IBM SPSS Modeler 17 应用程序指南

注释

在使用本信息及其支持的产品前，请阅读第 333 页的『声明』中的信息。

产品信息

本版本适用于 IBM(r) SPSS(r) Modeler V17.0.0 及所有后续发行版和修订版，直到在新版本中另有声明为止。

目录

第 1 章 关于 IBM SPSS Modeler

IBM® SPSS® Modeler 是一组数据挖掘工具，通过这些工具可以采用商业技术快速建立预测性模型，并将其应用于商业活动，从而改进决策过程。IBM SPSS Modeler 参照行业标准 CRISP-DM 模型设计而成，可支持从数据到更优商业成果的整个数据挖掘过程。

IBM SPSS Modeler 提供了各种来源于机器学习、人工智能和统计学的建模方法。“建模”选用板中提供的方法使您可以根据数据派生新信息，并开发预测模型。每种方法各有所长，并且最适合于解决特定类型的问题。

SPSS Modeler 可以作为独立产品购买，也可以作为客户机与 SPSS Modeler Server 配合使用。另外，还提供了很多其他选项，以下各节概述了这些选项。有关更多信息，请参阅 http://www.ibm.com/software/analytics/spss/products/modeler/。

IBM SPSS Modeler 产品

IBM SPSS Modeler 系列产品及相关联的软件由以下部分组成。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- 用于 IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 适配器

IBM SPSS Modeler

SPSS Modeler 是产品的完整功能版本，您可以在个人计算机上安装并运行此版本。可以在本地方式下将 SPSS Modeler 作为独立产品运行，也可以在分布式方式下将其与 IBM SPSS Modeler Server 配合使用，以提高大型数据集的性能。

借助 SPSS Modeler，您可以快速直观地构建准确的预测模型而无需进行编程。通过使用独特的可视界面，可以轻松地实现数据挖掘过程的可视化。在本产品中提供的高级分析的支持下，您可以发现数据中先前隐藏的模式和趋势。可以对结果进行建模并了解影响结果的因素，这使您可以利用业务机会并降低风险。

现已推出两个版本的 SPSS Modeler: SPSS Modeler Professional 和 SPSS Modeler Premium。请参阅主题 第 2 页的『IBM SPSS Modeler 的版本』以获取更多信息。

IBM SPSS Modeler Server

SPSS Modeler 使用客户端/服务器体系结构将资源集约型操作的请求分发给功能强大的服务器软件，因而使大数据集的传输速度大大加快。

SPSS Modeler Server 是需要单独许可证的产品，在分布式分析方式下它在服务器主机上与一个或多个 IBM SPSS Modeler 安装一起连续运行。通过此方式，SPSS Modeler Server 极大地提高了大型数据集的性能，因为可以在服务器上完成内存密集型操作，而无需将数据下载到客户端计算机。IBM SPSS Modeler Server 还提供对 SQL 优化和数据库内建模功能的支持，从而在性能和自动化方面提供更多优势。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是用于管理多个 SPSS Modeler Server 配置选项（这些选项还可以通过选项文件进行配置）的图形应用程序。此应用程序提供了用于监视和配置 SPSS Modeler Server 安装的控制台用户界面，并且可供当前的 SPSS Modeler Server 客户免费使用。应用程序只能安装在 Windows 计算机上；但是它可以管理安装在任何受支持平台上的服务器。

IBM SPSS Modeler Batch

数据挖掘通常是交互过程，因此，还可以从命令行运行 SPSS Modeler 而不需要图形用户界面。例如，您可能具有长时间运行或重复任务，并且希望在用户不进行干预的情况下执行这些任务。SPSS Modeler Batch 是该产品的一个特殊版本，可提供对 SPSS Modeler 完整分析性能的支持，而无需访问常规的用户界面。要使用 SPSS Modeler Batch，需要 SPSS Modeler Server。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher 是一个工具，它使您能够创建 SPSS Modeler 流的打包版本，该版本的流可以由外部运行时引擎运行或者可以嵌入在外部应用程序中。通过此方式，您可以发布和部署完整的 SPSS Modeler 流，以便在未安装 SPSS Modeler 的环境中进行使用。SPSS Modeler Solution Publisher 作为 IBM SPSS Collaboration and Deployment Services - Scoring 服务的组成部分进行分发，需要单独的许可证。通过此许可证，您将接收 SPSS Modeler Solution Publisher Runtime，它使您可以执行已发布的流。

有关 SPSS Modeler Solution Publisher 的更多信息，请参阅 IBM SPSS Collaboration and Deployment Services 文档。IBM SPSS Collaboration and Deployment Services Knowledge Center 包含名为“IBM SPSS Modeler Solution Publisher”和“IBM SPSS Analytics Toolkit”的部分。

用于 IBM SPSS Collaboration and Deployment Services 的 IBM SPSS Modeler Server 适配器

提供了一些用于 IBM SPSS Collaboration and Deployment Services 的适配器，这些适配器使 SPSS Modeler 和 SPSS Modeler Server 可以与 IBM SPSS Collaboration and Deployment Services 存储库进行交互。通过这种方式，部署到该存储库的 SPSS Modeler 流可以由多个用户共享，也可以从瘦客户机应用程序 IBM SPSS Modeler Advantage 进行访问。将在适配器安装在主管该存储库的系统上。

IBM SPSS Modeler 的版本

现已推出下列版本的 SPSS Modeler。

SPSS Modeler Professional

SPSS Modeler Professional 提供处理大多数类型的结构化数据（例如 CRM 系统中跟踪的行为和交互、人口统计信息、采购行为和销售数据）所需要的所有工具。

SPSS Modeler Premium

SPSS Modeler Premium 是需要单独许可证的产品，它将 SPSS Modeler Professional 扩展为处理专门的数据（例如用于实体分析或社交网络的数据）以及无结构文本数据。SPSS Modeler Premium 由下列组件组成。

IBM SPSS Modeler Entity Analytics 在 IBM SPSS Modeler 预测分析的基础上添加了额外的维度。鉴于预测性分析尝试根据过去的数据来预测未来行为，实体分析侧重于通过解决记录自身中的身份冲突来提高当前数据的连贯性和一致性。身份可以指个人、组织、对象或可能存在不确定性的任何其他实体的身份。在许多领域，身份识别至关重要，例如客户关系管理、欺诈检测、反洗黑钱以及国家和国际安全。

IBM SPSS Modeler Social Network Analysis 将关于关系的信息转换为字段，这些字段可描述个人和组社交行为的特征。使用介绍社交网络之下关系的数据，IBM SPSS Modeler Social Network Analysis 可识别影响网络中他人行为的社交领导。另外，您还可以确定哪些人员受其他网络参与者的影响最大。通过将这些结果与其他测量相结合，您可以创建复杂的个人档案，预测模型将以这些个人档案为基础。与未包含此社交信息的模型相比，包含此社交信息的模型表现更好。

IBM SPSS Modeler Text Analytics 使用先进的语言技术和自然语言处理 (NLP) 来快速处理各种各样的无结构文本数据、抽取和组织关键概念，以及将这些概念分组为类别。抽取的概念和类别可以和现有结构化数据中进行组合（例如人口统计学），并且可用于借助 IBM SPSS Modeler 的一整套数据挖掘工具来进行建模，以此实现更好更集中的决策。

IBM SPSS Modeler 文档

可以从 SPSS Modeler 的帮助菜单中获取在线帮助格式的文档。此文档包括 SPSS Modeler、SPSS Modeler Server 和 SPSS Modeler Solution Publisher 的文档以及《应用程序指南》和其他支持材料。

每个产品 DVD 的 *Documentation* 文件夹下都提供了该产品的 PDF 格式的完整文档（包括安装指示信息）。并且，还可以从 Web 下载安装文档: http://www.ibm.com/support/docview.wss?uid=swg27043831。

这两种格式的文档都可以从 SPSS Modeler Knowledge Center (http://www-01.ibm.com/support/knowledgecenter/SS3RA7_17.0.0.0) 获得。

SPSS Modeler Professional 文档

SPSS Modeler Professional 文档套件（安装指示信息除外）如下。

- **IBM SPSS Modeler 用户指南**。使用 SPSS Modeler 的一般使用介绍，包括如何构建数据流、处理缺失值、生成 CLEM 表达式、处理项目和报告以及将用于部署的流打包为 IBM SPSS Collaboration and Deployment Services、预测应用程序或 IBM SPSS Modeler Advantage。
- **IBM SPSS Modeler Source、Process 和 Output 节点**。描述用于以不同格式读取、处理和输出数据的所有节点。实际上这表示所有节点而非建模节点。
- **IBM SPSS Modeler Modeling 节点**。有关用于创建数据挖掘模型的所有节点的描述。IBM SPSS Modeler 提供了各种借助机器学习、人工智能和统计学的建模方法。
- **IBM SPSS Modeler 算法指南**。描述 IBM SPSS Modeler 中使用的建模方法的数学基础。本指南仅以 PDF 格式提供。
- **IBM SPSS Modeler 应用程序指南**。本指南中的示例旨在为具体的建模方法和技术提供具有针对性的简介。还可以在“帮助”菜单中查阅本指南的在线版本。请参阅主题第 4 页的『应用程序示例』，了解更多信息。
- **IBM SPSS Modeler Python 脚本编制和自动化**。通过编写 Python 脚本实现系统自动化的相关信息，其中包括可以用于处理节点和流的属性的信息。
- **IBM SPSS Modeler 部署指南**。有关在 IBM SPSS Collaboration and Deployment Services Deployment Manager 中以处理作业的步骤形式运行 IBM SPSS Modeler 流和方案的信息。
- **IBM SPSS Modeler CLEF 开发者指南**。CLEF 提供了将第三方程序（例如，数据处理例程或建模算法）作为节点集成到 IBM SPSS Modeler 的功能。
- **IBM SPSS Modeler 数据库内挖掘指南**。有关如何利用数据库的功能通过第三方算法来改进性能并增强分析功能的信息。
- **IBM SPSS Modeler Server 管理与性能指南**。提供有关如何配置和管理 IBM SPSS Modeler Server 的信息。

- **IBM SPSS Modeler Administration Console 用户指南**。有关安装和使用控制台用户界面以监视和配置 IBM SPSS Modeler Server 的信息。控制台实现为 Deployment Manager 应用程序的插件。
- **IBM SPSS Modeler CRISP-DM 指南**。借助 CRISP-DM 方法进行 SPSS Modeler 数据挖掘的分步指南。
- **IBM SPSS Modeler Batch 用户指南**。提供在批处理方式下使用 IBM SPSS Modeler 的完整指导，包括批处理方式执行和命令行自变量的详细信息。本指南仅以 PDF 格式提供。

SPSS Modeler Premium 文档

SPSS Modeler Premium 文档套件（安装指示信息除外）如下。

- **IBM SPSS Modeler Entity Analytics 用户指南。** 提供有关通过 SPSS Modeler 使用实体分析的信息，涵盖存储库安装和配置、实体分析节点和管理任务。
- **IBM SPSS Modeler Social Network Analysis 用户指南。** 这是使用 SPSS Modeler 执行社交网络分析（包括组分析和扩散分析）的指南。
- **SPSS Modeler Text Analytics 用户指南。**提供有关通过 SPSS Modeler 使用文本分析的信息，涵盖文本挖掘节点、交互式工作台、模板和其他资源。

应用程序示例

SPSS Modeler 中的数据挖掘工具可以帮助解决很多业务和组织问题，应用程序示例将提供有关特定建模方法和技术的简要的针对性说明。此处使用的数据集比某些数据挖掘器管理的大量数据存储要小得多，但涉及的概念和方法应可扩展到实际的应用程序。

可以通过在 SPSS Modeler 中的“帮助”菜单中单击**应用程序示例**来访问示例。数据文件和样本流安装在产品安装目录下的 *Demos* 文件夹中。请参阅主题『Demos 文件夹』，了解更多信息。

数据库建模示例。请参阅《IBM SPSS Modeler 数据库内挖掘指南》中的示例。

脚本编制示例。请参阅《IBM SPSS Modeler 脚本编制和自动化指南》中的示例。

Demos 文件夹

与应用程序示例一起使用的数据文件和样本流安装在产品安装目录下的 *Demos* 文件夹中。可以在 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组访问此文件夹，也可以通过单击“文件打开”对话框中最近访问的目录列表中的 *Demos* 来进行访问。

第 2 章 IBM SPSS Modeler 概述

新手入门

作为一种数据挖掘应用程序，IBM SPSS Modeler 提供了用以寻找大数据集中有用关系的策略性方法。与更传统的统计方法相比，您在开始时不必知道您要寻找什么。您可以通过拟合不同的模型和研究不同的关系来探索您的数据，直到发现有用的信息。

启动 IBM SPSS Modeler

要启动此应用程序，请单击:

开始 > 所有程序 > IBM SPSS Modeler 16 > IBM SPSS Modeler 16

主窗口将在几秒钟后显示。

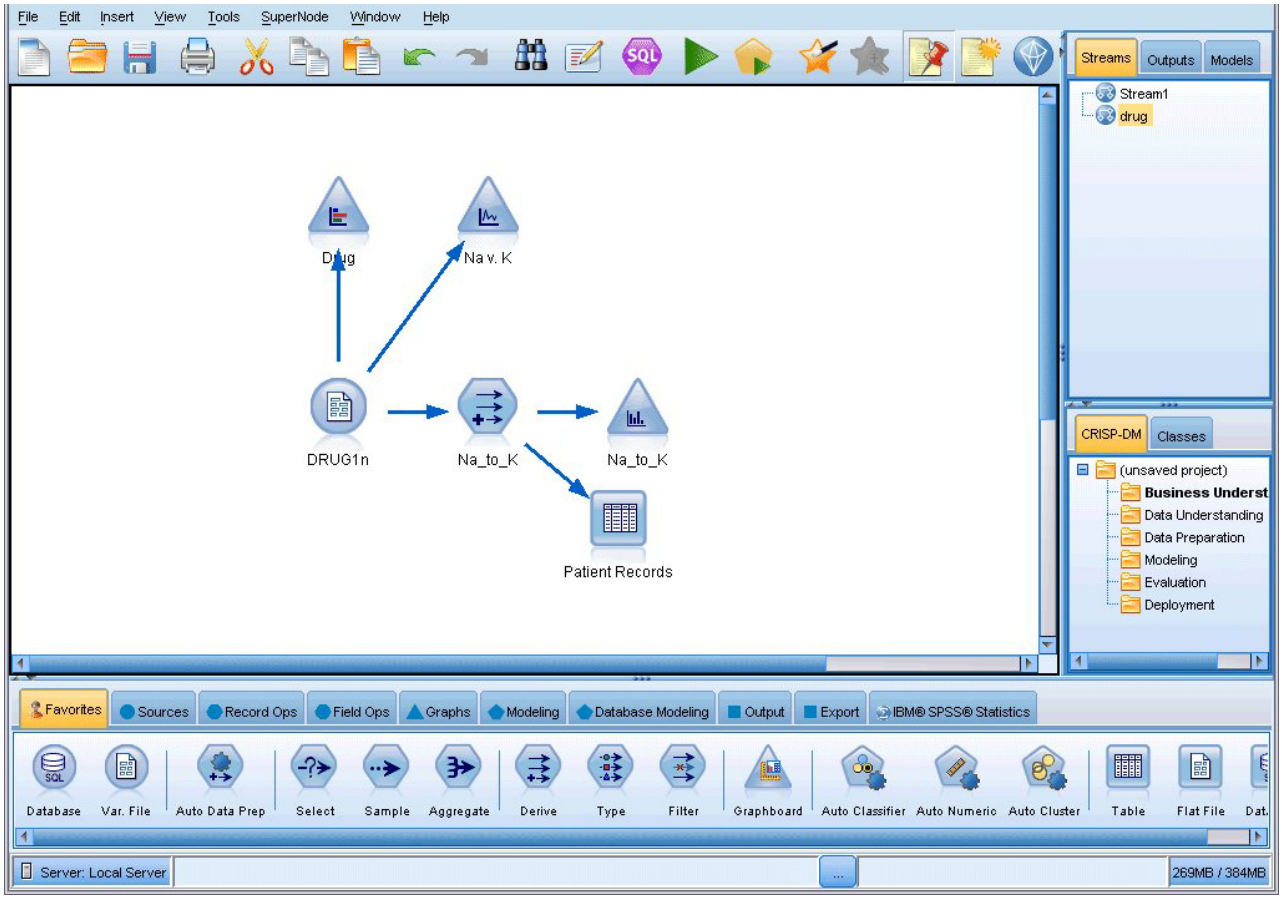


图 1. IBM SPSS Modeler 主应用程序窗口

从命令行启动

您可以使用操作系统的命令行来如下启动 IBM SPSS Modeler:

1. 在安装了 IBM SPSS Modeler 的计算机上，打开 DOS 或命令提示符窗口。
2. 要以交互方式启动 IBM SPSS Modeler 界面，请输入 `modelerclient` 命令，然后输入所需的参数；例如:

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

可用参数（标记）允许您连接到服务器、装入流、运行脚本或根据需要指定其他参数。

正在连接到 IBM SPSS Modeler Server

IBM SPSS Modeler 可作为独立的应用程序运行，或作为直接连接到 IBM SPSS Modeler Server 的客户端运行，或者作为通过进程协调器 (COP) 插件从 IBM SPSS Collaboration and Deployment Services 连接到 IBM SPSS Modeler Server 或服务器集群的客户端运行。当前连接状态显示在 IBM SPSS Modeler 窗口的左下角。

无论何时想连接到服务器，都请手动输入想要连接的服务器的名称或选择之前已定义的名称。但是，如果您拥有 IBM SPSS Collaboration and Deployment Services，则可以从“服务器登录”对话框搜索服务器列表或服务器集群列表。可以通过进程协调器执行浏览网络上运行的 Statistics 服务的功能。

连接到服务器

1. 在“工具”菜单上，单击**服务器登录**。这将打开“服务器登录”对话框。或者，双击 IBM SPSS Modeler 窗口的连接状态区域。
2. 使用该对话框指定要连接到本地服务器计算机的选项或从表中选择连接。
   - 单击 **添加** 或 **编辑** 以添加或编辑连接。请参阅主题第 7 页的『添加并编辑 IBM SPSS Modeler Server 连接』以获取更多信息。
   - 单击**搜索**以访问进程协调器中的服务器或服务器集群。请参阅主题第 7 页的『搜索 IBM SPSS Collaboration and Deployment Services 中的服务器』以获取更多信息。

   服务器表。此表包含已定义的服务器连接集。该表显示缺省连接、服务器名称、描述和端口号。您可以手动添加新的连接，以及选择或搜索现有连接。要将特定的服务器设置为缺省连接，请在表中“缺省”列中为此连接选择复选框。

   缺省数据路径。指定用于服务器计算机上的数据的路径。单击省略号按钮（**...**），以浏览至所需要的位置。

   设置凭证。不选中此复选框可启用**单点登录**功能，该功能尝试使您使用本地计算机用户名和密码详细信息登录服务器。如果无法使用单点登录，或您选中此复选框以禁用单点登录（例如，登录管理员帐户），则启用以下字段让您输入您的凭证。

   用户标识。输入用于登录服务器的用户名。

   密码。输入与指定用户名相关联的密码。

   域。指定用于登录服务器的域。只有服务器计算机与客户计算机处于不同的 Windows 域时，才需要域名。
3. 单击 **确定** 以完成此连接。

断开与服务器的连接

1. 在“工具”菜单上，单击**服务器登录**。这将打开“服务器登录”对话框。或者，双击 IBM SPSS Modeler 窗口的连接状态区域。
2. 在此对话框中，选择“本地服务器”，然后单击**确定**。

添加并编辑 IBM SPSS Modeler Server 连接

您可以在“服务器登录”对话框中手动编辑或添加服务器连接。单击“添加”可以访问空的“添加/编辑服务器”对话框，在此对话框中可以输入服务器连接的详细信息。在“服务器登录”对话框中选择现有连接并单击“编辑”，将打开“添加/编辑服务器”对话框，其中包含所选连接的详细信息，以便可以进行任何更改。

注: 不能编辑从 IBM SPSS Collaboration and Deployment Services 中添加的服务器连接，因为名称、端口及其他详细信息已在 IBM SPSS Collaboration and Deployment Services 中做过定义。最佳实践指出，应该使用相同的端口与 IBM SPSS Collaboration and Deployment Services 和 SPSS Modeler Client 进行通信。这些端口可以设置为 options.cfg 文件中的 `max_server_port` 和 `min_server_port`。

添加服务器连接

1. 在“工具”菜单上，单击**服务器登录**。这将打开“服务器登录”对话框。
2. 在此对话框中，单击 **添加**。将打开“服务器登录: 添加/编辑服务器”对话框。
3. 输入服务器连接的详细信息，然后单击**确定**保存此连接并返回“服务器登录”对话框。

- **服务器**。指定可用服务器或从列表选择一个服务器。服务器计算机的名称可以使用字母数字（例如 *myserver*）或指派给服务器计算机的 IP 地址（例如，202.123.456.78）。
- **端口**。指定服务器要侦听的端口号。如果缺省设置不可用，请向系统管理员索取正确的端口号。
- **描述**。输入此服务器连接的可选描述。
- **确保安全连接（使用 SSL）**。指定是否应该使用 SSL（**安全套接字层**）连接。SSL 是一个常用协议，用于确保通过网络发送的数据的安全。要使用此功能，必须在承载 IBM SPSS Modeler Server 的服务器中启用 SSL。必要时请与本地管理员联系，以了解详细信息。

编辑服务器连接

1. 在“工具”菜单上，单击**服务器登录**。这将打开“服务器登录”对话框。
2. 在此对话框中，选择希望编辑的连接，然后单击 **编辑**。将打开“服务器登录: 添加/编辑服务器”对话框。
3. 更改服务器连接详细信息，然后单击**确认**保存更改内容并返回至“服务器登录”对话框。

搜索 IBM SPSS Collaboration and Deployment Services 中的服务器

在 IBM SPSS Collaboration and Deployment Services 中，可以使用进程协调器选择网络上可用的服务器或服务器集群，从而代替手动输入服务器连接。服务器集群是一组服务器，进程协调器从这组服务器中确定最适合对处理要求作出响应的服务器。

尽管可在“服务器登录”对话框中手动添加服务器，但通过搜索可用的服务器，可在无需知道正确服务器名称和端口号的情况下连接到服务器。此信息是自动提供的。但仍需输入正确的登录信息，如用户名、域和密码。

注: 如果您无权访问进程协调器功能，那么仍然可以手动输入要连接的服务器名称或选择先前已定义的名称。请参阅主题『添加并编辑 IBM SPSS Modeler Server 连接』以获取更多信息。

搜索服务器和服务器集群

1. 在“工具”菜单上，单击**服务器登录**。这将打开“服务器登录”对话框。
2. 在此对话框中，单击**搜索**打开“搜索服务器”对话框。如果在尝试浏览进程协调器时未登录到 IBM SPSS Collaboration and Deployment Services，则系统会提示您执行此项操作。
3. 从列表中选择服务器或服务器集群。
4. 单击**确定**以关闭对话框，然后将此连接添加到“服务器登录”对话框的表中。

更改 Temp 目录

IBM SPSS Modeler Server 执行的某些操作可能需要创建临时文件。缺省情况下， IBM SPSS Modeler 在系统临时目录下创建临时文件。可通过以下步骤更改临时目录的位置。

1. 创建新目录 *spss* 及其子目录 *servertemp*。
2. 编辑 *options.cfg* ，该文件位于 IBM SPSS Modeler 安装目录的 */config* 目录下。在此文件中编辑 `temp_directory` 参数，将其更改为: `temp_directory, "C:/spss/servertemp"`。
3. 完成此操作后，必须重新启动 IBM SPSS Modeler Server 服务。可通过单击 Windows 控制面板中的 **服务** 选项卡进行此服务重启操作。只需停止该服务然后将其重新启动即可激活所作的更改。重新启动机器也会重新启动该服务。

所有临时文件此时将写入该新目录。

注: 上述操作中最常见的错误是使用不正确的斜杠；应使用正斜杠。

启动多个 IBM SPSS Modeler 会话

如果需要同时启动一个以上的 IBM SPSS Modeler 会话，则必须对 IBM SPSS Modeler 和 Windows 的设置做一些更改。例如，如果您有两个独立的服务器许可证，并且希望从同一台客户机针对两台不同的服务器运行两个流，则需要对上述设置做一些更改。

要启用多个 IBM SPSS Modeler 会话:

1. 单击:

   开始 > 所有程序 > IBM SPSS Modeler 16
2. 在 IBM SPSS Modeler 17 快捷键（带箭头的图标）上右键单击并选择**属性**。
3. 在**目标**文本框中，将 `-noshare` 添加到该字符串的结尾。
4. 在 Windows 资源管理器中选择:

   工具 > 文件夹
5. 在“文件类型”选项卡上选择“IBM SPSS Modeler 流”选项，然后单击 **高级**。
6. 在“编辑文件类型”对话框中，选择“使用 IBM SPSS Modeler 打开”，然后单击**编辑**。
7. 在**用于执行操作的应用程序**文本框中，在 **-stream** 参数前添加 `-noshare`。

IBM SPSS Modeler 界面概览

在数据挖掘过程中的每一个阶段，均可通过 IBM SPSS Modeler 易于使用的界面来邀请特定业务的专家。建模算法（如预测、分类、细分和关联检测）可确保得到强大而准确的模型。模型结果可以方便地部署和读入到数据库、 IBM SPSS Statistics 和各种其他应用程序中。

使用 IBM SPSS Modeler 即处理数据的三个步骤。

- 首先，将数据读入 IBM SPSS Modeler。
- 接着，通过一系列处理来运行数据。
- 最后，将数据发送至目标。

这一操作序列称为 **数据流** ，因为数据以一条条记录的形式，从数据源开始，依次经过各种操纵，最终到达目标（模型或某种数据输出）。

图 2. 简单流

IBM SPSS Modeler 流画布

流工作区是 IBM SPSS Modeler 窗口的最大区域，也是您构建和操纵数据流的位置。

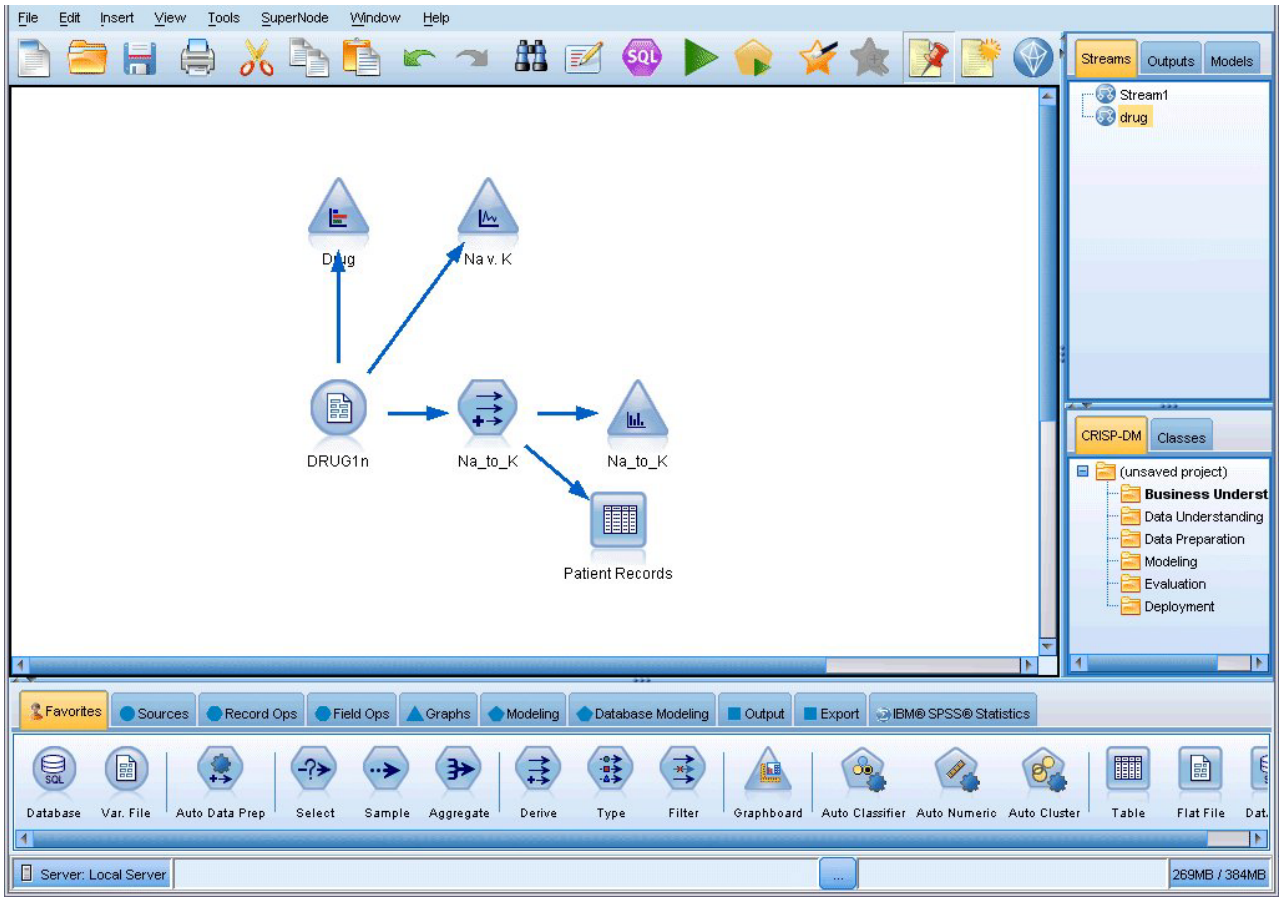


图 3. IBM SPSS Modeler 工作空间（缺省视图）

流是在界面的主画布中通过绘制与业务相关的数据操作图来创建的。每个操作都用一个图标或 **节点** 表示，这些节点通过 **流** 链接在一起，流表示数据在各个操作之间的流动。

在 IBM SPSS Modeler 中，可以在同一流工作区或通过打开新的流工作区来一次处理多个流。会话期间，流存储在 IBM SPSS Modeler 窗口右上角的“流”管理器中。

节点选用板(N)

IBM SPSS Modeler 中的大部分数据和建模工具位于 **节点选用板** 中，该选用板位于流工作区下方窗口的底部。

例如，可以使用“记录选项”选用板选项卡中包含的节点对数据**记录**执行操作，如选择、合并和追加等。

要将节点添加到画布中，请双击“节点”选用板中的图标或者将节点拖放到画布上。随后可将各个图标连接以创建一个表示数据流动的 **流** 。

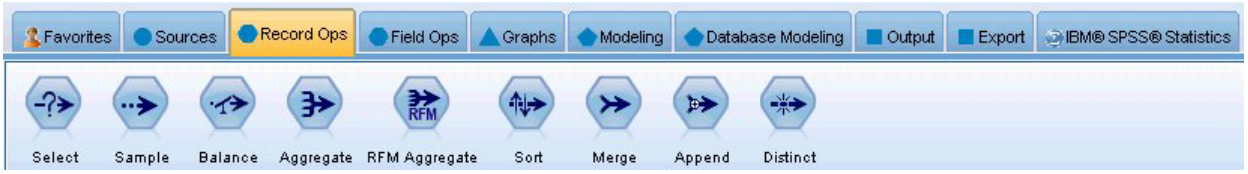


图 4. 节点选用板中的“记录选项”选项卡

每个选用板选项卡均包含一组不同的流操作阶段中使用的相关节点，如:

- **源**。此类节点将数据引入 IBM SPSS Modeler 中。
- **记录选项**。此类节点可对数据**记录**执行选择、合并和追加等操作。
- **字段选项**。此类节点可对数据**字段**执行操作，如过滤、导出新字段和确定给定字段的测量级别等。
- **图形**。此类节点可在建模前后以图表形式显示数据。图形包括散点图、直方图、网络节点和评估图表。
- **建模**此类节点可使用 IBM SPSS Modeler 中提供的建模算法，例如神经网络、决策树、聚类算法和数据序列等。
- **数据库建模**。此类节点使用 Microsoft SQL Server、IBM DB2 和 Oracle 以及 Netezza 数据库中提供的建模算法。
- **输出**。节点生成可在 IBM SPSS Modeler 中查看的数据、图表和模型等多种输出结果。
- **导出**。节点生成可在外部应用程序（如 IBM SPSS Data Collection 或 Excel）中查看的多种输出。
- **IBM SPSS Statistics.** 此类节点从 IBM SPSS Statistics 中导入数据或将数据导出到其中，并用于运行 IBM SPSS Statistics 过程。

随着对 IBM SPSS Modeler 的熟悉，您也可以自定义供自己使用的选用板内容。

“节点”选用板下方是一个报告窗格，此窗格提供各种操作的进度反馈，例如何时将数据读入数据流中。“节点”选用板下方还有一个状态窗格，此窗格提供有关应用程序当前正在执行的操作的信息以及何时需要用户反馈的指示信息。

IBM SPSS Modeler 管理器

管理器窗格位于窗口右上角。此窗格包含用于管理流、输出和模型的三个选项卡。

可以使用“流”选项卡打开、重命名、保存和删除会话中创建的流。

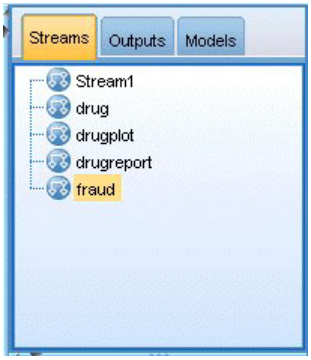


图 5. “流”选项卡

图 6. “输出”选项卡

“输出”选项卡中包含由 IBM SPSS Modeler 中的流操作生成的各类文件，如图形和表。您可以显示、保存、重命名和关闭此选项上列出的表、图形和报告。

图 7. 包含模型块的“模型”选项卡

在管理器选项卡中，“模型”选项卡具有最强大的功能。该选项卡中包含所有模型**块**，这些模型块包含针对当前会话在 IBM SPSS Modeler 中生成的模型。可以直接从“模型”选项卡浏览这些模型或者将它们添加到画布内的流中。

IBM SPSS Modeler 项目

窗口右侧底部是工程窗格，用于创建和管理数据挖掘**工程**（与数据挖掘任务相关的文件组）。可以通过两种方法来查看您在 IBM SPSS Modeler 中创建的项目：“类”视图和 CRISP-DM 视图。

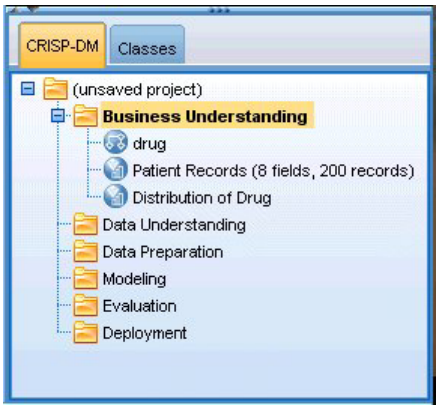


图 8. CRISP-DM 视图

依据业内认可的非专利方法“跨行业数据挖掘过程标准”，CRISP-DM 选项卡提供了一种项目组织方法。无论是有经验的数据挖掘人员还是新手，使用 CRISP-DM 工具都会使您事半功倍。

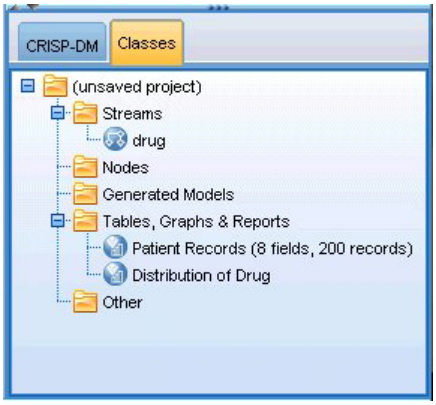


图 9. “类”视图

“类”选项卡提供了一种在 IBM SPSS Modeler 中按类别（即按照所创建对象的类别）组织您工作的方式。此视图在获取数据、流、模型的详尽目录时十分有用。

IBM SPSS Modeler工具栏

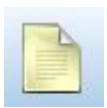
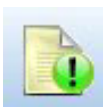
IBM SPSS Modeler 窗口顶部有一个图标工具栏，其中包含许多有用功能。下面是一些工具栏按钮及其功能。

创建新流

打开流

保存流

打印当前流

	剪切并移到剪贴板		复制到剪贴板
	粘贴选择		撤销上一次操作
	重做		搜索节点
	编辑流属性		预览 SQL 生成
	运行当前流		运行流选择
	停止流（仅在流处于运行状态时可用）		添加超节点
	放大（仅限于 SuperNodes）		缩小（仅限于 SuperNodes）
	流中无标记		插入注释
	隐藏流标记（如果有）		显示隐藏的流标记
	在 IBM SPSS Modeler Advantage 中打开流		

流标记由流注释、模型链接和评分分支指示组成。

在《*IBM SPSS 建模节点*》指南中介绍了模型链接。

定制工具栏

您可以更改工具栏的各个方面，例如:

- 是否显示
- 图标是否有可用工具提示
- 使用大或小图标

要打开或关闭工具栏显示，请执行以下操作:

1. 在主菜单中，单击:

查看 > 工具栏 > 显示

要更改工具提示或图标大小设置，请执行以下操作:

1. 在主菜单中，单击:

查看 > 定制 > 显示

根据需要单击**显示工具提示**或**大按钮**。

自定义 IBM SPSS Modeler 窗口

使用 IBM SPSS Modeler 界面各部分之间的分界线，可以调整工具的大小或关闭某些工具以满足个人偏好。例如，如果要处理大型流，那么可以使用每条分界线上的小箭头来关闭节点选用板、管理器窗格和项目窗格。这样可以最大化流画布，从而为处理大型流或多个流提供足够的工作空间。

此外，从“视图”菜单上，单击**节点选用板**、**管理器**或**工程**可打开或关闭这些项目的显示。

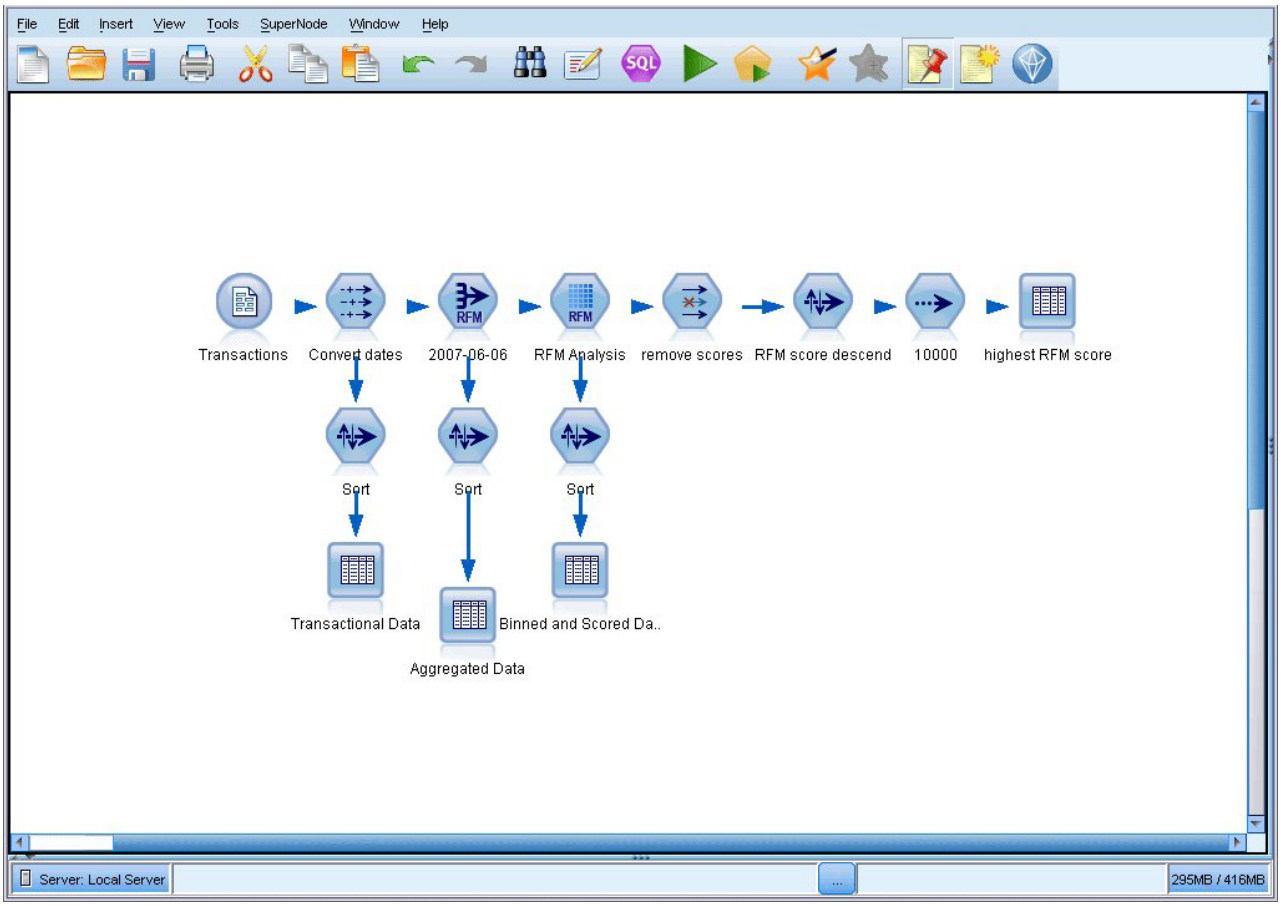


图 10. 最大化的流画布

另外一种关闭节点选用板和管理器以及工程窗格的方法是: 垂直或水平移动 IBM SPSS Modeler 窗口侧面或底部的滚动条，将流工作区当作滚动页面使用。

您也可以控制屏幕标记的显示，此标记由流注释、模型链接和评分分支指示组成。要打开或关闭此显示，请单击:

视图 > 流标记

更改流的图标尺寸

您可以通过下列方式更改流图标的大小。

- 流属性设置
- 流中的弹出菜单
- 使用键盘

您可以调整整个流视图的大小，将其调整为标准图标尺寸的 8% 至 200% 之间的某个尺寸。

要调整整个流的大小（流属性方法）

1. 从主菜单中，选择

   工具 > 流属性 > 选项 > 布局。
2. 从“图标大小”菜单中选择所需大小。
3. 单击**应用**以查看结果。
4. 单击**确定**保存更改。

要调整整个流的大小（菜单方法）

1. 右键单击画布上的流背景。
2. 选择**图标尺寸**，并选择所需的大小。

要调整整个流的大小（键盘方法）

1. 同时按住主键盘上的 Ctrl + [-] 来缩小至下一个较小的尺寸。
2. 同时按住主键盘上的 Ctrl + Shift + [+] 来放大至下一个较大的尺寸。

获取复杂流的总体视图时，此功能尤其有用。您还可以使用此功能来最大程度地减少打印流所需的页面数。

在 IBM SPSS Modeler 中使用鼠标

IBM SPSS Modeler 中最常见的鼠标用法如下所示:

- **单击。**使用鼠标右键或左键从菜单中选择选项、打开弹出菜单，以及访问其他标准控件和选项。单击并按住按键可移动和拖动节点。
- **双击。**双击鼠标左键可将节点放入于流画布中以及编辑现有节点。
- **单击鼠标中键。**单击鼠标中键并拖动光标可连接流画布中的节点。双击鼠标中键可断开某个节点的连接。如果没有三键鼠标，可在单击并拖动鼠标时通过按 Alt 键来模拟此功能。

使用快捷键

IBM SPSS Modeler 中的许多可视化编程操作均有与之关联的快捷键。例如，可通过单击某个节点并按键盘上的 Delete 键将此节点删除。同样地，可在按住 Ctrl 键的同时按 S 键来快速保存某个流。控制命令（例如此命令）由 Ctrl 和其他键的组合指定，例如 Ctrl+S。

标准 Windows 操作中使用了大量快捷键，例如使用 Ctrl+X 来执行剪切操作。IBM SPSS Modeler 不仅支持这些快捷键，而且还支持下列应用程序特定的快捷键。

注: 在某些情况下，IBM SPSS Modeler 中使用的旧快捷键与标准 Windows 快捷键相冲突。支持将这些旧快捷键与 Alt 键组合使用。例如，可以使用 Ctrl+Alt+C 来打开或关闭高速缓存。

表 1. 支持的快捷键

快捷键	函数
Ctrl+A	全选
Ctrl+X	剪切(T)
Ctrl+N	新建流
Ctrl+O	打开流
Ctrl+P	打印(P)
Ctrl+C	复制(C)
Ctrl+V	粘贴(P)
Ctrl+Z	撤销
Ctrl+Q	选择选定节点的所有下游节点
Ctrl+W	全部不选下游节点（使用 Ctrl+Q 进行切换）
Ctrl+E	从选定节点运行
Ctrl+S	保存当前流
Alt+箭头键	向所使用的箭头方向移动流画布上的选定节点
Shift+F10	打开选定节点的弹出菜单

表 2. 支持的旧热键快捷键

快捷键	函数
Ctrl+Alt+D	复制节点
Ctrl+Alt+L	加载节点
Ctrl+Alt+R	重命名节点
Ctrl+Alt+U	创建用户输入节点
Ctrl+Alt+C	切换高速缓存开关
Ctrl+Alt+F	刷新高速缓存
Ctrl+Alt+X	扩展超节点
Ctrl+Alt+Z	放大/缩小
删除	删除节点或连接

打印

可在 IBM SPSS Modeler 中打印下列对象:

- 流图表
- 图形(G)
- 表(T)
- 报告（来自报告节点和工程报告）
- 脚本（来自“流属性”、“独立脚本”或“超节点脚本”对话框）
- 模型（模型浏览器、包含当前内容的对话框选项卡、树查看器）
- 注解（使用输出的“注解”选项卡）

要打印对象:

- 要不预览就打印，请单击工具栏上的“打印”按钮。

- 要在打印前设置页面，请选择“文件”菜单中的**页面设置**。
- 要在打印前预览，请选择“文件”菜单中的**打印预览**。
- 要查看标准打印对话框中用于选择打印机以及指定外观的选项，请选择“文件”菜单中的**打印**。

实现 IBM SPSS Modeler 的自动化

由于高级数据挖掘往往是一个冗长的复杂过程，因此 IBM SPSS Modeler 包含对几种类型的编码和自动处理的支持。

- **表达式操作控制语言** (CLEM) 是一种用于分析和操作在 IBM SPSS Modeler 流中流动的数据的语言。数据挖掘人员可在流操作中广泛使用 CLEM 语言来执行根据成本和收入数据推导利润这样的简单任务，也可以执行将 Web 日志数据转换为具有有用信息的一系列字段和记录这样的复杂任务。
- **脚本编写**是用于在用户界面上实现过程自动化的强大工具。脚本可以执行用户使用鼠标或键盘执行的同一类操作。还可以指定输出并处理生成的模型。

第 3 章 建模简介

模型是一组规则、公式或方程式，可以使用它们来根据一组输入字段或变量预测输出。例如，金融机构可以使用模型来根据以往的申请人的已知相关信息预测贷款申请人具有较低风险还是较高风险。

能够预测结果是预测性分析的中心目标，并且了解建模过程是使用 IBM SPSS Modeler 的关键。

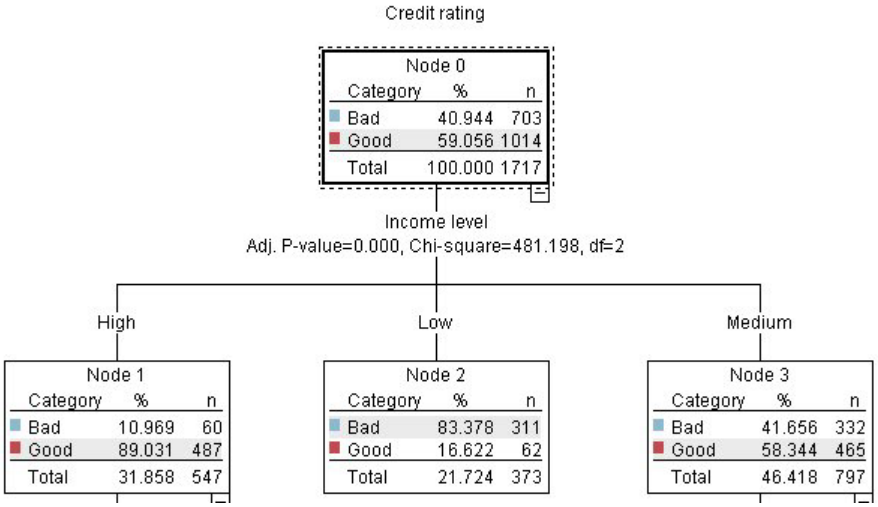


图 11. 简单的决策树模型

本示例使用**决策树**模型，该模型使用一系列决策规则对记录进行分类（并预测响应），例如:

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

本示例使用 CHAID（卡方自动交互效应检测）模型时，旨在进行常规的介绍，大部分概念会广泛应用于 IBM SPSS Modeler 中的其他建模类型。

无论要了解哪种模型，均需要首先了解进入该模型的数据。此示例中的数据包含有关银行客户的信息。其中使用了下列字段:

字段名称	描述
Credit_rating	信用评级: 0 = 不良，1 = 优良，9 = 缺失值
年龄	Age in years
收入	收入水平: 1 = 低，2 = 中，3 = 高
Credit_cards	持有的信用卡数: 1 = 少于五张，2 = 五张或更多
教育	教育程度: 1 = 高中，2 = 大学
Car_loans	申请的汽车贷款数: 1 = 没有或者一项，2 = 两项以上

对于已申请银行贷款的客户，银行维护其相关历史信息的数据库，这些信息包括客户是偿还了贷款（信用评级 = 优良）还是拖欠贷款（信用评级 = 不良）。通过使用此现有数据，银行将构建一个模型，该模型使他们能够预测未来的贷款申请者拖欠贷款的可能性。

通过使用决策树模型，您可以分析两组客户的特征并预测贷款拖欠的发生可能性。

本示例使用了名为 *modelingintro.str* 的流，该流位于 *streams* 子文件夹下的 *Demos* 文件夹中。数据文件是 *tree_credit.sav*。请参阅主题第 4 页的『Demos 文件夹』以获取更多信息。

我们来看一下流。

1. 从主菜单中选择下列选项:

   文件 > 打开流

2. 单击“打开”对话框的工具栏上的金块图标，然后选择 Demos 文件夹。
3. 双击 *streams* 文件夹。
4. 双击名为 *modelingintro.str* 的文件。

构建流

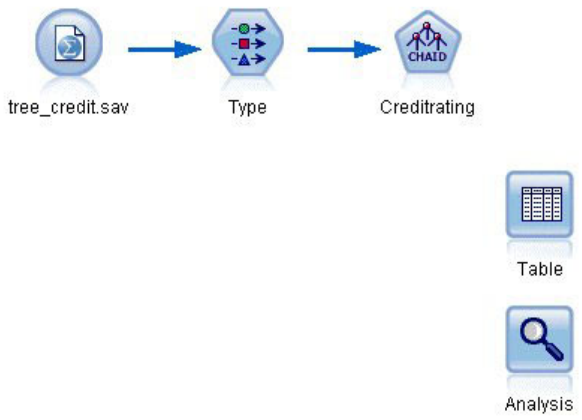


图 12. 建模流

要构建将创建模型的流，至少需要 3 个元素:

- 一个从某些外部源读取数据的源节点，在本示例中为 IBM SPSS Statistics 数据文件。
- 一个指定字段属性的源节点或“类型”节点，字段属性包括测量级别（字段包含的数据类型）以及每个字段在建模过程中的角色是目标还是输入等。
- 一个在运行流时生成模型块的建模节点。

在此示例中，将使用 CHAID 建模节点。CHAID（即，卡方自动交互检测）是一种分类方法，此方法通过使用称为卡方统计的特定类型统计信息确定决策树中的最佳分割位置来构建决策树。

如果在源节点中指定了测量级别，那么可以除去单独的“类型”节点。从功能上来说，结果是一样的。

此流还包含“表”节点和“分析”节点，创建模型块并将其添加到此流中之后将使用这两个节点查看评分结果。

Statistics 文件源节点从 *tree_credit.sav* 数据文件读取 IBM SPSS Statistics 格式数据，该文件安装在 *Demos* 文件夹中。（名为 *$CLEO_DEMOS* 的特殊变量用于引用位于当前 IBM SPSS Modeler 安装下的该文件。这样，无论当前的安装文件夹或版本是什么，均可以确保路径有效。）

图 13. 使用“Statistics 文件”源节点读取数据

类型节点指定每个字段的**测量级别**。测量级别是指示字段中数据的类型的类别。我们的源数据文件使用三种不同的测量级别。

连续字段（例如*年龄*字段）包含连续的数字值，而**名义**字段（例如*信用评价*字段）有两个或多个不同值，例如*不良*、*优良*或*无信用历史记录*。**有序**字段（例如*收入水平*字段）用于描述包含具有固有顺序的多个不同值的数据，在此个案中为*低*、*中*和*高*。

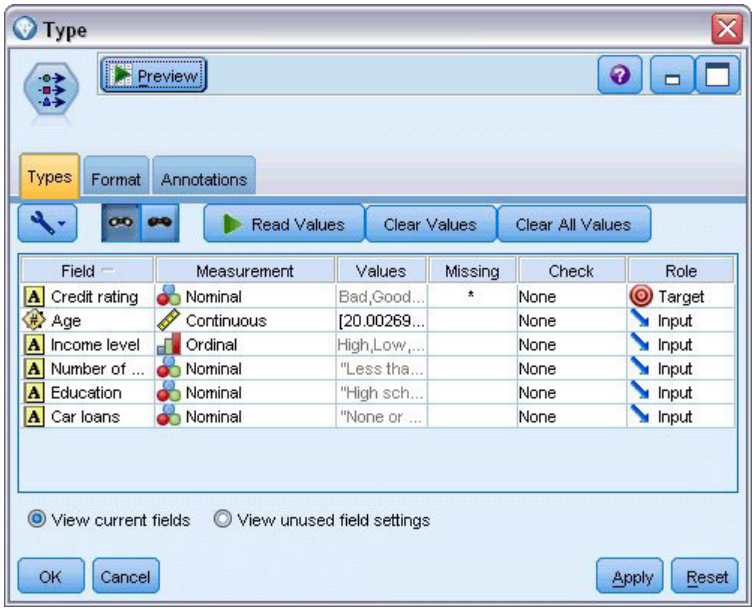


图 14. 使用“类型”节点设置目标和输入字段

对于每个字段，类型节点还指定**角色**，以指示每个字段在建模中扮演的部分。将字段*信用评价*的角色设置为*目标*，此字段指示指定的客户是否拖欠贷款。这是**目标**，或者是要预测其值的字段。

对于其他字段，将角色设置为*输入*。输入字段有时也称为**预测变量**，或建模算法用其值来预测目标字段值的字段。

CHAID 建模节点将生成模型。

在建模节点的“字段”选项卡中，已选中**使用预定义角色**，这意味着将按在类型节点中的指定使用目标和输入。此时，可以更改字段角色，但就此示例而言，将按原样使用这些字段角色。

1. 单击“构建选项”选项卡。

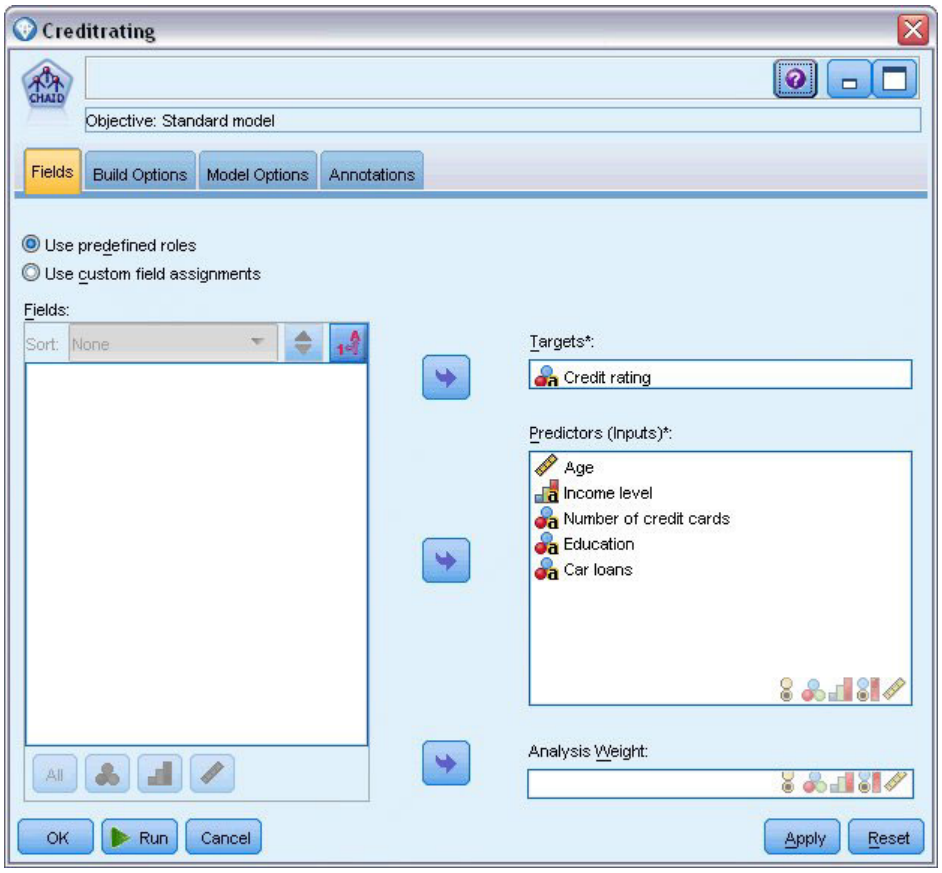


图 15. CHAID 建模节点，“字段”选项卡

下面是一些选项，可以在这些选项中指定要构建的模型种类。

由于我们想要一个全新的模型，因此使用缺省选项**构建新模型**。

我们还要求它为单个标准决策树模型，并且不包含任何增强，因此保留缺省目标选项**构建单个树**。

我们可以选择启动允许对模型进行微调的交互建模会话，本示例只使用缺省设置**生成模型**来生成模型。

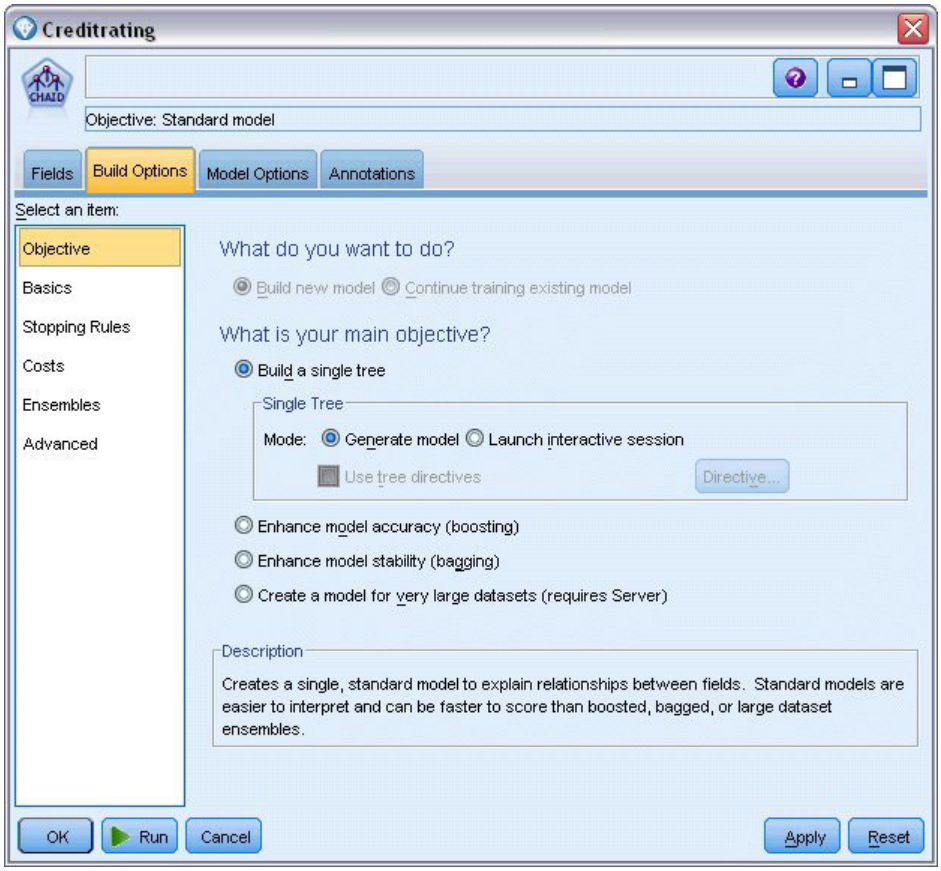


图 16. CHAID 建模节点，“构建选项”选项卡

对于此示例，我们希望保持树相当简单，因此，将通过增加父节点和子节点个案的最小数来限制树增长。

2. 在“构建选项”选项卡上，从左侧的导航器窗格选择**停止规则**。
3. 选择**使用绝对值**选项。
4. 将**父分支中的最小记录数**设置为 400。
5. 将**子分支中的最小记录数**设置为 200。

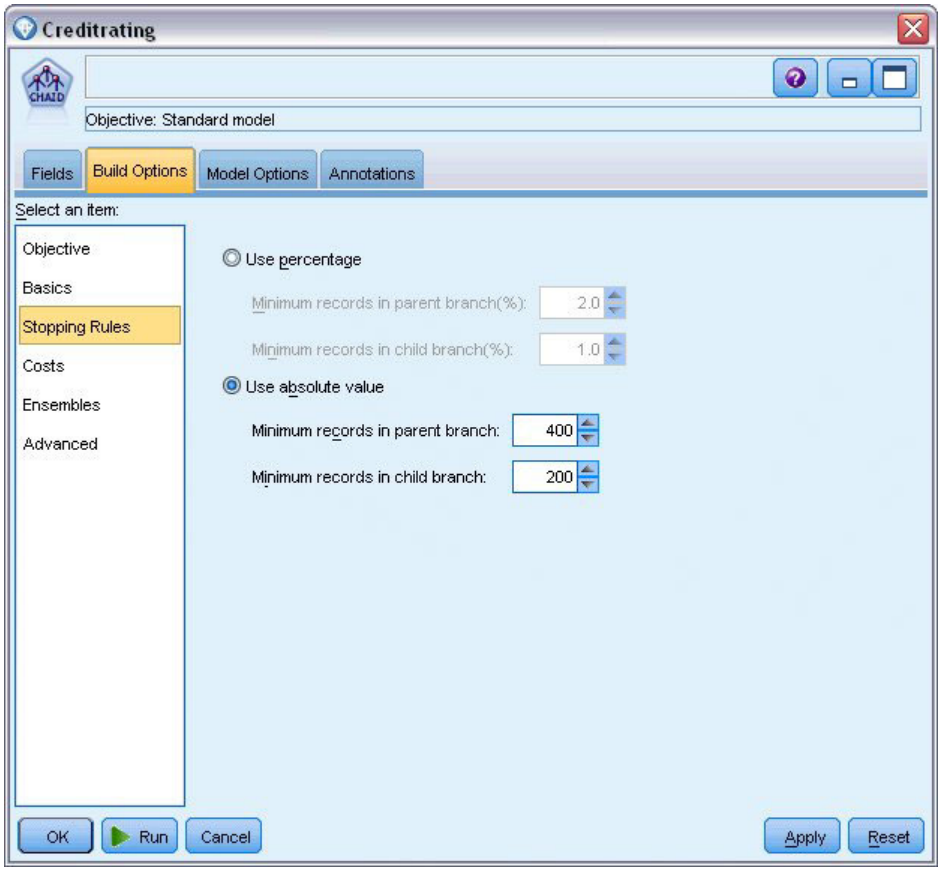


图 17. 设置用于决策树构建的中止条件

在本例中，我们可以使用所有其他缺省选项，因此单击**运行**以创建模型。（另外，也可以右键单击该节点，然后从上下文菜单中选择**运行**，或选择节点，并从“工具”菜单中选择**运行**。）

浏览模型

执行完成后，模型块将添加到应用程序窗口右上角的“模型”选用板中，并且还将放在流画布中，同时提供一个指向从中创建该模型块的建模节点的链接。要查看模型的详细信息，右键单击模型块并选择**浏览**（在模型选用板上）或**编辑**（在工作区上）。

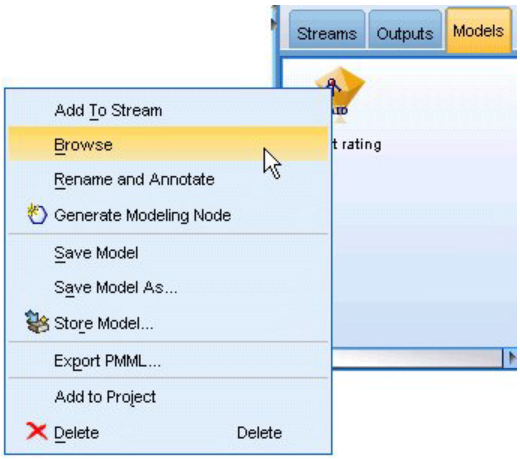


图 18. “模型”选用板

对于 CHAID 模型块，“模型”选项卡以规则集的形式显示详细信息，规则集实际上是可用于根据不同输入字段的值将各个记录分配给子节点的一系列规则。

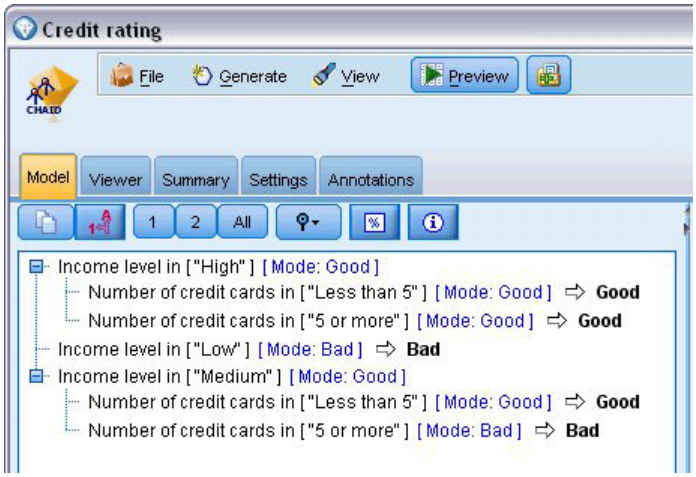


图 19. CHAID 模型块，规则集

对于每个决策树终端节点--意味着那些树节点没有进一步拆分--返回*优良*或*不良*的预测值。对于落在该节点内的记录，所有个案中的预测均由**模式**或最常见的响应决定。

在规则集的右侧，“模型”选项卡显示了预测变量重要性图表，该图表显示评估模型时每个预测变量的相对重要性。通过这一点，我们看到*收入水平*在此个案中最显著，而其他唯一显著的因子是*信用卡数量*。

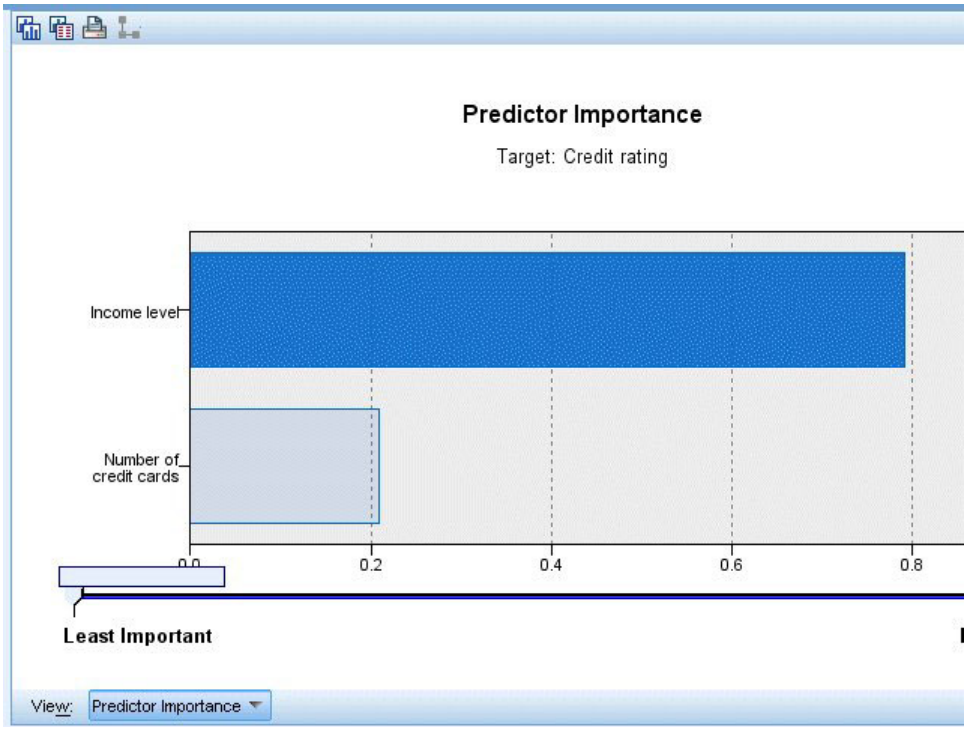


图 20. 预测变量重要性图表

模型块中的“查看器”选项卡以树的形式显示同一模型，其中每个决策点都包含一个节点。使用工具栏上的“缩放”控件对特定节点进行放大和缩小可以查看树的更多内容。

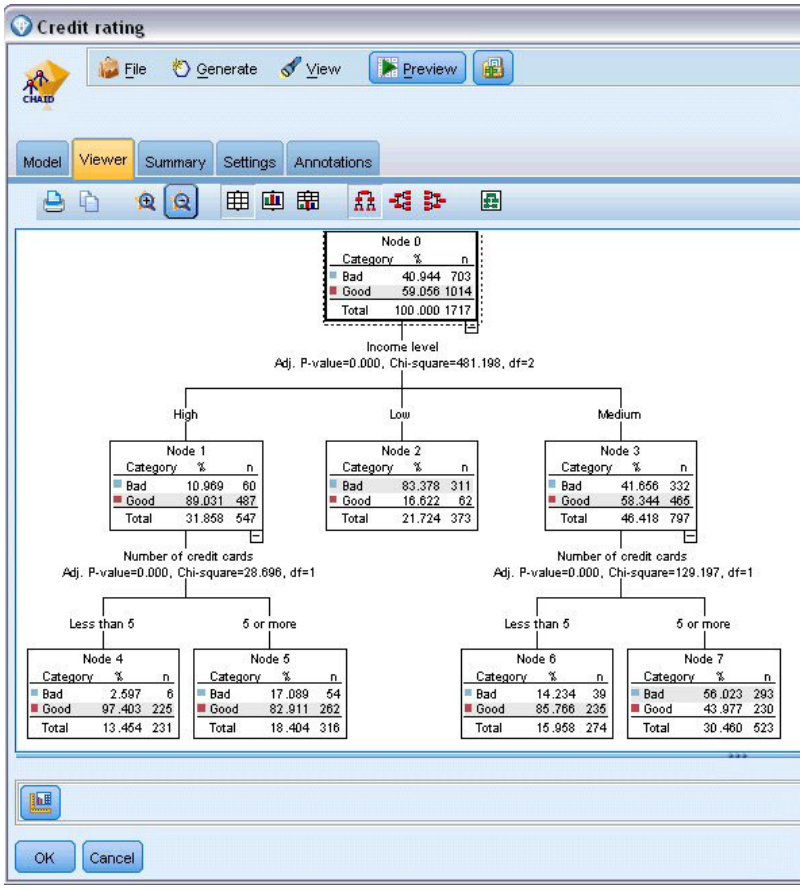


图 21. 模型块中的查看器选项卡，已选择缩小

查看树的上部，第一个节点（节点 0）为我们提供数据集中所有记录的摘要。数据集中超过 40% 的个案分类为风险较高。这是一个相当高的比例，因此我们需要了解此树是否可以提供关于哪些因素可能会导致出现此情况的任何线索。

我们可以看到第一个分割是根据*收入水平*。收入水平处于*低*类别的记录将分配给节点 2，所以此类别中的贷款拖欠者百分比最高不足为奇。很明显，向此类别中的客户提供贷款具有高风险。

但是，此类别中 16% 的客户实际上*未*拖欠贷款，因此预测并非始终准确。没有模型能够预测每一个响应，但好的模型能够根据可用数据预测对每一个记录作出的*最常见*的响应。

同样，如果我们查看高收入客户（节点 1），那么可以看到绝大部分 (89%) 风险较低。但是其中超过 10% 的客户也会拖欠贷款。是否可以优化我们的贷款标准来最大程度地降低此处的风险？

注意模型如何根据持有的信用卡数量将这些客户分成两个子类别（节点 4 和节点 5）。对于高收入客户，如果我们只向那些信用卡少于 5 张的客户贷款，则可以将我们的成功率从 89% 提高到 97%--甚至更满意的结果。

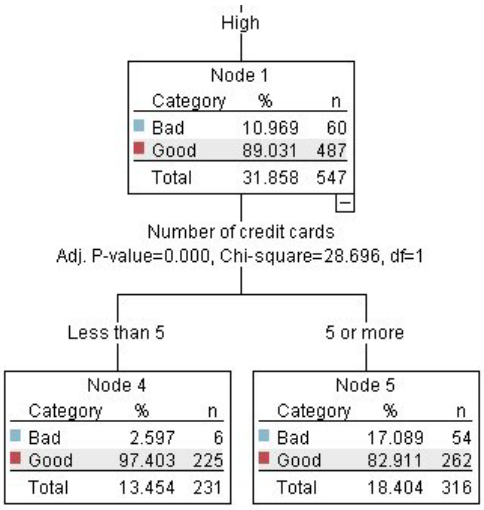


图 22. 高收入客户的树形视图

但中等收入类别（节点 3）中的那些客户是什么情况？他们更平均地划分为“优良”和“不良”评级。

子类别（此情况中是节点 6 和 7）仍然能帮助我们。这次，只向那些信用卡少于 5 张的中等收入客户贷款，可将优良评价的百分比从 58% 提高到 85%，这是显著的改进。

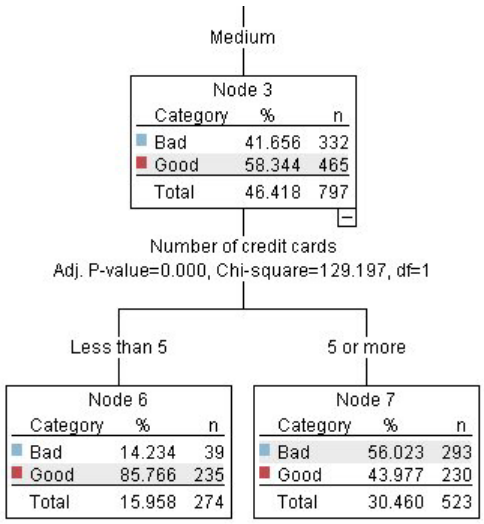


图 23. 中等收入客户的树形视图

因此，我们了解到对于输入此模型的每条记录，将向其分配一个特定节点，并且根据该节点最常见的响应分配预测值*优良*或*不良*。

为各个记录分配预测值的这一过程称为**评分**。通过对用于估算该模型的相同记录进行评分，我们可以评估该模型执行训练数据（已知道其结果的数据）的准确度。让我们来看看如何执行此操作。

评估模型

我们已通过浏览模型了解了评分方式。但是，如果要评估模型的*准确度*，那么需要对一些记录进行评分，并将模型预测的响应与实际结果进行比较。我们将对用于估算模型的同一记录进行评分，从而对观察到的响应与预测响应进行比较。

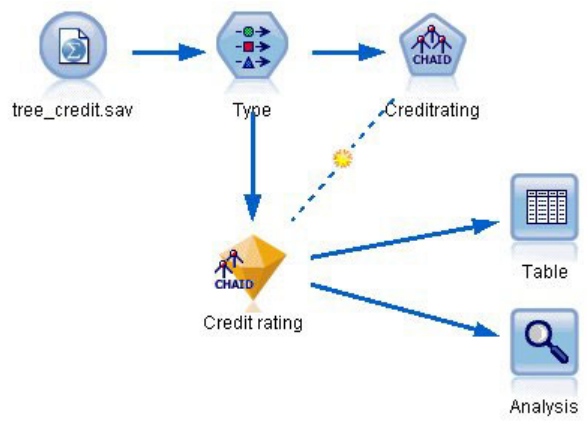
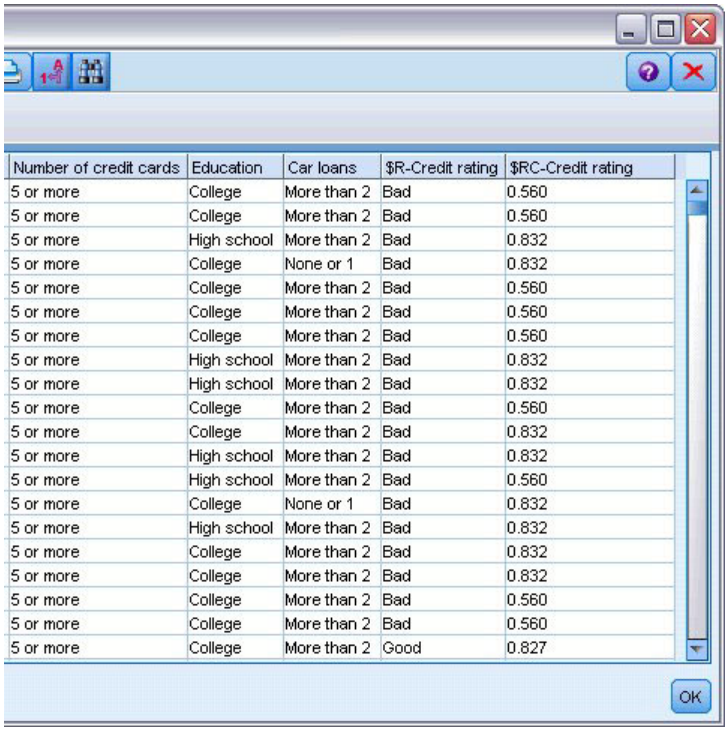


图 24. 将模型块附加到输出节点以进行模型评估

1. 要查看分数或预测值，请将表节点添加到模型块，然后双击“表”节点，并单击**运行**。

   表在名为 *$R-Credit rating* 的字段中显示预测分数，该字段由模型创建。我们可以将这些值与包含实际响应的原始*信用评价*字段进行比较。

   按照惯例，在评分过程中生成的字段的名称基于目标字段，但是要加上标准前缀。前缀 *$G* 和 *$GE* 由广义线性模型生成，*$R* 是用于本例中的 CHAID 模型所生成的预测的前缀，*$RC* 用于置信度值，*$X* 通常是使用整体生成的，而 *$XR*、*$XS* 和 *$XF* 在目标字段分别为“连续”、“分类”、“集合”或“标志”字段的情况下用作前缀。不同的模型类型使用不同的前缀集。**置信度值**是模型自身对每个预测值的准确度的估计，范围为 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

图 25. 显示已生成的评分和置信度值的表

与预期的一样，预测值与大多数（并非全部）记录的实际响应相匹配。出现此情况的原因是每个 CHAID 终端节点都具有混合响应。预期值与 *最常见* 的响应相匹配，但对于该节点中的其他响应，该预期值是错误的。（记住，16% 的少部分低收入客户没有拖欠。）

为了避免出现这种情况，可以继续将树拆分为越来越小的分支，直到每个节点都只包含*优良*或*不良*响应为止。但是，这样的模型可能会非常复杂，并且不易推广到其他数据集。

要查看具体有多少预测值正确，我们可通读表格，并计算预测字段 *$R-Credit rating* 的值匹配*信用评价*的值的记录数量。幸运的是，有更简单的方法 - 我们可以使用自动执行此操作的“分析”节点。

2. 将模型块连接到“分析”节点。
3. 双击“分析”节点，然后单击**运行**。

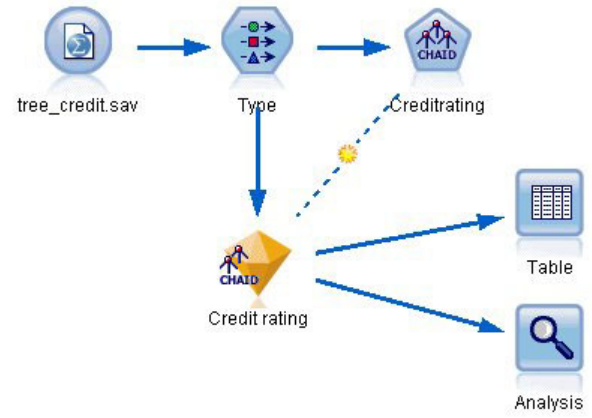


图 26. 附加“分析”节点

分析表明，对于 2464 条记录中的 1899 条记录（超过 77%），模型预测的值与实际响应相匹配。

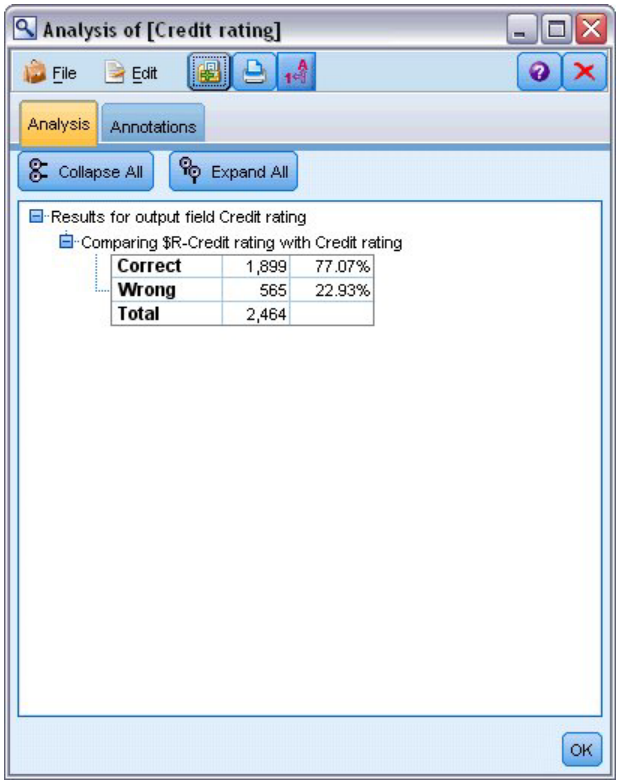


图 27. 观察到的响应与预测响应的比较分析结果

此结果受到评分的记录和用于评估模型的记录相同的事实的限制。在真实情况中，可使用分区节点将数据拆分为培训和评估的单独示例。

通过使用一个样本分区生成模型并使用另一个样本对模型进行检验，您会得到该模型推广到其他数据集的情况。

通过“分析”节点，我们可以根据已知道实际结果的记录来检验模型。下一阶段介绍如何使用模型对我们不知道结果的记录进行评分。例如，这可能包括当前不是银行客户的人员，但他们是促销邮寄的潜在目标。

对记录评分

先前我们对用于估算模型的相同记录进行了评分，以评价模型的准确度。现在，我们要了解如何对与用于创建模型的记录不同的记录集进行评分。以下是使用目标字段进行建模的目标：研究已知道结果的记录以确定使您可以预测未知结果的模式。

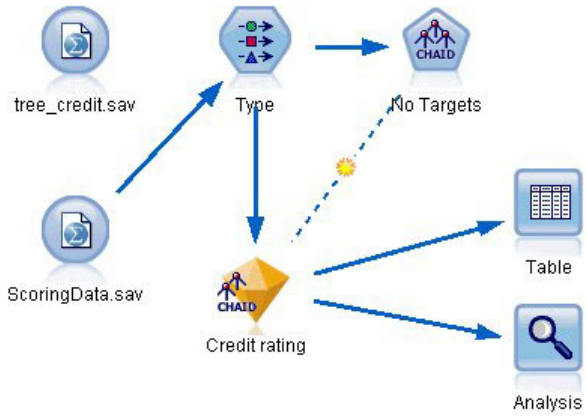


图 28. 附加用于评分的新数据

您可以更新“Statistics 文件”源节点，使它指向其他数据文件，也可以添加一个从中读取要进行评分的数据的新源节点。无论采用哪种方式，新数据集包含的输入字段必须与模型（*年龄*、*收入水平*、*教育*等）所使用的相同，但不包含目标字段*信用评价*。

另外，您也可以将模型块添加到包含预期输入字段的任何流中。无论是读取文件还是数据库，只要字段名和类型与模型使用的相匹配，源类型都无关紧要。

也可以将模型块保存为单独的文件、将模型导出为 PMML 格式以用于其他支持此格式的应用程序，或将模型存储到 IBM SPSS Collaboration and Deployment Services 存储库中，这样可以在企业范围对模型进行部署、评分和管理。

无论使用何种基础结构，模型自身都按同一方式工作。

摘要

本示例演示了创建模型、评估模型以及对模型评分的基本步骤。

- 建模节点通过研究已知道结果的记录来估算模型并创建模型块。这有时称为训练模型。
- 可将模型块添加到包含预期字段的任何流中，以对记录进行评分。通过对已知道结果的记录（例如现有客户）进行评分，您可以评估模型的运行情况。
- 如果您对模型的运行情况感到满意，则可以对新数据（如准客户）进行评分，以预测他们的响应。
- 用于训练或估算模型的数据可以称为分析数据或历史数据；评分数据也可以称为操作数据。

第 4 章 标志目标的自动建模

对客户响应建模（自动分类器）

通过“自动分类器”节点，您可以为标志（例如某个指定客户是否可能拖欠贷款或者是否对特定的报价做出响应）或名义（集合）目标自动创建和比较多个不同的模型。在本例中，我们将搜索标志（是或否）结果。在一个相对简单的流中，节点生成一组候选模型并对它们进行排序，选择最有效的模型，然后将它们合并为一个汇总（整体）模型。此方法将自动化操作的方便性与组合多个模型的优势融为一体，从而产生任何单一模型所不能带来的更为准确的预测。

本示例基于某个虚构的公司，该公司希望通过为每个客户提供适合的报价以实现更高收益。

此方法突出了自动操作的优势。有关使用连续（数值范围）目标的类似示例，请参阅。

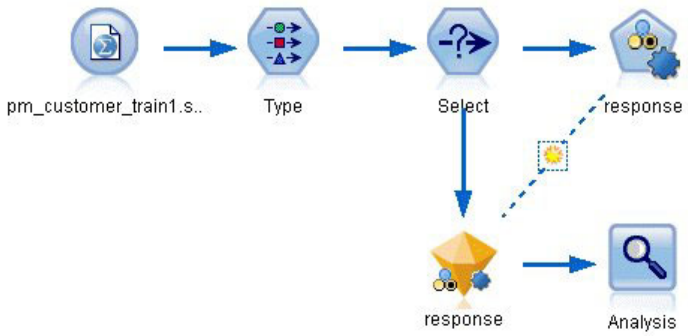


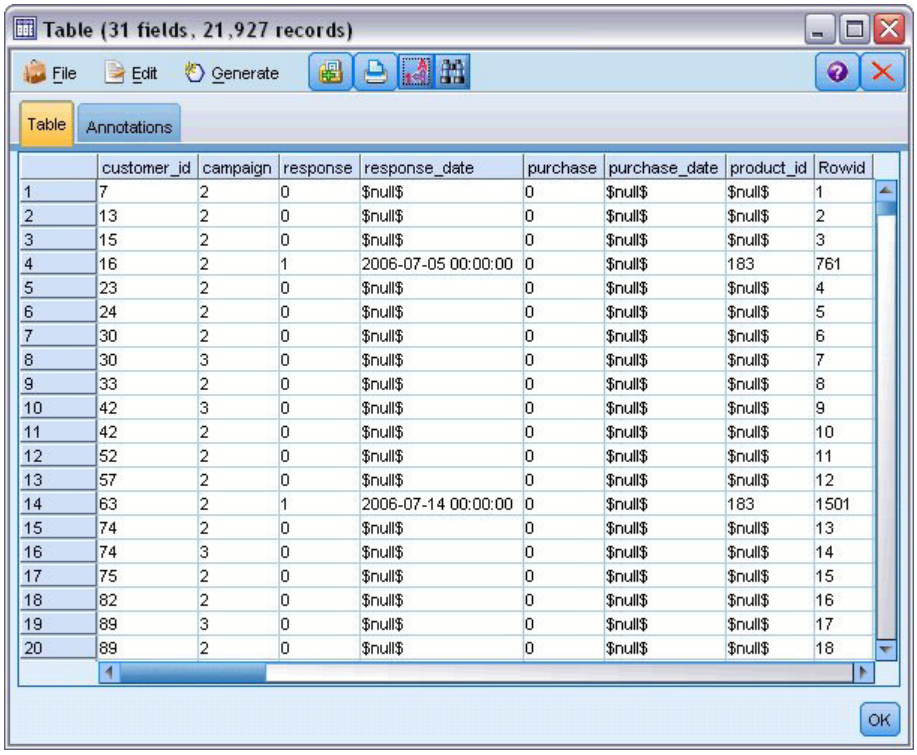
图 29. 自动分类器样本流

本示例使用安装在 *streams* 目录下 Demo 文件夹中的流 *pm_binaryclassifier.str*。所使用的数据文件为 *pm_customer_train1.sav* 。请参阅主题『历史数据』以获取更多信息。

历史数据

文件 *pm_customer_train1.sav* 的历史数据可跟踪过去的营销活动中为特定客户提供的报价，由 *campaign* 字段的值表示。*Premium account* 活动中的记录数最大。

campaign 字段的值在数据中实际编码为整数（例如 *2 = Premium account*）。稍后，您可为这些值定义标签以用于给出更有意义的输出。



	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

图 30. 先前促销活动的相关数据

此文件还包含一个 *响应* 字段，该字段表明所提供的报价是否被接受（0 = *否*，1 = *是*）。这将是您希望预测的 **目标字段** 或值。此外，还包括一些字段，这些字段包含有关每位客户的人口统计信息和财务信息。这些字段可用于构建或“训练”根据收入、年龄或每月交易次数等特征来预测个人或群体的响应率的模型。

构建流

1. 添加指向 *pm_customer_train1.sav* 的 Statistics 文件源节点，该文件位于 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中。（您可以在文件路径中指定 `$CLEO_DEMOS/` 作为引用此文件夹的快捷方式。请注意，路径中必须使用正斜杠而非反斜杠，如上文所示。）

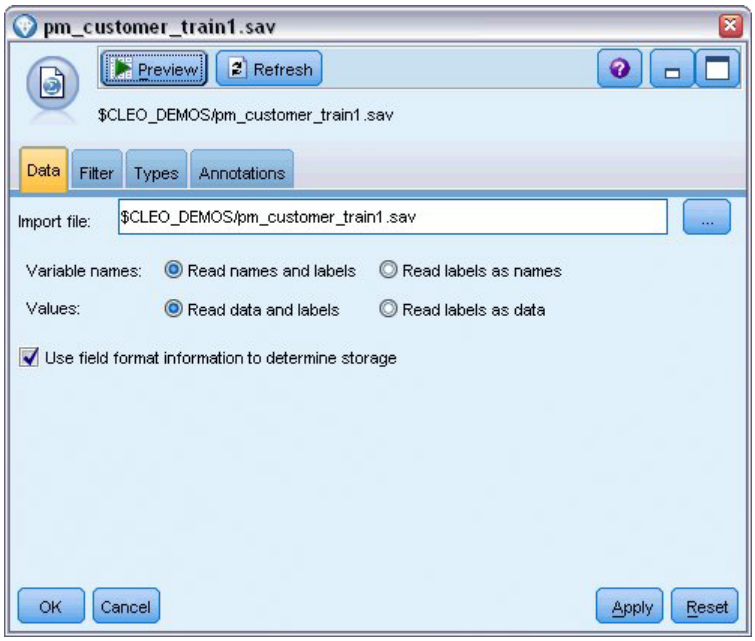


图 31. 读入数据

2. 添加类型节点，然后选择*响应*作为目标字段（“角色”为**目标**）。将此字段的“测量”设置为**标志**。

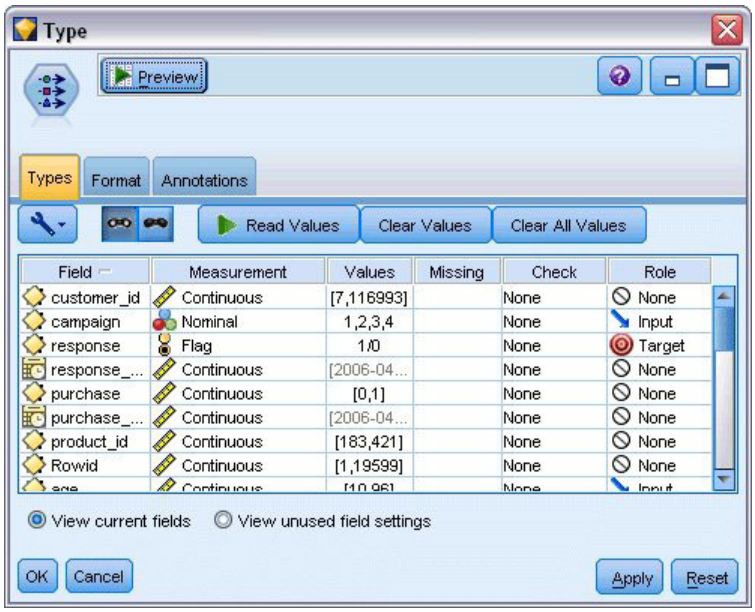


图 32. 设置测量级别和角色

3. 对于下列字段，将角色设置为**无**：*customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid* 和 *X_random*。当您构建模型时，将忽略这些字段。
4. 单击类型节点的 **读取值** 按钮以确保值获得实例化。

正如前述内容所示，我们的源数据包含有关四项不同活动的信息，每个活动针对不同类型的客户帐户。这些活动在数据中编码为整数，以方便记住每个整数所代表的帐户类型，让我们为每一个都定义标签。

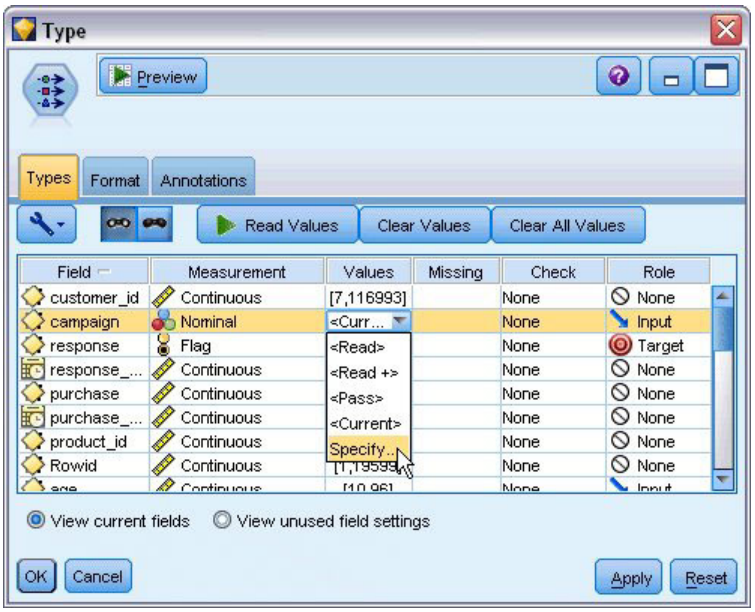


图 33. 选择以指定字段值

5. 在**活动**字段的行上，单击**值**列中的条目。
6. 从下拉列表选择**指定**。

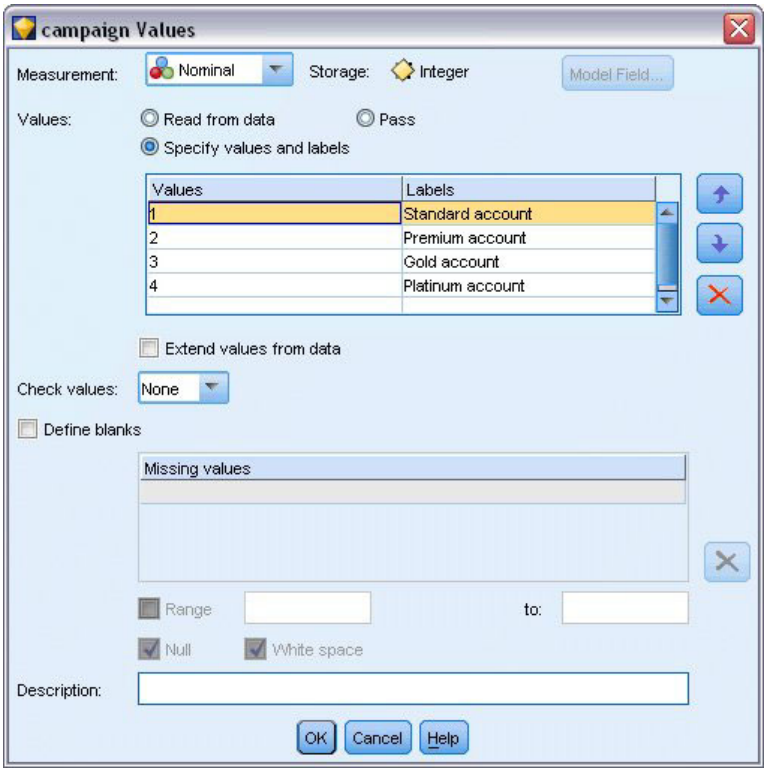
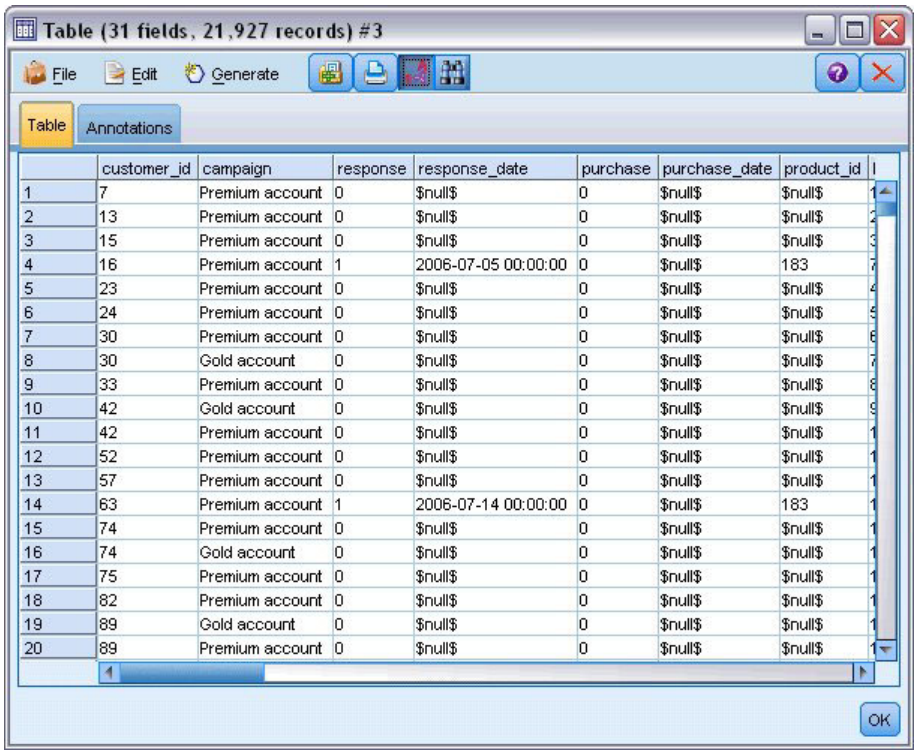


图 34. 定义字段值的标签

7. 在**标签**列中，键入**活动**字段四个值中每个值所显示的标签。
8. 单击**确定**。

现在您可在输出窗口中显示标签而非整数了。



	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

图 35. 显示字段值标签

9. 将表节点附加到类型节点。
10. 打开“表”节点，然后单击**运行**。
11. 在输出窗口上，单击**显示字段和值标签**工具栏按钮以显示标签。
12. 单击**确定**关闭输出窗口。

尽管数据包含有关四项不同活动的信息，但每一次的分析应侧重于其中一项活动。由于 Premium account 活动（在数据中编码为 *campaign=2*）中的记录数最大，因此可以使用选择节点实现仅在流中包含这些记录。

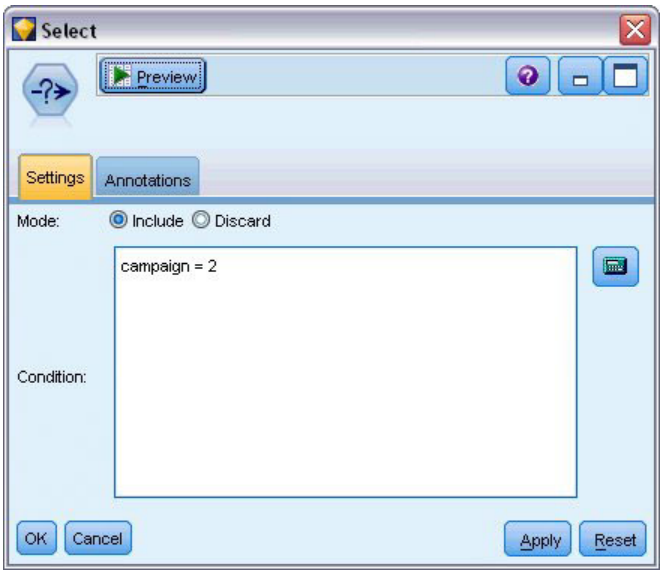


图 36. 为单项活动选择记录

生成和比较模型

1. 附加一个自动分类器节点，然后选择**总体准确性**作为对模型进行排序的度量。
2. 将**要使用的模型数**设置为 3。这意味着在执行节点时将构建三个最佳模型。

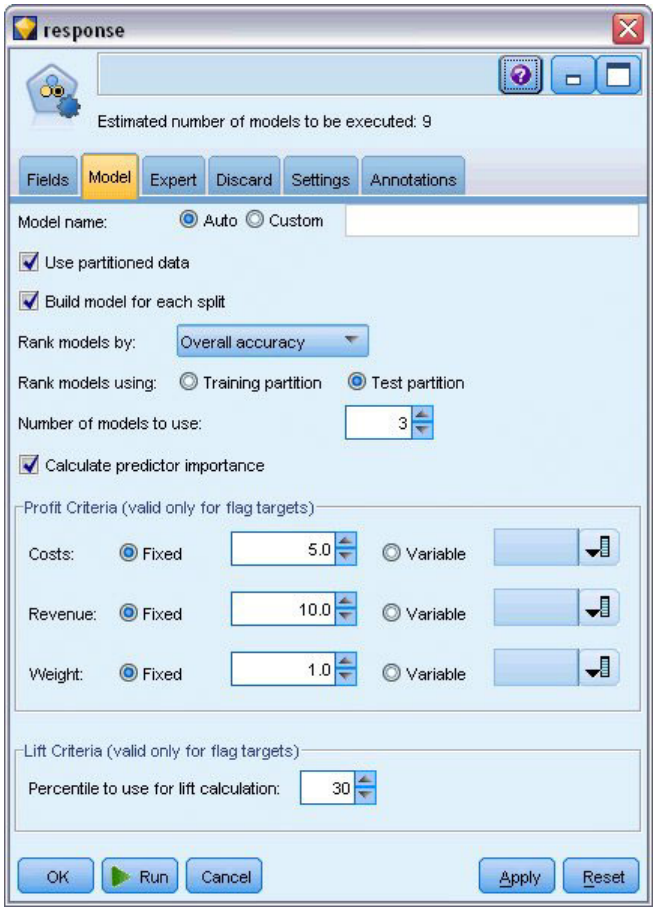


图 37. "自动分类器"节点的"模型"选项卡

在"专家"选项卡上，可从最多 11 种不同模型算法中进行选择。

3. 取消选择**判别**和 **SVM** 模型类型。（这些模型需要花费更多时间来训练这些数据，因此取消选中它们将加快示例的执行速度。如果您不介意稍等一下，也可以保留它们的选中状态。）

