

SAS[®] Enterprise Miner[™] High-Performance Data Mining Node Reference for SAS 9.3[®]

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2011. *SAS® Enterprise Miner™ High-Performance Data Mining Node Reference for SAS 9.3®*. Cary, NC: SAS Institute Inc.

SAS® Enterprise Miner™ High-Performance Data Mining Node Reference for SAS 9.3®

Copyright © 2011, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hardcopy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, December 2011

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/publishing or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

Chapter 1 • The HP Explore Node	1
HP Explore Node	1
Chapter 2 • The HP Impute Node	5
HP Impute Node	5
Chapter 3 • The HP Neural Node	15
HP Neural Node	15
Chapter 4 • The HP Regression Node	23
HP Regression Node	23
Chapter 5 • The HP Transform Node	37
HP Transform Node	37
Chapter 6 • The HP Variable Selection Node	43
HP Variable Selection Node	43

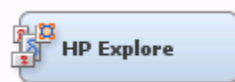
Chapter 1

The HP Explore Node

HP Explore Node	1
Overview of the HP Explore Node	1
HP Explore Node Properties	1
HP Explore Node Results	3

HP Explore Node

Overview of the HP Explore Node




The HP Explore node enables you to obtain descriptive information about a training data set. Statistics such as mean, standard deviation, minimum value, maximum value, and percentage of missing values are generated for each input variable and displayed in tabular and graphical form. These descriptive statistics are an important tool in the pre-model building process. For example, if the HP Explore node identifies variables with missing values, you can impute these values using the HP Impute node.


HP Explore Node Properties

HP Explore Node General Properties


The following general properties are associated with the HP Explore node:

- **Node ID** — The Node ID property displays the ID that SAS Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first HP Explore node that is added to a diagram will have a Node ID of HPEXpl. The second HP Explore node added to a diagram will have a Node ID of HPEXpl2, and so on.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Explore window. The Imported Data — HP Explore window contains a list of the ports that provide data sources to the HP Explore node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following:


- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Explore window. The Exported Data — HP Explore window contains a list of the output data ports that the HP Explore node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes, such as data or configuration information.

HP Explore Node Train Properties

The following train properties are associated with the HP Explore node:

- **Variables** — specifies the properties of each variable in the data source that you want to use. Select the  button to the right of the Variables property to open a variables table. You can set the variable status to either **Use** or **Don't Use** in the table, and you can set a variable's report status to **Yes** or **No**.
- **Summary Statistics** — select **Yes** to display the summary statistics for each of the input variables.
- **Max Num Levels** — specifies the maximum number of variable measurement levels. Variables with more levels than the value specified here are rejected. You can specify only positive integers in this property.

HP Explore Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.
- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.
- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.

- **User-Added Node** — specifies if the node was created by a user as a SAS Enterprise Miner extension node.

HP Explore Node Results

After the HP Explore node successfully runs, you can open the Results — Explore window by right-clicking the node in the Diagram Workspace and selecting Results from the pop-up menu. The Results — Explore window contains an imputation summary table and a window that displays the node's output.

Select **View** from the main menu of the Explore Results window to view the following information:

- **Properties**
 - **Settings** — displays a window with a Read-Only table of the HP Explore node properties configuration when the node was last run. Use the Show Advanced Properties check box at the bottom of the window to see all of the available properties.
 - **Run Status** — indicates the status of the HP Explore node run. The Run Start Time, Run Duration, and information about whether the run completed successfully are displayed in this window.
 - **Variables** — a table of variables properties in the node.
 - **Train Code** — the code that SAS Enterprise Miner used to train the node.
 - **Notes** — enables users to read or create notes.
- **SAS Results**
 - **Log** — the SAS log of the HP Explore node run.
 - **Output** — the SAS output of the HP Explore node run. The SAS output displays how many and which types of variables are in the training data set. If you run in a grid environment, the output displays limited information about the connection to the grid.
 - **Flow Code** — the SAS code that is used to produce the output that the HP Explore node passes on to the next node in the process flow diagram.
- **Scoring**
 - **SAS Code** — the SAS score code that was created by the node. The SAS score code can be used outside of the SAS Enterprise Miner environment in custom user applications.
 - **PMML Code** — The HP Explore node does not generate PMML code.
- **Summary Statistics**
 - **Class Plot** — is created for each input class variable. The Class Plot displays the frequency of each measurement level.
 - **Interval Plot** — displays the descriptive statistics for each of the interval inputs. The **Subset** menu enables you to select the following statistics:
 - **Coefficient of Variation**
 - **Kurtosis**
 - **Maximum**
 - **Mean**

- **Minimum**
- **Number of Missing Values**
- **Number of Non-Missing Values**
- **Percentage of Missing Values**
- **Skewness**
- **Standard Deviation**
- **Missing Values** — displays the total number of missing observations for each input variable.
- **Statistics Table** — provides a summary of the generated statistics for each variable. The information available in this table is as follows:
 - **Data Role** — the role of the data.
 - **Scale** — the level of the variable. Interval variables are assigned the label **VAR**, and class variables are assigned the label **CLASS**.
 - **Variable** — the name of the input variable.
 - **Missing** — the number of missing observations.
 - **Percent Missing** — the percentage of missing observations.
 - **Non Missing** — the number of nonmissing observations.
 - **Minimum** — the minimum value of the input variable.
 - **Mean** — the mean value of the input variable.
 - **Maximum** — the maximum value of the input variable.
 - **Standard Deviation** — the standard deviation of the input variable.
 - **Skewness** — the measure of skewness of the input variable.
 - **Kurtosis** — the kurtosis, also called steepness, of the input variable.
 - **Coefficient of Variation** — the standard deviation divided by the mean, a unit-less measure.
 - **Number of Levels** — the number of distinct levels for each class variable.
 - **Mode** — the modal value for each class variable.
 - **Mode Percent** — the percentage of observations that equal the mode.
- **Table** — opens the data table that corresponds to the graph that you have in focus.
- **Plot** — opens the Select a Chart plotting wizard so that you can plot the data in the table that you have in focus.

Chapter 2

The HP Impute Node

HP Impute Node	5
Overview of the HP Impute Node	5
HP Impute Node Properties	6
Using Indicator Variables with the HP Impute Node	10
Imputation Methods	11
HP Impute Node Results	12

HP Impute Node

Overview of the HP Impute Node



Use the HP Impute node to replace missing values in data sets that are used for data mining. The HP Impute node is typically used during the modification phase of the Sample, Explore, Modify, Model, and Assess (SEMMA) SAS data mining methodology.

Data mining databases often contain observations that have missing values for one or more variables. Missing values can result from the following: data collection errors; incomplete customer responses; actual system and measurement failures; or from a revision of the data collection scope over time (such as tracking new variables that were not included in the previous data collection schema).

If an observation contains a missing value, then by default that observation is not used for modeling by nodes such as Neural Network, or Regression. However, rejecting all incomplete observations might ignore useful or important information that is still contained in the nonmissing variables. Rejecting all incomplete observations might also bias the sample, since observations that are missing values might have other things in common as well.

Choosing the "best" missing value replacement technique inherently requires the researcher to make assumptions about the true (missing) data. For example, researchers often replace a missing value with the mean of the variable. This approach assumes that the variable's data distribution follows a normal population response. Replacing missing values with the mean, median, or another measure of central tendency is simple, but it can greatly affect a variable's sample distribution. You should use these replacement statistics carefully and only when the effect is minimal.

Another imputation technique replaces missing values with the mean of all other responses given by that data source. This assumes that the input from that specific data source conforms to a normal distribution. Another technique studies the data to see whether the missing values occur in only a few variables. If those variables are determined to be insignificant, the variables can be rejected from the analysis. The observations can still be used by the modeling nodes.

The HP Impute node provides the following imputations for missing interval variables:

- Default Constant Value
- Mean
- Maximum
- Minimum
- Midrange
- None

The HP Impute node provides the following imputations for missing class variables:

- Count
- Default Constant Value
- Distribution
- None


You can customize the default imputation statistics by specifying your own replacement values for missing and nonmissing data. You replace missing values for the training, validation, test, and score data sets by using imputation statistics that are calculated from the active training predecessor data set.

The HP Impute node must follow an Input Data, SAS Code, or other high-performance node.

HP Impute Node Properties


HP Impute Node General Properties

The following general properties are associated with the HP Impute node:


- **Node ID** — The Node ID property displays the ID that Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first HP Impute node that is added to a diagram will have a Node ID of HPImp. The second HP Impute node added to a diagram will have a Node ID of HPImp2, and so on.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Impute window. The Imported Data — HP Impute window contains a list of the ports that provide data sources to the HP Impute node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.


- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Impute window. The Exported Data — HP Impute window contains a list of the output data ports that the HP Impute node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes, such as data or configuration information.

HP Impute Node Train Properties

The following train properties are associated with the HP Impute node:

- **Variables** — specifies the properties of each variable in the data source that you want to use. Select the  button to the right of the Variables property to open a variables table. You can set the variable status to either **Use** or **Don't Use** in the table, and you can set a variable's report status to **Yes** or **No**.
- **Non Missing Variables** — Use the Non Missing Variables property of the HP Impute node to specify if you want to impute variables with no missing values. The default setting for the Non Missing Variables property is **No**.
- **Missing Cutoff** — Use the Missing Cutoff property of the HP Impute node to specify the maximum percent of missing values that are allowed for a variable to be imputed. Variables whose percentage of missing exceeds this cutoff are ignored. The default setting for the Missing Cutoff property is 50%.

