IBM SPSS Decision Trees 20

注意：使用本信息及其支持的产品之前，请阅读注意事项第 99 页码下的一般信息。

此版本适用于 IBM® SPSS® Statistics 20 及所有后续发布和修订，除非在新版本中另有说明。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。Decision Trees 可选附加模块提供本手册中描述的其他分析方法。此 Decision Trees 附加模块必须与 SPSS Statistics Core 系统一起使用，并已完全集成到了该系统中。

关于 IBM Business Analytics

IBM Business Analytics 软件提供决策者赖以提高业务绩效的完整、一致和准确的信息。包括业务智能、预测分析、财务状况和战略管理以及分析应用程序在内的一整套产品组合让您即刻、清楚地了解当前绩效并依此采取行动，以及能够预测未来的成果。结合丰富的行业解决方案、被证明的实践经验和专业的服务，无论公司规模大小，都能促使其获得最高的产能、自信自觉地做出决定并得到更好的成绩。

作为产品组合的一部分，IBM SPSS Predictive Analytics 软件帮助公司预测未来实践并采取积极行动，促使其获得更好的业务成果。全世界的商业政府和学术客户依赖 IBM SPSS 技术，因其具有竞争力的优势，能够吸引、留住和发展客户，同时减少欺诈和减轻风险。通过将 IBM SPSS 软件融入日常运营中，公司成为具有预测性的企业，能够引导和自觉做出决策，以满足业务目标，实现可观的竞争优势。欲知更多信息或联系代表，请访问 http://www.ibm.com/spss。

技术支持

我们提供有“技术支持”以维护客户。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。如要联系技术支持，请访问 IBM Corp. 网站，网址为 http://www.ibm.com/support。在请求协助时，请准备好您和您组织的 ID 以及支持协议。

针对学生的技术支持

如果您是使用任何学生版 IBM SPSS 软件产品的学生，请访问我们专为学生提供的在线教育解决方案 (http://www.ibm.com/spss/rd/students/)页面。如果您是使用大学提供的 IBM SPSS 软件副本的学生，请联系所在大学的 IBM SPSS 产品协调员。

客户服务

如果对发货或帐户存在任何问题，请联系您当地的办事处。请先准备好您的序列号以供识别。

培训讲座

IBM Corp. 提供公开的以及现场的培训讲座。所有讲座都是以实践小组为特色的。讲座将定期在各大城市开展。有关这些讲座的更多信息，请前往 http://www.ibm.com/software/analytics/spss/training。

附加出版物

SPSS Statistics：数据分析指南、SPSS Statistics：Statistical Procedures Companion 和 SPSS Statistics：Advanced Statistical Procedures Companion（由 Marija Norušis 编写，并已由 Prentice Hall 出版）作为建议的补充材料提供。这些出版物涵盖 SPSS Statistics Base 模块、Advanced Statistics 模块和 回归模块中的统计过程。无论您是刚开始从事数据分析工作，还是已准备好使用高级应用程序，这些书籍都将帮助您最有效地利用在 IBM® SPSS® Statistics 产品中找到的功能。有关其他信息，包括出版物的内容和示例章节，请参阅作者的网站： http://www.norusis.com

内容

部分 I: 用户指南

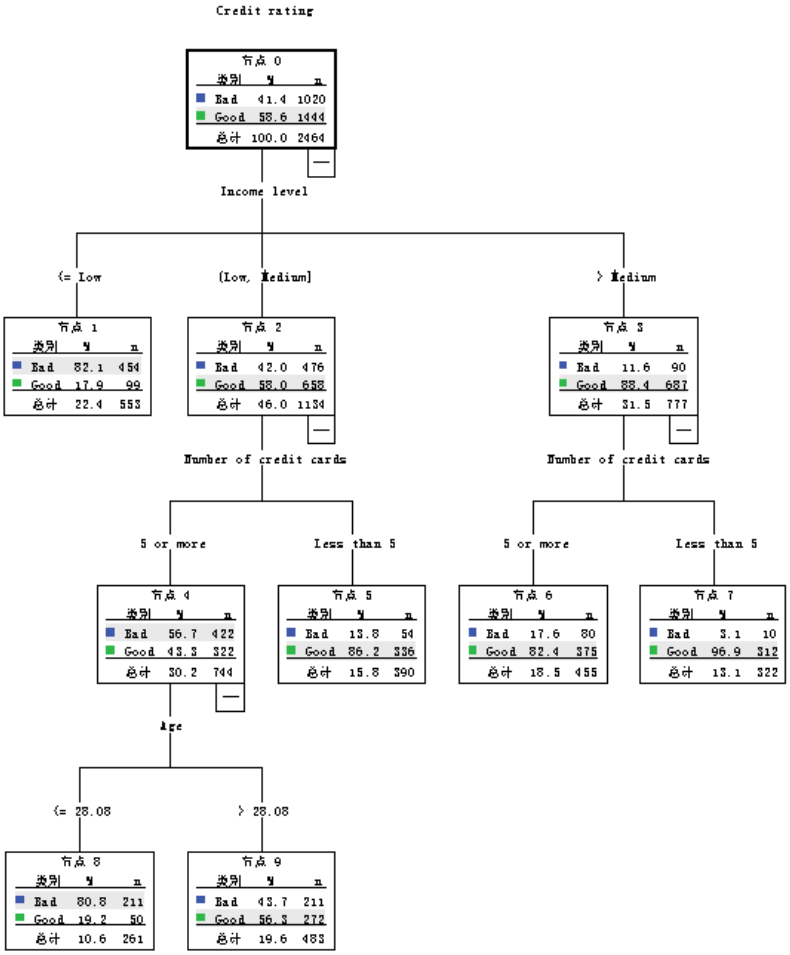
部分 II: 示例

部分 I:
用户指南

章
1

创建决策树

图片 1-1
决策树



“决策树”过程创建基于树的分类模型。它将个案分为若干组，或根据自变量（预测变量）的值预测因变量（目标变量）的值。此过程为探索性和证实性分类分析提供验证工具。

此过程可以用于：

分段。确定可能成为特定组成员的人员。

层次。将个案指定为几个类别之一，如高风险组、中等风险组和低风险组。

预测。创建规则并使用它们预测将来的事件，如某人将拖欠贷款或者车辆或住宅潜在转售价值的可能性。

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

交互确定。确定仅与特定子组有关的关系，并在正式的参数模型中指定这些关系。

类别合并和连续变量离散化。以最小的损失信息对组预测类别和连续变量进行重新编码。

示例。一家银行希望根据贷款申请人是否表现出合理的信用风险来对申请人进行分类。根据各种因素（包括过去客户的已知信用等级），您可以构建模型以预测客户将来是否可能拖欠贷款。

基于树的分析提供了一些引人注意的功能：

- 通过分析功能，您可以确定具有高风险或低风险的同类组。
- 还可轻松构建用于预测个别个案的规则。

数据注意事项

数据。因变量和自变量可以是：

- **标定.** 当变量值表示不具有内在等级的类别时，该变量可以作为名义变量；例如，雇员任职的公司部门。名义变量的示例包括地区、邮政编码和宗教信仰。
- **有序.** 当变量值表示带有某种内在等级的类别时，该变量可以作为有序变量；例如，从十分不满意到十分满意的服务满意度水平。有序变量的示例包括表示满意度或可信度的态度分数和优先选择评分。
- **刻度.** 当变量值表示带有有意义的度规的已排序类别时，该变量可以作为刻度（连续）变量对待，以便在值之间进行合适的距离比较。刻度变量的示例包括以年为单位的年龄和以千美元为单位的收入。

频率权重如果加权有效，则将分数权重四舍五入为最接近的整数；所以，为权重值小于 0.5 的个案指定权重 0，因而会从分析中排除它们。

假设。此过程假定已经为所有分析变量指定适当的测量级别，一些功能假定分析中包括的因变量的所有值都定义了值标签。

- **测量级别。**测量级别影响树计算；因此，应该为所有变量指定适当的测量级别。默认情况下，假定数值变量是刻度变量，而字符串变量假定为名义变量，这可能没有准确地反映真实的测量级别。变量列表中每个变量旁的图标标识变量类型。


尺度


名义


有序

可以暂时更改变量的测量级别，方法是在源变量列表中右键单击该变量，然后从上下文菜单中选择测量级别。

- **值标签。**此过程的对话框界面假定分类（名义、有序）因变量的所有非缺失值已定义值标签，或者它们都没有定义值标签。一些功能是不可用的，除非分类因变量的至少两个非缺失值具有值标签。如果至少两个非缺失值已经定义了值标签，则将从分析中排除带有其他没有值标签的值的所有个案。

获取决策树

▶ 从菜单中选择：

分析 > 分类 > 树...

图片 1-2
“决策树”对话框

▶ 选择一个因变量。

▶ 选择一个或多个自变量。

▶ 选择生长法。

根据需要，您可以：

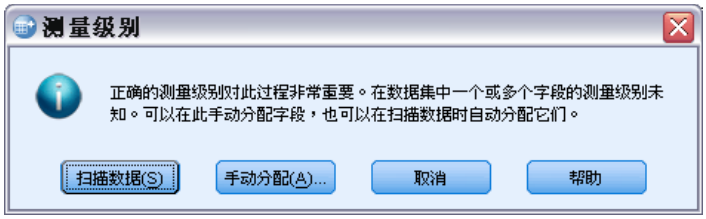
- 更改源列表中所有变量的测量级别。
- 强制自变量列表中的第一个变量作为第一个拆分变量进入模型。
- 选择定义个案对树生长过程的影响程度的影响变量。影响值较低的个案影响较小；而影响值较高的个案影响较大。影响变量值必须为正。
- 验证树。
- 自定义树生长条件。

- 将终端节点编号、预测值和预测概率保存为变量。
- 以 XML (PMML) 格式保存模型。

具有未知测量级别的字段

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。

图片 1-3
测量级别警报



- **扫描数据。** 读取活动数据集中的数据，并分配默认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。

更改测量级别

▶ 右键单击源列表中的变量。

▶ 从弹出的上下文菜单中选择测量级别。

这将暂时更改测量级别以用于“决策树”过程。

生长法

可用的生长法如下：

CHAID. 卡方自动交互检测。在每一步，CHAID 选择与因变量有最强交互作用的自变量（预测变量）。如果每个预测变量的类别与因变量并非显著不同，则合并这些类别。

穷举 CHAID. CHAID 的一种修改版本，其检查每个预测变量所有可能的拆分。

CRT. 分类和回归树。CRT 将数据拆分为若干尽可能与因变量同质的段。所有个案中因变量值都相同的终端节点是同质的“纯”节点。

QUEST. 快速、无偏、有效的统计树。一种快速方法，它可避免其他方法对具有许多类别的预测变量的偏倚。只有在因变量是名义变量时才能指定 QUEST。

每种方法都有其各自的优点和限制，其中包括：

	CHAID*	CRT	QUEST
基于卡方**	X		
替代自变量（预测变量）		X	X
树修剪		X	X
多阶节点拆分	X		
二元节点拆分		X	X
影响变量	X	X	
先验概率		X	X
误分类成本	X	X	X
快速计算	X		X

*包括穷举 CHAID。

**QUEST 也将卡方测量用于名义自变量。

选择类别

图片 1-4
“类别”对话框

对于分类（名义、有序）因变量，可以：

- 控制将哪些类别包括在分析中。
- 确定目标类别。

包括/排除类别

可以将分析限制为因变量的特定类别。

- 因变量的值在“排除”列表中的个案不会包括在分析中。
- 对于名义因变量，您也可以在分析中包括用户缺失的类别。（默认情况下，用户缺失的类别显示在“排除”列表中。）

目标类别

选定（选中）的类别被视为分析中主要对其感兴趣的类别。例如，如果主要对确定最可能拖欠贷款的那些人感兴趣，则可能选择“bad”信用等级类别作为目标类别。

- 没有默认的目标类别。如果未选定任何类别，则某些分类规则选项和与增益相关的输出将不可用。
- 如果选定了多个类别，则为每个目标类别生成单独的增益表和图表。
- 将一个或多个类别指定为目标类别，对树模型、风险估计或误分类结果没有影响。

类别和值标签

此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

包括/排除类别和选择目标类别

▶ 在“决策树”主对话框中，选择带有两个或更多已定义值标签的分类（名义、有序）因变量。

▶ 单击类别。

验证

图片 1-5
“验证”对话框

通过验证可以评估树结构广义化为更大总体的程度。可以使用两种验证方法：交叉验证和分割样本验证。

交叉验证

交叉验证将样本分割为许多子样本（或**样本群**）。然后，生成树模型，并依次排除每个子样本中的数据。第一个树基于第一个样本群的个案之外的所有个案，第二个树基于第二个样本群的个案之外的所有个案，依此类推。对于每个树，估计其误分类风险的方法是将树应用于生成它时所排除的子样本。

- 最多可以指定 25 个样本群。该值越大，每个树模型中排除的个案数就越小。
- 交叉验证生成单个最终树模型。最终树经过交叉验证的风险估计计算为所有树的风险的平均值。

分割样本验证

对于分割样本验证，模型是使用训练样本生成的，并在延续样本上进行测试。

- 您可以指定训练样本大小（表示为样本总大小的百分比），或将样本分割为训练样本和测试样本的变量。

- 如果使用变量定义训练样本和测试样本，则将变量值为 1 的个案指定给训练样本，并将所有其他个案指定给测试样本。该变量不能是因变量、权重变量、影响变量或强制的自变量。
- 您可以同时显示训练样本和测试样本的结果，或者仅显示测试样本的结果。
- 对于小的数据文件（个案数很少的数据文件），应该谨慎使用分割样本验证。训练样本很小可能会导致很差的模型，因为在某些类别中，可能没有足够的个案使树充分生长。

树生长条件

可用的生长条件可能取决于生长法、因变量的测量级别或这两者的组合。

增长限制

图片 1-6
“条件”对话框，“增长限制”选项卡

使用“增长限制”选项卡，可以限制树中的级别数和控制父节点和子节点的最小个案数。

最大树深度。控制根节点下的最大增长级别数。对于 CHAID 和穷举 CHAID 方法，**自动**设置将树限制为根节点下的三个级别；而对于 CRT 和 QUEST 方法，则限制为根节点下的五个级别。

最小个案数。控制节点的最小个案数。不会拆分不满足这些条件的节点。

- 增大最小值往往会生成具有更少节点的树。
- 而减小最小值则会生成具有更多节点的树。

对于个案数目很小的数据文件，父节点的默认值（100 个个案）和子节点的默认值（50 个个案）有时可能导致树在根节点下没有任何节点；在这种情况下，减小最小值可能产生更有用的结果。

CHAID 条件

图片 1-7
“条件”对话框，“CHAID”选项卡

对于 CHAID 和穷举 CHAID 方法，您可以控制：

显著性水平。您可以控制用于拆分节点和合并类别的显著性值。对于这两个条件，默认的显著性水平都是 0.05。

- 对于拆分节点，值必须大于 0 且小于 1。较小的值往往会产生具有较少节点的树。
- 对于合并类别，该值必须大于 0 且小于或等于 1。要阻止合并类别，请指定值 1。对于刻度自变量，这意味着最终树中变量的类别数是指定的区间数（默认值是 10）。 有关详细信息，请参阅第 10 页码CHAID 分析的刻度区间。

卡方统计。对于有序因变量，用于确定节点拆分和类别合并的卡方是使用似然比方法计算的。对于名义因变量，可以选择以下方法：

- **Pearson。**此方法提供更快的计算，但是对于小样本应该谨慎使用它。这是默认方法。
- **似然比。**此方法比 Pearson 方法更稳健，但是所用的计算时间更长。对于小样本，这是首选的方法。

模型估值。对于名义和有序因变量，可以指定：

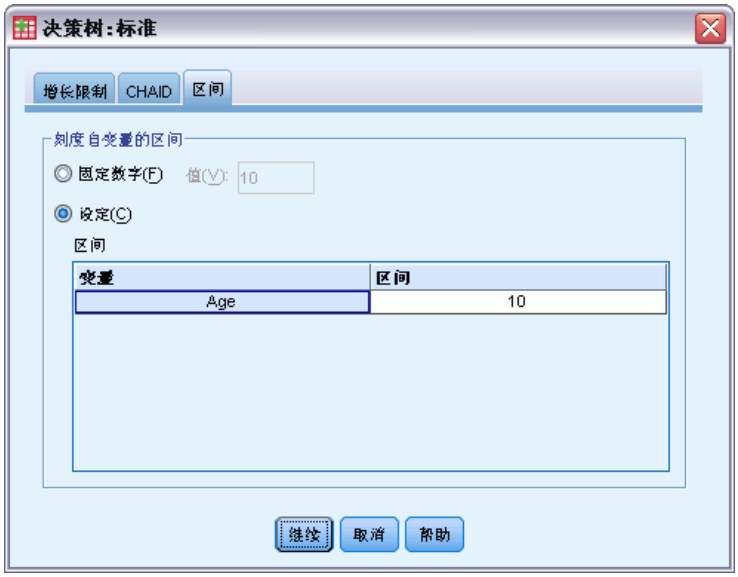
- **最大迭代次数。**默认值是 100。如果树由于达到最大迭代次数而停止生长，您可能希望增大最大值，或更改控制树生长的一个或多个其他条件。
- **期望单元格频率的最小更改。**该值必须大于 0 且小于 1。默认值是 0.05。较小的值往往会产生具有较少节点的树。

使用 Bonferroni 方法调整显著性值。对于多个比较，使用 Bonferroni 方法调整用于合并和拆分条件的显著性值。这是默认值。

允许重新拆分节点中的合并类别。除非明显阻止类别合并，否则该过程将尝试将自变量（预测变量）类别合并在一起，以产生描述模型的最简单的树。此选项允许该过程重新拆分合并的类别（如果这样可以提供更好的方案）。

CHAID 分析的刻度区间

图片 1-8
“条件”对话框，“区间”选项卡



在 CHAID 分析中，刻度自变量（预测变量）在分析之前始终分段到离散组（例如，0 - 10、11 - 20、21 - 30 等）中。您可以控制初始/最大组数（尽管该过程可能在初始拆分后合并连续组）：

- **固定数字。**所有的刻度自变量最初都分段到相同数量的组中。默认值为 10。
- **自定义。**每个刻度自变量最初都分段到该变量所指定数量的组中。

指定刻度自变量的区间

▶ 在“决策树”主对话框中，选择一个或多个刻度自变量。

▶ 选择 CHAID 或穷举 CHAID 作为生长法。

▶ 单击条件。

▶ 单击区间选项卡。

在 CRT 和 QUEST 分析中，所有拆分均为二元的，而且刻度和有序自变量的处理方式是相同的；因此，无法为刻度自变量指定多个区间。

CRT 条件

图片 1-9
“条件”对话框，“CRT”选项卡

CRT 生长法尝试最大化节点内的齐性。对不代表同质个案子集的节点，它的程度显示为**不纯值**。例如，其中所有个案都具有相同的因变量值的终端节点是无需进一步拆分（因为它是“纯的”）的同质节点。

