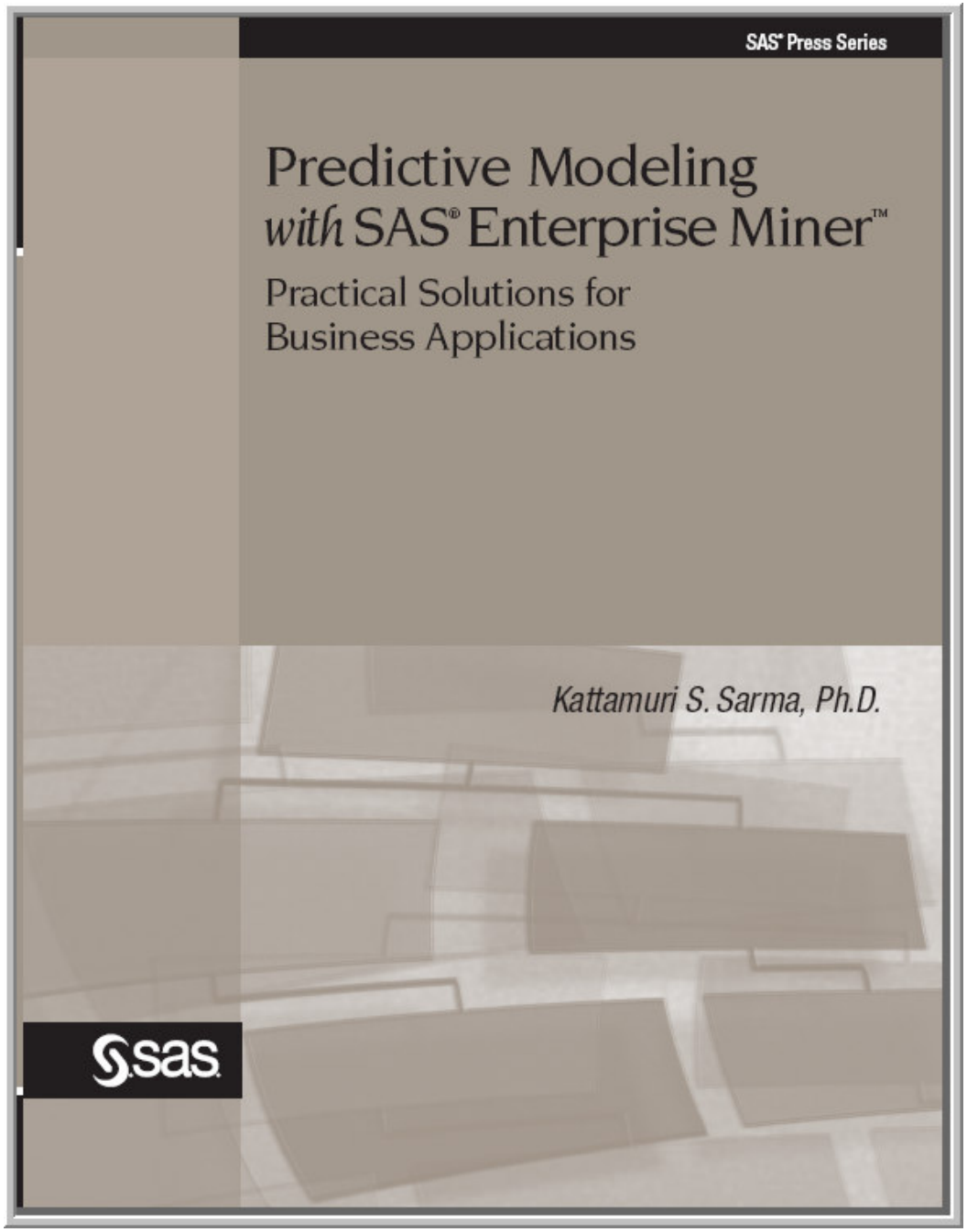
Variable Selection and Transformation of Variables in SAS® Enterprise Miner™

