IBM SPSS Modeler 18.5 建模节点

注

在使用本资料及其支持的产品之前，请阅读第 289 页的『注意事项』中的信息。

产品信息

本版本适用于的版本 18、发行版 4、IBM® SPSS Modeler 的修订 0 以及所有后续版本和修改，除非在新版本中另有说明

内容

前言

IBM SPSS Modeler 是 IBM Corp. 企业级数据挖掘工作平台。 SPSS Modeler 通过深度的数据分析帮助组织改进与客户和市民的关系。 组织通过借助源自 SPSS Modeler 的洞察力可以留住优质客户，识别交叉销售机遇，吸引新客户，检测欺诈，降低风险，促进政府服务交付。

SPSS Modeler 的可视化界面让用户可以应用他们自己的业务专长，这将生成更加强有力的预测模型，缩减实现解决方案所需时间。 SPSS Modeler 提供了多种建模技术，例如预测、分类、分割和关联检测算法。 模型创建成功后，通过 IBM SPSS Modeler Solution Publisher，在广泛的企业内交付给决策者，或通过数据库交付。

关于 IBM Business Analytics

IBM Business Analytics 软件提供完整、一致和正确的信息，决策人依据此信息来提高业务性能。 企业智能、预测分析、财务业绩和战略管理的完整产品组合，和分析应用程序一起提供对当前业绩的清晰、直接和实用的洞察力，以及预测未来结果的能力。 结合丰富的行业解决方案，久经证明的实践和专业服务以及各种规模的组织都能够实现最高生产力、确信地自动作出决策以及获取更好的结果。

作为此产品服务组合的组成部分，IBM SPSS Predictive Analytics 软件可帮助组织预测未来事件，并在该洞察的基础上提前行动以实现更好的业务结果。 减少欺诈和降低风险时，全球的商业、政府和学术客户都依赖 IBM SPSS 技术作为吸引、保留和增加客户的竞争优势。 通过在日常运营中融入 IBM SPSS 软件，组织将成为预测型企业 - 即可以指引并实现决策的自动化，以满足企业目标并实现可衡量的竞争优势。 有关详细信息或要联系一位代表，请访问 http://www.ibm.com/spss。

技术支持

技术支持可供维护客户使用。 客户可就 IBM Corp. 产品使用问题或某一受支持硬件环境的安装帮助寻求技术支持。 要寻求技术支持，请访问 IBM Corp. 网站 http://www.ibm.com/support。 请求帮助时，请准备好标识您自身、组织和支持协议。

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 提供了各种借助机器学习、人工智能和统计学的建模方法。 通过建模选项板中的方法，您可以根据数据生成新的信息以及开发预测模型。 每种方法各有所长，同时适用于解决特定类型的问题。

SPSS Modeler 可以作为独立产品购买，也可以作为客户端与 SPSS Modeler Server 一起使用。 同时提供了大量其他选项，以下各节将对这些选项进行概述。 有关更多信息，请参阅 https://www.ibm.com/analytics/us/en/technology/spss/。

IBM SPSS Modeler 产品

IBM SPSS Modeler 系列产品及关联的软件包括以下各项。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console（包含在 IBM SPSS Deployment Manager 中）
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS 协作和部署服务 的 IBM SPSS Modeler Server 适配器

IBM SPSS Modeler

SPSS Modeler 是具有完整功能的产品，它安装并运行于个人计算机上。 您可以在本地方式作为独立产品运行 SPSS Modeler，也可以在分布方式下将其与 IBM SPSS Modeler Server 一起使用来提高大型数据集的性能。

借助 SPSS Modeler，您可以快速直接地构建准确的预测模型，而不进行编程。 通过使用唯一可视界面，您可以轻松地查看数据挖掘过程。 借助该产品随附的高级分析支持，您可以发现数据中先前隐藏的模式和趋势。 您可以构建结果模型并了解影响结果的因素，从而利用业务机会并降低风险。

SPSS Modeler 推出了两个版本：SPSS Modeler Professional 和 SPSS Modeler Premium。 有关更多信息，请参阅主题 第 2 页的『IBM SPSS Modeler 版本』。

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 是一个单独授权的产品，在分布分析方式下，该产品在安装了一个或多个 IBM SPSS Modeler 的服务器主机上持续运行。 这种运行方式大大提高了 SPSS Modeler Server 对大型数据集的处理速度，因为在服务器上可以运行耗用内存的操作，并且无需将数据下载到客户端计算机上。 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带来更多优势。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是一个图形用户界面，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。 控制台包含在 IBM SPSS Deployment Manager，可以用于监视和配置 SPSS Modeler Server 安装，并且可供当前 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。 通过这种方式，您可以发布和部署完整的 SPSS Modeler 流以用于未安装 SPSS Modeler 的环境。 SPSS Modeler Solution Publisher 作为 IBM SPSS 协作和部署服务-评分 服务的组成部分分发，需要单独的许可证。 通过此许可证，您可以接收 SPSS Modeler Solution Publisher Runtime，它使您能够执行已发布的流。

有关 SPSS Modeler Solution Publisher 的更多信息，请参阅 IBM SPSS 协作和部署服务 文档。 IBM SPSS 协作和部署服务 IBM 文档包含名为“IBM SPSS Modeler Solution Publisher”和“IBM SPSS Analytics Toolkit”的部分。

IBM SPSS 协作和部署服务 的 IBM SPSS Modeler Server 适配器

IBM SPSS 协作和部署服务 的一些适配器使 SPSS Modeler 和 SPSS Modeler Server 能够与 IBM SPSS 协作和部署服务 存储库进行交互。 通过这种方式，部署到存储库的 SPSS Modeler 流可以由多个用户共享，或者从瘦客户端应用程序 IBM SPSS Modeler Advantage 进行访问。 请将适配器安装在托管存储库的系统上。

IBM SPSS Modeler 版本

SPSS Modeler 推出了下列版本。

SPSS Modeler Professional

SPSS Modeler Professional 提供处理大多数类型的结构化数据所需要的所有工具，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。

SPSS Modeler Premium

SPSS Modeler Premium 是一个单独授权的产品，它对 SPSS Modeler Professional 进行了扩展，以便后者能够处理专门的数据和非结构化文本数据。 SPSS Modeler Premium 包含 IBM SPSS Modeler 文本分析：

IBM SPSS Modeler 文本分析 采用先进语言技术和自然语言处理 (NLP)，以快速处理大量非结构化文本数据，提取和组织关键概念，以及将这些概念分为各种类别。 提取的概念和类别可以与现有的结构化数据（例如人口统计信息）相结合，并且可借助 IBM SPSS Modeler 的全套数据挖掘工具进行建模，以此实现更好更集中的决策。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription 提供与传统 IBM SPSS Modeler 客户端完全相同的预测性分析功能。 通过 Subscription 版本，您可以定期下载产品更新。

文档

可从 SPSS Modeler 中的**帮助**菜单获取文档。 这样会打开始终可在产品外部访问的在线 IBM 文档。

每个产品的完整文档（包括安装指示信息）也在以下位置以 PDF 格式提供：https://www.ibm.com/support/pages/spss-modeler-185-documentation。

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**。 对于使用 SPSS Modeler 的一般简介，包括如何构建数据流、处理缺失值、构建 CLEM 表达式处理项目和报告，以及将用于部署的流打包到 IBM SPSS 协作和部署服务 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**。 描述所有用于创建数据挖掘模型的节点。 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 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。 还可以从“帮助”菜单获取本指南的联机版本。 有关更多信息，请参阅主题 第 3 页的『应用程序示例』。
- **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 Deployment Manager 下以处理作业的步骤形式运行 IBM SPSS Modeler 流的信息。
- **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 管理和性能指南**。 提供有关如何配置和管理 IBM SPSS Modeler Server 的信息。
- **IBM SPSS Deployment Manager 用户指南**。 有关使用 Deployment Manager 应用程序中包含的管理控制台用户界面来监视和配置 IBM SPSS Modeler Server 的信息。
- **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。

- **SPSS Modeler 文本分析 用户指南**。 提供有关将文本分析与 SPSS Modeler 配合使用的信息，包括文本挖掘节点、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小得多，但涉及的概念和方法可扩展到实际应用程序。

要访问示例，请在 SPSS Modeler 中单击“帮助”菜单中的**应用程序示例**。

数据文件和样本流安装在产品安装目录下的 Demos 文件夹中。 有关更多信息，请参阅 第 3 页的『Demos 文件夹』。

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* 脚本编写与自动化指南 中的示例。

Demos 文件夹

与应用程序示例配合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（例如：`C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`）。也可以从 Windows"开始"菜单上的 IBM SPSS Modeler 程序组访问此文件夹，或者通过单击**文件** > **打开流**对话框中最近的目录列表中的 Demos 来进行访问。

许可证跟踪

当您使用 SPSS Modeler 时，系统会定期跟踪并记录许可证使用情况。 所记录的许可证度量为 *AUTHORIZED_USER* 和 *CONCURRENT_USER*，并且记录的度量类型取决于您针对 SPSS Modeler 具有的许可证类型。

产生的日志文件可由 IBM License Metric Tool 处理，通过该工具可生成许可证使用情况报告。

许可证日志文件是在记录 SPSS Modeler Client 日志文件的同一目录中创建的（缺省情况下，为 `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`）。

第 2 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。

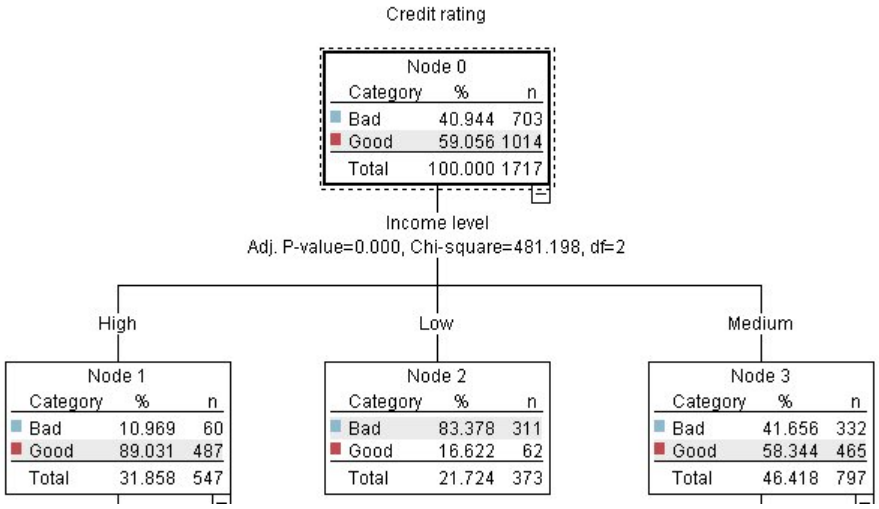


图 *1:* 简单的决策树模型

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	年龄
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*。 有关更多信息，请参阅主题 第 3 页的『Demos 文件夹』。

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

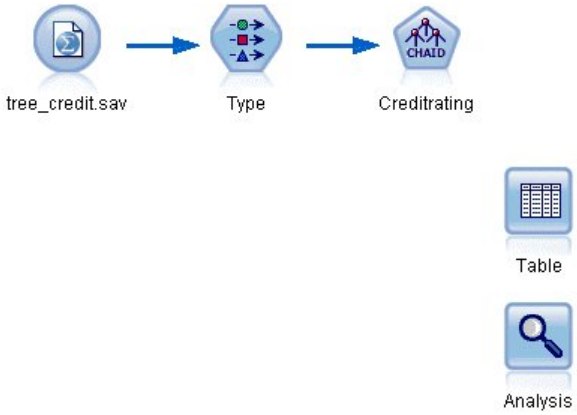


图 *2:* 建模流

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拆分位置来构建决策树。

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。）

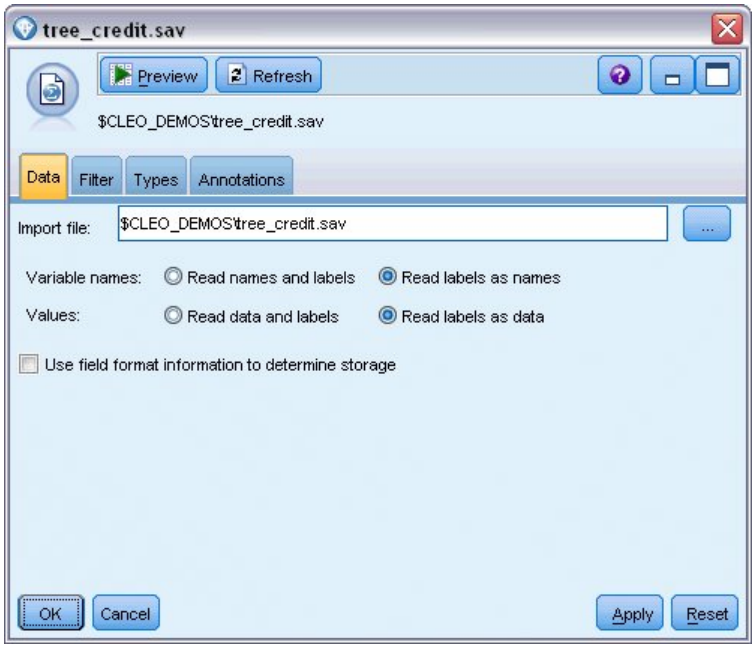


图 *3:* 使用*"Statistics* 文件*"*源节点读取数据

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。

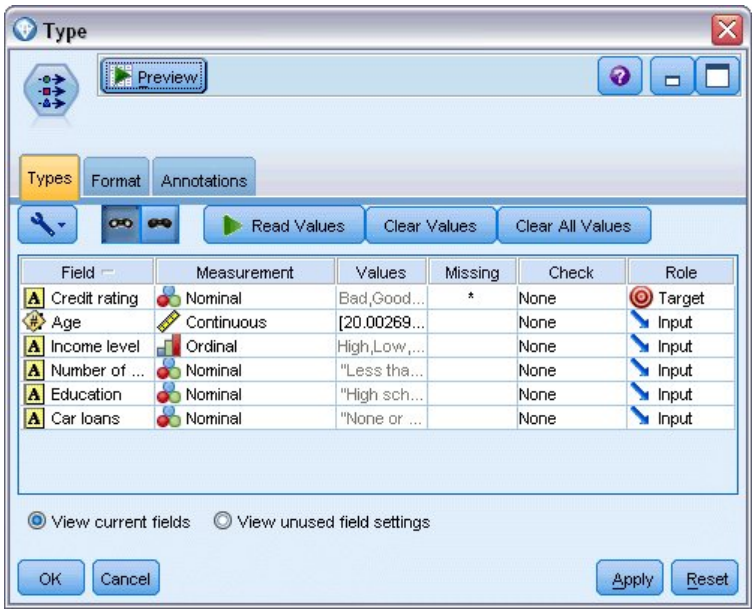


图 *4:* 使用*"*类型*"*节点设置目标和输入字段

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。

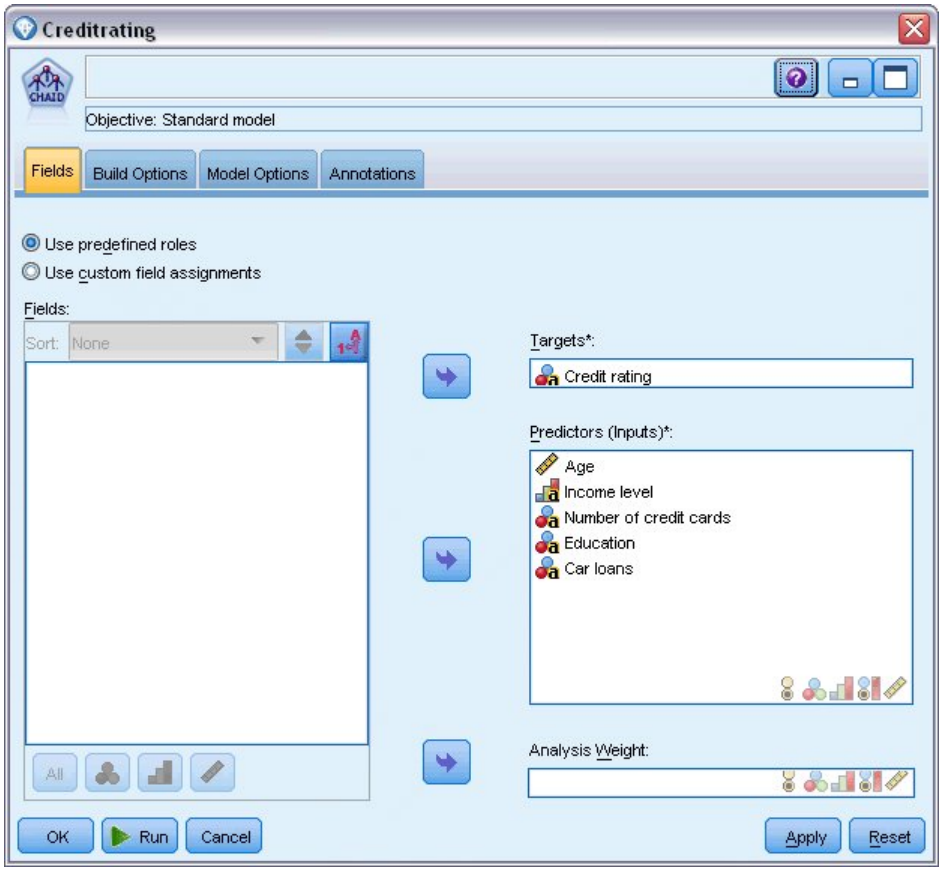


图 *5: CHAID* 建模节点，“字段”选项卡

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。

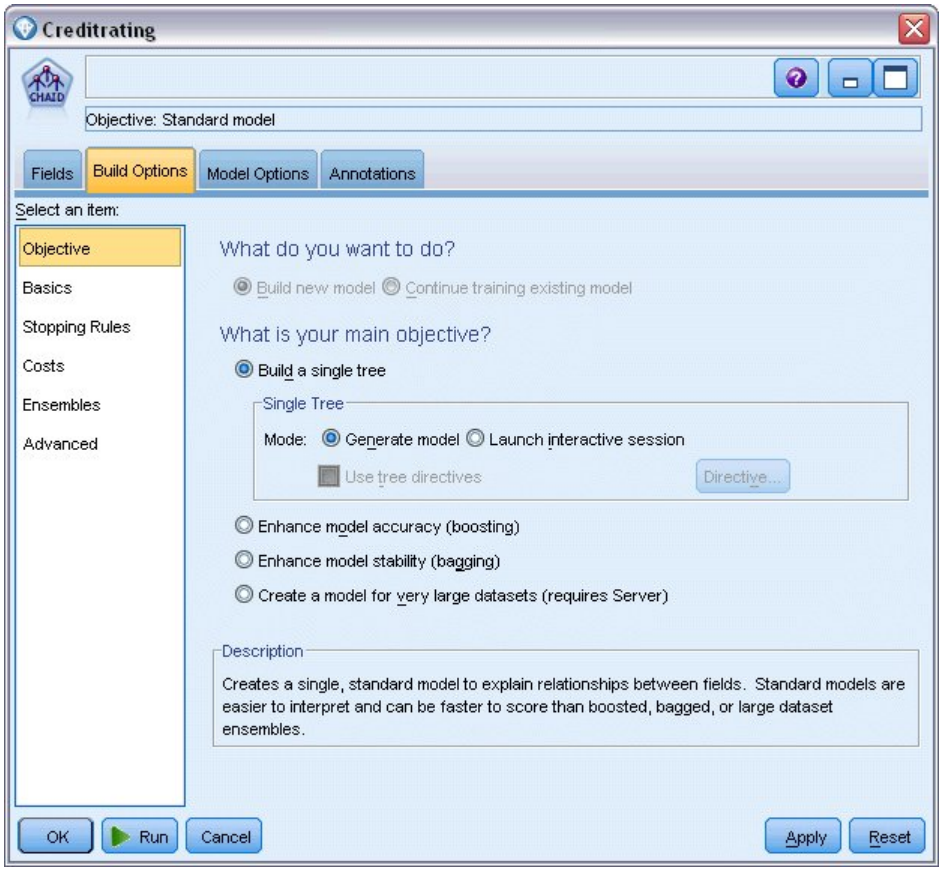


图 *6: CHAID* 建模节点，“构建选项”选项卡

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。

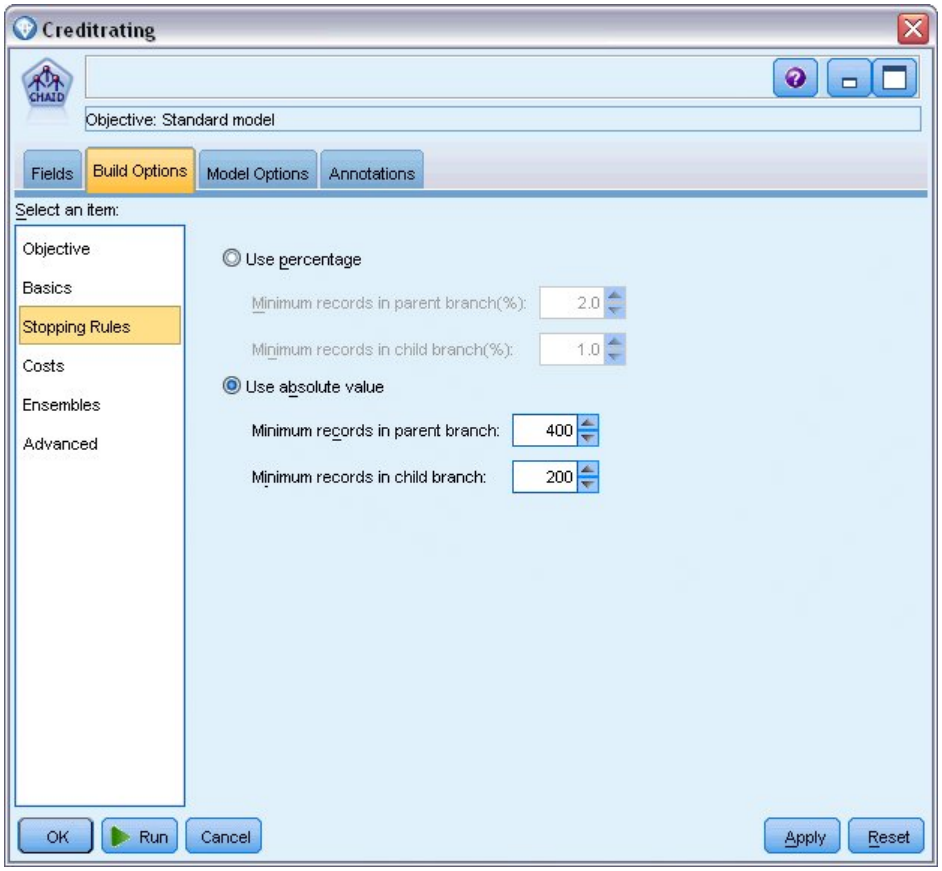


图 *7:* 设置用于决策树构建的中止条件

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）。

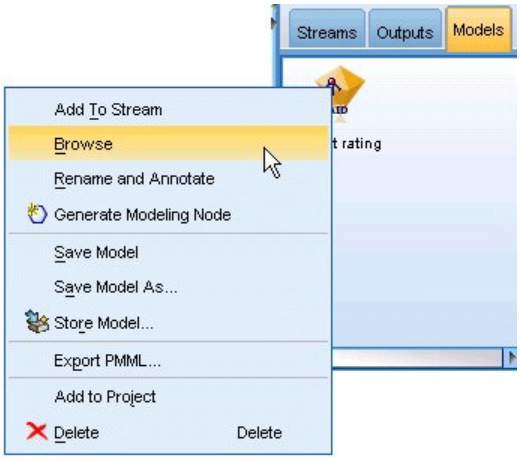


图 *8:* “模型”选用板

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。

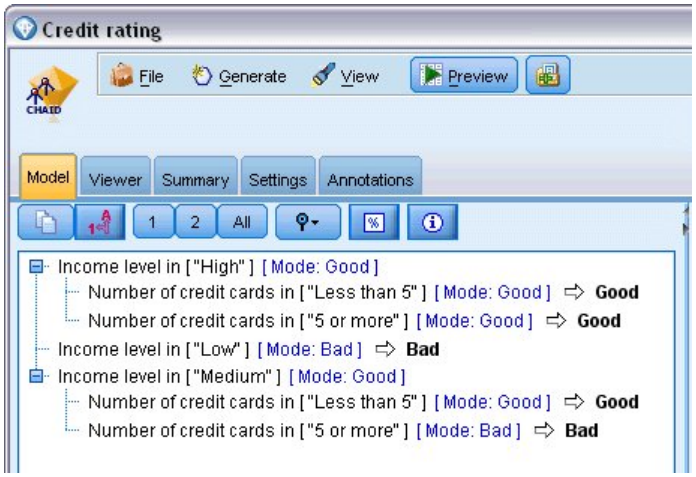


图 *9: CHAID* 模型块，规则集

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。

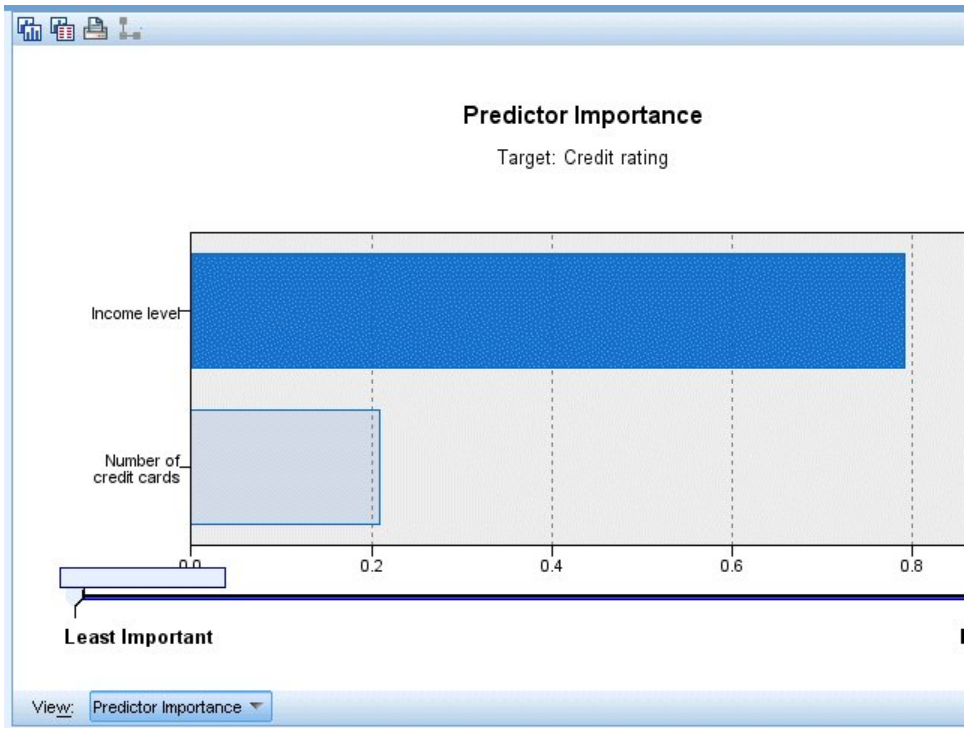


图 *10:* 预测变量重要性图表

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。

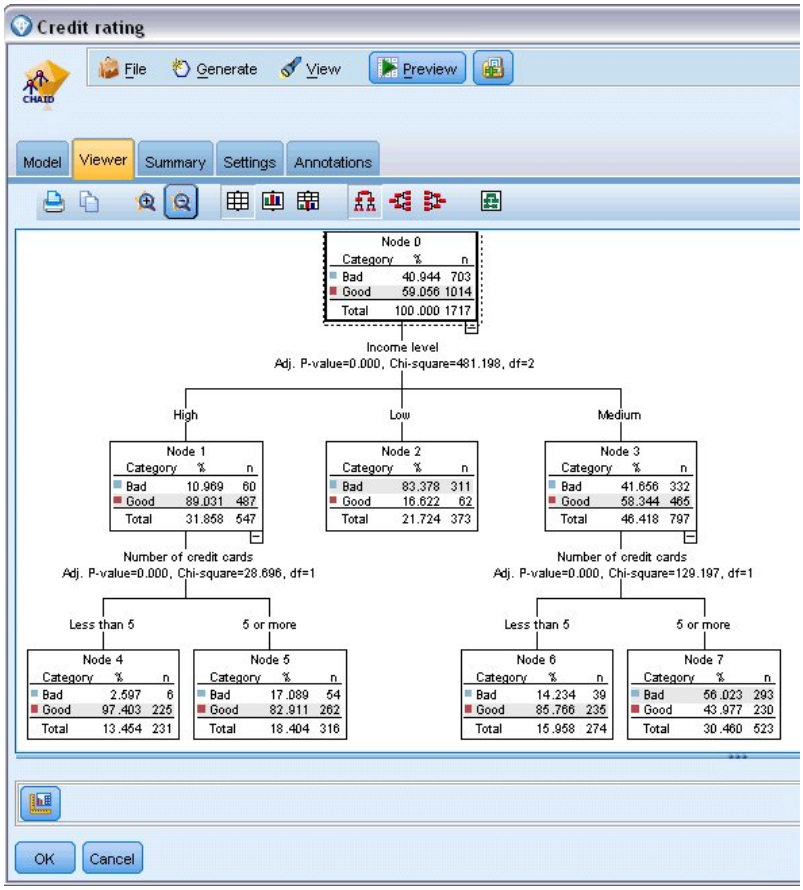


图 *11:* 模型块中的查看器选项卡，已选择缩小

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。

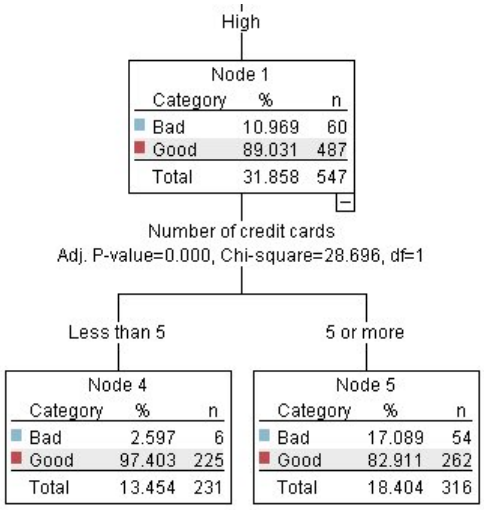


图 *12:* 高收入客户的树形视图

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。

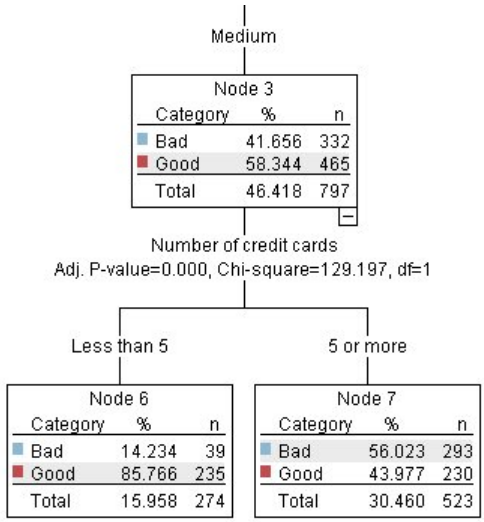


图 *13:* 中等收入客户的树形视图

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。

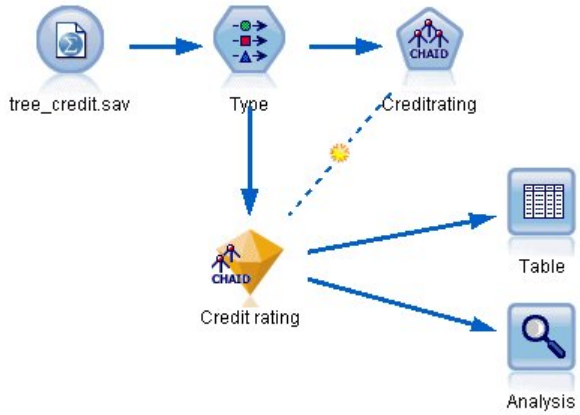
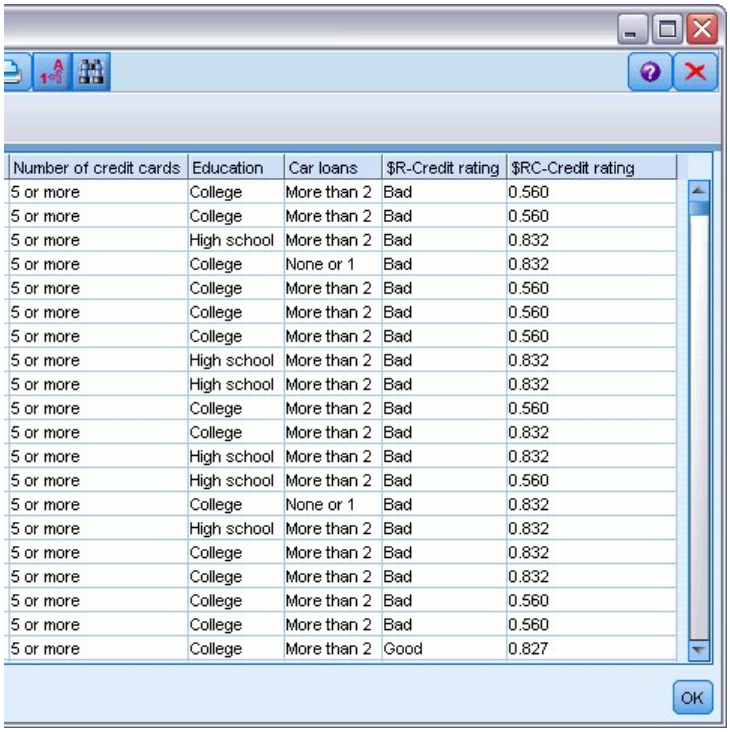


图 *14:* 将模型块附加到输出节点以进行模型评估

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

图 *15:* 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**。

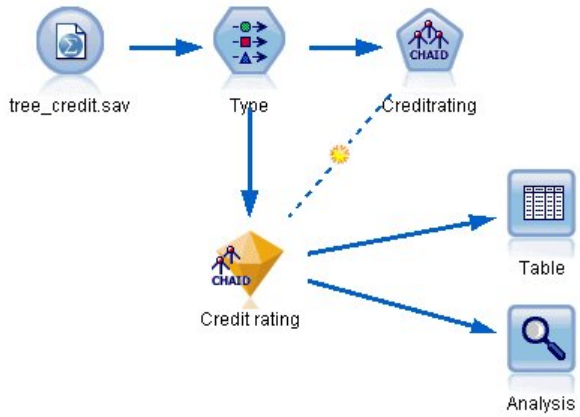


图 *16:* 附加“分析”节点

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。

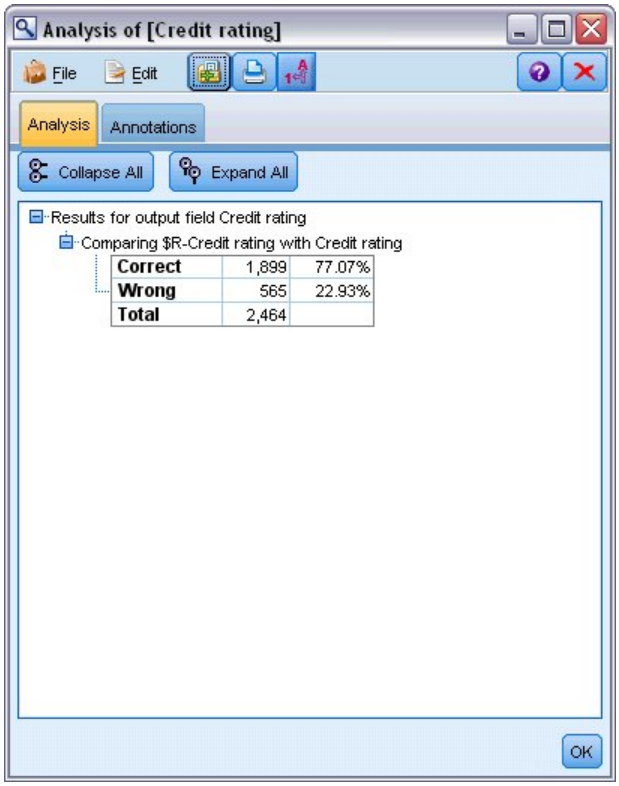


图 *17:* 观察到的响应与预测响应的比较分析结果

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。

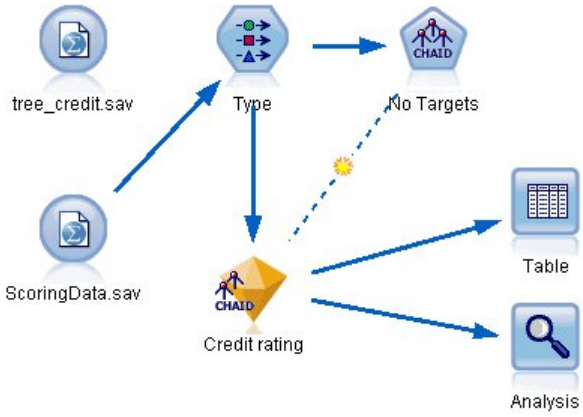


图 *18:* 附加用于评分的新数据

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 协作和部署服务 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。

目录

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 3 章 建模概述

建模节点概述

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。 通过建模选项板中的方法，您可以根据数据生成新的信息以及开发预测模型。 每种方法各有所长，同时适用于解决特定类型的问题。

IBM SPSS Modeler 应用程序指南 为上述多种方法提供了示例以及建模过程的一般介绍。 本指南作为联机教程提供，也有 PDF 格式。 请参阅主题第 3 页的『应用程序示例』，以获取更多信息。

建模方法分为以下类别：

- 受监督
- 关联
- 细分

受监督模型

受监督模型使用一个或多个**输入**字段的值来预测一个或多个输出（或**目标**）字段的值。 这些技术的一些示例包括：决策树（C&R 树、QUEST、CHAID 和 C5.0 算法）、回归（线性、logistic、广义线性和 Cox 回归算法）、神经网络、支持向量机和贝叶斯网络。

“受监督”模型可帮助组织预测已知的结果，例如顾客是否购买、流失或某交易是否符合某种已知的犯罪模式。 其建模方法包括机器学习、规则归纳、子组标识、统计方法和多模型生成。

受监督节点

“自动分类器”节点用于创建和对比二元结果（是或否，流失或不流失等）的若干不同模型，使用户可以选择给定分析的最佳处理方法。 由于支持 多种建模算法，因此可以对用户希望使用的方法、每种方法的特定选项以及对比结果的标准进行选择。 该节点根据指定的选项生成一组模型，并根据您指定的条件对最佳候选项进行排序。

自动数字节点使用多种不同方法估计和对比模型的连续数字范围结果。 此节点和自动分类器节点的工作方式相同，因此可以选择要使用和要在单个建模传递中使用多个选项组合进行测试的算法。 受支持的算法包括神经网络、C&R 树、CHAID、线性回归、广义线性回归以及支持向量机 (SVM)。 可以根据相关度、相对误差或使用的变量数来比较模型。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。 该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果树中某个节点中 100% 的观测值都属于目标字段的一个特定类别，那么该节点将被认定为“纯洁”。 目标和输入字段可以是数字范围或分类（名义、有序或标志）； 所有拆分都是二进制的（只有两个子组）。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。 输入字段可以是数字范围（连续），但目标字段必须是分类。 所有分割都是二元的。

CHAID 使用卡方统计来生成决策树，以确定最佳的分割。 CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。 目标和输入字段可以是数字范围（连续）或分类。 穷举 CHAID 是 CHAID 的修正版，它可以更彻底地检查所有可能的拆分，但计算时间较长。

C5.0 节点构建决策树或规则 集。 该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。 目标字段必须为分类字段。 允许进行多次多于两个子组的分割。

决策列表节点可标识子组或段，显示与总体相关的给定二元结果的似然度的高低。 例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。 通过定制段和并排预览备选模型来比较结果，您可以将自己的业务知识体现在模型中。 决策列表模型由一组规则构成，其中每个规则具备一个条件和一个结果。 规则依顺序应用，相匹配的第一个规则将决定结果。

线性回归模型根据目标与一个或多个预测变量之间的线性关系来预测连续目标。

“PCA/因子”节点提供用于降低数据复杂程度的强大数据降维技术。 主成份分析（PCA）可找出输入字段的线性组合，该组合最好地捕获了整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。 因子分析则尝试识别底层因素，这些因素说明了观测的字段集合内的相关性模式。 对于这两种方法，其共同的目标是找到可对原始字段集合中的信息进行有效总结的少量派生字段。

“特征选择”节点根据一组条件（例如缺失值百分比）筛选要移除的输入字段，然后，相对于指定目标对余下的输入的重要性进行排秩。 例如，假如某个给定数据集有上千个潜在输入，那么哪些输入最有可能用于对患者结果进行建模呢?

判别分析提出比 Logistic 回归更加严格的假设，但是在满足这些假设时可成为 Logistic 回归的有价值替代方案或补充。

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。 它与线性回归类似，但采用分类目标字段而不是数字范围。

广义线性模型对广义线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。 而且，该模型还允许因变量呈非正态分布。 它涵盖了大量统计模型的功能，包括线性回归、逻辑回归、计数数据的对数线性模型和区间删失生存模型。

广义线性混合模型 (GLMM) 扩展了线性模型，使得目标可以有非正态分布，通过指定的连接函数与因子和协变量线性相关，并且观测值可能相关。 广义线性混合模型涵盖了各种模型，从简单线性回归模型到非正态纵向模型数据的复杂多级模型。

使用 Cox 回归节点，您可以在已有的检查记录中建立时间事件的生存模型。 对于输入变量的给定值，该模型会生成一个生存函数，用来预测在给定时间 (*t*) 发生相关事件的概率。

使用支持向量机 (SVM) 节点，可以将数据分为两组，而无需过度拟合。 SVM 可以与宽数据集配合使用，例如那些含有大量输入字段的数据集。

通过贝叶斯网络节点，你可以利用对真实世界认知的判断力并结合所观察和记录的证据来构建概率模型。 该节点侧重于主要用于分类的树增强朴素贝叶斯 (TAN) 和马尔可夫毯网络。

自学响应模型（SLRM）节点可用于构建一个包含单个新观测值或少量新观测值的模型，通过此模型，无需使用全部数据对模型进行重新训练即可对模型进行重新评估。

时间序列节点估计时间序列数据的指数平滑模型、单变量自回归整合移动平均值 (ARIMA) 模型和多变量 ARIMA（即变换函数）模型，并生成未来性能的预测数据。 此“时间序列”节点类似于 SPSS Modeler V18 中不推荐使用的先前“时间序列”节点。 但是，此较新“时间序列”节点旨在利用 IBM SPSS Analytic Server 的能力来处理大数据，并在 SPSS Modeler V17 中添加的输出查看器中显示生成的模型。

The *k*-最近相邻元素 (KNN) 节点将新的观测值关联到预测变量空间中与其最邻近的 *k* 个对象的类别或值（其中 *k* 为整数）。 类似观测值相互靠近，而不同观测值相互远离。

空间-时间预测 (STP) 节点使用包含位置数据、预测输入字段（预测变量）、时间字段和目标字段的数据。 每个位置在数据中都有许多行，这些行表示每个预测变量在每个测量时间的值。 分析数据后，可以使用该数据来预测分析中使用的形状数据内任意位置处的目标值。

关联模型

关联模型查找您数据中的模式，其中一个或多个实体（如事件、购买或属性）与一个或多个其他实体相关联。 这些模型构建定义这些关系的规则集。 数据中的字段可以作为输入和目标。 您可以手动查找这些关联，但关联规则算法可以更快速地完成，并能探索更多复杂的模式。 Apriori 和 Carma 模型是使用此类算法的示例。 另一种类型的关联模型是序列检测模型，后者可以在按时间建立结构的数据中查找顺序模式。

在预测多个结果时，关联模型最为有用，例如，购买了产品 X 的客户同时也购买了 Y 和 Z。 关联模型可将特定结论（例如，购买某物的决策）与一组条件相关联。 关联规则算法相对于更标准的决策树算法（C5.0 和 C&RT）的优势在于，它可以找到任何属性间存在的关联。 决策树算法只使用单一结论来构建规则，而关联算法则试图找到更多规则，且每个规则具有不同的结论。

关联节点

“先验”节点从数据抽取一组规则，即抽取信息内容最多的规则。 Apriori 节点提供五种选择规则的方法并使用复杂的索引模式来高效地处理大数据集。 对于较大的问题，Apriori 训练的速度通常较快；它对可保留的规则数量没有任何限制，而且可处理最多带有 32 个前提条件的规则。 “先验”要求输入和输出字段均为分类型字段，但因为它专为处理此类型数据而进行优化，因而处理速度快得多。

CARMA 模型在不要求用户指定输入或目标字段的情况下从数据抽取一组规则。 与 Apriori 相反，CARMA 节点提供规则支持的构建设置（支持前项和后项），而不仅仅是前项支持。 这就意味着生成的规则可以用于更多应用程序，例如用于查找产品或服务（前项）的列表，这些产品或服务的后项为想在节日期间促销的商品。

序列节点可发现连续数据或与时间有关的数据中的关联规则。 序列是一系列可能会以可预测顺序发生的项目集合。 例如，一个购买了剃刀和须后水的顾客可能在下次购物时购买剃须膏。 序列节点基于 CARMA 关联规则算法，该算法使用一个有效的两次传递方法查找序列。

“关联规则”节点与 Apriori 节点类似；但是，与 Apriori 不同，“关联规则”节点能够处理列表数据。 另外，“关联规则”节点可以与 IBM SPSS Analytic Server 配合使用，以处理大型数据以及利用更快的并行处理功能。

细分模型

细分模型将数据划分为具有类似输入字段模式的记录段或聚类。 细分模型只对输入字段感兴趣，没有输出或目标字段的概念。 细分模型的示例为 Kohonen 网络、K-Means 聚类、二阶聚类和异常检测等。

在不知道特定结果的情况下（例如，需要识别新犯罪模式或在客户群中识别利益群体时），细分模型（也称为“聚类模型”）非常有用。 聚类模型主要用来确定相似记录的组并根据它们所属的组来为记录添加标签。 此方法的优点在于，不用提前了解这些组及其特征就可以使用，它使聚类模型（其中没有需要模型预测的预定义输出或目标字段）区别于其他的建模技术。 对于这些模型来说，没有正确或错误的结果之分。 模型的值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。 聚类模型通常用于创建在后续分析中用作输入的聚类或段（例如，将潜在用户分成几个相似的子组）。

细分节点

“自动聚类”节点估算和比较识别具有类似特征记录组的聚类模型。 节点工作方式与其他自动建模节点相同，使您在一次建模运行中即可试验多个选项组合。 可使用基本度量对模型进行比较，尝试对聚类模型进行过滤，对其有用性进行排名，并提供基于特定字段重要性的度量。

K-Means 节点将数据集聚类到不同分组（或聚类）。 此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。 *k*-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。 此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。 您可以通过查看模型块 中每个单元所捕获观测值的数量来找出规模较大的单元。 这将让您对聚类的相应数量有所估计。

TwoStep 节点使用二阶聚类方法。 第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。 第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。 TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。 它可以高效处理混合的字段类型和大型的数据集。

Anomaly Detection 节点确定不符合“正常”数据格式的异常观测值（离群值）。 通过此节点，即使离群值不符合任何先前已知的模式，即使您并不确定要查找的内容，也可以识别这些离群值。

注: KNN， SVM， GENLin， Cox， SLRM 和 Bayes Net 节点是 SPSS Modeler 中扩展 Classification Module 的一部分，在 Modeler Premium 许可证下受支持。

数据库内数据挖掘模型

IBM SPSS Modeler 支持与可从数据库供应商获取的数据挖掘和建模工具集成，包括 Oracle Data Miner 和 Microsoft Analysis Services。 可以在 IBM SPSS Modeler 应用程序内的所有数据库中构建、评分和存储模型。 有关完整详细信息，请参阅《 *IBM SPSS Modeler* 数据库内挖掘指南》。

IBM SPSS Statistics 模型

如果您在计算机上拥有 IBM SPSS Statistics 安装和许可的一个副本，您可以从 IBM SPSS Modeler 访问和运行某些 IBM SPSS Statistics 例程以构建模型和给模型评分。

构建分割模型

通过拆分建模，可以使用单个流来为标志、名义或连续输入字段的每个可能值构建单独的模型，并且生成的模型全部都可从单个模型块进行访问。 输入字段的可能值可能对模型具有非常不同的影响。 使用分割建模，您可以容易地在流的一次执行中为每个可能的字段值构建最佳拟合模型。

请注意，交互建模会话不能使用分割。 您通过互动建模单独指定每个模型，而使用分割会自动构建多个模型，所以使用分割没有优势。

分割建模会指定某个输入字段为分割字段。 您可以通过在“类型”规范中将字段角色设置为**分割**来执行此操作。

只能将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为拆分字段。

您可以将多个输入字段分配为分割字段。 但是这种情况下，所创建模型数量可能大增。 给所选分割字段值的每个可能组合构建一个模型。 例如，如果三个输入字段指定为分割字段，每个字段具有三个可能值，那么结果会创建 27 个不同模型。

即使在您将一个或多个字段指定为分割字段后，您仍然可以通过建模节点对话框上的复选框设置来选择创建多个分割模型还是单个模型。

如果定义了分割字段但未选择复选框，那么只生成一个模型。 同样，如果选择了复选框但未定义分割字段，那么分割被忽略，生成一个模型。

当您执行流时，在后台为分割字段的每个可能值构建单独的模型，但只有一个模型块置于模型选用板和流工作区中。 分割模型块由分割符号指示；这是叠加在模型块图像上的两个灰色矩形。

浏览分割模型块时，您会看到包含已构建的所有单独模型的列表。

您可以通过在查看器中双击块从列表中查看单个模型。 这样打开单个模型的标准浏览器窗口。 当块位于工作区中时，双击缩略图打开标准大小的图形。 有关更多信息，请参阅主题 第 35 页的『拆分模型查看器』。

一旦将模型创建为分割模型之后，就不能删除其分割处理，也不能从分割建模节点或模型块下游撤销分割。

示例。 某个国内零售商希望按产品类别估算其国内每家店铺的销售情况。 则其通过使用分割建模，将其输入数据的“店铺”字段指定为分割字段，这样能在一次操作中为每个店铺的每个分类构建单独的模型。 其然后可以使用所得信息比只使用一个模型更加准确地控制库存水平。

分割和分区

分割与分区共有某些特征，但其使用方式截然不同。

分区将数据集随机分为两部分或三部分：训练、检验和（可选）验证，并用于检验单个模型的性能。

分割将按分割字段的可能值的数目划分数据集，并用于构建多个模型。

分区和分割工作方式彼此完全不同。 您可以在建模节点中选择一个、两个或一个也不选。

支持分割模型的建模节点

大量建模节点可创建分割模型。 例外的情况是自动聚类、PCA/因子、特征选择、SLRM、随机树、树-AS、线性-AS、LSVM、关联模型（Apriori、Carma 和序列）、聚类模型（K 均值、Kohonen、二阶和异常）、统计信息 模型以及用于数据库内建模的节点。

支持拆分建模的建模节点是：

C&R 树


贝叶斯网络


线性

QUEST


GenLin


GLMM

CHAID


KNN


STP

C5.0

Cox


单类 SVM


神经网络


自动分类器


XGBoost Tree


决策列表


自动数字


XGBoost Linear


回归


Logistic


HDBSCAN


判别


SVM


时间序列

受分割影响的特征

使用分割模型以各种方式影响大量 IBM SPSS Modeler 特征。 本部分提供有关在流中将拆分模型与其他节点配合使用的指导。

记录选项节点

当在包含年“样本”节点的流中使用拆分模型时，按拆分字段对记录进行分层，以实现记录的平均采样。 当选择复杂作为样本方法时，此选项可用。

如果流包含“平衡”节点，那么平衡适用于输入记录的整体集合，而非拆分内的记录子集。

当通过“汇总”节点来汇总记录时，如果要计算每个拆分的汇总，请将拆分字段设置为关键字段。

字段选项节点

通过“类型”节点，可以指定将哪个或哪些字段用作拆分字段。

注: 尽管“整体”节点用于组合两个或多个模型块，但其无法用于撤销拆分操作，因为拆分模型包含在单个模型块内。

建模节点

分割模型不支持预测变量重要性（估算模型时预测变量输入字段的相对重要性）计算。 构建分割模型时会忽略预测变量重要性设置。

注: 使用拆分模型时，会忽略调整倾向分数设置。

KNN（最近相邻元素）节点只有在设置为预测目标字段时才支持拆分模型。 其他设置（只标识最近相邻元素）不创建模型。 如果选择选项**自动选择 k**，那么每个拆分模型可能具有不同数量的最近相邻元素。 因此，整体模型生成的列数等于所有拆分模型找到的最近相邻元素的最大数。 对于最近相邻元素数小于此最大值的拆分模型，存在对应数量的已填充 $null 值的列。 有关更多信息，请参阅主题 第 255 页的『KNN 节点』。

数据库建模节点

数据库内建模节点不支持分割模型。

模型块

不可能从分割模型块**导出到 PMML**，因为块包含多个模型，而 PMML 不支持这种包装。 可以导出到文本或 HTML。

建模节点字段选项

所有建模节点均有一个“字段”选项卡，在此选项卡中指定的字段将用于构建模型。

在构建模型之前，需要指定要将哪些字段用作目标和输入。 某些特殊情况下，所有建模节点将采用上游的类型节点的字段信息。 如果正在使用类型节点选择输入和目标字段，则不必在此选项卡上做任何更改。 （特殊情况包括序列节点和文本抽取节点，这两个节点需要在建模节点中指定字段设置。）

使用类型节点设置。 该选项通知节点使用来自上游类型节点的字段信息。 这是缺省选项。

使用定制设置。 该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。 选中此选项后，请根据需要指定下面的字段。

注：并非所有字段都对所有节点显示。

- **使用事务格式（仅限 Apriori、CARMA、MS 关联规则和 Oracle Apriori 节点）**。 如果源数据为**事务处理格式**，那么选中此复选框。 此格式的记录具有两个字段，一个为标识字段，一个为内容字段。 每条记录代表单个交易或单个项，关联项通过相同的标识得以链接。 如果数据为**表格格式**，请取消选中此复选框，表格格式中项目由独立标志表示，其中每个标志字段表示某个特定项目是否存在，且每条记录表示关联项目的完整集合。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。
  - **标识**。 对于事务处理数据，请从列表中选择标识字段。 数字字段或符号字段可用作标识字段。 此字段的每个唯一值都应该表明一个特定的分析单元。 例如，在市场购物篮的应用中，每个标识可能表示一个客户。 对于 Web 日志分析应用，每个标识可能代表一台计算机（以 IP 地址表示）或一个用户（以登录数据表示）。
  - **标识是连续的**。 （仅限 Apriori 和 CARMA 节点）如果您的数据进行了预先排序，以便所有标识相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。 如果您的数据未经预先排序（或者您不确定），请将此选项保持未选中状态，那么该节点将自动对数据进行排序。

    注: 如果数据未经排序，而您选择此选项，那么模型中可能会出现无效的结果。
  - **内容**。 指定模型的内容字段。 这些字段包含与关联建模有关的项目。 您可以指定多个标志字段（如果数据为表格格式）或者一个名义字段（如果数据为事务格式）。
- **目标**。 对于需要一个或多个目标字段的模型，请选择目标字段或字段。 此操作与在“类型”节点中将字段的角色设置为目标类似。
- **评估**。 （仅适合自动聚类模型。） 不为聚类模型指定目标，但可选择一个评估字段以确定其重要性等级。 此外，还可评估聚类区分此字段值的程度，从而指示是否可使用聚类来预测此字段。 注：评估字段必须是包含多个值的字符串。
  - **输入**。 选择输入字段或字段。 此操作与在“类型”节点中将字段的角色设置为输入类似。
  - **分区**。 通过此字段，您可以指定用于针对模型构建中的训练、检验和验证阶段将数据划分为不同样本的字段。 通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。 如果已使用类型或分区节点定义了多个分区字段，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。） 同时请注意，要在分析时应用选定分区，还必须启用节点的“模型选项”选项卡中的分区功能。 （取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）
- **拆分**。 对于分割模型，选择分割字段或字段。 此操作与在“类型”节点中将字段的角色设置为分割类似。 您仅可将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为分割字段。 选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关更多信息，请参阅主题 第 21 页的『构建分割模型』。

- **使用频率字段**。 此选项允许您选择一个字段作为频率权重。 如果训练数据中的每条记录代表多个单元（例如，您正在使用聚合的数据），那么可采用此项。 字段值应是每个记录代表的单位的数量。 有关更多信息，请参阅主题 第 24 页的『使用频率和权重字段』。

注：如果您看到错误消息**元数据（在输入/输出字段上）无效**，请确保已指定所有必填字段，例如“频率”字段。

- **使用权重字段**。 此选项允许您选择一个字段作为观测值权重。 观测值权重将作为对输出字段各个水平上方差的差异的一种考量。 有关更多信息，请参阅主题 第 24 页的『使用频率和权重字段』。
- **后项**。 对于规则归纳节点 (Apriori)，请选择在生成的规则集中用作结果的字段。 （这对应于“类型”节点中角色为目标或双向的字段。）
- **前项**。 对于规则归纳节点 (Apriori)，请选择在生成的规则集中用作前提条件的字段。 （这对应于“类型”节点中角色为输入或双向的字段。）

某些模型的“字段”选项卡与本节所述“字段”选项卡不同。

- 有关更多信息，请参阅主题 第 202 页的『序列节点字段选项』。
- 有关更多信息，请参阅主题 第 193 页的『CARMA 节点字段选项』。

使用频率和权重字段

频率和权重字段用于赋予某些记录高于其他记录的附加重要性，例如，因为您知道一部分人未在训练数据（权重）中表示出来，或者因为一个记录代表多个相同观测值（频率）。

- 频率字段的值应为正整数。 频率权重小于或等于零的记录将排除在分析之外。 非整数频率权重将四舍五入为最近的整数。
- 观测值权重应为正数但不一定是整数值。 观测值权重小于或等于零的记录将排除在分析之外。

评分频率和权重字段

频率和权重字段用于训练模型，但不用于评分，因为每条记录的分数基于该记录的特征，而与它代表的观测值个数无关。 例如，假设您有下表中的数据。

表 *1:* 数据示例

已婚	**已响应**
是	是
是	是
是	是
是	否
否	是
否	否
否	否

基于上表，可以得出这样的结论：四分之三的已婚者对促销作出响应；而三分之二的未婚者对此未作出响应。 因此，您将相应地对任何新记录进行评分，如下表所示。

表 *2:* 已评分的记录示例

已婚	**$-已响应**	**$RP-已响应**
是	是	0.75 (3/4)
否	否	0.67 (2/3)

另外，您可以使用频率字段更简洁地存储训练数据，如下表所示。

表 *3:* 已评分记录的替代项示例		
已婚	**已响应**	**频率**
是	是	3
是	否	1
否	是	1
否	否	2

因为此表完全代表同一数据集，因此可以构建相同的模型并仅根据婚姻状况预测响应率。 如果评分数据中包含 10 个已婚人员，那么无论这些人员是显示为 10 条单独的记录，还是一条频率值为 10 的记录，对于每个人员，都将预测为是。 权重（虽然通常不是整数）可以被认为以相似的方式表示记录的重要性。 这就是对记录进行评分时不使用频率和权重字段的原因。

评估和比较模型

某些模型类型可支持频率字段，某些可支持权重字段，还有一些可同时支持这两种字段。 但在使用这两种字段的所有情况中，它们仅用于构建模型，在使用“评估”节点或“分析”节点对模型进行评估时，或者在使用受“自动分类器”节点和“自动数值”节点支持的大部分方法进行模型排秩时，均不考虑使用这两种字段。

- 例如，在使用评估图表比较模型时将忽略频率和权重值。 这将在使用频率和权重字段的模型与不使用这些字段的模型之间进行级别比较，但同时意味着，必须使用不依赖频率或权重字段并且可以准确表示总体的数据集才能获得准确的评估。 实际上，您可以通过确保使用频率或权重字段值始终为空或 1 的测试样本对模型进行评估来完成此操作。 （此限制仅在评估模型时适用；如果训练样本和测试样本的频率或权重值始终为 1，那么一开始就没有任何理由使用这些字段。）
- 如果使用“自动分类器”基于“利润”对模型进行排秩，那么可考虑频率，在这种情况下推荐使用此方法。
- 如果有必要，可以使用分区节点，将数据分割为训练样本和检验样本。

建模节点分析选项

许多建模节点都包括“分析”选项卡，您可以通过该选项卡获取预测变量重要性信息以及原始倾向评分和调整后的倾向评分。

模型评估

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此缺省情况下，对于某些模型，预测变量重要性均处于关闭状态。 预测变量重要性对于决策列表模型不可用。 有关更多信息，请参阅第 32 页的『预测变量重要性』。

倾向评分

可以在建模节点中和模型块的“设置”选项卡上启用倾向评分。 该功能仅在所选目标为标志字段时才可用。 有关更多信息，请参阅主题 第 26 页的『倾向评分』。

计算原始倾向评分。 原始倾向评分仅派生自基于训练数据的模型。 如果模型预测值为 真（将响应），那么倾向与 P 相同，其中 P 为预测的可能性。 如果模型预测的值为假，那么计算出的倾向为 (1 – P)。

- 如果构建模型时选择了此选项，那么缺省情况下将在模型块中启用倾向评分。 不过，无论是否在建模节点中选择了原始倾向评分，都可以始终在模型块中选择启用原始倾向评分。
- 对模型进行评分时，原始倾向评分将被添加到将 *RP* 字母附加到标准前缀的字段中。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RRP-churn*。

计算调整后的倾向评分。 原始倾向仅基于由可能过度拟合的模型指定的估计，这将导致过于乐观地估计倾向。 调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。

- 与原始置信度分数不同，调整后的倾向评分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 *AP* 字母附加到标准前缀的字段中添加调整后的倾向评分。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RAP-churn*。 调整后的倾向评分不适用于 logistic 回归模型。
- 在计算调整后的倾向评分时，必须尚未平衡用于计算的检验或验证分区。 为避免这一点，请确保在任何上游平衡节点中选中 **仅平衡训练数据** 选项。 此外，如果已在上游获取了复杂样本，那么这将导致调整后的倾向评分无效。
- 调整后的倾向评分不适用于“增强型”树和规则集模型。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。

基于。 对于要进行计算的调整后的倾向评分，流中必须存在一个分区字段。 可以指定是使用检验分区还是验证分区进行此计算。 为获取最佳结果，检验或验证分区包含的记录数量应至少与用于训练原始模型的分区所包含的记录数相同。

倾向评分

对于返回预测为 是或 否的模型，您除了可以要求标准预测和置信度值以外，还可要求倾向评分。 倾向评分指示特定结果或响应的可能性。 下表提供了一个示例。

表 *4:* 倾向评分

客户	**要响应的倾向**
Joe Smith	35%
Jane Smith	15%

倾向评分仅适用于有标志目标的模型，并且指示为字段定义的值为 真的可能性，如在源节点或类型节点中指定的那样。

倾向评分与置信度评分

倾向评分与置信度评分不同，置信度评分应用于当前预测，即是或否。 例如，在预测为否时，高置信度实际表示不响应的可能性很高。 倾向评分可以回避此限制，从而轻松比较所有记录。 例如，置信度为 *0.85* 的 否预测将转换为 *0.15*（或 *1* 减 *0.85*）的原始倾向。

表 *5:* 置信度评分

客户	**预测**	**置信度**
Joe Smith	会响应	.35
Jane Smith	不会响应	.85

获得倾向评分

- 可以在建模节点中的“分析”选项卡或模型块中的“设置”选项卡上启用倾向评分。 该功能仅在所选目标为标志字段时才可用。 有关更多信息，请参阅主题 第 25 页的『建模节点分析选项』。
- 也可以通过整体节点计算倾向评分，具体取决于所用的整体方法。

计算调整后的倾向评分

计算调整后的倾向评分将作为构建模型过程的一部分，否则没有可用的调整后的倾向评分。 构建模型后，则可使用检验或验证分区中的数据对模型进行评分，同时通过在该分区上分析原始模型的性能，构建一个提供调整后的倾向评分的新模型。 根据模型的类型，可以使用两种方法之一来计算调整后的倾向评分。

- 对于规则集模型和树模型，要生成调整后的倾向分数，可通过重新计算每个树节点上每个类别的频率（适用于树模型）或重新计算每个规则的支持和置信度（适用于规则集模型）。 这样一来，请求调整后的倾向评分时将使用与原始模型一起存储的新规则集模型或树模型。 每次将原始模型应用到新数据时，都会随之将新模型应用到原始倾向分数以生成调整后的分数。

- 对于其他模型，通过对检验或验证分区上的原始模型进行评分而生成的记录将按其原始倾向评分进行分级。 接着，对定义非线性函数的神经网络模型进行训练，该函数从每个分级的平均原始倾向中映射到相同分级的平均观测倾向中。 正如之前对树模型的说明，得出的神经网络模型将与原始模型一起存储，并且在请求调整后的倾向分数时应用到原始倾向评分。

关于测试分区中缺失值的警告说明。 检验/验证分区中缺失输入值的处理方法随模型不同而有所差异（请参阅各个模型评分算法以获取详细信息）。 有缺失输入值时，C5 模型无法计算调整倾向。

误分类成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

除 C5.0 模型之外，在对模型进行评分时，误分类成本是不适用的；在使用自动分类器节点、评估图表或分析节点对模型进行排秩或比较时，误分类成本也不予以考虑。 将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的 成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

注: 仅“决策树”模型允许在构建时指定成本。

模型块

图 *19:* 模型块

模型块是模型的容器，其中包含一组规则、公式或方程式，它们代表在 SPSS Modeler 中模型构建操作的结果。 模型块的主要用途是对数据进行评分以生成预测，或者实现对模型属性进行进一步分析。 在屏幕上打开模型块后，可以查看有关模型的各类详细信息，例如，在模型创建中输入字段的相对重要性。 要查看预测，则需要进一步添加并执行处理或输出节点。 有关更多信息，请参阅主题 第 35 页的『使用流中的模型块』。

图 *20:* 从建模节点到模型块的模型链接

在成功地执行建模节点后，会在流工作区上放置对应的模型块，并以金色钻石形图标表示（因此称之为“块”）。 在流工作区上显示的模型块，带有到位于建模节点之前的最近合适节点的连接（实线），以及到建模节点本身的链接（虚线）。

此外，模型块也放置在位于 IBM SPSS Modeler 窗口右上角的“模型”选用板中。 从任一位置均可选中模型块，并浏览模型的详细信息。

在建模节点成功执行后，模型块始终位于“模型”选用板中。 可以设置用户选项来控制是否也将模型块置于流工作区上。

以下主题提供了使用 IBM SPSS Modeler 中模型块的相关信息。 要深入了解所使用的算法，请参阅产品下载过程中以 PDF 文件形式提供的《*IBM SPSS Modeler* 算法指南》。

模型链接

缺省情况下，在流工作区上显示的模型块带有指向创建它的建模节点的链接。 这在具有多个模型块的复杂流中特别有用，它使您能够识别将被每个建模节点更新的模型块。 每个链接包含一个指示当建模节点执行时是否替换模型的符号。 有关更多信息，请参阅主题 第 29 页的『替换模型』。

定义和删除模型链接

您可以在工作区上手动定义和删除模型链接。 在定义新的链接后，光标将变成链接光标。

图 *21:* 链接光标

定义新链接（上下文菜单）

1. 右键单击要作为链接起点的建模节点。
2. 从上下文菜单中选择**定义模型链接**。
3. 单击要作为链接终点的模型块。

定义新链接（主菜单）

1. 单击要作为链接起点的建模节点。
2. 在主菜单中，选择：

   编辑 > 节点 > 定义模型链接
3. 单击要作为链接终点的模型块。

删除现有链接（上下文菜单）

1. 右键单击位于链接终点的模型块。
2. 从上下文菜单中选择**删除模型链接**。

或者：

1. 右键单击位于链接中部的符号。
2. 从上下文菜单中选择**删除链接**。

删除现有链接（主菜单）

1. 单击要删除其链接的建模节点或模型块。
2. 在主菜单中，选择：

   编辑 > 节点 > 除去模型链接

复制和粘贴模型链接

如果复制了带链接的模型块，但未包括其建模节点，那么当将其粘贴到同一流中时，粘贴后的模型块将具有到建模节点的链接。 新链接具有与原始链接相同的模型替换状态（请参阅第 29 页的『替换模型』）。

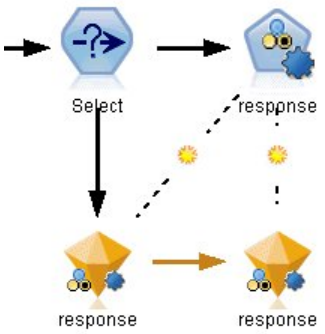


图 *22:* 复制和粘贴带链接的模型块

如果将模型块连同其链接的建模节点一起复制并粘贴，那么无论对象粘贴到同一流还是新流中，链接都将保留。

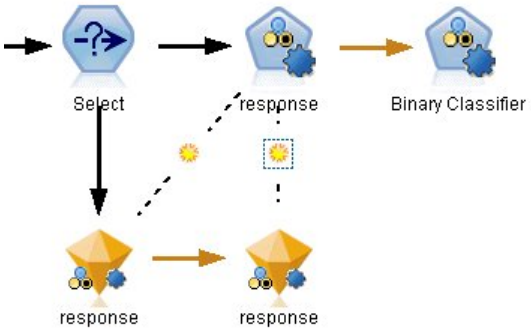


图 *23:* 复制和粘贴带链接的模型块

注：如果复制了带链接的模型块，但未包括其建模节点，那么将其粘贴到新流（或不包含建模节点的超节点）中时，链接将中断，并且将只粘贴模型块。

模型链接和超节点

如果定义超节点包含链接模型的建模节点或模型块（但未同时包含），链接将被破坏。 展开超节点不会恢复链接，只能通过撤销创建超节点来完成此操作。

替换模型

您可以选择在重新执行创建模型块的建模节点时是否替换（即更新）现有模型块。 如果关闭替换选项，那么重新执行建模节点时将创建新的模型块。

每个从建模节点到模型块的链接包含一个指示当建模节点重新执行时是否替换模型的符号。

图 *24:* 模型替换处于打开状态的模型链接

初始显示链接时，模型替换处于打开，并通过链接中的小旭日形符号指示。 在此状态下，重新执行位于链接一端的建模节点就会更新另一端的模型块。

图 *25:* 模型替换处于关闭状态的模型链接

如果模型替换处于关闭，那么链接符号替换为灰色点。 在此状态下，重新执行位于链接一端的建模节点会在工作区上新增一个更新后的模型块。

在任一情况下，在“模型”选用板中是更新现有模型块还是新增模型块，取决于**替换原有模型**系统选项的设置。

执行顺序

当执行具有包含模型块的多个分支的流时，首先对流进行评估，以确保先执行模型替换处于打开的分支，然后再执行使用结果模型块的任何分支。

如果您的需求更为复杂，那么可通过脚本手动设置执行顺序。

更改模型替换设置

1. 右键单击链接上的符号。
2. 根据情况选择**打开（关闭）模型替换**。

注: 模型链接上的模型替换设置将覆盖“用户选项”对话框（工具 > 选项 > 用户选项）的“通知”选项卡上的设置。

模型选用板

通过模型选用板（位于管理器窗口的“模型”选项卡中），您可以按各种方式使用、检查和修改模型块。

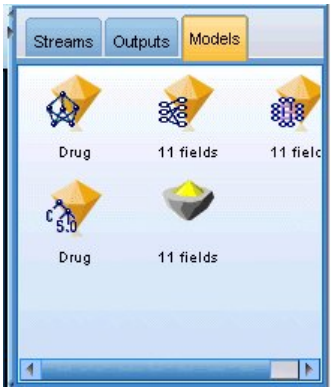


图 *26:* “模型”选用板

右键单击模型选用板中的模型块，打开带有以下选项的上下文菜单：

- **添加到流**。 将模型块添加到当前处于活动状态的流中。 如果流中存在选定节点，当可以连接时，模型块将连接到选定节点，否则链接到最近的可能节点。 如果创建模型的建模节点仍然在流中，那么显示的模型块将带有到建模节点的链接。
- **浏览**。 打开块的模型浏览器。
- **重命名并添加注解**。 通过此选项，您可以重命名模型块和/或修改模型块的注解。
- **生成建模节点**。 如果要修改或更新某个模型块，但无法使用用于创建该模型的流，那么可以使用此选项与创建原始模型相同的选项来重新生成一个建模节点。
- **保存模型，将模型另存为**。 将此模型块保存到外部生成模型 (.gm) 二进制文件。
- **存储模型**。 在 IBM SPSS Collaboration and Deployment Services Repository 中存储模型块。

- **导出 PMML**。 以预测模型标记语言 (PMML) 格式导出模型块，其可用于 IBM SPSS Modeler 之外的新数据评分。 **导出 PMML** 可用于所有生成的模型节点。
- **添加到项目**。 保存模型块并将其添加到当前工程。 在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。 在 CRISP-DM 选项卡上，此节点将被添加到缺省项目阶段。
- **删除**。 从选用板中删除模型块。

右键单击模型选用板中的未占用区域，打开带有以下选项的上下文菜单：

- **打开模型**。 加载先前在 IBM SPSS Modeler 中创建的模型块。
- **检索模型**。 从 IBM SPSS 协作和部署服务 存储库检索保存的模型。
- **装入选用板**。 从外部文件加载保存的模型选用板。
- **检索选用板**。 从 IBM SPSS 协作和部署服务 存储库检索保存的模型选用板。
- **保存选用板**。 将模型选用板的所有内容保存到外部生成模型选用板 (.gen) 文件。
- **存储选用板**。 将模型选用板的所有内容保存到 IBM SPSS 协作和部署服务 存储库中。
- **清除选用板**。 从选用板中删除所有模型块。
- **将选用板添加到项目**。 保存模型选用板并将其添加到当前工程。 在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。 在 CRISP-DM 选项卡上，此节点将被添加到缺省项目阶段。
- **导入 PMML**。 从外部文件加载模型。 可以打开、浏览由 IBM SPSS Statistics 或其他支持此格式的应用程序所创建的 PMML 模型并对其进行评分。 有关更多信息，请参阅主题 第 36 页的『导入和导出 PMML 模型』。

浏览模型块

通过模型块浏览器，您可以检查和使用模型的结果。 在浏览器中，您可以保存、打印或导出生成模型，检查模型摘要，查看或编辑模型注释。 对于某些类型的模型块，还可以生成新的节点，例如“过滤”节点或“规则集”节点。 对于某些模型，您还可以查看模型参数，如规则或聚类中心。 对于某些类型的模型（基于树的模型和聚类模型），您可以查看其模型结构的图表显示。 使用模型块浏览器的控件如下所述。

菜单

文件菜单。 所有模型块均有一个“文件”菜单，其中包括以下选项的子集：

- **保存节点**。 将模型块保存到某个节点 (.nod) 文件。
- **存储节点**。 在 IBM SPSS 协作和部署服务 存储库中保存模型块。
- **页眉和页脚**。 通过此选项，您可以编辑页面的页眉和页脚，以便从模型块进行打印。
- **页面设置**。 通过此选项，您可以更改页面设置，以便于从模型块进行打印。
- **打印预览**。 显示模型块的打印预览。 从子菜单中选择要预览的信息。
- **打印**。 打印模型块的内容。 从子菜单中选择要打印的信息。
- **打印视图**。 打印当前视图或所有视图。
- **导出文本**。 将模型块内容导出到某个文本文件中。 从子菜单中选择要导出的信息。
- **导出 HTML**。 将模型块内容导出到 HTML 文件中。 从子菜单中选择要导出的信息。
- **导出 PMML**。 以预测模型标记语言 (PMML) 格式导出模型，导出的文件可在其他 PMML 兼容软件中使用。 有关更多信息，请参阅主题 第 36 页的『导入和导出 PMML 模型』。
- **导出 SQL**。 以结构化查询语言 (SQL) 导出模型，可以通过其他数据库来编辑和使用导出的 SQL。

  注: 仅在下列模型中提供了 SQL 导出：C5、C&RT、CHAID、QUEST、线性回归、Logistic 回归、神经网络、PCA／因子以及决策列表模型。
- **为 Server Scoring Adapter 发布**。 将模型发布到安装有评分适配器的数据库中，可在数据库中进行模型评分。 有关更多信息，请参阅主题 第 38 页的『为评分适配器发布模型』。

“生成”菜单。 多数模型块还具有“生成”菜单，通过此菜单可以根据模型块生成新节点。 此菜单中的可用选项取决于您所浏览模型的类型。 请查看具体的模型块类型，以详细了解您可从特定模型中生成的内容。

“视图”菜单。 在模型块的“模型”选项卡上，此菜单允许您显示或隐藏在当前模式下可用的各类直观表示工具栏。 要使全部工具栏可用，可从“常规”工具栏中选择“编辑模式”（画笔图标）。

“预览”按钮。 某些模型块具有“预览”按钮，允许您查看模型数据的样本，包括由建模过程创建的额外字段。 缺省显示的行数为 10，不过可以在流属性中更改此值。

“添加到当前项目”按钮。 保存模型块并将其添加到当前工程。 在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。 在 CRISP-DM 选项卡上，此节点将被添加到缺省项目阶段。

模型块概要/信息

模型块的“概要”选项卡或“信息”视图显示了关于字段、构建设置和模型估计过程的信息。 结果以树形视图显示，通过单击指定项可以扩展或合并树形视图。

分析。 显示模型的相关信息。 具体详细信息因模型类型而异，这些信息可在每种模型块的相应章节中找到。 另外，如果执行了附加到此建模节点的“分析”节点，那么还会在此部分显示该分析中的信息。

字段。 列出构建模型时用作目标和输入的字段。 对于分割模型，也列出确定分割的字段。

注: 在具有增强或组装模式的神经网络模型、线性模型和其他模型的信息视图中，所显示的图标相同（名义图标），而与类型为标志、名义还是有序无关。

构建设置/选项。 包含构建模型时使用的设置的相关信息。

训练摘要。 显示模型类型、用于创建模型的流、创建模型的用户、模型构建时间和构建模型所用时间。 请注意，只有“摘要”选项卡上提供构建模型所耗用的时间，“信息”视图中不提供此时间。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。 预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

预测变量重要性对于可生成相应重要性统计标准的模型可用，包括神经网络模型、决策树（C&R 树、C5.0、CHAID 和 QUEST）、贝叶斯网络模型、判别分析模型、SVM 和 SLRM 模型、线性和 logistic 回归模型、广义线性模型以及最近邻元素 (KNN) 模型。 对于这些模型中的大部分而言，可以在建模节点的“分析”选项卡上启用预测变量重要性。 有关更多信息，请参阅主题 第 25 页的『建模节点分析选项』。 有关 KNN 模型，请参阅第 256 页的『相邻元素』。

注：对于分割模型，预测变量重要性不受支持。 构建分割模型时会忽略预测变量重要性设置。 有关更多信息，请参阅主题 第 21 页的『构建分割模型』。

计算预测变量重要性所用的时间远远大于构建模型的用时，特别当使用较大数据集时。 对于 SVM 和 logistic 回归模型，计算变量重要性的用时比对其他模型执行此操作的用时都要长，所以缺省情况下这两种模型均禁用此功能。 使用一个包含许多预测变量的数据集时，使用“特征选择”节点进行初始筛选可以较快地生成结果（请参阅以下内容）。

- 如果适用，可以从检验分区计算出预测变量重要性。 否则，就使用训练数据。
- 预测变量重要性也适用于 SLRM 模型，但需要使用 SLRM 算法进行计算。 有关更多信息，请参阅主题 第 246 页的『SLRM 模型块』。
- 可以使用 IBM SPSS Modeler 的图表工具进行交互、编辑，并保存图表。
- 还可以根据预测变量重要性图表中的信息生成“过滤”节点。 有关更多信息，请参阅主题 第 33 页的『基于重要性过滤变量』。

预测变量重要性和特征选择

在某些情况下，模型块中显示的预测变量重要性图表可能似乎给出与“特征选择”节点相似的结果。 当特征选择基于每个输入字段与特定目标（与其他输入无关）的关系强度对输入字段进行排序时，预测变量重要性图表将显示此特定模型中各个输入的相对重要性。 因此，在筛选输入时使用特征选择可能较为保守。 例如，如果 工作职务和 工作类别与薪资的关系强度相同，特征选择就会指示这两者都很重要。 但在建模时，还需考虑交互性和相关性。 这样，当两个输入的大部分信息都相同时，您可能会发现仅使用了两个输入之一。

在实际应用中，特征选择对预筛选最有用，特别是处理包含大量变量的较大数据集时，而预测变量重要性在微调模型时更为有用。

单一模型与自动化建模节点之间的预测变量重要性差异

根据您要从个别节点创建单一模型还是使用自动化建模节点来生成结果，您可能会看到预测变量重要性的细微差异。 这种实现上的差异是由一些工程限制所致。

例如，借助 CHAID 之类的单一分类器，此算法在计算重要性值时应用中止规则并使用概率值。 相反，自动分类器不使用中止规则，而是直接在计算中使用预测的标签。 这些差异可能意味着，如果您使用自动分类器来生成单一模型，那么与针对单一分类器计算出的值相比，重要性值可以被认为是粗略估算值。 要获取最准确的预测变量重要性值，我们建议使用单一节点来取代自动化建模节点。

基于重要性过滤变量

还可以根据预测变量重要性图表中的信息生成“过滤”节点。

标记要包括在图表上的预测变量（如果适用），然后从菜单中选择：

生成 > 过滤节点（预测变量重要性）

或者

> 字段选择（预测变量重要性）

最大变量数。 包括或排除等于指定数量的最重要预测变量。

重要性大于。 包括或排除所有相对重要性高于指定值的预测变量。

整体查看器

整体模型

整体模型提供了有关整体中的组件模型和整体性能的信息。

主 (独立视图) 工具栏允许您选择使用整体或参考模型来进行评分。 如果使用整体进行评分，您还可以选择组合规则。 这些更改不需要重新执行模型；但是，这些选择将保存到模型 (块) 以供评分和/或下游模型评估。 它们也会影响从整体查看器导出的 PMML。

组合规则。 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体得分值。

- 可以使用投票、最高概率或最高平均概率来组合**分类**目标的整体预测值。 **投票**选择在基本模型中最常具有最高概率的类别。 **最高概率**选择在所有基本模型中达到单一最高概率的类别。 **最高平均概率**选择当类别概率在基本模型中取平均值时具有最高值的类别。
- 可以通过对来自基本模型的预测值取平均值或中位数，对**连续**目标的整体预测值进行组合。

缺省值取自在建模过程中生成的指定。 更改组合规则会重新计算模型精确性并更新模型精确性的所有视图。 也会更新预测变量重要性图表。 如果选择参考模型用于评分，那么此控件将被禁用。

显示所有组合规则。 选择该选项时，所有可用组合规则的结果将显示在模型质量图表中。 组件模型精确性图表也将更新以显示每种投票方式的参考线。

模型摘要

“模型摘要”视图是整体质量和差异性的快照摘要。

质量。 该图表显示与参考模型和 navie 模型相比较的最终模型精确性。 精确性越大，模型越好的格式；“最佳”模型将具有最高精确性。 对于分类目标，精确性就是预测值与观测值匹配的记录百分比。 对于连续目标，精确性为 1 减去预测中的平均绝对误差 (预测值的绝对平均值减去观测值) 与预测值范围 (最大预测值减去最小预测值) 的比率。

对于 bagging 整体，参考模型是构建在整个培训分区上的标准模型。 对于 boosted 整体，参考模型是第一个组件模型。

如果未构建模型，那么由 Naive 模型代表精确性，并将所有记录分配给模态类别。 不会为连续目标计算 Naive 模型。

差异性。 该图表显示用于构建整体的组件模型间的“观点差异性”，以越大则差异性越大格式表示。 这是一种基本模型间预测差异程度的测量。 差异性对 boosted 整体模型不可用，同时也不会对连续目标显示。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。 预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

预测变量重要性对所有整体模型均不可用。 预测变量集在组件模型之间可能会有所不同，但可以为至少在一个组件模型中使用的预测变量计算重要性。

预测变量频率

由于选择的建模方法或预测变量选择不同，预测变量集在组件模型间也可能不同。 预测变量频率图是一个点图，显示了预测变量在整体组件模型中的分布。 每个点代表包含预测变量的一个或多个组件模型。 预测变量绘制在 y 轴上，并以频率的降序排序；因此，最顶端的是在最多组件模型中使用的预测变量，而最低端的是在最少组件模型中使用的预测变量。 将显示排在前 10 位的预测变量。

出现频率最高的预测变量通常是最重要的。 此图对于使预测变量集在组件模型间保持一致的方法没用。

组件模型精确性

该图表是组件模型预测精确性的点图。 每个点代表在 y 轴上绘制了精确性水平的一个或多个组件模型。 悬停在任意点上可获得对应的单独组件模型的信息。

参考线。 该图显示整体的颜色编码线以及参考模型和 naïve 模型。 对应于要用于评分的模型的线的旁边会显示一个复选标记。

互动。 该图表会在您更改组合规则时更新。

Boosted 整体。 为 boosted 整体显示一个线图。

组件模型详细信息

该表显示关于组件模型的信息，按行列出。 缺省情况下，组件模型按模型编号的升序排序。 您可以按任意列的值对这些行进行升序或降序排序。

模型。 代表组件模型创建顺序的数字。

精确性。 百分比形式的整体精确性。

方法。 建模方法。

预测变量。 组件模型中使用的预测变量数。

模型大小。 模型大小取决于建模方法：对于树，这是树中的节点数；对于线性模型，这是系数的数目；对于神经网络，这是突触的数目。

记录数。 训练样本中输入记录的加权数量。

自动数据准备

此视图显示在自动数据准备 (ADP) 步骤中排除了哪些字段，以及转换字段的派生方式等信息。 对于每个转换或排除字段，在此表中列出了字段名、在分析中的角色，以及 ADP 步骤所采取的操作。 字段是按照字段名称的字母升序排列的。

操作 **Trim outliers**（如果显示）表示位于分界值（平均值的 3 个标准差）之外的连续预测变量值被设为分界值。

分割模型的模型块

分割模型的模型块可以访问分割创建的所有单独模型。

分割模型块包含：

- 创建的所有分割模型列表，连同每个模型的统计量集合
- 有关整体模型的信息

从分割模型列表中，您可以打开单个模型以进一步检查。

拆分模型查看器

“模型”选项卡列出块中包含的所有模型，以各种形式提供有关分割模型的统计量。 它有两种一般形式，具体取决于建模节点。

排序依据。 使用此列表选择列出模型的顺序。 您可以根据任何显示列的值将列表按升序或降序排序。 或者，单击列标题，按该列将列表排序。 缺省是总精确性的降序。

显示/隐藏列菜单。 单击此按钮，以显示菜单，以便选择单个列以显示或隐藏。

视图。 如果您正在使用分区，您可以选择查看培训数据或测试数据的结果。

对于每个拆分，详细信息显示如下：

图形。 指示此模型数据分布的缩略图。 当块位于工作区中时，双击缩略图打开标准大小的图形。

模型。 模型类型图标。 双击图标打开此特定分割的模型块。

分割字段。 建模节点中指定为分割字段的字段及其各个可能值。

否。 拆分中的记录数。 此特定分割中涉及的记录数。

否。 使用的字段。 基于所用输入字段的数量排序分割模型。

总体准确性 (%)。拆分模型正确预测的记录数相对于该拆分中的记录总数的百分比。

拆分。 列标题显示用于创建拆分的字段，单元格为拆分值。 双击任意拆分以便为针对该拆分构建的模型打开“模型查看器”。

精确性。 百分比形式的整体精确性。

模型大小。 模型大小取决于建模方法：对于树，这是树中的节点数；对于线性模型，这是系数的数目；对于神经网络，这是突触的数目。

记录数。 训练样本中输入记录的加权数量。

使用流中的模型块

模型块置于流中，允许您对新数据进行评分并生成新节点。 通过对数据**进行评分**，您可以使用通过模型构建获得的信息来为新记录创建预测。 要查看评分结果，需要为模型块添加终端节点（即处理或输出节点）并执行终端节点。

对于某些模型而言，还可从模型块中获得有关预测质量的其他信息，例如置信度值或到聚类中心的距离。 通过生成新节点，您可以轻松地根据已生成模型的结构来创建新节点。 例如，您可以根据执行输入字段选择的多数模型生成“过滤”节点，此节点仅传递模型标识为“重要”的输入字段。

注: 在 IBM SPSS Modeler 的不同版本中执行时，给定模型为给定观测值指定的得分可能会有细小差别。 这通常是由于各个版本之间的软件增强所致。

使用模型块对数据进行评分

1. 将模型块连接到向其传递数据的数据源或流。
2. 将一个或多个处理或输出节点（如表或分析节点）添加或连接到模型块。
3. 执行模型块中的某个下游节点。

注：您无法使用“未优化规则”节点对数据进行评分。 要根据关联规则模型对数据进行评分，请使用“未优化规则”节点生成“规则集”模型块，然后使用“规则集”模型块进行评分。 有关更多信息，请参阅主题 第 198 页的『从关联模型块生成规则集』。

使用模型块生成处理节点

1. 在此选用板中浏览模型，或者在流工作区中编辑模型。

2. 在“模型块浏览器”窗口的“生成”菜单中选择所需节点类型。 可用选项将因模型块类型的不同而有所不同。 请查看具体的模型块类型，以详细了解您可从特定模型中生成的内容。

重新生成建模节点

如果要修改或更新某个模型块，但无法使用用于创建该模型的流，那么可以使用与创建原始模型相同的选项来重新生成一个建模节点。

要重新构建模型，右键单击模型选用板中的模型，然后选择**生成建模节点**。

此外，当浏览模型时，请选择“生成”菜单中的**生成建模节点**。

多数情况下，重新生成的建模节点应与创建原始模型的建模节点在功能上一致。

- 对决策树模型而言，还可以将交互式会话过程中的其他设置存储到节点，重新生成建模节点的过程中将启用 **使用树型指令** 选项。
- 对于决策列表模型而言，将启用 **使用保存的交互会话信息** 选项。 有关更多信息，请参阅主题 第 112 页的『决策列表模型选项』。
- 对于时间序列模型，将启用**使用现有模型继续估计**选项，通过该选项您可以使用当前数据重新生成先前的模型。

导入和导出 PMML 模型

PMML（也称为预测模型标记语言）是一种 XML 格式，用于描述数据挖掘和统计模型，包括模型的输入、用于为数据挖掘准备数据的变换，以及定义模型自身的参数。 IBM SPSS Modeler 可导入和导出 PMML，这使得其能够与其他支持此格式的应用程序（如 IBM SPSS Statistics）共享模型。

有关 PMML 的详细信息，请参阅数据挖掘组网站 (*http://www.dmg.org*)。

导出模型

PMML 导出支持大多数模型类型，这些模型类型生成在 IBM SPSS Modeler 中。 有关更多信息，请参阅主题 第 37 页的『支持 PMML 的模型类型』。

1. 右键单击模型选项板上的模型块。 （或者，双击工作区上的模型块并选择“文件”菜单。）
2. 在菜单上，单击**导出 PMML**。
3. 在“导出”（或“保存”）对话框中，指定此模型的目标目录及唯一名称。

注:

您可在“用户选项”对话框中更改 PMML 导出选项。 在主菜单中，单击:

工具 > 选项 > 用户选项

然后单击 PMML 选项卡。

导入以 PMML 格式保存的模型

以 PMML 格式从 IBM SPSS Modeler 或其他应用程序中导出的模型可以导入到模型选用板中。 有关更多信息，请参阅主题 第 37 页的『支持 PMML 的模型类型』。

1. 在模型选用板上，右键单击选用板并从菜单中选择**导入 PMML**。
2. 选择要导入的文件并根据需要为变量标签指定选项。
3. 单击**打开**。

如果模型中存在变量标签，请使用变量标签。 PMML 可为数据字典中的变量同时指定变量名和变量标签（例如，Referrer 标识，简称 *RefID*）。 如果在最初导出的 PMML 中存在变量标签，则选中此选项可以使用这些变量标签。

如果已选中变量标签选项但在 PMML 中没有变量标签，则按常规使用变量名。

支持 PMML 的模型类型

PMML 导出

IBM SPSS Modeler 模型。 可以将 IBM SPSS Modeler 中创建的以下模型导出为 PMML 4.3:

- C&R 树
- QUEST
- CHAID
- 神经网络
- C5.0
- Logistic 回归
- Genlin
- SVM
- Apriori
- Carma
- K-Means
- Kohonen
- TwoStep
- 二阶 AS
- GLMM（对于所有 GLMM 模型都会导出 PMML，但是 PMML 只有固定效应）
- 决策列表
- Cox
- 序列（不支持序列 PMML 模型评分）
- 随机树
- 树 AS
- 线性
- 线性 AS
- 回归
- Logistic
- GLE
- LSVM
- KNN
- 关联规则

数据库本机模型。 对于使用数据库本机算法生成的模型，PMML 导出不可用。 无法导出使用 Microsoft 的 Analysis Services 或 Oracle Data Miner 创建的模型。

PMML 导入

IBM SPSS Modeler 可以导入并评分由所有 IBM SPSS Statistics 产品的当前版本生成的 PMML 模型，包括从 IBM SPSS Modeler 导出的模型和由 IBM SPSS Statistics 17.0 或以后版本生成的模型或转换 PMML。 这实质上意味着评分引擎可评分的任何 PMML，以下除外:

- 无法导入 Apriori、CARMA、异常检测、序列和关联规则模型。
- 将 PMML 模型导入到 IBM SPSS Modeler 中后，虽然可以对其进行评分，但不能进行浏览。 （注意，其中包括最初从 IBM SPSS Modeler 中导出的模型。 为避免此限制，可将模型按生成的模型文件 (`*.gm`) 导出而不是按 PMML 导出。）

- 在导入时会执行有限的验证，但在试图对模型评分时会执行全面验证。 因此有可能导入成功，但评分却失败或产生不正确的结果。

注: 对于导入到 IBM SPSS Modeler 中的第三方 PMML，IBM SPSS Modeler 将尝试对可以识别并进行评分的有效 PMML 进行评分。 但是，无法保证将对所有 PMML 进行评分，也无法保证以应用程序生成 PMML 的方式对 PMML 进行评分。

为评分适配器发布模型

您可以将模型发布到安装有评分适配器的数据库服务器。 评分适配器可通过使用数据库的用户定义函数 (UDF) 功能在数据库中进行模型评分。 在数据库中进行评分可避免评分前提取数据的需求。 发布到评分适配器也将生成一些示例 SQL 以执行 UDF。

发布评分适配器

1. 双击模型块将其打开。
2. 从模型块菜单中选择:

   文件 > 针对 Server Scoring Adapter 发布
3. 填写对话框中的相关字段，然后单击**确定**。

数据库连接。 要为模型使用的数据库的连接详细信息。

发布标识。 （仅限于 Db2 for z/OS 数据库）模型的标识。 如果您重新构建同一模型并使用相同发布标识，那么生成的 SQL 也保持不变，所以无需更改使用之前生成的 SQL 的应用程序即可重新构建模型。 （对于其他数据库，生成的 SQL 对模型则是唯一。）

生成示例 SQL。 如果选择此项，将在**文件**字段中指定的文件中生成示例 SQL。

未优化模型

未优化模型包含从数据中抽取的信息，但并不用于直接生成预测。 即这些模型不能添加到流。 未优化模型在“已生成模型”选用板上显示为“未打磨的钻石”。

图 *27:* 未优化模型的图标

要查看未优化规则模型的详细信息，右键单击模型，然后选择上下文菜单中的 **浏览** 。 像其他在 IBM SPSS Modeler 中生成的模型一样，各种选项卡将提供所创建模型的相关概要和规则信息。

正在生成节点。 “生成“菜单允许您基于规则创建新节点。

- **选择节点**。 生成“选择”节点以选择要对其应用当前选定规则的记录。 如果未选择任何规则，此选项则禁用。
- **规则集**。 生成“规则集”节点以预测单个目标字段的值。 有关更多信息，请参阅主题 第 198 页的『从关联模型块生成规则集』。

第 4 章 筛选模型

筛选字段和记录

分析的预备阶段中可以使用多个建模节点来查找对建模最有用的字段和记录。 可使用特征选择节点来按照重要性筛选字段并为之排序，以及使用异常检测节点来查找不符合“正常”数据已知模式的异常记录。

“特征选择”节点根据一组条件（例如缺失值百分比）筛选要移除的输入字段，然后，相对于指定目标对余下的输入的重要性进行排秩。 例如，假如某个给定数据集有上千个潜在输入，那么哪些输入最有可能用于对患者结果进行建模呢?

Anomaly Detection 节点确定不符合“正常”数据格式的异常观测值（离群值）。 通过此节点，即使离群值不符合任何先前已知的模式，即使您并不确定要查找的内容，也可以识别这些离群值。

注意：异常检测并不考虑任何特定的目标（相关）字段，也不考虑这些字段是否与正在预测的模式相关，只是通过基于模型中所选字段集的聚类分析确定异常记录或观测值。 由于上述原因，您可能想将异常检测与特征选择或字段筛选和排秩的其他方法结合使用。 例如，您可以使用特征选择来确定与某个特定目标相关的最重要的字段，然后使用异常检测寻找针对这些字段而言最异常的记录。 （另外一个方法是构建一个决策树模型，然后将所有错误分类的记录视为可能的异常进行检查。 但是此方法很难用于进行大批量的复制和自动化。）

特征选择节点

数据挖掘问题可能包括成百甚至上千个可用作输入的备选字段。 从而花费大量的时间和精力来检查模型究竟应该包含哪些字段或变量。 为了缩小选择范围，可以使用特征选择算法来识别对某给定分析最为重要的字段。 例如，如果你试着根据多种因素来预测患者结果，那么哪些因素最为重要呢?