   由于在"模型"选项卡上将**要使用的模型数**设置为 3，因此节点将计算余下九个算法的准确性，并构建包含三个最准确算法的单个模型块。

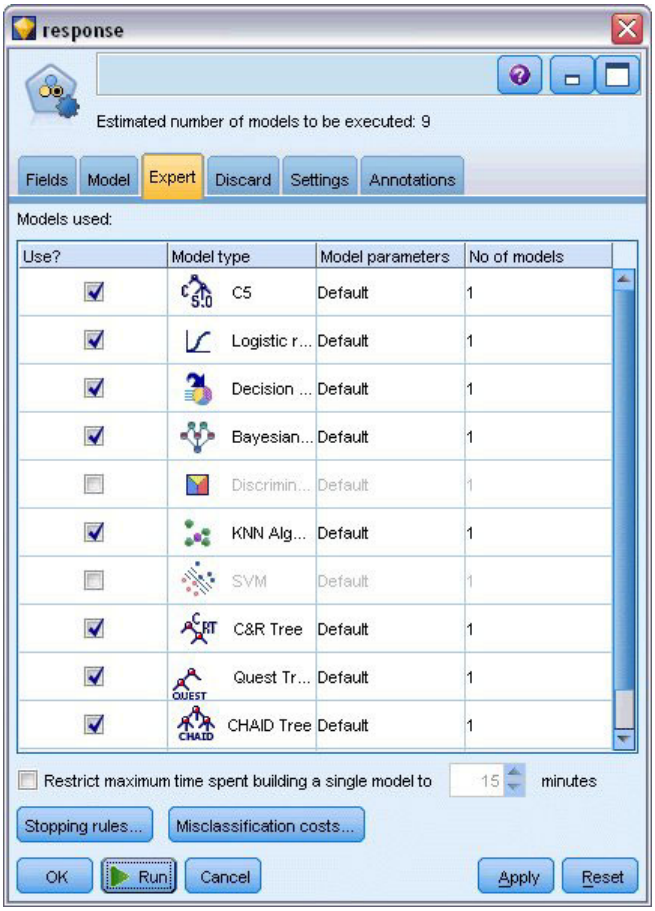


图 38. “自动分类器”节点的“专家”选项卡

4. 在“设置”选项卡上，对于整体方法，请选择**置信度加权投票**。此选项确定如何为每条记录生成一个汇总评分。

   使用简单投票方式时，若三个模型中有两个模型均预测 *是* ，则 *是* 将以 2 比 1 的投票结果取胜。在使用置信度加权投票方式的情况下，将基于各预测的置信度值进行加权投票。因此，如果一个预测 *否* 的模型的置信度比两个预测 *是* 的模型合在一起的置信度还高，则 *否* 取胜。

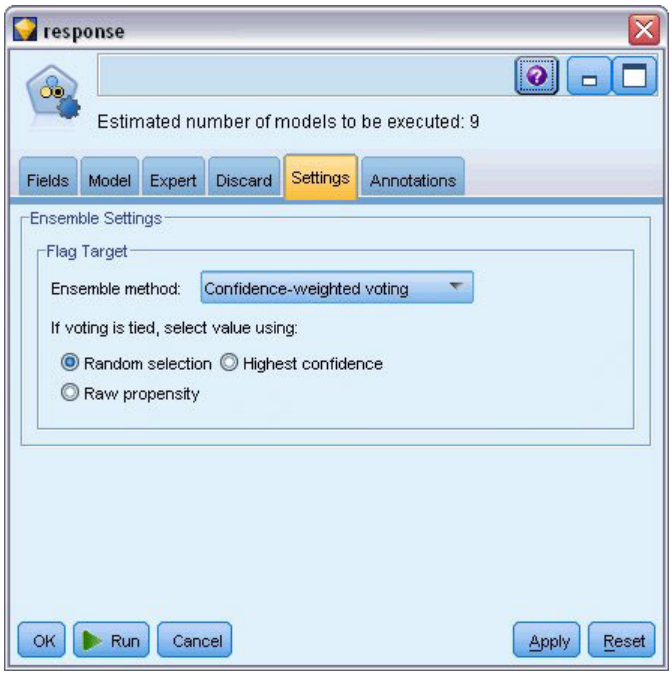


图 39. “自动分类器”节点：“设置”选项卡

5. 单击**运行**。

几分钟后，将构建生成的模型块，并将其放入画布和窗口右上角的“模型”选用板中。您可以浏览此模型块，或者以多种其他方式对其进行保存或部署。

打开模型块；它将列出在运行期间创建的每个模型的详细信息。（对于可能会在大型数据集中创建数百个模型的实际情况，这可能会花费数小时的时间。）请参阅第 33 页的图 29。

如果需要进一步探索任何单独的模型，可在**模型**列中双击此模型块图标，以向下浏览至单独模型结果，您可以从中生成建模节点、模型块或评估图表。在**图形**列中，可以双击缩略图生成标准大小的图形。

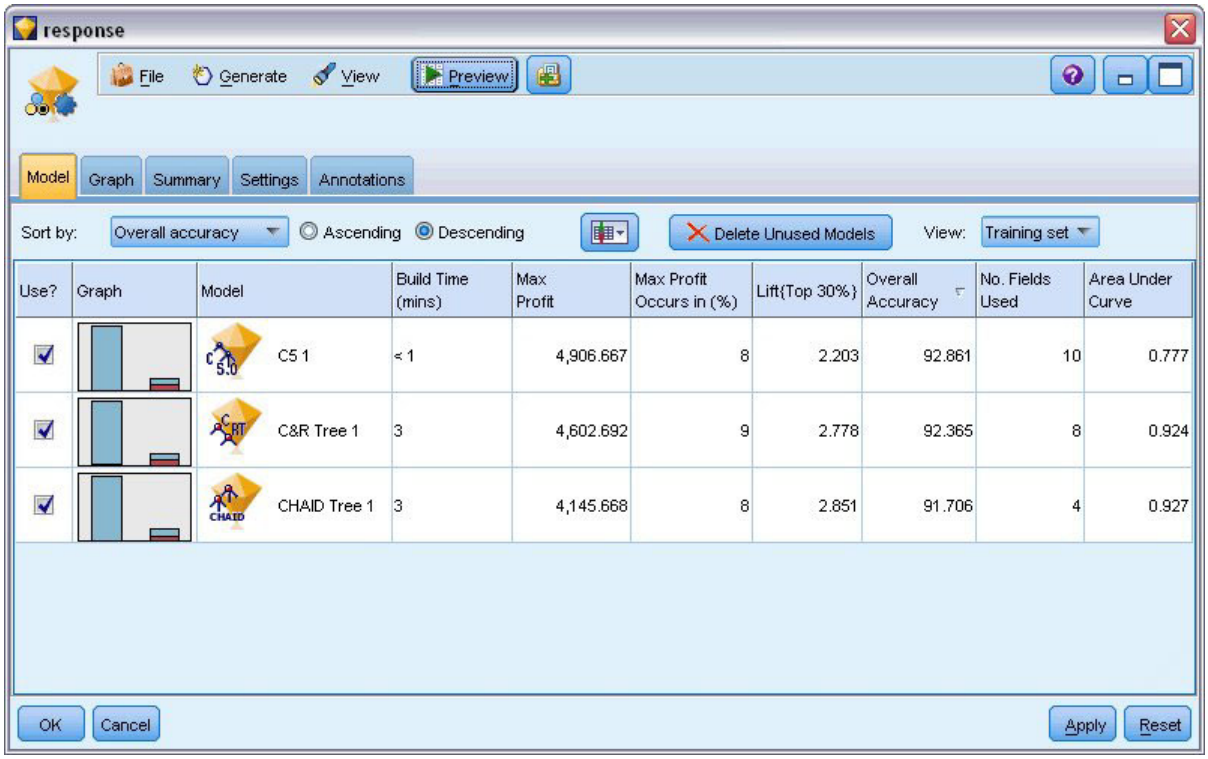


图 40. 自动分类器结果

缺省情况下，由于您在“自动分类器”节点的“模型”选项卡中选择了总体准确性度量，因此模型将根据此度量进行排序。根据这一度量，C51 模型的精确性最高，但 C&R 树和 CHAID 模型的精确性与之相差不大。

您可以通过单击其他列的标题对该列进行排序，或者也可以从工具栏的 **排序方式** 下拉列表中选择所需的度量。

基于这些结果，您可以决定使用所有这三个最准确的模型。通过结合多个模型的预测，可以避免单个模型的局限性，从而使总体准确性更高。

在**使用?** 列中，选择 C51，C&R 树和 CHAID 模型。

在模型块后面附加一个“分析”节点（“输出”选用板）。右键单击分析节点，然后选择**运行**以运行流。

由整体模型生成的汇总评分将显示在名为 *$XF-response* 的字段中。根据训练数据进行度量时，预测值与实际响应（如原始*响应*字段中的记录所示）相匹配的总体准确性为 92.82%。

尽管该准确性低于此个案的三个模型中的最高准确性（C51 为 92.86%），但它们之间的差异非常小，可以忽略不计。一般来说，在应用到除训练数据之外的数据集中时，整体模型通常更可能具有良好效果。

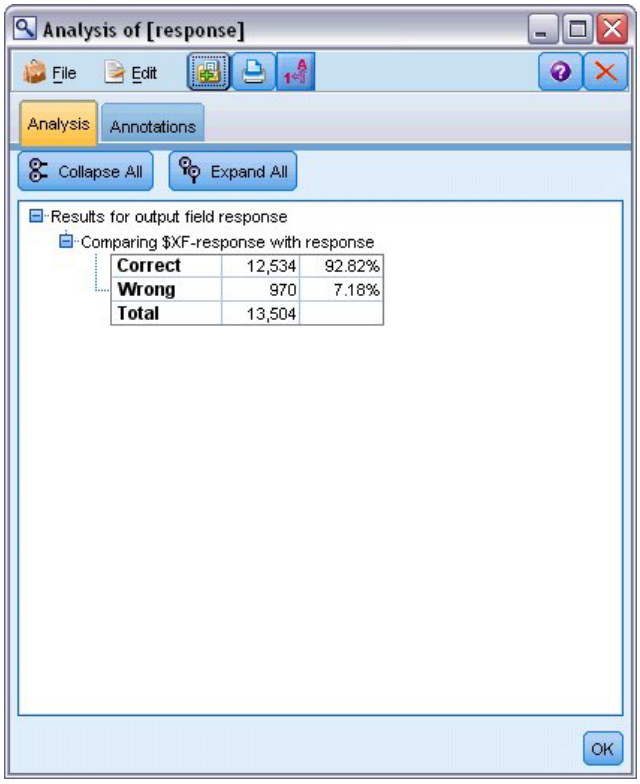


图 41. 对三个整体模型的分析

摘要

综上所述，您使用“字段分类器”节点对多种不同的模型进行了比较，然后使用三个最准确的模型并将它们添加到整体“自动分类器”模型块内的流中。

- 基于总体准确性，“C51”、“C&R 树”和 CHAID 模型对于训练数据效果最佳。
- 整体模型与最好的单个模型相比效果相差不大，而且在应用到其他数据集时可以起到更好的效果。如果您的目标是尽可能多地自动执行这一过程，您可以通过此方法获得在大多数情况下都很稳健的模型，而无需深入挖掘任意一个模型的细节。

第 5 章 连续目标的自动建模

属性值（自动数值）

通过“自动数值”节点，您可以为连续（数值范围）结果自动创建和比较不同的模型，例如预测某项财产的应征税值。借助于单独节点，可以估计和比较一组候选模型，并生成一个模型子集以进一步分析。这类节点与自动分类器节点工作方式相同，但连续不仅限于标志或名义目标。

该节点将候选模型中的最佳模型合并到单个汇总（整体）模型块中。此方法将自动化操作的方便性与组合多个模型的优势融为一体，从而产生任何单一模型所不能带来的更为准确的预测。

本示例主要讲述一个负责调整和评估房地产税的虚拟市政机构。为使预测更为准确，他们将构建一个根据建筑类型、周边状况、占地面积以及其他已知因素预测属性值的模型。

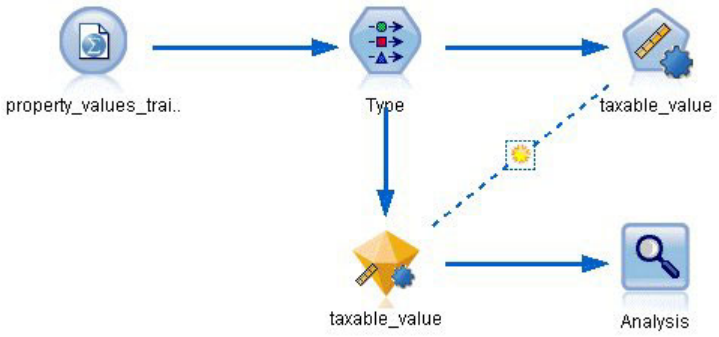


图 42. 自动数值样本流

此示例使用安装在 *streams* 下 Demos 文件夹中的流 *property_values_numericpredictor.str*。所使用的数据文件为 *property_values_train.sav*。请参阅主题第 4 页的『Demos 文件夹』以获取更多信息。

训练数据

数据文件包含一个名为 *taxable_value* 的字段，该字段就是要预测的 **目标字段** 或值。其他字段所包含的信息有周边情况、建筑类型以及内部体积，它们均可以用作预测变量。

字段名称	标签
property_id	属性标识
周边状况	城市内的区域
building_type	建筑物的类型
year_built	建造年代
volume_interior	内部体积
volume_other	车库和其他建筑所占的体积
lot_size	占地面积
taxable_value	应征税值

在 Demos 文件夹中还包括一个名为 *property_values_score.sav* 的评分数据文件。该文件中除了没有 *taxable_value* 字段之外，剩下的字段与数据文件相同。在训练模型使用已知应征税值的数据集之后，您就可以对仍不知晓应征税值的记录进行评分。

构建流

1. 添加指向 *property_values_train.sav* 的 Statistics 文件源节点，该文件位于 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中。（您可以在文件路径中指定 `$CLEO_DEMOS/` 作为引用此文件夹的快捷方式。请注意，如上所示，路径中必须使用正斜杠，而不是反斜杠。））

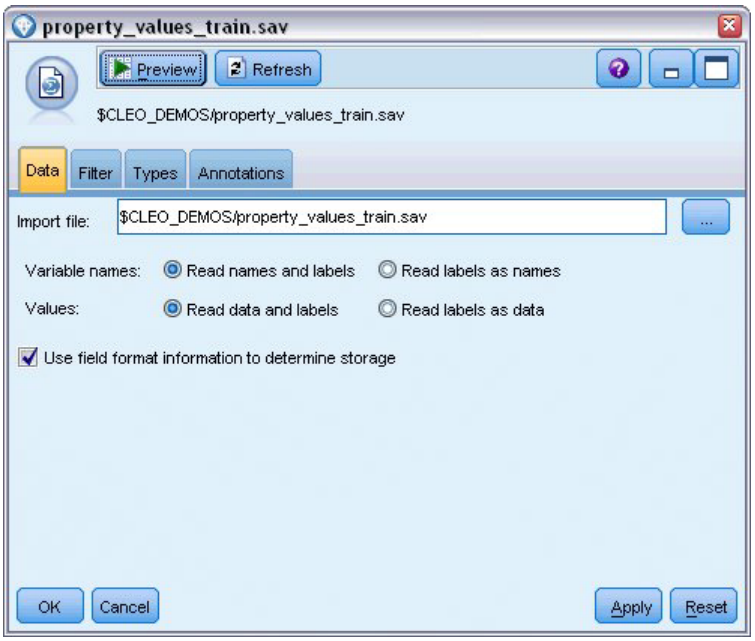


图 43. 读入数据

2. 添加类型节点，然后选择 *taxable_value* 作为目标字段（角色为**目标**）。所有其他字段的角色均应设置为**输入**，从而指示这些字段将用作预测变量。

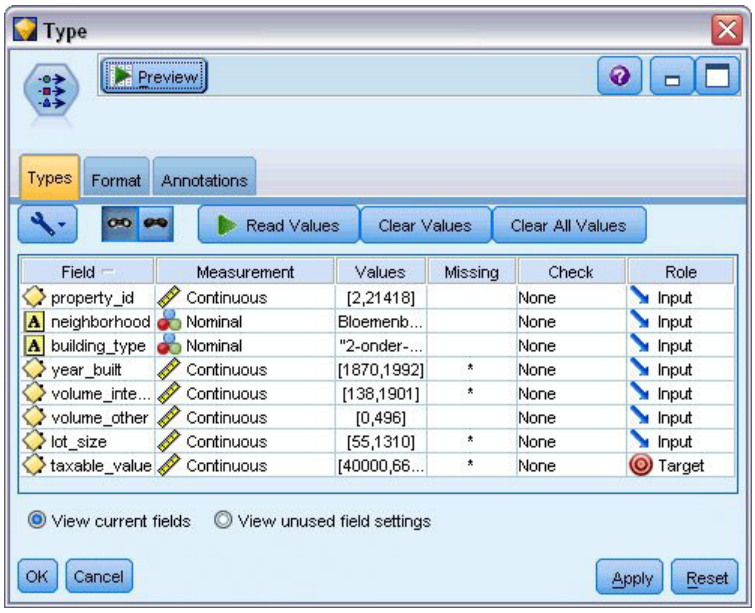


图 44. 设置目标字段

3. 附加自动数值节点，并选择**相关性**作为对模型排序的方法。
4. 将**要使用的模型数**设置为 3。这意味着在执行节点时将构建三个最佳模型。

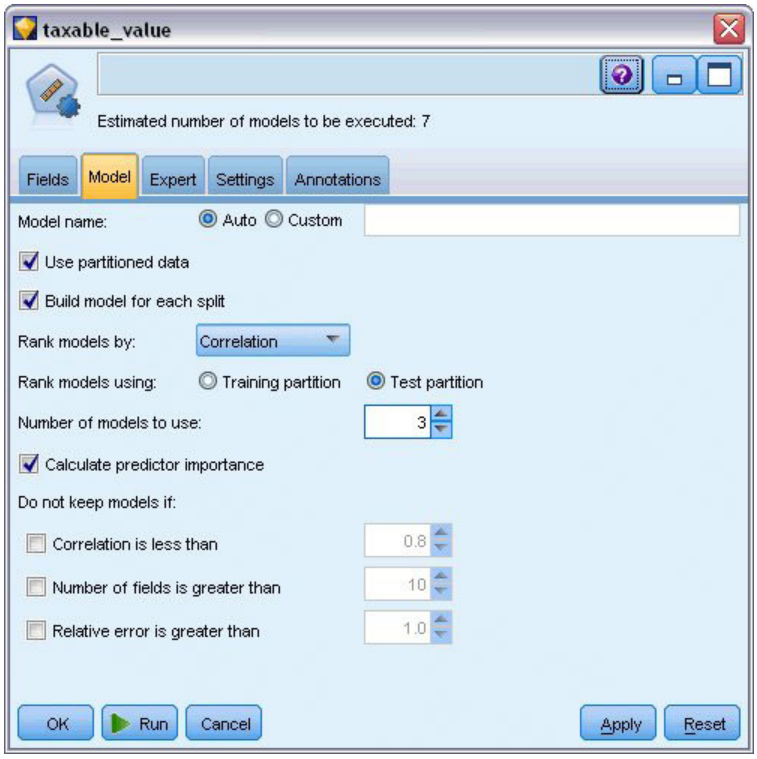


图 45. "自动数值"节点的"模型"选项卡

5. 在"专家"选项卡中，保留缺省设置；节点将为每个算法估算单个模型（共七个模型）。（或者，您可以修改这些设置，以对每个模型类型的多个变量进行比较。）

由于在“模型”选项卡上将**要使用的模型数**设置为 3，因此节点将计算七个算法的准确性，并构建包含三个最准确算法的单个模型块。

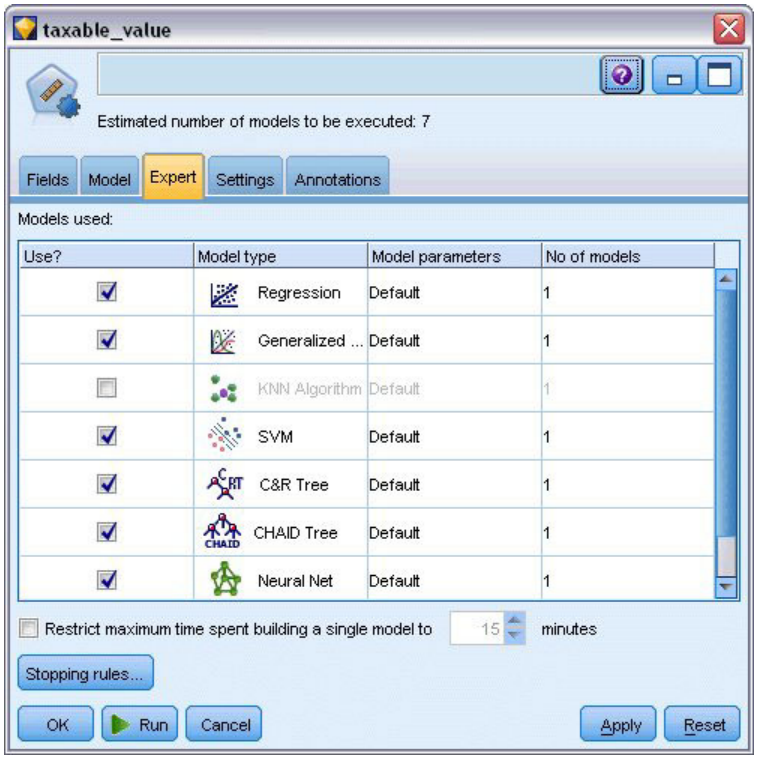


图 46. “自动数值”节点的“专家”选项卡

6. 在“设置”选项卡中，保留缺省设置。由于这是一个连续目标，因此会由各个模型的平均评分生成整体评分。

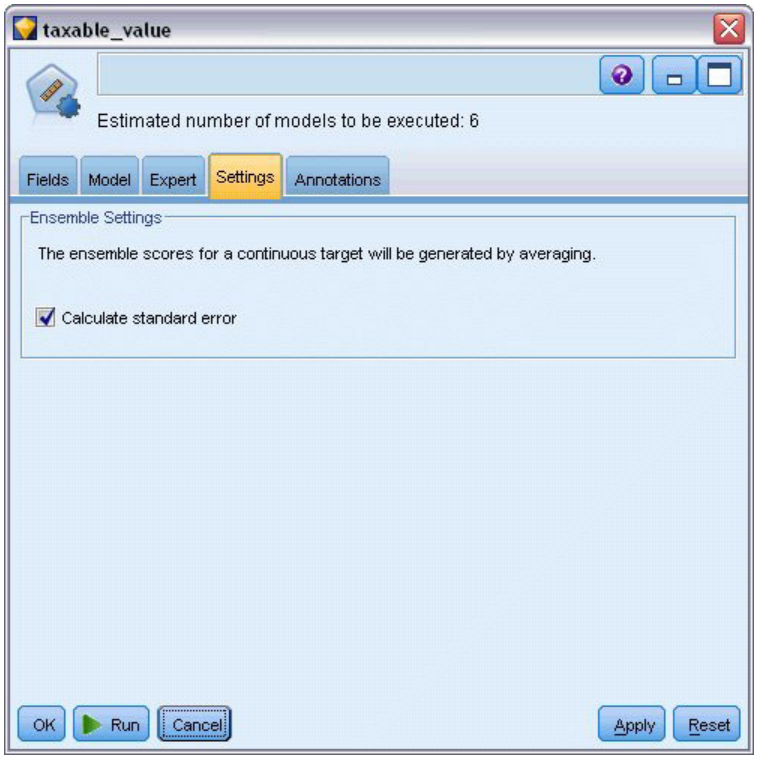


图 47. “自动数值”节点的“设置”选项卡

比较模型

1. 单击“运行”按钮。

将构建模型块，并将其放入画布和窗口右上角的“模型”选用板中。您可以浏览此模型块，或者以多种其他方式对其进行保存或部署。

打开模型块；它将列出在运行期间创建的每个模型的详细信息。（对于在大型数据集中估算数百个模型的实际情况，这可能会花费数小时的时间。）请参阅第 45 页的图 42。

如果需要进一步探索任何单独的模型，可在**模型**列中双击此模型块图标，以向下浏览至单独模型结果，您可以从中生成建模节点、模型块或评估图表。

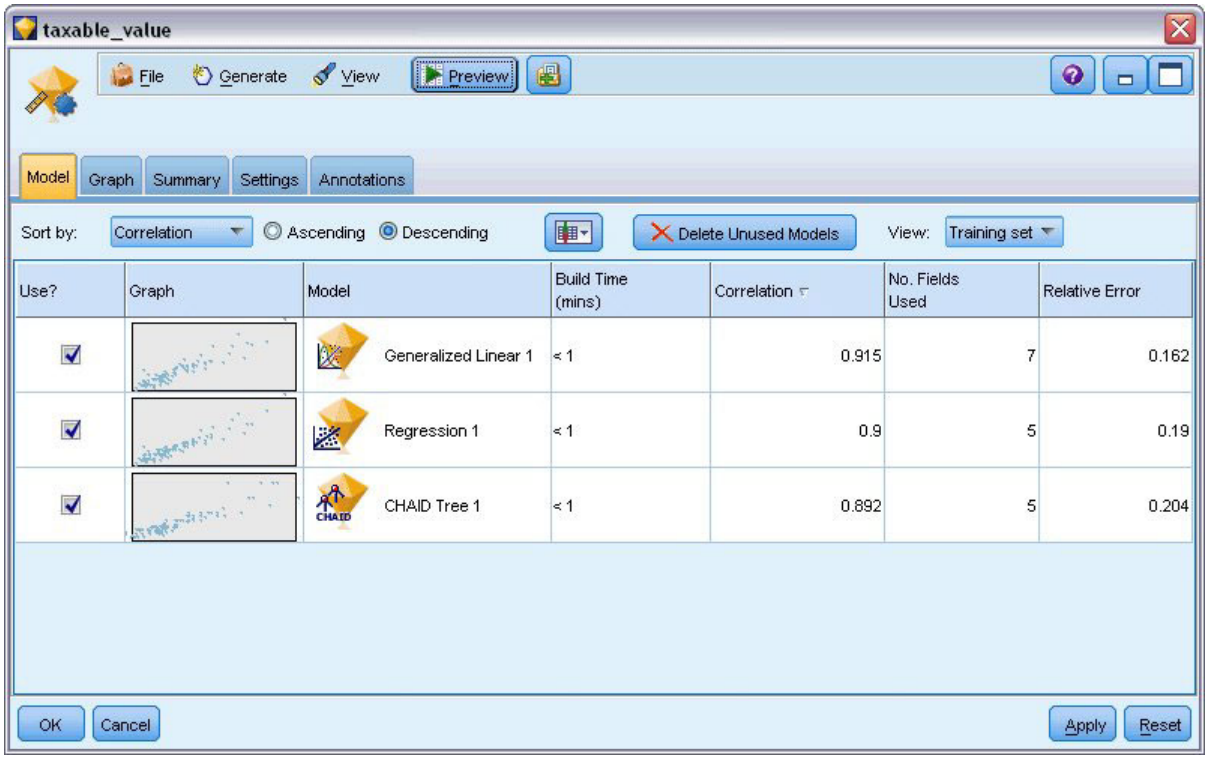


图 48. 自动数值结果

缺省情况下，由于您在“自动数值”节点中选择了相关性度量，因此模型将根据此度量进行排序。出于排序目的，将使用相关性的绝对值，值越接近于 1 表示关系越强。在本度量中，广义线性模型的排序最佳，但是还有几个模型也近乎准确。除此之外，广义线性模型还具有最低的相对错误。

通过单击列标题或从工具栏上的 **排序方式** 列表中选择所需的测量，您可以对不同的列进行排序。

每个图形都显示了相对于模型预测值的观察值图，从而可以快速直观地表示模型之间的相关性。对一个好的模型来说，所有的点都应聚集在对角线附近，在本例中所有模型都是如此。

在**图形**列中，可以双击缩略图生成标准大小的图形。

基于这些结果，您可以决定使用所有这三个最准确的模型。通过结合多个模型的预测，可以避免单个模型的局限性，从而使总体准确性更高。

在**使用?**列中，确保选中了所有三个模型。

在模型块后面附加一个“分析”节点（“输出”选用板）。右键单击分析节点，然后选择**运行**以运行流。

由整体模型生成的平均评分会添加到名为 *$XR-taxable_value* 且相关性为 0.922 的字段中，该相关性值高于三个单独模型中的这些相关性值。该整体节点还显示了较低的平均绝对误差，因此与任何单独模型相比，在应用到其他数据集时，执行效果可能会更好。

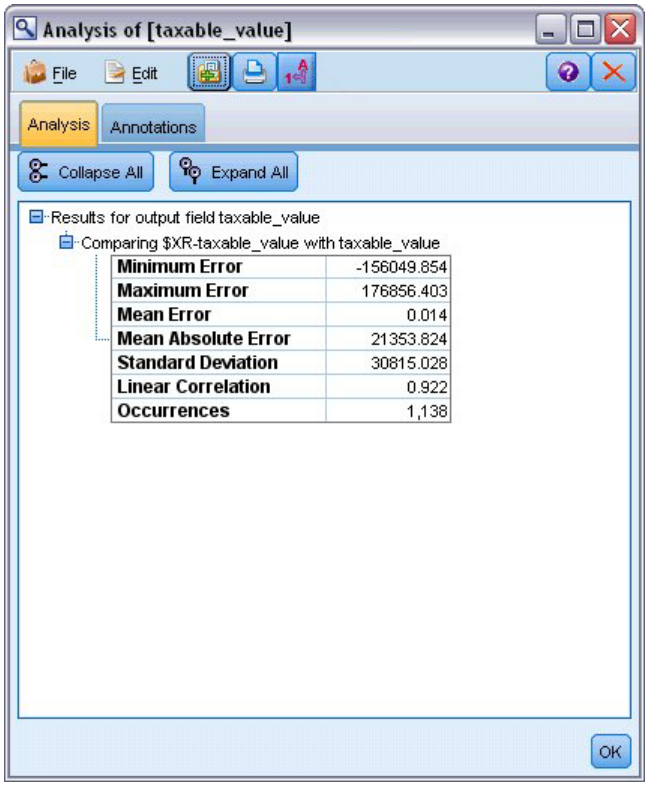


图 49. 自动数值样本流

摘要

综上所述，您使用自动数值节点比较了多种不同的模型，然后选定三个最准确的模型并将它们添加到位于一个整体自动数值模型块内的流中。

- 根据总体准确性，“广义线性”、“回归”和 CHAID 模型在训练数据方面表现最佳。
- 整体模型显示出优于两个单独模型的效果，并且在应用到其他数据集时，执行效果可能会更好。如果您的目标是尽可能多地自动执行这一过程，您可以通过此方法获得在大多数情况下都很稳健的模型，而无需深入挖掘任意一个模型的细节。

第 6 章 自动数据准备 (ADP)

准备数据以进行分析是任何数据挖掘项目中最重要的步骤之一，而从传统上说也是最耗时的步骤之一。自动数据准备 (ADP) 节点为您处理任务、分析您的数据并识别修订、筛选出有问题或者可能不可用的字段、在适当的时候派生新属性以及通过智能筛分技术提高性能。您可以完全自动化地使用节点，允许节点选择并应用修正，或者也可在修正前预览更改，按照需要接受或拒绝。

通过使用 ADP 节点，您可以快速、方便地准备数据以进行数据挖掘，而无需具备相关统计概念的预备知识。如果使用缺省设置运行此节点，那么您将可以更快速地进行构建和评分。

本示例使用名为 *ADP_basic_demo.str* 的流，该流引用名为 *telco.sav* 的数据文件来演示构建模型时通过使用缺省的 ADP 节点设置所增加的准确性。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *ADP_basic_demo.str* 位于 *streams* 目录下。

构建流

1. 要构建流，请添加指向 *telco.sav* 的“Statistics 文件”源节点，telco.sav 位于 IBM SPSS Modeler 安装程序的 *Demos* 目录中。

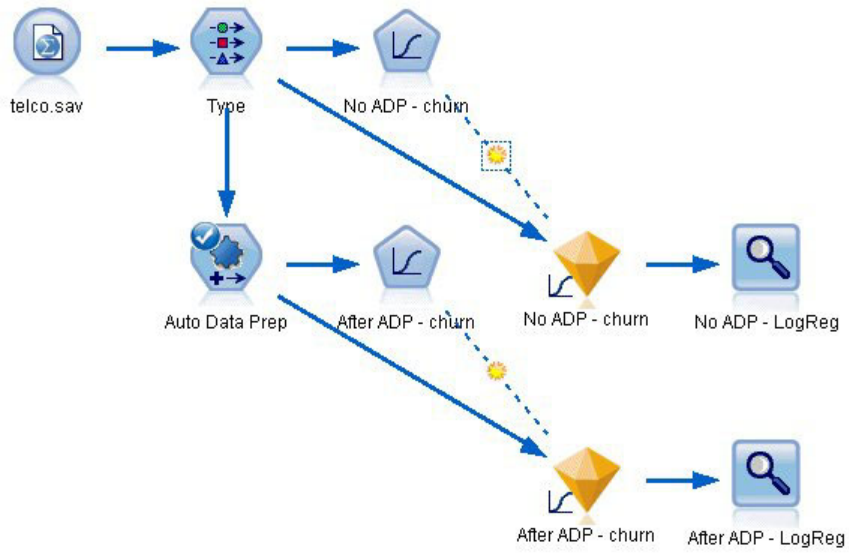


图 50. 构建流

2. 将一个类型节点附加到源节点，将 *churn* 字段的测量级别设置为**标志**，并将角色设置为**目标**。将所有其他字段的角色设置为 **Input**。

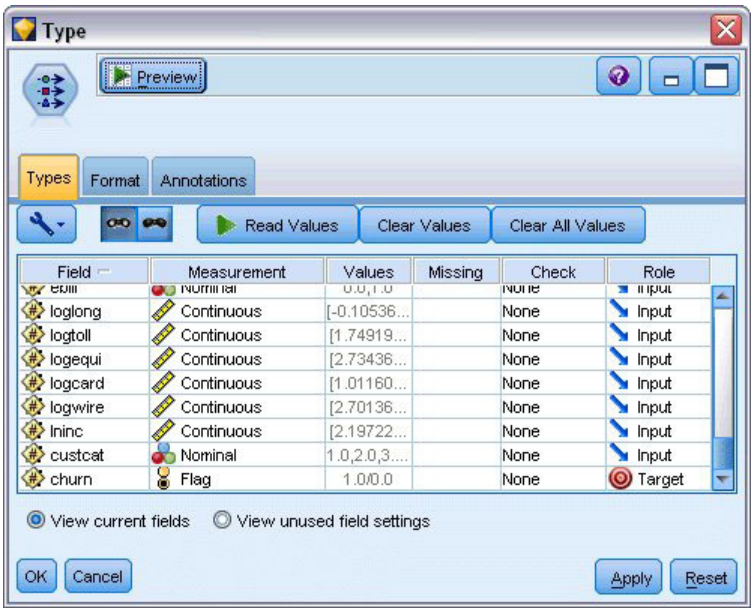


图 51. 选择目标

3. 将 Logistic 节点附加到“类型”节点。
4. 在 Logistic 节点上，单击“模型”选项卡并选择**二项**过程。在*模型名称*字段中，选择**自定义**并输入 `No ADP - churn`。

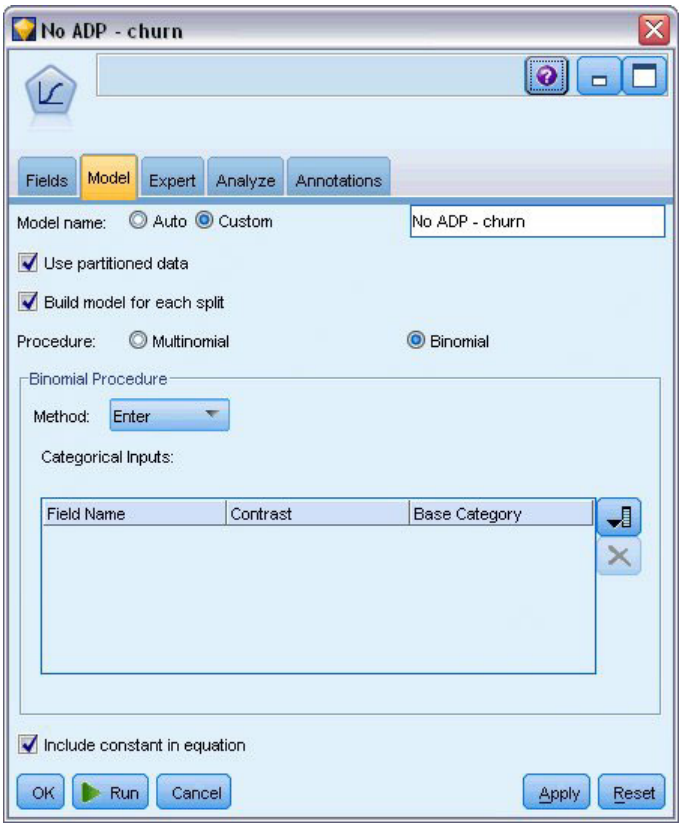


图 52. 选择模型选项

5. 将 ADP 节点附加到“类型”节点。在“目标”选项卡上保留缺省设置，以均衡速度与准确性的方式分析和准备数据。
6. 在“目标”选项卡顶部，单击**分析数据**以分析和处理数据。

ADP 节点上的其他选项使您可以指定优先关注准确性还是处理速度，或者要对许多数据准备处理步骤进行微调。

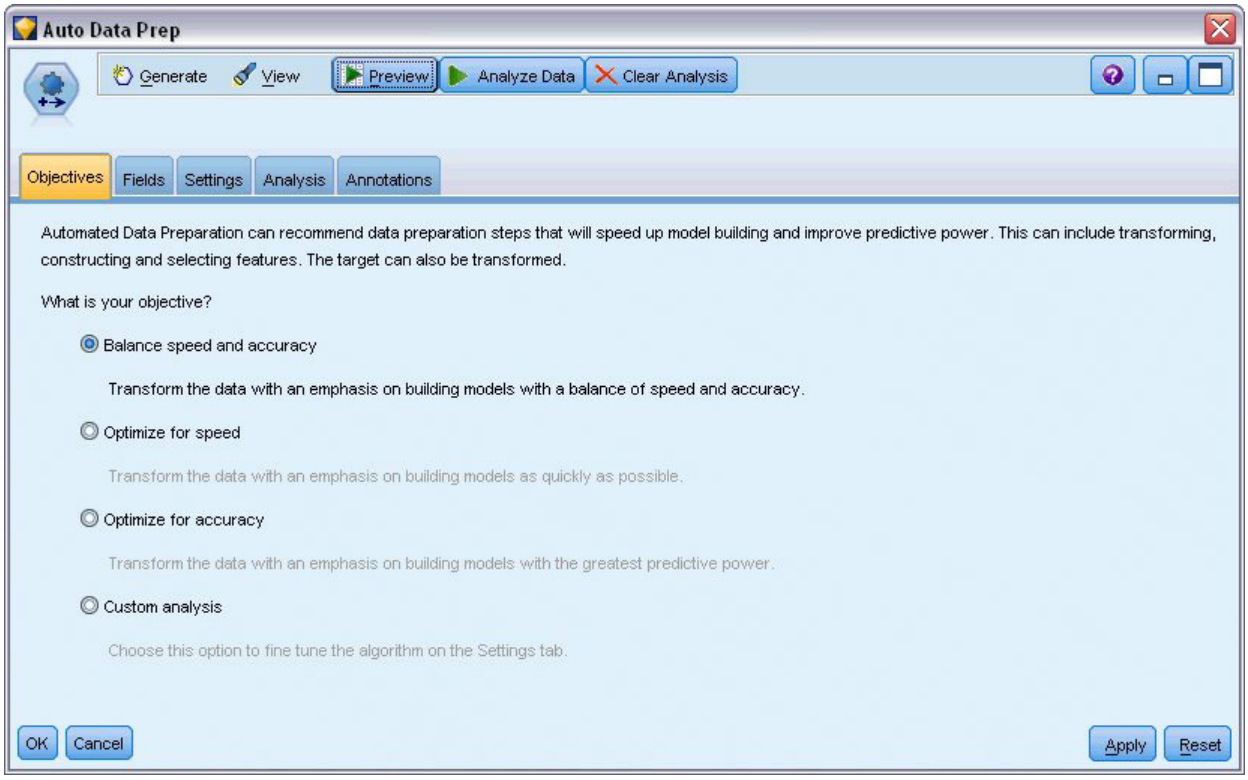


图 53. ADP 缺省目标

数据处理的结果将显示在“分析”选项卡上。**字段处理摘要**显示，在 41 项导入 ADP 节点的数据特征中，有 19 项已转换为辅助处理，而有 3 个因未使用而废弃。

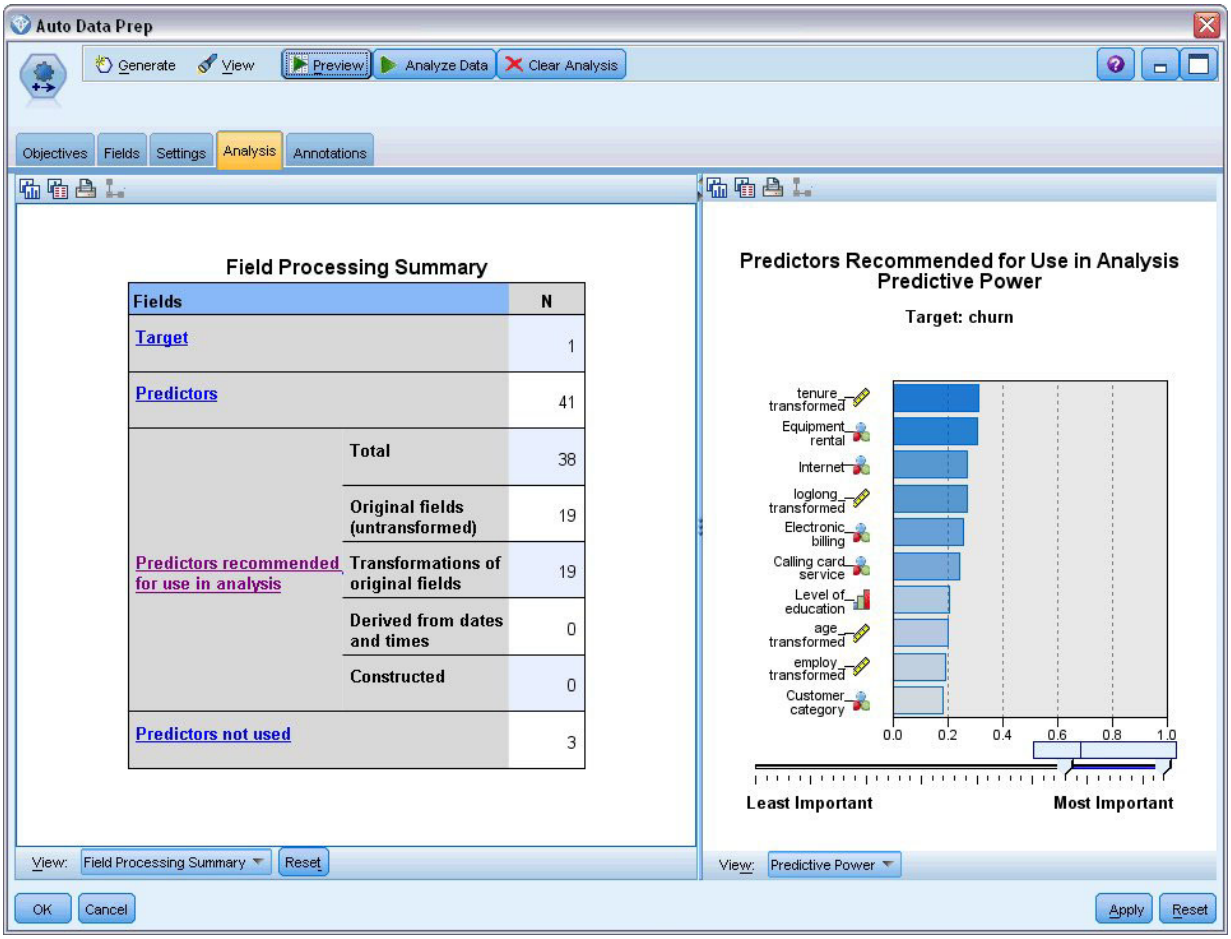


图 54. 数据处理摘要

7. 将 Logistic 节点附加到 ADP 节点。
8. 在 Logistic 节点上，单击“模型”选项卡并选择**二项**过程。在*建模名称*字段中，选择**自定义**并输入 `After ADP - churn`。

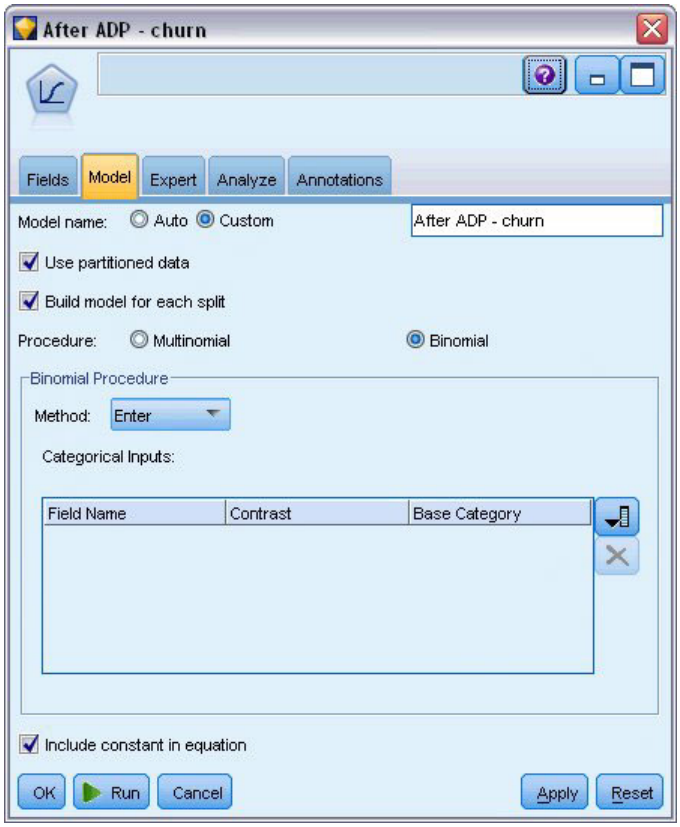


图 55. 选择模型选项

比较模型准确性

1. 同时运行两个 Logistic 节点以创建模型块，这些模型块将添加到流和右上角的“模型”选用板中。

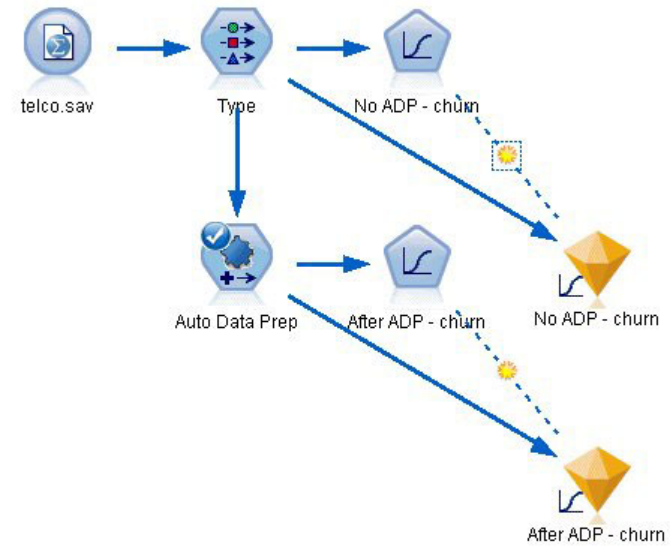


图 56. 附加模型块

2. 将“分析”节点附加到模型块，并使用其缺省设置运行“分析”节点。

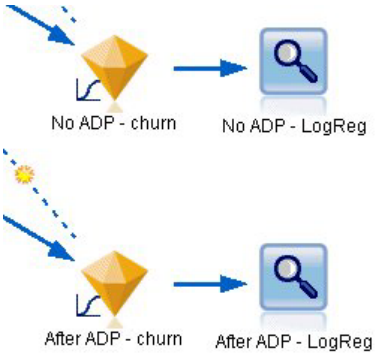


图 57. 附加“分析”节点

对非 ADP 派生模型的“分析”显示，在“Logistic 回归”节点中使用缺省设置运行数据会使模型的准确性较低 - 仅为 10.6%。

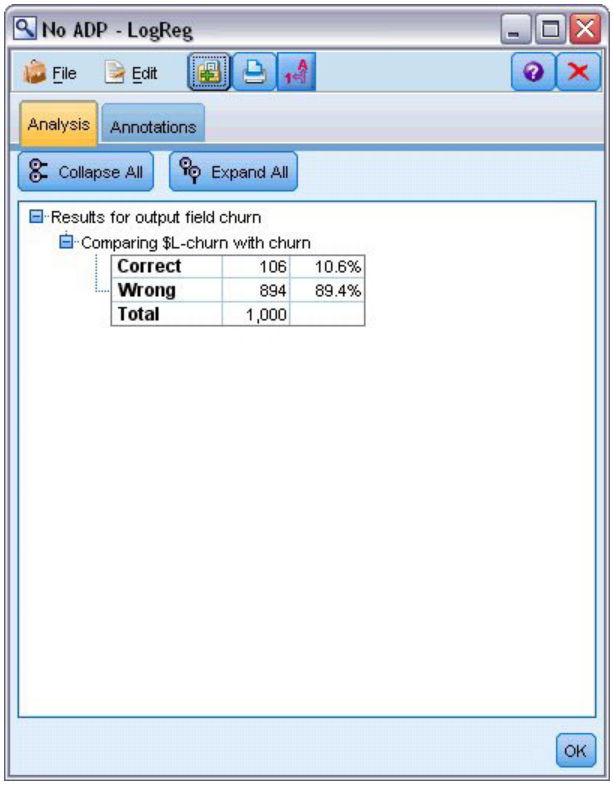


图 58. 非 ADP 派生模型结果

对 ADP 派生模型的“分析”显示，您已通过缺省 ADP 设置运行数据来构建了一个更准确的模型 - 正确率为 78.8%。

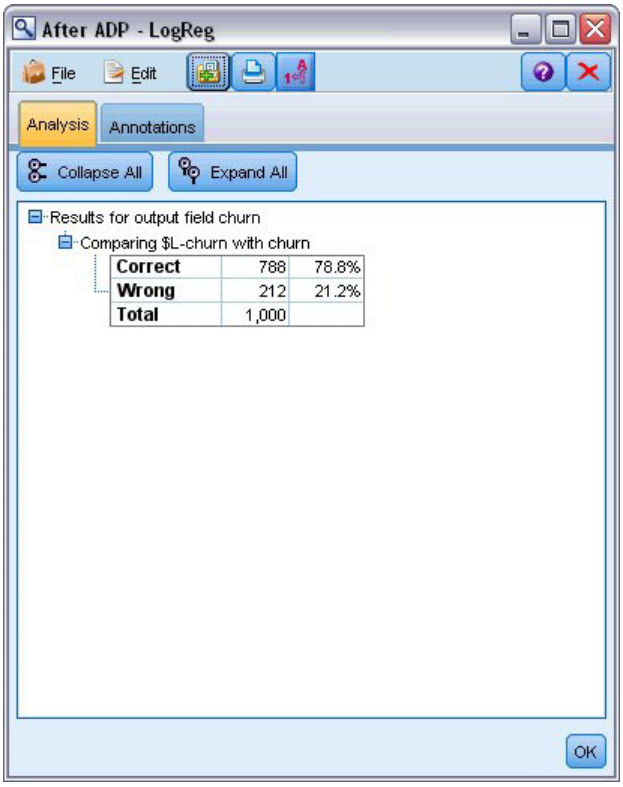


图 59. ADP 派生模型结果

总之，通过运行 ADP 节点对数据处理过程进行微调，您只需很少的直接数据操作便可构建更准确的模型。

当然，如果您有兴趣证明或推翻这一论断，或想构建特定模型，您会发现直接使用模型设置很有用；然而，对于那些构建时间短或有大量数据要处理的模型，ADP 节点可为您带来优势。

有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于安装光盘的 *\Documentation* 目录中。

请注意，本示例中的结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

第 7 章 准备分析数据（数据审核）

数据审核节点将首先全面检查用户导入到 IBM SPSS Modeler 的数据。在初始数据探究过程中经常会使用数据审核报告，该报告显示了摘要统计以及每个数据字段的直方图和分布图，并且它使您可以指定缺失值、离群值和极值的处理。

本例使用名为 *telco_dataaudit.str* 的流，该流引用名为 *telco.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *telco_dataaudit.str* 位于 *streams* 目录下。

构建流

1. 要构建流，请添加指向 *telco.sav* 的“Statistics 文件”源节点，telco.sav 位于 IBM SPSS Modeler 安装程序的 *Demos* 目录中。

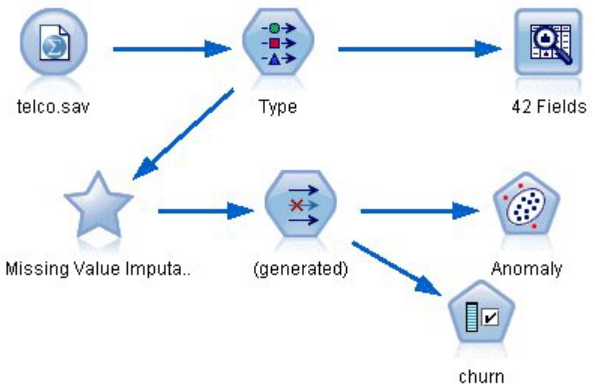


图 60. 构建流

2. 添加“类型”节点以定义字段，并将 *churn* 指定为目标字段（角色为**目标**）。为了使此字段成为唯一目标字段，应将所有其他字段的角色设置为**输入**。

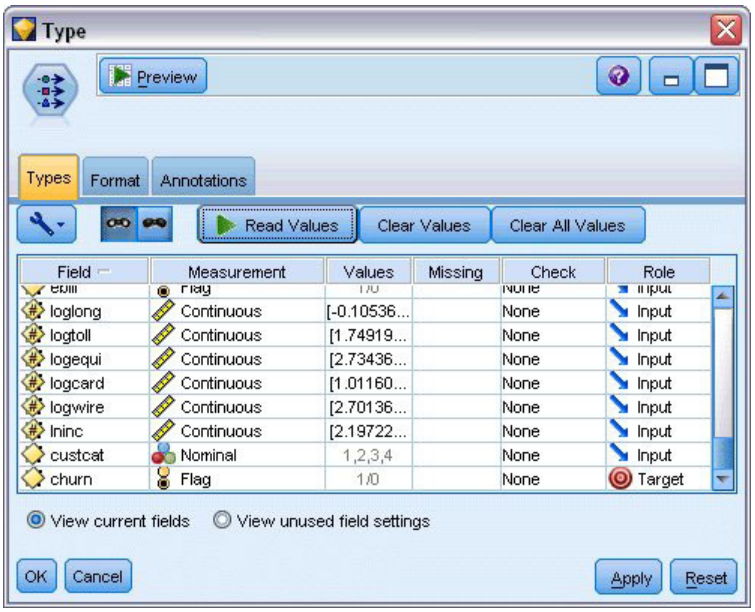


图 61. 设置目标

3. 确认已正确定义字段的测量级别。例如，大多数值为 0 和 1 的字段都可以用作标志字段，但某些字段，比如性别，作为包含两个值的名义字段会更加准确。

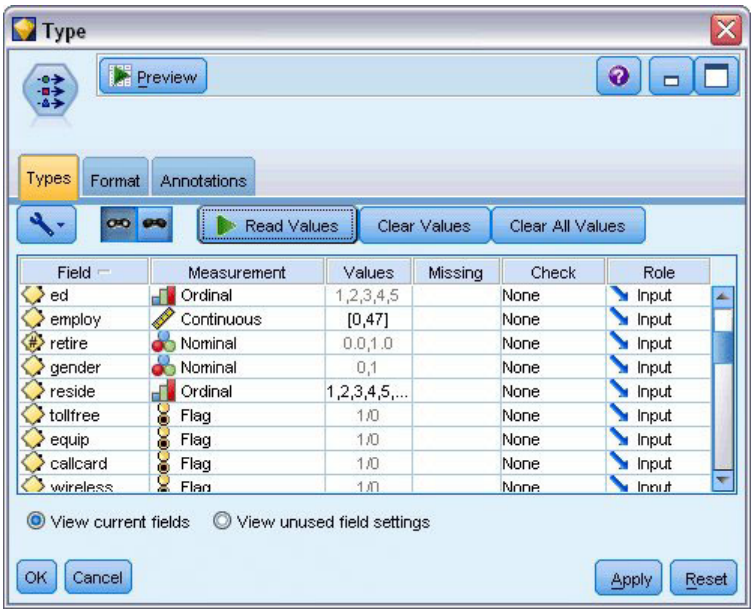


图 62. 设置测量级别

提示: 要更改多个具有相似值（例如 0/1）的字段的属性，请单击*值*列标题以按照该列对字段进行排序，然后使用 Shift 键选择所有要更改的字段。然后可以右键单击选定项，更改所有选定字段的测量级别或其他属性。

4. 将“数据审核”节点附加到流。在“设置”选项卡上，保留缺省设置以便在报告中包含所有字段。由于 *churn* 是类型节点中定义的唯一目标字段，系统会自动将其用作交叠字段。

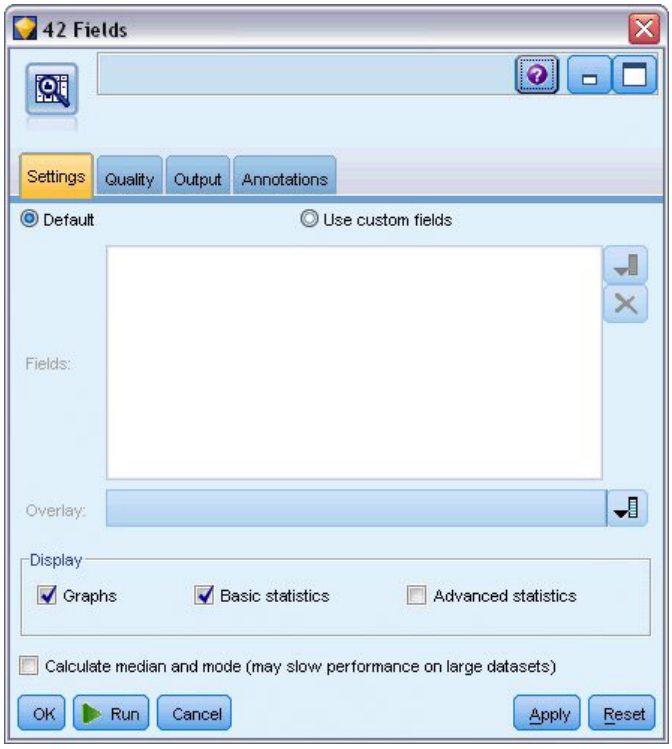


图 63. “数据审核”节点，“设置”选项卡

在“质量”选项卡上，保留检测缺失值、离群值和极值的所有缺省设置，然后单击**运行**。

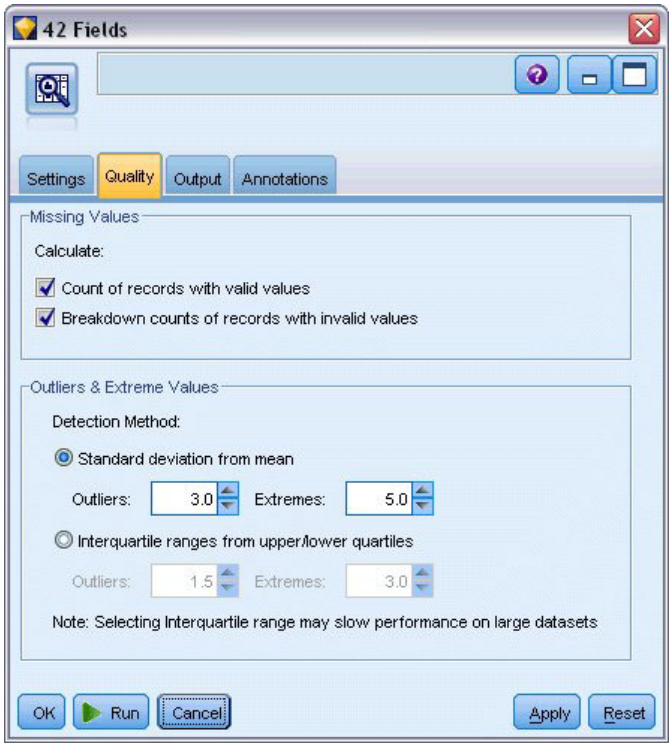


图 64. “数据审核”节点，“质量”选项卡

浏览统计量和图表

将显示“数据审核”浏览器，其中包含每个字段的缩略图和描述统计。

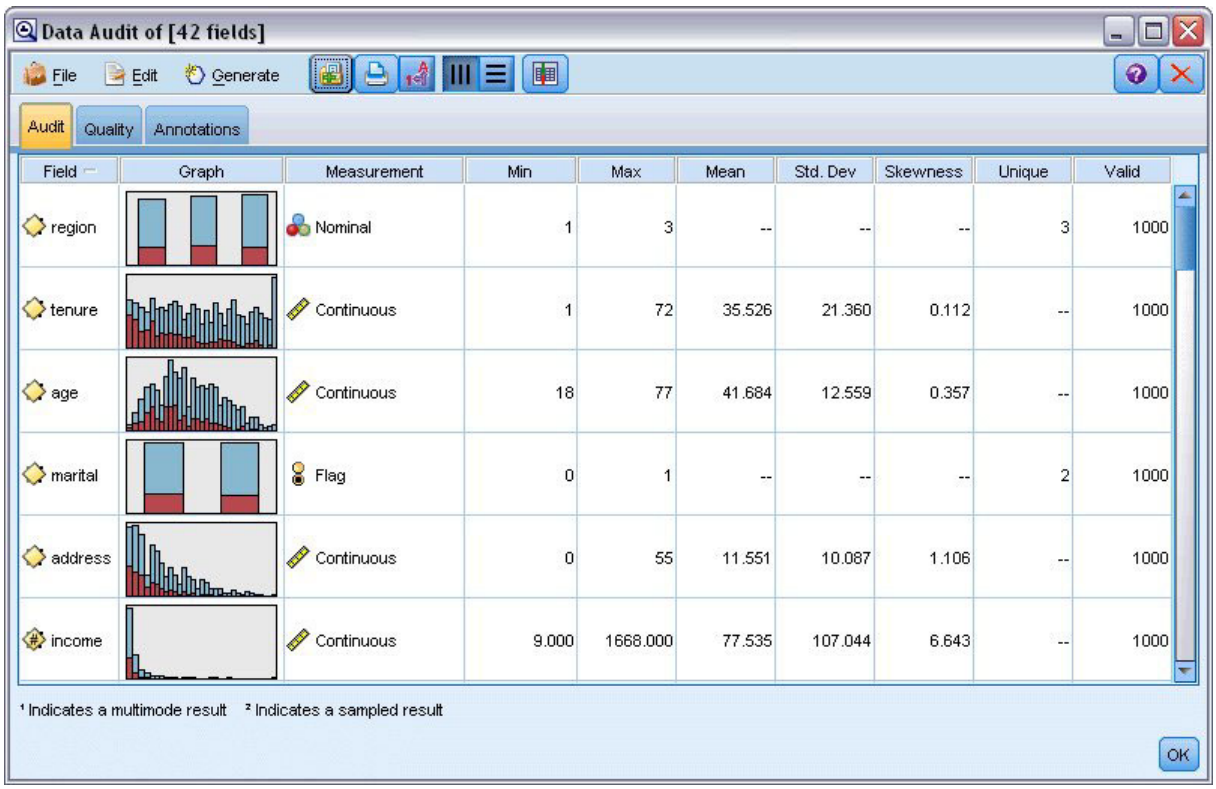


图 65. “数据审核”浏览器

使用工具栏显示字段和值标签，并将图表从水平对齐切换为垂直对齐（仅适用于分类字段）。

1. 也可以使用工具栏或“编辑”菜单选择要显示的统计量。

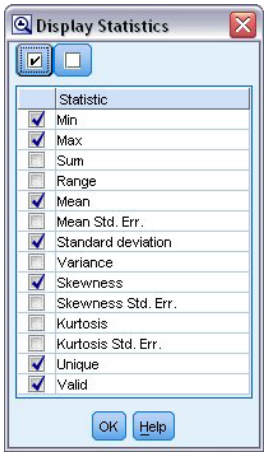


图 66. 显示统计量

双击审核报告中的任意缩略图可以查看该图表的全尺寸版本。由于 *churn* 是流中的唯一目标字段，系统会将其自动用作交叠字段。可以使用图形窗口工具栏来切换字段标签和值标签的显示，也可以单击“编辑方式”按钮对

该图表进行进一步定制。

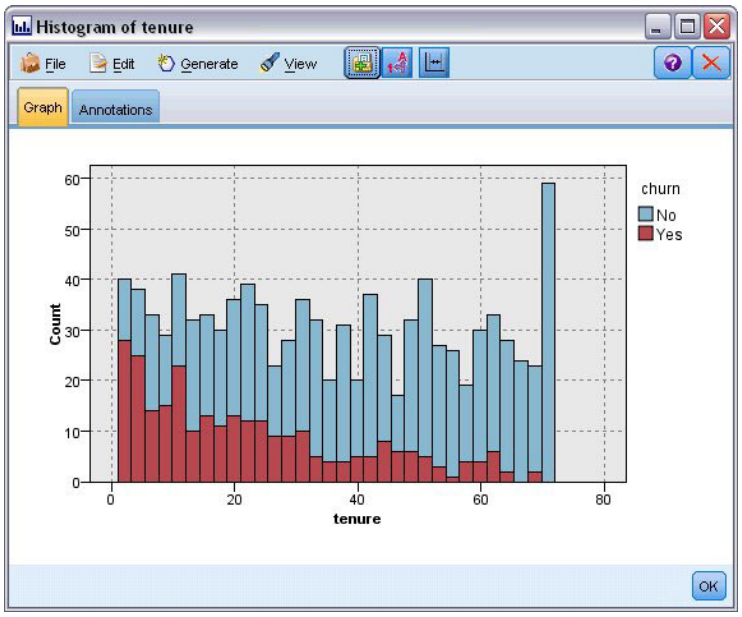


图 67. 保有期直方图

另外，您也可以选择一个或多个缩略图并为每个缩略图生成“图形”节点。生成的节点将放在流画布中，您可以将其添加到流中以重新创建该特定的图形。

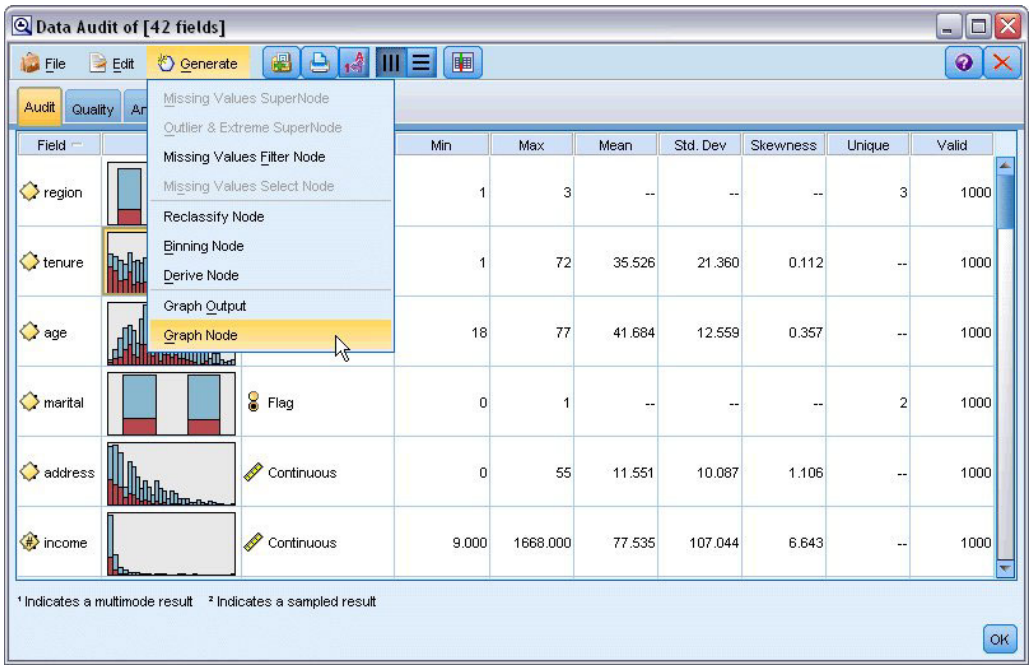


图 68. 生成“图形”节点

处理离群值和缺失值

审核报告中的“质量”选项卡显示有关离群值、极值和缺失值的信息。

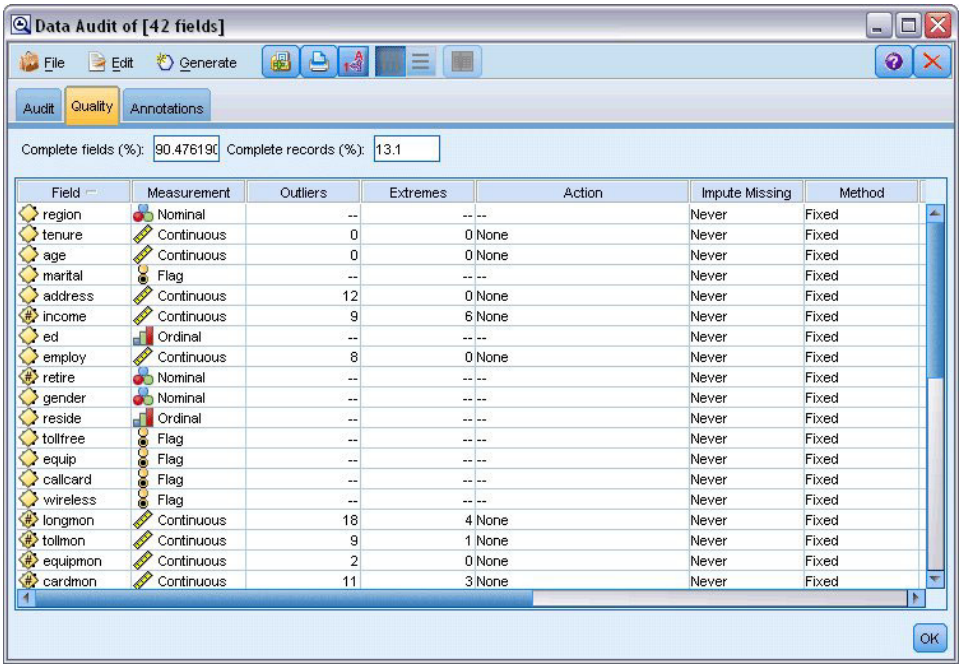


图 69. “数据审核”浏览器，“质量”选项卡

也可以指定处理这些值的方法并生成超节点，以自动应用各种变换。例如，您可以使用多种方法（包括 C&RT 算法）来选择一个或多个字段并选择插补或替换这些字段的缺失值。

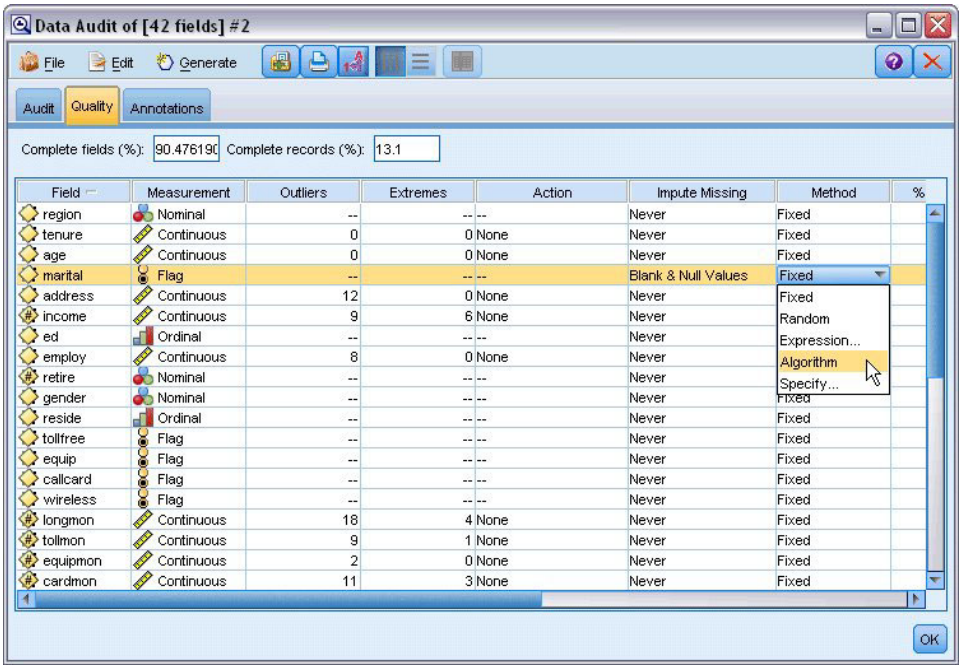


图 70. 选择插补方法

指定用于一个或多个字段的归因方法后，要生成缺失值超节点，请从菜单中选择:

生成 > 缺失值超节点

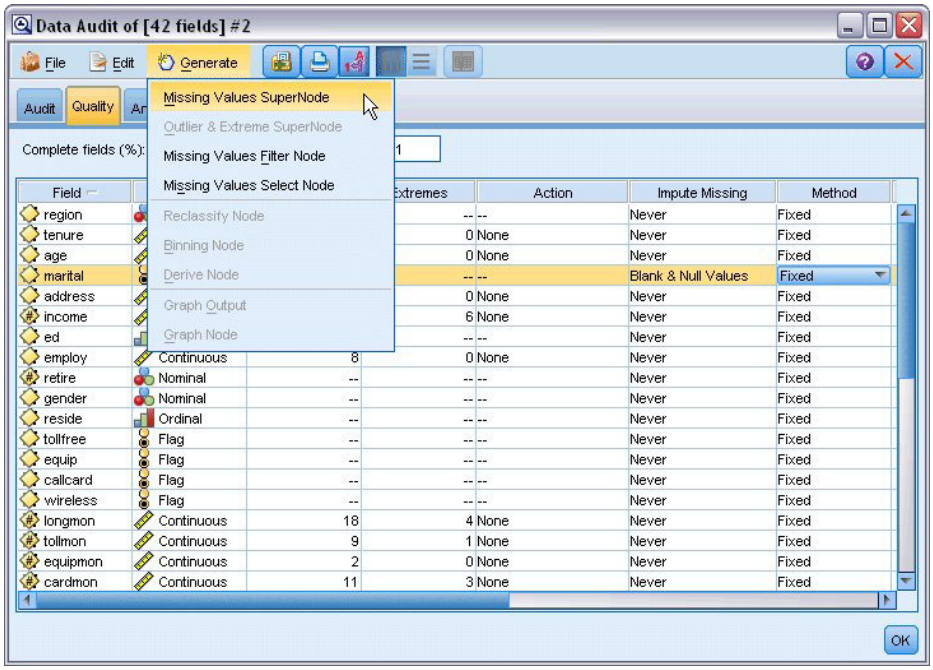


图 71. 生成超节点

生成的超节点将添加到流画布中，您可以在该流画布中将此超节点附加到流中以应用各种变换。

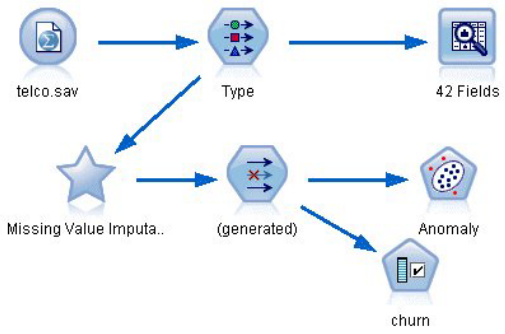


图 72. 具有缺失值超节点的流

实际上，超节点包含执行所需变换的一系列节点。要了解超节点的工作方式，可编辑超节点并单击 **放大** 。

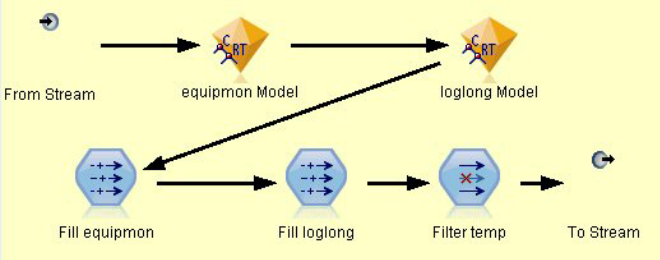


图 73. 放大超节点

例如，对于使用算法插补的每个字段，将有一个独立的 C&RT 模型，以及一个使用该模型预测的值来替换空白值和空值的“填充”节点。用户可以添加、编辑或除去超节点中的特定节点，从而对行为进行进一步定制。

另外，也可以生成“选择”节点或“过滤”节点，以除去具有缺失值的字段或记录。例如，您可以过滤掉质量百分比低于指定阈值的任何字段。

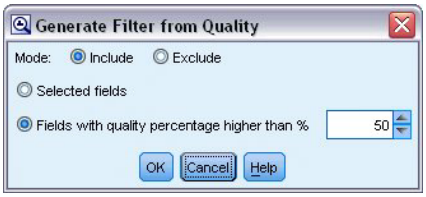


图 74. 生成“过滤”节点

也可以用类似的方法来处理离群值和极值。指定要对每个字段执行的操作（强制、废弃或取消）并生成超节点，以应用各种变换。

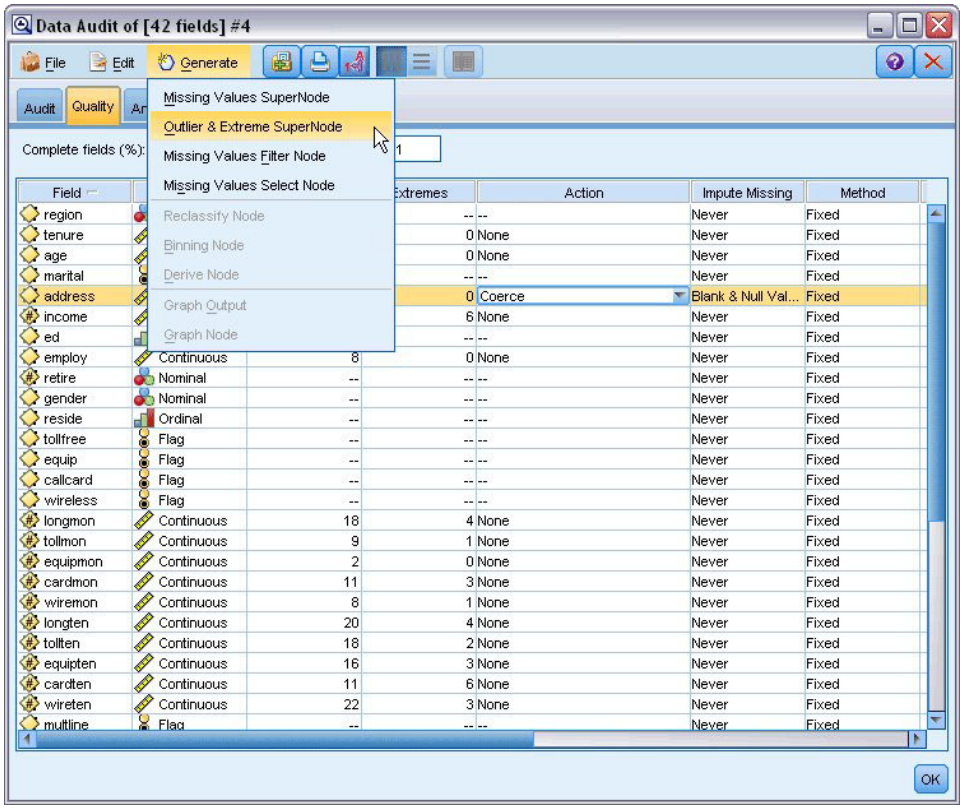


图 75. 生成“过滤”节点

完成审核并将生成的节点添加到流中之后，您可以继续进行分析。您可能会选择使用“异常检测”、“特征选择”或其他多种方法来进一步筛选数据。

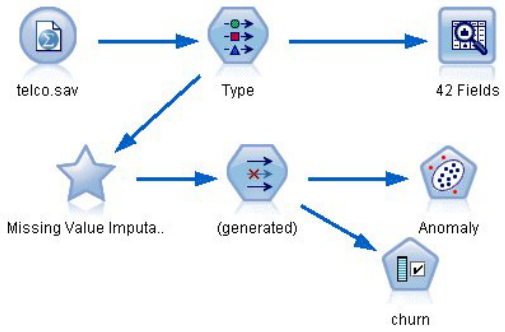


图 76. 具有缺失值超节点的流

第 8 章 药物治疗（勘察表/C5.0）

在本节中，假设您是一位收集研究数据的医学研究人员。您已收集了关于身患同一疾病的一组患者的数据。在治疗过程中，每位患者均对五种药物中的一种有明显反应。您的其中一项职责是通过数据挖掘找出适用于今后患有此疾病的患者的药物。

此示例使用名为 *druglearn.str* 的流，此流引用名为 *DRUG1n* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *druglearn.str* 位于 *streams* 目录中。

此 demo 中使用的数据字段包括:

数据字段	描述
年龄	（数值）
性别	*男*或*女*
BP	血压: *高*、*正常*或*低*
胆固醇	血胆固醇: *正常*或*高*
Na	血液中钠的浓度
K	血液中钾的浓度
Drug	对患者有效的处方药

读取文本数据

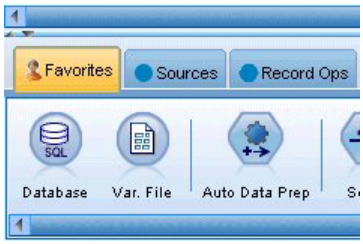


图 77. 添加“变量文件”节点

您可以使用**“变量文件”节点**读取定界文本数据。可以从选用板中添加“变量文件”节点，方法是单击**源**选项卡找到此节点，或者使用**收藏夹**选项卡（缺省情况下，其中包含此节点）。然后，双击新添加的节点以打开其对话框。

单击紧挨“文件”框右边以省略号“...”标记的按钮，浏览到您系统中的 IBM SPSS Modeler 安装目录。打开 *Demos* 目录，然后选择名为 *DRUG1n* 的文件。

确保选中了**从文件读取字段名称**，注意已加载此对话框中的字段和值。

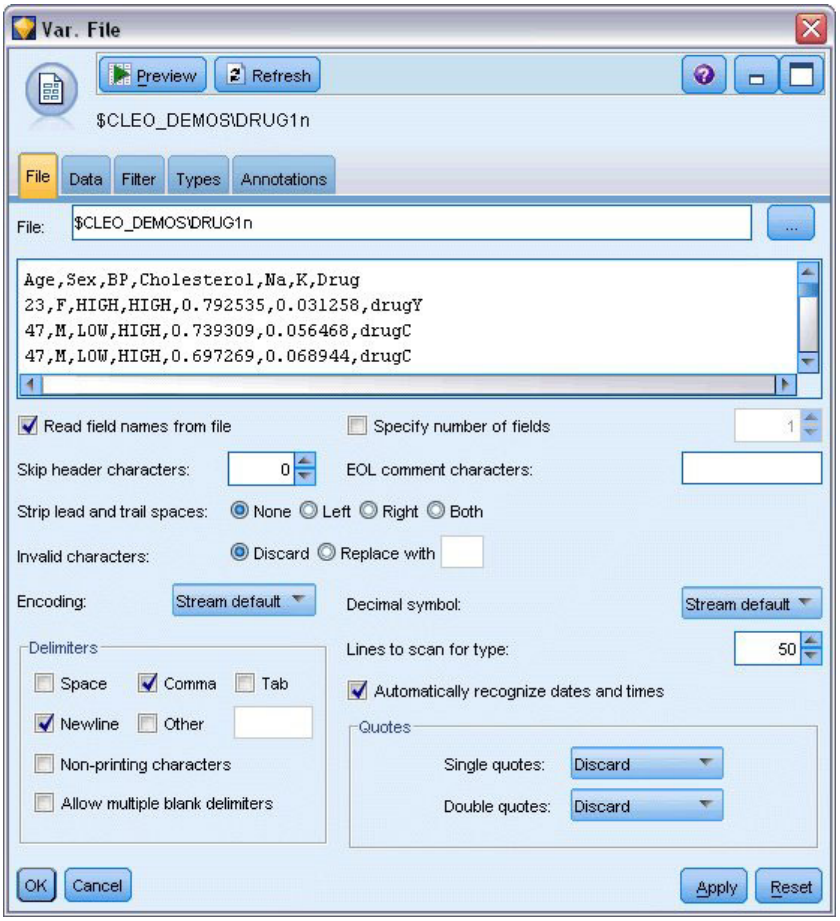


图 78. "变量文件"对话框

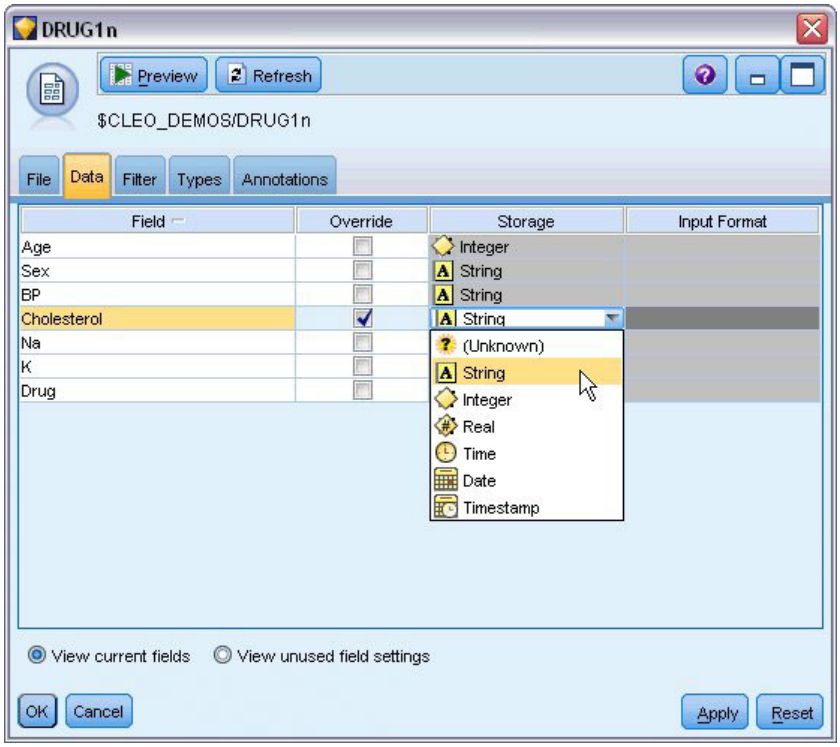


图 79. 更改字段的存储类型

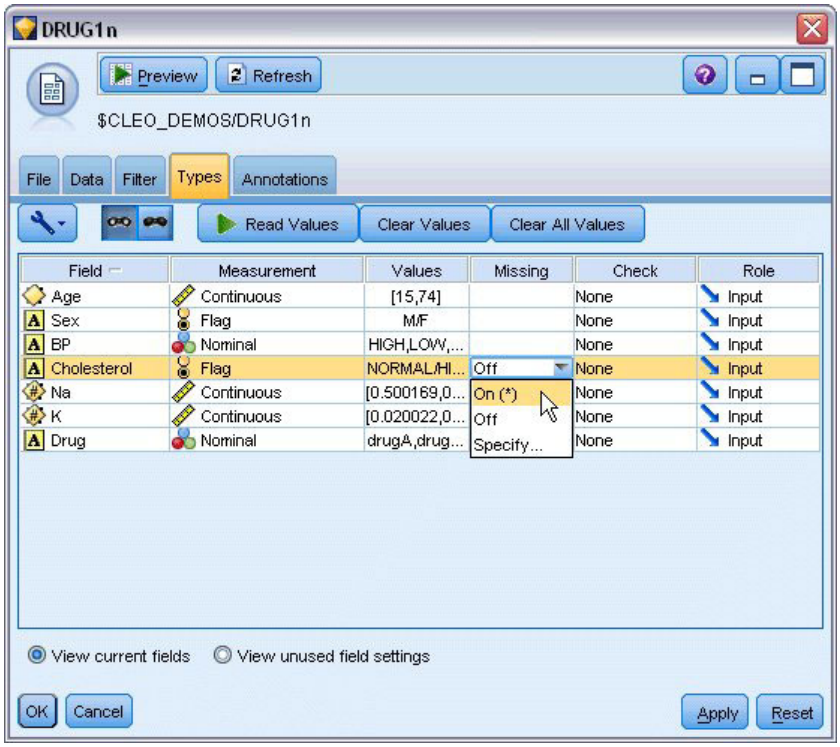


图 80. 选择“类型”选项卡中的“值”选项

单击**数据**选项卡，覆盖和更改某个字段的**存储**。注意，存储不同于**测量**，即，数据字段的测量级别（或用途类型）。**类型** 选项卡可帮助您了解数据中的更多字段类型。还可以选择 **读取值** 来查看各个字段的实际值，具体取决于您在 *值* 列中的选择。此过程称为 **实例化** 。

添加表

由于您已装入数据文件，因此可能希望浏览某些记录的值。其中一个方法就是构建一个包含“表”节点的流。要将“表”节点放入流中，请双击选用板中的图标或者将该节点拖放到画布上。

图 81. 已连接到数据源的“表”节点

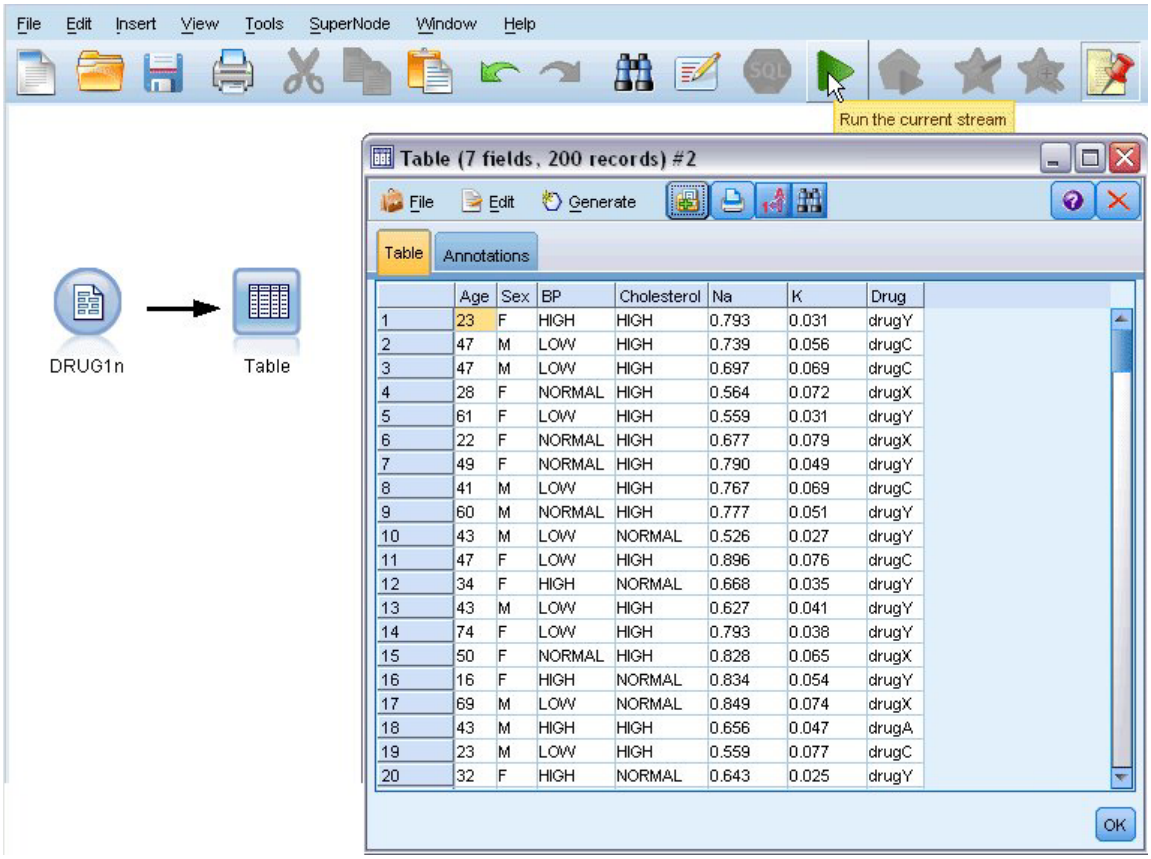


图 82. 从工具栏运行流

双击选用板中的某个节点会将该节点自动连接到流画布中的选定节点。另外，如果尚未连接节点，那么您可以使用鼠标中键将“源”节点连接到“表”节点。要模拟鼠标中键操作，请在使用鼠标时按下 Alt 键。要查看表，请单击工具栏上的绿色箭头按钮运行流，或者右键单击“表”节点，然后选择**运行**。

创建分布图

数据挖掘过程中，创建汇总视图通常有助于研究数据。IBM SPSS Modeler 提供了若干不同类型的图表供您选择，具体取决于您要汇总分析的数据类型。例如，要找出每种药物的对症患者的比例，请使用“分布”节点。

将“分布”节点添加到流，并将其与“源”节点相连接，然后双击该节点以编辑要显示的选项。

选择 *药物* 作为要显示其分布的目标字段。然后，在对话框中单击**运行**。

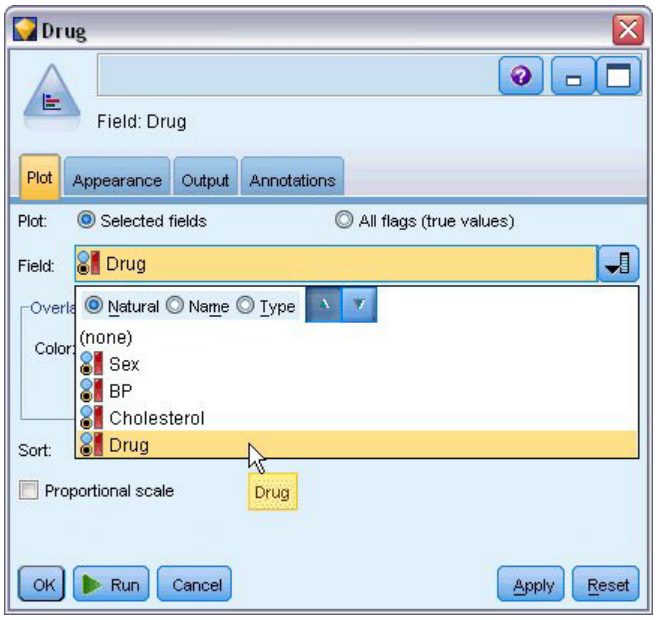


图 83. 选择药物作为目标字段

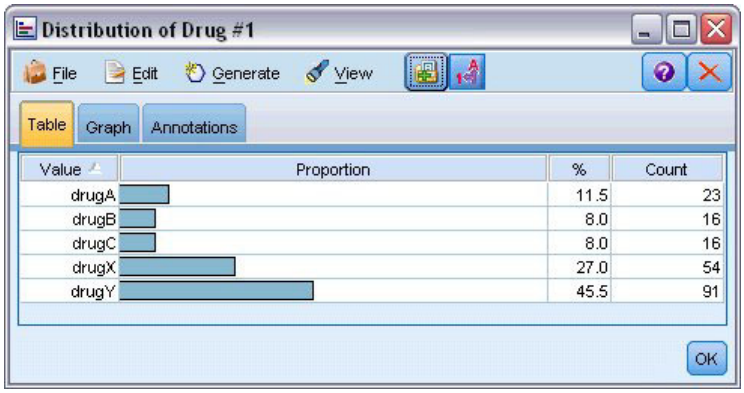


图 84. 对症药物类型分布

最终图形有助于您查看数据的“结构”。结果表明，药物 *Y* 的对症患者最多，而药物 *B* 和药物 *C* 的对症患者最少。

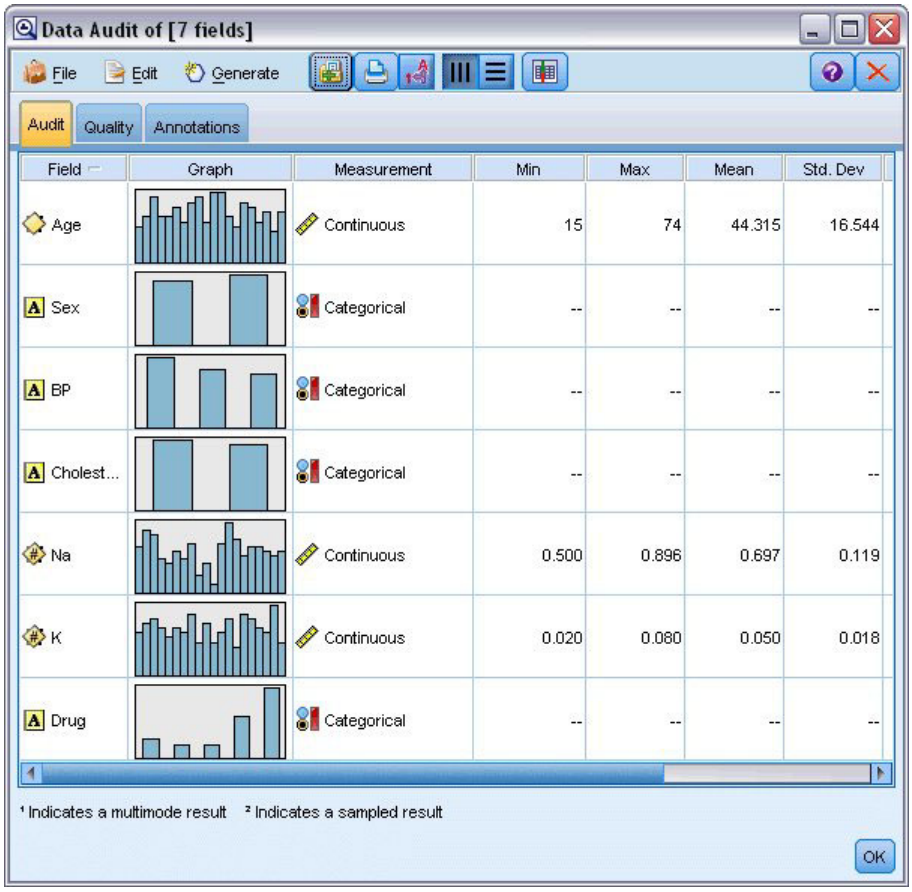


图 85. 数据审核结果

此外，您还可以附加并执行“数据审核”节点，以便立即快速浏览所有字段的分布图和直方图。可以在“输出”选项卡中找到数据审核节点。

创建散点图

现在我们来看一下有哪些因素会对*药物*（目标变量）产生影响。作为研究人员，您知道血液中钠和钾的浓度是两个重要因素。由于这两者都是数值，因此您可以使用药物类别作为颜色叠加来创建关于钠与钾的散点图。

将“图”节点放在工作空间中，并将其连接到“源”节点，然后双击以编辑该节点。

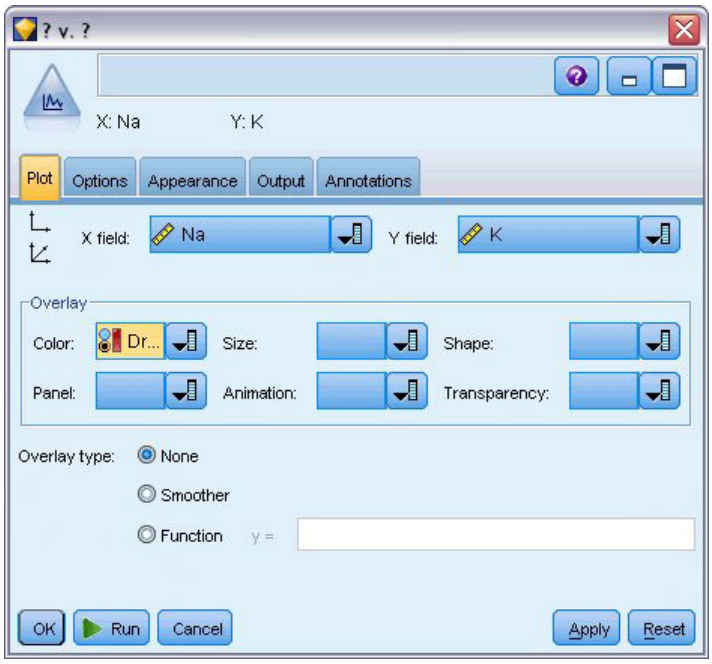


图 86. 创建散点图

在“散点图”选项卡中，选择 *Na* 作为 X 字段，选择 *K* 作为 Y 字段，并选择 *药物* 作为交叠字段。然后，单击**运行**。

此图清楚地显示了一个阈值，高于此阈值时的对症药物始终为药物 *Y*，而低于此阈值时的对症药物不是药物 *Y*。此阈值是一个比率，即钠 (*Na*) 与钾 (*K*) 的比率。

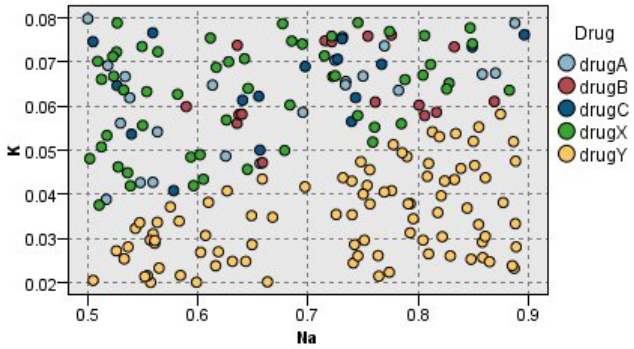


图 87. 药物分布散点图

创建网络图

由于许多数据字段都是分类字段，因此您也可尝试绘制反映不同类别之间的关联的网络图。首先，将 Web 节点连接到工作空间中的“源”节点。在“网络节点”对话框中，选择 *BP*（血压）和*药物*。然后，单击**运行**。

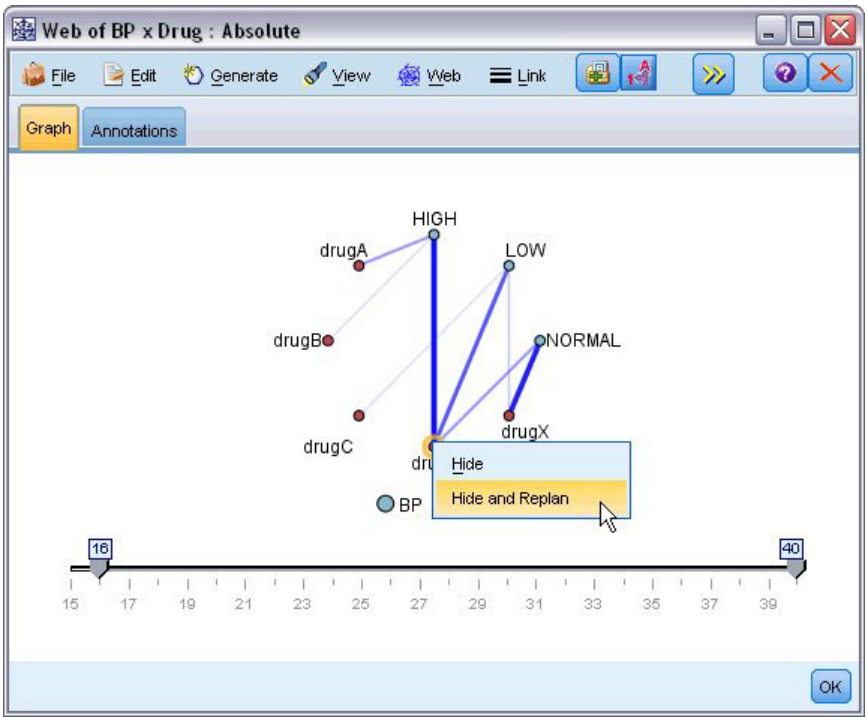


图 88. 药物和血压网络图

此图显示，药物 *Y* 与三种级别的血压均相关。这并不奇怪，因为您早已看出 *Y* 是最佳药物。要侧重于其他药物，您可以隐藏药物 *Y*。在**视图**菜单中，选择**编辑方式**，然后在药物 *Y* 点上单击鼠标右键，并选择**隐藏并重新规划**。

简图中隐藏了药物 *Y* 及其所有链接。现在您可以清楚地看到，只有药物 *A* 和 *B* 与高血压有关。只有药物 *C* 和 *X* 与低血压有关。并且，正常血压仅与药物 *X* 相关联。此时，对于指定的患者，您仍然无法在药物 *A* 与 *B* 之间或药物 *C* 与 *X* 之间作出选择。此时建模可以助您一臂之力。

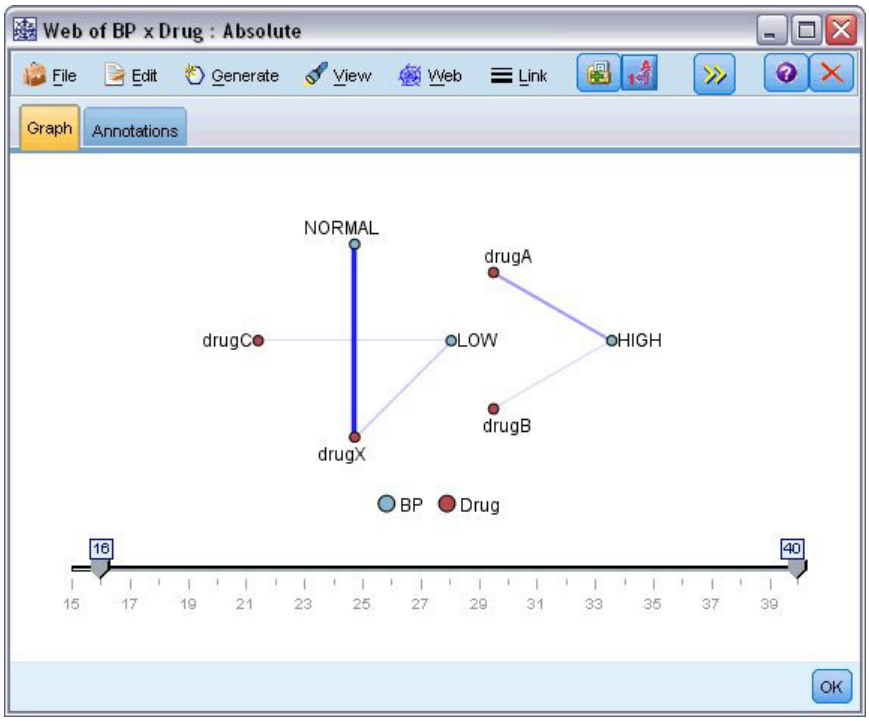


图 89. 已隐藏药物 Y 及其链接的网络图

导出新字段

由于钠与钾的比似乎可以用来预测何时可以使用药物 *Y*，因此您可以为每条记录导出一个包含此比值的字段。稍后在您构建模型以预测何时使用五种药物中的每种药物时可以使用此字段。为了简化流布局，请首先删除 DRUG1n 源节点外的所有节点。将“派生”节点（“字段选项”选项卡）附加到 DRUG1n，然后双击此节点以进行编辑。