HP Impute Node Train Properties: Class Variables

- **Default Input Method** — Use the Default Input Method property of the HP Impute node to specify the imputation statistic that you want to use to replace missing class variables. The choices are as follows:
 - **Count** — Use the Count setting to replace missing class variable values with the most frequently occurring class variable value.
 - **Default Constant Value** — Use the Default Constant setting to replace missing class variable values with the value that you enter in the Default Character Value property.
 - **Distribution** — Use the Distribution setting to replace missing class variable values with replacement values that are calculated based on the random percentiles of the variable's distribution. In this case, the assignment of values is based on the probability distribution of the nonmissing observations. The distribution imputation method typically does not change the distribution of the data very much.

- **None** — Missing class variable values are not imputed under the None setting.
- **Default Target Method** — Use the Default Target Method property of the HP Impute node to specify the imputation statistic that you want to use to replace missing class target variables. The choices are:
 - **Count** — Use the Count setting to replace missing target variable values with the most frequently occurring target variable value.
 - **Default Constant Value** — Use the Default Constant setting to replace missing target variable values with the value that you enter in the Default Character Value property.
 - **Distribution** — Use the Distribution setting to replace missing target variable values with replacement values that are calculated based on the random percentiles of the variable's distribution. In this case, the assignment of values is based on the probability distribution of the nonmissing observations. The distribution imputation method typically does not change the distribution of the data very much.
 - **None** — Missing target variable values are not imputed under the None setting.

HP Impute Node Train Properties: Interval Variables

- **Default Input Method** — Use the Method Interval property of the HP Impute node to specify the imputation statistic that you want to use to replace missing interval variables. The choices are:
 - **Mean** — Use the Mean setting to replace missing interval variable values with the arithmetic average, calculated as the sum of all values divided by the number of observations. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean. The mean is the preferred statistic to use to replace missing values if the variable values are at least approximately symmetric (for example, a bell-shaped normal distribution). Mean is the default setting for the Default Input Method for interval variables.
 - **Maximum** — Use the Minimum setting to replace missing interval variable values with the maximum value for the variable.
 - **Minimum** — Use the Minimum setting to replace missing interval variable values with the minimum value for the variable.
 - **Midrange** — Use the Midrange setting to replace missing interval variable values with the maximum value for the variable plus the minimum value for the variable divided by two. The midrange is a rough measure of central tendency that is easy to calculate.
 - **Default Constant Value** — Use the Default Constant Value setting to replace missing interval variable values with the value that you enter in the Default Character Value property.
 - **None** — Specify the None setting if you do not want to replace missing interval variable values.
- **Default Target Method** — Use the Default Target Method property of the HP Impute node to specify the imputation statistic that you want to use to replace missing target variables. The choices are:
 - **Mean** — Use the Mean setting to replace missing target variable values with the arithmetic average, calculated as the sum of all values divided by the number of observations. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean. The mean is the preferred statistic to use to replace missing values if the variable values are at

least approximately symmetric (for example, a bell-shaped normal distribution). Mean is the default setting for the Default Target Method for interval variables.

- **Maximum** — Use the Minimum setting to replace missing interval variable values with the maximum value for the variable.
- **Minimum** — Use the Minimum setting to replace missing interval variable values with the minimum value for the variable.
- **Midrange** — Use the Midrange setting to replace missing target variable values with the maximum value for the variable plus the minimum value for the variable divided by two. The midrange is a rough measure of central tendency that is easy to calculate.
- **Default Constant Value** — Use the Default Constant Value setting to replace missing target variable values with the value that you enter in the Default Character Value property.
- **None** — Specify the None setting if you do not want to replace missing target variable values.

HP Impute Node Train Properties: Default Constant Value

Use the Default Constant properties to specify how you want to manage numeric and character constant values to be used during imputation.

- **Default Character Value** — Use the Default Character Value property of the HP Impute node to specify the character string or value that you want to use during constant character imputation.
- **Default Number Value** — Use the Default Number Value property of the HP Impute node to specify the numeric value that you want to use as a constant value for numeric imputation. The Default Number Value property defaults to a setting of 0.0.

HP Impute Node Train Properties: Method Options

- **Random Seed** — Use the Random Seed property of the HP Impute node to specify the initial Random Seed value that you want to use for random number generation. The default Random Seed value is 12345.

HP Impute Node Score Properties

The following score properties are associated with the HP Impute node:

- **Hide Original Variables** — Set the Hide Original Variables property of the HP Impute node to **No** if you want to keep the original variables in the exported metadata from your imputed data set. In that case, the variables are exported with a role of Rejected. The default setting for the Hide Original Variables is **Yes**. When Set to **Yes**, the original variables are removed only from the exported metadata, and not removed from the exported data sets and data views.

HP Impute Node Score Properties: Indicator Variables

- **Type** — Use the Type property of the HP Impute node to specify the type of indicator variables used to flag the imputed observations for each variable.

You can choose from the following settings:

- **None** — (default setting) Do not create an indicator variable.
- **Single** — A single indicator variable is created that indicates one or more variables were imputed.

- **Unique** — Unique binary indicator variables are created for every imputed variable.
- **Source** — Use the Indicator Variable Source property of the HP Impute node to specify the role that you want to assign to the created indicator variables. You can choose between **Missing Variables** and **Imputed Variables**. The default setting is **Imputed Variables**.
- **Role** — Use the Indicator Variable Role property of the HP Impute node to specify the role that you want to assign to the created indicator variables. You can choose between the roles **Rejected** and **Input**. The default setting is **Rejected**.

HP Impute Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.
- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.
- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.
- **User-Added Node** — specifies if the node was created by a user as a SAS Enterprise Miner extension node.

Using Indicator Variables with the HP Impute Node

Use the Indicator Variables property to create flag variables that identify each variable's imputed observations. Select the check box to create flag variables for each imputed input variable. Flag variables are named by concatenating the prefix **M_** with the input variable's name. Each flag variable contains an indicator value of 1 if the observation was imputed and an indicator value of 0 if it was not. For example, assume that you have two input variables named **AGE** and **HEIGHT** that contain the following values prior to imputation:

Age	Height
25	73
.	66
30	.

When you run the node, imputed flag variable(s) are created, depending on the setting that is configured for the Indicator Variable property:

Table 2.1 Indicator Variable Setting Is Set to *UNIQUE*

Age	Height	M_AGE	M_HEIGHT
25	73	0	0
34	66	1	0
30	62	0	1

Table 2.2 Indicator Variable Setting Is Set to *SINGLE*

Age	Height	M_VARIABLE
25	73	0
34	66	1
30	62	1

Imputation Methods

Interval Imputation Methods

To set the default interval imputation statistic, click the **Default Input Method** or **Default Target Method** property drop-down arrow and select one of the following imputation methods:

- Mean — (default) or the arithmetic average, calculated as the sum of all values divided by the number of observations. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean. The mean is the preferred statistic to use to replace missing values if the variable values are at least approximately symmetric (for example, a bell-shaped normal distribution).
- Maximum — all replacements are equal to the maximum value for the variable.
- Minimum — all replacements are equal to the minimum value for the variable.
- Mid-range — the maximum plus the minimum divided by two. The midrange is a rough measure of central tendency that is easy to calculate.
- Default Constant — all replacements are equal to a fixed constant value. You configure the fixed constant value using the **Default Character Value** and the **Default Number Value** properties.
- None — do not impute the missing values.

Class Imputation Statistics

Missing values for class variables can be replaced with one of the following statistics:

- Count — missing values are replaced with the modal value.

- **Default Constant Value** — all replacements are equal to a fixed constant value. You configure the fixed constant value using the Default Character Value and the Default Number Value properties.
- **Distribution** — replacement values are calculated based on the random percentiles of the variable's distribution. In this case, the assignment of values is based on the probability distribution of the nonmissing observations. This imputation method typically does not change the distribution of the data very much.
- **None** — replacement values are not imputed, and are left as missing.

If several values occur with the same frequency and Count is the class imputation statistic, the smallest value is used as the Count. If the Count is a missing value, then the next most frequently occurring nonmissing value is used for data replacement.

HP Impute Node Results

After the HP Impute node successfully runs, you can open the Results — Impute window by right-clicking the node in the Diagram Workspace and selecting Results from the pop-up menu. The Results — Impute window contains an imputation summary table and a window displaying the node's output.

Select **View** from the main menu of the Impute Results window to view the following information:

- **Properties**
 - **Settings** — displays a window with a read-only table of the HP Impute node properties configuration when the node was last run. Use the Show Advanced Properties check box at the bottom of the window to see all of the available properties.
 - **Run Status** — indicates the status of the HP Impute node run. The Run Start Time, Run Duration, and information about whether the run completed successfully are displayed in this window.
 - **Variables** — a table of variables property of the node.
 - **Train Code** — the code that Enterprise Miner used to train the node.
 - **Notes** — allows users to read or create notes of interest.
- **SAS Results**
 - **Log** — the SAS log of the HP Impute node run.
 - **Output** — the SAS output of the HP Impute node run. The SAS output includes a variable summary, a distribution of missing observations in training data table, and an imputation summary by variable.
 - **Flow Code** — the SAS code used to produce the output that the HP Impute node passes on to the next node in the process flow diagram.
- **Scoring**
 - **SAS Code** — the SAS score code that was created by the node. The SAS score code can be used outside of the Enterprise Miner environment in custom user applications.
 - **PMML Code** — the HP Impute node does not generate PMML code.
- **Model**
 - **Imputation Summary** — a table of that shows a summary of the imputation run. The Imputation Summary table displays the variable name, the imputation

method, the imputed variable name, the variable role, the variable level, the variable type (numeric or character), the variable label (if any), and the number of missing values for the train data set, the validation data set, and the test data set. The Imputation Summary table is a subset of the SAS output from the HP Impute node run.

