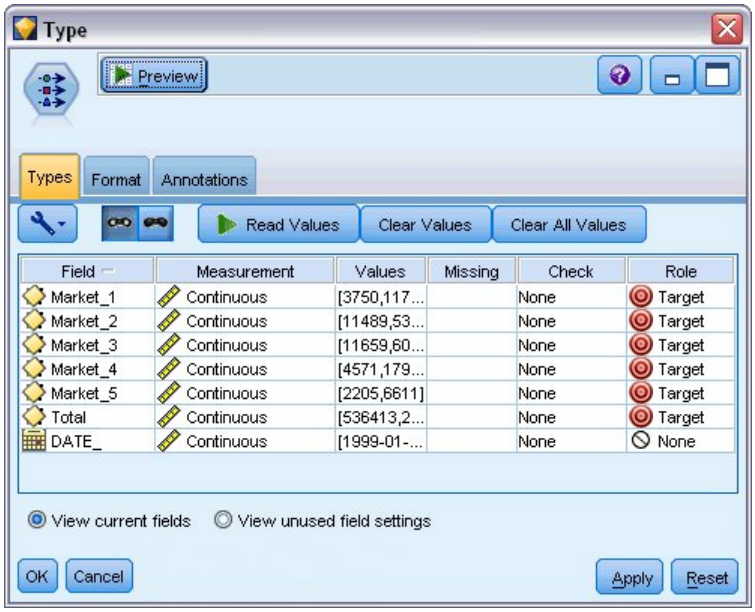


圖 *179.* 設定多個欄位的角色

設定時間間隔

1. 從「建模」選用區中，將「時間序列」節點新增至串流並將它連接至「類型」節點。
2. 在「資料規格」標籤的「觀察」窗格中，針對**日期/時間**欄位選取 DATE_。
3. 針對**時間間隔**選取 Months。

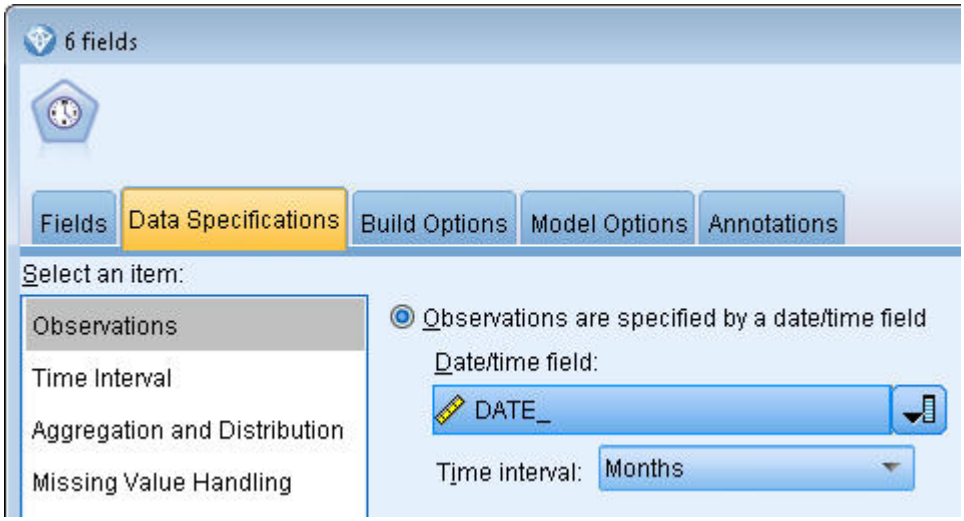


圖 *180.* 設定時間間隔

4. 在「模型選項」標籤上，選取**將記錄延伸至將來**勾選框。
5. 將值設為 **3**。

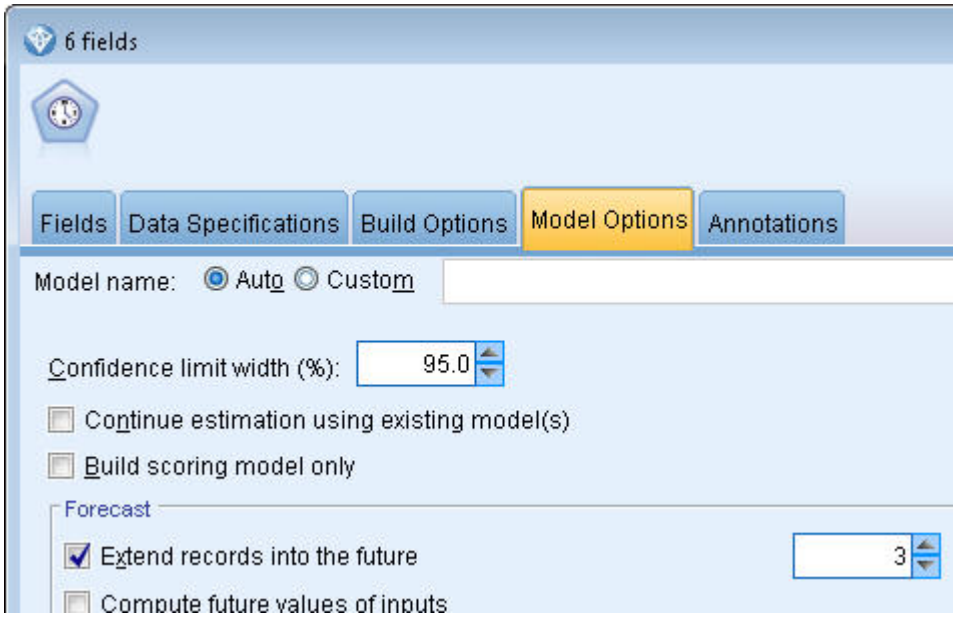


圖 *181.* 設定預測期間

建立模型

1. 在「時間序列」節點上，選擇「欄位」標籤。在**欄位**清單中，選取所有市場（5 個）並將其複製到**目標**和**候選輸入**清單中。此外，選取 Total 欄位並將其複製到**目標**清單。
2. 選擇「建置選項」標籤，然後在「一般」窗格上，確保已使用所有預設值選取了 Expert Modeler **方法**。這樣做可讓 Expert Modeler 決定每個時間序列最適合使用的模型。按一下「**執行**」。

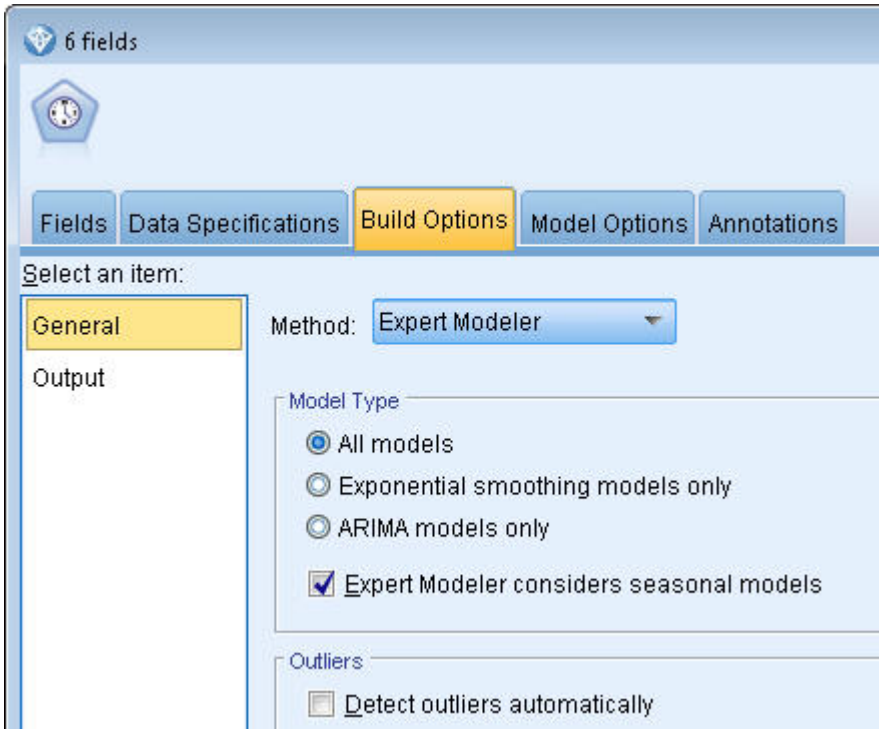


圖 *182.* 針對時間序列選擇 *Expert Modeler*

3. 將「時間序列」模型區塊連接至「時間序列」節點。
4. 將「表格」節點連接至「時間序列」模型區塊並按一下**執行**。

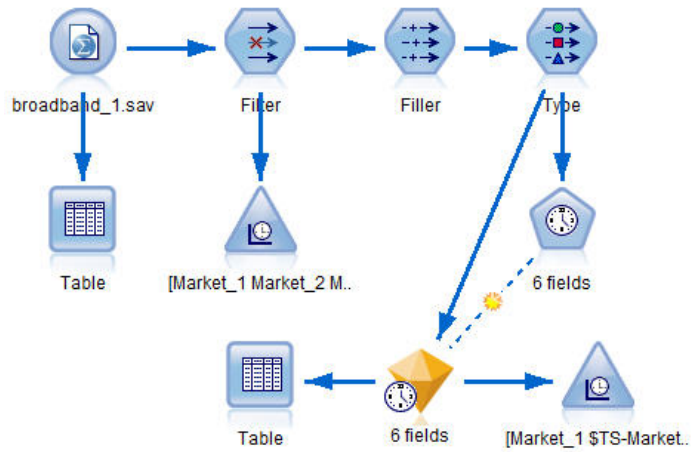


圖 *183.* 顯示時間序列建模的串流範例

現在，原始資料中附加了三個新列（61 到 63）。這些是預測期間的列，在本案例中為 2004 年 1 月到 3 月。

現在，還存在數個新直欄；*$TS-* 直欄由「時間序列」節點新增。對於每一列（亦即，時間序列資料中的每個時間間隔），這些直欄表示下列內容：

直欄	說明
$TS-*colname*	每欄原始資料產生的模型資料。

直欄	說明
$TSLCI-*colname*	每欄產生的模型資料的低信賴區間值。
$TSUCI-*colname*	每欄產生的模型資料的高信賴區間值。
$TS-Total	此列的 $TS-*colname* 值的總計。
$TSLCI-Total	此列的 $TSLCI-*colname* 值的總計。
$TSUCI-Total	此列的 $TSUCI-*colname* 值的總計。

預測作業的最重要直欄為直欄 *$TS-Market_n*、*$TSLCI-Market_n* 和 *$TSUCI-Market_n*。尤其是，這些直欄對應的列 61 到 63 包含每個本地市場的使用者訂閱預測資料和信賴區間。

檢查模型

1. 按兩下「時間序列」模型區塊，並選取「輸出」標籤以顯示為每個市場產生之模型的相關資料。

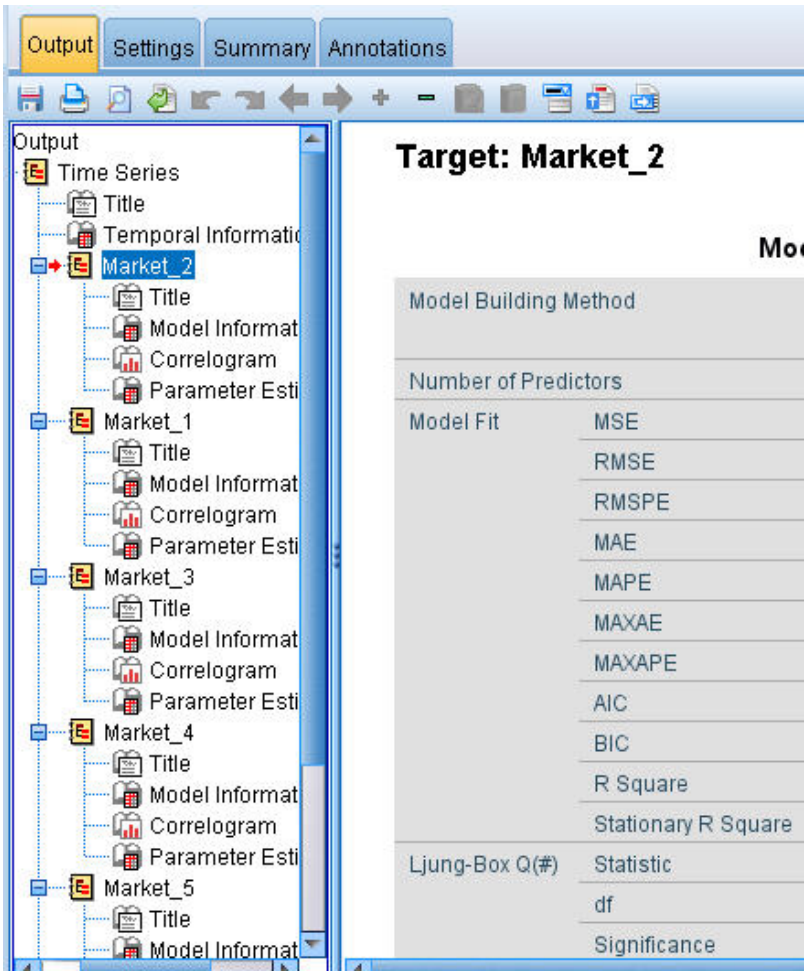


圖 *184.* 為市場產生的時間序列模型

在左側的「輸出」直欄中，選取任意市場的**模型資訊**。**預測工具數目**明細行顯示有多少欄位用作每個目標的預測工具；在本案例中為無。

模型資訊表格中剩餘的明細行顯示每個模型的各種適合度測量。**平穩 R 平方**值提供模型所說明之序列中變異數總計所佔比例的估計值。值越高（最大為 1.0），模型適合度越好。

Q(#) 統計量、**df** 和**顯著性**明細行與 Ljung-Box 統計量（檢定殘差在模型中的隨機性）相關；殘差越隨機，模型可能就越好。**Q(#)** 是 Ljung-Box 統計量本身，而 **df**（自由度）表示在估計特定目標時自由改變的模型參數數目。

顯著性明細行給定 Ljung-Box 統計量的顯著性值，另外指出是否正確指定了模型。小於 0.05 的顯著性值表示殘差不是隨機的，表示觀察到的序列中存在模型未納入的結構。

請同時納入**平穩 R 平方**和**顯著性**值，Expert Modeler 為 *Market_3* 和 *Market_4* 選擇的模型完全可以接受。*Market_1*、*Market_2* 和 *Market_5* 的**顯著性**值全部小於 0.05，表示可能需要對這些市場的更適模型進行一些試驗。

此顯示畫面顯示一些其他的適合度測量。**R 平方**值提供可由模型說明之時間序列中變異數總計的估計值。由於此統計量的最大值為 1.0，我們的模型在這方面很好。

RMSE 是均方根誤差，測量序列的實際值與模型預測值之間的差異，並以序列本身所用的單位來表示。由於這是對誤差的測量，我們希望此值盡可能地小。初看之下，*Market_2* 和 *Market_3* 的模型（根據我們到目前為止看到的統計量，這些模型仍然可以接受）相較於另外三個市場的模型成功性更低。

這些額外的適合度測量包括絕對百分比錯誤平均值 (**MAPE**) 及其上限值 (**MAXAPE**)。絕對百分比錯誤用來測量目標序列與其模型預測層級的差異程度，以百分比值表示。您可透過檢驗所有模式內的平均數值及最大數值，瞭解您預測值內的不確定性。

MAPE 值顯示所有模型都顯示 1% 左右的不確定性平均值，此值非常低。MAXAPE 值顯示絕對百分比錯誤上限，這對於想像您預測值的最糟情況非常有用。它顯示大部分模型的最大百分比錯誤位於大約 1.8 到 3.7% 的範圍內，這也是一組很低的數字，只有 *Market_4* 較高，接近 7%。

MAE（表示絕對錯誤）值顯示預測錯誤的絕對值的平均值。類似於 RMSE 值，它以序列本身所用的單位表示。**MAXAE** 以相同單位顯示最大預測錯誤，並指出預測值的最糟情況。

雖然這些絕對值很有趣，但是它是百分比錯誤（MAPE 和 MAXAPE）的值，這些值在本案例中更有用，因為目標序列代表不同規模的市場的訂閱者數目。

MAPE 和 MAXAPE 值代表可接受的模型不確定性數量嗎？它們一定會很低。在這種情況下商業意識開始起作用，因為可接受的風險將因問題不同而異。我們將假設適合度統計量在可接受的範圍內，然後繼續查看餘數錯誤。

相較於僅檢視適合度統計量，檢查模型殘差的自動相關性函數 (ACF) 和局部自動相關性函數 (PACF) 值可對模型提供更多的定量洞察。

良好指定的時間序列模型將會擷取所有非隨機變數，其中包括週期、趨勢以及重要的循環因素和其他因素。如果是這種情況，則所有錯誤在一段時間內都應該與它本身產生關聯（自動產生關聯）。任何一個自動相關性函數中的顯著性結構表示基礎模型不完整。

2. 針對第四個市場，在左欄中，按一下**相關圖**以顯示模型中殘差的自動相關性函數 (ACF) 和局部自動相關性函數 (PACF) 的值。

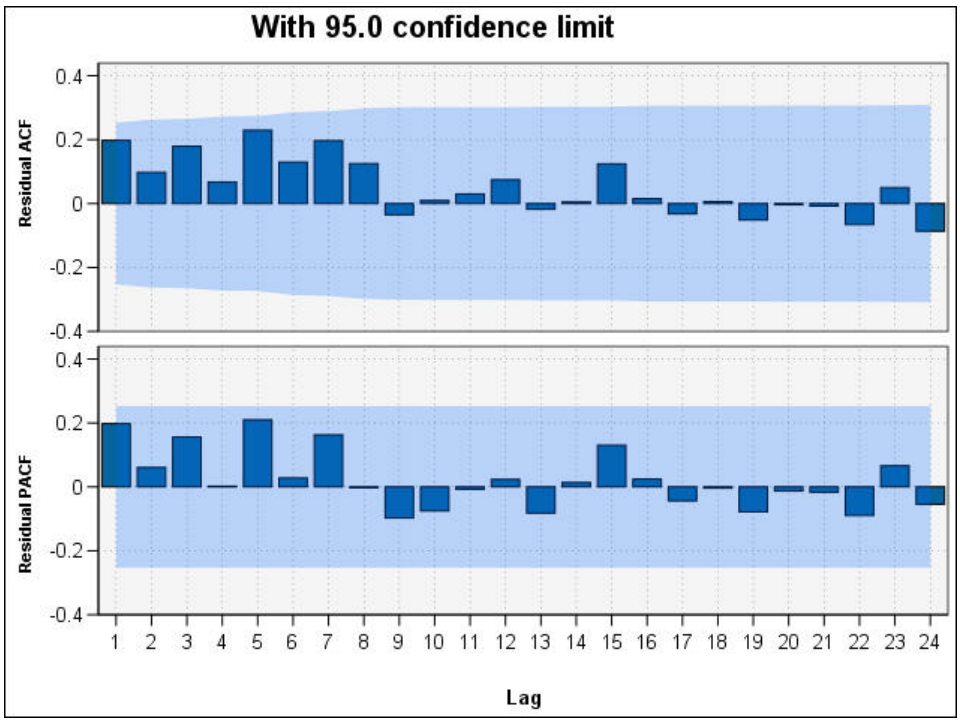


圖 *185.* 第四個市場的 *ACF* 和 *PACF* 值

在這些圖形中，錯誤變數的原始值已延遲高達 24 小時的時段，請與原始值比較以查看一段時間內是否有任何相關性。對於可接受的模型，在較上方 (ACF) 圖形中任何長條都不應該超出陰影區域（在正（上）或負（下）方向）。

如果發生此情況，您需要檢查較下方 (PACF) 圖形來查看其中的結構是否確認過。PACF 圖形在干預時間點控制序列值之後查看了相關性。

Market_4 的值全部位於陰影區域，因此我們可以繼續並檢查其他市場的值。

3. 按一下其他每個市場以及總計的**相關圖**。

   其他市場的值全部都顯示部分值超出陰影區域，確認我們之前從其**顯著性**值所懷疑的內容。我們需要在某個時間點對那些市場試驗一些不同的模型，以瞭解是否能夠得到更適合的模型，但對於本範例中的剩餘項目，我們將集中在能夠從 *Market_4* 模型學習的內容。
4. 從「圖形」選用區，將「時間圖」節點連接至「時間序列」模型區塊。
5. 在「圖形」標籤上，清除**在個別的畫面中顯示序列**勾選框。
6. 在**序列**清單中，按一下欄位選取器按鈕，選取 *Market_4* 和 *$TS-Market_4* 欄位，然後按一下**確定**以將其新增至清單中。
7. 按一下**執行**以顯示第一個本地市場之實際與預測資料的行式圖形。

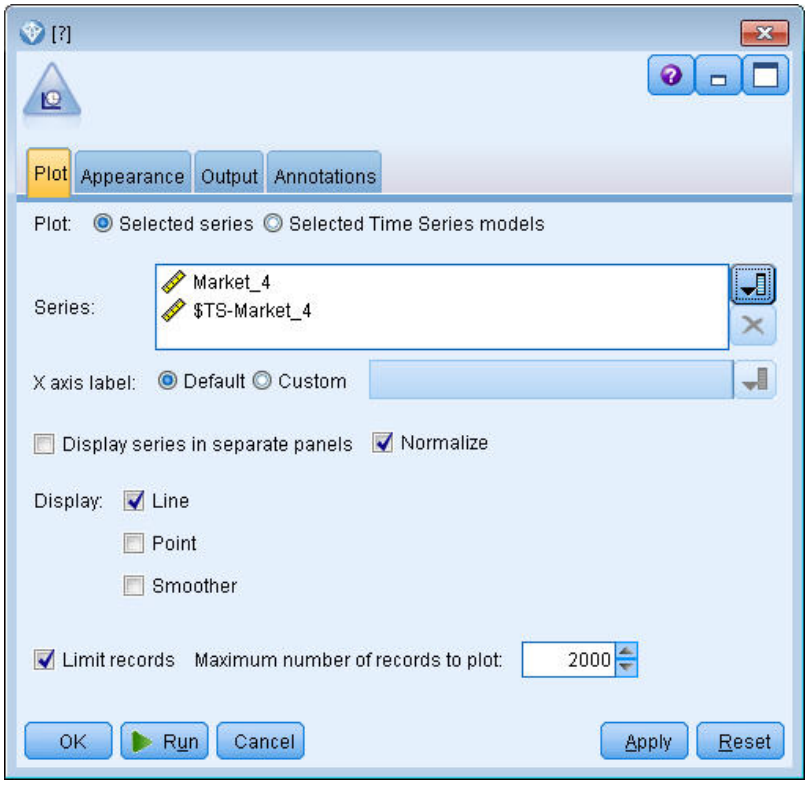


圖 *186.* 選取要繪圖的欄位

請注意預測 (*$TS-Market_4*) 明細行超出實際資料結尾的程度。您現在可以預測在此市場中接下來三個月的預期需求。

在整個時間序列期間實際資料與預測資料的行在圖形上緊接在一起，表示這是此特定時間序列的可靠模型。

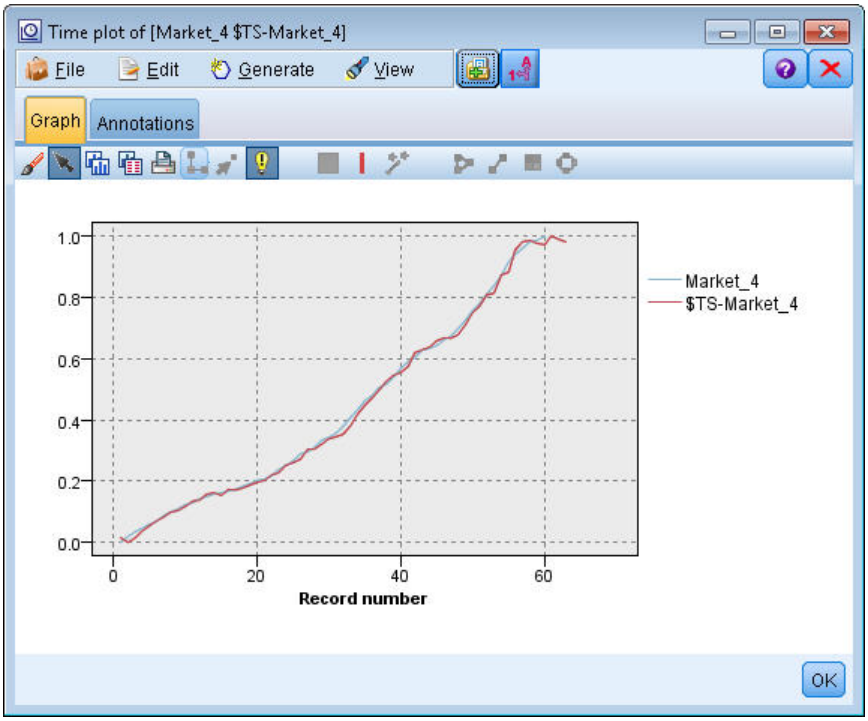


圖 *187. Market_4* 的實際與預測資料的時間圖

在檔案中儲存模型以用於未來的範例：

8. 按一下**確定**以關閉目前的圖形。
9. 開啟「時間序列」模型區塊。
10. 選擇**檔案 > 儲存節點**並指定檔案位置。
11. 按一下「**儲存**」。

針對這個特定市場，您具有一個可靠的模型，但預測有哪些誤差範圍呢？您可以透過檢查信賴區間獲得此情況的指示。

12. 按兩下串流中的最後一個「時間圖」節點（標示了 **Market_4 $TS-Market_4** 的節點）以再次開啟其對話框。
13. 按一下欄位選取器按鈕並將 *$TSLCI-Market_4* 和 *$TSUCI-Market_4* 欄位新增至**序列**清單。
14. 按一下「**執行**」。

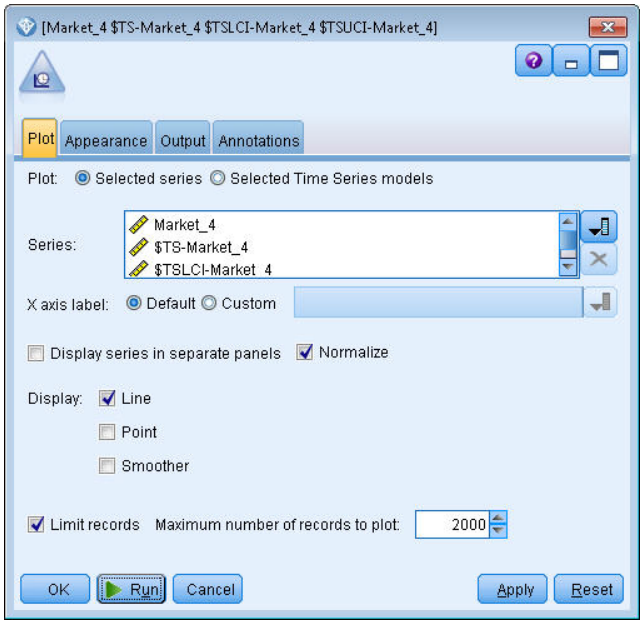


圖 *188.* 將更多欄位新增至圖形

現在，您擁有與先前相同的圖形，但新增了信賴區間的上限 (*$TSUCI*) 和下限 (*$TSLCI*)。

請注意在預測期間信賴區間的界限偏離程度，這表示隨著您對未來的進一步預測，不確定性也會增加。

但是，隨著每個時段的過去，您將會有其他（在本案例中）月份的實際使用資料，您可以據以進行您的預測。您可以將新資料讀取到串流中並重新套用您認為可靠的模型。請參閱第 162 頁的『重新套用時間序列模型』主題，以取得更多資訊。

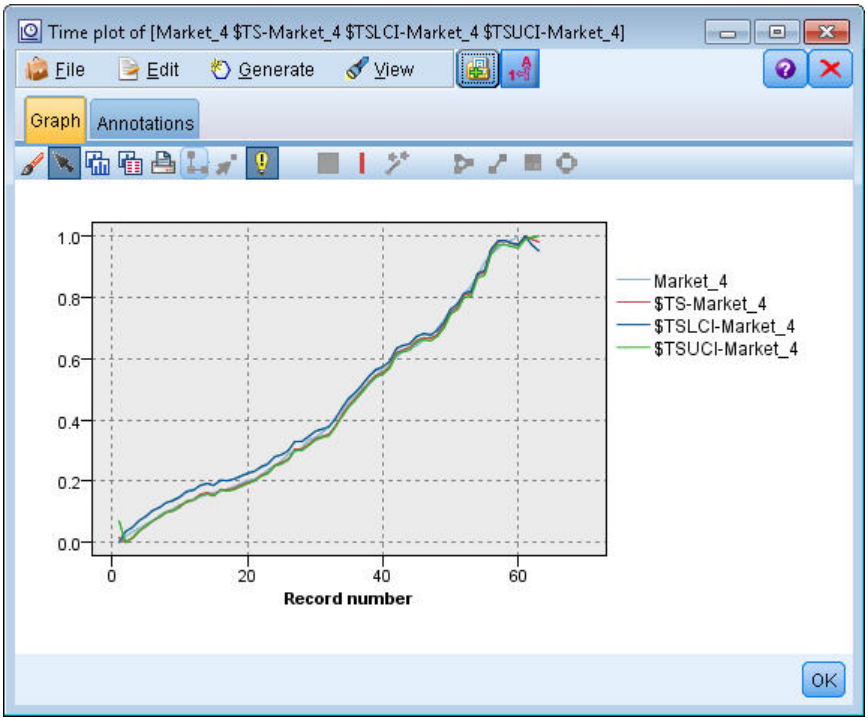


圖 *189.* 新增了信賴區間的時間圖

摘要

您已學習如何使用 Expert Modeler 來產生多個時間序列的預測，並且已將產生的模型儲存至外部檔案。

在下一個範例中，您將會看到如何將非標準時間序列資料輪換成適合輸入「時間序列」節點的格式。

重新套用時間序列模型

此範例套用第一個時間序列範例中的時間序列模型，但也可以獨立使用。請參閱第 145 頁的『使用時間序列節點來預測』主題，以取得更多資訊。

如原始實務範例所示，全國性寬頻服務提供業者的分析師必須針對數個本地市場的每一個市場產生每月使用者訂閱預測，以預測寬頻需求。您已經使用了 Expert Modeler 來建立模型及未來三個月的預測。

現在，已使用原始預測期間的實際資料更新您的資料倉儲，因此您可以使用該資料將預測水平線再延伸三個月。

此範例使用參照資料檔 *broadband_2.sav* 的串流 *broadband_apply_models.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*broadband_apply_models.str* 檔位於 *streams* 資料夾。

擷取串流

在此範例中，您將根據儲存在第一個範例中的「時間序列」模型重建「時間序列」節點。如果您沒有已儲存的模型，請勿擔心，我們已在 *Demos* 資料夾中提供了一個。

1. 開啟 *Demos* 之下 *streams* 資料夾中的 *broadband_apply_models.str*。

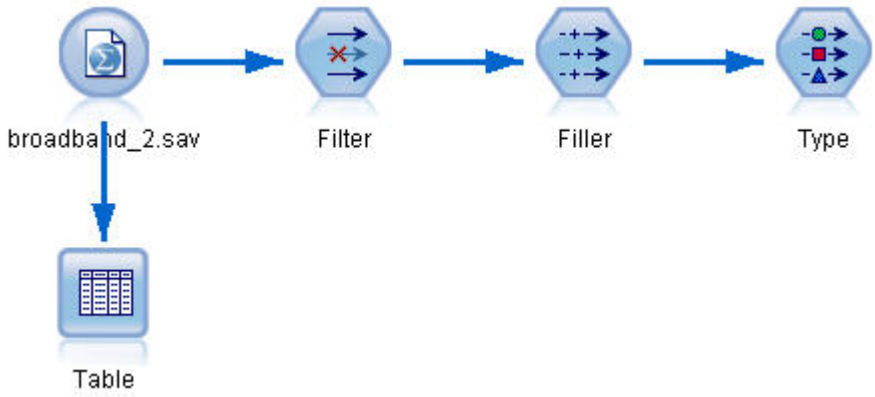
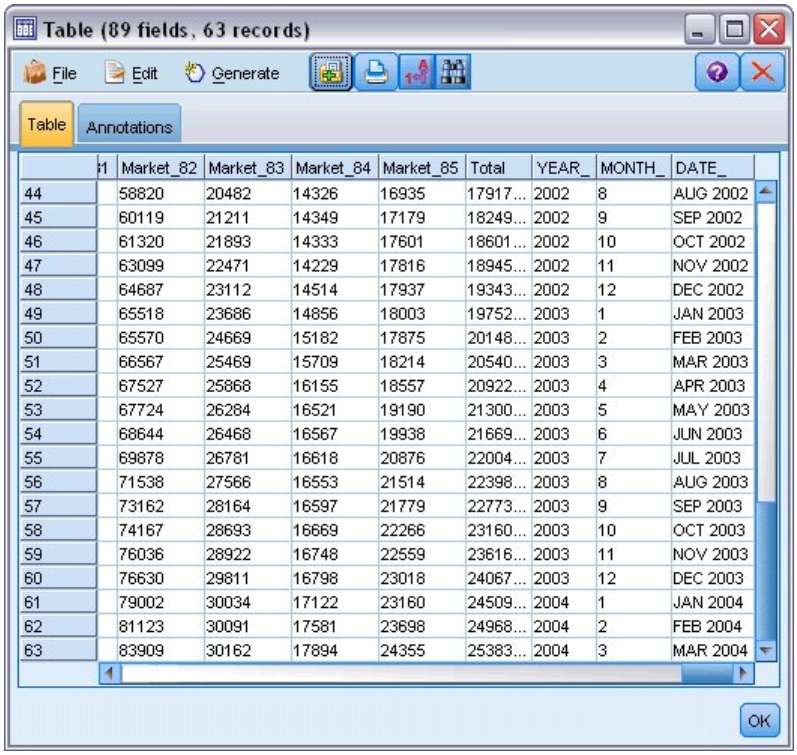


圖 *190.* 開啟串流

每月更新資料收集在 *broadband_2.sav* 中。

2. 將「表格」節點連接至 IBM SPSS Statistics 的「檔案」來源節點，開啟「表格」節點並按一下**執行**。

註：已使用 2004 年 1 月到 3 月的實際銷售資料（列 61 到 63 中的資料）更新資料檔案。



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

圖 *191.* 已更新的銷售資料

擷取儲存的模型

1. 在 IBM SPSS Modeler 功能表上，選擇**插入** > **檔案中的節點**並從 *Demos* 資料夾選取 *TSmodel.nod* 檔（或使用您儲存在第一個時間序列範例中的「時間序列」模型）。

此檔案包含前一個範例中的時間序列模型。插入作業會將對應的「時間序列」模型區塊放置在畫布上。

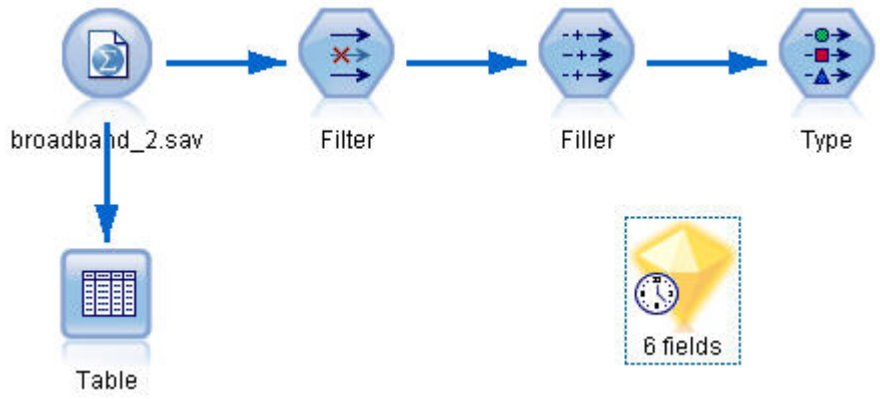


圖 *192.* 新增模型區塊

產生建模節點

1. 開啟「時間序列」模型區塊並選擇**產生** > **產生建模節點**。

這樣做會將「時間序列」建模節點放置在畫布上。

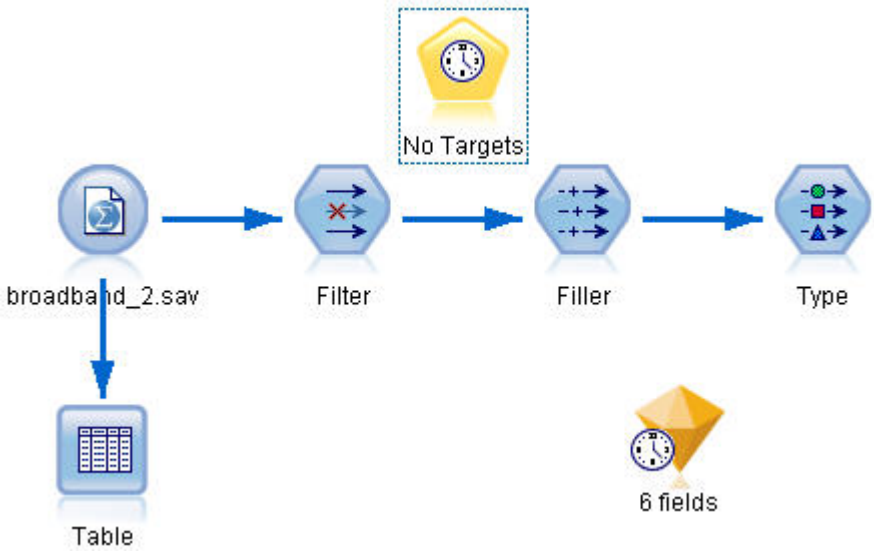


圖 *193.* 從模型區塊產生建模節點

產生新模型

1. 關閉「時間序列」模型區塊並從畫布中刪除它。

   舊模型是根據 60 列資料建置的。您需要根據更新的銷售資料（63 列）來產生新模型。
2. 將新產生的「時間序列」建置節點連接到串流。

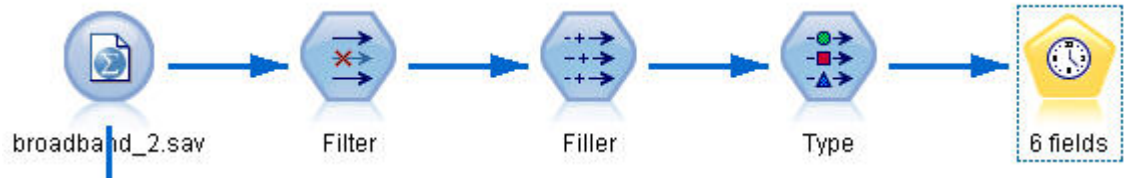


圖 *194.* 將建模節點連接到串流

3. 開啟「時間序列」節點。
4. 在**模型選項**標籤上，確保已勾選**繼續使用現有模型預估**。

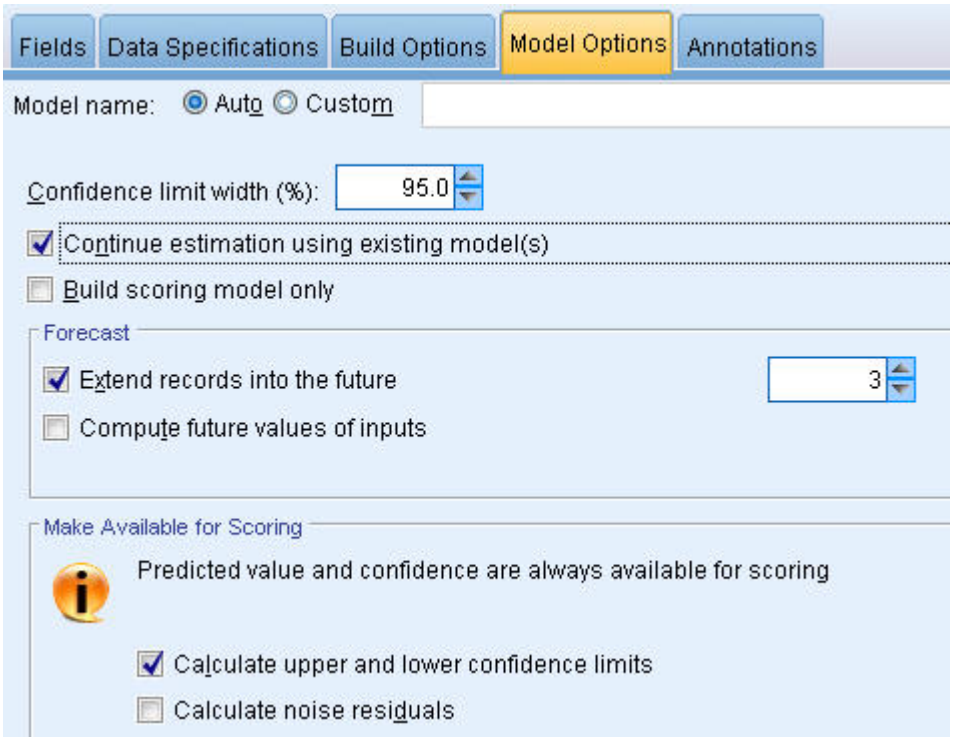


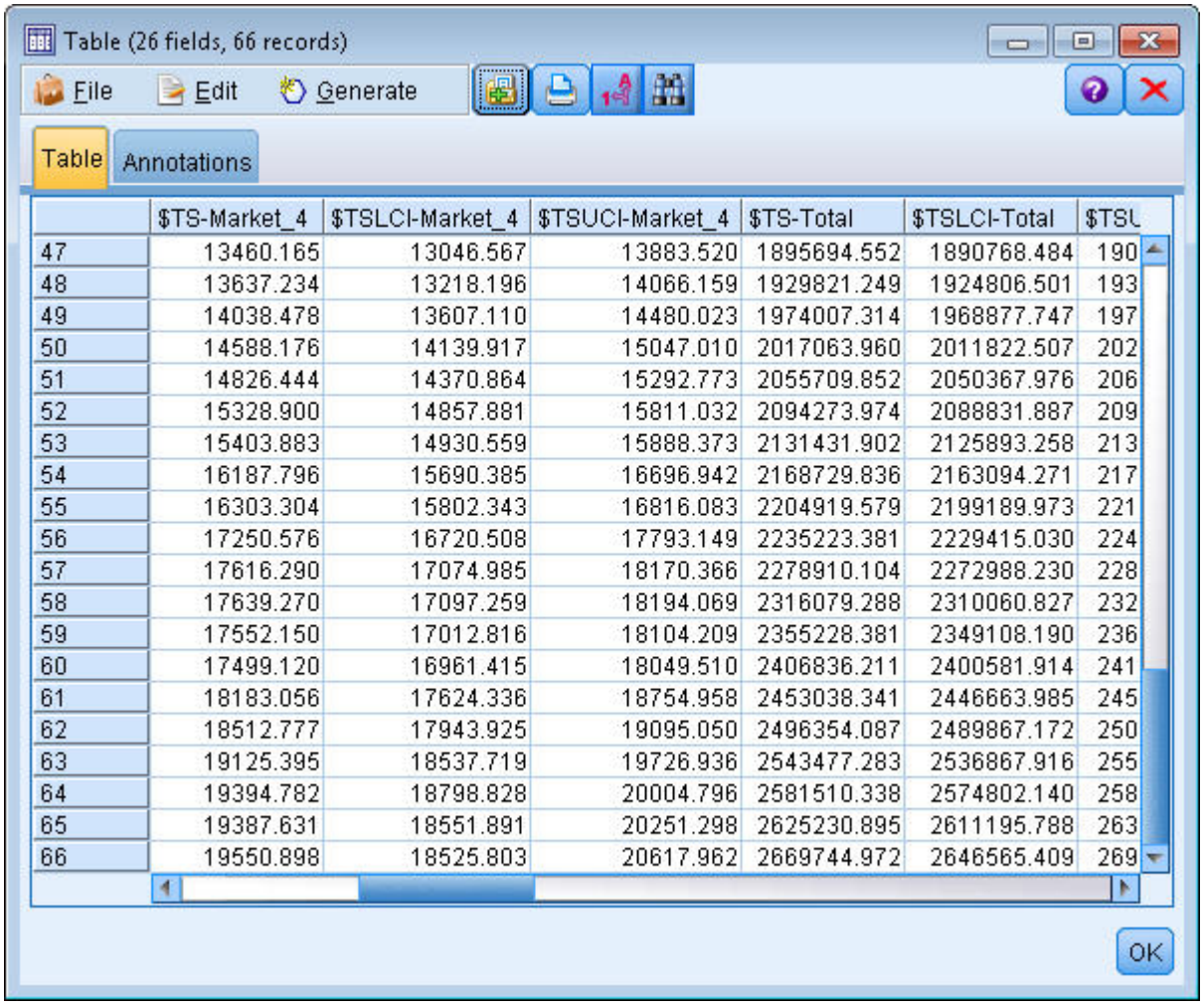
圖 *195.* 重複使用時間序列模型的儲存設定

5. 確保**將記錄延伸到未來**設為 **3**。
6. 按一下**執行**以將新的模型區塊放置在畫布和「模型」選用區中。

檢查新模型

1. 將「表格」節點連接至畫布上的新「時間序列」模型區塊。
2. 開啟「表格」節點並按一下**執行**。

新模型仍然會提前三個月預測，因為您正在重複使用儲存的設定。但是，這次它會預測四月份到六月份（第 64 行到 66 行），因為預估期間現在是在三月份而不是一月份結束。



	$TS-Market_4	$TSLCI-Market_4	$TSUCI-Market_4	$TS-Total	$TSLCI-Total	$TSL
47	13460.165	13046.567	13883.520	1895694.552	1890768.484	190
48	13637.234	13218.196	14066.159	1929821.249	1924806.501	193
49	14038.478	13607.110	14480.023	1974007.314	1968877.747	197
50	14588.176	14139.917	15047.010	2017063.960	2011822.507	202
51	14826.444	14370.864	15292.773	2055709.852	2050367.976	206
52	15328.900	14857.881	15811.032	2094273.974	2088831.887	209
53	15403.883	14930.559	15888.373	2131431.902	2125893.258	213
54	16187.796	15690.385	16696.942	2168729.836	2163094.271	217
55	16303.304	15802.343	16816.083	2204919.579	2199189.973	221
56	17250.576	16720.508	17793.149	2235223.381	2229415.030	224
57	17616.290	17074.985	18170.366	2278910.104	2272988.230	228
58	17639.270	17097.259	18194.069	2316079.288	2310060.827	232
59	17552.150	17012.816	18104.209	2355228.381	2349108.190	236
60	17499.120	16961.415	18049.510	2406836.211	2400581.914	241
61	18183.056	17624.336	18754.958	2453038.341	2446663.985	245
62	18512.777	17943.925	19095.050	2496354.087	2489867.172	250
63	19125.395	18537.719	19726.936	2543477.283	2536867.916	255
64	19394.782	18798.828	20004.796	2581510.338	2574802.140	258
65	19387.631	18551.891	20251.298	2625230.895	2611195.788	263
66	19550.898	18525.803	20617.962	2669744.972	2646565.409	269

圖 *196.* 顯示新預測的表格

3. 將「時間圖」圖形節點連接至「時間序列」模型區塊。

   這次我們將使用特別針對時間序列模型設計的時間圖顯示。

4. 在「圖形」標籤上，將 **X 軸標籤**設為**自訂**，然後選取 Date_。
5. 針對**圖形**選擇**選取的時間序列模型**選項。
6. 從**序列**清單中，按一下欄位選取器按鈕，選取 $TS-Market_4 欄位，然後按一下**確定**以將其新增至清單中。

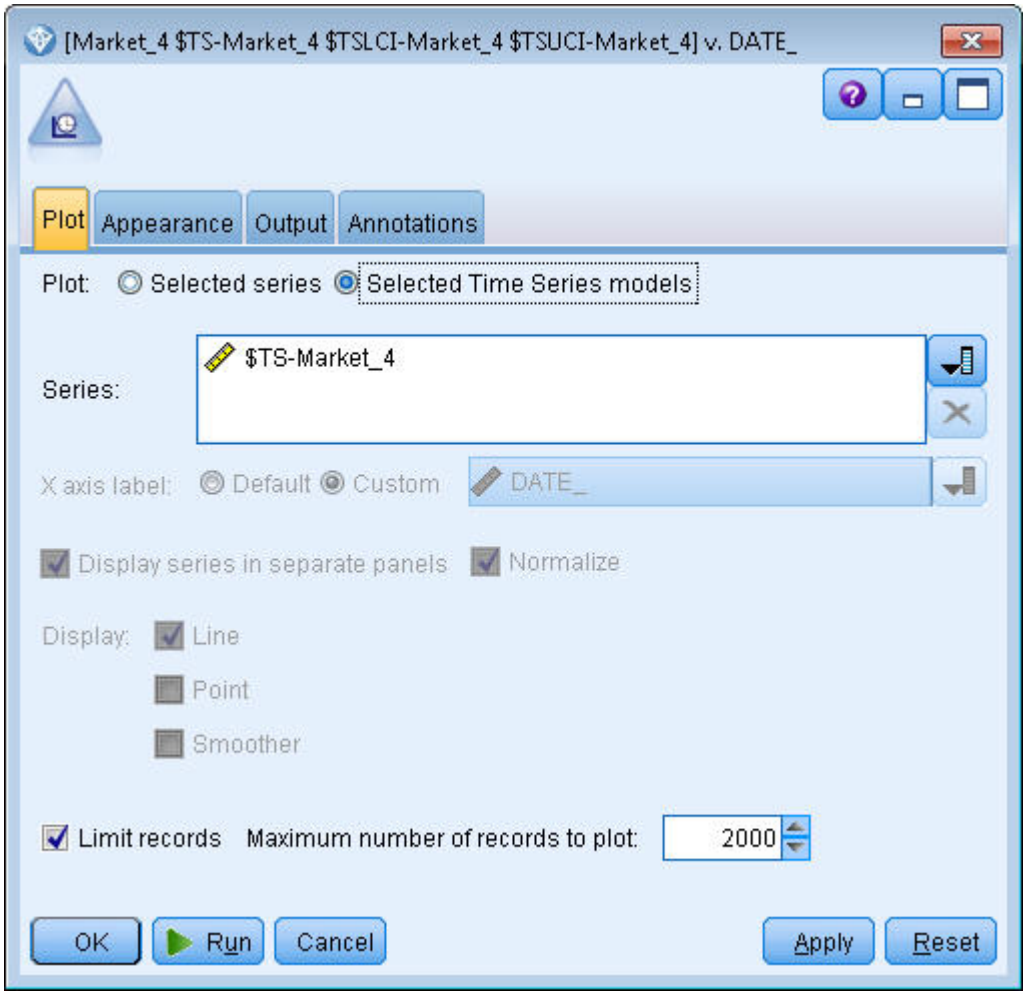


圖 *197.* 指定要繪製的欄位

7. 按一下「**執行**」。

現在，您有一個圖形顯示 Market_4 的實際銷售量（最高到 2004 年 3 月），以及預測銷售量和信賴區間（由藍色陰影區域表示）（最高到 2004 年 6 月）。

如第一個範例所示，在整個時段內預測值緊接在實際資料之後，又一次表示您擁有良好模型。

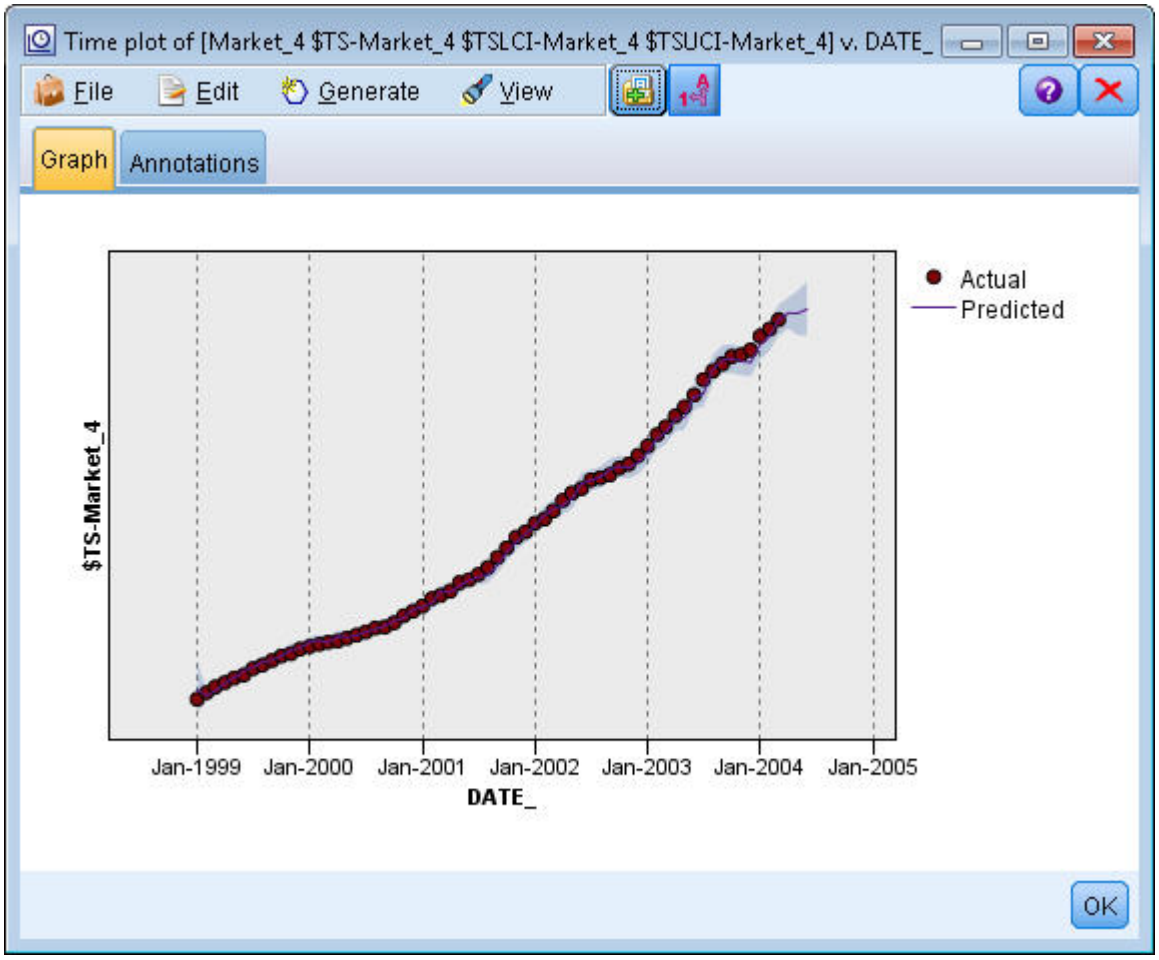


圖 *198.* 預測延伸到六月

摘要

您已瞭解在更多現行資料變成可用時如何套用儲存的模型以延伸您先前的預測，並且您已經在不重建模型的情況下執行此動作。當然，如果有理由認為模型已變更，則您應該重建模型。

第 15 章 預測型錄銷售量（時間序列）

型錄公司對於根據其在過去 10 年的銷售資料來預測其男士服裝線每月銷售量很感興趣。

此範例使用參照資料檔 *catalog_seasfac.sav* 的串流 *catalog_forecast.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*catalog_forecast.str* 檔位於 *streams* 目錄。

我們在較早的範例中已查看您如何讓 Expert Modeler 決定哪些模型最適合用於您的時間序列。現在，是時候仔細研究一下在您自行選擇模型時可用的兩種方法--指數平滑和 ARIMA。

若要協助您決定適當的模型，最好是先繪製時間序列的圖形。時間序列的視覺化檢驗通常會是協助您進行選擇的強大指引。特別是，您需要詢問您自己：

- 序列有整體趨勢嗎？如果有，趨勢是顯示不變還是隨時間推移消失？
- 序列顯示週期性嗎？如果是，則週期性變動是隨著時間而增長還是在連續時段內保持不變？

建立串流

1. 建立新串流並新增指向 *catalog_seasfac.sav* 的「統計量檔案」來源節點。

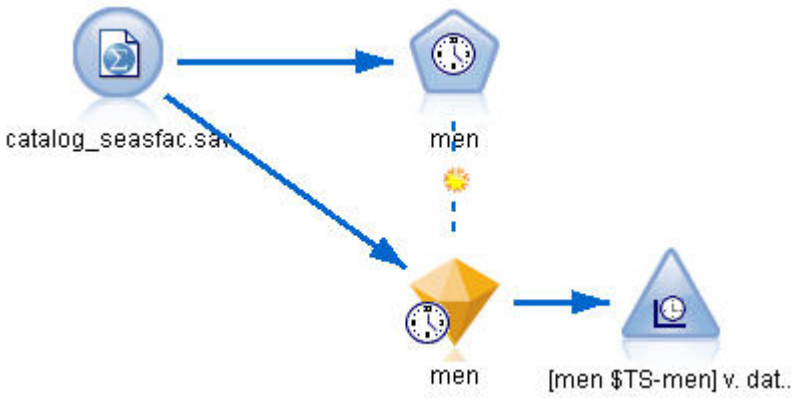


圖 *199.* 預測型錄銷售量

2. 開啟 IBM SPSS Statistics 的「檔案」來源節點並選取「類型」標籤。
3. 按一下**讀取值**，然後按一下**確定**。
4. 按一下 `men` 欄位的**角色**直欄，然後將角色設為**目標**。

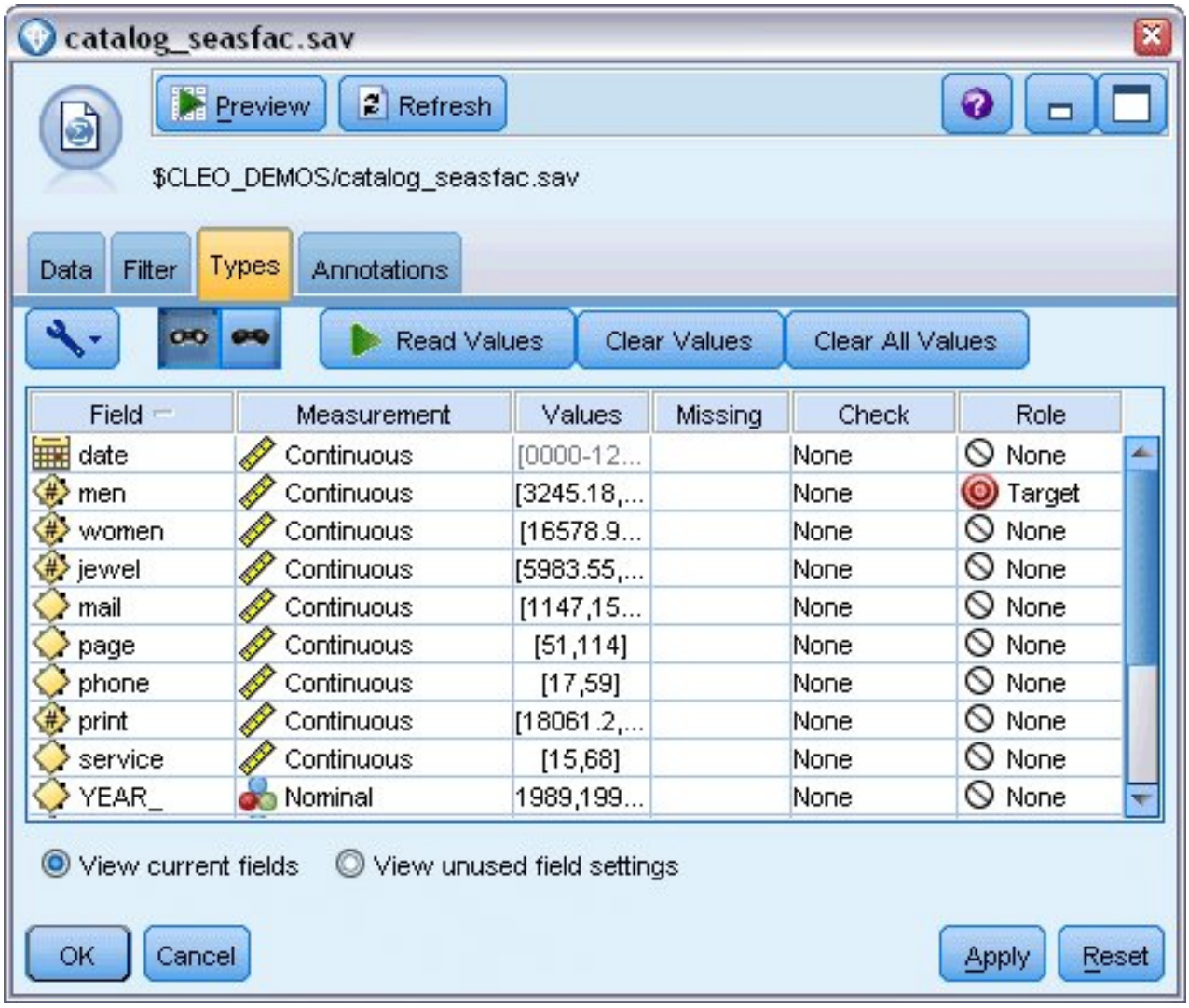


圖 *200.* 指定目標欄位

5. 將所有其他欄位的角色設為**無**，然後按一下**確定**。
6. 將「時間圖」圖形節點連接至 IBM SPSS Statistics 的「檔案」來源節點。
7. 開啟「時間圖」節點，然後在「圖形」標籤上，將 men 新增至**序列**清單。
8. 將 **X 軸標籤**設為**自訂**，然後選取 date。
9. 清除**正規化**勾選框。

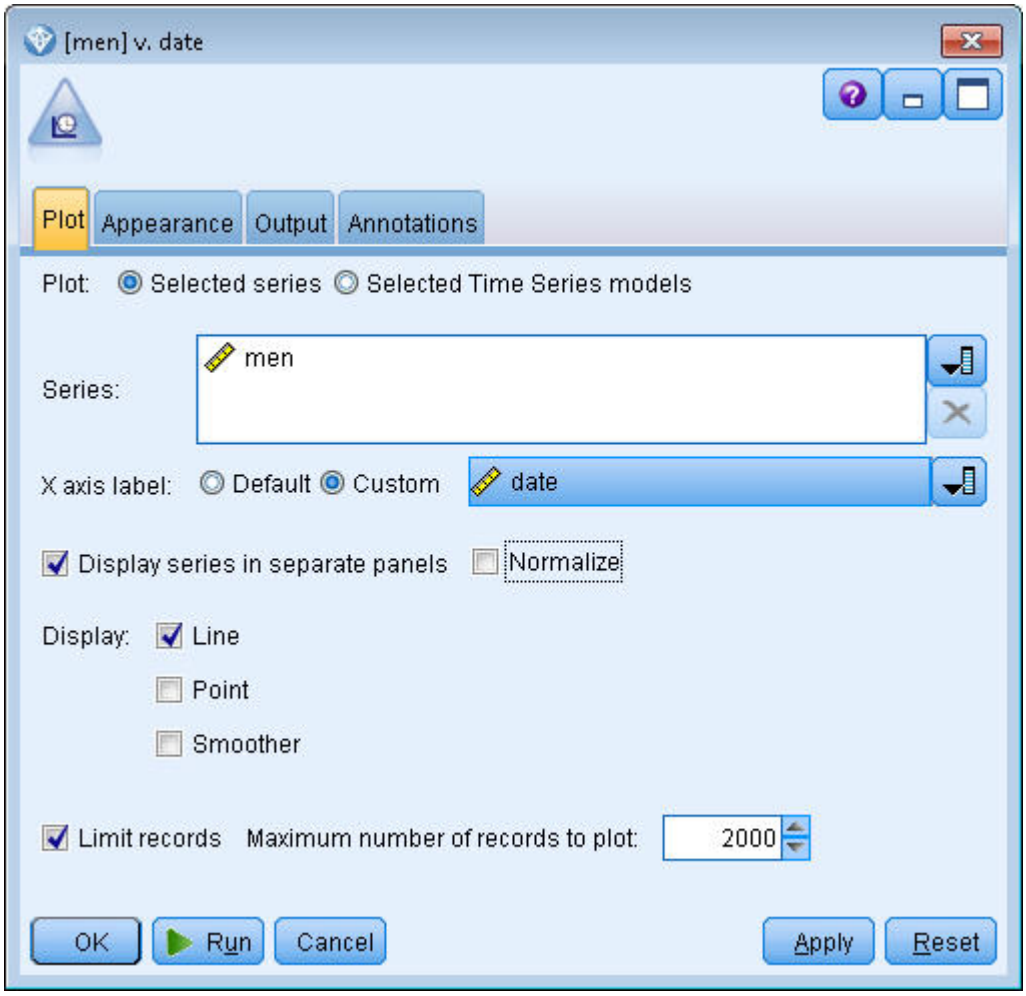


圖 *201.* 繪製時間序列的圖形

10. 按一下「**執行**」。

檢驗資料

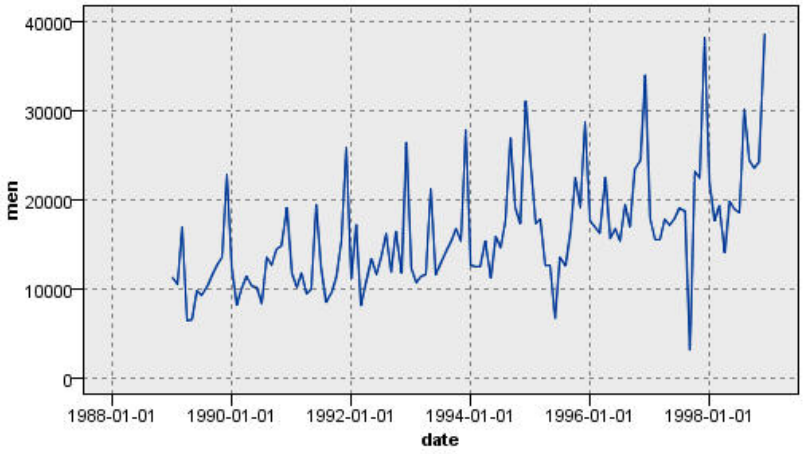


圖 *202.* 男士服裝的實際銷售量

該序列顯示一般上升趨勢；亦即，序列值在一段時間內傾向增長。上升趨勢似乎是不變的，表示這是線性趨勢。

該序列還有一個不同的周期性型樣，其年度高銷售量在 12 月，如圖形上的垂直線所指示。周期性變化似乎隨著上升序列趨勢而增長，表示這是相乘周期性而非加法周期性。

1. 按一下**確定**以關閉該圖形。

現在，您已識別序列的性質，您已準備好嘗試對其建模。指數平滑方法對于預測展示趨勢及/或周期性的序列很有用。如我們所見，您的資料會展示兩個性質。

指數平滑化

建置一個最適指數平滑模型涉及確定模型類型（模型是否需要包括趨勢及/或週期），然後取得所選模型的最適參數。

在一段時間內的男士服裝銷售量圖所建議的模型，包含一個線性趨勢成份和一個相乘性週期成份。這表示 Winters 模型。但是，首先我們將探索簡式模型（沒有趨勢和週期），然後再探索 Holt 模型（納入了線性趨勢，但沒有週期）。這將為您提供練習來識別模型不太適合資料的情況，並提供順利建模所需的技能。