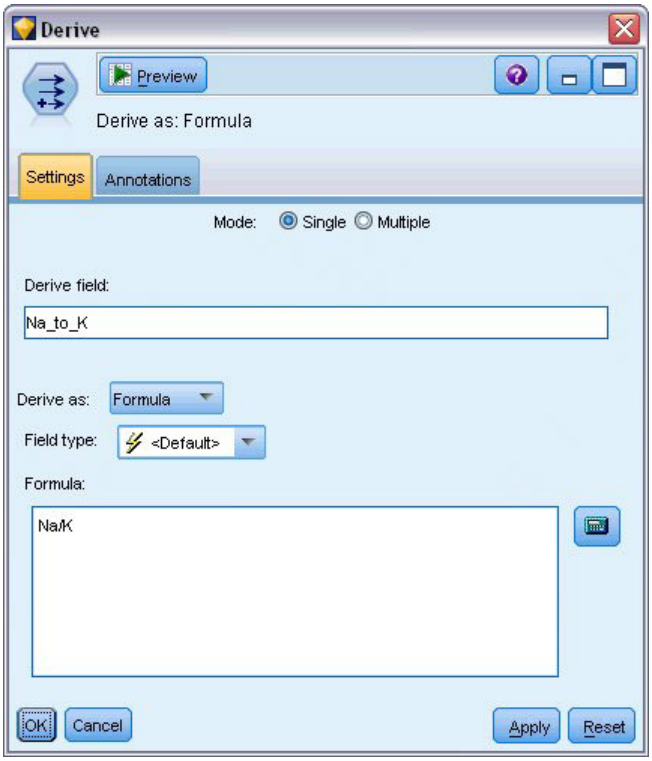


图 90. 编辑“派生”节点

将新字段命名为 *Na_to_K* 。由于是通过将钠值除以钾值获取新字段，所以请在公式中输入 `Na/K` 。您也可以通过单击该字段右侧的图标来创建公式。这将打开“表达式构建器”，这是一种使用函数、操作数、字段及其值的内置列表以交互方式创建表达式的方法。

您可以通过将“直方图”节点附加到“派生”节点来检查新字段的分布情况。在“直方图”节点对话框中，将 *Na_to_K* 指定为要绘制的字段，并将*药物*指定为交叠字段。

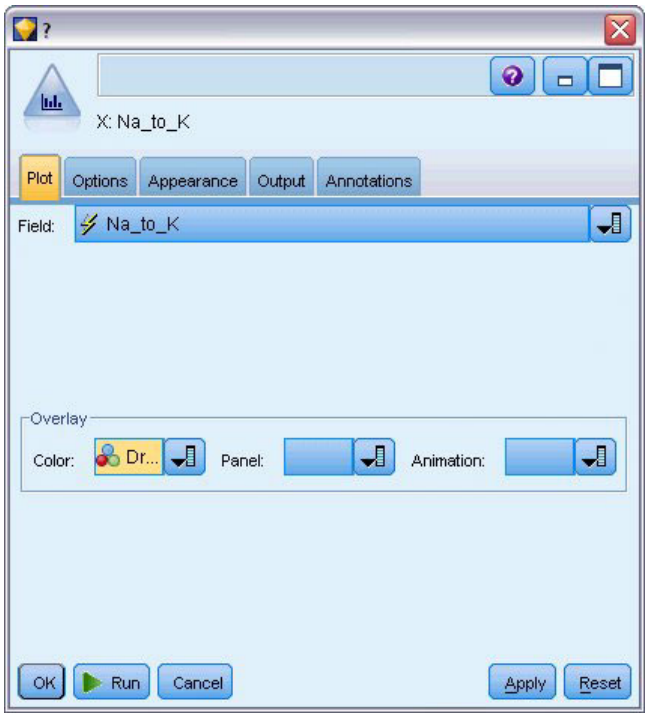


图 91. 编辑“直方图”节点

运行流时，将在此处显示该图形。您可以根据显示结果得出以下结论：当 *Na_to_K* 字段的值等于或大于 15 时，应选择药物 *Y* 。

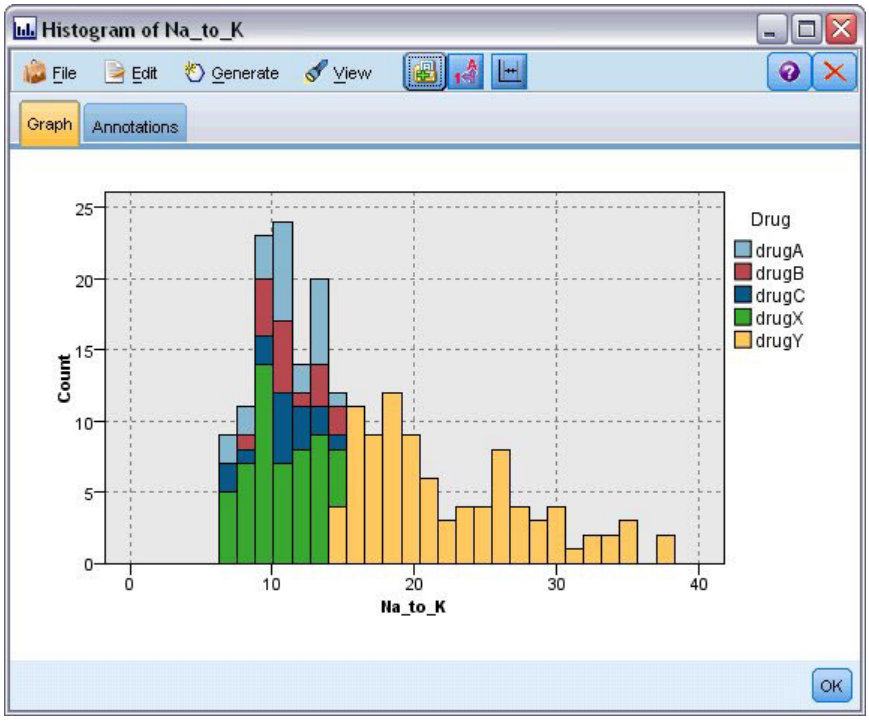


图 92. 直方图显示

构建模型

通过探索和处理数据，您可以提出一些假设。血液中钠与钾的比率以及血压似乎都会影响药物的选择。但您尚无法完全解释清楚所有关系。此时似乎可以通过建模找出某些答案。此种情况下，您可以尝试使用规则构建模型 (C5.0) 来拟合数据。

由于使用的是导出字段 *Na_to_K*，您可以过滤掉原始字段 *Na* 和 *K*，以避免在建模算法中重复操作。上述操作可通过过滤节点完成。

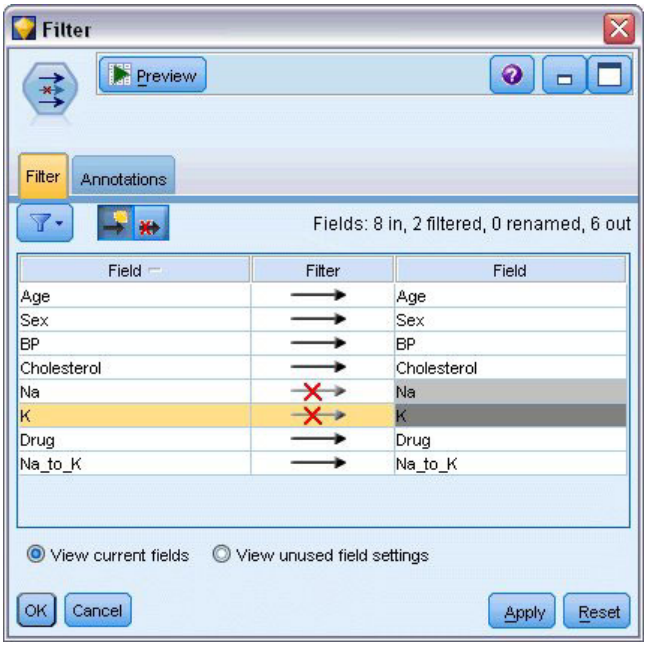


图 93. 编辑“过滤”节点

在“过滤”选项卡上，单击 *Na* 和 *K* 旁边的箭头。箭头上显示的红色 X 指示现在要过滤掉这些字段。

然后，附加一个已连接到“过滤”节点的“类型”节点。“类型”节点允许您指示要使用的字段类型以及如何使用这些字段预测结果。

在“类型”选项卡上，将*药物*字段的角色设置为**目标**，表明您要预测该*药物*字段。将其他字段的角色设置为**输入**，表示这些字段将用作预测变量。

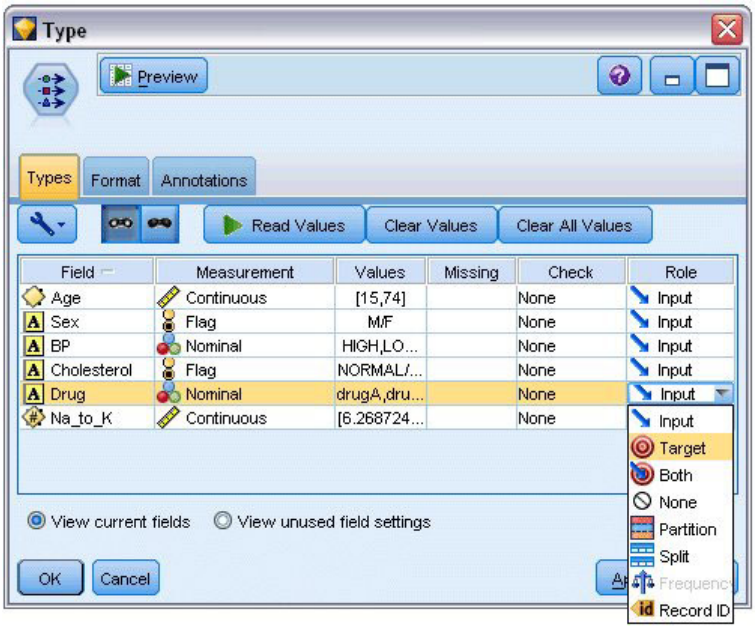


图 94. 编辑“类型”节点

要估算此模型，请将节点 C5.0 放在工作空间中，然后将此节点附加到流的末端（如图所示）。单击绿色**运行**工具栏按钮运行流。

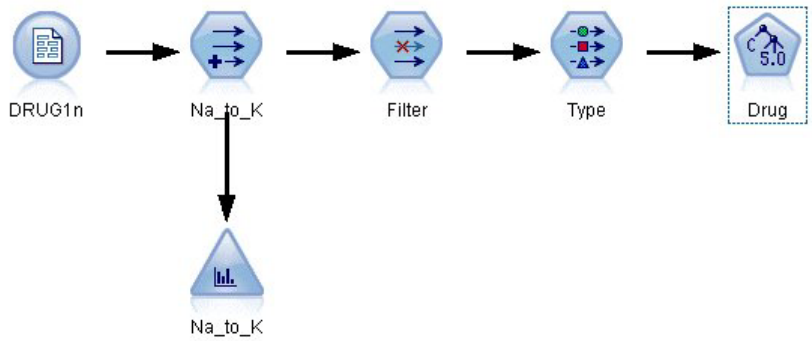


图 95. 添加 C5.0 节点

浏览模型

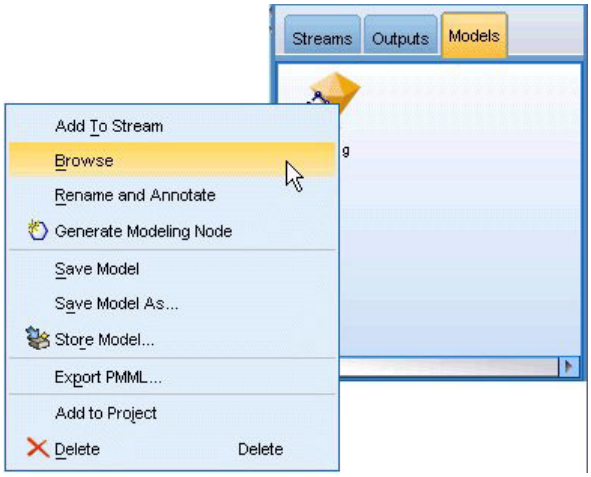


图 96. 浏览模型

执行 C5.0 节点时，模型块将添加到流和窗口右上角的“模型”选用板中。要浏览模型，右键单击任一图标并从上下文菜单选择**编辑**或**浏览**。

规则浏览器以决策树形式显示 C5.0 节点所生成的规则集。最初，决策树处于折叠状态。要展开决策树，请单击 **所有** 按钮显示所有层。

图 97. 规则浏览器

现在，您可以看到此决策树的缺失部分。对于 *Na* 与 *K* 的比率小于 14.64 的高血压人员，年龄将决定药物的选择。对于低血压患者，胆固醇含量似乎是最有力的预测变量。

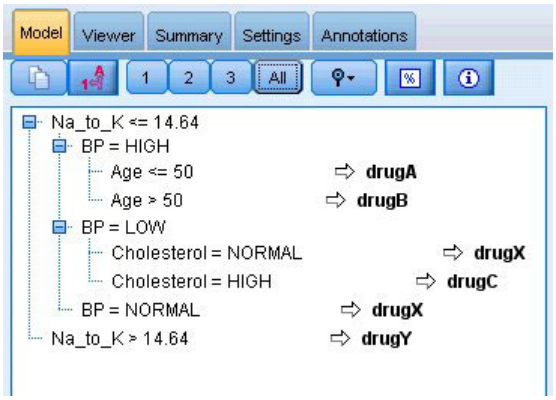


图 98. 完全展开的规则浏览器

通过单击 **查看器** 选项卡，还可以更复杂的图表形式查看同一决策树。通过此图表形式，您可以更轻松地查看各个血压类别的观测值数量以及各个观测值的百分比。

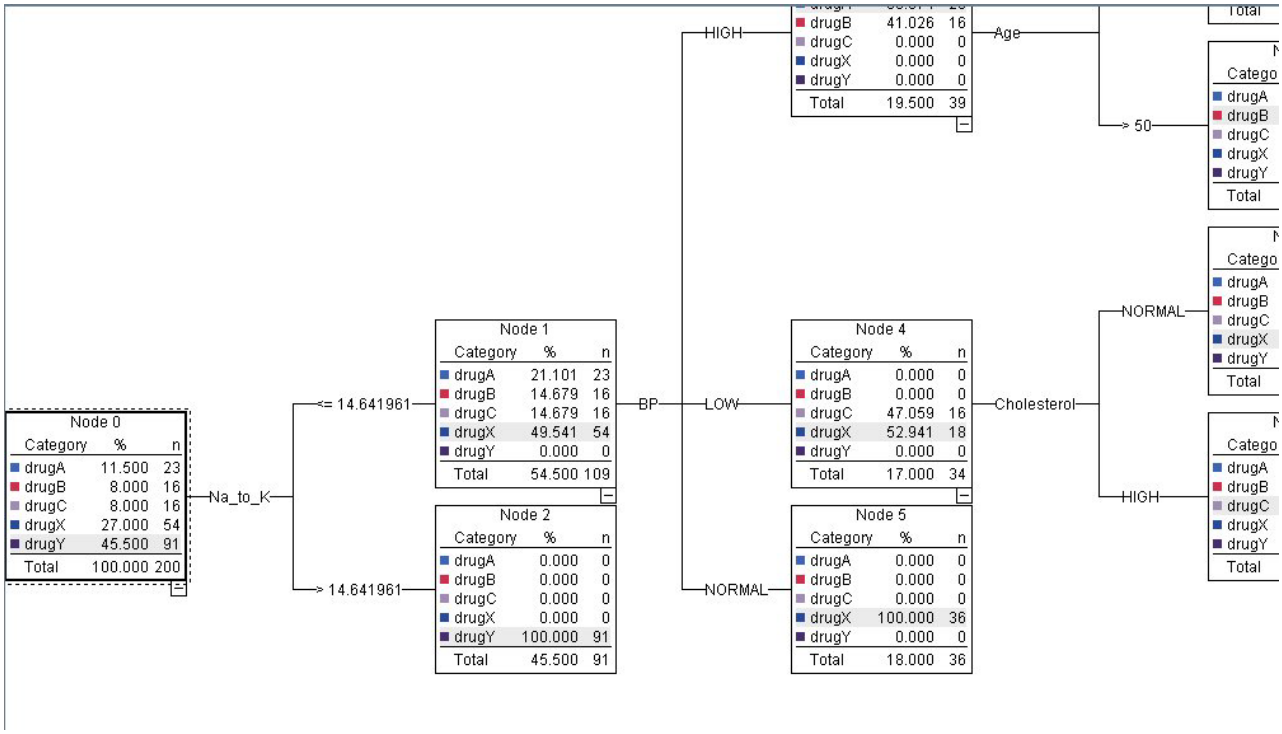


图 99. 图形形式的决策树

使用“分析”节点

可以使用“分析”节点来评估模型的准确性。将“分析”节点（从“输出”节点选用板）附加到模型块，打开该节点并单击**运行**。

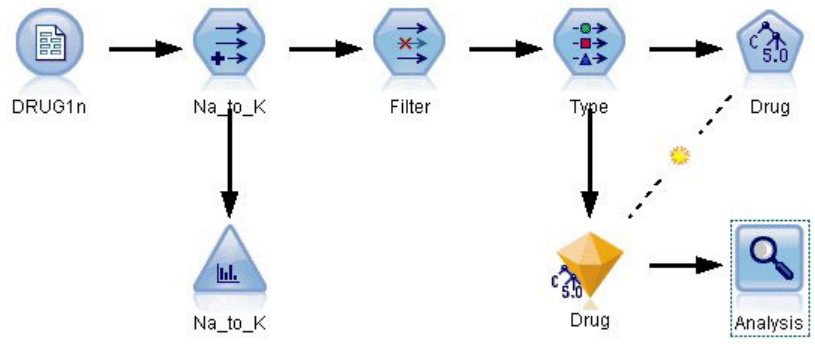


图 100. 添加“分析”节点

“分析”节点输出显示，通过此假设数据集，该模型已正确预测该数据集中每个记录的药物选择。在真正的数据集中，未必能做到完全准确，但分析节点可帮您确定模型的精确度能否满足特殊使用要求。

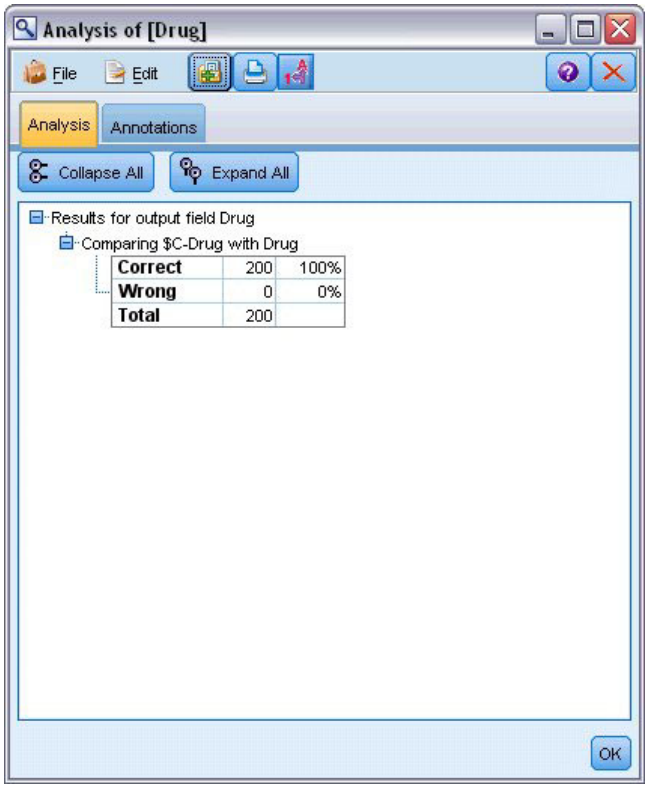


图 101. “分析”节点输出

第 9 章 筛选预测变量（特征选择）

“特征选择”节点有助于识别预测特定结果时最重要的字段。在包含成百乃至上千个预测变量的集合中，“特征选择”节点可以执行筛选和排序，并选出可能最重要的预测变量。最后，将生成一个速度更快且更高效的模型，此模型使用较少的预测变量、执行速度更快且更易于理解。

本示例中使用的数据表示某个虚构电话公司的数据仓库，并包含该公司 5,000 名客户对特殊促销活动的响应的相关信息。这些数据包含大量的字段，其中包括客户年龄、职业、收入、电话使用情况等统计信息。三个“目标”字段显示客户是否对这三个报价做出了响应。该公司希望使用这些数据来帮助预测哪些客户最可能在将来对类似报价做出响应。

此示例使用名为 *featureselection.str* 的流，此流引用名为 *customer_dbase.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *featureselection.str* 位于 *streams* 目录下。

本示例仅主要讲述其中一种促销活动，并将其作为目标。本示例使用 CHAID 树构建节点来开发模型，以描述最有可能对促销活动做出响应的客户。其中对以下两种方法作了对比:

- 不使用特征选择。数据集中的所有预测变量字段均可用作 CHAID 树的输入。
- 使用特征选择。使用“特征选择”节点来选择前 10 个预测变量。然后将它们输入 CHAID 树中。

通过比较两个生成的树模型，可以看到特征选择如何产生有效的结果。

构建流

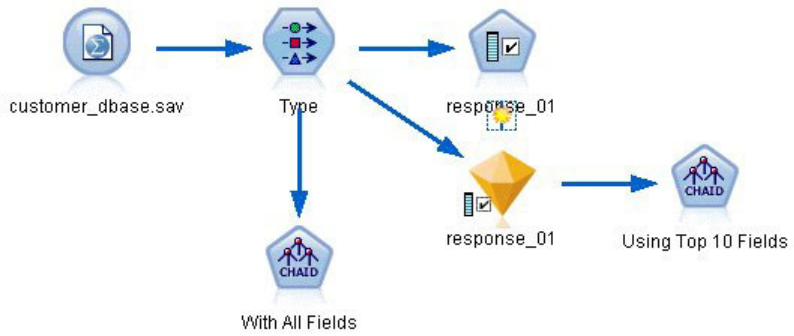


图 102. "特征选择"示例流

1. 将"Statistics 文件"源节点放入空白的流画布中。将此节点指向示例数据文件 *customer_dbase.sav*，该文件位于 IBM SPSS Modeler 安装程序的 *Demos* 目录下。（或者，可打开位于 *streams* 目录下的示例流文件 *featureselection.str*。）
2. 添加"类型"节点。在"类型"选项卡上，向下滚动到底部并将 *response_01* 的角色更改为*目标*。将其他响应字段（*response_02*）和（*response_03*）以及客户标识（列表顶部的 *custid*）的角色更改为无。将所有其他字段的角色设置为*输入*，并单击**读取值**按钮，然后单击**确定**。

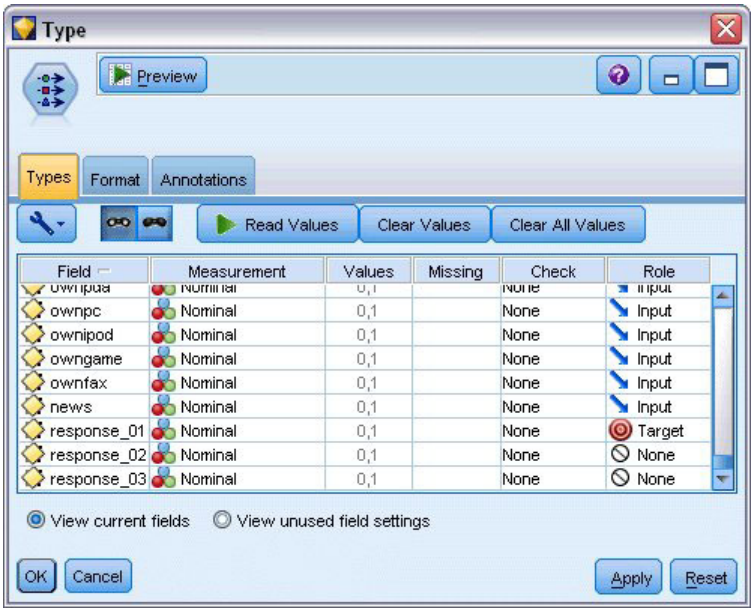


图 103. 添加"类型"节点

3. 将"特征选择"建模节点添加到流中。在此节点上，您可以指定要筛选的规则和标准，或要筛选的字段。
4. 运行流以创建"特征选择"模型块。
5. 右键单击流上或"模型"选用板中的模型块并选择**编辑**或**浏览**以查看结果。

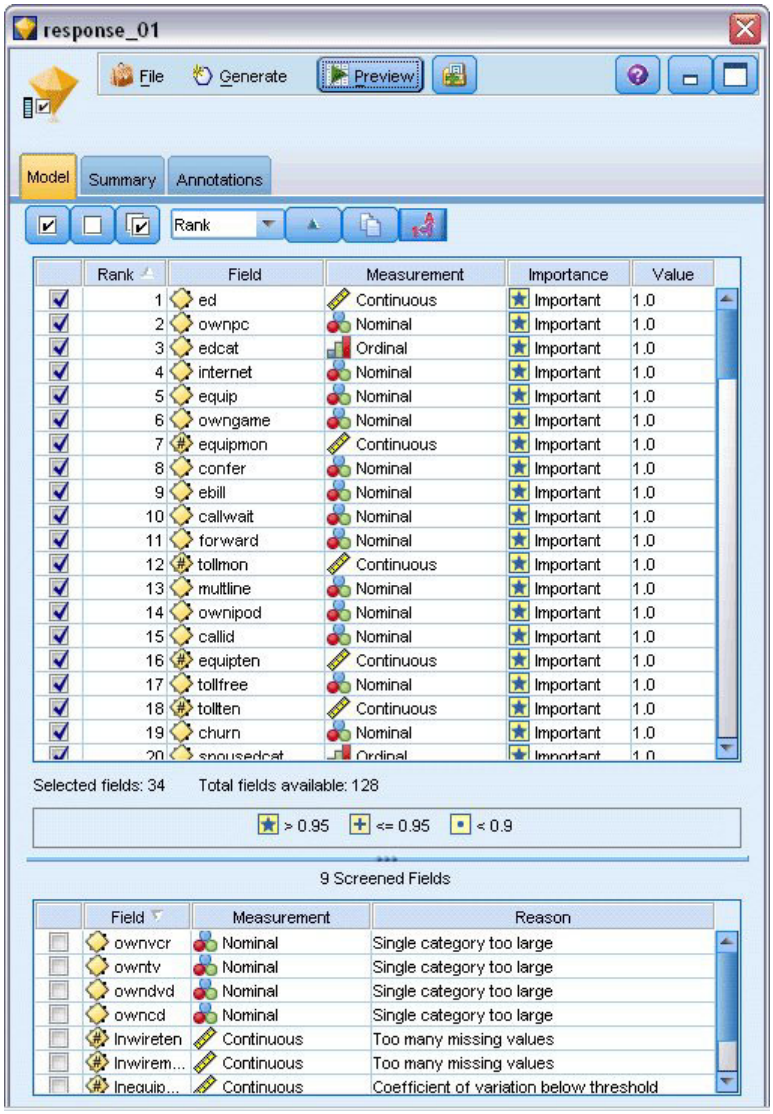


图 104. "特征选择"模型块中的"模型"选项卡

顶部面板显示了被认为对预测有用的字段。这些字段根据重要性进行排列。底部面板显示了从分析中筛选出来的字段及筛选的原因。通过检查顶部面板中的字段，可以确定在随后的建模会话中要使用哪些字段。

6. 现在，可以选择要在下游使用的字段。虽然最初已将 34 个字段识别为重要字段，但我们希望进一步精简预测变量集合。
7. 使用第一列中的勾选标记来取消选中不需要的预测变量，以便仅选中前 10 个预测变量。（单击行 11 中的选中标记，按住 Shift 键并单击行 34 中的选中标记。）关闭模型块。
8. 要在不使用特征选择的情况下比较结果，必须向流添加两个 CHAID 建模节点：一个使用特征选择，另一个不使用。
9. 将一个 CHAID 节点连接到"类型"节点，并将另一个节点连接到"特征选择"模型块。
10. 打开每个 CHAID 节点，选择"构建选项"选项卡，确保在"目标"窗格中选中了选项**构建新模型**、**构建单个树**和**启动交互会话**。

在"基本"窗格上，确保将**最大树深度**设置为 5。

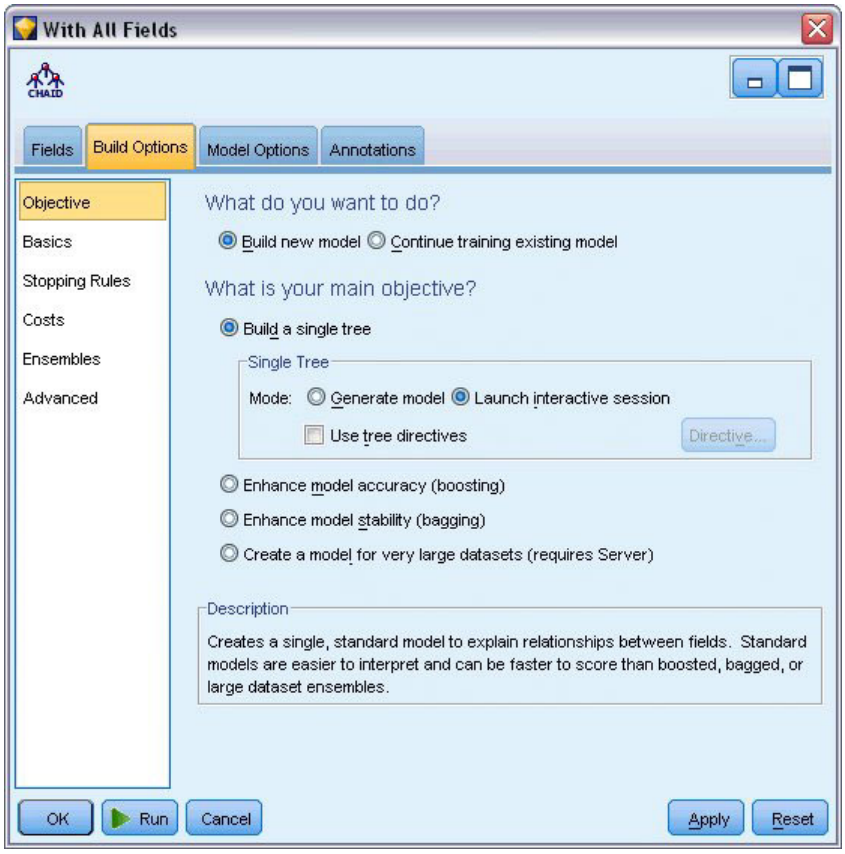


图 105. CHAID 建模节点针对所有预测变量字段的目标设置

构建模型

1. 执行使用数据集中所有预测变量的 CHAID 节点（即连接到“类型”节点的节点）。节点运行时，请注意其执行所用时间。表会显示在结果窗口中。
2. 从菜单中，选择 **树 > 生长树**，可生成并显示展开的树。

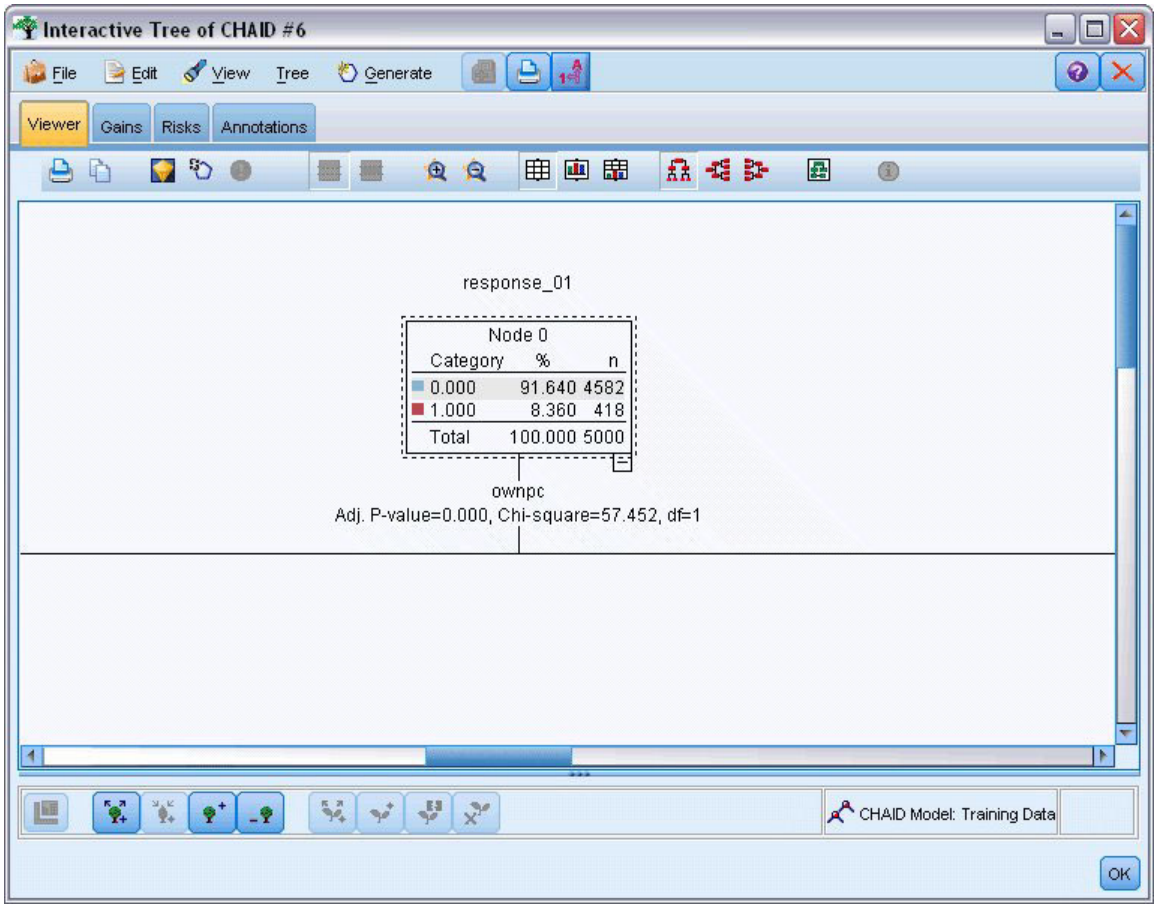


图 106. 在树构建器中生成树

3. 现在，对仅使用 10 个预测变量的另一个 CHAID 节点执行相同操作。打开“树构建器”后再次创建树。

第二个模型的执行速度应该比第一个模型快。由于此数据集非常小，因此执行时间上的差别可能只有几秒钟；但对于实际应用中更大的数据集，此差别可能非常明显（几分钟甚至几小时）。使用特征选择可以显著加快处理速度。

第二个树包含的树节点也少于第一个树。因此更易于理解。但在决定使用此模型之前，需要查明此模型是否有效，并查明其与使用所有预测变量的模型相比较的结果。

比较结果

要比较两个结果，需要进行有效性测量。为此，将使用树构建器中的“增益”选项卡。我们将查看 **提升** ，该图可测量节点中的记录与数据集中的所有记录相比时，其落入目标类别的可能性究竟提升多少。例如，提升值 148% 表示与数据集中的所有记录相比，节点中的记录落在目标类别的可能性是其 1.48 倍。提升值显示在“增益”选项卡的*指数*列中。

1. 在预测变量的完整集合的树构建器中，单击“增益”选项卡。将“目标类别”更改为 1.0。首先单击“分位数”工具栏按钮将显示更改为分位数。然后从此按钮右侧的下拉列表中选择 **四分位数** 。
2. 在具有 10 个预测变量集合的树构建器中重复此步骤，就可以对两个类似的增益表进行比较，如下图中所示。

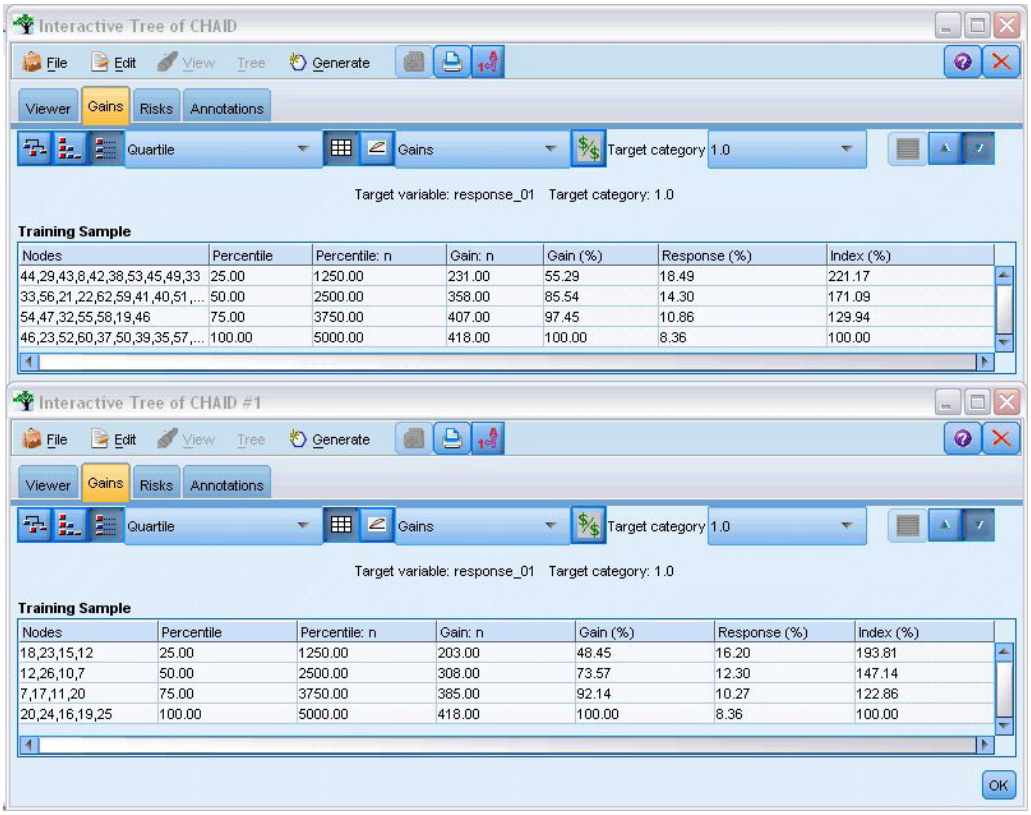


图 107. 两个 CHAID 模型的增益图

每个增益表都将其树的终端节点分组为分位数。要比较两个模型的有效性，可查看每个表中 25% 分位数的提升（ *指数* 值）。

包括所有预测变量时，模型显示提升值 221%。即，具有这些节点中的特征的个案，其响应目标促销活动的可能性是其他个案的 2.2 倍。要查看这些特征的内容，请单击以选择最上面的一行。然后切换到“查看器”选项卡，其中相应的节点现在以黑色框突出显示。沿着树向下寻找每个突出显示的终端节点，以了解这些预测变量的分割方式。第一个分位数仅包含 10 个节点。如果转换为实际应用中的评分模型，那么 10 个不同的客户特征难以进行管理。

如果仅包括前 10 个预测变量（由特征选择识别），则提升值接近为 194%。虽然此模型不如使用所有预测变量的模型那样有效，但它无疑也是有用的。此处第一个分位数仅包含四个节点，因此它更简单。因此，我们可以确定特征选择模型比使用所有预测变量的模型更合适。

摘要

我们来看一下特征选择的优势。使用较少的预测变量会降低成本。这意味着您要收集、处理和输入模型的数据减少。并且节省了计算时间。在本示例中，即使有额外的特征选择步骤，模型构建的速度也明显提高，因为使用了较小的预测变量集合。如果使用较大的实际数据集，则节省的时间应大大增加。

使用较少的预测变量会使评分更加简单。如本示例所示，您可以只识别有可能响应促销活动的客户的 4 个特征。请注意，预测变量越多，过度拟合模型的风险越大。生成的模型越简单，则对其他数据集会越有利（尽管可能需要通过测试确定该模型）。

您可能已使用树构建算法来进行特征选择，这将使树可以识别对您最重要的预测变量。实际上，CHAID 算法经常用于完成此操作，并且使用此算法甚至可以逐层创建树以控制树的深度和复杂性。但是，使用“特征选择”节

点更快且更简单。此节点通过单一步骤快速对所有预测变量进行排序，使您可以迅速识别最重要的字段。使用此节点还可以更改要包括的预测变量数。可以使用前 15 或 20 个而不是前 10 个预测变量再次轻松运行此示例，并比较其结果以确定最佳模型。

第 10 章 减少输入数据字符串长度（重新分类节点）

减少输入数据字符串长度（重新分类）

对于二项 logistic 回归模型和包含二项 logistic 回归模型的自动分类器模型，字符串字段被限制为最多不得超过八个字符。如果字符串超过八个字符，则可以使用重新分类节点对其重新编码。

本示例使用名为 *reclassify_strings.str* 的流，该流所引用的数据文件名为 *drug_long_name* 。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *reclassify_strings.str* 位于 *streams* 目录下。

本示例主要讲述流中的一小部分，以显示可能因字符串过长而生成的某种错误；并且解释了如何使用“重新分类”节点将字符串的详细信息更改为可接受的长度。尽管示例中使用了二项 Logistic 回归节点，但这相当于使用自动分类器节点来生成二项 Logistic 回归模型。

重新分类数据

1. 使用一个变量文件源节点，连接到 *Demos* 文件夹下的数据集 *drug_long_name* 。

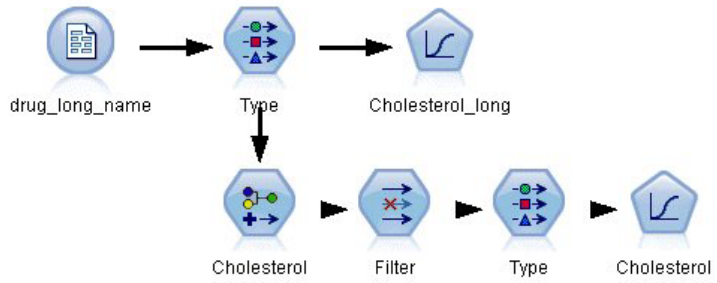


图 108. 显示对二项 logistic 回归的字符串重新分类的样本流

2. 将类型节点添加至源节点，然后选择 **Cholesterol_long** 作为目标。
3. 将 Logistic 回归节点添加到类型节点中。
4. 在 Logistic 回归节点上，单击“模型”选项卡并选择**二项**过程。

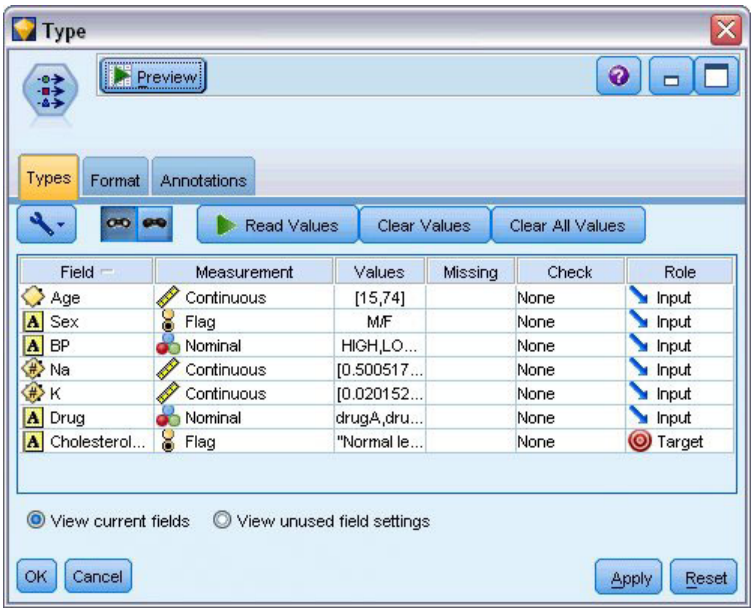


图 109. "Cholesterol_long"字段中的长字符串详细信息

5. 在 *reclassify_strings.str* 中执行 Logistic 回归节点时，会显示一个错误消息，警告您 **Cholesterol_long** 字符串值过长。

   如果遇到此类型的错误消息，请按照本示例其他部分所说明的步骤来修改数据。

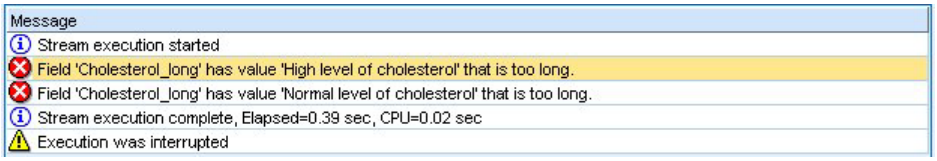


图 110. 执行二项 logistic 回归节点时所显示的错误消息

6. 为类型节点添加一个重新分类节点。
7. 在"重新分类"字段中，选择 **Cholesterol_long**。
8. 键入 **Cholesterol** 作为新的字段名称。
9. 单击 **获取** 按钮，将 **Cholesterol_long** 值添加至原始值列。
10. 在新的值列中，在 **高胆固醇水平** 原始值旁边，键入 **高** ，在 **正常胆固醇水平** 原始值的旁边，键入 **正常** 。

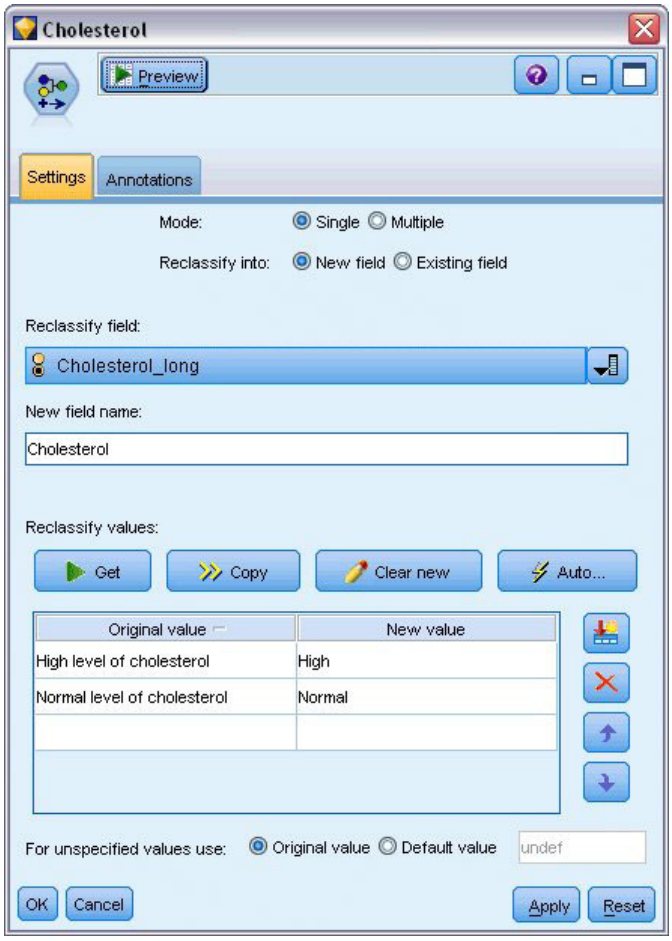


图 111. 对长字符串进行重新分类

11. 为重新分类节点添加一个过滤节点。
12. 在“过滤”列中，单击以删除 **Cholesterol_long**。

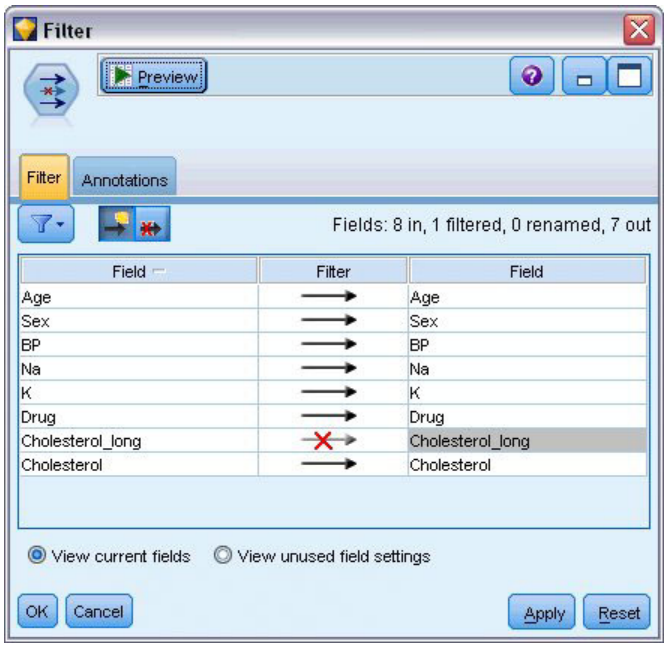


图 112. 过滤数据中的“Cholesterol_long”字段

13. 将类型节点添加至过滤节点并选择 **Cholesterol** 作为目标。

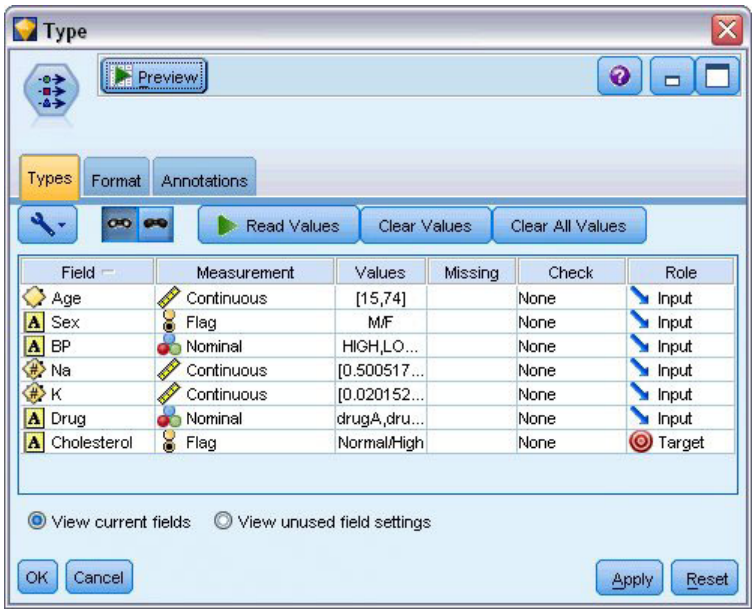


图 113. “Cholesterol”字段中的短字符串详细信息

14. 将一个 Logistic 节点添加到类型节点中。
15. 在 Logistic 节点上，单击“模型”选项卡并选择**二项**过程。
16. 您现在可以执行二项 Logistic 节点，并生成一个不会显示错误消息的模型。

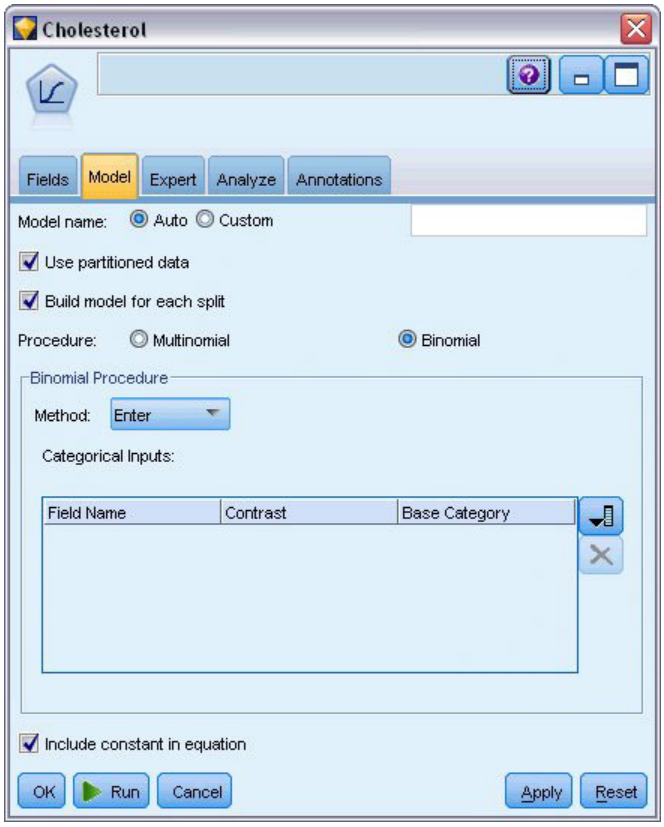


图 114. 选择"二项"作为过程

此示例仅显示了流的一部分。如果您需要更多有关该类型流（可能需要重新分类长字符串）的信息，请参见以下示例:

- "自动分类器"节点。请参阅主题第 33 页的『对客户响应建模（自动分类器）』以获取更多信息。
- "二项 Logistic 回归"节点。请参阅主题第 135 页的第 13 章，『电信客户流失（二项 Logistic 回归）』以获取更多信息。

有关如何使用 IBM SPSS Modeler（如《用户指南》、《节点参考》和《算法指南》）的详细信息，可从安装光盘的 *\Documentation* 目录下找到。

第 11 章 对客户响应建模（决策列表）

决策列表算法可以生成表示给定的二元结果（是/否）的可能性上限和下限的规则。决策列表模型广泛用于客户关系管理，例如客户服务中心或市场营销应用程序。

本示例基于某个虚构的公司，该公司希望为每个客户提供适合的报价，以便在未来的市场竞销活动中实现更高收益。特别地，该示例根据先前的促销活动使用决策列表模型来识别积极响应的客户的特征，并根据识别结果生成邮件发送列表。

决策列表模型尤其适用于交互建模，这将允许您调整模型中的参数并立即看到结果。对于允许您自动创建多个不同模型并对结果进行排序的其他方法，可以改用自动分类器节点。

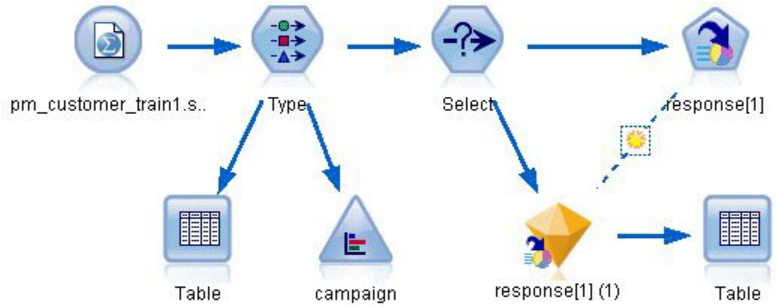
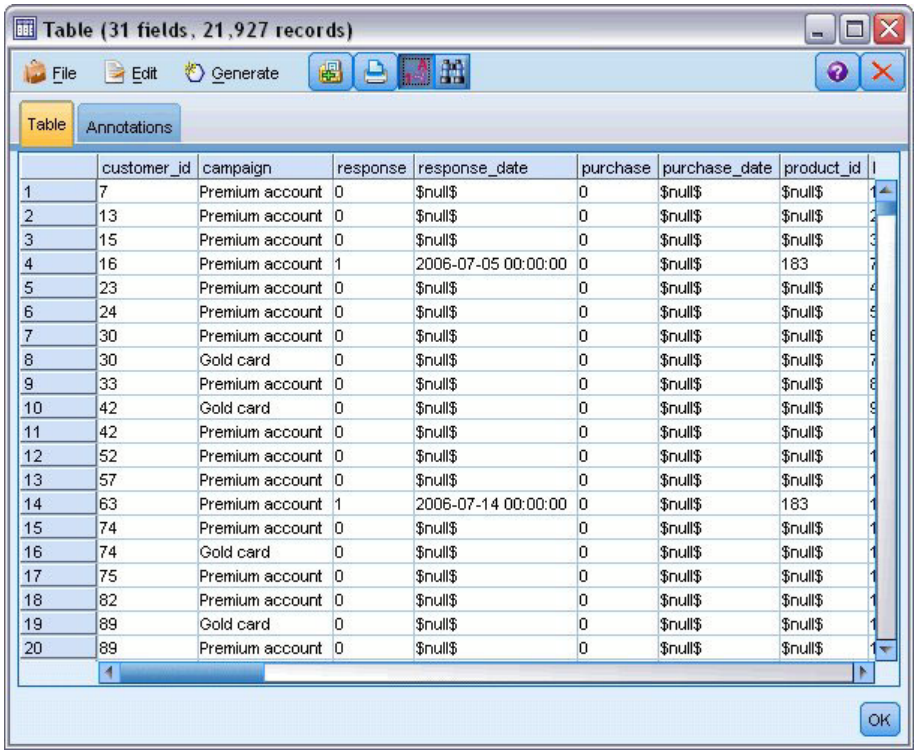


图 115. 决策列表样本流

本示例使用流 *pm_decisionlist.str* ，该流引用数据文件 *pm_customer_train1.sav* 。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *pm_decisionlist.str* 位于 *streams* 目录下。

历史数据

文件 *pm_customer_train1.sav* 的历史数据可跟踪过去的营销活动中为特定客户提供的报价，由 *campaign* 字段的值表示。*Premium account* 活动中的记录数最大。



	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 card	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold card	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 card	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 card	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

图 116. 先前促销活动的相关数据

campaign 字段的值在数据中实际编码为整数，并带有类型节点中定义的标签（例如，*2 = Premium account*）。可以使用工具栏切换表中值标签的显示。

该文件还包括若干包含每位客户的相关人口统计和金融信息的字段，这些字段可用于构建或“训练”依据特定特征针对不同组预测响应率的模型。

构建流

1. 添加指向 *pm_customer_train1.sav* 的 Statistics 文件节点，该文件位于 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中。（您可以在文件路径中指定 `$CLEO_DEMOS/` 作为引用此文件夹的快捷方式。）

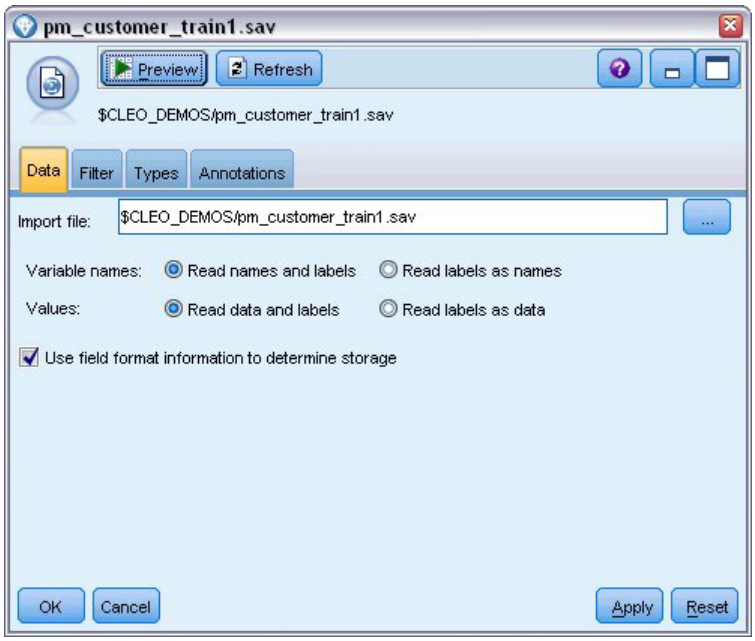


图 117. 读入数据

2. 添加类型节点，然后选择*响应*作为目标字段（“角色”为**目标**）。将此字段的测量级别设置为**标志**。

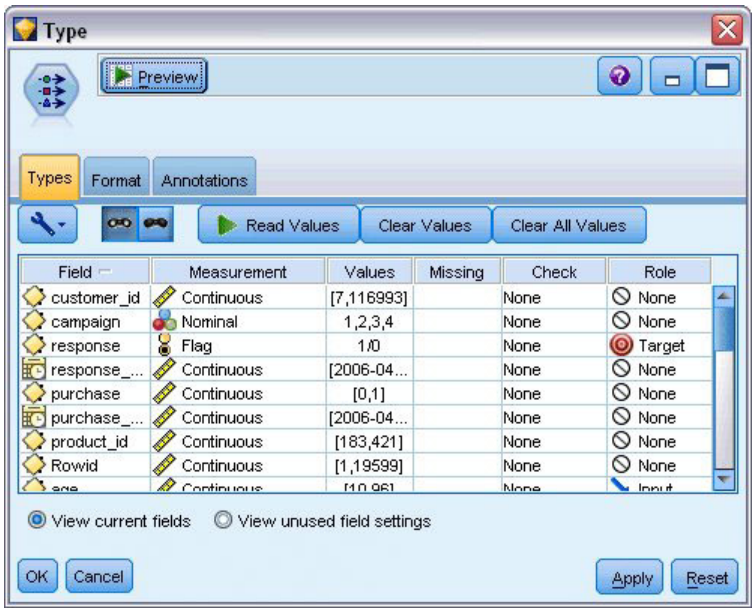


图 118. 设置测量级别和角色

3. 对于下列字段，将角色设置为**无**：*customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid* 和 *X_random*。这些字段在数据中均有用途，但不会在实际模型的构建中使用。
4. 单击类型节点的 **读取值** 按钮以确保值获得实例化。

尽管数据包含有关四项不同活动的信息，但每一次的分析应侧重于其中一项活动。由于 Premium 活动（在数据中编码为 *campaign=2* ）中的记录数最大，因此可以使用选择节点实现仅在流中包含这些记录。

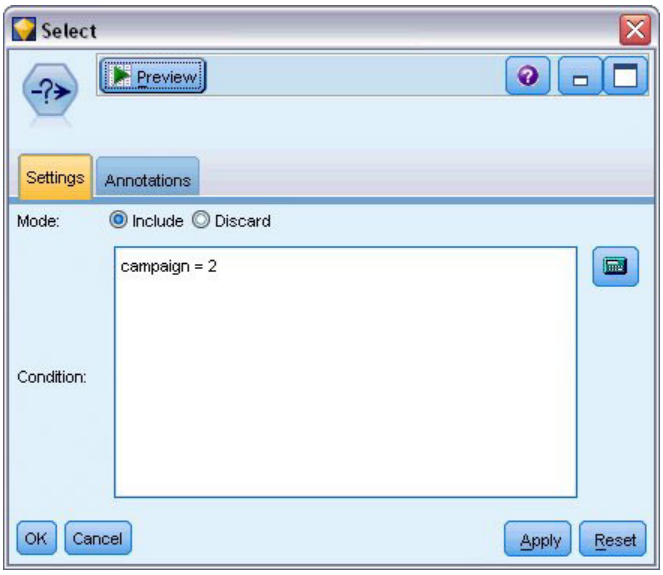


图 119. 为单项活动选择记录

创建模型

1. 在流中添加决策列表节点。在“模型”选项卡中，将**目标值**设为 1 以表示要搜索的结果。在这种情况下，您正在搜索对以前的报价发出 是 响应的客户。

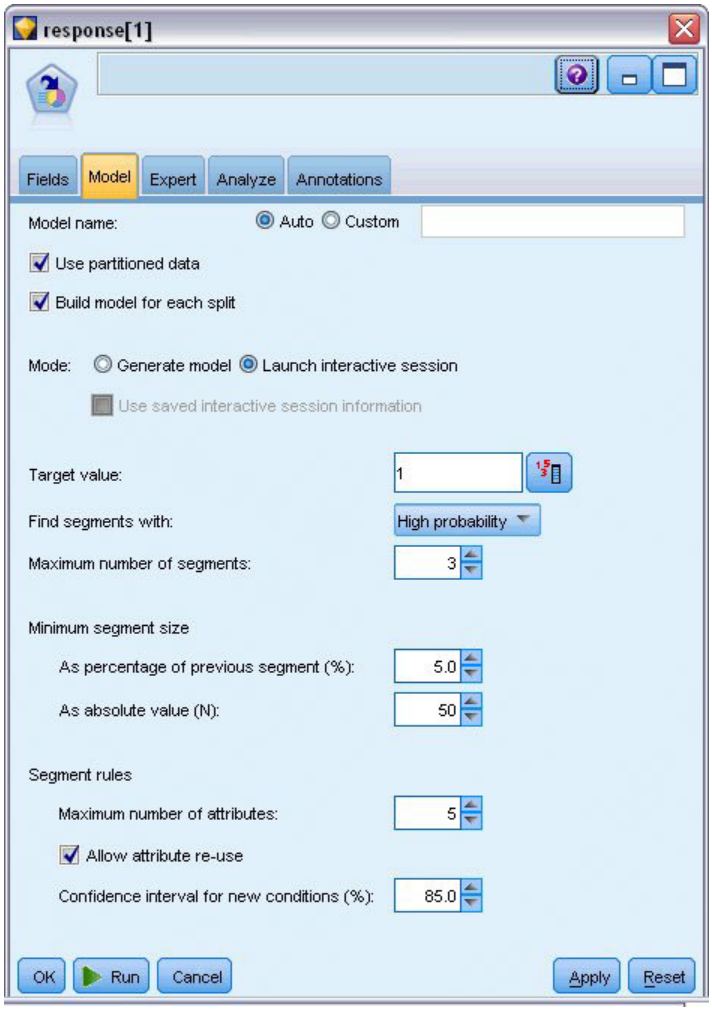


图 120. "决策列表"节点，"模型"选项卡

2. 选择**启动交互会话**。
3. 为简化本例中的模型，请将最大段数设为 3。
4. 将新条件的置信度区间更改为 85%。
5. 在"专家"选项卡上，将**模式**设置为**专家**。

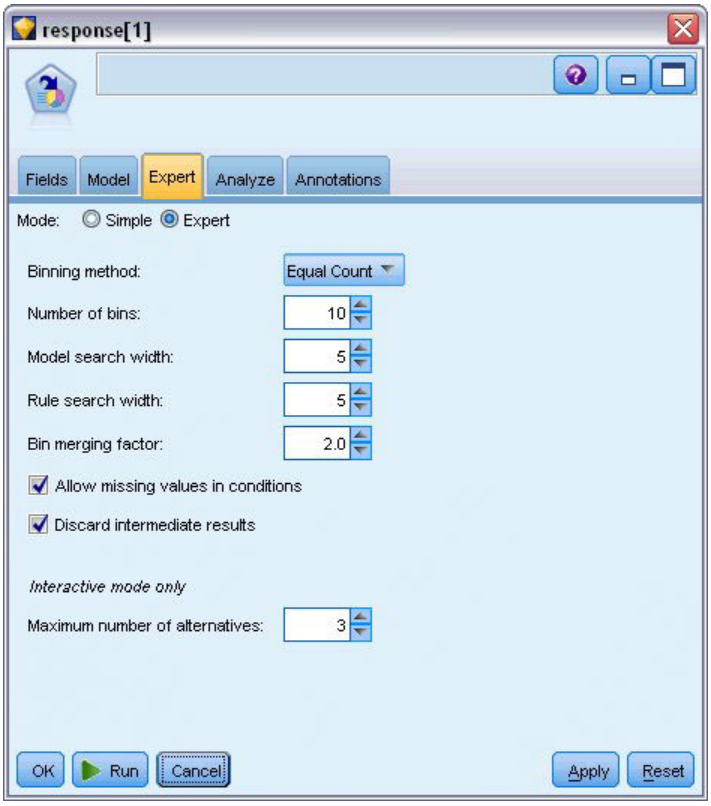


图 121. "决策列表"节点，"专家"选项卡

6. 将**最大替代值数**设为 3。此选项与在"模型"选项卡上所选的**启动交互会话**设置一起使用。
7. 单击**运行**显示"交互列表"查看器。

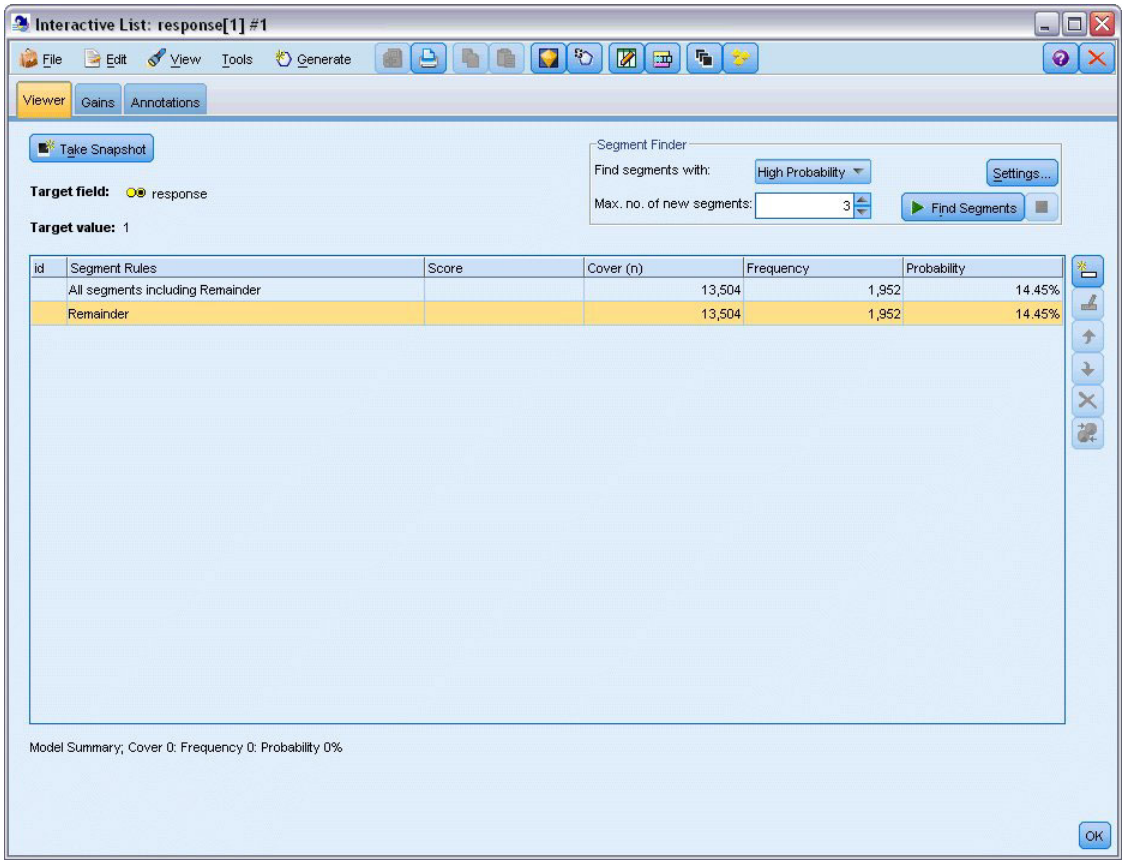


图 122. 交互列表查看器

尚未定义任何段，因此所有记录都位于其余段中。在示例的 13,504 条记录中，有 1,952 条记录的响应为是，总匹配率为 14.45%。需要通过识别更可能（或不太可能）作出积极响应的客户段来提高此匹配率。

8. 在交互列表查看器中，从菜单中选择以下选项:

   工具 > 查找段

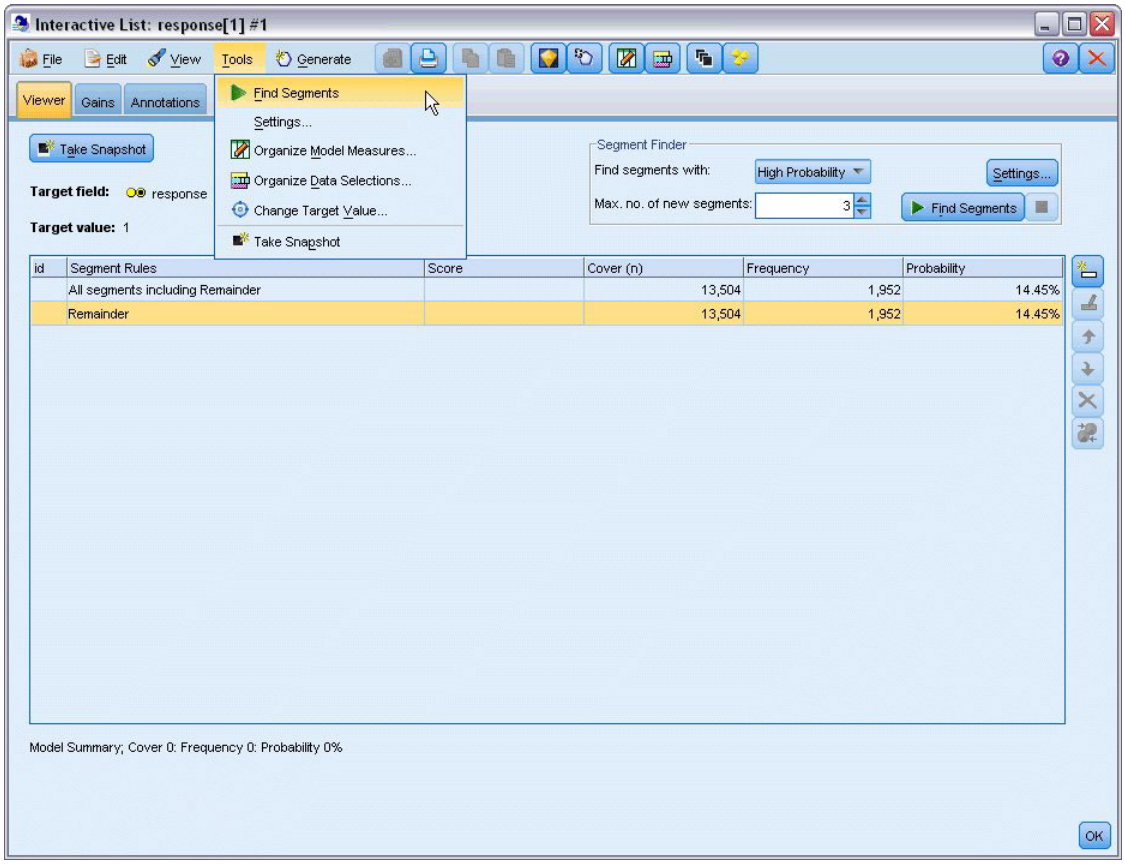


图 123. 交互列表查看器

此操作将根据“决策列表”节点中指定的设置来运行缺省挖掘任务。完成的任务将返回三个替代模型，这三个模型在“模型作品集”对话框的“替代”选项卡中列出。

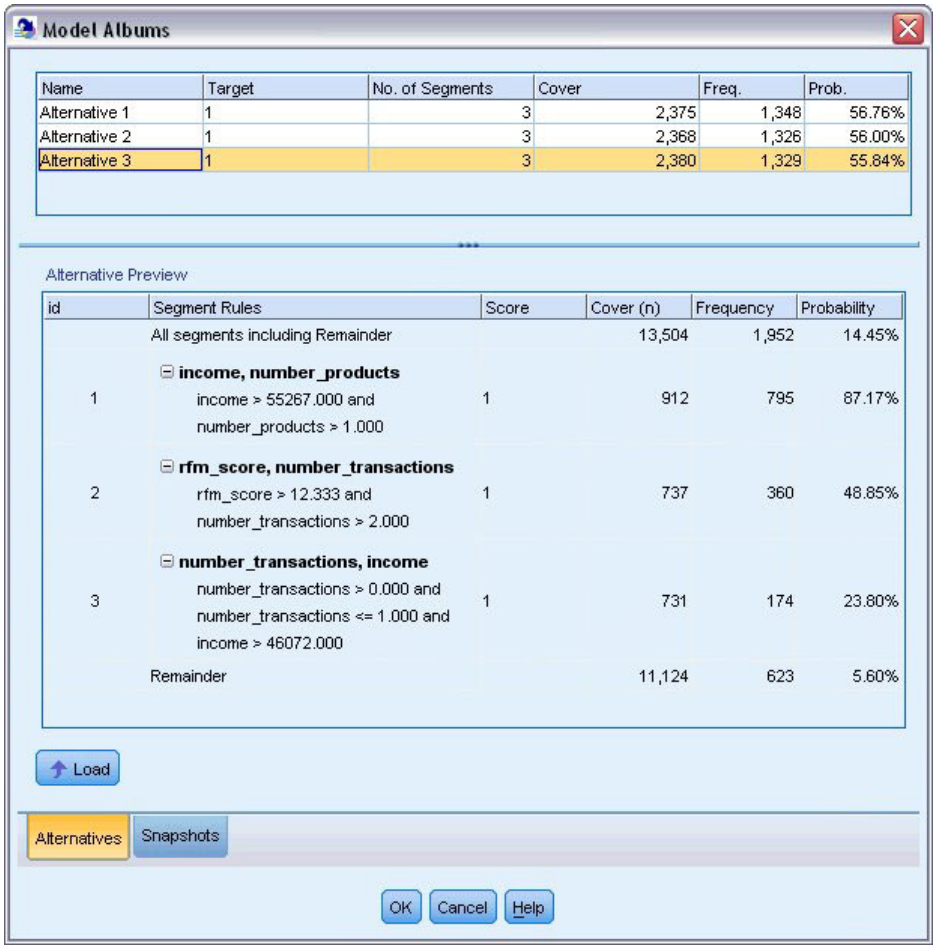


图 124. 可用替代模型

9. 从列表中选择第一个替代模型；其详细信息显示在“替代预览”窗格中。

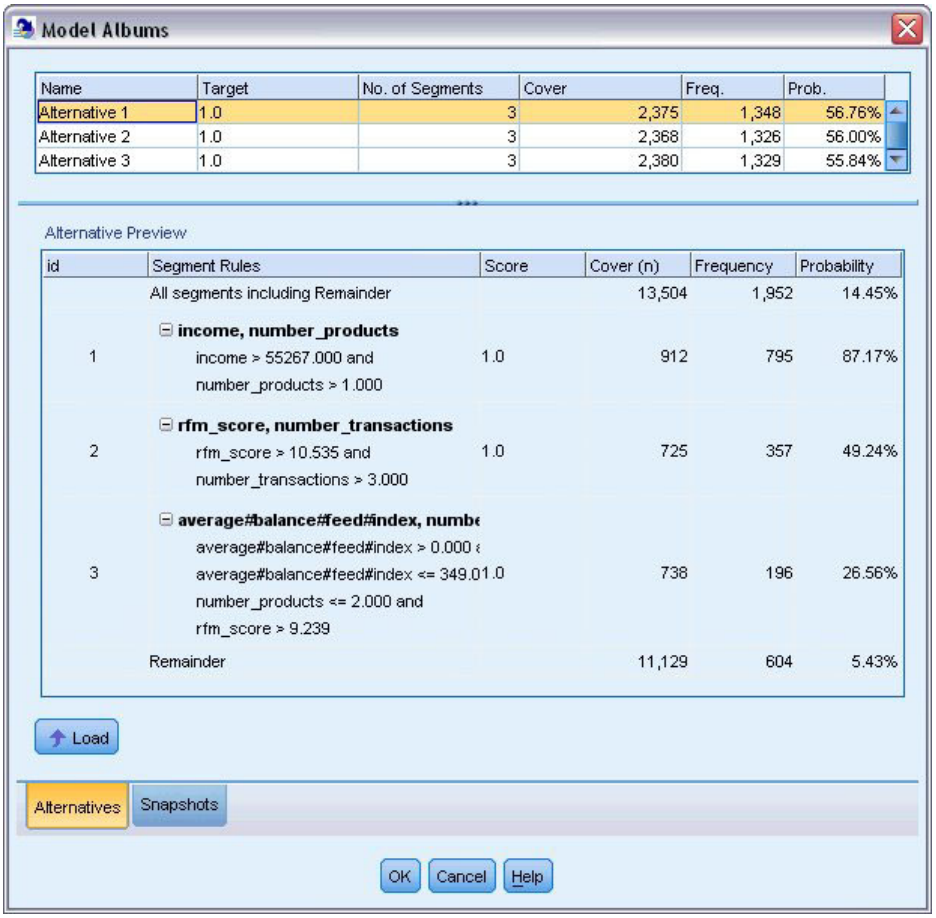


图 125. 所选替代模型

通过“替代预览”窗格您可以在不更改工作模型的情况下快速浏览任意数量的替代模型，从而简化尝试不同方法的过程。

注: 为了更好地查看模型，可能需要将对话框中的“替代预览”窗格最大化，如下所示。您可以通过拖动窗格边框来进行此操作。

通过使用基于预测变量（例如收入、每月事务数和 RFM 评分）的规则，模型可以识别响应率高于总体样本响应率的段。组合段后，该模型会提示您可以将匹配率提高至 56.76%。但该模型只占总体样本的一小部分，还有超过 11,000 条记录（其中有几百条匹配记录）位于其余段中。您希望模型在排除低响应率段的同时捕获更多的匹配记录。

10. 要尝试使用不同的建模方法，可从菜单中选择以下项:

工具 > 设置

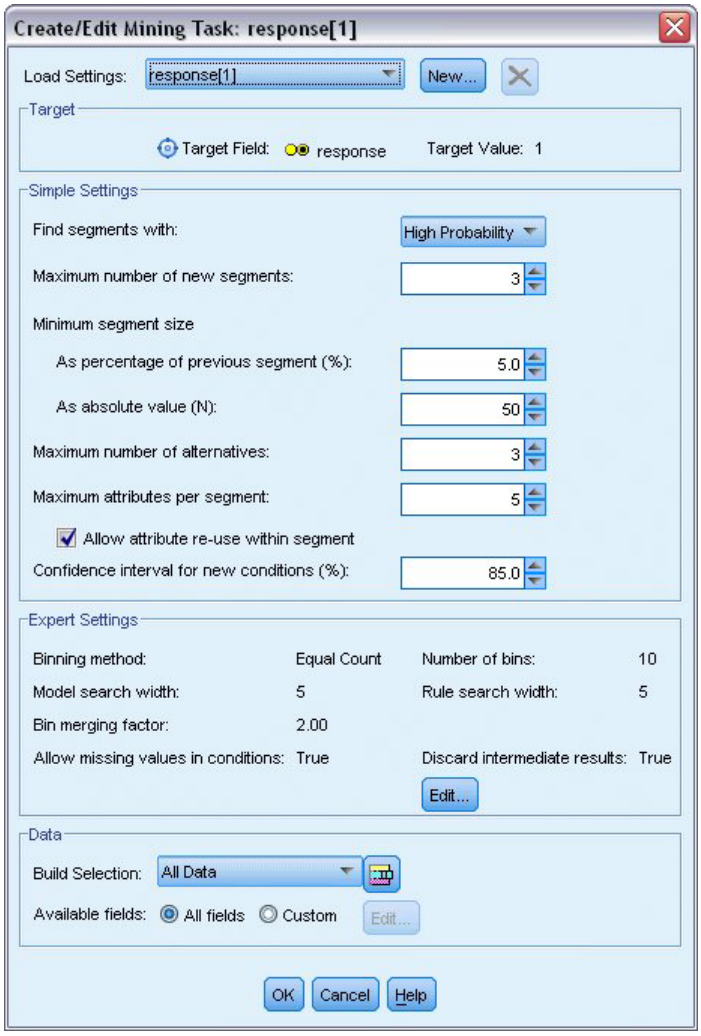


图 126. "创建/编辑挖掘任务"对话框

11. 单击**新建**按钮（右上角）以创建第二个挖掘任务，并指定*向下搜索*作为"新建设置"对话框中的任务名称。

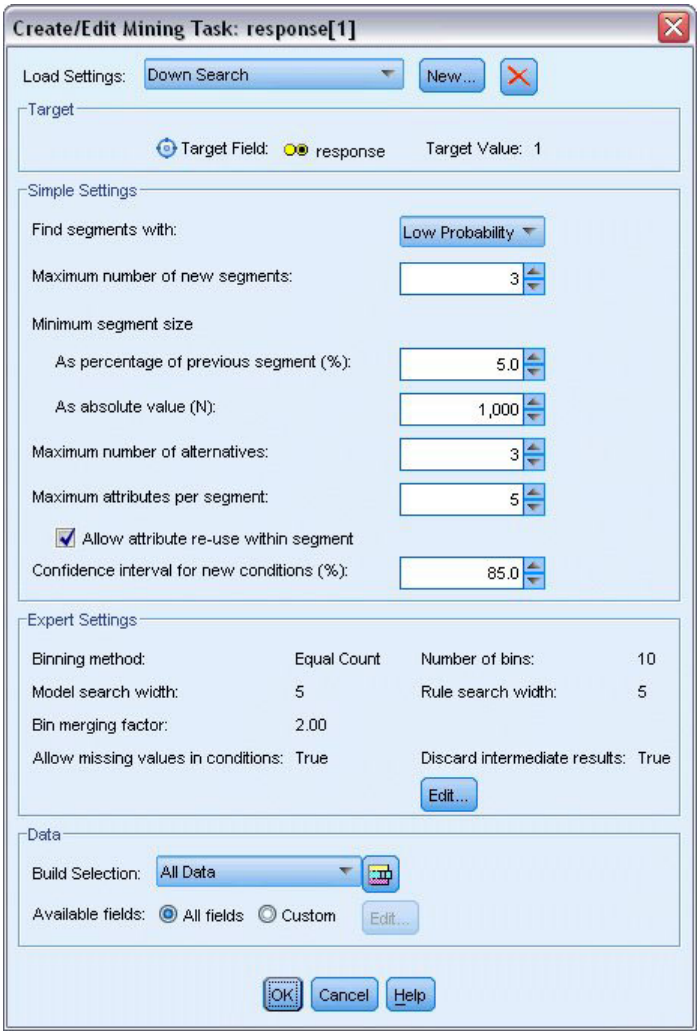


图 127. "创建/编辑挖掘任务"对话框

12. 将任务的搜索方向更改为**低概率**。此操作将使算法搜索具有 *最低* （而不是最高）响应率的段。
13. 将最小段大小增至 1,000。单击**确定**返回到"交互列表"查看器。
14. 在"交互列表"查看器中，确保*段查找器*窗格正在显示新任务详细信息并单击**查找段**。

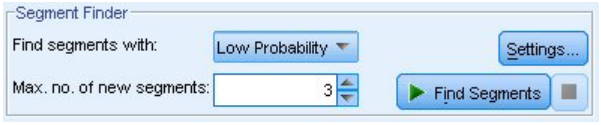


图 128. 在新挖掘任务中查找段

该任务返回一组新的替代模型，这些模型显示在"模型作品集"对话框的"替代"选项卡中，并且预览方法与先前的结果相同。

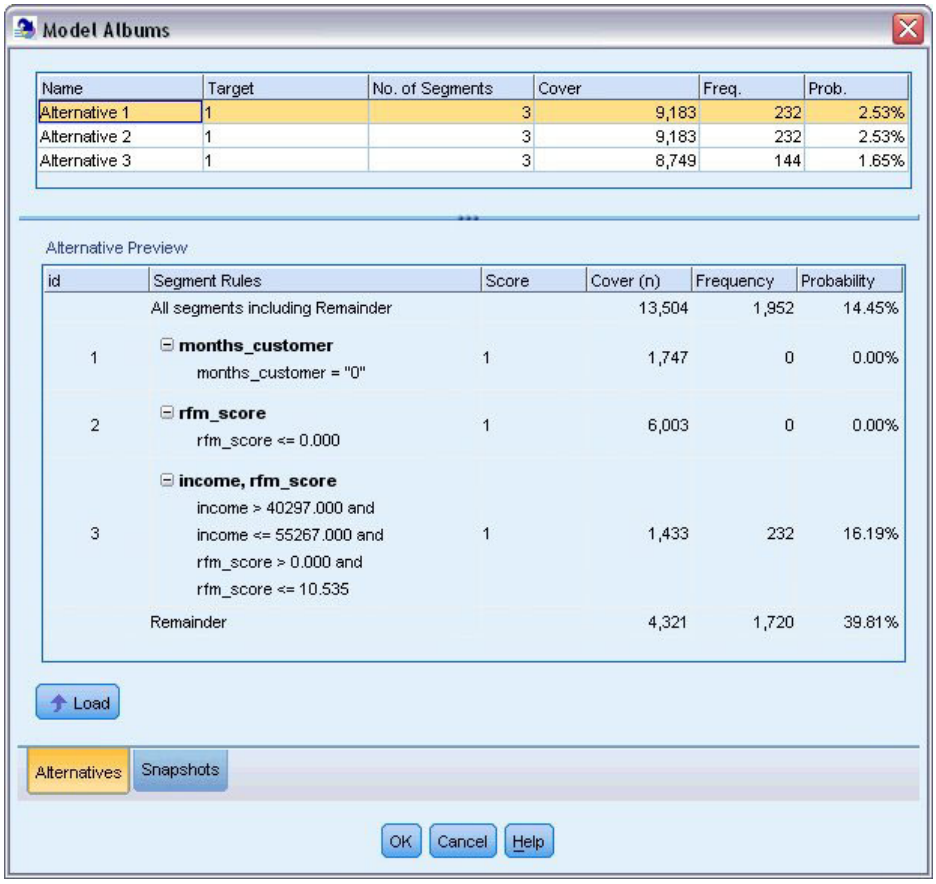


图 129. 向下搜索模型结果

这一次，每个模型都可识别具有低响应率而不是高响应率的段。浏览第一个替代模型，只需排除这些低响应率段就可以将其余段中的匹配率提高到 39.81%。此值低于前面浏览的模型的值，但其覆盖率更高（总匹配率更高）。

将这两种方法组合使用：使用“低概率”搜索剔除不感兴趣的记录，然后使用“高概率”搜索可以改进此结果。

15. 单击**加载**以使其（第一个向下搜索替代模型）成为工作模型并单击**确定**以关闭“模型作品集”对话框。

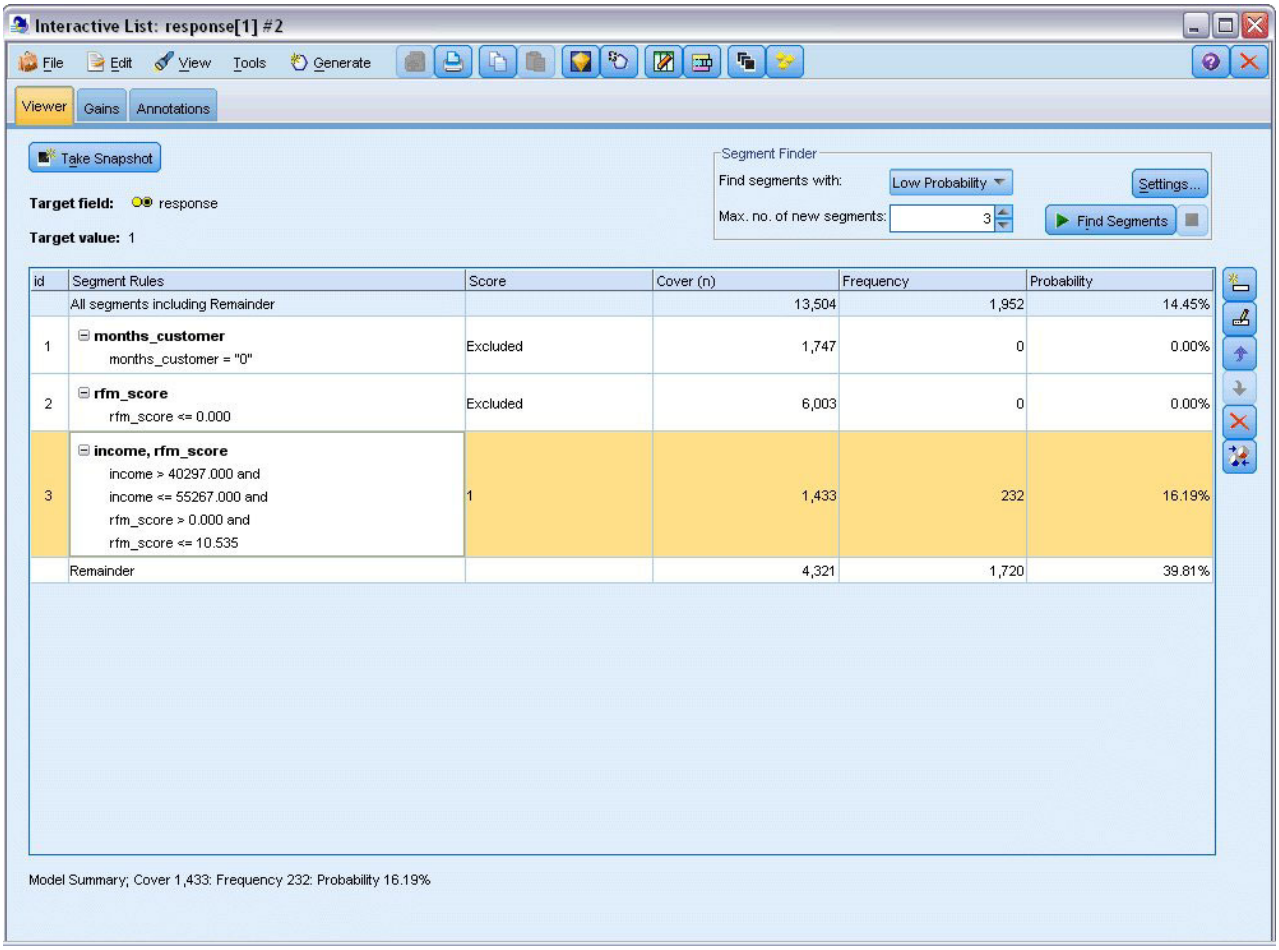


图 130. 排除段

16. 分别右键单击前两个段，然后选择 **排除段** 。这些段共同捕获将近 8,000 条记录，但之间的匹配率为零，因此需要将它们从未来的报价中排除。（排除的段的评分将为空，并以此来表示这些段。）
17. 右键单击第三个段并选择 **删除段** 。此段的匹配率 16.19% 与基准匹配率 14.45% 的差别不是很大，因此不足以证明应将其保留。

    注：删除段与排除段是两种不同的操作。排除段只是更改其评分方式，而删除段会将段从模型中彻底删除。

    排除最低响应率段后，便可以在剩余段中搜索高响应率段。
18. 单击表中的其余行以将其选中，使下一项挖掘任务仅应用于该行。

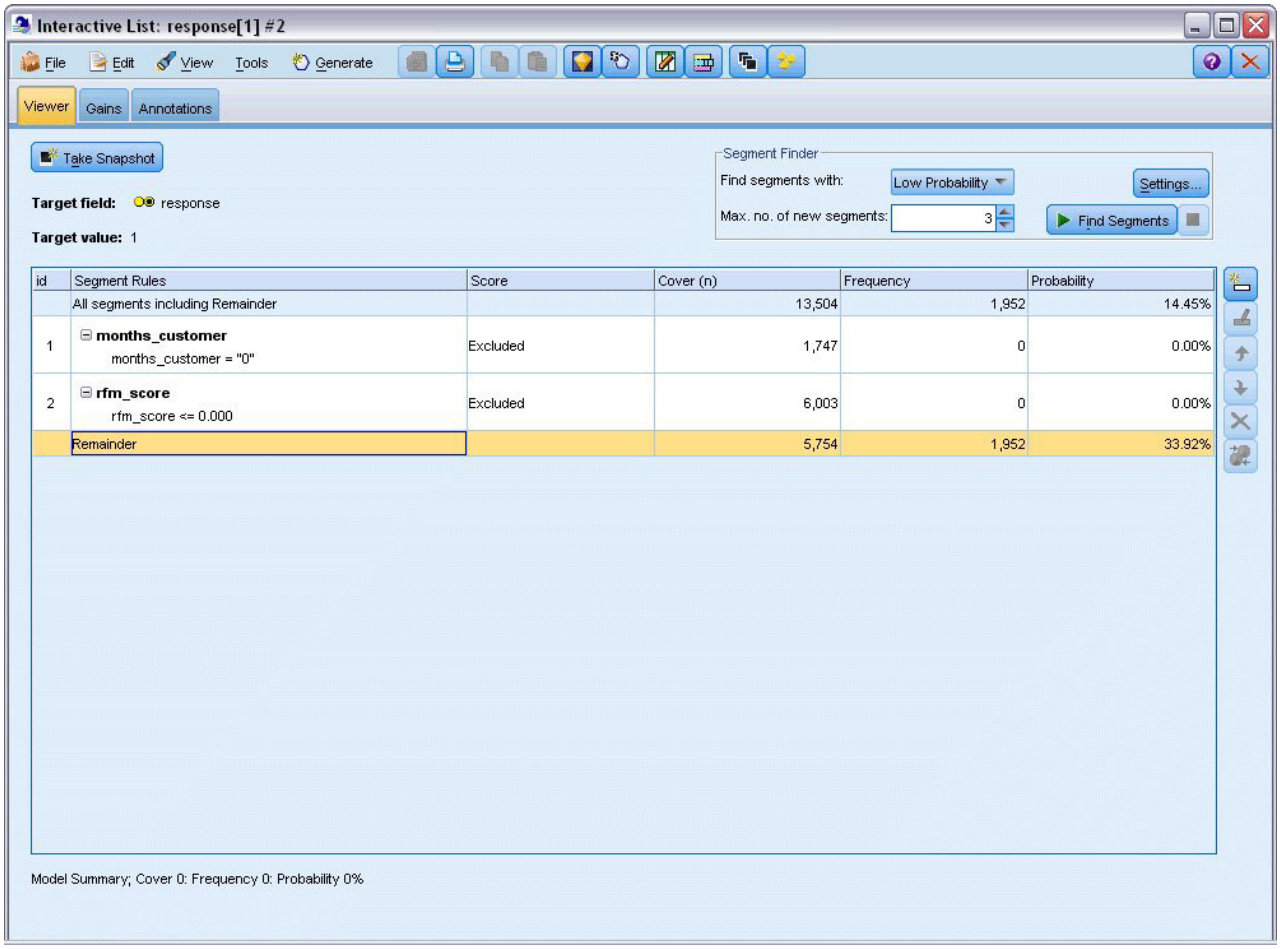


图 131. 选择段

19. 使用所选剩余模型，单击**设置**以重新打开“创建/编辑挖掘任务”对话框。
20. 在顶部的**装入设置**中，选择缺省挖掘任务 **response[1]**。
21. 编辑**简单设置**以将新段数增加到 5 并将最小段大小增加到 500。
22. 单击**确定**返回到“交互列表”查看器。

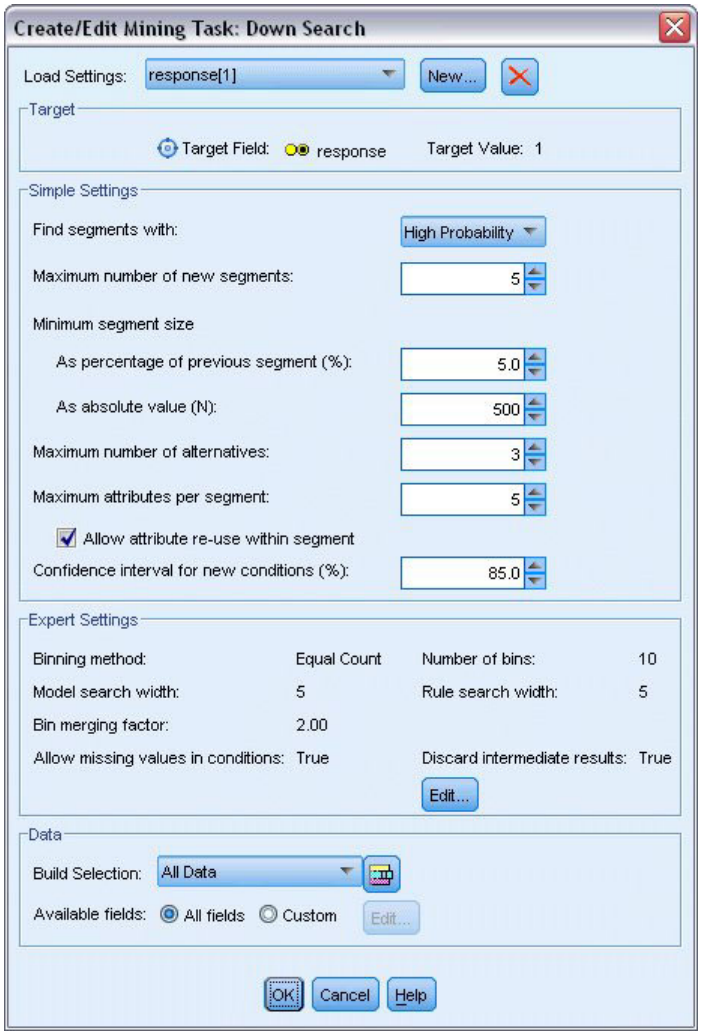


图 132. 选择缺省挖掘任务

23. 单击**查找段**。

这将显示另一组替代模型。通过将一项挖掘任务的结果反馈给另一项挖掘任务，这些最新模型将同时包含高响应率段和低响应率段。具有低响应率的段将被排除，这意味着其评分为空，而包含的段评分将为 1。总体统计量反映了上述排除结果，其中第一个替代模型具有 45.63% 的匹配率，覆盖率（3,456 条记录中有 1,577 条匹配记录）高于先前任何一个模型。

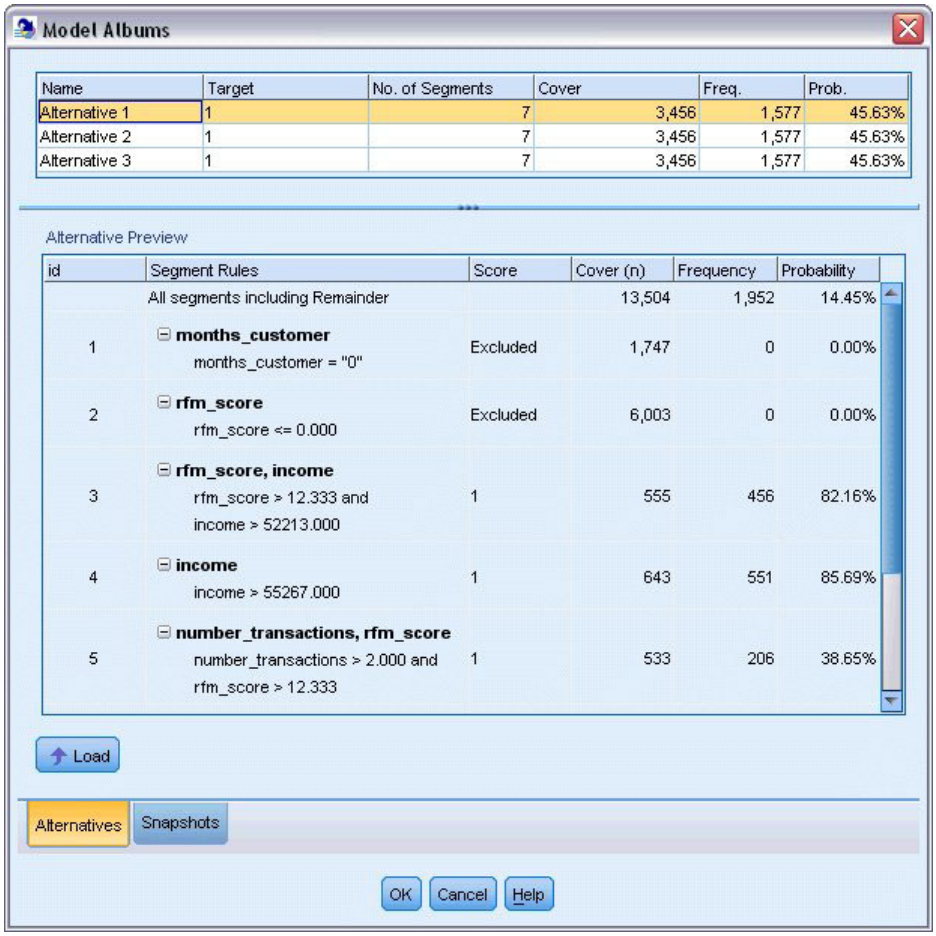


图 133. 组合模型的替代模型

24. 预览首个替代模型然后单击**加载**以使其成为工作模型。

使用 Excel 计算自定义测量量

1. 为更清楚地了解模型的实际性能，可以在“工具”菜单中选择**组织模型测量**。

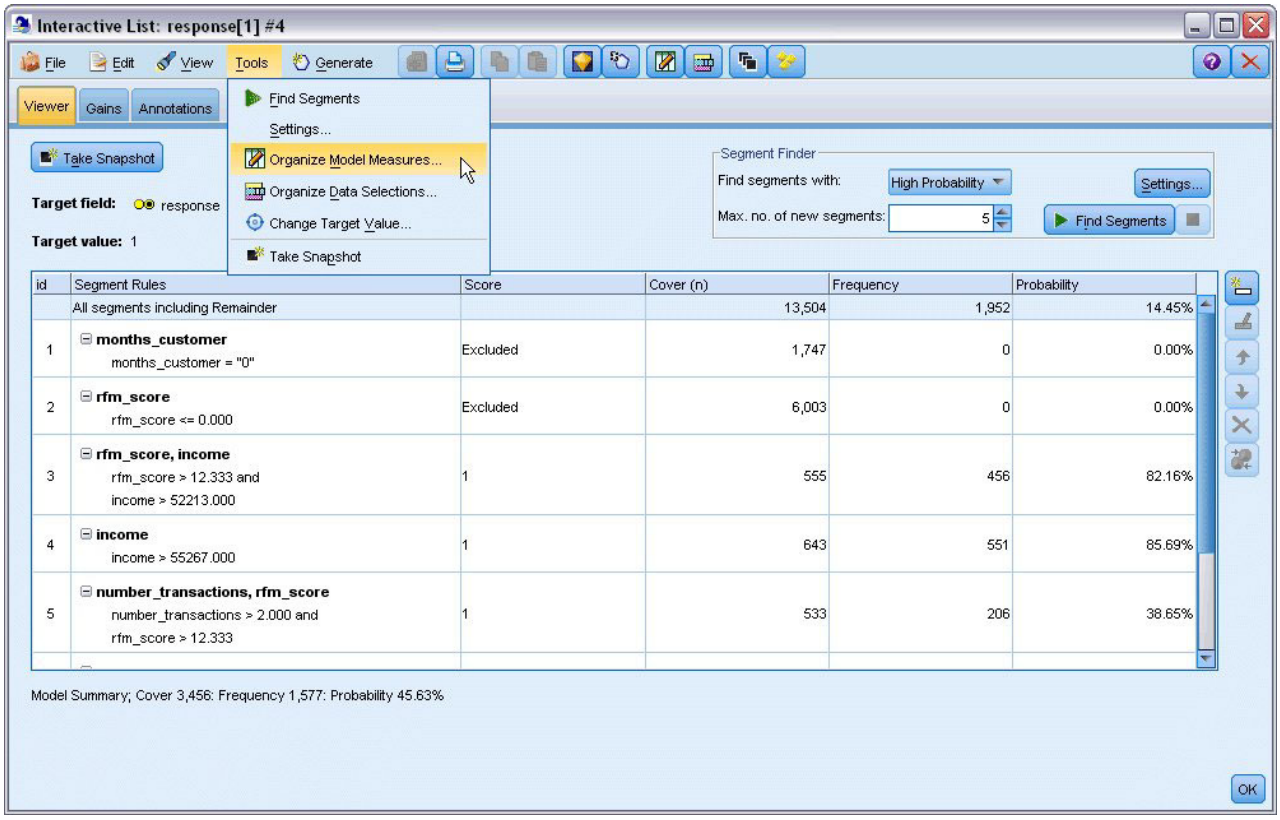


图 134. 组织模型测量

使用“组织模型测量”对话框，可以选择要在交互列表查看器中显示的测量量（或列）。还可以指定是根据所有记录还是选定的子集来计算测量，并且在适用的情况下您可以选择显示饼图而不是数字。

