ECLT5810 E-Commerce Data Mining Technique
SAS Enterprise Miner -- Decision Tree

I. Tree Node
Setting Tree Node Defaults
- define default options that you commonly use to build trees
- right-click Tree node in Project Navigator → select Edit Defaults button

There are totally 3 Tabs enabled:
1. Basic Tab
   A. Splitting criterion

   ![Splitting criterion tab](image)

   - For interval target
     - F test
     - Variance reduction
   - For ordinal target
     - Entropy reduction
     - Gini reduction
   - For binary or nominal target
     - Chi-square test
     - Entropy reduction
     - Gini reduction

   Splitting criterion: General options

B. General options

   ![General options tab](image)

   - Minimum number of observations in a leaf: 1
   - Observations required for a split search: .
   - Maximum number of branches from a node: 2
   - Maximum depth of tree: 6
   - Splitting rules saved in each node: 5
   - Surrogate rules saved in each node: 0

   Treat missing as an acceptable value

- Observations required for a split search: minimum no. of data in a node to perform split
- Maximum depth of tree: max. tree height
- Splitting rules saved in each node: no. of highest score rules stored in the node
- Surrogate rules saved in each node: if observations with missing splitting attribute are encountered, surrogate rules will be used instead of the main rules
- Treat missing as an acceptable value: whether use observations with missing splitting attribute to assess the split test or not
2. Advanced Tab  
A. Model assessment

The Tree node supports the following model assessment criteria:

<table>
<thead>
<tr>
<th>Interval Targets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Profit or loss</td>
<td>evaluates the tree using the maximum average profit or minimal average loss when Profit or Loss Matrix is defined</td>
</tr>
<tr>
<td>Average Square Error</td>
<td>evaluates the tree that has the smallest average square error between the actual class and predicted class</td>
</tr>
<tr>
<td>Average, profit, or loss in the top 10, 25, or 50%</td>
<td>this measure evaluates the tree based on the average predicted values for the target, the maximum average profit, or the minimum average loss for the top n% of the cases. You use this model assessment criterion when your overall goal is to create the tree with the best lift value</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ordinal Targets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinal Proportion correctly classified</td>
<td>evaluates the tree that has the best classification rate when weighted for the ordinal distances</td>
</tr>
<tr>
<td>Proportion of event, profit, or loss in top 10, 25, or 50%</td>
<td>evaluates the tree that results in the maximum profit or minimum loss in the top n% of the data</td>
</tr>
<tr>
<td>Total Leaf Impurity (Gini Index)</td>
<td>chooses the tree that has the greatest total leaf impurity (smallest Gini Index)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binary or Nominal Target</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion correctly classified</td>
<td>evaluates the tree based on the proportion correctly classified</td>
</tr>
<tr>
<td>Average Profit or Loss</td>
<td>(default if you defined a decision matrix) evaluates the tree based on the maximum average profit or minimal average loss</td>
</tr>
<tr>
<td>Proportion of event, profit, or loss in top 10, 25, or 50%</td>
<td>evaluates the tree that results in the maximum profit or minimum loss in the top n% of the data</td>
</tr>
<tr>
<td>Total Leaf Impurity (Gini Index)</td>
<td>chooses the tree that has the greatest total leaf impurity (smallest Gini Index)</td>
</tr>
</tbody>
</table>
B. General options

- Sub-tree: choose method to select subtree
  1. Best assessment value  
     Use model assessment criteria specified before
  2. The most leaves  
     Use the fully built tree
  3. At most indicated number of leaves  
     Use tree with \( n \) leaves

- Observations sufficient for split search: max. no. of data to perform splitting test
  - more data \( \rightarrow \) more memory and longer CPU times

- Maximum tries in an exhaustive split search: max. no. of splitting tests perform in a node

---

**Input Data Source** node:
- Data Set: SAMPSIO. DMAGECR
- Target: GOOD_BAD

**Data Partition** node
- Method: stratified use GOOD_BAD variable
- Training: 70%
- Validate: 30%

**Tree** node
- Advanced setting \( \rightarrow \) Model Assessment Measure \( \rightarrow \) Misclassification

Running the **Tree** Node
Open **Tree Results Browser**

**Model Tab**
In tree node, the Model tab contains the following sub tabs:

<table>
<thead>
<tr>
<th>Sub tab</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td>lists the model name, description, date created, last modified, and the target variable. To browse the target profile, select the <em>Profile information</em> button</td>
</tr>
<tr>
<td><strong>Basic</strong></td>
<td>Displays the <em>Basic</em> tab settings that were used to create the tree in browse mode</td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Displays the <em>Advanced</em> tab settings that were used to create the tree in browse mode</td>
</tr>
<tr>
<td><strong>Import</strong></td>
<td>This tab is grayed out if you did not import a tree from a predecessor <em>Variable Selection</em> node, <em>Tree</em> node, or <em>SAS Code</em> node. To view the properties and a table view of the import tree data set, select the <em>Properties</em> button</td>
</tr>
</tbody>
</table>
After running the tree node, you can investigate the result of the tree.

