Chapter 8: Tree Analysis Using SAS Enterprise Miner

Introduction

In several earlier chapters, we focused on understanding the customers at DMR Publishing. Understanding your customers is always a good first step in any comprehensive analysis. In this chapter, you will learn how to discover and measure patterns to predict future behavior from a combination of past behaviors and characteristics. The main technique you will use in this process is tree analysis. One of the benefits of tree analysis is that in addition to being a powerful predictive technique, it also has a strong descriptive component. The descriptive results of a tree analysis are easily interpreted, making this technique a favorite among business analysts and marketers.

Project Overview

The leadership team at DMR Publishing Company is interested in understanding the characteristics of customers that subscribe to more than one magazine or journal when compared to customers that have only one subscription. For this analysis, you will use SAS Enterprise Miner.

This project has ten steps:

1. Initiate the project in SAS Enterprise Miner 13.1.
2. Input the data source.
3. Define the variable roles.
4. Define the target variable.
5. Partition the data.
6. Set the tree properties.
7. Build the decision tree.
8. Adjust your graph properties.
9. Interpret your findings.
10. Display Node Rules.

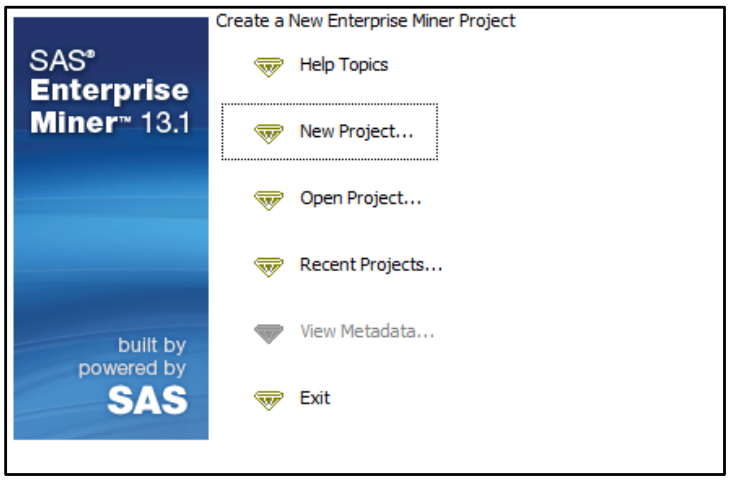
Decision Tree Analysis

The decision tree technique is designed to segment data according to series of simple rules. For a full documentation of the process, access the **Help** section of SAS Enterprise Miner. Once you have a project open, go to **Help ▶ Contents.** When the window opens, look for **Node Reference ▶ Model ▶ Decision Tree Node.**

Initiate the Project

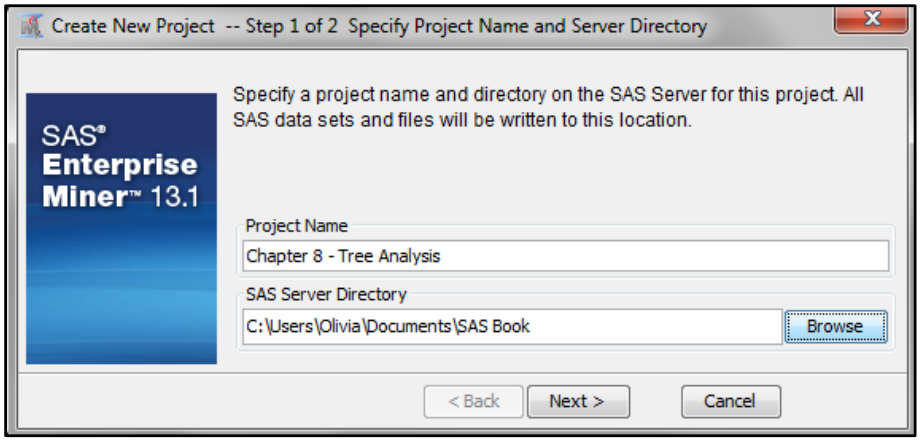
To open SAS Enterprise Miner (EM), click on the node on your desktop or Start menu. Your first choice is to open an existing project or create a new project. Highlight and click **New Project** (Figure 8.1).

Figure 8.1: Initialize SAS Enterprise Miner



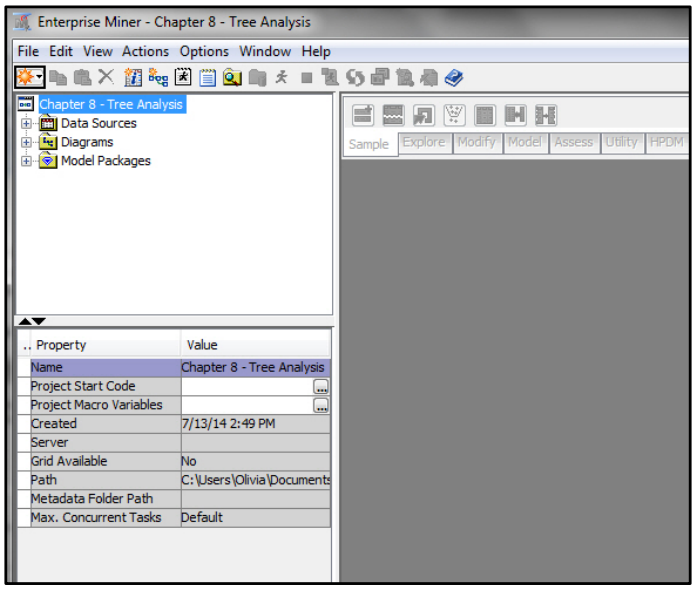
A window opens that asks you to name your project and select a SAS Server Directory (Figure 8.2). Depending on your setup, it may require additional connections. If that is the case, contact your IT department. Otherwise, click browse and select a folder where you would like to save your project files.