特征选择由以下三个步骤组成：

- **筛选**。 删除不重要或有问题的输入、记录或观测值（例如输入字段含有过多缺失值，或者输入字段的变异太大或太少而变得无用）。
- **分级**。 对剩余输入进行排序并根据重要性进行分级。
- **正在选择**。 确定要在后续模型中使用的功能子集，例如通过仅保留最重要的输入以及过滤或排除所有其他输入来进行确定。

当下，许多组织的数据均已超载，因此简化和加快建模过程是特征选择的根本优势。 通过将注意力迅速集中到最重要的字段上，可以降低所需的计算量，并且可以方便地找到因某种原因被忽略的小而重要的关系，最终获得更简单、精确和易于解释的模型。 通过减少模型中的字段数量，可以减少评分时间以及未来迭代中所收集的数据量。

示例。 某电话公司有一个数据仓库，其中包含 5000 名公司客户对某次特别促销活动的响应的相关信息。 数据包含有客户年龄、职业、收入、电话使用情况的统计数据等大量数据。 三个目标字段表示客户是否对三个报价做出响应。 该公司希望使用这些数据来帮助预测哪些客户最有可能在将来对类似报价做出响应。

需求。 单个目标字段（其角色设置为目标），以及要根据目标进行筛选或排序的多个输入字段。 目标和输入字段均具有连续（数值范围）或分类的测量级别。

特征选择模型设置

“模型”选项卡上的设置包含标准模型选项以及用于对输入字段筛选条件进行微调的设置。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

筛选输入字段

筛选就是剔除不提供关于输入/目标关系的任何有用信息的输入或观测值。 筛选选项只依据在问题中使用字段的属性，而不考虑该字段针对于选定目标字段的预测能力。 被筛选出来的字段将不参与有关输入排序的计算，同时还可选择将这些字段过滤掉，或是从用于建模的数据中删除。

可根据以下标准筛选字段：

- **缺失值的最大百分比**。 筛选具有过多缺失值的字段，以占记录总数的百分比表示。 缺失值百分比大的字段几乎不提供任何预测信息。
- **单个类别中记录的最大百分比**。 筛选相对于记录总数而言同一类别中具有过多记录的字段。 例如，如果数据库中 95% 的客户开同一类型的车，那么此信息无助于区分客户。 任何超过指定最大值的字段都将被筛选掉。 此选项仅适用于分类字段。
- **最大类别数（以记录的百分比表示）**。 筛选相对于记录总数而言具有过多类别的字段。 如果很高百分比的类别只含有一个观测值，那么该字段用处有限。 例如，如果每名客户都戴不同的帽子，那么此信息在建立行为模式模型时就不太可能有用。 此选项仅适用于分类字段。
- **最小变异系数**。 筛选变异系数小于或等于指定最小值的字段。 此度量值是输入字段标准偏差与输入字段均值之间的比值。 如果此值接近 0，那么变量值的变异性就不高。 此选项仅适用于连续（数字范围）字段。
- **最小标准差**。 筛选标准差小于或等于指定最小值的字段。 此选项仅适用于连续（数字范围）字段。

缺少数据的记录。 目标字段具有缺失值或所有输入都具有缺失值的记录或观测值将被从用于排秩的计算式中排除。

特征选择选项

“选项”选项卡用于指定在模型块中选择或排除输入字段的缺省设置。 然后可以将模型添加到流，以选择用于后续模型构建的字段子集。 或者，也可以通过在生成模型后在模型浏览器中选择或弃选其他字段，以覆盖这些设置。 但是，缺省设置下，无需更多修改即可应用模型块，这点在脚本编写方面特别有用。

有关更多信息，请参阅主题 第 41 页的『特征选择模型结果』。

可用选项有：

所有排名的字段。 根据字段的重要、边际或不重要排秩等级来选择字段。 可编辑每项排序的标签及用于指派记录的排秩等级的分界值。

前几个字段。 根据重要性选择前 *n* 个字段。

重要性大于。 选择重要性大于指定值的所有字段。

不管如何选择，目标字段总是被保留。

重要性排秩选项

所有分类。 当所有输入和目标均为分类字段时，可以根据以下任何一个测量对重要性进行排序：

- **Pearson 卡方**。 无需现有关系的强度或方向即可检验目标和输入的独立性。
- **似然比卡方统计**。 与 Pearson 卡方类似，也用于检验目标 - 输入的独立性。
- **Cramer V**。 基于 Pearson 卡方统计的关联度量。 值范围为 0 到 1，0 表示无关联，1 表示完全关联。
- **Lambda**。 这是反映变量用于预测目标值时误差降低比例的相关性测量。 值为 1 表示输入字段完美地预测了目标，值为 0 则表示输入未提供目标的任何有用信息。

部分分类。 当部分但并非所有输入为分类字段且目标也为分类字段时，可以根据 Pearson 或似然比卡方对重要性进行排序。 （除非所有输入均为分类变量，否则 Cramer's *V* 和 lambda 均不可用。）

分类与连续。 针对连续目标来为分类输入排秩或与之相反的情形时（即其中之一为分类字段，但不能两者均为分类字段），则使用 *F* 统计量。

均连续。 针对连续目标来为连续输入排秩时，将使用基于相关系数的 *t* 统计量。

特征选择模型块

“特征选择”模型块显示每个输入相对于选定目标的重要性（遵循“特征选择”节点的排序）。 排秩前已筛选掉的所有字段也将被列出。 有关更多信息，请参阅主题 第 39 页的『特征选择节点』。

运行含有特“特征选择”型块的流时，模型行为将如同过滤器，仅保留“模型”选项卡上当前选中的输入。 例如，可以选择评定为“重要”的所有字段（缺省选项之一）或在“模型”选项卡上手动选择一个字段子集。 不管如何选择，目标字段总是被保留。 所有其他字段将被排除。

过滤仅基于字段名称；例如，如果选择 年龄和 收入，那么匹配其中一个名称的任何字段都将被保留。 该模型不是基于新数据更新字段排秩，而只是根据选定的名称来过滤字段。 所以，将模型应用到新的或更新过的数据时应多加注意。 存有疑问时，最好重新生成模型。

特征选择模型结果

“特征选择”模型块的“模型”选项卡在顶部窗格中显示所有输入的排序和重要性，并且使您可以通过左侧栏中的复选框来选择用于过滤的字段。 运行流时，将只保留选定的字段；其他字段将被废弃。 缺省选择是基于模型构建节点中指定的选项，但可以根据需要选择或弃选其他字段。

底部窗格列出依据缺失值百分比或建模节点中指定的其他标准而从排秩中排除的输入。 与其他排序字段一样，可以通过左栏复选框来选择包含或丢弃这些字段。 有关更多信息，请参阅主题 第 39 页的『特征选择模型设置』。

- 要按秩、字段名称、重要性或任何其他显示的列对列表进行排序，请单击列标题。 如果要使用工具栏，那么可以从“排序方式”列表选择需要的项，并使用“向上”和“向下”箭头来更改排序方向。
- 您可以使用工具栏来选中或取消选中所有字段以及访问“选中字段”对话框，您可以通过该对话框根据排序或重要性来选择字段。 也可以按住 Shift 和 Ctrl 键并单击字段，以选择更多的字段，并使用空格键来切换选定的字段组。 有关更多信息，请参阅主题 第 41 页的『按照重要性选择字段』。
- 将输入排秩为“重要”、“边际”或“不重要”的阈值显示在表下方的图注中。 这些值是在建模节点中指定的。 有关更多信息，请参阅主题 第 40 页的『特征选择选项』。

按照重要性选择字段

使用特征选择模型块对数据进行评分时，由排序或筛选字段选中的所有字段都将被保留，如左栏复选框所示。 其他字段将被丢弃。 要更改选择，您可以使用工具栏访问“选中字段”对话框，该对话框使您可以根据排序或重要性来选择字段。

所有标记的字段。 选择标记为“重要”、“边际”和“不重要”的所有字段。

前几个字段。 允许您根据重要性来选择前 *n* 个字段。

重要性大于。 选择重要性大于指定阈值的所有字段。

从特征选择模型中生成过滤器

根据“特征选择”模型的结果，您可以使用“根据特征生成过滤”对话框来生成一个或多个“过滤”节点，该节点根据相对于指定目标的重要性包含或排除字段子集。 虽然模型块也可以用于过滤，但使用此方法可以在不复制或不修改模型的情况下自由地尝试不同的字段子集。 不管是选择包含还是选择排除，过滤时将总是保留目标字段。

包含/排除。 您可以选择包括或排除字段，例如包括前 10 个字段或排除所有标记为“不重要”的字段。

选定字段。 包括或排除表中当前选定的所有字段。

所有已标记字段。 选择标记为“重要”、“边际”和“不重要”的所有字段。

前几个字段。 允许您根据重要性来选择前 *n* 个字段。

重要性大于。 选择重要性大于指定阈值的所有字段。

异常检测节点

异常检测模型用于识别数据中的离群值或异常观测值。 与存储有关异常观测值的规则的其他建模方法不同，异常检测模型存储有关正常行为的信息。 因此即使在离群值不符合任何已知模式的情况下，异常检测模型也使识别离群值成为可能，在新模式可能不断涌现的应用（如缺陷检测）中，该模型可能尤其有用。 异常检测是一种不受监督的方法，这就意味着它不需要包含已知缺陷观测值的训练数据集作为开始点。

识别离群值的传统方法通常是一次检查一个或两个变量，而异常检测可以检查大量字段以识别相似记录所属的聚类或对等组。 然后，可将每条记录与其对等组中的其他记录进行比较，以识别出可能的异常值。 观测值与正常中心值离得越远，它越有可能是异常观测值。 例如，该算法可能会将记录聚合为三个不同的聚类，并对离任何一个聚类的中心值较远的那些记录进行标记。

每条记录都指定了一个异常指数，该指数是组偏差指数与该观测值所属聚类中平均值的比。 此指数的值越大，观测值与平均值的偏差就越大。 通常情况下，异常指数值小于 1 甚至小于 1.5 的观测值都不会被视为异常值，因为该偏差与平均值相同或者只是大一点。 但是，指数值大于 2 的观测值有可能是异常观测值，因为该偏差至少是平均值的两倍。

异常检测是一种探索性方法，它是为对应该进行进一步分析的可能异常观测值或记录进行快速检测而设计的。 这些观测值应视为 疑似 异常值，在进行进一步检查后，可以证明它们是或不是真正的异常值。 您可能会发现某个记录完全有效，但无法选择从数据中将其筛选出来用于模型构建。 另外，如果算法重复检测出虚假异常值，那么可能表示数据收集过程中存在错误或假象。

注意：异常检测并不考虑任何特定的目标（相关）字段，也不考虑这些字段是否与正在预测的模式相关，只是通过基于模型中所选字段集的聚类分析确定异常记录或观测值。 由于上述原因，您可能想将异常检测与特征选择或字段筛选和排秩的其他方法结合使用。 例如，您可以使用特征选择来确定与某个特定目标相关的最重要的字段，然后使用异常检测寻找针对这些字段而言最异常的记录。 （另外一个方法是构建一个决策树模型，然后将所有错误分类的记录视为可能的异常进行检查。 但是此方法很难用于进行大批量的复制和自动化。）

示例。 对农业发展补贴进行审查以确定是否可能存在内部欺诈观测值时，异常检测可用于发现有悖于标准值的偏差，并突出显示值得进一步调查的异常记录。 特别值得关注的是那些相对农场类型和规模而言似乎申请了过多（或过少）补助金的补贴申请。

需求。 一个或多个输入字段。 请注意，只有其角色使用源节点或“类型”节点设置为**输入**的字段才能用作输入。 目标字段（角色设置为**目标**或**两者**）将被忽略。

强度。 通过标记不符合已知规则集（而不是符合已知规则集）的观测值，异常检测模型可以确定异常观测值，即使这些观测值不符合先前已知的模式也是如此。 与特征选择结合使用时，异常检测可用于筛选大量数据，以便更快地确定相对最需要关注的记录。

异常检测模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

确定异常分界值的基于方法。 指定用于确定分界值以标记异常的方法。 可用选项有：

- **最小异常指标级别**。 指定用于标记异常的最小分界值。 达到或超过此阈值的记录将进行标记。
- **训练数据中最异常记录的百分比**。 用于自动设置一个阈值，其级别标记为训练数据中记录的指定百分比。 所生成的分界值作为参数包含在模型中。 请注意，此选项确定了分界值的设置方式，而并非确定评分期间要标记的记录的实际百分比。 实际评分结果可能根据数据的不同而有所变化。
- **训练数据中最异常记录的数量**。 用于自动设置一个阈值，其级别标记为训练数据中的指定记录数。 所生成的临界值作为参数包含在模型中。 请注意，此选项确定了分界值的设置方式，而并非确定评分期间要标记的记录的具体数目。 实际评分结果可能根据数据的不同而有所变化。

注：无论如何确定分界值，这都不会影响对每条记录报告的潜在异常指标值。 它仅在对模型进行估算和评分时指定用于将记录标记为异常的阈值。 如果您稍后要检查较大数量或较小数量的记录，那么可以使用“选择”节点以根据异常指标值 (`$O-AnomalyIndex > X`) 来标识记录子集。

要报告的异常字段数。 指定要报告的字段数，用于指示将特定记录标记为异常的原因。 将报告最异常字段，此类字段定义为与记录所分配到的聚类的字段标准值偏差最大的字段。

异常检测专家选项

要指定缺失值和其他设置的选项，请在“专家”选项卡上将模式设置为**专家**。

调整系数。 用于平衡计算距离时指定给连续（数字范围）和分类字段的相对权重的值。 值越大，对连续字段的影响也越大。 它必须为非零值。

自动计算对等组数。 异常检测可用于快速分析大量可行的解决方案，以选择训练数据的最佳对等组数。 可通过设置对等组的最大数和最小数来扩大或缩小范围。 较大的值将使系统可以探究更多的可行解决方案，但相应的代价是处理时间增加。

指定对等组数。 如果您知道要在模型中包含的聚类数，请选择此选项并输入对等组数。 通常，选择此选项可提高性能。

噪声级别和比率。 这些设置用于确定二阶聚类期间离群值的处理方式。 在第一阶段中，使用聚类特征 (CF) 树将数据从大量单项记录浓缩成数量可管理的聚类。 该树基于相似性度量构建，并且当树的某个节点中记录过多时，它会分割为子节点。 在第二阶段中，将从 CF 树的终端节点开始创建分层聚类。 噪声处理在第一次数据传递时开启，并在第二次数据传递时关闭。 第一次数据传递时，噪声聚类中的观测值将分配给第二次数据传递中的常规聚类。

- **噪声级别**。 请指定介于 0 到 0.5 之间的值。 只有在下列情况下此设置才相关：CF 树在增长阶段进行填充，即该树不再接收叶节点中的观测值并且叶节点无法分割。

  如果 CF 树进行填充并且噪声级别设置为 0，那么阈值将增大并且 CF 树将使用所有观测值重新生长。 最终聚类之后，不能分配到聚类的变量标记为离群值。 将对离群值聚类指定标识号 -1。 离群值聚类不包括在聚类数的计数中；即，如果您指定 *n* 个聚类和噪声处理，那么算法将输出 *n* 个聚类和一个噪声聚类。 实际上，增大此值可使算法在将异常记录纳入树中时有更大余地，而不是将它们分配给单独的离群值聚类。

  如果 CF 树进行填充并且噪声级别大于 0，那么在将稀疏叶片中的任何数据放入其自身的噪声叶片中之后，该 CF 树会重新生长。 如果稀疏叶片中的观测值数与最大叶片中的观测值数的比率小于噪声级别，那么认为该叶片是稀疏叶片。 在树生长完成后，系统会在可能的情况下将离群值放入 CF 树中。 如果未放入 CF 树中，那么对于第二阶段聚类，将废弃离群值。
- **噪声比率**。 指定分配给组件的应该用于噪声缓存的内存量。 此值必须介于 0.0 到 0.5 之间。 如果将特定观测值插入树的叶片中之后，所产生的紧性小于阈值，那么叶片将不再分割。 如果紧性超过阈值，那么叶片将进行分割，同时将另一个小聚类添加至 CF 树。 实际上，增大此设置可能会导致算法更快速地向较简单的树倾斜。

插补缺失值(I)。 对于连续字段，请用字段均值替换缺失值。 对于分类字段，多个缺失值类别将进行组合并被视为一个有效类别。 如果取消选中此选项，那么将从分析中排除任何具有缺失值的记录。

异常检测模型块

异常检测模型块包含异常检测模型所捕获的所有信息以及有关训练数据和估算过程的信息。

运行包含异常检测模型块的流时，多个新字段将按照模型块中“设置”选项卡上的选择添加至流。 有关更多信息，请参阅主题 第 44 页的『异常检测模型设置』。 新字段名称基于模型名称，并带有前缀 *$O*，下表对这些名称进行了概述。

表 *6:* 新字段名称的生成

字段名称	**描述**
$O-Anomaly	指示记录是否异常的标志字段。
$O-AnomalyIndex	记录的异常索引值。
$O-PeerGroup	指定记录分配给哪个对等组。
$O-Field-n	与聚类标准值偏差相关的第 *n* 个最异常字段的名称。
$O-FieldImpact-n	字段的变量偏差指数。 此值用于度量与记录分配到的聚类字段标准值的偏差。

也可以选择抑止非异常记录的评分，以使结果更易于读取。 有关更多信息，请参阅主题 第 44 页的『异常检测模型设置』。

异常检测模型详细信息

所生成的异常检测模型的“模型”选项卡显示模型中对等组的相关信息。

请注意，所报告的对等组大小和统计信息是基于训练数据的估算值，并且可能与实际评分结果略有不同，即使对相同数据运行也是如此。

未报告原因的残差是 1 减去识别为异常的记录的每个异常列的平均异常指标值之和。 此百分比可用于指示报告的字段在多大程度上解释了异常。 此可以指导您确定要报告的异常字段数。

异常检测模型摘要

异常检测模型块的“摘要”选项卡显示字段、构建设置和估算过程的相关信息。 另外，还显示了对等组数以及用于将记录标记为异常的分界值。

异常检测模型设置

使用“设置”选项卡可以指定用于对模型块进行评分的选项。

异常记录指示方法 指定在输出中处理异常记录的方式。

- **标志和指标** 创建标志字段，对于模型中包含的所有超过分界值的记录，此字段设置为 *True*。 另外，将报告另一个字段中每条记录的异常指标。 有关更多信息，请参阅主题 第 42 页的『异常检测模型选项』。
- **仅标志** 创建标志字段，但不报告每条记录的异常指标。
- **仅指标** 报告异常指标但不创建标志字段。

要报告的异常字段数：指定要报告的字段数，用于指示将特定记录标记为异常的原因。 将报告最异常字段，此类字段定义为与记录所分配到的聚类的字段标准值偏差最大的字段。

废弃记录 选择此选项可废弃流中的所有**非异常**记录，从而更容易地专注于任何下游节点中的潜在异常。 另外，您也可以丢弃所有**异常**记录，以便将后续分析限制为那些未根据模型标记为潜在异常的记录。

注: 由于取整造成的细微差异，评分期间标记的实际记录数可能与训练模型时标记的记录数不同，即使对相同数据运行也是如此。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

第 5 章 自动建模节点

自动建模节点对多种不同的建模方法进行估算和比较，这使您可以在一次建模运行中尝试多种方法。 您可以选择所使用的建模算法，以及每个建模算法的具体选项，包括可能互斥的组合。 例如，您无需为神经网络选择快速、动态或修剪之中的某个方式，完全可以全部尝试。 节点研究选项的每个可能组合，根据您指定的测量为每个候选模型排序，并保存最佳模型用于评分或将来的分析。

您可以根据分析需要从三个自动建模节点中进行选择：

“自动分类器”节点用于创建和对比二元结果（是或否，流失或不流失等）的若干不同模型，使用户可以选择给定分析的最佳处理方法。 由于支持 多种建模算法，因此可以对用户希望使用的方法、每种方法的特定选项以及对比结果的标准进行选择。 该节点根据指定的选项生成一组模型，并根据您指定的条件对最佳候选项进行排序。

自动数字节点使用多种不同方法估计和对比模型的连续数字范围结果。 此节点和自动分类器节点的工作方式相同，因此可以选择要使用和要在单个建模传递中使用多个选项组合进行测试的算法。 受支持的算法包括神经网络、C&R 树、CHAID、线性回归、广义线性回归以及支持向量机 (SVM)。 可以根据相关度、相对误差或使用的变量数来比较模型。

“自动聚类”节点估算和比较识别具有类似特征记录组的聚类模型。 节点工作方式与其他自动建模节点相同，使您在一次建模运行中即可试验多个选项组合。 可使用基本度量对模型进行比较，尝试对聚类模型进行过滤，对其有用性进行排名，并提供基于特定字段重要性的度量。

最佳模型保存在一个组合模型块中，可对其进行浏览和比较，并选择评分中使用的模型。

- 只有对于二元、名义和数字目标，您才可以选择多个评分模型，并将评分组合在一个模型整体中。 通过结合多个模型的预测，可以避免单个模型的局限性，使所得的整体准确性通常比从任一模型中获得的准确性要高。
- 您还可以选择向下钻取结果，或为要使用或进一步探索的所有单独模型生成建模节点或模型块。

模型和执行时间

根据模型的数据集和数量，自动建模节点执行时间可能为数小时或甚至更长。 在选择选项时，请注意正在生成的模型个数。 如果现实条件允许，您可能希望将建模运行的时间安排在夜晚或周末，因为此时对系统资源的需求可能比较小。

- 必要的话，可以使用分区节点或样本节点减少包括在初始训练传递中的记录数。 一旦将选择限制在几个生成的候选模型内，就可以恢复全部数据集。
- 要减少输入字段数，请使用特征选择。 有关更多信息，请参阅主题 第 39 页的『特征选择节点』。 另外，您可以使用初始建模运行来识别需要进一步探索的字段和选项。 例如，如果性能最佳的模型似乎都使用了相同的三个字段，那么有力地说明这些字段值得保留。
- 您还可以限制评估任一模型所需的时间并且指定用于过滤和排序模型的评估尺度。

自动建模节点算法设置

对于每个模型类型，可以使用缺省设置，或为每个模型类型选择选项。 这些特定选项类似于独立建模节点中可用的选项，不同之处在于并非只能选择一种设置而是大多数情况下可以根据应用需要选择多种。 例如，如果对比神经网络模型，可以选择几种不同的训练方法，并且在使用随机种子和不使用随机种子的情况下尝试每种方法。 选定选项的所有可能组合都将使用，从而使得在单次遍历中生成许多不同模型变得更容易。 但是，使用时要小心，因为选择多个设置会引起模型数非常快速地增加。

要为每个模型类型选择选项：

1. 在自动建模节点上，选择**专家**选项卡。

2. 单击模型类型的**模型参数**列。
3. 从下拉菜单中，选择**指定**。
4. 在**算法设置**对话框上，从**选项**列中选择选项。

注：**算法设置**对话框的“专家”选项卡上提供了更多选项。

自动建模节点停止规则

为自动建模节点指定的停止规则不仅与节点所构建的个别模型的停止有关，还与所有节点执行有关。

限制总执行时间。（仅限神经网络、K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和 C&R 模型）在指定小时数后停止执行。所有在该时间点之前（包括该点）生成的模型都将包括在模型块中，但这之后不会再生成模型。

在生成有效模型后立即停止。当模型传递了所有在“丢弃”选项卡（自动分类器或自动聚类节点的）和“模型”选项卡（自动数值节点的）上指定的标准时将停止执行。有关更多信息，请参阅主题 第 50 页的『自动分类器节点丢弃选项』。有关更多信息，请参阅主题 第 56 页的『自动聚类节点丢弃选项』。

自动分类器节点

“自动分类器”节点使用多种不同的方法来估算和比较名义（集合）或二元（是/否）目标的模型，这使您可以在一次建模运行中尝试多种方法。您可以选择所用算法，并试验选项的多个组合。例如，您无需在径向基函数、多项式、sigmoid 或线性方法中选择一种来用于 SVM，您可以全部都尝试一下。该节点将探究每种可能的选项组合，并根据您指定的测量对每个候选模型进行排序，然后保存最佳模型以用于评分或进行进一步分析。有关更多信息，请参阅 第 45 页的『第 5 章 自动建模节点』。

示例

某零售公司具有历史数据，可用于追踪以前营销活动中向特定客户提供的商品推荐信息。公司现在希望通过向每个客户提供合适的报价来获取更多的利润。

要求

一个测量级别为名义或标志（角色设置为**目标**）的目标字段和至少一个输入字段（角色设置为**输入**）。对于“标志”字段，假定为目标字段定义的真值表示计算利润、提升和相关统计量时的匹配项。输入字段的测量级别可以是连续或分类，但具有限制，即某些输入可能不适合一些模型类型。例如，在 C&R 树、CHAID 和 QUEST 模型中用作输入的有序字段必须是数字存储类型（而不是字符串），如果指定了其他类型，将被这些模型忽略。类似地，在某些情况下可对连续输入字段进行分级。这和使用单个建模节点时的要求一样；例如，不管是从贝叶斯网络节点还是自动分类器节点生成，贝叶斯网络模型都以同样的方式工作。

频率和权重字段

频率和权重用于为某些记录提供高于其他记录的附加重要性，原因可能是用户知道构建数据集省略父总体的一部分（加权）或一个记录代表一些相同的观测值（频率）等。如果指定了频率字段，那么 C&R 树、CHAID、QUEST、决策列表和贝叶斯网络模型可以使用此字段。C&RT、CHAID 和 C5.0 模型可以使用权重字段。其他模型类型将省略这些字段并以任意方式构建模型。频率和权重字段仅用于模型构建，并且在评估和评分模型时不予以考虑。有关更多信息，请参阅 第 24 页的『使用频率和权重字段』。

前缀

如果您将表节点附加到自动分类器节点块，那么表中存在多个名称以前缀 $ 开头的新变量。

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。不同的模型类型使用不同的前缀集。

例如，前缀 $G、$R 和 $C 分别用作广义线性模型、CHAID 模型和 C5.0 模型生成的预测的前缀。$X 通常是使用整体生成的，如果目标字段为“连续”、“分类”或“标志”字段，那么分别使用 $XR、$XS 和 $XF 作为前缀。

$...C 前缀用于“分类”或“标志”目标的预测置信度；例如，$XFC 用作整体标志预测置信度的前缀。$RC 和 $CC 分别为 CHAID 模型和 C5.0 模型的单个置信度预测的前缀。

支持的模型类型

支持的模型类型包括神经网络、C&R 树、QUEST、CHAID、C5.0、Logistic 回归、决策列表、贝叶斯网络、判别、最近邻元素、SVM、XGBoost Tree 和 XGBoost-AS。 有关更多信息，请参阅主题 第 48 页的『自动分类器节点专家选项』。

连续机器学习

建模的不便之处在于，由于随时间推移对数据的更改，模型会变得过时。 这通常称为模型漂移或概念漂移。为了有效地帮助克服模型漂移，SPSS Modeler 提供了连续的自动化机器学习。 此功能可用于“自动分类器”节点和“自动数值”节点模型块。

自动分类器节点模型选项

通过“自动分类器”节点的“模型”选项卡，您可以指定要创建的模型数以及用于比较模型的标准。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

交叉验证。 交叉验证可以为模型提供已知数据的数据集（即训练数据集，用于运行训练），以及未知数据的数据集（即验证数据集或测试集，用于测试模型）。 交叉验证的目标是测试模型预测未用于模型估计的新数据的能力，以指出过度拟合或选择偏差之类的问题。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

模型评级依据。 指定用于比较和排序模型的标准。 选项包括总体精确性、ROC 曲线下的区域、利润、提升和字段的数量。 请注意，无论在此处选定哪些尺度，所有这些尺度都能在汇总报告中使用。

注: 对于名义（设置）目标，排秩限制为**总体准确性**或**字段数**。

计算利润、提升和相关统计量时，将假定目标字段定义为 真 值以表示匹配项。

- **总体准确性**。 模型正确预测出的记录相对于记录总数的百分比。
- **ROC 曲线下面积**。 ROC 曲线提供模型的性能指标。 曲线位置距参考线越远，则检验准确度越高。
- **利润（累积）**。 根据指定的成本、收入和权重标准计算出的各个累积百分位数（按预测的置信度排序）的利润总和。 通常，顶部百分位数的初始利润接近于零，然后逐步增加，最后减少。 对于构建完好的模型，利润将显示一个正确定义的峰值，并在峰值位置报告百分位数。 对于不包含任何信息的模型，利润曲线将相对较直，可能显示为增加、减少或保持水平，具体取决于所采用的成本/收入结构。
- **增益（累积）**。 相对于整个样本（其中分位数按预测的置信度排序）的累积分位数匹配率。 例如，顶部分位数提升值 3 表示其匹配率是整个样本的三倍。 对于构建完好的模型，顶部分位数应从稍高于 1.0 的位置开始提升，然后径直向分位数下限 1.0 跌落。 对于不包含任何信息的模型，提升将在 1.0 上下徘徊。
- **字段数**。 基于所用输入字段的数量对模型进行排序。

模型评级方式。 如果正在使用分区，那么可以指定根据训练数据集还是检验数据集进行排序。 对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。

要使用的模型数。 指定要在节点生成的模型块中列出的最大模型数。 按照指定的排秩标准将顺次列出排秩靠前的模型。 注意，增大此限制会降低性能。 允许的最大值为 100。

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，预测变量重要性可能会增加计算某些模型所需的时间，如果仅仅希望对许多不同的模型进行广泛对比，那么不建议评估变量重要性。 将分析限制在要进一步探索的几个模型上会更有用。 有关更多信息，请参阅 第 32 页的『预测变量重要性』。

计算整体分布图。 控制是否将整体分布图包括在生成的自动模型输出中。 关闭时，将提高自动建模性能。

利润标准。 仅适合标志目标。 利润等于每条记录的收入减去该记录的成本。 也就是说，分位数的利润就是位于该分位数内的所有记录的利润总和。 这里假定利润仅应用于匹配项，但成本可应用于所有的记录。

- **成本**。 指定与每个记录关联的成本。 您可以选择 **固定** 或 **可变** 成本。 对于固定成本，请指定成本值。 对于可变成本，请单击“字段选择器”按钮，将某个字段选择为成本字段。 （**成本**不适用于 ROC 图表。）
- **收入**。 指定与表示匹配项的每个记录关联的收入。 您可以选择 **固定** 或 **可变** 成本。 对于固定收入，请指定收入值。 对于可变收入，请单击“字段选择器”按钮，将某个字段选择为收入字段。 （**收入**不适用于 ROC 图表。）
- **权重**。 如果数据中的记录代表多个单元，那么可以使用频率权重来调整结果。 使用**固定**或**可变**加权，指定与每个记录关联的加权。 对于固定加权，请指定加权值（每个记录的单元数）。 对于可变加权，请单击“字段选择器”按钮，将某个字段选择为权重字段。 （**权重**不适用于 ROC 图表。）

增益标准。 仅适合标志目标。 指定提升计算使用的百分位数。 注意，在比较结果时也可以更改此值。 有关更多信息，请参阅主题 第 56 页的『自动模型块』。

自动分类器节点专家选项

通过“自动分类器”节点的“专家”选项卡，您可以应用分区（如果可用），选择要使用的算法以及指定停止规则。

选择模型。 缺省情况下，将选中所有模型进行构建；但是，如果您拥有 Analytic Server，那么您可以选择将模型限制为能够在 Analytic Server 上运行的模型，并对模型进行预设，使其构建拆分模型，或准备好处理超大型数据集。

注: 不支持在“自动分类器”节点中本地构建 Analytic Server 模型。

使用的模型。 使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。 选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。 列出可用的算法（请参阅下面的内容）。

模型参数。 对于每个模型类型，可以使用缺省设置，或选择 **指定** 为每个模型类型选择选项。 这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。 例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数量。 列出根据当前设置为每个算法生成的模型的数目。 当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制构建单个模型所花费的最长时间。 （仅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。 例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，那么您大概不希望它使得整个的建模运行停滞。

注: 如果目标为名义（集合）字段，那么“决策列表”选项不可用。

支持的算法

使用支持向量机 (SVM) 节点，可以将数据分为两组，而无需过度拟合。 SVM 可以与宽数据集配合使用，例如那些含有大量输入字段的数据集。

The *k*-最近相邻元素 (KNN) 节点将新的观测值关联到预测变量空间中与其最邻近的 *k* 个对象的类别或值（其中 *k* 为整数）。 类似观测值相互靠近，而不同观测值相互远离。

判别分析提出比 Logistic 回归更加严格的假设，但是在满足这些假设时可成为 Logistic 回归的有价值替代方案或补充。

通过贝叶斯网络节点，你可以利用对真实世界认知的判断力并结合所观察和记录的证据来构建概率模型。 该节点侧重于主要用于分类的树增强朴素贝叶斯 (TAN) 和马尔可夫毯网络。

决策列表节点可标识子组或段，显示与总体相关的给定二元结果的似然度的高低。 例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。 通过定制段和并排预览备选模型来比较结果，您可以将自己的业务知识体现在模型中。 决策列表模型由一组规则构成，其中每个规则具备一个条件和一个结果。 规则依顺序应用，相匹配的第一个规则将决定结果。

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。 它与线性回归类似，但采用分类目标字段而不是数字范围。

CHAID 使用卡方统计来生成决策树，以确定最佳的分割。 CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。 目标和输入字段可以是数字范围（连续）或分类。 穷举 CHAID 是 CHAID 的修正版，它可以更彻底地检查所有可能的拆分，但计算时间较长。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。 输入字段可以是数字范围（连续），但目标字段必须是分类。 所有分割都是二元的。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。 该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果树中某个节点中 100% 的观测值都属于目标字段的一个特定类别，那么该节点将被认定为“纯洁”。 目标和输入字段可以是数字范围或分类（名义、有序或标志）； 所有拆分都是二进制的（只有两个子组）。

C5.0 节点构建决策树或规则 集。 该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。 目标字段必须为分类字段。 允许进行多次多于两个子组的分割。

神经网络节点使用的模型是对人类大脑处理信息的方式简化了的模型。 此模型通过模拟大量类似于神经元的抽象形式的互连简单处理单元而运行。 神经网络是功能强大的一般函数估计器，只需要最少的统计或数学知识就可以对其进行训练或应用。

线性回归模型根据目标与一个或多个预测变量之间的线性关系来预测连续目标。

通过线性支持向量机 (LSVM) 节点，您可以将数据分为两组，而无需过度拟合。 LSVM 是线性的，并且可以与大量数据集配合使用，例如包含大量记录的数据集。

“随机树”节点与现有 C&RT 节点相似；但是，“随机树”节点旨在处理大数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示产生的模型。 “随机树”节点将生成您可以对未来观测值进行预测或分类的决策树。 通过在每个步骤最大限度降低不纯洁度，此方法使用递归分区将训练记录分割为多个段。如果树中某个节点的全部观测值都属于目标字段的一个特定类别，那么系统会将该节点视为纯洁。 目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。

树 AS 节点类似于现有的 CHAID 节点；但是，“树 AS”节点旨在处理大量数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示生成的模型。 此节点通过使用卡方统计 (CHAID) 来识别最优拆分，从而生成决策树。 对 CHAID 的这一使用可生成非二元树，意味着某些拆分将具有两个以上的分支。 目标和输入字段可以是数字范围（连续）或分类。 Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。

XGBoost Tree© 是将树模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost Tree 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost Tree 节点仅显示了核心功能和常用参数。 此节点使用 Python 进行实现。

XGBoost© 是实现提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost-AS 节点仅显示了核心功能和常用参数。 在 Spark 中实现 XGBoost-AS 节点。

注: 如果选择“树-AS”以在 Analytic Server 上运行，当存在“分区”节点上游时，它将无法构建模型。 在此情况下，为使“自动分类器”能够与 Analytic Server 上的其他建模节点一起工作，请取消选择“树-AS”模型类型。

误分类成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

除 C5.0 模型之外，在对模型进行评分时，误分类成本是不适用的；在使用自动分类器节点、评估图表或分析节点对模型进行排秩或比较时，误分类成本也不予以考虑。 将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的 成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

自动分类器节点丢弃选项

通过“自动分类器”节点的“废弃”选项卡，您可以自动废弃不符合特定标准的模型。 这些模型将不会列在汇总报告中。

可以为总准确性指定最小阈值，为模型中使用的变量数指定最大阈值。 此外，对于标志目标，可以为提升、利润和曲线下方区域指定最小阈值，提升和利润由在“模型”选项卡上指定的内容所确定。 有关更多信息，请参阅主题 第 47 页的『自动分类器节点模型选项』。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关更多信息，请参阅主题 第 46 页的『自动建模节点停止规则』。

“自动分类器”节点设置

在“自动分类器”节点的**设置**选项卡上，可以预先配置模型块上可用的“评分-时间”选项。

过滤掉由整体模型生成的字段。 从输出中移除由各个模型生成的所有附加字段，这些模型均输入到“整体”节点中。 如果只想关注所有输入模型中的综合评分，请选中此复选框。 如果希望使用分析节点或评估节点将综合评分的准确性与各个输入模型评分的准确性进行比较，则请确保取消选中此选项。

提示: 持续机器学习设置还可用于“自动分类器”节点和“自动数值”节点。

自动数值节点

“自动数值”节点使用多种不同方法来估算和比较模型以得出连续数值范围结果，这使您可以在一次建模运行中尝试多种方法。 您可以选择所用算法，并试验选项的多个组合。 例如，您可以使用神经网络、线性回归、C&RT 和 CHAID 模型预测住房价值，以确定哪种模型的性能最好，并且可以尝试步进、向前和向后回归法的不同组合。 节点研究选项的每个可能组合，根据您指定的测量为每个候选模型排序，并保存最佳模型用于评分或将来的分析。 有关更多信息，请参阅主题 第 45 页的『第 5 章 自动建模节点』。

示例

市政当局需要更准确地估计房地产税以及无需检查每个属性就可以按需要调整特定属性的值。 通过使用“自动数值”节点，分析人员可以生成并比较多个模型，这些模型根据构建类型、近邻、大小和其他已知因素来预测属性值。

要求

一个目标字段（角色设置为**目标**）和至少一个输入字段（角色设置为**输入**）。 目标必须为连续（数值范围）字段，如年龄或收入。 输入字段可以是连续或分类，但具有限制，即某些输入可能不适合一些模型类型。 例如，C&R 树模型能将分类字符串字段作为输入使用，而线性回归模型不能使用这些字段并将在指定这些字段后省略它们。 这和使用单独建模节点时的要求相同。 例如，不管 CHAID 模型是在 CHAID 节点中还是在自动数值节点中生成，其工作方式都相同。

频率和权重字段

频率和权重用于为某些记录提供高于其他记录的附加重要性，原因可能是用户知道构建数据集省略父总体的一部分（加权）或一个记录代表一些相同的观测值（频率）等。 如果指定频率字段，那么 C&R 树和 CHAID 算法可以使用该字段。 C&RT、CHAID 回归和 GenLin 算法可以使用权重字段。 其他模型类型将省略这些字段并以任意方式构建模型。 频率和权重字段仅用于模型构建，并且在评估和评分模型时不予以考虑。 有关更多信息，请参阅主题 第 24 页的『使用频率和权重字段』。

前缀

如果您将表节点附加到自动数字节点块，那么表中存在多个名称以前缀 $ 开头的新变量。

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。 不同的模型类型使用不同的前缀集。

例如，前缀 $G、$R 和 $C 分别用作广义线性模型、CHAID 模型和 C5.0 模型生成的预测的前缀。$X 通常是使用整体生成的，如果目标字段为"连续"、"分类"或"标志"字段，那么分别使用 $XR、$XS 和 $XF 作为前缀。

$...E 前缀用于连续目标的预测置信度；例如，$XRE 用作整体连续预测置信度的前缀。$GE 是广义线性模型的单一置信度预测的前缀。

支持的模型类型

支持的模型类型包括神经网络、C&R 树、CHAID、回归、GenLin、最近相邻元素、SVM、XGBoost Linear、GLE 和 XGBoost-AS。 有关更多信息，请参阅 第 52 页的『自动数值节点专家选项』。

连续机器学习

建模的不便之处在于，由于随时间推移对数据的更改，模型会变得过时。 这通常称为模型漂移或概念漂移。为了有效地帮助克服模型漂移，SPSS Modeler 提供了连续的自动化机器学习。 此功能可用于“自动分类器”节点和“自动数值”节点模型块。

自动数值节点模型选项

通过“自动数值”节点的“模型”选项卡，您可以指定要保存的模型数以及用于比较模型的标准。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

交叉验证。 交叉验证可以为模型提供已知数据的数据集（即训练数据集，用于运行训练），以及未知数据的数据集（即验证数据集或测试集，用于测试模型）。 交叉验证的目标是测试模型预测未用于模型估计的新数据的能力，以指出过度拟合或选择偏差之类的问题。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

模型评级依据。 指定用于比较模型的标准。

- **相关性**。 这是每条记录的观测值和模型预测的值之间的 Pearson 相关性。 相关性是两种变量之间的线性关联尺度，值越接近 1 说明变量之间的关系越强。 （相关性的值在 -1 和 +1 之间，-1 代表完全负关系，+1 代表完全正关系。 值为 0 表示无线性关系，但具有负相关性的模型将排在最后。）
- **字段数**。 模型中用作预测变量的字段的数目。 在某些情况下，选择使用较少字段的模型可简化数据准备过程并提高性能。
- **相对误差**。 相对误差是模型观测值相对于预测值的方差与观测值相对于平均值的方差的比率。 在实际应用的角度，它对比模型相对于 **空** 或 **截距** 模型（仅返回目标字段的平均值作为预测值）的性能。 对于好的模型，此值应小于 1，说明此模型比空模型更精确。 相对错误大于 1 的模型不如空模型精确，因此这样的模型没有意义。 对于线性回归模型，相对错误等同于相关性的平方并且未添加任何新的信息。 对于非线性模型，相对错误与相关性无关并且为评估模型性能提供了附加尺度。

模型评级方式。 如果正在使用分区，那么可以指定根据训练分区还是检验分区进行排序。 对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。

要使用的模型数。 指定要在节点生成的模型块中显示的最大模型数。 按照指定的排秩标准将顺次列出排秩靠前的模型。 通过增加此限制，您可以对比更多模型的结果，但这可能会降低性能。 允许的最大值为 100。

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，预测变量重要性可能会增加计算某些模型所需的时间，如果仅仅希望对许多不同的模型进行广泛对比，那么不建议评估变量重要性。 将分析限制在要进一步探索的几个模型上会更有用。 有关更多信息，请参阅 第 32 页的『预测变量重要性』。

计算整体分布图。 控制是否将整体分布图包括在生成的自动模型输出中。 关闭时，将提高自动建模性能。

如果符合以下条件，请勿保留模型。 指定相关性、相对误差和所用字段数的阈值。 无法满足这些标准中的任意一个的模型将被丢弃，并且不会在汇总报告中列出。

- **相关性小于**。 这是要包含在摘要报告中的模型的最小相关性（以绝对值表示）。
- **使用的字段数大于**。 这是要包含的任意模型将使用的最大字段数。
- **相对误差大于**。 这是要包含的任意模型的最大相对误差。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关更多信息，请参阅主题 第 46 页的『自动建模节点停止规则』。

自动数值节点专家选项

通过“自动数值”节点的“专家”选项卡，您可以选择要使用的算法和选项并指定中止规则。

选择模型。 缺省情况下，将选中所有模型进行构建；但是，如果您拥有 Analytic Server，那么您可以选择将模型限制为能够在 Analytic Server 上运行的模型，并对模型进行预设，使其构建拆分模型，或准备好处理超大型数据集。

注: 不支持在“自动数值”节点中本地构建 Analytic Server 模型。

使用的模型。 使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。 选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。 列出可用的算法（请参阅下面的内容）。

模型参数。 对于每个模型类型，可以使用缺省设置，或选择 **指定** 为每个模型类型选择选项。 这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。 例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数量。 列出根据当前设置为每个算法生成的模型的数目。 当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制构建单个模型所花费的最长时间。 （仅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。 例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，那么您大概不希望它使得整个的建模运行停滞。

支持的算法

神经网络节点使用的模型是对人类大脑处理信息的方式简化了的模型。 此模型通过模拟大量类似于神经元的抽象形式的互连简单处理单元而运行。 神经网络是功能强大的一般函数估计器，只需要最少的统计或数学知识就可以对其进行训练或应用。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。 该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果树中某个节点中 100% 的观测值都属于目标字段的一个特定类别，那么该节点将被认定为“纯洁”。 目标和输入字段可以是数字范围或分类（名义、有序或标志）； 所有拆分都是二进制的（只有两个子组）。

CHAID 使用卡方统计来生成决策树，以确定最佳的分割。 CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。 目标和输入字段可以是数字范围（连续）或分类。 穷举 CHAID 是 CHAID 的修正版，它可以更彻底地检查所有可能的拆分，但计算时间较长。

线性回归是一种常用的统计技术，用于汇总数据并通过拟合直线或曲面来进行预测，从而最大限度地减少预测输出值与实际输出值之间的差异。

广义线性模型对广义线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。 而且，该模型还允许因变量呈非正态分布。 它涵盖了大量统计模型的功能，包括线性回归、逻辑回归、计数数据的对数线性模型和区间删失生存模型。

The *k*-最近相邻元素 (KNN) 节点将新的观测值关联到预测变量空间中与其最邻近的 *k* 个对象的类别或值（其中 *k* 为整数）。 类似观测值相互靠近，而不同观测值相互远离。

使用支持向量机 (SVM) 节点，可以将数据分为两组，而无需过度拟合。 SVM 可以与宽数据集配合使用，例如那些含有大量输入字段的数据集。

线性回归模型根据目标与一个或多个预测变量之间的线性关系来预测连续目标。

通过线性支持向量机 (LSVM) 节点，您可以将数据分为两组，而无需过度拟合。 LSVM 是线性的，并且可以与大量数据集配合使用，例如包含大量记录的数据集。

“随机树”节点与现有 C&RT 节点相似；但是，“随机树”节点旨在处理大数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示产生的模型。 “随机树”节点将生成您可以对未来观测值进行预测或分类的决策树。 通过在每个步骤最大限度降低不纯洁度，此方法使用递归分区将训练记录分割为多个段。如果树中某个节点的全部观测值都属于目标字段的一个特定类别，那么系统会将该节点视为纯洁。 目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。

树 AS 节点类似于现有的 CHAID 节点；但是，“树 AS”节点旨在处理大量数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示生成的模型。 此节点通过使用卡方统计 (CHAID) 来识别最优拆分，从而生成决策树。 对 CHAID 的这一使用可生成非二元树，意味着某些拆分将具有两个以上的分支。 目标和输入字段可以是数字范围（连续）或分类。 Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。

XGBoost Linear© 是将线性模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 SPSS Modeler 中的 XGBoost Linear 节点使用 Python 进行实现。

GLE 扩展了线性模型，以便目标可以有非正态分布，通过指定的连接函数与因子和协变量线性相关，并且观测值可能相关。 广义线性混合模型涵盖了各种模型，从简单线性回归模型到非正态纵向模型数据的复杂多级模型。

XGBoost© 是实现提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost-AS 节点仅显示了核心功能和常用参数。 在 Spark 中实现 XGBoost-AS 节点。

自动数值节点设置

通过“自动数值”节点的**设置**选项卡，您可以预先配置模型块中可用的“分数-时间”选项。

过滤掉由整体模型生成的字段。 从输出中移除由各个模型生成的所有附加字段，这些模型均输入到“整体”节点中。 如果只想关注所有输入模型中的综合评分，请选中此复选框。 如果希望使用分析节点或评估节点将综合评分的准确性与各个输入模型评分的准确性进行比较，则请确保取消选中此选项。

计算标准误差。 对于连续（数值范围）目标，缺省情况下会运行标准误差计算以计算测量或估算值与真值之间的差值；并显示这些估算值的相近匹配程度。

提示: 持续机器学习设置还可用于“自动分类器”节点和“自动数值”节点。

自动聚类节点

自动聚类节点估算和比较识别具有类似特征记录组的聚类模型。 节点的工作方式与其他自动建模节点相同，这使您可以在一次建模运行中试验多个选项组合。 模型可使用基本测量进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个基于特定字段的重要性的测量。

聚类模型常常用于识别在后续分析中可用作输入的组。 例如，您可能希望基于如收入的统计特征来针对客户群，或基于客户过去购买的服务而针对客户群。 可以在不了解客户群及其特征的情况下进行此操作 -- 您可能不知道要寻找多少个客户群，或该用什么特征去定义客户群。 聚类模型常称作不受监督的学习模型，因为其不使用目标字段，且不返回可估算为真或假的具体预测。 聚类模型的值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。 有关更多信息，请参阅第 171 页的『第 11 章 聚类模型』。

需求。 这是用于定义兴趣特征的一个或多个字段。 聚类模型使用目标字段的方式与其他模型不同，因为其不作出能被评估为真或假的特定预测。 相反，其用于识别可能相关的观测值组。 例如，您无法使用预测给定客户会流失还是对预订作出积极响应的聚类模型。 但您可以使用基于客户对此类事物的倾向性将客户分组的聚类模型。 不使用权重字段和频率字段。

评估字段。 虽然不使用目标，但是您可以选择性地指定要在比较模型中使用的一个或多个评估字段。 可通过衡量聚类是否能有效区分这些字段，评估聚类模型的效果。

支持的模型类型

支持的模型类型包括二阶、K 均值、Kohonen、单类 SVM 和 K-Means-AS。

自动聚类节点模型选项

使用“自动聚类”节点的“模型”选项卡可以指定要保存的模型数，以及用于比较模型的标准。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

模型评级依据。 指定用于比较和排序模型的标准。

- **轮廓**。 这是用于衡量聚类结合与分离特性的指数。 有关详细信息，请参阅下面的 *Silhouette* 排秩测量。
- **聚类数**。 模型中的聚类数。
- **最小聚类的大小**。 最小聚类的大小。
- **最大聚类的大小**。 最大聚类的大小。
- **最小/最大聚类**。 最小聚类与最大聚类的大小比率。
- **重要性**。 **字段**选项卡上的**评估**字段的重要性。 注意只有在**评估**字段已指定时，才能计算。

模型评级方式。 如果正在使用分区，那么可以指定根据训练数据集还是检验数据集进行排序。 对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。

要保留的模型数。 指定要在节点生成的块中列出的最大模型数。 按照指定的排秩标准将顺次列出排秩靠前的模型。 注意，增大此限制会降低性能。 允许的最大值为 100。

Silhouette 排秩测量

缺省排秩测量，Silhouette，缺省值为 0，这是因为小于 0 的值（即负值）表示其分配的聚类中的观测值与点之间的平均距离大于观测值与另一个聚类中点的最小平均距离。 因此，具有负 Silhouette 值的模型可以安全地丢弃。

排秩测量实际上为修改的 silhouette 系数，它结合了聚类结合（偏向包含紧密结合聚类的模型）和聚类分离（偏向包含高度分离聚类的模型）的概念。 平均 Silhouette 系数是在所有观测值上的简单平均，每个单独观测值应用下列计算：

(*B* - *A*) / max(*A*, *B*)

其中 *A* 为从观测值到其所属聚类的矩心的距离，*B* 为从观测值到每个其他聚类矩心的最小距离。

Silhouette 系数（及其平均值）大小在 -1（表示极差的模型）与 1（表示极好的模型）之间。 可以在总体观测值级别上求平均值（得到总体 Silhouette），也可在聚类级别上求平均值（得到聚类 Silhouette）。 距离可以使用 Euclidean 距离进行计算。

自动聚类节点专家选项

通过“自动聚类”节点的“专家”选项卡，您可以应用分区（如果可用），选择要使用的算法以及指定停止规则。

选择模型。 缺省情况下，将选中所有模型进行构建；但是，如果您拥有 Analytic Server，那么您可以选择将模型限制为能够在 Analytic Server 上运行的模型，并对模型进行预设，使其构建拆分模型，或准备好处理超大型数据集。

注: 不支持在“自动聚类”节点中本地构建 Analytic Server 模型。

使用的模型。 使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。 选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。 列出可用的算法（请参阅下面的内容）。

模型参数。 对于每个模型类型，可以使用缺省设置，或选择 **指定** 为每个模型类型选择选项。 这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。 例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数量。 列出根据当前设置为每个算法生成的模型的数目。 当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制构建单个模型所花费的最长时间。 （仅限 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。 例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，那么您大概不希望它使得整个的建模运行停滞。

支持的算法

K-Means 节点将数据集聚类到不同分组（或聚类）。 此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。 *k*-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。 您可以通过查看模型块 中每个单元所捕获观测值的数量来找出规模较大的单元。 这将让您对聚类的相应数量有所估计。

TwoStep 节点使用二阶聚类方法。 第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。 第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。 TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。 它可以高效处理混合的字段类型和大型的数据集。

自动聚类节点丢弃选项

使用“自动聚类”节点的“废弃”选项卡可以自动废弃不符合特定标准的模型。 这些模型将不会列在模型块中。

您可以指定最小 silhouette 值、聚类数、聚类大小和模型中所用评估字段的重要性。 Silhouette 以及聚类的数量和大小根据建模节点中指定的值确定。 有关更多信息，请参阅主题 第 55 页的『自动聚类节点模型选项』。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关更多信息，请参阅主题 第 46 页的『自动建模节点停止规则』。

自动模型块

执行自动建模节点时，节点评估每个可能选项组合的候选模型，基于您指定的测量为每个候选模型排序，并将最佳模型保存在复合自动模型块中。 此模型块实际上包含该节点生成的一个或多个模型的集合，其中模型可单独被浏览或选中用于评分。 每个模型列有模型类型和构建时间，以及适合该模型类型的多个其他测量。可以按照这些列中的任意一列对表进行排序，以便快速确定最关注的模型。

- 要浏览任何一个单独的模型块，请双击模型块图标。 然后，可以从这里生成该模型的建模节点到流工作区，或生成模型块副本到模型选用板。
- 使用缩略图图形可以快速而直观地评估每个模型类型，总结如下。 可以双击缩略图生成标准大小的图形。标准大小的散点图可以最多显示 1000 个点并且会在数据集包含更多点时基于样本。 （仅对于散点图，图表每显示一次就重新生成一次，所以上游数据中的任意更改（例如在未选中 **设置随机种子** 时更新随机样本或分区）在每次重新绘制散点图时都会反映出来。
- 使用工具栏在“模型”选项卡上显示或隐藏特定的列或更改用于对表排序的列。 （也可以通过单击列标题更改排序列。）

- 使用“删除”按钮以永久删除任何未用的模型。
- 要重新为列排序，请单击列标题并将该列拖放到所需位置。
- 如果正在使用分区，那么可选择查看可应用的训练分区或检验分区的结果。

特定的列取决于要对比的模型的类型，下文已详细列出。

二元目标

- 对于二元模型，缩略图图表显示实际值的分布和与预测值的重叠，来快速直观地表示每个类别中正确预测的记录条数。
- 排秩标准与“自动分类器”建模节点中的选项匹配。 有关更多信息，请参阅主题 第 47 页的『自动分类器节点模型选项』。
- 对于最大利润，还会报告产生的最大数的百分位数。
- 对于累积提升，可以使用工具栏更改选定的百分位数。

名义目标

- 对于名义（集合）模型，缩略图图表显示实际值的分布和与预测值的重叠，来快速直观地表示每个类别中正确预测的记录条数。
- 排秩标准与“自动分类器”建模节点中的选项匹配。 有关更多信息，请参阅主题 第 47 页的『自动分类器节点模型选项』。

连续目标

- 对于连续（数值范围）模型，将根据每个模型的观测值预测图形散点，从而快速直观地表示模型之间的相关性。 对于好的模型，点应趋向于聚集在对角线周围，而不是在整个图形中随机分布。
- 排秩标准与“自动数值”建模节点中的选项匹配。 有关更多信息，请参阅主题 第 52 页的『自动数值节点模型选项』。

聚类目标

- 对于聚类模型，将根据每个模型的聚类计算图形散点，从而快速直观地表示聚类分布。
- 排秩标准与“自动聚类”建模节点中的选项匹配。 有关更多信息，请参阅主题 第 55 页的『自动聚类节点模型选项』。

选择评分模型

使用？ 列使您能够选择要在评分中使用的模型。

- 对于二元、名义和数字目标，您可以选择多个评分模型，并将评分组合在一个整体模型块中。 通过结合多个模型的预测，可以避免单个模型的局限性，使所得的整体准确性通常比从任一模型中获得的准确性要高。
- 对于聚类模型，一次只能选择一个评分模型。 缺省情况下，首先选择顶级模型。

生成节点和模型

可以从复合自动模型块的构建位置生成其副本，或自动建模节点。 例如，当您没有从中构建自动模型块的原始流时，这可能非常有用。 此外，还可以为自动模型块中列出的任何单独模型生成模型块或建模节点。

自动建模块

从“生成”菜单中，选择**模型至选用板**将自动模型块添加到模型选用板上。 可对生成的模型进行保存，或者在不重新运行流的情况下使用它。

或者，可以从“生成”菜单中选择**生成建模节点**以便将建模节点添加到流工作区。 可以不用重复完整的建模运行，而使用此节点重新估计选定的模型。

单独模型块

1. 在**模型**菜单中，双击所需的单独模型块。 块副本在新的对话框中打开。
2. 从新对话框中的“生成”菜单中，选择**模型至选用板**将单独建模块添加到模型选用板上。
3. 或者，可以从新对话框中的“生成”菜单中选择**生成建模节点**以便将单独建模节点添加到流工作区。

生成评估图表

对于二元模型，可以生成评估图表以直观评价和对比每个模型的性能。 评估图表不适用于自动数值或自动聚类节点生成的模型。

1. 在“自动分类器”自动化模型块中的使用？下， 选择要评估的模型。
2. 从“生成”菜单中，选择**评估图表**。 这将显示“评估图表”对话框。
3. 选择图表类型和其他需要的选项。

评估图形

在自动模型块的“模型”选项卡上，可以向下浏览以显示所示每个模型的单独图形。 对于自动分类器和自动数值块，“图形”选项卡同时显示反映所有模型组合结果的图形和预测变量重要性。 有关更多信息，请参阅主题第 32 页的『预测变量重要性』。

对于“自动分类器”，会显示一个分布图，而对于“自动数值”则显示一个多重散点图（也称为散点图）。

第 6 章 决策树

决策树模型

决策树模型可用于开发分类系统，此分类系统可以基于一组决策规则来预测或分类未来的观测值。 如果已将数据分成您感兴趣的类别（例如，高风险和低风险贷款、订户和非订户、投票人和非投票人或细菌类型），那么您可以使用自己的数据来构建用于对具有最高准确性的旧观测值或新观测值进行分类的规则。 例如，可以基于年龄和其他因素构建对信用风险或购买意向进行分类的树。

此方法（有时称为 规则归纳 ）有多个优点。 首先，浏览树的同时可以明显地看出模型背后的推论过程。 这与其他“黑箱”建模技术不同的地方，在其他“黑箱”建模技术中，您很难了解其内部逻辑。

其次，此过程只会将真正影响决策的属性自动包含在其规则中。 不会提高树的准确性的属性将被忽略。 此方法可获得非常有用的数据信息，并且可用于在训练其他学习方法（如神经网络）之前将数据缩减到相关字段。

决策树模块可转换成 if-then 规则的集合（ 规则集 ），在多数情况下此规则集以更为复杂的形式显示信息。 决策树表示法可以让您知道数据属性是如何将总体 分割 或 分区 成与问题相关的子集。 树-AS 节点输出不同于其他决策树节点，因为它在块中直接包含规则列表，无需创建规则集。 规则集表示法可以让您知道特定项目组与具体结论是如何关联的。 例如，以下规则就提供了关于值得购买的一组汽车的概要：

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

树构建算法

有多种算法可用于执行分类和分段分析。 这些算法执行的操作基本相同，检查数据集中的所有字段，通过将数据分割为多个子组来找到能够实现最佳分类或预测的字段。 此过程将重复应用以将子组分割成越来越小的单位，直到树结束生长（由特定的停止条件所定义）。 构建树的过程中所用的目标和输入字段可以是连续（数字范围）或分类（这取决于所采用的算法）。 如果使用的是连续目标，那么生成回归树；如果使用的是分类目标，那么生成分类树。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。 该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果树中某个节点中 100% 的观测值都属于目标字段的一个特定类别，那么该节点将被认定为“纯洁”。 目标和输入字段可以是数字范围或分类（名义、有序或标志）； 所有拆分都是二进制的（只有两个子组）。

CHAID 使用卡方统计来生成决策树，以确定最佳的分割。 CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。 目标和输入字段可以是数字范围（连续）或分类。 穷举 CHAID 是 CHAID 的修正版，它可以更彻底地检查所有可能的拆分，但计算时间较长。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。 输入字段可以是数字范围（连续），但目标字段必须是分类。 所有分割都是二元的。

C5.0 节点构建决策树或规则 集。 该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。 目标字段必须为分类字段。 允许进行多次多于两个子组的分割。

树 AS 节点类似于现有的 CHAID 节点；但是，“树 AS”节点旨在处理大量数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示生成的模型。 此节点通过使用卡方统计 (CHAID) 来识别最优拆分，从而生成决策树。 对 CHAID 的这一使用可生成非二元树，意味着某些拆分将具有两个以上的分支。 目标和输入字段可以是数字范围（连续）或分类。 Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。

“随机树”节点与现有 C&RT 节点相似；但是，“随机树”节点旨在处理大数据以创建单个树，并在 SPSS Modeler V17 中添加的输出查看器中显示产生的模型。 “随机树”节点将生成您可以对未来观测值进行预测或分类的决策树。 通过在每个步骤最大限度降低不纯洁度，此方法使用递归分区将训练记录分割为多个段。如果树中某个节点的全部观测值都属于目标字段的一个特定类别，那么系统会将该节点视为纯洁。 目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。

基于树的分析的一般用法

以下为一些基于树的分析的多个用法：

细分：确定可能隶属于特定类别的人员。

分层：将观测值分配到多个类别中的一个，例如高风险组、中等风险组和低风险组。

预测：创建规则，并使用这些规则来预测未来事件。 预测还可能意味着尝试将预测属性与连续变量值相关联。

数据降维和变量筛选：从大型变量集选择有用的预测变量子集，以用于构建正式的参数模型。

交互识别：确定仅与特定子组有关的关系，并在正式的参数模型中指定这些关系。

类别合并和带状化连续变量：以最小的信息损失，对组预测变量类别和连续变量进行重新编码。

交互树构建器

可以自动生成树模型，由运用算法在其中决定每一级的最佳分割，也可以使用交互树构建器来控制模型的生成，并在保存模型块之前运用专业知识精练或简化树。

1. 创建流并添加以下任一决策树节点：C&R 树、CHAID 或 QUEST。

   注: 树-AS 或 C5.0 树都不支持交互式树构建。
2. 打开节点，并在“字段”选项卡上选择目标字段和预测变量字段，然后在需要时指定其他模型选项。 有关具体说明，请参阅各树构建节点文档。
3. 在“构建选项”选项卡的“目标”面板上，选择**启动交互会话**。
4. 单击**运行**以启动树构建器。

其中显示了从根节点开始的当前树。 可以逐层编辑和修剪树，并在生成一个或多个模型之前访问增益、风险和相关的信息。

注释

- 使用 C&R 树、CHAID 和 QUEST 节点时，模型中使用的所有有序字段的存储类型都必须是数字（而非字符串）。 必要的话，可以使用重新分类节点对存储类型进行转换。
- 还可以选择使用分区字段将数据分隔到训练样本和测试样本中。
- 作为使用树构建器的另一种替代方法，也可以直接从建模节点中生成树模型或其他 IBM SPSS Modeler 模型。 有关更多信息，请参阅主题 第 68 页的『直接构建树模型』。

生成和修剪树

使用树构建器的“查看器”选项卡可以查看从根节点开始的当前树。

1. 要生成树，请从菜单中选择：

   树 > 生成树

   系统将通过递归分割每个分支直到符合一个或多个停止标准来构建树。 然后，可根据使用的建模方法在每个分割处自动选择最合适的预测变量。
2. 也可以选择 **生成树的第一层** 添加一个层。
3. 要在一个特定节点下添加分支，可选择该节点，然后选择 **生成分支** 。
4. 要选择某个分割所使用的预测变量，请选择所需的节点，然后选择**使用定制分割生成分支**。 有关更多信息，请参阅主题 第 61 页的『定义定制分割』。
5. 要修剪分支，可选择某个节点，然后选择 **移除分支** 以清除所选择的节点。
6. 要移除树的最底层，可选择 **移除第一层** 。
7. 仅对于 C&R 树和 QUEST 树，选择**生成树并修剪**，以基于成本复杂性算法进行修剪，该算法根据终端节点的数量调整风险估计，通常会生成更简单的树。 有关更多信息，请参阅主题 第 69 页的『C&R 树节点』。

在查看器选项卡上读取分割规则

查看“查看器”选项卡上的分割规则时，方括号表示临界值包含在范围中，而圆括号表示临界值不包含在范围中。 因此，表达式 (23,37] 表示从 23（不包括）到 37（包括）；也就是说，从 23 之上到 37。 在“模型”选项卡上，相同的条件将显示为：

```
Age > 23 and Age <= 37
```

中断树生成。 要中断树生成操作（例如，如果此操作所用的时间比预期的长），可单击工具栏上的“停止执行”按钮。

图 *28: “*停止执行”按钮

此按钮仅在树生成期间启用。 此按钮将使当前的生成操作在其当前点停止，并保留所有已添加的节点，但不保存所作的更改，也不关闭该窗口。 树构建器将保持打开状态，以便您根据需要生成模型、更新指令，或以适当的格式导出输出。

定义定制分割

通过“定义分割”对话框，您可以选择预测变量并为每个分割指定条件。

1. 在树构建器的“查看器”选项卡上选择一个节点，然后从菜单中选择：

   树 > 使用定制拆分扩展分支
2. 从下拉列表中选择所需的预测变量，或单击**预测变量**按钮，以查看每个预测变量的详细信息。 有关更多信息，请参阅主题 第 62 页的『查看预测变量详细信息』。
3. 可接受为每个分割选择的缺省条件，或选择**定制**为分割指定适当的条件。
   - 对于连续（数值范围）的预测变量，可以使用**编辑范围值**字段以指定落在每个新节点中的值的范围。
   - 对于分类预测变量，可使用**编辑集合值**或**编辑有序值**字段，以指定映射到每个新节点的特定值（如果是有序预测变量，那么指定值的范围）。
4. 选择 **生成** ，使用选定的预测变量重新生成分支。

在不考虑中止规则的情况下，通常可使用任何预测变量分割树。 唯一的例外情况是当节点是纯节点（即所有观测值都落在相同的目标类中，从而没有可分割的观测值），或所选择的预测变量是常数（即没有可分割的预测变量）时无法分割树。

缺失值分配到。 仅对于 CHAID 树，如果给定的预测变量中有缺失值，那么可以在定义定制分割时选择将这些缺失值分配给特定的子节点。 （对于 C&R 树和 QUEST，可使用替代项按算法中定义的方式处理缺失值。有关更多信息，请参阅主题 第 62 页的『分割的详细信息和代用项』。）

查看预测变量详细信息

“选择预测变量”对话框中显示了可用于当前分割的预测变量（有时称为“代替变量”）的统计量。

- 对于 CHAID 和 Exhaustive CHAID，列出了每个分类预测变量的卡方统计量；如果预测变量是数字范围类型，那么显示 *F* 统计量。 卡方统计量可用来测量目标字段与分割字段的不相关程度。 较高的卡方统计量通常与较低的概率有关，这意味着两个字段间不相关的机率较低 - 表示此分割情况良好。 这里也将自由度包括在内，因为自由度考虑了以下事实，即与双向分割相比，三向分割更易具有较高的统计量和较低的概率。
- 对于 C&R 树和 QUEST，显示了每个预测变量的改进值。 如果使用此预测变量，那么改进值越大，父节点和子节点间的纯度差异越大。 （纯节点指其中所有的观测值都落在一个目标类别中的节点；树中的杂质越少，此模型拟合数据的效果就越好。） 换句话说，较高的改进值通常表示对此类型的树进行了有用的分割。 所使用的杂质测量在树构建节点中指定。

分割的详细信息和代用项

可在“查看器”选项卡中选择任意节点，然后选择位于工具栏右侧的分割信息按钮查看有关该节点的分割详细信息。 此时将显示所使用的分割规则及相关的统计量。 对于 C&R 树分类树，将显示改进值和关联值。 关联值可用于测量代用项与原始分割字段间的一致性，其中“最佳”代用项通常是对分割字段模拟得最像的字段。 对于 C&R 树和 QUEST，还将列出所有用于代替主要预测变量的替代项。

要编辑选定节点的分割，可单击位于代用项面板左侧的图标以打开“定义分割”对话框。 （作为快捷方式，可以在单击图标选择代用项作为原始分割字段之前，从列表中选择此代用项。）

代用项。 如果适用，那么会针对所选节点显示主要分割字段的所有代用项。 代用项是在给定记录的主要预测变量值缺失时使用的替代字段。 给定分割允许的最大代用项数在树构建节点中指定，但实际数量取决于训练数据。 一般来讲，缺失数据越多，可能使用的代用项越多。 对于其他决策树模型，此选项卡为空。

注: 要在模型中包含替代项，必须在训练阶段对其进行标识。 如果训练样本没有缺失值，那么不会标识任何代用项；在测试或评分过程中遇到的具有缺失值的所有记录将自动落入记录数最大的子节点。 如果在测试或评分过程中预期出现缺失值，请确保值在训练样本中也处于缺失状态。 代用项对于 CHAID 树不可用。

虽然 CHAID 树中不使用代用项，但当定义定制分割时，仍可选择将这些代用项分配给特定的子节点。 有关更多信息，请参阅主题 第 61 页的『定义定制分割』。

定制树形视图

在树构建器的“查看器”选项卡中显示当前的树。 缺省情况下，将展开树中所有的分支，但也可以按照需要展开和折叠分支并定制其他设置。

- 单击父节点右下角的减号 (**–**) 可以隐藏其所有子节点。 单击父节点右下角的加号 (**+**) 显示其子节点。
- 使用“视图”菜单或工具栏更改树的方向（从上至下、从左至右或从右至左）。
- 单击主工具栏上的“显示字段和值标签”按钮以显示或隐藏字段和值标签。
- 使用放大镜按钮放大或缩小视图，或单击工具栏右侧的树状图按钮以查看完整树的图。
- 如果正在使用分区字段，那么可以在训练分区和测试分区之间交换树形视图（**视图> 分区**）。 显示测试样本时，可以查看但不能编辑树。 （将在窗口右下角的状态栏中显示当前分区。）
- 单击分割信息按钮（工具栏最右侧的“i”按钮）以查看当前分割的详细信息。 有关更多信息，请参阅主题 第 62 页的『分割的详细信息和代用项』。
- 将在每个节点中显示统计量、图形或同时显示两者（请参见下文）。

显示统计量和图形

节点统计信息。 对于分类目标字段，每个节点中的表显示每个分类中的记录数和百分比以及该节点代表的整个样本的百分比。 对于连续（数值范围）目标字段，该表显示目标字段的平均值、标准差、记录数和预测值。

节点图。 对于分类目标字段，图形为目标字段的每个类别中的百分比条形图。 表中每行的前面是一个颜色样本，其对应的颜色表示该节点图形中的每个目标字段类别。 对于连续（数字范围）目标字段，该图形显示节点中记录的目标字段的直方图。

收益

“增益”选项卡可显示树中所有终端节点的统计量。 增益可用于测量给定节点上的平均值或比例与总平均值之间的差异大小。 一般来说，此差异越大，作为决策工具的树就越有效。 例如，某个节点的指数或“提升”值为 148% 表示，该节点中的记录落在目标类别中的可能性大概是其作为一个整体用于数据集的可能性的 1.5 倍。

对于 C&R 树和指定防止过度拟合集合的 QUEST 节点，显示两组统计信息：

- 树生成集合 - 已移除防止过度拟合集合的训练样本
- 防止过度拟合集合

对于其他 C&R 树和 QUEST 交互树以及所有 CHAID 交互树，只显示树生长组统计信息。

通过“增益”选项卡，您可以执行下列操作：

- 显示每个节点统计量、累积数统计量或分位数统计量。
- 显示增益或利润。
- 将视图在表和图表间进行交换。
- 选择目标类别（仅分类目标）。
- 根据指数百分比对表按升序或降序排序。 如果显示的是多个分区的统计量，那么通常将排序应用于训练样本而不是测试样本。

一般来说，在增益表中选定的内容也会在树形视图中得到更新，反之亦然。 例如，如果在表中选择某个行，那么也会在树中选中相应的节点。

分类增益

对于分类树（指使用分类目标变量的树），从增益指数百分比可看出每个节点上给定目标类别的比例与总比例间的差异有多大。

依次显示节点统计量

在此视图的表中，将为每个终端节点显示一行。 例如，如果直邮活动的总响应是 10%，但有 20% 的记录落在节点 X 内并且做出积极的响应，那么该节点的指数百分比应为 200%，表示该组中的响应者进行购买的可能性大概是总人数的两倍。

对于 C&R 树和指定防止过度拟合集合的 QUEST 节点，显示两组统计信息：

- 树生成集合 - 已移除防止过度拟合集合的训练样本
- 防止过度拟合集合

对于其他 C&R 树和 QUEST 交互树以及所有 CHAID 交互树，只显示树生长组统计信息。

节点。 当前节点的标识（显示在“查看器”选项卡上）。

节点：n。 此节点中的记录总数。

节点 (%)。 数据集中属于此节点的所有记录的百分比。

收益：n。 具有所选目标类别且属于此节点的记录数。 换句话说，在数据集的所有落在目标类别的记录中，有多少记录落在该节点?

收益 (%)。 整个数据集的目标类别中属于此节点的所有记录的百分比。

响应 (%)。 当前节点中落在目标类别下的记录的百分比。 该上下文中的响应有时也称为“匹配项”。

指数 (%)。 当前节点的响应百分比，表示为整个数据集的响应百分比的百分比。 例如，指数值为 300% 表示该节点中的记录落在目标类别中的可能性大概是其作为一个整体用于数据集的可能性的三倍。

累积统计

在累积视图中，表的每行显示一个节点，但统计量是累积的，并按指数百分比以升序或降序顺序排序。 例如，如果按降序排序，那么首先列出指数百分比最高的节点，并且接下来的行中的统计量是对该行及上面的行的累积数。

随着所添加节点的响应百分比越来越低，累积指数百分比将逐行降低。 最后一行的累积指数通常是 100%，因为此时将包括整个数据集。

分位数

在此视图中，表中的每一行都表示一个分位数而不是节点。 分位数可以是四分位数 (4)、五分位数 (5)、十分位数 (10)、二十分位数 (20) 或百分位数 (100)。 如果需要多个节点以补足此百分比（例如，如果显示四分位数时，而前两个节点包含的观测值不到所有观测值的 50%），那么可在一个分位数中列出多个节点。 可以对表的其余部分进行累积，且与累积视图的解释方式相同。

分类利润和投资回报率

对于分类树，增益统计量也可按利润和投资回报率显示。 通过“定义利润”对话框，您可以为每个类别指定收入和支出。

1. 在“增益”选项卡上，单击工具栏上的“利润”按钮（标注为 $/$）以访问该对话框。
2. 输入目标字段的每个类别的收入和支出值。

例如，如果为每个客户邮寄报价的成本是 $0.48，而从接受三个月预订的积极响应中获得的收入是 $9.95，那么每个 *no* 响应将花费 $0.48，而每个 *yes* 响应将赚取 $9.47（按 9.95–0.48 计算）。

在增益表中， **利润** 的计算方式为终端节点的每条记录中的总收入减去支出。 **ROI** 是某个节点的总利润除以总支出得到的值。

注释

- 利润值仅影响在增益表中显示的平均利润和投资回报率，可以明确查看统计量，尤其适合查看利润。 但是，它们不会影响基本的树模型结构。 不应将利润与误分类成本相混淆，误分类成本在树构建节点中指定，且可化为模型中的因子（作为避免高成本错误的一种方式）。
- 在两个交互树构建会话之间不会保留利润说明。

回归增益

对于回归树，可以选择依次显示节点视图、累积节点视图和分位数视图。 表中可显示平均值。 只有在分位数视图中才可使用图表。

增益图

在“增益”选项卡上，图表可作为表的替代项显示。

1. 在“增益”选项卡上，选择“分位数”图标（工具栏从左数第三个图标）。 （对于依次显示节点统计量或累积统计量，不可使用图表。）
2. 选择“图表”图标。
3. 按照需要从下拉列表中选择所显示的单位（百分位数、十分位数等等）。
4. 选择 **增益** 、 **响应** 或 **提升** 更改所显示的测量量。

增益图

增益图绘制的是表中 增益 *(%)* 列值的散点图。 增益定义为每个增量中匹配项数与树中匹配项总数的比例，它使用下列等式：

（增量中匹配项数/匹配项总数）x 100%

该图有效说明了您需要撒出多大范围的网络，才能获取树中所有匹配项的给定百分比。 对角线绘制的是整个样本的预期响应（如果未使用模型的话）。 这种情况下，响应率应该为常量，因为一个人响应的可能性与另一个人相同。 为了使您的收益加倍，您需要询问两倍数量的人。 曲线表明通过将那些秩（基于增益排序）位于较高百分比的人员包括在内，您可以使得响应得到多大程度的改善。 例如，包括最高的 50% 可能会网罗超过 70% 的正面响应。 该曲线越陡，增益越高。

提升图表

提升图表对表中 指数 *(%)* 列中的值进行了绘制。 此图表将每个增量中具有积极响应的记录的百分比与训练数据集中具有积极响应的记录的总百分比作了比较，其方程式为：

(增量中匹配项数/增量中记录数) / (匹配项总数/记录总数)

响应图表

响应图表对表中 响应 *(%)* 列中的值进行了绘制。 响应是增量中具有积极响应的记录的百分比，其方程式为：

（增量中具有积极响应的记录/增量中的记录） x 100%

基于增益的选择

通过“基于增益的选择”对话框，您可以根据指定的规则或阈值来自动选择具有最佳（或最差）增益的终端节点。 然后可以根据该选择生成一个选择节点。

1. 在“增益”选项卡上，选择依次显示节点视图或累积视图，然后选择该选择所基于的目标类别。 （该选择基于当前的表显示，不可用于分位数视图。）
2. 从“增益”选项卡的菜单中选择以下项：

   编辑 > 选择终端节点 > 基于增益的选择

   仅选择。 可以选择匹配节点或不匹配节点 - 例如，选择前 100 条记录以外的所有记录。

   按增益信息匹配。 根据当前目标类别的增益统计量来匹配节点，包括：

   - 其增益、响应或提升（指数）与指定的阈值相匹配的节点，例如，响应大于或等于 50%。
   - 基于目标类别的增益的顶部 *n* 个节点。
   - 上限为指定记录数的顶部节点。
   - 上限为指定训练数据百分比的顶部节点。
3. 单击**确定**更新“查看器”选项卡上的选择。
4. 要根据“查看器”选项卡上的当前选择新建“选择”节点，请从“生成”菜单中选择**选择节点**。 有关更多信息，请参阅主题 第 68 页的『生成过滤节点和选择节点』。

注：由于实际上选择的是节点而不是记录或百分比，因此无法始终获取与选择标准完全匹配的结果。 系统选择 上限为指定等级的完整节点。 例如，如果选择顶部 12 个观测值，而第一个节点中有 10 个观测值，第二个节点中有 2 个观测值，那么将只选择第一个节点。

风险

风险指任意等级上误分类的机率。 “风险”选项卡可显示某点的风险估计和（分类输出的）误分类表。

- 对于数字预测，风险是每个终端节点上的合并方差评估。
- 对于分类预测，风险是错误分类观测值的比例，可根据任意先验分布或误分类成本进行调整。

保存树模型和结果

可以用以下多种方式保存或导出交互树构建会话的结果：

- 生成基于当前树的模型（**生成> 生成模型**）。
- 保存用于生成当前树的指令。 下次执行树构建节点时，将自动重新生成当前树（包括已定义的任何定制分割）。
- 导出模型、增益和风险信息。 有关更多信息，请参阅主题 第 67 页的『导出模型、增益和风险信息』。

通过树构建器或树模型块，可以执行下列操作：

- 根据当前的树生成过滤节点或选择节点。 有关更多信息，请参阅第 68 页的『生成过滤节点和选择节点』。
- 生成一个规则集块，该节点将树结构表示成一组定义了树的终端分支的规则。 有关更多信息，请参阅第 68 页的『从决策树中生成规则集』。
- 此外，还可以按 PMML 格式导出模型（仅限树模型块）。 有关更多信息，请参阅第 30 页的『模型选用板』。 如果模型包含任何定制分割，那么不会在导出的 PMML 中保留此信息。 （保留分割，但不保留它是定制分割而不是通过算法选择的分割这一事实。）

- 基于当前树的所选部分生成图形。 请注意，仅当块附加到流中的其他节点时，此操作才有效。 有关更多信息，请参阅第 89 页的『生成图形』。

注：无法保存交互树自身。 为了避免丢失所执行的操作，请在关闭树构建器窗口之前生成模型和/或更新树指令。

从树构建器生成模型

要基于当前树生成模型，可从树构建器菜单中选择以下项：

生成 > 模型

在“生成新模型”对话框中，您可以从下列选项中进行选择：

模型名称。 可以指定定制名称或根据建模节点的名称自动生成模型名称。

创建节点于。 可以在**画布**、**GM 选用板**或**两者**上添加节点。

包含 tree 指令。 要在生成模型中包括来自当前树的指令，选择此选项。 通过此选项，您可以根据需要重新生成树。 有关更多信息，请参阅主题 第 66 页的『树生长指令』。

树生长指令

对于 C&R 树、CHAID 和 QUEST 模型，树指令用于指定生成树（一次一级）的条件。 每当从节点中启动交互树构建器时，都会应用指令。

- 指令可作为一种最安全的方法用来重新生成在以前的交互会话中创建的树。 有关更多信息，请参阅主题 第 67 页的『更新树指令』。 也可以手动编辑指令，但操作时需要格外小心。
- 指令与其所描述的树结构高度相关。 因此，对原始数据或建模选项的任何更改都可能会导致以前有效的一组指令失效。 例如，如果 CHAID 算法基于更新的数据将双向分割更改为三向分割，那么基于以前的双向分割的所有指令都将失效。

注：如果选择直接生成模型（不使用树构建器），那么将忽略所有的树指令。

编辑指令

1. 要查看或编辑已保存的指令，请打开树构建节点并选择“构建选项”选项卡的“目标”面板。
2. 选择 **启动交互会话** 以启用控件，选中 **使用树指令**，然后单击 **指令**。

指令语法

指令可指定从根节点开始生成树的条件。 例如，生成树的第一层：

```
Grow Node Index 0 Children 1 2
```

由于未指定任何预测变量，算法将选择最佳分割。

注：第一个拆分必须始终位于根节点 (`Index 0`) 上，并且必须指定两个子代的索引值 （在本例中为 1 和 2）。 除非首先生成创建的节点 2 的根，否则指定 `Grow Node Index 2 Children 3 4` 是无效操作。

要生成树，请使用：

```
Grow Tree
```

要生成并修剪树（仅限 C&R 树），请使用：

```
Grow_And_Prune Tree
```

要为连续预测变量指定定制分割，请使用：

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

要分割具有两个值的名义预测变量，请使用：

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

要分割具有多个值的名义预测变量，请使用：

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

要分割有序预测变量，请使用：

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

注：指定定制拆分时，字段名称和值（EDUCATE、GENDER、CHILDS 等） 区分大小写。

CHAID 树的指令

CHAID 树的指令对数据或模型中的更改非常敏感，因为这些指令与 C&R 树和 QUEST 中的不同，它们并非只能使用二元分割。 例如，下面的语法看起来很有效，但如果算法将根节点分割为两个以上的子节点时，这些语法将失效：

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

对于 CHAID，节点 0 可能具有 3 个或 4 个子节点，这种情况将使上述第二行语法失效。

在脚本中使用指令

也可使用三重引号将指令嵌入到脚本中。

更新树指令

要保留在交互树构建会话中执行的操作，可以保存用于生成当前树的指令。 与保存无法进一步进行编辑的模型块不同的是，您可以通过保存指令来按树的当前状态重新生成该树以进行进一步编辑。

要更新指令，请从树构建器菜单中选择以下项：

文件 > 更新指令

指令保存在用于创建树（C&R 树、QUEST 或 CHAID）的建模节点中，并可用于重新生成当前树。 有关更多信息，请参阅主题 第 66 页的『树生长指令』。

导出模型、增益和风险信息

可以从树构建器中根据需要以文本、HTML 或图像格式导出模型、增益和风险统计量。

1. 在树构建器窗口中，选择要导出的选项卡或视图。
2. 从菜单中选择：

   文件 > 导出
3. 根据需要选择 **文本** 、 **HTML** 或 **图形** ，并从子菜单中选择要导出的特定项目。

在适用的情况下，导出基于当前的选择。

导出文本或 HTML 格式。 您可以为训练分区或检验分区（如果已定义）导出增益统计量或风险统计量。 导出基于“增益”选项卡上的当前选择 - 例如，可以选择依次显示节点统计量、累积统计量或分位数统计量。

正在导出图形。 可以导出在“查看器”选项卡上显示的当前树，或为训练分区或测试分区（如果已定义）导出增益图。 可用的格式包括 *.JPEG*、 *.PNG* 和 *.BMP*。 对于增益，导出基于“增益”选项卡上的当前选择（仅当显示图表时可用）。

生成过滤节点和选择节点

在树构建器窗口中，或在浏览决策树模型块时，从菜单中选择以下项：

生成 > 过滤节点

或者

> “选择”节点

"过滤"节点。 生成用于过滤当前树不使用的任何字段的节点。 此方法可以快速削减数据集，使其仅包括那些算法选择为重要字段的字段。 如果此决策树节点的上游存在“类型”节点，那么“过滤”模型块将传递所有角色为目标的字段。

"选择"节点。 生成用于选择所有落在当前节点中的记录的节点。 此选项需要在“查看器”选项卡中选择一个或多个树分支。

该模型块位于流工作区中。

从决策树中生成规则集

生成的规则集模型块可作为定义树的终端分支的一组规则来表示树的结构。 通常，规则集可保留完整的决策树中的大部分重要信息，但其使用的模型比较简单。 最重要的区别是，使用规则集时，可以为任意特定记录应用多个规则，也可以不应用任何规则。 例如，可以看到所有预测结果为否的规则，紧随其后是所有预测为是的规则。 如果应用多个规则，那么每个规则将根据与此规则关联的置信度获得一个加权“投票”，并通过组合应用到所讨论记录的所有规则的加权投票来确定最终的预测。 如果没有规则可应用，那么会将缺省预测分配到该记录。

注: 对规则集进行评分时，您可能会注意到此评分相比于针对树的评分存在差异；这是由于树种每个终端分支都是独立评分的。 如果在数据中存在缺失值，那么您可以明显注意到此差异。

仅可从具有分类目标字段的树（不是回归树）中生成规则集。

在树构建器窗口中，或在浏览决策树模型块时，从菜单中选择以下项：

生成 > 规则集

规则集名称：指定新的规则集模型块的名称。

节点的创建位置：控制新的规则集模型块的位置。 选择 **工作区** 、 **GM 选用板** 或 **两者** 。

最少实例数：指定在规则集模型块中保留的最小实例数（已应用规则的记录数）。 支持度小于指定值的规则将不会包含在新的规则集中。

最小置信度：指定规则集模型块中要保留的规则的最小置信度。 置信度小于指定值的规则将不会包含在新的规则集中。

直接构建树模型

作为使用交互式树构建器的替代方法，您可以在运行流时直接从节点构建决策树模型。 这与大多数其他模型构建节点相一致。 对于交互树构建器所不支持的 C5.0 树模型和树-AS 模型来说，这是唯一可以使用的方法。

1. 创建流并添加其中一个决策树节点 - C&R 树、CHAID、QUEST、C5.0 或树-AS。
2. 对于 C&R 树、QUEST 或 CHAID，在“构建选项”选项卡的“目标”面板上，选择一个主目标。 如果您选择**构建单个树**，请确保将**模式**设为**生成模型**。

   对于 C5.0，在“模型”选项卡上，将**输出类型**设为**决策树**。

   对于树-AS，在“构建选项”选项卡的“基本”面板上，选择**树生长算法**类型。
3. 选择目标字段和预测变量字段，并在需要时指定其他模型选项。 有关具体说明，请参阅各树构建节点文档。
4. 运行流以生成模型。

关于构建树的注释

- 使用此方法生成树时，会忽略树生长指令。
- 无论使用交互模式还是直接模式，这两种创建决策树的方法最终都会生成相似的模型。 只需考虑希望在此过程中执行多大程度的控制。

决策树节点

IBM SPSS Modeler 中的决策树节点提供对以下树构建算法的访问：

- C&R 树
- QUEST
- CHAID
- C5.0
- 树 AS
- 随机树

有关更多信息，请参阅主题 第 59 页的『决策树模型』。

这些算法有一点很相似：它们可以通过将数据递归分割为越来越小的子组来构造决策树。 但是，存在一些重要的差异。

输入字段。 输入字段（预测变量）可以是以下任何类型（测量级别）：连续、分类、标志、名义或有序。

目标字段。 仅可指定一个目标字段。 对于 C&R 树、CHAID、树 AS 和随机树，目标可以为连续、分类、标志、名义或有序。 对于 QUEST，目标字段可以是分类、标志或名义。 对于 C5.0，目标字段可以是标志、名义或有序。

分割类型。 C&R 树、QUEST 和随机树仅支持二元分割（即，每个树节点不能分割成两个以上的分支）。 相比之下，CHAID、C5.0 和树-AS 支持一次分割为两个以上的分支。

用于分割的方法。 不同算法在用于确定分割的标准上有所不同。 C&R 树在预测分类输出时使用离差测量（缺省为 Gini 系数，不过您可以进行更改）。 对于连续目标，将使用最小平方偏差法。 CHAID 和树-AS 使用卡方检验；QUEST 将卡方检验用于分类预测变量并将方差分析用于连续输入。 对于 C5.0，将使用信息论度量，即信息增益率。

缺失值处理。 所有算法均允许预测变量字段缺失值，但它们使用不同的缺失值处理方法。 C&R 树和 QUEST 根据需要使用替代预测字段，以确保具有缺失值的记录在训练期间通过树。 CHAID 将缺失值分为单独的类别，并使它们可以用于树构建。 C5.0 使用分离方法，该方法将记录的分离部分从节点（此节点中的分割取决于具有缺失值的字段）向下传递到树的各个分支。

修剪。 C&R 树、QUEST 和 C5.0 提供的选项允许完全生成树，然后删除对于树的精确性没有显著影响的底层分割以进行修剪。 但是，所有决策树算法都允许控制最小子组大小，这有助于避免出现数据记录较少的分支。

交互树构建。 C&R 树、QUEST 和 CHAID 提供了启动交互式会话的选项。 通过此选项，您可以构建树（一次一级），编辑分割并在创建模型之前对树进行修剪。 C5.0、Tree-AS 和随机树没有交互选项。

先验概率。 C&R 树和 QUEST 支持在预测分类目标字段时为类别指定先验概率。 先验概率是对总体（从中抽取训练数据）中每个目标分类的总相对频率的估计。 换而言之，先验概率是对预测变量值有任何了解之前对每个可能的目标值的概率估计。 CHAID、C5.0、树 AS 和随机树不支持指定先验概率。

规则集。 不适用于树 AS 或随机树。 对于具有分类目标字段的模型，决策树节点提供了以规则集形式创建模型的选项，这有时比复杂决策树更容易解释。 对 C&R 树、QUEST 和 CHAID，您可以从交互式会话中生成规则集；对于 C5.0，可以在建模节点上指定此选项。 另外，所有决策树模型都支持根据模型块生成规则集。 有关更多信息，请参阅主题 第 68 页的『从决策树中生成规则集』。

C&R 树节点

分类和回归 (C&R) 树节点是一种基于树的分类和预测方法。 与 C5.0 类似，此方法可使用递归分区将训练记录分割为具有相似输出字段值的段。 “C&R 树”节点首先检查输入字段以找到最佳拆分（以拆分所引起的杂质指标下降情况来衡量）。 分割可定义两个子组，其中每个子组随后又被分割为两个子组，依此类推，直到触发其中一个停止标准为止。 所有分割都是二元的（仅有两个子组）。

修剪

C&R 树允许您先生成树，然后根据成本复杂性算法（该算法可根据终端节点数调整风险估计）修剪此树。通过此方法（此方法可以使树在长大后再根据更复杂的标准进行修剪）可生成交叉验证属性更佳的小型树。增加终端节点数通常会降低当前（训练）数据的风险，但当模型扩展为适用不可见数据时，实际的风险可能会更大。 假设在一种极端的情况下，训练集合中的每条记录都有一个单独的终端节点。 风险估算将为 0%，因为每条记录都属于其自己的节点，但不可见（测试）数据的误分类风险几乎肯定大于 0。 成本复杂性度量尝试弥补这一不足。

示例。 某有线电视公司委托进行市场营销研究，以确定有意预订有线电视互动新闻服务的用户。 使用研究中得来的数据可创建流，其中的目标字段为有意预订有线电视服务，预测变量字段则包括年龄、性别、教育、收入类别、每天看电视的时间和子女数。 通过将“C&R 树”节点应用于流，您将能够预测和分类响应以获取营销活动的最高响应率。

需求。 要训练“C&R 树”模型，您需要一个或多个输入字段以及正好一个目标字段。 目标字段和输入字段可以是连续字段（数字范围），也可以是分类字段。 设置为 双向或 无的字段将忽略。 对于模型中使用的字段，必须将它们的类型完全实例化，并且模型中使用的所有有序（有序集合）字段的存储类型必须是数字（而不是字符串）。 必要的话，可以使用重新分类节点对存储类型进行转换。

强度。 遇到缺少数据及大量字段等问题时，C&R 树模型的表现十分稳健。 这些模型通常不需要花费很长的训练时间用于估计。 另外，C&R 树模型似乎比某些其他模型类型更易于理解 - 派生自模型的规则解释起来更简明易懂。 与 C5.0 不同的是，C&R 树可同时兼容连续字段和分类输出字段。

CHAID 节点

CHAID 或卡方自动交互效应检测是一种通过使用卡方统计量识别最优分割来构建决策树的分类方法。

CHAID 首先检查每个输入字段和结果之间的交叉表，然后使用卡方独立性检验来检验显著性。 如果以上多个关系具有显著的统计意义，那么 CHAID 将选择最重要（*p* 值最小）的输入字段。 如果输入具有两个以上的类别，那么将会对这些类别进行比较，然后将结果中未显示出差异的类别合并在一起。 此操作通过将显示的显著性差异最低的类别对相继合并在一起来实现。 当所有剩余类别在指定的检验级别上存在差异时，此类别合并过程将终止。 对于名义输入字段，可以合并任何类别；对于有序集合，只能合并相邻的类别。

Exhaustive CHAID 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

需求。 目标和输入字段可以是连续字段，也可以是分类字段；节点在每一层上都可以分割为两个或多个子组。 模型中使用的所有顺序字段的存储类型都必须是数字类型（不是字符串）。 必要的话，可以使用重新分类节点对存储类型进行转换。

强度。 CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。 因此，与二元生成方法相比，CHAID 倾向于创建范围更广的树。 CHAID 适用于所有类型的输入，并且接受观测值权重和频率变量。

QUEST 节点

QUEST，或称快速、无偏倚、高效率统计树，是一种用于构建决策树的二元分类法。 开发此方法的一个主要原因是减少包含很多变量或观测值的大型 C&R 树分析所需的处理时间。 QUEST 的第二个目的是减少在分类树方法中发现的趋势以便支持允许有多个分割的输入，即连续（数值范围）输入或具有多个类别的输入。

- 根据显著性检验，QUEST 使用一系列规则来评估节点上的输入字段。 为了进行选择，可能需要对节点的每个输入执行一次检验。 与 C&R 树不同，所有的分割都不用检查，而与 C&R 树和 CHAID 都不同的是，在评估输入字段以供选择时不会检验类别组合。 因此可加快分析的速度。
- 通过使用由目标类别形成的组中选定的输入来运行二次判别分析可以确定分割。 而且，与穷举搜索（C&R 树）相比，此方法确定最优分割的速度得到了改进。

需求。 输入字段可以是连续（数值范围）的，但目标字段必须是分类的。 所有分割都是二元的。 不能使用权重字段。 模型中使用的所有有序（有序集合）字段的存储类型都必须是数值类型（不是字符串）。 必要的话，可以使用重新分类节点对存储类型进行转换。

强度。 与 CHAID 相似但与 C&R 树不同的是，QUEST 可使用统计检验确定是否使用输入字段。 QUEST 还可以将输入的选择与分割问题分开，分别为其应用不同的标准。 不过在 CHAID 中，确定变量选择的统计检验结果还可生成分割。 同样，C&R 树也可采用杂质更改测量在选择输入字段的同时确定分割。

决策树节点字段选项

在“字段”选项卡上，可以选择是要使用在上游节点中定义的字段角色设置，还是手动进行字段分配。

使用预定义角色：此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标、预测变量和其他角色，请选中此选项。

字段。 使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。 图标表示每个角色字段的有效测量级别。

要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择一个字段作为预测目标。

预测变量（输入）。 选择一个或多个字段作为预测输入。

分析权重。 （仅限 CHAID、C&RT 和树-AS）要使用字段作为观测值权重，在此处指定。 观测值权重将作为对输出字段各个水平上方差的差异的一种考量。 有关更多信息，请参阅主题 第 24 页的『使用频率和权重字段』。

决策树节点构建选项

通过“构建选项”选项卡，您可以设置构建模型的所有选项。 当然，您只需单击**运行**按钮，即可采用所有缺省选项来构建模型；不过，通常您需要根据具体用途定制构建选项。

该选项卡包含多个不同的窗口，您可以在其中设置特定于自己的模型的定制。

“决策树”节点 - 目标

对于“构建选项”选项卡上的“目标”窗格中的 C&R 树节点、QUEST 节点和 CHAID 节点，您可以选择是构建新模型还是更新现有模型。 还可以设置节点的主要目标：构建标准模型、构建准确性或稳定性增强的模型，或者构建用于超大型数据集的模型。

您要执行什么操作?

构建新模型。 （缺省）每次运行包含此建模结点的流时，就会创建一个全新模型。

继续训练现有模型。 缺省情况下，每次执行建模节点时，将创建一个全新的模型。 如果选中该选项，那么会继续训练该节点成功生成的最后一个模型。 这样就可以在无需访问原始数据的情况下更新或刷新现有的模型， 并可能会显著提升性能，这是因为 只有新的或更新后的记录被反馈到流中。 上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型调色板中不再可用的情况下，也可以使用该选项。

注: 仅当您选择**构建单一树**（针对 C&R 树、CHAID 和 QUEST）、**创建标准模型**（针对神经网络和线性模型）或**针对超大型数据集创建模型**以作为目标时，才会激活该选项。

您的主要目标是什么?