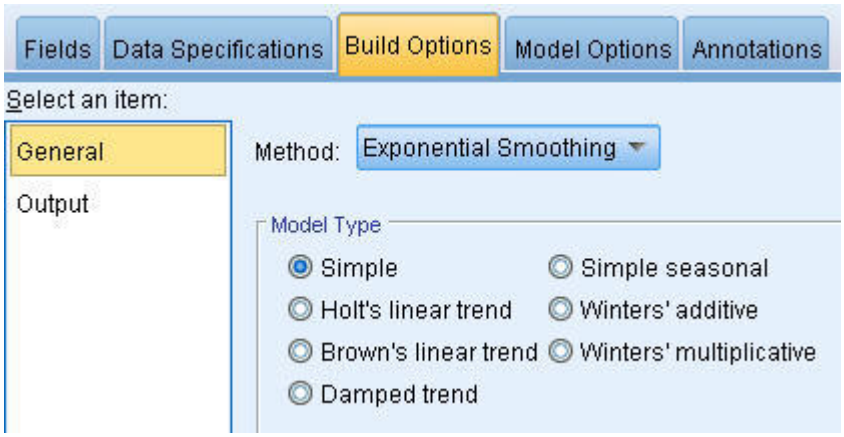


圖 *203.* 指定指數平滑

我們將從簡式指數平滑模型開始。

1. 將「時間序列」節點新增至串流並將其連接至來源節點。
2. 在「資料規格」標籤的「觀察」窗格中，針對**日期/時間**欄位選取 `date`。
3. 針對**時間間隔**選取 `Months`。

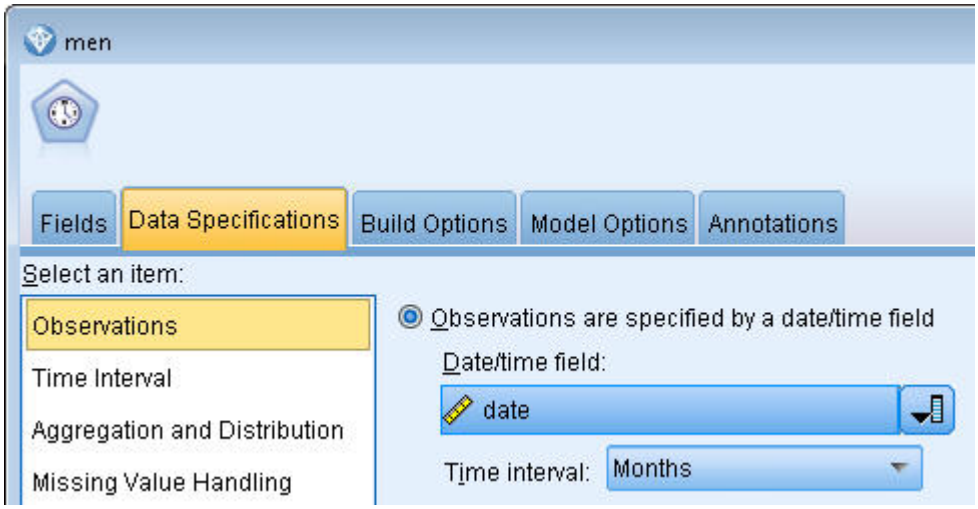


圖 *204.* 設定時間間隔

4. 在「建置選項」標籤的「一般」窗格中，將**方法**設為**指數平滑**。
5. 將**模型類型**設為**簡式**。

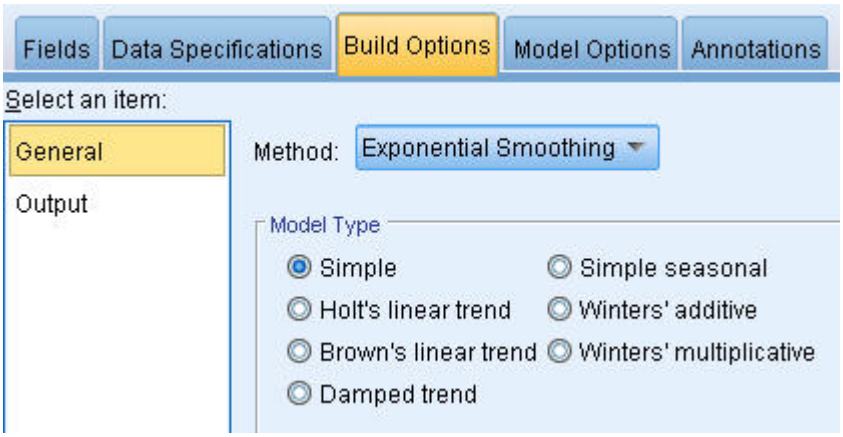


圖 *205.* 設定模型建置方法

6. 按一下**執行**以建立模型區塊。

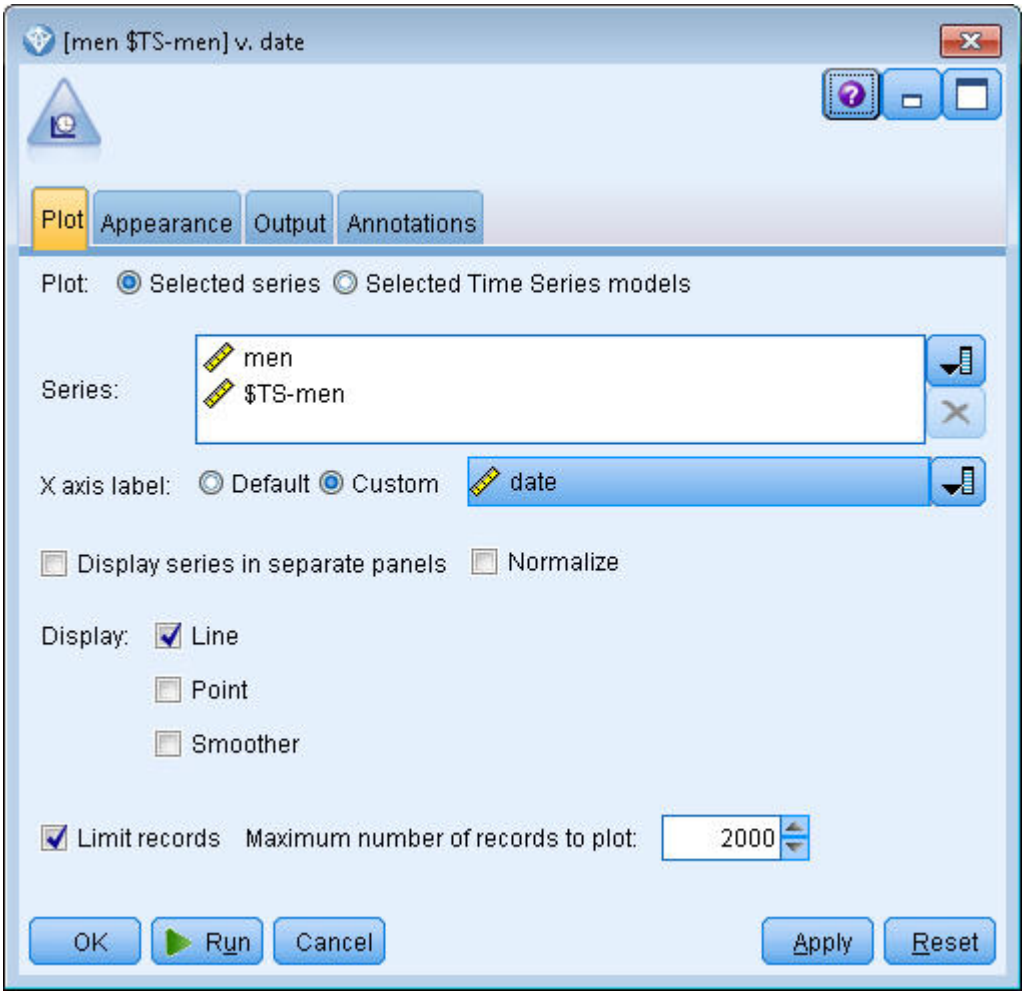


圖 *206.* 繪製時間序列模型的圖形

7. 將「時間圖」節點連接至模型區塊。
8. 在**圖形**標籤上，將 men and $TS-men 新增至**序列**清單。

9. 將 **X 軸標籤**設為**自訂**，然後選取 date。
10. 清除**在個別畫面中顯示序列**和**正規化**勾選框。
11. 按一下「**執行**」。

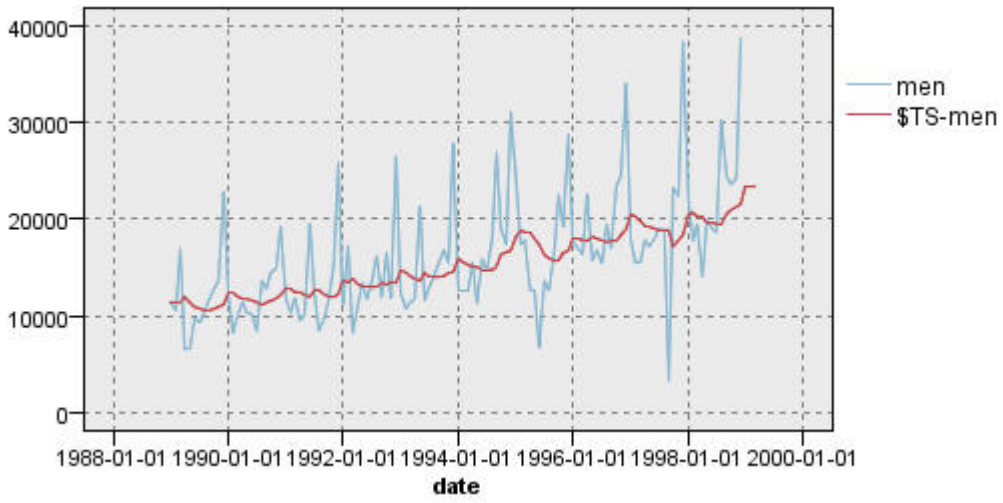


圖 *207.* 簡式指數平滑模型

men 圖代表實際資料，而 **$TS-men** 表示時間序列模型。

雖然簡式模型實際上展示逐漸（而非遲緩）上升趨勢，但它未考量週期。您可以安全地拒絕此模型。

12. 按一下**確定**以關閉時間圖視窗。

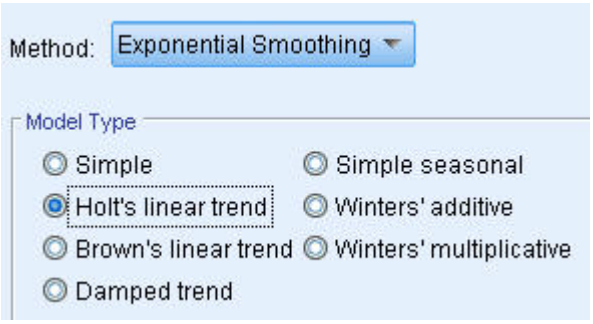


圖 *208.* 選取 *Holt* 模型

嘗試使用 Holt 線性模型。此模型在對趨勢建模方面至少要好於簡式模型，雖然它也不太可能擷取週期。

13. 重新開啟「時間序列」節點。
14. 在「建置選項」標籤上的「一般」窗格中，仍然選取**指數平滑**作為**方法**，選取 **Holt 線性趨勢**作為**模型類型**。
15. 按一下**執行**以重建模型區塊。
16. 重新開啟「時間圖」節點，然後按一下**執行**。

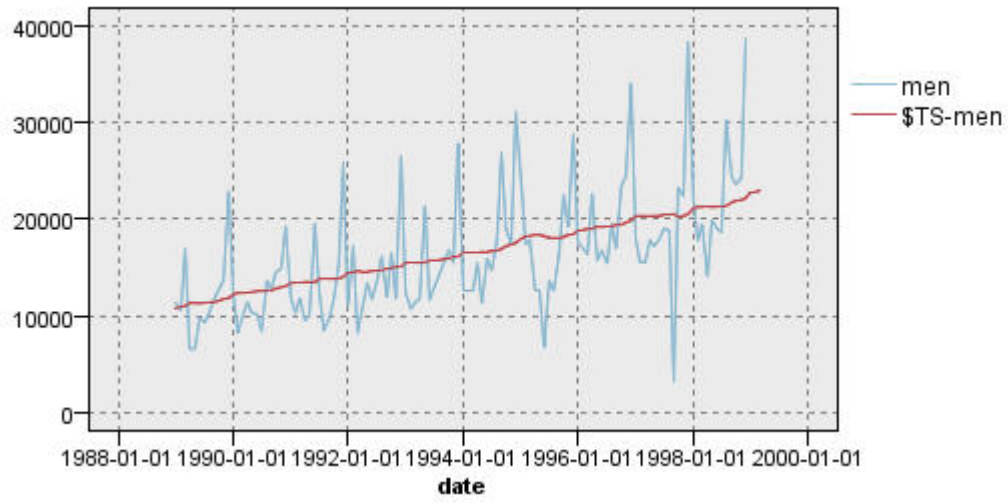


圖 *209. Holt* 線性趨勢模型

Holt 模型相比簡式模型，顯示更平滑的上升趨勢，但它仍然不考量週期，因此您也可以捨棄此模型。

17. 關閉時間圖視窗。

您可能記起在一段時間內的男士服裝銷售量起始圖所建議的模型包含一個線性趨勢和一個相乘性週期。因此，更合適的候選模型可能為 Winter 模型。

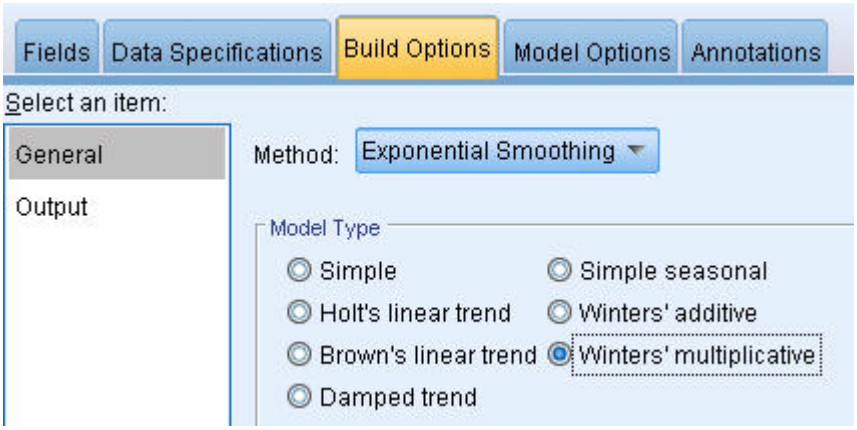


圖 *210.* 選取 *Winter* 模型

18. 重新開啟「時間序列」節點。
19. 在「建置選項」標籤上的「一般」窗格中，仍然選取**指數平滑**作為**方法**，選取 **Winter 相乘性**作為**模型類型**。
20. 按一下**執行**以重建模型區塊。
21. 開啟「時間圖」節點，然後按一下**執行**。

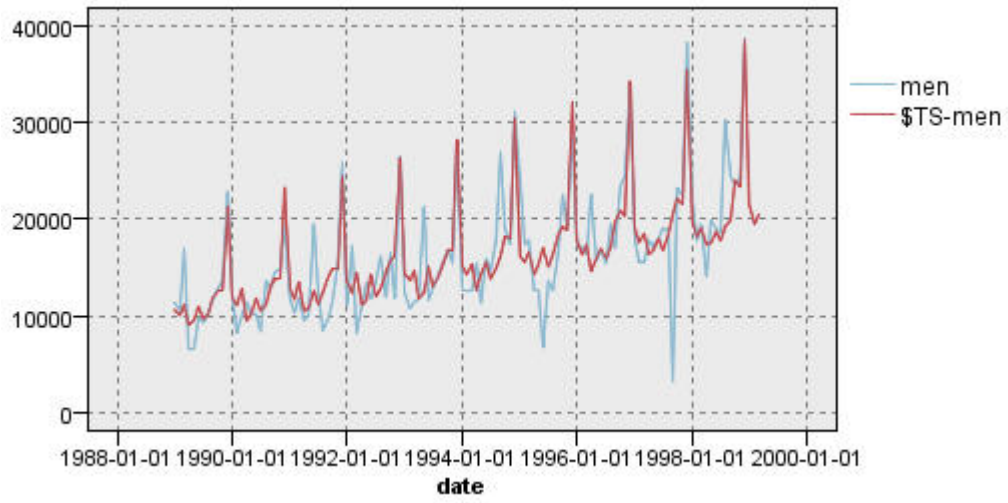


圖 *211. Winter* 相乘性模型

此模型看上去更佳；它同時反映了資料的趨勢和週期。

資料集涵蓋了 10 年的資料，並且包括 10 個週期性尖峰（在每年的 12 月出現）。預測結果中的 10 個尖峰與實際資料的 10 個年度尖峰很相符。

但是，結果也強調了指數平滑程序的限制。同時查看上升與下降尖峰，存在重要的結構未考量。

如果您的主要興趣是對具有週期性變動的長期趨勢建模，則指數平滑可能是一個不錯的選擇。若要對更複雜的結構（例如這個結構）建模，我們必須考量使用 ARIMA 程序。

ARIMA 程序

可使用 ARIMA 程序來建立自動迴歸整合移動平均 (ARIMA) 模型，該模型適合對時間序列進行精細建模。與指數平滑化模型相比，ARIMA 模型在對趨勢和週期性成份建模方面提供更準確的方法，並且增加了可在模型中包含預測工具變數的優勢。

繼續查看想要開發預測模型的型錄公司範例，我們瞭解了公司如何收集男士服裝每月銷售量的相關資料，以及可用來說明銷售量中部分變異的數個序列的相關資料。可能的預測工具包括郵寄的型錄數目、型錄中的頁數、開放用於訂購的電話線路數目、列印廣告花費的金額以及客戶服務代表數目。

還有任何預測工具對預測有幫助嗎？具有預測工具的模型真的好於沒有預測工具的模型嗎？我們可以使用 ARIMA 程序來建立具有預測工具的預測模型，並查看透過沒有預測工具的指數平滑模型進行預測的能力是否有顯著的差異。

利用 ARIMA 方法，您可以透過向這些成份指定自動迴歸、差異分析、移動平均值以及週期性對應項目的順序來細部調整模型。手動確定這些成份的最佳值非常耗時，其中涉及很多試錯，因此對於此範例，我們將讓 Expert Modeler 選擇 ARIMA 模型。

我們將嘗試透過將資料集中的部分其他變數視為預測工具變數來建置更好的模型。看起來最有可能併入作為預測工具的變數包括郵寄的型錄數目 (mail)、型錄中的頁數 (page)、開放用於訂購的電話線路數目 (phone)、列印廣告花費的金額 (print) 以及客戶服務代表數目 (service)。

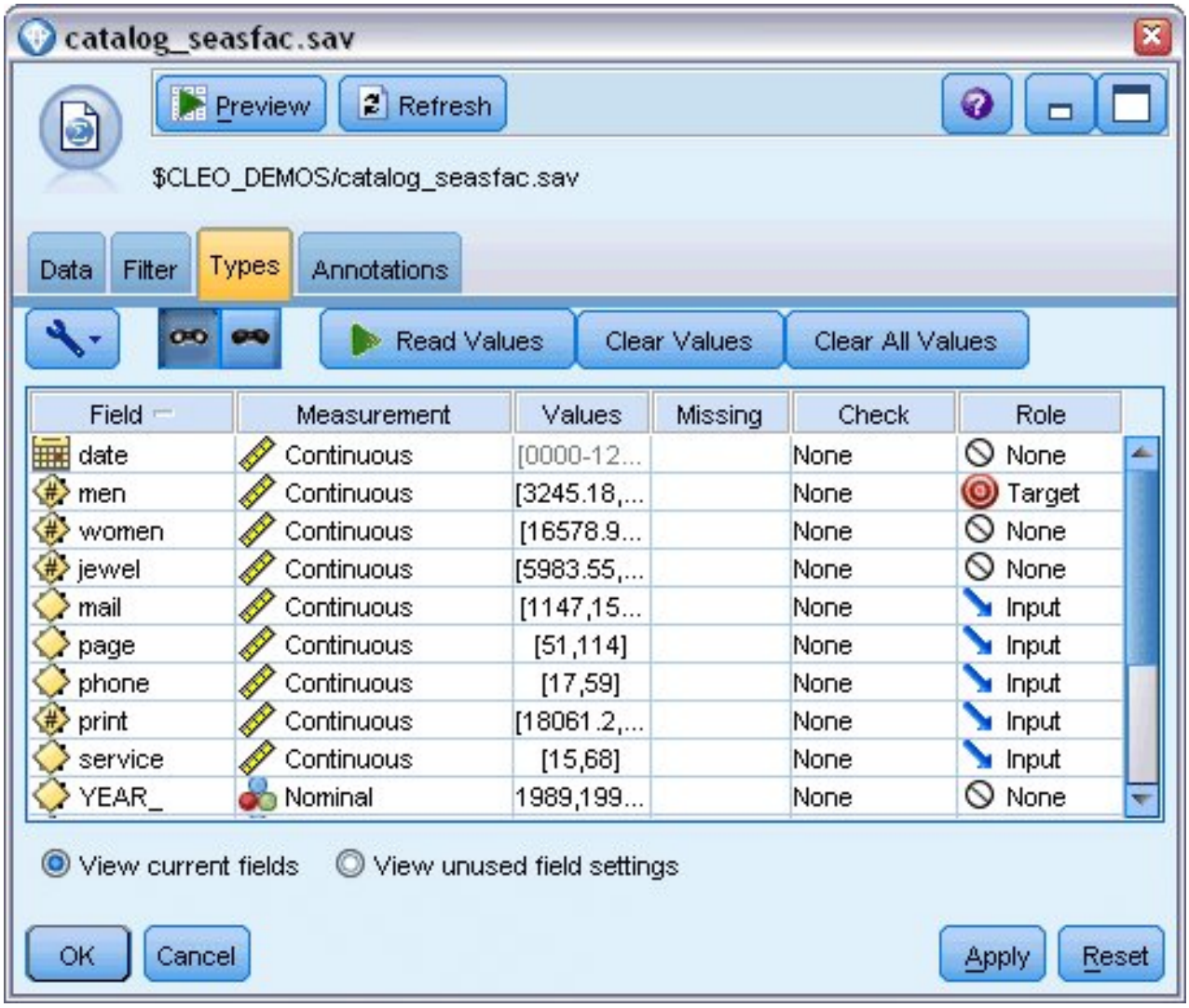


圖 *212.* 設定預測工具欄位

1. 開啟 IBM SPSS Statistics 的「檔案」來源節點。
2. 在「類型」標籤上，將 `mail`、`page`、`phone`、`print` 和 `service` 的**角色**設為**輸入**。
3. 確保 `men` 的角色設為**目標**，且所有剩餘欄位設為**無**。
4. 按一下**「確定」**。
5. 開啟「時間序列」節點。
6. 在「建置選項」標籤上的「一般」窗格中，將**方法**設為 **Expert Modeler**。
7. 選取**僅限 ARIMA 模型**選項並確保已勾選 **Expert Modeler 考量週期性模型**。

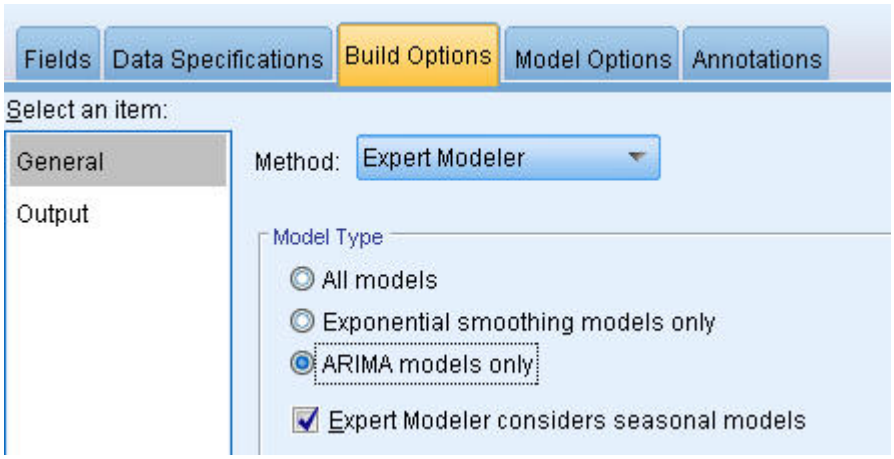


圖 *213.* 僅選擇 *ARIMA* 模型

8. 按一下**執行**以重建模型區塊。
9. 開啟模型區塊。

   在「輸出」標籤的左欄中，選取**模型資訊**。請注意，Expert Modeler 是如何只選擇五個指定預測工具中的兩個作為對模型而言很重要的預測工具。

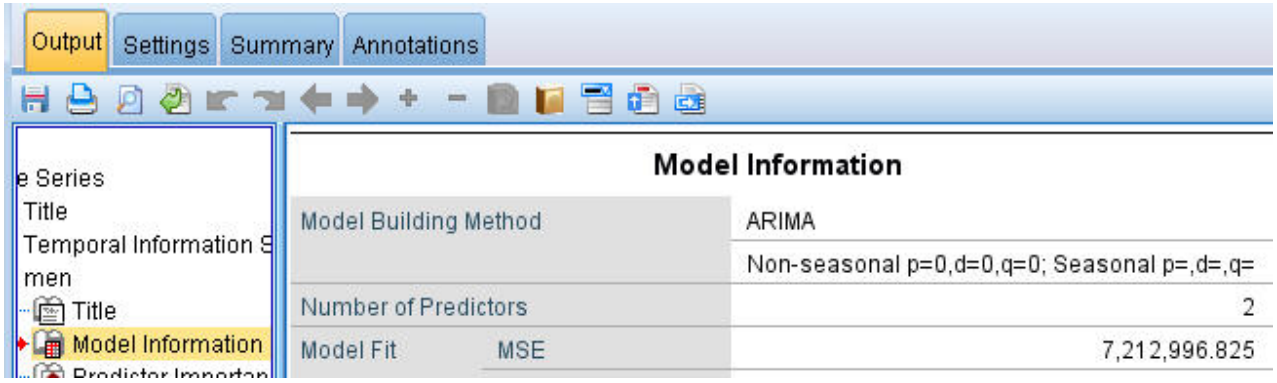


圖 *214. Expert Modeler* 選擇兩個預測工具

10. 按一下**確定**以關閉模型區塊。
11. 開啟「時間圖」節點，然後按一下**執行**。

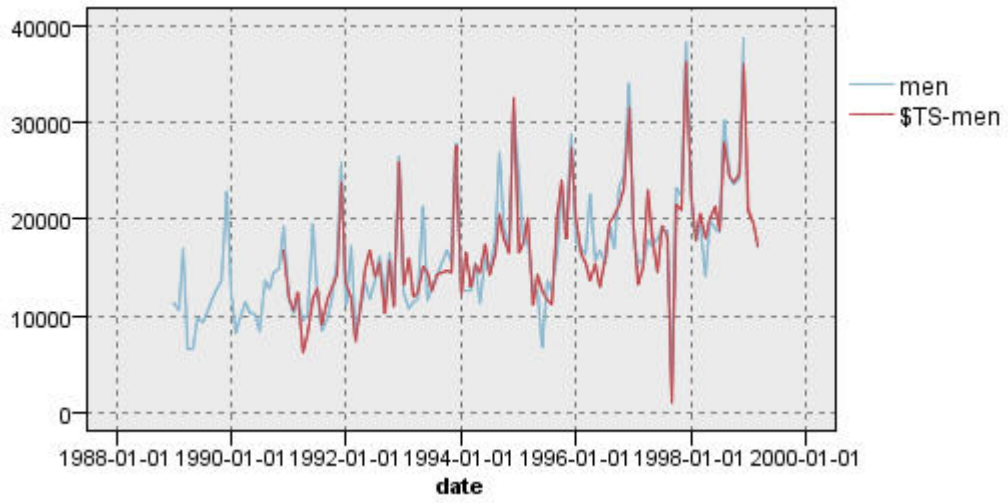


圖 *215.* 指定了預測工具的 *ARIMA* 模型

此模型在前一個模型上有所改進，方法是同時擷取大的下降峰值，使模型成為目前為止最適合的模型。

我們會嘗試進一步精簡模型，但從此時開始進行的任何改進可能都很小。我們已確定具有預測工具的 ARIMA 模型是偏好的模型，讓我們使用剛建置的模型吧。此範例的目的是為了預測來年的銷售量。

12. 按一下**確定**以關閉時間圖視窗。
13. 開啟「時間序列」節點並選取「模型選項」標籤。
14. 選取**將記錄延伸到未來**勾選框並將其值設為 12。
15. 選取**計算輸入的未來值**勾選框。
16. 按一下**執行**以重建模型區塊。
17. 開啟「時間圖」節點，然後按一下**執行**。

針對 1999 的預測看上去良好；如預期所示，在 12 月份的峰值過後迴歸到正常的銷售層次，在下半年的上升趨勢穩定，銷售量總體上高於去年。

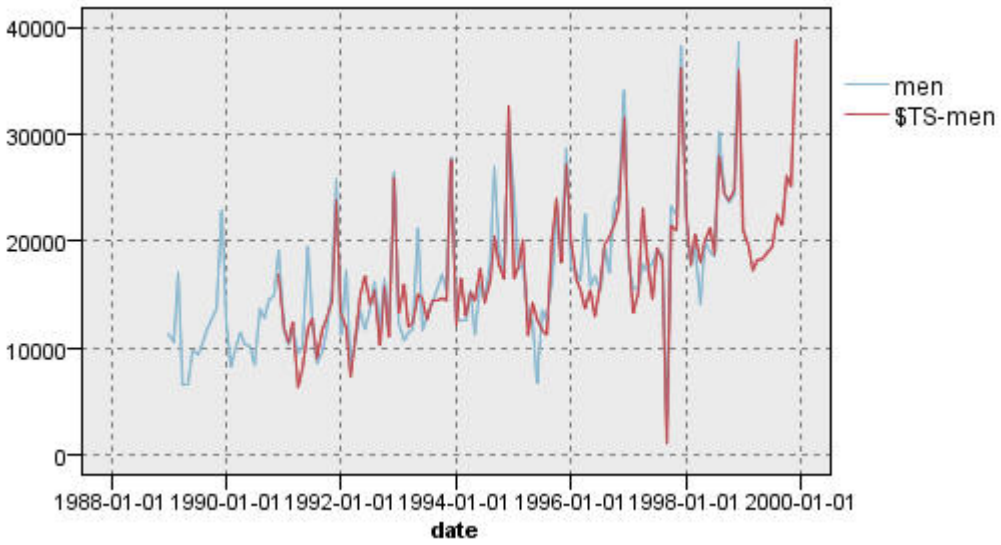


圖 *216.* 銷售量預測延伸 *12* 個月

摘要

您已順利為複雜的時間序列建模，其中不僅納入了上升趨勢，還包括週期性和其他變數。您也已瞭解如何透過試錯越來越靠近正確的模型，然後用來預測未來的銷售量。

實際上，您將需要重新套用模型，因為實際銷售資料已更新（例如每月或每季更新）並產生更新的預測。請參閱第 162 頁的『重新套用時間序列模型』主題，以取得更多資訊。

第 16 章 向客戶提供優惠（自我學習）

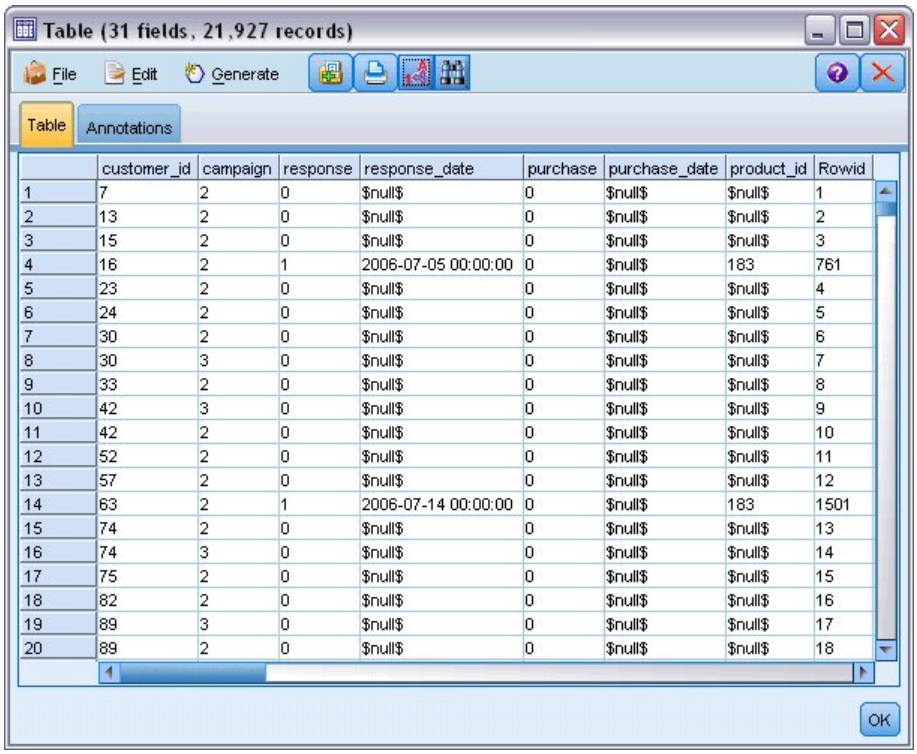
「自我學習回應模型 (SLRM)」節點會產生並啟用模型更新，讓您能夠預測哪個優惠最適合客戶以及預測接受優惠的機率。這些類型的模型最有利于客戶關係管理，例如市場行銷或呼叫中心。

本範例基於虛構銀行業公司。市場行銷部門希望通過向每個客戶提供合適的金融服務優惠來在未來的行銷活動中獲取更多利潤。尤其是，該範例使用「自我學習回應模型」來根據以前的優惠和回應識別最有可能做出積極回應之客戶的特性，並根據結果推出目前最好的優惠。

此範例使用串流 *pm_selflearn.str*，該串流參照 *pm_customer_train1.sav*、*pm_customer_train2.sav* 和 *pm_customer_train3.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾中獲取。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*pm_selflearn.str* 檔案位于 *streams* 資料夾中。

現有資料

公司具有歷程資料，可用於追蹤過去行銷活動中向客戶提供的優惠以及客戶對那些優惠作出的回應。這些資料還包括個人背景資訊和財務資訊，可用來預測不同客戶的回應率。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

圖 *217.* 對先前優惠的回應

建置串流

1. 新增指向 *pm_customer_train1.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「統計量檔案」來源節點。

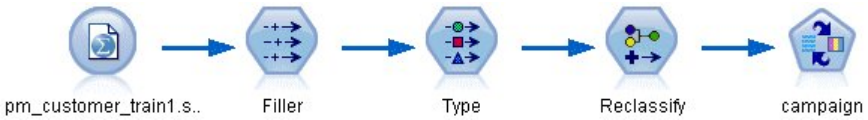


圖 *218. SLRM* 串流範例

2. 新增「填充值」節點並選取 `campaign` 作為欄位中的「填充」。
3. 選取**一律**作為「取代」類型。
4. 在「取代為」文字框中，輸入 `to_string(campaign)` 並按一下**確定**。

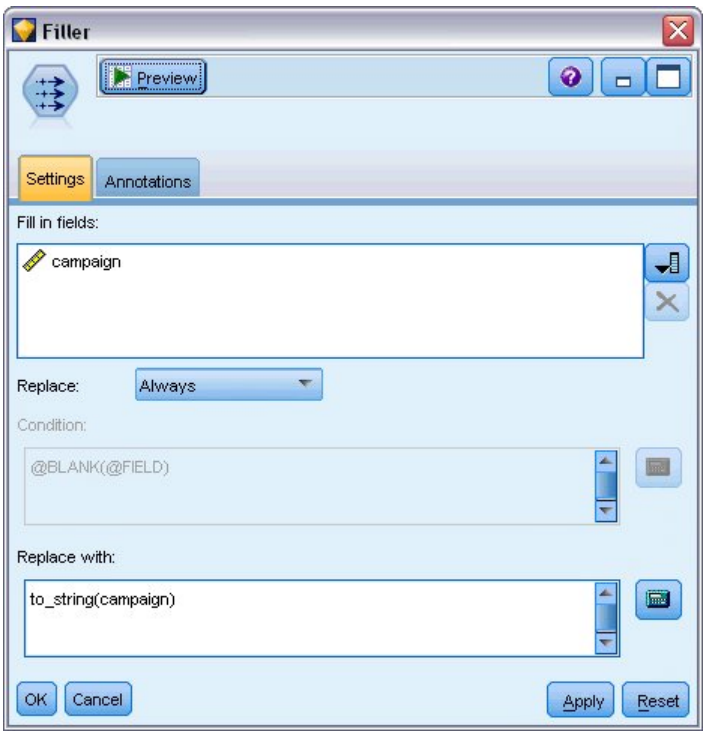


圖 *219.* 衍生行銷活動欄位

5. 新增「類型」節點，並將 *customer_id*、*response_date*、*purchase_date*、*product_id*、*Rowid* 和 *X_random* 欄位的角色 設為**無**。

圖 *220.* 變更類型節點設定

6. 將 *campaign* 和 *response* 欄位的角色設為**目標**。這些是您想要據以進行預測的欄位。

   將 *response* 欄位的**測量**設為**旗標**。
7. 按一下**讀取值**，然後按一下**確定**。

   由於行銷活動欄位資料顯示為數字清單（1、2、3 和 4），因此您可以重新分類欄位以具有更有意義的標題。
8. 將「重新分類」節點新增至「類型」節點。
9. 在**重新分類為**欄位中，選取**現有欄位**。
10. 在**重新分類欄位**清單中，選取**行銷活動**。
11. 按一下**取得**按鈕；行銷活動值即會新增至原始值直欄。
12. 在新值直欄中，在前四列中輸入下列行銷活動名稱：
    - **抵押**
    - **汽車貸款**
    - **節約**
    - **津貼**
13. 按一下**確定**。

圖 *221.* 重新分類行銷活動名稱

14. 將 SLRM 建模節點連接至「重新分類」節點。在「欄位」標籤上，針對「目標」欄位選取**行銷活動**，針對「目標回應」欄位選取**回應**。

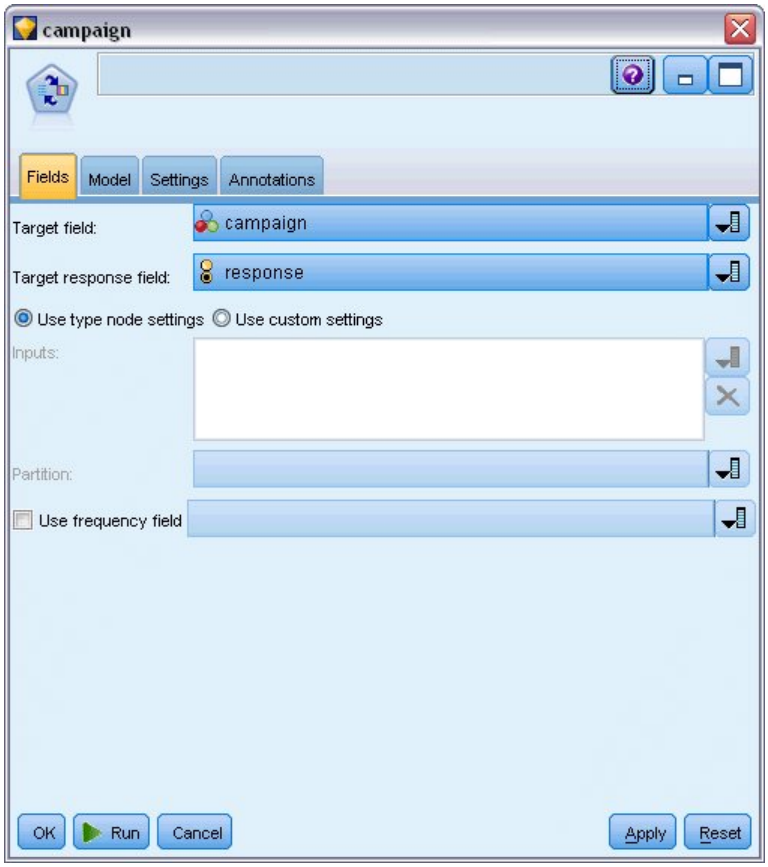


圖 222. 選取目標和目標回應

15. 在「設定」標籤的「每筆記錄的最大預測數」欄位中，將數目減到 2。

    這表示對於每位客戶，將有兩個識別的優惠被客戶接受的可能性最高。

16. 確保選取了**考量模型可靠性**，然後按一下**執行**。

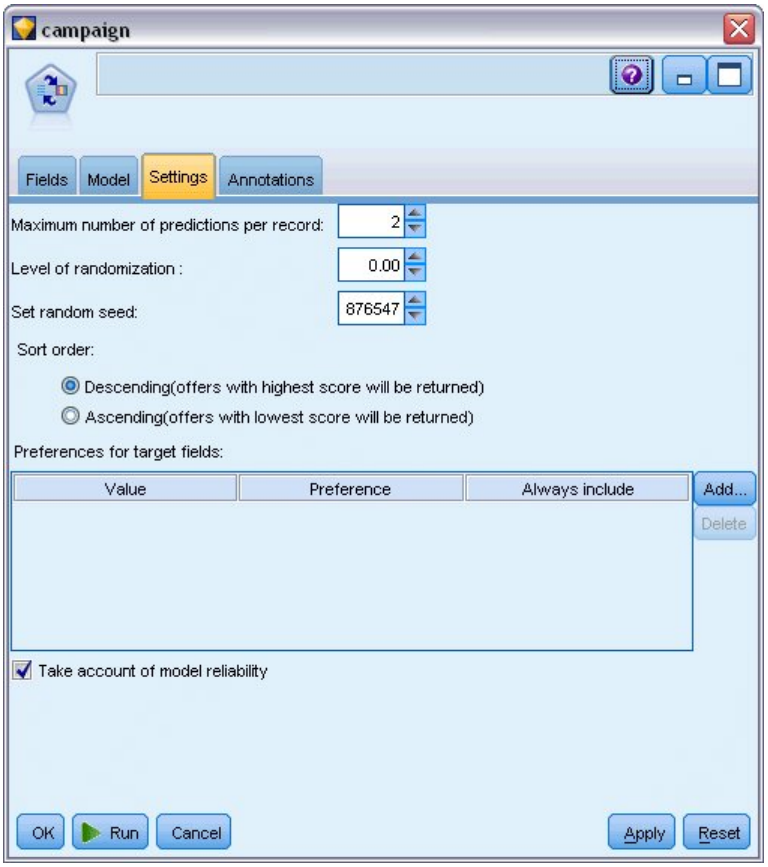


圖 *223. SLRM* 節點設定

瀏覽模型

1. 開啟模型區塊。「模型」標籤顯示每個優惠的預測正確性預估值，以及在估計模型時每個預測工具的相對重要性。

   若要顯示具有目標變數之每個預測工具的相關性，請從右窗格的**視圖**清單選擇**與回應的相關性**。
2. 若要在存在預測值的四個優惠之間切換，請從左窗格的**視圖**清單中選取所需的優惠。

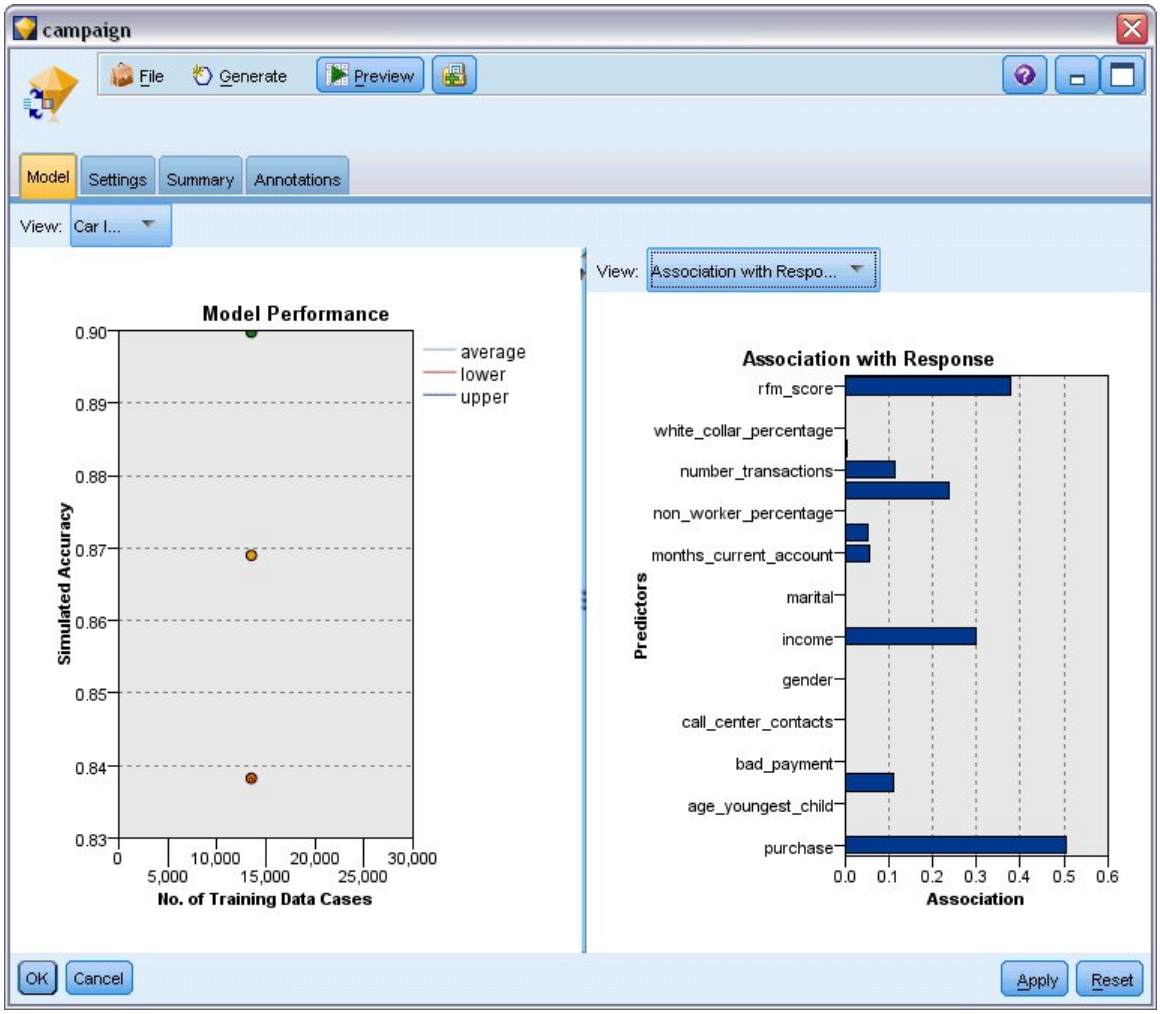


圖 *224. SLRM* 模型區塊

3. 關閉模型區塊視窗
4. 在串流畫布上，中斷指向 *pm_customer_train1.sav* 之 IBM SPSS Statistics 檔案來源節點的連線。
5. 新增指向 *pm_customer_train2.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「統計量檔案」來源節點，並將它連接至「填充值」節點。

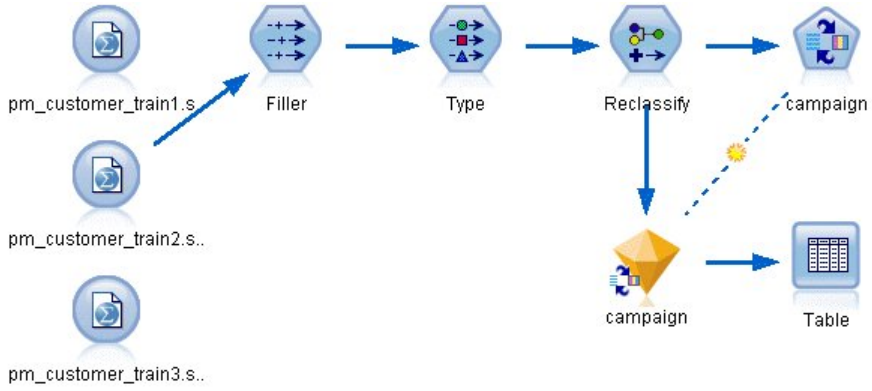


圖 *225.* 將第二個資料來源連接至 *SLRM* 串流

6. 在 SLRM 節點的「模型」標籤上，選取**繼續訓練現有模型**。

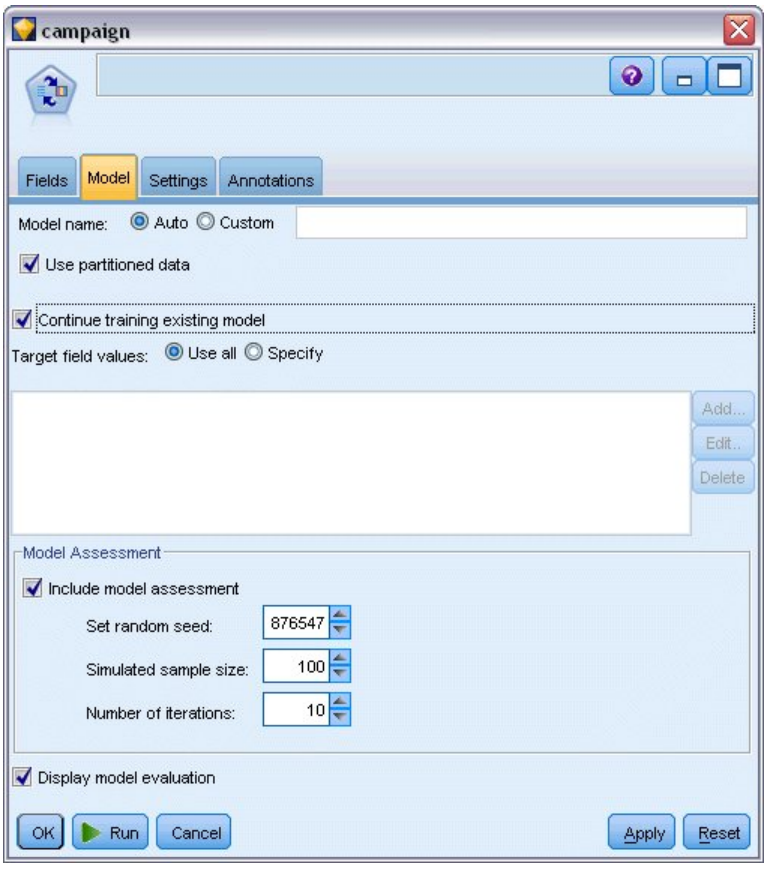


圖 *226.* 繼續訓練模型

7. 按一下**執行**以重建模型區塊。若要檢視其詳細資料，請在畫布上按兩下該區塊。

   「模型」標籤現在會顯示每個優惠的修訂的預測正確性預估值。

8. 新增指向 *pm_customer_train3.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「統計量檔案」來源節點，並將它連接至「填充值」節點。

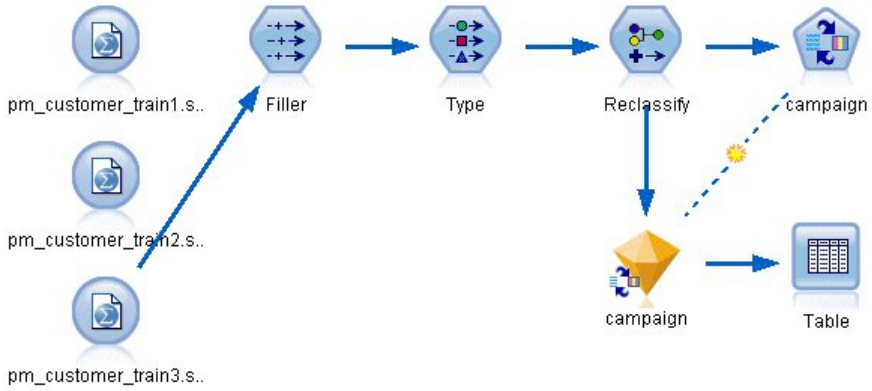


圖 *227.* 將第三個資料來源連接至 *SLRM* 串流

9. 按一下**執行**以再次重建模型區塊。若要檢視其詳細資料，請在畫布上按兩下該區塊。

10. 「模型」標籤現在會顯示每個優惠的最終預測正確性預估值。

    如您所見，當您新增額外的資料來源時，平均正確性輕微下降（從 86.9% 下降到 85.4%）；然而，這是最小波動量，可能導致可用資料內發生輕微異常。

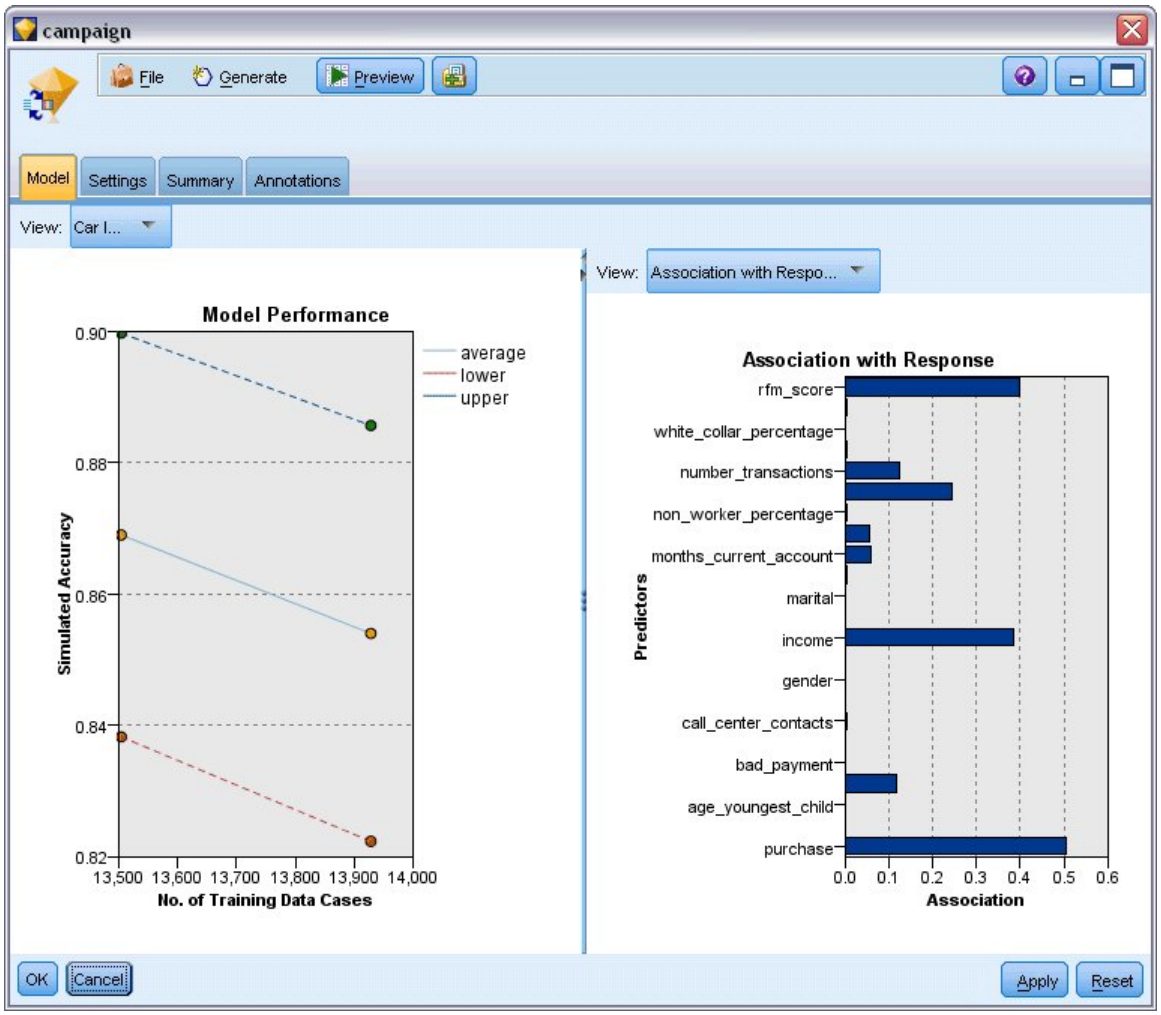
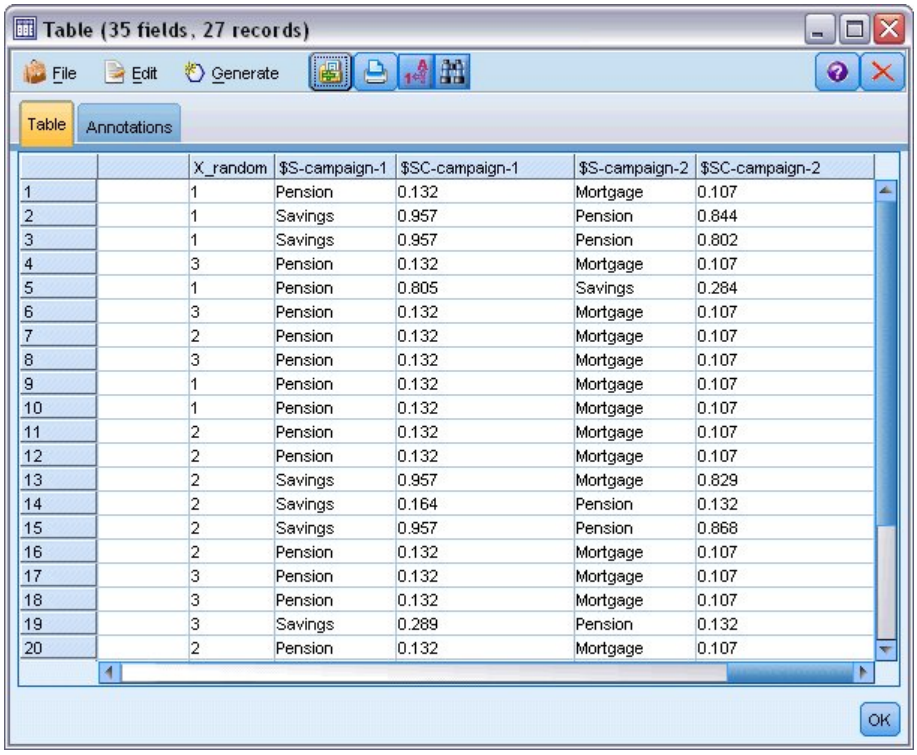


圖 *228.* 已更新 *SLRM* 模型區塊

11. 將「表格」節點附加至最後一個（即第三個）產生的模型並執行「表格」節點。
12. 捲動到表格右側。預測顯示客戶最有可能接受的優惠，以及客戶將接受的信賴度，視每個客戶的詳細資料而定。

例如，在所示表格的第一行，僅存在一個 13.2% 的信賴度（在 *$SC-campaign-1* 欄中以值 0.132 表示），先前有汽車貸款的客戶將接受津貼（如有提供的話）。但是，第二行和第三行顯示的另外兩個客戶也有汽車貸款；在他們的案例中，信賴度為 95.7% 的他們以及具有類似歷程的其他客戶可開啟儲蓄帳戶（如有提供的話），如果信賴度超過了 80%，則可以接受津貼。



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

圖 *229.* 模型輸出 - 預測的優惠和信賴度

在《*IBM SPSS Modeler* 演算法手冊》中列出了在 IBM SPSS Modeler 中所使用建模方法的數學基礎說明，該手冊在產品下載過程中以 PDF 檔案形式提供。

另請注意，僅根據訓練資料得出這些結果。若要評量模型推廣到真實世界中的其他資料的程度，您可以使用「分割區」節點來送出一部分記錄用於測試和驗證。

第 17 章 預測貸款違約者（貝氏網路）

貝氏網路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。

此範例使用參照資料檔 *bankloan.sav* 的串流 *bayes_bankloan.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取，並且可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*bayes_bankloan.str* 檔位於 *streams* 目錄。

例如，假設某銀行擔心有潛在的貸款不會償還。如果先前的貸款拖欠資料可用來預測哪些潛在客戶容易有償還貸款的問題，則可以拒絕向這些「不良風險」客戶貸款或為其提供替代產品。

此範例著重於使用現有貸款拖欠資料來預測潛在的未來違約者，並查看三個不同的貝氏網路模型類型來建立在此情況下進行預測的較佳模型。

建置串流

1. 新增指向 *Demos* 資料夾中 *bankloan.sav* 的「統計量檔案」來源節點。

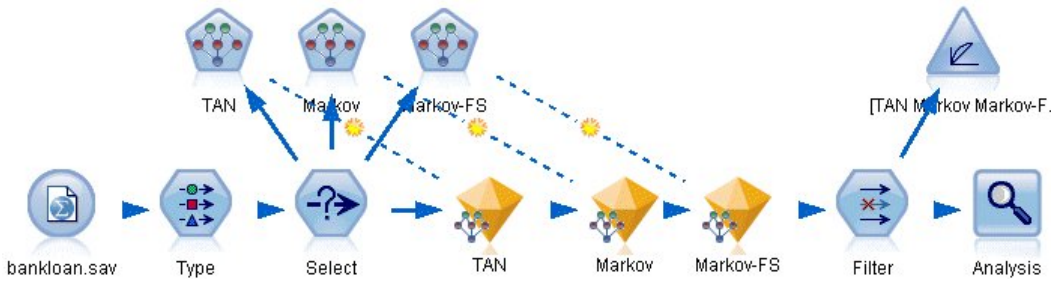


圖 *230.* 貝氏網路串流範例

2. 將「類型」節點新增至來源節點，並將**預設**欄位的角色設為**目標**。所有其他欄位應該將其角色設為**輸入**。
3. 按一下**讀取值**按鈕，在值 直欄中移入值。

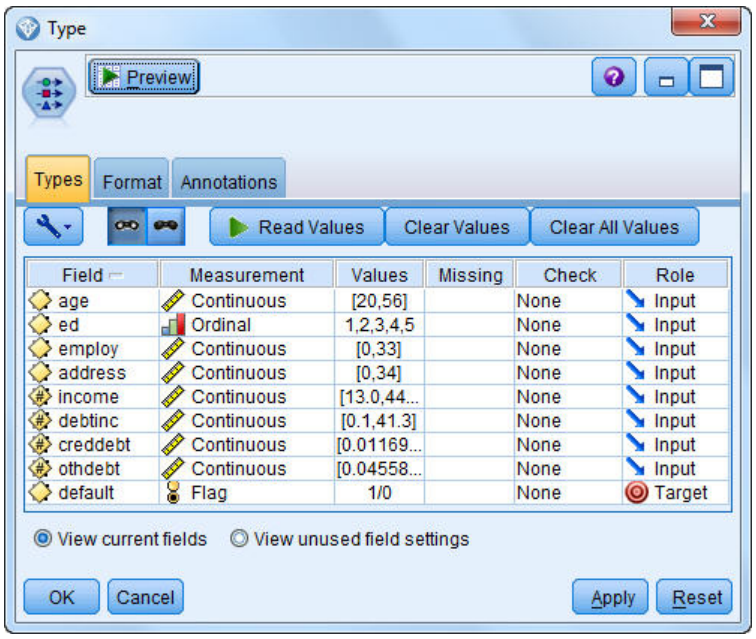


圖 *231.* 選取目標欄位

在建置模型時，目標具有空值的觀察值沒有用處。您可以排除這些觀察值以防止它們用在模型評估中。

4. 將「選取」節點新增至「類型」節點。
5. 針對模式選取**捨棄**。
6. 在「條件」方框中輸入 **default = '$null$'**。

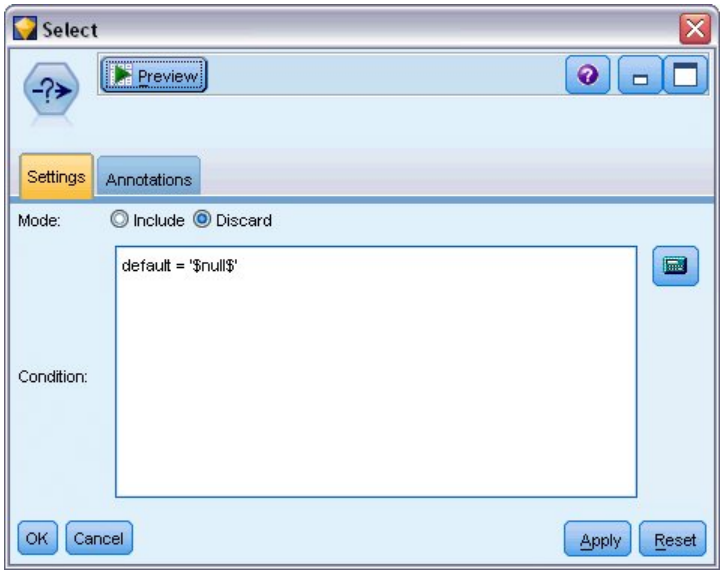


圖 *232.* 捨棄空值目標

由於您可以建置數種不同類型的貝氏網路，因此值得比較數個模型以查看哪個模型提供的預測最佳。要建立的第一個模型是 Tree Augmented Naïve Bayes (TAN) 模型。