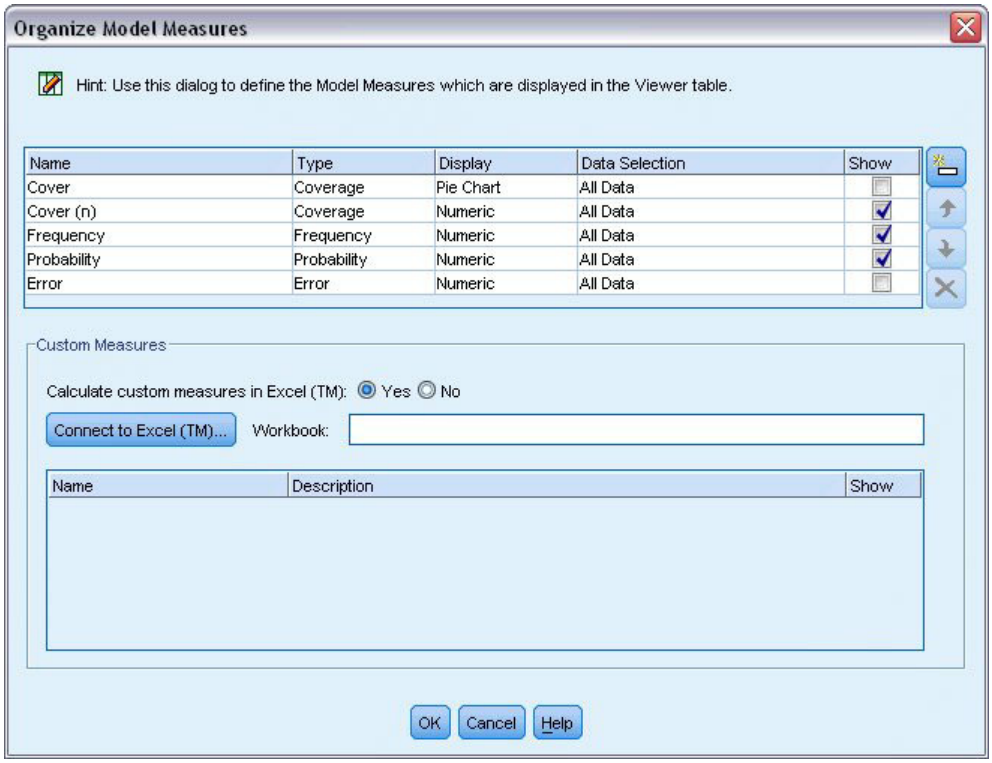


图 135. "组织模型测量"对话框

此外，如果已安装 Microsoft Excel，那么您可以链接到 Excel 模板，该模板将用于计算定制测量并将这些测量添加到交互显示中。

2. 在"组织模型测量"对话框中，将**计算 Excel 中的自定义测量量**设置成**是**。
3. 单击**连接到 Excel (TM)**
4. 选择 *template_profit.xlt* 工作簿（位于 IBM SPSS Modeler 安装位置的 *Demos* 文件夹中的 *streams* 下），然后单击**打开**启动电子表格。

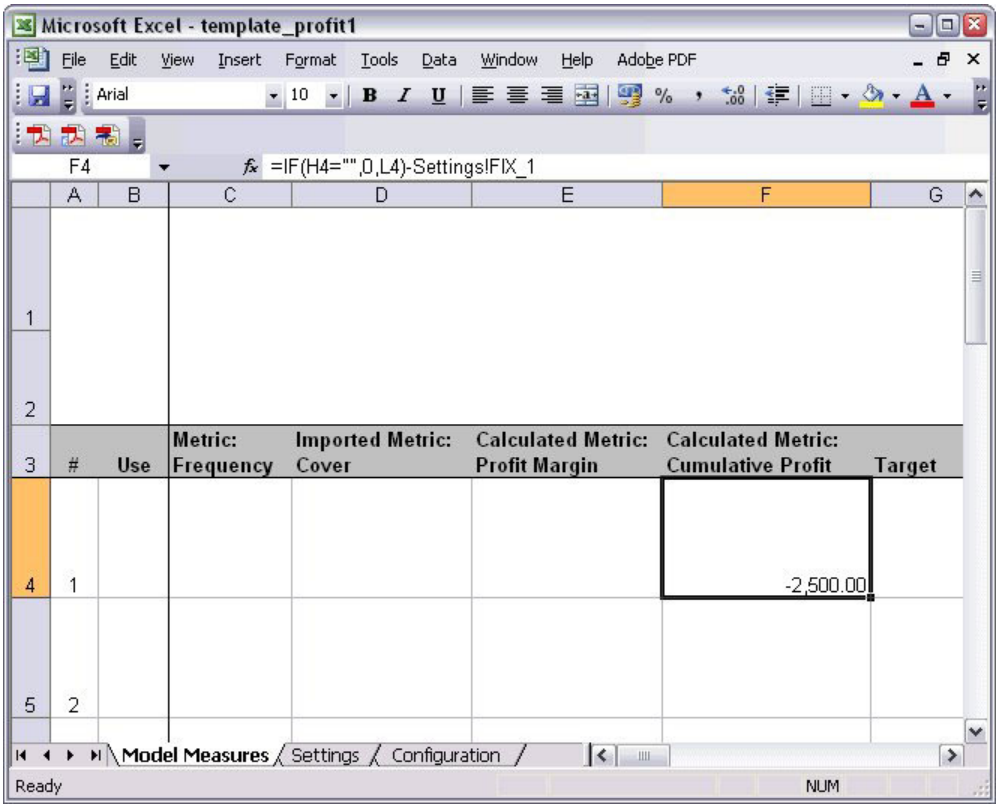


图 136. "Excel 模型测量"工作表

Excel 模板包含以下三个工作表:

- **模型测量**: 显示从模型中导入的模型测量，并计算用于重新导出至模型的定制测量。
- **设置**: 包含要用于计算定制测量的参数。
- **配置**: 定义要从模型中导入以及导出至模型的测量。

重新导出至模型的矩阵如下:

- **边际利润**。来自段的净收入
- **累积利润**。来自活动的总利润

通过以下公式定义:

```
毛利 = 频率 * 每个响应者的收入 - 涉及范围 * 可变成本
累积利润 = 总毛利 - 固定成本
```

请注意，"频率和涉及范围"将从模型中导入。

成本和收入参数由用户在"设置"工作表中指定。

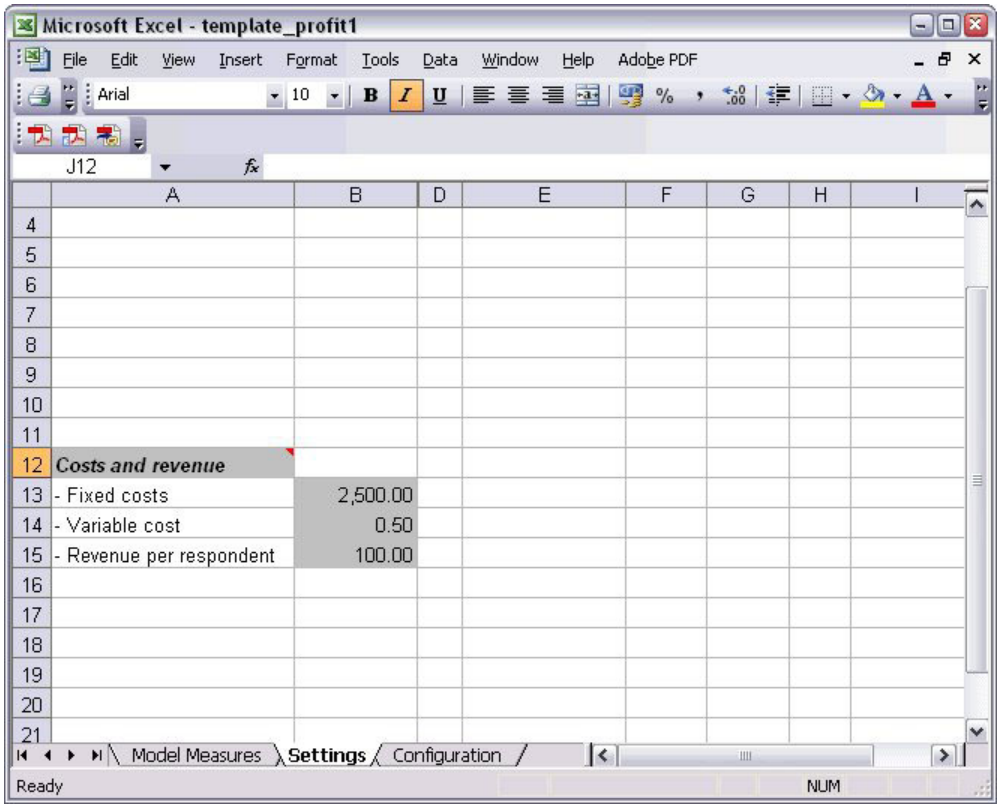


图 137. "Excel 设置"工作表

固定成本是活动的准备成本，例如设计和规划的成本。

可变成本是将报价发送给每位客户的成本，如信封和邮票的成本。

每个响应者的收入是响应报价的客户的净收入。

5. 要完成返回到模型的链接，可以使用 Windows 任务栏（或者按 Alt+Tab）返回到"交互列表"查看器。

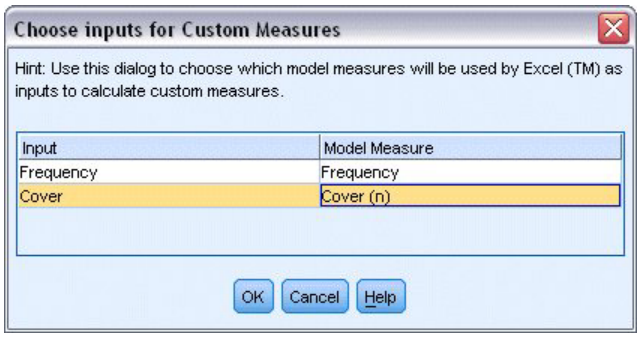


图 138. 选择定制测量的输入

此时将显示"选择定制测量的输入"对话框，您可在其中将模型中的输入映射为模板中定义的特定参数。左侧的列列出了可用测量，右侧的列根据"配置"工作表中的定义将这些测量映射到电子表格参数。

6. 在**模型测量量**列中，选择相对于各自输入的**频率**和**涉及范围 (n)** 并单击**确定**。

在本例中，"频率和涉及范围 (n)"模板中的参数名恰好与输入相匹配，但也可以使用其他名称。

7. 在"组织模型测量"对话框中单击**确定**以更新交互列表查看器。

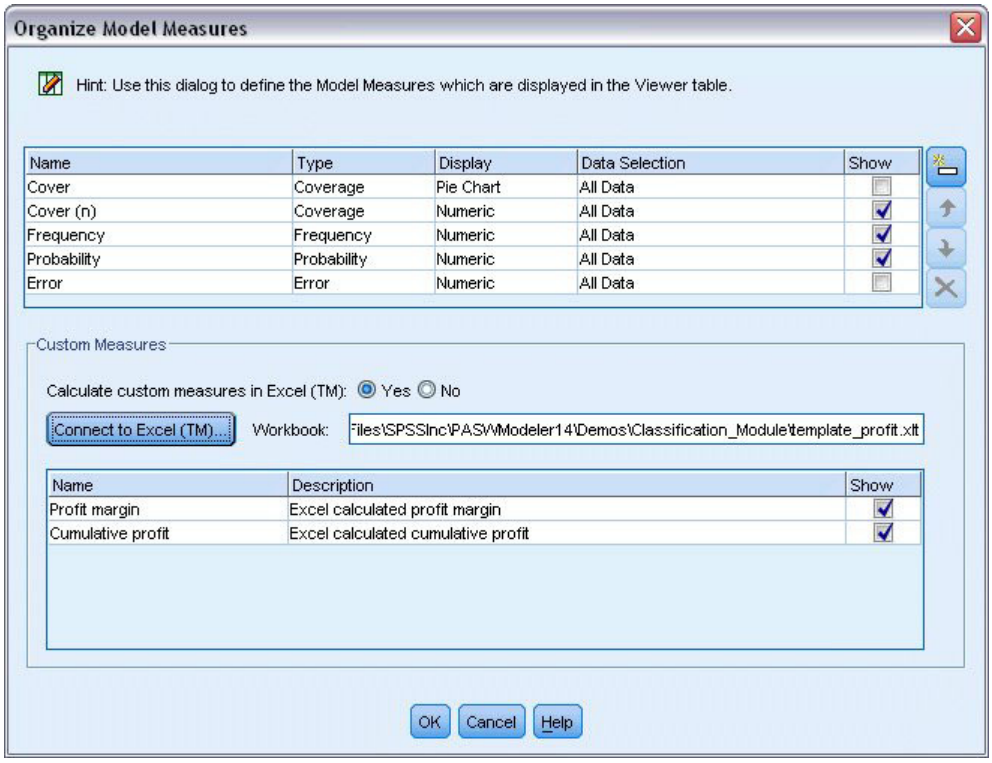


图 139. 显示 Excel 的定制测量的“组织模型测量”对话框

现在，新的测量将作为新列添加到窗口中，并将在模型每次更新时进行重新计算。

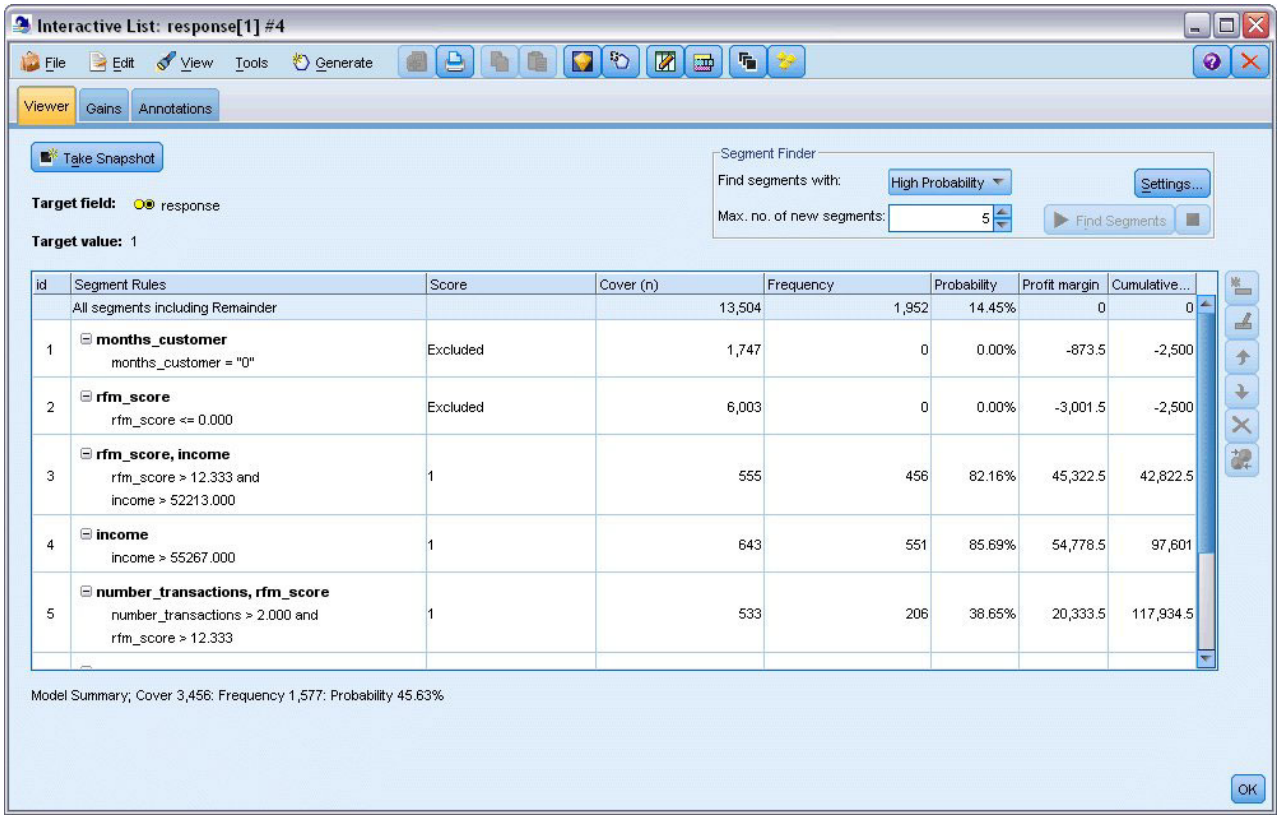


图 140. "交互列表"查看器中显示的 Excel 定制测量

通过编辑 Excel 模板，可以创建任意数量的定制测量。

修改 Excel 模板

尽管 IBM SPSS Modeler 提供了可以与交互列表查看器一起使用的缺省 Excel 模板，但您可能希望更改设置或添加自己的设置。例如，对您的组织而言，模板中的成本可能并不正确并且需要修改。

注：如果确实要修改现有模板或创建自己的模板，请记住应使用 Excel 2003 *.xlt* 后缀保存文件。

要使用新成本和收入细节修改缺省模板，并使用新数据更新交互列表查看器：

1. 请在交互列表查看器中，选择"工具"菜单中的**组织模型测量**。
2. 在"组织模型测量"对话框中，单击**连接到 Excel™**。
3. 选择 *template_profit.xlt* 工作簿并单击 **打开** 以启动电子表格。
4. 选择"设置"工作表。
5. 将 **固定成本** 编辑为 3,250.00，并将 **每个响应者的收入** 编辑为 150.00。

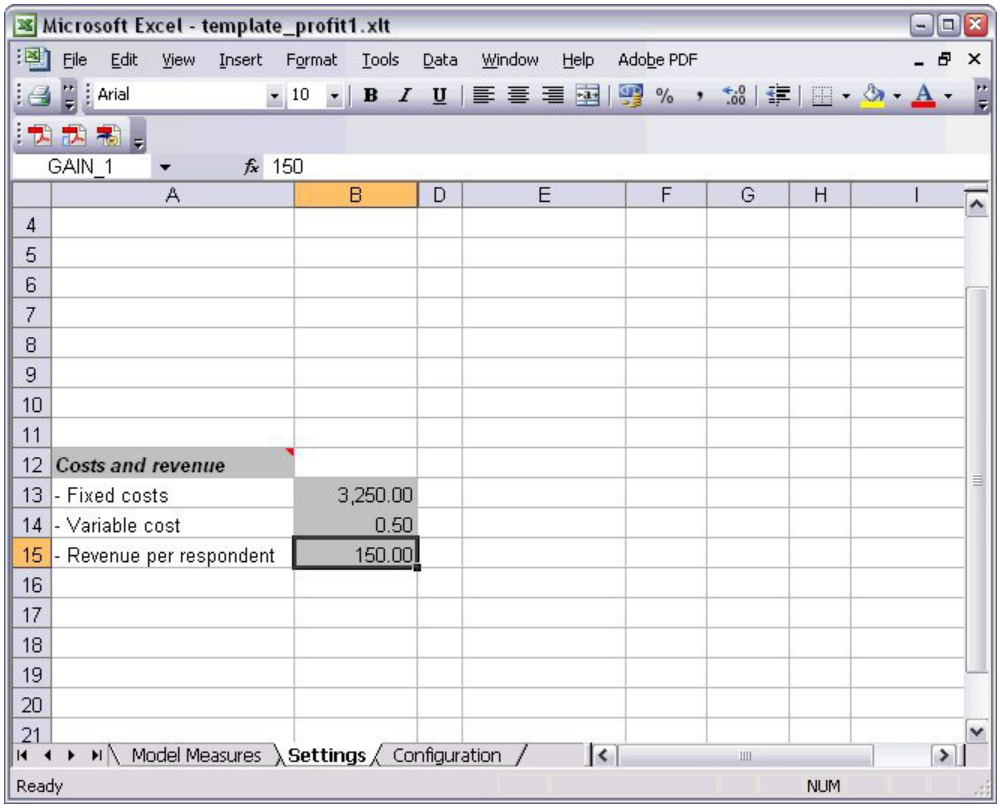


图 141. "Excel 设置"工作表中的已修改值

6. 使用唯一且相关的文件名保存已修改的模板。请确保文件的扩展名为 Excel 2003 *.xlt* 。

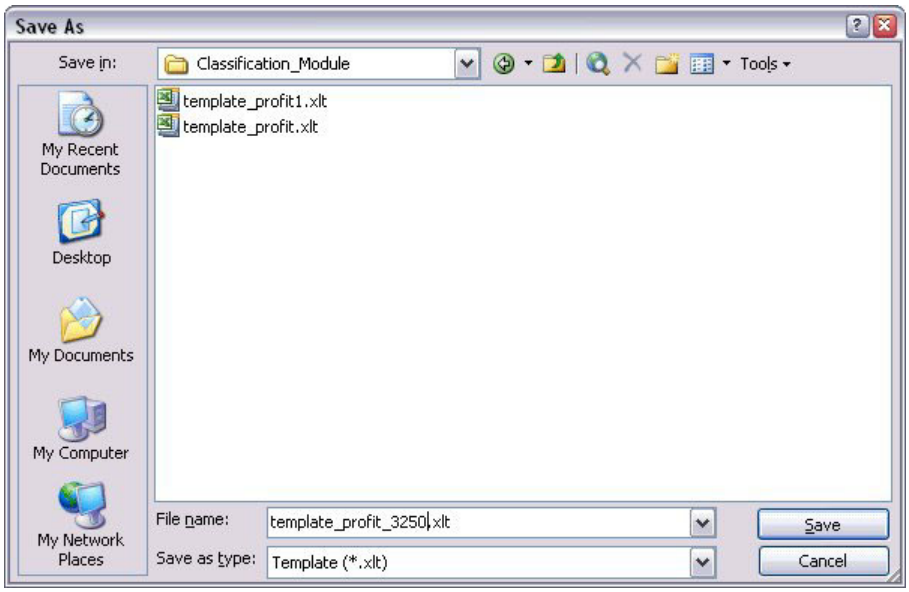


图 142. 保存已修改的 Excel 模板

7. 使用 Windows 任务栏（或者按 Alt+Tab）返回到"交互列表"查看器。

在"选择自定义测量的输入"对话框中，选择要显示的测量量并单击**确定**。

8. 在"组织模型测量"对话框中，单击**确定**以更新交互列表查看器。

显然，本模型仅显示了一种修改 Excel 模板的简单方式；您还可以进一步更改，例如，从交互列表查看器提取数据或向其中传递数据；也可以在 Excel 中生成其他输出（如图形）。

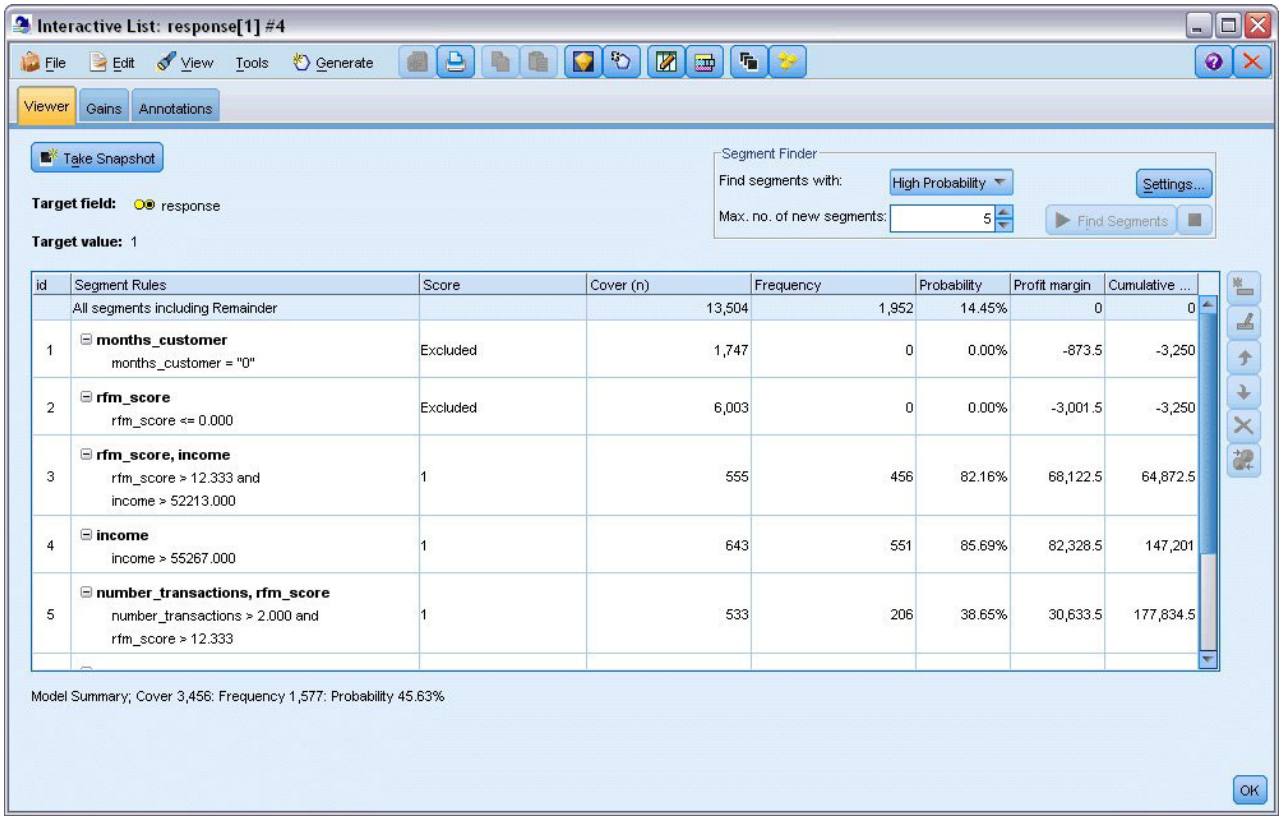


图 143. 交互列表查看器中显示的已修改的 Excel 定制测量量

保存结果

为了保存模型以供稍后在交互式会话期间使用，您可以创建该模型的快照，此快照将在“快照”选项卡上列出。可以在交互式会话期间随时返回到任何已保存的快照。

按此方式继续时，您可以尝试使用其他挖掘任务来搜索其他段。还可以编辑现有段、根据自己的业务规则插入自定义段、创建数据选择以优化特定组的模型以及以多种其他方式定制模型。最后，可以根据情况明确包含或排除每个段，以指定每个段的评分方式。

如果对结果感到满意，那么可以使用“生成”菜单生成模型，此模型可以添加到流中或进行部署以用于评分。

此外，要保存交互会话的当前状态以备他日使用，可以在“文件”菜单中选择**更新建模节点**。此操作将决策列表建模节点更新为当前设置，包括挖掘任务、模型快照、数据选择和自定义测量量。下次运行流时，只需确保在决策列表建模节点中选中了**使用保存的会话信息**，就可将会话恢复到其当前状态。

第 12 章 电信业客户分类（多项 Logistic 回归）

Logistic 回归是一种统计方法，它可根据输入字段的值对记录进行分类。这种技术与线性回归类似，但用分类目标字段代替了数值字段。

例如，假设某个电信提供商根据服务使用情况模式对其客户群进行了细分，将这些客户分为了四个组。如果人口统计数据可用于预测组成员资格，那么您可以为各个潜在客户定制报价。

此示例使用名为 *telco_custcat.str* 的流，此流引用名为 *telco.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *telco_custcat.str* 位于 *streams* 目录下。

本示例主要讲述使用人口统计数据预测使用情况模式。目标字段 *custcat* 有四个可能的值对应于四个客户组，如下所示:

值(V)	标签
1	基本服务
2	电子服务
3	增值服务
4	全套服务

由于目标具有多个类别，因此将使用多项模型。如果目标具有两个截然不同的类别（例如，是/否、真/假或者流失/未流失），那么可以改为创建二项模型。请参阅主题第 135 页的第 13 章，『电信客户流失（二项 Logistic 回归）』以获取更多信息。

构建流

1. 在 *Demos* 文件夹中添加指向 *telco.sav* 的“Statistics 文件”源节点。

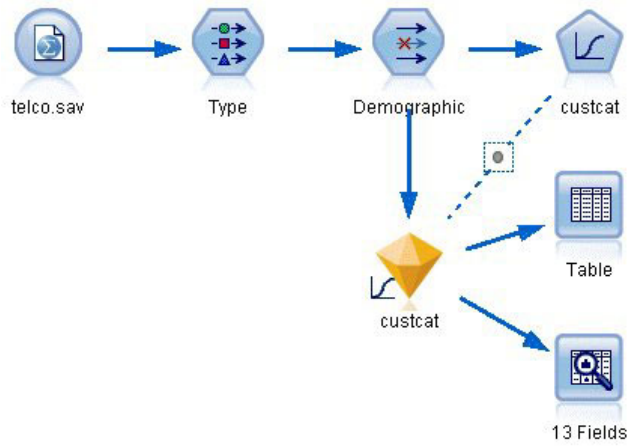


图 144. 用于通过多项 Logistic 回归对客户进行分类的样本流

a. 添加类型节点并单击**读取值**，确保所有测量级别设置正确。例如，具有值 0 和 1 的多数字段可视为标志。

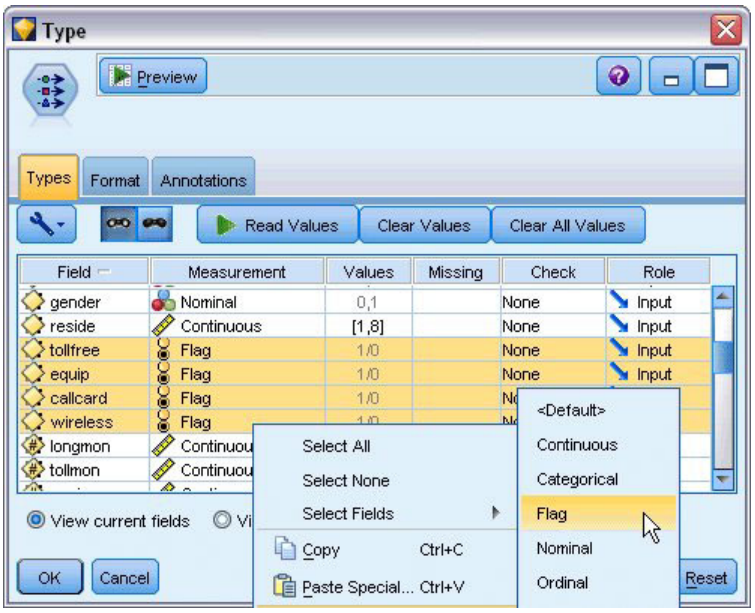


图 145. 设置多个字段的测量级别

提示：要更改具有相似值（例如 0/1）的多个字段的属性，请单击*值*列标题以按照值对字段进行排序，然后在按住 Shift 键的同时使用鼠标或箭头键选择所有要更改的字段。然后可以右键单击选定的内容以更改选定字段的测量级别或其他属性。

注意，*性别*更准确而言应视为具有两个值的集合的字段，而不是标志，所以将其测量值保留为**名义**。

b. 将*客户类别*字段的角色设置为**目标**。将所有其他字段的角色设置为 **Input**。

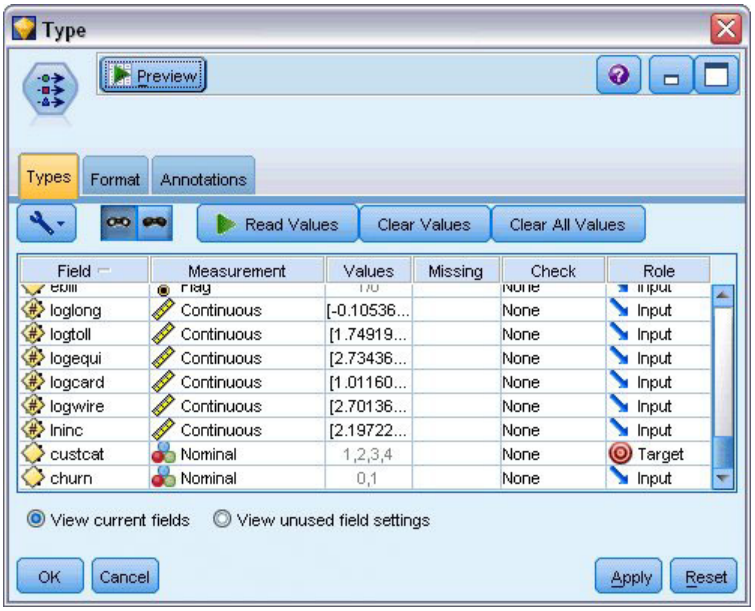


图 146. 设置字段角色

因为本示例主要讲述人口统计，所以请使用过滤节点以选取相关字段（ *地区* 、 *年龄* 、 *婚姻状况* 、 *地址* 、 *收入* 、 *教育程度* 、 *行业* 、 *退休* 、 *性别* 、 *居住地* 和 *客户类别* ）。其他字段可以排除在此分析之外。

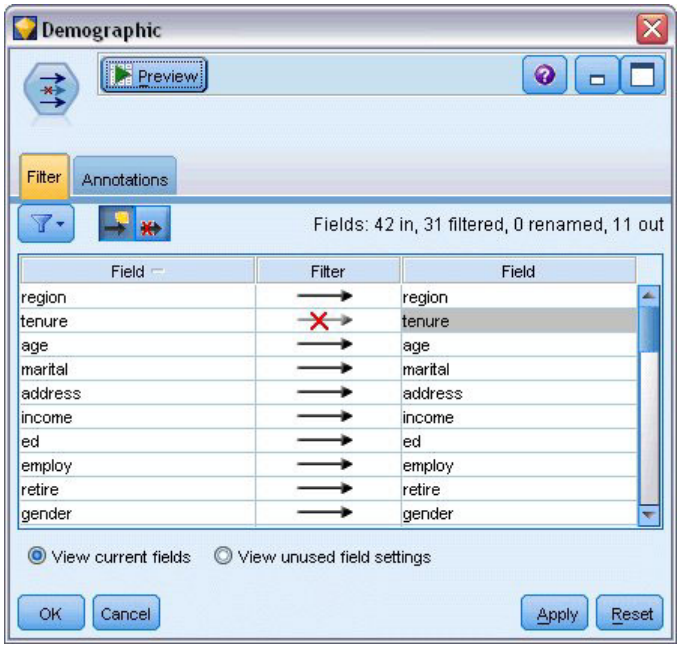


图 147. 过滤人口统计字段

（另外，您可以将这些字段的角色更改为**无**，而不要排除这些字段，或者选择要在建模节点中使用的字段。）

2. 在 Logistic 节点上，单击 **模型** 选项卡并选择 **逐步** 法。选中 **多项** 、 **主效应** 和 **将常量纳入方程式** 。

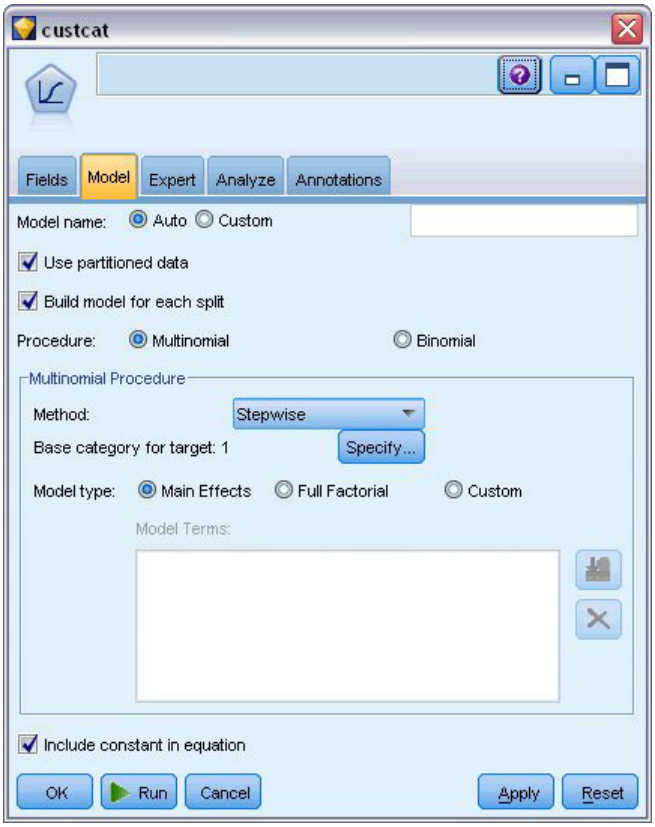


图 148. 选择模型选项

将目标的底数类别保留为 1。模型将对其他客户与预订基本服务的客户进行比较。

3. 在“专家”选项卡上，选中**专家**模式，选中**输出**，然后在“高级输出”对话框中选中**分类表**。

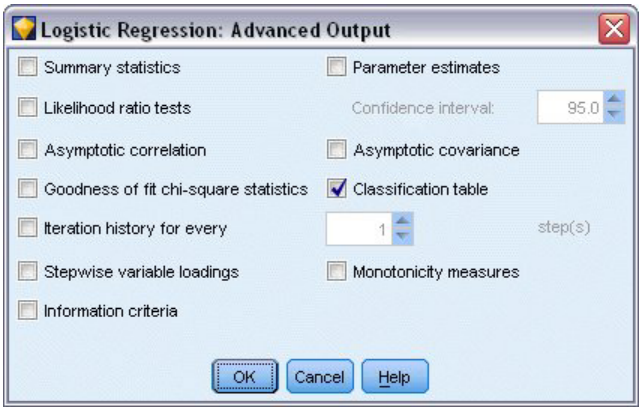


图 149. 选择输出选项

浏览模型

1. 执行节点以生成将添加到右上角的“模型”选用板中的模型。要查看其详细信息，请在生成的模型节点上用右键单击并选择 **浏览** 。

“模型”选项卡中显示了用于将记录分配到目标字段的每个类别的方程式。有四个可能的类别，其中一个是未对其显示方程详细信息的基准类别。在余下的三个方程式中显示了详细信息，其中类别 3 表示增值服务，依此类推。

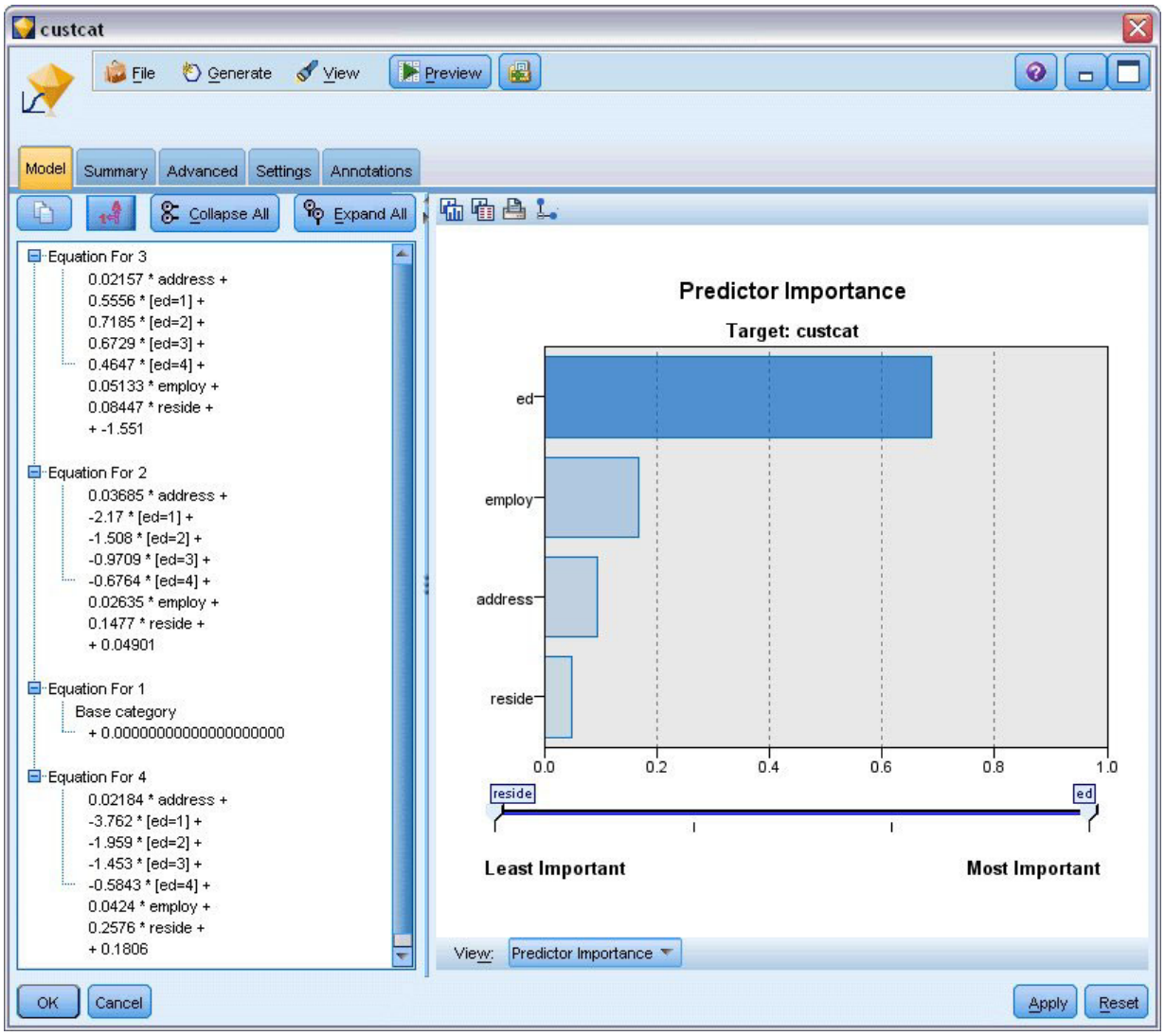


图 150. 浏览模型结果

“摘要”选项卡显示了模型使用的目标字段和输入字段（预测变量字段）以及其他内容。请注意，这是根据步进法实际选择的字段，而不是提交以供考虑的完整列表。

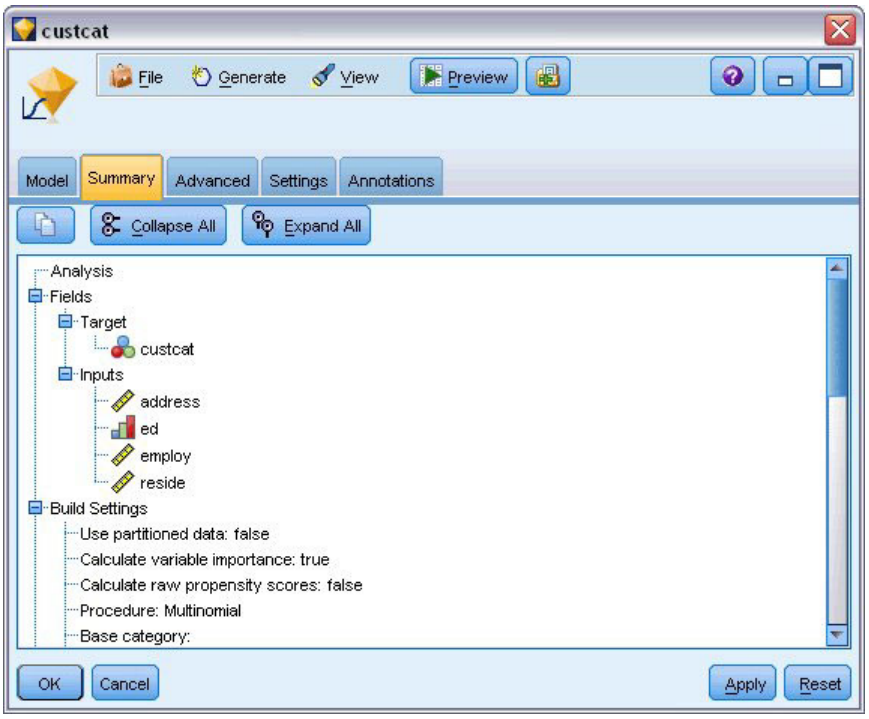


图 151. 显示目标字段和输入字段的模型摘要

“高级”选项卡上显示的项取决于在建模节点的“高级输出”对话框中选择的选项。

始终显示的一项为“个案处理摘要”，此摘要显示了落在目标字段的每个类别中的记录百分比。这将生成一个空模型用作比较的基础。

在不构建使用预测变量的模型的情况下，最佳预测是将所有客户都分配到最常用的组（用于增值服务的组）。

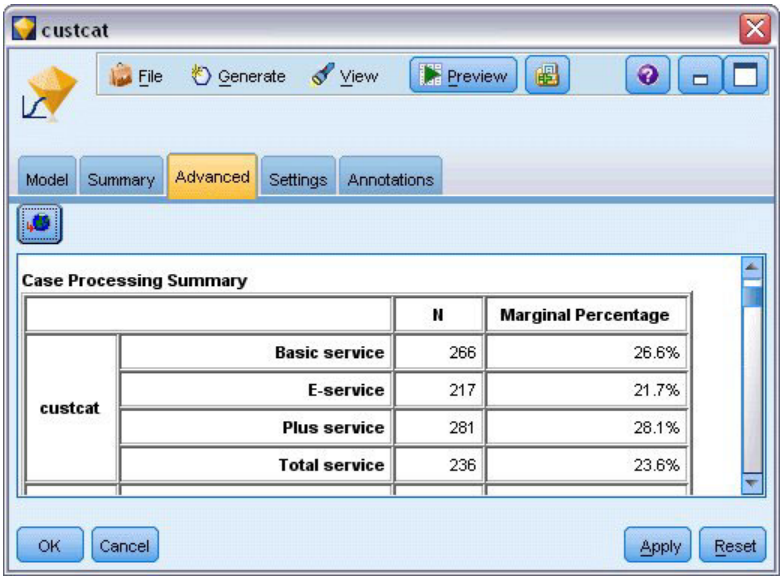


图 152. 个案处理摘要

如果基于训练数据将所有客户分配到空模型，则得到的正确率将是 281/1000 = 28.1%。“高级”选项卡还包括其他信息，这些信息使您可以检查模型的预测。然后，可将这些预测与空模型的结果相比，以查看使用此数据的模型的执行效果。

在“高级”选项卡底部，分类表显示了此模型的结果，其正确率为 39.9%。

特别是，此模型在识别全套服务客户（类别 4）时表现优异，而在识别电子服务客户（类别 2）时表现很差。对于类别 2 中客户，如果您希望提高准确性，那么可能需要找到另一个预测变量来识别这些客户。

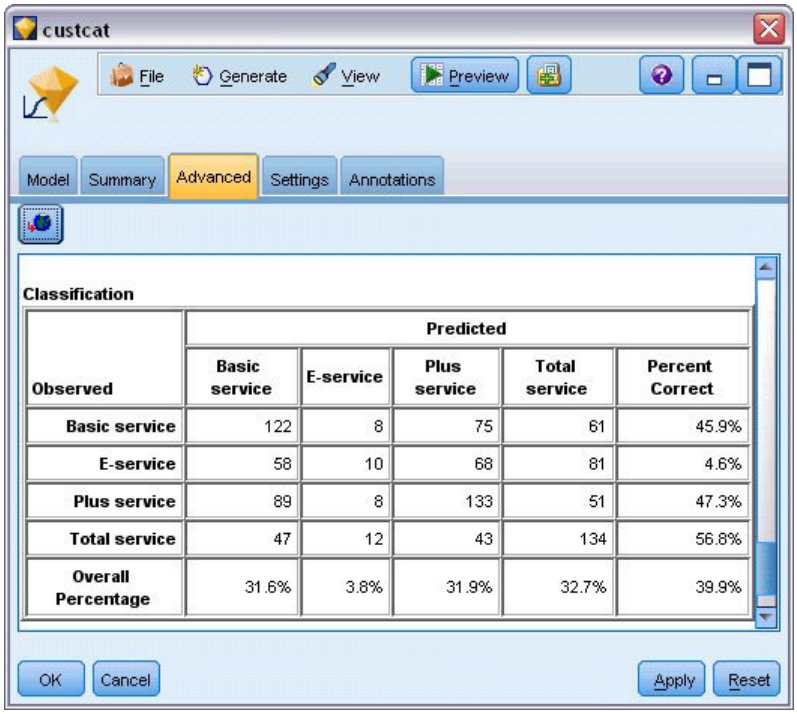


图 153. 分类表

根据您要预测的内容，模型可以充分地满足您的需求。例如，如果您不关心识别类别 2 中的客户，那么该模型的准确性足以满足需求。这种情况可能是，电子服务仅是一种为吸引顾客而出售且获利微薄的产品。

例如，如果投资的最高回报来自于落在类别 3 或类别 4 中的客户，则该模型能够提供所需的信息。

要评估模型对数据的实际拟合度，可以在构建该模型时使用“高级输出”对话框中提供的一些诊断结果。有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于安装光盘的 *\Documentation* 目录中。

另请注意，这些结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

第 13 章 电信客户流失（二项 Logistic 回归）

Logistic 回归是一种统计方法，它可根据输入字段的值对记录进行分类。这种技术与线性回归类似，但用分类目标字段代替了数值字段。

此示例使用名为 *telco_churn.str* 的流，此流引用名为 *telco.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *telco_churn.str* 位于 *streams* 目录下。

例如，假设某个电信服务提供商关心流失到竞争对手那里的客户数。如果可以将服务使用情况数据用于预测可能会转移到其他提供商的客户，那么可通过定制报价来尽可能多地保留客户。

本示例主要讲述利用使用情况数据来预测客户流失（顾客流失率）。由于目标具有两个截然不同的类别，因此将使用二项模型。如果目标具有多个类别，那么将改为创建多项模型。请参阅主题第 127 页的第 12 章，『电信业客户分类（多项 Logistic 回归）』以获取更多信息。

构建流

1. 在 *Demos* 文件夹中添加指向 *telco.sav* 的“Statistics 文件”源节点。

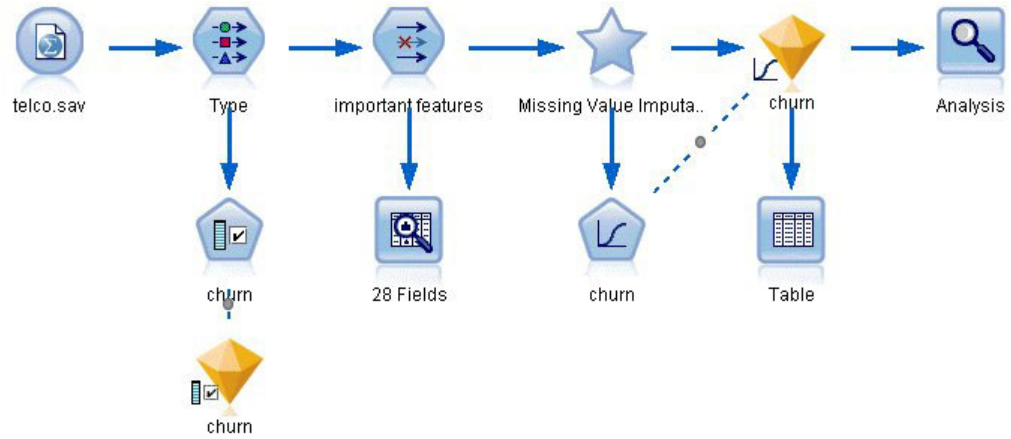


图 154. 用于通过二项 Logistic 回归对客户进行分类的样本流

2. 添加“类型”节点以定义字段，从而确保所有测量级别都已正确设置。例如，大多数值为 0 和 1 的字段都可以用作标志字段，但某些字段，比如性别，作为包含两个值的名义字段会更加准确。

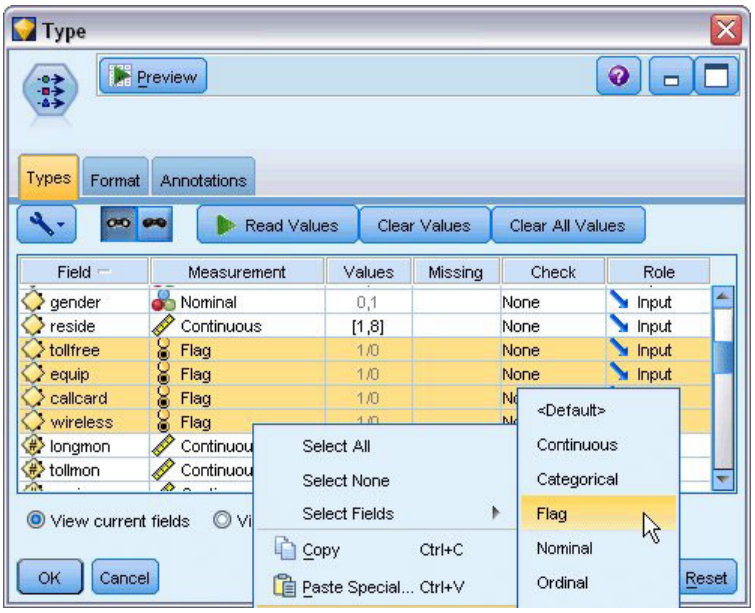


图 155. 设置多个字段的测量级别

提示: 要更改多个具有相似值（例如 0/1）的字段的属性，请单击*值*列标题以按照值对字段进行排序，然后在按住 Shift 键的同时使用鼠标或箭头键选择所有要更改的字段。然后可以右键单击选定的内容以更改选定字段的测量级别或其他属性。

3. 将*流失*字段的测量级别设置为**标志**，并将其角色设置为**目标**。将所有其他字段的角色设置为 **Input**。

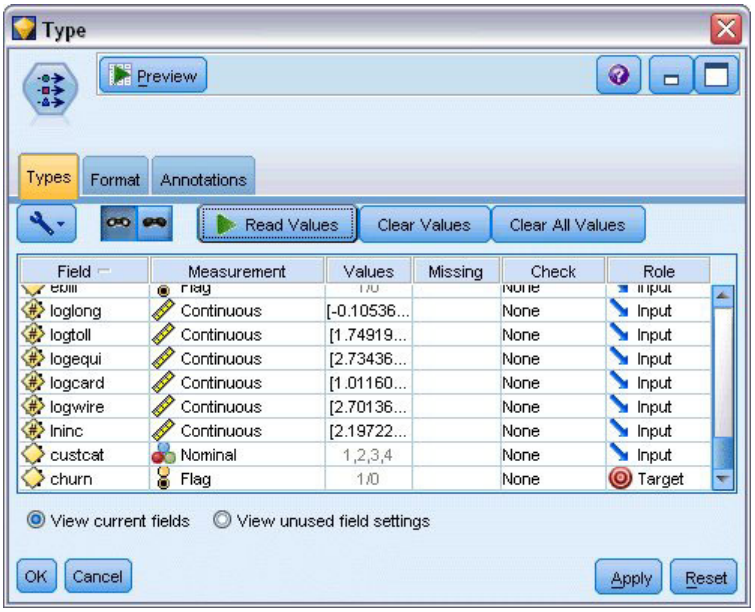


图 156. 设置顾客流失率字段的测量级别和角色

4. 为类型节点添加“特征选择”建模节点。

   通过使用特征选择节点，对于不能为预测变量/目标之间的关系添加任何有用信息的预测变量或数据，可以将其删除。

5. 运行流。

6. 打开结果模型块，并从**生成**菜单中，选择**过滤**以创建过滤节点。

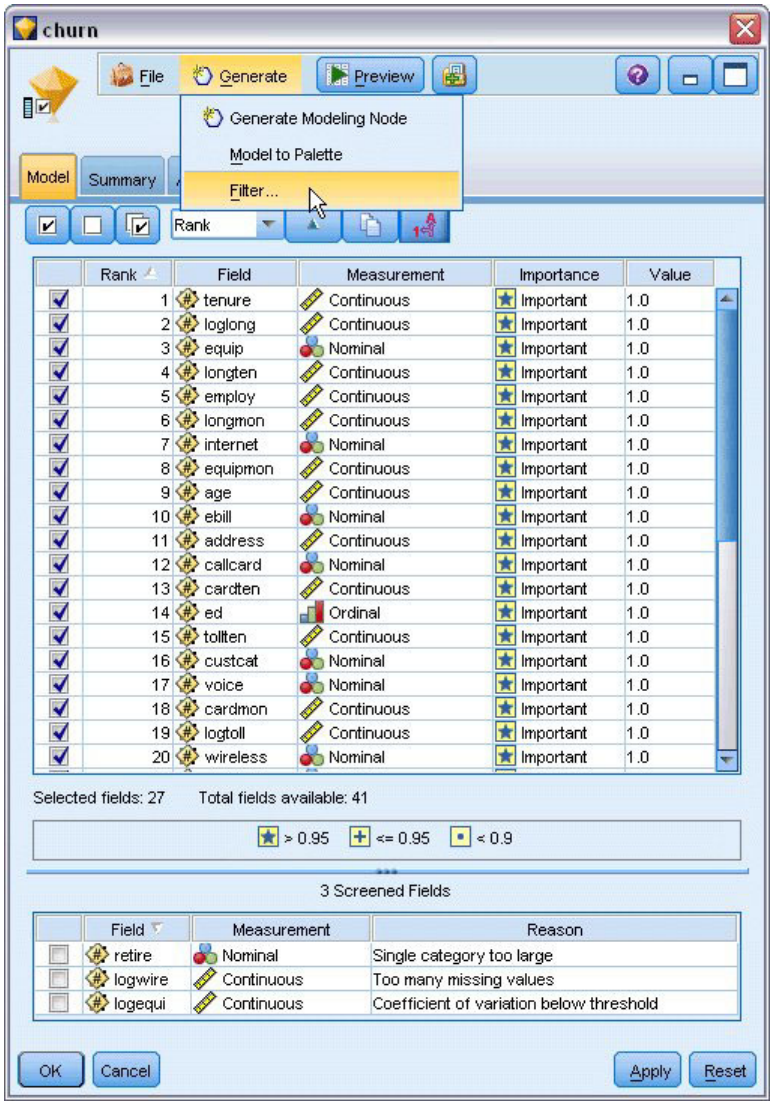


图 157. 从“特征选择”节点生成“过滤”节点

不是 *telco.sav* 文件中的所有数据都对预测客户流失有用。可以使用过滤器仅选择被认为很重要的数据来用作预测变量。

7. 在“生成过滤”对话框中，选择**所有已标记的字段**：**重要**，然后单击**确定**。
8. 将生成过滤节点附加到类型节点。

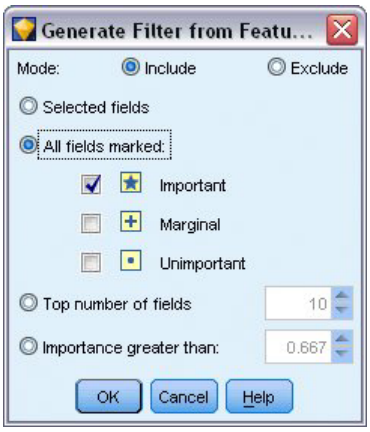


图 158. 选择重要字段

9. 将数据审核节点附加到生成的“过滤”节点。

   打开数据审核节点，然后单击**运行**。

10. 在“数据审核”浏览器的“质量”选项卡上，单击 *% 完成*列以便按数值升序顺序对此列进行排序。这样就可以识别所有含有大量缺失数据的字段；在本示例中，唯一需要修改的字段是 *logtoll*，其完成值比例小于 50%。

11. 在 *logtoll* 的 *归因于缺失* 列中，单击 **指定**。

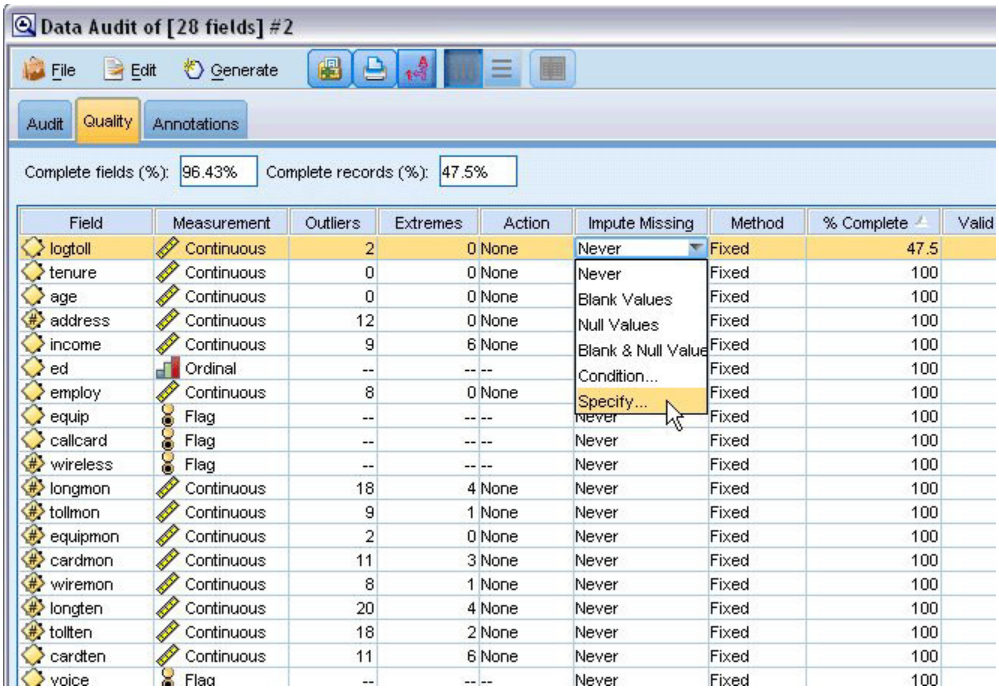


图 159. logtoll 的插补缺失值

12. 对于 **归因条件**，选择 **空白值和空值**。对于**固定为**，选择**平均值**，然后单击**确定**。

    选择 **平均值** 可确保归因值不会反过来影响总数据中所有值的平均值。

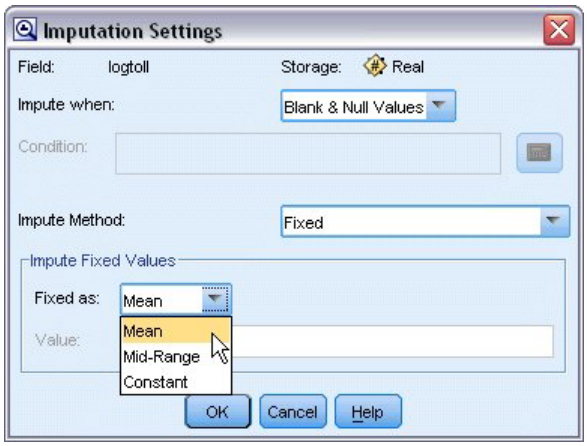


图 160. 选择插补设置

13. 在“数据审核”浏览器的“质量”选项卡上，生成缺失值超节点。为完成此操作，可从菜单中选择以下项:

    生成 > 缺失值超节点

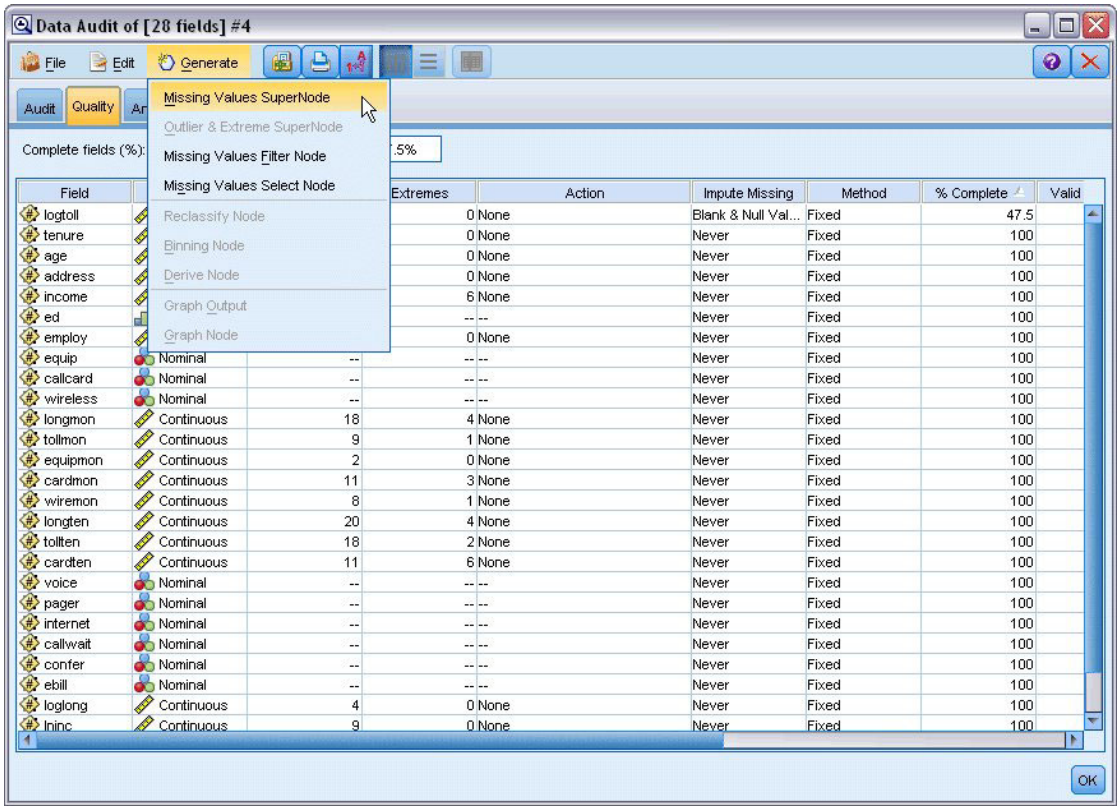


图 161. 生成缺失值超节点

在“缺失值超节点”对话框中，将**样本大小**增加到 50%，然后单击**确定**。

超节点将显示在流画布中，其标题为: *缺失值插补*。

14. 将超节点附加到过滤节点。

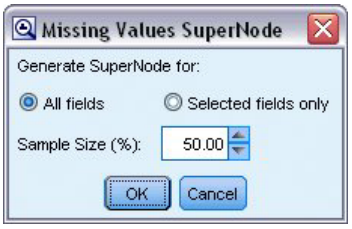


图 162. 指定样本大小

15. 将 Logistic 节点添加到超节点。
16. 在 Logistic 节点上，单击“模型”选项卡并选择**二项**过程。在 *二项过程* 区域，选择 **前进** 法。

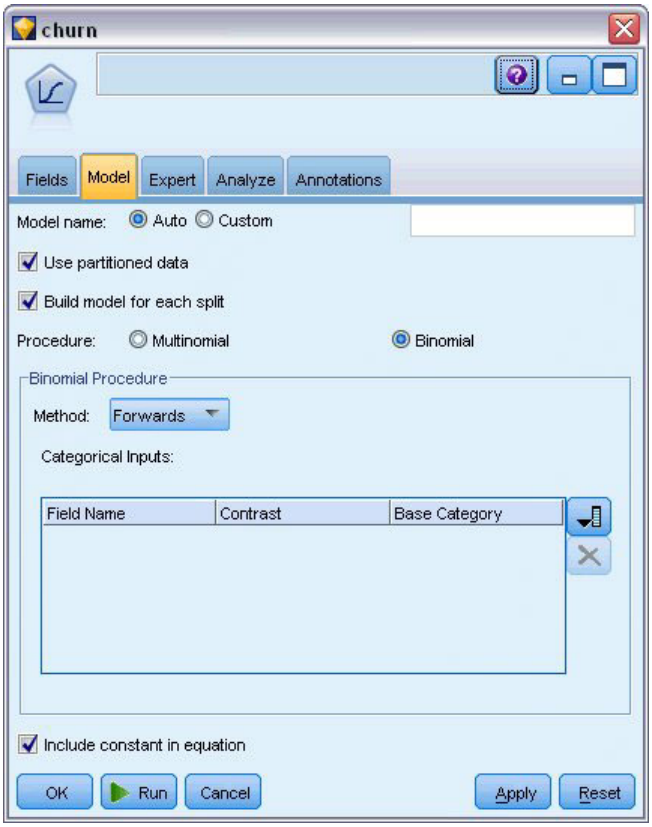


图 163. 选择模型选项

17. 在“专家”选项卡上，选择**专家**模式，然后单击**输出**。此时显示“高级输出”对话框。
18. 在“高级输出”对话框中，选择**在每个步骤**作为显示类型。选择 **迭代历史记录** 和 **参数估计** ，然后单击 **确定** 。

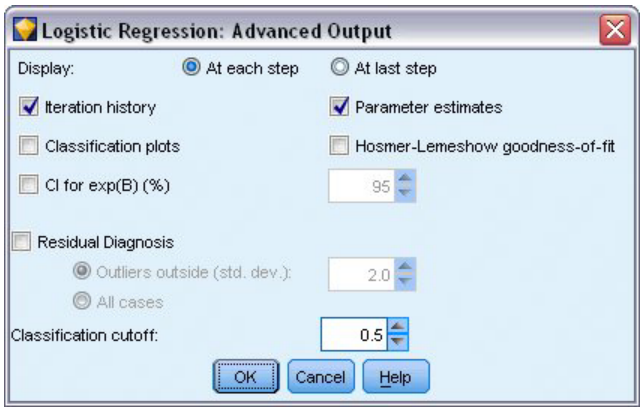


图 164. 选择输出选项

浏览模型

1. 在 Logistic 节点上，单击**运行**创建模型。

   模型块将添加到流画布和右上角的“模型”选用板中。要查看其详细信息，右键单击模型块并选择**编辑**或**浏览**。

   “摘要”选项卡显示了模型使用的目标字段和输入字段（预测变量字段）以及其他内容。请注意，这些根据前向法实际选择出来的字段，而不是提交以供考虑的完整列表。

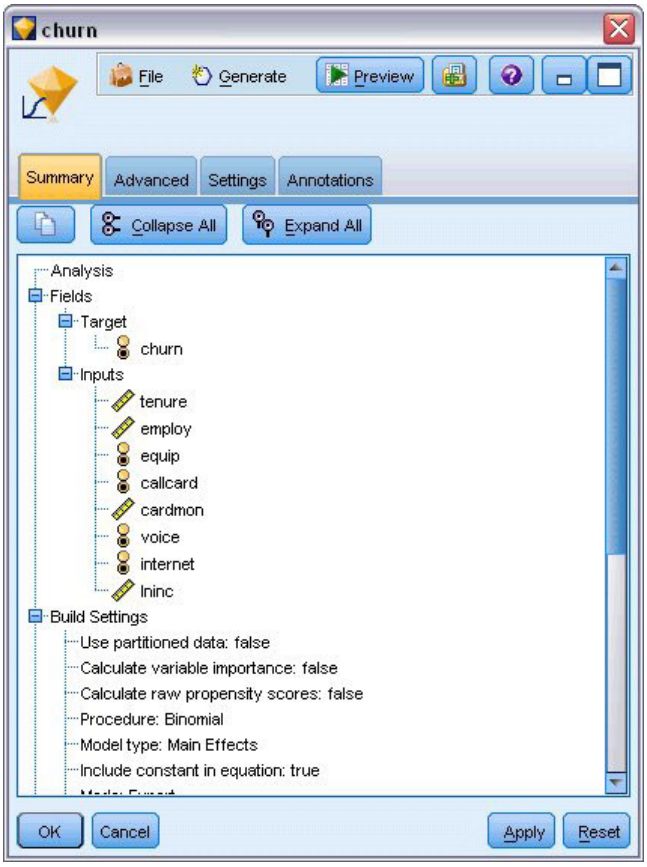


图 165. 显示目标字段和输入字段的模型摘要

“高级”选项卡上显示的项取决于在 Logistic 节点的“高级输出”对话框中选择的选项。始终显示的一项为“个案处理摘要”，此摘要显示了分析中包括的记录数及百分比。另外，此摘要还列出了其中有一个或多个输入字段不可用的缺失个案数（如果有的话）以及所有未选定的个案数。

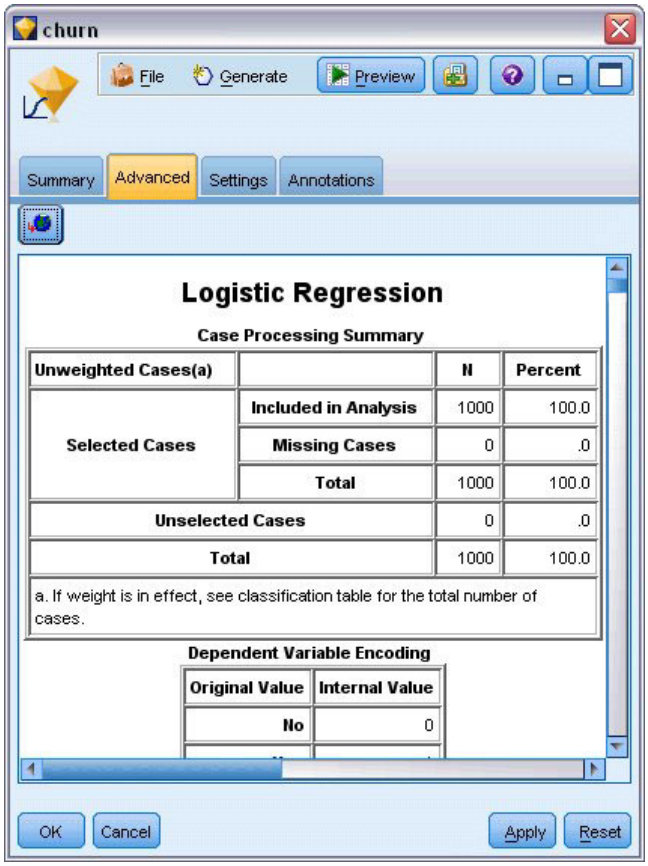


图 166. 个案处理摘要

2. 在“个案处理摘要”中向下滚动，以显示“块 0: 起始块”下的分类表。

   向前步进法从空模型（即，没有预测变量的模型）开始，可以将此空模型用作与最终构建的模型进行比较的基础。按照惯例，此空模型会将所有值都预测为 0，因此其准确度为 72.6%，这完全是因为已正确预测到 726 个未流失的客户。但是，根本没有正确预测到已流失的客户。

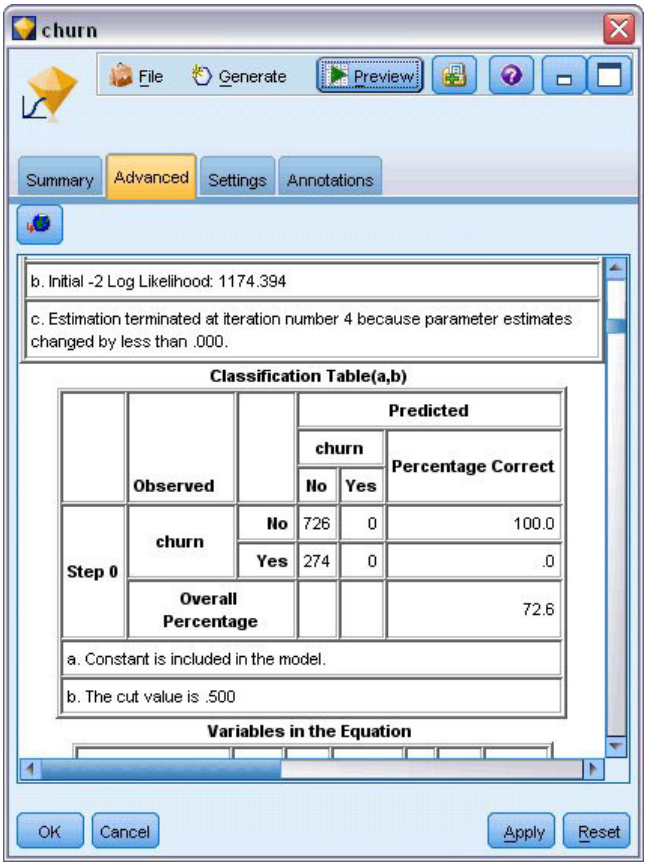


图 167. 起始分类表 - 块 0

3. 现在向下滚动以显示“块 1: 方法 = 向前步进法”下的分类表。

   此分类表显示了在每个步骤中添加预测变量之后模型的结果。在第一个步骤中（在仅使用了一个预测变量之后），模型预测流失的准确性就已从 0.0% 增加到 29.9%。

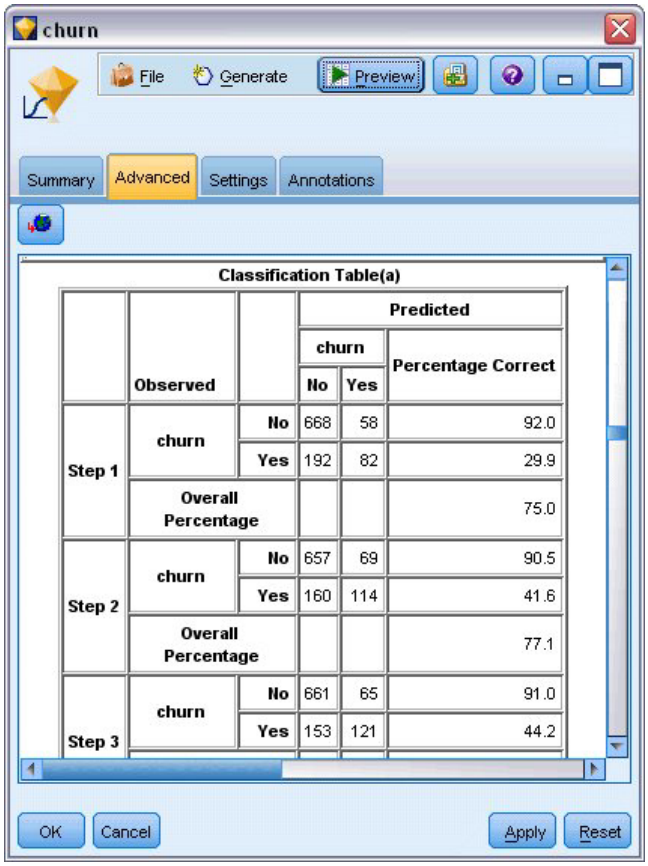


图 168. 分类表 - 块 1

4. 向下滚动到此分类表的底部。

分类表显示步骤 8 为最后一步。在此阶段，算法已确定不再需要向模型添加任何其他预测变量。虽然预测非流失客户的准确性有所下降，达到了 91.2%，但预测已流失客户的准确性却从原来的 0% 上升到了 47.1%。这相比原来不使用任何预测变量的空模型其有效性显著提高。

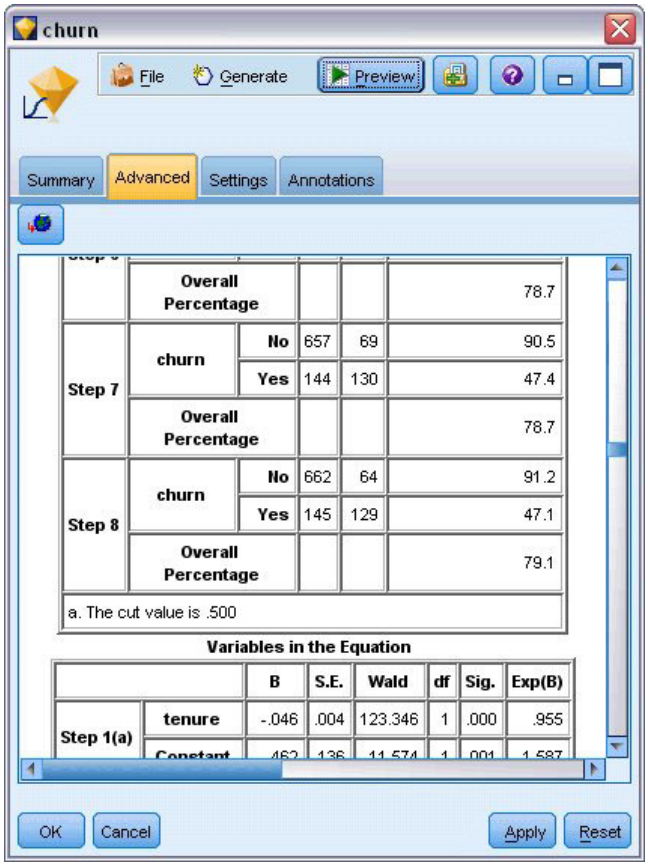


图 169. 分类表 - 块 1

对于希望减少流失的客户，能够将流失率减少接近一半将会成为保护其收入流的主要步骤。

注: 此示例还显示了将总体百分比作为判断模型准确性的依据在某些情况下如何会导致错误结论。原来空模型的总准确性为 72.6%，而最终预测模型的总准确性为 79.1%；但是，正如我们所看到的，其实际单个类别的预测准确性的差别极大。

要评估模型对数据的实际拟合度，可以在构建该模型时使用“高级输出”对话框中提供的一些诊断结果。有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于安装光盘的 *\Documentation* 目录中。

另请注意，这些结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

第 14 章 预测带宽利用率（时间序列）

使用时间序列节点进行预测

为了预测带宽利用率，某个国内宽带提供商的一位分析员需要对用户预订进行预测。分析师需要对各地市场进行预测，才能得出全国注册用户数量。您将使用时间序列建模来生成对随后三个月多个地区市场的预测。第二个示例则说明源数据的格式不适合作为时间序列节点的输入时应如何转换源数据。

这两个示例均使用名为 *broadband_create_models.str* 的流，该流引用名为 *broadband_1.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *broadband_create_models.str* 位于 *streams* 文件夹中。

最后一个示例演示如何将保存的模型应用于更新过的数据集，以将预测时间延长三个月。

在 IBM SPSS Modeler 中，可以在单一操作中生成多个时间序列模型。将要使用的源文件具有 85 个不同市场的时间序列数据，但为简便起见，只为其中五个市场以及总体市场的数据建模。

broadband_1.sav 数据文件具有全部 85 个地区市场的月度带宽使用率数据。在本示例中，将只使用前五个序列；将为这五个序列各创建一个独立的模型，并创建一个总计模型。

该文件还包含指示每个记录的年份和月份的日期字段。此字段将在“时间间隔”节点中用于标注记录。日期字段会以字符串格式读入到 IBM SPSS Modeler 中，但为了在 IBM SPSS Modeler 中使用该字段，必须使用填充节点将存储类型转换为数字日期格式。

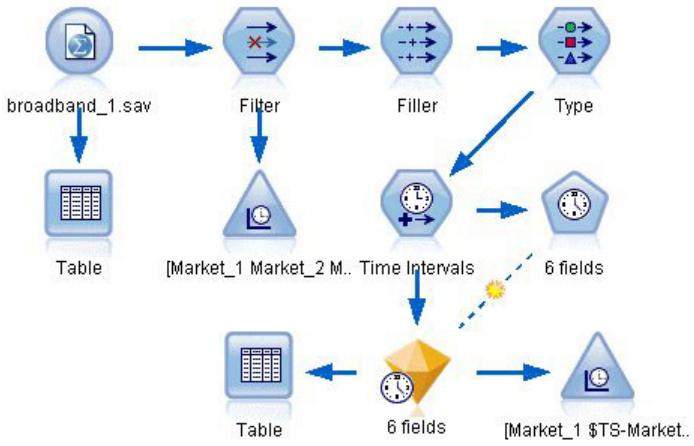


图 170. 用于显示时间序列建模的样本流

时间序列节点要求每个序列各占一列，每个区间各占一行。IBM SPSS Modeler 提供用于变换数据的方法，以使其在需要时符合此格式。

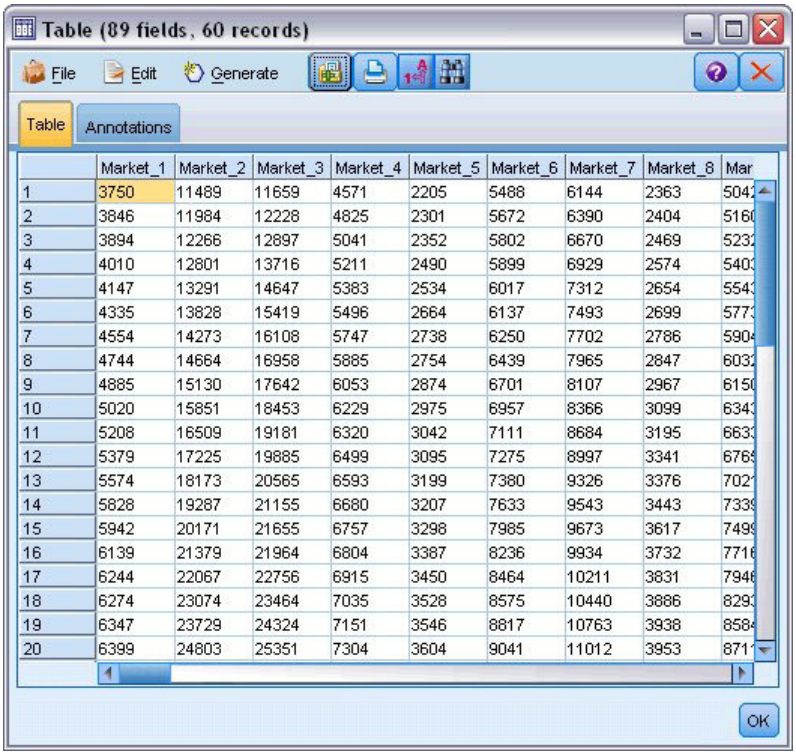


图 171. 宽带地区市场的月度预订数据

创建流

1. 新建流并添加指向 *broadband_1.sav* 的“Statistics 文件”源节点。
2. 使用过滤节点过滤掉 *Market_6* 至 *Market_85* 字段以及 *MONTH_* 和 *YEAR_* 字段，以简化模型。

提示: 要一次选定多个相邻字段，请单击 *Market_6* 字段，然后按住鼠标左键并向下拖动至 *Market_85* 字段。选定字段将以蓝色突出显示。要添加其他字段，请按住 Ctrl 键，然后单击 *MONTH_* 和 *YEAR_* 字段。

图 172. 简化模型

检查数据

构建模型之前，详细了解数据的性质始终是一个好主意。数据是否呈现季节性变化？虽然专家建模器可以自动找出每个序列的最佳季节性或非季节性模型，但是当数据中不存在季节性时，通常可以通过将搜索对象限制为非季节性模型，从而更快速地获得结果。虽然未检查各地区市场的数据，但我们通过标绘这五个市场的总订户数，可大体了解是否存在季节性的因素。

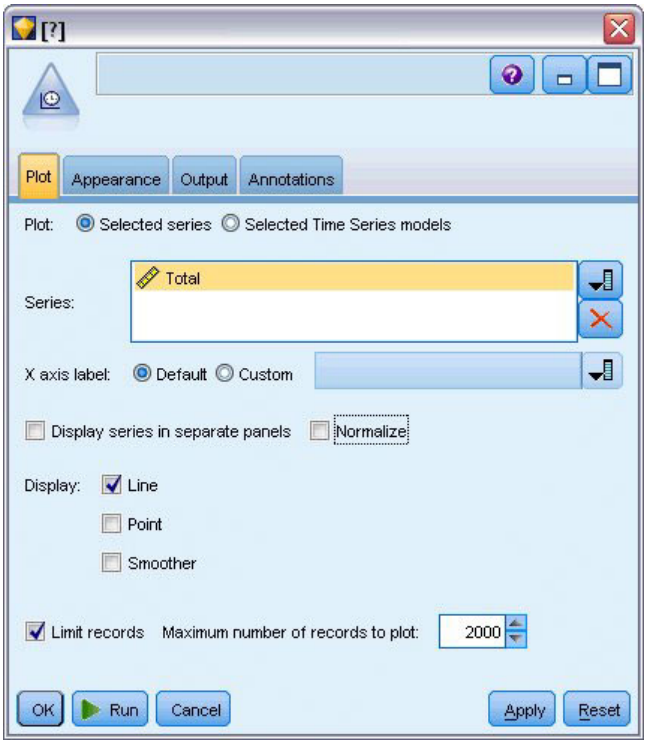


图 173. 绘制总订户数

1. 通过“图形”选用板，可将时间散点图节点附加到过滤节点中。
2. 将*总计*字段添加到“序列”列表。
3. 取消选择 **在单独面板中显示序列** 和 **标准化** 复选框。
4. 单击**运行**。

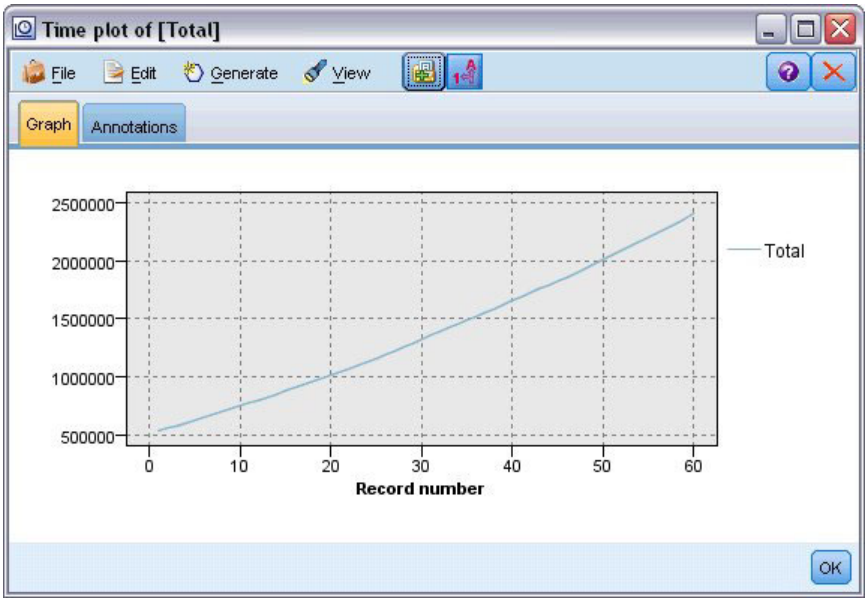


图 174. “总计”字段的时间图

该序列表现出非常平滑的上升趋势，并且无季节性变化的迹象。可能个别序列具有季节性，但通常季节性不是数据的突出特点。

当然，排除季节性模型前应检查每个序列。然后，可将表现出季节性的序列分离出来，并单独为它们建模。

通过 IBM SPSS Modeler，可以轻松地同时绘制多个序列。

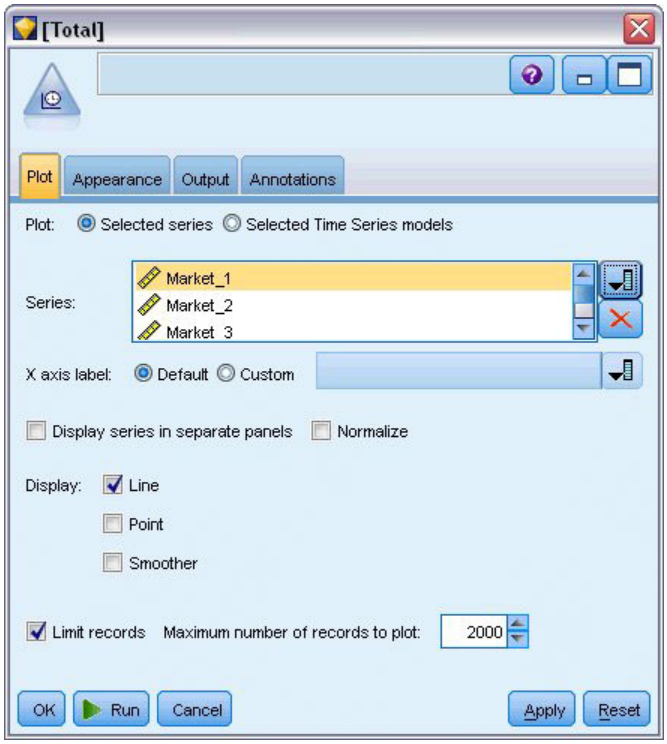


图 175. 绘制多个时间序列

5. 重新打开时间散点图节点。
6. 从“序列”列表中删除*总计*字段（将其选中，然后单击红色的 X 按钮）。
7. 将 *Market_1* 至 *Market_5* 字段添加到列表中。
8. 单击**运行**。

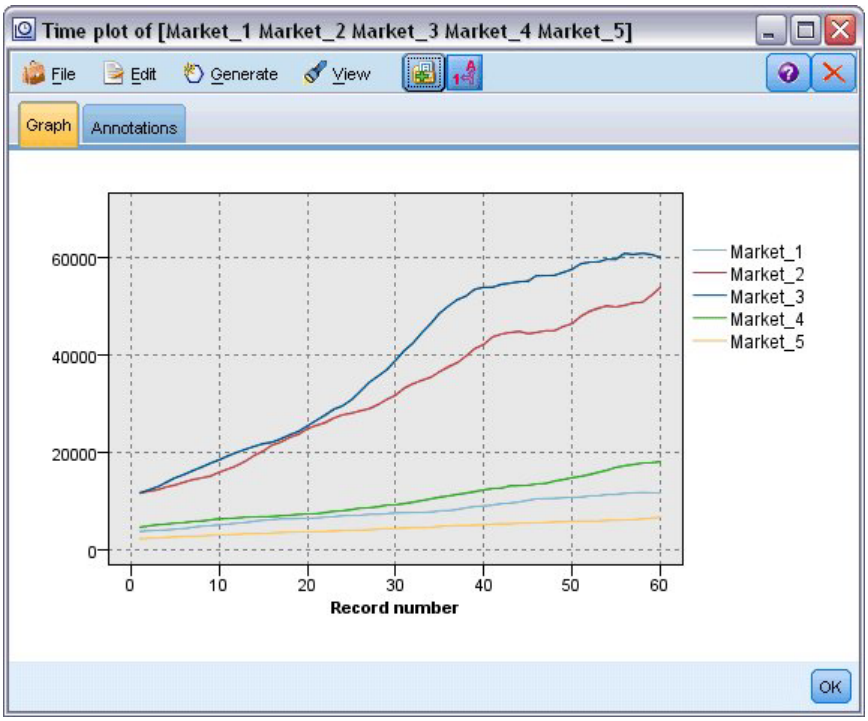


图 176. 多个字段的时间图

检查后发现每个市场的曲线均呈稳定上升趋势。虽然一些市场的曲线上升不如其他市场那么稳定，但也未表现出任何季节性趋势。

定义日期

现在需要将 *DATE_* 字段的存储类型更改为日期格式。

1. 将填充节点附加到过滤节点。
2. 打开填充节点并单击字段选择器按钮。
3. 选择 **DATE_** 并将它添加到 **填入字段** 。
4. 将 **替换** 条件设置为 **始终** 。
5. 将 **替换为** 的值设置为 **to_date(DATE_)** 。

图 177. 设置日期存储类型

更改缺省日期格式以与“日期”字段的格式相匹配。要确保“日期”字段的转换按预期执行，此操作是必需的。

6. 在菜单上，选择**工具 > 流属性 > 选项**，以显示“流选项”对话框。
7. 将缺省 **日期格式** 设置为 **MON YYYY** 。

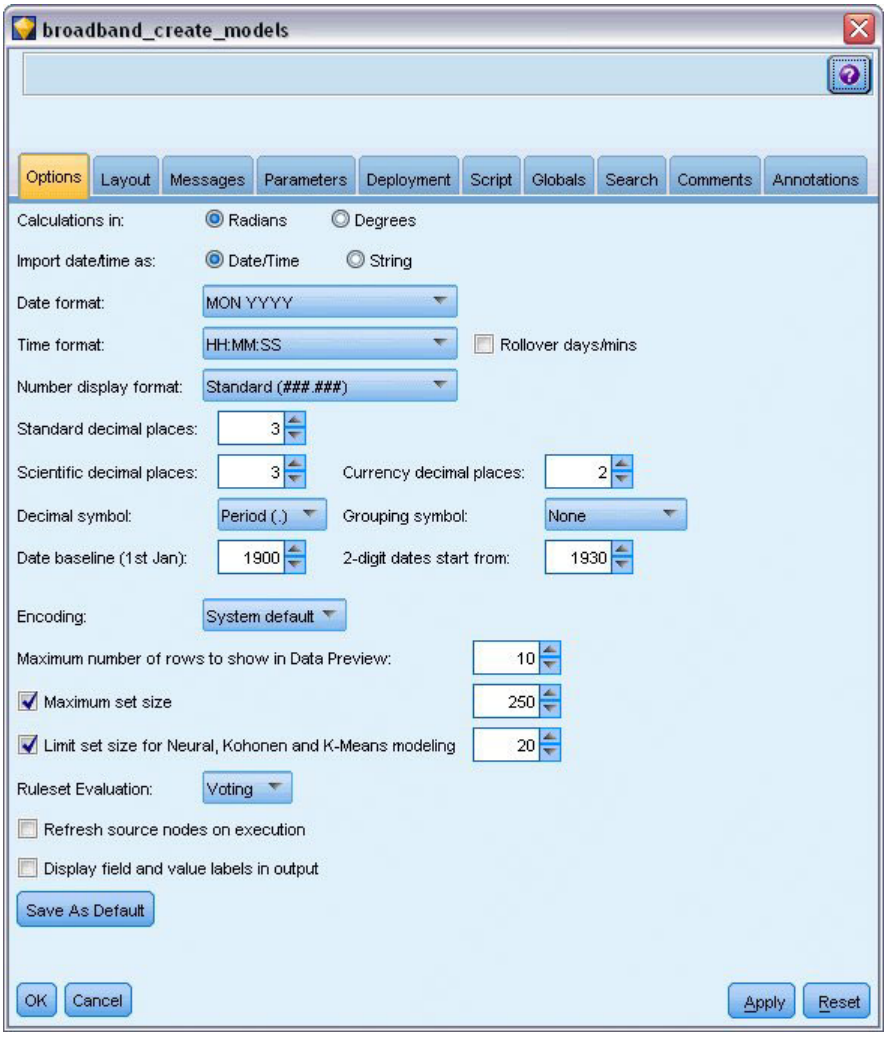


图 178. 设置日期格式

定义目标

1. 添加“类型”节点并将 *DATE_* 字段的角色设置为**无**。将所有其他字段（*Market_n* 字段以及*合计*字段）的角色设置为**目标**。
2. 单击**读取值**按钮以填充“值”列。

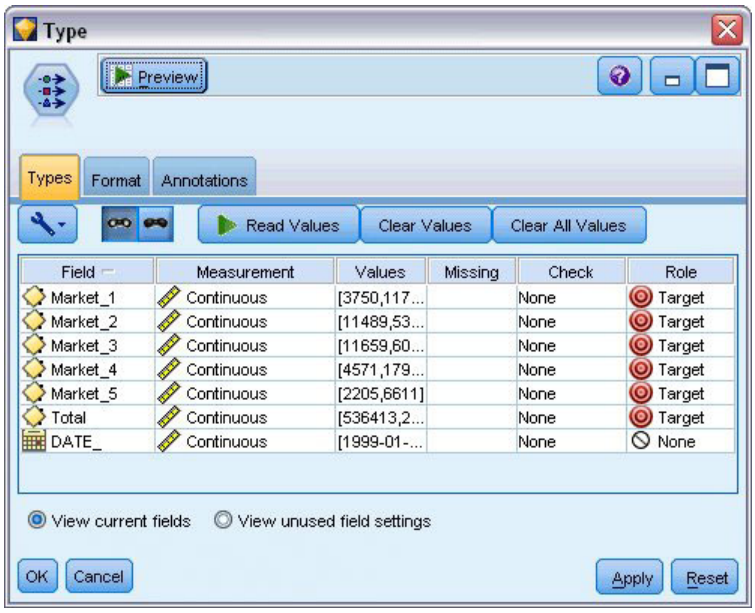


图 179. 设置多个字段的角色

设置时间区间

1. 添加时间区间节点（通过“字段操作”选用板）。
2. 在“区间”选项卡上，选择**月**作为时间区间。
3. 选中 **根据数据构建** 选项。
4. 选择 **DATE_** 作为构建字段。

图 180. 设置时间间隔

5. 在“预测”选项卡上，选中**将记录扩展到未来**复选框。
6. 将值设置为 **3** 。
7. 单击**确定**。

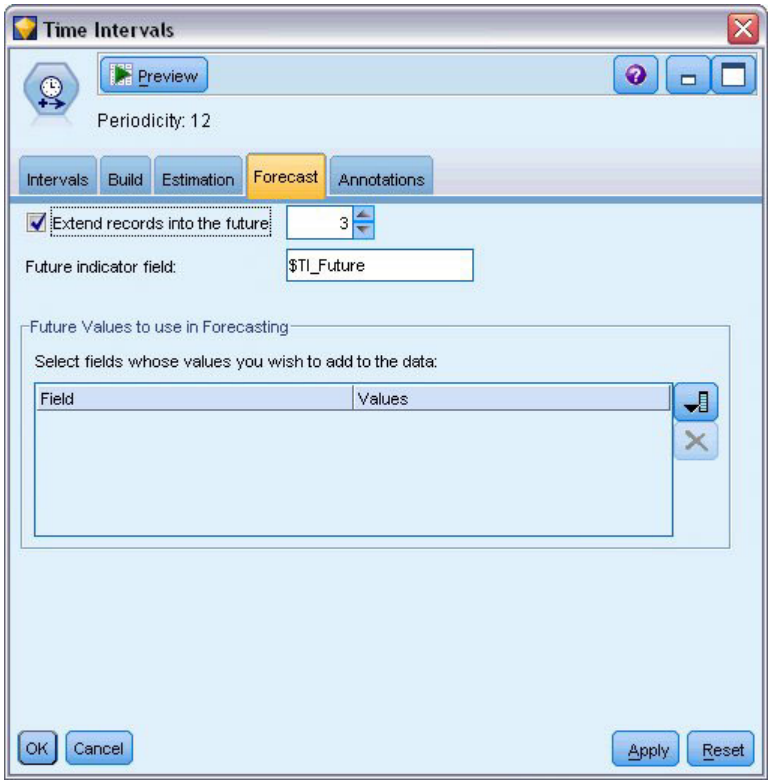


图 181. 设置预测期

创建模型

1. 从“建模”选用板中，将“时间序列”节点添加到流中，并将此节点附加到“时间间隔”节点。
2. 单击使用全部缺省设置的时间序列节点上的**运行**。此操作使专家建模器能够将最合适的模型用于每个时间序列。

