Chapter 5
Build Decision Trees

About the Tasks That You Will Perform

Now that you have verified the input data, it is time to build predictive models. You perform the following tasks to model the input data using nonparametric decision trees:

1. You enable SAS Enterprise Miner to automatically train a full decision tree and to automatically prune the tree to an optimal size. When training the tree, you select split rules at each step to maximize the split decision logworth. Split decision logworth is a statistic that measures the effectiveness of a particular split decision at differentiating values of the target variable. For more information about logworth, see the SAS Enterprise Miner Help.
2. You interactively train a decision tree. At each step, you select from a list of candidate rules to define the split rule that you deem to be the best.
3. You use a Gradient Boosting node to generate a set of decision trees that form a single predictive model. Gradient boosting is a boosting approach that resamples the analysis data set several times to generate results that form a weighted average of the re-sampled data set.

Automatically Train and Prune a Decision Tree

Decision tree models are advantageous because they are conceptually easy to understand, yet they readily accommodate nonlinear associations between input variables and one or more target variables. They also handle missing values without the need for imputation. Therefore, you decide to first model the data using decision trees. You will compare decision tree models to other models later in the example.

However, before you add and run the Decision Tree node, you will add a Control Point node. The Control Point node is used to simplify a process flow diagram by reducing the number of connections between multiple interconnected nodes. By the end of this example, you will have created five different models of the input data set, and two Control Point nodes to connect these nodes. The first Control Point node, added here, will distribute the input data to each of these models. The second Control Point node will collect the models and send them to evaluation nodes.

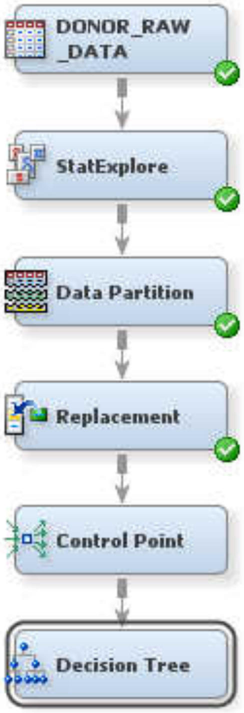
To use the Control Point node:

1. Select the **Utility** tab on the Toolbar.
2. Select the **Control Point** node icon. Drag the node into the Diagram Workspace.
3. Connect the **Replacement** node to the **Control Point** node.

SAS Enterprise Miner enables you to build a decision tree in two ways: automatically and interactively. You will begin by letting SAS Enterprise Miner automatically train and prune a tree.

To use the **Decision Tree** node to automatically train and prune a decision tree:

1. Select the **Model** tab on the Toolbar.
2. Select the **Decision Tree** node icon. Drag the node into the Diagram Workspace.
3. Connect the **Control Point** node to the **Decision Tree** node.



4. Select the **Decision Tree** node. In the Properties Panel, scroll down to view the **Train** properties:
   - Click on the value of the **Maximum Depth** splitting rule property, and enter `10`. This specification enables SAS Enterprise Miner to train a tree that includes up to ten generations of the root node. The final tree in this example, however, will have fewer generations due to pruning.
   - Click on the value of the **Leaf Size** node property, and enter `8`. This specification constrains the minimum number of training observations in any leaf to eight.

- Click on the value of the **Number of Surrogate Rules** node property, and enter `4`. This specification enables SAS Enterprise Miner to use up to four surrogate rules in each non-leaf node if the main splitting rule relies on an input whose value is missing.

