IBM SPSS Data Preparation 23

注释

使用本信息及其支持的产品之前，请阅读 第 27 页的『声明』 中的信息。

产品信息

此版本适用于 IBM® SPSS® Statistics V23R0M0 及所有后续发行版和修改版，除非在新版本中另有说明。

目录

第 1 章 数据准备简介

随着计算系统能力的提高，对信息的需要成比例增长，导致收集的数据中出现更多的个案、更多的变量以及更多的数据输入错误。这些错误会损害作为数据仓储最终目标的预测模型的预测，因此您需要使数据保持“干净”。不过，数据仓储中的数据量的增长已经大大超出了手动验证个案的能力，而这对于实现自动化的数据验证过程来说十分关键。

“数据准备”附加模块允许您标识活动数据集中的异常个案和无效个案、变量和数据值，并准备建模数据。

″数据准备″过程的用法

“数据准备”过程的用法取决于您的特定需要。加载数据后，典型的过程是:

- **元数据准备。** 复查数据文件中的变量并确定其有效值、标签和测量级别。标识不太可能但经常存在编码错误的变量值的组合。根据这些信息定义验证规则。这是一项极为耗时的任务，不过，如果您需要定期验证具有类似属性的数据文件，则完成这项任务是十分值得的。
- **数据验证。** 运行基本检查并针对定义的验证规则进行检查，标识无效个案、变量和数据值。找到无效数据时，调查并更正原因。这可能需要另一个通过元数据准备的步骤。
- **模型准备。** 使用自动数据准备获得将改进模型构建的原始字段的转换。标识可能导致许多预测模型出现问题的潜在统计离群值。有些离群值是尚未标识的无效变量值导致的结果。这可能需要另一个通过元数据准备的步骤。

数据文件变成“干净”的之后，就可以从其他附加模块构建模型了。

第 2 章 验证规则

规则用于确定个案是否有效。有两种类型的验证规则:

- **单变量规则。** 单变量规则包含一组应用于单个变量的固定检查，例如范围外值的检查。对于单变量规则，有效值可以表示为一个值范围，也可以表示为一个可接受值列表。
- **交叉变量规则。** 交叉变量规则是可应用于单一变量或变量组合的用户定义规则。交叉变量规则由标记无效值的逻辑表达式定义。

验证规则保存在数据文件的数据字典中。这样指定一次规则后就可以重用规则。

加载预定义的验证规则

通过从安装中所包含的外部数据文件加载预定义规则可以快速获取一组可供使用的验证规则。

加载预定义的确认规则

1. 从菜单中选择:

   数据 > 验证 > 装入预定义规则...

或者，您也可以使用复制数据属性向导从任何数据文件加载规则。

定义验证规则

“定义验证规则”对话框允许您创建和查看单变量和交叉变量验证规则。

创建和查看验证规则

1. 从菜单中选择:

   数据 > 验证 > 定义规则...

   该对话框中包含从数据字典读取的单变量和交叉变量验证规则。如果不存在任何规则，则会自动创建一个新的占位符规则，您可以对其进行修改以满足您的要求。
2. 在“单变量规则”和“交叉变量规则”选项卡上选择各个规则可查看和修改其属性。

定义单变量规则

“单变量规则”选项卡允许您创建、查看和修改单变量验证规则。

规则。 该列表按名称和规则适用的变量类型显示单变量验证规则。该对话框打开时，它显示在数据字典中定义的规则，或者，如果当前未定义任何规则，则显示名为“Single-Variable Rule 1”的占位符规则。下列按钮将显示在“规则”列表下方:

- **新建。** 在“规则”列表底部添加一个新的条目。该规则会被选中，并分配名称“SingleVarRule *n*”，其中 *n* 是一个整数，这使得新规则的名称在单变量和交叉变量规则中是唯一的。
- **复制。** 在“规则”列表底部添加一个所选规则的副本。规则的名称会进行调整，使其在单变量和交叉变量规则中是唯一的。例如，如果复制“SingleVarRule 1”，则第一个复制规则的名称将是“Copy of SingleVarRule 1”，第二个将是“Copy (2) of SingleVarRule 1”，依此类推。

- **删除。** 删除所选规则。

规则定义。 通过这些控件，可以查看和设置所选规则的属性。

- **名称。** 规则的名称在单变量和交叉变量规则中必须是唯一的。
- **类型。** 这是规则适用的变量类型。请从**数值**、**字符串**和**日期**中进行选择。
- **格式。** 这允许您为可应用于日期变量的规则选择日期格式。
- **有效值。** 您可以以值范围或值列表的形式指定有效值。

  范围定义

  范围定义控件允许您指定有效范围。该范围以外的值会被标记为无效。

  要指定范围，请输入最小值和/或最大值。复选框控件允许您标记范围内的未标注值和非整数值。

  列表定义

  列表定义控件允许您定义有效值的列表。未包含在列表中的值会被标记为无效。

  在网格中输入列表值。该复选框确定针对可接受值列表检查字符串数据值时是否区分大小写。
- **允许使用用户缺失值。** 控制是否将用户缺失值标记为无效。
- **允许使用系统缺失值。** 控制是否将系统缺失值标记为无效。这不适用于字符串规则类型。
- **允许使用空值。** 控制是否将空白（也就是完全为空）字符串值标记为无效。这不适用于非字符串规则类型。

定义交叉变量规则

“交叉变量规则”选项卡允许您创建、查看和修改交叉变量验证规则。

规则。该列表按名称显示交叉变量验证规则。该对话框打开时，它显示名为“CrossVarRule 1”的占位符规则。下列按钮将显示在“规则”列表下方:

- **新建。**在“规则”列表底部添加一个新的条目。该规则会被选中，并分配名称“CrossVarRule *n*”，其中 *n* 是一个整数，这使得新规则的名称在单变量和交叉变量规则中是唯一的。
- **复制。**在“规则”列表底部添加一个所选规则的副本。规则的名称会进行调整，使其在单变量和交叉变量规则中是唯一的。例如，如果复制“CrossVarRule 1”，则第一个复制规则的名称将是“Copy of CrossVarRule 1”，第二个将是“Copy (2) of CrossVarRule 1”，依此类推。
- **删除。**删除所选规则。

规则定义。通过这些控件，可以查看和设置所选规则的属性。

- **名称。**规则的名称在单变量和交叉变量规则中必须是唯一的。
- **逻辑表达式。**这实际上就是规则定义。您应该编写表达式以使无效个案的计算结果为 1。

构建表达式

1. 要构建一个表达式，可以将成分粘贴到“表达式”字段中或是在“表达式”字段中直接输入。

- 通过从“函数组”列表中选择组，然后双击“函数和特殊变量”列表中的函数或变量（或者选择函数或变量，然后单击**插入**），可以粘贴函数或常用的系统变量。对问号指示的所有参数输入值（适用于函数）。标记为**全部**的函数组提供所有可用函数和系统变量的列表。对话框的保留区域中显示对当前所选函数或变量的简要描述。
- 字符串常数必须包含在引号或撇号中。
- 如果值包含小数，则必须使用句号 (.) 作为小数指示符。

