TIBCO®

The Power of Now®

TIBCO Spotfire Miner™ 8.2 Getting Started Guide

November 2010

TIBCO Software Inc.

IMPORTANT INFORMATION

Reference

The correct bibliographic reference for this document is as follows:

TIBCO Spotfire Miner™ 8.2 Getting Started Guide, TIBCO Software Inc.

Technical Support

For technical support, please visit http://spotfire.tibco.com/support and register for a support account.

CONTENTS

A QUICK TOUR

1

INTRODUCTION

TIBCO Spotfire Miner™ is a tool for enterprise-wide data mining that is designed to work seamlessly with the software you already use. You can import data from and export data to many sources, including spreadsheets such as Excel and Lotus, databases such as DB2, Oracle, and Sybase, and analytical software such as SAS and SPSS. After you have accessed your data, you can do any of the following:

- Explore your data via charts, tabular displays, and descriptive statistics.
- Use Spotfire Miner's tools for data cleaning and data manipulation to prepare your data for analytic modeling.
- Fit a variety of statistical models, including linear and logistic regression, and classification trees.
- Evaluate the effectiveness of your models with standard tools, such as lift charts.

This Quick Tour briefly introduces you to the notion of a Spotfire Miner *network*, and then explores a simple network to show you how you can use Spotfire Miner to solve a real-world data mining problem.

OVERVIEW OF THE SPOTFIRE MINER INTERFACE

The Spotfire Miner interface contains a palette of nodes to use in data mining, plus a canvas for designing visual *networks*. When you start Spotfire Miner by loading a new worksheet, the interface looks like Figure 1.1.

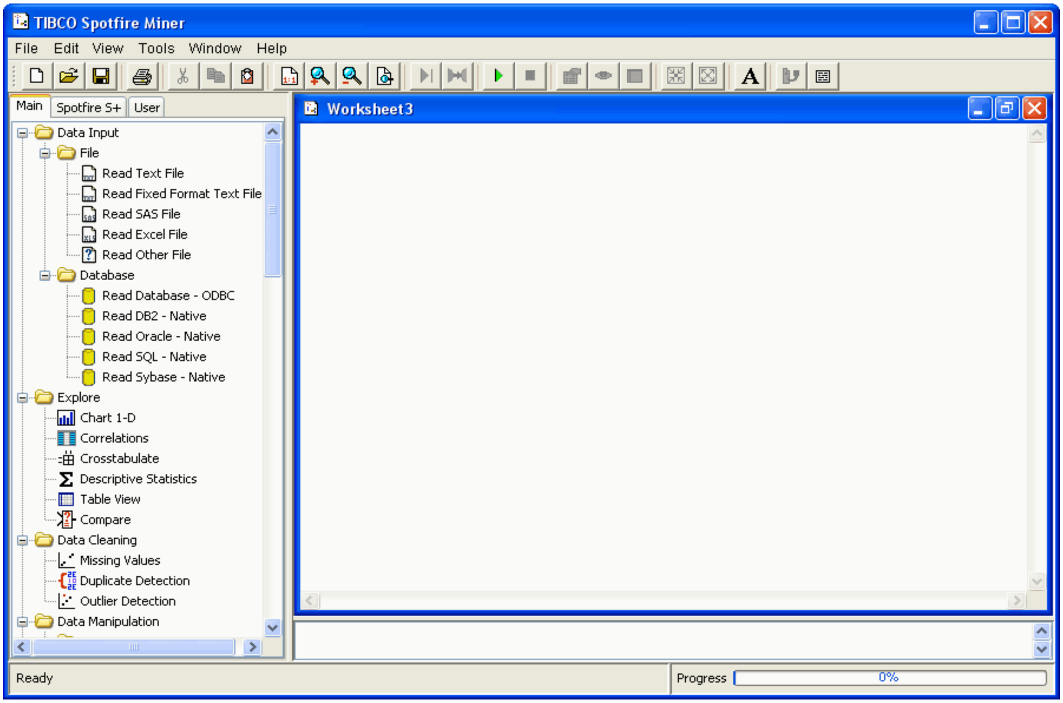


Figure 1.1: *The Spotfire Miner graphical user interface.*

Create a network by dragging and dropping *components* from the *explorer pane* on the left to a *worksheet* in the *desktop pane* on the right to create the *nodes* of the network. Then establish *links* between the nodes and set the *properties* of the nodes.

Below the desktop pane is a *message pane*, which displays messages on the status of the nodes as they are evaluated. Watch this pane for error and warning messages from Spotfire Miner.

When you run the network, Spotfire Miner evaluates the nodes by passing data through the Spotfire Miner *pipeline* architecture, where it processes the data node by node. Temporary files cache the results of each node in a binary format for quick processing. By default, data are passed through the pipeline 10,000 rows at a time, but you can adjust this number either globally or for individual nodes.

In the sections that follow, use a simple network to see both its essential features and how these features combine to solve a data mining problem in the pharmaceutical industry.

A DATA ANALYSIS PROBLEM

The data set used in this example is from the Duke University Cardiovascular Disease Databank and consists of 3504 patients and 6 variables. The patients were referred to Duke University Medical Center for chest pain. The goal of this exercise is simple:

Predict the probability a patient has *significant* coronary disease, defined as greater than or equal to a 75% diameter narrowing in at least one important coronary artery.

The six variables used in this dataset are as follows:

`sex`	0 = male, 1 = female
`age`	of the patient, in years
`cad.dur`	the duration of the coronary event, in months
`cholesterol`	the measurement of the patient's cholesterol level
`sigdz`	the presence (or absence) of *significant* coronary disease
`tvdlm`	the presence (or absence) of *severe* coronary disease. This is also called "three vessel" or "left main" disease.

This analysis uses significant coronary disease as a response variable (`sigdz`). To run this analysis, create a Spotfire Miner network to evaluate two resulting models and determine which is a better predictor of the probability of significant coronary disease.

To begin this example, launch Spotfire Miner. The Spotfire Miner splash screen appears, followed by the dialog shown in Figure 1.2.

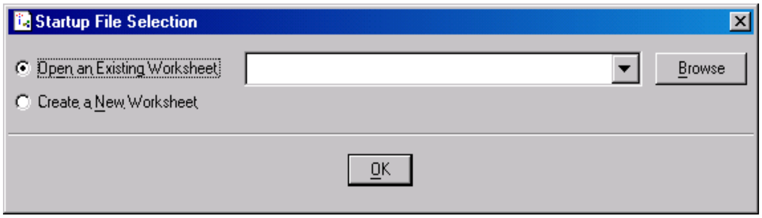


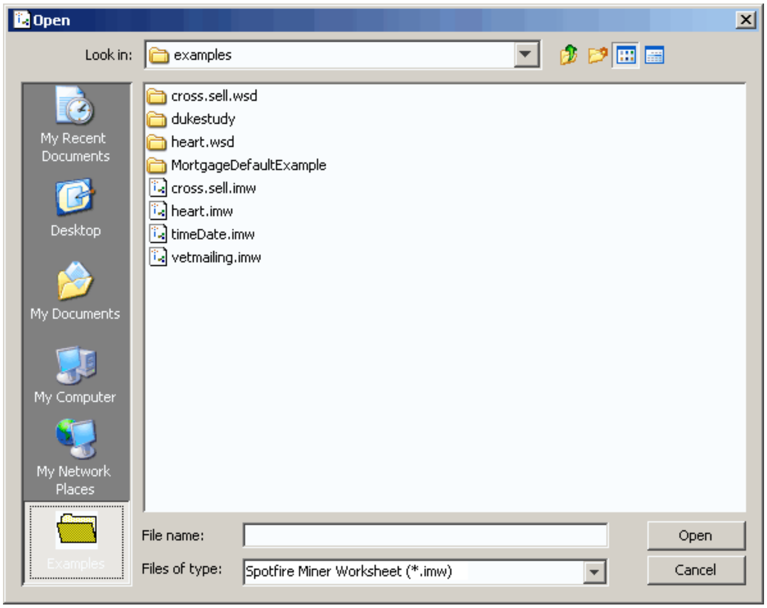
Figure 1.2: *The **Startup File Selection** dialog.*

To open the example worksheet for this problem:

1. Click **Browse** to display the **Open** dialog.

2. At the bottom of the **Open** file selection dialog, click the **Examples** folder icon. (Clicking this icon copies all files in the installation **examples** folder to the **examples**[1] directory and preserves the original worksheets and datasets in the installation examples directory if you need them.)

3. Note the **Open** dialog now displays the new **examples** folder, as shown in Figure 1.3. Double-click the **dukestudy** folder, select **dukecath.imw**, and click **Open**.



Look in: examples

cross.sell.wsd
dukestudy
heart.wsd
MortgageDefaultExample
cross.sell.imw
heart.imw
timeDate.imw
vetmailing.imw

File name:

Files of type: Spotfire Miner Worksheet (*.imw)

Figure 1.3: *Clicking the **Examples** folder icon (lower left) copies the files in the installation **examples** directory .*

[1] The location of the ***%Documents%*** folder depends on which version of Microsoft Windows® you are running. by default:

on Windows XP®: **C:\Documents and Settings\*username*\My Documents\Spotfire Miner\examples**

on Windows Vista®: **C:\users\*username*\Documents\Spotfire Miner\examples**

4. In the **Startup File Selection** dialog, click **OK** to open the worksheet containing the example network shown in Figure 1.4.

5. Notice that all the nodes of the network appear with red *status indicators*, which means the nodes are connected but the data is missing. After you import the data, the nodes that you can run change to yellow. After you complete the dialog for the node and run the network, all the status indicators change to green to show that the nodes havecompleted successfully.

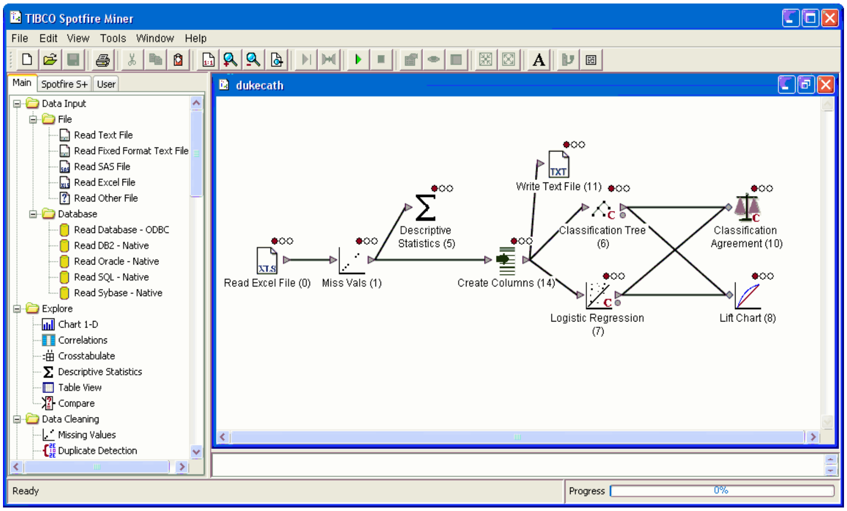


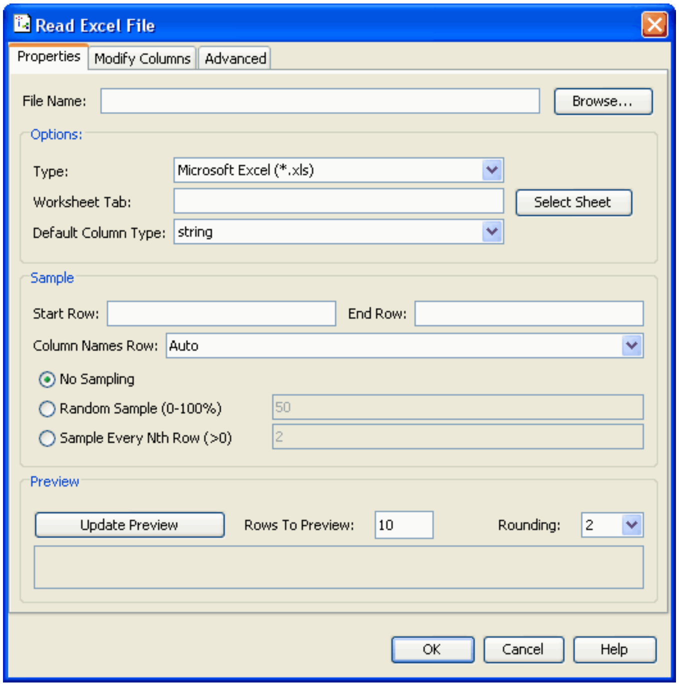
Figure 1.4: *Red status indicators mean the nodes are not ready to be run or are missing data.*

The network in Figure 1.4 shows some data mining steps. Next, examine each of the nodes in the network to determine what they do.

ACCESS DATA

The example network shown in Figure 1.4 begins with a **Read Excel File** node, one of many ways to enter data in the Spotfire Miner pipeline. You can use any of the **Data Input** components for this purpose, including **Read Text File**, **Read Fixed Format Text File**, **Read SAS File**, **Read Other File**, or one of the **Database** components.

1. Double-click the **Read Excel File** node to open its properties dialog. The dialog is shown in Figure 1.5.



File Name:

Type: Microsoft Excel (*.xls)

Worksheet Tab:

Default Column Type: string

Start Row:

End Row:

Column Names Row: Auto

● No Sampling

○ Random Sample (0-100%) 50

○ Sample Every Nth Row (>0) 2

Rows To Preview: 10

Rounding: 2

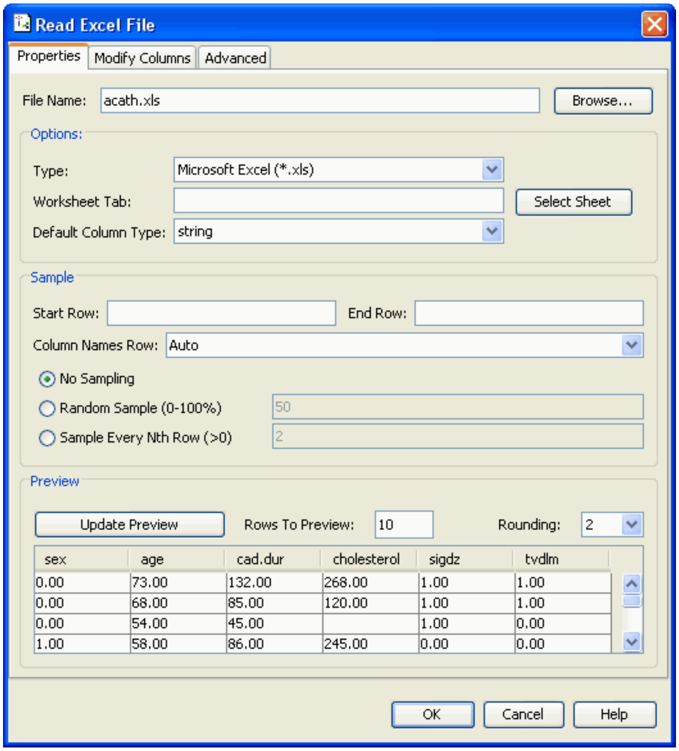
Figure 1.5: *The* ***Properties*** *page of the* ***Read Excel File*** *dialog.*

2. Click **Browse** to display the **Open** dialog.

3. Because you previously clicked the **Examples** folder icon (at the lower left of the dialog), the **Open** dialog should display the **examples/dukestudy** folder.

4. In the **dukestudy** folder, select the data file **acath.xls**, and click **Open**. (If your options are set to hide file extensions, the file name is displayed as **acath**.)

   In the **Preview** group, click **Update Preview** to display the first ten rows of the data (the default).