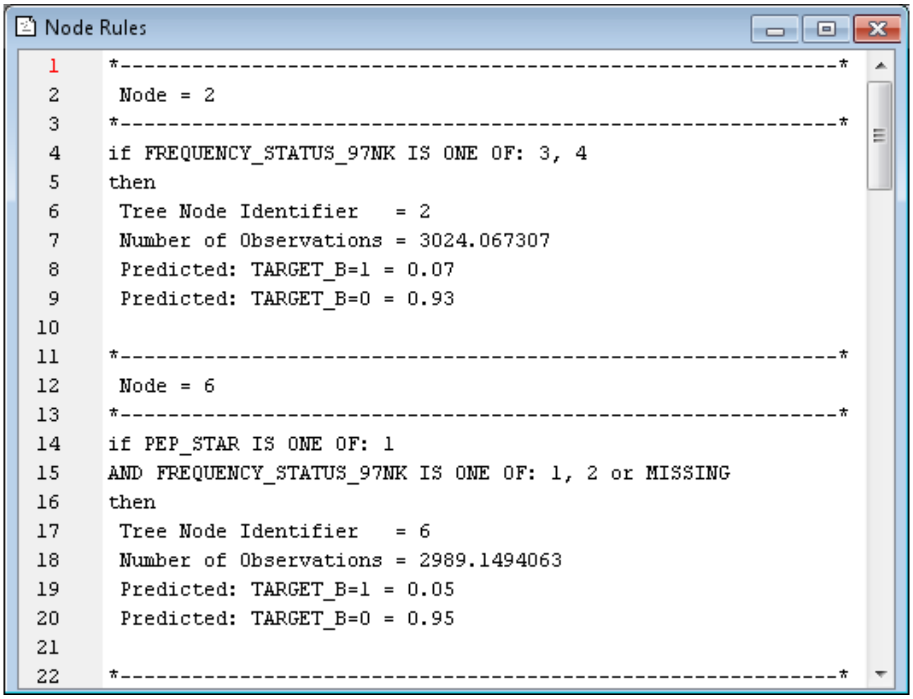
Note: The **Assessment Measure** subtree property is automatically set to **Decision** because you defined a profit matrix in "Create a Data Source" on page 11. Accordingly, the Decision Tree node will build a tree that maximizes profit in the validation data.

5. In the Diagram Workspace, right-click the Decision Tree node, and select **Run** from the resulting menu. Click **Yes** in the Confirmation window that opens.

6. In the window that appears when processing completes, click **Results**. The Results window appears.

   a. On the **View** menu, select **Model** ⇨ **Node Rules**. The English Rules window appears.

   b. Expand the Node Rules window. This window contains the IF-THEN logic that distributes observations into each leaf node of the decision tree.



```
*------------------------------------------------------------*
 Node = 2
*------------------------------------------------------------*
if FREQUENCY_STATUS_97NK IS ONE OF: 3, 4
then
 Tree Node Identifier   = 2
 Number of Observations = 3024.067307
 Predicted: TARGET_B=1 = 0.07
 Predicted: TARGET_B=0 = 0.93

*------------------------------------------------------------*
 Node = 6
*------------------------------------------------------------*
if PEP_STAR IS ONE OF: 1
AND FREQUENCY_STATUS_97NK IS ONE OF: 1, 2 or MISSING
then
 Tree Node Identifier   = 6
 Number of Observations = 2989.1494063
 Predicted: TARGET_B=1 = 0.05
 Predicted: TARGET_B=0 = 0.95

*------------------------------------------------------------*
```

In the Output window, the **Tree Leaf Report** indicates that there are seven leaf nodes in this tree. For each leaf node, the following information is listed:

- node number
- number of training observations in the node
- percentage of training observations in the node with TARGET_B=1 (did donate), adjusted for prior probabilities
- percentage of training observations in the node with TARGET_B=0 (did not donate), adjusted for prior probabilities