- **Table** — open the data table that corresponds to the graph that you have in focus.
- **Plot** — opens the Select a Chart plotting wizard so that you can plot the data in the table that you have in focus.

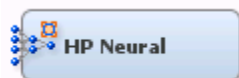
Chapter 3

The HP Neural Node

HP Neural Node	15
Overview of the HP Neural Node	15
Data Requirements of the HP Neural Node	16
HP Neural Node Properties	16
HP Neural Node Results	18
HPDM Assessment	20
HP Neural Node Example	21

HP Neural Node

Overview of the HP Neural Node



The HP Neural node creates multilayer neural networks that pass information from one layer to the next in order to map an input to a specific category or predicted value. The HP Neural node enables this mapping to take place in a distributed computing environment, which enables you to build neural networks on massive data sets in a relatively short amount of time.

A neural network consists of units (neurons) and connections between those units. There are three types of units:

- Input Units — obtain the values of input variables and standardize those values.
- Hidden Units — perform internal computations and provide the nonlinearity that makes neural networks powerful.
- Output Units — compute predicted values and compare those values with the values of the target variables.

Units pass information to other units through connections. Connections are directional and indicate the flow of computation within the network. Connections cannot form loops, because the HP Neural node permits only feed-forward networks. The following restrictions apply to feed-forward networks:

- Input units can be connected to hidden units or to output units.
- Hidden units can be connected to other hidden units or to output units.

- Output units cannot be connected to other units.

Each unit produces a single computed value. For input and hidden units, this computed value is passed along the connections to other hidden or output units. For output units, the computed value is the predicted value. The predicted value is compared with the target value to compute the error function, which the training method attempts to minimize.

The HP Neural node was designed with two goals. First, the HP Neural node aims to perform efficient, high-speed training of neural networks. Second, the HP Neural node attempts to create accurate, generalizable models in an easy to use manner. With these goals in mind, most parameters for the neural network are selected automatically. This includes standardization of input and target variables, activation and error functions, and termination of model training.


Data Requirements of the HP Neural Node

The HP Neural node requires one or more input variables and one or more target variables. The inputs and targets can be binary, nominal, or interval. All ordinal variables are ignored during training. If an observation has missing values for any of the specified target variables, the observation is not used for training or for computing validation error.


HP Neural Node Properties

HP Neural Node General Properties


The following general properties are associated with the HP Neural node:

- **Node ID** — The Node ID property displays the ID that SAS Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first HP Neural node that is added to a diagram has a Node ID of HPDMNeural. The second HP Neural node added to a diagram has a Node ID of HPDMNeural2, and so on.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Neural window. The Imported Data — HP Neural window contains a list of the ports that provide data sources to the HP Neural node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and click for the desired option:


- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a **Table** tab and a **Variables** tab. The tabs contain summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Neural window. The Exported Data — HP Neural window contains a list of the output data ports that the HP Neural node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.

If data exists for an imported data source, you can select the row in the imported data table and click:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a **Table** tab and a **Variables** tab. The tabs contain summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.

HP Neural Node Train Properties

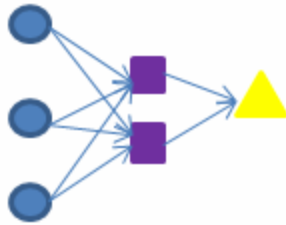
The following train properties are associated with the HP Neural node:

- **Variables** — specifies the properties of each variable in the data source that you want to use. Select the  button to the right of the Variables property to open a variables table. You can set the variable status to either **Use** or **Don't Use** in the table, and you can set a variable's report status to **Yes** or **No**.
- **Use Inverse Priors** — specifies whether inverse priors are used to weight training observations in the presence of a class target. This is especially helpful when you have rare events. When you specify **Yes**, a weight is calculated as the inverse of the fraction of time that the target class occurs in the input data set. It is applied to the prediction error of each nominal target variable. The default value for this property is **No**.
- **Create Validation** — specifies whether a validation data set is created. When you specify **Yes**, the incoming data is split into training and validation subsets to control when training stops and to prevent over-fitting. The validation data set contains every fourth observation from the input data set, starting with the first observation.

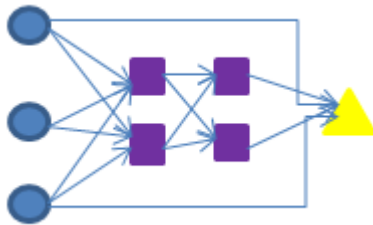
There are cases when it is interesting to determine how well a neural network can fit a set of data with a particular number of neurons. This is accomplished by setting **Create Validation** to **No**. In this case, training does not stop when the validation error no longer improves. It continues until the maximum number of iterations have been met, or until no further improvement in total training error is seen.

- **Number of Hidden Neurons** — specifies the number of hidden neurons within the network. The number of neurons is split equally across the hidden layers. A good strategy is to start with a small number of hidden neurons and slowly increase the number until the validation error stops improving.
- **Architecture** — specifies the network architecture that you want to use to train the neural network. You can specify **One Layer**, **One Layer with Direct**, **Two Layers**, or **Two Layers with Direct**. The options **One Layer with Direct** and **Two Layers with Direct** enable direct connections from the input units to the output units.

The simplest network that can be modeled is the One Layer network. Each input unit is connected to each hidden unit and each hidden unit is connected to the output unit, as demonstrated in the image below.



The most complex network available is the Two Layers with Direct. This adds an additional hidden layer to the network as well as direct connections between inputs and targets, as demonstrated in the image below.



- **Number of Tries** — specifies the number of times, or tries, that the network is to be retrained, using a different set of initial weights with each try. Because training involves optimizing a nonlinear objective function, this provides one way to be reasonably sure that a good set of weights is found. The default value is **2**.
- **Maximum Iterations** — specifies the maximum number of iterations allowed within each try. The default value is **50**.

HP Neural Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.
- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.
- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.
- **User-Added Node** — specifies whether the node was created by a user as a SAS Enterprise Miner extension node.

HP Neural Node Results

After the HP Neural node successfully runs, you can open the Results — Neural window. Right-click the node in the Diagram Workspace and select **Results** from the pop-up menu. The Results — Neural window contains an imputation summary table and a window that displays the node's output.

Select **View** from the main menu of the Neural Results window to view the following information:

- **Properties**
 - **Settings** — displays a window with a read-only table of the HP Neural node properties configuration when the node was last run. Use the **Show Advanced Properties** check box at the bottom of the window to see all of the available properties.
 - **Run Status** — indicates the status of the HP Neural node run. The Run Start Time, Run Duration, and information about whether the run completed successfully are displayed in this window.
 - **Variables** — a table of variables property of the node.
 - **Train Code** — the code that SAS Enterprise Miner used to train the node.
 - **Notes** — enables users to read or create notes of interest.
- **SAS Results**
 - **Log** — the SAS log of the HP Neural node run.
 - **Output** — the SAS output of the HP Neural node run. The SAS output includes a variable summary, a distribution of missing observations in the training data table, and an imputation summary by variable.
 - **Flow Code** — the SAS code used to produce the output that the HP Neural node passes on to the next node in the process flow diagram.
- **Scoring**
 - **SAS Code** — the SAS score code that was created by the node. The SAS score code can be used outside of the SAS Enterprise Miner environment in custom user applications.
 - **PMML Code** — the HP Neural node does not generate PMML code.
- **Assessment**
 - **Fit Statistics**
 - **Classification Chart**
 - **Score Rankings Overlay**
 - **Score Distribution**
- **Model**
 - **Model Information** — includes information about the model itself. This includes statistics on the number of inputs and targets as well as details about the generated network.
 - **Misclassification Plot** — displays the percentage of correct and incorrect percentages for the different levels of the class target variable.
 - **Class Levels** — displays the number of measurement levels and the values of those levels for each of the class variables.
 - **Number of Observations** — provides statistics on the number of observations read and used as well as how the original observations were split across training and validation tables.
 - **Performance Information** — displays the host, execution mode, and number of threads used by the HP Neural node.

- **Link Graph** — graphically displays the neural network architecture. All input variables are listed on the far left of the Link Graph. The target variable is placed at the far right of the Link Graph. Hidden units are found between the input variables and the target variable. The width and color of the lines indicate the magnitude of the weight for that particular connection. The thinner, blue values indicate a smaller magnitude link weight, and the thicker, red values indicate a larger magnitude link weight.
- **Weights** — displays the value of the weights for each connection. As in the Link Graph table, the colors represent the magnitude of the weight.
- **Training History** — summarizes each try made during retraining. For each try, the report contains the number of completed iterations, the root mean square error for training and validation, the reason for stopping, and which try was identified as the best.
- **Iteration History Plot** — shows the distribution of the root mean square training and validation error across all iterations for the best try that was identified in the Training History table.
- **Table** — opens the data table that corresponds to the graph that you have in focus.
- **Plot** — opens the Select a Chart plotting wizard so that you can plot the data in the table that you have in focus.

HPDM Assessment

Overview

The High-Performance Data Mining (HPDM) assessment plots display a range of rank order statistics for model assessment of the HPDM models. The HPDM assessment statistics are different from the regular SAS Enterprise Miner assessment statistics when you run HPDM models using data sets in a grid mode. The HPDM assessment statistics are computed on the whole data set, and the regular SAS Enterprise Miner assessment statistics are computed on a sample of the data set. When running HPDM models in solo mode, only the regular SAS Enterprise Miner assessment plots are displayed in the result window. More details about HPDM assessment can be found in “The %EM_new_assess Macro” in *SAS Enterprise Miner High-Performance Data Mining Procedures and Macro Reference for SAS 9.3* and the %HPDM_node_assess macro documentation.

