IBM SPSS Modeler 18.5 数据库内挖掘指南

注

在使用本资料及其支持的产品之前，请阅读第 81 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 企业强度的数据挖掘工作台。 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，世界范围的商业、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。 通过在日常活动中融入 IBM SPSS 软件，成为预测企业的组织可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 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 Web 站点：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 数据库内数据挖掘

数据库建模概述

IBM SPSS Modeler Server 支持与数据库供应商提供的数据挖掘和建模工具相集成，包括 IBM Netezza、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。 通过集成，可将 IBM SPSS Modeler 的分析功能和易用性将与数据库的功能和性能相结合，同时还兼备数据库供应商提供的数据库自有算法。 模型在数据库内创建，然后可以借助 IBM SPSS Modeler 界面以正常方式浏览模型并为之评分，必要时还可使用 IBM SPSS Modeler Solution Publisher 来对模型进行部署。 在 IBM SPSS Modeler 的“数据库建模”选用板中列出了支持的算法。

使用 IBM SPSS Modeler 访问数据库自有算法的若干优势：

- 数据库内的算法常常与数据库服务器紧密集成，这可能有助于提高性能。
- 在“数据库内”构建和存储的模型不仅由可访问该数据库的应用程序共享，且更易于在这些应用程序中部署。

SQL 生成。 数据库内建模与 SQL 生成（又称为“SQL 回送”）存在明显区别。 使用此功能可以生成原生 IBM SPSS Modeler 操作的 SQL 语句，这些语句可以“回送”到数据库（即，在其中执行）以提高性能。 例如，“合并”、“汇总”和“选择”节点均可生成可以通过上述方式回送到数据库的 SQL 代码。 将 SQL 生成与数据库建模结合使用可以使流自始至终在数据库中运行，相比于在 IBM SPSS Modeler 中运行流，前者具有极大的性能优势。

注: 数据库建模和 SQL 优化需要在 IBM SPSS Modeler 计算机上启用 IBM SPSS Modeler Server 连接。 通过启用此设置，您可以访问数据库算法，直接从 IBM SPSS Modeler 回送 SQL 以及访问 IBM SPSS Modeler Server。 要验证当前许可证的状态，请从 IBM SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项**服务器启用**。

关于所支持的算法的更多信息，请参阅针对指定供应商的后续章节。

您需要的内容

进行数据库建模，需要进行以下设置：

- 与安装了所需分析组件（Microsoft Analysis Services 或 Oracle Data Miner）的相应数据库的 ODBC 连接。
- 在 IBM SPSS Modeler 中，必须在“帮助应用程序”对话框（**工具 > 帮助应用程序**）中启用数据库建模。
- 应该启用 IBM SPSS Modeler 以及 IBM SPSS Modeler Server（如果采用）中“用户选项”对话框内的**生成 SQL** 和 **SQL 优化**设置。 请注意，进行数据库建模并非必须启用 SQL 优化，但强烈建议您启用此功能以提高性能。

注: 数据库建模和 SQL 优化需要在 IBM SPSS Modeler 计算机上启用 IBM SPSS Modeler Server 连接。 通过启用此设置，您可以访问数据库算法，直接从 IBM SPSS Modeler 回送 SQL 以及访问 IBM SPSS Modeler Server。 要验证当前许可证的状态，请从 IBM SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项**服务器启用**。

关于详细信息，请参阅针对指定供应商的后续章节。

模型构建

采用数据库算法构建模型和对模型评分的过程类似于 IBM SPSS Modeler 中其他类型的数据挖掘。 节点和建模“块”的一般处理过程类似于 IBM SPSS Modeler 中其他的流处理过程。 唯一的区别是，实际处理和模型构建回送到数据库中进行。

数据库建模流在概念上与 IBM SPSS Modeler 中的其他数据流完全相同；但是，这个流的所有操作均在数据库中执行，例如，使用“Microsoft 决策树”节点进行模型构建便是如此。 运行流时，IBM SPSS Modeler 会指示数据库构建和存储最终模型，而且详细信息将下载到 IBM SPSS Modeler。 数据库中的执行由流中使用的紫色阴影节点指示。

数据准备

无论是否使用了数据库自有算法，为了提高性能，应该尽可能将数据准备工作回送到数据库完成。

- 如果原始数据存储在数据库中，那么目标就是通过确保所有必需上游操作均可转换为 SQL 使数据留在数据库中。 这样可以避免将数据下载到 IBM SPSS Modeler，从而避免可能抵消增益的瓶颈，并允许在数据库中运行整个流。
- 如果原始数据 没有存储于数据库，那么仍可使用数据库建模。 此种情况下，将在 IBM SPSS Modeler 中进行数据准备，所准备的数据集将自动上载到此数据库并进行模型构建。

模型评分

采用数据内数据挖掘在 IBM SPSS Modeler 中生成的模型与常规的 IBM SPSS Modeler 模型不同。 虽然这些模型作为生成的模型“块”显示在模型管理器中，但实际上，它们是保存在远程数据挖掘或数据库服务器上的远程模型。 您在 IBM SPSS Modeler 中所看到的其实是对这些远程模型的引用。 换言之，您所看到的 IBM SPSS Modeler 模型是“空”模型，其中包含数据库服务器主机名、数据库名和模型名等信息。 当对采用数据库自有算法创建的模型进行浏览和评分时，您应当清楚这个明显差别。

创建模型后，您可以将其添加到流并像其他所有在 IBM SPSS Modeler 中生成的模型一样进行评分。 所有评分均在数据库中完成，即使上游操作并非如此。 （如果可以提高性能，上游操作仍可能会被推回数据库，但评分时并不一定要求这样。） 在大多数情况下，您还可以使用数据库供应商提供的标准浏览器来浏览生成的模型。

对于浏览和评分，需要与运行 Oracle Data Miner 或 Microsoft Analysis Services 的服务器的活动连接。

查看结果和指定设置

要查看结果以及指定评分设置，请在流工作区中双击模型。 您还可以选择右键单击此模型，然后选择**浏览**或**编辑**。 具体设置取决于模型的类型。

导出并保存数据库模型

借助“文件”菜单中的选项，可以从模型浏览器中导出数据库模型和摘要，就像导出在 IBM SPSS Modeler 中创建的模型一样。

1. 在模型浏览器的“文件”菜单中，选择以下某项：

- **导出文本**将模型摘要导出到文本文件
- **导出 HTML** 将模型摘要导出到 HTML 文件
- **导出 PMML**（仅支持 IBM Db2 IM 模型）以预测模型标记语言 (PMML) 格式导出模型，导出的模型可以与其他 PMML 兼容软件配合使用。

注: 还可通过从“文件”菜单中选择**保存节点**来保存某个生成的模型。

模型一致性

对于生成的每个数据库模型，IBM SPSS Modeler 会存储一个模型结构说明，同时会以数据库中的模型名称来保存一个模型引用。 生成模型的“服务器”选项卡将显示为此模型所生成的唯一键，此键与数据库中的实际模型相匹配。

IBM SPSS Modeler 使用这个随机生成的键来检查模型是否仍然一致。 这个键会在创建模型时存储在模型描述中。 最好在运行部署流之前检查键匹配情况。

1. 要通过将数据库中存储的模型的描述与 IBM SPSS Modeler 存储的随机键进行比较来检查该模型的一致性，请单击**检查**按钮。 如果未找到数据库模型或键不匹配，那么系统将报错。

查看和导出生成的 SQL

可以在执行前预览所生成的 SQL 代码，这可能有助于您进行调试。

第 3 章 使用 Microsoft Analysis Services 进行数据库建模

IBM SPSS Modeler 与 Microsoft Analysis Services

IBM SPSS Modeler 支持与 Microsoft SQL Server Analysis Services 的集成。 此功能作为 IBM SPSS Modeler 中的建模节点实现，并且可以从“数据库建模”选用板上使用此功能。 如果此选用板不可见，您可以通过启用 MS Analysis Services 集成（位于“帮助应用程序”对话框的“Microsoft”选项卡上）将其激活。 有关更多信息，请参阅主题 第 11 页的『启用与 Analysis Services 的集成』。

IBM SPSS Modeler 支持集成下列 Analysis Services 算法：

- 决策树
- 聚类
- 关联规则
- 朴素贝叶斯
- 线性回归
- 神经网络
- Logistic 回归
- 时间序列
- 序列聚类

下图说明了从客户端到服务器的数据流，其中数据库内挖掘由 IBM SPSS Modeler Server 管理。 模型构建使用 Analysis Services 进行。 生成的模型由 Analysis Services 存储。 对此模型的引用在 IBM SPSS Modeler 流中维护。 然后，该模型从 Analysis Services 下载到 Microsoft SQL Server 或 IBM SPSS Modeler 中进行评分。

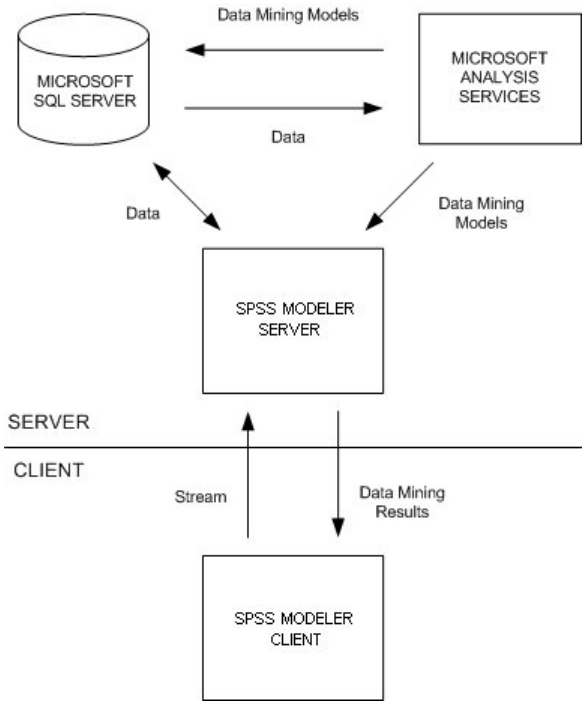

图 *1:* 模型构建过程中，*IBM SPSS Modeler*、*Microsoft SQL Server* 与 *Microsoft Analysis Services* 之间的数据流

注意：尽管可以使用 IBM SPSS Modeler Server，但它不是必需的。 IBM SPSS Modeler 客户端本身就能够处理数据库内挖掘计算。

与 Microsoft Analysis Services 进行集成的需求

以下是在 IBM SPSS Modeler 中使用 Analysis Services 算法执行数据库内建模的必备条件。 您可能需要咨询数据库管理员以确保满足这些条件。

- 在 Windows 上安装 IBM SPSS Modeler Server 后（分布式模式）运行 IBM SPSS Modeler。 与 Analysis Services 的集成不支持 UNIX 平台。

  要点: IBM SPSS Modeler 用户必须使用 Microsoft 提供的 SQL Native Client 驱动程序配置 ODBC 连接，该驱动程序的 URL 在其他 *IBM SPSS Modeler Server* 需求下列出。 建议您不要将 *IBM SPSS Data Access Pack* 提供的驱动程序（一般推荐用于 *IBM SPSS Modeler* 的其他用途）用于此用途。 驱动程序应配置为在启用**与 Windows 认证集成**的条件下使用 SQL Server，因为 IBM SPSS Modeler 不支持 SQL Server 认证。 如果您有关于创建或设置 ODBC 数据源权限方面的疑问，请与数据库管理员联系。
- 必须安装 SQL Server，但不一定与 IBM SPSS Modeler 安装在同一主机上。 IBM SPSS Modeler 用户必须具有足够的权限以读写数据以及删除和创建表和视图。

  注: 建议您使用 SQL Server Enterprise Edition。 Enterprise Edition 提供了用于调整算法结果的高级参数，从而提供了更大的灵活性。 Standard Edition 版本提供了相同的参数但不允许用户编辑某些高级参数。
- Microsoft SQL Server Analysis Services 必须安装在与 SQL Server 相同的主机上。

其他 IBM SPSS Modeler Server 需求

要在 IBM SPSS Modeler Server 中使用 Analysis Services 算法，那么必须在 IBM SPSS Modeler Server 主机上安装以下组件。

注: 如果 SQL Server 安装在 IBM SPSS Modeler Server 所在的主机上，那么这些组件已经可用。

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider（确保选择适合您操作系统的正确版本）
- Microsoft SQL Server Native Client（确保选择适合您操作系统的正确版本）
- 如果您使用的是 Microsoft SQL Server 2008 或 2012，那么可能还需要安装 Microsoft Core XML Services (MSXML) 6.0。

要下载这些组件，转到 *www.microsoft.com/downloads*，搜索 **.NET Framework** 或（对于所有其他组件）**SQL Server Feature Pack**，并选择您的 SQL Server 版本的最新软件包。

这些组件可能需要首先安装其他软件包，此类软件包也可从 Microsoft 下载站点获得。

其他 IBM SPSS Modeler 需求

要在 IBM SPSS Modeler 中使用 Analysis Services 算法，必须安装以上组件，同时在客户端添加以下组件：

- Microsoft SQL Server Datamining Viewer Controls（确保选择了适合您操作系统的正确版本）- 这还需要：
- Microsoft ADOMD.NET

要下载这些组件，转到 *www.microsoft.com/downloads*，搜索 **SQL Server Feature Pack**，并选择您的 SQL Server 版本的最新软件包。

注: 数据库建模和 SQL 优化需要在 IBM SPSS Modeler 计算机上启用 IBM SPSS Modeler Server 连接。 通过启用此设置，您可以访问数据库算法，直接从 IBM SPSS Modeler 回送 SQL 以及访问 IBM SPSS Modeler Server。 要验证当前许可证的状态，请从 IBM SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项**服务器启用**。

启用与 Analysis Services 的集成

要启用 IBM SPSS Modeler 与 Analysis Services 的集成，需要配置 SQL Server 和 Analysis Services，创建 ODBC 源，在 IBM SPSS Modeler 的“帮助应用程序”对话框中启用集成，并启用 SQL 生成和优化。

注意：Microsoft SQL Server 和 Microsoft Analysis Services 必须可用。 有关更多信息，请参阅主题 第 10 页的『与 Microsoft Analysis Services 进行集成的需求』。

配置 SQL Server

配置 SQL Server 以便可以在数据库内进行评分。

1. 在 SQL Server 主机上创建以下注册表键：

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

2. 为该键添加如下 DWORD 键值：

```
AllowInProcess 1
```

3. 完成上述更改后，重新启动 SQL Server。

配置 Analysis Services

必须首先在 Analysis Server“属性”对话框中手动配置两项设置后，IBM SPSS Modeler 才能与 Analysis Services 进行通信：

1. 通过 MS SQL Server Management Studio 登录到 Analysis Server。
2. 要访问“属性”对话框，请右键单击服务器名称，然后选择**属性**。
3. 选中**显示高级（所有）属性**复选框。
4. 更改以下属性：

- 将 DataMining\AllowAdHocOpenRowsetQueries 的值更改为 True（缺省值为 False）。
- 将 DataMining\AllowProvidersInOpenRowset 的值更改为 [all]（无缺省值）。

为 SQL Server 创建 ODBC DSN

要对数据库执行读取或写入，您必须为相关数据库安装并配置 ODBC 数据源，并根据需要配置读许可权或写许可权。 Microsoft SQL Native Client ODBC 驱动程序是必需的，并且自动随 SQL Server 一起安装。 建议您不要将 *IBM SPSS Data Access Pack* 提供的驱动程序（一般推荐用于 *IBM SPSS Modeler* 的其他用途）用于此用途。 如果 IBM SPSS Modeler 和 SQL Server 驻留在不同的主机上，可以下载 Microsoft SQL Native Client ODBC 驱动程序。 有关更多信息，请参阅主题 第 10 页的『与 Microsoft Analysis Services 进行集成的需求』。

如果您有关于创建或设置 ODBC 数据源权限方面的疑问，请与数据库管理员联系。

1. 通过使用 Microsoft SQL Native Client ODBC 驱动程序，创建一个指向数据挖掘过程中使用的 SQL Server 数据库的 ODBC DSN。 余下的驱动程序设置应使用缺省设置。
2. 对于此 DSN，请确保选中了**使用集成的 Windows 认证**。

- 如果 IBM SPSS Modeler 和 IBM SPSS Modeler Server 运行在不同的主机上，请在每个主机上创建相同的 ODBC DSN。 确保每台主机上使用的 DSN 名称相同。

在 IBM SPSS Modeler 中启用 Analysis Services 集成

要使 IBM SPSS Modeler 能够使用 Analysis Services，首先必须在“帮助应用程序”对话框中提供服务器指定信息。

1. 从 IBM SPSS Modeler 菜单中选择：

   工具 > 选项 > 帮助应用程序
2. 单击 **Microsoft** 选项卡。

- **启用 Microsoft Analysis Services 集成**。 启用 IBM SPSS Modeler 窗口底部的“数据库建模”选用板（如果尚未显示）并为 Analysis Services 算法添加节点。
- **分析服务器主机**。 指定运行 Analysis Services 的机器的名称。

- **分析服务器数据库**。 通过单击省略号 (...) 按钮打开一个子对话框，在该对话框中，您可以从可用数据库中选择所需的数据库。 列表中的数据库都是可供指定分析服务器使用的数据库。 由于 Microsoft Analysis Services 在指定数据库中存储数据挖掘模型，因此应选择在其中存储了由 IBM SPSS Modeler 构建的 Microsoft 模型的相应数据库。
- **SQL Server 连接**。 指定 DSN 信息，SQL Server 数据库使用此信息来存储要传递到分析服务器的数据。 请选择用来提供用于构建 Analysis Services 数据挖掘模型的数据的 ODBC 数据源。 如果您要根据平面文件或 ODBC 数据源提供的数据构建 Analysis Services 模型，那么此类数据将自动上载到在此 ODBC 数据源所指向的 SQL Server 数据库中创建的临时表。
- **在即将覆盖数据挖掘模型时发出警告**。 选中此选项可以确保数据库中存储的模型不会在未经警告的情况下被 IBM SPSS Modeler 覆盖。

注意：可以在各个 Analysis Services 节点中覆盖“帮助应用程序”对话框中所作的设置。

启用 SQL 生成和优化

1. 从 IBM SPSS Modeler 菜单中选择：

   工具 > 流属性 > 选项
2. 在导航窗格中单击**优化**选项。
3. 确认是否已启用**生成 SQL** 选项。 要使数据库建模正常发挥作用，此设置是必需的。
4. 选中**优化 SQL 生成**和**优化其他执行**（非严格必需但强烈推荐使用，以使性能更优）。

使用 Analysis Services 构建模型

Analysis Services 模型构建要求训练数据集位于 SQL Server 数据库的表或视图中。 如果数据不在 SQL Server 中，或者需要通过无法在 SQL Server 中执行的数据准备过程在 IBM SPSS Modeler 中进行处理，那么此类数据将在模型构建前自动上载到 SQL Server 中的临时表。

管理 Analysis Services 模型

通过 IBM SPSS Modeler 构建 Analysis Services 模型会在 IBM SPSS Modeler 中创建一个模型，然后在 SQL Server 数据库中创建一个模型或替换其中一个模型。 IBM SPSS Modeler 模型会引用数据库服务器上存储的数据库模型的内容。 IBM SPSS Modeler 可以通过在 IBM SPSS Modeler 模型和 SQL Server 模型中存储相同的已生成模型密钥字符串来执行一致性检查。

MS 决策树建模节点可同时用于分类属性和连续属性的预测建模。 对于分类属性，此节点根据数据集中输入列之间的关系进行预测。 例如，某方案要预测哪些顾客最有可能购买自行车，如果在年轻顾客中购买自行车的比例是十分之九，而在年纪较大的顾客中购买比例仅为十分之二，那么该节点可推断出年龄是有关自行车购买行为的良好预测变量。 决策树可以根据此特定输出结果的趋势进行预测。 对于连续属性，此算法将使用线性回归来确定决策树分割位置。 如果有一个以上的列被设置为可预测的列，或如果输入数据包含一个被设置为可预测的嵌套表，那么该节点可为每个可预测的列构建单独的决策树。

MS Clustering 建模节点采用迭代技术将某个数据集中的观测值分组归入具有类似特征的聚类。 这些分组对于探索数据、识别数据异常和创建预测而言非常有用。 聚类模型可以识别您无法通过表面观测进行逻辑推导而获得的数据集中的关系。 例如，在逻辑上，您可以判断骑自行车上下班的人的工作地点通常离家不远。 但是，此算法可以找出骑自行车上下班的人员的其他不明显特征。 聚类节点区别于其他未指定目标字段的数据挖掘节点。 聚类节点将通过数据中的关系和节点所识别聚类的关系对模型进行严格训练。

MS 关联规则建模节点对于推荐引擎十分有用。 推荐引擎将根据顾客已采购的商品或者其表示感兴趣的商品向客户推荐产品。 将根据同时包含各个观测值的标识以及这些观测值所含项目的标识的数据集来创建关联模型。 观测值中的一组项目称为**项目集**。 关联模型由一系列项目集以及用于描述如何在观测值中将这些项目分组到一起的规则构成。 此算法所发现的规则可用于根据客户购物车中已有的商品预测该客户未来可能购买的商品。

MS 朴素贝叶斯建模节点可计算目标字段和预测变量字段之间的条件概率，并假定这些列是相互独立的。 此模型将所有建议预测变量视为相互独立，因此被称为“朴素”。 此方法比其他 Analysis Services 算法的计算量小，因此对于在建模初期迅速发现关系非常有用。 您可以使用此节点对数据执行初始探索，然后应用结果，以便创建含有其他计算时间可能更长但结果更为准确的节点的附加模型。

MS 线性回归建模节点是决策树节点的变异，其中 MINIMUM_LEAF_CASES 参数被设置为大于或等于节点用来训练挖掘模型的数据集中的观测值总数。 如果按上述方法设置参数，那么该节点将永远不会创建分割，因此可执行线性回归。

MS 神经网络建模节点类似于 MS 决策树节点，即，当给定可预测属性的每个状态时，MS 神经网络节点会为输入属性的每个可能的状态计算概率。 之后，可以根据已输入的属性，使用这些概率预测属性的结果。

MS 逻辑回归建模节点是 MS 神经网络节点的一个变体，其中 HIDDEN_NODE_RATIO 参数设置为 0。 此设置将创建一个神经网络模型，该模型不包含隐藏层，因此等同于逻辑回归。

