

## Tip: Fit Multivariate Adaptive Regression Splines (MARS) in SAS® Enterprise Miner™

Multivariate Adaptive Regression Splines (Friedman, 1991) is a nonparametric technique that combines regression splines and model selection methods. It is a powerful predictive modeling tool because 1) it extends linear models to analyze nonlinear dependencies 2) it produces parsimonious models that do not overfit the data and thus have good predictive power. MARS constructs spline basis functions in an adaptive way by automatically selecting appropriate knot values for different variables. This can help E-miners to identify linear and nonlinear variables, and the interactions of them as well. When excluding higher order terms, MARS is really good at identifying the effects of single variables in a multivariate setting. This makes it highly usable in process control and for identifying experimental designs. MARS also has its application in forecasting as a variable screening tool.

It has always been a desirable tool for our E-miners and now you have MARS as an extension node in Enterprise Miner by just following a few simple steps.

1. Download all the files from the Github repository (<https://github.com/sassoftware/dm-flow/MARS>), including a XML file (MARS.xml) defining the node properties, a SAS catalog (emextn.sas7bcat), and two GIF files (MARS\_16.gif and MARS\_32.gif) for the node icon.
2. To deploy the extension node, you need to follow the steps as instructed in Chapter 5 “Deploying an Extension Node” in [“SAS® Enterprise Miner™ 14.1 Extension Nodes: Developer’s Guide”](#).
3. After store the files in the proper directories, restart the Enterprise Miner server if necessary.
4. The MARS extension node runs with SAS Enterprise Miner 13.1 or any later version.

Once deployed, you can find the MARS node under the Applications tab as shown in Figure 1.

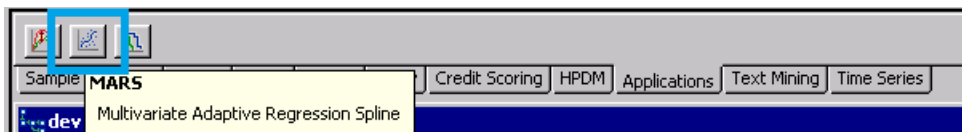


Figure 1: MARS node under the Applications Tab on the toolbar

### MARS Node Requirements

One or more input variables are required for the MARS node. The data set can contain at most one target variable, either interval or categorical.

If the input data set contains a frequency variable, the frequency variable must be an interval variable and all observations must be positive integers.

### MARS Node Properties

Drag a MARS node onto an open diagram, and you will see the property panel as shown in Figure 2.

Here are the descriptions of main properties.

- **Main Effects Only** – Specifies whether to include main effects only. If No is selected, then two-way or higher order interaction between spline basis functions are included.

General	
Node ID	MARS10
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
<input checked="" type="checkbox"/> Regression Splines	
Main Effects Only	Yes
Interaction Orders	2
Keep Effects	...
Effects Without Transformation	...
Exclude Missing	No
Spline Options	...
Penalty	No Penalty
<input checked="" type="checkbox"/> Modeling	
Probability Distribution	Default
Link Function	Default
Selection Method	Default
Use Fast Algorithm	No
Fast Algorithm Options	...
Optimization Options	...
Convergence Options	...
<input checked="" type="checkbox"/> Cross Validation	
Cross Validation	No
Number of Folds	2
Random Seed	12345
<input checked="" type="checkbox"/> Output Data Set	
Output Design Matrix	No
Selected Model	After Backward Selection
Score	
Excluded Rejected Variable	Reject

Figure 2: MARS node properties panel

- **Interaction Orders** – Specifies higher order interaction when **Main Effects Only** is set to “No”.
- **Keep Effects** – Specifies a list of variables to be included in the final model.
- **Effects Without Transformation** – Specifies a list of variables to be considered without nonparametric transformation. Variables should appear in the linear form if they are selected.
- **Exclude Missing** – Specifies whether to exclude missing from train data.
- **Spline Options** – asdfasdfsdf
  - **Maximum Number of Basis** – Uses default the maximum number of basis functions in the final model or specifies in the Maximum Basis Number property. Default is the larger value between 21 and one plus two times the number of non-intercept effects specified in the MODEL statement.
  - **Maximum Basis Number** – Specifies the number of maximum number of basis functions that can be used in the final model when **Maximum Number of Basis** is set to “User Specify”.