HPDM Assessment Plots for Binary Target Variables

In the HPDM assessment for binary target variables, the data is binned by descending value of the nonmissing estimated probability of the event level for the binary target variable. For each bin (group) of observations, several statistics are computed and plotted against the depth of bins in the HPDM assessment plots. By default the horizontal axis of an HPDM assessment plot displays the depth (binning) of the observations, and the vertical axis of an HPDM assessment plot displays one or several statistics. The available plots are as follows:

- Lift and Cumulative Lift
- Event and Non-Event Rate
- Classification Rates — CR
- Separation Curve — KS
- Cumulative Captured Events
- Receiver-Operator Characteristic

HPDM Assessment Plots for Interval Target Variables

In the HPDM assessment for interval target variables, the data is binned by a descending order of the nonmissing predicted values of the interval target variable. For each bin (group) of observations, several statistics are computed and plotted against the depth of bins in the HPDM assessment plots. By default, the horizontal axis of an HPDM assessment plot displays the depth (binning) of the observations. The vertical axis of an HPDM assessment plot displays the actual target means, the predicted target means, and the residual means in each bin.

HPDM Bin Statistics Table

This table displays the summary and fit statistics for the binning of the target variable.

HP Neural Node Example

This example uses the sample SAS data set called `Sampsio.Hmeq`. You must use the data set to create a SAS Enterprise Miner Data Source. Right-click the **Data Sources** icon in the Project Navigator, and select **Create Data Source** to launch the Data Source wizard.

1. Choose SAS Table as your metadata source and click **Next**.
2. Enter `sampsio.hmeq` in the **Table** field and click **Next**.
3. Continue to the Metadata Advisor step, and choose the **Basic Metadata Advisor**.
4. In the Column Metadata window, set the role of the variable `Bad` to **Target** and set the level of the variable `Bad` to **Binary**. Click **Next**.
5. There is no decision processing. Click **Next**.
6. In the Create Sample window, you are asked if you want to create a sample data set. Select **No**. Click **Next**.
7. Set the role of the HMEQ data set to **Train**, and then click **Finish**.

Drag the HMEQ data set and the **HP Neural** node onto your diagram workspace. Connect them as shown in the diagram below.



Select the **HP Neural** node and make the following changes:

1. Set the **Number of Hidden Neurons** to 6.
2. Select **Two Layers with Direct** for the **Architecture** property.
3. Set the **Number of Tries** to 5.
4. Set the **Maximum Iterations** property to 10.

Right-click on the **HP Neural** node and select **Run** from the drop-down menu. In the Confirmation window, select **Yes**. After the HP Neural node has successfully run, select **Results** from the Run Status window. Notice the following results:

- The Model Information table indicates that the limited memory BFGS algorithm was used, and that there were 12 input variables, 1 output variable, 6 hidden neurons, and 87 weights.

- The Number of Observations table indicates that 2512 observations were used for training and 852 were used for validation.
- The Misclassification Plot shows a large number of observations were incorrectly assigned a target value of **1**, especially when compared to the number incorrectly assigned to **0**.
- The Link Graph table provides a visual representation of the neural network. Note that the thinner, blue values indicate a smaller link weight, and the thicker, red values indicate a larger link weight. In this neural network, there are six hidden neurons, three in each of the two hidden layers.
- The Weights graph provides an alternative view of the weights for each connection.
- The Training History table provides the root mean square error for both the training and validation data sets, in addition to the reason each try was stopped. In this example, the second try produced the best neural network.
- The Iteration History Plot graphs the root mean square error for the training and validation data against the iteration number. This plot corresponds to the second try. Notice that this graph ends at the tenth iteration because you specified **10** in the **Maximum Iterations** property, which is also the best iteration.

Chapter 4

The HP Regression Node

HP Regression Node	23
Overview of the HP Regression Node	23
HP Regression Node Properties	24
HP Regression Node Model Selection Options Tables	30
HP Regression Node Results	31
HPDM Assessment	34
HP Regression Node Example	35

HP Regression Node

Overview of the HP Regression Node



The HP Regression node fits a linear regression or a logistic regression for an interval or binary target variable that is specified in the training data set. Linear regression attempts to predict the value of an interval target as a linear function of one or more independent inputs. Logistic regression attempts to predict the probability that a binary target will acquire the event of interest based on a specified link function of one or more independent inputs.


The HP Regression node supports binary and interval target variables. For example, when modeling customer profiles, a variable named Purchase that indicates whether a customer made a purchase can be modeled as a binary target variable. If your customer profiles also contain a variable named Value that indicates the amount spent, this variable can be modeled as an interval target variable. The HP Regression node does not support the modeling of more than one target variable.

The Regression node supports forward, backward and stepwise selection methods for interval targets, and forward, backward, stepwise, LAR, and LASSO selection methods.


HP Regression Node Properties

HP Regression Node General Properties


The following general properties are associated with the HP Regression node:

- **Node ID** — The Node ID property displays the ID that SAS Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first HP Regression node that is added to a diagram will have a Node ID of HPReg. The second HP Regression node that is added to a diagram will have a Node ID of HPReg2, and so on.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Regression window. The Imported Data — HP Regression window contains a list of the ports that provide data sources to the HP Regression node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following tasks:


- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a **Table** tab and a **Variables** tab. The tabs contains summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Regression window. The Exported Data — HP Regression window contains a list of the output data ports that the HP Regression node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.

If data exists for an imported data source, you can select the row in the imported data table and do any of the following:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a **Table** tab and a **Variables** tab. The tabs contain summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes, such as data or configuration information.

HP Regression Node Train Properties

The following train properties are associated with the HP Regression node:

- **Variables** — specifies the properties of each variable in the data source that you want to use. Select the  button to the right of the Variables property to open a variables table. You can set the variable status to either **Use** or **Don't Use** in the table, and you can set a variable's report status to **Yes** or **No**.

HP Regression Node Train Properties: Equation


- **Main Effects** — Set the Main Effects property to **No** if you want to suppress the input and rejected variables with status of Use in the regression analysis. The default setting for this property is **Yes**.
- **Two-Factor Interactions** — Set the Two-Factor Interactions property to **Yes** if you want to include all two-factor interactions for class variables that have a status of Use. The default setting for this property is **No**.
- **Polynomial Terms** — Set the Polynomial Terms property to **Yes** if you want to include polynomial terms for interval variables with status of Use in the regression analysis. When this property is set to **Yes**, you must specify an integer value for the Polynomial Degree property. The default setting for Polynomial Terms is **No**.
- **Polynomial Degree** — When the Polynomial Terms property of the Regression node is set to **Yes**, use the Polynomial Degree property to specify the highest degree of polynomial terms (for interval variables with status set to **Use**) to be included in the regression analysis. The Polynomial Degree property can be set to 2 or 3.
- **Suppress Intercept** — Set the Suppress Intercept property to **Yes** to suppress intercepts when you are coding class variables. The Suppress Intercept property is ignored for ordinal targets. The default setting for the Suppress Intercept property is **No**.

Note: If **Main Effects**, **Two-Factor Interactions**, and **Polynomial Terms** are all set to **No**, then SAS Enterprise Miner will report an exception message. In the event where all three properties are set to **No**, there are no effects specified for the regression analysis.

HP Regression Node Train Properties: Modeling

- **Regression Type** — Use the Regression Type property to specify the type of regression that you want to run.
 - **Logistic Regression** — the default regression type for binary targets. For logistic regression, the event level of the binary target is determined by the sorting order of this variable in the preceding input data set node. The default sorting order of the binary target is **Descending**. For example, if the binary target has two levels, 1 and 0, then the HP Regression node chooses 1 as the event level and 0 as the non-event level.
 - **Linear Regression** — the default regression type for interval targets.
- **Link Function** — Use the Link Function property of the Regression node to specify the link function that you want to use in your regression analysis. Link functions link the response mean to the linear predictor. In a linear regression, the identity link function $g(M) = X\beta$ is used.

In a logistic regression, you can select one of the link functions:

- **Complementary log-log**
- **Logit** (default)
- **Log-Log**
- **Probit**
- **Optimization Options** — specifies the optimization options for the regression model. Select the  button to the right of the Optimization Options property to open the Optimization Options window.

The following properties are available in the Optimization Options window:

- **Optimization Technique** — specifies the optimization technique used by the HP Regression node.
 - **Conjugate-Gradient**
 - **Double-Dogleg**
 - **Newton-Raphson**
 - **Nelder-Mead Simplex**
 - **Newton-Raphson with Ridging** — default
 - **Qual Quasi-Newton**
 - **Trust-Region**
- **Maximum Number of Iterations** — Use the Maximum Number of Iterations property to specify the maximum number of iterations to be used in the optimization technique. To use the default value, leave the value as blank or a dot. The default value for the Maximum Number of Iterations property varies according to the selected optimization technique:

Optimization Technique	Default Max Iterations
Conjugate-Gradient	400
Double-Dogleg	200
Newton-Raphson	50
Nelder-Mead Simplex	1000
Newton-Raphson with Ridging	50
Qual Quasi-Newton	200
Trust-Region	50


- **Maximum Number of Function Calls** — Use the Maximum Number of Function Calls property of the HP Regression node to specify the maximum number of function calls to allow in the optimization technique. To use the default value, leave the value as blank or a dot.