The **MS 时间序列**建模节点提供的回归算法对连续值（如产品销售）在时间上的预测进行了优化。 虽然其他 Microsoft 算法（例如决策树）需要更多的新信息列作为输入才能预测趋势，但时间序列模型却非如此。 时间序列模型可以仅根据用于创建模型的原始数据集来预测趋势。 您还可以在进行预测时向模型中添加新数据，并将新数据自动并入趋势分析。 有关更多信息，请参阅主题 第 15 页的『MS 时间序列节点』。

MS 序列聚类建模节点标识数据中的顺序序列，并将此分析的结果与聚类技术结合以基于序列和其他属性生成聚类。 有关更多信息，请参阅主题 第 16 页的『MS 序列聚类节点』。

您可以从 IBM SPSS Modeler 窗口底部的“数据库建模”选用板中访问每个节点。

对所有算法节点通用的设置

以下设置通用于所有 Analysis Services 算法。

服务器选项

在“服务器”选项卡上，可以配置分析服务器主机、数据库和 SQL Server 数据源。 此处指定的选项将覆盖“帮助应用程序”对话框的“Microsoft”选项卡上指定的选项。 有关更多信息，请参阅主题 第 11 页的『启用与 Analysis Services 的集成』。

注意：对 Analysis Services 模型进行评分时，还可以使用此选项卡的变体。 有关更多信息，请参阅主题 第 17 页的『Analysis Services 模型块服务器选项卡』。

模型选项

要构建最基本的模型，在进行处理前，需要在“模型”选项卡上指定选项。 评分方法和其他高级选项可在“专家”选项卡上找到。

提供以下基本建模选项：

模型名称。 指定对执行节点时创建的模型赋予的名称。

- **自动**。 基于目标或标识字段名自动生成模型名称，在未指定目标的情况下（例如聚类模型），基于模型类型名称自动生成模型名称。
- **定制**。 允许您为所创建模型指定定制名称。

使用分区数据。 将数据分割成多个不同的子集或样本，以根据当前分区字段进行训练、检验和验证。 通过使用一个样本来创建模型并使用另一个样本对模型进行检验，可以确定此模型适用于与当前数据类似的更大型数据集的程度。 如果未在流中指定分区字段，那么将忽略此选项。

含穿透钻取。 如果显示此选项，那么您可以查询模型以了解模型中所包含观测值的详细信息。

唯一字段。 从下拉列表中，选择用于唯一地标识每个观测值的字段。 通常，这个字段为标识字段，例如 **CustomerID**。

MS 决策树专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 聚类专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 朴素贝叶斯专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 线性回归专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 神经网络专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS Logistic 回归专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 关联规则节点

“MS 关联规则”建模节点对于推荐引擎十分有用。 推荐引擎将根据顾客已采购的商品或者其表示感兴趣的商品向客户推荐产品。 将根据同时包含各个观测值的标识以及这些观测值所含项目的标识的数据集来创建关联模型。 观测值中的一组项目称为**项目集**。

关联模型由一系列项目集以及用于描述如何在观测值中将这些项目分组到一起的规则构成。 此算法所发现的规则可用于根据客户购物车中已有的商品预测该客户未来可能购买的商品。

对于表格格式数据，该算法创建代表每个生成推荐 ($M-*field*) 的概率 ($MP-*field*) 的评分。 对于事务处理格式数据，为支持 ($MS-*field*)、每个生成推荐 ($M-*field*) 的概率 ($MP-*field*) 和调整后概率 ($MAP-*field*) 创建评分。

requirements

事务处理关联模型的需求如下所示：

- **唯一字段**。 关联规则模型需要一个用于唯一地标识记录的键。
- **“标识”字段**。 在构建具有事务处理格式数据的 MS 关联规则模型时，用于标识每个事务的标识字段为必填项。 标识字段可以设置为与唯一字段相同。
- **至少一个输入字段**。 关联规则算法至少需要一个输入字段。
- **目标字段**。 当构建具有事务处理数据的 MS 关联模型时，目标字段必须为事务字段，例如用户购买的产品。

MS 关联规则专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 时间序列节点

“MS 时间序列”建模节点支持两种类型的预测：

- 未来
- 历史记录

未来预测评估在历史数据结束之外若干指定时间段的目标字段值，并总是得到执行。 **历史预测**是在历史数据中具有实际值的若干指定时间段的评估目标字段值。 通过使用历史预测，可以将实际历史值与预测值进行比较，从而评估模型质量。 预测起始点的值确定了是否执行历史预测。

与 IBM SPSS Modeler 时间序列节点不同，MS 时间序列节点不需要提前的时间间隔节点。 另一项区别是，缺省情况下，仅针对预测的行生成评分，而不会针对时间序列数据中的所有历史行生成评分。

requirements

MS 时间序列模型的需求如下所示：

- **单个键时间字段**。 每个模型必须包含一个数值或日期字段，该字段将用作观测值序列并定义模型使用的时间块。 键时间字段的数据类型可以是日期时间数据类型或数字数据类型。 但是，此字段必须包含连续的值，并且这些值对于每个系列必须唯一。
- **单个目标字段**。 在每个模型中，只能指定一个目标字段。 目标字段的数据类型必须具有连续的值。 例如，可以预测数字属性（例如收入、销售额或温度）随时间推移的变化情况。 但是，无法使用包含分类值的字段（例如采购状态或教育程度）作为目标字段。
- **至少一个输入字段**。 MS 时间序列算法需要至少一个输入字段。 输入字段的数据类型必须具有连续的值。 构建模型时，将忽略不连续的输入字段。
- **必须对数据集进行排序**。 输入数据集必须排序（在键时间字段上），否则模型构建会因错误而中断。

MS 时间序列模型选项

模型名称。 指定对执行节点时创建的模型赋予的名称。

- **自动**。 基于目标或标识字段名自动生成模型名称，在未指定目标的情况下（例如聚类模型），基于模型类型名称自动生成模型名称。
- **定制**。 允许您为所创建模型指定定制名称。

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。

含穿透钻取。 如果显示此选项，那么您可以查询模型以了解模型中所包含观测值的详细信息。

唯一字段。 从下拉列表选择键时间字段，该字段用于构建时间序列模型。

MS 时间序列专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

如果要进行历史预测，那么可以包括在评分结果中的历史步骤数由 (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP) 的值确定。 缺省情况下，此限制为 10，这表示只进行 10 项历史预测。 此时，例如当您在模型块的“设置”选项卡上为**历史预测**输入小于 -10 的值时，会发生错误（参见 第 18 页的『MS 时间序列模型块设置选项卡』）。 如果要查看更多历史预测，可以增大 HISTORIC_MODEL_COUNT 或 HISTORIC_MODEL_GAP 的值，但这将导致模型的构建时间延长。

MS 时间序列设置选项

开始估算。 指定预测开始的时间段。

- **起始日期：新预测**。 未来预测的开始时间段，表示为相对于最后一个历史数据时间段的偏移值。 例如，如果您的历史数据结束于 12/99，且您想在 01/00 开始预测，那么应使用值 1；但如果您想在 03/00 开始预测，那么应使用值 3。
- **起始日期：历史预测**。 历史预测的开始时间段，表示为相对于最后一个历史数据时间段的负偏移值。 例如，如果历史数据结束于 12/99，并且要对数据的最后五个时间段进行历史预测，请使用值 -5。

结束估算。 指定预测停止的时间段。

- **预测的结束步骤**。 预测的停止时间段，表示为相对于最后一个历史数据时间段的偏移值。 例如，如果历史数据结束于 12/99，并且您希望预测停止于 6/00，请在这里使用值 6。 对于未来预测，值必须总是大于或等于**开始于**值。

MS 序列聚类节点

MS 序列聚类节点使用一种序列分析算法，该算法探索包含可由以下路径链接的事件的数据或序列。 这方面的一些示例包括用户在 Web 站点中进行导航和浏览时创建的单击路径，或者顾客在网上零售店将商品添加到购物车的顺序。 算法按照分组或聚类找出最常见的序列和等同的序列。

requirements

Microsoft Sequence Clustering 模型的需求如下所示：

- **“标识”字段**。 Microsoft 序列聚类算法要求序列信息以事务处理格式存储。 因此，用于标识每个事务的标识字段为必填。
- **至少一个输入字段**。 此算法至少需要一个输入字段。
- **序列字段**。 算法还需要序列标识字段，该字段必须具有“连续”测量级别。 例如，您可以使用 Web 页面标识、整数或文本字符串，前提是此字段按顺序标识事件。 每个序列只允许使用一个序列标识，并且每个模型中只允许存在一种序列。 序列字段不得与标识字段和唯一字段相同。
- **目标字段**。 构建序列聚类模型时，目标字段为必填。
- **唯一字段**。 序列聚类模型需要一个用于唯一地标识记录的键字段。 可以将唯一字段设置为与标识字段相同。

MS 序列聚类字段选项

所有建模节点都有一个“字段”选项卡，您可以在其中指定要用于构建模型的字段。

必须先指定要用作目标和输入的字段，然后才能构建序列聚类模型。 请注意，对于“MS 序列聚类”节点，无法使用来自上游“类型”节点的字段信息；必须在此处指定字段设置。

标识。 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）。

输入。 请为模型选择一个或多个输入字段。 这些是包含序列建模感兴趣事件的字段。

序列。 请从列表中选择一个字段用作序列标识字段。 例如，您可以使用 Web 页面标识、整数或文本字符串，前提是此字段按顺序标识事件。 每个序列只允许使用一个序列标识，并且每个模型中只允许存在一种序列。 序列字段不得与此选项卡上指定的标识字段以及“模型”选项卡上指定的唯一字段相同。

目标。 选择一个字段用作目标字段，即您将基于序列数据尝试预测其值的字段。

MS 序列聚类专家选项

“专家”选项卡上提供的选项根据所选流的结构不同而有所变化。 有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

对 Analysis Services 模型评分

模型评分发生在 SQL Server 中，并由 Analysis Services 执行。 如果数据源自 IBM SPSS Modeler 内或需要在 IBM SPSS Modeler 内准备，那么可能需要将数据集上载到临时表。 您使用数据库内挖掘从 IBM SPSS Modeler 创建的模型实际是保存在远程数据挖掘或数据库服务器上的远程模型。 对使用 Microsoft Analysis Services 算法创建的模型进行浏览和评分时，这是一项需要了解的重要区别。

在 IBM SPSS Modeler 中，通常只提供一次预测以及关联的概率或置信度。

要获取模型评分示例，请参阅第 19 页的『Analysis Services 挖掘示例』。

对所有 Analysis Services 模型通用的设置

下列设置是所有 Analysis Services 模型的公共设置。

Analysis Services 模型块服务器选项卡

“服务器”选项卡用于为数据库内挖掘指定连接。 此选项卡还提供了唯一的模型键。 此键是在构建模型时随机生成的，并存储于 IBM SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

在“服务器”选项卡上，可以为评分操作配置分析服务器主机和数据库及 SQL Server 数据源。 在 IBM SPSS Modeler 中，此处指定的选项将覆盖那些在“帮助应用程序”或“构建模型”对话框中指定的选项。 有关更多信息，请参阅主题 第 11 页的『启用与 Analysis Services 的集成』。

模型 GUID。 模型键显示在此处。 此键是在构建模型时随机生成的，并存储于 IBM SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

选中这一项。 单击此按钮将根据 Analysis Services 数据库中存储的模型中的键检查模型键。 此操作有助于验证模型是否仍存在于分析服务器中，并表示模型的结构未更改。

注: “检查”按钮仅适用于在准备评分时添加到流工作区中的模型。 如果检查失败，可调查此模型是否已被删除或被服务器上的其他模型替换。

视图。 单击此项可以打开决策树模型的图形视图。 决策树查看器由 IBM SPSS Modeler 中的其他决策树算法所共享，且功能相同。

Analysis Services 模型块汇总选项卡

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。 查看完成后要隐藏结果时，请使用展开控件来折叠想要隐藏的具体结果，或者单击**全部折叠**按钮来折叠所有结果。

分析。 显示指定模型的相关信息。 如果已执行附加到此模型块的分析节点，那么分析中的信息也将显示在此部分中。

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。

MS 时间序列模型块

MS 时间序列模型仅针对预测的时间段生成评分，而不针对历史数据生成评分。

下表显示添加到模型中的字段。

表 *1:* 添加到模型中的字段

字段名称	**描述**
$M-*field*	*field* 的预测值
$Var-*field*	*field* 的计算方差
$Stdev-*field*	*field* 的标准差

MS 时间序列模型块服务器选项卡

“服务器”选项卡用于为数据库内挖掘指定连接。 此选项卡还提供了唯一的模型键。 此键是在构建模型时随机生成的，并存储于 IBM SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

在“服务器”选项卡上，可以为评分操作配置分析服务器主机和数据库及 SQL Server 数据源。 在 IBM SPSS Modeler 中，此处指定的选项将覆盖那些在“帮助应用程序”或“构建模型”对话框中指定的选项。 有关更多信息，请参阅主题 第 11 页的『启用与 Analysis Services 的集成』。

模型 GUID。 模型键显示在此处。 此键是在构建模型时随机生成的，并存储于 IBM SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

选中这一项。 单击此按钮将根据 Analysis Services 数据库中存储的模型中的键检查模型键。 此操作有助于验证模型是否仍存在于分析服务器中，并表示模型的结构未更改。

注: “检查”按钮仅适用于在准备评分时添加到流工作区中的模型。 如果检查失败，可调查此模型是否已被删除或被服务器上的其他模型替换。

视图。 单击此项可以打开时间序列模型的图形视图。 Analysis Services 将整个模型显示为树。 您还可以查看图，该图将显示目标字段在一段时间内的历史值以及预测的未来值。

有关更多信息，请参阅 MSDN 库中对时间序列查看器的说明，位置在 http://msdn.microsoft.com/en-us/library/ms175331.aspx。

MS 时间序列模型块设置选项卡

开始估算。 指定预测开始的时间段。

- **起始日期：新预测**。 未来预测的开始时间段，表示为相对于最后一个历史数据时间段的偏移值。 例如，如果您的历史数据结束于 12/99，且您想在 01/00 开始预测，那么应使用值 1；但如果您想在 03/00 开始预测，那么应使用值 3。
- **起始日期：历史预测**。 历史预测的开始时间段，表示为相对于最后一个历史数据时间段的负偏移值。 例如，如果历史数据结束于 12/99，并且要对数据的最后五个时间段进行历史预测，请使用值 -5。

结束估算。 指定预测停止的时间段。

- **预测的结束步骤**。 预测的停止时间段，表示为相对于最后一个历史数据时间段的偏移值。 例如，如果历史数据结束于 12/99，并且您希望预测停止于 6/00，请在这里使用值 6。 对于未来预测，值必须总是大于或等于**开始于**值。

MS 序列聚类模型块

下表显示添加到 MS 序列聚类模型中的字段（其中 *field* 是目标字段的名称）。

表 *2:* 添加到模型中的字段

字段名称	**描述**
$MC-*field*	此序列所属的聚类的预测。

表 2: 添加到模型中的字段 (继续)	
字段名称	**描述**
$MCP-*field*	此序列属于所预测聚类的概率。
$MS-*field*	*field* 的预测值
$MSP-*field*	$MS-*field* 值正确的概率。

导出模型和生成节点

可以将模型摘要和结构导出到文本格式的文件和 HTML 格式的文件。 适当时，可以生成相应的“选择”和“过滤”节点。

与 IBM SPSS Modeler 中的其他模型块类似，Microsoft Analysis Services 模型块支持直接生成记录和字段操作节点。 使用模型块的“生成”菜单选项，可以生成下列节点：

- 选择节点（仅当在“模型”选项卡上选中某项时）
- 过滤节点

Analysis Services 挖掘示例

其中包含多个样本流，这些样本流演示了如何使用 IBM SPSS Modeler 进行 MS Analysis Services 数据挖掘。 这些流位于 IBM SPSS Modeler 安装文件夹中，该文件夹目录为：

\Demos\Database_Modelling\Microsoft

注意：您可以通过 Windows“开始”菜单中的 IBM SPSS Modeler 程序组来访问 Demos 文件夹。

示例流：决策树

下列流按顺序一起使用可作为使用由 MS Analysis Services 提供的决策树算法的数据库挖掘过程的示例。

表 3: 决策树 - 示例流	
流	**描述**
1_upload_data.str	用于净化数据和将数据从平面文件上载到数据库。
2_explore_data.str	提供关于 IBM SPSS Modeler 数据探索的示例
3_build_model.str	采用数据库自有算法构建模型。
4_evaluate_model.str	用作 IBM SPSS Modeler 模型评估的示例
5_deploy_model.str	部署用于数据库内评分的模型。

注意：要运行此示例，必须按此顺序执行各个流。 另外，必须更新每个流中的源节点和建模节点，以便将想使用的数据库作为有效数据源供您引用。

这些示例流中使用的数据集与信用卡申请有关，演示了同时带有分类和连续预测变量的分类问题。 关于此数据集的更多信息，请参阅示例流中同一文件夹下的 *crx.names* 文件。

此数据集可从位于 *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* 的 UCI Machine Learning Repository 中获得。

示例流：上载数据

第 1 个示例流，即 *1_upload_data.str*，用于清除数据和将数据从纯文本文件上载到 SQL 服务器。

由于 Analysis Services 数据挖掘需要键字段，因而这个初始流通过 IBM SPSS Modeler 的 `@INDEX` 函数，使用“派生”节点将名为 *KEY* 的新字段添加到数据集中，其唯一值为 *1*、*2* 和 *3*。

随后的填充节点用于缺失值处理，并将从文本文件 *crx.data* 中读取的空字段替换为空值。

示例流：探索数据

第二个示例流 *2_explore_data.str* 用于演示如何使用“数据审核”节点获取数据（包括汇总统计量和图形）的一般概述。

双击“数据审核报告”中的图形可显示一个更为详细的图形，用于更深入地探索给定字段。

示例流：构建模型

第 3 个示例流，即 *3_build_model.str*，演示 IBM SPSS Modeler 中的模型构建。 可将数据库模型附加到流并通过双击指定构建设置。

在此对话框的“模型”选项卡上，可以指定以下设置：

1. 选择 **Key** 字段作为唯一标识字段。

在“专家”选项卡上，可以微调设置以构建模型。

在运行之前，请确保指定正确的数据库用于模型构建。 使用“服务器”选项卡调整设置。

示例流：评估模型

第 4 个示例流，即 *4_evaluate_model.str*，演示构建数据库内模型时使用 IBM SPSS Modeler 的优点。 执行该模型之后，可以将其添加回数据流，并使用 IBM SPSS Modeler 中提供的几个工具评估该模型。

查看建模结果

您可以双击模型块以探索结果。 “摘要”选项卡提供了结果的规则树视图。 还可以单击**视图**按钮（位于“服务器”选项卡上）来查看决策树模型的图形视图。

评估建模结果

样本流中的“分析”节点创建一个重合矩阵，以显示每个预测字段与其目标字段之间的匹配模式。 然后，执行“分析”节点以查看结果。

样本流中的“评估”节点可以创建一个增益图，用于显示模型对精确性的提高。 然后，执行“评估”节点以查看结果。

示例流：部署模型

对模型的精确性感到满意后，可以对其进行部署以便与外部应用程序配合使用，或者用于发布回到数据库中。 在最后一个示例流 *5_deploy_model.str* 中，将从表 CREDIT 中读取数据，然后使用“数据库导出”节点对数据进行评分，并将其发布到表 CREDITSCORES。

运行这个流将生成以下 SQL：

```
DROP TABLE CREDITSCORES

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

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

```
            PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
        FROM [CREDIT1] PREDICTION JOIN
            openrowset(''MSDASQL'',
            ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
            T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
            T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
            T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
            T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
            T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
        ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
            and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
            and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
            and [T].[C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

第 4 章 使用 Oracle Data Mining 构建数据库模型

关于 Oracle Data Mining

IBM SPSS Modeler 支持与 Oracle Data Mining (ODM) 的集成，ODM 提供了紧密内嵌于 Oracle RDBMS 中的一系列数据挖掘算法。 这些功能可通过访问 IBM SPSS Modeler 的图形用户界面和面向工作流的开发环境加以使用，使客户可以充分利用 ODM 提供的数据挖掘算法。

IBM SPSS Modeler 支持集成 Oracle Data Mining 的下列算法:

- 朴素贝叶斯
- Adaptive Bayes
- 支持向量机 (SVM)
- 广义线性模型 (GLM)*
- 决策树
- O-Cluster
- K-Means
- 非负矩阵分解 (NMF)
- Apriori
- 最小描述符长度 (MDL)
- 属性重要性 (AI)

* 仅限于 11*g* R1

与 Oracle 进行集成的需求

以下是使用 Oracle Data Mining 执行数据库内建模的先决条件。 您可能需要咨询数据库管理员以确保满足这些条件。

- 以本地模式或在 Windows 或 UNIX 上安装 IBM SPSS Modeler Server 后运行 IBM SPSS Modeler。
- 带有 Oracle Data Mining 选项的 Oracle 10 *g* R2 或 11 *g* R1（10.2 或更高版本的数据库）。

注: 10*g*R2 支持除广义线性模型（需要 11*g*R1）以外的所有数据库建模算法。

- 用于连接至 Oracle 的 ODBC 数据源，如下所述。

注: 数据库建模和 SQL 优化需要在 IBM SPSS Modeler 计算机上启用 IBM SPSS Modeler Server 连接。 通过启用此设置，您可以访问数据库算法，直接从 IBM SPSS Modeler 回送 SQL 以及访问 IBM SPSS Modeler Server。 要验证当前许可证的状态，请从 IBM SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项**服务器启用**。

启用 Oracle 集成

要启用 IBM SPSS Modeler 与 Oracle Data Mining 的集成，需要配置 Oracle，创建 ODBC 源，在 IBM SPSS Modeler 的“帮助应用程序”对话框中启用集成，并启用 SQL 生成和优化。

配置 Oracle

要安装和配置 Oracle Data Mining，请参阅 Oracle 文档（特别是 *Oracle Administrator's Guide*）以获得更多详细信息。

为 Oracle 创建 ODBC 源

要启用 Oracle 和 IBM SPSS Modeler 之间的连接，您需要创建 ODBC 系统数据源名称 (DSN)。

在创建 DSN 之前，您应当对 ODBC 数据源和驱动程序，以及 IBM SPSS Modeler 中的数据库支持有基本的了解。

如果以分布式模式运行 IBM SPSS Modeler Server，请在服务器计算机上创建 DSN。 如果以本地（客户机）模式运行，请在客户计算机上创建 DSN。

1. 安装 ODBC 驱动程序。 您可在此版本随附的 IBM SPSS Data Access Pack 安装盘上找到这些驱动程序。 运行 *setup.exe* 文件以启动安装程序，并选择所有相关的驱动程序。 请按屏幕上的指示信息执行操作，以安装驱动程序。
   a. 创建 DSN。