Figure 8.2: Create and Name New Project



When you are finished, click **Next** ▶ and you will see window that displays your choices. Click **Finish**. You are now in the EM workspace, as shown in Figure 8.3.

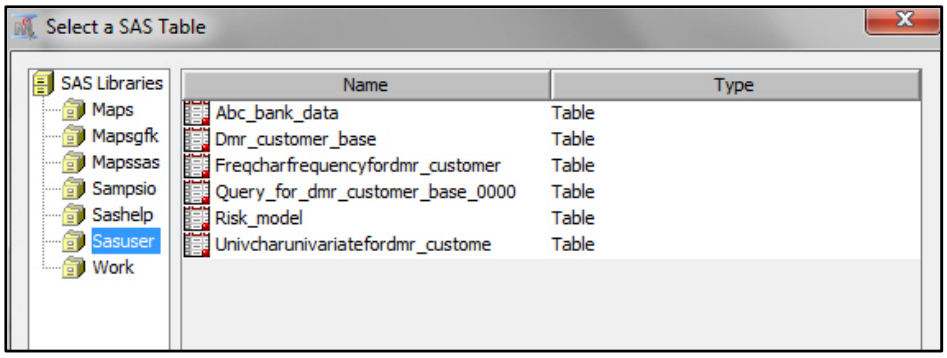
Figure 8.3: SAS Enterprise Miner Workspace



Input the Data Source

Next, click on the **Create Data Source** node (upper left menu right under the word 'Actions.'). The Data Source Wizard will open and ask you to locate a SAS Table. Click **Next** ▶ and browse in the SASUSER library for the **DMR_CUSTOMER_DATA** data set. Once you locate the data, select the data and click **OK** (Figure 8.4).

Figure 8.4: Locate and Input Data Source



Click **Next**. The next window shows you a summary. Click **Next** again and select **Advanced** to open a table that will allow you to change various aspects of the data.

Change the **Role** for CUSTOMER_REVENUE, CUSTOMER_SUBSCRIPTION_COUNT, and YEARS_AS_CUSTOMER to **Rejected** (Figure 8.5). The role for these three variables is set to **Rejected**

because they are a measure of performance with DMR Publishing. So, the tree will be built using AGE, GENDER, and HOUSEHOLD_INCOME.

Click **Next** several times until you see **Finish**. Click **Finish.** The target variable will be built using the CUSTOMER_SUBSCRIPTION_COUNT variable in a later step.

Figure 8.5: Assign Variable Roles

Name	Role	Level	Report	Order	Drop
AGE	**Input**	**Interval**	**No**		**No**
CUSTOMER_ID	**ID**	**Interval**	**No**		**No**
CUSTOMER_REVENUE	**Rejected**	**Interval**	**No**		**No**
CUSTOMER_SUBSCRIPTION_COUNT	**Rejected**	**Nominal**	**No**		**No**
GENDER	**Input**	**Nominal**	**No**		**No**
HOUSEHOLD_INCOME	**Input**	**Interval**	**No**		**No**
YEARS_AS_CUSTOMER	**Rejected**	**Interval**	**No**		**No**

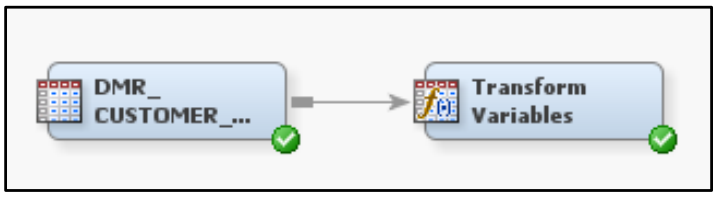
The **DMR_CUSTOMER_DATA** data now appears under **Data Sources** in your Project Tree display in the upper left corner of your workspace.

Next, you want to create a new diagram in your workspace. Go to **File ▶ New ▶ Diagram**. Name the diagram 'Tree Analysis.' Then, drag the DMR_CUSTOMER_DATA data into the new diagram. Right-click on the DMR_CUSTOMER_DATA icon and select **Update.** When it completes, click **OK.**

