

Assignment #2

Due Date: 11/11/2009 (Wed)

The Data

A supermarket is beginning to offer a line of organic products. The supermarket's management would like 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 their loyalty program participants and have now collected data that includes whether or not these customers have purchased any of the organic product.

The data set ORGANICS contains 22,000 observations and 17 variables. The variables in the data set are shown below.

Name	Model Role	Measurement	Variable Label
CUSTID	id	nominal	Customer Loyalty ID
GENDER	input	nominal	M=male, F=female, U=unknown
DOB	rejected	interval	Date of Birth
EDATE	rejected	unary	Date data taken from data base
AGE	input	interval	Age, in years
AGEGRP1	input	nominal	Age Group 1
AGEGRP2	input	nominal	Age Group 2
TV_REG	input	nominal	TV Region
NGROUP	input	nominal	Neighborhood Group
NEIGHBORHOOD	input	nominal	Type of Residential Neighborhood
LCDATE	rejected	interval	Loyalty card application date
ORGANICS	input	ordinal	Number of Organic Products Purchased
BILL	input	interval	Total Amount Spent
REGION	input	nominal	Geographic Region
CLASS	input	nominal	Customer Loyalty Status
ORGYN	input	binary	Organics Purchased?
AFFL	input	interval	Affluence grade
LTIME	input	interval	Years as Loyalty Card Member

There are two target variables:

- ORGANICS – number of organic products purchased.
- ORGYN – Organics purchased? 1 = yes, 0 = no.

However, we are mainly interested in ORGYN.

Initial Data Exploration

1. Set the model role for the target variable and examine its distribution. What is the proportion of individuals who purchased organic products?
2. The variables AGE, AGEGRP1, and AGEGRP2 are all different measurements for the same information. Presume that, based on previous experience, you know that

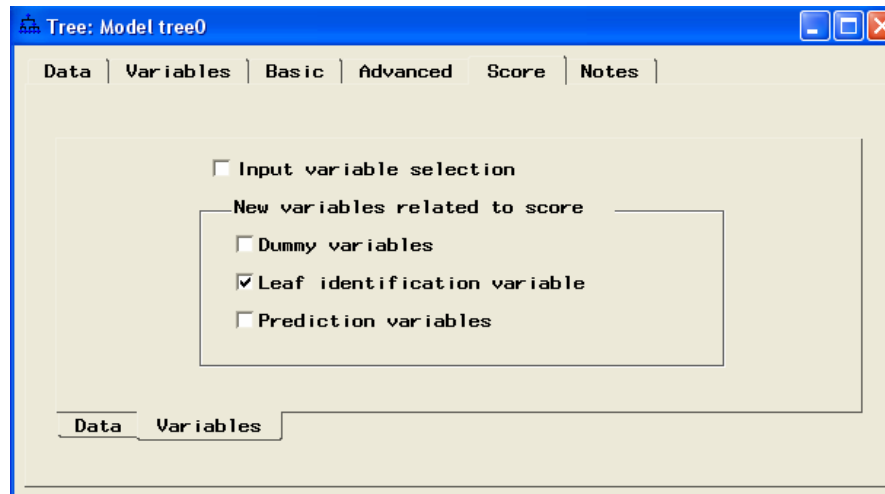
AGE should be used for this types of modeling. Set the model role for AGEGRP1 and AGEGRP2 to **rejected**.

3. 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.
4. Is there missing involved? Which variable has the heaviest missing rate?

Collapsing the Levels of NEIGHBORHOOD Using Trees

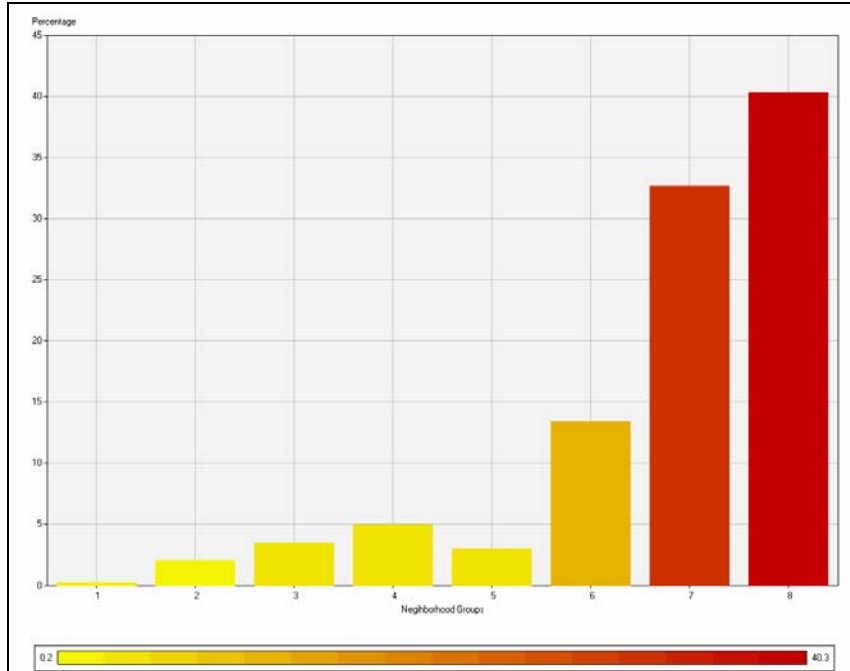
(Note that this part can be also done in Variable Transform node.)

