



## Using the Variable Selection Node

The Variable Selection tool provides a selection based on one of two criteria.

When you use the *R-square variable selection criterion*, a two-step process is followed.

1. SAS Enterprise Miner computes the squared correlation for each variable and then assigns the *Rejected* role to those variables that have a value less than the squared correlation criterion. (The default is 0.005.)
2. SAS Enterprise Miner evaluates the remaining (not rejected) variables using a forward stepwise R-square regression. Variables that have a stepwise R-square improvement less than the threshold criterion (default=0.0005) are assigned the *Rejected* role.

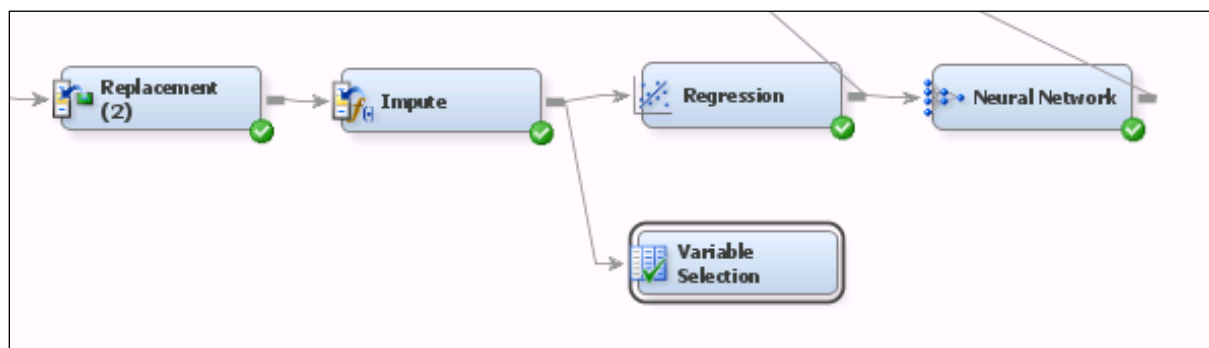
When you use the *chi-square selection criterion*, variable selection is performed using binary splits for maximizing the chi-square value of a 2x2 frequency table. The rows of the 2x2 table are defined by the (binary) target variable. The columns of the table are formed by a partition of the training data using a selected input.

Several partitions are considered for each input. For an  $L$ -level class input (binary, ordinal, or nominal), partitions are formed by comparing each input level separately to the remaining  $L-1$  input levels, creating a collection of  $L$  possible data partitions. The partition with the highest chi-square value is chosen as the input's best partition. For interval inputs, partitions are formed by dividing the input range into (a maximum of) 50 equal-length bins and splitting the data into two subsets at one of the 49 bin boundaries. The partition with the highest chi-square statistic is chosen as the interval input's best partition. The partition-and-variable combination with the highest chi-square statistic is used to split the data, and the process is repeated within both subsets. The partitioning stops when no input has a chi-square statistic in excess of a user-specified threshold. All variables not used in at least one partition are rejected.

The Variable Selection node's chi-square approach is quite similar to a decision tree algorithm with its ability to detect nonlinear and nonadditive relationships between the inputs and the target. However, the method for handling categorical inputs makes it sensitive to spurious input and target correlations. Instead of the Variable Selection node's chi-square setting, you might want to try the Decision Tree node, properly configured for input selection.

The following steps show how to use the Variable Selection tool with the R-square setting.

1. Click the **Explore** tab.
2. Drag a **Variable Selection** tool into the diagram workspace.
3. Connect the **Impute** node to the **Variable Selection** node. (Only a portion of the flow is shown.)



The default Target Model (input selection method) property is set to **Default**. With this default setting, if the target is binary and the total parameter count for a regression model exceeds 400, the chi-square method is used. Otherwise, the R-square method of variable selection is used.