File Name: acath.xls

Type: Microsoft Excel (*.xls)

Worksheet Tab:

Default Column Type: string

Start Row:

End Row:

Column Names Row: Auto

● No Sampling

○ Random Sample (0-100%) 50

○ Sample Every Nth Row (>0) 2

Rows To Preview: 10

Rounding: 2

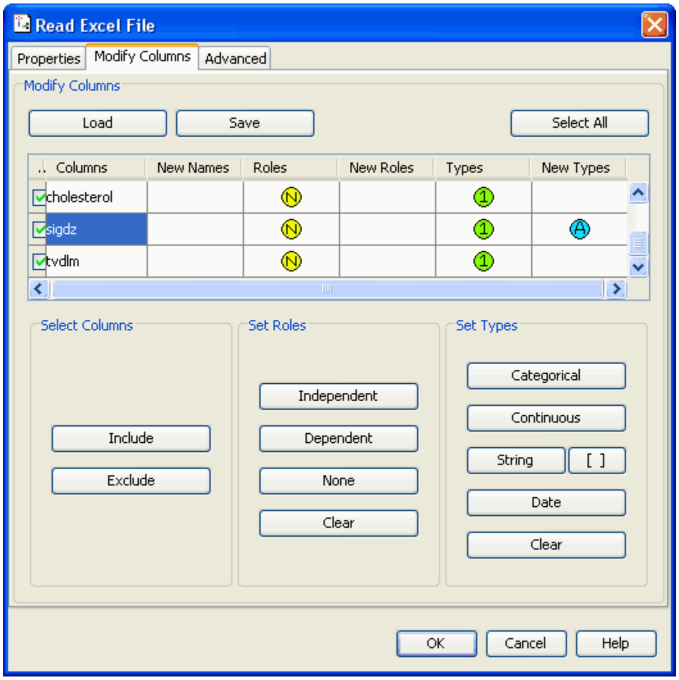
sex	age	cad.dur	cholesterol	sigdz	tvdlm
0.00	73.00	132.00	268.00	1.00	1.00
0.00	68.00	85.00	120.00	1.00	1.00
0.00	54.00	45.00		1.00	0.00
1.00	58.00	86.00	245.00	0.00	0.00

Figure 1.6: *The completed* ***Read Excel File*** *node.*

Because the goal of this example is to predict the probability of significant coronary disease (`sigdz`), you want to build a model that uses `sigdz` as a dependent variable, which requires that it be set as a categorical variable. However, it was imported as a continuous (numeric) variable. To change `sigdz` to a categorical variable:

5. Click the **Modify Columns** tab.
6. Scroll down until you find `sigdz` and select it.
7. In the **Set Types** group, click **Categorical**.

8. Click **OK** to close the dialog.



Columns	New Names	Roles	New Roles	Types	New Types
☑ cholesterol		N		1	
☑ sigdz		N		1	A
☑ tvdlm		N		1	

Figure 1.7: *Changing the* `sigdz` *variable type from continuous to categorical.*

9. On the Spotfire Miner toolbar, click the **Run to Here** ▶| to run the network so far.

The status indicator for the **Read Excel File** node now turns green, indicating that it completed reading in the data successfully.

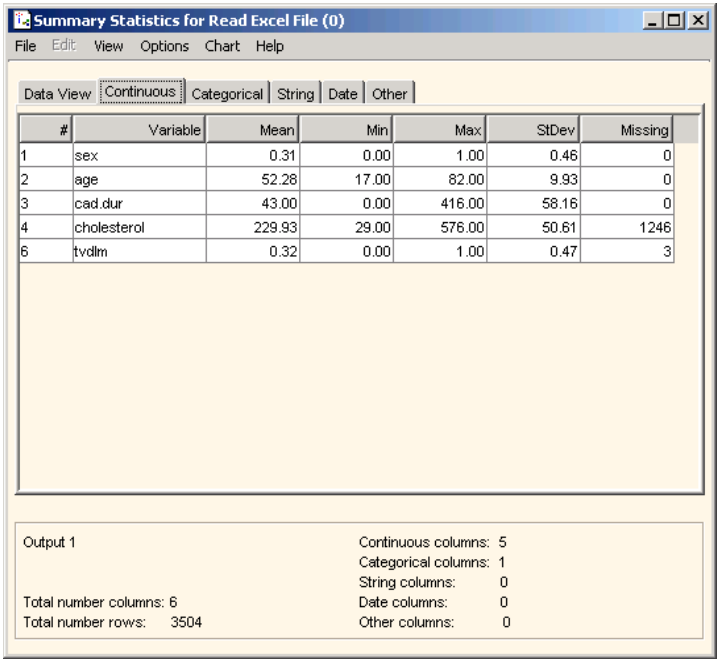
Note

To see the help file for the example dataset, from the Spotfire Miner main menu, click **Help** ▶ **Help Index**, and then select **acath.xls Data Set**.

EXPLORE DATA

Open the viewer for the **Read Excel File** node and examine the data you imported.

1. Click the **Read Excel File** node to select it, and then click the **Viewer** button on the Spotfire Miner toolbar.



#	Variable	Mean	Min	Max	StDev	Missing
1	sex	0.31	0.00	1.00	0.46	0
2	age	52.28	17.00	82.00	9.93	0
3	cad.dur	43.00	0.00	416.00	58.16	0
4	cholesterol	229.93	29.00	576.00	50.61	1246
6	tvdlm	0.32	0.00	1.00	0.47	3

Output 1

Total number columns: 6
Total number rows: 3504

Continuous columns: 5
Categorical columns: 1
String columns: 0
Date columns: 0
Other columns: 0

Figure 1.8: *The viewer for the* ***Read Excel File*** *node, displaying the* ***Continuous*** *tab.*

As shown in Figure 1.8, the viewer for the **Read Excel File** node is the generic *node viewer*, the common viewer for many of the nodes in Spotfire Miner, including all the input/output and data manipulation nodes. The node viewer consists of six tabbed pages:

- The first displaying the entire data set.
- The second through fifth displaying the four data types (*continuous*, *categorical*, *string*, and *date*).
- The sixth displaying any other data types.

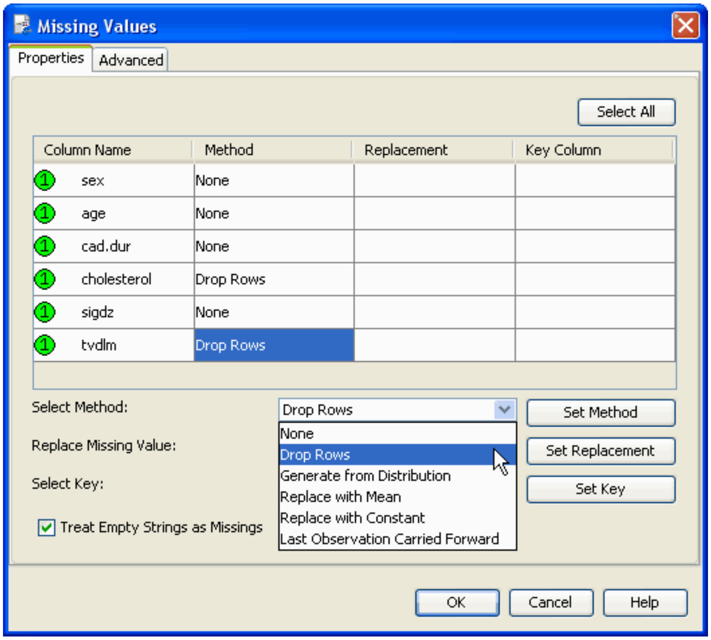
The bottom of each page of the node viewer displays summary data for the node's output: 5 continuous columns (or variables) and 3,504 observations. Figure 1.8 displays the five variables for the default *continuous* variables.

2. Click the **Continuous** tab to examine these variables. This chart shows an interesting characteristic about the data: The **Missing** column shows the `cholesterol` variable missing 1246 values and the `tvdlm` variable missing three (3) values.
3. Click the **Categorical** tab to see the sole categorical variable, `sigdz`. To see its levels, click anywhere in its row.
4. When you are finished examining the data, close the node viewer by clicking the button (☒) at the top right corner of the window.

Clean the Data

Next, use the **Missing Values** node to drop the missing rows, because they add nothing to the analysis.

1. Right-click the **Missing Values** node in the worksheet, and then select **Properties**.
2. Click **cholesterol**, and then CTRL+click **tvdlm** to select just the two columns. In **Select Method** box, select **Drop Rows**, and then click **Set Method**.



Column Name	Method	Replacement	Key Column
① sex	None		
① age	None		
① cad.dur	None		
① cholesterol	Drop Rows		
① sigdz	None		
① tvdlm	Drop Rows		

Select Method: Drop Rows

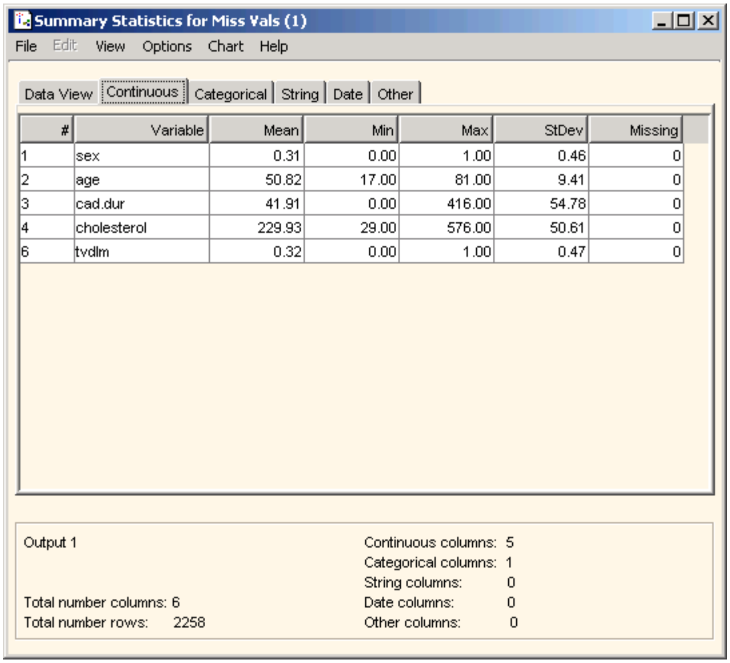
Replace Missing Value:

Select Key:

☑ Treat Empty Strings as Missings

Figure 1.9: *Selecting **Drop Rows** as the method for handling missing values.*

3. Click **OK**, and then click **Run to Here** (▶) to run the network so far.
4. Right-click the **Missing Values** node and select **Viewer**.
5. Click the **Continuous** tab and examine the summary data at the bottom of the dialog. As shown in Figure 1.10, the data set now contains only 2258 rows.



#	Variable	Mean	Min	Max	StDev	Missing
1	sex	0.31	0.00	1.00	0.46	0
2	age	50.82	17.00	81.00	9.41	0
3	cad.dur	41.91	0.00	416.00	54.78	0
4	cholesterol	229.93	29.00	576.00	50.61	0
6	tvdlm	0.32	0.00	1.00	0.47	0

Output 1

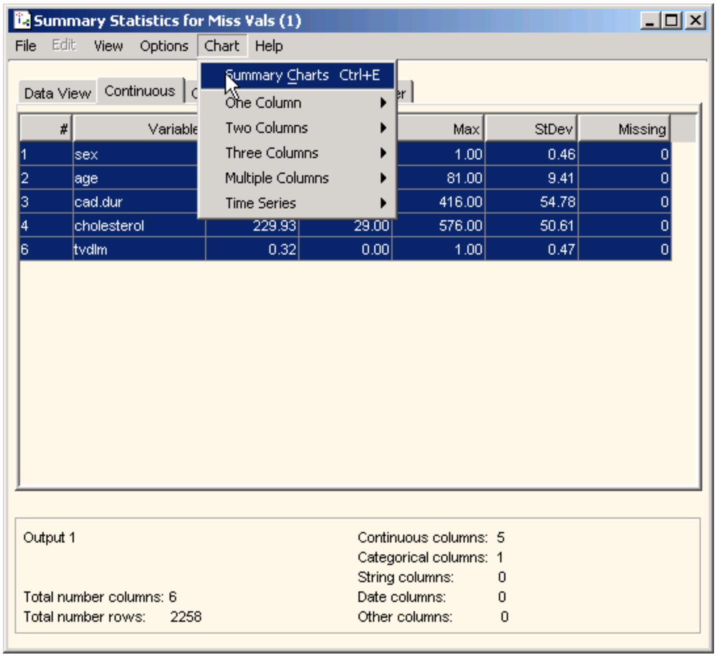
Total number columns: 6
Total number rows: 2258

Continuous columns: 5
Categorical columns: 1
String columns: 0
Date columns: 0
Other columns: 0

Figure 1.10: *Running the* ***Missing Values*** *node drops the rows with no data for the* `cholesterol` *or* `tvdlm` *(severe coronary disease) variables.*

To get a visual representation of the data, you can plot each of these continuous variables:

6. Select the first row in the grid view by clicking anywhere in the row.
7. SHIFT-click the last row in the grid view to select all the continuous variables in the data set.
8. From the menu at the top of the node viewer window, select **Chart ▶ Summary Charts**, as shown in Figure 1.11.

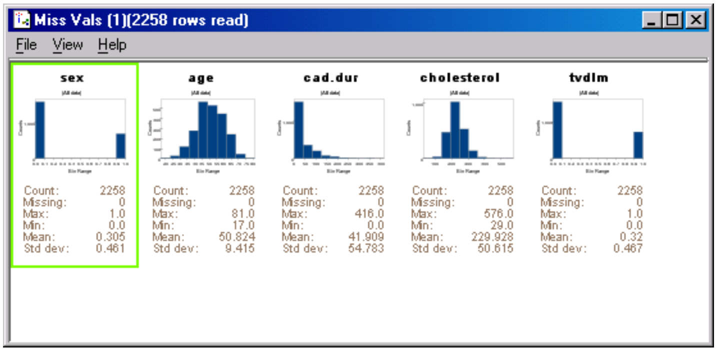


#	Variable			Max	StDev	Missing
1	sex			1.00	0.46	0
2	age			81.00	9.41	0
3	cad.dur			416.00	54.78	0
4	cholesterol	229.93	29.00	576.00	50.61	0
6	tvdlm	0.32	0.00	1.00	0.47	0

```
Output 1                                  Continuous columns:  5
                                          Categorical columns: 1
                                          String columns:      0
Total number columns: 6                   Date columns:        0
Total number rows:    2258                Other columns:       0
```

Figure 1.11: *Creating univariate charts of the data from within the node viewer.*

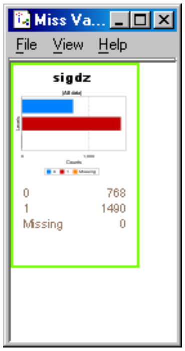
The chart viewer shown in Figure 1.12 opens, displaying a data summary and plot for each of the selected variables:



```
Count:      2258   Count:      2258   Count:      2258   Count:      2258   Count:      2258
Missing:       0   Missing:       0   Missing:       0   Missing:       0   Missing:       0
Max:         1.0   Max:        81.0   Max:       416.0   Max:       576.0   Max:         1.0
Min:         0.0   Min:        17.0   Min:         0.0   Min:        29.0   Min:         0.0
Mean:      0.305   Mean:     50.824   Mean:     41.909   Mean:    229.928   Mean:       0.32
Std dev:   0.461   Std dev:   9.415   Std dev:  54.783   Std dev:  50.615   Std dev:   0.467
```

Figure 1.12: *A data summary and plot for each continuous variable in the dataset. For continuous data, histograms are displayed.*

9. Return to the viewer window, click the **Categorical** tab, and then repeat for the `sigdz` variable in the data:

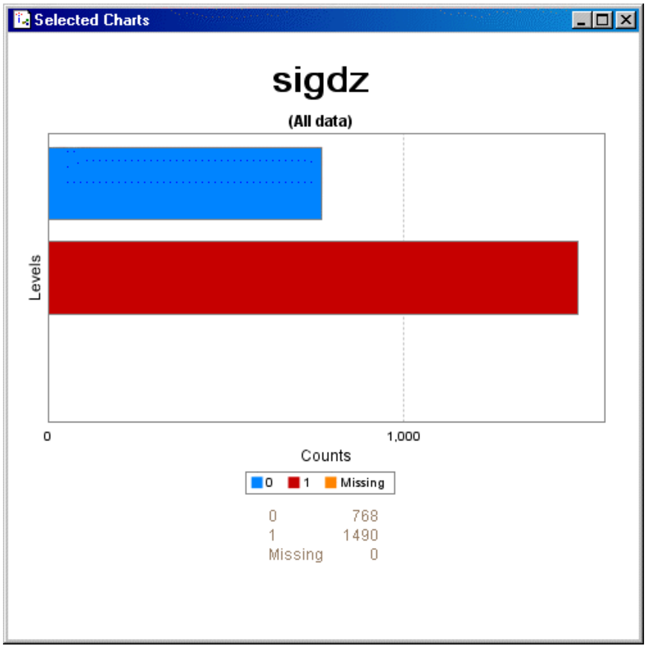


```
0            768
1           1490
Missing        0
```

Figure 1.13: *A data summary and plot for each categorical variable is also displayed. For categorical data, Spotfire Miner displays a bar chart.*

To enlarge the categorical plot, double-click the `sigdz` plot. As Figure 1.14 shows, a **Selected Charts** window opens, displaying the data summary and a bar chart for the `sigdz` variable.