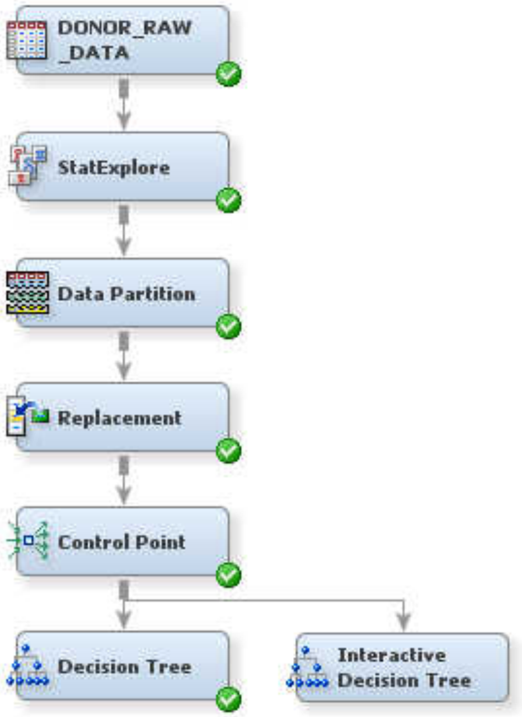
This tree has been automatically pruned to an optimal size. Therefore, the node numbers that appear in the final tree are not sequential. In fact, they reflect the positions of the nodes in the full tree, before pruning.

7. Close the Results window.

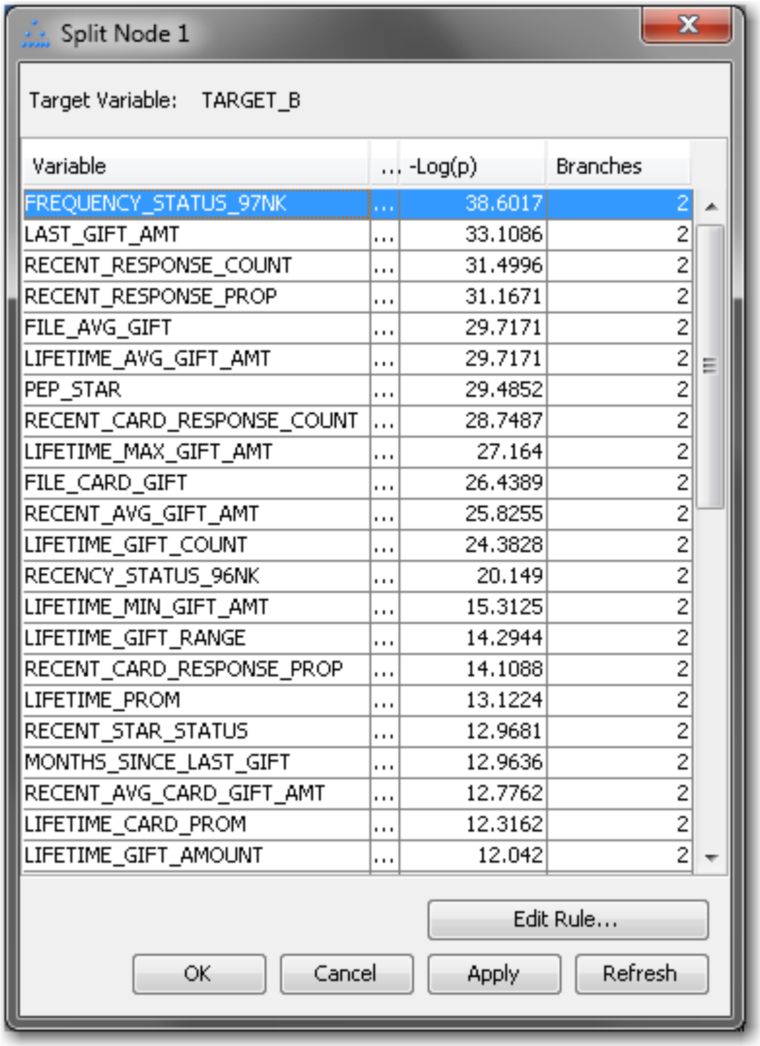
Interactively Train a Decision Tree

To use the Decision Tree node to interactively train and prune a decision tree:

1. From the **Model** tab on the Toolbar, select the **Decision Tree** node icon. Drag the node into the Diagram Workspace.
2. In the Diagram Workspace, right-click the **Decision Tree** node, and select **Rename** from the resulting menu. Enter `Interactive Decision Tree` and then click **OK** in the window that opens.
3. Connect the Control Point node to the Interactive Decision Tree node.