Optimization Technique	Default Max Function Calls
Conjugate-Gradient	1000
Double-Dogleg	500
Newton-Raphson	125
Nelder-Mead Simplex	3000
Newton-Raphson with Ridging	125

Optimization Technique	Default Max Function Calls
Qual Quasi-Newton	500
Trust-Region	125

- **Maximum CPU Time in seconds** — specifies an upper limit of CPU time (in seconds) for the optimization process. The default value is the largest floating-point double representation of your computer. The time specified by this property is checked only once at the end of each iteration. Therefore, the actual run time can be longer than that which the property specifies. To use the default value, leave the value as blank or a dot.

Note: The **Maximum CPU Time in seconds** property governs only the optimization process time for the HPLOGISTIC procedure. It does not govern the maximum overall execution time for the HP Regression node.

- **Minimum Number of Iterations** — specifies the minimum number of iterations. The default value is 1. If you request more iterations than are actually needed for convergence, the optimization algorithms can behave unpredictably. To use the default value, leave the value as blank or a dot.
- **Normalize Objective Function** — Use the Normalize Objective Function property to determine whether the objective function should be normalized during the optimization. The reciprocal of the used frequency count is used for normalization. The default value is **Yes**.
- **Convergence Options** — specifies the convergence options for the regression model. Select the  button to the right of the Convergence Options property to open the Convergence Options window.

The following properties are available in the Convergence Options window:

- **Absolute Function Convergence** — specifies the threshold for absolute function convergence. The default value is the negative square root of the largest double precision value that is available on your computer. To use the default value, leave the value as blank or a dot.
- **Absolute Function Difference Convergence** — specifies the threshold for absolute function difference convergence. The default value is 0. To use the default value, leave the value as blank or a dot.
- **Absolute Gradient Convergence** — specifies the threshold for absolute gradient convergence. The default value is 1E-5. To use the default value, leave the value as blank or a dot.
- **Relative Function Difference Convergence** — specifies the threshold for relative function difference convergence. The default value is twice the machine precision. To use the default value, leave the value as blank or a dot.
- **Relative Gradient Convergence** — specifies the threshold for relative gradient convergence. The default value is 1E-8. To use the default value, leave the value as blank or a dot.

HP Regression Node Train Properties: Model Selection

This section details the model selection options that are available in the HP Regression node. The [Model Selection Options Table on page 30](#) provides a quick reference to determine which options are available for each model selection method.

- **Selection Model** — Use the Selection Model property to specify the model selection method that you want to use during training.

You can choose from the following effect selection methods:

- **Backward** — begins with all candidate effects in the model and removes effects until the Stay Significance Level or the Stop Criterion is met.
- **Forward** — begins with no candidate effects in the model and adds effects until the Entry Significance Level or the Stop Criterion is met.
- **Stepwise** — begins as in the forward model but might remove effects already in the model. Continues until Stay Significance Level or Stepwise Stopping Criteria are met.
- **None** — (default setting) uses all inputs to fit the model.
- **LAR** — performs least angle regression selection. This method, like forward selection, starts with no effects in the model and adds effects. The parameter estimates at any step are shrunk when compared to the corresponding least squares estimates. If the model contains classification variables, then these classification variables are split. The **LAR** selection method is supported only for interval target variables.
- **Lasso** — adds and deletes parameters based on a version of ordinary least squares, where the sum of the absolute regression coefficients is constrained. If the model contains classification variables, then these classification variables are split. The **Lasso** selection method is supported only for interval target variables.

Note: The **LAR** and **Lasso** methods are available only for interval target variables.

If you specify either **LAR** or **Lasso** for a training set with a binary target variable, then **None** will be used instead.

- **Selection Criterion** — Use this property to determine the order in which effects enter or leave at each step of the selection method. If you use either **LAR** or **Lasso** as the selection method, then you can specify only **Default** or **Significance Level** here.

The following section criteria are available:

- **Default** — uses the significance levels of the effects.
- **Adjusted R-square** — available only for linear regressions. If you have a binary target variable and specify this option, then **Default** is used instead.
- **AIC** — Akaike's Information Criterion
- **AICC** — Corrected Akaike's Information Criterion
- **SBC** — Schwarz Bayesian Information Criterion
- **Mallow's C_p** — Mallow's C_p statistic, available only for linear regressions. If you have a binary target variable and specify this option, then **Default** is used instead.
- **Significance Level** — uses the significance levels of the effects.


Note: If you select **Logistic Regression** as the **Regression Type**, then **Significance Level** is always used as the **Selection Criterion**. Similarly, if you select **LAR** or **Lasso** as the **Selection Model**, then **Significance Level** is always used as the **Selection Criterion**.

- **Stop Criterion** — specifies the criterion that is used to stop the selection process.

The following section criteria are available:

- **Default** — uses the significance levels of the effects.
- **Adjusted R-square** — available only for linear regressions.
- **AIC** — Akaike's Information Criterion
- **AICC** — Corrected Akaike's Information Criterion
- **SBC** — Schwarz Bayesian Information Criterion
- **Mallow's C_p** — Mallow's C_p statistic.
- **Significance Level** — uses the significance levels of the effects.
- **None** — no criterion for stopping selection is used. The selection process stops when no suitable addition or removal of candidate effects is found or if a size-based limit, such as **Maximum Number of Effects**, is reached.

Note: If you specify a criterion other than **Significance Level** or **None**, then the selection process stops when a local extremum is found or if a size-based limit is reached. The determination of whether a local minimum is achieved is made on the basis of a stop horizon at the next three steps.

- **Selection Options** — specifies the selection options for the regression model. Select the  button to the right of the Convergence Options property to open the Selection Options window.

The following selection options are available:

- **Entry Significance Level** — significance level for adding variables in forward, stepwise, LAR, or Lasso regression. The default value for the Entry Significance Level is 0.05. Values must be between 0 and 1.
- **Stay Significance Level** — significance level for removing variables in backward, stepwise, or LASSO regression. The default value for the Entry Significance Level is 0.05. Values must be between 0 and 1.
- **Maximum Number of Effects** — the maximum number of effects in any model considered during the selection process. This option is ignored with the backward regression. If a model at some step of the selection process contains the specified maximum number of effects, then no candidate effects are considered for addition. The default value is zero, which indicates that this option should be ignored.
- **Minimum Number of Effects** — the minimum number of effects in any model considered during the backward or stepwise selection process. For backward regression, the selection process terminates if a model at some step of the selection process contains the specified minimum number of effects. The default value is zero, which indicates that this option should be ignored.
- **Hierarchy** — specifies whether no variable, only class variables, or both class and interval variables are subject to hierarchy rules. You can specify **None** for no hierarchy rules, **Class Variables** for just class variables, and **All Variables** for class and interval variables.
- **Maximum Number of Steps** — Maximum number of selection steps that are performed. The default value is zero, which indicates that this option is ignored.

HP Regression Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.

- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.
- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.
- **User-Added Node** — specifies if the node was created by a user as a SAS Enterprise Miner extension node.

HP Regression Node Model Selection Options Tables

The two tables below provide a quick reference to determine which options are available for each model selection method. The available model selection methods are given in the first row and the model selection options are given in the first column. A Y indicates that the corresponding model selection option is available for that method, and an N indicates it is not available.

Table 4.1 Linear Regression Model Selection Options

	Forward	Backward	Stepwise	LAR	Lasso
Selection Criterion	Y	Y	Y	Significance Level	Significance Level
Stop Criterion	Y	Y	Y	Y	Y
Entry Significance Level	Y	N	Y	Y	Y
Stay Significance Level	N	Y	Y	N	Y
Maximum Number of Effects	Y	N	Y	Y	Y
Minimum Number of Effects	N	Y	N	N	N
Hierarchy	Y	Y	Y	N	N
Maximum Number of Steps	Y	Y	Y	Y	Y

Table 4.2 Logistic Regression Model Selection Options

	Forward	Backward	Stepwise
Selection Criterion	Significance Level	Significance Level	Significance Level
Stop Criterion	Y	Y	Y
Entry Significance Level	Y	N	Y
Stay Significance Level	N	Y	Y
Maximum Number of Effects	Y	N	Y
Minimum Number of Effects	N	Y	N
Hierarchy	Y	Y	Y
Maximum Number of Steps	Y	Y	Y

HP Regression Node Results

HP Regression Node Results Window

After a successful node run, you can open the Results window of the HP Regression node by right-clicking the node and selecting **Results** from the pop-up menu.

Select **View** from the main menu in the Results window to view the following:

- **Properties**
 - **Settings** — displays a window with a read-only table of the HP Regression node properties configuration when the node was last run.
 - **Run Status** — displays the status of the HP Regression node run. Information about the run start time, run duration, and completion status are displayed in this window.
 - **Variables** — displays a table of the variables in the training data set.
 - **Train Code** — displays the code that SAS Enterprise Miner used to train the node.
 - **Notes** — displays notes of interest, such as data or configuration information.
- **SAS Results**
 - **Log** — the SAS log of the HP Regression run.
 - **Output** — the SAS output of the HP Regression run.
 - **Flow Code** — the SAS code used to produce the output that the Regression node passes on to the next node in the process flow diagram.

- **Scoring**
 - **SAS Code** — the SAS score code that was created by the node. The SAS score code can be used outside of the SAS Enterprise Miner environment in custom user applications.
 - **PMML Code** — the HP Regression node does not generate PMML code.
- **Assessment** — The selections listed here are available when HP Regression node is run solo mode or on the grid. Additional assessment plots are available when HP Regression runs on the grid. See [HPDM Assessment on page 34](#) for more information about these assessment plots.
 - **Fit Statistics** — a table of the fit statistics from the model.
 - **Classification Chart** — The Classification chart displays a stacked bar chart of the classification results for a categorical target variable. The horizontal axis displays the target levels that observations actually belong to. The color of the stacked bars identifies the target levels that observations are classified into. The height of the stacked bars represents the percentage of total observations. This selection is available only for logistic regressions.
 - **Decision Chart** — displays a bar chart of the proportion of correctly classified observations and the proportion of misclassified observations in the training and validation data sets.
 - **Score Rankings Overlay** — In a score rankings chart, several statistics for each decile (group) of observations are plotted on the vertical axis. For a binary target, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order. Then the sorted observations are grouped into deciles, and observations in a decile are used to calculate the statistics that are plotted in deciles charts. The Score Rankings Overlay plot displays both train and validation statistics on the same axis.