- **构建单个树**。 创建单个标准决策树模型。 通常，与使用其他目标选项构建的模型相比，标准模型更易于说明并可以更快速地进行评分。

  注: 只有**构建单一树**拆分模型中才支持**继续训练现有模型**，并且必须连接到 Analytic Server。

  方式。 指定用于构建模型的方法。 **生成模型**可在运行流时自动创建模型。 **启动交互式会话**将打开树构建器，您可以通过该构建器在创建模型块之前构建树（一次一级）、编辑分割并根据需要进行修剪。

  使用 tree 伪指令。 选中此选项可以指定从节点中生成交互树时应用的指令。 例如，可以指定第一级分割和第二级分割，当启动树构建器时会自动应用这些分割。 还可以保存交互树构建会话中的指令，以便将来重新创建树时使用。 有关更多信息，请参阅主题 第 67 页的『更新树指令』。
- **增强模型准确度（提升）**。 如果您要使用一种名为**增强**的特殊方法来提高模型准确率，请选择此项。 增强的工作原理是在序列中构建多个模型。 第一个模型按常规方式进行构建。 然后，以这种方式构建第二个模型时，重点集中于被第一个模型误分类的记录。 构建第三个模型时，将焦点集中于第二个模型的错误，依此类推。 最后，通过将整个模型集应用到观测值，并使用加权投票过程将单独的预测组合为一个总预测来分类观测值。 增强方法可以显著提高决策树模型的准确性，但也需要更长的训练时间。

- **增强模型稳定性（组装）**。 如果您要使用一种名为 **组装**（Bootstrap 汇总）的特殊方法来提高模型稳定性并避免过度拟合，请选择此项。 此选项将创建多个模型并将其进行组合，以获取更加可靠的预测。 与标准模型相比，使用此选项获取的模型构建和评分所花费的时间更长。
- **为非常大的数据集创建模型**。 如果您的数据集过大，而无法使用任何上述目标选项构建模型，请选择此项。 此选项用于将数据划分为更小的数据块，并对每个块构建一个模型。 然后，将自动选择最准确的模型并将它们合并到单一模型块中。 如果您在此屏幕上选择**继续训练现有模型**选项，可以执行增量式模型更新。 仅在**为超大型数据集创建模型**的模型中支持**继续训练现有模型**选项，不需要连接到 Analytic Server。但是，无法使用拆分来创建超大型数据集的模型。

  注: 此选项适合超大型数据集，需要连接到 IBM SPSS Modeler Server。

“决策树”节点 - 基本

指定有关如何构建决策树的基本选项。

树生长算法（仅限 CHAID 和树-AS）选择您要使用的 **CHAID** 算法类型。 **Exhaustive CHAID** 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

最大树深度指定根节点以下的最大级数（对样本进行递归分割的次数）。 缺省值为 5；选择**定制**，并输入值以指定其他级数。

修剪（仅限 C&RT 和 QUEST）

对树进行修剪以避免过度拟合修剪包括删除对于树的准确性没有显著影响的底层分割。 修剪有助于简化树，使树更容易被理解，在某些情况下还可提高广义性。 如果需要未修剪的完整树，请取消选中此选项。

- **设置风险最大差分（在标准误差范围内）**通过此选项，您可以指定更自由的修剪规则。 标准误差规则使算法可以选择最简单的树，该树的风险估计接近于（但也可能大于）风险最小的子树的风险估计。 该值表示已修剪树和风险最小的树之间所允许的风险估计差异大小。 例如，如果指定 2 ，那么将选择其风险估计（2 × 标准误差）大于完整树的风险估计的树。

最大代用项。 替代项是用于处理缺失值的方法。 对于树中的每个分割，算法都会对与选定的分割字段最相似的输入字段进行识别。 这些被识别的字段就是该分割的 代用项 。 当必须对某个记录进行分类，但此记录中的分割字段中具有缺失值时，可以使用代用项字段的值填补此分割。 增加此设置将可以更加灵活地处理缺失值，但也会导致内存使用量和训练时间增加。

“决策树”节点 - 中止规则

这些选项可控制树的构建方式。 停止规则可确定何时停止分割树的特定分支。 设置最小分支大小可阻止通过分割创建非常小的子组。 如果节点（父级）中要分割的记录数小于指定值，那么**父分支中的最小记录数**将阻止进行分割。 如果由拆分创建的任何分支（子级）中的记录数小于指定值，那么**子分支中的最小记录数**将阻止进行分割。

- **使用百分比**：按总训练数据的百分比指定大小。
- **使用绝对值**：按绝对记录数指定大小。

“决策树”节点 - 整体

这些设置决定了在“目标”中请求 Boosting、Bagging 或超大型数据集时发生的整体行为。 将忽略不适用于选定目标的选项。

Bagging 和大型数据集 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体评分值。

- **分类目标的缺省合并规则**。 可以通过投票、最高概率或最高均值概率来对分类目标的整体预测值进行组合。 **投票**选择在基本模型中最常具有最高概率的类别。 **最高概率**选择在所有基本模型中达到单一最高概率的类别。 **最高平均概率**选择当类别概率在基本模型中取平均值时具有最高值的类别。
- **连续目标的缺省合并规则**。 可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，那么组合规则选择将被忽略。 Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

提升和组装。 当以增强模型精确性或稳定性为目标时，指定要构建的基本模型数；对于 Bagging 方法，此为 bootstrap 样本数。 它应为正整数。

C&R 树和 QUEST 节点 - 成本和先验

误分类成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

除 C5.0 模型之外，在对模型进行评分时，误分类成本是不适用的；在使用自动分类器节点、评估图表或分析节点对模型进行排秩或比较时，误分类成本也不予以考虑。 将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的 成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

先验

通过这些选项可以在预测分类目标字段时为分类指定先验概率。 **先验概率**是对总体（从中抽取训练数据）中每个目标分类的总相对频率的估计。 换句话说，先验概率是对预测变量值有任何了解 之前对每个可能的目标值的概率估计。 有三种方法用来设置先验概率：

- **基于训练数据**。 这是缺省选项。 先验概率基于训练数据中分类的相对频率。
- **对所有的类都相等**。 所有类别的先验概率都定义为 1/*k*，其中 *k* 是目标分类数。
- **定制**。 您可以自行指定先验概率。 对于所有类，都将先验概率的初值设置为相等。 可以将单个分类的概率调整为用户定义的值。 要调整特定分类的概率，可在表中对应于所需分类的概率单元格中，先清除其内容，然后输入所需的值。

所有分类的先验概率之和应为 1.0（ **概率约束** ）。 如果权重之和不为 1.0，将出现一个警告，显示带有自动标准化这些值的选项。 此自动调整操作可在强制执行概率约束时保留分类中的比例。 通过单击 **标准化** 按钮，可在任何时间执行此调整。 将表中所有分类重置为相同的值，可单击 **均衡** 按钮。

使用误分类成本调整先验概率。 通过此选项可以根据误分类成本（在“成本”选项卡中指定）调整先验概率。 从而可为使用两分杂质测量的树将损失信息直接合并到树生成过程中。 （未选中此选项时，损失信息仅用于为基于两分测量的树分类记录和计算风险估计。）

CHAID 节点 - 成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

除 C5.0 模型之外，在对模型进行评分时，误分类成本是不适用的；在使用自动分类器节点、评估图表或分析节点对模型进行排秩或比较时，误分类成本也不予以考虑。 将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的 成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

C&R 树节点 - 高级

通过高级选项，您可以对树构建过程进行微调。

杂质最小变化。 指定最小杂质改变以便在树中创建新的分割。 **杂质**是指由树定义的子组在每个组中所具有的输出字段值的广度。 对于分类目标，如果节点中 100% 的观测值都落在目标字段的特定类别中，那么该节点被认为是“纯节点”。 树构建的目的是创建具有相似输出值的子组 - 换句话说，是为了减少每个节点中的杂质。 如果某个分支的最佳分割按小于指定值的数量减少杂质，那么不会进行此分割。

分类目标的杂质度量。 对于分类目标字段，指定用于测量树的杂质的方法。 （对于连续目标，将忽略此选项，而通常会使用 **最小平方差** 杂质测量。）

- **吉尼**是基于分支的类别成员资格概率的一般杂质测量。
- **两分**是强调二元分割并更有可能导致从分割中生成大小近似相同的分支的杂质测量。
- **有序**添加了额外的限制，即只有相邻的目标类才可以组成一组，此选项仅适用于有序目标。 如果对于名义目标选中此选项，将缺省使用标准的两分测量。

防止过度拟合集合。 该算法在内部将记录划分为模型构建集合和防止过度拟合集合，后者作为独立的数据记录集，用于跟踪训练过程中的错误，以防止该方法对数据中的几率变异进行建模。 指定记录的百分比。 缺省值为 30。

复制结果。 通过设置随机种子，您可以复制分析。 指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。

QUEST 节点 - 高级

通过高级选项，您可以对树构建过程进行微调。

分割的显著性水平。 指定用于分割节点的显著性水平 (Alpha)。 此值必须在 0 到 1 之间。 较小的值往往会产生具有较少节点的树。

防止过度拟合集合。 该算法在内部将记录划分为模型构建集合和防止过度拟合集合，后者作为独立的数据记录集，用于跟踪训练过程中的错误，以防止该方法对数据中的几率变异进行建模。 指定记录的百分比。 缺省值为 30。

复制结果。 通过设置随机种子，您可以复制分析。 指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。

CHAID 节点 - 高级

通过高级选项，您可以对树构建过程进行微调。

分割的显著性水平。 指定用于分割节点的显著性水平 (Alpha)。 此值必须在 0 到 1 之间。 较小的值往往会产生具有较少节点的树。

合并的显著性水平。 指定用于合并类别的显著性水平 (Alpha)。 值必须大于 0 且小于或等于 1。 要阻止对类别进行任何合并，请指定值 1。 对于连续目标，这意味着最终树中变量的类别数与指定的间隔数相匹配。 此选项对于 Exhaustive CHAID 不适用。

使用 Bonferroni 方法调整显著性值。 在检验预测变量的各种类别组合时调整显著性值。 显著相关值可基于检验次数进行调整，而检验次数直接与预测变量的类别数及测量级别相关。 通常需要选中此选项，因为它可以更好地控制假阳性错误率。 禁用此选项将提高您的分析能力以找到真差分，但以增加假阳性率为代价。 建议您禁用此选项，尤其对于较小的样本。

允许在节点内重新拆分已合并的类别。 CHAID 算法尝试合并类别以生成用于描述模型的最简单的树。 如果选中此选项，并且合并后的结果能够比较好地描述模型，那么可以重新分割已合并的类别。

分类目标的卡方。 对于类别目标，您可以指定用于计算卡方统计量的方法。

- **Pearson**。 此方法提供更快的计算，但是对于小样本应该谨慎使用它。

- **似然比**。 与 Pearson 方法相比，此方法更加稳健，但计算时间更长。 对于小样本，这是首选的方法。 对于连续目标，将始终使用此方法。

期望单元格频率的最小更改。 （为名义模型和行效应顺序模型）估计单元格频率时，迭代过程 (epsilon) 用于对最优估计（在特定分割的卡方检验中使用）进行收敛。 Epsilon 可确定必须对迭代进行多大的更改才可使其继续；如果对最后一个迭代的更改小于指定的值，那么迭代将停止。 如果因算法中存在问题而无法收敛，那么可以增加该值或增加最大迭代次数，直到发生收敛为止。

收敛的最大迭代次数。 指定停止前的最大迭代次数，而不考虑是否已进行收敛。

防止过度拟合集合。 （只有在使用交互树构建器时，此选项才可用。） 该算法在内部将记录划分为模型构建集合和防止过度拟合集合，后者作为独立的数据记录集，用于跟踪训练过程中的错误，以防止该方法对数据中的几率变异进行建模。 指定记录的百分比。 缺省值为 30。

复制结果。 通过设置随机种子，您可以复制分析。 指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。

决策树节点模型选项

在“模型选项”选项卡上，您可以选择是指定模型名称，还是自动生成名称。 还可以选择获取预测变量重要性信息，以及标志目标的原始倾向评分和调整后的倾向评分。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

模型评估

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此缺省情况下，对于某些模型，预测变量重要性均处于关闭状态。 预测变量重要性对于决策列表模型不可用。 有关更多信息，请参阅第 32 页的『预测变量重要性』。

倾向评分

可以在建模节点中和模型块的“设置”选项卡上启用倾向评分。 该功能仅在所选目标为标志字段时才可用。 有关更多信息，请参阅主题 第 26 页的『倾向评分』。

计算原始倾向评分。 原始倾向评分仅派生自基于训练数据的模型。 如果模型预测值为 真（将响应），那么倾向与 P 相同，其中 P 为预测的可能性。 如果模型预测的值为假，那么计算出的倾向为 (1 – P)。

- 如果构建模型时选择了此选项，那么缺省情况下将在模型块中启用倾向评分。 不过，无论是否在建模节点中选择了原始倾向评分，都可以始终在模型块中选择启用原始倾向评分。
- 对模型进行评分时，原始倾向评分将被添加到将 *RP* 字母附加到标准前缀的字段中。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RRP-churn*。

计算调整后的倾向评分。 原始倾向仅基于由可能过度拟合的模型指定的估计，这将导致过于乐观地估计倾向。 调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。
- 与原始置信度分数不同，调整后的倾向评分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 *AP* 字母附加到标准前缀的字段中添加调整后的倾向评分。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RAP-churn*。 调整后的倾向评分不适用于 logistic 回归模型。
- 在计算调整后的倾向评分时，必须尚未平衡用于计算的检验或验证分区。 为避免这一点，请确保在任何上游平衡节点中选中 **仅平衡训练数据** 选项。 此外，如果已在上游获取了复杂样本，那么这将导致调整后的倾向评分无效。
- 调整后的倾向评分不适用于“增强型”树和规则集模型。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。

基于。 对于要进行计算的调整后的倾向评分，流中必须存在一个分区字段。 可以指定是使用检验分区还是验证分区进行此计算。 为获取最佳结果，检验或验证分区包含的记录数量应至少与用于训练原始模型的分区所包含的记录数相同。

C5.0 节点

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

该节点使用 C5.0 算法构建 **决策树** 或 **规则集** 。 C5.0 模型的工作原理是根据提供最大 **信息增益** 的字段分割样本。 然后通常会根据不同的字段再次分割由第一次分割定义的每个子样本，且此过程会重复下去直到无法继续分割子样本。 最后，将重新检查最底层分割，并删除或 **修剪** 对模型值没有显著影响的分割。

注：C5.0 节点只能预测分类目标。 分析包含分类（名义或有序）字段的数据时，与 11.0 版以前的 C5.0 版本相比将类别组合在一起的可能性更大。

C5.0 可以生成两种模型。 **决策树** 是对由算法建立的分割的简单描述。 每个终端（或“叶”）节点可描述训练数据的特定子集，而训练数据中的每个观测值都完全属于树中的某个终端节点。 换句话说，对于在决策树中显示的任何特定数据记录，仅可能有一个预测。

反过来， **规则集** 则是尝试对单个记录进行预测的一组规则。 规则集源自决策树，并且在某种程度上表示在决策树中建立的经简化或提取的信息版本。 通常，规则集可保留完整的决策树中的大部分重要信息，但其使用的模型比较简单。 由于规则集的这种工作方式，其属性与决策树的属性不同。 最重要的区别是，使用规则集时，可以为任意特定记录应用多个规则，也可以不应用任何规则。 如果应用多个规则，则每个规则将根据与此规则关联的置信度获得一个加权“投票”，并通过组合应用到所讨论记录的所有规则的加权投票来确定最终的预测。 如果没有规则可应用，那么会将缺省预测分配到该记录。

示例。 医学研究员已收集一组患有相同疾病的患者的相关数据。 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。 您可以将 C5.0 模型与其他节点结合使用，以帮助找出可能适用于今后患有相同疾病的患者的药物。

需求。 要训练 C5.0 模型，必须有一个分类（即名义或有序）目标字段和一个或多个任意类型的输入字段。 设置为 双向或 无的字段将忽略。 必须对模型中使用的字段的类型完全实例化。 另外，还可以指定权重字段。

强度。 遇到缺少数据及存在大量输入字段等问题时，C5.0 模型的表现十分稳健。 这些模型通常不需要花费很长的训练时间用于估计。 此外，C5.0 模型与某些其他模型类型相比似乎更容易理解，因为源自模型的规则解释起来更简明易懂。 C5.0 还提供功能强大的 **增强** 方法来提高分类的准确性。

注: 启用并行处理有助于加快 C5.0 模型构建速度。

C5.0 节点模型选项

模型名称。 指定要生成的模型的名称。

- **自动**。 在选中此选项的情况下，将根据目标字段名称自动生成模型名称。 这是缺省选项。
- **定制**。 选中此选项可以为此节点将创建的模型块指定定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

输出类型。 在此处指定您希望生成的模型块是**决策树**还是**规则集**。

组符号。 如果选中了此选项，那么 C5.0 将尝试对输出字段具有相似模式的符号值进行组合。 如果未选中此选项，C5.0 将为用于分割父节点的符号字段的每个值创建一个子节点。 例如，如果 C5.0 分割的是 颜色字段（其值为 红色、绿色和 蓝色），则它将缺省创建一个三向分割。 但是，如果选中此选项，且 颜色 = 红色的记录与 颜色 = 蓝色的记录非常相似，则 C5.0 将创建一个双向分割，其中所有 绿色记录在一个组中，而所有蓝色记录连同所有 红色记录在另一个组中。

使用增强。 C5.0 算法包含一个用于提高其准确率的特殊方法，称为**增强**。 它的工作原理是在序列中构建多个模型。 第一个模型按常规方式进行构建。 然后，以这种方式构建第二个模型时，重点集中于被第一个模型误分类的记录。 构建第三个模型时，将焦点集中于第二个模型的错误，依此类推。 最后，通过将整个模型集应用到观测值，并使用加权投票过程将单独的预测组合为一个总预测来分类观测值。 增强方法可以显著

提高 C5.0 模型的准确性，但也需要更长的训练时间。 通过**尝试次数**选项，您可以控制用于增强型模型的模型数。 该功能基于 Freund 和 Schapire 的研究，包含一些用于更好地处理噪声数据的专利改进。

交叉验证。 如果选中此选项，那么 C5.0 将使用一组根据训练数据的子集构建的模型来估算根据整个数据集构建的模型的准确性。 如果数据集太小以致于无法将其分割为传统的训练集合和测试集合，此选项非常有用。 在计算准确性评估后，交叉验证模型将被丢弃。 可以指定用于交叉验证的 **折叠次数** 或模型数。 注意，在 IBM SPSS Modeler 以前的版本中，构建模型和交叉验证模型是两个单独的操作。 在当前的版本中，则无需执行单独的模型构建步骤。 模型构建和交叉验证将同时执行。

方式。 对于**简单**训练，大多数 C5.0 参数是自动设置的。 **专家**训练允许更直接地控制训练参数。

简单模式选项

偏爱。 缺省情况下，C5.0 将尝试尽可能生成最准确的树。 在某些情况下，此操作可能会导致过度拟合，从而在将此模型应用于新数据时导致性能偏低。 选择**普遍性**以使用受此问题影响较小的算法设置。

注：不保证在选中**普遍性**选项的情况下构建的模型的适用性优于其他模型。 当普遍性问题比较严重时，通常可使用保留检验样本验证模型。

预期噪声 (%)。 指定训练集中噪声数据或错误数据的预期比例。

专家模式选项

修剪严重性。 确定决策树或规则集的修剪程度。 增加该值可获得一个更简洁的小型树。 减小该值可获得一个更精确的树。 此设置仅影响本地修剪（请参见下面的“使用全局修剪”）。

每个子分支的最小记录数。 可以使用子组的大小来限制树的任何分支中的分割数。 仅当两个或多个生成的子分支中至少包含从训练集合得到的这一最小记录数时，才可分割树的分支。 缺省值为 2。 增大此值以帮助防止使用噪声数据进行**过度训练**。

使用全局修剪。 树的修剪分为两个阶段：第一个阶段是本地修剪，将检查子树并折叠分支以提高模型的准确性。 第二个阶段是全局修剪，在此阶段中将把树视作一个整体并折叠虚弱的子树。 缺省情况下将执行全局修剪。 要忽略全局修剪阶段，请取消选中此选项。

辨别属性。 如果选中此选项，那么 C5.0 将在开始构建模型前检查预测变量的有效性。 如果发现不相关的预测变量，则会将其从模型构建过程中排除。 此选项对于具有许多预测变量字段的模型非常有用，并且有助于防止过度拟合。

注: 启用并行处理有助于加快 C5.0 模型构建速度。

树-AS 节点

“树 AS”节点可以与分布式环境中的数据配合使用。 在此节点中，您可以使用 CHAID 或 Exhaustive CHAID 模型来构建决策树。

CHAID 或卡方自动交互效应检测是一种通过使用卡方统计量识别最优分割来构建决策树的分类方法。

CHAID 首先检查每个输入字段和结果之间的交叉表，然后使用卡方独立性检验来检验显著性。 如果以上多个关系具有显著的统计意义，那么 CHAID 将选择最重要（*p* 值最小）的输入字段。 如果输入具有两个以上的类别，那么将会对这些类别进行比较，然后将结果中未显示出差异的类别合并在一起。 此操作通过将显示的显著性差异最低的类别对相继合并在一起来实现。 当所有剩余类别在指定的检验级别上存在差异时，此类别合并过程将终止。 对于名义输入字段，可以合并任何类别；对于有序集合，只能合并相邻的类别。

Exhaustive CHAID 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

需求。 目标和输入字段可以是连续字段，也可以是分类字段；节点在每一层上都可以分割为两个或多个子组。 模型中使用的所有顺序字段的存储类型都必须是数字类型（不是字符串）。 必要的话，可以使用重新分类节点来对其进行转换。

强度。 CHAID 可以生成非二元树，这意味着有些分割将有多于两个的分支。 因此，与二元生成方法相比，CHAID 倾向于创建范围更广的树。 CHAID 适用于所有类型的输入，并且接受观测值权重和频率变量。

树-AS 节点字段选项

在“字段”选项卡上，可以选择是要使用在上游节点中定义的字段角色设置，还是手动进行字段分配。

使用预定义角色：此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标、预测变量和其他角色，请选中此选项。

字段。 使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。 图标表示每个角色字段的有效测量级别。

要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择一个字段作为预测目标。

预测变量：选择一个或多个字段作为预测输入。

分析权重：要使用字段作为观测值权重，在此处指定。 观测值权重将作为对输出字段各个水平上方差的差异的一种考量。 有关更多信息，请参阅 第 24 页的『使用频率和权重字段』。

树-AS 节点构建选项

通过“构建选项”选项卡，您可以设置构建模型的所有选项。 当然，您只需单击**运行**按钮，即可采用所有缺省选项来构建模型；不过，通常您需要根据具体用途定制构建选项。

该选项卡包含多个不同的窗口，您可以在其中设置特定于自己的模型的定制。

树-AS 节点 - 基础

指定有关如何构建决策树的基本选项。

树生长算法：选择您要使用的 **CHAID** 算法类型。 **Exhaustive CHAID** 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

最大数深度 指定根节点以下的最大级别数（以递归方式拆分样本的次数）；缺省值为 5。 最大级别数（也称为节点数）为 50,000 个。

分级：如果您使用连续数据，您必须对输入进行分级。 您可以在前一个节点中执行此操作；但是，树-AS 节点会对任何连续输入进行自动分级。 如果您使用树-AS 节点来自动对数据进行分级，请选择要对输入进行分割的**分级数**。 数据将按等频率分级；可用选项为 2、4、5、10、20、25、50 或 100。

树-AS 节点 - 生长

使用生长选项来对树构建过程进行微调。

从 p 值切换至效应大小的记录阈值：指定在构建树时，模型将从使用 **P 值设置**切换至**效应大小设置**的记录数。 缺省值为 1,000,000。

拆分的显著性水平：指定用于拆分节点的显著性水平 (Alpha)。 该值必须介于 0.01 和 0.99 之间。 较小的值往往会产生具有较少节点的树。

合并的显著性水平：指定用于合并类别的显著性水平 (Alpha)。 该值必须介于 0.01 和 0.99 之间。 此选项对于 Exhaustive CHAID 不适用。

使用 Bonferroni 法调整显著性值：测试预测变量的各种类别组合时，调整显著性值。 显著相关值可基于检验次数进行调整，而检验次数直接与预测变量的类别数及测量级别相关。 通常需要选中此选项，因为它可以更好地控制假阳性错误率。 禁用此选项将提高您的分析能力以找到真差分，但以增加假阳性率为代价。 建议您禁用此选项，尤其对于较小的样本。

效应大小阈值（仅限连续目标）：设置使用连续目标时，拆分节点和合并类别时要使用的效应大小阈值。 该值必须介于 0.01 和 0.99 之间。

效应大小阈值（仅限分类目标）：设置使用分类目标时，拆分节点和合并类别时要使用的效应大小阈值。 该值必须介于 0.01 和 0.99 之间。

允许重新拆分节点内的已合并类别：CHAID 算法尝试合并类别以生成用于描述模型的最简单的树。 如果选中此选项，并且合并后的结果能够比较好地描述模型，那么可以重新分割已合并的类别。

叶节点分组的显著性水平：指定确定如何形成叶节点分组或者如何识别不正常的叶节点的显著性水平。

分类目标的卡方：对于类别目标，您可以指定用于计算卡方统计量的方法。

- **Pearson**：此方法提供更快的计算，但是对于小样本应该谨慎使用它。
- **似然比检验**：与 Pearson 方法相比，此方法更加稳健，但计算时间更长。 对于小样本，这是首选的方法。 对于连续目标，将始终使用此方法。

树-AS 节点 - 停止规则

这些选项可控制树的构建方式。 停止规则可确定何时停止分割树的特定分支。 设置最小分支大小可阻止通过分割创建非常小的子组。 如果节点（父级）中要分割的记录数小于指定值，那么**父分支中的最小记录数**将阻止进行分割。 如果由拆分创建的任何分支（子级）中的记录数小于指定值，那么**子分支中的最小记录数**将阻止进行分割。

- **使用百分比**：按总训练数据的百分比指定大小。
- **使用绝对值**：按绝对记录数指定大小。

期望单元格频率的最小变化值：（为名义模型和行效应顺序模型）估计单元格频率时，迭代过程 (epsilon) 用于对最优估计（在特定分割的卡方检验中使用）进行收敛。 Epsilon 可确定必须对迭代进行多大的更改才可使其继续；如果对最后一个迭代的更改小于指定的值，那么迭代将停止。 如果因算法中存在问题而无法收敛，那么可以增加该值或增加最大迭代次数，直到发生收敛为止。

收敛的最大迭代次数：指定停止前的最大迭代次数，而不考虑是否已进行收敛。

树-AS 节点 - 成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

（仅限有序目标）您可以选择**针对有序目标的缺省成本增加**，并在成本矩阵中设置缺省值。 以下列表中描述了可用选项。

- **无增加** - 针对每个正确的预测使用缺省值 1.0。
- **线性** - 每个后续非正确预测增加成本 1。
- **平方** - 每个后续非正确预测为线性值的平方。 在此情况下，值包括：1、4、9 等。
- **定制** - 如果您在表中手动编辑任何值，那么下拉选项会自动更改为**定制**。 如果您将下拉选项更改为任何其他选项，那么所选选项的值会替代您编辑的值。

树-AS 节点模型选项

在“模型选项”选项卡上，您可以选择是指定模型名称，还是自动生成名称。 您也可以选择计算置信度值，并在对模型评分期间添加标识。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

计算置信度：要在对模型进行评分时添加置信度字段，请选中此复选框。

规则标识：要在对模型进行评分时添加包含记录分配到的叶节点标识的字段，请选中该字段。

树-AS 模型块

树-AS 模型块输出

在创建树-AS 模型后，在输出查看器中提供了以下信息。

模型信息表

“模型信息”表提供了有关模型的关键信息。 该表标识了一些高级模型设置，例如：

- 使用的算法类型；CHAID 或 Exhaustive CHAID。
- “类型”节点或树-AS 节点“字段”选项卡中选中的目标字段的名称。
- 在“类型”节点或树-AS 节点“字段”选项卡中，选择作为预测变量的字段名称。
- 数据中的记录数。 如果您使用频率权重构建模型，那么此值为已加权的有效计数，它表示树基于的记录数。
- 生成的树中叶节点的数量。
- 树种的层数：即，树深度。

预测变量重要性

预测变量重要性图形以条形图的形式显示模型中前 10 个输入（预测变量）的重要性。

如果图表中存在超过 10 个字段，那么可以使用图表下的滑块来调整图表中包含的预测变量的选择。 滑块上的指示符标记为固定宽度，滑块上每个标记表示 10 个字段。 您可以沿滑块移动指示符标记，以显示后 10 个或前 10 个字段（按预测变量重要性排序）。

您可以双击图表以打开单独的对话框，您可以在其中编辑图形设置。 例如，您可以修改项目（如图形大小以及使用的字体大小和颜色）。 关闭此单独的编辑对话框后，更改会应用于“输出”选项卡中显示的图表。

主要决策规则表

缺省情况下，此交互式表格会基于叶节点中包含的合计记录的百分比，显示输出中前五个叶节点的规则的统计信息。

您可以双击该表，以打开单独的对话框，您可以在其中编辑表中显示的规则信息。 对话框中显示的信息以及提供的选项取决于目标类型；例如，分类目标或连续目标。

在表中会显示以下规则信息：

- 规则标识(U)
- 规则应用和组成的方式的详细信息
- 每项规则的记录计数。 如果您使用频率权重构建模型，那么此值为已加权的有效计数，它表示树基于的记录数。
- 每条规则的记录百分比

此外，对于连续目标，表中包含一个额外的列，其中显示每条规则的**平均**值。

您可以使用以下**表内容**选项来更改规则表布局：

- **主要决策规则**：按叶节点中包含的合计记录百分比排序的前五条决策规则。
- **所有规则**：该表包含模型生成的所有叶节点，但每页仅显示 20 条规则。 选中此布局时，可以使用额外的**按标识查找规则**和**页面**选项来搜索规则。

此外，对于分类目标，您可以通过使用**按类别分类的主要规则**选项来更改规则表的布局。 前五条决策规则是按照针对您选择的**目标类别**的合计记录百分比进行排序的。

如果您更改规则表的布局，那么可以通过单击位于对话框左上角的“复制到查看器”按钮来讲修改后的规则表复制回“输出查看器”。

树-AS 模型块设置

在树-AS 模型块的“设置”选项卡上，可以在模型评分期间指定用于置信度及 SQL 生成的选项。 只有将模型块添加到流之后，此选项卡才可用。

计算置信度：要在评分操作中包括置信度，请选中此复选框。 在数据库中评分模型时，排除置信度意味着可以生成更有效的 SQL。 不会为回归树分配置信度。

规则标识：要在评分输出中添加一个字段，表示每个记录分配到的终端节点的标识，请选中此复选框。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定生成 SQL 的方式：

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**：如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

“随机树”节点

“随机树”节点可以与分布式环境中的数据配合使用。 此节点中，您可以构建包含多个决策树的整体模型。

“随机树”节点是一种基于树的分类和预测方法，此方法根据分类和回归方法构建。 与 C&R 树类似，此预测方法使用递归分区将训练记录分割为具有相似输出字段值的段。 首先，此节点通过检查可供其使用的输入字段来查找最佳分割（以分割所引起的杂质指标下降情况进行测量）。 分割可定义两个子组，其中每个子组随后又分割为两个子组，依此类推，直到触发其中一项中止条件为止。 所有分割都是二元的（仅有两个子组）。

“随机树”节点使用进行替换的拔靴法采样来生成样本数据。 样本数据用于生成树模型。 在树生长期间，“随机树”不会再次进行数据采样。 相反，它会随机选择部分预测变量，并使用最佳的预测变量来分割树节点。 分割每个树节点时，都会重复此过程。 这是在随机林中生成树的基本构想。

“随机树”使用类似于 C&R 树的树。 由于此类树是二叉树，用于分割的每个字段都会产生两个分支。 对于具有多个类别的分类字段，各个类别将根据内部分割条件分为两组。 每个树都尽可能成长到最大范围（不进行修剪）。 进行评分时，“随机树”通过多数表决（对于分类）或平均值（对于回归）来组合各个树的分数。

随机树与 C&R 树有所差异，如下所示：

- “随机树”节点随机选择指定数目的预测变量，并使用所选变量中最佳的变量来分割节点。 与之相对，“C&R 树”从所有预测变量中寻找最佳变量。
- “随机树”中的每个树都充分生成，直到每个叶节点都包含单个记录为止。 因此，树深度可能会非常大。 但是，标准的“C&R 树”对于树生长使用不同的中止规则，这通常会使树的深度较浅。

与 C&R 树相比，随机树将添加两项功能：

- 第一项功能是组装，其中训练数据集的副本是通过对原始数据集进行放回抽样来创建的。 此操作将大小与原始数据集相等的 Bootstrap 样本，在此操作执行后将根据每个副本构建组件模型。 这些成分模型共同构成一个整体模型。
- 第二项功能是，在树的每个分割处仅考虑将输入字段采样进行杂质测量。

需求。 要训练“随机树”模型，您需要一个或多个输入字段以及一个目标字段。 目标字段和输入字段可以是连续字段（数字范围），也可以是分类字段。 将忽略设置为两者或无的字段。 对于模型中使用的字段，必须将它们的类型完全实例化，并且模型中使用的任何有序（有序集合）字段的存储类型必须是数字类型（而不是字符串）。 必要的话，可以使用重新分类节点对存储类型进行转换。

强度。 处理大型数据集和许多字段时，“随机树”模型是稳健的模型。 由于使用组装和字段采样，因此它们更不容易过度拟合，并且测试中看到的结果更可能在您使用新数据时重复。

“随机树”节点字段选项

在“字段”选项卡上，可以选择是要使用在上游节点中定义的字段角色设置，还是手动进行字段分配。

使用预定义角色：此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标、预测变量和其他角色，请选中此选项。

字段。 使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。 图标表示每个角色字段的有效测量级别。

要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择一个字段作为预测目标。

预测变量：选择一个或多个字段作为预测输入。

分析权重：要使用字段作为观测值权重，在此处指定。 观测值权重将作为对输出字段各个水平上方差的差异的一种考量。 有关更多信息，请参阅 第 24 页的『使用频率和权重字段』。

“随机树”节点构建选项

通过“构建选项”选项卡，您可以设置构建模型的所有选项。 当然，您只需单击**运行**按钮，即可采用所有缺省选项来构建模型；不过，通常您需要根据具体用途定制构建选项。

该选项卡包含多个不同的窗口，您可以在其中设置特定于自己的模型的定制。

“随机树”节点 - 基本

指定有关如何构建决策树的基本选项。

要构建的模型树。 指定节点可以构建的最大树数量。

样本大小。 缺省情况下，Bootstrap 样本的大小等于原始训练数据。 处理大型数据集时，减少样本大小可以提高性能。 这是从 0 到 1 的比率。 例如，将样本大小设置为 0.6 以使其降低至原始训练数据大小的 60%。

处理不平衡数据。 如果模型的目标是标志结果（例如，购买或不购买）并且所需结果与非所需结果的比率很小，那么数据是不平衡数据并且模型所执行的 Bootstrap 采样可能会影响模型精确性。 要提高准确性，请选中此复选框；模型随后会捕获所需结果中的更大比例部分并生成更好的模型。

将加权采样用于变量选择。 缺省情况下，每个叶节点的变量是使用同一概率随机选择的。 要将加权用于变量并改进选择过程，请选中此复选框。 权重是由“随机树”节点本身计算的。 字段的重要性越高（权重越高），越容易被选作预测变量。

最大节点数。 指定允许各个树中存在的最大叶节点数。 如果下一次分割时将超过此数字，那么树增长将在进行拆分之前停止。

最大树深度。 指定根节点下方的最大叶节点级别数；即，样本进行递归分割的次数。

最小子节点大小。 指定分割父节点之后必须包含在子节点中的最小记录数。 如果子节点包含的记录数少于您输入的数目，那么不会拆分父节点。

指定要用于分割的预测变量数。 如果是构建分割模型，请设置要用于构建每次分割的最小预测变量数。 这防止拆分创建过小的子组。 如果不选择此选项，分类的缺省值为 $\lfloor sqrt(M) \rfloor$，回归的缺省值为 $\lfloor M/3 \rfloor$，其中 M 是预测变量总数。 如果选择此选项，将使用指定数量的预测变量。

注: 用于分割的预测变量数不能大于数据中的预测变量总数。

在准确性无法再提高时停止构建。 “随机树”使用特定过程来决定何时停止训练。 具体来说，如果当前整体准确性的提高程度小于指定阈值，那么将停止添加新的树。 这可能导致生成的模型中所包含的树少于您为**要构建的模型数**选项指定的值。

“随机树”节点 - 成本

在某些环境中，特定错误类别的成本高于其他错误的成本。 例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。 使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。 这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。 缺省情况下，所有误分类成本都设置为 1.0。 要输入定制成本值，可选择 **使用误分类成本** 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。 成本不会自动均摊。 例如，如果将 *A* 误分类为 *B* 的成本设置为 2.0，那么将 *B* 误分类为 *A* 的成本将仍是缺省值 1.0，除非也明确地对它进行更改。

（仅限有序目标）您可以选择**针对有序目标的缺省成本增加**，并在成本矩阵中设置缺省值。 以下列表中描述了可用选项。

- **无增加** - 对于每次不正确的预测，缺省值为 1.0。
- **线性** - 每个后续非正确预测增加成本 1。
- **平方** - 每个后续非正确预测为线性值的平方。 在此情况下，值包括：1、4、9 等。
- **定制** - 如果您在表中手动编辑任何值，那么下拉选项会自动更改为**定制**。 如果您将下拉选项更改为任何其他选项，那么所选选项的值会替代您编辑的值。

“随机树”节点 - 高级

指定有关如何构建决策树的高级选项。

缺失值的最大百分比。 指定任何输入中允许的缺失值的最大百分比。 如果该百分比超过了此数字，那么将从模型构建中排除此输出。

排除单个类别多数值高于以下值的字段。 指定单个类别可以在某个字段中具有的最大记录百分比。 如果任何类别值表示的记录百分比高于指定值，那么将从模型构建中排除整个字段。

字段类别的最大数目。 指定字段中可以包含的最大类别数。 如果类别数超过了此数字，那么将从模型构建中排除此字段。

最小字段变异。 如果某个连续字段的变异系数小于您在此处指定的值（换言之，该字段几乎不变），那么将从模型构建中排除此字段。

分级数。 请指定要用于连续输入的均等频率分级数。 可用选项包括：2、4、5、10、20、25、50 或 100。

要报告的有趣规则数量。 指定要报告的规则数量（最小值为 1，最大值为 1000，缺省值为 50）。

“随机树”节点模型选项

在“模型选项”选项卡上，您可以选择是指定模型名称，还是自动生成名称。 对模型进行评分的过程中，您还可以选择计算预测变量的重要性。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

随机树模型块

随机树模型块输出

创建随机树模型之后，输出查看器中提供了以下信息：

模型信息表

模型信息表提供关于模型的关键信息。 此表始终包含以下高级模型设置：

- 在“类型”节点或“随机树”节点字段选项卡中选择的目标字段的名称。
- 模型构建方法 - 随机树。
- 输入模型中的预测变量数。

表中显示的其他详细信息取决于您构建的是分类模型还是回归模型以及构建模型是否旨在处理不平衡数据：

- 分类模型（缺省设置）
  - 模型精确性
  - 误分类规则
- 分类模型（已选择**处理不平衡数据**）
  - Gmean
  - 真阳性率（细分为多个类别）。
- 回归模型
  - 均方根误差
  - 相对误差
  - 解释的方差

记录摘要

此摘要显示用于拟合模型的记录数以及排除的记录数。 将显示这两种记录数以及整数所占百分比。 如果构建模型旨在包括频率权重，那么还将显示包括和排除的记录的未加权数目。

预测变量重要性

预测变量重要性图形以条形图的形式显示模型中前 10 个输入（预测变量）的重要性。

如果图表中存在超过 10 个字段，那么可以使用图表下的滑块来调整图表中包含的预测变量的选择。 滑块上的指示符标记为固定宽度，滑块上每个标记表示 10 个字段。 您可以沿滑块移动指示符标记，以显示后 10 个或前 10 个字段（按预测变量重要性排序）。

您可以双击图表以打开单独的对话框，您可以在其中编辑图形大小。 关闭此单独的编辑对话框后，更改会应用于“输出”选项卡中显示的图表。

主要决策规则表

缺省情况下，此交互表显示按相关度排序的主要规则的统计信息。

您可以双击该表，以打开单独的对话框，您可以在其中编辑表中显示的规则信息。 对话框中显示的信息以及提供的选项取决于目标类型；例如，分类目标或连续目标。

在表中会显示以下规则信息：

- 规则应用和组成的方式的详细信息
- 如果结果属于最频繁的类别
- 规则准确性
- 树准确性
- 相关度索引

相关度索引将使用以下公式进行计算：

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

在此公式中：

- P(A(t)) 是树准确性
- P(B(t)) 是规则准确性
- P(B(t)|A(t)) 表示树和节点由树和节点进行的正确预测
- 此公式的其余部分表示由树和节点进行的不正确预测

您可以使用下列**表内容**选项来变更规则表：

- **主要决策规则** 按相关度索引排序的前五条主要决策规则。
- **所有规则** 该表包含由模型生成的所有规则，但每页仅显示 20 条规则。 选中此布局时，可以使用额外的**按标识查找规则**和**页面**选项来搜索规则。

另外，对于分类目标，您可以使用**按类别列出的主要规则**选项来变更规则表布局。 前五条决策规则是按照针对您选择的**目标类别**的合计记录百分比进行排序的。

注: 对于分类目标，仅当未在构建选项的“基本”选项卡中选择**处理不平衡数据**时，此表才可用。

如果更改了规则表的布局，那么通过单击对话框左上角的“复制到查看器”按钮，您可以将修改后的规则表复制回输出查看器。

混淆矩阵

对于分类模型，混淆矩阵显示预测结果数与实际观测结果数，包括正确预测所占的比例。

注: 混淆矩阵不适用于回归模型，并且在构建选项的“基本”选项卡上选择了**处理不平衡数据**时，也无法使用混淆矩阵。

随机树模型块设置

在随机树模型块的“设置”选项卡上，您可以指定模型评分期间用于置信度和 SQL 生成的选项。 只有将模型块添加到流之后，此选项卡才可用。

计算置信度：要在评分操作中包括置信度，请选中此复选框。 在数据库中评分模型时，排除置信度意味着可以生成更有效的 SQL。 不会为回归树分配置信度。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定生成 SQL 的方式：

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**：如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

C&R 树、CHAID、QUEST 和 C5.0 决策树模型块

决策树模型块表示用于预测其中一个决策树建模节点（C&R 树、CHAID、QUEST 或 C5.0）所发现的特定输出字段的树结构。 树模型可以直接从树构建节点中生成，也可以从交互式树构建器中间接生成。 有关更多信息，请参阅主题 第 60 页的『交互树构建器』。

评分树模型

运行包含树模型块的流时，特定的结果取决于树的类型。

- 对于分类树（分类目标），会将两个新字段（其中分别包含每条记录的预测值和置信度）添加到数据中。 预测取决于为其分配记录的终端节点的使用最频繁的类别；如果在给定节点中大多数响应为 是，那么对分配到该节点的所有记录的预测也为是。
- 对于回归树，仅生成预测值；而不会分配置信度。
- 另外，对于 CHAID、QUEST 和 C&R 树模型，也可以添加表示节点标识的附加字段，每条记录都将分配到此节点中。

新的字段名称将通过为模型名称添加前缀生成。 对于 C&R 树、CHAID 和 QUEST，前缀是 *$R-*（表示预测字段）、*$RC-*（表示置信度字段）和 *$RI-*（表示节点标识符字段）。 对于 C5.0 树，预测字段的前缀是 *$C-*，而置信度字段的前缀是 *$CC-*。 如果存在多个树模型节点，那么必要时新字段名称的前缀中将包含数字以进行区分 - 例如，*$R1-*、*$RC1-* 和 *$R2-*。

使用树模型块

可以多种方式保存或导出与模型相关的信息。

注: 树构建器窗口中也提供了其中的许多选项。

通过树构建器或树模型块，可以执行下列操作：

- 根据当前的树生成过滤节点或选择节点。 有关更多信息，请参阅第 68 页的『生成过滤节点和选择节点』。
- 生成一个规则集块，该节点将树结构表示成一组定义了树的终端分支的规则。 有关更多信息，请参阅第 68 页的『从决策树中生成规则集』。
- 此外，还可以按 PMML 格式导出模型（仅限树模型块）。 有关更多信息，请参阅第 30 页的『模型选用板』。 如果模型包含任何定制分割，那么不会在导出的 PMML 中保留此信息。 （保留分割，但不保留它是定制分割而不是通过算法选择的分割这一事实。）
- 基于当前树的所选部分生成图形。 请注意，仅当块附加到流中的其他节点时，此操作才有效。 有关更多信息，请参阅第 89 页的『生成图形』。
- 仅在增强型 C5.0 模型中，可以选择 **单一决策树（工作区）** 或 **单一决策树（GM 选用板）** 以根据当前选定的规则创建一个新的规则集。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。

注: 虽然 C&R 树节点已替代“构建规则”节点，但现有流中最初使用“构建规则”节点创建的决策树节点仍然正常工作。

单个树模型块

如果在建模节点上选择**构建单个树**作为主目标，那么结果模型块包含下列选项卡。

表 7: 单个树模型块上的选项卡

制表符	描述	其他信息
模型	显示用于定义模型的规则。	有关更多信息，请参阅主题 第 86 页的『决策树模型规则』。
查看器	显示模型的树形视图。	有关更多信息，请参阅主题 第 88 页的『决策树模型查看器』。
目录	显示字段、构建设置和模型估算过程的相关信息。	有关更多信息，请参阅主题 第 32 页的『模型块概要/信息』。
设置	使您可以在模型评分期间为置信度及 SQL 生成指定选项。	有关更多信息，请参阅主题 第 88 页的『决策树/规则集模型块设置』。
注释	使您可以添加描述性注释、指定定制名称、添加工具提示文本以及指定模型的搜索关键字。	

决策树模型规则

决策树模型块的“模型”选项卡显示定义该模型的规则。 此外，还可以显示预测变量重要性的图形和包含有关历史记录、频率和代用项信息的第三个面板。

注: 如果您在 CHAID 节点的“构建选项”选项卡（“目标”面板）上选中**为超大型数据集创建模型**选项，那么“模型”选项卡只显示树规则详细信息。

树规则

左侧面板显示了条件列表，这些条件定义算法发现的数据的分区 - 本质上是一系列规则，可基于不同预测变量的值将单个记录分配给子节点。

决策树基于输入字段值的对数据进行递归分区。 数据分区称为 分支 。 初始分支（有时称为 根 ）包含所有数据记录。 根将根据特定输入字段的值被分成多个子集或 子分支 。 每个子分支可以进一步分割成次级子分

支，次级子分支还可进一步分割，如此类推。 不再分割的分支是树的最底层分支。 这样的分支称为 终端分支 （或 叶片 ）。

树规则详细信息

规则浏览器显示了输入值，输入值定义了每个分区或分支以及这些分割中记录的输出字段值概要。 有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』。

对于基于数值型字段的分割，分支将以下行所示的形式显示：

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

这里的 *relation* 是数值型关系。 例如，由 *revenue* 字段大于 100 的值所定义的分支将显示为如下形式：

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

对于基于符号型字段的分割，分支将以下行所示的形式显示：

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

这里的 *values* 表示定义分支的字段值。 例如，包含 *region* 字段值为 *North*、*West* 或 *South* 的记录的分支将以如下形式表示：

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

终端分支也将进行预测，同时会在规则条件的尾部添加箭头和预测值。 例如，预测输出字段的 *high* 值的 *revenue* > 100 所定义的叶将显示为：

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

数值型和符号型输出字段的分支 概要 定义有所不同。 对于含有数值型输出字段的树，分支的 平均 值便是概要，分支的 效应 便是分支平均值与其父分支平均值的差。 对于含有符号型输出字段的树，分支中记录的 中位数 （或出现频率最高的值）便是概要。

要完全描述分支，需要包含定义分支的条件以及定义树中更深层分割的条件。 例如，在树中：

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

第二条线所代表的分支由条件 *revenue > 100* 和 *region = "North"* 进行定义。

如果单击工具栏上的 **显示实例/置信度**，那么每条规则还将显示其所适用的记录数（实例数）和规则为真的记录所占的比例（置信度）。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。

注: 只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

其他模型信息

如果单击工具栏中的 **显示其他信息面板** ，您将在窗口底部看到显示选定规则详细信息的面板。 信息面板包含三个选项卡。

历史记录。 此选项卡追踪从根节点至选定节点的分割条件。 从而给出了一个条件列表，据此可以判断出记录何时分配给了选定节点。 所有条件均为真的记录将分配给此节点。

频率。 对于含符号型目标字段的模型而言，此选项卡（为每个可能的目标值）显示了分配给包含目标值（训练数据中）节点的记录的数量。 还将显示频率图（显示为最多三位小数的百分比）。 对于含数值型目标的模型，此选项卡为空。

代用项。 如果适用，那么会针对所选节点显示主要分割字段的所有代用项。 代用项是在给定记录的主要预测变量值缺失时使用的替代字段。 给定分割允许的最大代用项数在树构建节点中指定，但实际数量取决于训练数据。 一般来讲，缺失数据越多，可能使用的代用项越多。 对于其他决策树模型，此选项卡为空。

注: 要在模型中包含替代项，必须在训练阶段对其进行标识。 如果训练样本没有缺失值，那么不会标识任何代用项；在测试或评分过程中遇到的具有缺失值的所有记录将自动落入记录数最大的子节点。 如果在测试或评分过程中预期出现缺失值，请确保值在训练样本中也处于缺失状态。 代用项对于 CHAID 树不可用。

效应

节点的效应是平均值的增加或减少（预测值与父节点相比较）。 例如，如果某节点的平均值是 0.2，其父代的平均值是 0.6，那么该节点的效应是 0.2-0.6=-0.4。 此统计仅适用于连续目标。

决策树模型查看器

决策树模型块的“查看器”选项卡类似于树构建器中的显示。 主要的区别是当浏览模型块时，无法生成或修改树。 两个组件中用于查看和定制显示的其他选项都类似。 有关更多信息，请参阅主题 第 62 页的『定制树形视图』。

注: 如果您在“构建选项”选项卡的“目标”面板上选中了**为超大型数据集创建模型**选项，那么对于已构建的 CHAID 模型块，将不会显示“查看器”选项卡。

查看“查看器”选项卡上的分割规则时，方括号表示临界值包含在范围中，而圆括号表示临界值不包含在范围中。 因此，表达式 (23,37] 表示从 23（排除）到 37（包含），即从刚好高于 23 到 37。 在“模型”选项卡上，相同的条件将显示为：

```
Age > 23 and Age <= 37
```

决策树/规则集模型块设置

通过决策树或规则集模型块的“设置”选项卡，您可以在模型评分期间指定用于置信度及 SQL 生成的选项。 只有将模型块添加到流之后，此选项卡才可用。

计算置信度：选中此选项以便在评分操作中包括置信度。 在数据库中评分模型时，排除置信度有助于生成更有效的 SQL。 不会为回归树分配置信度。

注: 如果您在 CHAID 模型的“方法”面板上的“构建选项”选项卡中选中**为超大型数据集创建模型**选项，那么此复选框仅在名义或标志分类目标的模型块中可用。

计算原始倾向评分：对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

注: 如果您在 CHAID 模型的“方法”面板上的“构建选项”选项卡中选中**为超大型数据集创建模型**选项，那么此复选框仅在标志分类目标的模型块中可用。

计算调整后的倾向评分：原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

注: 调整后的倾向评分不适用于增强型树和规则集模型。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。

规则标识 对于 CHAID、QUEST 和 C&R 树模型，此选项会在评分输出中添加一个字段，以指明每个记录所分配到的终端节点的标识。

注: 选中此选项后，SQL 生成将不可用。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **通过转换至无缺失值支持的本机 SQL 来进行评分**：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分，而没有处理缺失值的开销。 此选项仅在对个案进行评分时遇到缺失值的情况下才将预测设置为 null ($null$)。

  注: 此选项对于 CHAID 模型不适用。 对于其他模型类型，此选项仅适用于决策树（而非规则集）。
- **过转换为本机 SQL，使用缺失值支持度进行评分** 对于 CHAID、QUEST 和 C&R 树模型，可生成本机 SQL，使用完全缺失值支持度对数据库中的模型进行评分。 因此，生成 SQL 意味着已按模型中指定的方式处理缺失值。 例如，C&R 树使用替代规则和最大子退回。

  注: 对于 C5.0 模型，此选项仅可用于规则集（而非决策树）。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

增强型 C5.0 模型

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

创建增强型 C5.0 模型（规则集或决策树）时，实际上创建了一组相关模型。 增强型 C5.0 模型的模型规则浏览器显示位于层次结构顶层的模型的列表，以及每个模型的估计准确性和增强型模型的整体的准确性。 要检查特定模型的规则或分割，可选择并根据在单模型中扩展规则或分支的方式扩展该模型。

也可以从增强型模型集中提取特定的模型并创建恰好包含此模型的新规则集模型块。 要从增强型 C5.0 模型中创建新的规则集，可选择所需规则集或树，并从“生成”菜单中选择**单一决策树（GM 选用板）**或**单一决策树（工作区）**。

生成图形

“树”节点提供了大量信息；但是，对于业务用户，此信息可能并非始终采用可轻松访问的格式。 要通过可以轻松纳入业务报告、简报等的方式提供数据，您可以生成所选数据的图形。 例如，您可以从模型块的“模型”或“查看器”选项卡，或者从交互树的“查看器”选项卡为树的选定部分生成图形，从而仅为选定的树或分支节点中的观测值创建图形。

注：仅当块附加到流中的其他节点时，您才能根据该块生成图形。

生成图形

第一步是选择要在图形上显示的信息：

- 在块的“模型”选项卡上，展开左侧窗格中的条件和规则列表，然后选择感兴趣的项。
- 在块的“查看器”选项卡上，展开分支列表并选择感兴趣的节点。
- 在交互树的“查看器”选项卡上，展开分支列表并选择感兴趣的节点。

  注：您无法在上述“查看器”选项卡中选择顶级节点。

创建图形的方式相同，而与选择显示数据的方式无关：

1. 从“生成”菜单选择**图形（从选择）**；或者在“查看器”选项卡上单击左下角处的**图形（从选择）**按钮。 这将显示“图形板基本”选项卡。

   注：通过此方式显示“图形板”时，仅“基本”和“详细”选项卡可用。
2. 使用“基本”或“详细”选项卡设置来指定要在图形上显示的详细信息。
3. 单击“确定”以生成图形。

图形标题标识已选择要包含的节点或规则。

用于增强、组装和超大型数据集的模型块

如果在建模节点上选择**提高模型准确性（增强）**、**提高模型稳定性（组装）**或**为超大型数据集创建模型**作为主目标，那么 IBM SPSS Modeler 会构建多个模型的整体。 有关更多信息，请参阅主题 第 33 页的『整体模型』。

生成的模型块包括下列选项卡。 “模型”选项卡提供多个不同的模型视图。

表 *8:* 模型块中的可用选项卡

制表符	视图	描述	其他信息
模型	模型摘要	显示整体质量和（增强型模型和连续目标除外）差异性的摘要，以及有关预测变量在不同模型中的变化程度的度量。	有关更多信息，请参阅主题 第 33 页的『模型摘要』。
	预测变量重要性	显示了指示估计模型时每个预测变量（输入字段）的相对重要性的图表。	有关更多信息，请参阅主题 第 34 页的『预测变量重要性』。
	预测变量频率	显示了表示每个预测变量在模型集中的相对使用频率的图表。	有关更多信息，请参阅主题 第 34 页的『预测变量频率』。
	成分模型准确度	绘制关于整体中每个不同模型的预测准确性的图表。	
	成分模型详细信息	显示整体中每个不同模型的信息。	有关更多信息，请参阅主题 第 34 页的『组件模型详细信息』。
	信息	显示字段、构建设置和模型估算过程的相关信息。	有关更多信息，请参阅主题 第 32 页的『模型块概要/信息』。
设置		使您可以在评分操作中包括置信度。	有关更多信息，请参阅主题 第 88 页的『决策树/规则集模型块设置』。
注释		使您可以添加描述性注释、指定定制名称、添加工具提示文本以及指定模型的搜索关键字。	

C&R 树、CHAID、QUEST、C5.0 和 Apriori 规则集模型块

对于关联规则建模节点 (Apriori) 或某个树构建节点（C&R 树、CHAID、QUEST 或 C5.0）所发现的特定输出字段，规则集模型块表示用于预测此字段的规则。 对于关联规则，必须从未优化规则块中生成规则集。 对于树，可以从交互式树构建器、C5.0 模型构建节点或任何树模型块中生成规则集。 与未优化规则块不同，可将规则集块放置在流中以生成预测。

运行包含规则集块的流时，会将两个新字段（分别包含每条记录对数据的预测值和置信度）添加到流中。 新的字段名称将通过为模型名称添加前缀生成。 对于关联规则集，预测字段的前缀是 *$A-*，而置信度字段的前缀是 *$AC-*。 对于 C5.0 规则集，预测字段的前缀是 *$C-*，而置信度字段的前缀是 *$CC-*。 对于 C&R 树规则集，预测字段的前缀为 *$R-*，置信度字段的前缀为 *$RC-*。 在一个序列（可预测相同的输出字段）中具有多个规则集块的流中，新的字段名称将包括数字 前缀，以便将这些名称区别开来。 流中的第一个关联规则集块将使用常用名称，第二个节点将使用以 *$A1-* 和 *$AC1-* 开头的名称，第三个节点将使用以 *$A2-* 和 *$AC2-* 开头的名称，依此类推。

如何应用规则。 从关联规则中生成的规则集与其他模型块不同，因为对于任何特定记录，都可以生成多个预测并且这些预测可能并不一致。 可使用两种方法从规则集中生成预测。

注: 不论使用哪种方法，从决策树中生成的规则集都会返回相同的结果，因为从决策树派生的规则是互斥的。

- **投票**。 此方法尝试组合对记录应用的所有规则的预测。 对于每条记录，会检查所有的规则，并使用应用于该记录的每个规则生成一个预测和一个关联置信度。 为每个输出值计算置信度图表的总和，具有最大置信度总和的值将被选作最终预测。 最终预测的置信度是该值（由应用于该记录的规则数划分）的置信度总和。
- **首个命中项**。 此方法仅仅是按顺序检验规则，并且对记录应用的第一项规则即为用于生成预测的规则。

可在流选项中控制所使用的方法。

正在生成节点。 “生成”菜单使您可以根据规则集创建新节点。

- **“过滤”节点**创建新的“过滤”节点以过滤规则集中的规则不使用的字段。
- **“选择”节点**创建新的“选择”节点以选择对其应用选定规则的记录。 生成的节点将选择所应用规则的所有条件均为真的记录。 此选项需要选定一个规则。
- **规则跟踪节点**创建将计算字段（用于表示对每条记录进行预测时所使用的规则）的新超节点。 当使用第一个匹配方法评估规则集时，仅用一个表明将触发第一个规则的符号来表示。 当使用投票方法评估规则集时，则用一个显示投票机制的输入的复杂字符串来表示。
- **单一决策树**（画布）/**单一决策树**（GM 选用板）。 创建从当前选定的规则中派生的单个新规则集块。 仅适用于 **增强型** C5.0 模型。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。
- **从模型到选用板**将模型返回到模型选用板。 当有同事发给您包含模型的流而不是模型本身时，该功能很有用。

注: 规则集块中的“设置”和“摘要”选项卡与决策树模型完全相同。

规则集模型选项卡

规则集块的“模型”选项卡中显示由算法从数据中提取的规则列表。

规则按后项（预测类别）划分，并按下列格式显示:

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

其中 consequent 和 antecedent_1 直到 antecedent_n 是所有条件。 该规则可解释为“对于其中 antecedent_1 直到 antecedent_n 都为真的记录，consequent 也可能为真。” 如果单击工具栏上的 **显示实例/置信度** 按钮，则每个规则还将显示有关应用该规则的条件为真的记录的数目信息（ **实例** ），及整个规则为真的记录的比例信息（ **置信度** ）。

注意，对于 C5.0 规则集，置信度的计算方式有些不同。 C5.0 使用下列公式计算规则的置信度:

```
(1 + number of records where rule is correct)
/
(2 + number of records for which the rule's antecedents are true)
```

这一置信度估计计算方式可调整从决策树中生成规则（即 C5.0 创建规则集时所执行的操作）的过程。

第 7 章 贝叶斯网络模型

贝叶斯网络节点

通过 **贝叶斯网络** 节点，您可以利用对真实世界认知的判断力并结合所观察和记录的证据，通过使用看似不相关的属性建立事件发生的几率，从而构建概率模型。 该节点侧重于树扩展朴素贝叶斯 (TAN) 网络和马尔可夫覆盖网络，这些网络主要用于分类。

贝叶斯网络可用于在许多不同的情况下进行预测，示例如下：

- 选择违约风险较低的贷款时机。
- 根据传感器输入数据和现有记录，估算设备是否需要维修、增加零配件或更换。
- 借助在线故障排除工具解决客户问题。
- 实时诊断并排除移动电话网络故障。
- 评估研发项目的潜在风险和回报，以在最佳时机集中资源。

贝叶斯网络是一种图形模型，可显示数据集中的变量（通常称之为 **节点** ）以及概率，还可以显示这些变量之间的条件和独立性。 贝叶斯网络可呈现节点之间的因果关系；但是，网络中的链接（也称为 **arcs** ）没有必要呈现直接因果关系。 例如，如果图形中所显示的症状和疾病之间的概率独立性成立，贝叶斯网络可根据特定症状和其他相关数据是否存在，计算患者患有某种特殊疾病的几率。 这种网络非常稳健，即使在信息缺失时，也可以利用现有的任何信息作出最佳预测。

标准的基础贝叶斯网络示例由 Lauritzen 和 Spiegelhalter 于 1988 年创建。 该网络示例是一种简化的网络版本，通常称作“Asia”模型，医生可用它来诊断新患者的病情，所有链接的方向可大体指示因果关系。 每个节点代表与患者状况相关的一个方面，例如“吸烟”表示这些患者确为吸烟者，而“VisitAsia”表示他们最近是否去过亚洲。 概率关系由所有节点之间的链接指示，例如，吸烟会增大患者患有支气管炎和肺癌的几率，而年龄仅与肺癌的患病率相关。 同样地，肺部 x 光检查出的异常可能由肺结核或肺癌引起。同时，如果患者本身患有支气管炎或肺癌，那么他们更有可能出现呼吸急促（呼吸困难）的症状。

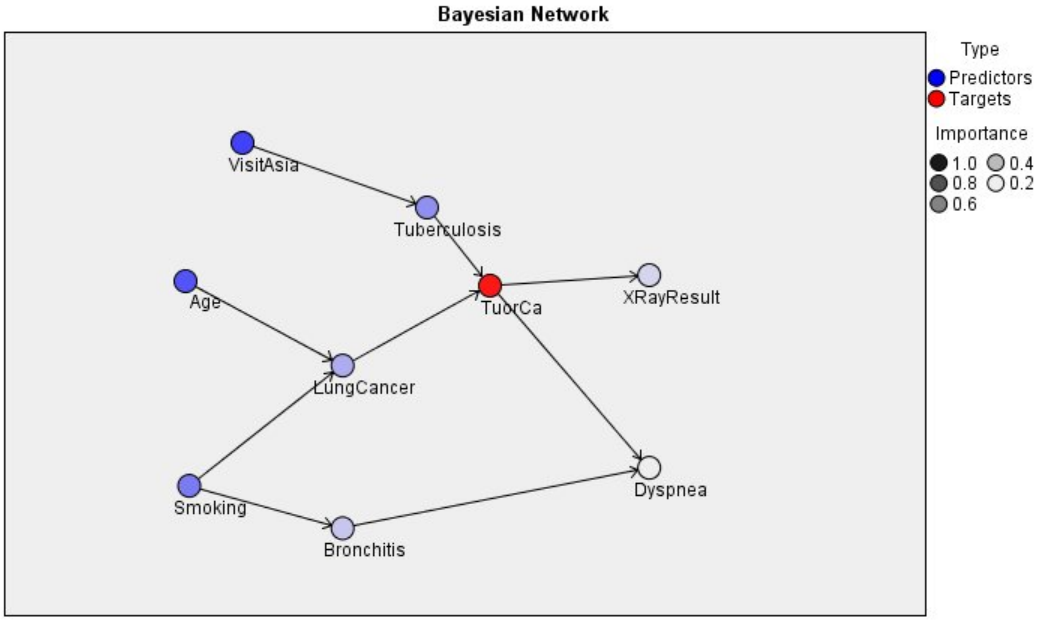


图 *29: Lauritzen* 和 *Spegelhalter* 的 *Asia* 网络示例

以下是您有可能决定使用贝叶斯网络的几点原因：

- 它可帮助您了解因果关系。 由此，您可以了解出现问题的地方并可预测任何干涉可能引发的后果。

- 该网络可提供避免数据过度拟合的有效方法。
- 可以轻松地观测到所涉及关系的清晰视图。

需求。 目标字段必须为分类且测量级别为名义、有序或标志。 输入内容可以为任何类型的字段。 连续（数值范围）输入字段将自动分级；但是，如果分布出现不对称，则可使用贝叶斯网络节点之前的分级节点对字段进行手动分级，从而获得更佳的效果。 例如，在**主管字段**与贝叶斯网络节点**目标**字段相同的位置处，使用最优分级。

示例。 银行分析师希望可以预测可能拖欠偿还贷款的客户或潜在客户。 您可使用贝叶斯网络模型来标识最有可能拖欠还款的客户的特征，并构建几种不同类型的模型，以确定哪种模型可以最好地预测潜在的贷款拖欠者。

示例。 一位电信运营商希望减少中断服务（又称为“流失”）的客户数量，并使用上一个月的数据对模型每月进行更新。 您可以使用贝叶斯网络模型确定最有可能流失的客户的特征，然后每月使用新数据继续训练该模型。

贝叶斯网络节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

为每个拆分构建模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅主题 第 21 页的『构建分割模型』。

分区。 通过此字段，您可以指定用于针对模型构建中的训练、检验和验证阶段将数据划分为不同样本的字段。 通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。 如果已使用类型或分区节点定义了多个分区字段，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。）同时请注意，要在分析时应用选定分区，还必须启用节点的“模型选项”选项卡中的分区功能。 （取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

拆分。 对于分割模型，选择分割字段或字段。 此操作与在“类型”节点中将字段的角色设置为分割类似。 您仅可将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为分割字段。 选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关更多信息，请参阅主题 第 21 页的『构建分割模型』。

继续训练现有模型。 如果选择此选项，则在模型块“模型”选项卡上显示的结果，将在每次运行模型时重新生成和更新。 例如，如果已为现有模型添加新的或更新的数据源，则需要执行此操作。

注：此操作只能更新现有网络；它无法添加或者移除节点或连接。 每次重新训练模型时，网络的形状都将保持不变，只会更改条件概率和预测变量重要性。 如果新数据与旧数据大致相似也无妨，因为您所期望的是关注相同的内容；但是，如果您希望检查或更新 重要的内容（针对其重要程度），则需要构建新模型，即构建新网络。

结构类型。 选择构建贝叶斯网络时要使用的结构：

- **TAN**。 树扩展朴素贝叶斯模型 (TAN) 用于创建简单的贝叶斯网络模型，后者是对标准朴素贝叶斯模型的改进。 这是由于该模型允许每一个预测变量除了依赖于目标变量之外，还依赖于其他预测变量，由此增加分类的准确度。
- **马尔可夫覆盖**。 此结构用于选择数据集中的节点的集合，这些节点包含目标变量的父项、其子项以及子项的父项。 马尔可夫覆盖基本可以确定网络中预测目标变量的所需的所有变量。 用户认为这种构建网络的方法更为准确；但是，当处理大型数据集时，由于所包含的变量数较多，所以可能会消耗许多处理时间。 要减少处理工作量，可以使用“专家”选项卡上的**特征选择**选项，选择与目标变量有重大相关性的变量。

包含特征选择预处理步骤。 选择该框，您可以使用“专家”选项卡上的**特征选择**选项。

参数学习方法。 贝叶斯网络参数是指给定每个节点的父项值时，该节点具有的条件概率。 有两种可能的选择，您可以用来控制估算节点（此处父项值已知）间条件概率表这一任务。

- **最大似然**。 使用大型数据集时，请选中此框。 这是缺省选项。
- **小单元格计数的贝叶斯调整**。 对于较小的数据集，存在模型过度拟合的风险以及出现大量零计数的可能性。 选中此选项可通过应用平滑来减少任何零计数以及不可靠的估计结果带来的影响，从而解决这些问题。

贝叶斯网络节点专家选项

使用节点专家选项可微调模型构建过程。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

缺失值。 缺省情况下，IBM SPSS Modeler 仅使用对模型中所用的全部字段具有有效值的记录。 （这种方式有时称为缺失值的 **成列删除** 。） 如果有很多缺失数据，您可能会发现这种方式去除的记录过多，剩余记录不足以生成较好的模型。 在这种情况下，可以取消选中**仅使用完整记录**选项。 IBM SPSS Modeler 随后将尝试使用尽可能多的信息来估算模型，包括其中一些字段具有缺失值的记录。 （这种方式有时称为缺失值的成对删除 。） 但在某些情形下，以这种方式使用不完整记录可能会在模型的估计过程中产生计算问题。

附加所有概率。 指定是否将输出字段每个类别的概率添加到该节点所处理的每条记录。 如果未选中此选项，那么仅添加预测类别的概率。

独立检验。 一种独立评估检验，用于评估两个变量中成对的观测值是否相互独立。 请从以下可用选项中选择要使用的检验类型：

- **似然比**。 通过计算两种不同假设下结果的最大概率之间的比率来检验目标-预测变量的独立性。
- **Pearson 卡方**。 通过使用零假设（所观察事件的相对出现频率遵循指定的频率分布）来检验目标-预测变量的独立性。

贝叶斯网络模型可在检验对之外使用附加变量执行独立性的条件检验。 此外，模型不仅可以研究目标和预测变量之间的关系，还可研究预测变量自身之间的关系。

注: 只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，独立性检验选项才可用。

显著性水平。 可以与独立性检验设置结合使用，通过此选项，您可以在执行检验时设置要使用的分界值。 该值越小，网络中的链接就越少；缺省水平值为 0.01。

注: 只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，此选项才可用。

最大条件集大小。 该算法用于创建马尔可夫覆盖结构，它使用大小不断增加的条件集来执行独立性检验并从网络中移除不需要的链接。 由于包含大量条件变量的检验需要更多的时间和内存进行处理，因此您可以限制要包括的变量数目。 在处理众多变量间具有较强独立性的数据时，这种操作非常有用。 但请注意，最终形成的网络可能包含一些多余链接。

指定执行独立性检验时要使用的条件变量的最大数目。 缺省设置为 5。

注: 只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，此选项才可用。

特征选择。 通过这些选项，您可以限制在处理模型时所使用的输入数，以便加快模型构建过程。 由于在创建马尔可夫覆盖结构时存在大量的潜在输入，因此该操作特别有用；通过此项操作，您可以选择与目标变量有重大关联的输入。

注: 只有在“模型”选项卡上选中**包括特征选择预处理步骤**时，特征选择选项才可用。

- **始终选择输入**。 通过使用“字段选择器”（文本字段右侧的按钮），从数据集中选择构建贝叶斯网络模型时始终使用的字段。 目标字段始终处于选中状态。 请注意，如果其他检验认为某些项不重要，那么在模型构建过程中，贝叶斯网络可能仍然会从此列表中删除这些项。 因此，该选项仅确保列表中的项用在模型构建过程中，而不确保它们绝对显示在生成的贝叶斯模型中。
- **最大输入数**。 指定构建贝叶斯网络模型时要使用的来自数据集中的总输入数。 您可以输入的最大数目为数据集中的总输入量。

  注: 如果在**始终选择输入**中选择的字段数超过**最大输入数**的值，那么将显示一条错误消息。

贝叶斯网络模型块

注: 如果在建模节点的“模型”选项卡中选中了**继续训练现有参数**，那么将在每次重新生成模型时更新模型块的“模型”选项卡上显示的信息。

模型块“模型”选项卡分为两个面板：

左侧窗格

基本：此视图包含节点网络图表，此图表显示目标与其最重要的预测变量之间的关系，以及各预测变量之间的关系。 各预测变量的重要性可通过其颜色的深浅显示；颜色越深表示变量越重要，反之亦然。

当您将鼠标指针悬停在节点上时，工具提示中会显示代表范围的节点的分级值。

可以使用 IBM SPSS Modeler 中的图表工具进行交互、编辑，并保存图表。 例如，可以在其他应用程序如 MS Word 中使用图表。

提示: 如果网络包含大量节点，那么可以单击以选中某个节点，然后拖动它以使图表更加清晰。

分布：此视图将以微型图表显示网络中每个节点的条件概率。 将鼠标悬停在图形上方，可在工具提示中显示图形值。

右窗格

预测变量重要性：这将显示一个图表，以指示在估计模型时所使用的各个预测变量的相对重要性。 有关更多信息，请参阅 第 32 页的『预测变量重要性』。

条件概率：当在左窗格中选择了某个节点或微型分布图时，右窗格则会显示相关的条件概率表。 该表包含各个节点值的条件概率值，以及各节点的父节点中的值组合。 此外，该表还包含为每个记录值和父节点中各个值组合所观测的记录数量。

贝叶斯网络模型设置

在贝叶斯网络模型块的“设置”选项卡中可指定选项以修改已构建的模型。 例如，可以通过贝叶斯网络节点使用相同的数据和设置构建几个不同的模型，然后使用每个模型中的此选项卡对设置稍做修改以查看其对结果的影响。

注: 只有将模型块添加到流中之后，此选项卡才可用。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

追加所有概率：指定是否将输出字段每个类别的概率添加到该节点所处理的每条记录。 如果未选中此选项，那么仅添加预测类别的概率。

此复选框的缺省设置由建模节点的“专家”选项卡上的相应复选框确定。 有关更多信息，请参阅主题 第 95 页的『贝叶斯网络节点专家选项』。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

贝叶斯网络模型摘要

模型块的“摘要”选项卡显示了有关模型的下列信息：模型本身（分析）、模型中使用的字段（字段）、构建模型时使用的设置（构建设置）和模型训练（训练概要）。

当第一次浏览此节点时，“摘要”选项卡的结果是折叠起来的。 要查看感兴趣的结果，可使用项目左侧的展开控件展开项目，或单击 **全部展开** 按钮显示所有结果。 当结束对项目的查看时，为了隐藏结果，可使用展开控件折叠要隐藏的特定结果，或单击 **全部折叠** 按钮折叠所有结果。

分析。 显示指定模型的相关信息。

字段。 列出构建模型时用作目标和输入的字段。

构建设置。 包含构建模型时使用的设置的相关信息。

训练摘要。 显示模型类型、用于创建模型的流、模型创建者、模型构建时间和构建模型所耗用的时间。

第 8 章 神经网络

神经网络可以近似多种预测模型，而对模型结构和假设只有最小需求。 关系的形式是在学习过程中确定的。 如果目标与预测变量之间的线性关系合适，那么神经网络的结果应该与传统线性模型的结果非常相似。 如果非线性关系更适合，神经网络将自动接近“正确”模型结构。

此灵活性的缺点是，不容易对神经网络进行解释。 如果要尝试解释在目标与预测变量之间生成关系的底层过程，那么最好使用更为传统的统计模型。 但是，如果模型可解释性并不重要，那么使用神经网络可以获得良好的预测。

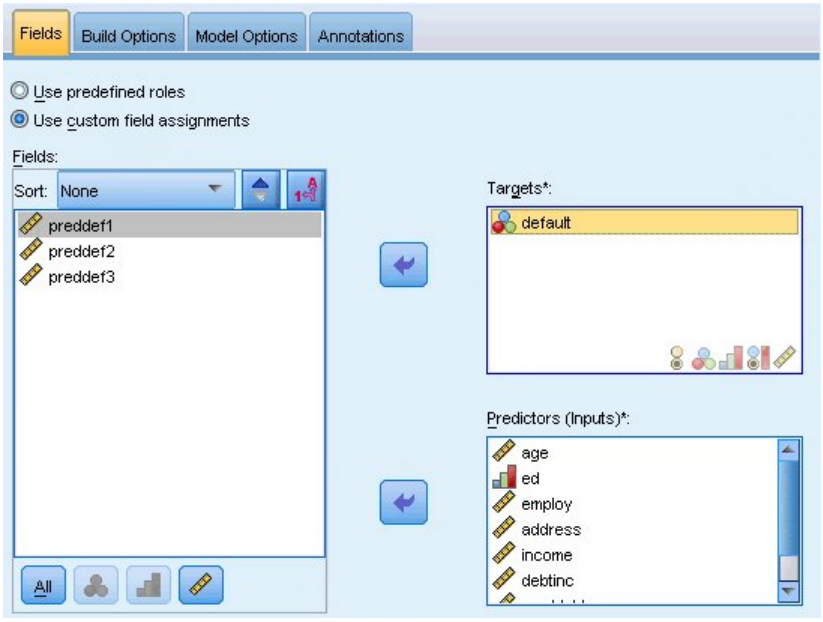


图 *30:* “字段”选项卡

字段要求。 必须至少有一个目标字段和一个输入字段。 将忽略设置为“两者”或“无”的字段。 对于目标或预测变量（输入），没有测量级别限制。 有关更多信息，请参阅第 23 页的『建模节点字段选项』。

在模型构建期间分配给神经网络（从而分配给所生成的最终模型）的初始权重取决于数据中字段的顺序。 SPSS Modeler 先按字段名称对数据进行自动排序，再将该数据提供给神经网络进行训练。 这意味着，在模型构建器中设置随机种子后，显式更改上游数据中的字段顺序不会影响所生成的神经网络模型。 但是，如果更改输入字段名称造成字段排序顺序出现变化，那么将会生成不同的神经网络模型，即使在模型构建器中设置了随机种子也是如此。 考虑到字段名称有不同排序顺序，模型质量将不会受到明显的影响。

神经网络模型

神经网络是神经系统运转方式的简单模型。 其基本单元是 **神经元**，通常将其组织到 **层** 中，如下面的图所示。

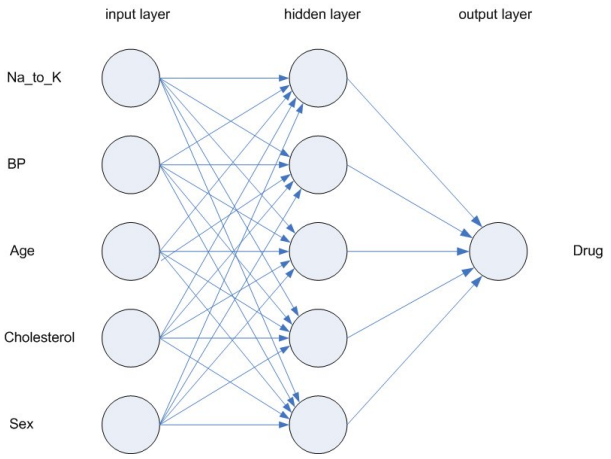


图 *31:* 神经网络的结构

神经网络是模拟人类大脑处理信息方式的简化模型。 此模型的工作方式为模拟大量类似于神经元的抽象形式的互连处理单元。

这些处理单元都位于层中。 神经网络通常包含三个部分：**输入层**，其中的单元表示输入字段；一个或多个**隐藏层**；一个**输出层**，带有一个或多个表示目标字段的单元。 这些单元通过可变的连接强度（或 **权重** ）连接。 输入数据 显示在第一层，其值从每个神经元传播到下一层的每个神经元。 最终从输出层中输出结果。

该网络可通过以下过程进行学习，即检查单个记录，然后为每条记录生成预测，并且当生成的预测不正确时，对权重进行调整。 在满足一个或多个停止标准之前，此过程会不断重复，而网络会持续提高其预测准确度。

最初，所有的权重都是随机生成的，并且从网络输出的结果很可能没有意义的。 网络可通过 **训练** 来学习。 向该网络重复应用已知道结果的示例，并将网络给出的结果与已知的结果进行比较。 从此比较中得出的信息会传递回网络，并逐渐改变权重。 随着训练的进行，该网络对已知结果的复制会变得越来越准确。 一旦训练完毕，就可以将网络应用到未知结果的未来案例中。

将神经网络与遗存流配合使用

IBM SPSS Modeler V14 引入了新的“神经网络”节点，支持增强和组装技术，并可针对超大型数据集进行优化。 在以后的发行版中，包含旧节点的现有流仍可构建模型以及进行模型评分。 但是，未来的发行版将移除这项支持，我们建议您使用新版本。

从 V13 以后，带有未知值的字段（即，值在培训数据中不存在）不再自动按照缺失值进行处理，而是使用 $null$ 值进行评分。 因此，如果要在 V13 或更高版本中使用 V13 以前的旧神经网络模型将具有未知值的字段评分为非空，那么应该将未知值标记为缺失值（例如，通过使用“类型”节点完成此任务）。

请注意，为了实现兼容，仍包含旧节点的任何原有流可能仍在通过**工具 > 流属性 > 选项**使用限制集大小选项；从 V14 起，此选项仅适用于 Kohonen 网络和 *K*-Means。

目标

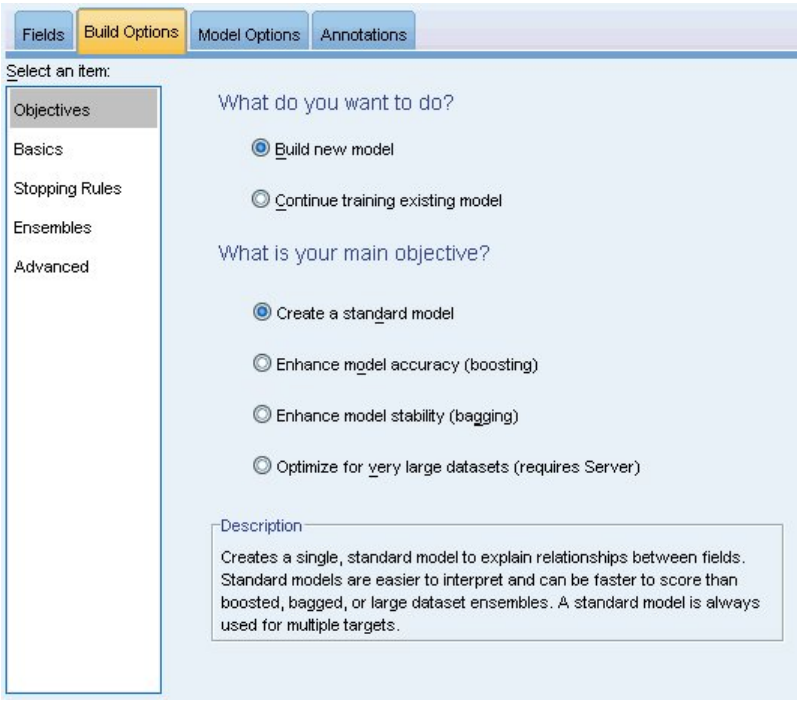


图 *32:* 目标设置

您要执行什么操作?

- **构建新模型**。 构建全新的模型。 这是节点的常见操作。
- **继续训练现有的模型**。 继续训练此节点成功生成的最后一个模型。 这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，因为只有新的或更新后的记录被传入流中。 上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型调色板中不再可用的情况下，也可以使用该选项。

  注: 启用此选项后，“字段”和“构建选项”选项卡上的所有其他控件将处于禁用状态。

您的主要目标是什么? 请选择适当的目标。

- **创建标准模型**。 使用一种方法来构建一个可以使用预测变量预测目标的模型。 一般来说，标准模型更易于理解，而且评分速度比 boosted、bagged 或大型数据集整体更快。

  注: 只有**构建单一树**拆分模型中才支持**继续训练现有模型**，并且必须连接到 Analytic Server。
- **增强模型准确度（提升）**。 使用 Boosting 构建整体模型的方法，可生成一个模型序列来获得更多精确预测值。 与标准模型相比，整体模型需要更长的时间来构建和评分。

  Boosting 将生成一连串的“组件模型”，其中的每个模型都基于整个数据集进行构建。 在构建每个连续的组件模型之前，将根据上一个组件模型的残值确定记录的权重。 残值较大的个案将被赋予相对较高的分析权重，以使下一个组件模型还侧重于预测这些记录。 这些成分模型共同构成一个整体模型。 这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。
- **增强模型稳定性（组装）**。 使用 Bagging (bootstrap 汇总) 构建整体模型的方法，可生成多个模型来获得更多可靠的预测值。 与标准模型相比，整体模型需要更长的时间来构建和评分。

  Bootstrap 汇总 (Bagging) 通过对原始数据集进行放回方式的取样，生成训练数据集的副本。 这将创建大小与原始数据集相同的 Bootstrap 样本。 然后，在每个副本上构建“成分模型”。 这些成分模型共同构成一个整体模型。 这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。
- **为非常大的数据集创建模型**。 通过将数据集拆分成单独的数据块来构建整体模型的方法。 如果您的数据集非常大，无法构建任何上述模型，或希望用于增量建模，请选择该选项。 该选项可使用更短的时间来构建模型，但需要比标准模型更长的时间来评分。

存在多个目标时，此方法将仅创建标准模型，而不考虑选定的目标。

基本

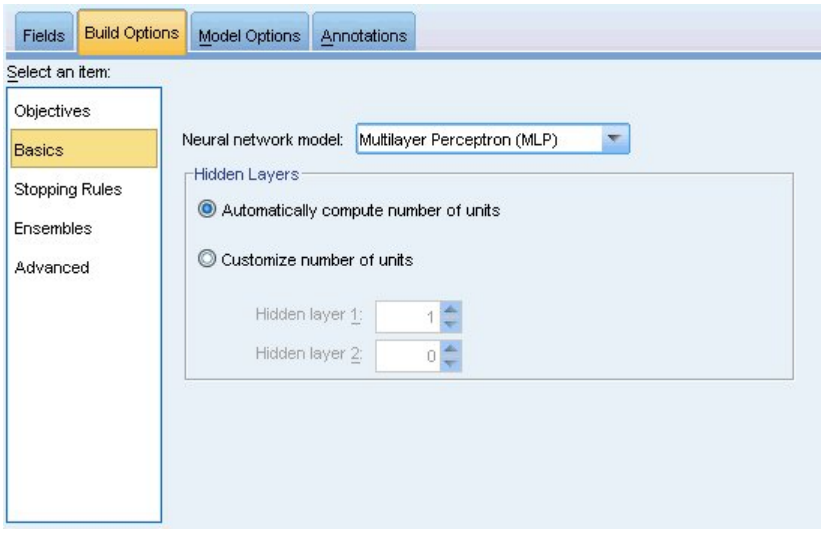


图 *33:* 基本设置

神经网络模型。 此类模型用于确定神经网络如何通过隐藏层将预测变量连接到目标。 **多层感知器(MLP)**允许构建较为复杂的关系，但代价是更长的训练与评分时间。 **径向基函数(RBF)**可以缩短训练与评分时间，但与 MLP 相比其预测能力要差些。

隐藏层。 神经网络的隐藏层包含无法观察到的单元。 每个隐藏单元的值都是预测变量的某个函数；此函数的准确形式部分取决于网络类型。 多层感知器可能有一个或两个隐藏层；径向基函数网络可以有一个隐藏层。

- **自动计算单元数**。 此选项构建具有单个隐藏层的网络，并计算隐藏层中的“最佳”单元数。
- **定制单元数**。 此选项允许您指定每个隐藏层中的单元数。 第一个隐藏层必须至少具有一个单元。 对第二个隐藏层指定 0 个单元将构建具有单个隐藏层的多层感知器。

  注: 选择值时，您应确保节点数不超过连续预测变量数加上所有分类（标志、名义和有序）预测变量间类别总数之和。

中止规则

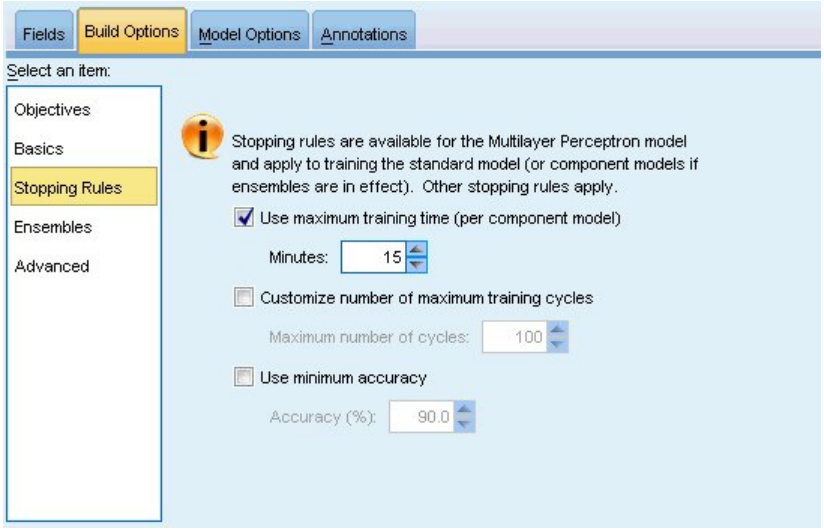


图 *34:* “中止规则”设置

有一些用于确定何时中止训练多层感知器网络的规则；使用径向基函数算法时，将忽略这些设置。 训练将至少进行一个周期（数据遍历），然后可以根据以下条件中止训练。

使用最长训练时间（每个组件模型）。 选择是否指定运行算法的最大分钟数。 指定大于 0 的数。 构建整体模型时，这是该整体的每个组件模型所允许的训练时间。 请注意，为了完成当前周期，训练可能会比指定的时间限制延长一点。

定制最大训练周期数。 允许的最大训练周期数。 如果超过最大周期数，那么训练将中止。 指定大于 0 的整数。

使用最低准确性。 如果使用此选项，那么在达到指定的准确性前，训练将持续进行。 这种情况可能永远不会出现，但您可以随时中断训练，以截止到目前所达到的最佳精确性保存该网络。

如果防止过度拟合集合中的误差并未在每个周期后减小，训练误差中的相对变化较小，或者当前训练误差与初始误差相比较小，那么训练算法也将中止。

整体

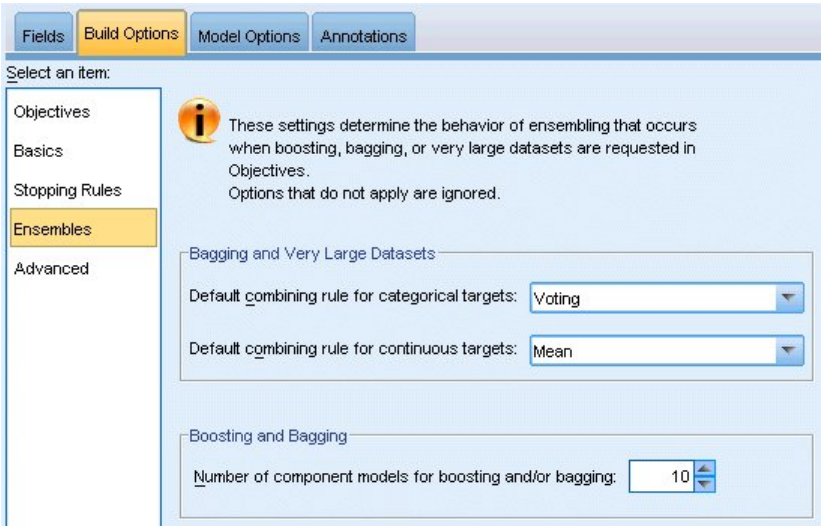


图 *35:* 整体设置

这些设置决定了在“目标”中请求 Boosting、Bagging 或超大型数据集时发生的整体行为。 将忽略不适用于选定目标的选项。

Bagging 和大型数据集 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体评分值。

- **分类目标的缺省合并规则**。 可以通过投票、最高概率或最高均值概率来对分类目标的整体预测值进行组合。 **投票**选择在基本模型中最常具有最高概率的类别。 **最高概率**选择在所有基本模型中达到单一最高概率的类别。 **最高平均概率**选择当类别概率在基本模型中取平均值时具有最高值的类别。
- **连续目标的缺省合并规则**。 可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，那么组合规则选择将被忽略。 Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

提升和组装。 当以增强模型精确性或稳定性为目标时，指定要构建的基本模型数；对于 Bagging 方法，此为 bootstrap 样本数。 它应为正整数。

高级

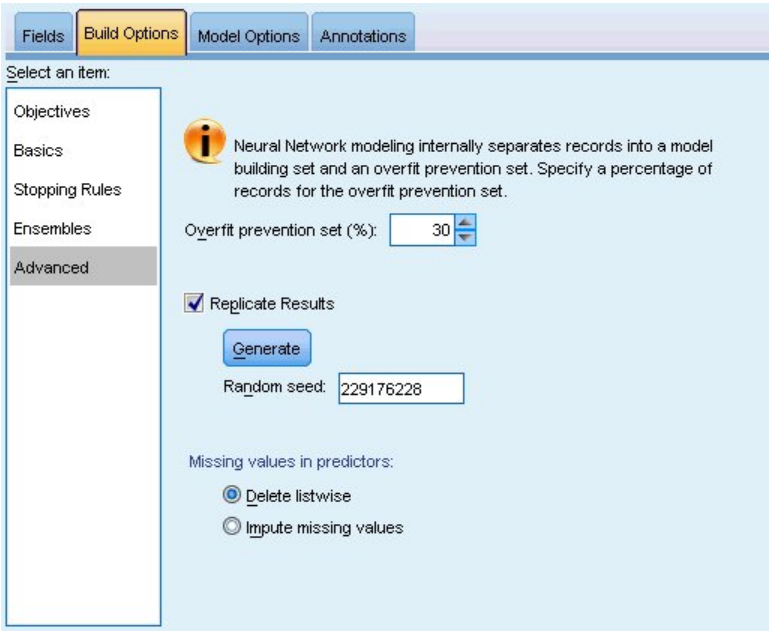


图 *36:* 高级设置

高级设置提供对无法很好地归入其他设置组的选项的控制。

防止过度拟合集合。 神经网络方法在内部将记录划分为模型构建集合和防止过度拟合集合，后者作为独立的数据记录集，用于跟踪训练过程中的错误，以防止该方法对数据中的几率变异进行建模。 指定记录的百分比。 缺省值为 30。

复制结果。 设置随机种子允许您复制分析。 指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。 缺省情况下，这些分析以种子值 229176228 进行复制。

预测变量中的缺失值。 这将指定如何处理缺失值。 **成列删除**将在预测变量上存在缺失值的记录从模型构建中排除。 **插补缺失值**将替换预测变量中的缺失值，并在分析中使用这些记录。 连续字段插补最小和最大观测值的平均值；分类字段插补最经常出现的类别。 请注意，将始终从模型构建中移除“字段”选项卡上指定的任何其他字段中包含缺失值的记录。

模型选项

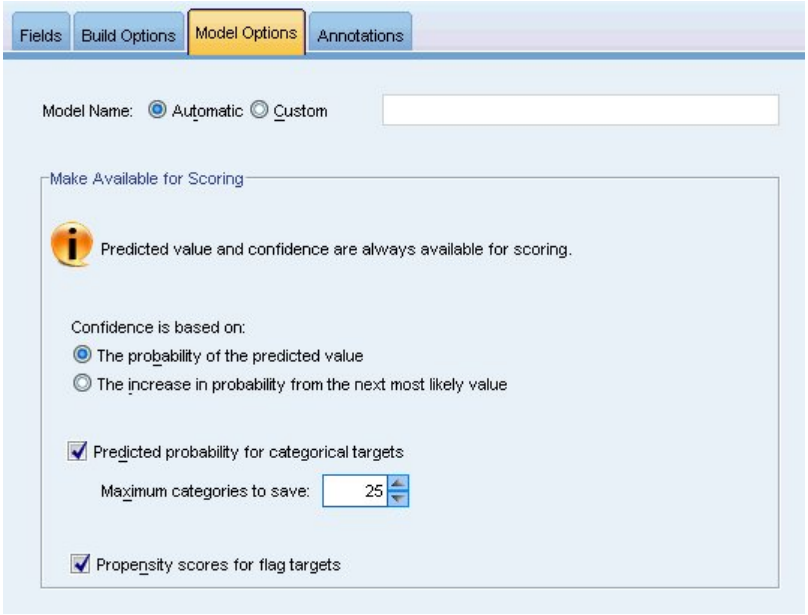


图 *37: “模型选项”选项卡*

模型名称。 可以基于目标字段来自动生成模型名称，或指定定制名称。 自动生成的名称为目标字段名。 如果存在多个目标，那么模型名称将由这些字段名按顺序排列组成，且字段名之间通过“与” (&) 符号连接。 例如，如果 *field1 field2 field3* 是目标，那么模型名称是：*field1 & field2 & field3*。

可用于评分。 对模型进行评分时，应生成此组中的选定项目。 在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。 计算的置信度可基于预测值的概率（最高的预测概率）或最高预测概率和次高预测概率之间的差异。

- **分类目标的预测概率**。 这将生成分类目标的预测概率。 为每个类别创建一个字段。
- **标志目标的倾向评分**。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 该模型产生原始倾向评分；如果分区处于有效，那么模型还会根据测试分区产生调整后的倾向评分。

模型摘要

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

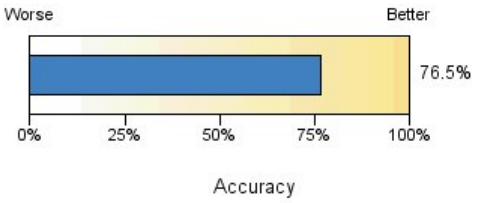


图 *38:* 神经网络模型摘要视图

“模型摘要”视图是一个快照，即神经网络预测或分类准确性的概览摘要。

模型摘要。 此表标识目标、已训练的神经网络类型、中止训练的中止规则（已训练多层感知器网络时显示），以及网络的每个隐藏层中的神经元数。

神经网络质量。 此图表显示最终模型的准确性，数值越大越好。 对于分类目标，此指数只是预测值与实测值匹配的记录所占的百分比。 对于连续目标，准确性指定为 R^2 值。

多个目标。 如果有多个目标，那么每个目标都将显示在表的**目标**行中。 图表中显示的准确性是各个目标准确性的平均值。

预测变量重要性

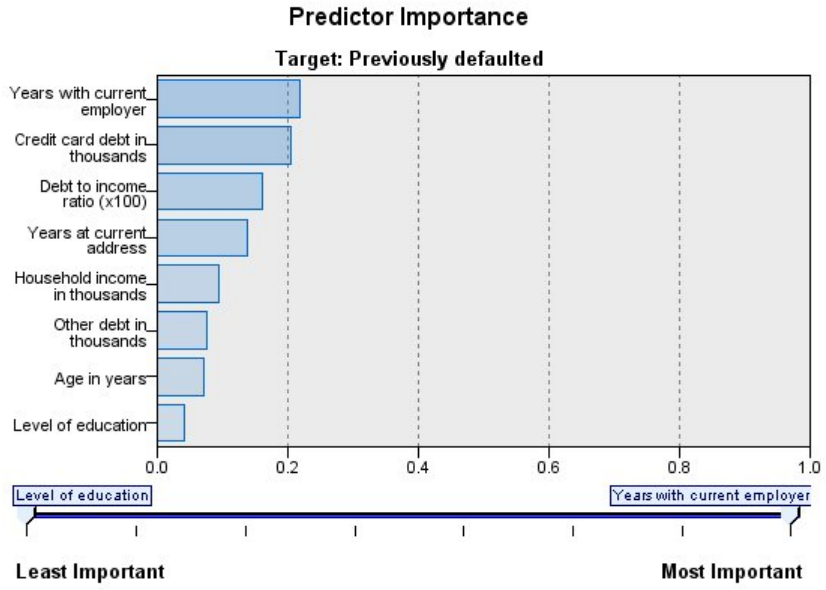


图 *39:* “预测变量重要性”视图

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。预测变量重要性与模型精度无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

多个目标。如果存在多个目标，那么每个目标都会显示在单独的图表中，并且会提供**目标**下拉列表，用于控制要显示的目标。

按已观测进行预测

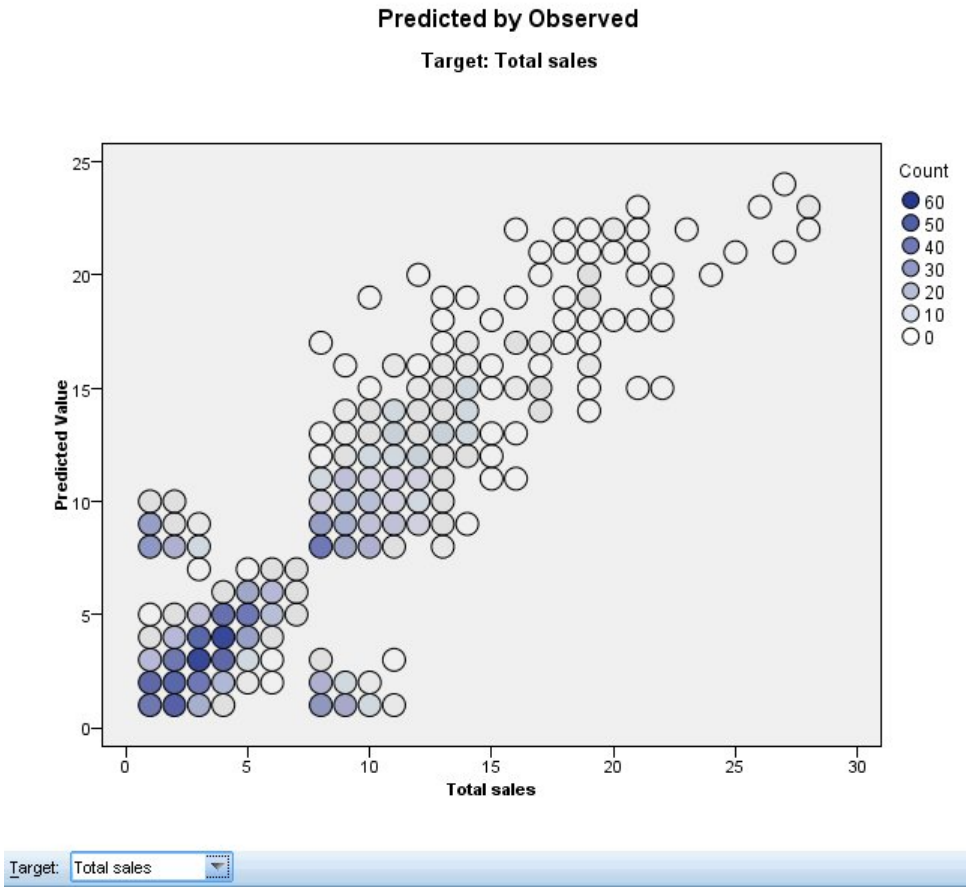


图 *40: "按已观测进行预测"视图*

对于连续目标，这将显示预测值位于垂直轴上，而观测值位于水平轴上的离散化散点图。

多个目标。如果存在多个连续目标，那么每个目标都会显示在单独的图表中，并且会提供**目标**下拉列表，用于控制要显示的目标。

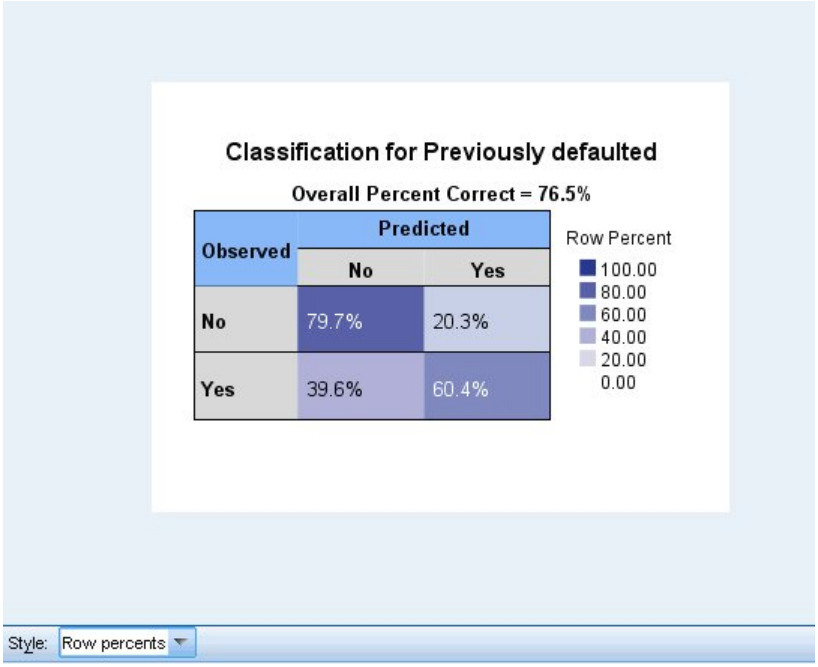


图 *41:* 分类视图，行百分比样式

对于分类目标，将以热图显示已观测和已预测值的交叉分类，以及整体正确百分比。

表样式。 有几个不同的显示样式，可以从**样式**下来列表中访问这些样式。

- **行百分比**。 这将在单元格中显示行百分比（单元格计数以行总计百分比表示）。 这是缺省选项。
- **单元格计数**。 这将显示单元格中的单元格计数。 热图的阴影还是基于行百分比。
- **热图**。 这将在单元格中不显示任何值，只显示阴影。
- **压缩**。 这将在单元格中不显示任何行或列标题，或值。 在目标具有许多分类时，此样式将十分有用。

缺失。 如果目标上的任何记录缺失值，那么会显示在所有有效行下的 **（缺失）**行中。 具有缺失值的记录不会对整体正确百分比作出贡献。

多个目标。 如果有多个分类目标，那么每个目标将显示在单个表中，同时有一个**目标**下拉列表控制要显示的目标。

大型表。 如果所显示的目标有 100 多个类别，将不显示任何表。

网络

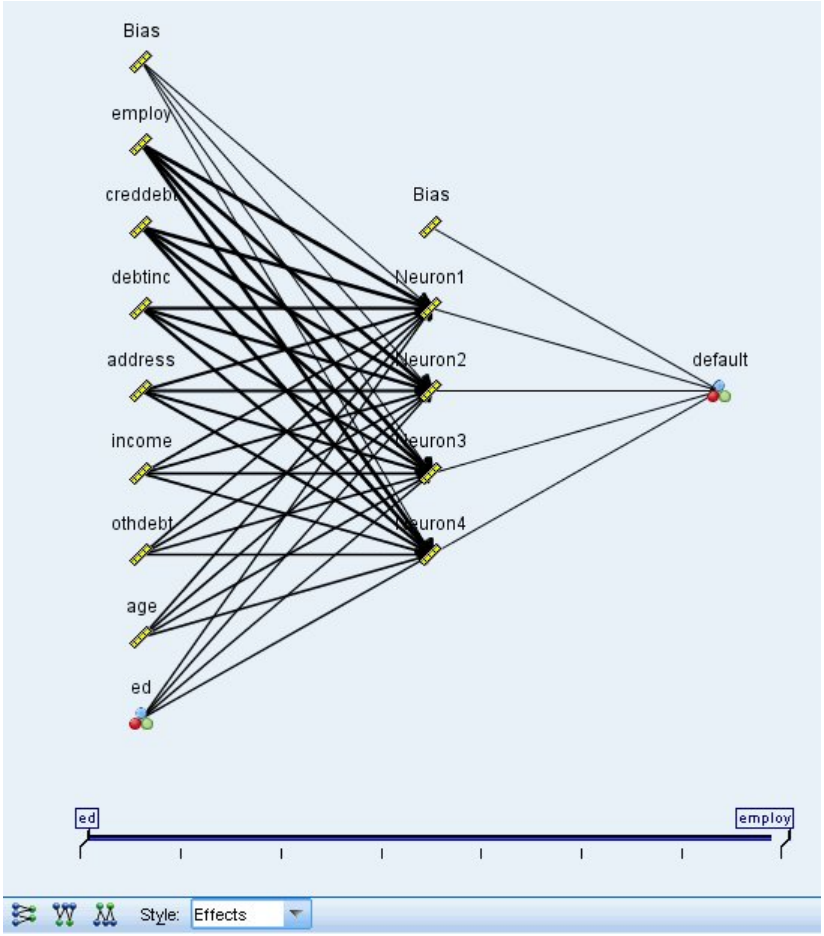


图 *42:* 网络视图，左侧的输入，效应样式

这将显示神经网络的图形表示。

图表样式。 有两种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **效应**。 这会在图表中将每个预测变量与目标显示为单个节点，而不考虑测量尺度是连续还是分类。 这是缺省选项。
- **系数**。 这将为分类预测变量与目标显示多个指示节点。 系数样式图中的连接线根据突触权重的估算值进行着色。

图表方向。 缺省情况下，输入位于网络图的左侧，而目标位于右侧。 通过使用工具栏控件，您可以更改方向，以使输入位于顶部而目标位于底部，或者输入位于底部而目标位于顶部。

预测变量重要性。 在图中，连接线条根据预测变量的重要性进行加权，粗线条表示重要性较高。 工具栏中有一个“预测变量重要性”滑块，用于控制网络图中显示的预测变量。 这不会改变模型，只是帮助您重点关注最重要的预测变量。

多个目标。 如果存在多个目标，那么所有目标都将显示在图表中。

设置

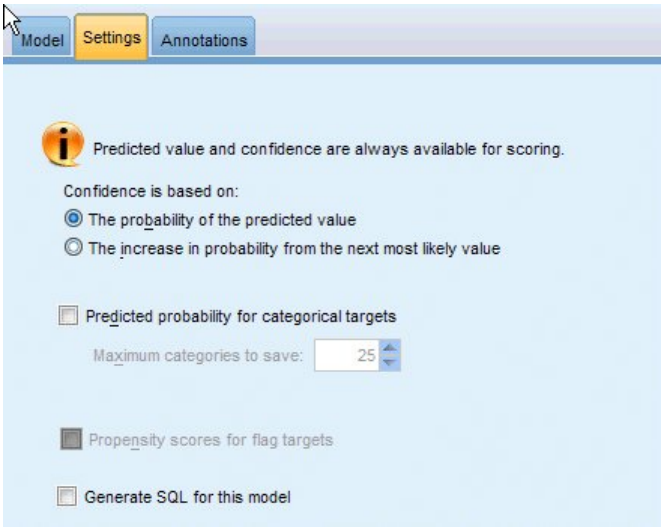


图 *43: “设置”选项卡*

在对模型评分时，应生成此选项卡中的选定项目。 在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。 计算的置信度可基于预测值的概率（最高的预测概率）或最高预测概率和次高预测概率之间的差异。

- **分类目标的预测概率**。 这将生成分类目标的预测概率。 为每个类别创建一个字段。
- **标志目标的倾向评分**。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 该模型产生原始倾向评分；如果分区处于有效，那么模型还会根据测试分区产生调整后的倾向评分。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

通过转换至本机 SQL 来进行评分：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分。

注: 虽然该选项可以更快速获得结果，但是本机 SQL 的大小和复杂性会随着模型复杂性的增加而增加。

在数据库外进行评分此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

第 9 章 决策列表

Decision List 模型标识了子组或**段**，即，显示了与整体样本相关的二值（yes 或 no）结果的似然度的高低。例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。 通过 Decision List Viewer 可以实现对模型的完全控制，它允许您编辑段、添加自己的业务规则、指定每个段的评分方式，以及采用其他多种方式定制模型从而对所有段的匹配比例进行优化。 因此，它尤其适用于生成邮件列表，或确定作为特定活动目标的记录。 此外，还可以使用多个 **挖掘任务** 对不同建模方法进行组合，例如，确定同一模型中性能较高和较低的段，并根据需要在评分阶段包含或排除每个段。

段、规则和条件

模型由段列表组成，每个段由选择匹配记录的规则进行定义。 给定的规则可以有多个条件，例如：

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

规则的列表顺序即为应用顺序，第一个匹配规则将决定给定记录的输出结果。 如果单独采用，那么规则或条件可能会发生重叠，但规则的顺序排除了二义性。 如果规则不匹配，那么记录将会分配给其余规则。

完全控制评分

通过 Decision List Viewer，您可以查看、修改和重组段，并且可以出于评分目的来选择包括或排除哪些段。例如，您可以选择在未来报价中排除某组客户和包含其他客户，并且可以立即查看这对于整体匹配率的影响。 对于包括的段和所有其他段（包括剩余段），Decision List 模型分别返回评分 *Yes* 和 *$null$*。 对评分的这种直接控制使得 Decision List 模型成为生成邮件发送清单的理想工具，而这些模型被广泛应用于客户关系管理中，包括呼叫中心或市场应用方面。

挖掘任务、测量和选择

建模过程由 **挖掘任务** 实现。 每项挖掘任务可以有效地启动一次新的建模，并且会返回一组新的备选模型。缺省任务基于 Decision List 节点的初始规范，您可以定义任意数量的定制任务。 您还可以重复应用任务，例如您可以在整个训练集合中运行高概率搜索，然后在剩余集合中运行低概率搜索以移除性能较低的段。

数据选择

可以定义数据选择和定制模型测量以进行模型构建和评估。 例如，可以在挖掘任务中指定数据选择以裁剪模型，使之符合具体区域的要求，并且可以创建定制测量以评估其就整个国家范围而言的性能优劣。 不同于挖掘任务的是，测量并不改变底层模型而是以其他视角对其性能进行评估。

添加您的业务知识

通过微调或扩展由算法识别的段，Decision List Viewer 使您可以将业务知识并入模型。 您可以编辑模型所生成的段或添加基于指定规则的其他段。 然后可以应用更改并预览结果。

为了进行深入了解，Excel 动态链接使您可以将数据导出到 Excel，这些数据可用于在 Excel 中创建演示图表和计算定制测量（例如综合利润和 ROI），您可在构建模型的同时在 Decision List Viewer 中查看这些定制测量。

示例。 金融机构的市场营销部门希望通过向每个客户提供合适的报价以在未来的营销活动中获取更多的利润。 您可以使用决策列表模型来根据以前的促销活动确定最有可能做出积极响应的客户所具备的特征，并根据结果生成邮件发送列表。

需求。 一个测量级别类型为标志或名义的分类目标字段（指示想要预测的二元结果（是/否））和至少一个输入字段。 当目标字段类型为名义时，必须手动选择一个值作为**匹配**或**响应**；所有其他值集中在一起作为**不匹配**。 还可以指定一个可选的频率字段。 连续日期/时间字段将被忽略。 使用在建模节点的“专家”选项卡上指定的算法对连续数字范围的输入自动分级。 为了更好地控制分级，可添加上游分级节点并使用已分级的字段作为测量级别为有序的输入。

决策列表模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

方式。 指定用于构建模型的方法。

- **生成模型**。 执行节点时自动在模型选用板上生成模型。 可将生成的模型添加到流中以便评分，但是此模型无法继续编辑。
- **启动交互式会话**。 打开 Decision List Viewer 交互式建模（输出）窗口，在此窗口中您可以从多个替代项中进行选择并重复应用具有不同设置的算法以逐步生成或修改模型。 有关更多信息，请参阅主题第 113 页的『Decision List Viewer』。
- **使用保存的交互式会话信息**。 使用先前保存的设置来启动交互式会话。 可以使用 Decision List Viewer 中的“生成”菜单（用于创建模型或建模节点）或“文件”菜单（用于更新从中启动会话的节点）保存交互设置。

目标值。 指定目标字段的值，该字段表示要对其进行建模的结果。 例如，如果目标字段 churn 编码为 `0` = no 和 `1` = `yes`，则指定 `1` 以确定用于指示哪些记录可能流失的规则。

查找段的方式。 表示搜索目标变量是否应该查找出现的**高概率**或**低概率**。 查找和排除这些段可能对于改善您的模型非常有帮助，当剩下的段为低概率段时尤其有用。

最大段数。 指定要返回的最大段数。 创建顶部的 *N* 个段，其中最好的段是概率最高的段，如果多个模型具有相同的概率，那么为覆盖率最高的段。 允许的最小设置为 1；没有最大设置。

最小段大小。 下面的两项设置指定了最小段大小。 两个值中的较大者优先。 例如，如果百分比值等于比绝对值高的数字，那么百分比设置优先。

- **占先前段的百分比 (%)**。 以记录的百分比形式指定最小组大小。 允许的最小设置为 0；允许的最大设置为 99.9。
- **作为绝对值 (N)**。 以记录的绝对数量指定最小组大小。 允许的最小设置为 1；没有最大设置。

段规则。

最大属性数。 指定每个段规则的最大条件数。 允许的最小设置为 1；没有最大设置。

- **允许属性复用**。 如果启用，那么每个周期可以使用所有属性，即使以前的周期已使用过这些属性。 段的条件是在周期内构建的，每个周期都会增加一个新条件。 周期数使用 **最大属性数** 设置定义。

新条件的置信区间 (%)。 指定用于检验段显著性的置信水平。 此设置在返回的段数（如果存在）以及每个段规则的条件数中具有非常重要的作用。 值越高，返回的结果集越小。 允许的最小设置为 50；允许的最大设置为 99.9。

决策列表节点专家选项

通过“专家”选项，您可以对模型构建过程进行微调。

分箱方法。 这是用于对连续字段（计数或宽度相等）进行分级的方法。

分箱数。 要为连续字段创建的分级的数目。 允许的最小设置为 2；没有最大设置。

模型搜索宽度。 这是每个周期中可用于下一周期的模型结果的最大数。 允许的最小设置为 1；没有最大设置。

规则搜索宽度。 这是每个周期中可用于下一周期的规则结果的最大数。 允许的最小设置为 1；没有最大设置。

分箱合并因子。 段与其相邻段合并时必须增加的最小量。 允许的最小设置为 1.01；没有最大设置。

- **允许条件中出现缺失值**。 `True` 表示允许规则中的 `IS MISSING` 检验。

- **废弃中间结果**。 如果为 True，那么将仅返回搜索过程的最终结果。 最终结果是不在搜索过程中进行任何进一步细化的结果。 如果为 False，那么还会返回中间结果。

  最大替代项数。 指定运行挖掘任务时可以返回的最大替代项数。 允许的最小设置为 1；没有最大设置。

  注意，挖掘任务将只返回替代值的实际数量，最大为指定的最大数量。 例如，如果最大数量设为 100，但只找到 3 个替代值，那么只显示这 3 个替代值。

决策列表模型块

模型包括一个 **段** 列表，每个段都由 **规则** 进行定义，从而可以选择匹配的记录。 在生成模型前可轻松查看或修改这些段，并选择包括哪些段或不包括哪些段。 用于评分时，决策列表模型对于包含的段返回是，对于所有其他段（包括余数）返回 *$null$*。 对评分的这种直接控制使得决策列表模型成为生成邮件发送清单的理想工具，而这些模型被广泛应用于客户关系管理中，包括呼叫中心或市场应用方面。

运行包含决策列表模型的流时，节点将添加三个新字段，包括评分字段，其中对于包含的字段评分为 *1*（表示是），对于不包含的字段评分为 *$null*，用于其中含有记录的段的概率（匹配率）字段，及段的标识编号字段。 新字段的名称派生自要预测的输出字段的名称，并带有表示评分的前缀 *$D-*、表示概率的前缀 *$DP-* 或表示段标识的前缀 *$DI-*。

按照构建模型时指定的目标值对模型进行评分。 您可以手动排除段，以便将其评分为 *$null$*。例如，如果您运行低概率搜索以查找命中率低于平均水平的段，那么除非您手动排除这些段，否则这些“低”段将评分为 *Yes*。 如果必要，可以使用导出节点或过滤节点将空值重新编码为 否。