7. 將「貝氏網路」節點新增至「選取」節點。
8. 在「模型」標籤上，針對模型名稱選取**自訂**，並在文字框中輸入 TAN。

9. 針對「結構」類型選取 **TAN**，然後按一下**確定**。

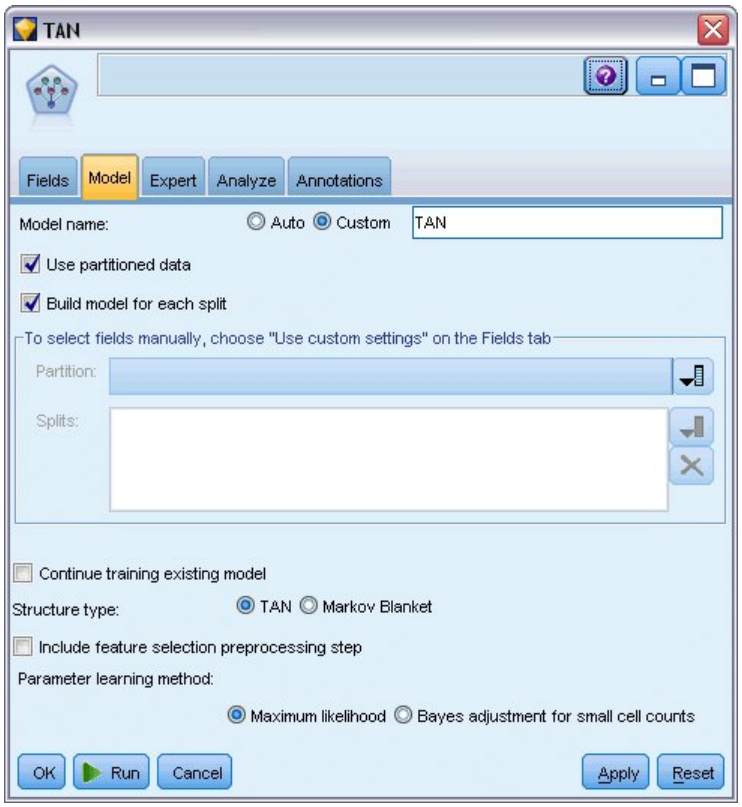


圖 *233.* 建立 *Tree Augmented Naïve Bayes* 模型

要建置的第二個模型類型是 Markov Blanket 結構。

10. 將第二個「貝氏網路」節點新增至「選取」節點。
11. 在「模型」標籤上，針對模型名稱選取**自訂**，並在文字框中輸入 `Markov`。
12. 針對「結構」類型選取 **Markov Blanket**，然後按一下**確定**。

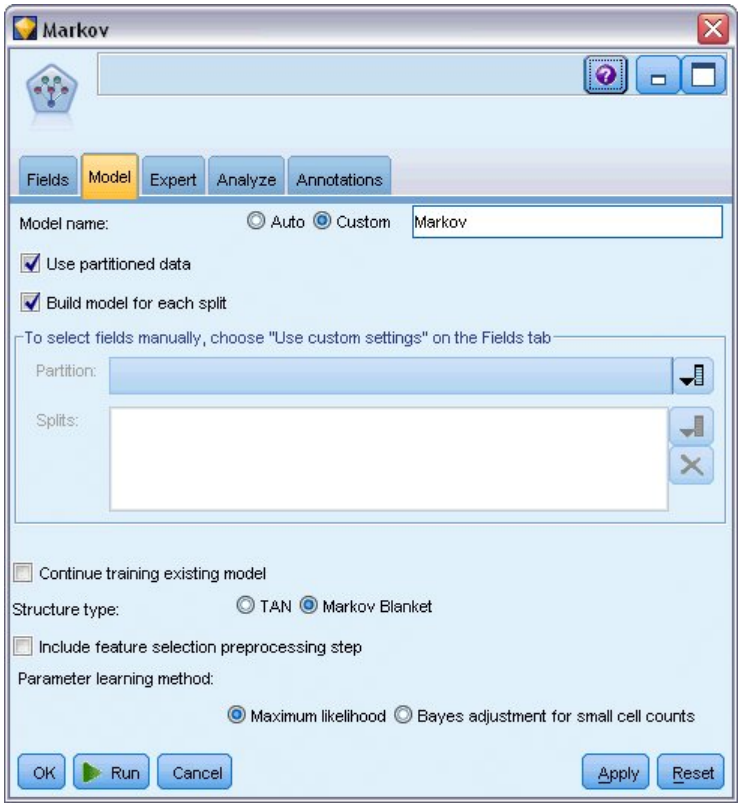


圖 *234. 建立 Markov Blanket 模型*

要建置的第三個模型類型具有 Markov Blanket 結構，還會使用功能選擇預先處理來選取與目標變數顯著相關的輸入。

13. 將第三個「貝氏網路」節點新增至「選取」節點。
14. 在「模型」標籤上，針對模型名稱選取**自訂**，並在文字框中輸入 `Markov-FS`。
15. 針對「結構」類型選取 **Markov Blanket**。
16. 選取**包括功能選擇預先處理步驟**，然後按一下**確定**。

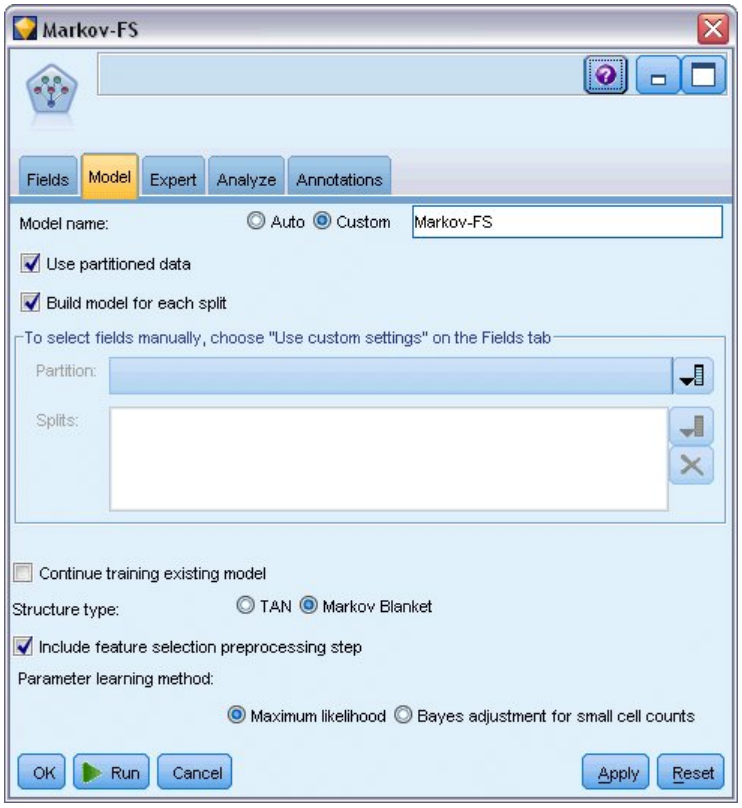


圖 *235.* 建立具有功能選擇預先處理的 *Markov Blanket* 模型

瀏覽模型

1. 執行串流來建立模型區塊，模型區塊將會新增至串流以及右上角的「模型」選用區中。若要檢視其詳細資料，請在串流中的任意模型區塊上按兩下。

   模型區塊「模型」標籤分為兩個窗格：左窗格包含節點網路圖形，此圖形顯示目標與其最重要的預測工具之間的關係，以及各預測工具之間的關係。

   右窗格顯示預測工具重要性，其指出在預估模型時每個預測工具的相對重要性，或顯示條件機率，其中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。

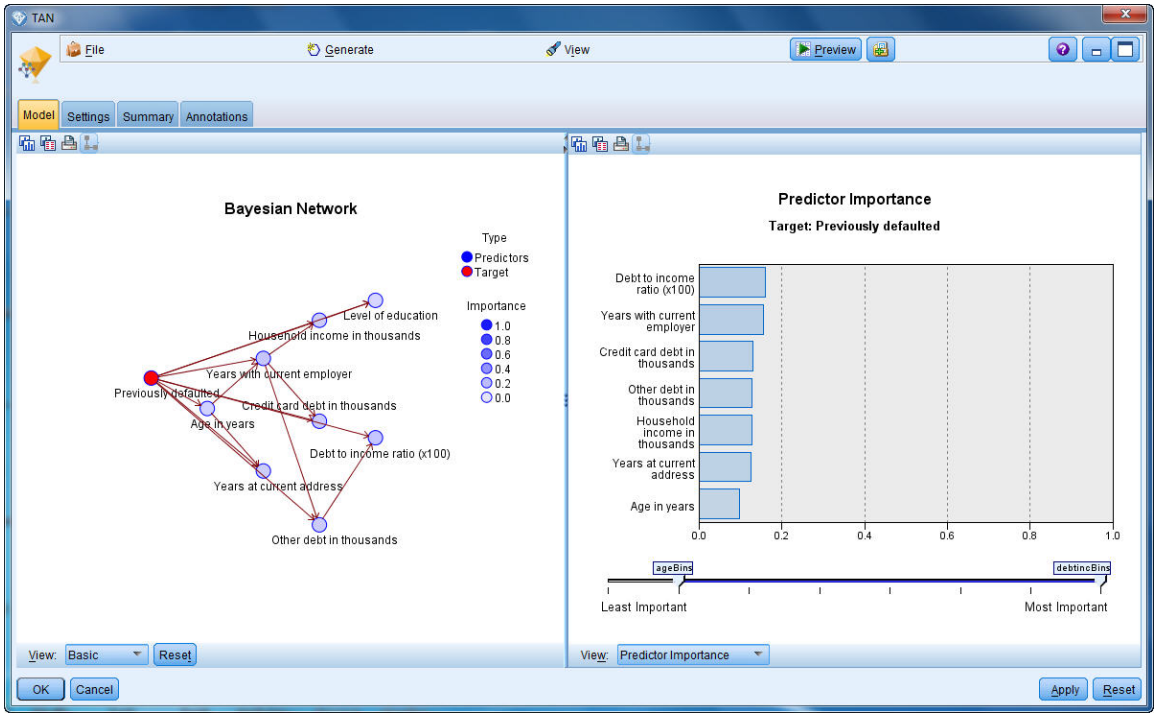


圖 *236.* 檢視 *Tree Augmented Naïve Bayes* 模型

2. 將 TAN 模型區塊連接至 Markov 區塊（選擇警告對話框上的**取代**）。
3. 將 Markov 區塊連接至 Markov-FS 區塊（選擇警告對話框上的**取代**）。
4. 將三個區塊與「選取」節點對齊以便於檢視。

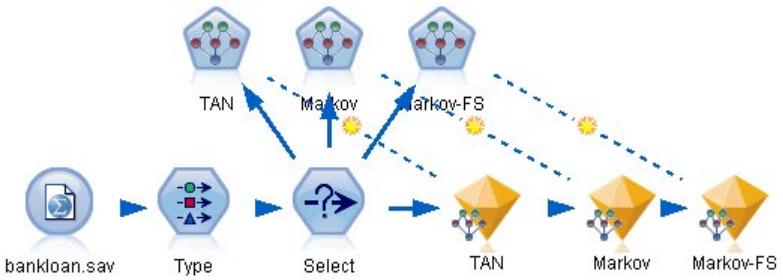


圖 *237.* 對齊串流中的區塊

5. 若要在您將建立的「評估」圖形上重新命名模型輸出以便於清楚說明，請將「過濾器」節點連接至 Markov-FS 模型區塊。
6. 在右邊的欄位 直欄中，將 $B-default 重新命名為 TAN，將 $B1-default 重新命名為 Markov，將 $B2-default 重新命名為 Markov-FS。

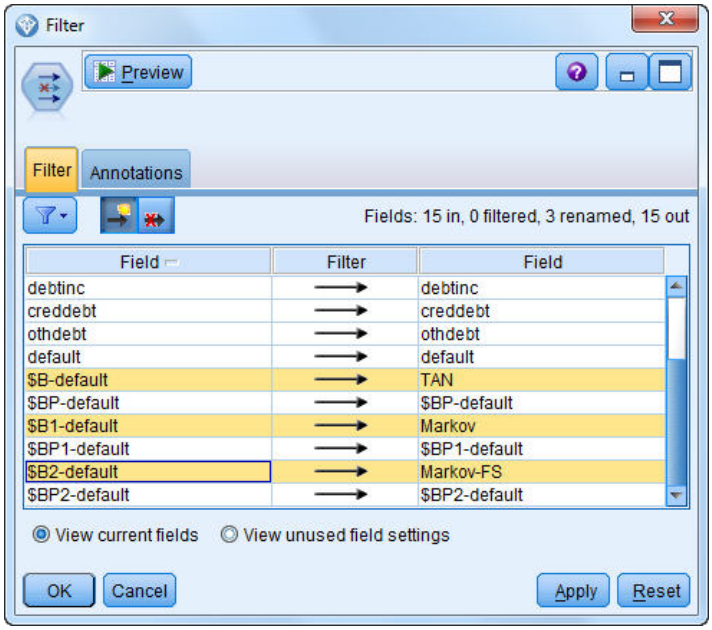


圖 *238.* 重新命名模型欄位名稱

若要比較模型的預測正確性，您可以建置一個增益圖表。

7. 將「評估」圖形節點連接至「過濾器」節點並使用其預設值來執行該圖形節點。

圖形顯示每個模型類型產生類似的結果；但是，Markov 模型略勝一籌。

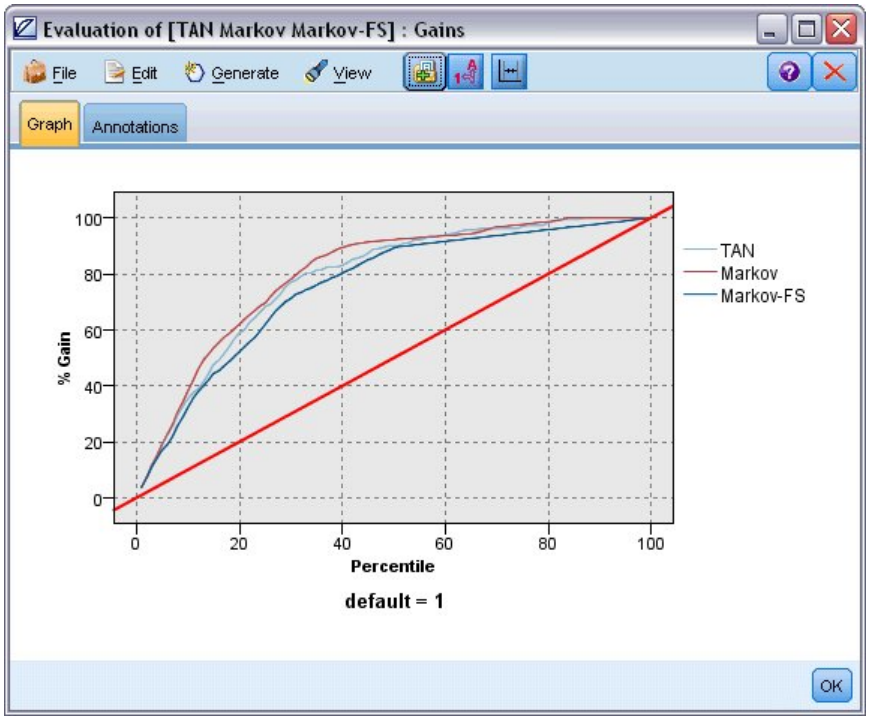


圖 *239.* 評估模型正確性

若要檢查每個模型的預測效果，您可以使用「分析」節點而非「評估」圖形。這會按照正確預測和不正確預測百分比來顯示正確性。

8. 將「分析」節點連接至「過濾器」節點並使用其預設值來執行「分析」節點。

就「評估」圖形而言，這顯示 Markov 模型在正確地預測方面略勝一籌；但是，Markov-FS 模型只是 Markov 模型後面的幾個百分比點。這可能表示使用 Markov-FS 模型更好，因為它使用更少的輸入來計算其結果，因此節省了資料收集、輸入和處理時間。

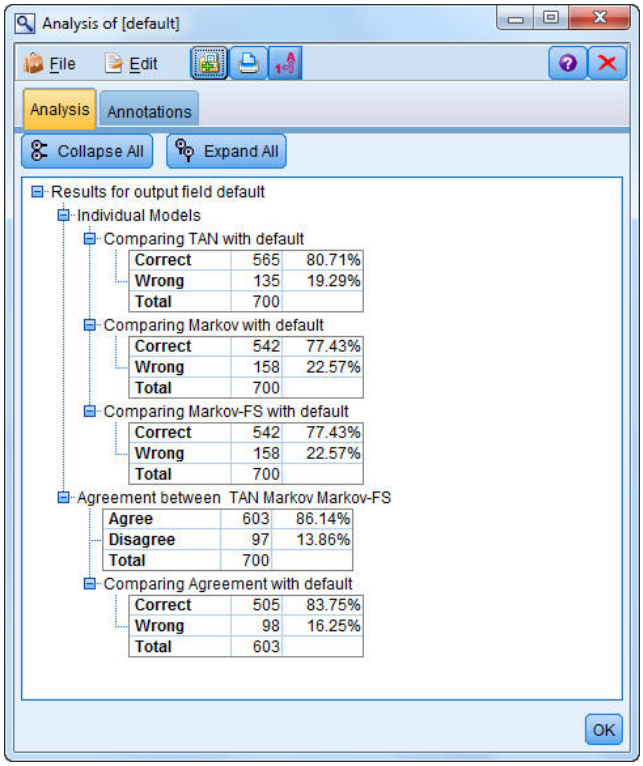


圖 *240.* 分析模型正確性

在《*IBM SPSS Modeler* 演算法手冊》中列出了在 IBM SPSS Modeler 中所使用建模方法的數學基礎說明，該手冊位於安裝磁碟的 *\Documentation* 目錄中。

另請注意，僅根據訓練資料得出這些結果。若要評量模型推廣到真實世界中的其他資料的程度，您可以使用「分割區」節點來送出一部分記錄用於測試和驗證。

第 18 章 每月重新訓練模型（貝氏網路）

貝氏網路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。

此範例使用參照資料檔案 *telco_Jan.sav* 和 *telco_Feb.sav* 的串流 *bayes_churn_retrain.str*。這些檔案可從任何 IBM SPSS Modeler 安裝架構的 *Demos* 目錄中獲取，並且可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*bayes_churn_retrain.str* 檔案位於 *streams* 目錄。

例如，假設某電信提供者擔心流失到競爭者的客戶數目（流失）。如果歷程客戶資料可用來預測未來哪些客戶更有可能會流失，則針對這些客戶可以使用獎勵或其他優惠來阻止他們轉向其他服務提供者。

此範例著重於使用現有月份的流失資料來預測未來可能會流失哪些客戶，然後新增下列月份資料以精簡和重新訓練模型。

建置串流

1. 新增指向 *Demos* 資料夾中 *telco_Jan.sav* 的「統計量檔案」來源節點。

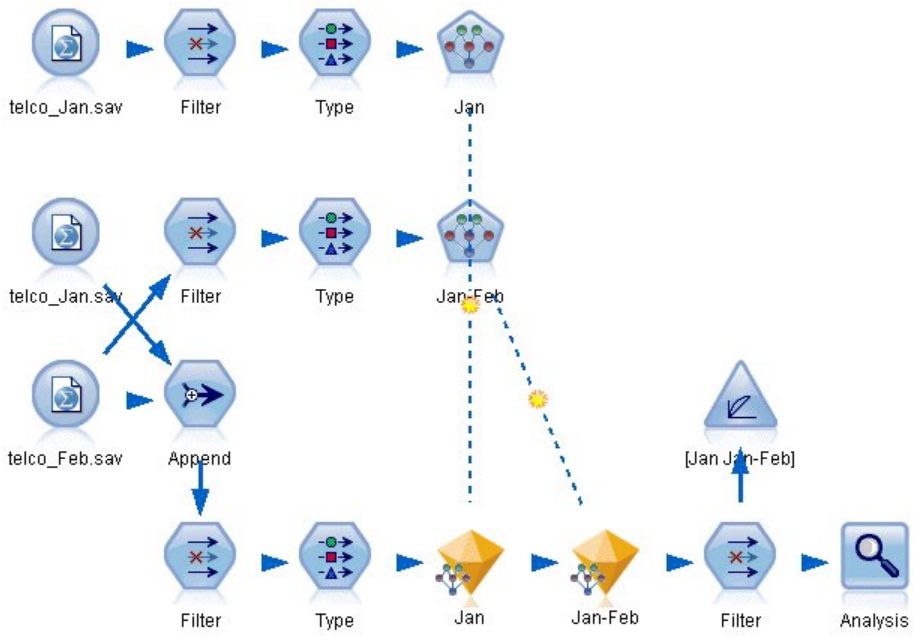


圖 *241.* 貝氏網路串流範例

先前的分析已向您顯示在預測流失時數個不太重要的資料欄位。可以從您的資料集中過濾出這些欄位，以便當您建置評分模型時增加處理速度。

2. 將「過濾器」節點新增至「來源」節點。
3. 排除 *address*、*age*、*churn*、*custcat*、*ed*、*employ*、*gender*、*marital*、*reside*、*retire* 和 *tenure* 以外的所有欄位。
4. 按一下**確定**。

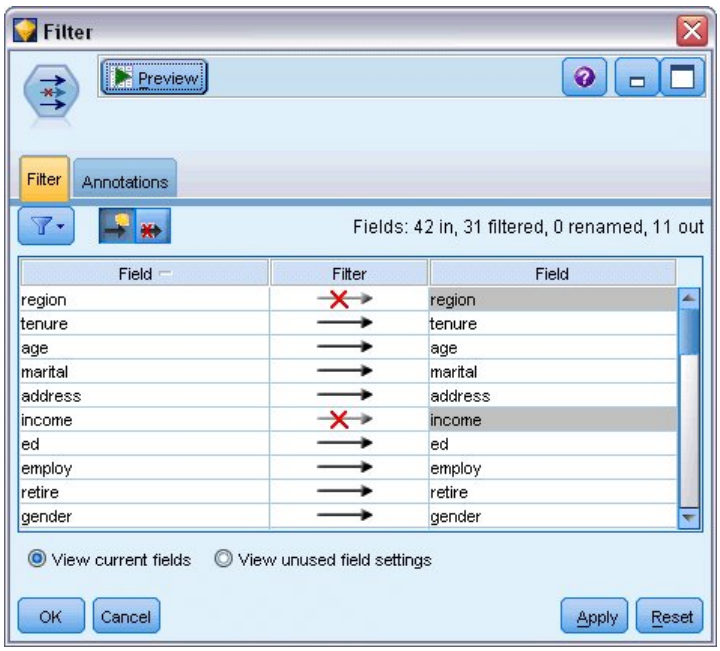


圖 *242.* 過濾不必要的欄位

5. 將「類型」節點新增至「過濾器」節點。
6. 開啟「類型」節點，按一下**讀取值**按鈕以在值 直欄中移入值。
7. 為了讓「評估」節點能夠評量哪個值為 true 哪個值為 false，請將 *churn* 欄位的測量層次設為**旗標**，並將其角色設為**目標**。按一下**確定**。

圖 *243.* 選取目標欄位

您可以建置數個不同類型的貝氏網路；但是，對于這個範例，您將建置一個 Tree Augmented Naïve Bayes (TAN) 模型。這樣做會建立一個大型網路，並確保您已包括資料變數之間的所有可能鏈結，從而建置一個強大的起始模型。

8. 將「貝氏網路」節點新增至「類型」節點。

9. 在「模型」標籤上，針對模型名稱選取**自訂**，並在文字框中輸入 Jan。
10. 針對「參數學習」方法，選取**小儲存格計數的貝氏調整**。
11. 按一下**「執行」**。模型區塊即會新增至串流，也會新增至右上角的「模型」選用區。

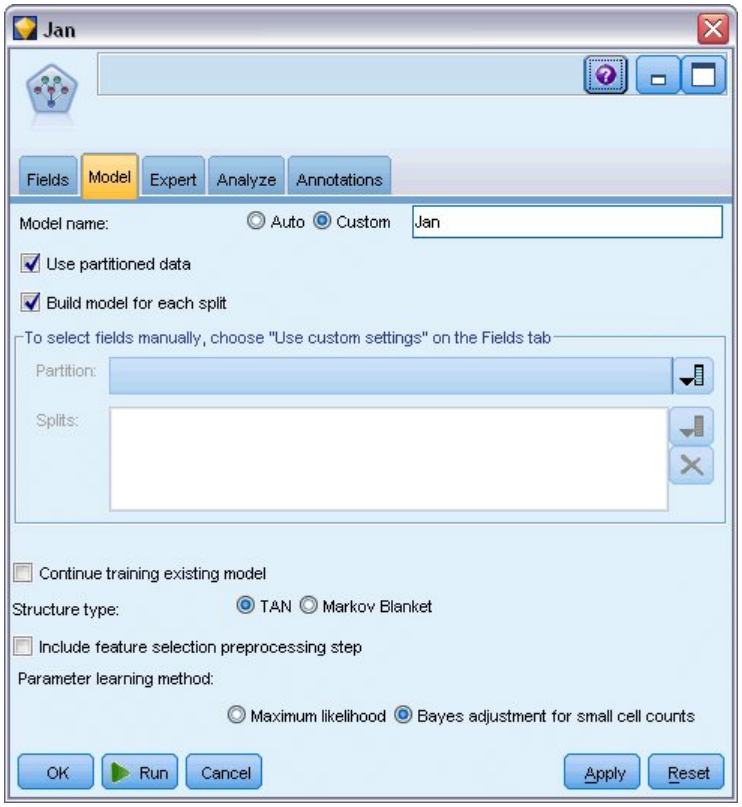


圖 *244.* 建立 *Tree Augmented Naïve Bayes* 模型

12. 新增指向 *Demos* 資料夾中 *telco_Feb.sav* 的「統計量檔案」來源節點。
13. 將這個新的來源節點連接至「過濾器」節點（在警告對話框上，選擇**取代**以取代與前一個來源節點的連線）。

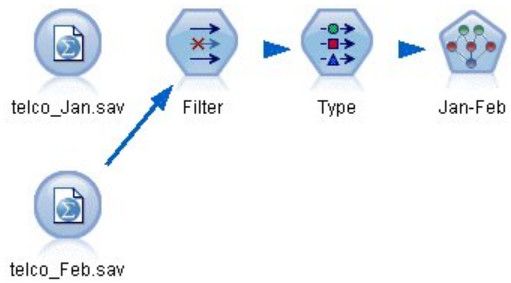


圖 *245.* 新增第二個月的資料

14. 在「貝氏網路」節點的「模型」標籤上，針對模型名稱選取**自訂**，並在文字框中輸入 Jan-Feb。
15. 選取**繼續訓練現有模型**。
16. 按一下**「執行」**。模型區塊會改寫串流中的現有模型區塊，但也會新增至右上角的「模型」選用區。

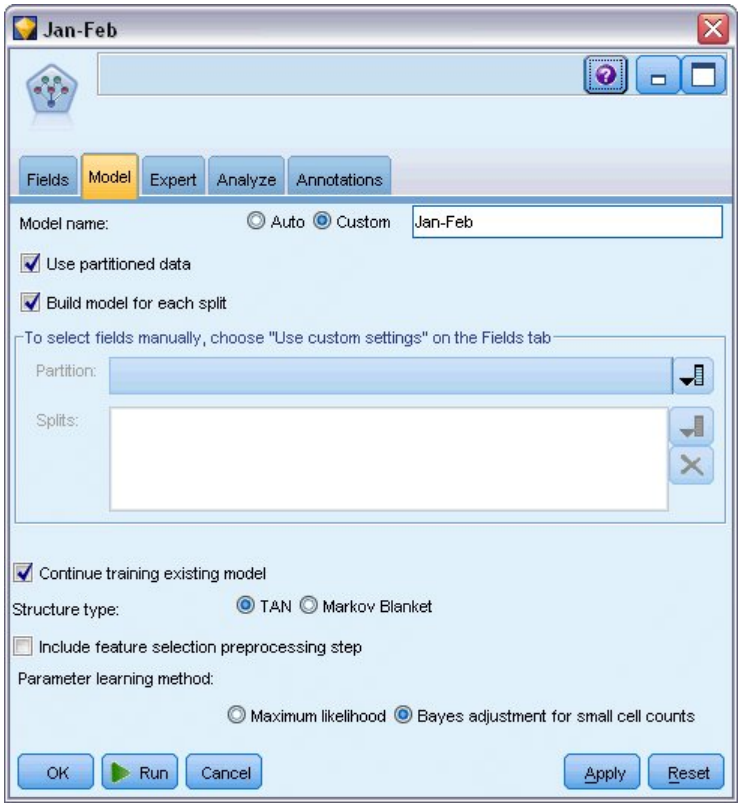


圖 *246.* 重新訓練模型

評估模型

若要比較模型，必須結合兩個資料庫。

1. 新增「附加」節點並在其中連接 *telco_Jan.sav* 和 *telco_Feb.sav* 來源節點。

圖 *247.* 附加兩個資料來源

2. 在串流中從先前項目複製「過濾器」和「類型」節點，並將其貼至串流畫布。
3. 將「附加」節點連接至新複製的「過濾器」節點。

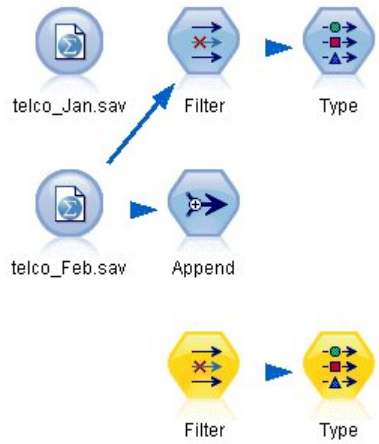


圖 *248.* 將複製的節點貼至串流

兩個貝氏網路模型的區塊位於右上角的「模型」選用區中。

4. 按兩下 Jan 模型區塊以將其帶入串流，並將它連接至新複製的「類型」節點。
5. 將已在串流中的 Jan-Feb 模型區塊連接到 Jan 模型區塊。
6. 開啟 Jan 模型區塊。

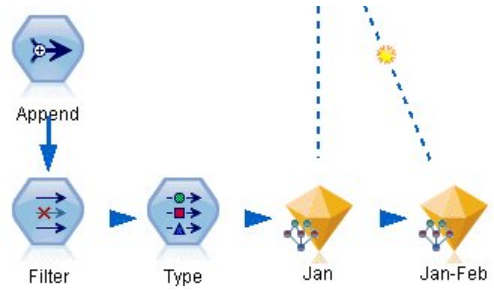


圖 *249.* 將區塊新增至串流

「貝氏網路」模型區塊的「模型」標籤分割成兩個直欄。左欄包含節點的網路圖形，此圖形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。

右欄顯示預測工具重要性，其指出在預估模型時每個預測工具的相對重要性，或顯示條件機率，其中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。

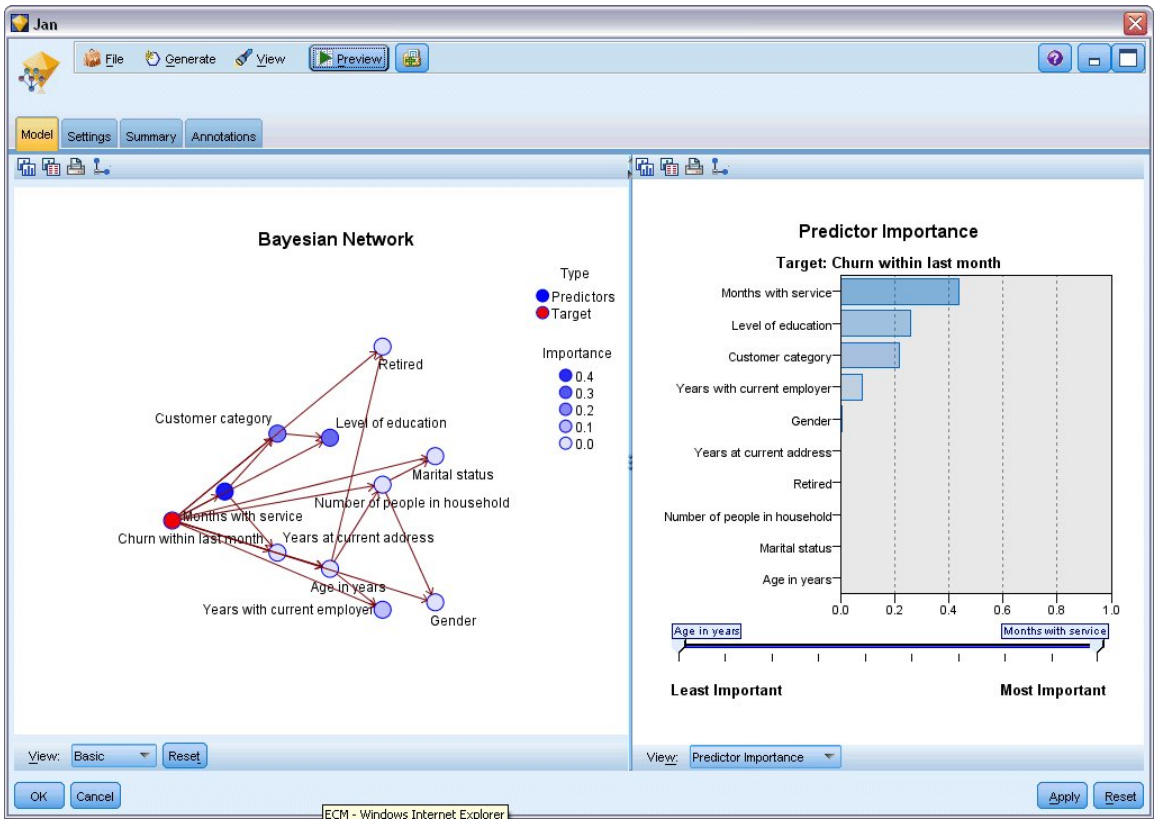


圖 *250.* 顯示預測工具重要性的貝氏網路模型

若要顯示任何節點的條件式機率，請按一下左欄中的節點。即會更新右欄以顯示所需的詳細資料。

對於資料值劃分成相對於節點的母項節點和同層級節點的每個 Bin，會顯示條件式機率。

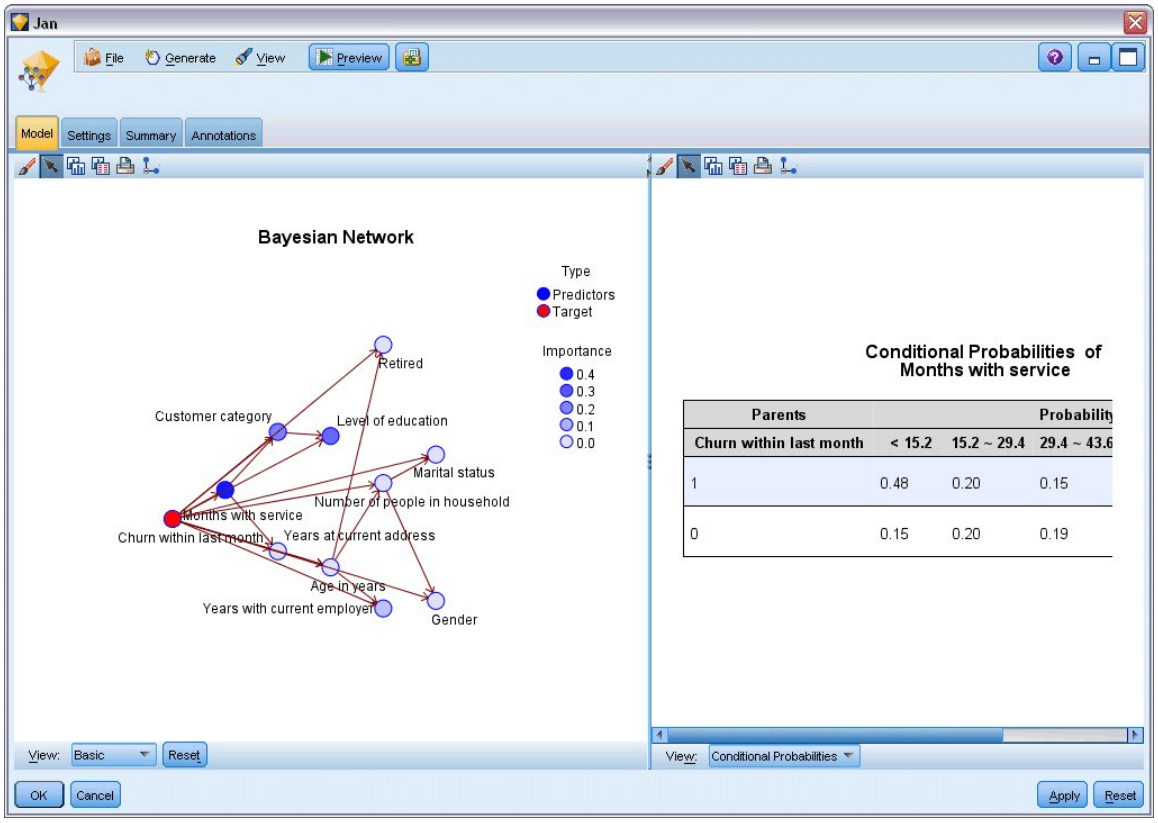


圖 *251.* 顯示條件式機率的貝氏網路模型

7. 若要重新命名模型輸出以便於清楚說明，請將「過濾器」節點連接至 Jan-Feb 模型區塊。
8. 在右側欄位 直欄中，將 $B-churn 重新命名為 Jan，將 $B1-churn 重新命名為 Jan-Feb。

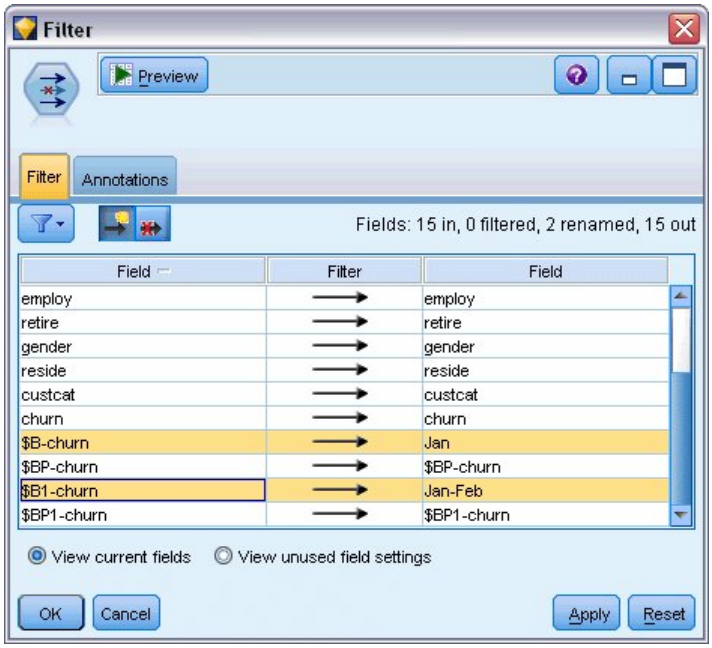


圖 *252.* 重新命名模型欄位名稱

若要檢查每個模型預測流失的效果，請使用「分析」節點；這會根據正確和不正確的預測百分比來顯示正確性。

9. 將「分析」節點連接至「過濾器」節點。
10. 開啟「分析」節點並按一下**執行**。

這會顯示兩個模型在預測流失時擁有類似的準確度。

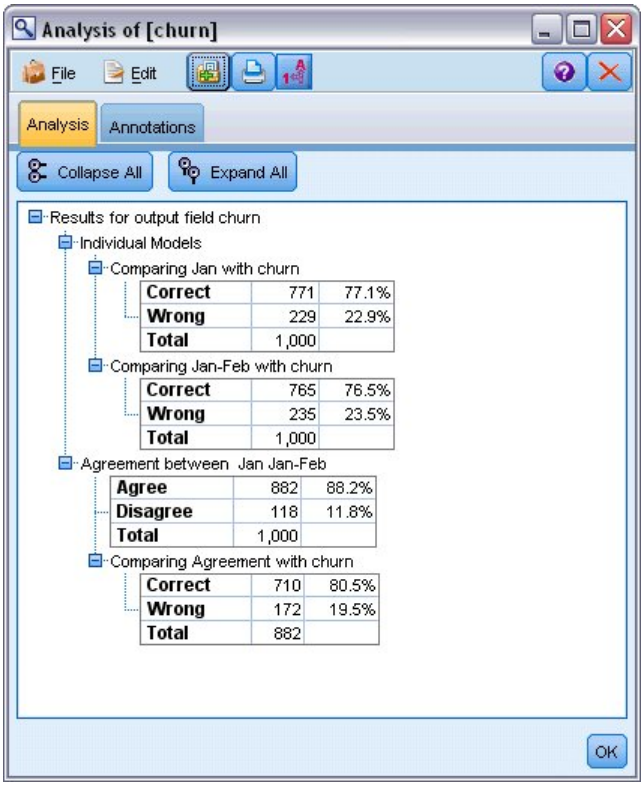


圖 *253.* 分析模型正確性

作為「分析」節點的替代，您可以使用「評估」圖形透過建置增益圖表來比較模型的預測正確性。

11. 將「評估」圖形節點連接至「過濾器」節點。

並使用其預設值來執行圖形節點。

就「分析」節點而言，圖形顯示每個模型類型產生相似的結果；但是，使用兩個月份資料重新訓練過的模型略勝一籌，因為它在其預測中的信賴度更高。

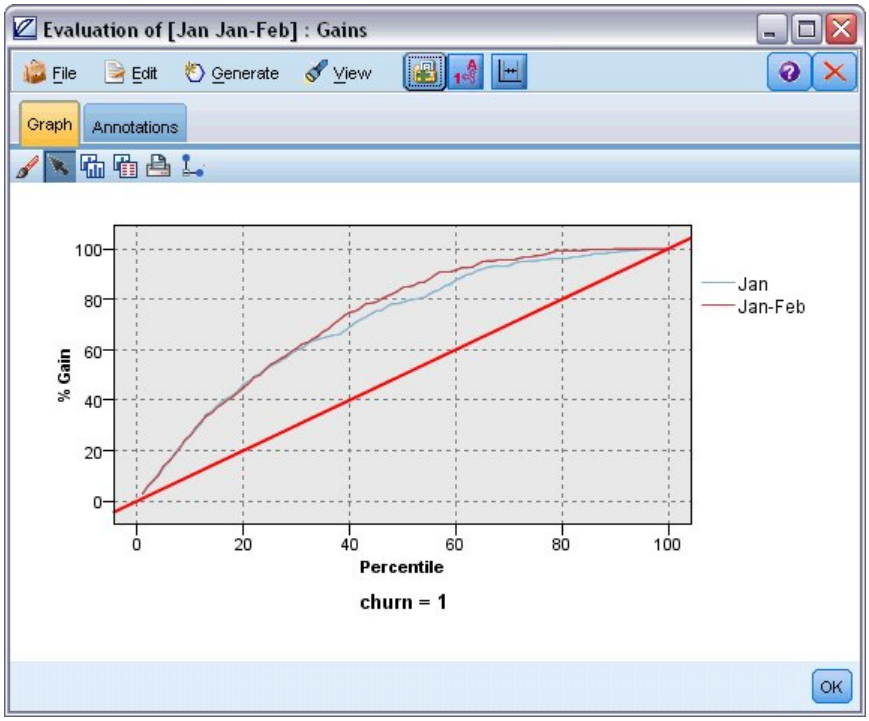


圖 *254.* 評估模型正確性

在《*IBM SPSS Modeler* 演算法手冊》中列出了在 IBM SPSS Modeler 中所使用建模方法的數學基礎說明，該手冊位於安裝磁碟的 *\Documentation* 目錄中。

另請注意，僅根據訓練資料得出這些結果。若要評量模型推廣到真實世界中的其他資料的程度，您可以使用「分割區」節點來送出一部分記錄用於測試和驗證。

第 19 章 零售銷售促銷（神經網路/C&RT）

此範例處理的資料會對零售產品線及促銷對銷售的影響進行說明。（此資料是虛構的。）您在此範例中的目標是預測未來銷售促銷的效果。與狀況監視範例類似，資料採礦處理程序由探索、資料準備、訓練及測試階段組成。

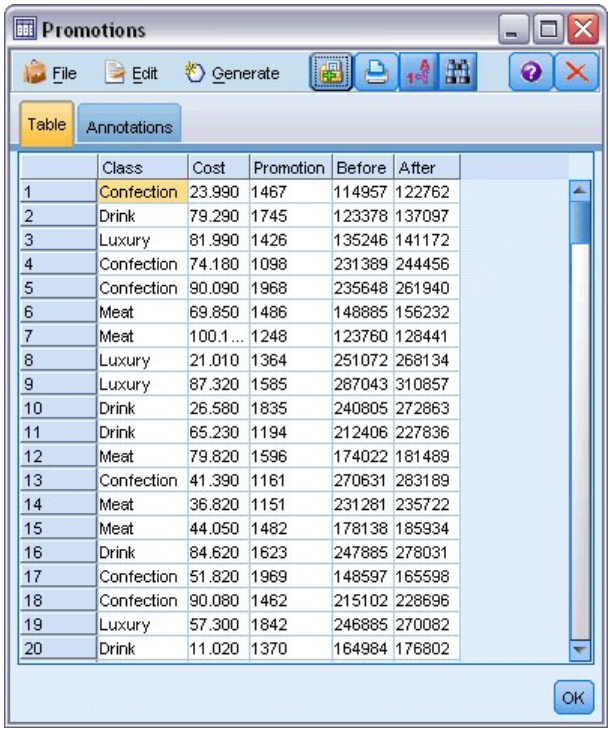
此範例使用名為 *goodsplot.str* 及 *goodslearn.str* 的串流，其參照的資料檔名為 *GOODS1n* 及 *GOODS2n*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。串流 *goodsplot.str* 位於 *streams* 資料夾中，而 *goodslearn.str* 檔案位於 *streams* 目錄中。

檢驗資料

每一個記錄都包含：

- 類別。產品類型。
- 成本。單價。
- 促銷。在特定促銷上花費的金額指標。
- 之前。促銷之前的營收。
- 之後。促銷之後的營收。

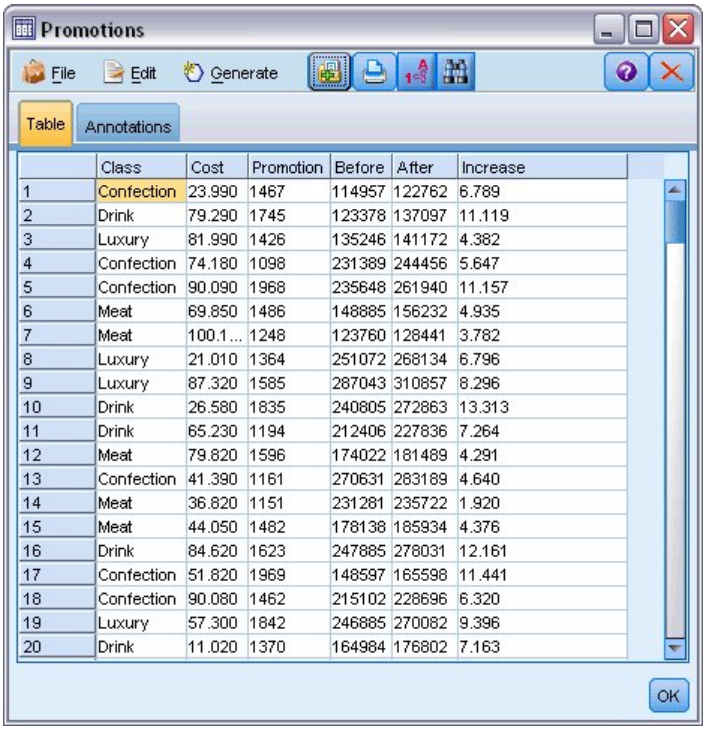
串流 *goodsplot.str* 包含一個簡單的串流，用來顯示表格中的資料。兩個營收欄位（之前和之後）以絕對項表示；但促銷之後的營收增長（可能是促銷的結果）似乎可能會是更有用的數字。



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

圖 *255.* 促銷對產品銷售的影響

goodsplot.str 還包含一個節點，用來在稱為增量的欄位中衍生此值（表示為促銷之前營收的百分比），並顯示一個表格來顯示此欄位。



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

圖 *256.* 促銷之後的營收增長

此外，該串流還會顯示針對促銷成本支出的增量直方圖及增量散佈圖，上面覆蓋了所包含產品的種類。

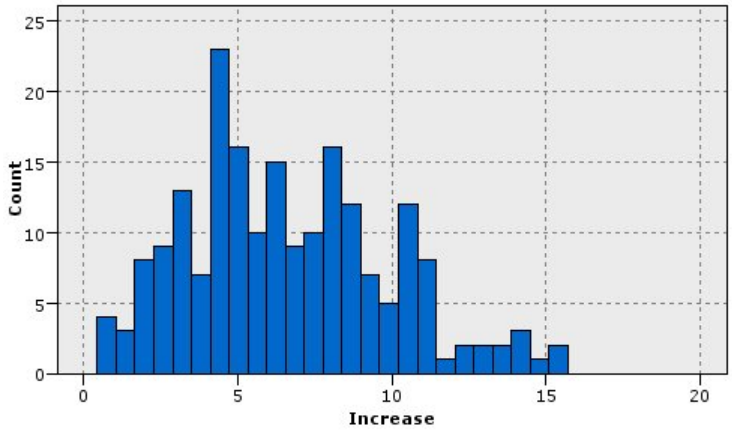


圖 *257.* 營收增量的直方圖

散佈圖針對每一個產品類別顯示增量，在營收增長與促銷成本之間存在幾乎為線性的關係。因此，決策樹狀結構或神經網路似乎可能以合理的精確度預測其他可用欄位中的營收增長。

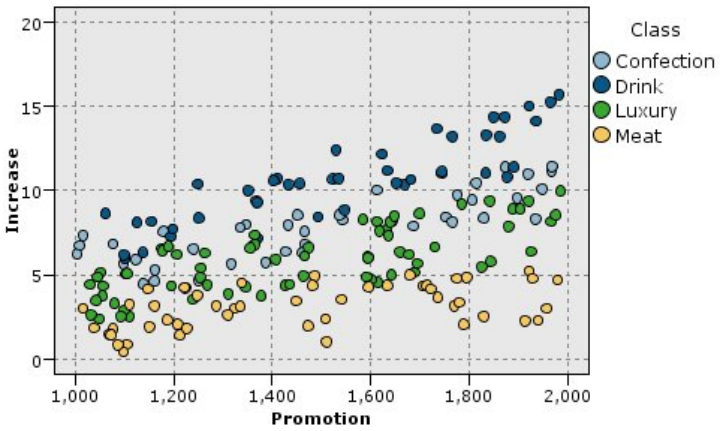


圖 *258.* 營收增長與促銷支出

學習和測試

串流 `goodslearn.str` 會訓練神經網路及決策樹狀結構來對營收增量進行此預測。

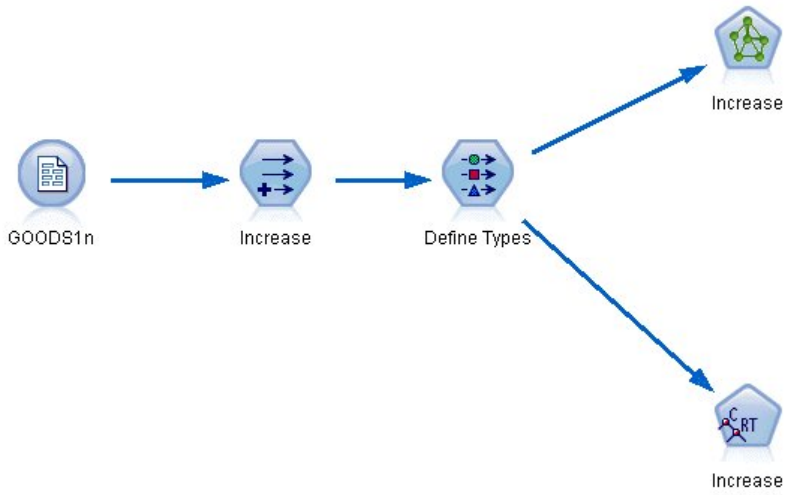


圖 *259.* 為串流 *goodslearn.str* 建模

執行模型節點並產生實際模型之後，您可以測試學習過程的結果。在「類型」節點與新的「分析」節點之間連接數列中的決策樹狀結構及網路，將輸入（資料）檔案變更為 `GOODS2n`，並執行「分析」節點，即可完成該測試。從這個節點的輸出中，特別是預測增長與正確答案之間的線性相關性中，您會發現受訓練系統預測營收增長的成功率較高。

進一步探索可以集中於受訓練系統犯了相對較大錯誤的觀察值上；透過繪製預測營收增長與實際增長可以識別這些錯誤。可以使用 SPSS Modeler 內的互動式圖形選取此圖形上的偏離值，並且可能可以透過其內容調整資料說明或學習過程來提高精確度。

第 20 章 狀況監視（神經網路/C5.0）

此範例涉及機器的監視狀態資訊以及識別和預測錯誤狀態的問題。資料透過虛構模擬建立，由一段時間內測量的多個連結數列組成。就下列各項而言，每一個記錄都是機器上的 Snapshot 報告：

- 時間。整數。
- 冪次。整數。
- 溫度。整數。
- 壓力。正常情況下為 0，1 代表瞬間壓力警告。
- 執行時間。自前次提供服務以來的時間。
- 狀態。通常是 0，發生錯誤時變更為錯誤碼（101、202 或 303）。
- 結果。在此時間數列中出現的錯誤碼，如果沒有錯誤發生，則為 0。（這些錯誤碼僅在事後提供。）

此範例使用名為 *condplot.str* 及 *condlearn.str* 的串流，其參照的資料檔名為 *COND1n* 及 *COND2n*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*condplot.str* 及 *condlearn.str* 檔案位於 *streams* 目錄中。

針對每一個時間數列，都有一系列記錄來自正常作業週期，然後來自導致錯誤的週期，如下表中所示：

時間(M)	電源	溫度	壓力	執行時間	狀態	結果
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

大部分資料採礦專案通常使用下列處理程序：

- 檢查資料來判斷哪些屬性可能與感興趣狀態的預測或識別相關。
- 保留那些屬性（如果已存在），如有必要，可以衍生那些屬性並將其新增至資料。
- 使用產生的資料來訓練規則及神經網路。
- 使用獨立測試資料測試受訓練系統。

檢驗資料

檔案 *condplot.str* 說明了處理程序的第一部分。它包含繪製多個圖形的串流。如果溫度或電源的時間數列包含可見型樣，則可以區分即將出現的錯誤條件，也可以預測這些條件的出現。針對溫度和電源，下面的串流在不同的圖形上繪製了與三個不同錯誤碼相關聯的時間數列，產生了六個圖形。「選取」節點會分隔與不同錯誤碼相關聯的資料。

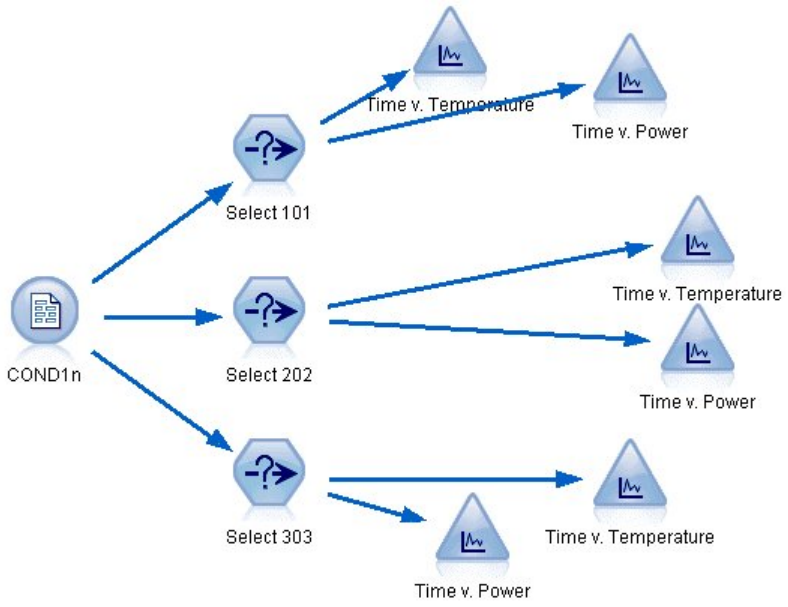


圖 *260. Condplot* 串流

此串流的結果會在此圖中顯示。

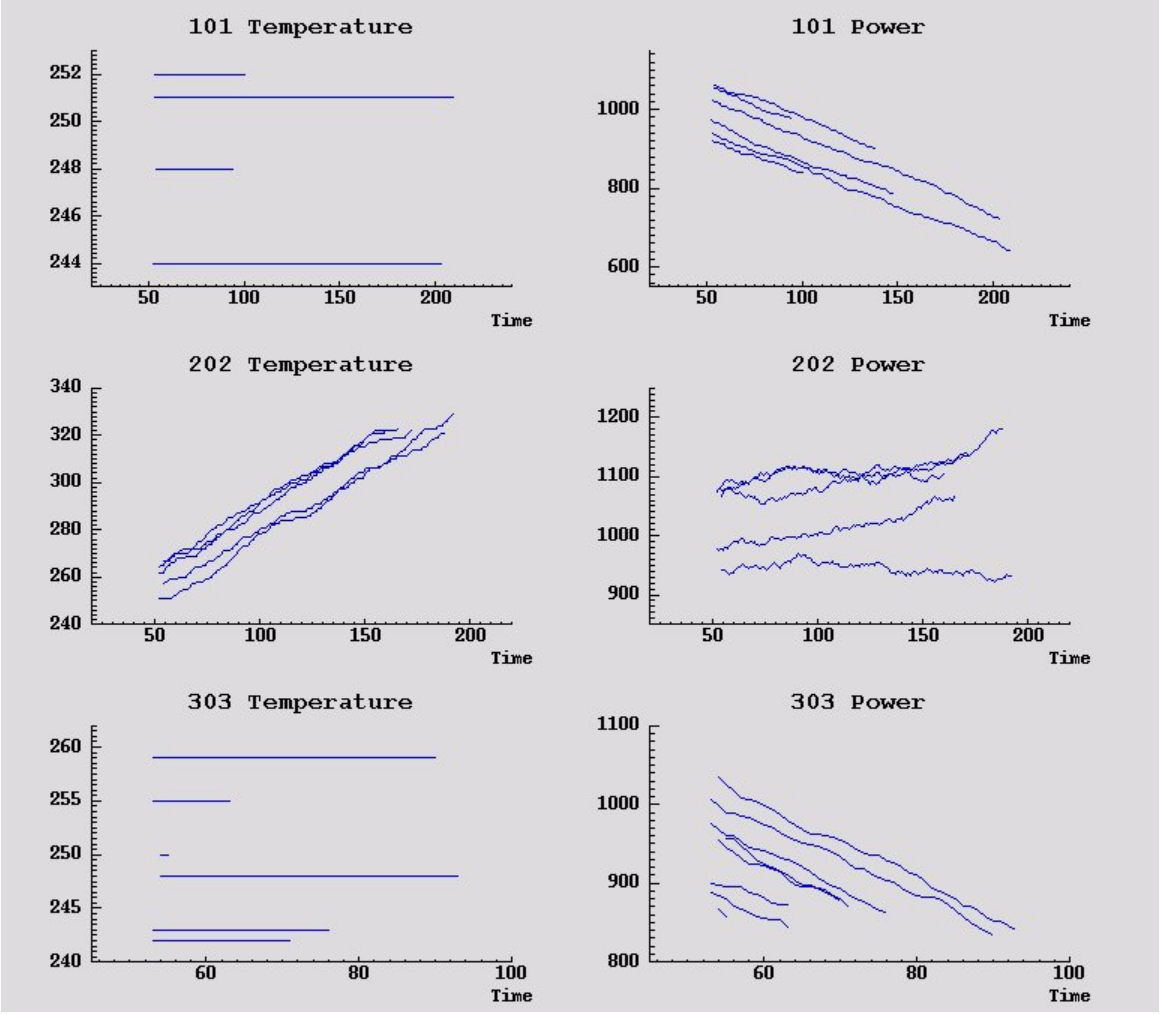


圖 *261.* 一段時間內的溫度和電源

這些圖形清晰地顯示了區別 202 錯誤與 101 及 303 錯誤的型樣。202 錯誤會顯示一段時間內溫度升高和電源波動；而其他錯誤卻不會。但是，區分 101 與 303 錯誤的型樣則不太清晰。這兩個錯誤都會顯示溫度穩定和電源下降，但 303 錯誤的電源下降似乎更急劇。

基於這些圖形，溫度和電源出現變更和變更率，以及出現波動和波動度似乎與預測和區別錯誤相關。因此應該先將這些屬性新增至資料，然後再套用學習系統。

資料預備

基於探索資料的結果，串流 *condlearn.str* 會衍生相關資料並學會預測錯誤。

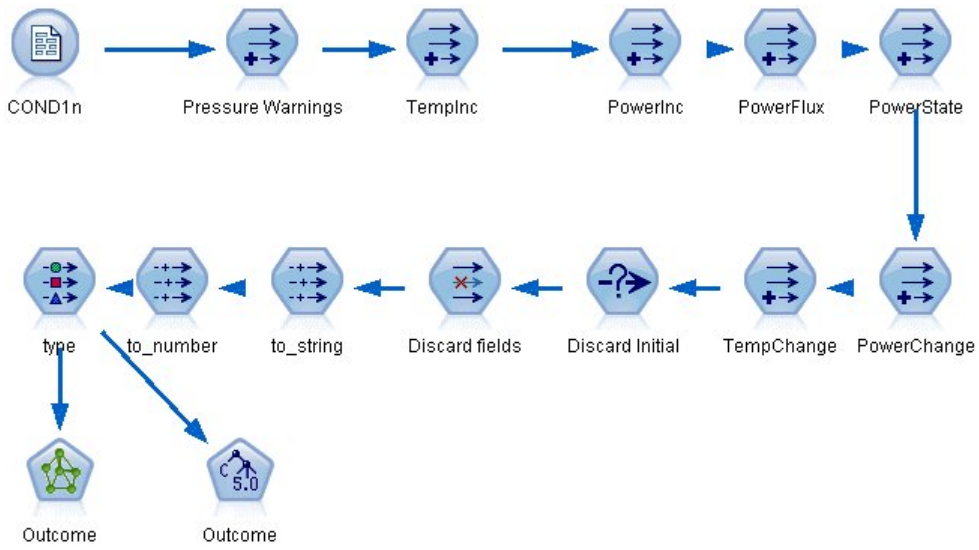


圖 *262. Condlearn* 串流

串流會使用多個「衍生」節點來準備用於建模的資料。

- **變數檔案節點**。讀取資料檔 *COND1n*。
- **衍生壓力警告**。計算瞬間壓力警告次數。當時間回到 0 時重設。
- **衍生 TempInc**。使用 `@DIFF1` 計算瞬間溫度變更率。
- **衍生 PowerInc**。使用 `@DIFF1` 計算瞬間電源變更率。
- **衍生 PowerFlux**。一個旗標，如果電源在最後一個記錄和這個記錄中按相反方向改變則為 true；即針對電源尖峰或低谷。
- **衍生 PowerState**。一種狀態，以穩定開始並在連續兩次偵測到不穩定電源時切換至波動。僅在五個時間間隔沒有不穩定電源或重設時間時切換回穩定。
- **PowerChange**。最後五個時間間隔內 *PowerInc* 的平均數。
- **TempChange**。最後五個時間間隔內 *TempInc* 的平均數。
- **捨棄起始（選取）**。捨棄每一個時間數列的第一個記錄來避免電源及溫度在邊界的大型（不正確）跳躍。
- **捨棄欄位**。將記錄減少為執行時間、狀態、結果、壓力警告、*PowerState*、*PowerChange* 及 *TempChange*。
- **類型**。將結果的角色定義為 **目標**（要預測的欄位）。此外，將結果的測量層次定義為**名義**，將壓力警告定義為**連續**，將 *PowerState* 定義為**旗標**。

教學

在 *condlearn.str* 中執行串流會訓練 C5.0 規則及神經網路。該網路可能需要花一段時間進行訓練，但可以提前岔斷訓練來儲存產生合理結果的網路。學習完成後，管理程式視窗右上角的「模型」標籤會閃動來警示您建立了兩個新區塊：一個代表神經網路，一個代表規則。

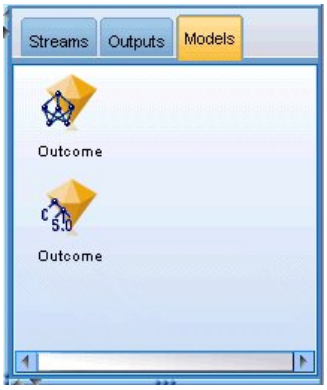


圖 *263.* 包含模型區塊的模型管理程式

模型區塊還會新增至現有串流，這樣我們就能夠進行系統測試或匯出模型的結果。在此範例中，我們將測試模型的結果。

測試

模型區塊會新增至串流，二者都會連接至「類型」節點。

1. 如圖所示重新定位區塊，使得「類型」節點連接至神經網路區塊，而該區塊連接至 C5.0 區塊。
2. 將「分析」節點連接至 C5.0 區塊。
3. 編輯原始來源節點來讀取檔案 *COND2n*（而非 *COND1n*），因為 *COND2n* 包含看不見的測試資料。

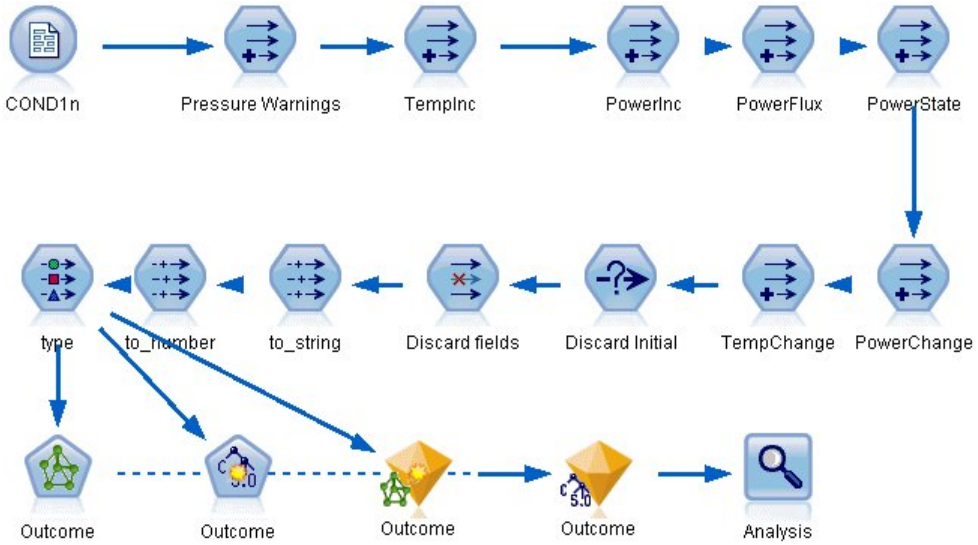


圖 *264.* 測試受訓練網路

4. 開啟「分析」節點並按一下「執行」。

這樣做會產生一些數字，可反映受訓練網路及規則的精確度。

第 21 章 分類電信客戶（區別分析）

區別分析是一種統計技術，它可根據輸入欄位的值對記錄進行分類。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。