1. The variable NEIGHBORHOOD contains information about the type of residential neighborhood. Therefore, what measurement scale should we assign?
2. It has 55 levels. To facilitate subsequent modeling, it is desirable to collapse its levels. Explain why this action is also desirable if we are going to build a tree model.
3. Suppose that, based on expert's suggestion, we want to collapse NEIGHBORHOOD into **8** levels only using decision trees.
 - Target: ORGYN
 - Use Entropy as splitting criterion.
 - Score tab => check the Process or Score Training, validation and Test.
 - Score tab => Variable => uncheck the Input Variable Selection option. (Why do we need to do so?)

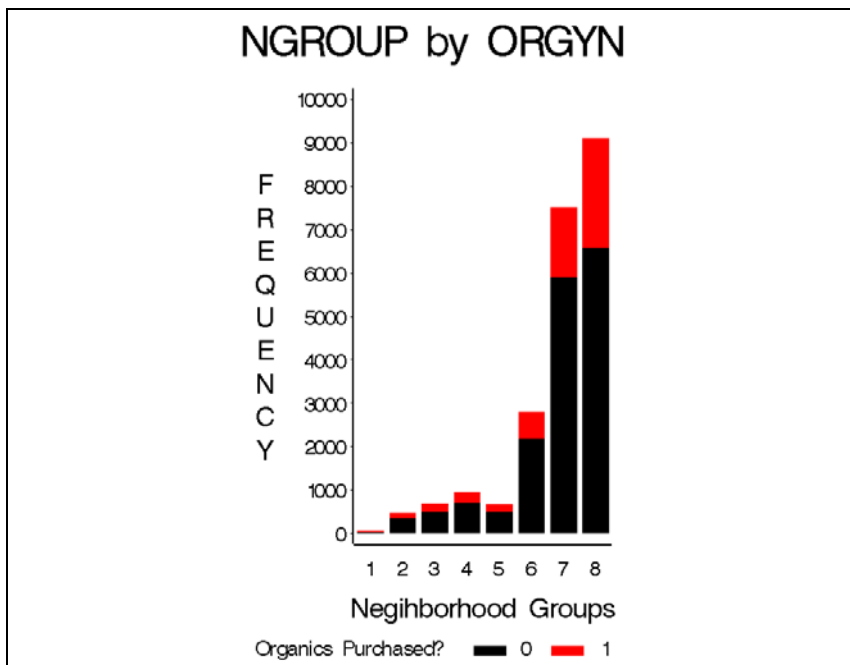


What other actions are necessary in order to get a final tree structure with exactly 8 terminal nodes?

4. Add a Transform Variable node. Let NGROUP denote the new variable that contains collapsed levels of the variable NEIGHBORHOOD. Change the measurement scale of NGROUP as nominal and format it as \$2. (character variables with 2 letter length). Label it as "Neighborhood Group". Then get a snapshot as below.



5. If we are solely interested in tree modeling, do we need to do any further variable transformation on inputs? Why?
6. Add a Multiplot node and explore the distribution of ORGYN at each level of *class* inputs: GENDER, TV_REG, REGION, CLASS, NGROUP. Which of these class inputs do you think are highly related to ORGYN? Base your answers on plots as below. Note that if you use stratified random sampling, then these plots may give you some idea on which variables should be used to stratify the data.



Partition the Data

1. Add a Data Partition node to the diagram and connect it to the Transform Variables node. Assign 40% of the data for training, 30% for validation, and 30% for test. Explain why this step is necessary if we are about to fit a tree model.