4. Select **Target Model** ⇨ **R-Square**.

Property	Value
<b>General</b>	
Node ID	Varsel
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Max Class Level	100
Max Missing Percent	50
Target Model	R-Square
Manual Selector	...
Rejects Unused Inputs	Yes
<input checked="" type="checkbox"/> Bypass Options	
Variable	None
Role	Input



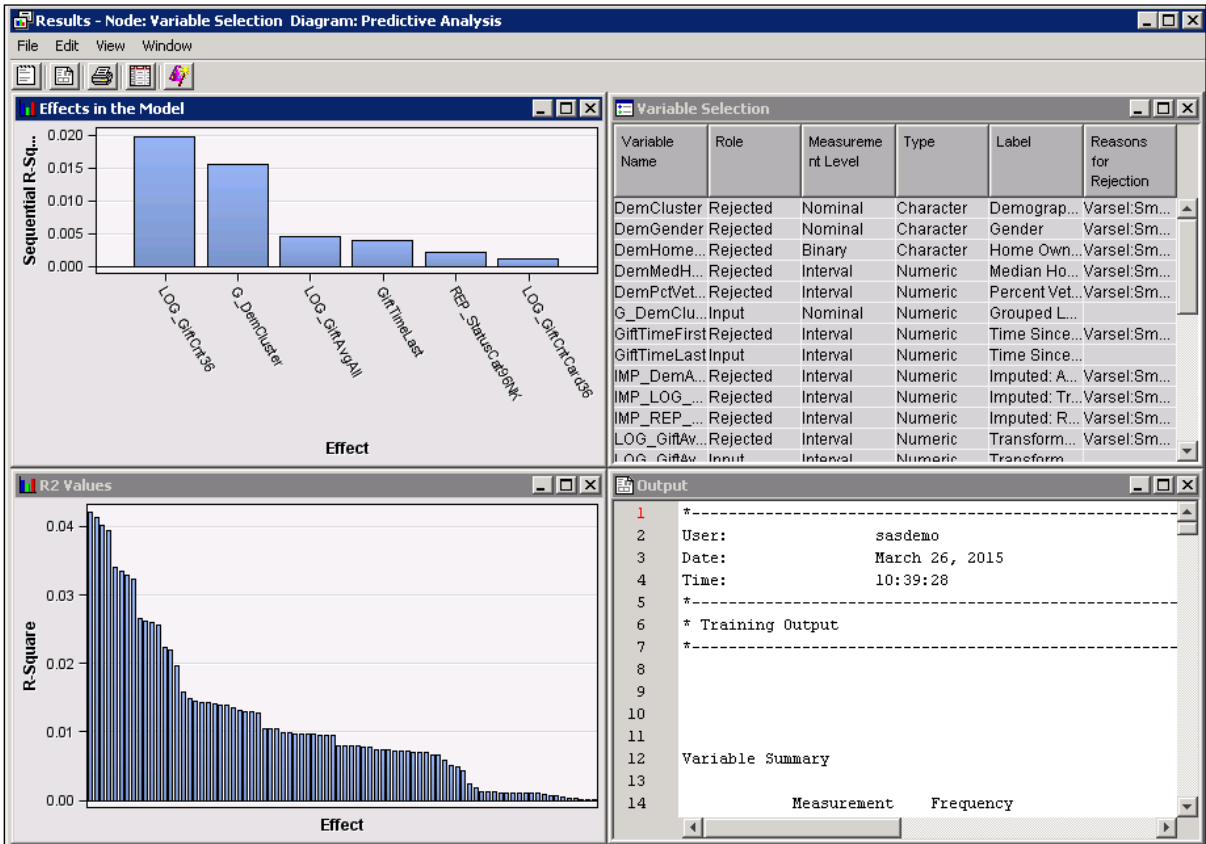
Recall that a two-step process is performed when you apply the R-square variable selection criterion to a binary target. The Properties panel enables you to specify the maximum number of variables that can be selected, the cutoff minimum squared correlation measure for an individual variable to be selected, and the necessary R-square improvement for a variable to remain as an input variable.

Additional available options include the following:

- **Use AOV16 Variables** – When selected, this option requests SAS Enterprise Miner to bin interval variables into 16 equally spaced groups (AOV16). The AOV16 variables are created to help identify nonlinear relationships with the target. Bins with zero observations are eliminated, which means that an AOV16 variable can have fewer than 16 bins.
- **Use Group Variables** – When set to **Yes**, this option enables the number of levels of a class variable to be reduced, based on the relationship of the levels to the target variable.
- **Use Interactions** – When this option is selected, SAS Enterprise Miner evaluates two-way interactions for categorical inputs.