例如，假設某電信公司根據服務使用型樣來切割客戶群，將客戶分成四組。如果可以使用人口資料來預測群組成員資格，則可以針對個別潛在客戶自訂報價。

此範例使用名為 *telco_custcat_discriminant.str* 的串流，其參照的資料檔名為 *telco.sav*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*telco_custcat_discriminant.str* 檔案位於 *streams* 目錄中。

該範例的重點集中在使用人口資料來預測使用型樣。目標欄位 *custcat* 有四個可能的值，對應於四組客戶，如下所示：

數值	標籤
1	基本服務
2	電子服務
3	附加服務
4	總服務

建立串流

1. 首先，設定串流內容以在輸出中顯示變數和值標籤。從功能表中選擇：

   檔案 > **串流內容...** > **選項** > **一般**
2. 請確保選取了**在輸出中顯示欄位和值標籤**，然後按一下**確定**。

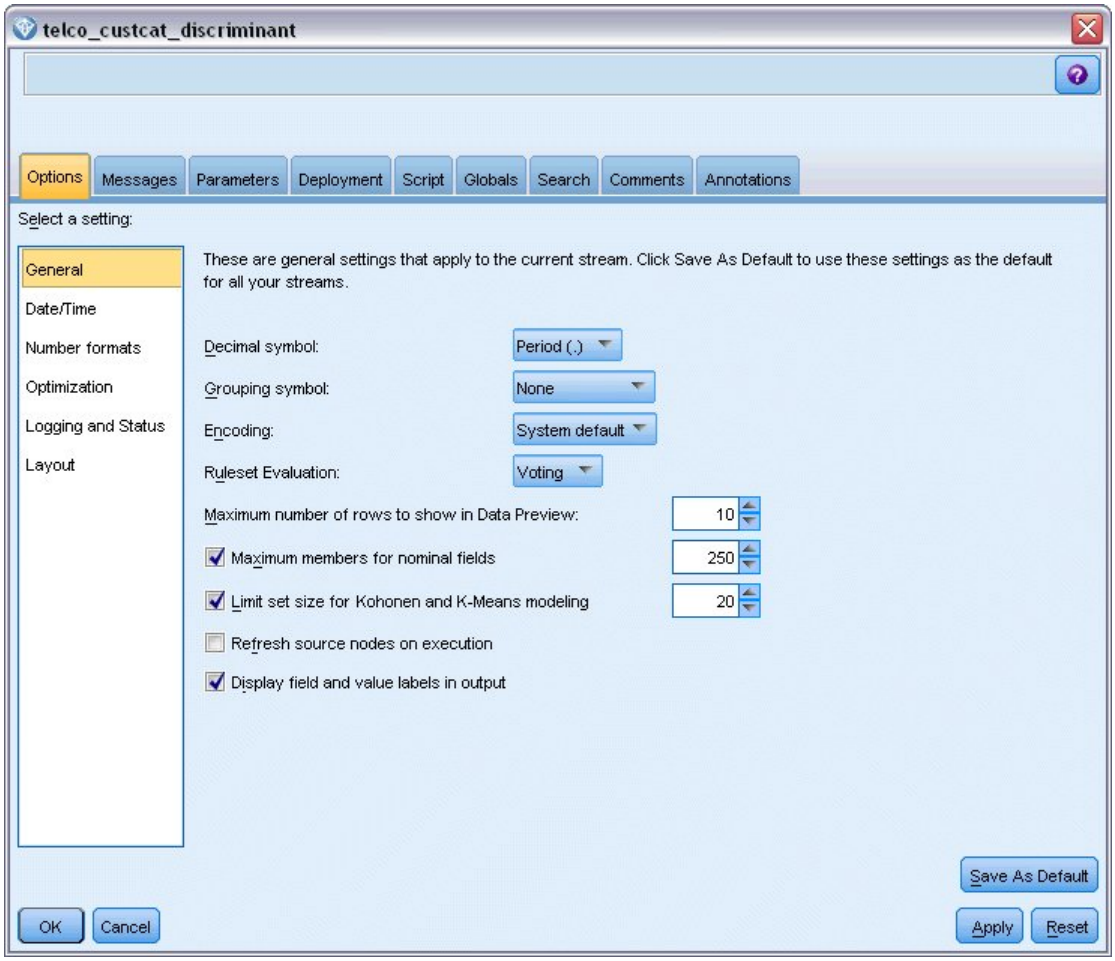


圖 *265.* 串流內容

3. 新增指向 *Demos* 資料夾中 *telco.sav* 的「統計量檔案」來源節點。

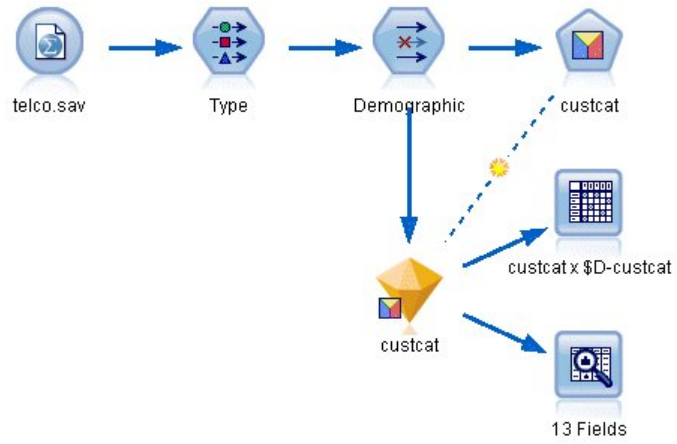


圖 *266.* 使用區別元件分析來分類客戶的串流範例

a. 新增「類型」節點並按一下**讀取值**，確保已正確設定所有測量層次。例如，值為 0 和 1 的大部分欄位可以視為旗標。

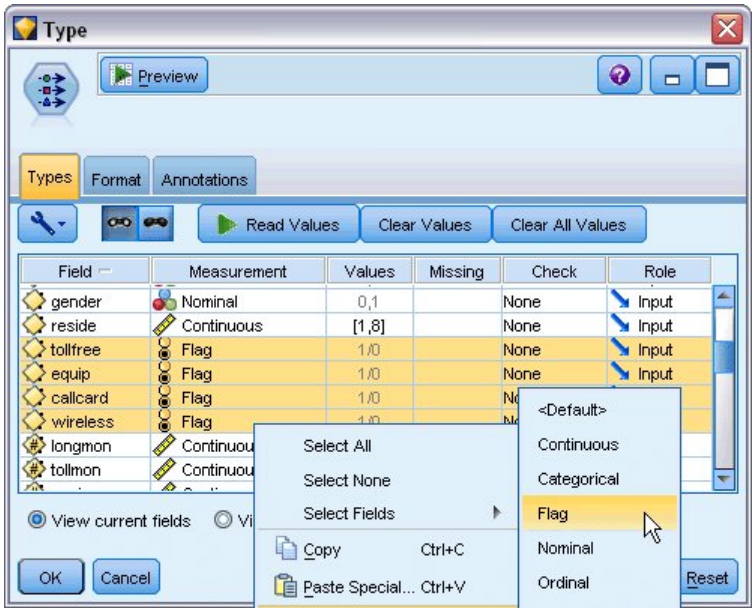


圖 *267.* 設定多個欄位的測量層次

提示：若要變更具有類似值（例如 0/1）的多個欄位的內容，請按一下值 直欄標頭以依值排序欄位，然後在使用滑鼠或方向鍵的同時按住 shift 鍵，以選取您要變更的所有欄位。然後，您可以在選項上按一下滑鼠右鍵以變更所選欄位的測量層次或其他屬性。

請注意，將 *gender* 視為具有一組值（包含兩個值）的欄位而不是一個旗標更為正確，因此請將其測量值保留為**名義**。

b. 將 *custcat* 欄位的角色設為**目標**。所有其他欄位應該將其角色設為**輸入**。

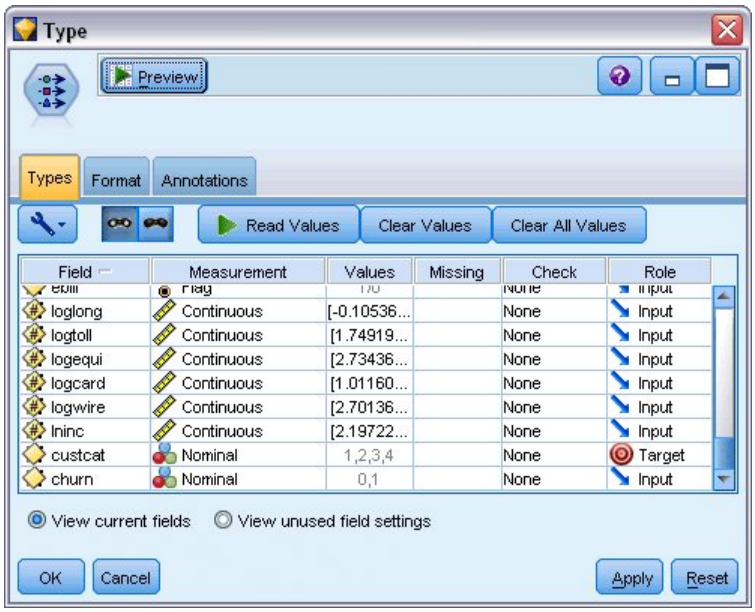


圖 *268.* 設定欄位角色

由於此範例的焦點是個人背景資訊，因此使用「過濾器」節點可以只包括相關欄位（*region*、*age*、*marital*、*address*、*income*、*ed*、*employ*、*retire*、*gender*、*reside* 和 *custcat*）。可以排除其他欄位以便進行此分析。

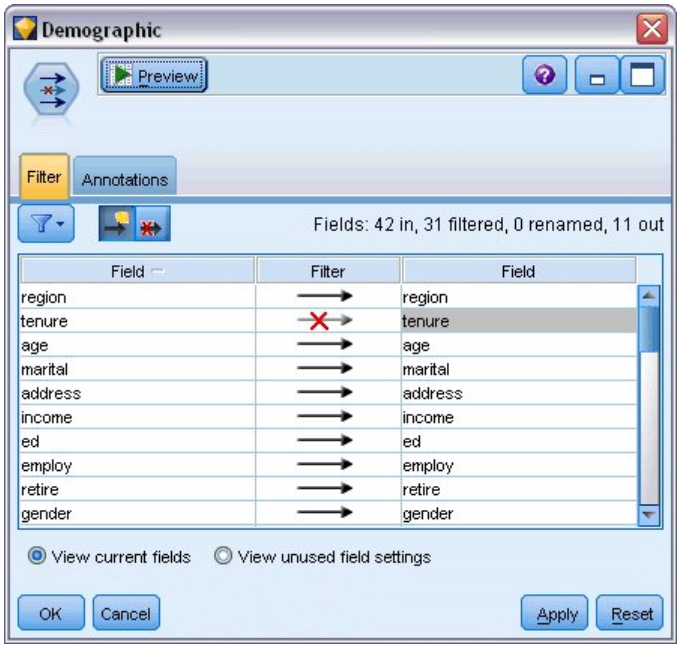


圖 *269.* 根據個人背景資訊欄位過濾

（此外，您可以針對這些欄位將角色變更為**無**而非排除他們，或是選取您要用在建模節點中的欄位。）

4. 在「區別元件」節點中，按一下「模型」標籤並選取**逐步**方法。

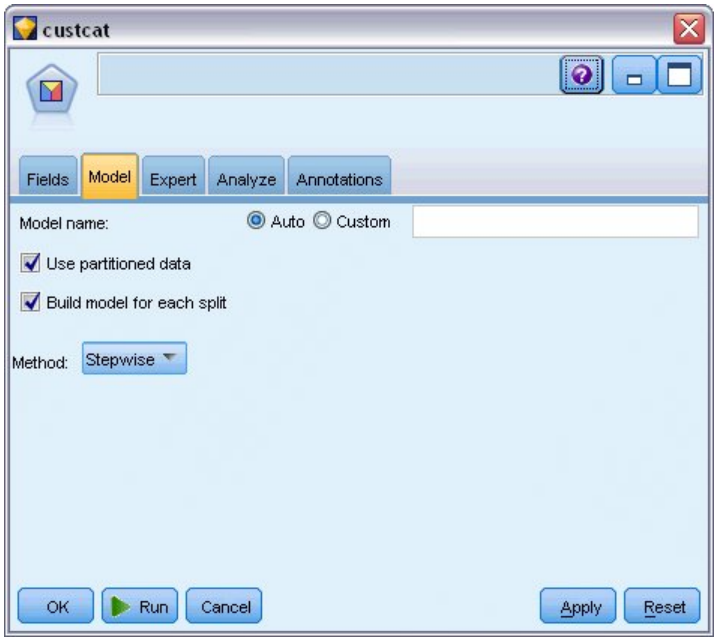


圖 *270.* 選擇模型選項

5. 在「專家」標籤上將模式設為**專家**，然後按一下**輸出**。
6. 在「進階輸出」對話框中選取**摘要表格**、**地域圖**和**步驟摘要**，然後按一下**確定**。

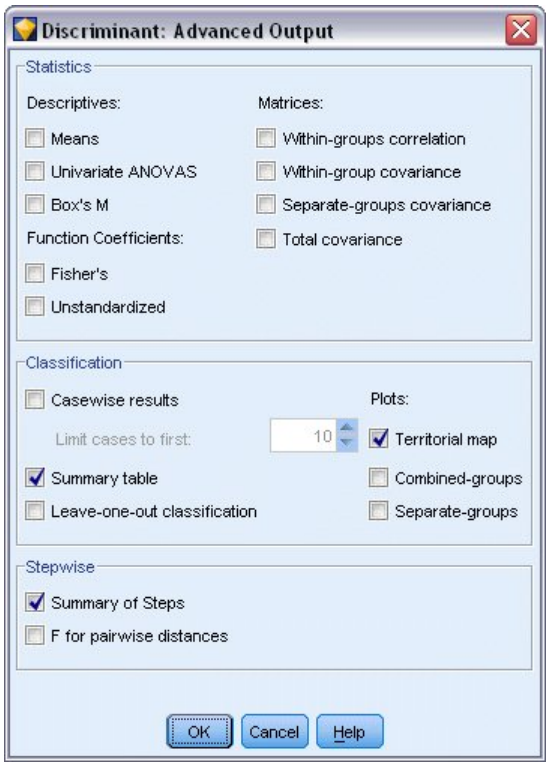


圖 *271.* 選擇輸出選項

檢查模型

1. 按一下**執行**以建立模型，該模型會新增至串流以及右上角的「模型」選用區中。若要檢視其詳細資料，請按兩下串流中的模型區塊。

   「摘要」標籤顯示（其他內容）以及目標和已提交進行考量之輸入（預測工具欄位）的完整清單。

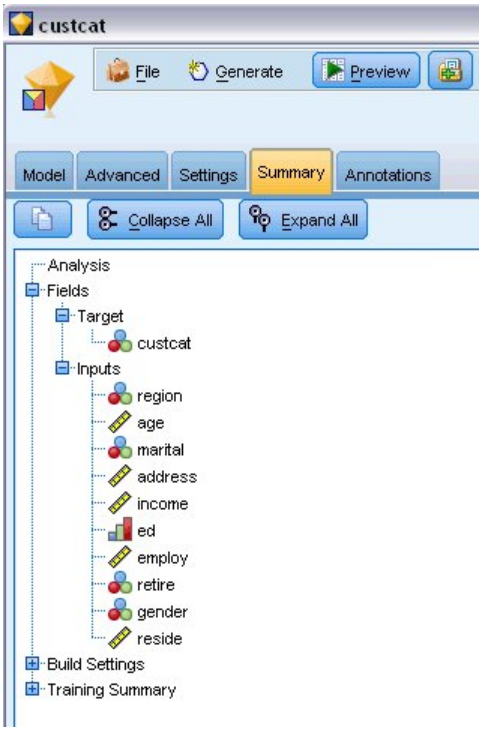


圖 272. 顯示目標和輸入欄位的模型摘要

如需區別元件分析結果的詳細資料，請執行下列作業：

2. 按一下「進階」標籤。
3. 按一下「在外部瀏覽器中啟動」按鈕（位於「模型」標籤下）以在 Web 瀏覽器中檢視結果。

分析使用區別分析分類電信客戶的輸出

逐步迴歸分析區別分析

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	Age in years	1.000	1.000	7.521	.978
	Marital status	1.000	1.000	3.500	.990
	Years at current address	1.000	1.000	8.433	.975
	Household income in thousands	1.000	1.000	6.689	.980
	Level of education	1.000	1.000	61.454	.844
	Years with current employer	1.000	1.000	16.976	.951
	Retired	1.000	1.000	3.005	.991
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988
1	Age in years	.980	.980	6.125	.829
	Marital status	.999	.999	3.803	.834
	Years at current address	.983	.983	8.487	.823
	Household income in thousands	.989	.989	6.022	.829
	Years with current employer	.953	.953	14.933	.807
	Retired	.992	.992	1.432	.840
	Gender	1.000	1.000	.358	.843
	Number of people in household	1.000	1.000	3.967	.834
2	Age in years	.563	.548	.352	.807
	Marital status	.999	.952	3.903	.798
	Years at current address	.798	.773	2.913	.800
	Household income in thousands	.689	.664	.634	.806
	Retired	.927	.891	.528	.806
	Gender	.998	.951	.391	.807
	Number of people in household	.979	.934	4.841	.796
3	Age in years	.535	.535	.252	.795
	Marital status	.605	.593	1.507	.792
	Years at current address	.776	.771	3.514	.787
	Household income in thousands	.688	.657	.687	.794
	Retired	.917	.880	.353	.795
	Gender	.997	.931	.395	.795

圖 *273.* 不在分析中的變數

當您有大量預測值時，可以使用逐步迴歸分析法自動選取要在模型中使用的「最佳」變數。逐步迴歸分析法從不包括任何預測值的模型開始。在每一個步驟，具有超過輸入準則（依預設為 3.84）的最大 *F* 輸入值的預測值會新增至模型。

在最後一個步驟離開分析之變數的 *F* 輸入值全都小於 3.84，因此不會再新增。

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Level of education	1.000	61.454	
2	Level of education	.953	59.108	.951
	Years with current employer	.953	14.933	.844
3	Level of education	.951	60.046	.940
	Years with current employer	.934	15.824	.834
	Number of people in household	.979	4.841	.807

圖 *274.* 分析中的變數

此表格會顯示在每一個步驟分析中變數的統計量。容差是未由方程式中其他自變數說明之變數變異的比例。容差非常小的變數幾乎不會對模型貢獻資訊，而且會造成計算上的問題。

若從目前模型中移除變數（假定其他變數仍然存在），則 *F* 移除值對於情況說明會非常有用。用於輸入變數的 *F* 移除與上一步的 *F* 輸入相同（在「不在分析中的變數」表格中顯示）。

逐步迴歸分析法的相關注意事項

逐步迴歸分析法很方便，但有其限制。請注意，因為逐步迴歸分析僅基於統計價值選取模型，所以可能會選擇沒有實際顯著性的預測值。如果您有一些資料經驗並能夠預期重要的預測值，則應該使用該知識並避免使用逐步迴歸分析法。但如果您有多個預測值，不知道從哪裡入手，則執行逐步分析並調整選取的模型比完全沒有模型要好。

檢查模型適合度

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198[a]	80.2	80.2	.407
2	.048[a]	19.4	99.6	.214
3	.001[a]	.4	100.0	.031

a. First 3 canonical discriminant functions were used in the analysis.

圖 *275.* 特徵值

幾乎模型的所有變異都由前兩個區別函數解釋。有三個函數自動適合，但由於特徵值極小，您可以相當安全地忽略第三個函數。

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

圖 *276. Wilks' Lambda (λ)* 值*(W)*

Wilks' lambda 同意僅前兩個函數有用。針對每一組函數，這會對所列函數平均數在各個群組之間相等的假設進行測試。函數 3 測試的顯著性值大於 0.10，因此這個函數對模型幾乎沒有貢獻。

結構矩陣

Structure Matrix

	Function		
	1	2	3
Level of education	.966*	-.090	-.244
Years with current employer	-.182	.964*	-.193
Age in years[b]	-.162	.598*	-.285
Household income in thousands[b]	.109	.514*	-.190
Years at current address[b]	-.151	.394*	-.214
Retired[b]	-.108	.230*	-.137
Gender[b]	.008	.054*	.009
Number of people in household	.232	.097	.968*
Marital status[b]	.132	.134	.600*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

b. This variable not used in the analysis.

圖 *277.* 結構矩陣

有多個區別函數時，星號 (*) 會標示每一個變數與其中一個標準函數的最大絕對相關性。然後，在每一個函數內，這些標示的變數會按相關性大小進行排序。

- 教育程度與第一個函數的相關性最強，並且它是唯一一個與此函數相關性最強的變數。
- 現任雇主在職年數、年齡（歲）、家庭收入（千元）、現址居住年數、已退休及性別與第二個函數相關性最強，盡管性別和已退休比其他變數的相關性要弱一些。其他變數將此函數標示為「穩定性」函數。
- 家庭人數及婚姻狀態與第三個區別函數的相關性最強，但這是無用的函數，因此這些預測值幾乎是無用的。

地域圖

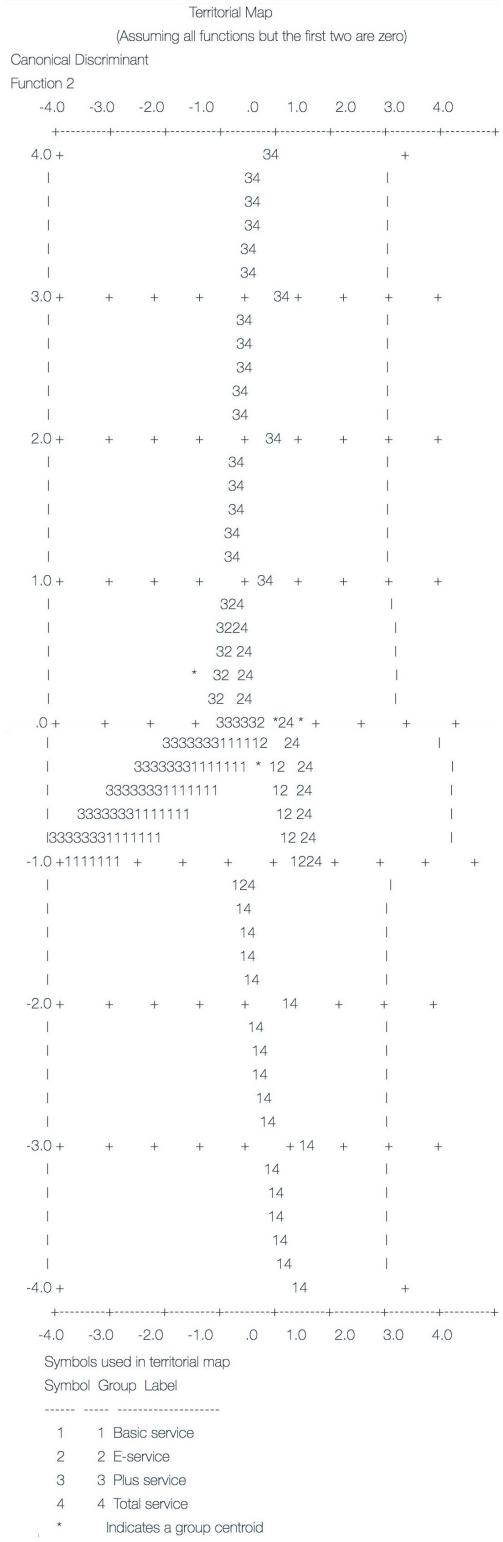


圖 *278.* 地域圖

地域圖可協助您研究群組與區別函數之間的關係。與結構矩陣結果結合，它可對預測值與群組之間的關係進行圖形解譯。水平軸上顯示的第一個函數將第 4 組（總服務客戶）與其他群組分離。由於教育程度與第一個函數

強烈正相關，這表示您的總服務客戶一般受教育程度最高。第二個函數將第 1 組和第 3 組（基本服務和附加服務客戶）分離。附加服務客戶通常比基本服務客戶參加工作的時間更長，年齡也更大一些。電子服務客戶並未與其他客戶很好地分離，盡管地域圖顯示他們往往受過良好的教育且具有一定的工作經驗。

一般情況下，用星號 (*) 標示的群組重心與地域線的接近度暗示著所有群組之間的分離不是很強。

僅繪製前兩個區別函數，但由於大家認為第三個函數太不顯著，地域圖提供了區別函數的綜合性視圖。

分類結果

Classification Results[a]

			Predicted Group Membership			
	Customer category	Basic service	E-service	Plus service	Total service	Total
Original	Count	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47.0	4.1	22.9	25.9	100.0
		E-service	22.6	6.9	26.7	43.8	100.0
		Plus service	36.3	5.0	39.9	18.9	100.0
		Total service	16.9	6.8	15.7	60.6	100.0

a. 39.5% of original grouped cases correctly classified.

圖 *279.* 分類結果

您透過 Wilks' lambda 知道模型比猜測效果好，但必須轉到分類結果才能判定有多好。假設提供觀察資料，則「虛無」模型（亦即，沒有預測值的模型）會將所有客戶分類至限制模式的群組附加服務。因此，虛無模型會是正確的 281/1000 = 28.1%。您的模型會獲取額外 11.4% 的客戶，或者說獲取 39.5% 的客戶。事實上，您的模型的優點在於可以識別出總服務客戶。但是，它在分類電子服務客戶上的效果非常差。您可能需要找到另一個預測值，才能區分這些客戶。

摘要

您已建立區別模型，可根據每一個客戶的人口資訊將客戶分為四個預先定義的「服務使用」群組。使用結構矩陣及地域圖，您可識別哪些變數對切割客戶群最為有用。最後，分類結果顯示模型在分類電子服務客戶方面做得很差。需要進行更多的研究來判定另一個預測值變數，以便更好地分類這些客戶，但視您想要預測的內容而定，該模型可能會完全滿足您的需要。例如，如果您對識別電子服務客戶並不關心，則該模型對您來說是足夠精確的。在這種情況下，電子服務可能是低價促銷，幾乎不會帶來利潤。如果，例如，您的最高投資報酬率來自附加服務或總服務客戶，則該模型可能會為您提供需要的資訊。

另請注意，僅根據訓練資料得出這些結果。若要評量模型推廣到其他資料的程度，您可以使用「分割區」節點來送出一部分記錄用於測試和驗證。

《IBM SPSS Modeler 演算法手冊》中列出了在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明。可從安裝磁碟的 *\Documentation* 目錄取得。

第 22 章 分析區間受限存活資料（廣義線性模型）

分析區間受限的存活資料時 — 亦即，當感興趣事件的確切時間不明但僅知道在給定區間內發生時 — 則按區間將 Cox 模型套用至事件風險會產生互補對數存活函數的對數迴歸模型。

旨在比較阻止潰瘍復發之兩種療法功效的研究中的局部資訊收集在 *ulcer_recurrence.sav* 中。此資料集已在其他位置 [1] 呈現及分析。使用廣義線性模型，您可以抄寫互補對數存活函數的對數迴歸模型的結果。

此範例使用名為 *ulcer_genlin.str* 的串流，其參照資料檔 *ulcer_recurrence.sav*。資料檔位於 *Demos* 資料夾中，串流檔位於 *streams* 子資料夾中。

建立串流

1. 新增指向 *Demos* 資料夾中 *ulcer_recurrence.sav* 的「統計量檔案」來源節點。

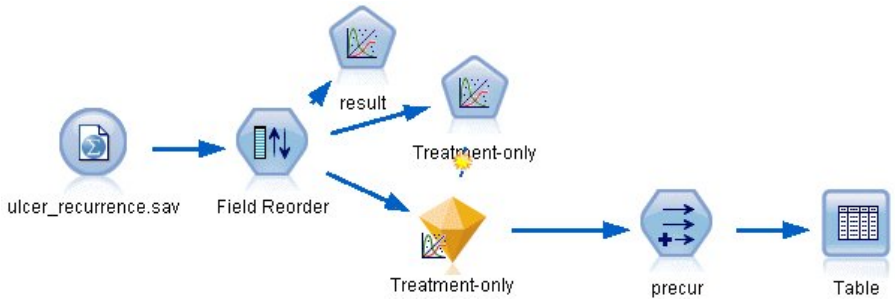


圖 *280.* 用來預測潰瘍復發的樣本串流

2. 在來源節點的「過濾器」標籤上，過濾出 *id* 及 *time*。

1. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

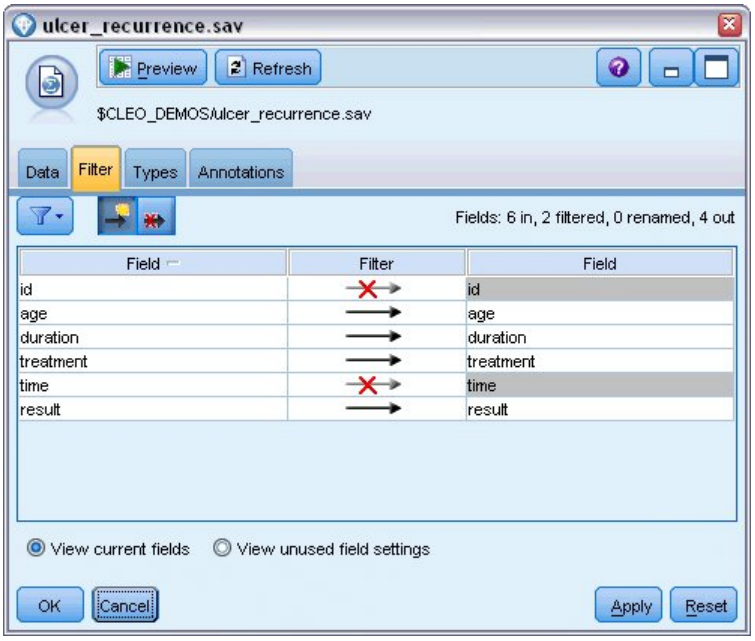


圖 *281.* 過濾不需要的欄位

3. 在來源節點的「類型」標籤上，將 *result* 欄位的角色設定為**目標**，並將其測量層次設定為**旗標**。結果為 1 表示潰瘍復發。所有其他欄位應該將其角色設為**輸入**。
4. 按一下**讀取值**以將資料實例化。

圖 *282.* 設定欄位角色

5. 新增「欄位重新排序」節點並指定 *duration*、*treatment* 及 *age* 作為輸入順序。這可判定在模型中輸入欄位的順序，並協助您嘗試抄寫 Collett 的結果。

圖 *283.* 將欄位重新排序，以便根據需要將其輸入模型

6. 將 GenLin 節點連接至來源節點；在 GenLin 節點上，按一下**模型**標籤。
7. 選取**第一個（最低）**作為目標的參照種類。這表示第二個種類是感興趣事件，它對模型的影響存在於對參數估計值的解譯中。如果連續預測值的係數為正，則指出復發機率隨預測值的增加而增長；如果種類的名義預測值係數較大，則指出復發機率相對於其他集種類增加。

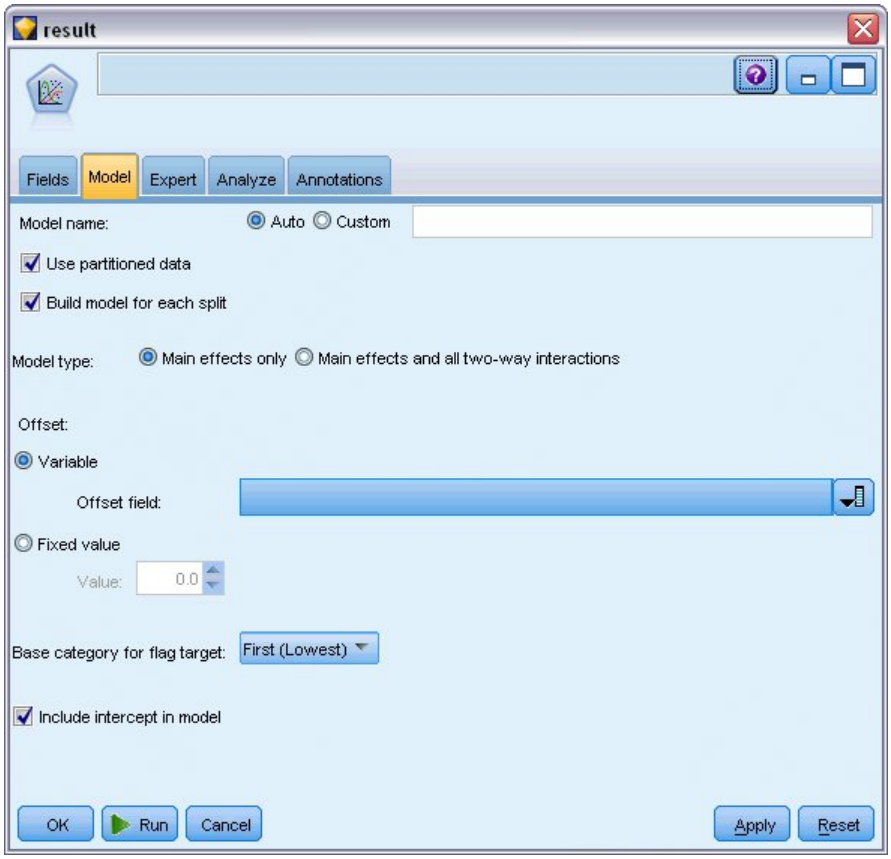


圖 *284.* 選擇模型選項

8. 按一下**專家**標籤並選取**專家**以啟動專家建模選項。
9. 選取**二項式**作為分佈，選取**互補對數存活函數的對數**作為鏈結函數。
10. 選取 **固定值**作為估計尺度參數的方法，並保留預設值 1.0。
11. 選取**遞減**作為因素的種類順序。這表示每一個因素的第一個種類是其參照種類；此選項對模型的影響存在於對參數估計值的解譯中。

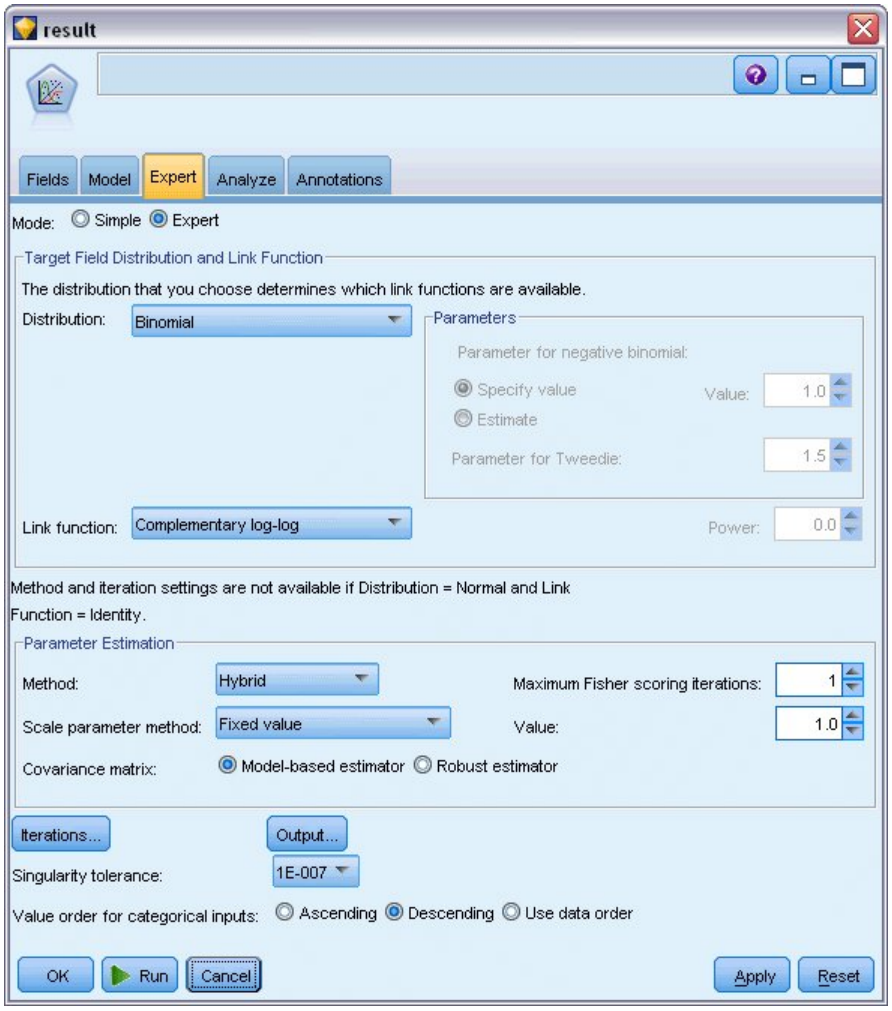


圖 *285.* 選擇專家選項

12. 執行串流以建立模型區塊，該區塊會新增至串流畫布，還會新增至右上角的「模型」選用區。若要檢視模型詳細資料，請用滑鼠右鍵按一下區塊並選擇**編輯**或**瀏覽**。

模型效應的檢定

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
Age in years	.358	1	.550
Duration of disease	.003	1	.958
Treatment group	.382	1	.537

Dependent Variable: Result

Model: (Intercept), Age in years, Duration of disease, Treatment group

圖 *286.* 主效應模型的模型效應檢定

所有模型效應都不具有統計顯著性；但治療效果中可觀察的任何差異都具有臨床意義，因此我們將僅使用治療方案作為模型項來配適縮減的模型。

配適僅含治療方案的模型

1. 在 GenLin 節點的「欄位」標籤上，按一下**使用自訂設定**。
2. 選取 *result* 作為目標。
3. 選取 *treatment* 作為唯一輸入。

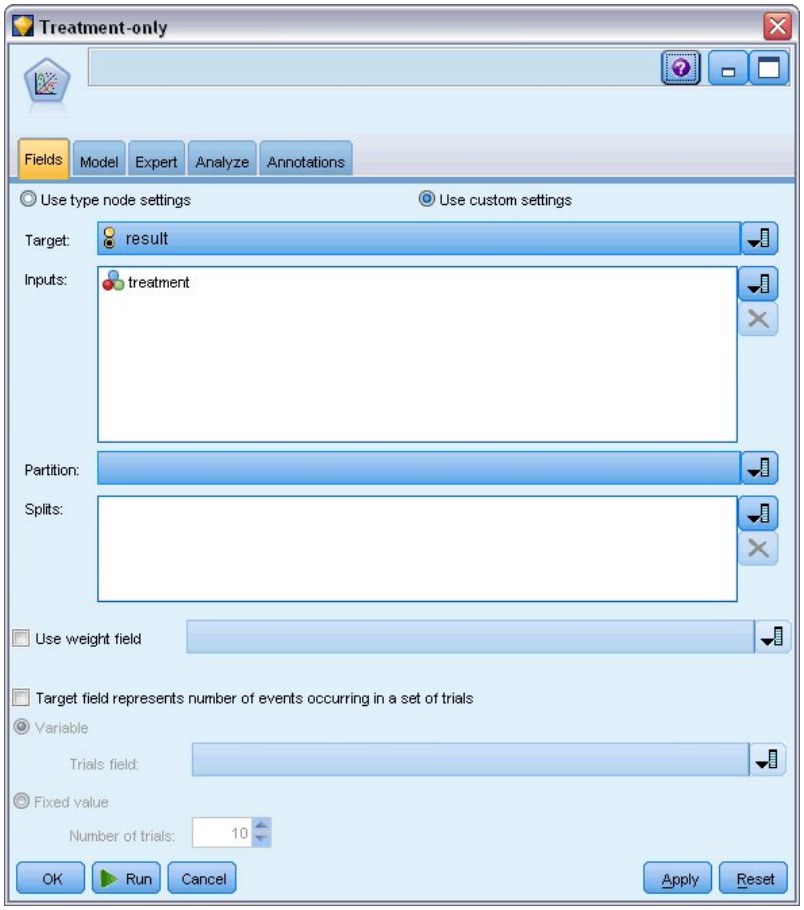


圖 *287.* 選擇欄位選項

4. 執行串流並開啟產生的模型區塊。

在模型區塊上，選取**進階**標籤並捲動至底部。

參數估計值

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[Treatment group=1]	.378	.6288	-.855	1.610	.361	1	.548
[Treatment group=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result
Model: (Intercept), Treatment group

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

圖 *288.* 僅含治療方案之模型的參數估計值

治療效果（兩個治療層次之間線性預測值的差異；亦即 *[treatment=1]* 的係數）仍不具有統計顯著性，但只能表示治療方案 *A [treatment=0]* 可能優於 *B [treatment=1]*，因為治療 *B* 的參數估計值大於 *A*，因此與前 12 個月中的復發機率增長相關聯。線性預測值（截距 + 治療效果）是 $\log(-\log(1-P(recur_{12,t}))$ 的估計值，其中 $P(recur_{12,\ t})$ 是治療方案 t（=*A* 或 *B*）在 12 個月時的復發機率。這些預測機率針對資料集中的每一個觀察產生。

預測的復發及存活機率

圖 *289.* 衍生節點設定選項

1. 針對每一個病患，模型會對預測結果及該預測結果的機率進行評分。為了查看預測的復發機率，請將產生的模型複製到選用區並連接「衍生」節點。
2. 在「設定」標籤中，鍵入 `precur` 作為衍生欄位。
3. 選擇將其衍生為**條件式**。
4. 按一下計算機按鈕以開啟 **If** 條件的表示式建置器。

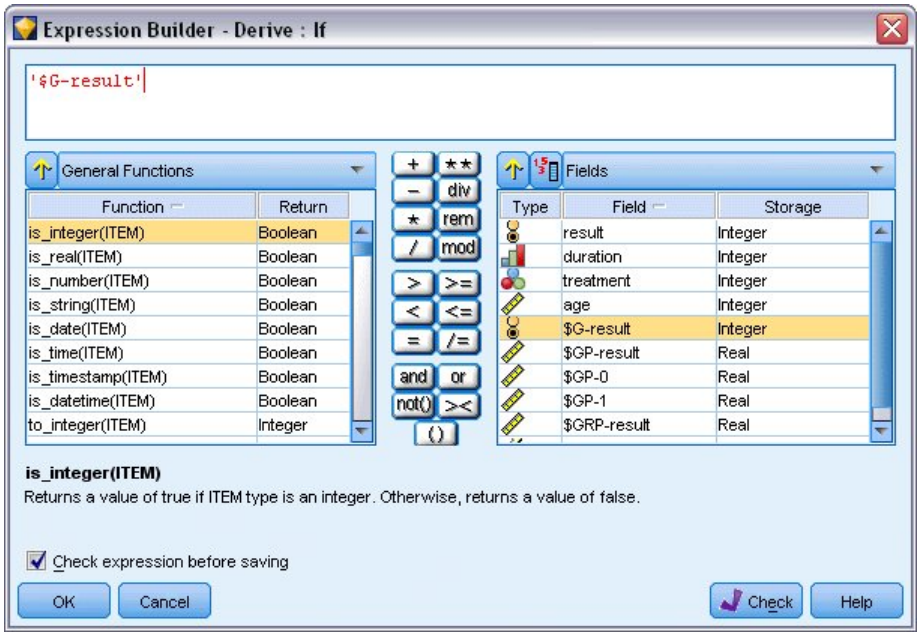


圖 *290.* 衍生節點：*If* 條件的表示式建置器

5. 將 *$G-result* 欄位插入表示式。
6. 按一下**確定**。

當 *$G-result* 等於 1 時，衍生欄位 *precur* 會採取 **Then** 表示式的值，當它為 0 時，則會採取 **Else** 表示式的值。

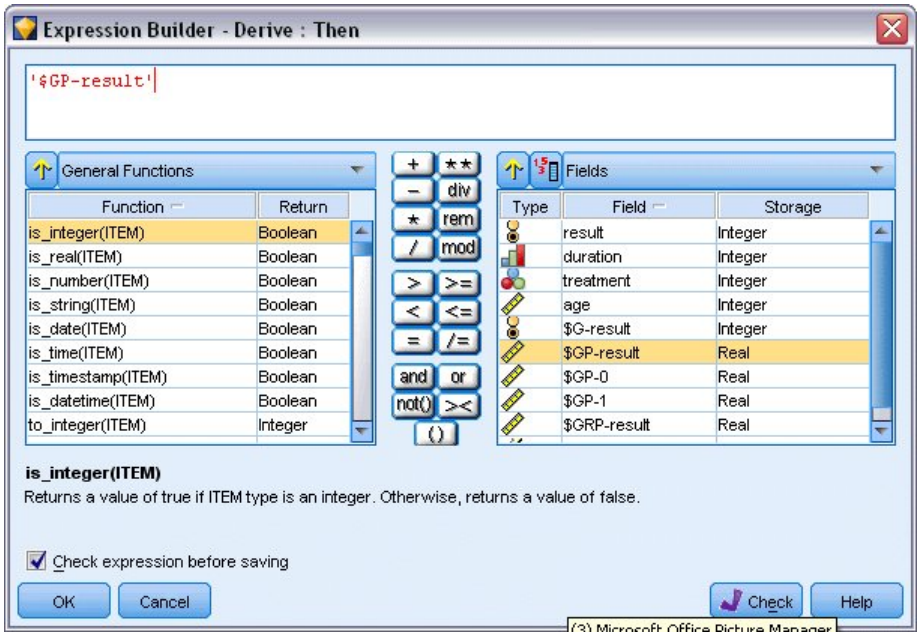


圖 *291.* 衍生節點：*Then* 表示式的表示式建置器

7. 按一下計算機按鈕以開啟 **Then** 表示式的表示式建置器。
8. 將 *$GP-result* 欄位插入表示式。

9. 按一下「**確定**」。

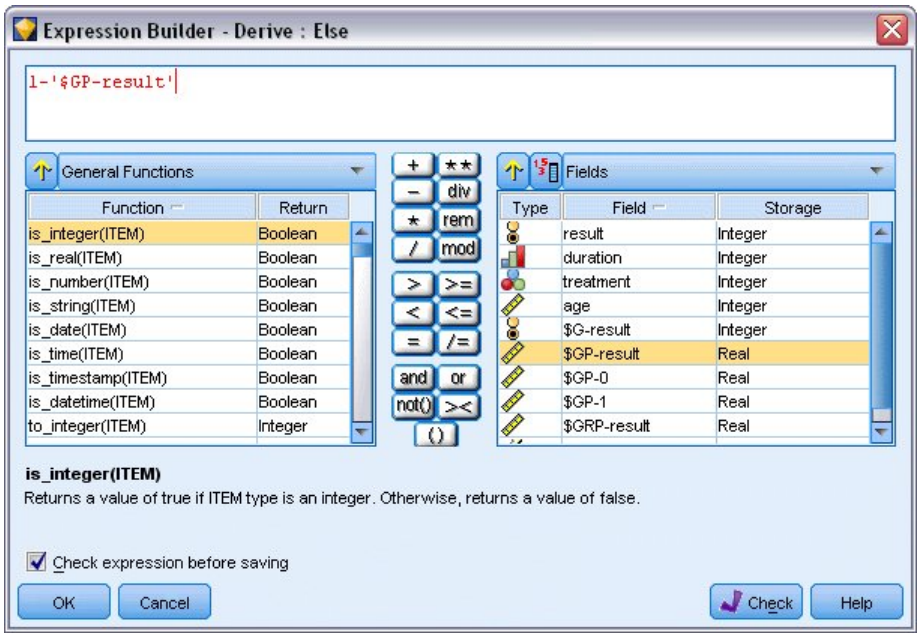
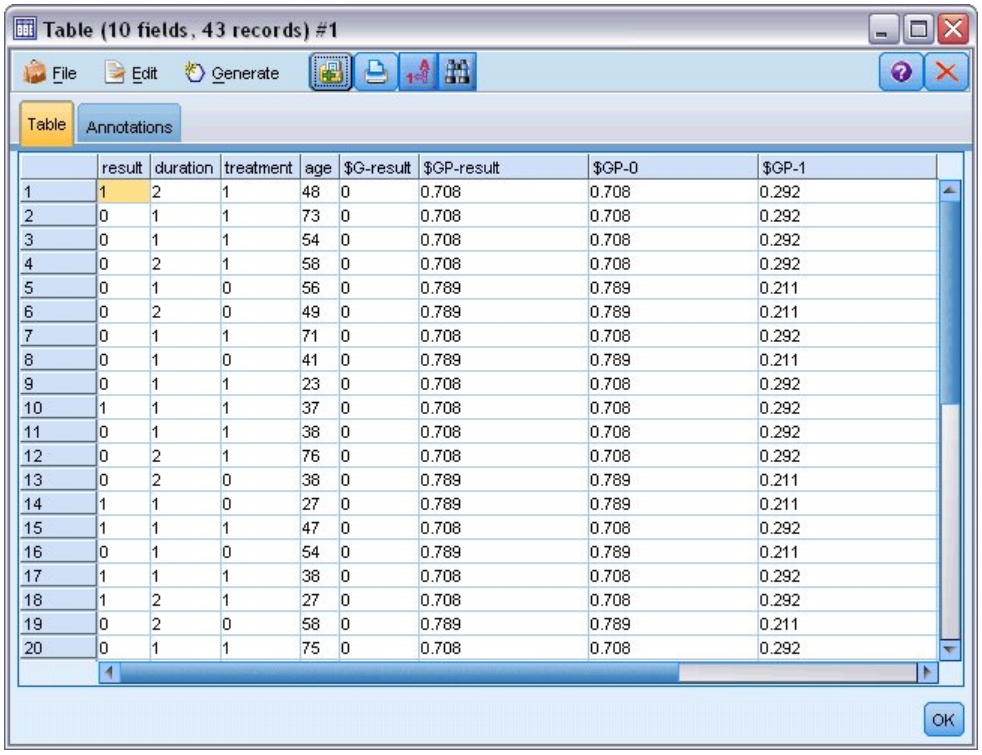


圖 *292.* 衍生節點：*Else* 表示式的表示式建置器

10. 按一下計算機按鈕以開啟 **Else** 表示式的表示式建置器。
11. 在表示式中鍵入 1-，然後將 *$GP-result* 欄位插入表示式。
12. 按一下「**確定**」。

圖 *293.* 衍生節點設定選項

13. 將表格節點連接至「衍生」節點並執行它。



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

圖 *294.* 預測機率

指派給治療方案 *A* 的病患在前 12 個月中經歷復發的機率估計值為 0.211；治療方案 *B* 則為 0.292。請注意，$1–P(recur_{12,\ t})$ 是 12 個月時的存活機率，存活分析師可能對此更感興趣。

依週期為復發機率建模

模型目前的問題是它會忽略在第一次檢查時收集的資訊；亦即，很多病患在前六個月未經歷過復發。「更好的」模型會為記錄事件是否會在每一個區間發生的二進位回應建模。配適此模型需要重新建構可在 *ulcer_recurrence_recoded.sav* 中找到的原始資料集。 此檔案包含另外兩個變數：

- 週期，它會記錄觀察值是對應於第一個檢查期還是第二個。
- 依週期排序的結果，它會記錄給定病患在給定週期內是否復發。

每一個原始觀察值（病患）每個區間貢獻一個觀察值，在此區間，它仍然處於風險集中。因此，例如，病患 1 貢獻兩個觀察值；一個觀察值對應於沒出現復發的第一個檢查期，一個對應於記錄了復發的第二個檢查期。而病患 10 則貢獻單一觀察值，因為在第一個週期中記錄了復發。病患 16、28 及 34 在六個月之後退出研究，因此僅對新資料集貢獻了單一觀察值。

1. 新增指向 *Demos* 資料夾中 *ulcer_recurrence_recoded.sav* 的「統計量檔案」來源節點。

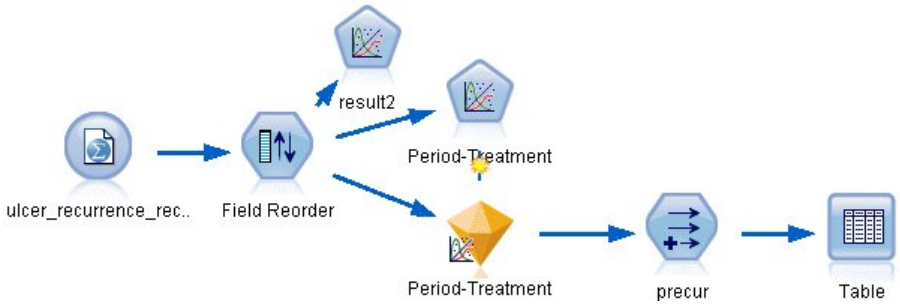


圖 *295.* 用來預測潰瘍復發的樣本串流

2. 在來源節點的「過濾器」標籤上，過濾出 *id*、*time* 及 *result*。

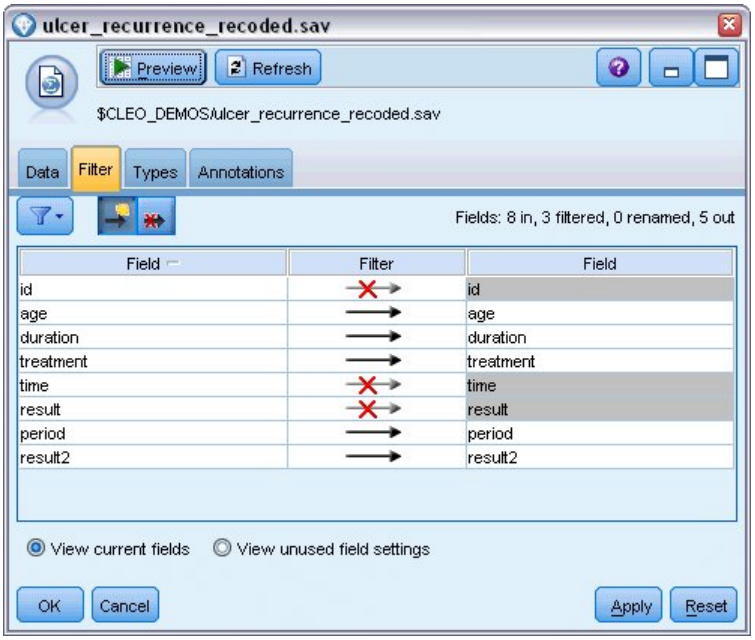


圖 *296.* 過濾不需要的欄位

3. 在來源節點的「類型」標籤上，將 *result2* 欄位的角色設定為**目標**，並將其測量層次設定為**旗標**。所有其他欄位都應該將其角色設定為**輸入**。

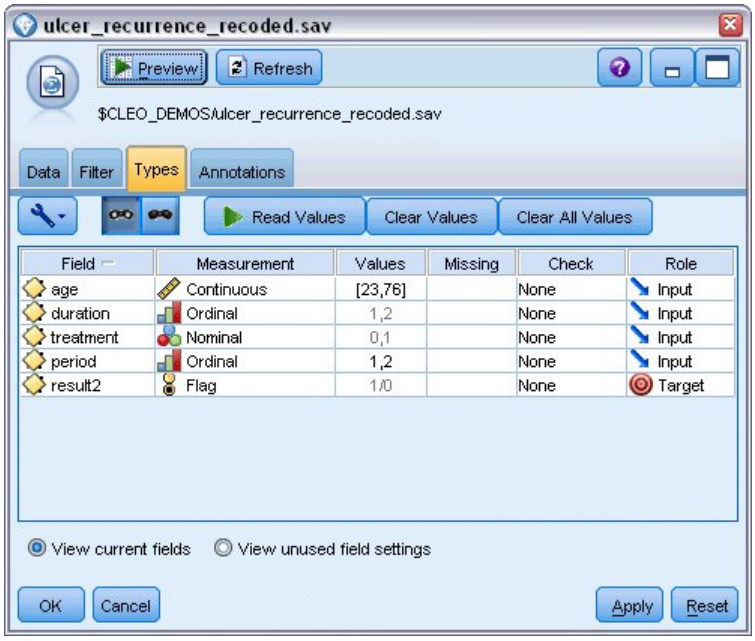


圖 *297.* 設定欄位角色

4. 新增「欄位重新排序」節點並指定 *period*、*duration*、*treatment* 及 *age* 作為輸入順序。讓 *period* 成為第一個輸入（並且模型中不包括截距項）將容許您配適一組完整的虛擬變數來擷取週期效果。

圖 *298.* 將欄位重新排序，以便根據需要將其輸入模型

5. 在 GenLin 節點上，按一下**模型**標籤。

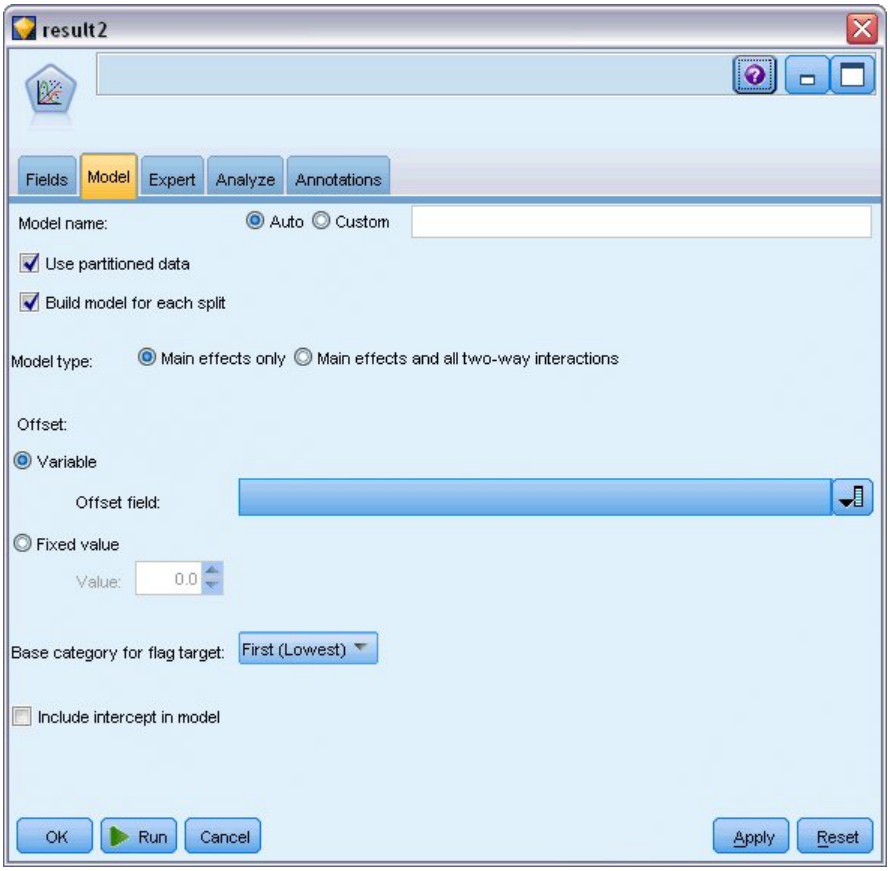


圖 *299.* 選擇模型選項

6. 選取**第一個（最低）**作為目標的參照種類。這表示第二個種類是感興趣事件，它對模型的影響存在於對參數估計值的解譯中。
7. 取消選擇「**模式中包括截距**」。
8. 按一下**專家**標籤並選取**專家**以啟動專家建模選項。

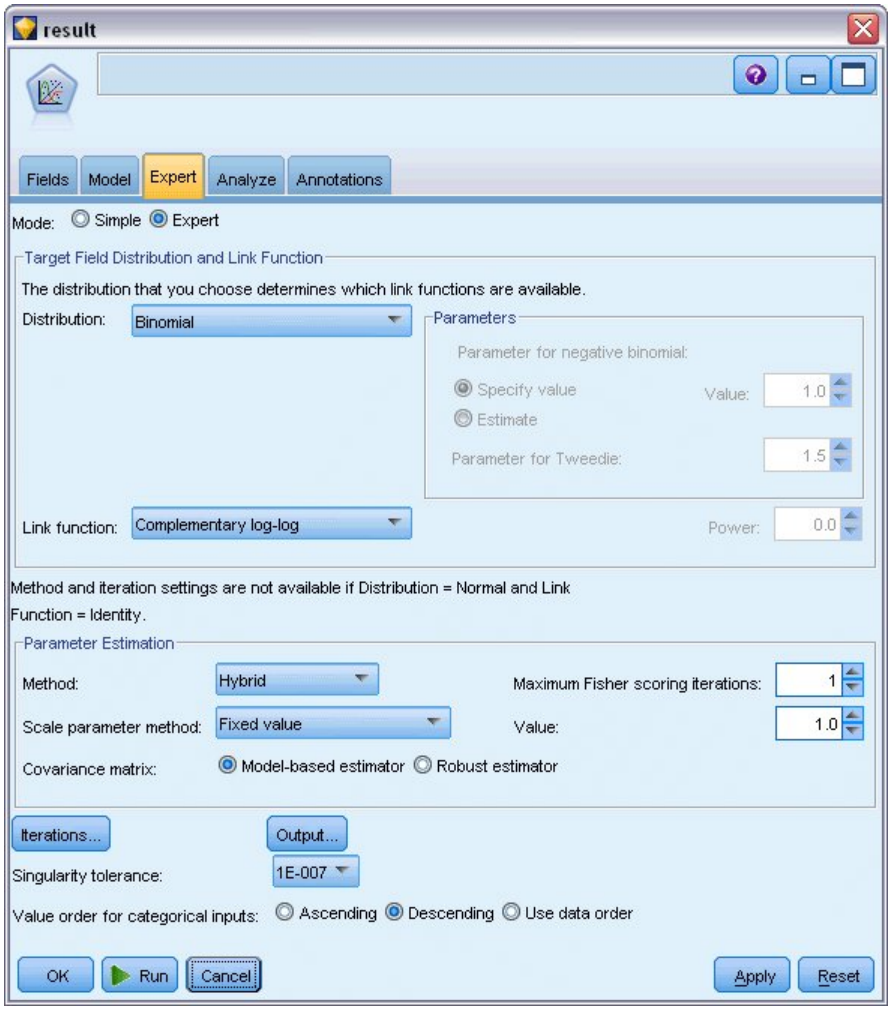


圖 *300.* 選擇專家選項

9. 選取**二項式**作為分佈，選取**互補對數存活函數的對數**作為鏈結函數。
10. 選取 **固定值**作為估計尺度參數的方法，並保留預設值 1.0。
11. 選取**遞減**作為因素的種類順序。這表示每一個因素的第一個種類是其參照種類；此選項對模型的影響存在於對參數估計值的解譯中。
12. 執行串流以建立模型區塊，該區塊會新增至串流畫布，還會新增至右上角的「模型」選用區。若要檢視模型詳細資料，請用滑鼠右鍵按一下區塊並選擇**編輯**或**瀏覽**。

模型效應的檢定

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
Period	.464	1	.496
Age in years	.314	1	.575
Duration of disease	.000	1	.988
Treatment group	.117	1	.732

Dependent Variable: Result by period

Model: Period, Age in years, Duration of disease, Treatment group

圖 *301.* 主效應模型的模型效應檢定

所有模型效應都不具有統計顯著性；但週期及治療效果中可觀察的任何差異都具有臨床意義，因此我們將僅使用那些模型項來配適縮減的模型。

配適縮減的模型

1. 在 GenLin 節點的「欄位」標籤上，按一下**使用自訂設定**。
2. 選取 *result2* 作為目標。
3. 選取 *period* 及 *treatment* 作為輸入。

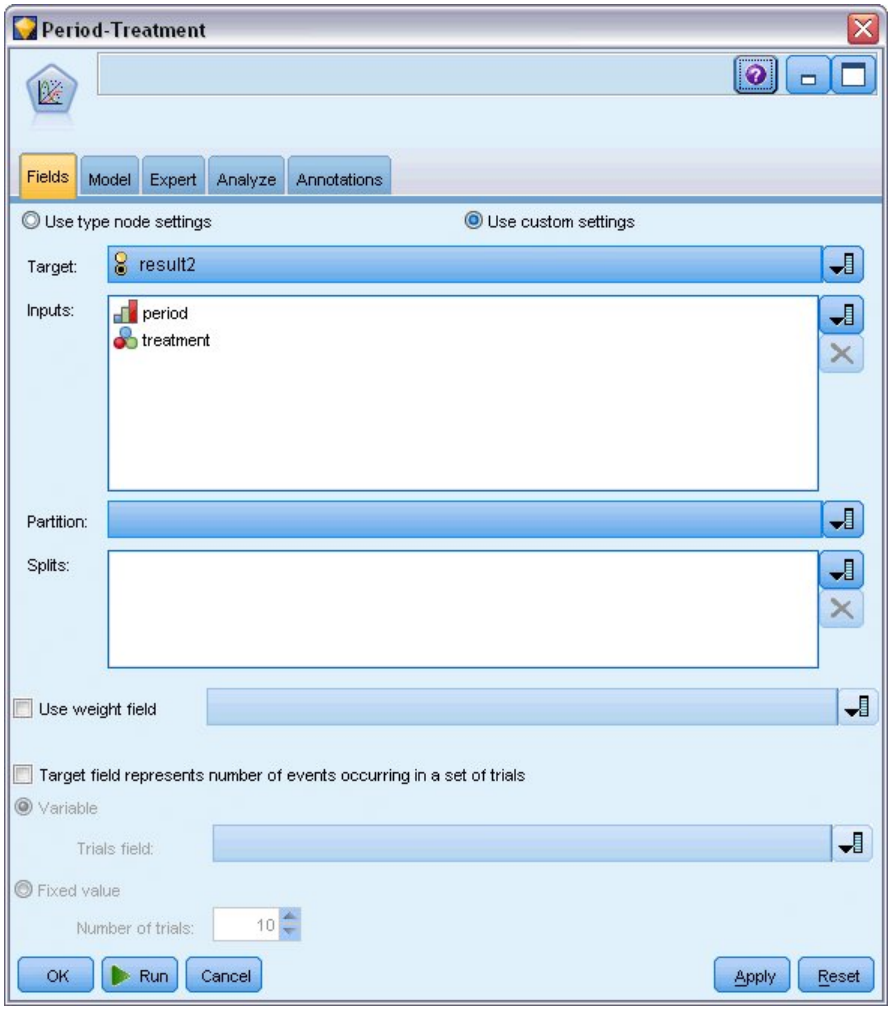


圖 *302.* 選擇欄位選項

4. 執行節點並瀏覽產生的模型，然後將產生的模型複製到選用區，連接表格節點，並執行它。

參數估計值

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[Period=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[Period=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[Treatment group=1]	.195	.6279	-1.035	1.426	.097	1	.756
[Treatment group=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period