Decision Tree I

1. Add a Tree node to the workspace and connect it to the Data Partition node. Suppose now we want to develop a tree model that
 - a. Has binary splits only;
 - b. Favors balanced splits;
 - c. Has maximum depth 10;
 - d. Provides variable importance ranking;
 - e. Selects the best tree size based on minimum misclassification rate;
 - f. Treats missing values as acceptable values

What options should you choose? Provides snapshots of the **Variables, Basic, Advanced** tabs with your selections. You might want to be careful with the usage of input variables. There should be only 9 inputs.

Name	Status	Model Role	Measurement
NGROUP	use	input	nominal
ORGYN	use	target	binary
LTIME	use	input	interval
AFFL	use	input	interval
CLASS	use	input	nominal
REGION	use	input	nominal
BILL	use	input	interval
TV_REG	use	input	nominal
AGE	use	input	interval
GENDER	use	input	nominal
LEAF	don't use	input	ordinal
NODE	don't use	input	ordinal
ORGANICS	don't use	target	ordinal
LCDATE	don't use	rejected	interval
EDATE	don't use	rejected	unary
DOB	don't use	rejected	interval
NEIGHBORHOOD	don't use	input	nominal
AGEGRP2	don't use	rejected	nominal
AGEGRP1	don't use	rejected	nominal

2. Examine the tree results.
 - a. How many leaves are in the tree that is selected based on the validation data set?
 - b. According to the selected tree model, how many 1's are misclassified as 0's? And what's the percentage of false positive errors?
 - c. Examine the tree ring. Define the color so that it corresponds to the proportion of 1's. Get a snapshot of the tree ring.

- d. What is the percentage of 1's in node 5 (see tree ring)? Obtain the rules that lead to node 5.
 - e. Which inputs are selected according to its importance ranking?
3. Explain the main idea of 1-SE. Which subtree would you select applying 1-SE?
 4. View the tree. Which variable was used for the first split? What were the competing and surrogate splits for the first split?
 5. Suppose that we would like to view the tree structure in the following manner:
 - has 5 depths;
 - display splitting variable only for internal nodes;
 - display node information for terminals (leaves) only;
 - within each terminal node, suppress the percentage and frequency of responses with values 0.

What actions should be taken in order to achieve the desired view?

Decision Tree II – 3-Way Splits

1. Add another Tree node to the workspace and connect it to the Data Partition node. Suppose we would like to grow a tree that wallows for 3-way splits. Make appropriate selections and remark on why Gini or entropy based goodness-of-split criteria can not be used to evaluate 3-way split.
2. Examine the tree results.
 - a. How many leaves are there in the tree that is selected based on the validation data set?
 - b. What inputs are selected according to their importance ranking?

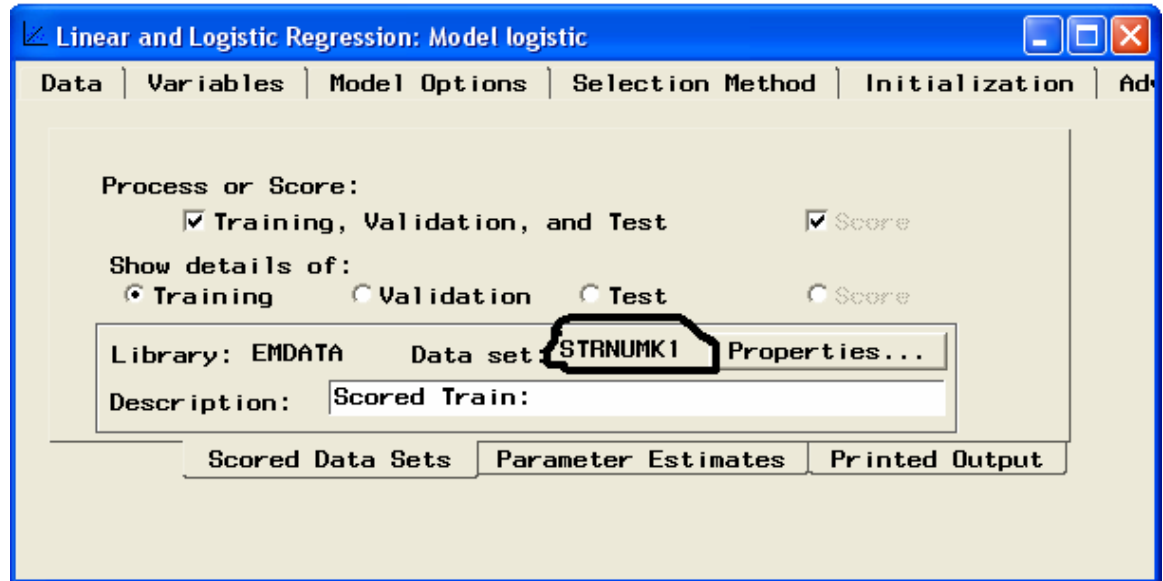
Decision Tree III – Interactive

1. Add another Tree node to the workspace and connect it to the Data Partition node. We would like to grow a more interpretable tree that
 - Has 3-way splits;
 - Has first split induced by GENDER.
 - Has all the second-depth splits induced by AGE.
 - Has all the rest nodes trained by SAS EM.
 - Applies 1-SE to select the best tree size.