第 3 章 验证数据

“验证数据”对话框允许您标识活动数据集中可疑的和无效的个案、变量和数据值。

示例。数据分析人员每个月必须向客户提供客户满意度报告。她每个月接收到的数据需要进行质量检查，看是否存在不完整的客户标识、超出范围的变量值以及经常错误输入的变量值组合。“验证数据”对话框允许分析人员指定唯一标识客户的变量，为有效变量范围定义单变量规则，并定义交叉变量规则以找出不可能的组合。该过程返回问题个案和变量的报告。此外，每个月的这些数据都具有相同的数据元素，因此分析人员可以将规则应用于下个月的新数据文件。

统计信息。该过程生成多项检查失败的变量、个案和数据值的列表，违反单变量和交叉变量规则的次数计数，以及分析变量的简单描述摘要。

权重。该过程忽略权重变量规范，而是像对待任何其他分析变量一样对待权重变量。

验证数据

1. 从菜单中选择:

   数据 > 验证 > 验证数据...
2. 选择一个或多个分析变量，以便由基本变量检查或单变量验证规则进行验证。

   或者，您可以:
3. 单击**交叉变量规则**选项卡并应用一个或多个交叉变量规则。

根据需要，您可以:

- 选择一个或多个个案标识变量以便检查重复的或不完整的 ID。个案标识变量还可用于标记个案输出。如果指定了两个或更多个案标识变量，则可将其值的组合视为个案标识。

测量级别未知的字段

当数据集中的一个或多个变量（字段）的测量级别未知时，将显示测量级别警告。由于测量级别会影响该过程的计算结果，因此所有变量必须都定义有测量级别。

扫描数据。 读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。

手动分配。 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

验证数据: 基本检查

“基本检查”选项卡允许您为分析变量、个案标识和全部个案选择基本检查。

分析变量。如果在“变量”选项卡上选择了任何分析变量，则可选择以下任意有效性检查。复选框允许您打开或关闭检查。

- **缺失值的最大百分比。**报告缺失值百分比大于指定值的分析变量。指定的值必须是一个小于等于 100 的正数。

- **单个类别中个案所占的最大百分比。**如果任何分析变量是分类变量，则此选项报告表示单个非缺失类别的个案的百分比大于指定值的分类分析变量。指定的值必须是一个小于等于 100 的正数。百分比基于具有非缺失变量值的个案。
- **计数为 1 的类别的最大百分比。**如果任何分析变量是分类变量，则此选项报告仅包含一个个案的变量类别的百分比大于指定值的分类分析变量。指定的值必须是一个小于等于 100 的正数。
- **最小变异系数。**如果任何分析变量是刻度变量，则此选项报告变异系数的绝对值小于指定值的刻度分析变量。此选项仅适用于平均值非零的变量。指定的值必须是一个非负数。指定 0 会关闭变异系数检查。
- **最小标准差。**如果任何分析变量是刻度变量，则此选项报告标准差小于指定值的刻度分析变量。指定的值必须是一个非负数。指定 0 会关闭标准差检查。

个案标识。如果在“变量”选项卡上选择了任何个案标识变量，则可选择以下任意有效性检查。

- **标记不完整的标识。**此选项报告具有不完整个案标识的个案。对于特定个案，如果任何标识变量的值为空或者缺失，则该标识被视为不完整。
- **标记重复的标识。**此选项报告具有重复个案标识的个案。不完整的标识会从可能重复的组中排除。

标记空个案。此选项报告所有变量均为空或空白的个案。为了标识空个案，您可以选择使用文件中的所有变量（不包括任何标识变量）或者仅使用在“变量”选项卡上定义的分析变量。

验证数据: 单变量规则

“单变量规则”选项卡显示可用的单变量验证规则，并允许您应用这些规则分析变量。要定义其他单变量规则，请单击**定义规则**。请参阅主题第 3 页的『定义单变量规则』以获取更多详细信息。

分析变量。该列表显示分析变量，汇总其分布，并显示应用于每个变量的规则的数量。注意，用户缺失值和系统缺失值不包含在摘要中。“显示”下拉列表控制显示哪些变量；您可以从**所有变量**、**数值变量**、**字符串变量**和**日期变量**中选择。

规则。要对分析变量应用规则，请选择一个或多个变量，然后在“规则”列表中选中要应用的所有规则。“规则”列表仅显示适用于所选分析变量的规则。例如，如果选择了数值分析变量，则仅显示数值规则；如果选择了字符串变量，则仅显示字符串规则。如果未选择任何分析变量，或者选择的分析变量具有混合数据类型，则不显示任何规则。

变量分布。“分析变量”列表中显示的分布摘要可基于所有个案或基于前 *n* 个个案的扫描，这在“个案数”文本框中指定。单击**重新扫描**可更新分布摘要。

验证数据: 交叉变量规则

“交叉变量规则”选项卡显示可用的交叉变量规则，并允许将其应用于您的数据。要定义其他交叉变量规则，请单击**定义规则**。请参阅主题第 4 页的『定义交叉变量规则』以获取更多详细信息。

验证数据: 输出

个案情况报告。如果您应用了任何单变量或交叉变量验证规则，则可请求列出每个个案的确认违反规则的报告。

- **最小违规数。**此选项指定要包括在报告中的个案的最小违规数。指定一个正整数。
- **个案的最大数量。**此选项指定个案报告中包含的个案的最大数量。请指定一个小于等于 1000 的正整数。

单变量验证规则。如果已经应用了任何单变量验证规则，则可选择如何显示结果或者是否显示结果。

- **依据分析变量汇总违规数。**对于每个分析变量，此选项均显示违反的所有单变量验证规则以及违反每个规则的值的数量。它还报告每个变量违反单变量规则的总次数。
- **依据规则汇总违规数。**对于每个单变量验证规则，此选项均报告违反了该规则的变量以及每个变量的无效值的数量。它还报告全部变量违反每个规则的值的总数。

显示分析变量的描述统计。 此选项允许您请求分析变量的描述统计。会为每个分类变量生成一个频率表。为刻度变量生成包括平均值、标准差、最小值和最大值的汇总统计表。

将具有验证违规数的个案移到活动数据集的顶部。 此选项将违反了单变量规则或交叉变量规则的个案移动到活动数据集的顶部以便于查阅。

验证数据: 保存

“保存”选项卡允许将记录违规的变量保存到活动数据集。

摘要变量。这些是可以保存的单个变量。选中一个框可保存该变量。为这些变量提供了缺省名称；您可以进行编辑。

- **空个案指示符。**空个案会分配值 1。所有其他个案都具有代码 0。变量的值反映在“基本检查”选项卡上指定的范围。
- **双标识组。**具有相同个案标识的个案（具有不完整标识的个案除外）会分配有相同的组号。具有唯一标识或不完整标识的个案都具有代码 0。
- **标识指示符不完整。**具有空的或不完整的个案标识的个案将分配值 1。所有其他个案的代码都为 0。
- **验证规则违反（总数）。**这是按个案计数的违反单变量和交叉变量验证规则的总数。