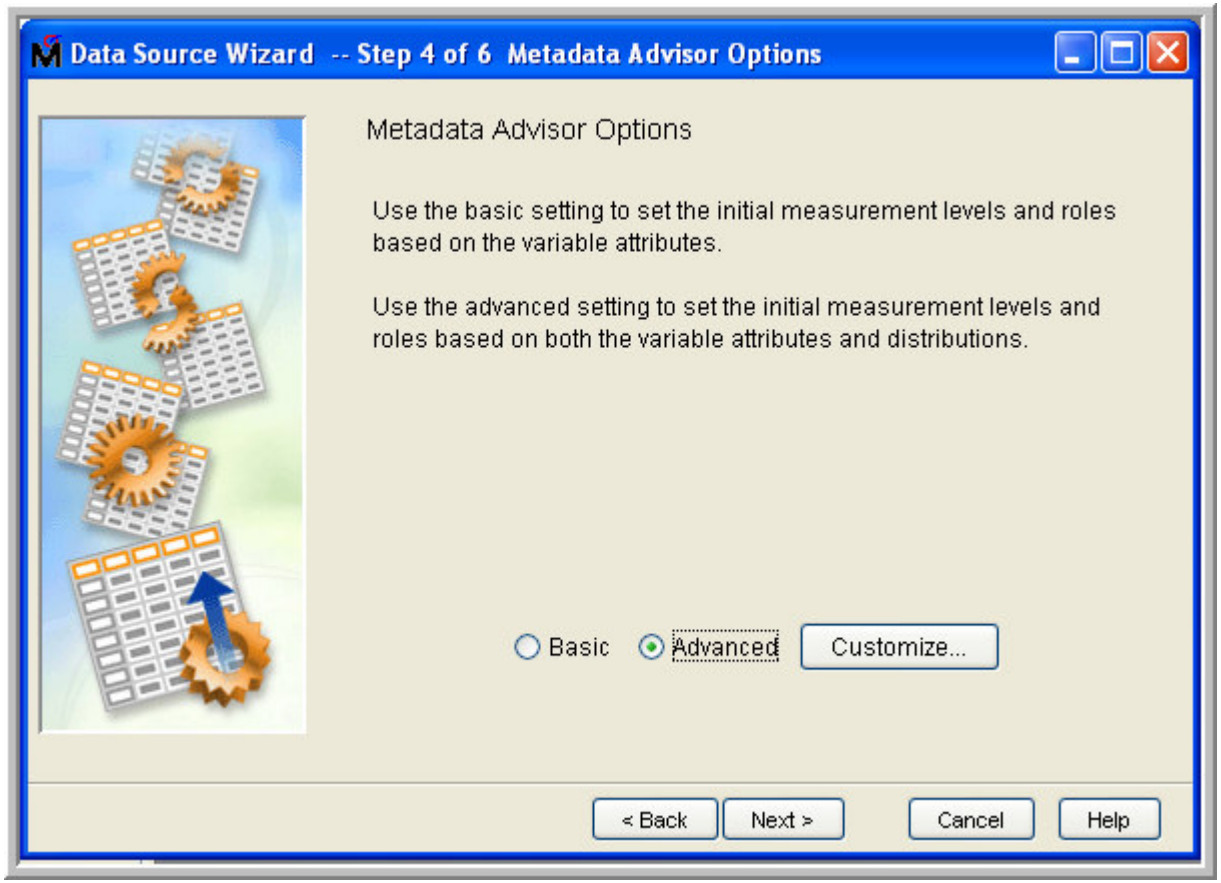
Kattamuri S. Sarma, Ph.D
Ecostat Research Corp.,
White Plains NY
kssarma@worldnet.att.net
kssarma@ecostat-research.com


SAS® Press Series
Predictive Modeling
with SAS® Enterprise Miner™
Practical Solutions for
Business Applications
Kattamuri S. Sarma, Ph.D.
sas

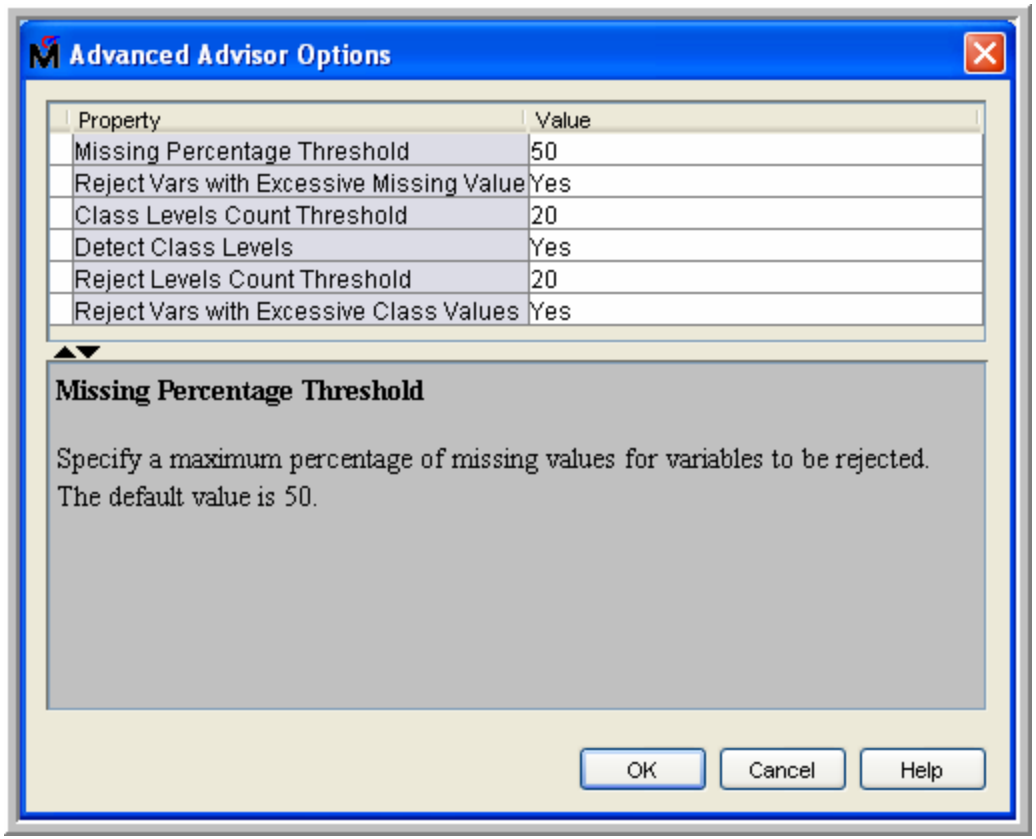
Issues confronting every predictive modeler