      注: 菜单序列随 Windows 版本不同而有所变化。

      - **Windows XP**。 从“开始”菜单中选择**控制面板**。 双击**管理工具**，然后双击**数据源 (ODBC)**。
      - **Windows Vista**。 从“开始”菜单中选择**控制面板**，然后选择**系统维护**。 双击**管理工具**，选择**数据源 (ODBC)**，然后单击**打开**。
      - **Windows 7**。 从“开始”菜单，依次选择**控制面板**、**系统和安全**和**管理工具**。 选择**数据源 (ODBC)**，然后单击**打开**。

   b. 转到**系统 DSN** 选项卡，然后单击**添加**。
2. 选择 **SPSS OEM 6.0 Oracle Wire Protocol** 驱动程序。
3. 单击**完成**。
4. 在 ODBC Oracle Wire Protocol 驱动程序安装屏幕中，输入选择的数据源名称、Oracle 服务器的主机名、连接端口号及使用的 Oracle 示例的 SID。

   如果已使用 *tnsnames.ora* 文件配置了 TNS，那么可以从服务器计算机的 *tnsnames.ora* 文件获取主机名、端口和 SID。 要获取更多信息，请与 Oracle 管理员联系。
5. 请单击**测试** 按钮，以测试连接。

在 IBM SPSS Modeler 中启用 Oracle Data Mining 集成

1. 从 IBM SPSS Modeler 菜单中选择：

   工具 > 选项 > 帮助应用程序
2. 单击 **Oracle** 选项卡。

启用 Oracle Data Mining 集成。 启用 IBM SPSS Modeler 窗口底部的“数据库建模”选用板（如果尚未显示）并为 Oracle Data Mining 算法添加节点。

Oracle 连接。 请指定用于构建和存储模型的缺省 Oracle ODBC 数据源以及有效的用户名和密码。 可在各个建模节点和模型块上覆盖此设置。

注意：用于建模的数据库连接可以与用于访问数据的连接相同，也可以不相同。 例如，可能有一个流用于访问一个 Oracle 数据库的数据，将数据下载到 IBM SPSS Modeler 以进行清理或执行其他操作，然后将数据上载到另一个 Oracle 数据库以用于建模。 另外，原始数据也可以位于平面文件或其他（非 Oracle）源中，但在这种情况下，需要将数据上载到 Oracle 才能进行建模。 所有情况下数据都将自动上载到在用于建模的数据库中创建的一个临时表格中。

覆盖 Oracle Data Mining 模型前发出警告。 选中此选项可以确保数据库中存储的模型不会在未经警告的情况下被 IBM SPSS Modeler 覆盖。

列出 Oracle Data Mining 模型。 显示可用的数据挖掘模型。

允许启动 Oracle Data Miner。 （**可选**）启用该选项后，IBM SPSS Modeler 便可以启动 Oracle Data Miner 应用程序。 请参阅 第 36 页的『Oracle Data Miner』 以了解更多信息。

Oracle Data Miner 可执行文件的路径。 （**可选**） 用于指定 Oracle Data Miner for Windows 可执行文件的物理位置（例如 *C:\odm\bin\odminerw.exe*）。 Oracle Data Miner 不会随着 IBM SPSS Modeler 一起安装，必须从 Oracle Web 站点 (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) 下载正确的版本并在客户端进行安装。

启用 SQL 生成和优化

1. 从 IBM SPSS Modeler 菜单中选择：

   工具 > 流属性 > 选项
2. 在导航窗格中单击**优化**选项。
3. 确认是否已启用**生成 SQL** 选项。 要使数据库建模正常发挥作用，此设置是必需的。
4. 选中**优化 SQL 生成**和**优化其他执行**（非严格必需但强烈推荐使用，以使性能更优）。

使用 Oracle Data Mining 构建模型

Oracle 建模节点的工作方式与 IBM SPSS Modeler 中其他建模节点的一样，不过也有几个例外。 可通过横向显示在 IBM SPSS Modeler 窗口底部的数据库建模选用板来访问这些节点。

数据注意事项

Oracle 要求以字符串格式（CHAR 或 VARCHAR2）存储分类数据。 因此，IBM SPSS Modeler 不允许将测量级别为标志或名义（分类）的数字存储字段指定为 ODM 模型的输入。 如有必要，可在 IBM SPSS Modeler 中使用“重新分类”节点将数字转换为字符串。

目标字段。 只能选择一个字段作为 ODM 分类模型的输出（目标）字段。

模型名称。 从 Oracle 11gR1 开始，名称 `unique` 已成为关键字，不能用作定制模型名称。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

一般评论

- 对于 Oracle Data Mining 创建的模型，IBM SPSS Modeler 不提供 PMML 导出/导入功能。
- 模型评分始终在 ODM 中进行。 如果数据来自于 IBM SPSS Modeler 或需要在其中准备数据，那么需要将数据集上载到临时表。
- 在 IBM SPSS Modeler 中，通常只提供一次预测以及关联的概率或置信度。
- IBM SPSS Modeler 将可以用于模型构建和评分的字段数限制为 1000。
- IBM SPSS Modeler 可以从使用 IBM SPSS Modeler Solution Publisher 发布执行的流中对 ODM 模型进行评分。

Oracle 模型服务器选项

指定用于上传建模数据的 Oracle 连接。 如果需要，您可以在“服务器”选项卡上为每个建模节点都选择一个连接，以覆盖在“帮助应用程序”对话框中指定的缺省 Oracle 连接。 有关更多信息，请参阅主题 第 23 页的『启用 Oracle 集成』。

注释

- 用于建模的连接可以与流的源节点中使用的连接相同，也可以不相同。 例如，可能有一个流用于访问一个 Oracle 数据库的数据，将数据下载到 IBM SPSS Modeler 以进行清理或执行其他操作，然后将数据上载到另一个 Oracle 数据库以用于建模。
- ODBC 数据源名称可有效地内嵌于每个 IBM SPSS Modeler 流中。 如果在一台主机上创建的流在另一台主机上执行，那么该数据源在两台主机上的名称必须相同。 另外，也可以在各个源或建模节点的“服务器”选项卡上选择另一个数据源。

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。

注意：仅“决策树”模型允许在构建时指定损失。

Oracle 朴素贝叶斯

朴素贝叶斯是广泛用于处理分类问题的算法。 此模型将所有建议预测变量视为相互独立，因此被称为朴素。朴素贝叶斯是一种可伸缩的快速算法，用于计算各个属性与目标属性的组合的条件概率。 使用训练数据确定独立的概率。 给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

- 交叉验证用于检验模型拟合（用于构建模型的）数据的准确性。 如果可用于构建模型的观测值的数量很小，那么该交叉验证特别有用。
- 模型输出可用矩阵格式浏览。 矩阵中的数字为条件概率，这些条件概率与预测的类（列）和预测变量值的组合（行）相关联。

朴素贝叶斯模型选项

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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

朴素贝叶斯专家选项

除非给定的值或值对在训练数据中具有足够高的发生率，否则在模型构建后，单个预测变量属性值或值对将被忽略。 基于训练数据中的记录数而计算出来的分数值，可指定用于忽略值的发生率临界值。 调整此临界值可减少噪声并改进模型拟合其他数据集的能力。

- **单项阈值**。 指定给定的预测变量属性值的临界值。 给定值的出现次数必须等于或大于指定的分数，否则该值将被忽略。
- **成对阈值**。 指定给定属性和预测变量值对的临界值。 给定值对的出现次数必须等于或大于指定的分数，否则该值对将被忽略。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。 要启用此功能，请选择**选择**，单击**指定**按钮，选择其中一个可能的结果，然后单击**插入**。

使用预测集。 对于目标字段的所有可能输出结果，生成所有可能结果的表。

Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) 使用最小描述符长度 (MDL) 和自动特征选择来构造 Bayesian Network 分类符。 尽管 ABN 的执行速度慢些，但在朴素贝叶斯表现糟糕的某些情况中它仍有良好表现，而在其他大多数情况下也至少不比朴素贝叶斯差。 ABN 算法能够用于构建三种高级的、基于 Bayesian 的模型，包括简化的决策树（单功能）、修剪的朴素贝叶斯和增强型多功能模型。

注: Oracle 12C 中已丢弃 Oracle Adaptive Bayes 算法，而且使用 Oracle 12C 时此算法在 IBM SPSS Modeler 中不受支持。 请参阅 http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726。

已生成的模型

在单功能构建模式中，ABN 可根据一组人可读规则生成一个简化的决策树，使业务用户或分析人员可以了解模型预测的基础并据此向其他人演示或解说。 相比于朴素贝叶斯和多功能模型，这是一个突出的优势。 这些规则可以像 IBM SPSS Modeler 中的标准规则集一样进行浏览。 如下所示的是一个简单的规则集:

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

修剪的朴素贝叶斯和多功能模型无法在 IBM SPSS Modeler 中浏览。

Adaptive Bayes 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

模型类型

构建模型时有三种不同模式可供选择。

- **多特征**。 构建和对比若干个模型，包括 NB (朴素贝叶斯) 模型、单功能产品概率模型和多功能产品概率模型。 这是最详尽的模式，但通常所需的计算时间也最长。 只有单功能模型胜出而成为最佳模型时，才会产生规则。 如果选择了多特征或 NB 模型，那么不会生成任何规则。
- **单特征**。 根据规则集创建简化决策树。 每个规则均含有一个条件以及与每个结果关联的概率。 各规则互相排斥且其为人可读格式，这可能是相比于朴素贝叶斯和多功能模型的重要优点。
- **朴素贝叶斯**。 构建单一 NB 模型并将它与全局样本先验分布进行对比（全局样本中目标值的分布）。 只有 NB 模型胜出而成为比全局先验分布更好的目标值预测变量时，才产生 NB 模型作为输出。 否则，将不会输出任何模型。

Adaptive Bayes 专家选项

限制执行时间。 请选择此选项来指定以分钟表示的最长构建时间。 此选项可用于缩短模型生成时间，不过这样一来，所生成的模型准确性较差。 该算法将在建模过程中的每个重要步骤检验是否能够在指定的时间内完成下一个重要步骤，然后再继续下一步，并在达到限制时返回可用的最佳模型。

最大预测变量数。 此选项可用于通过限制使用的预测变量的数量，来限制模型的复杂性和提高执行速度。预测变量将根据预测变量与目标相关性的 MDL 度量值来进行排序，此排序度量了预测变量包含在模型中的可能性。

最大朴素贝叶斯预测变量数。 此选项指定朴素贝叶斯模型中使用的预测变量的最大数。

Oracle 支持向量机 (SVM)

支持向量机 (SVM) 是一种分类和回归算法，它使用机器学习理论在不过度拟合数据的同时，最大限度地提高预测准确性。 SVM 使用训练数据的可选非线性变换，接着在变换后的数据中搜索回归方程以分隔类（对于分类目标）或拟合目标（对于连续目标）。 Oracle 上配置了 SVM 后，就可以使用这两个可用核函数（线性和高斯）中的其中一个来构建模型。 线性核函数完全忽略了非线性转换，使得生成的模型本质上为回归模型。

有关详细信息，请参阅《Oracle Data Mining 应用程序开发人员指南》和《Oracle Data Mining 概念》。

Oracle SVM 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

主动学习。 提供处理大型建模数据集的方法。 算法可使用主动学习，根据小样本创建一个初始模型，随后将初始模型应用到完整的训练数据集中，再根据结果递增地更新样本和模型。 更新循环将不断重复，直到模型在训练数据上收敛，或支持向量的数量达到了允许的最大值。

内核函数。 选择**线性**或**高斯**，或保留缺省的**系统已确定**允许系统选择最适合的内核。 高斯核函数模拟更复杂的关系，但一般来说，耗费的计算时间更长。 可先使用线性核函数，然后如果线性核函数未能找到合适的拟合，再尝试使用高斯核函数。 这种方法在回归模型中更常用，因为回归模型中核函数的选择更重要。 同时请注意，用高斯核函数构建的 SVM 模型在 IBM SPSS Modeler 中无法浏览。 用线性核函数构建的模型则可以像浏览标准回归模型一样在 IBM SPSS Modeler 中进行浏览。

标准化方法。 指定用于连续输入字段和目标字段的标准化方法。 可选择 **Z-Score**、**最值法**或**无**。 如果选中**自动数据准备**复选框，Oracle 将自动执行标准化。 取消选中此复选框以选择手动标准化方法。

Oracle SVM 专家选项

内核高速缓存大小。 指定以字节表示的缓存大小，该缓存用于保存构建操作期间计算的核函数。 如所预期，较大的缓存通常构建速度更快。 缺省值为 50 MB。

收敛容差。 指定模型构建终止前允许的容差值。 此值必须在 0 到 1 之间。 缺省值为 0.001。 值较大，构建速度也较快，但模型准确率较低。

指定标准差。 指定高斯核函数使用的标准差参数。 此参数影响着模型的复杂度和拓展到其他数据集的能力（即数据的过度拟合和失度拟合）之间的平衡。 标准差值越高，越容易倾向于失度拟合。 此参数值缺省通过训练数据估算得出。

指定 Epsilon。 仅适用于回归模型，用于指定构建对 epsilon 不敏感的模型时可允许错误的区间的值。 换言之，它用于区分小错误（忽略）与大错误（不可忽略）。 此值必须在 0 到 1 之间。 缺省情况下，这是根据训练数据估算的。

指定复杂性因子。 指定复杂性因子，复杂性因子用于平衡模型错误（通过训练数据测量出）和模型复杂度，以防止数据的过度拟合和失度拟合。 该值越高则对错误的罚分就越高，数据过度拟合的风险也越高；值越低则对错误的罚分就越低，也就越容易数据的失度拟合。

指定离群值比率。 指定训练数据中期望的离群值比率。 只对一级 SVM 模型有效。 不能与**指定复杂性因子**设置一起使用。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。 要启用此功能，请选择**选择**，单击**指定**按钮，选择其中一个可能的结果，然后单击**插入**。

使用预测集。 对于目标字段的所有可能输出结果，生成所有可能结果的表。

Oracle SVM 权重选项

在分类模型中，通过使用权重，可以指定各个可能的目标值的相对重要性。 这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。 权重可以使模型产生偏差，以便弥补那些在数据中没有得到很好表示的类别。 增加目标值的权重会增加该类别获得正确预测的百分比。

有三种方法可用来设置权重：

- **基于训练数据**。 这是缺省选项。 权重以训练数据中类别的相对频率为基础。
- **对所有的类都相等**。 所有类别的权重都定义为 1/*k*，其中 *k* 是目标类别数。
- **自定义**。 您可以指定自己的权重。 所有的类的权重起始值设置为相等。 您可以将各个类别的权重调整为用户定义的值。 要调整特定分类的权重，可在表中对应于所需类别的权重单元格中，先清除其内容，然后输入所需的值。

所有类别的权重之和应为 1.0。 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。 此项自动调整操作可以保留各类别的比例，同时实施权重约束。 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。

Oracle 广义线性模型 (GLM)

（仅限于 11g）“广义线性模型”放宽了线性模型所作的限制假设。 例如，这包括假定目标变量具有正态分布，以及假定预测变量对目标变量的效应在本质上是线性效应。 广义线性模型适合于目标分布可能是非正态分布的预测，例如多项式分布或泊松分布。 同样，广义线性模型在预测变量与目标之间的关系或链接有可能是非线性关系或链接的情况下非常有用。

有关详细信息，请参阅《Oracle Data Mining 应用程序开发人员指南》和《Oracle Data Mining 概念》。

Oracle GLM 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

标准化方法。 指定用于连续输入字段和目标字段的标准化方法。 可选择 **Z-Score**、**最值法**或**无**。 如果选中 **自动数据准备**复选框，Oracle 将自动执行标准化。 取消选中此复选框以选择手动标准化方法。

缺失值处理。 指定如何处理输入数据中的缺失值：

- **替换为均值或众数**将数值属性的缺失值替换为均值，并将分类属性的缺失值替换为众数。
- **仅使用完整记录**忽略带有缺失值的记录。

Oracle GLM 专家选项

使用行权重。 选中此复选框以激活相邻下拉列表，从中可以为行选择包含权重因子的列。

将行诊断保存到表。 选中此复选框以激活相邻文本字段，在此可以输入表格名称以包含行级别诊断。

系数置信水平。 目标的预测值在模型计算的置信度区间内的确定性程度，从 0.0 到 1.0。 置信度边界随系数统计信息一起返回。

目标的参考类别。 选择**定制**为用作参考类别的目标字段选择值或保留缺省值**自动**。

岭回归。 岭回归是一种补偿在变量中有太高相关性程度的情况的方法。 您可以使用**自动** 选项，允许算法控制此方法的使用，或者也可通过**禁用** 和**启用** 选项手动控制。 如果您选择手动启用岭回归，那么可以通过在相邻字段中输入值来覆盖岭参数的系统缺省值。

为岭回归生成 VIF。 如果您想当 Ridge 正在用于线性回归时生成方差膨胀因子 (VIF) 统计量，选中此复选框。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。 要启用此功能，请选择**选择**，单击**指定**按钮，选择其中一个可能的结果，然后单击**插入**。

使用预测集。 对于目标字段的所有可能输出结果，生成所有可能结果的表。

Oracle GLM 权重选项

在分类模型中，通过使用权重，可以指定各个可能的目标值的相对重要性。 这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。 权重可以使模型产生偏差，以便弥补那些在数据中没有得到很好表示的类别。 增加目标值的权重会增加该类别获得正确预测的百分比。

有三种方法可用来设置权重：

- **基于训练数据**。 这是缺省选项。 权重以训练数据中类别的相对频率为基础。
- **对所有的类都相等**。 所有类别的权重都定义为 1/*k*，其中 *k* 是目标类别数。
- **自定义**。 您可以指定自己的权重。 所有的类的权重起始值设置为相等。 您可以将各个类别的权重调整为用户定义的值。 要调整特定分类的权重，可在表中对应于所需类别的权重单元格中，先清除其内容，然后输入所需的值。

所有类别的权重之和应为 1.0。 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。 此项自动调整操作可以保留各类别的比例，同时实施权重约束。 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。

Oracle 决策树

Oracle Data Mining 根据常用的分类和回归树算法，提供了一种经典的决策树功能。 ODM 决策树模型含有每个节点的完整信息，包括置信度、支持和分割标准。 可以显示每个节点的完整规则，而且还提供每个节点的替代属性，该替代属性用于在将模型应用到具有缺失值的观测数据时作为替代。

决策树的广泛应用是因为它适用性广、便于应用及易于理解。 决策树将对所有可能的输入属性进行筛选，以查找最佳“分割器”，即属性切割点（例如，`AGE > 55`），以便将下游数据记录分割成若干更均质的总体。 每次分割决策后，ODM 将重复长出整个树和创建终端“叶子”的过程，该叶子代表具有类似记录、项目或人员的总体。 从根树节点（例如，总人数）向下查看，决策树提供了 `IF A, then B` 语句的人类可读规则。 这些决策树规则还提供每个树节点的支持和置信度。

Adaptive Bayes Networks 也可以提供用于解释每项预测的简单规则，但每个分割决策的 Oracle Data Mining 完整规则是由决策树提供。 决策树还可以用于生成最佳客户、已恢复健康病人、与欺骗关联的因子等的详细配置信息。

决策树模型选项

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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11g）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

杂质度量。 指定寻求分割每个节点数据的最佳测试问题时使用的度量。 最佳分割器和分隔值是那些能最大限度提高节点中各实体的目标值均一性的分割器和分隔值。 均一性通过一个度量值来衡量。 受支持的度量值为 **吉尼** 和 **熵** 。

决策树专家选项

最大深度。 设置要构建的树模型的最大深度。

节点中记录的最小百分比。 设置节点中记录的最小百分比。

分割的记录的最小百分比。 设置父节点中记录的最小数，该最小数以用于训练模型的记录总数的百分比表示。 如果记录数小于此百分比，那么不会尝试进行任何分割。

节点中的最小记录数。 设置返回记录的最小数。

分割的最小记录数。 设置父节点中记录的最小数，该最小数以数字表示。 如果记录数小于此值，那么不会尝试进行任何分割。

规则标识。 如果选中，模型中会包含一个字符串以在已进行特定分割的树中标识节点。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。 要启用此功能，请选择**选择**，单击**指定**按钮，选择其中一个可能的结果，然后单击**插入**。

使用预测集。 对于目标字段的所有可能输出结果，生成所有可能结果的表。

Oracle O-Cluster

Oracle O-Cluster 算法确定数据中自然发生的分组。 正交分区聚类 (O-Cluster) 是 Oracle 专有的聚类算法，它创建基于分层网格的聚类模型，也就是说，它在输入属性空间中创建轴平行（正交）分区。 该算法递归式地运行。 所产生的分层结构为一个不规则的网格，该网格将属性空间分割成各个聚类。

O-Cluster 算法可处理数字属性和分类属性，且 ODM 将自动选择最佳的聚类定义。 ODM 提供聚类详细信息，聚类规则和聚类矩心值，并可以用于根据总体的聚类成员资格对总体进行评分。

O-Cluster 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11g）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

最大聚类数。 设置生成的聚类的最大数目。

O-Cluster 专家选项

最大缓冲区。 设置最大缓冲区大小。

敏感度。 设置一个分数，该分数指定分割新聚类所要求的最高密度。 该分数与全局均匀密度相关联。

Oracle k-Means

Oracle k-Means 算法确定数据中自然发生的分组。 k-Means 算法是基于距离的聚类算法，该算法将数据分区为预定数量的聚类（条件是存在足够的不同观测值）。 基于距离的算法根据距离度量（函数）来衡量数据点之间的相似性。 根据所使用的距离度量，数据点被指派到与之距离最近的聚类。 ODM 提供增强版的 k-Means。

k-Means 算法支持分层聚类，处理数字和分类属性并将总体分割为用户指定数量的聚类。 ODM 提供聚类详细信息，聚类规则和聚类矩心值，并可以用于根据总体的聚类成员资格对总体进行评分。

k-Means 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

聚类数。 设置生成聚类的数量。

距离函数。 指定 k-Means 聚类使用的距离函数。

分割标准。 指定 k-Means 聚类使用的分割标准。

标准化方法。 指定用于连续输入字段和目标字段的标准化方法。 可选择 **Z-Score**、**最值法**或**无**。

k-Means 专家选项

迭代次数。 设置 k-Means 算法的迭代次数。

收敛容差。 设置 k-Means 算法的收敛容差。

分级数。 指定 k-Means 生成的属性直方图中的图条数。 每个属性的图条边界都是通过对整个训练数据集进行全局计算得到的。 分箱方法为等宽法。 具有单一值的属性只有一个分类，除此以外，其他所有属性均具有同样数量的图条。