A closer look at this categorical variable shows a large number of patients who have significant coronary arterial disease (by a factor of roughly 2:1), as revealed by the counts of the levels under the chart. The next section returns to this observation.



```
0          768
1         1490
Missing      0
```

Figure 1.14: *An enlarged view of the* `sigdz` *plot, showing a 2:1 ratio of those with significant coronary disease and those without.*

10. Close the **Selected Charts** window.
11. When you are finished viewing the data, close the node viewer and both chart viewers.

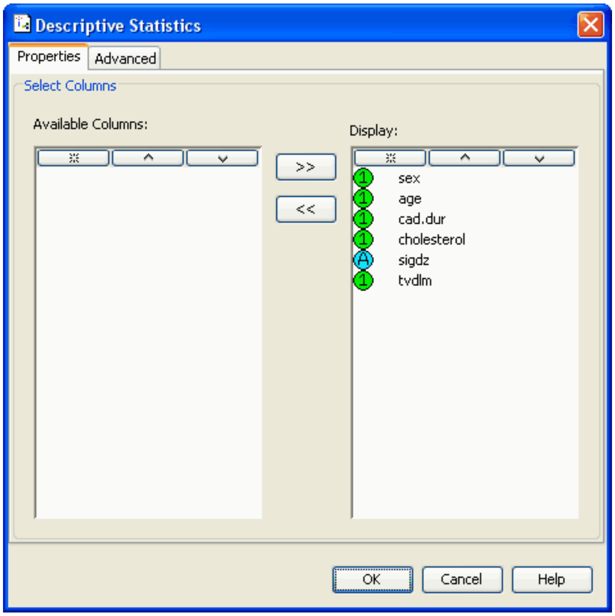
Further Data Exploration

You can get a better understanding of the data by examining the summary statistics of the data, now that you have modified columns and removed missing values. By running the **Descriptive Statistics** node, you can get the mean, standard deviation, and the extreme values of the data.

Set the properties for the **Descriptive Statistics** node as follows:

1. Right-click the **Descriptive Statistics** node and select **Properties**.
2. Select all the variables in the **Available Columns** list, and then click the right double-arrow button >> to move the variables to the **Display** list box, as shown in Figure 1.15.

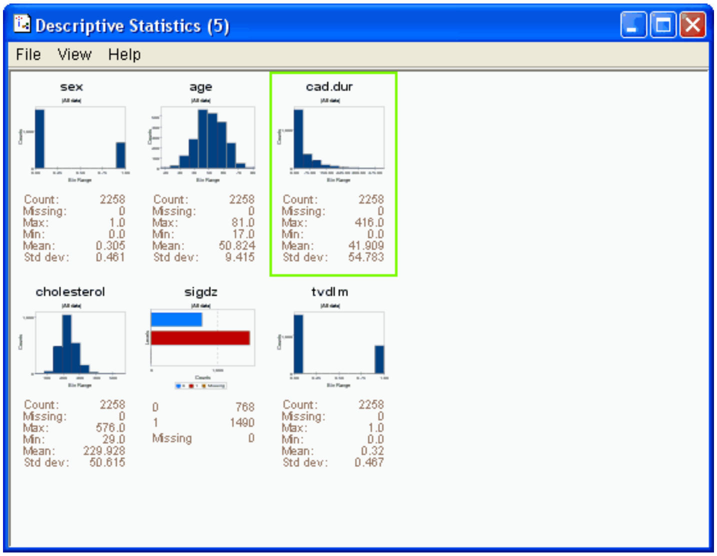
Click **OK**, and then click **Run to Here** on the toolbar.



Display:
sex
age
cad.dur
cholesterol
sigdz
tvdlm

Figure 1.15: *Use the **Descriptive Statistics** node to select variables for which you want to calculate statistics, such as the mean, the standard deviation, and the extreme values.*

3. Right-click the **Descriptive Statistics** node and select **Viewer**. The statistics for the variables are shown in Figure 1.16. Notice that the histogram for `cad.dur` shows that the levels are highly skewed.



```
sex                     age                     cad.dur
Count:       2258       Count:       2258       Count:       2258
Missing:        0       Missing:        0       Missing:        0
Max:          1.0       Max:         81.0       Max:        416.0
Min:          0.0       Min:         17.0       Min:          0.0
Mean:       0.305       Mean:      50.824       Mean:      41.909
Std dev:    0.461       Std dev:    9.415       Std dev:   54.783

cholesterol             sigdz                   tvdlm
Count:       2258       0             768       Count:       2258
Missing:        0       1            1490       Missing:        0
Max:        576.0       Missing         0       Max:          1.0
Min:         29.0                               Min:          0.0
Mean:     229.928                               Mean:        0.32
Std dev:   50.615                               Std dev:    0.467
```

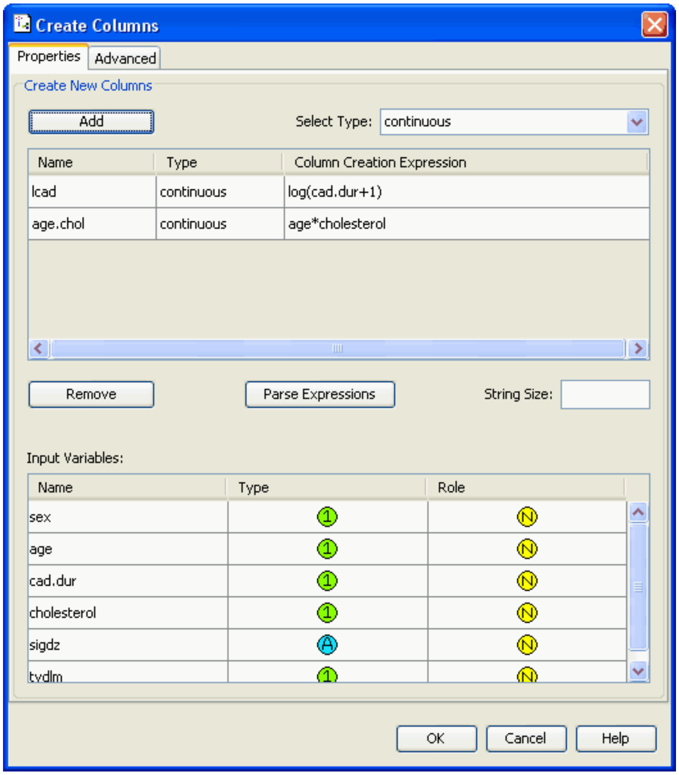
Figure 1.16: *The output of the* ***Descriptive Statistics*** *node, showing the statistics for the variables.*

Next, in the section Manipulate the Data, you can address the skewed nature of `cad.dur` by creating a log transformation of that variable, and, for demonstration's sake, create a new variable, `age*cholesterol`, which you can include in the model. Later, in the section Create Model, create a logistic regression test and a classification tree test for the `sigdz` prediction.

4. Close the **Descriptive Statistics** viewer.

Manipulate the Data

To create the log transformation of `cad.dur`, use an expression to create a new variable called `lcad`. Likewise, use an expression to create the variable `age.chol`. You can create these two new variables using the **Create Columns** node.



Select Type: continuous

Name	Type	Column Creation Expression
lcad	continuous	log(cad.dur+1)
age.chol	continuous	age*cholesterol

String Size:

Input Variables:

Name	Type	Role
sex	1	N
age	1	N
cad.dur	1	N
cholesterol	1	N
sigdz	A	N
tvdlm	1	N

Figure 1.17: *Create new columns (`lcad` and `age.chol`) by manipulating existing variables using the **Create Columns** node.*

To create these columns:

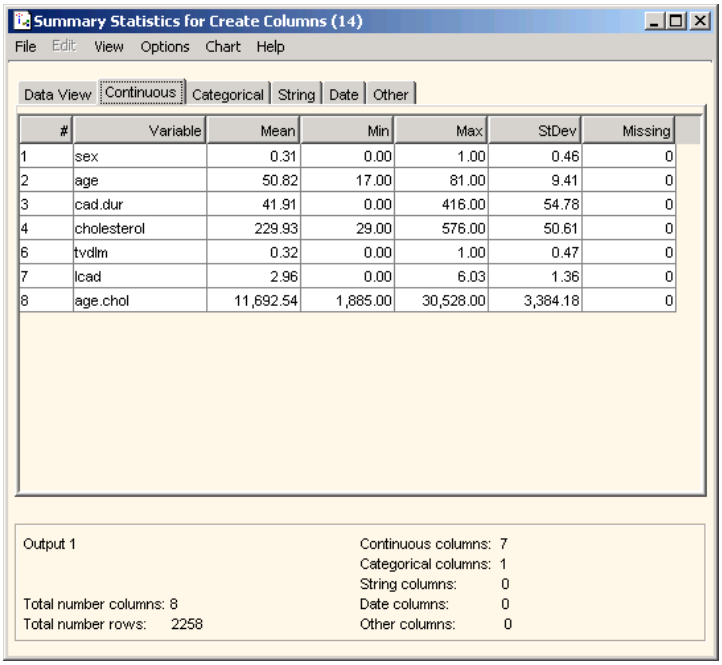
1. Right-click the **Create Columns** node and select **Properties**.
2. From the **Select Type** drop-down list, click **continuous**.
3. Click **Add**.
4. Under **Name**, type `lcad`, and under **Column Creation Expression**, type `log(cad.dur+1)`.
5. Click **Add** again.

6. Under **Name**, type `age.chol`, and under **Column Creation Expression**, type `age*cholesterol`, as shown in Figure 1.17.

7. Click **OK**.

8. From the toolbar, click **Run to Here** (▶|) to run the **Create Columns** node, and then click **Viewer** (👁) to see the data set with the two new columns.

9. Click the **Continuous** tab to see the new variables displayed, as in Figure 1.18.



Summary Statistics for Create Columns (14)
File Edit View Options Chart Help
Data View
Continuous
Categorical
String
Date
Other

#	Variable	Mean	Min	Max	StDev	Missing
1	sex	0.31	0.00	1.00	0.46	0
2	age	50.82	17.00	81.00	9.41	0
3	cad.dur	41.91	0.00	416.00	54.78	0
4	cholesterol	229.93	29.00	576.00	50.61	0
6	tvdlm	0.32	0.00	1.00	0.47	0
7	lcad	2.96	0.00	6.03	1.36	0
8	age.chol	11,692.54	1,885.00	30,528.00	3,384.18	0

```
Output 1                              Continuous columns:  7
                                      Categorical columns: 1
                                      String columns:      0
Total number columns: 8               Date columns:        0
Total number rows:      2258          Other columns:       0
```

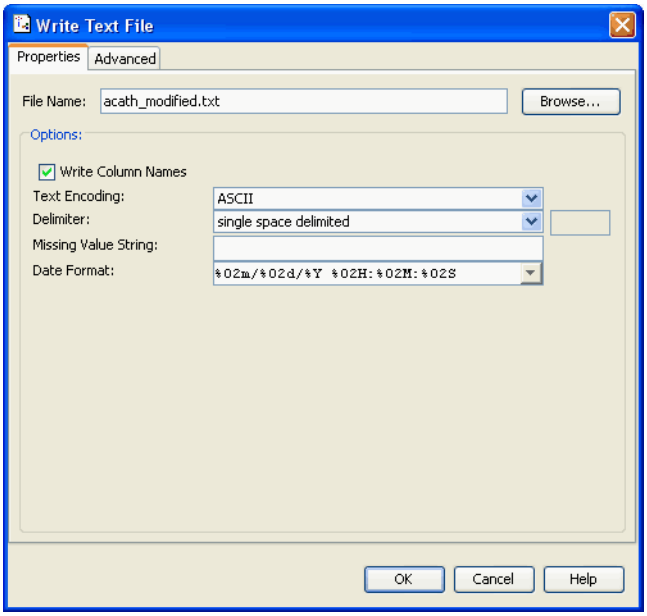
Figure 1.18: *Two new columns,* `lcad` *and* `age.chol`*, are created when you run the* ***Create Columns*** *node.*

10. Close the node viewer.

With the addition of the new variables, you have all the data that you need to create a model and make a prediction for the `sigdz` response variable.

Before you create the model, save the modified data set by writing it to a text file. This way, you can retrieve the data for future reference:

1. Just above the **Create Columns** node, double-click the **Write Text File** node.
2. For **File Name**, type **acath_modified.txt** and for **Delimiter**, select **single space delimited**. Click **OK**, and then click **Run to Here** (▶|). The file is saved to the **examples** directory. If the file **acath_modified.txt** already exists, you can write over it by saving it.



File Name: acath_modified.txt

☑ Write Column Names

Text Encoding: ASCII

Delimiter: single space delimited

Missing Value String:

Date Format: %02m/%02d/%Y %02H:%02M:%02S

Figure 1.19: *Use the* ***Write Text File*** *node to output the modified* ***acath.xls*** *Excel data to a text file,* ***acath_modified.txt****. You can now use this data for other analyses.*

CREATE MODEL

Spotfire Miner provides tools for predicting the response variables based on the independent variables. This example demonstrates two methods, a *classification tree* and a *logistic regression*, to determine which predicts `sigdz` better.

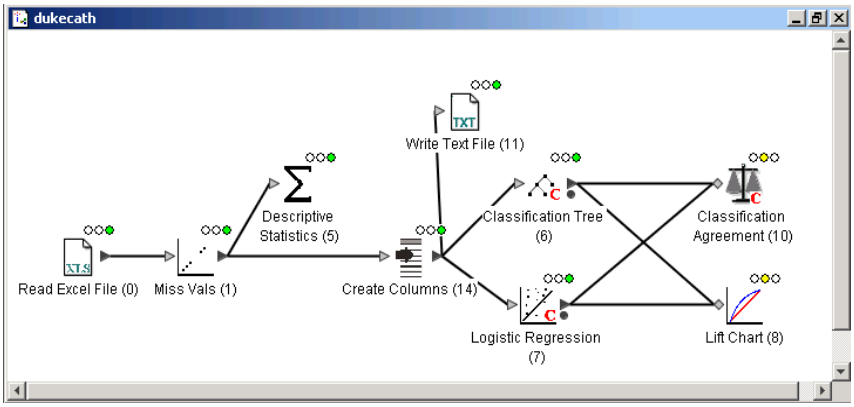


Figure 1.20: *The latter part of the network focuses on comparing the predictions generated by running the* ***Classification Tree*** *and the* ***Logistic Regression*** *nodes. After running these, use the* ***Classification Agreement*** *and* ***Lift Chart*** *nodes to assess the performance of the nodes.*

The `sigdz` variable is binary; that is, it shows that either a patient who comes into the hospital for chest pain actually has significant coronary arterial disease or does not. Binary response data are most commonly modeled by one of two methods: a classification tree or a logistic regression model. The example network illustrates both models.