1. Too many inputs (as many as 3700 or more)
2. “Unary” inputs
   - The Drop node can be used after creating a meta data with “advanced advisor ” option
3. Extremely skewed or “near-unary” inputs
   - Use StatExplore node or Multiplot node to identify “near unary inputs”, and set them to “reject” status, if necessary. Can be programmed in SAS base
4. “Moderately” Skewed Distributions
   - Use Transform Variables node which allows you to test alternative transformations
5. Too many missing Values
   - Specify a missing value threshold while creating a data source. Variables with missing values greater than the threshold are dropped.
6. “Moderate” number of missing value
   - Determine the source of missing values and impute using the Impute node
7. Many inputs without information value or relevance
   - The Variable Selection node can be used to eliminate irrelevant variables
   - Two methods of selection are available for binary targets – Chi-Square Method, R-Square Method
   - Non-linear relationships detected by means of “AOV16” Variables or “binned” variables

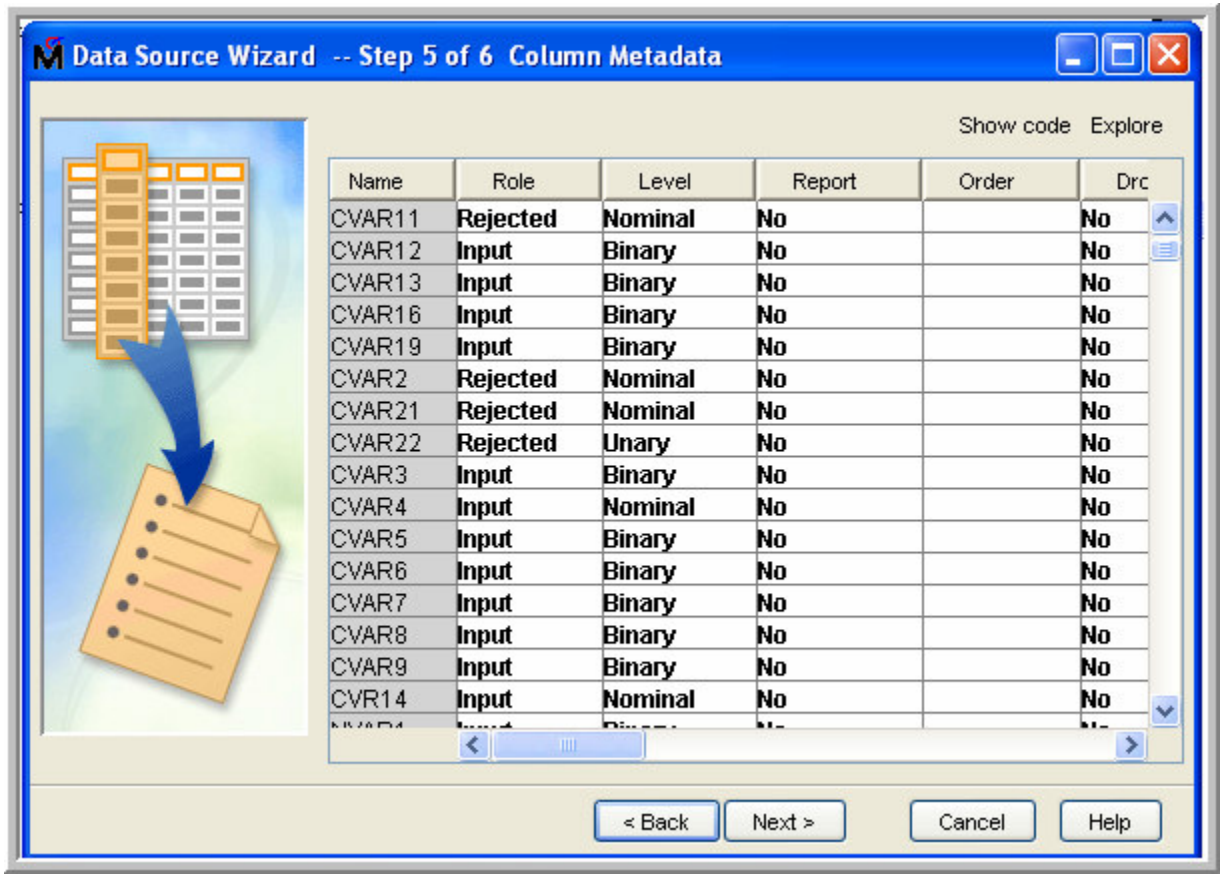
METADATA


Data Source Wizard -- Step 4 of 6 Metadata Advisor Options
Metadata Advisor Options
Use the basic setting to set the initial measurement levels and roles based on the variable attributes.
Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.
Basic
Advanced
Customize...
< Back
Next >
Cancel
Help

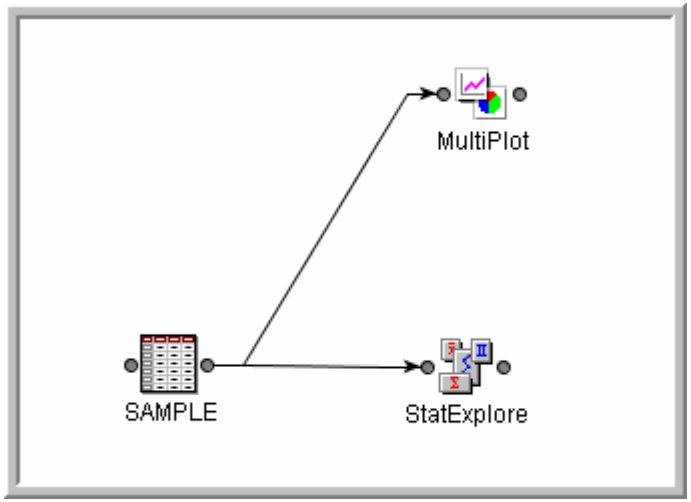
METADATA


Advanced Advisor Options
Property
Value
Missing Percentage Threshold
50
Reject Vars with Excessive Missing Value
Yes
Class Levels Count Threshold
20
Detect Class Levels
Yes
Reject Levels Count Threshold
20
Reject Vars with Excessive Class Values
Yes
Missing Percentage Threshold
Specify a maximum percentage of missing values for variables to be rejected.
The default value is 50.
OK
Cancel
Help