5. Run the Variable Selection node and view the results.

The Results window appears.



6. Maximize the Variable Selection window.

7. Select the **ROLE** column heading to sort the variables by their assigned roles.
8. Select the **Reasons for Rejection** column heading.

Variable Selection					
Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection ▲
G_DemCluster	Input	Nominal	Numeric	Grouped Levels fo...	
GiftTimeLast	Input	Interval	Numeric	Time Since Last Gift	
LOG_GiftAvgAll	Input	Interval	Numeric	Transformed: Gift ...	
LOG_GiftCnt36	Input	Interval	Numeric	Transformed: Gift ...	
LOG_GiftCntCard36	Input	Interval	Numeric	Transformed: Gift ...	
REP_StatusCat96...	Input	Nominal	Character	Replacement: Stat...	
DemGender	Rejected	Nominal	Character	Gender	Varsel:Small R-sq...
DemHomeOwner	Rejected	Binary	Character	Home Owner	Varsel:Small R-sq...
DemMedHomeVal...	Rejected	Interval	Numeric	Median Home Val...	Varsel:Small R-sq...
DemPctVeterans	Rejected	Interval	Numeric	Percent Veterans ...	Varsel:Small R-sq...
GiftTimeFirst	Rejected	Interval	Numeric	Time Since First Gift	Varsel:Small R-sq...
IMP_DemAge	Rejected	Interval	Numeric	Imputed: Age	Varsel:Small R-sq...
IMP_LOG_GiftAvg...	Rejected	Interval	Numeric	Imputed: Transfor...	Varsel:Small R-sq...
IMP_REP_DemMe...	Rejected	Interval	Numeric	Imputed: Replace...	Varsel:Small R-sq...
LOG_GiftAvg36	Rejected	Interval	Numeric	Transformed: Gift ...	Varsel:Small R-sq...
LOG_GiftAvgLast	Rejected	Interval	Numeric	Transformed: Gift ...	Varsel:Small R-sq...
LOG_GiftCntAll	Rejected	Interval	Numeric	Transformed: Gift ...	Varsel:Small R-sq...
LOG_GiftCntCardAll	Rejected	Interval	Numeric	Transformed: Gift ...	Varsel:Small R-sq...
M_DemAge	Rejected	Binary	Numeric	Imputation Indicat...	Varsel:Small R-sq...
M_LOG_GiftAvgCa...	Rejected	Binary	Numeric	Imputation Indicat...	Varsel:Small R-sq...
M_REP_DemMedl...	Rejected	Binary	Numeric	Imputation Indicat...	Varsel:Small R-sq...
PromCnt12	Rejected	Interval	Numeric	Promotion Count 1...	Varsel:Small R-sq...
PromCnt36	Rejected	Interval	Numeric	Promotion Count 3...	Varsel:Small R-sq...
PromCntAll	Rejected	Interval	Numeric	Promotion Count A...	Varsel:Small R-sq...
PromCntCard12	Rejected	Interval	Numeric	Promotion Count ...	Varsel:Small R-sq...
PromCntCard36	Rejected	Interval	Numeric	Promotion Count ...	Varsel:Small R-sq...
PromCntCardAll	Rejected	Interval	Numeric	Promotion Count ...	Varsel:Small R-sq...
StatusCatStarAll	Rejected	Binary	Numeric	Status Category St...	Varsel:Small R-sq...
DemCluster	Rejected	Nominal	Character	Demographic Clu...	Varsel:Small R-sq...

The Variable Selection node finds that most inputs have insufficient target correlation to justify keeping them. You can try these inputs in a subsequent model or adjust the R-square settings to be less severe. Notice the input **G\_DemCluster** is a grouping of the original **DemCluster** input.



Binary targets generate notoriously low R-square statistics. A more appropriate association measure might be the likelihood chi-square statistic found in the Regression node.