Model: Period, Treatment group

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

圖 *303.* 僅含治療方案之模型的參數估計值

治療效果仍不具有統計顯著性，但只能表示治療方案 *A* 可能優於 *B*，因為治療方案 *B* 的參數估計值與前 12 個月中的復發機率增長相關聯。週期值在統計上明顯不同於 0，但這是因為截距項實際不適合。週期效果（*[period=1]* 和 *[period=2]* 的線性預測值之間的差異）在統計上不顯著，正如我們可以在模型效應檢定中所看見的一樣。線性預測值（週期效果 + 治療效果）是 $\log(-\log(1-P(recur_{p,\ t}))$ 的估計值，其中 $P(recur_{p,\ t})$ 是治療方案 t（=*A* 或 *B*）在週期 p（=1 或 2，代表六個月或 12 個月）時的復發機率。這些預測機率針對資料集中的每一個觀察產生。

預測的復發及存活機率

圖 *304.* 衍生節點設定選項

1. 針對每一個病患，模型會對預測結果及該預測結果的機率進行評分。為了查看預測的復發機率，請將產生的模型複製到選用區並連接「衍生」節點。
2. 在「設定」標籤中，鍵入 `precur` 作為衍生欄位。
3. 選擇將其衍生為**條件式**。
4. 按一下計算機按鈕以開啟 **If** 條件的表示式建置器。

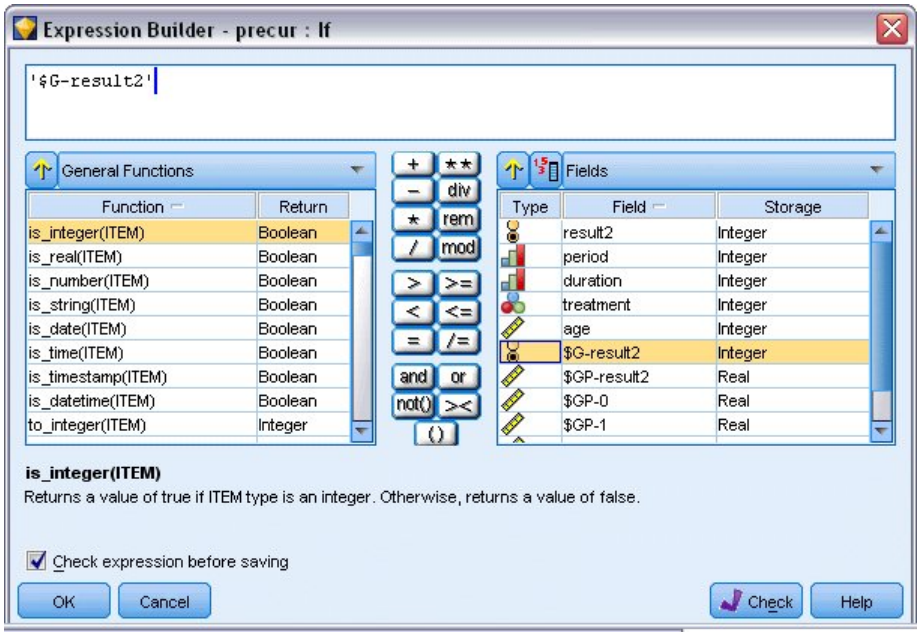


圖 *305.* 衍生節點：*If* 條件的表示式建置器

5. 將 *$G-result2* 欄位插入表示式。
6. 按一下**確定**。

   當 *$G-result2* 等於 1 時，衍生欄位 *precur* 會採取 **Then** 表示式的值，當它為 0 時，則會採取 **Else** 表示式的值。

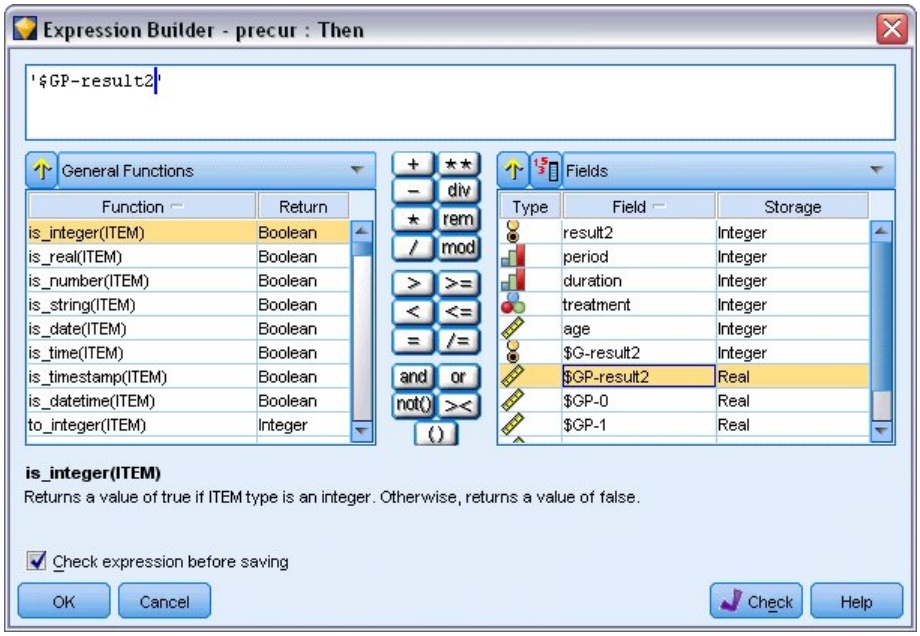


圖 *306.* 衍生節點：*Then* 表示式的表示式建置器

7. 按一下計算機按鈕以開啟 **Then** 表示式的表示式建置器。
8. 將 *$GP-result2* 欄位插入表示式。

9. 按一下「**確定**」。

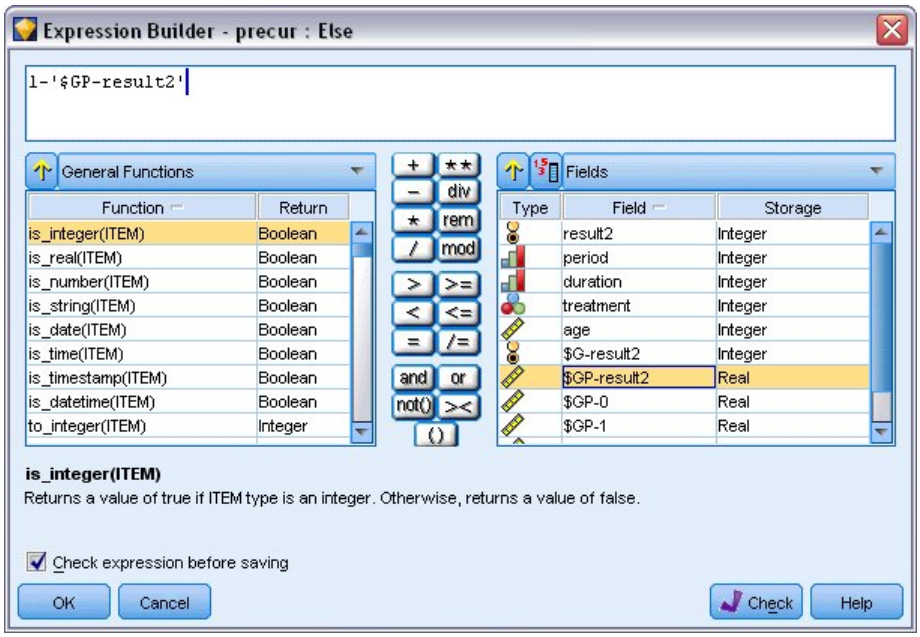


圖 *307.* 衍生節點：*Else* 表示式的表示式建置器

10. 按一下計算機按鈕以開啟 **Else** 表示式的表示式建置器。
11. 在表示式中鍵入 1-，然後將 *$GP-result2* 欄位插入表示式。
12. 按一下「**確定**」。

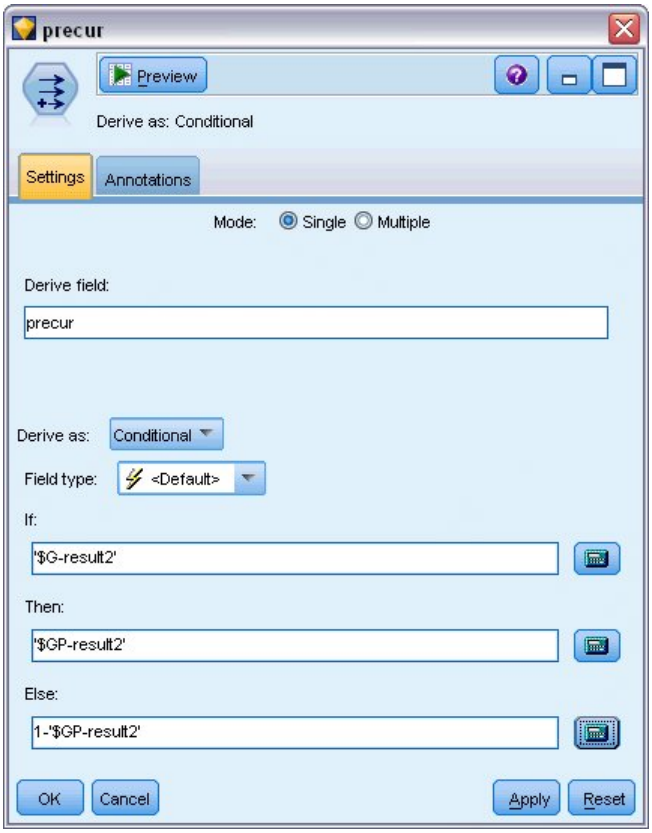
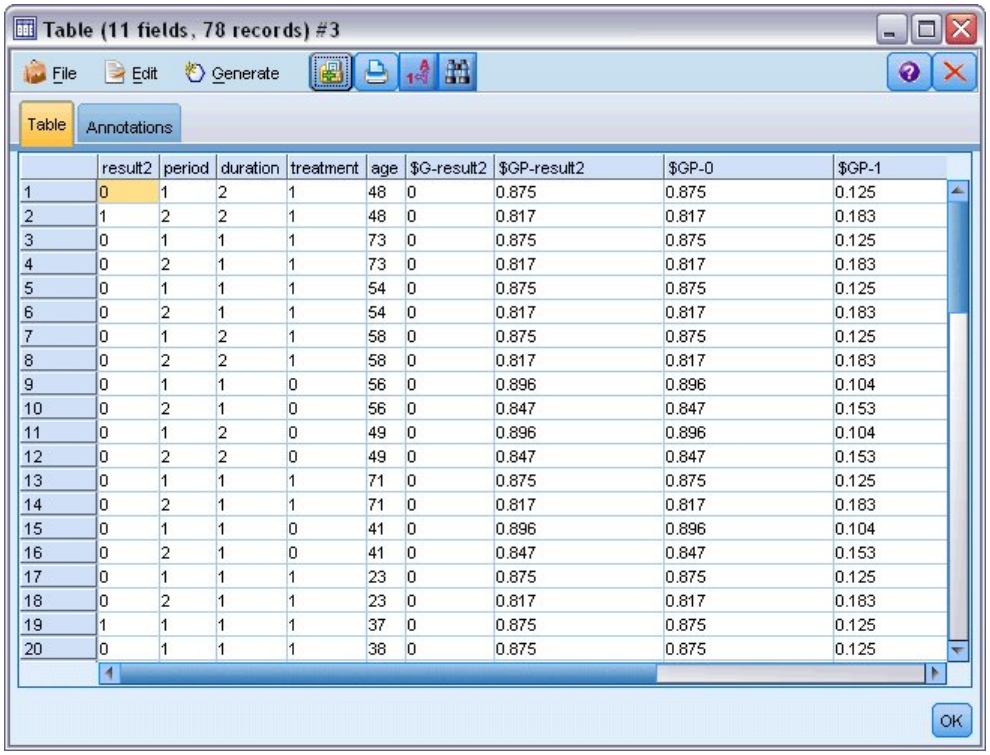


圖 *308.* 衍生節點設定選項

13. 將表格節點連接至「衍生」節點並執行它。



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

圖 *309.* 預測機率

表 *3.* 估計的復發機率

治療方案	6 個月	12 個月
A	0.104	0.153
B	0.125	0.183

從估計的復發機率，可以將 12 個月內的存活機率估計為 $1–(P(recur_{1,\ t}) + P(recur_{2,\ t})\times(1–P(recur_{1,\ t})))$；因此，針對每一種治療方案：

A：1 – (0.104 + 0.153*0.896) = 0.759

B：1 – (0.125 + 0.183*0.875) = 0.715

再次對 *A* 作為較好的治療方案顯示出非統計上的顯著性支援。

摘要

使用「廣義線性模型」，您為區間受限存活資料配適了一系列互補對數存活函數的對數迴歸模型。雖然對選擇治療方案 *A* 有某種程度的支援，但要達到統計顯著性結果可能需要更大型的研究。然而，針對現有資料還有進一步的探索方法。

- 也許可以使用互動效應重新配適模型，特別是週期與治療群組之間的互動。

在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出。

相關程序

「廣義線性模型」程序是一種功能強大的工具，可配適各種模型。

- 「廣義估計方程式」程序可延伸廣義線性模型來容許重複測量。
- 「線性混合模型」程序可容許您針對含隨機成分的尺度應變數及/或重複測量配適模型。

閱讀資料推薦

如需廣義線性模型的相關資訊，請參閱下列文字：

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 23 章 使用卜瓦松 (Poisson) 迴歸來分析船隻損壞率（廣義線性模型）

您可以使用廣義線性模型為計數資料分析配適卜瓦松 (Poisson) 迴歸。例如，在其他位置 [2] 呈現並分析的資料集涉及由波浪造成的貨船損壞。事件計數可以建模為在給定預測值的情況下以卜瓦松 (Poisson) 比率出現，產生的模型可協助您判斷哪些船隻類型最容易損壞。

此範例使用串流 *ships_genlin.str*，其參照資料檔 *ships.sav*。資料檔位於 *Demos* 資料夾中，串流檔位於 *streams* 子資料夾中。

在這種狀況下為原始儲存格計數建模可能會產生誤導，因為聚集服務月數會隨船隻類型而變。像這樣可測量風險「暴露」量的變數會在廣義線性模型內作為偏移變數處理。而且，卜瓦松 (Poisson) 迴歸還會假設應變數的對數在預測值中為線性。因此，若要使用廣義線性模型針對事故率配適卜瓦松 (Poisson) 迴歸，則需要使用聚集服務月數的對數。

配適「過度離散」的卜瓦松 (Poisson) 迴歸

1. 新增指向 *Demos* 資料夾中 *ships.sav* 的「統計量檔案」來源節點。

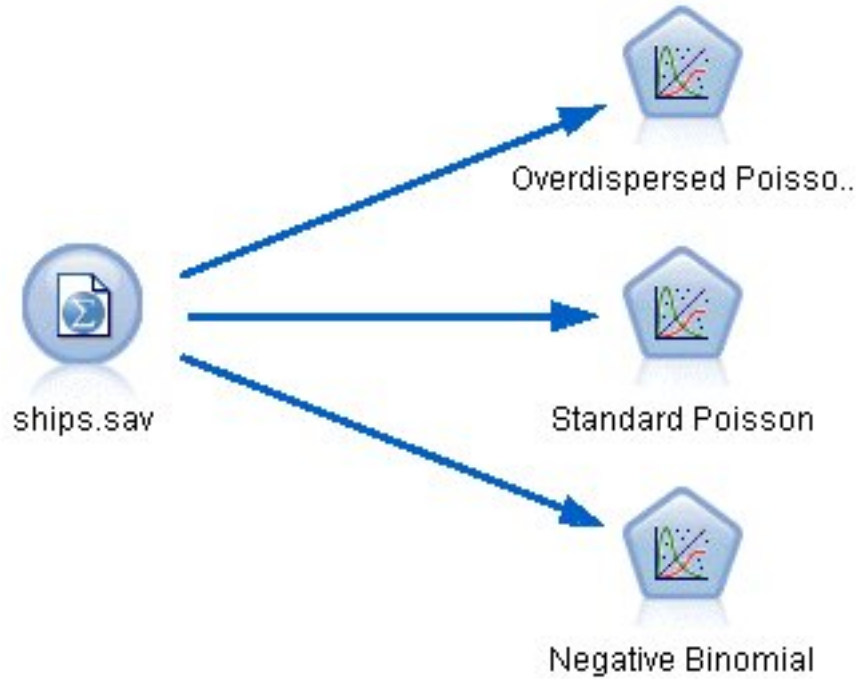


圖 *310.* 用來分析損壞率的樣本串流

2. 在來源節點的「過濾器」標籤上，排除欄位 *months_service*。此變數的對數轉換值包含在 *log_months_service* 中，並會在分析中使用。

2. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

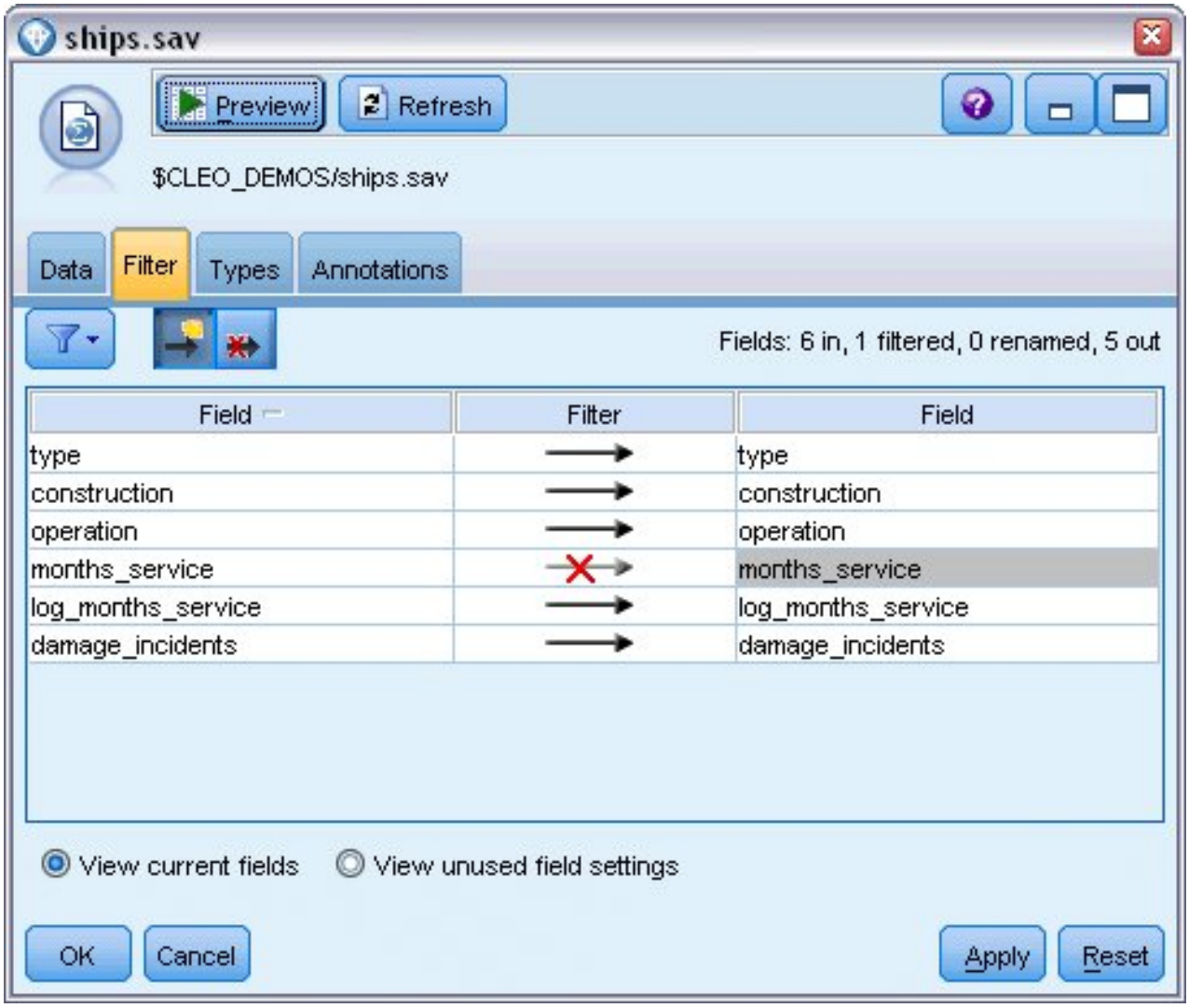


圖 *311.* 過濾不需要的欄位

（或者，您可以將「類型」標籤上此欄位的角色變更為**無**而不是排除它，或選取要在建模節點中使用的欄位。）

3. 在來源節點的「類型」標籤上，將 *damage_incidents* 欄位的角色設定為**目標**。所有其他欄位都應該將其角色設定為**輸入**。
4. 按一下**讀取值**以將資料實例化。

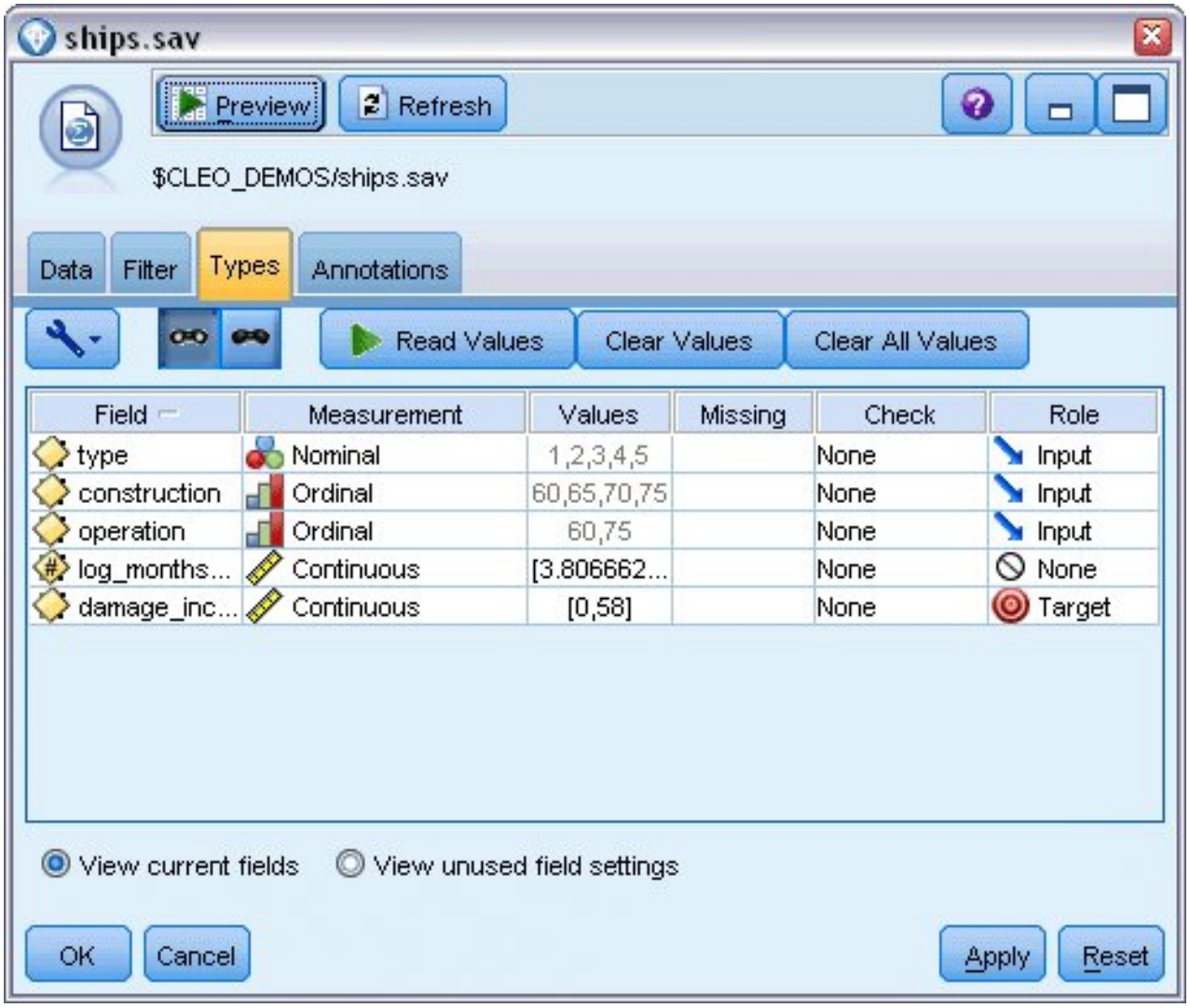


圖 *312.* 設定欄位角色

5. 將 Genlin 節點連接至來源節點；在 Genlin 節點上，按一下**模型**標籤。
6. 選取 *log_months_service* 作為偏移變數。

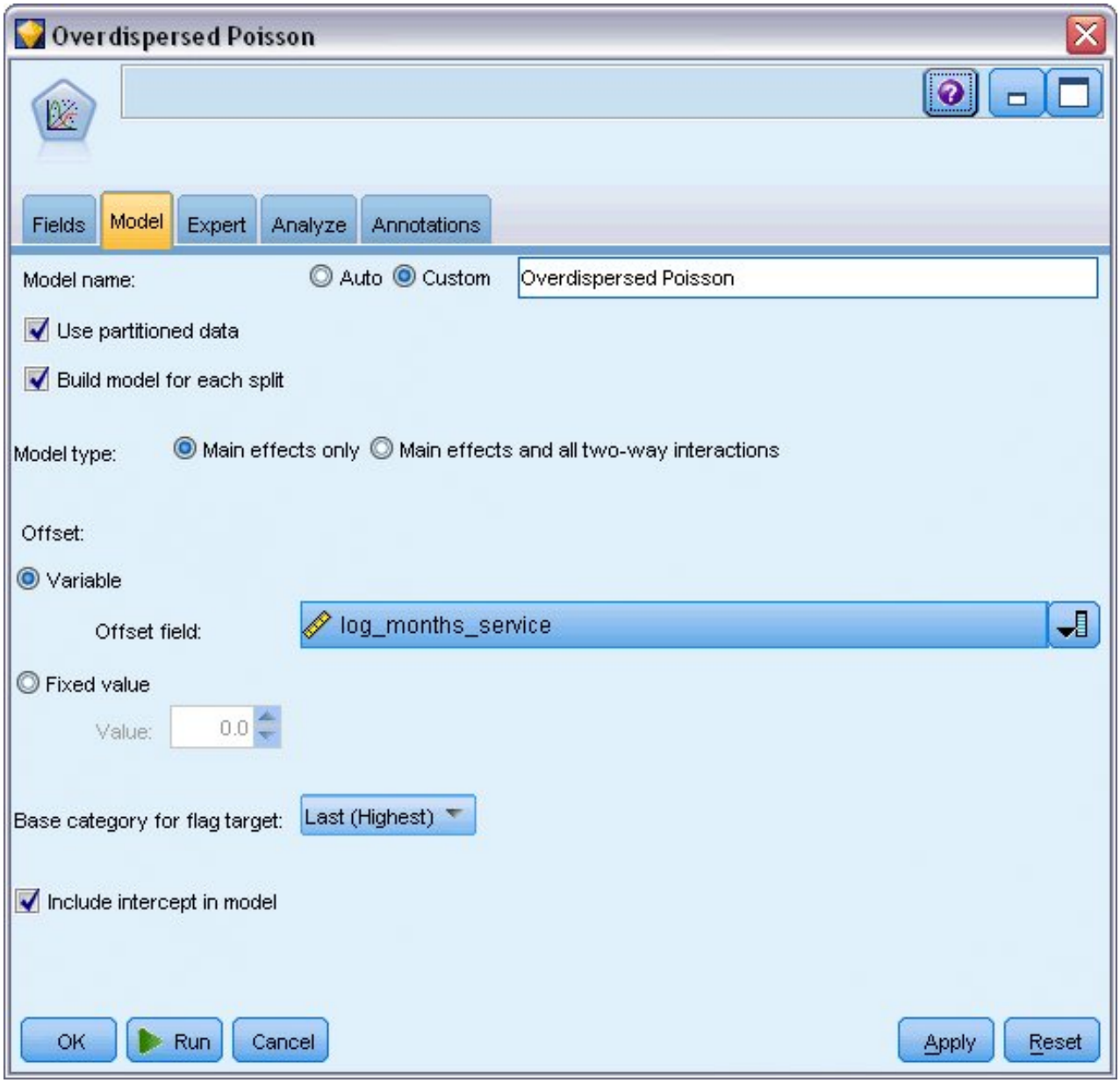


圖 *313.* 選擇模型選項

7. 按一下**專家**標籤並選取**專家**以啟動專家建模選項。

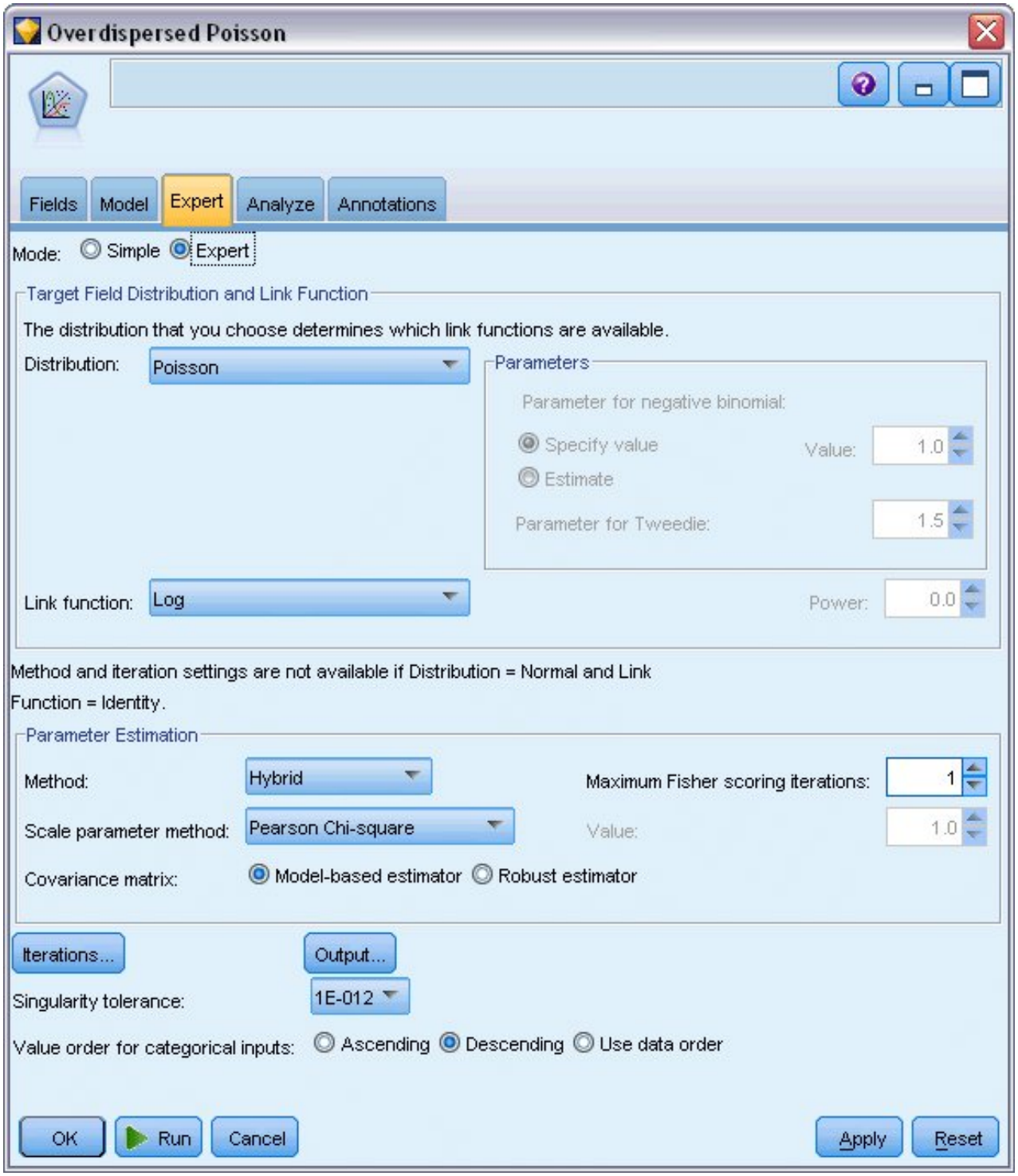


圖 *314.* 選擇專家選項

8. 選取 **Poisson** 作為回應的分配，並選取**對數**作為鏈結函數。
9. 選取 **Pearson 卡方**作為估計尺度參數的方法。在卜瓦松 (Poisson) 迴歸中尺度參數通常假設為 1，但 McCullagh 和 Nelder 使用 Pearson 卡方估計可取得更保守的變異估計值及顯著性層次。
10. 選取**遞減**作為因素的種類順序。這表示每一個因素的第一個種類是其參照種類；此選項對模型的影響存在於對參數估計值的解譯中。

11. 按一下**執行**以建立模型區塊，該區塊會新增至串流畫布，還會新增至右上角的「模型」選用區。若要檢視模型詳細資料，請用滑鼠右鍵按一下區塊並選擇**編輯**或**瀏覽**，然後按一下**進階**標籤。

適合度統計量

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

圖 *315.* 適合度統計量

適合度統計量表會提供用來比較競爭模型的量數。此外，離差和 Pearson 卡方統計量的值*/df* 會提供尺度參數的對應估計值。卜瓦松 (Poisson) 迴歸的這些值應接近 1.0；實際大於 1.0 則表示配適過度離散的模型可能是合理的。

綜合測試

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
63.650	8	.000

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

a. Compares the fitted model against the intercept-only model.

圖 *316.* 綜合測試

綜合測試是對現行模型與虛無（在此情況下為截距）模型概似比進行的卡方測試。顯著性值小於 0.05 表示現行模型優於虛無模型。

模型效應的檢定

Tests of Model Effects

	Type III		
Source	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
Year of construction	17.242	3	.001
Period of operation	6.249	1	.012
Ship type	15.415	4	.004

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

圖 *317.* 模型效應的檢定

檢定模型中的每一項是否具有任何效應。顯著性值小於 0.05 的項具有一定明顯的效應。每一個主效應項都對模型有貢獻。

參數估計值

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2828	-6.960	-5.852	513.238	1	.000
[Year of construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[Year of construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[Year of construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[Year of construction=60]	0[a]	.	.	.	.	.	.
[Period of operation=75]	.384	.1538	.083	.686	6.249	1	.012
[Period of operation=60]	0[a]	.	.	.	.	.	.
[Ship type=5]	.326	.3067	-.276	.927	1.127	1	.288
[Ship type=4]	-.076	.3779	-.817	.665	.040	1	.841
[Ship type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[Ship type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[Ship type=1]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

圖 *318.* 參數估計值

參數估計值的表格會將各個預測值的效果做成摘要。雖然此模型中的係數因為鏈結函數的本質而很難解譯，但共變數的係數符號及因素層次的係數相對值可能會對模型中的預測值效應提供重要的見解。

- 針對共變數，正（負）係數表示預測值與結果之間的關係為正向（反向）。係數為正的共變數的值增長會對應於損壞事件率的增長。
- 針對因素，因素層次的係數越大表示損壞發生率越大。因素層次的係數符號取決於因素層次相對於參照種類的影響。

您可以基於參數估計值進行下列解譯：

- 船隻類型 *B [type=2]* 的損壞率在統計上顯著（*p* 值為 0.019）低於（估計係數為 –0.543）參照種類類型 *A [type=1]*。類型 *C [type=3]* 的估計參數實際比 *B* 低，但 *C* 估計值中的變異性減弱了效果。請參閱因素層次之間所有關係的估計邊際平均數。
- 在 1965–69 *[construction=65]* 之間與 1970–74 *[construction=70]* 之間建造的船隻損壞率在統計上顯著（*p* 值 <0.001）高於（估計係數分別為 0.697 和 0.818）參照種類 1960–64 *[construction=60]* 之間建造的船隻。請參閱因素層次之間所有關係的估計邊際平均數。
- 在 1975–79 *[operation=75]* 之間作業的船隻損壞率在統計上顯著（*p* 值為 0.012）高於（估計係數為 0.384）在 1960–1974 *[operation=60]* 之間作業的船隻。

配適替代模型

「過度離散」的卜瓦松 (Poisson) 迴歸的一個問題是沒有正式的方法對它與「標準」卜瓦松 (Poisson) 迴歸進行測試。但有一種建議的正式測試可用來判定是否存在過度離散，即在「標準」卜瓦松 (Poisson) 迴歸與所有其他設定都相同的負二項式迴歸之間執行概似比測試。如果卜瓦松 (Poisson) 迴歸中不存在過度離散，則統計量 –2×（卜瓦松 (Poisson) 模型的對數概似值 – 負二項式模型的對數概似值）應具有混合分配，其中一半機率群位於 0，其餘機率則位於自由度為 1 的卡方分配中。

1. 選取 **固定值**作為估計尺度參數的方法。依預設，此值為 1。

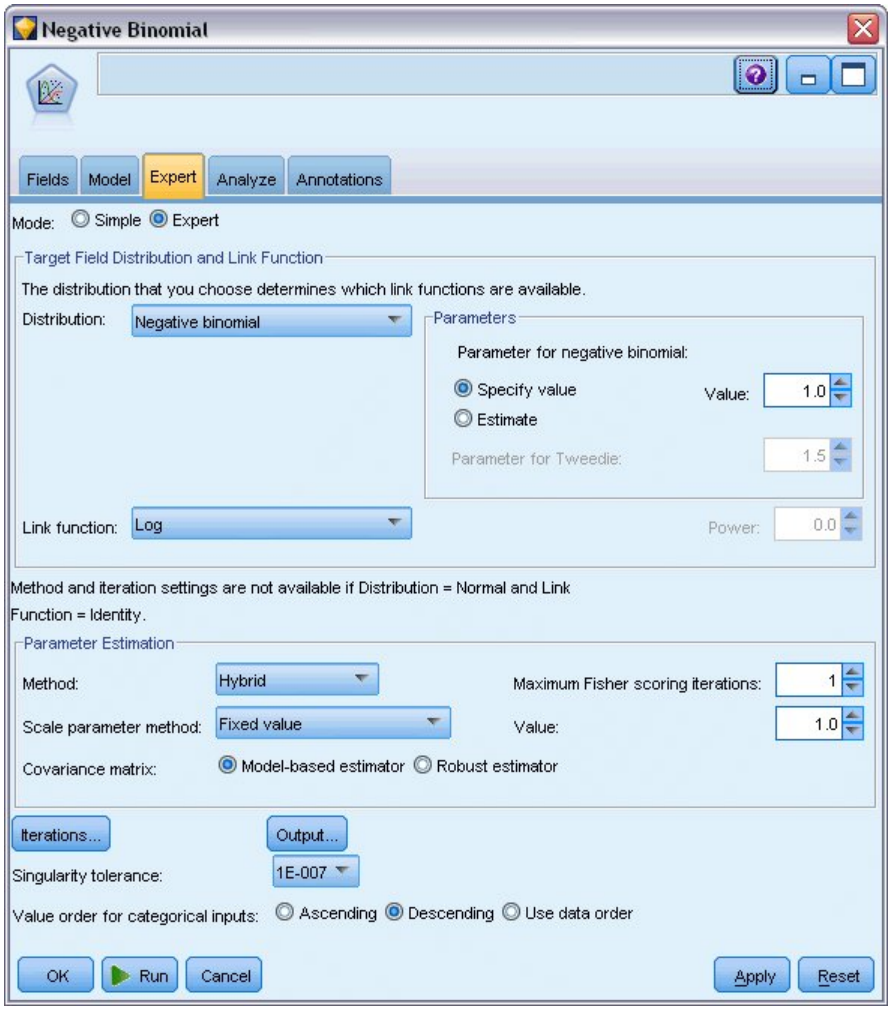


圖 *319.* 專家標籤

2. 若要配適負二項式迴歸，請複製並貼上 Genlin 節點，將其連接至來源節點，開啟新節點並按一下**專家**標籤。
3. 選取**負二項式**作為分配。保留輔助參數的預設值 1。
4. 在新建立的模型區塊上執行串流並瀏覽「進階」標籤。

適合度統計量

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

圖 *320.* 標準卜瓦松 *(Poisson)* 迴歸的適合度統計量

針對標準卜瓦松 (Poisson) 迴歸報告的對數概似值為 –68.281。將此值與負二項式模型進行比較。

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

圖 *321.* 負二項式迴歸的適合度統計量

針對負二項式迴歸報告的對數概似值為 –83.725。此值實際上小於卜瓦松 (Poisson) 迴歸的對數概似值，這指出（無需進行概似比測試）此負二項式迴歸對卜瓦松 (Poisson) 迴歸沒有任何改進。

但負二項式分佈輔助參數的選用值 1 對於此資料集來說可能不是最理想的。測試過度離散可以使用的另一種方法是配適輔助參數等於 0 的負二項式模型，並在「專家」標籤的「輸出」對話框上要求進行 Lagrange 乘數測試。如果測試不顯著，則對於此資料集來說過度離散不應該是個問題。

摘要

使用「廣義線性模型」，您為計數資料配適了三個不同的模型。事實證明，負二項式迴歸對卜瓦松 (Poisson) 迴歸沒有任何改進。過度離散的卜瓦松 (Poisson) 迴歸似乎成為標準卜瓦松 (Poisson) 模型的合理替代方案，但並沒有正式的測試可用於在二者之間進行選擇。

在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出。

相關程序

「廣義線性模型」程序是一種功能強大的工具，可配適各種模型。

- 「廣義估計方程式」程序可延伸廣義線性模型來容許重複測量。
- 「線性混合模型」程序可容許您針對含隨機成分的尺度應變數及/或重複測量配適模型。

閱讀資料推薦

如需廣義線性模型的相關資訊，請參閱下列文字：

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 24 章 對汽車保險索賠配適伽瑪迴歸（廣義線性模型）

您可以使用廣義線性模型為正範圍資料分析配適伽瑪迴歸。例如，在其他位置 [3] 呈現並分析的資料集涉及汽車損壞索賠。平均索賠金額可以建模為具有伽瑪分配，使用反向鏈結函數將應變數平均數與預測值的線性組合關聯起來。為了說明用來計算平均索賠金額的可變索賠數，請指定索賠數作為調整比例加權。

此範例使用名為 *car-insurance_genlin.str* 的串流，其參照的資料檔名為 *car_insurance_claims.sav*。資料檔位於 *Demos* 資料夾中，串流檔位於 *streams* 子資料夾中。

建立串流

1. 新增指向 *Demos* 資料夾中 *telco.sav* 的「統計量檔案」來源節點。

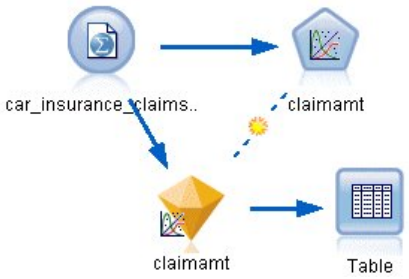


圖 *322.* 用來預測汽車保險索賠的樣本串流

2. 在來源節點的「類型」標籤上，將 *claimamt* 欄位的角色設定為**目標**。所有其他欄位應該將其角色設為**輸入**。
3. 按一下**讀取值**以將資料實例化。

3. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

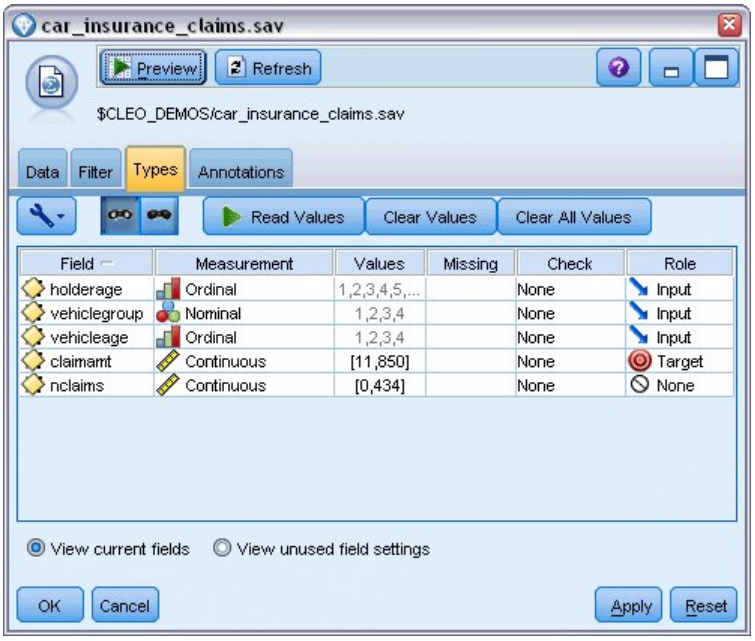


圖 *323.* 設定欄位角色

4. 將 Genlin 節點連接至來源節點；在 Genlin 節點中，按一下「欄位」標籤。
5. 選取 *nclaims* 作為尺度加權欄位。

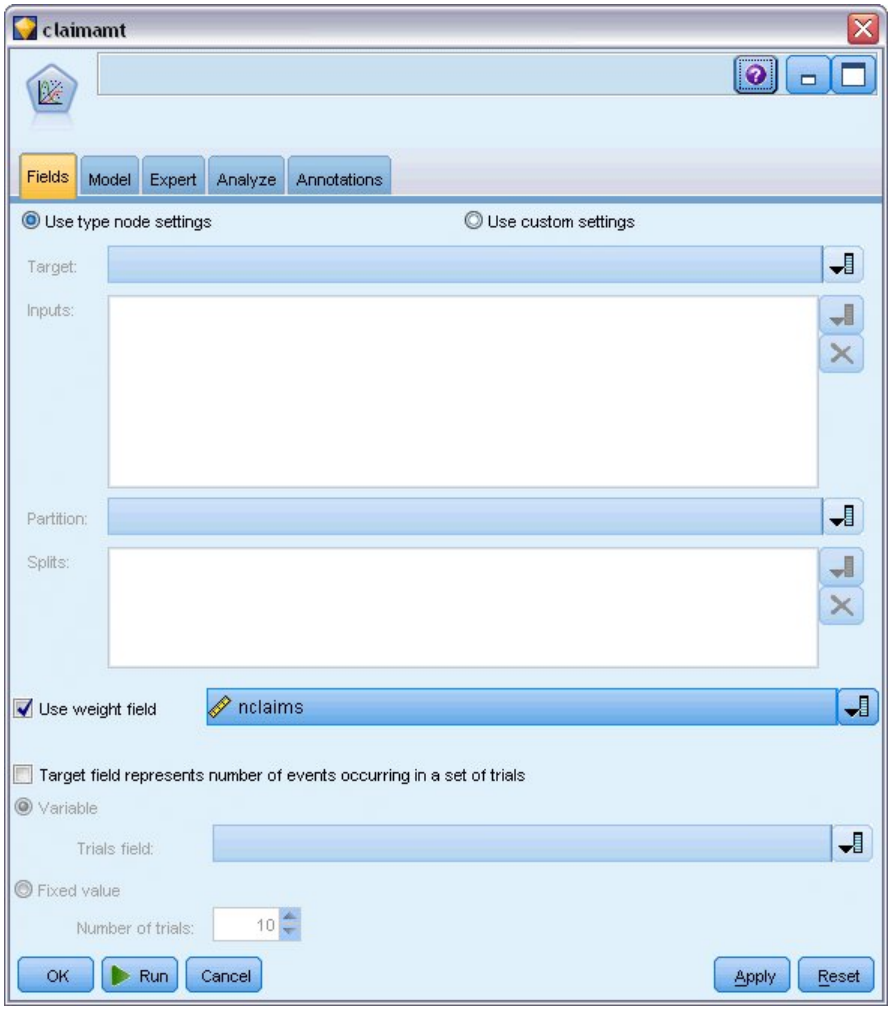


圖 *324.* 選擇欄位選項

6. 按一下「專家」標籤並選取**專家**以啟動專家建模選項。

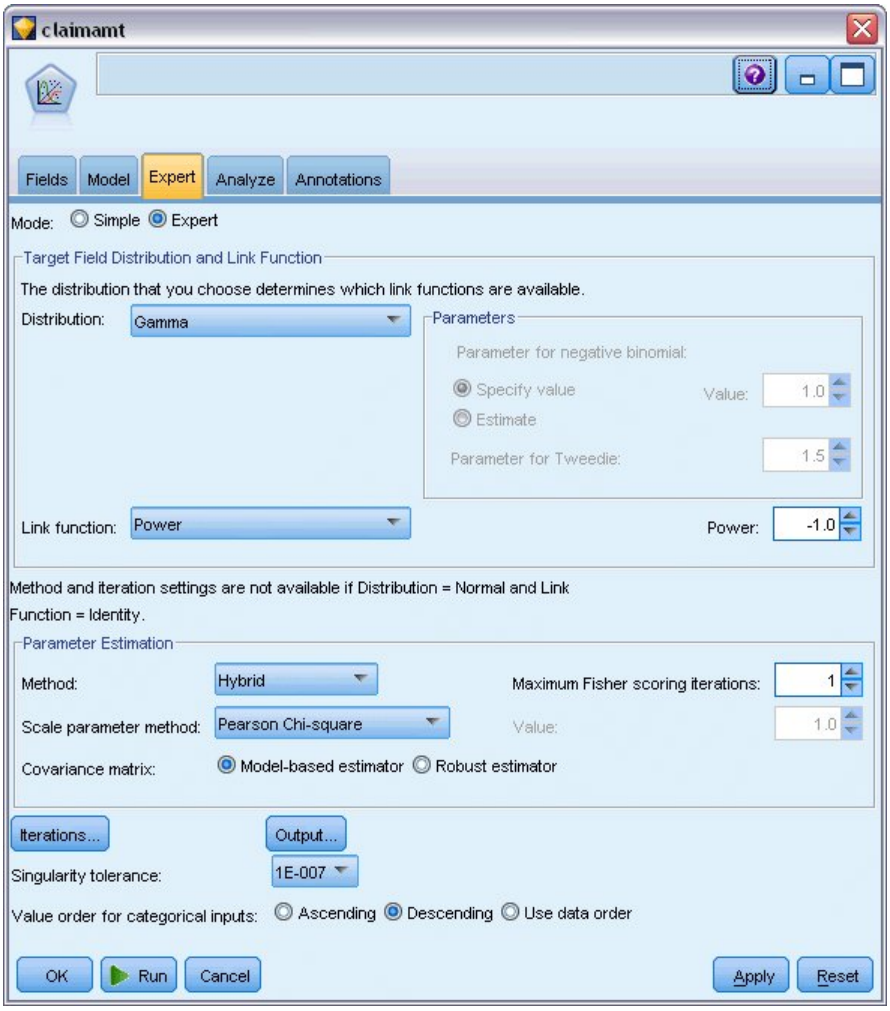


圖 *325.* 選擇專家選項

7. 選取**伽瑪**作為回應分佈。
8. 選取**冪次**作為鏈結函數，並鍵入 `-1.0` 作為冪次函數的指數。這是一個反向鏈結。
9. 選取 **Pearson 卡方**作為估計尺度參數的方法。這是 McCullagh 和 Nelder 使用的方法，因此我們在這裡遵循該方法以便抄寫其結果。
10. 選取**遞減**作為因素的種類順序。這表示每一個因素的第一個種類是其參照種類；此選項對模型的影響存在於對參數估計值的解譯中。
11. 按一下**執行**以建立模型區塊，該區塊會新增至串流畫布，還會新增至右上角的「模型」選用區。若要檢視模型詳細資料，請用滑鼠右鍵按一下模型區塊並選擇**編輯**或**瀏覽**，然後選取「進階」標籤。

參數估計值

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003	.0004	.003	.004	66.593	1	.000
[Policyholder age=8]	.001	.0004	.000	.002	4.898	1	.027
[Policyholder age=7]	.001	.0004	.000	.002	5.046	1	.025
[Policyholder age=6]	.001	.0004	.000	.002	5.740	1	.017
[Policyholder age=5]	.001	.0004	.001	.002	10.682	1	.001
[Policyholder age=4]	.000	.0004	.000	.001	1.268	1	.260
[Policyholder age=3]	.000	.0004	.000	.001	.720	1	.396
[Policyholder age=2]	.000	.0004	-.001	.001	.054	1	.816
[Policyholder age=1]	0[a]	.	.	.	.	.	.
[Vehicle age=4]	.004	.0004	.003	.005	88.175	1	.000
[Vehicle age=3]	.002	.0002	.001	.002	53.013	1	.000
[Vehicle age=2]	.000	.0001	.000	.001	13.191	1	.000
[Vehicle age=1]	0[a]	.	.	.	.	.	.
[Vehicle group=4]	-.001	.0002	-.002	-.001	61.883	1	.000
[Vehicle group=3]	-.001	.0002	-.001	.000	13.039	1	.000
[Vehicle group=2]	3.765E-5	.0002	.000	.000	.050	1	.823
[Vehicle group=1]	0[a]	.	.	.	.	.	.
(Scale)	1.209[b]						

Dependent Variable: Average cost of claims

Model: (Intercept), Policyholder age, Vehicle age, Vehicle group

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

圖 *326.* 參數估計值

綜合測試及模型效應（未顯示）測試指出該模型優於虛無模型，並且每一個主效應項都對模型有貢獻。參數估計值表格顯示 McCullagh 和 Nelder 針對因素層次及尺度參數取得的相同值。

摘要

您使用「廣義線性模型」針對索賠資料配適了伽瑪迴歸。請注意，雖然在此模型中使用了伽瑪分配的標準鏈結函數，但對數鏈結也會提供合理的結果。一般來說，很難直接將模型與不同的鏈結函數進行比較；但對數鏈結是指數為 0 的特殊冪次鏈結，因此您可以比較含對數鏈結的模型與含冪次鏈結的模型離差，來判斷哪個模型的適合度更好（例如，請參閱 McCullagh 和 Nelder 的 11.3 節）。

在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出。

相關程序

「廣義線性模型」程序是一種功能強大的工具，可配適各種模型。

- 「廣義估計方程式」程序可延伸廣義線性模型來容許重複測量。
- 「線性混合模型」程序可容許您針對含隨機成分的尺度應變數及/或重複測量配適模型。

閱讀資料推薦

如需廣義線性模型的相關資訊，請參閱下列文字：

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 25 章 分類細胞樣本 (SVM)

支援向量機器 (SVM) 是一個分類與迴歸技術，特別適合用於大量資料集。大量資料集指的是具有大量預測工具的資料集，例如可能會在生物資訊學（對生物化學和生物學資料套用資訊技術）欄位中遇到的預測工具。

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。研究人員想要開發一個 SVM 模型，該模型可使用其他病患樣本中的這些細胞性質值來提前指出其樣本是良性還是惡性。

本樣本使用串流 *svm_cancer.str*，其位於 *Demos* 資料夾的 *streams* 子資料夾之下。資料檔案是 *cell_samples.data*。請參閱第 4 頁的『Demos 資料夾』主題，以取得更多資訊。

該樣本基於可公開從 UCI 機器學習儲存庫取得的資料集。資料集包含數百個人類細胞樣本記錄，每筆記錄包含一組細胞性質的值。每筆記錄中的欄位如下：

欄位名稱	說明
ID	病患 ID
腫塊	腫塊厚度
UnifSize	細胞大小的均勻性
UnifShape	細胞形狀的均勻性
MargAdh	邊緣粘黏
SingEpiSize	單一上皮細胞大小
BareNuc	裸核
BlandChrom	Bland 染色質
NormNucl	正常核仁
Mit	有絲分裂
類別	良性或惡性

出於本樣本的目的，我們將使用在每筆記錄中具有相對較少數目的預測工具的資料集。

建立串流

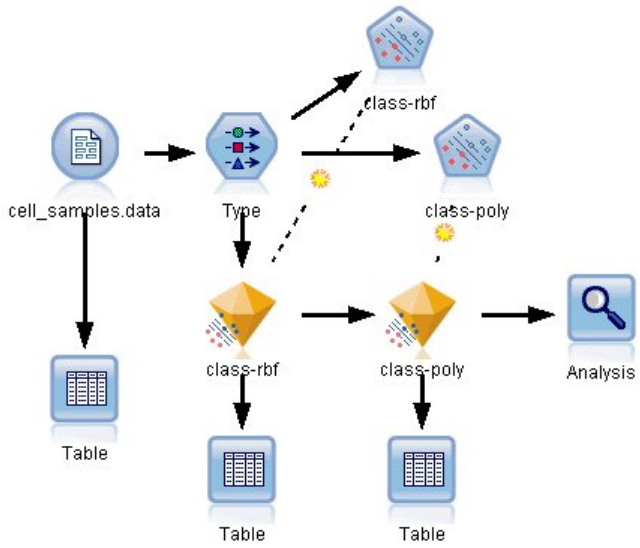
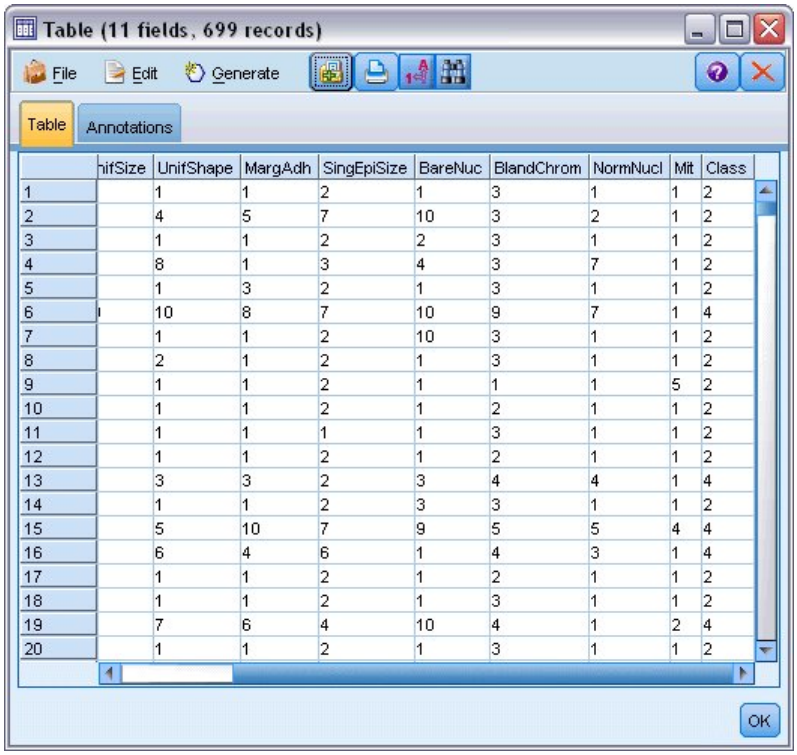


圖 *327.* 顯示 *SVM* 建模的串流範例

1. 建立新串流並新增指向 *cell_samples.data*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「變數檔案」來源節點。

   我們來看看原始檔中的資料。
2. 將「表格」節點新增到串流中。
3. 將「表格」節點連接至「變數檔案」節點，然後執行串流。



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

圖 *328. SVM* 的來源資料

ID 欄位包含病患 ID。每個病患的細胞樣本性質包含在欄位 *Clump* 到 *Mit* 中。值的分級從 1 到 10，值為 1 表示最先開始。

Class 欄位包含診斷，由個別醫療程序確認樣本是良性（值 = 2）還是惡性（值 = 4）。

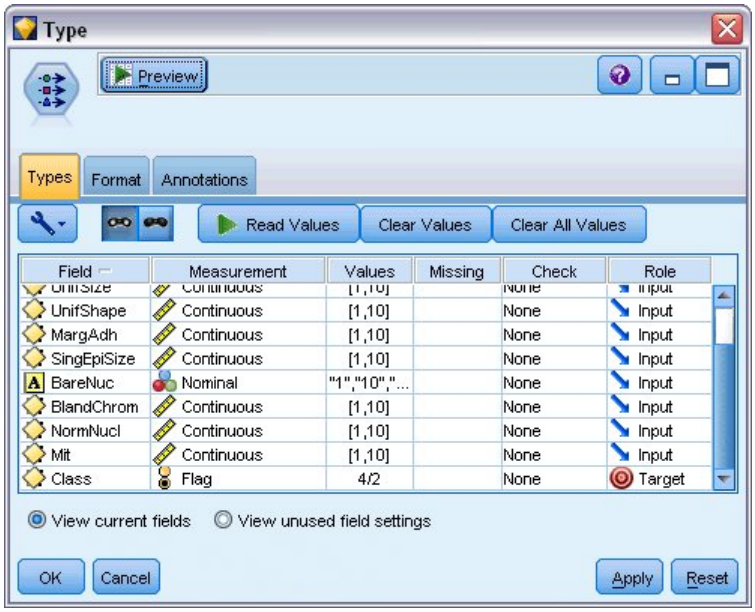


圖 *329.* 類型節點設定

4. 新增「類型」節點並將它連接至「變數檔案」節點。

5. 開啟「類型」節點。

   我們想要建模來預測 *Class* 的值（即良性 (=2)，或惡性 (=4)）。由於此欄位只能有兩個可能值的其中一個，因此我們需要將其測量層級變更為反映此結果。
6. 在 *Class* 欄位（清單中的最後一個欄位）的**測量**直欄中，按一下**繼續**並將它變更為**旗標**。
7. 按一下**讀取值**。
8. 在**角色**直欄中，將 *ID*（病患 ID）的角色設為**無**，因為它將不會用來作為模型的預測工具或目標。
9. 將目標 *Class* 的角色設為**目標**，並將所有其他欄位（預測工具）的角色保留為**輸入**。
10. 按一下**「確定」**。

    SVM 節點可讓您選擇核心函數來執行其處理。由於無法輕鬆得知哪個函數對任何給定的資料集執行結果最佳，因此我們將依次選擇不同的函數並比較結果。我們從預設的 RBF（徑向基底函數）開始吧。

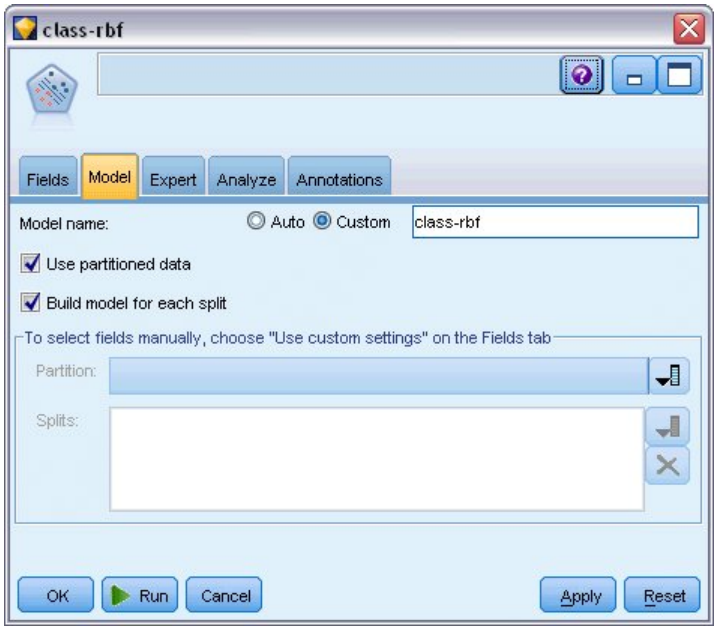


圖 *330.* 模型標籤設定

11. 從「建模」選用區將 SVM 節點連接至「類型」節點。
12. 開啟 SVM 節點。在**模型**標籤上，針對**模型名稱**按一下**自訂**選項，然後在相鄰的文字欄位中輸入 *class-rbf*。

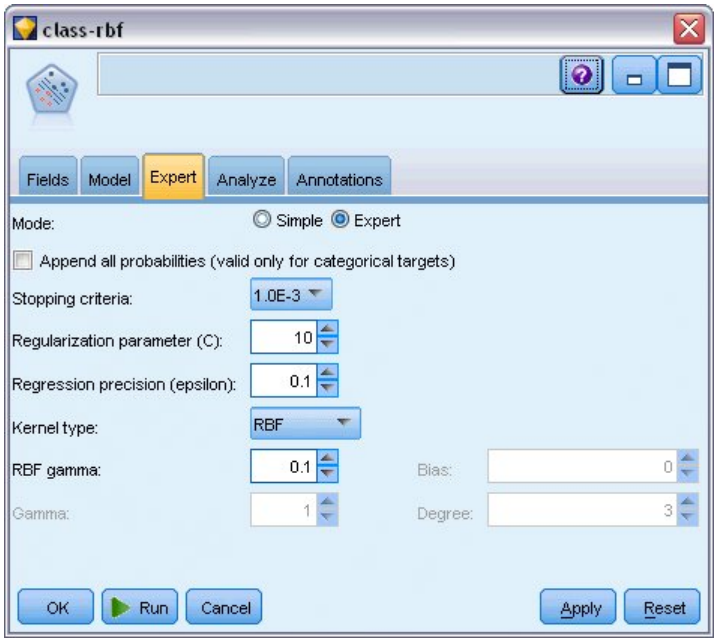


圖 *331.* 預設專家標籤設定

13. 在**專家**標籤上，將**模式**設為**專家**以方便讀取，但將所有預設選項保留原樣。請注意，依預設**核心類型**設為 **RBF**。在簡式模式下所有選項都變成灰色。

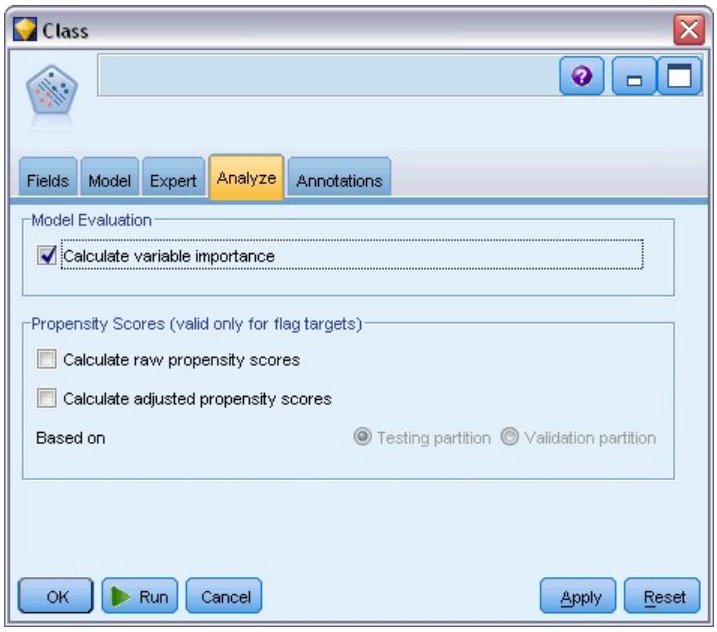


圖 *332.* 分析標籤設定

14. 在**分析**標籤上，選取**計算變數重要性**勾選框。
15. 按一下「**執行**」。模型區塊放置在串流和畫面右上方的「模型」選用區中。
16. 按兩下串流中的模型區塊。