Create Target Variable

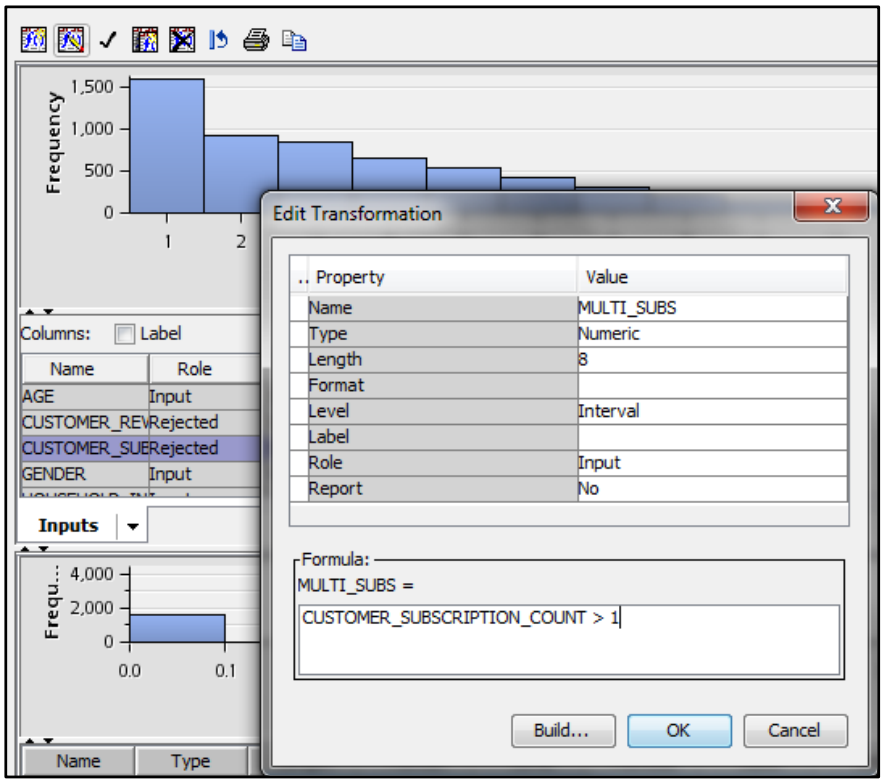
You are now ready to create the target variable. Click the **Modify** tab and select drag the **Transform Variables** icon into the diagram. Connect the DMR_CUSTOMER_DATA icon to the **Transform Variables** icon (Figure 8.6). Right-click the **Transform Variables** icon and select **Update.** When it completes, click **OK.**

Figure 8.6: Connect Transform Variables to DMR Customer Data



With the **Transform Variables** icon highlighted, go to the **Property** window in the lower left of the screen. Under **Train**, double-click on the three dots to the right of the **Formulas** to open a window that will allow you to create new variables (Figure 8.7). Highlight the variable CUSTOMER_SUBSCRIPTION_COUNTS. To create the target variable, click the Create icon in the upper left corner. This icon will open the window shown in the smaller window in Figure 8.7.

Figure 8.7: Build Target Variable Equation



In the bottom window under Formula, type CUSTOMER_SUBSCRIPTION_COUNT > 1. In the upper portion of the smaller box to the right of Name and under Value, type MULTI_SUBS as the name of the target variable. Click **OK** ▶ **OK**. The name of the new target variable is now MULTI_SUBS.

Your next step is to assign MULTI_SUBS as a target variable. Click on the **Utility** tab, drag the **Metadata** icon onto the diagram, and connect it to the **Transform Variables** icon (Figure 8.8).

Figure 8.8: Attach Metadata to Transform Variables

Right-click on the **Metadata** icon and select **Update.** When it completes, click **OK.**

Next, highlight the **Metadata** icon and go to the **Property** window in the lower left corner. Click on the three dots under **Variables** to the right of **Train**. The window shown in Figure 8.9 opens. Under **New Role**, change the role for MULTI_SUBS to Target. Click **OK**. Right-click on the **Metadata** icon and select **Run**. When it completes, click **OK**.

Figure 8.9: Assign Target Variable Role Using Metadata

Name	Hidden	Hide	Role	New Role
AGE	N	**Default**	Input	**Default**
CUSTOMER_ID	N	**Default**	ID	**Default**
CUSTOMER_REV	N	**Default**	Rejected	**Default**
CUSTOMER_SUB	Y	**Default**	Rejected	**Default**
GENDER	N	**Default**	Input	**Default**
HOUSEHOLD_IN	N	**Default**	Input	**Default**
MULTI_SUBS	N	**Default**	Input	**Target**
YEARS_AS_CUS	N	**Default**	Rejected	**Default**

Your next step is to partition the data.

Partition the Data

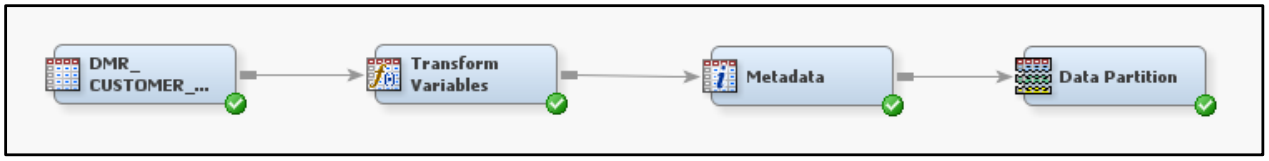
The purpose of partitioning the data is to allow you to develop the model on one subset of the data and validate the model on another subset of the data. The subsets of the data are mutually exclusive in that they do not share any common observations.

In decision tree modeling, the subsets are used as follows:

- *Train:* The train subset for the initial model fitting.
- *Validation:* The validation data set is used to evaluate the model during the tree building process and create the best subtree.

From the **Sample** tab, select the second icon from the left, **Data Partition,** and drag it onto the diagram. Connect it to the **Metadata** icon (Figure 8.10).