可以选择用于度量不纯值的方法，以及拆分节点所需的不纯值中的最小减少值。

杂质测量。对于刻度因变量，使用最小二乘偏差 (LSD) 测量杂质。它为节点内的方差，并根据任意频率权重或影响值进行调整。

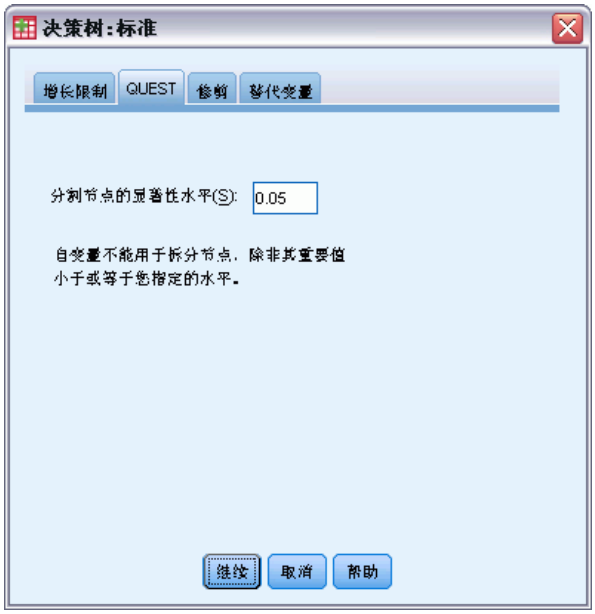
对于分类（名义、有序）因变量，可以选择杂质测量：

- **Gini。**找到可以根据因变量的值最大化子节点齐性的拆分。对于因变量的每个类别，Gini 基于成员身份的平方概率。它在节点中的所有个案都属于单个类别时达到其最小值（零）。这是默认度量。
- **两分法。**因变量的类别分组为两个子类。找到最适合于分隔两个组的拆分。
- **顺序两分法。**与两分法相似，但它只能对相邻类别进行分组。此度量仅可用于有序因变量。

最小改进更改。这是拆分节点所需的不纯值中的最小减少值。默认值为 .0001。较大的值往往会产生节点较少的树。

QUEST 条件

图片 1-10
“条件”对话框，“QUEST”选项卡



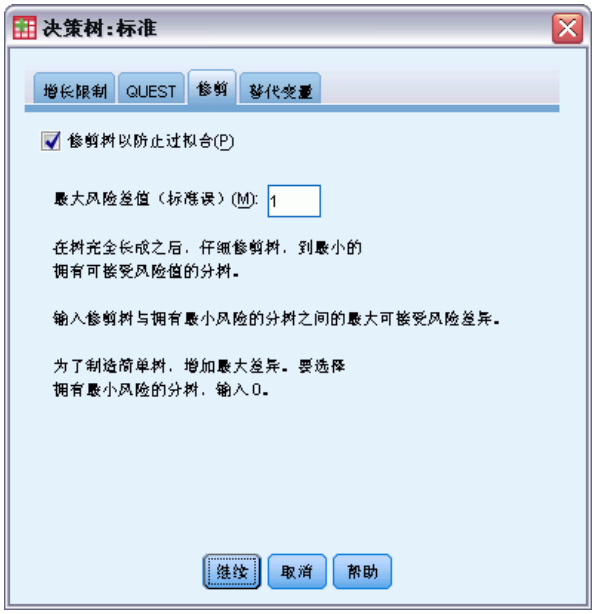
对于 QUEST 方法，可以指定用于拆分节点的显著性水平。除非显著性水平小于或等于指定的值，否则自变量不能用于拆分节点。该值必须大于 0 且小于 1。默认值为 0.05。较小的值往往会从最终模型中排除较多的自变量。

指定 QUEST 条件

- 在“决策树”主对话框中，选择一个名义因变量。
- 选择 QUEST 作为生长法。
- 单击条件。
- 单击 QUEST 选项卡。

修剪树

图片 1-11
“条件”对话框，“修剪”选项卡



对于 CRT 和 QUEST 方法，可以通过**修剪**树避免过拟合模型：在满足停止生长的条件之前保持树处于生长状态，然后根据指定的最大风险差值，自动将其修剪到最小子树。以标准误表示风险值。默认值为 1。该值必须为非负数。要获取具有最小风险的子树，请指定 0。

修剪与隐藏节点

创建修剪树时，从树中修剪的任何节点在最终树中都是不可用的。您可以以交互方式隐藏和显示最终树中的选定子节点，但是不能显示在创建树的过程中修剪的节点。 有关详细信息，请参阅第 35 页码第 2 章中的树编辑器。

替代变量

图片 1-12
“条件”对话框，“替代变量”选项卡

CRT 和 QUEST 可以将**替代变量**用于自变量（预测变量）。对于缺失该变量的值的个案，将使用与原始变量高度相关的其他自变量进行分类。这些备用预测变量称为替代变量。可以指定要在模型中使用的最大替代变量数。

- 默认情况下，最大替代变量数比自变量数小 1。换句话说，针对每个自变量，其他的所有自变量均可能被用作替代变量。
- 如果不希望模型使用替代变量，请指定 0 作为替代变量数。

选项

可用选项可能取决于生长法、因变量的测量级别和/或为因变量的值定义的值标签是否存在。

误分类成本

图片 1-13
“选项”对话框，“误分类成本”选项卡

对于分类（名义、有序）因变量，通过误分类成本，可以包括有关与错误分类关联的相对惩罚的信息。例如：

- 拒绝为信用良好的客户发放贷款的成本，可能与为之后拖欠贷款的客户提供贷款的成本不同。
- 将患有心脏病的高风险个人误分类为低风险的成本，可能比将低风险的个人误分类为高风险的成本要高得多。
- 向也许不会回复的个人发送大量邮件的成本可能非常低，但不向可能回复的个人发送邮件的成本却相对较高（在失去的收入方面）。

误分类成本和值标签

除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定误分类成本

- ▶ 在“决策树”主对话框中，选择带有两个或更多已定义值标签的分类（名义、有序）因变量。
- ▶ 单击**选项**。
- ▶ 单击**误分类成本**选项卡。
- ▶ 单击**自定义**。
- ▶ 在网格中输入一个或多个误分类成本。值必须为非负数。（在对角线上表示的正确分类始终为 0。）

填充矩阵。 在许多情况下，可能希望成本是对称的，—即，将 A 误分类为 B 的成本与将 B 误分类为 A 的成本是相同的。使用以下控件可以更轻松地指定对称的成本矩阵：

- **复制下三角形。**将矩阵的下三角形中的值（在对角线之下）复制到对应的上三角形单元格中。
- **复制上三角形。**将矩阵的上三角形中的值（在对角线之上）复制到对应的下三角形单元格中。
- **使用单元格平均值。**对于矩阵的每一半中的每个单元格，对这两个值（上三角形和下三角形）进行平均，并用平均值替换这两个值。例如，如果将 A 误分类为 B 的成本为 1，而将 B 误分类为 A 的成本为 3，则此控件会将这两个值都替换为平均值 (1+3)/2 = 2。

利润

图片 1-14
“选项”对话框，“利润”选项卡

对于分类因变量，可以将收入值和支出值指定给因变量的水平。

- 利润是通过收入减去支出计算出来的。
- 利润值影响增益表中的平均利润值和 ROI（投资回报）值。但它们不影响树模型的基础结构。
- 收入值和支出值必须为数值型，且必须为网格中显示的因变量的所有类别指定它们。

利润和值标签

此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定利润

- ▶ 在“决策树”主对话框中，选择带有两个或更多已定义值标签的分类（名义、有序）因变量。
- ▶ 单击**选项**。
- ▶ 单击**利润**选项卡。
- ▶ 单击**自定义**。
- ▶ 输入网格中列出的所有因变量类别的收入值和支出值。

先验概率

图片 1-15
“选项”对话框，“先验概率”选项卡

对于具有分类因变量的 CRT 和 QUEST 树，可以指定组成员身份的先验概率。**先验概率**是在了解有关自变量（预测变量）值的任何信息之前，对因变量的每个类别的总体相对频率的评估。使用先验概率有助于更正由不代表整个总体的样本中的数据导致的树的任何生长。

从训练样本获取（先验）。如果数据文件中因变量值的分布代表总体分布，则使用此设置。如果使用的是分割样本验证，则使用训练样本中的个案分布。

注意：由于在分割样本验证中个案是随机指定给训练样本的，因此您事先不知道训练样本中个案的实际分布。 有关详细信息，请参阅第 7 页码验证。

各类别之间相等。如果因变量的类别在总体中是以相等方式表示的，则使用此设置。例如，如果有四个类别，则每个类别中的个案约为 25%。

自定义。对于网格中列出的每个因变量类别，输入一个非负值。这些值可以是比例、百分比、频率计数或表示各类别之间值分布的任何其他值。

使用误分类成本调整先验。如果定义自定义的误分类成本，则可以根据这些成本调整先验概率。 有关详细信息，请参阅第 15 页码误分类成本。

利润和值标签

此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定先验概率

- 在“决策树”主对话框中，选择带有两个或更多已定义值标签的分类（名义、有序）因变量。
- 选择 CRT 或 QUEST 作为生长法。
- 单击选项。
- 单击先验概率选项卡。

得分

图片 1-16
“选项”对话框，“得分”选项卡

对于具有有序因变量的 CHAID 和穷举 CHAID，可以将自定义得分指定给每个因变量类别。得分定义因变量类别的顺序以及它们之间的距离。您可以使用得分来增大或减小有序值之间的相对距离或更改值的顺序。

- **为每个类别使用序数秩。**指定给最低的因变量类别的得分是 1，为次最高类别指定的得分是 2，依此类推。这是默认值。
- **自定义。**对于网格中列出的每个因变量类别，输入一个数值型得分值。

示例

值标签	初始值	得分
Unskilled	1	1
Skilled manual	2	4
Clerical	3	4.5
Professional	4	7
Management	5	6

- 得分增大了 Unskilled 和 Skilled 之间的相对距离，而减小了 Skilled manual 和 Clerical 之间的相对距离。
- 得分交换了 Management 和 Professional 的顺序。

得分和值标签

此对话框要求因变量具有已定义的值标签。除非分类因变量至少有两个值定义了值标签，否则此对话框不可用。

指定得分

▶ 在“决策树”主对话框中，选择带有两个或更多已定义值标签的有序因变量。

▶ 选择 CHAID 或**穷举 CHAID** 作为生长法。

▶ 单击**选项**。

▶ 单击**得分**选项卡。

缺失值

图片 1-17
“选项”对话框，“缺失值”选项卡

“缺失值”选项卡控制名义值、用户缺失值、自变量（预测变量）值的处理。

- 用户缺失的有序和刻度自变量值的处理随生长法的不同而不同。
- 名义因变量的处理在“类别”对话框中指定。 有关详细信息，请参阅第 5 页码选择类别。
- 对于有序和刻度因变量，始终排除具有系统缺失或用户缺失的因变量值的个案。

视为缺失值。用户缺失值当作系统缺失值处理。系统缺失值的处理随生长法的不同而不同。

视为有效值。名义自变量的用户缺失值在树生长和分类中被视为普通值。

方法从属规则

如果一些（而不是所有）自变量值是系统缺失或用户缺失的：

- 对于 CHAID 和穷举 CHAID，系统缺失和用户缺失的自变量值作为单个组合类别包括在分析中。对于刻度和有序自变量，算法首先使用有效值生成类别，然后确定是将缺失类别与其最类似的（有效的）类别合并,还是将其作为单独的类别保留。
- 对于 CRT 和 QUEST，从树生长过程中排除具有缺失自变量值的个案，但如果方法中包括替代变量，则使用替代变量对其进行分类。如果将名义用户缺失值视为缺失，同样也按此方式对其进行处理。 有关详细信息，请参阅第 14 页码替代变量。

指定名义自变量用户缺失处理

▶ 在“决策树”主对话框中，选择至少一个名义自变量。

- ▶ 单击选项。
- ▶ 单击缺失值选项卡。

保存模型信息

图片 1-18
“保存”对话框

可以将模型中的信息另存为工作数据文件中的变量，也可以将整个模型以 XML (PMML) 格式保存到外部文件中。

保存变量

终端节点编号。为其指定每个个案的终端节点。该值是树节点编号。

预测值。模型所预测的因变量的分类（组）或值。

预测概率。与模型的预测关联的概率。为每个因变量类别保存一个变量。对刻度因变量不可用。

样本分配（训练/检验）。对于分割样本验证，此变量指示在训练或检验样本中是否使用了某个案。对于训练样本，值为 1；对于检验样本，值为 0。只在选择了分割样本验证时才可用。 有关详细信息，请参阅第 7 页码验证。

将树模型导出为 XML

可以以 XML (PMML) 格式保存整个树模型。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。

训练样本。将模型写入指定的文件。对于分割样本验证的树，这是训练样本的模型。

检验样本。将检验样本的模型写入指定文件。只在选择了分割样本验证时才可用。

输出

可用的输出选项取决于生长法、因变量的测量级别和其他设置。

树显示

图片 1-19
“输出”对话框，“树”选项卡

可以控制树的初始外观或完全取消树的显示。

树。默认情况下，树形图包括在“查看器”显示的输出中。取消选择（取消选中）此选项可以从输出中排除树形图。

显示。这些选项控制“查看器”中树形图的初始外观。也可以通过编辑生成的树修改所有这些属性。

- **方向。**可以自上而下（根节点在顶部）、从左向右或从右向左地显示树。
- **节点内容。**节点可以显示表、图表或这两者。对于分类因变量，表显示频率计数和百分比，而图表则是条形图。对于刻度因变量，表显示均值、标准差、个案数和预测值，而图表则是直方图。
- **标度。**默认情况下，大树会自动按比例缩小以适合页上的树。可以指定最大为 200% 的自定义缩放百分比。

- **自变量统计量。**对于 CHAID 和穷举 CHAID，统计量包括 F 值（对于刻度因变量）或卡方值（对于分类因变量）以及显著性值和自由度。对于 CRT，显示改进值。对于 QUEST，显示 F、显著性值和自由度（对于刻度和有序自变量）；对于名义自变量，显示卡方、显著性值和自由度。
- **节点定义。**节点定义显示在每个节点拆分中使用的自变量的值。

表格式树。树中每个节点的摘要信息，包括该节点的父节点编号、自变量统计量、自变量值，刻度因变量的均值和标准差，或者分类因变量的计数和百分比。

图片 1-20
表格式树(F)

节点		Bad		Good		总计		预测类别	父节点	主自变量				
		N	百分比	N	百分比	N	百分比			变量	Sig.	卡方	df	拆分值
0	1020	1020	41.4%	1444	58.6%	2464	100.0%	Good						
1	454	454	82.1%	99	17.9%	553	22.4%	Bad	0	Income level	.000	662.457	2	<= Low
2	476	476	42.0%	658	58.0%	1134	46.0%	Good	0	Income level	.000	662.457	2	(Low, Medium]
3	90	90	11.6%	687	88.4%	777	31.5%	Good	0	Income level	.000	662.457	2	> Medium
4	422	422	56.7%	322	43.3%	744	30.2%	Bad	2	Number of credit cards	.000	193.113	1	5 or more
5	54	54	13.8%	336	86.2%	390	15.8%	Good	2	Number of credit cards	.000	193.113	1	Less than 5
6	80	80	17.6%	375	82.4%	455	18.5%	Good	3	Number of credit cards	.000	38.587	1	5 or more
7	10	10	3.1%	312	96.9%	322	13.1%	Good	3	Number of credit cards	.000	38.587	1	Less than 5
8	211	211	80.8%	50	19.2%	261	10.6%	Bad	4	Age	.000	95.299	1	<= 28.0792
9	211	211	43.7%	272	56.3%	483	19.6%	Good	4	Age	.000	95.299	1	> 28.0792

统计量

图片 1-21
“输出”对话框，“统计量”选项卡

可用的统计量表取决于因变量的测量级别、生长法和其他设置。

模型

摘要。摘要包括所用的方法、模型中包括的变量以及已指定但未包括在模型中的变量。

图片 1-22
模型摘要表

指定	增长方法	CHAID
	因变量	Credit rating
	自变量	Age, Income level, Number of credit cards, Education, Car loans
	验证	无
	最大树深度	3
	父节点中的最小个案	400
	子节点中的最小个案	200
结果	自变量已包括	Income level, Number of credit cards, Age
	节点数	10
	终端节点数	6
	深度	3

风险。风险估计及其标准误。对树的预测准确性的测量。

- 对于分类因变量，风险估计是在为先验概率和误分类成本调整后不正确分类的个案的比例。
- 对于刻度因变量，风险估计是节点中的方差。

分类表。对于分类（名义、有序）因变量，此表显示每个因变量类别的正确分类和不正确分类的个案数。对刻度因变量不可用。

图片 1-23
风险和分类表

风险

估计	标准误差
.205	.008

增长方法: CHAID
因变量列表: Credit rating

分类

	预测		
观测	Bad	Good	百分比更正
Bad	665	355	65.2%
Good	149	1295	89.7%
整体百分比	33.0%	67.0%	79.5%

增长方法: CHAID
因变量列表: Credit rating

成本、先验概率、得分和利润值。对于分类因变量，此表显示在分析中使用的成本、先验概率、得分和利润值。对刻度因变量不可用。

自变量

对模型的重要性。对于 CRT 生长法，根据每个自变量（预测变量）对模型的重要性对其进行分类。对 QUEST 或 CHAID 方法不可用。

替代变量（按分割）。对于 CRT 和 QUEST 生长法，如果模型包括替代变量，则在树中列出每个分割的替代变量。对 CHAID 方法不可用。 有关详细信息，请参阅第 14 页码替代变量。

节点性能

摘要。对于刻度因变量，该表包括因变量的节点编号、个案数和均值。对于带有已定义利润的分类因变量，该表包括节点编号、个案数、平均利润和 ROI（投资回报）值。对不带已定义利润的分类因变量不可用。 有关详细信息，请参阅第 16 页码利润。

图片 1-24
节点的增益摘要表和百分位数

节点的收益汇总

节点	N	百分比	利润	投资回报率
6	712	35.2%	4.688	5.1%
8	455	21.4%	4.360	5.0%
4	86	3.6%	3.410	4.5%
10	495	19.8%	3.116	4.3%
9	249	7.3%	.188	.5%
3	467	12.7%	-.684	-2.5%

百分位数的收益汇总

百分点	多个节点	N	利润	投资回报率
10	23 ; 18	246	93.649	1837.2%
20	18 ; 21 ; 22	493	91.808	1755.3%
30	22 ; 17	739	90.317	1692.0%
40	17 ; 15 ; 20	986	87.090	1563.3%
50	20 ; 16 ; 19 ; 5	1232	82.712	1405.1%
60	5 ; 14	1478	78.061	1254.5%
70	14 ; 13	1725	73.533	1122.8%
80	13 ; 11 ; 12	1971	66.720	947.6%
90	12 ; 10	2218	60.567	809.2%
100	10	2464	54.846	694.5%

按目标类别。对于带有已定义目标类别的分类因变量，该表包括按节点或百分位组显示的百分比增益、响应百分比和指标百分比（提升）。将对每个目标类别生成一个单独的表。对于不带已定义目标类别的刻度因变量或分类因变量不可用。 有关详细信息，请参阅第 5 页码选择类别。

图片 1-25
节点的目标类别增益和百分位数

收益的收益

节点	节点		收益		响应	指数
	N	百分比	N	百分比		
6	712	35.2%	648	44.9%	95.5%	127.4%
8	455	21.4%	375	26.0%	90.9%	121.1%
4	86	3.6%	53	3.7%	77.3%	103.1%
10	495	19.8%	278	19.3%	73.1%	97.4%
9	249	7.3%	44	3.0%	31.3%	41.7%
3	467	12.7%	46	3.2%	18.8%	25.1%

百分位数的收益

百分点	多个节点	N	收益		响应	指数
			N	百分比		
10	23 ; 18	246	243	16.8%	98.6%	168.3%
20	18 ; 21 ; 22	493	477	33.0%	96.7%	165.0%
30	22 ; 17	739	704	48.7%	95.2%	162.4%
40	17 ; 15 ; 20	986	905	62.7%	91.8%	156.7%
50	20 ; 16 ; 19 ; 5	1232	1076	74.5%	87.3%	149.0%
60	5 ; 14	1478	1220	84.5%	82.5%	140.8%
70	14 ; 13	1725	1343	93.0%	77.9%	132.9%
80	13 ; 11 ; 12	1971	1396	96.7%	70.8%	120.9%
90	12 ; 10	2218	1430	99.1%	64.5%	110.1%
100	10	2464	1444	100.0%	58.6%	100.0%