By default, the horizontal axis of a score rankings chart displays the deciles (groups) of the observations.

The vertical axis displays the following values:

- Cumulative Lift
- Lift
- Gain
- % Response
- Cumulative % Response
- % Captured Response
- Mean for Predicted — for interval targets
- Maximum for Predicted — for interval targets
- Minimum for Predicted — for interval targets
- **Score Rankings Overlay** — displays various assessment statistics plots for each model. The drop-down menu in the upper left corner of the window enables you to select different plots. The available assessment statistics plots are as follows:
 - Cumulative Lift
 - Lift
 - Gain

- Percent Response
- Cumulative Percent Response
- Percent Captured Response
- Cumulative Percent Captured Response
- **Score Distribution** — The Score Distribution chart plots the proportions of events (by default), nonevents, and other values on the vertical axis. The values on the horizontal axis represent the model score of a bin. The model score depends on the prediction of the target and the number of buckets used. For categorical targets, observations are grouped into bins, based on the posterior probabilities of the event level and the number of buckets. The Score Distribution chart of a useful model shows a higher percentage of events for higher model score and a higher percentage of nonevents for lower model scores. For interval targets, observations are grouped into bins, based on the actual predicted values of the target. The default chart choice is Percentage of Events. Multiple chart choices are available for the Score Distribution Chart. This selection is available only for logistic regressions. The chart choices are as follows:
 - Percentage of Events — for categorical targets.
 - Number of Events — for categorical targets.
 - Cumulative Percentage of Events — for categorical targets.
 - Expected Profit — for categorical targets.
 - Report Variables — for categorical targets.
 - Mean for Predicted — for interval targets.
 - Max. for Predicted — for interval targets.
 - Min. for Predicted. — for interval targets.
- **HPDM Assessment** — Additional assessment plots are available when HP Regression runs on the grid. See [HPDM Assessment on page 34](#) for more information about these assessment plots.
- **Residual Statistics** — displays a box plot for the residual variable VALUE measurements when the target is interval. This selection is available only for linear regressions.
- **Model** — graphs and tables with information about the variables in the model. The available graphs and tables are as follows:
 - **Model/Performance Information** — displays modeling and performance information for the HP Regression node.
 - **Parameter Estimation** — displays bar charts for the coefficients in the final model. The bars are color coded to indicate the algebraic signs of the coefficients. The chart choices are as follows:
 - T-values
 - Estimates
 - Standard Errors
 - P-Values
 - **Selection Details** — available when model selection is used.
 - **Odds Ratio Plot** — displays the odds ratio, e^{β} , for changing either one unit of an interval input variable or between the specified level and the reference level of

a categorical input variable. Formally, let y be the binary outcome variable where 0 indicates failure and 1 indicates success, and let p be the probability that y is a success. Let x_1, x_2, \dots, x_k be a set of predictor variables. The logistic regression of y on x_1, x_2, \dots, x_k (without the interaction term) estimates the parameter values for $\beta_0, \beta_1, \dots, \beta_k$. This estimate is made via the maximum likelihood method, given as $\text{logit}(p) = \log(p / (1 - p)) = \beta_0 + \beta_1 * x_1 + \dots + \beta_k * x_k$. This plot is displayed e^{β} for each variable.

- **Table** — displays a table that contains the underlying data that is used to produce a chart. The **Table** menu item is dimmed and unavailable unless a results chart is open and selected.
- **Plot** — use the Graph Wizard to modify an existing Results plot or create a Results plot of your own. The **Plot** menu item is dimmed and unavailable unless a Results chart or table is open and selected.

HPDM Assessment

Overview

The High-Performance Data Mining (HPDM) assessment plots display a range of rank order statistics for model assessment of the HPDM models. The HPDM assessment statistics are different from the regular SAS Enterprise Miner assessment statistics when you run HPDM models using data sets in a grid mode. The HPDM assessment statistics are computed on the whole data set, and the regular SAS Enterprise Miner assessment statistics are computed on a sample of the data set. When running HPDM models in solo mode, only the regular SAS Enterprise Miner assessment plots are displayed in the result window. More details about HPDM assessment can be found in “The %EM_new_assess Macro” in *SAS Enterprise Miner High-Performance Data Mining Procedures and Macro Reference for SAS 9.3* and the %HPDM_node_assess macro documentation.

HPDM Assessment Plots for Binary Target Variables

In the HPDM assessment for binary target variables, the data is binned by descending value of the nonmissing estimated probability of the event level for the binary target variable. For each bin (group) of observations, several statistics are computed and plotted against the depth of bins in the HPDM assessment plots. By default the horizontal axis of an HPDM assessment plot displays the depth (binning) of the observations, and the vertical axis of an HPDM assessment plot displays one or several statistics. The available plots are as follows:

- Lift and Cumulative Lift
- Event and Non-Event Rate
- Classification Rates — CR
- Separation Curve — KS
- Cumulative Captured Events
- Receiver-Operator Characteristic

HPDM Assessment Plots for Interval Target Variables

In the HPDM assessment for interval target variables, the data is binned by a descending order of the nonmissing predicted values of the interval target variable. For each bin (group) of observations, several statistics are computed and plotted against the depth of bins in the HPDM assessment plots. By default, the horizontal axis of an HPDM

assessment plot displays the depth (binning) of the observations. The vertical axis of an HPDM assessment plot displays the actual target means, the predicted target means, and the residual means in each bin.

HPDM Bin Statistics Table

This table displays the summary and fit statistics for the binning of the target variable.

HP Regression Node Example

This example uses the sample SAS data set called `Sampsio.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 `Sampsio.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 `Bad` to **Target** and set the level of the variable `Bad` to **Binary**. 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 HP Regression node onto your diagram workspace. Connect them as shown in the diagram below.



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

Notice the following information:

- **Model/Performance Information** — The Model/Performance Information window contains information about the input data set, the regression model, and the high-performance computing environment. In this example, a logistic regression and a Logit link function were used to model the target variable `BAD`. The **Procedure Task Timing** group provides the timing information for various computational stages of the HP Regression node.
- **Odds Ratio Plot** — In the Logit model, the logarithm of the odds of the outcome is modeled as a linear combination of the input variables. Displays the odds ratio, e^{β} , for changing either one unit of an interval input variable or between the specified level and the reference level of a categorical input variable.
- **Parameter Estimation** — The Parameter Estimation plot displays the parameter estimates and associated p-value, t-value, and standard error for each input variable that is in the model. The drop-down menu in the upper left corner of the window enables you to select different parameters. Select **Estimates** from this menu. In this plot, we can see that `Delinq` and `Derog` have strong positive coefficients. Also, the

class level Sales for the Job variable has a strong positive coefficient. All other levels of Job have negative coefficients.

- Classification Chart — The Classification Chart displays the percentage of correct and incorrect classifications for the different levels of the class target variable. This example displays a large percentage of incorrect classification for target level 1 compared to target level 0.

Chapter 5

The HP Transform Node

HP Transform Node	37
Overview of the HP Transform Node	37
HP Transform Node Properties	37
HP Transform Node Results	40

HP Transform Node

Overview of the HP Transform Node




The HP Transform node enables you to make transformations to your interval input variables. Interval input transformations are important for improving the fit of your model. Transformation of variables can be used to stabilize variances, remove nonlinearity, and correct non-normality in variables.


HP Transform Node Properties

HP Transform Node General Properties


The following general properties are associated with the HP Transform node:

- **Node ID** — The Node ID property displays the ID that Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first Transform Variable node added to a diagram will have a Node ID of HPTrans. The second Transform Variable node added to the diagram will have a Node ID of HPTrans2.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Transform window. The Imported Data — HP Transform window contains a list of the ports that provide data sources to the Transform node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and click:


- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Transform window. The Exported Data — HP Transform window contains a list of the output data ports that the Transform node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.


If data exists for an imported data source, you can select the row in the imported data table and click:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.

HP Transform Node Train Properties

The following train properties are associated with the HP Transform node:

- **Variables** — Use the Variables property to specify the properties of each variable that you want to use in the data source. Select the  button to open a variables table. You set the transformation method in the Variables table.
- **Interval Inputs** — Use the Interval Inputs property to specify the default transformation method that you want to apply to interval input variables. Each variable in the Variables table that uses the Default method will use the method that you specify in this property. The available methods are
 - **Log** — transformed using the logarithm of the variable.
 - **Log 10** — transformed using the base-10 logarithm of the variable.
 - **Square Root** — transformed using the square root of the variable.
 - **Inverse** — transformed using the inverse of the variable.
 - **Square** — transformed using the square of the variable.
 - **Exponential** — transformed using the exponential of the variable.
 - **Centering** — centers variable values by subtracting the mean from each variable.
 - **Standardize** — standardizes the variable by subtracting the mean and dividing by the standard deviation.
 - **Range** — transformed using a scaled value of a variable equal to $(x - \min) / (\max - \min)$, where x is current variable value, \min is the minimum value for that variable, and \max is the maximum value for that variable.