PMML

使用“第一个匹配”选择标准可将决策列表模型存储为 PMML RuleSetModel。 但是，希望所有的规则具有相同的评分。 为允许对目标字段或目标值进行更改，可将多个规则集模型存储到一个文件中按顺序进行应用，无法与第一个模型匹配的案例将传递到第二个模型，依此类推。 算法名称 *DecisionList* 用于表示此非标准的行为，且仅具有该名称的规则集模型可被识别为决策列表模型并如上所述进行评分。

决策列表模型块设置

通过决策列表模型块的“设置”选项卡，您可以获取倾向评分，还可以启用或禁用 SQL 优化。 只有将模型块添加到流之后，才可以使用此选项卡。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **通过转换至本机 SQL 来进行评分**：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分。

  注: 虽然该选项可以更快速获得结果，但是本机 SQL 的大小和复杂性会随着模型复杂性的增加而增加。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

Decision List Viewer

基于任务的 Decision List Viewer 图形界面简单易用，它消除了模型构建过程的复杂性，使您可以摆脱数据挖掘技术的低层详细信息而将全部精力投入需要用户参与的分析内容上，如设置目标、选择目标组、分析结果，以及选择最优模型。

工作模型窗格

工作模型窗格将显示当前模型，包括挖掘任务和适用于该工作模型的其他操作。

标识。 标识连续段顺序。 模型段根据其标识号按顺序进行计算。

段规则。 提供段名称和已定义的段条件。 缺省情况下，段名称是字段名或条件中使用的连接字段名（以逗号为分隔符）。

分数。 表示要预测的字段，假定其值与其他字段（预测变量）的值相关。

注：以下选项可切换为通过第 121 页的『组织模型测量』对话框显示。

覆盖。 该饼图直观地标识出每个段的覆盖范围与整个覆盖范围的对比情况。

覆盖范围 (n)。 列出每个段相对于整个覆盖范围的覆盖范围量。

频率。 列出接收到的相对于覆盖范围的匹配项数。 例如，如果涉及范围为 79，频率为 50，那么表示在 79 个之中有 50 个对所选段进行了响应。

概率。 指示段概率。 例如，如果涉及范围为 79，频率为 50，那么表示该段的概率为 63.29%（50 除以 79）。

错误。 指示段错误。

窗格底部的信息显示整个模型的涉及范围、频率和概率。

工作模型工具栏

工作模型窗格的工具栏提供了以下功能。

注：也可以通过右键单击模型段来访问某些功能。

表 *9:* 工作模型工具栏按钮

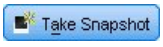
工具栏按钮	**描述**
	启动生成新模型对话框，该对话框提供用于创建新模型块的选项。
	保存交互会话的当前状态。 这会将“决策列表”建模节点更新为当前设置，包括挖掘任务、模型快照、数据选择和定制测量。 要将会话恢复至此状态，选中建模节点的“模型”选项卡中的**使用保存的会话信息**对话框，然后单击**运行**。
	显示“组织模型测量”对话框。 有关更多信息，请参阅主题 第 121 页的『组织模型测量』。
	显示“组织数据选择”对话框。 有关更多信息，请参阅主题 第 118 页的『组织数据选择』。
	显示“快照”选项卡。 有关更多信息，请参阅主题 第 115 页的『“快照”选项卡』。
	显示“替代”选项卡。 有关更多信息，请参阅主题 第 115 页的『“替代”选项卡』。
	获取当前模型结构的快照。 快照显示在“快照”选项卡中，通常用于模型比较。
	启动插入段对话框，该对话框提供用于创建新模型段的选项。
	启动编辑段规则对话框，该对话框提供的选项可用于将条件添加到模型段，或更改先前定义的模型段条件。
	在模型层次中将所选段上移。

表 9: 工作模型工具栏按钮 (继续)	
	在模型层次中将所选段下移。
	删除所选段。
	在模型中包括/排除所选段的情况之间进行切换。 排除时，段结果将计入余数。 不同于删除段的是，排除段允许您选择重新激活段。

“替代”选项卡

单击**查找段**生成“替代”选项卡，该选项卡将针对工作模型窗格中的选定模型或段列出所有替代挖掘结果。

要将替代模型提升为工作模型，突出显示所需替代模型并单击**加载**；则替代模型显示在工作模型窗格中。

注：只有当您已在决策列表建模节点的“专家”选项卡上设置了**最大替代项数**时，才会显示“替代”选项卡以创建多个替代项。

每个已生成的模型替代项会显示特定的模型信息：

名称。 每个替代模型都有顺序编号。 第一个替代项通常包含最佳结果。

目标。 指明目标值。 例如：1，等于“true”。

分段数。 替代模型中所使用的段规则数。

覆盖。 替代模型的涉及范围。

频率 相对于涉及范围的匹配项的数量。

概率 指明替代模型的概率百分比。

注：替代结果不会随模型保存；结果仅在活动会话期间有效。

“快照”选项卡

快照是模型在特定时间点的视图。 例如，如果您需要将另一个替代模型加载工作模型窗格、但不希望失去当前模型的相关工作，那么可以获取模型快照。 “快照”选项卡将列出在任意数量的工作模型状态下手动获取的所有模型快照。

注：快照将随模型保存。 我们建议在您加载首个模型时执行快照。 该快照用于保存原始模型结构，从而确保您可随时返回原始模型状态。 生成的快照名称显示为时间戳，指示其生成时间。

创建模型快照

1. 选择要在工作模型窗格中显示的适当的模型/替代项。
2. 对该工作模型进行必要的更改。
3. 单击 **执行快照** 。 此时将在“快照”选项卡中显示一个新快照。

   名称。 快照名称。 您可以双击快照名称对其进行更改。

   目标。 指明目标值。 例如：1，等于“true”。

   分段数。 模型中所使用的段规则数。

   覆盖。 模型的涉及范围。

   频率 相对于涉及范围的匹配项的数量。

   概率 指明模型的概率百分比。
4. 要将快照提升为工作模型，突出显示所需快照并单击**加载**；则快照模型显示在工作模型窗格中。
5. 可通过以下方法删除快照：单击 **删除** ，或右键单击快照，然后在菜单中选择 **删除** 。

使用 Decision List Viewer

将以最佳方式预测客户响应和行为的模型是通过多个阶段进行构建的。 启动 Decision List Viewer 时，工作模型将填入已定义的模型段和测量，并且准备就绪，等待您启动挖掘任务、根据需要修改段/测量，并生成新的模型或建模节点。

您可添加一个或多个段规则，直到获得满意的模型。 可以通过运行挖掘任务或使用 **编辑段规则** 功能为模型添加段规则。

在模型构建过程中，您可以对模型的性能进行评估，方法是根据测量数据验证模型、在图表中对图形进行可视化处理，或生成定制 Excel 测量量。

肯定模型的质量后，您可以生成新模型并将其置于 IBM SPSS Modeler 工作区或模型选用板中。

挖掘任务

挖掘任务 是确定新规则生成方式的参数的集合。 其中某些参数是可以选择的，以便为您提供使模型适应新状况的灵活性。 任务由任务模板（类型）、目标和构建选择（挖掘数据集）组成。

下列各部分详细介绍各种挖掘任务操作：

- 第 116 页的『运行挖掘任务』
- 第 116 页的『创建和编辑挖掘任务』
- 第 118 页的『组织数据选择』

运行挖掘任务

通过 Decision List Viewer，您可以运行挖掘任务或在模型之间复制和粘贴段规则以手动向模型添加段规则。 挖掘任务包含有关如何生成新段规则的信息（数据挖掘参数设置，如搜索策略、源属性、搜索宽度、置信度级别等）、待预测的客户行为，以及要调查的数据。 挖掘任务的目标是搜索可能的最佳段规则。

要通过运行挖掘任务生成模型段规则，请执行下列操作：

1. 单击 **余数** 行。 如果工作模型窗格中已有显示的段，您也可以选择其中某一个，根据所选段查找其他规则。 选择余数或段之后，可采用下列方法之一生成模型或替代模型：

- 从“工具”菜单选择**查找段**。
- 右键单击**余数**行/段，然后选择**查找段**。
- 单击工作模型窗格上的**查找段**按钮。

在任务处理过程中，进度将在工作区底部显示，并在任务完成时提示您。 任务完成所用的时间完全取决于挖掘任务的复杂性以及数据集的大小。 如果结果中只有一个模型，那么任务完成后它将立即显示在工作模型窗格上；但是，如果结果包含多个模型，那么模型显示在“替代”选项卡上。

注：任务结果将为：完成并更新模型、完成但不更新模型或失败。

可以重复查找新段规则的过程，直到不再有新规则添加到模型中。 这表示已找到所有有意义的客户组。

可以对任何现有的模型段运行挖掘任务。 如果对任务的结果不满意，您可以选择对同一模型段启动另一个挖掘任务。 此操作将基于所选段提供找到的其他规则。 位于所选段“下方”的段（即，在所选段之后添加到模型的段）将被新段替代，因为每个段都取决于其前项。

创建和编辑挖掘任务

挖掘任务是搜索组成数据模型的规则集合的机制。 除所选模板中定义的搜索条件外，任务还会定义目标（激发分析的实际问题，如有多少客户可能对邮件做出响应），并标识要使用的数据集。 挖掘任务的目标是搜索可能的最佳模型。

创建挖掘任务

要创建挖掘任务，请执行下列操作：

1. 选择要在其中挖掘其他段条件的段。
2. 单击**设置**。 此时将打开“创建/编辑挖掘任务”对话框。 该对话框提供用于定义挖掘任务的选项。

3. 进行必要的更改，然后单击**确定**以返回到工作模型窗格。 Decision List Viewer 使用这些设置作为每个任务运行的缺省值，除非选择了其他任务或设置。
4. 单击**查找段**以启动选定段上的挖掘任务。

编辑挖掘任务

“创建/编辑挖掘任务”对话框提供的选项可用于定义新的挖掘任务或编辑现有挖掘任务。

可用于挖掘任务的大部分参数与决策列表节点中提供的参数类似。 例外显示如下。 有关更多信息，请参阅主题 第 112 页的『决策列表模型选项』。

加载设置：创建多个挖掘任务后，请选择所需任务。

新建... 单击此项以根据当前显示的任务的设置创建新的挖掘任务。

目标

目标字段：表示要预测的字段，假定其值与其他字段（预测变量）的值相关。

目标值。 指定目标字段的值，该字段表示要对其进行建模的结果。 例如，如果目标字段 churn 编码为 `0 = no` 和 `1 = yes`，则指定 `1` 以确定用于指示哪些记录可能流失的规则。

简单设置

最大替代项数。 指定运行挖掘任务后将显示的替代项数。 允许的最小设置为 1；没有最大设置。

专家设置

编辑... 打开**编辑高级参数**对话框，以定义高级设置。 有关更多信息，请参阅主题 第 117 页的『编辑高级参数』。

数据

构建选择。 提供的选项用于指定 Decision List Viewer 应对其进行分析以查找新规则的评估度量方式。 列出的评估尺度在“组织数据选择”对话框中进行创建/编辑。

可用字段。 提供用于显示所有字段或手动选择要显示的字段的选项。

编辑... 如果选择了**定制**选项，那么将打开**定制可用字段**对话框，以选择哪些字段可用作挖掘任务找到的段属性。 有关更多信息，请参阅主题 第 117 页的『定制可用字段』。

编辑高级参数

“编辑高级参数”对话框提供以下配置选项。

分级方法。 这是用于对连续字段（计数或宽度相等）进行分级的方法。

分级数。 要为连续字段创建的分级的数目。 允许的最小设置为 2；没有最大设置。

模型搜索宽度。 这是每个周期中可用于下一周期的模型结果的最大数。 允许的最小设置为 1；没有最大设置。

规则搜索宽度。 这是每个周期中可用于下一周期的规则结果的最大数。 允许的最小设置为 1；没有最大设置。

分级合并因子。 段与其相邻段合并时必须增加的最小量。 允许的最小设置为 1.01；没有最大设置。

- **允许条件中出现缺失值**。 `True` 表示允许规则中的 `IS MISSING` 检验。
- **废弃中间结果**。 如果为 `True`，那么将仅返回搜索过程的最终结果。 最终结果是不在搜索过程中进行任何进一步细化的结果。 如果为 `False`，那么还会返回中间结果。

定制可用字段

通过“定制可用字段”对话框，您可以选择可用作通过挖掘任务找到的段属性的字段。

可用。 列出当前可用作段属性的字段。 要从列表中除去字段，请选择相应的字段并单击**除去 >>**。所选字段将从“可用”列表移至“不可用”列表。

不可用。 列出不可用作段属性的字段。 要将字段包含在可用列表中，请选择相应的字段并单击 **<< 添加**。 此时所选字段将从“不可用”列表移至“可用”列表。

组织数据选择

通过组织数据选择（挖掘数据集），可以指定 Decision List Viewer 应对哪些评估尺度进行分析以查找新规则，并选择要用作尺度基准的数据选择。

要组织数据选择，请执行下列操作：

1. 从“工具”菜单中选择**组织数据选择**，或右键单击某个段并选择该选项。 此时将打开“组织数据选择”对话框。

   注：通过“组织数据选择”对话框，您也可以编辑或删除现有数据选择。
2. 单击 **添加新的数据选择** 按钮。 此时会将一个新的数据选择条目添加到现有的表中。
3. 单击 **名称** 并输入适当的选择名称。
4. 单击 **分区** 并选择适当的分区类型。
5. 单击 **条件** 并选择适当的条件选项。 如果选择**指定**，那么会打开“指定选择条件”对话框，其中包含定义特定字段条件的选项。
6. 定义适当的条件，然后单击 **确定** 。

通过“创建/编辑挖掘任务”对话框中的“构建选择”下拉列表可访问这些数据选择。 通过该列表，您可以选择用于特定挖掘任务的评估度量方式。

分段规则

通过运行基于任务模板的挖掘任务，可以查找模型段规则。 您可以使用“插入段”或“编辑段规则”功能手动为模型添加段规则。

如果选择挖掘新的段规则，结果（如果有）将在“交互列表”对话框的“查看器”选项卡中显示。 通过从“模型作品集”对话框中选择替代结果，并单击**加载**，可以快速精练您的模型。 这样，您可以尝试不同结果，直到准备好构建出准确描述最佳目标组的模型。

插入段

您可以使用“插入段”功能手动为模型添加段规则。

要将段规则条件添加到模型，请执行下列操作：

1. 在交互列表对话框中，选择您要添加新段的位置。 新段将直接插在所选段的上方。
2. 在“编辑”菜单中，选择**插入段**或通过右键单击段访问此选项。

   这将打开“插入段”对话框，您可以在其中插入新的段规则条件。
3. 单击 **插入** 。 这将打开“插入条件”对话框，您可以在其中定义新规则条件的属性。
4. 从下拉列表中选择字段和运算符。

   注：如果选择 **Not in** 运算符，那么所选条件将用作排除条件，并且在“插入规则”对话框中显示为红色。 例如，当条件 `region = 'TOWN'` 显示为红色时，表示已从结果集中排除 TOWN。
5. 输入一个或多个值，或者单击**插入值**图标，以显示“插入值”对话框。 通过此对话框，您可以选择为选定字段定义的值。 例如，字段**已婚**将提供值**是**和**否**。
6. 单击**确定**返回“插入段”对话框。 再次单击 **确定** 将所创建的段添加到模型中。

此时该新段将显示在指定的模型位置。

编辑段规则

通过“编辑段规则”功能，您可以添加、更改或删除段规则条件。

要更改段规则条件，请执行下列操作：

1. 选择要编辑的模型段。
2. 从“编辑”菜单选择**编辑段规则**，或右键单击规则以访问此选项。

   此时将打开“编辑段规则”对话框。

3. 选择适当的条件，然后单击 **编辑** 。

   这将打开“编辑条件”对话框，您可以在其中定义所选规则条件的属性。

4. 从下拉列表中选择字段和运算符。

   注：如果选择 **Not in** 运算符，那么所选条件将用作排除条件，并且在“编辑段规则”对话框中显示为红色。 例如，当条件 `region = 'TOWN'` 显示为红色时，表示已从结果集中排除 TOWN。

5. 输入一个或多个值，或单击**插入值**按钮以显示“插入值”对话框。 通过此对话框，您可以选择为选定字段定义的值。 例如，字段**已婚**将提供值**是**和**否**。
6. 单击**确定**返回到“编辑段规则”对话框。 再次单击 **确定** 返回工作模型。

此时所选择的段将与更新的规则条件一起显示。

删除段规则条件

要删除段规则条件：

1. 选择包含要删除的规则条件的模型段。
2. 从“编辑”单中选择**编辑段规则**，或右键单击段以访问此选项。

   这将打开“编辑段规则”对话框，您可在其中删除一项或多项段规则条件。

3. 选择适当的规则条件，然后单击 **删除** 。
4. 单击**确定**。

删除一个或多个段规则条件将使工作模型窗格刷新其测量度量。

复制段

Decision List Viewer 为您提供了一种复制模型段的简便方法。 如果要将一个模型中的段应用于另一个模型时，只需将该段从一个模型复制（或剪切）并粘贴到另一个模型中即可。 此外，您还可以从“替代预览”窗格中显示的模型复制段并将其粘贴到工作模型窗格中显示的模型中。 这些剪切、复制和粘贴功能使用系统剪贴板存储或检索临时数据。 这意味着将在剪贴板中复制条件和目标。 剪贴板内容不仅仅保留用于 Decision List Viewer，也可以粘贴在其他应用程序中。 例如，在文本编辑器中粘贴剪贴板内容时，会以 XML 格式粘贴条件和目标。

要复制或剪切模型段，请执行下列操作：

1. 选择要在其他模型中使用的模型段。
2. 从“编辑”菜单中选择**复制**（或**剪切**），或右键单击模型段并选择**复制**或**剪切**。
3. 打开适当的模型（将在其中粘贴模型段的模型）。
4. 选择某个模型段，然后单击 **粘贴** 。

注：除了**剪切**、**复制**和**粘贴**命令外，您还可以使用以下组合键：**Ctrl+X**、**Ctrl+C** 和 **Ctrl+V**。

复制（剪切）的段将插入先前选择的模型段上方。 粘贴的段和下方段的测量量将重新计算。

注：此过程中的两个模型必须基于同一基础模型模板，并包含同一目标，否则将显示错误消息。

替代模型

当有多个结果时，“替代”选项卡显示每个挖掘任务的结果。 每个结果包含所选数据中与目标最接近匹配的条件，以及所有“相当匹配”的替代项。 显示的替代项总数取决于分析过程中采用的搜索条件。

要查看替代模型，请执行下列操作：

1. 单击“替代”选项卡上的替代模型。 在“替代预览”窗格中，替代模型段显示或替代当前模型段。
2. 要在工作模型窗格中使用替代模型，在“替代预览”窗格中选择模型并单击**加载**，或在“替代”选项卡上右键单击替代模型名称并选择**加载**。

注：生成新模型时，不会保存替代模型。

定制模型

数据不是静态的。 客户会迁移、结婚和更换工作。 产品会随之失去市场焦点并作废。

Decision List Viewer 为业务用户提供了使模型方便迅速地适应新状况的灵活性。 您可通过编辑、设置优先级、删除或停用特定模型段来更改模型。

为段设置优先级

您可选择任意顺序，对模型规则进行排列。 缺省情况下，模型段按优先级顺序显示，第一个段具有最高优先级。 当您为一个或多个段指定不同的优先级时，模型会发生相应的更改。 您可以根据需要通过将段移至较高或较低的优先级位置来更改模型。

要为模型段设置优先级，请执行下列操作：

1. 选择要为其指定不同优先级的模型段。
2. 单击工作模型窗格工具栏中的两个箭头按钮之一，将所选模型段在列表中上移或下移。

设置优先级后，会重新计算先前的所有评估结果，并显示新值。

删除段

要删除一个或多个段，请执行下列操作：

1. 选择模型段。
2. 从“编辑”菜单中选择**删除段**，或在工作模型窗格的工具栏中单击删除按钮。

测量量将针对修改后的模型重新计算，模型也会发生相应的更改。

排除段

在搜索特定组时，您可能会将一部分模型段作为商业操作的基准。 部署模型时，您可能会选择排除模型中的某些段。 排除的段作为空值进行评分。 排除某个段并不代表不使用该段，而是从邮件列表中排除与该规则匹配的所有记录。 该规则仍在应用，但方式不同。

要排除特定的模型段，请执行下列操作：

1. 在工作模型窗格中选择一个段。
2. 在工作模型窗格的工具栏中单击 **切换段排除** 按钮。 此时将在所选段的所选“目标”列中显示**已排除**。

注：与删除的段不同，已排除的段在最终模型中仍然可供重复使用。 已排除的段仍将影响图表结果。

更改目标值

通过“更改目标值”对话框，您可以更改当前目标字段的目标值。

与工作模型具有不同目标值的快照和会话结果会通过将该行的表背景变为黄色进行标识。 这表示该快照/会话结果已过时。

创建/编辑挖掘任务对话框将显示当前工作模型的目标值。 该目标值不会随挖掘任务保存， 而是取自工作模型的值。

当您将某个与当前工作模型具有不同目标值的已保存模型提升为工作模型（例如，通过编辑替代结果或编辑快照副本）时，已保存模型的目标值将更改为工作模型的目标值（工作模型窗格中显示的目标值不会更改）。 模型度量将根据新目标重新计算。

生成新的模型

“生成新模型”对话框提供的选项可用于命名模型并选择创建新节点的位置。

模型名称。 选择**定制**可调整自动生成的名称，或为流工作区中显示的节点创建唯一名称。

创建节点于。 选择**工作区**会将新模型置于工作区中；选择 **GM 选用板**会将新模型置于“模型”选用板中；选择**两者**会将新模型同时置于工作区和“模型”选用板中。

包含交互式会话状态。 启用此选项后，交互式会话状态将保留在已生成的模型中。 稍后根据模型生成建模节点时，该状态将继续传递并用于初始化交互会话。 无论是否选择此选项，模型本身对新数据的评分方式都是相同的。 如果未选择此选项，模型仍然可以创构建建节点，但该节点将更为一般化，它会启动新的交互会话而不是从原有会话停止的位置继续前进。 如果更改节点设置但以保存的某种状态执行，那么会忽略已更改的设置以采用保存状态的设置。

注：标准度量是属于模型的唯一度量。 其他度量将保留在交互状态。 生成的模型不会显示已保存的交互挖掘任务状态。 启动 Decision List Viewer 时，它会显示通过查看器所做的初始设置。

有关更多信息，请参阅主题 第 36 页的『重新生成建模节点』。

模型评估

成功建模需要先对模型进行仔细评估，然后才能在生产环境中实施。 Decision List Viewer 提供了一些统计和业务度量，可用于评估模型在现实世界中的影响。 其中包括增益图和与 Excel 的全面互操作，从而实现成本/增益方案的模拟，以便评估部署的作用。

您可采用以下方式评估自己的模型：

- 使用 Decision List Viewer 中提供的预定义的统计测量和商业模型测量（概率、频率）。
- 评估从 Microsoft Excel 中导入的测量。
- 使用增益图对模型进行可视化处理。

组织模型测量

Decision List Viewer 提供了用于定义按列计算并显示的测量的选项。 每个段可包括缺省的涉及范围、频率、概率和错误等测量量，按列显示。 此外，您也可以创建将按列显示的新测量量。

定义模型测量

要为模型添加测量量或定义现有的测量量，请执行下列操作：

1. 从“工具”菜单中选择**组织模型测量**，或右键单击模型以选择此选项。 此时将打开“组织模型测量”对话框。
2. 单击**添加新的模型测量**按钮（位于“显示”列右侧）。 此时将在表中显示一个新的测量量。
3. 提供测量量名称，并选择适当的类型、显示选项和选择。 “显示”列指示是否为工作模型显示测量。 定义现有测量量时，请选择适当的度量和选择，并指定该度量是否将在工作模型中显示。
4. 单击 **确定** 返回 Decision List Viewer 工作区。 如果已选中新测量的“显示”列，那么会为工作模型显示该新测量。

Excel 中的定制度量

有关更多信息，请参阅主题 第 121 页的『Excel 中的评估』。

刷新测量

在某些特定情况下，可能需要重新计算模型测量，例如对一组新客户应用现有模型时。

要重新计算（刷新）模型测量，请执行下列操作：

在“编辑”菜单中选择**刷新所有测量量**。

或者

按 F5。

此时将重新计算所有测量量，并针对工作模型显示新值。

Excel 中的评估

Decision List Viewer 可以与 Microsoft Excel 进行集成，使您可以在模型构建过程中直接使用自己的值计算和利润公式来模拟成本/收益方案。 通过与 Excel 的链接，您可以将数据导出至 Excel（数据在其中可用于创建演示图表）、计算定制测量（如复杂利润和 ROI 测量），并且可以在构建模型时通过 Decision List Viewer 查看这些测量。

注：要使用 Excel 电子表格，CRM 分析专家必须针对 Decision List Viewer 与 Microsoft Excel 的同步定义配置信息。 该配置包含于 Excel 电子表格文件中，用于指明 Decision List Viewer 与 Excel 之间相互传输的信息。

以下步骤仅在已安装 MS Excel 的情况下有效。 如果未安装 Excel，那么不会显示使模型与 Excel 同步的选项。

要使模型与 MS Excel 同步，请执行下列操作：

1. 打开模型，运行交互会话，并从“工具”菜单中选择**组织模型测量**。
2. 为 **计算 Excel 中的定制测量量** 选项选择 **是** 。 这将激活**工作簿**字段，使您可以选择预先配置的 Excel 工作簿模板。
3. 单击 **连接到 Excel** 按钮。 这将打开“打开”对话框，使您可以导航至本地或网络文件系统中预先配置的模板所在的位置。
4. 选择适当的 Excel 模板，然后单击 **打开** 。 此时将启动所选的 Excel 模板；使用 Windows 任务栏（或按 Alt+Tab）返回到“选择定制测量的输入”对话框。
5. 在 Excel 模板中定义的度量名称与模型度量名称之间选择适当的映射，然后单击 **确定** 。

建立链接后，Excel 将立即采用预先配置的 Excel 模板启动，该模板以电子表格显示模型规则。 Excel 中的计算结果在 Decision List Viewer 中显示为新列。

注：保存模型时，不会保留 Excel 度量；度量仅在活动会话期间有效。 但是，您可以创建包括 Excel 度量的快照。 在快照视图中保存的 Excel 度量仅适用于历史记录比较，在重新打开时不会刷新。 有关更多信息，请参阅主题 第 115 页的『“快照”选项卡』。 重新建立与 Excel 模板的连接前，Excel 度量将不会显示在快照中。

MS Excel 集成设置

Decision List Viewer 与 Microsoft Excel 的集成是通过使用预先配置的 Excel 电子表格模板实现的。 该模板由以下三个工作表组成：

模型度量。 显示导入的 Decision List Viewer 测量、定制 Excel 测量，以及计算总计（在“设置”工作表中定义）。

“设置”。 提供用于根据已导入的 Decision List Viewer 测量和定制 Excel 测量生成计算的变量。

配置。 提供用于指定从 Decision List Viewer 导入哪些测量以及用于定义定制 Excel 测量的选项。

警告：“配置”工作表的结构已严格定义。 请 **勿** 编辑绿色阴影区域中的任何单元。

- **模型中的度量**。 指示在计算中使用哪些 Decision List Viewer 度量。
- **要建模的度量**。 指示 Excel 生成的哪些度量将返回到 Decision List Viewer。 Excel 生成的度量在 Decision List Viewer 中显示为新的测量列。

注：生成新模型时，Excel 度量不会随模型一起保留；这些度量仅在活动会话期间有效。

更改模型测量

下列示例演示如何通过多种方法更改模型测量：

- 更改现有测量。
- 从模型导入其他标准测量。
- 将其他定制测量导出到模型。

更改现有测量

1. 打开模板并选择“配置”工作表。
2. 通过突出显示并重写名称或说明来编辑任何 **名称** 或 **说明** 。

   请注意，如果要更改测量（例如，为了提示用户概率而非频率），只需更改 **来自模型的度量** 中的名称和说明，该名称和说明随后将显示在模型中并且用户可以选择要映射的恰当测量。

从模型导入其他标准测量

1. 打开模板并选择“配置”工作表。
2. 从菜单中选择：

   工具 > 保护 > 取消保护工作表
3. 选择 A5 单元格，该单元格有黄色阴影且包含 **结束** 字。
4. 从菜单中选择：

   插入 > 行
5. 在新测量的 **名称** 和 **说明** 中键入相应内容。 例如， **错误** 和 **段的相关错误** 。
6. 在单元格 C5 中，输入公式 **=COLUMN('Model Measures'!N3)**.
7. 在单元格 D5 中，输入公式 **=ROW('Model Measures'!N3)+1**。

   这些公式会使新的测量显示在模型测量工作表的 N 列中，此列目前为空。
8. 从菜单中选择：

   工具 > 保护 > 保护工作表
9. 单击**确定**。
10. 在模型测量工作表中，确保 N3 单元格已将 **错误** 作为新列的标题。
11. 选择整个 N 列。
12. 从菜单中选择：

    格式 > 单元格
13. 缺省情况下，所有单元均有一个 **一般** 数字类别。 单击 **百分比** 可更改数字显示的方式。 此方法可帮助您检查 Excel 中的数字；此外，也提供给您另外一种使用数字的方法，例如可将数字用作图表的输出。
14. 单击**确定**。
15. 将电子表格保存为 Excel 2003 模板，该模板具有唯一的名称且文件扩展名为 *.xlt*。 为了易于定位新模板， 建议您将其保存在本地或网络文件系统上的预先配置的模板中。

将其他定制测量导出到模型

1. 打开之前示例中已添加“错误”列的模板；选择“配置”工作表。
2. 从菜单中选择：

   工具 > 保护 > 取消保护工作表
3. 选择 A14 单元格，该单元格有黄色阴影且包含 **结束** 字。
4. 从菜单中选择：

   插入 > 行
5. 在新测量的 **名称** 和 **说明** 中键入相应内容。 例如， **定比变换错误** 和 **应用于 Excel 错误的定比变换** 。
6. 在单元格 C14 中，输入公式 **=COLUMN('Model Measures'!O3)**.
7. 在单元格 D14 中，输入公式 **=ROW('Model Measures'!O3)+1**。

   这些公式指定 O 列将提供模型的新测量。
8. 选择“设置”工作表。
9. 在 A17 单元格中输入说明 **'- 定比变化错误** 。
10. 在 B17 单元格中输入 **10** 的定比变换因子。
11. 在“模型测量”工作表中，在 O3 单元格中输入说明**定比变换错误**作为新列的标题。
12. 在 O4 单元格中输入公式 **=N4*Settings!B17** 。
13. 选择 O4 单元格的右下角并将其向下拖动到 O22 单元格，以将公式复制到每一个单元格中。
14. 从菜单中选择：

    工具 > 保护 > 保护工作表
15. 单击**确定**。

16. 将电子表格保存为 Excel 2003 模板，该模板具有唯一的名称且文件扩展名为 *.xlt*。 为了易于定位新模板， 建议您将其保存在本地或网络文件系统上的预先配置的模板中。

当使用该模板连接 Excel 时，“错误”值可用作新的定制测量。

对模型进行可视化处理

了解模型作用的最佳方式是对其进行可视化处理。 使用增益图，可以通过研究多个替代项的实际效果深入掌握有关模型商业增益和技术增益的有价值的日常信息。 第 124 页的『增益图』部分显示了某个模型在随机决策过程中的增益，并且可以在存在替代模型时实现对多个图表的直接比较。

增益图

增益图绘制的是表中 增益 % 列值的散点图。 增益定义为每个增量中匹配项数与树中匹配项总数的比例，它使用下列等式：

（增量中匹配项数/匹配项总数）x 100%

增益图有效地为您说明需要怎样的撒网广度才能捕获树中所有匹配项的给定百分比。 斜线绘制整个样本在未使用模型的情况下的预期响应。 这种情况下，响应率应该为常量，因为一个人响应的可能性与另一个人相同。 为了使您的收益加倍，您需要询问两倍数量的人。 曲线表明通过将那些秩（基于增益排序）位于较高百分比的人员包括在内，您可以使得响应得到多大程度的改善。 例如，包括最高的 50% 可能会网罗超过 70% 的正面响应。 该曲线越陡，增益越高。

要查看增益图，请执行下列操作：

1. 打开包含决策列表节点的流，并从该节点启动一个交互会话。
2. 单击 **增益** 选项卡。 根据指定的分区，您会看到一个或两个图表（例如，如果同时为模型测量定义了训练分区和检验分区，那么会显示两个图表）。

缺省情况下，图表会显示为段。 您可以将图表切换为分位数显示，方法是选择 **分位数** ，然后在下拉菜单中选择适当的分位数方法。

图表选项

“图表选项”功能提供的选项可用于选择以图表显示哪些模型和快照、绘制哪些分区，以及是否显示段标签。

要绘制的模型

当前模型。 使您可以选择要绘制的模型。 您可以选择工作模型或任何已创建的快照模型。

要绘制的分区

左侧图表的分区。 该下拉列表提供用于显示所有已定义分区或所有数据的选项。

右侧图表的分区。 该下拉列表提供用于显示所有已定义分区、所有数据或仅显示左侧图表的选项。 如果选择 **只绘制左侧图** ，那么仅显示左侧图表。

显示段标签。 启用此选项后，将在图表中显示全部段标签。

第 10 章 统计模型

统计模型使用数学方程式对从数据中提取的信息进行编码。 在某些情况下，统计建模方法可以非常快速地给出合适的模型。 甚至对于那些只有更加灵活的机器学习方法（例如神经网络）才能最终给出更好结果的问题，仍然可以将某些统计模型用作基线预测模型以判断更先进方法的性能。

提供了下列统计建模节点。

线性回归模型根据目标与一个或多个预测变量之间的线性关系来预测连续目标。

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。 它与线性回归类似，但采用分类目标字段而不是数字范围。

“PCA/因子”节点提供用于降低数据复杂程度的强大数据降维技术。 主成份分析（PCA）可找出输入字段的线性组合，该组合最好地捕获了整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。 因子分析则尝试识别底层因素，这些因素说明了观测的字段集合内的相关性模式。 对于这两种方法，其共同的目标是找到可对原始字段集合中的信息进行有效总结的少量派生字段。

判别分析提出比 Logistic 回归更加严格的假设，但是在满足这些假设时可成为 Logistic 回归的有价值替代方案或补充。

广义线性模型对广义线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。 而且，该模型还允许因变量呈非正态分布。 它涵盖了大量统计模型的功能，包括线性回归、逻辑回归、计数数据的对数线性模型和区间删失生存模型。

广义线性混合模型 (GLMM) 扩展了线性模型，使得目标可以有非正态分布，通过指定的连接函数与因子和协变量线性相关，并且观测值可能相关。 广义线性混合模型涵盖了各种模型，从简单线性回归模型到非正态纵向模型数据的复杂多级模型。

使用 Cox 回归节点，您可以在已有的检查记录中建立时间事件的生存模型。 对于输入变量的给定值，该模型会生成一个生存函数，用来预测在给定时间 (*t*) 发生相关事件的概率。

线性节点

线性回归是一种常见的统计方法，用于根据数字输入字段的值对记录进行分类。 线性回归拟合一条直线或一个平面，该直线或平面将预测输出值与实际输出值之间的差异最小化。

需求。 在线性回归模型中只能使用数字字段。 必须有且仅有一个目标字段（角色设置为**目标**），但可以有一个或多个预测变量（角色设置为**输入**）。 角色为**两者**或**无**的字段将被忽略，就像对待非数值字段一样。（如有必要，可以使用“派生”节点对非数字字段进行重新编码。）

强度。 线性回归模型相对简单，用来形成预测的数学公司易于解释。 由于线性回归是一种由来已久的统计方法，因此这些模型的属性已广为人所熟知。 通常，线性模型的训练速度也非常快。 “线性”节点提供了自动字段选择方法，以排除方程中不重要的输入字段。

注: 如果目标字段为分类（例如**是/否**或**流失/未流失**）而非连续范围，那么可以将 Logistic 回归用作替代项。 Logistic 回归还支持非数值输入，因而无需对这些字段进行重新编码。 有关更多信息，请参阅主题 第 133 页的『Logistic 节点』。

线性模型

线性模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。

线性模型相对简单，用于评分的数学公式也易于解释。 这些模型的属性比较好理解，与同一数据集上的其他模型类型（如神经网络或决策树）相比能够非常快速构建。

示例。 在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来估计理赔成本。 通过在服务中心部署该模型，客服代表可以在接听客户电话的同时输入理赔信息，并立即获得基于以往数据的“预期”成本。

字段要求。 必须有一个目标和至少一个输入。 缺省情况下，不使用带“两者”或“无”预定义角色的字段。 目标必须是连续的（刻度）。 对预测变量（输入）没有测量级别限制。分类（标志、名义和有序）字段用作模型中的因子，同时连续字段用作协变量。

目标

您要执行什么操作?

- **构建新模型**。 构建全新的模型。 这是节点的常见操作。
- **继续训练现有的模型**。 继续训练此节点成功生成的最后一个模型。 这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，因为只有新的或更新后的记录被传入流中。 上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型调色板中不再可用的情况下，也可以使用该选项。

  注: 启用此选项后，“字段”和“构建选项”选项卡上的所有其他控件将处于禁用状态。

您的主要目标是什么? 请选择适当的目标。

- **创建标准模型**。 使用一种方法来构建一个可以使用预测变量预测目标的模型。 一般来说，标准模型更易于理解，而且评分速度比 boosted、bagged 或大型数据集整体更快。

  注: 只有**构建单一树**拆分模型中才支持**继续训练现有模型**，并且必须连接到 Analytic Server。
- **增强模型准确度（提升）**。 使用 Boosting 构建整体模型的方法，可生成一个模型序列来获得更多精确预测值。 与标准模型相比，整体模型需要更长的时间来构建和评分。

  Boosting 将生成一连串的“组件模型”，其中的每个模型都基于整个数据集进行构建。 在构建每个连续的组件模型之前，将根据上一个组件模型的残值确定记录的权重。 残值较大的个案将被赋予相对较高的分析权重，以使下一个组件模型还侧重于预测这些记录。 这些成分模型共同构成一个整体模型。 这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。
- **增强模型稳定性（组装）**。 使用 Bagging (bootstrap 汇总) 构建整体模型的方法，可生成多个模型来获得更多可靠的预测值。 与标准模型相比，整体模型需要更长的时间来构建和评分。

  Bootstrap 汇总 (Bagging) 通过对原始数据集进行放回方式的取样，生成训练数据集的副本。 这将创建大小与原始数据集相同的 Bootstrap 样本。 然后，在每个副本上构建“成分模型”。 这些成分模型共同构成一个整体模型。 这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。
- **为非常大的数据集创建模型**。 通过将数据集拆分成单独的数据块来构建整体模型的方法。 如果您的数据集非常大，无法构建任何上述模型，或希望用于增量建模，请选择该选项。 该选项可使用更短的时间来构建模型，但需要比标准模型更长的时间来评分。

有关增强、组装和超大型数据集的设置，请参阅第 127 页的『整体』。

基本

自动准备数据。 该选项允许在内部转换目标和预测变量，以使模型的预测能力最大化；将保存模型的任何转换并应用到新数据用于评分。 转换字段的原始版本将从模型中排除。 缺省情况下，将执行以下自动数据准备。

- **日期与时间处理**。 每个日期预测变量被转换成新的连续预测变量，其中包含自参考日期 (1970-01-01) 以来经过的时间。 每个时间预测变量被转换成新的连续预测变量，其中包含自参考时间 (00:00:00) 以来经过的时间。
- **调整测量级别**。 具有少于 5 个不同值的连续预测变量将被重新强制转换为有序预测变量。 具有多于 10 的不同值的有序预测变量将重新强制转换为连续预测变量。
- **离群值处理**。 如果连续预测变量的值位于分界值（平均值的 3 个标准差）之外，那么将其设为分界值。
- **缺失值处理**。 名义预测变量的缺失值被替换为训练分区的众数。 有序预测变量的缺失值被替换为训练分区的中位数。 连续预测变量的缺失值被替换为训练分区的平均值。
- **受监督的合并**。 这将减少与目标关联的需处理的字段数，得到更简约的模型。 通过输入与目标间的关系可以确定类似的类别。 无显著差异（即 p 值大于 0.1）的类别则被合并。 如果所有类别合并为一个类别，那么字段的原始和派生版本将从模型中排除，因为它们没有作为预测变量的值。

置信度级别。 此为用于在系数视图中计算模型系数的区间估计值的置信度。 请指定大于 0 且小于 100 的值。 缺省值为 95。

模型选择

模型选择方法。 选择一种模型选择方法（下面将详细介绍）或**包括所有预测变量**，后者简单地输入所有可用预测变量作为主效应模型项。 缺省使用**前向逐步**。

向前逐步选择。 在开始时模型中没有任何效应，然后在每个步骤中添加和删除效应，直到根据逐步选择标准不能再添加或删除效应为止。

- **输入/移除条件**。 此为用于决定是将某个效应添加到还是剔除出模型的统计。 **信息标准 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。 **F 统计**基于有关模型错误改进情况的某个统计检验。 **调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。 **防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。 防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

  如果选择了 **F 统计**以外的标准，那么在每步中将对应于选择标准的最大正增长的效应添加到模型。 对应于标准中减少情况的任何模型效应将被移除。

  如果选择了 **F 统计**作为标准，那么在每步中将具有低于指定阈值（**纳入 p 值小于此值的效应**）的最小 *p* 值的效应添加到模型。 缺省值为 0.05。 任何具有大于指定阈值**移除 p 值大于此值的效应**的 *p* 值的模型效应将被移除。 缺省值为 0.10。
- **在最终模型中定制最大效应数**。 缺省情况下，所有可用效应都将被输入模型中。 或者，如果逐步选择算法在具有指定最大效应数的某个步骤结束，那么此算法将以当前效应集合结束。
- **定制最多步骤数**。 逐步选择算法在达到特定步骤数后停止。 此值缺省为可用效应数的 3 倍。 或者，指定一个正整数作为最大步骤数。

最佳子集选择。 这将检查“所有可能的”模型，或至少检查可能模型的较大子集（大于“前向逐步”方法），以选择满足相应标准的最佳子集。 **信息标准 (AICC)** 基于给定模型的训练集的似然性，并进行调整以惩罚过于复杂的模型。 **调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。 **防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。 防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

选择具有最大标准值的模型作为最佳模型。

注: 与向前逐步选择相比，最佳子集选择涉及更密集的计算。 在与 Boosting、Bagging 或超大型数据集配合执行最佳子集时，花费的时间比使用向前逐步选择构建标准模型要长得多。

整体

这些设置决定了在“目标”中请求 Boosting、Bagging 或超大型数据集时发生的整体行为。 将忽略不适用于选定目标的选项。

Bagging 和大型数据集 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体评分值。

- **连续目标的缺省合并规则**。 可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，那么组合规则选择将被忽略。 Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

提升和组装。 当以增强模型精确性或稳定性为目标时，指定要构建的基本模型数；对于 Bagging 方法，此为 bootstrap 样本数。 它应为正整数。

高级

复制结果。 设置随机种子允许您复制分析。 随机数生成器用于选择哪个记录在过度拟合集中。 指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。 缺省值为 54752075。

模型选项

模型名称。 可以基于目标字段来自动生成模型名称，或指定定制名称。 自动生成的名称为目标字段名。

请注意，在对模型评分时，始终会计算预测值。 新字段的名称是目标字段的名称，前缀为 *$L-*。例如，对于名为销售的目标字段，新字段将命名为 *$L* - 销售。

模型摘要

“模型摘要”视图是一个快照，提供模型及其拟合的一览摘要。

表
表中列出一些高级模型设置，包括：

- 字段选项卡上指定的目标名称，
- 是否已按基本设置中指定的方式执行执行自动数据准备，
- 模型选择设置中指定的模型选择方法和选择标准。 还显示了最终模型的选择标准值，并以较小、较佳的格式显示。

图表
此图表显示最终模型的准确性，数值越大越好。 对于最终模型，此值为 100 × 调整后的 R^2。

自动数据准备

此视图显示在自动数据准备 (ADP) 步骤中排除了哪些字段，以及转换字段的派生方式等信息。 对于每个转换或排除字段，在此表中列出了字段名、在分析中的角色，以及 ADP 步骤所采取的操作。 字段是按照字段名称的字母升序排列的。 对每个字段执行的可能操作包括：

- **派生持续时间：月份**计算从包含日期的字段中的值到当前系统日期经过的时间（以月份为单位）。
- **派生持续时间：小时**计算从包含时间的字段中的值到当前系统时间经过的时间（以小时为单位）。
- **将测量级别从连续改为有序**将不到 5 个唯一值的连续字段重新强制转换为有序字段。
- **将测量级别从有序改为连续**将超过 10 个唯一值的有序字段重新强制转换为连续字段。
- **删除离群值**如果连续预测变量的值位于分界值（平均值的 3 个标准差）之外，那么将其设为分界值。
- **替换缺失值**分别使用众数、中位数和平均值替换名义字段、有序字段和连续字段的缺失值。
- **合并类别以最大化与目标的关联**根据输入与目标间的关系确定“类似”的预测变量类别。 无显著差异（即 *p* 值大于 0.05）的类别则被合并。
- **排除常量预测变量/在离群值处理之后/在合并类别之后**删除具有单个值的预测变量，可能在执行其他 ADP 操作之后。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。 预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

按已观测进行预测

这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。 理想情况下，该点应在 45 度线上；您可以从该视图上判断出任何被模型预测为较差的纪录。

残差

这将显示模型残差的诊断图。

图表样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **直方图**。 此为 Student 化的残差的分级直方图，并带有正态分布交叠。 线性模型假设残差具有正态分布，因此理想情况下直方图应相当接近平滑线。
- **P-P 图**。 此为分级概率-概率 (P-P) 图，将 Student 化的残差与正态分布进行对比。 如果绘制点的斜率比正态线更平缓，那么残差显示出比正态分布更显著的可变性；如果更陡峭，那么残差的可变性低于正态分布。 如果绘制点呈 S 型曲线，那么残差为偏斜分布。

离群值

此表列出对模型施加过度影响的记录，并显示记录 ID（如果在“字段”选项卡上指定）、目标值，以及 Cook 距离。 Cook 距离是在特定记录从模型系数的计算中排除的情况下，所有记录的残差变化幅度的测量。 较大的 Cook 距离表示在排除记录后系数会发生显著变化，因此应被视为有一定影响。

应仔细检查有影响的记录，以确定是在模型估计中给予较低权重，按照特定可接受阈值截断离群值，还是彻底移除有影响的记录。

效应

此视图显示模型中每个效应的大小。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **图表**。 在此图表中，将按预测变量重要性递减顺序，从上到下排列显示效应。 在图表中，连接线条根据效应的显著性进行加权，粗线条表示较显著的效应（*p* 值较小）。 悬停在连接线条上将显示工具提示，以指示效应的 *p* 值和重要性。 这是缺省选项。
- **表**。 此为总体模型与单独模型效应的 ANOVA 表。 各个效应将按预测变量重要性递减顺序，从上到下排列显示。 请注意，缺省情况下该表处于折叠状态，以便仅显示整体模型的结果。 要查看单独模型效应的结果，在表中单击**校正的模型**单元格。

预测变量重要性。 提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。 这不会改变模型，只是帮助您重点关注最重要的预测变量。 缺省情况下，将显示前 10 个效应。

显著性。 提供有一个“显著性”滑块，以便在按预测变量重要性显示效应的基础上，进一步控制在视图中显示哪些效应。 将隐藏显著性值大于滑块值的效应。 这不会改变模型，只是帮助您重点关注最重要的效应。 缺省情况下此值为 1.00，因此不会根据显著性来过滤效应。

系数

此视图显示模型中每个系数的值。 注意，由于因子（分类预测变量）在模型内部经过指示符编码，因此包含因子的**效应**通常具有多个关联**系数**；每种类别一个关联系数，但对应于冗余（参考）参数的类别除外。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **图表**。 在此图表中，首先显示截距，然后按预测变量重要性递减顺序，从上到下排列显示效应。 在包含因子的效应中，系数按照数据值的升序进行排列。 在图表中，连接线条根据系数的显著性（参见图表键）而具有不同颜色，粗线条表示较显著的系数（*p* 值较小）。 悬停在连接线条上将显示工具提示，以指示与参数关联的效应的系数值、*p* 值和重要性。 这是缺省样式。
- **表**。 这将显示单独模型系数的值、显著性检验，以及置信区间。 在截距后面，各个效应将按预测变量重要性递减顺序，从上到下排列显示。 在包含因子的效应中，系数按照数据值的升序进行排列。 请注意，缺省情况下该表将处于折叠状态，以便仅显示每个模型参数的系数、显著性和重要性。 要查看标准误差、*t* 统计和置信区间，在表中单击**系数**单元格。 悬停在表中的模型参数名称上，将显示工具提示，以指示参数名

称、与参数关联的效应以及与模型参数关联的值标签（对于分类预测变量）。 自动数据准备合并分类预测变量的类似类别时，使用此工具提示查看创建的新类别非常有用。

预测变量重要性。 提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。 这不会改变模型，只是帮助您重点关注最重要的预测变量。 缺省情况下，将显示前 10 个效应。

显著性。 提供有一个“显著性”滑块，以便在按预测变量重要性显示系数的基础上，进一步控制在视图中显示哪些系数。 显著性值大于滑块值的系数将被隐藏。 这不会改变模型，只是帮助您重点关注最重要的系数。 缺省情况下，值为 1.00，因此不会根据显著性对系数进行过滤。

估计平均值

这些是针对显著预测变量显示的图表。 在图表中，目标的模型估计值位于垂直轴上，预测变量的每个值位于水平轴上，所有其他预测变量保持恒定。 它提供了有关每个预测变量系数在目标上的效应的可视化，非常有用。

注: 如果没有显著的预测变量，那么不会生成估计均值。

模型构建摘要

如果在“模型选择”设置中选择了**无**以外的模型选择算法，这将提供有关模型构建过程的一些详细信息。

前向逐步。 如果选择算法为前向逐步，此表将显示逐步选择算法中的最近 10 步。 对于其中每个步骤，显示在此步骤上选择标准的值与模型中的效应。 这允许您了解每个步骤对模型的贡献大小。 每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

最佳子集。 如果选择算法为最佳子集，此表将显示前 10 个模型。 对于每个模型，显示选择标准的值与模型中的效应。 您可以从中了解这些最佳模型的稳定性；如果它们倾向于具有存在少量差异的相似效应，那么您可以充分确信它们的确是“最佳”模型；如果它们倾向于具有迥异的效应，那么某些效应可能太相似，需要进行合并（或删除一些）。 每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

设置

请注意，在对模型评分时，始终会计算预测值。 新字段的名称是目标字段的名称，前缀为 *$L-*。例如，对于名为销售的目标字段，新字段将命名为 *$L* - 销售。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **通过转换至本机 SQL 来进行评分**：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分。

  注: 虽然该选项可以更快速获得结果，但是本机 SQL 的大小和复杂性会随着模型复杂性的增加而增加。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

线性-AS 节点

IBM SPSS Modeler 有两个不同版本的线性节点：

- **线性**是在 IBM SPSS Modeler Server 上运行的传统节点。
- 连接到 IBM SPSS Analytic Server 时，可以运行**线性 AS**。

线性回归是一种常见的统计方法，用于根据数字输入字段的值对记录进行分类。 线性回归拟合一条直线或一个平面，该直线或平面将预测输出值与实际输出值之间的差异最小化。

需求。 在线性回归模型中只能使用数字字段和分类预测变量。 必须有且仅有一个目标字段（角色设置为**目标**），但可以有一个或多个预测变量（角色设置为**输入**）。 角色为**两者**或**无**的字段将被忽略，就像对待非数值字段一样。 （如有必要，可以使用“派生”节点对非数字字段进行重新编码。）

强度。 线性回归模型相对简单，用来形成预测的数学公司易于解释。 由于线性回归是一种由来已久的统计方法，因此这些模型的属性已广为人所熟知。 通常，线性模型的训练速度也非常快。 “线性”节点提供了自动字段选择方法，以排除方程中不重要的输入字段。

注: 如果目标字段为分类（例如**是/否**或**流失/未流失**）而非连续范围，那么可以将 Logistic 回归用作替代项。 Logistic 回归还支持非数值输入，因而无需对这些字段进行重新编码。 有关更多信息，请参阅主题 第 133 页的『Logistic 节点』。

线性-AS 模型

线性模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。

线性模型相对简单，用于评分的数学公式也易于解释。 这些模型的属性比较好理解，与同一数据集上的其他模型类型（如神经网络或决策树）相比能够非常快速构建。

示例。 在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来估计理赔成本。 通过在服务中心部署该模型，客服代表可以在接听客户电话的同时输入理赔信息，并立即获得基于以往数据的“预期”成本。

字段要求。 必须有一个目标和至少一个输入。 缺省情况下，不使用带“两者”或“无”预定义角色的字段。 目标必须是连续的（刻度）。 对预测变量（输入）没有测量级别限制。分类（标志、名义和有序）字段用作模型中的因子，同时连续字段用作协变量。

基本

包括截距。 当 x 轴为 0 时，此选项包含 y 轴上的偏移量。 截距通常包括在模型中。 如果您可以假设数据穿过原点，那么可以排除截距。

考虑双向交互。 该选项会告诉模型比较每个可能的输入对，以了解各输入对的趋势之间是否会互相影响。 如果会互相影响，那么这些输入更有可能包含在设计矩阵中。

系数估算值的置信区间 (%)。 这是用于在“系数”视图中计算模型系数估算值的置信区间。 请指定大于 0 且小于 100 的值。 缺省值为 95。

分类预测变量的排序顺序。 这些控件用于确定因子（分类输入）类别的顺序，以确定“最后一个”类别。 如果输入不是分类目标或者指定了定制参考类别，那么将忽略排序顺序设置。

模型选择

模型选择方法。 选择一种模型选择方法（下面将详细介绍）或**包括所有预测变量**，后者简单地输入所有可用预测变量作为主效应模型项。 缺省使用**前向逐步**。

向前逐步选择。 在开始时模型中没有任何效应，然后在每个步骤中添加和删除效应，直到根据逐步选择标准不能再添加或删除效应为止。

- **输入/移除条件**。 此为用于决定是将某个效应添加到还是剔除出模型的统计。 **信息标准 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。 **F 统计**基于有关模型错误改进情况的某个统计检验。 **调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。 **防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。 防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

  如果选择了 **F 统计**以外的标准，那么在每步中将对应于选择标准的最大正增长的效应添加到模型。 对应于标准中减少情况的任何模型效应将被移除。

  如果选择了 **F 统计**作为标准，那么在每步中将具有低于指定阈值（**纳入 p 值小于此值的效应**）的最小 *p* 值的效应添加到模型。 缺省值为 0.05。 任何具有大于指定阈值**移除 p 值大于此值的效应**的 *p* 值的模型效应将被移除。 缺省值为 0.10。
- **在最终模型中定制最大效应数**。 缺省情况下，所有可用效应都将被输入模型中。 或者，如果逐步选择算法在具有指定最大效应数的某个步骤结束，那么此算法将以当前效应集合结束。
- **定制最多步骤数**。 逐步选择算法在达到特定步骤数后停止。 此值缺省为可用效应数的 3 倍。 或者，指定一个正整数作为最大步骤数。

最佳子集选择。 这将检查“所有可能的”模型，或至少检查可能模型的较大子集（大于“前向逐步”方法），以选择满足相应标准的最佳子集。 **信息标准 (AICC)** 基于给定模型的训练集的似然性，并进行调整以惩罚过于复杂的模型。 **调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。 **防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。 防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

选择具有最大标准值的模型作为最佳模型。

注: 与向前逐步选择相比，最佳子集选择涉及更密集的计算。 在与 Boosting、Bagging 或超大型数据集配合执行最佳子集时，花费的时间比使用向前逐步选择构建标准模型要长得多。

模型选项

模型名称。 可以基于目标字段来自动生成模型名称，或指定定制名称。 自动生成的名称为目标字段名。

请注意，在对模型评分时，始终会计算预测值。 新字段的名称是目标字段的名称，前缀为 *$L-*。例如，对于名为销售的目标字段，新字段将命名为 *$L* - 销售。

交互式输出

运行线性-AS 模型后，以下输出可用。

模型信息

“模型信息”视图提供了有关模型的关键信息。 该表标识了一些高级模型设置，例如：

- 字段选项卡上指定的目标名称
- 回归权重字段
- 模型选择设置上指定的模型构建方法
- 预测变量输入的数量。
- 最终模型中预测变量的数量
- 校正赤池信息准则 (AICC)。 AICC 是一种用于基于 -2（受限）对数似然选择和比较混合模型的度量方法。值越小，表示模型拟合得越好。 AICC 用于更正小样本的 AIC。 随样本大小的增加，AICC 将收敛为 AIC。
- R 方。 这是线性模型的拟合优度测量，有时称为决定系数。 它是因变量中由回归模型解释的变异的比例。它的取值范围从 0 到 1。 较小的值表示模型不适合数据。
- 调整后的 R 方

记录摘要

“记录摘要”视图提供了有关模型中包括和排除的记录（观测值）的数目和百分比的信息。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

按已观测进行预测

这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。 理想情况下，该点应在 45 度线上；您可以从该视图上判断出任何被模型预测为较差的纪录。

设置

请注意，在对模型评分时，始终会计算预测值。 新字段的名称是目标字段的名称，前缀为 *$L-*。例如，对于名为销售的目标字段，新字段将命名为 *$L* - 销售。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用 Server Scoring Adapter（如果已安装）进行评分，否则在过程中进行评分。** 如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库之外进行评分。** 此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

Logistic 节点

Logistic 回归（也称为**名义回归**）是一种用于依据输入字段的值对记录进行分类的统计方法。 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。 同时支持二项模型（用于具有两种离散类别的目标）和多项式模型（用于具有两种以上类别的目标）。

Logistic 回归的工作原理是构建一组方程式，使输入字段值与每个输入字段类别所关联的概率相关。 生成模型后，便可以用它来估计新数据的概率。 对于每条记录，将计算每种可能输出类别的成员资格概率。 具有最高概率的目标类别将被指定为该记录的预测输出值。

二项模型示例。 某个电信服务提供商关心流失到竞争对手那里的客户数。 使用服务利用率数据，可以创建二项模型以预测哪些客户有可能转向其他提供商，并定制服务以保留尽可能多的客户。 由于目标具有两个不同的类别（可能转移或不转移），因此使用了二项模型。

注: 字符串字段的长度限制为 8 个字符（仅适用于二项模型）。 如有必要，可以使用“重新分类”节点或使用“匿名化”节点对较长的字符串进行重新编码。

多项示例。 电信提供商按照服务用途模式划分客户群，将客户分类成四组。 通过使用人口统计数据来预测组成员资格，您可以创建多项模型，以将潜在客户归入不同的组，然后为个别客户定制产品。

需求。 一个或多个输入字段和唯一一个具有两个或多个类别的分类目标字段。 对于二项模型，目标必须具有标志测量级别。 于多项式模型，目标可以具有标志，或名义的测量级别，以及两个或多个类别。 设置为双向或 无的字段将忽略。 必须对模型中使用的字段的类型完全实例化。

强度。 通常，Logistic 回归模型非常准确。 它们可处理符号和数字类型的输入字段。 它们可以给出所有目标类别的预测概率，从而能够轻松识别出第二最佳推测值。 当组成员资格是真正分类字段时，Logistic 模型最为有效；如果组成员资格基于连续范围字段的值（例如，高 IQ 与低 IQ），那么应考虑使用线性回归，以利用整个范围的值所提供的更丰富的信息。 Logistic 模型还可以执行自动字段选择，但其他方法（例如树模型或特征选择）在对大型数据集执行此操作时可能速度更快。 最后，由于 Logistic 模型被很多分析人员和数据挖掘人员所熟知，因此他们可能会将其用作比较其他建模技术的基准。

处理大型数据集时，可以禁用高级输出选项似然比检验，从而显著提高性能。 有关更多信息，请参阅主题第 136 页的『Logistic 回归高级输出』。

要点: 如果临时磁盘空间较少，二项 Logistic 回归可能无法构建，并会显示错误。 当根据大型数据集（10GB 或更多）进行构建时，需要相同的可用磁盘空间量。 您可以使用环境变量 SPSSTMPDIR 来设置临时目录的位置。

Logistic 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

过程。 指定将创建二项模型还是多项式模型。 对话框中提供的选项会因所选建模过程的类型而异。

- **二项式。** 当目标字段是具有两个离散（二分）值（如是/否、启动/关闭或男/女）的标志或名义字段时使用。
- **多项式。** 当目标字段是具有两个以上值的名义字段时，应使用此选项。 可以指定**主效应**、**全析因**或**定制**。

在方程中包含常量。 此选项用于确定生成的方程中是否将包含常数项。 在大多数情况下，应将此选项保持为选中状态。

二项模型

对于二项模型，可用的方法和选项如下：

方法。 指定构建 Logistic 回归模型时要使用的方法。

- **进入法**。 这是缺省方法，用于将所有项直接输入方程中。 构建模型时不进行字段选择。
- **向前步进法**。 顾名思义，字段选择向前步进法用于分步构建方程。 初始模型是可能的最简单模型，其方程式中不含任何模型项（除常量外）。 在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，那么将该项添加到模型中。 此外，还会重新评估当前包含在模型中的项，以确定能否在不对模型功能造成重大减损的情况下删除其中任何项。 如果可以，那么会将其删除。 然后重复此过程，添加并/或删除其他项。 当无法再添加任何项来改进模型、且无法再删除任何项而不对模型功能造成减损时，最终模型便已生成。
- **向后步进法**。 向后步进法与向前步进法在本质上是相反的。 采用这种方法时，初始模型将包含作为预测变量的所有项。 每个步骤会评估模型中的项，并且将可以删除而不对模型功能造成重大减损的项删除。 此外，还会对先前删除的项进行重新评估，以确定其中的最佳项是否对模型的预测功能起到显著作用。 如果是，那么会将其重新添加到模型中。 当无法再删除任何项而不对模型功能造成重大减损、且无法再添加任何项以改进模型时，最终模型便已生成。

分类输入。 列出标识为分类字段的字段，即具有标志、名义或有序的测量级别。 可以为每个分类字段指定对比和基准类别。

- **字段名称**。 此列包含分类输入的字段名称。 要向此列中添加连续输入字段或数值输入字段，请单击列表右边的“添加字段”图标，然后选择所需输入字段。
- **对比**。 分类字段的回归系数的解释取决于使用的对比。 对比决定如何设定假设检验以比较估计平均值。 例如，如果已知某个分类字段具有隐含顺序（如模式或分组），那么可以使用对比为该顺序建模。 可用的对比如下：

  指示符。 这些对比指示类别成员资格是否存在。 这是缺省方法。

  简式。 将预测变量字段的每个类别（参考类别除外）与参考类别进行比较。

  差分。 将预测变量字段的每个类别（第一个类别除外）与先前类别的平均效果进行比较。 也称为逆 Helmert 对比。

  Helmert。 将预测变量字段的每个类别（最后一个类别除外）与后续类别的平均效果进行比较。

  重复。 将预测变量字段的每个类别（第一个类别除外）与前一个类别进行比较。

  多项式。 正交多项式对比。 假设类别均匀分布。 多项式对比仅适用于数字字段。

  偏差。 将预测变量字段的每个类别（参考类别除外）与总体效果进行比较。
- **基本类别**。 指定如何针对选定的对比类型确定参考类别。 选择 **第一个** 以使用输入字段的第一个类别（按字母顺序排列），或选择 **最后一个** 以使用最后一个类别。 缺省基本类别适用于**分类输入**区域中列出的变量。

  注: 如果对比设置为“差分”、Helmert、“重复”或“多项式”，那么此字段不可用。

每个字段对整体响应影响的估计，可以计算为其他各个类别相对于参考类别的似然增量或减量。 这有助于确定比较有可能给出特定响应的字段和值。

基准类别在输出中显示为 0.0。 这是因为将其与自己进行比较会产生空的结果。 所有其他类别均显示为与基准类别相关的方程式。 有关更多信息，请参阅主题 第 138 页的『Logistic 模型块详细信息』。

多项式模型

对于多项式模型，可用的方法和选项如下：

方法。 指定构建 Logistic 回归模型时要使用的方法。

- **进入法**。 这是缺省方法，用于将所有项直接输入方程中。 构建模型时不进行字段选择。

- **步进法** 顾名思义，字段选择步进法用于分步构建方程。 初始模型是可能的最简单模型，其方程式中不含任何模型项（除常量外）。 在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，那么将该项添加到模型中。 此外，还会重新评估当前包含在模型中的项，以确定能否在不对模型功能造成重大减损的情况下删除其中任何项。 如果可以，那么会将其删除。 然后重复此过程，添加并/或删除其他项。 当无法再添加任何项来改进模型、且无法再删除任何项而不对模型功能造成减损时，最终模型便已生成。
- **前进**。 字段选择向前法与分步构建模型的步进法类似。 但采用这种方法时，初始模型是最简单的模型，只能向模型中添加常量和项。 每个步骤会对尚未纳入到模型中的项进行检验，看它们对模型的改进起多大作用，然后将其中的最佳项添加到模型中。 当无法再添加任何项、或最佳备选项无法对模型产生足够的改进时，最终模型便已生成。
- **后退**。 向后法与向前法在本质上是相反的。 但采用这种方法时，初始模型包含作为预测变量的所有项，只能从模型中删除项。 对模型影响较小的模型项将被逐一删除，直到无法再删除任何项而不对模型功能造成重大损害，从而生成最终模型。
- **向后步进法**。 向后步进法与步进法在本质上是相反的。 采用这种方法时，初始模型将包含作为预测变量的所有项。 每个步骤会评估模型中的项，并且将可以删除而不对模型功能造成重大减损的项删除。 此外，还会对先前删除的项进行重新评估，以确定其中的最佳项是否对模型的预测功能起到显著作用。 如果是，那么会将其重新添加到模型中。 当无法再删除任何项而不对模型功能造成重大减损、且无法再添加任何项以改进模型时，最终模型便已生成。

注: 自动方法（包括步进法、向前步进法和先后步进法）是适应性强的学习方法，并且特别容易过度拟合训练数据。 使用这些方法时，用新数据或使用分区节点创建的保留测试样本对结果模型的有效性进行验证尤为重要。

目标的基本类别。 指定如何确定参考类别。 这将用作对目标中所有其他类别的回归方程式进行估计的基准。 选择 **第一个** 以使用当前目标字段的第一个类别（按字母顺序排列），或选择 **最后一个** 以使用最后一个类别。 或者，可以选择 **指定** 以选择特定类别，并从列表中选择所需的值。 可以在类型节点中为每个字段定义可用值。

通常应将关注程度最低的类别指定为基准类别，例如低价促销产品。 然后再以相对方式将其他类别与该基准类别相关，从而确定什么使它们更有可能自成类别。 这有助于确定比较有可能给出特定响应的字段和值。

基准类别在输出中显示为 0.0。 这是因为将其与自己进行比较会产生空的结果。 所有其他类别均显示为与基准类别相关的方程式。 有关更多信息，请参阅主题 第 138 页的『Logistic 模型块详细信息』。

模型类型。 有三个选项用于定义模型中的项。 **主效应**模型仅包括各个输入字段，而不检验输入字段之间的交互（乘法效应）。 **全因子**模型包括所有交互以及输入字段主效应。 全析因模型捕获复杂关系的能力较强，但也比较难以解释，而且更有可能出现过度拟合情况。 由于有可能出现大量可能组合，因此对于全析因模型，自动字段选择方法（进入法以外的方法）处于禁用状态。 **定制**模型仅包括您指定的项（主效应和交互效应）。 选择此选项时，应使用“模型项”列表在模型中添加或删除项。

模型项。 构建定制模型时，将需要明确指定模型中的项。 此列表显示了模型项的当前集合。 “模型项”列表右边的按钮用于添加和移除模型项。

- 要将项添加到模型中，请单击 添加新的模型项按钮。
- 要删除项，请选定所需项，然后单击 删除选定模型项按钮。

将项添加到 Logistic 回归模型

请求定制 Logistic 回归模型时，可以通过单击“Logistic 回归模型”选项卡中的添加新的模型项按钮将项添加到模型中。 这将打开“新建项”对话框，您可在其中指定项。

要添加的术语的类型。 有几种将项添加到模型的方法，具体取决于在“可用字段”列表中对输入字段的选择。

- **单向交互效应**。 插入表示所有选定字段的交互的项。
- **主效应**。 针对每个选定的输入字段插入一个主效应项（该字段本身）。
- **所有双向交互效应**。 针对每个可能的选定输入字段对插入一个双向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B* 和 *C*，此方法将插入项 *A * B*、*A * C* 和 *B * C*。
- **所有三向交互效应**。 针对每个可能的选定输入字段组合（一次取三个字段）插入一个三向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B*、*C* 和 *D*，此方法将插入项 *A * B * C*、*A * B * D*、*A * C * D* 和 *B * C * D*。

- **所有四向交互效应**。 针对每个可能的选定输入字段组合（一次取四个字段）插入一个四向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B*、*C*、*D* 和 *E*，此方法将插入项 *A * B * C * D*、*A * B * C * E*、*A * B * D * E*、*A * C * D * E* 和 *B * C * D * E*。

可用字段。 列出要在构造模型项时使用的可用输入字段。

预览。 根据上述所选字段和项类型，显示单击**插入**时将添加到模型中的项。

Insert 键。 将项插入模型（根据当前选择的字段和项类型）并关闭对话框。

Logistic 节点专家选项

如果您对 Logistic 回归有详细了解，那么可以通过专家选项对训练过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

尺度（仅限多项式模型）。 您可以指定将用于更正参数协方差矩阵的估计值的离差尺度值。 **Pearson** 使用 Pearson 卡方统计量来估算尺度值。 **偏差**使用偏差函数（似然比卡方）统计量来估算尺度值。 您也可以指定自己的用户定义尺度值。 必须是正数值。

附加所有概率。 如果选中此选项，那么会将输出字段的每个类别的概率添加到节点所处理的每条记录中。 如果未选中此选项，那么仅添加预测类别的概率。

例如，包含具有三个类别的多项式模型结果的表将包括五个新列。 一个列将列出预测正确的结果的概率，第二个列将显示该预测准确或失误的概率，第三个列将显示每个类别的预测失误或准确的概率。 有关更多信息，请参阅主题 第 138 页的『Logistic 模型块』。

注：对于二项模型，此选项始终处于选中状态。

奇异性容差。 指定检查奇异值时使用的容差。

收敛。 通过这些选项，您可以控制用于模型收敛的参数。 当您执行模型时，收敛设置将控制重复运行不同参数以观察其拟合程度的次数。 参数的尝试次数越多，结果将越接近（即，结果将会收敛）。 有关更多信息，请参阅主题 第 136 页的『Logistic 回归收敛选项』。

输出。 通过这些选项，可以请求将出现在由节点构建的模型块的高级输出中的附加统计量。 有关更多信息，请参阅主题 第 136 页的『Logistic 回归高级输出』。

步进。 通过这些选项，您可以控制采用步进、向前、向后或向后步进估计法添加和移除字段的标准。 （如果已选择进入法，该按钮将处于禁用状态。） 有关更多信息，请参阅主题 第 137 页的『Logistic 回归步进选项』。

Logistic 回归收敛选项

您可设置用于 Logistic 回归模型估计的收敛参数。

最大迭代次数。 指定用于估算模型的最大迭代次数。

最大逐步二分法。 逐步二分法是 Logistic 回归用于处理估算过程的复杂性的方法。 在通常情况下，应使用缺省设置。

对数似然收敛。 如果对数似然的相对变化小于此值，那么迭代将停止。 如果值为 0，则不使用该标准。

参数收敛。 如果参数估计中的绝对变化或相对变化小于此值，那么迭代将停止。 如果值为 0，则不使用该标准。

变化值（仅限多项式模型）。 您可以指定要添加到每个空单元格（输入字段和输出字段值的组合）中的值，该值介于 0 和 1 之间。 当相对于数据中的记录数有许多可能的字段值组合时，此值有助于估计算法处理数据。 缺省值为 0。

Logistic 回归高级输出

选择要在回归模型块的高级输出中显示的可选输出。 要查看高级输出，请浏览模型块并单击 **高级** 选项卡。 有关更多信息，请参阅主题 第 139 页的『Logistic 模型块高级输出』。

二项式选项

选择要为模型生成的输出的类型。 有关更多信息，请参阅主题 第 139 页的『Logistic 模型块高级输出』。

显示。 选择是在每个步骤中显示结果还是等到所有步骤已完成时再显示结果。

Exp(B) 的 CI。 选择表达式中每个系数（显示为 Beta）的置信区间。 指定置信区间的水平（缺省值为 95%）。

残差诊断。 请求残差的“观测值诊断”表。

- **外部离群值（标准 差）**。 仅列出这样的残差观测值：所列变量的绝对标准化值至少与您指定的值一样大。 缺省值为 2。
- **全部个案**。 在残差的“观测值诊断”表中包括所有观测值。

  注：由于此选项将列出每条输入记录，因此可能在报告中生成异常巨大的表，其中每条记录占一行。

分类分界值。 此选项可用于确定对观测值进行分类的分割点。 具有大于分类分界值的预测值的个案被分类为正，具有小于分类分界值的预测值的个案分类为负。 要更改缺省值，请输入一个 0.01 到 0.99 之间的值。

多项式选项

选择要为模型生成的输出的类型。 有关更多信息，请参阅主题 第 139 页的『Logistic 模型块高级输出』。

注：选择**似然比检验**选项将极大地增加构建 Logistic 回归模型所需的处理时间。 如果模型构建时间过长，可以考虑禁用此选项，或利用 Wald 统计量和评分统计量。 有关更多信息，请参阅主题 第 137 页的『Logistic 回归步进选项』。

每项的迭代历史记录。 选择在高级输出中打印迭代状态的分步间隔。

置信区间。 方程中系数的置信区间。 指定置信区间的水平（缺省值为 95%）。

Logistic 回归步进选项

通过这些选项，您可以控制采用步进、向前、向后或向后步进估计法添加和移除字段的标准。

模型中的项数（仅限多项式模型）。 您可以指定模型中的最小项数（针对向后法和向后步进法模型）和最大项数（针对向前法和步进法模型）。 如果指定大于 0 的最小值，模型将包括该数量的项，即使根据统计标准应将其中某些项删除也是如此。 对于前进法、逐步法和进入法模型，将忽略最小值设置。 如果指定最大值，可能会省略模型中的某些项，即使根据统计标准应将其选中也是如此。 对于后退法、后退逐步法和进入法模型，将忽略 **指定最大值** 设置。

纳入标准（仅限多项式模型）。 选择**评分**可以最大程度地提高处理速度。 **似然比** 选项可能会稍微多提供一些有力的估计值，但所需的计算时间较长。 缺省设置是使用评分统计量。

剔除标准。 选择**似然比**可以获得更稳健的模型。 要缩短构建模型所需的时间，可以尝试选择 **Wald** 。 但是，如果数据中有完全或半完全分隔（可使用模型块的“高级”选项卡确定），Wald 统计量将变得极不可靠，不应采用。 缺省设置是使用似然比统计量。 对于二项模型，还有附加选项 **条件** 。 此选项提供以基于条件参数估计值的似然比统计量的概率为依据的移除检验。

标准的显著性阈值。 通过此选项，您可以根据与每个字段相关联的统计概率（*p* 值）来指定选择标准。 仅当关联的 *p* 值小于 **纳入标准** 值时，才会将字段添加到模型中；仅当 *p* 值大于 **剔出标准** 值时，才会将字段删除。 **纳入标准** 值必须小于 **剔出标准** 值。

纳入或剔除的需求（仅限多项式模型）。 对于某些应用程序，除非模型也包含交互项所涉及字段的低阶项，否则将交互项添加到模型中在数学上没有意义。 例如，除非 *A* 和 *B* 也纳入到模型中，否则将 *A * B* 纳入到模型中没有意义。 使用这些选项，可以确定如何在逐步模型项选择过程中处理这些依赖关系。

- **用于离散效应的层次**。 仅当相关字段的所有低阶效应（涉及较少字段的主效应或交互）均位于模型中时，高阶效应（涉及较多字段的交互）才会进入模型，并且如果涉及相同字段的高阶效应位于模型中，那么将不会删除低阶效应。 此选项仅适用于分类字段。
- **用于所有效应的层次**。 此选项的工作原理与上一选项相同，但它适用于所有输入字段。
- **用于所有效应的包含**。 仅当效应中包含的所有效应也纳入到模型中时，该效应才能纳入到模型中。 此选项与**用于所有效果的层次**选项类似，只是连续字段的处理方式略有不同。 要让一个效应包含另一个效应，被包含（低阶）效应必须包括包含（高阶）效应中涉及的所有连续字段，且被包含效应的分类字段必须是包含效应中离散字段的子集。 例如，如果 *A* 和 *B* 是分类字段，*X* 是连续字段，那么项 *A * B * X* 将包含项 *A * X* 和 *B * X*。
- **无**。 没有任何强制关系；模型中项的添加和删除是独立的。

Logistic 模型块

Logistic 模型块表示由 Logistic 节点估计的方程式。 其中包含 Logistic 回归模型捕获的所有信息，以及有关模型结构和性能的信息。 这种类型的方程式也可以通过其他模型（如 Oracle SVM）生成。

运行包含 Logistic 模型块的流时，该节点将添加两个包含模型预测和相关概率的新字段。 新字段的名称派生自所预测的输出字段的名称，并带有表示预测类别的前缀 *$L-* 或表示相关概率的前缀 *$LP-*。 例如，对于名为 *colorpref* 的输出字段，新字段将命名为 *$L-colorpref* 和 *$LP-colorpref*。 此外，如果在 Logistic 节点中选中了 **追加所有概率** 选项，那么会针对输出字段的每个类别添加一个附加字段，其中包含属于每条记录对应类别的概率。 这些附加字段根据输出字段的值命名，以 *$LP-* 为前缀。例如，如果 *colorpref* 的合法值为 *Red*、*Green* 和 *Blue*，那么将添加三个新的字段：*$LP-Red*、*$LP-Green* 和 *$LP-Blue*。

正在生成“过滤”节点。 通过“生成”菜单，您可以创建新的“过滤”节点，以根据模型结果传递输入字段。 因多重共线性而从模型中删除的字段以及模型中未使用的字段将被生成的节点过滤。

Logistic 模型块详细信息

对于多项式模型，Logistic 模型块的“模型”选项卡采用分屏显示，模型方程显示在左侧窗格中，而预测变量重要性显示在右侧窗格中。 对于二项模型，此选项卡只显示预测变量重要性。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

模型方程式

对于多项式模型，左窗格显示为 logistic 回归模型估计的实际方程式。 在目标字段中，除基准类别之外，每种类别均有一个方程式。 这些方程式以树格式显示。 这种类型的方程式也可以通过某些其他模型（如 Oracle SVM）生成。

方程式。 显示用于在给定一组预测变量值的情况下推导出目标类别概率的回归方程。 目标字段的最后一个类别将被视为 **基准类别**；显示的方程式将针对一组特定预测变量值给出其他类别相对于基准类别的对数优势比。 给定预测变量模式的每个类别的预测概率根据这些对数优势比值推导得出。

如何计算概率

每个方程式会计算一个特定目标类别相对于基准类别的对数优势比。 **对数优势比**（也称为**分对数**）是指定目标类别相对于基准类别的概率比，并对结果取自然对数。 对于基准类别，该类别相对于自身的几率为 1.0，因此对数优势比为 0。 您可以将其视为所有系数均为 0 的基准类别的隐式方程式。

要根据特定目标类别的对数优势比推导出概率，需要取该类别的方程式计算的分对数值，并应用以下公式：

$P(group_i) = \exp(g_i) / \Sigma_k \exp(g_k)$

其中 *g* 是计算的对数优势比， *i* 是类别参考号， *k* 为 1 至目标类别数之间的数字。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

注：与其他类型的模型相比，计算 Logistic 回归的预测变量重要性可能需要更长时间，因此缺省情况下在“分析”选项卡中未选中预测变量重要性。 选中该选项可能会降低性能，对大数据集尤为明显。

Logistic 模型块概要

Logistic 回归模型的概要显示用于生成该模型的字段和设置。 此外，如果已执行附加到该建模节点的分析节点，那么还会在此部分显示该分析中的信息。 有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』。

Logistic 模型块设置

Logistic 模型块中的“设置”选项卡用于指定模型评分过程中的置信度、概率、倾向评分和 SQL 生成选项。 该选项卡仅在已将模型块添加到流中之后才可用，而且可以根据模型和目标的类型显示不同选项。

多项式模型

对于多项式模型，可用的选项如下。

计算置信度：指定是否在评分期间计算置信度。

计算原始倾向评分（仅限标志目标）：（仅限于具有标志目标的模型）您可以请求生成原始倾向评分，这些评分指示对目标字段指定的 `true` 结果的似然值。 除此之外，标准预测及置信度值也是如此。 调整后的倾向评分不可用。 有关更多信息，请参阅主题 第 25 页的『建模节点分析选项』。

追加所有概率：指定是否将输出字段每个类别的概率添加到该节点所处理的每条记录。 如果未选中此选项，那么仅添加预测类别的概率。 例如，对于具有三种类别的名义目标，评分输出针对三种类别的每一种都包含一列，并包含指示任何时候预测类别的概率的第四列。 例如，如果类别 红色、 绿色和 蓝色的概率分别是 0.6、0.3 和 0.1，那么预测类将为 红色，其中概率为 0.6。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **通过转换至本机 SQL 来进行评分**：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分。

  注: 虽然该选项可以更快速获得结果，但是本机 SQL 的大小和复杂性会随着模型复杂性的增加而增加。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

注: 对于多项式模型，如果已选中**追加所有概率**，那么 SQL 生成不可用；或者，对于具有名义目标的模型，如果已选中**计算置信度**，那么 SQL 生成不可用。 仅仅对具有标志目标的多项式模型，支持具有置信度计算的 SQL 生成。 SQL 生成不适用于二项模型。

二项模型

对于二项模型，置信度和概率始终处于启用状态，并且用于禁用这些选项的设置不可用。 SQL 生成不适用于二项模型。 对于二项模型，唯一可以更改的设置是计算原始倾向评分的功能。 正如以上针对多项式模型的说明，此内容适用于只具有标志目标的模型。 有关更多信息，请参阅主题 第 25 页的『建模节点分析选项』。

Logistic 模型块高级输出

Logistic 回归（也称为 **名义回归** ）的高级输出将提供有关估计模型及其性能的详细信息。 高级输出包含的大部分信息技术含量很高，需要具备 Logistic 回归分析方面的广泛知识才能正确理解该输出。

警告。 指示结果中存在的任何警告或潜在问题。

案例处理摘要。 列出由模型中的每个符号字段处理和细分的记录数。

步骤摘要（可选）。 列出使用自动字段选择时在模型创建过程的每个步骤中添加或移除的效应。

注：仅针对步进法、向前法、向后法或向后步进法显示此选项。

迭代历史记录（可选）。 显示从初始估计值开始每 *n* 次迭代的参数估计的迭代历史记录，其中 *n* 是打印间隔值。 缺省设置是打印每次迭代 (n=1)。

模型拟合信息（多项模型）。 显示根据其中所有参数系数均为 0（仅有截距）的模型对模型（最终模型）进行的似然比检验。

分类（可选）。 显示输出字段预测值和实际值的百分比矩阵。

拟合优度卡方统计（可选）。 显示 Pearson 和似然比卡方统计量。 这些统计量可检验模型对训练数据的总体拟合度。

Hosmer 和 Lemeshow 拟合优度（可选）。 显示将观测值分组为风险的十分位数并对每个十分位数中的观测概率与预期概率进行比较的结果。 此拟合优度统计量比多项式模型中采用的传统拟合优度统计量更为稳健，尤其适用于具有连续协变量的模型和小样本的研究。

伪 R 方（可选）。 显示模型拟合度的 Cox 和 Snell、Nagelkerke 以及 McFadden *R* 平方度量。 这些统计量在某些方面与线性回归中的 *R* 平方统计量类似。

单调性测量（可选）。 显示数据中一致对、不一致对和约束对的数目，以及每类占总对数的百分比。 此表中还显示 Somers 的 D、Goodman 和 Kruskal 的伽玛、Kendall 的 tau-a 以及协调索引 C。

信息准则（可选）。 显示赤池信息准则 (AIC) 和 Schwarz 贝叶斯信息准则 (BIC)。

似然比检验（可选）。 显示统计检验，表明模型效应的系数是否在统计意义上不同于 0。 重要输入字段是那些在输出中具有非常低的显著性水平的字段（标注为 *Sig.*）。

参数估计值（可选）。 显示方程系数的估计值、这些系数的检验、派生自标注为 *Exp(B)* 的系数的几率比及其置信区间。

渐近协方差/相关矩阵（可选）。 显示系数估计值的渐近协方差和/或相关性。

观测频率和预测频率（可选）。 对于每个协变量模式，显示每个输出字段值的显示观测频率和预测频率。 此表可能很大，对于具有数字输入字段的模型来说尤其如此。 如果结果表过大而无法应用，那么将省略该表，并显示一条警告。

主成分分析/因子节点

“PCA/因子”节点提供用于降低数据复杂程度的强大数据降维技术。 该技术提供以下两种相似但不同的方法。

- **主成分分析 (PCA)** 可以找出输入字段的线性组合，这些组合能够出色地捕获整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。 主成分分析集中关注所有方差，包括共享方差和独有方差。
- **因子分析**尝试确定可以解释一组观测字段中的相关性模式的基本概念（即**因子**）。 因子分析只集中关注共享方差。 估计模型时不考虑特定字段独有的方差。 因子/主成分分析节点提供几种因子分析方法。

这两种方式的目标都是找到有效概括原始字段集中的信息的少量派生字段。

需求。 主成分分析因子模型中只能使用数值字段。 要估计因子分析或主成分分析，需要一个或多个角色设置为输入字段的字段。 角色设置为目标、双向或无的字段将被忽略，就像对待非数值字段一样。

强度。 因子分析和 PCA 可以在不牺牲太多信息内容的情况下有效地降低数据的复杂性。 这些技术可帮助您构建更稳健的模型，并实现比原始输入字段更高的执行速度。

主成分分析/因子节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

抽取方法。 指定要用于数据降维的方法。

- **主成分**。 这是缺省方法，它将使用 PCA 来查找对输入字段进行汇总的成分。
- **未加权最小二乘法**。 此因子分析方法的工作原理是找出最能重现输入字段之间的关系（相关性）模式的因子集合。
- **广义最小二乘法**。 此因子分析方法与未加权最小二乘法类似，区别在于它利用加权降低具有大量独有（非共享）方差的字段的重要程度。
- **最大似然**。 这种因子分析方法基于对这些关系形式的假设，生成最有可能在输入字段中产生观察到的关系模式（相关性）的因子方程。 具体而言，此方法假定训练数据遵循多变量正态分布。
- **主轴因式分解**。 此因子分析方法与主成分法十分类似，区别在于它仅侧重于共享方差。
- **Alpha 因式分解**。 此因子分析方法将分析中的字段视为潜在输入字段范围内的样本。 它会将因子的统计可靠性最大化。
- **映像因式分解**。 此因子分析方法使用数据估计来隔离公共方差，并查找描述该方差的因子。

主成份分析（PCA）/因子节点专家选项

如果您对因子分析和 PCA 有详细了解，那么可以通过专家选项对训练过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

缺失值。 缺省情况下，IBM SPSS Modeler 仅使用对模型中所用的全部字段具有有效值的记录。 （这种方式有时称为缺失值的 **成列删除** 。） 如果有很多缺失数据，您可能会发现这种方式去除的记录过多，剩余记录不足以生成较好的模型。 在这种情况下，可以取消选中 **仅使用完整记录** 选项。 IBM SPSS Modeler 随后将尝试使用尽可能多的信息来估算模型，包括其中一些字段具有缺失值的记录。 （这种方式有时称为缺失值的 **成对删除** 。） 但在某些情形下，以这种方式使用不完整记录可能会在模型的估计过程中产生计算问题。

字段。 指定估算模型时是使用输入字段的相关性矩阵（缺省设置）还是使用其协方差矩阵。

收敛的最大迭代次数。 指定用于估算模型的最大迭代次数。

抽取因子。 选择要从输入字段中抽取的因子数的方法有两种。

- **特征值超出**。 此选项将保留特征值大于指定条件的所有因子或成分。 **特征值**用于度量每个因子或成分对输入字段集合中的方差进行汇总的能力。 使用相关性矩阵时，模型将保留特征值大于指定值的所有因子或成分。 使用协方差矩阵时，标准是指定的乘以平均特征值。 该尺度变换使此选项对于两种类型的矩阵具有类似的意义。
- **最大数目**。 此选项将保留指定数目的因子或成分，按特征值的降序排列。 换言之，将保留 *n* 个最高特征值所对应的因子或成分，其中 *n* 为指定标准。 缺省提取标准为五个因子/成分。

组件/因子矩阵格式。 这些选项用于控制因子矩阵（或 PCA 模型的成分矩阵）的格式。

- **对值进行排序**。 如果选中此选项，那么将按数字顺序对模型输出中的因子加载进行排序。
- **隐藏以下值**。 如果选中此选项，那么将在矩阵中隐藏小于指定阈值的评分，以便于查看矩阵中的模式。

旋转。 通过这些选项，您可以控制模型的旋转方法。 有关更多信息，请参阅主题 第 141 页的『主成分分析（PCA）/因子节点旋转选项』。

主成分分析（PCA）/因子节点旋转选项

许多情况下，对保留的因子集合进行数学旋转可提高其实用性，尤其可以降低其解释难度。 选择一种旋转方法：

- **无旋转**。 缺省选项。 不使用旋转。
- **最大方差法**。 这是可以将每个因子上负荷较高的字段的数目降至最低的正交旋转法。 它简化了因子的解释过程。
- **直接斜交旋转**。 一种斜交（非正交）旋转法。 当 **Delta** 等于 0（缺省值）时，解将采用斜交法。 Delta 越偏向负值，因子斜交度越小。 要覆盖缺省的 Delta 值 0，请输入小于或等于 0.8 的数字。
- **最大四次方值法**。 这是可以将解释每个字段所需的因子的数量降至最低的正交旋转法。 它简化了被观测字段的解释过程。
- **等量最大法**。 此旋转法结合了 Varimax 法与 Quartimax 法，前者用于简化因子，后者用于简化字段。 可将某个因子上载荷较高的字段数量和解释某个字段所需的因子数量降至最低。
- **最优斜交法**。 这是实现了因子关联的斜交旋转法。 它计算起来比斜交旋转更快，因此适用于大型数据集。 **Kappa** 用于控制解的倾斜度（因子相关的程度）。

主成分分析/因子模型块

主成分分析/因子模型块表示由主成分分析/因子节点创建的因子分析和主成分分析 (PCA) 模型。 其中包含被训练模型捕获的所有信息，以及有关模型性能和特征的信息。

当您运行包含因子方程模型的流时，节点会为模型中的每个因子或成分添加一个新字段。 新字段名称派生自模型名称并带有前缀和后缀，前缀为 *$F-*，而后缀为 *-n*，其中 *n* 是因子或成分的编号。 例如，如果模型名为 *Factor* 且包含三个因子，新字段将命名为 *$F-Factor-1*、 *$F-Factor-2* 和 *$F-Factor-3*。

为更好地了解因子模型的编码内容，可以进一步执行一些下游分析。 查看因子模型结果的一种实用方法是使用统计量节点查看因子与输入字段之间的相关性。 这种方法可显示哪些输入字段对哪些因子的载荷较重，并帮助您发现因子是否具有潜在的意义或解释。

您还可以使用高级输出中提供的信息对因子模型进行评估。 要查看高级输出，请单击模型块浏览器的 **高级** 选项卡。 高级输出包含大量详细信息，适合于在因子分析或主成分分析方面具有广泛知识的用户。 有关更多信息，请参阅主题 第 142 页的『主成分分析/因子模型块高级输出』。

主成分分析/因子模型块方程式

因子模型快的“模型”选项卡显示每个因子的因子得分方程式。 因子或成分的评分是通过将每个输入字段值乘以其系数并将结果相加计算得出的。

主成分分析/因子模型块概要

因子模型的“概要”选项卡显示因子/主成分分析模型中保留的因子数，以及有关用于生成模型的字段和设置的其他信息。 有关更多信息，请参阅主题 第 31 页的『浏览模型块』。

主成分分析/因子模型块高级输出

因子分析的高级输出提供有关所估计模型及其性能的详细信息。 高级输出中包含的大部分信息技术含量很高，需要具备因子分析方面的广泛知识才能正确理解该输出。

警告。 指示结果中存在的任何警告或潜在问题。

公因子方差。 显示因子或成分占每个字段的方差的比例。 初始指定具有整个因子集合（最初，模型的因子数与输入字段数相同）的初始公因子方差，提取指定基于保留因子集合的公因子方差。

解释的总方差。 显示由模型中的因子解释的总方差。 初始特征值显示由整个初始因子集合解释的方差。 提取平方和载入显示由模型中保留的因子解释的方差。 旋转平方和载入显示由旋转因子解释的方差。 请注意，对于斜交旋转法， 旋转加载平方和仅显示加载平方和，而不显示方差百分比。

因子（或成分）矩阵。 显示输入字段与未旋转因子之间的相关性。

旋转因子（或成分）矩阵。 显示输入字段与正交旋转的旋转因子之间的相关性。

模式矩阵。 显示输入字段与斜交旋转法的旋转因子之间的偏相关。

结构矩阵。 显示输入字段与斜交旋转法的旋转因子之间的简单相关性。

因子相关性矩阵。 显示斜交旋转法的因子之间的相关性。

判别节点

判别分析用于为组成员资格构建预测模型。 该模型中包含一个判别函数（或者，存在两个以上的组时，将会包含一组判别函数），后者基于在各个组之间提供最佳判别的预测变量的线性组合。 这些函数根据组成员身份已知的个案样本生成；然后，可以将这些函数应用于具有预测变量测量值，但具有未知组成员身份的新个案。

示例。 根据使用情况数据，电信公司可以使用判别分析来对用户进行分组。 此操作使电信公司可以对潜在的用户进行评分，并将最有可能属于最有价值的组的客户作为目标。

需求。 您需要一个或多个输入字段，且只需要一个目标字段。 目标必须为带有字符串或整数存储的分类字段（测量级别为标志或名义）。 （如有必要，可以使用“填充”或“派生”节点来转换存储。） 设置为 双向或无的字段将忽略。 必须对模型中使用的字段的类型完全实例化。

强度。 判别分析和 Logistic 回归都是合适的分类模型。 但是，判别分析会对输入字段进行更多的假设，例如，假设这些字段为正态分布且连续，当满足这些要求时它们能提供更好的结果，在样本量比较小时尤其如此。

判别节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

方法。 以下选项可用于向模型中输入预测变量：

- **进入法**。 这是缺省方法，用于将所有项直接输入方程中。 不能显著增加模型预测能力的项将不被添加。
- **步进法** 初始模型是可能的最简单模型，其方程式中不含任何模型项（除常量外）。 在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，那么将该项添加到模型中。

注：步进法特别容易过度拟合训练数据。 当使用这些方法时，用保留测试样本或新数据对结果模型的有效性进行验证尤其重要。

判别节点专家选项

如果对判别分析有详尽了解，可用专家选项调整训练过程。 要访问专家选项，请在“专家”选项卡中将**模式**设置为**专家**。

先验概率。 此选项确定对于组成员身份的先验知识，是否调整分类系数。

- **所有组均相等**。 假设所有组的先验概率相等；这对系数没有影响。
- **根据组大小计算**。 样本中的观察组大小决定组成员身份的先验概率。 例如，如果分析中包括的 50% 的观测值属于第一组，25% 属于第二组，25% 属于第三组，那么会调整分类系数以增加第一组相对于其他两组的成员身份可能性。

使用协方差矩阵。 您可用选择使用组内协方差矩阵或独立组协方差矩阵对个案进行分类。

- 组内。汇聚的组内协方差矩阵用来对观测值分类。
- 独立组。分组协方差矩阵用于分类。 由于分类基于判别函数（而不是基于原始变量），因此该选项并不始终等同于二次判别。

输出。 通过这些选项，可以请求将出现在由节点构建的模型块的高级输出中的附加统计量。 有关更多信息，请参阅主题 第 143 页的『判别节点输出选项』。

步进。 通过这些选项，您可以使用步进估计法控制针对添加和删除字段的标准。 （如果已选择进入法，该按钮将处于禁用状态。） 有关更多信息，请参阅主题 第 144 页的『判别节点步进选项』。

判别节点输出选项

选择要在 Logistic 回归模型块的高级输出中显示的可选输出。 要查看高级输出，请浏览模型块并单击 **高级** 选项卡。 有关更多信息，请参阅主题 第 145 页的『判别分析模型块高级输出』。

描述性。 可用选项为均值（包括标准差）、单变量 ANOVA 以及 Box 的 *M* 检验。

- 均值 *(Means)*. 显示自变量的总均值、组均值以及标准差。
- 单变量 *ANOVA*. 为每个自变量的组均值的等同性执行单向方差检验分析。
- *Box's M*。一种用于检查组协变量矩阵是否相等的检验。 对于十分大的样本，不显著的 p 值表示矩阵不同的证据不足。 这个检验对于违反多变量常态的情况很敏感。

函数系数。 可用的选项有 Fisher 的分类系数和未标准化的系数。

- *Fisher (Fisher's)*. 显示可以直接用于分类的 Fisher 分类函数系数。 将为每个组获取一个单独的分类函数系数集，并且会对该组分配一个具有最大判别评分的观测值（分类函数值）。
- 未标准化 . 显示未标准化的判别函数系数。

矩阵。 可用的自变量系数矩阵有组内相关性矩阵、组内协方差矩阵、分组协方差矩阵和总体协方差矩阵。

- 组内相关性。显示合并组内协方差矩阵，该矩阵是通过在计算相关性之前对所有组的各个协方差矩阵求平均值得到的。
- 组内协方差。显示合并组内协方差矩阵，该矩阵可能与总协方差矩阵不同。 该矩阵是通过对所有组的各个协方差矩阵求平均值得到的。
- 分组协方差。显示每组的独立协方差矩阵。
- 总体协方差。显示所有观测值均来自一个样本时的协方差矩阵。

分类。 以下输出属于分类结果。

- 观测值结果。针对每个观测值显示的实际组、预测组、后验概率和判别评分的代码。
- 摘要表。基于判别分析正确或不正确分配给每组数据的个案数。 有时也称为“混淆矩阵”。
- 留一分类。分析中的每个个案均通过从该个案以外的所有其他个案衍生的函数进行分类。 也称为“U 方法”。
- 区域图。用于根据变量值对个案进行分组的边界的图。 与观测值被划分到的组相对应的编号。 在每个组的边界内使用星号标记该组的均数。 当只有一个判别函数时不显示此图。
- 组合组。创建前两个判别函数值的全组散点图。 如果只有一个函数，则会显示一个直方图。
- 分组。创建前两个判别函数值的独立组散点图。 如果仅有一个函数，则显示直方图。

步进法 步骤摘要显示每个步骤后的所有变量的统计信息；**F 表示成对距离**，显示成对距离矩阵 *F* 比率表示每对组。 *F* 比可用于组之间马氏距离的显著性检验。

判别节点步进选项

方法。 选择用于输入或移去新变量的统计。 可用的替代方法是 Wilk lambda、未解释方差、Mahalanobis 距离、最小 *F* 比率以及 Rao *V*。 通过 Rao *V*，可以在 *V* 中指定要输入的变量的最小增量。

- *Wilks Lambda*。一种用于逐步判别分析的变量选择方法，该方法根据变量的威克斯 λ 下限的多少来选择要输入到方程中的变量。 在每一步中，输入使整体威尔克斯 λ 最小化的变量。
- 未解释方差。在每一步，均是输入能使组间未解释变动合计最小的变量。
- 马氏距离。自变量上观测值的值与所有观测值的平均值相差多少的测量值。 较大马氏距离表示某一观测值在一个或多个自变量上有极值。
- 最小 *F* 比。一种在逐步分析中选择变量的方法，该方法基于最大化从两组间的 Mahalanobis 距离计算得到的 F 比。
- *Rao V*。组均数之间差异的测量。 也称为 Lawley-Hotelling 跟踪。 每一步都需要输入最大化 Rao 的 V 中的增量的变量。 选择此选项后，输入用于分析的变量的最小值。

标准。 可用的替代方法是 **使用 F 值** 和 **使用 F 的概率**。 输入用于输入和除去变量的值。

- 使用 *F* 值。如果变量的 F 值大于 Entry 值，那么将该变量输入到模型中，如果 F 值小于 Removal 值，那么将其除去。 纳入值必须大于移除值，并且这两个值都必须为正数。 要将更多变量输入到模型中，需要减小纳入值。 要从模型中删除更多变量，需要增大移除值。
- 使用 *F* 的概率。如果变量 F 值的显着性水平小于 "输入" 值，那么会将该变量输入到模型中; 如果显着性水平大于 "剔除" 值，那么会将该变量移除。 纳入值必须小于移除值，并且这两个值都必须为正数。 要将更多变量输入到模型中，需要增大纳入值。 要将更多的变量从模型中移去，请降低“剔除”值。

判别分析模型块

判别分析模型块表示由判别节点估计的方程式。 这些方程式包含由判别分析模型所捕获的所有信息及有关模型结构和性能的信息。

当运行包含判别分析模型块的流时，该节点可添加包含模型预测和相关概率的两个新字段。 新字段的名称派生自要预测的输出字段的名称，并带有表示预测类别的前缀 *$D-* 或表示相关概率的前缀 *$DP-*。 例如，对于名为 *colorpref* 的输出字段，会将新字段命名为 *$D-colorpref* 和 *$DP-colorpref*。

生成"过滤"节点。 使用“生成”菜单可以创建新的过滤节点，用于根据模型结果传递输入字段。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

判别分析模型块高级输出

判别分析的高级输出给出了有关估计模型及其性能的详细信息。 在高级输出中包含的多数信息具有很强的技术性，需要具有广泛的判别分析方面的知识才能够对此输出作出正确地解释。 有关更多信息，请参阅主题第 143 页的『判别节点输出选项』。

判别分析模型块设置

通过判别分析模型块中的“设置”选项卡，您可以在对模型进行评分时获取倾向评分。 此选项卡在只带有标志目标的模型中提供，并且仅在已将模型块添加到流中后可用。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

判别分析模型块汇总

判别分析模型块的“摘要”选项卡显示了用于生成模型的字段和设置。 此外，如果已执行附加到该建模节点的分析节点，那么还会在此部分显示该分析中的信息。 有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』。

GenLin 节点

“广义线性”模型对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。 而且，该模型还允许因变量呈非正态分布。 它涵盖广泛使用的统计模型，例如用于正态分布响应的线性回归、用于二分类数据的 Logistic 模型、用于计数数据的对数线性模型、用于间隔检查生存数据的互补双对数模型，以及许多其他通过其非常通用的模型规划的统计模型。

示例。 运输公司可以使用广义线性模型，对在不同期间建造的一些轮船类型的损坏统计采用泊松回归，其结果模型可帮助确定哪些轮船类型最容易损坏。

汽车保险公司可以使用广义线性模型，对汽车损坏理赔采用伽玛回归，其结果模型可帮助确定对理赔额度贡献最大的因素。

医学研究者可以使用广义线性模型对区间型删失的生存数据应用互补双对数回归，从而预测病理状况重现的时间。

广义线性模型的工作原理是构建一个方程式，从而使输入字段值与输出字段值关联起来。 生成模型后，便可以将其用于为新数据估值。 对于每条记录，将计算每种可能输出类别的成员资格概率。 具有最高概率的目标类别将被指定为该记录的预测输出值。

需求。 您需要一个或多个输入字段，同时有且仅有一个具有两个或多个类别的目标字段（其测量级别可以为连续或标记）。 必须对模型中使用的字段的类型完全实例化。

强度。 广义线性模型极为灵活，但选择模型结构的过程并未自动化，因此您需要对数据有一定的了解（这在“黑盒”算法中是不需要的）。

GenLin 节点字段选项

除建模节点的“字段”选项卡通常提供的目标、输入和分区定制选项外（请参阅第 23 页的『建模节点字段选项』），GenLin 节点还提供以下附加功能。

使用权重字段。 刻度参数是与响应方差相关的估计模型参数。 刻度权重是“已知”值，可能因观测值的不同而异。 如果指定了刻度权重变量，那么对每个观测值，都会用与响应方差相关的刻度参数除以该刻度权重变量。 分析中不会使用尺度权重值小于等于 0 或缺失的记录。

目标字段表示在一组试验中发生的事件数。 如果响应是一组试验中发生的事件数，那么目标字段将包含该事件数，并且您可选择包含试验次数的附加变量。 或者，如果试验数在所有主体中都相同，那么可以使用固定值指定试验。 试验数应大于等于每个记录的事件数。 事件应为非负整数，试验应为正整数。

GenLin 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

模型类型。 有两个选项用于要构造的模型类型。 **仅主效应** 使模型仅分别包括各个输入字段，而不检验输入字段之间的交互效应（乘法效应）。 **主效应和所有双向交互**包括所有双向交互以及输入字段主效应。

偏移量。 偏移量项是“结构性”预测变量。 它的系数不通过模型估计而假定其值为 1；因此，偏移量的值只与目标的线性预测变量简单相加。 这对于泊松回归模型尤其有用，在这种模型中，每个观测值对于相关事件可以具有不同的揭示级别。

例如，对各个驾驶员的事故率建模时，三年驾驶经验的驾驶员在一次事故中的过错率与 25 年驾驶经验的驾驶员在一次事故中的过错率存在重大差别。 如果将驾驶员的经验的自然对数作为偏移量项包括在内，那么可以按照泊松响应或负二项式响应对事故数进行建模。