- **Degree of Freedom** – Specifies the degree of freedom. Larger value of degree of freedom lead to fewer spline knots and thus smoother function estimates.
  - **Alpha** – Specifies the number of knots considered for each variable. The value must be from 0 to 1.
- **Penalty** – Specifies the penalty for increasing number of variables in the multivariate adaptive regression spline model.
- **Probability Distribution** – Specifies the probability distribution of Generalized Linear Model. Normal is for interval target by default, Binary for classification if character variable.
  - **Default:** the Normal distribution for continuous response variables and to the Binary distribution for classification or character variables
  - **Poisson**
  - **Negative Binomial**
  - **Gamma**
  - **Binary**
  - **Normal**
- **Link Function** – Specifies the probability distribution of Generalized Linear Model. Normal is for interval target by default, Binary for classification if character variable.
  - **Default:** corresponding to the probability distribution
  - **Log**
  - **Reciprocal**
  - **Identity**
  - **Logit**
  - **Probit**
  - **Power with exponent -2**
  - **Complementary log-log**
- **Selection Method** – Specifies the method of selection process. The default algorithm of MARS contains two stages: forward selection and backward selection. During the forward selection process, bases are created from interactions between existing parent bases and nonparametric transformation of continuous or classification variables as candidate effects. After the model grows to a certain size, the backward selection process begins by deleting selected based. The deletion continues until the null model is reached, and then the overall best model is chosen based on some goodness-of-fit criterion. The Forward Only selection skips the backward selection step after forward selection is finished.
- **Use Fast Algorithm** – The fast algorithm improves the speed of the forward selection by tuning several parameters.
- **Cross Validation** – Specifies whether to perform cross validation.
- **Number of Folds** – Specifies the number of cross validation fold when **Cross Validation** is set to “Yes”.
- **Random Seed** – Specifies the seed to start the pseudorandom number generator for random cross validation when **Cross Validation** is set to “Yes”. If 0 is specified, the seed is generated from the time of day, which is read from the computer's clock.

- **Output Design Matrix** – Specifies whether to create a data set that contains the design matrix of constructed basis functions.
- **Selected Model** – Specifies the selected model to produce the design matrix when Output Design Matrix is set to “Yes”.
  - **After Backward Selection**
  - **After Forward Selection**
  - **Initial Model**
- **Exclude Rejected Variable** – Excluded Rejected Variable" description="Specifies what action should be taken for variables excluded from the final model. This option is only in effect when using a variable selection method. When set to “None”, the roles of these variables remain unchanged. When set to Hide, these variables are dropped from the metadata exported by the node. When set to “Reject”, the roles of these variables are set to REJECTED.


### MARS Node Example

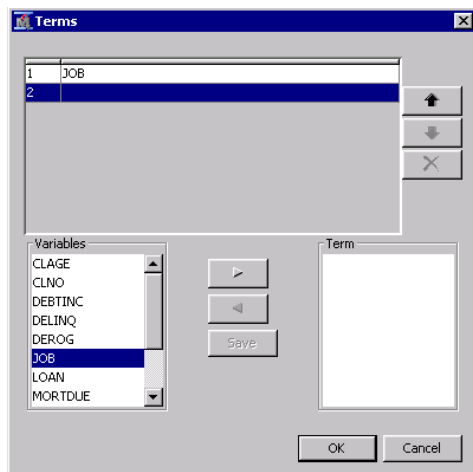
This example uses the sample SAS data set SAMPPIO.HMEQ. You must use the data set to create a SAS Enterprise Miner Data Source. Right-click the **Data Sources** folder in the Project Navigator and select **Create Data Source** to launch the Data Source wizard.

- Select **SAS Table** as your metadata source and click **Next**.
- Enter SAMPPIO.HMEQ in the Table field and click **Next**.
- Continue to the Metadata Advisor step and select the **Basic Metadata Advisor**.
- In the Column Metadata window, set the role of the variable Value to **Target** and set the level of the variable Value to **Interval**. Click **Next**.
- There is no decision processing. Click **Next**.
- In the Create Sample window, you are asked if you want to create a sample data set. Select **No**. Click **Next**.
- Set the role of the HMEQ data set to **Train**, and then click **Finish**.

Drag the HMEQ data set and the MARS node to your diagram workspace. Connect them as shown in the diagram below.



Select the  button next to the **Keep Effects** property to open a term editor. Specify variable **Job** to be included in the final model as shown in the diagram below, and then click **OK**.



Run the MARSnode with other settings as default by right-clicking on the MARS node and selecting **Run**. In the Confirmation window, select **Yes**. After a successful run of the MARSnode, select **Results** in the Run Status window.