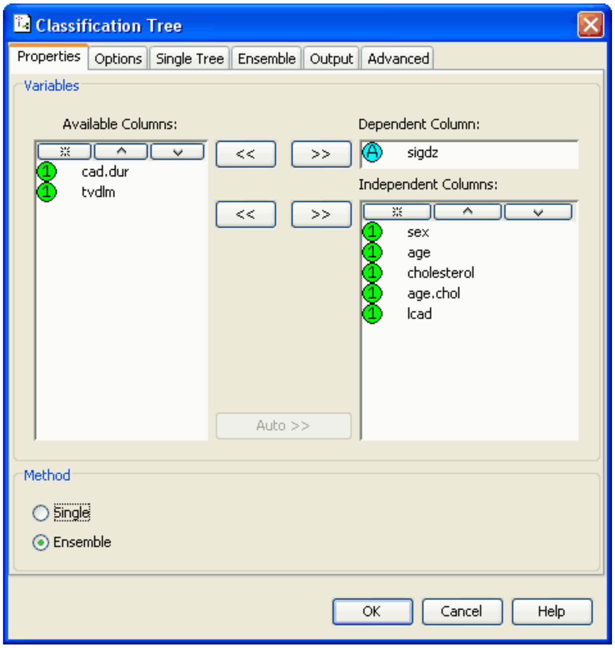
For both modeling components shown, use the same dependent (`sigdz`) and independent variables (`sex`, `lcad`, `age`, `cholesterol`, and `age.chol`) to predict the response.

The example does not use `cad.dur` and `tvdlm` for specific reasons: It uses `lcad` (log of `cad.dur`) instead of `cad.dur`, and `tvdlm` contains information that would not be available in practice to predict `sigdz`.

For each model, specify the categorical variable `sigdz` as the response, or *dependent* variable, and all remaining variables in the data as predictors, or *independent* variables.

Creating the Classification Tree

1. Right-click the **Classification Tree** node and select **Properties** from the shortcut menu. The completed dialog page for this section is shown in Figure 1.21.



Available Columns:
- cad.dur
- tvdlm

Dependent Column: sigdz

Independent Columns:
- sex
- age
- cholesterol
- age.chol
- lcad

Method
- ○ Single
- ● Ensemble

Figure 1.21: *The* ***Properties*** *page for the* ***Classification Tree*** *node allows you to select the dependent and independent variables used in the prediction. Because you are interested in predicting significant coronary disease (*`sigdz`*), it is the dependent variable used in both of the modeling node predictions.*

2. In the **Available Columns** list box, click `sigdz` to select it.

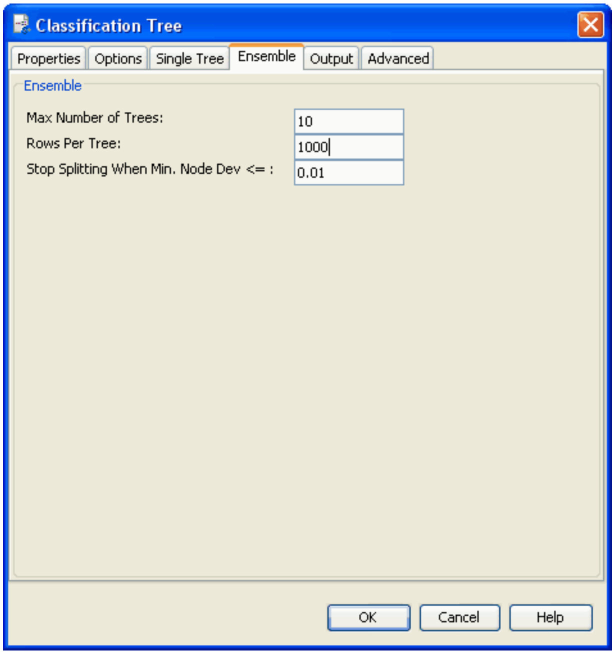
3. Click the >> button to the left of the **Dependent Column** box.

4. In the **Available Columns** list box, CTRL+click `sex`, `age`, `cholesterol`, `age.chol`, and `lcad`.

5. Click the >> button to the left of the **Independent Columns** box.

6. In the **Method** group at the bottom of the page, select **Ensemble**. (A collection of trees is called an *ensemble*. Predictions for the tree model are based on the average from the ensemble.)

7. Click the **Ensemble** tab.

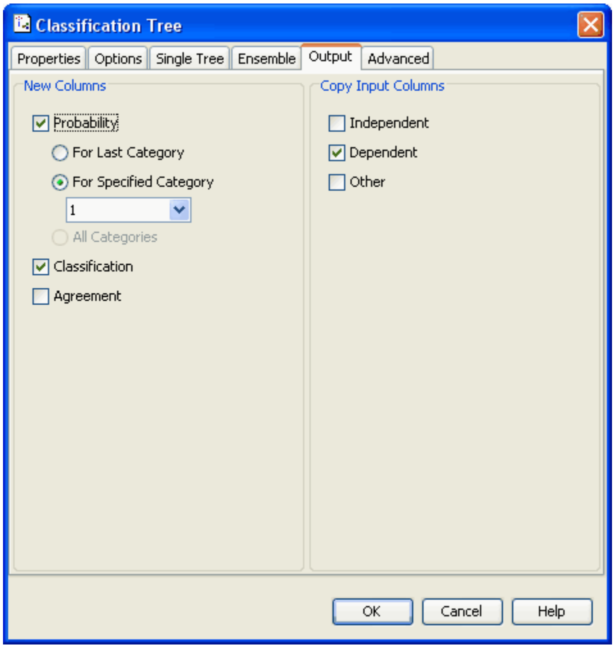


Max Number of Trees: 10
Rows Per Tree: 1000
Stop Splitting When Min. Node Dev <= : 0.01

Figure 1.22: *The* ***Ensemble*** *page of the* ***Classification Tree*** *dialog.*

8. In **Rows Per Tree**, type **1000**. (The **acath.xls** data set has 3504 rows. Your specified value must be fewer than the total number of rows in the data set.) Figure 1.22 shows the completed dialog page for this section.

9. Click the **Output** tab.



New Columns

☑ Probability

○ For Last Category

● For Specified Category

1

○ All Categories

☑ Classification

□ Agreement

Copy Input Columns

□ Independent

☑ Dependent

□ Other

Figure 1.23: *The **Output** page of the **Classification Tree** dialog.*

10. In the **New Columns** group, under **Probability**, click **For Specified Category**, and then select **1** from the drop-down list.

    By default, Spotfire Miner returns the computed probabilities for the last level in the dependent variable. To display the probabilities for level **1**, you must choose the variable **1** by selecting this option explicitly.

    Figure 1.23 shows the completed dialog page for this section.

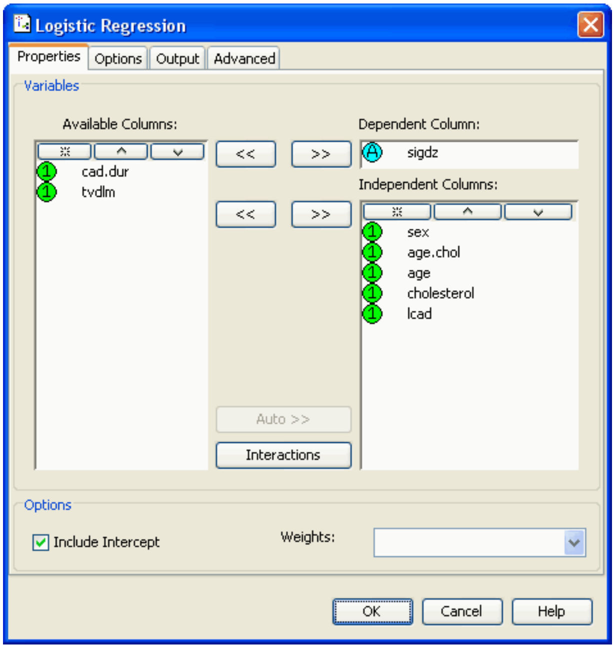
11. Click **OK** to close the dialog.

Creating the Logistic Regression Test

Next, specify the properties of the **Logistic Regression** node.

1. Repeat the Classification Tree steps 1-5 and steps 9-10 for the **Logistic Regression** node.

Figure 1.24 shows the completed properties page for the **Logistic Regression** node.



Available Columns:
cad.dur
tvdlm

Dependent Column: sigdz

Independent Columns:
sex
age.chol
age
cholesterol
lcad

☑ Include Intercept

Weights:

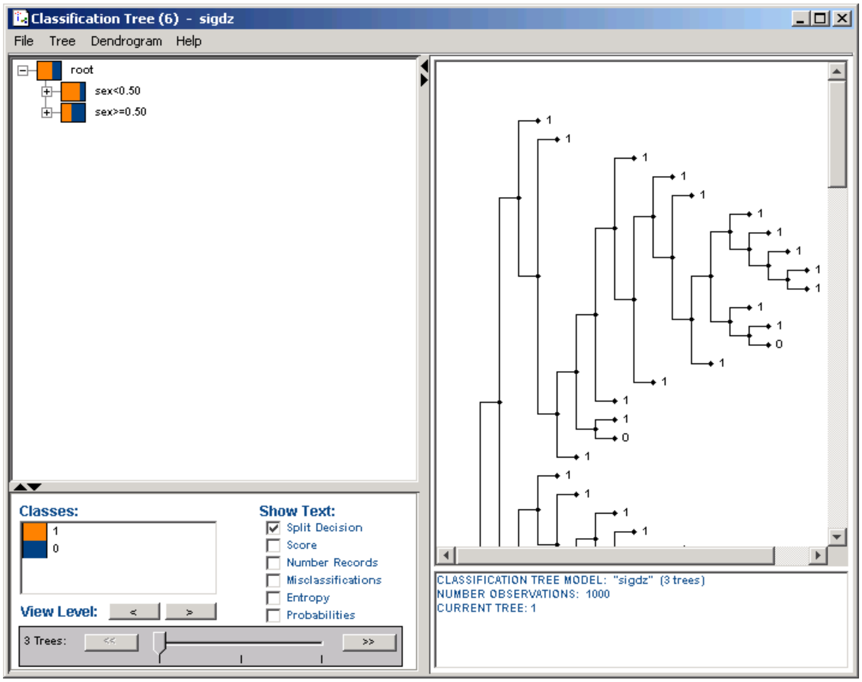
Figure 1.24: *Selecting the variables for the **Logistic Regression** dialog.*

2. Click **OK** to accept the changes.

3. Click the ▶ button to run the network.

View both models:

4. Right-click the **Classification Tree** node and select **Viewer** from the shortcut menu. The **Classification Tree Viewer** opens, as shown in Figure 1.25.



root
sex<0.50
sex>=0.50

Classes:
- 🟧 1
- 🟦 0

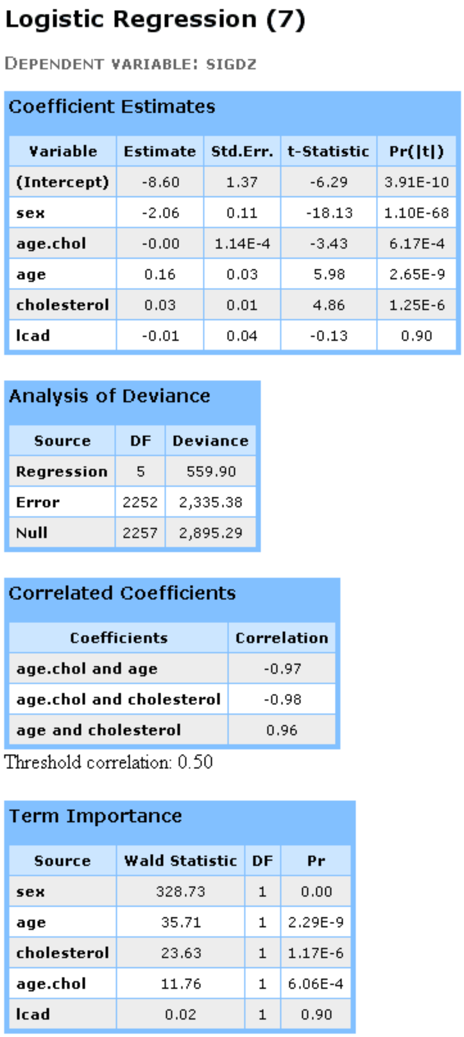
Show Text:
- ☑ Split Decision
- □ Score
- □ Number Records
- □ Misclassifications
- □ Entropy
- □ Probabilities

```
CLASSIFICATION TREE MODEL:  "sigdz"  (3 trees)
NUMBER OBSERVATIONS:  1000
CURRENT TREE: 1
```

Figure 1.25: *The viewer for the **Classification Tree** node.*

The classification tree algorithm fits a separate tree for each *block* of data sent through the pipeline. For these data, the chunk size you chose (1,000 rows) results in three (3) trees, because the dataset contains 2258 rows. The collection of these three trees is called an *ensemble*. To scroll through the trees, click the double-arrow buttons (<< and >>) in the gray pane at the bottom left of the viewer. Predictions made from the ensemble are calculated as averages from the three tree models. This is known as *block model averaging*.

5. Open the viewer for the **Logistic Regression** node. A Web browser window opens, displaying a table of coefficients for the model, as shown in Figure 1.26. (Maximize the browser to see the best results.)



Logistic Regression (7)

DEPENDENT VARIABLE: SIGDZ

Coefficient Estimates

Variable	Estimate	Std.Err.	t-Statistic	Pr(|t|)
(Intercept)	-8.60	1.37	-6.29	3.91E-10
sex	-2.06	0.11	-18.13	1.10E-68
age.chol	-0.00	1.14E-4	-3.43	6.17E-4
age	0.16	0.03	5.98	2.65E-9
cholesterol	0.03	0.01	4.86	1.25E-6
lcad	-0.01	0.04	-0.13	0.90

Analysis of Deviance

Source	DF	Deviance
Regression	5	559.90
Error	2252	2,335.38
Null	2257	2,895.29

Correlated Coefficients

Coefficients	Correlation
age.chol and age	-0.97
age.chol and cholesterol	-0.98
age and cholesterol	0.96

Threshold correlation: 0.50

Term Importance

Source	Wald Statistic	DF	Pr
sex	328.73	1	0.00
age	35.71	1	2.29E-9
cholesterol	23.63	1	1.17E-6
age.chol	11.76	1	6.06E-4
lcad	0.02	1	0.90

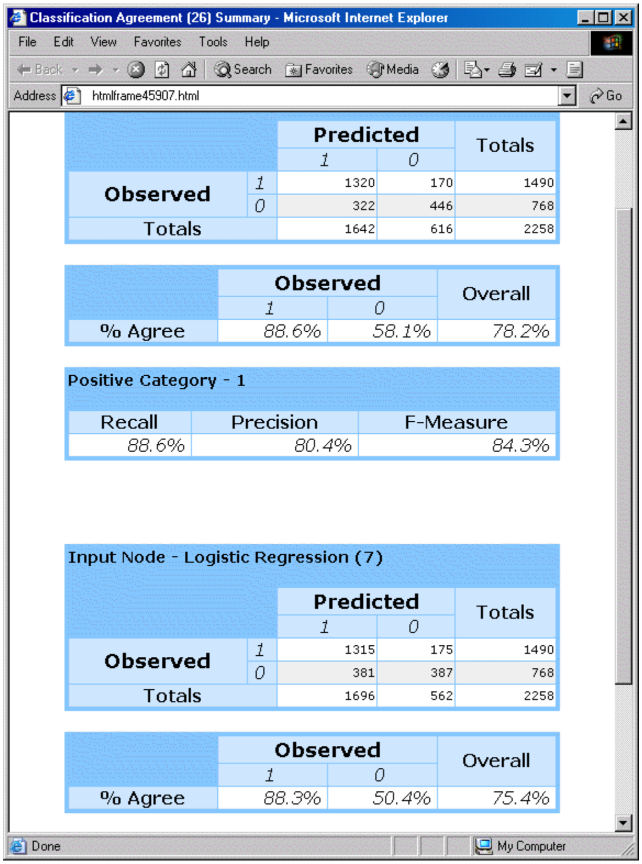
Figure 1.26: *The viewer for the **Logistic Regression** node.*

6. When you are finished examining the nodes, close both viewers.

Comparing Models

To compare the classification tree and logistic regression models from the previous section, use two different nodes: a *classification agreement* node and a *lift chart* node.

The **Classification Agreement** node compares the accuracy of multiple classification models; in this case, it's the output from the **Classification Trees** and the **Logistic Regression** nodes. It uses the predicted values from a model to produce a *confusion* matrix, which indicates the number and proportion of observations that are classified correctly by the model, as shown in Figure 1.27.