替换现有的摘要变量。保存到数据文件的变量必须具有唯一的名称，否则就会替换具有相同名称的变量。

保存指示符变量。此选项允许保存确认违反规则的完整记录。每个变量都对应着验证规则的一次应用，如果个案违反了该规则，则值为 1，如果未违反，则值为 0。

第 4 章 自动数据准备

准备分析数据是任何项目中最重要的步骤之一，而从一般来说也是最耗时的步骤之一。“自动数据准备 (ADP)”为您处理任务，分析您的数据并识别修正，过滤出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能过滤技术改进性能。您可以通过完全**自动**方式使用算法，这种方式可以允许选择并应用修正；或者也可以通过**交互式**方式使用算法，这种方式可以在做出更改前对其进行预览，并按照需要进行接受或拒绝。

通过使用 ADP，您可以快速、轻松地准备数据以供建模，无需具备相关统计概念的预备知识。您可以更快速地构建模型并进行评分。此外，使用 ADP 还能提高自动化建模过程。

注: 当 ADP 准备字段进行分析时，它将创建包含调整或转换的新字段，而不是替换旧字段的现有值和属性。旧字段不用于进一步分析，其角色被设置为“无”。 同时请注意，任何用户缺失值信息都不会转移到这些新建的字段，而新字段中的任何缺失值将成为系统缺失值。

示例。在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来标记具有潜在欺骗性的可疑理赔。构建模型前，他们将使用自动数据准备来准备数据进行建模。由于他们希望能够在应用转换前查看建议的转换，他们将在交互模式下使用自动数据准备。

某汽车集团希望跟踪各类私人汽车的销售情况。为了能够标识表现良好和表现不好的型号，他们希望建立汽车销售和汽车特性之间的关系。他们将使用自动数据准备来准备数据进行分析，同时使用准备“之前”和“之后”的数据构建模型以查看结果的差别。

您的目标是什么? 自动数据准备可以推荐能够加快其他算法的建模速度、并增强这些模型的预测能力的数据准备步骤。可包括转换、构建和选择功能。也可对目标进行转换。您可以指定数据准备过程应遵循的建模优先级次序。

- **均衡速度和精确度。**该选项可以准备数据，以使建模算法处理数据的速度和预测的精确度具有同等优先级。
- **优化速度。**该选项可以准备数据，以使建模算法处理数据的速度具有较高优先级。如果您处理非常大的数据集，或要求快速得到结果时，则选择此选项。
- **优化精确度。**该选项可以准备数据，以使建模算法生成的预测结果的精确度具有较高优先级。
- **自定义分析。**如果您希望手动修改“设置”选项卡上的算法，请选择此选项。注意，如果您随后在“设置”选项卡上更改了与其他目标之一不一致的选项，则会自动选择该设置。

获得自动数据准备

从菜单中选择:

1. 从菜单中选择:

   转换 > 准备数据以供建模 > 自动...

2. 单击**运行**。

根据需要，您可以:

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

获得交互式数据准备

1. 从菜单中选择:

   转换 > 准备数据以供建模 > 交互式...

2. 在对话框顶部工具栏中单击**分析**。
3. 单击“分析”选项卡，并复核建议的数据准备步骤。
4. 如果满意，单击**运行**。否则，单击**清除分析**，更改所需设置，并单击**分析**。

根据需要，您可以:

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。
- 单击**保存 XML**，将建议的数据准备步骤保存到 XML 文件。

字段选项卡

“字段”选项卡指定应准备哪些字段以供进一步分析。

使用预定义角色。 此选项使用现有的字段信息。如果存在具有“目标”角色的单个字段，它将用作目标，否则将不存在目标。所有具有预定义角色“输入”的字段将用作输入。需要至少一个输入字段。

使用自定义字段分配。 当您通过将字段从其缺省列表中移走来覆盖字段角色时，对话框会自动切换到该选项。当进行自定义字段分配时，请指定以下字段:

- **目标（可选）。** 如果计划构建需要目标的模型，请选择目标字段。这类似于将字段角色设为“目标”。
- **输入字段。** 选择一个或多个输入字段。这类似于将字段角色设为“输入”。

设置选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以微调算法如何处理数据。如果您对与其他目标不一致的缺省设置进行了更改，则“目标”选项卡会自动更新为选择**自定义分析**选项。

准备日期和时间

许多建模算法无法直接处理日期和时间细节。这些设置允许您从现有数据中的日期和时间派生新的持续时间数据，以用作模型输入。该字段包含必须采用日期或时间存储类型预定义的日期和时间。不建议在自动数据准备后将原始日期和时间字段用作模型输入。

准备日期和时间以供建模。取消选择该选项将在保持选择的同时禁用所有其他“准备日期&时间”控件。

计算到参考日期为止已过去的时间。这将为包含日期的每个变量生成自参考日期后的年/月/日数。

- **参考日期。**指定以该日期为参考，根据输入数据中的日期信息计算持续时间的日期。如果选择**当前日期**，则 ADP 执行时始终使用当前系统日期。要使用特定日期，选择**固定日期**，并输入所需日期。
- **持续日期的单位。**指定 ADP 是自动确定持续日期的单位，还是从**固定单位**（年、月或日）中选择。

计算到参考时间为止已过去的时间。这将为包含时间的每个变量生成自参考日期后的小时/分钟/秒数。

- **参考时间。**指定以该时间为参考，根据输入数据中的日期信息计算持续的时间。如果选择**当前时间**，则 ADP 执行时始终使用当前系统时间。要使用特定时间，选择**固定时间**，并输入所需具体时间。
- **持续时间的单位。**指定 ADP 是自动确定持续时间的单位，还是从**固定单位**（小时、分或秒）中选择。

提取循环时间元素。使用这些设置将单个日期或时间字段拆分成一个或多个字段。例如，如果您选择了全部三个日期复选框，则输入日期字段 ″1954-05-23″ 会被拆分成三个字段：1954、5 和 23，分别使用在**字段名称**面板中定义的后缀，原始日期字段则被忽略。

- **从日期提取。**对于任何日期输入，请指定是否要提取年、月、日或任意组合。
- **从时间提取。**对于任何时间输入，请指定是否要如果要提取小时、分、秒或任意组合。

排除字段

质量较差的数据会影响到预测的准确性，因此需要为输入特征指定可接受的质量级别。所有为常量或缺失值达 100% 的字段自动被排除。

排除低质量的输入字段。取消选择该选项将在保持选择的同时禁用所有其他“排除字段”控件。

排除缺失值过多的字段。删除缺失值超过指定百分比的字段，而不会用于进一步分析。指定大于或等于 0 的值等同于取消选择该选项，同时指定小于或等于 100 的值将自动排除具有所有缺失值的字段。缺省值是 50。

排除唯一类别过多的名义字段。删除类别超过个数的字段，而不会用于进一步分析。指定一个正整数。缺省值为 100。这对于自动从建模中删除包含记录特有信息（如 ID、地址或名称）的字段非常有用。

排除单个类别中值过多的分类字段。删除在单个类别中包含超过指定百分比的记录的有序和名义字段，而不会用于进一步分析。指定大于或等于 0 的值等同于取消选择该选项，同时指定小于或等于 100 的值将自动排除常数字段。缺省值为 95。