行。节点性能表可以按终端节点、百分位数或这两者显示结果。如果选择这两者，则为每个目标类别生成两个表。百分位数表根据排序顺序显示每个百分位数的累计值。

百分位数增量。对于百分位数表，可以选择以下百分位数增量：1、2、5、10、20 或 25。

显示累积统计量。对于终端节点表，在具有累积结果的每个表中显示附加列。

图表

图片 1-26
“输出”对话框，“图”选项卡

可用的图表取决于因变量的测量级别、生长法和其他设置。

模型自变量重要性。按自变量（预测变量）显示的模型重要性的条形图。仅对 CRT 生长法可用。

节点性能

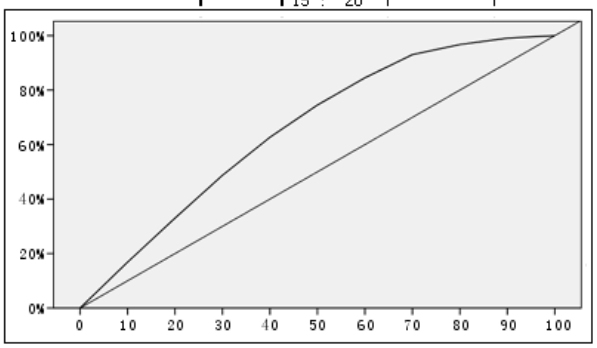
增益。增益是每个节点的目标类别中的总个案百分比，它的计算方法如下：（节点目标 n /总目标 n）x 100。增益图表是累积百分位数增益的折线图，它的计算方法如下：（累积百分位数目标 n /总目标 n）x 100。将为每个目标类别生成单独的折线图。只对带有已定义目标类别的分类因变量可用。 有关详细信息，请参阅第 5 页码选择类别。

增益图表绘制将在百分位数增益表（它还报告累计值）的增益百分比列中看到的相同值。

图片 1-27
百分位数增益表和增益图表

百分位数的收益

百分点	多个节点	N	收益 N	收益 百分比	响应	指数
10	23 ; 18	246	243	16.8%	98.6%	168.3%
20	18 ; 21 ; 22	493	477	33.0%	96.7%	165.0%
30	22 ; 17	739	704	48.7%	95.2%	162.4%
40	17 ; 15 ; 20	986	905	62.7%	91.8%	156.7%
50				74.5%	87.3%	149.0%
60				84.5%	82.5%	140.8%
70				93.0%	77.9%	132.9%
80				96.7%	70.8%	120.9%
90				99.1%	64.5%	110.1%
100				100.0%	58.6%	100.0%



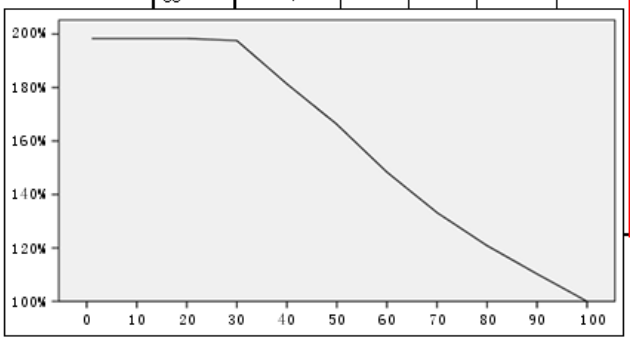
指标. 指标是目标类别的节点响应百分比与整个样本的总体目标类别响应百分比的比率。 指标图表是累积百分位数指标值的折线图。仅对分类因变量可用。累积百分位数指标的计算方法如下：（累积百分位数响应百分比/总响应百分比）x 100。将为每个目标类别生成单独的图表，且必须定义目标类别。

指标图表绘制将在百分位数增益表的指标列中看到的相同值。

图片 1-28
百分位数增益表和指标图表

百分位数的收益

百分点	多个节点	N	收益 N	收益 百分比	响应	指数
10	23 ; 18	246	243	16.8%	98.6%	168.3%
20	18 ; 21 ; 22	493	477	33.0%	96.7%	165.0%
30	22 ; 17	739	704	48.7%	95.2%	162.4%
40	17 ; 15 ; 20	986	905	62.7%	91.8%	156.7%
50	20 ;					149.0%
60						140.8%
70						132.9%
80						120.9%
90						110.1%
100						100.0%



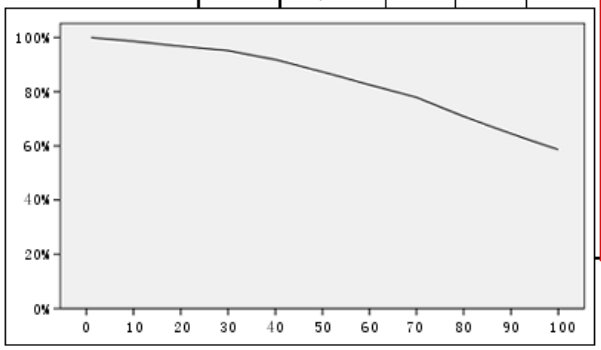
响应. 节点中的个案在指定目标类别中的百分比。 响应图表是累积百分位数响应的折线图，它的计算方法如下：（累积百分位数目标 n /累积百分位数合计 n）x 100。仅对带有已定义目标类别的分类因变量可用。

响应图表绘制将在百分位数增益表的响应列中看到的相同值。

图片 1-29
百分位数增益表和响应图表

百分位数的收益

百分点	多个节点	N	收益		响应	指数
			N	百分比		
10	23 ; 18	246	243	16.8%	98.6%	168.3%
20	18 ; 21 ; 22	493	477	33.0%	96.7%	165.0%
30	22 ; 17	739	704	48.7%	95.2%	162.4%
40	17 ; 15 ; 20	986	905	62.7%	91.8%	156.7%
					87.3%	149.0%
					82.5%	140.8%
					77.9%	132.9%
					70.8%	120.9%
					64.5%	110.1%
					58.6%	100.0%



均值。因变量的累积百分位数均值的折线图。仅对刻度因变量可用。

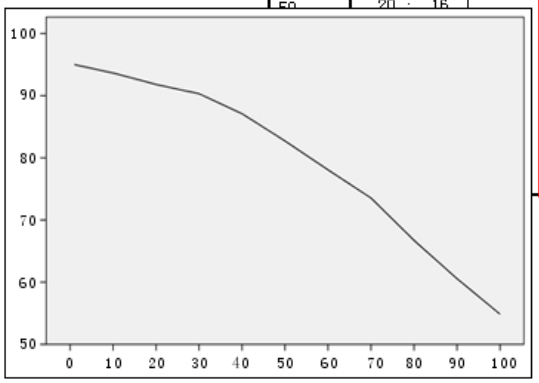
平均利润。累积平均利润的折线图。只对带有已定义利润的分类因变量可用。 有关详细信息，请参阅第 16 页码利润。

平均利润图表绘制将在百分位数增益摘要表的利润列中看到的相同值。

图片 1-30
百分位数增益摘要表和平均利润图表

百分位数的收益汇总

百分点	多个节点	N	利润	投资回报率
10	23 ; 18	246	4.903	5.2%
20	18 ; 21 ; 22	493	4.770	5.2%
30	22 ; 17	739	4.662	5.1%
40	17 ; 15 ; 20	986	4.429	5.0%
50	20 ; 16		4.113	4.9%
			3.778	4.7%
			3.451	4.5%
			2.959	4.2%
			2.515	3.8%
			2.102	3.5%



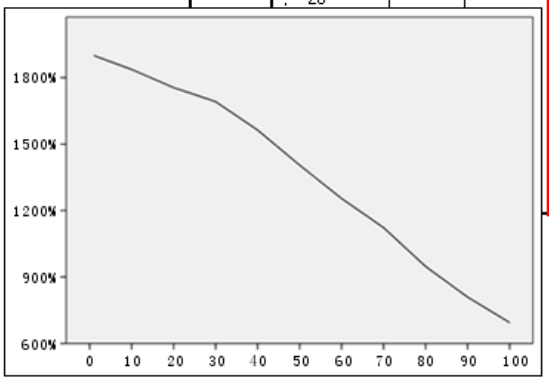
投资回报（ROI）。累积 ROI（投资回报）的折线图。ROI 计算为利润与支出之比。只对带有已定义利润的分类因变量可用。

ROI 图表绘制将在百分位数增益摘要表的 ROI 列中看到的相同值。

图片 1-31
百分位数增益摘要表和 ROI 图表

百分位数的收益汇总

百分点	多个节点	N	利润	投资回报率
10	23 ; 18	246	93.649	1837.2%
20	18 ; 21 ; 22	493	91.808	1755.3%
30	22 ; 17	739	90.317	1692.0%
40	17 ; 15 ; 20	986	87.090	1563.3%
				1405.1%
				1254.5%
				1122.8%
				947.6%
				809.2%
				694.5%



百分位数增量。对于所有的百分位数图表，此设置控制在图表上显示的百分位数增量：1、2、5、10、20 或 25。

选择规则和评分规则

图片 1-32
“输出”对话框，“规则”选项卡

使用“规则”选项卡，能够生成命令语法、SQL 或简单（纯英文）文本形式的选择或分类/预测规则。可以在“查看器”中显示这些规则并/或将其保存到外部文件。

语法。控制查看器中显示的输出和保存到外部文件的选择规则中的选择规则的形式。

- **IBM SPSS Statistics.** 命令语法语言。规则表示为一组定义过滤条件以用于选择个案子集的命令，或表示为可用于对个案评分的 COMPUTE 语句。
- **SQL。**生成标准的 SQL 规则，以便从数据库中选择或提取记录，或者将值指定给那些记录。生成的 SQL 规则不包含任何表名称或其他数据源信息。
- **简单文本。**纯英文的伪代码。规则表示为一组“if...then”逻辑语句，而这些语句描述了模型的分类或每个节点的预测。此形式的规则可以使用已定义变量和值标签或者变量名称和数据值。

类型。 对于 SPSS Statistics 和 SQL 规则，控制生成的规则类型：选择规则或评分规则。

- **为个案指定值。**此规则可用于为满足节点成员条件的个案指定模型的预测值。将为满足节点成员条件的每个节点生成单独的规则。
- **选择个案。**此规则可用于选择满足节点成员条件的个案。对于 SPSS Statistics 和 SQL 规则，将生成单个规则用于选择满足选择条件的所有个案。

在 SPSS Statistics 和 SQL 规则中包含替代变量。对于 CRT 和 QUEST，可以在规则中包含来自模型的替代预测变量。包含替代变量的规则可能非常复杂。一般来说，如果只想获得有关树的概念信息，请排除替代变量。如果某些个案有不完整的自变量（预测变量）数据并且您需要规则来模拟树，请包含替代变量。 有关详细信息，请参阅第 14 页码替代变量。

节点。控制已生成规则的范围。为范围中包含的每个节点生成单独的规则。

- **所有终端节点。**为每个终端节点生成规则。
- **最佳终端节点。**基于指标值为排在前面的 n 个终端节点生成规则。如果该数超过树中的终端节点数，则为所有终端节点生成规则。（请参见下面的注解。）
- **达到指定个案百分比的最佳终端节点。**基于指标值为排在前面的 n 个个案百分比的终端节点生成规则。（请参见下面的注解。）
- **其指标值达到或超过分界值的终端节点。**为指标值大于或等于指定值的所有终端节点生成规则。大于 100 的指标值表示，该节点中目标类别的个案百分比超过根节点中的百分比。（请参见下面的注解。）
- **所有节点。**为所有节点生成规则。

注 1：基于指标值的节点选，仅对带有已定义目标类别的分类因变量可用。如果已指定多个目标类别，则为每个目标类别生成一组单独的规则。

注 2：对于用于选择个案的 SPSS Statistics 和 SQL 规则（而不是用于指定值的规则），**所有节点**和**所有终端节点**将有效地生成选择在分析中使用的所有个案的规则。

将规则导出到文件。在外部文本文件中保存规则。

也可以基于最终树模型中的选定节点，以交互方式生成和保存选择规则或评分规则。有关详细信息，请参阅第 42 页码第 2 章中的个案选择和评分规则。

注意：如果将命令语法形式的规则应用到其他数据文件，该数据文件包含的变量必须与最终模型中的自变量同名，以相同的单位度量并且具有相同的用户定义的缺失值（如果存在）。

章
2

树编辑器

通过树编辑器，您可以：

- 隐藏和显示选定的树分支。
- 控制节点内容、在节点拆分中显示的统计量以及其他信息的显示。
- 更改节点、背景、边框、图表和字体颜色。
- 更改字体样式和大小。
- 更改树的对齐方式。
- 选择个案的子集以基于所选节点进一步进行分析。
- 创建和保存用于根据所选节点对个案进行选择或评分的规则。

编辑树模型：

▶ 在浏览器窗口中双击树模型。

或

▶ 从“编辑”菜单或右键单击上下文菜单中选择：
编辑内容 > 在单独窗口中

隐藏和显示节点

要隐藏（折叠）某个父节点下某分支中的所有子节点，请：

▶ 单击该父节点右下角下面的小框中的减号（-）。

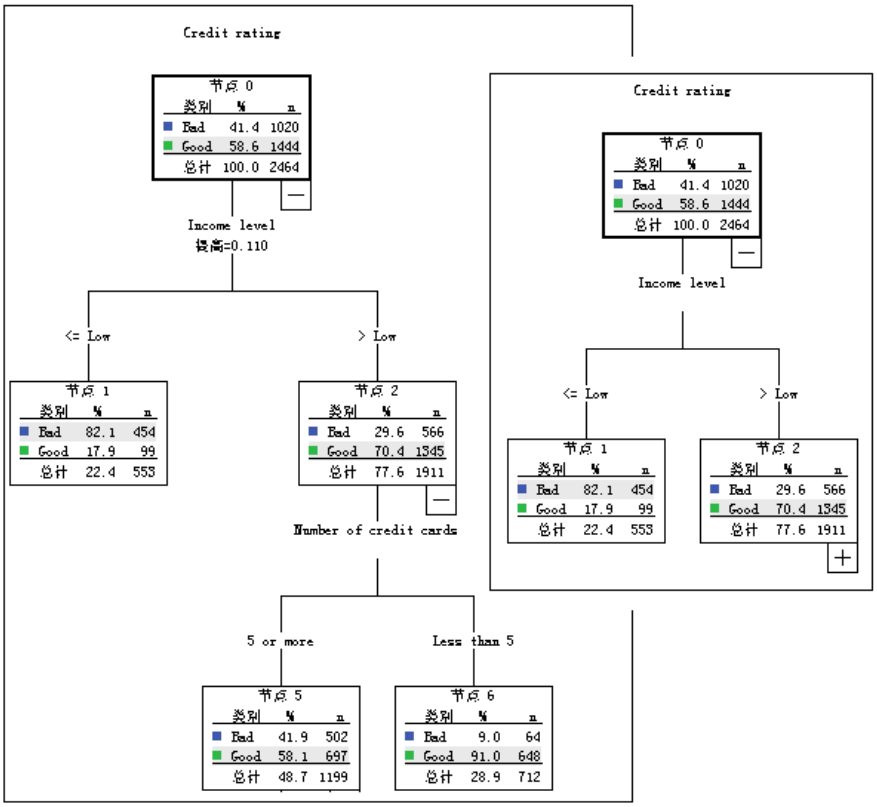
这样将隐藏该父节点下该分支上的所有节点。

要显示（展开）某个父节点下某分支中的子节点，请：

▶ 单击父节点右下角下面的小框中的加号（+）。

注意：隐藏分支上的子节点与修剪树不同。如果需要已修剪的树，则在创建树之前必须请求进行修剪，并且修剪的分支将不包含在最终树中。有关详细信息，请参阅第 13 页码第 1 章中的修剪树。

图片 2-1
展开和折叠的树



选择多个节点

您可以选择个案，生成评分和选择规则，以及基于当前选定的节点执行其他操作。要选择多个节点，请：

- 单击要选择的节点。
- 按住 Ctrl 并单击要选择的其他节点。

您可以选择一个分支中的多个同级节点和/或父节点以及其他分支中的子节点。但是，不能同时选择父节点以及同一节点分支的子节点/后代节点。

使用大型树

树模型有时可能包含过多节点和分支，导致难以乃至无法查看完整尺寸的整个树。如果使用大型树，您可能会发现许多有用的功能：

- **树地图。**您可以使用树地图（树的小型简化版本）来浏览树并选择节点。 有关详细信息，请参阅第 37 页码树状图。

- **尺度。**您可以通过更改树显示的刻度百分比进行放大和缩小。 有关详细信息，请参阅第 37 页码缩放树显示。
- **节点和分支显示。**通过在节点中仅显示表或仅显示图表和/或取消显示节点标签或自变量信息，可以使树更紧凑。 有关详细信息，请参阅第 39 页码控制树中显示的信息。

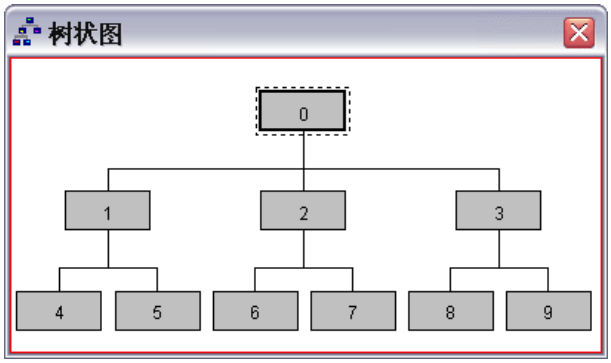
树状图

树地图提供了一个紧凑而简化的树视图，您可以用其浏览树和选择节点。

要使用树地图窗口，请：

▶ 从“树编辑器”菜单中选择：
视图 > 树状图

图片 2-2
树地图窗口



- 当前选定的节点在“树模型编辑器”和树地图窗口中将突出显示。
- 当前位于“树模型编辑器”视图区域中的树的部分在树地图中用红色矩形来指示。右键单击该矩形并拖动其以更改在视图区域中显示的树的部分。
- 如果在树地图中选择当前未位于“树编辑器”视图区域中的节点，则该视图将转为包括选定的节点。
- 在树地图中选择多个节点与在“树编辑器”中的操作方式相同：按住 Ctrl 并单击以选择多个节点。您不能同时选择父节点以及同一节点分支的子节点/后代节点。

缩放树显示

缺省情况下，树将自动缩放为适合“浏览器”窗口，这样会导致某些树在初始状态下十分难以读取。您可以选择预设的刻度设置或输入介于 5% 与 200% 之间的定制刻度值。

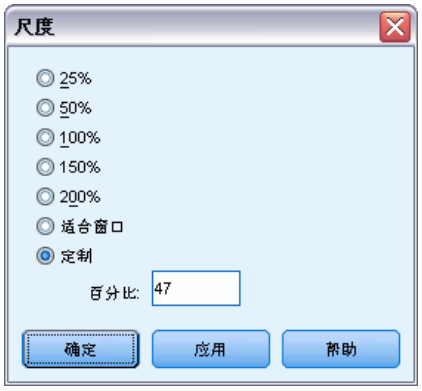
更改树的刻度：

▶ 从工具栏上的下拉列表中选择刻度百分比，或输入定制的百分比值。

或

▶ 从“树编辑器”菜单中选择：
视图 > 尺度...