Classification Agreement (26) Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address htmlframe45907.html
Go
Done
My Computer

		Predicted		Totals
		1	0	
Observed	1	1320	170	1490
	0	322	446	768
Totals		1642	616	2258

	Observed		Overall
	1	0	
% Agree	88.6%	58.1%	78.2%

Positive Category - 1

Recall	Precision	F-Measure
88.6%	80.4%	84.3%

Input Node - Logistic Regression (7)

		Predicted		Totals
		1	0	
Observed	1	1315	175	1490
	0	381	387	768
Totals		1696	562	2258

	Observed		Overall
	1	0	
% Agree	88.3%	50.4%	75.4%

Figure 1.27: *The output from the **Classification Agreement** node, which compares the output from the **Classification Trees** node and the **Logistic Regression** nodes.*

Note that the overall accuracy of the **Classification Tree** node is 78.2%, while that of the **Logistic Regression** node is 75.4%. The **Classification Tree** node is only a slightly better predictor overall, but note that it also predicts the absence of significant coronary disease better (58.1% vs. 50.4%), so it is a better model overall.

The other node used for comparison is the classic *lift chart*, which compares the gain in response, or *lift*, of one model to that of another model. The lift is also compared to doing nothing, which is shown as a straight reference line on the lift chart.

1. Open the viewer for the **Lift Chart** node.
2. Under **Chart Type**, select **Cumulative Gain**, as shown in Figure 1.28.

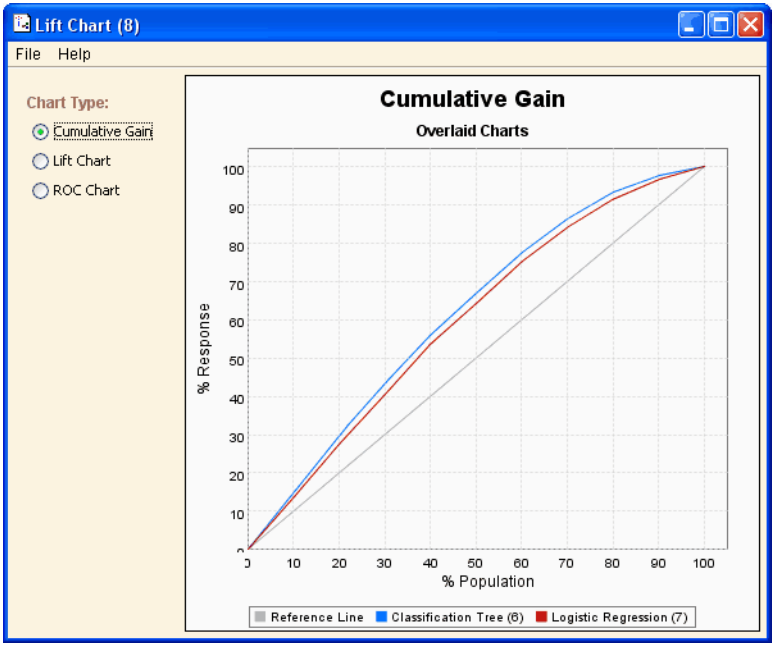


Chart Type:

● Cumulative Gain
○ Lift Chart
○ ROC Chart

Figure 1.28: *The viewer for the **Lift Chart** node.*

Notice that the classification tree, displayed as the blue (upper) line, shows slightly more lift than the logistic regression model (the red line). The classification tree might be the slightly more preferred model for predicting `sigdz`.

SUMMARY

You can draw at least two conclusions from this example using this dataset:

1. The classification tree is slightly more accurate for predicting significant coronary disease than logistic regression,
2. You can predict reliably the probability a patient has significant coronary arterial disease 88.6% of the time, if you use the classification tree model.

Note that the model did not use the `tvdlm` variable, which is *severe* coronary arterial disease. Another study you could perform is the probability a patient has severe coronary arterial disease, given he has exhibited *significant* coronary arterial disease. You could run this analysis by subsetting the data, or by using the significant coronary arterial disease cases as an indicator the patient will develop severe coronary arterial disease.

It is important to note that, for the purpose of simplicity, this model was run on only one specific set of data, which is called the *training* data. Ideally, you would use a **Partition** node to use a percentage of the data for *training* (running the model) and another part for *testing* (predicting the model), and finally using a new dataset for *validating* (confirming the model).

The example in Chapter 2, Integrating Spotfire Miner with Spotfire, demonstrates visualizing model predictions from Spotfire Miner using Spotfire, and how you can move between Spotfire Miner and Spotfire.

The example in Chapter 3, An Extended Tour, illustrates a more complex example, using training and testing data to create a model to predict the probability of home mortgage loan defaulting by customers. This model is then used to score a new data set. As you get more familiar with this tool, you can see how to customize the capabilities of Spotfire Miner for your data mining application.

INTEGRATING SPOTFIRE MINER WITH SPOTFIRE

2

INTRODUCTION

This chapter continues where the last chapter left off, with the successful completion of the dukecath worksheet. In this quick start chapter, we introduce integrating the results of a Spotfire Miner analysis with TIBCO Spotfire®.

The TIBCO Spotfire package is a flexible visualization tool that neatly complements the capabilities of Spotfire Miner. You can use the two smoothly; together, they constitute a powerful analysis combination.

This chapter demonstrates how you can visualize the model predictions using Spotfire, and how you can move between Spotfire Miner and Spotfire. Changes that you make to data preparation and models can be displayed instantly in Spotfire.

Note

To walk through this exercise, you must have both Spotfire Miner and Spotfire installed.

PREPARE THE DATA FILE FOR SPOTFIRE

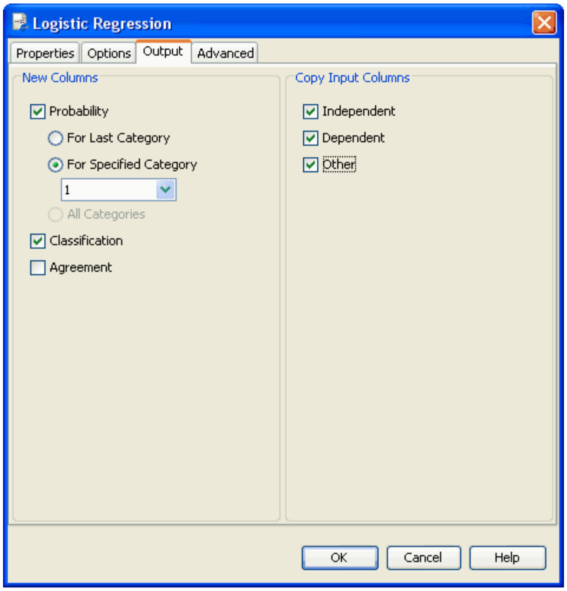
In this section, start with the completed worksheet from the Duke study, and then output the model results as a text file to import into Spotfire.

Saving the File

First, change the model nodes so they output all of the original data as well as the model results. This provides some useful visualization options when you view the results in Spotfire.

1. If Spotfire Miner is not running, start it.
2. From the main menu, choose **File ▶ Open** to open the example worksheet.
3. In the **Open** dialog, click **Examples**, and then open the **SpotfireIntegration** folder.
4. Select the **dukecath.SpotfireEnhancement** worksheet, and then click **Open**.
5. Open the **Logistic Regression** dialog.

6. To replicate the original data set and append the model results, in the **Output** tab, select all three check boxes in the **Copy Input Columns** group. (See Figure 2.1). .



New Columns

☑ Probability

○ For Last Category

● For Specified Category

1

○ All Categories

☑ Classification

□ Agreement

Copy Input Columns

☑ Independent

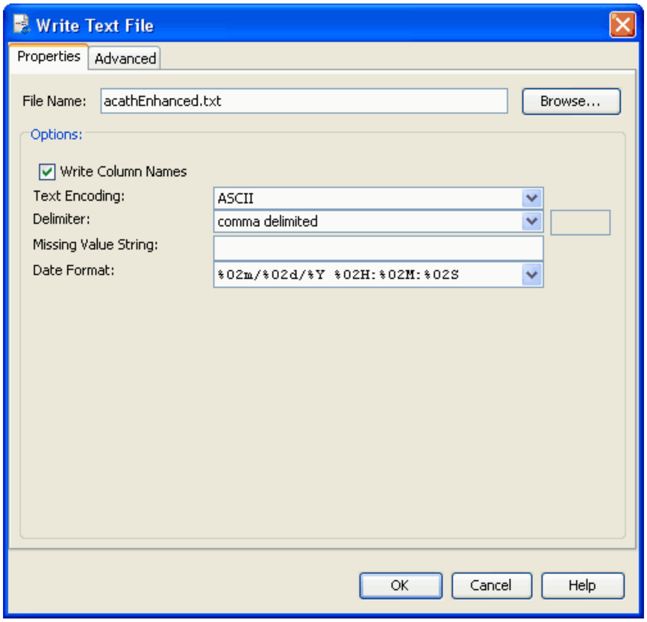
☑ Dependent

☑ Other

Figure 2.1: ***Logistic Regression Output*** *options to copy input columns.*

7. In the Explorer pane, expand the **Data Output ▶ File** folder, and then drag a **Write Text File** node to the worksheet. Position it below and to the right of the **Logistic Regression** node, and connect the two nodes.

8. Double-click the **Write Text File** node to open its properties dialog.



File Name: acathEnhanced.txt

Options:

☑ Write Column Names

Text Encoding: ASCII

Delimiter: comma delimited

Missing Value String:

Date Format: %02m/%02d/%Y %02H:%02M:%02S

Figure 2.2: ***Write Text File*** *dialog.*

9. Create a new text file called **acathEnhanced.txt** to contain the original data with additional model information appended to it. (This exercise uses a relative path name; although, you could provide the fully-qualified path name instead, if you prefer. Later, you will overwrite this file with output from the tree model.)

10. Click **OK** to close the dialog, and then right-click the node and select **Rename**.

11. Rename the node with the file name. (See Figure 2.3.) This practice makes it easier for you to remember which file is involved. The new file is saved to your working directory.

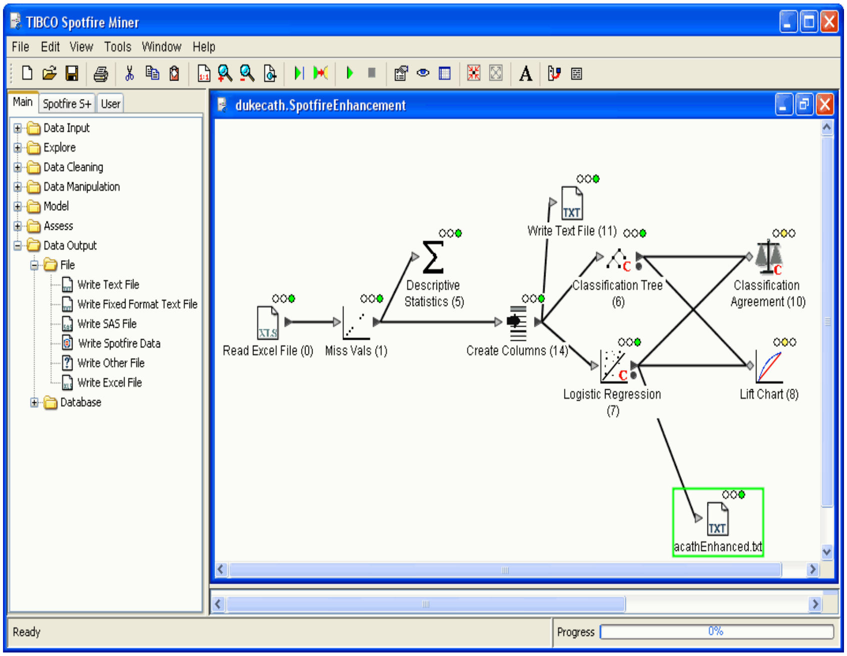


Figure 2.3: ***Write Text File*** *node.*

Now, you are ready to import this file into Spotfire and create a visualization.

CREATING A VISUALIZATION

In this section, work with Spotfire to import the Spotfire Miner data set and create an initial visualization. Then, manipulate the visualization using Spotfire's tools to provide clear and useful results.

1. Without closing Spotfire Miner, start a session of Spotfire.
2. From the menu, select **File** ▶ **Open**.
3. Browse to the working directory where your exported file **acathEnhanced.txt** is saved. Click **Open**.

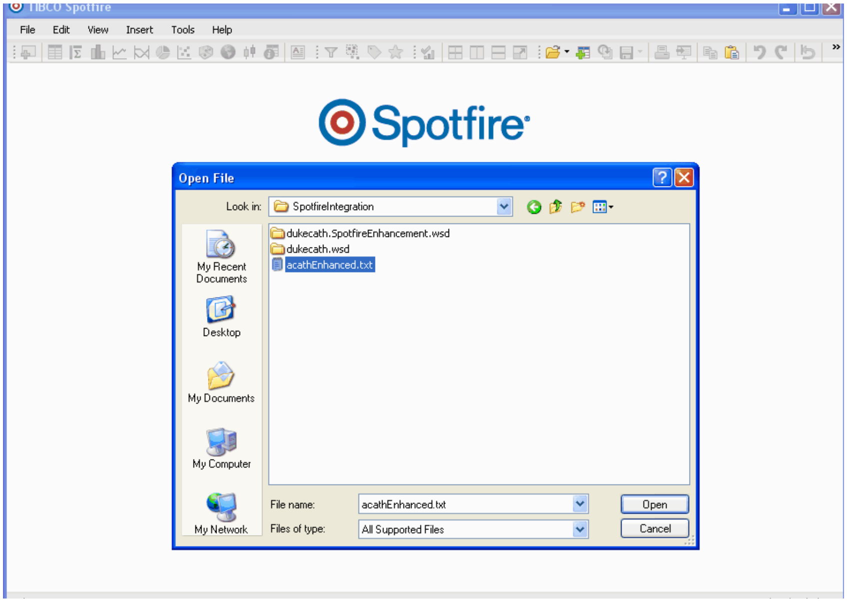
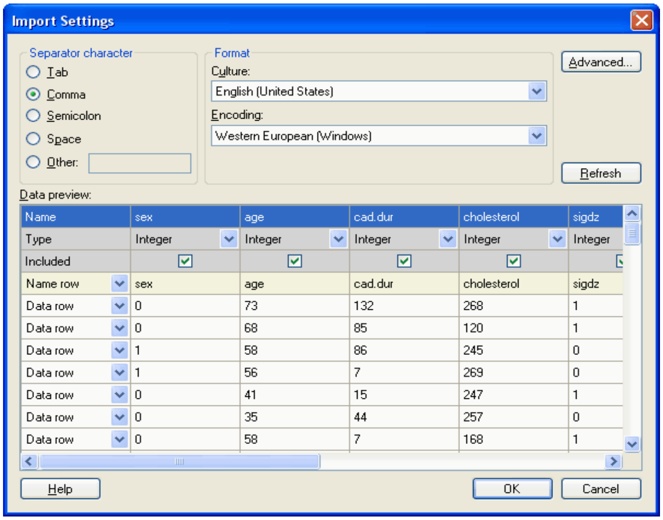


Figure 2.4: *Spotfire's **Open File** dialog.*

Spotfire samples the data and displays a viewer with data choices. (See Figure 2.5.)