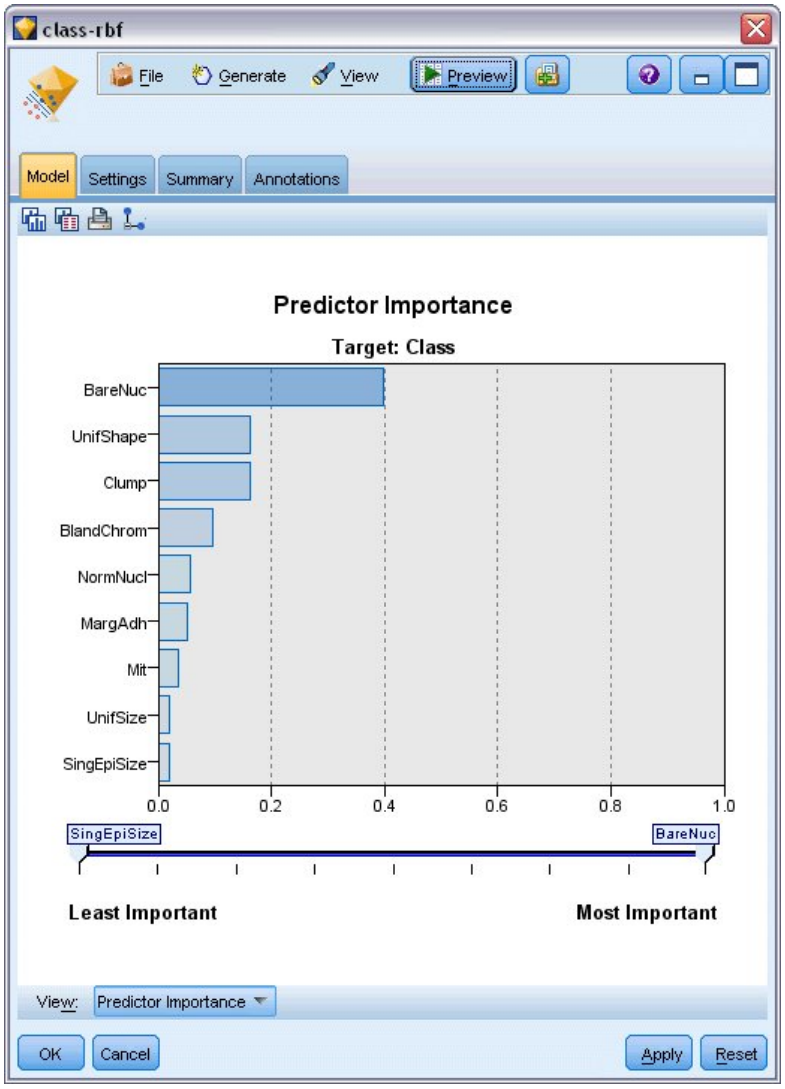
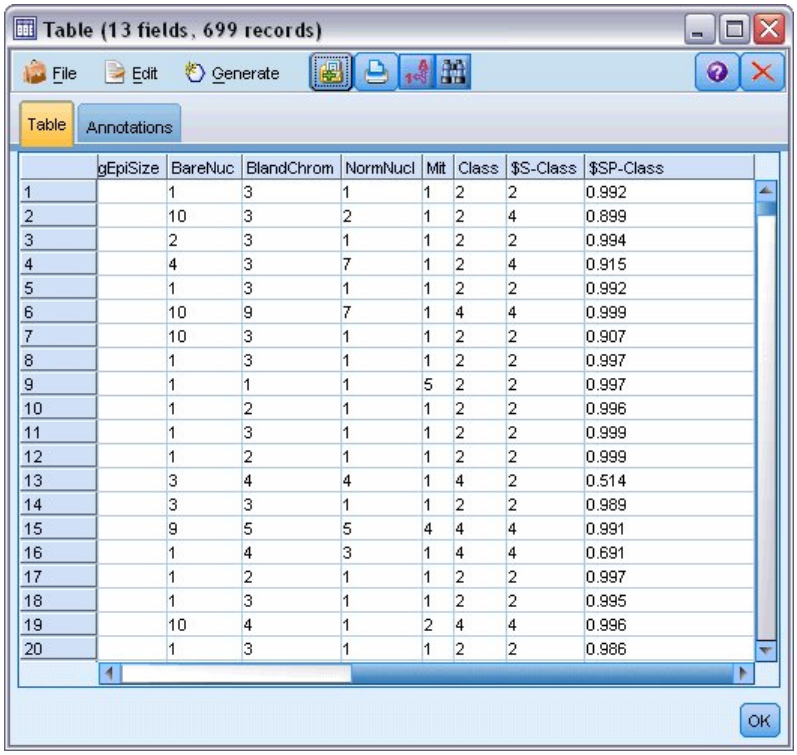


圖 *333.* 預測工具重要性圖形

在「模型」標籤上，「預測工具重要性」圖形顯示在預測各種欄位時的相對效果。這會向我們顯示 *BareNuc* 很容易產生最大影響，而 *UnifShape* 和 *Clump* 也相當重要。

1. 按一下**確定**。
2. 將「表格」節點連接至 *class-rbf* 模型區塊。
3. 開啟「表格」節點並按一下**執行**。



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

圖 *334.* 針對預測和信賴度值新增的欄位

4. 模型已建立兩個額外的欄位。捲動至表格輸出的右側以查看它們：

新欄位名稱	**說明**
$S-Class	模型預測的類別 的值。
$SP-Class	此預測的傾向評分（此預測為 true 的可能性值為 0.0 到 1.0）。

只需查看表格便能看到大部分記錄的傾向評分（在 *$SP-Class* 直欄中）相當的高。

但是存在一些顯著的例外；例如第 14 行的 1041801，其中的值 0.514 低得難以接受。此外，比較 *Class* 與 *$S-Class*，清楚地得知此模型進行了很多錯誤的預測，即使傾向評分相當的高也是如此（例如 第 2 行和第 4 行）。

我們來看看選擇不同的函數類型會不會取得更好的結果。

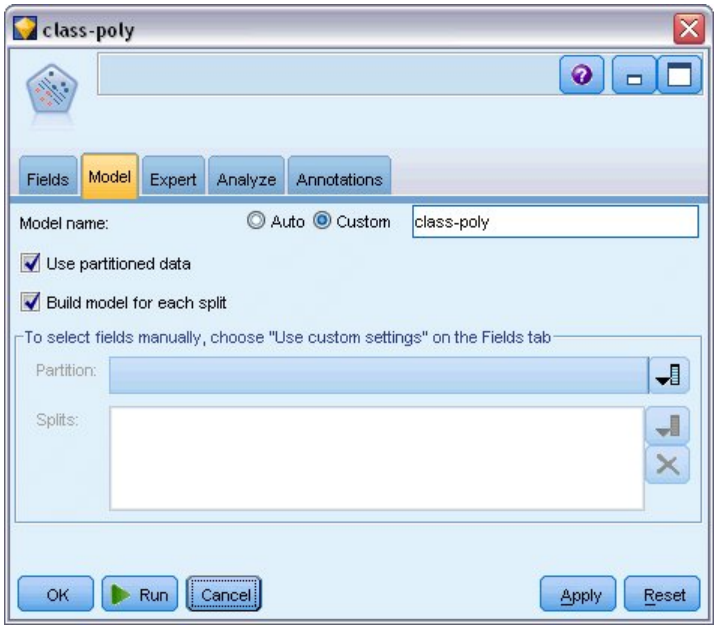


圖 *335.* 為模型設定新名稱

1. 關閉「表格」輸出視窗。
2. 將第二個 SVM 建模節點連接至「類型」節點。
3. 開啟新 SVM 節點。
4. 在**模型**標籤上，選擇「自訂」並輸入 *class-poly* 作為模型名稱。

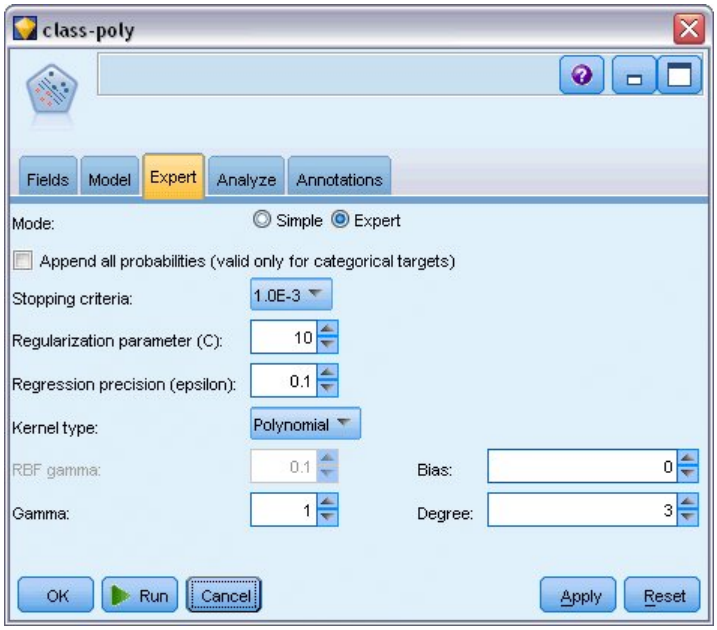


圖 *336.* 多項式的專家標籤設定

5. 在**專家**標籤上，將**模式**設為**專家**。

6. 將**核心類型**設為**多項式**並按一下**執行**。*class-poly* 模型區塊即會新增到串流中以及畫面右上方的「模型」選用區中。
7. 將 *class-rbf* 模型區塊連接至 *class-poly* 模型區塊（在警告對話框中選擇**取代**）。
8. 將「表格」節點連接至 *class-poly* 區塊。
9. 開啟「表格」節點並按一下**執行**。

比較結果

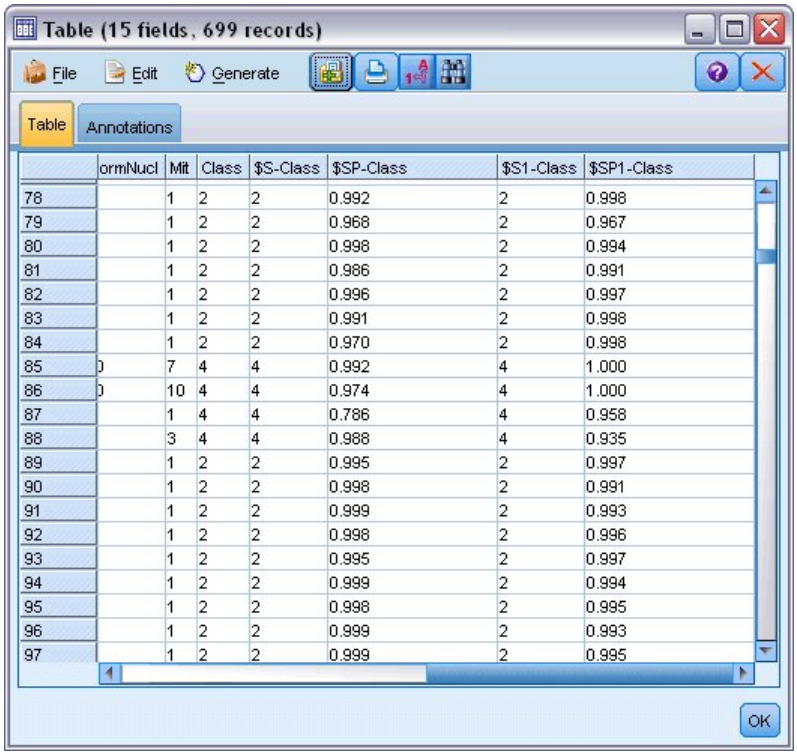


圖 *337.* 針對多項式函數新增的欄位

1. 捲動至表格輸出的右側以查看新增的欄位。

   針對多項式函數類型產生的欄位名稱為 *$S1-Class* 和 *$SP1-Class*。

   多項式的結果要好得多。許多傾向評分為 0.995 或更佳，讓我們倍受鼓舞。
2. 若要確認對模型的改進，請將「分析」節點連接至 *class-poly* 模型區塊。

開啟「分析」節點並按一下**執行**。

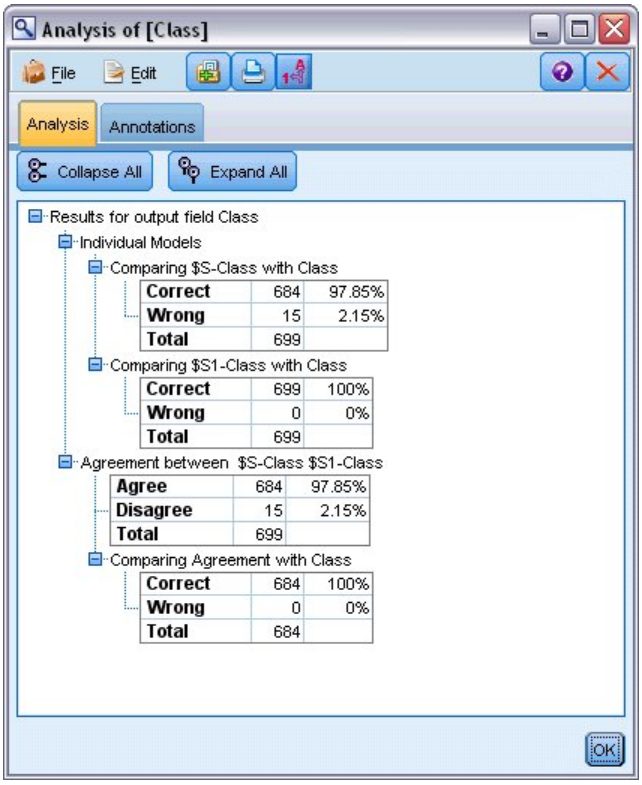


圖 *338.* 分析節點

具有「分析」節點的這項技術可讓您比較同一類型的兩個以上模型區塊。「分析」節點中的輸出顯示 RBF 函數正確預測 97.85% 的案例，這個結果仍然相當不錯。但是，輸出顯示多項式函數已在每個單一案例中正確預測診斷。實際上，您不可能看到 100% 的正確性，但是您可以使用「分析」節點來協助判定模型的正確性對於您的特定應用程式而言是否可以接受。

事實上，在這個特定的資料集上，其他函數類型（Sigmoid 和線性）的執行效能都比不上多項式函數。但是，對於不同的資料集，結果很可能是不同的，因此一律值得嘗試全範圍的選項。

摘要

您已使用不同類型的 SVM 核心函數來透過多個屬性預測分類。您已瞭解不同的核心如何針對相同的資料集提供不同的結果，以及如何測量一個模型對另一個模型的改進。

第 26 章 使用 Cox 迴歸為客戶流失時間建模

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，以及各種其他欄位。

此範例使用串流 *telco_coxreg.str*，其參照資料檔 *telco.sav*。資料檔位於 *Demos* 資料夾中，串流檔位於 *streams* 子資料夾中。請參閱第 4 頁的『Demos 資料夾』主題，以取得更多資訊。

建置適合的模型

1. 新增指向 *Demos* 資料夾中 *telco.sav* 的「統計量檔案」來源節點。

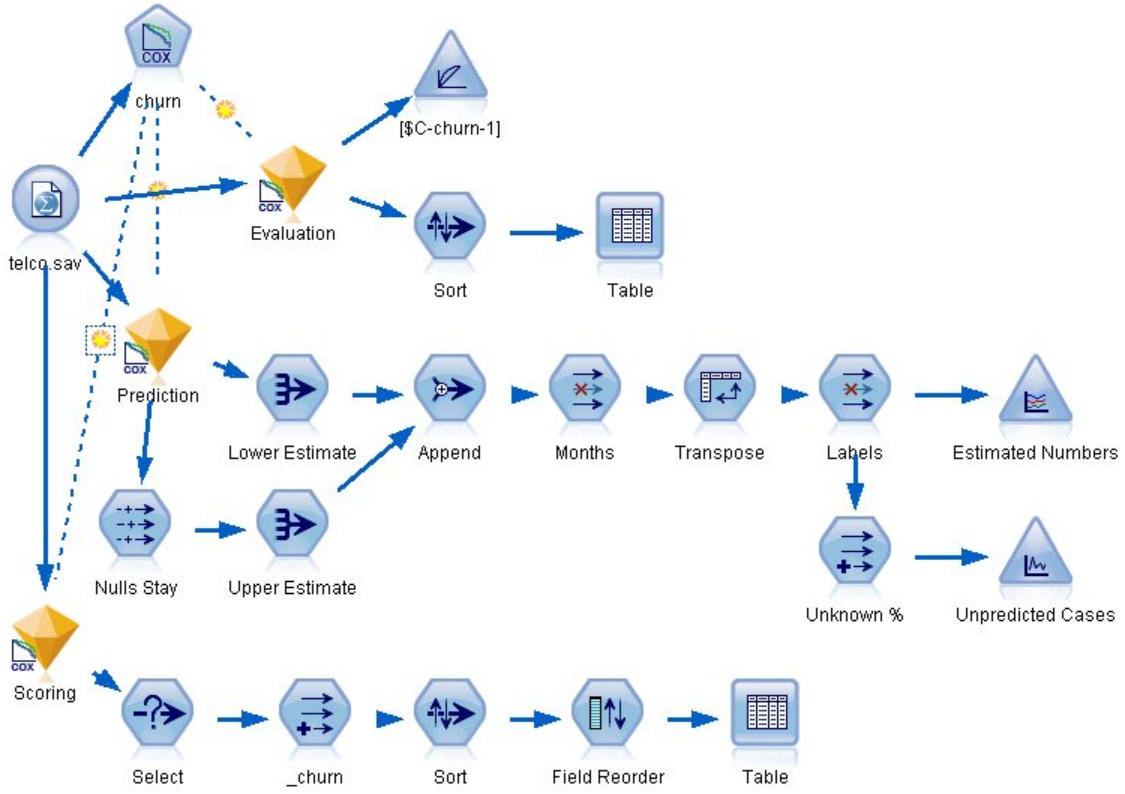


圖 *339.* 用來分析流失時間的樣本串流

2. 在來源節點的「過濾器」標籤上，排除欄位 *region*、*income*、*longten* 到 *wireten*，以及 *loglong* 到 *logwire*。

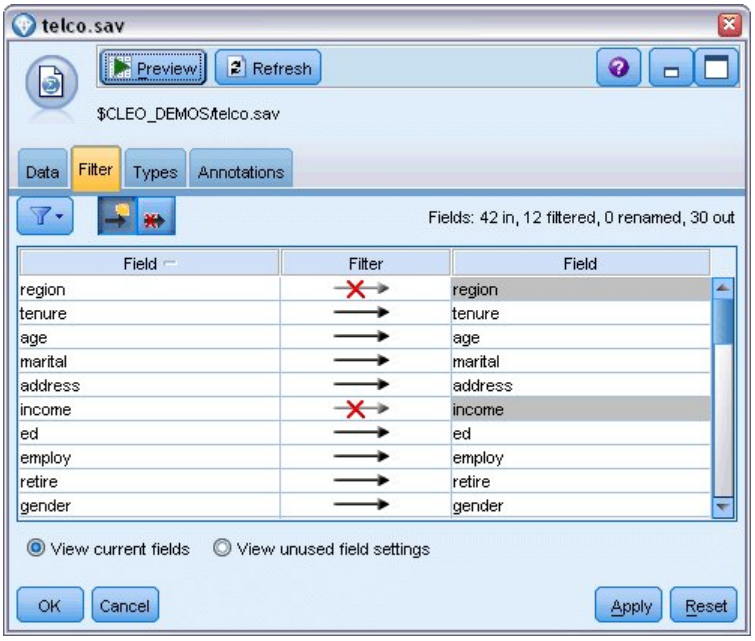


圖 *340.* 過濾不需要的欄位

（或者，您可以將「類型」標籤上這些欄位的角色變更為**無**而不是排除它，或選取要在建模節點中使用的欄位。）

3. 在來源節點的「類型」標籤上，將 *churn* 欄位的角色設定為**目標**，並將其測量層次設定為**旗標**。所有其他欄位都應該將其角色設定為**輸入**。
4. 按一下**讀取值**以將資料實例化。

圖 *341.* 設定欄位角色

5. 將 Cox 節點連接至來源節點；在**欄位**標籤中，選取 *tenure* 作為存活時間變數。

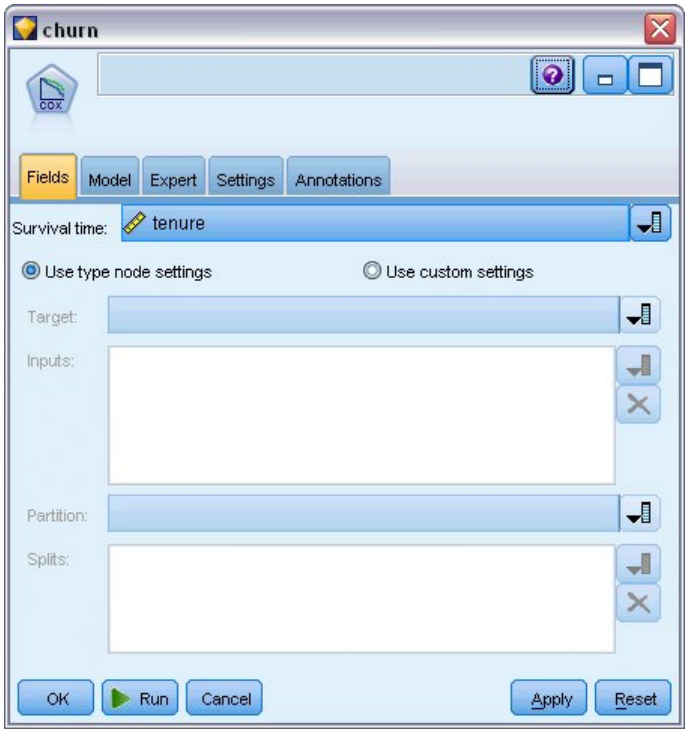


圖 *342.* 選擇欄位選項

6. 按一下**模型**標籤。
7. 選取**逐步迴歸分析**作為變數選取方法。

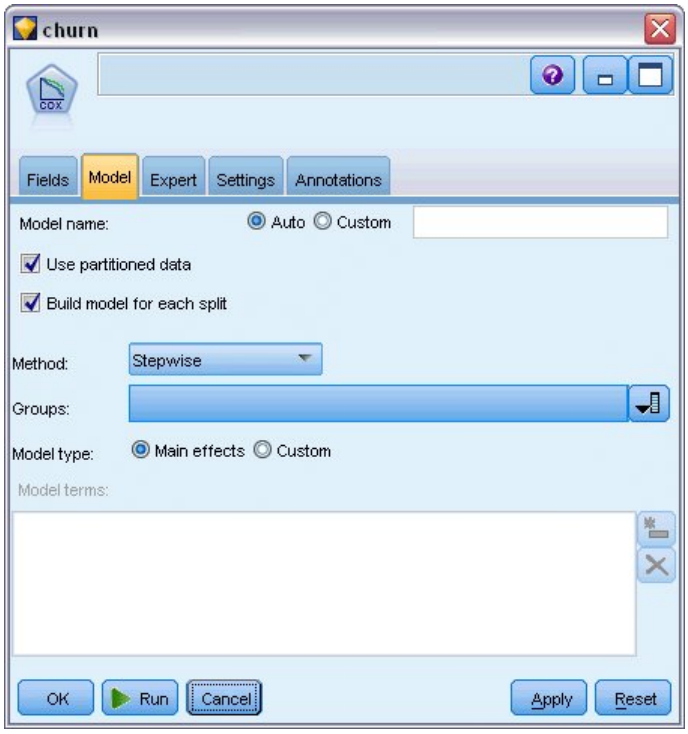


圖 *343.* 選擇模型選項

8. 按一下**專家**標籤並選取**專家**以啟動專家建模選項。
9. 按一下**輸出**。

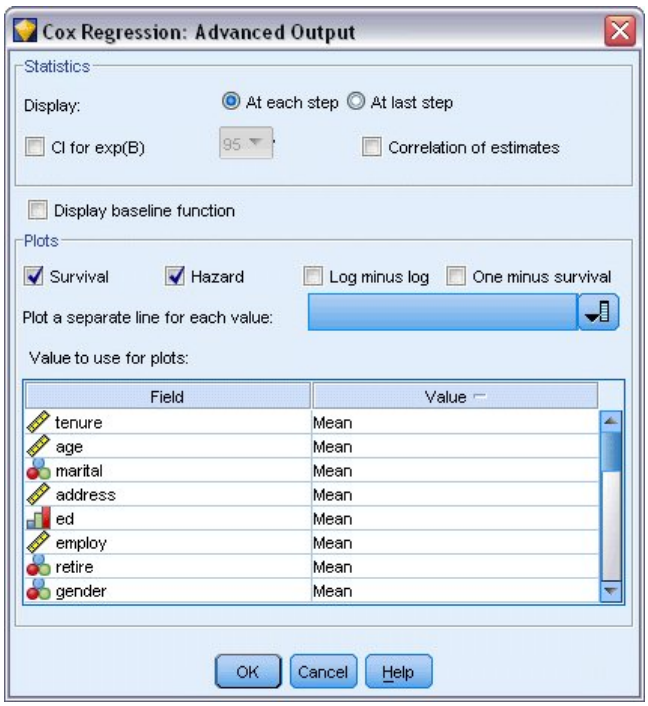


圖 *344.* 選擇進階輸出選項

10. 選取**存活**及**風險**作為要產生的圖形，然後按一下**確定**。
11. 按一下**執行**以建立模型區塊，該區塊會新增至串流，以及右上角的「模型」選用區。若要檢視其詳細資料，請按兩下串流上的區塊。首先，看一下「進階」輸出標籤。

受限觀察值

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	0.0%
	Cases with negative time	0	0.0%
	Censored cases before the earliest event in a stratum	0	0.0%
	Total	0	0.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

圖 *345.* 觀察值處理摘要

狀態變數可識別給定觀察值是否發生了事件。如果事件尚未發生，則觀察值被說成受限。受限觀察值不會在計算迴歸係數時使用，但會用來計算基準線風險。觀察值處理摘要顯示 726 個觀察值受限。這些是未流失的客戶。

種類變數編碼

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

圖 *346.* 種類變數編碼

種類變數編碼是有用的參照，可用來解譯種類共變數，特別是二分變數的迴歸係數。依預設，參照種類是「最後一個」種類。因此，例如，即使已婚客戶在資料檔中的變數值為 1，為了進行迴歸，仍將其編碼為 0。

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

圖 *347.* 綜合測試

模型建置處理程序採用逐步向前演算法。綜合測試可測量模型的效能。上一步的卡方變更是 –2 模型對數概似值在上一步與現行步驟之間的差異。如果該步驟要新增變數且變更顯著性小於 0.05，則可以併入。如果該步驟要移除變數且變更顯著性大於 0.10，則可以排除。在十二個步驟中，十二個變數會新增至模型。

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	address	-.035	.009	14.543	1	.000	.966
	employ	-.051	.010	25.767	1	.000	.950
	reside	-.103	.046	5.037	1	.025	.902
	equip	-1.948	.381	26.180	1	.000	.143
	callcard	.777	.151	26.451	1	.000	2.175
	longmon	-.233	.022	115.619	1	.000	.792
	equipmon	-.042	.011	15.377	1	.000	.959
	multline	.612	.145	17.854	1	.000	1.844
	voice	-.501	.157	10.197	1	.001	.606
	internet	-.362	.160	5.114	1	.024	.697
	callid	-.464	.148	9.790	1	.002	.629
	ebill	-.399	.156	6.557	1	.010	.671

圖 *348.* 方程式中的變數（僅第 *12* 步）

最終模型包括 *address*、*employ*、*reside*、*equip*、*callcard*、*longmon*、*equipmon*、*multline*、*voice*、*internet*、*callid* 及 *ebill*。若要瞭解個別預測值的效果，請查看指數 Exp(B)，可以將其解譯為預測值中增長一個單位的風險預測變更。

- *address*的 Exp(B) 值表示客戶在相同的位址每居住一年，流失風險減少 100%–(100%×0.966)=3.4%。在相同位址居住五年的客戶的流失風險會減少 $100\%–(100\%\times0.966^5)=15.88\%$。
- *callcard* 的 Exp(B) 值表示未訂閱呼叫卡服務的客戶的流失風險為訂閱該服務客戶的 2.175 倍。從種類變數編碼重新呼叫 *No* = 1 進行迴歸。

- *internet* 的 Exp(B) 值表示未訂閱網際網路服務的客戶的流失風險為訂閱該服務客戶的 0.697 倍。這一點有些令人擔憂，因為這表示使用該服務的客戶會比不使用該服務的客戶更快地離開公司。

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	ed	6.328	4	.176
	ed(1)	.007	1	.934
	ed(2)	.203	1	.652
	ed(3)	.835	1	.361
	ed(4)	5.773	1	.016
	retire	.013	1	.908
	gender	.214	1	.644
	tollfree	3.243	1	.072
	wireless	.668	1	.414
	tollmon	.000	1	.987
	cardmon	3.163	1	.075
	wiremon	1.084	1	.298
	pager	1.808	1	.179
	callwait	.266	1	.606
	forward	2.201	1	.138
	confer	2.568	1	.109
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

圖 *349.* 不在模型中的變數（僅第 *12* 步）

離開模型之變數的分數統計量顯著性值全都大於 0.05。但 *tollfree* 和 *cardmon* 的顯著性值雖然不小於 0.05，卻相當接近。這些值可能非常有趣，值得進一步研究。

共變數平均數

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
ed(1)	.204
ed(2)	.287
ed(3)	.209
ed(4)	.234
employ	10.987
retire	.953
gender	.483
reside	2.331
tollfree	.526
equip	.614
callcard	.322
wireless	.704
longmon	11.723
tollmon	13.274
equipmon	14.220
cardmon	13.781
wiremon	11.584
multline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

圖 *350.* 共變數平均數

此表格會顯示每一個預測值變數的平均值。此表格在查看針對平均值而建構的存活圖形時是有用的參照。但請注意，當您查看種類預測值之指標變數的平均數時，「平均」客戶實際不存在。即使使用所有尺度預測值，您也不可能找到共變數值全部接近平均數的客戶。如果要查看特定觀察值的存活曲線，則可以變更共變數值，存活曲線在「圖形」對話框中的這些值處繪製。如果要查看特定觀察值的存活曲線，則可以變更共變數值，存活曲線在「進階輸出」對話框之「圖形」群組中的這些值處繪製。

存活曲線

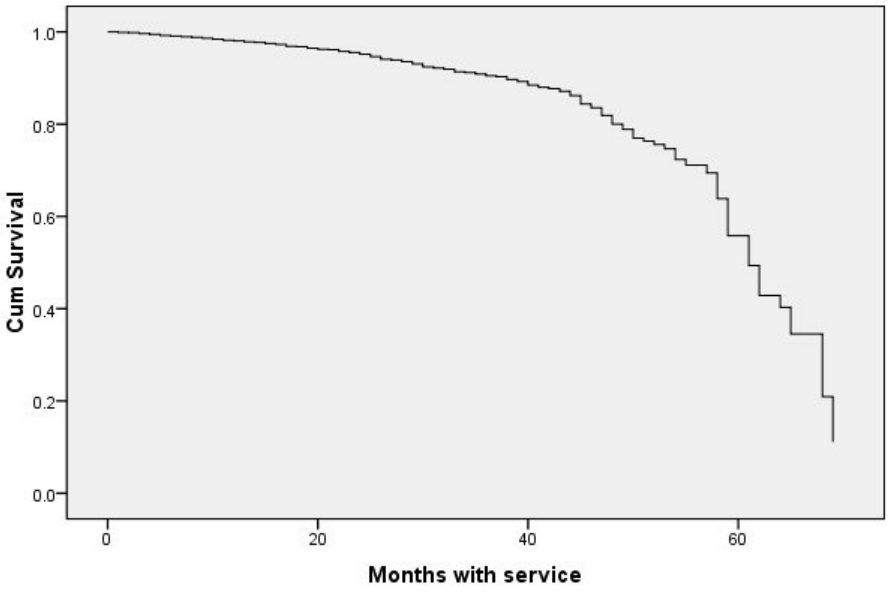


圖 *351.* 「平均」客戶的存活曲線

基本存活曲線是「平均」客戶之模型預測流失時間的視覺化顯示。水平軸顯示事件發生時間。垂直軸顯示存活機率。因此，存活曲線上的任何點都會顯示「平均」客戶在超過該時間保留客戶的機率。過去 55 個月，存活曲線變得不太平滑。在這麼長時間內與公司合作的客戶減少了，可用的資訊也隨之減少，因此曲線為塊狀。

風險曲線

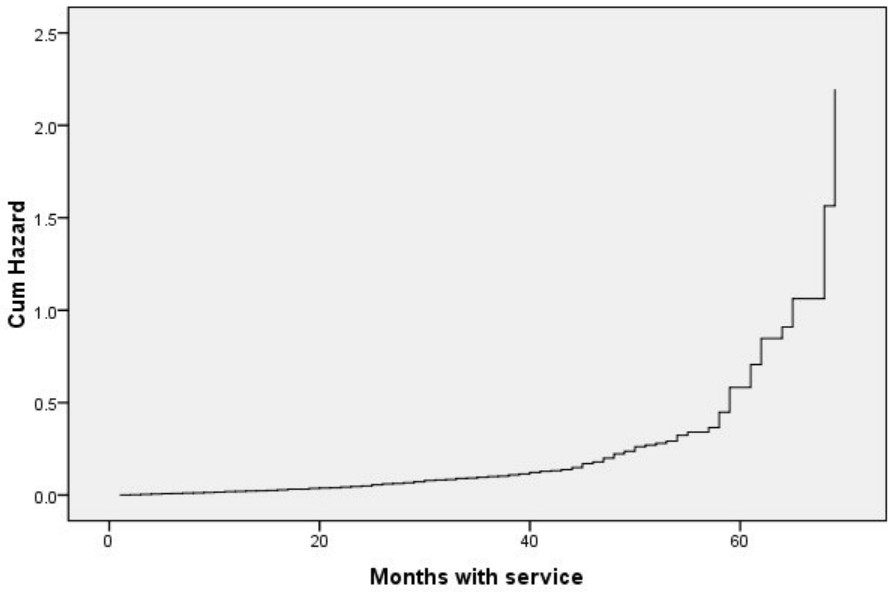


圖 *352.* 「平均」客戶的風險曲線

基本風險曲線是對「平均」客戶流失可能性之累積模型預測的視覺化顯示。水平軸顯示事件發生時間。垂直軸顯示累積風險，等於存活機率的負對數。過去的 55 個月，風險曲線像存活曲線一樣，出於相同的原因變得不太平滑。

評估

逐步選取方法可保證您的模型僅具有「統計顯著性」預測值，但不保證模型實際預測目標時表現良好。若要做到這一點，您需要分析評分的記錄。

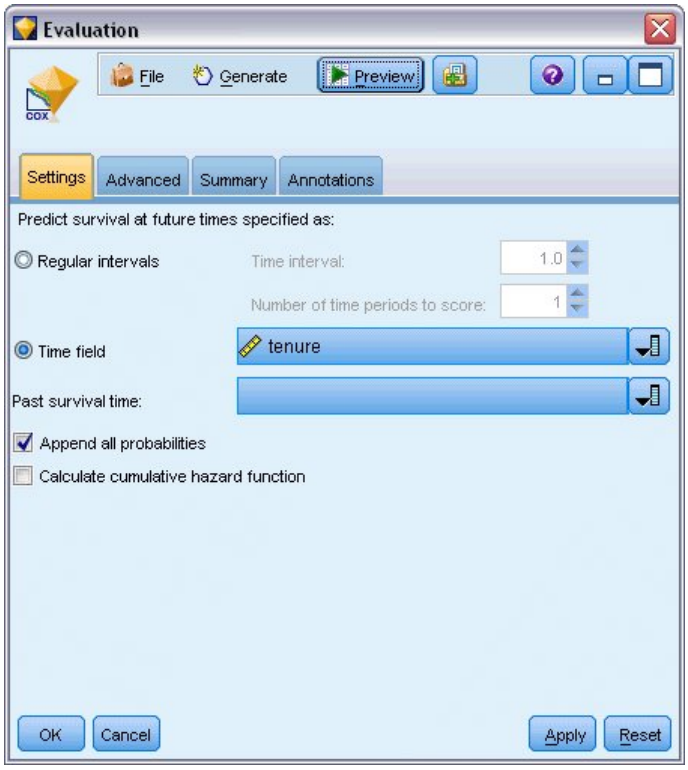


圖 *353. Cox* 區塊：設定標籤

1. 將模型區塊放在畫布上，將其連接至來源節點，開啟區塊並按一下「設定」標籤。
2. 選取**時間欄位**並指定 *tenure*。每一個記錄都對其保有期長短進行評分。
3. 選取**附加所有機率**。

   這會使用 0.5 作為客戶是否流失的分割值建立評分；如果流失傾向大於 0.5，則會被評為流失者。這個數字並沒有什麼神奇之處，其他分割值可能會產生更需要的結果。針對考慮選擇分割值的一種方式，請使用「評估」節點。

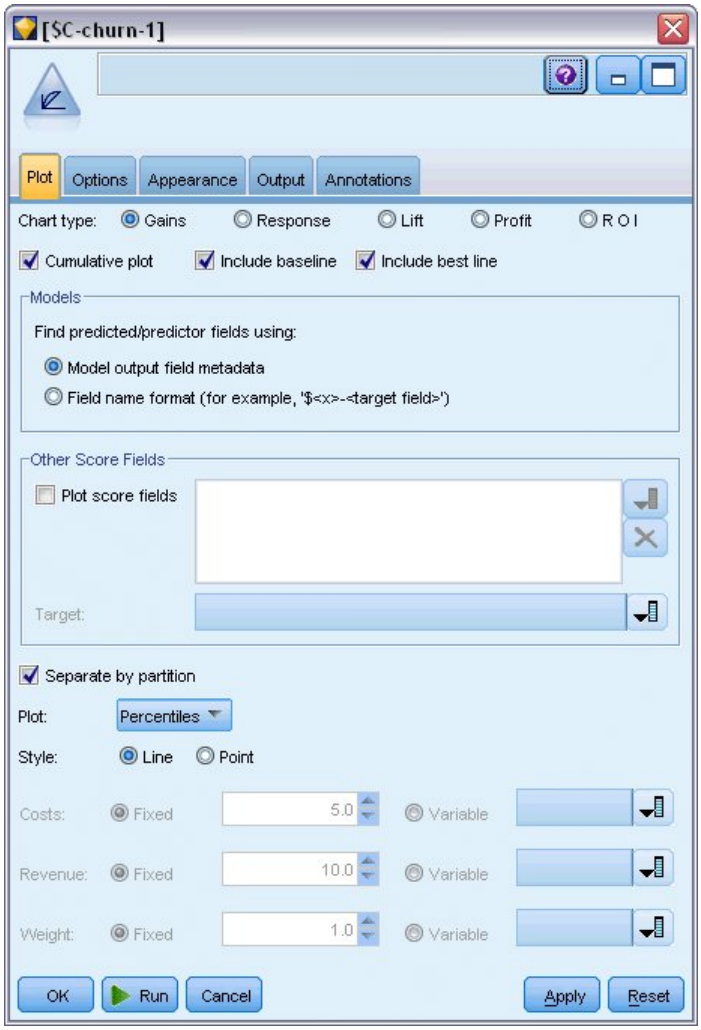


圖 *354.* 評估節點：圖形標籤

4. 將「評估」節點連接至模型區塊；在「圖形」標籤上，選取**含最佳線**。
5. 按一下「**選項**」標籤。

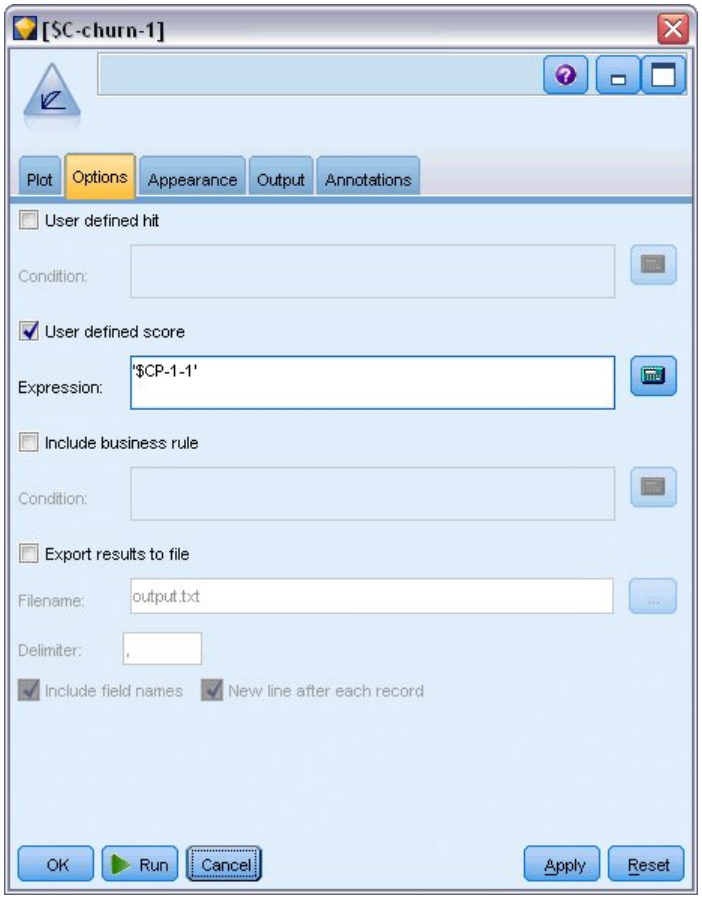


圖 *355.* 評估節點：選項標籤

6. 選取**使用者定義的分數**並鍵入 `'$CP-1-1'` 作為表示式。這是模型產生的欄位，對應於流失傾向。
7. 按一下「**執行**」。

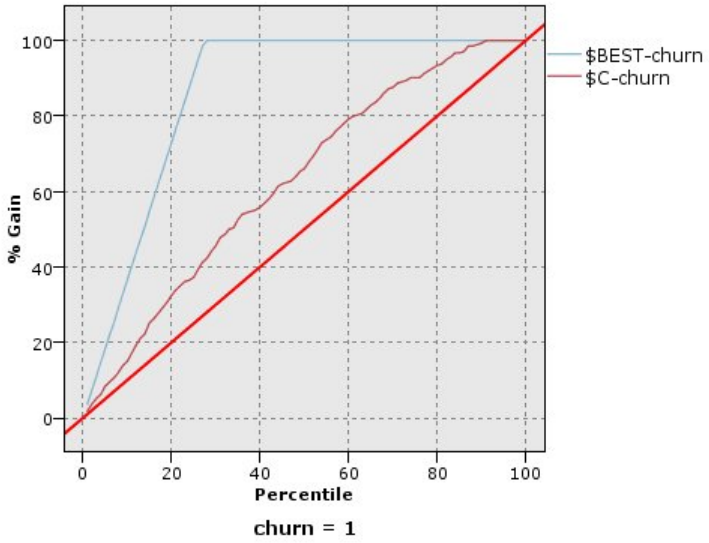


圖 *356.* 增益圖

累積收益圖以觀察值總數的百分比為目標，顯示指定類別「增益」中觀察值的總數百分比。例如，曲線上的一個點位於 (10%, 15%)，表示如果使用模型為資料集評分，並依據預測的流失傾向將所有觀察值排序，則預期前 10% 會包含約 15% 的所有觀察值，實際採取種類 *1*（流失者）。同樣，前 60% 包含約 79.2% 的流失者。如果您選取 100% 的評分資料集，則會取得資料集中的所有流失者。

對角線是「基準線」曲線；如果您從評分資料集中隨機選取 20% 的記錄，則會預期「增益」約 20% 的所有記錄，實際採取種類 *1*。曲線位於基準線上方的位置越遠，增益越大。「最佳」線顯示「完美」模型的曲線，為每個流失者指定的流失傾向評分比非流失者高。您可以使用累積收益圖選擇對應於所需增益的百分比，然後將此百分比對應至適當的分割值，以協助選擇類別分割值。

「所需」增益的內容為何，需視類型一和類型二錯誤的成本而定。即，將流失者分類為非流失者（類型一）的成本是多少？將非流失者分類為流失者（類型二）的成本是多少？如果客戶保留是主要問題，則應該降低類型一錯誤；在累積增益圖上，這可能對應於對預測傾向為 *1* 的前 60% 的客戶增加客戶關懷，這會擷取 79.2% 的可能流失者，但可能會花費時間和資源來獲得新客戶。如果降低維護現行客戶群的成本是優先事項，則應降低類型二錯誤。在圖表上，這可能對應於對前 20% 增加客戶關懷，這會擷取 32.5% 的流失者。通常兩件事都很重要，因此不得不選擇決策規則將客戶分類，該規則會提供靈敏度及特定性的最佳組合。

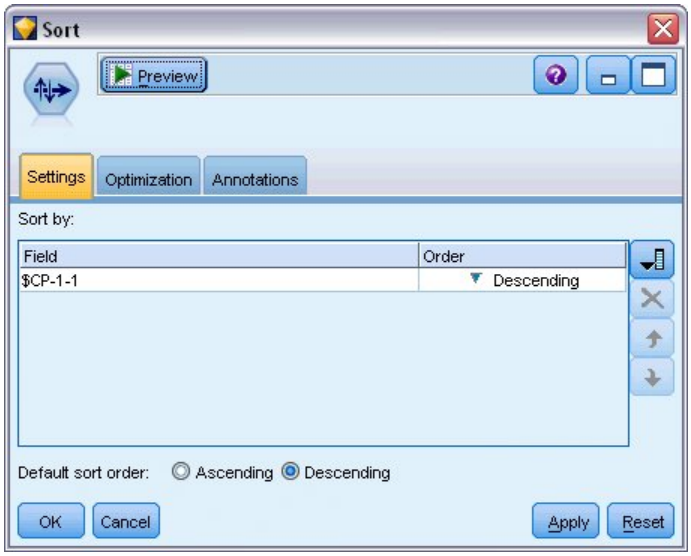
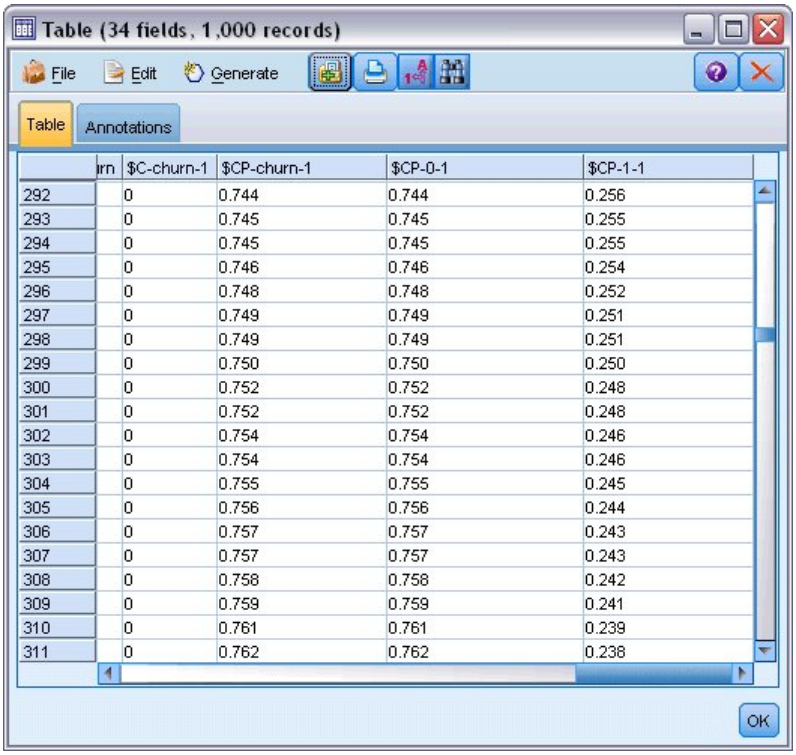


圖 *357.* 排序節點：設定標籤

8. 假設您已決定 45.6% 是需要的增益，這對應於採取前 30% 的記錄。若要尋找適當的分類分割值，請將「排序」節點連接至模型區塊。
9. 在「設定」標籤上，選擇依 *$CP-1-1* 降序排序，然後按一下**確定**。



	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292	0	0.744	0.744	0.256
293	0	0.745	0.745	0.255
294	0	0.745	0.745	0.255
295	0	0.746	0.746	0.254
296	0	0.748	0.748	0.252
297	0	0.749	0.749	0.251
298	0	0.749	0.749	0.251
299	0	0.750	0.750	0.250
300	0	0.752	0.752	0.248
301	0	0.752	0.752	0.248
302	0	0.754	0.754	0.246
303	0	0.754	0.754	0.246
304	0	0.755	0.755	0.245
305	0	0.756	0.756	0.244
306	0	0.757	0.757	0.243
307	0	0.757	0.757	0.243
308	0	0.758	0.758	0.242
309	0	0.759	0.759	0.241
310	0	0.761	0.761	0.239
311	0	0.762	0.762	0.238

圖 *358.* 表格

10. 將「表格」節點連接至「排序」節點。
11. 開啟「表格」節點，然後按一下**執行**。

向下捲動輸出，您會看到第 300 個記錄的 *$CP-1-1* 值是 0.248。使用 0.248 作為分類分割值應會導致約 30% 的客戶評分為流失者，擷取約 45% 的實際客戶總計。

追蹤保留的預期客戶數

對模型滿意後，您應該追蹤今後兩年內在資料集中保留的預期客戶數。空值即總保有期（未來時間 + *tenure*）超出資料中用來訓練模型之存活時間範圍的客戶，是個有趣的挑戰。有一種處理空值的方式是建立兩組預測，一組假設空值已流失，另一組則假設已保留。透過這種方式，您可以建立預期保留客戶數的上限及下限。

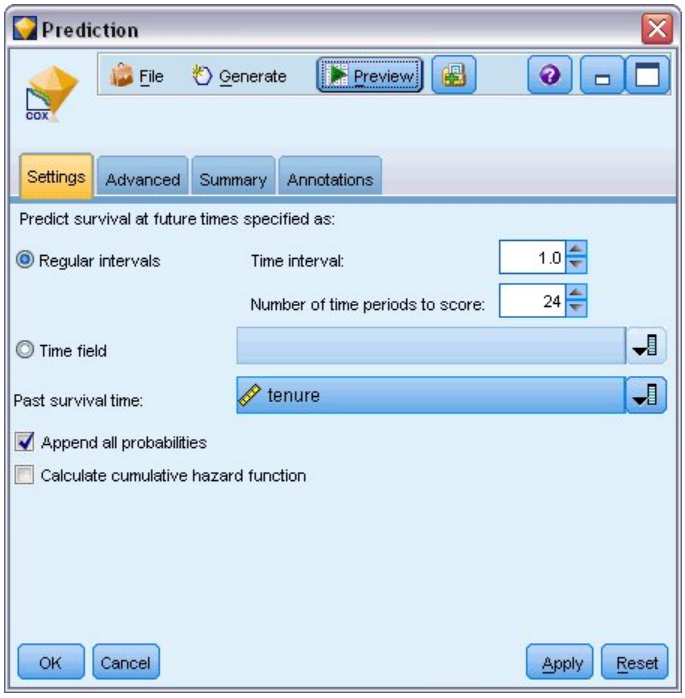


圖 *359. Cox* 區塊：設定標籤

1. 在「模型」選用區中按兩下模型區塊（或將區塊複製並貼上串流畫布）並將新區塊連接至來源節點。
2. 開啟區塊的「設定」標籤。
3. 請確定**定期**已選取，並指定 1.0 作為時間間隔，指定 24 作為週期數進行評分。這會指定在接下來 24 個月的每個月對每一個記錄進行評分。
4. 選取 *tenure* 作為用來指定過去存活時間的欄位。評分演算法將考慮每一個客戶作為公司客戶的時間長度。
5. 選取**附加所有機率**。

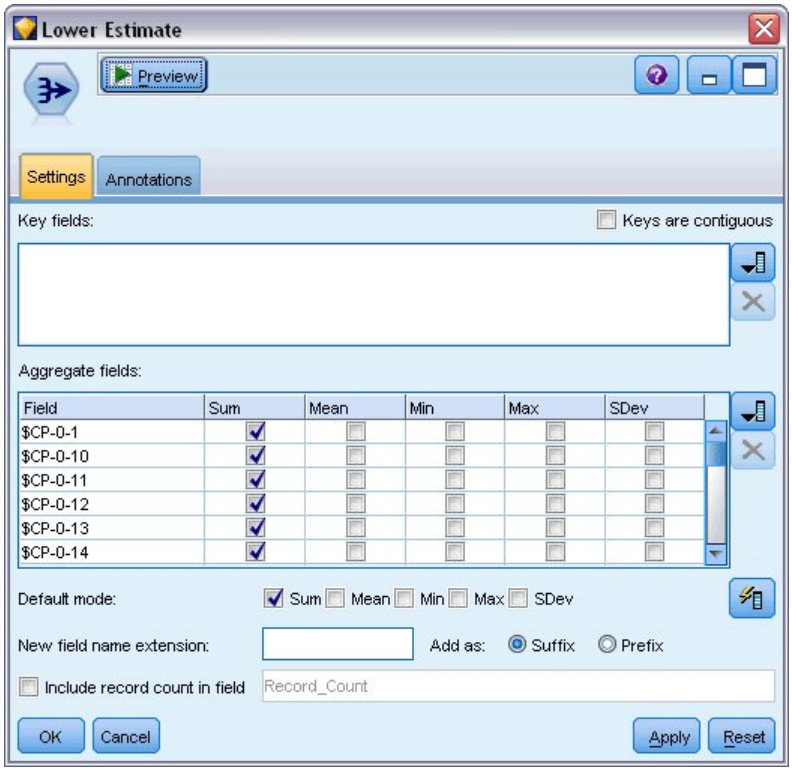


圖 *360.* 聚集節點：設定標籤

6. 將「聚集」節點連接至模型區塊；在「設定」標籤上，取消選取**平均數**作為預設模式。
7. 選取形式為 *$CP-0-n* 的欄位 *$CP-0-1* 到 *$CP-0-24*，作為要聚集的欄位。如果在「選取欄位」對話框上依據名稱（亦即按字母順序）排序欄位，那麼這是最簡單的。
8. 取消選取**在欄位中包含記錄計數**。
9. 按一下**「確定」**。此節點會建立「下限」預測。

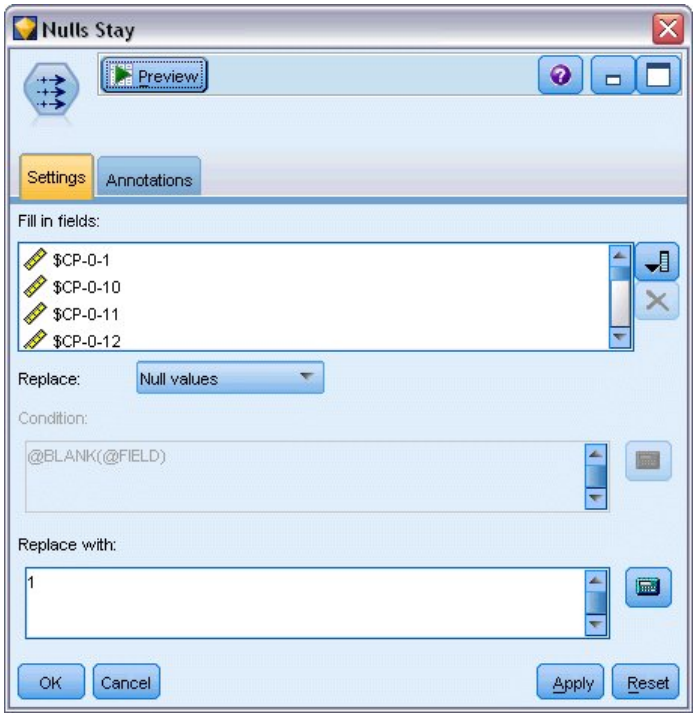


圖 *361.* 填入器節點：設定標籤

10. 將「填入器」節點連接至剛剛連接了「聚集」節點的 Coxreg 區塊；在「設定」標籤上，選取形式為 *$CP-0-n* 的欄位 *$CP-0-1* 到 *$CP-0-24*，作為要填入的欄位。如果在「選取欄位」對話框上依據名稱（亦即按字母順序）排序欄位，那麼這是最簡單的。
11. 選擇用值 1 取代**空值**。
12. 按一下**「確定」**。

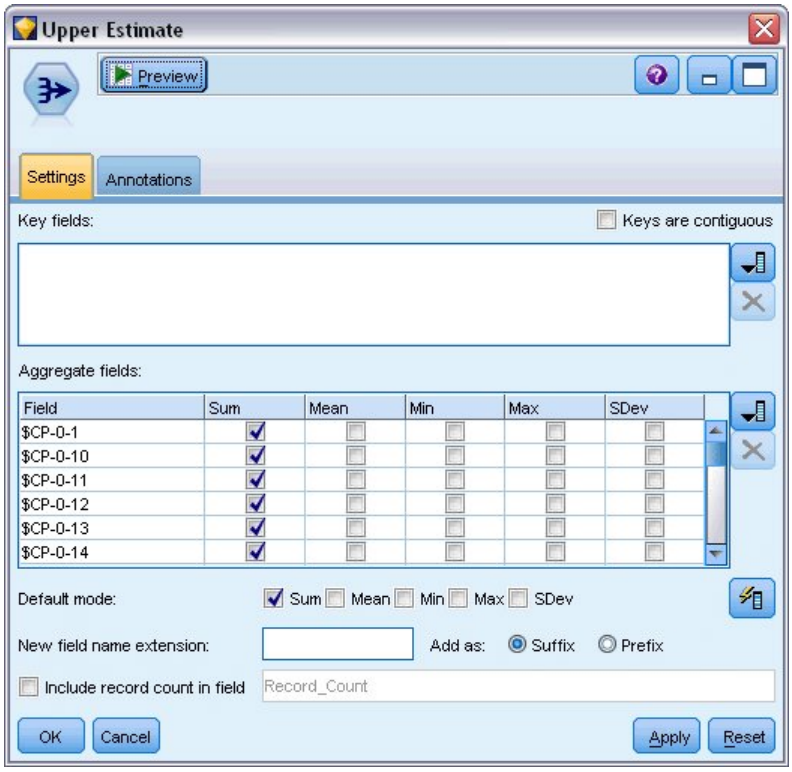


圖 *362.* 聚集節點：設定標籤

13. 將「聚集」節點連接至「填入器」節點；在「設定」標籤上，取消選取**平均數**作為預設模式。
14. 選取形式為 *$CP-0-n* 的欄位 *$CP-0-1* 到 *$CP-0-24*，作為要聚集的欄位。如果在「選取欄位」對話框上依據名稱（亦即按字母順序）排序欄位，那麼這是最簡單的。
15. 取消選取**在欄位中包含記錄計數**。
16. 按一下**「確定」**。此節點會建立「上限」預測。

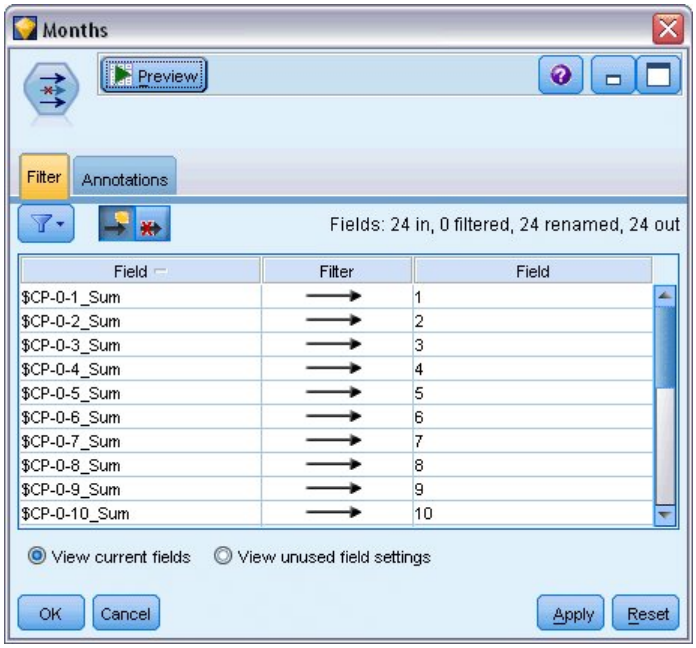


圖 *363.* 過濾器節點：設定標籤

17. 將「附加」節點連接至兩個「聚集」節點，然後將「過濾器」節點連接至「附加」節點。
18. 在「過濾器」節點的「設定」標籤上，將欄位重新命名為 *1* 到 *24*。透過使用「跳至」節點，這些欄位名稱會變成下游圖表中 *x* 軸的值。

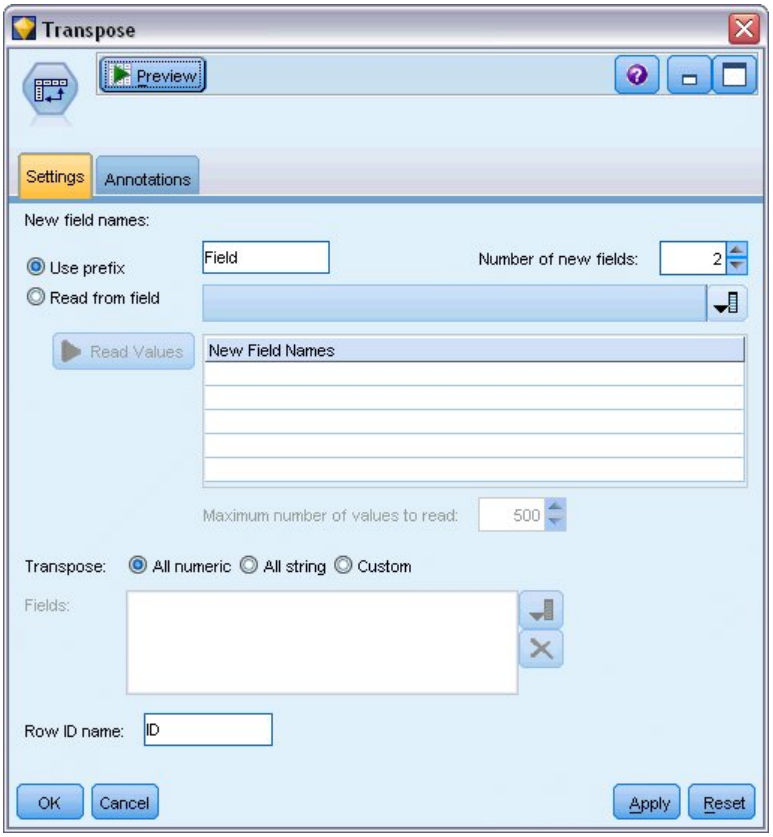


圖 *364.* 跳至節點：設定標籤

19. 將「跳至」節點連接至「過濾器」節點。
20. 鍵入 2 作為新欄位數。

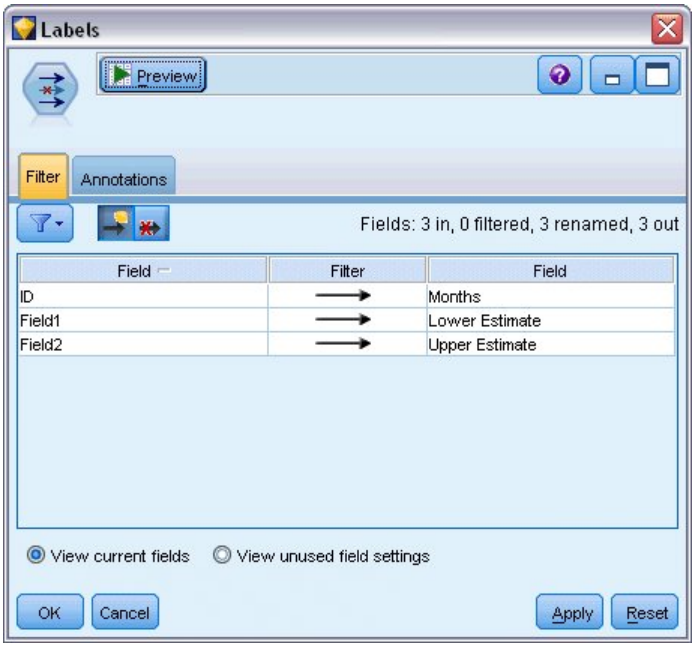


圖 *365.* 過濾器節點：過濾器標籤

21. 將「過濾器」節點連接至「跳至」節點。
22. 在「過濾器」節點的「設定」標籤上，將 *ID* 重新命名為月，將 *Field1* 重新命名為估計值下限，將 *Field2* 重新命名為估計值上限。