图片 2-3
“刻度”对话框



也可以在创建树模型之前指定刻度值。 有关详细信息，请参阅第 22 页码第 1 章中的输出。

节点摘要窗口

节点摘要窗口提供了所选节点的较大视图。也可以使用摘要窗口来查看、应用或保存基于所选节点的选择或评分规则。

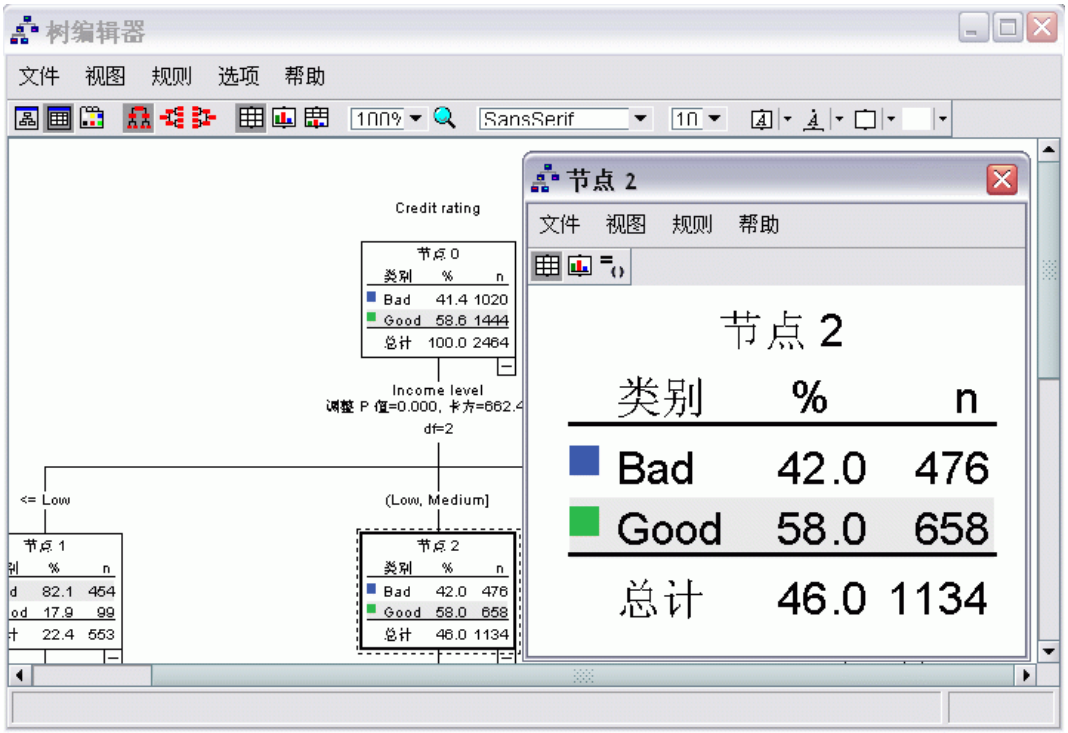
- 使用节点摘要窗口中的“查看”菜单可以在摘要表、图表或规则的视图之间进行切换。
- 使用节点摘要窗口中的“规则”菜单可以选择要查看的规则类型。 有关详细信息，请参阅第 42 页码个案选择和评分规则。
- 节点摘要窗口中的所有视图反映了所有选定节点的组合摘要。

要使用节点摘要窗口，请：

▶ 在“树编辑器”中选择节点。要选择多个节点，请按住 Ctrl 并单击。

▶ 从菜单中选择：
视图 > 摘要

图片 2-4
摘要窗口

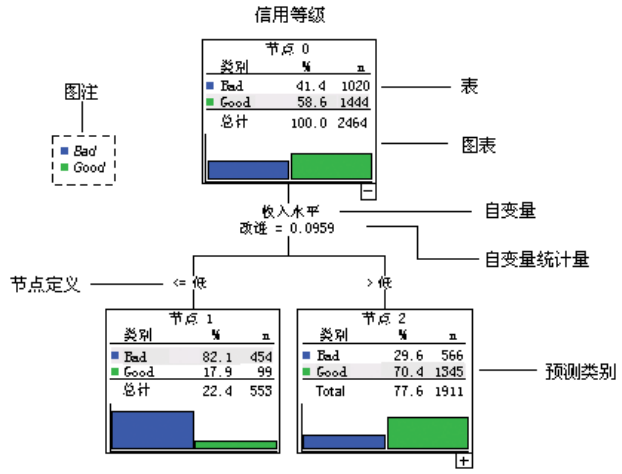


控制树中显示的信息

使用树编辑器中的“选项”菜单，您可以控制节点内容、自变量（预测变量）名称和统计量、节点定义和其他设置的显示。这些设置中的大部分还可以从工具栏中进行控制。

设置	“选项”菜单选择
突出显示预测类别（分类因变量）	突出显示预测值
节点中的表和/或图表	节点内容
显著性检验值和 p 值	自变量统计量
自变量（预测变量）名称	自变量
节点的自变量（预测变量）值	节点定义
对齐（从上到下、从左到右、从右到左）	方向
图表图注	图例

图片 2-5
树元素



更改树颜色和文本字体

您可以更改树中的以下颜色：

- 节点边框、背景和文本颜色
- 分支颜色和分支文本颜色
- 树背景颜色
- 预测类别突出显示颜色（分类因变量）
- 节点图表颜色

还可以更改树中所有文本的类型字体、样式和大小。

注意：不能更改单个节点或分支的颜色或字体属性。颜色更改应用于同一类型的所有元素，字体更改（不同于颜色）将应用于所有图表元素。

更改颜色和文本字体属性：

▶ 使用工具栏来更改整个树的字体属性或更改各个树元素的颜色。（将鼠标光标置于工具栏上的每个控件上时，工具提示将描述该控件。）

或

▶ 在“树编辑器”中的任意位置双击可打开“属性”窗口，或从菜单中选择：
视图 > 属性

▶ 对于边框、分支、节点背景、预测类别和树背景，请单击**颜色**选项卡。

▶ 对于字体颜色和属性，请单击**文本**选项卡。

▶ 对于节点图表颜色，请单击**节点图表**选项卡。

图片 2-6
“属性”窗口，“颜色”选项卡

图片 2-7
“属性”窗口，“文本”选项卡

图片 2-8
“属性”窗口，“节点图表”选项卡

个案选择和评分规则

您可以使用树编辑器来执行以下操作：

- 基于所选节点来选择个案子集。 有关详细信息，请参阅第 42 页码过滤个案。
- 以 IBM® SPSS® Statistics 命令语法或 SQL 格式生成个案选择规则或评分规则。 有关详细信息，请参阅第 43 页码保存选择和评分规则。

当运行“决策树”过程以创建树模型时，还可以基于各种条件自动保存规则。 有关详细信息，请参阅第 33 页码第 1 章中的选择规则和评分规则。

过滤个案

如果希望了解有关特定节点或节点组中个案的详细信息，您可以选择个案的子集以基于所选节点进一步进行分析。

▶ 在树编辑器中选择节点。要选择多个节点，请按住 Ctrl 并单击。

▶ 从菜单中选择：
规则 > **过滤个案...**

▶ 输入一个过滤变量名称。所选节点中的个案将接收此变量值 1。其他所有个案将接收值 0，这些个案将从后续分析中排除，直至您更改过滤状态。

▶ 单击**确定**。

图片 2-9
“过滤个案”对话框

保存选择和评分规则

您可以将个案选择或评分规则保存在外部文件中，然后将这些规则应用于其他数据源。这些规则基于“树编辑器”中的所选节点。

语法。控制浏览器中显示的输出和保存到外部文件的选择规则中的选择规则的形式。

- **IBM SPSS Statistics.** 命令语法语言。规则表示为一组定义过滤条件以用于选择个案子集的命令，或表示为可用于对个案评分的 COMPUTE 语句。
- **SQL。**生成标准 SQL 规则以从数据库中选择/提取记录或为这些记录指定值。生成的 SQL 规则不包含任何表名称或其他数据源信息。

类型。您可以创建选择或评分规则。

- **选择个案。**此规则可用于选择满足节点成员条件的个案。对于 SPSS Statistics 和 SQL 规则，将生成单个规则用于选择满足选择条件的所有个案。
- **为个案指定值。**此规则可用于为满足节点成员条件的个案指定模型的预测值。将为满足节点成员条件的每个节点生成单独的规则。

包括替代变量。 对于 CRT 和 QUEST，可以在规则中包含来自模型的替代预测变量。包含替代变量的规则可能非常复杂。一般来说，如果只想获得有关树的概念信息，请排除替代变量。如果某些个案有不完整的自变量（预测变量）数据并且您需要规则来模拟树，请包含替代变量。 有关详细信息，请参阅第 14 页码第 1 章中的替代变量。

保存个案选择或评分规则：

▶ 在树编辑器中选择节点。要选择多个节点，请按住 Ctrl 并单击。

▶ 从菜单中选择：
规则 > 导出...

▶ 选择所需的规则类型并输入文件名。

图片 2-10
“导出规则”对话框

注意：如果将命令语法形式的规则应用到其他数据文件，该数据文件包含的变量必须与最终模型中的自变量同名，以相同的单位度量并且具有相同的用户定义的缺失值（如果存在）。

部分 II:
示例

章
3

数据假设和要求

“决策树”过程假设：

- 已为所有分析变量指定相应的测量级别。
- 对于分类（**名义**、**有序**）因变量，已为应包括在分析中的所有类别定义值标签。

我们将使用文件 tree_textdata.sav 来说明这两种要求的重要性。此数据文件反映了定义任何属性，例如测量级别或值标签之前，读入或输入的数据的默认状态。 有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Decision Trees 20。

测量级别对树模型的影响

此数据文件中的两个变量均为数值变量，并且均已指定**刻度**测量级别。但是（我们在后面将看到）这两个变量实际上是依赖于数值代码表示类别值的分类变量。

▶ 要运行决策树分析，请从菜单中选择：
分析 > **分类** > **树...**

源变量列表中两个变量旁边的图标表示它们将被视为刻度变量。

图片 3-1
具有两个刻度变量的“决策树”主对话框

▶ 选择 dependent 作为因变量。

- 选择 independent 作为自变量。
- 单击**确定**以运行该过程。
- 再次打开“决策树”对话框，并单击**重置**。
- 右键单击源列表中的 dependent，然后从上下文菜单中选择**名义**。
- 对源列表中的变量 independent 执行相同的操作。

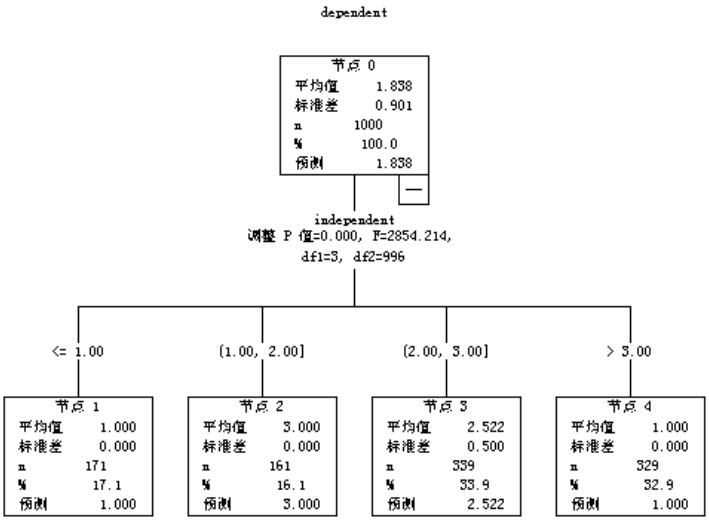
现在每个变量旁边的图标均表示它们将被视为名义变量。

图片 3-2
源列表中的名义图标

- 选择 dependent 作为因变量并选择 independent 作为自变量，然后单击**确定**再次运行该过程。

现在，我们来比较一下这两个树。首先，我们查看其中两个数值变量均被视为刻度变量的树。

图片 3-3
将两个变量均视为刻度变量的树

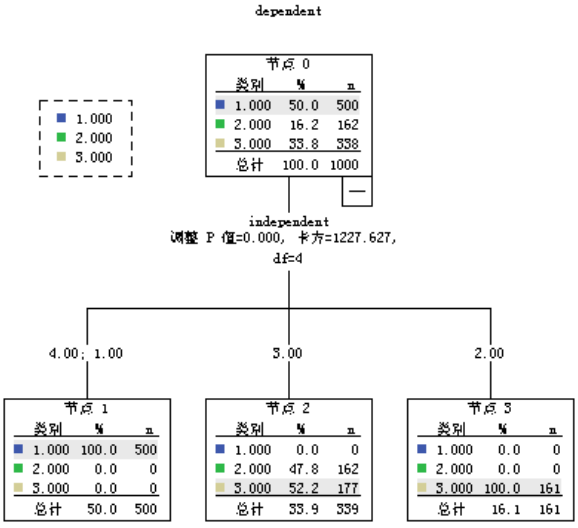


- 树的每个节点显示“预测”值，该值是该节点处因变量的均值。对于实际为分类变量的变量，均值统计方法可能并不具有实际意义。
- 该树具有四个子节点，每个节点代表自变量的一个值。

树模型经常合并相似节点，但对于刻度变量来说，只能合并连续值。本示例中，不存在视为相似的连续值，因此不能合并任何节点。

其中两个变量被视为名义变量的树在许多方面都有所不同。

图片 3-4
将两个变量都视为名义变量的树



- 不再显示预测值，每个节点均包含一个频率表，显示每个类别因变量的个案数（计数和百分比）。

- “预测”类别（各个节点中具有最高计数的类别）将突出显示。例如，节点 2 的预测类别为类别 3。
- 仅存在三个子节点，而不是四个子节点，其中自变量的两个值合并到一个节点中。

合并到同一节点的两个自变量值为 1 和 4。按照定义，由于名义值没有任何固有顺序，因此允许合并不连续值。

永久指定测量级别

在“决策树”对话框中更改变量的测量级别时，所做更改只是临时性更改，不会与数据文件一起保存。此外，您可能并非总是能够知道所有变量应具有的正确测量级别。

“定义变量属性”可以帮助您确定每个变量的正确测量级别，并永久更改指定的测量级别。使用“定义变量属性”：

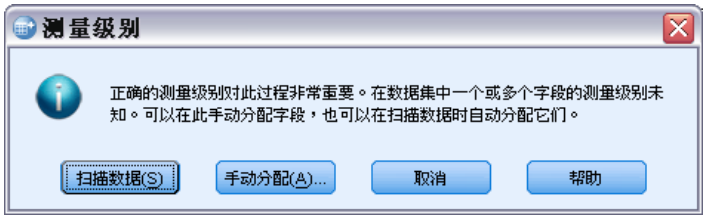
▶ 从菜单中选择：

数据 > 定义变量属性...

具有未知测量级别的变量

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。

图片 3-5
测量级别警报



- **扫描数据。** 读取活动数据集中的数据，并分配默认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。

值标签对树模型的影响

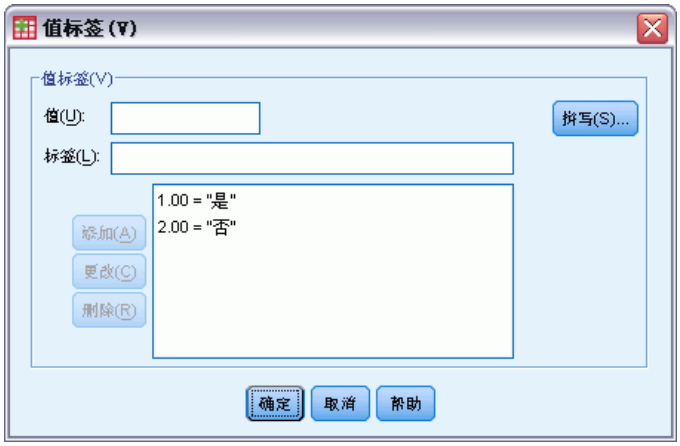
“决策树”对话框界面假设分类（名义、有序）因变量的所有非缺失值均已定义值标签或未定义值标签。 除非分类因变量至少有两个非缺失值具有值标签，否则某些功能将不可用。如果至少两个非缺失值已经定义了值标签，则将从分析中排除带有其他没有值标签的值的所有个案。

本示例中的原始数据文件不包含定义的值标签，如果将因变量视为名义变量，则树模型将在分析中使用所有非缺失值。本示例中，这些值为 1、2 和 3。

但是当我们定义某些（而非全部）因变量值的值标签时，会发生什么情况？

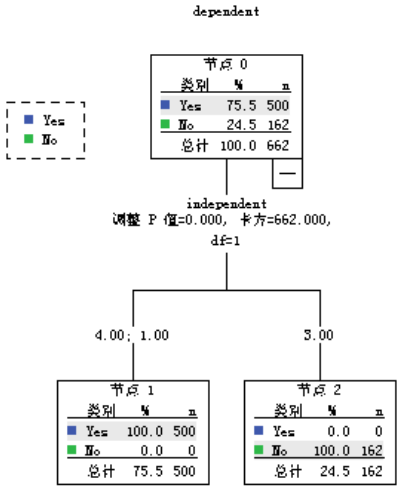
- 在“数据编辑器”窗口中，单击**变量视图**选项卡。
- 单击变量 dependent 的**值**单元。

图片 3-6
为 dependent 变量定义值标签



- 首先，在“值”中输入 1，在“值标签”中输入 Yes，然后单击**添加**。
- 接下来，在“值”中输入 2，在“值标签”中输入 No，然后再次单击**添加**。
- 随后，单击**确定**。
- 再次打开“决策树”对话框。该对话框应仍将 dependent 选择为因变量，并具有名义测量级别。
- 单击**确定**再次运行该过程。

图片 3-7
具有偏值标签的名义因变量的树



现在，树模型中只包含两个具有所定义值标签的因变量值。因变量具有值 3 的所有个案均已排除，如果您不熟悉该数据，则这种情况可能不会很明显。

为所有值指定值标签

要避免从分析中意外遗漏有效的分类值，请使用“定义变量属性”为数据中发现的所有因变量值指定值标签。

如果变量 name 的数据字典信息显示在“定义变量属性”对话框中，则可看到尽管该变量的值 3 具有 300 多个个案，但并未为该值定义任何值标签。

图片 3-8
“定义变量属性”对话框中带有偏值标签的变量

章 4

使用决策树评估信用风险

银行维护一个有关已从银行获取贷款的客户的历史信息数据库，其中包含这些客户是否偿还贷款或拖欠贷款。使用数模型，您可以分析两组客户的特征，并构建模型以预测贷款申请者拖欠其贷款的可能性。

信用数据存储在 tree_credit.sav 中。 有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Decision Trees 20。

创建模型

“决策树”过程提供了多种用于创建树模型的不同方法。对于本示例，我们将使用缺省方法：

CHAID. 卡方自动交互检测。在每一步，CHAID 选择与因变量有最强交互作用的自变量（预测变量）。如果每个预测变量的类别与因变量并非显著不同，则合并这些类别。