4. For this example, click **OK** to accept all defaults.



Separator character

○ Tab
● Comma
○ Semicolon
○ Space
○ Other:

Format

Culture: English (United States)

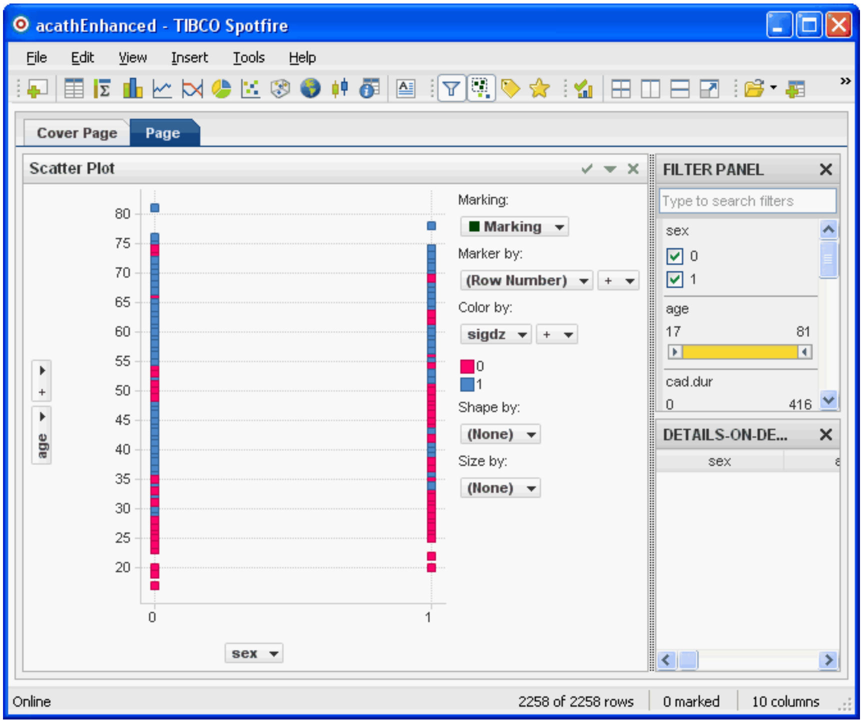
Encoding: Western European (Windows)

Data preview:

Name	sex	age	cad.dur	cholesterol	sigdz
Type	Integer	Integer	Integer	Integer	Integer
Included	☑	☑	☑	☑	☑
Name row	sex	age	cad.dur	cholesterol	sigdz
Data row	0	73	132	268	1
Data row	0	68	85	120	1
Data row	1	58	86	245	0
Data row	1	56	7	269	0
Data row	0	41	15	247	1
Data row	0	35	44	257	0
Data row	0	58	7	168	1

Figure 2.5: *Spotfire's* ***Import Settings*** *dialog.*

By default, Spotfire displays a scatter plot of possibly interesting results.



Marking: Marking
Marker by: (Row Number)
Color by: sigdz
🟥 0
🟦 1
Shape by: (None)
Size by: (None)

sex
☑ 0
☑ 1

age
17 81

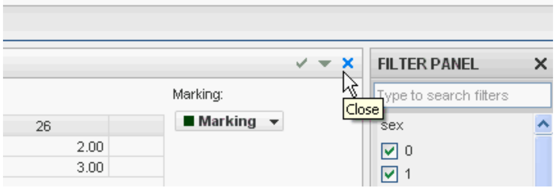
cad.dur
0 416

2258 of 2258 rows | 0 marked | 10 columns

Figure 2.6: *Spotfire's default display of the data.*

Instead of using this default representation, we will create a 3D presentation.

5. Close the scatter plot by clicking the close button for the visualization (see Figure 2.7).



Marking: Marking

sex
☑ 0
☑ 1

Figure 2.7: *Close button for the visualization.*

6. On the toolbar, click the **New 3D Scatter Plot** button to display the new visualization.

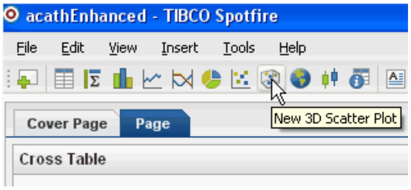
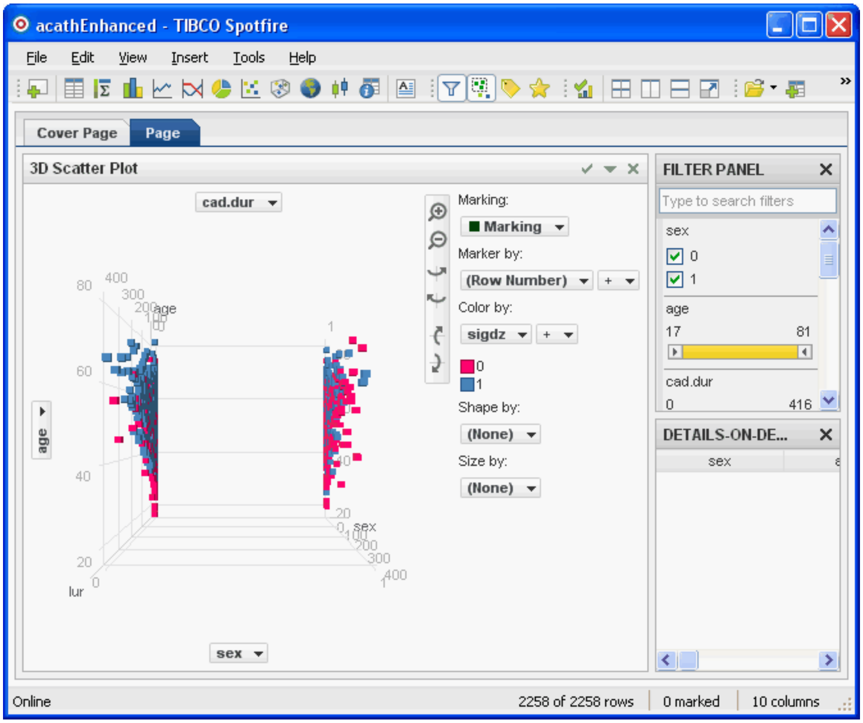

acathEnhanced - TIBCO Spotfire
File Edit View Insert Tools Help
New 3D Scatter Plot
Cover Page
Page
Cross Table

Figure 2.8: *New 3D Scatter Plot button.*

7. Review the results.



Marking: Marking
Marker by: (Row Number)
Color by: sigdz
🟥 0
🟦 1
Shape by: (None)
Size by: (None)

sex
☑ 0
☑ 1
age
17 81
cad.dur
0 416

Online 2258 of 2258 rows 0 marked 10 columns

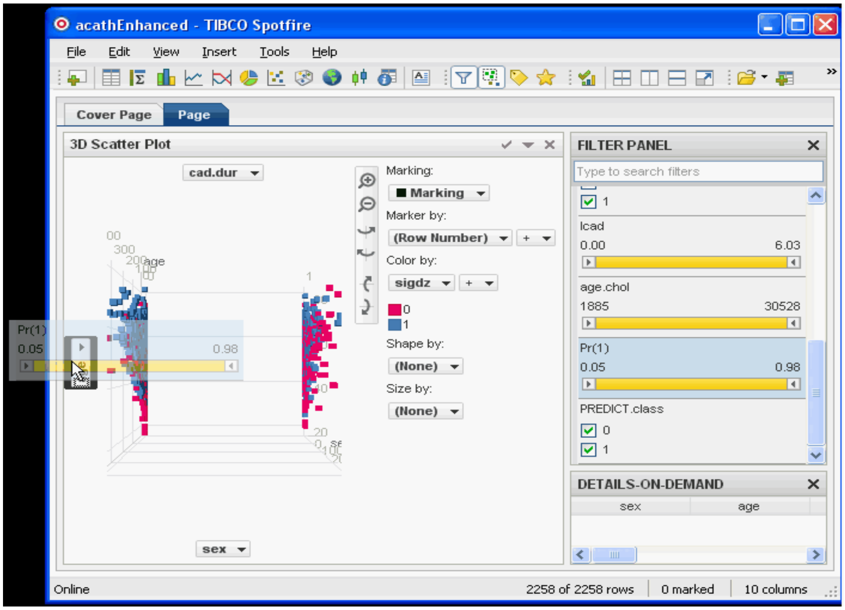
Figure 2.9: *Initial 3D Scatter Plot visualization.*

The 3-D scatter plot display also chooses variables that might be interesting. The choice of color coding is displayed in the legend. In this case, Spotfire has chosen the actual dependent variable to control the red/blue colors that are useful. Note that the `sex` variable is binary. This display perspective looks down one of the dimensions of the cube; already the dependent appears to be strongly sex linked, as seen by the predominance of red in the right group & blue in the left.

Spotfire Controls

The series of check boxes and sliders in the right pane are Spotfire's filter controls. You can click any of the filter variables and drag them into the visualization. For this exercise, show the probability variable, `Pr(1)`, in the axis that is vertical in the plot.

8. Click the `Pr(1)` variable and drag it into the workspace until it overlays the age tab (the vertical, or Y axis), as shown in Figure 2.10.



Marking: Marking
Marker by: (Row Number)
Color by: sigdz
🟥 0
🟦 1
Shape by: (None)
Size by: (None)

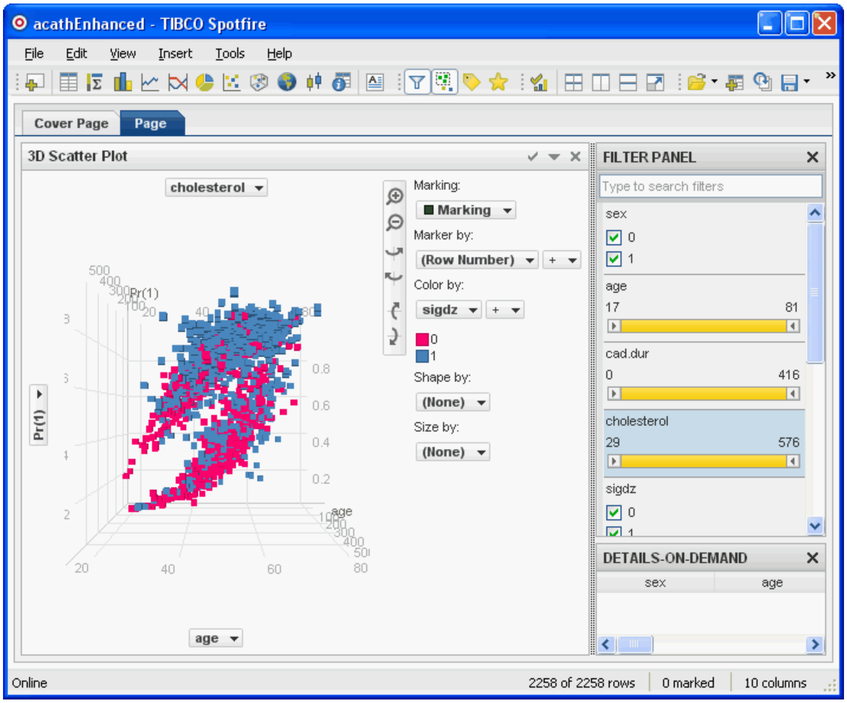
☑ 1
lcad 0.00 6.03
age.chol 1885 30528
Pr(1) 0.05 0.98
PREDICT.class
☑ 0
☑ 1

Online 2258 of 2258 rows 0 marked 10 columns

Figure 2.10: *Dragging the `Pr(1)` variable to the vertical axis.*

9. Review the results and note that the visualization is rearranged. Note that `sex` is still on the horizontal (X) axis and `cad.dur` is on the depth (Z) axis.

10. Repeat step 8, dragging `age` to the X axis and `cholesterol` to the Z axis. (See Figure 2.11).



acathEnhanced - TIBCO Spotfire

3D Scatter Plot

Marking: Marking

Marker by: (Row Number)

Color by: sigdz

Shape by: (None)

Size by: (None)

FILTER PANEL

sex	
☑ 0	
☑ 1	
age	
17	81
cad.dur	
0	416
cholesterol	
29	576
sigdz	
☑ 0	
☑ 1	

DETAILS-ON-DEMAND

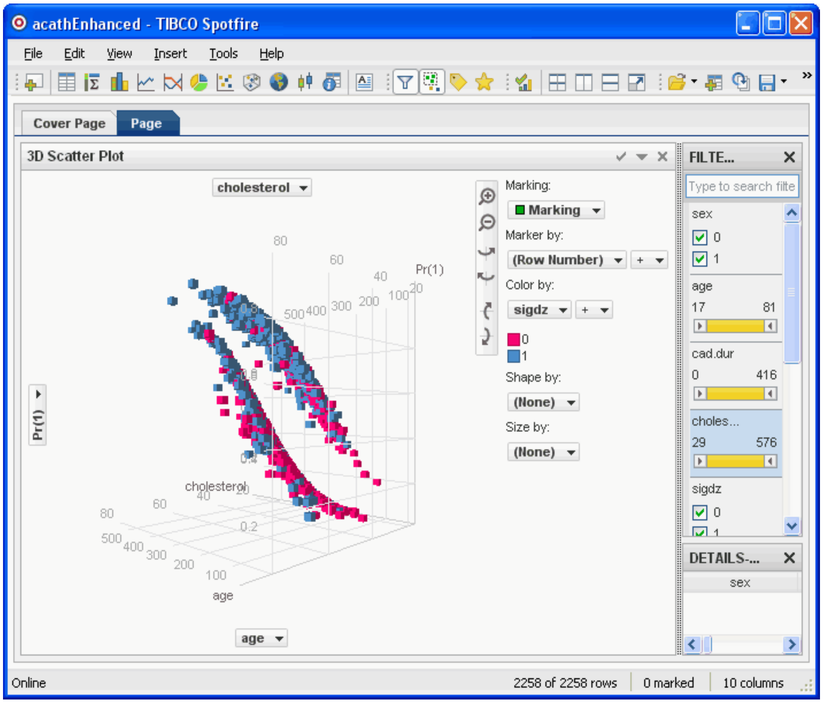
sex	age

Online 2258 of 2258 rows 0 marked 10 columns

Figure 2.11: *Visualization after changing the axes.*

Each point of the visualization represents one patient, and the location is plotted by the patient's age, cholesterol level, and (from the model) calculated probability of having significant coronary disease. The points are colored by the actual presence or absence of disease.

Figure 2.12 demonstrates using the rotation controls to change the viewpoint of the 3D image. Note that the points fall into two broad surfaces.


acathEnhanced - TIBCO Spotfire

3D Scatter Plot

Marking: Marking

Marker by: (Row Number)

Color by: sigdz

🟥 0

🟦 1

Shape by: (None)

Size by: (None)

FILTE...

sex: ☑ 0 ☑ 1

age: 17 81

cad.dur: 0 416

choles...: 29 576

sigdz: ☑ 0 ☑ 1

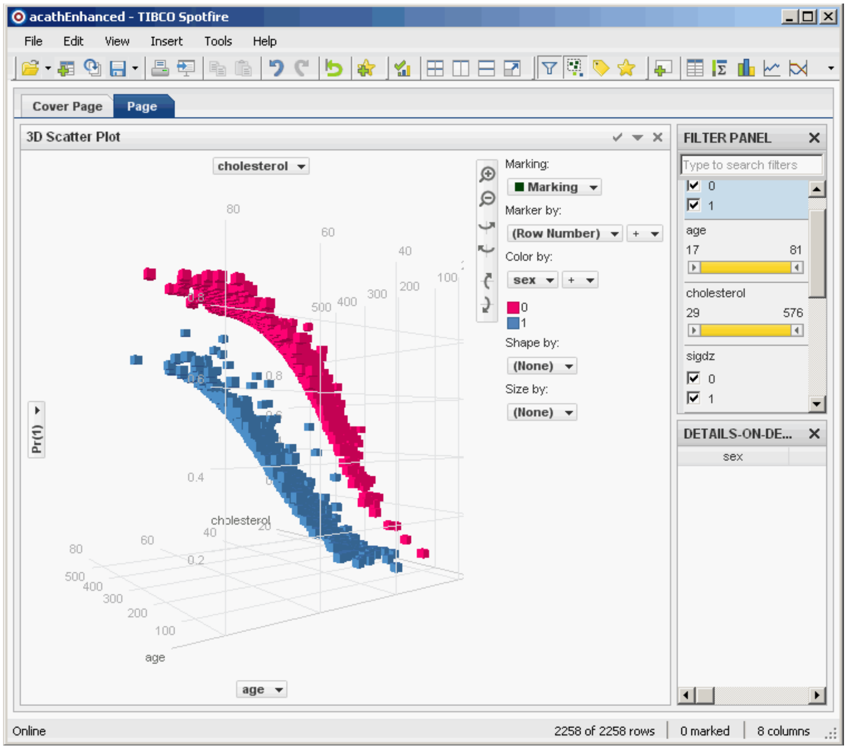
DETAILS-...

sex

Online | 2258 of 2258 rows | 0 marked | 10 columns

Figure 2.12: *The visualization after rotating.*

The two broad surfaces turn out to correspond to the two sexes. You can demonstrate this correspondence by setting the **Color by** control to `sex`. (See Figure 2.13.).