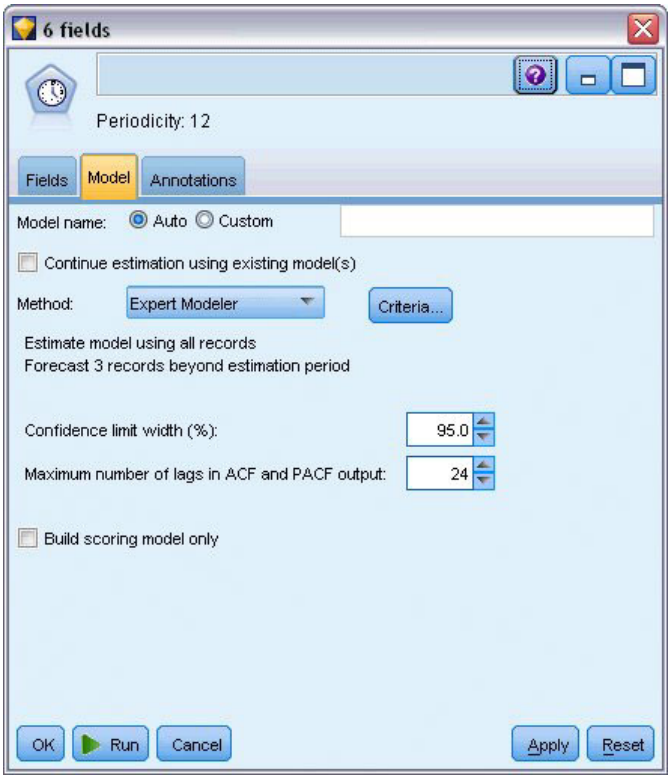


图 182. 针对时间序列选择专家建模器

3. 将时间序列模型块附加到时间区间节点。
4. 将“表”节点附加到时间序列模型并单击**运行**。

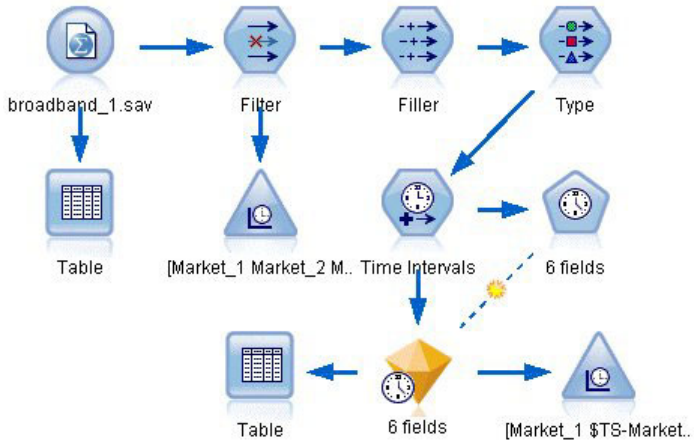


图 183. 用于显示时间序列建模的样本流

现在有三个新行（第 61 至 63 行）附加到原始数据中。这三行用于预测时限，在本例中为 2004 年 1 至 3 月。

现在还提供了几个新列，即“时间间隔”节点添加的 *$TI_* 列以及“时间序列”节点添加的 *$TS-* 列。这些列表示每行（也就是时间序列数据中的每个区间）的以下内容:

Column	描述
$TI_TimeIndex	此行的时间区间索引值。
$TI_TimeLabel	此行的时间区间标签。
$TI_Year	此行中生成数据的年份和月份指示符。
$TI_Month	
$TI_Count	确定此行的新数据时所涉及记录的数量。
$TI_Future	指明此行是否包含预测数据。
$TS-*colname*	由每列原始数据生成的模型数据。
$TSLCI-*colname*	每列生成的模型数据中的置信度区间下限值。
$TSUCI-*colname*	每列生成的模型数据中的置信度区间上限值。
$TS-Total	此行的 $TS-*colname* 值的总计。
$TSLCI-Total	此行的 $TSLCI-*colname* 值的总计。
$TSUCI-Total	此行的 $TSUCI-*colname* 值的总计。

对预测操作最重要的列是 *$TS-Market_n* 、 *$TSLCI-Market_n* 和 *$TSUCI-Market_n* 。特别是这些列的第 61 至 63 行，它们包含各个地区市场的用户预订预测数据和置信度区间。

检查模型

1. 双击时间序列模型块以显示有关为每个市场生成的模型的数据。

   请注意专家建模器如何选择通过为其他市场生成的类型来为市场 5 生成不同类型的模型。

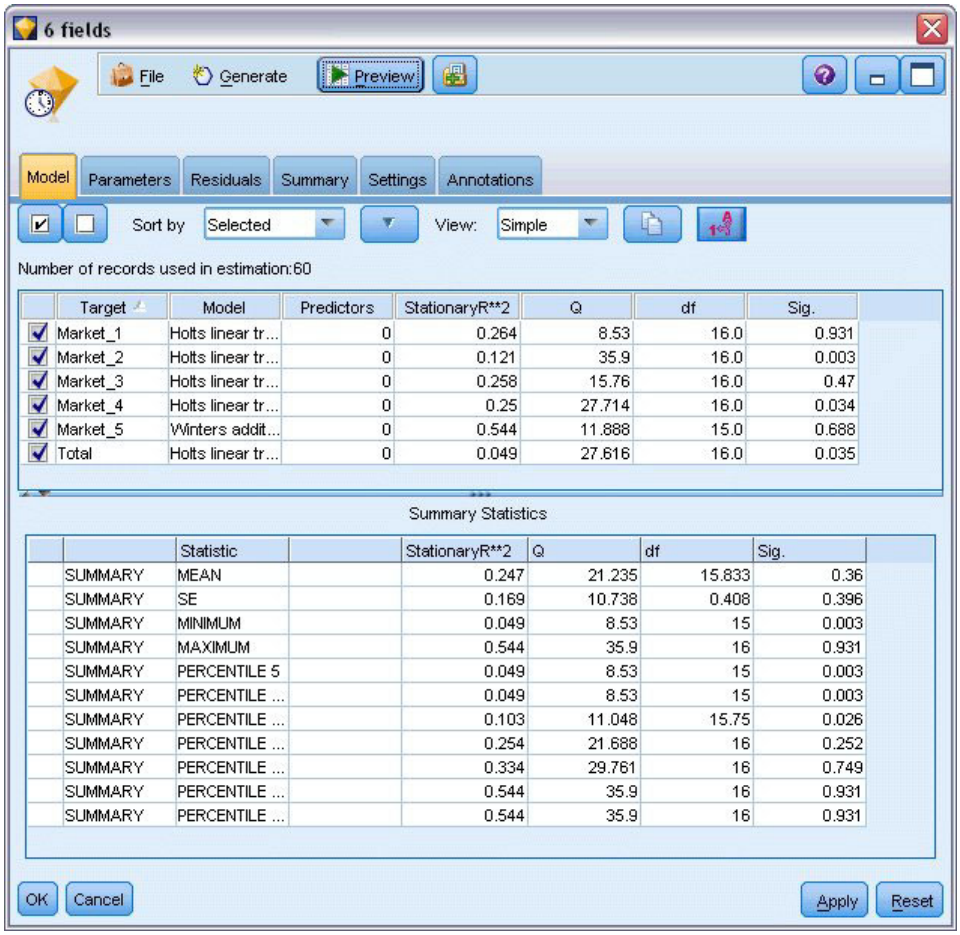


图 184. 为市场生成的时间序列模型

预测变量列用于显示用作各个目标预测变量的字段数量，在本例中为 0。

此视图中余下的列显示每个模型的不同拟合度测量值。**StationaryR**2** 列显示的是固定的 *R* 平方值。此统计量是序列中由模型解释的总变异所占比例的估计值。该值越高（最大值为 1.0），则模型拟合会越好。

Q、**df** 和 **Sig.** 列与 Ljung-Box 统计量相关联，该检验是对模型中残差错误的随机检验；错误的随机性越大，则模型会变得越好。**Q** 是 Ljung-Box 统计量自身，而 **df**（自由度）表示估算特定目标时可以自由改变的模型参数的数目。

Sig. 列给出了 Ljung-Box 统计量的显著性值，从而以另一种方式来表示指定的模型是否正确。显著性值小于 0.05 表示残差误差不是随机的，则意味着所观测的序列中存在模型无法解释的结构。

如果将固定 *R* 平方值和显著性值考虑在内，则专家建模器为 *Market_1*、*Market_3* 和 *Market_5* 选择的模型完全可以接受。*Market_2* 和 *Market_4* 的 **Sig.** 值均小于 0.05，表明可能还必须进行一些实验，以便为这些市场找到拟合度更好的模型。

屏幕下方的汇总值提供了有关这些统计量在所有模型中的分布情况的信息。例如，所有模型的固定 *R* 平方均值为 0.247，而此值的最小值为 0.049（ *总计* 模型的值），最大值为 0.544（ *Market_5* 的值）。

SE 表示每个统计量在所有模型中的标准误差。例如，固定 *R* 平方值在所有模型中的标准误差为 0.169。

汇总部分还包括百分位值，这些值提供了有关统计量在模型中的分布情况的信息。对于每个百分位数，该百分比模型的拟合统计量具有比所述值低的值。

例如，仅 25% 的模型的固定 *R* 平方值低于 0.121。

2. 单击“视图”下拉列表并选择**高级**。

屏幕上将显示多个其他拟合度测量值。**R**2** 为 *R* 的平方值，即时间序列中可由模型解释的总变化估算值。由于此统计量的最大值为 1.0，因此在这一点上，我们的模型表现不错。

图 185. 时间序列模型的高级显示

RMSE 指均方根误差，它用于度量序列的实际值与模型预测值之间的差异，以序列自身所用单位表示。由于这是误差测量值，因此该值越小越好。乍一看， *Market_2* 和 *Market_3* 的模型的成功率低于其他三个市场，但根据目前所观察的统计量仍可以接受。

其他的拟合度测量值包括均值绝对百分比误差（ **MAPE** ）和其最大值（ **MaxAPE** ）。绝对误差百分比用于度量目标序列与其模型预测水平的差异程度，以百分比值表示。通过审查所有模型中的均值和最大值，可以大概知道预测的不确定性程度。

MAPE 值显示所有模型的平均不确定性都低于 1%，这是一个很低的值。MaxAPE 值显示最大绝对误差百分比，并且对设想预测的最坏情况很有帮助。它显示，每个模型的最大百分比误差大约在 1.8% 至 2.5% 之间，仍然是一组很低的数字。

MAE（平均绝对误差）值用于显示预测误差绝对值的均值。类似于 RMSE 值，此值以序列自身所用单位表示。**MaxAE** 显示以同一单位表示的最大预测误差，并指示预测的最坏情况。

尽管关注的是这些绝对值，但由于目标序列表示的是不同规模市场的订户量，因此在此情况下百分比误差值（MAPE 和 MaxAPE）更有用。

MAPE 和 MaxAPE 值可以使用模型表示可接受的不确定性吗？毫无疑问这些值很小。由于可接受风险因问题的不同而有所变化，因此商业意识可以在此派上用场。假设拟合度统计在可接受的范围之内，然后继续查看残差错误。

相对于仅查看拟合度统计，检查模型残差的自相关函数 (ACF) 和偏自相关函数 (PACF) 的值可以通过更量化的方式深入了解模型。

合理指定的时间模型将捕获所有非随机的变异，其中包括季节性、趋势、循环周期以及其他重要的因素。如果是这种情况，那么任何误差都不应随时间的推移与其自身相关联（自相关）。这两个自相关函数中的显著结构都可以表明基础模型不完整。

3. 单击“残差”选项卡以显示第一个地区市场的模型中残差的自相关函数 (ACF) 和偏自相关函数 (PACF) 的值。

图 186. 市场的 ACF 值和 PACF 值

在这些图中，已将误差变量的原始值延迟多达 24 个时间段并与原始值进行比较，以确定随着时间推移是否存在任何相关性。要使模型可接受，上 (ACF) 图中的条形在正（上）方向或负（下）方向均不应扩展到阴影区之外。

如果出现此情况，您需要检查下 (PACF) 散点图，以了解是否已确认此处的结构。PACF 散点图主要关注在控制插入时间点的序列值之后的相关性。

Market_1 的值都位于阴影区之内，因此我们可以继续检查其他市场的值。

4. 单击 **显示模型散点图** 下拉列表可显示其他地区市场和总体市场的 ACF 值和 PACF 值。

由于 *Market_2* 和 *Market_4* 的值不太重要，因此可以确认早先对 **Sig.** 值所产生的可疑因素。我们需要在某些时间试验这些市场的其他一些不同的模型，以了解是否可以获取更佳的拟合，但对于该示例的剩余部分，应更多地考虑可以从 *Market_1* 模型中获取的其他内容。

5. 在“图形”选用板中，将“时间图”节点附加到时间序列模型块。
6. 在“图”选项卡上，取消选中**在单独面板中显示系列**复选框。
7. 在 **系列** 列表上，单击字段选择器按钮，选定 *Market_1* 和 *$TS-Market_1* 字段，然后单击 **确定** 将它们添加到列表中。
8. 单击**运行**，以显示第一个地区市场的实际数据和预测数据的线图。

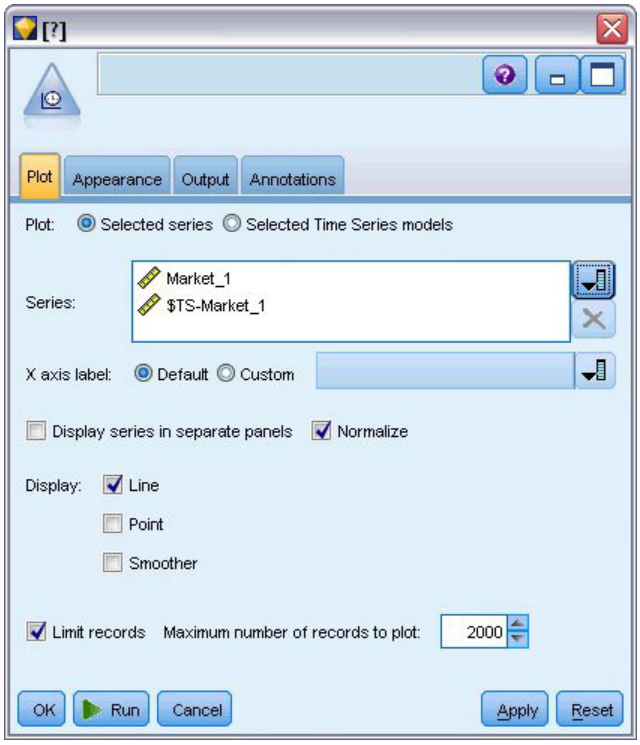


图 187. 选择要绘制的字段

请注意预测（ *$TS-Market_1* ）线如何通过实际数据的末端向外延伸。目前已得出对此市场未来三个月的预期需求的预测。

整个时间序列上的实际数据线和预测数据线在图上非常接近，表明对此特定时间序列这是一个可靠的模型。

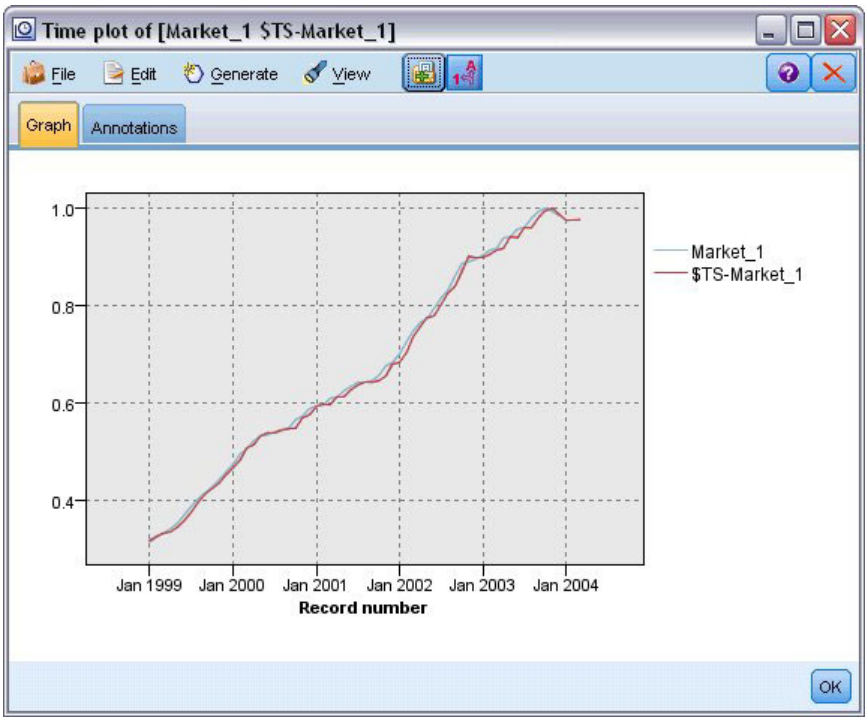


图 188. Market_1 的实际数据和预测数据的时间图

将模型保存在文件中，以便在将来的示例中使用:

9. 单击 **确定** 关闭当前图形。
10. 打开时间序列模型块。
11. 选择 **文件 > 保存节点** 并指定文件位置。
12. 单击**保存**。

现在虽然有了此特定市场的可靠模型，但该预测的误差到底有多大呢？可通过检查置信度区间得到预测的误差大小。

13. 双击流中最后的时间散点图（标注为 **Market_1 $TS-Market_1** ），以重新打开该节点的对话框。
14. 单击字段选择器按钮并将 *$TSLCI-Market_1* 和 *$TSUCI-Market_1* 字段添加到 **系列** 列表中。
15. 单击**运行**。

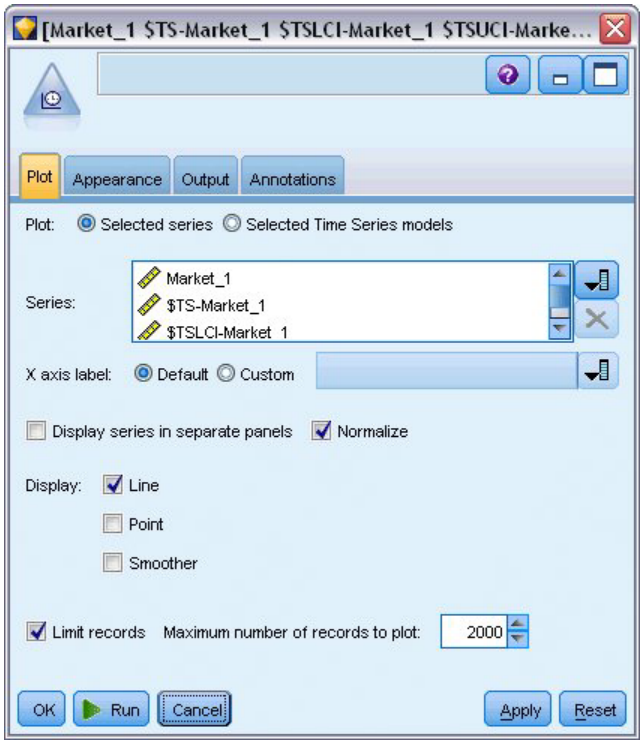


图 189. 添加更多要绘制的字段

现在有了与以前一样的图形，但添加了置信度区间上限（ *$TSUCI* ）和下限（ *$TSLCI* ）。

请注意置信度区间的边界如何随预测时限而分叉，这表示预测越指向更远的将来，不确定性就变得越来越大。

但是，随着每个时段的流逝，您就会多一个时段（在本例中为月）的实际使用率数据作为预测的依据。您可以将这些新数据读入流中，并再次应用您的模型，因为您知道它是可靠的。请参阅主题第 166 页的『重新应用时间序列模型』以获取更多信息。

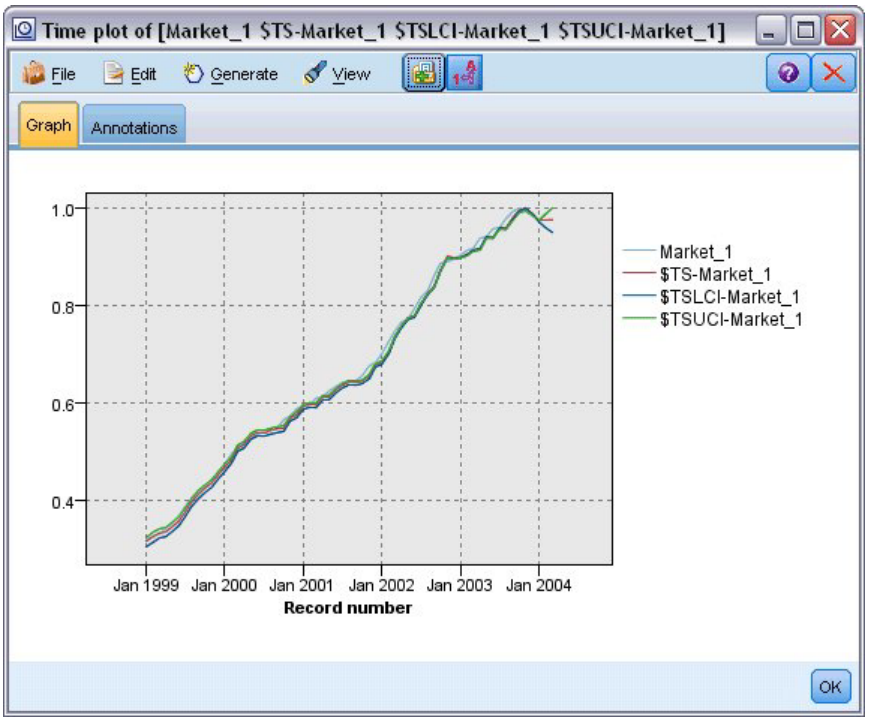


图 190. 添加了置信区间的时间图

摘要

您已学习了如何使用专家建模器为多个时间序列生成预测，并已将得到的模型保存到外部文件中。

在下一个示例中，您将看到如何将非标准时间序列数据转换为适合输入到时间序列节点的格式。

重新应用时间序列模型

本示例应用第一个时间序列示例中的时间序列模型，但也可以独立使用。请参阅主题第 147 页的『使用时间序列节点进行预测』以获取更多信息。

与原来的情形一样，为了预测带宽需求，某个国内宽带提供商的一位分析员需要为多个地区市场中的每个市场生成对用户预订的月度预测。已使用专家建模器创建了模型并且要对未来三个月进行预测。

由于现在已使用原始预测期的实际数据更新了您的数据仓库，因而您想使用这些数据将预测时间范围再延长三个月。

此示例使用名为 *broadband_apply_models.str* 的流，该流引用名为 *broadband_2.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *broadband_apply_models.str* 位于 *streams* 文件夹中。

检索流

在此示例中，您将根据第一个示例中保存的时间序列模型重新创建“时间序列”节点。如果未保存模型也不用担心，因为 *Demos* 文件夹中提供了一个模型。

1. 从 *Demos* 下的 *streams* 文件夹中打开流 *broadband_apply_models.str*。

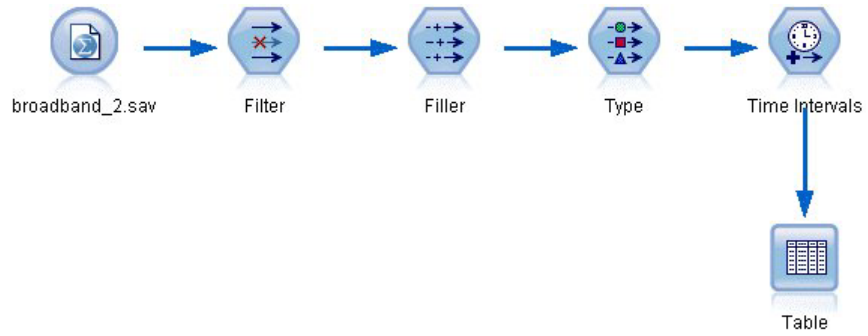
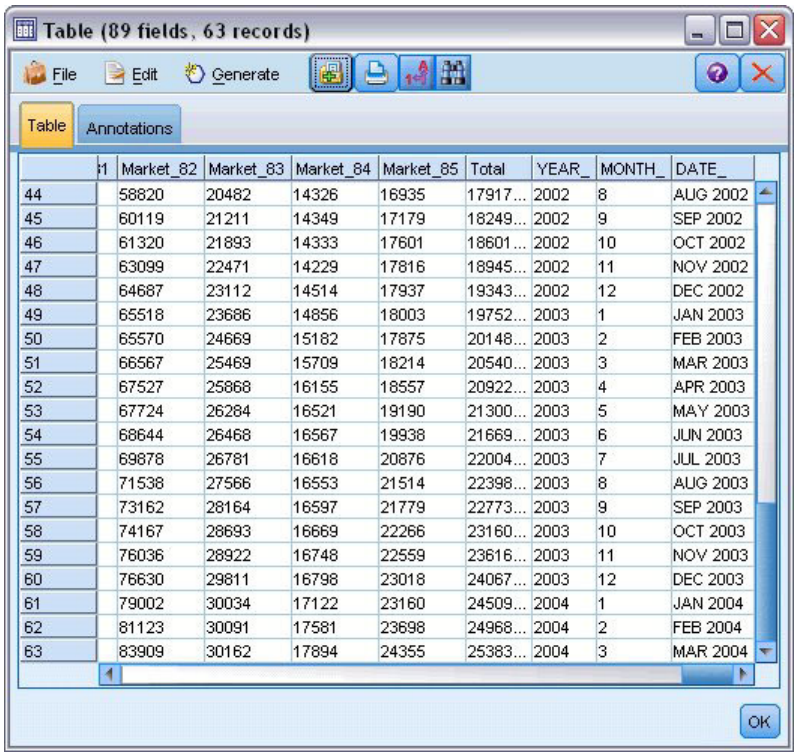


图 191. 打开流



	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

图 192. 已更新的销售数据

更新过的月度数据收集在 *broadband_2.sav* 中。

2. 将表节点附加到 IBM SPSS Statistics 文件源节点，打开表节点并单击**运行**。

注: 已使用 2004 年 1 至 3 月份（第 61 至 63 行）的实际销售数据对数据文件进行了更新。

3. 打开流上的时间区间节点。
4. 单击 **预测** 选项卡。
5. 确保 **将记录扩展到未来** 设置为 **3** 。

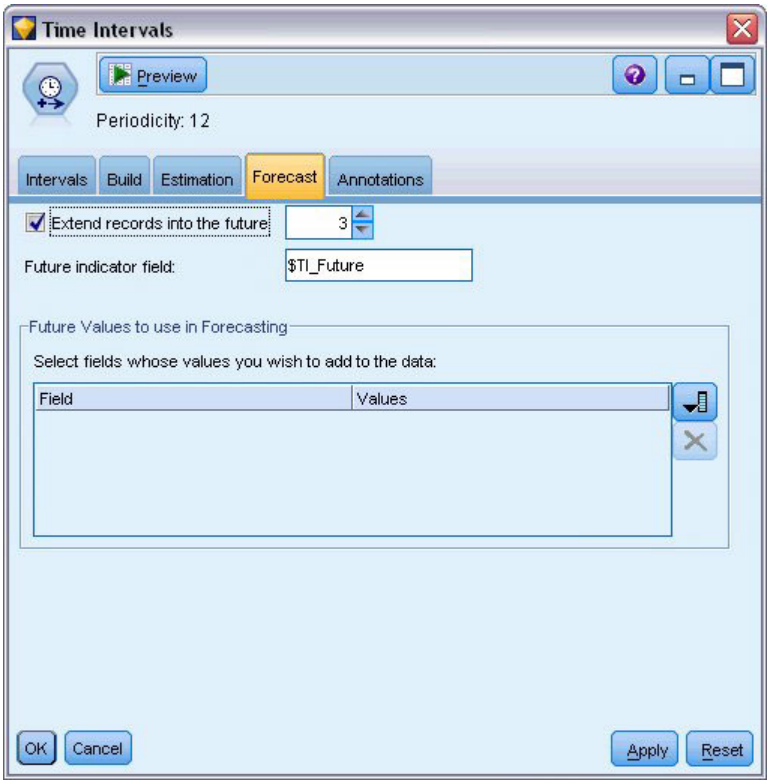


图 193. 检查预测期的设置

检索保存的模型

1. 在 IBM SPSS Modeler 菜单上，选择 **插入 > 来自文件的节点** ，并从 *Demos* 文件夹中选择 *TSmodel.nod* 文件（或使用在第一个时间序列示例中保存的时间序列模型）。

此文件包含来自上一个示例的时间序列模型。插入操作将把对应的时间序列模型块放置在工作区上。

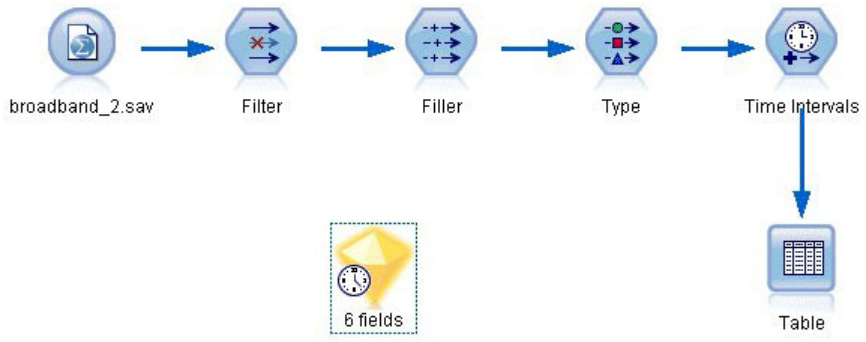


图 194. 添加模型块

生成建模节点

1. 打开时间序列模型块并选择 **生成 > 生成建模节点** 。

此操作将把时间序列建模节点放置在工作区上。

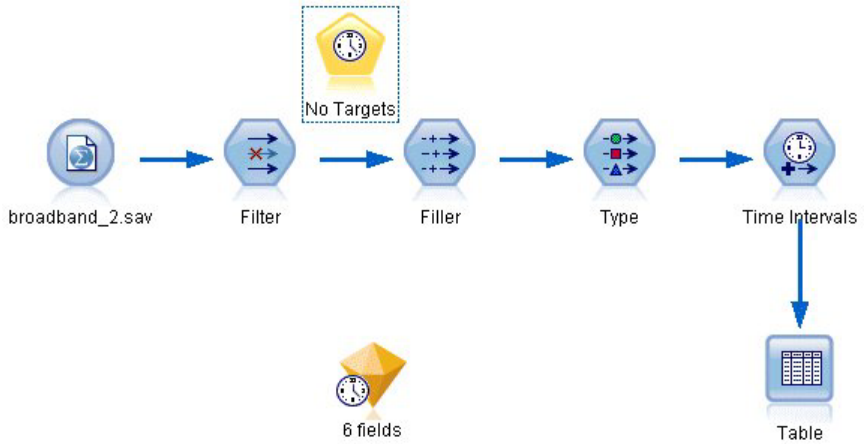


图 195. 从模型块生成建模节点

生成新模型

1. 关闭时间序列模型块并将它从工作区中删除。

   旧模型以 60 行数据为基础构建。现在需要基于更新过的销售数据（63 行）来生成新模型。

2. 将新生成的时间序列构建节点附加到流。

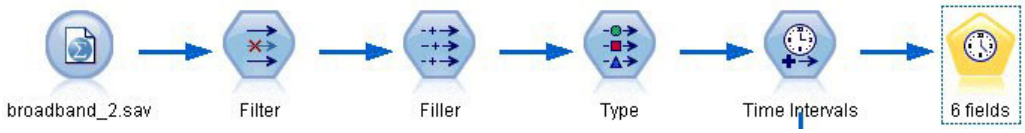


图 196. 将建模节点附加到流

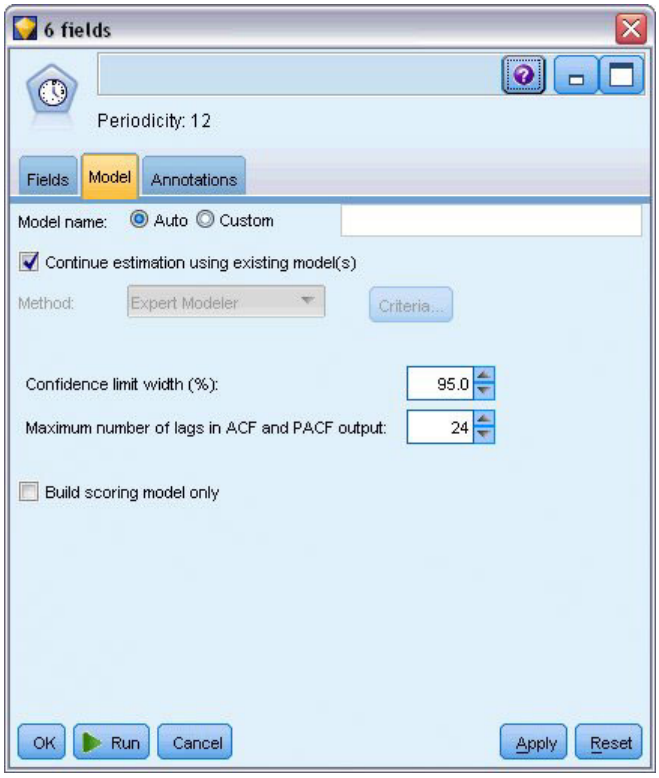
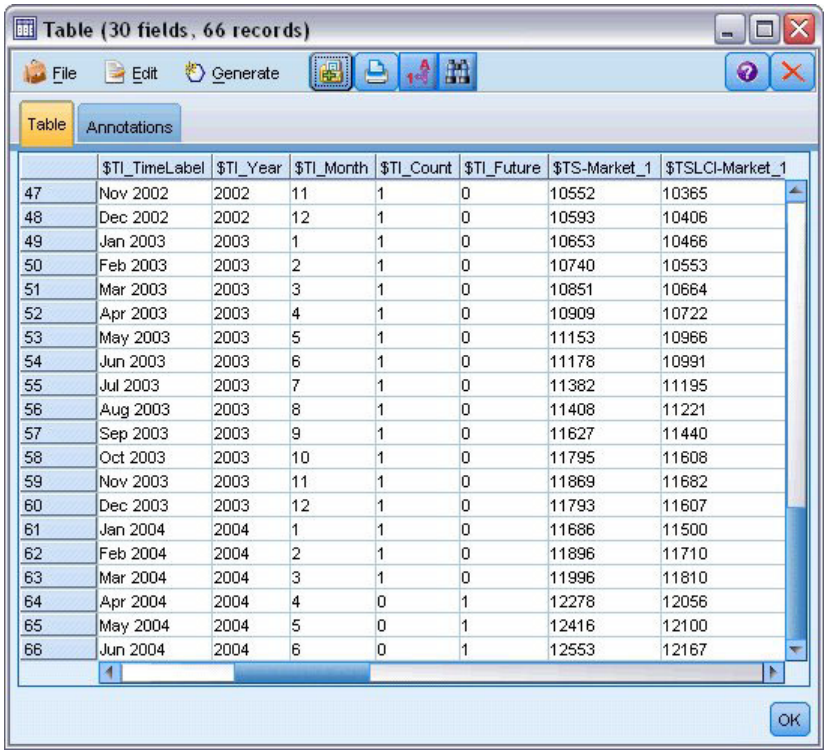


图 197. 复用已存储的时间序列模型设置

3. 打开时间序列节点。
4. 在 **模型** 选项卡上，请确保已选中 **使用现有模型继续估计** 。
5. 单击**运行**以将新的模型块放在工作区及“模型”选用板中。

检查新模型



	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	Nov 2002	2002	11	1	0	10552	10365
48	Dec 2002	2002	12	1	0	10593	10406
49	Jan 2003	2003	1	1	0	10653	10466
50	Feb 2003	2003	2	1	0	10740	10553
51	Mar 2003	2003	3	1	0	10851	10664
52	Apr 2003	2003	4	1	0	10909	10722
53	May 2003	2003	5	1	0	11153	10966
54	Jun 2003	2003	6	1	0	11178	10991
55	Jul 2003	2003	7	1	0	11382	11195
56	Aug 2003	2003	8	1	0	11408	11221
57	Sep 2003	2003	9	1	0	11627	11440
58	Oct 2003	2003	10	1	0	11795	11608
59	Nov 2003	2003	11	1	0	11869	11682
60	Dec 2003	2003	12	1	0	11793	11607
61	Jan 2004	2004	1	1	0	11686	11500
62	Feb 2004	2004	2	1	0	11896	11710
63	Mar 2004	2004	3	1	0	11996	11810
64	Apr 2004	2004	4	0	1	12278	12056
65	May 2004	2004	5	0	1	12416	12100
66	Jun 2004	2004	6	0	1	12553	12167

图 198. 显示新预测的表

1. 将表节点附加到工作区中新的时间序列模型块。
2. 打开“表”节点，然后单击**运行**。

由于复用的是已存储的设置，因此新模型仍然会对未来三个月进行预测。不过，此次的预测时限为 4 至 6 月，因为估计期（在时间区间节点上指定）现在是在 3 月而不是 1 月结束。

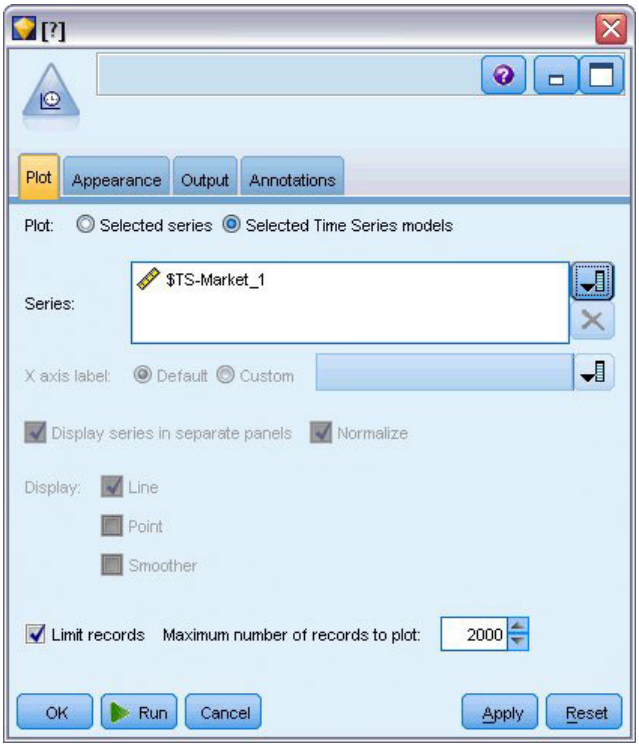


图 199. 指定要绘制的字段

3. 将“时间图”图形节点附加到时间序列模型块。

   此次将使用专为时间序列模型设计的时间散点图显示方式。

4. 在“散点图”选项卡上，选择**选定的时间序列模型**选项。
5. 在 **系列** 列表上，单击字段选择器按钮，选定 *$TS-Market_* 1 字段，然后单击 **确定** 将它添加到列表中。
6. 单击**运行**。

现在的图形显示的是截止 2004 年 3 月 *Market_1* 的实际销售量，以及截止 2004 年 6 月的预测销售量及置信度区间（蓝色阴影区）。

与第一个示例中一样，在整个预测时限内，预测值与实际数据贴得很紧，再次表明构建的是一个比较理想的模型。

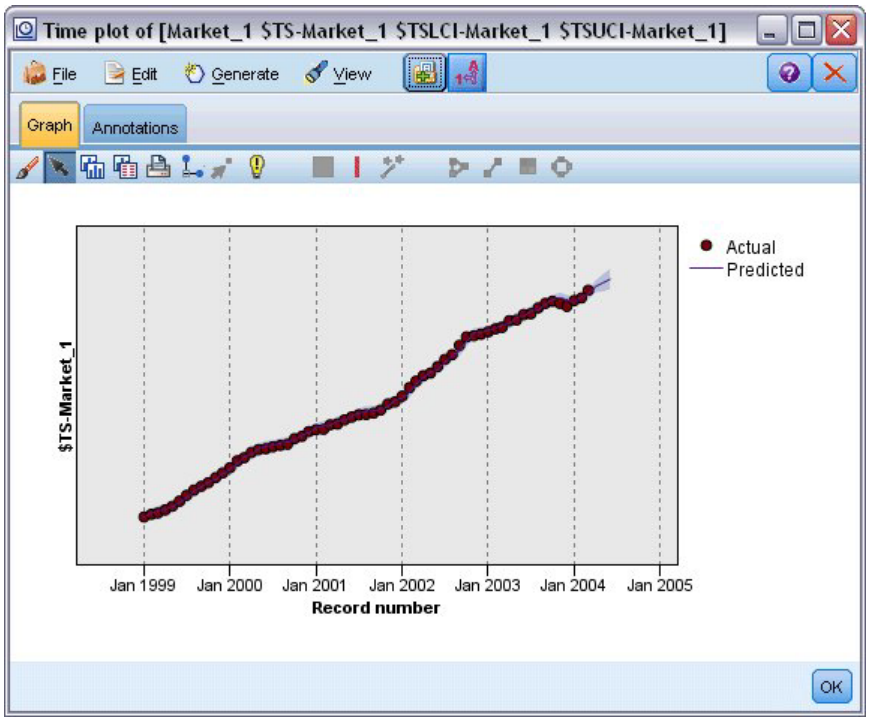


图 200. 已延长到六月份的预测

摘要

上面学习了有更多当前数据可用时如何应用保存的模型以扩展以前的预测，并可在无需重新构建模型的情况下实现这一点。当然，如果有理由认为模型已改变，则应重新构建模型。

第 15 章 预测产品分类销售情况（时间序列）

产品分类公司会根据过去 10 年的销售数据来预测其男装类的月度销售情况。

此示例使用名为 *catalog_forecast.str* 的流，此流引用名为 *catalog_seasfac.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *catalog_forecast.str* 位于 *streams* 目录中。

在前一个示例中，我们已了解如何让专家建模器确定最适合于您的时间序列的模型。现在我们来深入了解两种可用于选择模型的方法 – 指数平滑与 ARIMA。

为了帮您找到适当的模型，最好先绘制时间序列。时间序列的可视化检查通常可以有效地指导并帮助您进行选择。另外，您需要弄清以下几点:

- 此序列是否具有总体趋势？如果是，趋势是持续存在还是将随时间推移而消逝？
- 此序列是否显示季节性？如果是，那么这种季节的波动是随时间而加剧还是持续稳定存在？

创建流

1. 新建流并添加指向 *catalog_seasfac.sav* 的“Statistics 文件”源节点。

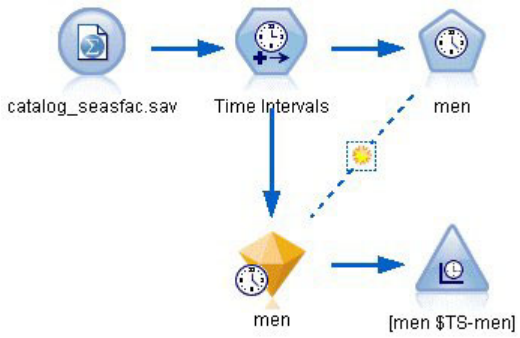


图 201. 预测产品分类销售情况

图 202. 指定目标字段

2. 打开 IBM SPSS Statistics 文件源节点并选择“类型”选项卡。
3. 单击 **读取值** ，然后单击 **确定** 。
4. 单击*角色*列（在*男*字段中），将角色设置为**目标**。
5. 将所有其他字段的角色设置为**无**，然后单击**确定**。

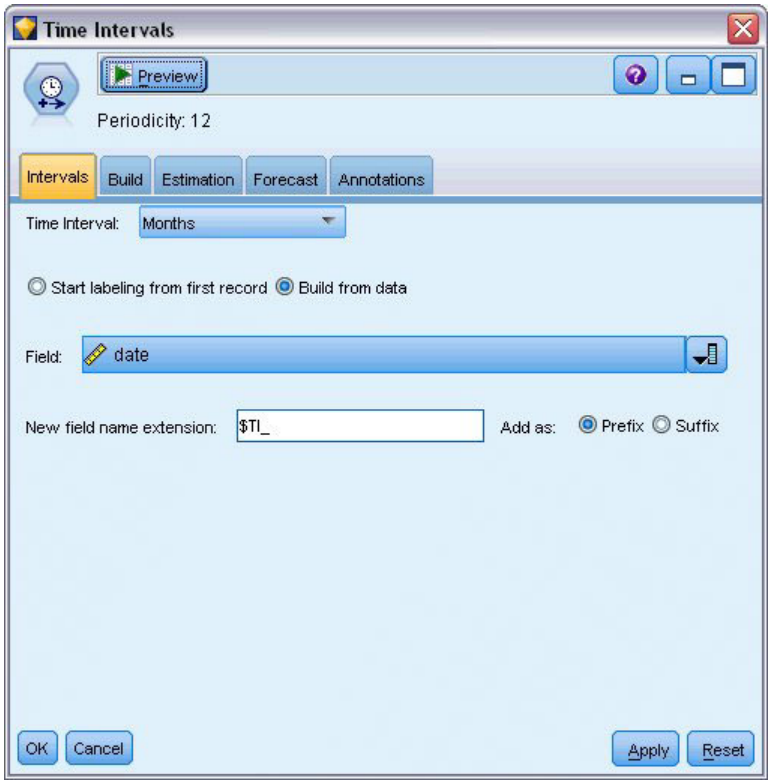


图 203. 设置时间间隔

6. 将时间区间节点添加到 IBM SPSS Statistics 文件源节点。
7. 打开时间区间节点，然后将 **时间区间** 设置为 **月** 。
8. 选择 **从数据构建** 。
9. 将 **字段** 设置为 **日期** ，然后单击 **确定** 。

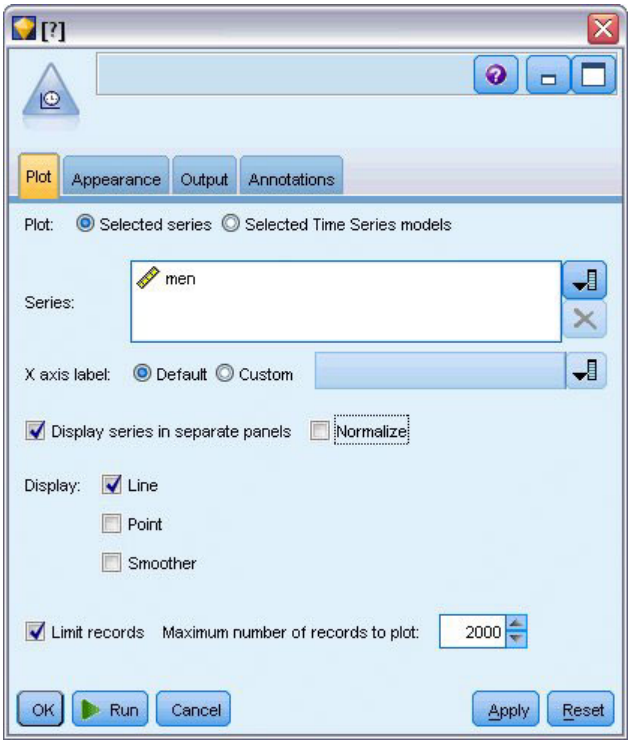


图 204. 绘制时间序列

10. 将时间散点图节点添加到时间区间节点。
11. 在“散点图”选项卡上，将**男**添加到序列列表。
12. 取消选择 **标准化** 复选框。
13. 单击**运行**。

检查数据

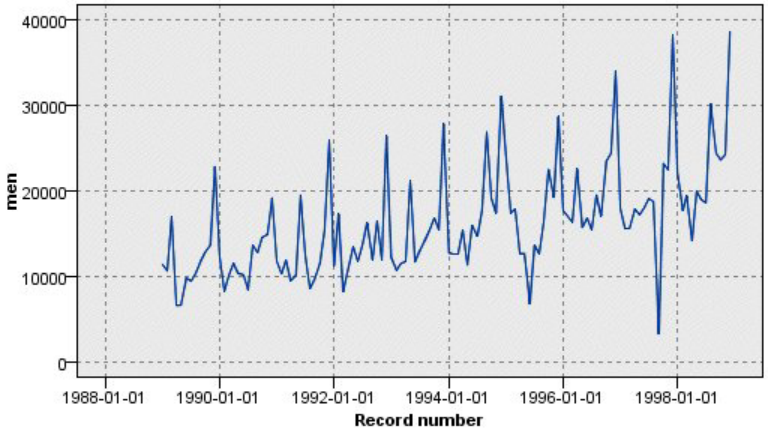


图 205. 男装的实际销售情况

此序列显示整体上升趋势，即序列值趋向于随时间变化而增加。上升趋势似乎将持续，即为线性趋势。

此序列还有一个明显的季节模式，即年度高点在十二月（如图形中的垂直线所示）。季节变化显示随上升序列而增长的趋势，表明是乘法季节模型而不是加法季节模型。

1. 单击 **确定** 以关闭此散点图。

由于您已了解此序列的特征，因此可以开始尝试对其进行建模。指数平滑法有助于预测存在趋势和/或季节性的序列。如您所见，此处数据同时体现上述两种特征。

指数平滑法

构建最佳指数平滑法模型包括确定模型类型（此模型是否需要包含趋势和/或季节性），然后获取选定模型的最佳参数。

随着时间的推移，男装销售情况图建议您使用同时包含线性趋势成分和乘法季节性成分的模型。这暗示使用温特斯模型。但是，我们先探究一个简单模型（既无趋势也无季节性），然后探究一个霍特模型（包含线性趋势，但无季节性）。此操作将让您了解模型在什么时候不适合数据，这是成功构建模型的基本技巧。

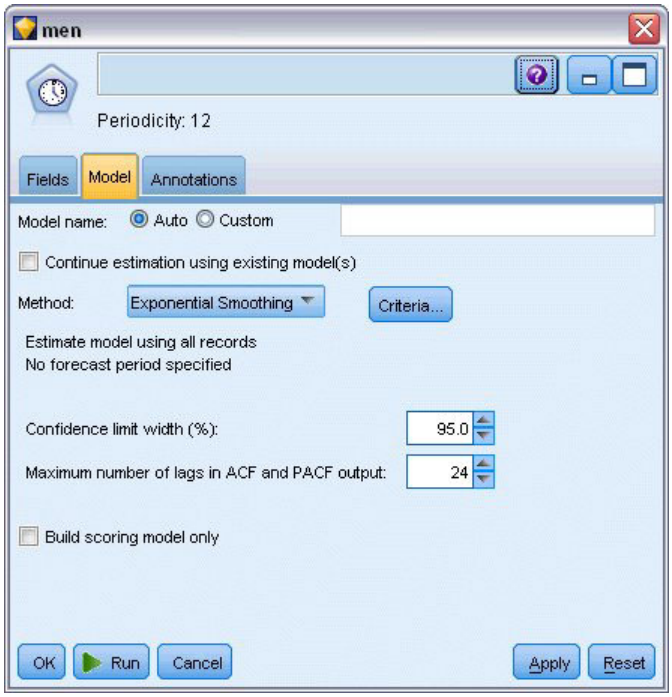


图 206. 指定指数平滑法

下面我们将开始构建一个简单的指数平滑法模型。

1. 将时间序列节点添加到时间区间节点。
2. 在 **模型** 选项卡中，将 **方法** 设置为 **指数平滑** 。
3. 单击**运行**创建模型块。

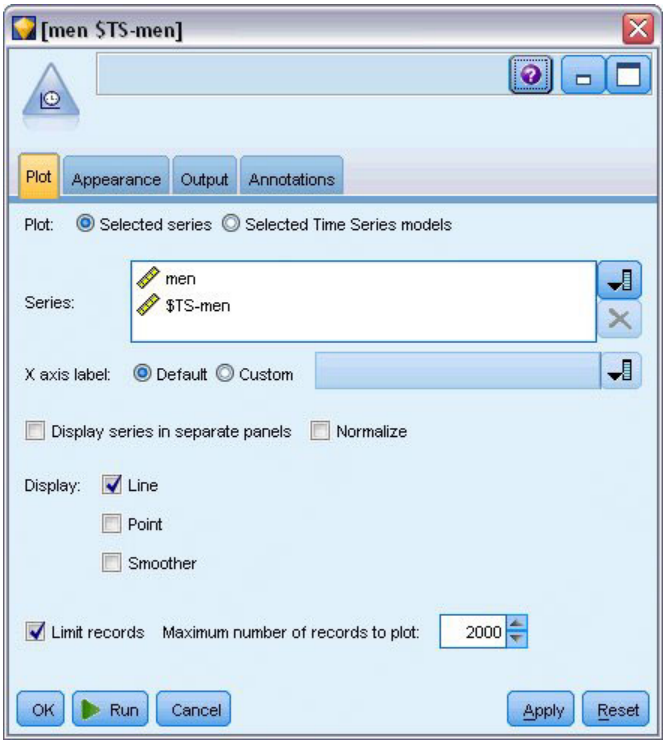


图 207. 绘制时间序列模型

4. 将时间散点图节点附加到模型块。
5. 在 **散点图** 选项卡中，将 *男* 和 *$TS-men* 添加到 **序列** 列表。
6. 取消选择 **在单独面板中显示序列** 和 **标准化** 复选框。
7. 单击**运行**。

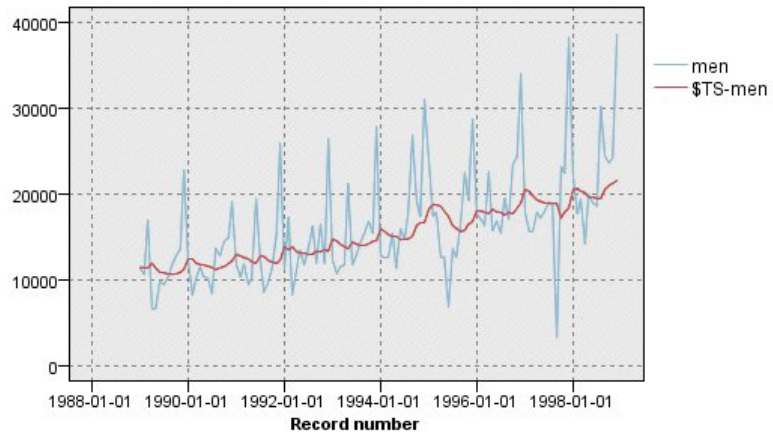


图 208. 简单指数平滑法模型

在散点图中， **男** 表示实际数据， **$TS-men** 则表示时间序列模型。

实际上，虽然简单模型确实显示了渐进（并且十分冗长的）上升趋势，但它并未考虑季节性。您完全可以拒绝此模型。

8. 单击 **确定** 关闭时间散点图窗口。

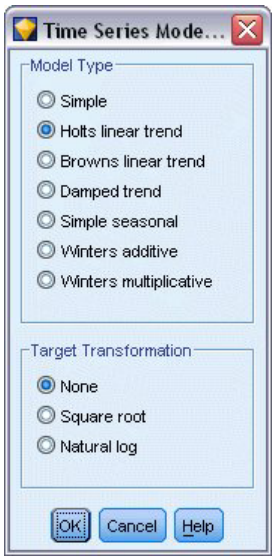


图 209. 选择霍特模型

下面试着做一个霍特线性模型。虽然，此模型的趋势性会比简单模型稍强，但它同样无法捕捉季节。

9. 重新打开时间序列节点。
10. 在 **模型** 选项卡中（依然选择 **指数平滑** 方法），单击 **标准** 。
11. 在“指数平滑标准”对话框中，选择**霍特线性趋势**。
12. 单击 **确定** 关闭此对话框。
13. 单击**运行**以再次创建模型块。
14. 再次打开时间散点图节点并单击**运行**。

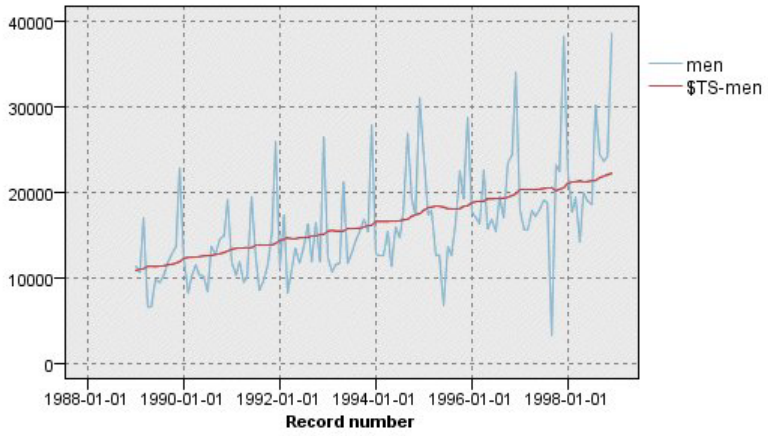


图 210. 霍特线性趋势模型

虽然，霍特模型显示比简单模型更强的平滑趋势，但它仍未考虑季节，所以还应放弃此模型。

15. 关闭时间图窗口。

随着时间的推移，男装销售散点图建议您使用同时包含线性趋势和乘法季节的模型，您可以恢复最初的男装销售散点图。因此 Winters 模型才是更适合的备选方案。

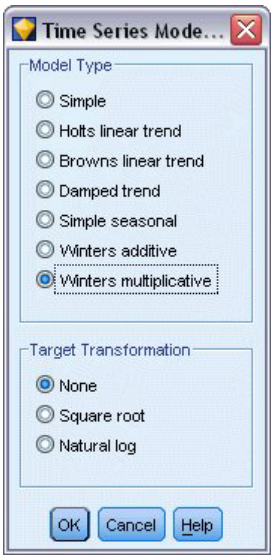


图 211. 选择温特斯模型

16. 重新打开时间序列节点。
17. 在 **模型** 选项卡中（依然选择 **指数平滑** 方法），单击 **标准** 。
18. 在“指数平滑标准”对话框中，选择 **Winters 乘法**。
19. 单击 **确定** 关闭此对话框。
20. 单击**运行**以再次创建模型块。
21. 打开时间散点图节点并单击**运行**。

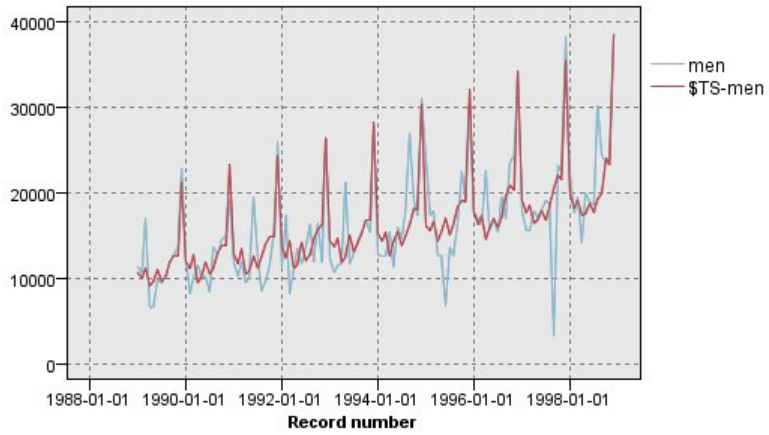


图 212. 温特斯乘法模型

此模型看起来更合适，因为它同时反映了数据的趋势和季节性。

此数据集的时间跨度为 10 年，并且包含 10 个季节峰值（出现在每年十二月份）。这 10 个峰值表示与实际数据中的 10 个年度峰值完全匹配的预测结果。

但此结果同时突出了“指数平滑法”步骤的局限性。看看上升和下降的峰值，还有一些重要结构没有得到解释。

如果您主要关注对包含季节性变化的长期趋势进行建模，那么指数平滑法是明智的选择。要构建此类结构较复杂的模型，则需要考虑使用 ARIMA 步骤。

ARIMA

ARIMA 过程使您可以创建一个适用于时间序列的微调建模的差分自回归移动平均值 (ARIMA) 模型。ARIMA 模型构建趋势和季节模型的方法比构建指数平滑模型更复杂，并新增了在模型中包含预测变量的功能。

继续以要开发预测模型的产品分类公司为例，我们了解了公司如何通过多个可用于解释某些销售变化情况的序列来收集男装的月度销售数据。预测变量可包括：邮递的产品目录数和产品目录的页数、开通的订购热线数目、印刷广告投入额以及客户服务代表人数。

这些预测变量是否对预测都有用？包含预测变量的模型是否优于不包含预测变量的模型？通过 ARIMA 步骤，我们可以创建一个包含预测变量的预测模型，然后看一下此模型是否与不包含预测变量的指数平滑模型在预测能力上存在巨大差别。

ARIMA 方法使您可以通过指定自回归、差分和移动平均值的顺序以及这些组件的季节对应物来对模型进行微调。由于手动确定上述各部分的最佳值时需要有大量的试错，从而可能变得十分耗时，因此在本例中，我们将为专家建模器选择 ARIMA 模型。

我们会通过将数据集中的某些其他变量视为预测变量来尝试构建更好的模型。最适合作为预测变量包含在内的变量有：邮递的产品目录数（ *邮件* ）、产品目录的页数（ *页* ）、开通的订购热线数目（ *电话* ）、印刷广告投入额（ *印刷* ）和客户服务代表人数（ *服务* ）。

图 213. 设置预测变量字段

1. 打开 IBM SPSS Statistics 文件源节点。
2. 在“类型”选项卡中，将*邮件*、*页*、*电话*、*印刷*和*服务*的*角色*设置为**输入**。
3. 确保将**男**的角色设置为**目标**，并将所有剩余字段的角色设置为**无**。

4. 单击**确定**。

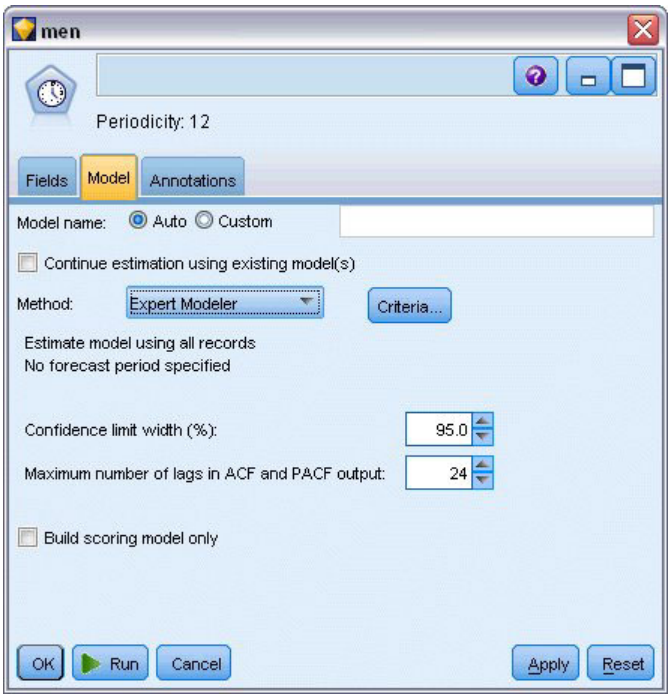


图 214. 选择专家建模器

5. 打开时间序列节点。
6. 在“模型”选项卡中，将**方法**设置为 **Expert Modeler**，然后单击**标准**。

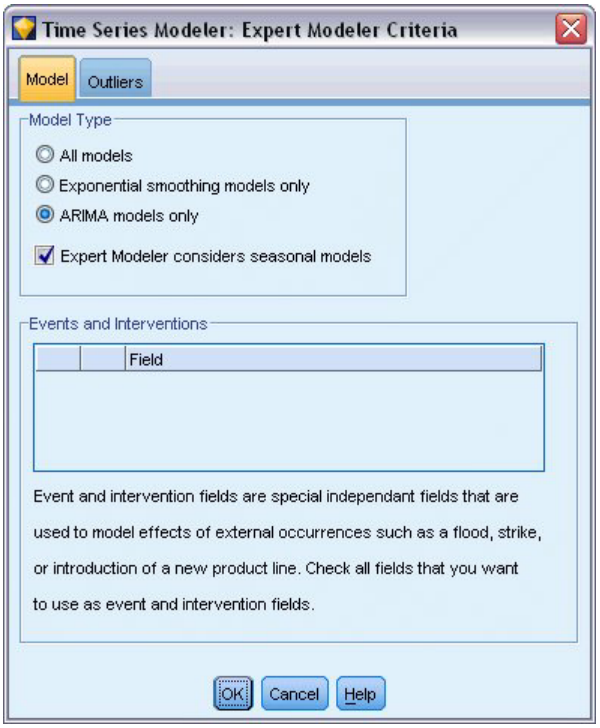


图 215. 仅选择 ARIMA 模型

7. 在“专家建模器标准”对话框中，选择**仅 ARIMA 模型**选项，并确保选中 **专家建模器考虑季节模型**。
8. 单击 **确定** 关闭此对话框。
9. 单击“模型”选项卡上的**运行**以再次创建模型块。

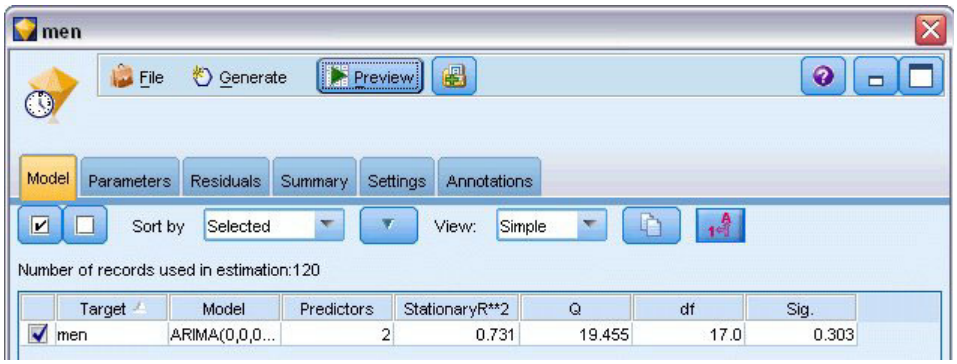


图 216. 专家建模器可选择两个预测变量

10. 打开模型块。

    注意专家建模器如何仅选择了 5 个指定预测变量中的 2 个作为模型的重大预测变量。
11. 单击**确定**关闭模型块。
12. 打开时间散点图节点并单击**运行**。

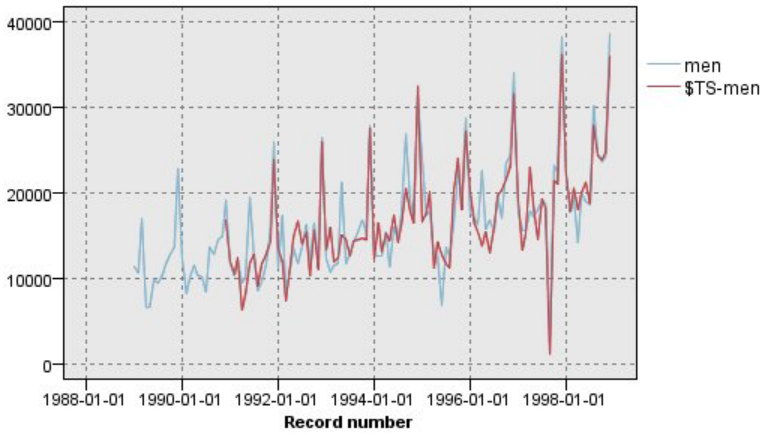


图 217. 包含指定的预测变量的 ARIMA 模型

比前一模型有所改进，此模型可以捕捉大型的下降峰值，并将其保持为当前最适合的值。

我们可以进一步优调此模型，但从此以后所进行的任何调整都将微乎其微。我们已确定包含预测变量的 ARIMA 模型更合适，因此使用刚刚构建的这一模型即可。此示例的目的是预测明年的销售情况。

13. 单击 **确定** 关闭时间散点图窗口。
14. 打开时间区间节点，然后选择 *预测* 选项卡。
15. 选择 *将记录扩展到未来* 复选框，然后将值设置为 12。

预测时使用预测变量时，您需要为预测时使用的字段指定估计值，这样建模器可以较为准确地预测目标字段。

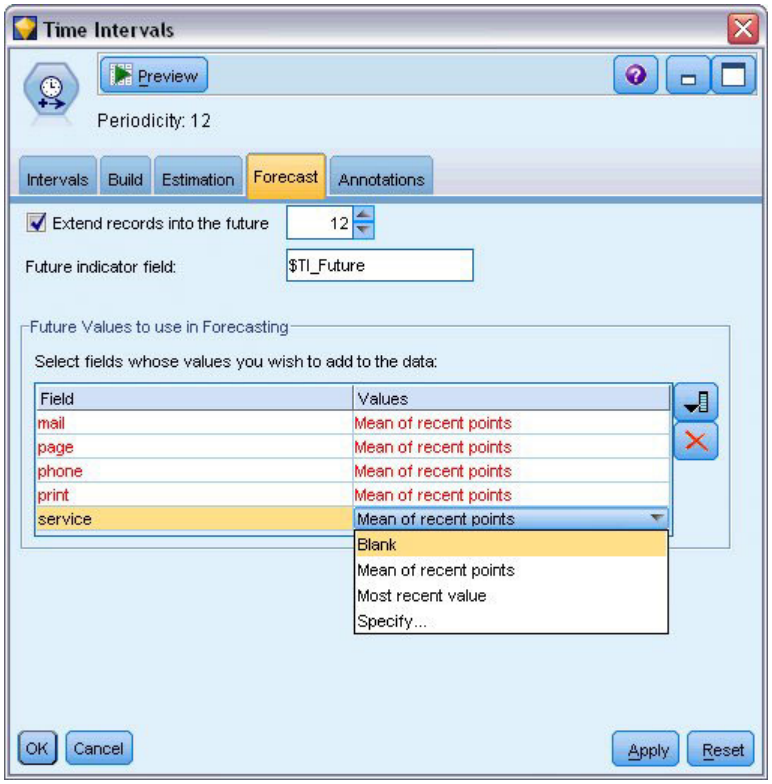


图 218. 指定预测变量字段的未来值

16. 在**要在预测中使用的未来值**组中，单击“值”列右侧的“字段选择器”按钮。
17. 在“选择字段”对话框中，通过**服务**选择**邮件**，然后单击**确定**。

    在实际应用中，您可以在此时手动指定未来值，因为这五个预测变量与您所控制的项都有关。此示例旨在使用某个预定义函数来避免必须为每个预测变量指定 12 个值。（如果对此示例比较熟悉，您可能会试着测试不同的未来值来了解其对模型有何影响。）
18. 对于每个字段，依次单击 **值** 字段以显示可能值列表，然后选择 **最近点的平均值** 。此选项将计算该字段后三个数据点的平均值，并将其用作每个案例的估计值。
19. 单击**确定**。
20. 打开时间序列节点并单击**运行**以再次创建模型块。
21. 打开时间散点图节点并单击**运行**。

1999 年的预测形势良好：如预期的那样，销售水平继十二月高峰期后再次回复正常，下半年一直保持平稳的上升趋势，整体销售情况比去年有明显的好转。

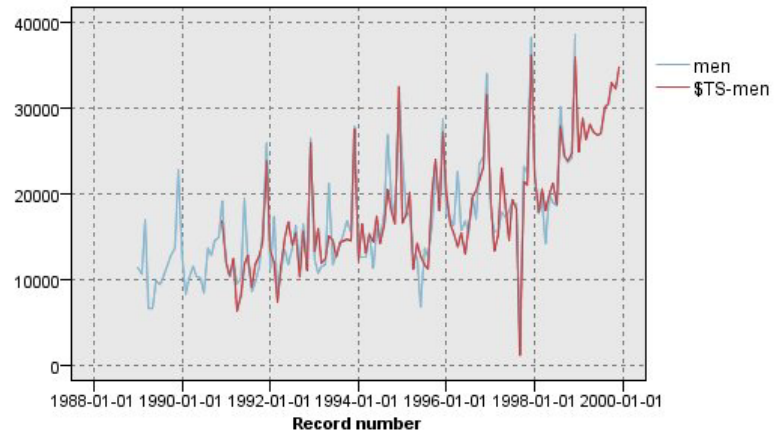


图 219. 包含指定的预测变量的销售预测

摘要

您已成功对复杂的时间序列进行建模，不仅包含上升趋势，还包含季节变化以及其他变化。通过试错，您还知道如何一步步得到精确模型，然后借助此模型预测未来的销售情况。

事实上，在更新实际销售数据时（例如 每月或每季度），您需要重新应用此模型并生成最新的预测。请参阅主题第 166 页的『重新应用时间序列模型』以获取更多信息。

第 16 章 向客户报价（自学）

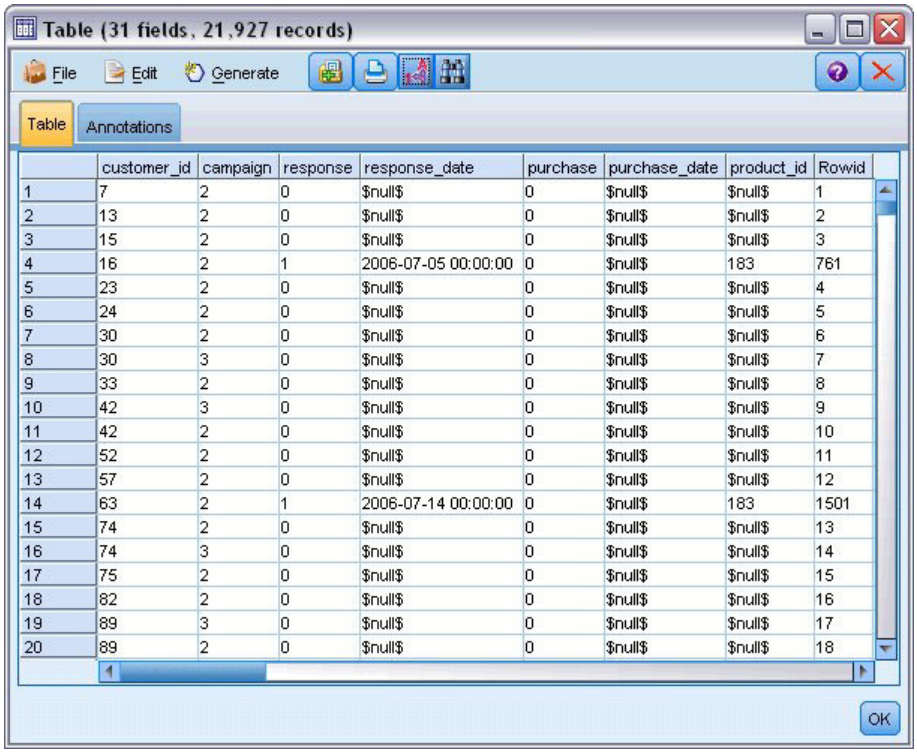
“自学响应模型”(SLRM) 节点用于生成和启用模型更新，该模型让您可以预测最适合于客户的报价以及报价被接受的概率。这种模型在客户关系管理（例如市场营销或呼叫中心）中非常有用。

此示例中提到的金融公司纯属虚构。市场营销部希望通过为各个客户匹配合适的报价，在未来的营销活动中创造更好的业绩。示例具体使用自学响应模型，根据以前的报价和响应确定最可能做出正面响应的客户的特征，并根据结果提高当前报价被接受的概率。

本示例使用流 *pm_selflearn.str* ，该流引用数据文件 *pm_customer_train1.sav* 、 *pm_customer_train2.sav* 和 *pm_customer_train3.sav* 。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *pm_selflearn.str* 位于 *streams* 文件夹中。

现有数据

公司拥有追踪以前营销活动中向客户做出的报价及客户对报价的响应的历史数据。这些数据还含有可用于预测不同客户响应率的人口统计和金融信息。



	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

图 220. 以前报价的响应

构建流

1. 添加指向 *pm_customer_train1.sav* 的“Statistics 文件”源节点，该文件位于 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中。

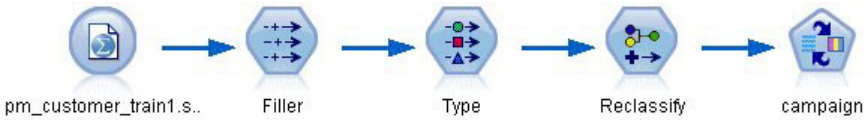


图 221. SLRM 样本流

2. 添加一个填充节点，并选择 `campaign` 作为字段的填充内容。
3. 选择 **始终** 作为替换类型。
4. 在“替换为”文本框中输入 `to_string(campaign)` 并单击**确定**。

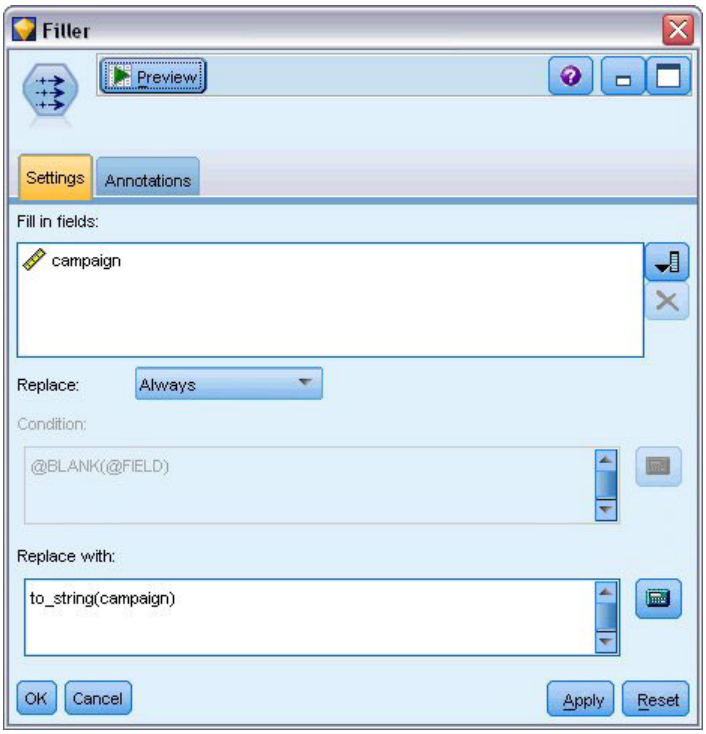


图 222. 导出竞销活动字段

5. 添加一个类型节点，然后将 *customer_id*、*response_date*、*purchase_date*、*product_id*、*Rowid* 和 *X_random* 字段的*角色*设置为**无**。

图 223. 更改“类型”节点设置

6. 将 *campaign* 和 *response* 字段的*角色*设置为**目标**。这些字段将作为预测的基准。

   将*响应*字段的**测量**设置为**标志**。

7. 单击 **读取值** ，然后单击 **确定** 。

   由于 campaign 字段数据显示为一系列数字（1、2、3 和 4），因此可以对字段进行重新分类以显示更有意义的标题。

8. 为类型节点添加一个重新分类节点。
9. 在**重新分类为**字段中，选择**现有字段**。
10. 在**重新分类字段**列表中，选择**campaign**。
11. 单击 **获取** 按钮；campaign 的值将添加到 *原始值* 列。
12. 在*新值*列中，在前四行中输入以下活动名称:
    - **抵押**
    - **汽车贷款**
    - **储蓄**
    - **退休金**
13. 单击**确定**。

图 224. 对竞销活动名称进行重新分类

14. 将 SLRM 建模节点附加到“重新分类”节点。在“字段”选项卡上，将 **campaign** 和 **response**分别选择为目标字段和目标响应字段。

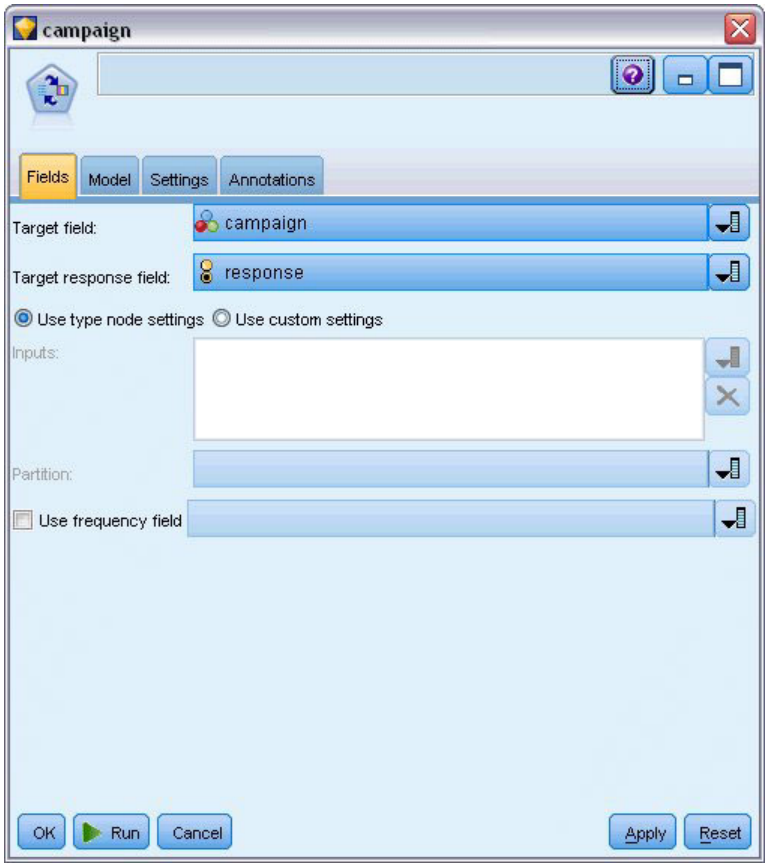


图 225. 选择目标和目标响应

15. 在“设置”选项卡的“每条记录的最大预测数”字段中，将数字减为 2。

    这表示对于每位客户，将确定两项具有最高接受概率的报价。

16. 确保选中了**考虑模型可靠性**，并单击**运行**。

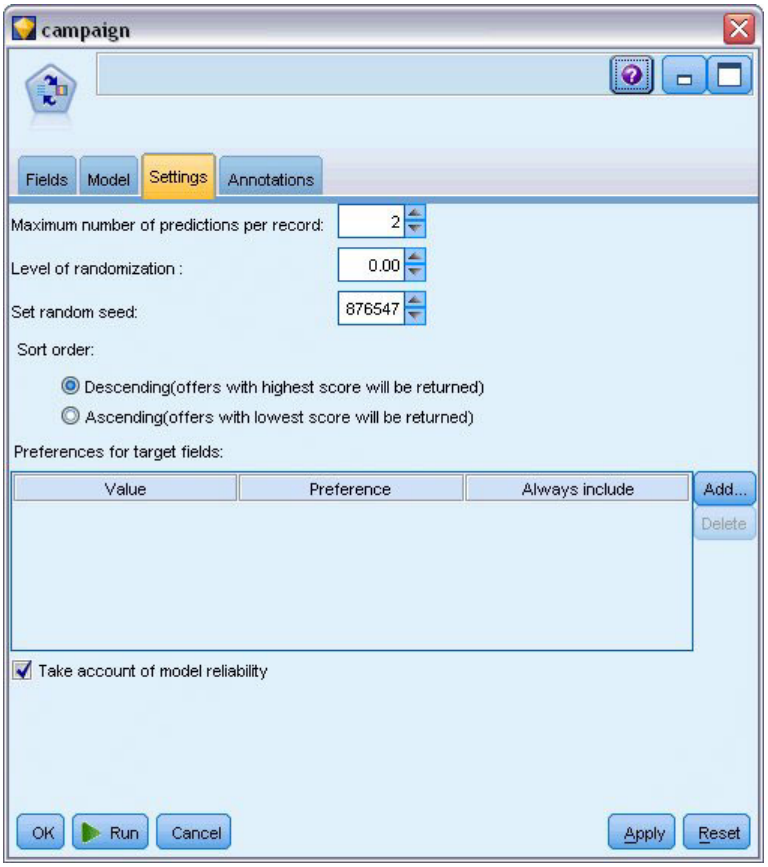


图 226. SLRM 节点设置

浏览模型

1. 打开模型块。“模型”选项卡最初显示每项报价的预测准确性估计值，以及每个预测变量在估计模型时的相对重要性。

   要显示每个预测变量与目标变量的相关性，从右侧窗格的**视图**列表中选择**与响应关联**。
2. 要在具有预测值的四个报价之间进行切换，从左侧窗格的**视图**列表中选择所需报价。

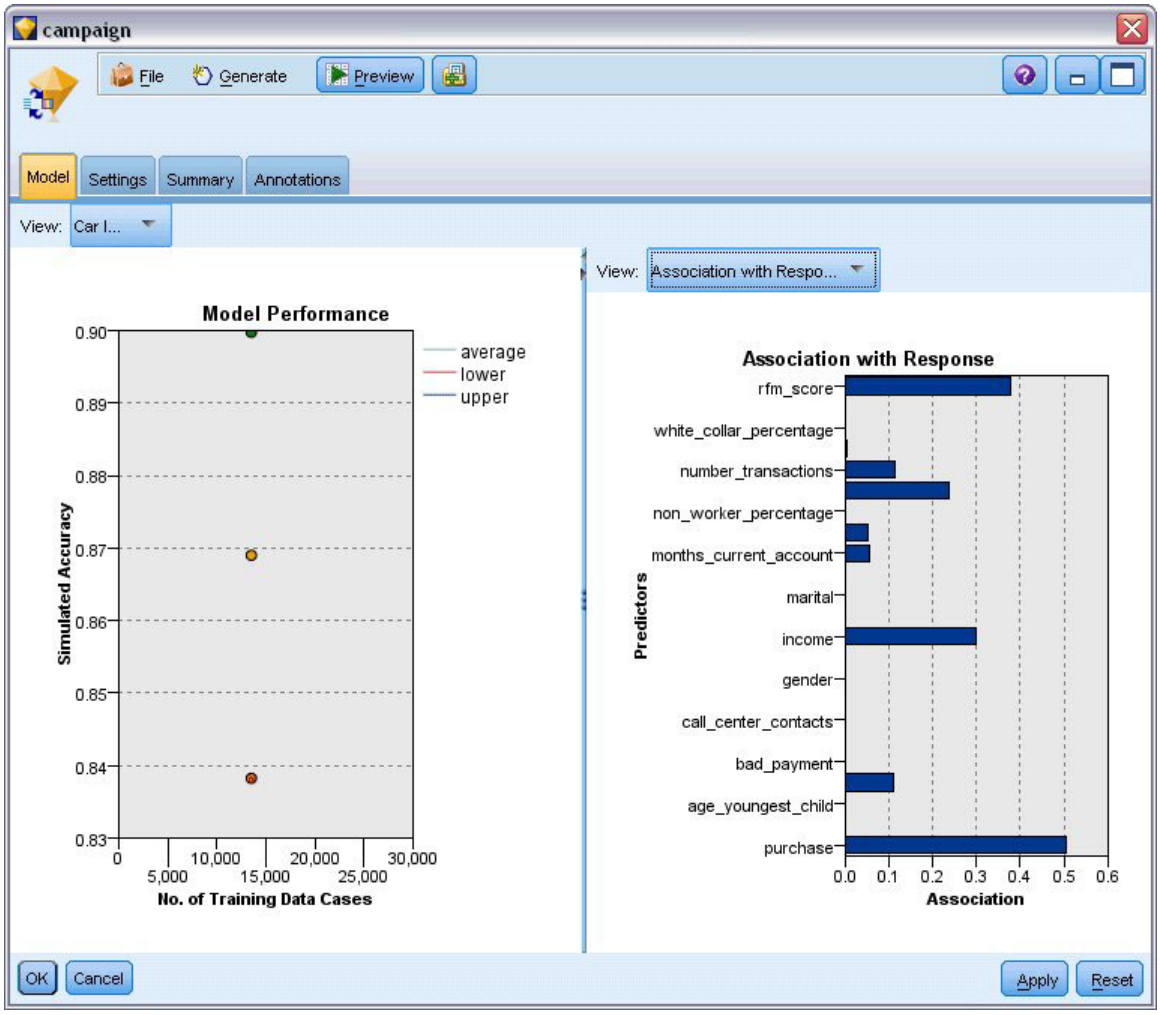


图 227. SLRM 模型块

3. 关闭模型块窗口。
4. 在流工作区上，将指向 *pm_customer_train1.sav* 的 IBM SPSS Statistics 文件源节点断开连接。
5. 添加指向 *pm_customer_train2.sav* 的“Statistics 文件”源节点（该文件位于 IBM SPSS Modeler 安装目录的 *Demos* 文件夹中），并将其连接到过滤节点。

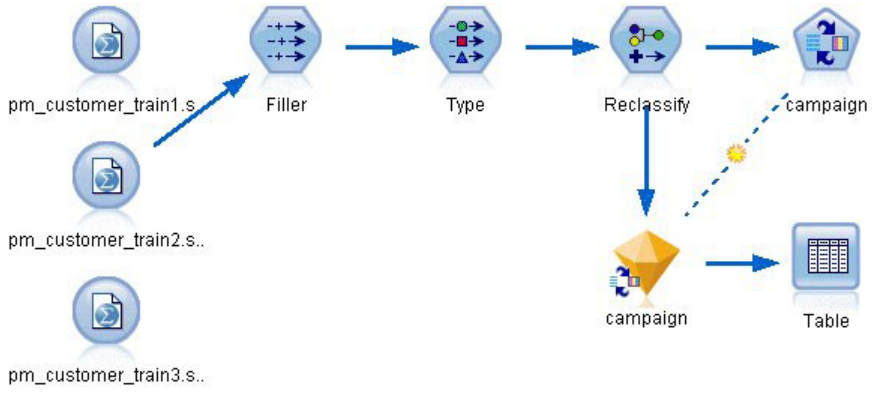


图 228. 将第二个数据源附加到 SLRM 流

6. 在 SLRM 节点的“模型”选项卡中，选择**继续训练现有模型**。

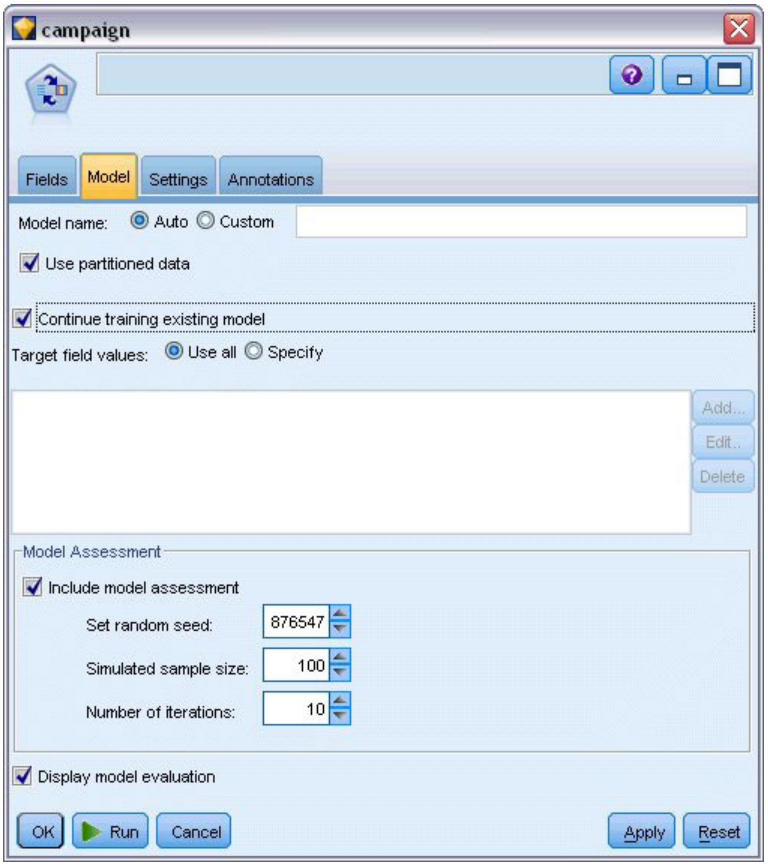


图 229. 继续训练模型

7. 单击**运行**以再次创建模型块。要查看其详细信息，请双击画布中的模型块。

   此时“模型”选项卡将显示每项报价的预测准确性修正估计值。

8. 添加指向 *pm_customer_train3.sav* 的“Statistics 文件”源节点（该文件位于 IBM SPSS Modeler 安装目录的 *Demos* 文件夹中），并将其连接到过滤节点。

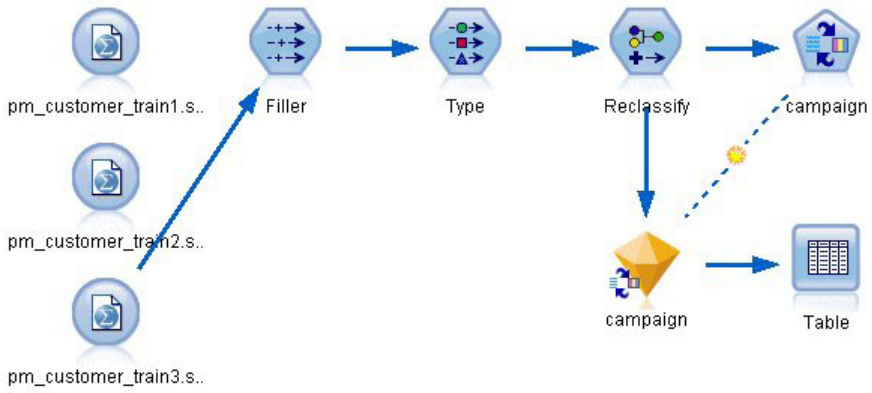


图 230. 将第三个数据源附加到 SLRM 流

9. 单击**运行**以再次创建模型块。要查看其详细信息，请双击画布中的模型块。

10. 此时“模型”选项卡将显示每项报价的预测准确性最终估计值。

    如您所见，添加其他数据源时，平均准确性稍有下降（从 86.9% 降至 85.4%）；但这种波动幅度很小，并且可能是可用数据中的轻微异常造成的。

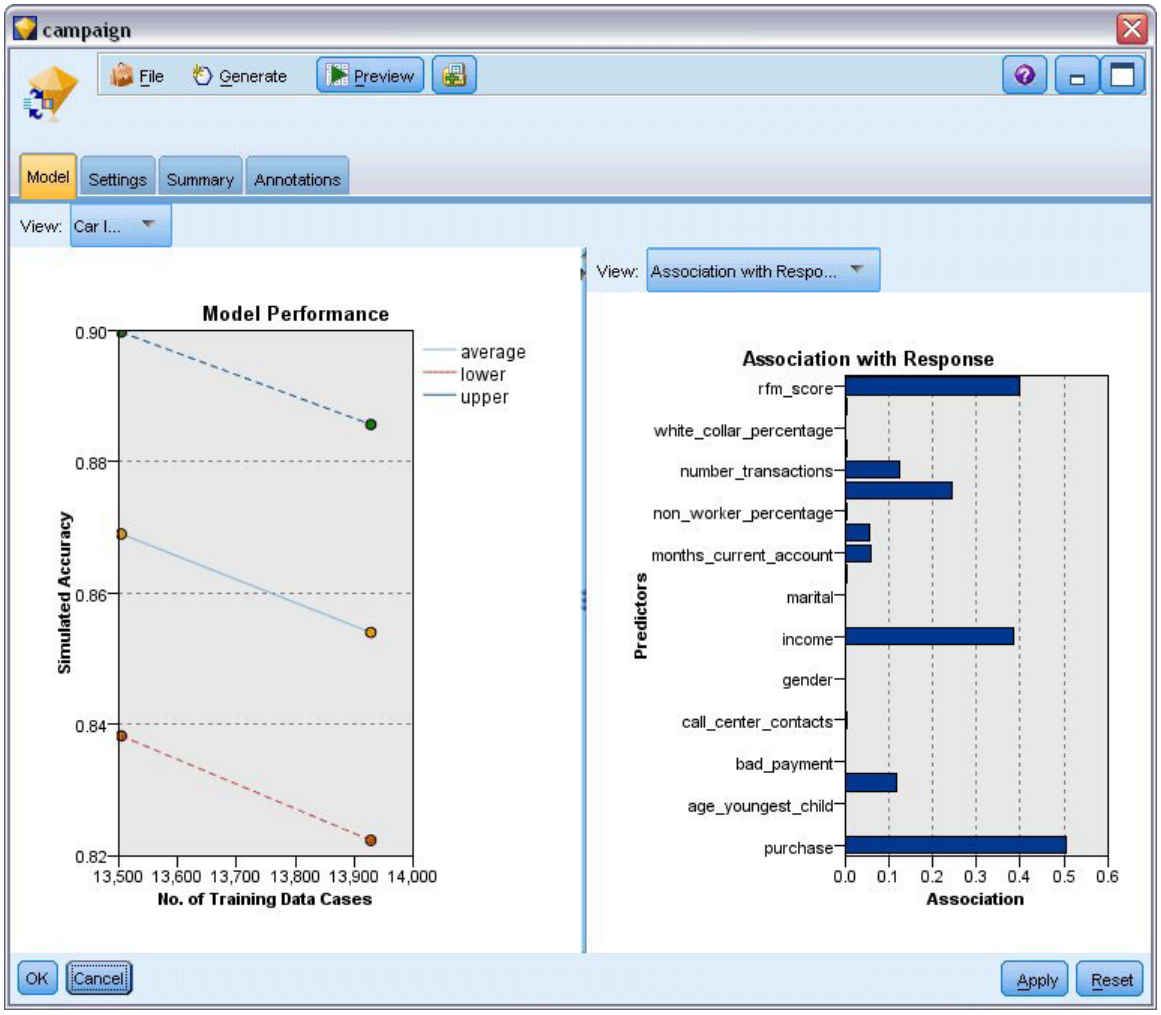
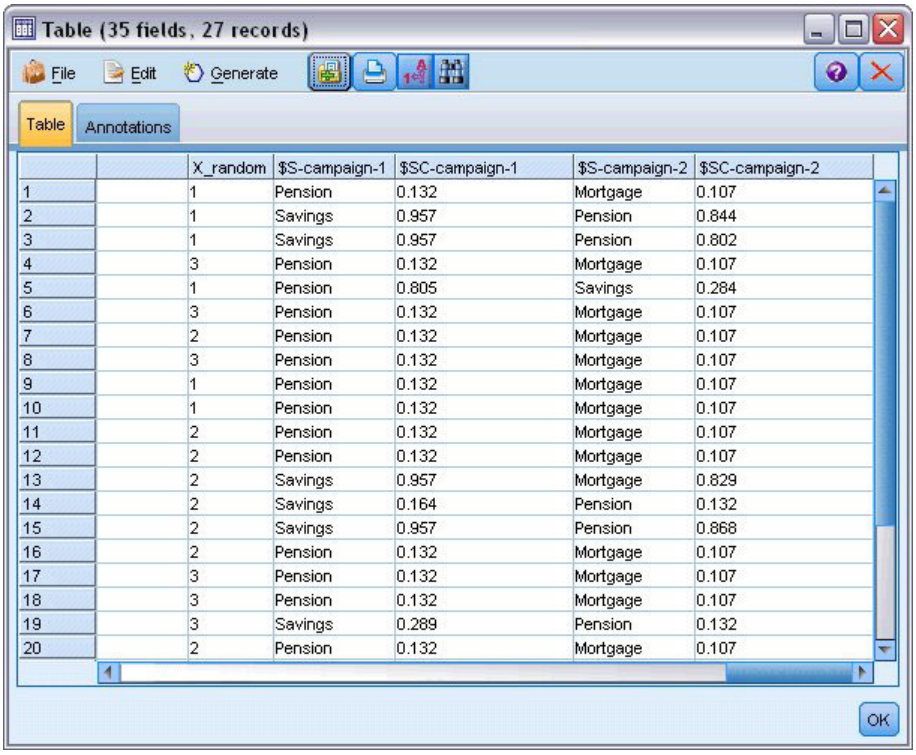


图 231. 更新后的 SLRM 模型块

11. 将表节点附加到生成的最后一个（第三个）模型，然后执行该表节点。
12. 滚动至表的右侧。预测将显示客户最有可能接受哪些报价，以及他们将会接受的置信度，具体取决于每位客户的详细信息。

例如，在显示表格的第一行中，以前取得汽车贷款的某位客户将根据提供的报价接受退休金的置信度比率只有 13.2%（表示为 *$SC-campaign-1* 列中的 0.132）。但是，第二行和第三行显示了另外两位也曾取得汽车贷款的客户；在他们的案例中，他们以及具有类似历史记录的其他客户将根据提供的储蓄报价开立储蓄帐户的置信度为 95.7%，并且接受退休金的置信度高于 80%。



		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

图 232. 模型输出 - 预测报价和置信度

有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于产品 DVD 的 *\Documentation* 目录中。

另请注意，这些结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。.