构建 CHAID 树模型

▶ 要运行决策树分析，请从菜单中选择：
分析 > **分类** > **树...**

图片 4-1
“决策树”对话框

- 选择 Credit rating 作为因变量。
- 选择所有剩余变量作为自变量。（该过程将自动排除任何对最终模型无明显作用的变量。）

  此时，您可以运行该过程并生成基本树模型，但我们将选择某些附加输出，并对用于生成模型的标准稍做调整。

选择目标类别

- 单击所选因变量右下方的**类别**按钮。

这将打开“类别”对话框，您可以在其中指定所需的因变量目标类别。目标类别不影响树模型本身，但某些输出和选项仅在已选择目标类别后才可用。

图片 4-2
“类别”对话框

- 选择（选中）Bad 类别的目标复选框。具有不良信用等级（拖欠贷款）的客户将被视为感兴趣的目标类别。
- 单击**继续**。

指定树生长条件

对于本示例，我们需要保持非常简单的树结构；因此，我们将通过提高父节点和子节点的最小个案数来限制树生长。

- 在“决策树”主对话框中，单击**条件**。

图片 4-3
“条件”对话框，“增长限制”选项卡

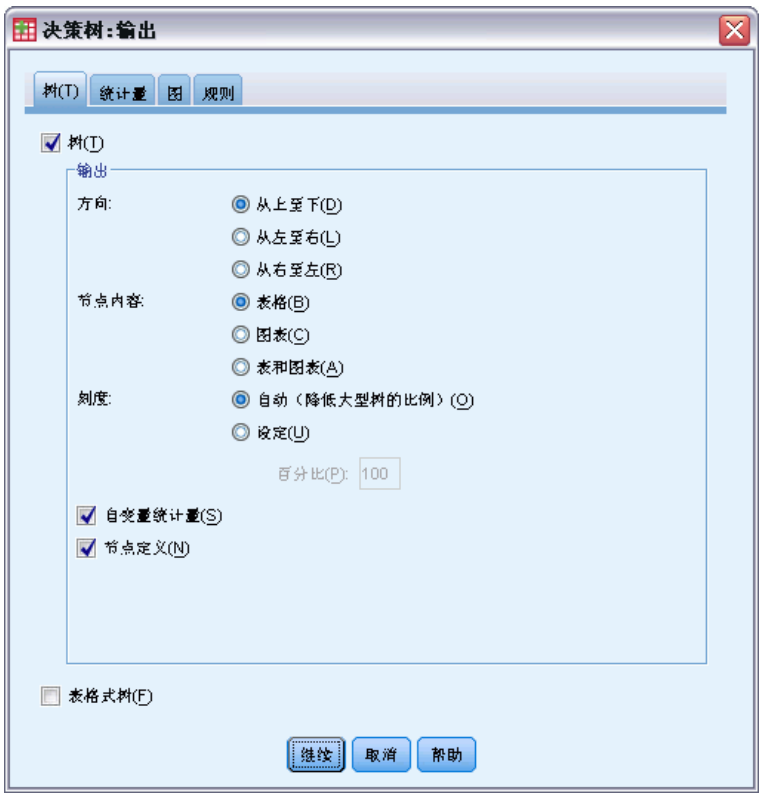
- 在“最小个案数”组中，为父节点键入 400，并为子节点键入 200。
- 单击**继续**。

选择附加输出

- 在“决策树”主对话框中，单击**输出**。

这将打开一个标签式对话框，您可以在其中选择各种类型的附加输出。

图片 4-4
“输出”对话框，“树”选项卡



- 在“树”选项卡上，选择（选中）**表格式树**。
- 然后单击**图**选项卡。

图片 4-5
“输出”对话框，“图”选项卡

▶ 选择（选中）**增益**和**索引**。

注意：这些图表需要因变量的目标类别。在此示例中，除非已指定一个或多个目标类别，否则将无法访问“图”选项卡。

▶ 单击**继续**。

保存预测值

可以保存包含有关模型预测值信息的变量。例如，可以保存为每个个案预测的信用等级，然后将这些预测值与实际的信用等级进行比较。

▶ 在“决策树”主对话框中，单击**保存**。

图片 4-6
保存对话框

- 选择（选中）**终端节点编号**、**预测值**和**预测概率**。
- 单击**继续**。
- 在“决策树”主对话框中，单击**确定**以运行该过程。

评估模型

对于本示例，模型结果包括：

- 提供有关模型信息的表。
- 树形图。
- 提供模型性能指示的图表。
- 添加到活动数据集的模型预测变量。

模型摘要表

图片 4-7
模型汇总

指定	增长方法	CHAID
	因变量	Credit rating
	自变量	Age, Income level, Number of credit cards, Education, Car loans
	验证	无
	最大树深度	3
	父节点中的最小个案	400
	子节点中的最小个案	200
结果	自变量已包括	Income level, Number of credit cards, Age
	节点数	10
	终端节点数	6
	深度	3

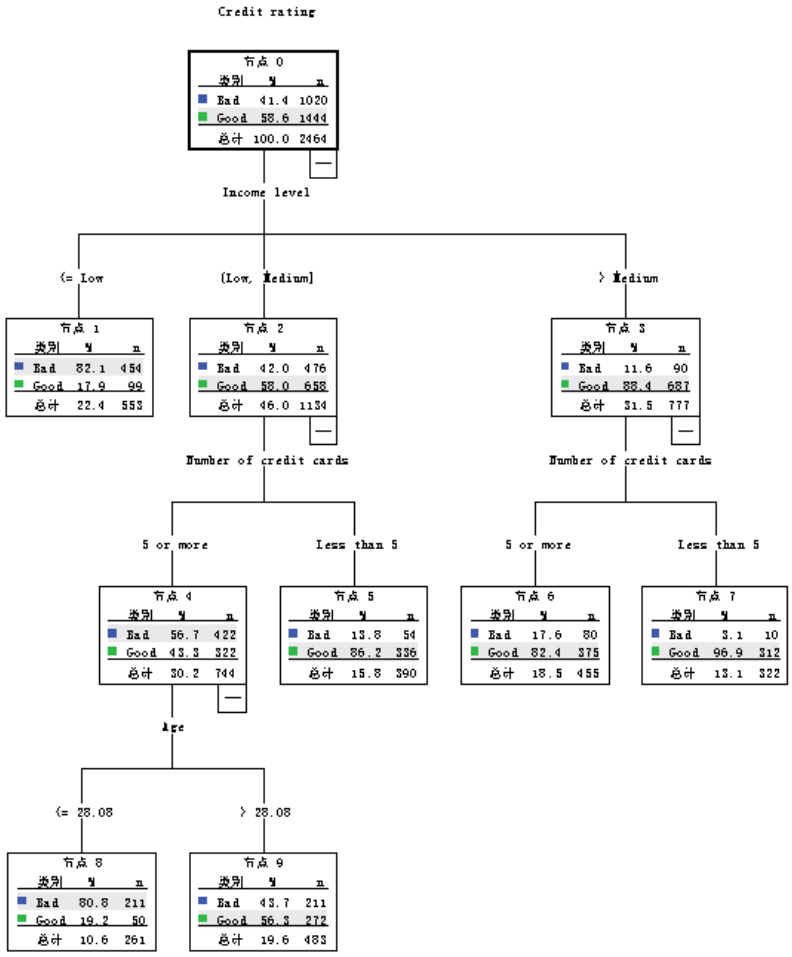
模型摘要表提供了有关用于构建模型的规范以及生成的模型的一些非常广泛的信息。

- “指定”部分提供有关用于生成树模型的设置的信息，包括在分析中使用的变量。
- “结果”部分显示有关最终模型中的总计数和终端节点数、树深度（根节点下的级别数），以及所包含的自变量的信息。

指定了五个自变量，但最终模型中只包含三个自变量。education 的变量和当前 car loans 的数目对模型无明显作用，因此自动将其从最终模型中删除。

树形图

图片 4-8
credit rating 模型的树形图



树形图是树模型的图形表示形式。此树形图显示：

- 使用 CHAID 方法，income level 是 credit rating 的最佳预测变量。
- 对于低收入类别，income level 是对 credit rating 唯一起作用的预测变量。此类别中的银行客户有 82% 拖欠了贷款。由于其下没有任何子节点，因而此节点被视为**终端**节点。
- 对于中等和高收入类别，另一个最佳预测变量是 number of credit cards。
- 对于拥有五个或更多信用卡的中等收入客户，该模型包含另一个预测变量：age。28 岁或以下的客户中超过 80% 都具有不良的信用等级，而 28 岁以上的客户中仅有不到半数具有不良的信用等级。

您可以使用树编辑器来隐藏和显示选定的分支、更改颜色和字体，以及选择基于所选节点的个案子集。 有关详细信息，请参阅第 67 页码选择节点中的个案。

树表

图片 4-9
信用等级的树表

	Bad		Good		总计			
节点	N	百分比	N	百分比	N	百分比	预测类别	父节点
0	1020	41.4%	1444	58.6%	2464	100.0%	Good	
1	454	82.1%	99	17.9%	553	22.4%	Bad	0
2	476	42.0%	658	58.0%	1134	46.0%	Good	0
3	90	11.6%	687	88.4%	777	31.5%	Good	0
4	422	56.7%	322	43.3%	744	30.2%	Bad	2
5	54	13.8%	336	86.2%	390	15.8%	Good	2
6	80	17.6%	375	82.4%	455	18.5%	Good	3
7	10	3.1%	312	96.9%	322	13.1%	Good	3
8	211	80.8%	50	19.2%	261	10.6%	Bad	4
9	211	43.7%	272	56.3%	483	19.6%	Good	4

树表，顾名思义，即以表格形式提供了大多数基本的树形图信息。对于每个节点，该表显示：

- 每个因变量类别中的个案数及百分比。
- 因变量的预测类别。在此示例中，预测类别是在该节点中具有 50% 以上个案的 credit rating 类别，因为只有两种可能的信用等级。
- 树中每个节点的父节点。请注意，节点 1（低收入水平节点）不是任何节点的父节点。由于该节点为终端节点，因此没有子节点。

图片 4-10
信用等级的树表（续）

主自变量				
变量	Sig.	卡方	df	拆分值
Income level	.000	662.457	2	<= Low
Income level	.000	662.457	2	(Low, Medium]
Income level	.000	662.457	2	> Medium
Number of credit cards	.000	193.113	1	5 or more
Number of credit cards	.000	193.113	1	Less than 5
Number of credit cards	.000	38.587	1	5 or more
Number of credit cards	.000	38.587	1	Less than 5
Age	.000	95.299	1	<= 28.0792
Age	.000	95.299	1	> 28.0792

- 用于分割节点的自变量。

- 卡方值（因为该树是使用 CHAID 方法生成的）、自由度 (df) 以及分割的显著性水平 (Sig.)。对于多数实际用途，您可能仅对显著性水平（对此模型中的所有分割，均低于 0.0001）感兴趣。
- 该节点的自变量的值。

注意：对于有序和刻度自变量，您可能看到树中的范围和以一般格式 (value1, value2] 表示的树表，该格式基本上表示“大于 value1 且小于或等于 value2”。 在此示例中，income level 仅有三个可能值：Low、Medium和High，并且 (Low, Medium] 只表示 Medium。类似地，>Medium 表示 High。

节点的增益

图片 4-11
节点的收益

	节点		收益			
节点	N	百分比	N	百分比	响应	指数
1	553	22.4%	454	44.5%	82.1%	198.3%
8	261	10.6%	211	20.7%	80.8%	195.3%
9	483	19.6%	211	20.7%	43.7%	105.5%
6	455	18.5%	80	7.8%	17.6%	42.5%
5	390	15.8%	54	5.3%	13.8%	33.4%
7	322	13.1%	10	1.0%	3.1%	7.5%

增长方法: CHAID
因变量列表: Credit rating

节点收益表提供了有关模型中终端节点的信息摘要。

- 在此表中仅列出终端节点,即树停止生长处的节点。通常，您只会对终端节点感兴趣，因为它们代表模型的最佳分类预测。
- 由于收益值提供了有关目标类别的信息，因而此表仅在指定了一个或多个目标类别时才可用。在此示例中，只有一个目标类别，因此只有一个节点收益表。
- Node N 是每个终端节点中的个案数，Node Percent 是每个节点中个案总数的百分比。
- Gain N 是目标类别的每个终端节点中的个案数，Gain Percent 是目标类别中的个案数相对于目标类别中的整体个案数的百分比。在此示例中，个案数及百分比具有不良的信用等级。
- 对于分类因变量，响应是所指定目标类别的节点中的个案百分比。在此示例中，这些是在树形图中为 Bad 类别显示的相同百分比。
- 对于分类因变量，指标是目标类别的响应百分比与整个样本的响应百分比的比率。

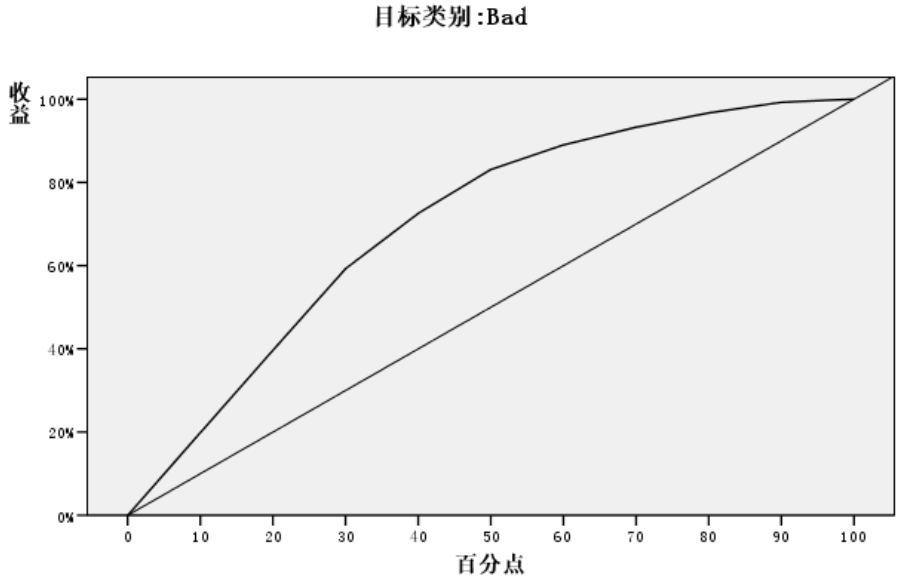
索引值

指标值基本上指示了该节点的观察目标类别百分比与目标类别的期望百分比的不同程度。 根节点中的目标类别百分比表示在考虑任何自变量效果之前的期望百分比。

大于 100% 的指标值表示目标类别中的个案数多于目标类别中的整体百分比。 相反，小于 100% 的指标值表示目标类别中的个案数少于整体百分比。

收益图表

图片 4-12
不良信用等级目标类别的收益图表

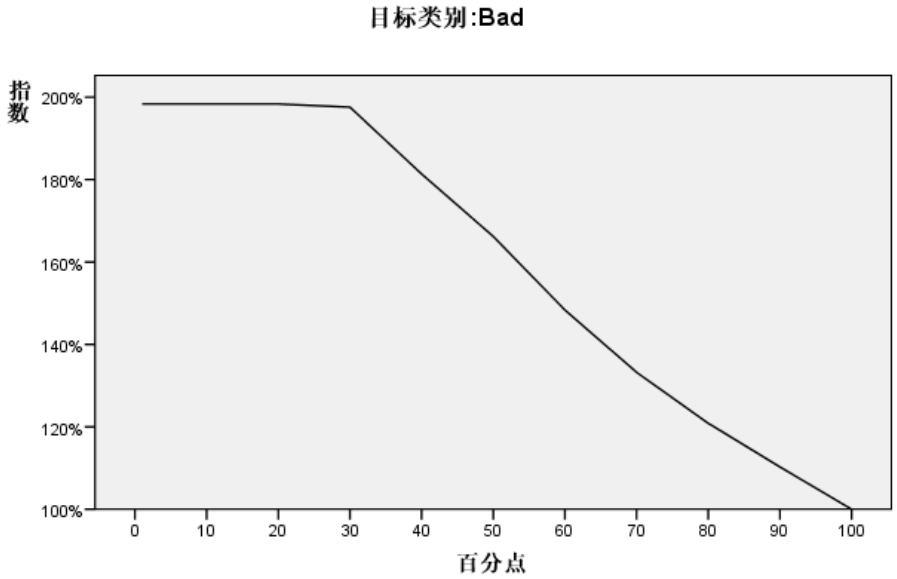


此收益图表显示该模型十分良好。

从一端转到另一端时，累积收益图表将始终以 0% 开始且以 100% 结束。对于良好的模型，收益图表向 100% 突增，然后趋于平稳。未提供任何信息的模型将沿着对角参考线。

指数图表

图片 4-13
不良信用等级目标类别的指数图表



此指数图表也指示该模型十分良好。累积指数图表趋向于从 100% 以上开始，然后逐渐下降到 100%。

对于良好的模型，指数值应正好从高于 100% 开始，在移动过程中保持较高的稳定水平，然后骤降至 100%。对于未提供任何信息的模型，整个图表的线将保持在 100% 左右。

风险估计和分类

图片 4-14
风险和分类表

风险

估计	标准误差
.205	.008

增长方法: CHAID
因变量列表: Credit rating

分类

	预测		
观测	Bad	Good	百分比更正
Bad	665	355	65.2%
Good	149	1295	89.7%
整体百分比	33.0%	67.0%	79.5%

增长方法: CHAID
因变量列表: Credit rating

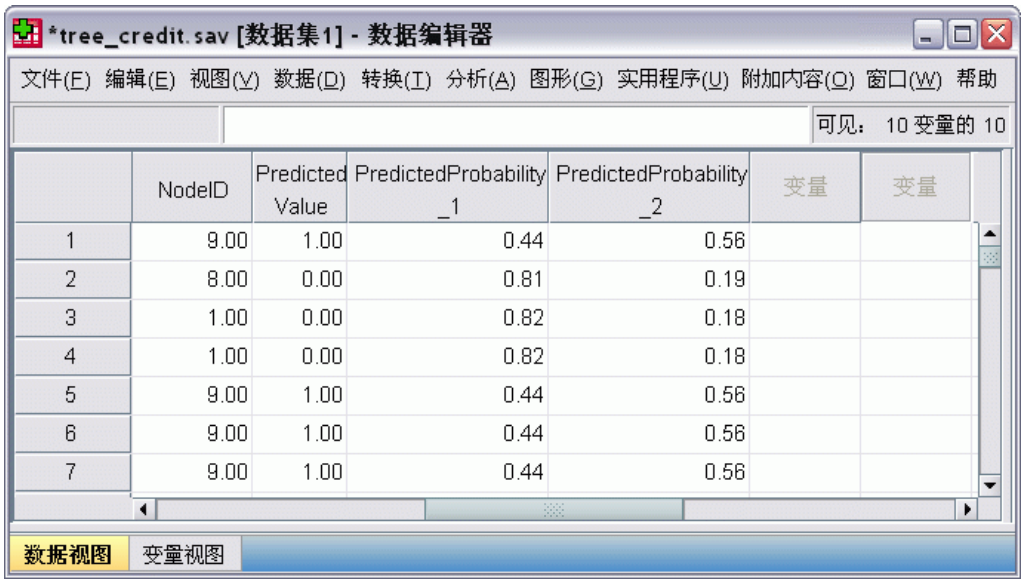
风险和分类表提供了模型运行状况的快速评估。

- 0.205 的风险估计值表明该模型（良好或不良信用等级）所预测类别的个案错误率为 20.5%。因此对客户进行误分类的“风险”约为 21%。
- 分类表中的结果与风险估计一致。该表显示模型对约 79.5% 的客户进行了正确分类。

但是，分类表揭示了此模型的一个潜在问题：对于具有不良信用等级的客户，此模型仅为其中的 65% 预测了不良等级，这意味着在具有不良信用等级的客户中有 35% 被错误地分类为“良好”客户。

预测值

图片 4-15
预测值和概率的新变量



活动数据集中已创建四个新变量：

NodeID。每个个案的终端节点数。

PredictedValue。每个个案的因变量预测值。由于因变量编码为 0 = Bad 1 = Good，因此预测值 0 表示将该个案预测为具有不良信用等级。

PredictedProbability。个案属于每个因变量类别的概率。由于因变量仅有两个可能值，因此创建了两个变量：

- **PredictedProbability_1。**个案属于不良信用等级分类的概率。
- **PredictedProbability_2。**个案属于良好信用等级分类的概率。