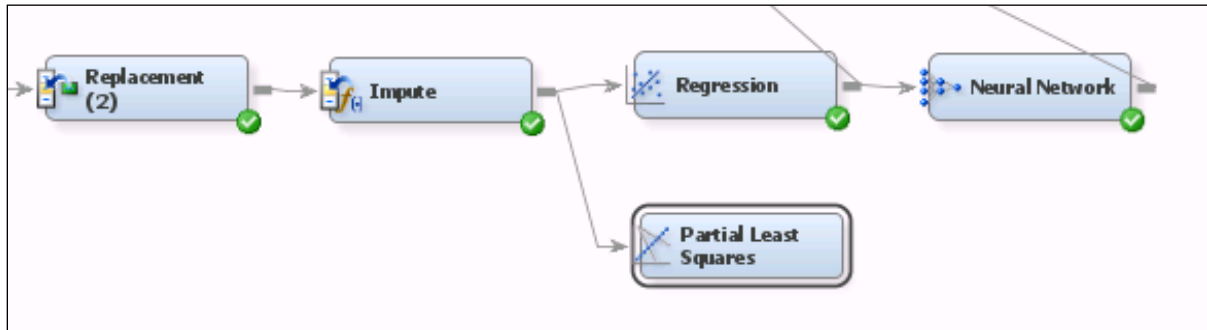
## Using Partial Least Squares for Input Selection

Partial least squares (PLS) regression can be thought of as a merging of multiple and principal components regression. In multiple regression, the goal is to find linear combinations of the inputs that account for as much (linear) variation in the target as possible. In principal component regression, the goal is to find linear combinations of the inputs that account for as much (linear) variation in the input space as possible, and then use these linear combinations (called *principal component vectors*) as the inputs to a multiple regression model. In PLS regression, the goal is to have linear combinations of the inputs (called *latent vectors*) that account for variation in **both** the inputs and the target. The technique can extract a small number of latent vectors from a set of correlated inputs that correlate with the target.

A useful feature of the PLS procedure is the inclusion of an input importance metric named *variable importance in the projection* (VIP). VIP quantifies the relative importance of the original input variables to the latent vectors. A sufficiently small VIP for an input (less than 0.8, by default) plus a small parameter estimate (less than 0.1, by default) permits removal of the input from the model.

The following steps demonstrate the use of the PLS tool for variable selection:

1. Click the **Model** tab.
2. Drag a **Partial Least Squares** tool into the diagram workspace.
3. Connect the **Impute** node to the **Partial Least Squares** node.