调整测量

调整测量级别。 取消选择该选项将在保持选择的同时禁用所有其他“调整测量”控件。

测量级别。 指定是否将“值太少”的连续字段的测量级别调整为有序，并将“值太多”的有序字段的测量级别调整为连续。

- **有序字段值的最大数量。** 具有超过指定类别数目的有序字段将被重新强制转换为连续字段。指定一个正整数。缺省值为 10。该值必须大于或等于连续字段值的最小数目。
- **连续字段值的最小数量。** 具有少于指定唯一值数目的连续字段将被重新强制转换为有序字段。指定一个正整数。缺省值为 5。该值必须小于或等于有序字段值的最大数量。

提高数据质量

准备要提高数据质量的字段。 取消选择该选项将在保持选择的同时禁用所有其他“提高数据质量”控件。

处理离群值。 指定是否为输入和目标替换离群值；如果是，则指定离群值分界值标准（采用标准差测量）和离群值替换方法。可以通过修整（设置为分界值）或将其设置为缺失值来替换离群值。在任何离群值被设置为缺失值后，将按照下面所选的缺失值处理设置进行处理。

替换缺失值。 指定是否替换连续、名义或有序字段的缺失值。

重新排序名义字段。 选中此选项，以按从小（最少出现）到大（最常出现）的类别顺序重新编码名义（集合）字段值。新字段值从 0 开始作为最少出现的类别。注意，如果原始字段为字符串，新字段将为数值。例如，如果名义字段的数据值为 ″A″、″A″、″A″、″B″、″C″、″C″，那么自动数据准备将把 ″B″ 重新编码为 0、将 ″C″ 编码为 1，同时将 ″A″ 编码为 2。

重新调整字段

重新调整字段。 取消选择该选项将在保持选择的同时禁用所有其他“重新调整字段”控件。

分析权重。 此变量包含分析（回归或抽样）权重。分析权重将作为对目标字段各个水平上方差的差异的一种考量。选择一个连续字段。

连续输入字段。 这将使用 **Z 得分转换**或**最小/最大转换**来标准化连续输入字段。当您在“选择和构建”设置中选择了**执行特征构建**时，重新调整输入特别有用。

- **Z 得分转换。** 以观察到的平均值和标准差作为总体参数估计，将字段标准化，然后将 *z* 得分映射到具有指定**最终平均值**和**最终标准差**的正态分布的对应值。为**最终平均值**指定一个数字并为**最终标准差**指定一个正数。缺省值为 0 和 1，分别对应于标准化重新调整。
- **最小/最大转换。** 以观察到的最小值和最大值作为总体参数估计，将字段映射到具有指定**最小值**和**最大值**的均匀分布的对应值。在指定数字值时，确保**最大值**大于**最小值**。

连续目标。 这将使用 Box-Cox 转换来转换连续目标，其结果字段为近似正态分布，且具有指定的**最终平均值**和**最终标准差**。为**最终平均值**指定一个数字并为**最终标准差**指定一个正数。缺省值分别为 0 和 1。

注: 如果目标已被 ADP 转换，则使用转换后目标构建的后续模型将针对转换后的单位评分。要解释和使用结果，您必须将预测值转换回原始刻度。请参阅 主题以获取更多信息。请参阅 第 19 页的『逆转换得分』主题以获取更多信息。

转换字段

为提高数据预测能力，您可以转换输入字段。

转换建模字段。取消选择该选项将在保持选择的同时禁用所有其他“转换字段”控件。

分类输入字段 以下选项可用:

- **合并松散类别以最大化与目标的关联。**选中此选项，可以减少与目标关联的需处理的字段数目，得到更简约的模型。通过输入与目标间的关系可以确定类似的类别。无显著差异（即 *p* 值大于指定值）的类别则被合并。指定一个大于 0 且小于或等于 1 的值。如果将所有类别合并为单个类别，则会从进一步分析中排除字段的原始和派生版本，因为它们没有值作为预测变量。
- **没有目标时，根据以下计数合并松散类别。**如果数据集没有目标，您可以选择合并有序和名义字段的松散类别。等频法用于合并具有低于指定的总记录数最小百分比的类别。指定一个大于等于 0 且小于等于 100 的值。缺省值为 10。当不存在具有低于指定最小个案百分比的类别，或只剩下两个类别时，合并停止。

连续输入字段。如果数据集包含类别目标，则可以采用强关联对连续输入分级，以改进处理性能。分箱是根据“齐次子集”的属性来创建，后者通过 Scheffe 方法进行确定，并使用指定的 *p* 值作为确定齐次子集的临界值 alpha。指定一个大于 0 且小于或等于 1 的值。缺省值为 0.05。如果特定字段的分箱化结果为单个分箱，则会排除字段的原始和分级版本，因为它们没有值作为预测变量。

注: ADP 中的分箱化与最佳分箱化不同。最佳分箱化使用熵信息将连续字段转换为分类字段。这需要在内存中对全部数据进行排序和存储。ADP 使用齐次子集来分箱化连续字段，这意味着 ADP 分箱化不需要在内存中对全部数据进行排序和存储。通过使用齐次子集方法分箱化连续字段，分箱化后的类别数总是小于或等于目标中的类别数。

选择和构建

为提高数据预测能力，您可以根据现有字段构建新的字段。

执行特征选择。 如果某个连续输入与目标关联的 *p* 值大于指定的 *p* 值，则从分析中删除此连续输入。

执行特征构建。 选择该选项从若干现有特征组合派生出新特征。旧特征将不用于进一步分析。该选项仅适用于目标为连续或不存在目标的连续输入特征。

字段名称

为方便识别新的和转换后的特征，ADP 可以创建并应用基本新名称、前缀或后缀。您可以更改这些名称，以使其与您的要求和数据更相关。

转换字段和构建字段。指定要应用到转换目标和输入字段的名称扩展。

此外，还需要指定要应用到通过“选择和构建”设置所构建的任何特征的前缀名称。新名称将通过为此前缀根名称添加数字后缀生成。数字格式取决于生成的特征数目，例如:

- 第 1-9 个构建的特征将命名为: feature1 到 feature9
- 第 10-99 个构建的特征将命名为: feature10 到 feature99
- 第 100-999 个构建的特征将命名为: feature001 到 feature999，依次类推。

这可以确保不论有多少个特征，都将按有意义的顺序排列。

从日期和时间计算得出的持续时间。指定要应用到从日期和时间计算的持续时间的名称扩展。

从日期和时间提取的循环元素。指定要应用到从日期和时间提取的循环元素的名称扩展。

应用和保存转换

根据您在使用交互式还是自动数据准备对话框，应用和保存转换的设置也略有差异。

交互式数据准备应用转换设置

已转换数据。这些设置指定已转换数据的保存位置。

- **将新字段添加到活动数据集。**自动数据准备所创建的任何字段都将作为新字段添加到活动数据集。**更新已分析字段的角色**可将由自动数据准备从进一步分析中排除的任何字段的角色设置为“无”。
- **新建包含已转换数据的数据集或文件。** 自动数据准备所建议的字段都将添加到新数据集或文件中。**包括未分析字段**会将在“字段”选项卡上未指定的原始数据集字段添加到新数据集。这对于将包含未在建模中使用的信息（如 ID、地址或名称）的字段转移到新数据集中非常有用。