Figure 8.10: Attach Data Partition to Metadata Icon



Within the **Property** window in the bottom left, look for **Data Set Allocations** under **Train.** The default shows 40% of the file allocated to **Training** and 30% each to **Validation** and **Testing**. For tree models, you only need **Training** and **Validation**. So, set **Training** and **Validation** to 50% each, and **Test** to 0% (Figure 8.11).

Figure 8.11: Reassign Data Allocation Percentages

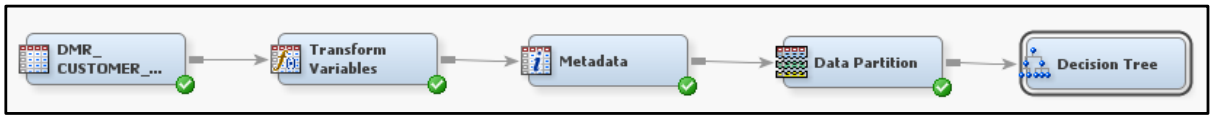
.. Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	No
Class Targets	Yes
Status	
Create Time	7/26/14 7:17 AM
Run ID	ac584a31-9599-480f-84b8-3
Last Error	
Last Status	Complete
Last Run Time	7/26/14 7:20 AM
Run Duration	0 Hr. 0 Min. 1.90 Sec.
Grid Host	
User-Added Node	No

Right-click on the **Data Partition** icon and select **Update.** When it completes, click **OK.** You are now ready to build your tree.

Build the Decision Tree

In the **Model** menu above the diagram, drag the **Decision Tree** icon (second from left) into the workspace, and connect it to the **Data Partition** icon (Figure 8.12).

Figure 8.12: Connect Decision Tree Icon to Metadata Icon

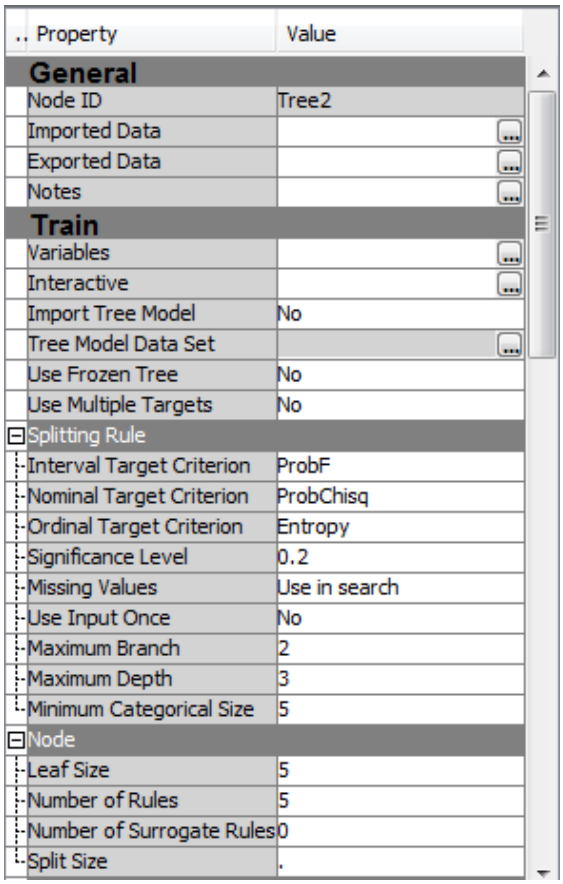


To update the path to the decision tree node, right-click on it and select **Update**.

Highlight the **Decision Tree** node and go to the lower left menu (Figure 8.13). The defaults are good, but you might want to consider several settings. For each, click the three dots to the right of your choice:

- *Variables:* This setting is a chance to suppress variables from the decision tree. To do so, go to the **Use** column and change the value to **No**. This setting is not necessary for your current tree, but it is a way to explore other options.
- *Interactive*: This setting enables you to hand-build your tree by allowing you to create splits and prune your tree on the fly. It is beyond the scope of this book. But once you've mastered the basic steps, you might want to explore some of these options.
- *Maximum Branch:* Although this setting remains at 2 for this demonstration, in practice, one might typically use 4 or 5.
- *Maximum Depth:* Normally, you might leave this setting at 6; it is set to 3 for this demonstration.