对于包含每个个案的终端节点，预测概率只是每个因变量类别中的个案比例。例如，在节点 1 中，82% 的个案属于不良类别，18% 属于良好类别，从而分别导致 0.82 和 0.18 的预测概率。

对于分类因变量，预测值是在每个个案的终端节点中具有最高个案比例的类别。例如，对于第一个个案，预测值为 1（良好信用等级），因为其终端节点中约 56% 的个案具有良好信用等级。 相反，对于第二个个案，预测值为 0（不良信用等级），因为其终端节点中约 81% 的个案具有不良信用等级。

但是，如果已定义成本，则预测类别与预测概率之间的关系可能不那么简单。 有关详细信息，请参阅第 70 页码为结果分配成本。

改进模型

整体而言，模型仅具有 80% 以下的正确分类比率。这是在多数终端节点中反映出来的，其中预测类别（在节点中突出显示的类别）与 80% 或更多个案的实际类别相同。

但是，存在一个终端节点，其中在良好与不良信用等级之间几乎平均分割了个案。在节点 9 中，预测信用等级为“良好”，但该节点中只有 56% 的个案实际具有良好信用等级。 这意味着该节点中几乎一半的个案 (44%) 将具有错误的预测类别。如果主要问题是识别不良信用风险，则此节点不会正常执行。

选择节点中的个案

我们来看一看节点 9 中的个案，以查看数据是否揭示了任何有用的附加信息。

- 双击浏览器中的树打开树编辑器。
- 单击节点 9 将其选中。（如果您要选择多个节点，请按住 Ctrl 并单击）。
- 从树编辑器菜单中选择：
  规则 > **过滤个案...**

图片 4-16
“过滤个案”对话框

“过滤个案”对话框将创建过滤变量，并基于该变量的值应用过滤设置。缺省的过滤变量名称为 filter_$。

- 所选节点中的个案将接收过滤变量值 1。
- 其他所有个案将接收值 0，这些个案将从后续分析中排除，直至您更改过滤状态。

在此示例中，这表示目前会将未位于节点 9 中的个案过滤掉（但不删除）。

- 单击**确定**创建过滤变量并应用过滤条件。

图片 4-17
数据编辑器中的已过滤个案

在数据编辑器中，使用贯穿行号的对角线来指示已过滤掉的个案。未位于节点 9 中的个案已被过滤掉。 未过滤节点 9 中的个案，因此后续分析将只包含节点 9 中的各个案。

检查所选个案

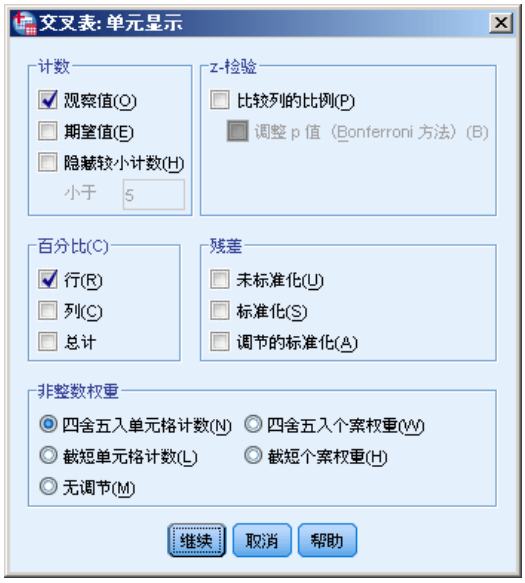
作为对节点 9 中的个案进行检查的第一步，您可能想要查看未在该模型中使用的变量。在此示例中，数据文件中的所有变量均包含在分析中，但其中的两个变量未包含在最终模型中：education 和 car loans。 为何过程会将这些变量从最终模型中省略，这可能出于一个充分的理由，因而它们可能不会告之我们过多信息，但无论如何，我们还是来看一看此原因。

▶ 从菜单中选择：
分析 > 描述统计 > 交叉表...

图片 4-18
“交叉表”对话框

- 选择 Credit rating 作为行变量。
- 选择 Education 和 Car loans 作为列变量。
- 单击**单元**。

图片 4-19
“交叉表：单元显示”对话框



- 在“百分比”组中，选择（选中）**行**。
- 然后单击**继续**，并在“交叉表”主对话框中单击**确定**以运行该过程。

检查交叉制表，您可以看到对于未包含在该模型中的两个变量，良好和不良信用等级类别中的个案之间没有太大差别。

图片 4-20
所选节点中个案的交叉制表

Credit rating* Education 交叉制表

			Education		合计
			High school	College	
Credit rating	Bad	计数	110	101	211
		Credit rating 的 %	52.1%	47.9%	100.0%
	Good	计数	128	144	272
		Credit rating 的 %	47.1%	52.9%	100.0%
合计		计数	238	245	483
		Credit rating 的 %	49.3%	50.7%	100.0%

Credit rating* Car loans 交叉制表

			Car loans		合计
			None or 1	More than 2	
Credit rating	Bad	计数	18	193	211
		Credit rating 的 %	8.5%	91.5%	100.0%
	Good	计数	39	233	272
		Credit rating 的 %	14.3%	85.7%	100.0%
合计		计数	57	426	483
		Credit rating 的 %	11.8%	88.2%	100.0%

- 对于 education，略超过半数的具有不良信用等级的个案仅为中学教育程度，而略超过半数的具有良好信用等级的个案为大学教育程度，但此差别在统计上并不明显。
- 对于 car loans，仅有一项或没有汽车贷款的良好信用个案的百分比高于相应的不良信用个案的百分比，但这两组中的绝大多数个案都有两项或更多项汽车贷款。

因此，尽管您对这些变量未包含在最终模型中的原因具有一定了解，但您并未深入了解如何才能更好地对节点 9 进行预测。如果未为分析指定其他变量，则在继续之前您可能希望对部分变量进行检查。

为结果分配成本

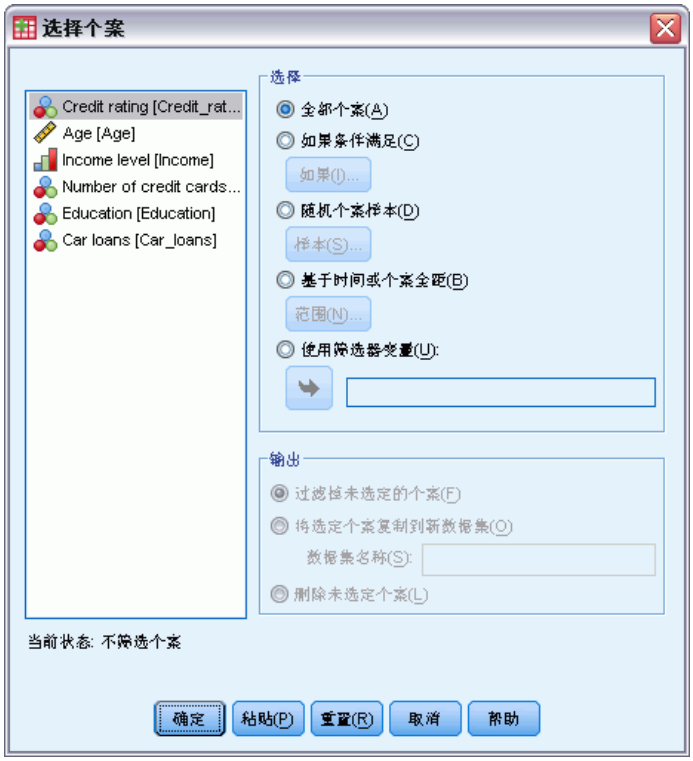
如前面所述，实际上节点 9 中几乎一半的个案分别位于两个信用等级类别中，除此之外，如果您的主要目标是要构建可正确识别不良信用风险的模型，则预测类别为“良好”类型的事实将会出现问题。 尽管您可能无法改善节点 9 的性能，但仍可以改进模型以提高对不良信用等级个案进行正确分类的比率，虽然这也将导致对良好信用等级个案进行误分类的比率较高。

首先，您需要关闭个案过滤，以便在分析中重新使用所有个案。

▶ 从菜单中选择：
数据 > 选择个案...

▶ 在“选择个案”对话框中，选择**全部个案**，然后单击**确定**。

图片 4-21
“选择个案”对话框



▶ 再次打开“决策树”对话框，并单击**选项**。

▶ 单击**误分类成本**选项卡。

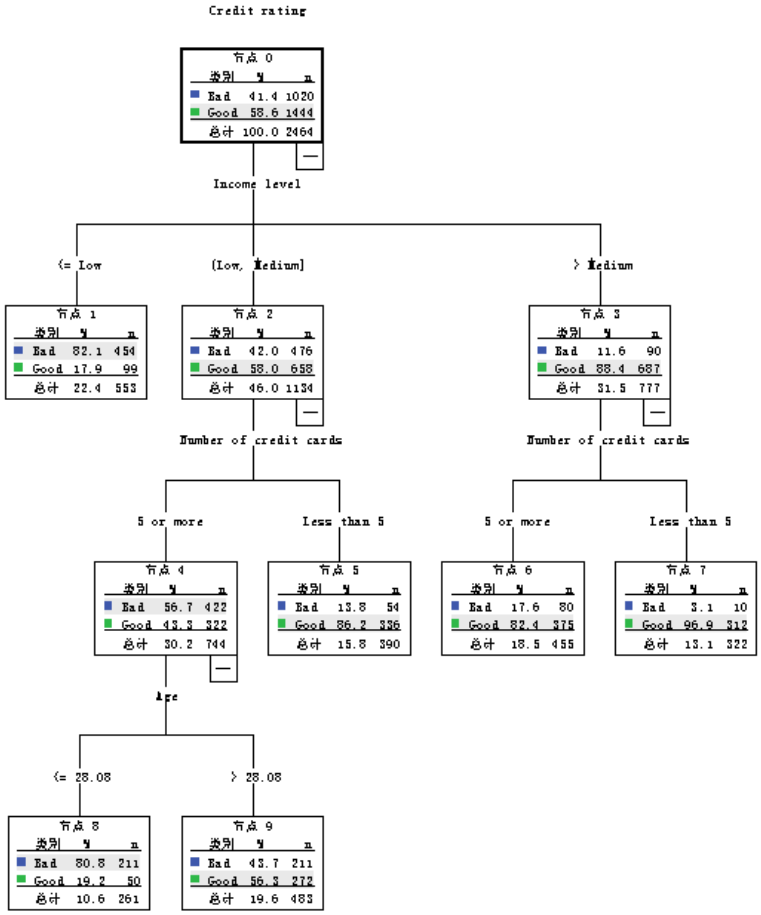
图片 4-22
“选项”对话框，“误分类成本”选项卡

▶ 选择**定制**，并为 Bad 实际类别/ Good 预测分类输入值 2。

这样，该过程即会了解，将不良信用风险错误地分类为良好信用风险的“成本”是将良好信用风险错误地分类为不良信用风险的“成本”的两倍。

▶ 单击**继续**，然后单击主对话框中的**确定**以运行该过程。

图片 4-23
具有调整的成本值的树模型



乍一看来，该过程生成的树基本上与原始树相同。但是，如果进一步检查，则会发现尽管每个节点中的个案分布并未更改，但某些预测类别已经更改。

对于终端节点，预测类别在所有节点中都保持相同，但一个节点除外：节点 9。预测类别当前为 Bad，即使略超过半数的个案位于 Good 类别中。

由于我们告之该过程将不良信用风险误分类为良好信用风险的成本较高，因此其中两种类别的个案当前几乎为平均分布的任何节点都具有 Bad 预测类别，即使绝大多数个案位于 Good 类别中。

预测类别的此更改将反映在分类表中。

图片 4-24
基于调整成本的风险和分类表

风险

估计	标准误差
.276	.011

增长方法: CHAID
因变量列表: Credit rating

分类

	预测		
观测	Bad	Good	百分比更正
Bad	902	118	88.4%
Good	443	1001	69.3%
整体百分比	54.6%	45.4%	77.2%

增长方法: CHAID
因变量列表: Credit rating

- 与以前的 65% 相比，现在几乎 86% 的不良信用风险已正确分类。
- 另一方面，良好信用风险的正确分类已从 90% 下降到 71%，且整体正确分类已从 79.5% 下降到 77.1%。

还要注意，风险估计与整体正确分类比率之间不再一致。如果整体正确分类比率为 77.1%，则您将需要 0.229 的风险估计值。 在此示例中，增加不良信用个案误分类的成本已提高了风险值，从而难以简明易懂地对其进行解释。

摘要

您可以使用树模型来将个案分类为由特定特征（例如，与具有良好和不良信用记录的银行客户相关的特征）识别的各组。如果特定预测结果比其他可能的结果更重要，您可以改进模型，,以使该结果与较高的误分类成本相关联，但如果降低一个结果的误分类比率，将会增加其他结果的误分类比率。

章 5

建立评分模型

“决策树”过程最强大、最有用的一项功能是能够建立随后可应用于其他数据文件的模型以预测结果。例如，基于包含人口统计信息和有关交通工具购买价格信息的数据文件，我们可以建立相关模型，用于预测有多少具有类似人口统计特征的人员有可能购置新车，然后，将该模型应用到其他数据文件，在这些文件中，人口统计信息可用,但有关以前交通工具购买情况的信息不可用。

对于本示例，我们将使用数据文件 tree_car.sav。 有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Decision Trees 20。

建立模型

- 要运行决策树分析，请从菜单中选择：
  分析 > **分类** > **树...**

图片 5-1
“决策树”对话框

- 选择 Price of primary vehicle 作为因变量。
- 选择所有剩余变量作为自变量。（该过程将自动排除任何对最终模型无明显作用的变量。）
- 选择 CRT 作为生长法。

▶ 单击**输出**。

图片 5-2
“输出”对话框，“规则”选项卡

▶ 单击**规则**选项卡。

▶ 选择（选中）**生成分类规则**。

▶ 选择 IBM SPSS Statistics 作为语法。

▶ 选择**为个案指定值**作为类型。

▶ 选择（选中）**将规则导出到文件**，然后输入文件名和目录位置。

请记住文件名和位置或者用笔记录下来，因为随后我们将用到它。如果不包含目录路径，则可能不知道文件的保存位置。您可以使用**浏览**按钮导航到特定（且有效）的目录位置。

▶ 单击**继续**，然后单击**确定**以运行该过程，并建立树模型。

评估模型

在将该模型应用到其他数据文件之前，您可以希望确保该模型与建立该模型所用的原始数据和谐匹配。

模型摘要

图片 5-3
模型摘要表

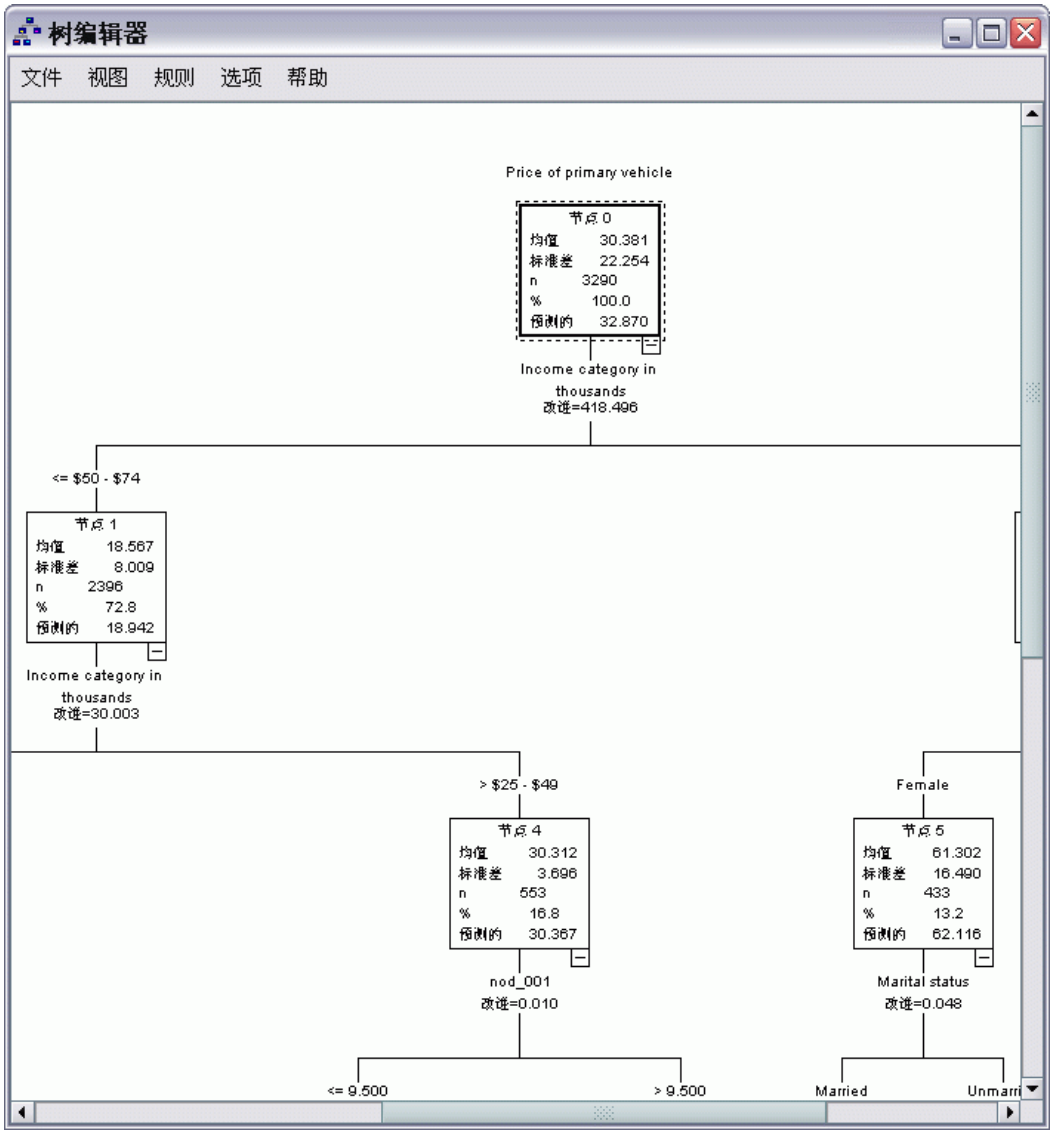
指定	增长方法	CRT
	因变量	Price of primary vehicle
	自变量	Age in years, Gender, Income category in thousands, Level of education, Marital status
	验证	无
	最大树深度	5
	父节点中的最小个案	100
	子节点中的最小个案	50
结果	自变量已包括	Income category in thousands, Age in years, Level of education
	节点数	29
	终端节点数	15
	深度	5

模型摘要表表明：在选定的自变量中，只有三个自变量对模型具有明显作用，需要将其包含在最终模型中：income、age 和 education。若要将此模型应用到其他数据文件，则需要了解这一重要信息，因为该模型中使用的自变量必须存在于要应用该模型的所有数据文件中。

该摘要表还表明树模型本身可能不是特别简单的模型，因为它拥有 29 个节点和 15 个终端节点。如果您需要的是能够实际应用的可靠模型而不是便于描述或说明的简单模型，这可能并不是问题。当然，对于实际用途，您可能还要求模型不要依赖于过多自变量（预测变量）。在本例中，由于最终模型中仅包含三个自变量，因此没有问题。

树模型图

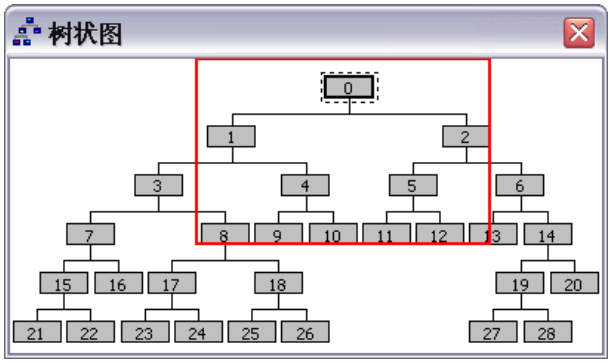
图片 5-4
树编辑器中的树模型图



树模型图具有如此多的节点，可能很难在保证能够阅读节点内容信息的情况下总览整个模型。您可以使用树状图来查看整个树：

- 双击浏览器中的树打开树编辑器。
- 从树编辑器菜单中选择：
  视图 > 树状图

图片 5-5
树状图



- 树状图显示整个树。您可以更改树状图窗口的大小，它将放大或缩小树状图的显示,以适合窗口大小。
- 树状图中突出显示的区域是树编辑器中当前显示的树区域。
- 您可以使用树状图在树中导航并选择节点。

有关详细信息，请参阅第 37 页码第 2 章中的树状图。

对于刻度因变量，各个节点显示因变量的平均值和标准差。节点 0 显示交通工具购买价格的总体平均值约为 29.9（以千为单位），标准差约为 21.6。

- 节点 1 表示收入低于 75（也是以千为单位）的个案，交通工具价格的平均值仅有 18.7。
- 相反，节点 2 表示收入等于或高于 75 的个案，交通工具价格的平均值为 60.9。

对树的进一步调查将显示 age 和 education 也显示与交通工具购买价格具有某种关系，但是现在我们主要关注模型的实际应用，而不是对其组件进行详细检查。

风险估计

图片 5-6
风险表

风险

估计	标准误差
68.485	2.985

增长方法: CRT
因变量列表: Price of primary vehicle

迄今为止，我们已检查的所有结果均未说明该模型是否为特别好的模型。衡量模型性能的一项指标是风险估计。对于刻度因变量，风险估计是度量节点内方差，其本身可能不会提供多少信息。方差越低表示模型越好，但方差是相对于测量单位的。例如，如果价格以个位而非以千位记录，则风险估计将放大一千倍。

对于刻度因变量，为了对风险估计进行有意义的合理解释，还需要进行如下工作：

- 总方差等于节点内（误差）方差加上节点间（解释的）方差。
- 节点内方差为风险估计值： 68.485.