块增长。 设置分配用于容纳聚类数据的内存的增长因子。

最小百分比属性支持。 设置属性值分数，该属性值必须为非空，才能使该属性包含在聚类的规则说明中。如果参数值在具有缺失值的数据中设置得过高，那么可能导致规则过短，或甚至为空。

Oracle 非负矩阵分解 (NMF)

非负矩阵分解 (NMF) 用于将大数据集简化为若干具有代表性的属性。 它与主成分分析 (PCA) 的原理类似，但可以处理更大量的属性，在加法表示模型中，NMF 是功能强大的先进数据挖掘算法，而且用途广泛。

NMF 可以用于将大量数据（比如文本数据）简化为小的、稀疏得多的表示，NMF 降低了数据的维度，即用少得多的变量保存了等量的信息。 NMF 模型的输出可用有监督的学习方法（比如 SVM）或没有监督的学习方法（比如 聚类）来进行分析。 Oracle Data Mining 用 NMF 和 SVM 算法来挖掘尚未结构化的文本数据。

NMF 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

标准化方法。 指定用于连续输入字段和目标字段的标准化方法。 可选择 **Z-Score**、**最值法**或**无**。 如果选中**自动数据准备**复选框，Oracle 将自动执行标准化。 取消选中此复选框以选择手动标准化方法。

NMF 专家选项

指定特征数。 指定要提取的特征的数量。

随机种子值。 设置 NMF 算法的随机种子。

迭代数。 设置 NMF 算法的迭代次数。

收敛容差。 设置 NMF 算法的收敛容差。

显示所有特征。 显示所有特征的特征标识和置信度，而不是仅显示最佳特征的特征标识和置信度。

Oracle Apriori

Apriori 算法会发现数据中的关联规则。 例如，“如果客户购买剃须刀和须后产品，那么该客户还会购买剃须膏，并且置信度为 80%。”关联挖掘问题可以分解为两个子问题：

- 找到所有称为频繁项集合的项组合，即支持度大于最小支持度的项组合。
- 使用频繁项集合来生成所需要的规则。 举例说明规则的生成原理，例如，ABC 和 BC 为频繁项，如果 `support(ABC)` 与 `support(BC)` 的比例大于等于最小置信度时，那么可使用“从规则 A 推导出 BC”。 注意：如果 ABCD 为频繁项，该规则将具有最小支持度。 ODM 关联仅支持单一后项规则（从 ABC 推导出 D）。

频繁项集合的数量取决于最小支持度参数。 生成规则的数量取决于频繁项集合的数量和置信度参数。 如果置信度参数设得过高，那么关联模型中可能存在频繁项集合，但不存在规则。

ODM 将基于 SQL 来执行 Apriori 算法。 候选生成和支持计数步骤使用 SQL 查询来执行。 不使用专门的内存储数据结构。 SQL 查询将使用各种提示进行优化，以便能在数据库服务器中高效运行。

Apriori 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 Apriori 模型之前，需要指定要将哪些字段用作与关联建模有关的项目。

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。 选择此选项后，根据是否正在使用事务处理格式来指定对话框中的剩余字段。

如果 没有 使用事务处理格式，请指定：

- **输入**。 选择输入字段。 此操作类似于在“类型”节点中将字段角色设置为输入。
- **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。

如果 正在 使用事务处理格式，请指定：

使用事务格式。 如果希望将每个项目行中的数据转换为每个观测值行中的数据，请使用此选项。

选择此选项会更改该对话框下半部分中的字段控件：

对于事务处理格式，请指定：

- **标识**。 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）。
- **内容**。 指定模型的内容字段。 该字段包含与关联建模有关的项目。
- **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。 通过用某个样本创建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。）

Apriori 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

最大规则长度。 为任何规则设置最大预条件数，该值为从 2 到 20 的整数。 这是一种用来限制规则复杂性的方式。 如果规则过于复杂或过于具体，或者训练规则集所需的时间过长，请尝试减小此设置。

最小置信度。 设置最小置信度级别，介于 0 到 1 之间的值。 置信度低于指定标准的规则将被放弃。

最小支持度。 设置最小支持阈值，介于 0 到 1 之间的值。 Apriori 发现频率高于最小支持阈值的模式。

Oracle 最小描述长度 (MDL)

Oracle 最小描述长度 (MDL) 算法用于确定对目标属性具有最大影响的属性。 通常情况下，知道哪个是最有影响的属性可以更好地了解和管理业务并且有助于简化建模操作。 另外，这些属性可以指示为扩大模型而希望添加的数据的类型。 例如，MDL 可用于找到与以下预测内容最相关的属性：制造的零件的质量、与流失相关联的因素以及最有可能用于治疗特定疾病的基因等等。

Oracle MDL 将废弃它认为对于预测目标而言不重要的输入字段。 然后，它使用余下的输入字段构建与 Oracle Data Miner 中显示的 Oracle 模型相关联的未优化模型块。 在 Oracle Data Miner 中浏览模型将显示一个图表，该图表显示了余下的输入字段，这些字段按其在预测目标方面的重要性顺序排秩。

负排秩指示噪声。 排秩为零或更小值的输入字段不影响预测，应从数据中移除。

要显示图表

1. 在“模型”选用板中右键单击非优化模型块并选择**浏览**。
2. 在模型窗口中，单击按钮以启动 Oracle Data Miner。
3. 连接到 Oracle Data Miner。 有关更多信息，请参阅主题 第 36 页的『Oracle Data Miner』。
4. 在 Oracle Data Miner 导航面板中，展开**模型**，然后展开**属性重要性**。
5. 选择相关的 Oracle 模型（其名称与您在 IBM SPSS Modeler 中指定的目标字段名称相同）。 如果您不确定哪个正确，请选择“属性重要性”文件夹并按创建日期查找模型。

MDL 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。

唯一字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 IBM SPSS Modeler 限制这个键字段必须为数字。

注: 对于除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外的所有 Oracle 节点，此字段均为可选项。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

Oracle 属性重要性 (AI)

属性重要性的目标是找出数据集中的哪些属性与结果相关，以及 它们影响最终结果的程度。 “Oracle 属性重要性”节点将分析数据、查找模式并预测具有相关联置信度的结果。

AI 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。

自动数据准备。 （仅 11*g*）启用（缺省）或禁用 Oracle Data Mining 的自动数据准备模式。 如果选中此框，那么 ODM 将自动执行算法所需的数据转换。 有关详细信息，请参阅 *Oracle Data Mining* 概念。

AI 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。

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。

AI 模型块模型选项卡

针对 Oracle AI 模型块的“模型”选项卡显示所有输入的排名和重要性，并允许您使用左侧列中的复选框来选择用于进行过滤的字段。 运行这个流时，将只保留选中的字段以及目标预测。 其他输入字段将被废弃。 缺省选择基于建模节点中指定的选项，但您可以根据需要选择或取消选择其他字段。

- 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。 另外，可以从“排序依据”按钮旁的列表中选择期望的项目，并使用向上和向下箭头来更改排序方向。
- 可使用工具栏来选中或弃选所有字段和访问“选中字段”对话框，可在该对话框上根据排序或重要性来选择字段。 另外，还可以在按住 Shift 或 Ctrl 键的情况下单击各字段以扩展选择。
- 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。

管理 Oracle 模型

Oracle 模型添加到模型选用板的方式与其他 IBM SPSS Modeler 模型的添加方式一样，而且使用方法也大致相同。 但是，也有几点重大差异，比如 IBM SPSS Modeler 中生成的每个 Oracle 模型实际引用的是存储在数据库服务器上的模型。

Oracle 模型块服务器选项卡

通过 IBM SPSS Modeler 构建 ODM 模型即可在 IBM SPSS Modeler 中创建一个模型，并创建或替代 Oracle 数据库中的一个模型。 这种 IBM SPSS Modeler 模型将引用数据库服务器上存储的数据库模型的内容。 IBM SPSS Modeler 可以通过将完全相同的生成**模型键**字符串存储在 IBM SPSS Modeler 模型和 Oracle 模型中执行一致性检查。

每个 Oracle 模型的键字符串显示在“列出模型”对话框中的模型信息列下。 IBM SPSS Modeler 模型的键字符串在 IBM SPSS Modeler 模型的“服务器”选项卡上显示为**模型键**（放置在流中时）。

模型块“服务器”选项卡上的“检验”按钮，可用于检验 IBM SPSS Modeler 模型中的模型键和 Oracle 模型是否匹配。 如果 Oracle 中无法找到名称相同的模型，或者模型键不匹配，那么 Oracle 模型已被删除或在 IBM SPSS Modeler 模型构建后重新构建。

Oracle 模型块汇总选项卡

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。 查看完成后要隐藏结果时，请使用展开控件来折叠想要隐藏的具体结果，或者单击**全部折叠**按钮来折叠所有结果。

分析。 显示指定模型的相关信息。 如果已执行附加到此模型块的分析节点，那么分析中的信息也将显示在此部分中。

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。

Oracle 模型块设置选项卡

模型块的“设置”选项卡允许您覆盖建模节点上某些选项的设置，以便进行评分。

Oracle 决策树

使用误分类成本。 确定是否在 Oracle 决策树模型中使用误分类成本。 有关更多信息，请参阅主题 第 25 页的『误分类成本』。

规则标识。 如果选择（选中），将规则标识列添加到 Oracle 决策树模型中。 规则标识用于标识树中进行特定分割的节点。

Oracle NMF

显示所有特征。 如果选择（选中），显示所有特征的特征标识和置信度，而不是仅在 Oracle NMF 模型中显示最佳特征的特征标识和置信度。

列出 Oracle 模型

“列出 Oracle Data Mining 模型”按钮用于启动一个对话框，该对话框列出现有数据库模型并允许删除模型。此对话框可以从“帮助应用程序”对话框中打开，也可以通过 ODM 相关节点的构建、浏览和应用对话框打开。

将显示每个模型的以下信息：

- **模型名称**。 模型的名称，用于对列表进行排序
- **模型信息**。 模型键信息，由构建日期/时间和目标列名组成
- **模型类型**。 构建此模型的算法的名称

Oracle Data Miner

Oracle Data Miner 是 Oracle Data Mining (ODM) 的用户界面，并替代以前 IBM SPSS Modeler 的 ODM 用户界面。 Oracle Data Miner 旨在提高分析人员在使用 ODM 算法方面的成功率。 该目标通过以下方式来实现：

- 用户在应用能同时处理数据准备和算法选择的方法学方面需要更多帮助。 Oracle Data Miner 通过提供可引导应用使用正确方法学的数据挖掘操作，解决此用户需求。
- Oracle Data Miner 为模型构建提供改进的和扩展的试探法，为指定模型和转换设置提供可降低错误几率的转换向导。

定义 Oracle Data Miner 连接

1. Oracle Data Miner 可通过任何版本的 Oracle 进行启动，可通过 **启动 Oracle Data Miner** 按钮应用节点和输出对话框。

图 *2: 启动“Oracle Data Miner”按钮*

2. 如果正确设置的“帮助应用程序”选项，那么 Oracle Data Miner 的**编辑连接**对话框将在 Oracle Data Miner 外部应用程序启动之前显示在用户面前。

   注意：此对话框仅在不存在已定义连接名称时显示。

   - 提供一个 Data Miner 连接名称并输入对应的 Oracle 10gR1 或 10gR2 服务器信息。 Oracle 服务器应与 IBM SPSS Modeler 中指定的服务器一样。
3. Oracle Data Miner 的 **选择连接** 对话框提供用于指定使用哪个（以上步骤中定义的）连接名称的选项。

关于 Oracle Data Miner 需求、安装和使用的详细信息，请参阅 Oracle Web 站点上的 Oracle Data Miner。

准备数据

使用 Oracle Data Mining 算法的朴素贝叶斯、Adaptive Bayes 和支持向量机来建模时，可以使用两种类型的数据准备：

- **分箱**，即，对于无法接受连续数据的算法，将连续数字范围字段转换为类别。
- **标准化**，即应用于数字范围的变换，以使这些数字范围具有类似的平均值和标准差。

离散化

IBM SPSS Modeler 的“分级”节点提供了许多执行分级操作的技术。 定义了可以应用于一个或多个字段的分箱操作。 如在数据集上执行分级操作，那么将创建临界值并允许创建 IBM SPSS Modeler 的“派生”节点。 “派生”操作可转换为 SQL 并模型构建和评分前被应用。 此方法将在模型与执行分箱的“派生”节点之间创建依赖关系，但允许分箱规范由多个建模任务重复使用。

规范化

用作支持矢量机模型的输入的连续（数字范围）字段应该先进行标准化，然后再用于模型构建。 对于回归模型，还必须反转标准化，以根据模型输出重新构建评分。 SVM 模型设置用于选择 **Z-Score**、**最值法** 或**无**。 通过 Oracle 构建标准化系数是模型构建过程中的一个步骤，这些系数将被上载到 IBM SPSS Modeler 并保存在模型中。 应用时，这些系数将被转换为 IBM SPSS Modeler 派生表达式，并用于准备（评分时使用的）数据，然后再将数据传输到模型。 此情况中，标准化与建模任务紧密关联。

Oracle Data Mining 示例

提供若干样本流，以演示如何在 IBM SPSS Modeler 中使用 ODM。 这些流位于 *\Demos\Database_Modelling\Oracle Data Mining* 目录下的 IBM SPSS Modeler 安装文件夹中。

注意：您可以通过 Windows“开始”菜单中的 IBM SPSS Modeler 程序组来访问 Demos 文件夹。

下表中的流是数据库挖掘过程的示例，通过使用 Oracle Data Mining 提供的支持向量机 (SVM) 算法，依次使用这些样本流。

表 *4:* 数据库挖掘 - 示例流

流	**描述**
1_upload_data.str	用于净化数据和将数据从平面文件上载到数据库。
2_explore_data.str	提供关于 IBM SPSS Modeler 数据探索的示例
3_build_model.str	采用数据库自有算法构建模型。
4_evaluate_model.str	用作 IBM SPSS Modeler 模型评估的示例
5_deploy_model.str	部署用于数据库内评分的模型。

注意：要运行此示例，必须按此顺序执行各个流。 另外，必须更新每个流中的源节点和建模节点，以便将想使用的数据库作为有效数据源供您引用。

这些示例流中使用的数据集与信用卡申请有关，演示了同时带有分类和连续预测变量的分类问题。 关于此数据集的更多信息，请参阅示例流中同一文件夹下的 *crx.names* 文件。

此数据集可从位于 *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* 的 UCI Machine Learning Repository 中获得。

示例流：上载数据

第一个示例流 *1_upload_data.str* 用于清理平面文件中的数据并将其上载到 Oracle。

由于 Oracle Data Mining 需要唯一的标识字段，因而这个初始流通过 IBM SPSS Modeler 的 `@INDEX` 函数，使用“派生”节点将名为标识的新字段添加到数据集中，其唯一值为 1、2 和 3。

“填充”节点用于处理缺失值，并将从文本文件 *crx.data* 读取的空字段替换为 空 值。

示例流：探索数据

第二个示例流 *2_explore_data.str* 用于演示如何使用“数据审核”节点获取数据（包括汇总统计量和图形）的一般概述。

双击“数据审核报告”中的图形可显示一个更为详细的图形，用于更深入地探索给定字段。

示例流：构建模型

第 3 个示例流，即 *3_build_model.str*，演示 IBM SPSS Modeler 中的模型构建。 双击“数据库源”节点（标注为 CREDIT）以指定数据源。 要指定构建设置，请双击构建节点（最初标注为 CLASS，指定数据源后将更改为 FIELD16）。

在对话框的“模型”选项卡上：

1. 确保选择**标识**作为唯一字段。
2. 确保选择**线性**作为核函数，选择 **Z 评分**作为标准化方法。

示例流：评估模型

第 4 个示例流，即 *4_evaluate_model.str*，演示构建数据库内模型时使用 IBM SPSS Modeler 的优点。 一旦执行完模型，即可将它添加回数据流中并使用 IBM SPSS Modeler 提供的多种工具来评估模型。

查看建模结果

将一个“表”节点附加到模型块以探索结果。 **$O-field16** 字段显示每个观测值中 *field16* 的预测值，而 **$OC-field16** 字段显示该预测的置信度值。

评估建模结果

您可以使用“分析”节点创建重合矩阵，以显示每个预测字段与其目标字段之间的匹配模式。 然后，运行“分析”节点以查看结果。

您可以使用“评估”节点创建增益图，用于显示模型对精确性的提高。 然后，运行“评估”节点以查看结果。

示例流：部署模型

对模型的精确性感到满意后，可以对其进行部署以便与外部应用程序配合使用，或者用于发布回到数据库中。 最后一个示例流，即 *5_deploy_model.str* 中，数据从表格 `CREDITDATA` 读取，然后进行评分并使用名称为部署解决方案的 Publisher 节点将数据发布到表格 `CREDITSCORES`。

第 5 章 使用 IBM Data Warehouse 和 IBM Netezza Analytics 的数据库建模

使用 IBM Data Warehouse 和 IBM Netezza Analytics 的 SPSS Modeler

IBM SPSS Modeler 支持 IBM Data Warehouse 和 IBM Netezza Analytics 集成，这提供了在这些 IBM 服务器上运行高级分析的能力。 这些功能可通过访问 IBM SPSS Modeler 图形用户界面和面向工作流的开发环境加以使用，使您可以在 IBM Netezza 或 IBM Data Warehouse 环境中运行数据挖掘算法。

SPSS Modeler 支持集成来自 **IBM Netezza Analytics** 的以下算法：

- 决策树
- K-Means
- 二阶
- 贝叶斯网络
- 朴素贝叶斯
- KNN
- 分裂式聚类
- PCA
- 回归树
- 线性回归
- 时间序列
- 广义线性

有关这些算法的更多信息，请参阅 *IBM Netezza Analytics* 开发人员指南和 *IBM Netezza Analytics* 参考指南。

SPSS Modeler 支持集成来自 **IBM Data Warehouse** 的以下算法（不支持贝叶斯网络、分裂式聚类和时间序列）：

- 决策树
- K-Means
- 二阶
- 朴素贝叶斯
- KNN
- PCA
- 回归树
- 线性回归
- 广义线性

注: 不支持 AIX。

集成需求

以下是使用 IBM Netezza Analytics 或 IBM Data Warehouse 执行数据库内建模的必备条件。 您可能需要咨询数据库管理员以确保满足这些条件。

- 对在 Windows 或 UNIX（不包括 zLinux，未提供可用的 IBM Netezza ODBC 驱动程序）上安装 IBM SPSS Modeler Server 运行的 IBM SPSS Modeler。
- 运行 IBM Netezza Analytics 数据包的 IBM Netezza Performance Server。

  注：所需的最低 Netezza Performance Server (NPS) 版本取决于所需的 INZA 版本，如下所示：

  – NPS 6.0.0 P8 以上的所有版本都支持 2.0 以前的 INZA 版本。

  – 要使用 INZA 2.0 或更高版本，需要 NPS 6.0.5 P5 或更高版本。

  “Netezza 广义线性”和“Netezza 时间序列”需要 INZA 2.0 及更高版本才能正常运行。 所有其他 Netezza 数据库内节点都需要 INZA 1.1 或更高版本。
- 连接到 IBM Netezza 数据库所需的 ODBC 数据源。 有关更多信息，请参阅主题 第 40 页的『启用集成』。
- 连接到 IBM Data Warehouse 数据库所需的 ODBC 数据源。
- IBM SPSS Modeler 中启用的 SQL 生成和优化。 有关更多信息，请参阅主题 第 40 页的『启用集成』。

注: 数据库建模和 SQL 优化需要在 IBM SPSS Modeler 计算机上启用 IBM SPSS Modeler Server 连接。 通过启用此设置，您可以访问数据库算法，直接从 IBM SPSS Modeler 回送 SQL 以及访问 IBM SPSS Modeler Server。 要验证当前许可证的状态，请从 IBM SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项**服务器启用**。

启用集成

启用与 IBM Netezza Analytics 或 IBM Data Warehouse 的集成包括以下步骤。

- 配置 IBM Netezza Analytics 或 IBM Data Warehouse
- 创建 ODBC 源
- 在 IBM SPSS Modeler 中启用集成
- 在 IBM SPSS Modeler 中启用 SQL 生成和优化

在以下部分中将介绍这些内容。

配置 IBM Netezza Analytics 或 IBM Data Warehouse

要安装和配置 IBM Netezza Analytics 或 IBM Data Warehouse，请参阅相应的 IBM 文档。 例如，对于 IBM Netezza Analytics，请参阅此产品随附的 *IBM Netezza Analytics* 安装指南。 该指南中的设置数据库权限部分包含需要运行以允许 IBM SPSS Modeler 流读取数据库的脚本的详细信息。

注: 如果您将使用依赖于矩阵计算的节点，那么必须通过运行 `CALL NZM..INITIALIZE();` 初始化“矩阵引擎”，否则存储过程的执行将失败。 对于每个数据库，该初始化为一次性设置步骤。

为 IBM Netezza Analytics 创建 ODBC 源

要启用 IBM Netezza 数据库和 IBM SPSS Modeler 之间的连接，您需要创建 ODBC 系统数据源名称（DSN）。

在创建 DSN 之前，您应当对 ODBC 数据源和驱动程序，以及 IBM SPSS Modeler 中的数据库支持有基本的了解。

如果以分布式模式运行 IBM SPSS Modeler Server，请在服务器计算机上创建 DSN。 如果以本地（客户机）模式运行，请在客户计算机上创建 DSN。

Windows 客户端

1. 从您的 *Netezza Client* CD 上，运行 *nzodbcsetup.exe* 文件以启动安装程序。 请按屏幕上的指示信息执行操作，以安装驱动程序。 有关详细说明，请参阅《IBM Netezza ODBC、JDBC 和 OLE DB 安装与配置指南》。

a. 创建 DSN。

注: 菜单序列随 Windows 版本不同而有所变化。

- **Windows XP**。 从“开始”菜单中选择**控制面板**。 双击**管理工具**，然后双击**数据源 (ODBC)**。
- **Windows Vista**。 从“开始”菜单中选择**控制面板**，然后选择**系统维护**。 双击**管理工具**，选择**数据源 (ODBC)**，然后单击**打开**。
- **Windows 7**。 从“开始”菜单，依次选择**控制面板**、**系统和安全**和**管理工具**。 选择**数据源 (ODBC)**，然后单击**打开**。

b. 转到**系统 DSN** 选项卡，然后单击**添加**。

2. 从列表中选择 **NetezzaSQL**，然后单击**完成**。
3. 在 Netezza ODBC 驱动程序设置屏幕的 **DSN 选项**选项卡上，键入选择的数据源名称、IBM Netezza 服务器的主机名或 IP 地址、连接端口号、使用的 Netezza 实例的数据库，以及用于数据库连接的用户名和密码信息。 单击**帮助**按钮获得字段说明。
4. 单击**测试连接**按钮并确保您连接到数据库。
5. 在成功连接后，重复单击**确定**以退出 ODBC 数据源管理器屏幕。