第 17 章 预测贷款拖欠者（贝叶斯网络）

使用贝叶斯网络，可以通过将观察到并记录下的证据与实际常识结合起来构建概率模型，以通过使用表面看上去不相关的属性确定发生的可能性。

此示例使用名为 *bayes_bankloan.str* 的流，它引用名为 *bankloan.sav* 的数据文件。这些文件位于任意 IBM SPSS Modeler 安装程序中的 *Demos* 目录下，并可从 Windows“开始”菜单上的 IBM SPSS Modeler 程序组进行访问。文件 *bayes_bankloan.str* 位于 *streams* 目录下。

例如，假设某个银行希望了解不偿还贷款的潜在情况。如果先前的贷款拖欠数据可用于预测哪些潜在客户可能难以偿还贷款，那么可以对这些“风险大”的客户减少贷款或者为他们提供替代产品。

本示例主要讲述使用现有贷款拖欠数据来预测今后出现的潜在贷款拖欠者，并观察了三个不同的贝叶斯网络模型类型，从而确定在这种情况下哪个类型的预测效果更好。

构建流

1. 在 *Demos* 文件夹中添加指向 *bankloan.sav* 的“Statistics 文件”源节点。

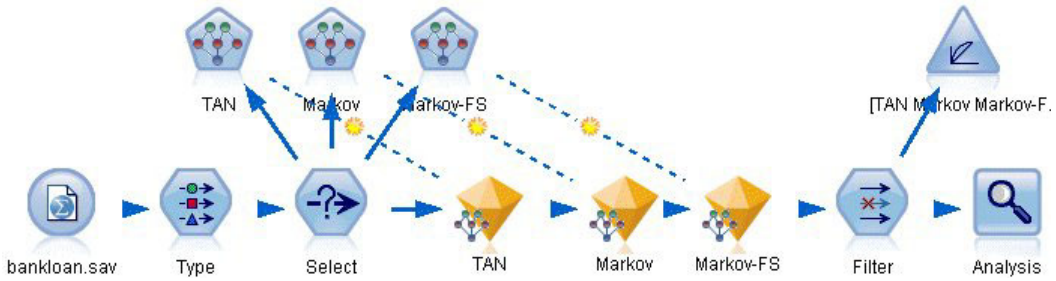


图 233. 贝叶斯网络样本流

2. 将类型节点添加到源节点，并将**缺省**字段的角色设为**目标**。将所有其他字段的角色设置为 **Input**。
3. 单击**读取值**按钮以填充值列。

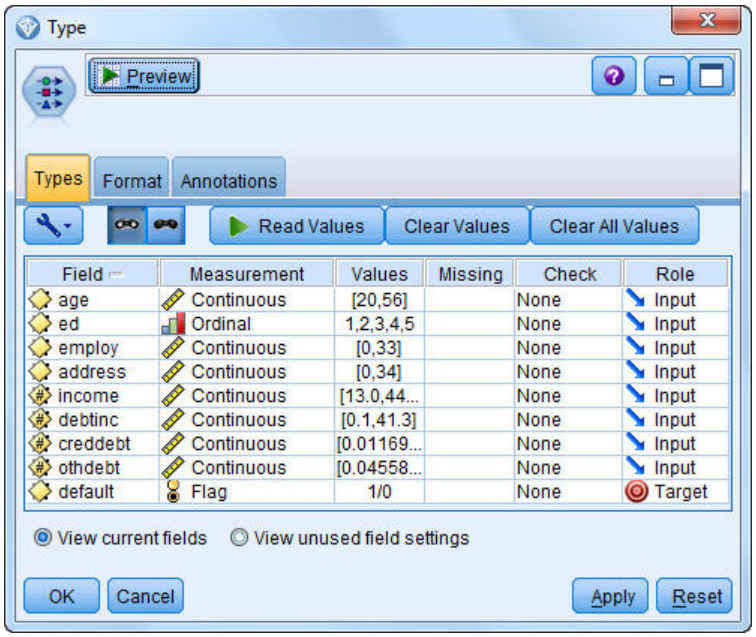


图 234. 选择目标字段

构建模型时，其目标具有空值的个案没有用处。您可以排除这些观测值以防止在模型评估中使用它们。

4. 为类型节点添加一个选择节点。
5. 对于模型，请选择 **丢弃** 。
6. 在“条件”框中，输入 **default = '$null$'**。

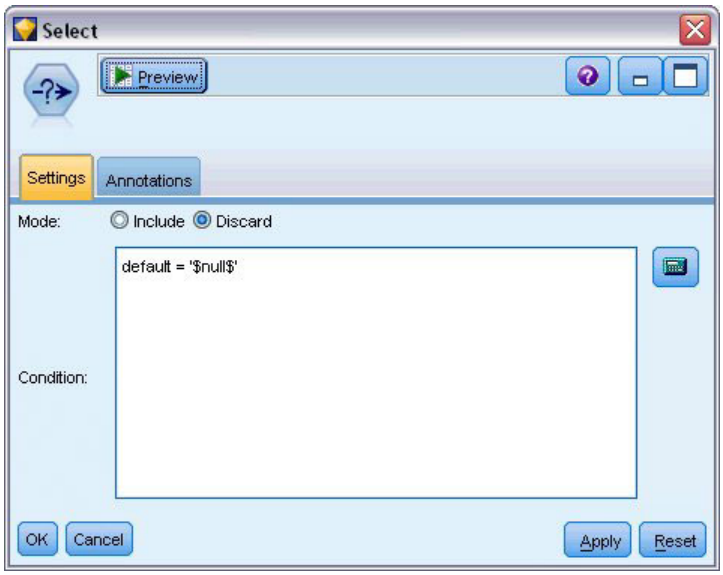


图 235. 废弃空的目标

由于您可以构建多种不同类型的贝叶斯网络，因此最好对它们进行比较，以确定哪种模型提供最好的预测。第一个要创建的模型是树扩展朴素贝叶斯 (TAN) 模型。

7. 将贝叶斯网络节点附加到选择节点上。
8. 对于“模型”选项卡上的模型名称，请选择**自定义**，并在文本框中输入 TAN。

9. 对于结构类型，请选择 **TAN** 并单击**确定**。

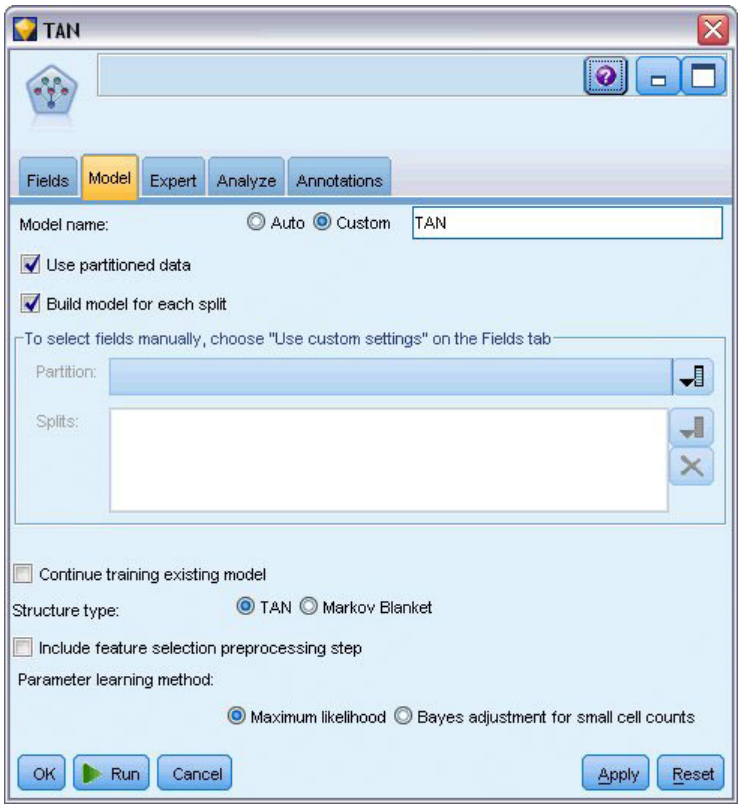


图 236. 创建树扩展朴素贝叶斯模型

第二个要创建的模型具有马尔可夫覆盖结构。

10. 将第二个贝叶斯网络节点添加到选择节点上。
11. 对于“模型”选项卡上的模型名称，请选择**自定义**，并在文本框中输入马尔可夫。
12. 对于结构类型，请选择**马尔可夫覆盖**并单击**确定**。

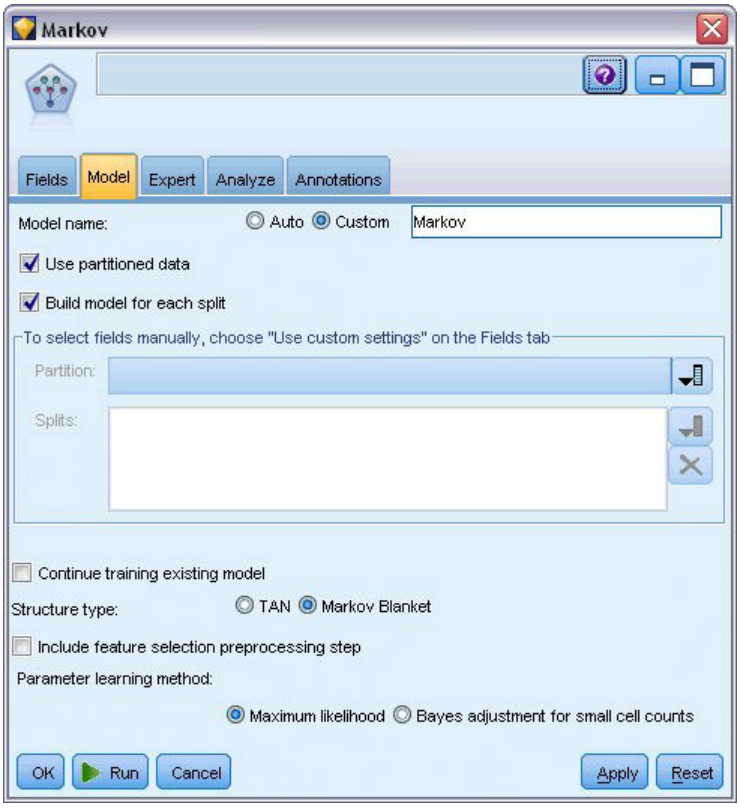


图 237. 创建马尔可夫覆盖模型

要创建的第三个模型具有马尔可夫覆盖结构，同时也使用了特征选择预处理来选择与目标变量有重大关联的输入。

13. 将第三个贝叶斯网络节点添加到选择节点上。
14. 对于“模型”选项卡上的模型名称，请选择**自定义**，并在文本框中输入马尔可夫 FS。
15. 对于结构类型，请选择**马尔可夫覆盖**。
16. 选择**包括特征选择预处理步骤**并单击**确定**。

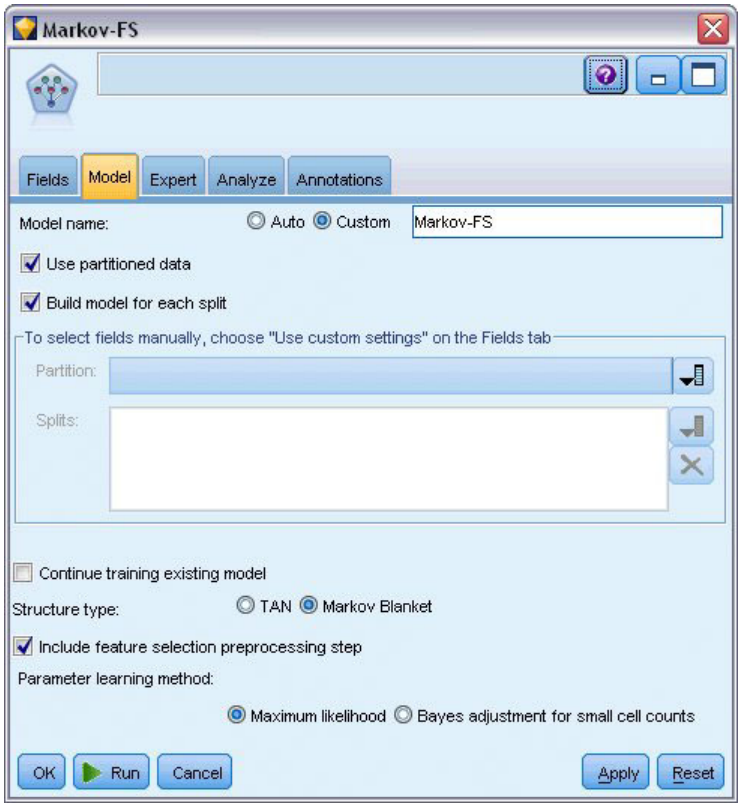


图 238. 使用特征选择预处理来创建马尔可夫覆盖模型

浏览模型

1. 运行流以创建模型块，这些模型块将添加到流和右上角的“模型”选用板中。要查看其详细信息，请双击流中的任一模型块。

   模型块“模型”选项卡分为两个窗格。左窗格包含节点网络图，可显示目标与其最重要预测变量之间的关系，以及各预测变量之间的关系。

   右窗格可能显示*预测变量重要性*，它表示评估模型时每个预测变量的相对重要性，右窗格也可能显示 *条件概率*，它包含各个节点值的条件概率值，以及各节点的父节点中的所有值组合。

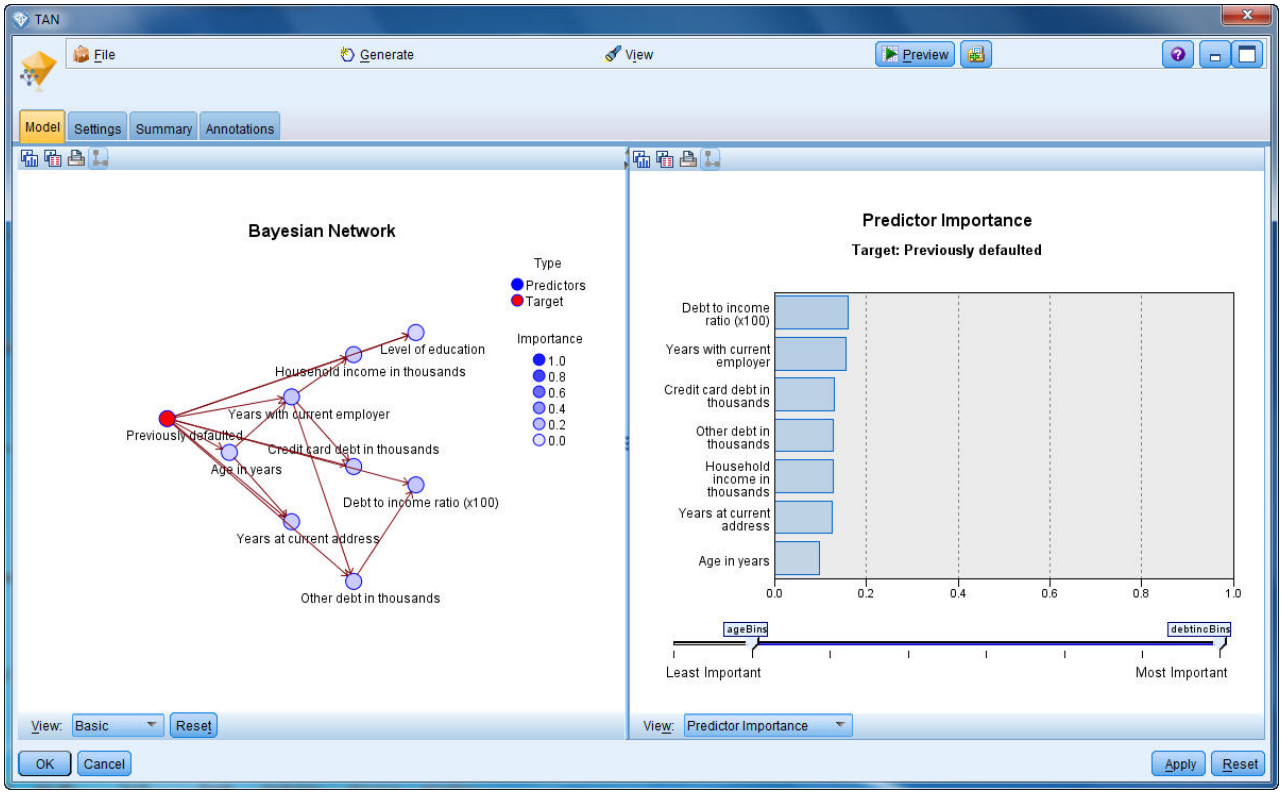


图 239. 查看树扩展朴素贝叶斯模型

2. 将 TAN 模型块连接到马尔可夫模型块（选择警告对话框上的**替换**）。
3. 将马尔可夫模型块连接到马尔可夫 FS 模型块（选择警告对话框上的**替换**）。
4. 通过选择节点对齐三个模型块以方便查看。

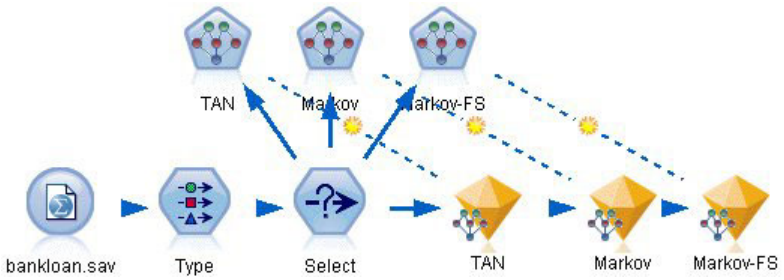


图 240. 对齐流中的块

5. 要重新命名您正在创建的评估图形上的模型输出以避免混淆，请将过滤节点附加到马尔可夫 FS 模型块。
6. 在右侧的字段栏中，将 $B-default、$B1-default 和 $B2-default 分别重新命名为 TAN、马尔可夫和马尔可夫 FS。

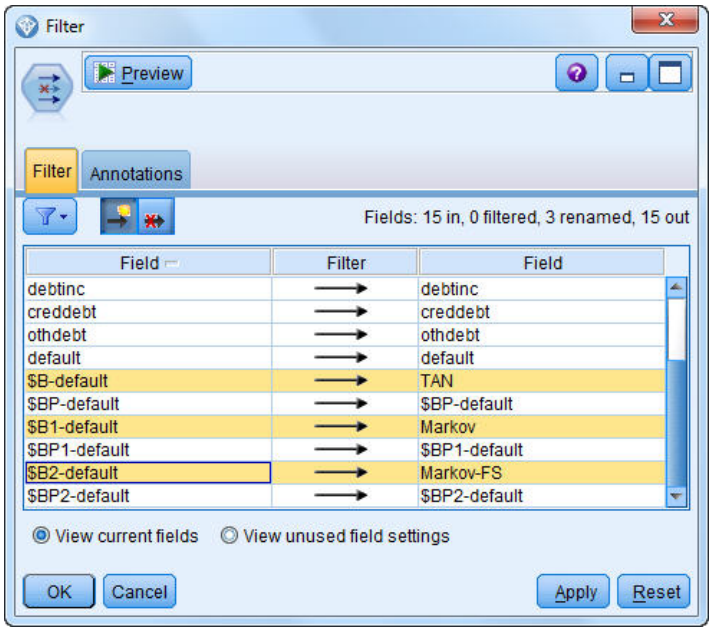


图 241. 重新命名模型字段名称

要比较模型的预测准确性，您可以构建一个增益图。

7. 将评估图形节点附加到过滤节点上，然后使用图形节点的缺省设置来执行它。

该图形显示，每个模型类型都生成了相似的结果，但是马尔可夫模型要稍微好一些。

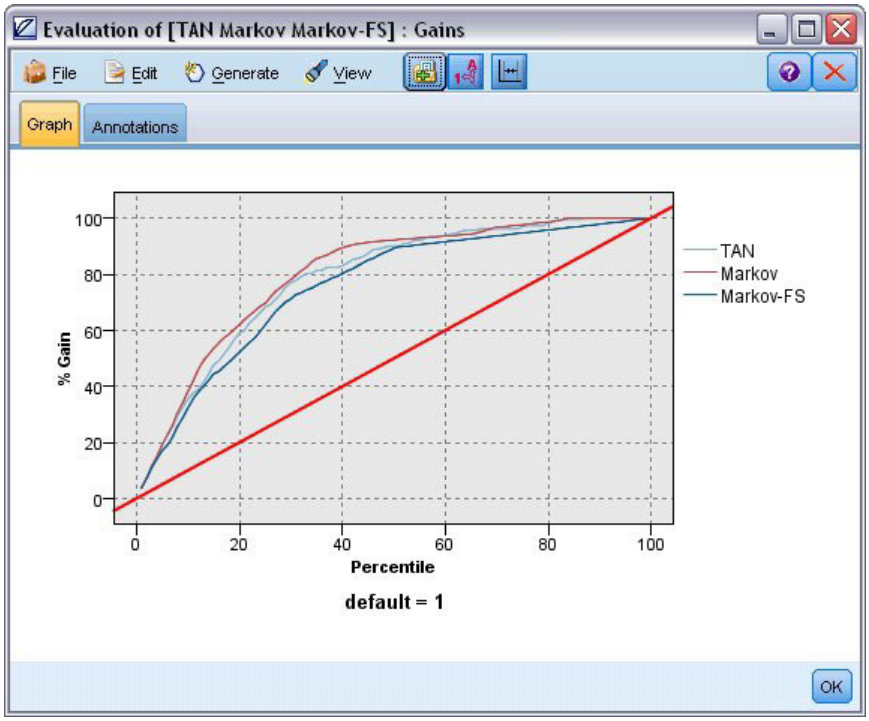


图 242. 评估模型准确性

要检查每个模型的预测效果，您可能会使用“分析”节点而不是“评估”图形。下图显示了依据正确和不正确的预测百分比得出的准确性。

8. 将分析节点附加到过滤节点上，然后使用分析节点的缺省设置来执行它。

与“评估”图形一样，本图形说明马尔可夫模型在预测的正确性方面稍微好一些；但马尔可夫 FS 模型仅落后马尔可夫模型几个百分点。这可能就意味着使用马尔可夫 FS 模型要更为方便一些，因为它计算结果所需的输入更少，因此节省了数据收集和输入的时间以及处理时间。

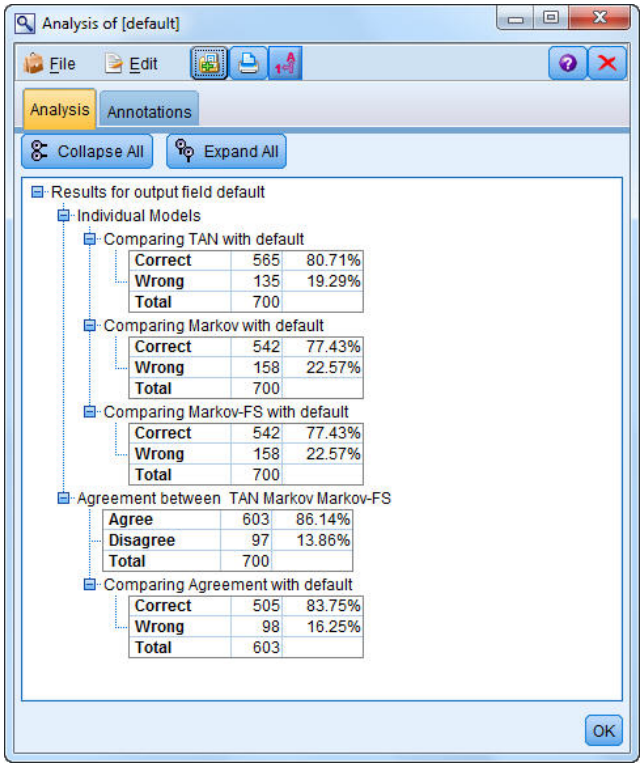


图 243. 分析模型准确性

有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于安装光盘的 *\Documentation* 目录中。

另请注意，这些结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

第 18 章 每个月重新训练模型（贝叶斯网络）

使用贝叶斯网络，可以通过将观察到并记录下的证据与实际常识结合起来构建概率模型，以通过使用表面看上去不相关的属性确定发生的可能性。

此示例使用名为 *bayes_churn_retrain.str* 的流，此流引用名为 *telco_Jan.sav* 和 *telco_Feb.sav* 的数据文件。这些文件位于任意 IBM SPSS Modeler 安装程序中的 *Demos* 目录下，并可从 Windows“开始”菜单上的 IBM SPSS Modeler 程序组进行访问。文件 *bayes_churn_retrain.str* 位于 *streams* 目录下。

例如，假设某个电信服务提供商关心流失到竞争对手那里的客户数（顾客流失率）。如果历史记录的客户数据用作预测哪些客户更可能在今后流失，那么这些客户可能划定为出于刺激性的动机或其他一些使它们失去信心的意图而倒向了另一个服务提供商。

本示例主要讲述使用现有的每月流失数据来预测哪些客户在今后流失，然后将这些数据添加到模型中，从而精炼和重新训练模型。

构建流

1. 在 *Demos* 文件夹中添加指向 *telco_Jan.sav* 的“Statistics 文件”源节点。

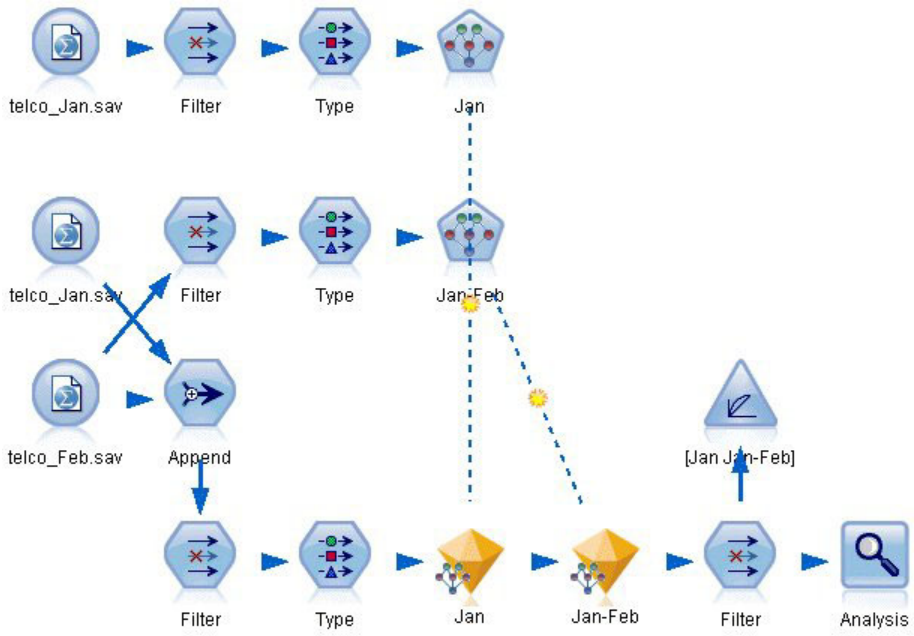


图 244. 贝叶斯网络样本流

先前的分析已说明在预测顾客流失率时有几个数据字段不太重要。这些字段可从数据集中滤出，从而在构建模型以及对模型评分时提高处理速度。

2. 为源节点添加一个过滤节点。
3. 排除除 *地址* 、 *年龄* 、 *流失* 、 *客户类别* 、 *教育程度* 、 *行业* 、 *性别* 、 *婚姻状况* 、 *居住地* 、 *退休* 和 *保有期* 外的所有字段。

4. 单击**确定**。

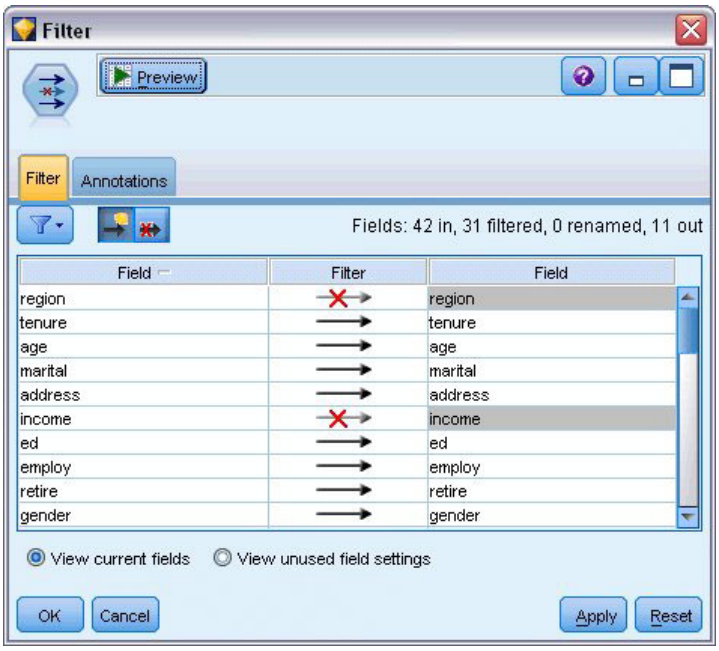


图 245. 过滤不必要的字段

5. 为过滤节点添加一个类型节点。
6. 打开类型节点并单击**读取值**按钮以填充值列。
7. 为了使“评估”节点可以评估值的真假，需要将流失字段的测量级别设置为**标志**，并将其角色设置为**目标**。单击**确定**。

图 246. 选择目标字段

您可以构建多个不同类型的贝叶斯网络；但在本示例中，您将构建树扩展朴素贝叶斯 (TAN) 模型。该模型创建了一个大型网络，可以确保其中已经囊括了数据变量间所有可能存在的连接关系，从而构建一个稳健的初始模型。

8. 将贝叶斯网络节点附加到类型节点上。
9. 对于“模型”选项卡上的模型名称，请选择**自定义**，并在文本框中输入 Jan。
10. 对于参数学习方法，请选择 **对小单元格计数的贝叶斯调整** 。
11. 单击**运行**。模型块被添加到流，同时添加到右上角的“模型”选用板。

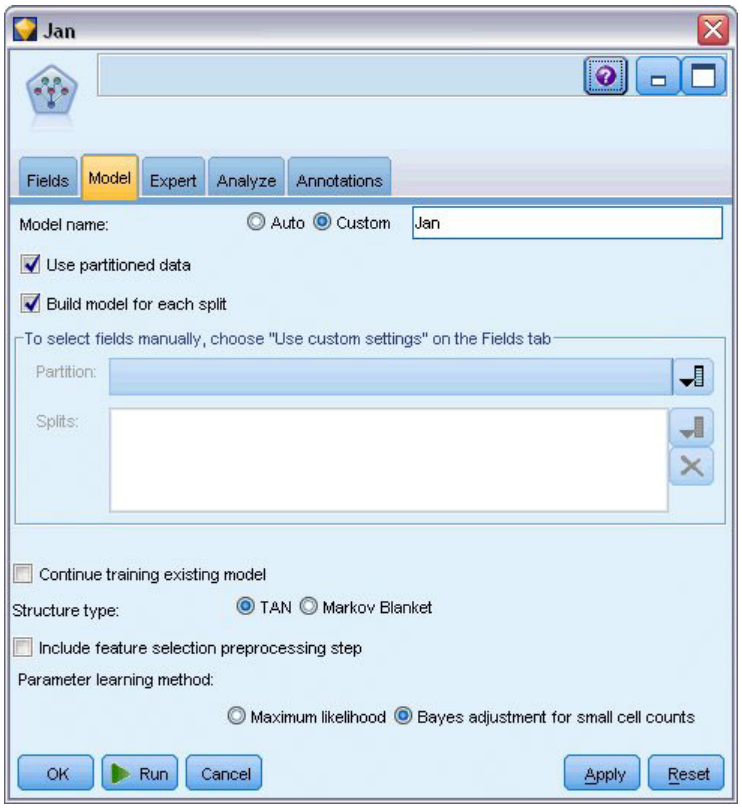


图 247. 创建树扩展朴素贝叶斯模型

12. 在 *Demos* 文件夹中添加指向 *telco_Feb.sav* 的“Statistics 文件”源节点。
13. 将此新的源节点附加到过滤节点（在警告对话框上，选择**替换**以替换到前一源节点的连接）。

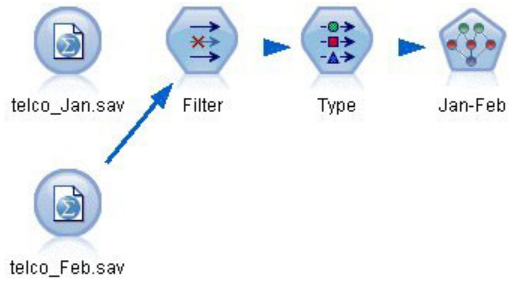


图 248. 添加第二个月的数据

14. 对于贝叶斯网络节点的“模型”选项卡上的模型名称，请选择**自定义**，并在文本框中输入 Jan-Feb。

15. 选择 **继续训练现有模型** 。
16. 单击**运行**。模型块覆盖流中的现有模型块，但同时也将添加到右上角的“模型”选用板。

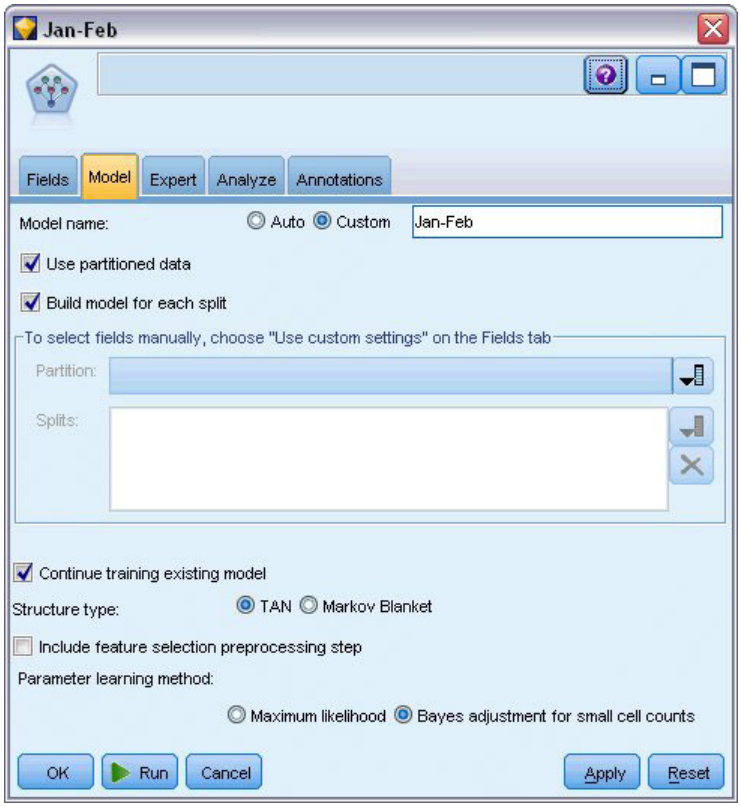


图 249. 重新训练模型

评估模型

要对模型进行比较，必须将两个数据集合并到一起。

1. 添加一个追加节点并将*telco_Jan.sav* 和 *telco_Feb.sav* 源节点都附加到该节点上。

图 250. 追加两个数据源

2. 从早期的流中复制过滤节点和类型节点，然后将它们粘贴到流工作区上。
3. 将追加节点附加到最新复制的过滤节点上。

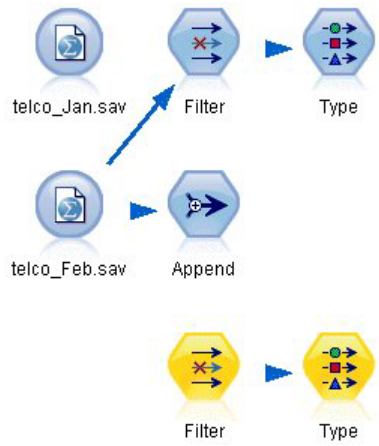


图 251. 将复制的节点粘贴到流中

两个贝叶斯网络模型的模型块位于右上角的“模型”选用板中。

4. 双击 Jan 模型块以将其导入流，并将其附加到新复制的类型节点。
5. 将流中现有的 Jan-Feb 模型块附加到 Jan 模型块。
6. 打开 Jan 模型块。

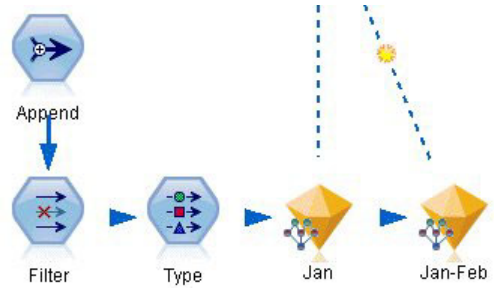


图 252. 将模型块添加到流中

贝叶斯网络模型块“模型”选项卡分为两列。左列包含节点网络图，可显示目标与其最重要预测变量之间的关系，以及各预测变量之间的关系。

右列可能显示*预测变量重要性*，它表示评估模型时每个预测变量的相对重要性，右窗格也可能显示 *条件概率*，它包含各个节点值的条件概率值，以及各节点的父节点中的所有值组合。

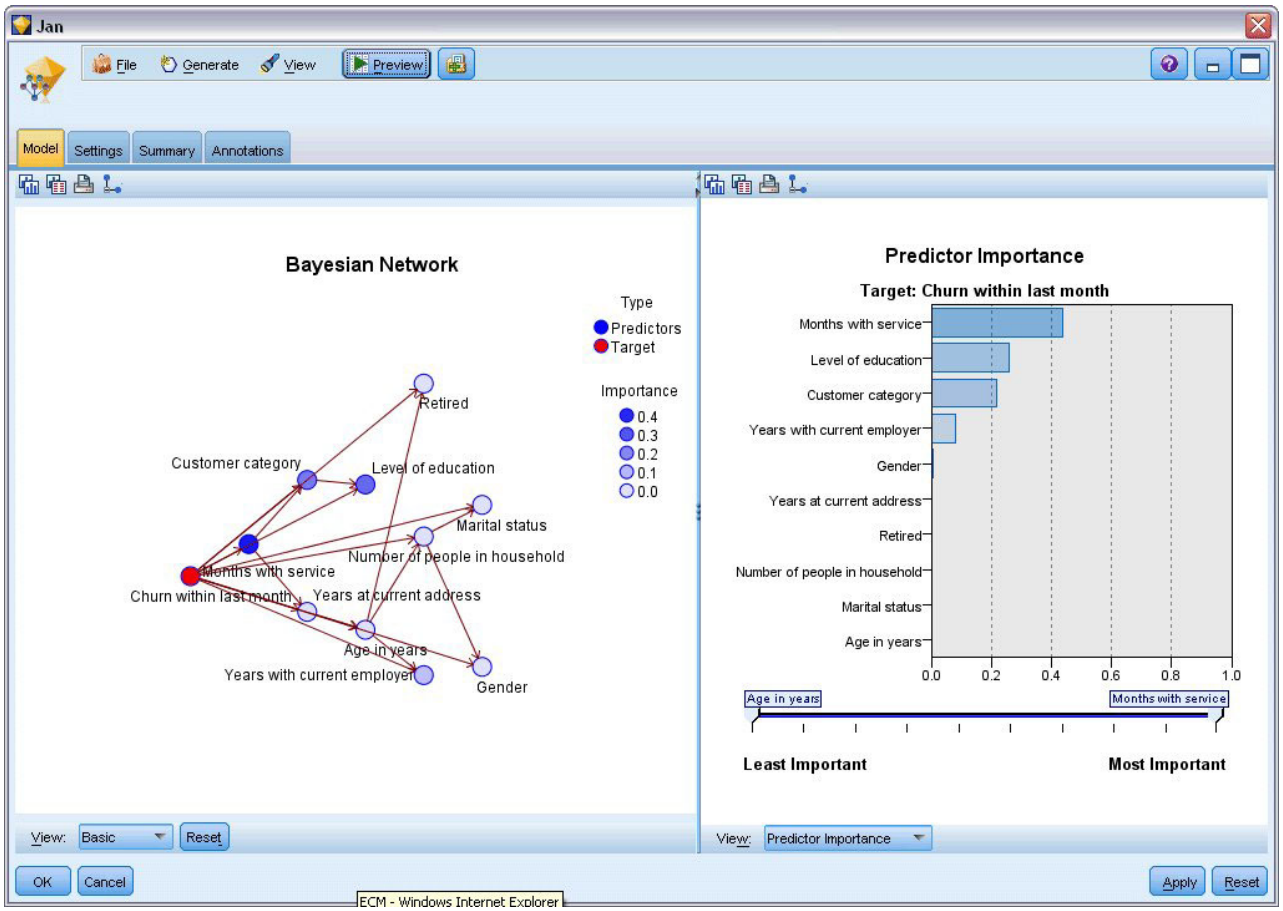


图 253. 显示预测变量重要性的贝叶斯网络模型

要显示所有节点的条件概率，请单击左列中的节点。相应的右列会更新以显示所需的详细信息。

显示每个分级的条件概率，这些分级中的数据值已划分为与该节点的父节点和同级节点相关。

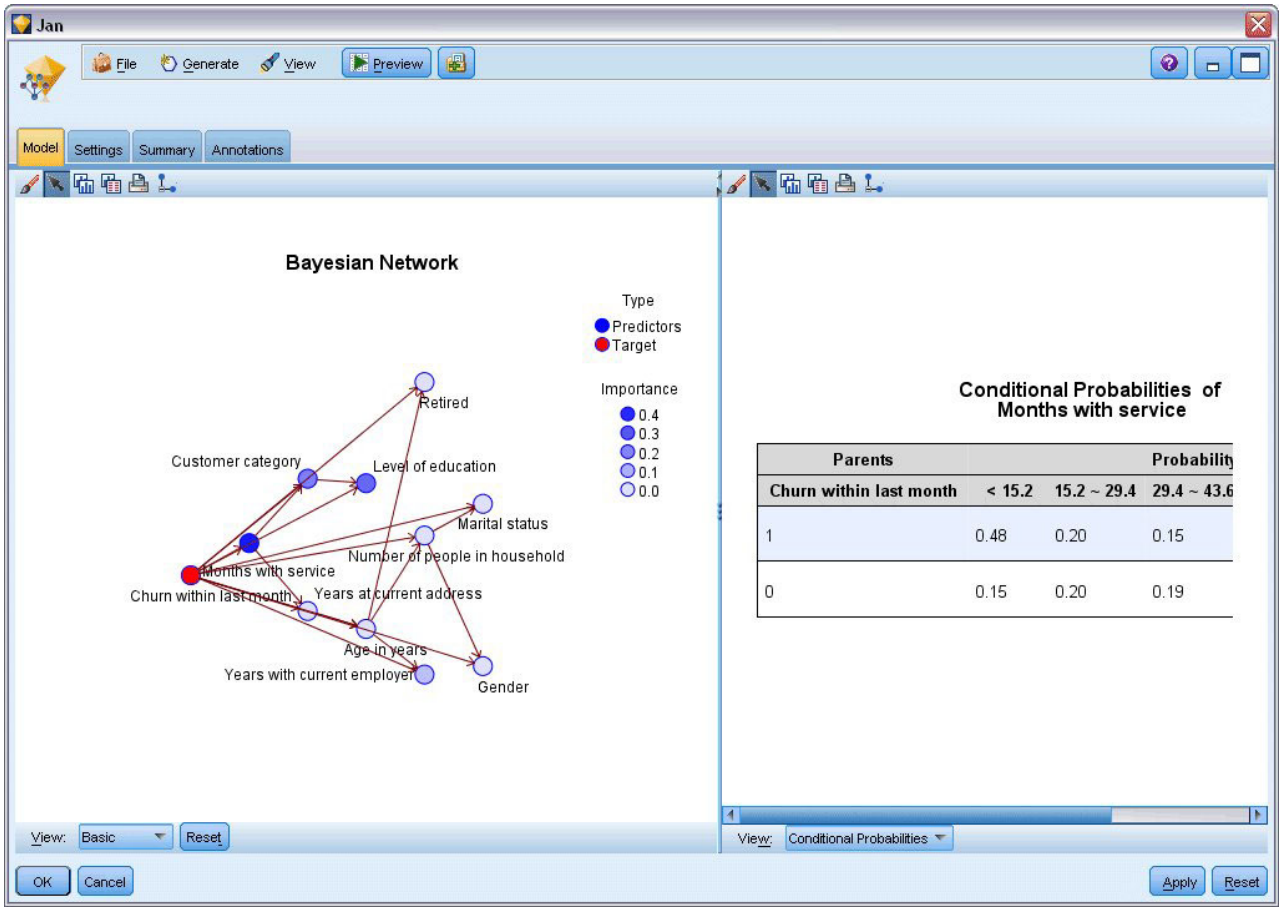


图 254. 显示条件概率的贝叶斯网络模型

7. 要重新命名模型输出以避免混淆，请将过滤节点附加到 Jan-Feb 模型块。
8. 在右侧的 *字段* 列，将 $B-churn 和 $B1-churn 分别重新命名为 Jan 和 Jan-Feb。

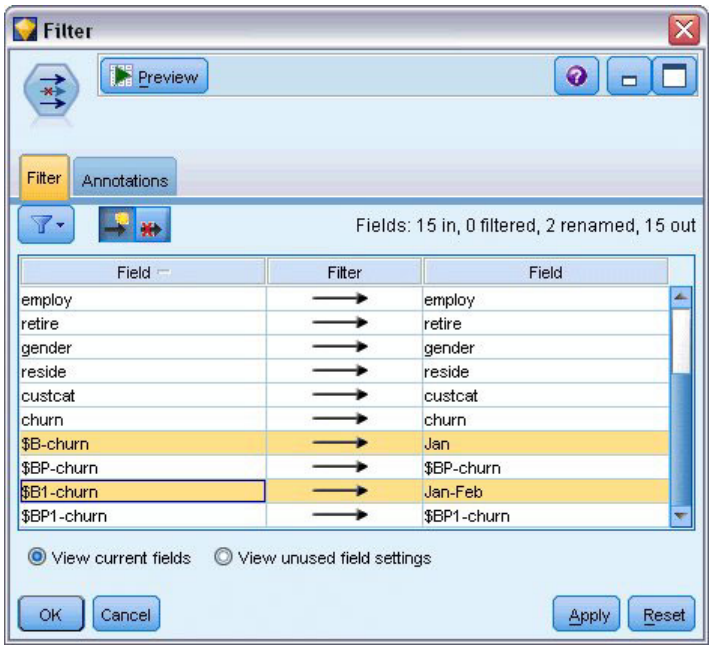


图 255. 重新命名模型字段名称

要检查每个模型预测流失的好坏，请使用分析节点；这样将显示依据正确和不正确的预测百分比得出的准确性。

9. 将“分析”节点附加到“过滤”节点。
10. 打开分析节点并单击**运行**。

这表明两个模型在预测流失时具有类似精确度。

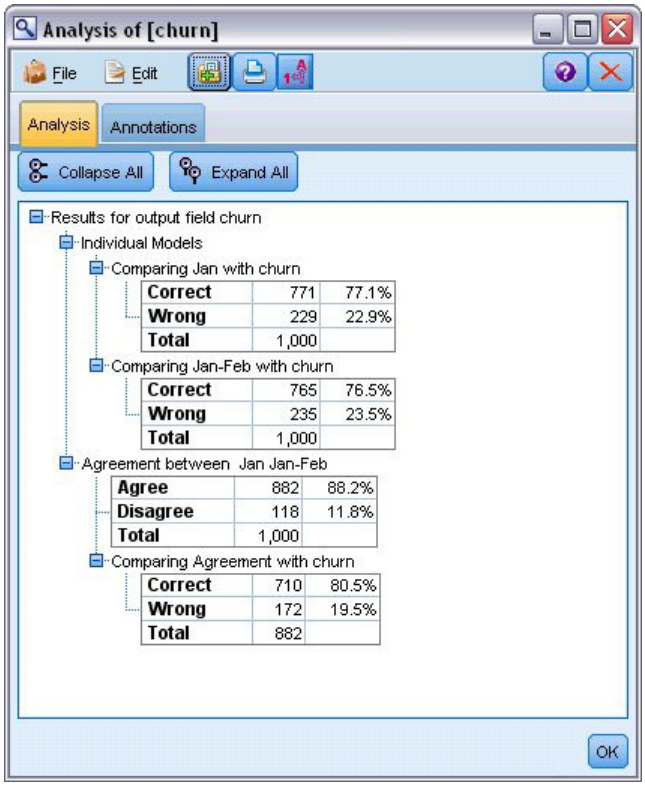


图 256. 分析模型准确性

您可以使用评估图形代替分析节点，通过构建一个增益图比较模型的预测准确性。

11. 将“评估”图形节点附加到“过滤”节点。

并使用其缺省设置执行图形节点。

与分析节点相同，该图形显示两个模型类型都生成了相似的结果；但是，因为使用两个月数据的重新训练模型在预测中具有更高水平的置信度，所以要稍微好一些。

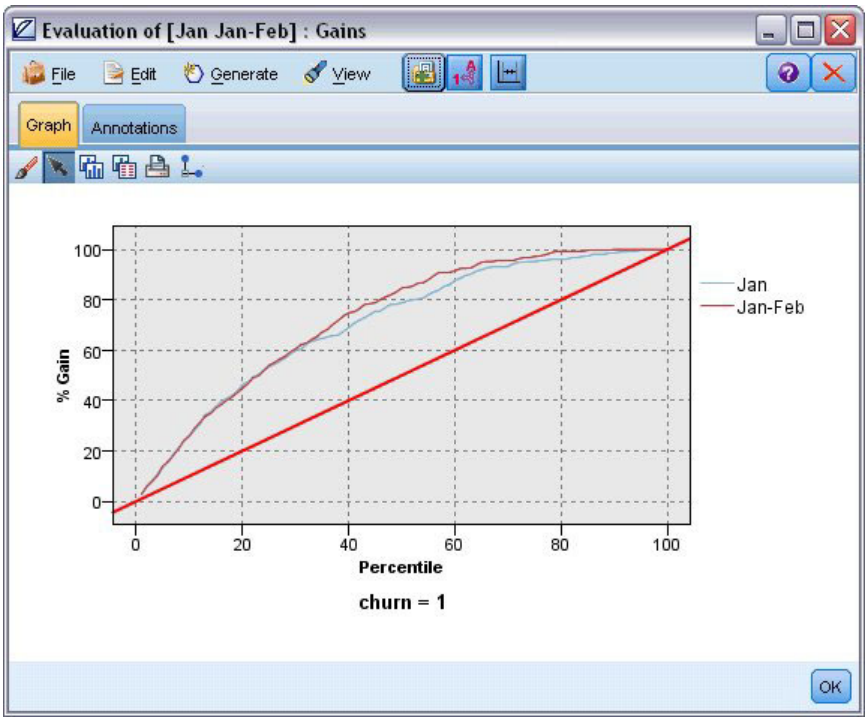


图 257. 评估模型准确性

有关 IBM SPSS Modeler 中所用建模方法的数学原理的说明，请参阅 *IBM SPSS Modeler Algorithms Guide*，该指南位于安装光盘的 *\Documentation* 目录中。

另请注意，这些结果仅基于训练数据。要评估模型适用于实际应用中的其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

第 19 章 零售促销（神经网络/C&RT）

此示例使用数据来说明零售产品线和促销对销售的影响。（此数据纯为虚构。）在此示例，您的目标在于预测未来促销活动的影响。与条件监视示例类似，数据挖掘过程包括探索、数据准备、训练和检验阶段。

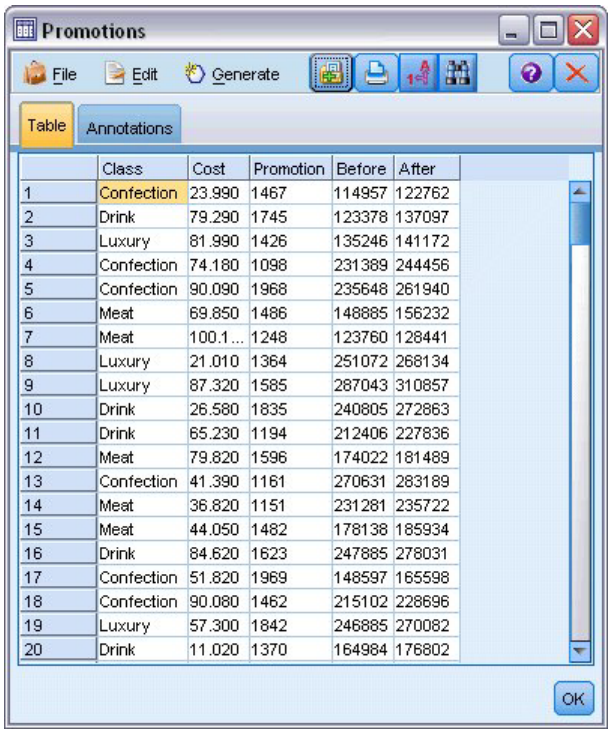
此示例使用名为 *goodsplot.str* 和 *goodslearn.str* 的流，这些流引用名为 *GOODS1n* 和 *GOODS2n* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。流 *goodsplot.str* 在 *streams* 文件夹中，而 *goodslearn.str* 文件在 *streams* 目录中。

检查数据

每条记录含有:

- *类别*。模型类型。
- *成本*。单价。
- *促销*。特定促销上所花费金额的指数。
- *促销前*。促销之前的收入。
- *促销后*。促销之后的收入。

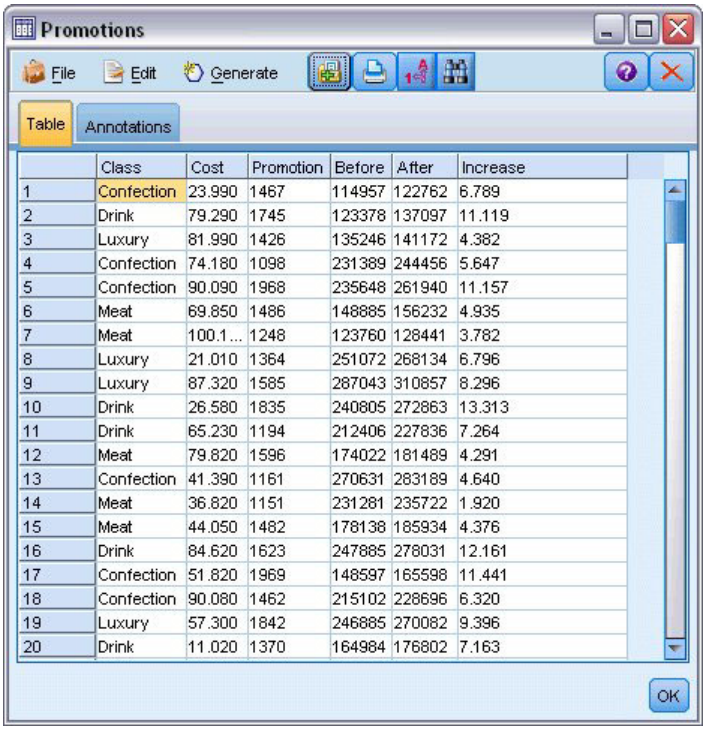
流 *goodsplot.str* 含有一个用于在表格中显示数据的简单流。两个收入字段（即 *Before* 和 *After*）用绝对值来表示；但是，可能促销后收入的增长量（并假定收入增长源于促销）是更有用的数据。



	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

图 258. 促销对产品销售的影响

goodsplot.str 也包含引导出该值的节点，然后在名称为*增长量*的字段中用促销前的收入百分比来表达该值，并显示一个带有该字段的表格。



	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

图 259. 促销后的收入增长量

另外，流将显示一个增长量的直方图和一个以促销费用为参照的增长量的散点图，产品的各个类别的散点图将叠放在一起。

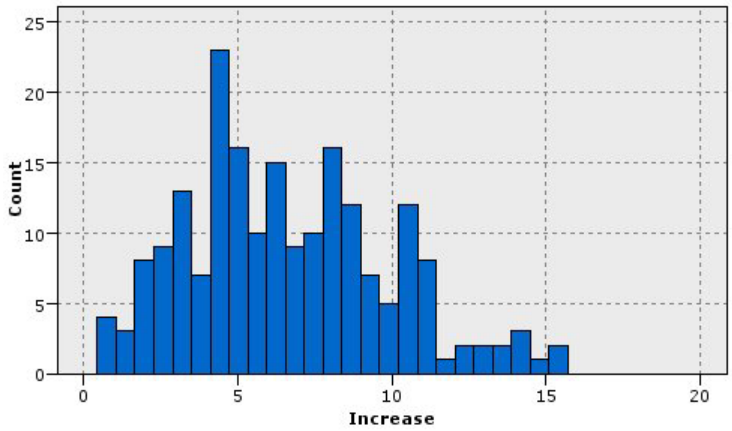


图 260. 收入增长量直方图

散点图显示对于每类产品，收入增长量和促销成本之间存在准线性关系。因此，决策树或神经网络似乎可以合理和准确地预测其他可用字段上的收入增长量。

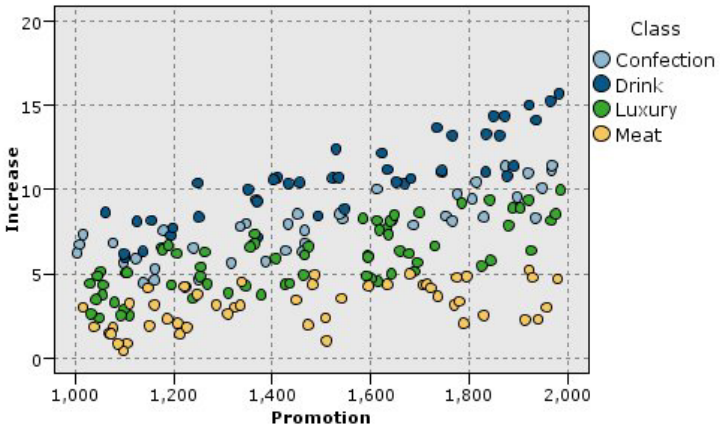


图 261. 收入增长量与促销费用

学习和检验

流 *goodslearn.str* 将训练神经网络和决策树，以对收入增长量做出预测。

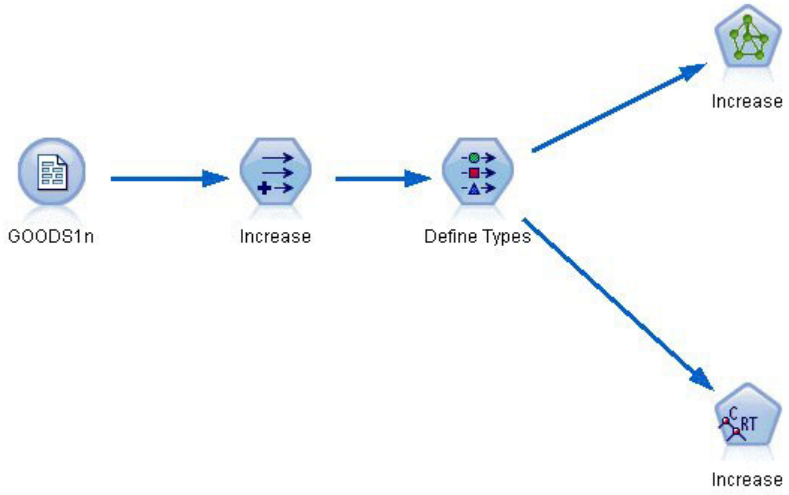


图 262. 对流 goodslearn.str 进行建模

执行模型节点并生成实际模型后，您可以检验学习过程的效果。检验方法如下：将“类型”节点和新“分析”节点之间的决策树和网络串联起来，接着将输入（数据）文件更改为 *GOODS2n*，然后执行“分析”节点。按照此节点的输出数据，特别是按照预测的增长量与正确答案之间的线性相关进行判断，可以发现已训练系统对收入增长量的预测成功率颇高。

进一步的探索可以侧重于已训练系统产生相对较大误差的个案；可通过绘制收入的预测增长量与实际增长量的对比图来确定这些个案。可使用 IBM SPSS Modeler 的迭代图来选择图上的离群值，而依据离群值的属性，通过调整数据说明和学习过程，提高预测的准确性将成为可能。

第 20 章 状态监测（神经网络/C5.0）

本示例涉及监测机器的状态信息以及与识别和预测故障状态相关的问题。其中的数据通过虚构模拟创建得到并包括大量按时间测量的连续序列。每个记录都是与计算机的以下方面相关的快照报告:

- *时间*。整数。
- *功率*。整数。
- *温度*。整数。
- *压力*。0 表示正常，1 表示瞬时压力报警。
- *正常运行时间*。上次运行时间。
- *状态*。正常情况下是 0，发生错误时更改为错误代码（101、202 或 303）。
- *结果*。在此时间序列中显示的错误代码，或者为 0（如果未发生错误）。（提供这些代码唯一的好处是可在事后了解出现的错误。）

此示例使用名为 *condplot.str* 和 *condlearn.str* 的流，这些流引用名为 *COND1n* 和 *COND2n* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *condplot.str* 和 *condlearn.str* 都位于目录 *streams* 下。

对于每个时间序列，都会对应一列正常运行期间产生的记录，后跟一列非正常运行期间产生的故障记录，如下表所示:

时间	幂	温度	压力	正常工作时间	状态	结果
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

通常，大多数数据挖掘工程都会经历以下过程:

- 检查数据以确定哪些属性可能与相关状态的预测或识别有关。
- 保留这些属性（如果已存在），或者在必要时导出这些属性并将其添加到数据中。
- 使用结果数据训练规则和神经网络。
- 使用独立测试数据测试经过训练的系统。

检查数据

文件 *condplot.str* 说明上述过程的第一部分。该文件包含绘制大量图形的流。如果温度或功率的时间序列包含可见的特性曲线，那么您可以区别即将发生的错误情况，或者可以预测这些错误情况的发生。对于温度和功率，下面的流可绘制与单独图形中的三个不同错误代码相关联的时间序列，并生成六个图形。选择节点可分隔与不同错误代码关联的数据。

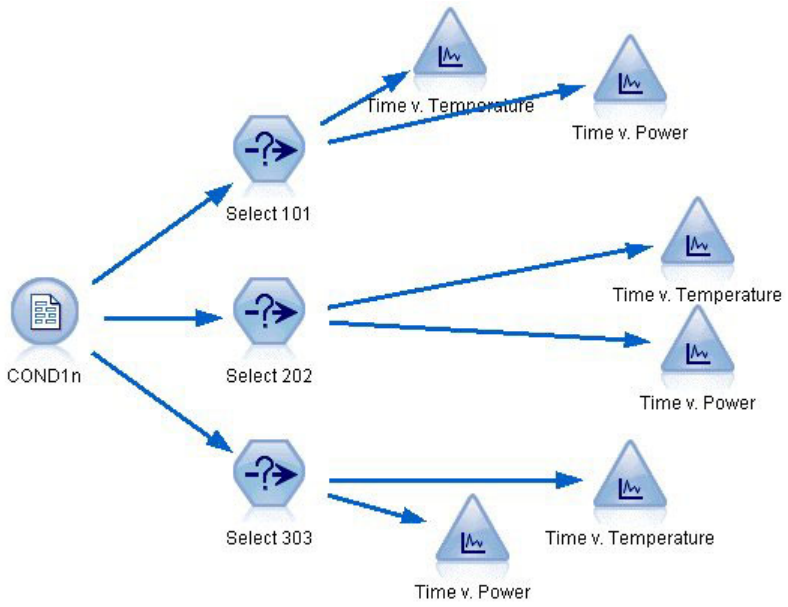


图 263. Condplot 流

该流的结果显示在下图中。

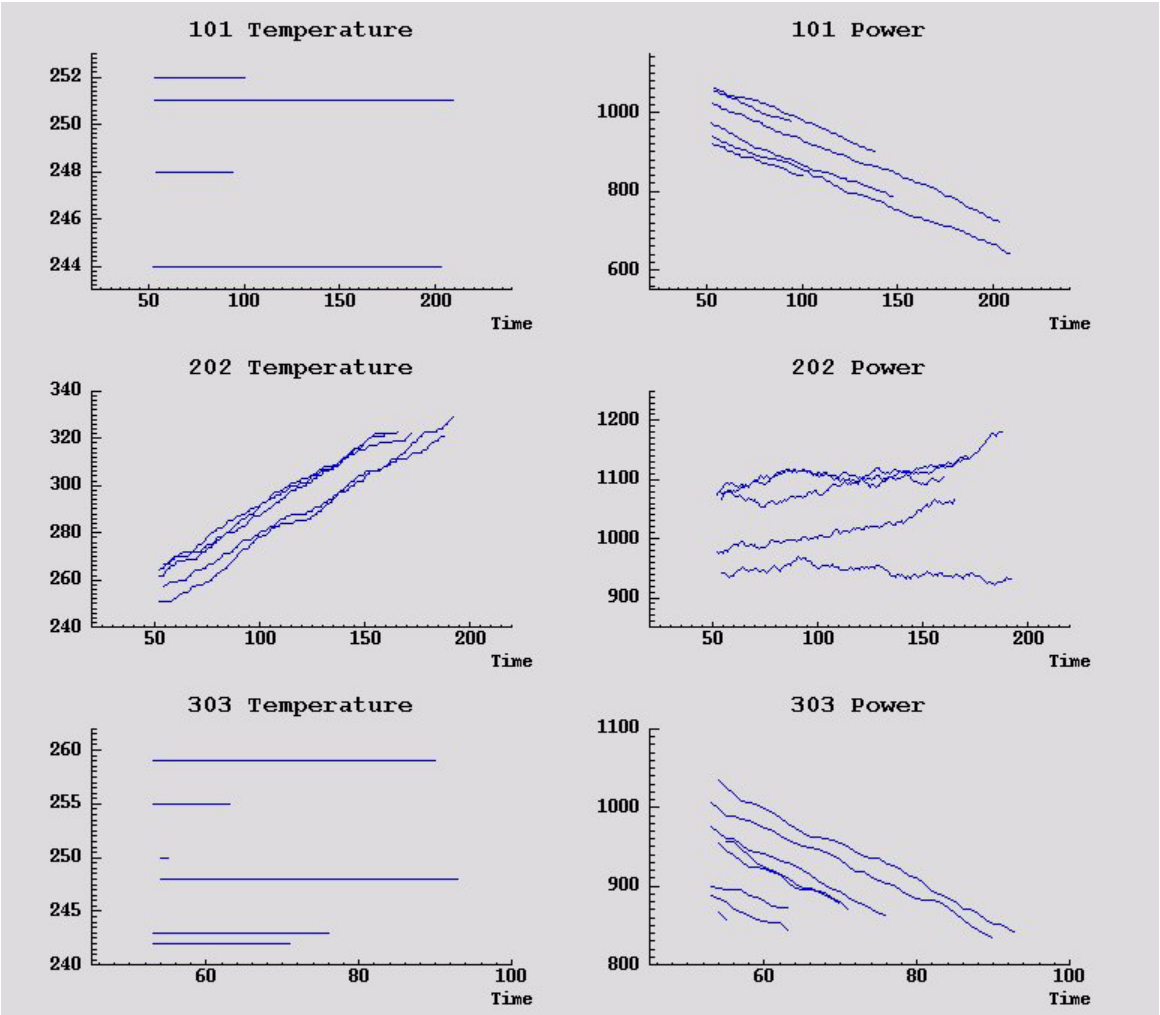


图 264. 随时间推移而变化的温度和功率

这些图形清楚地显示了将 202 错误与 101 和 303 错误区分开来的特性曲线。202 错误显示随着时间的推移温度不断上升并且功率发生波动；而其他两个错误未显示此内容。但是，用于区分 101 和 303 错误的特性曲线不够明确。这两个错误图都显示了平滑的温度曲线和功率的下降，但 303 错误图中的功率下降显得更加急剧一些。

根据上述图形，似乎温度和功率的变化和变化率以及波动的存在和波动程度都与预测故障及区别故障相关。因此应先将这些属性添加到数据，然后再应用学习系统。

数据准备

根据数据研究结果，流 *condlearn.str* 可导出相关数据并学习如何预测故障。

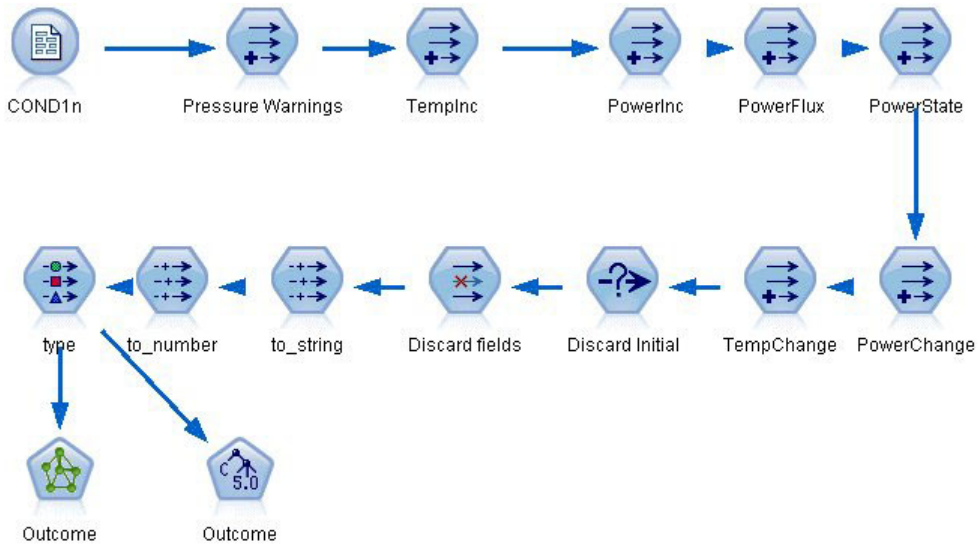


图 265. Condlearn 流

此流使用大量“派生”节点来执行数据的建模准备工作。

- **“变量文件”节点**。读取数据文件 *COND1n* 。
- **Derive Pressure Warnings**。对瞬时压力警报数进行计数。当时间返回到 0 时重置。
- **Derive TempInc**。使用 `@DIFF1` 计算温度瞬时变化率。
- **Derive PowerInc**。使用 `@DIFF1` 计算功率瞬时变化率。
- **Derive PowerFlux**。这是一个标志，如果在上一个记录和本记录中功率按相反的方向变化（即功率的峰值或波谷值），则此值为真。
- **Derive PowerState**。起始为 *稳定* ，当检测到两个连续的功率波动时，切换为 *波动* 的状态。仅当五个时间区间内都没有出现功率波动或重置 *时间* 时，才切换回 *稳定* 状态。
- **PowerChange**。最近五个时间区间内 *PowerInc* 的平均值。
- **TempChange**。最近五个时间区间内 *TempInc* 的平均值。
- **Discard Initial（选择）**。丢弃每个时间序列的第一个记录，以避免在 *功率* 和 *温度* 的边界处出现大的（不正确的）跳跃。
- **Discard fields**。削减记录字段，只保留 *正常工作时间* 、 *状态* 、 *结果* 、 *压力报警* 、 *PowerState* 、 *PowerChange* 和 *TempChange* 字段。
- **Type**。将*结果*的角色定义为**目标**（要预测的字段）。此外，将*结果*的测量级别定义为**名义**、将*压力报警*的测量级别定义为**连续**，将 *PowerState* 的测量级别定义为**标志**。

学习

运行 *condlearn.str* 中的流可训练 C5.0 规则和神经网络。训练网络需要一段时间，但可以提前中断训练以保存生成合理结果的网络。学习完成后，管理器窗口右上角的“模型”选项卡将闪烁，以提醒您创建了两个新的模型块：一个表示神经网络，另一个表示规则。

图 266. 包含模型块的模型管理器

还可以将模型块添加到现有的流中，这允许我们测试系统，或导出模型结果。在此示例中，将测试模型结果。

检验

模型块将添加到流中，这两者均已连接到“类型”节点。

1. 如图所示，重新定位模型块，以使类型节点连接到神经网络模型块，后者连接到 C5.0 模型块。
2. 将分析节点附加到 C5.0 模型块。
3. 编辑原始源节点以读取文件 *COND2n*（而不是 *COND1n*），因为 *COND2n* 包含隐藏的测试数据。

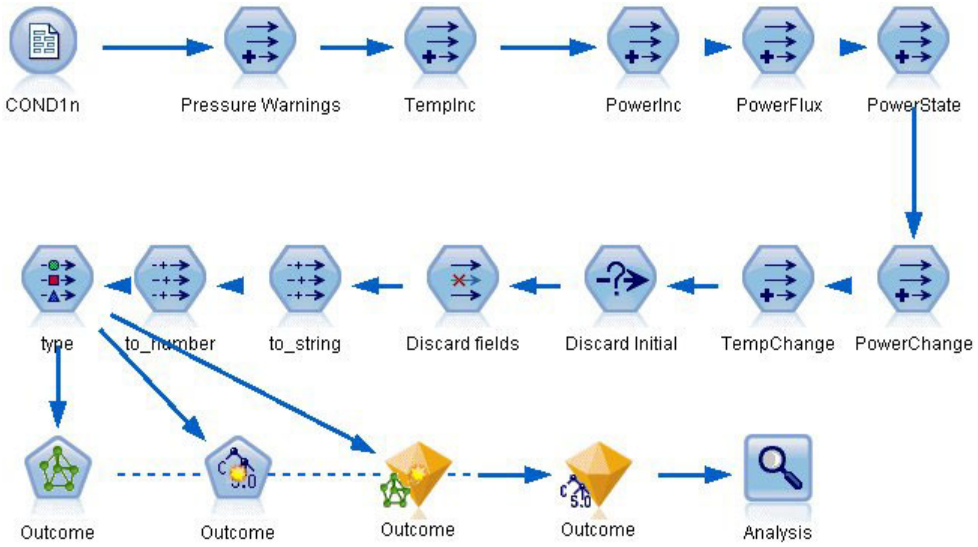


图 267. 测试经过训练的网络

4. 打开分析节点并单击“运行”。

这将生成可反映经过训练的网络和规则的准确性的图表。

第 21 章 电信客户分类（判别分析）

判别分析是一项根据输入字段值对记录进行分类的统计技术。这种技术与线性回归类似，但用分类目标字段代替了数值字段。

例如，假设某个电信提供商根据服务使用情况模式对其客户群进行了细分，将这些客户分为了四个组。如果人口统计数据可用于预测组成员资格，那么您可以为各个潜在客户定制报价。

本示例使用的流名为 *telco_custcat_discriminant.str* ，该流引用名为 *telco.sav* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *telco_custcat_discriminant.str* 位于 *streams* 目录下。

本示例主要讲述使用人口统计数据预测使用情况模式。目标字段 *客户类别* 具有四个可能的值，分别对应四个客户组，如下所示:

值(V)	**标签**
1	基本服务
2	电子服务
3	增值服务
4	全套服务

创建流

1. 首先，设置流属性，以便在输出中显示变量标签和值标签。在菜单中选择:

   文件 > **流属性...** > **选项** > **常规**
2. 确保选择**在输出中显示字段和值标签**，然后单击**确定**。

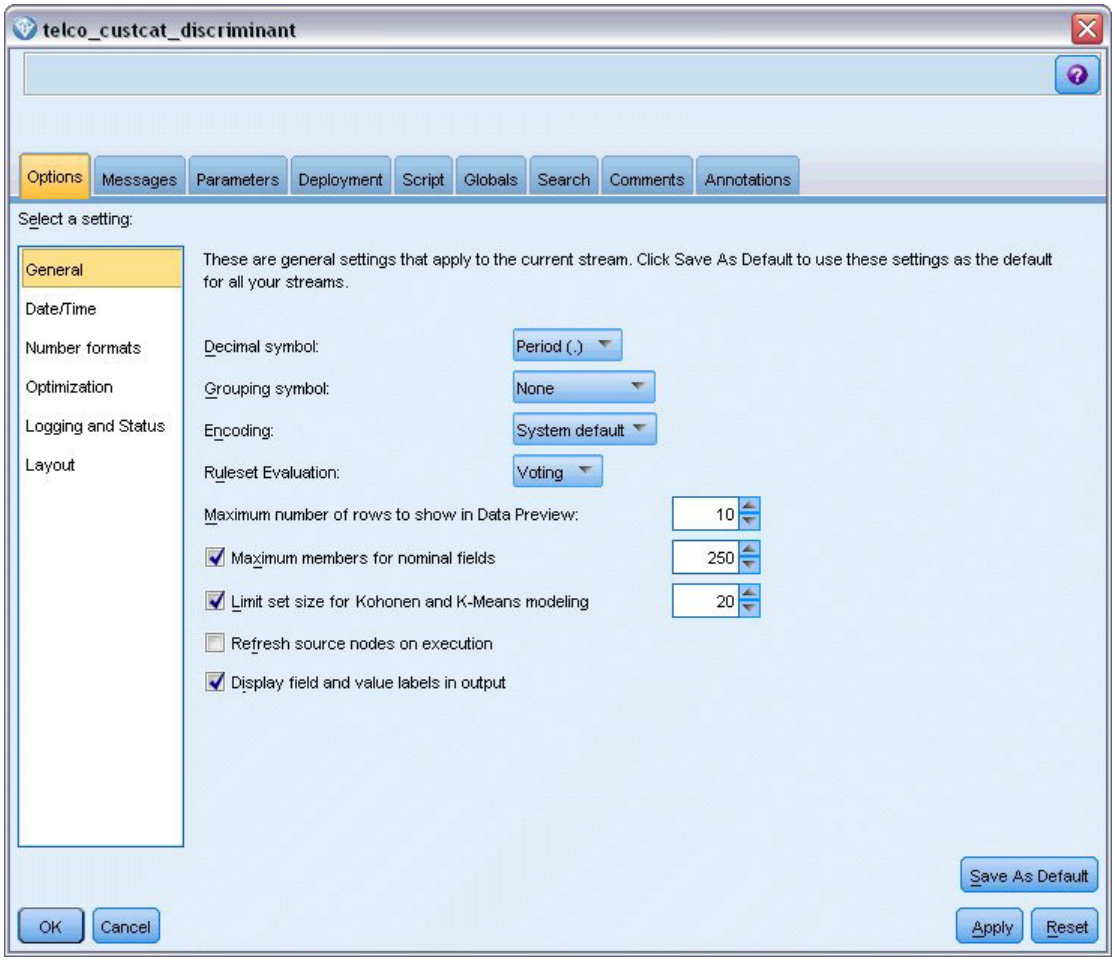


图 268. 流属性

3. 在 *Demos* 文件夹中添加指向 *telco.sav* 的“Statistics 文件”源节点。

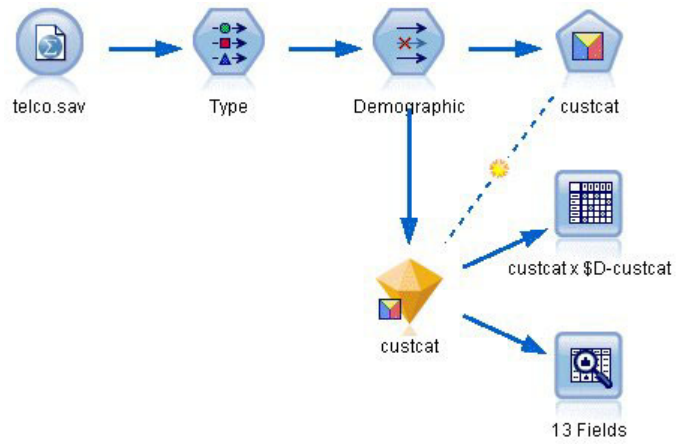


图 269. 使用判别分析对客户进行分类的样本流

a. 添加类型节点并单击**读取值**，确保所有测量级别设置正确。例如，具有值 0 和 1 的多数字段可视为标志。

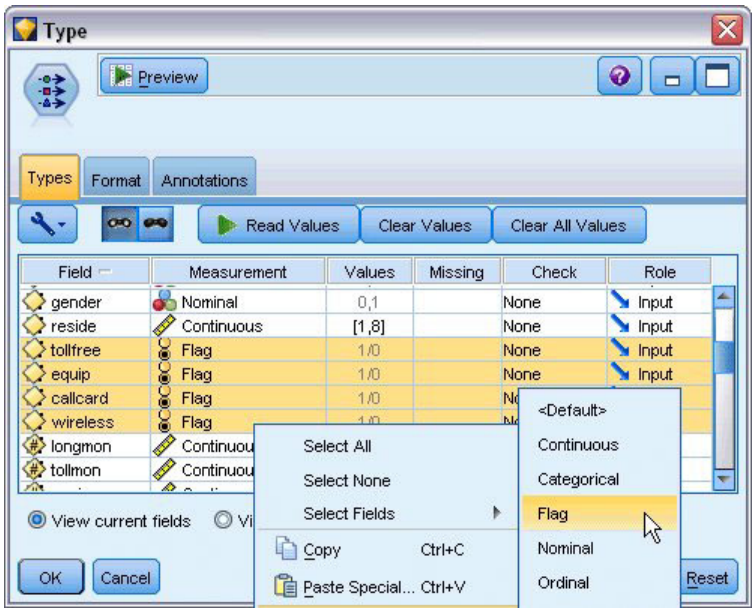


图 270. 设置多个字段的测量级别

提示: 要更改具有相似值（例如 0/1）的多个字段的属性，请单击*值*列标题以按照值对字段进行排序，然后在按住 Shift 键的同时使用鼠标或箭头键选择所有要更改的字段。然后可以右键单击选定的内容以更改选定字段的测量级别或其他属性。

注意，*性别*更准确而言应视为具有两个值的集合的字段，而不是标志，所以将其测量值保留为**名义**。

b. 将*客户类别*字段的角色设置为**目标**。将所有其他字段的角色设置为 **Input**。

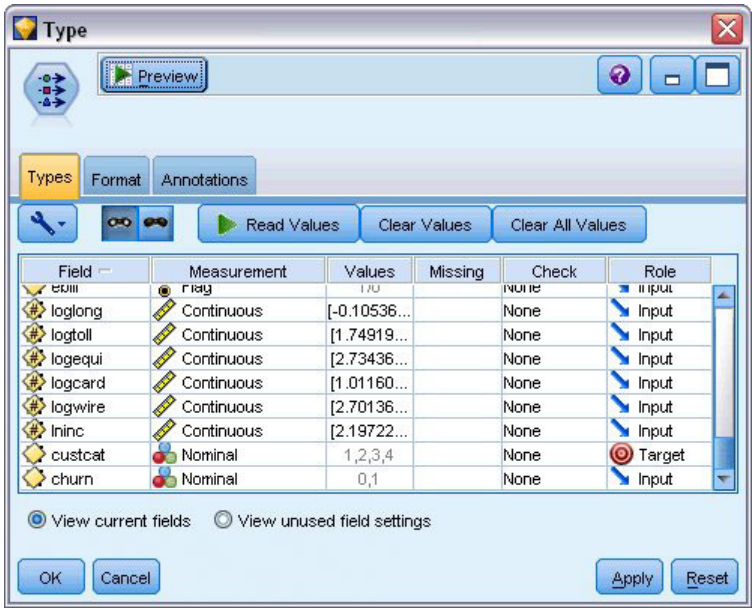


图 271. 设置字段角色

因为本示例主要讲述人口统计，所以请使用过滤节点以选取相关字段（ *地区* 、 *年龄* 、 *婚姻状况* 、 *地址* 、 *收入* 、 *教育程度* 、 *行业* 、 *退休* 、 *性别* 、 *居住地* 和 *客户类别* ）。其他字段可以排除在此

分析之外。

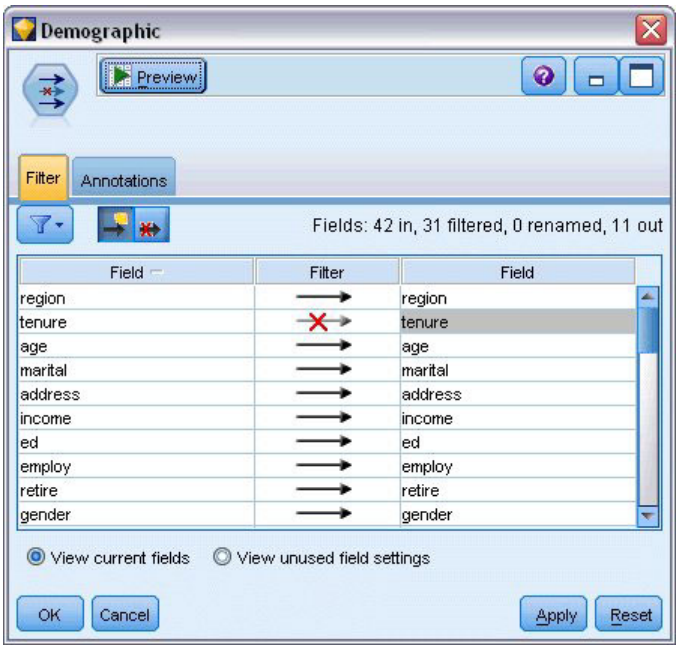


图 272. 过滤人口统计字段

（另外，您可以将这些字段的角色更改为**无**，而不要排除这些字段，或者选择要在建模节点中使用的字段。）

4. 在判别节点中，单击“模型”选项卡，然后选择**步进法**。

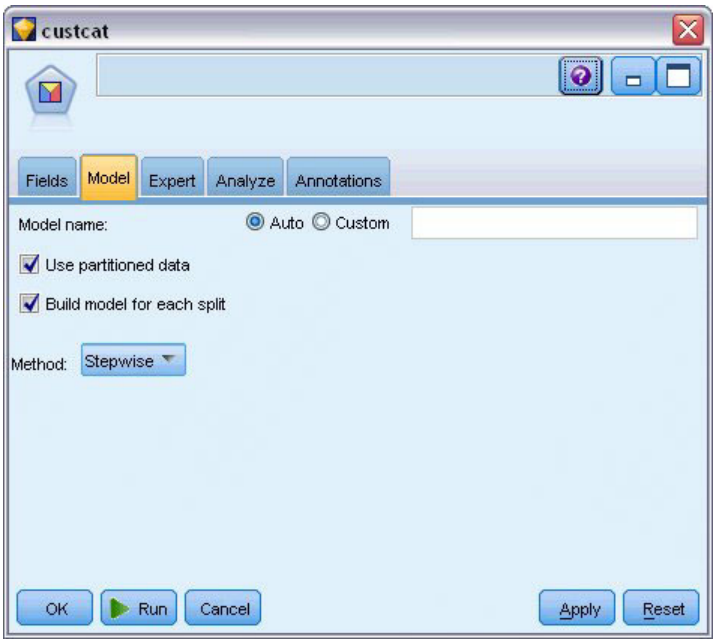


图 273. 选择模型选项

5. 在“专家”选项卡上，将模式设置为**专家**，然后单击**输出**。
6. 在“高级输出”对话框中，选择**汇总表**、**区域图**和**步骤汇总**，然后单击**确定**。

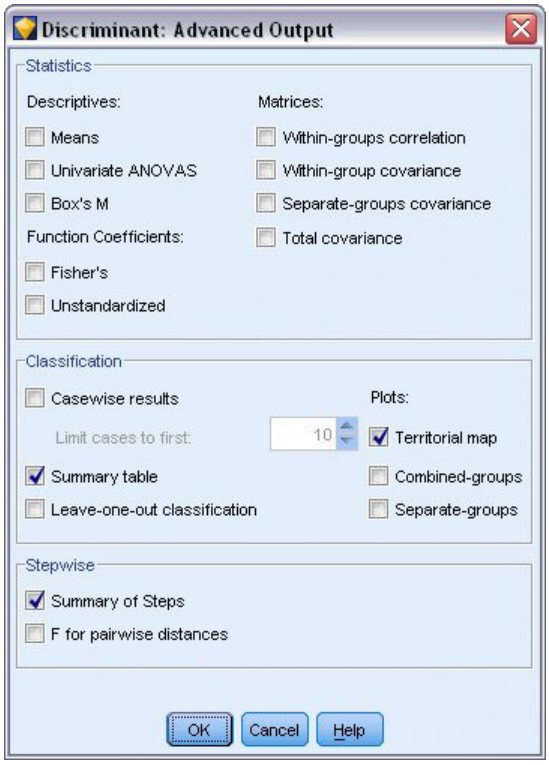


图 274. 选择输出选项

检查模型

1. 单击**运行**以创建模型，该模型将添加到流和右上角的“模型”选用板中。要查看其详细信息，双击流中的模型块。

   “汇总”选项卡显示目标（还有其他内容），以及针对考虑事项提交的完整输入（预测变量字段）列表。

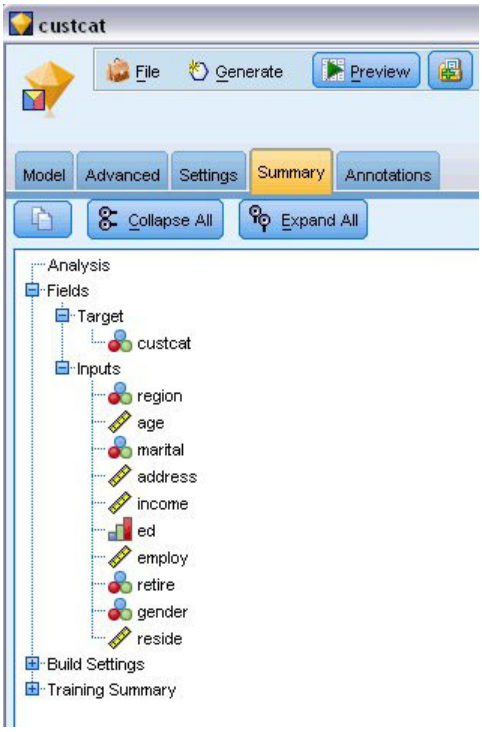


图 275. 显示目标字段和输入字段的模型摘要

有关判别分析结果的详细信息:

2. 单击“高级”选项卡。
3. 单击“在外部浏览器中启动”按钮（就在“模型”选项卡下）以在您的 Web 浏览器中查看结果。

使用判别分析对电信业客户进行分类的分析结果

逐步判别分析

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
	Retired	1.000	1.000	3.005	.991
	Years with current employer	1.000	1.000	16.976	.951
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988

图 276. 未包含在分析的步骤 0 中的变量

拥有大量预测变量时，步进法有助于自动选择“最适合的”用于模型的变量。步进法的最初模型不包括任何预测变量。在每个步骤中，会将具有超出输入标准值（缺省为 3.84）的最大 *F to Enter* 值的预测变量添加到模型中。

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
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

图 277. 未包含在分析的步骤 3 中的变量

在最后一个步骤中保留在分析之外的变量具有的 *F to Enter* 值都小于 3.84，因此不再向分析中添加其他变量。

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

图 278. 分析中包含的变量

此表显示了包括在每个步骤的分析中的变量的统计信息。*容差* 指该变量的方差中不能由方程式的其他自变量解释的部分所占比例。容差很小的变量可以向模型提供的信息很少，并且可能导致计算问题。

F to Remove 值有助于描述从当前模型中删除某个变量（假设保留其他变量）时发生的情况。输入变量的 *F to Remove* 与上述步骤中的 *F to Enter* 相同（显示于“不包括在分析中的变量”表）。

有关步进法的警告说明

步进法很方便，但也有其局限。请注意，因为步进法仅根据统计意义选择模型，所以它有可能选择不具有 *实际意义* 的预测变量。如果您比较熟悉数据并对有重要意义的预测变量有所预期，那么应该利用您的经验而不是使用步进法。但是，如果存在多个预测变量而您不知道从何处着手，则运行逐步分析法并调整选定的模型比完全没有模型要好。

检查模型拟合

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

图 279. 特征值

几乎所有由模型解释的方差都源于前两个判别函数。三个函数可自动拟合，但由于第三个函数特征值极小，可以完全忽视此函数而不用担心安全性。

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

图 280. Wilks' lambda

Wilks' lambda 认同仅有前两个函数是有用的。对于每一个函数集合,该判别将检验各组所列函数的均值相等的假设。对第 3 个函数的检验具有的显著性值大于 0.10，因此该函数对模型而言意义甚微。

结构矩阵

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

a. This variable not used in the analysis.

图 281. 结构矩阵

如果存在多个判别函数，那么将使用星号来标记每个变量与某典范函数的最大绝对相关度。在每个函数内部，这些标记星号 (*) 的变量将按相关度大小排序。

- *教育程度*与第一个函数的相关性最强，并且它是与此函数的相关性最强的唯一变量。
- 虽然*受雇于现任雇主的年数*、*年龄*、*家庭收入（以千计）*、*现住址居住年数*、*是否退休*以及*性别*与第二个函数的相关性最强，但是*性别*和*是否退休*与该函数的相关性弱于其他变量。其他变量将该函数标记为"稳定"函数。
- *家庭成员数*和*婚姻状况*与第三个判别函数的相关性最强，但该函数是无用函数，因此这些变量是几乎无用的预测变量。

区域图

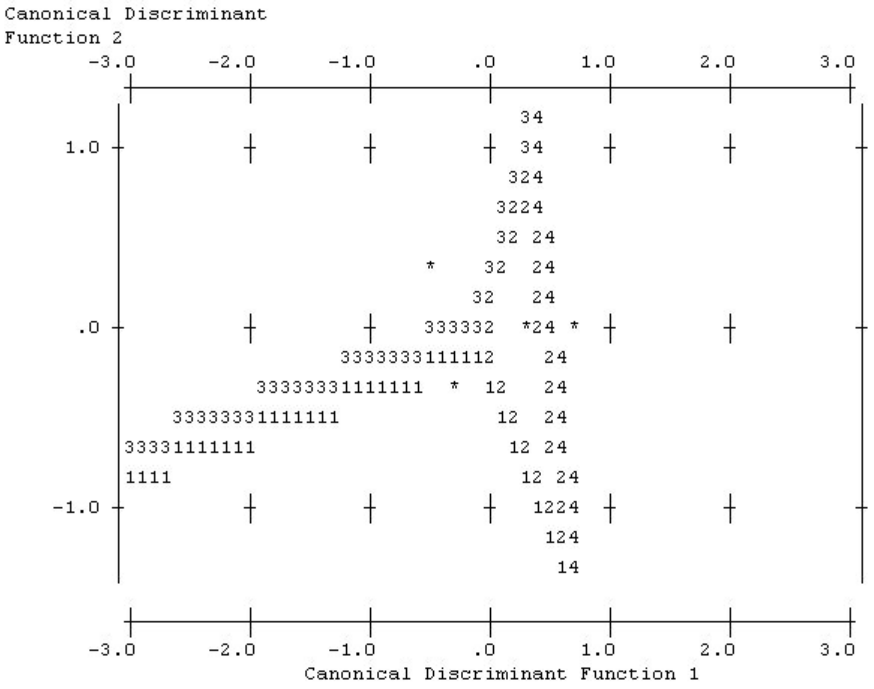


图 282. 区域图

区域图有助于研究组与判别函数之间的关系。结合结构矩阵的结果，区域图能够对预测变量和组之间的关系提供图形化的解释。第一个函数，显示在水平轴上，将组 4（ *全套服务* 用户）从其他组中区分开来。因为 *教育程度* 与第一个函数具有很强的明确的关联度，这表明 *全套服务* 用户通常具有最高的教育程度。第二个函数将组 1 和 3（ *基本服务* 和 *增值服务* 用户）区分开来。*增值服务*客户的工作时间和年龄往往大于*基本服务*客户。尽管区域图表明 *电子服务* 用户受过良好教育并且具有中等工作经验，但无法很好地将它与其他组区分开来。

通常，，标记有星号 (*) 的组的矩心靠近区域边界时，表明所有组间的分隔不是非常强。

区域图仅绘制了前两个判别函数，但由于第三个函数无关紧要,因此区域图提供了判别分析模型的全面视图。

分类结果

			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.