- **Bucket** — buckets are created by dividing the data into evenly spaced intervals based on the difference between the maximum and minimum values.
- **None** — (default setting) No transformation is performed.
- **Interval Targets** — Use the Interval Targets property to specify the default transformation method that you want to use for interval target variables. Each interval target variable in the Variables table that uses the Default method will use the method that you specify in this property. The available methods are
 - **Log** — transformed using the logarithm of the variable.
 - **Log 10** — transformed using the base-10 logarithm of the variable.
 - **Square Root** — transformed using the square root of the variable.
 - **Inverse** — transformed using the inverse of the variable.
 - **Square** — transformed using the square of the variable.
 - **Exponential** — transformed using the exponential of the variable.
 - **Centering** — centers variable values by subtracting the mean from each variable.
 - **Standardize** — standardizes the variable by subtracting the mean and dividing by the standard deviation.
 - **Range** — transformed using a scaled value of a variable equal to $(x - \text{min}) / (\text{max} - \text{min})$, where x is current variable value, min is the minimum value for that variable, and max is the maximum value for that variable.
 - **Bucket** — buckets are created by dividing the data into evenly spaced intervals based on the difference between the maximum and minimum values.
 - **None** — (default setting) No transformation is performed.
- **SAS Code** — Select the  button to open the SAS Code window. You enter SAS code statements to create your own custom variable transformation in the SAS Code window. If you want to use code from a SAS catalog or external file, use the SAS Code window to submit a filename statement and a %include statement.

HP Transform Node Train Properties: Binning

- **Number of Bins** — specifies the number of bins to use when performing bucket or quantile transformations. When **Variables** is selected, the **Number of Bins** property specified in the Variables Editor is used.
- **Missing Values** — Use the Missing property to specify how to handle missing values when you use an optimal binning transformation. Select from any of the available missing value policies.
 - **Separate** — assigns missing values to its own separate branch.
 - **Ignore** — assigns all missing values to a missing value.
 - **First** — assigns the observations that contain missing values to the first bin.

HP Transform Node Score Properties

The following score properties are associated with the HP Transform node:

- **Hide** — Use the Hide property to specify how to handle the original variables after a transformation is performed. Setting the Hide property to **No** if you want to keep the original variables in the exported metadata from your transformed data set. The default setting for the Hide property is **Yes**. When this property is set to **Yes**, the

original variables are removed only from the exported metadata, and not removed from the exported data sets and data views.

- **Reject** — Use the Reject property to specify whether the model role of the original variables should be changed to Rejected or not. The default value for this property is **Yes**. To change the Reject value from **Yes** to **No**, you must set the Hide property value to **No**.

HP Transform Node Report Properties

The following report property is associated with the HP Transform node:

- **Summary Statistics** — Use the Summary Variables property to specify which variables have summary statistics computed. The default setting of **Yes** generates summary statistics on the transformed and new variables in the data set. The **No** setting does not generate any summary statistics.

HP Transform Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.
- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.
- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.
- **User-Added Node** — specifies if the node was created by a user as a SAS Enterprise Miner extension node.

HP Transform Node Results

After the HP Transform node successfully runs, you can open the Results — Transform window by right-clicking the node in the Diagram Workspace and selecting Results from the pop-up menu. The Results — Transform window contains an imputation summary table and a window displaying the node's output.

Select **View** from the main menu of the Transform Results window to view the following information:

- **Properties**
 - **Settings** — displays a window with a read-only table of the HP Transform node properties configuration when the node was last run. Use the Show Advanced Properties check box at the bottom of the window to see all of the available properties.
 - **Run Status** — indicates the status of the HP Transform node run. The Run Start Time, Run Duration, and information about whether the run completed successfully are displayed in this window.
 - **Variables** — a table of variables property of the node.
 - **Train Code** — the code that Enterprise Miner used to train the node.

- **Notes** — allows users to read or create notes of interest.
- **SAS Results**
 - **Log** — the SAS log of the HP Transform node run.
 - **Output** — the SAS output of the HP Transform node run. The SAS output displays how many and what type of variables are in the training data set. If you run in a grid environment, the output displays limited information about the connection to the grid.
 - **Flow Code** — the SAS code used to produce the output that the HP Transform node passes on to the next node in the process flow diagram.
- **Scoring**
 - **SAS Code** — the SAS score code that was created by the node. The SAS score code can be used outside of the Enterprise Miner environment in custom user applications.
 - **PMML Code** — the HP Transform node does not generate PMML code.
- **Summary Statistics**
 - **Statistics Table** — the Statistics Table provides a summary of the generated statistics for each variable. The information available in this table is
 - **Missing** — the number of missing observations
 - **Non Missing** — the number of nonmissing observations
 - **Minimum** — the minimum value of the input variable
 - **Mean** — the mean value of the input variable
 - **Maximum** — the maximum value of the input variable
 - **Standard Deviation** — the standard deviation of the input variable
 - **Skewness** — the measure of skewness of the input variable
 - **Kurtosis** — the kurtosis, also called steepness, of the input variable
 - **Table** — open the data table that corresponds to the graph that you have in focus.
 - **Plot** — opens the Select a Chart plotting wizard so that you can plot the data in the table that you have in focus.

Chapter 6

The HP Variable Selection Node

HP Variable Selection Node	43
Overview of the HP Variable Selection Node	43
Input Data Requirements for the HP Variable Selection Node	44
HP Variable Selection Node Properties	44
HP Variable Selection Node Results	47
HP Variable Selection Node Example	49

HP Variable Selection Node

Overview of the HP Variable Selection Node



Many data mining databases have hundreds of potential model inputs (independent or explanatory variables) that can be used to predict the target (dependent or response variable). The HP Variable Selection node reduces the number of inputs by identifying the input variables that are not related to the target variable, and rejecting them. Although rejected variables are passed to subsequent nodes in the process flow, these variables are not used as model inputs by a successor modeling nodes.

The HP Variable Selection node quickly identifies input variables that are useful for predicting the target variables. The use status **Input** is assigned to these variables. You can override the automatic selection process by assigning the status **Input** to a rejected variable or the status **Rejected** to an input variable. The information rich inputs are then evaluated in more detail by one of the modeling nodes.

The HP Variable Selection node provides both unsupervised and supervised variable selection. The input interval variables that have more missing data than desired or input class variables that have more levels than desired are excluded in the pre-selection stage. The unsupervised model performs variable selection by identifying a set of variables that jointly explain the maximal amount of data variance. Supervised variable selection includes LASSO, LARS, and stepwise regression analysis. The node can be run prior to any other analysis and the results passed to any Enterprise Miner node or any procedure in the SAS System.


Input Data Requirements for the HP Variable Selection Node

One or more input variables are required for the HP Variable Selection node. The data set can contain at most one target variable, and if a target variable is missing then only unsupervised selection is available. The HP Variable Selection node does not support multiple target variables.


HP Variable Selection Node Properties

HP Variable Selection Node General Properties


The following general properties are associated with the HP Variable Selection Node:

- **Node ID** — The Node ID property displays the ID that Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first HP Variable Selection Node that is added to a diagram will have a Node ID of HPVS. The second HP Variable Selection Node added to a diagram will have a Node ID of HPVS2, and so on.
- **Imported Data** — The Imported Data property provides access to the Imported Data — HP Variable Selection window. The Imported Data — HP Variable Selection window contains a list of the ports that provide data sources to the HP Variable Selection Node. Select the  button to the right of the Imported Data property to open a table of the imported data.

If data exists for an imported data source, you can select the row in the imported data table and click:


- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Exported Data** — The Exported Data property provides access to the Exported Data — HP Variable Selection window. The Exported Data — HP Variable Selection window contains a list of the output data ports that the HP Variable Selection Node creates data for when it runs. Select the  button to the right of the Exported Data property to open a table that lists the exported data sets.

If data exists for an imported data source, you can select the row in the imported data table and click:

- **Browse** to open a window where you can browse the data set.
- **Explore** to open the Explore window, where you can sample and plot the data.
- **Properties** to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.
- **Notes** — Select the  button to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.

HP Variable Selection Node Train Properties

The following train properties are associated with the HP Variable Selection node:

- **Variables** — specifies the properties of each variable in the data source that you want to use. Select the  button to the right of the Variables property to open a variables table. You can set the variable status to either **Use** or **Don't Use** in the table, and you can set a variable's report status to **Yes** or **No**.
- **Pre-screening** — set the Pre-screening property to **Yes** if you want to exclude variables that meet either of the following criteria:
 - The number of measurement levels is larger than the value specified in the **Maximum Level** property.
 - The percentage of missing data is larger than the value specified in the **Maximum Missing Percent** property.
- **Maximum Level** — Class variables with more measurement levels than specified here will have their use status set to **Rejected**. This property is available only when the **Pre-screening** property is set to **Yes**.
- **Maximum Missing Percent** — Variables with a greater percentage of missing data than the value specified here will have their use status set to **Rejected**. Valid values for this property are real numbers between 0 and 100. This property is available only when the **Pre-screening** property is set to **Yes**.
- **Target Model** — specifies the variable selection method.
 - **Unsupervised Selection** — Unsupervised selection identifies a set of variables that jointly explains the maximal amount of data variance. This method does not require a target variable. The HPREDUCE procedure is used in unsupervised selection.
 - **Supervised Selection** — Supervised selection requires exactly one target variable, and includes LASSO, LARS, and stepwise regression analysis. The HPREG procedure is used for binary or interval target variables and the HPLOGISTIC procedure is used for nominal or ordinal target variables.
 - **Sequential Selection** — Sequential selection runs unsupervised selection and supervised selection, sequentially.

