Assignment 5 - SAS #2 – Decision Trees

Getting Data Ready in SAS

1. Initial Data Exploration

A supermarket is offering a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The **ORGANICS** data set contains 13 variables and over 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels:

Name	Model Role	Measurement Level	Description
ID	ID	Nominal	Customer loyalty identification number
DemAffl	Input	Interval	Affluence grade on a scale from 1 to 30
DemAge	Input	Interval	Age, in years
DemCluster	Rejected	Nominal	Type of residential neighborhood
DemClusterGroup	Input	Nominal	Neighborhood group
DemGender	Input	Nominal	M = male, F = female, U = unknown
DemRegion	Input	Nominal	Geographic region
DemTVReg	Input	Nominal	Television region
PromClass	Input	Nominal	Loyalty status: tin, silver, gold, or platinum
PromSpend	Input	Interval	Total amount spent
PromTime	Input	Interval	Time as loyalty card member
TargetBuy	Target	Binary	Organics purchased? 1 = Yes, 0 = No
TargetAmt	Rejected	Interval	Number of organic products purchased

✎ Although two target variables are listed, this exercise concentrates on the binary variable **TargetBuy**.

a. Create a new diagram named **Organics**.

b. Define the data set **AAEM61.ORGANICS** as a data source for the project.

1) Set the roles for the analysis variables as shown above.
(You can go back and modify variable roles even after you complete the wizard by right-clicking on the **Organics** data source and selecting **Edit Variables…**)

The variable **DemClusterGroup** contains collapsed levels of the variable **DemCluster**. Presume that, based on previous experience you believe that **DemClusterGroup** is sufficient for this type of modeling effort. Set the model role for **DemCluster** to Rejected.

2) Examine the distribution of the target variable **TargetBuy**. You can do this by clicking on that variable in the Column Metadata (step 6 of 9 in the wizard) and then clicking the **Explore** button.

What is the proportion of individuals who purchased organic products (hint: take a look at the "Sample Statistics" window)?

3) Finish the **Organics** data source definition.

c. Add the **AAEM61.ORGANICS** data source to the **Organics** diagram workspace.

Decision Trees

You'll be working on the project you just. Remember, this project used the "Organics" data set. When you open SAS Enterprise Miner, you should be able to find your work under the File/Recent Projects. If you can't find it there, go to File/Open Projects… and search for your project.

Create a Decision Tree based on the Organics Data Set

2. Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 50% of the data for training and 50% for validation. Run it.

Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.

3. Create a decision tree model autonomously (i.e., just run the Decision Tree node).

Answer the two questions below and attach the screenshot(s) in your solution document where you found the answer.

1) How many leaves are in the optimal tree? _____

2) Which variable was used for the first split? _____

4. Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.

In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to **3** to allow for three-way splits.

Create a decision tree model using average square error as the model assessment statistic.

Answer the two questions below and attach the screenshot(s) in your solution document where you found the answer.

1) How many leaves are in the optimal tree? _____
   (HINT: In your iteration plot, you can click near the "Number of leaves" label and drag right to zoom in)

2) Based on average square error, which of the decision two tree models appears to be better (the first one or the second one)? _____

<<GO TO THE NEXT PAGE!!>>

2. Start at the top of the decision tree and work your way downwards to answer the following questions (you don't need to include screen shots for these questions – just provide the answer):

Question	**Answer**
What is the probability that a 33 year old man with affluence grade 10 buys Organics?	
What is the probability that a 23 year old woman with affluence grade 3 buys Organics?	
What is the probability that a 55 year old man with affluence grade 6 buys Organics?	
What is the probability that a 64 year old woman with affluence grade 4 buys Organics?	
What is the probability that a 65 year old woman with affluence grade 20 buys Organics?	