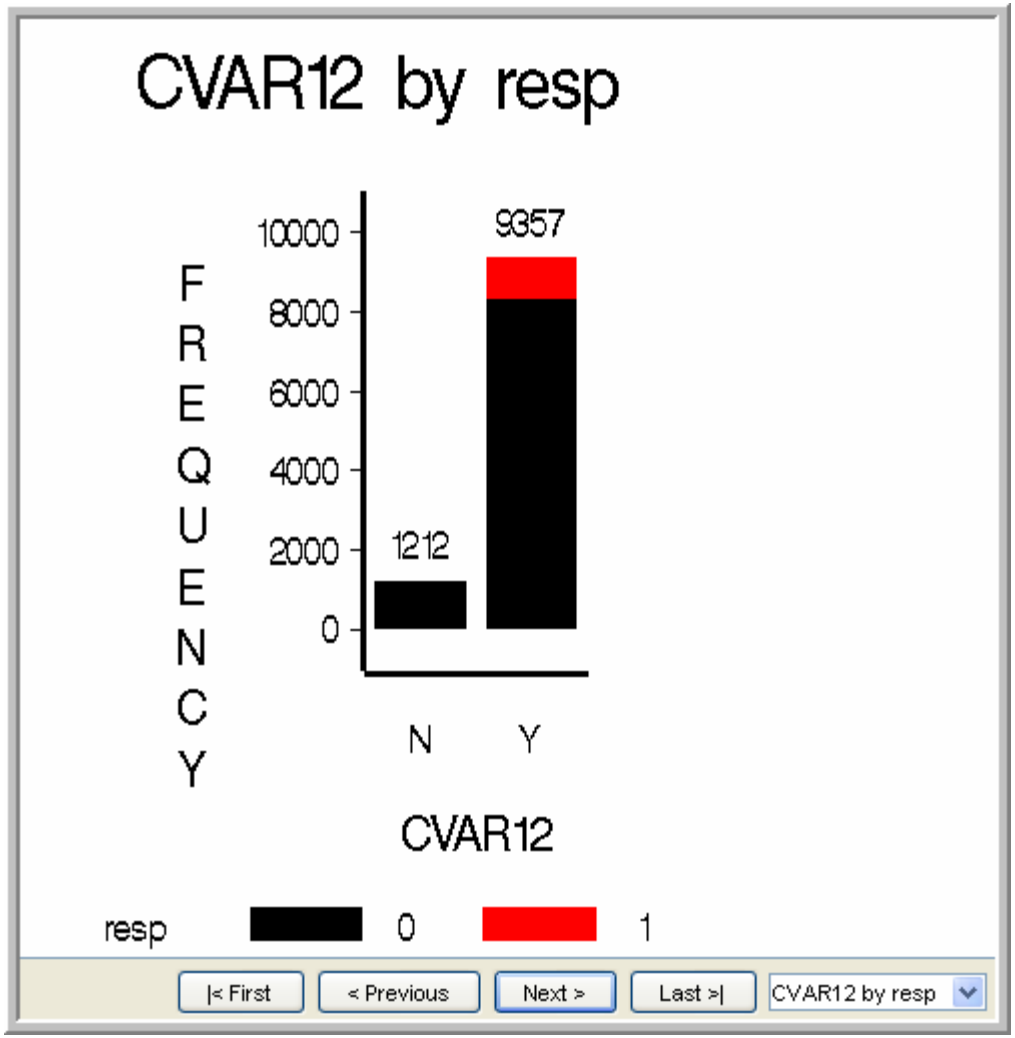
METADATA


Data Source Wizard -- Step 5 of 6 Column Metadata
Show code
Explore
Name
Role
Level
Report
Order
Drc
CVAR11 Rejected Nominal No No
CVAR12 Input Binary No No
CVAR13 Input Binary No No
CVAR16 Input Binary No No
CVAR19 Input Binary No No
CVAR2 Rejected Nominal No No
CVAR21 Rejected Nominal No No
CVAR22 Rejected Unary No No
CVAR3 Input Binary No No
CVAR4 Input Nominal No No
CVAR5 Input Binary No No
CVAR6 Input Binary No No
CVAR7 Input Binary No No
CVAR8 Input Binary No No
CVAR9 Input Binary No No
CVR14 Input Nominal No No
< Back
Next >
Cancel
Help