Windows 服务器

对于 Windows Server，该程序与 Windows XP 客户端的程序相同。

UNIX 或 Linux 服务器

以下程序适用于 UNIX 或 Linux 服务器（不包括 zLinux，未提供适用的 IBM Netezza ODBC 驱动程序）。

1. 从您的 Netezza Client CD/DVD 上，将对应的 `<platform>cli.package.tar.gz` 文件复制到服务器上的临时位置。
2. 通过 **gunzip** 和 **untar** 命令，提取存档内容。
3. 为提取的 *unpack* 脚本添加执行权限。
4. 运行脚本，并在屏幕提示时给出回答。
5. 编辑 `modelersrv.sh` 文件以包括以下行。

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

例如:

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

6. 找到文件 `/usr/local/nz/lib64/odbc.ini` 并将其内容复制到随 SDAP 安装的 `odbc.ini` 文件（由环境变量 $ODBCINI 定义）中。

注意：对于 64 位 Linux 系统，**Driver** 参数错误地引用了 32 位驱动程序。 当您在上一步骤中复制 `odbc.ini` 内容时，应相应地编辑该参数中的路径，例如:

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

7. 编辑 Netezza DSN 定义中的参数，以反映要使用的数据库。
8. 重新启动 IBM SPSS Modeler Server，并在客户端上测试使用 Netezza 数据库内挖掘节点。

在 SPSS Modeler 中启用集成

1. 在 IBM SPSS Modeler 主菜单中，选择

工具 > 选项 > 帮助应用程序。

2. 单击 **IBM Data Warehouse** 选项卡。

启用 IBM Data Warehouse Analytics Integration。 启用 IBM SPSS Modeler 窗口底部的“数据库建模”选用板（如尚未显示）并添加 IBM Data Warehouse 和 Netezza Data Mining 算法的建模节点。

IBM Data Warehouse 连接。 单击**编辑**按钮，并选择创建 ODBC 源时设置的 IBM Data Warehouse 连接字符串。 有关更多信息，请参阅 IBM Data Warehouse 管理控制台。

启用 SQL 生成和优化

由于使用超大型数据集的可能性，出于性能的原因，您应在 IBM SPSS Modeler 中启用 SQL 生成和优化选项。

1. 从 IBM SPSS Modeler 菜单中选择：

   工具 > 流属性 > 选项
2. 在导航窗格中单击**优化**选项。
3. 确认是否已启用**生成 SQL** 选项。 要使数据库建模正常发挥作用，此设置是必需的。
4. 选中**优化 SQL 生成**和**优化其他执行**（非严格必需但强烈推荐使用，以使性能更优）。

使用 IBM Netezza Analytics 和 IBM Data Warehouse 构建模型

每种受支持的算法均具有对应的建模节点。 您可以从节点选用板上的**数据库建模**选项卡中访问 IBM Data Warehouse 和 IBM Netezza 建模节点。

数据注意事项

数据源中的字段可以包含各种数据类型的变量，具体取决于建模节点。 在 IBM SPSS Modeler 中，数据类型称为测量级别。 建模节点的“字段”选项卡通过图标来指示其输入字段和目标字段所允许的测量级别类型。

目标字段 目标字段是您尝试预测其值的字段。 在可以指定目标的情况下，只能选择一个源数据字段作为目标字段。

记录标识字段 指定用于唯一标识每个观测值的字段。 例如，标识字段，比如客户标识。 如果源数据不包含标识字段，您可以通过“派生”节点来创建此字段，如下所示。

1. 选择源节点。
2. 在节点选用板的“字段选项”选项卡中，双击“派生”节点。
3. 在工作区上双击“派生”节点的图标可将其打开。
4. 在**派生字段**字段中，输入（例如） **标识**。
5. 在**公式**字段中，输入 `@INDEX` 并单击**确定**。
6. 将“派生”节点连接到流的其余部分。

注: 如果您使用 `NUMERIC(18,0)` 数据类型从 Netezza 数据库检索长数字数据，在导入期间 SPSS Modeler 有时会向上舍入数据。 为避免此问题，请使用 `BIGINT` 或 `NUMERIC(36,0)` 数据类型存储数据。

注: 由于存在针对可以使用的字段类型的限制，因此具有无类型测量级别和记录标识角色的字段不会显示在 Netezza 数据库内建模节点（例如，K-Means）中。

处理空值

如果输入数据包含空值，那么使用某些 Netezza 节点可能会导致产生错误消息或者长时间运行的流，因此我们建议移除包含空值的记录。 请使用以下方法。

1. 将“选择”节点附加到源节点。
2. 将“选择”节点的**模式**选项设置为**丢弃**。
3. 在**条件**字段中输入以下内容：

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

   确保包括每个输入字段。

4. 将“选择”节点连接到流的其余部分。

模型输出

包含 Data Warehouse 或 Netezza 建模节点的流有可能每次运行都产生略微不同的结果。 这是因为数据在建模之前被读入临时表，因此节点读取源数据的顺序并不始终相同。 但是，这种影响产生的差异可以忽略不计。

一般评论

- 在 IBM SPSS 协作和部署服务 中，不能使用包含 IBM Data Warehouse 或 IBM Netezza 数据库建模节点的流来创建评分配置。
- Data Warehouse 或 Netezza 节点构建的模型无法进行 PMML 导出或导入。

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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

目标。 请选择一个字段作为预测目标。 对于广义线性模型，请另查看此屏幕中的**试验**字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

服务器选项

在“服务器”选项卡上，指定用于构建模型的 IBM Data Warehouse 数据库。

IBM Data Warehouse Server 详细信息。 在这里，可以指定要用于模型的数据库的连接详细信息。

- **使用上游连接**。 （缺省）使用上游节点（例如“数据库源”节点）中指定的连接详细信息。 仅当所有上游节点都能够使用 SQL 回送功能时，此选项才有效。 在此情况下，无需将数据移出数据库，因为 SQL 完全实现所有的上游节点。
- **移动数据到连接**。 将数据移动到此处指定的数据库。 这样，即使数据位于另一个 IBM Data Warehouse 数据库或者另一供应商的数据库中，甚至位于平面文件中，也仍然可以进行建模。 另外，如果由于某个节点未执行 SQL 回送而导致数据已被提取，那么数据将移回到此处指定的数据库中。 单击**编辑**按钮以浏览并选择连接。

警告: IBM Netezza Analytics 和 IBM Data Warehouse 通常与非常大型的数据集配合使用。 在数据库之间传输大量数据，或者从数据库中取出或存入大量数据，可能非常耗时，应尽可能避免。

注: ODBC 数据源名称可有效地内嵌于每个 IBM SPSS Modeler 流中。 如果在一台主机上创建的流在另一台主机上执行，那么该数据源在两台主机上的名称必须相同。 另外，也可以在各个源或建模节点的“服务器”选项卡上选择另一个数据源。

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。 您还可以设置评分选项的缺省值。

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。

名称已使用时替换现有项。 如果选中此复选框，那么将覆盖所有的同名现有模型。

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。 有关这些选项的详细信息，请参阅特定模型块的“设置”选项卡的帮助主题。

管理模型

通过 SPSS Modeler 构建 IBM Netezza 或 IBM Data Warehouse 模型将在 SPSS Modeler 中创建一个模型，并在 IBM Data Warehouse 数据库中创建或替换一个模型。 这种 SPSS Modeler 模型将引用数据库服务器上存储的数据库模型的内容。 SPSS Modeler 可以通过将完全相同的生成模型键字符串存储在 SPSS Modeler 模型和 Netezza 或 Data Warehouse 模型中来执行一致性检查。

每个 Netezza 或 Data Warehouse 模型的模型名称都显示在“列出数据库模型”对话框的模型信息列下面。 SPSS Modeler 模型的模型名称在 SPSS Modeler 模型的“服务器”选项卡上显示为“模型键”（放置在流中时）。

“检查”按钮可用于检查 SPSS Modeler 模型和 Netezza 或 Data Warehouse 模型中的模型键是否匹配。 如果在 Netezza 后 Data Warehouse 中找不到具有相同名称的模型，或模型键不匹配，那么说明在构建 SPSS Modeler 模型之后删除或重新构建了该 Netezza 或 Data Warehouse 模型。

列出数据库模型

SPSS Modeler 提供了一个用于列出 IBM Data Warehouse 中存储的模型的对话框，并允许删除模型。 此对话框可以从“IBM 帮助应用程序”对话框以及 IBM Data Warehouse 和 IBM Netezza Data Mining 相关节点的构建、浏览和应用对话框中进行访问。 将显示每个模型的以下信息：

- 模型名称（模型的名称，用于对列表进行排序）。
- 所有者名称。
- 模型中使用的算法。
- 模型的当前状态；例如“已完成”。
- 模型的创建日期。

IBM Data WH 回归树

回归树是一种基于树的算法，它根据数字目标字段的值来重复分割观测值样本，以派生同一类型的子集。 与决策树一样，回归树将数据分解为子集，其中，树叶对应于足够小或足够均匀的子集。 通过选择分割来降低目标属性值的离差，以便采用树叶处的平均值来足够合理地预测它们。

IBM Data WH 回归树构建选项 - 树增长

可以设置用于树增长和树修剪的构建选项。

下列构建选项可用于树增长：

最大树深度。 这是在根节点以下树可以增长到的最大层数，即，递归分割样本的次数。 缺省值为 62，这是建模所允许的最大树深度。

注: 如果模型块中的查看器显示模型的文本表示，那么最多可以显示 12 层。

分割条件。 这些选项用于控制何时停止分割树。 如果您不想使用缺省值，请单击**定制**并更改这些值。

- **分割评估测量**。 这个类评估测量用于评估分割树的最佳位置。

  注: 目前唯一可能的选项是“方差”。
- **分割的最小改进**。 在树中创建新的分割之前，必须减少的最小杂质量。 树构建的目的是创建具有相似输出值的子组，以最大程度地减少每个节点中的杂质。 如果某个分支的最佳分割按小于分割标准所指定的数量来减少杂质，那么不会分割此分支。
- **分割的最小实例数**。 可以分割的最小记录数。 如果剩余的未分割记录数小于此数目，那么将不执行进一步分割。 您可以使用此字段来防止在树中创建小型子组。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

IBM Data WH 树构建选项 - 树修剪

您可以使用修剪选项来指定回归树的修剪标准。 修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

修剪度量。 修剪测量确保从树中移除树叶后，模型的估算精确度仍处于可接受的限度之内。 可以选择以下测量之一:

- **mse**。 均方误差 - （缺省）测量拟合线与数据点的接近程度。
- **r2.** R 平方 - 测量因变量的偏差比例（由回归模型解释）。
- **Pearson**。 Pearson 相关系数 - 测量正态分布的线性因变量之间的关系强度。
- **Spearman**。 Spearman 相关系数 - 检测根据 Pearson 相关性看起来较弱，但实际可能较强的非线性关系。

用于修剪的数据。 您可以使用部分或全部训练数据来估算新数据的预期精确性。 或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据**。 此选项（缺省）使用所有训练数据来估算模型精确度。
- **使用特定百分比的训练数据来进行修剪**。 使用此选项可以将数据分为两个集合，分别用于训练和修剪，从而使用此处指定的百分比来修剪数据。

  如果您要指定随机种子，以确保在您每次运行流时，数据以相同方式分区，请选择**复制结果**。 您可以在**用于修剪的种子**字段中指定一个整数，或单击**生成**来创建伪随机整数。
- **使用现有表中的数据**。 指定用于估算模型精确性的独立修剪数据集的表名称。 这种做法被视为比使用训练数据更为可靠。 不过，此选项可能导致从删除训练集中去除较大的数据子集，因而会降低决策树的质量。

Netezza 分裂式聚类

分裂式聚类是一种聚类分析方法，它通过重复运行算法，使聚类分裂为子聚类，直至达到规定的停止点。

聚类的构造以包含全部训练实例（记录）的单个聚类开始。 此算法的第一次迭代将数据集分为两个子聚类，后续迭代将这些子聚类划分为进一步的子聚类。 停止标准指定为最大迭代次数、数据集细分为的最大层数以及进行进一步分区所需的最小实例数。

产生的层次聚类树可以用于将实例从根聚类向下传播，以便对它们进行分类，如下例所示。

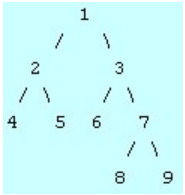

图 *3:* 分裂式聚类树示例

在每一层，根据实例与子聚类中心之间的距离来选择最佳匹配子聚类。

在应用了层次结构层 -1（缺省值）的情况下对实例进行评分时，此评分将只返回一个叶聚类，这是因为叶子由负数指定。 在示例中，这将是聚类 4、5、6、8 或 9 之一。 但是，例如，如果将层次结构层设置为 2，那么评分将返回根聚类下第二层的一个聚类，也就是 4、5、6 或 7。

Netezza 分裂式聚类字段选项

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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

Netezza 分裂式聚类构建选项

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。

距离测量。 这是用于测量数据点之间的距离的方法；距离越大，表示非相似性越大。 选项为：

- **欧式距离**。 （缺省）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **曼哈顿距离**。 两点之间的距离计算为其坐标之间的绝对差总和。
- **堪培拉距离**。 类似于曼哈顿距离，但对更加靠近原点的数据点更加敏感。
- **最大值**。 两点之间的距离计算为任何坐标尺寸之差的最大值。

最大迭代次数。 此算法通过执行同一过程的多次迭代来完成操作。 使用此选项可在指定的迭代次数后停止模型训练。

聚类树的最大深度。 这是数据集可以细分为的最大层数。

重复结果。 如果您要设置随机种子，请选中此复选框，这将允许您重复进行分析。 可指定一个整数或单击**生成**来创建伪随机整数。

用于分割的最小实例数。 可以分割的最小记录数。 如果剩余的未分割记录数小于此数目，那么将不执行进一步分割。 您可以使用此字段来防止在聚类树中创建非常小的子组。

IBM Data WH 广义线性

线性回归是一种广为接受的统计技术，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。 线性模型由于训练简单且模型应用方便，在构建各种真实世界现象的模型方面用途甚广。 然而，线性模型假设因变量（对象）呈正态分布，且自变量（预测变量）对因变量的影响是线性的。

线性回归在许多情况下非常有用，但是上述假设并不适用。 例如，对顾客从给定数量的商品中进行选择的行为建模时，因变量可能呈多项式分布。 同样，对年龄与收入的关系建模时，收入通常随年龄增长而增加，但这二者的关联却不像一条直线那么简单。

对于这些情况，可以使用广义线性模型。 广义线性模型扩展了线性回归模型，使因变量与预测变量之间通过特定的关联函数建立关联，由预测变量选择合适的函数。 另外，此模型允许因变量呈非正态分布，例如泊松分布。

此算法以迭代方式（次数可达指定的迭代次数）求出拟合度最佳的模型。 在计算最佳拟合时，误差由因变量的预测值和实际值之间的差异的平方和来表示。

IBM Data WH 广义线性模型字段选项

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”节点或者上游源节点的“类型”选项卡中的角色设置（例如目标或预测变量）。

使用定制字段分配。 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。 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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

目标。 请选择一个字段作为预测目标。

记录标识。 这是要用作唯一记录标识的字段。 此字段的值对于每个记录而言必须是唯一的（例如，客户标识号）。

实例权重。 指定字段以使用实例权重。 实例权重是每行输入数据的权重。 缺省情况下，假定所有输入记录具有相同的相对重要性。 可以通过向输入记录分配各项权重来更改重要性。 指定的字段必须包含每行输入数据的数字权重。

预测变量（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。

IBM Data WH 广义线性模型选项 - 常规

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。 您可以进行有关模型的多项设置，像是关联函数、输入字段的交互（如果有的话）以及设置评分选项的缺省值。

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。

字段选项。 可以指定用于构建模型的输入字段的角色。

常规设置。 这些设置关系到算法的停止标准。

- **最大迭代次数**。 算法最多进行迭代的次数；最小值为 1 次，缺省为 20 次。
- **最大误差 (1e)**。 最大误差值（以科学记数法表示），达到此值后，此算法应停止查找最佳拟合模型。 最小值为 0，缺省值为 -3，表示 1E-3 或 0.001。
- **无意义误差值阈值 (1e)**。 这是一个值（以科学记数法表示），所有小于此值的误差均被视为零值。 最小值为 -1，缺省值为 -7，表示误差值若低于 1E-7 （或 0.0000001），那么被视为不显著。

分布设置。 这些设置与因变量（目标变量）的分布相关。

- **响应变量的分布**。 分布类型；这是下列其中一项：**伯努利**（缺省），**高斯**、**泊松**、**二项**、**负二项式**、**Wald**（逆高斯）和**伽玛**。
- **参数**。 （仅限泊松或二项式分布）您必须在**指定参数**字段中指定下列其中一个选项：
  - 要自动从数据估算参数，请选择**缺省值**。
  - 要允许对分布拟似然进行优化，请选择 **Quasi**。
  - 要明确指定参数值，请选择**显式**。

  （仅限二项式分布）您必须按二项式分布的要求指定将用作试验字段的输入表列。 此列包含二项式分布的试验数。

  （仅限负二项式分布）您可以使用缺省值 -1 或指定不同的参数值。

关联函数设置。 这些设置与关联函数相关，后者用于使因变量与预测变量相关。

- **关联函数**。 要使用的函数，这是下列其中一项：**Identity**、**Inverse**、**Invnegative**、**Invsquare**、**Sqrt**、**Power**、**Oddspower**、**Log**、**Clog**、**Loglog**、**Cloglog**、**Logit**（缺省）、**Probit**、**Gaussit**、**Cauchit**、**Canbinom**、**Cangeom** 和 **Cannegbinom**。
- **参数**。 （仅限于 Power 或 Oddspower 关联函数），如果关联函数为 **Power** 或 **Oddspower**，您可以指定其参数值。 请选择是指定一个值，还是使用缺省值 1。

IBM Data WH 广义线性模型选项 - 交互

“交互”控制面板包含了选项，可以指定交互行为（即，输入字段间的乘积效应）。

列交互。 选择此选项框来指定输入字段的交互性。 若无交互行为，请将选项框留空。

通过从源列表中选择一个或多个字段，并将其拖动到交互列表，在模型中输入交互。 所创建的交互类型取决于将选项拖放到何种热点值。

- **主要**。 拖入的字段作为单独的主要交互，显示在交互列表的底部。
- **双向**。 所有可能配对的拖入字段作为双向交互，显示在交互列表的底部。
- **三向**。 所有可能配成三元组的拖入字段，均作为三向交互，显示在交互列表的底部。
- ***.** 所有已删除字段的组合会作为单个交互显示在交互列表的底部。

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。

对话框按钮

显示在右侧的按钮允许您对模型中使用的项进行更改。

图 *4: "*删除"按钮

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。

图 *5: "*重新排序"按钮

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。

图 *6: "*定制交互"按钮

添加定制项

您可以 *n1*x1*x1*x1..* 形式指定自定义交互。 从**字段**列表中选择一个字段，单击右方向箭头按钮将字段添加到**自定义项**，单击**按***，选择下一个字段，再单击右方向箭头按钮，以此类推。 当您已完成构建定制交互，可单击**添加项**将其返回"交互"面板。

IBM Data WH 广义线性模型选项 - 评分选项

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。 有关更多信息，请参阅主题 第 67 页的『IBM Data WH 广义线性模型块 -"设置"选项卡』。

- **包含输入字段**。 如果您需要将该输入字段连同预测值一同显示在模型输出中，请选择该选框。

IBM Data WH 决策树

决策树是代表分类模型的层次结构。 使用决策树模型，您可以开发分类系统，以便根据一组训练数据来预测未来观测值或者对其进行分类。 分类采用树结构形式，其中的分支表示分类中的分割点。 这些分割以递归方式将数据划分为子组，直至到达停止点为止。 停止点处的树节点称为**叶片**。 每片树叶分配一个标签（称为 **类标签**）给其子组或类成员。

实例权重和类权重

缺省情况下，假定所有输入记录和类具有相同的相对重要性。 通过对其中任意一个或全部项目的各个成员分配不同的权重，您可以对此进行更改。 这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。 权重可以使模型产生偏差，以便弥补那些在数据中没有得到很好表示的类别。 增加目标值的权重会增加该类别获得正确预测的百分比。

在“决策树”建模节点中，可以指定两种类型的权重。 **实例权重**分配一个权重给每一行输入数据。 对于大部分观测值，权重通常指定为 1.0，同时仅对那些比大部分观测值更加重要或更加不重要的观测值指定较大或较小的值，如下表所示。

表 *5:* 实例权重示例

记录标识	**目标**	**实例权重**
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

类权重对目标字段的每个类别分配一个权重，如下表所示。

表 *6:* 类权重示例

类	**类权重**
drugA	1.0
drugB	1.5

可以同时使用这两种类型的权重，在这种情况下，它们将相乘并用作实例权重。 因此，如果将之前的两个示例一起使用，那么此算法将使用下表所示的实例权重。

表 *7:* 实例权重计算示例

记录标识	**计算**	**实例权重**
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Netezza 决策树字段选项

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。

记录标识。 这是要用作唯一记录标识的字段。 该字段的值对于每个记录必须是唯一的（例如，客户标识号）。

实例权重。 在此处指定字段将允许您使用实例权重（每一行输入数据具有一个权重）来代替缺省的类权重（目标字段的每个类别具有一个权重）或者同时使用这两种权重。 在此处指定的字段必须是包含每行输入数据的数字权重的字段。 有关更多信息，请参阅主题 第 48 页的『实例权重和类权重』。

预测变量（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。

IBM Data WH 决策树构建选项

下列构建选项可用于树增长：

增长测量。 这些选项用于控制测量树增长的方式。

- **杂质测量**。 此测量用于评估分割树的最佳位置。 这是对数据子组或分段中的可变性的测量。 较小的杂质测量值指示这样一个组，该组中的大多数成员的标准或目标字段具有相似的值。