4. Select the **Interactive Decision Tree** node. In the Properties Panel, in the **Train** properties group, click on the ellipses that represent the value of **Interactive**. The Interactive Decision Tree window appears.
   a. Select the root node (at this point, the only node in the tree), and then from the **Action** menu, select **Split Node**. The Split Node window appears that lists the candidate splitting rules ranked by logworth (-Log(p)). The FREQUENCY_STATUS_97NK rule has the highest logworth. Ensure that this row is selected, and click **OK**.

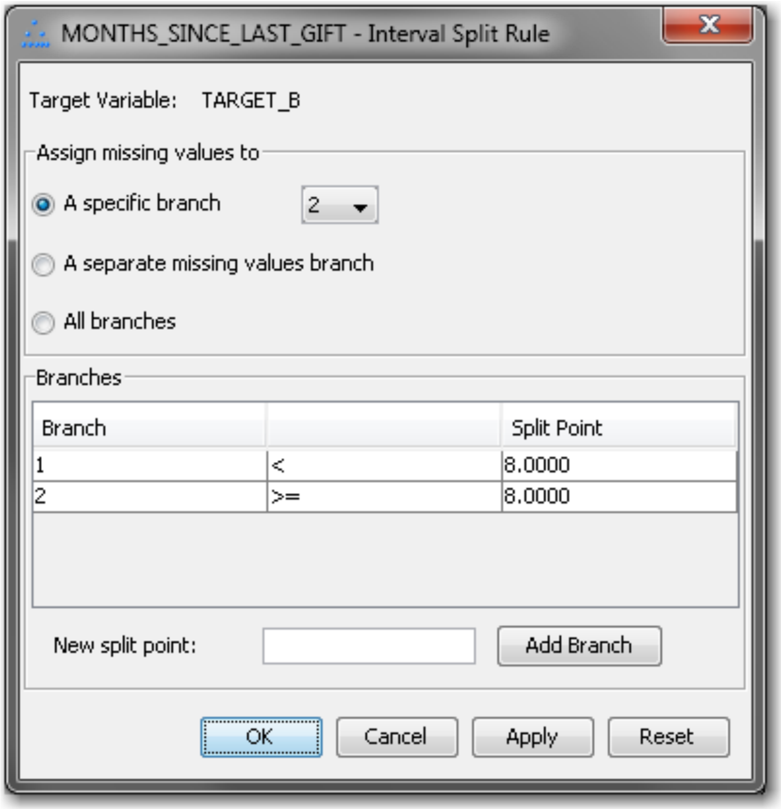


Target Variable: TARGET_B

Variable	...	-Log(p)	Branches
FREQUENCY_STATUS_97NK	...	38.6017	2
LAST_GIFT_AMT	...	33.1086	2
RECENT_RESPONSE_COUNT	...	31.4996	2
RECENT_RESPONSE_PROP	...	31.1671	2
FILE_AVG_GIFT	...	29.7171	2
LIFETIME_AVG_GIFT_AMT	...	29.7171	2
PEP_STAR	...	29.4852	2
RECENT_CARD_RESPONSE_COUNT	...	28.7487	2
LIFETIME_MAX_GIFT_AMT	...	27.164	2
FILE_CARD_GIFT	...	26.4389	2
RECENT_AVG_GIFT_AMT	...	25.8255	2
LIFETIME_GIFT_COUNT	...	24.3828	2
RECENCY_STATUS_96NK	...	20.149	2
LIFETIME_MIN_GIFT_AMT	...	15.3125	2
LIFETIME_GIFT_RANGE	...	14.2944	2
RECENT_CARD_RESPONSE_PROP	...	14.1088	2
LIFETIME_PROM	...	13.1224	2
RECENT_STAR_STATUS	...	12.9681	2
MONTHS_SINCE_LAST_GIFT	...	12.9636	2
RECENT_AVG_CARD_GIFT_AMT	...	12.7762	2
LIFETIME_CARD_PROM	...	12.3162	2
LIFETIME_GIFT_AMOUNT	...	12.042	2

b. The tree now has two additional nodes. Select the lower left node (where FREQUENCY_STATUS_97NK is 3 or 4), and then from the **Action** menu, select **Split Node**. In the Split Node window that opens, select MONTHS_SINCE_LAST_GIFT, which ranks second in logworth, and click **Edit Rule** to manually specify the split point for this rule. The Interval Split Rule window appears.