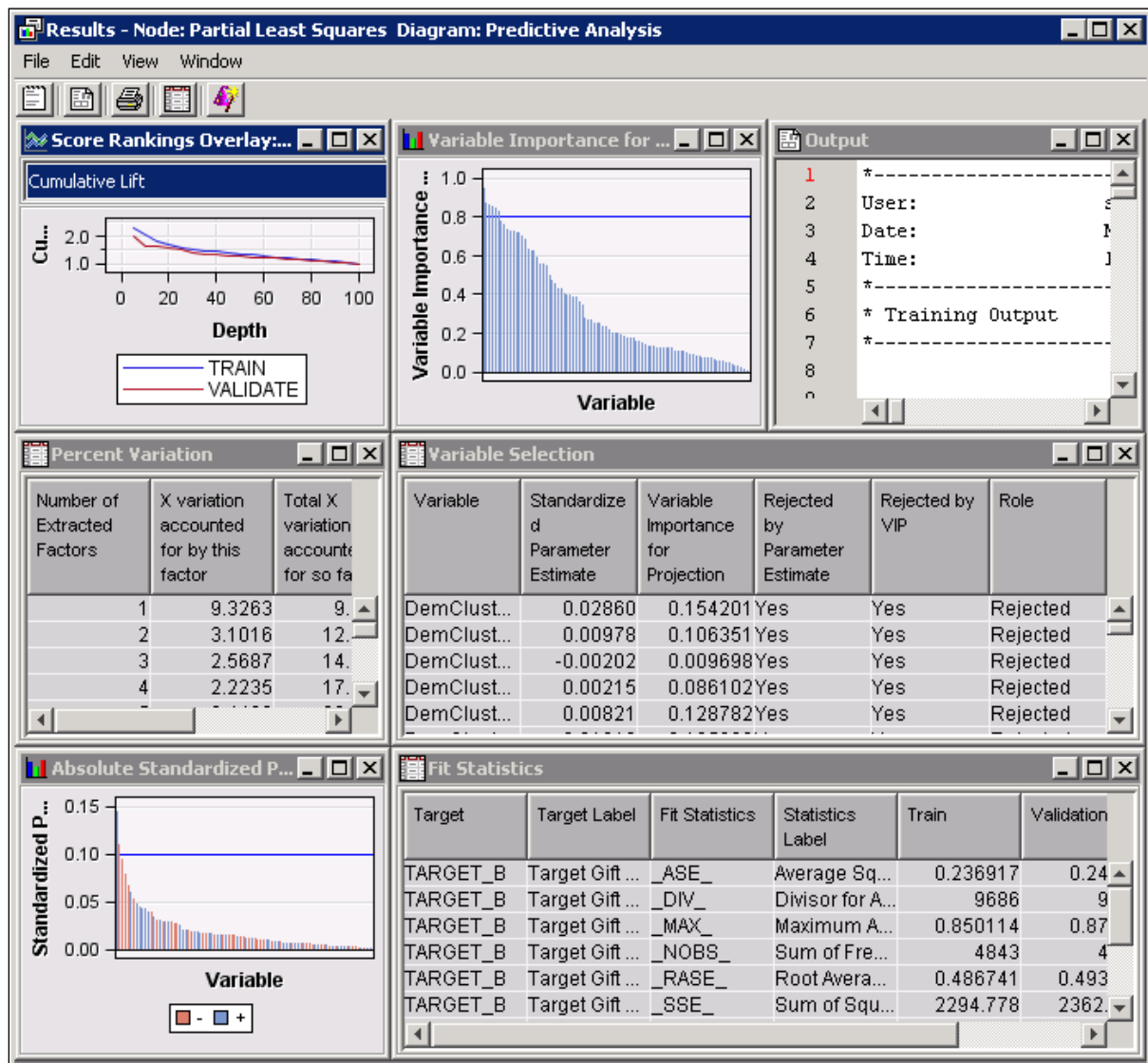


4. Select **Export Selected Variables** ⇒ Yes.

Train	
Variables	...
Modeling Techniques	
Regression Model	PLS
PLS Algorithm	NIPALS
Maximum Iteration	200
Epsilon	1.0E-12
Number of Factors	
Default	Yes
Number of Factors	15
Cross Validation	
CV Method	None
CV N Parameter	7
Random CV Options	
Number of Iterations	10
Default No. of Test Obs.	Yes
No. of Test Obs.	100
Default Random Seed	Yes
Random Seed	1234
Score	
Variable Selection	
Variable Selection Criterion	Both
Para. Est. Cutoff	0.1
VIP Cutoff	0.8
Export Selected Variables	Yes
Hide Rejected Variables	Yes



- Run the Partial Least Squares node and view the results.



6. Maximize the Percent Variation window.

Percent Variation				
Number of Extracted Factors	X variation accounted for by this factor	Total X variation accounted for so far	Y variation accounted for by this factor	Total Y variation accounted for so far
1	9.3263	9.3263	2.9889	2.9889
2	3.1016	12.4279	1.5930	4.5818
3	2.5687	14.9966	0.4026	4.9844
4	2.2235	17.2201	0.0909	5.0754
5	3.4493	20.6694	0.0223	5.0977
6	1.7022	22.3716	0.0290	5.1267
7	1.8648	24.2364	0.0134	5.1400
8	2.5094	26.7458	0.0113	5.1513
9	1.0899	27.8357	0.0238	5.1751
10	1.5489	29.3846	0.0096	5.1847
11	1.1964	30.5809	0.0086	5.1934
12	0.7255	31.3065	0.0158	5.2091
13	1.2303	32.5368	0.0087	5.2178
14	1.3137	33.8505	0.0065	5.2243
15	0.8789	34.7294	0.0089	5.2332

By default, the Partial Least Squares tool extracts 15 latent vectors or factors from the training data set. These factors account for 35% and 5.2% of the variation in the inputs and target, respectively.