Figure 8.13: Assign Decision Tree Settings



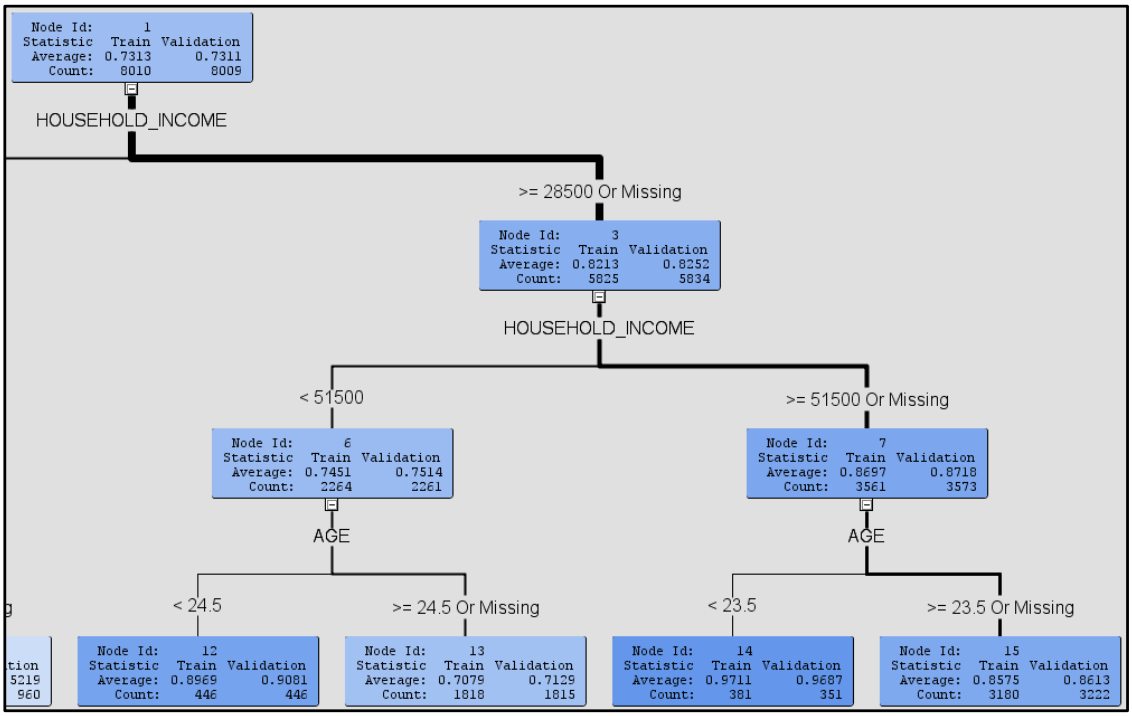
Property	Value
General	
Node ID	Tree2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	3
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

Right-click the decision tree node and click **Run** and **Yes**. When the **Run Completed** box opens, click **Results**.

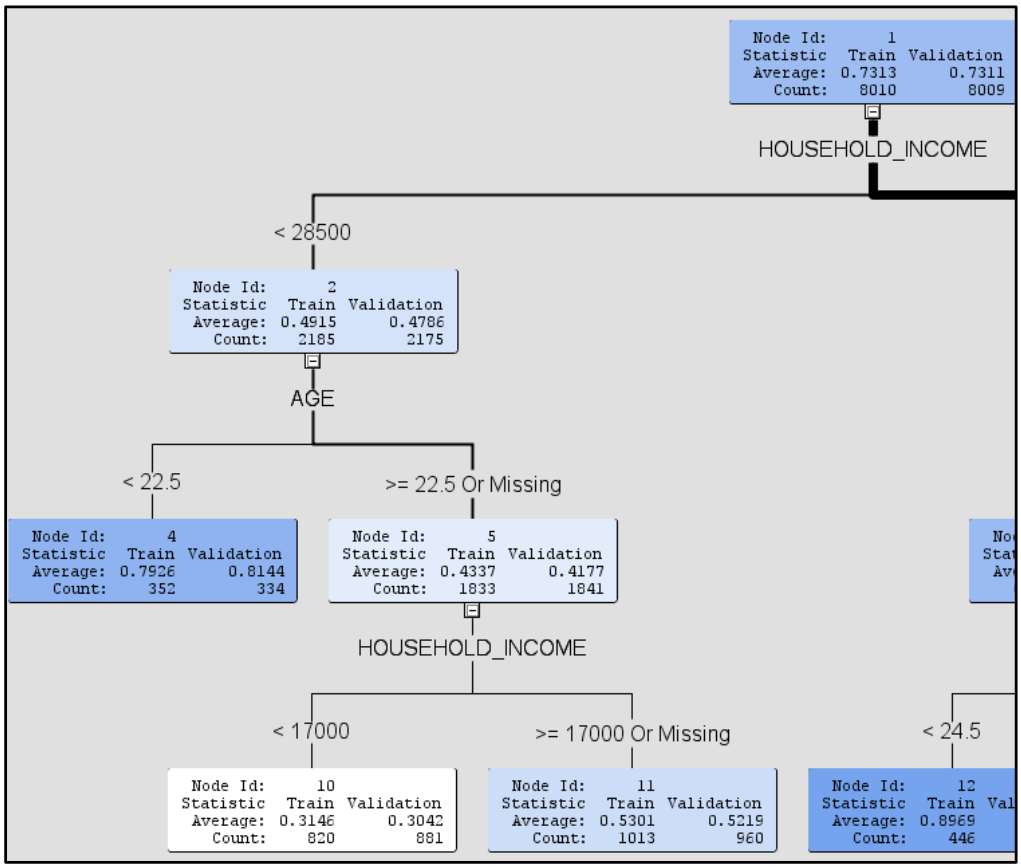
View the Decision Tree Output

The first window contains six panels. Maximize the panel in the upper right corner to display the tree shown in Outputs 8.1 and 8.2. The tree is clearly visible and you can view it in its entirety by scrolling to the left and right.