  受支持的测量为**熵**和**吉尼**。 这些测量基于分支的类别成员资格概率。

- **最大树深度**。 这是在根节点以下树可以增长到的最大层数，即，递归分割样本的次数。 此属性的缺省值为 10，您可以为此属性设置的最大值为 62。

  注: 如果模型块中的查看器显示模型的文本表示，那么最多可以显示 12 层。

分割条件。 这些选项用于控制何时停止分割树。

- **分割的最小改进**。 在树中创建新的分割之前，必须减少的最小杂质量。 树构建的目的是创建具有相似输出值的子组，以最大程度地减少每个节点中的杂质。 如果某个分支的最佳分割按小于分割标准所指定的数量来减少杂质，那么不会分割此分支。
- **分割的最小实例数**。 可以分割的最小记录数。 如果剩余的未分割记录数小于此数目，那么将不执行进一步分割。 您可以使用此字段来防止在树中创建小型子组。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。

- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

IBM Data WH 决策树节点 - 类权重

在这里，可以对各个类分配权重。 在缺省情况下，将向所有的类分配值 1，从而使它们具有相同的权重。 通过为不同类标签指定不同的数值权重，将引导算法相应地对特定类的训练集进行加权。

要更改权重，在**权重**列双击权重并进行所需更改。

值。 类标签集源自目标字段的可能值。

权重。 要分配给特定类的权重。 如果为某个类分配较高权重，那么模型将对此类比其他类更为敏感。

可以将类权重与实例权重配合使用。 有关更多信息，请参阅主题 第 48 页的『实例权重和类权重』。

IBM Data WH 决策树节点 - 树修剪

您可以使用修剪选项来指定决策树的修剪标准。 修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

修剪测量。 缺省的修剪测量为**精确度**，它确保在从树上去掉一个树叶后，模型的估算精确度仍保持在可接受的限度内。 如果您要在应用修剪时将类权重考虑在内，那么可以使用**加权精确度**选项。

用于修剪的数据。 您可以使用部分或全部训练数据来估算新数据的预期精确性。 或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据**。 此选项（缺省）使用所有训练数据来估算模型精确度。
- **使用特定百分比的训练数据来进行修剪**。 使用此选项可以将数据分为两个集合，分别用于训练和修剪，从而使用此处指定的百分比来修剪数据。

  如果您要指定随机种子，以确保在您每次运行流时，数据以相同方式分区，请选择**复制结果**。 您可以在**用于修剪的种子**字段中指定一个整数，或单击**生成**来创建伪随机整数。

- **使用现有表中的数据**。 指定用于估算模型精确性的独立修剪数据集的表名称。 这种做法被视为比使用训练数据更为可靠。 不过，此选项可能导致从删除训练集中去除较大的数据子集，因而会降低决策树的质量。

IBM Data WH 线性回归

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。 线性回归模型仅限于直接建模线性关系，但它相对简单，用于评分的数学公式也易于解释。 与其他更优化的回归算法产生的模型相比，线性模型快速、高效，并且简单易用，但其应用范围有限。

IBM Data WH 线性回归构建选项

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。

使用奇异值分解来对方程求解。 使用奇异值分解矩阵而不是原始矩阵，不但能够更有效地应对数字误差，并且可以加快计算过程。

在模型中包含截距。 包含截距可以提高解的总体准确性。

计算模型诊断。 此选项将导致对模型计算大量诊断信息。 结果将存储在矩阵或表 中以供稍后复查。 诊断选项包括 r 平方、残差平方和、估算方差、标准差、*p* 值和 *t* 值。

这些诊断信息与模型的有效性和可用性相关。 您应当针对底层数据运行其他诊断，以确保其满足线性假设。

IBM Data WH KNN

“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。

IBM Data WH KNN 模型选项 - 常规

在“模型选项 - 常规”选项卡上，您可以选择是指定模型名称，还是自动生成名称。 您还可以设置那些控制如何计算最近相邻元素数量的选项，并设置相关选项以获得增强的模型性能和准确度。

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。

相邻元素

距离测量。 这是用于测量数据点之间的距离的方法；距离越大，表示非相似性越大。 选项为：

- **欧式距离**。 （缺省）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **曼哈顿距离**。 两点之间的距离计算为其坐标之间的绝对差总和。
- **堪培拉距离**。 类似于曼哈顿距离，但对更加靠近原点的数据点更加敏感。
- **最大值**。 两点之间的距离计算为任何坐标尺寸之差的最大值。

最近邻元素数目 (k)。 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。

通过选择 *k*，您可以控制在防止过度拟合（这可能很重要，尤其对于“噪声”数据）和求解（针对类似实例产生不同预测结果）之间的平衡。 您通常需要针对每个数据集来调整 *k* 值，其典型值在 1 至几十之间。

增强性能和准确度

在计算距离之前标准化测量。 如果选中，该选项将标准化连续输入字段的测量结果，然后再计算距离值。

使用核心集以提高大型数据集的性能。 如果选中，该选项将针对大型数据集采用核心集抽样以加快计算过程。

IBM Data WH KNN 模型选项 - 评分选项

在“模型选项 - 评分选项”选项卡上，您可以设置评分选项的缺省值，并为单独类指定相对权重。

使其可用于评分

包括输入字段。 指定缺省情况下是否将输入字段包括在评分中。

类权重

如果您要更改单独类在构建模型中的相对重要性，请使用此选项。

注意：仅当您使用 KNN 进行分类时，此选项才处于启用状态。 如果您要执行回归（即，目标字段类型为连续），此选项将被禁用。

在缺省情况下，将向所有的类分配值 1，从而使它们具有相同的权重。 通过为不同类标签指定不同的数值权重，将引导算法相应地对特定类的训练集进行加权。

要更改权重，在**权重**列双击权重并进行所需更改。

值。 类标签集源自目标字段的可能值。

权重。 要分配给特定类的权重。 如果为某个类分配较高权重，那么模型将对此类比其他类更为敏感。

IBM Data WH K-Means

K-Means 节点实现 *k*-Means 算法，这提供了聚类分析的方法。 您可使用该节点来聚类数据集为不同的组。

此算法是基于距离的聚类算法，它依赖于距离度量（函数）以测量数据点之间的相似性。 根据使用的距离度量，将数据点分配到与之距离最近的聚类。

此算法通过执行同一基本过程的多次迭代完成操作，在该过程中，将每个训练实例分配到最接近（由应用于实例和聚类中心的指定距离函数确定）的聚类。 然后，重新计算所有聚类中心，作为分配给特定聚类实例的平均属性值向量。

IBM Data WH K-Means 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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Data WH K-Means 构建选项选项卡

通过设置构建选项，您可以根据具体用途定制模型的构建。

如果要使用缺省选项构建模型，请单击**运行**。

距离测量。 此参数定义用于测量数据点之间的距离的方法。 距离越大，表示非相似性越大。 请选择下列其中一个选项：

- **欧式**。 欧式距离测量是两个数据点之间的直线距离。

- **标准化欧式**。 标准化欧式距离测量类似于欧式距离测量，但已通过标准差的平方标准化。 与欧式距离测量不同的是，标准化欧式距离测量还具有尺度不变性。
- **马氏**。 马氏距离测量是考虑输入数据的相关性的广义欧式距离测量。 与标准化欧式距离测量一样，马氏距离测量具有尺度不变性。
- **曼哈顿**。 曼哈顿距离测量是计算为其坐标的绝对差总和的两个数据点之间的距离。
- **堪培拉**。 堪培拉距离测量类似于曼哈顿距离测量，但对距离原点越近的数据点越敏感。
- **最大值**。 最大值测量是计算为任何坐标尺寸之差的最大值的两个数据点之间的距离。

聚类数。 此参数定义要创建的聚类数。

最大迭代次数。 此算法执行同一过程的多次迭代。 此参数定义迭代次数，在此迭代次数后模型训练停止。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

复制结果。 如果要设置随机种子以重复进行分析，请选中此复选框。 您可以指定一个整数，也可以通过单击**生成**来创建一个伪随机整数。

IBM Data WH 朴素贝叶斯

朴素贝叶斯是广泛用于处理分类问题的算法。 此模型将所有建议预测变量视为相互独立，因此被称为朴素。 朴素贝叶斯是一种可伸缩的快速算法，用于计算各个属性与目标属性的组合的条件概率。 使用训练数据确定独立的概率。 给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

Netezza 贝叶斯网络

贝叶斯网络是一个模型，它显示数据集中的变量以及概率，或者显示这些变量之间有条件的独立性。 使用“Netezza 贝叶斯网络”节点，可以通过将记录的实测证据与实际常识相结合来构建概率模型，以使用表面上不相关的属性确定发生可能性。

Netezza 贝叶斯网络字段选项

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。

对于此节点，只有评分才需要目标字段，所以此字段未显示在此选项卡上。 您可以在“类型”节点、此节点的“模型选项”选项卡或模型块的“设置”选项卡上设置或更改目标。 有关更多信息，请参阅主题 第 62 页的『Netezza 贝叶斯网络块 -“设置”选项卡』。

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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

Netezza 贝叶斯网络构建选项

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。

基本索引。 为第一个属性（输入字段）分配的数字标识，以方便内部管理。

样本大小。 当属性数量过多并可能导致处理时间过长时，要采用的样本大小。

执行期间显示其他信息。 如果选中此框（缺省情况），那么将在消息对话框中显示附加的进度信息。

Netezza 时间序列

时间序列 是一个数值序列，以时间上前后接续的（但不必是规律的）点计量 - 例如，每日股票价格或每周销售数据。 分析此类数据有时会很有用，例如，用于突显某些行为，像是趋势或季节性变动（一项重复性的模式），或是通过过去的事件预测未来的行为的时候。

Netezza 时间序列支持下列时间序列算法。

- 谱分析
- 指数平滑法
- 自回归整合移动平均值 (ARIMA)
- 季节性趋势分解

这些算法将时间序列分解成一个趋势和一个季节性成分。 再对这些成分进行分析，以构建出一个可用于预测的模型。

谱分析用于识别时间序列中的周期性行为。 对于包含多个底层周期性的时间序列，或者在数据中存在大量随机噪声时，谱分析提供了最为明确的方法来识别周期性成分。 此方法通过将序列从时间域变换为一系列频率域，检测周期性行为的频率。

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。 采用指数平滑法，观测所造成的影响随时间推移而以指数级减少。 该方法一次预测一个点，当有新数据进入时再对预测作出调整，对资料的加入、趋势以及季节性变化作出整体性考虑。

ARIMA 模型提供了比指数平滑模型更复杂的方法进行趋势建模和季节性成分建模。 此方法涉及明确指定自回归阶数和移动平均值阶数以及差分度。

注：在实际应用中，如果要包括预测变量（该变量有助于解释要预测的序列的行为，例如邮寄的目录数或某公司网页的点击数），ARIMA 模型将非常有用。 而指数平滑模型在说明时间序列的行为时，并不试图解释其行为原因。

季节性趋势分解先将周期性行为从时间序列中删除，以便进行趋势分析，之后再为趋势选择一个基本形状，例如一个二次函数。 这些基本形状带有若干参数，应为这些参数值确定一个值，以尽量减少平均残差均方误差（即时间序列拟合值与观测值之间的差异）。

Netezza 时间序列值的插值

插值是估算时间序列中缺失的数据并插补一个值的过程。

如果时间序列的时间间隔有规律，但某些值不存在，那么可以使用线性插值来估算这些缺失值。 考虑如下示例序列中某机场航厦每月的乘客抵达人数。

表 *8:* 某航厦的每月抵达人数

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

在此例中，通过线性插值可以估算出第 5 个月的缺失值为 3,650,000 （第 4 个月与第 6 个月的中间点）。

不规律间隔的处理方式有所不同。 考虑如下序列中的温度读数。

表 *9:* 温度读数

日期	时间	温度
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

这里，我们有 3 天内从 3 个点所取得的一系列读数，但除了少数读数外，大部分读数的获取时间并不相同。另外，其中只有 2 天是连续的。

这种情况可以通过下列两种方法中的一种来处理：计算汇总，或者确定步长。

汇总可能是根据对数据语义的了解，使用公式计算得出的每日汇总。 执行这一步会得到如下的数据集。

表 *10:* 温度读数（汇总）

日期	时间	温度
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	null
2011-07-27	24:00	72

另外，此算法可以将该序列视为差异序列以确定适当的步长。 在此例中，算法所确定的步长可能是 8 个小时，这样会得到如下结果。

表 *11:* 使用步长计算的温度读数

日期	时间	温度
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	

表 *11:* 使用步长计算的温度读数 *(*继续*)*		
日期	**时间**	**温度**
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

在这里，只有 4 个读数与原始测量值对应，但借着原始序列中的其他已知值，缺失的值可再次通过插值计算出来。

Netezza 时间序列字段选项

在“字段”选项卡上，指定源数据输入字段的角色。

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。

目标。 请选择一个字段作为预测目标。 这必须是测量级别为“连续”的字段。

（预测变量）时间点。 （必填）这是包含时间序列的日期或时间值的输入字段。 这必须是测量级别为“连续”或“分类”且数据存储类型为日期、时间、时间戳记或数字的字段。 此处指定的字段的数据存储类型同时也定义了此建模节点的其他选项卡上某些字段的输入类型。

（预测变量）时间序列标识（依据）。 包含时间序列标识的字段。如果输入项包含一个以上的时间序列，那么使用这个字段。

Netezza 时间序列构建选项

构建选项分两个级别：

- 基本 - 设定算法选择、插值以及所采用的时间范围。
- 高级 - 设置预测

本节描述基本选项。

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。

算法

这些是有关所要采用的时间序列算法的设置。

算法名称。 选择要使用的时间序列算法。 可选的算法包括**谱分析**、**指数平滑法**（缺省）、**ARIMA** 或**季节趋势分解**。 有关更多信息，请参阅主题 第 54 页的『Netezza 时间序列』。

趋势。 （仅限于指数平滑法）如果时间序列呈现出一种趋势，那么简单指数平滑法效果不佳。 若有趋势，使用该字段来指定它，以使算法可将它纳入考量。

- **系统确定**。 （缺省）系统尝试为该参数找到最佳值。
- **无 (N)**。 时间序列未呈现趋势。
- **加法 (A)**。 随着时间推移而稳定增加的趋势。
- **阻尼加法 (DA)**。 随着时间推移最终会消失的加法趋势。
- **乘法 (M)**。 该趋势也是随时间而增加，但速度通常比稳定加法趋势快。
- **阻尼乘法 (DM)**。 随着时间推移最终会消失的乘法趋势。

季节性。 （仅限于指数平滑法）使用该字段指定时间序列中的数据是否呈现季节性特征。

- **系统确定**。 （缺省）系统尝试为该参数找到最佳值。
- **无 (N)**。 时间序列未呈现季节性模式。
- **加法 (A)**。 季节性浮动模式呈现随时间推移稳定上行的趋势。

- **乘法 (M)**。 具有与加法季节性相同的特点，但除此之外，其季节性浮动的振幅（高低点间的距离）围绕着总体的上行趋势而上下浮动。

将系统确定的设置用于 ARIMA。 （仅限于 ARIMA）如果您希望由系统来确定 ARIMA 算法的设置，请选择此选项。

指定。 （仅限于 ARIMA）选择此选项并单击按钮可以手动指定 ARIMA 设置。

插值法

时间序列源数据包含缺失值时，选择一种方法来插入估算值以填补数据中的间隔。 有关更多信息，请参阅主题 第 54 页的『Netezza 时间序列值的插值』。

- **线性**。 时间序列的间隔有规律，仅仅是某些值缺失时，请选择此方法。
- **指数样条**。 把数据值以高速增加或减少的已知点拟合成一条平滑曲线。
- **三次样条**。 将已知数据点拟合成一条平滑曲线来估算缺失的值。

时间范围

在此可选择是否使用全范围的时间序列数据，或时间序列数据的一个连续的子集来建立模型。 这些字段的有效输入由“字段”选项卡上对“时间点”指定的字段的数据存储类型定义。 有关更多信息，请参阅主题 第 56 页的『Netezza 时间序列字段选项』。

- **使用数据中提供的最早和最新时间**。 如果您想要使用全范围的时间序列数据，请选择此选项。
- **指定时间窗口**。 如果您希望只使用时间序列的一部分，请选择此选项。 使用**最早时间（自）**与**最晚时间（至）**字段来界定边界。

ARIMA 结构

指定 ARIMA 模型中各种非季节性成分及季节性成分的值。 在每种情况下，将运算符设置为 **=**（等于）或 **<=**（小于或等于），然后在相邻字段中指定该值。 指定度数的所有值都必须为非负整数。

非季节性。 模型中各非季节性成分的值。

- **自相关度 (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，自回归阶为 2 时，指定序列中过去两个时段的值用于预测当前值。
- **派生 (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。 差分的阶与序列趋势度相对应，一阶差分导致线性趋势，二阶差分导致二次趋势，等等。
- **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。 季节性自相关 (SP)、派生 (SD) 以及移动平均值 (SQ) 成分扮演与其非季节性对应成分相同的角色。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以前序列值的影响。 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。

仅当在数据中检测到季节性趋势时，或您从“高级”选项卡中指定了“周期设置”时，才需用到季节性设置。

Netezza 时间序列构建选项 - 高级

您可以使用高级设置来指定预测选项。

对模型构建选项使用系统确定的设置。 如果您希望由系统来作高级设置，请选择此选项。

指定。 如果您希望手动指定高级选项，请选择此选项。 （算法为谱分析时，该选项不可选。）

- **时间段/时间段的单位**。 这是一个时间周期，在此之后，时间序列的一些特征行为不断重复。 例如，对于一个每周销售数字，您可以指定周期为 1，单位为 **星期**。 **周期**必须为非负整数；**周期单位**可以是**毫秒**、**秒**、**分**、**小时**、**天**、**星期**、**季**或者**年**之一。 如果未设置**周期**，或时间类型不为数字，请勿设置**周期单位**。但是，如果您指定**周期**，您必须也指定**周期单位**。

用于预测的设置。 您可以选择预测特定时间点之前的整个时段，或者预定特定时间点的情况。 这些字段的有效输入由“字段”选项卡上对“时间点”指定的字段的数据存储类型定义。 有关更多信息，请参阅主题 第 56 页的『Netezza 时间序列字段选项』。

- **预测范围**。 仅当您只想指定预测结束点时，才应选择此选项。 预测将到此时间点为止。
- **预测时间**。 选择此选项可以指定一个或多个时间点，作为预测时间点。 单击**添加**在时间点的表中增加一行。 要删除一行，请选定该行，再单击**删除**。

Netezza 时间序列模型选项

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。 您还可以设置模型输出选项的缺省值。

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。

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。

- **在结果中包含历史值**。 按照缺省，模型输出不包含数据的历史数据值（之前用来进行预测的值）。 选择此复选框以包含这些值。
- **在结果中包含内插值**。 如果您选择在输出中包括历史值，且希望同时包括内插值（如果有），请选中此框。 请注意，插值仅对历史数据起作用，所以如果未选择**在输出中包含历史记录值**，那么此框不可用。 有关更多信息，请参阅主题 第 54 页的『Netezza 时间序列值的插值』。

IBM Data WH 二阶

“二阶”节点实现了“二阶”算法，该算法提供了一种对大型数据集进行数据聚类的方法。

您可以使用此节点在考虑可用资源（例如内存和时间约束）的情况下对数据进行聚类。

“二阶”算法是一种数据库挖掘算法，此算法以如下方式对数据进行聚类：

1. 创建聚类特征 (CF) 树。 这个高度平衡的树存储聚类特征，以便执行类似输入记录成为相同树节点的组成部分的分层聚类。
2. CF 树的叶子在内存中以分层方式进行聚类，以生成最终的聚类结果。 最佳聚类数目是自动确定的。 如果您指定了最大聚类数目，那么将确定所指定限制之内的最佳聚类数目。
3. 聚类结果在第二个步骤中进行优化，在该步骤中，将对数据应用与 K-Means 算法类似的算法。

IBM Data WH 二阶字段选项

通过设置字段选项，您可以指定使用上游节点中定义的字段角色设置。 并且，还可以手动进行字段分配。

选择一个项目。 选择此选项表示使用上游“类型”节点中的角色设置或者上游源节点的“类型”选项卡中的角色设置。 例如，角色设置包括目标和预测变量。

使用定制字段分配。 如果您希望手动指定目标、预测变量和其他角色，请选择此选项。

字段。 使用箭头可以将此列表中的项手动分配到右侧的角色字段。 图标表示每个角色字段的有效测量级别。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Data WH 二阶构建选项

通过设置构建选项，您可以根据具体用途定制模型的构建。

如果要使用缺省选项构建模型，请单击**运行**。

距离测量。 此参数定义用于测量数据点之间的距离的方法。 距离越大，表示非相似性越大。 选项为：

- **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**。 欧式距离测量是两个数据点之间的直线距离。
- **标准化欧式**。 标准化欧式距离测量类似于欧式距离测量，但已通过标准差的平方标准化。 与欧式距离测量不同的是，标准化欧式距离测量还具有尺度不变性。

聚类数。 此参数定义要创建的聚类数。 选项为：

- **自动计算聚类数**。 聚类数目自动计算。 您可以在**最大值**字段中指定最大聚类数目。
- **指定聚类数**。 指定应该创建的聚类数。

统计。 此参数定义将多少统计信息包括在模型中。 选项为：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

复制结果。 如果要设置随机种子以重复进行分析，请选中此复选框。 您可以指定一个整数，也可以通过单击**生成**来创建一个伪随机整数。

IBM Data WH PCA

主成分分析 (PCA) 是一种强大的数据削减技术，用于降低数据复杂性。 PCA 可以找出输入字段的线性组合，这些组合能够最好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（不相关）。 其目标在于找到有效概括原始输入字段集中的信息的一小部分派生字段（主成分）。

注: 如果使用小写字段名称，那么在对模型评分时可能会发生错误。 这是已知的 Db2 Data Warehouse 缺陷，变通方法是在评分之前将所有字段重命名为大写。

IBM Data WH PCA 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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Data WH PCA 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。

在计算 PCA 之前执行数据中心化。 如果选中此选项（缺省情况），那么将在进行分析前执行数据中心化（也称为“均值消去法”）。 为了确保第一主成分描述最大方差的方向，有必要进行数据中心化，否则该成分可能更接近于数据的平均值。 如果已采用这种方式来准备数据，您通常可以取消选中此选项以提升性能。

在计算 PCA 之前执行数据换算。 该选项将在分析之前执行数据换算。 这样做可以减低以不同单位测量不同变量时的分析任意性。 作为最简单的形式，可以通过将每个变量除以其标准差来完成数据换算。