7. Maximize the Variable Selection window.
8. Select the **Role** column heading to sort the table by variable role.

Variable Selection					
Variable	Standardized Parameter Estimate	Variable Importance for Projection	Rejected by Parameter Estimate	Rejected by VIP	Role ▲
GiftTimeLast	-0.07806	0.842938	Yes	No	Input
LOG_GiftAvgAll	-0.05216	0.824256	Yes	No	Input
LOG_GiftCnt36	0.06015	0.948709	Yes	No	Input
LOG_GiftCntAll	-0.09391	0.871212	Yes	No	Input
LOG_GiftCntCard...	0.04138	0.851113	Yes	No	Input
LOG_GiftCntCard...	0.01129	0.863856	Yes	No	Input
PromCntAll	0.14428	0.717499	No	Yes	Input
PromCntCardAll	-0.10896	0.756161	No	Yes	Input
DemCluster 00	0.02860	0.154201	Yes	Yes	Rejected
DemCluster 01	0.00978	0.106351	Yes	Yes	Rejected
DemCluster 02	-0.00202	0.009698	Yes	Yes	Rejected
DemCluster 03	0.00215	0.086102	Yes	Yes	Rejected
DemCluster 04	0.00821	0.128782	Yes	Yes	Rejected
DemCluster 05	-0.01016	0.105993	Yes	Yes	Rejected
DemCluster 06	-0.01654	0.236457	Yes	Yes	Rejected
DemCluster 07	0.02488	0.271819	Yes	Yes	Rejected
DemCluster 08	-0.00984	0.137011	Yes	Yes	Rejected
DemCluster 09	-0.00225	0.053855	Yes	Yes	Rejected
DemCluster 10	-0.01887	0.266047	Yes	Yes	Rejected
DemCluster 11	0.01489	0.19543	Yes	Yes	Rejected
DemCluster 12	0.00299	0.057599	Yes	Yes	Rejected
DemCluster 13	0.02883	0.456834	Yes	Yes	Rejected
DemCluster 14	0.01939	0.267769	Yes	Yes	Rejected
DemCluster 15	0.00304	0.001325	Yes	Yes	Rejected
DemCluster 16	0.00268	0.078152	Yes	Yes	Rejected
DemCluster 17	0.00653	0.070584	Yes	Yes	Rejected
DemCluster 18	0.00049	0.004551	Yes	Yes	Rejected
DemCluster 19	0.01337	0.14982	Yes	Yes	Rejected
DemCluster 20	-0.01075	0.124651	Yes	Yes	Rejected
DemCluster 21	0.01681	0.1812	Yes	Yes	Rejected

The majority of the selected inputs relate to donation count, promotion count, time since donation, and amount of donation.

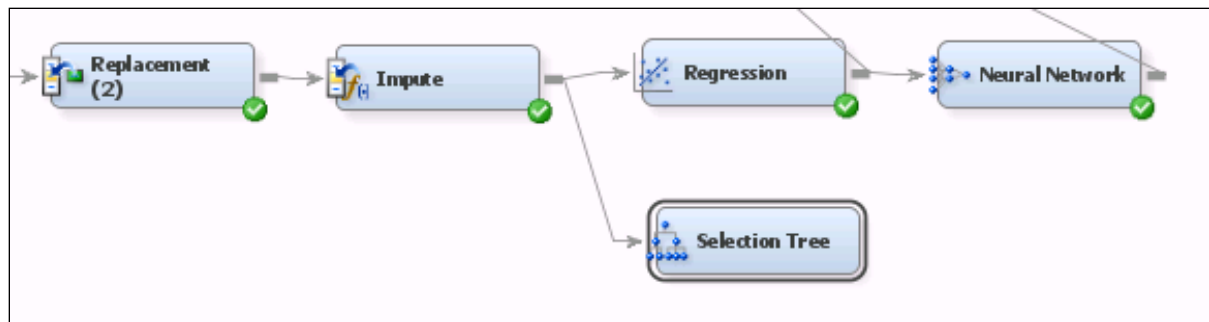


## Using the Decision Tree Node for Input Selection


Decision trees can be used to select inputs for flexible predictive models. They have an advantage over using a standard regression model or the Variable Selection tool's R-square method if the inputs' relationships to the target are nonlinear or nonadditive.

1. Connect a **Decision Tree** node to the **Impute** node. Rename the Decision Tree node **Selection Tree**.





You can use the Tree node with default settings to select inputs. However, this tends to select too few inputs for a subsequent model. Two changes to the Tree defaults result in more inputs being selected. Generally, when you use trees to select inputs for flexible models, it is better to err on the side of too many inputs rather than too few. The model's complexity optimization method can usually compensate for the extra inputs.