Explore your data with MultiPlot



Exploring data with MultiPlot



Variable Selection

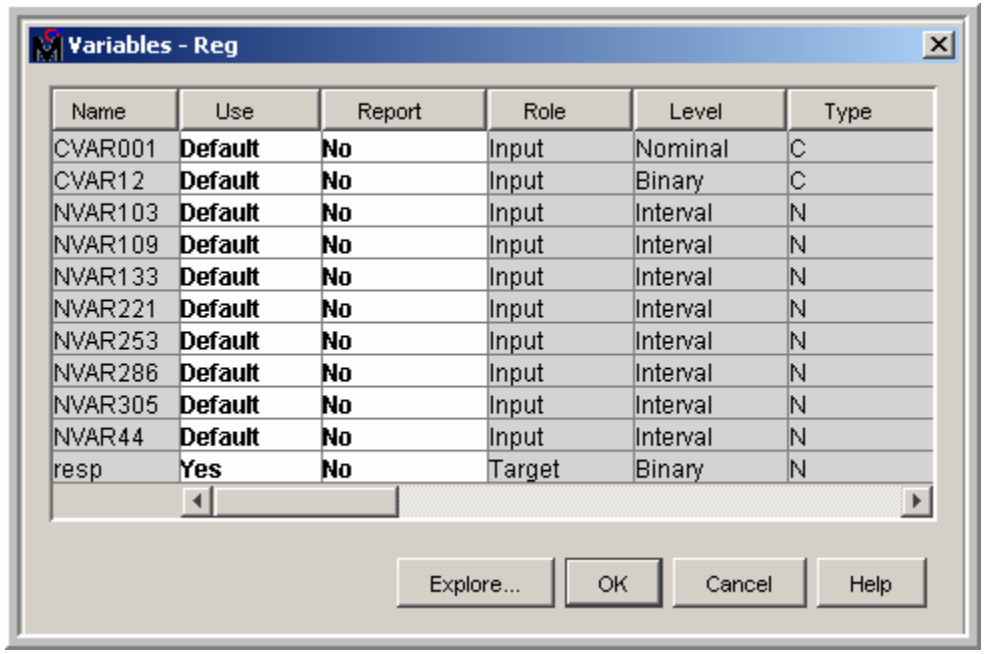
- Enterprise Miner
  - R-Square Method
    - A two step procedure
      - In the first step an initial selection is done based on simple R-Square
      - In the second step a “forward selection procedure is used
    - Non-linear Relationships are detected by means of “binned” variables known as AOV16 variables
    - All possible two-way Interactions can be tested
  - Chi-Square Method
    - Develop a CHAID type of tree, and select all variables which produced significant splits
  - Decision Tree Node
    - Choice of different splitting criteria
    - Variables are selected based on their contribution in producing splits
    - A categorical variable which captures all interactions is created and passed to the next node. The categories are the terminal nodes (leaf nodes) are created and passed.
- Other
  - Calculate information value (IV) of each input, and select those with high IV
    - (Less than 0.02..unpredictive, 0.02 to 0.1.. Weak, 0.1 to 0.3.. Medium, 0.3+ strong)