自动数据准备应用和保存设置

“已转换数据”组与“交互式数据准备”中的相同。在自动数据准备中，有下列其他选项可用:

应用转换。在“自动数据准备”对话框中，取消选择本选项将在保持选择的同时禁用所有其他“应用和保存”控件。

将转换另存为语法。这可将建议的转换作为命令语法保存到外部文件。“交互式数据准备”对话框则不包含此控件，因为它可在您单击**粘贴**时将转换作为命令语法粘贴到语法窗口。

将转换另存为 XML。这可将建议的转换作为 XML 保存到外部文件，后者可通过 TMS MERGE 与模型 PMML 合并，或通过 TMS IMPORT 应用到其他数据集。“交互式数据准备”对话框则不包含此控件，因为它可在您单击对话框顶部工具栏中的**保存 XML** 时将转换另存为 XML。

分析选项卡

注: 在“交互式数据准备”对话框中使用“分析”选项卡，您可以审核建议的转换。“自动数据准备”对话框不包含此步骤。

1. 在完成对 ADP 的设置（包括对“目标”、“字段”和“设置”选项卡所作的任何更改）后，单击**分析数据**。算法将设置应用到数据输入，并在“分析”选项卡上显示结果。

“分析”选项卡包含表格和图形输出，其中显示数据处理概要，并显示有关如何修改或改进数据以提高得分的建议。您可以审核这些建议，并加以接受或拒绝。

“分析”选项卡包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。有三个主视图:

- 字段处理概要（缺省视图）。请参阅主题『字段处理概要』以获取更多详细信息。
- 字段。请参阅主题第 15 页的『字段』以获取更多详细信息。
- 操作摘要。请参阅主题第 16 页的『操作摘要』以获取更多详细信息。

有四个链接/辅助视图:

- 预测能力（缺省视图）。请参阅主题第 16 页的『预测能力』以获取更多详细信息。
- 字段表。请参阅主题第 16 页的『字段表』以获取更多详细信息。
- 字段详细信息。请参阅主题第 16 页的『字段详细信息』以获取更多详细信息。
- 操作详细信息。请参阅主题第 17 页的『操作详细信息』以获取更多详细信息。

视图间链接

在主视图内，表格中的下划线文本控制链接视图中的显示。单击文本将显示有关特定字段、字段集合或处理步骤的详细信息。您最近一次选择的链接显示为深色，这可帮助您识别两个视图面板内容间的联系。

重置视图

要重新显示原始分析建议，并放弃对分析视图的任何更改，请单击主视图面板底部的**重置**。

字段处理概要

“字段处理摘要”表格提供了有关字段处理的预计总体影响的快照，包括对特征状态的更改和构建的特征数目。

请注意，这里不会实际构建模型，因此并不存在总体预测能力在数据准备前后的变化测量或图表，您只能显示单个建议预测变量的预测能力图表。

该表格显示以下信息:

- 目标字段数。
- 原始 (输入) 预测变量数。
- 在分析和建模中建议使用的预测变量数。其中包括建议的字段总数、建议的原始和未转换的字段数、建议的转换字段数（排除任何字段的中间版本、从日期/时间预测变量派生的字段以及构建的预测变量）、从日期/时间字段派生的建议字段数，以及建议的构建预测变量数。
- 不建议以任何形式（原始、派生字段和构建预测变量的输入）使用的输入预测变量数。

如果任何**字段**信息带有下划线，单击可在链接视图中显示更多信息。在“字段表链接视图”中显示**目标**、**输入特征**和**未使用输入特征**的详细信息。请参阅主题第 16 页的『字段表』以获取更多信息。**在“预测能力”链接视图中显示建议在分析中使用的特征**。请参阅主题第 16 页的『预测能力』以获取更多详细信息。

字段

“字段”主视图显示处理过的字段，以及 ADP 是否建议在下游模型中使用它们。您可以覆盖任何字段建议。例如，排除构建的特征或包含 ADP 建议排除的特征。如果字段已转换，您可以决定是接受建议转换，还是使用原始版本。

“字段”视图由两个表格组成，分别显示目标和处理或创建的预测变量。

目标表

仅当数据中定义有目标时，才会显示**目标**表。

该表包含两列:

- **名称。**此为目标字段的名称或标签。不论字段是否已转换，始终使用原始名称。
- **测量级别。**此列显示代表测量级别的图标。将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。

  如果目标已转换，则**测量级别**列将反映最终转换版本。*注*: 您不能关闭目标转换。

预测变量表格

预测变量表格总是显示。表格的每一行代表一个字段。缺省情况下，按预测能力的降序来排列行。

对于普通特征，原始名称始终用作行名称。表格中以单独行显示日期/时间字段的原始和派生版本，此外，还包括构建的预测变量。

注意，在表格中显示的字段转换后版本始终代表最终版本。

缺省情况下，在预测变量表中只显示建议的字段。要显示其余字段，选中表格上方的**在表中包括非推荐字段**复选框，这些字段随即显示在表格底部。

该表包含以下列:

- **使用的版本。**此列显示一个下拉列表，以控制字段是否将在下游使用，以及是否使用建议的转换。缺省情况下，下拉列表将反映建议。

  对于已转换的普通预测变量，下拉列表有三个选项: **已转换**、**原始**和**不使用**。

  对于未转换的普通预测变量，下拉列表的选项为: **原始**和**不使用**。

  F对于派生的日期/时间字段和构建的预测变量，选项为: **已转换**和**不使用**。

  对于原始日期字段，下拉列表被禁用，并设置为**不使用**。

  注: 对于同时具有原始和已转换版本的预测变量，如果切换**原始**和**已转换**版本，则会自动更新这些特征的**测量级别**和**预测能力**设置。
- **名称。**每个字段的名称均为链接。单击名称可以在链接视图中显示有关该字段的更多信息。请参阅主题第 16 页的『字段详细信息』以获取更多详细信息。

- **测量级别。**此列显示代表数据类型的图标。将鼠标悬停在图标上可以显示数据描述标签（连续、有序、名义等）。
- **预测能力。**只会对 ADP 建议的字段显示预测能力。如果未定义目标，则不会显示此列。预测能力范围从 0 到 1，其中较大的值表示“更好的”预测变量。通常，预测能力对于比较一个 ADP 分析内的预测变量有用，但不应跨分析比较预测能力值。

操作摘要

对于自动数据准备的每个操作，将会转换和/或过滤掉输入预测变量。保留的字段将用于下一个操作。在最后步骤中保留的字段将被建议用于建模，转换和构建预测变量的输入则被过滤掉。

“操作摘要”是一张简单列表，列出了 ADP 所执行的处理操作。如果任何**操作**带有下划线，单击可在链接视图中显示有关所执行操作的更多信息。请参阅 第 17 页的『操作详细信息』主题以获取更多信息。

注：只会显示每个字段的原始和最终转换版本，而不会显示在分析过程中使用的任何中间版本。

预测能力

在首次运行分析时缺省显示，或者在“字段处理概要”主视图中选择了**建议在分析中使用的预测变量**时显示，该图表显示建议预测变量的预测能力。字段按其预测能力排序，预测能力值最高的字段显示在顶端。