使用不太准确但速度更快的方法来计算 PCA。 该选项将导致算法使用低准确度但快速的方法 (forceEigensolve) 来寻找主成分。

管理 IBM Data WH 和 Netezza 模型

IBM Data Warehouse 和 IBM Netezza Analytics 模型像其他 IBM SPSS Modeler 模型一样添加到工作区和“模型”选用板中，并以几乎相同的方式使用。 但是，也有几点重大差异，比如 IBM SPSS Modeler 中创建的每个 IBM Data Warehouse 或 IBM Netezza Analytics 模型实际引用的是存储在数据库服务器上的模型。 因此，要使流正常工作，必须将其连接到创建模型所在的数据库，并且模型表未被外部进程修改。

对 IBM Data Warehouse 和 IBM Netezza Analytics 模型评分

在工作区上使用金色模型块图标来代表模型。 模型块的主要用途是对数据进行评分以生成预测，或者允许进一步分析模型属性。 评分以一个或多个附加数据字段的形式添加，如本节的随后内容所述，通过将一个“表”节点附加到模型块并运行流的此分支，可以使这些字段可见。 某些模型块对话框（例如，决策树或回归树的模型块对话框）还包含“模型”选项卡，其中提供了模型的直观表示。

这些附加的字段由目标字段名中添加的前缀 `$<id>-` 加以区分，其中 `<id>` 取决于模型，用于标识所添加的信息类型。 在每个模型块的主题中描述了不同的标识。

要查看评分，按以下步骤操作：

1. 将表节点附加到模型块。
2. 打开表节点。
3. 单击**运行**。
4. 滚动到表输出窗口的右侧，以查看附加字段及其评分。

IBM Data WH 和 Netezza 模型块“服务器”选项卡

在“服务器”选项卡上，可以设置模型评分的服务器选项。 您可以继续使用在上游指定的服务器连接，也可将数据移动到在此指定的其他数据库。

IBM Data Warehouse Server 详细信息。 在这里，可以指定要用于模型的数据库的连接详细信息。

- **使用上游连接**。 （缺省）使用上游节点（例如“数据库源”节点）中指定的连接详细信息。 仅当所有上游节点都能够使用 SQL 回送功能时，此选项才有效。 在此情况下，无需将数据移出数据库，因为 SQL 完全实现所有的上游节点。
- **移动数据到连接**。 将数据移动到此处指定的数据库。 这样，即使数据位于另一个 IBM Data Warehouse 数据库或者另一供应商的数据库中，甚至位于平面文件中，也仍然可以进行建模。 另外，如果由于某个节点未执行 SQL 回送而导致数据已被提取，那么数据将移回到此处指定的数据库中。 单击**编辑**按钮以浏览并选择连接。

警告: IBM Netezza Analytics 和 IBM Data Warehouse 通常与非常大型的数据集配合使用。 在数据库之间传输大量数据，或者从数据库中取出或存入大量数据，可能非常耗时，应尽可能避免。

模型名称。 模型的名称。 该名称的显示仅供参考；无法在此对其进行更改。

IBM Data WH 决策树模型块

决策树模型块显示建模操作的输出，还允许您设置一些选项来为模型评分。

在您运行包含决策树模型块的流时，该节点会缺省添加一个新的字段，其名称将从目标名称派生。

表 *12:* 决策树的模型评分字段

新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。

如果您在建模节点或模型块上选择选项**计算所分配类用于记录评分的概率**，并运行流，那么会再添加一个字段。

表 *13:* 决策树的模型评分字段 - 更多	
新增字段的名称	**含义**
$IP-*target_name*	预测结果的置信度值 (0.0 - 1.0)。

IBM Data WH 决策树块 -“模型”选项卡

模型选项卡以图形格式显示决策树模型的预测变量重要性。 条形的长度表示预测变量的重要性。

注: 使用 IBM Netezza Analytics V2.x 或更早版本时，决策树模型的内容仅以文本格式显示。

对于这些版本，将显示以下信息：

- 文本的每一行都对应于一个节点或叶子。
- 缩进反映树的层。
- 对于节点，将显示分割条件。
- 对于叶子，将显示所分配的类标签。

IBM Data WH 决策树块 -“设置”选项卡

通过“设置”选项卡，可以设置模型评分的某些选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

计算所分配类用于记录评分的概率。 （仅限于决策树和朴素贝叶斯）如果选中此选项，那么表示附加的建模字段包括置信度（即，概率）字段和预测字段。 如果您取消选中该复选框，将只生成预测字段。

使用确定性输入数据。 如果选中此选项，那么将确保任何对同一个视图多次运行遍历的 Netezza 算法都将使用同一组数据进行遍历。 如果取消选中此复选框以表明正在使用非确定性数据，那么将创建一个临时表以存放要处理的数据输出，例如由分区节点生成的删除；创建模型后，将删除这个表。

IBM Data WH 决策树块 -“查看器”选项卡

查看器选项卡以 SPSS Modeler 显示其决策树模型的方式显示树模型的树表示。

注: 如果模型是使用 IBM Netezza Analytics V2.x 或之前版本构建的，那么**查看器**选项卡为空。

IBM Data WH 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 模型块的流时，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。 名为 $KM-K-Means 的新字段用于聚类成员信息，而名为 $KMD-K-Means 的新字段用于与聚类中心的距离。

IBM Data WH K-Means 块 -“模型”选项卡

模型选项卡包含各种图形视图，这些视图显示聚类的汇总统计量和字段分布。 您可以从模型中导出数据，也可以将视图作为图形导出。

使用 IBM Netezza Analytics V2.x 或更早版本时，或者使用马氏距离作为距离度量来构建模型时，K-Means 模型的内容仅以文本格式显示。

对于这些版本，将显示以下信息：

- **汇总统计**。 对于最小聚类和最大聚类，汇总统计量显示记录数量。 另外，汇总统计量还显示这些聚类所拥有的数据集的百分比。 该列表还显示了最大聚类与最小聚类的比值。
- **聚类汇总**。 聚类汇总列出了算法所创建的聚类。 对于每个聚类，该表显示了该聚类中的记录数量，以及这些记录离聚类中心的平均距离。

IBM Data WH K-Means 块 -“设置”选项卡

通过“设置”选项卡，可以设置模型评分的某些选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

距离测量。 这是用于测量数据点之间的距离的方法；距离越大，表示非相似性越大。 选项为：

- **欧式距离**。 （缺省）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **曼哈顿距离**。 两点之间的距离计算为其坐标之间的绝对差总和。
- **堪培拉距离**。 类似于曼哈顿距离，但对更加靠近原点的数据点更加敏感。
- **最大值**。 两点之间的距离计算为任何坐标尺寸之差的最大值。

Netezza 贝叶斯网络模型块

贝叶斯网络模型块提供了一种设置模型评分选项的方法。

在您运行包含贝叶斯网络模型块的流时，该节点会添加一个新的字段，其名称将从目标名称派生。

表 *14:* 贝叶斯网络的模型评分字段	
新增字段的名称	**含义**
$BN-*target_name*	当前记录的预测值。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这个附加字段。

Netezza 贝叶斯网络块 -“设置”选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

目标。 如果您要对不同于当前目标的目标字段评分，在此选择新的目标。

记录标识。 如果未指定记录标识字段，在此选择要使用的字段。

预测类型。 您要使用的预测算法的变异：

- **最佳（最相关的相邻元素）**。 （缺省）使用最相关的相邻元素节点。
- **相邻元素（相邻元素的加权预测）**。 使用所有相邻元素节点的加权预测。
- **NN 相邻元素（非空相邻元素）**。 与上一选项相同，不同之处在于它将忽略具有空值的节点（即，对于要计算预测结果的实例，该节点对应的属性存在缺失值）。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

IBM Data WH 朴素贝叶斯模型块

朴素贝叶斯模型块提供了一种设置模型评分选项的方法。

在您运行包含朴素贝叶斯模型块的流时，该节点会缺省添加一个新的字段，其名称将从目标名称派生。

表 *15:* 朴素贝叶斯的模型评分字段 - 缺省	
新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。

如果您在建模节点或模型块上选择选项**计算所分配类用于记录评分的概率**，并运行此流，那么会再添加两个字段。

表 *16:* 朴素贝叶斯的模型评分字段 - 更多	
新增字段的名称	**含义**
$IP-*target_name*	实例类的 Bayesian 分子（即，先验类概率与条件实例属性值概率的乘积）。
$ILP-*target_name*	后者的自然对数。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

IBM Data WH 朴素贝叶斯块 -“设置”选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

计算所分配类用于记录评分的概率。 （仅限于决策树和朴素贝叶斯）如果选中此选项，那么表示附加的建模字段包括置信度（即，概率）字段和预测字段。 如果您取消选中该复选框，将只生成预测字段。

提高较小或严重失衡数据集的概率准确性。 在计算概率时，该选项将调用 *m* 估算技术，以避免在估算期间出现零概率。 这种类型的概率估算可能速度较慢，但可针对较小或严重失衡数据集提供更好的结果。

IBM Data WH KNN 模型块

KNN 模型块提供了一种设置模型评分选项的方法。

在您运行包含 KNN 模型块的流时，该节点会添加一个新的字段，其名称将从目标名称派生。

表 *17: KNN* 模型评分字段	
新增字段的名称	**含义**
$KNN-*target_name*	当前记录的预测值。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这个附加字段。

IBM Data WH KNN 块 -“设置”选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

距离测量。 这是用于测量数据点之间的距离的方法；距离越大，表示非相似性越大。 选项为：

- **欧式距离**。 （缺省）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **曼哈顿距离**。 两点之间的距离计算为其坐标之间的绝对差总和。
- **堪培拉距离**。 类似于曼哈顿距离，但对更加靠近原点的数据点更加敏感。
- **最大值**。 两点之间的距离计算为任何坐标尺寸之差的最大值。

最近邻元素数目 (k)。 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。

通过选择 *k*，您可以控制在防止过度拟合（这可能很重要，尤其对于“噪声”数据）和求解（针对类似实例产生不同预测结果）之间的平衡。 您通常需要针对每个数据集来调整 *k* 值，其典型值在 1 至几十之间。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

在计算距离之前标准化测量。 如果选中，该选项将标准化连续输入字段的测量结果，然后再计算距离值。

使用核心集以提高大型数据集的性能。 如果选中，该选项将针对大型数据集采用核心集抽样以加快计算过程。

Netezza 分裂式聚类模型块

分裂式聚类模型块提供了一种设置模型评分选项的方法。

运行包含分裂式聚类模型块的流时，该节点将添加两个新字段，这两个新字段的名称从目标名称派生。

表 *18:* 分裂式聚类模型评分字段

新增字段的名称	**含义**
$DC-*target_name*	当前记录所分配到的子聚类的标识。
$DCD-*target_name*	从当前记录到子聚类中心的距离。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

Netezza 分裂式聚类块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

距离测量。 这是用于测量数据点之间的距离的方法；距离越大，表示非相似性越大。 选项为：

- **欧式距离**。 （缺省）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **曼哈顿距离**。 两点之间的距离计算为其坐标之间的绝对差总和。
- **堪培拉距离**。 类似于曼哈顿距离，但对更加靠近原点的数据点更加敏感。
- **最大值**。 两点之间的距离计算为任何坐标尺寸之差的最大值。

应用的层次结构层。 应用于数据的层次结构层。

IBM Data WH PCA 模型块

PCA 模型块提供了一种设置模型评分选项的方法。

在您运行包含 PCA 模型块的流时，该节点会缺省添加一个新的字段，其名称将从目标名称派生。

表 *19: PCA* 模型评分字段

新增字段的名称	**含义**
$F-*target_name*	当前记录的预测值。

如果您在建模节点或模型块上的**主成分数...**字段中指定大于 1 的值， 并运行流，那么节点将为每个成分添加一个新字段。 在此情况下，字段名称带有后缀 *-n*，其中 *n* 是成分的编号。 例如，如果模型名为 *pca* 且包含三个成分，那么新字段将命名为 *$F-pca-1*、*$F-pca-2* 和 *$F-pca-3*。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

注: 如果使用小写字段名称，那么在对模型评分时可能会发生错误。 这是已知的 Db2 Data Warehouse 缺陷，变通方法是在评分之前将所有字段重命名为大写。

IBM Data WH PCA 块 -“设置”选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

要在预测中使用的主成分个数。 您要用来减小数据集的主成分个数。 该值不得超过属性（输入字段）数量。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

Netezza 回归树模型块

回归树模型块提供了一种设置模型评分选项的方法。

在您运行包含回归树模型块的流时，该节点会缺省添加一个新的字段，其名称将从目标名称派生。

表 *20:* 回归树的模型评分字段	
新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。

如果您在建模节点或模型块上选择选项**计算估算方差**，并运行流，那么会再添加一个字段。

表 *21:* 回归树的模型评分字段 - 更多	
新增字段的名称	**含义**
$IV-*target_name*	所预测的值的估算方差。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

Netezza 回归树块 - 模型选项卡

模型选项卡以图形格式显示回归树模型的预测变量重要性。 条形的长度表示预测变量的重要性。

注: 使用 IBM Netezza Analytics V2.x 或更早版本时，回归树模型的内容仅以文本格式显示。

对于这些版本，将显示以下信息：

- 文本的每一行都对应于一个节点或叶子。
- 缩进反映树的层。
- 对于节点，将显示分割条件。
- 对于叶子，将显示所分配的类标签。

Netezza 回归树块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

计算估算方差。 表示是否应在输出中包含所分配类的方差。

Netezza 回归树块 -“查看器”选项卡

查看器选项卡以 SPSS Modeler 显示其回归树模型的方式显示树模型的树表示。

注: 如果模型是使用 IBM Netezza Analytics V2.x 或之前版本构建的，那么**查看器**选项卡为空。

IBM Data WH 线性回归模型块

线性回归模型块提供了一种设置模型评分选项的方法。

在您运行包含线性回归模型块的流时，该节点会添加一个新的字段，其名称将从目标名称派生。

表 *22:* 线性回归的模型评分字段	
新增字段的名称	**含义**
$LR-*target_name*	当前记录的预测值。

IBM Data WH 线性回归块 -“设置”选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。 如果选中此选项，那么将向下游传递所有原始输入字段，从而对每行数据追加一个或多个附加的建模字段。 如果您取消选中该复选框，那么只会传递记录标识字段和额外建模字段，而使流能够更加快速地运行。

Netezza 时间序列模型块

此模型块使您能够访问时间序列建模操作的输出。 输出项由下列字段组成。

表 *23:* 时间序列模型输出字段

字段	描述
TSID	时间序列的标识；这是在建模节点的“字段”选项卡上对“时间序列标识”指定的字段中的内容。 有关更多信息，请参阅主题 第 56 页的『Netezza 时间序列字段选项』。
TIME	当前时间序列内的时间周期。
HISTORY	数据曾使用过的历史数据值（曾用于预测）。 仅当模型块的“设置”选项卡中的**在输出中包含历史记录值**选项被选中时，该字段才被包含在内。
$TS-INTERPOLATED	内插值（如果使用）。 仅当模型块的“设置”选项卡中的**在输出中包含插补的值**选项被选中时，该字段才被包含在内。 插值是建模节点的“构建选项”选项卡上的一个选项。
$TS-FORECAST	时间序列的预测值。

要查看模型输出，请从节点选用板的“输出”选项卡将一个“表”节点附加到模型块，并运行这个“表”节点。

Netezza 时间序列块 - "设置”选项卡

在“设置”选项卡中，您可以指定选项来自订模型输出。

模型名称。 模型名称，在建模节点的“模型选项”选项卡上定义。

其他选项与建模节点的“建模选项”选项卡上的选项相同。

IBM Data WH 广义线性模型块

此模型块使您能够访问建模操作的输出。

运行包含广义线性模型块的流时，该节点将添加一个新的字段，其名称是从目标名称派生。

表 *24:* 广义线性的模型评分字段

新增字段的名称	含义
$GLM-*target_name*	当前记录的预测值。

“模型”选项卡显示各种与模型有关的统计量。

输出项由下列字段组成。

表 *25:* 广义线性模型输出字段

输出字段	描述
参数	模型使用的参数（即，预测变量）。 这些是数值和名义列，以及截距（回归模型中的常数项）。

表 *25:* 广义线性模型输出字段 *(*继续*)*	
输出字段	**描述**
Beta	相关系数（即，模型的线性成分）。
标准误差	Beta 的标准差。
测试	用于评估参数有效性的检验统计量。
p 值	假定参数显著时，误差的概率。
残差汇总	
残差类型	显示汇总值的预测残差类型。
RSS	残差的值。
df	残差的自由度。
p 值	误差的概率。 高值表示拟合度差的模型；低值表示拟合度佳。

IBM Data WH 广义线性模型块 -“设置”选项卡

在“设置”选项卡中，您可以自订模型的输出。

此选项与建模节点的评分选项中显示的选项相同。 有关更多信息，请参阅主题 第 48 页的『IBM Data WH 广义线性模型选项 - 评分选项』。

IBM Data WH 二阶模型块

运行包含二阶模型块的流时，该节点将添加两个新字段，这两个字段包含聚类成员信息以及与该记录所分配到的聚类中心的距离。 名为 $TS-Twostep 的新字段用于聚类成员信息，而名为 $TSP-Twostep 的新字段用于与聚类中心的距离。

IBM Data WH 二阶块 -“模型”选项卡

模型选项卡包含各种图形视图，这些视图显示聚类的汇总统计量和字段分布。 您可以从模型中导出数据，也可以将视图作为图形导出。

第 6 章 使用 IBM Db2 for z/OS 进行数据库建模

IBM SPSS Modeler 和 IBM Db2 for z/OS

SPSS Modeler 支持与 Db2 for z/OS 进行集成，这提供了在 Db2 for z/OS 服务器上运行高级分析的能力。 您可以通过 SPSS Modeler 图形用户界面以及面向工作流程的开发环境访问这些功能。 这样，就可以直接在 Db2 for z/OS 环境中运行数据挖掘算法，从而利用 IBM Db2 Analytics Accelerator。

SPSS Modeler 支持与 Db2 for z/OS 中的下列算法集成。

- 决策树
- K-Means
- 朴素贝叶斯
- 回归树
- 二阶

与 IBM Db2 for z/OS 进行集成的需求

以下是使用 Db2 for z/OS 和 IBM Db2 Analytics Accelerator for z/OS 执行数据库内建模的先决条件。 为了确保满足这些条件，您可能需要咨询数据库管理员。 有关详细需求，包括受支持的版本，请参阅软件产品兼容性报告。

- 以本地方式运行或者针对 Windows 或 UNIX 上的 SPSS Modeler Server 安装来运行的 IBM SPSS Modeler
- Db2 for z/OS 与 Db2 Analytics Accelerator for z/OS
- IBM SPSS Data Access Pack
- 在运行 SPSS Modeler Server 的服务器上，下列其中一个系统：
  - IBM Db2 Data Server Driver for ODBC and CLI
  - 带有配置用于 Db2 for z/OS 的 ODBC 数据源的任何 Db2 for Linux®、UNIX and Windows 版本
- Db2 Connect for System z 许可证
- 在 SPSS Modeler 中启用 SQL 生成和优化
- Db2 z/OS 数据库内数据挖掘需要仅加速器表 (AOT) 或加速表以及 INZA 支持。 IDAA 5.1 中引入了 IDAA INZA。这意味着 Db2 z/OS 数据库数据挖掘节点不适用于先前版本的 IDAA。

  如果在 Modeler 中使用 IDAA 支持的 DSN，在使用 DSN 的“数据库源”节点中返回的表列表中，仅显示 AOT 或加速表。

启用 IBM Db2 Analytics Accelerator for z/OS 集成

启用 Db2 Analytics Accelerator for z/OS 集成的过程由下列步骤组成：

- 配置 Db2 for z/OS 和 Db2 Analytics Accelerator for z/OS
- 创建 ODBC 源
- 在 IBM SPSS Modeler 中启用 IBM Db2 for z/OS 集成
- 在 SPSS Modeler 中启用 SQL 生成和优化
- 启用 IBM SPSS Modeler Server Scoring Adapter for Db2 for z/OS
- 在 IBM SPSS Modeler 中使用 IBM Db2 Client 配置 DSN

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

以下 Web 站点上的内容描述了配置 Db2 for z/OS 和 Analytics Accelerator for z/OS 的方法：

Db2 Analytics Accelerator for z/OS。

为 IBM Db2 for z/OS 和 IBM Db2 Analytics Accelerator 创建 ODBC 源

有关如何在 Db2 for z/OS 与 IBM Db2 Analytics Accelerator 之间启用连接的信息，请参阅下列 Web 站点：

- 对于 V4：Db2 Analytics Accelerator for z/OS 4.1.0
- 对于 V3：Db2 Analytics Accelerator for z/OS 3.1.0
- Enabling query acceleration with IBM Db2 Analytics Accelerator for ODBC and JDBC applications without modifying the applications
- SQL error from ODBC driver when running a query in Db2 Analytics Accelerator for z/OS

在 IBM SPSS Modeler 中启用 IBM Db2 for z/OS 集成

要在 SPSS Modeler 中启用 Db2 for z/OS 集成，请执行下列步骤：

1. 从 SPSS Modeler `config` 目录，打开 `odbc-db2-accelerator-names.cfg` 文件。

   如果此文件不存在，那么必须进行创建。
2. 添加所有数据源的名称和所有加速器的名称。 例如：

   ```
   dsn1, acceleratorname1
   dsn2, acceleratorname2
   ```
3. 用于仅加速器表 (AOT) 的缺省 CCSID 为 Unicode；要覆盖此设置，向加速器名称中添加编码字符串来修改条目。 例如：

   ```
   dsn1, acceleratorname1, EBCDIC
   dsn2, acceleratorname2, UNICODE
   ```
4. 保存并关闭 `odbc-db2-accelerator-names.cfg` 文件，然后打开同一目录中的 `odbc-db2-custom-properties.cfg` 文件。
5. SPSS Modeler 使用 SQL 设置 IDAA 注册。 如果需要，可以将 SQL 更改为所需值来覆盖这些条目。 例如：

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```
6. 缺省情况下， SPSS Modeler 使用 SQL 为数据库缓存创建临时表。 如果需要，可以指定预期数据库名称来覆盖此设置。 例如：

   ```
   [OSZ]
   table_create_temp_sql_acc, 'CREATE TABLE <table-name> <(table-columns)> IN DATABASE
   NAME_OF_DATABASE_FOR_AOT'
   ```
7. 缺省情况下，SPSS Modeler 认为 ODBC 源节点中编写的 SQL 查询是不可重放的，意味着在多次执行时，会认为查询返回不同的结果。 但是，在某些场景中，这可能会阻止 Modeler 为下游节点生成 SQL，并且可以通过将相关值更改为 Y 进行覆盖。 例如：