Marking: Marking

Marker by: (Row Number)

Color by: sex

🟥 0

🟦 1

Shape by: (None)

Size by: (None)

☑ 0

☑ 1

age

17 81

cholesterol

29 576

sigdz

☑ 0

☑ 1

2258 of 2258 rows | 0 marked | 8 columns

Figure 2.13: *Rotated visualization with the* ***Color by*** *control set to* `sex`.

In the next section, use Spotfire Miner to change the model, and then reload the data to see how the new model affects the visualization.

DYNAMICALLY UPDATE YOUR DATA

This section demonstrates a useful interactive feature that greatly smooths the interactive nature between Spotfire Miner and Spotfire.

1. Without closing Spotfire, return to your open instance of Spotfire Miner.
2. Open the **Logistic Regression** dialog, and in the **Properties** page, select `age.chol` in the **Independent Columns** list box. Move it to the **Available Columns** list box.

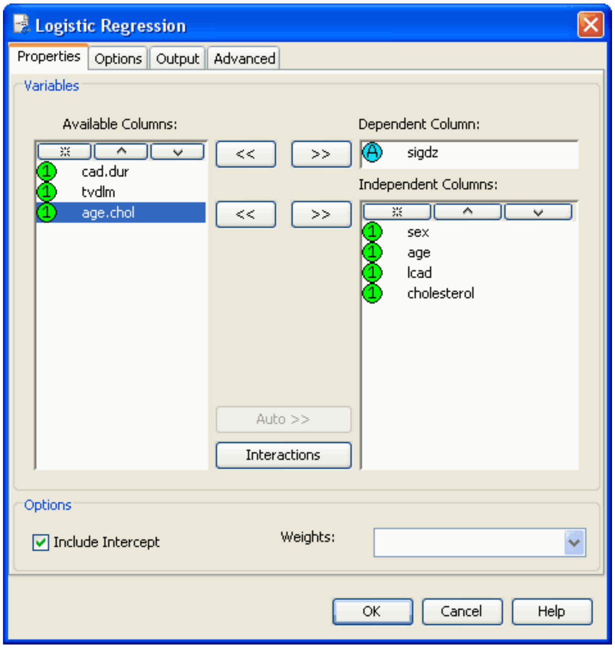


Figure 2.14: *Moving* `age.chol` *out of the model.*

3. Click **OK** to save the change.
4. Rerun the **Logistic Regression** and **acathEnhanced.txt** nodes.
5. Return to Spotfire without closing Spotfire Miner.

6. On the toolbar, click the **Reload Data** button.

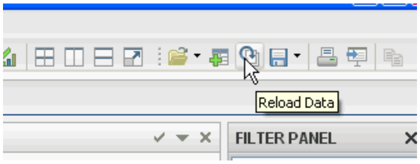


Figure 2.15: *The **Reload Data** button.*

After the data refreshes, the visualization is redrawn. (Note that the data have the same number and names of variables, so refreshing the data is very straightforward.)

7. Note that the surfaces simplify in their shape, reflecting the simplified model.

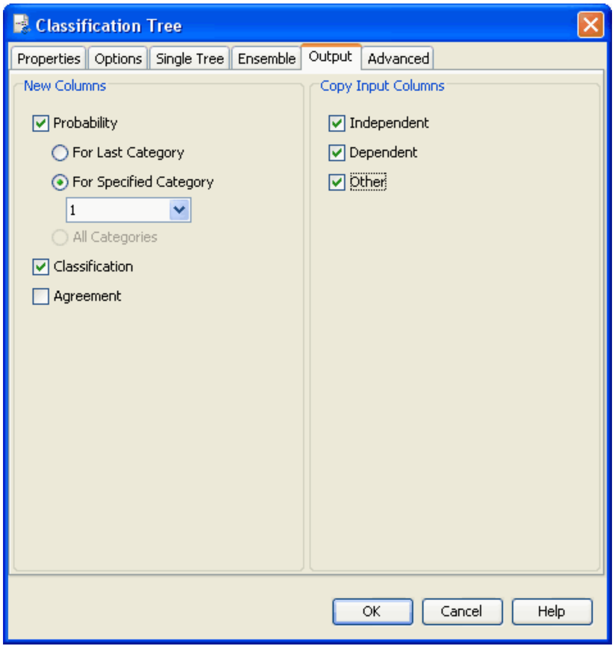
   This interactivity is very useful when you want to change data preparation or a model, and you want to view the effects of the change.

CHANGE TO A DIFFERENT MODEL

In this section, explore how the tree model predictions compare.

Return to Spotfire Miner and double-click the **Classification Tree** node to display its dialog.

To replicate the number and names of the variables the **Classification Tree** node produces, in the **Output** tab, select all three check boxes in the **Copy Input Columns** group. (See Figure 2.16.)



New Columns

☑ Probability

○ For Last Category

● For Specified Category

1

○ All Categories

☑ Classification

□ Agreement

Copy Input Columns

☑ Independent

☑ Dependent

☑ Other

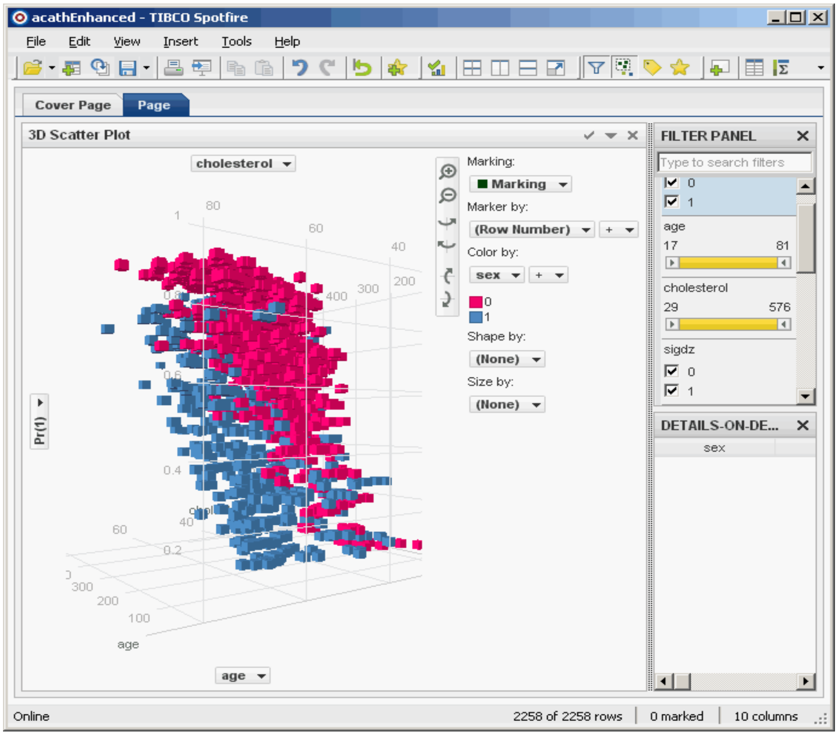
Figure 2.16: ***Classification Tree Output*** *options to copy input columns.*

Right-click the **Write Text File** node named **acathEnhanced.txt** and from the drop-down menu, select **Copy**.

Paste the copy above and to the right of the **Classification Tree** node.

Connect the node copy to the output of the **Classification Tree** and run the worksheet through this node. (Note that doing so overwrites the output file with results from the classification tree.)

Return to Spotfire and reload the data again. Note how the visualization changes. Figure 2.17 shows the new visualization from approximately the same angle as Figure 2.12.



Marking: Marking

Marker by: (Row Number)

Color by: sex

🟥 0
🟦 1

Shape by: (None)

Size by: (None)

☑ 0
☑ 1

age
17 81

cholesterol
29 576

sigdz
☑ 0
☑ 1

Online 2258 of 2258 rows 0 marked 10 columns

Figure 2.17: *Classification Tree visualization.*

The tree model's results (as expressed by the `Pr(1)` levels) are more complex than the logistic model. By setting the **Color by** control to `sex`, you can confirm that the general location of the corresponding clouds of points still correspond roughly to the logistic's model. (That is, the males still have higher overall predicted probability of coronary disease.)

This example demonstrates how Miner and Spotfire neatly complement each other to constitute a powerful analysis platform.

AN EXTENDED TOUR 3

INTRODUCTION

In this Extended Tour, use the modeling utilities in Spotfire Miner to forecast financial data. In this example, develop a model to predict which customers will default on their home mortgage loan. For example, imagine you work for a private mortgage company and want to buy loans from another mortgage company, but you have decided that you want to be conservative with the risk that you take. You want to buy home mortgage loans for which the probability of not defaulting is greater than .98.

Home loan mortgages are a big business in the U.S. Not only is there a huge primary market for home financing and refinancing, but there is also a very active secondary market, in which loan portfolios are actively traded. Loans are valued according to the risk of default and no-default. A key problem is to build a model that can predict loan default based on known attributes of the customer or pool of customers (credit score, loan history, house value, and so on). This model is valuable not only in the secondary market, where the problem is to accurately value the loans, but also in the primary market, where the problem is to build a successful loan origination strategy.

Statistical modeling offers huge benefits in this area. The relationships between the response and predictors (for example, probability of default given the customer history) are strong and interpretable. Spotfire Miner is ideally suited to this problem, because it offers advanced semi-parametric and non-parametric methods that do not assume a specific parametric (for example, linear) form between the response and predictors.

In this Extended Tour example, you develop several models to fit a set of training data, compare the models using new data (testing data), and then choose the best performing model. Using the model that you determine to be the best, predict or score a list of potential loan customers. Filter this list according to the risk you are willing to take and decide which loans to buy.

Data Mining the TIBCO Spotfire Miner Way

TIBCO has defined a process for data mining based on experience acquired in developing and deploying real-world data-mining solutions. Figure 3.1 presents a high-level view of this process. The example follows these steps.

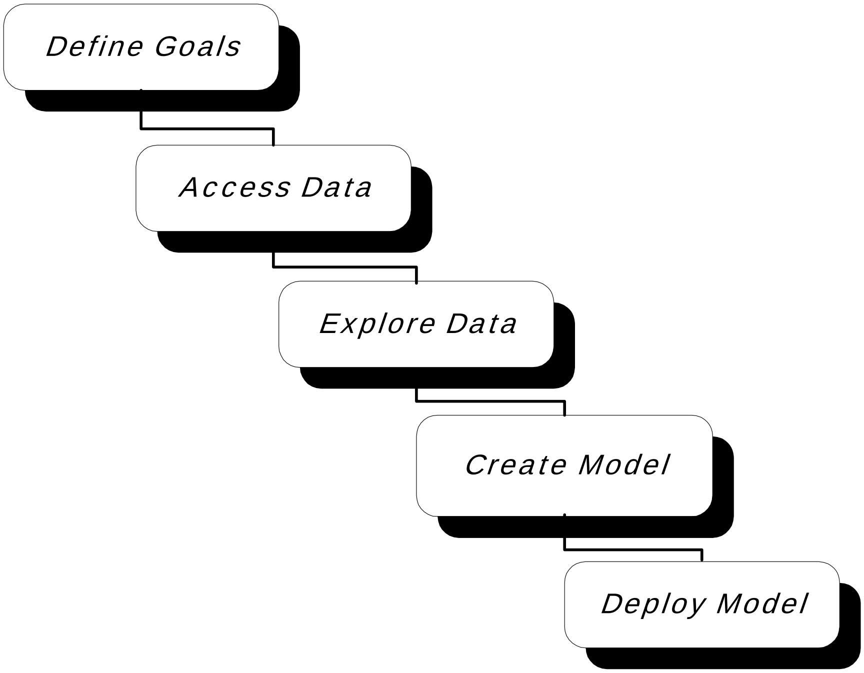
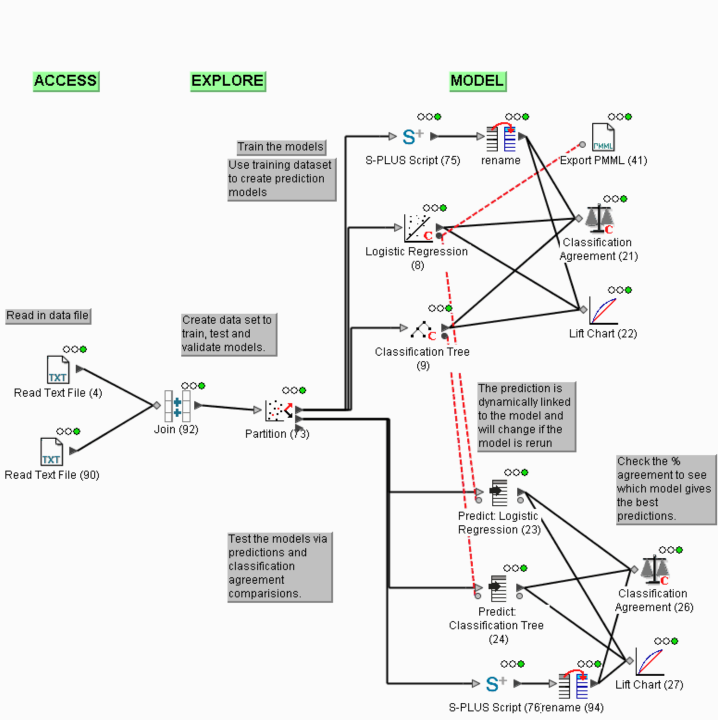


Figure 3.1: *The data mining process: Define Goal, Access Data, Explore Data, Create Model, and Deploy Model.*

In the sections that follow, you can divide these steps further to see how to translate this high-level view into a practical approach for solving a real-world data mining problem using Spotfire Miner's advanced modeling and analysis capabilities.

The Spotfire Miner worksheet below shows the example of the phases outlined in Figure 3.1. The upper network shown in Figure 3.2 represents the accessing, exploring and modeling phases, while the lower network represents scoring (deployment) using a Spotfire S+ model.



Train the models

Use training dataset to create prediction models

Read in data file

Create data set to train, test and validate models.

The prediction is dynamically linked to the model and will change if the model is rerun

Check the % agreement to see which model gives the best predictions.

Test the models via predictions and classification agreement comparisons.

Figure 3.2: *The completed networks are in a worksheet called **MortgageDefault.complete.imw**. This predicts the probability of customers not defaulting on home mortgage loans.*

You can find the data and example worksheets created in this Extended Tour chapter in the **examples/ MortgageDefaultExample** folder. You can solve the entire problem in one worksheet, as shown in Figure 3.2; however, the example builds the solution through a series of worksheets.

DEFINE GOALS

As discussed in the section Introduction on page 52, the problem is to predict the probability of obtaining no-default loans given the available data on customers. The final result will be a text file containing the customers least likely to default (`Pr(NoDefault)>.98`). (That is, the probability of not defaulting is greater than 98%.)

This example contains two data sets from which to create a predictive model. The first data set includes the variables in Table 3.1.

Table 3.1: *Variables in **mortdef.txt** data file.*

Variable	Description
`ID`	Integer number for customer identification.
`Status`	Categorical: Default or NoDefault.
`Delinquency`	Delinquency score.
`PercPastDue`	Past due as a percent of principal plus interest.
`MonthsPastDue`	Number of months past due.
`CurrentLTV`	Current loan-to-value.
`PaymentDiff`	Payment differential.

The second data set provides a credit score from an independent credit reporting organization. The variables in the second data file are shown below.

Table 3.2: *Variables in **mortdef.creditscore.txt** data file.*

Variable	Description
`ID`	Integer number for customer identification.
`CreditScore`	Credit score.

The data is based on real home mortgage data modified for this example. In reality, the percentage of default to no-default loans in the training and testing data would be much lower.

With the goal clearly defined, you can begin to create the Spotfire Miner networks to solve the problem. The first step is to load the data.

ACCESS DATA

The data is in three text files, as described in Table 3.3.