对于普通预测变量的转换后版本，其字段名称将反映您在“设置”选项卡的“字段名称”面板中选择的后缀，例如：*_transformed*

测量级别图标显示在各个字段名后面。

每个建议预测变量的预测能力通过线性回归或 naïve Bayes 模型进行计算，具体取决于目标是连续还是分类。

字段表

在“字段处理摘要”主视图中单击**目标**、**预测变量**或**未使用预测变量**时显示，“字段表”视图显示一个简单表，其中列出了相关特征。

该表包含两列:

- **名称。**预测变量名。

  对于目标变量，不论其是否已转换，始终使用字段的原始名称或标签。

  对于普通预测变量的转换后版本，其名称将反映您在“设置”选项卡的“字段名称”面板中选择的后缀，例如: *_transformed*

  对于从日期和时间派生的字段，将使用最终转换版本的名称，例如: *bdate_years*

  对于构建的预测变量，将使用构建预测变量的名称，例如: *Predictor1*。
- **测量级别。**此列显示代表数据类型的图标。

  对于目标，**测量级别**始终反映转换后的版本（如果目标已转换）。例如，从有序（有序集合）转换为连续（范围、刻度），反之亦然。

字段详细信息

在“字段”主视图中单击任何**名称**时显示，“字段详细信息”视图包括选定字段的分布、缺失值和预测能力图表（如果适用）。此外，字段的处理历史和转换后的字段名称也将显示（如果适用）。

对于每个图表集，两个版本将并排显示，以比较字段在应用转换前后的情况。如果字段的转换后版本不存在，则只显示原始版本的图表。对于派生的日期或时间字段和构建的预测变量，只显示新预测变量的图表。

注: 如果字段因为类别太多而被排除，则只显示处理历史。

分布图

连续字段分布显示为直方图，并叠放一条正态分布曲线，还有一条平均值垂直参考线。类别字段显示为条形图。

直方图带有标签以显示标准差和偏度。不过，如果值个数等于或低于 2，或原始字段的方差低于 10-20，则不会显示偏度。

将鼠标悬停在图表的上方，可以显示直方图的平均值，或条形图中类别计数与占记录总数的百分比。

缺失值图表

该图表显示为饼图，以比较在应用转换前后的缺失值百分比。图表标签显示百分比。

如果 ADP 执行了缺失值处理，则转换后的饼图还应包含替换值作为标签，即用于替换缺失值的值。

将鼠标悬停在图表的上方，可以显示缺失值计数和占记录总数的百分比。

预测能力图表

对于建议的字段，以条形图形式显示转换前后的预测能力。如果目标已经过转换，则计算的预测能力对应于转换后的目标。

注: 如果未定义目标，或在“主视图”面板中单击目标，将不会显示预测能力图表。

将鼠标悬停在图表的上方，可以显示预测能力值。

处理历史表

该表格显示字段的转换后版本是如何派生的。ADP 采取的操作按照其执行顺序列出。不过，对于某些步骤，可能对特定字段执行了多个操作。

注: 该表格不显示未转换字段的处理历史。

表中的信息分为二或三列:

- **操作。**操作的名称。例如，连续预测变量。请参阅主题『操作详细信息』以获取更多详细信息。
- **详细信息。**所执行处理的列表。例如，转换成标准单位。
- **函数。**针对构建的预测变量显示，其中显示输入字段的线性组合，例如，0.06*age + 1.21*height。

操作详细信息

在“操作摘要”主视图中选择任何带有下划线的**操作**时显示，“操作详细信息”链接视图显示所执行的每个处理步骤的操作相关与通用信息。首先显示操作相关的详细信息。

对于每个操作，描述用作标题位于链接视图的顶部。操作相关详细信息显示在标题下方，可能包括派生预测变量数目、字段重新设计、目标转换、类别合并或重新排序和预测变量构建或排除等详细信息。

在处理每个操作时，在处理过程中使用的预测变量数可能会变化，例如，排除或合并预测变量。

注: 如果某个操作已关闭，或未指定目标，则在“操作摘要”主视图中单击该操作时，会在操作详细信息位置显示一条错误消息。

有 9 个可能的操作，不过对于每个分析而言，这些操作并非都有必要使用。

文本字段表

该表显示下列项的数目:

- 从分析中排除的预测变量。

日期和时间预测变量表

该表显示下列项的数目:

- 从日期和时间预测变量派生的持续时间。
- 日期和时间元素。
- 派生的日期和时间预测变量总数。

如果已计算了任何日期持续时间，则参考日期或时间将显示为脚注。

预测变量过滤表

该表显示从处理中排除的以下预测变量数目:

- 常量。
- 缺失值过多的预测变量。
- 在单个类别中有太多个案的预测变量。
- 类别过多的名义字段（集合）。
- 过滤出的预测变量总数。

检查测量级别表

该表显示重新设计、分解成以下项的字段数目:

- 重新强制转换为连续字段的有序字段（有序集合）。
- 重新强制转换为有序字段的连续字段。
- 重新设计总数。

如果输入字段（目标或预测变量）并非连续或有序，这将显示为脚注。

离群值表

该表显示离群值处理方式的计数。

- 发现并修整其离群值的连续字段数，或发现离群值并将其设为缺失值的连续字段数，具体取决于您在“设置”选项卡的“准备输入和目标”面板上的设置。
- 由于在离群值处理后为常量，而被排除的连续字段数。

离群值分界值显示为脚注。如果输入字段（目标或预测变量）不是连续的，还会显示另一个脚注。

缺失值表

该表显示已替换缺失值、分解为以下项目的字段数:

- 目标。如果未指定目标，则不显示此行。
- 预测变量。它将进一步分解为名义（集合）、有序（有序集合）和连续特征数。
- 被替换的缺失值总数。

目标表

该表显示目标是否被转换，显示为:
- 到正态的 Box-Cox 转换。这将进一步分解为显示指定标准（平均值和标准差）和 Lambda 的列。
- 对其重新排序以提高稳定性的目标类别。

分类预测变量表

该表显示以下分类预测变量的数目:
- 按最低到最高重新排序其类别以提高稳定性。
- 合并其类别以最大化目标关联。
- 合并其类别以处理松散类别。
- 由于与目标关联程度过低而被排除。
- 由于在合并后为常量而被排除。

如果没有分类预测变量，则显示相应脚注。

连续预测变量表

有两个表。第一个表格显示以下转换数之一:
- 转换成标准单位的预测变量值。此外，还会显示转换的预测变量数、指定的平均值和标准差。
- 映射到通用范围的预测变量值。此外，还会显示通过最值法转换的预测变量数，以及指定的最小值和最大值。
- 分箱化的预测变量值和预测变量数。

第二个表显示预测变量空间构建详细信息，显示为以下预测变量的数目:
- 已构建。
- 由于与目标关联程度过低而被排除。
- 由于在分箱化后为常量而被排除。
- 由于在构建后为常量而被排除。

如果未输入连续预测变量，则显示相应脚注。

逆转换得分

如果目标已被 ADP 转换，则使用转换后目标构建的后续模型将针对转换后的单位评分。要解释和使用结果，您必须将预测值转换回原始刻度。

1. 要逆转换得分，从菜单中选择:

   转换 > 准备数据以供建模 > 逆转换得分...
2. 选择要逆转换的字段。此字段应包含转换后目标的模型预测值。
3. 为新字段指定后缀。此新字段将包含采用未转换目标的原始刻度的模型预测值。

4. 指定包含 ADP 转换的 XML 文件位置。这应当是从交互式或自动数据准备对话框中保存的文件。请参阅主题第 13 页的『应用和保存转换』以获取更多详细信息。

第 5 章 标识异常个案

“异常检测”过程查找基于聚类组标准值偏差的异常个案。该过程设计为在探索性数据分析步骤中，快速检测到用于数据审核的异常个案，并优先于任何推论性数据分析。此算法设计为一般“异常检测”；即异常个案的定义不被指定为任何特定应用程序，例如对保健行业中异常付款模式的检测或对金融业中洗钱行为的检测，其中对异常的定义可以被很好地界定。

示例。雇用的构建中风治疗效果预测模型的数据分析人员对数据质量非常关注，因为这类模型对异常观察值十分敏感。某些偏离的观察值表示真正唯一的个案，因此不适合用于预测，而其他观察值是由数据输入错误导致的，其值从技术上说是“正确”的，因此不能被数据验证过程捕获。“标识异常个案”过程找出并报告这些离群值，以便分析人员能够确定如何处理这些值。

统计信息。该过程生成对等组、连续和分类变量的对等组标准值、基于对等组标准值偏差的异常指标，以及对被视为异常的个案影响最大的变量影响值。

数据注意事项

数据。 此过程既处理连续变量也处理分类变量。每行表示一个不同观察值，每列表示一个对等组以其为基础的不同变量。个案标识变量可在用于标记输出的数据文件中获得，但不能用于分析中。允许缺失值。被指定的权重变量可以忽略。

检测模型可用于新检验数据文件。检验数据元素必须与培训数据元素一致。并且，根据算法设置，用于创建模型的缺失值处理方法可适用于优先于评分的检验数据文件。

个案顺序。 注意，解决方案可取决于个案顺序。要使顺序的影响降至最低程度，可随机个案等级排序的顺序。想要验证给定解的稳定性，您可能想要通过以不同随机顺序排序的案例来得到多个不同的解。在文件非常大的情况，可使用以不同随机顺序排序的个案样本运行多次。

假设。 算法假设所有变量都为不恒定且独立的，并且没有个案具有含有任何输入变量的缺失值。假设每个连续变量具有正态（高斯）分布，假设每个分类变量具有多项分布。经验内部检验表明，该过程对于违反独立性假设和分布假设均相当稳健，但应了解这些假设符合的程度。

标识异常个案

1. 从菜单中选择:

   数据 > 标识异常个案...
2. 选择至少一个分析变量。
3. 还可以选择一个个案标识变量用于标记输出。

测量级别未知的字段

当数据集中的一个或多个变量（字段）的测量级别未知时，将显示测量级别警告。由于测量级别会影响该过程的计算结果，因此所有变量必须都定义有测量级别。

扫描数据。 读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，该过程可能需要一些时间。

手动分配。 打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

由于测量级别对该过程很重要，因此您无法访问运行该过程的对话框，除非所有字段均定义了测量级别。

标识异常个案: 输出

异常个案及其被视为异常的原因的列表。此选项可生成三个表:

- 异常个案指标列表显示标识为异常的个案，并显示其相应的异常指标值。
- 异常个案 Peer ID 列表显示异常个案及其相应对等组的相关信息。
- 异常原因列表显示个案号、原因变量、变量影响值、变量值以及每个原因的变量的标准值。

所有表都根据异常指标按降序排列。此外，如果在“变量”选项卡上指定了个案标识变量，则会显示个案的 ID。

摘要。此组中的控件可生成分布摘要。

- **对等组标准值。**此选项显示连续变量标准值表（如果分析中使用了任何连续变量）以及分类变量标准值表（如果分析中使用了任何分类变量）。连续变量标准值表显示每个对等组的每个连续变量的平均值和标准差。分类变量标准值表显示每个对等组的每个分类变量的众数（最大类别）、频率和频率百分比。连续变量的平均值和分类变量的众数在分析中用作标准值。
- **异常指标。**异常指标摘要显示标识为最不正常个案的异常指标的描述统计。
- **按分析变量列出出现的原因。**对于每个原因，该表将每个变量的出现频率和频率百分比显示为原因。该表还报告每个变量的影响的描述统计。如果在“选项”选项卡上将最大的原因数量设置为 0，则此选项不可用。
- **已处理的个案数。**个案处理摘要显示活动数据集中所有个案的计数和计数百分比、分析中包含和排除的个案，以及每个对等组中的个案。

标识异常个案: 保存

保存变量。此组中的控件允许您将模型变量保存到活动数据集。您也可以选择将存在名称冲突的现有变量替换为要保存的变量。

- **异常指标。**将每个个案的异常指标值保存到具有指定名称的变量中。
- **对等组。**将对等组 ID、个案计数以及每个个案的以百分比表示的大小保存到具有指定根名称的变量中。例如，如果指定了根名称 *Peer*，则会生成变量 *Peerid*、*PeerSize* 和 *PeerPctSize*。*Peerid* 为个案的对等组 ID，*PeerSize* 为组的大小，而 *PeerPctSize* 为用百分比表示的组大小。
- **原因。**使用指定的根名称保存原因变量集。原因变量集包含作为原因的变量的名称、变量影响测量、变量自身的值以及标准值。变量集的数量取决于在“选项”选项卡上请求的原因的数目。例如，如果指定根名称 *Reason*，则会生成变量 *ReasonVar_k*、*ReasonMeasure_k*、*ReasonValue_k* 和 *ReasonNorm_k*，其中 *k* 是第 *k* 个原因。如果原因数量设置为 0，则此选项不可用。

导出模型文件。允许以 XML 格式保存模型。

标识异常个案: 缺失值

“缺失值”选项卡用于控制对用户缺失值和系统缺失值的处理。

- **从分析中排除缺失值。**具有缺失值的个案会从分析中排除。
- **在分析中包括缺失值。**连续变量的缺失值将替换为它们对应的总平均值，分类变量的缺失类别将分组并视为有效类别。处理过的变量随后在分析中使用。或者，您也可以请求创建表示每个个案中缺失变量的比例的附加变量并在分析中使用该变量。

标识异常个案: 选项

异常个案的标识条件。这些选择确定在异常列表中包括多少个个案。

- **具有最高异常指标值的个案所占的百分比。**指定一个小于或等于 100 的正数。
- **具有最高异常指标值的个案的固定数量。**指定一个正整数，该整数小于或等于分析中使用的活动数据集的个案总数。
- **仅标识异常指标值符合或超过最小值的个案。** 指定一个非负数。如果某个个案的异常指标值大于或等于指定分界点，则将该个案视为异常个案。此选项与**个案百分比**和**个案的固定数量**选项一起使用。例如，如果指定 50 作为固定数量，并指定 2 作为分界值，则异常列表最多可包含 50 个个案，每个个案的异常指标值都大于等于 2。

对等组的数量。该过程将搜索指定的最小值和最大值之间的最佳对等组数量。该值必须为正整数，并且最小值不能超过最大值。如果指定的值相等，则该过程假定对等组的数量是固定的。