Variable Selection: R-Square Criterion
Step 2: Final Selection

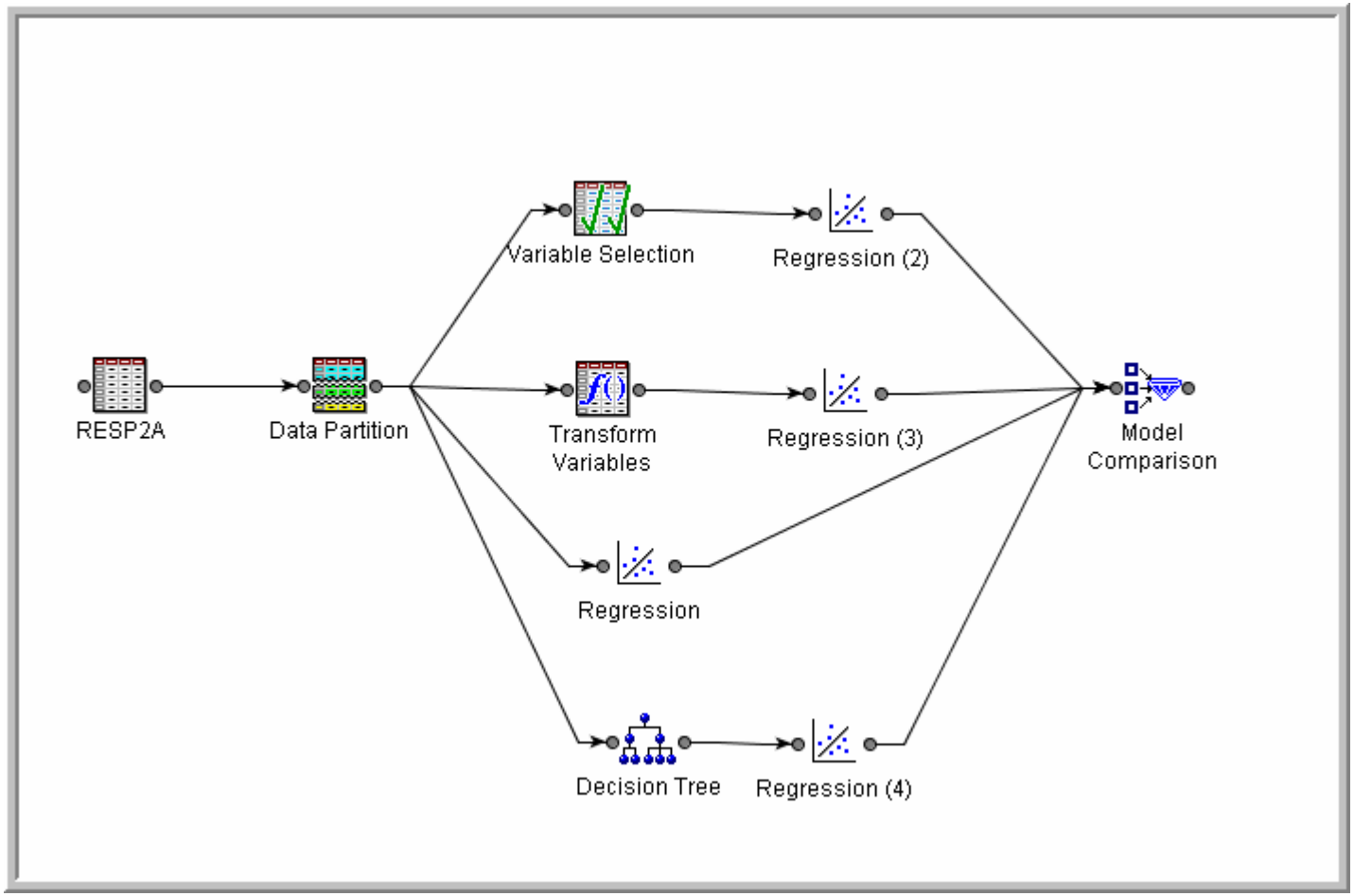
Effects Chosen for Target: resp

Effect	DF	R-Square	F Value	p-Value
AOV16: NVAR253	15	0.051527	22.904008	<.0001
AOV16: NVAR103	15	0.042570	19.764629	<.0001
Group: CVAR001	4	0.031354	56.511905	<.0001
Group: nvar001	5	0.012859	18.802502	<.0001
Class: CVAR12	1	0.009494	70.173878	<.0001
Group: NVAR262	6	0.004830	5.978402	<.0001
AOV16: NVAR4	15	0.004625	2.297146	0.0030
AOV16: NVAR238	13	0.004001	2.299057	0.0050
AOV16: NVAR28	15	0.003080	1.535839	0.0837
AOV16: NVAR163	12	0.002773	1.730899	0.0541
AOV16: NVAR47	12	0.002568	1.604919	0.0828
Group: NVAR73	5	0.001450	2.176465	0.0539
Var: NVAR305	1	0.001090	8.190036	0.0042
Group: NVAR287	5	0.001116	1.678044	0.1362
AOV16: NVAR137	5	0.000898	1.350747	0.2398
Group: NVAR17	3	0.000528	1.322686	0.2650
Group: NVAR16	4	0.001128	2.122426	0.0753

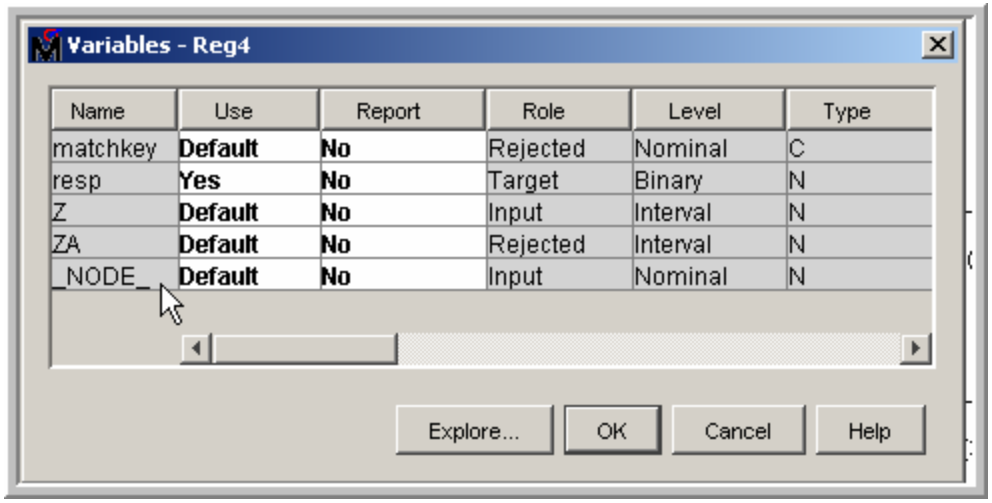
Variables Selected by Chi-Square criterion