- 在考虑任何自变量之前，总方差是因变量的方差，也就是根节点处的方差。
- 根节点处显示的标准差为 21.576；因此总方差是该值的平方： 465.524.
- 由于误差（未解释的方差）产生的方差比例为 68.485 / 465.524 = 0.147。
- 由该模型解释的方差比例为 1 - 0.147 = 0.853 或 85.3%，该值表明这是一个相当良好的模型。（这对分类因变量的整体正确分类率有类似的解释。）

将该模型应用到其他数据文件

已经确定该模型十分良好，现在我们可以将该模型应用到包含类似 age、income 和 education 变量的其他数据文件，并生成表示该文件中每个个案的预测交通工具购买价格的新变量。 此过程通常称为**评分**。

在生成模型后，我们指定为个案指定值的“规则”应该以命令语句形式保存在文本文件中。现在，我们将使用该文件中的命令在另一个数据文件中生成得分。

▶ 打开数据文件 tree_score_car.sav。有关详细信息，请参阅附录 A 中的样本文件中的 IBM SPSS Decision Trees 20。

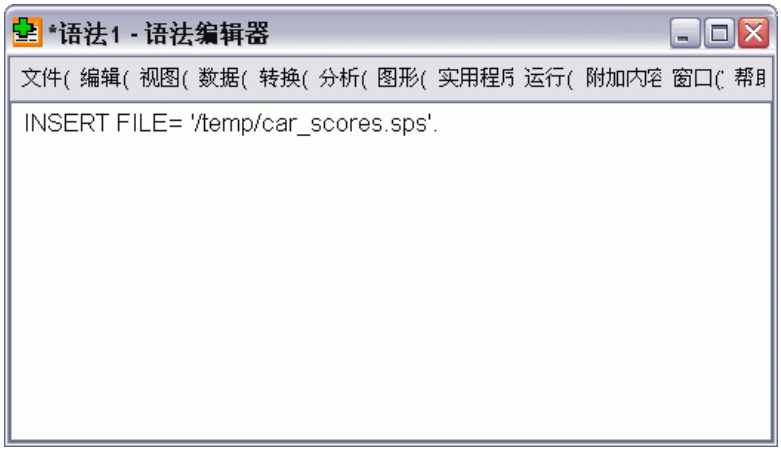
▶ 下一步，从菜单中选择：
文件 > 新建 > 语法

▶ 在命令语法窗口中，键入：

```
INSERT FILE=
 '/temp/car_scores.sps'.
```

如果您使用了其他文件名或位置，请进行相应的更改。

图片 5-7
使用 INSERT 命令运行命令文件的语法窗口



INSERT 命令将在指定的文件（创建该模型时生成的?规则?文件）中运行这些命令。

▶ 从命令语法窗口菜单中选择：
运行 > 全部

图片 5-8
添加到数据文件的预测值

这会向数据文件中添加两个新变量：

- nod_001 包含由模型为每个个案预测的终端节点编号。
- pre_001 包含每个个案的交通工具购买价格的预测值。

由于我们应用了用于为终端节点指定值的规则，因此可能的预测值的编号与终端节点的编号相同，在本例中为 15。例如，具有预测节点编号 10 的每个个案均具有相同的预测交通工具购买价格：30.56. 这并不是巧合，这是为原始模型中终端节点 10 报告的平均值。

虽然您通常会将该模型应用到其因变量值未知的数据，但在本示例中，我们实际应用该模型的数据文件包含了该信息，您可以将模型预测值与实际值进行比较。

▶ 从菜单中选择：

分析 > 相关 > 双变量...

▶ 选择 Price of primary vehicle 和 pre_001。

图片 5-9
?双变量相关?对话框

▶ 单击确定以运行该过程。

图片 5-10
实际交通工具价格和预测交通工具价格的相关性

		Price of primary vehicle	pre_001
Price of primary vehicle	Pearson 相关性	1	.919**
	显著性(双侧)		.000
	N	3290	3290
pre_001	Pearson 相关性	.919**	1
	显著性(双侧)	.000	
	N	3290	3290

**. 在 .01 水平(双侧)上显著相关。

相关性 0.92 表明实际交通工具价格和预测交通工具价格之间具有非常高的正相关性，这表明该模型运作正常。

摘要

可以使用“决策树”过程建立随后可应用于其他数据文件的模型，以预测结果。目标数据文件必须包含与最终模型中所包含的自变量名称相同的变量，以同一度规进行度量，且具有相同的用户定义的缺失值（如果有）。但是，从最终模型中排除的因变量和自变量可以不必在目标数据文件中出现。

章 6

树模型中的缺失值

不同的生长法以不同方式处理自变量（预测变量）的缺失值：

- CHAID 和穷举 CHAID 将每个自变量的所有系统缺失值和用户缺失值视为单个类别。对于刻度自变量和有序自变量，该类别以后可能（也可能不会）与该自变量的其他类别合并，具体取决于生长条件。
- CRT 和 QUEST 尝试对自变量（预测变量）使用**替代变量**。对于缺失该变量的值的个案，将使用与原始变量高度相关的其他自变量进行分类。这些备用预测变量称为替代变量。

本示例展示了用于该模型中的自变量存在缺失值时，CHAID 和 CRT 之间的差异。

对于本示例，我们将使用数据文件 tree_missing_data.sav。有关详细信息，请参阅附录 A 中的样本文件中的IBM SPSS Decision Trees 20。

注意：对于名义自变量和名义因变量，您可以选择将**用户缺失**值视为有效值，在该个案中像任何其他非缺失值那样对待这些值。 有关详细信息，请参阅第 20 页码第 1 章中的缺失值。

具有 CHAID 的缺失值

图片 6-1
存在缺失值的信用数据

类似于信用风险示例（有关详细信息，请参见第 4 章），本示例将尝试构建一个模型，以便对良好和较差的信用风险分类。主要不同之处在于此数据文件包含该模型中使用的某些自变量的缺失值。

- 要运行决策树分析，请从菜单中选择：
  分析 > **分类** > **树...**

图片 6-2
“决策树”对话框

- 选择 Credit rating 作为因变量。
- 选择所有其他变量作为自变量。（该过程将自动排除任何对最终模型无明显作用的变量。）
- 选择 CHAID 作为生长法。

对于本示例，我们需要保持非常简单的树结构。因此，我们将通过提高父节点和子节点的最小个案数来限制树的生长。

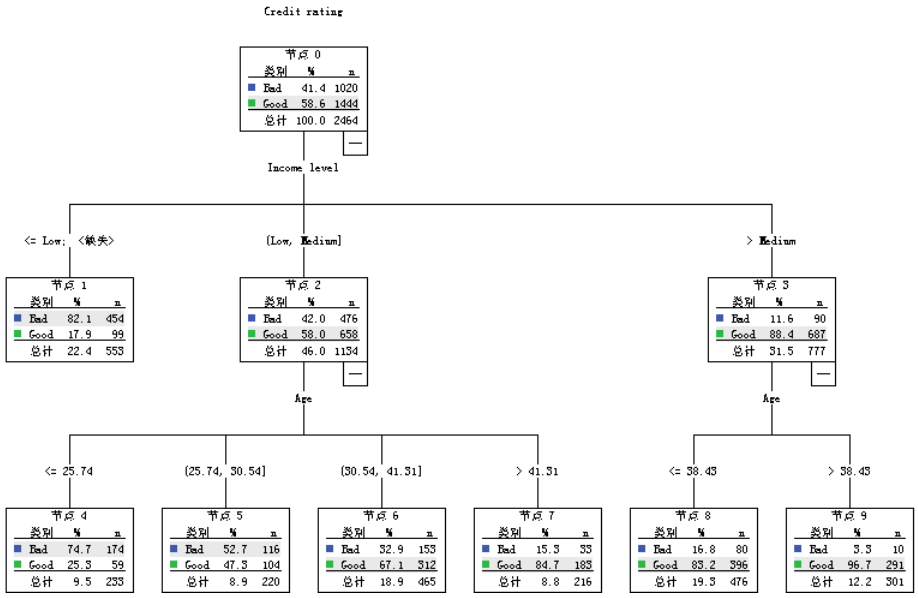
- 在“决策树”主对话框中，单击**条件**。

图片 6-3
“条件”对话框，“增长限制”选项卡

- 对于“最小个案数”，请为父节点键入 400，为子节点键入 200。
- 单击继续，然后单击确定运行该过程。

CHAID 结果

图片 6-4
具有缺失自变量值的 CHAID 树



对于节点 3，income level 的值显示为>Medium;<missing>。这意味着该节点包含高收入类别中的个案和 income level 存在缺失值的所有个案。

终端节点 10 包含 number of credit cards 为缺失值的个案。如果您对确定良好的信用风险感兴趣，该节点实际上不是最好的终端节点，如果您希望将该模型用于预测良好的信用风险，则该节点可能会出现问题。您可能不需要预测良好信用等级的模型，只因为您不了解关于一个个案具有多少信用卡的任何信息，并且其中的某些个案还可能缺失收入水平的信息。

图片 6-5
CHAID 模型的风险和分类表

风险

估计	标准误差
.218	.008

增长方法: CHAID
因变量列表: Credit rating

分类

	预测		
观测	Bad	Good	百分比更正
Bad	744	276	72.9%
Good	262	1182	81.9%
整体百分比	40.8%	59.2%	78.2%

增长方法: CHAID
因变量列表: Credit rating

风险和分类表指示 CHAID 模型当前约对 75% 的个案正确地进行了分类。这不算糟，但也并不太好。此外，我们可能有理由怀疑良好信用个案的正确分类比率可能过于乐观，因为它部分依赖于这样一种假设，即缺少有关两种自变量（income level 和 number of credit cards）的信息是良好信用的指示。

具有 CRT 的缺失值

除了我们将使用 CRT 作为生长法之外，现在让我们尝试相同的基本分析。

- 在“决策树”主对话框中，选择 CRT 作为生长法。
- 单击**条件**。
- 确保父节点的最小个案数仍设置为 400 子节点的最小个案数设置为 200。
- 单击**替代变量**选项卡。

注意：除非已选择 CRT 或 QUEST 作为生长法，否则将不会看到“替代变量”选项卡。

图片 6-6
“条件”对话框，“替代变量”选项卡

对于每个自变量节点分割，**自动**设置将把为该模型指定的每个其他自变量视为可能的替代变量。由于本示例中没有特别多的自变量，使用**自动**设置即可。

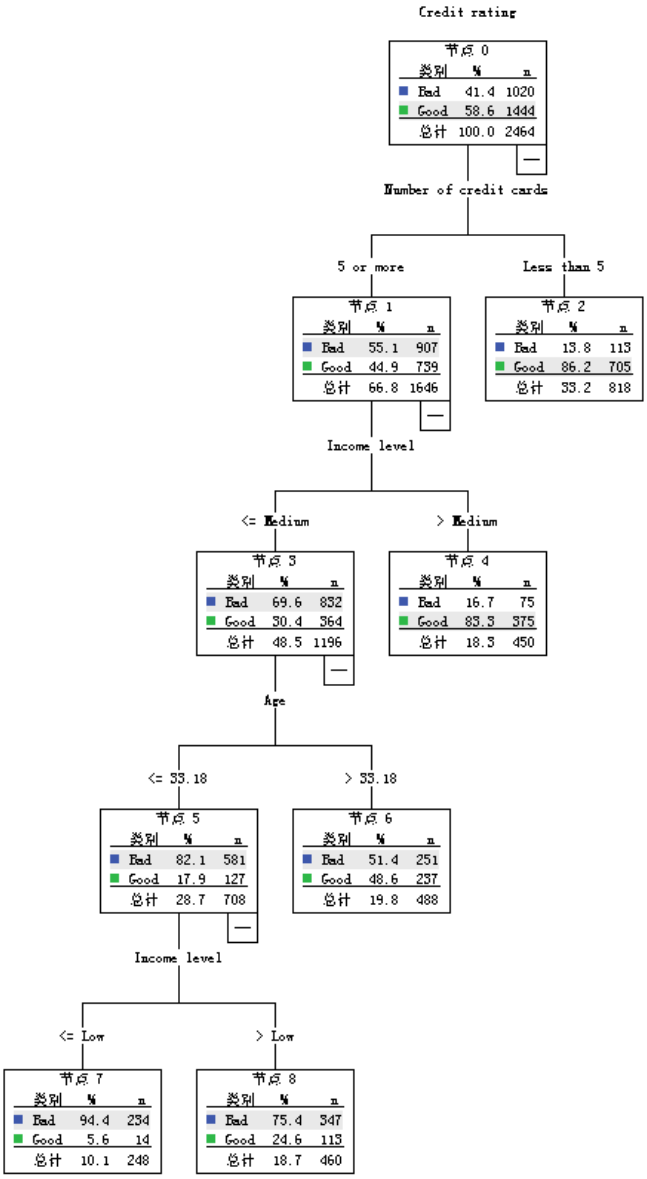
- 单击**继续**。
- 在“决策树”主对话框中，单击**输出**。

图片 6-7
“输出”对话框，“统计量”选项卡

- 单击**统计量**选项卡。
- 选择**替代变量（按分割）**。
- 单击**继续**，然后单击**确定**运行该过程。

CRT 结果

图片 6-8
具有缺失自变量值的 CRT 树



您可能会立即注意到该树的外观不太像 CHAID 树。就其本身而言，这并不能说明什么。在 CRT 树模型中，所有分割均为二元的，也就是说，每个父节点仅被分割为两个子节点。在 CHAID 模型中，父节点可分割为多个子节点。因此，即使这些树代表相同的基础模型，但它们的外观通常看起来也不会相同。

但还是有一些重要的不同之处：

- CRT 模型中最重要的自变量（预测变量）为 number of credit cards，而在 CHAID 模型中，最重要的预测变量为 income level。

- 对于少于五个信用卡的个案，number of credit cards 是信用等级的唯一重要预测变量，节点 2 为终端节点。
- 对于 CHAID 模型，income level 和 age 也包含于该模型，尽管 income level 目前是第二个（而非第一个）预测变量。
- 没有任何包含<缺失>类别的节点，因为在该模型中 CRT 使用替代预测变量而非缺失值。

图片 6-9
CRT 模型的风险和分类表

风险

估计	标准误差
.224	.008

增长方法: CRT
因变量列表: Credit rating

分类

	预测		
观测	Bad	Good	百分比更正
Bad	832	188	81.6%
Good	364	1080	74.8%
整体百分比	48.5%	51.5%	77.6%

增长方法: CRT
因变量列表: Credit rating

- 风险和分类表显示了几乎 78% 的整体正确分类比率比 CHAID 模型稍有增加 (75%)。
- 与 CHAID 模型仅有的 64.3% 相比，CRT 模型的较差信用个案的正确分类比率要高得多，达到81.6%。
- 但是，良好信用个案的正确分类比率从 CHAID 的 82.8% 降低到 CRT 的 74.8%。

替代变量

CHAID 和 CRT 模型之间差异的部分原因为在 CRT 模型中使用了替代变量。替代变量表指示了在该模型中使用替代变量的方式。

图片 6-10
替代变量表

父节点	自变量		提高	关联性
0	主	Number of credit cards	.090	
	替代	Car loans	.052	.643
		Age	.001	.004
1	主	Income level	.071	
	替代	Age	.001	.004
3	主	Age	.022	
5	主	Income level	.006	
	替代	Age	.000	.009

增长方法: CRT
因变量: Credit rating

- 在根节点（节点 0）处，最佳自变量（预测变量）是 number of credit cards。
- 对于任何 number of credit cards 存在缺失值的个案，使用 car loans 作为替代预测变量，因为该变量与 number of credit cards 具有相当高的关联性 (0.643)。

- 如果个案的 car loans 也存在缺失值，则使用 age 作为替代变量（尽管它仅具有相当低的关联值 0.004）。
- 还使用 Age 作为节点 1 和 5 处 income level 的替代变量。

摘要

不同的生长法采用不同的方式来处理缺失数据。如果数据用于创建包含多个缺失值的模型（或者如果您希望将该模型应用到包含多个缺失值的其他数据文件），则应评估缺失值对各种模型的影响。如果您希望在该模型中使用替代变量以补偿缺失值，请使用 CRT 或 QUEST 方法。

附录
A

样本文件

随产品一起安装的样本文件可以在安装目录的 Samples 子目录中找到。对于以下每种语言在“样本”子目录中有单独的文件夹：英语、法语、德语、意大利语、日语、韩语、波兰语、俄语、简体中文、西班牙语和繁体中文。