圖 *366.* 多圖節點：圖形標籤

23. 將「多圖」節點連接至「過濾器」節點。
24. 在「圖形」標籤上，月作為 X 欄位，估計值下限及估計值上限作為 Y 欄位。

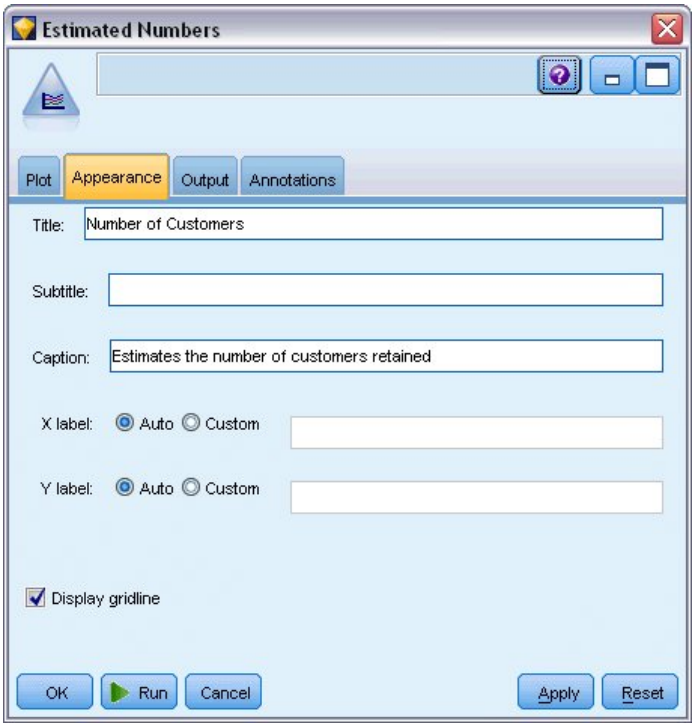


圖 *367.* 多圖節點：外觀標籤

25. 按一下「外觀」標籤。
26. 鍵入客戶數作為標題。
27. 鍵入估計保留的客戶數作為標題。
28. 按一下「**執行**」。

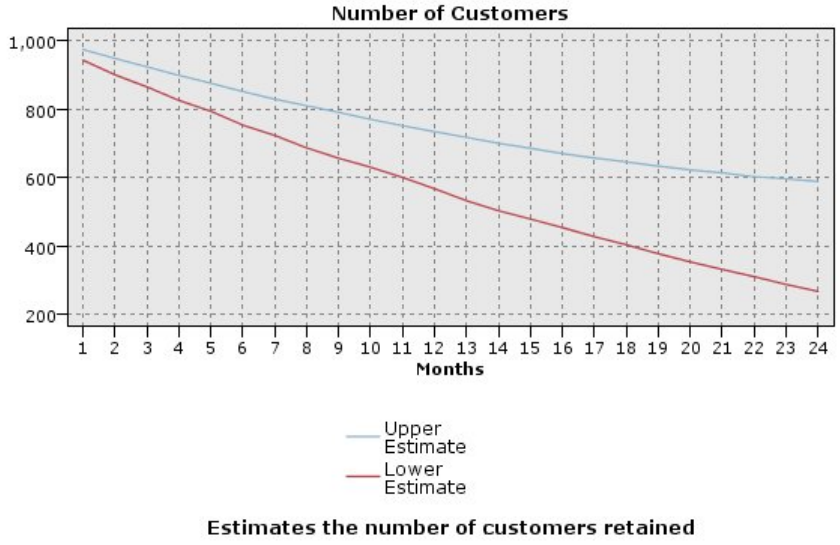


圖 *368.* 估計保留客戶數的多圖

將繪製估計保留客戶數的上限及下限。兩條線之間的差額是評分為空值的客戶數，因此其狀態高度不確定。隨著時間的推移，這些客戶數會增加。 12 個月之後，您可以預期在資料集中保留 601 到 735 位原

始客戶；24 個月之後則保留 288 到 597 位。

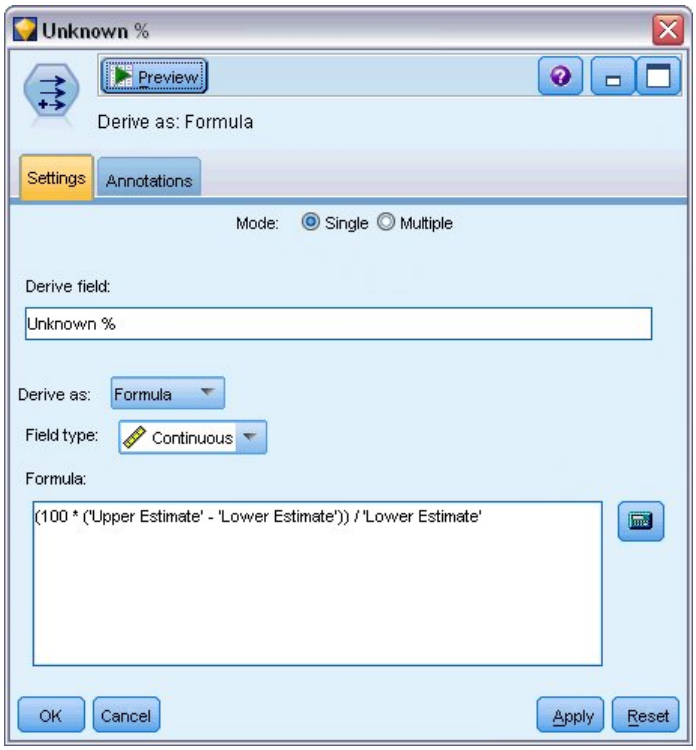


圖 *369.* 衍生節點：設定標籤

29. 若要再看一下保留客戶數估計值有多麼不確定，請將「衍生」節點連接至「過濾器」節點。
30. 在「衍生」節點的「設定」標籤上，鍵入不明 %作為衍生欄位。
31. 選取**連續**作為欄位類型。
32. 鍵入 `(100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate'` 作為公式。不明 % 是「不確定」客戶數佔估計值下限的百分比。
33. 按一下「**確定**」。

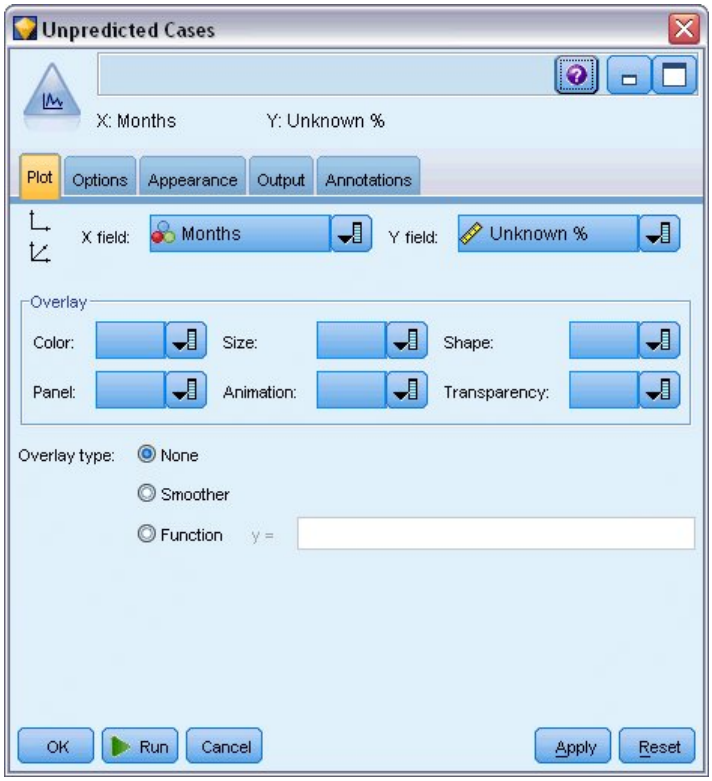


圖 *370.* 多圖節點：圖形標籤

34. 將「圖形」節點連接至「衍生」節點。
35. 在「圖形」節點的「圖形」標籤上，選取月作為 X 欄位，選取不明 % 作為 Y 欄位。
36. 按一下**外觀**標籤。

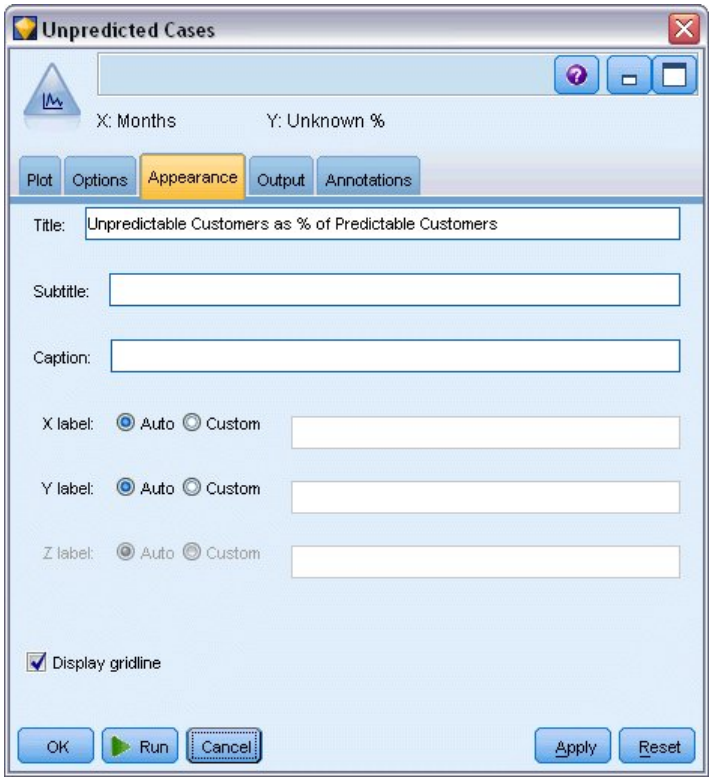


圖 *371.* 圖形節點：外觀標籤

37. 鍵入無法預期客戶數佔可預期客戶數的百分比作為標題。
38. 執行節點。

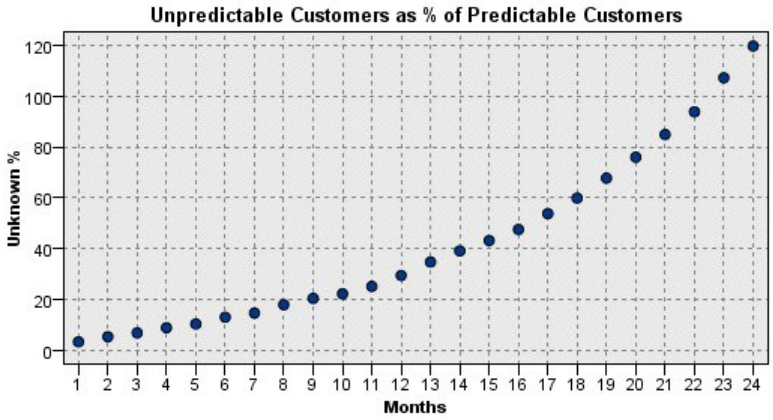


圖 *372.* 無法預期客戶數的圖形

在第一年中，無法預期客戶的百分比基本上以線性速率增長，但增長率在第二年期間激增，直到第 23 個月，具有空值的客戶數會超過保留的預期客戶數。

評分

對模型滿意後，您應該對客戶評分以按季度識別最有可能在下一年內流失的個體。

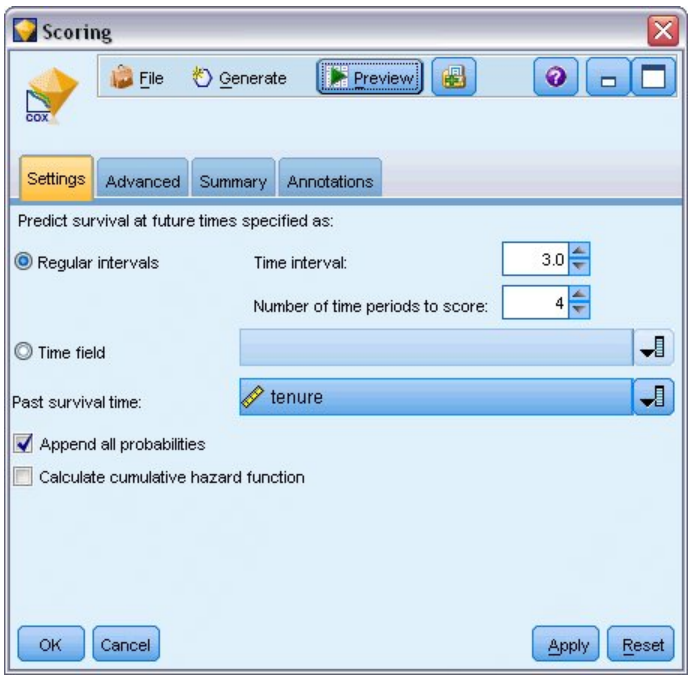


圖 *373. Coxreg* 區塊：設定標籤

1. 將第三個模型區塊連接至來源節點並開啟模型區塊。
2. 請確定**定期**已選取，並指定 `3.0` 作為時間間隔，指定 `4` 作為週期數進行評分。這會指定在接下來的四個季度對每一個記錄進行評分。
3. 選取 *tenure* 作為用來指定過去存活時間的欄位。評分演算法將考慮每一個客戶作為公司客戶的時間長度。
4. 選取**附加所有機率**。透過這些額外的欄位可以更輕鬆地將記錄排序以在表格中檢視。

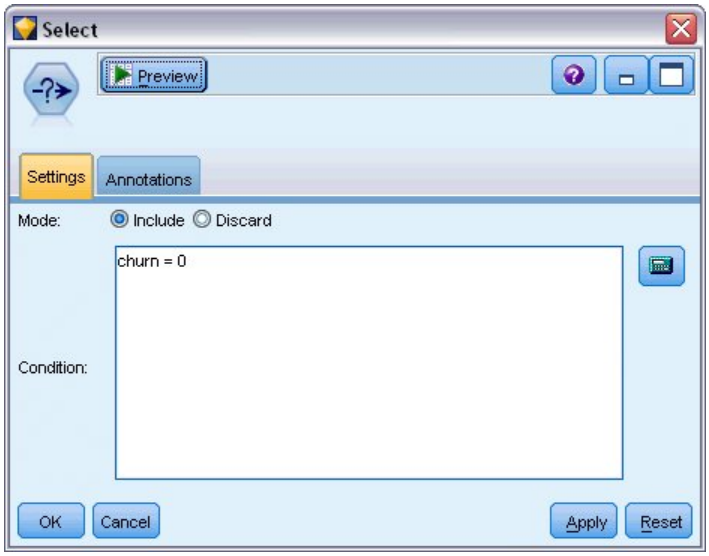


圖 *374.* 選取節點：設定標籤

5. 將「選取」節點連接至模型區塊；在「設定」標籤上，鍵入 `churn=0` 作為條件。這會從結果表中移除已流失的客戶。

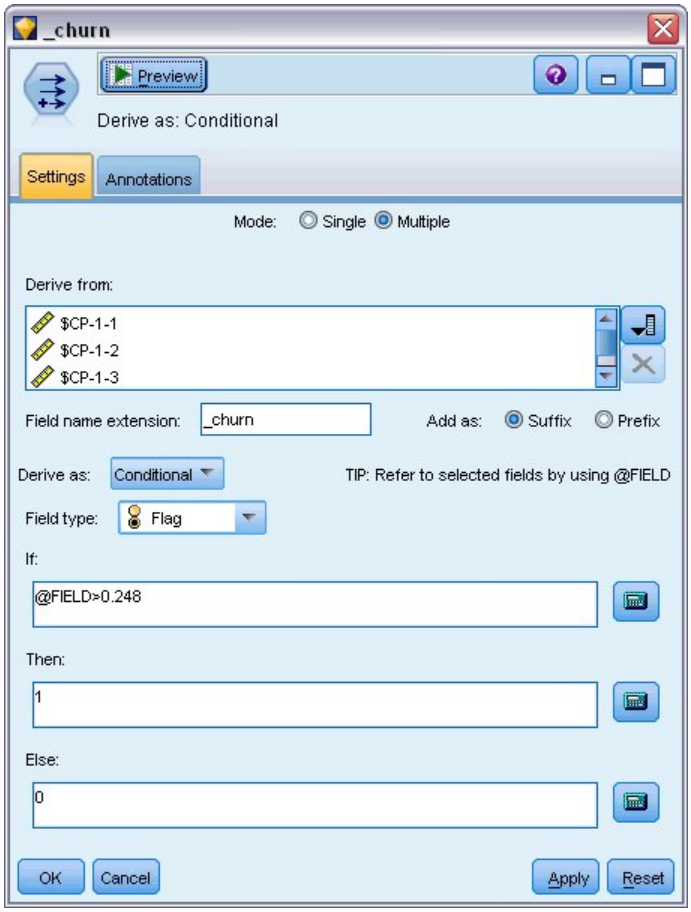


圖 *375.* 衍生節點：設定標籤

6. 將「衍生」節點連接至「選取」節點；在「設定」標籤上，選取**多個**作為模式。
7. 選擇從形式為 *$CP-1-n* 的欄位 *$CP-1-1* 到 *$CP-1-4* 衍生，並鍵入 `_churn` 作為要新增的字尾。如果在「選取欄位」對話框上依據名稱（亦即按字母順序）排序欄位，那麼這是最簡單的。
8. 選擇將欄位衍生為**條件式**。
9. 選取**旗標**作為測量層次。
10. 鍵入 `@FIELD>0.248` 作為 **If** 條件。請記住，這是評估期間識別的分類分割值。
11. 鍵入 `1` 作為 **Then** 表示式。
12. 鍵入 `0` 作為 **Else** 表示式。
13. 按一下「**確定**」。

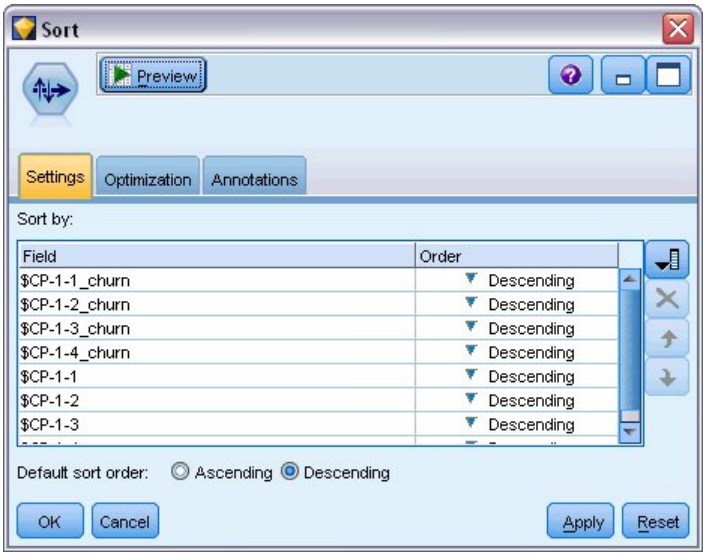
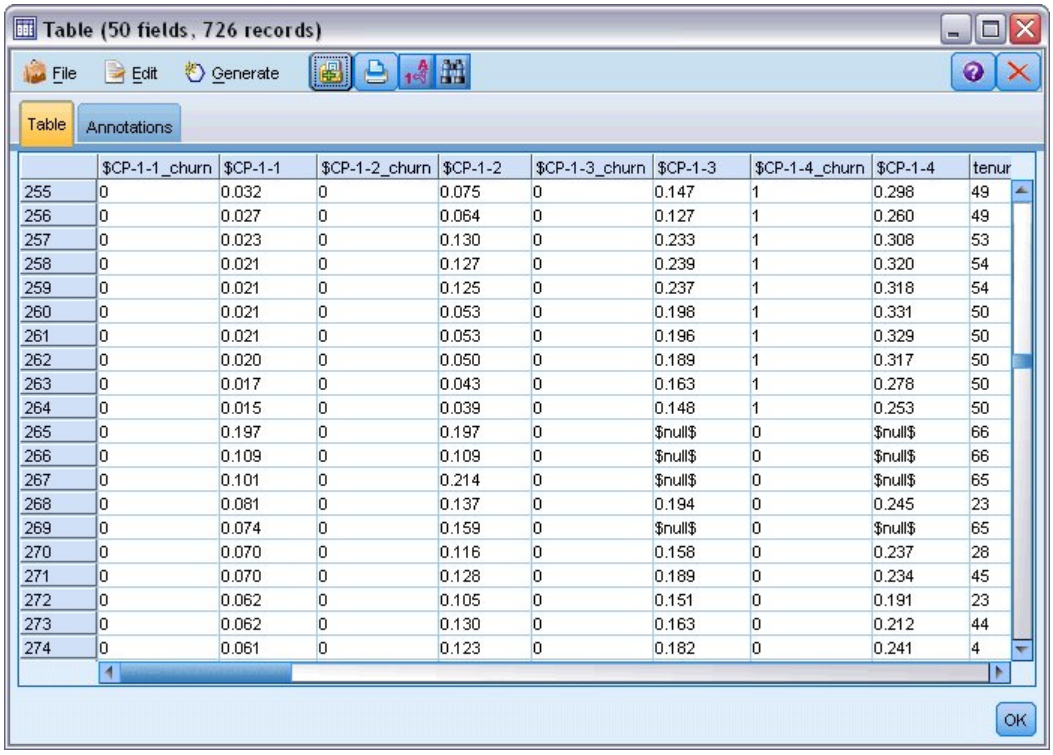


圖 *376.* 排序節點：設定標籤

14. 將「排序」節點連接至「衍生」節點；在「設定」標籤上，選擇按 *$CP-1-1_churn* 到 *$CP-1-4-churn*，然後按 *$CP-1-1* 到 *$CP-1-4* 排序，全部按降序排列。預測流失的客戶會出現在頂部。

圖 *377.* 欄位重新排序節點：重新排序標籤

15. 將「欄位重新排序」節點連接至「排序」節點；在「重新排序」標籤上，選擇將 *$CP-1-1_churn* 到 *$CP-1-4* 放在其他欄位前面。這樣只是會讓結果表格更易於讀取，因此為選用。您將需要使用按鈕來將欄位移入圖中顯示的位置。

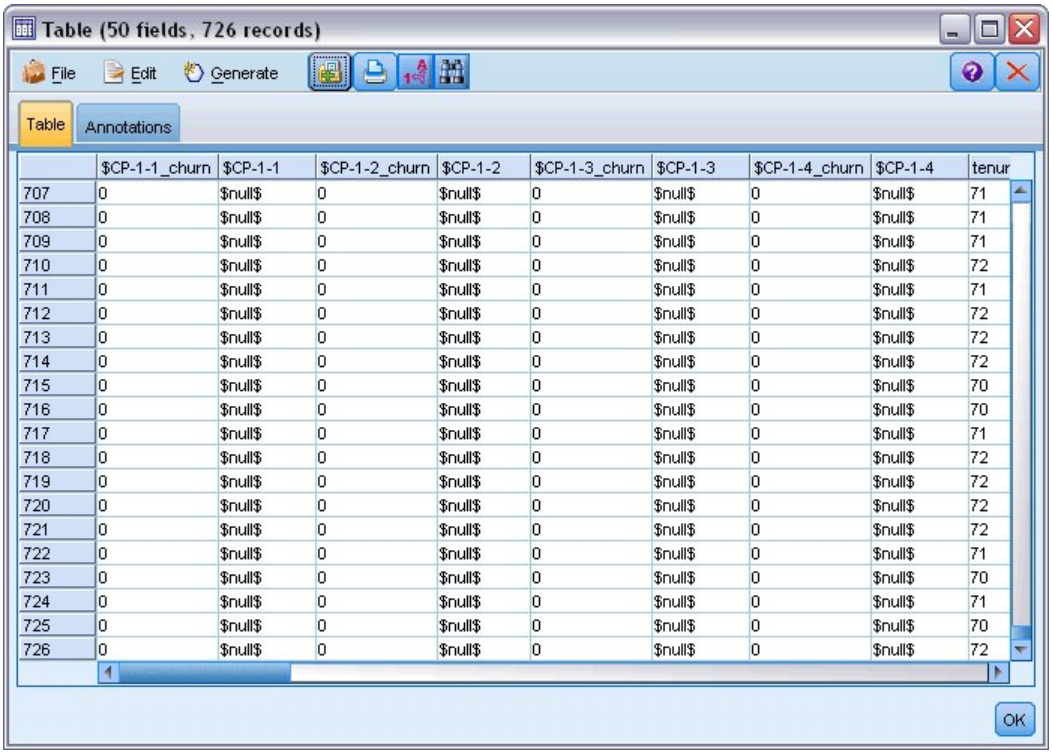


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

圖 *378.* 顯示客戶分數的表格

16. 將「表格」節點連接至「欄位重新排序」節點並執行它。

預期會有 264 個客戶在年底流失，184 個在第三季度末流失，103 個在第二季度末流失，31 個在第一季度末流失。請注意，假設有兩個客戶，在第一季度流失傾向較高的客戶在後來幾個季度中的流失傾向不一定較高；例如，請參閱記錄 256 和 260。這可能是因為在客戶現行保有期之後幾個月內風險函數的形狀；例如，因促銷而加入的客戶在早期轉移的可能性比因他人推薦而加入的客戶更大，但如果他們不轉移，則可能在剩餘的保有期內更忠誠。您可能應該對客戶重新排序以取得最可能流失客戶的不同視圖。



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

圖 *379.* 顯示空值客戶的表格

表格底部是具有預測空值的客戶。這些客戶的總保有期（未來時間 + *tenure*）超出了資料中用來訓練模型的存活時間範圍。

摘要

使用 Cox 迴歸，您找到了一個可接受的流失時間模型，繪製了在今後兩年內保留的預期客戶數，並識別出今後一年內最可能流失的個別客戶。請注意，雖然這是可接受的模型，但可能不是最佳模型。理想情況下，您應至少將使用逐步向前法取得的此模型與使用逐步向後法建立的模型進行比較。

在 IBM SPSS Modeler 中使用的建模方法的數學基礎說明會在《*IBM SPSS Modeler* 演算法手冊》中列出。

第 27 章 購物籃分析（規則歸納/C5.0）

此範例處理的虛構資料說明了超市購物籃的內容（即一起購買的物品集合）以及相關聯的購買者個人資料，可透過尊榮卡計劃獲得。目標是為了探索購買相似產品的客戶群組，並以人口統計方式描述特徵（例如，年齡、收入等）。

此範例說明了資料採礦的兩個階段：

- 關聯規則建模和顯示購買物品之間鏈結的 Web 顯示
- 側寫已識別產品群組購買者的 C5.0 規則歸納

附註：此應用程式並未直接利用預測建模，因此產生的模型沒有精確度測量，並且在資料採礦處理程序中也沒有相關聯的訓練/測試區別。

此範例使用名為 *baskrule* 的串流，其參照的資料檔名為 *BASKETS1n*。這些檔案可從任何 IBM SPSS Modeler 安裝的 *Demos* 目錄取得。您可從 Windows「開始」功能表的 IBM SPSS Modeler 程式集存取。*baskrule* 檔案位於 *streams* 目錄中。

存取資料

使用「變數檔案」節點，連接至資料集 *BASKETS1n*，選取以從該檔案中讀取欄位名稱。將「類型」節點連接至資料來源，然後將該節點連接至「表格」節點。 將欄位 *cardid* 的測量層次設定為無類型（因為每一個尊榮卡 ID 僅在資料集中出現一次，因此可能對建模毫無用處）。選取名義作為 *sex* 欄位的測量層次（這是為了確保 Apriori 建模演算法不會將 *sex* 視為旗標）。

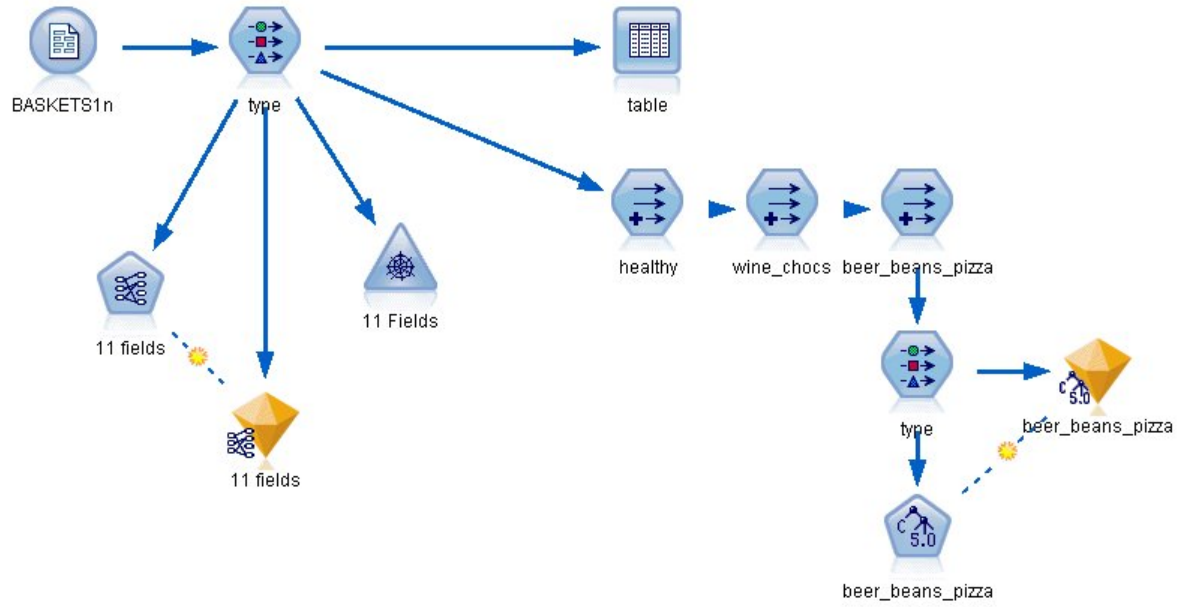


圖 *380. baskrule* 串流

現在執行該串流來實例化「類型」節點並顯示表格。該資料集包含 18 個欄位，每一個記錄代表一個購物籃。

18 個欄位會在下列標題中呈現。

購物籃摘要：

- *cardid*。購買此購物籃的客戶的尊榮卡 ID。
- *value*。購物籃的總採購價格。
- *pmethod*。購物籃的付款方式。

持卡人的個人詳細資料：

- *sex*
- *homeown*。持卡人是否為屋主。
- *income*
- *age*

購物籃內容 — 用來顯示產品種類的旗標：

- *fruitveg*
- *freshmeat*
- *dairy*
- *cannedveg*
- *cannedmeat*
- *frozenmeal*
- *beer*
- 葡萄酒
- *softdrink*
- *fish*
- *confectionery*

在購物籃內容中探索親緣性

首先，您需要使用 Apriori 來產生關聯規則，以對購物籃內容中的親緣性（關聯）有個全面的瞭解。透過編輯「類型」節點，並將所有產品種類的角色設定為兩者，將所有其他角色設定為無，選取要在此建模程序中使用的欄位。（兩者表示該欄位可以是所產生模型的輸入或輸出。）

附註：您可以設定多個欄位的選項，具體方法為按住 Shift 鍵並按一下來選取欄位，然後再從直欄中指定選項。

圖 *381.* 選取用於建模的欄位

指定用於建模的欄位之後，將 Apriori 節點連接至「類型」節點，編輯它，選取選項**僅包含旗標欄位的 true 值**，然後在 Apriori 節點上按一下執行。 結果，即位於管理程式視窗右上角之「模型」標籤上的模型，會包含您可以透過使用快速功能表及選取**瀏覽**來檢視的關聯規則。

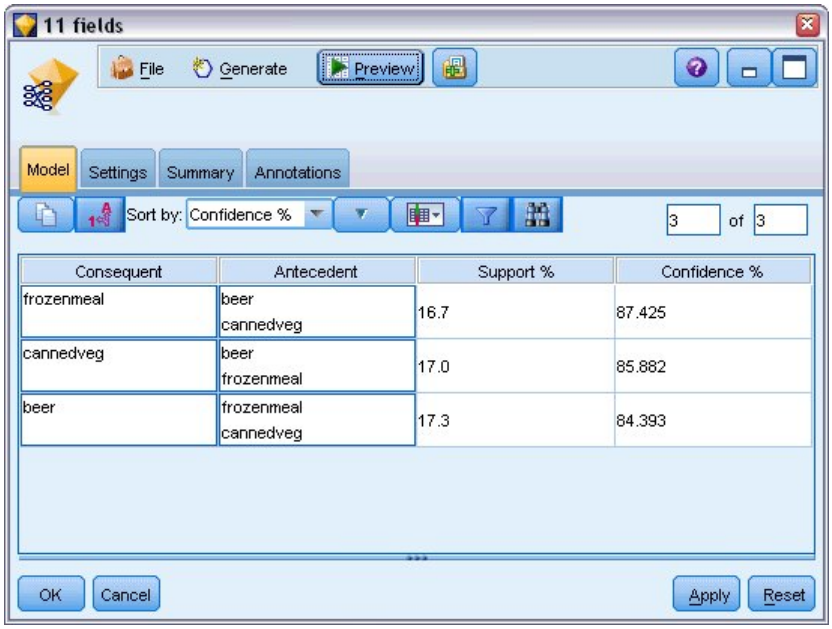


圖 *382.* 關聯規則

這些規則會顯示冷凍餐食、蔬菜罐頭及啤酒之間的各種關聯。出現雙向關聯規則，例如：

```
frozenmeal -> beer
beer -> frozenmeal
```

表示 Web 顯示（僅顯示雙向關聯）可能會強調顯示此資料中的部分型樣。

將 Web 節點連接至「類型」節點，編輯 Web 節點，選取所有購物籃內容欄位，選取**僅顯示 true 值旗標**，然後在 Web 節點上按一下執行。

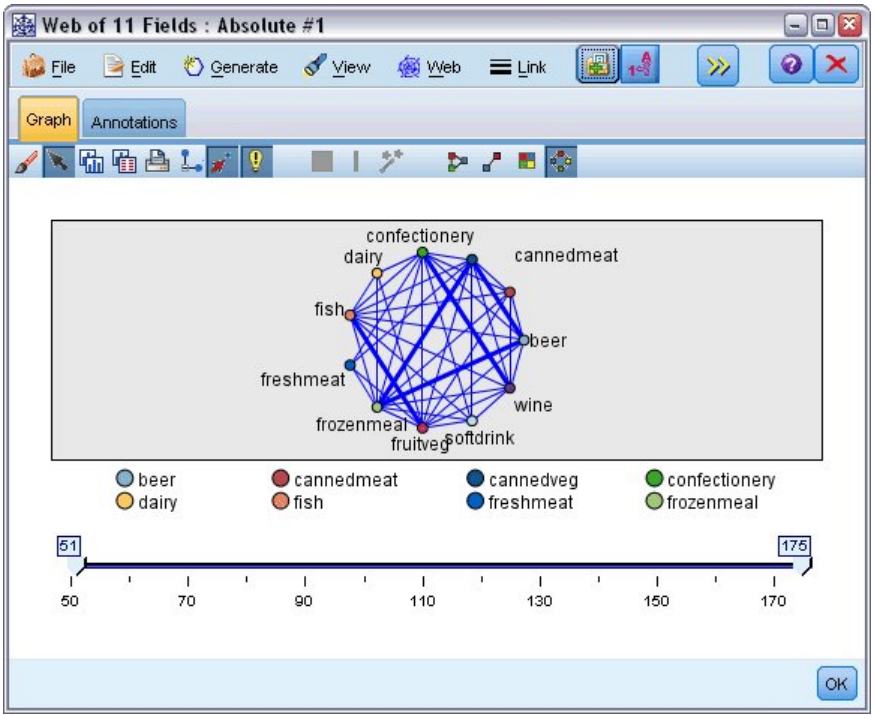


圖 *383.* 產品關聯的 *Web* 顯示

因為大部分產品種類組合在數個購物籃中出現，此 Web 上的強鏈結數太多，無法顯示模型建議的客戶群組。

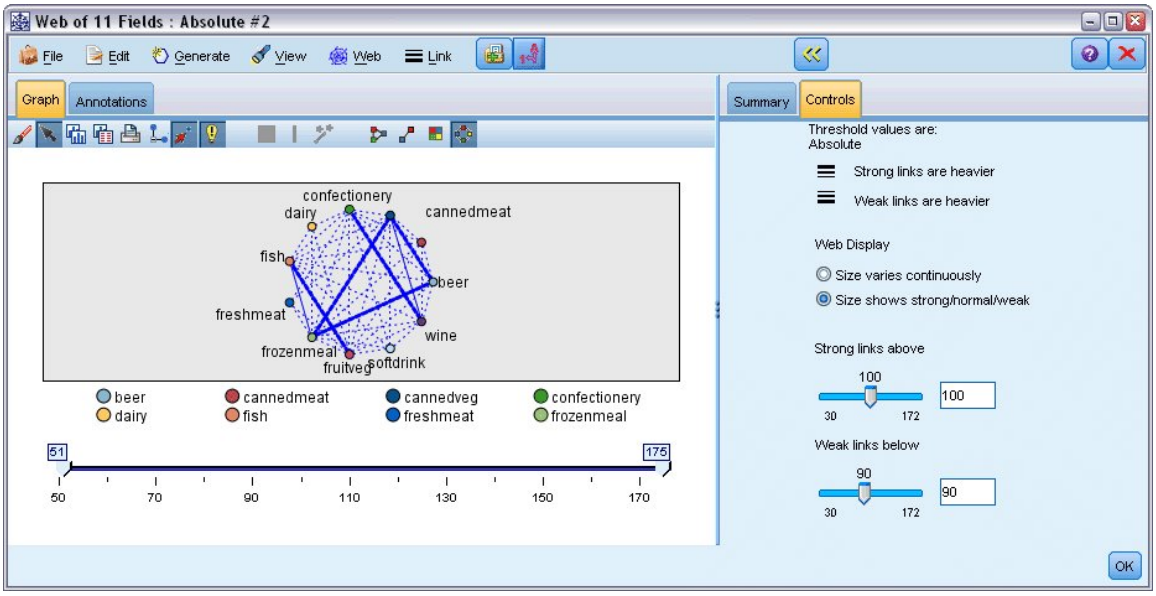


圖 *384.* 受限的 *Web* 顯示

1. 若要指定弱連線和強連線，請按一下工具列上的黃色雙箭頭按鈕。這會展開顯示 Web 輸出摘要和控制項的對話框。
2. 選取**大小顯示強/正常/弱**。

3. 將弱鏈結設為低於 90。
4. 將強鏈結設為高於 100。

在產生的顯示中，三組客戶脫穎而出：

- 購買魚和水果和蔬菜的客戶，可能被稱為「健康飲食者」
- 購買葡萄酒和糕點的客戶
- 購買啤酒、冷凍餐食和蔬菜罐頭（「啤酒、豆類和披薩」）的客戶

側寫客戶群組

您現在已根據客戶購買的產品類型識別出三組客戶，但您還想瞭解這些客戶是誰，即，他們的人口統計資訊。這可以透過針對每一組使用旗標標記每一個客戶，並使用規則歸納 (C5.0) 來建置這些旗標的規則型設定檔來實現。

首先，您必須針對每一組衍生一個旗標。這可以使用您剛剛建立的 Web 顯示自動產生。使用滑鼠右鍵，按一下 *fruitveg* 和 *fish* 之間的鏈結來強調顯示它，然後按一下滑鼠右鍵並選取**為鏈結產生衍生節點**。

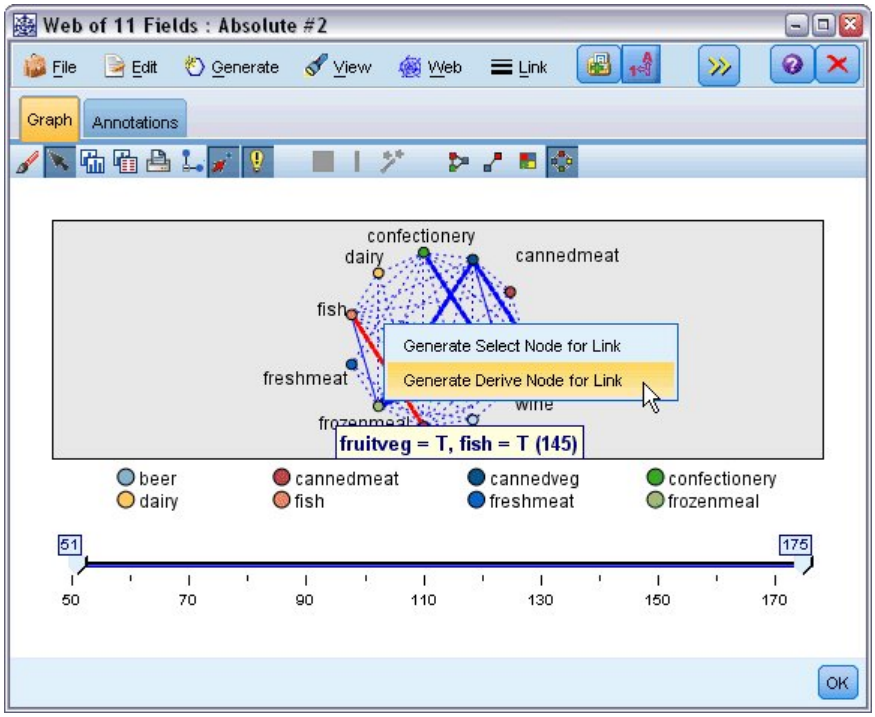


圖 *385.* 針對每一組客戶衍生旗標

編輯產生的「衍生」節點以將「衍生」欄位名稱變更為 *healthy*。使用從 *wine* 到 *confectionery* 的鏈結重複練習，將產生的「衍生」欄位命名為 *wine_chocs*。

針對第三組（包含三個鏈結），先確定未選取任何鏈結。然後透過按住 shift 鍵時按一下滑鼠左鍵，選取 *cannedveg*、*beer* 及 *frozenmeal* 三角形中的全部三個鏈結。（您務必處於「互動」模式而非「編輯」模式。）然後從 Web 顯示功能表選擇：

產生 > 衍生節點（「And」）

將產生的「衍生」欄位的名稱變更為 *beer_beans_pizza*。

若要側寫這些客戶群組，請將現有的「類型」節點連接至系列中的這三個「衍生」節點，然後連接另一個「類型」節點。在新的「類型」節點中，將所有欄位的角色設定為無，但 *value*、*pmethod*、*sex*、*homeown*、*income* 及 *age* 除外，這些欄位應設定為輸入，而相關的客戶群組（例如，*beer_beans_pizza*）應設定為目標。連接 C5.0 節點，將「輸出」類型設定為**規則集**，然後在該節點上按一下執行。產生的模型（針對 *beer_beans_pizza*）包含此組客戶清晰易懂的人口統計資訊。

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

透過選取其他客戶群組旗標作為第二個「類型」節點中的輸出，可以對其套用相同的方法。在此環境定義中透過使用 Apriori 而非 C5.0 可以產生更廣泛的替代設定檔；還可以使用 Apriori 來同步側寫所有客戶群組旗標，因為它不會受限於單個輸出欄位。

摘要

此範例顯示如何使用 IBM SPSS Modeler 透過建模（使用 Apriori）及視覺化（使用 Web 顯示）在資料庫中探索親緣性或鏈結。這些鏈結對應於資料中的觀察值分組，並且這些群組可以詳細調查並透過建模進行側寫（使用 C5.0 規則集）。

在零售領域中，此類客戶分組可以，例如，用來以特價優惠為目標來改進直接郵寄的回應率，或自訂某個分支庫存的產品範圍來符合人口基數的需求。

第 28 章 評量新的車輛產品與服務 (KNN)

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

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

您可以指定要檢查的最近鄰接項數目；此值稱為 *k*。圖片顯示如何使用兩個不同的 *k* 值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。

汽車製造商已為兩個新的車輛（小汽車和卡車）開發了原型。在將新的模型引入此範圍之前，製造商想要判定市場上哪些現有車輛與原形最相似，亦即，哪些車輛是其「最近鄰接項」，進而判定將與哪些模型競爭。

製造商已收集數個種類下現有模型的相關資料，並已新增其原型的詳細資料。要比較之模型所在的種類包括價格（以千為單位）(*price*)、引擎尺寸 (*engine_s*)、 馬力 (*horsepow*)、軸距 (*wheelbas*)、寬度 (*width*)、長度 (*length*)、空車重量 (*curb_wgt*)、燃油容量(*fuel_cap*)和燃油效率 (*mpg*)。

本範例使用位於 *Demos* 資料夾中 *streams* 子資料夾之下的串流 *car_sales_knn.str*。資料檔案為 *car_sales_knn_mod.sav*。請參閱第 4 頁的『Demos 資料夾』主題，以取得更多資訊。

建立串流

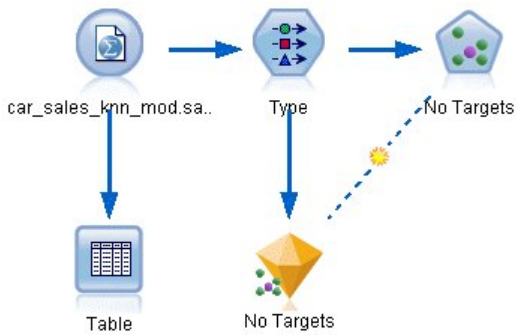


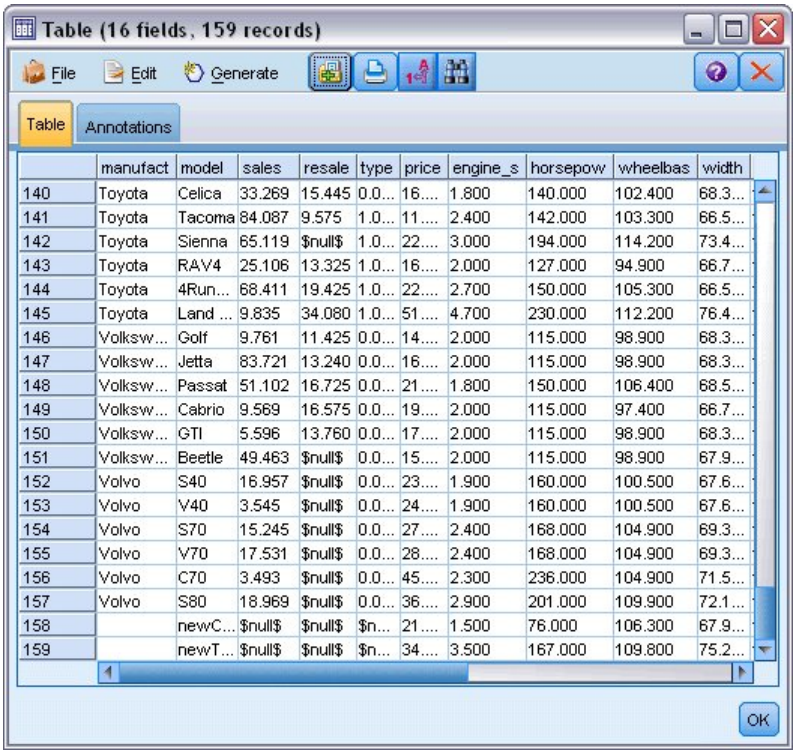
圖 *386. 用於 KNN 建模的串流範例*

建立新串流並新增指向 *car_sales_knn_mod.sav*（位於 IBM SPSS Modeler 安裝架構的 *Demos* 資料夾）的「統計量檔案」來源節點。

首先，我們來看看製造商收集了哪些資料。

1. 將「表格」節點連接至「統計量檔案」來源節點。

2. 開啟「表格」節點並按一下**執行**。



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

圖 *387.* 汽車和卡車的來源資料

已在檔案末尾新增兩個原型（*newCar* 和 *newTruck*）的詳細資料。

我們可以從來源資料中看到製造商使用「卡車」分類（*type* 直欄中的值 1）而不是簡單的車輛的任何非機動車類型。

當我們要識別兩個原形的最近鄰接項時，為了可以將兩個原型指定為保留項，必須要有最後一欄 *partition*。這樣一來，共資料將不會影響計算，因為它是我們要考量的市場剩餘部分。當在此欄位中所有其他記錄都具有 0 時，將兩筆保留記錄的 *partition* 值設為 1 可讓我們稍後在要設定焦點記錄時使用此欄位，焦點記錄即我們要計算其最近鄰接項的記錄。

暫時保留表格輸出視窗開啟，因為我們稍後將參照它。

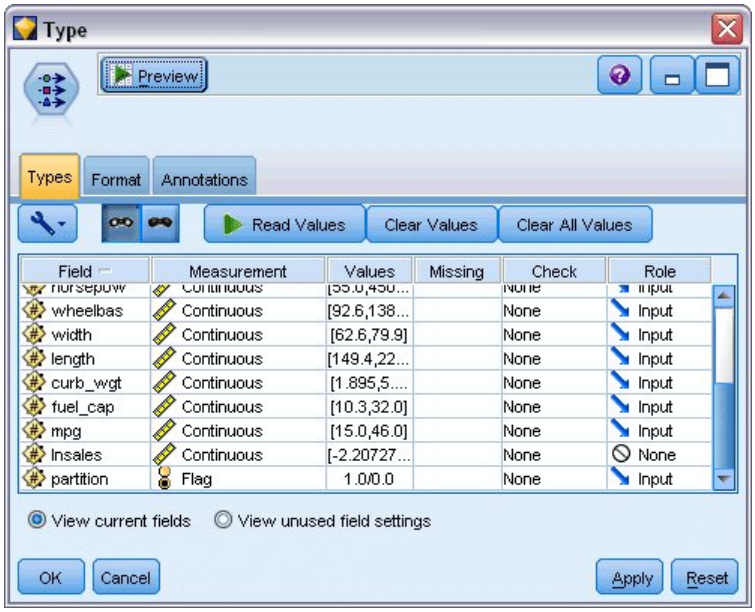


圖 *388.* 類型節點設定

3. 將「類型」節點新增到串流中。
4. 將「類型」節點連接至「統計量檔案」來源節點。
5. 開啟「類型」節點。

   我們只想要對欄位 *price* 到 *mpg* 進行比較，因此保留將所有這些欄位的角色都設為**輸入**。
6. 將所有其他欄位（從 *manufact* 到 *type* 再加上 *lnsales*）的角色設為**無**。
7. 將最後一個欄位 *partition* 的測量層次設為**旗標**。確保其角色設為**輸入**。
8. 按一下**讀取值**以將資料值讀入串流中。
9. 按一下**「確定」**。

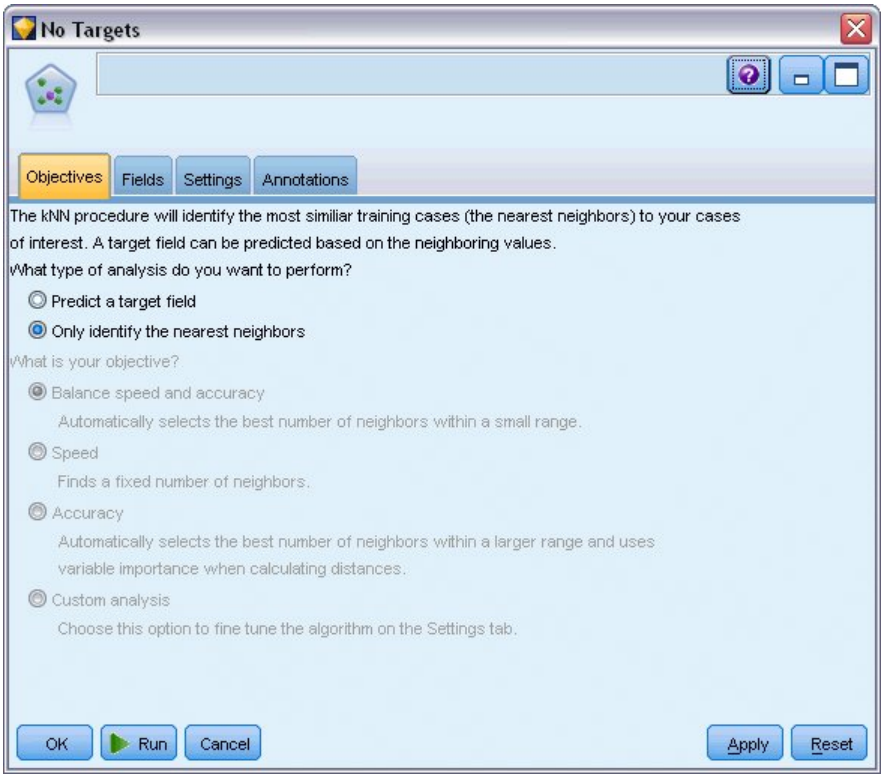


圖 *389.* 選擇以識別最近的鄰接項

10. 將 KNN 節點連接至「類型」節點。
11. 開啟 KNN 節點。

    我們這次不會預測目標欄位，因為我們只想要找出兩個原型的最近鄰接項。
12. 在**目標**標籤上，選擇**只識別最近鄰接項**。
13. 按一下**設定**標籤。

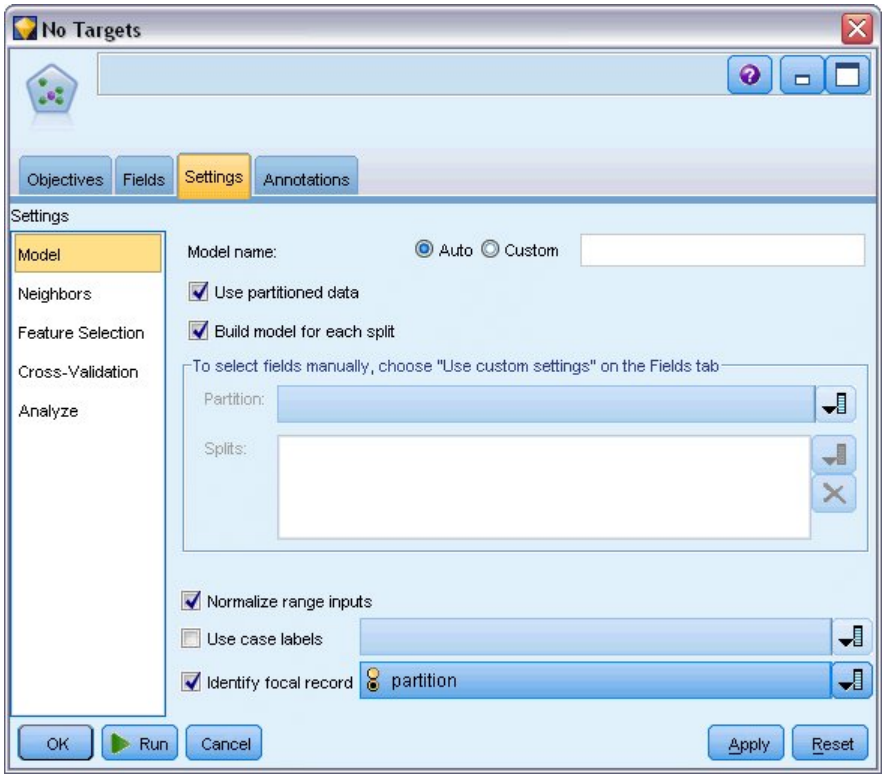


圖 *390.* 使用分割區欄位來識別焦點記錄

現在，我們可使用 *partition* 欄位來識別焦點記錄，亦即我們要識別其最近鄰接項的記錄。透過使用旗標欄位，可確保其中的這個欄位值設為 1 的記錄會成為焦點記錄。

如我們所見，在此欄位中值為 1 的記錄只有 *newCar* 和 *newTruck*，因此這些記錄將是焦點記錄。

14. 在**設定**標籤的**模型**畫面上，選取**識別焦點記錄**勾選框。
15. 從欄位的下拉清單中，選擇 **partition**。
16. 按一下**執行**按鈕。

檢查輸出

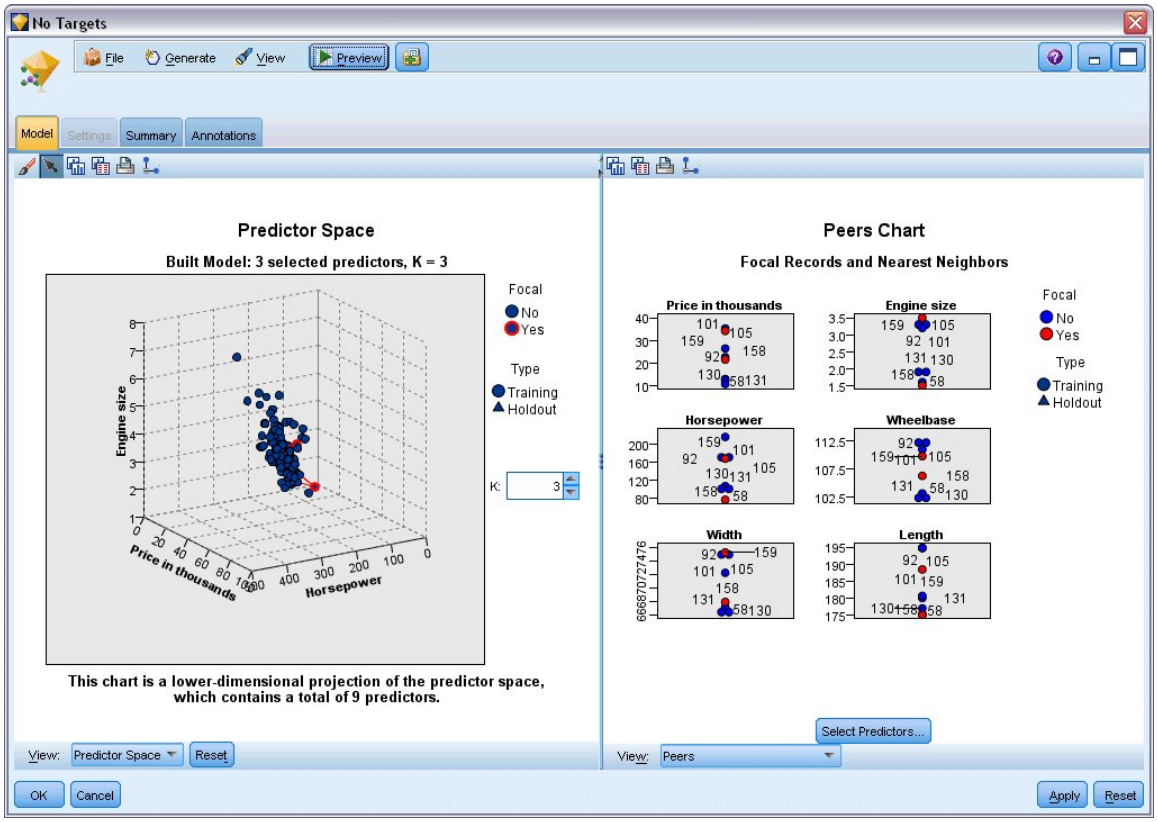


圖 *391.* 模型檢視器視窗

已在串流畫布和「畫布」選用區中建立模型區塊。開啟任一區塊以查看「模型檢視器」顯示畫面，其具有雙畫面視窗：

- 第一個畫面會顯示模型的概述，稱為主要視圖。「最近鄰接項」模型的主要視圖稱為**預測工具空間**。
- 第二個面板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。

  鏈結視圖顯示當您往下探查主要視圖的一部分時模型的某個特性的相關詳細資料。

預測工具空間

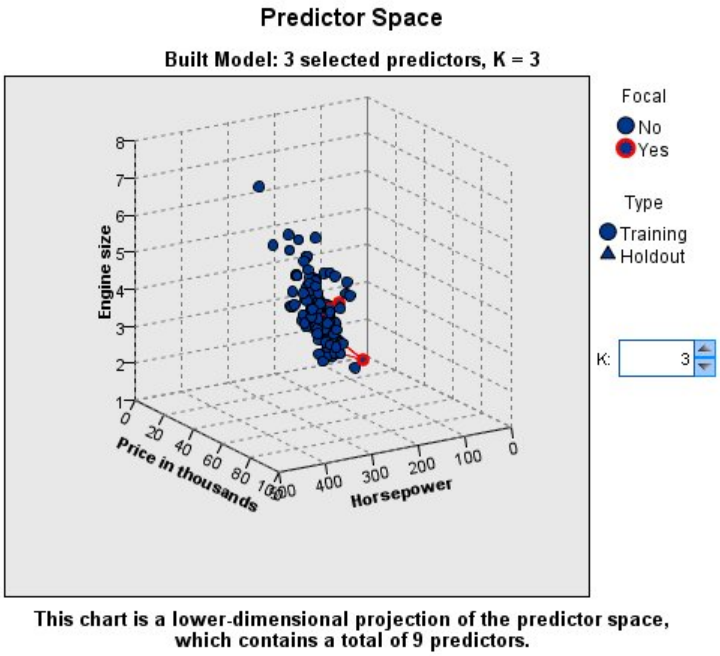


圖 *392.* 預測工具空間圖表

預測工具空間圖表是一個互動式 3-D 圖，它繪制三個特性（實際上是來源資料的前三個輸入欄位）的資料點、代表價格、引擎大小及馬力的圖形。

兩個焦點記錄以紅色強調顯示，利用線條將它們連接至 *k* 個最近的鄰接項。

透過按一下並拖曳圖表，您可以旋轉以便更好地檢視各個點在預測工具空間中的分佈。按一下**重設**按鈕將其回復成預設視圖。

對等圖表

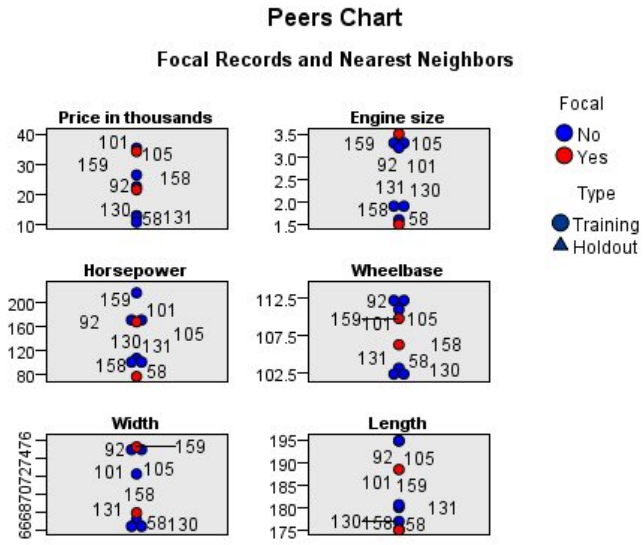


圖 *393.* 對等圖表

預設輔助視圖是對等圖表，其強調顯示在預測工具空間中選取的兩筆焦點記錄，以及六個特性（來源資料的前六個輸入欄位）中每個特性上的 *k* 個最近鄰接項。

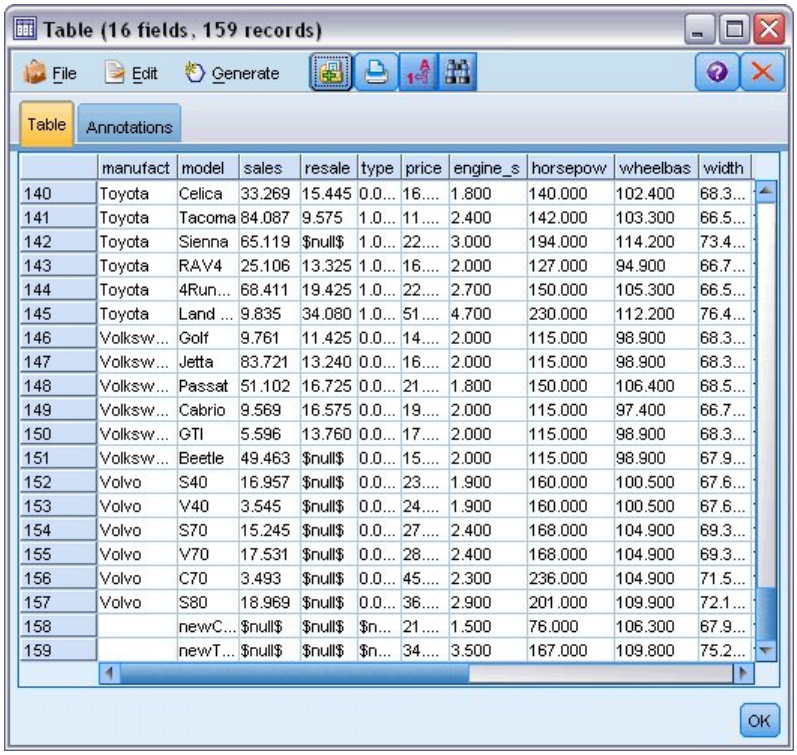
車輛由其在來源資料中的記錄編號表示。在這裡，我們需要「表格」節點中的輸出來協助識別它們。

如果「表格」節點輸出仍然可用：

1. 按一下位於主要 IBM SPSS Modeler 右上方的管理程式窗格的**輸出**標籤。
2. 按兩下項目**表格（16 個欄位，159 筆記錄）**。

如果表格輸出不再可用：

3. 在主要 IBM SPSS Modeler 視窗上，開啟「表格」節點。
4. 按一下**「執行」**。



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

圖 *394.* 依記錄編號識別記錄

向下捲動到表格底端，我們可以看到 *newCar* 和 *newTruck* 是資料中的最後兩筆記錄，編號分別為 158 和 159。

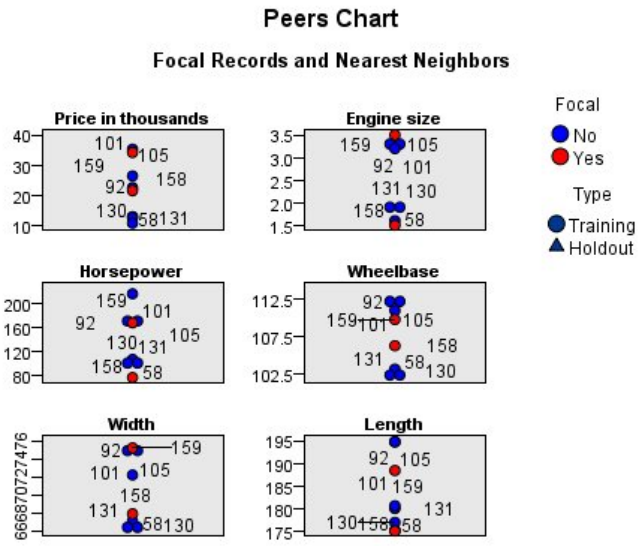


圖 *395.* 比較對等圖表上的特性

我們可以在對等圖表上看到，例如，*newTruck* (159) 的引擎大小大於它的任何最近鄰接項，而 *newCar* (158) 的引擎小於它的 任何最近鄰接項。

對於六個特性中的每一個，您都可以將滑鼠移至個別點上來查看該特定案例的每個特性的實際值。

但哪些車輛是 *newCar* 和 *newTruck* 的最近鄰接項呢？

對等圖表有點擁擠，讓我們變更為更簡單的視圖吧。

5. 按一下對等圖表（目前顯示**對等**的項目）底端的**視圖**下拉清單。
6. 選取**鄰接項與距離表格**。

鄰接項與距離表格

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

圖 *396.* 鄰接項與距離表格

這樣更好。現在，我們將在市場中看到兩個原型都最接近的三個模型。

如果是 *newCar*（焦點記錄為 158），則它們是 Saturn SC (131)、 Saturn SL (130) 和 Honda Civic (58)。

沒有大的驚喜--所有這三個都是中型的轎車，因此 *newCar* 應該非常適合，尤其是其卓越的燃油效率。

如果是 *newTruck*（焦點記錄為 159），則最近的鄰接項為 Nissan Quest (105)、 Mercury Villager (92) 和 Mercedes M-Class (101)。

如先前所見，這些都不是傳統意義上的卡車，而是被分類為非汽車的車輛。查看「表格」節點輸入以尋找其最近的鄰接項，我們可以看到 *newTruck* 相對較貴，並且是其類型中最重的卡車之一。但是，燃油效率優於其最近的競爭對手，所以這應該對它有利。

摘要

我們已查看如何在特定資料集的觀察值中，使用最近鄰接項分析來比較一組廣泛的特性。我們還針對兩個截然不同的保留記錄計算了與那些保留項最為相似的觀察值。

第 29 章 揭示商業度量 (TCM) 中的因果關係

一個企業會追蹤可說明一段時間內企業財務狀態的各種關鍵績效指標，他們還會追蹤可以控制的多種度量值。他們對使用時間原因建模來揭示可控制度量值與關鍵績效指標之間的因果關係很感興趣。他們還想瞭解關鍵績效指標之間的因果關係。