注: 根据数据中的变动量，有时数据可支持的对等组的数量可能小于指定的最小数量。在这种情况下，该过程可能会生成数量较少的对等组。

最大的原因数量。原因包括变量影响测量、此原因的变量名、变量的值以及相应对等组的值。指定一个非负整数，如果此值等于或超过分析中使用的已处理变量的数量，则会显示所有变量。

DETECTANOMALY 命令的附加功能

使用命令语法语言还可以:

- 在分析中省略活动数据集中的一些变量，而不显式指定所有分析变量（使用 EXCEPT 子命令）。
- 指定通过调整平衡连续和分类变量的影响（使用 CRITERIA 子命令的 MLWEIGHT 关键字）。

请参阅*命令语法参考*以获取完整的语法信息。

第 6 章 最优分箱化

“最优分箱化”过程通过将每个变量的值分布到分箱中离散化一个或多个刻度变量（因此称为**分箱化输入**变量）。分箱的构成根据“监督”分箱化过程的分类向导变量得以最优化。然后，可以使用分箱而非原始数据值进行进一步的分析。

示例。减少变量具有的不同值的数量具有多种用途，包括:

- 其他过程的数据要求。离散化变量可作为分类变量用于需要分类变量的过程。例如，“交叉表格”过程要求所有变量均为分类变量。
- 数据隐私。报告分箱化值而不是实际值可帮助保护数据源的隐私。“最优分箱”过程可指导分箱的选择。
- 速度性能。有些过程在处理较少数量的不同值时更加有效。例如，使用离散化变量时“多项 Logistic 回归”的速度会提高。
- 揭示数据的完全分离或准完全分离。

最优分箱化与可视分箱化。“可视分箱化”对话框提供了多种不使用向导变量创建分箱的自动方法。这些“未受监督”的规则对于生成描述统计（例如频率表）十分有用，但如果最终目标是生成预测模型，则“最优分箱化”更好。

输出。该过程生成分箱的分割点以及每个分箱化输入变量的描述统计的表。此外，您可以将新变量保存到包含分箱化输入变量的分箱化值的活动数据集中，并将分箱化规则作为命令语法保存以便用于分箱化新数据。

最优分箱化数据注意事项

数据。此过程需要分箱化输入变量是数值型刻度变量。向导变量应是分类变量，可以是字符串或数值。

获取最优分箱化

1. 从菜单中选择:

   转换 > 最优分箱化...
2. 选择一个或多个分箱化输入变量。
3. 选择一个向导变量。

缺省情况下不会生成包含分箱化数据值的变量。使用保存选项卡以保存这些变量。

最优分箱化: 输出

“输出”选项卡控制结果的显示。

- **分箱的端点。**显示每个分箱化输入变量的端点集。
- **已分箱化变量的描述统计。**对于每个分箱化输入变量，此选项显示具有有效值的个案数、具有缺失值的个案数、不同有效值的数量以及最小值和最大值。对于向导变量，此选项显示每个相关分箱化输入变量的类分布。
- **已分箱化变量的模型熵。**对于每个分箱化输入变量，此选项显示相对于向导变量的变量预测准确性的测量。

最优分箱化: 保存

将变量保存到活动数据集。包含分箱化数据值的变量可在进一步分析中代替初始变量。

将分箱化规则另存为语法。生成可用于分箱化其他数据集的命令语法。记录的规则基于分箱化算法确定的分割点。

最优分箱化: 缺失值

"缺失值"选项卡指定是通过成列删除还是成对删除处理缺失值。用户缺失值始终被视为无效。将初始变量值记录到新变量中时，用户缺失值会转换为系统缺失值。

- **成列。**此选项针对每个向量和分箱化输入变量对进行操作。该过程将利用向导和分箱化输入变量的具有非缺失值的所有个案。
- **列表**此选项跨"变量"选项卡上指定的所有变量进行操作。如果某个个案的任何变量缺失，则排除整个个案。

最优分箱化: 选项

预处理。"预分箱化"具有许多不同值的分箱化输入变量可缩短处理时间，而不会使最终分箱的质量发生大幅度下滑。分箱的最大数量为创建的分箱的数量设置了一个上限。这样，如果指定 1000 作为最大值，但分箱化输入变量的不同值的数量少于 1000，则为分箱化输入变量创建的预处理分箱的数量将等于分箱化输入变量中不同值的数量。

稀疏填充的分箱。有时候，该过程可能会生成仅具有很少个案的分箱。下面的方案会删除这些伪分割点:

对于给定的变量，假定该算法找到了 n_{final} 个分割点，从而有 $n_{final}+1$ 个分箱。对于分箱 $i = 2$、.....、n_{final}（从值第二低的分箱到值第二高的分箱），计算

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

其中 *sizeof(b)* 是分箱中的个案数。

当此值小于指定的合并阈值时，b_i 被认为是稀疏填充的，并将与 b_{i-1} 或 b_{i+1} 合并，具体取决于哪一个具有较低的类信息熵。

该过程仅穿过这些分箱一次。

分箱端点。此选项指定如何定义区间下限。因为该过程自动确定分割点的值，所以这主要是偏好的问题。

第一个(最低)分箱/最后一个(最高)分箱。这些选项指定如何定义每个分箱化输入变量的最小和最大分割点。通常情况下，该过程假设分箱化输入变量可采用实数线上的任何值，但是，如果由于某些理论或实际的原因需要限制该范围，则可通过最低值/最高值进行限制。

OPTIMAL BINNING 命令的附加功能

使用命令语法语言还可以:

- 通过均等频率方法执行未受监督的分箱化（使用 CRITERIA 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

声明

本信息是为在美国提供的产品和服务编写的。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

本条款不适用英国或任何这样的条款与当地法律不一致的国家或地区: International Business Machines Corporation“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可以随时对本资料中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的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 Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

此处所含的性能数据均在受控环境下决定。因此，在其他操作环境中获得的结果可能差异较大。有些测量可能在开发级的系统中进行，不保证这些测量结果与常用系统上的测量结果相同。此外，有些测量结果可能通过推断来估计得出。实际结果可能有所差异。此文档的用户应针对其具体环境验证适用的数据。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。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。

有关 IBM 未来方向或意向的所有声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

版权许可:

本信息包括源语言形式的样本应用程序，这些样本说明不同操作平台上的编程方法。如果是为按照在编写样本程序的操作平台上的应用程序编程接口 (API) 进行应用程序的开发、使用、经销或分发为目的，您可以任何形式对这些样本程序进行复制、修改、分发，而无须向 IBM 付费。这些示例并未在所有条件下作全面测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。本样本程序仍然是“按现状”提供的，不附有任何种类的保证。对于因使用样本程序所引起的任何损害，IBM 概不负责。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明:

© （贵公司的名称）（年）。此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

© Copyright IBM Corp. _（输入年份）_. All rights reserved.

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。当前的 IBM 商标列表，可从 Web 站点 www.ibm.com/legal/copytrade.shtml 上“版权和商标信息”部分获取。

Adobe、Adobe 徽标、PostScript 和 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 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

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 及/或其分支机构的商标和注册商标。

索引

M

Printed in China