Logistic Regression Model

Add a logistic regression node.

- Use stepwise selection method
- Validation Misclassification as Criteria.
- Check the Process or Score – Training, Validation, and Test option in the Output tab. Record the data set name for scored train, e.g., EMDATA.STRNUMK1. This part is useful for the following diagnostic task.



- View the results. Which variables are selected in the final logistic regression model?

Assessment

Add an Assessment node and connect it to the three tree models that we have built. Run it and then view results.

- According to the **Test** data, which model has the best **Root ASE** and the smallest **Misclassification Rate**.
- Draw the lift chart and compare the four models.

Yet Another Auxiliary Use of Trees

Tree methods can also be used to assess the goodness-of-fit of the fitted logistic regression model. The main idea is that if the logistic regression provides a good fit, then its residuals should have no structural pattern. When we try to develop a tree structure, it would lead to a tree structure containing the root node only. On the other hand, if the fit is not good, a tree model on the residuals would discover some additional patterns that may have been missed by the regression model.

1. Add an Input Data Source node. And then open the scored train data by the logistic regression. This data set is available in the EMDATA library - EMDATA.STRNUMK1.
2. Set R_ORGYN1 as the target. And carefully include appropriate input variables. You may include the predicted value as input. Recall that the plot of residuals vs predicted values is commonly used for diagnostic purpose.

Name	Model Role	Measurement	Type	Format	Inform
CUSTID	id	nominal	char	\$10.	\$10.
P_ORGYN1	input	interval	num	BEST12.	12.
GENDER	input	nominal	char	\$1.	\$1.
AGE	input	interval	num	BEST12.	12.
TV_REG	input	nominal	char	\$12.	\$12.
BILL	input	interval	num	BEST12.	12.
REGION	input	nominal	char	\$10.	\$10.
CLASS	input	nominal	char	\$8.	\$8.
AFFL	input	interval	num	BEST12.	12.
LTIME	input	interval	num	BEST12.	12.
NGROUP	input	nominal	char	\$2.	\$2.
AGEGRP1	rejected	nominal	char	\$5.	\$5.
AGEGRP2	rejected	nominal	char	\$5.	\$5.
NEIGHBORHOOD	rejected	nominal	char	\$2.	\$2.
NODE	rejected	ordinal	num	BEST12.	12.
LEAF	rejected	ordinal	num	BEST12.	12.
DOB	rejected	interval	date	DDMMYY8.	DDMMYY
EDATE	rejected	unary	date	DDMMYY8.	DDMMYY
LCDATE	rejected	interval	date	DDMMYY8.	DDMMYY
WARN	rejected	unary	char	\$4.	\$4.
I_ORGYN	rejected	binary	char	\$12.	\$12.
U_ORGYN	rejected	binary	num	BEST12.	12.
F_ORGYN	rejected	binary	char	\$12.	\$12.
R_ORGYN0	rejected	interval	num	BEST12.	12.
D_ORGYN_	rejected	unary	char	\$5.	\$5.
EP_ORGYN_	rejected	interval	num	BEST12.	12.
BP_ORGYN_	rejected	binary	num	BEST12.	12.
CP_ORGYN_	rejected	binary	num	BEST12.	12.
P_ORGYN0	rejected	interval	num	BEST12.	12.
ORGANICS	rejected	ordinal	num	BEST12.	12.
ORGYN	rejected	binary	num	BEST12.	12.
R_ORGYN1	target	interval	num	BEST12.	12.

3. Add the Partition node and partition the data into Training and Validation with proportion 70% and 30%, respectively.
4. Add a Tree node. Make your selections appropriately. Make sure that the variable importance ranking result is usable. This is because we expect the diagnostic tree structure to provide clues on which variables have been inadequately represented in the logistic model.
5. View the results. What is the final tree size after applying 1-SE? Which variables seem important?

Final Diagram

Your final diagram should look like this.