其他分布和关联类型的组合将需要偏移变量的其他转换。

注：如果使用了变量偏移量字段，那么不应同时将指定字段用作输入。 如果需要，可在上游源节点或“类型”节点中将偏移量字段的角色设置为**无**。

标志目标的基本类别。

对于二元响应，可以为因变量选择参考类别。 这会影响某些输出，如参数估计值和保存的值，但不应更改模型拟合度。 例如，如果您的二元响应取值 0 和 1：

- 缺省情况下，此过程将最后一个值（最高值）或 1 作为参考类别。 在这种情况下，模型保存的概率估计给定个案取值 0 的概率，参数估计值应解释为与类别 0 的似然估计相关。
- 如果您指定第一个值（最低值）或 0 作为参考类别，那么模型保存的概率估计给定个案取值 1 的概率。
- 如果您指定定制值并且变量已定义了标签，那么可以通过从列表中选择值来设置参考类别。 在指定模型的过程中并不确定某一特定变量的编码方式时，这种方法非常方便。

在模型中包含截距。 截距通常包括在模型中。 如果您可以假设数据穿过原点，那么可以排除截距。

GenLin 节点专家选项

如果具备广义线性模型的深入知识，那么可以使用专家选项对训练过程进行微调。 要访问专家选项，请在“专家”选项卡中将**模式**设置为**专家**。

目标字段分布和关联函数

分布。

此选项指定因变量的分布。 指定非正态分布和非恒等关联函数的功能是广义线性模型相对一般线性模型的重要改进。 存在许多可能的分布-关联函数组合，其中有多个组合适用于任何给定的数据集，因此，您可根据先验理论考虑事项或哪个组合看起来拟合得最好来指导您的选择。

- **二项式**。 此分布仅适合表示二元响应或事件数量的变量。

- **伽玛**。 该分布适用于具有正刻度值并向更大的正值偏度的变量。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。
- **逆高斯**。 该分布适用于具有正刻度值并向更大的正值偏度的变量。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。
- **负二项式**。 该分布可以视为观察 *k* 成功所需的试验次数，适合具有非负整数值的变量。 如果数据值是非整数、小于 0 或缺失，那么分析中不会使用相应的个案。 负二项式分布辅助参数的固定值可以是任何大于或等于 0 的数字。 辅助参数设置为 0 时，使用该分布等效于使用泊松分布。
- **正态**。 该分布适合围绕某个中间值（平均值）呈对称钟型分布的标度变量。 因变量必须是数值型变量。
- **泊松**。 此分布可以被认为是在固定时间段内感兴趣的事件发生的次数，适用于具有非负整数值的变量。 如果数据值是非整数、小于 0 或缺失，那么分析中不会使用相应的个案。
- **Tweedie**。 此分布适合由伽玛分布泊松混合表示的变量；之所以称为?混合?分布，是因为它兼具连续（取非负实数值）和离散分布（在单个值 0 处为正概率质量）的属性。 因变量必须是数值型变量，数据值大于或等于零。 如果数据值小于零或缺失，那么分析中不会使用相应的个案。 Tweedie 分布参数的固定值可以是任何大于 1 且小于 2 的数字。
- **多项**。 此分布适合表示序数响应的变量。 因变量可以是数值或字符串，它必须至少有两个不同有效数据值。

链接函数。

关联函数是允许模型估计的因变量的转换。 有下列函数可用：

- **Identity.** $f(x)=x$。 不会对因变量进行转换。 此链接函数可用于任何分布。
- **Complementary log-log.** $f(x)=\log(-\log(1-x))$。 该函数只适用于二项分布。
- **累积概率**。 $f(x) = \tan(\pi (x - 0.5))$，用于响应每个类别的累积概率。 此函数仅适用于多项分布。
- **累积补充日志记录**。 $f(x)=\ln(-\ln(1-x))$，用于响应每个类别的累积概率。 此函数仅适用于多项分布。
- **累积逻辑**。 $f(x)=\ln(x / (1-x))$，用于响应每个类别的累积概率。 此函数仅适用于多项分布。
- **累积负面日志记录**。 $f(x)=-\ln(-\ln(x))$，用于响应每个类别的累积概率。 此函数仅适用于多项分布。
- **累积 probit**。 $f(x)=\Phi^{-1}(x)$，应用于响应的每个类别的累积概率，其中 Φ^{-1} 是逆标准正常累积分布函数。 此函数仅适用于多项分布。
- **Log.** $f(x)=\log(x)$。 此链接函数可用于任何分布。
- **Log complement.** $f(x)=\log(1-x)$。 该函数只适用于二项分布。
- **Logit.** $f(x)=\log(x / (1-x))$。 该函数只适用于二项分布。
- **负二项式**。 $f(x)=\log(x / (x+k^{-1}))$，其中 *k* 是负二项式分布的辅助参数。 此函数仅适用于负二项式分布。
- **Negative log-log.** $f(x)=-\log(-\log(x))$。 该函数只适用于二项分布。
- **赔率** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$，如果 $\alpha \neq 0$。 $f(x)=\log(x)$，如果 $\alpha=0$。 α 是必需的数字规范，而且必须是实数。 该函数只适用于二项分布。
- **概率值**。 $f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是标准正态累积分布的反函数。 该函数只适用于二项分布。
- **Power.** $f(x)=x^{\alpha}$, if $\alpha \neq 0$. $f(x)=\log(x)$，如果 $\alpha=0$。α 是必需的数字规范，必须是实数。 此链接函数可用于任何分布。

参数。 通过此组中的控件，可以在选中某些分布选项时指定参数值。

- **负二项式的参数**。 对于负二项式分布，选择以指定一个值或允许系统提供估计值。
- **Tweedie 的参数**。 对于 Tweedie 分布，给固定值指定在 1.0 与 2.0 之间的一个数字。

参数估计。 该组中的控件允许指定估计方法并提供参数估计值的初始值。

- **方法**。 可以选择参数估计方法。 可以选择的方法包括 Newton-Raphson、Fisher 评分方法以及先执行 Fisher 评分迭代再切换为 Newton-Raphson 方法的混合方法。 如果在混合方法的 Fisher 评分方法阶段期间，在达到 Fisher 迭代的最大次数之前实现了收敛性，那么算法将继续执行 Newton-Raphson 方法。
- **刻度参数方法**。 可以选择刻度参数估计方法。 最大似然法可联合估计刻度参数和模型效应；请注意，如果响应具有负二项式、泊松或二项式分布，那么此选项无效。 偏差和 Pearson 卡方选项从这些统计的值估计刻度参数。 或者，可以为刻度参数指定固定值。

- **协方差矩阵**。 基于模型的估计量是海森矩阵的广义逆负矩阵。 健壮性（也称为 Huber/White/sandwich）估计是“改正”的基于模型的估计量，即使错误地指定了方差和关联函数，也能提供对协方差的一致估计。

迭代。 通过这些选项，您可以控制用于模型收敛的参数。 有关更多信息，请参阅主题 第 148 页的『广义线性模型迭代』。

输出。 通过这些选项，可以请求将出现在由节点构建的模型块的高级输出中的附加统计量。 有关更多信息，请参阅主题 第 148 页的『广义线性模型高级输出』。

奇异性容差。 奇异（非可逆）矩阵具有线性相关列，对估计算法可能产生严重问题。 即使近似奇异的矩阵也可导致不良结果，因此该过程会将行列式小于容差的矩阵作为奇异矩阵对待。 指定一个正值。

广义线性模型迭代

您可设置用于对广义线性模型进行估计的收敛参数。

迭代。 可用选项有：

- **最大迭代次数**。 算法将执行的最大迭代次数。 指定非负整数。
- **最大逐步二分法**。 每次迭代时，步长都会减去因子 0.5，直到对数似然估计增加或者达到最大步骤对分。 请指定正整数。
- **检查数据点的分离**。 如果选择此项，算法将执行检验以确保参数估计值具有唯一值。 当过程可生成一个正确对每个个案进行分类的模型时，将发生分离。 此选项可用于二元格式的二项式响应 。

收敛条件。 下列选项可用：

- **参数估计变化**。 如果选择此项，算法将在参数估计值的绝对或相对更改小于指定值（必须为正值）的迭代之后停止。
- **对数似然变化**。 如果选择此项，算法将在对数似然估计函数的绝对或相对更改小于指定值（必须为正值）的迭代之后停止。
- **Hessian 收敛**。 对于“绝对值”指定，如果基于 Hessian 收敛性的统计小于指定的正值，那么假设收敛性。 对于“相对”指定，如果统计小于指定的正值和对数似然估计的绝对值的乘积，那么假设收敛性。

广义线性模型高级输出

选择要在广义线性模型块的高级输出中显示的可选输出。 要查看高级输出，请浏览模型块并单击 **高级** 选项卡。 有关更多信息，请参阅主题 第 149 页的『GenLin 模型块高级输出』。

可用的输出如下所示：

- **观测值处理摘要**。 显示分析和“相关数据摘要”表中包含的个案和从中排除的个案的数量和百分比。
- **描述统计**。 显示有关因变量、协变量和因子的描述统计和摘要信息。
- **模型信息**。 显示数据文件名称、因变量或事件和试验变量、偏移变量、刻度权重变量、概率分布和关联函数。
- **拟合度统计量**。 显示离差和刻度化离差、Pearson 卡方和刻度化 Pearson 卡方、对数似然估计、AIC 准则 (AIC)、有限样本校正 AIC (AICC)、BIC 准则 (BIC) 以及 CAIC 准则 (CAIC)。
- **模型汇总统计**。 显示模型拟合度检验，包括用于模型拟合度 Omnibus 检验的似然比统计，以及每种效应的类型 I 或 III 对比的统计。
- **参数估计**。 显示参数估计值和相应的检验统计和置信区间。 除了显示原始参数估计值之外，还可以选择显示取幂参数估计值。
- **参数估计值的协方差矩阵**。 显示估计参数协方差矩阵。
- **参数估计值的相关性矩阵**。 显示估计参数相关性矩阵。
- **对比系数 声明 矩阵**。 为缺省效应和估计边际均值显示对比系数（如果“EM 均值”选项卡上要求）。
- **常规可估计函数**。 显示生成对比系数 (L) 矩阵的矩阵。
- **迭代历史记录**。 显示参数估计值和对数似然估计的迭代历史记录，并打印对梯度向量和海森矩阵的最后一次评估。 迭代历史记录表从第 0 次迭代（初始估计值）开始每隔 *n* 次迭代就显示一次参数估计值，其中 *n* 代表打印区间的值。 如果请求显示迭代历史记录，那么无论 *n* 为多少都会显示最后一次迭代。

- **拉格朗日乘数检验**。 显示拉格朗日乘数检验统计量，用于为正态、伽玛和逆高斯分布评估使用离差或 Pearson 卡方计算或者设置为固定值的尺度参数的有效性。 对于负二项式分布，这种统计量会对固定辅助参数进行检验。

模型效应。 可用选项有：

- **分析类型**。 指定要生成的分析类型。 类型 I 分析通常适合于对模型中预测变量的排序具有先验理由的情况，而类型 III 应用范围更广。 根据在“卡方统计”组中的选择，计算 Wald 或似然比统计。
- **置信区间水平 (%)**。 指定大于 50 且小于 100 的置信水平。 Wald 区间基于参数呈渐近正态分布的假设之上；截面似然区间更加准确但需要进行大量的计算。 截面似然区间的容差水平可作为标准，用以停止在区间计算所采用的迭代算法。
- **对数似然函数**。 它控制对数似然函数的显示格式。 整个函数包含一个相对于参数估计恒定的额外项；它对参数估计没有任何影响，并且在某些软件产品中不会显示。

GenLin 模型块

GenLin 模型块表示由 GenLin 节点估计的方程式。 这些方程式包含由模型所捕获的所有信息及有关模型结构和性能的信息。

当您运行包含 GenLin 模型块的流时，该节点会添加一些新字段，这些字段的内容取决于目标字段的性质：

- **标记目标**。 添加包含预测类别和相关概率的字段，并为每个类别添加概率。 前两个新字段的名称派生自要预测的输出字段的名称，并带有表示预测类别的前缀 *$G-* 或表示相关概率的前缀 *$GP-*。 例如，对于名为 *default* 的输出字段，新字段将命名为 *$G-default* 和 *$GP-default*。 后两个附加字段是根据输出字段的值命名的，以 *$GP-* 为前缀。例如，如果 *default* 的合法值为 *Yes* 和 *No*，那么会将新字段命名为 *$GP-Yes* 和 *$GP-No*。
- **连续目标**。 添加包含预测平均值和标准误差的字段。
- **连续目标（表示一系列试验中的事件数）**。 添加包含预测平均值和标准误差的字段。
- **有序目标**。 为有序集合的每个值添加包含预测类别和相关概率的字段。 字段的名称派生自要预测的有序集合的值，并带有表示预测类别的前缀 *$G-* 或表示相关概率的前缀 *$GP-*。

正在生成"过滤"节点。 使用“生成”菜单可以创建新的过滤节点，用于根据模型结果传递输入字段。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

GenLin 模型块高级输出

广义线性模型的高级输出可提供有关估计模型及其性能的详细信息。 高级输出中包含的大部分信息的技术性含量都很高，需要进行此类分析所需的丰富知识才能够对此输出作出正确解释。 有关更多信息，请参阅主题 第 148 页的『广义线性模型高级输出』。

GenLin 模型块设置

使用 GenLin 模型块的“设置”选项卡，您可以在对模型进行评分时获取倾向评分，并且也适用于对模型进行评分期间生成 SQL。 此选项卡在只带有标志目标的模型中提供，并且仅在已将模型块添加到流中后可用。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

GenLin 模型块汇总

GenLin 模型块的“摘要”选项卡显示了用于生成模型的字段和设置。 此外，如果已执行附加到该建模节点的分析节点，那么还会在此部分显示该分析中的信息。 有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』。

广义线性混合模型

GLMM 节点

使用此节点可以创建广义线性混合模型 (GLMM)。

广义线性混合模型

广义线性混合模型扩展了线性模型，使得：

- 目标通过指定的关联函数与因子和协变量线性相关。
- 目标可以有非正态分布。
- 观测可能相关。

广义线性混合模型涵盖了各种模型，从简单线性回归模型到非正态纵向模型数据的复杂多级模型。

示例。 地区教育委员会可以使用广义线性混合模型来判定一种实验教学方法对于提高数学成绩的有效性。 由于来自同一班级的学生由同一个老师教导，因此这些学生是相关的，而且同一学校的各班级之间也是相关的，因此我们可以在学校和班级级别包含随机效应，以说明可变性的不同来源。

医学研究者可以用广义线性混合模型来判定一种新型抗惊厥药能否降低患者癫痫发作的比率。 对同一患者重复测量通常是正相关的，因此适合使用具有随机效应的混合模型。 目标字段为发作次数，其值为正整数，因此可能适合使用具有泊松分布和对数关联的广义线性混合模型。

电视、电话和网络服务的电缆供应商高管可以使用广义线性混合模型了解潜在的客户。 由于可能的回答具有名义测量级别，因此公司分析人员使用具有随机截距的广义 logit 混合模型，以确定在给定调查响应者回答中不同服务类型（电视、电话、互联网）的服务使用情况回答之间的相关性。

“数据结构”选项卡用于在观测值具有相关性时指定数据集中记录间的结构关系。 如果数据集中的记录代表独立的观测值，那么无需在此选项卡上指定任何设置。

主体。 指定分类字段的值组合应唯一定义数据集中的主体。 例如，单个患者标识字段应足以在一个医院内定义主体，但如果患者标识号在医院间不唯一，那么需要使用医院标识和患者标识的组合。 在重复测量设置中，将为每个主体记录多个观测值，因此每个主体可能在数据集中占用多条记录。

主体是可视为独立于其他主体的观测单元。 例如，在医学研究中，一位患者的血压读数可视为独立于其他患者的读数。 当每个主体都有重复测量，而又要对这些观测值的相关性建立模型，定义主体就变得尤为重要。 例如，您可能期望同一个患者在连续多次就医时得到的血压读数是相关的。

在“数据结构”选项卡上所有指定为**主体**的字段都用于定义残差协方差结构的主体，并提供可能字段的列表，用于定义随机效应块上的随机效应协方差结构主体。

重复测量。 此处指定的字段用于标识重复观测值。 例如，单个变量周可以标识医学研究中 10 周内的观测值，而月和天可共同用于标识一年内的每一天的观测值。

定义协方差分组依据。 此处指定的分类字段用于定义独立的重复效应协方差参数集；每个由分组字段的交叉分类定义的类别都具有此字段。 所有主体具有相同的协方差类型；相同协方差组内的主体具有相同的参数值。

空间协方差坐标。 该列表中的变量指定为重复的协方差类型选择了其中一种空间协方差类型时重复观察的坐标。

重复协方差类型。 这指定残差的协方差结构。 可用结构如下：

- 一阶自回归 (AR1)
- 自回归移动平均值 (1,1) (ARMA11)
- 复合对称
- 对角线
- 标度恒等
- 空间：幂
- 空间：指数
- 空间：高斯
- 空间：线性
- 空间：线性对数
- 空间：球面
- Toeplitz
- 非结构化
- 方差成分

目标

这些设置通过关联函数定义目标、其分布以及其到预测变量的关系。

目标。 需要目标。 它可具有任何测量级别，同时目标的测量级别可限制适当的分布和关联函数。

- **使用试验次数作为分母**。 当目标响应是一系列试验中发生的事件数时，目标字段包含事件数，您可以额外选择一个包含试验数的字段。 例如，在试验新型杀虫剂时，可以对蚂蚁样本施用不同浓度的杀虫剂，然后记录每个样本中杀灭的蚂蚁数量以及被施用杀虫剂的蚂蚁数量。 在本例中，记录杀灭的蚂蚁数量的字段应指定为目标（事件）字段，记录每个样本中蚂蚁数量的字段应指定为试验字段。 如果在每个样本中蚂蚁数量都相同，那么可以将试验数指定为固定值。

  试验数应大于等于每个记录的事件数。 事件应为非负整数，试验应为正整数。
- **定制参考类别**。 对于分类目标，您可以选择参考类别。 这会影响某些输出，如参数估计值，但不应更改模型拟合度。 例如，如果您的目标使用 0、1 和 2 的值，缺省情况下，过程将最后（最高值）的类别或 2 作为参考类别。 在这种情况下，参数估计值应解释为与类别 0 或 1 的似然估计相关，此似然估计相对于类别 2 的似然估计。 如果您指定定制类别并且目标已定义了标签，那么可以通过从列表中选择值来设置参考类别。 在指定模型的过程中并不确定某一特定字段的编码方式时，这种方法非常方便。

目标分布以及与线性模型的关系（链接）。 给定预测变量的值，该模型将预期目标的值分布会遵循指定的形状，同时目标值通过指定关联函数与预测变量线性相关。 可以使用现有的一些通用模型的快捷方式，或者如果您要拟合的特殊分布与关联函数组合不在此快捷列表上，还可以选择**定制**设置。

- **线性模型**。 指定恒等关联的正态分布，在可使用线性回归或 ANOVA 模型预测目标时非常有用。
- **伽玛回归**。 指定对数关联的伽玛分布，应在目标包含所有正值并偏斜到更大值时使用。
- **对数线性**。 指定对数关联的泊松分布，应在目标表示固定时间段内的发生计数时使用。
- **负二项式回归**。 指定对数关联的负二项式分布，应在目标和分母表示观察 *k* 次成功所需试验次数时使用。
- **多项式逻辑回归**。 指定多项分布，应在目标是多类别响应时使用。 它使用累积 logit 关联（有序结果）或广义 logit 关联（多类别名义响应）。
- **二元逻辑回归**。 指定 logit 关联的二元分布，应在目标是通过 logistic 回归模型预测的二元响应时使用。
- **二元概率单位**。 指定 probit 关联的二元分布，应在目标是基础正态分布的二元响应时使用。
- **区间删失生存**。 指定互补双对数关联的二项式分布，在一些观测没有终止事件的生存分析中有用。

分布

此选项指定目标的分布。 指定非正态分布和非恒等关联函数的功能是广义线性混合模型相对线性混合模型的重要改进。 分布-关联函数可能存在多种组合，其中一些适合任何给定的数据集，因此可以根据先验理论的要求进行选择，或选择最合适的组合。

二项式

此分布仅适合表示二元响应或事件数量的目标。

伽玛

该分布适用于具有正刻度值并向更大的正值偏度的目标。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。

逆高斯

该分布适用于具有正刻度值并向更大的正值偏度的目标。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。

多项式

此分布适用于表示多类别响应的目标。 模型形式取决于目标的测量级别。

名义目标将产生名义多项模型，其中为目标的每个类别（参考类别除外）分别估计一组模型参数。 给定预测变量的参数估计显示预测变量与目标每个类别相对于参考类别的似然之间的关系。

有序目标将产生有序多项模型，其中传统截距项被替换为一组与目标类别累积概率相关的**阈值**参数。

负二项式

负二项式回归使用带对数关联的负二项式分布，它在目标代表具有较高方差的出现次数时使用。

正态

该分布适合围绕某个中间值（平均值）呈对称钟型分布的连续目标。

泊松

该分布可视为被观察事件在固定时间段内发生的次数，适合具有非负整数值的变量。 如果数据值是非整数、小于 0 或缺失，那么分析中不会使用相应的个案。

关联函数

关联函数是目标的转换形式，可用于模型估计。 可用函数有：

恒等

$f(x)=x$。 目标不会被转换。 此关联函数可用于任何分布，多项式分布除外。

互补双对数

$f(x)=\log(-\log(1-x))$。 该函数只适用于二项式或多项式分布。

Cauchit

$f(x) = \tan(\pi (x - 0.5))$。 该函数只适用于二项式或多项式分布。

日志

$f(x)=\log(x)$。 此关联函数可用于任何分布，多项式分布除外。

对数互补

$f(x)=\log(1-x)$。 该函数只适用于二项式分布。

分对数

$f(x)=\log(x / (1-x))$。 该函数只适用于二项式或多项式分布。

负双对数

$f(x)=-\log(-\log(x))$。 该函数只适用于二项式或多项式分布。

概率

$f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是逆标准正态累积分布函数。 该函数只适用于二项式或多项式分布。

幂

$f(x)=x^{\alpha}$，如果 $\alpha \neq 0$。$f(x)=\log(x)$，如果 $\alpha=0$。α 是必需的数字规范，并且必须是实数。 此关联函数可用于任何分布，多项式分布除外。

固定效应

固定效应因子通常被视为字段，其所需的值都表示在数据集中，同时可用于评分。 缺省情况下，在模型固定效应部分输入未在对话框中指定的具有预定义输入角色的字段。 分类（标志、名义和有序）字段用作模型中的因子，连续字段则用作协变量。

通过在源列表中选择一个或多个字段并拖到效应列表来将效应输入至模型。 所创建的效应类型取决于您放置选择的热区。

- **主要**。 放置的字段在效应列表底部显示为单独的主效应。
- **双向**。 放置字段的所有可能对在效应列表底部显示为双向交互。
- **三向**。 放置字段的所有可能三元组在效应列表底部显示为三向交互。
- ***.** 所有放下字段的组合显示为效应列表底部的单个交互。

效应构建器右侧的按钮可用于执行各种操作。

表 *10:* 效应构建器按钮描述

图标	描述
	通过选择要删除的项并单击删除按钮，可以从固定效应模型中删除项。
	通过选择要重新排序的项并单击向上或向下箭头，可以在固定效应模型中对项进行重新排序。
	使用第 153 页的『添加定制项』对话框并通过单击“添加定制项”按钮，可以向模型添加嵌套项。

包括截距。 模型中通常包含截距。 如果您可以假设数据穿过原点，那么可以排除截距。

添加定制项

在此过程中，可为您的模型建立嵌套项。 嵌套项在创建因子或协变量的效应模型时非常有用，其值不会与其他因子级别交互。 例如，杂货连锁商店可能会追踪几个店址的顾客消费习惯。 由于每位顾客只经常光顾某一位置的商店，因此可以说客户效应**嵌套在**商店位置效应中。

此外，还可以加入交互效应，比如包括相同协变量的多项项，或是将多重嵌套级别添加到嵌套项。

限制: 嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。 因此，如果 *A* 是因子，那么指定 *A*A* 是无效的。
- 嵌套效应内的所有因子必须是唯一的。 因此，如果 *A* 是因子，那么指定 *A(A)* 是无效的。
- 效应不可嵌套在协变量中。 因此，如果 *A* 是因子且 *X* 是协变量，那么指定 *A(X)* 是无效的。

构建嵌套项

1. 选择嵌入另一个因子的因子或协变量，然后单击方向按钮。
2. 单击（**内部**）。
3. 选择前一个因子或协变量嵌套在其中的因子，然后单击箭头按钮。
4. 单击**添加项**。

（可选）可包含交互效应或者将多层嵌套添加到嵌套项中。

随机效应

随机效应因子是其在数据文件中的值可视为来自较大总体值的随机样本的字段。 这对解释目标中的多余可变性十分有用。 缺省情况下，如果您已在“数据结构”选项卡中选择多个主体，那么将为最内部主体以外的每个

主体创建一个随机效应块。 例如，如果您在“数据结构”选项卡上选择“学校”、“班级”和“学生”作为主体，那么将会自动创建下列随机效应块：

- 随机效应 1：主体是学校（没有效应，只有截距）
- 随机效应 2：主体是学校 * 班级（没有效应，只有截距）

您可以通过以下方式使用随机效应块：

1. 要添加新的块，请单击**添加块...** 这将打开 第 154 页的『随机效应块』 对话框。
2. 要编辑现有块，请选择要编辑的块，然后单击**编辑块...** 这将打开 第 154 页的『随机效应块』 对话框。
3. 要删除一个或多个块，选择您要删除的块并单击删除按钮。

随机效应块

通过在源列表中选择一个或多个字段并拖到效应列表来将效应输入至模型。 所创建的效应类型取决于您放置选择的热区。 分类（标志、名义和有序）字段用作模型中的因子，连续字段则用作协变量。

- **主要**。 放置的字段在效应列表底部显示为单独的主效应。
- **双向**。 放置字段的所有可能对在效应列表底部显示为双向交互。
- **三向**。 放置字段的所有可能三元组在效应列表底部显示为三向交互。
- ***.** 所有放下字段的组合显示为效应列表底部的单个交互。

效应构建器右侧的按钮可用于执行各种操作。

表 *11:* 效应构建器按钮描述

图标	描述
	通过选择您要删除的项并单击删除安妮，可以从模型中删除项。
	通过选择要重新排序的项并单击向上或向下箭头，可以在模型中对项进行重新排序。
	使用第 153 页的『添加定制项』对话框并通过单击“添加定制项”按钮，可以向模型添加嵌套项。

包括截距。 缺省情况下截距通常不包括在随机效应中。 如果您可以假设数据穿过原点，那么可以排除截距。

显示该块的参数预测。 指定显示随机效应参数估计。

定义协方差分组依据。 此处指定的分类字段用于定义独立的随机效应协方差参数集；每个由分组字段的交叉分类定义的类别都具有此字段。 可为每个随机效应区组指定不同的分组字段集。 所有主体具有相同的协方差类型；相同协方差组内的主体具有相同的参数值。

主体组合。 您可以从“数据结构”选项卡的预设主体组合中指定随机效应主体。 例如，如果学校、班级和学生定义为“数据结构”选项卡上的主体，那么按照这种顺序，“主体”组合下拉列表将包含选项**无**、**学校**、**学校 * 班级**和**学校 * 班级 * 学生**。

随机效应协方差类型。 这指定残差的协方差结构。 可用结构如下：

- 一阶自回归 (AR1)
- 异质自回归 (ARH1)
- 自回归移动平均值 (1,1) (ARMA11)
- 复合对称
- 异质复合对称 (CSH)
- 对角线

- 标度恒等
- Toeplitz
- 非结构化
- 方差成分

权重和偏移量

分析权重。 尺度参数是与响应方差相关的估计模型参数。 分析权重是“已知”值，可能因观测值的不同而异。 如果指定了分析权重字段，那么对每个观测值，都会用与响应方差相关的尺度参数除以该分析权重值。 分析中不使用分析权重值小于等于 0 或缺失的记录。

偏移量。 偏移量项是“结构性”预测变量。 它的系数不通过模型估计而假定其值为 1；因此，偏移量的值只与目标的线性预测变量简单相加。 这对于泊松回归模型尤其有用，在这种模型中，每个观测值对于相关事件可以具有不同的揭示级别。

例如，对各个驾驶员的事故率建模时，三年驾驶经验的驾驶员在一次事故中的过错率与 25 年驾驶经验的驾驶员在一次事故中的过错率存在重大差别。 如果将驾驶员的经验的自然对数作为偏移量项包括在内，那么可以按照泊松响应或负二项式响应对事故数进行建模。

其他分布和关联类型的组合将需要偏移变量的其他转换。

常规构建选项

这些选择指定一些用于构建模型的更高级条件。

排列顺序
这些控件用于确定目标和因子（分类输入）类别的顺序，以确定“最后一个”类别。 如果目标未分类或者在第 151 页的『目标』设置上指定了定制参考类别，那么目标排序顺序设置将被忽略。

停止规则
您可指定算法执行的最大迭代次数。 此算法使用包含内层循环和外层循环的双迭代式过程。 指定的最大迭代次数值适用于这两个循环。 请指定非负整数。 缺省值为 100。

估计后设置
这些设置确定如何计算一些用于查看的模型输出。

置信度级别 (%)
这是用于计算模型系数的区间估计值的置信度级别。 请指定大于 0 且小于 100 的值。 缺省值为 95。

自由度
用来指定如何计算显著性检验的自由度。 如果样本够大、数据平衡或模型使用较简单的协方差类型（例如，调整恒等式或对角线），可选择**残差方法**。 这是缺省设置。 如果样本很小、数据不平衡或模型使用复杂的协方差类型（例如非结构化），请选择 **Satterthwaite 近似法**。 如果样本较小，并且您使用的是受限最大似然 (REML) 模型，请选择 **Kenward-Roger 近似**。

固定效应和系数检验
这是用于计算参数估计值协方差矩阵的方法。 如果担心实验数据违反模型假设，可选择稳健估计。

估计

模型构建算法使用包含内层循环和外层循环的双迭代式过程。 下列设置适用于内层循环。

排序顺序
这些控件确定目标和因子（分类输入）的类别顺序，以确定“最后一个”类别。 如果目标未分类或者在第 151 页的『目标』设置上指定了定制参考类别，那么目标排序顺序设置将被忽略。

参数收敛。
如果参数估计值的最大绝对变化或最大相对变化小于指定的非负值，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

对数似然收敛。
如果对数似然函数的绝对变化或相对变化小于指定的非负值，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

Hessian 收敛。
对于**绝对**指定，如果基于 Hessian 的统计小于指定的值，那么假设收敛性。 对于**相对**指定，如果统计小于指定值与对数似然估计的绝对值的乘积，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

最大 Fisher 评分步长。
指定非负整数。 值为 0 指定 Newton-Raphson 方法。 如果值大于 0，那么指定最多使用迭代次数 *n* 次 Fisher 评分算法，其中 *n* 是指定的整数，其后是 Newton-Raphson。

奇异性容差。
此值在检查奇异性时用作容错。 请指定一个正值。

中止规则
您可指定算法执行的最大迭代次数。 此算法使用包含内层循环和外层循环的双迭代式过程。 指定的最大迭代次数值适用于这两个循环。 指定非负整数。 缺省值为 100。

估计后设置
这些设置可确定计算某些模型输出供查看的方式。

置信度级别 (%)
这是用于计算模型系数的区间估计值的置信度级别。 请指定大于 0 且小于 100 的值。 缺省值为 95。

自由度
用来指定如何为显著性检验计算自由度。 如果样本够大、数据平衡或模型使用较简单的协方差类型（例如，标度恒等或对角线），请选择**残差法**。 这是缺省设置。 如果样本较小、数据不平衡或模型使用较复杂的协方差类型（例如，非结构化），请选择 **Satterthwaite 近似**。 如果样本较小，并且您使用的是受限最大似然 (REML) 模型，请选择 **Kenward-Roger 近似**。

固定效应与系数检验
这是用于计算参数估计值协方差矩阵的方法。 如果担心实验数据违反模型假设，可选择稳健估计。

注: 缺省情况下，将使用参数收敛算法，此算法选中容差为 1E-6 的最大**绝对**更改。 此设置所生成的结果可能不同于在版本 22 之前的版本中获得的结果。 要重新生成版本 22 之前的版本中的结果，请对参数收敛准则使用**相对**，并保留缺省容差值 1E-6。

常规

模型名称。 可以基于目标字段来自动生成模型名称，或指定定制名称。 自动生成的名称为目标字段名。 如果存在多个目标，那么模型名称将由这些字段名按顺序排列组成，且字段名之间通过“与” (&) 符号连接。 例如，如果 *field1 field2 field3* 是目标，那么模型名称是：*field1 & field2 & field3*。

可用于评分。 对模型进行评分时，应生成此组中的选定项目。 在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。 计算的置信度可基于预测值的概率（最高的预测概率）或最高预测概率和次高预测概率之间的差异。

- **分类目标的预测概率**。 这将生成分类目标的预测概率。 为每个类别创建一个字段。
- **标志目标的倾向评分**。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 该模型产生原始倾向评分；如果分区处于有效，那么模型还会根据测试分区产生调整后的倾向评分。

估算的平均值

此选项卡可以显示因子级别和因子交互效应的估计边际均值。 估计边际均值对多项模型不可用。

项
此处列出完全由分类字段组成的固定效应中的模型项。 选中您希望模型生成估计边际均值的每个项。

对比类型
这将指定对比字段级别使用的对比类型。

无
不生成对比。

成对
生成所指定因子所有级别组合的成对比较。 这是因子交互的唯一可用对比方法。

偏差
对比是因子的每个级别与总均值的比较。

简单
对比是因子的每个级别（最后一个除外）与最后一个级别的比较。 “最后一个”级别是根据构建选项上指定的因子排序顺序来确定。 注意，所有这些对比类别不正交。

对比字段
这将指定因子，其水平使用所选的对比类型进行比较。 如果选择**无**作为对比类型，那么无法（无需）选择任何对比字段。

连续字段
列出的连续字段提取自使用连续字段的固定效应中的项。 当计算估计边际均值时，协方差固定为指定值。 选择平均值或或指定定制值。

使用以下方法针对多重比较进行调整
执行含多重对比的假设检验时，总体显著性水平可根据所含对比的显著性水平进行调整。 这允许您选择调整方法。

最小显著性差异
此方法并不控制拒绝“某些线性对比有别于原假设值”这一假设的总体概率。

连续 *Bonferroni (Sequential Bonferroni)*
这是一种连续下降的拒绝的 Bonferroni 过程，该过程在拒绝单个假设方面保守程度很低，但保持相同的整体显著性水平。

连续 *Sidak (Sequential Sidak)*
这是一个逐步下降的排斥性 Sidak 过程，就排斥单个假设而言，其保守性小得多，且保持了相同的总体显著性水平。

最低显著性差异方法的保守程度不及连续 Sidak 方法，后者的保守程度又不及顺序 Bonferroni 方法；这意味着，最低显著性差异会拒绝至少与连续 Sidak 一样多的单个假设，而后者又会拒绝至少与顺序 Bonferroni 一样多的单个假设。

显示估计均值的方式
这将指定是否基于目标原始刻度或关联函数转换计算估计边际均值。

原始目标标度
计算目标的估计边际均值。 注意，当使用事件/试验选项指定目标时，这将给出事件/试验比例而不是事件数量的估计边际均值。

关联函数转换
计算线性预测变量的估计边际均值。

模型视图

缺省情况下，显示“模型摘要”视图。 要查看另一个模型视图，从视图缩略图中选择。

模型摘要

此视图是模型及其拟合的快照和概览摘要。

表。 表可标识目标、概率分布和在目标设置上指定的关联函数。 如果目标是由事件和试验定义，那么单元格将会拆分，以显示事件字段以及试验字段或者固定次数的试验。 此外，还会显示经有限样本校正的 Akaike 信息准则 (AICC) 和 Bayesian 信息准则 (BIC)。

- *Akaike* 已校正。用于根据 -2 (受限) 对数似然选择和比较混合模型的度量。 值越小，表示模型拟合得越好。 AICC 用于更正小样本的 AIC。 随样本大小的增加，AICC 将收敛为 AIC。
- *Bayesian*。用于根据 -2 对数似然选择和比较模型的度量。 值越小，表示模型拟合得越好。 BIC 也会“惩罚”过多参数模型（例如，具有大量输入的复杂模型），但比 AIC 更严格。

图。 如果目标为分类目标，那么图表将显示最终模型的准确性，即正确分类的百分比。

数据结构

该视图提供您指定的数据结构的摘要，并帮助您检查是否正确指定主体和重复测量。 将为每个主体字段和重复测量字段以及目标显示所观察到的第一个主体信息。 此外，将显示每个主体字段和重复测量字段的级别数。

预测-实测

对于连续目标（包括指定为事件/试验的目标），这将显示已分箱的散点图，其中，在纵轴上显示预测值，在横轴上显示实测值。 理想情况下，这些点应该位于 45 度线上；此视图可以告诉您是否有任何记录被模型预测得特别糟糕。

分类

对于分类目标，这将显示热图中观测值与预测值的交叉分类以及整体正确百分比。

表样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **行百分比**。 此样式将在单元格中显示行百分比（单元格计数表示为行总数的百分比）。 这是缺省选项。
- **单元格计数**。 这将显示单元格中的单元格计数。 热图的阴影部分仍然基于行百分比。
- **热图**。 此样式将只显示阴影，而不会在单元格中显示值。
- **压缩**。 此样式不会在单元格中显示行或列标题，也不会显示值。 在目标具有许多分类时，此样式将十分有用。

缺失。 如果目标上的任何记录缺失值，则会显示在所有有效行下的 **（缺失）**行中。 具有缺失值的记录不会对整体正确百分比有任何影响。

多个目标。 如果有多个分类目标，那么每个目标将显示在单个表中，同时有一个**目标**下拉列表控制要显示的目标。

大型表。 如果所显示的目标有 100 多个类别，将不显示任何表。

固定效应

此视图显示模型中每个固定效应的大小。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **图**。 此图表中的效应按照“固定效应”设置中指定的顺序从上到下排序。 在图表中，连接线条根据效应的显著性进行加权，粗线条表示较显著的效应（*p* 值较小）。 这是缺省选项。
- **表**。 这是 ANOVA 表，其中包含总体模型与各个模型效应。 效应按照其在固定效应设置里指定的顺序从上到下排序。

显著性。 这里有一个“显著性”滑块，用于控制在视图中显示哪些效应。 显著性值大于滑块值的效应将被隐藏。 这不会更改模型，而只是让您可以关注最重要的系数。 缺省情况下此值为 1.00，因此不会根据显著性来过滤系数。

固定系数

此视图显示模型中每个固定系数的值。 注意，由于因子（分类预测变量）在模型内部经过指示符编码，因此包含因子的**效应**通常具有多个关联**系数**；每种类别一个关联系数，但对应于冗余系数的类别除外。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **图**。 此图表首先显示截距，然后按照“固定效应”设置中指定的顺序从上到下对效应进行排序。 在包含因子的效应中，系数按数据值升序进行排序。 在图表中，连接线的颜色和粗细取决于系数的显著性，粗线对应于较显著的系数（*p* 值较小）。 这是缺省样式。
- **表**。 这将显示各个模型系数的值、显著性检验和置信区间。 在截距之后，效应按照其在固定效应设置里指定的顺序从上到下排序。 在包含因子的效应中，系数按数据值升序进行排序。

多项式。 如果多项分布起作用，那么多项式下拉列表将控制要显示的目标分类。 这些值在列表中的排序顺序由“构建选项”设置上的规范决定。

指数。 用于显示某些模型类型的指数分布系数估计值和置信区间，包括二元 Logisitic 回归（二项式分布和分对数关联函数），名义 Logistic 回归（多项式分布和分对数关联函数），负二项式回归（负二项式分布和分对数关联函数）以及对数线性模型（泊松分布和分对数关联函数）。

显著性。 这里有一个“显著性”滑块，以控制在视图中显示哪些系数。 显著性值大于滑块值的系数将隐藏。 这不会更改模型，而只是让您可以关注最重要的系数。 缺省情况下此值为 1.00，因此不会根据显著性来过滤系数。

随机效应协方差

该视图显示随机效应协方差矩阵（**G**）。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **协方差值**。 这是协方差矩阵热图，效应在其中按照“固定效应”设置中指定的顺序从上到下进行排序。 corrgram 图中的颜色对应于键中所显示的单元格值。 这是缺省选项。
- **相关图**。 这是协方差矩阵的热图。
- **压缩**。 这是没有行和列标题的协方差矩阵热图。

区组。 如果有多个随机效应块，那么将有一个“块”下拉列表来选择要显示的块。

组。 如果随机效应块具有组规范，那么会有“组”下拉列表，用于选择要显示的组级别。

多项式。 如果多项分布起作用，那么多项式下拉列表将控制要显示的目标分类。 这些值在列表中的排序顺序由"构建选项"设置上的规范决定。

协方差参数

此视图显示协方差参数估计值，还有残差和随机效应的相关统计。 这些是高级而重要的结果，提供协方差结构是否合适的信息。

摘要表。 这是一个关于残差 (**R**) 和随机效应 (**G**) 协方差矩阵的参数数量，固定效应 (**X**) 和随机效应 (**Z**) 设计矩阵中的排序（列数）以及由定义数据结构的主体字段所定义的主体数量的快速参考。

协方差参数表。 对于选定的效应，将为每个协方差参数显示估计值、标准误差和置信区间。 所显示的参数数量取决于效应的协方差结构，对于随机效应块，那么取决于块中的效应数量。 如果您看到非对角线参数不显著，您可以使用更加简单的协方差结构。

效应。 如果有多个随机效应块，那么将有一个“块”下拉列表来选择要显示的残差或随机效应块。 残差效应总是可用。

组。 如果残差或随机效应区组是分组指定的，那么会有“组”下拉列表来选择要显示的分组级别。

多项式。 如果多项分布起作用，那么多项式下拉列表将控制要显示的目标分类。 这些值在列表中的排序顺序由“构建选项”设置上的规范决定。

估计均值：显著效应

这些是为 10 个“最显著”的固定全因子效应显示的图表，首先显示的是三向交互，然后显示的是双向交互，最后显示的是主效应。 该图表在横轴上显示主效应（或者交互中第一个列出的效应）的每个值的纵轴上目标的模型估计值；为交互中第二个列出的效应的每个值生成单独的折线；对于三向交互中第三个列出的效应的每个值，将生成单独的图表；所有其他预测变量保持不变。 它提供了目标上每个预测变量的系数的效应的有用可视化显示。 请注意，如果没有显著的预测变量，那么不会生成估计均值。

置信度。 这将使用指定为“构建选项”一部分的置信水平显示边际均值的置信上限和下限。

估计平均值：定制效应

这些是用户请求的固定所有因子效应的表和图表。

样式。 有多种不同的显示样式，可以从**样式**下拉列表中进行访问。

- **图**。 该样式显示水平轴上主效应（或交互中首个列出的效应）的每个值在垂直轴上的目标模型估计值线图；为交互中的第二个列出效应的每个值生成单独的线；为三阶交互中的第三个列出效应的每个值生成单独的图表；所有其他预测变量保持恒定。

如果请求了对比，将显示另一个图表以比较对比字段的级别；对于交互，将为对比字段之外的每个效应级别组合显示图表。 对于**成对**对比，会是一个距离网络图表；也就是图形化表示的比较表，表中网络节点间的距离与样本间的差异相对应。 黄线对应于统计上的显著差异；黑线对应于不显著的差异。 将鼠标悬停在该网络中的某条线上，可显示该线所连节点间的调整差异显著性工具提示。

对于**偏差**对比，会显示目标模型估计值位于垂直轴上，而对比字段值位于水平轴上的条形图；对于交互效应，每个效应级别组合（而非对比字段）会显示一个图表。 这些条形图将以黑色水平线表示对比字段的每个水平和总平均值之间的差异。

对于**简单**对比，会显示目标模型估计值位于垂直轴上，而对比字段值位于水平轴上的条形图；对于交互效应，每个效应级别组合（而非对比字段）会显示一个图表。 这些条形图将以黑色水平线表示对比字段每个水平（最后一个水平除外）和最后一个水平之间的差异。

- **表**。 该样式显示目标模型估计值、其标准误差和效应中每个字段水平组合的置信区间；所有其他预测变量保持恒定。

  如果请求了对比，另一个表将显示每个对比的估计、标准误差、显著性检验和置信区间；对于交互，除对比字段之外，效应的每个效应水平组合都有一组单独的行。 此外，将显示带整体检验结果的表；对于交互除了对比字段之外，效应的每个效应水平组合都有单独的整体检验。

置信度。 这将切换使用指定为“构建选项”一部分的置信度的边际平均值的置信上限和下限的显示。

布局。 用于切换对对比图表的布局。 圆形布局不如网状布局能使对比明显，但可避免重叠线。

设置

在对模型评分时，应生成此选项卡中的选定项目。 在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。 计算的置信度可基于预测值的概率（最高的预测概率）或最高预测概率和次高预测概率之间的差异。

- **分类目标的预测概率**。 这将生成分类目标的预测概率。 为每个类别创建一个字段。
- **标志目标的倾向评分**。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 该模型产生原始倾向评分；如果分区处于有效，那么模型还会根据测试分区产生调整后的倾向评分。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

GLE 节点

GLE 模型通过指定的关联函数标识与因子和协变量线性相关的因变量。 而且，该模型还允许因变量呈非正态分布。 它涵盖广泛使用的统计模型，例如适用于正态分布响应的线性回归、适用于二元数据的 Logistic 模型、适用于计数数据的对数线性模型、适用于区间删失生存数据的互补双对数模型，以及使用其非常通用的模型公式的许多其他统计模型。

示例。 运输公司可以使用广义线性模型，对在不同期间建造的一些轮船类型的损坏统计采用泊松回归，其结果模型可帮助确定哪些轮船类型最容易损坏。

汽车保险公司可以使用广义线性模型，对汽车损坏理赔采用伽玛回归，其结果模型可帮助确定对理赔额度贡献最大的因素。

医学研究人员可以使用广义线性模型来拟合区间删失生存数据的互补对数回归，以预测疾病复发时间。

GLE 模型的工作原理是构建一个公式，从而使输入字段值与输出字段值进行关联。 生成模型后，便可以将其用于为新数据估值。

对于分类目标、每条记录，将计算每个可能的输出类别的成员资格概率。 具有最高概率的目标类别将被指定为该记录的预测输出值。

需求。 您需要一个或多个输入字段，同时有且仅有一个具有两个或多个类别的目标字段（其测量级别可以为连续、分类或标志）。 必须对模型中使用的字段的类型完全实例化。

目标

这些设置通过关联函数定义目标、其分布以及其到预测变量的关系。

目标 目标为必要设置。 目标可以具有任何测量级别，并且目标的测量级别会影响适合的分布和关联函数。

- **使用预定义目标** 要使用上游“类型”节点 （或上游源节点的“类型”选项卡）中的目标设置，请选择此选项。
- **使用定制目标** 要手动分配目标，请选择此选项。
- **使用试验次数作为分母** 如果目标响应是一组试验中发生的事件数量，目标字段将包含该事件数量，您可选择包含试验次数的附加字段。 例如，在试验新型杀虫剂时，可以对蚂蚁样本施用不同浓度的杀虫剂，然后记录每个样本中杀灭的蚂蚁数量以及被施用杀虫剂的蚂蚁数量。 在本例中，记录杀灭的蚂蚁数量的字段应指定为目标（事件）字段，记录每个样本中蚂蚁数量的字段应指定为试验字段。 如果在每个样本中蚂蚁数量都相同，那么可以将试验数指定为固定值。

  试验数应大于等于每个记录的事件数。 事件应为非负整数，试验应为正整数。
- **定制参考类别**。 对于分类目标，您可以选择参考类别。 这会影响某些输出，如参数估计值，但不应更改模型拟合度。 例如，如果您的目标使用 0、1 和 2 的值，缺省情况下，过程将最后（最高值）的类别或 2 作为参考类别。 在这种情况下，参数估计值应解释为与类别 0 或 1 的似然估计相关，此似然估计相对于类别 2 的似然估计。 如果您指定定制类别并且目标已定义了标签，那么可以通过从列表中选择值来设置参考类别。 在指定模型的过程中并不确定某一特定字段的编码方式时，这种方法非常方便。

目标分布以及与线性模型的关系（关联） 给定预测变量的值，模型的预期为目标值按照指定的形状分布，并在指定的关联函数中与预测变量呈线性相关。 可以使用现有的一些通用模型的快捷方式，或者如果您要拟合的特殊分布与关联函数组合不在此快捷列表上，还可以选择**定制**设置。

- **线性模型** 使用恒等关联函数指定正态分布，在目标可用线性回归或 ANOVA 模型来预测时特别有用。
- **伽玛回归** 使用对数关联函数指定伽玛分布，在目标包含所有正值并向更大值偏斜时使用。
- **对数线性** 使用对数关联函数指定泊松分布，在目标代表某个固定时段内事件发生的次数时使用。
- **负二项式回归** 使用对数关联函数指定负二项式分布，在目标和分母代表观察到第 *k* 次成功所需的试验次数时使用。
- **Tweedie 回归** 指定包含恒等式、对数或幂关联函数的 Tweedie 分布，用于混合有零及正实数值的建模响应。 这些分布也称为复合泊松、复合伽玛及泊松伽玛分布。
- **多项式 Logistic 回归** 指定多项式分布，在目标为多类别响应时使用。 它使用累积 logit 关联（有序结果）或广义 logit 关联（多类别名义响应）。
- **二元 Logistic 回归** 使用分对数关联函数指定二项式分布，在目标为 logistic 回归模型预测的二元响应时使用。
- **二元概率值** 使用概率值关联函数指定二项式分布，在目标为具基本正态分布的二元响应时使用。
- **区间删失生存** 使用互补双对数关联函数指定二项式分布，在生存分析的观察没有终止事件时特别有用。
- **定制** 指定您自己的分布和关联函数组合。

分布

此选项指定目标的**分布**。 能够指定非正态分布和非恒等关联函数是广义线性模型相对于线性模型的重大改进。 分布-关联函数可能存在多种组合，其中一些适合任何给定的数据集，因此可以根据先验理论的要求进行选择，或选择最合适的组合。

- **自动** 如果您不确定要使用哪个分布，请选择此选项；节点将分析您的数据以估计并应用最佳的分布方法。
- **二项** 此分布仅适用于表示二元响应或事件数的目标。
- **伽玛** 此分布适用于具有向更大正值偏斜的正尺度值的目标。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。

- **逆高斯** 此分布适用于具有向更大正值偏斜的正尺度值的目标。 如果数据值小于等于 0 或缺失，那么分析中不会使用相应的个案。
- **多项式** 此分布适用于表示多类别响应的目标。 模型形式取决于目标的测量级别。

  名义目标将产生名义多项模型，其中为目标的每个类别（参考类别除外）分别估计一组模型参数。 给定预测变量的参数估计显示预测变量与目标每个类别相对于参考类别的似然之间的关系。

  有序目标将产生有序多项模型，其中传统截距项被替换为一组与目标类别累积概率相关的**阈值**参数。
- **负二项式** 负二项式回归使用带对数关联的负二项式分布，它在目标代表具有较高方差的出现次数时使用。
- **正太** 此分布适用于其值围绕中心值（均值）呈对称钟形分布的连续目标。
- **泊松** 此分布可视为固定时间段内发生感兴趣事件的次数，并且适用于具有非负整数值的变量。 如果数据值是非整数、小于 0 或缺失，那么分析中不会使用相应的个案。
- **Tweedie** 此分布适用于可以由泊松分布与伽玛分布混合表示的变量；在某种意义上，此分布是“混合”分布，因为它同时具备连续分布（采用非负实数值）和离散分布（正概率群位于单个值 0）的属性。 因变量必须是数值型变量，数据值大于或等于零。 如果数据值小于零或缺失，那么分析中不会使用相应的个案。 Tweedie 分布参数的固定值可以是任何大于 1 且小于 2 的数字。

关联函数

关联函数是允许进行模型估计的目标变换。 可用函数有：

- **自动** 如果您不确定要使用哪个关联函数，请选择此选项；节点将分析您的数据以估计并应用最佳的关联函数。
- **恒等** $f(x)=x$。 目标不会被转换。 此关联函数可用于任何分布，多项式分布除外。
- **互补双对数** $f(x)=\log(-\log(1-x))$。 该函数只适用于二项式或多项式分布。
- **Cauchit** $f(x) = \tan(\pi (x - 0.5))$。 该函数只适用于二项式或多项式分布。
- **对数** $f(x)=\log(x)$。 此关联函数可用于任何分布，多项式分布除外。
- **对数互补** $f(x)=\log(1-x)$。 该函数只适用于二项式分布。
- **分对数** $f(x)=\log(x / (1-x))$。 该函数只适用于二项式或多项式分布。
- **负双对数** $f(x)=-\log(-\log(x))$。 该函数只适用于二项式或多项式分布。
- **概率** $f(x)=\Phi^{-1}(x)$，其中 Φ^{-1} 是逆标准正态累积分布函数。 该函数只适用于二项式或多项式分布。
- **Power** $f(x)=x^{\alpha}$，如果 $\alpha \neq 0$。$f(x)=\log(x)$，如果 $\alpha=0$。α 是必需的数字规范，必须是实数。 此关联函数可用于任何分布，多项式分布除外。

Tweedie 参数 仅当您已选中 **Tweedie 回归**单选按钮或选择 Tweedie 作为**分布**方法时才可用。 选择介于 1 与 2 之间的某个值。

模型效应

固定效应因子通常被视为字段，其所需的值都表示在数据集中，同时可用于评分。 缺省情况下，在模型固定效应部分输入未在对话框中指定的具有预定义输入角色的字段。 分类（标志、名义和有序）字段可用作模型中的因子，连续字段可用作协变量。

通过在源列表中选择一个或多个字段并拖到效应列表来将效应输入至模型。 所创建的效应类型取决于您放置选择的热区。

- **主要** 拖入的字段显示为独立主效应，列在效应列表的底部。
- **双向** 所有可能的拖入字段配对显示为双向交互效应，列在效应列表的底部。
- **三向** 所有可能的三组拖入字段显示为三向交互效应，列在效应列表的底部。
- ***** 所拖动的全部字段组合起来，作为单一的交互显示在效应列表的底部。

效应构建器右侧的按钮可用于执行各种操作。

表 *12:* 效应构建器按钮描述

图标	描述
	通过选择要删除的项并单击删除按钮，可以从固定效应模型中删除项。
	通过选择要重新排序的项并单击向上或向下箭头，可以在固定效应模型中对项进行重新排序。
	通过单击“添加定制项”按钮，使用“添加定制项”对话框向模型中添加嵌套项。

包括截距 截距通常包括在模型中。 如果您可以假设数据穿过原点，那么可以排除截距。

添加定制项

在此过程中，可为您的模型建立嵌套项。 嵌套项在创建因子或协变量的效应模型时非常有用，其值不会与其他因子级别交互。 例如，杂货连锁商店可能会追踪几个店址的顾客消费习惯。 由于每个顾客只经常光顾其中一个店址，因此“顾客”效应可视为嵌入在“店址”效应中。

此外，还可以加入交互效应，比如包括相同协变量的多项项，或是将多重嵌套级别添加到嵌套项。

限制。 嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。 因此，如果 *A* 是因子，那么指定 *A*A* 是无效的。
- 嵌套效应内的所有因子必须是唯一的。 因此，如果 *A* 是因子，那么指定 *A(A)* 是无效的。
- 效应不可嵌套在协变量中。 因此，如果 *A* 是因子且 *X* 是协变量，那么指定 *A(X)* 是无效的。

构建嵌套项

1. 选择嵌入另一个因子的因子或协变量，然后单击方向按钮。
2. 单击（**内部**）。
3. 选择前一个因子或协变量嵌套在其中的因子，然后单击箭头按钮。
4. 单击**添加项**。

（可选）可包含交互效应或者将多层嵌套添加到嵌套项中。

权重和偏移量

分析权重 尺度参数是与响应方差相关的估计模型参数。 分析权重是“已知”值，可能因观测值的不同而异。 如果指定了**分析权重**字段，那么对每个观测值，都会用与响应方差相关的尺度参数除以该分析权重值。 分析中不会使用分析权重值小于等于 0 或缺失的记录。

偏移量 偏移量项是一个结构预测变量。 模型不估计该预测变量的系数，但假设其值为 1；因此，偏移量值只是简单地加到目标的线性预测变量中。 这对于泊松回归模型特别有用，在这种模型中，每个观测值对感兴趣事件的暴露程度可能不同。

例如，对各个驾驶员的事故率建模时，三年驾驶经验的驾驶员在一次事故中的过错率与 25 年驾驶经验的驾驶员在一次事故中的过错率存在重大差别。 如果将驾驶员经历的自然对数纳入偏移量项，那么事故数可以建模为具有对数关联的泊松或负二项式响应。

其他分布和关联类型的组合将需要偏移变量的其他转换。

构建选项

这些选项指定了构建模型的更高级标准。

排序顺序 这些控件用于确定目标和因子（分类输入）类别的顺序，以确定“最后一个”类别。 如果目标未分类或者在第 161 页的『目标』设置上指定了定制参考类别，那么目标排序顺序设置将被忽略。

估算后设置 这些设置可确定计算某些模型输出供查看的方式。

- **置信度级别 %** 这是用于计算模型系数的区间估计值的置信度级别。 请指定大于 0 且小于 100 的值。 缺省值为 95。
- **自由度** 用来指定如何为显著性检验计算自由度。 如果样本够大、数据平衡或模型使用如调整恒等式或对角线这类较简单的协方差类型，可选择**为全部检验固定（残差法）**。 这是缺省选项。 如果样本小、数据不平衡或模型使用如未结构化这类复杂的协方差类型，可选择**各检验不同（Satterthwaite 近似法）**。
- **固定效应和系数的检验**。 这是用于计算参数估计值协方差矩阵的方法。 如果担心实验数据违反模型假设，可选择稳健估计。

检测影响离群值 对于除多项式分布之外的所有分布，请选择此选项以确定影响离群值。

执行趋势分析 对于散点图，选择此选项可以执行趋势分析。

估算

方法 选择要使用的极大似然估计方法；可用选项包括：

- Fisher 评分方法
- Newton-Raphson
- 混合

最大 Fisher 迭代次数 请指定非负整数。 值为 0 指定 Newton-Raphson 方法。 如果值大于 0，那么指定最多使用迭代次数 *n* 次 Fisher 评分算法，其中 *n* 是指定的整数，其后是 Newton-Raphson。

尺度参数方法 选择估算尺度参数的方法；可用选项包括：

- 极大似然估计
- 固定值。 您还可以设置要使用的**值**。
- 偏差
- Pearson 卡方

负二项式方法 选择估算负二项式辅助参数的方法；可用选项包括：

- 极大似然估计
- 固定值。 您还可以设置要使用的**值**。

执行非负最小平方。 选择此选项以执行非负最小平方 (NNLS) 估计。 NNLS 是一种受约束最小平方问题类型，其中不允许系数变为负值。 并非所有数据集都适用于 NNLS，它需要预测变量和目标之间存在正相关性或没有相关性。

参数收敛 (Parameter Convergence)如果参数估计值的最大绝对变化或最大相对变化小于指定的非负值，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

对数似然收敛 (Log-likelihood Convergence)如果对数似然函数的绝对变化或相对变化小于指定的非负值，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

Hessian 收敛性 对于**绝对**指定，如果基于 Hessian 的统计小于指定的值，那么假设收敛性。 对于**相对**指定，如果统计小于指定值与对数似然估计的绝对值的乘积，那么假设收敛性。 如果指定的值等于 0，那么不使用该准则。

最大迭代次数 您可指定算法执行的最大迭代次数。 此算法使用包含内层循环和外层循环的双迭代式过程。 指定的最大迭代次数值适用于这两个循环。 指定非负整数。 缺省值为 100。

奇异性容差 此值在检查奇异性时用作容错。 请指定一个正值。

注: 缺省情况下使用**参数收敛**，在此设置中，将检查容差为 1E-6 的最大**绝对**更改。 此设置可能会生成与 V17 之前的版本中获取的结果不同的结果。 要重现 V17 之前的版本中的结果，请对“参数收敛”准则使用**相对**，并保留缺省容差值 1E-6。

模型选择

使用模型选择或规则化。 要激活此窗格上的控件，请选中此复选框。

方法。 选择模型选择方法或 (如果使用 **Ridge**) 要使用的规则化方法。 您可以从以下选项中进行选择：

- **套索**。 也称为 L1 规则化，如果存在大量预测变量，那么此方法比 "向前步进" 方法更快。 此方法将通过减小（即，施加惩罚）参数来避免过度拟合。 它可以将某些参数减少为零，从而执行变量选择套索。
- **海岭** 也称为 L2 规则化，此方法通过对参数进行缩小 (即施加惩罚) 来防止过度拟合。 它会按相同比例缩小所有参数，但不会清除任何参数，并且不是变量选择方法。
- **弹性网络**。 也称为 L1 + L2 规则化，此方法通过对参数进行缩小 (即，施加惩罚) 来防止过度拟合。 它可以将某些参数减少为零，从而执行变量选择。
- **前向逐步**。 此方法从模型中没有任何效应开始，并且一次添加或移除一个效应，直到无法根据逐步条件添加或移除其他效应为止。

自动检测双向交互。 要自动检测双向交互，请选择此选项。 请注意， GLE 仅检测两个分类变量和连续变量平方的交互。 它不会检测两个连续变量的交互以及分类变量和数字变量的交互。

惩罚参数

只有您选择了 Lasso 或弹性网络**方法**时，这些选项才可用。

自动选择惩罚参数。 如果您不确定要设置的参数惩罚，请选中此复选框，节点将识别并应用惩罚。

Lasso 惩罚参数。 输入要由套索模型选择 **方法**使用的惩罚参数。

弹性网络惩罚参数 1。 输入要由弹性网络模型选择 **方法**使用的 L1 惩罚参数。

弹性网络惩罚参数 2。 输入要由弹性网络模型选择 **方法**使用的 L2 惩罚参数。

向前步进

只有您选择了向前步进**方法**时，这些选项才可用。

包含 p 值不小于的效应。 指定效应必须包含在计算中的最小概率值。

除去 p 值大于以下值的效应。 指定效应必须包含在计算中的最大概率值。

在最终模型中定制最大效应数。 要激活 **最大效应数** 选项，请选中此复选框。

最大效应数。 指定使用向前步进构建方法时的最大效应数。 为了优化性能， **10** 是受支持的最高数量。

定制最多步骤数。 要激活 **最大步骤数** 选项，请选中此复选框。

最大步骤数。 指定使用正向逐步构建方法时的最大步骤数。

模型选项

模型名称 您可以自动根据目标字段生成模型名称，也可以指定**定制**名称。 自动生成的名称为目标字段名。 如果存在多个目标，那么模型名称将由这些字段名按顺序排列组成，且字段名之间通过“与” (&) 符号连接。 例如，如果 field1、field2 和 field3 是目标，那么模型名称为：*field1 & field2 & field3*。

计算预测变量重要性 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此缺省情况下，对于某些模型，预测变量重要性均处于关闭状态。

有关更多信息，请参阅 第 32 页的『预测变量重要性』。

GLE 模型块

GLE 模型块输出

创建 GLE 模型之后，在输出中就会提供以下信息。

模型信息表

“模型信息”表提供了有关模型的关键信息。 该表标识了一些高级模型设置，例如：

- 在“类型”节点或 GLE 节点字段选项卡中选择的目标字段的名称。
- 已建模和引用目标类别百分比。
- 概率分布和关联的关联函数。
- 所使用的模型构建方法。
- 最终模型中输入的预测变量数和数目。
- 分类准确性百分比。
- 模型类型。
- 模型的百分比准确性（如果目标是连续目标）。

记录摘要

摘要表显示用于拟合模型的记录数以及排除的记录数。 显示的详细信息包含所包括和排除的记录数和所占百分比以及未加权数目（如果您使用了频率加权）。

预测变量重要性

预测变量重要性图形以条形图的形式显示模型中前 10 个输入（预测变量）的重要性。

如果图表中存在超过 10 个字段，那么可以使用图表下的滑块来调整图表中包含的预测变量的选择。 滑块上的指示符标记为固定宽度，滑块上每个标记表示 10 个字段。 您可以沿滑块移动指示符标记，以显示后 10 个或前 10 个字段（按预测变量重要性排序）。

您可以双击图表以打开单独的对话框，您可以在其中编辑图形设置。 例如，您可以修改项目（如图形大小以及使用的字体大小和颜色）。 关闭此单独的编辑对话框后，更改会应用于“输出”选项卡中显示的图表。

残差（按预测图列出）

您可以使用此图来标识离群值，也可以使用它来诊断非线性或非恒定误差方差。 理想图将显示随机分布在基准线四周的点。

期望的模式为标准化偏差残差在线性预测变量的预测值之间的分布具有平均值零和恒定范围。 期望的模式是穿过零的水平线。

GLE 模型块设置

在 GLE 模型块的“设置”选项卡上，您可以指定模型评分期间用于原始倾向的选项和用于 SQL 生成的选项。 只有将模型块添加到流之后，此选项卡才可用。

计算原始倾向评分 对于仅具有标志目标的模型，您可以请求原始倾向评分，这些评分指示为目标字段指定的 true 结果的发生可能性。 除此之外，标准预测及置信度值也是如此。 调整后的倾向评分不可用。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定生成 SQL 的方式：

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**：如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

Cox 节点

Cox 回归为时间事件数据建立预测模块。 该模块生成生存函数用于为预测变量的给定值预测被观察事件在给定时间内 *t* 发生的概率。 从观察主体中估计预测的生存函数形状与回归系数；该方法稍后可应用于具有预测变量测量的新个案。 注意，已检查主体中的信息，即未在观察时间内经历被观察事件的信息，为模块估计做出巨大贡献。

示例。 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，选择了一个随机的客户样本，并且从数据库中抽取了他们作为客户的时间（无论他们是否仍为活跃客户）以及各种人口统计字段。

需求。 您需要一个或多个输入字段，只需一个目标字段，且必须在 Cox 节点中指定生存时间字段。 应对目标字段进行编码，使得“false”值表示生存时间，“true”值表示感兴趣事件已发生；目标字段的测量级别必须为标志，且带有字符串或整数存储。 （如有必要，可以使用"填充"或"派生"节点来转换存储。） 设置为 双向 或 无的字段将忽略。 必须对模型中使用的字段的类型完全实例化。 生存时间可以是任意数字字段。

注: 在对 Cox 回归模型进行评分时，如果将分类变量中的空字符串用作对模型构建的输入，将报告错误。 请避免使用空字符串作为输入。

日期和时间。 “日期和时间”字段不能直接用于定义生存时间；如果有“日期和时间”字段，那么应根据输入研究的日期和观测日期之间的差值，使用这些字段创建包含生存时间的字段。

Kaplan-Meier 分析。 可以在没有输入字段的情况下执行 Cox 回归。 这等效于 Kaplan-Meier 分析。

Cox 节点字段选项

生存时间。 选择数值字段（测量级别为连续的字段）以使节点可执行。 生存时间表示所预测记录的寿命。 例如，当模型化客户流失时间时，它可能是记录客户在组织内的时间长度的字段。 客户加入或流失的日期不会影响该模型；只有客户保有期的持续时间与其相关。

生存时间为无单位的持续时间。 您必须确保输入字段与生存时间相匹配。 例如，在按月测量流失的研究中，您可将月销售量而非年销售量用作输入。 如果您的数据具有开始日期和结束日期而不是持续时间，您必须在 Cox 代码上游将这些日期重新编码为持续时间。

此对话框中的剩余字段是整个 IBM SPSS Modeler 中通用的标准字段。 有关更多信息，请参阅主题 第 23 页的『建模节点字段选项』。

Cox 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

方法。 以下选项可用于向模型中输入预测变量：

- **进入法**。 这是缺省方法，用于将所有项直接输入模型中。 构建模型时不进行字段选择。
- **步进法** 顾名思义，字段选择步进法用于分步构建模型。 初始模型可能是最简单的模型，其模型中不含任何模型项（除常量外）。 在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，那么将该项添加到模型中。 此外，还会重新评估当前包含在模型中的项，以确定能否在不对模型功能造成重大减损的情况下删除其中任何项。 如果可以，那么会将其删除。 然后重复此过程，添加并/或删除其他项。 当无法再添加任何项来改进模型、且无法再删除任何项而不对模型功能造成减损时，最终模型便已生成。
- **向后步进法**。 向后步进法与步进法在本质上是相反的。 采用这种方法时，初始模型将包含作为预测变量的所有项。 每个步骤会评估模型中的项，并且将可以删除而不对模型功能造成重大减损的项删除。 此外，还会对先前删除的项进行重新评估，以确定其中的最佳项是否对模型的预测功能起到显著作用。 如果是，那么会将其重新添加到模型中。 当无法再删除任何项而不对模型功能造成重大减损、且无法再添加任何项以改进模型时，最终模型便已生成。

注：自动方法（包括步进法和向后步进法）是适应性强的学习方法，并且特别容易过度拟合训练数据。 使用这些方法时，用新数据或使用分区节点创建的保留测试样本对结果模型的有效性进行验证尤为重要。

组。 指定组字段会导致节点为该字段的每个类别计算单独的模型。 该字段可以是存储类型为字符串或整数的分类字段（标志或名义）。

模型类型。 有两个选项用于定义模型中的项。 **主效应**模型仅包括各个输入字段，而不检验输入字段之间的交互（乘法效应）。 **定制**模型仅包括您指定的项（主效应和交互效应）。 选择此选项时，应使用“模型项”列表在模型中添加或删除项。

模型项。 构建定制模型时，将需要明确指定模型中的项。 此列表显示了模型项的当前集合。 “模型项”列表右边的按钮可用于添加或删除模型项。

- 要将项添加到模型中，请单击 添加新的模型项按钮。
- 要删除项，请选定所需项，然后单击 删除选定模型项按钮。

将项添加到 Cox 回归模型

在请求定制模型时，可以通过单击“模型”选项卡中的添加新的模型项按钮将各项添加到模型中。 此时将打开一个新的对话框，您可在其中指定项。

要添加的术语的类型。 有几种将项添加到模型的方法，具体取决于在“可用字段”列表中对输入字段的选择。

- **单向交互效应**。 插入表示所有选定字段的交互的项。
- **主效应**。 针对每个选定的输入字段插入一个主效应项（该字段本身）。
- **所有双向交互效应**。 针对每个可能的选定输入字段对插入一个双向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B* 和 *C*，此方法将插入项 *A * B*、*A * C* 和 *B * C*。
- **所有三向交互效应**。 针对每个可能的选定输入字段组合（一次取三个字段）插入一个三向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B*、*C* 和 *D*，此方法将插入项 *A * B * C*、*A * B * D*、*A * C * D* 和 *B * C * D*。
- **所有四向交互效应**。 针对每个可能的选定输入字段组合（一次取四个字段）插入一个四向交互项（输入字段的积）。 例如，如果已在“可用字段”列表中选定输入字段 *A*、*B*、*C*、*D* 和 *E*，此方法将插入项 *A * B * C * D*、*A * B * C * E*、*A * B * D * E*、*A * C * D * E* 和 *B * C * D * E*。

可用字段。 列出要在构造模型项时使用的可用输入字段。 请注意，列表中可能包含非法输入字段，因此务必确保所有的模型项都只包含输入字段。

预览。 根据上述所选字段和项类型，显示单击**插入**时将添加到模型中的项。

Insert 键。 将项插入模型（根据当前选择的字段和项类型）并关闭对话框。

Cox 节点专家选项

收敛。 通过这些选项，您可以控制用于模型收敛的参数。 当您执行模型时，收敛设置将控制重复运行不同参数以观察其拟合程度的次数。 参数的尝试次数越多，结果将越接近（即，结果将会收敛）。 有关更多信息，请参阅主题 第 168 页的『Cox 节点收敛标准』。

输出。 通过这些选项，可以请求将显示在由节点构建的生成模型的高级输出中的附加统计量和散点图（包括生存曲线）。 有关更多信息，请参阅主题 第 168 页的『Cox 节点高级输出选项』。

步进。 通过这些选项，您可以使用步进估计法控制针对添加和删除字段的标准。 （如果已选择进入法，该按钮将处于禁用状态。） 有关更多信息，请参阅主题 第 169 页的『Cox 节点步进标准』。

Cox 节点收敛标准

最大迭代次数。 允许您指定模型的最大迭代次数，用于控制过程求解的时间。

对数似然收敛。 如果对数似然的相对变化小于此值，那么迭代将停止。 如果值为 0，则不使用该标准。

参数收敛。 如果参数估计中的绝对变化或相对变化小于此值，那么迭代将停止。 如果值为 0，则不使用该标准。

Cox 节点高级输出选项

统计。 您可以获得模型参数的统计，包括 exp(*B*) 的置信区间和估计值的相关性。 您可以在每一步或者仅在最后一步请求这些统计。

显示基线函数。 允许您显示协变量平均值下的基线风险函数和累积生存。

图

图有助于评估估计的模型和解释结果。 可以绘制生存函数、风险函数、对数累积函数和一减生存函数。

- 生存。在线性刻度上显示累积生存函数。

- 风险。以线性比例显示累积风险函数。
- **对数减对数**。 显示对估计运用了 ln(-ln) 转换之后的累积剩余估计。
- *1* 减生存函数。用于根据线性尺度按照被一减的方式绘制生存函数的散点图。

为每个值单独绘制一条线。 此选项仅可用于分类字段。

用于绘图的值。 由于这些函数都依赖于预测变量的值，因此您必须使用预测变量的常量值来绘制函数随时间推移的变化情况。 缺省情况下，将使用各个预测变量的平均数作为常数值，但您可以使用网格为散点图输入自己的值。 对于分类输入，使用指示符编码，因此每个类别都具有回归系数（最后一个类别除外）。 因此，分类输入具有每个指示符对比度的平均值，等于类别中对应于指示符对比度的观测值比例。

Cox 节点步进标准

剔除标准。 选择**似然比**可以获得更稳健的模型。 要缩短构建模型所需的时间，可以尝试选择 **Wald** 。 还有附加选项 **条件** ，此选项提供以基于条件参数估计值的似然比统计量的概率为依据的移除检验。

标准的显著性阈值。 通过使用此选项，您可以根据与每个字段关联的统计概率（*p* 值）来指定选择标准。 仅当关联的 *p* 值小于 **纳入标准** 值时，才会将字段添加到模型中；仅当 *p* 值大于 **剔出标准** 值时，才会将字段删除。 **纳入标准** 值必须小于 **剔出标准** 值。

Cox 节点设置选项

预测未来的生存状况。 指定一个或多个未来时间。 即在未发生终端事件的情况下，无论每个观测值是否可能至少在此时间长度（从现在开始）内生存，都将在每个时间值为每条记录预测生存时间，一个时间值对应一个预测值。 请注意，生存时间为目标字段的“false”值。

- **定期**。 生存时间值是根据指定的**时间间隔**和**要评分的时间段数**生成的。 例如，如果请求 3 个时间段，每次间隔为 2，那么将预测未来时间 2、4、6 的生存状况。 每条记录都以相同的时间值进行求值。
- **时间字段**。 在所选的时间字段中为每条记录提供生存时间（生成一个预测字段），因此可以在不同的时间评估各条记录。

过去生存时间。 将迄今为止的记录的生存时间指定为一个字段，例如将现有客户的保有期指定为一个字段。在未来时间对生存的似然进行评分取决于过去生存时间。

注：未来和过去生存时间的值必须在用于训练模型的数据的生存时间范围内。 时间超出此范围的记录将标记为空。

附加所有概率。 指定是否将输出字段每个类别的概率添加到该节点所处理的每条记录。 如果未选中此选项，那么仅添加预测类别的概率。 为每个未来时间计算概率。

计算累积风险函数。 指定是否将累积风险的值添加到每条记录中。 为每个未来时间计算累积风险。

Cox 模型块

Cox 回归模型表示由 Cox 节点所估计的方程式。 这些方程式包含由模型所捕获的所有信息及有关模型结构和性能的信息。

运行包含生成的 Cox 回归模型的流时，该节点可添加包含模型预测和相关概率在内的两个新字段。 新字段的名称派生自要预测的输出字段的名称并带有前缀和后缀，前缀为表示预测类别的 *$C-* 或表示相关概率的 *$CP-*，而后缀为未来时间间隔的数目或用于定义时间间隔的时间字段的名称。 例如，对于名为 *churn* 的输出字段，以及以规则区间定义的两个未来时间间隔，新字段命名为 *$C-churn-1*、 *$CP-churn-1*、 *$C-churn-2* 和 *$CP-churn-2*。 如果使用时间字段 *tenure* 定义未来时间，那么新字段为 *$C-churn_tenure* 和 *$CP-churn_tenure*。

如果在 Cox 节点中选中了 **追加所有概率** 设置选项，那么会针对每个未来时间添加两个附加字段，其中包含每条记录生存和失败的概率。 这些附加字段根据输出字段的名称进行命名，前缀为 *$CP-<false value>-*（表示生存概率）和 *$CP-<true value>-*（表示事件的发生概率），后缀为未来时间间隔的编号。 例如，对于“false”值为 0，“true”值为 1 的输出字段和以规则区间定义的两个未来时间间隔，新字段命名为 *$CP-0-1*、*$CP-1-1*、*$CP-0-2* 和 *$CP-1-2*。 如果使用单个时间字段 *tenure* 定义未来时间，由于存在单个的未来区间，那么新字段为 *$CP-0-1* 和 *$CP-1-1*

如果在 Cox 节点中选中了 **计算累积风险函数** 设置选项，那么会针对每个未来时间添加附加字段，其中包含每条记录的累计风险函数。 这些附加字段是根据输出字段的名称进行命名的并带有前缀和后缀，前缀为 *$CH-*，而后缀为未来时间间隔的数目或用于定义时间间隔的时间字段的名称。 例如，对于名为 *churn* 的输出字段，以及以规则区间定义的两个未来时间间隔，新字段命名为 *$CH-churn-1* 和 *$CH-churn-2*。 如果使用时间字段 *tenure* 定义未来时间，那么新字段为 *$CH-churn-1*。

Cox 回归输出设置

除生成 SQL 外，块的“设置”选项卡与模型节点的“设置”选项卡包含相同的控件。 块控件的缺省值由模型节点中设置的值决定。 有关更多信息，请参阅主题 第 169 页的『Cox 节点设置选项』。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

Cox 回归高级输出

Cox 回归的高级输出可提供有关所估计模型及其性能的详细信息，其中包括生存曲线。 高级输出中包含的大部分信息的技术含量都很高，需要具备 Cox 回归方面的广泛知识才能正确理解该输出。

第 11 章 聚类模型

聚类模型主要用来确定相似记录的组并根据它们所属的组来为记录添加标签。 不需事先了解组信息及组特征即可完成该操作。 事实上，甚至无法确切知道要查找多少个组。 这点将聚类模型与其他机器学习方法区别开来，即不存在供模型预测的预定义输出或目标字段。 由于不存在用于判断模型的分类效果的外部标准，因而这些模型通常被称作 **不受监督学习** 模型。 对于这些模型而言，不存在 对或 错的答案。 模型的 值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。

聚类方法基于对记录间距离和聚类间距离的测量。 将记录指派给聚类时将尽量缩短属于同一个聚类的记录之间的距离。

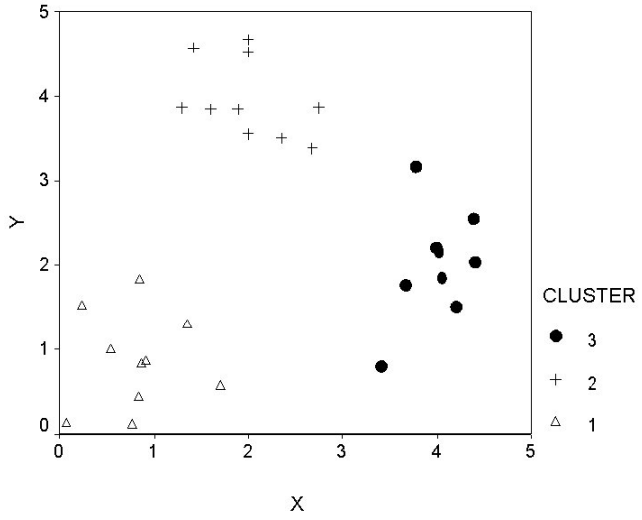


图 *44:* 简单聚类模型

提供了下列聚类方法：

K-Means 节点将数据集聚类到不同分组（或聚类）。 此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。 *k*-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。

TwoStep 节点使用二阶聚类方法。 第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。 第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。 TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。 它可以高效处理混合的字段类型和大型的数据集。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。 您可以通过查看模型块 中每个单元所捕获观测值的数量来找出规模较大的单元。 这将让您对聚类的相应数量有所估计。

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非监督学习来查找数据集的聚类或密集区域。 SPSS Modeler 中的 HDBSCAN 节点公开 HDBSCAN 库的核心特征和常用参数。 此节点以 Python 实现，当您一开始不了解数据集的分组时，可以使用此节点将数据集聚类为不同的组。

通常使用聚类模型来创建聚类或段，然后将聚类或段用作后续分析的输入。 常见例子如营销人员常使用市场分段来将整个市场划分为多个类似的子组。 每个市场分段都有自己的特征，该特性将影响到针对该分段的市

场营销努力是否能取得成功。 如果您使用数据挖掘来优化市场营销战略，通常可以通过识别合适的市场分段和在预测模型中使用分段信息来显著改进模型。

Kohonen 节点

Kohonen 网络是一种执行聚类的神经网络类型，也称为 **knet** 或 **自组织映射** 。 如果在开始时没有分组的相关信息，那么可使用此类型的网络将数据集聚类到有明显区别的不同分组。 对记录进行分组，以便组或聚类中的记录趋于相似，而不同组中的记录则有所差异。

基本单元为**神经元**，并且它们分为两层：**输入层**和**输出层**（也称为**输出映射**）。 所有输入神经元都和所有输出神经元相连接，这些连接有与其相关的 **强度** 或 **权重** 。 训练过程中，每个单元会与所有其他单元进行竞争以“赢得”每条记录。

输出映射是神经元的二维网络（单元之间无连接）。

输入数据会显示在输入层，相应值将传播到输出层。 响应最强的输出神经元将称为 **胜利者** 并且会成为输入的结果。

最初的权重随机产生。 如果某个单元赢得一条记录，那么其权重（与其附近单元的权重一起统称为 **近邻** ）将作调整以尽可能地与此条记录的预测变量值的模式相匹配。 显示所有输入记录，并且权重将相应更新。 将重复此过程，直到变化非常小为止。 当进行训练时，网格单元的权重将作调整从而形成聚类的一个二维“映射”（所以会有术语**自组织映射**）。

此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异很大的记录则应彼此远离。

与 IBM SPSS Modeler 中的大多数学习方法不同的是，Kohonen 网络 不 使用目标字段。 这种没有目标字段的学习称为 **无监督学习** 。 Kohonen 网络试图揭示输入字段集中的模式，而不是预测结果。 通常，Kohonen 网络最终会形成几个汇总许多观测数据的单元（ **强** 单元），以及几个实际不对应任何观测数据的单元（ **弱** 单元）。 强单元（有时也包括网格中与其邻近的其他单元）代表可能的聚类中心。

Kohonen 网络的另一种用途是 **降维** 。 二维网格的空间特性可提供从 *k* 个原始预测变量到保留了原始预测变量中相似性关系的两个派生特征的映射。 在某些情况下，此方法的作用与因子分析或主成分分析的作用相同。

请注意，计算输出网格缺省大小的方法与 IBM SPSS Modeler 以前的版本相比已发生了变化。 通常，新方法将生成更小的输出层，这些输出层训练起来更快且通用性更强。 如果您发现使用缺省大小得到的结果不理想，可以尝试在“专家”选项卡上增加输出网格的大小。 有关更多信息，请参阅主题 第 173 页的『Kohonen 节点专家选项』。

需求。 要训练 Kohonen 网络，您需要一个或多个角色设置为输入的字段。 角色设置为目标、两者或无的字段将被忽略。

强度。 您不需要关于组成员资格的数据即可构建 Kohonen 网络模型。 您甚至不需要知道要寻找的组的个数。 Kohonen 网络刚开始会有大量的单元，随着训练的进行，这些单元会向数据中的自然聚类集中。 可通过查看模型块中每个单元捕获的观测值数来识别强单元，进而了解适当的聚类数。

Kohonen 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

继续训练现有模型。 缺省情况下，每次执行 Kohonen 节点时，都会创建一个全新的网络。 如果选中此选项，那么会继续训练该节点成功生成的最后一个网络。

显示反馈图。 如果选中此选项，那么将在训练期间显示二维数组的直观表示。 每个节点的强度用颜色表示。 红色表示聚集了许多记录的单元（ **强** 单元），白色表示聚集的记录较少或没有记录的单元（ **弱** 单元）。 如果构建模型所花费的时间相对较短，可能不会显示反馈。 注意，此功能会减慢训练进度。 要加快训练进度，请取消选中此选项。

停止于。 缺省中止条件将根据内部参数中止训练。 也可以指定时间作为中止条件。 以分钟为单位输入网络训练的时间。

设置随机种子值。 如果未设置随机种子，那么每次执行节点时，用于初始化网络权重的随机值的序列都是不同的。 这将导致即使节点设置和数据值都完全相同，节点也会在不同的运行中创建不同的模型。 通过选择该选项，可以将随机种子设置为特定值，从而使结果模型具有精确的可再现性。 特定的随机种子通常会生成相同的随机值序列，在这种情况下执行节点通常会产生相同的生成模型。

注: 对从数据库中读取的记录使用**设置随机种子**选项时，可能需要在抽样前使用“排序”节点以确保每次执行节点时都获得相同的结果。 这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。

注：如果要在模型中包括名义（集合）字段，但在构建模型时遇到内存问题，或者构建模型所需的时间过长，那么可以考虑对大型集合字段进行重新编码以减少值的数量，或者考虑使用包含较少值的其他字段作为该大型集合的代理。 例如，如果包含个别产品值的 *product_id* 字段出现问题，可以考虑将其从模型中删除并改为添加信息不是很详细的 *product_category* 字段代替。

优化。 根据您的具体需求，选择旨在提高模型构建性能的选项。

- 选择 **速度** 可指示算法从不使用磁盘溢出，以便提高性能。
- 选择 **内存** 可指示算法在合适的时候，以牺牲某些速度为代价使用磁盘溢出。 缺省情况下，此选项处于选中状态。

  注: 以分布式方式运行时，`options.cfg` 中指定的管理员选项可能会覆盖此设置。

附加集群标签。 缺省对新模型选中此选项，但对从较早版本的 IBM SPSS Modeler 加载的模型取消选中。该选项会创建一个由 K-Means 和“二阶聚类”节点共同创建的相同类型的分类评分字段。 在计算不同模型类型的排秩测量时，该字符串字段用于“自动聚类”节点。 有关更多信息，请参阅主题 第 54 页的『自动聚类节点』。

Kohonen 节点专家选项

对于对 Kohonen 网络有详尽了解的用户，可使用专家选项对训练过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

宽度和长度。 将二维输出图的大小（宽度和长度）指定为每个维上的输出单元数。

学习速率衰减。 选择线性或指数学习速率衰减。 **学习速率** 是随时间递减的加权因子，使得网络可以从数据的大尺度特征开始进行编码，然后逐渐集中于更细微的数据信息。

阶段 1 和阶段 2。 Kohonen 网络训练分为两个阶段。 阶段 1 是粗略估计阶段，用于捕获数据中的大致模式。 阶段 2 是调整阶段，用于调整图以便为数据更精细的特征建模。 每个阶段都有以下三个参数：

- **近邻**。 设置近邻的起始大小（半径）。 此参数确定在训练期间与赢得单元一起被更新的“邻近”单元数。在阶段 1，近邻大小以 阶段 *1* 近邻为起始值，然后减少到（ 阶段 *2* 近邻 + 1）。 在阶段 2，近邻大小起始为 阶段 *2* 近邻，然后减少到 1.0。 阶段 *1* 近邻应该大于阶段 *2* 近邻。
- **初始 Eta**。 设置学习速率 **eta** 的起始值。 在阶段 1，eta 起始为 阶段 *1* 初始 *Eta*，然后减少到 阶段 *2* 初始 *Eta*。 在阶段 2 期间，eta 从阶段 *2* 初始 *Eta* 开始，并递减到 0。 阶段 *1* 初始 *Eta* 应该比 阶段 *2* 初始 *Eta* 大。
- **周期**。 为训练的每个阶段设置周期数。 每个阶段均会进行指定次数的数据处理。

Kohonen 模型块

Kohonen 模型块包含由经过训练的 Kohonen 网络捕获的所有信息，还包含有关网络体系结构的信息。

当运行包含 Kohonen 模型块的流时，节点将添加两个新字段，这两个字段包含 Kohonen 输出网格中对该记录反应最强烈的单元的 *X* 坐标和 *Y* 坐标。 新字段名称源自模型名称，前缀 *$KX-* 和 *$KY-*。例如，如果模型命名为 *Kohonen*，那么新字段将命名为 *$KX-Kohonen* 和 *$KY-Kohonen*。

为了更好地了解 Kohonen 网络编码的内容，可单击模型块浏览器上的“模型”选项卡。 此时会显示聚类浏览器，提供聚类、字段和重要性等级的图形表示。 有关更多信息，请参阅主题 第 183 页的『聚类查看器 - 模型选项卡』。

如果要将聚类可视化为网格，那么可以通过使用图节点绘制 *$KX-* 和 *$KY-* 字段来查看 Kohonen 网络的结果。 （应在散点图节点中选择 **X-Agitation** 和 **Y-Agitation** 以防止每个单元的记录彼此覆盖。） 在散点图中，也可以重叠符号字段以调查 Kohonen 网络是如何聚类数据的。

深入了解 Kohonen 网络的一种强大的方法是使用规则归纳法，确定用于区分网络所发现聚类的特征。 请参阅主题第 76 页的『C5.0 节点』，以获取更多信息。

有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』

Kohonen 模型摘要

Kohonen 模型块的“摘要”选项卡显示有关网络的体系结构或拓扑结构的信息。 二维 Kohonen 特征图（输出层）的长度和宽度显示为 **$KX-** *model_name* 和 **$KY-** *model_name*。 对于输入层和输出层，将列出该层的单元数。

K-Means 节点

K-Means 节点提供一种进行 **聚类分析** 的方法。 它可以用于在最初不知道有哪些组时，将数据集聚类为不同的组。 与 IBM SPSS Modeler 中的大多数学习方法不同的是，K-Means 模型 不 使用目标字段。 这种没有目标字段的学习称为 **无监督学习** 。 K-Means 模型试图揭示输入字段集的模式而不是预测结果。 记录会进行分组，以使一个组或聚类中的记录彼此相似，而不同组中的记录则互不相同。

K-Means 的工作原理是根据数据定义一组起始聚类中心。 然后根据记录的输入字段值，将每条记录分配到与其最相似的聚类中。 在分配完所有记录后，更新聚类中心以反映分配到每个聚类的新记录集。 然后再次检查记录，以确定是否应将这些记录重新分配到不同的聚类中，这个记录分配/聚类迭代过程将一直持续，直到达到最大迭代次数或一次迭代与下次迭代之间的改变不超过指定阈值为止。

注：生成的模型在一定程度上取决于训练数据的顺序。 重排数据顺序并重建模型有可能会生成不同的最终聚类模型。

需求。 要训练 K-Means 模型，您需要一个或多个角色设置为输入的字段。 角色设置为输出、两者或无的字段将被忽略。

强度。 您不需要关于组成员资格的数据即可构建 K 平均值模型。 通常，K-Means 模型是进行大型数据集聚类的最快方法。

K-Means 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

指定的聚类数。 指定要生成的聚类数。 缺省值是 5。

生成距离字段。 如果选中此选项，那么模型块将包括一个字段，该字段包含每条记录与所分配到的聚类的中心之间的距离。

聚类标签。 为生成的聚类成员资格字段的值指定格式。 可将聚类成员资格指示为具有指定**标签前缀**的**字符串**（例如 `"Cluster 1"`、`"Cluster 2"`等），也可指示为 **数字**。

注：如果要在模型中包括名义（集合）字段，但在构建模型时遇到内存问题，或者构建模型所需的时间过长，那么可以考虑对大型集合字段进行重新编码以减少值的数量，或者考虑使用包含较少值的其他字段作为该大型集合的代理。 例如，如果包含个别产品值的 *product_id* 字段出现问题，可以考虑将其从模型中删除并改为添加信息不是很详细的 *product_category* 字段代替。

优化。 根据您的具体需求，选择旨在提高模型构建性能的选项。

- 选择 **速度** 可指示算法从不使用磁盘溢出，以便提高性能。
- 选择 **内存** 可指示算法在合适的时候，以牺牲某些速度为代价使用磁盘溢出。 缺省情况下，此选项处于选中状态。

  注: 以分布式方式运行时，`options.cfg` 中指定的管理员选项可能会覆盖此设置。

K-Means 节点专家选项

对于对 *k*-means 聚类有详尽了解的用户，可使用专家选项对训练过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

停止于。 指定训练模型时要使用的中止条件。 **缺省**停止标准是 20 次迭代或变化 < 0.000001，以首先发生者为准。 选中**定制**可指定自己的停止标准。

- **最大迭代次数**。 使用此选项可在指定的迭代次数后停止模型训练。
- **更改容差**。 通过此选项，您可以在某次迭代的聚类中心中的最大差异小于指定的级别时中止模型训练。

集合编码值。 指定 0 到 1.0 之间的值，以用于将集合字段重新编码为数字字段组。 缺省值是 0.5 的平方根（大约为 0.707107），它可为重新编码的标志字段提供适当的加权。 值越接近 1.0，对集合字段的加权就越高于对数值字段的加权。

K-Means 模型块

K-Means 模型块包含由聚类模型捕获的所有信息，还包含有关训练数据和估计过程的信息。

当运行包含 K-Means 建模节点的流时，该节点将添加两个新字段，这两个字段包含聚类成员资格以及与该记录所分配到的聚类中心的距离。 新字段名得自模型名称，即为聚类成员资格加上 *$KM-* 前缀，为与聚类中心的距离加上 *$KMD-* 前缀。 例如，如果模型名称为 *Kmeans*，那么新字段的名称应是 *$KM-Kmeans* 和 *$KMD-Kmeans*。

深入了解 K-Means 模型的一种强大的方法是使用规则归纳法，确定用于区分模型所发现聚类的特征。 请参阅主题第 76 页的『C5.0 节点』，以获取更多信息。 您也可以单击模型块浏览器上的“模型”选项卡，以显示聚类浏览器，其中会提供聚类、字段和重要性级别的图形表示。 有关更多信息，请参阅主题第 183 页的『聚类查看器 - 模型选项卡』。

有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』

K 平均值模型摘要

K 平均值模型块的“摘要”选项卡包含有关训练数据、估计过程和由模型定义的聚类的信息。 显示的信息有聚类数，还有迭代历史记录。 如果已执行附加到此建模节点的分析节点，那么分析信息也将显示在此选项卡上。

“二阶聚类”节点

“二阶聚类”节点提供一种形式的**聚类分析**。 它可以用于在最初不知道有哪些组时，将数据集聚类为不同的组。 与 Kohonen 节点和 K-Means 节点一样，“二阶聚类”模型也不使用目标字段。 二阶聚类模型试图揭示输入字段集的模式而不是预测结果。 记录会进行分组，以使一个组或聚类中的记录彼此相似，而不同组中的记录则互不相同。

二阶聚类是一种分两步进行聚类的方法。 第一步对数据进行一次遍历，在这个过程中，将原始输入数据压缩为一组容易处理的子聚类。 第二步采用分层聚类方法，将这些子聚类逐渐合并成越来越大的聚类，在此过程中无需再次遍历数据。 分层聚类的优点在于不需要事先选择聚类数。 许多分层聚类方法一开始将单个的记录作为最初的聚类，然后递归合并这些记录以生成更大的聚类。 虽然此类方法常因数据量巨大而失败，但二阶聚类的初始预聚类会使分层聚类的速度非常快，即使数据集巨大也是如此。

注: 生成的模型在一定程度上取决于训练数据的顺序。 重排数据顺序并重建模型有可能会生成不同的最终聚类模型。

需求。 要训练“二阶聚类”模型，您需要一个或多个角色设置为输入的字段。 角色设置为目标、两者或无的字段将被忽略。 二阶聚类算法不处理缺失值。 构建模型时将忽略任意输入字段包含空白的记录。

强度。 二阶聚类可以处理混合字段类型，并能高效处理大型数据集。 它还能检验多个聚类解并选择其中最佳的解，因此您开始时不必知道应有多少个聚类。 可将“二阶聚类”设置为自动排除**离群值**或能对结果造成损害的极其异常情况。

IBM SPSS Modeler 有两个不同版本的“二阶聚类”节点：

- **二阶聚类**是在 IBM SPSS Modeler Server 上运行的传统节点。
- 连接到 IBM SPSS Analytic Server 之后，可以运行**二阶 AS 聚类**。

二阶聚类节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

标准化数字字段。 缺省情况下， TwoStep 会将所有数字输入字段标准化为同一比例，平均值为 0，方差为 1。 要保留数字字段的原始比例，请取消选择此选项。 符号字段不受影响。

排除离群值。 如果选中此选项，那么那些与主要聚类似乎格格不入的记录将自动排除在分析之外。 这样可以防止此类情况歪曲结果。

离群值检测在预聚类步骤进行。 选中此选项时，会将相对于其他子聚类具有较少记录的子聚类视为潜在离群值，且重新构建不包括这些记录的子聚类树。 子聚类被视为包含潜在离群值的下限大小由**百分比**选项控制。 如果其中某些潜在离群值记录与任何新子聚类配置足够相似，那么可将其添加到重新构建的子聚类中。 将其余无法合并的潜在离群值视为离群值添加到“噪声”聚类中并排除在分层聚类步骤之外。

使用经过离群值处理的“二阶”模型对数据进行评分时，会将与最近主要聚类的距离大于特定阈值距离（基于对数似然）的新观测值视为离群值分配到“噪声”聚类中，名称为 -1。

聚类标签。 为生成的聚类成员资格字段指定格式。 可将聚类成员资格指示为具有指定**标签前缀**的**字符串**（例如 "Cluster 1"、"Cluster 2"等），也可指示为**数字**。

自动计算聚类数。 “二阶聚类”可以非常迅速地对大量聚类解进行分析并为训练数据选择最佳聚类数。 通过设置**最大**和**最小**聚类数来指定要尝试的解决方案的范围。 “二阶聚类”通过一个两阶段过程确定最佳聚类数。 在第一个阶段，随着所添加聚类的增多，可基于贝叶斯信息准则 (BIC) 中的差异选择模型中聚类数的上限。 在第二个阶段，为聚类数比最小 BIC 解决方案还少的所有模型找出聚类间最小距离的差异。 距离的最大差异用于标识最终聚类模型。

指定聚类数。 如果知道模型中要包含的聚类的数目，请选中此选项并输入聚类数。

距离测量。 此选项确定如何计算两个聚类之间的相似性。

- **对数似然**。 该似然度量假设变量服从某种概率分布。 假设连续变量是正态分布，而假设分类变量是多项分布。 假设所有变量均是独立的。
- **欧式**。 欧几里德距离测量是两个聚类之间的“直线”距离。 它只能用于所有变量连续的情况。

聚类标准。 此选项确定自动聚类算法如何确定聚类数。 可以指定 Bayesian 信息准则 (BIC) 或 Akaike 信息准则 (AIC)。

二阶聚类模型块

二阶聚类模型块包含由聚类模型捕获的所有信息，还包含有关训练数据和估计过程的信息。

当运行包含“二阶聚类”模型块的流时，节点将为该记录添加包含聚类成员资格的新字段。 新字段名称派生自模型名称，并以 *$T-* 为前缀。例如，如果您的模型名为 *TwoStep*，那么新字段将命名为 *$T-TwoStep*。

深入了解 TwoStep 模型的一种强大的方法是使用规则归纳法，确定用于区分模型所发现聚类的特征。 请参阅主题 第 76 页的『C5.0 节点』 以获取更多信息。 您也可以单击模型块浏览器上的“模型”选项卡，以显示聚类浏览器，其中会提供聚类、字段和重要性级别的图形表示。 有关更多信息，请参阅主题 第 183 页的『聚类查看器 - 模型选项卡』。

有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』

两步模型摘要

二阶聚类模型块的“摘要”选项卡显示找出的聚类数以及有关训练数据、估计过程和所使用的构建设置的信息。

有关更多信息，请参阅主题 第 31 页的『浏览模型块』。

二阶 AS 聚类节点

IBM SPSS Modeler 有两个不同版本的“二阶聚类”节点：

- 二**阶聚类**是在 IBM SPSS Modeler Server 上运行的传统节点。
- 连接到 IBM SPSS Analytic Server 之后，可以运行二**阶 AS 聚类**。

二阶 AS 聚类分析

“二阶聚类”是一个探索工具，用于揭示数据集中原本不明显的自然分组（即聚类）。 此过程使用的算法有多个不错的特征使其有别于传统聚类技术：

- **分类变量和连续变量的处理**。 通过假设变量是独立的，可以假设分类变量和连续变量服从联合多项正态分布。
- **自动选择聚类的数量**。 通过跨不同的聚类解比较模型选择准则的值，此过程可以自动确定最优的聚类数。
- **可伸缩性**。 通过构造对记录进行摘要的聚类特征 (CF) 树，二阶算法能够分析大型数据文件。

例如，零售和消费者产品公司定期地对描述客户的购买习惯、性别、年龄、收入水平和其他属性的信息应用聚类技术。 这些公司针对每个消费者群体定制其市场营销和产品开发战略，以提高销售量并建立品牌忠诚度。

“字段”选项卡

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 选中所有具有已定义的“输入”角色的字段。

使用定制字段分配。 添加和移除字段，而不考虑对其定义的角色分配。 您可以选择具有任意角色的字段，并将其移入或移出**预测变量（输入）**列表。

基本

聚类数

自动确定
此过程确定指定范围内的最佳聚类数。 **最小值**必须大于 1。 这是默认选项。

指定固定值
此过程生成指定的聚类数。 **数目**必须大于 1。

聚类条件

此选项控制自动聚类算法如何确定聚类数。

贝叶斯信息准则 (BIC)
用于根据 -2 对数似然选择和比较模型的度量。 值越小，表示模型拟合得越好。 BIC 也会“惩罚”过多参数模型（例如，具有大量输入的复杂模型），但比 AIC 更严格。

赤池信息准则 (AIC)
用于根据 -2 对数似然选择和比较模型的度量。 值越小，表示模型拟合得越好。 AIC“惩罚”过多参数模型（例如，具有大量输入的复杂模型）。

自动聚类方法

如果您选择了**自动确定**，请从下面用于自动确定聚类数的聚类方法中选择：

使用聚类条件设置
信息条件收敛是对应于两个当前聚类解的信息条件与第一个聚类解的比率。 所使用的条件是在“聚类条件”组中选择的条件。

距离跳转
距离跳转是与两个连续聚类解相对应的距离的关系。

最大值
对信息条件收敛法的结果和距离跳跃法的结果进行组合，以生成与第二次跳跃相对应的聚类数。

最小值
对信息条件收敛法的结果和距离跳跃法的结果进行组合，以生成与第一次跳跃相对应的聚类数。

特征重要性方法

特征重要性方法确定特征（字段）在聚类解中的重要性。 输出包含有关整体特征重要性和每个聚类中的每个特征字段的重要性的信息。 将排除不满足最小阈值的特征。

使用聚类条件设置
这是缺省方法，此方法基于在“聚类条件”组中选择的条件。

效应大小
特征重要性基于效应大小而不是显著性值。

特征树条件

这些设置确定如何构建聚类特征树。 通过构建聚类特征树并对记录进行摘要，二阶算法能够分析大型数据文件。 换而言之，二阶聚类使用聚类特征树来构建聚类，从而使其能够处理许多观测值。

距离测量

此选项确定如何计算两个聚类之间的相似性。

对数似然

似然测量假设字段服从某种概率分布。 假设连续字段呈正态分布，而假设分类字段呈多项式分布。 假设所有字段都是独立的。

欧几里德

欧几里德距离测量是两个聚类之间的“直线”距离。 使用平方欧几里德距离测量和 Ward 法来计算聚类之间的相似性。 仅当所有字段都是连续字段时，才能使用此测量。

离群值聚类

包括离群值聚类
包括作为常规聚类离群值的观测值的聚类。 如果未选中此选项，那么所有观测值都将包括在常规聚类中。

特征树叶中的观测值数小于
如果特征树叶中的观测值数小于指定的值，那么将该叶片视为离群值。 该值必须是大于 1 的整数。 如果您更改此值，那么较高的值可能会导致更多界外值聚类。

离群值的最高百分比
构建聚类模型时，离群值将按离群值强度进行排名。 进入离群值主要百分比所需的离群值强度作为阈值，用于确定是否将观测值分类为离群值。 较大的值意味着将较多的观测值分类为离群值。 此值必须介于 1 与 100 之间。