**All Tab**

**Summary Tab**
- displays a table that provides summary statistics for the currently selected tree. The information provided in the table depends on whether the target is a categorical or an interval variable and on the model assessment measure you chose in the Advanced tab.

**Summary Table for Categorical Targets**
- confusion matrix: pull-down menu View ➔ confusion matrix
- leaf statistics: pull-down menu View ➔ leaf statistics

**Tree Ring Tab**
- contains the tree ring, which represents a graphical display of the tree complexity, split balance, and discriminatory power
- controls the table view in Table Tab

Center region - root node (entire data set)
Successive ring - represent each split
Segments - represent each node
Size of segments - proportional to the no. of data in the node
Segment color - categorical ➔ assessment value
- interval target ➔ target value
Tree Ring Pop-Up Menu
- right-click the tree ring:

  Define colors: choose your own customized colors to provide better contrast between segments and/or choose a different color criterion.

  Node definition: displays the node definition (the English rule) for the selected node in the tree ring.

  Probe: displays summary information in a textbox when you click on a node (or slide your cursor over the surface) in the tree ring.

  Pick: selects a node in the tree ring. The tree diagram is automatically updated based on the node you pick in the tree ring.

  Redraw tree: redraws the tree diagram from the currently selected node of the tree ring.

Table Tab
- contains a table that provides a measure of how well the tree describes the data. Assessment statistics for each sub-tree are listed for the training and validation data sets; validation statistics are not listed if you did not use a validation data set. The type of statistic that is provided in the table depends on the model assessment measure that you chose in the Advanced tab.
- For a categorical target, the default table statistic is the proportion of observations correctly classified.
- For an interval target, the default table statistic is the average sum of squared differences of an observation from its predicted value.

Plot Tab
- displays a plot of the assessment values on the vertical axis for the different sub-trees. The Tree node automatically chooses the sub-tree that optimizes the model assessment measure that you chose in the Advanced tab. The vertical reference line identifies this sub-tree.

Set the vertical axis statistic for the assessment plot
- right-click on the plot:

  Categorical targets:
  - Proportion misclassified
  - Proportion of event in either the top 10, 25, or 50%
  - Total leaf impurity (Gini index)

  Interval targets:
  - Average square error
  - Average profit or loss
  - Proportion of event in the either the top 10, 25, or 50%

Tree Diagram
- displays node (segment) statistics, the names of variables used to split the data into nodes, and the variable values for several levels of nodes in the tree.

  View pull-down menu → Tree

A Tree diagram contains the following items:

**Root node**: top node in the tree that contains all observations

**Internal nodes**: non terminal nodes (also includes the root node) that contains the splitting rule

**Leaf nodes**: terminal nodes that contain the final classification for a set of observations

**Numeric labels**: indicate at which point the Tree node found significant splits in interval level variable distributions or in categorical splits for nominal or ordinal level distributions

**Character labels**: name of variable used to split
Right click the tree diagram and choose Statistics. Select all the statistics to display in the tree diagram.

Then each node of the tree will contain the following information.

- **Assigned Decision**
- **Percentages for each target level**
- **Counts for each target level and total counts**
- **Percentages of observations correctly classified within the node**

**Tree Diagram Pop-up Menus**
- specify tree customizations, save the tree, and print the tree
- right-click on a node:

  - **View competing splits** opens the Input Selection window, which lists the splitting rules and their measure of worth. You must first select a node in the diagram before you can view competing splits
  - **View surrogate splits** displays the surrogate splits in the Input Selection window. You must first select a node in the diagram before you can view surrogate splits
  - **View split values** displays the split values (right click on each splitting value)
  - **Define colors** define tree diagram and tree ring colors
  - **Statistics** set the node statistics
  - **Diagram-node types** specify the information displayed in the diagram, such as node statistics, or node statistics and variables
  - **Tree options** set the depth of the tree
  - **Node definition** displays a text box that contains the rule that defines the node