 The changes below to the defaults act independently. You can experiment to discover which method generalizes best with your data.

2. Enter **1** as the Number of Surrogate Rules value.
3. Select **Subtree** ⇒ **Method** ⇒ **Largest**.

Property	Value
<b>Splitting Rule</b>	
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	6
Minimum Categorical Size	5
<b>Node</b>	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	1
Split Size	.
<b>Split Search</b>	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
<b>Subtree</b>	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25



Changing the number of surrogates enables inclusion of surrogate splits in the variable selection process. By definition, surrogate inputs are typically correlated with the selected split input. Although it is usually a bad practice to include redundant inputs in predictive models, many flexible models tolerate some degree of input redundancy. The advantage of including surrogates in the variable selection is to enable inclusion of inputs that do not appear in the tree explicitly but are still important predictors of the target.

Changing the Subtree method causes the tree algorithm to not prune the tree. As with adding surrogate splits to the variable selection process, it tends to add (possibly irrelevant) inputs to the selection list.

4. Run the Selection Tree node and view the results.
5. To view a complete list of variables selected by the tree, including surrogate variables, select **View** ⇒ **Model** ⇒ **Variable Importance** from the Selection Tree results. Only variables with nonzero values for Importance are the variables selected by the Selection Tree. This list includes surrogate variables that do not appear in the tree. (The same list can also be viewed in the Output window.)

Variable Importance						
Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_GiftC...	Transforme...	1	0	1.0000	1.0000	1.0000
LOG_GiftC...	Transforme...	0	1	0.8994	0.8994	1.0000
PromCnt12	Promotion ...	2	1	0.6637	0.3618	0.5451
DemCluster	Demograp...	3	2	0.6591	0.0978	0.1483
DemMedH...	Median Ho...	2	1	0.6435	0.1086	0.1688
LOG_GiftAv...	Transforme...	0	3	0.6273	0.7094	1.1310
LOG_GiftAv...	Transforme...	1	0	0.5241	0.6719	1.2819
GiftTimeLast	Time Since ...	1	0	0.4809	0.4453	0.9259
IMP_LOG_...	Imputed: Tr...	2	0	0.4000	0.3210	0.8024
LOG_GiftAv...	Transforme...	1	0	0.3936	0.0000	0.0000
PromCnt36	Promotion ...	0	1	0.3782	0.0000	0.0000
PromCntCa...	Promotion ...	0	1	0.3654	0.0000	0.0000
DemHome...	Home Owner	1	0	0.3077	0.0000	0.0000
StatusCatSt...	Status Cate...	1	0	0.2946	0.0660	0.2240
IMP_REP_...	Imputed: R...	0	1	0.2665	0.0000	0.0000
LOG_GiftC...	Transforme...	0	1	0.2625	0.0588	0.2240
DemPctVet...	Percent Vet...	0	1	0.0287	0.0000	0.0000
LOG_GiftC...	Transforme...	0	0	0.0000	0.0000	.
IMP_DemA...	Imputed: Age	0	0	0.0000	0.0000	.
DemGender	Gender	0	0	0.0000	0.0000	.
PromCntCa...	Promotion ...	0	0	0.0000	0.0000	.
GiftTimeFirst	Time Since ...	0	0	0.0000	0.0000	.
PromCntCa...	Promotion ...	0	0	0.0000	0.0000	.
REP_Statu...	Replaceme...	0	0	0.0000	0.0000	.
PromCntAll	Promotion ...	0	0	0.0000	0.0000	.
M_DemAge	Imputation I...	0	0	0.0000	0.0000	.
M_LOG_Gif...	Imputation I...	0	0	0.0000	0.0000	.
M_REP_De...	Imputation I...	0	0	0.0000	0.0000	.

The Importance column quantifies approximately how much of the overall variability in the target each input explains. The values are normalized by the amount of variability explained by the input with the highest importance. The variable importance definition not only considers inputs selected as split variables, but it also accounts for surrogate inputs (if a positive number of surrogate rules are selected in the Properties pane). For example, the second most important input (**LOG\_GiftCntCard36**) accounts for almost the same variability as the most important input (**LOG\_GiftCnt36**) even though it does not explicitly appear in the tree.