其他设置

初始距离变动阈值

这是用于使聚类特征树增长的初始阈值。 如果将叶片插入到树中的叶片后，所产生的紧性小于此阈值，那么该叶片将不再拆分。 如果紧性超过此阈值，那么该叶片将进行拆分。

叶节点最大分支数

叶节点可以具有的最大子节点数。

非叶节点最大分支数

非叶节点可以具有的最大子节点数。

最大树深度

聚类树可以具有的最大级别数。

测量级别的调整权重
通过提高连续字段的权重降低分类字段的影响。 此值表示用于降低分类字段权重的分母。 例如，缺省值 6 使分类字段的权重为 1/6。

内存分配

聚类算法使用的最大内存量，以兆字节 (MB) 计。 如果此过程使用的空间量超过此最大值，那么将使用磁盘来存储内存中放不下的信息。

延迟拆分
延迟聚类特征树的重建。 聚类算法在评估新观测值时，将多次重建聚类特征树。 此选项将延迟该操作并减少重建该树的次数，从而提高性能。

标准化

聚类算法处理已标准化的连续字段。 缺省情况下，所有连续字段都已标准化。 为了节省部分时间和计算工作，您可以将已标准化的连续字段移到**不标准化**列表。

特征选择

在“特征选择”屏幕上，可以设置规则以确定何时排除字段。 例如，可以排除包含许多缺失值的字段。

用于排除字段的规则

缺失值百分比大于
在分析中，将排除缺失值百分比大于指定值的字段。 此值必须是大于零且小于 100 的正数。

分类字段的类别数大于
在分析中，将排除类别数大于指定数目的分类字段。 此值必须是大于 1 的正整数。

趋向于单个值的字段

连续字段的变异系数小于
在分析中，将排除变异系数小于指定值的连续字段。 变异系数是标准差与均值之比。 较小的值常常表示这些值的变异程度较小。 此值必须在 0 到 1 之间。

分类字段的单个类别中的观测值百分比大于
在分析中，将排除单个类别中的观测值百分比大于指定值的分类字段。 值必须大于 0 且小于 100。

自适应特征选择

此选项将运行额外的数据遍历，以查找并移除最不重要的字段。

模型输出

模型构建汇总

模型规范
模型规范数、最终模型中的聚类数以及最终模型中包含的输入数（字段数）的摘要。

记录摘要
模型中包括和排除的记录（观测值）的数目和百分比。

排除的输入
对于任何未包括在最终模型中的字段，显示字段被排除的原因。

评估

模型质量
这个表显示每个聚类的优度和重要性以及整体模型拟合度。

特征重要性条形图
这个条形图显示特征（字段）在所有聚类中的重要性。 在条形图中，条形较长的特征（字段）比条形较短的特征（字段）更为重要。 特征（字段）还按重要性以降序排序（最前面的条形最重要）。

特征重要性字云
这个字云显示特征（字段）在所有聚类中的重要性。 文本较大的特征（字段）比文本较小的特征（字段）更为重要。

离群值聚类
如果您选择不包括离群值，那么这些选项将处于禁用状态。

交互式表和图表
这个表和图表显示离群值强度以及离群值聚类与常规聚类的相对相似性。 在表中选择不同的行，将会在图表中显示不同离群值聚类的信息。

透视表
这个表显示离群值强度以及离群值聚类与常规聚类的相对相似性。 这个表与交互式显示包含相同的信息。 这个表支持所有的标准透视表功能和编辑表功能。

最大数
输出中要显示的最大离群值数。 如果有超过 20 个离群值聚类，将会改为显示透视表。

解释

聚类间的特征重要性概要文件

交互式表和图表。
这些表和图表显示聚类解中使用的每个输入（字段）的特征重要性和聚类中心。 在表中选择不同的行将会显示一个不同的图表。 对于分类字段，显示条形图。 对于连续字段，将显示平均值和标准偏差的图表。

透视表。
这个表显示每个输入（字段）的特征重要性和聚类中心。 这个表与交互式显示包含相同的信息。 这个表支持所有的标准透视表功能和编辑表功能。

聚类中的特征重要性
对于每个聚类，显示每个输入（字段）的聚类中心和特征重要性。 每个聚类都有一个单独的表。

聚类距离
这个面板图表显示聚类之间的距离。 每个聚类都有一个单独的面板。

聚类标签

Text
每个聚类的标签由为**前缀**指定的值和后跟的序列号组成。

数字
每个聚类的标签是一个序列号。

模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

两阶 AS 聚类模型块

两阶 AS 模型块显示输出查看器的“模型”选项卡中的模型的详细信息。 有关使用此查看器的更多信息，请参阅《Modeler 用户指南》(ModelerUsersGuide.pdf) 中标题为『处理输出』的部分。

两阶 AS 聚类模型块包含由聚类模型捕获的所有信息，以及有关训练数据和估算过程的信息。

运行包含两阶 AS 聚类模型块的流时，节点将为该记录添加包含聚类成员资格的新字段。 新字段名称派生自模型名称，并以 ***$AS-*** 为前缀。例如，如果模型命名为 TwoStep，那么新字段将命名为 ***$AS-TwoStep***。

深入了解 TwoStep-AS 模型的一种强大的方法是使用规则归纳法，确定用于区分模型所发现聚类的特征。 请参阅主题第 76 页的『C5.0 节点』，以获取更多信息。

有关使用模型浏览器的一般信息，请参阅第 31 页的『浏览模型块』。

注: 在 TwoStep-AS 模型查看器中的**评估 > 模型质量**部分下，将显示每个聚类的记录数。 如果您连接“分布”节点以计算评分结果，那么可能会发现每个聚类中的记录数与您在**模型质量**部分下看到的记录数不同。 这将发挥应有的作用。 从算法的角度来看，评估结果来自于一个分层聚类过程，而评分表则来自于直接将数据案例与最终聚类的分布进行比较。 如果聚类模型不完善，那么这两个不同的评分过程可能会产生不同的结果。 但在大多数情况下，差异很小。

二阶-AS 聚类模型块设置

“设置”选项卡为二阶-AS 模型块提供了额外的选项。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **通过转换至本机 SQL 来进行评分**：如果选择此项，将生成本机 SQL 在数据库中对模型进行评分。

  注: 虽然该选项可以更快速获得结果，但是本机 SQL 的大小和复杂性会随着模型复杂性的增加而增加。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

K-Means-AS 节点

K-Means 是最常用的聚类算法之一。 它将数据点聚类到预定义数量的 聚类中。[1] SPSS Modeler 中的 K-Means-AS 节点使用 Spark 进行实现。

有关 K-Means 算法的详细信息，请参阅 https://spark.apache.org/docs/2.2.0/ml-clustering.html。

请注意，K-Means-AS 节点自动对分类变量执行独热编码。

[1] “聚类。” *Apache Spark*. MLlib: Main Guide. Web. 2017 年 10 月 3 日。

K-Means-AS 节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 该选项通知节点使用来自上游类型节点的字段信息。 缺省情况下，此选项处于选定状态。

使用定制字段分配。 如果要手动分配输入字段，请选择此选项，然后选择一个或多个输入字段。 使用此选项类似于在“类型”节点中将字段角色设置为**输入**。

K-Means-AS 节点构建选项

使用“构建选项”选项卡可以指定 K-Means-AS 节点的构建选项，包括用于模型构建的常规选项、用于初始化聚类中心的初始化选项以及用于计算迭代和随机种子的高级选项。 有关更多信息，请参阅 JavaDoc for K-Means on SparkML。[1]

常规

模型名称。 对特定聚类评分后生成的字段的名称。 选择**自动**（缺省）或选择**定制**并输入名称。

聚类数。 指定要生成的聚类数。 缺省值为 **5**，最小值为 **2**。

初始化

初始化方式。 指定用于初始化聚类中心的方法。 缺省值为 **K-Means||**。 有关这两种方法的详细信息，请参阅可扩展 K-Means + +。[2]

初始化步骤。 如果选择 **K-Means||** 初始化模式，请指定初始化步骤数。 缺省值为 **2**。

高级

高级设置。 如果要按如下设置高级选项，请选择此选项。

最大迭代次数。 指定在搜索聚类中心时要执行的最大迭代次数。 缺省值为 **20**。

容差。 指定迭代算法的汇合容差。 缺省值为 **1.0E-4**。

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

显示

显示图形。 如果要在输出中包含图形，请选择此选项。

下表显示 SPSS Modeler K-Means-AS 节点中的设置与 K-Means Spark 参数之间的关系。

表 *13:* 映射到 *Spark* 参数的节点属性

SPSS Modeler 设置	脚本名称（属性名称）	K-Means SparkML 参数
输入字段	`features`	
聚类数	`clustersNum`	`k`
初始化模式	`initMode`	`initMode`
初始化步骤数	`initSteps`	`initSteps`
最大迭代次数	`maxIter`	`maxIter`
容差	`toleration`	`tol`
随机种子	`randomSeed`	`seed`

[1] “Class KMeans。” *Apache Spark*. JavaDoc。 Web. 2017 年 10 月 3 日。

[2] Bahmani, Moseley 等著，“Scalable K-Means++。”2012 年 2 月 28 日。http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf。

聚类查看器

聚类模型通常用于根据所检查变量查找具有类似记录的组（聚类），其中同组成员间的相似性高而不同组成员间的相似性低。 结果可用于识别原本不明显的关联。 例如，通过对客户偏好、收入水平和购物习惯的聚类分析，可以识别出对某种市场营销活动更可能做出反应的客户类型。

有两种方法可以解释聚类显示中的结果：

- 检查聚类以确定该聚类的唯一特征。 是否有一个聚类包含所有高收入借款人？ 此聚类是否包含比其他聚类更多的记录？
- 检查各聚类上的字段以确定值在聚类间的分布情况。 个人的教育水平是否决定其在聚类中的成员资格？ 高信用得分是否在一个聚类或另一个聚类的成员资格之间加以区分？

使用“聚类查看器”中的主视图和各个链接视图，可以清楚回答这些问题。

可在 IBM SPSS Modeler 中生成以下聚类模型块：

- Kohonen 网络模型块
- K-Means 模型块
- 二阶聚类模型块

要查看有关聚类模型块的信息，右键单击模型节点并从上下文菜单中选择**浏览**（或选择流中节点的**编辑**）。 或者，如果您正使用“自动聚类”建模节点，双击“自动聚类”模型块中的所需聚类模型块。 有关更多信息，请参阅主题 第 54 页的『自动聚类节点』。

聚类查看器 - 模型选项卡

聚类模型的“模型”选项卡显示各聚类之间字段的汇总统计和分布的图形显示，也称为**聚类查看器**。

注：“模型”选项卡对于使用 IBM SPSS Modeler 13 之前版本构建的模型不可用。

“聚类查看器”包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。 有两个主视图：

- 模型摘要（缺省视图）。 有关更多信息，请参阅主题 第 183 页的『模型摘要视图』。
- 聚类。 有关更多信息，请参阅主题 第 183 页的『聚类视图』。

有四个链接/辅助视图：

- 预测变量的重要性. 有关更多信息，请参阅主题 第 184 页的『聚类预测变量重要性视图』。
- 聚类大小（缺省视图）。 有关更多信息，请参阅主题 第 184 页的『聚类大小视图』。
- 单元格分布。 有关更多信息，请参阅主题 第 185 页的『单元格分布视图』。
- 聚类比较。 有关更多信息，请参阅主题 第 185 页的『聚类比较视图』。

模型摘要视图

“模型摘要”视图显示聚类模型的快照或摘要，包括加阴影以表示结果较差、尚可或良好的聚类结合和分离的 Silhouette 测量。 该快照可让您快速检查质量是否较差，如果较差，您可返回建模节点修改聚类模型设置以生成较好的结果。

结果较差、尚可和良好是基于 Kaufman 和 Rousseeuw (1990) 关于聚类结构解释的研究成果来判定的。 在“模型摘要”视图中，良好的结果表示数据将 Kaufman 和 Rousseeuw 的评级反映为聚类结构的合理迹象或强迹象，尚可的结果将其评级反映为弱迹象，而较差的结果将其评级反映为无明显迹象。

针对所有记录计算 (B-A) / max(A,B) 的平均值，其中 A 是记录与其聚类中心的距离，而 B 是记录与非所属最近聚类中心的距离。 Silhouette 系数为 1 表示所有观测值直接位于其聚类中心上。 值为 -1 表示所有个案都位于另外某些聚类的聚类中心。 值为 0 表示在正常情况下观测值到其自身聚类中心与到最近其他聚类中心是等距的。

摘要所包含的表格具有以下信息：

- **算法**。 所使用的聚类算法，例如“二阶”。
- **输入功能部件**。 字段数量，也称为**输入**或**预测变量**。
- **聚类**。 解中聚类的数量。

聚类视图

“聚类”视图包含一个聚类-特征网格，其中包括每个聚类的名称、大小和概要文件。

网格中的列包含以下信息：

- **集群**。 算法生成的聚类编号。
- **标签**。 应用于每个聚类的任何标签（缺省为空白）。 双击单元格输入描述聚类内容的标签，例如“豪华汽车买家”。
- **描述**。 聚类内容的任何描述（缺省为空白）。 双击单元格输入聚类描述；例如“年龄超过 55 岁、专业人员、收入超过 100,000 美元”。
- **大小**。 每个聚类的大小，表示为总体聚类样本的百分比。 网格中的每个大小单元格显示一个垂直条，其中显示聚类中的大小百分比、数值格式的大小百分比和聚类个案计数。
- **特征**。 单个输入或预测变量，缺省按总体重要性排序。 如果有列的大小相等，那么其以聚类编号的升序显示。

  总体特征重要性由单元格背景阴影的颜色表示；最重要的特征颜色最深；最不重要的特征则没有阴影。 表格上方的向导指示与每个特征单元格颜色关联的重要性。

当鼠标悬停在单元格上时，会显示特征的全名/标签和单元格的重要性值。 根据视图和特征类型，可能会显示其他信息。 在“聚类中心”视图中，这包括单元格统计量和单元格值；例如“Mean: 4.32”。 对于分类特征，单元格显示最常见（模态）类别的名称及其百分比。

在“聚类”视图中，您可以选择多种显示聚类信息的方式：

- 转置聚类和特征。 有关更多信息，请参阅主题 第 184 页的『变换聚类和特征』。
- 排序特征。 有关更多信息，请参阅主题 第 184 页的『排序特征』。
- 排序聚类。 有关更多信息，请参阅主题 第 184 页的『排序聚类』。
- 选择单元格内容。 有关更多信息，请参阅主题 第 184 页的『单元格内容』。

变换聚类和特征

缺省情况下，聚类显示为列，特征显示为行。 为翻转这种显示，单击**特征排序方式**按钮左侧的**变换聚类和特征**按钮。 例如，当显示许多聚类时，您可能想要进行此操作，以减少查看数据所需的水平滚动量。

排序特征

特征排序方式按钮可使您选择特征单元格的显示方式：

- **总体重要性**。 这是缺省的排序方式。 特征以总体重要性的升序进行排序，排序方式在各聚类间相同。 如果有特征具有同数重要性值，那么按照特征名称的升序列出同数特征。
- **聚类内重要性**。 特征按照其相对于每个聚类的重要性进行排序。 如果有特征具有同数重要性值，那么按照特征名称的升序列出同数特征。 当选中此选项时，排序顺序通常因聚类而异。
- **名称**。 特征按照名称的字母顺序进行排序。
- **数据顺序**。 特征按照其在数据集中的顺序进行排序。

排序聚类

缺省情况下，聚类按照大小的降序排序。 **聚类排序方式**按钮可使您按照名称的字母顺序对其进行排序，或如果您创建了唯一标签，那么按照标签的字母顺序对其进行排序。

具有相同标签的特征按照聚类名称排序。 如果聚类按照标签排序且您编辑了聚类的标签，那么自动更新排序顺序。

单元格内容

单元格按钮使您能够更改特征和评估字段的单元格内容的显示。

- **聚类中心**。 缺省情况下，单元格显示特征名称/标签和每个聚类/特征组合的集中倾向。 对于连续字段和具有分类字段的类别百分比的模式（最频繁出现的类别）显示均值。
- **绝对分布**。 显示特征名称/标签和每个聚类中特征的绝对分布。 对于类别特征，显示条形图，其中叠放了按数据值的升序排序的类别。 对于连续特征，显示平滑密度图，其对每个聚类使用相同的端点和间隔。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。

- **相对分布**。 显示特征名称/标签和单元格中的相对分布。 总体而言，显示类似于绝对分布的显示，不同之处在于所显示的是相对分布。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。

- **基本视图**。 如果聚类很多，不滚动很难看到所有详细信息。 要减少滚动量，选择此视图将显示更改为更紧凑的表格。

聚类预测变量重要性视图

“预测变量重要性”视图显示评估模型时每个字段的相对重要性。

聚类大小视图

“聚类大小”视图显示包含每个聚类的饼图。 每个聚类的百分比大小显示在每个分区上；鼠标悬停在每个分区上显示该分区中的计数。

图表下方的表格列出以下大小信息：

- 最小聚类的大小（总体计数和百分比）。
- 最大聚类的大小（总体计数和百分比）。
- 最大聚类与最小聚类的大小比率。

单元格分布视图

“单元格分布”视图显示您在“聚类”主面板的表格中选择的任意特征单元格数据分布的展开的详图。

聚类比较视图

“聚类比较”视图由网格式布局构成，行中为特征，列中为选定聚类。 此视图帮助您更好地理解组成聚类的因素；同时使您能够看到各聚类间的差异，不但与总体数据比较，而且还在彼此之间比较。

选择要显示的聚类，单击“聚类”主面板中聚类列的顶部。 使用 Ctrl+单击或 Shift+单击选择或取消选择多个聚类进行比较。

注：您可以选择最多五个要显示的聚类。

聚类以选择时的顺序显示，而字段顺序则由**特征排序方式**选项决定。 当您选择**聚类内重要性**时，将始终按总体重要性顺序排序字段。

背景图显示每个特征的总体分布：

- 类别特征显示为点图，其中点的大小代表每个聚类最频繁出现的（模态）类别（按特征）。
- 连续特征显示为箱图，其显示整体中位数和四分位距。

叠放在这些背景视图上的是所选聚类的箱图：

- 对于连续特征，方点标记和水平线表示每个聚类的中位数和四分位距。
- 每个聚类由不同颜色表示，显示在视图顶部。

浏览聚类查看器

“聚类浏览器”为交互式显示。 您可以：

- 选择字段或聚类以查看更多详细信息。
- 比较聚类以选择感兴趣的项目。
- 更改显示。
- 变换轴。
- 使用“生成”菜单生成“派生”节点、“过滤”节点和“选择”节点。

使用工具栏

您可使用工具栏选项控制在左右两侧面板中显示的信息。 您可使用工具栏控件更改显示的方向（从上至下、从左至右或从右至左）。 另外，您还可以将查看器重置为缺省设置，并打开对话框以在主面板中指定“聚类”视图的内容。

仅当您在主面板中选择**聚类**视图时，**特征排序方式**、**聚类排序方式**、**单元格**和**显示**选项才可用。 有关更多信息，请参阅主题 第 183 页的『聚类视图』。

表 *14:* 工具栏图标

图标	主题
	请参阅变换聚类和特征
	请参阅特征排序方式
	请参阅聚类排序方式

表 *14:* 工具栏图标 *(*继续*)*	
图标	**主题**
	请参阅单元格

从聚类模型生成节点

“生成”菜单可基于聚类模型新建节点。 可从生成模型的“模型”选项卡访问该选项，它可基于当前显示或选择（即所有可见聚类或所有选定聚类）生成节点。 例如，您可选择一个特征，然后生成“过滤”节点以丢弃所有其他（非可见）特征。 生成的节点放置在工作区上（未连接）。 另外，您还可以在模型调色板上生成模型块的副本。 记住，在执行之前连接节点并进行所需编辑。

- **生成建模节点**。 在流工作区上创建建模节点。 例如，如果您想在某个流中使用这些模型设置但您不再拥有用来生成这些设置的建模节点，该功能会很有用。
- **建模到选用板**。 在“模型”调色板上创建模型块。 当有同事发给您包含模型的流而不是模型本身时，该功能很有用。
- **“过滤”节点**。 创建新的“过滤”节点以过滤聚类模型不使用的过滤字段和/或当前“聚类查看器”显示中不可见的字段。 如果此聚类节点上游有“类型”节点，那么所生成的“过滤”节点会丢弃具有角色目标的任何字段。
- **“过滤”节点（来自选择）**。 基于“聚类查看器”中的选择创建用于过滤字段的新“过滤”节点。 使用 Ctrl+单击的方法选择多个字段。 在下游丢弃“聚类查看器”中选择的字段，但您可在执行之前通过编辑“过滤”节点更改此行为。
- **“选择”节点**。 创建新的“选择”节点以基于在当前“聚类查看器”显示中可见的任一聚类中的成员资格选择记录。 自动生成选择条件。
- **“选择”节点（来自选择）**。 创建新的“选择”节点以基于在“聚类查看器”中选择的聚类中的成员资格选择记录。 使用 Ctrl+单击的方式选择多个聚类。
- **“派生”节点**。 创建新的“派生”节点，其派生出标记字段，该字段基于“聚类查看器”中所有可见聚类的成员资格分配给记录 *True* 或 *False* 值。 自动生成派生条件。
- **“派生”节点（来自选择）**。 创建新的“派生”节点，该节点基于“聚类查看器”中选择的聚类中的成员资格派生出标记字段。 使用 Ctrl+单击的方式选择多个聚类。

除了生成节点之外，您还可以从“生成”菜单创建图形。 有关更多信息，请参阅主题 第 186 页的『从聚类模型生成图形』。

控制聚类视图显示

要控制主面板的聚类视图中显示的内容，单击**显示**按钮；打开“显示”对话框。

特征。 缺省选定。 要隐藏所有输入特征，取消选择该复选框。

评估字段。 选择要显示的评估字段（不用于创建聚类模型的字段，但被发送至模型查看器以评估聚类）；缺省不显示任何字段。 注：评估字段必须是包含多个值的字符串。 如果没有评估字段可用，那么此复选框不可用。

集群描述。 缺省选定。 要隐藏所有聚类描述单元格，取消选择该复选框。

聚类大小。 缺省选定。 要隐藏所有聚类大小单元格，取消选择该复选框。

最大类别数。 指定在类别特征图表中显示的最大类别数量；缺省值是 20。

从聚类模型生成图形

聚类模型提供许多信息，但其格式有时不便于业务用户访问。 要通过可以轻松纳入业务报告、简报等的方式提供数据，您可以生成所选数据的图形。 例如，可从“聚类浏览器”生成所选聚类的图形，这样可以只创建该聚类中观测值的图形。

注：仅当模型块附加到流中的其他节点时，您才能从聚类浏览器中生成图形。

生成图形

1. 打开包含“聚类浏览器”的模型块。

2. 在“模型”选项卡上，从**视图**下拉列表选择聚类。
3. 在主面板上，选择您要为其生成图形的一个或多个聚类。
4. 从“生成”菜单，选择**图形（从选择创建）**；显示“图形板基本”选项卡。

   注：通过此方式显示“图形板”时，仅“基本”和“详细”选项卡可用。
5. 使用“基本”或“详细”选项卡设置来指定要在图形上显示的详细信息。
6. 单击“确定”以生成图形。

图形标题标识模型类型和选择包含在内的一个或多个聚类。

第 12 章 关联规则

关联规则将特定结论（例如，购买特定产品）与一组条件（例如，购买多个其他产品）关联起来。 例如，规则

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

表述的是： 啤酒经常与 罐装蔬菜和 冷冻食品一起成对出现。 该规则可靠率为 84% 并适用于 17% 的数据或 173 条记录。 关联规则算法自动找到可使用可视方法（比如 Web 节点）手动找到的关联。

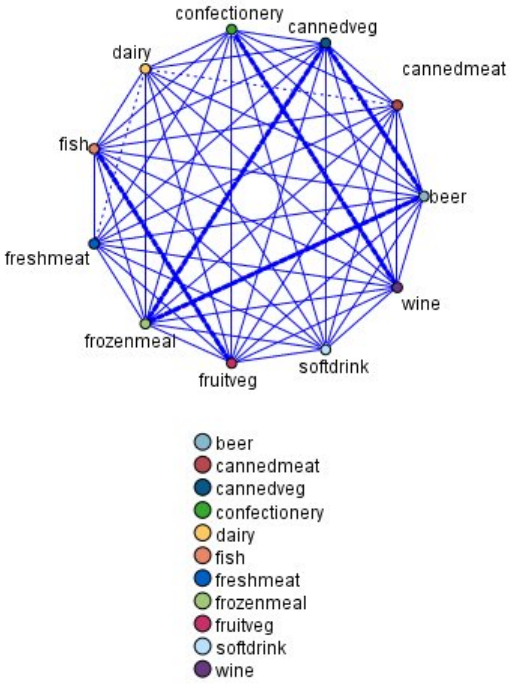


图 *45:* 显示市场购物篮商品之间的关联的网络节点

关联规则算法相较于更为标准的决策树算法（C5.0 和 C&R 树）的优势在于，关联可以存在于任何属性之间。 决策树算法只使用单一结论来构建规则，而关联算法则试图找到更多规则，且每个规则具有不同的结论。

关联算法的缺点是试图在可能非常大的搜索空间中查找规则，因而运行时间比决策树算法长得多。 关联算法使用 **生成与检验** 方法来查找规则（简单规则将初始生成）并对照数据集来验证这些规则。 好的规则会保存，根据各种限制，然后所有规则都会进行专业化处理。**专业化**是将条件添加到规则的过程。 然后这些新规则将对照数据进行验证，并且验证过程中将迭代保存最符合条件和最有用的规则。 用户通常会对允许进入规则的前提条件的可能的数量给出一定限制，并根据信息理论和高效索引方式使用各种方法来缩小原来可能很大的搜索空间。

处理结束后，将给出最符合条件的规则的列表。 此组关联规则不能直接用于做出预测，这点与标准的模型（比如决策树或神经网络）不同。 这是由于规则可能有许多不同的结论。 需要将关联规则转换为分类规则集的另外一层转换。 因此，关联算法生成的关联规则被称作 **未优化模型** 。 虽然用户可以浏览这些未优化模型，但除非用户指令系统从未优化模型生成分类模型，否则无法明确地将这些模型用作分类模型。 用户可通过浏览器的“生成”菜单选项来完成这种转换。

支持两种关联规则算法：

“先验”节点从数据抽取一组规则，即抽取信息内容最多的规则。 Apriori 节点提供五种选择规则的方法并使用复杂的索引模式来高效地处理大数据集。 对于较大的问题，Apriori 训练的速度通常较快；它对可保留的规则数量没有任何限制，而且可处理最多带有 32 个前提条件的规则。 “先验”要求输入和输出字段均为分类型字段，但因为它专为处理此类型数据而进行优化，因而处理速度快得多。

序列节点可发现连续数据或与时间有关的数据中的关联规则。 序列是一系列可能会以可预测顺序发生的项目集合。 例如，一个购买了剃刀和须后水的顾客可能在下次购物时购买剃须膏。 序列节点基于 CARMA 关联规则算法，该算法使用一个有效的两次传递方法查找序列。

表格数据与事务处理数据

关联规则模型使用的数据可能是事务处理格式，也可能表格格式，如下所述。 下面的内容是一般描述；具体的要求可能有所不同，请参见每种模型类型文档中的讨论。 请注意，对模型进行评分时，要评分的数据必须反映用于构建该模型的数据格式。 使用表格数据构建的模型只能用于对表格数据进行评分；使用事务处理数据构建的模型只能对事务处理数据进行评分。

事务处理格式

事务处理数据对于每个交易或项目具有一个单独的记录。 例如，如果客户进行了多次采购，那么每次采购都是一项单独的具有通过客户标识链接的相关联商品的记录。 有时，这也称为**行穷尽**格式。

客户	采购
1	jam
2	milk
3	jam
3	bread
4	jam
4	bread
4	milk

Apriori、CARMA 和序列节点都可使用事务处理数据。

表格数据

表格数据（也称为 **篮子** 数据或 **真值表** 数据）由单独的标志表示项目，其中每个标志字段表示一个特定项目的存在或不存在。 每条记录表示一个相关项目的完整集合。 标志字段可以是分类的也可以是数字的，但某些模型具有更具体的要求。

客户	Jam	Bread	Milk
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Apriori、CARMA、GSAR 和序列节点都可使用表格数据。

Apriori 节点

Apriori 节点还会发现数据中的关联规则。 Apriori 提供了五种不同的规则选择方法，并使用一种复杂的索引编制方案来高效处理大型数据集。

需求。 要创建 Apriori 规则集，您需要一个或多个输入字段和一个或多个目标字段。 输入字段和输出字段（角色为输入、目标或两者的字段）必须是符号型字段。 角色为无的字段将被忽略。 执行节点之前字段类型必须完全实例化。 数据可以是表格格式，也可以是事务格式。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。

强度。 对于大型问题， Apriori 训练的速度通常更快。 它对于可以包含的规则数也没有任何限制，可以处理最多带有 32 个预条件的规则。 Apriori 提供了五种不同的训练方法，因此将数据挖掘方法与当前问题相匹配时可以实现更强的灵活性。

Apriori 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

最小前项支持度。 您可以指定针对在规则集中保留规则的支持度标准。 **支持度**指的是训练数据中条件（规则中的“if”部分）为真的记录的百分比。 （请注意，此支持度定义与 CARMA 和序列节点中使用的定义不同。 有关更多信息，请参阅主题 第 202 页的『序列节点模型选项』。） 如果您要获取适用于非常小的数据子集的规则，请尝试增大此设置。

注: Apriori 的支持度定义基于带有前提条件的记录的数目。 这与 CARMA 和序列算法不同，对于这两种算法，支持度定义基于具有规则中所有项（即条件和结果）的记录的数量。 关联模型的结果显示（条件）支持度和规则支持度两个测量。

最小规则置信度。 您还可以指定置信度标准。 **置信度**基于其规则的前提条件为 true 的记录，并且是其结果也为 true 的记录的百分比。 换句话说，置信度是基于规则的正确预测的百分比。 置信度低于指定标准的规则将被放弃。 如果您获得的规则太多，请尝试增加此设置。 如果您获得的规则太少（甚至根本无法获得规则），请尝试降低此设置。

注: 如果需要，您可以在自己的值中突出显示值和类型。 请注意，如果将置信度值减小至小于 1.0，那么构建规则的过程不仅需要使用大量可用内存，而且极为耗时。

最大前项数。 您可以为任何规则指定最大前置条件数。 这是一种用来限制规则复杂性的方式。 如果规则太复杂或者太具体，请尝试降低此设置。 此设置对于训练时间也具有很大的影响。 如果规则集训练所需的时间过长，请尝试降低此设置。

仅标志的 true 值。 如果对于表格（数据表）格式的数据选择了此选项，则在生成的规则中只会包括真值。这样可能有助于使得规则更容易理解。 该选项不适用于事务格式的数据。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。

注: 如果字段类型为标志，那么 CARMA 模型构建节点在构建模型时会忽略空的记录，而 Apriori 模型构建节点会包含空的记录。 空的记录是模型构建中使用的所有字段值均为 false 的记录。

优化。 根据您的具体需求，选择旨在提高模型构建性能的选项。

- 选择 **速度** 可指示算法从不使用磁盘溢出，以便提高性能。
- 选择 **内存** 可指示算法在合适的时候，以牺牲某些速度为代价使用磁盘溢出。 缺省情况下，此选项处于选中状态。

  注: 以分布式方式运行时，*options.cfg* 文件中指定的管理员选项可能会覆盖此设置。 有关更多信息，请参阅《*IBM SPSS Modeler Server* 管理员指南》。

Apriori 节点专家选项

对于那些详细了解 Apriori 操作的人员来说，通过下列专家选项可以对归纳过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

评估度量。 Apriori 支持五种评估潜在规则的方法。

- **规则置信度**。 此缺省方法使用规则的置信度（或准确性）来评估规则。 对于此评估尺度，**评估尺度下限**为禁用状态，因为此选项对于“模型”选项卡上的**最小规则置信度**选项来说是多余的。 有关更多信息，请参阅主题 第 191 页的『Apriori 节点模型选项』。
- **置信度差**。 （也称为**与先验相比的绝对置信度差**。） 此评估尺度是规则的置信度与其先验置信度之间的绝对差。 此选项会防止出现偏差，即结果分布不均匀。 这有助于防止保留“明显的”规则。 例如，可能会出现这样的情况，80% 的客户会购买您最受欢迎的产品。 某项以 85% 的准确性预测购买该受欢迎产品的规则不会使您的了解加深，尽管 85% 的准确性对于绝对尺度来说似乎已经相当不错。 请将该评估尺度下限设置为您希望保留的规则的置信度最小差。
- **置信比**。 （也称为**置信度商数与 1 之间的差**。） 此评估度量是从 1 中减去规则置信度与先验置信度的比率（或者，如果比率大于 1，则为其倒数）。 与“置信度差”类似，此方法会考虑不均匀分布。 它尤其擅长寻找用于预测罕见事件的规则。 例如，假设有一种罕见的病理状况只在 1% 的病人中出现。 如果某个规则有 10% 的几率预测出这种病理状况，那么它与随机猜测相比是一种很大的提高，尽管从绝对尺度角度来看 10% 的准确性好像非常不起眼。 请将该评估尺度下限设置为您希望保留的规则的最小差。
- **信息差**。 （也称为**与先验的信息差**。） 此评估尺度基于 **信息增益** 测量。 如果某个特定结果的概率被视为一个逻辑值（一个 **数位**），则信息增益为基于条件可以确定的该数位的比例。 信息差是给定条件的情况下信息增益与只给定了结果的先验置信度的情况下信息增益之间的差。 此方法的一个重要特征在于，它考虑了支持度，因此对于给定水平的置信度，它倾向于覆盖更多记录的规则。 请将该评估尺度下限设置为您希望保留的规则的信息差。

  注：此评估尺度的尺度与其他尺度相比直观性较差，因此您可能需要试验各种下限才能获得满意的规则集。
- **标准化卡方**。 （也称为**标准化卡方度量方式**。） 此评估尺度是条件与结果之间关联的一个统计学指数。 该度量标准化为采用 0 到 1 之间的值。 该度量比信息差度量更依赖于支持度。 请将该评估尺度下限设置为您希望保留的规则的信息差。

  注：与信息差评估度量方式相同，此评估度量方式的尺度与其他尺度相比直观性较差，因此您可能需要试验各种下限才能获得满意的规则集。

允许没有前项的规则。 选择此选项可允许规则只包括结果（项目或项目集）。 如果您对确定常见项目或项目集感兴趣，那么此选项非常有用。 例如，cannedveg 是一个没有前项的单项规则，它表明购买 *cannedveg* 是数据中的常见情况。 在某些情况下，如果您只对最可信的预测感兴趣，则可能希望包括这样的规则。 缺省情况下，此选项处于关闭状态。 按照惯例，没有条件的规则的条件支持度表示为 100%，规则支持度与置信度相同。

CARMA 节点

CARMA 节点使用关联规则发现算法来发现数据中的关联规则。 关联规则是下列形式的语句：

if *antecedent(s)* **then** *consequent(s)*

例如，如果某个 Web 客户购买了无限网卡和高端无线路由器，那么该客户还可能购买无线音乐播放器（如果提供该产品的话）。 CARMA 模型在不要求用户指定输入或目标字段的情况下从数据抽取一组规则。 这就意味着生成的规则可用于很多种应用程序。 例如，您可以使用此节点生成的规则来查找一系列产品或服务（前项），其后项是您要在此假期内进行促销的商品。 使用 IBM SPSS Modeler，您可以确定哪些客户购买了这些条件产品，然后举办一个旨在促销这些结果产品的营销活动。

需求。 与 Apriori 不同，CARMA 节点不需要输入字段或目标字段。 这是该算法工作方式的重要组成部分，相当于在将所有字段设置为 双向的情况下构建 Apriori 模型。 您可以通过在构建模型后对该模型进行过滤来限制仅作为前提条件或结果列出的项目。 例如，您可以使用模型浏览器来查找一系列产品或服务（条件），其结果是您要在此假期内进行促销的项目。

要创建 CARMA 规则集，您需要指定一个标识字段以及一个或多个内容字段。 该标识字段可以是任意角色或测量级别。 角色为无的字段将被忽略。 执行节点之前字段类型必须完全实例化。 与 Apriori 相似，数据可以是表格格式，也可以是事务格式。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。

强度。 CARMA 节点基于 CARMA 关联规则算法。 与 Apriori 相比，CARMA 节点为规则支持度（对前提条件和结果的支持度）提供构建设置，而不是为前提条件支持度提供构建设置。 CARMA 还允许带有多个结果的

规则。 与 Apriori 相似，CARMA 节点生成的模型可以插入到数据流中用来创建预测。 有关更多信息，请参阅主题 第 27 页的『模型块』。

CARMA 节点字段选项

执行 CARMA 节点之前，必须在 CARMA 节点的“字段”选项卡上指定输入字段。 虽然大多数建模节点的字段选项卡选项都相同，但 CARMA 节点有几个独特的选项。 所有选项均在下面讨论。

使用“类型”节点设置。 此选项用于告知节点使用上游 Type 节点中的字段信息。 这是缺省选项。

使用定制设置。 该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。 选择了此选项之后，请根据您要读取事务格式的数据还是表格格式的数据来指定下面的字段。

使用事务格式。 此选项将根据您的数据是事务处理格式还是表格格式来更改此对话框中的其他字段控件。 如果您使用带有事务处理数据的多个字段，那么认为在某个特定记录中，这些字段中指定的项目表示着可以在一个带有时间戳的事务中找到的项目。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。

表格数据

如果未选中**使用事务格式**，则显示以下字段。

- **输入**。 选择输入字段或字段。 此操作与在“类型”节点中将字段的角色设置为输入类似。
- **分区**。 通过此字段，您可以指定用于针对模型构建中的训练、检验和验证阶段将数据划分为不同样本的字段。 通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。 如果已使用类型或分区节点定义了多个分区字段，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。） 同时请注意，要在分析时应用选定分区，还必须启用节点的“模型选项”选项卡中的分区功能。 （取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

事务处理数据

如果选中了**使用事务格式**，则显示以下字段。

- **标识**。 对于事务处理数据，请从列表中选择标识字段。 数字字段或符号字段可用作标识字段。 此字段的每个唯一值都应该表明一个特定的分析单元。 例如，在市场购物篮的应用中，每个标识可能表示一个客户。 对于 Web 日志分析应用，每个标识可能代表一台计算机（以 IP 地址表示）或一个用户（以登录数据表示）。
- **标识是连续的**。 （仅限 Apriori 和 CARMA 节点）如果您的数据进行了预先排序，以便所有标识相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。 如果您的数据未经预先排序（或者您不确定），请将此选项保持未选中状态，那么该节点将自动对数据进行排序。

  注: 如果数据未经排序，而您选择此选项，那么模型中可能会出现无效的结果。
- **内容**。 指定模型的内容字段。 这些字段包含与关联建模有关的项目。 您可以指定多个标志字段（如果数据为表格格式）或者一个名义字段（如果数据为事务格式）。

CARMA 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

最小规则支持度 (%)。 您可以指定支持条件。 **规则支持度**指的是训练数据中包含整个规则的标识所占的比例。 （请注意，此支持度定义不同于 Apriori 节点中使用的前提条件支持度。） 如果您要关注更常见的规则，请增大此设置的值。

最小规则置信度 (%)。 您可以指定用于在规则集中保留规则的置信度条件。 **置信度**是指预测正确的标识在所有由规则进行了预测的标识中所占的百分比。 基于训练数据，该百分比的计算如下：包含整个规则的标识数量除以其中包含条件的标识数量。 置信度低于指定标准的规则将被放弃。 如果您获得的规则无关或者太多，请尝试增加此设置。 如果您获得的规则太少，请尝试降低此设置。

注: 如果需要，您可以在自己的值中突出显示值和类型。 请注意，如果将置信度值减小至小于 1.0，那么构建规则的过程不仅需要使用大量可用内存，而且极为耗时。

最大规则大小。 您可以在规则中设置不同 item sets 的最大数目（相对于 items）。 如果相关规则相对较短，那么可以降低此设置，以加快规则集构建速度。

注: 如果字段类型为标志，那么 CARMA 模型构建节点在构建模型时会忽略空的记录，而 Apriori 模型构建节点会包含空的记录。 空的记录是模型构建中使用的所有字段值均为 false 的记录。

CARMA 节点专家选项

对于那些详细了解 Apriori 操作的人员来说，通过下列专家选项可以对建模过程进行微调。 要访问专家选项，请将“专家”选项卡上的“模式”设置为**专家**。

排除具有多个结果的规则。 选择此选项可以排除“双头”结果，即包含两个项目的结果。 例如，规则 bread & cheese & fish -> wine&fruit 包含一个双头结果 (wine&fruit)。 缺省情况下，这样的规则包括在内。

设置修剪值。 为了节省内存，所使用的 CARMA 算法在处理期间会定期从其潜在项目集的列表中移除（**修剪**）不常用的项目集。 选择此选项可调整修剪频率，您指定的数字将决定修剪频率。 输入较小的值可降低该算法的内存要求（但可能会延长所需的训练时间），输入较大的值会加快训练速度（但可能会提高内存要求）。 缺省值是 500。

改变支持。 选择该选项会排除因为纳入不平均而好像表现为非常频繁的不频繁项目集合，从而提高效率。 这是通过这样的方式实现的：首先从较高的支持度水平开始，然后逐渐下降到“模型”选项卡上指定的水平。 对于 **事务的估计数量** 输入一个值可指定支持度水平应采用的下降速度。

允许没有前项的规则。 选择此选项可允许规则只包括结果（项目或项目集）。 如果您对确定常见项目或项目集感兴趣，那么此选项非常有用。 例如，cannedveg 是一个没有前项的单项规则，它表明购买 *cannedveg* 是数据中的常见情况。 在某些情况下，如果您只对最可信的预测感兴趣，则可能希望包括这样的规则。 此选项缺省为不选中状态。

关联规则模型块

关联规则模型块代表由下列关联规则建模节点之一所发现的规则：

- Apriori
- CARMA

模型块包含建模期间从数据提取的规则的相关信息。

注: 如果未按 ID 对事务处理数据进行排序，那么关联规则块评分可能不正确。

查看结果

您可以使用该对话框上的“模型”选项卡来浏览关联模型（Apriori 和 CARMA）以及序列模型生成的规则。 在生成新节点或对模型评分之前浏览模型块会使您看到规则的相关信息，还会提供用于过滤结果和对结果进行排序的选项。

模型评分

精炼模型块（Apriori、CARMA 和序列）可以添加到流中，用于进行评分。 有关更多信息，请参阅主题 第 35 页的『使用流中的模型块』。 用于评分的模型块在其各自的对话框中包括一个额外的“设置”选项卡。 有关更多信息，请参阅主题 第 197 页的『关联规则模型块设置』。

无法以其原始格式将未优化模型块用于评分。 而您可以生成一个规则集，并将该规则集用于评分。 有关更多信息，请参阅主题 第 198 页的『从关联模型块生成规则集』。

“关联规则”模型块详细信息

在关联规则模型块的“模型”选项卡上，您可以看到一个表，其中包含了该算法提取的规则。 表中的每行都代表一个规则。 第一列代表结果（规则的“then”部分），而下一列代表条件（规则的“if”部分）。 后面的列包含规则信息，如置信度、支持度和提升。

关联规则通常以下表中的格式显示。

表 *15:* 关联规则示例	
结果	**前提条件**
`Drug = drugY`	`Sex = F` `BP = HIGH`

该示例规则的解释为 如果 *Sex = "F" and BP = "HIGH"*，则 *Drug* 很可能为 *drugY*；或者以另一种方式解释 对于 *Sex = "F" and BP = "HIGH"* 的记录，*Drug* 很可能为 *drugY*。 使用对话框工具栏，可以选择显示其他信息，如置信度、支持度和实例数。

“排序”菜单。 工具栏上的“排序”菜单按钮控制着规则的排序。 排序的方向（升序或降序）可以使用排序方向按钮（上箭头或下箭头）进行更改。

您可以按照下列条件对规则进行排序：

- 支持度
- 置信度
- 规则支持(R)
- 结果
- 评估
- 增益
- 部署能力

“显示/隐藏”菜单。 “显示/隐藏”菜单（标准工具栏按钮）用于控制规则的显示选项。

图 *46:* “显示/隐藏”按钮

可用的显示选项如下：

- **规则标识**，显示模型构建期间分配的规则标识。 通过规则标识，可以标识哪些规则要应用于某个给定的预测。 通过规则标识，还可以在以后合并附加的规则信息，如部署能力、产品信息或条件。
- **实例数**，显示规则所适用的唯一标识数（即，前提条件为 true 的标识）的相关信息。 例如，假设规则为 `bread -> cheese`，训练数据中包含条件 *bread* 的记录数量称为**实例数**。
- **支持度**，显示前提条件支持度，即其前提条件为 true 的标识在训练数据中所占的比例。 例如，如果 50% 的训练数据包含购买面包的情况，那么规则 `bread -> cheese` 将具有 50% 的前项支持度。 注：此处定义的支持度与实例数相同，但以百分比的形式表示。
- **置信度**，显示规则支持度与前提条件支持度的比率。 此比值表明了带有指定条件、并且其结果也为真的标识的比例。 例如，如果 50% 的训练数据包含 bread（指示前项支持度），但只有 20% 同时包含 bread 和 cheese（指示规则支持度），那么规则 `bread -> cheese` 的置信度将为 `Rule Support / Antecedent Support`，在本例中为 40%。
- **规则支持度**，显示其整个规则、前提条件和结果均为 true 的标识所占的比例。 例如，如果 20% 的训练数据同时包含 bread 和 cheese，那么规则 `bread -> cheese` 的规则支持度为 20%。
- 如果您选择了其中一个专家关联规则标准（置信度差、置信度比率、信息差或标准化卡方），那么将会包括**评估**。 这些专家标准测量将与用户所设置的**评估测量下限**数值进行比较（仅当选择了专家标准规则时，才会应用）。 对于每个专家关联规则标准，“评估”统计的含义如下所示：
  - 置信度差：后验置信度 - 先验置信度
  - 置信度比率：(后验置信度 - 先验置信度)/后验置信度
  - 信息差：信息增益测量
  - 标准化卡方：标准化卡方统计

  其中每一项统计都会与用户所设置的**评估测量下限**数值进行比较，如果统计超过此数值，就会选择规则。

- **增益**，显示规则置信度与具有结果的先验概率的比率。 例如，如果整个人口中有 10% 购买面包，那么以 20% 的置信度预测人们是否会购买面包的规则将具有 20/10 = 2 的增益。 如果另一条规则告诉您，人们购买面包的置信度为 11%，那么该规则将具有接近 1 的增益，这意味着具有前项不会对具有结果的概率造成很大差异。 总之，提升度不为 1 的规则比提升度接近 1 的规则的相关性更强。
- **部署能力**，这是对训练数据中满足前提条件但不满足结果的部分所占百分比的度量。 在产品购买领域，它的意思大致为：总的客户群中有多少百分比拥有了（或已经购买了）条件，但尚未购买结果。 可部署性统计信息定义为 `((Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100`，其中前项支持度表示前项为 true 的记录数，规则支持度表示前项和后项均为 true 的记录数。

“过滤器”按钮。 菜单上的“过滤器”按钮（漏斗图标）会扩展对话框的底部，从而显示一个面板，其中将显示活动的规则过滤器。 过滤器用于减少“模型”选项卡上显示的规则数量。

图 *47: “过滤”按钮*

要创建过滤器，请单击位于扩展面板右侧的过滤器图标。 这样将打开一个单独的对话框，您可以在其中指定用于显示规则的约束条件。 请注意，“过滤”按钮通常与“生成”菜单一起使用，以便首先过滤规则，然后生成一个包含部分规则的模型。 有关更多信息，请参阅以下第 196 页的『为规则指定过滤器』。

“查找规则”按钮。 “查找规则”按钮（望远镜图标）使您可以搜索对指定规则标识显示的规则。 相邻的显示框指示可用规则数中当前显示的规则数。 规则标识由模型按照发现时间的顺序指定，并且会在评分期间添加到数据中。

图 *48: “查找规则”按钮*

要对规则标识重新排序：

1. 您可以在 IBM SPSS Modeler 中对规则标识进行重新排序，方法是，首先根据所需的测量标准（如置信度或提升）对规则显示表进行排序。
2. 然后使用“生成”菜单中的选项，创建一个经过过滤的模型。
3. 在“已过滤的模型”对话框中，选择**对规则重新进行连续编号的起始号码**，然后指定一个开始号码。

有关更多信息，请参阅第 198 页的『生成已过滤的模型』。

为规则指定过滤器

缺省情况下，规则算法（如 Apriori、CARMA 和序列）可能会生成非常大量的规则。 为了在浏览时增强明确度，或者为了简化规则评分，您应该考虑过滤规则，以便更加显著地显示相关的结果和条件。 使用规则浏览器“模型”选项卡上的过滤选项，可以打开一个用于指定过滤条件的对话框。

后项。 选择**启用过滤器**可激活根据包含或排除指定结果来对规则进行过滤的选项。 选择 **包括任意** 可创建一个过滤器，该过滤器中的规则至少包含一个指定结果。 另外，选择 **排除** 可创建一个排除指定结果的过滤器。 您可以使用列表框右侧的选取器图标选择结果。 这样将打开一个对话框，其中列出生成的规则中包含的所有结果。

注：结果可能包含多个项目。 过滤器只会检查结果是否包含一个指定项目。

前项。 选择**启用过滤器**可激活根据包含或排除指定前提条件来对规则进行过滤的选项。 您可以使用列表框右侧的选取器图标选择项目。 这样将打开一个对话框，其中列出生成的规则中包含的所有条件。

- 选择 **包括所有** 可将过滤器设置为一个包含过滤器，其中的规则必须包括指定的所有条件。
- 选择 **包括任意** 可创建一个过滤器，该过滤器中的规则至少包含一个指定条件。
- 选择 **排除** 可创建一个排除包含指定条件的规则的过滤器。

置信度。 选择**启用过滤器**可激活根据规则的置信度级别来对规则进行过滤的选项。 您可以使用 **最小** 和 **最大** 控件来指定置信度范围。 当您浏览生成的模型时，置信度将以百分比的形式列出。 当您对输出评分时，置信度则表示为一个介于 0 和 1 之间的数字。

前提条件支持。 选择**启用过滤器**可激活根据规则的前提条件支持度级别来对规则进行过滤的选项。 条件支持度指的是训练数据中与当前规则包含相同条件的比例，因此与普及性指数有点类似。 您可以使用 **最小** 和 **最大** 控件，根据支持度水平来指定过滤规则的范围。

增益。 选择**启用过滤器**可激活根据规则的增益度量来对规则进行过滤的选项。 注：增益过滤仅可用于发行版 8.5 之后构建的关联模型，或包含增益度量的先前版本的模型。 序列模型不包含此选项。

单击 **确定** 可应用已在此对话框中启用的所有过滤器。

为规则生成图形

关联节点提供了大量信息，但对业务用户来说，它可能并不始终是一种方便访问的格式。 要通过可以轻松纳入业务报告、简报等的方式提供数据，您可以生成所选数据的图形。 从“模型”选项卡上，可以为选定规则生成图形，从而只为该规则中的观测值创建图形。

1. 在“模型”选项卡上，选择感兴趣的规则。
2. 从“生成”菜单中，选择**图形（从选定内容）**。 这将显示“图形板基本”选项卡。

   注：通过此方式显示“图形板”时，仅“基本”和“详细”选项卡可用。
3. 使用“基本”或“详细”选项卡设置来指定要在图形上显示的详细信息。
4. 单击“确定”以生成图形。

图形标题标识所包含的选定规则和条件详细信息。

关联规则模型块设置

此“设置”选项卡用于为关联模型（Apriori 和 CARMA）指定评分选项。 此选项卡仅在模型块添加到用于评分的流后才可用。

注: 用于浏览未优化的 模型的对话框不包含“设置”选项卡，因为无法对其进行评分。 要对 “未优化”模型进行评分，必须先生成规则集。 有关更多信息，请参阅主题 第 198 页的『从关联模型块生成规则集』。

最大预测数：指定为每个购物篮项集合包括的最大预测数。 此选项与下面的“规则标准”一起使用可生成“最佳”预测，其中 最佳指的是置信度、支持度、提升等的最高水平，如下面的内容所述。

规则条件：选择用于确定规则强度的度量。 规则按照此处选择的标准强度进行排序，以便返回项目集合的最佳预测。 以下列表中显示了可用条件。

- 置信度
- 支持度
- 规则支持度（支持度 * 置信度）
- 增益
- 部署能力

允许重复预测：选择此选项可在评分时包括具有相同结果的多项规则。 例如，选择此选项可允许对下列规则进行评分：

```
bread & cheese -> wine
cheese & fruit -> wine
```

关闭此选项可在评分时排除重复的预测。

注: 仅当先前已预测所有结果 (`wine & pate`) 时，才会将具有多个结果 (`bread & cheese & fruit -> wine & pate`) 的规则视为重复预测。

忽略不匹配的购物篮项目：选择此选项可忽略项目集中附加项的存在。 例如，为包含 `[tent & sleeping bag & kettle]` 的购物篮选择此选项时，尽管该购物篮中存在额外的商品 (`kettle`)，但规则 `tent & sleeping bag -> gas_stove` 仍将适用。

可能存在一些情况应该排除额外的项目。 例如，很可能出现这样的情况，某人购买了 tent（帐篷）、sleeping bag（睡袋）和 kettle（水壶），而此人已经拥有了 gas stove（燃气炉），这点通过 kettle（水

壶）的存在表明。 换句话说，gas stove（燃气炉）可能不是最佳预测。 这种情况下，您应该取消选择 **忽略不匹配的购物篮项目** 以确保规则条件与购物篮内容精确匹配。 缺省情况下，不匹配的项目将被忽略。

检查购物篮中是否不包含预测。 选择此选项可确保结果也不存在于购物篮中。 例如，如果进行评分的目的是为了进行一项家具产品推荐，那么已经包含餐桌的购物篮可能不会购买另一个这样的家具。 这种情况下，您应该选择此选项。 另一方面，如果是易变质或一次性产品（如奶酪、婴儿配方奶粉或者卫生纸），那么后项已存在于购物篮的规则可能有些价值。 在后面一种情况下，最有用的选项可能是下面的 **不检查购物篮中是否存在预测值** 。

检查购物篮中是否存在预测值：选择此选项可确保结果也存在于购物篮中。 当您尝试深入了解现有的客户或事务时，此方法非常有用。 例如，您可能希望确定提升最高的规则，然后探索哪些客户符合这些规则。

不检查购物篮中是否存在预测值：选择此选项可在评分时包括所有规则，而与购物篮中是否存在结果无关。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

关联规则模型块概要

关联规则模型块的“概要”选项卡显示发现的规则数量，以及规则集中规则的最大和最小支持度、提升值、置信度和部署能力。

从关联模型块生成规则集

关联模型块（如 Apriori 和 CARMA）可用于直接对数据评分，您也可以首先生成一个规则子集，称为 **规则集** 。 </all>GRI END FILTER ALL -->处理无法直接用于评分的未优化模型时，规则集尤其有用。 有关更多信息，请参阅主题 第 38 页的『未优化模型』。

要生成规则集，请从模型块浏览器的“生成”菜单中选择**规则集**。 您可以指定下列选项，将规则转换为规则集：

规则集名称。 通过此选项，您可以指定新生成的“规则集”节点的名称。

创建节点于。 控制新生成的“规则集”节点的位置。 选择 **工作区** 、 **GM 选用板** 或 **两者** 。

目标字段。 确定哪个输出字段将用于生成的“规则集”节点。 从列表中选择一个输出字段。

最小支持度。 指定要在生成的规则集中保留的规则的最小支持度。 支持度小于指定值的规则将不会包含在新的规则集中。

最小置信度。 指定要在生成的规则集中保留的规则的最小置信度。 置信度小于指定值的规则将不会包含在新的规则集中。

缺省值。 通过此选项，您可以为分配到不会触发任何规则的已评分记录的目标字段指定缺省值。

生成已过滤的模型

要从关联模型块（如 Apriori、CARMA 或序列规则集节点）生成已过滤的模型，请从模型块浏览器的“生成”菜单中选择**已过滤的模型**。 这样将创建一个子集模型，其中只包含浏览器中当前显示的那些规则。 注：无法对未优化模型生成已过滤的模型。

您可以指定下列用于过滤规则的选项：

新模型的名称。 通过此选项，您可以指定新的“已过滤模型”节点的名称。

创建节点于。 控制新的“已过滤模型”节点的位置。 选择 **工作区** 、 **GM 选用板** 或 **两者** 。

规则编号。 指定规则标识在已过滤模型所包含的规则子集中的编号方式。

- **保留原始规则标识号**。 选择此选项可以保持原始的规则编号。 缺省情况下，会为规则提供一个与算法发现它们的顺序相对应的标识。 该顺序可能会因所采用算法的不同而有所差别。
- **对规则进行连续重新编号，开始于**。 选择此选项可以为已过滤规则指定新的规则标识。 新的标识将根据“模型”选项卡上规则浏览器表中显示的排序顺序进行指定，从您在此处指定的数字开始。 您可以使用右侧的箭头指定标识的开始号码。

关联规则评分

通过关联规则模型块运行新数据生成的评分会返回到不同的字段中。 将为每个预测添加三个新字段，其中 *P* 表示预测，*C* 表示置信度，*I* 表示规则标识。 这些输出字段的组织取决于输入数据是采用事务格式还是表格格式。 请参阅第 190 页的『表格数据与事务处理数据』以获取有关这些格式的概述。

例如，假设您要使用一个基于下面三个规则生成预测的模型对购物篮数据进行评分：

```
Rule_15  bread&wine -> meat (confidence 54%)
Rule_22  cheese -> fruit (confidence 43%)
Rule_5  bread&cheese -> frozveg (confidence 24%)
```

表格数据。 对于表格数据，这三个预测（3 为缺省值）将在单一记录中返回。

表 *16:* 表格格式的评分

标识	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat	0.54	15	fruit	0.43	22	frozveg	.24	5

事务数据。 对于事务处理数据，将对每个预测生成一个单独的记录。 预测仍然会添加到单独的列中，但评分在计算时返回。 这样会生成带有不完整预测的记录，如下面的示例输出所示。 第二个和第三个预测（P2 和 P3）在第一个记录中是空白值，同时还会显示相关的置信度和规则标识。 但返回评分时，最后一个记录将包含所有三个预测。

表 *17:* 事务处理格式的评分

标识	项	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	meat	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	Cheese	meat	0.54	14	fruit	0.43	22	$null$	$null$	$null$
Fred	wine	meat	0.54	14	fruit	0.43	22	frozveg	0.24	5

要只包括用于报告或部署目的的完整预测，请使用选择节点选择完整的记录。

注：为了清楚起见，这些示例中使用的字段名称都是缩写。 在实际应用中，关联模型的结果字段将按下表所示进行命名。

表 *18:* 关联模型的结果字段的名称

新字段	字段名示例
预测	*$A-TRANSACTION_NUMBER-1*
置信度（或其他标准）	*$AC-TRANSACTION_NUMBER-1*
规则标识(U)	*$A-Rule_ID-1*

带有多个结果的规则

CARMA 算法允许规则带有多个结果，例如：

```
bread -> wine&cheese
```

对这样的“双头”规则进行评分时，预测将以下表中显示的格式返回。

表 *19:* 对包括带有多个结果预测的结果进行评分												
标识	**Bread**	**Wine**	**Cheese**	**P1**	**C1**	**I1**	**P2**	**C2**	**I2**	**P3**	**C3**	**I3**
Fred	1	1	1	meat&veg	0.54	16	fruit	0.43	22	frozveg	.24	5

在某些情况下，您可能需要在部署之前分割这样的评分。 要分割带有多个结果的预测，您需要使用 CLEM 字符串功能解析该字段。

部署关联模型

对关联模型进行评分时，预测和置信度将输出到单独的列中（其中 *P* 表示预测， *C* 表示置信度， *I* 表示规则标识）。 这种情况要区分输入数据是表格格式还是事务格式。 有关更多信息，请参阅主题 第 199 页的『关联规则评分』。

准备评分进行部署时，您可能会发现您的应用程序需要将输出数据转换为预测位于行中的格式，而不是位于列中的格式（每行一个预测，有时称为“行穷尽”格式）。

转置表格评分

您可以使用 IBM SPSS Modeler 中的一些步骤将表格评分从列转置为行，如下面的步骤所示。

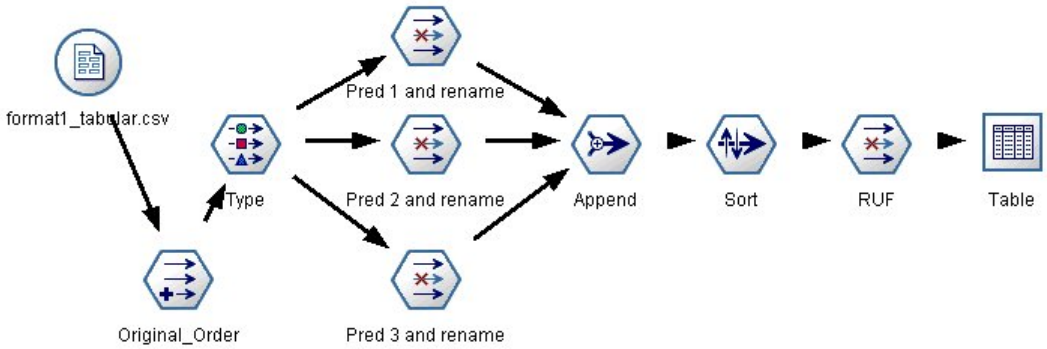


图 *49:* 用于将表格数据转置为行穷尽格式的示例流

1. 在“派生”节点中使用 `@INDEX` 函数来确定当前的预测顺序，并将此指标保存在新字段中，例如 *Original_order*。
2. 添加一个类型字段，确保所有字段均实例化。
3. 使用过滤节点将缺省的预测、置信度和标识字段（ *P1*、*C1*、*I1*）重命名为普通字段，如 *Pred*、 *Crit* 和 *Rule_ID*，这些字段将用于在以后追加记录。 对于每个生成的预测都需要一个过滤节点。

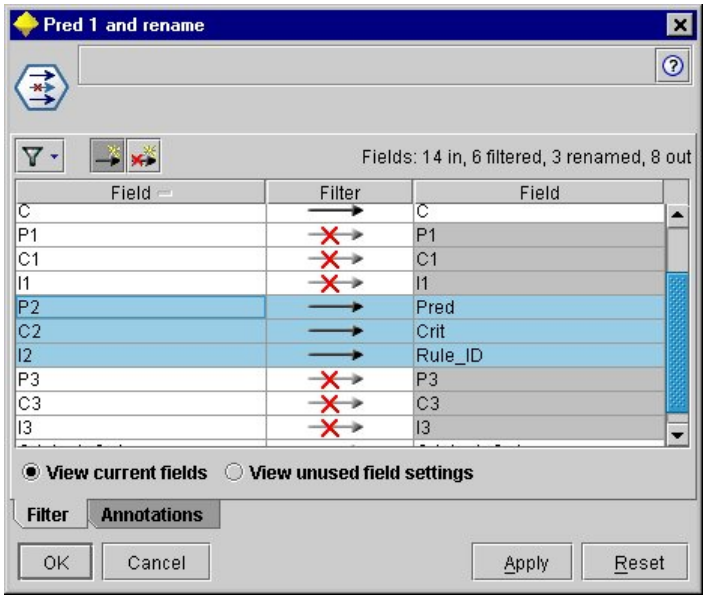


图 *50:* 重命名预测 *2* 的字段时过滤预测 *1* 和预测 *3* 的字段。

4. 使用追加节点追加共享 *Pred*、 *Crit* 和 *Rule_ID* 的值。
5. 连接一个排序节点，以便按照字段 *Original_order* 的升序对记录进行排序，按照 *Crit* 的降序对记录进行排序，后面一个字段是用于按标准（如置信度、提升和支持度）对预测进行排序的字段。
6. 使用另一个过滤节点将字段 *Original_order* 从输出中过滤掉。

此时，数据就可以进行部署了。

转置事务评分

转置事务评分的过程与上面的过程相似。 例如，下面显示的流会根据部署需要，将评分转置为每行一个预测的格式。

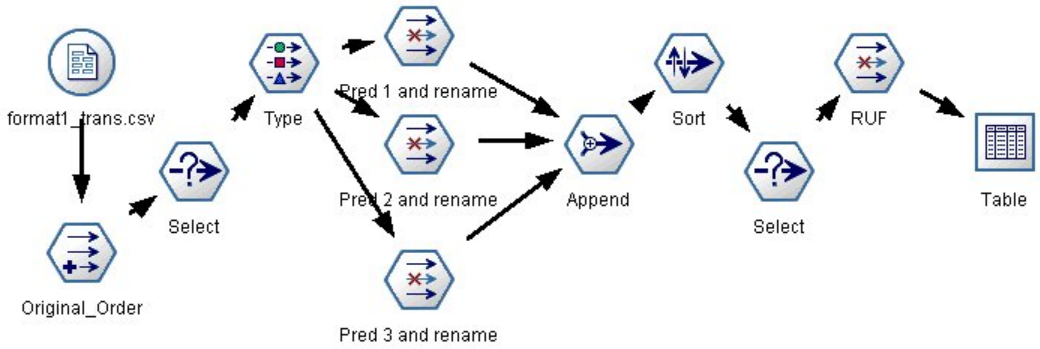


图 *51:* 用于将事务处理数据转置为行穷尽格式的示例流

除了添加两个“选择”节点之外，该过程与先前对表格数据说明的过程完全相同。

- 第一个选择节点用于对相邻记录的规则标识进行比较，以便只包括唯一的或未优化的记录。 此“选择”节点使用 CLEM 表达式来选择记录：`ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`。
- 第二个“选择”节点用于废弃无关的规则，或 `Rule_ID` 具有空值的规则。 此“选择”节点使用以下 CLEM 表达式来废弃记录：`not(@NULL(Rule_ID))`。

有关转置评分进行部署的详细信息，请联系技术支持部门。

序列节点

“序列”节点可发现序列数据或面向时间的数据中 `bread -> cheese` 格式的模式。 序列的元素为组成一个事务的 **项目集合** 。 例如，如果某人进入商店，购买了面包和牛奶，几天之后返回了该商店，购买了一些奶

酪，那么这个人的购买活动可以表示为两个项目集合。 第一个项目集合包含面包和牛奶，第二个包含奶酪。 **序列** 是一系列可能会以可预测顺序发生的项目集合。 序列节点会检测频繁出现的序列，并创建一个可用于生成预测的生成模型节点。

需求。 要创建序列规则集，您需要指定一个标识字段、一个可选的时间字段，以及一个或多个内容字段。 请注意，这些设置必须在建模节点的“字段”选项卡上进行；不能从上游“类型”节点中读取。 该标识字段可以是任意角色或测量级别。 如果指定时间字段，那么该字段可以是任意角色，但其存储必须是数字、日期、时间或时间戳。 如果不指定时间字段，序列节点则会使用隐含的时间戳，实际上是使用行号作为时间值。 内容字段可具有任意测量级别和角色，但所有内容字段的类型必须相同。 如果这些字段是数字型的，那么必须为整数范围（不是实数范围）。

强度。 序列节点基于 CARMA 关联规则算法，该算法使用一个有效的两次传递方法查找序列。 另外，序列节点创建的生成的模型节点可以插入到数据流中来创建预测。 生成的模型节点还可生成超节点用于检测或计数特定的序列，以及基于特定的序列作出预测。

序列节点字段选项

执行序列节点之前，必须在序列节点的“字段”选项卡上指定标识字段和内容字段。 如果您要使用时间字段，也需要在此处指定。

“标识”字段。 从列表中选择标识字段。 数字字段或符号字段可用作标识字段。 此字段的每个唯一值都应该表明一个特定的分析单元。 例如，在市场购物篮的应用中，每个标识可能表示一个客户。 对于 Web 日志分析应用，每个标识可能代表一台计算机（以 IP 地址表示）或一个用户（以登录数据表示）。

- **标识是连续的**。 如果您的数据进行了预先排序，以便所有标识相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。 如果您的数据未经预先排序（或者您不确定），请将此选项保持不选中状态，序列节点将自动对该数据进行排序。

注：如果您的数据未经过排序而您选择了此选项，那么可能会在序列模型中得到无效结果。

时间字段。 如果您要在数据中使用字段来指示事件时间，请选择**使用时间字段**并指定要使用的字段。 时间字段必须是数字、日期、时间或时间戳型的。 如果未指定时间字段，那么将假定记录按顺序从数据源到达，并将记录号用作时间值（第一条记录的发生时间为 "1"；第二条记录的发生时间为： "2"；如此类推）。

内容字段。 指定模型的内容字段。 这些字段包含与序列建模有关的事件。

序列节点可以处理表格格式的数据，也可以处理事务格式的数据。 如果您使用带有事务处理数据的多个字段，那么认为在某个特定记录中，这些字段中指定的项目表示着可以在一个带有时间戳的事务中找到的项目。 有关更多信息，请参阅主题 第 190 页的『表格数据与事务处理数据』。

分区。 通过此字段，您可以指定用于针对模型构建中的训练、检验和验证阶段将数据划分为不同样本的字段。 通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。 如果已使用类型或分区节点定义了多个分区字段，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。）同时请注意，要在分析时应用选定分区，还必须启用节点的“模型选项”选项卡中的分区功能。 （取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

序列节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

最低规则支持度 (%) 您可以指定支持度标准。 规则支持度指的是训练数据中包含整个序列的标识所占的比例。 如果您要关注更常见的序列，请增加此设置。

最低规则置信度 (%) 您可以指定针对在序列集中保留序列的置信度标准。 置信度是指预测正确的标识在所有由规则进行了预测的标识中所占的百分比。 基于训练数据，该百分比的计算如下：包含整个序列的标识数量除以其中包含条件的标识数量。 置信度低于指定标准的序列将被放弃。 如果您获得的序列太多或者不是非常相关，请尝试增加此设置。 如果您获得的序列太少，请尝试降低此设置。

注: 如果需要，您可以在自己的值中突出显示值和类型。 请注意，如果将置信度值减小至小于 1.0，那么构建规则的过程不仅需要使用大量可用内存，而且极为耗时。

最大序列大小 您可以设置序列中不同值的最大数目。 如果相关序列相对较短，那么可以降低此设置，以加快序列集构建速度。

要添加到流中的预测数 指定最终生成的“模型”节点要添加到流中的预测的数目。 有关更多信息，请参阅 第 204 页的『序列模型块』。

序列节点专家选项

对于那些详细了解序列节点操作的人员来说，通过下列专家选项可以对建模过程进行微调。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

设置最长持续时间。 如果选中了此选项，那么会将序列限制为持续时间（第一个项目集与最后一个项目集之间的时间）少于或等于指定值的序列。 如果没有指定时间字段，该持续时间则以原始数据中的行数（记录数）表示。 如果使用的时间字段为时间、日期或时间戳型字段，该持续时间则表示为秒数。 对于数字字段，持续时间则使用与字段相同的单位数表示。

设置修剪值。 为了节省内存，“序列”节点中使用的 CARMA 算法在处理期间会定期从其潜在项目集的列表中移除（**修剪**）不常用的项目集。 选择此选项可调整修剪的频率。 指定的数字决定了修剪频率。 输入较小的值可降低该算法的内存要求（但可能会延长所需的训练时间），输入较大的值会加快训练速度（但可能会提高内存要求）。

设置内存中的最大序列。 如果选中了此选项，那么 CARMA 算法会将建模期间候选序列的内存存储限制为指定的序列数。 如果 IBM SPSS Modeler 在序列建模期间使用的内存过多，请选择此选项。 请注意，您在此处指定的最大序列值指的是在构建模型期间进行内部跟踪的备选序列数。 此数字应该比最终模型中预期的序列数大很多。

约束项目集之间的间隔。 通过此选项，您可以指定对分隔项目集的时间间隔的约束。 如果选择了此选项，那么不会考虑时间间距小于您所指定的**最小间距**或大于**最大间距**的项目集作为序列的组成部分。 使用此选项可避免考虑包括较长时间间隔或者在很短的时间跨度内发生的那些序列。

注：如果使用的时间字段为时间、日期或时间戳记字段，那么时间间隔以秒为单位。 对于数字型字段，时间间距则使用与时间字段相同的单位数表示。

例如，请考虑下面的事务列表。

表 *20:* 事务的示例列表

标识	时间	内容
1001	1	apples
1001	2	bread
1001	5	Cheese
1001	6	dressing

如果您针对这些数据建模时指定的最小间距为 2，那么会得到下列序列：

```
apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing
```

您将看不到诸如 `apples -> bread` 之类的序列，因为 `apples` 与 `bread` 之间的间隔小于最小间隔。 同样地，请考虑以下替代数据。

表 *21:* 事务的示例列表

标识	时间	内容
1001	1	apples

表 *21:* 事务的示例列表 *(*继续*)*

标识	时间	内容
1001	2	bread
1001	5	Cheese
1001	20	dressing

如果最大间隔设置为 10，那么您将不会看到任何带有 dressing 的序列，因为 cheese 与 dressing 之间的间隔过大，而无法将它们视为同一序列的组成部分。

序列模型块

“序列”模型块表示“序列”节点针对某个特定输出字段发现的序列，可以添加到流中以生成预测。

当您运行包含“序列”节点的流时，“序列”节点会将包含预测的一对字段，以及序列模型中每个预测的相关置信度值添加到数据中。 缺省情况下，会添加包含三个最佳预测的三对字段（以及它们相关联的置信度值）。 您既可以通过在构建时设置序列节点模型选项更改构建模型时生成的预测数，也可以在将模型块添加到流之后在“设置”选项卡上更改此数量。 有关更多信息，请参阅主题 第 206 页的『序列模型块设置』。

新的字段名称派生自模型名称。 预测字段的字段名称为 *$S-sequence-n*（其中 *n* 表示第 *n* 个预测）置信度字段的字段名称为 *$SC-sequence-n*。 在一个序列中具有多个序列规则节点的流中，新的字段名称将包括数字前缀，以便将它们区别开来。 流中的第一个“序列”节点将使用常用名称，第二个节点将使用以 *$S1-* 和 *$SC1-* 开头的名称，第三个节点将使用以 *$S2-* 和 *$SC2-* 开头的名称，依此类推。 预测按照置信度的顺序显示，因此 *$S-sequence-1* 所包含预测的置信度最高，*$S-sequence-2* 所包含预测的置信度次高，依此类推。 对于可用预测数小于所请求预测数的记录，其余预测包含值 $null$。 例如，如果对于特定记录只能进行两项预测，那么 *$S-sequence-3* 和 *$SC-sequence-3* 的值将为 $null$。

对于每条记录，会将模型中的规则与目前对于当前标识已经处理的事务集合（包括当前记录和具有相同标识和较早时间戳的所有以前记录）进行比较。 将使用适用于此事务集合的、置信度值最高的 *k* 个规则为该记录生成 *k* 个预测，其中 *k* 为模型添加到流之后在“设置”选项卡上指定的预测数。 （如果多个规则对于该事务集合预测了相同的结果，那么只使用置信度最高的规则。） 有关更多信息，请参阅主题 第 206 页的『序列模型块设置』。

与其他类型的关联规则模型相同，数据格式必须与构建序列模型时使用的格式相匹配。 例如，使用表格数据构建的模型只能用于对表格数据进行评分。 有关更多信息，请参阅主题 第 199 页的『关联规则评分』。

注：在流中使用生成的“序列集”节点对数据进行评分时，您在构建模型时选择的任何容差或间隔设置都将被忽略，不会用于评分目的。

根据序列规则进行的预测

该节点以与时间相关（如果在构建模型时未使用时间戳字段的话，那么与顺序相关）的方式处理记录。 记录应该按照标识字段和时间戳字段（如果存在的话）排序。 但是，预测与添加到其中的记录的时间戳没有关系。 它们只是在给出到当前记录为止当前标识的事务历史记录的情况下，指出最可能在将来的某个时间出现的项目。

请注意，每条记录的预测不一定与该记录的事务相关。 如果当前记录的事务未触发特定规则，那么将根据当前标识的先前事务来选择规则。 换言之，如果当前记录未将任何有用的预测信息添加到序列中，那么此标识的最后一个有用事务的预测就会被继承到当前记录。

例如，假设您拥有的序列模型具有一个规则

```
Jam -> Bread (0.66)
```

然后您将其传递到下列记录。

表 *22:* 示例记录

标识	采购	预测
001	jam	bread

表 *22:* 示例记录 *(*继续*)*		
标识	**采购**	**预测**
001	milk	bread

请注意，与您的预期相同，第一个记录生成了预测 *bread*。 第二条记录还包含 *bread* 的预测，因为没有先 *jam* 后 *milk* 的规则；因此，*milk* 事务不会添加任何有用信息，并且规则 `Jam -> Bread` 仍适用。

生成新节点

通过“生成”菜单可以基于序列模型创建新的超节点。

- **规则超节点**。 创建一个可以检测和计算已评分数据中序列发生次数的超节点。 如果未选择任何规则，此选项则禁用。 有关更多信息，请参阅主题 第 206 页的『从序列模型块生成规则超节点』。
- **建模到选用板**。 将模型返回到模型选用板。 当有同事发给您包含模型的流而不是模型本身时，该功能很有用。

序列模型块详细信息

序列模型块的“模型”选项卡显示算法提取的规则。 表中的每行都代表一个规则，其中条件（规则“if”部分）位于第一列，结果（规则的“then”部分）位于后面的第二列。

每项规则都以下列格式显示。

表 *23:* 规则格式	
前提条件	**结果**
beer and cannedveg	beer
fish fish	fish

第一个规则示例解释为：对于在同一事务中具有*“beer”*和*“cannedveg”*的标识，后面可能会出现*“beer”*。 第二个规则示例解释为：对于在一个事务中具有*“fish”*，而在另一个事务中也具有*“fish”*的标识，后面可能会出现*“fish”*。 请注意在第一个规则中， *beer* 和 *cannedveg* 是同时购买的；在第二个规则中， *fish* 是在两个不同的事务中购买的。

“排序”菜单。 工具栏上的“排序”菜单按钮控制着规则的排序。 排序的方向（升序或降序）可以使用排序方向按钮（上箭头或下箭头）进行更改。

您可以按照下列条件对规则进行排序：

- 支持度百分比
- 置信度百分比
- 规则支持百分比
- 结果
- 第一个前提条件
- 最后一个前提条件
- 项目数（前提条件）

例如，下表按照项目数，以降序进行排序。 条件集合中具有多个项目的规则排在条件集合中项目数较少的规则前面。

表 *24:* 按项目数排序的规则	
前提条件	**结果**
beer and cannedveg and frozenmeal	frozenmeal

表 *24:* 按项目数排序的规则 *(*继续*)*	
前提条件	**结果**
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

“显示/隐藏条件”菜单。 显示/隐藏标准菜单按钮（网格图标）控制着规则的显示选项。 可用的显示选项如下：

- **实例数**显示出现完整序列（同时包含前提条件和结果）的唯一标识数的相关信息。 （请注意，此内容与关联模型不同，后者的实例数指的是其中 仅 条件适用的标识数。 例如，在给定规则 bread -> cheese 的情况下，同时包含 *bread* 和 *cheese* 的训练数据中的标识数称为**实例数**。
- **支持度**显示训练数据中前提条件为 true 的标识所占的比例。 例如，如果 50% 的训练数据包含前提条件 *bread* ，那么 bread -> cheese 规则的支持度为 50%。 （与关联模型不同，支持度不基于实例数，如前面所述）。
- **置信度**显示预测正确的标识在所有由规则进行了预测的标识中所占的百分比。 基于训练数据，该百分比的计算如下：包含整个序列的标识数量除以其中包含条件的标识数量。 例如，如果 50% 的训练数据包含 cannedveg（指示前提条件支持度），但只有 20% 同时包含 cannedveg 和 frozenmeal，那么规则 cannedveg -> frozenmeal 的置信度将为 Rule Support / Antecedent Support，或者在本例中，为 40%。
- 序列模型的**规则支持度**基于实例数，并显示整个规则、前提条件和结果都为 true 的训练记录的所占比例。 例如，如果有 20% 的训练数据同时包含 *bread* 和 *cheese*，那么规则 bread -> cheese 的规则支持度为 20%。

请注意，这些比例基于有效事务（至少具有一个观测项或真值的事务），而不基于总的事务。 在这些计算中不会考虑无效事务（没有项目或真值的事务）。

“过滤器” 按钮。 菜单上的“过滤器”按钮（漏斗图标）会扩展对话框的底部，从而显示一个面板，其中将显示活动的规则过滤器。 过滤器用于减少“模型”选项卡上显示的规则数量。

图 *52: “*过滤*”*按钮

要创建过滤器，请单击位于扩展面板右侧的过滤器图标。 这样将打开一个单独的对话框，您可以在其中指定用于显示规则的约束条件。 请注意，“过滤”按钮通常与“生成”菜单一起使用，以便首先过滤规则，然后生成一个包含部分规则的模型。 有关更多信息，请参阅以下第 196 页的『为规则指定过滤器』。

序列模型块设置

序列模型块的“设置”选项卡显示模型的评分选项。 此选项卡仅在模型添加到流工作区用于评分之后可用。

最大预测数。 指定每个购物篮项集合包括的最大预测数。 适用于此事务集合的、置信度值最高的规则将用于为记录生成预测，预测的数量不超过指定的上限。

序列模型块概要

序列规则模型块的“概要”选项卡显示发现的规则数量，以及规则的最大和最小支持度和置信度。 如果已执行附加到此建模节点的分析节点，那么分析信息也将显示在此选项卡上。

有关更多信息，请参阅主题 第 31 页的『浏览模型块』。

从序列模型块生成规则超节点

要基于序列规则生成规则超节点：

1. 在序列规则模型块的“模型”选项卡上，单击表中的某行以选择所需的规则。
2. 从规则浏览器菜单中选择：

   生成 > 规则超节点

重要：要使用生成的超节点，您必须在将数据传入超节点之前按标识字段（和时间字段，如果有的话）对数据进行排序。 超节点无法在未排序的数据中正确检测序列。

您可以指定下列用于生成规则超节点的选项：

检测。 指定如何为传入超节点的数据定义匹配项。

- **仅前项。** 每当超节点在具有同一标识的一组记录中发现选定规则的前提条件的顺序正确时，它都会确定一个匹配项，而不考虑是否还找到了结果。 请注意，此选项不考虑原始序列建模节点中的时间戳容差或项目间距约束设置。 在流中检测到最后一个条件项目集合（所有其他条件均以正确顺序发现）后，具有当前标识的所有后续记录都将包含下面选择的概要。
- **整个序列。** 每当超节点在具有同一标识的一组记录中发现选定规则的前提条件和结果的顺序正确时，它都会确定一个匹配项。 此选项不考虑原始序列建模节点中的时间戳记容差或项目间隔约束设置。 在流中检测到最后一个结果（所有条件均以正确顺序发现）后，当前记录和具有当前标识的所有后续记录都将包含下面选择的概要。

显示。 控制如何将匹配项摘要添加到规则超节点输出中的数据内。

- **第一次出现的结果值。** 添加到数据中的值是根据匹配项的首次出现预测的结果值。 这些值将作为一个名为 *rule_n_consequent* 的新字段进行添加，其中 *n* 为规则编号（基于流中规则超节点的创建顺序）。
- **第一次出现的 true 值。** 如果对于该标识至少存在一个匹配项，那么添加到数据中的值为 true；如果没有任何匹配项，那么添加的值为 false。 这些值将作为一个名为 *rule_n_flag* 的新字段添加。
- **出现次数。** 添加到数据中的值为该标识的匹配项数。 值将添加为新字段 *rule_n_count*。
- **规则编号。** 添加的值为选定规则的规则编号。 **规则编号**是根据超节点添加到流中的顺序指定的。 例如，第一个规则超节点被视为规则 *1*，第二个规则超节点被视为规则 *2*，以此类推。 如果流中包含多个规则超节点，那么此选项最有用。 这些值将作为一个名为 *rule_n_number* 的新字段添加。
- **包含置信度数字。** 如果选中此选项，那么会将规则置信度以及选定概要添加到数据流中。 这些值将作为一个名为 *rule_n_confidence* 的新字段添加。

“关联规则”节点

关联规则是以下格式的语句。

例如，“如果顾客购买了剃须刀和须后水，那么该顾客还会购买剃须膏，并且置信度为 80%”。 “关联规则”节点从数据中抽取一组规则，抽出的规则具有出现频率最高的信息内容。 “关联规则”节点与 Apriori 节点非常类似，但是，存在一些明显的差异：

- “关联规则”节点无法处理事务性数据。
- “关联规则”节点能够处理存储类型为“列表”且测量级别为“集合”的数据。
- “关联规则”节点可以与 IBM SPSS Analytic Server 配合使用。 这提供了可伸缩性，并且意味着您可以处理大型数据并利用速度更快的并行处理。
- “关联规则”节点提供了更多设置，例如能够限制生成的规则数，从而提高处理速度。
- 模型块的输出将显示在输出查看器中。

注: “关联规则”节点不支持 IBM SPSS 协作和部署服务 中的“模型评估”步骤或“冠军参选者”步骤。

注: 如果字段类型为标志，“关联规则”节点在构建模型时会忽略空的记录。 空的记录是模型构建中使用的所有字段值均为 false 的记录。

在 IBM SPSS Modeler 安装的 Demos 目录中，提供了名为 `geospatial_association.str` 的流，这个流显示了有关如何使用“关联规则”的有效示例，并引用数据文件 `InsuranceData.sav`、`CountyData.sav` 和 `ChicagoAreaCounties.shp`。 您可以从 Windows“开始”菜单上的 IBM SPSS Modeler 程序组访问 Demos 目录。 `geospatial_association.str` 文件位于 `streams` 目录中。

关联规则 - 字段选项

在**字段**选项卡上，您可以选择是使用上游节点（例如上一个“类型”节点）中已定义的字段角色设置，还是手动进行字段分配。

使用预定义角色
此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的角色设置（例如目标或预测变量）。具有输入角色的字段被视为条件，具有目标角色的字段被视为预测，而那些同时用作输入和目标的字段被视为具有这两种角色。

使用定制字段分配(C)
如果要在此屏幕上手动分配目标、预测变量和其他角色，请选中此选项。

字段
如果您选中了**使用定制字段分配**，那么使用方向按钮可以将此列表中的项手动分配到屏幕右侧的框。 图标指示每个字段的有效测量级别。

两者（条件或预测）
添加到此列表中的字段可以在模型所生成的规则中充当条件或预测角色。 这基于每条规则，因此，某个字段可能是一条规则中的条件，并且是另一条规则中的预测。

仅预测
添加到此列表中的字段只能显示为规则的预测（也称为“结果”）。 字段出现在此列表中并不表示在任何规则中使用了该字段，只要使用了该字段，那么它只能是预测。

仅条件
添加到此列表中的字段只能显示为规则的条件（也称为“前提条件”）。 字段出现在此列表中并不表示在任何规则中使用了该字段，只要使用了该字段，那么它只能是条件。

关联规则 - 规则构建

每条规则的项数

使用这些选项可以指定每条规则中可以使用的项或值的数目。

注: 这两个字段的总和不得超过 10。

最大条件数
选择单条规则中可以包含的最大条件数。

最大预测数
选择单条规则中可以包含的最大预测数。

规则构建

使用这些选项可以指定要构建的规则的数目和类型。

最大规则数
指定在为模型构建规则时可以考虑使用的最大规则数。

针对前 N 个的规则条件
选择用于建立前 N 条规则的条件，其中 N 是在**最大规则数**字段中输入的值。 您可以从下列条件中选择。

- **confidence**
- **规则支持度**
- **条件支持度**
- **Lift**
- **部署能力**

仅包含标志变量的真值
当数据采用表格格式时，选中此选项只会将标志字段的真值包括在生成的规则中。 选择真值可能有助于使规则更容易理解。 该选项不适用于事务格式的数据。 有关更多信息，请参阅 第 190 页的『表格数据与事务处理数据』。

规则条件

如果您选中**启用规则条件**，那么可以使用这些选项来选择最小强度，只有那些满足此强度的规则才能在模型中使用。

- **置信度** 指定模型所生成的规则的置信度级别的最小百分比值。 如果模型所生成的规则的置信度级别小于此数量，那么该规则将被废弃。
- **规则支持度** 指定模型所生成的规则的规则支持度级别的最小百分比值。 如果模型所生成的规则的置信度级别小于此数量，那么该规则将被废弃。
- **条件支持度** 指定模型所生成的规则的条件支持度级别的最小百分比值。 如果模型所生成的规则的条件支持度级别小于指定的数量，那么该规则将被废弃。
- **增益** 指定模型所生成的规则允许的最小增益值。 如果模型所生成的规则的值小于指定的数量，那么该规则将被废弃。

排除规则

在某些情况下，两个或两个以上字段之间的关联已知或者不证自明，在这种情况下，您可以排除其中的字段预测彼此的规则。 通过排除包含这两个值的规则，您可以减少不相关的输入并增加找到有用结果的机会。

字段
选择不想在规则构建中同时使用的关联字段。 例如，关联字段可能是“制造商”和“汽车型号”，或者是“学年”和“学生时代”。 当模型创建规则时，如果规则至少包含在规则的任一侧（条件或预测）选择的其中一个字段，那么该规则将被废弃。

关联规则 - 转换

离散化

使用这些选项可以指定如何对连续（数字范围）字段进行分箱。

分级数
所有设置为自动分箱的连续字段都划分为您指定的间距相等的箱数。 您可以选择 2 - 10 范围内的任意数字。

列表字段

最大列表长度
要在列表字段的长度未知时限制要包括在模型中的项数，请输入最大列表长度。 您可以选择 1 到 100 范围内的任何数字。 如果列表长度超过输入的数字，那么模型仍将使用此字段，但仅包括截至此数字为止的值；此字段中的所有其他值都将被忽略。

关联规则 - 输出

使用此窗格中的选项可以控制在构建模型时所生成的输出。

规则表

使用这些选项可以创建一种或多种表类型，这些表类型针对每个选定的条件显示最佳规则数（基于您指定的数目）。

置信度
置信度是规则支持度与条件支持度的比率。 在具有所列示条件值的项中，具有预测的结果值的百分比。 将创建一个包含最佳 N 条关联规则的表，这些规则基于输出中要包括的置信度（其中 N 是**要显示的规则数**值）。

规则支持(R)
整个规则、条件和预测均为 true 的项所占的比例。 对于数据集中所有的项，规则所正确解释并预测的百分比。 此度量给出规则的总体重要性。 将创建一个包含最佳 N 条关联规则的表，这些规则基于输出中要包括的规则支持度（其中 N 是**要显示的规则数**值）。

增益
规则置信度与具有预测的先验概率的比率。 规则的置信度值与结果值出现在总体中的百分比的比率。 此比率度量规则对机会的改进程度。 将创建一个包含最佳 N 条关联规则的表，这些规则基于输出中要包括的增益（其中 N 是**要显示的规则数**值）。

条件支持度
条件为 true 的项所占的比例。 将创建一个包含最佳 N 条关联规则的表，这些规则基于输出中要包括的前提条件支持度（其中 N 是**要显示的规则数**值）。

部署能力
用于度量训练数据中满足条件但不满足预测的部分所占的百分比。 此度量显示规则未命中的频率。 它实际上与置信度是相反的。 将创建一个包含最佳 N 条关联规则的表，这些规则基于输出中要包括的部署能力（N 是**要显示的规则数**值）。

要显示的规则数
设置表中要显示的最大规则数。

模型信息表

使用这些选项中的一个或多个选项可以选择输出中要包括的模型表。

- **字段转换**
- **记录摘要**
- **规则统计信息**
- **最频率的值**
- **最频率的字段**

规则的可排序字云。

使用这些选项可以创建显示规则输出的字云。 字以不断递增的文本大小显示，以表明其重要性。

创建可排序的字云。
选中此框将在输出中创建可排序的字云。

缺省排序
选择最初创建字云时要使用的排序类型。 字云是交互式的，您可以在模型查看器中更改条件以查看不同的规则和排序。 可以从下列排序选项中选择：

- 置信度。
- 规则支持(R)
- 增益
- 条件支持。
- 部署能力

要显示的最大规则数
设置字云中要显示的规则数；可以选择的最大值为 20。

关联规则 - 模型选项

使用此选项卡上的设置可以指定“关联规则”模型的评分选项。

模型名称 可以根据目标字段自动生成模型名称（未指定此类字段时，将根据模型类型生成模型名称），也可以指定定制名称。

最大预测数 指定可以包括在分数结果中的预测的最大数目。 将此选项与**规则条件**条目配合使用可生成“最佳”预测，其中“最佳”表示最高级别的置信度、支持度、增益等等。

规则条件 选择用于确定规则强度的度量。 规则将按此处选择的条件强度进行排序，以返回项集合的最佳预测。 您可以从 5 个不同的标准中选择。

- **置信度** 置信度是规则支持度与条件支持度的比率。 在具有所列示条件值的项中，具有预测的结果值的百分比。

- **条件支持度** 条件为 true 的项所占的比例。
- **规则支持度** 整个规则、条件和预测均为 true 的项所占的比例。 用**条件支持度**值乘以**置信度**值计算得出。
- **增益** 规则置信度与具有预测的先验概率的比率。
- **部署能力** 用于度量训练数据中满足条件但不满足预测的部分所占的百分比。

允许重复预测 要在评分期间包括多条具有相同预测的规则，请选中此复选框。 例如，选中此框将允许对下列规则进行评分。

```
bread & cheese -> wine
cheese & fruit -> wine
```

注: 仅当先前已预测所有预测 (`wine & pate`) 时，才会将具有多个预测 (`bread & cheese & fruit -> wine & pate`) 的规则视为重复预测。

只有在预测未出现在输入中时才对规则进行评分 要确保预测不会也出现在输入中，请选中此选项。 例如，如果进行评分的目的是为了进行一项家具产品推荐，那么已经包含餐桌的输入可能不会购买另一个这样的家具。 在这种情况下，请选中此选项。 但是，如果产品易腐烂或者是一次性的（如奶酪、婴儿代乳品或者卫生纸），那么结果已存在于输入中的规则可能有些价值。 在后面一种情况下，最有用的选项可能是**对所有规则进行评分**。

只有在预测出现在输入中时才对规则进行评分 要确保预测也出现在输入中，请选中此选项。 当您尝试深入了解现有的客户或事务时，此方法非常有用。 例如，您可能希望确定增益最高的规则，然后探索哪些客户符合这些规则。

对所有规则进行评分 要在评分期间包括所有规则，而无论是否存在预测，请选中此选项。

“关联规则”模型块

模型块包含模型构建期间从数据中提取的规则的相关信息。

查看结果

您可以使用此对话框的“模型”选项卡来浏览“关联规则”模型所生成的规则。 在生成新节点或者对模型进行评分前浏览模型块将显示有关规则的信息。

模型评分

经过优化的模型块可以添加到流中，用于进行评分。 有关更多信息，请参阅主题 第 35 页的『使用流中的模型块』。 用于评分的模型块在其各自的对话框中包括一个额外的“设置”选项卡。 有关更多信息，请参阅主题 第 212 页的『关联规则模型块设置』。

关联规则模型块详细信息

关联规则模型块显示输出查看器的“模型”选项卡中的模型的详细信息。 有关使用此查看器的更多信息，请参阅《Modeler 用户指南》(ModelerUsersGuide.pdf) 中标题为『处理输出』的部分。

GSAR 建模操作将创建多个具有前缀 $A 的新字段，如下表所示。

表 *25:* 关联规则建模操作所创建的新字段

字段名称	描述
$A-<prediction>#	此字段包含模型对已评分记录的预测。 <prediction> 是包括在模型的“预测”角色中的字段名称，而 # 是输出规则的编号序列（例如，如果将得分设置为包括 3 条规则，那么编号序列将为 1 到 3）。

表 *25:* 关联规则建模操作所创建的新字段 *(*继续*)*	
$AC-<prediction>#	此字段包含预测中的置信度。 <prediction> 是包括在模型的“预测”角色中的字段名称，而 # 是输出规则的编号序列（例如，如果将得分设置为包括 3 条规则，那么编号序列将为 1 到 3）。
$A-Rule_ID#	此字段包含针对已评分数据集中每个记录进行预测的规则的标识。 # 是输出规则的编号序列（例如，如果将得分设置为包括 3 条规则，那么编号序列将从 1 到 3）。

关联规则模型块设置

“关联规则”模型块的“设置”选项卡显示模型的评分选项。 只有在将模型添加到流工作区中用于评分之后，此选项卡才可用。

最大预测数 指定为每个项集合包括的最大预测数。 适用于此事务集合的、置信度值最高的规则将用于为记录生成预测，预测的数量不超过指定的上限。 将此选项与**规则条件**选项配合使用可生成“最佳”预测，其中最佳指的是最高级别的置信度、支持度、增益等等。

规则条件 选择用于确定规则强度的度量。 规则将按此处选择的条件强度进行排序，以返回项集合的最佳预测。 您可以从下列标准中选择。

- **confidence**
- **规则支持度**
- **Lift**
- **条件支持度**
- **部署能力**

允许重复预测 要在评分时包括多条具有相同结果的规则，请选中此复选框。 例如，选中此选项表示可以对下列规则进行评分：

```
bread & cheese -> wine
cheese & fruit -> wine
```

要在评分时排除重复预测，请取消选中此复选框。

注: 仅在先前已预测所有结果 (`wine & pate`) 时，才会将具有多个结果 (`bread & cheese & fruit -> wine & pate`) 的规则视为重复预测。

只有在预测未出现在输入中时才对规则进行评分 选中此框可确保结果不会也出现在输入中。 例如，如果进行评分的目的是为了进行一项家具产品推荐，那么已经包含餐桌的输入可能不会购买另一个这样的家具。 在这种情况下，请选中此选项。 另一方面，如果产品易腐烂或者是一次性的（如奶酪、婴儿代乳品或者卫生纸），那么结果已存在于输入中的规则可能有些价值。 在后面一种情况下，最有用的选项可能是**对所有规则进行评分**。

只有在预测出现在输入中时才对规则进行评分 选中此框可确保结果也出现在输入中。 当您尝试深入了解现有的客户或事务时，此方法非常有用。 例如，您可能希望确定增益最高的规则，然后探索哪些客户符合这些规则。

对所有规则进行评分 要在评分时包括所有规则，而无论结果是否出现在输入中，请选中此选项。

第 13 章 时间序列模型

为何进行预测?

预测的意思就是对一个或多个序列在一定时间内的值进行预言。 例如，您可能希望预测某个系列产品或服务的预期需求，以便分配资源进行制造或配送。 因为计划决策的实施需要时间，所以预测在很多计划过程中都是一个必不可少的工具。

时间序列建模方法假定历史记录总会自我重演，即使不是完全一样也会非常接近，足以通过研究过去对将来作出更好的决策。 例如，为了预测下一年的销售量，您可能得从分析今年的销售量开始，看看近年来都有哪些发展趋势或模式（如果存在的话）。 但模式可能很难测量。 例如，如果您的销售量在几周之内连续上升，那么这是季节性原因呢还是一种长期趋势的开始?