Output 8.1: View Initial Tree (Left Branch)



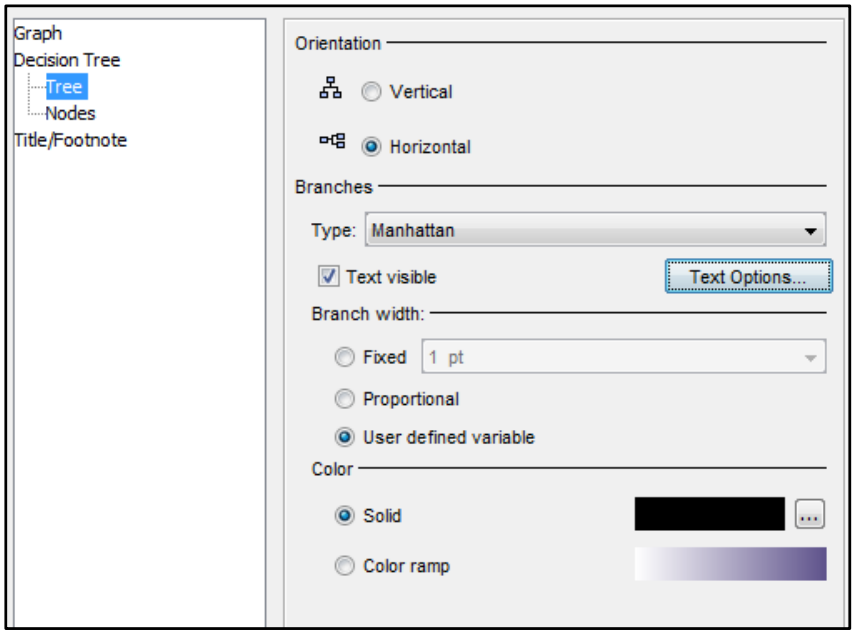
Output 8.2: View Initial Tree (Right Branch)



Graph Properties

Before you begin analyzing the results, you have some options for changing the view in accordance with your personal preferences. To explore the many options within the tree output, right-click anywhere on the screen shown in Output 8.1 and select **Graph Properties**. When the window opens, click **Tree.** Change the **Orientation** to **Horizontal,** and click **OK** (Figure 8.14).

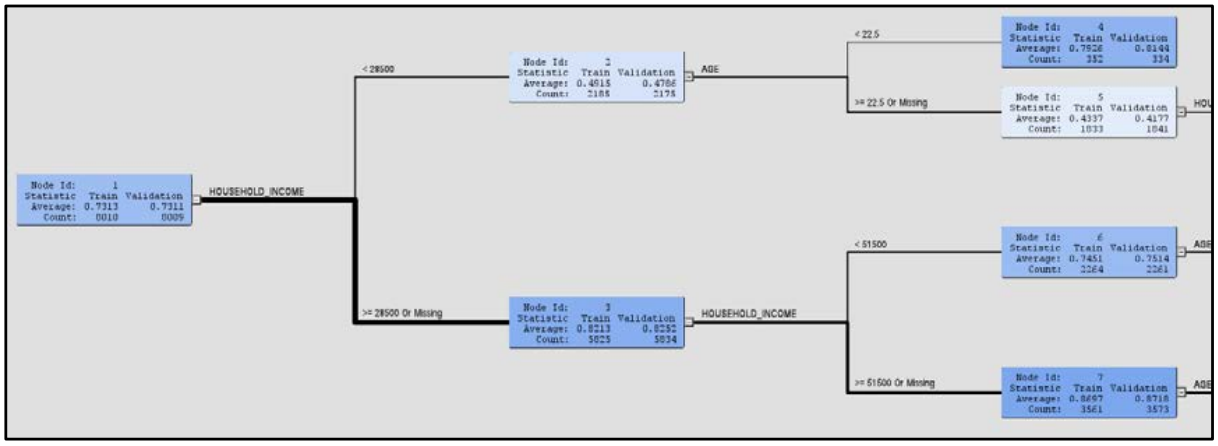
Figure 8.14: Change Graph Properties for Decision Tree Output



You can also click **Text Options** and change the font to **Arial Narrow** and the size to **10 pt**. Doing so is useful if the variable names are long. Click **Apply ▶ OK ▶ OK.** The tree is now in a horizontal view**.**

The horizontal view of the tree is shown in Output 8.3. The tree results read from left to right. Each rectangular box is called a *node*. Each node location contains information about the group of DMR customers within that node.