Notice the following information:

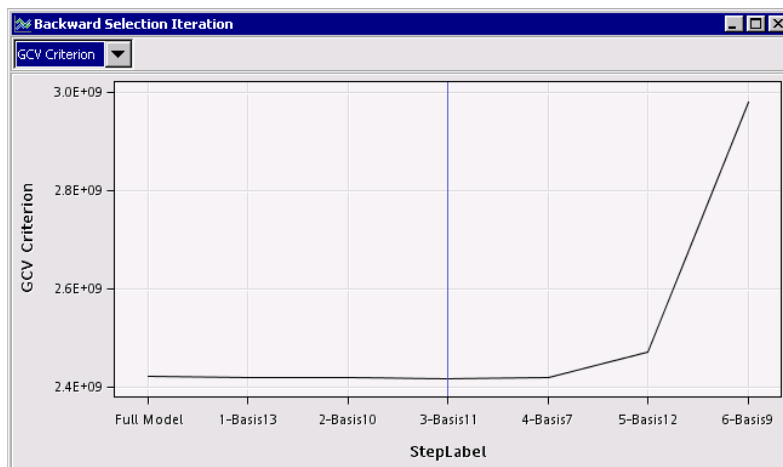
**Bases Transformation Information** is a table of the transformations that are used to generate the basis matrix. The first basis function, Basis0, is the intercept. The second basis function, Basis1, is 1 when variable Job has level 'Sales' and 0 otherwise. The eleventh basis function, Basis11, is  $\text{Loan} - 40800$  when  $\text{loan} > 40800$  and 0 otherwise, and 40800 here is a knot value. Other basis functions are constructed in a similar manner by using other knot values. The knots are chosen automatically.

Bases Transformation Information	
Name	Transformation
Basis0	1
Basis1	$\text{Basis0} * (\text{JOB} = \text{'Sales'})$
Basis2	$\text{Basis0} * (\text{JOB} = \text{'Self'})$
Basis3	$\text{Basis0} * (\text{JOB} = \text{'Mgr'})$
Basis4	$\text{Basis0} * (\text{JOB} = \text{'Office'})$
Basis5	$\text{Basis0} * (\text{JOB} = \text{'ProfExe'})$
Basis6	$\text{Basis0} * (\text{JOB} = \text{'Other'})$
Basis7	$\text{Basis0} * \text{NOT}(\text{MISSING}(\text{MORTDUE}))$
Basis8	$\text{Basis0} * \text{MISSING}(\text{MORTDUE})$
Basis9	$\text{Basis7} * \text{MAX}(\text{MORTDUE} - 31807, 0)$
Basis10	$\text{Basis7} * \text{MAX}(31807 - \text{MORTDUE}, 0)$
Basis11	$\text{Basis0} * \text{MAX}(\text{LOAN} - 40800, 0)$
Basis12	$\text{Basis0} * \text{MAX}(40800 - \text{LOAN}, 0)$
Basis13	$\text{Basis0} * \text{MAX}(\text{LOAN} - 3800, 0)$
Basis14	$\text{Basis0} * \text{MAX}(3800 - \text{LOAN}, 0)$

**Parameter Estimates** is a table of parameter estimates and the selected variables.

Parameter Estimates					
Name	Coefficient	Parent	Variable	Knot	Levels
Basis0	92186		Intercept		
Basis1	16264		JOB		.0
Basis2	57548		JOB		.1
Basis3	16442		JOB		.2
Basis4	1895.75		JOB		.3
Basis5	38932		JOB		.4
Basis6	0		JOB		.5
Basis7	-10643Basis0		MORTDUE		
Basis9	1.1757Basis7		MORTDUE	31807	
Basis12	-1.2250Basis0		LOAN	40800	

**Backward Selection Iteration** is a plot displays the progression of the backward elimination phase. The GCV criterion provides an estimate of how well the model will perform with new data, so the final model should have good predictive power. The figure below shows that the backward elimination step eliminates basis functions 13, 10, and 11.



**ANOVA** is an Analysis of Variance (ANOVA) table for the target variable.

**Classification Variables** is a table of classification variable levels information.

**Fit Control Parameters** is a table of parameters of spline fitting controls.

**Fit Statistics** is a table of the fit statistics from the model.

**Model Information** is a table of MARS model settings.

**Variable Importance** is a table of input variables, scaled by their relative importance as predictors for the target variable.

**Dependent Variable vs. Fitted Values** is a plot displays the raw dependent variable overlaid with the fitted values. This plot is not produced for dependent variable with nonnormal distribution.

**Residuals vs. Fitted Values** is a plot displays the residuals overlaid with the fitted values. This plot is not produced for dependent variable with nonnormal distribution.

\*Note: Special thanks to Paal Navestad, Senior Data Scientist @ ConocoPhillips for providing valuable feedbacks on this article.