使用统计建模技术，可以分析过去数据中存在的模式并加以预测，以确定该序列的未来值可能属于的范围。 其结果是您的决策所依据的预测更为准确。

时间序列数据

时间序列 是以规律的时间间隔采集的测量值的有序集合，例如，每日的股票价格或每周的销售数据。 测量值可以是您感兴趣的任何内容，每个序列通常可以归为下列类别之一：

- **从属**。 要预测的序列。
- **预测变量**。 这是可能有助于解释目标的序列，例如使用广告预算来预测销售量。 预测变量只能用于 ARIMA 模型。
- **事件**。 一种特殊的预测变量序列，用于说明可预测的重复发生事件，例如促销活动。
- **干预**。 一种特殊的预测变量序列，用于说明一次性事件，例如停电或员工罢工。

时间间隔可以代表任何时间单位，但所有测量值的时间间隔必须相同。 而且，没有测量值的任何时间间隔必须设置为缺失值。 因此，有测量值的时间间隔数（包括测量值为缺失值的时间间隔）定义数据历史记录范围的时间长度。

时间序列的特征

研究序列过去的行为有助于辨别其中的模式从而作出更好的预测。 将其绘制成图时，许多时间序列就会表现出下列一种或多种特征：

- 趋势
- 季节周期和非季节周期
- 脉冲和步进
- 离群值

趋势

趋势 是指序列水平的逐渐上升或下降或序列值随时间的推移而增大或减小的趋势。

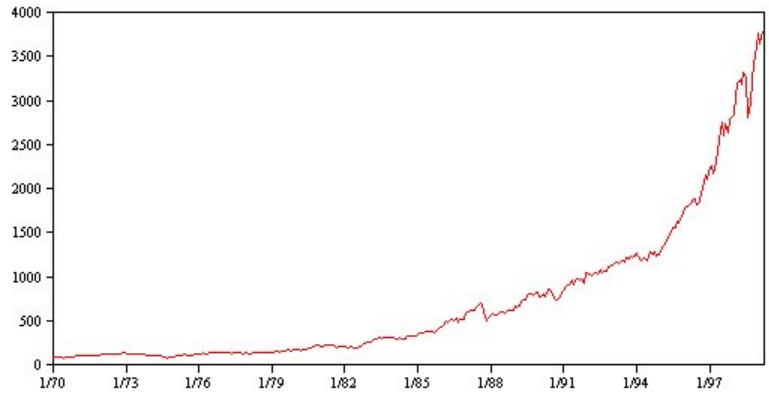


图 *53:* 趋势

趋势既可以是 **局部** 的，也可以是 **全局** 的，而一个序列可以同时体现这两种趋势。 从历史记录来看，股票市场指数的序列图总的趋势是上升的。 经济萧条时期所表现出的是局部下降趋势，而经济繁荣时期表现出的是局部上升趋势。

趋势既可以是 **线性** 的，也可以是 **非线性** 的。 线性趋势是指序列水平表现为正增加或负增加，就和本金以单利计息差不多。 非线性趋势通常表现为倍增，即相对于以前的序列值成比例地增长。

全局线性趋势可通过指数平滑模型和 ARIMA 模型很好地拟合和预测。 在构建 ARIMA 模型的过程中，通常会对表现出趋势的序列进行区分，以消除趋势的影响。

季节性周期

季节性周期 是序列值中可预测的重复模式。

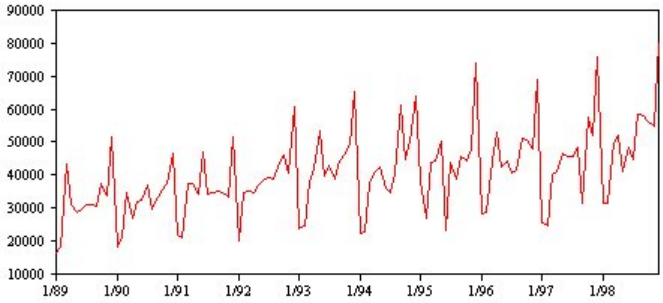


图 *54:* 季节性周期

季节性周期与序列的时间间隔相联系。 例如，月度数据通常会随季度和年度而循环。 月度序列可能会表现出第一个季度较低的明显季度周期或每年十二月份都出现峰值的年度周期。 表现出季节性周期的序列称之为具有 **季节性** 。

季节模式对于获取良好的拟合和预测非常有用，用来捕获季节性的有指数平滑模型和 ARIMA 模型。

非季节性周期

非季节性周期是序列值中可能无法预测的重复模式。

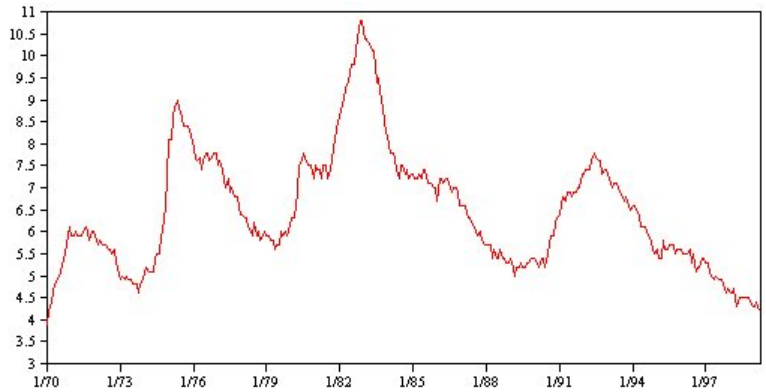


图 *55:* 非季节性周期

某些序列（如失业率）明显地表现出周期性行为；但这种周期性的周期会随时间而变化，因此很难预测何时高何时低。 其他序列可能具有可预测的周期，但可能与阳历并不完全吻合，或者其周期比一年长。 例如，潮汐遵循阴历，与奥林匹克运动会相关的国际旅游和贸易每隔四年膨胀一次，还有许多宗教节日，其阳历日期每年都会变化。

非季节周期模式很难建模，通常会增加预测的不确定性。 例如，股票市场的许多序列实例就常使预测者的努力无功而返。 即便如此，当存在非季节模式时，还是有必要加以说明。 在许多情况下，您仍然可以找出与历史数据拟合得很好的模型，从而最大限度地减小预测中的不确定性。

脉冲和步进

许多序列都会出现水平突变。 它们通常分为两种类型：

- 序列水平突然、 临时性的变动，或称 **脉冲**
- 序列水平突然、 永久性的变动，或称 **步进**

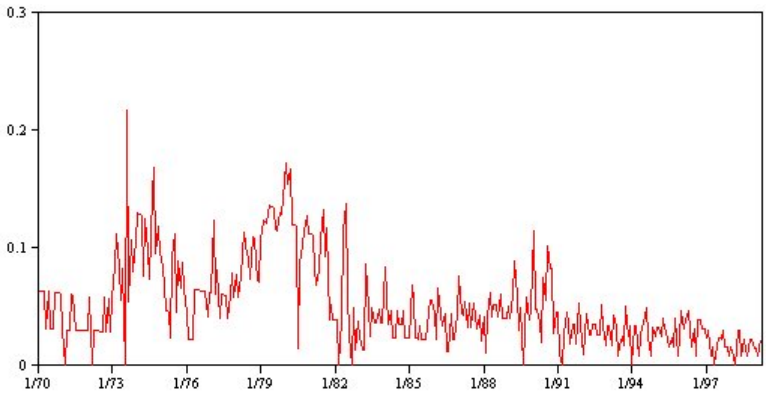


图 *56:* 脉冲序列

观测到步进或脉冲时，找到一种貌似合理的解释很重要。 时间序列模型是用来说明渐变而非突变的。 因此，它们往往低估脉冲并为步进所瓦解，导致模型拟合差强人意，增加预测的不确定性。 （某些季节性实例可能表现为突然的水平变化，但该水平在不同的季节周期之间则保持稳定。）

如果扰动是可以解释的，那么可以使用 **干预** 或 **事件** 为其建模。 例如，1973 年 8 月，石油输出国组织 (OPEC) 颁布的石油禁运导致了通货膨胀率的急剧变化，经过数月之后才恢复到正常水平。 通过为该禁运月指定一个 **点干预** ，可以改善模型的拟合度，因此可以间接提高预测的准确性。 例如，某个零售店可能会发现，所有商品均标记降价 50% 的当天销售量比平时高出很多。 通过将降价 50% 的促销指定为一个定期的 **事件** ，可以改善模型的拟合度，估计将来重复该项促销措施的影响。

离群值

时间序列水平中无法解释的变动称为 **离群值** 。 这些观测值与序列中的其他值不一致，可能会显著影响分析，从而影响时间序列模型的预测能力。

下图显示了时间序列中常见的几种离群值。 蓝线表示没有离群值的序列。 红线表示如果序列包含离群值情况下可能存在的模式。 这些离群值全部归为 **确定性** 离群值，因为它们只影响序列的均值水平。

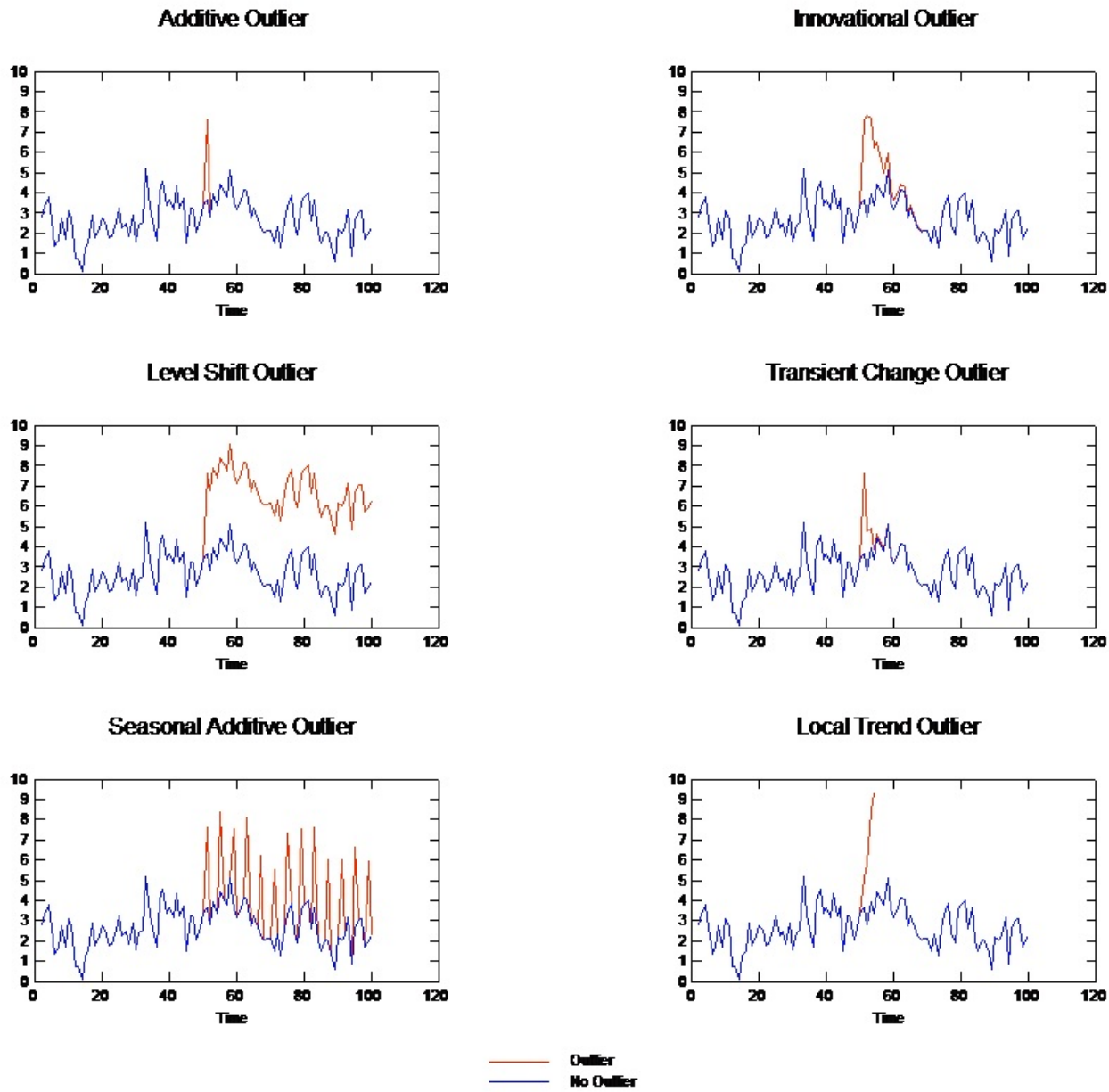


图 *57:* 离群值类型

- **加性离群值**。 加性离群值表现为一次观测中出现的异常大或异常小的值。 后续观测不受加性离群值的影响。 连续的加性离群值通常称为**加法离群值修补**。
- **创新离群值**。 创新离群值的特征为初始影响一直对后续观测产生作用。 这些离群值的影响可能会随着时间的推移而不断增强。
- **水平变动离群值**。 对于水平变动，离群值之后出现的所有观测值均移动到新水平。 与加性离群值相反，水平变动离群值会影响许多观测值，并且具有永久性影响。
- **瞬时变化离群值**。 瞬时变化离群值类似水平变动离群值，只是这种离群值对后续观测的影响呈指数递减。最终，该序列会恢复到正常水平。
- **季节加性离群值**。 季节加性离群值表现为以固定时间间隔重复出现的异常大或异常小的值。
- **局部趋势离群值**。 局部趋势离群值会在出现初始离群值之后，在序列中产生一个由离群值中的模式所导致的整体漂移。

时间序列中的离群值检测包括确定存在的任何离群值的位置、类型和大小。 Tsay (1988) 提出了一个用于检测均值水平变化以识别出确定性离群值的迭代过程。 此过程是将一个假设不存在离群值的时间序列模型与另一个具有离群值的模型进行比较。 从两个模型之间的差异得到将任何给定点视为离群值的影响的估计。

自相关函数和偏自相关函数

自相关和偏自相关用于测量当前序列值和过去序列值之间的相关性，并指示预测将来值时最有用的过去序列值。 了解了此内容，您就可以确定 ARIMA 模型中过程的顺序。 更具体来说，

- **自相关函数 (ACF)**。 延迟为 *k* 时，这是相距 *k* 个时间间隔的序列值之间的相关性。
- **偏自相关函数 (PACF)**。 延迟为 *k* 时，这是相距 *k* 个时间间隔的序列值之间的相关性，同时考虑了间隔之间的值。

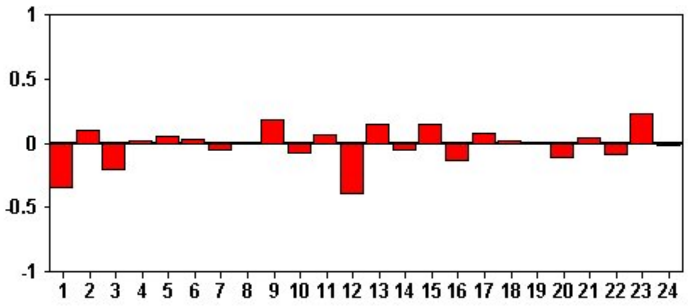


图 *58:* 序列的 *ACF* 图

ACF 图的 *x* 轴表示计算自相关处的延迟； *y* 轴表示相关值（介于 –1 和 1 之间）。 例如，ACF 图中延迟 1 处的峰值表示每个序列值与前面的值强相关，延迟 2 处的峰值表示每个值与以前两个点之间的值强相关，依此类推。

- 正相关表示较大的当前值与指定延迟处较大的值相对应；负相关表示较大的当前值与指定延迟处较小的值相对应。
- 相关的绝对值是关联强度的测量，绝对值越大表明关系越强。

序列转换

变换对在模型估计之前稳定序列常常有用。 这对 ARIMA 模型尤其重要，因为估计这类模型之前需要序列保持 **稳定** 。 如果在整个序列中，全局水平（均值）以及与该水平的平均偏差（方差）保持不变，那么该序列是稳定的。

尽管多数令人感兴趣的序列都不稳定，但只要能够通过应用变换（如，自然对数、差分或季节差分）使序列保持稳定，那么 ARIMA 就是有效的。

方差稳定变换。 方差随时间变化的序列通常可以使用自然对数变换或平方根变换来保持稳定。 这些变换也称为函数变换。

- **自然对数**。 对序列值取自然对数。
- **平方根**。 对序列值应用平方根函数。

自然对数变换和平方根变换不能用于具有负值的序列。

水平稳定变换。 ACF 中值的缓慢下降表示每个序列值都与上一个值具有很强的相关性。 通过分析序列值的变化，您可以获得一个稳定水平。

- **简单差分**。 计算序列中每个值与上一个值之间的差，序列中最旧的值除外。 这意味着经过差分的序列将比原始序列少一个值。
- **季节差分**。 除计算每个值与上一个季节值之间的差值外，其他均与简单差分相同。

将简单差分或季节差分同时用于对数变换或平方根变换时，总是先应用方差稳定变换。 同时使用简单差分和季节差分时，无论首先应用简单差分还是季节差分，得到的序列值均相同。

预测变量序列

预测变量序列包括可能有助于解释要预测序列的行为的相关数据。 例如，一个网上零售商或目录零售商可能会根据邮寄的目录数量、开通的电话数量或公司网页的点击次数来预测销售量。

任何序列都可以作为预测变量，条件是该序列须延伸到要预测的将来时间，并且具有不存在缺失值的完整数据。

向模型中添加预测变量时以慎重为宜。 添加大量预测变量会增加估计模型所需的时间。 虽然添加预测变量可以提高模型拟合历史数据的能力，但并不意味着该模型就一定能产生更好的预测结果，因为增加的复杂怕有可能及不上所造成的麻烦。 理想的目标是，找出的模型既是最简单的，同时又能作出很好的预测。

一般而言，建议预测变量的数量应小于样本大小除以 15（即最多每 15 个观测值一个预测变量）。

缺失数据的预测变量。 不能在预测中使用包含不完整数据或缺失数据的预测变量。 这适用于历史数据和将来值。 在某些情况下，可通过设置模型的估计范围以便在估计模型时排除最旧数据来避免上述限制。

空间-时间预测建模节点

空间-时间预测 (STP) 有许多潜在的应用，例如紧急管理建筑物或设施、对机械服务工程师进行绩效分析和预测或者进行公共交通规划。 在这些应用中，通常要对空间和时间进行能耗等测量。 可能与记录这些测量值相关的问题包括哪些因子影响未来的观测值、如何实现所需的变化或者如何更好地管理系统？ 为了回答这些问题，您可以在不同位置使用能够预测未来值的统计技术，并可以显式地对可调因子进行建模以执行假设情况分析。

STP 分析使用包含位置数据、预测输入字段（预测变量）、时间字段和目标字段的数据。 每个位置在数据中都有许多行，这些行表示每个预测变量在每个测量时间的值。 分析数据后，可以使用该数据来预测分析中使用的形状数据内任意位置处的目标值。 并且，还可以预测何时能够获知未来时间点的输入数据。

注: STP 节点不支持 IBM SPSS 协作和部署服务 中的“模型评估”步骤或“冠军参选者”步骤。

IBM SPSS Modeler 安装的 Demos 目录中提供一个流，该流显示使用 STP 的已处理示例（名为 `stp_server_demo.str`），并且引用数据文件 `room_data.csv` 和 `score_data.csv`。 您可以从 Windows"开始"菜单上的 IBM SPSS Modeler 程序组访问 Demos 目录。 `stp_server_demo.str` 文件位于 `streams` 目录中。

空间-时间预测 - 字段选项

在“字段”选项卡上，您可以选择是使用上游节点中已定义的字段角色设置还是手动进行字段分配。

使用预定义角色
此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的角色设置（仅限目标和预测变量）。

使用定制字段分配(C)
要在此屏幕上手动分配目标、预测变量和其他角色，请选中此选项。

字段
显示数据中所有可以选择的字段。 使用方向按钮可以将此列表中的项手动分配到屏幕右侧的各种框。 图标指示每个字段的有效测量级别。

注: 对于每个位置每个时间间隔，STP 都需要 1 条记录才能正常运行；因此，这些是必需字段。

在**字段**窗格底部，单击**全部**按钮将选中所有字段（这与测量级别无关），单击单独的测量级别按钮将选中所有具有该测量级别的字段。

目标
选择一个字段作为预测目标。

注: 您只能选择测量级别为“连续”的字段。

位置
选择要在模型中使用的位置类型。

注: 您只能选择测量级别为“地理空间”的字段。

位置标签
形状数据通常包含一个表明层特征的名称的字段，例如，这可能是省/自治区/直辖市或者国家或地区的名称。 使用此字段可以将名称或标签与位置相关联，方法是选择一个分类字段来标注输出中的所选**位置**字段。

时间字段
选择要在预测中使用的时间字段。

注: 您只能选择测量级别为“连续”且存储类型为时间、日期、时间戳记或整数的字段。

预测变量（输入）
选择一个或多个字段作为预测输入。

注: 您只能选择测量级别为“连续”的字段。

空间-时间预测 - 时间间隔

在“时间间隔”窗格中，您可以选择用于设置时间间隔和随时间推移进行的任何必需汇总的选项。

在您可以构建 STP 模型之前，需要进行数据准备以便将时间字段转换为索引；要使得能够进行这种转换，时间字段中的记录之间必须有固定的区间。 如果数据尚未包含此信息，请使用此窗格中的选项来设置此区间，然后才能使用建模节点。

时间间隔 请选择要将数据集转换为的时间间隔。 可用的选项取决于在“字段”选项卡上选择作为模型的**时间字段**的字段存储类型。

- **周期** 仅适用于整数时间字段；这是一系列与任何其他可用时间间隔均不匹配的时间间隔，并且每项测量之间的间距一致。
- **年** 仅可用于“日期”或“时间戳记”时间字段。
- **季度** 仅可用于“日期”或“时间戳记”时间字段。 如果您选择了此选项，那么系统将提示您选择第一个季度的**开始月份**。
- **月** 仅可用于“日期”或“时间戳记”时间字段。
- **周** 仅可用于“日期”或“时间戳记”时间字段。
- **天** 仅可用于“日期”或“时间戳记”时间字段。
- **小时** 仅可用于“时间”或“时间戳记”时间字段。
- **分钟** 仅可用于“时间”或“时间戳记”时间字段。
- **秒** 仅可用于“时间”或“时间戳记”时间字段。

当您选择了**时间间隔**时，系统将提示您填写更多字段。 可用的字段同时取决于时间间隔和存储类型。 以下列表显示了可能会显示的字段。

- **每周天数**
- **每天小时数**
- **一周的开始日期** 一周的第一天
- **一天的开始时间** 新的一天的开始时间。
- **时间间隔值** 您可以选择下列其中一个选项：1、2、3、4、5、6、10、12、15、20 或 30。
- **开始月份** 财年的开始月份。
- **开始期次** 如果使用了**期间**，请选择开始期次。

数据与指定的时间间隔设置匹配 如果数据已包含正确的时间间隔信息，并且不需要进行转换，请选中此复选框。 选中此框后，**汇总**区域中的字段将不可用。

汇总

只有在您取消选中**数据与指定的时间间隔设置匹配**复选框时才可用；请指定用于汇总字段以便与指定区间匹配的选项。 例如，如果有以周和月为单位的混合数据，那么可以对周值进行汇总或“累计”，以获得均匀的月间隔。 请选择要用于汇总不同字段类型的缺省设置，并创建您希望用于任何特定字段的任何定制设置。

- **连续** 设置要应用于未逐个指定的所有连续字段的缺省汇总方法。 您可以从下列方法中选择：

- 总和
- 平均值
- 最小值
- 最大值
- 中位数
- 第一个四分位数
- 第三个四分位数

所指定字段的定制设置 要将特定汇总函数应用于个别字段，请在此表中选择字段并选择汇总方法。

- **字段** 使用**添加字段**按钮以显示**"选择字段"**对话框并选择所需字段。 所选择的字段将显示在此列中。
- **汇总函数** 从下拉列表中，选择用于将字段转换为指定时间间隔的汇总函数。

空间-时间预测 - 基本构建选项

使用此对话框中的设置可以设置基本模型构建选项。

模型设置

包括截距
包括截距（模型中的常数项）可以提高解的总体准确度。 如果您可以假设数据穿过原点，那么可以排除截距。

最大自回归阶

自回归阶指定使用哪些先前值来预测当前值。 使用此选项可以指定用于计算新值的先前记录数。 您可以选择介于 1 与 5 之间的任意整数。

空间协方差

估计方法
选择要使用的估计方法；您可以选择**参数**或**非参数**。 对于**参数**方法，您可以从三种**模型**类型中进行选择：

- **Gaussian**
- **Exponential**
- **幂指** 如果选择此选项，那么还必须指定要使用的**幂**级别。 此级别可以是介于 1 与 2 之间的任意值，并以 0.1 为增量变动。

空间-时间预测 - 高级构建选项

熟悉 STP 的用户可以使用下列选项对模型构建过程进行微调。

缺失值的最大百分比
指定模型中可以包括的包含缺失值的记录所占的最大百分比。

模型构建中用于假设检验的显著性水平
指定用于 STP 模型估计的所有检验（包括两项拟合优度检验、效应 F 检验和系数 T 检验）的显著性水平值。 此级别可以是 0 与 1 之间的任何值，并以 0.01 为增量变动。

空间-时间预测 - 输出

在构建模型之前，请使用此窗格中的选项来选择要包括在输出查看器中的输出。

模型信息

模型规范
选中此选项表示将模型规范信息包括在模型输出中。

时间信息摘要
选中此选项表示将时间信息摘要包括在模型输出中。

评估

模型质量
选中此选项表示将模型质量包括在模型输出中。

均值结构模型中的效应检验
选中此选项表示将效应检验信息包括在模型输出中。

解释

模型系数的均值结构
选中此选项表示将均值结构模型系数信息包括在模型输出中。

自回归系数
选中此选项表示将自回归系数信息包括在模型输出中。

空间衰变检验
选中此选项表示将空间协方差（即空间衰变）检验信息包括在模型输出中。

参数空间协方差模型参数图
选中此选项表示将参数空间协方差模型参数图信息包括在模型输出中。

注: 仅当您在“基本”选项卡上选择了**参数**估计方法时，此选项才可用。

相关性热图
选中此选项表示将目标值的图包括在模型输出中。

注: 如果模型中包含 500 个以上的位置，那么不会创建图输出。

相关性图
选中此选项表示将相关性的图包括在模型输出中。

注: 如果模型中包含 500 个以上的位置，那么不会创建图输出。

位置聚类
选中此选项表示将位置聚类输出包括在模型输出中。 只有不需要访问图数据的输出才会包括在聚类输出中。

注: 只能为非参数空间协方差模型创建此输出。

如果选择了此选项，那么可以设置下列各项：

- **相似度阈值** 请选择阈值，达到此阈值的输出聚类将被视为足够相似，从而合并到单个聚类中。
- **要显示的最大聚类数** 请设置模型输出中可以包括的聚类数的上限。

空间-时间预测 - 模型选项

模型名称 可以根据目标字段自动生成模型名称，或者指定定制名称。 自动生成的名称为目标字段名。

不确定性因子 (%) 不确定性因子是一个百分比值，表示预测未来时的不确定性增幅。 随着进入未来的每个步骤，预测不确定性的上下限将按此百分比递增。 请设置要应用于模型输出的不确定性因子；这将设置预测值的上下边界。

空间-时间预测模型块

空间-时间预测 (STP) 模型块在输出查看器的“模型”选项卡中显示模型的详细信息。 有关使用此查看器的更多信息，请参阅《Modeler 用户指南》(ModelerUsersGuide.pdf) 中标题为『处理输出』的部分。

空间-时间预测 (STP) 建模操作将创建多个具有 $STP- 前缀的新字段，如下表所示。

表 *26: STP* 建模操作创建的新字段

字段名称	描述

表 *26: STP* 建模操作创建的新字段 *(继续)*	
$STP-<Time>	在模型构建过程中创建的时间字段。 “构建选项”选项卡的“时间间隔”窗格中的设置确定了创建此字段的方式。 <Time> 是“字段”选项卡上选择作为**时间字段**的字段的原始名称。 **注:** 只有在模型构建过程中转换了原始**时间字段**的情况下，才会创建此字段。
$STP-<Target>	此字段包含目标值的预测。 <Target> 是模型的原始**目标**字段的名称。
$STPVAR-<Target>	此字段包含 VarianceOfPointPrediction 值。 <Target> 是模型的原始**目标**字段的名称。
$STPLCI-<Target>	此字段包含 LowerOfPredictionInterval 值（即置信度下限）。 <Target> 是模型的原始**目标**字段的名称。
$STPUCI-<Target>	此字段包含 UpperOfPredictionInterval 值（即置信度上限）。 <Target> 是模型的原始**目标**字段的名称。

空间/时间预测模型设置

使用“设置”选项卡可以控制您认为建模操作中可接受的不确定性级别。

不确定性因子 (%) 不确定性因子是一个百分比值，表示预测未来时的不确定性增幅。 随着进入未来的每个步骤，预测不确定性的上下限将按此百分比递增。 请设置要应用于模型输出的不确定性因子；这将设置预测值的上下边界。

TCM 节点

使用此节点可以创建时间因果模型 (TCM)。

时间因果模型

时间因果建模会尝试发现时间序列数据中的重要因果关系。 在时间因果建模中，您指定一组目标序列以及这些目标的候选输入集。 这样，过程将为每个目标构建一个自回归时间序列模型，并且仅包括那些与目标具有因果关系的输入。 这种方法不同于传统的时间序列建模，在传统建模方法中，您必须明确指定目标序列的预测变量。 由于时间因果建模通常涉及为多个相关的时间序列构建模型，因此结果称为模型系统。

在时间因果建模上下文中，术语因果指的是格兰杰因果。 对于时间序列 X 和时间序列 Y，如果同时依据 X 和 Y 的过去值对 Y 进行回归所产生的 Y 模型比仅对 Y 的过去值进行回归所产生的模型要好，那么将时间序列 X 称为时间序列 Y 的“格兰杰原因”。

注: 时间因果模型节点不支持在 IBM SPSS 协作和部署服务 中执行模型评估或优胜参选者步骤。

示例

业务决策制定者可以使用时间因果建模来发现描述业务的大量基于时间的度量中的因果关系。 分析可能会揭示几个可控输入，这些输入对关键绩效指标的影响最大。

大型 IT 系统的管理者可以使用时间因果建模在大量相互关联的操作度量中检测异常。 这样，通过因果模型不仅可以检测异常，还能发现这些异常的最有可能的根本原因。

字段要求

必须至少有一个目标。 缺省情况下，将不使用预定义角色为“无”的字段。

数据结构

时间因果建模支持两种类型的数据结构。

基于列的数据
对于基于列的数据，每个时间序列字段都包含单个时间序列的数据。 此结构是时间序列建模器所使用的传统时间序列数据结构。

多维数据
对于多维数据，每个时间序列字段都包含多个时间序列的数据。 这样，特定字段内的单独时间序列将由类别字段（称为维度字段）的一组值标识。例如，两种不同的销售渠道（零售和 Web）的销售数据可能存储在单个 *sales* 字段中。 名为 *channel* 且值为“retail”和“web”的维度字段标识与这两种销售渠道中的每种渠道关联的记录。

注: 要构建时间因果模型，您需要足够的数据点。 产品使用以下约束：

```
m>(L + KL + 1)
```

其中，m 是数据点的数量，L 是延迟数，而 K 是预测变量数。 确保数据集足够大，从而使数据点的数量 (m) 满足条件。

要建模的时间序列

在“字段”选项卡上，请使用**时间序列**设置来指定要包括在模型系统中的序列。

请选择应用于数据的数据结构选项。 对于多维数据，请单击**选择维度**以指定维度字段。 指定维度字段的顺序定义了这些字段在所有后续对话框和输出中的显示顺序。 请使用“选择维度”子对话框上的向上和向下箭头按钮对维度字段进行重新排序。

对于基于列的数据，术语序列的含义与术语字段相同。 对于多维数据，包含时间序列的字段称为度量字段。 对于多维数据，时间序列由度量字段以及每个维度字段的值定义。 对于基于列的数据和多维数据，注意事项如下所示。

- 将对指定为候选输入或同时作为目标和输入的序列加以考虑，以便将其包括在每个目标的模型中。 每个目标的模型都始终包含该目标自身的延迟值。
- 指定为强制输入的序列将始终包括在每个目标的模型中。
- 必须将至少一个序列指定为目标或者同时指定为目标和输入。
- 如果选择了**使用预定义角色**，那么角色为“输入”的字段将设置为候选输入。 没有任何预定义角色映射到强制输入。

多维数据

对于多维数据，请在网格中指定度量字段及相关角色，网格中的每一行都指定单个度量和角色。 缺省情况下，模型系统包含此网格中每一行的所有维度字段组合的序列。 例如，如果存在 *region* 维度和 *brand* 维度，那么在缺省情况下，指定度量 *sales* 作为目标意味着对于 *region* 与 *brand* 的每个组合，都存在单独的 sales 目标序列。

对于网格中的每一行，您可以通过单击维度的省略号按钮来定制任何维度字段的值集合。 此操作将打开“选择维度值”子对话框。 另外，您还可以添加、删除或复制网格行。

序列计数列显示当前对相关度量指定的维度值集合的数目。 显示的值可能大于序列的实际数目（每个集合各有一个对应的序列）。 当指定的某些维度值组合未与相关度量所包含的序列相对应时，将发生这种情况。

选择维度值

对于多维数据，您可以通过指定哪些维度值将应用于具有特定角色的特定度量字段来定制分析。 例如，如果 *sales* 是一个度量字段，*channel* 是值为“retail”和“web”的维度，那么您可以指定“web”销售是输入，“retail”销售是目标。 另外，还可以指定将应用于分析中使用的所有度量字段的维度子集。 例如，如果 *region* 是指示地理区域的维度字段，那么您可以将分析限制在特定区域。

所有值
指定包括当前维度字段的所有值。 这是缺省选项。

选择要包括或排除的值

使用此选项可以指定当前维度字段的值集。 如果对**方式**选择了**包括**，那么将仅包括**选择的值**列表中指定的值。 如果对**方式**选择了**排除**，那么将包括除**选择的值**列表中指定的值之外的所有值。

您可以对要从中进行选择的值集进行过滤。 满足过滤条件的值将显示在**匹配**选项卡中，不满足过滤条件的值将显示在**未选择的值**列表的**不匹配**选项卡中。 **全部**选项卡列示所有未选择的值，而无论指定了什么过滤条件。

- 指定过滤器时，可以使用星号 (*) 来表示通配符。
- 要清除当前过滤器，请在“过滤显示的值”对话框中对搜索项指定空值。

观测值

在“字段”选项卡上，使用**观测**设置来指定用于定义观测的字段。

由日期/时间定义的观测值

您可以指定观测值由日期、时间或时间戳记字段定义。 除了用于定义观测值的字段以外，请选择用于描述观测值的适当时间间隔。 根据指定的时间间隔，您还可以指定其他设置，例如两次观测之间的时间间隔（增量）或每周的天数。 以下注意事项适用于时间间隔：

- 如果各个观测值之间的时间间距不定期（例如处理销售订单的时间），请使用**不定期**值。 选择**不规则**时，必须在“数据规范”选项卡上的**时间间隔**设置中指定用于分析的时间间隔。
- 如果观测值表示日期和时间，并且时间间隔为小时、分钟或秒，请使用**每天的小时数**、**每天的分钟数**或**每天的秒数**。 如果观测值表示时间（持续时间）并且未引用日期，而时间间隔为小时、分钟或秒，请使用**小时数（非周期性）**、**分钟数（非周期性）**或**秒数（非周期性）**。
- 根据选择的时间间隔，此过程可以检测缺失的观测值。 由于此过程假定所有观测值之间的时间间距相等，并假定未缺失观测值，因此有必要检测缺失的观测值。 例如，如果时间间隔为“天”，并且日期 2014-10-27 后面跟着 2014-10-29，那么表明缺失 2014-10-28 的观测值。 对于任何缺失的观测值，将插补值。 您可以在“数据规范”选项卡上指定用于处理缺失值的设置。
- 指定的时间间隔使此过程能够检测到同一时间间隔内的多个需要汇总到一起的观测值，并使各个观测值使用统一的时间间隔边界（例如每个月的第一天），以确保各个观测值之间的间距相等。 例如，如果时间间隔为“月”，那么同一个月内的多个日期将聚集到一起。 此类汇总称为分组。 缺省情况下，分组时将计算观测值的总和。 通过“数据指定项”选项卡上的**汇总和分布**设置，您可以指定另一种分组方法，例如计算各个观测值的平均值。
- 对于某些时间间隔，附加设置可以定义正常等间距时间间隔中的中断。 例如，如果时间间隔为“天”，但仅工作日有效，那么可以指定一周有五天，每周第一天为星期一。

周期或循环周期定义的观测

观测可以由一个或多个表示周期或周期反复循环（直至达到任意数目的循环级别为止）的整数字段定义。 借助此结构，您可以描述任何标准时间间隔都无法支持的观测值序列。 例如，要描述只有 10 个月的财年，可以使用表示年的循环字段和表示月的周期字段，并且一个循环的长度为 10。

指定循环周期的字段定义周期性级别层次结构，最低级别由**周期**字段定义。 次高级别由级别为 1 的循环字段指定，接着由级别为 2 的循环字段指定，依此类推。 除最高级别以外，每个级别的字段值对于次高级别都必须具有周期性。 最高级别的值不得具有周期性。 例如，如果是 10 个月的财年，年中的月份是周期性值，而年不是周期性的。

- 特定级别的循环长度是下一个最低级别的周期。 在财年示例中，只有一个循环级别，并且循环长度为 10，这是因为次低级别表示月，而指定的财年包含 10 个月。
- 指定不从 1 开始的任何周期字段的起始值。 此设置是检测缺失值的必备步骤。 例如，如果周期性字段起始于 2，但起始值指定为 1，那么此过程将假定该字段的每个循环中的第一个周期都有一个缺失值。

分析的时间间隔

用于分析的时间间隔可以与观测的时间间隔不同。 例如，如果观测时间间隔为“天”，可以为分析时间间隔选择“月”。 然后，系统在构建模型之前将数据从每日数据汇总为每月数据。 您还可以选择将时间间隔较长的数据拆分到较短的时间间隔内。 例如，如果观测值是每季度数据，那么您可以将每季度数据分布为每月数据。

执行分析的时间间隔的可用选项取决于观测的定义方式以及这些观测的时间间隔。 特别是，如果观测值由循环周期定义，那么仅支持汇总。 在这种情况下，分析时间间隔必须大于或等于观测值的时间间隔。

分析时间间隔在“数据指定项”选项卡上的**时间间隔**设置中指定。 汇总或分布数据的方法在“数据指定项”选项卡上的**汇总和分布**设置中指定。

聚集和分布

聚集函数
如果用于分析的时间间隔比观测值的时间间隔长，那么将对输入数据进行聚集。 例如，当观测值的时间间隔为“天”并且分析时间间隔为“月”时，将执行聚集。 可用的聚集函数如下所示：mean、sum、mode、min 或 max。

分布函数
如果用于分析的时间间隔比观测值的时间间隔短，那么将对输入数据进行分布。 例如，当观测值的时间间隔为“季度”并且分析时间间隔为“月”时，将执行分布。 可用的分布函数如下所示：mean 或 sum。

分组函数
当观测值由日期/时间定义，并且同一个时间间隔内存在多个观测值时，将进行分组。 例如，如果观测值的时间间隔为“月”，那么同一个月的多个日期将分组到一起，并与它们所在的月份相关联。 可用函数有：mean、sum、mode、min 或 max。当观测值由日期/时间定义，并且观测值的时间间隔指定为“不定期”时，将始终执行分组。

注: 尽管分组是一种聚集形式，但在对缺失值进行任何处理之前执行，而正式的聚集是在对所有缺失值进行处理之后执行。 如果观测值的时间间隔指定为“不规则”，那么将仅使用分组函数来执行聚集。

将跨天观测值聚集到前一天
指定是否将时间跨天边界的观测值汇总到前一天的值。 例如，如果每一天的时间范围是从 20:00 开始的 8 小时，那么对于每小时观测值，此设置指定是否将介于 00:00 与 04:00 之间的观测值包括在前一天的聚集结果中。 仅当观测值的时间间隔为“每天的小时数”、“每天的分钟数”或“每天的秒数”，并且分析时间间隔为“天”时，此设置才适用。

所指定字段的定制设置
您可以对每个字段指定聚集函数、分布函数和分组函数。 这些设置将覆盖聚集函数、分布函数和分组函数的缺省设置。

缺失值

输入数据中的缺失值将替换为插补值。 可用的替换方法如下所示：

线性插值
使用线性插值法替换缺失值。 缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果序列中的第一个或最后一个观测值具有缺失值，那么将使用序列开头或结尾的两个最近的非缺失值。

序列平均值
将缺失值替换为整个序列的平均值。

临近点的平均值
使用有效周围值的平均值替换缺失值。 邻近点的跨度为缺失值前后用于计算平均值的有效值数目。

邻近点的中值
使用有效周围值的中值替换缺失值。 邻近点的跨度为缺失值前后用于计算中位数的有效值数目。

线性趋势
此选项使用序列中的所有非缺失观测值来拟合简单线性回归模型，该模型随后用于插补缺失值。

其他设置：

缺失值的最大百分比 (%)
指定针对任何序列允许的缺失值的最大百分比。 将从分析中排除缺失值数量超过指定最大值的序列。

常规数据选项

每个维度字段的不同值的最大数目
此设置适用于多维数据，它指定任何一个维度字段所允许的不同值的最大数量。 缺省情况下，此限制设置为 10000，但可以增大到任意大的数字。

常规构建选项

置信区间宽度 (%)
此设置控制预测及模型参数的置信区间。 您可以指定任何小于 100 的正数值。 缺省情况下，将使用 95% 置信区间。

每个目标的最大输入数目
此设置指定每个目标的模型中允许的最大输入数目。 您可以指定 1 到 20 范围内的整数。 每个目标的模型都始终包含自身的延迟值，因此将此值设置为 1 表示唯一的输入是目标自身。

模型容差
此设置控制用于确定每个目标的最佳输入集合的迭代式过程。 您可以指定任何大于零的值。 缺省值为 0.001。 模型容差是预测变量选择的停止条件。 它可以影响最终模型中包含的预测变量数。 但是，如果目标自身预测良好，那么最终模型中可能不包含其他预测变量。 可能需要一些试验和错误（例如，如果将此设置位置为较高的值，那么可尝试将其设置为较小的值以查看是否可包含其他预测变量）。

离群值阈值 (%)
如果根据模型计算而得的可能性指出观测值为超出此阈值的离群值，那么会将该观测值标记为离群值。 您可以指定 50 到 100 范围内的值。

每个输入的延迟项数
此设置指定每个目标的模型中每个输入的延迟项数。 缺省情况下，延迟项数根据用于分析的时间间隔自动确定。 例如，如果时间间隔为“月”（增量为 1 个月），那么延迟项数为 12。 您可以选择性地明确指定延迟项数。 指定的值必须是 1 - 20 范围内的整数。

使用现有模型继续估算
如果已生成时间因果模型，那么选择此选项将复用对该模型指定的条件设置，而不构建新模型。 这样，就可以基于先前模型设置但使用较新的数据来重新估算并生成新预测，从而节省时间。

要显示的序列

这些选项指定要显示其输出的序列（目标或输入）。 所指定序列的输出内容由**输出选项**设置确定。

显示与最佳拟合的模型相关联的目标
缺省情况下，将显示与 10 个最佳拟合的模型（由 R 平方值确定）相关联的目标的输出。 您可指定不同的最佳拟合模型固定数目，也可以指定最佳拟合模型的百分比。 您还可以从下列拟合度度量中进行选择：

R 平方
线性模型的拟合优度度量，有时称为决定系数。 这是目标变量的变动中，由模型解释的比例。 它的取值范围从 0 到 1。 较小的值表示模型不适合数据。

均方根百分比误差
这是对序列的模型预测值与观测值之间差异程度的度量。 它与使用的单位无关，因此可用于比较具有不同单位的序列。

均方根误差
均方误差的平方根。 这是对相依序列与其模型预测水平相差程度的度量，以相依序列的单位表示。

BIC
Bayesian 信息准则。这是一种基于 -2 扣减对数似然来选择和比较模型的度量。 值越小，表示模型拟合得越好。 BIC 也会“惩罚”过多参数模型（例如，具有大量输入的复杂模型），但比 AIC 更严格。

AIC
Akaike 信息准则。 这是一种基于 -2 扣减对数似然来选择和比较模型的度量。 值越小，表示模型拟合得越好。 AIC“惩罚”过多参数模型（例如，具有大量输入的复杂模型）。

指定各个序列

您可以指定需要其输出的各个序列。

- 对于基于列的数据，可以指定其中包含您需要的序列的字段。 所指定字段的顺序定义这些字段在输出中的显示顺序。
- 对于多维数据，可以通过针对包含序列的度量字段，在网格中添加相应条目来指定特定序列。 然后，指定用于定义序列的维度字段的值。

- 您可以直接在网格中输入每个维度字段的值，也可以从可用维度值列表中选择值。 要从可用维度值列表中进行选择，请单击所需维度的单元格中的省略符按钮。 此操作将打开“选择维度值”子对话框。
- 您可以通过在“选择维度值”子对话框中单击双筒望远镜图标并指定搜索项来搜索维度值列表。 空格被视为搜索项的组成部分。 搜索项中的星号 (*) 并不表示通配符。
- 网格中序列的顺序定义这些序列在输出中的显示顺序。

对于基于列的数据和多维数据，输出都仅限于 30 个序列。 此限制包括您指定的各个序列（输入或目标）以及与最佳拟合的模型相关联的目标。 逐个指定的序列优先于与最佳拟合的模型相关联的目标。

输出选项

这些选项指定输出内容。 **目标的输出**组中的选项为**要显示的序列**设置所指定的最佳拟合模型的相关目标生成输出。 **序列的输出**组中的选项用于为**要显示的序列**设置所指定的各个序列生成输出。

总体模型系统

显示模型系统中序列之间因果关系的图形表示法。 所显示目标的模型拟合度统计表和离群值表都将包括在输出项中。 如果在**序列的输出**组中选择了此选项，那么将为**要显示的序列**设置所指定的每个序列创建一个单独的输出项。

序列之间的因果关系具有相关的显著性水平，其中，较小的显著性水平表示关系更为显著。 您可以选择隐藏显著性水平大于指定值的关联。

模型拟合度统计和离群值

这些表列出选定显示的目标序列的模型拟合统计量和离群值。 这些表与“总体模型系统”显示内容中的表包含相同的信息。 这些表支持所有的标准透视表功能和编辑表功能。

模型效应和模型参数

这些表列示选定显示的目标序列的模型效应检验和模型参数。 模型效应检验包括模型中每个输入的 F 统计量和相关显著性值。

影响图

显示感兴趣的序列与其所影响或影响其的其他序列之间因果关系的图形表示。 影响所关注序列的序列称为原因。 选择**效应**将生成初始化为显示效应的影响图。 选择**原因**将生成初始化为显示原因的影响图。 选择**包括原因和效应**将生成两个单独的影响图，其中一个影响图初始化为显示原因，另一个影响图初始化为显示效应。 您可以在显示影响图的输出项中来回切换原因和效应。

您可以指定要显示的原因或效应级别数，其中，第一个级别只是感兴趣的系列。 每个附加级别显示所关注序列的更多间接原因或效应。 例如，效应显示中第三级别的序列将第二级别的序列作为直接输入。 第三个级别中的序列受所关注序列间接影响，这是因为，所关注序列是第二个级别中的序列的直接输入。

序列图

这些图列出选择显示的目标序列的观测值和预测值。 请求进行预测时，此图还显示这些预测的预测值和置信区间。

残值图

这些图列出选择显示的目标序列的模型残差。

主要输入

这些图列出所显示的每个目标的长期状况以及该目标的 3 个主要输入。 主要输入是显著性值最小的输入。 为了支持输入和目标采用不同的刻度，y 轴表示每个序列的 z 得分。

预测表

这些表列出选定显示的目标序列的预测值及这些预测的置信区间。

离群值根本原因分析

确定哪些序列最有可能是所关注序列中各个离群值的原因。 针对**要显示的序列**设置中指定的各个序列的列表中包含的每个目标序列，都会执行离群值根本原因分析。

输出

交互式离群值表格和图表
这个表和图表列示每个所关注序列的离群值及离群值根本原因。 对于每个离群值，表中都包含一行。 此图表是影响图。 在表中选择一行将在影响图中突出显示一条路径，该路径从所关注序列通往最有可能引起相关联离群值的序列。

离群值的透视表
此表列示每个所关注序列的离群值及离群值根本原因。 此表与交互式显示中的表包含相同的信息。 这个表支持所有的标准透视表功能和编辑表功能。

因果级别
您可指定在搜索根本原因时要包括的级别数。 此处使用的级别概念与针对影响图描述的概念相同。

所有模型之间的模型拟合度
此直方图根据所有模型及所选拟合统计来显示模型拟合度。 提供了下列拟合统计：

R 平方
线性模型的拟合优度度量，有时称为决定系数。 这是目标变量的变动中，由模型解释的比例。 它的取值范围从 0 到 1。 较小的值表示模型不适合数据。

均方根百分比误差
这是对序列的模型预测值与观测值之间差异程度的度量。 它与使用的单位无关，因此可用于比较具有不同单位的序列。

均方根误差
均方误差的平方根。 这是对相依序列与其模型预测水平相差程度的度量，以相依序列的单位表示。

BIC
Bayesian 信息准则。这是一种基于 -2 扣减对数似然来选择和比较模型的度量。 值越小，表示模型拟合得越好。 BIC 也会“惩罚”过多参数模型（例如，具有大量输入的复杂模型），但比 AIC 更严格。

AIC
Akaike 信息准则。 这是一种基于 -2 扣减对数似然来选择和比较模型的度量。 值越小，表示模型拟合得越好。 AIC“惩罚”过多参数模型（例如，具有大量输入的复杂模型）。

某段时间的离群值
这个条形图显示估计期的每个时间间隔内所有目标的离群值数目。

序列变换
这个表列出已对模型系统中的序列应用的所有转换。 可能的转换包括缺失值插补、汇总和分布。

估计期

缺省情况下，估计期从所有序列中的最早观测值的时间开始，并以最晚观测值的时间结束。

按开始时间和结束时间
您可以同时指定估计期的开始时间和结束时间，也可以仅指定开始时间或者仅指定结束时间。 如果省略了估计期的开始时间或结束时间，那么将使用缺省值。

- 如果观测值由日期/时间字段定义，请以用于该日期/时间字段的格式输入开始时间值和结束时间值。
- 对于由循环周期定义的观测值，请对每个循环周期字段指定值。 每个字段都将显示在单独的列中。

按最晚或最早时间间隔

将估计期定义为指定数目的时间间隔，这些时间间隔从数据中的最早时间间隔开始或者以最晚时间间隔结束，并可以使用可选的偏移量。 在此上下文中，时间间隔是指分析时间间隔。 例如，假定观测值按月获取，但分析时间间隔为季度。 指定**最晚**并对**时间间隔数目**指定值 24 表示最晚的 24 个季度。

您可以选择性地排除指定数目的时间间隔。 例如，指定最晚的 24 个时间间隔并对排除数目指定 1 表示估计期由最后一个时间间隔之前的 24 个时间间隔组成。

模型选项

模型名称
您可以对模型指定定制名称，也可以接受自动生成的名称，即 *TCM*。

预测

将记录扩展至将来选项用于设置时间间隔数目，以预测估计期结束之后的情况。 在这种情况下，时间间隔为“数据指定项”选项卡上指定的分析时间间隔。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。 此设置没有最大限制。

交互式输出

时间因果建模的输出包括许多交互式输出对象。 在“输出查看器”中激活（双击）对象，即可使用交互式功能。

总体模型系统

显示模型系统中各个序列之间的因果关系。 将特定目标连接到其输入的所有线条都具有相同的颜色。 线条的粗细表示因果联系的重要性，粗线代表更重要的联系。 不是目标的输入用黑色方块表示。

- 您可以显示主要模型、指定序列、所有序列或没有输入的模型的关系。 主要模型是满足您通过**要显示的序列**设置对最佳拟合模型指定的条件的模型。
- 通过在图表中选择序列名称，单击鼠标右键，然后从上下文菜单中选择**创建影响图**，您可以为一个或多个序列生成影响图。
- 您可以选择隐藏显著性水平大于指定值的因果关系。 显著性水平越低表示因果关系越重要。
- 通过在图表中选择序列名称，单击鼠标右键然后从上下文菜单中选择**突出显示序列的关系**，您可以显示特定序列的关系。

影响图

显示感兴趣的序列与其所影响或影响其的其他序列之间因果关系的图形表示。 影响所关注序列的序列称为原因。

- 您可以通过指定所需的序列名称来更改所关注的序列。 在影响图中双击任何节点都会将所关注序列更改为该节点的相关联序列。
- 您可以在原因和结果之间切换显示，并且可以更改要显示的原因或结果的级别数。
- 单击任何节点将打开与此节点相关联的序列的详细时序图。

离群值根本原因分析

确定哪些序列最有可能是所关注序列中各个离群值的原因。

- 您可以通过在离群值表中选择任何离群值对应的行来显示该离群值的根本原因。 另外，还可以通过在序列图中单击离群值的图标来显示根本原因。
- 单击任何节点将打开与此节点相关联的序列的详细时序图。

总体模型质量

模型的直方图适合所有模型，适合特定的拟合统计量。 单击条形图中的条形可过滤点图，以使其仅显示与所选条形关联的模型。 通过指定序列名称，可以在点图中找到特定目标序列的模型。

离群值分布

这个条形图显示估计期的每个时间间隔内所有目标的离群值数目。 在条形图中单击某个条形可以对点图进行过滤，以便它仅显示与所选条形关联的离群值。

TCM 模型块

TCM 建模操作将创建多个具有前缀 $TCM 的新字段，如下表所示。

表 *27: TCM* 建模操作创建的新字段

字段名称	描述
$TCM-*colname*	每个目标序列模型预测的值。
$TCMLCI-*colname*	每个已预测序列的置信区间下限值。
$TSUCI-*colname*	每个已预测序列的置信区间上限值。
$TCMResidual-*colname*	每列生成的模型数据中的噪声残差值。

TCM 模型块设置

“设置”选项卡为 TCM 模型块提供了额外的选项。

预测

将记录扩展至将来选项用于设置时间间隔数目，以预测估计期结束之后的情况。 在这种情况下，时间间隔为 TCM 节点的“数据指定项”选项卡上指定的分析时间间隔。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。

使其可用于评分

为要评分的每个模型创建新字段。 使您可以指定为每个要进行评分的模型创建新字段。

- **噪声残值**。 如果选中了此选项，那么对于每个目标字段，将为模型残差创建新字段（带有缺省前缀 $TCM-），并同时创建这些值的总计。
- **置信度上限和下限**。 如果选中了此选项，那么对于每个目标字段，将分别为置信区间上限和下限创建新字段（带有缺省前缀 $TCM-），并同时创建这些值的总计。

包含用于评分的目标。 选择要包含在模型评分中的可用目标。

时间因果模型方案

“时间因果模型方案”过程使用活动数据集中的数据，对时间因果模型系统运行用户定义的方案。 方案是由称为根序列的时间序列以及指定时间范围内针对此序列的一组用户定义的值来定义的。 指定的值随后将用来生成受根序列影响的时间序列的预测。 此过程需要一个由“时间因果建模”过程创建的模型系统文件。 假设活动数据集就是用来创建模型系统文件的那些数据。

示例

通过使用“时间因果建模”过程，业务决策制定者发现了影响许多重要绩效指标的关键度量。 该度量可控制，因此决策制定者希望调查该度量的各组值对下一个季度的影响。 将模型系统文件装入到“时间因果模型方案”过程中，并对该关键度量指定值集，即可轻松完成此调查。

定义方案期

方案期是您指定用于运行方案的值所在的周期。 方案期可以在估计期结束之前或之后开始。 您可以选择性地指定预测范围超出方案期结束时间。 缺省情况下，将生成截至方案期结束为止的预测。 所有方案将使用相同的方案期和指定项来确定预测时间范围。

注: 预测从预测期开始后的第一个时间段开始。 例如，如果方案期从 2014 年 11 月 01 日开始，并且时间间隔为月，那么第一个预测针对 2014 年 12 月 01 日。

按开始时间、结束时间和预测时间范围来指定

- 如果观测值由日期/时间字段定义，请以用于该日期/时间字段的格式输入开始时间值、结束时间值和预测时间范围值。 日期/时间字段的值与相关联时间间隔的开始时间一致。 例如，如果分析时间间隔为“月”，那么值 10/10/2014 将调整为 10/01/2014，即当月的第一天。
- 对于循环周期定义的观测，为每个循环周期字段指定值。 每个字段都将显示在单独的列中。

指定与估算周期终点相关的时间间隔

以相对于估计期结束时间的时间间隔数方式定义开始时间和结束时间，其中，时间间隔是指分析时间间隔。 估计期的结束定义为时间间隔 0。 估计期结束前的时间间隔具有负值，估计期结束后的时间间隔具有正值。 另外，您还可以指定方案期结束后要预测的时间间隔数。 缺省值为 0。

例如，假定分析时间间隔为“月”，并且您指定了 1 作为开始时间间隔，指定了 3 作为结束时间间隔，并指定了 1 作为在此之后要预测的时间范围。 这样，方案期是估计期结束后的 3 个月。 因此，将为方案期的第 2 个月和第 3 个月生成预测，并为方案期结束后的 1 个月生成预测。

添加方案和方案组

“方案”选项卡指定要运行的方案。 要定义方案，您必须先通过单击**定义方案期**来定义方案期。 要创建方案和方案组（仅适用于多维数据），请单击相关联的**添加方案**或**添加方案组**按钮。 通过在相关联的网格中选择特定方案或方案组，可以对其进行编辑、复制或删除。

基于列的数据

网格中的**根字段**列指定一个时间序列字段，此字段的值将替换为方案值。 **方案值**列按最早到最晚顺序显示指定的方案值。 如果方案值由表达式定义，那么此列将显示该表达式。

多维数据

各个方案

“各个方案”网格中的每一行都指定一个时间序列，该时间序列的值将替换为指定的方案值。 此序列由**根度量**列中指定的字段与每个维度字段的指定值的组合定义。 **方案值**列的内容与基于列的数据相同。

方案组

方案组定义一组方案，这些方案基于单个根度量字段和多个维度值集合。 对于指定的度量字段，每个维度值集合（每个维度字段各有一个对应的值）都定义了一个时间序列。 对于这样的每一个时间序列，都将生成一个单独的方案，然后，这些时间序列的值替换为方案值。 方案组的方案值由一个表达式指定，该表达式将应用于该组中的每个时间序列。

序列计数列显示与方案组相关联的维度值集合的数目。 显示的值可能大于与该方案组相关联的时间序列的实际数目（每个集合各有一个对应的序列）。 指定的某些维度值组合未与该组的根度量所包含的序列相对应时，将发生这种情况。

作为方案组示例，假定存在度量字段 *advertising* 以及两个维度字段 *region* 和 *brand*。 您可以定义以 *advertising* 作为根度量并且包含 *region* 与 *brand* 的所有组合的方案组。 然后，您可以指定 `advertising*1.2` 作为表达式，以调查将 *advertising* 增大 20% 对每个与 *advertising* 字段相关联的时间序列的影响。 如果 *region* 有 4 个值，并且 *brand* 有 2 个值，那么共有 8 个这样的时间序列，因此这个组将定义 8 个方案。

方案定义

用于定义方案的设置取决于您的数据是基于列的数据还是多维数据。

根序列

指定方案的根序列。 每个方案都基于单个根序列。 对于基于列的数据，请选择用于定义根序列的字段。 对于多维数据，请通过在网格中针对包含序列的度量字段添加条目来指定根序列。 然后，指定用于定义根序列的维度字段的值。 在指定维度值时，下列操作适用：

- 您可以直接在网格中输入每个维度字段的值，也可以从可用维度值列表中选择值。 要从可用维度值列表中进行选择，请单击所需维度的单元格中的省略符按钮。 此操作将打开“选择维度值”子对话框。
- 您可以通过在“选择维度值”子对话框中单击双筒望远镜图标并指定搜索项来搜索维度值列表。 空格被视为搜索项的组成部分。 搜索项中的星号 (*) 并不表示通配符。

指定受影响的目标

如果您知道受根序列影响的特定目标，并且希望仅调查对这些目标的影响，请使用此选项。 缺省情况下，将自动确定根序列所影响的目标。 您可以使用“选项”选项卡上的设置来指定此方案所影响的序列的广度。

对于基于列的数据，请选择所需的目标。 对于多维数据，请通过在网格中针对包含目标序列的目标度量字段添加条目来指定目标序列。 缺省情况下，将包括所指定度量字段中包含的所有序列。 通过对一个或多个维度字段定制所包括的值，可以对包括的序列的集合进行定制。 要对包括的维度值进行定制，请单击所需维度的省略号按钮。 此操作将打开“选择维度值”对话框。

序列计数列（对于多维数据）显示当前对相关目标度量指定的维度值集合的数目。 显示的值可能大于受影响目标序列的实际数目（每个集合各有一个对应的序列）。 当指定的某些维度值组合未与相关目标度量所包含的序列相对应时，将发生这种情况。

方案标识
每个方案都必须具有唯一标识。 此标识将显示在与该方案相关联的输出中。 对于此标识的值，除了唯一性以外，没有任何限制。

指定根序列的方案值
使用此选项可以指定根序列在方案期中具有的显式值。 必须对网格中列出的每个时间间隔指定一个数字值。 通过单击**读取**、**预测**或**读取\预测**，可以获取方案期中根序列在每个时间间隔的的值（实际值或预测值）。

指定根序列的方案值的表达式
您可以定义表达式，以计算根序列在方案期中的值。 您可以直接输入表达式，也可以单击计算器按钮并从方案值表达式构建器中创建表达式。

- 表达式可以包含模型系统中的任何目标或输入。
- 如果方案期延伸到现有数据的时间范围以外，那么此表达式将应用于此表达式中各个字段的预测值。
- 对于多维数据，此表达式中的每个字段都指定一个时间序列，该时间序列由该字段以及您对根度量指定的维度值定义。 这些时间序列用于对此表达式进行求值。

例如，假设根字段为 *advertising* 并且表达式为 `advertising*1.2`。 此方案中使用的各个 *advertising* 值表示在现有值的基础上增大 20%。

注: 要创建方案，请在“方案”选项卡上单击**添加方案**。

选择维度值

对于多维数据，您可以定制用于定义方案或方案组所影响的目标的维度值。 另外，您还可以定制用于为方案组定义根序列集合的维度值。

所有值
指定包含当前维度字段的所有值。 这是缺省选项。

选择值
使用此选项可以指定当前维度字段的值集。 您可以对要从中进行选择的值集进行过滤。 满足过滤条件的值将显示在**匹配**选项卡中，不满足过滤条件的值将显示在**未选择的值**列表的**不匹配**选项卡中。 **全部**选项卡列示所有未选择的值，而无论指定了什么过滤条件。

- 指定过滤器时，可以使用星号 (*) 来表示通配符。
- 要清除当前过滤器，请在“过滤显示的值”对话框中将搜索项指定为空值。

要定制受影响目标的维度值，请完成下列步骤:

1. 在“方案定义”或“方案组定义”对话框中，选择要为其定制维度值的目标度量。
2. 单击要定制的维度的相应列中的省略符按钮。

要对方案组的根序列的维度值进行定制，请完成下列步骤:

1. 在“方案组定义”对话框中，单击要定制的维度的省略号按钮（位于根序列网格中）。

方案组定义

根序列
指定方案组的根序列集合。 对于此集合中的每个时间序列，都将生成一个单独的方案。 请通过在网格中针对包含所需序列的度量字段添加条目来指定根序列。 然后，指定用于定义集合的维度字段的值。 缺省情况下，将包括所指定根度量字段中包含的所有序列。 通过对一个或多个维度字段定制所包括的值，可以对包括的序列的集合进行定制。 要对包括的维度值进行定制，请单击某个维度的省略号按钮。 此操作将打开“选择维度值”对话框。

序列计数列显示当前为相关根度量包括的维度值集合的数目。 显示的值可能大于该方案组的根序列的实际数目（每个集合各有一个对应的序列）。 当指定的某些维度值组合未与根度量所包含的序列相对应时，将发生这种情况。

指定受影响的目标序列
如果您知道受根序列集合影响的特定目标，并且希望仅调查对这些目标的影响，请使用此选项。 缺省情况下，将自动确定每个根序列所影响的目标。 您可以使用“选项”选项卡上的设置来指定每个方案所影响的序列的广度。

请通过在网格中针对包含序列的度量字段添加条目来指定目标序列。缺省情况下，将包括所指定度量字段中包含的所有序列。 通过对一个或多个维度字段定制所包括的值，可以对包括的序列的集合进行定制。 要对包括的维度值进行定制，请单击所需维度的省略号按钮。 此操作将打开“选择维度值”对话框。

序列计数列显示当前对相关目标度量指定的维度值集合的数目。 显示的值可能大于受影响目标序列的实际数目（每个集合各有一个对应的序列）。 当指定的某些维度值组合未与相关目标度量所包含的序列相对应时，将发生这种情况。

方案标识前缀
每个方案组都必须具有唯一的前缀。 此前缀用于构造标识，该标识将显示在此方案组中每个方案的相关输出中。 对于一个方案，其标识依次由前缀、下划线以及每个用于标识根序列的维度字段的值组成。 各个维度值之间以下划线分隔。 对于前缀的值，除了唯一性以外，没有任何限制。

根序列的方案值的表达式
方案组的方案值由表达式指定，该表达式用于计算该组中各个根序列的值。 您可以直接输入表达式，也可以单击计算器按钮并从方案值表达式构建器中创建表达式。

- 表达式可以包含模型系统中的任何目标或输入。
- 如果方案期延伸到现有数据的时间范围以外，那么此表达式将应用于此表达式中各个字段的预测值。
- 对于组中的每个根序列，此表达式中的字段指定了时间序列，这些时间序列由这些字段以及用于定义根序列的维度值定义。 这些时间序列用于对此表达式进行求值。 例如，如果根序列由 `region='north'` 和 `brand='X'` 定义，那么表达式中使用的时间序列由同样的维度值定义。

例如，假设根度量字段为 *advertising*，并假设存在两个度量字段 *region* 和 *brand*。 此外，假设方案组包含维度字段值的所有组合。 然后，您可以指定 `advertising*1.2` 作为表达式，以调查将 *advertising* 增大 20% 对每个与 *advertising* 字段相关联的时间序列的影响。

注: 方案组仅应用于多维数据，要创建方案组，请在“方案”选项卡上单击**添加方案组**。

选项

受影响的目标的最大级别
指定受影响目标的最大级别数。 后续每个级别（最多 5 个级别）都包含间接受根序列影响的目标。 具体而言，第一个级别包含将根序列用作直接输入的目标。 第二个级别中的目标将第一个级别中的目标用作直接输入，依此类推。 增大此设置的值将增加计算的复杂程度，并可能影响性能。

自动检测的最大目标数
指定针对每个根系列自动检测到的受影响目标的最大数量。增大此设置的值将增加计算的复杂程度，并可能影响性能。

影响图
显示每个方案的根序列与其影响的目标序列之间的因果关系的图形表示。 受影响目标的方案值和预测值的表都将包括在输出项中。 图形包含受影响目标的预测值的图。 单击影响图中的任何节点都将打开与该节点相关联的序列的详细时序图。 对于每个方案，将生成单独的影响图。

序列图
为每个方案中每个受影响目标的预测值生成序列图。

预测表和方案表
这些表列出每个方案的预测值和方案值。 这些表与影响图中的表包含的信息相同。 这些表支持所有的标准透视表功能和编辑表功能。

在图表中包含置信区间
指定是否将方案预测置信区间同时包括在图表输出中。

置信区间宽度 (%)
此设置用于控制方案预测置信区间。 您可以指定任何小于 100 的正数值。 缺省情况下，将使用 95% 置信区间。

“时间序列”节点

“时间序列”节点可以在本地或分布式环境中与数据配合使用；在分布式环境中，可以利用 IBM SPSS Analytic Server 的能力。 通过此节点，可以选择对时间序列的指数平滑法模型、单变量自回归整合移动平均值 (ARIMA) 及多变量 ARIMA（或转换函数）模型进行估计和构建，并根据时间序列数据产生预测。

指数平滑是一种使用先前的序列观察的加权值来预测未来值的预测方法。 因此，指数平滑不是以对数据的理论理解为基础的。 指数平滑每次预测一个点，在输入新数据时可调整其预测。 此技术有助于预测可展示趋势和/或季节性的序列。 您可从以不同方式处理趋势和季节性的各种指数平滑法模型中进行选择。

与指数平滑法模型相比，**ARIMA** 模型在对趋势和季节组件建模方面提供更成熟的方法，特别是，增加了可在模型中包括自变量（预测变量）的优势。 这包括明确指定自回归阶数和移动平均值阶数以及差分次数。 可以包含预测变量并为任意或所有预测变量定义变换函数以及指定对离群值的自动检测或精确设置。

注: 在实际应用中，如果要包括预测变量（这些变量可能有助于说明正在预测的序列的行为，例如邮寄的目录的数目或某公司网页的点击数），那么 ARIMA 模型最有用。 而指数平滑模型在说明时间序列的行为时，并不试图去了解其行为的原因。 例如，过去每 12 个月达到最大值的序列可能继续保持该行为，即使您不了解其原因也是如此。

还提供了**专家建模器**选项，此选项将尝试自动识别和估计对一个或多个目标变量拟合度最高的 ARIMA 模型或指数平滑法模型，从而无需通过试错来识别适当的模型。 如果您有任何疑问，请使用“专家建模器”选项。

如果指定了预测变量，那么专家建模器会选择将那些在统计意义上与相依序列具有显著关系的变量包括在 ARIMA 模型中。 适当时，使用差分和/或平方根或自然对数转换对模型变量进行转换。 缺省情况下，Expert Modeler 会考虑所有指数平滑模型和所有 ARIMA 模型并为每个目标字段选择其中最适合的模型。 不过，可以将 Expert Modeler 限制为仅选择最适合的指数平滑模型或仅选择最适合的 ARIMA 模型。 也可以指定对离群值进行自动检测。

“时间序列”节点 - 字段选项

在“字段”选项卡上，可以选择是要使用在上游节点中定义的字段角色设置，还是手动进行字段分配。

使用预定义角色：此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标、预测变量和其他角色，请选中此选项。

字段。 使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。 图标表示每个角色字段的有效测量级别。

要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择一个或多个字段作为预测目标。

候选输入。 选择一个或多个字段作为预测输入。 您只能选择测量级别为“连续”的字段。

事件和干预。 使用该区域来指定某些输入字段以作为事件或干预字段。 此指定会将字段标识为包含可受事件（可预测的重现情况，例如促销）或干预（一次性事件，例如停电或员工罢工）所影响的时间序列数据。 选择的字段必须是具有整数存储的标志。

“时间序列”节点 - 数据规范选项

通过“数据规范”选项卡，您可以设置用于将数据包含在模型中的所有选项。 只要同时指定**日期/时间字段**和**时间间隔**，便可以单击**运行**按钮来构建包含所有缺省选项的模型，但通常您会想要根据自己的用途定制构建。

该选项卡包含多个不同的窗口，您可以在其中设置特定于自己的模型的定制。

“时间序列”节点 - 观测值

使用此窗格中的设置可以指定用于定义观测值的字段。

由日期/时间字段指定的观测值

您可以指定观测值由日期、时间或时间戳记字段定义。 除了用于定义观测值的字段以外，请选择用于描述观测值的适当时间间隔。 根据指定的时间间隔不同，您还可以指定其他设置，例如观测值之间的间隔（增量）或者每周的天数。 对于时间间隔，注意事项如下所示：

- 如果各个观测值之间的时间间距不定期（例如处理销售订单的时间），请使用**不定期**值。 如果选择了**不定期**，那么必须从“数据指定项”选项卡上的**时间间隔**设置中指定用于分析的时间间隔。

- 如果观测值表示日期和时间，并且时间间隔为小时、分钟或秒，请使用**每天的小时数**、**每天的分钟数**或**每天的秒数**。 如果观测值表示时间（持续时间）并且未引用日期，而时间间隔为小时、分钟或秒，请使用**小时数（非周期性）**、**分钟数（非周期性）**或**秒数（非周期性）**。
- 根据选择的时间间隔，此过程可以检测缺失的观测值。 由于此过程假定所有观测值之间的时间间距相等，并假定未缺失观测值，因此有必要检测缺失的观测值。 例如，如果时间间隔为“天”，并且日期 2015-10-27 后跟 2015-10-29，那么表示缺失 2015-10-28 的观测值。 对于任何缺失观测值，将插补值；使用“数据规范”选项卡的**缺失值处理**区域可以指定用于处理缺失值的设置。
- 指定的时间间隔使此过程能够检测到同一时间间隔内的多个需要汇总到一起的观测值并使各个观测值在时间间隔边界（例如每个月的第一天）处对齐，以确保各个观测值之间的间距相等。 例如，如果时间间隔为“月”，那么同一个月内的多个日期将汇总到一起。 此类汇总称为分组。 缺省情况下，进行分组时，将计算观测值的总和。 通过“数据指定项”选项卡上的**汇总和分布**设置，您可以指定另一种分组方法，例如计算各个观测值的平均值。
- 对于某些时间间隔，附加设置可以定义正常等间距时间间隔中的中断。 例如，如果时间间隔为“天”，但只有工作日有效，那么您可以指定每周有 5 天，并且每周从星期一开始。

定义为周期或循环周期的观测值

观测值可以由一个或多个表示周期或周期反复循环（直至达到任意数目的循环级别为止）的整数字段定义。 借助此结构，您可以描述任何标准时间间隔都无法支持的观测值序列。 例如，可以使用表示年份的循环字段和表示月份的周期字段（一个循环的长度为 10）来描述仅包含 10 个月的财年。

用于指定循环周期的字段定义了周期性级别的层次结构，在此层次结构中，最低级别由**周期**字段定义。 次高级别由级别为 1 的循环字段指定，接着由级别为 2 的循环字段指定，依此类推。 除最高级别以外，每个级别的字段值对于次高级别都必须具有周期性。 最高级别的值不得具有周期性。 例如，对于由 10 个月组成的财年，月在年中具有周期性，而年不具有周期性。

- 在特定级别，循环长度是次低级别的周期长度。 在财年示例中，只有一个循环级别，并且循环长度为 10，这是因为次低级别表示月，而指定的财年包含 10 个月。
- 指定不从 1 开始的任何周期字段的起始值。 此设置是检测缺失值所必需的。 例如，如果周期性字段起始于 2，但起始值指定为 1，那么此过程将假定该字段的每个循环中的第一个周期都有一个缺失值。

“时间序列”节点 - 分析时间间隔

用于分析的时间间隔可以与观测值的时间间隔不同。 例如，观测值的时间间隔为“天”时，您可以选择“月”用作进行分析的时间间隔。 系统先将每日数据汇总为每月数据，然后再构建模型。 您还可以选择将时间间隔较长的数据分布到较短的时间间隔内。 例如，如果观测值是按季度的，那么您可以将数据从季度分发到月度数据。

使用此窗格中的设置指定用于分析的时间间隔。 汇总或分布数据的方法是在“数据指定项”选项卡上的**汇总和分布**设置中指定的。

执行分析所采用的时间间隔的可用选项取决于定义观测值的方式以及这些观测值的时间间隔。 特别是，如果观测值由循环周期定义，那么仅支持汇总。 在此情况下，分析的时间间隔必须大于或等于观测值的时间间隔。

“时间序列”节点 - 汇总和分布选项

使用此窗格中的设置可以指定用于对观测值的时间间隔的相关输入数据进行汇总或分布的设置。

汇总函数
如果用于分析的时间间隔比观测值的时间间隔长，那么将对输入数据进行汇总。 例如，当观测值的时间间隔为“天”并且分析时间间隔为“月”时，将执行汇总。 可用的汇总函数如下所示：mean、sum、mode、min 或 max。

分布函数
如果用于分析的时间间隔比观测值的时间间隔短，那么将对输入数据进行分布。 例如，当观测值的时间间隔为“季度”并且分析时间间隔为“月”时，将执行分布。 可用的分布函数如下所示：mean 或 sum。

分组函数
当观测值由日期/时间定义，并且同一个时间间隔内存在多个观测值时，将进行分组。 例如，如果观测值的时间间隔为“月”，那么同一个月的多个日期将分组到一起，并与它们所在的月份相关联。 以下是可用的分组函数：mean、sum、mode、min 或 max。当观测值由日期/时间定义，并且观测值的时间间隔指定为“不定期”时，将始终执行分组。

注: 尽管分组是一种汇总形式，但在对缺失值进行任何处理之前执行，而正式的汇总是在对所有缺失值进行处理之后执行。 如果观测值的时间间隔指定为“不定期”，那么将仅使用分组函数来执行汇总。

将跨天观测值汇总到前一天
指定是否将时间跨天边界的观测值汇总到前一天的值。 例如，对于在 20:00 开始的 8 小时一天的每小时观测值，此设置指定是否将介于 00:00 与 04:00 之间的观测值包含在前一天的汇总结果中。 仅当观测值的时间间隔为“每天的小时数”、“每天的分钟数”或“每天的秒数”，并且分析时间间隔为“天”时，此设置才适用。

所指定字段的定制设置
您可以对每个字段指定汇总函数、分布函数和分组函数。 这些设置将覆盖汇总函数、分布函数和分组函数的缺省设置。

“时间序列”节点 - 缺失值选项

使用此窗格中的设置可以指定输入数据中要替换为插补值的缺失值数。 提供了下列替换方法：

线性插值
使用线性插值替换缺失值。 缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果序列中的第一个或最后一个观测值具有缺失值，那么将使用序列开头或末尾的两个最近邻非缺失值。

使用线性插值替换缺失值。

- 对于非季节数据，缺失值之前的最后一个有效值以及之后的第一个有效值用于插值。 如果缺失值位于时间序列的开始或结束位置，那么基于两个最近的有效值使用线性推断方法。
- 对于季节数据，使用缺失值之前相同时间段的最后一个有效值和缺失值之后相同时间段的第一个有效值，线性内插缺失值。 如果无法针对缺失值找到相同时间段的两个值之一，那么数据将被视为非季节数据，并且使用非季节数据的线性插值来插补缺失值。

序列平均值
将缺失值替换为整个序列的平均值。

邻近点的平均值
使用有效周围值的均值替换缺失值。 邻近点的跨度为缺失值前后用于计算平均值的有效值数目。

邻近点的中位数
使用有效周围值的中位数替换缺失值。 邻近点的跨度为缺失值前后用于计算中位数的有效值数目。

线性趋势
此选项使用序列中的所有非缺失观测值来拟合简单线性回归模型，该模型随后用于插补缺失值。

其他设置：

最低数据质量评分 (%)
针对时间变量以及对应于每个时间序列的输入数据计算数据质量度量。 如果数据质量评分低于此阈值，那么将丢弃对应的时间序列。

“时间序列”节点 - 估计期

在“估计期”窗格中，您可以指定要在模型估计中使用的记录的范围。 缺省情况下，估计期从所有序列中的最早观测值的时间开始，并以最晚观测值的时间结束。

按开始时间和结束时间
您可以同时指定估计期的开始时间和结束时间，也可以仅指定开始时间或者仅指定结束时间。 如果省略了估计期的开始时间或结束时间，那么将使用缺省值。

- 如果观测值由日期/时间字段定义，请以用于该日期/时间字段的格式输入开始时间值和结束时间值。
- 对于由循环周期定义的观测值，请对每个循环周期字段指定值。 每个字段都将显示在单独的列中。

按最晚或最早时间间隔

将估计期定义为指定数目的时间间隔，这些时间间隔从数据中的最早时间间隔开始或以最晚时间间隔结束，并具有可选偏移量。 在此上下文中，时间间隔是指分析时间间隔。 例如，假定观测值按月获取，但分析时间间隔为季度。 指定**最晚**并对**时间间隔数目**指定值 24 表示最晚的 24 个季度。

您可以选择性地排除指定数目的时间间隔。 例如，指定最晚的 24 个时间间隔并对排除数目指定 1 表示估计期由最后一个时间间隔之前的 24 个时间间隔组成。

“时间序列”节点 - 构建选项

通过“数据规范”选项卡，您可以设置用于构建模型的所有选项。 当然，您只需单击**运行**按钮，即可采用所有缺省选项来构建模型；不过，通常您需要根据具体用途定制构建选项。

此选项卡包含两种不同的窗格，您可以在这些窗格中设置特定于模型的定制内容。

“时间序列”节点 - 常规构建选项

此窗格中的可用选项取决于您从**方法**列表中选择以下三项设置中的哪一项：

- **专家建模器**。 选择此选项以使用“专家建模器”，此组件将自动为每个相依序列查找拟合度最高的模型。
- **指数平滑法**。 使用此选项可指定定制的指数平滑法模型。
- **ARIMA**。 使用此选项可指定定制的 ARIMA 模型。

专家建模器

在**模型类型**下，选择您要构建的模型的类型：

- **所有模型**。 专家建模器同时考虑 ARIMA 模型和指数平滑法模型。
- **仅指数平滑法模型**。 “专家建模器”仅考虑指数平滑法模型。
- **仅 ARIMA 模型**。 “专家建模器”仅考虑 ARIMA 模型。

专家建模器考虑季节性模型。 只有在为活动数据集定义了周期性时才启用此选项。 选中此选项时，Expert Modeler 将同时考虑季节模型和非季节模型。 如果未选择此选项，那么专家建模器将仅考虑非季节性模型。

专家建模器考虑复杂的指数平滑法模型。 选择了此选项时，“专家建模器”将搜索所有 13 个指数平滑法模型（其中 7 个存在于原始时间序列节点中，而剩下 6 个是 V18.1 中新增的节点）。 如果 未选择此选项，那么“专家建模器”将搜索原始的 7 个指数平滑法模型。

在**离群值**下，从以下选项中进行选择

自动检测离群值。 缺省情况下，不自动检测离群值。 选中此选项以执行离群值自动检测，然后选择所需的离群值类型。

输入字段必须具有标志、名义或有序测量级别，并且必须是数字（例如，对于标志字段，必须为 1/0，而非 True/False），才能包含在此列表中。

对于在**字段**选项卡上标识为事件字段或干预字段的输入，Expert Modeler 仅考虑简单回归而不是任意变换函数。

指数平滑法

模型类型。 指数平滑法模型分类为季节性模型或非季节性模型。[1] 仅当使用“数据规范”选项卡上的“时间间隔”窗格定义的周期性为季节性时，季节性模型才可用。 季节性周期如下：循环周期、年、季度、月、每周的天数、每天的小时数、每天的分钟数以及每天的秒数。 提供了以下模型类型：

- **简式**。 此模型适合于没有趋势或季节性的序列。 其唯一的相关平滑参数是水平。 简单的指数平滑法非常类似于自回归阶数为 0、差分阶数为 1、移动平均值阶数为 1 且没有常量的 ARIMA 模型。
- **Holt 线性趋势**。 此模型适合于其中有线性趋势但没有季节性的序列。 其相关的平滑参数是水平和趋势，并且在此模型中，这些参数的值不会彼此限制。 Holt's 模型比 Brown's 模型更加常用，但在计算大型序列

[1] Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

的估计值时会花费更多的时间。 霍特指数平滑非常类似于自回归阶数为 0、差分阶数为 2 而且移动平均值阶数为 2 的 ARIMA 模型。

- **阻尼趋势**。 此模型适合于具有逐渐消失的线性趋势但没有季节性的序列。 其相关的平滑参数是水平、趋势和阻尼趋势。 阻尼指数平滑模型非常类似于自回归阶数为一、差分阶数为一且移动平均值阶数为二的 ARIMA 模型。
- **乘法趋势**。 该模型适合于具有一种随序列量级而变的趋势且没有季节性的序列。 其相关的平滑参数是水平和趋势。 乘性趋势指数平滑与任何 ARIMA 模型都不相似。
- **Brown 线性趋势**。 此模型适合于其中有线性趋势但没有季节性的序列。 其相关的平滑参数是水平和趋势，但在此模型中，这些参数的值假设相等。 因此，Brown 模型是 Holt 模型的特例。 布朗指数平滑法非常类似于自回归阶数为 0、差分阶数为 2 且移动平均值阶数为 2 的 ARIMA 模型，其第二阶移动平均值的系数等于第一阶系数一半的平方。
- **简单季节性**。 此模型适合于没有趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平和季节。 季节指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均值阶数为 1、*p* 和 *p*+1 的 ARIMA 模型，其中 *p* 是一个季节区间中的周期数。 对于以月为时间单位的数据， *p* = 12。
- **Winters 加法**。 此模型适合于具有线性趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势和季节。 Winters 加法指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均值阶数为 *p*+1 的 ARIMA 模型，其中 *p* 是一个季节区间中的周期数。 对于以月为时间单位的数据， *p* = 12。
- **具有加性季节性的阻尼趋势**。 此模型适合于具有逐渐消退的线性趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势、阻尼趋势和季节。 阻尼趋势和加性季节性指数平滑与任何 ARIMA 模型都不相似。
- **具有加性季节性的乘法趋势**。 该模型适合于具有随序列量级而变的趋势且季节效应不随时间变化的序列。 其相关的平滑参数是水平、趋势和季节。 乘性趋势和加性季节性指数平滑与任何 ARIMA 模型都不相似。
- **乘法季节性**。 此模型适合于不具有趋势且季节效应随序列量级而变的序列。 其相关的平滑参数是水平和季节。 乘性季节性指数平滑与任何 ARIMA 模型都不相似。
- **Winters 乘法**。 此模型适合于具有线性趋势且季节效应随序列的大小而变化的序列。 其相关的平滑参数是水平、趋势和季节。 Winters 的可乘指数平滑法与任何 ARIMA 模型都不相似。
- **具有乘法季节性的阻尼趋势**。 此模型适合于具有逐渐消退的线性趋势且季节效应随序列的大小而变化的序列。 其相关的平滑参数是水平、趋势、阻尼趋势和季节。 阻尼趋势和乘性季节性指数平滑与任何 ARIMA 模型都不相似。
- **具有乘法季节性的乘法趋势**。 此模型适合于具有随序列的量级发生变化的趋势和季节效应的序列。 其相关的平滑参数是水平、趋势和季节。 乘性趋势和乘性季节性指数平滑与任何 ARIMA 模型都不相似。

目标转换。 您可以指定在对每个因变量建模前对其执行的变换。

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

ARIMA

指定定制 ARIMA 模型的结构。

ARIMA 阶数。 在网格的相应单元格中，输入模型的各个 ARIMA 成分的值。 所有的值都必须是非负整数。 对于自回归和移动平均值组件来说，该值表示最大阶数。 所有较低的正阶数都将包括在模型中。 例如，如果指定 2，那么模型包含顺序 2 和 1 。 只有在为活动数据集定义了周期性的情况下，才会启用季节列中的单元格。

- **自回归的 (p)**。 模型中的自回归阶数。 自回归阶数指定序列中哪些以前的值用于预测当前值。 例如，自回归阶数 2 指定序列中过去两个时间段的值用于预测当前值。
- **差分 (d)**。 指定在估计模型之前应用于序列的差分的阶。 当趋势出现时（具有趋势的序列通常是不稳定的，而 ARIMA 建模时假定是稳定的），差分是必需的并可用于去除这些趋势的影响。 差分阶数对应于序列趋势的程度；第一阶差分表示线性趋势，第二阶差分表示二次趋势，依此类推。

- **移动平均值 (q)**。 模型中移动平均值阶数的值。 移动平均值阶数指定如何使用与序列以前值均值之间的偏差来预测当前值。 例如，移动平均值阶数 1 和 2 指定在预测序列的当前值时，可考虑与序列（来自过去两个时限中的每一个）均值之间的偏差。

季节性。 季节性自回归成分、移动平均值成分和差分成分与其非季节性对应成分起着相同的作用。 但是，对于季节阶数，当前的序列值会受到由一个或多个季节周期分隔的序列值的影响。 例如，对于月度数据（季节性周期为 12），季节性阶 1 表示当前序列值将受到当前周期之前 12 个周期的序列值的影响。 因此，对于以月为时间单位数据，将季节阶数指定为 1 相当于将非季节阶数指定为 12。

自动检测离群值。 选中此选项可以对离群值执行自动检测，并选择可用的一个或多个离群值类型。

要检测的离群值的类型。 选择要检测的离群值类型。 支持的类型有：

- 加性（缺省值）
- 水平变动（缺省值）
- 创新的
- 瞬时的
- 季节加性
- 局部趋势
- 可加的修补

转换函数顺序和变换。 要指定变换并为 ARIMA 模型中的任何或所有输入字段定义转换函数，请单击**设置**；这将显示另一个对话框，您可以在其中输入转换和变换详细信息。

在模型中包含常量。 除非您确定整体平均序列值为 0，否则包含常量是标准。 如果应用差分，那么建议排除常量。

其他详细信息

- 有关离群值类型的更多信息，请参阅第 215 页的『离群值』。
- 有关传输和变换函数的更多信息，请参阅第 239 页的『转换函数和变换函数』。

转换函数和变换函数

使用“转换函数顺序和变换”对话框可以指定变换以及为 ARIMA 模型中的任何或所有输入字段定义转换函数。

目标变换。 在此窗格中，您可以指定对每个目标变量进行建模之前要对其执行的变换。

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

候选输入转换函数和变换。 通过使用转换函数，您可指定以何种方式使用输入字段的过去值来预测目标序列的未来值。 左侧窗格的列表中显示了所有的输入字段。 此窗格中的其余信息特定于您选择的输入字段。

转换函数的阶数。 将转换函数的各个组件的值输入到**结构** 网格的相应单元格中。 所有的值都必须是非负整数。 对于分子和分母组件来说，该值表示最大阶数。 所有较低的正阶数都将包括在模型中。 此外，对于分子组件通常会包括阶数 0。 例如，如果为分子指定 2，那么模型包含阶数 2、1 和 0。 如果为分母指定 3，那么模型包含阶数 3、2 和 1。 只有在为活动数据集定义了周期性的情况下，才会启用季节列中的单元格。

分子。 转换函数的分子阶数指定选定的独立（预测变量）序列中哪些先前的值用于预测相依序列的当前值。 例如，分子阶数 1 指定使用过去某个时间段的独立序列的值（以及独立序列的当前值）来预测每个相依序列的当前值。

分母。 转换函数的分母阶数指定如何使用与选定独立（预测变量）序列的先前值均值之间的偏差来预测相依序列的当前值。 例如，分母阶数 1 指定在预测每个相依序列的当前值时，需要考虑与过去一个时间周期的独立序列的均值偏差。

差分。 指定在估计模型之前应用于所选独立（预测变量）序列的差分的阶数。 当趋势出现时，差分是必需的并可用于去除这些趋势的影响。

季节性。 季节性分子、分母和差分成分与其非季节性对应成分起着相同的作用。 但是，对于季节阶数，当前的序列值会受到由一个或多个季节周期分隔的序列值的影响。 例如，对于以月为时间单位的数据（季节周期为 12），季节阶数 1 表示当前序列值会受到当前序列之前的 12 个周期内的序列值的影响。 因此，对于以月为时间单位数据，将季节阶数指定为 1 相当于将非季节阶数指定为 12。

延迟。 设置延迟会将输入字段的影响延迟，延迟的时间为指定的时间间隔数。 例如，如果延迟设置为 5，那么输入字段在时间 *t* 不会产生影响，直到此后五个时限后 (*t* + 5) 才会对预测产生影响。

变换。 为一组自变量指定的转换函数还包括要对这些变量执行的可选变换。

- **无**。 未执行变换。
- **平方根**。 将执行平方根变换。
- **自然对数**。 执行自然对数变换。

“时间序列”节点 - 构建输出选项

ACF 和 PACF 输出中的最大延迟数。 自相关 (ACF) 和偏自相关 (PACF) 用于测量当前序列值和过去序列值之间的相关性，并指示预测将来值时最有用的过去序列值。 您可以设置自相关和偏自相关表及图中显示的最大延迟数。

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常，您希望将建模工作的主要精力放在最重要的预测变量上，并考虑删除或忽略那些最不重要的预测变量。 对于某些模型，计算预测变量重要性（特别在处理大型数据集时）可能需要花费较长时间，因此缺省情况下，预测变量重要性对某些模型处于关闭状态。

“时间序列”节点 - 模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

置信限制宽度 (%)。 为模型预测和残差自相关计算置信区间。 您可以指定任何小于 100 的正数值。 缺省情况下，将使用 95% 置信区间。

继续使用现有模型进行估算。 如果已生成一个时间序列模型，那么选择此选项可以重新使用为该模型指定的标准设置，并在模型选用板中生成一个新的模型节点，而不必从头构建一个新模型。 这样，您可以基于先前的模型设置但使用较新的数据来重新估算并生成新预测，从而节省时间。 例如，如果特定时间序列的原始模型是 Holt's 线性趋势，那么会使用相同的模型类型来重新估计和预测该数据。 系统不会重新尝试为新数据查找最佳模型类型。

仅构建评分模型。 要减少模型中存储的数据量，请选中此框。 使用此选项可以在使用许多时间序列（数万个）构建模型时提高性能。 您仍可以按照常规方法对数据评分。

将记录扩展到未来。 启用以下**要在预测中使用的未来值**部分，在该部分中可以设置估计期结束后要预测的时间间隔数量。 在这种情况下，时间间隔为您在“数据指定项”选项卡上指定的分析时间间隔。 此设置没有最大限制。 通过使用以下选项，您可以自动计算输入的未来值，或者手动为一个或多个预测变量指定预测值。

要在预测中使用的未来值

- **计算输入的未来值** 如果选择此选项，那么会自动计算预测变量、噪声预测、差异估算和未来时间值的预测值。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。
- **选择要将其值添加到数据中的字段**。 对于要预测的每条记录（不包括保留值），如果您使用的是预测变量字段（角色设置为 `Input`），那么可以为每个预测变量指定预测周期的估计值。 您可以手动指定值，也可以从列表中选择值。
  - **字段**。 单击“字段选择器”按钮并选择可用作预测变量的任何字段。 请注意，在此选择的字段可能用于建模，也可能不用于建模；要将某个字段实际用作预测变量，必须在某个下游建模节点中选择该字段。 此对话框为您简单提供了指定未来值的便捷环境，这样可以在多个下游建模节点之间共享这些值，而无需在每个节点中单独进行指定。 另请注意，可用字段的列表可能会受到“构建选项”选项卡中选项的约束。

请注意，如果为流中不再可用（因为已将其删除或由于“构建选项”选项卡中的选项更新）的字段指定了未来值，那么此字段将显示为红色。

– **值**。 对于每个字段，可以在函数列表中进行选择，也可以单击**指定**手动输入值或从预定义值列表中选择值。 如果预测变量字段与您所控制的项目或其他预先可知的项目相关，则应手动输入值。 例如，如果要根据房间预订数目预测饭店下个月的收入，可以指定在该期间实际具有的预订数目。 相反，如果预测变量字段与您无法控制的某些因素（如股票价格）相关，那么可以使用函数，如“最近值”或“最近点的均数”。

可用的函数取决于字段的测量级别。

表 *28:* 可用于测量级别的函数

测量级别	**函数**
连续或名义字段。	空 最近点的均数 最近的值 指定
标志字段	空 最近的值 True False 指定

最近点的均数根据最后三个数据点的均数计算未来值。

最近值将未来值设置为最近数据点的值。

真/假将标志字段的未来值设置为指定的真值或假值。

指定打开一个对话框，用于手动指定未来值或从预定义列表中选择未来值。

使其可用于评分

您可以在此设置模型块的对话框中显示的评分选项的缺省值。

- **计算置信度的上限和下限**。 如果选择了此选项，那么对于每个目标字段，将为置信区间上限和下限创建新字段（带有缺省前缀 $TSLCI- 和 $TSUCI-）。
- **计算噪声残差**。 如果选中了此选项，那么对于每个目标字段，此选项将为模型残差创建新字段（带有缺省前缀 $TSResidual-），并同时创建这些值的总计。

模型设置

要在输出中显示的最大模型数。 指定您要包含在输出中的最大模型数。 请注意，如果构建的模型数超过了此阈值，那么模型不会显示在输出中，但它们仍可用于评分。 缺省值为 10。 显示大量模型可能会导致性能不佳或不稳定。

时间序列模型块

“时间序列”模型块输出

创建时间序列模型后，输出查看器中会提供以下信息。 请注意，时间序列模型的“输出”查看器中可以显示的模型数量限制为 10 个。

时间信息摘要

此摘要显示以下信息：

- 时间字段
- 增量
- 起始点和结束点
- 唯一点的数目

此摘要适用于所有目标。

模型信息表

（对于每个目标重复）模型信息表提供关于模型的关键信息。 此表始终包含以下高级模型设置：

- 在“类型”节点或“时间序列”节点字段选项卡中选择的目标字段的名称。
- 模型构建方法 - 例如，指数平滑或 ARIMA。
- 输入模型中的预测变量数。
- 用于拟合模型类型的记录数。 不同类型的模型的示例可能包括：RMSE、MAE、AIC、BIC 和 R 方。

另外，如果数据满足所需的条件，那么还可能显示 Ljunq-Box Q 统计信息。 在下列情况下，此统计信息**不**可用：

- 非缺失数据点的数目小于或等于所需的总计项的数目（固定值 18）。
- 参数数目大于或等于所需的总计项的数目。
- 所计算的总计项的数目小于可接受的最小 k 值（固定值 7）。
- 对于每个目标，表重复显示。

预测变量重要性

（对于每个目标重复）预测变量重要性图形以条形图的形式显示模型中前 10 个输入（预测变量）的重要性。

如果图表中存在超过 10 个字段，那么可以使用图表下的滑块来调整图表中包含的预测变量的选择。 滑块上的指示符标记为固定宽度，滑块上每个标记表示 10 个字段。 您可以沿滑块移动指示符标记，以显示后 10 个或前 10 个字段（按预测变量重要性排序）。

您可以双击图表以打开单独的对话框，您可以在其中编辑图形设置。 例如，您可以修改项目（如图形大小以及使用的字体大小和颜色）。 关闭此单独的编辑对话框后，更改会应用于“输出”选项卡中显示的图表。

相关图

将对每个目标显示相关图（即，自相关图），并且该图显示了残值（期望值与实际值之间的差值）与时间延迟的自相关函数 (ACF) 或偏自相关函数 (PACF)。 置信区间在整个图表中突出显示。

参数估计

对于每个目标，参数估计值重复显示（适用时），其中包含以下详细信息：

- 目标名称
- 所应用的变换
- 对模型 (ARIMA) 中此参数使用的延迟
- 系数值
- 参数估计值的标准误差
- 参数估计值除以标准误差后的值
- 参数估计的显著性水平。

“时间序列”模型块设置

“设置”选项卡为“时间序列”模型块提供其他选项。

预测

用于**将记录扩展到未来**的选项。 设置在估计期结束之后要预测的时间间隔数量。 在这种情况下，时间间隔是在“时间序列”节点的“数据规范”选项卡上指定的分析时间间隔。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。

计算输入的未来值。 如果选择此选项，那么会计算预测变量、噪声预测、差异估算和未来时间值的预测值。

要在预测中使用的未来值

- **计算输入的未来值** 如果选择此选项，那么会自动计算预测变量、噪声预测、差异估算和未来时间值的预测值。 请求进行预测时，将为所有并非同时作为目标的输入序列自动构建自回归模型。 然后，使用这些模型生成这些输入序列在预测期内的值。
- **选择要将其值添加到数据中的字段**。 对于要预测的每条记录（不包括保留值），如果您使用的是预测变量字段（角色设置为 Input），那么可以为每个预测变量指定预测周期的估计值。 您可以手动指定值，也可以从列表中选择值。
  - **字段**。 单击“字段选择器”按钮并选择可用作预测变量的任何字段。 请注意，在此选择的字段可能用于建模，也可能不用于建模；要将某个字段实际用作预测变量，必须在某个下游建模节点中选择该字段。 此对话框为您简单提供了指定未来值的便捷环境，这样可以在多个下游建模节点之间共享这些值，而无需在每个节点中单独进行指定。 另请注意，可用字段的列表可能会受到“构建选项”选项卡中选项的约束。

    请注意，如果为流中不再可用（因为已将其删除或由于“构建选项”选项卡中的选项更新）的字段指定了未来值，那么此字段将显示为红色。
  - **值**。 对于每个字段，可以在函数列表中进行选择，也可以单击**指定**手动输入值或从预定义值列表中选择值。 如果预测变量字段与您所控制的项目或其他预先可知的项目相关，则应手动输入值。 例如，如果要根据房间预订数目预测饭店下个月的收入，可以指定在该期间实际具有的预订数目。 相反，如果预测变量字段与您无法控制的某些因素（如股票价格）相关，那么可以使用函数，如“最近值”或“最近点的均数”。

    可用的函数取决于字段的测量级别。

表 *29:* 可用于测量级别的函数

测量级别	**函数**
连续或名义字段。	空 最近点的均数 最近的值 指定
标志字段	空 最近的值 True False 指定

最近点的均数根据最后三个数据点的均数计算未来值。

最近值将未来值设置为最近数据点的值。

真/假将标志字段的未来值设置为指定的真值或假值。

指定打开一个对话框，用于手动指定未来值或从预定义列表中选择未来值。

使其可用于评分

为要评分的每个模型创建新字段。 使您可以指定为每个要进行评分的模型创建新字段。

- **噪声残值**。 如果选中了此选项，那么对于每个目标字段，将为模型残差创建新字段（带有缺省前缀 $TSResidual-），并同时创建这些值的总计。
- **置信度上限和下限**。 如果选中了此选项，那么对于每个目标字段，此选项将分别为置信区间上限和下限创建新字段（带有缺省前缀 $TSLCI- 和 $TSUCI-），并同时创建这些值的总计。

包含用于评分的目标。 选择要包含在模型评分中的可用目标。

第 14 章 自学响应节点模型

SLRM 节点

使用 **自学响应模型** (SLRM) 节点，可以构建这样的模型：随着数据集的增长，可以不断对其进行更新或重新估计，而不必每次使用整个数据集重新构建该模型。 例如，如果有多个产品，而您希望确定某位客户获得报价后最有可能购买的产品，那么这种模型将十分有用。 此模型可用于预测最适合客户的报价，以及该报价被接受的概率。

最初构建模型时，可以使用较小的数据集，其中的报价和对这些报价的响应可以随机选择。 随着数据集的增长，模型可得到更新，从而越发能够根据其他输入字段（如年龄、性别、职业和收入）预测最适合客户的报价以及这些客户接受报价的概率。 可以通过在节点对话框中添加或删除这些可用报价对其进行更改，而不必更改数据集的目标字段。

如果与 IBM SPSS 协作和部署服务 一起使用，那么可以为模型设立自动定期更新。 该过程不需要人工监督或操作就可以为不可能或没必要由数据挖掘者定制干预的组织和应用程序提供灵活且成本低的解决方案。

示例。 某金融机构希望通过向每个客户提供最有可能接受的报价来获取更多的利润。 您可以使用自学模型来根据先前的促销活动确定最有可能对活动作出积极响应的客户的特征，并根据最近的客户响应实时更新该模型。

SLRM 节点字段选项

执行 SLRM 节点之前，必须在节点的“字段”选项卡上同时指定目标字段和目标响应字段。

目标字段。 从列表中选定目标字段；例如，包含要为客户提供的不同产品的名义（集合）字段。

注：目标字段的存储格式必须为字符串而不是数字。

目标响应字段。 从列表中选择目标响应字段。 例如，接受或拒绝。

注：此字段必须是标志字段。 标志的真值表示报价接受，假值表示报价拒绝。

此对话框中的剩余字段是整个 IBM SPSS Modeler 中通用的标准字段。 有关更多信息，请参阅主题 第 23 页的『建模节点字段选项』。

注：如果源数据包括要用作连续（数值范围）输入字段的范围，那么您必须确保元数据包含每个范围的最小值和最大值详细信息。

SLRM 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

继续训练现有模型。 缺省情况下，每次执行建模节点时，将创建一个全新的模型。 如果选中该选项，那么会继续训练该节点成功生成的最后一个模型。 这样就可以在无需访问原始数据的情况下更新或刷新现有的模型， 并可能会显著提升性能，这是因为 只有新的或更新后的记录被反馈到流中。 上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型调色板中不再可用的情况下，也可以使用该选项。

目标字段值。缺省情况下，此选项设置为**使用全部**，表示将构建其中包含与选定目标字段值相关联的每个报价的模型。 如果希望生成仅包含目标字段的某些报价的模型，请单击 **指定** ，并使用 **添加** 、**编辑** 和 **删除** 按钮输入或修改要为其构建模型的报价的名称。 例如，如果选择的目标是列出提供的所有产品，那么可以使用此字段将提供报价的产品限制为在此输入的产品。

模型评估。 此面板中的字段与模型无关，因为这些字段不会影响评分。 不过，这些字段有助于形成一个直观表示，显示模型预测结果的准确程度。

注：要在模型块中显示模型评估结果，您还必须选中**显示模型评估**复选框。

- **包括模型评估**。 选中此复选框可以创建针对每项选定报价显示模型的预测准确性的图形。

- **设置随机种子值**。 根据随机百分比估算模型的准确性时，您可以通过此选项在另一会话中复制相同结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。
- **模拟样本大小**。 指定评估模型时要在样本中使用的记录数。 缺省值为 100。
- **迭代数**。 通过此选项，您可以在迭代次数达到指定值后停止构建模型评估。 指定最大迭代次数；缺省值为 20。

  注：请记住，如果样本大小较大并且迭代次数较多，那么这将增加构建模型所用的时间。

显示模型评估。 选中此选项将在模型块中显示结果的图形表示。

SLRM 节点设置选项

使用节点设置选项可微调模型构建过程。

每个记录的最大预测数。 通过此选项，您可以限制对数据集中每条记录进行的预测数。 缺省是 3。

例如，您可能有六个产品（如储蓄、抵押、车贷、养老金、信用卡和保险），但您只想了解其中最好的两个来进行推荐；在这种情况下，您可以将此字段设置为 2。 当您构建模型并将其附加到表时，会看到每条记录有两个预测列（以及相关的产品被接受概率的置信度）。 预测可以由六种可能报价中的任意报价组成。

随机化级别。 为了避免出现任何偏差（例如，在较小或不完整的数据集中）并且平等对待所有可能的报价，您可以对报价的选择及其成为推荐报价的概率添加随机化级别。 随机化表示为百分比，以 0.0（无随机化）与 1.0（完全随机化）之间小数值的形式显示。 缺省值为 0.0。

设置随机种子值。 向选择的报价添加随机化级别时，您可以通过此选项在另一个会话中复制相同结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。

注: 对从数据库中读取的记录使用**设置随机种子**选项时，可能需要在抽样前使用“排序”节点以确保每次执行节点时都获得相同的结果。 这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。

排序顺序。 选择报价在已构建模型中的显示顺序：

- **降序**。 模型首先显示评分最高的报价。 这些报价被接受的概率最高。
- **升序**。 模型首先显示评分最低的报价。 这些报价被拒绝的概率最高。 例如，在决定要从某种特定报价的营销活动中删除哪些客户时，这种顺序相当实用。

目标字段的首选项。 构建模型时，数据中可能存在您希望主动提升或移除的特定方面。 例如，如果构建用于选择为某个客户推荐的最佳财务报价的模型，您可能需要确保始终包含一种特定报价（无论其对于每个客户的评分如何）。

要在此面板中包含某项报价并编辑其首选项，请单击**添加**，键入报价的名称（例如，“储蓄”或“抵押”），然后单击**确定**。

- **值**。 此选项将显示您添加的报价的名称。
- **首选项**。 指定要对报价应用的首选度级别。 首选度表示为百分比，以 0.0（非首选）与 1.0（最首选）之间小数值的形式显示。 缺省值为 0.0。
- **始终包含**。 要确保某项特定报价始终包括在预测中，请选中此框。

  注: 如果**首选率**设置为 0.0，那么将忽略**始终包括**设置。

考虑模型可靠性。 与包含少量数据的全新模型相比，已通过多次重新生成来进行微调的结构良好、数据丰富的模型应当始终生成更准确的结果。 要利用较成熟模型具有的较高可靠性，请选中此框。

SLRM 模型块

注：如果在“模型选项”选项卡上同时选中**包括模型评估**和**显示模型评估**，那么结果只会显示在此选项卡。

在运行包含 SLRM 模型的流时，该节点会估计每个目标字段值（报价）的预测准确性，以及所用的每个预测变量的重要性。

注：如果在建模节点的“模型”选项卡上选中了**继续训练现有模型**，那么将在每次重新生成模型时更新模型块上显示的信息。

对于使用 IBM SPSS Modeler 12.0 或更高版本构建的模型，模型块的“模型”选项卡分为两列：

左列。

- **视图**。 有多项报价时，请选择要显示其结果的一项报价。
- **模型性能**。 此部分显示每项报价的估计模型准确性。 测试集合通过模拟生成。

右列。

- **视图**。 选择要显示**与响应的关联**还是**变量重要性**详细信息。
- **与响应相关联**。 显示每个预测变量与目标变量之间的关联（相关性）。
- **预测变量的重要性.** 表示在估计模型过程中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 虽然在使用 SLRM 的情况下，图形是由 SLRM 算法模拟生成的，但该图表可用解释其他显示预测变量重要性的模型的方式进行解释。 方法是：依次从模型中删除每个预测变量，然后查看此操作对模型准确性的影响如何。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

SLRM 模型设置

在 SLRM 模型块的“设置”选项卡中可指定选项以修改已构建的模型。 例如，可以通过 SLRM 节点使用相同的数据和设置构建几个不同的模型，然后使用每个模型中的此选项卡对设置稍做修改以查看其对结果的影响。

注: 只有将模型块添加到流中之后，此选项卡才可用。

每个记录的最大预测数。 通过此选项，您可以限制对数据集中每条记录进行的预测数。 缺省是 3。

例如，您可能有六个产品（如储蓄、抵押、车贷、养老金、信用卡和保险），但您只想了解其中最好的两个来进行推荐；在这种情况下，您可以将此字段设置为 2。 当您构建模型并将其附加到表时，会看到每条记录有两个预测列（以及相关的产品被接受概率的置信度）。 预测可以由六种可能报价中的任意报价组成。

随机化级别。 为了避免出现任何偏差（例如，在较小或不完整的数据集中）并且平等对待所有可能的报价，您可以对报价的选择及其成为推荐报价的概率添加随机化级别。 随机化表示为百分比，以 0.0（无随机化）与 1.0（完全随机化）之间小数值的形式显示。 缺省值为 0.0。

设置随机种子值。 向选择的报价添加随机化级别时，您可以通过此选项在另一个会话中复制相同结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。

注: 对从数据库中读取的记录使用**设置随机种子**选项时，可能需要在抽样前使用“排序”节点以确保每次执行节点时都获得相同的结果。 这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。

排序顺序。 选择报价在已构建模型中的显示顺序：

- **降序**。 模型首先显示评分最高的报价。 这些报价被接受的概率最高。
- **升序**。 模型首先显示评分最低的报价。 这些报价被拒绝的概率最高。 例如，在决定要从某种特定报价的营销活动中删除哪些客户时，这种顺序相当实用。

目标字段的首选项。 构建模型时，数据中可能存在您希望主动提升或移除的特定方面。 例如，如果构建用于选择为某个客户推荐的最佳财务报价的模型，您可能需要确保始终包含一种特定报价（无论其对于每个客户的评分如何）。

要在此面板中包含某项报价并编辑其首选项，请单击**添加**，键入报价的名称（例如，“储蓄”或“抵押”），然后单击**确定**。

- **值**。 此选项将显示您添加的报价的名称。
- **首选项**。 指定要对报价应用的首选度级别。 首选度表示为百分比，以 0.0（非首选）与 1.0（最首选）之间小数值的形式显示。 缺省值为 0.0。
- **始终包含**。 要确保某项特定报价始终包括在预测中，请选中此框。

  注: 如果**首选率**设置为 0.0，那么将忽略**始终包括**设置。

考虑模型可靠性。 与包含少量数据的全新模型相比，已通过多次重新生成来进行微调的结构良好、数据丰富的模型应当始终生成更准确的结果。 要利用较成熟模型具有的较高可靠性，请选中此框。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

第 15 章 支持向量机模型

关于 SVM

支持向量机 (SVM) 是一项功能强大的分类和回归技术，可最大化模型的预测准确度，而不会过度拟合训练数据。 SVM 特别适用于分析预测变量字段非常多（如数千个）的数据。

SVM 适用于多个学科，例如客户关系管理 (CRM)、面部图像和其他图像识别、生物信息学、文本挖掘概念提取、入侵检测、蛋白质结构预测以及语音识别。

SVM 如何运行

SVM 的工作原理是将数据映射到高维特征空间，这样即使数据不是线性可分，也可以对该数据点进行分类。找到类别之间的分隔符，然后以将分隔符绘制成超平面的方式变换数据。 之后，可用新数据的特征预测新记录所属的组。

例如，请考虑下图，图中的数据点落在两个不同的类别中。

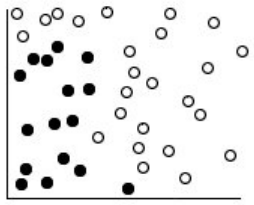


图 *59:* 原始数据集

可以使用一条曲线分隔这两个类别，如下图所示。

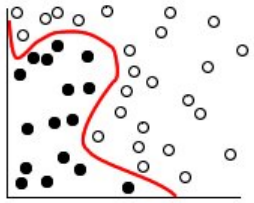


图 *60:* 添加了分隔符的数据

转换后，可以使用超平面定义这两个类别之间的边界，如下图所示。

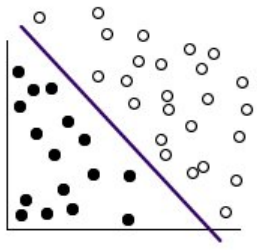


图 *61:* 转换后的数据

用于变换的数学函数称为 **核** 函数。 IBM SPSS Modeler 中的 SVM 支持下列核类型：

- 线性

- 多项式
- 径向基函数 (RBF)
- Sigmoid

如果数据的线性分隔比较简单，那么建议使用线性核函数。 在其他情况下，应当使用其他函数。 在所有情况下，您都需要尝试使用不同的函数才能获得最佳模型，因为每一个函数均使用不同的算法和参数。

调整 SVM 模型

除了类别之间的分隔线，分类 SVM 模型还会查找用于定义两个类别之间的空间的边际线。

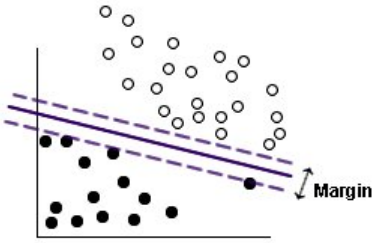


图 *62:* 使用初步模型的数据

位于边距上的数据点称为 **支持向量** 。

两个类别之间的边距越宽，模型在预测新记录所属的类别方面性能越佳。 在上一个示例中，边距不是很宽，因此称该模型 **过度拟合** 。 为了增加边界的宽度，可以接受少量的误分类；下图中显示了一个这样的示例。

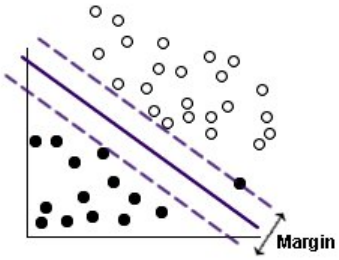


图 *63:* 使用改进模型的数据

在某些情况下，线性分隔难度较大；下图中显示了一个这样的实例。

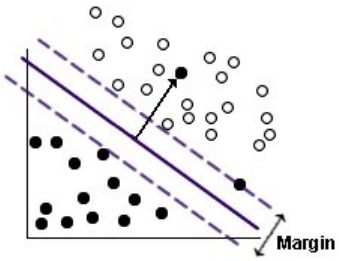


图 *64:* 线性分隔存在的问题

在类似这种情况中，目标是找到宽边距和少量误分类数据点之间的最佳平衡。 核函数有一个 **规则化参数** （称为 C），该参数控制这两个值之间的平衡。 如果要获得最佳模型，您可能需要对该参数和其他核参数尝试使用不同的值。

SVM 节点

通过 SVM 节点，可以使用支持向量机对数据进行分类。 SVM 特别适合于大型数据集，即具有大量预测变量字段的数据集。 可以对节点使用缺省设置以相对较快地生成基本模型，也可以使用“专家”设置以尝试使用不同类型的 SVM 模型。

生成模型后，您可以：

- 浏览模型块，以显示生成模型过程中相对比较重要的输入字段。
- 将表节点附加到模型块中，以查看模型输出。

示例。 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。 通过对原始数据进行分析，发现良性样本与恶性样本之间的许多特征显著不同。 该研究人员希望开发一个 SVM 模型，该模型可以使用其他患者的样本中相似细胞特征的值，以尽早发现他们的样本是良性还是恶性。

SVM 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

SVM 节点专家选项

如果您对支持向量机具有深入了解，那么可以使用专家选项对训练过程进行调整。 要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

追加所有概率（仅对分类目标有效）。 如果选中了该选项，那么将指定针对节点所处理的每条记录显示名义或标志目标字段的每个可能值的概率。 如果未选中该选项，那么仅为名义或标志目标字段显示预测值的概率。 该复选框的设置将决定模型块上的相应复选框的缺省状态。

中止条件。 确定何时停止优化算法。 值范围从 1.0E–1 到 1.0E–6;，缺省值为 1.0E–3。 降低值会生成更准确的模型，但该模型需要更长时间进行训练。

规则化参数(C)。 控制最大化边距和最小化训练错误项之间的平衡。 通常情况下，值应当介于 1 和 10（含本数）之间；缺省值为 10。 增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

回归精度 (epsilon)。 仅当目标字段的测量级别为连续时才使用。 如果错误数小于此处指定的值，那么可以接受错误数。 增加该值可能会加快建模速度，但要以准确度为代价。

内核类型。 确定用于变换的核函数的类型。 核类型不同，计算分隔符的方法也将不同，因此建议尝试使用不同的选项。 缺省值为 **RBF** （径向基函数）。

RBF 伽玛。 仅在核类型设置为 **RBF** 时才启用。 通常情况下，值应当介于 3/ *k* 和 6/ *k* 之间，其中 *k* 为输入字段的数量。 例如，如果有 12 个输入字段，那么应当尝试使用介于 0.25 和 0.5 之间的值。 增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

伽玛。 仅在核类型设置为 **多项式** 或 **Sigmoid** 时才启用。 增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

偏差。 仅在核类型设置为 **多项式** 或 **Sigmoid** 时才启用。 在内核函数中设置 `coef0` 值。 大多数情况下可以使用缺省值 0。

度。 仅在核类型设置为 **多项式** 时才启用。 控制映射空间的复杂性（维度）。 通常情况下，不使用大于 10 的值。

SVM 模型块

SVM 模型会创建许多新字段。 其中最重要的是 **$S-fieldname** 字段，该字段显示由模型预测的目标字段值。

模型创建的新字段的数量和名称取决于目标字段的测量级别（此字段在下表中由字段名指示）。

要查看这些字段及其值，请将表节点添加到 SVM 模型块中，然后执行表节点。

表 *30:* 目标字段测量级别为“名义”或“标记”	
新字段名	**描述**
$S-*fieldname*	目标字段预测值。
$SP-*fieldname*	预测值概率。
$SP-*value*	名义或标志的各个可能值的概率（仅在选中模型块中“设置”选项卡上的**追加所有概率**时才显示）。
$SRP-*value*	（仅适用于标志目标）原始 (SRP) 和调整后的 (SAP) 倾向评分，表示目标字段结果为“真”的可能性。 仅当在生成模型之前选中 SVM 建模节点的“分析”选项卡上的相应复选框之后，才显示这些评分。 有关更多信息，请参阅主题 第 25 页的『建模节点分析选项』。
$SAP-*value*	

表 *31:* 目标字段测量级别为“连续”	
新字段名	**描述**
$S-*fieldname*	目标字段预测值。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

注：与其他类型的模型相比，计算 SVM 的预测变量重要性可能需要更长时间，因此缺省情况下在“分析”选项卡中未选中预测变量重要性。 选中该选项可能会降低性能，对大数据集尤为明显。

SVM 模型设置

通过“设置”选项卡可以指定在查看结果时显示的附加字段（例如，通过执行表节点附加到块）。 通过选择这些选项可以查看每个选项的效果,并且单击“预览”按钮（滚动至“预览”输出右侧）可以查看附加字段。

追加所有概率（仅对分类目标有效）。 如果选中该选项，那么为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。 如果未选中该选项，那么仅为名义或标志目标字段显示预测值及其概率。

此复选框的缺省设置由建模节点的相应复选框确定。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选中下列其中一个选项以指定 SQL 生成的执行方式。

- **缺省值：使用服务器评分适配器（如果已安装）进行评分，否则在过程中进行评分**如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库外进行评分**此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

LSVM 节点

通过 LSVM 节点，您可以使用支持向量机对数据进行分类。 LSVM 特别适用于大型数据集，即具有大量预测变量字段的数据集。 可以对节点使用缺省设置以便相对较快地生成基本模型，也可以使用构建选项来试用不同的设置。

LSVM 节点类似于 SVM 节点，但它是线性的，更擅长处理大量记录。

生成模型后，您可以：

- 浏览模型块，以显示生成模型过程中相对比较重要的输入字段。
- 将表节点附加到模型块中，以查看模型输出。

示例。 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。 通过对原始数据进行分析，发现良性样本与恶性样本之间的许多特征显著不同。 该研究人员希望开发一种 LSVM 模型，该模型可以使用其他患者的样本中相似细胞特征的值，以尽早发现他们的样本是良性还是恶性。

LSVM 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

计算预测变量重要性。 对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。 通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。 请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此缺省情况下，对于某些模型，预测变量重要性均处于关闭状态。 预测变量重要性对于决策列表模型不可用。 有关更多信息，请参阅第 32 页的『预测变量重要性』。

LSVM 构建选项

模型设置

包括截距。 包括截距（模型中的常数项）可以提高解的总体准确度。 如果您可以假设数据穿过原点，那么可以排除截距。

分类目标的排序顺序。 指定分类目标的排序顺序。 对于连续目标，将忽略此设置。

回归精度 (epsilon)。 仅当目标字段的测量级别为连续时才使用。 如果错误数小于此处指定的值，那么可以接受错误数。 增加该值可能会加快建模速度，但要以准确度为代价。

排除具有任何缺失值的记录。 设置为 **True** 时，如果任何单个值缺失，那么将排除记录。

惩罚设置

罚函数。 指定用于降低过度拟合可能性的罚函数的类型。 选项为 **L1** 或 **L2**。

L1 和 **L2** 通过增加系数罚分来降低过度拟合的几率。 二者的区别是当有大量特征时，在模型构建期间，**L1** 通过将某些系数设置为 0 来使用特征选择。 **L2** 不具备此能力，因此在有大量特征时不应使用该选项。

惩罚参数 (lambda)。 指定惩罚（规则化）参数。 如果设置了**罚函数**，那么将启用此设置。

LSVM 模型块（交互式输出）

运行 LSVM 模型后，以下输出可用。

模型信息

“模型信息”视图提供了有关模型的关键信息。 该表标识了一些高级模型设置，例如：

- 在“字段”选项卡中指定的目标的名称
- 模型选择设置上指定的模型构建方法

- 预测变量输入的数量。
- 最终模型中预测变量的数量
- 规则化类型（L1 或 L2）
- 惩罚参数 (lambda)。 这是规则化参数。
- 回归精度 (epsilon)。 如果错误小于此值，那么将接受这些错误。 更高的值可能会加快建模速度，但要以准确性为代价。 仅当目标字段的测量级别为连续时，此项才可用。
- 分类准确性百分比。 这仅适用于分类。
- 平均平方误差。 这仅适用于回归。

记录摘要

“记录摘要”视图提供了有关模型中包括和排除的记录（观测值）的数目和百分比的信息。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

按已观测进行预测

这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。 理想情况下，该点应在 45 度线上；您可以从该视图上判断出任何被模型预测为较差的纪录。

注: 与其他类型的模型相比，计算 LSVM 和 SVM 的预测变量重要性可能需要更长时间。 选中该选项可能会降低性能，对大数据集尤为明显。

混淆矩阵

混淆矩阵（有时也称为摘要表）显示了根据 LSVM 分析正确和不正确分配给每个组的观测值数。

LSVM 模型设置

在 SVLM 模型块的“设置”选项卡上，您可以指定模型评分期间用于原始倾向的选项和用于 SQL 生成的选项。只有将模型块添加到流之后，此选项卡才可用。

计算原始倾向评分 对于仅具有标志目标的模型，您可以请求原始倾向评分，这些评分指示为目标字段指定的 true 结果的发生可能性。 除此之外，标准预测及置信度值也是如此。 调整后的倾向评分不可用。

为此模型生成 SQL：使用数据库中的数据时，可以将 SQL 代码推回到数据库中以进行执行，这可以极大地提高许多操作的性能。

选择下列其中一个选项以指定 SQL 的生成方式。

- **缺省值：使用 Server Scoring Adapter（如果已安装）进行评分，否则在过程中进行评分**。 如果连接到安装有评分适配器的数据库，将使用评分适配器和用户定义的功能 (UDF) 生成 SQL，并在数据库中对您的模型进行评分。 如果没有可用的评分适配器，那么此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。
- **在数据库之外进行评分**。 此选项会从数据库访存回您的数据，并在 SPSS Modeler 中对其进行评分。

第 16 章 最近相邻元素模型

KNN 节点

“最近邻元素分析”方法是根据个案间的相似性来对个案进行分类。 在 machine learning 中，它被开发为一种识别数据模式而不需要与任何存储的模式或个案完全匹配的方法。 类似观测值相互靠近，而不同观测值相互远离。 因此，通过两个个案之间的距离可以测量他们的非相似性。

相互靠近的个案称为“邻元素”。 当提出新的观测值（holdout 观测值）时，计算其到模型中每个观测值的距离。 计算最相似观测值（最近邻元素）的分类，并将新观测值放在包含最多最近邻元素的类别中。

您可以指定要检查的最近邻元素数目，该值称为 *k*。 这些图显示了如何使用两个不同的 *k* 值对新案例进行分类。 当 *k* = 5 时，新案例将放在类别 *1* 中，因为大多数最近邻元素都属于类别 *1*。 但当 *k* = 9 时，新观测值将放在类别 *0* 中，因为大多数最近邻元素都属于类别 *0*。

最近邻元素分析也可用于计算连续目标的值。 在此情况下，最近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。

KNN 节点目标选项

您可以在“对象”选项卡输入数据中根据最近相邻元素的值选择构建预测目标字段值的模型，或者只是查找特定感兴趣观测值的最近相邻元素。

您要执行那种类型的分析?

