**Carnegie Mellon University, Heinz College**

**Business Intelligence & Data Mining with SAS Suite**

**Fall 2012 (94832 Mini 1)**

**Assignment 2 - Predictive Modeling**

**(Due 9/27/2012)**

**A supermarket is offering a new line of organic products. The supermarket’s manager wants to determine which customers are likely to purchase these products in order to conduct future promotions. He hires you as his business intelligence consultant (with a good salary) to help solve this problem.**

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of their loyalty program participants and collected data that includes whether or not these customers purchased any of the organic products. The **ORGANICS** data set contains over 22,000 observations and 18 variables. The variables in the data set are shown below with the appropriate roles and levels. Note that although two target variables are listed, this analysis concentrates on only the binary variable **ORGYN**.

****

**Questions:**

**(Note: It is recommended that you provide screen shots of your analysis along with the discussion.)**

**1. Initial Data Exploration**

**a.** Create a new diagram named **Organics**. Define the data set **AAEM.ORGANICS (Metadata Repository--> Shared Data --> Libraries --> AAEMSPEL--> Organics)** as a data source for the project.

**b.** Set the model role for the target variable and examine the distribution of the variable. What is the proportion of individuals who purchased organic products?

**c.** Are there any variables that should not be included as input variables in your analysis?

**(**The variables **AGE**, **AGEGRP1**, and **AGEGRP2** are all different measurements for the same information. Presume that, **AGE** should be used for this type of modeling. Set the model role for **AGEGRP1** and **AGEGRP2** to Rejected.)

**(**The variable **NGROUP** contains collapsed levels of the variable **NEIGHBORHOOD**. Therefore, only

one of these variables should be used in a model. Presume that, **NGROUP** is sufficient for this type of modeling effort. Set the model role for **NEIGHBORHOOD** to Rejected.)

**(**The variables **LCDATE** and **LTIME** essentially measure the same thing. Set the model role for **LCDATE** to Rejected, retaining the variable **LTIME** as an input variable.)

(The variable **ORGANICS** contains information that would not be known at the time that you develop a model to predict the purchase of organic products. Set the model role for **ORGANICS** to Rejected.)

**d.** Add the **AAEM.ORGANICS** data source to the diagram workspace. Add a Data Partition node to the diagram and connect it to the Data Source node. Assign 70% of the data for training and 30% for validation.

**2. Predictive Modeling Using Decision Trees**

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

**b.** Create a decision tree model either interactively, automatically, or autonomously. Prune the tree using average square error. Create the tree and examine the tree results. How many leaves are in the pruned tree? Explain why you decide to prune the tree this way.

**e.** View the tree. Which variable was used for the first split? What were the competing splits for this first split? Why? Which variables were important in growing this tree? How would you interpret your results to the supermarket manager?

**f.** 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. Again use average squared error to prune the tree. Run the diagram from the new Decision Tree node and examine the tree results. How many leaves are in the pruned tree?Which variables were important in growing this tree?

**j.** View the tree. Which variable was used for the first split? Based on average square error, which of the decision tree models appears to be better, binary or three-way? Explain your reason.