HP Variable Selection Node Train Properties: Unsupervised Selection

- **Maximum Steps** — specifies the maximum number of steps to take for variable selection.
- **Maximum Effects** — specifies the maximum number of effects to select.
- **Correlation Statistics** — specifies the statistics that determine variable selection.
 - **Covariance** — selects variables based on the covariance matrix.
 - **Correlation** — selects variables based on the correlation matrix.
 - **Sum of Squares and Crossproducts** — selects variables based on the sum of squares and cross-product matrices.
- **Cumulative Cutoff** — specifies the fraction of the total variance to be explained by the selected variables.
- **Increment** — specifies the minimal increment of explained variance allowed after the cumulative cutoff value is reached.

HP Variable Selection Node Train Properties: Supervised Selection

- **Intercept** — Set the Intercept property to **Yes** if you want to include the intercept. The intercept is required for nominal and ordinal target variables. The default setting is **No**.
- **Selection Method** — specifies the method that is used to select effects
 - **Fast Selection** — For binary or interval target variables, fast selection starts with no effects in the model and adds effects until the entry significance level is met. For nominal or ordinal target variables, fast selection starts with all effects in the model and deletes effects until the exit significance level is met.
 - **LAR** — The least angle regression (LAR) method starts with no effects in the model and adds effects. The estimates at any step are reduced when compared to the corresponding least squares estimates. The **LAR** option is available only for binary and interval target variables.
 - **LASSO** — The LASSO method adds and deletes variables based on a version of ordinary least squares where the sum of the absolute regression coefficients is constrained. The **LASSO** option is available only for binary and interval target variables.
- **Choose Criteria** — During the selection process, variables are chosen based on the selection criterion specified here.
 - **SBC** — chooses the model that has the smallest Schwarz Bayesian Criterion value.
 - **AIC** — chooses the model that has the smallest Akaike Information Criterion value.
 - **AICC** — chooses the model that has the smallest Corrected Akaike Information Criterion value.
- **Stop Criteria** — specifies the criterion that is used to stop the selection process.
 - **Max Steps** — the maximum number of steps to take for variable selection
 - **SBC** — Schwarz Bayesian Criterion value
 - **AIC** — Akaike Information Criterion value
 - **AICC** — Corrected Akaike Information Criterion value
 - **Significance Level** — Significance Level
- **Maximum Steps** — specifies the maximum number of selection steps that are performed.
- **Collinearity Diagnostics** — Set the Collinearity Diagnostics property to **Yes** if you want the model to produce variance inflation factors with the parameter estimates. The Collinearity Diagnostics property is available only for binary and interval target variables.

HP Variable Selection Node Status Properties

The following status properties are associated with this node:

- **Create Time** — displays the time that the node was created.
- **Run ID** — displays the identifier of the node run. A new identifier is created every time the node runs.
- **Last Error** — displays the error message from the last run.
- **Last Status** — displays the last reported status of the node.

- **Last Run Time** — displays the time at which the node was last run.
- **Run Duration** — displays the length of time of the last node run.
- **Grid Host** — displays the grid server that was used during the node run.
- **User-Added Node** — specifies if the node was created by a user as a SAS Enterprise Miner extension node.

HP Variable Selection Node Results

HP Variable Selection Node Results Window

After a successful node run, you can open the Results window of the HP Variable Selection node by right-clicking the node and selecting **Results** from the pop-up menu.

Select **View** from the main menu in the Results window to view the following:

- **Properties**
 - **Settings** — displays a window with a read-only table of the HP Variable Selection node properties configuration when the node was last run.
 - **Run Status** — displays the status of the HP Variable Selection node run. Information about the run start time, run duration, and completion status are displayed in this window.
 - **Variables** — displays a table of the variables in the training data set.
 - **Train Code** — displays the code that Enterprise Miner used to train the node.
 - **Notes** — displays notes of interest, such as data or configuration information.
- **SAS Results**
 - **Log** — the SAS log of the HP Variable Selection run.
 - **Output** — the SAS output of the HP Variable Selection run.
 - **Flow Code** — the HP Variable Selection node does not generate flow code.
- **Scoring**
 - **SAS Code** — the HP Variable Selection node does not generate SAS code.
 - **PMML Code** — the HP Variable Selection node does not generate PMML code.
- **Model** — graphs and tables with information about the variables in the model. The available graphs and tables are
 - **Variable Explained by HPReduce** — A selection summary table displays what variable (or effects for class variables) is selected in each step. Also, the total variance that is explained by the variables selected is given. This table is available for both the unsupervised selection model and sequential selection model.
 - **Variable Selection by HPReduce** — This table provides the use status, role, measurement level, and reason for rejection or inclusion for the variables in the input data set. This table is available for both the unsupervised selection model and sequential selection model.
 - **Parameter Estimation by HPReg** — This table displays the parameters (or effects for class variables) in the selected model. Also, the estimates, degrees of freedom (DF), standard error, standardized estimates, t-value, and two-tailed significance probability ($\text{Pr} > |t|$) are provided. This table is available for both the

supervised selection model and sequential selection model, but only for binary or interval target variables.

- **Variable Selection by HPReg** — This table provides the use status, role, measurement level, and reason for rejection or inclusion for the variables in the input data set. This table is available for both the supervised selection model and sequential selection model, but only for binary or interval target variables.
- **Fit Statistics by HPReg** — A table of fit statistics for the selected model, such as root mean square error, R-square, Adjusted R-square, AIC, AICC, SBC, and ASE. This table is available for both the supervised selection model and sequential selection model, but only for binary or interval target variables.
- **ANOVA by HPReg** — This table displays an analysis of variance for the selected model. This table is available for both the supervised selection model and sequential selection model, but only for binary or interval target variables.
- **Parameter Estimation by HPLgistic** — This table displays the parameters (or effects for class variables) in the selected model. Also, the estimates, degrees of freedom (DF), standard error, standardized estimates, t-value, and two-tailed significance probability ($Pr > |t|$) are provided. This table is available for both the supervised selection model and sequential selection model, but only for nominal or ordinal target variables.
- **Variable Selection by HPLogistic** — This table provides the use status, role, measurement level, and reason for rejection or inclusion for the variables in the input data set. This table is available for both the supervised selection model and sequential selection model, but only for nominal or ordinal target variables.
- **Global Test by HPLogistic** — This table provides a statistical test for the hypothesis of whether the final model provides a better fit than a model without effects (an “intercept-only” model). This table is available for both the supervised selection model and sequential selection model, but only for nominal or ordinal target variables.
- **Fit Statistics by HPLogistic** — A table of fit statistics for the selection model, such as the log-likelihood, AIC, AICC, and BIC. This table is available for both the supervised selection model and sequential selection model, but only for nominal or ordinal target variables.
- **Plots**
 - **Parameter Estimates** — A bar chart that displays the absolute values of the parameter estimates. The color of the bar indicates the sign of the parameter estimate.
 - **Variance Explained** — A bar chart that displays the percentage of variance that is explained by the variables selected. This plot is available for both the unsupervised selection model and sequential selection model.
 - **Solution Path** — This plot displays the standardized coefficients for all of the effects selected at each step in the stepwise selection method. This plot is available for both the supervised selection model and sequential selection model, but only for binary or interval target variables.
 - **Iteration Plot** — This plot displays the change in selection criterion as effects enter the model. This plot is available for both the supervised selection model and sequential selection model.
- **Table** — displays a table that contains the underlying data that is used to produce a chart. The Table menu item is dimmed and unavailable unless a results chart is open and selected.

- **Plot** — use the Graph Wizard to modify an existing Results plot or create a Results plot of your own. The Plot menu item is dimmed and unavailable unless a Results chart or table is open and selected.

HP Variable Selection Node Example

This example uses the sample SAS data set called `Sampsio.Hmeq`. You must use the data set to create an 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.

- Choose SAS Table as your metadata source and click **Next**.
- Enter `Sampsio.Hmeq` in the Table field and click **Next**.
- Continue to the Metadata Advisor step and choose the **Basic Metadata Advisor**.
- In the Column Metadata window, set the role of the variable `Bad` to **Target** and set the level of the variable `Bad` to **Binary**. 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 HP Variable Selection node onto your diagram workspace. Connect them as shown in the diagram below.



Select the HP Variable Selection node and change the following properties:

- Set the value for **Correlation Statistics** to **Correlation**.
- Set the value for **Selection Method** to **LAR**.

Run the HP Variable Selection node by right-clicking on the HP Variable Selection node and selecting **Run**. In the Confirmation window, select **Yes**. After a successful run of the HP Variable Selection node, select **Results** in the Run Status window.

Notice the following results:

- **Variable Selection by HPReg** — In this example, the variables `Mortdue`, `Reason`, and `Value` are rejected. In this example, the `HPREDUCE` procedure and the `HPREG` procedure are responsible for identifying the role of each variable.
- **Variance Explained** — The variance explained plot displays the percentage of variance that is explained by the effects in the model at each step. Each bar indicates the base variance explained and the incremental variance explained. The base variance explained is the total variance explained by the effects in the model before a new effect enters the model. The incremental variance explained is the total variance explained by the new effect that entered the model in that iteration.
- **Parameter Estimates** — The parameter estimates bar chart is color-coded to indicate the sign of the parameter estimate. Notice that `Job_Sales` has a relatively strong, positive parameter estimate and `Job_Office` has a relatively strong, negative parameter estimate.

- Iteration Plot — The iteration plot shows how the criterion used to choose the selected model changes as the effects enter the model. Notice that the graph achieves a minimum value at the 12th step, which is where model selection terminates.
- Solution Path — The solution path plot enables you to assess the relative importance of the effects selected at any step in the selection process. Also, it provides information as to when effects entered the model. The blue vertical line appearing at step 12 indicates that the model was selected at that step because the optimal value of the Schwarz-Bayesian Coefficient was reached.