图 283. 分类结果

根据 Wilks' lambda 检验，可以得知模型的预测能力比猜测要强大，但需要借助于分类结果才能确定其强大的程度。对于给定的观测数据，“空”模型（即不包括任何预测变量的模型）将把所有用户分类到*增值服务*模型

组。因此，空模型的正确率将是 281/1000 = 28.1%。模型可获得较之空模型多 11.4% 即 39.5% 的用户。特别是模型在鉴别 *全套服务* 用户时表现优异。然而，它在对 *电子服务* 用户进行分类时表现得格外糟糕。可能需要寻找新的预测变量来区分这些用户。

摘要

已创建了一个判别模型，用于根据每个用户的人口统计学信息将用户分类到四个预定义的“服务使用”组之一。利用结构矩阵和区域图，能够鉴别出那些最有助于分割客户群的变量。最后，分类结果显示模型对 *电子服务* 用户进行分类时表现欠佳。需要进一步研究来确定另一个预测变量，以便更好地对这些用户进行分类，但该模型可能完全能够满足您的需求，这取决于您希望预测的内容。例如，如果您对 *电子服务* 用户的鉴别并不关心，那么该模型可足以满足需求。这种情况可能是，将电子服务作为一种仅为吸引顾客而出售并产生微薄利润的产品。例如，如果投资的最高回报来自于 *增值服务* 或 *全套服务* 用户，则该模型能够提供所需的信息。

另请注意，这些结果仅基于训练数据。要评估该模型适用于其他数据的程度，可以使用“分区”节点提供部分记录以用于测试和验证。

IBM SPSS Modeler Algorithms Guide 中列出了对 IBM SPSS Modeler 中用到的建模方法的数学原理的说明。此文件可在安装光盘的 *\Documentation* 目录中找到。

第 22 章 分析区间型删失的生存数据（广义线性模型）

当分析区间型删失的生存数据时，即不知道所关注事件的准确时间，而只知道事件发生在给定的时间间隔内则可将 Cox 模型应用到时间间隔内的事件危险性并生成互补重对数回归模型。

从研究（此研究的设计目的是比较两种防止溃疡复发的疗法的功效）中获取的部分信息位于 *ulcer_recurrence.sav* 中。此数据集已在其他地方给出并分析 [1]。使用广义线性模型，可以复制互补重对数回归模型的结果。

此示例使用名称为 *ulcer_genlin.str* 的流，此流参考的是数据文件 *ulcer_recurrence.sav* 。数据文件和流文件分别位于 *Demos* 文件夹和 *streams* 子文件夹中。

创建流

1. 在 *Demos* 文件夹中添加一个指向 *ulcer_recurrence.sav* 的“Statistics 文件”源节点。

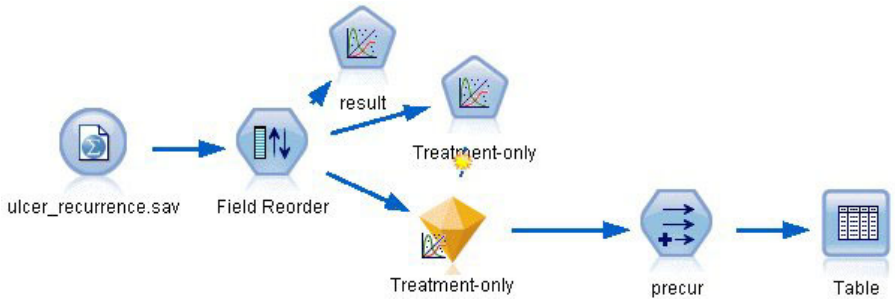


图 284. 用于预测溃疡复发的样本流

2. 在源节点的“过滤”选项卡上，过滤掉 *id* 和*时间*。

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

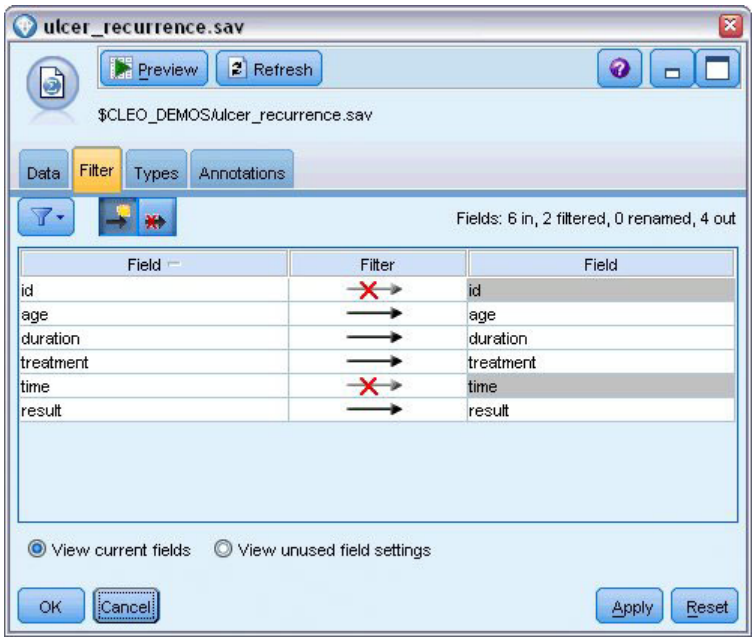


图 285. 过滤不需要的字段

3. 在源节点的“类型”选项卡上，将 *result* 字段的角色设置为 **Target**，将其测量级别设置为 **Flag**。结果为 1 表示溃疡已复发。将所有其他字段的角色设置为 **Input**。
4. 单击**读取值**以实例化数据。

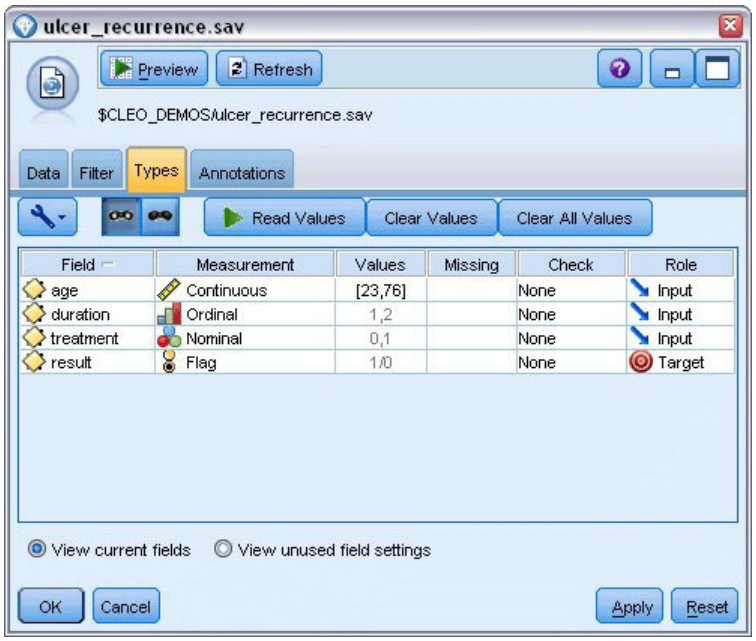


图 286. 设置字段角色

5. 添加字段重排节点并指定 *持续时间* 、 *治疗* 和 *年龄* 作为输入的顺序。此操作将确定在模型中输入字段的顺序并会帮助您尝试复制 Collett 的结果。

图 287. 对字段进行重新排序以使其按预期输入到模型中

6. 将 GenLin 节点附加到源节点；在 GenLin 节点中，单击 **模型** 选项卡。
7. 选择 **第一个（最低值）** 作为目标的参考类别。此操作表示第二个类别是所关注的事件，它对模型的影响在参数估计中进行解释。系数为正的连续预测变量表示复发概率随着预测变量值的增加而增加；相对于其他设置的类别，系数越大的名义预测变量的类别表示复发概率越大。

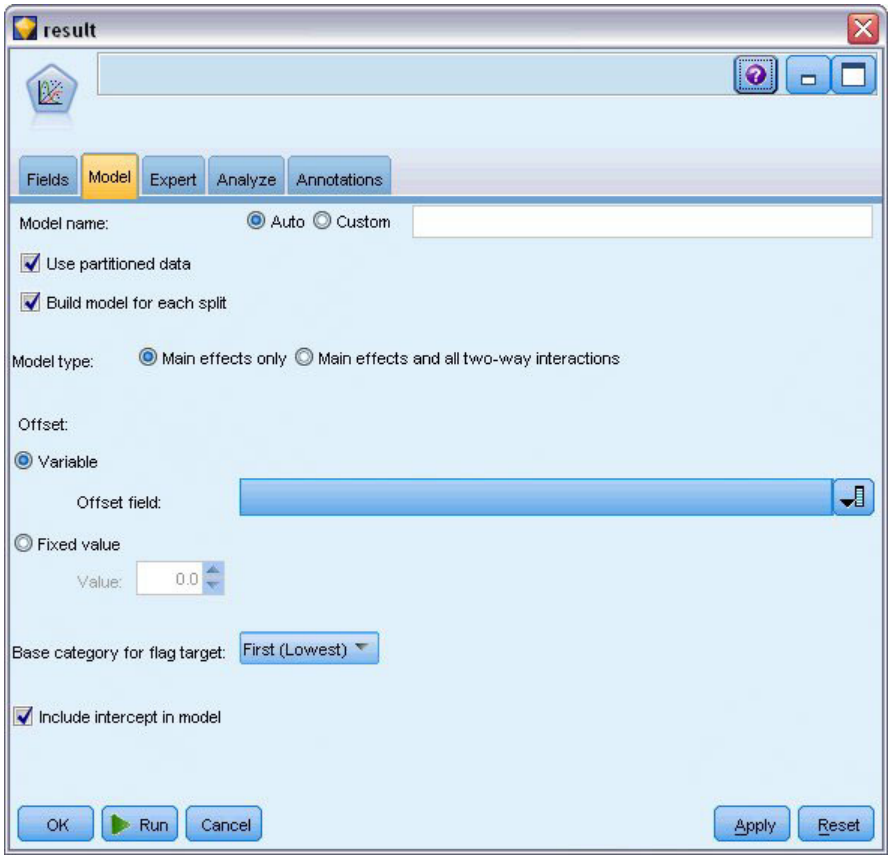


图 288. 选择模型选项

8. 单击 **专家** 选项卡并选择 **专家** 以激活专家建模选项。
9. 选择 **二项** 作为分布， **互补重对数** 作为连接函数。
10. 选择 **固定值** 作为估计尺度参数的方法，并选择缺省值 1.0。
11. 选择**降序**作为因子的类别顺序。这指示每个因子的第一个类别将是其参考类别；模型中此项选择的效应由参数估计解释。

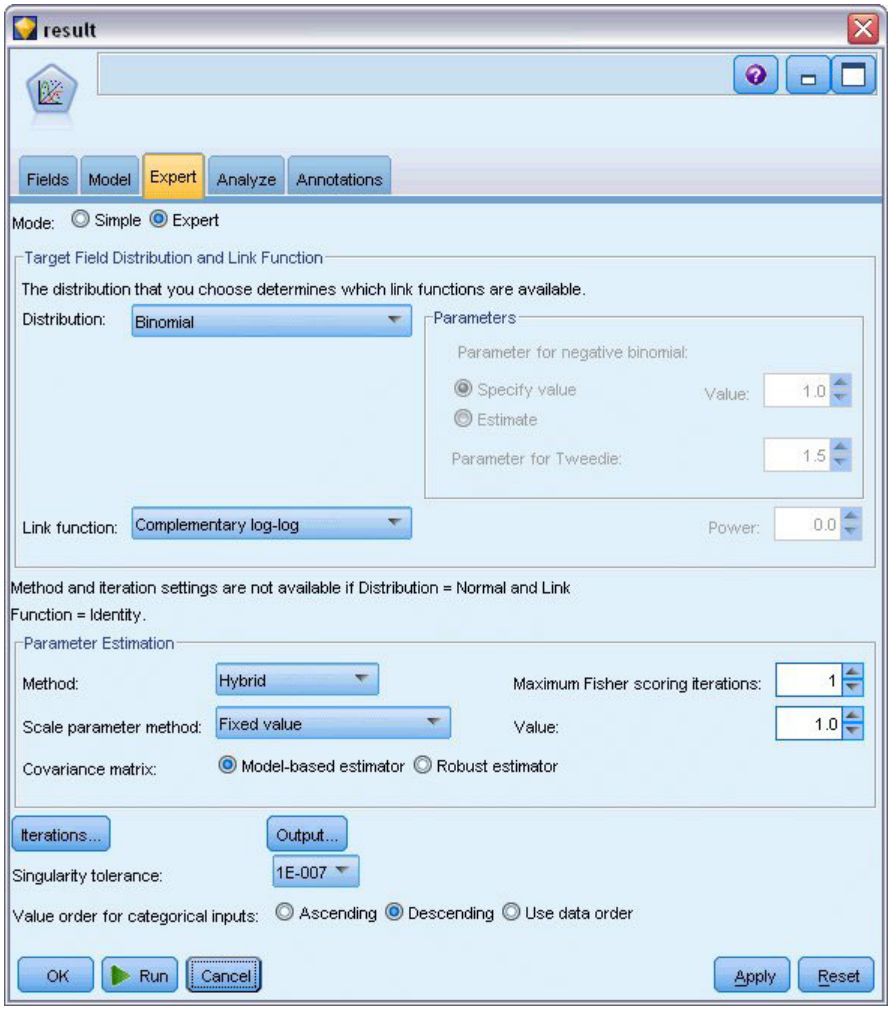


图 289. 选择专家选项

12. 运行流以创建模型块，此模型块将被添加到流工作区和位于右上角的“模型”选用板中。要查看模型详细信息，请右键单击此模型块并选择**编辑**或**浏览**。

模型效应检验

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
duration	.003	1	.958
treatment	.382	1	.537
age	.358	1	.550

Dependent Variable: Result
Model: (Intercept), duration, treatment, age

图 290. 主效应模型的模型效应检验

没有任何的模型效果是在统计意义下显著的；但是，任何可观察到的治疗效果上的差异都具有临床意义，因此我们将仅以治疗作为模型项拟合一个简化模型。

拟合仅治疗模型

1. 在 GenLin 节点的“字段”选项卡上，单击**使用自定义设置**。
2. 选择 *结果* 作为目标。
3. 选择 *治疗* 作为唯一的输入。

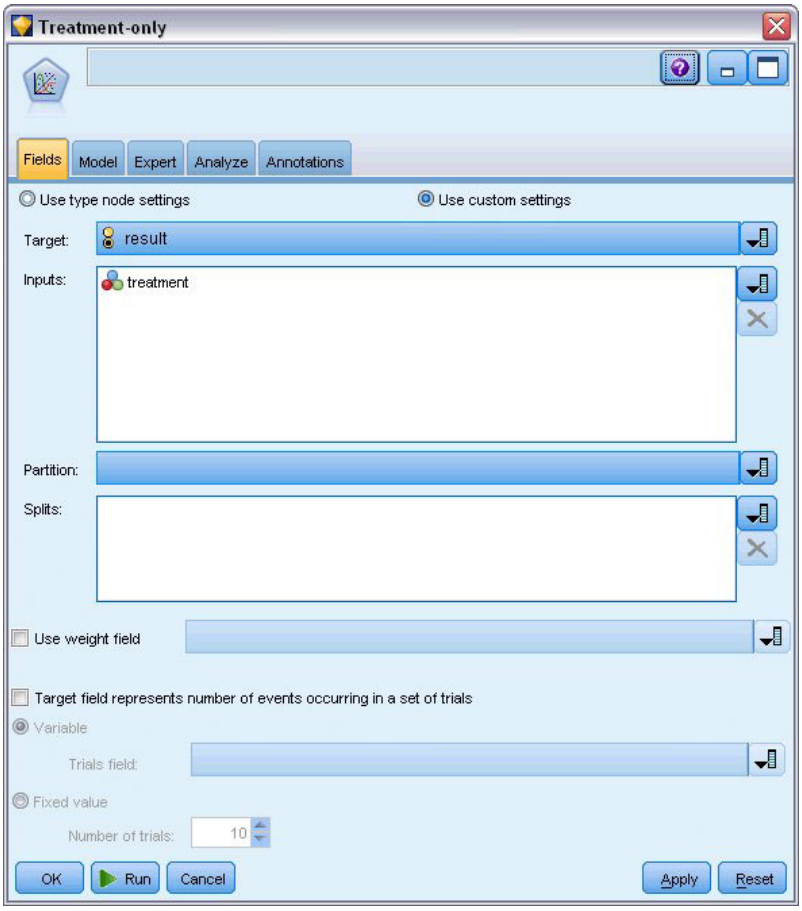


图 291. 选择字段选项

4. 运行流，并打开生成的模型块。

在模型块上，选择**高级**选项卡，并滚动到底部。

参数估计

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=1]	.378	.6288	-.855	1.610	.361	1	.548
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result
Model: (Intercept), treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

图 292. 仅治疗模型的参数估计

治疗效果（两种治疗水平之间的线性预测变量的差异；即 *[治疗 =1]* 的系数）仍然不是统计意义下显著的，而是仅能估计出治疗 *A [治疗 =0]* 可能比 *B [治疗 =1]* 的效果好，因为治疗 *B* 的参数估计大于 *A* 的参数估计，因而与前 12 个月内复发概率增加相关联。线性预测变量（截距 + 治疗效果）是 $\log(-\log(1-P(recur_{12,t}))$ 的估计值，其中 $P(recur_{12,\ t})$ 是治疗（t = *A* 或 *B*）12 个月后的复发概率。可为数据集中的每个观测数据生成这些预测的概率。

预测复发和生存的概率

图 293. “派生”节点设置选项

1. 对于每位患者，模型都可对预测结果和该预测结果的概率进行评分。为查看预测的复发概率，可将生成的模型复制到选用板并附加导出节点。
2. 在“设置”选项卡中，键入 `precur` 作为导出字段。
3. 选择将它作为**条件**导出。
4. 单击计算器按钮可打开 **If** 条件的表达式构建器。

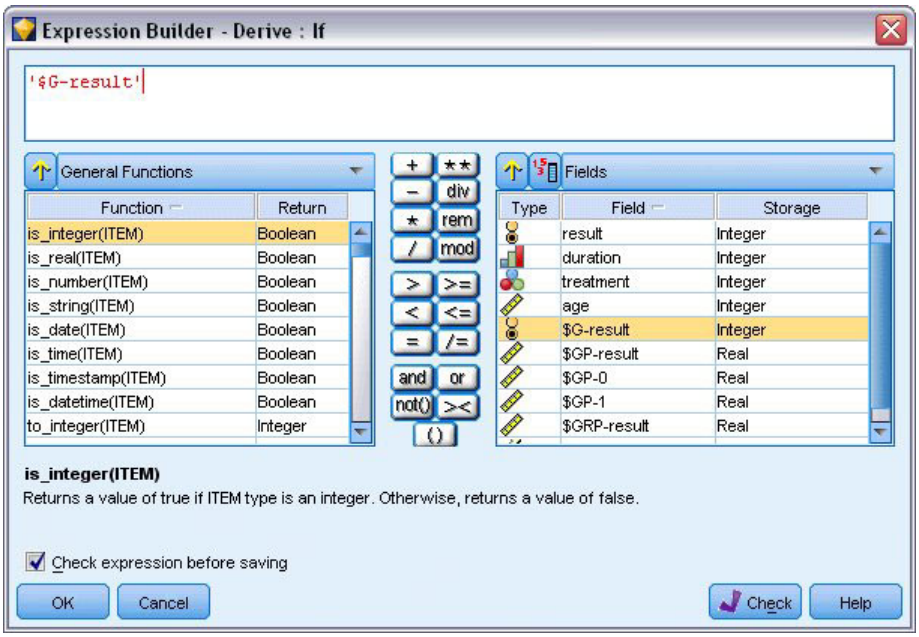


图 294. “派生”节点: If 条件的表达式构建器

5. 将 *$G-result* 字段插入表达式中。
6. 单击**确定**。

导出字段 *precur* 在 *$G-result* 等于 1 时和 0 时分别取 **Then** 表达式和 **Else** 表达式的值。

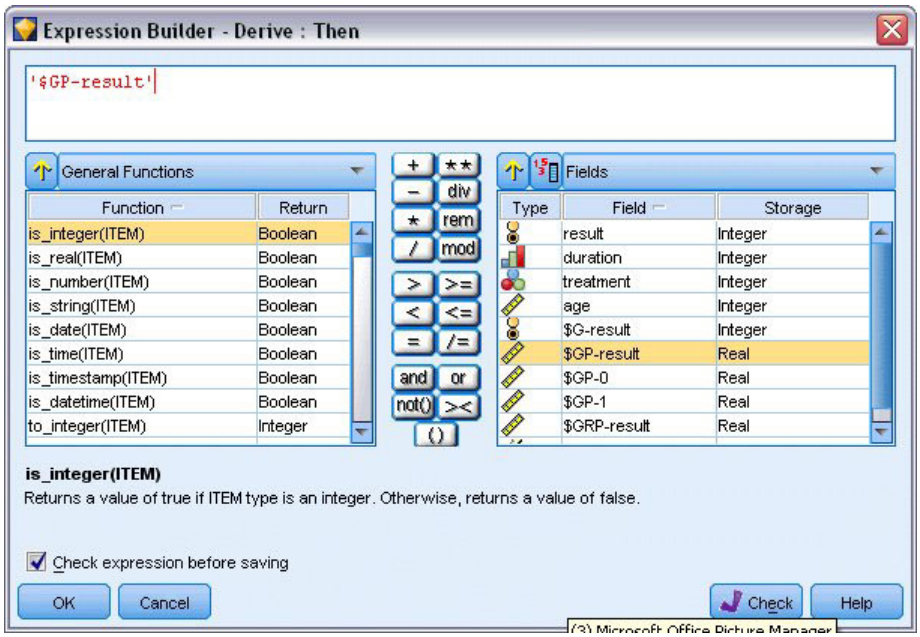


图 295. "派生"节点: Then 表达式的表达式构建器

7. 单击计算器按钮可打开 **Then** 表达式的表达式构建器。
8. 将 *$GP-result* 字段插入表达式中。
9. 单击**确定**。

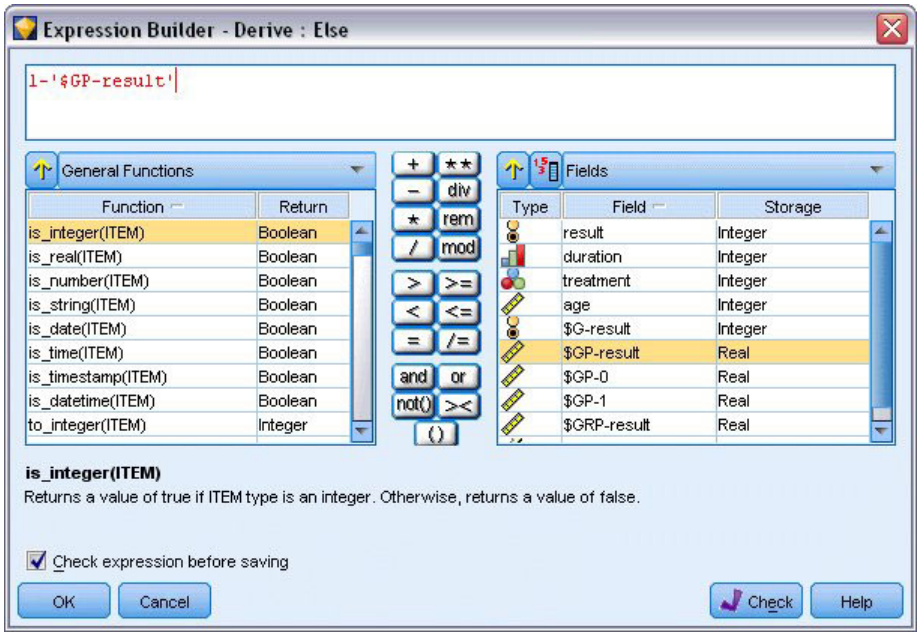
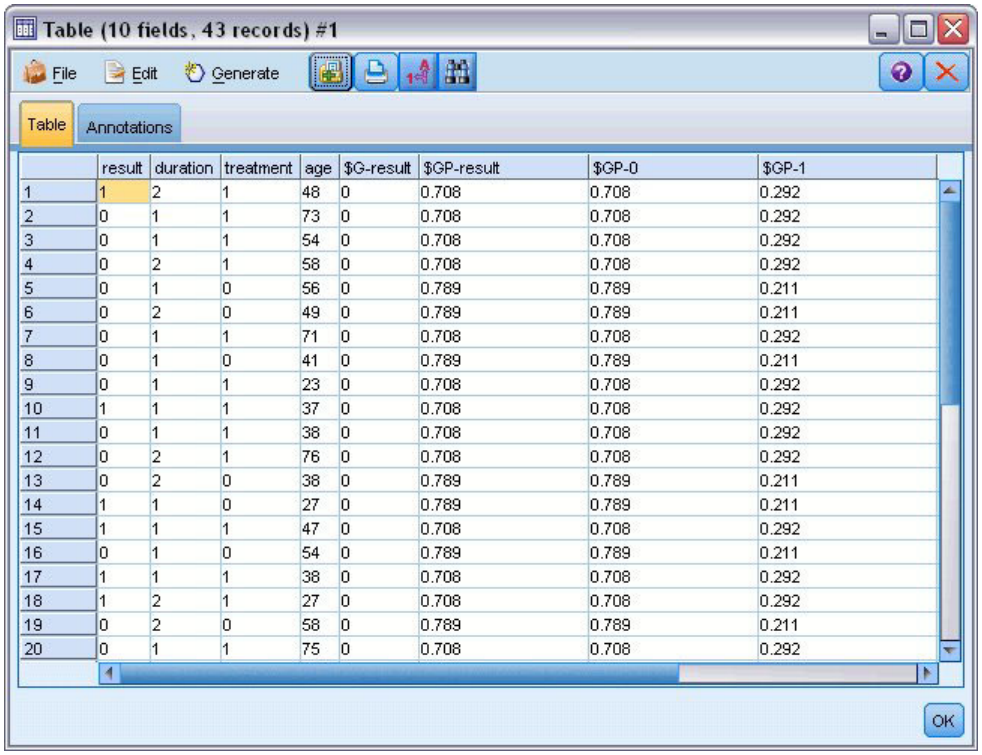


图 296. "派生"节点: Else 表达式的表达式构建器

10. 单击计算器按钮可打开 **Else** 表达式的表达式构建器。
11. 在表达式中输入 1-，然后将 *$GP-result* 字段插入表达式。
12. 单击**确定**。

图 297. “派生”节点设置选项

13. 将表节点附加到“派生”节点并进行执行。



	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

图 298. 预测概率

对于分配到治疗方案 *A* 的患者，他在前 12 个月内病情复发概率的估计值为 0.211；对于分配到治疗方案 *B* 的患者，复发概率的估计值为 0.292。请注意，1-P($recur_{12,\ t}$) 是 12 个月的生存概率，生存分析员将更加关注此概率。

按周期对复发概率进行建模

模型建立时出现的一个问题是它忽略了在第一次检查时所收集的信息；即，许多患者在前六个月内没有病情的复发。“更理想”的模型会模拟二元响应，该响应可记录事件是否会在每个时间间隔内发生。对此模型进行拟合需要重新构造原始数据集，可以在 *ulcer_recurrence_recoded.sav* 中找到此数据集。此文件包含两个附加变量:

- *Period*: 记录病例对应于第一个检查周期还是第二个检查周期。
- *Result by period*: 记录给定患者在指定周期内是否出现病情复发。

在风险集中，每个原始病历（病人）都为其所存在的每个时间间隔提供一个实例。因而，例如，患者 1 提供两个实例；一个在第一次检查周期内，此时病情没有复发，另一个在第二次检查周期内，此时记录到一次病情的复发。另一方面，患者 10 仅提供了一个实例，因为已在第一个周期内记录到病情的复发。患者 16、28 和 34 在六个月后放弃参加研究，因此仅向新数据集提供一个实例。

1. 在 *Demos* 文件夹中添加一个指向 *ulcer_recurrence_recoded.sav* 的“Statistics 文件”源节点。

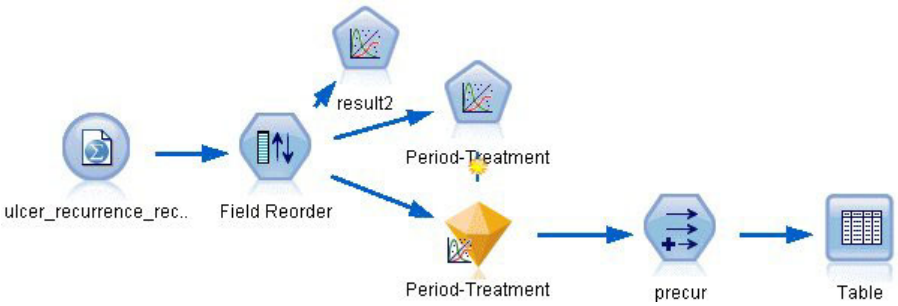


图 299. 用于预测溃疡复发的样本流

2. 在源节点的“过滤”选项卡上，过滤掉 *id*、*时间*和*结果*。

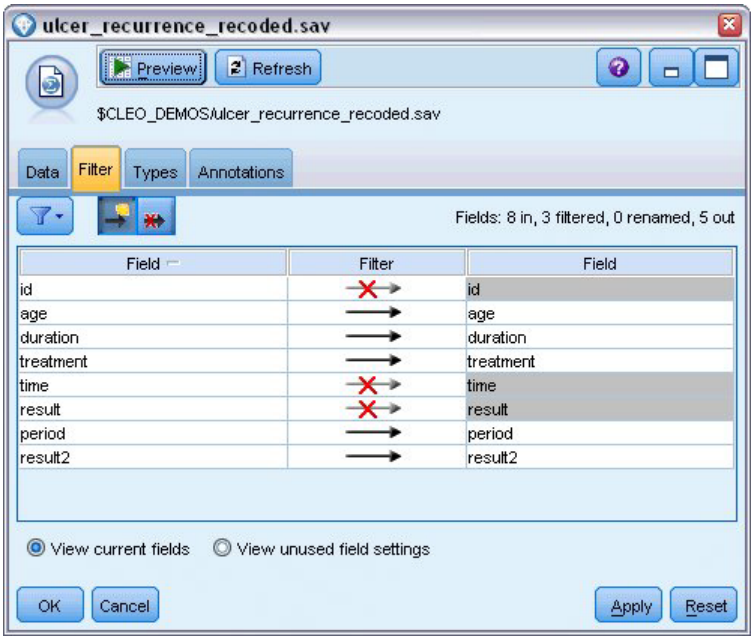


图 300. 过滤不需要的字段

3. 在源节点的“类型”选项卡上，将 *result2* 字段的角色设置为 **Target**，将其测量级别设置为 **Flag**。将所有其他字段的角色设置为 **Input**。

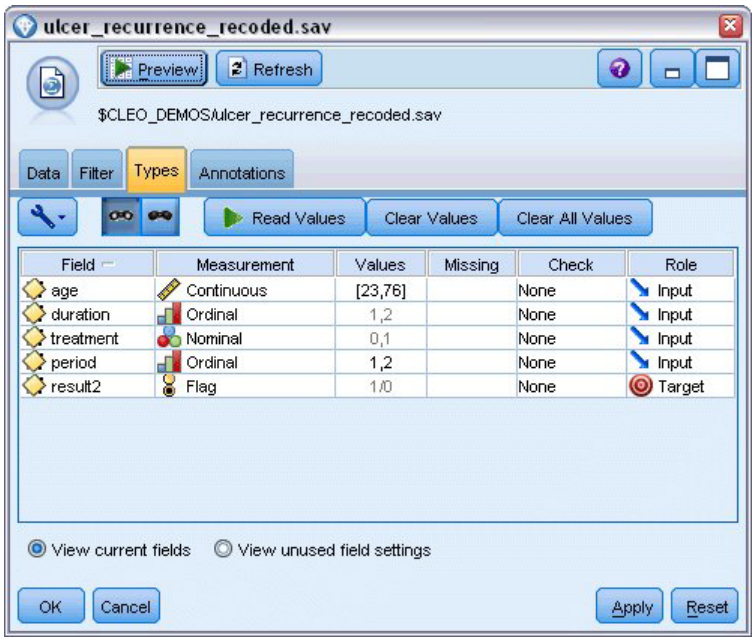


图 301. 设置字段角色

4. 添加字段重排节点并指定 *周期* 、 *持续时间* 、 *治疗* 和 *年龄* 作为输入的顺序。将 *周期* 作为第一个输入（不包括模型中的截距项）使您能够拟合完整的虚设变量集以捕获周期效果。

图 302. 对字段进行重新排序以使其按预期输入到模型中

5. 在 GenLin 节点中，单击 **模型** 选项卡。

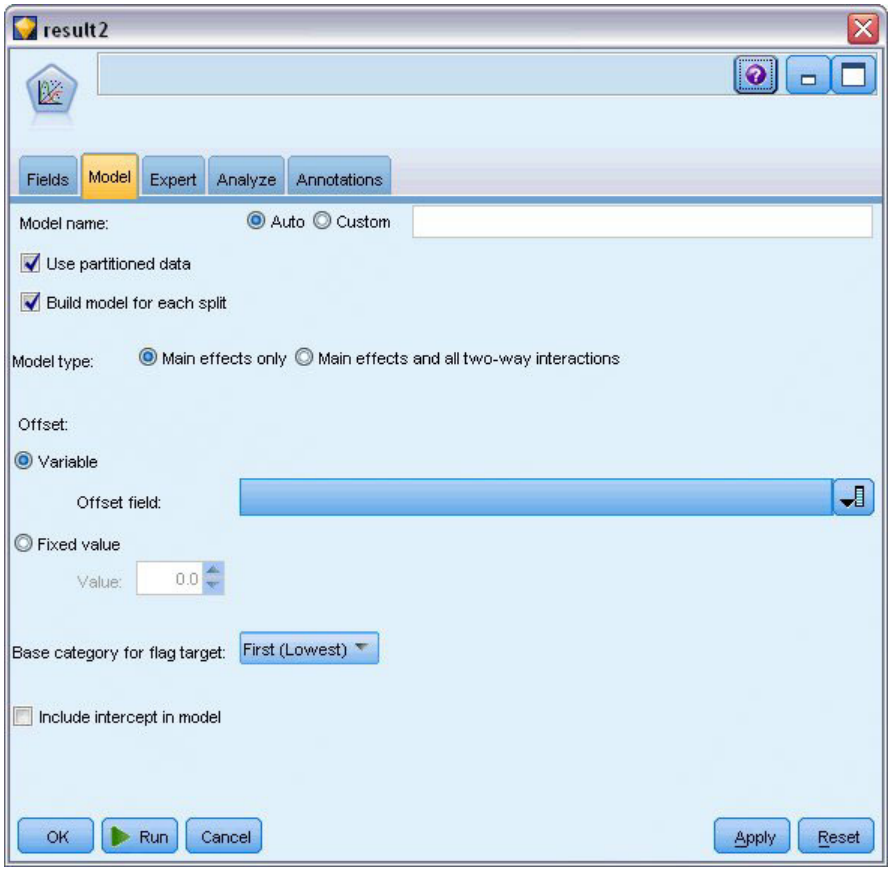


图 303. 选择模型选项

6. 选择 **第一个（最低值）** 作为目标的参考类别。此操作表示第二个类别是所关注的事件，它对模型的影响在参数估计中进行解释。
7. 取消选择 **Include intercept in model**。
8. 单击 **专家** 选项卡并选择 **专家** 以激活专家建模选项。

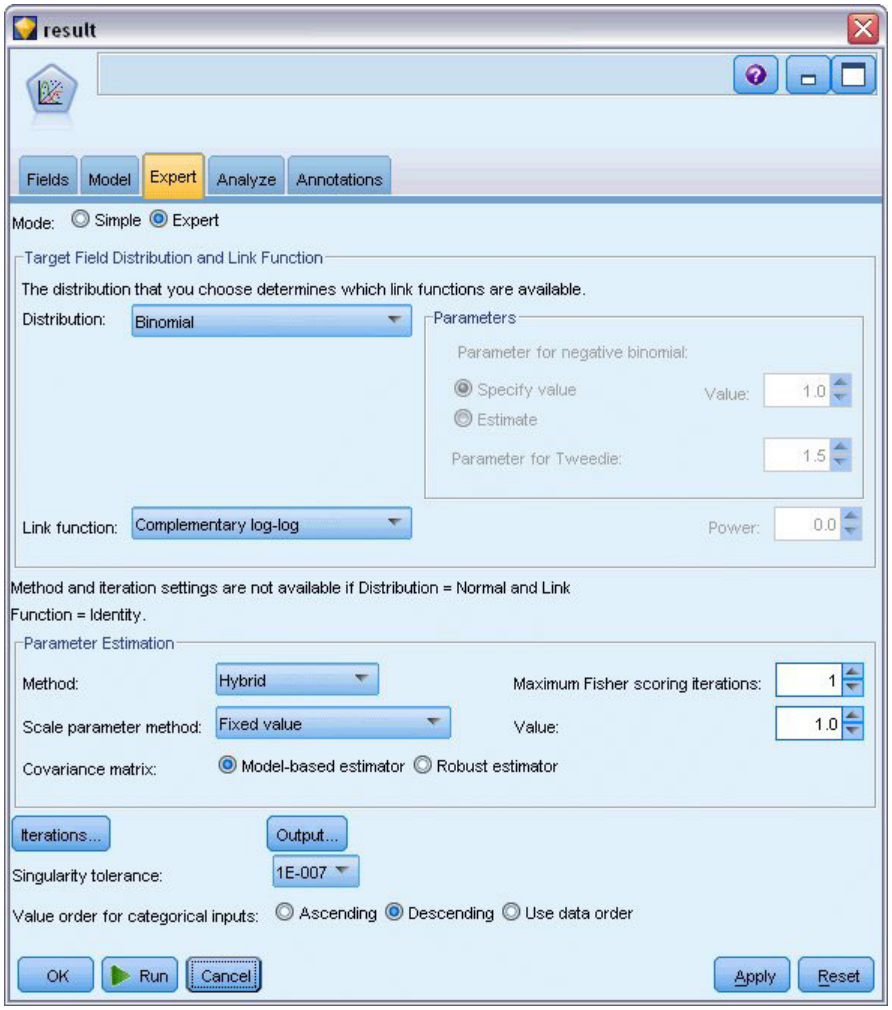


图 304. 选择专家选项

9. 选择 **二项** 作为分布， **互补重对数** 作为连接函数。
10. 选择 **固定值** 作为估计尺度参数的方法，并选择缺省值 1.0。
11. 选择**降序**作为因子的类别顺序。这指示每个因子的第一个类别将是其参考类别；模型中此项选择的效应由参数估计解释。
12. 运行流以创建模型块，此模型块将被添加到流工作区和位于右上角的“模型”选用板中。要查看模型详细信息，请右键单击此模型块并选择**编辑**或**浏览**。

模型效应检验

Source	Type III Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

Dependent Variable: Result by period
Model: period, duration, treatment, age

图 305. 主效应模型的模型效应检验

没有任何的模型效果是在统计意义下显著的；但是，任何可观察到的周期和治疗效果上的差异都具有临床意义，因此我们将仅为这些模型项拟合一个简化模型。

拟合简化模型

1. 在 GenLin 节点的“字段”选项卡上，单击**使用自定义设置**。
2. 选择 *结果 2* 作为目标。
3. 选择 *周期* 和 *治疗* 作为输入。

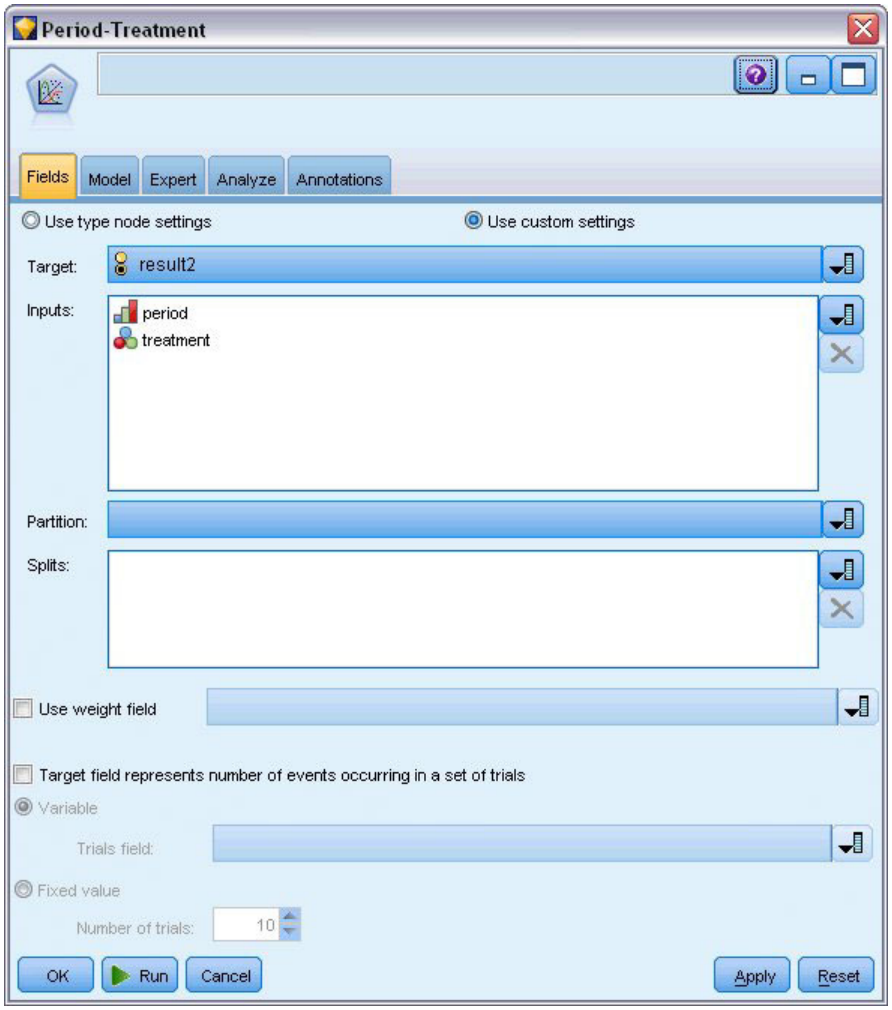


图 306. 选择字段选项

4. 运行节点并浏览生成的模型，然后将生成的模型复制到选用板，附加表节点后再次运行。

参数估计

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=1]	.195	.6279	-1.035	1.426	.097	1	.756
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: period, treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

图 307. 仅治疗模型的参数估计

治疗效果仍然不是统计意义下显著的，而是仅能估计出治疗 *A* 可能比 *B* 的效果好，因为治疗 *B* 的参数估计与前 12 个月内复发的概率增加相关联。周期值在统计意义下显著地不为 0，但这是因为截距项没有拟合的缘故。周期效果（ *[周期 =1]* 和 *[周期 =2]* 的线性预测变量的值之间的差异）不是统计意义下显著的，这一点可以在模型效应检验中看到。线性预测变量（周期效果 + 治疗效果）是 $\log(-\log(1-P(recur_{p,\ t}))$ 的估计值，其中 $P(recur_{p,\ t})$ 是在给定治疗（t = *A* 或 *B*）的周期（p = 1 或 2，表示 6 个月或 12 个月）内复发的概率。可为数据集中的每个观测数据生成这些预测的概率。

预测复发和生存的概率

图 308. “派生”节点设置选项

1. 对于每位患者，模型都可对预测结果和该预测结果的概率进行评分。为查看预测的复发概率，可将生成的模型复制到选用板并附加导出节点。
2. 在“设置”选项卡中，键入 `precur` 作为导出字段。
3. 选择将它作为**条件**导出。
4. 单击计算器按钮可打开 **If** 条件的表达式构建器。

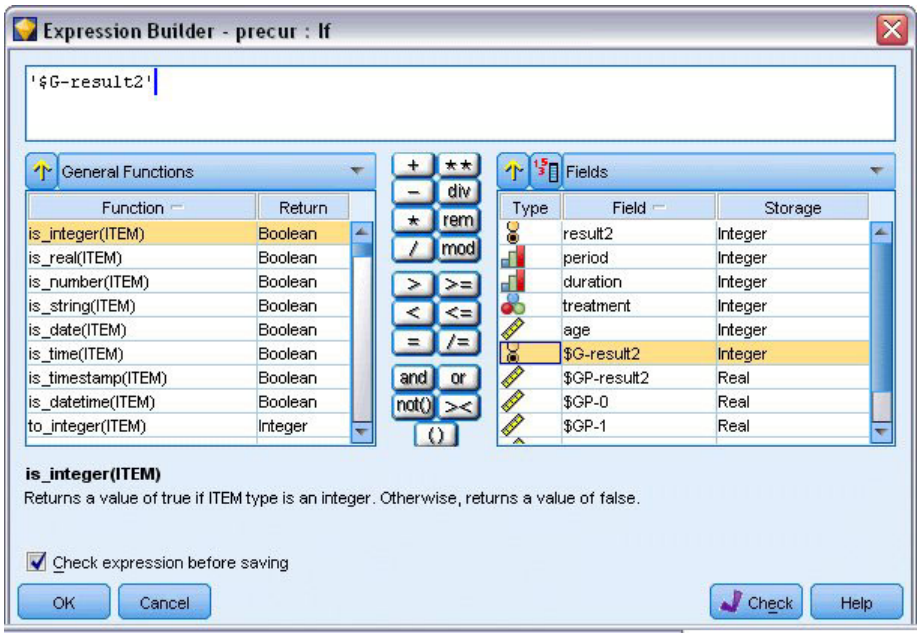


图 309. "派生"节点: If 条件的表达式构建器

5. 将 *$G-result2* 字段插入表达式中。
6. 单击**确定**。

导出字段 *precur* 在 *$G-result2* 等于 1 时和 0 时分别取 **Then** 表达式和 **Else** 表达式的值。

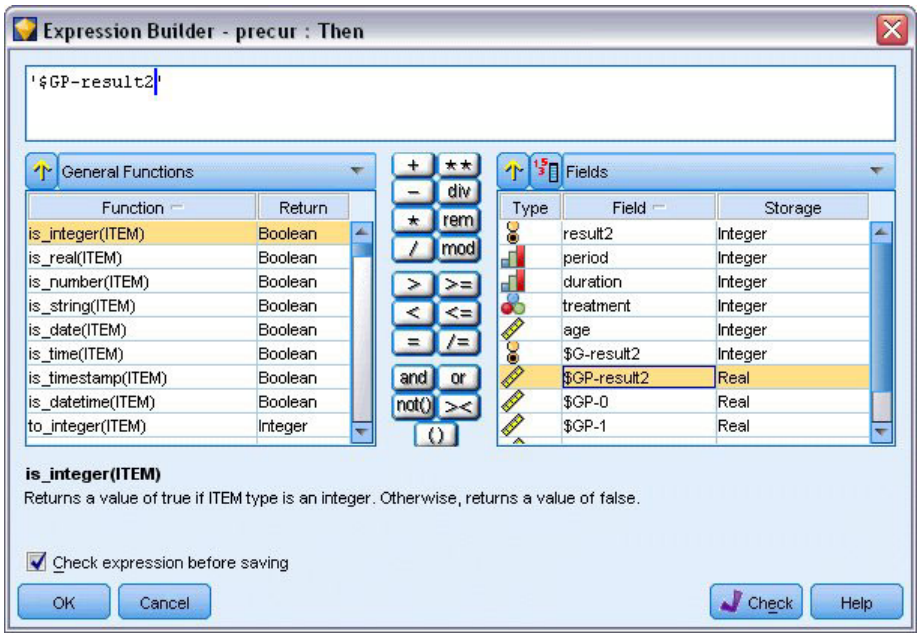


图 310. "派生"节点: Then 表达式的表达式构建器

7. 单击计算器按钮可打开 **Then** 表达式的表达式构建器。
8. 将 *$GP-result2* 字段插入表达式中。
9. 单击**确定**。

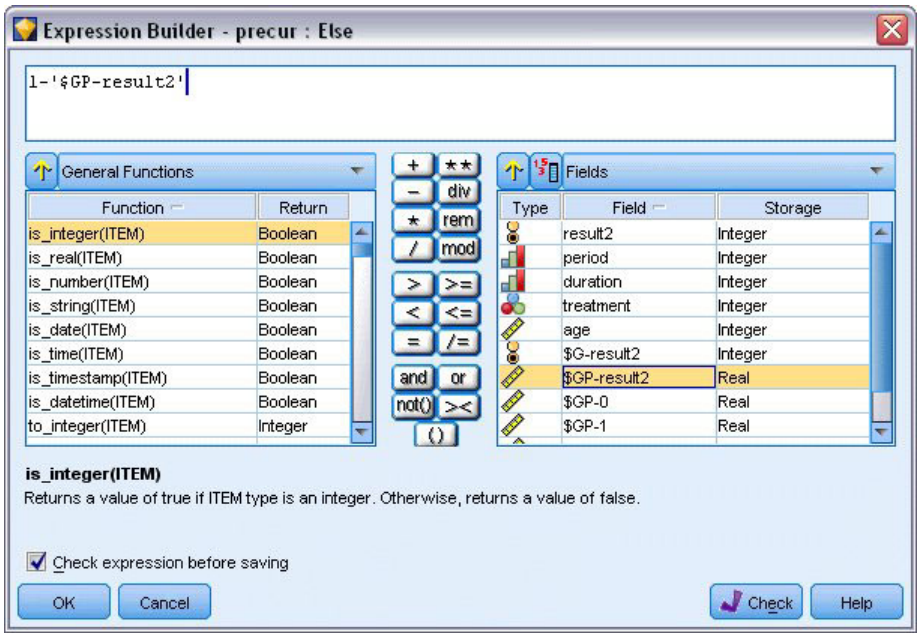
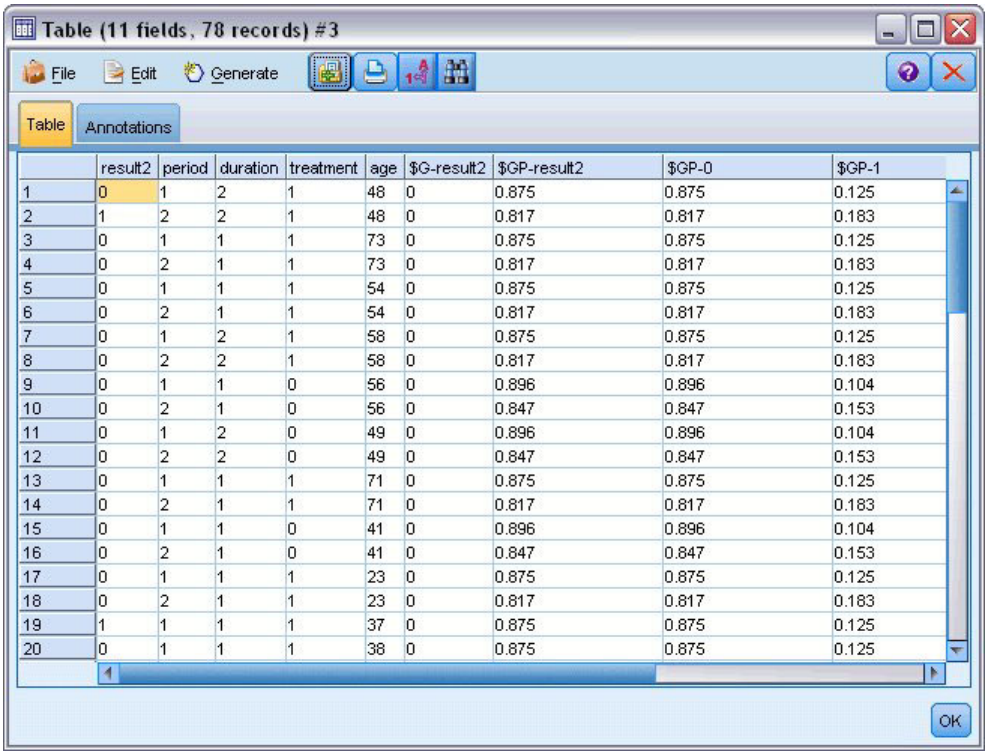


图 311. "派生"节点: Else 表达式的表达式构建器

10. 单击计算器按钮可打开 **Else** 表达式的表达式构建器。
11. 在表达式中输入 1-，然后将 *$GP-result2* 字段插入表达式。
12. 单击**确定**。

图 312. "派生"节点设置选项

13. 将表节点附加到"派生"节点并进行执行。



	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

图 313. 预测概率

表 3. 复发估计概率

治疗	6 个月	12 个月
A	0.104	0.153
B	0.125	0.183

根据复发估计概率，12 个月内的估计生存概率为 $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$；因此，对于每种治疗方案:

A: 1 -–(0.104 + 0.153*0.896) = 0.759

B: 1 -–(0.125 + 0.183*0.875) = 0.715

再一次显示出 *A* 作为更理想的治疗不是统计意义下显著的。

摘要

使用广义线性模型，已通过一系列互补重对数回归模型拟合了区间型删失的生存数据。虽然对于选择治疗 *A* 显示出一定的支持，但要取得统计意义下显著的结果还需要更大量的研究。不过，研究现有的数据还有一些其他方法。

- 值得一试的是使用模型重新拟合交互效应，尤其是 *周期* 和 *治疗组* 之间的交互效应。

IBM SPSS Modeler Algorithms Guide 中列出了对 IBM SPSS Modeler 中所使用的建模方法的数学原理的说明。

相关过程

广义线性模型过程是用于拟合各种模型的强大工具。

- 广义估计方程过程用于扩展广义线性模型以实现重复度量。
- 线性混合模型过程使您可以针对包含随机分量和/或重复度量的刻度因变量来拟合模型。

推荐读物

请参阅下列文本以了解有关广义线性模型的更多信息:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002, 2005. *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 使用泊松回归来分析船只损坏率（广义线性模型）

使用广义线性模型可以拟合泊松回归以便对计数数据进行分析。例如，在别处被提出和分析的[2]关于波浪对货船造成的损坏的数据集。如果有预测变量的值，便可将事件计数的模型建为以泊松比率发生，而且结果模型可以帮助您确定哪种类型的船最容易损坏。

本示例使用流 *ships_genlin.str*，该流引用了数据文件 *ships.sav*。数据文件和流文件分别位于 *Demos* 文件夹和 *streams* 子文件夹中。

由于*分类汇总服务月数*会随船只类型而变化，因此，在这种情况下为原始单元格计数建模会使人产生误解。这种测量承受风险程度的变量将在广义线性模型中作为偏移变量处理。此外，泊松回归假设因变量的对数在预测变量中为线性。因此，要使用广义线性模型来将泊松回归拟合到事故率，您需要使用 *Logarithm of aggregate months of service*。

拟合“高度离散的”泊松回归

1. 在 *Demos* 文件夹中添加指向 *ships.sav* 的“Statistics 文件”源节点。

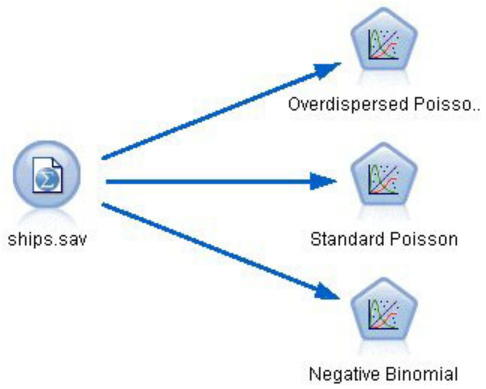


图 314. 用于分析损坏率的样本流

2. 在源节点的“过滤”选项卡上，排除字段 *months_service*。该变量的经对数转换的值包含在 *log_months_service* 中，这些值将在分析中使用。

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

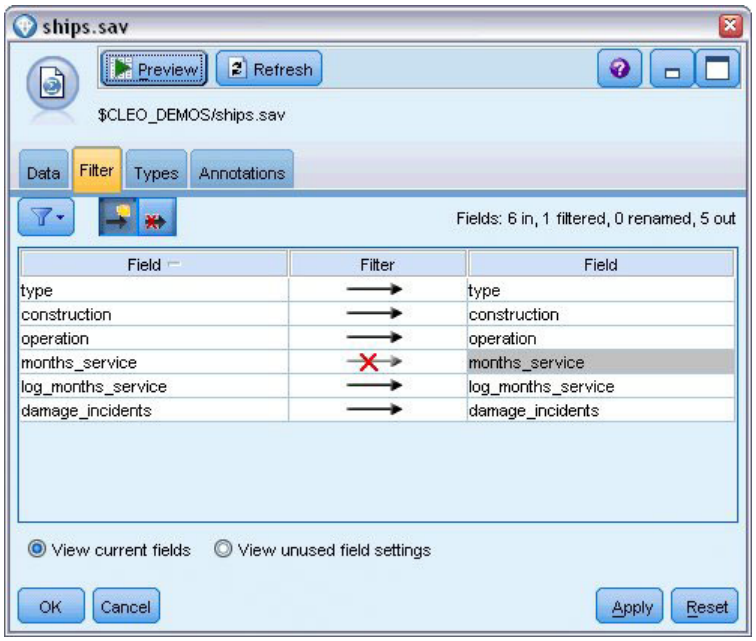


图 315. 过滤不需要的字段

（或者，也可以将“类型”选项卡上该字段的角色改为**无**而不是排除该字段，或选择您要在模型节点中使用的字段。）

3. 在源节点的类型选项卡中，将 *damage_incidents* 字段的角色设置为 **Target**。将所有其他字段的角色设置为 **Input**。
4. 单击**读取值**以实例化数据。

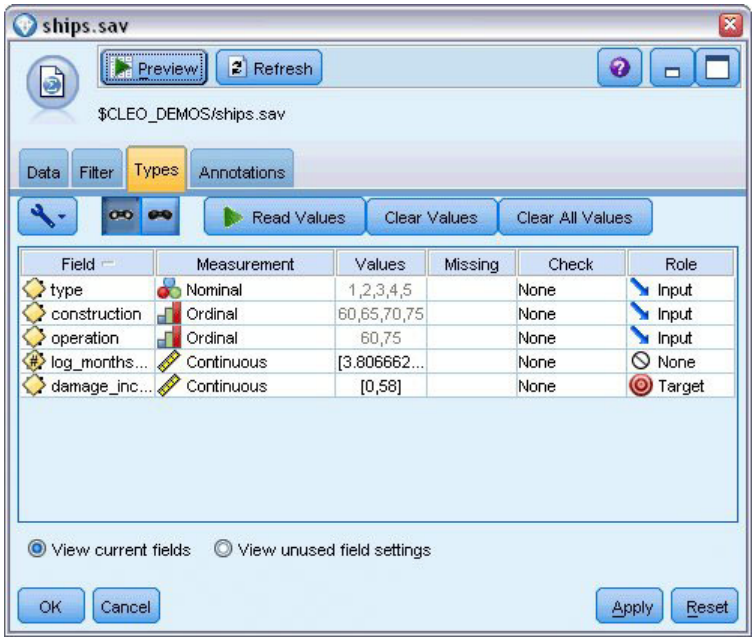


图 316. 设置字段角色

5. 将 GenLin 节点附加到源节点；在 GenLin 节点中，单击 **模型** 选项卡。

6. 选择 *log_months_service* 作为偏移变量。

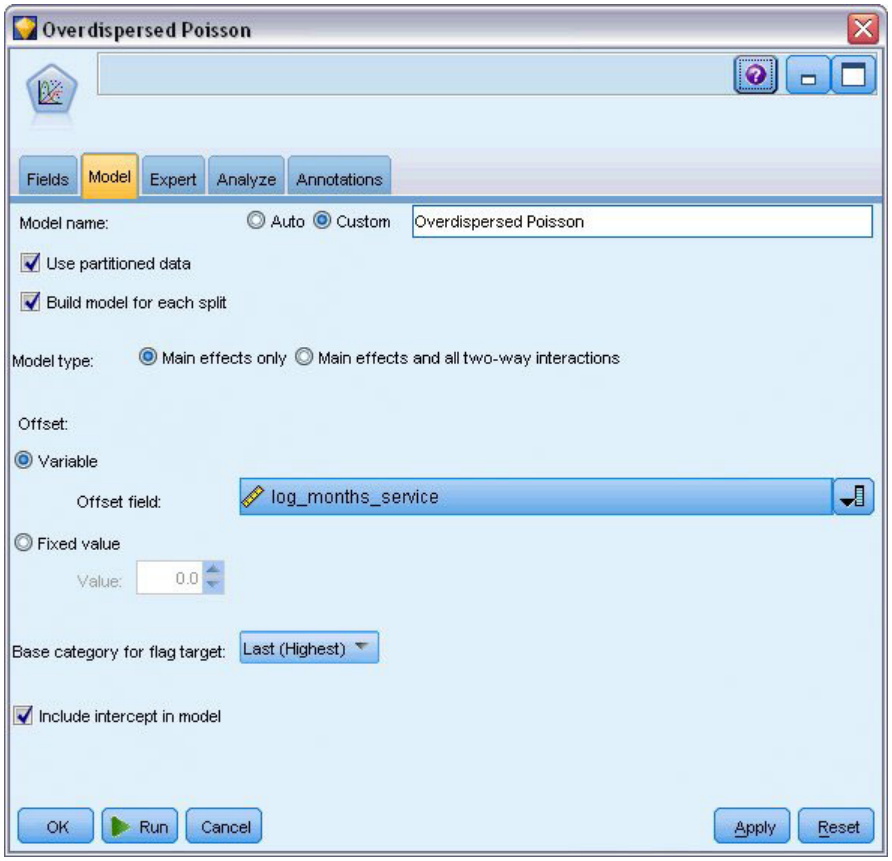


图 317. 选择模型选项

7. 单击 **专家** 选项卡并选择 **专家** 以激活专家建模选项。

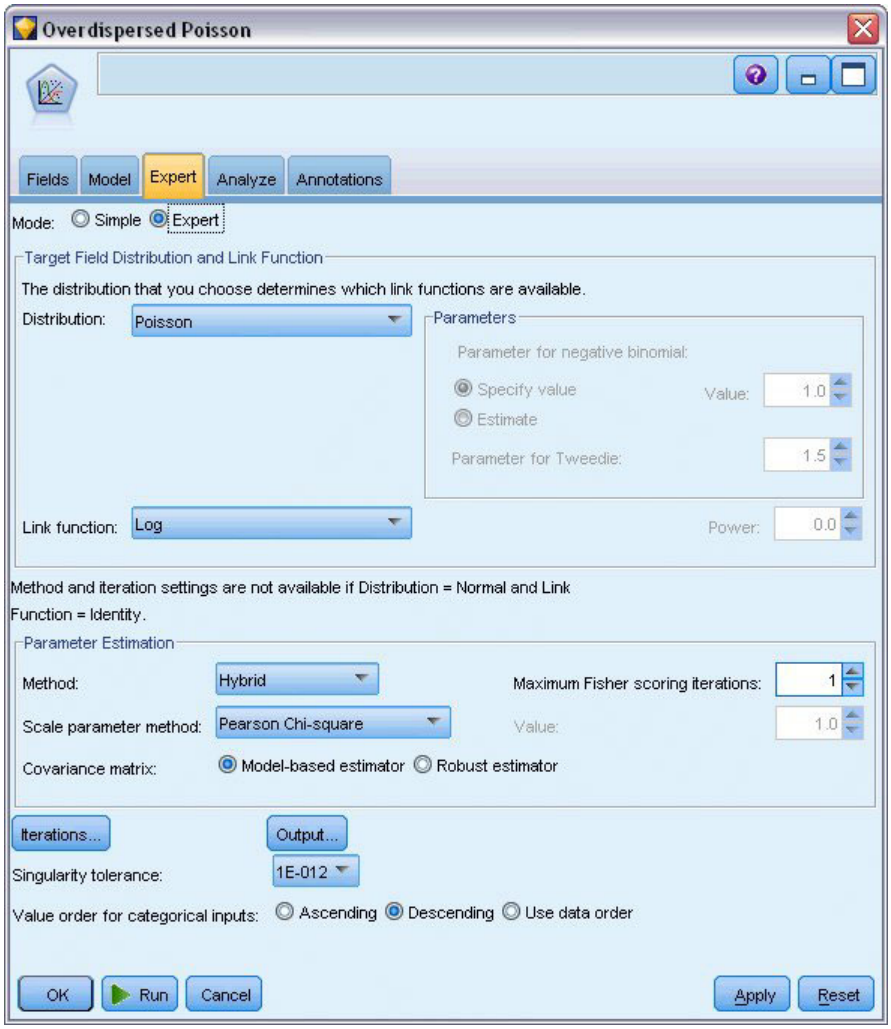


图 318. 选择专家选项

8. 选择**泊松**作为响应的分布，并选择**对数**作为关联函数。
9. 选择 **Pearson 卡方**作为估计尺度参数的方法。尺度参数在泊松回归中通常假设为 1，但 McCullagh 和 Nelder 却用 Pearson 卡方估计来获得更保守的方差估计值和显著性水平。
10. 选择**降序**作为因子的类别顺序。这指示每个因子的第一个类别将是其参考类别；模型中此项选择的效应由参数估计解释。
11. 单击**运行**以创建模型块，此模型块将被添加到流工作区和位于右上角的“模型”选用板中。要查看模型详细信息，请右键单击此模型块并选择**编辑**或**浏览**，然后单击**高级**选项卡。

拟合度统计

	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.

图 319. 拟合度统计

拟合度统计表提供了对于比较竞争模型很有用的度量。此外，偏差和 Pearson 卡方统计量的 *Value/df* 值给出了对尺度参数的相应估计。泊松回归中的这些值应该接近 1.0；这些值大于 1.0 的事实表明拟合高度离散的模型也许是合理的。

Omnibus 检验

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

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

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

图 320. Omnibus 检验

Omnibus 检验是当前模型与零（此个案中为截距）模型的似然比卡方检验。小于 0.05 的显著性值表明当前模型的性能要高于零模型的性能。

模型效应检验

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

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

图 321. 模型效应检验

检验模型中的每一项是否具有效应。显著性值小于 0.05 的项具有一定可辨别效应。每个主效应项都对模型有贡献。

参数估计

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
[type=5]	.326	.3067	-.276	.927	1.127	1	.288
[type=4]	-.076	.3779	-.817	.665	.040	1	.841
[type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[type=1]	0[a]	.	.	.	.	.	.
[construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[construction=60]	0[a]	.	.	.	.	.	.
[operation=75]	.384	.1538	.083	.686	6.249	1	.012
[operation=60]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

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

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

图 322. 参数估计

参数估计表总结了每个预测变量的影响。关联函数的性质决定了此模型的系数难于解释，但是，通过协变量系数的符号和因子水平系数的相对值，还是可以获得模型预测变量效应的重要信息。

- 对于协变量来说，正（负）系数表示预测变量与结果成正向（反向）关系。系数为正的协变量值的增加对应于损坏事件比率的增加。
- 对于因子，系数越大的因子水平表示损坏的发生率越高。因子水平系数的符号取决于因子水平相对于参考类别的影响作用。

基于参数估计可以做出如下解释:

- 船只类型 *B [类型 =2]* 的损伤率（估计系数为 –0.543）在统计意义下显著（ *p* 值为 0.019）低于类型 *A [类型 =1]*（参考类别）。类型 *C [type=3]* 的估计参数实际要低于 *B*，但是 *C* 的估计值中的变异性遮盖了效应。有关因子水平间所有关系的解释，请参阅估计边际均值。

- 根据统计，1965–69 *[construction=65]* 和 1970–74 *[construction=70]* 年间建造的船只的损坏率显著（*p* 值 < 0.001）高于（估计系数分别为 0.697 和 0.818）在 1960–64 *[construction=60]* 年间建造的船只（参考类别）。有关因子水平间所有关系的解释，请参阅估计边际均值。
- 航运时间为 1975–79 *[航运时间 =75]* 的船只的损伤率（估计系数为 0.384）在统计意义下显著（ *p* 值为 0.012）高于航运时间为 1960–1974 *[航运时间 =60]* 的船只。

拟合其他模型

“高度离散的”泊松回归存在一个问题，即没有正式的方式来检验它与“标准”泊松回归。但是，建议的用来确定是否具有高度离散的正式检验是在所有其他设置相同的情况下执行“标准”泊松回归和负二项式回归之间的似然比检验。如果泊松回归中未出现过度离散，那么统计量 - 2×（泊松模型的对数似然 - 负二项式模型的对数似然）应该具有混合分布（其一半的概率质量位于 0 处，而其余的概率质量位于卡方分布中且自由度为 1）。

1. 选择**固定值**作为估计尺度参数的方法。缺省情况下，该值为 1。

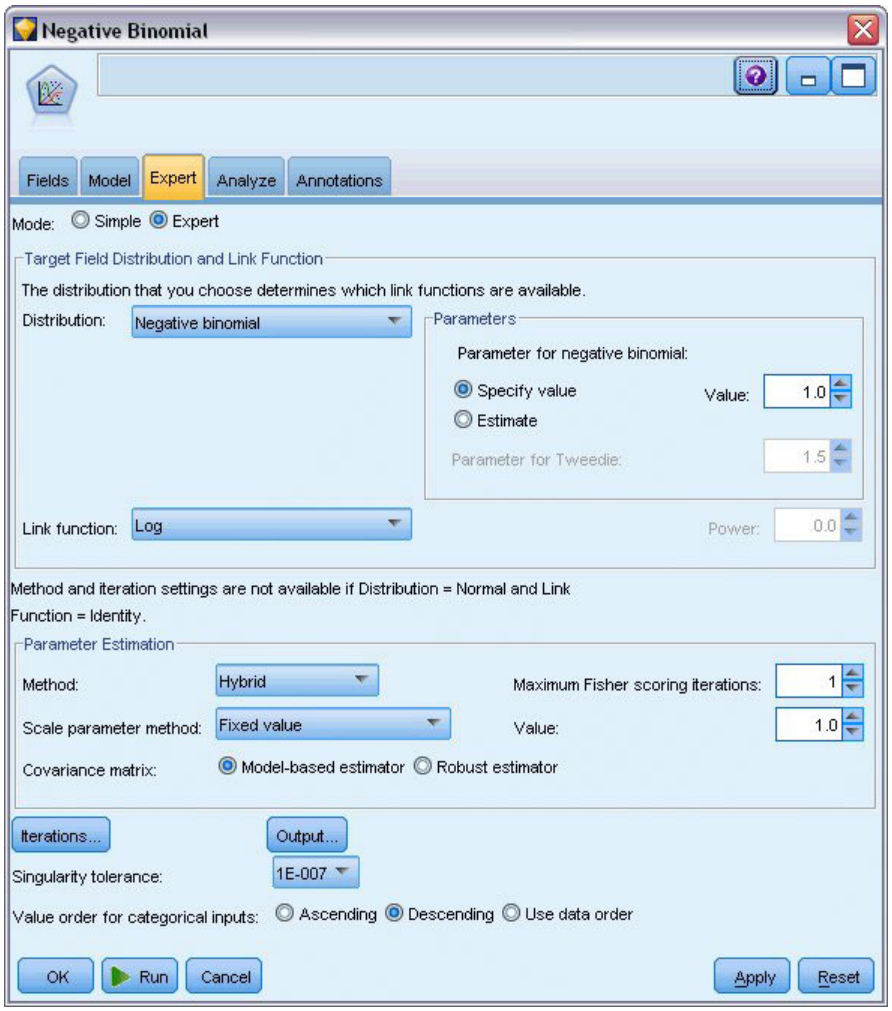


图 323. “专家”选项卡

2. 要拟合负二项回归，可复制并粘贴 GenLin 节点，将其附加到源节点，然后打开新节点，并单击**专家**选项卡。
3. 选择**负二项式**作为分布。保留辅助参数的缺省值为 1。
4. 运行流，并在新创建的模型块上浏览“高级”选项卡。

拟合度统计

	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.

图 324. 标准泊松回归的拟合度统计

标准泊松回归报告的对数似然为 –68.281。将该值与负二项式模型进行比较。

	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.

图 325. 负二项式回归的拟合度统计

负二项回归报告的对数似然为 –83.725。而实际上负二项式回归的对数似然要小于泊松回归的对数似然，这就表示该负二项式回归没有提供优于泊松回归的改进。

但是，选择负二项式分布的辅助参数为 1 也许不是非常适合该数据集。检验高度离散的另一种方法是以辅助参数为 0 来拟合负二项式模型，并请求专家选项卡的输出对话框上的拉格朗日乘数检验。如果该检验不显著，则过散布对于该数据集不是问题。

摘要

通过使用“广义线性模型”，您为计数数据拟合了三个不同的模型。显示了负二项式回归没有提供任何优于泊松回归的改进。高度离散的泊松回归似乎可以合理地替代标准泊松模型，但对于两者间的选择还没有正式的检验。

IBM SPSS Modeler Algorithms Guide 中列出了对 IBM SPSS Modeler 中所使用的建模方法的数学原理的说明。

相关过程

广义线性模型过程是用于拟合各种模型的强大工具。

- 广义估计方程过程用于扩展广义线性模型以实现重复度量。
- 线性混合模型过程使您可以针对包含随机分量和/或重复度量的刻度因变量来拟合模型。

推荐读物

请参阅下列文本以了解有关广义线性模型的更多信息:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002, 2005. *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 将 Gamma 回归拟合至汽车保险理赔（广义线性模型）

广义线性模型可以被应用于拟合正范围数据分析的 Gamma 回归。例如，在其他地方出现和分析的数据集 [3] 与汽车的损伤理赔有关。平均理赔金额可以当作其具有伽玛分布来建模，通过使用逆联接函数将因变量的均值与预测变量的线性组合关联。出于考虑到用于计算平均理赔金额的理赔数目不同，指定*理赔数*作为尺度权重。

本示例使用命名为 *car-insurance_genlin.str* 的流，它引用命名为 *car_insurance_claims.sav* 的数据文件。数据文件和流文件分别位于 *Demos* 文件夹和 *streams* 子文件夹中。

创建流

1. 在 *Demos* 文件夹中添加指向 *car_insurance_claims.sav* 的“Statistics 文件”源节点。

图 326. 用于预测汽车保险索赔的样本流

2. 在源节点的“类型”选项卡中，将 *claimamt* 字段的角色设置为 **Target**。将所有其他字段的角色设置为 **Input**。
3. 单击**读取值**以实例化数据。

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

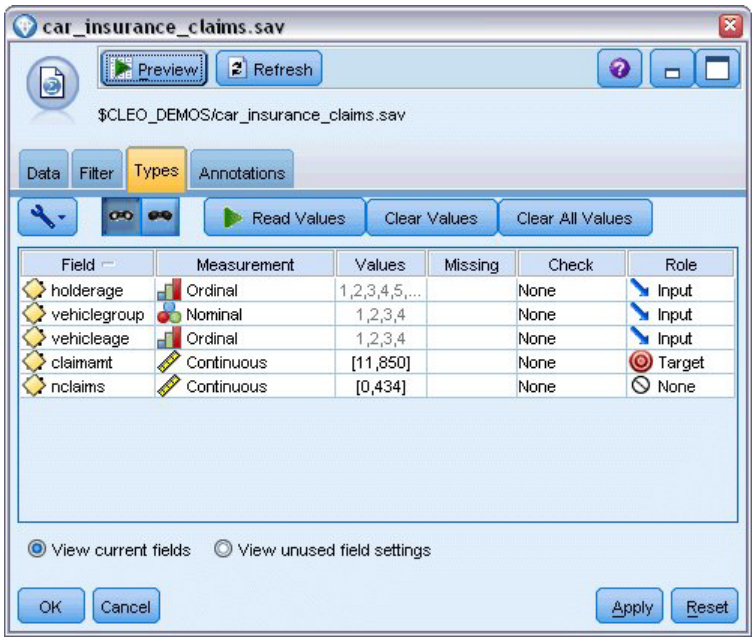


图 327. 设置字段角色

4. 将 GenLin 节点附加到源节点；在 GenLin 节点中，单击“字段”选项卡。
5. 选择 *nclaims* 作为尺度权重字段。

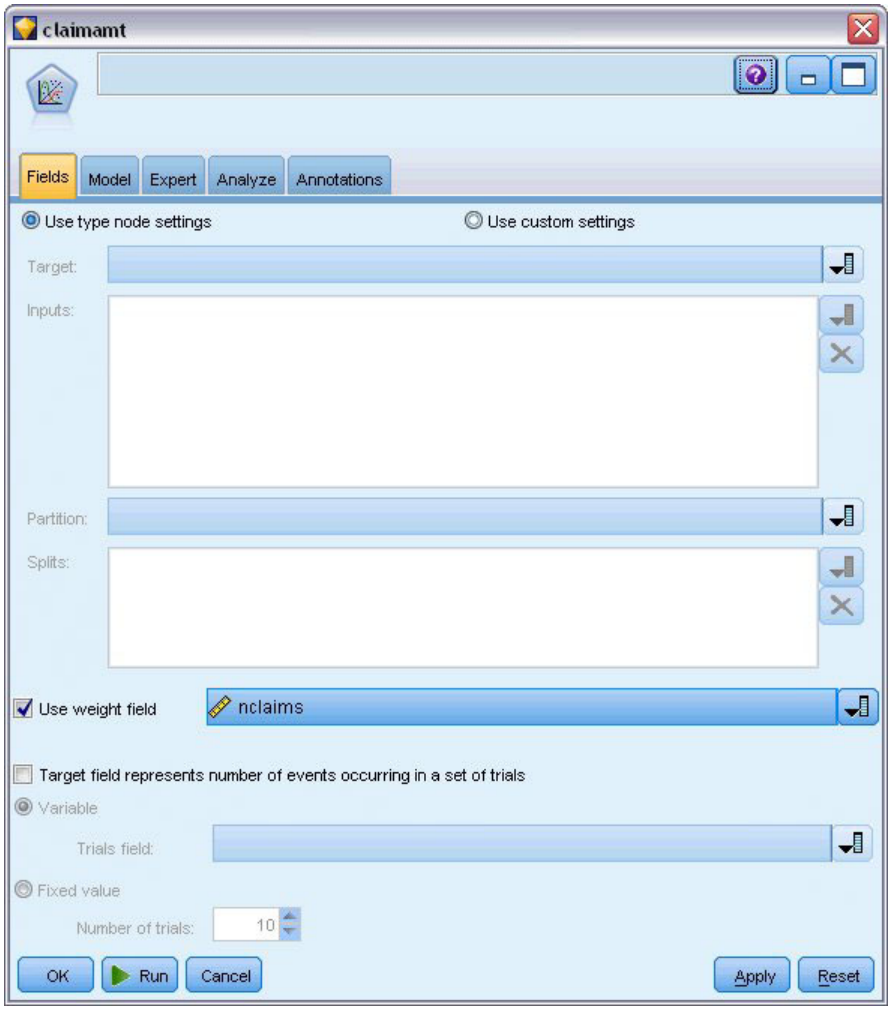


图 328. 选择字段选项

6. 单击专家选项卡并选择**专家**以激活专家建模器选项。

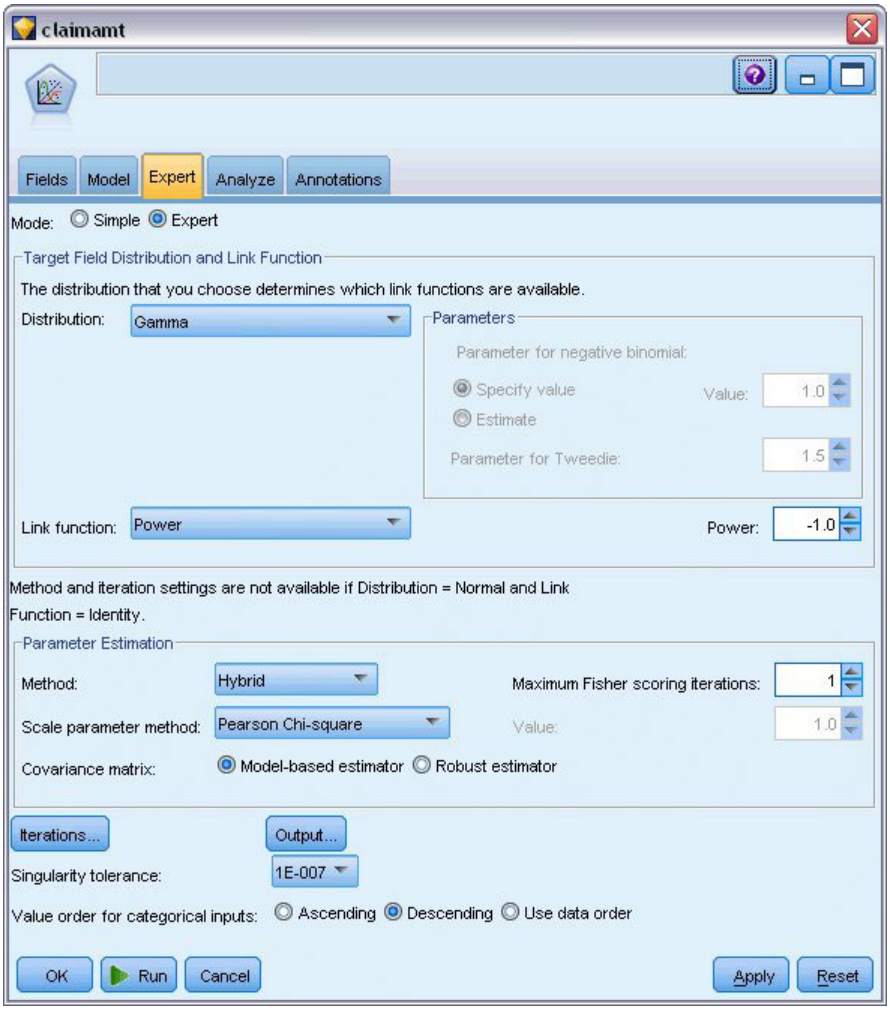


图 329. 选择专家选项

7. 选择 **伽玛** 作为响应分布。
8. 选择**幂**作为关联函数，并输入 `-1.0` 作为幂函数的指数。这是一个逆联接。
9. 选择 **Pearson 卡方**作为估计尺度参数的方法。这是 McCullagh 和 Nelder 应用的方法，因此我们在此沿用它来精确重现其结果。
10. 选择**降序**作为因子的类别顺序。这指示每个因子的第一个类别将是其参考类别；模型中此项选择的效应由参数估计解释。
11. 单击**运行**以创建模型块，此模型块将被添加到流工作区和位于右上角的“模型”选用板中。要查看模型详细信息，请右键单击此模型块并选择**编辑**或**浏览**，然后选择“高级”选项卡。

参数估计

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000[a]	.	.	.	.	.	.
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000[a]	.	.	.	.	.	.
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000[a]	.	.	.	.	.	.
(Scale)	1.209[b]			.001	.0004	.000	.002

Dependent Variable: Average cost of claims
Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

图 330. 参数估计

Omnibus 检验和模型效应检验（未显示）表明，这个模型的性能要高于空模型的性能而且每个主效应项都对此模型起作用。参数估计表显示与 McCullagh 和 Nelder 获得的因子水平相同的值和刻度参数。

摘要

通过使用广义线性模型，可以将伽玛回归拟合至理赔数据。请注意：伽玛分布的典型关联函数被应用于此模型中的同时，对数链接也将给出合理结果。通常，很难甚至根本就不可能直接将模型与不同的关联函数进行比较；不过，对数链接是一个特例，在对数链接中，指数为零，因此，您可以将一个模型中的偏差与一个对数链接以及一个具有幂关联模型进行比较，以确定哪一个能更好的拟合（例如参阅 McCullagh 和 Neldersection 的第 11.3 部分）。

IBM SPSS Modeler Algorithms Guide 中列出了对 IBM SPSS Modeler 中所使用的建模方法的数学原理的说明。

相关过程

广义线性模型过程是用于拟合各种模型的强大工具。

- 广义估计方程过程用于扩展广义线性模型以实现重复度量。
- 线性混合模型过程使您可以针对包含随机分量和/或重复度量的刻度因变量来拟合模型。

推荐读物

请参阅下列文本以了解有关广义线性模型的更多信息:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002, 2005. *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 细胞样本分类（SVM）

支持矢量机 (SVM) 是一项特别适合于广泛数据集的分类和回归技术。广泛数据集包含大量预测变量，例如可能会在生物信息学领域遇到（对生物化学数据和生物学数据应用信息技术）的预测变量。

一位医学研究人员获得了一个包含大量人体细胞样本的特征的数据集，这些样本是从被认为可能会患上癌症的患者身上提取的。对原始数据的分析表明，良性样本与恶性样本之间的很多特征显著不同。该研究人员希望开发一个 SVM 模型，使该模型可以使用其他患者样本中的这些细胞特征值尽早发现他们的样本是良性还是恶性。

本示例使用了名为 *svm_cancer.str* 的流，该流位于 *Demos* 文件夹下的 *streams* 子文件夹中。数据文件为 *cell_samples.data* 。请参阅主题第 4 页的『Demos 文件夹』以获取更多信息。

本示例基于可以从 UCI Machine Learning Repository 公开获取的数据集。数据集由数百条人体细胞样本记录组成，每条记录都包含一组细胞特征的值。每条记录中包含的字段包括:

字段名称	描述
标识	患者标识
Clump	肿块的厚度
UnifSize	细胞大小的均匀度
UnifShape	细胞大小的均匀度
MargAdh	边际粘连
SingEpiSize	单层上皮细胞的大小
BareNuc	裸核
BlandChrom	温和的染色质
NormNucl	正常的核仁
Mit	有丝分裂
Class	良性或恶性

为达到本示例的目的，我们使用的是每条记录包含相对较少预测变量的数据集。

创建流

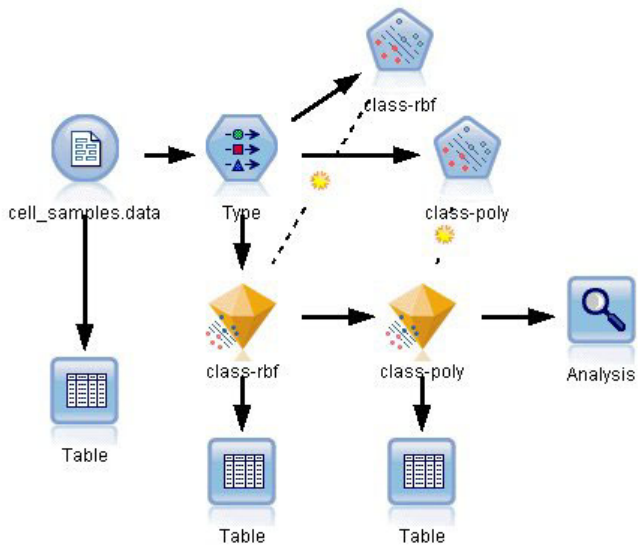
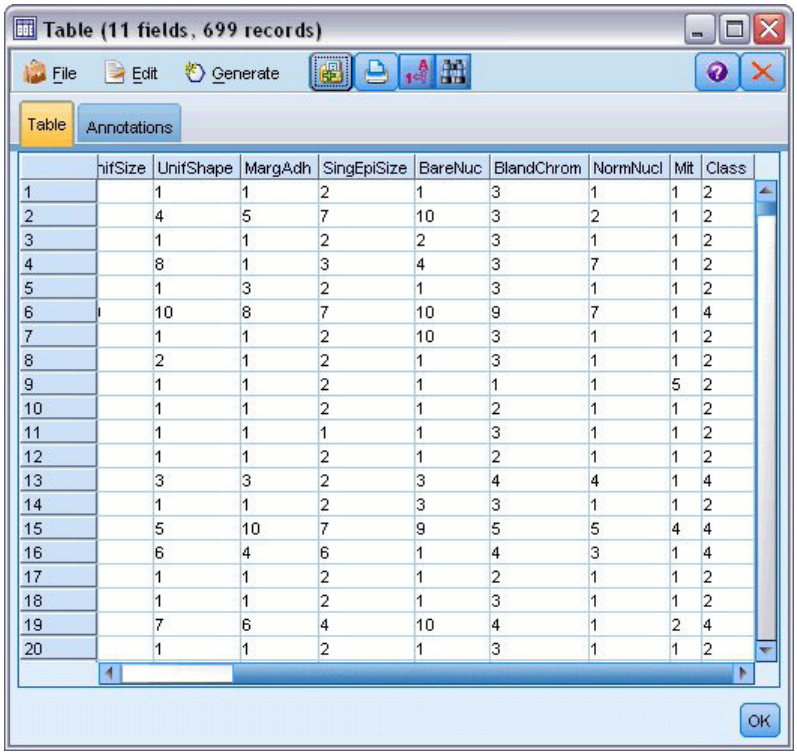


图 331. 用于显示 SVM 建模的样本流

1. 创建一个新流，然后在 IBM SPSS Modeler安装程序的 *Demos* 文件夹中添加一个指向 *cell_samples.data* 的变量文件源节点。

   让我们看一下源文件中的数据。
2. 为流添加表节点。
3. 将“表”节点附加到“变量文件”节点并运行流。



	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

图 332. SVM 的源数据

标识字段包含患者的标识。来自每位患者的细胞样本特征包含在从 *Clump* 到 *Mit* 的字段中。这些字段的值按照 1 到 10 进行分级，1 表示最接近于良性。

Class 字段包含诊断，由多步独立的医疗程序确认，用于表明样本是良性（值 = 2）还是恶性（值 = 4）。

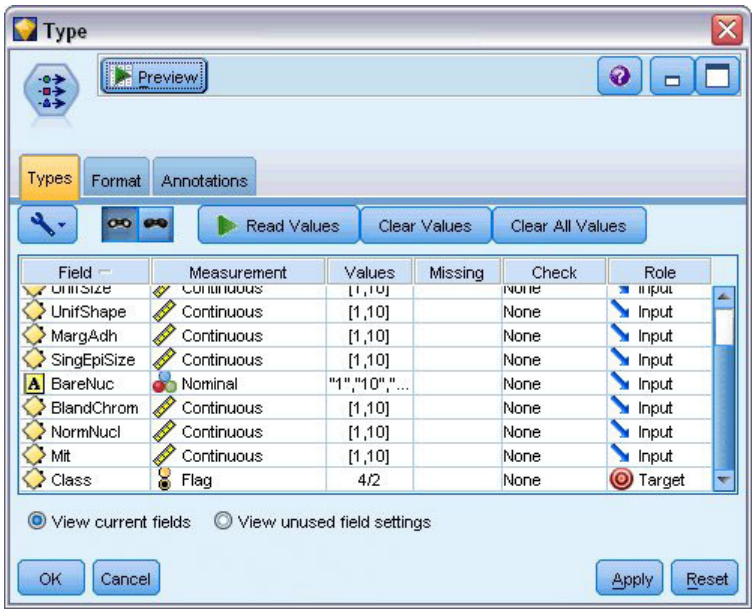


图 333. "类型"节点设置

4. 添加一个类型节点并将它附加到变量文件节点。

5. 打开此“类型”节点。

   我们希望模型预测 *Class* 的值（即，良性 (=2) 还是恶性 (=4)）。由于此字段可以为仅有的这两个可能值之一，我们需要更改其测量级别以反映这一情况。

6. 在类字段的**测量**列（列表中最后一个），单击值**连续**并将其更改为**标志**。
7. 单击 **读取值** 。
8. 在**角色**列中，将*标识*（患者的标识）的角色设置为**无**，因为此字段将不会用作预测变量或模型的目标。
9. 将目标 *Class* 的角色设置为**目标**，并将所有其他字段（预测变量）的角色保留为**输入**。
10. 单击**确定**。

    SVM 节点提供多个用于执行其处理的内核函数供您选择。由于难以确定哪个函数对于任何给定数据集都能产生最佳效果，因此我们将依次选择不同的函数并对结果进行比较。我们从缺省函数开始，即 RBF（径向基函数）。

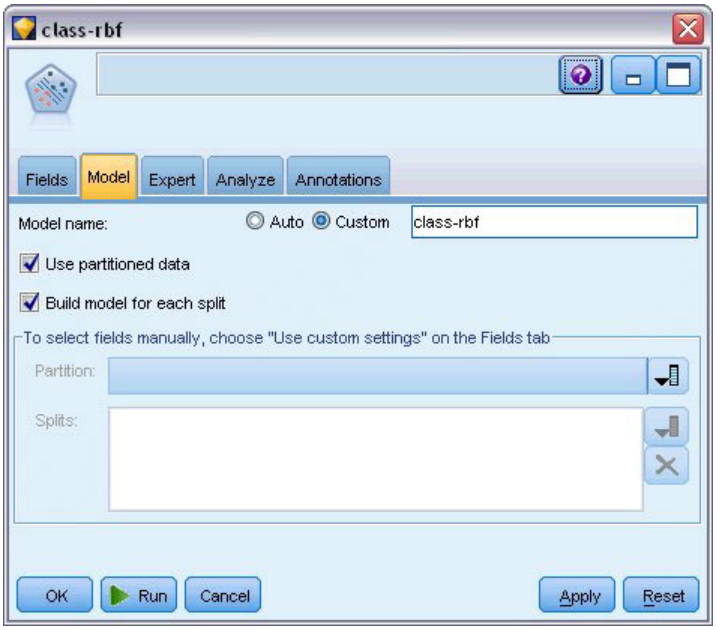


图 334. “模型”选项卡设置

11. 在“建模”选用板中，将 SVM 节点附加到类型节点。
12. 打开此 SVM 节点。在 **模型** 选项卡中，单击 **模型名称** 的 **定制** 选项，然后在相邻的文本字段中键入 *class-rbf* 。

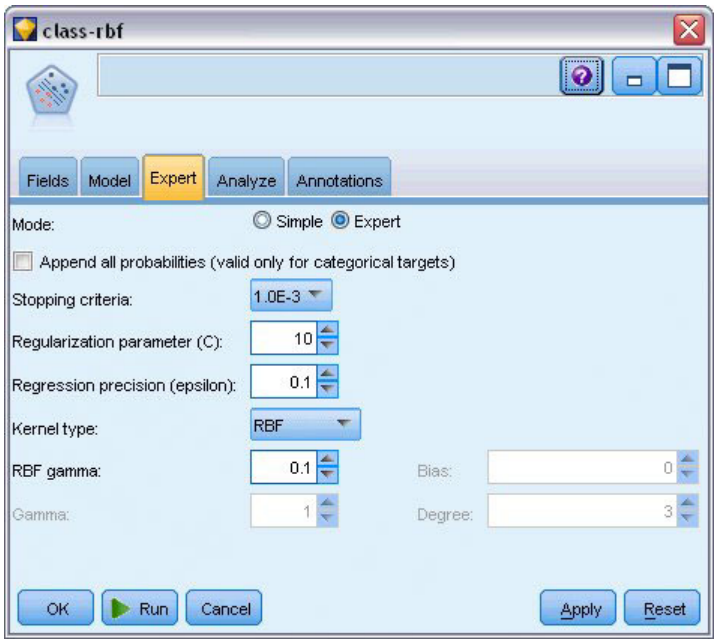


图 335. 缺省“专家”选项卡设置

13. 在**专家**选项卡上，将**模式**设为**专家**以获得可靠性，但保持所有缺省选项不变。注意，**Kernel 类型**缺省设为 **RBF**。在“简单”方式下，所有选项都显示为灰色。

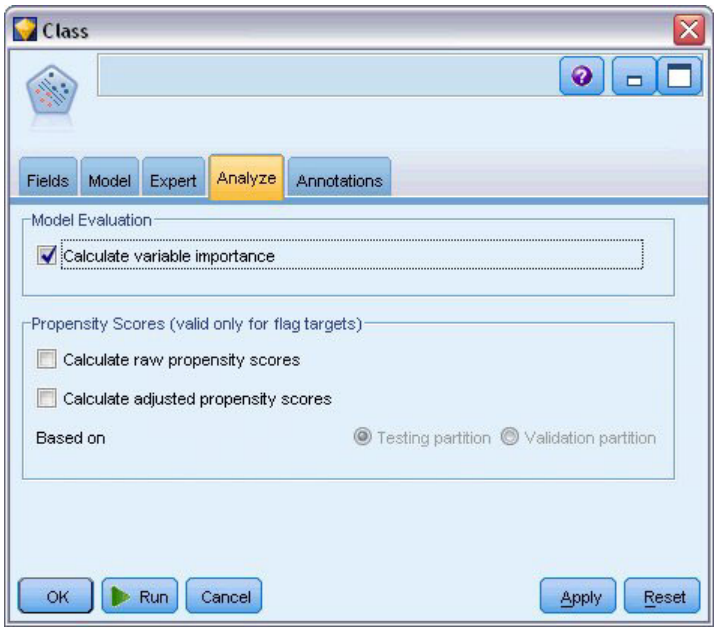


图 336. “分析”选项卡设置

14. 在 **分析** 选项卡上，选中 **计算变量重要性** 复选框。
15. 单击**运行**。模型块将位于流和屏幕右上角的“模型”选用板中。
16. 双击流中的模型块。

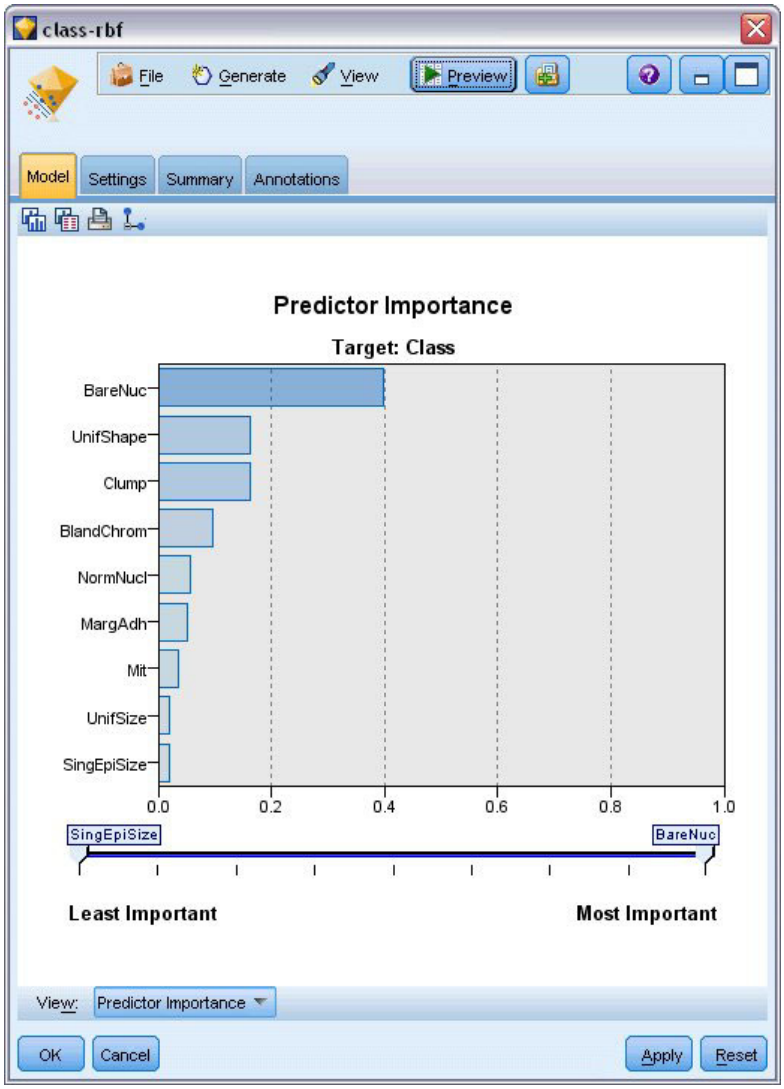
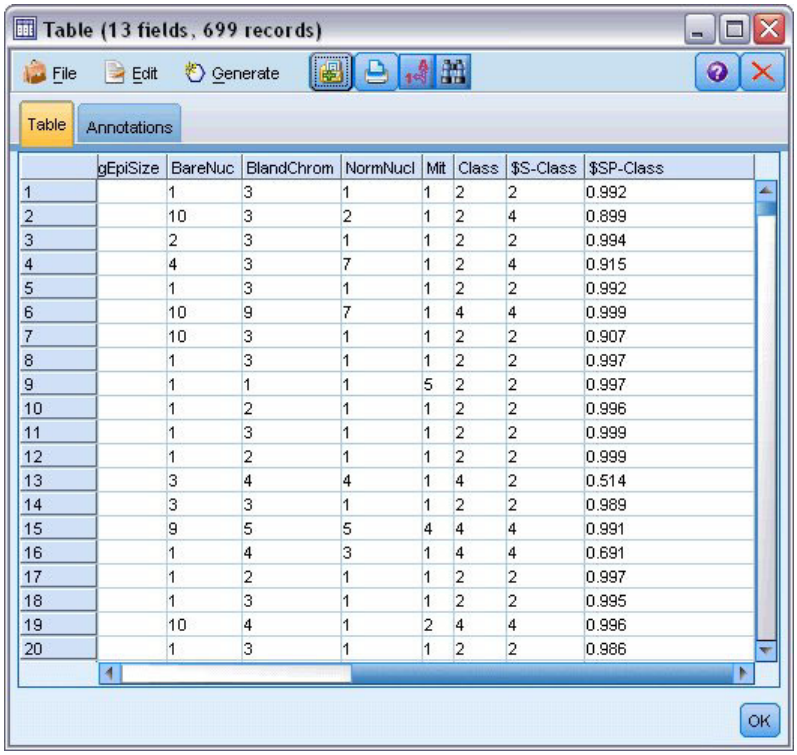


图 337. "预测变量重要性"图形

在"模型"选项卡上，"预测变量重要性"图形显示了不同字段对预测的相对影响。此图向我们显示了 *BareNuc* 无疑具有最大的影响，而 *UnifShape* 和 *Clump* 的影响也很大。

1. 单击**确定**。
2. 将表节点附加到 *class-rbf* 模型块。
3. 打开"表"节点，然后单击**运行**。



	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

图 338. 为预测和置信度值添加的字段

4. 此模型创建了两个额外的字段。向右滚动表输出可看到这两个字段:

新字段名	描述
$S-Class	由模型预测的 *Class* 值。
$SP-Class	此预测值的倾向评分（即此预测值为真的似然，其值介于 0.0 到 1.0 之间）。

只需查看上表，我们就可以看到大多数记录的倾向评分（在 *$SP-Class* 列）都相当高。

但是，也有一些明显的例外情况；例如，对于患者 1041801 的记录（位于第 13 行），其倾向评分为无法接受的低分 0.514。同时，通过比较 *Class* 和 *$S-Class*，可以清楚地看到此模型作出了许多不正确的预测，即使是倾向评分相对高的地方也是如此（例如，第 2 行和第 4 行）。

让我们看一下是否可以通过选择另一种函数类型获得较好的效果。

尝试另一种函数

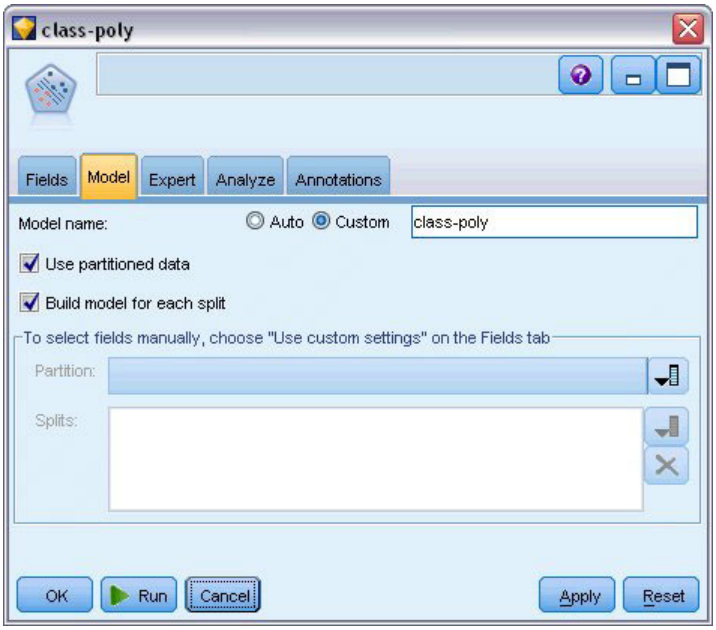


图 339. 为模型设置新名称

1. 关闭“表格”输出窗口。
2. 将第二个 SVM 建模节点附加到类型节点。
3. 打开新的 SVM 节点。
4. 在**模型**选项卡上，选择自定义和类型 *class-poly* 作为模型名称。

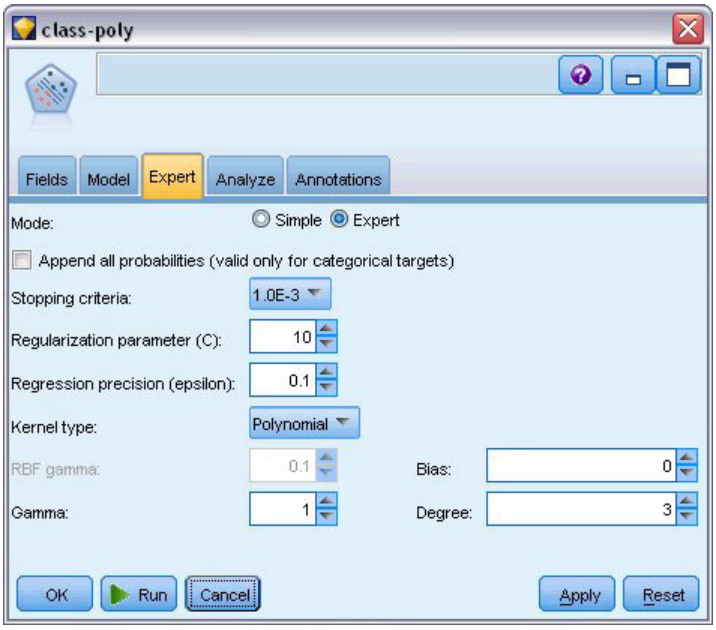


图 340. 多项式的“专家”选项卡设置

5. 在 **专家** 选项卡上，将 **模式** 设置为 **专家** 。

6. 将**核类型**设置为**多项式**并单击**运行**。*class-poly* 模型块被添加到流，同时还添加到屏幕右上角的“模型”选用板。
7. 将 *class-rbf* 模型块连接到 *class-poly* 模型块（在警告对话框上选择**替换**）。
8. 将表节点附加到 *class-poly* 模型块。
9. 打开“表”节点，然后单击**运行**。

比较结果

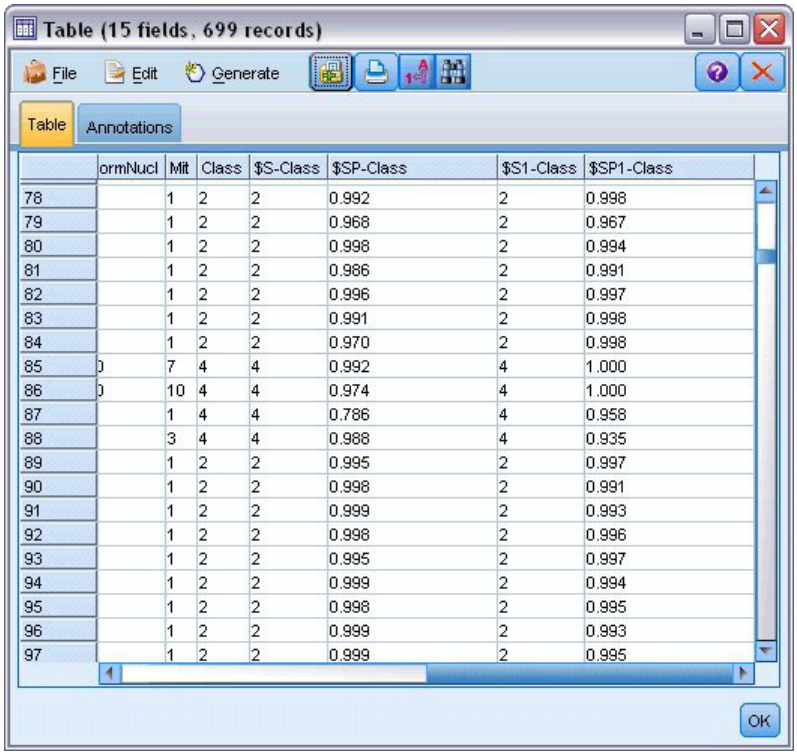


图 341. 为多项式函数添加的字段

1. 向右滚动表输出可看到新添加的字段。

   为多项式函数类型生成的字段分别名为 *$S1-Class* 和 *$SP1-Class* 。

   多项式的结果看起来好多了。许多倾向评分均为 0.995 或更高，这些结果非常令人鼓舞。
2. 要确认此模型的性能更为优异，请将分析节点附加到 *class-poly* 模型块。

打开分析节点并单击**运行**。

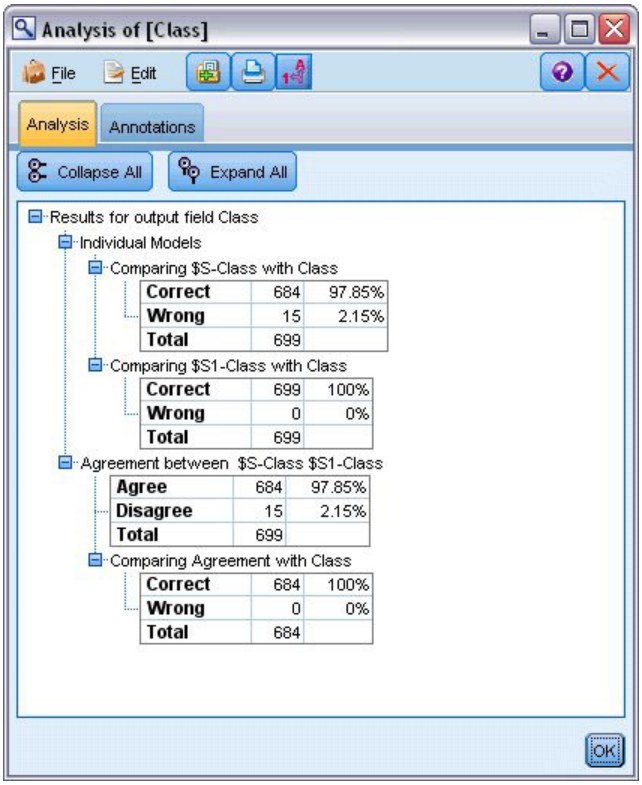


图 342. "分析"节点

此方法使用"分析"节点，它使您可以比较同一类型的两个或更多模型块。来自"分析"节点的输出显示 RBF 函数可以正确地预测 97.85% 的个案，这仍是一个不错的结果。但是，输出显示多项式函数已正确预测每个个案中的诊断。实际使用中，未必能做到完全准确，但分析节点可帮您确定模型的精确度能否满足特殊使用要求。

实际上，对于这个特定的数据集，其他两种函数类型（Sigmoid 和线性）的效果都不如多项式函数。但用于其他数据集时，其结果可能会明显不同，因此始终应该尝试所有选项。

摘要

您已使用多种不同类型的 SVM 内核函数根据多项属性来预测分类情况。此外您还了解了不同类型的核对于相同的数据集给出的结果存在多大差异，以及如何相对其他模型来度量某个模型是否性能更佳。

第 26 章 将 Cox 回归用于客户流失时间模型

作为减少客户流失计划的一部分，电信公司对建模“流失时间”很感兴趣，以便确定客户快速切换到其他服务的相关因素。为此，随机选取了一些客户样本，和他们作为客户所花费的时间（无论他们是否仍为活动客户）以及从数据库中抽取的其他各种字段。

此示例使用流 *telco_coxreg.str* ，此流参考的是数据文件 *telco.sav* 。数据文件和流文件分别位于 *Demos* 文件夹和 *streams* 子文件夹中。请参阅主题第 4 页的『Demos 文件夹』以获取更多信息。

构建合适的模型

1. 在 *Demos* 文件夹中添加指向 *telco.sav* 的“Statistics 文件”源节点。

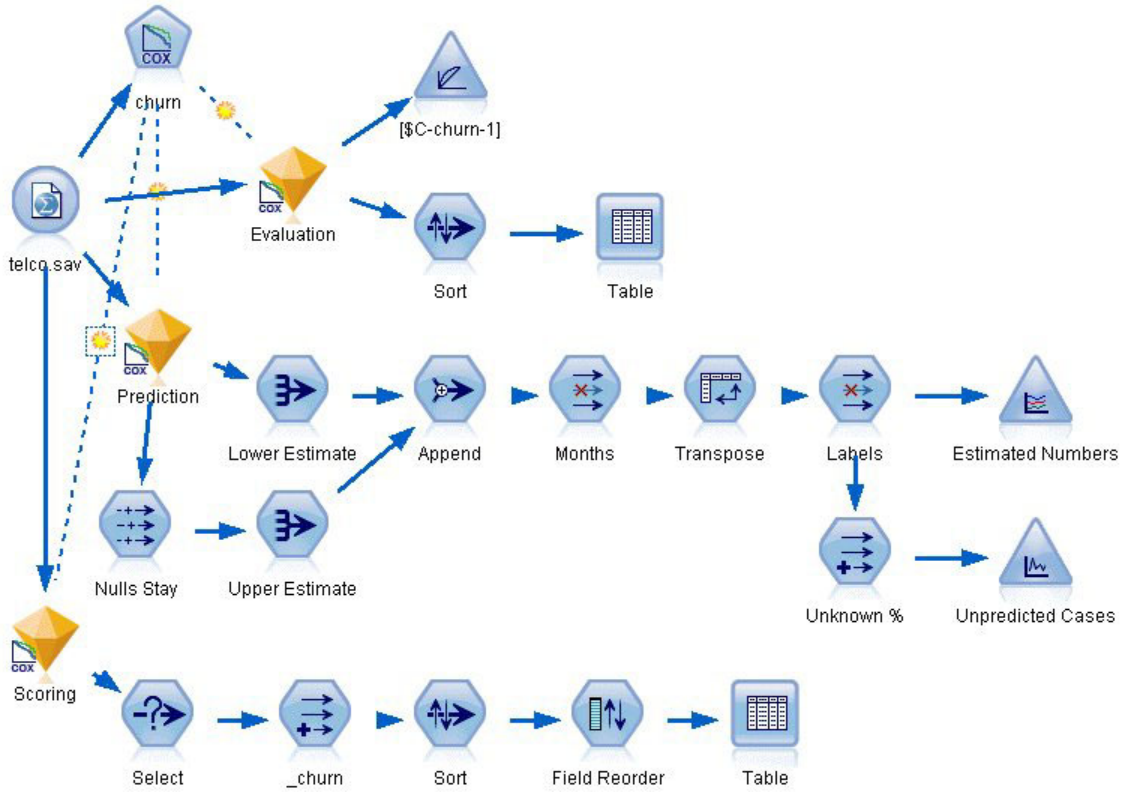


图 343. 用于分析流失时间的样本流

2. 在源节点的“过滤”选项卡上，排除*区域*、*收入字段*、从 *longten* 到 *wireten* 的字段，以及从 *loglong* 到 *logwire* 的字段。

图 344. 过滤不需要的字段

（或者，可以在“类型”选项卡上将这些字段的角色更改为**无**，而不用排除它，或者在建模节点中选择要使用的字段。）

3. 在源节点的“类型”选项卡上，将*流失*字段的角色设置为**目标**，将其测量级别设置为**标志**。将所有其他字段的角色设置为 **Input**。
4. 单击**读取值**以实例化数据。

图 345. 设置字段角色

5. 将 Cox 节点添加到源节点中；在 **字段** 选项卡中，选择 *保有期* 作为生存时间变量。

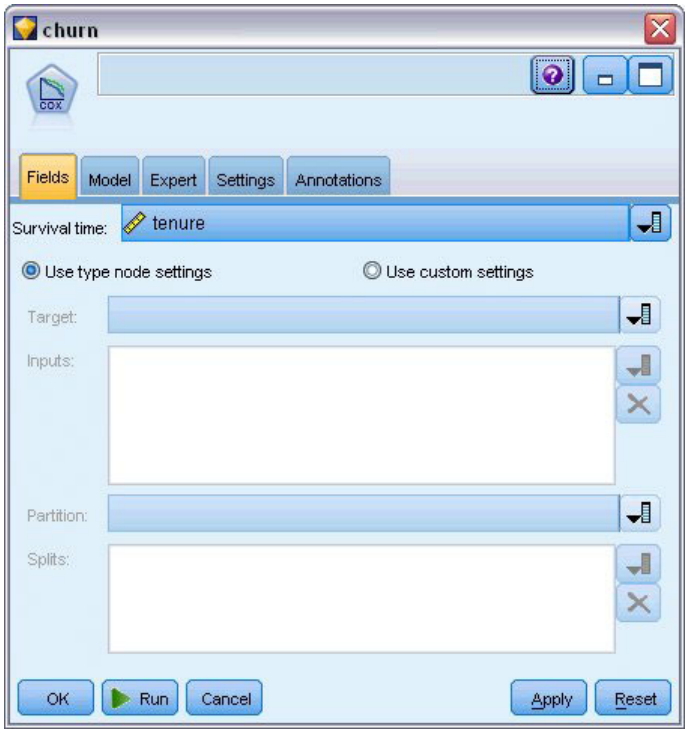


图 346. 选择字段选项

6. 单击 **模型** 选项卡。
7. 选择 **步进法** 作为变量选择方法。

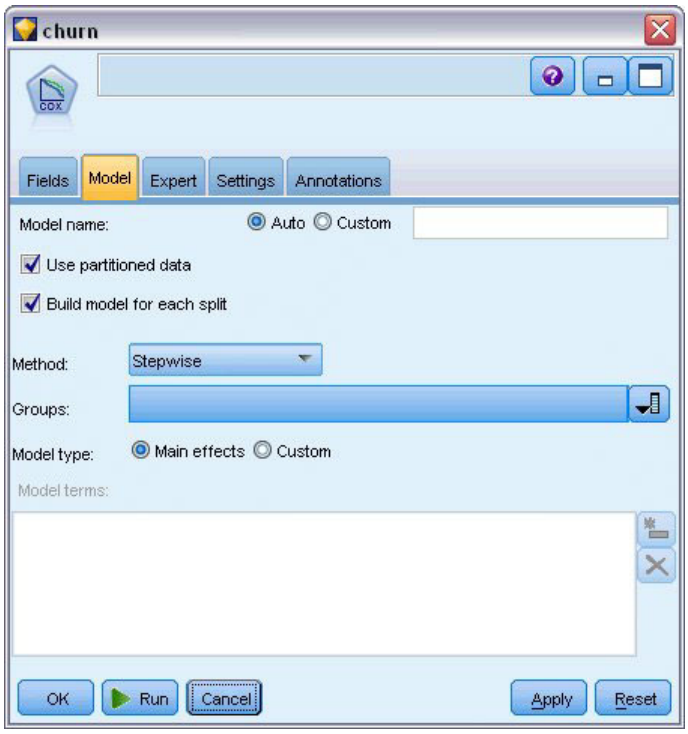


图 347. 选择模型选项

8. 单击 **专家** 选项卡并选择 **专家** 以激活专家建模选项。
9. 单击**输出**。

图 348. 选择高级输出选项

10. 选择**生存**和**风险**作为要生成的散点图，然后单击**确定**。
11. 单击**运行**以创建模型块，该模型块将被添加到流和位于右上角的“模型”选用板中。要查看其详细信息，请双击流上的模型块。首先，请查看“高级输出”选项卡。

删失的观测值

		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%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

图 349. 个案处理摘要

状态变量确定是否已发生给定观测值的事件。如果事件尚未发生，那么认为已审查该个案。已删失的观测值不能用于计算回归系数，但可用于计算基线风险。个案处理摘要显示 726 个个案已删失。该数字表示尚未流失的客户量。

分类变量编码

		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	

图 350. 分类变量编码

分类变量编码是解释分类协变量（特别是二分变量）回归系数的有用参考。缺省情况下，参考类别是“最后一个”类别。例如，即使已婚客户在数据文件中的变量值为 1，但为了回归的目的，这些变量值都编码为 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)

图 351. Omnibus 检验

模型构建过程采用向前步进算法。Omnibus 检验是指对模型执行情况的测量。上一步骤的卡方更改是上一步骤和当前步骤中模型的 2 对数似然之间的差值。如果在某一步中要添加变量，则在更改的显著性小于 0.05 时才能进行此包含操作。如果某一步中要移除变量，则在更改的显著性大于 0.10 时才能进行此排除操作。在 12 个步骤中，12 个变量将添加到模型中。

		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

图 352. 方程中的变量（仅步骤 12）

最终的模型包含 *地址* 、 *雇用状况* 、 *居住地址* 、 *equip* 、 *电话卡* 、 *longmon* 、 *equipmon* 、 *multline* 、 *声音* 、 *因特网* 、 *callid* ，以及 *电子帐单* 。要了解单个预测变量的效果，请查看 Exp(B)，可将 Exp(B) 解释为预测变量中单元增量风险中的预测更改。