Enter `8` as the **New split point**, and click **Add Branch**. Then, select Branch 3 (>= 8.5) and click **Remove Branch**. Click **OK**.



MONTHS_SINCE_LAST_GIFT - Interval Split Rule

Target Variable: TARGET_B

Assign missing values to

● A specific branch 2

○ A separate missing values branch

○ All branches

Branches

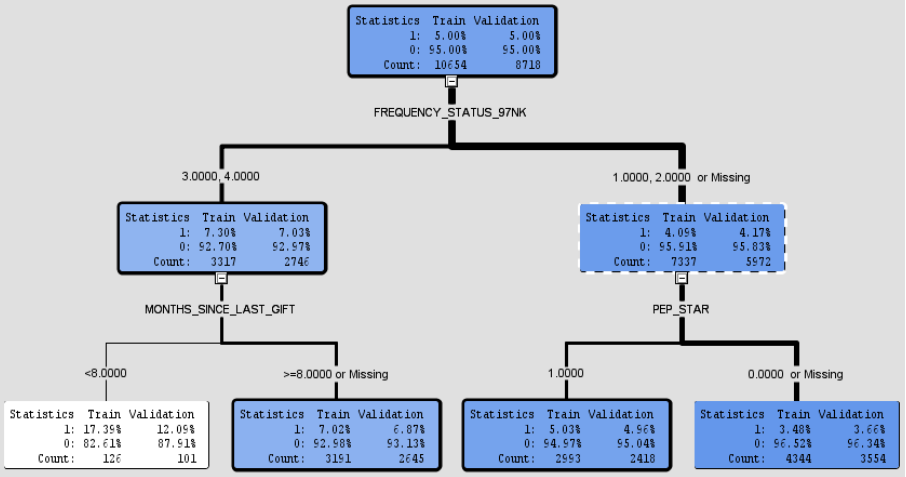
Branch		Split Point
1	<	8.0000
2	>=	8.0000

New split point:

Ensure that MONTHS_SINCE_LAST_GIFT is selected in the Split Node window, and click **OK**.

c. Select the first generation node that you have not yet split (where FREQUENCY_STATUS_97NK is 1, 2, or Missing). From the **Action** menu, select **Split Node**. In the Split Node window that opens, ensure that PEP_STAR is selected, and click **OK**.

The tree now has seven nodes, four of which are leaf nodes. The nodes are colored from light to dark, corresponding to low to high percentages of correctly classified observations.



```
Statistics  Train  Validation
        1:   5.00%      5.00%
        0:  95.00%     95.00%
    Count:  10654       8718
```

```
Statistics  Train  Validation
        1:   7.30%      7.03%
        0:  92.70%     92.97%
    Count:   3317       2746
```

```
Statistics  Train  Validation
        1:   4.09%      4.17%
        0:  95.91%     95.83%
    Count:   7337       5972
```

```
Statistics  Train  Validation
        1:  17.39%     12.09%
        0:  82.61%     87.91%
    Count:    126        101
```

```
Statistics  Train  Validation
        1:   7.02%      6.87%
        0:  92.98%     93.13%
    Count:   3191       2645
```

```
Statistics  Train  Validation
        1:   5.03%      4.96%
        0:  94.97%     95.04%
    Count:   2993       2418
```

```
Statistics  Train  Validation
        1:   3.48%      3.66%
        0:  96.52%     96.34%
    Count:   4344       3554
```

d. Select the lower right node (where FREQUENCY_STATUS_97NK is 1, 2, or Missing and PEP_STAR is 0 or Missing). From the **Action** menu, select **Train Node**. This selection causes SAS Enterprise Miner to continue adding generations of this node until a stopping criterion is met. For more information about stopping criteria for decision trees, see the SAS Enterprise Miner Help.