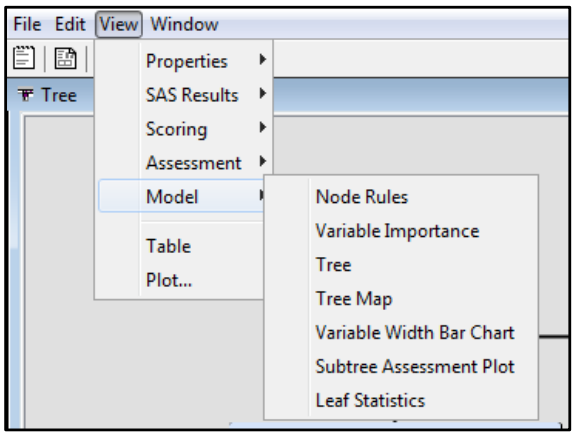
Output 8.3: View Horizontal Tree



Diagnostics

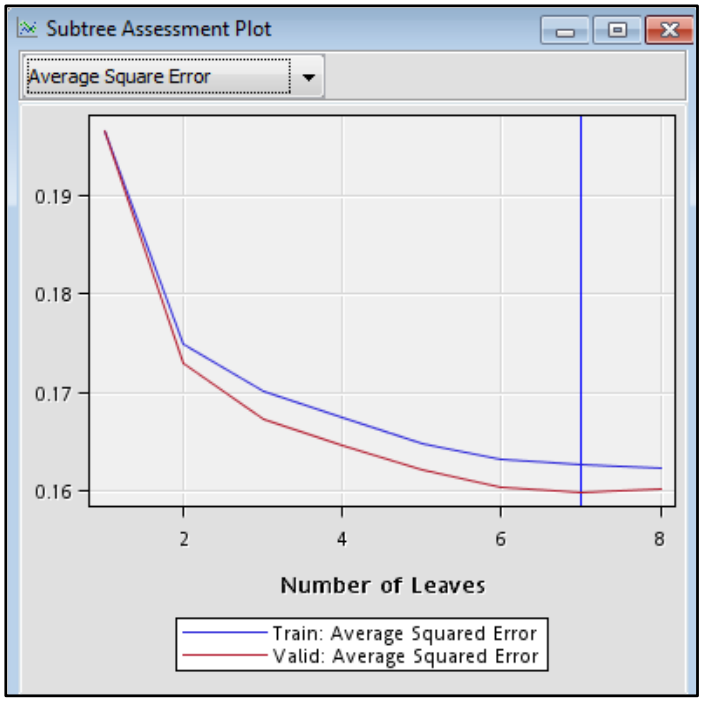
With all predictive models, there is a risk of over-fitting. In other words, your model could fit your data so well that it does not work well on new data when you implement the model. You can check for over-fitting by viewing the Subtree Sequence Plot. To view the Subtree Sequence Plot, go to the upper left menu and select **View ▶ Model ▶ Subtree Assessment Plot** (Figure 8.15).

Figure 8.15: Access Subtree Assessment Plot



The Subtree Assessment Plot shows that the Average Squared Error continues to decrease as the number of leaves increases (Output 8.4).

Output 8.4: View Subtree Assessment Plot



If the error begins to increase, you should limit the number of leaves or prune the tree. For more information on pruning decision trees, see *Decision Trees for Analytics Using SAS® Enterprise Miner*, by Barry de Ville and Padraic Neville (2013).

Node Characteristics

Three dimensions of the tree output help you visually interpret the results:

- *Shading:* When the shading is darker, the percentage of the target group MULTI_SUBS > 1 is higher.
- *Line Width:* The width or thickness of the connecting lines represents the volume of records going to the node.
- *Node Values:* Each node displays values for the records for both the model and validation data subsets within that node.
    - *Node Id:* This simple identifier makes it easy to identify the exact node.
    - *Percentage:* The percentage of the target variable represents DMR customers who have more than one subscription.
    - *Count:* The count is the number of DMR customers in that node.

Node Interpretation

In Output 8.5, the topmost node on the left represents the whole sample of 8,010 DMR customers. Both the Train and Validation data subsets show that approximately 73% of DMR customers have more than a single subscription (MULTI_SUBS > 1) as shown in Output 8.5.

The first-level split is on HOUSEHOLD INCOME at $28,500. The upper box (Node 2) is lighter and represents 2,185 DMR customers with income less than $28,500. Within Node 2, 47.7% of DMR customers have more than one subscription. The lower box (Node 3) is darker and represents customers with annual income greater than $28,500. The percentage of DMR customers with more than a single subscription (MULTI_SUBS > 1) is 82.5%.

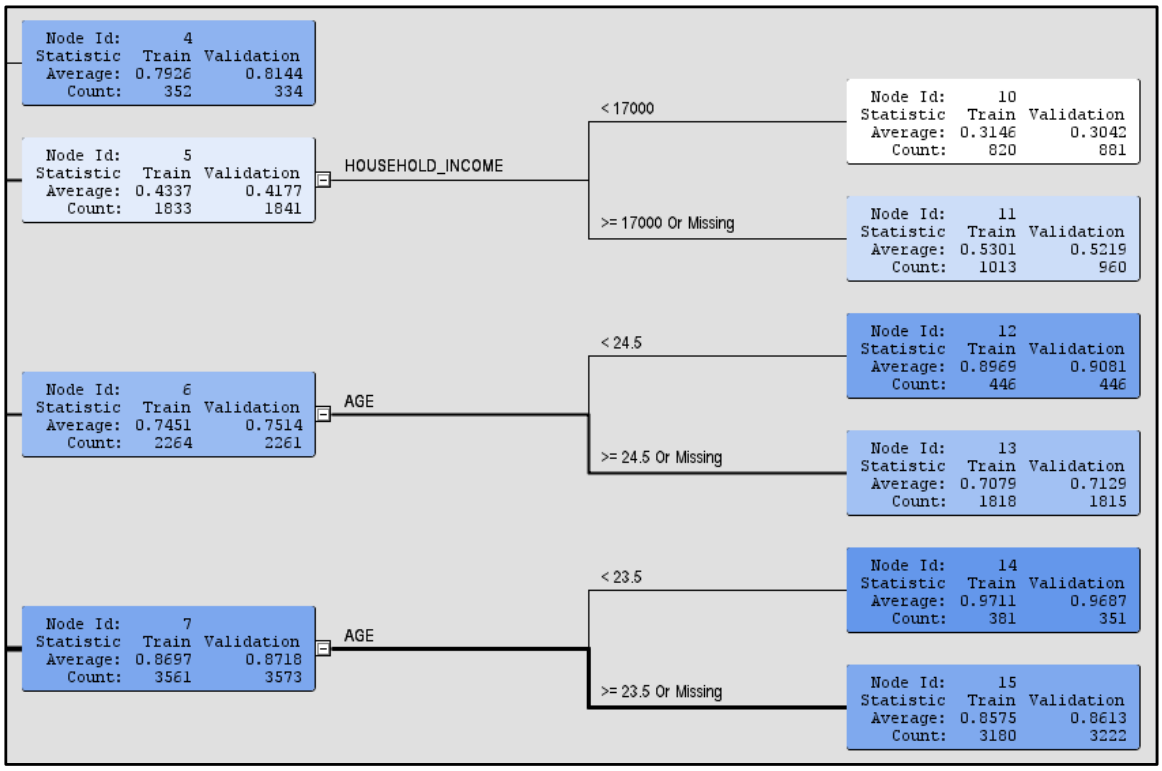
The second-level splits on a different variable for each node in the first level. Node 2 splits on AGE into Node 4 and Node 5. Node 4 shows that 81.4% of DMR customers younger than 22.5 years have more than one subscription. Node 5 shows that only 41.8% of DMR customers who are older than 22.5 years have more than one subscription.