并非所有样本文件均提供此处的全部语言版本。如果样本文件未提供某种语言的版本，则相应语言文件夹中包含该样本文件的英语版本。

描述

以下是对在整个文档的各种示例中使用的样本文件的简要描述。

- **accidents.sav。** 该假设数据文件涉及某保险公司，该公司正在研究给定区域内汽车事故的年龄和性别风险因子。每个个案对应一个年龄类别和性别类别的交叉分类。
- **adl.sav。**该假设数据文件涉及在确定针对脑卒中患者的建议治疗类型的优点方面的举措。医师将女性脑卒中患者随机分配到两组中的一组。第一组患者接受标准的物理治疗，而第二组患者则接受附加的情绪治疗。在进行治疗的三个月时间里，将为每个患者进行一般日常生活行为的能力评分并作为原始变量。
- **advert.sav。**该假设数据文件涉及某零售商在检查广告支出与销售业绩之间的关系方面的举措。为此，他们收集了过去的销售数据以及相关的广告成本..
- **aflatoxin.sav。**该假设数据文件涉及对谷物的黄曲霉毒素的检测，该毒素的浓度会因谷物产量的不同（不同谷物之间及同种谷物之间）而有较大变化。谷物加工机从 8 个谷物产量的每一个中收到 16 个样本并以十亿分之几 (PPB) 为单位来测量黄曲霉毒素的水平。
- **anorectic.sav。**在研究厌食/暴食行为的标准症状参照时，研究人员对 55 名已知存在进食障碍的青少年进行了调查。其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。
- **bankloan.sav。**该假设数据文件涉及某银行在降低贷款拖欠率方面的举措。该文件包含 850 位过去和潜在客户的财务和人口统计信息。前 700 个个案是以前曾获得贷款的客户。剩下的 150 个个案是潜在客户，银行需要按高或低信用风险对他们进行分类。
- **bankloan_binning.sav。**该假设数据文件包含 5,000 位过去客户的财务和人口统计信息。
- **behavior.sav。**在一个经典示例中，52 名学生被要求以 10 分的标度对 15 种情况和 15 种行为的组合进行评价，该 10 分的标度介于 0 =平均值在个人值之上，值被视为相异性。
- **behavior_ini.sav。**该数据文件包含 behavior.sav 的二维解的初始配置。

- **brakes.sav。**该假设数据文件涉及某生产高性能汽车盘式制动器的工厂的质量控制。该数据文件包含对 8 台专用机床中每一台的 16 个盘式制动器的直径测量。盘式制动器的目标直径为 322 毫米。
- **breakfast.sav。**在一项经典研究中，21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价，从 1 =他们的喜好根据六种不同的情况加以记录，从“全部喜欢”到“只带饮料的快餐”。
- **breakfast-overall.sav。**该数据文件只包含早餐食品喜好的第一种情况，即“全部喜欢”。
- **broadband_1.sav。**该假设数据文件包含各地区订制了全国宽带服务的客户的数量。该数据文件包含 4 年期间 85 个地区每月的订户数量。
- **broadband_2.sav。**该数据文件和 broadband_1.sav 一样，但包含另外三个月的数据。
- **car_insurance_claims.sav。**在别处被提出和分析的关于汽车损坏赔偿的数据集。平均理赔金额可以当作其具有 gamma 分布来建模，通过使用逆联接函数将因变量的均值与投保者年龄、车辆类型和车龄的线性组合关联。提出理赔的数量可以作为尺度权重。
- **car_sales.sav。**该数据文件包含假设销售估计值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。
- **car_sales_uprepared.sav。**这是 car_sales.sav 的修改版本，不包含字段的任何已转换版本。
- **carpet.sav。**在一个常用示例中，一家公司非常重视一种新型地毯清洁用品的市场营销，希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品 标志和退货保证。包装设计有三个因子水平，每个因子水平因刷体位置而不同；有三个品牌名称（K2R、Glory 和 Bissell）；有三个价格水平；最后两个因素各有两个级别（有或无）。十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。
- **carpet_prefs.sav。**该数据文件所基于的示例和在 carpet.sav 中所描述的一样，但它还包含从 10 位消费者的每一位中收集到的实际排列顺序。消费者被要求按照从最喜欢到最不喜欢的顺序对 22 个产品特征进行排序。carpet_plan.sav 中定义了变量 PREF1 到 PREF22 包含相关特征的标识符。
- **catalog.sav。**该数据文件包含某编目公司出售的三种产品的假设每月销售数据。同时还包括 5 个可能的预测变量的数据。
- **catalog_seasfac.sav。**除添加了一组从“季节性分解”过程中计算出来的季节性因子和附带的日期变量外，该数据文件和 catalog.sav 是相同的。
- **cellular.sav。**该假设数据文件涉及某便携式电话公司在减少客户流失方面的举措。客户流失倾向分被应用到帐户，分数范围从 0 到 100。得到 50 分或更高分数的帐户可能会更换提供商。
- **ceramics.sav。**该假设数据文件涉及某制造商在确定新型优质合金是否比标准合金具有更高的耐热性方面的举措。每个个案代表对一种合金的单独检验；个案中会记录合金的耐热极限。
- **cereal.sav。**该假设数据文件涉及一份 880 人参于的关于早餐喜好的民意调查，该调查记录了参与者的年龄、性别、婚姻状况以及生活方式是否积极（根据他们是否每周至少做两次运动）。每个个案代表一个单独的调查对象。

- **clothing_defects.sav。**这是关于某服装厂的质量控制过程的假设数据文件。检验员要对工厂中每次大批量生产的服装进行抽样检测并清点不合格的服装的数量。
- **coffee.sav。**这是关于六种冰咖啡的认知品牌形象的数据文件。对于 23 种冰咖啡特征属性中的每种属性，人们选择了由该属性所描述的所有品牌。为保密起见，六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。
- **contacts.sav。**该假设数据文件涉及一组公司计算机销售代表的联系方式列表。根据这些销售代表所在的公司部门及其公司的秩来对每个联系方式进行分类。同时还记录了最近一次的销售量、最近一次销售距今的时间和所联系公司的规模。
- **creditpromo.sav。**该假设数据文件涉及某百货公司在评价最新信用卡促销的效果方面的举措。为此，随机选择了 500 位持卡人。其中一半收到了宣传关于在接下来的三个月内降低消费利率的广告。另一半收到了标准的季节性广告。
- **customer_dbase.sav。** 该假设数据文件涉及某公司在使用数据仓库中的信息来为最有可能回应的客户提供特惠商品方面的举措。随机选择客户群的子集并为其提供特惠商品，同时记录下他们的回应。
- **customer_information.sav。**该假设数据文件包含客户邮寄信息，如姓名和地址。
- **customer_subset.sav。**来自 customer_dbase.sav 的拥有 80 个个案的子集。
- **debate.sav。**该假设数据文件涉及在某政治辩论前后对该辩论的参与者所做的调查的成对回答。每个个案对应一个单独的调查对象。
- **debate_aggregate.sav。**该假设数据文件分类汇总了 debate.sav 中的回答。每个个案对应一个辩论前后的偏好的交叉分类。
- **demo.sav。**这是关于购物客户数据库的假设数据文件，用于寄出每月的商品。将记录客户对商品是否有回应以及各种人口统计信息。
- **demo_cs_1.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第一步。每个个案对应不同的城市，并记录地区、省、区和城市标识。
- **demo_cs_2.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第二步。每个个案对应来自第一步中所选城市的不同的家庭单元，并记录地区、省、区、市、子区和单元标识。还包括设计前两个阶段的抽样信息。
- **demo_cs.sav。**该假设数据文件包含用复杂抽样设计收集的调查信息。每个个案对应不同的家庭单元，并记录各种人口统计和抽样信息。
- **dmdata.sav。**该假设数据文件包含直销公司的人口统计学和购买信息。dmdata2.sav 包含收到测试邮件的联系人子集的信息，dmdata3.sav 包含未收到测试邮件的其余联系人的信息。
- **dietstudy.sav。**该假设数据文件包含对 “Stillman diet” 的研究结果。每个个案对应一个单独的主体，并记录其在实行饮食方案前后的体重（磅）以及甘油三酸酯的水平（毫克/100 毫升）。
- **dvdplayer.sav。**这是关于开发新的 DVD 播放器的假设数据文件。营销团队用原型收集了焦点小组数据。每个个案对应一个单独的被调查用户，并记录他们的人口统计信息及其对原型问题的回答。
- **german_credit.sav。**该数据文件取自加州大学欧文分校的 Repository of Machine Learning Databases 中的 “German credit” 数据集。
- **grocery_1month.sav。**该假设数据文件是在数据文件 grocery_coupons.sav 的基础上加上了每周购物“累计”，所以每个个案对应一个单独的客户。所以，一些每周更改的变量消失了，而且现在记录的消费金额是为期四周的研究过程中的消费金额之和。

- **grocery_coupons.sav。**该假设数据文件包含由重视顾客购物习惯的杂货连锁店收集的调查数据。对每位顾客调查四周，每个个案对应一个单独的顾客周，并记录有关顾客购物地点和方式的信息（包括那一周里顾客在杂货上的消费金额）。
- **guttman.sav。** Bell 创建了一个表，用来阐释可能的社会群体。Guttman 引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。
- **health_funding.sav。**该假设数据文件包含关于保健基金（每 100 人的金额）、发病率（每 10,000 人的比率）以及保健提供商拜访率（每 10,000 的比率）的数据。每个个案代表不同的城市。
- **hivassay.sav。**该假设数据文件涉及某药物实验室在开发用于检测 HIV 感染的快速化验方面的举措。化验结果为八个加深的红色阴影，如果有更深的阴影则表示感染的可能性很大。用 2,000 份血液样本来进行实验室试验，其中一半受到 HIV 感染而另一半没有受到感染。
- **hourlywagedata.sav。**该假设数据文件涉及在政府机关和医院工作的具有不同经验水平的护士的时薪。
- **insurance_claims.sav。**该假设数据文件涉及某保险公司，该公司希望构建一个模型用于标记可疑的、具有潜在欺骗性的理赔。每个个案代表一次单独的理赔。
- **insure.sav。**该假设数据文件涉及某保险公司，该公司正在研究指示客户是否会根据 10 年的人寿保险合同提出理赔的风险因子。数据文件中的每个个案代表一副根据年龄和性别进行匹配的合同，其中一份记录了一次理赔而另一份则没有。
- **judges.sav。**该假设数据文件涉及经过训练的裁判（加上一个体操爱好者）对 300 次体操表演给出的分数。每行代表一次单独的表演；裁判们观看相同的表演。
- **kinship_dat.sav。**Rosenberg 和 Kim 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”。每个源对应一个 15×15 的相似性矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。
- **kinship_ini.sav。**该数据文件包含 kinship_dat.sav 的三维解的初始配置。
- **kinship_var.sav。**该数据文件包含自变量 gender、gener(ation) 和 degree (of separation)，这些变量可用于解释 kinship_dat.sav 的解的维数。具体而言，它们可用来将解的空间限制为这些变量的线性组合。
- **marketvalues.sav。**该数据文件涉及 1999 - 2000 年间 Algonquin，Ill. 地区新的房屋开发中的住房销售。这些销售仅仅来自公众记录。
- **nhis2000_subset.sav。**美国健康访问调查 (NHIS) 是针对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态的人口统计信息和观察数据。该数据文件包含取自 2000 年调查信息的子集。国家健康统计中心。2000 年美国健康访问调查。公用数据文件和文档。

ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年发布。

- **ozone.sav。**这些数据包含了用来根据其余变量预测臭氧浓度的六个气象变量的 330 个观察值。在以前的研究人员中， 和 发现了这些变量之间的非线性，这妨碍了标准回归方法。
- **pain_medication.sav。**该假设数据文件包含用于治疗慢性关节炎疼痛的抗炎药的临床试验结果。我们感兴趣的是该药见效的时间以及它和现有药物的比较。
- **patient_los.sav。**该假设数据文件包含被医院确诊为疑似心肌梗塞（即 MI 或“心脏病发作”）的患者的治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **patlos_sample.sav。**该假设数据文件包含在治疗心肌梗塞（即 MI 或“心脏病发作”）期间收到溶解血栓剂的患者的样本治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **poll_cs.sav。**该假设数据文件涉及民意测验专家在确定正式立法前公众对法案的支持水平方面的举措。个案对应注册的选民。每个个案记录选民居住的县、镇、区。
- **poll_cs_sample.sav。**该假设数据文件包含在 poll_cs.sav 中列出的选民的样本。该样本是根据 poll.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。请注意，由于该抽样计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (poll_jointprob.sav) 包含联合选择概率。在选取了样本之后，对应于选民人群统计信息及其对提交法案的意见的附加变量将被收集并添加到数据文件。
- **property_assess.sav。**该假设数据文件涉及某县资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应过去一年中县里所出售的资产。数据文件中的每个个案记录资产所在的镇、最后评估资产的评估员、该次评估距今的时间、当时的估价以及资产的出售价格。
- **property_assess_cs.sav。**该假设数据文件涉及某州资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应该州的资产。数据文件中的每个个案记录资产所在的县、镇和区，最后一次评估距今的时间以及当时的估价。
- **property_assess_cs_sample.sav。**该假设数据文件包含在 property_assess_cs.sav 中列出的资产的样本。该样本是根据 property_assess.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。在选取了样本之后，附加变量 Current value 将被收集并添加到数据文件。
- **recidivism.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料；如果在第一次被捕后两年内又第二次被捕，则还将记录两次被捕间隔的时间。
- **recidivism_cs_sample.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应在 2003 年 6 月期间第一次被捕释放的先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料，及其第二次被捕的数据（如果发生在 2006 年 6 月底之前）。根据 recidivism_cs.csplan 中指定的抽样计划从抽样部门选择罪犯；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (recidivism_cs_jointprob.sav) 包含联合选择概率。
- **rfm_transactions.sav。** 此假设数据文件包含购买交易数据，即每笔交易的购买日期、购买商品和消费金额。

- **salesperformance.sav。**这是关于评估两个新的销售培训课程的假设数据文件。60 名员工被分成 3 组且都接受标准的培训。另外，组 2 接受技术培训；组 3 接受实践教程。在培训课程结束时，对每名员工进行测验并记录他们的分数。数据文件中的每个个案代表一名单独的受训者，并记录其被分配到的组以及测验的分数。
- **satisf.sav。**该假设数据文件涉及某零售公司在 4 个商店位置所进行的满意度调查。总共对 582 位客户进行了调查，每个个案代表一位单独客户的回答。
- **screws.sav。**该数据文件包含关于螺钉、螺栓、螺母和图钉的特征的信息。
- **shampoo_ph.sav**这是关于某发制品厂的质量控制的假设数据文件。在规定的时间间隔对六批独立输出的产品进行检测并记录它们的 pH 值。目标范围是 4.5 - 5.5。
- **ships.sav。**在别处被提出和分析的关于波浪对货船造成的损坏的数据集。在给定了船的类型、建造工期和服务期后，可以根据以泊松比率发生来为事件计数建模。在因子交叉分类构成的表格中，每个单元格的分类汇总服务月数提供遇到风险的值。
- **site.sav。**该假设数据文件涉及某公司在为扩展业务而选择新址方面的举措。该公司聘请了两名顾问分别对选址进行评估，除了提供长期报告外，他们还要以“前景颇佳”、“前景良好”或“前景不佳”来对每个选址进行总结。
- **smokers.sav。**该数据文件摘自 1998 年全国家庭药物滥用调查并且是美国家庭的概率样本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，分析该数据文件的第一步应该是对数据进行加权以反映总体趋势。
- **stocks.sav** 该假设数据文件包含某一年的股票价格和成交量。
- **stroke_clean.sav。**该假设数据文件包含某医学数据库在经过“数据准备”选项中的过程清理后的状态。
- **stroke_invalid.sav。**该假设数据文件包含某医学数据库的初始状态及一些数据输入错误。
- **stroke_survival。**此假设数据文件涉及正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。中风后，记录心肌梗塞、缺血性中风或出血性中风的发生及其时间。样本为左侧截短，因为只包含在中风后管理的复元计划结束后存活的患者。
- **stroke_valid.sav。**该假设数据文件包含在使用“验证数据”过程检查值后，某医学数据库的状态。它仍包含潜在异常个案。
- **survey_sample.sav。**此数据文件包含调查数据，包括人口统计学数据和各种态度测量。它基于 1998 NORC 综合社会调查的变量子集，但某些数据值已经过修改，并添加了其他虚拟变量以供演示用途。
- **telco.sav。**该假设数据文件涉及某电信公司在减少客户群中的客户流失方面的举措。每个个案对应一个单独的客户，并记录各类人口统计和服务用途信息。
- **telco_extra.sav。**该数据文件与 telco.sav 数据文件类似，但删除了 “tenure” 和经对数转换的客户消费变量，代替它们的是标准化的对数转换客户消费变量。
- **telco_missing.sav。** 该数据文件是 telco.sav 数据文件的子集，但某些人口统计数据值已被缺失值替换。
- **testmarket.sav。**该假设数据文件涉及某快餐连锁店为其菜单添加新项目的计划。有三种可能的促销新产品的活动，所以会在多个随机选择的市场中的地点引入新的项目。在每个地点采用不同的促销方式，并记录新项目前四周的每周销售情况。每个个案对应单独地点的一周。

- **testmarket_1month.sav。**该假设数据文件是在数据文件 testmarket.sav 的基础上加上了每周销售“累计”，所以每个个案对应一个单独的地点。所以，一些每周更改的变量消失了，而且现在记录的销售是为期四周的研究过程中的销售之和。
- **tree_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_credit.sav。**该假设数据文件包含人口统计和银行贷款历史数据。
- **tree_missing_data.sav** 该假设数据文件包含具有大量缺失值的人口统计和银行贷款历史数据。
- **tree_score_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_textdata.sav。**这是一个只有两个变量的样本数据文件，主要打算在指定测量级别和值标签之前显示变量的默认状态。
- **tv-survey.sav。**该假设数据文件涉及由某电视演播室进行的一项关于是否要继续制作一档成功的节目的调查。906 位调查对象被问及他们在各种情况下是否会收看该节目。每行代表一位单独的调查对象；每列代表一种单独的情况。
- **ulcer_recurrence.sav。**此文件包含某项研究的部分信息，该研究旨在比较两种用来防止溃疡复发的治疗的功效。它提供了区间数据的优秀示例并且已在别处被提出和分析。
- **ulcer_recurrence_recoded.sav。**该文件重新组织 ulcer_recurrence.sav 中的信息以允许为研究的每个区间的事件概率建模而不是简单地为研究结束事件概率建模。它已在别处被提出和分析。
- **verd1985.sav。**该数据文件涉及某项调查。该调查记录了 15 个主体对 8 个变量的响应。需要处理的变量被分成 3 个集。数据集 1 包含 年龄 和 婚姻；数据集 2 包含 宠物 和 新闻；数据集 3 包含 音乐 和 居住。宠物被尺度化为多名义而年龄被尺度化为有序；所有其他变量都被尺度化为单名义。
- **virus.sav。**该假设数据文件涉及某因特网服务提供商 (ISP) 在确定病毒对其网络的影响方面的举措。他们从发现病毒到威胁得以遏制这段时间内跟踪其网络上受感染的电子邮件的流量的（近似）百分比。
- **wheeze_steubenville.sav。**这是关于空气污染对儿童健康影响的纵向研究的一个子集。这些数据包含儿童的气喘状况的重复二分类测量（这些儿童来自 Steubenville，Ohio，年龄为 7 到 10 岁），以及母亲在研究的第一年中是否为吸烟者的固定记录。
- **workprog.sav。**该假设数据文件涉及一份尝试为弱势群体提供较好的工作的政府工作计划。文件后还有一个潜在计划参与者的样本，其中一些参与者是被随机选择来参加该计划的，而其他参与者则不是。每个个案代表一位单独的计划参与者。
- **worldsales.sav** 该假设数据文件包含按不同大洲和产品列出的销售收入。

附录
B

注意事项

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。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。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。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.

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

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

以下段落不适用于英国或此类条款与当地法律不符的其他国家/地区： INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。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 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

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。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

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 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

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。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 http://www.ibm.com/legal/copytrade.shtml。

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。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

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。

此产品使用 Polar 工程咨询公司的 WinWrap Basic，版权 1993 年-2007 年，http://www.winwrap.com。

其他产品和服务名称可能是 IBM 或其他公司的商标。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

索引