   ```
   assume_custom_sql_replayable, Y
   ```
8. 在 SPSS Modeler 主菜单中，单击**工具** > **选项** > **帮助应用程序**。
9. 单击 **IBM Db2 for z/OS** 选项卡。
10. 选中**启用 IBM Db2 for z/OS 数据挖掘集成**，然后单击**确定**。

注: 您无法在 Modeler 中同时查看 IDAA 和非 IDAA 表。

启用 SQL 生成和优化

由于使用超大型数据集的可能性，出于性能的原因，您应在 IBM SPSS Modeler 中启用 SQL 生成和优化选项。

要配置 SPSS Modeler，请完成下列步骤：

1. 从 IBM SPSS Modeler 菜单中选择**工具** > **流属性** > **选项**。
2. 在导航窗格中单击**优化**选项。
3. 确认是否已启用**生成 SQL** 选项。 要使数据库建模正常发挥作用，此设置是必需的。
4. 选择**优化 SQL 生成和优化其他执行**（非严格必需但强烈推荐使用，以使性能更优）。

在 IBM SPSS Modeler 中使用 IBM Db2 Client 配置 DSN

如果需要在 SPSS Modeler 中配置使用 Db2 Client for Db2 的数据源名称 (DSN)，请完成以下步骤：

1. 如果尚未安装，请在安装了 Modeler Server 的操作系统上安装 Db2 客户端。
2. 通过使用 **db2 catalog** 命令，对数据库编目，并将新数据源添加到 Db2 客户端中的 `db2cli.ini` 文件。 确保指向定义的数据库别名。
3. 配置数据访问权；Modeler 文档中提供了详细步骤。

   有关更多信息，请参阅《Modeler Server 管理与性能指南》(ModelerServerAdminPerformance.pdf) 中的主题**体系结构和硬件建议** > **数据访问权**。
4. 通过引用步骤 2 中定义的数据库别名在 `odbc.ini` 中创建新的 ODBC 数据源。
5. 对于 Linux 或 UNIX 用户：
   a. 确保使用了驱动程序库 `libdb2o.so`（而不是 `libdb2.so`），并确保为新数据源定义了 `'DriverUnicodeType=1'`。
   b. 在 IBM SPSS 数据访问包安装中，确保将 Db2 客户端的库路径添加到 `odbc.sh`。
   c. 确保 Modeler Server 使用具有 UTF-16 编码的 ODBC 驱动程序包装程序库（这称为 `'libspssodbc_datadirect_utf16.so'`）。
6. 确保连接到 Db2 的用户具有运行以下查询的必需权限：

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

使用 IBM Db2 for z/OS 来构建模型

每种受支持的算法均具有对应的建模节点。 您可以从节点选用板上的“数据库建模”选项卡中访问 Db2 for z/OS 建模节点。

数据注意事项

数据源中的字段可以包含各种数据类型的变量，具体取决于建模节点。 在 SPSS Modeler 中，数据类型称为测量级别。 建模节点的“字段”选项卡通过图标来指示其输入字段和目标字段所允许的测量级别类型。

目标字段。 目标字段是您尝试预测值的字段。 在可以指定目标的情况下，只能选择一个源数据字段作为目标字段。

记录标识字段。 指定用来唯一地标识各个观测值的字段。 例如，标识字段，比如客户标识。 如果源数据不包含标识字段，您可以通过“派生”节点来创建此字段，如下所示。

1. 选择源节点。
2. 在节点选用板的“字段选项”选项卡中，双击“派生”节点。
3. 在工作区上双击“派生”节点的图标可将其打开。
4. 在**派生字段**字段中，输入（例如） **标识**。
5. 在**公式**字段中，输入 `@INDEX` 并单击**确定**。
6. 将“派生”节点连接到流的其余部分。

处理空值

如果输入数据包含空值，那么使用某些 Db2 for z/OS 节点可能会导致产生错误消息或者长时间运行的流，因此我们建议移除包含空值的记录。 请使用以下方法。

1. 将“选择”节点附加到源节点。
2. 将“选择”节点的**模式**选项设置为**丢弃**。
3. 在**条件**字段中输入以下内容：

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

   确保包括每个输入字段。
4. 将“选择”节点连接到流的其余部分。

模型输出

包含 Db2 for z/OS 建模节点的流有可能每次运行都产生略微不同的结果。 这是因为数据在建模之前被读入临时表，因此节点读取源数据的顺序并不始终相同。 但是，这种影响产生的差异可以忽略不计。

一般评论

- 在 SPSS Collaboration and Deployment Services 中，不能使用包含 Db2 for z/OS 建模节点的流来创建评分配置。
- Db2 for z/OS 节点构建的模型无法进行 PMML 导出或导入。

IBM Db2 for z/OS 模型 - 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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

目标。 请选择一个字段作为预测目标。 对于广义线性模型，请另查看此屏幕中的**试验**字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Db2 for z/OS 模型 - 服务器选项

在“服务器”选项卡上，指定要在其中构建模型的 Db2 for z/OS 系统。

- **使用上游连接**。 （缺省）使用上游节点（例如“数据库源”节点）中指定的连接详细信息。 注：仅当所有上游节点都能够使用 SQL 回送功能时，此选项才有效。 在这种情况下，由于 SQL 完全实现了所有的上游节点，因此无需将数据移出数据库。
- **移动数据到连接**。 将数据移动到此处指定的数据库。 这样，即使数据位于另一个 IBM 数据库或者另一供应商的数据库中，甚至位于平面文件中，也仍然可以进行建模。 另外，如果由于某个节点未执行 SQL 回送而导致数据已被提取，那么数据将移回到此处指定的数据库中。 单击**编辑**按钮以浏览并选择连接。

注： 实际上，ODBC 数据源名称嵌入在每个 SPSS Modeler 流中。 如果在一台主机上创建的流在另一台主机上执行，那么该数据源在两台主机上的名称必须相同。 另外，也可以在各个源或建模节点的“服务器”选项卡上选择另一个数据源。

IBM Db2 for z/OS 模型 - 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。

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。

如果现有名称已被使用，请替换该名称。 如果选中此复选框，那么将覆盖所有的同名现有模型。

IBM Db2 for z/OS 模型 - K-Means

K-Means 节点实现 *k*-Means 算法，这提供了聚类分析的方法。 您可使用该节点来聚类数据集为不同的组。

此算法是基于距离的聚类算法，它依赖于距离度量（函数）以测量数据点之间的相似性。 根据使用的距离度量，将数据点分配到与之距离最近的聚类。

此算法通过执行同一基本过程的多次迭代完成操作，在该过程中，将每个训练实例分配到最接近（由应用于实例和聚类中心的指定距离函数确定）的聚类。 然后，重新计算所有聚类中心，作为分配给特定聚类实例的平均属性值向量。

IBM Db2 for z/OS 模型 - K-Means 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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Db2 for z/OS 模型 - K-Means 构建选项

通过设置构建选项，您可以根据具体用途定制模型的构建。

如果要使用缺省选项构建模型，请单击**运行**。

距离测量。 此参数定义用于测量数据点之间的距离的方法。 距离越大，表示非相似性越大。 请选择下列其中一个选项：

- **欧式**。 欧式距离测量是两个数据点之间的直线距离。
- **标准化欧式**。 标准化欧式距离测量类似于欧式距离测量，但已通过标准差的平方标准化。 与欧式距离测量不同的是，标准化欧式距离测量还具有尺度不变性。

聚类数。 此参数定义要创建的聚类数。

最大迭代次数。 此算法执行同一过程的多次迭代。 此参数定义迭代次数，在此迭代次数后模型训练停止。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注： 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

复制结果。 如果要设置随机种子以重复进行分析，请选中此复选框。 您可以指定一个整数，也可以通过单击**生成**来创建一个伪随机整数。

IBM Db2 for z/OS 模型 - 朴素贝叶斯

朴素贝叶斯是广泛用于处理分类问题的算法。 此模型将所有建议预测变量视为相互独立，因此被称为“朴素”。 朴素贝叶斯是一种可伸缩的快速算法，用于计算各个属性与目标属性的组合的条件概率。 使用训练数据确定独立的概率。 给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

IBM Db2 for z/OS 模型 - 决策树

决策树是代表分类模型的层次结构。 使用决策树模型，您可以开发分类系统，以便根据一组训练数据来预测未来观测值或者对其进行分类。 分类采用树结构形式，其中的分支表示分类中的分割点。 这些分割以递归方式将数据划分为子组，直至到达停止点为止。 停止点处的树节点称为叶片。 每片树叶分配一个标签（称为 类标签）给其子组或类成员。

IBM Db2 for z/OS 模型 - 决策树字段选项

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。 如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

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。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

目标。 请选择一个字段作为预测目标。

记录标识。 这是要用作唯一记录标识的字段。 该字段的值对于每个记录必须是唯一的（例如，客户标识号）。

实例权重。 在此处指定字段将允许您使用实例权重（每一行输入数据具有一个权重）来代替缺省的类权重（目标字段的每个类别具有一个权重）或者同时使用这两种权重。 在此处指定的字段必须是包含每行输入数据的数字权重的字段。

预测变量（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。

IBM Db2 for z/OS 模型 - 决策树构建选项

下列构建选项可用于树增长：

增长测量。 这些选项用于控制测量树增长的方式。

- **杂质测量**。 此测量用于评估分割树的最佳位置。 这是对数据子组或分段中的可变性的测量。 较小的杂质测量值指示这样一个组，该组中的大多数成员的标准或目标字段具有相似的值。

  受支持的测量为**熵**和**吉尼**。 这些测量基于分支的类别成员资格概率。
- **最大树深度**。 这是在根节点以下树可以增长到的最大层数，即，递归分割样本的次数。 此属性的缺省值为 10，您可以为此属性设置的最大值为 62。

  注: 如果模型块中的查看器显示模型的文本表示，那么最多可以显示 12 层。

分割条件。 这些选项用于控制何时停止分割树。

- **分割的最小改进**。 在树中创建新的分割之前，必须减少的最小杂质量。 树构建的目的是创建具有相似输出值的子组，以最大程度地减少每个节点中的杂质。 如果某个分支的最佳分割按小于分割标准所指定的数量来减少杂质，那么不会分割此分支。
- **分割的最小实例数**。 可以分割的最小记录数。 如果剩余的未分割记录数小于此数目，那么将不执行进一步分割。 您可以使用此字段来防止在树中创建小型子组。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。

- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

IBM Db2 for z/OS 模型 - 决策树节点 - 类权重

在这里，可以对各个类分配权重。 在缺省情况下，将向所有的类分配值 1，从而使它们具有相同的权重。 通过为不同类标签指定不同的数值权重，将引导算法相应地对特定类的训练集进行加权。

要更改权重，在**权重**列双击权重并进行所需更改。

值。 类标签集源自目标字段的可能值。

权重。 要分配给特定类的权重。 如果为某个类分配较高权重，那么模型将对此类比其他类更为敏感。

可以将类权重与实例权重配合使用。

IBM Db2 for z/OS 模型 - 决策树节点 - 树修剪

您可以使用修剪选项来指定决策树的修剪标准。 修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

修剪测量。 缺省的修剪测量为**精确度**，它确保在从树上去掉一个树叶后，模型的估算精确度仍保持在可接受的限度内。 如果您要在应用修剪时将类权重考虑在内，那么可以使用**加权精确度**选项。

用于修剪的数据。 您可以使用部分或全部训练数据来估算新数据的预期精确性。 或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据**。 此选项（缺省）使用所有训练数据来估算模型精确度。
- **使用 % 的训练数据进行修剪**。 使用此选项可以将数据分为两个集合，分别用于训练和修剪，从而使用此处指定的百分比来修剪数据。
- 如果您要指定随机种子，以确保在您每次运行流时，数据以相同方式分区，请选择**复制结果**。 您可以在**用于修剪的种子**字段中指定一个整数，或单击**生成**来创建伪随机整数。
- **使用现有表中的数据**。 指定用于估算模型精确性的独立修剪数据集的表名称。 这种做法被视为比使用训练数据更为可靠。

IBM Db2 for z/OS 模型 - 回归树

回归树是一种基于树的算法，它根据数字目标字段的值来重复分割观测值样本，以派生同一类型的子集。 与决策树一样，回归树将数据分解为子集，其中，树叶对应于足够小或足够均匀的子集。 通过选择分割来降低目标属性值的离差，以便采用树叶处的平均值来足够合理地预测它们。

IBM Db2 for z/OS 模型 - 回归树构建选项 - 树增长

可以设置用于树增长和树修剪的构建选项。

下列构建选项可用于树增长：

最大树深度。 这是在根节点以下树可以增长到的最大层数，即，递归分割样本的次数。 缺省值为 62，这是建模所允许的最大树深度。

注: 如果模型块中的查看器显示模型的文本表示，那么最多可以显示 12 层。

分割条件。 这些选项用于控制何时停止分割树。

- **分割评估度量**。 这个类评估测量用于评估分割树的最佳位置。

  注: 目前唯一可能的选项是“方差”。
- **分割的最小改进**。 在树中创建新的分割之前，必须减少的最小杂质量。 树构建的目的是创建具有相似输出值的子组，以最大程度地减少每个节点中的杂质。 如果某个分支的最佳分割按小于分割标准所指定的数量来减少杂质，那么不会分割此分支。

- **用于分割的最小实例数**。 可以分割的最小记录数。 如果剩余的未分割记录数小于此数目，那么将不执行进一步分割。 您可以使用此字段来防止在树中创建小型子组。

统计。 此参数定义将多少统计信息包括在模型中。 请选择下列其中一个选项：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注： 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

IBM Db2 for z/OS 模型 - 回归树构建选项 - 树修剪

您可以使用修剪选项来指定回归树的修剪标准。 修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

修剪度量。 修剪测量确保从树中移除树叶后，模型的估算精确度仍处于可接受的限度之内。 可以选择以下测量之一：

- **mse**。 均方误差 - （缺省）测量拟合线与数据点的接近程度。
- **r2.** R 平方 - 测量因变量的偏差比例（由回归模型解释）。
- **Pearson**。 Pearson 相关系数 - 测量正态分布的线性因变量之间的关系强度。
- **Spearman**。 Spearman 相关系数 - 检测根据 Pearson 相关性看起来较弱，但实际可能较强的非线性关系。

用于修剪的数据。 您可以使用部分或全部训练数据来估算新数据的预期精确性。 或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据**。 此选项（缺省）使用所有训练数据来估算模型精确度。
- **使用 % 的训练数据进行修剪**。 使用此选项可以将数据分为两个集合，分别用于训练和修剪，从而使用此处指定的百分比来修剪数据。

  如果您要指定随机种子，以确保在您每次运行流时，数据以相同方式分区，请选择**复制结果**。 您可以在**用于修剪的种子**字段中指定一个整数，或单击**生成**来创建伪随机整数。
- **使用现有表中的数据**。 指定用于估算模型精确性的独立修剪数据集的表名称。 这种做法被视为比使用训练数据更为可靠。

IBM Db2 for z/OS 模型 - 二阶

“二阶”节点实现了“二阶”算法，该算法提供了一种对大型数据集进行数据聚类的方法。

您可以使用此节点在考虑可用资源（例如内存和时间约束）的情况下对数据进行聚类。

“二阶”算法是一种数据库挖掘算法，此算法以如下方式对数据进行聚类：

1. 创建聚类特征 (CF) 树。 这个高度平衡的树存储聚类特征，以便执行类似输入记录成为相同树节点的组成部分的分层聚类。
2. CF 树的叶子在内存中以分层方式进行聚类，以生成最终的聚类结果。 最佳聚类数目是自动确定的。 如果您指定了最大聚类数目，那么将确定所指定限制之内的最佳聚类数目。
3. 聚类结果在第二个步骤中进行优化，在该步骤中，将对数据应用与 K-Means 算法类似的算法。

IBM Db2 for z/OS 模型 - 二阶字段选项

通过设置字段选项，您可以指定使用上游节点中定义的字段角色设置。 并且，还可以手动进行字段分配。

选择一个项目。 选择此选项表示使用上游“类型”节点中的角色设置或者上游源节点的“类型”选项卡中的角色设置。 例如，角色设置包括目标和预测变量。

使用定制字段分配。 如果您希望手动指定目标、预测变量和其他角色，请选择此选项。

字段。 使用箭头可以将此列表中的项手动分配到右侧的角色字段。 图标表示每个角色字段的有效测量级别。

记录标识。 这是要用作唯一记录标识的字段。

预测变量（输入）。 选择一个或多个字段作为预测输入。

IBM Db2 for z/OS 模型 - 二阶构建选项

通过设置构建选项，您可以根据具体用途定制模型的构建。

如果要使用缺省选项构建模型，请单击**运行**。

距离测量。 此参数定义用于测量数据点之间的距离的方法。 距离越大，表示非相似性越大。 选项为：

- **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。

聚类数。 此参数定义要创建的聚类数。 选项为：

- **自动计算聚类数**。 聚类数目自动计算。 您可以在**最大值**字段中指定最大聚类数目。
- **指定聚类数**。 指定应该创建的聚类数。

统计。 此参数定义将多少统计信息包括在模型中。 选项为：

- **全部**。 将包括所有与列相关的统计信息和所有与值相关的统计信息。

  注: 此参数将包括最大数目的统计信息，并可能因此而影响系统的性能。 如果您不想以图形格式查看模型，请指定**无**。
- **列**。 将包括与列相关的统计信息。
- **无**。 仅包括对模型进行评分所需的统计信息。

复制结果。 如果要设置随机种子以重复进行分析，请选中此复选框。 您可以指定一个整数，也可以通过单击**生成**来创建一个伪随机整数。

IBM Db2 for z/OS 模型 - 二阶块 - 模型选项卡

模型选项卡包含各种图形视图，这些视图显示聚类的汇总统计量和字段分布。 您可以从模型中导出数据，也可以将视图作为图形导出。

管理 IBM Db2 for z/OS 模型

Db2 for z/OS 模型像其他 IBM SPSS Modeler 模型一样添加到工作区和“模型”选用板中，并以几乎相同的方式使用。

要直接在 Db2 for z/OS 中对数据进行评分，请完成下列步骤：

1. 在数据所在的 Db2 for z/OS 数据库中安装 SPSS Scoring Adapter。
2. 确保流连接到数据所在的 Db2 for z/OS 数据库。

对 IBM Db2 for z/OS 模型进行评分

在工作区上使用金色模型块图标来代表模型。 模型块的主要用途是对数据进行评分以生成预测，或者允许进一步分析模型属性。 评分以一个或多个附加数据字段的形式添加，如本节的随后内容所述，通过将一个“表”节点附加到模型块并运行流的此分支，可以使这些字段可见。 某些模型块对话框（例如，决策树或回归树的模型块对话框）还包含“模型”选项卡，其中提供了模型的直观表示。

这些附加的字段由目标字段名中添加的前缀 `$<id>-` 加以区分，其中 `<id>` 取决于模型，用于标识所添加的信息类型。 在每个模型块的主题中描述了不同的标识。

要查看评分，按以下步骤操作：

1. 将表节点附加到模型块。
2. 打开表节点。

3. 单击**运行**。
4. 滚动到表输出窗口的右侧，以查看附加字段及其评分。

注: 评分过程不是在加速器中运行，而是在 Db2 中运行，因此要求用于评分的输入表必须物理上位于 Db2 中。 因此，作为评分输入，只能使用基于 Db2 的表或加速表。 如果流使用仅加速器表，将发生以下错误：`"THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR."`

IBM Db2 for z/OS 决策树模型块

决策树模型块显示建模操作的输出，还允许您设置一些选项来为模型评分。

运行包含决策树模型块的流时，该节点将添加两个新字段，这两个新字段的名称从目标派生。

表 *26:* 决策树的模型评分字段

新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。
$IP-*target_name*	预测结果的置信度值 (0.0 - 1.0)。

注: 由于 Db2 for z/OS 的限制，可能会截断列名。

IBM Db2 for z/OS 决策树块 - 模型选项卡

模型选项卡以图形格式显示决策树模型的预测变量重要性。 条形的长度表示预测变量的重要性。

IBM Db2 for z/OS 决策树块 - 查看器选项卡

查看器选项卡以 SPSS Modeler 显示其决策树模型的方式显示树模型的树表示。

IBM Db2 for z/OS 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。

运行包含 K-Means 模型块的流时，该节点将添加两个新字段，这两个字段包含聚类成员信息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`。

注: 由于 Db2 for z/OS 的限制，可能会截断列名。

IBM Db2 for z/OS K-Means 块 - 模型选项卡

模型选项卡包含各种图形视图，这些视图显示聚类的汇总统计量和字段分布。 您可以从模型中导出数据，也可以将视图作为图形导出。

IBM Db2 for z/OS 朴素贝叶斯模型块

运行包含朴素贝叶斯模型块的流时，该节点将添加两个新字段，这两个新字段的名称从目标名称派生。

表 *27:* 朴素贝叶斯的模型评分字段

新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。
$IP-*target_name*	预测结果的置信度值 (0.0 - 1.0)。

注: 由于 Db2 for z/OS 的限制，可能会截断列名。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

IBM Db2 for z/OS 回归树模型块

运行包含回归树模型块的流时，该节点将添加两个新字段，这两个新字段的名称从目标名称派生。

表 *28:* 回归树的模型评分字段

新增字段的名称	**含义**
$I-*target_name*	当前记录的预测值。
$IS-*target_name*	所预测的值的估算标准差。

注: 由于 Db2 for z/OS 的限制，可能会截断列名。

您可以将一个“表”节点附加到模型块并运行这个“表”节点，以便查看这些附加字段。

IBM Db2 for z/OS 回归树块 - 模型选项卡

模型选项卡以图形格式显示回归树模型的预测变量重要性。 条形的长度表示预测变量的重要性。

IBM Db2 for z/OS 回归树块 - 查看器选项卡

查看器选项卡以 SPSS Modeler 显示其回归树模型的方式显示树模型的树表示。

IBM Db2 for z/OS 二阶模型块

运行包含二阶模型块的流时，该节点将添加两个新字段，这两个字段包含聚类成员信息以及与该记录所分配到的聚类中心的距离。 新字段名称派生自模型名称，前缀为 $TS-（表示聚类成员资格）和 $TSD-（表示与聚类中心的距离）。 例如，如果模型名称为 MDL，那么新字段的名称将是 $TS-MDL 和 $TSD-MDL。

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。

索引

Special Characters

A

D

E

I

K