The second split on the second level (Node 3) is on Household Income. Node 6 shows that 75.1% of DMR Customers with less than $51,500 in annual income have more than one subscription. Node 7 shows that 87.2% of DMR Customers with more than $51,500 in income have more than one subscription.

In Output 8.4, the final nodes show a wide range of percentages for DMR Customers having more than one subscription. Node 10 is the lightest background and has the lowest percentage of DMR Customers with more than one subscription (30.0%). Nodes 12 and 14 are the darkest, with the highest percentages of DMR Customers with more than one subscription (90.8% and 97.9%, respectively).

NOTE: “Or Missing” shows up in the bottom brackets whether or not you have missing values. In the current data, you have no missing values, so you can ignore “Or Missing” in this interpretation.

Output 8.5: View Horizontal Tree End Points



Interpret the Findings

The results of this analysis offer meaningful insights into the demographic characteristics of DMR customers with more than one subscription. If you consider DMR Customers with multiple subscriptions to be your best customers, there are several approaches you can take:

- *Buy more prospect names:* You may decide to buy names from a list company with the same demographics as in the highest performing nodes. So for example, Node 12 and Node 14 identify the best performing customers. Nodes 13 and 15 identify reasonably high performing customers. So, you would look to purchase names that have the characteristics of Nodes 13 and 15.

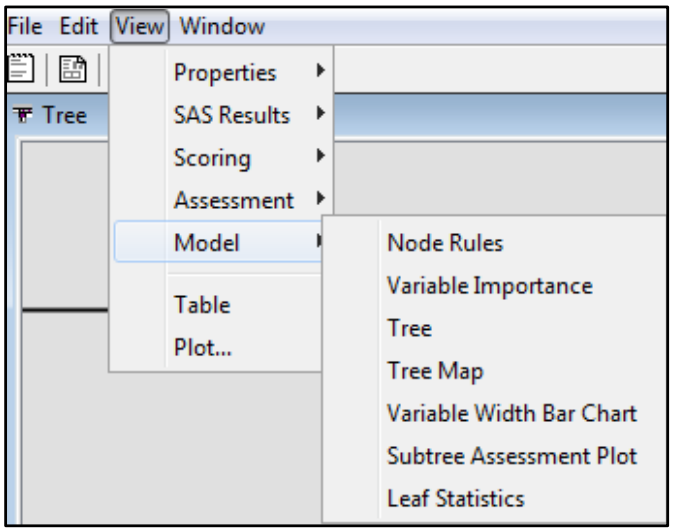
- *Target your worst customers:* Based on the known demographics of the worst performing customers, you might want to create additional products that would meet their needs. Node 10 represents a group of customers that show untapped potential. Developing and testing new products might prove to be a good business decision.

- *Leverage your best customers:* These customers are already doing a lot of business with you. A simple survey of their other interests or preferred delivery channels may open new opportunities to increase profits within this customer group.

Your next question may be, "How can I precisely identify the customers in these nodes?" Understanding how to define your subscribers using the **Node Rules** is the first step.

The **Node Rules** define each node in simple language based on the characteristics of the customers within a specific node. Within the Results window, go to **View ▶ Model ▶ Node Rules** (Figure 8.16).

Figure 8.16: Access and View Node Rules



In Output 8.6, the **Node Rules** for Node 14 are shown. Node 14 has the highest percentage of DMR multi-subscribers.

Output 8.6: View Node Rules—Highest Performing Node

```
51   *-------------------------------------------------------------*
52    Node = 14
53   *-------------------------------------------------------------*
54   if HOUSEHOLD_INCOME >= 51500 or MISSING
55   AND AGE < 23.5
56   then
57    Tree Node Identifier   = 14
58    Number of Observations = 381
59    Predicted: MULTI_SUBS = 0.9711286089
60
```

Node 14 customers, with a 97.1% rate of multiple subscriptions, can be described as having an annual HOUSEHOLD_INCOME greater than $51,500 and being younger than 23.5 years. This group of customers seems to represent young men and women in affluent households. Using these very specific demographic characteristics, you can purchase more prospect names that have similar characteristics.

Another view of the nodes with the lowest multiple-subscriber rate can give you an idea of how to grow your business. In Output 8.7, you can learn the characteristics of these low-performing groups.