Note: In the Interactive Decision Tree window, you can prune decision trees. However, in this example, you will leave the tree in its current state.

e. Close the Interactive Decision Tree window.

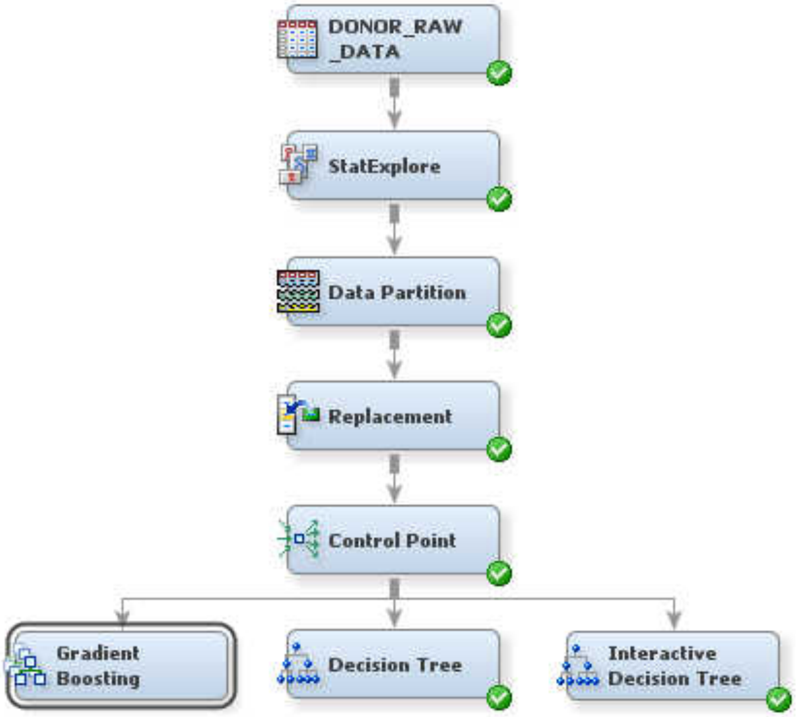
Create a Gradient Boosting Model of the Data

The Gradient Boosting node uses a partitioning algorithm to search for an optimal partition of the data for a single target variable. Gradient boosting is an approach that resamples the analysis data several times to generate results that form a weighted average of the resampled data set. Tree boosting creates a series of decision trees that form a single predictive model.

Like decision trees, boosting makes no assumptions about the distribution of the data. Boosting is less prone to overfit the data than a single decision tree. If a decision tree fits the data fairly well, then boosting often improves the fit. For more information about the Gradient Boosting node, see the SAS Enterprise Miner help documentation.

To create a gradient boosting model of the data:

1. Select the **Model** tab on the Toolbar.
2. Select the **Gradient Boosting** node icon. Drag the node into the Diagram Workspace.
3. Connect the **Control Point** node to the **Gradient Boosting** node.



4. Select the **Gradient Boosting** node. In the Properties Panel, set the following properties:

- Click on the value for the **Maximum Depth** property, in the **Splitting Rule** subgroup, and enter `10`. This property determines the number of generations in each decision tree created by the Gradient Boosting node.

- Click on the value for the **Number of Surrogate Rules** property, in the **Node** subgroup, and enter `2`. Surrogate rules are backup rules that are used in the event of missing data. For example, if your primary splitting rule sorts donors based on their ZIP codes, then a reasonable surrogate rule would sort based on the donor's city of residence.

5. In the Diagram Workspace, right-click the Gradient Boosting node, and select **Run** from the resulting menu. Click **Yes** in the Confirmation window that opens.

6. In the Run Status window, select **OK**.

TIP The book "Decision Trees for Analytics Using SAS Enterprise Miner" offers additional information about alternative measures of the effectiveness of a split, options for training and pruning, suggestions for guiding tree growth, and examples of multiple tree and gradient boosting models.

TIP "Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications" offers examples that automatically and interactively train and prune decision tree models and examples that create gradient boosting models.