- *地址*的 Exp(B) 值表示，对于居住在同一地址的客户，每年的流失风险会降低 100%–(100%× 0.966)=3.4% 。在同一地址居住五年的客户的流失风险会降低 $100\%-(100\%\times 0.966^5)=15.88\%$。

- *电话卡* 的 Exp(B) 值表示没有订购电话卡服务的客户流失的风险比率是订购此服务的客户的 2.175 倍。重新调用分类变量编码，其回归的 *No* = 1。
- *因特网* 的 Exp(B) 值表示未订购因特网服务的客户流失的风险比率是订购此服务的客户的 0.697 倍。这有点令人担忧，因为这表明使用该种服务的客户比不使用的客户取消公司服务的速度更快。

		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

图 353. 模型中没有的变量（仅步骤 12）

模型左侧变量的评分统计量的显著性值均大于 0.05。但是， *tollfree* 和 *cardmon* 的显著性值不小于 0.05，且与该值很接近。二者在今后的研究中有待进一步考证。

协变量平均值

	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

图 354. 协变量平均值

此表格显示了每个预测变量的平均值。在查看为均值构建的生存散点图时，此表格是很有用的参考。但请注意，在您查看分类预测变量的指示符变量均值时，“平均”客户实际上并不存在。即使使用所有刻度预测变量，您也无法找到一位其所有协变量值都接近均值的客户。如果要查看特定观测值的存活曲线，那么您可以更改协变量值，这些协变量用于在“散点图”对话框中绘制存活曲线。如果要查看特定观测值的存活曲线，那么您可以更改协变量值，这些值用于在“高级输出”对话框的散点图组中绘制存活曲线。

存活曲线

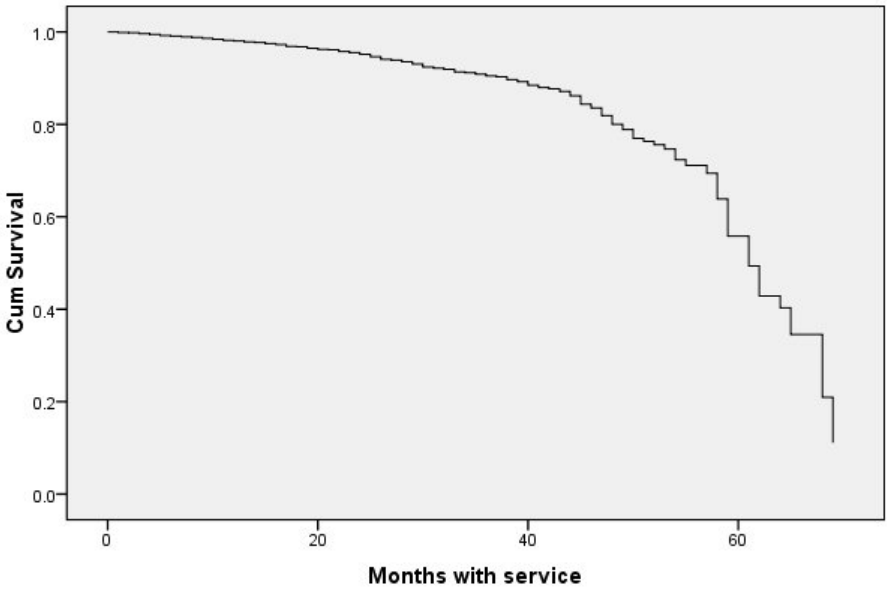


图 355. “平均”客户的存活曲线

基本存活曲线是“平均”客户的模型预测流失时间的可视化显示。水平轴显示事件发生的时间。垂直轴显示生存概率。所以，存活曲线上的任何一点表示“平均”客户经过某段时间仍未流失的概率。55 个月过后，存活曲线将变得不平滑。很少有客户那么长的时间使用公司服务，这样可获取的信息变少，导致曲线变成块状。

风险曲线

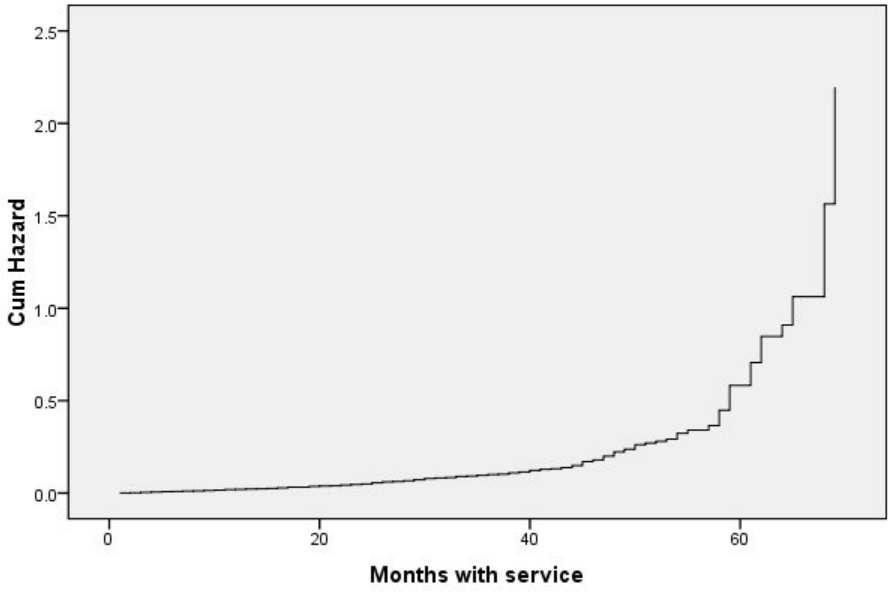


图 356. “平均”客户的风险曲线

基本风险曲线是“平均”客户的累积模型预测流失可能性的可视化显示。水平轴显示事件发生的时间。垂直轴显示累积风险，等于生存概率的负对数。55 个月过后，同存活曲线一样，风险曲线也变得不平滑，变化原因相同。

评估(E)

逐步选择法保证模型仅包含“统计意义下显著”的预测变量，但不保证模型实际上非常适用于预测目标。为此，需要分析已评分的记录。

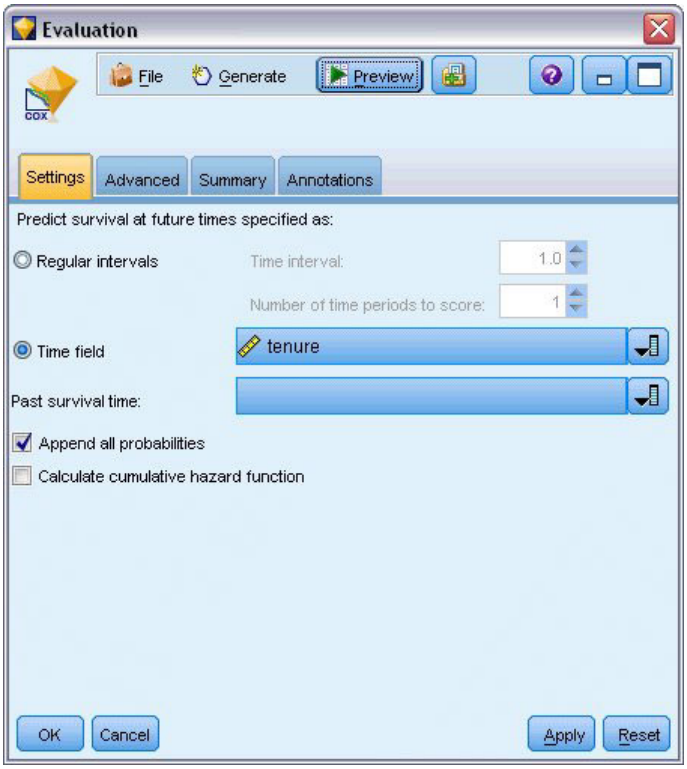


图 357. Cox 块: “设置”选项卡

1. 将模型块放置在工作区上并将其附加到源节点，然后打开该模型块并单击“设置”选项卡。
2. 选择 **时间字段** 并指定 *保有期* 。将根据每个记录的保有期长度对其进行评分。
3. 选择 **追加所有概率** 。

   这样可使用 0.5 作为分界值区分客户是否流失来创建评分；如果客户流失的倾向值大于 0.5，那么这些客户将被标记为流失的客户。该数值并不特殊，但不同的分界值可能产生不同的期望结果。在考虑选择分界值时，有一种方法是使用评估节点。

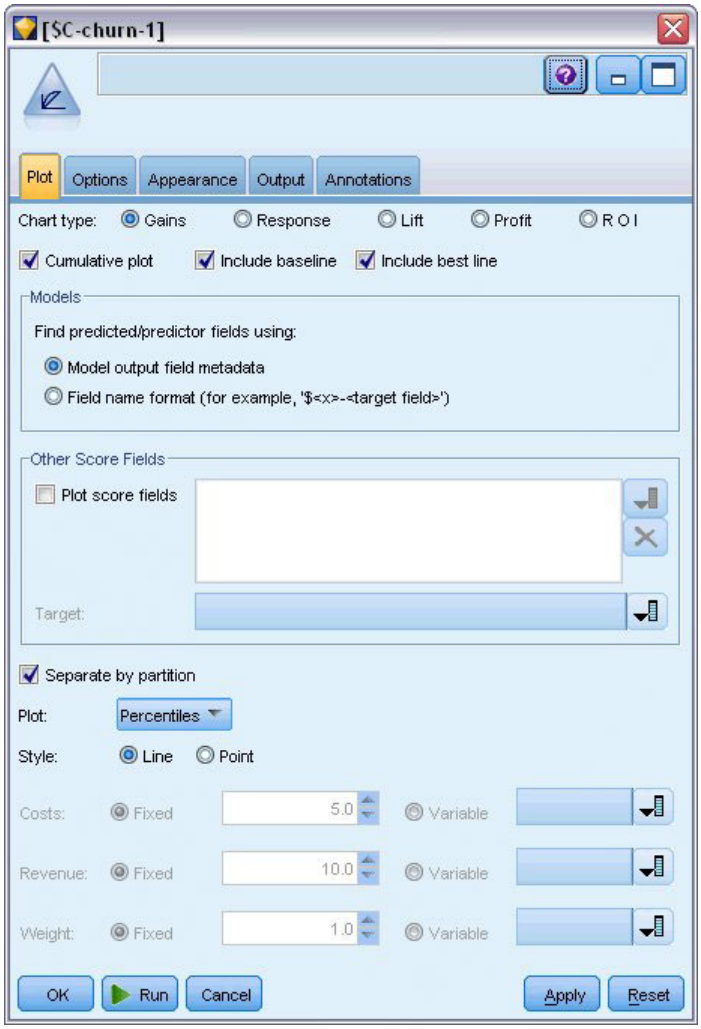


图 358. "评估"节点: "图"选项卡

4. 将评估节点附加到模型块; 在"图"选项卡上, 选择**包含最佳线**。
5. 单击**选项**选项卡。

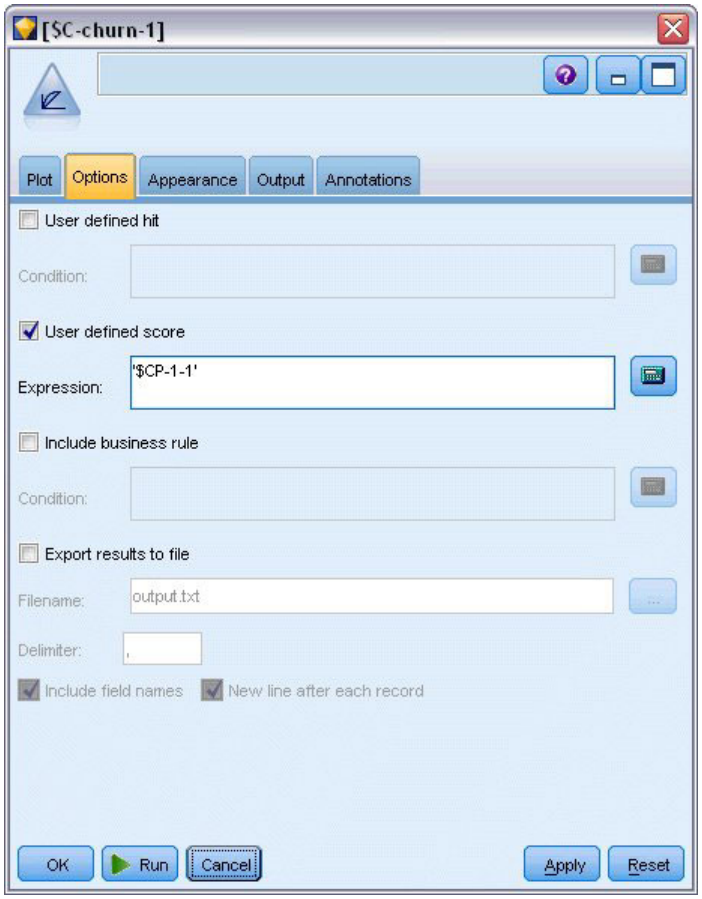


图 359. “评估”节点: “选项”选项卡

6. 选择 **用户定义的评分** ，然后键入 '$CP-1-1' 作为表达式。这是与流失倾向相对应的模型生成的字段。
7. 单击**运行**。

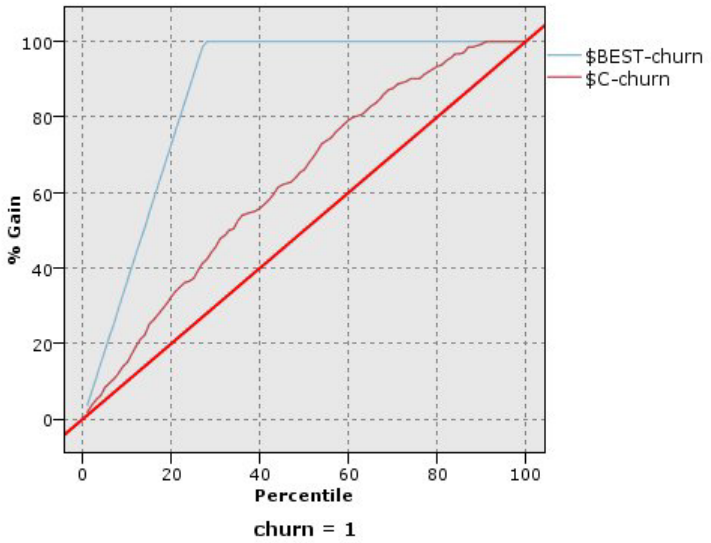


图 360. 增益图

累积增益图会在给定的类别中显示通过把个案总数的百分比作为目标而“增益”的个案总数的百分比。例如，曲线上的一点 (10%, 15%)，表示当您使用模型对数据集进行评分，并使用预测的流失倾向对所有观测值进行排序时，您可能希望前 10% 中包含实际类别为 *1*（流失者）的所有观测值中约 15% 的观测值。同样，前 60% 包含大约 79.2% 的流失顾客。如果选择整个已评分数据集，那么将获得数据集中所有流失者。

对角线是“基线”曲线；如果从已评分数据集中随机选择 20% 的记录，那么将期望从实际具有类别 *1* 的所有记录中“获得”约 20% 的记录。曲线所处的位置越高于基线，增益越大。“最佳”线显示将较高的倾向评分分配到每个流失者（而非每个非流失者）的“完美”模型的曲线。可以通过选择与所需增益相对应的百分比，然后将此百分比映射到相应的分界值，从而选择有助于选择分类分界值的累积增益图。

“期望”增益的组成取决于类型 I 与类型 II 错误的成本。即，将流失客户分类为非流失客户（类型 I）的成本是多少？将非流失客户分类为流失客户（类型 II）的成本是多少？当客户保持成为首要考虑的问题时，您会希望降低第一类错误；在累积增益图上，这对应于预测倾向值为 *1* 的前 60% 中的客户已增加的客户维护，这将捕获 79.2% 的可能流失者，但将花费时间和资源用于获得新客户。如果降低维护当前客户群成本是首先要考虑的事，那么您会希望降低类型 II 错误。在图表上，这可能对应于前 20% 增加的客户需求，这将捕获 32.5% 的流失客户。通常，这两方面都非常重要，因此，您必须选择一个决策规则，用于对具有最大敏感度和特异性的客户进行分类。

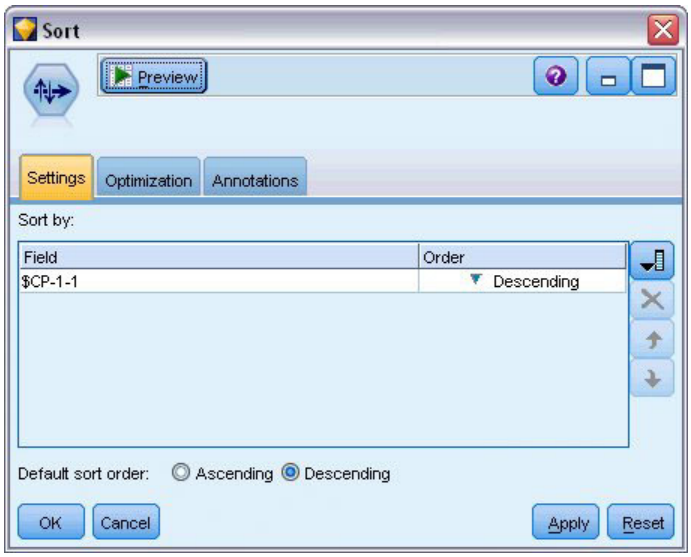
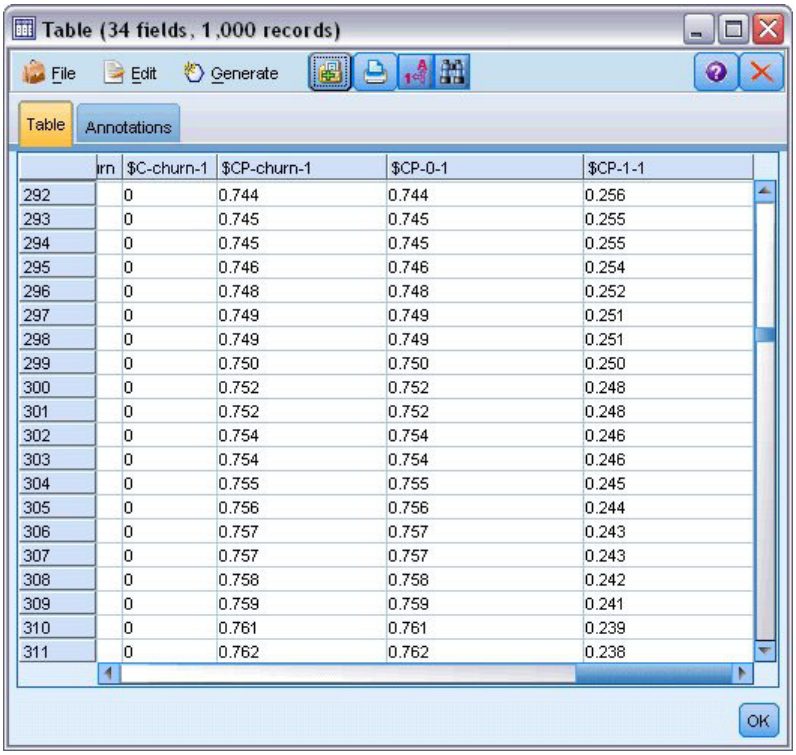


图 361. “排序”节点: “设置”选项卡

8. 假定您所定的期望增益为 45.6%，则相应会显示前面 30% 的记录。要找到合适的分类分界值，请将排序节点附加到模型块。
9. 在“设置”选项卡上，选择以降序按 *$CP-1-1* 排序，然后单击**确定**。



	$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

图 362. 表

10. 将“表”节点附加到“排序”节点。
11. 打开“表”节点，然后单击**运行**。

将输出向下滚动，就会看到第 300 个记录的 *$CP-1-1* 值为 0.248。使用 0.248 作为分类分界值，就会导致约 30% 的客户被评定为流失者，由此捕获实际总流失者的个数约占 45%。

跟踪仍在的预期客户数

如果对模型感到满意，那么您会希望跟踪数据集中未来两年内会保留的预期客户数。空值代表客户的总保有期（未来时间 + *保有期* ）不在用于训练模型的数据中的生存时间范围内，呈现出一种很有趣的挑战。处理空值的一种方法是创建两个预测集合，其中一个集合中的空值被假定为已流失，另一个集合中的空值被假定为已保留。使用这种方法，可以创建保留的预测客户数的上限和下限。

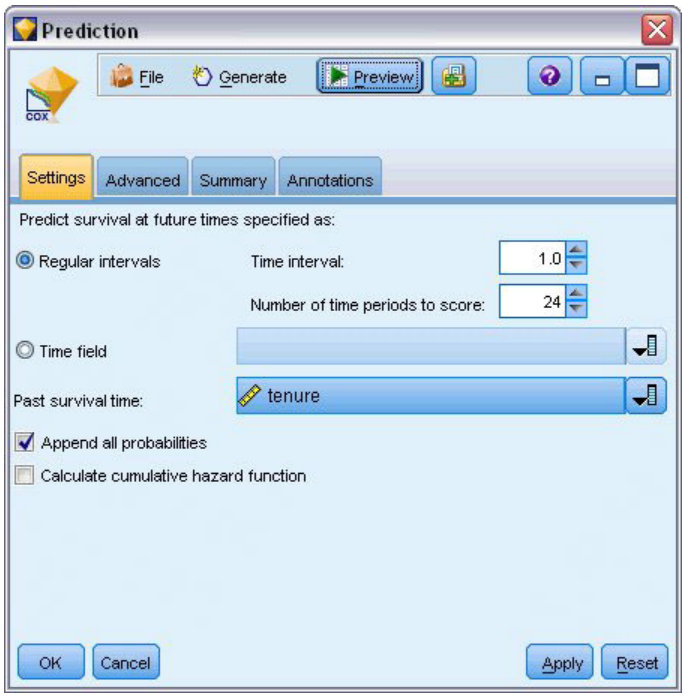


图 363. Cox 块: “设置”选项卡

1. 双击“模型”选用板中的模型块（或者复制并粘贴流画布上的模型块），并将新的模型块附加到“源”节点。
2. 打开模型块的“设置”选项卡。
3. 确保选中**定期**，指定 1.0 作为时间间隔并指定 24 作为评分周期数。此项操作指定在未来 24 个月中每个月都会对每个记录进行评分。
4. 选择 *tenure* 作为指定过去生存时间的字段。评分算法将考虑每个客户持续作为公司客户的时间长度。
5. 选择 **追加所有概率** 。

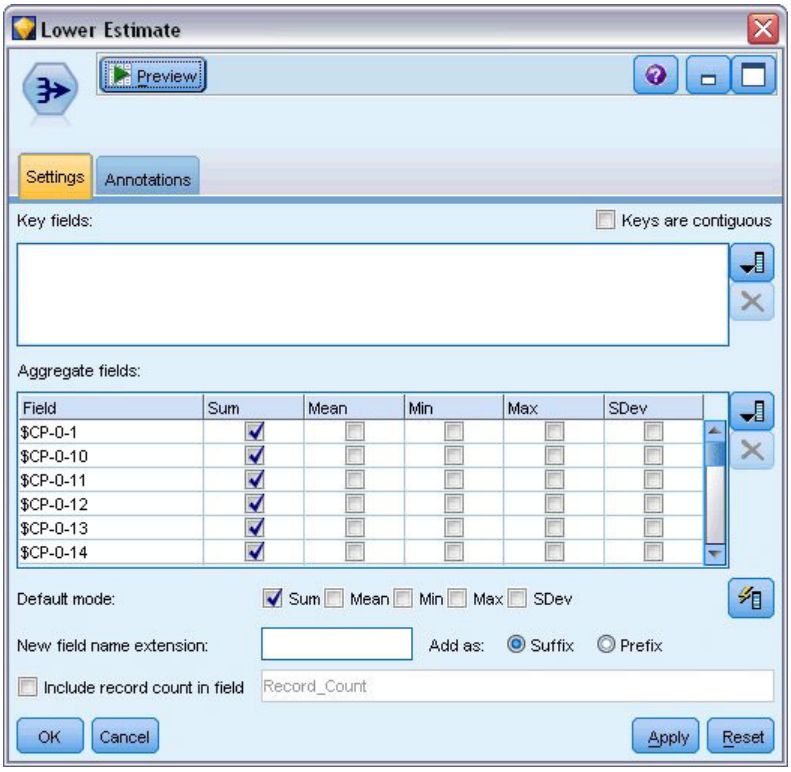


图 364. "汇总"节点: "设置"选项卡

6. 将汇总节点添加到模型块; 在"设置"选项卡上, 取消选中**均值**作为缺省模式。
7. 选择从 *$CP-0-1* 到 *$CP-0-24* 的字段（即格式为 *$CP-0-n* 的字段）作为要汇总的字段。如果在"选择字段"对话框上, 按名字（即字母顺序）对字段进行排序, 则这种方法最容易。
8. 取消选中 **在字段中包含记录计数** 。
9. 单击**确定**。此节点可创建"下限"预测。

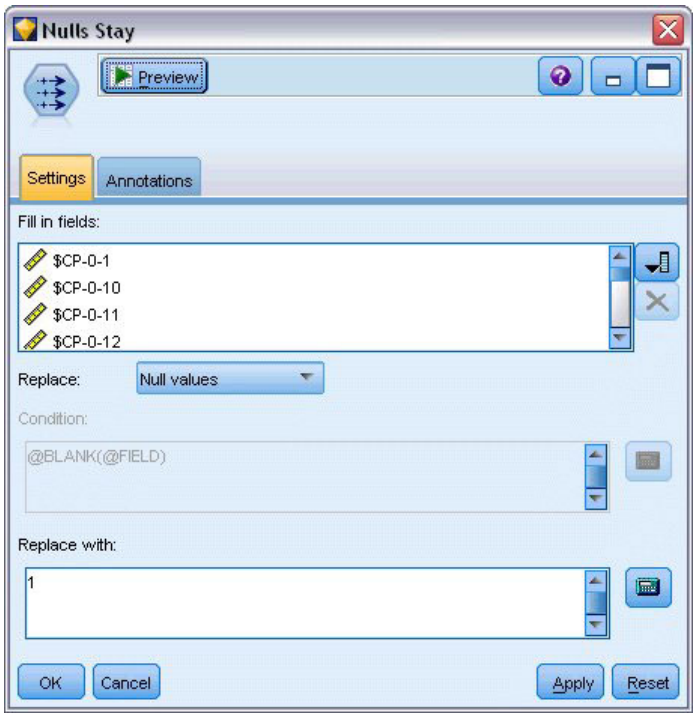


图 365. "填充"节点: "设置"选项卡

10. 将"填充"节点附加到刚才已附加"汇总"节点的 Coxreg 块; 在"设置"选项卡上, 选择格式为 *$CP-0-n* 的字段 *$CP-0-1* 到 *$CP-0-24* 作为要填充的字段。如果在"选择字段"对话框上, 按名字(即字母顺序)对字段进行排序, 则这种方法最容易。
11. 选择使用值 1 替换 **空值** 。
12. 单击**确定**。

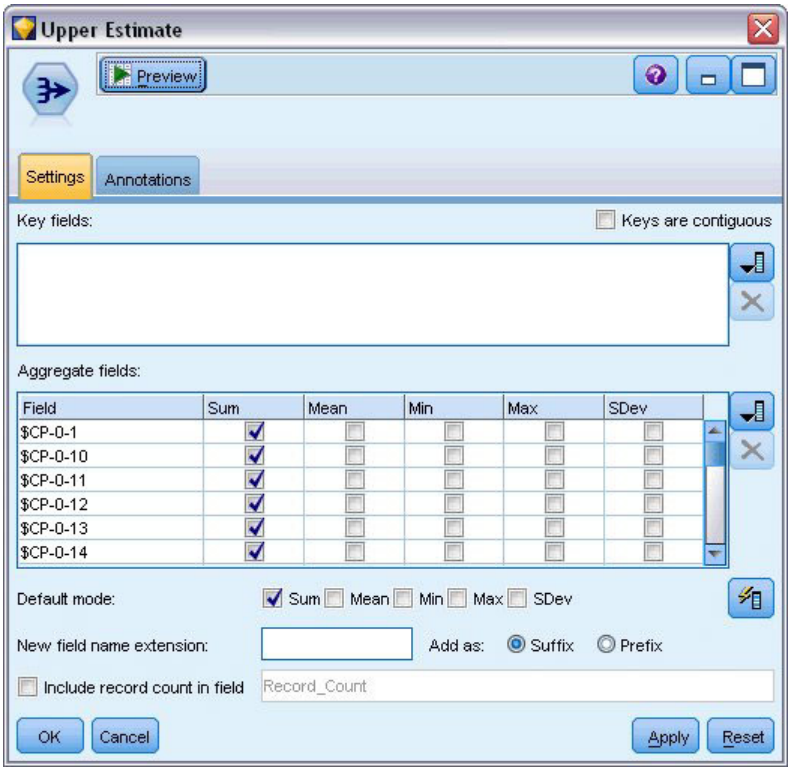


图 366. "汇总"节点: "设置"选项卡

13. 将汇总节点附加到填充节点；在"设置"选项卡上，取消选中**均值**作为缺省模式。
14. 选择从 *$CP-0-1* 到 *$CP-0-24* 的字段（即格式为 *$CP-0-n* 的字段）作为要汇总的字段。如果在"选择字段"对话框上，按名字（即字母顺序）对字段进行排序，则这种方法最容易。
15. 取消选中 **在字段中包含记录计数** 。
16. 单击**确定**。此节点可创建"上限"预测。

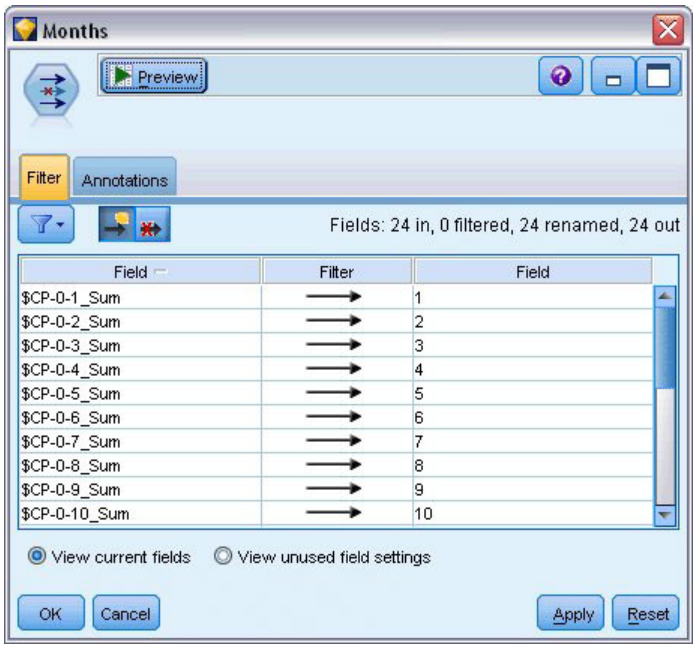


图 367. "过滤"节点: "设置"选项卡

17. 将追加节点附加到两个汇总节点; 然后将过滤节点附加到追加节点。
18. 在过滤节点的"设置"选项卡上, 将字段重新命名为 *1* 到 *24*。使用转置节点, 这些字段名称将变为图表下游中 *x* 轴上的值。

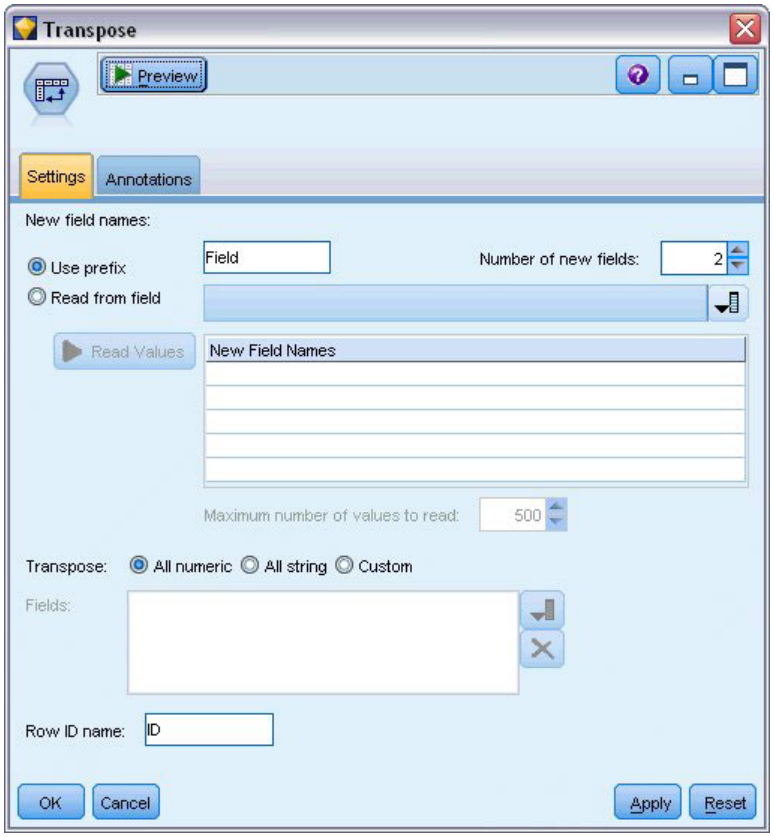


图 368. “变换”节点: “设置”选项卡

19. 将转置节点附加到过滤节点。
20. 第 2 类作为新字段数。

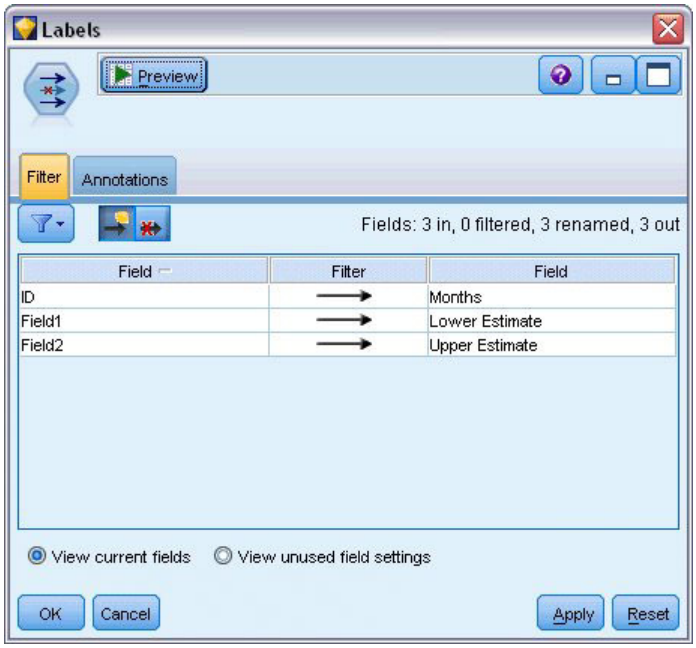


图 369. “过滤”节点: “过滤”选项卡

21. 将过滤节点附加到转置节点。
22. 在过滤节点的“设置”选项卡上，将*标识*重新命名为*月*，将*字段1*重新命名为*较低估计*，将*字段2*命名为*较高估计*。

图 370. “多图”节点：“图”选项卡

23. 将多重散点图节点附加到过滤节点。
24. 在“散点图”选项卡上，*月*是 X 字段，*较低估计*和*较高估计*是 Y 字段。

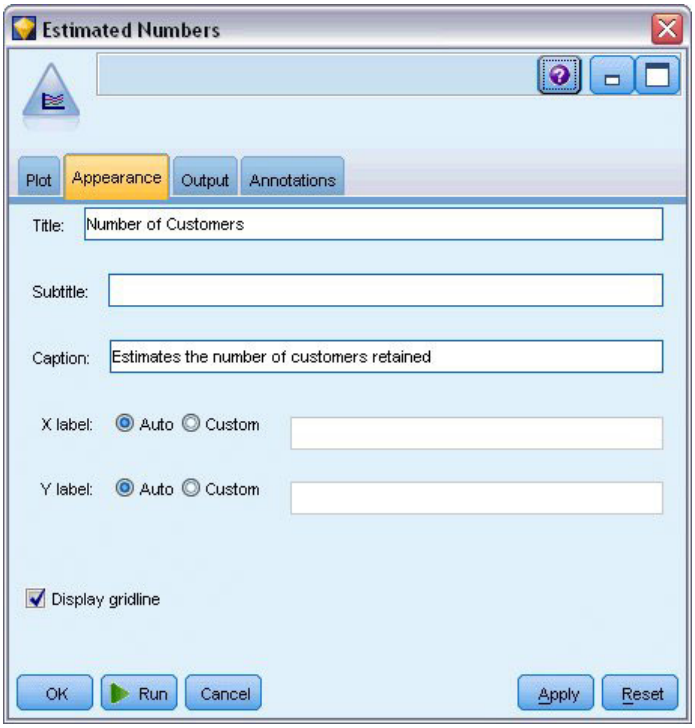


图 371. "多图"节点: "外观"选项卡

25. 单击"外观"选项卡。
26. 键入 客户数 作为标题。
27. 键入 估计保留的客户数 作为标注。
28. 单击**运行**。

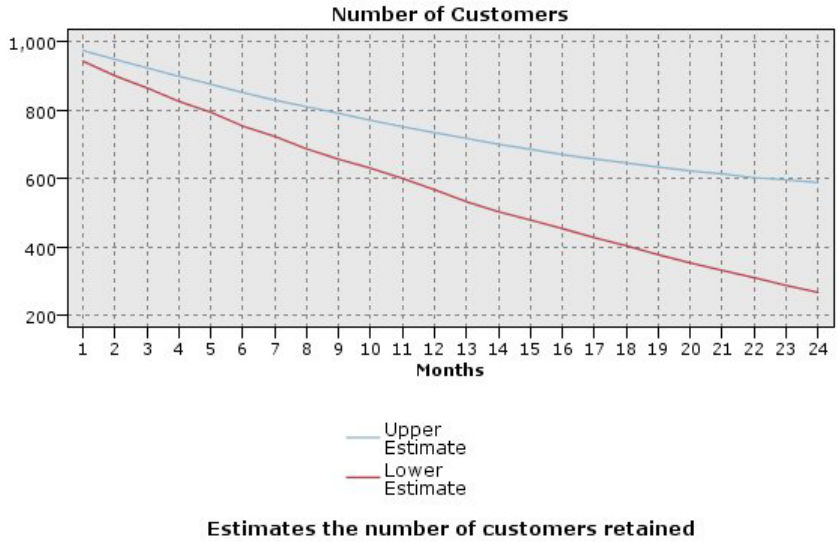


图 372. 预测保留的客户数的多图。

已绘制出估计的保留客户数的上限和下限。这两条线的差值是评分为空值的客户数，因此其状态很难确定。这些客户数量随时间增加。12 个月过后，可以预计数据集中保留的原始客户数介于 601 到 735 之间;

24 个月过后，该值介于 288 到 597 之间。

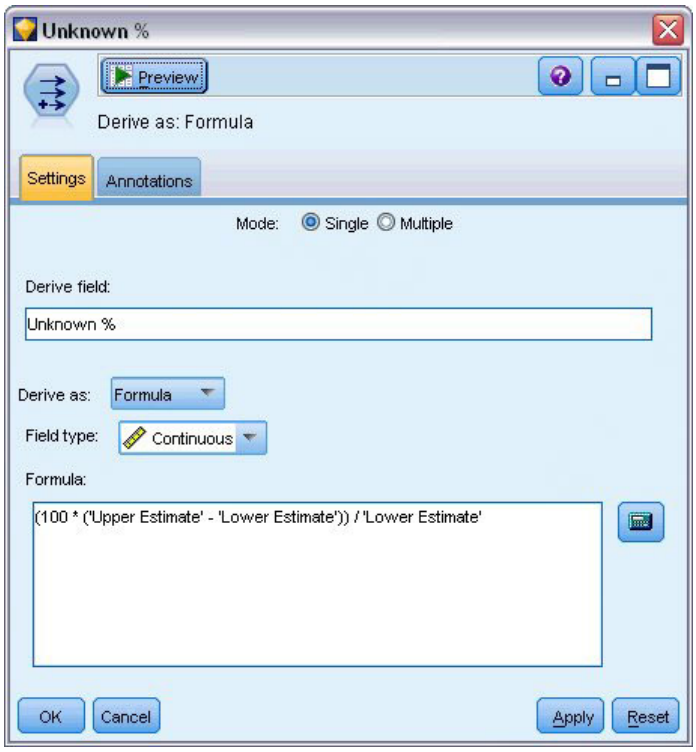


图 373. “派生”节点: “设置”选项卡

29. 要查看保留的客户数估计的不确定程度，请将导出节点附加到过滤节点。
30. 在导出节点的“设置”选项卡上，键入*未知百分比*作为导出字段。
31. 选择**连续**作为字段类型。
32. 键入 `(100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate'` 作为公式。*未知百分比*是作为较低估计百分比的“质疑”客户数。
33. 单击**确定**。

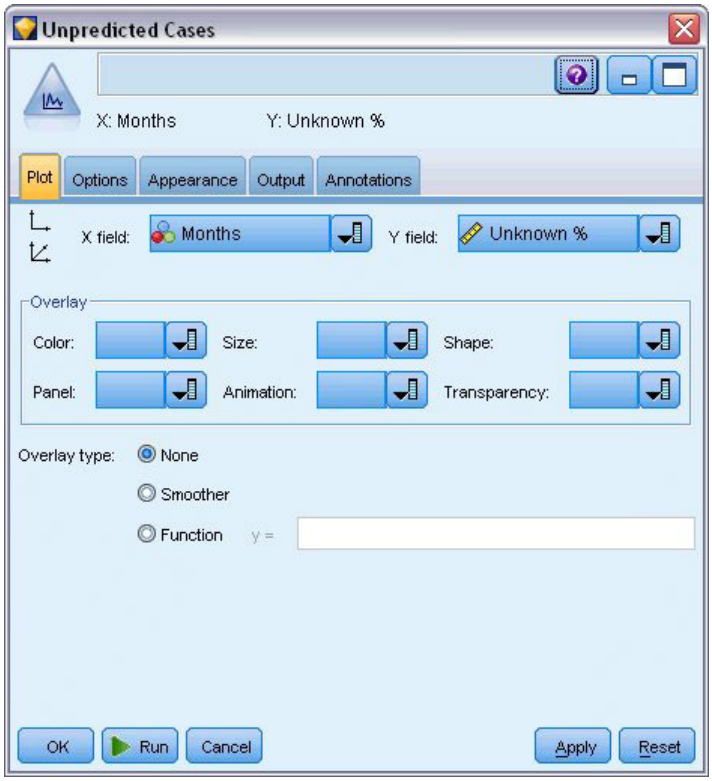


图 374. “图”节点: “图”选项卡

34. 将散点图节点附加到导出节点。
35. 在散点图节点的“散点图”选项卡上，选择*月*作为 X 字段，*未知百分比*作为 Y 字段。
36. 单击“**外观**”选项卡。

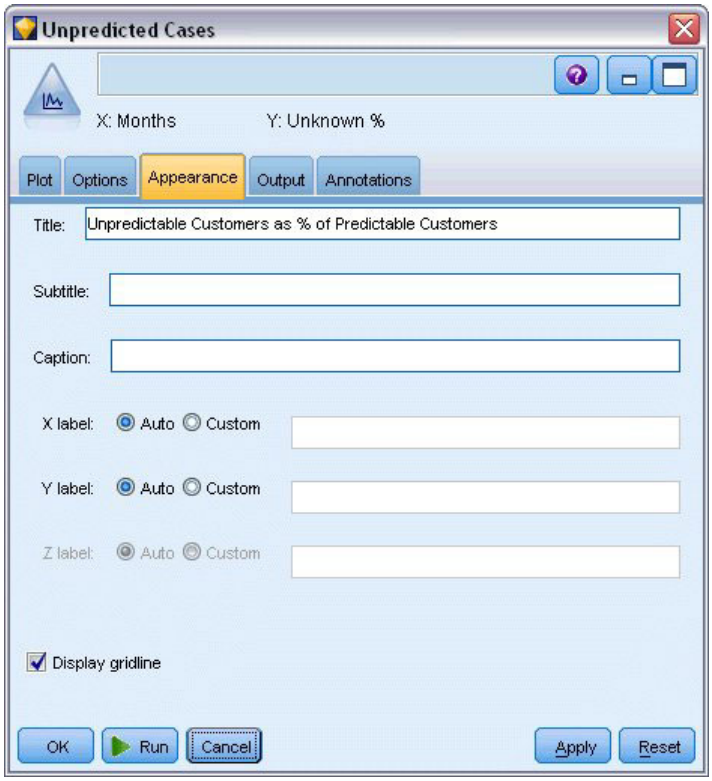


图 375. “图”节点：“外观”选项卡

37. 键入 `Unpredictable Customers as % of Predictable Customers` 作为标题。
38. 执行节点。

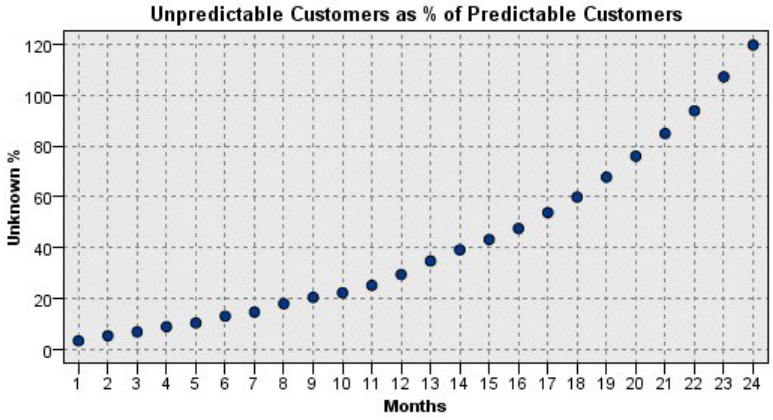


图 376. 关于不可预测的客户的图

第一年里，无法预测客户的百分比以明显的线性速率增长。但第二年增长速率猛增，一直到第 23 个月，具有空值的客户数超过了保留的预测客户数。

评分

如果对模型感到满意，那么您会希望对客户进行评分以确认下一年一个季度内最可能流失的客户。

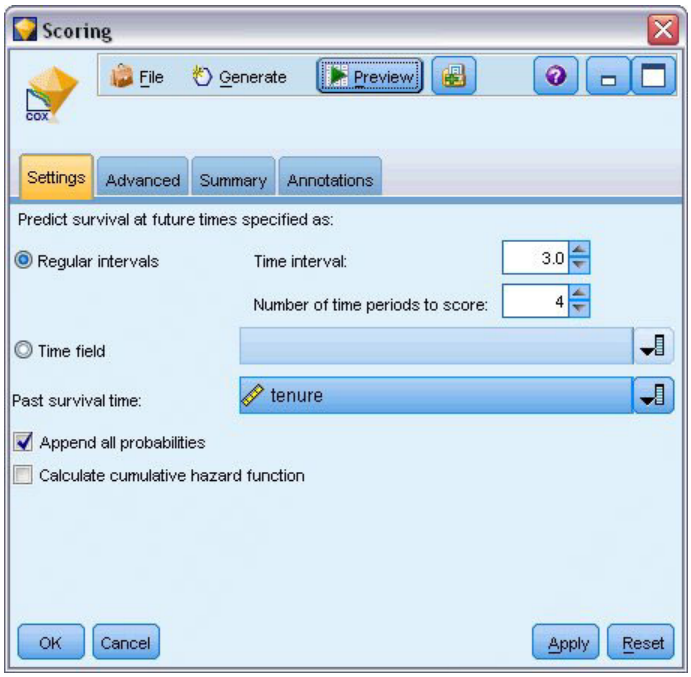


图 377. Coxreg 块: "设置"选项卡

1. 将第三个模型块附加到源节点并打开模型块。
2. 确保选中 **规则区间** ，并指定 `3.0` 为时间区间， `4` 为要对其评分的时段数。此项操作指定在未来四个季度中将对每个记录进行评分。
3. 选择 *tenure* 作为指定过去生存时间的字段。评分算法将考虑每个客户持续作为公司客户的时间长度。
4. 选择 **追加所有概率** 。使用这些附加字段更容易对表中要查看的记录进行排序。

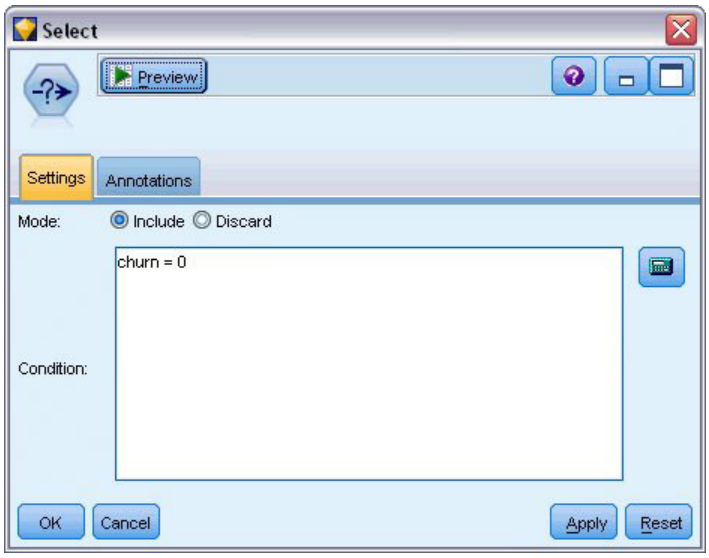


图 378. "选择"节点: "设置"选项卡

5. 将选择节点附加到模型块; 在"设置"选项卡上，键入 `churn=0` 作为条件。这将除去已从结果表中流失的客户。

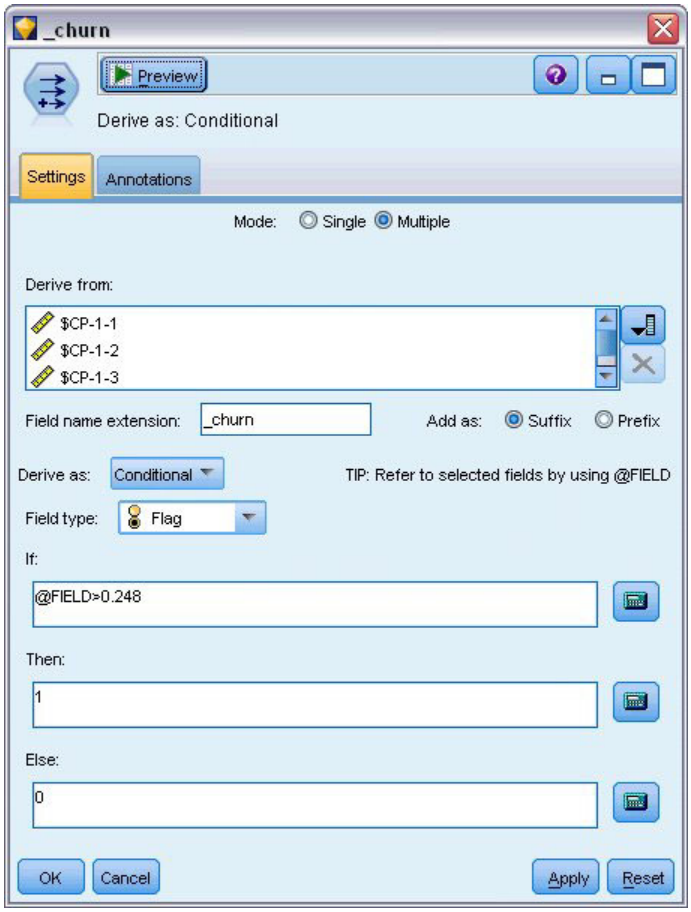


图 379. "派生"节点: "设置"选项卡

6. 将导出节点附加到选择节点; 在"设置"选项卡上, 选择**多个**作为模式。
7. 选择从格式为 *$CP-1-n* 的字段 *$CP-1-1* 到 *$CP-1-4* 进行派生, 然后输入 `_churn` 作为要添加的后缀。如果在"选择字段"对话框上, 按名字(即字母顺序)对字段进行排序, 则这种方法最容易。
8. 选择将字段导出为 **条件** 。
9. 选择**标志**为测量级别。
10. 键入 `@FIELD >0.248` 作为 **If** 条件。请记住, 这是评估期间确定的分类分界值。
11. 键入 `1` 作为 **Then** 表达式。
12. 键入 `0` 作为 **Else** 表达式。
13. 单击**确定**。

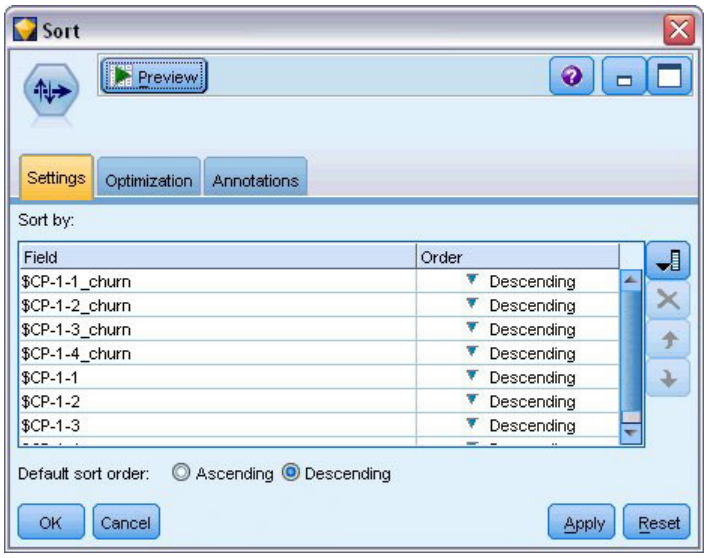
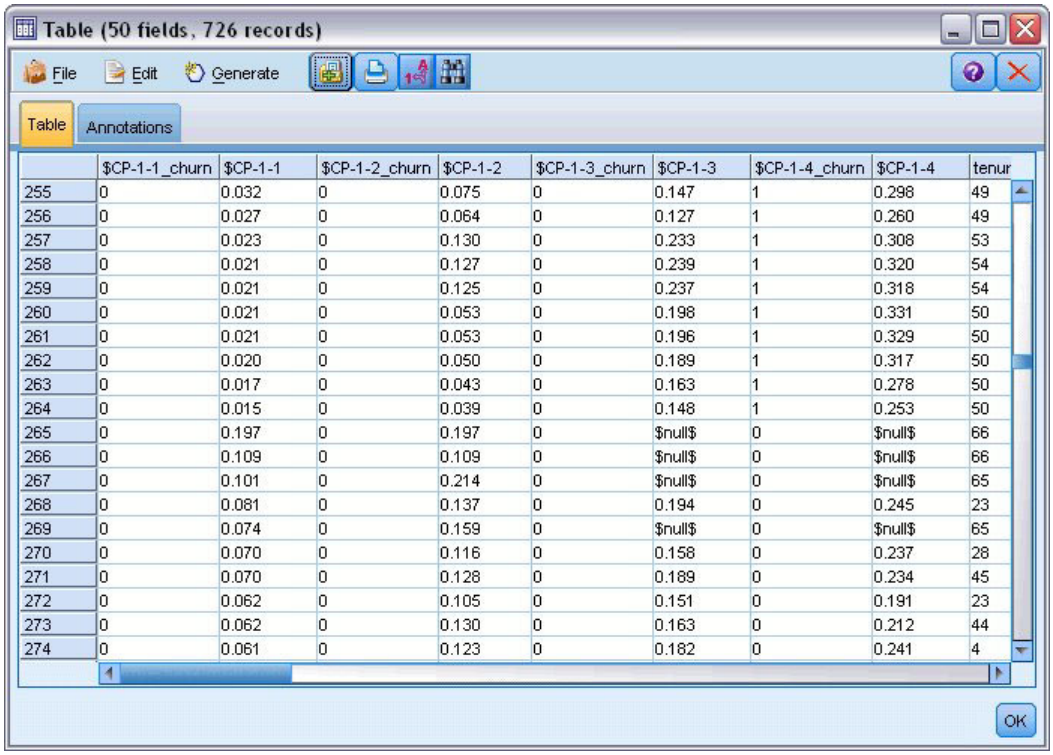


图 380. "排序"节点: "设置"选项卡

14. 将排序节点附加到导出节点；在"设置"选项卡上，选择先按 *$CP-1-1_churn* 到 *$CP-1-4-churn*、再按 *$CP-1-1* 到 *$CP-1-4* 进行排序，所有顺序都按降序排列。预测要流失的客户会出现在顶端。

图 381. "字段重新排序"节点: "重新排序"选项卡

15. 将字段重排节点附加到排序节点；在"重排"选项卡上，选择将 *$CP-1-1_churn* 到 *$CP-1-4* 放置在其他字段之前。此项操作会使结果表更易于读取，且是可选的。您将需要使用按钮将字段移动到图中显示的位置。

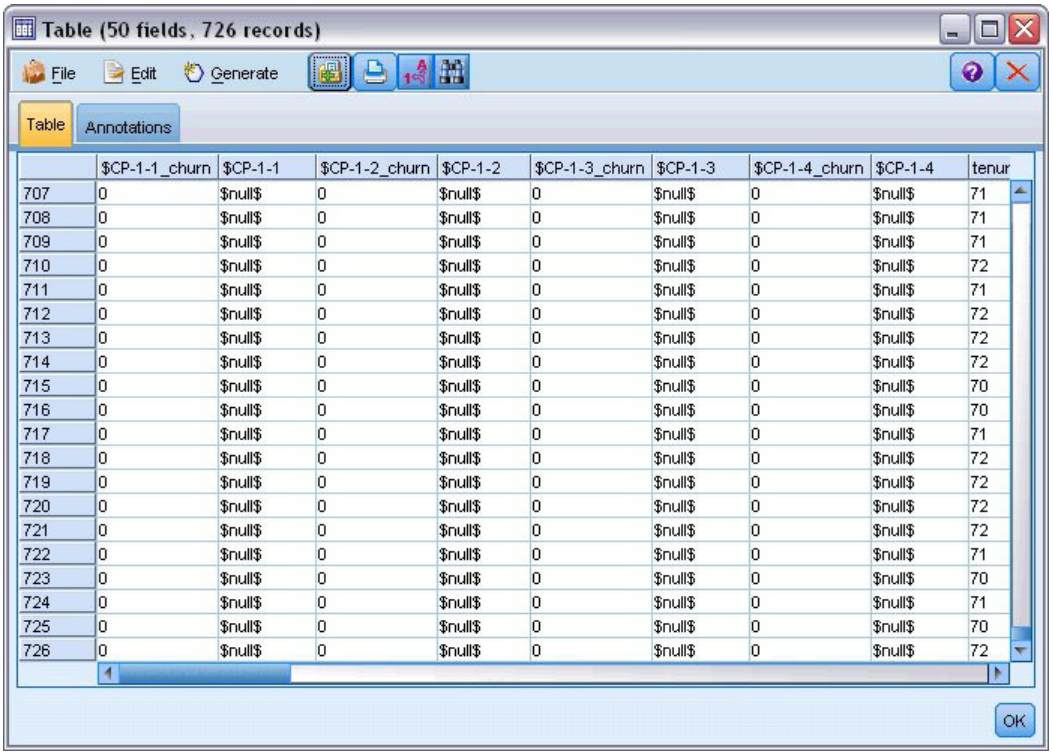


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	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

图 382. 显示客户评分的表

16. 将表节点附加到字段重排节点并运行。

预计年末将流失 264 位客户，第三个季度末流失 184 位，第二个季度末流失 103 位，第一个季度末流失 31 位。请注意，给定了两类客户，在第一季度中有较高流失倾向的一类客户不一定在以后的季度中有较高的流失倾向；例如，查看记录 256 和 260。这可能是由于客户当前保有期之后的几个月内风险函数的形状所致；例如，与那些因个人推荐而加入的客户相比，因促销而加入的客户有可能更早流失，但如果他们未流失，那么他们在其保有期的余下时间内实际上可能更忠诚。您可能希望对这些客户重新排序以获得最可能流失的客户的不同视图。



	$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

图 383. 显示具有空值的客户的表

该表的底部是具有预测空值的客户。这些客户的总保有期（未来时间 + *保有期* ）在用于训练模型的数据中的生存时间范围内。

摘要

通过使用 Cox 回归，您已找到适用于流失时间的模型，该模型绘制接下来两年中保留的预测客户数，并标识下一年中最可能流失的客户。请注意，这只是适用模型，它不一定是最佳模型。理想的做法是，您至少应当将使用向前步进法获取的该模型与使用后退步进法创建的模型进行比较。

IBM SPSS Modeler Algorithms Guide 中说明了 IBM SPSS Modeler 中使用的建模方法的数学基础。

第 27 章 市场购物篮分析（规则归纳/C5.0）

本示例处理描述超级市场购物篮内容（即，所购买的全部商品的集合）的虚构数据，以及购买者的相关个人数据（可通过忠诚卡方案获得）。目的是寻找购买相似产品并且可按人口统计学方式（如按年龄、收入等）刻画其特征的客户群。

本示例说明了数据挖掘的两个阶段:

- 关联规则建模和一个揭示所购买商品之间联系的 Web 显示
- C5.0 规则归纳（描绘已标识产品组的购买者的特征）

注: 此应用不直接使用预测建模，因此不会对最终模型进行准确性度量，并且在数据挖掘过程中也不存在相关联的训练/检验区分。

本例使用名为 *baskrule* 的流，该流引用名为 *BASKETS1n* 的数据文件。这些文件可在任何 IBM SPSS Modeler 安装程序的 *Demos* 目录中找到。此目录可通过 Windows 的“开始”菜单中的 IBM SPSS Modeler 程序组进行访问。文件 *baskrule* 位于 *streams* 目录下。

访问数据

使用“变量文件”节点连接到数据集 *BASKETS1n*，选择要从该文件读取的字段名称。将“类型”节点连接到数据源，然后将该节点连接到“表”节点。将字段*卡标识*的测量级别设置为*无类型*（因为每个忠诚卡标识在数据集中只出现一次，因此对于建模没有用处）。选择*名义*作为字段*性别*的测量级别（这是为了确保 Apriori 建模算法不会将*性别*视为标志）。

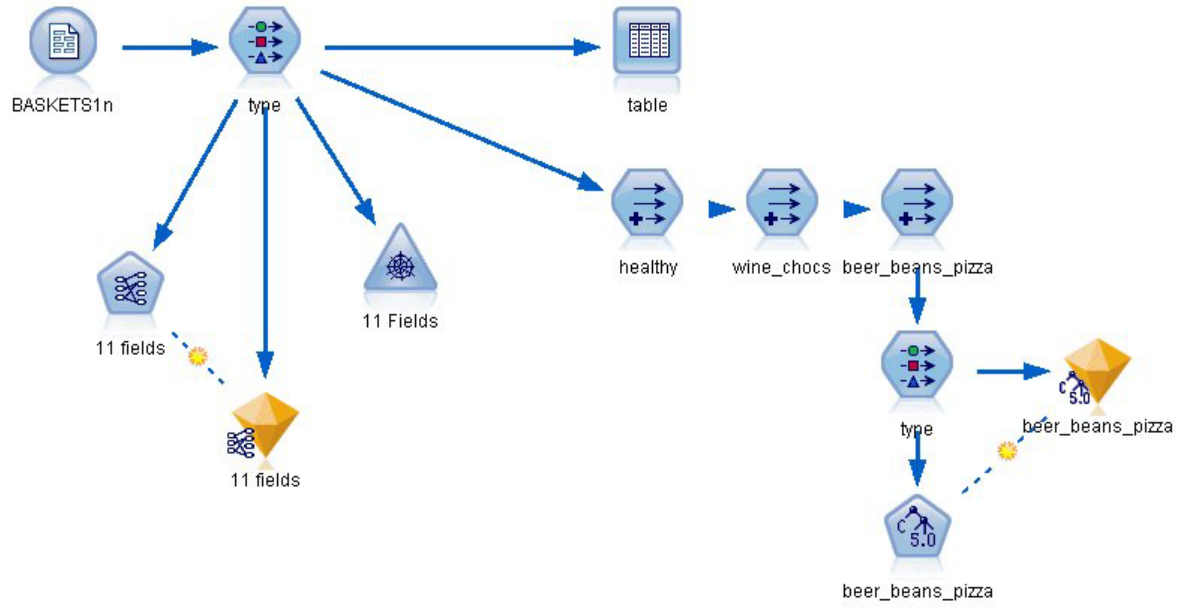


图 384. 购物规则流

现在，运行该流以将“类型”节点实例化并显示表。数据集包含 18 个字段，其中每条记录表示一个购物篮。

下列标题中会显示 18 个字段。

购物篮摘要:

- *cardid*。购买此篮商品的客户的忠诚卡标识。
- *value*。购物篮的总购买价格。
- *pmethod*。购物篮的支付方法。

持卡人的个人详细信息:

- *sex*
- *homeown*。卡持有者是否拥有住房。
- *收入*
- *age*

购物篮内容 - 产品类别的出现标志:

- *fruitveg*
- *freshmeat*
- *dairy*
- *cannedveg*
- *cannedmeat*
- *frozenmeal*
- *beer*
- *wine*
- *softdrink*
- *fish*
- *confectionery*

发现购物篮内容的关系

首先，需要使用 Apriori 大致了解购物篮内容的亲缘关系（关联）以生成关联规则。选择要在此建模过程中使用的字段，方法是：编辑“类型”节点，将所有产品类别的角色设置为*两者*，并将所有其他角色设置为*无*。（*两者*表示该字段可以是结果模型的输入，也可以是输出。）

注：通过按住 Shift 键并单击以选择多个字段，然后指定列中的选项，您可以为多个字段设置选项。

图 385. 选择用于建模的字段

指定了用于建模的字段后，请将 Apriori 节点附加到“类型”节点，编辑它，选择选项**只显示值为真的标志变量**，然后在 Apriori 节点上单击“运行”。结果（管理器窗口右上角“模型”选项卡上的模型）包含您可以查看（使用上下文菜单，然后选择**浏览**）的关联规则。

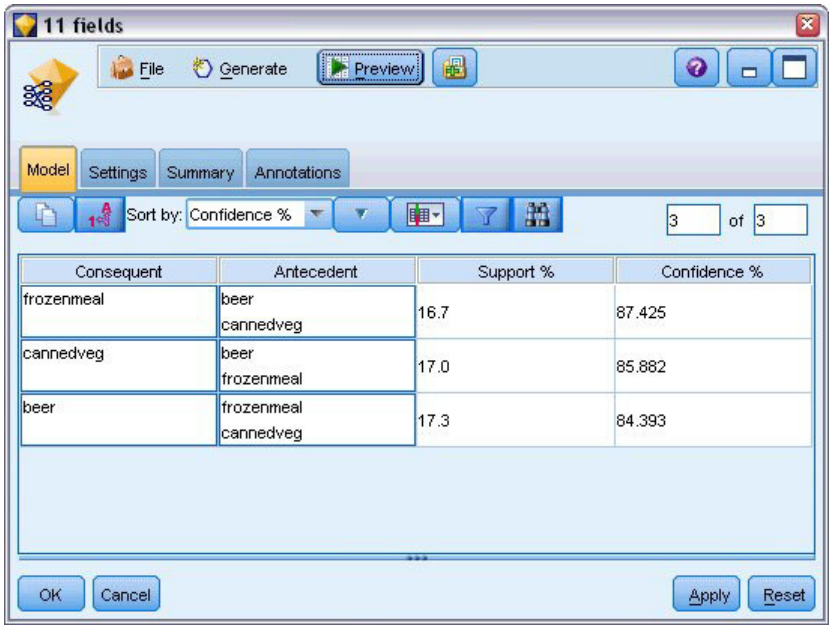


图 386. 关联规则

这些规则显示冻肉、罐装蔬菜和啤酒之间存在多种关联。出现双向关联规则（如:

```
frozenmeal -> beer
beer -> frozenmeal
```

表明 Web 显示（只显示双向关联）可能会突出显示此数据中的一些模式。

将 Web 节点附加到“类型”节点，编辑 Web 节点，选择所有购物篮内容字段，选择**仅显示 true 标志**，然后在 Web 节点上单击“运行”。

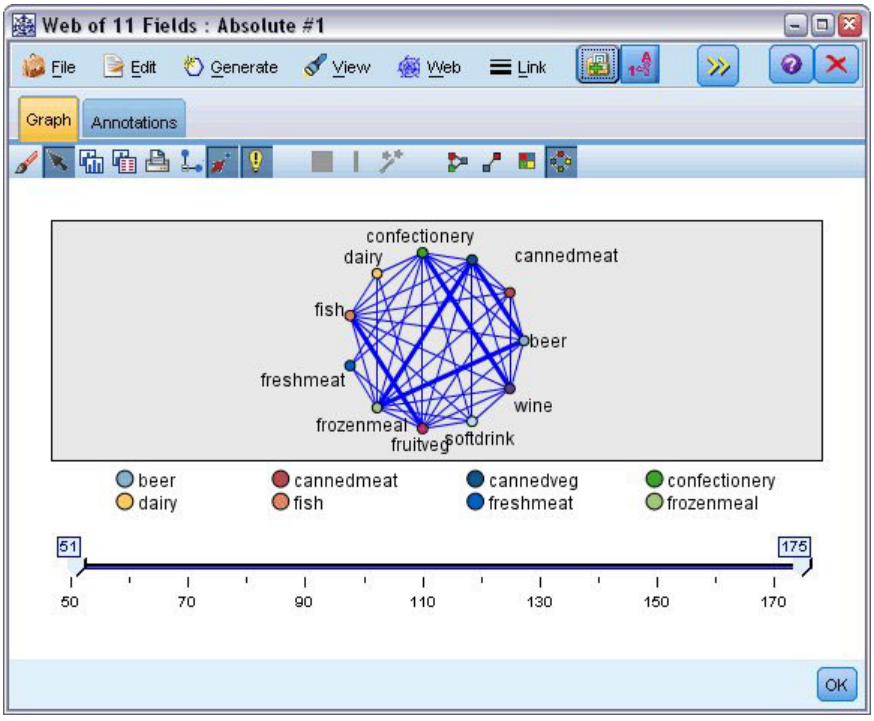


图 387. 产品关联的 Web 显示

因为大多数产品类别组合都会出现在多个购物篮中，所以此 Web 上的强链接太多，无法显示模型表示的客户群。

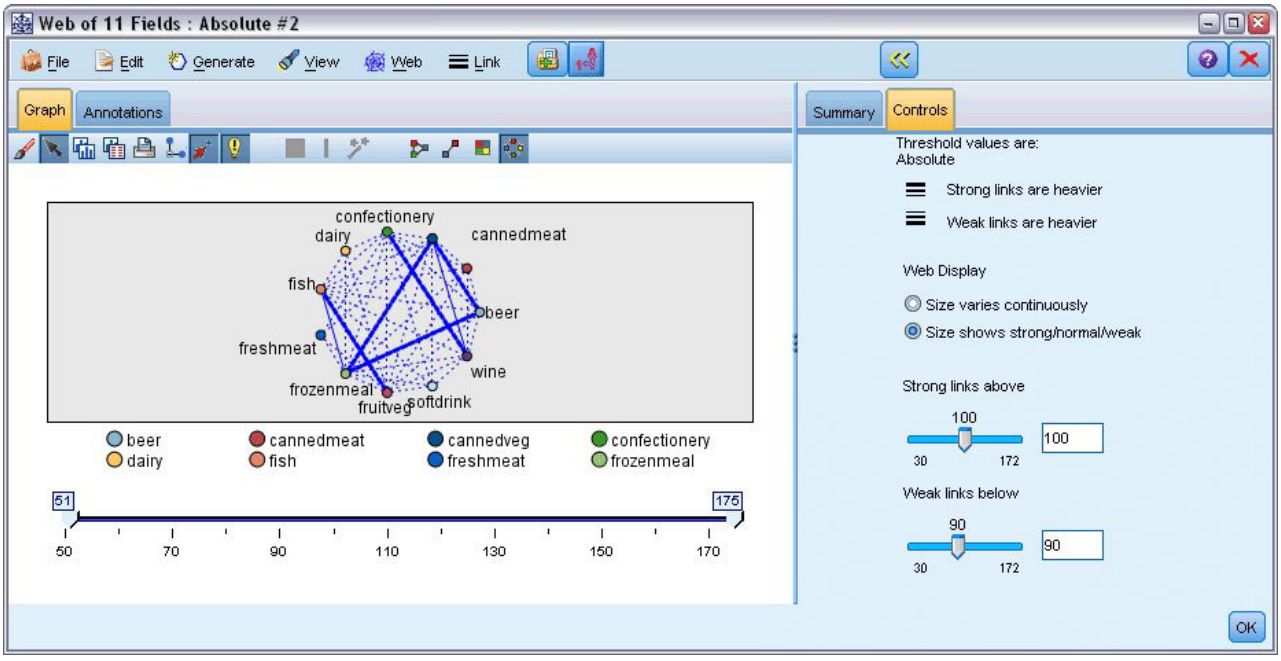


图 388. 限制性 Web 显示

1. 要指定弱连接和强连接，请单击工具栏上的黄色双箭头按钮。这会展开显示 Web 输出摘要和控件的对话框。
2. 选择 **大小表示强/正常/弱** 。
3. 将弱链接设置为低于 90。
4. 将强链接设置为高于 100。

在最终显示中，会有三个客户群突出显示:

- 购买鱼和果蔬的客户，可将这类客户称为“健康食客”
- 购买酒和粮果的客户
- 购买啤酒、冻肉和罐装蔬菜（“啤酒、豆类和比萨”）的客户

描绘客户群的特征

现在，您已根据客户购买的产品类型标识了三组客户，但是还想知道这些客户是谁，即他们的人口统计特征概况。通过为每个群中的每名客户添加标志，并使用规则归纳 (C5.0) 来基于规则描绘这些标志的特征，可以实现这一点。

首先，必须获取每个群的标志。使用刚才创建的 Web 显示可以自动生成每个群的标志。使用鼠标右键，单击 *fruitveg* 和 *fish* 之间的链接以突出显示该链接，然后右键单击并选择**为链接生成“派生”节点。**

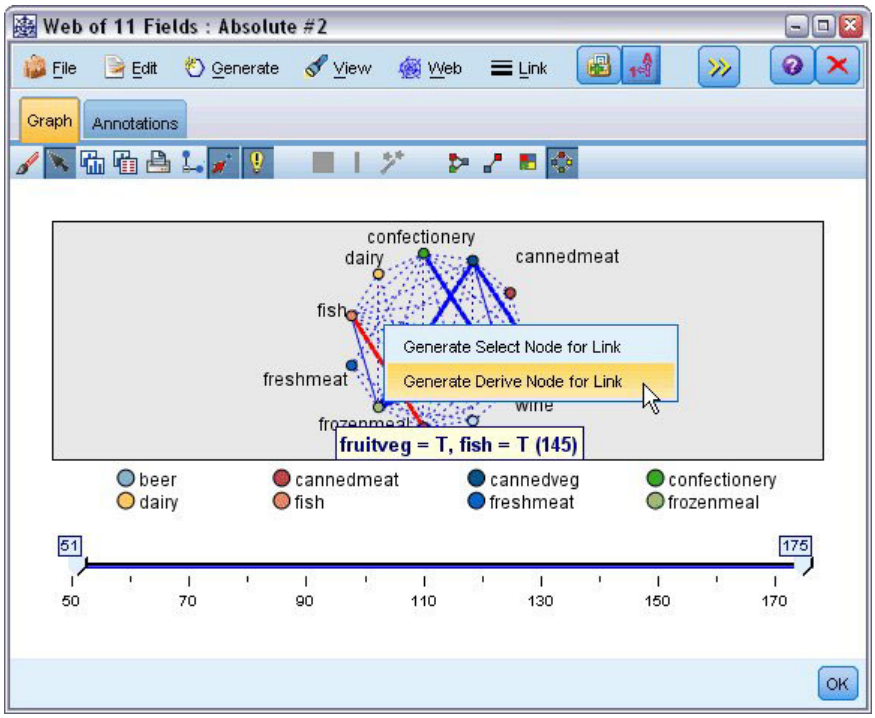


图 389. 派生每个客户组的标志

编辑最终的“派生”节点以将“派生”字段名称更改为*健康*。使用从*wine*到*confectionery*的链接重复该练习，并将最终的“派生”字段命名为 *wine_chocs*。

对于第三个组（涉及三个链接），首先要确保未选择任何链接。然后，在按住 shift 键的同时单击鼠标左键，从而选择 *cannedveg* 、 *beer* 和 *frozenmeal* 中的全部三个链接。（您一定要处于“交互”模式而不是“编辑”模式。）然后，从 Web 显示菜单中选择:

生成 > “派生”节点（“And”）

将最终“派生”字段的名称更改为 *beer_beans_pizza*。

要描述这些客户组的特征，请将现有的“类型”节点按顺序连接到这三个“派生”节点，然后附加另一个“类型”节点。在新“类型”节点中，请将除以下字段外的所有字段的角色都设置为*无*：*value*、*pmethod*、*sex*、*homeown*、*income* 和 *age*（这些字段的角色应该设置为*输入*），以及相关的客户群（例如，*beer_beans_pizza*，它们的角色应该设置为*目标*）。附加 C5.0 节点，将输出类型设置为**规则集**，然后在节点上单击“运行”。最终模型（用于 *beer_beans_pizza* ）包含此客户群的明确人口统计学特征：

```
Rule 1 for T:if sex = M
and income <= 16,900
then T
```

通过在第二个“类型”节点中选择其他客户组标志作为输出，可以对这些标志应用同一方法。通过在此上下文中使用 Apriori 代替 C5.0，可生成更多替代特征描绘；Apriori 也可用于同时描绘所有客户群标志的特征，原因是，Apriori 并非被限制到一个输出字段。

摘要

此示例说明如何使用 IBM SPSS Modeler 通过建模（使用 Apriori）和直观化（使用 Web 显示）发现数据库中的关系（即链接）。这些链接与数据中的案例组相对应，并且，通过建模（使用 C5.0 规则集）可详细研究这些组并描绘其特征。

例如，在零售领域，可能会使用这种客户组确定特殊优惠目标，以提高直接邮寄的响应率，或自定义某分部的存货产品范围以与其人口统计学基础的需求匹配。

第 28 章 评估新车辆产品 (KNN)

“最近相邻元素分析”是根据观测值与其他观测值的类似程度分类观测值的方法。在机器学习中，将其开发为识别数据模式而不需要与任何存储模式或观测值完全匹配的方法。类似观测值相互靠近，而不同观测值则相互远离。因此，两个个案之间的距离是其非相似性的度量方式。

将靠近彼此的个案称为“相邻元素”。提出新的观测值（holdout 观测值）时，将计算其到模型中每个观测值的距离。对最相似个案（即最近相邻元素）的分类进行计数，并将新的个案放入包含最多最近相邻元素的类别。

您可以指定要检查的最近相邻元素的数目；此值称为 *k*。图片显示了如何使用两个不同的 *k* 值对新的个案进行分类。*k* = 5 时，由于大多数最近相邻元素都属于类别 *1*，因此新的个案将放入类别 *1*。但是，*k* = 9 时，由于大多数最近相邻元素都属于类别 *0*，因此新的个案将放入类别 *0*。

最近相邻元素分析也可用于计算连续目标的值。在此情况下，最近相邻元素的平均值或中间目标值用于获得新观测值的预测值。

某家汽车制造商开发了两款新车（轿车和货车）的原型。在将新车型引入其产品系列前，该制造商想确定市场上哪些现有车辆与原型产品最接近，即哪些车辆是它们的“最近相邻元素”，并以此确定它们将与哪些车型展开竞争。

该制造商收集了有关现有车型的不同类别的数据，并添加了其原型产品的详细信息。需要在不同车型间进行比较的类别包括以千为单位的价格 (*price*)、发动机尺寸 (*engine_s*)、马力 (*horsepow*)、轴距 (*wheelbas*)、车宽 (*width*)、车长 (*length*)、整车重量 (*curb_wgt*)、油箱容量 (*fuel_cap*) 和燃油效率 (*mpg*)。

本示例使用了名为 *car_sales_knn.str* 的流，该流位于 *Demos* 文件夹下的 *streams* 子文件夹中。数据文件为 *car_sales_knn_mod.sav*。请参阅主题第 4 页的『Demos 文件夹』以获取更多信息。

创建流

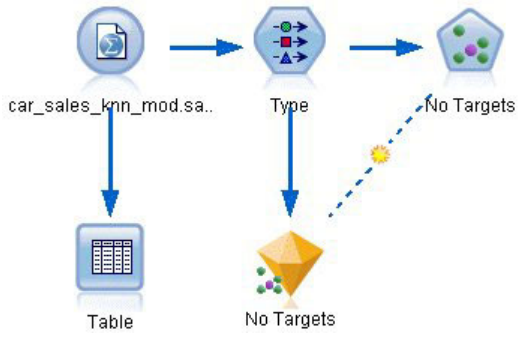


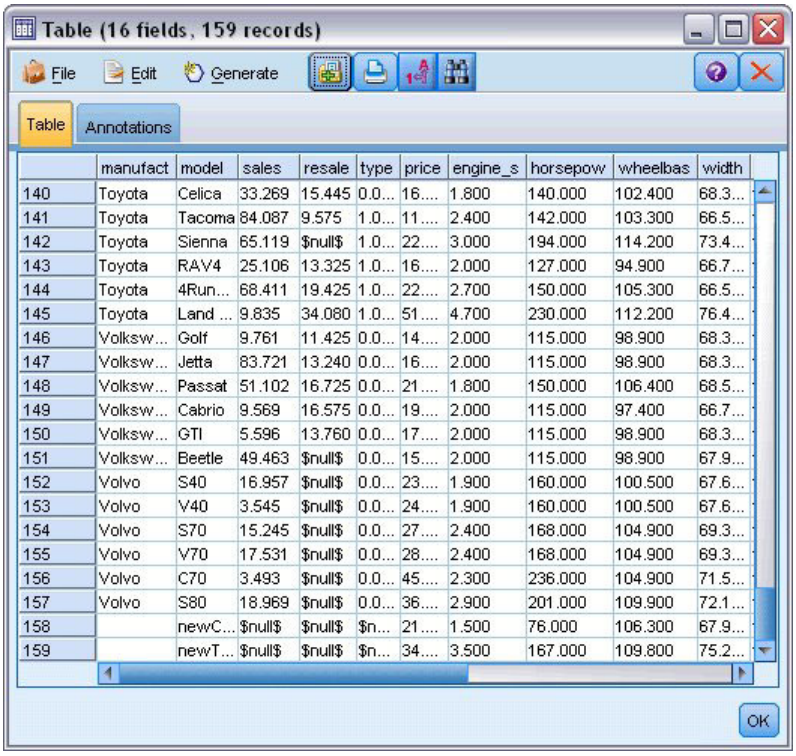
图 390. KNN 建模的样本流

创建一个新流，然后在 IBM SPSS Modeler 安装程序的 *Demos* 文件夹中添加一个指向 *car_sales_knn_mod.sav* 的“Statistics 文件”源节点。

现在，我们来看制造商收集的数据。

1. 将“表”节点附加到“Statistics 文件”源节点。

2. 打开“表”节点，然后单击**运行**。



	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...

图 391. 轿车与货车的源数据

两个原型（分别名为 *newCar* 和 *newTruck*）的详细信息，已被添加到文件末尾。

从源数据中我们可以看到，制造商使用了“货车”分类（在*类型*列中值为 1）而不是粗略地指代任何非客车车辆。

最后一列分区是必需的，以便我们在确定两个原型的最近相邻元素时可以将其指定为 holdout。在这种情况下，它们的数据不会影响计算，因为这是我们要考虑的市场的其余部分。将两个 holdout 记录的*分区*值设置为 1，同时所有其他记录在此字段上为 0 值，这允许我们稍后在设置焦点记录时使用该字段，焦点记录即我们要为其计算最近相邻元素的记录。

现在保持表输出窗口处于打开，稍后会引用它。

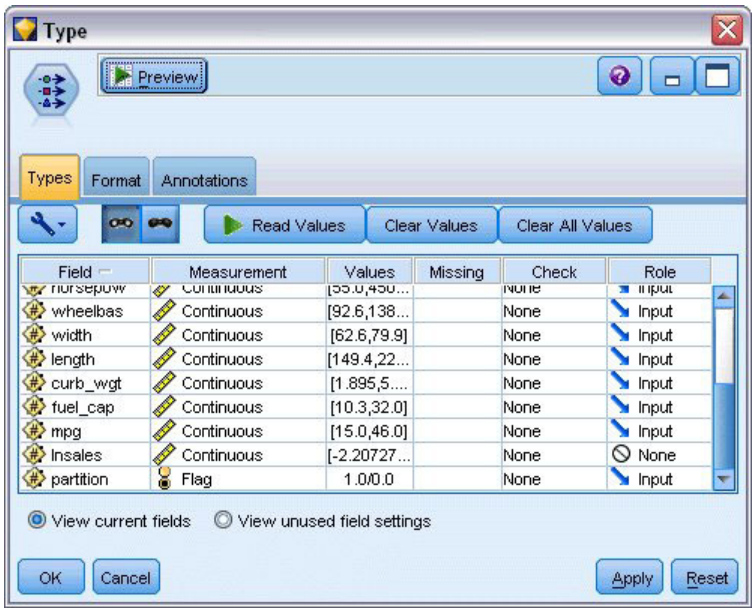


图 392. "类型"节点设置

3. 为流添加类型节点。
4. 将"类型"节点附加到"Statistics 文件"源节点。
5. 打开此"类型"节点。

   我们只想在字段 *price* 至 *mpg* 上进行比较，因此保持所有这些字段的角色设置为**输入**。
6. 将所有其他字段 (*manufact* 至 *type*，以及 *lnsales*) 的角色设置为**无**。
7. 将最后一个字段分区的测量级别设置为**标志**。确保将其角色设置为**输入**。
8. 单击**读取值**以读取数据值到流中。
9. 单击**确定**。

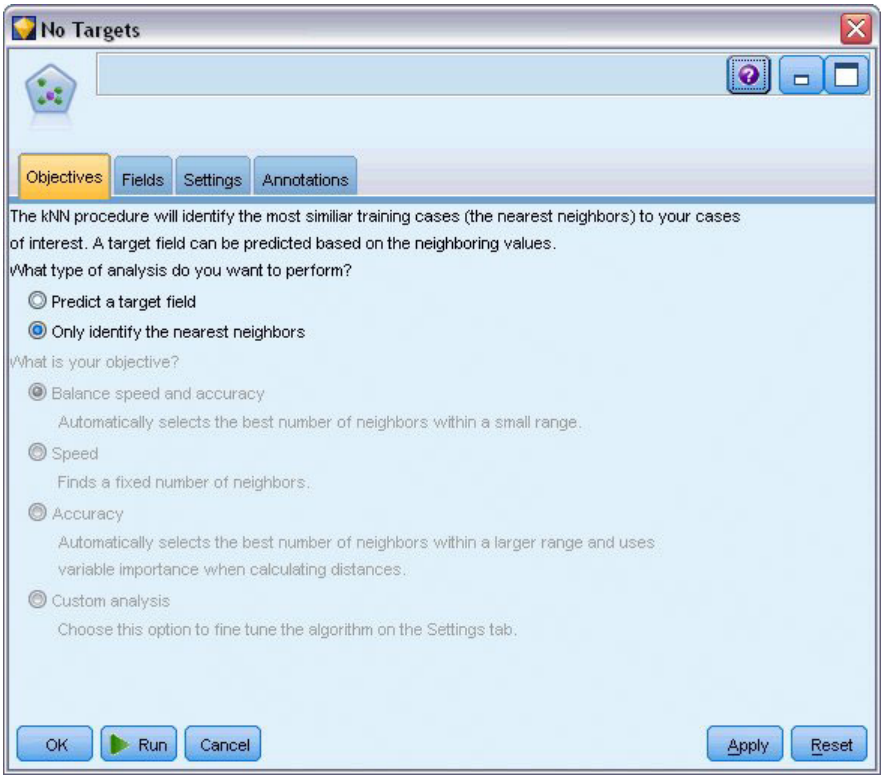


图 393. 选择识别最近相邻元素

10. 将 KNN 节点附加到“类型”节点。
11. 打开此 KNN 节点。

    由于我们只想为两个原型寻找最近相邻元素，因此这次不会预测目标字段。
12. 在**目标**选项卡上，选择**只识别最近相邻元素**。
13. 单击**设置**选项卡。

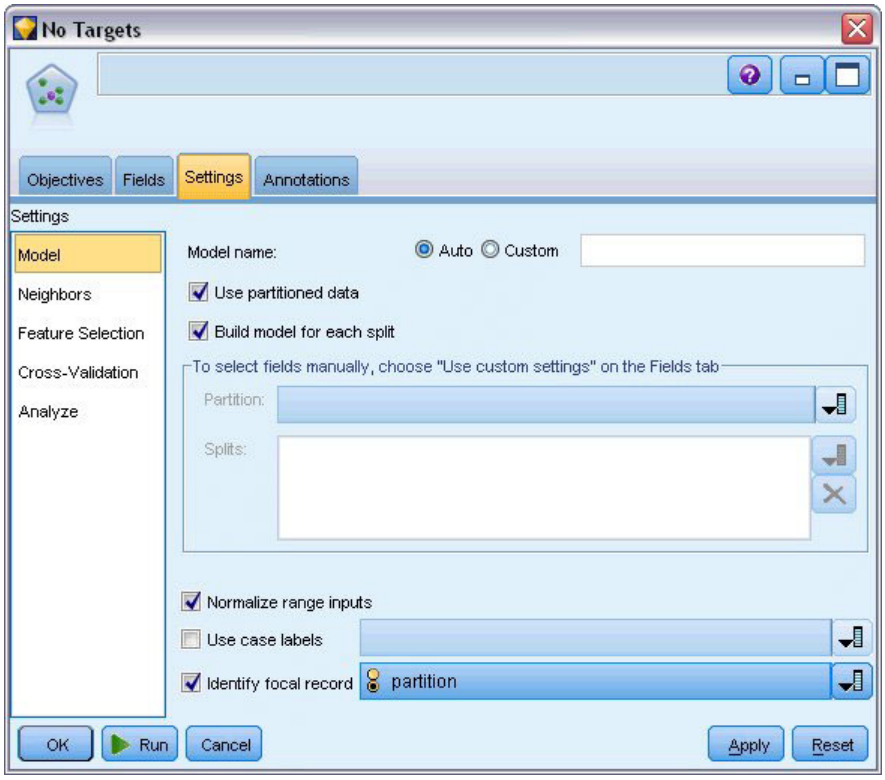


图 394. 使用分区字段识别焦点记录

现在，可以使用分区字段来识别焦点记录，即我们要为其确定最近相邻元素的记录。通过使用标志字段，可以确保此字段的值设置为 1 的记录成为焦点记录。

如您所见，该字段值为 1 的记录只有 *newCar* 和 *newTruck*，因此它们将作为我们的焦点记录。

14. 在**设置**选项卡的**模型**面板上，选中**识别焦点记录**复选框。
15. 从该字段的下拉列表中，选择**分区**。
16. 单击**运行**按钮。

检查输出

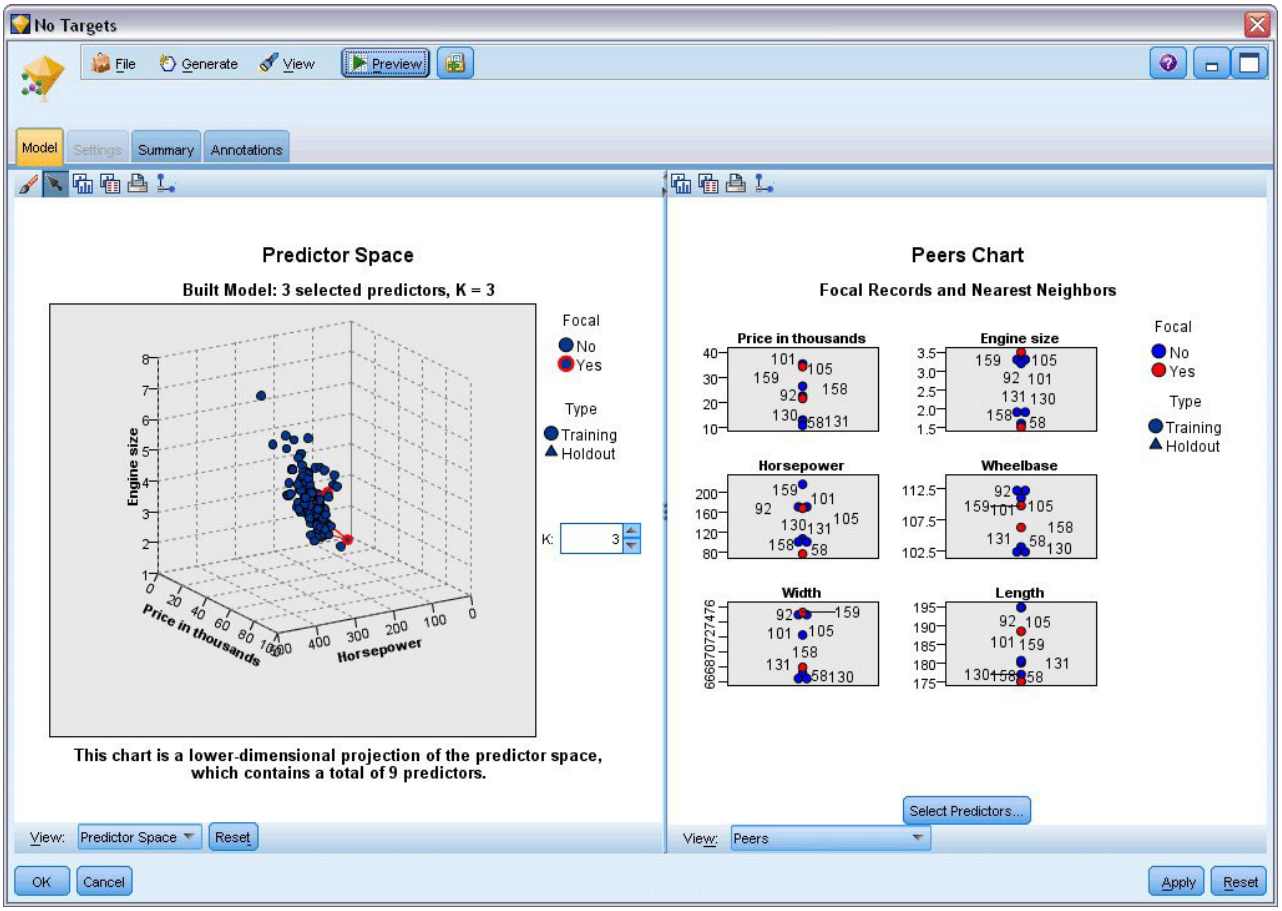


图 395. "模型查看器"窗口

已在流画布和"模型"选用板中创建模型块。打开任一模型块即可显示"模型查看器"，它包含一个双面板窗口：

- 第一个面板显示关于模型的概述，称为主视图。"最近相邻元素"模型的主视图称为**预测变量空间**。
- 第二个面板显示两种视图类型之一：

  辅助模型视图显示有关此模型的更多信息，但不侧重于模型本身。

  链接视图是在您深入查看主视图的某个部分时显示有关此模型某个特征的详细信息的视图。

预测变量空间

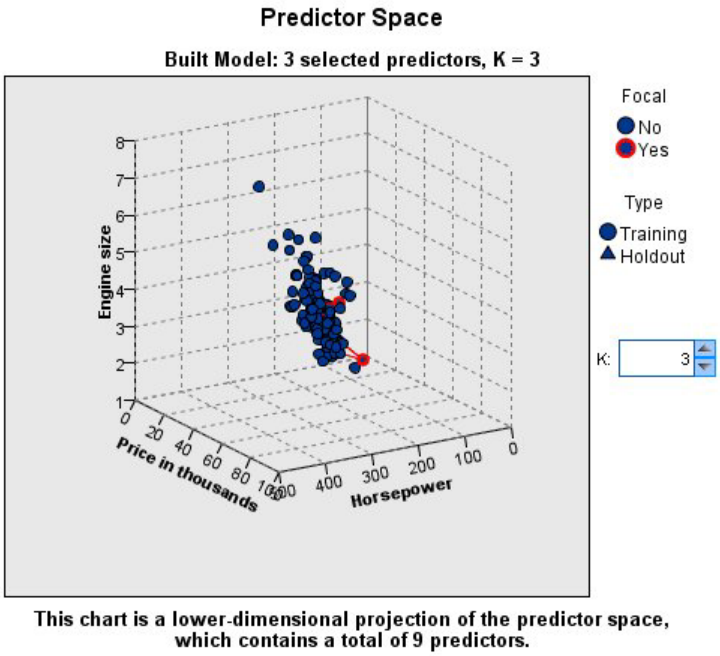


图 396. 预测变量空间图表

预测变量空间图表为交互式三维图形，它绘制了三个特征（实际为源数据的前三个输入字段）的数据点，分别表示价格、发动机尺寸和马力。

两个焦点记录突出显示为红色，通过线条连接到其 *k* 个最近相邻元素。

通过单击并拖动该图表，您可以对其进行旋转，从而更好地了解预测变量空间中数据点的分布。单击**重置**按钮，将其恢复到缺省视图。

对等图

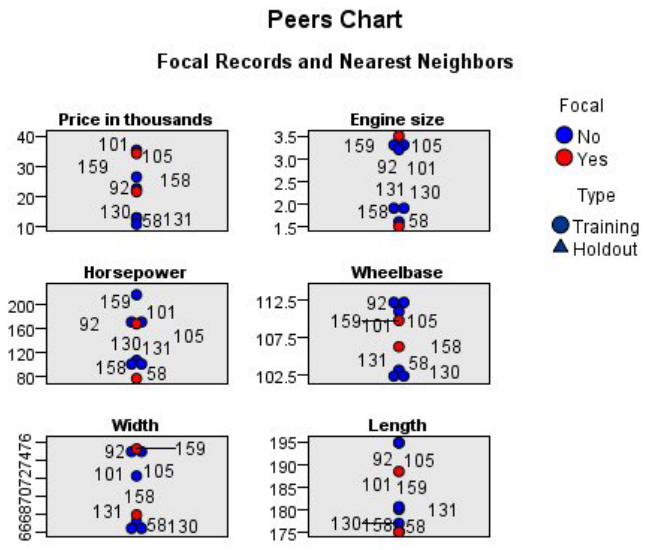


图 397. 对等图表

缺省辅助视图为对等图表，其中突出显示在预测变量空间中选中的两个焦点记录，及其在六个特征（源数据的前六个输入字段）上的 *k* 个最近相邻元素。

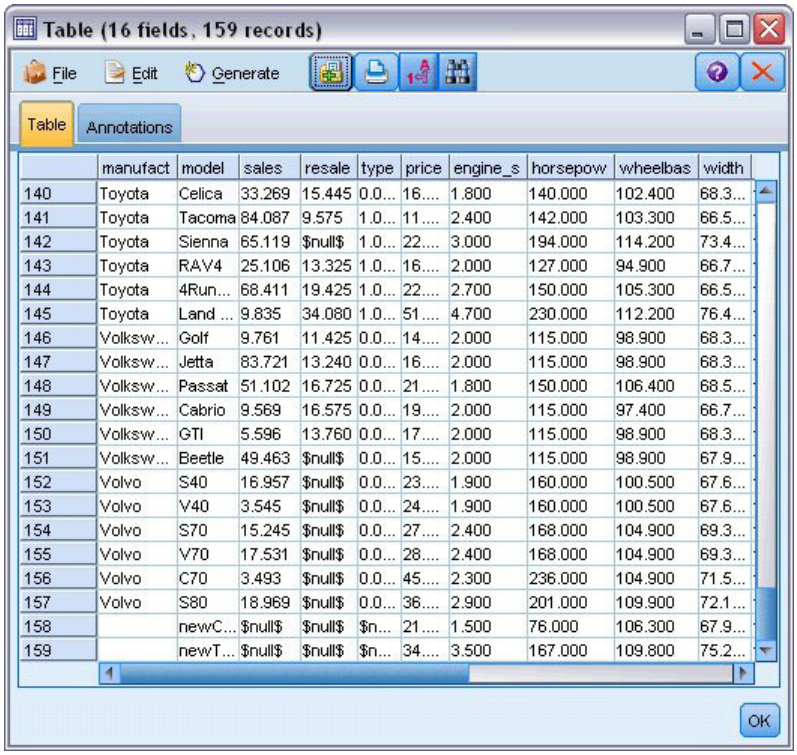
车辆通过它们在源数据中的记录号来表示。我们需要从“表”节点获得输出以帮助识别它们。

如果“表”节点输出仍然可用:

1. 单击位于主 IBM SPSS Modeler 窗口右上角的管理器窗格的**输出**选项卡。
2. 双击条目**表（16 个字段，159 个记录）**。

   如果表输出不再可用:

3. 在主 IBM SPSS Modeler 窗口上，打开“表”节点。
4. 单击**运行**。



	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...

图 398. 按记录号标识记录

向下滚动到表的底部，可以看到 *newCar* 和 *newTruck* 为数据的最后两条记录，编号分别为 158 和 159。

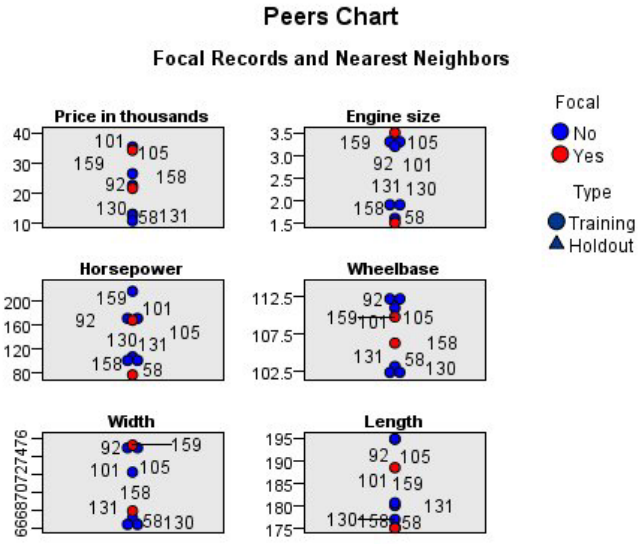


图 399. 比较对等图表上的特征

例如，在这里，可以从对等图表上看到 *newTruck* (159) 的发动机尺寸大于它的任何最近相邻元素，而 *newCar* (158) 的发动机尺寸则小于它的任何最近相邻元素。

对于六个特征中的每个特征，可以将鼠标移至单独点上，查看特定个案的每个特征的实际值。

但哪些车辆是 *newCar* 和 *newTruck* 的最近相邻元素呢？

对等图表稍微有点拥挤，现在我们转换到较简单的视图。

5. 单击位于对等图表底部的**视图**下拉列表（当前名为**对等**的条目）。
6. 选择**邻元素和距离表**。

相邻元素和距离表

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

图 400. 相邻元素和距离表

这样较好。现在可以看到在市场上与两种产品原型最接近的三种车型。

对于 *newCar*（焦点记录 158），它们是 Saturn SC (131)、Saturn SL (130) 和 Honda Civic (58)。

很正常，它们均为中型轿车，因此 *newCar* 的市场定位不错，特别是它具有优秀的燃油效率。

对于 *newTruck*（焦点记录 159），最近相邻元素为 Nissan Quest (105)、Mercury Villager (92) 和 Mercedes M-Class (101)。

如前所述，它们并不一定是传统意义上的货车，而只是分类为非客车的车辆。来看最近相邻元素的“表”节点输出，可以看到 *newTruck* 相对较为昂贵，并且是同类中最重的车辆。不过，其燃油效率优于最接近的对手，因此值得青睐。

摘要

我们已了解如何使用最近相邻元素分析来比较来自特定数据集的个案的广泛特征集。我们还为两个明显不同的保留记录计算了个案，它们最接近地呈现了这些 holdout。

声明

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务，则由用户自行负责。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并未授予用户使用这些专利的任何许可。您可以用书面方式将许可查询寄往:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

本条款不适用英国或任何这样的条款与当地法律不一致的国家或地区: International Business Machines Corporation“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。因此本条款可能不适用于您。

本信息中可能包含技术方面不够准确的地方或印刷错误。此处的信息将定期更改；这些更改将编入本资料的新版本中。IBM 可以随时对本出版物中描述的产品进行改进和/或更改，而不另行通知。

本信息中对非 IBM Web 站点的任何引用都只是为了方便起见才提供的，不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的许可证持有者如果要了解有关程序的信息以达到如下目的: (i) 允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及 (ii) 允许对已经交换的信息进行相互使用，请与下列地址联系:

IBM Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

此处包含的任何性能数据都是在受控环境中测得的。因此，在其他操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。此外，有些测量是通过推算而估计的，实际结果可能会有差异。本文档的用户应当验证其特定环境的适用数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

如果您正在查看本信息的软拷贝，图片和彩色图例可能无法显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp. 在全球许多行政管辖地区的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。Web 页面“Copyright and trademark information” (www.ibm.com/legal/copytrade.shtml) 提供了 IBM 商标的最新列表。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和/或其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 和/或其子公司的商标或注册商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

索引

S

T

V

W

Printed in China