Comparing Alternative Transformations of a Single Variable



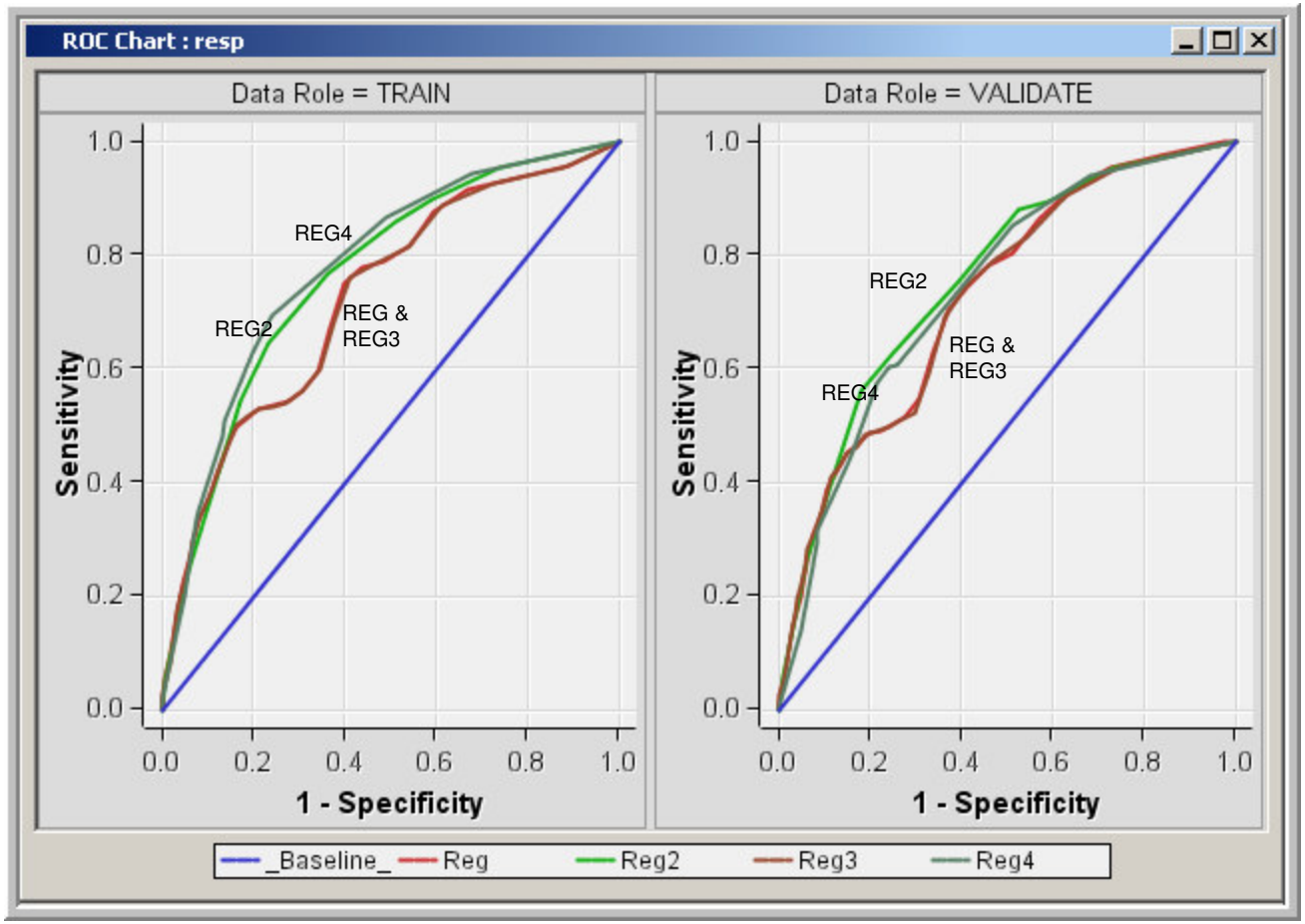
Transformation done by the Decision Tree node



Variable Selection and Transformation by Decision Tree

- Variables which produced significant splits and used in the tree are passed to the next node
- One can run a separate logistic regression for each segment created by the decision tree, or
- Use the special categorical variable “_NODE_” as input to logistic regressions

Comparison of Transformations



Key to the charts

- Reg 2 : The first process flow
  - Uses Variable Selection for testing non linear relationship by means of the AOV16 variable. Non-lenear relationship is stronger than the linear one. Hence it is selected and used in the regression
- Reg3: The second process flow
  - The Transform Variables node is used to select a transformation based on the criterion of “maximizing normality”
- REG: The third process flow
  - The input is used directly in the regression
- REG4: The last process flow at the bottom segment of the process flow diagram
  - The Decision Tree node is used to “bin” the explanatory variable,and the “binned” variable is used in the regression. Each bin represents a leaf node of the decision tree. The decision tree node passes the “binned” or categorical variable to the next node, namely the regression node.

Transformations in Enterprise Miner

Transformation Methods

The Transform Variables node supports various transformation methods. The available methods depends on the type and the role of a variable. See the

- For interval variables:
  - Simple Transformations
  - Binning Transformations
  - Best Power Transformations
- For class variables:
  - Group Rare Levels Transformation
  - Dummy Indicators Transformation

Transformations in Enterprise Miner

Simple Transformations

You can choose from the following simple transformations in the Transform Variables node:

- Log — Variable is transformed by taking the natural log of the variable.
- Square Root — Variable is transformed by taking the square root of the variable.
- Inverse — Variable is transformed by using the inverse of the variable.
- Square — Variable is transformed by using the square of the variable.
- Exponential — Variable is transformed by using the exponential logarithm of the variable.
- Standardize — Variable is standardized by subtracting the mean and dividing by the standard deviation.

Transformations in Enterprise Miner

Binning Transformations

Binning transformations enable you to collapse an interval variable, such as debt to income ratio, into an ordinal grouping variable. There are three types of binning transformations.

- Bucket — Buckets are created by dividing the data values into equally spaced interval based on the difference between the maximum and the minimum values.
- Quantile — Data is divided into groups that have approximately the same frequency in each group.
- Optimal Binning for Relationship to Target — Data is binned in order to optimize the relationship to the target. This method requires a binary target.