資料檔 `tcm_kpi.sav` 包含關於關鍵績效指標及可控制度量值的每週資料。關鍵績效指標的資料會儲存在字首為 *KPI* 的欄位中。可控制度量值的資料會儲存在字首為 *Lever* 的欄位中。

建立串流

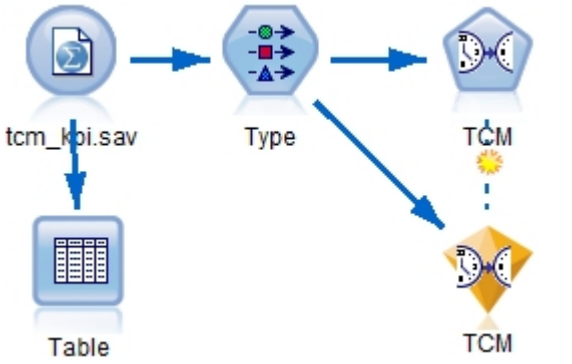
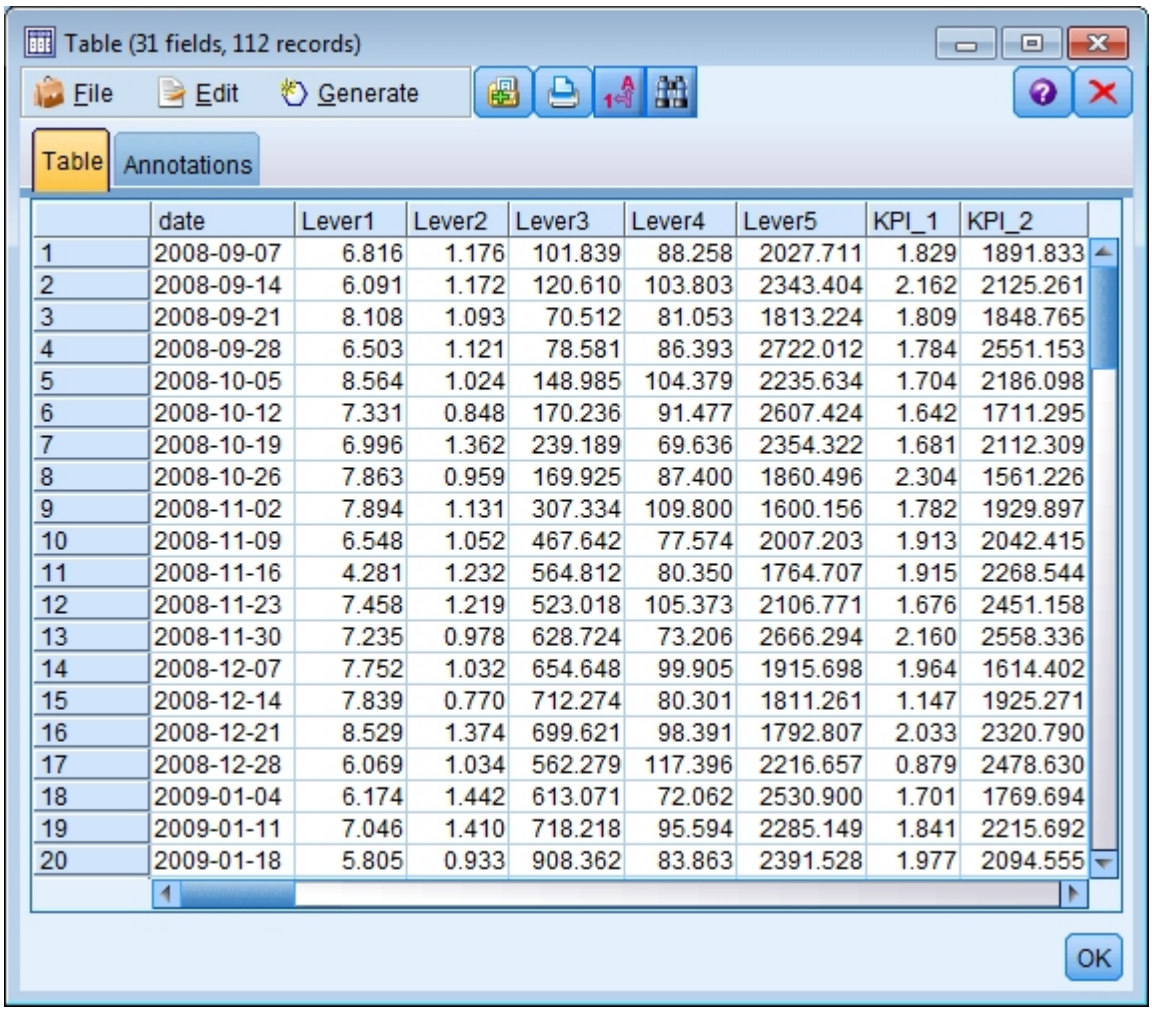


圖 *397.* 用於 *TCM* 建模的樣本串流

1. 建立新串流並新增指向 *tcm_kpi.sav*（位於 folder of your IBM SPSS Modeler 安裝的 *Demos* 資料夾中）的「統計量檔案」來源節點。 installation.
2. 將「表格」節點連接至「統計量檔案」來源節點。
3. 開啟「表格」節點，然後按一下**執行**來查看資料。它包含關於關鍵績效指標及可控制度量值的每週資料。關鍵績效指標的資料儲存在字首為 *KPI* 的欄位中，可控制度量值的資料儲存在字首為 *Lever* 的欄位中。



	date	Lever1	Lever2	Lever3	Lever4	Lever5	KPI_1	KPI_2
1	2008-09-07	6.816	1.176	101.839	88.258	2027.711	1.829	1891.833
2	2008-09-14	6.091	1.172	120.610	103.803	2343.404	2.162	2125.261
3	2008-09-21	8.108	1.093	70.512	81.053	1813.224	1.809	1848.765
4	2008-09-28	6.503	1.121	78.581	86.393	2722.012	1.784	2551.153
5	2008-10-05	8.564	1.024	148.985	104.379	2235.634	1.704	2186.098
6	2008-10-12	7.331	0.848	170.236	91.477	2607.424	1.642	1711.295
7	2008-10-19	6.996	1.362	239.189	69.636	2354.322	1.681	2112.309
8	2008-10-26	7.863	0.959	169.925	87.400	1860.496	2.304	1561.226
9	2008-11-02	7.894	1.131	307.334	109.800	1600.156	1.782	1929.897
10	2008-11-09	6.548	1.052	467.642	77.574	2007.203	1.913	2042.415
11	2008-11-16	4.281	1.232	564.812	80.350	1764.707	1.915	2268.544
12	2008-11-23	7.458	1.219	523.018	105.373	2106.771	1.676	2451.158
13	2008-11-30	7.235	0.978	628.724	73.206	2666.294	2.160	2558.336
14	2008-12-07	7.752	1.032	654.648	99.905	1915.698	1.964	1614.402
15	2008-12-14	7.839	0.770	712.274	80.301	1811.261	1.147	1925.271
16	2008-12-21	8.529	1.374	699.621	98.391	1792.807	2.033	2320.790
17	2008-12-28	6.069	1.034	562.279	117.396	2216.657	0.879	2478.630
18	2009-01-04	6.174	1.442	613.071	72.062	2530.900	1.701	1769.694
19	2009-01-11	7.046	1.410	718.218	95.594	2285.149	1.841	2215.692
20	2009-01-18	5.805	0.933	908.362	83.863	2391.528	1.977	2094.555

圖 *398.* 關鍵績效指標及可控制度量值的來源資料

4. 將「類型」節點新增到串流中。
5. 將「類型」節點連接至「統計量檔案」來源節點。

執行分析

1. 將 TCM 節點連接至「類型」節點，然後開啟 TCM 節點並跳至**欄位**標籤的**觀察**區段。

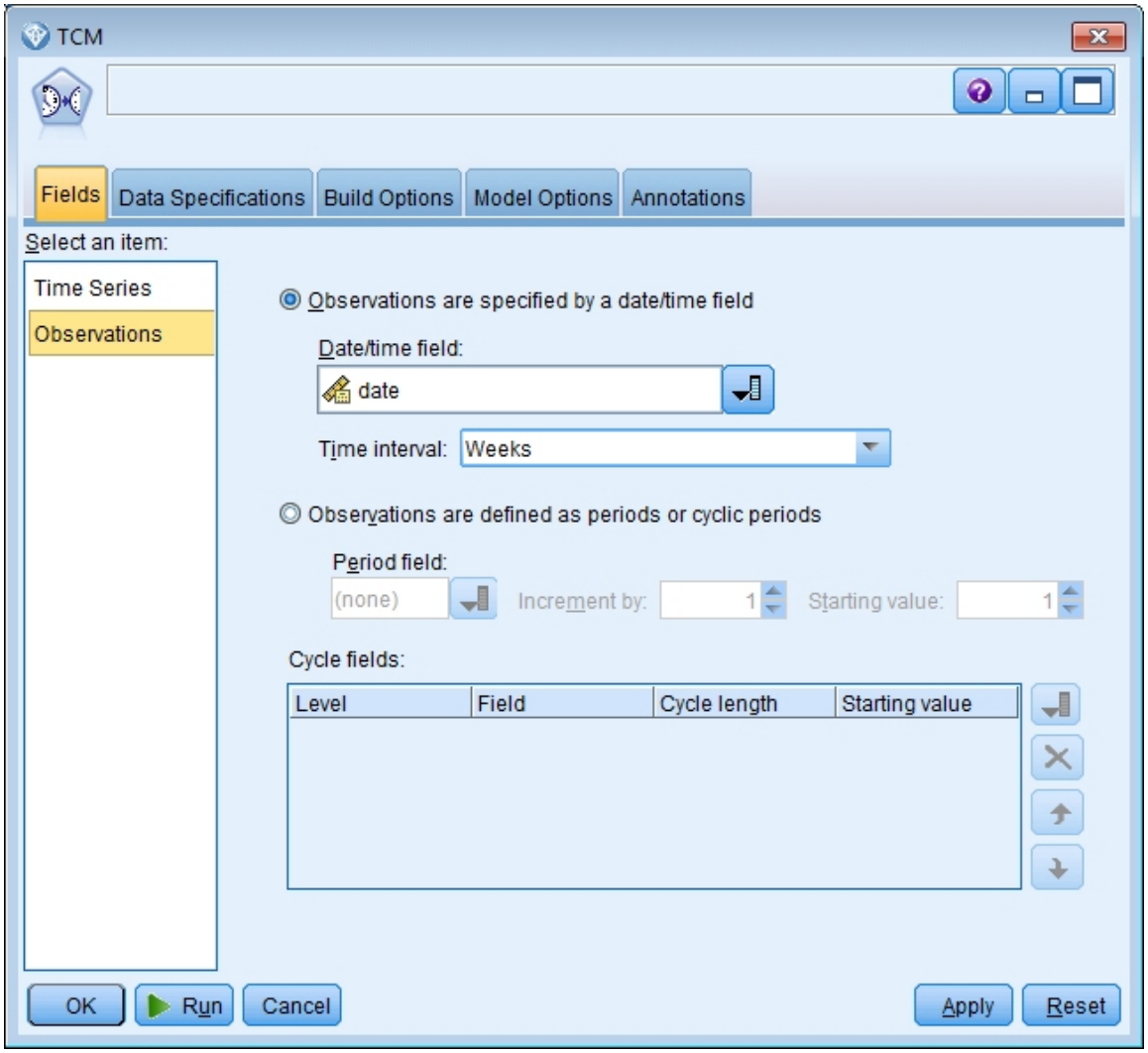


圖 *399.* 時間原因建模，觀察

2. 從「日期/時間」欄位中選取 *date*，然後從「時間間隔」欄位中選取 *Weeks*。
3. 按一下**時間數列**，然後選取**使用預先定義的角色**。

   在範例資料集 `tcm_kpi.sav` 中，欄位 *Lever1* 到 *Lever5* 的角色為「輸入」，*KPI_1* 到 *KPI_25* 的角色為「兩者」。選取**使用預先定義的角色**後時，會將角色為「輸入」的欄位視為時間原因建模的候選輸入，將角色為「兩者」的欄位視為時間原因建模的候選輸入及目標。

   時間原因建模程序會從候選輸入集中判定每一個目標的最佳輸入。在此範例中，候選輸入是欄位 *Lever1* 到 *Lever5* 以及欄位 *KPI_1* 到 *KPI_25*。
4. 按一下「**執行**」。

整體模型品質圖

「整體模型品質」輸出項目依預設產生，會顯示所有模型的長條圖及相關聯的模型配適點形圖。每一個目標數列都有一個個別的模型。模型適合度透過所選的適合度統計量進行測量。此範例使用預設的適合度統計量，即 R 平方。

「整體模型品質」項目包含互動式特性。若要啟用這些特性，請透過按兩下「檢視器」中的「整體模型品質」圖來啟動該項目。

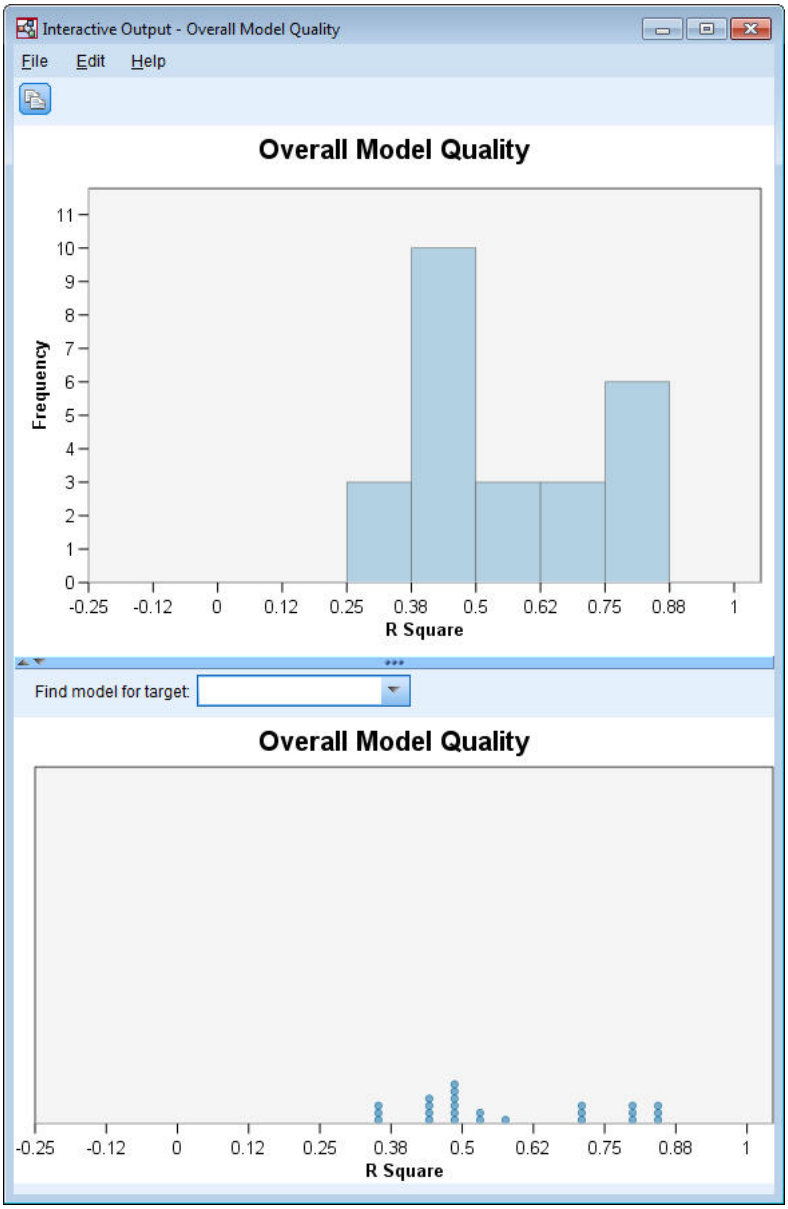


圖 *400.* 整體模型品質

按一下長條圖中的條欄，以過濾點圖，使其僅顯示與所選條欄相關聯的模型。將滑鼠移至點形圖中點的上方會顯示工具提示，其中包含相關聯數列的名稱以及適合度統計量的值。您可以透過在**為目標尋找模型**方框中指定數列名稱，為點形圖中的特定目標數列尋找模型。

整體模型系統

「整體模型系統」輸出項目依預設產生，會顯示模型系統中數列之間因果關係的圖形表示法。依預設，會顯示由 R 平方適合度統計量的值判定的前 10 個模型的關係。頂級模型（也稱為最適模型）數及適合度統計量在「時間原因建模」對話框的「要顯示的數列」設定（在「建置選項」標籤上）上指定。

「整體模型系統」項目包含互動式特性。若要啟用這些特性，請透過按兩下「檢視器」中的「整體模型系統」圖來啟動該項目。在此範例中，最重要的是查看系統中所有數列之間的關係。在互動式輸出中，從**強調顯示數列關係**下拉清單中選取**所有數列**。

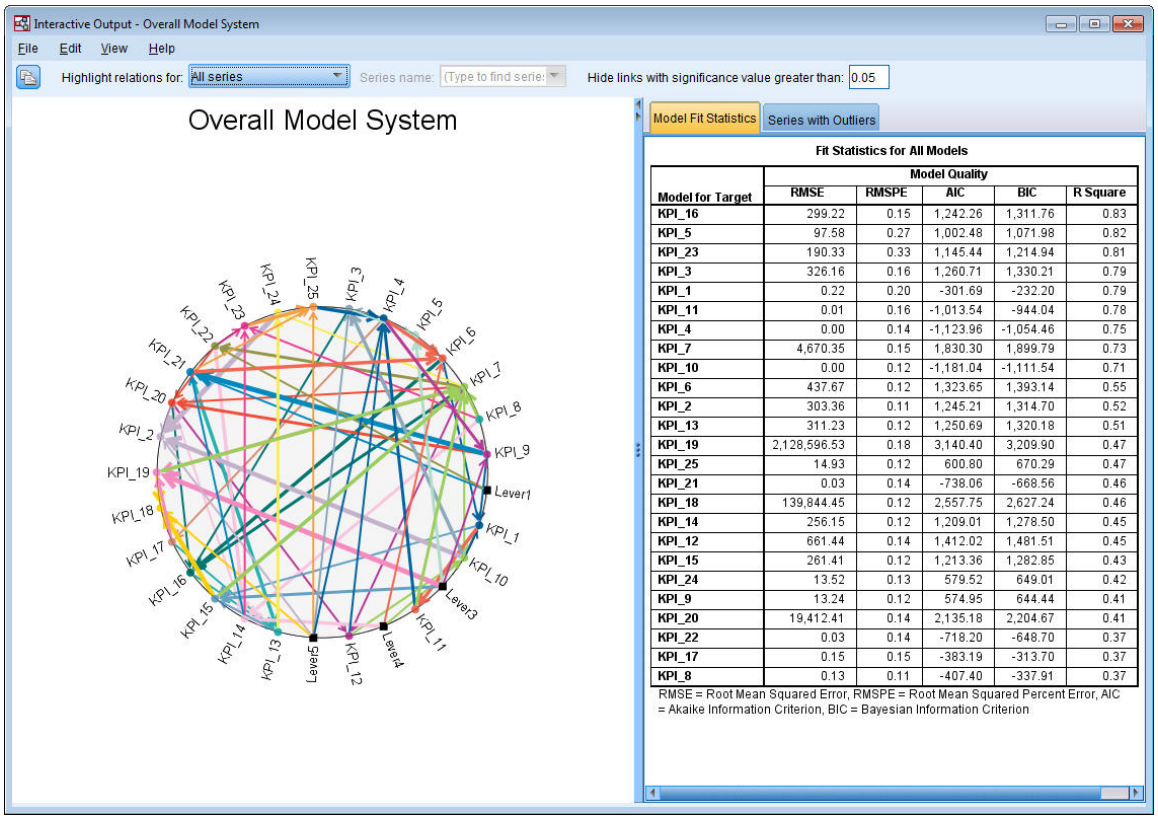


圖 *401.* 整體模型系統，所有數列的視圖

將特定目標連接至其輸入的所有線條的顏色都是相同的，每條線上的箭頭都從輸入指向該輸入的目標。例如，*Lever3* 輸入至 *KPI_19*。

每一個線條的粗細指示因果關係的顯著性，其中線條越粗表示關係越顯著。依預設會隱藏顯著性值大於 0.05 的因果關係。在 0.05 層次，只有 *Lever1*、*Lever3*、*Lever4* 及 *Lever5* 與關鍵績效指標欄位具有顯著的因果關係。您可以透過在標有**隱藏顯著性值大於下列值的鏈結**的欄位中輸入值來變更臨界值顯著性層次。

除了揭示 *Lever* 欄位與關鍵績效指標欄位之間的因果關係以外，分析還揭示了關鍵績效指標欄位之間的關係。例如，選取 *KPI_10* 來輸入至 *KPI_2* 的模型。

您可以過濾視圖以僅顯示單一數列的關係。例如，若要僅檢視 *KPI_19* 的關係，請按一下 *KPI_19* 的標籤，按一下滑鼠右鍵，然後選取**強調顯示數列關係**。

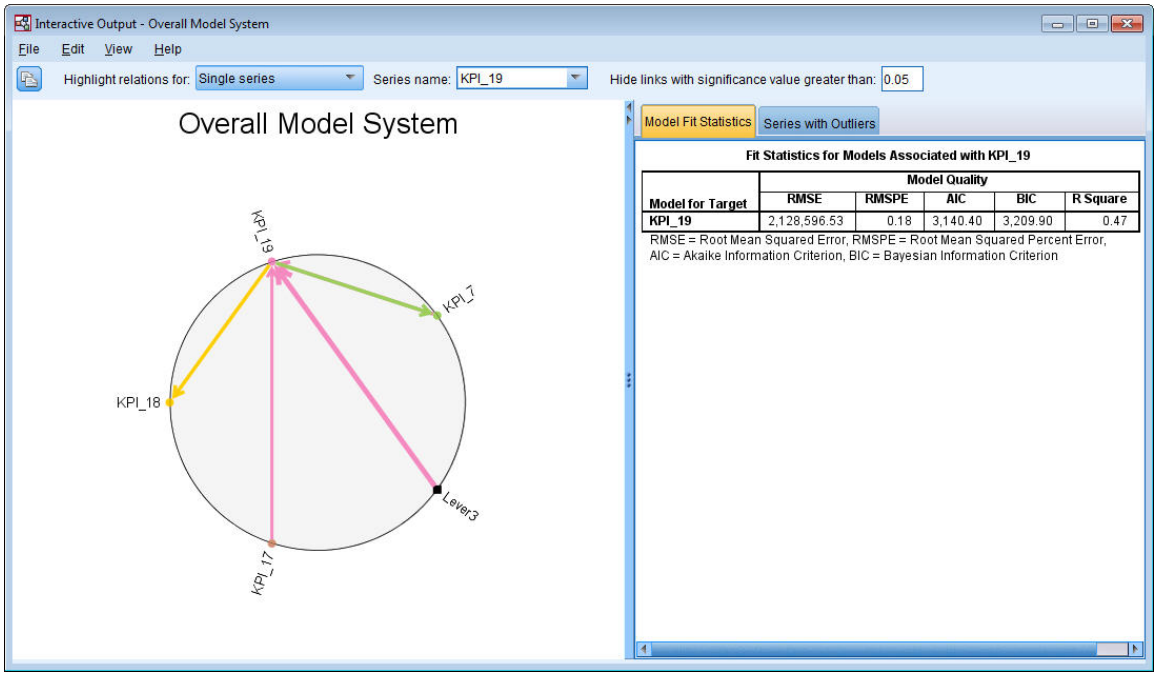


圖 *402.* 整體模型系統，單一數列的視圖

此視圖顯示 *KPI_19* 的顯著性值小於或等於 0.05 的輸入。它還顯示在 0.05 顯著性層次，選取 *KPI_19* 來輸入至 *KPI_18* 及 *KPI_7*。

除了顯示所選數列的關係以外，輸出項目還包含針對數列偵測到的任何偏離值的相關資訊。按一下**含偏離值的數列**標籤。

Series with Outliers for KPI_19

Series	Time	Observed Value
KPI_19	2008-10-12	7,358,201.68
	2009-04-05	2.10E+007
	2010-09-19	6,492,157.97

圖 *403. KPI_19* 的偏離值

偵測到 *KPI_19* 的三個偏離值。如果模型系統包含所有探索到的連線，則可能會超出偏離值偵測範圍，並判定出最有可能導致特定偏離值的數列。此分析類型稱為偏離值主要原因分析，並會在此個案研討後面的主題中進行說明。

影響圖表

您可以透過產生影響圖表來取得與特定數列相關聯的所有關係的完整視圖。按一下「整體模型系統」圖中 *KPI_19* 的標籤，按一下滑鼠右鍵，並選取**建立影響圖表**。

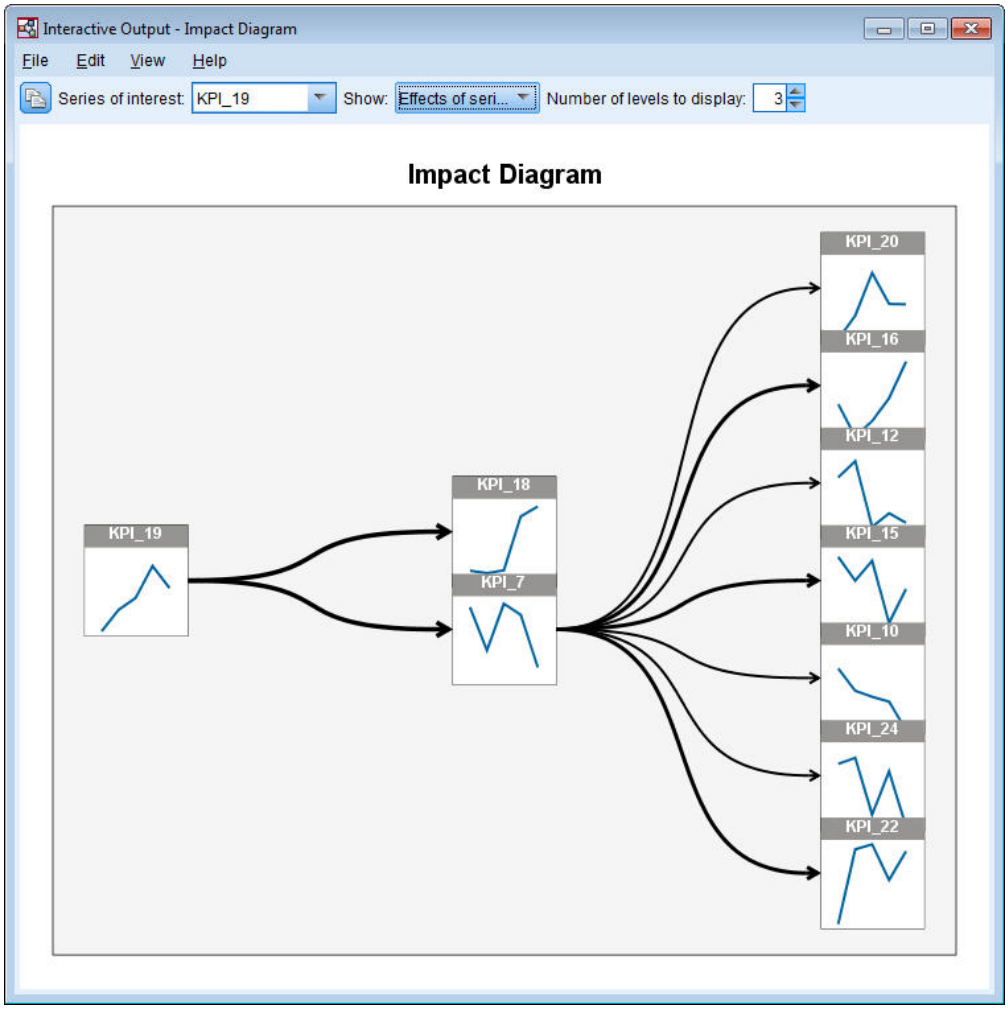


圖 *404.* 效應的影響圖表

從「整體模型系統」建立影響圖表時，如此範例中所示，它一開始會顯示受選定數列影響的數列。依預設，影響圖表會顯示三個效應層次，其中第一個層次僅是感興趣數列。其他每一個層次都會顯示感興趣數列之更間接的效應。您可以變更**要顯示的層次數**的值來顯示更多或更少的效果層次。此範例的影響圖表顯示 *KPI_19* 直接輸入至 *KPI_18* 及 *KPI_7*，但它會透過其對數列 *KPI_7* 的影響間接影響多個數列。如在整體模型系統中一樣，線條的粗細指示因果關係的顯著性。

在影響圖表的每一個節點中顯示的圖表會顯示估計週期結束時關聯數列的最後 L+1 個值以及任何預測值，其中 L 是每一個模型中包括的落階項目數。您可以透過按一下關聯節點來取得這些值的詳細序列圖。

按兩下節點會將關聯的數列設定為感興趣數列，並會基於該數列重新產生影響圖表。您還可以在**感興趣數列**方框中指定數列名稱來選取不同的感興趣數列。

影響圖表還可以顯示影響感興趣數列的數列。這些數列稱為原因。若要查看影響 *KPI_19* 的數列，請從**顯示**下拉清單中選取**數列原因**。

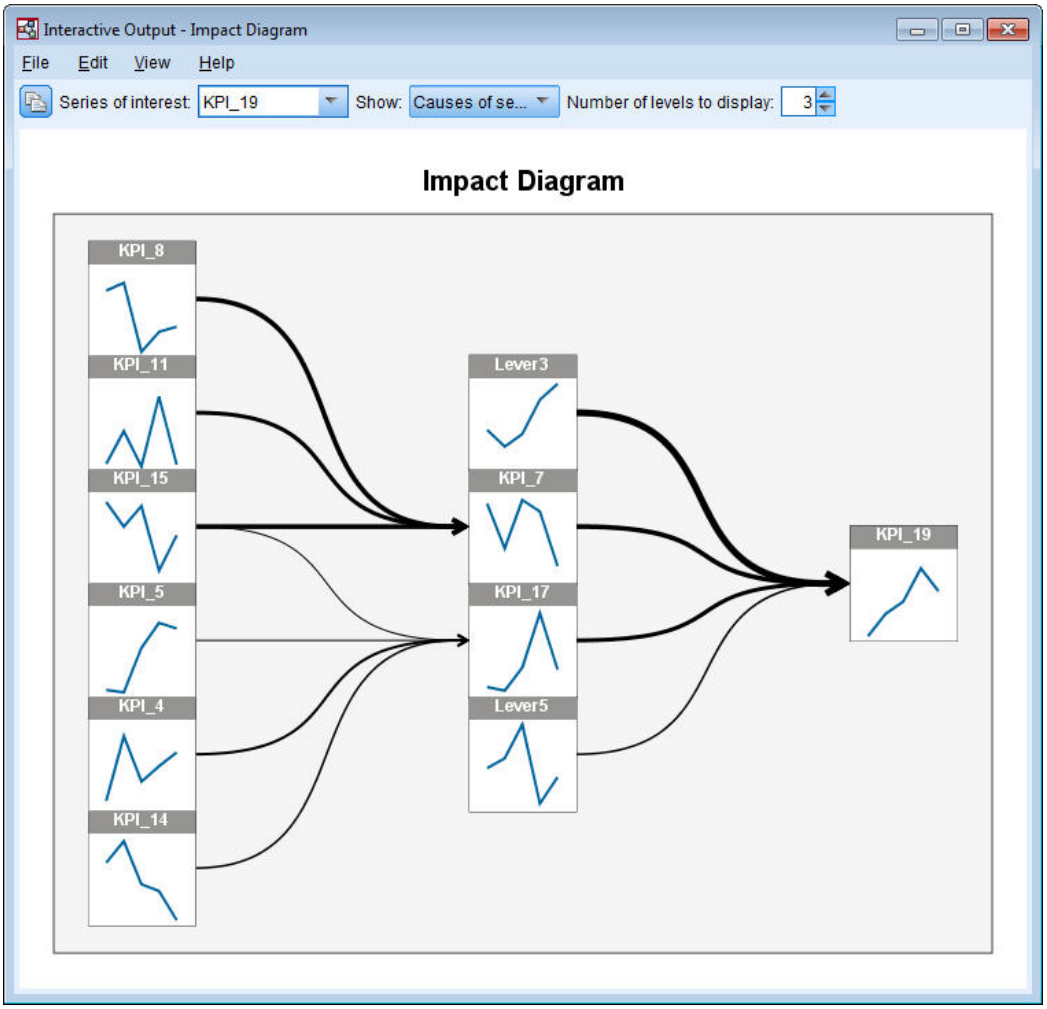


圖 *405.* 原因的影響圖表

此視圖顯示 *KPI_19* 的模型有四個輸入，並且 *Lever3* 與 *KPI_19* 的因果聯繫最顯著。它還會顯示透過對 *KPI_7* 及 *KPI_17* 的影響間接影響 *KPI_19* 的數列。針對效應討論的層次概念同樣也適用於原因。同樣，您可以變更**要顯示的層次數**的值來顯示更多或更少的原因層次。

判定偏離值的主要原因

若為時間原因模型系統，則可能會超出偏離值偵測範圍，並判定出最有可能導致特定偏離值的數列。此處理程序稱為偏離值主要原因分析，必須按一個數列接一個數列的方式來要求。分析需要時間原因模型系統及用來建置該系統的資料。在此範例中，作用中的資料集是用來建置模型系統的資料。

若要執行偏離值主要原因分析，請執行下列動作：

1. 在 TCM 對話框中，跳至**建置選項**標籤，然後按一下**選取項目**清單中的**要顯示的數列**。

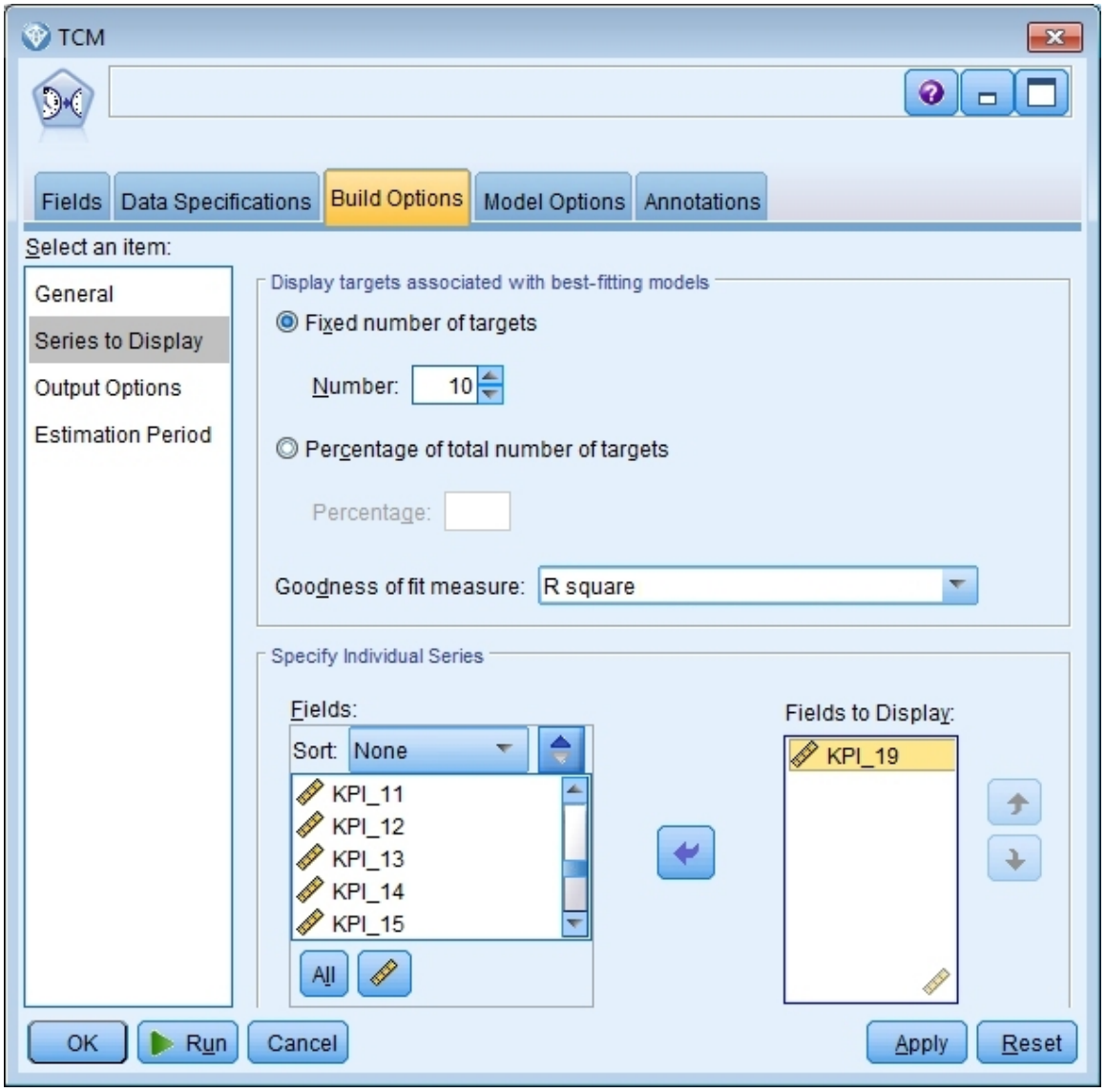


圖 *406.* 要顯示的時間原因模型數列

2. 將 *KPI_19* 移至**要顯示的欄位**清單。
3. 按一下「選項」標籤上**選取項目**清單中的**輸出選項**。

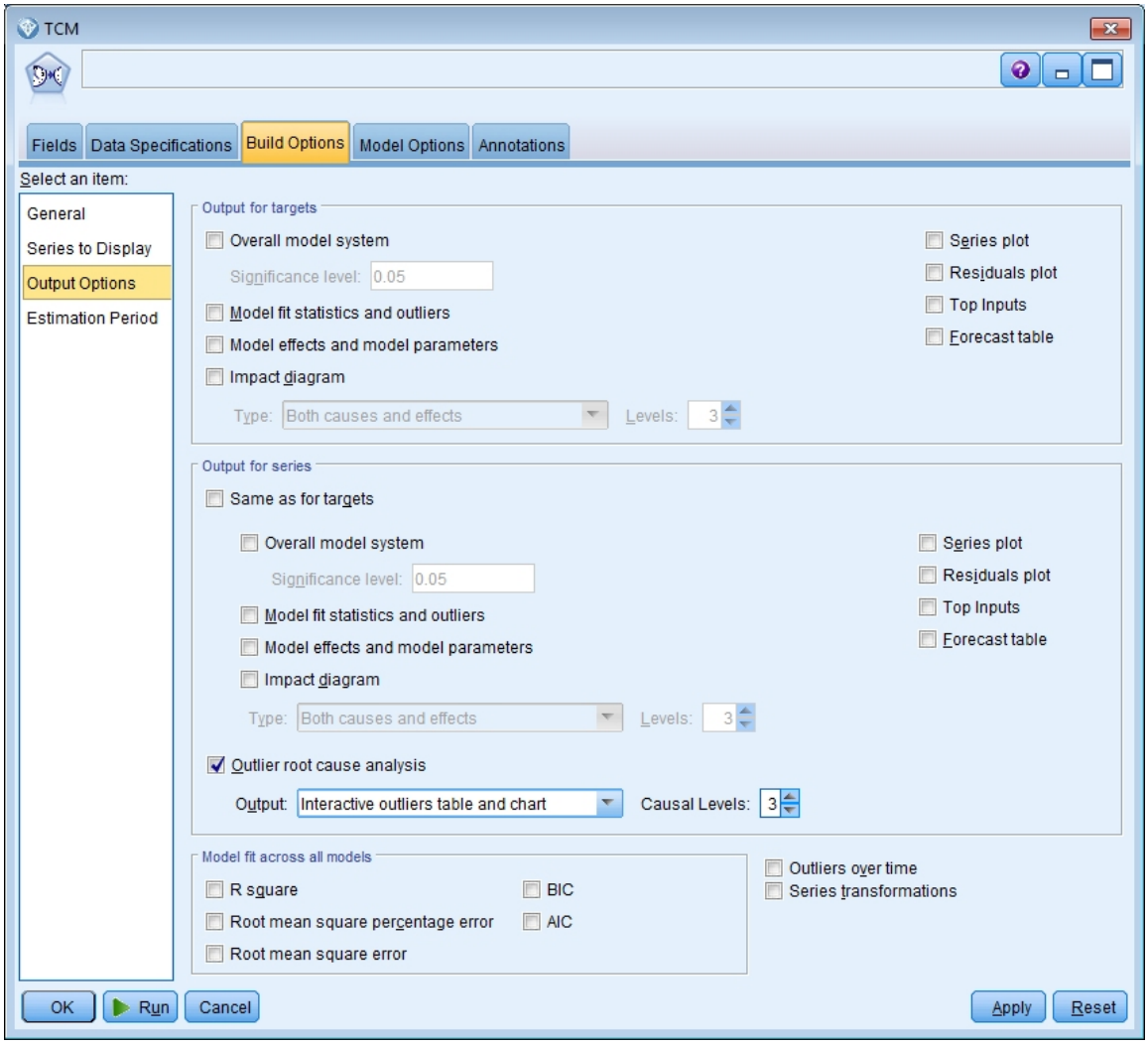
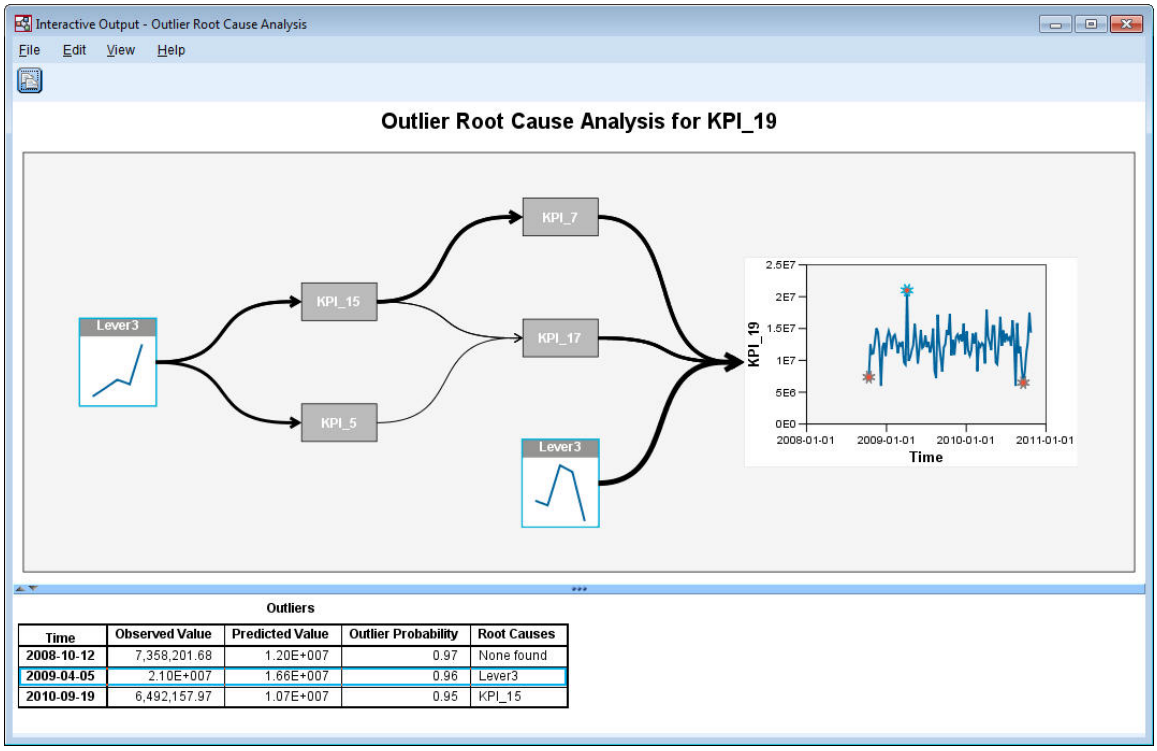


圖 *407.* 時間原因模型輸出選項

4. 取消選取**整體模型系統**、**與目標相同**、**R 平方**及**數列轉換**。
5. 選取**偏離值主要原因分析**，並保持**輸出**及**原因層次**的現有設定。
6. 按一下「**執行**」。
7. 按兩下「檢視器」中 *KPI_19* 的「偏離值主要原因分析」圖以啟動它。



Time	Observed Value	Predicted Value	Outlier Probability	Root Causes
2008-10-12	7,358,201.68	1.20E+007	0.97	None found
2009-04-05	2.10E+007	1.66E+007	0.96	Lever3
2010-09-19	6,492,157.97	1.07E+007	0.95	KPI_15

圖 *408. KPI_19* 的偏離值主要原因分析

分析結果在「偏離值」表格中彙總。該表格顯示找到了 2009-04-05 及 2010-09-19 偏離值的主要原因，但未找到 2008-10-12 偏離值的主要原因。按一下「偏離值」表格中的列會強調顯示主要原因數列的路徑，這裡顯示的是 2009-04-05 偏離值的主要原因數列。此動作還會強調顯示在序列圖中選取的偏離值。您還可以直接按一下序列圖中偏離值的圖示來強調顯示該偏離值之主要原因數列的路徑。

針對 2009-04-05 的偏離值，主要原因是 *Lever3*。該圖表顯示 *Lever3* 直接輸入至 *KPI_19*，但它還會透過其對影響 *KPI_19* 的其他數列的影響間接影響 *KPI_19*。偏離值主要原因分析的其中一個可配置參數是用來搜尋主要原因的原因層次數。依預設，會搜尋三個層次。主要原因數列的出現會顯示最多指定的原因層次數。在此範例中，*Lever3* 在第一個原因層次及第三個原因層次都會出現。

強調顯示的偏離值路徑中的每一個節點皆包含一個圖表，其時間範圍視節點出現的層次而定。針對第一個原因層次中的節點，該範圍是 T-1 到 T-L，其中 T 是偏離值出現的時間，L 是每一個模型中包括的落階項目數。針對第二個原因層次中的節點，該範圍是 T-2 到 T-L-1；針對第三個層次，該範圍是 T-3 到 T-L-2。您可以透過按一下關聯節點來取得這些值的詳細序列圖。

執行實務

若為時間原因模型系統，則可以執行使用者定義的實務。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。分析需要時間原因模型系統及用來建置該系統的資料。在此範例中，作用中的資料集是用來建置模型系統的資料。

若要執行實務，請執行下列動作：

1. 在 TCM 輸出對話框中，按一下**實務分析**按鈕。
2. 在「時間原因模型實務」對話框中，按一下**定義實務週期**。

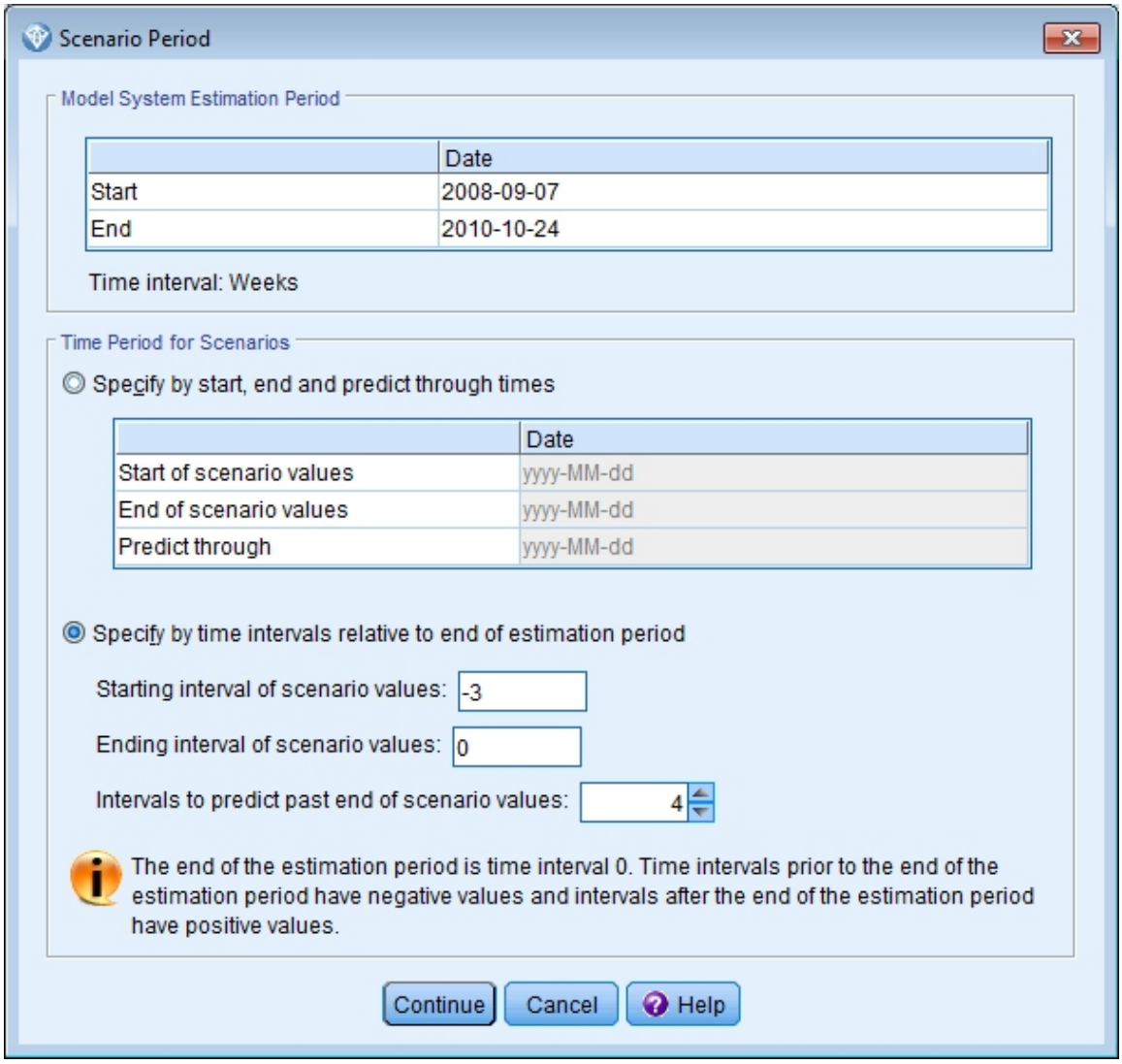


圖 *409.* 實務週期

3. 選取**依相對於估計週期結束的時間間隔來指定**。
4. 針對起始間隔輸入 `-3`，針對結束間隔輸入 `0`。

   這些設定指定每一個實務都基於針對估計週期中最後四個時間間隔指定的值。針對此範例，最後四個時間間隔表示最後四週。指定實務值的那段時間範圍稱為實務週期。
5. 針對間隔輸入 `4` 以超過實務值的結束進行預測。

   此設定指定針對超過實務週期結束的四個時間間隔產生預測。
6. 按一下「**繼續**」。
7. 在「實務」標籤上按一下**新增實務**。

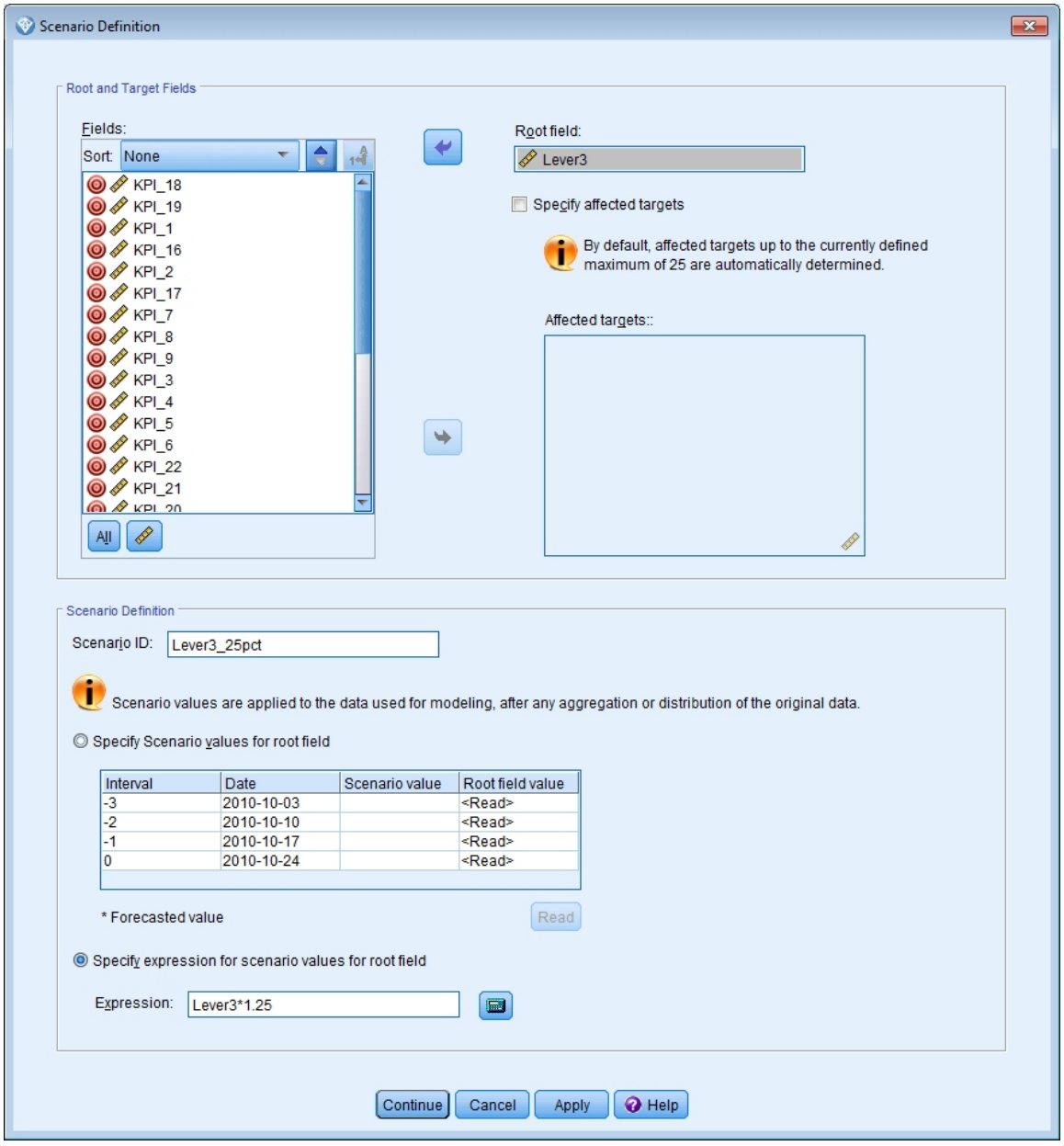


圖 *410.* 實務定義

8. 將 *Lever3* 移至**根欄位**方框來檢查實務週期中 *Lever3* 的指定值如何影響受 *Lever3* 因果影響的其他數列的預測。
9. 針對實務 ID 輸入 `Lever3_25pct`。
10. 選取**為根欄位的實務值指定表示式**並針對該表示式輸入 `Lever3*1.25`。

    此設定指定實務週期中 *Lever3* 的值比觀察值大 25%。針對更複雜的表示式，您可以透過按一下計算機圖示來使用表示式建置器。
11. 按一下「繼續」。
12. 重複步驟 10 - 14 來定義實務，其根欄位為 *Lever3*，實務 ID 為 `Lever3_50pct`，表示式為 `Lever3*1.5`。

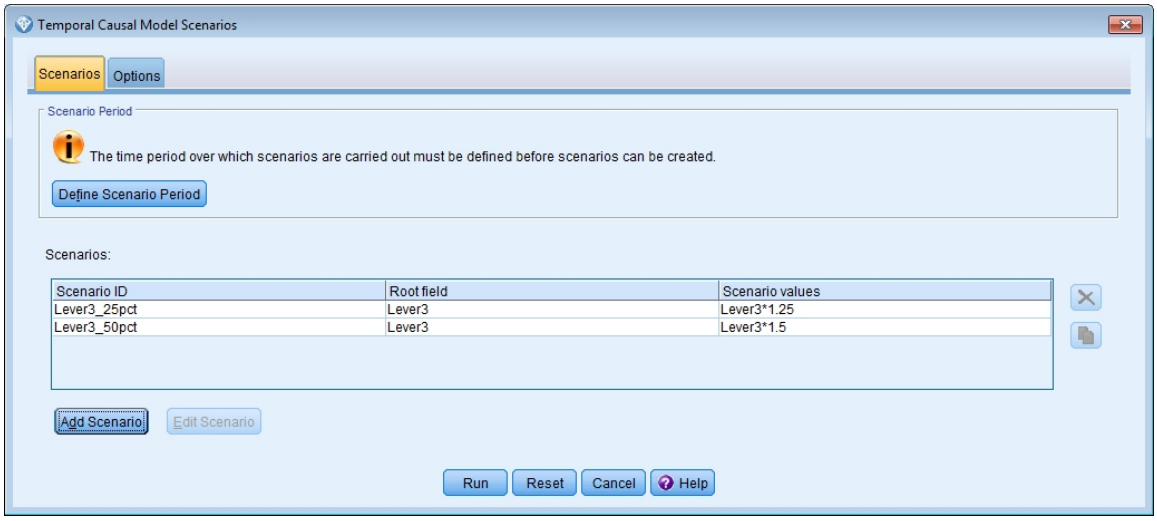
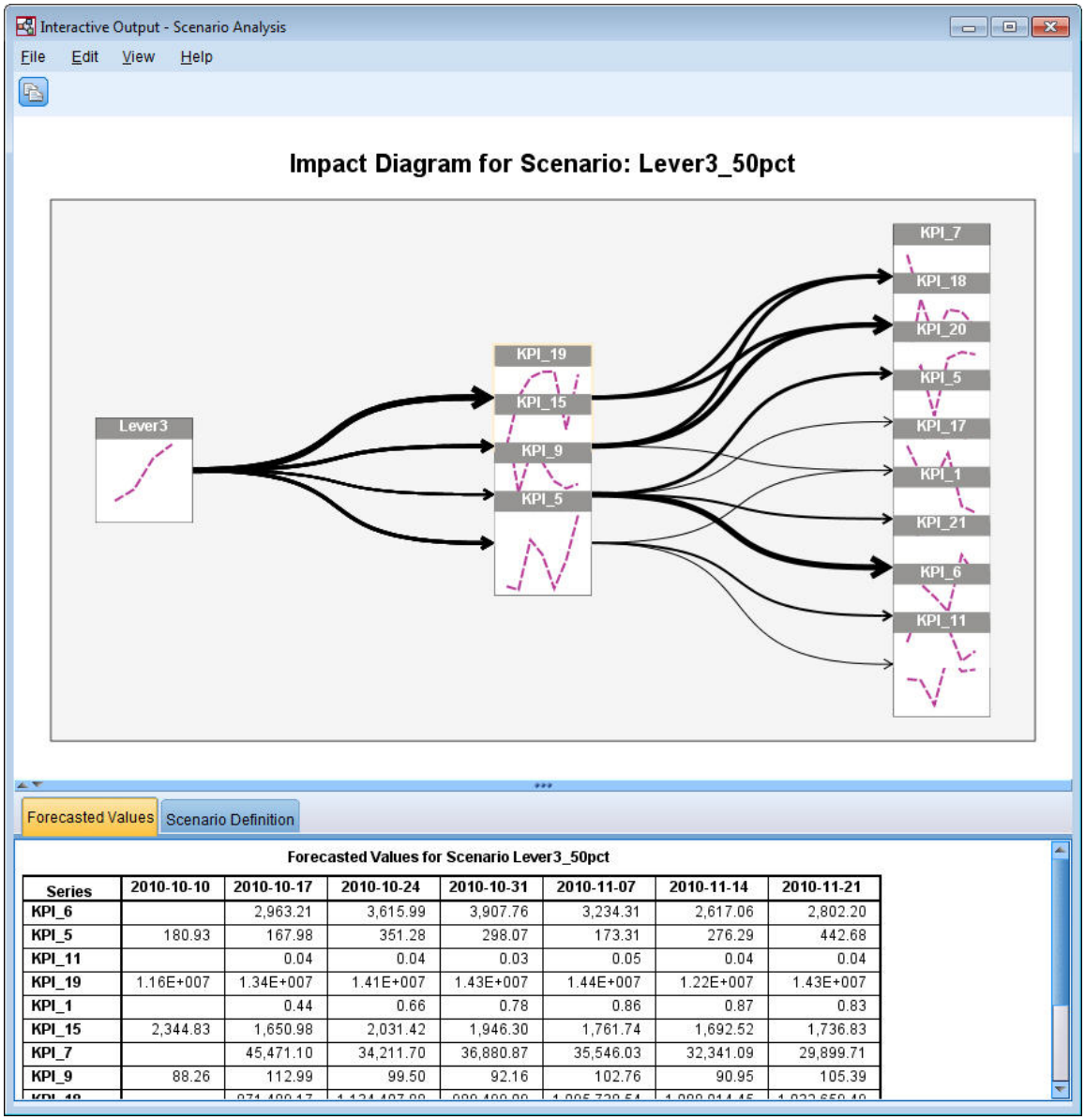


圖 *411.* 實務

13. 按一下**選項**標籤，並針對受影響目標的最大層次輸入 *2*。
14. 按一下「**執行**」。
15. 按兩下「檢視器」中 *Lever3_50pct* 的「影響圖表」以啟動它。



Series	2010-10-10	2010-10-17	2010-10-24	2010-10-31	2010-11-07	2010-11-14	2010-11-21
KPI_6		2,963.21	3,615.99	3,907.76	3,234.31	2,617.06	2,802.20
KPI_5	180.93	167.98	351.28	298.07	173.31	276.29	442.68
KPI_11		0.04	0.04	0.03	0.05	0.04	0.04
KPI_19	1.16E+007	1.34E+007	1.41E+007	1.43E+007	1.44E+007	1.22E+007	1.43E+007
KPI_1		0.44	0.66	0.78	0.86	0.87	0.83
KPI_15	2,344.83	1,650.98	2,031.42	1,946.30	1,761.74	1,692.52	1,736.83
KPI_7		45,471.10	34,211.70	36,880.87	35,546.03	32,341.09	29,899.71
KPI_9	88.26	112.99	99.50	92.16	102.76	90.95	105.39

圖 *412.* 實務 *Lever3_50pct* 的影響圖表

影響圖表會顯示受根數列 *Lever3* 影響的數列。會顯示兩個效果層次，因為您針對受影響目標的最大層次指定了 *2*。

「預測值」表格包括受 *Lever3* 影響的所有數列的預測，最多到第二個效果層次。第一個效果層次中的目標數列預測開始於實務週期開始之後的第一個時段。在此範例中，第一個層次中的目標數列預測開始於 2010-10-10。第二個效果層次中的目標數列預測開始於實務週期開始之後的第二個時段。在此範例中，第二個層次中的目標數列預測開始於 2010-10-17。預測的交錯本質反映出時間數列模型基於輸入落階值的事實。

16. 按一下 *KPI_5* 的節點以產生詳細的序列圖。

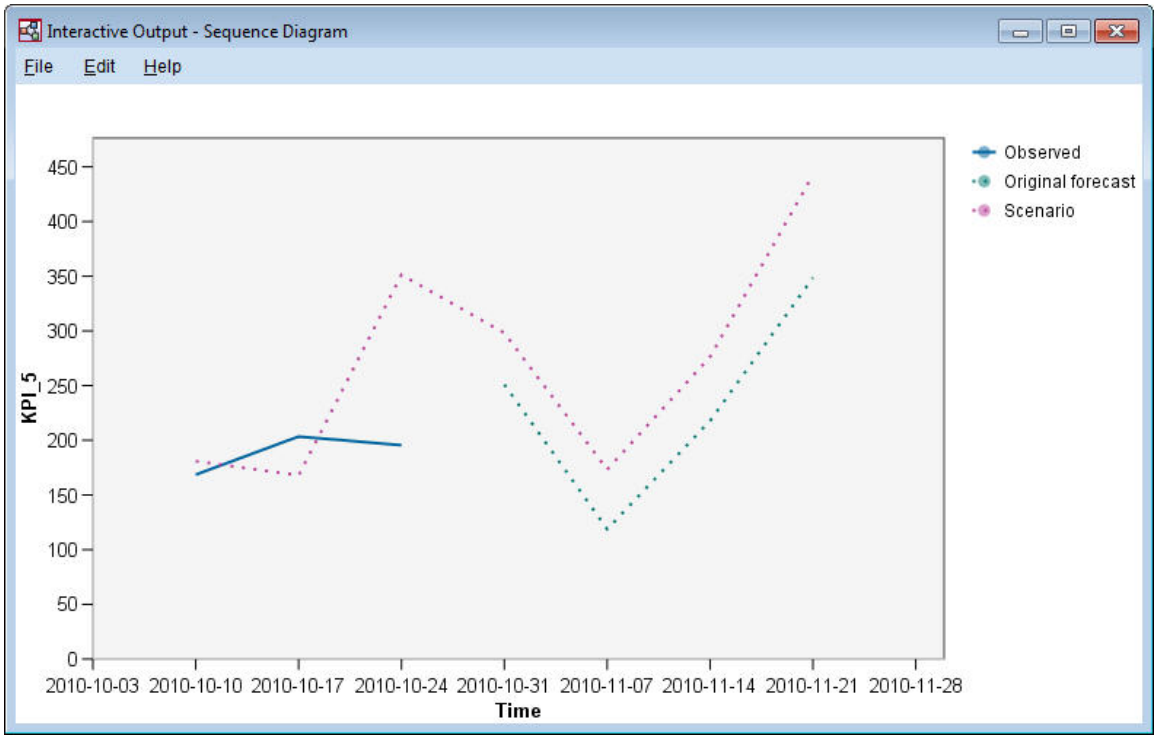


圖 *413. KPI_5* 的序列圖

序列圖會顯示實務中的預測值，它還會顯示沒有實務時數列的值。當實務週期包含估計週期內的時間時，會顯示數列的觀察值。針對超出估計週期結束的時間，會顯示原始預測。

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。

索引

索引順序以中文字，英文字，及特殊符號之次序排列。

〔十二劃〕

〔十三劃〕

〔十四劃〕

〔十五劃〕

〔十六劃〕

〔十七劃〕

〔十八劃〕

〔十九劃〕

〔二十一劃〕

〔二十三劃〕

C

E

I

M

S

U

W

Printed in Taiwan