预测目标字段。 如果您想根据最近相邻元素的值预测目标字段的值，请选择此选项。

仅识别最近相邻元素。 如果您只希望看到特定输入字段的最近相邻元素，请选择此选项。

如果您选择只识别最近相邻元素，在此选项卡上与准确性和速度相关的剩余选项将被禁用，因为其只与预测目标相关。

您的目标是什么?

预测目标字段时，您可以通过此组选项来决定速度、准确性或这二者的组合是否为最重要的因素。 或者您可以选择自己定制设置。

如果您选择平衡、速度或准确性选项，那么算法预先选择该选项的最合适设置组合。 高级用户可能希望覆盖这些选择；可在“设置”选项卡上的各个面板上进行此操作。

平衡速度与准确度。 选择小范围内相邻元素的最佳数量。

速度。 找出固定数量的相邻值。

准确性。 选择较大范围内的相邻元素的最佳数量，并在计算距离时使用预测变量重要性。

定制分析。 选择该选项以微调“设置”选项卡上的算法。

注: 与大多数其他模型不同的是，生成的 KNN 模型的大小随着训练数据量的增大呈线性增加。 如果在尝试构建 KNN 模型时看到报告“内存溢出”错误的出错信息，那么尝试增加 IBM SPSS Modeler 所使用的最大系统内存。 要进行此操作，请选择

工具 > 选项 > 系统选项

并在**最大内存**字段中输入新大小。 “系统选项”对话框中所作的更改要在重新启动 IBM SPSS Modeler 之后才能生效。

KNN 节点设置

在“设置”选项卡上您可以指定最近相邻元素分析特有的选项。 屏幕左侧的侧栏列出了用于指定选项的面板。

模型

“模型”窗格提供控制如何构建模型的选项，例如是否使用分区或分割模型、是否变换数值输入字段以使其落入相同范围内和如何管理感兴趣观测值。 您也可以给模型选择一个定制名称。

注: 使用分区数据和**使用观测值标签**不能使用同一字段。

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

创建拆分模型。 给指定为分割字段的输入字段的每个可能值构建一个单独模型。 有关更多信息，请参阅第 21 页的『构建分割模型』。

要手动选择字段... 缺省情况下，该节点使用分区，拆分来自“类型”节点的字段设置（如果有的话），但您可以在此处覆盖这些设置。 要激活**分区**与**分割**字段，请选择**字段**选项卡，并选择**使用定制设置**，然后返回此处。

- **分区**。 通过此字段，您可以指定用于针对模型构建中的训练、检验和验证阶段将数据划分为不同样本的字段。 通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。 如果已使用类型或分区节点定义了多个分区字段，那么必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。 （如果仅有一个分区字段，则将在启用分区后自动引入此字段。） 同时请注意，要在分析时应用选定分区，还必须启用节点的“模型选项”选项卡中的分区功能。 （取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）
- **拆分**。 对于分割模型，选择分割字段或字段。 此操作与在“类型”节点中将字段的角色设置为分割类似。 您可以仅将类型为**标志**、**名义**或**有序**的字段指定为分割字段。 选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关更多信息，请参阅主题 第 21 页的『构建分割模型』。

标准化范围输入。 选中此复选框为连续输入字段标准化值。 标准化特征具有相同的值范围，这可改进估计算法的性能。 使用调整后规范化 [2*(x-min)/(max-min)]-1。 调整后的标准化值介于 -1 与 1 之间。

使用案例标签。 选中此复选框以启用下拉列表，从这里您可以选择字段并将其值用作标签，以在“模型查看器”中标识在预测变量空间图表、对等图表和象限图中所需的观测值。 您可以选择测量级别为名义、有序或标志的任何字段用作标签字段。 如果您不在这里选择字段，那么用以源数据中行号标识的最近邻元素在“模型查看器”图表中显示记录。 如果您在构建模型之后要操作数据，可使用观测值标签，以避免每次需要参考源数据在显示中标识观测值。

识别焦点记录。 选中此复选框启用下拉列表，允许您标记感兴趣的输入字段（仅针对标志字段）。 如果在此处指定了一个字段，那么当构建模型时会在模型查看器中初始选中代表该字段的点。 在此处选择焦点记录是可选的；任何点都可以暂时成为焦点记录，只要在“模型查看器”中手动选中它。

相邻元素

“相邻元素”窗格具有一组控制如何计算最近相邻元素数量的选项。

最近邻元素数目 (k)。 指定特定观测值的最近相邻元素数量。 注意，使用大量的邻元素不一定会得到更准确的模型。

如果目标是预测目标，那么您具有两个选择：

- **指定固定 K**。 如果要指定要查找的最近邻元素的固定数目，请使用此选项。
- **自动选择 k**。 您也可以使用**最小值**和**最大值**字段以指定一个数值范围，并允许该过程选择该范围内相邻元素的“最佳”数量。 确定最近相邻元素数目的方法依赖于“特征选择”窗格上要求的特征选择。

  如果特征选择有效，那么针对请求范围中每个 *k* 值执行特征选择，并选择具有最低误差率（如果目标为连续，那么为最低平方和误差）的 *k* 值和特征集。

  如果特征选择未生效，那么使用 *V* 折交叉验证来选择“最佳”的邻元素数目。 请参阅“交叉验证”窗格以控制折叠指定。

距离计算。 该度规用于指定在测量个案相似性中使用的距离度规。

- **欧几里得度量**。 两个个案 x 和 y 之间的距离，为个案值之间的平方差在所有维度上之和的平方根。
- **城市街区度量**。 两个个案之间的距离是个案值之间绝对差在所有维度上之和。 又称为 Manhattan 距离。

或者，如果目标是预测目标，您可以选择在计算距离时按照其标准化重要性计算特征权重。 预测变量的特征重要性的计算方法为：不含预测变量的模型的误差率或平方和误差与完整模型的误差率或平方和误差之比。 通过重新对特征重要性值指定权重，来计算标准化的重要性，因此其总和为 1。

计算距离时按重要性对特征进行加权。 （只有当目标是预测目标时才显示。） 选中此复选框，当计算相邻元素之间距离时，使用预测变量重要性。 预测变量重要性将在模型块中显示，并用于预测（因此影响评分）。 有关更多信息，请参阅主题 第 32 页的『预测变量重要性』。

范围目标的预测。 （只有当目标是预测目标时才显示。） 如果指定了连续（数值范围）目标，这可指定预测值是基于最近相邻元素的均值还是中值来计算的。

特征选择

只有在目标是预测目标时才激活此窗格。 使您能够为特征选择请求和指定选项。 缺省情况下，特征选择会考虑所有特征，但可以选择特征子集以强制纳入模型。

执行特征选择。 选中此复选框启用特征选择选项。

- **强制进入**。 单击此框旁的字段选择按钮并选择一个或多个特征以强制纳入模型。

停止标准。 在每一步上，如果添加特征可以使误差最小（计算为分类目标的误差率和连续目标的平方和误差），那么考虑将其纳入模型中。 继续向前选择，直到满足指定的条件。

- **在选择了指定数量的特征后停止**。 除了那些强制纳入模型的特征外，算法还会添加固定数目的特征。 请指定正整数。 减少所选择的数目值可以创建更简约的模型，但存在缺失重要特征的风险。 增加所选择的数目值可以涵盖所有重要特征，但又存在因特征添加而增加模型误差的风险。
- **当绝对误差比率中的变化小于或等于最小值时停止**。 当绝对误差比率变化表明无法通过添加更多特征来进一步改进模型时，算法会停止。 指定一个正数。 减少最小变化值将倾向于包含更多特征，但存在包含对模型价值不大的特征的风险。 增加最小变化值将倾向于排除更多特征，但存在丢失对模型较重要的特征的风险。 “最佳”的最小变化值取决于数据和具体应用。 请参阅输出中的“特征选择误差日志”，以帮助您评估哪些特征最重要。 有关更多信息，请参阅主题 第 260 页的『预测变量选择错误日志』。

交叉验证

只有在目标是预测目标时才激活此窗格。 该窗格上的选项控制计算最近相邻元素时是否使用交叉验证。

交叉验证将样本划分为许多子样本，或**折叠**。 然后，生成最近邻元素模型，并依次排除每个子样本中的数据。 第一个模型基于第一个样本折的个案之外的所有个案，第二个模型基于第二个样本折的个案之外的所有个案，依此类推。 对于每个模型，估计其错误的方法是将模型应用于生成它时所排除的子样本。 “最佳”最近邻元素数为在折中产生最小误差的数量。

交叉验证折数。 *V* 折交叉验证用于确定“最佳”邻元素数目。 因性能原因，它无法与特征选择结合使用。

- **将个案随机分配到折**。 指定应当用于交叉验证的折数。 此过程将个案随机分配到折，从 1 编号到 *V*（折数）。
- **设置随机种子值**。 根据随机百分比估算模型的准确性时，您可以通过此选项在另一会话中复制相同结果。 通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。 输入所需的种子值。 如果未选中该选项，则每次执行节点时会生成不同的抽样。
- **使用字段分配个案**。 指定一个将活动数据集中的每个观测值分配到折中的数值字段。 该字段必须为数字，取值范围从 1 到 *V*。 如果此范围内的任何值缺失，并且这些缺失值位于拆分模型有效的任何拆分字段上，那么这将导致错误。

分析

只有在目标是预测目标时才激活“分析”窗格。 您可以使用它指定模型是否要纳入附加变量以包含：

- 每个可能目标字段值的概率
- 观测值和最近邻元素之间的距离
- 原始和调整后的倾向评分（仅适用于标志目标)。

附加所有概率。 如果选中该选项，那么为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。 如果未选中该选项，那么仅为名义或标志目标字段显示预测值及其概率。

保存个案与 k 个最近邻元素之间的距离。 对于每个焦点记录，将为其 *k* 个最近相邻元素（来自培训样本）和对应的 *k* 个最近距离创建单独的变量。

倾向评分

可以在建模节点中和模型块的“设置”选项卡上启用倾向评分。 该功能仅在所选目标为标志字段时才可用。 有关更多信息，请参阅主题 第 26 页的『倾向评分』。

计算原始倾向评分。 原始倾向评分仅派生自基于训练数据的模型。 如果模型预测值为 真（将响应），那么倾向与 P 相同，其中 P 为预测的可能性。 如果模型预测的值为假，那么计算出的倾向为 (1 – P)。

- 如果构建模型时选择了此选项，那么缺省情况下将在模型块中启用倾向评分。 不过，无论是否在建模节点中选择了原始倾向评分，都可以始终在模型块中选择启用原始倾向评分。
- 对模型进行评分时，原始倾向评分将被添加到将 *RP* 字母附加到标准前缀的字段中。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RRP-churn*。

计算调整后的倾向评分。 原始倾向仅基于由可能过度拟合的模型指定的估计，这将导致过于乐观地估计倾向。 调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。
- 与原始置信度分数不同，调整后的倾向评分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 *AP* 字母附加到标准前缀的字段中添加调整后的倾向评分。 例如，如果预测位于名为 *$R-churn* 的字段中，那么倾向评分字段的名称将是 *$RAP-churn*。 调整后的倾向评分不适用于 logistic 回归模型。
- 在计算调整后的倾向评分时，必须尚未平衡用于计算的检验或验证分区。 为避免这一点，请确保在任何上游平衡节点中选中 **仅平衡训练数据** 选项。 此外，如果已在上游获取了复杂样本，那么这将导致调整后的倾向评分无效。
- 调整后的倾向评分不适用于“增强型”树和规则集模型。 有关更多信息，请参阅主题 第 89 页的『增强型 C5.0 模型』。

KNN 模型块

KNN 模型会创建许多新字段，如下表所示。 要查看这些字段及其值，请将表节点添加到 KNN 模型块中，然后执行表节点，或单击模型块上的“预览”按钮。

表 *32: KNN* 模型字段

新字段名	**描述**
$KNN-*fieldname*	目标字段预测值。
$KNNP-*fieldname*	预测值概率。
$KNNP-*value*	名义或标志字段的每个可能值的概率。 只有在模型块的“设置”选项卡上选中了 **追加所有概率**才会被纳入。
$KNN-neighbor-*n*	焦点记录的第 *n* 个最近邻元素名称。 只有当模型块的“设置”选项卡上的**显示最近**设为非零值时才会被纳入。
$KNN-distance-*n*	焦点记录第 *n* 个最近邻元素到焦点记录的相对距离。 只有当模型块的“设置”选项卡上的**显示最近**设为非零值时才会被纳入。

最近相邻元素模型视图

模型视图

此模型视图有 2 个面板窗口：

- 第一个面板显示模型概览，称为主视图。

- 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模型的更多信息，但并不专注于模型本身。

  当用户深入查看主视图某个部分时，链接视图显示有关某个模型特征的详细信息。

缺省情况下，第一个面板显示预测变量空间，第二个面板显示预测变量重要性图表。 如果预测变量重要性图表不可用；即如果未在“设置”选项卡的“相邻元素”面板上选中**按照重要性计算特征权重**，那么将显示“视图”下拉列表中的第一个可用视图。

如果视图不具有可用信息，它将从“视图”下拉列表中省略。

预测变量空间

预测变量空间图表是有关预测变量空间（如果存在 3 个以上预测变量，那么为子空间）的交互式图形。 每条轴代表模型中的某个预测变量，图表中的点位置显示观测值这些预测变量在训练和 holdout 分区中的值。

密钥。 除了预测变量值外，图中的点还传递其他信息。

- 其形状表示点所属的分区，即训练或坚持分区。
- 点的颜色/阴影表示该个案的目标值，不同的颜色值等于分类目标的类别，阴影则表示连续目标的值范围。 训练分区的指示值为观测值；对于坚持分区，那么为预测值。 如果未指定目标，那么不会显示此键。
- 较粗的概要表示个案为焦点个案。 显示的焦点记录链接到它们的 *k* 个最近邻元素。

控件和交互性。 使用图表中的一些控件可以探索预测变量空间。

- 可以选择在图表中显示哪个预测变量子集，还可更改在维度上表示哪些预测变量。
- “焦点记录”仅仅是在“预测变量空间”图表中选定的点。 如果指定了焦点记录变量，那么初始情况下会选中代表焦点记录的点。 不过，如果选中了任何点，那么它都可以暂时成为焦点记录。 可以使用用于选择点的“常规”控件，即，单击一个点将选中该点并取消选中所有其他点；按下 Ctrl 键并单击一个点会将其添加到选择的点集合。 链接的视图，如对等图表，将根据在预测变量空间中选择的观测值自动更新。
- 您可以更改为焦点记录显示的最近邻元素数目 (*k*)。
- 在图表中的点上方悬停，可以显示工具提示以及个案标签值，或个案编号（如果未定义个案标签），以及观察和预测目标值。
- 通过“重置”按钮，您可以将“预测变量空间”恢复到其原始状态。

更改预测变量空间图表上的轴

您可以控制在预测变量空间图表的轴上显示的特征。

要更改轴设置：

1. 单击左侧面板上的“编辑模式”按钮（画笔图标），为预测变量空间选择编辑模式。
2. 在右侧面板中更改视图。 在两个主面板之间出现**显示区域**面板。
3. 单击**显示区域**复选框。
4. 单击预测变量空间中的任何数据点。
5. 要使用具有相同数据类型的预测变量替换某个轴：
   - 将新预测变量拖到您要替换的预测变量的区域标签（带有小 X 按钮）上。
6. 要使用具有不同数据类型的预测变量替换某个轴：
   - 在您要替换的预测变量的区域标签上，单击小 X 按钮。 预测变量空间变为二维视图。
   - 将新预测变量拖到**添加维度**区域标签上。
7. 单击左侧面板上的“探索模式”按钮（箭头图标），退出编辑模式。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。 预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。 由于它们

是相对值，因此显示的所有预测变量的值总和为 1.0。 预测变量重要性与模型精度无关。 它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

最近邻元素距离

该表只显示焦点记录的 *k* 个最近邻元素与距离。 如果焦点记录标识指定在 建模节点上，那么它为可用，且只显示此变量标识的焦点记录。

每行：

- **焦点记录**列包含焦点记录的观测值标签变量值；如果未定义观测值标签，那么此列包含焦点记录的观测值编号。
- 在**最近相邻元素**组下的第 *i* 列包含焦点记录的第 *i* 个最近相邻元素的观测值标签变量值；如果未定义观测值标签，那么此列包含焦点记录第 *i* 个最近相邻元素的观测值号。
- 在**最近距离**组下的第 *i* 列包含第 *i* 个最近相邻元素与焦点记录的距离。

对等

该图表显示焦点观测值及其在每个预测变量和目标上 *k* 个最近邻元素。 它仅在预测变量空间图表中选择了焦点观测值时可用。

对等图表以两种方式链接到预测变量空间。

- 在预测变量空间中所选的观测值（焦点观测值）显示在对等图表中，也包括其 *k* 个最近邻元素。
- 在对等图表中使用在预测变量空间中所选的 *k* 值。

选择预测变量。 使您可选择在对等图表中显示的预测变量。

象限图

该图表显示焦点观测值及其在散点图（或点图，取决于目标的测量级别）上 *k* 个最近邻元素。目标在 *y* 轴上，刻度预测变量在 *x* 轴上，按预测变量划分面板。 它仅当存在目标，且在预测变量空间图表中选择了焦点观测值时可用。

- 在训练分区的变量平均值处，为连续变量绘制了参考线。

选择预测变量。 使您可选择在象限图中显示的预测变量。

预测变量选择错误日志

对于该图表上的点，其 *y* 轴值为模型的误差（误差率或平方和误差，取决于目标的测量级别），*x* 轴上列出模型的预测变量（加上 *x* 轴左侧的所有特征）。 该图表仅当存在目标，且特征选择有效时可用。

分类表

该表显示按分区对目标观察与预测值的交叉分类。 它仅当存在分类目标（标志、名义或有序）时可用。

- 坚持分区中的**（缺失）**行包含在目标上具有缺失值的坚持个案。 这些个案对“坚持样本：整体百分比”值有贡献，但对“正确百分比”值无影响。

误差摘要

该表仅当存在目标变量时可用。 它显示模型的相关误差；即，连续目标的平方和误差以及分类目标的误差率（100% - 总体正确百分比）。

KNN 模型设置

通过“设置”选项卡可以指定在查看结果时显示的附加字段（例如，通过执行表节点附加到块）。 通过选择这些选项可以查看每个选项的效果,并且单击“预览”按钮（滚动至“预览”输出右侧）可以查看附加字段。

追加所有概率（仅对分类目标有效）。 如果选中该选项，那么为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。 如果未选中该选项，那么仅为名义或标志目标字段显示预测值及其概率。

此复选框的缺省设置由建模节点的相应复选框确定。

计算原始倾向评分。 对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向评分，这些评分指示为目标字段指定结果为真的可能性。 除了这些评分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向评分。 原始倾向评分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该评分可能会过度优化。 调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。 此选项要求在流中定义分区字段并且在建模节点中启用调整后的倾向评分后再生成模型。

显示最近。 如果您将此值设为 *n*，其中 *n* 是非零正整数，那么焦点记录的第 *n* 个最近邻元素与其到焦点记录的相对距离一起纳入在模型中。

第 17 章 Python 节点

SPSS Modeler 提供了用于使用 Python 本机算法的节点。 节点选用板上的 **Python** 选项卡包含您可用于运行 Python 算法的下列节点。 这些节点在 Windows 64、Linux64 和 Mac 上受支持。

合成少数类过采样技术 (Synthetic Minority Over-sampling Technique, SMOTE) 节点提供了用于处理不平衡数据集的过采样算法。 它提供了用于均衡数据的高级方法。 SPSS Modeler 中的 SMOTE 流程节点在 Python 中实现，并且需要 `imbalanced-learn`© Python 库。

XGBoost Linear© 是将线性模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 SPSS Modeler 中的 XGBoost Linear 节点使用 Python 进行实现。

XGBoost Tree© 是将树模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost Tree 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost Tree 节点仅显示了核心功能和常用参数。 此节点使用 Python 进行实现。

t-分布随机邻域嵌入 (t-SNE) 是用于可视化高维数据的工具。 其将数据点亲缘关系转换为可能性。 此 t-SNE 节点在 SPSS Modeler 中使用 Python 进行实现并且需要 `scikit-learn`© Python 库。

Gaussian Mixture© 模型是一个概率模型，其假定从有限数量的高斯分布和未知参数混合中生成所有数据点。 可以将混合模型认为是广义 K-Means 聚类以包含有关数据的协方差结构以及潜伏高斯分布的中心的信息。 SPSS Modeler 中的 Gaussian Mixture 节点公开 Gaussian Mixture 库的核心特征和常用参数。 此节点使用 Python 进行实现。

Kernel Density Estimation (KDE)© 使用 Ball Tree 或 KD Tree 算法以进行效率查询，并结合无监督学习、特征工程和数据建模等概念。 基于相邻元素的方法（例如，KDE）是最流行且最有用的一些密度估算方法。 SPSS Modeler 中的 KDE 建模和 KDE 模拟节点公开 KDE 库的核心特征和常用参数。 节点使用 Python 进行实现。

随机森林节点使用将树模型作为基本模型的组装算法的高级实现。 SPSS Modeler 中的此“随机森林”建模节点是在 Python 中实现的，并且需要 `scikit-learn`© Python 库。

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非监督学习来查找数据集的聚类或密集区域。 SPSS Modeler 中的 HDBSCAN 节点公开 HDBSCAN 库的核心特征和常用参数。 此节点以 Python 实现，当您一开始不了解数据集的分组时，可以使用此节点将数据集聚类为不同的组。

单类 SVM 节点使用无监督学习算法。 此节点可用于新内容检测。 它将检测指定样本集的软边界，以便按是否属于该集合对新点进行分类。 SPSS Modeler 中的这个单类 SVM 建模节点是在 Python 中实现的，并且需要使用 `scikit-learn`© Python 库。

SMOTE 节点

合成少数类过采样技术 (Synthetic Minority Over-sampling Technique, SMOTE) 节点提供了用于处理不平衡数据集的过采样算法。 它提供了用于均衡数据的高级方法。 SMOTE 过程节点使用 Python 进行实现并且需要 `imbalanced-learn`© Python 库。 有关 imbalanced-learn 库的详细信息，请参阅 https://imbalanced-learn.org/stable/[1]。

节点选用板上的 Python 选项卡包含 SMOTE 节点和其他 Python 节点。

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*，卷 18，第 17 号，2017 年，第 1-5 页。 (http://jmlr.org/papers/v18/16-365.html)

SMOTE 节点设置

在 SMOTE 节点的**设置**选项卡上定义下列设置。

目标设置

目标字段。 选择目标字段。 支持所有“标志”、“名义”、“有序”和“独立”测量类型。 如果在“分区”部分中选择了**使用分区数据**选项，那么将对训练数据进行过采样。

过采样比率

选择**自动**以自动选择过采样比率，或者选择**设置比率（少数对多数）**以设置定制比率值。 此比率是少数类中的样本数与多数类中的样本数之比。 此值必须大于 **0** 并小于或等于 **1**。

随机种子

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

方法

算法种类。 选择您要使用的 SMOTE 算法的类型。

样本规则

K 近邻。 指定要用于构建合成样本的最近邻居的数量

M 近邻。 指定要用于确定是否少数样本处于危险状态的最近邻居的数量。 仅当选择 **Borderline1** 或 **Borderline2** SMOTE 算法类型时，才会使用此选项。

分区

使用分区数据。 如果您仅希望对训练数据进行过采样，请选择此选项。

此 SMOTE 节点需要 `imbalanced-learn`© Python 库。 下表显示 SPSS Modeler SMOTE 节点对话框中的设置和 Python 算法之间的关系。

表 *33:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	脚本名称（属性名称）	Python API 的参数名称
过采样比率（数字输入控制）	`sample_ratio_value`	`ratio`
随机种子(D)	`random_seed`	`random_state`
K 邻居	`k_neighbours`	`k`
M 邻居	`m_neighbours`	`m`
算法种类	`algorithm_kind`	`kind`

XGBoost Linear 节点

XGBoost Linear© 是将线性模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 SPSS Modeler 中的 XGBoost Linear 节点使用 Python 进行实现。

有关提升算法的更多信息，请参阅 XGBoost 教程，网址为 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。 [1]

请注意，SPSS Modeler 中不支持 XGBoost 交叉验证功能。 您可以将 SPSS Modeler 分区节点用于此功能。另外，请注意，XGBoost 在 SPSS Modeler 中用于自动对分类变量执行独热编码。

[1] "XGBoost Tutorials。" *Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

XGBoost Linear 节点字段

"字段"选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游类型节点（或上游源节点的"类型"选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标和预测变量，请选择此选项。

字段。 使用方向按钮可以将项目从列表中手动分配给屏幕右侧的"目标"和"预测变量角色"字段。 图标表示每个角色字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择要用作预测的目标的字段。

预测变量。 选择一个或多个字段作为预测输入。

XGBoost Linear 节点的"构建选项"选项卡

使用"构建选项"选项卡可以指定 XGBoost Linear 节点的构建选项，包括线性提升参数和模型构建之类的**基本选项**以及用于目标的**学习任务选项**。 有关这些选项的更多信息，请参阅以下在线资源：

- XGBoost 参数引用 [1]
- XGBoost Python API[2]
- XGBoost 主页 [3]

基本

超参数优化（基于 Rbfopt）。 选择此选项以启用基于 Rbfopt 的超参数优化，这将自动发现最佳参数组合，从而使模型在样本上实现期望或更低的错误率。 有关 Rbfopt 的详细信息，请参阅 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

Alpha。 这是有关权重的 L1 规则化术语。增大此值将使模型更保守。

Lambda。 这是有关权重的 L2 规则化术语。 增大此值将使模型更保守。

Lambda 偏差。 这是有关基本选项的 L2 规则化术语。 （没有关于偏差的 L1 规则化术语，因为它不重要。）

提升舍入次数。 这是提升迭代的次数。

学习任务

目标。 请从以下学习任务目标类型中进行选择：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic** 或 **multi**。

随机种子值。 您可以单击**生成**来生成随机数字生成器所使用的种子。

下表显示 SPSS Modeler XGBoost Linear 节点对话框中的设置与 Python XGBoost 库参数之间的关系。

表 *34:* 映射到 *Python* 库参数的节点属性		
SPSS Modeler 设置	**脚本名称（属性名称）**	**XGBoost 参数**
目标	`TargetField`	
预测变量	`InputFields`	
Lambda	`lambda`	`lambda`
Alpha	`alpha`	`alpha`
Lambda 偏差	`lambdaBias`	`lambda_bias`
提升舍入次数	`numBoostRound`	`num_boost_round`
目标	`objectiveType`	`objective`
随机种子	`random_seed`	`seed`

[1]"XGBoost Parameters"*Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

[2]"Plotting API"*Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

[3]"Scalable and Flexible Gradient Boosting"。 Web. © 2015-2016 DMLC。

XGBoost Linear 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

XGBoost Tree 节点

XGBoost Tree© 是将树模型用作基本模型的梯度提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost Tree 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost Tree 节点仅显示了核心功能和常用参数。 此节点使用 Python 进行实现。

有关提升算法的更多信息，请参阅 XGBoost 教程，网址为 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。 [1]

请注意，SPSS Modeler 中不支持 XGBoost 交叉验证功能。 您可以将 SPSS Modeler 分区节点用于此功能。另外，请注意，XGBoost 在 SPSS Modeler 中用于自动对分类变量执行独热编码。

[1] "XGBoost Tutorials。" *Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

XGBoost Tree 节点的"字段"选项卡

"字段"选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游类型节点（或上游源节点的"类型"选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标和预测变量，请选择此选项。

字段。 使用方向按钮可以将项目从列表中手动分配给屏幕右侧的"目标"和"预测变量角色"字段。 图标表示每个角色字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择要用作预测的目标的字段。

预测变量。 选择一个或多个字段作为预测输入。

XGBoost Tree 节点的“构建选项”选项卡

使用“构建选项”选项卡可以指定 XGBoost Tree 节点的构建选项，包括用于模型构建和树增长的**基本选项**、用于目标的**学习任务选项**以及用于控制不平衡数据集的过度拟合及处理的**高级选项**。 有关这些选项的更多信息，请参阅以下在线资源：

- XGBoost 参数引用 [1]
- XGBoost Python API[2]
- XGBoost 主页 [3]

基本

超参数优化（基于 Rbfopt）。 选择此选项以启用基于 Rbfopt 的超参数优化，这将自动发现最佳参数组合，从而使模型在样本上实现期望或更低的错误率。 有关 Rbfopt 的详细信息，请参阅 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

树方法。 选择要使用的 XGBoost Tree 构建算法。

提升迭代次数。 指定提升迭代的次数。

最大深度。 指定树的最大深度。 增大此值将导致模型更复杂，并且很可能出现过度拟合。

最小子代权重。 指定子代中需要的实例权重 (hessian) 的最小总和。 如果树分区步骤生成实例权重总和小于此**最小子代权重**的叶节点，那么构建过程将停止进行进一步分区。 在线性回归模式下，此项简单地对应于每个节点中所需的最小实例数。 权重越大，算法越保守。

最大增量步骤。 指定允许用于每个树的权重估计的最大增量步骤。 如果设置为 **0**，那么没有约束。 如果设置为正值，那么它可以使更新步骤更为保守。 通常不需要此参数，但是在某个类极度不平衡的情况下，它可以用在 Logistic 回归中。

学习任务

目标。 请从以下学习任务目标类型中进行选择：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic** 或 **multi**。

提前停止。 如果想要使用提前停止功能，请选择此选项。 对于**停止舍入**，验证错误必须在每个提前停止舍入处降低才能继续培训。 **评估数据比率**是用于验证错误的输入数据的比率。

随机种子值。 您可以单击**生成**来生成随机数字生成器所使用的种子。

高级

子样本。 子样本是训练实例的比率。 例如，如果您将此项设置为 **0.5**，那么 XGBoost 将随机收集一半的数据实例以生成树，并且这将防止过度拟合。

Eta。 这是更新步骤期间用于防止过度拟合的步长收缩。 在每个提升步骤后，可以直接获取新功能的权重。 Eta 也会缩小功能权重，以使提升过程更保守。

伽玛。 这是对树的某个叶节点进行进一步分区所需的最小损失减小。 伽玛设置越大，算法越保守。

按树进行列采样。 这是构建每个树时列的子样本比率。

按级别进行列采样。 这是在每个级别每个分割的列的子样本比率。

Lambda。 这是有关权重的 L2 规则化术语。 增大此值将使模型更保守。

Alpha。 这是有关权重的 L1 规则化术语。增大此值将使模型更保守。

标度位置权重。 用于控制正权重和负权重的平衡。 这对于不平衡类非常有用。

下表显示 SPSS Modeler XGBoost Tree 节点对话框中的设置与 Python XGBoost 库参数之间的关系。

表 *35:* 映射到 *Python* 库参数的节点属性		
SPSS Modeler 设置	**脚本名称（属性名称）**	**XGBoost 参数**
目标	TargetField	
预测变量	InputFields	
树方法	treeMethod	tree_method
提升舍入次数	numBoostRound	num_boost_round
最大深度	maxDepth	max_depth
最小子代权重	minChildWeight	min_child_weight
最大增量步骤	maxDeltaStep	max_delta_step
目标	objectiveType	objective
提前停止	earlyStopping	early_stopping_rounds
停止舍入	stoppingRounds	
评估数据比率	evaluationDataRatio	
随机种子	random_seed	seed
子样本	sampleSize	subsample
Eta	eta	eta
伽玛	gamma	gamma
列样本（按树列出）	colsSampleRatio	colsample_bytree
列样本（按级别列出）	colsSampleLevel	colsample_bylevel
Lambda	lambda	lambda
Alpha	alpha	alpha
刻度位置权重	scalePosWeight	scale_pos_weight

[1]"XGBoost Parameters"*Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

[2]"Plotting API"*Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

[3]"Scalable and Flexible Gradient Boosting"。 Web. © 2015-2016 DMLC。

XGBoost Tree 节点的"构建选项"选项卡

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

t-SNE 节点

t 分布随机邻域嵌入 (t-Distributed Stochastic Neighbor Embedding, t-SNE)© 是用于将高维数据可视化的工具。 其将数据点亲缘关系转换为可能性。 原始空间中的亲缘关系通过高斯联合概率表示，并且内嵌空间中的亲缘关系通过 Student t 分布表示。 这可使 t-SNE 对于本地结构特别敏感，而且与现有技术相比具有以下优点：[1]

- 在单个映射的多个尺度上揭示结构
- 揭示位于多个不同集群中的数据
- 降低在点在中心拥挤在一起的趋势

t-SNE 节点在 SPSS Modeler 中使用 Python 进行实现并且需要 `scikit-learn`© Python 库。 有关 t-SNE 和 scikit-learn 库的详细信息，请参阅：

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

节点选用板上的 Python 选项卡包含此节点和其他 Python 节点。 “图形”选项卡上还提供 t-SNE 节点。

[1] 引用：

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE 节点专家选项

根据想要针对 t-SNE 节点设置的选项，选择**简单**方式或**专家**方式。

可视化类型。 选择 **2D** 或 **3D** 以指定是将图像绘制为二维还是三维。

方法。 选择 **Barnes Hut** 或 **Exact**。 缺省情况下，梯度计算算法使用 Barnes-Hut 近似值，其运行速度必须大幅快于 Exact 方法。 Barnes-Hut 近似值允许将 t-SNE 技术应用于大型现实世界数据集。 Exact 算法在避免最近邻元素错误方面更好一些。

初始化。 对于嵌套初始化选择**随机**或 **PCA**。

目标字段。 选择目标字段以显示为输出图形上的颜色映射图。 如果此处未指定目标字段，那么图像将使用一种颜色。

优化

困惑度。 困惑度与其他各种学习算法中使用的最近邻元素数量相关。 通常，数据集越大，需要的困惑度越大。 请考虑选择 **5** 和 **50** 之间的值。 缺省值为 **30**，范围为 **2 - 9999999**。

早期夸大。 此设置控制原始空间中的自然聚类将在内嵌空间中的紧密程度以及两者之间的空间量。 缺省值为 **12**，范围为 **2 - 9999999**。

学习速率。 如果学习速率太高，那么数据可能看起来好像一个“球”，其中任意点与其最近的邻域大致等距。 如果学习速率太低，那么大多数点可能看起来好像压缩在一个具有极少离群值的密集云中。 如果成本函数陷入错误的局部最小值中，那么提高学习速率可能会有所帮助。 缺省值为 **200**，范围为 **0 - 9999999**。

最大迭代次数。 优化的最大迭代次数。 缺省值为 **1000**，范围为 **250 - 9999999**。

角度大小。 从一个点度量的远距离节点的角度大小。 输入 **0** 到 **1** 之间的值。 缺省值为 **0.5**。

随机种子(D)

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

优化停止条件

不含进度的最大迭代次数。 在停止优化之前要执行的不含进度的最大迭代次数，在 250 次包含早期夸大 (early exaggeration) 的初始迭代后将使用该迭代数。 请注意，每隔 50 次迭代后才会检查进度，因此该值舍入为 50 的下一个倍数。 缺省值为 **300**，范围为 **0 - 9999999**。

最小梯度标准值。 如果梯度标准值低于此最低阈值，那么优化将停止。 缺省值为 **1.0E-7**。

度量。 在计算特征数组中实例之间的距离时要使用的度量。 如果度量是字符串，那么它必须是 `scipy.spatial.distance.pdist` 针对其度量参数允许的选项之一，或是 `pairwise.PAIRWISE_DISTANCE_FUNCTIONS` 中列出的度量。 选择其中一个可用度量类型。 缺省值为 **euclidean**。

当记录数大于以下值时。 为大型数据集指定一种绘制方法。 可以指定数据集大小上限，或使用缺省的 2000 点。 如果选择 **分隔** 或 **抽样** 选项，则处理大数据集的性能将显著提高。 另外，您也可以选择 **使用所有数据**，但必须要注意，这一选项可能大幅降低软件的执行效率。

- **分级**。 选择此选项可对所包含记录数超过指定数字的数据集进行分级。 “分级”使图形在实际绘制前被分散在较小的网格中，并计算在每个单元格中将出现的连接数。 在最终图形中，每个网格中的分级矩心处将使用一个连接（该连接即代表分级中所有连接点位置的平均数）。
- **样本**。 选择此选项将随机抽取指定记录数的数据样本。

下表显示 SPSS Modeler t-SNE 节点对话框的“专家”选项卡上的设置与 Python t-SNE 库参数之间的关系。

表 *36:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**Python t-SNE 参数**
方式	mode_type	
可视化类型	n_components	n_components
Method	method	method
嵌套初始化	init	init
目标	target_field	target_field
困惑度	perplexity	perplexity
早期夸大	early_exaggeration	early_exaggeration
学习速率	learning_rate	learning_rate
最大迭代次数	n_iter	n_iter
角度大小	angle	angle
设置随机种子	enable_random_seed	
随机种子(D)	random_seed	random_state
不含进度的最大迭代次数	n_iter_without_progress	n_iter_without_progress
最小梯度标准值	min_grad_norm	min_grad_norm
使用多个困惑度执行 t-SNE	isGridSearch	

t-SNE 节点输出选项

在**输出**选项卡上指定 t-SNE 节点输出的选项。

输出名称。 指定在节点运行时生成的输出的名称。 如果选择**自动**，那么将自动设置输出的名称。

输出到屏幕。 选择此选项以在新窗口中生成并显示输出。 这还会将输出添加到输出管理器。

输出到文件。 选择此选项可将输出保存到文件。 执行此操作将启用**文件名**和**文件类型**字段。 如果要使用其他字段创建绘图以进行比较，或者要使用输出文件的输出作为分类或回归模型中的预测变量，那么 t-SNE 节点需要此输出文件的访问权。 t-SNE 模型创建 x、y（和 z）坐标字段的结果文件，使用“固定文件”源节点可非常轻松地对其进行访问。

t-SNE 模型块

t-SNE 模型块包含 t-SNE 模型捕获的所有信息。 以下选项卡可用。

图形

图形选项卡显示 t-SNE 节点的图表输出。 Pyplot 分布图表显示低纬度结果。 如果未在 t-SNE 节点的专家选项卡上选择**使用多个困惑度执行 t-SNE** 选项，那么仅包含一个图形，而不是使用不同困惑度的六个图形。

文本输出

文本输出选项卡显示 t-SNE 算法的结果。 如果在 t-SNE 节点的专家选项卡上选择 **2D** 可视化类型，那么此处的结果是两个维度中的点值。 如果选择 **3D**，那么结果是三个维度中的点值。

高斯混合节点

Gaussian Mixture© 模型是一个概率模型，其假定从有限数量的高斯分布和未知参数混合中生成所有数据点。 可以将混合模型认为是广义 K-Means 聚类以包含有关数据的协方差结构以及潜伏高斯分布的中心的信息。[1]

SPSS Modeler 中的 Gaussian Mixture 节点公开 Gaussian Mixture 库的核心特征和常用参数。 此节点使用 Python 进行实现。

有关高斯混合建模算法和参数的更多信息，请参阅以下位置的高斯混合文档：http://scikit-learn.org/stable/modules/mixture.html 和 https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html。[2]

[1] “User Guide”。 *Gaussian mixture models*。 Web. © 2007 - 2017. scikit-learn developers.

[2] Scikit-learn: Python 中的机器学习，Pedregosa 等， JMLR 12，第 2825-2830 页，2011 年。

高斯混合节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的输入设置。

使用定制字段分配。 要手动分配输入，请选择此选项。

字段。 使用方向按钮可以将此列表中的项手动分配到屏幕右侧的“预测变量”列表。 图标指示每个字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

预测变量。 选择一个或多个字段作为预测变量。

高斯混合节点构建选项

使用“构建选项”选项卡可以指定高斯混合节点的构建选项，包括**基本选项**和**高级选项**。 有关此部分中未涉及的这些选项的详细信息，请参阅以下联机资源：

- 高斯混合参数参考 [1]
- 高斯混合节点用户指南 [2]

基本

协方差类型。 选择下列其中一个协方差矩阵：

- **完整**。 每个组件具有其自己的一般协方差矩阵。
- **绑定**。 所有组件共享相同的一般协方差矩阵。
- **对角**。 每个组件具有其自己的对角线协方差矩阵。
- **球面**。 每个组件具有其自己的单个方差。

组件数。 指定在构建模型时要使用的混合组件的数量。

聚类标签。 指定聚类标签是数值还是字符串。 如果您选择**字符串**，请指定聚类标签的前缀（例如，缺省前缀为 `cluster`，这将生成 **cluster-1** 和 **cluster-2** 等聚类标签）。

随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

高级

容差。 指定收敛阈值。 缺省值为 **0.001**。

迭代数。 指定要执行的最大迭代次数。 缺省值为 **100**。

初始化参数。 选择初始化参数 **Kmeans**（表示使用 k-means 进行初始化）或**随机**（表示随机初始化）。

热启动。 如果选择 **True**，那么将使用最新拟合的解作为下一个拟合的初始化。 在针对类似问题多次调用拟合时，这可加速收敛。

下表显示 SPSS Modeler 高斯混合节点对话框中的设置与 Python 高斯混合库参数之间的关系。

表 *37:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	脚本名称（属性名称）	高斯混合参数
使用预定义角色/使用定制字段分配	role_use	
输入	predictors	
使用分区数据	use_partition	
协方差类型	covariance_type	covariance_type
组件数	number_component	n_components
聚类标签	component_lable	
标签前缀	label_prefix	
设置随机种子	enable_random_seed	
随机种子	random_seed	random_state
容差	tol	tol
迭代数	max_iter	max_iter
初始化参数	init_params	init_params
热启动	warm_start	warm_start

[1] Scikit-learn: Python 中的机器学习，Pedregosa 等， JMLR 12，第 2825-2830 页，2011 年。

[2] “User Guide”。 *Gaussian mixture models*。 Web. © 2007 - 2017. scikit-learn developers.

高斯混合节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

KDE 节点

Kernel Density Estimation (KDE)© 使用 Ball Tree 或 KD Tree 算法以进行效率查询，并且游走于无监督学习、特征工程和数据建模。 基于相邻元素的方法（例如，KDE）是最流行且最有用的一些密度估算方法。 可在任意数量的维度执行 KDE，但是在实践当中，高维数可能导致性能下降。 SPSS Modeler 中的 KDE 建模和 KDE 模拟节点公开 KDE 库的核心特征和常用参数。 节点使用 Python 进行实现。 [1]

要使用 KDE 节点，必须设置上游“类型”节点。 KDE 节点将从“类型”节点（或者上游源节点的“类型”选项卡）读取输入值。

KDE 建模节点位于 SPSS Modeler 的“建模”选项卡和 Python 选项卡上。 “KDE 建模”节点生成一个模型块，并且块的评分值是来自输入数据的核心密度值。

KDE 模拟节点位于“输出”选项卡和 Python 选项卡上。 “KDE 模拟”节点生成 KDE Gen 源节点，后者可创建一些使用相同分布作为输入数据的记录。 KDE Gen 节点包含“设置”选项卡，可在其中指定节点将创建的记录数（缺省值为 1）并生成随机种子。

有关 KDE 的更多信息，包括示例，请参阅 KDE 文档 (http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation)。 [1]

[1] “User Guide”。 *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

KDE 建模节点和 KDE 模拟节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的输入设置。

使用定制字段分配。 要手动分配输入，请选择此选项。

字段。 使用方向按钮可以将此列表中的项手动分配到屏幕右侧的“输入”列表。 图标指示每个字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

输入。 选择一个或多个字段作为聚类输入。 KDE 只能处理连续字段。

KDE 节点构建选项

使用“构建选项”选项卡以指定 KDE 节点的构建选项，包括用于内核密度参数和集群标签的**基本选项**，以及**高级选项**，例如，容差、叶大小以及是否使用广度优先方法。 有关这些选项的更多信息，请参阅以下在线资源：

- 内核密度估算 Python API 参数参考 [1]
- 内核密度估算用户指南 [2]

基本

带宽。 指定内核的带宽。

内核。 选择要使用的内核。 KDE 建模节点的可用内核为 **Gaussian**、**Tophat**、**Epanechnikov**、**Eponential**、**Linear** 或 **Cosine**。 KDE 模拟节点的可用内核为 **Gaussian** 或 **Tophat**。 有关这些可用内核的详细信息，请参阅内核密度估算用户指南。[2]

算法。 对于要使用的树算法，选择 **Auto**、**Ball Tree** 或 **KD Tree**。 有关更多信息，请参阅 Ball Tree[3] 和 KD Tree。[4]

度量。 选择距离度量。 可用度量为 **Euclidean**、**Braycurtis**、**Chebyshev**、**Canberra**、**Cityblock**、**Dice**、**Hamming**、**Infinity**、**Jaccard**、**L1**、**L2**、**Matching**、**Manhattan**、**P**、**Rogerstanimoto**、**Russellrao**、**Sokalmichener**、**Sokalsneath**、**Kulsinski** 或 **Minkowski**。 如果选择 **Minkowski**，那么根据需要设置 **P 值**。

此下拉列表中可用的度量将根据选择的算法的不同而不同。 另外，请注意，密度输出的标准化仅针对 Euclidean 距离度量正确。

高级

绝对容差。 指定期望的结果的绝对容差。 较大的容差通常将导致更快的运行时间。 缺省值为 **0.0**。

相对容差。 指定期望的结果的相对容差。 较大的容差通常将导致更快的运行时间。 缺省值为 **1E-8**。

叶大小。 指定底层树的叶大小。 缺省值为 **40**。 更改叶大小可能会显著影响性能和所需的内存。 有关 Ball Tree 和 KD Tree 算法的更多信息，请参阅 Ball Tree[3] 和 KD Tree。[4]

广度优先。 如果想要使用广度优先方法，那么选择 **True**，或者选择 **False** 以使用深度优先方法。

下表显示 SPSS Modeler KDE 节点对话框中的设置与 Python KDE 库参数之间的关系。

表 *38:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**KDE 参数**
输入	`inputs`	

表 *38:* 映射到 *Python* 库参数的节点属性 *(*继续*)*		
SPSS Modeler 设置	**脚本名称（属性名称）**	**KDE 参数**
带宽(B)	bandwidth	bandwidth
内核(K)	kernel	kernel
算法	algorithm	algorithm
指标	metric	metric
P 值	pValue	pValue
绝对容差	atol	atol
相对容差	rtol	Rtol
叶大小	leafSize	leafSize
广度优先	breadthFirst	breadthFirst

[1] "API Reference." *sklearn.neighbors.KernelDensity*。 Web. © 2007-2018, scikit-learn developers.

[2] “User Guide”。 *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

[3] “Ball Tree”。 *Five balltree construction algorithms*。 © 1989, Omohundro, S.M.， 国际计算机科学研究所技术报告。

[4] “K-D Tree”。 *Multidimensional binary search trees used for associative searching*。 © 1975, Bentley, J.L.， ACM 的通信。

KDE 建模节点和 KDE 模拟节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

随机森林节点

随机森林 (Random Forest©) 是将树模型用作基本模型的组装算法的高级实现。 在随机森林中，整体中的每个树都是从使用训练集合的替换（例如，bootstrap 样本）绘制的样本构建的。 在树构造期间分割节点时，选中的分割不再是所有特征之间的最佳分割。 相反，选取的分割是特征的随机子集中的最佳分割。 由于此随机性，森林偏差通常略微增加（针对单个非随机树的偏差），但是由于平均化，其方差会降低，通常超过偏差增加的补偿，因此生成整体更优秀的模型。[1]

SPSS Modeler 中的随机森林节点使用 Python 进行实现。 节点选用板上的 Python 选项卡包含此节点和其他 Python 节点。

有关随机森林算法的更多信息，请参阅 https://scikit-learn.org/stable/modules/ensemble.html#forest。

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

随机森林节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标和预测变量，请选择此选项。

字段。 使用方向按钮可以将项目从列表中手动分配给屏幕右侧的“目标”和“预测变量角色”字段。 图标表示每个角色字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择要用作预测的目标的字段。

预测变量。 选择一个或多个字段作为预测输入。

随机森林节点构建选项

使用“构建选项”选项卡可以指定随机森林节点的构建选项，包括**基本选项**和**高级选项**。 有关这些选项的更多信息，请参阅 https://scikit-learn.org/stable/modules/ensemble.html#forest

基本

要构建的树数量。 选择森林中树的数量。

指定最大深度。 如果未选择，那么将展开节点直至所有叶片均纯净或者直至所有叶片包含的样本书小于 `min_samples_split`。

最大深度。 树的最大深度。

最小叶节点大小。 需要位于一个叶节点上的样本的最小数量。

用于拆分的特征数。 在查找最佳分割时要考虑的特征数目。

- 如果为 `auto`，那么对于分类器为 `max_features=sqrt(n_features)` 且对于回归为 `max_features=n_features`。
- 如果为 `sqrt`，那么为 `max_features=sqrt(n_features)`。
- 如果为 `log2`，那么为 `max_features=log2(n_features)`。

高级

在构建树时，使用 Bootstrap 样本。 如果选中，那么在构建树时使用 bootstramp 样本。

使用袋外样本来估算泛化关系准确性。 如果选中，那么将使用袋外样本来估算泛化关系准确性。

使用仅限随机树。 如果选中，那么将使用极限随机树代替常规随机森林。 在极限随机树中，在计算分割时，随机性更进一步。 在随机森林中，将使用一组随机的候选特征子集，但是不查找差异性最大的阈值，针对每个候选特征随机绘制阈值，并且将挑选这些随机生成的阈值中的最佳项作为分割规则。 这通常可使模型方差降低一点，代价是偏差略微增加。 [1]

复制结果。 如果选中，那么会复制模型构建过程以实现相同的评分结果。

随机种子值。 您可以单击**生成**来生成随机数字生成器所使用的种子。

超参数优化（基于 Rbfopt）。 选择此选项以启用基于 Rbfopt 的超参数优化，这将自动发现最佳参数组合，从而使模型在样本上实现期望或更低的错误率。 有关 Rbfopt 的详细信息，请参阅 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

目标。 您想要实现的目标函数值（基于样本的模型的错误率）（例如，未知最佳值）。 设置为可接受的值，例如，`0.01`。

最大迭代次数。 用于尝试模型的最大迭代次数。 缺省值为 `1000`。

最大求值次数。 以精确模式尝试模型的功能评估的最大次数。 缺省值为 `300`。

下表显示 SPSS Modeler 随机森林节点对话框中的设置与 Python 随机森林库参数之间的关系。

表 *39:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**随机森林参数**
目标	`target`	
预测变量	`inputs`	
要构建的树数量	`n_estimators`	`n_estimators`
指定最大深度	`specify_max_depth`	`specify_max_depth`
最大深度	`max_depth`	`max_depth`

表 *39:* 映射到 *Python* 库参数的节点属性 *(*继续*)*		
SPSS Modeler 设置	**脚本名称（属性名称）**	**随机森林参数**
最小叶节点大小	min_samples_leaf	min_samples_leaf
用于分割的特征数量	max_features	max_features
在构建树时，使用 Bootstrap 样本	bootstrap	bootstrap
使用袋外样本来估算泛化关系准确性	oob_score	oob_score
使用仅限随机树	extreme	
复制结果(C)	use_random_seed	
随机种子(D)	random_seed	random_seed
超参数优化（基于 Rbfopt）	enable_hpo	
目标（用于 HPO）	target_objval	
最大迭代次数（用于 HPO）	max_iterations	
最大评估次数（用于 HPO）	max_evaluations	

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

随机森林节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

随机森林模型块

随机森林模型块包含随机森林模型捕获的所有信息。 以下部分可用。

模型信息

此视图提供有关模型的关键信息，包括输入字段、独热编码值和模型参数。

预测变量重要性

此视图显示一个图表，以指示在估计模型时所使用的各个预测变量的相对重要性。 有关更多信息，请参阅第 32 页的『预测变量重要性』。

HDBSCAN 节点

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© 使用非监督学习来查找数据集的聚类或密集区域。 SPSS Modeler 中的 HDBSCAN 节点公开 HDBSCAN 库的核心特征和常用参数。 此节点以 Python 实现，当您一开始不了解数据集的分组时，可以使用此节点将数据集聚类为不同的组。 与 SPSS Modeler 中的大多数学习方法不同，HDBSCAN 模型不使用目标字段。 这种没有目标字段的学习称为 无监督学习 。 HDBSCAN 试图揭示输入字段集的模式，而不是预测结果。 记录会进行分组，以使一个组或聚类中的记录彼此相似，而不同组中的记录则互不相同。 HDBSCAN 算法将聚类视为高密度区域，这些区域之间由低密度区域分隔。 因此，HDBSCAN 所发现的聚类可以是任意形状，这与假定聚类为凸面形的 K 均值算法完全不同。 对于单独位于低密度区域中的离群点，也会予以标记。 HDBSCAN 还支持对新样本进行评分。[1]

要使用 HDBSCAN 节点，必须设置上游“类型”节点。 HDBSCAN 节点将从“类型”节点（或者上游源节点的“类型”选项卡）读取输入值。

有关 HDBSCAN 聚类算法的更多信息，请参阅 HDBSCAN 文档 (http://hdbscan.readthedocs.io/en/latest/)。[1]

[1] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN 节点字段

“字段”选项卡指定要在分析中使用的字段。

要点: 要训练 HDBSCAN 模型，您必须使用角色设置为**输入**的一个或多个字段。 角色设置为**输出**、**两者**或**无**的字段将被忽略。

使用预定义角色。 此选项使用上游“类型”节点（或上游源节点的“类型”选项卡）中的输入设置。

使用定制字段分配。 要手动分配输入，请选择此选项。

字段。 使用方向按钮可以将此列表中的项手动分配到屏幕右侧的“输入”列表。 图标指示每个字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

输入。 选择一个或多个字段作为聚类输入。

HDBSCAN 节点构建选项

使用“构建选项”选项卡可以指定 HDBSCAN 节点的构建选项，包括聚类参数和聚类标签的**基本选项**，以及高级参数和图表输出选项的**高级选项**。 有关这些选项的更多信息，请参阅以下在线资源：

- HDBSCAN Python API 参数参考 [1]
- HDBSCAN 主页 [2]

基本

超参数优化（基于 Rbfopt）。 选择此选项以启用基于 Rbfopt 的超参数优化，这将自动发现最佳参数组合，从而使模型在样本上实现期望或更低的错误率。 有关 Rbfopt 的详细信息，请参阅 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

最小聚类大小。 指定聚类的最小大小。 单联结分割中包含的点数少于此处指定的值时，将会认为点落在聚类之外，而不是认为一个聚类分割为两个新聚类。

最小样本数。 指定将一个点视为核心点之前，近邻中的最小样本数。 如果设置为 **0**，那么缺省值为最小聚类大小值。

算法。 选择要使用的算法。 HDBSCAN 具有专用于数据不同特征的变体。 缺省情况下，将会使用 **BEST**，这会根据数据的性质，自动选择最佳的算法。 有关这些算法类型的详细信息，请参阅 HDBSCAN 文档。[1] 请注意，您选择的算法将会影响性能。 例如，对于大型数据，我们建议尝试使用 Boruvka KDTree 或 Boruvka BallTree。

距离度量。 选择在计算特征数组中实例之间的距离时要使用的度量。

聚类标签。 指定聚类标签是数值还是字符串。 如果您选择**字符串**，请指定聚类标签的前缀（例如，缺省前缀为 `cluster`，这将生成 **cluster-1** 和 **cluster-2** 等聚类标签）。

高级

近似最小生成树。 如果要接受近似最小生成树，请选择 **True**。 对于某些算法，这可以提高性能，但是生成的聚类质量可能会稍有下降。 如果您愿意牺牲速度来换取正确性，您可尝试 **False** 选项。 在大多数情况下，建议选择 **True**。

聚类选择方法。 选择要使用何种方法从密集树中选择聚类。 对于 HDBSCAN，标准方法是使用“质量过剩”(**EOM**) 算法来查找最持久的聚类。 或者，您可选择位于树叶的聚类，这样会提供最细颗粒度的同质聚类。

接受单个聚类。 将此设置更改为 **True** 表示允许单聚类结果（前提为这是数据集的有效结果）。

P 值。 使用闵可夫斯基距离度量时（在**基本**构建选项下），如果您愿意的话，可以更改这个 p 值。

叶大小。 使用空间树算法（Boruvka KDTree 或 Boruvka BallTree）时，这是树叶节点中的点数。 此设置对生成的聚类没有影响，但可能会影响算法的运行时间。

有效性指标。 选择此选项可将“有效性指标”图表包括在模型块输出中。

密集树。 选择此选项可将“密集树”图表包括在模型块输出中。

单联结树。 选择此选项可将“单联结树”图表包括在模型块输出中。

最小生成树。 选择此选项可将“最小生成树”图表包括在模型块输出中。

下表显示 SPSS Modeler HDBSCAN 节点对话框中的设置与 Python HDBSCAN 库参数之间的关系。

表 *40:* 映射到 *Python* 库参数的节点属性

SPSS Modeler 设置	脚本名称（属性名称）	HDBSCAN 参数
输入	inputs	inputs
超参数优化	useHPO	
最小聚类大小	min_cluster_size	min_cluster_size
最小样本数	min_samples	min_samples
算法	algorithm	algorithm
距离度量	metric	metric
聚类标签	useStringLabel	
标签前缀	stringLabelPrefix	
近似最小生成树	approx_min_span_tree	approx_min_span_tree
选择聚类的方法	cluster_selection_method	cluster_selection_method
接受单个聚类	allow_single_cluster	allow_single_cluster
P 值	p_value	p_value
叶大小	leaf_size	leaf_size
有效性指标	outputValidity	
密集树	outputCondensed	
单联结树	outputSingleLinkage	
最小生成树	outputMinSpan	

[1] "API Reference." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

[2] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

单类 SVM 节点

单类 SVM© 节点使用无监督学习算法。 此节点可用于新内容检测。 它将检测指定样本集的软边界，以便按是否属于该集合对新点进行分类。 此单类 SVM 建模节点使用 Python 进行实现并且需要 `scikit-learn`© Python 库。 有关 scikit-learn 库的详细信息，请参阅 http://contrib.scikit-learn.org/imbalanced-learn/about.html[1]。

节点选用板上的 Python 选项卡包含单类 SVM 节点和其他 Python 节点。

注: 单类 SVM 用于无监督的离群值和新内容检测。 在大多数情况下，我们建议使用已知的“正常”数据集来构建模型，以使算法可以为指定样本设置正确的边界。 模型的参数（例如，nu、伽玛和内核）会显著影响结果。 因此，您可能需要试验这些选项，直到找到适合于您的情况的最佳设置为止。

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression." *Statistics and Computing Archive*，第 14 卷，第 3 期，2004 年 8 月 3 日，第 199-222 页。 (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

单类 SVM 节点的“字段”选项卡

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 选择此选项以选择具有已定义的“输入”角色的所有字段。

使用定制字段分配。 要手动选择字段，请选择此选项并选择输入字段和分割字段：

输入。 选择要在分析中使用的输入字段。 除了无类型或未知以外，支持所有存储类型和测量类型。 如果某个字段具有“字符串”存储类型，那么将通过独热编码算法以单个对全部的方式对此字段的值进行二进制化。

分割。 选择要用作分割字段的一个或多个字段。 支持所有“标志”、“名义”、“有序”和“独立”测量类型。

使用分区数据 如果定义了分区字段，那么此选项用于确保仅来自训练分区的数据用于构建模型。

单类 SVM 节点的“专家”选项卡

在单类 SVM 节点的“专家”选项卡上，您可以从**简单**模式或**专家**模式中进行选择。 如果您选择**简单**，那么将使用缺省值设置所有参数，如下所示。 如果您选择**专家**，那么可以为这些参数指定定制值。 有关这些选项的详细信息，请参阅 http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM。

停止条件。 指定中止条件的容差。 缺省值为 **1.0E-3** (0.001)。

回归精确度 (nu)。 训练错误和支持向量的尾数边界。 缺省值为 **0.1**。

内核类型。 要用在算法中的内核类型。 选项包括 **RBF**、**多项式**、**Sigmoid**、**线性**或**预先计算**。 缺省值为 **RBF**。

指定伽玛。 选择此选项以指定伽玛。 否则，将应用自动伽玛。

伽玛。 伽玛设置仅适用于 RBF、多项式和 Sigmoid 内核类型。

Coef0. Coef0 仅适用于多项式和 Sigmoid 内核类型。

度。 度仅适用于多项式内核类型。

使用缩小启发式。 选择此选项以使用缩小启发式。 缺省情况下，未选中此选项。

指定内核高速缓存的大小 (MB)。 选择此选项以指定内核缓存的大小。 缺省情况下，未选中此选项。 选中此选项时，缺省值为 **200** MB。

超参数优化（基于 Rbfopt）。 选择此选项以启用基于 Rbfopt 的超参数优化，这将自动发现最佳参数组合，从而使模型在样本上实现期望或更低的错误率。 有关 Rbfopt 的详细信息，请参阅 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html。

目标。 我们想要实现的目标函数值（基于样本的模型的错误率）（例如，未知最佳值）。 设置为可接受的值，例如，**0.01**。

最大迭代次数。 尝试模型的最大迭代次数。 缺省值为 **1000**。

最大求值次数。 尝试模型的函数求值的最大次数，其中焦点是速度准确性。 缺省值为 **300**。

单类 SVM 节点需要 `scikit-learn`© Python 库。 下表显示 SPSS Modeler SMOTE 节点对话框中的设置和 Python 算法之间的关系。

表 *41:* 映射到 *Python* 库参数的节点属性		
参数名称	**脚本名称（属性名称）**	**Python API 的参数名称**
停止条件	stopping_criteria	tol
回归精度	precision	nu
内核类型	kernel	kernel
伽玛	gamma	gamma
Coef0	coef0	coef0
硕士	degree	degree
使用缩小启发式	shrinking	shrinking
指定内核缓存的大小（数字输入框）	cache_size	cache_size
随机种子(D)	random_seed	random_state

单类 SVM 节点选项

在单类 SVM 节点的选项卡“选项”上，您可以设置以下选项。

平行坐标图形的类型。 SPSS Modeler 可以绘制平行坐标图形来表示已构建的模型。 有时，显示一些数据列/功能的值时它们远大于其他值，这会导致难以看到图形的一些其他部分。 对于 这种情况，您可以选择**独立纵轴**选项为所有纵轴提供独立的轴刻度，也可以选择**一般纵轴**强制所有纵轴共享同一轴刻度。

图形上的最大行数。 指定要在图形输出中显示的数据行的最大数目。 缺省值为 100。 为了提高性能，最多将显示 20 个字段。

在图形上绘制所有输入字段。 选择此选项可以在图形输出中显示所有输入字段。 缺省情况下，会将每个数据字段绘制为一条纵轴。 为了提高性能，最多将显示 30 个字段。

要在图形上绘制的定制字段。 您可以选择此选项并选择要显示的一部分字段，而不是在图形输出中显示所有输入字段。 这可以提高性能。 为了提高性能，最多将显示 20 个字段。

第 18 章 Spark 节点

SPSS Modeler 提供了用于使用 Spark 本机算法的节点。 节点选用板上的 **Spark** 选项卡包含您可用于运行 Spark 算法的下列节点。 这些节点在 Windows 64、Mac 64 和 Linux 64 上受支持。 请注意，这些节点不支持构建模型时指定整数/双精度列作为标志/名义列。 因此，必须将列值转换为 0/1 或 0、1、2、3、4...

保序回归属于回归算法系列。 SPSS Modeler 中的 Isotonic-AS 节点使用 Spark 进行实现。 有关保序回归算法的详细信息，请参阅 https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html。

XGBoost© 是实现提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost-AS 节点仅显示了核心功能和常用参数。 在 Spark 中实现 XGBoost-AS 节点。

K-Means 是最常用的聚类算法之一。 它将数据点聚集成多个预定义聚类。 SPSS Modeler 中的 K-Means-AS 节点使用 Spark 进行实现。 有关 K-Means 算法的详细信息，请参阅 https://spark.apache.org/docs/2.2.0/ml-clustering.html。 请注意，K-Means-AS 节点自动对分类变量执行独热编码。

多层感知器是基于前馈人工神经网络的分类器，其由多个层组成。 每层完全连接到下一层络。 SPSS Modeler 中的 MultiLayerPerceptron-AS 节点使用 Spark 进行实现。 有关多层感知器分类器 (MLPC) 的详细信息，请参阅 https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier。

Isotonic-AS 节点

保序回归属于回归算法系列。 SPSS Modeler 中的 Isotonic-AS 节点使用 Spark 进行实现。

有关保序回归算法的详细信息，请参阅 https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html。 [1]

[1] “Regression - RDD-based API。” *Apache Spark*. MLlib: Main Guide. Web. 2017 年 10 月 3 日。

Isotonic-AS 节点字段

“字段”选项卡指定要在分析中使用的字段。

字段。 列出数据源中的所有字段。 使用方向按钮可以将项目从此列表手动分配到屏幕右侧的“目标”、“输入”和“权重”字段。 图标表示每个角色字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择要用作目标的字段。

输入。 选择输入字段或字段。

权重。 选择表示指数权重的权重字段。 如果未设置，那么将使用缺省权重值 **1**。

Isotonic-AS 节点构建选项

使用“构建选项”选项卡可以指定 Isotonic-AS 节点的构建选项，包括功能索引和保序类型。 有关更多信息，请参阅 http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.html。 [1]

输入字段索引。 指定输入字段的索引。 缺省值为 **0**。

保序类型。 此设置确定输出序列应是保序/增大还是反序/减小。 缺省值为**保序**。

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc。 Web. 2017 年 10 月 3 日。

Isotonic-AS 模型块

Isotonic-AS 模型块包含保序回归模型捕获的所有信息。 以下部分可用。

模型摘要

此视图提供有关模型的关键信息，包括输入字段、目标字段和模型构建选项。

模型图

此视图显示分布图。

XGBoost-AS 节点

XGBoost© 是实现提升算法的高级实现。 提升算法以迭代方式学习弱分类器，然后将它们添加到最终的强分类器中。 XGBoost 具有很高的灵活性，并提供了很多对于大多数用户来说过于复杂的参数，因此 SPSS Modeler 中的 XGBoost-AS 节点仅显示了核心功能和常用参数。 在 Spark 中实现 XGBoost-AS 节点。

有关提升算法的更多信息，请参阅 XGBoost 教程，网址为 http://xgboost.readthedocs.io/en/latest/tutorials/index.html。 [1]

请注意，SPSS Modeler 中不支持 XGBoost 交叉验证功能。 您可以将 SPSS Modeler 分区节点用于此功能。另外，请注意，XGBoost 在 SPSS Modeler 中用于自动对分类变量执行独热编码。

注: 在 Mac 上，构建 XGBoost-AS 模型需要 V10.12.3 或更高版本。

[1] “XGBoost Tutorials。” *Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

XGBoost-AS 节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。

使用定制字段分配。 要手动分配目标和预测变量，请选择此选项。

字段。 使用方向按钮可以将项目从列表中手动分配给屏幕右侧的“目标”和“预测变量角色”字段。 图标表示每个角色字段的有效测量级别。 要选中列表中的所有字段，请单击**全部**按钮，或者单击个别的测量级别按钮以选中该测量级别的所有字段。

目标。 选择要用作预测的目标的字段。

预测变量。 选择一个或多个字段作为预测输入。

XGBoost-AS 节点构建选项

使用“构建选项”选项卡以指定 XGBoost-AS 节点的构建选项，包括用于模型构建和处理不平衡的数据集的**常规选项**，用于目标和评估度量的**学习任务选项**，以及用于特定提升系数的**提升系数参数**。 有关这些选项的更多信息，请参阅以下在线资源：

- XGBoost 主页 [1]
- XGBoost 参数引用 [2]
- XGBoost Spark API[3]

常规

工作程序数。 用于训练 XGBoost 模型的工作程序数量。

线程数。 每个工作程序使用的线程数。

使用外部内存。 是否使用外部内存作为缓存。

提升系数类型。 要使用的提升系数（**gbtree**、**gblinear** 或 **dart**）。

提升轮数。 用于提升的舍入次数。

标度位置权重。 此设置控制正权重和负权重的平衡，并且用于不平衡类。

随机种子值。 单击**生成**可以生成随机数字生成器所使用的种子。

学习任务

目标。 请从以下学习任务目标类型中进行选择：**reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**rank:pairwise**、**binary:logistic** 或 **multi**。

评估度量。 用于验证数据的评估度量。 将根据目标指定缺省度量（**rmse** 用于回归，**error** 用于分类，或者 **mean average precision** 用于排名）。 可用选项为 **rmse**、**mae**、**logloss**、**error**、**merror**、**mlogloss**、**uac**、**ndcg**、**map** 或 **gamma-deviance**（缺省值为 **rmse**）。

提升系数参数

Lambda。 这是有关权重的 L2 规则化术语。 增大此值将使模型更保守。

Alpha。 这是有关权重的 L1 规则化术语。增大此值将使模型更保守。

Lambda 偏差。 这是有关基本选项的 L2 规则化术语。 （没有关于偏差的 L1 规则化术语，因为它不重要。）

树方法。 选择要使用的 XGBoost Tree 构建算法。

最大深度。 指定树的最大深度。 增大此值将导致模型更复杂，并且很可能出现过度拟合。

最小子代权重。 指定子代中需要的实例权重 (hessian) 的最小总和。 如果树分区步骤生成实例权重总和小于此**最小子代权重**的叶节点，那么构建过程将停止进行进一步分区。 在线性回归模式下，此项简单地对应于每个节点中所需的最小实例数。 权重越大，算法越保守。

最大增量步骤。 指定允许用于每个树的权重估计的最大增量步骤。 如果设置为 **0**，那么没有约束。 如果设置为正值，那么它可以使更新步骤更为保守。 通常不需要此参数，但是在某个类极度不平衡的情况下，它可以用在 Logistic 回归中。

子样本。 子样本是训练实例的比率。 例如，如果您将此项设置为 **0.5**，那么 XGBoost 将随机收集一半的数据实例以生成树，并且这将防止过度拟合。

Eta。 这是更新步骤期间用于防止过度拟合的步长收缩。 在每个提升步骤后，可以直接获取新功能的权重。 Eta 也会缩小功能权重，以使提升过程更保守。

伽玛。 这是对树的某个叶节点进行进一步分区所需的最小损失减小。 伽玛设置越大，算法越保守。

按树进行列采样。 这是构建每个树时列的子样本比率。

按级别进行列采样。 这是在每个级别每个分割的列的子样本比率。

标准化算法。 在“常规”选项下选择 dart 提升系数类型时要使用的标准化算法。 可用选项为 **tree** 或 **forest**（缺省值为 **tree**）。

采样算法 在“常规”选项下选择 dart 提升系数类型时要使用的采样算法。 **均匀**算法均匀地选择已删除的树。 **加权**算法按权重比例选择已删除的树。 缺省值为**均匀**。

丢码率。 在“常规”选项下选择 dart 提升系数类型时要使用的丢码率。

跳过丢码的可能性。 在“常规”选项下选择 dart 提升系数类型时要使用的跳过丢码可能性。 如果跳过丢码，那么将按照与 **gbtree** 相同的方式添加新树。

下表显示 SPSS Modeler XGBoost-AS 节点对话框中的设置与 XGBoost Spark 参数之间的关系。

表 *42:* 映射到 *Spark* 参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**XGBoost Spark 参数**
目标	target_fields	
预测变量	input_fields	
Lambda	lambda	lambda
工作程序数	nWorkers	nWorkers
线程数	numThreadPerTask	numThreadPerTask
使用外部内存	useExternalMemory	useExternalMemory
提升系数类型	boosterType	boosterType
提升舍入次数	numBoostRound	round
刻度位置权重	scalePosWeight	scalePosWeight
目标	objectiveType	objective
评估度量	evalMetric	evalMetric
Lambda	lambda	lambda
Alpha	alpha	alpha
Lambda 偏差	lambdaBias	lambdaBias
树方法	treeMethod	treeMethod
最大深度	maxDepth	maxDepth
最小子代权重	minChildWeight	minChildWeight
最大增量步骤	maxDeltaStep	maxDeltaStep
子样本	sampleSize	sampleSize
Eta	eta	eta
伽玛	gamma	gamma
列样本（按树列出）	colsSampleRation	colSampleByTree
列样本（按级别列出）	colsSampleLevel	colsSampleLevel
标准化算法	normalizeType	normalizeType
采样算法	sampleType	sampleType
丢码率	rateDrop	rateDrop
跳过丢码的可能性	skipDrop	skipDrop

[1]“Scalable and Flexible Gradient Boosting”。 Web. © 2015-2016 DMLC。

[2]“XGBoost Parameters”*Scalable and Flexible Gradient Boosting*。 Web. © 2015-2016 DMLC。

[3]“ml.dmlc.xgboost4j.scala.spark Params”。 *DMLC for Scalable and Reliable Machine Learning*。 Web. 2017 年 10 月 3 日。

XGBoost-AS 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

K-Means-AS 节点

K-Means 是最常用的聚类算法之一。 它将数据点聚类到预定义数量的 聚类中。[1] SPSS Modeler 中的 K-Means-AS 节点使用 Spark 进行实现。

有关 K-Means 算法的详细信息，请参阅 https://spark.apache.org/docs/2.2.0/ml-clustering.html。

请注意，K-Means-AS 节点自动对分类变量执行独热编码。

[1] “聚类。” *Apache Spark*. MLlib: Main Guide. Web. 2017 年 10 月 3 日。

K-Means-AS 节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 该选项通知节点使用来自上游类型节点的字段信息。 缺省情况下，此选项处于选定状态。

使用定制字段分配。 如果要手动分配输入字段，请选择此选项，然后选择一个或多个输入字段。 使用此选项类似于在“类型”节点中将字段角色设置为**输入**。

K-Means-AS 节点构建选项

使用“构建选项”选项卡可以指定 K-Means-AS 节点的构建选项，包括用于模型构建的常规选项、用于初始化聚类中心的初始化选项以及用于计算迭代和随机种子的高级选项。 有关更多信息，请参阅 JavaDoc for K-Means on SparkML。[1]

常规

模型名称。 对特定聚类评分后生成的字段的名称。 选择**自动**（缺省）或选择**定制**并输入名称。

聚类数。 指定要生成的聚类数。 缺省值为 **5**，最小值为 **2**。

初始化

初始化方式。 指定用于初始化聚类中心的方法。 缺省值为 **K-Means||**。 有关这两种方法的详细信息，请参阅可扩展 K-Means + +。[2]

初始化步骤。 如果选择 **K-Means||** 初始化模式，请指定初始化步骤数。 缺省值为 **2**。

高级

高级设置。 如果要按如下设置高级选项，请选择此选项。

最大迭代次数。 指定在搜索聚类中心时要执行的最大迭代次数。 缺省值为 **20**。

容差。 指定迭代算法的汇合容差。 缺省值为 **1.0E-4**。

设置随机种子值。 选择此信息并单击**生成**可以生成由随机数字生成器使用的种子。

显示

显示图形。 如果要在输出中包含图形，请选择此选项。

下表显示 SPSS Modeler K-Means-AS 节点中的设置与 K-Means Spark 参数之间的关系。

表 *43:* 映射到 *Spark* 参数的节点属性		
SPSS Modeler 设置	**脚本名称（属性名称）**	**K-Means SparkML 参数**
输入字段	`features`	
聚类数	`clustersNum`	`k`
初始化模式	`initMode`	`initMode`
初始化步骤数	`initSteps`	`initSteps`
最大迭代次数	`maxIter`	`maxIter`
容差	`toleration`	`tol`
随机种子	`randomSeed`	`seed`

[1] “Class KMeans。” *Apache Spark*. JavaDoc。 Web. 2017 年 10 月 3 日。

[2] Bahmani, Moseley 等著，“Scalable K-Means++。”2012 年 2 月 28 日。http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf。

MultiLayerPerceptron-AS 节点

多层感知器是基于前馈人工神经网络的分类器，其由多个层组成。 每层完全连接到下一层络。 有关多层感知器分类器 (MLPC) 的详细信息，请参阅 https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier。 [1]

SPSS Modeler 中的 MultiLayerPerceptron-AS 节点使用 Spark 进行实现。 要使用此节点，必须设置上游“类型”节点。 MultiLayerPerceptron-AS 节点将从“类型”节点（或者上游源节点的“类型”选项卡）读取输入值。

[1] “多层感知器分类器。” *Apache Spark*. MLlib: Main Guide. Web. 5 Oct 2018.

MultiLayerPerceptron-AS 节点字段

“字段”选项卡指定要在分析中使用的字段。

使用预定义角色。 该选项通知节点使用来自上游类型节点的字段信息。 这是缺省选项。

使用定制字段分配。 要手动分配目标和预测变量，请选择此选项。

目标。 选择要用作预测的目标的字段。

预测变量。 选择一个或多个字段以用作预测输入。

MultiLayerPerceptron-AS 节点构建选项

使用“构建选项”选项卡可以指定 MultiLayerPerceptron-AS 节点的构建选项，包括性能、建模构建和专家选项。 有关这些选项的更多信息，请参阅 http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/classification/MultilayerPerceptronClassifier.html。 [1]

性能

感知器层。 使用此设置以定义要包含的感知器层数。 该值必须大于感知器字段的数量。 缺省值为 **1**。

隐藏层。 指定隐藏层数。 在多个隐藏层之间使用逗号。 缺省值为 **1**。

输出层。 指定输出层数。 缺省值为 **1**。

随机种子值。 如果想要生成随机数字生成器所使用的种子，请单击**生成**。

模型构建

最大迭代次数。 指定要执行的最大迭代次数。 缺省值为 **10**。

仅专家

块大小。 如果想要指定用于堆积矩阵中的输入数据的块大小，请选择“模型构建”部分中的**专家方式**选项。 这可加速计算。 缺省块大小为 **128**。

下表显示 SPSS Modeler MultiLayerPerceptron-AS 节点对话框中的设置与 Spark KDE 库参数之间的关系。

表 *44:* 映射到 *Spark* 参数的节点属性

SPSS Modeler 设置	**脚本名称（属性名称）**	**Spark 参数**
预测变量	`features`	
目标	`label`	
感知器层	`layers[0]`	`layers[0]`
隐藏层	`layers[1…<latest-1>]`	`layers[1…<latest-1>]`
输出层	`layers[<latest>]`	`layers[<latest>]`
随机种子	`seed`	`seed`
最大迭代次数	`maxiter`	`maxiter`

[1] "Class MultilayerPerceptronClassifier." *Apache Spark*. JavaDoc。 Web. 5 Oct 2018.

MultiLayerPerceptron 节点模型选项

模型名称。 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。

注意事项

本信息是为在美国提供的产品和服务编写的。 IBM 可能会提供其他语言形式的本资料。 但是，您可能需要拥有该语言的产品副本或产品版本，才能对其进行访问。

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。 有关您所在区域当前可获得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 可能已拥有或正在申请与本文档内容有关的各项专利。 提供本文档并不意味着授予用户使用这些专利的任何许可。 您可以以书面形式将许可查询寄往：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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，或以书面形式将查询寄往：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

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。 某些管辖区域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和/或程序进行改进和/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 站点中的资料不是本 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。

本程序的被许可方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 Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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。

本文档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 国际程序许可协议或任何同等协议中的条款提供。

所引用的性能数据和客户示例仅作说明用途。 实际的性能结果可能会因特定的配置和运营条件而异。

涉及非 IBM 产品的信息是从这些产品的供应商、已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。

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，若与实际个人或业务企业相似，纯属巧合。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域注册的商标或注册商标。 其他产品和服务名称可能是 IBM 或其他公司的商标。 IBM 商标的最新列表可从 Web 上的“Copyright and trademark information”处获得，网址为：www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 徽标、PostScript 以及 PostScript 徽标是 Adobe Systems Incorporated 在美国和／或其他国家或地区的注册商标或商标。

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 或其子公司在美国或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。

产品文档的条款和条件

根据以下条款和条件授予这些出版物的使用许可权。

适用性

这些条款和条件是对 IBM Web 站点的任何使用条款的补充。

个人使用

您可以复制这些出版物供个人非商业性使用，但前提是保留所有专有权声明。 未经 IBM 明确同意，您不可以分发、展示或制作这些出版物或其中任何部分的演绎作品。

商业性使用

您仅可在贵公司内部复制、分发和显示这些出版物，但前提是保留所有专有权声明。 未经 IBM 明确同意，您不可以制作这些出版物的演绎作品，或者在您的企业外部复制、分发或展示这些出版物或其中的任何部分。

权利

除非本许可权中明确授予，否则不得授予对这些出版物或其中包含的任何信息、数据、软件或其他知识产权的任何许可权、许可证或权利，无论明示的还是暗含的。

只要 IBM 认为这些出版物的使用会损害其利益或者 IBM 判定未正确遵守上述指示信息，IBM 将有权撤销本文授予的许可权。

只有您完全遵循所有适用的法律和法规，包括所有的美国出口法律和法规，您才可以下载、出口或再出口该信息。

IBM 对这些出版物的内容不作任何保证。 这些出版物“按现状”提供，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适销性、非侵权和适用于某种特定用途的保证。

词汇表

A

AICC

用于根据 -2 (受限) 对数似然选择和比较混合模型的度量。 值越小，表示模型拟合得越好。 AICC 用于更正小样本的 AIC。 随样本大小的增加，AICC 将收敛为 AIC。

B

Bayesian 信息标准 *(Bayesian Information Criterion, BIC)*

用于根据 -2 对数似然选择和比较模型的度量。 值越小，表示模型拟合得越好。 BIC 也会“惩罚”过多参数模型（例如，具有大量输入的复杂模型），但比 AIC 更严格。

Box's M 检验

一种用于检查组协变量矩阵是否相等的检验。 对于十分大的样本，不显著的 p 值表示矩阵不同的证据不足。 这个检验对于违反多变量常态的情况很敏感。

C

案例

针对每个观测值显示的实际组、预测组、后验概率和判别评分的代码。

分类结果 *(Classification Results)*

基于判别分析正确或不正确分配给每组数据的个案数。 有时也称为“混淆矩阵”。

联合组散点图 *(Combined-Groups Plots)*

创建前两个判别函数值的全组散点图。 如果只有一个函数，则会显示一个直方图。

协方差 *(Covariance)*

两个变量间关联性的非标准化测量值，等于叉积偏差除以 N-1。

F

Fisher (Fisher's)

显示可以直接用于分类的 Fisher 分类函数系数。 将为每个组获取一个单独的分类函数系数集，并且会对该组分配一个具有最大判别评分的观测值（分类函数值）。

H

风险图 *(Hazard Plot)*

以线性比例显示累积风险函数。

K

峰度 *(Kurtosis)*

存在离群值的程度的测量。 对于正态分布，峰度统计量的值为零。 正峰度值表示数据呈现比正态分布更极端的离群值。 负峰度值表示数据呈现比正态分布极端程度较低的离群值。

L

留一分类 *(Leave-one-out Classification)*

分析中的每个个案均通过从该个案以外的所有其他个案衍生的函数进行分类。 也称为“U 方法”。

M

MAE

平均误差。 测量序列与其模型预期水平的变化程度。 MAE 在原始序列单元中报告。

马氏距离 *(Mahalanobis Distance)*

自变量上观测值的值与所有观测值的平均值相差多少的测量值。 较大马氏距离表示某一观测值在一个或多个自变量上有极值。

MAPE

平均绝对误差百分比。 测量相关序列距其模型预测等级的变化程度的测量。 它使用独立单元，因此可以用于比较具有不同单元的序列。

MaxAE

最大绝对误差。 最大预测误差作为独立序列以相同单位表示。 与 MaxAPE 相似，它对设想预测的最坏情形很有帮助。 最大绝对误差和最大绝对百分比误差可以出现在不同的序列点。例如，当某一较大序列值的绝对误差略大于某一较小序列值的绝对误差时。 在这种情况下，最大绝对误差将出现在较大序列值处，最大绝对百分比误差将出现在较小序列值处。

MaxAPE

最大绝对百分比误差。 最大预测误差，以百分比形式表示。 此测量对设想预测的最坏情形很有帮助。

最大化最小 *F* 比纳入法 *(Maximizing the Smallest F Ratio Method of Entry)*

一种在逐步分析中选择变量的方法，该方法基于最大化从两组间的 Mahalanobis 距离计算得到的 F 比。

Maximum

数值变量的最大值。

平均值

集中趋势的测量。 算术平均值，等于总和除以观测值数。

均值 *(Means)*

显示自变量的总均值、组均值以及标准差。

中位数

大于或小于一半观测值的值，即 50th 个百分位。 如果有偶数个观测值，则中位数为它们以升序或降序排列时两个中间观测值的平均值。 中位数是集中趋势的一种测量，对离群值不敏感（与平均值不同，平均值会受部分极高或极低值的影响）。

最小化威尔克斯 λ *(Minimize Wilks' Lambda)*

一种用于逐步判别分析的变量选择方法，该方法根据变量的威克斯 λ 下限的多少来选择要输入到方程中的变量。 在每一步中，输入使整体威尔克斯 λ 最小化的变量。

Minimum

数值变量的最小值。

众数 *(Mode)*

最常出现的值。 如果多个值共享最大出现频率，则每个值都是一个众数。

N

标准化 *BIC (Normalized BIC)*

标准化贝叶斯信息准则。 尝试说明模型复杂性的模型总体拟合度的一般性测量。 它是一种基于均方误差的评分，包含模型中参数数量的罚分和序列的长度。 罚分会去除具有更多参数的模型的优势，使得统计量易于通过不同模型针对相同序列进行比较。

O

1 减生存函数

用于根据线性尺度按照被一减的方式绘制生存函数的散点图。

R

范围

数字变量的最大值与最小值的差值就是用最大值减最小值后得出的值。

Rao V（判别分析）(Rao's V (Discriminant Analysis))

组均数之间差异的测量。 也称为 Lawley-Hotelling 跟踪。 每一步都需要输入最大化 Rao 的 V 中的增量的变量。 选择此选项后，输入用于分析的变量的最小值。

RMSE

均方根误差。 均方误差的平方根。 用于测量因变量序列与其模型预测水平的差异程度，用与因变量序列相同的单位数表示。

R 方 *(R-Squared)*

线性模型的拟合优度度量，有时称为决定系数。 它是因变量中由回归模型解释的变异的比例。 它的取值范围从 0 到 1。 较小的值表示模型不适合数据。

S

分组 *(Separate-Groups)*

分组协方差矩阵用于分类。 由于分类基于判别函数（而不是基于原始变量），因此该选项并不始终等同于二次判别。

独立组协方差 *(Separate-Groups Covariance)*

显示每组的独立协方差矩阵。

独立组散点图 *(Separate-Groups Plots)*

创建前两个判别函数值的独立组散点图。 如果仅有一个函数，则显示直方图。

连续 *Bonferroni (Sequential Bonferroni)*

这是一种连续下降的拒绝的 Bonferroni 过程，该过程在拒绝单个假设方面保守程度很低，但保持相同的整体显著性水平。

连续 *Sidak (Sequential Sidak)*

这是一个逐步下降的排斥性 Sidak 过程，就排斥单个假设而言，其保守性小得多，且保持了相同的总体显著性水平。

偏度 *(Skewness)*

分布不对称性的测量。 正态分布是一种对称性分布，其偏度值为 0。 具有显著性正偏度的分布右侧尾部较长。 具有显著负偏态的分布具有向左延伸的长尾。 提示：取大于其标准误差两倍的偏度值指示离开对称的距离。

标准差

围绕平均值的离差的测量，等于方差的平方根。 以和原始变量相同的单位度量标准差。

标准差

对围绕平均值的离差的测量。 在正态分布中，68% 的观测值落入与均值相距不到一个标准差的范围内，95% 落入两个标准差的范围内。 例如，如果平均年龄值 45，标准差为 10，则 95% 的观测值将介于正态分布的 25 到 65 之间。

标准错误 *(Standard Error)*

检验统计量值因样本而异的测量。 这是统计量抽样分布的标准差。 例如，均数的标准误差是样本均数的标准差。

峰度标准误差 *(Standard Error of Kurtosis)*

峰度与其标准误差的比率可用作正态性检验 (即，如果比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 峰度较大的正值表示该分布的尾部比正态分布的尾部长；峰度的负值表示较短的尾部（与箱形均匀分布的尾部变得相似）。

均值标准误差 *(Standard Error of Mean)*

对取自相同分布的样本之间的平均值可能有多大差异的测量。 用于粗略将观测到的均数与假设值对比（即，如果差异与标准误差的比率小于 -2 或大于 +2，则可以得出此均数与假设值不同的结论）。

偏度标准误差 *(Standard Error of Skewness)*

偏度与其标准误差的比率可用作正态性检验 (即，如果该比率小于 -2 或大于 + 2 ，那么可以拒绝正态性)。 偏度正值越大表示长尾向右越长；负极值表示向左的长尾。

固定 *R* 方 *(Stationary R-squared)*

将模型的固定部分与简单均值模型进行比较的测量。 当存在趋势或季节模式时该度量值对普通 R 平方更具优势。 固定的 R 方可以为负值，其范围为负无穷到 1。 负值表示考虑中的模型比基准模型要差。 正值表示所检验的模型比基准模型好。

Sum

所有带有非缺失值的观测值的值的合计或总计。

生存图

在线性刻度上显示累积生存函数。

T

区域图

用于根据变量值对个案进行分组的边界的图。 与观测值被划分到的组相对应的编号。 在每个组的边界内使用星号标记该组的均数。 当只有一个判别函数时不显示此图。

总协方差

显示所有观测值均来自一个样本时的协方差矩阵。

U

无法解释的方差

在每一步，均是输入能使组间未解释变动合计最小的变量。

UNIQUE

同步评估所有效应，同时为任意类型的所有其他效应调整每一个效应。

单变量 *ANOVA*

为每个自变量的组均值的等同性执行单向方差检验分析。

未标准化

显示未标准化的判别函数系数。

使用 *F* 的值

如果变量的 F 值大于 Entry 值，那么将该变量输入到模型中，如果 F 值小于 Removal 值，那么将其除去。 纳入值必须大于移除值，并且这两个值都必须为正数。 要将更多变量输入到模型中，需要减小纳入值。 要从模型中删除更多变量，需要增大移除值。

使用 *F* 的概率

如果变量 F 值的显着性水平小于 "输入" 值，那么会将该变量输入到模型中; 如果显着性水平大于 "剔除" 值，那么会将该变量移除。 纳入值必须小于移除值，并且这两个值都必须为正数。 要将更多变量输入到模型中，需要增大纳入值。 要将更多的变量从模型中移去，请降低“剔除”值。

V

有效

有效观测值既不包含系统缺失值，也不包含定义为用户缺失的值。

偏差

对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。 方差按单元计量，即变量自身单元数的平方。

W

组内 *(Within-Groups)*

汇聚的组内协方差矩阵用来对观测值分类。

组内相关

显示合并组内协方差矩阵，该矩阵是通过在计算相关性之前对所有组的各个协方差矩阵求平均值得到的。

组内协方差

显示合并组内协方差矩阵，该矩阵可能与总协方差矩阵不同。 该矩阵是通过对所有组的各个协方差矩阵求平均值得到的。

索引

Special Characters

A

B

C

P

Q

R

S

T

W

X