Transformations in Enterprise Miner

Best Power Transformations

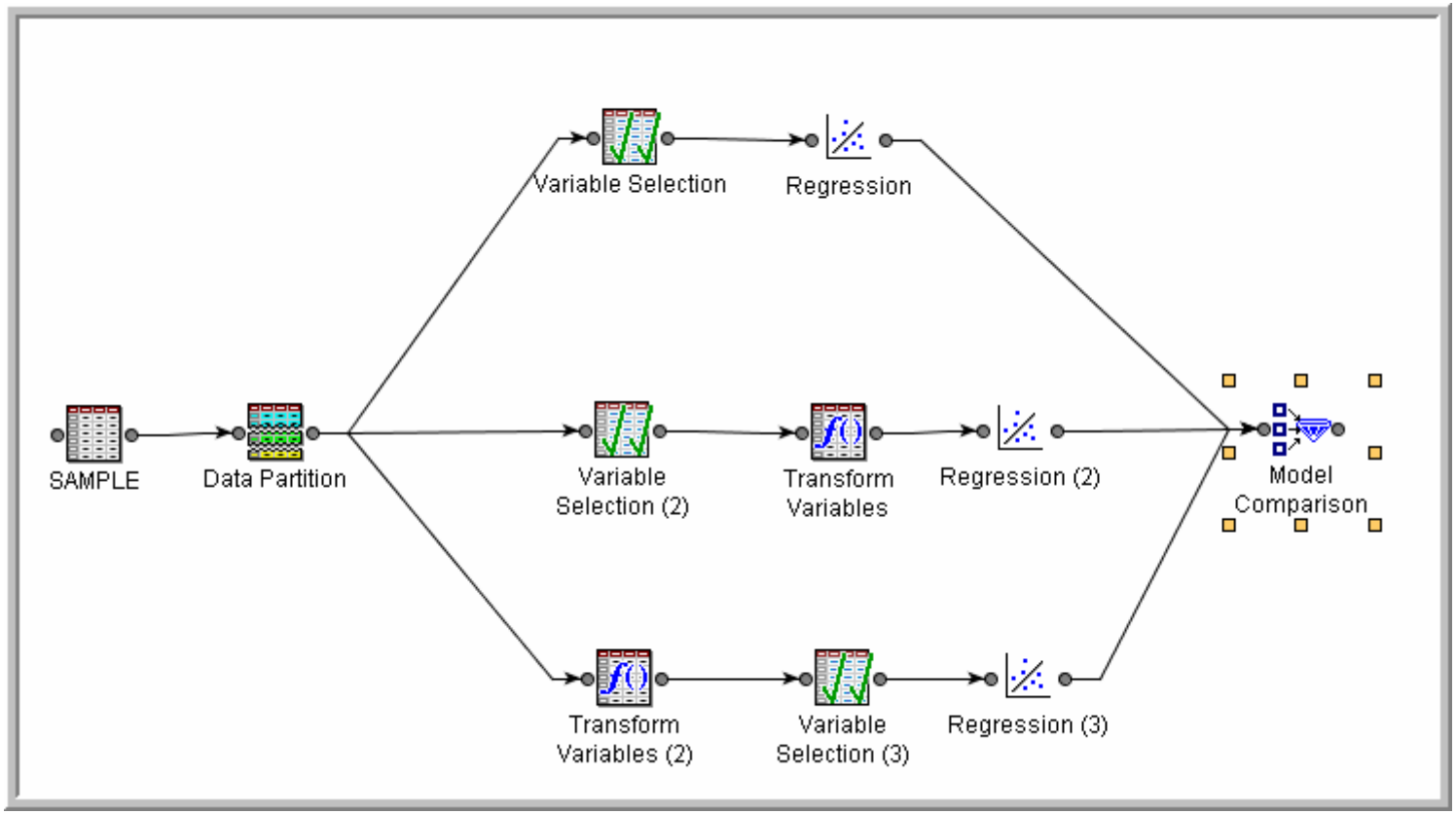
The best power transformations are a subset of the general class of transformations that are known as Box-Cox transformations. The Transform Variables node supports the following best power transformations:

- Maximize Normality — This method chooses the transformation that yields sample quantiles that are closest to the theoretical quantiles of a normal distribution. This method requires an interval target.
- Maximize Correlation with Target — This method chooses the transformation that has the best squar ed correlation with the target. This method requires an interval target. If the maximize correlation transformation is attempted with a non-interval target, the data will will not be transformed.
- Equalize Spread with Target Levels — This method chooses the transformation that has the smal lest variance of the variances between the target levels. This method requires a class target.
- Optimal Maximum Equalize Spread with Target Level — This method chooses the transformation that equalizes spread with target levels. This method requires a class target.

All of the Best Power transformations evaluate the transformation subset listed below, and choose the transformation which has the best results for the specified criterion. In the table below, x represents the transformed variable:

- x
- log(x)
- sqrt(x)
- e^x
- $x^{1/4}$
- x^2
- x^4

Alternative Configurations of variable selection and transformation



Variable Selection: Chi-Square

Property	Value
Node ID	Varsel
Imported Data	...
Exported Data	...
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	Chi-Square
Hide Rejected Variables	Yes
Reject Unused Variables	Yes
Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.0050
Stop R-Square	5.0E-4
Use AOV16 Variables	Yes
Use Group Variables	Yes
Use Interactions	No
SPDS	Yes
Print Option	Default
Status	
Time of Creation	1/13/08 10:47 AM
Run Id	d4b1c178-149c-4e0c-b535-ef0871744
Last Error	
Last Status	Complete
Needs Updating	No
Needs to Run	No
Time of Last Run	1/13/08 2:06 PM
Run Duration	0 Hr. 0 Min. 35.17 Sec.
Grid Host	

Transformation of Variables : Maximize Normality

Property	Value
Node ID	Trans
Imported Data	
Exported Data	
Variables	
Formula Builder	
Interactions Editor	
Default Methods	
Interval Inputs	Maximum Normal
Interval Targets	None
Class Inputs	Dummy Indicators
Class Targets	None
Sample Properties	
Method	Top
Size	Default
Random Seed	12345
Grouping Method	
Cutoff Value	0.5
Group Missing	No
Original Variables	
Hide	Yes
Reject	Yes
Missing Value	Use in Search
Add Minimum Value to Offset Value	Yes
Offset Value	1.0
Summary Variables	Transformed and New Variables
Status	
Time of Creation	1/13/08 11:49 AM
Run Id	20c24e08-3a86-4c7f-8c92-11b14f04
Last Error	
Last Status	Complete
Needs Updating	No
Needs to Run	No
Time of Last Run	1/15/08 11:08 AM
Run Duration	0 Hr. 0 Min. 26.94 Sec.
Grid Host	

Comparative Performance of Alternative Configurations