Table 3.3: *Data files available for modeling and predicting mortgage loan defaults.*

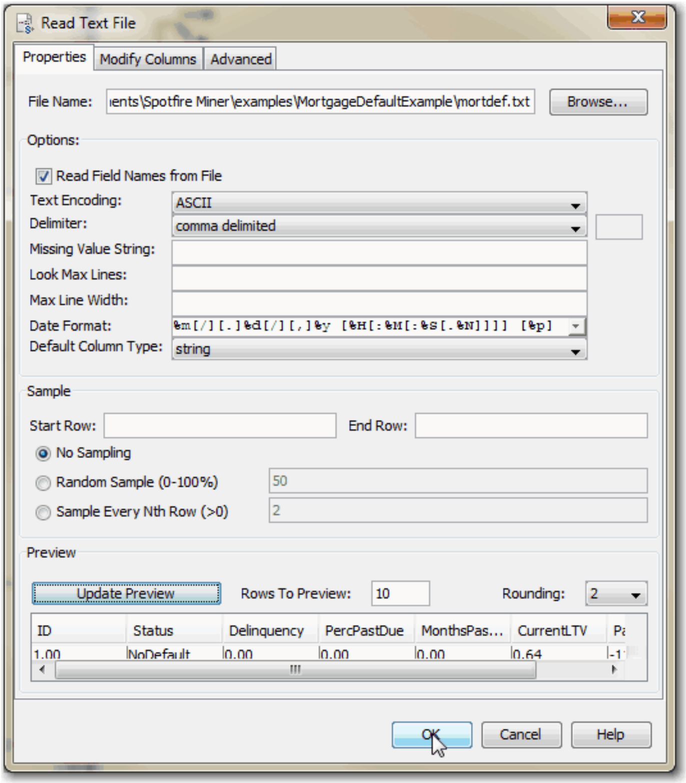
File Name	Description
examples/ MortgageDefaultExample/ mortdef.txt	List of customers, information about their loans, their payment histories, and the status of their loans.
examples/ MortgageDefaultExample/ mortdef.creditscore.txt	A credit score for each customer listed in **mortdef.txt**.
examples/ MortgageDefaultExample/ mortdef.score.txt	Data to predict which customers will default.

For the first phases of this example, you need only the first data file, **mortdef.txt**. If you have not done so, close any worksheets or windows still open from the preceding Quick Tour.

1. If Spotfire Miner is not running, start it.
2. From the main menu, choose **File ▶ New** to open a new Spotfire Miner worksheet.
3. In the explorer pane, click **Read Text File**, drag the mouse pointer to the worksheet in the desktop pane and release.

The **Read Text File** node status indicator is red, showing that it is not ready to be run. Before you can run the node, set its properties.

4. Double-click the **Read Text File** node to open its **Properties** dialog.



File Name: ıents\Spotfire Miner\examples\MortgageDefaultExample\mortdef.txt

☑ Read Field Names from File

Text Encoding: ASCII

Delimiter: comma delimited

Missing Value String:

Look Max Lines:

Max Line Width:

Date Format: %m[/][.]%d[/][,]%y [%H[:%M[:%S[.%N]]]] [%p]

Default Column Type: string

Start Row:

End Row:

● No Sampling

○ Random Sample (0-100%) 50

○ Sample Every Nth Row (>0) 2

Rows To Preview: 10

Rounding: 2

ID	Status	Delinquency	PercPastDue	MonthsPas...	CurrentLTV	Pa
1.00	NoDefault	0.00	0.00	0.00	0.64	-1

Figure 3.3: *The* ***Properties*** *page of the* ***Read Text File*** *dialog.*

5. Click **Browse**, and then click the **Examples** folder (in the left lower corner of the browser). You are prompted to copy the contents of the **examples** folder from the installation directory to an **examples** folder under your default user directory (as defined by your operating system[1]) and preserve the original **examples** folder, should you need to access it.

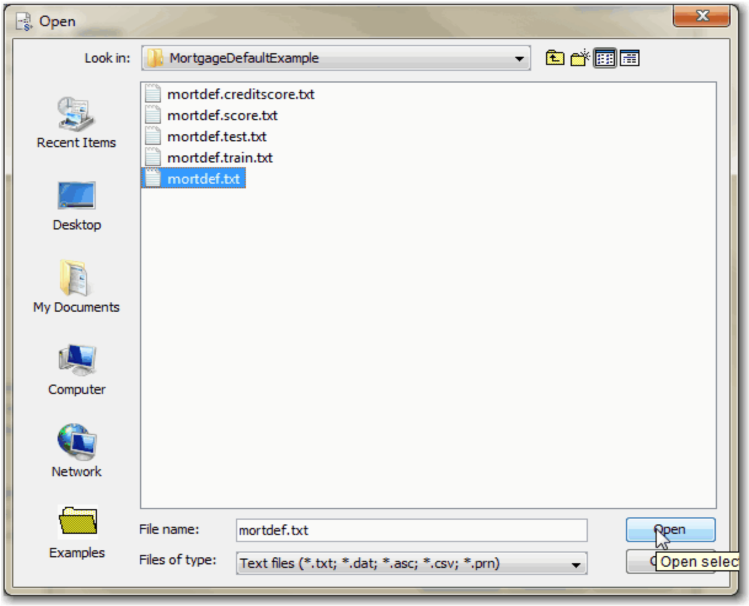
[1] The location of the default user directory depends on which version of Microsoft Windows® you are running. by default:

on Windows XP®: **C:\Documents and Settings\*username*\My Documents\Spotfire Miner\examples**

on Windows Vista®and Windows 7®: **C:\users\*username*\Documents\Spotfire Miner\examples**

Click **OK** to accept this option.

6. Double-click the **MortgageDefaultExample** folder, and then from this folder, select the data file **mortdef.txt**.
7. Click **Open**.



Look in: MortgageDefaultExample

mortdef.creditscore.txt
mortdef.score.txt
mortdef.test.txt
mortdef.train.txt
mortdef.txt

File name: mortdef.txt

Files of type: Text files (*.txt; *.dat; *.asc; *.csv; *.prn)

Figure 3.4: *Opening the **Examples** folder through the **Browser** dialog.*

Hint

As an alternative to browsing for a file, you can type the file name in the **File Name** text box. If you do not specify the full path to the file, Spotfire Miner looks for the file in the same folder as the worksheet.

8. For a preview of the first ten rows of data in the data file, click **Update Preview** in the **Preview** group. (The completed dialog page for this section is shown in Figure 3.3.)

By default, columns with numeric values are read in as *continuous* columns, and columns with nonnumeric characters are read as *string* columns. String columns are best used for storing identifying information that is typically different for each row and is not used in modeling.

To learn more about the columns, examine the values in the preview for more information about the kind of values each column contains. Alternatively, read the columns as string columns, and then examine them to determine the appropriate type.

In this example, read the `ID` column as string and the `Status` column as categorical. All the other columns in the data set contain numeric values, so read them as continuous. Use the **Modify Columns** page of the **Read Text File** dialog to change the column types for these variables.

9. Click the **Modify Columns** tab. (The completed dialog page for this section is shown in Figure 3.5.

Hint

If the columns are not wide enough to display the column names, you can expand them by positioning the mouse cursor over the vertical line between two columns, and then, when the pointer becomes a two-headed arrow, click and drag the mouse to the left or right until the column is the desired width. Release the mouse button.

Depending upon how much you widen the first column, you might have to scroll to the right in the grid view to see the **New Types** column.

10. Click anywhere in the row containing the variable name **Status** to select it.
11. In the **Set Types** group at the bottom right of the dialog, click **Categorical**.

Notice that a visual cue now appears in the **New Types** column reflecting the change in the data type of `Status` from string (Ⓢ) to categorical (Ⓐ).

This tab is also a convenient place to set the dependent variable.

12. In the **Set Roles** group, click **Dependent**.

Notice that a (Ⓓ) appears in the **New Roles** column to show **Status** is now set as a dependent variable. This information is carried along through the network.

13. Click anywhere in the `ID` row and select **String** from the **Set Types** group.
14. Click **OK** to close the dialog.

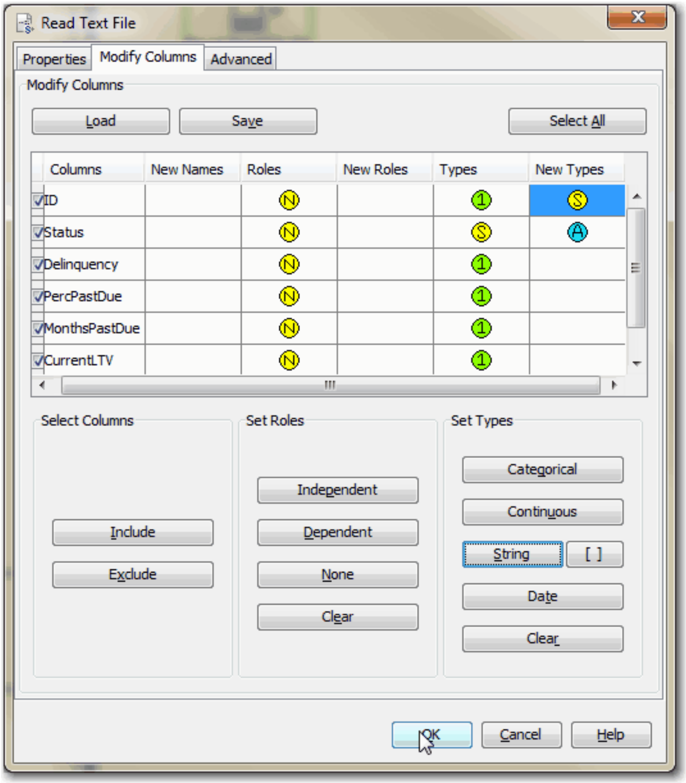
Note that the **Read Text File** node status indicator is yellow, showing that it is ready to run. But first, read in a second data file.

15. Right-click a blank space in the worksheet and select **Create New Node**.

A scrabble view of the explorer pane appears, and you can select a node to add to the current worksheet.

16. Select a **Read Text File** node and click **OK**.
17. Click the node and drag it to move it below the first node.
18. Select the **Properties** tool ()from the tool bar.
19. Click **Browse** and select **mortdef.creditscore.txt**.
20. Click **Open**.
21. Click the **Modify Columns** tab.
22. Click anywhere in the `ID` row and select **String** from the **Set Types** group.

23. Click **OK** to close the dialog. Note that the **Read Text File** node status indicator is yellow, showing that it is ready to run.)



Columns	New Names	Roles	New Roles	Types	New Types
☑ID		N		1	S
☑Status		N		S	A
☑Delinquency		N		1	
☑PercPastDue		N		1	
☑MonthsPastDue		N		1	
☑CurrentLTV		N		1	

Figure 3.5: *The **Modify Columns** page of the **Read Text File** dialog.*

24. Click the **Run** button ▶ on the Spotfire Miner toolbar.

As both nodes are executing, watch the message pane (below the worksheet) for information about execution time and cache size, as well as any errors or warning messages. After the nodes successfully complete, the status indicators change to green. Figure 3.6 shows the worksheet after the first two nodes have run.

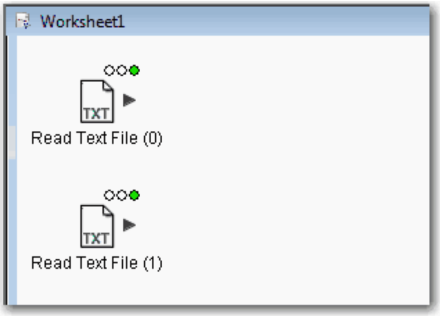


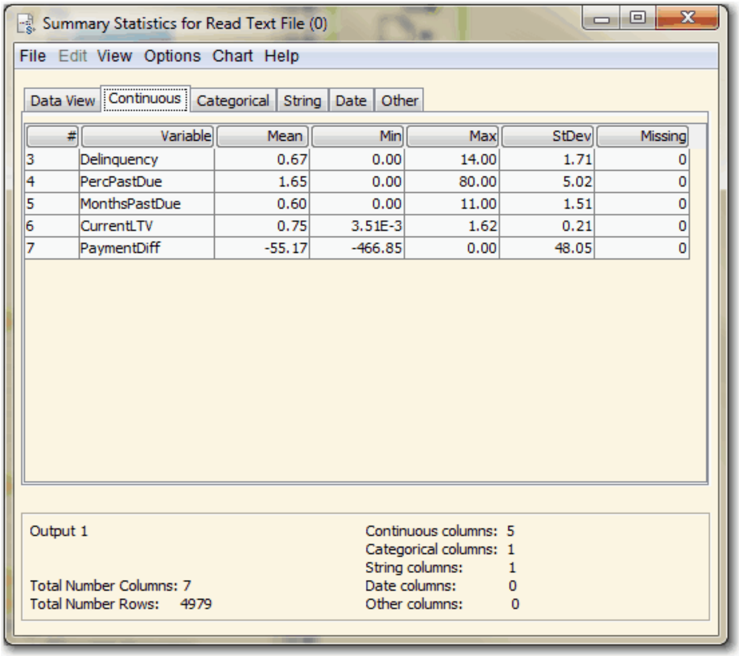
Figure 3.6: *Worksheet after running first two nodes.*

EXPLORE DATA

With the data read in, you can examine it in greater detail and prepare it for the model building phase of the problem.

Launch the viewer for the first **Read Text File** node.

1. Click **Read Text File (0)** to select it, and then click the Viewer button on the Spotfire Miner toolbar. (Figure 3.7 shows the open viewer, with the **Continuous** page displayed.)


Summary Statistics for Read Text File (0)
File Edit View Options Chart Help
Data View
Continuous
Categorical
String
Date
Other

#	Variable	Mean	Min	Max	StDev	Missing
3	Delinquency	0.67	0.00	14.00	1.71	0
4	PercPastDue	1.65	0.00	80.00	5.02	0
5	MonthsPastDue	0.60	0.00	11.00	1.51	0
6	CurrentLTV	0.75	3.51E-3	1.62	0.21	0
7	PaymentDiff	-55.17	-466.85	0.00	48.05	0

Output 1

Total Number Columns: 7
Total Number Rows: 4979

Continuous columns: 5
Categorical columns: 1
String columns: 1
Date columns: 0
Other columns: 0

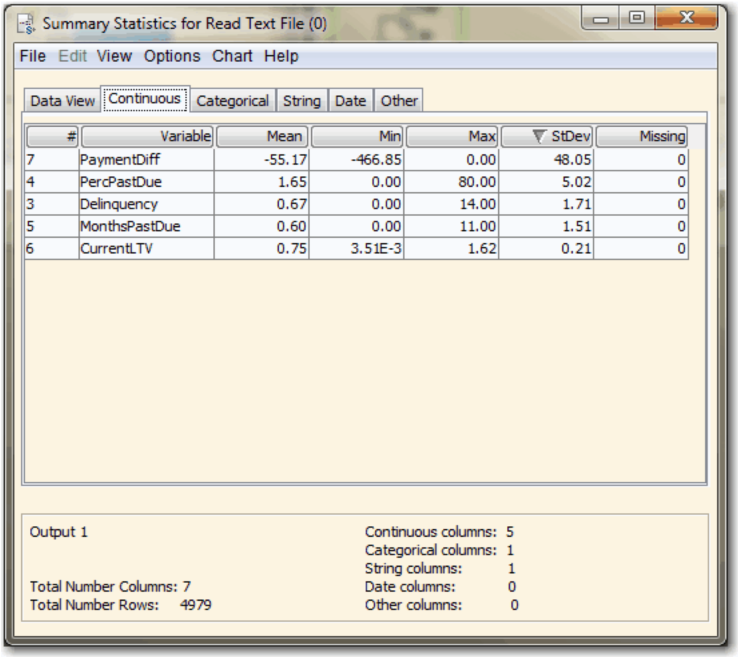
Figure 3.7: *The **Continuous** page of the node viewer.*

At the bottom of the node viewer, notice that the data file has 7 columns and 4979 rows. The number of variables (columns) of each type is shown in the bottom right corner. Each tab of this viewer summarizes a different type of data. The first tab shows the full data set.

The **Continuous** page of the node viewer shows the minimum, maximum, mean, and standard deviations for each continuous variable in the data. The number of missing values is shown in the last column (see Figure 3.7).

You can sort rows in the variable summary pages based on any single column.

2. To sort based on the **StDev** column in descending order, click its column header, as shown in Figure 3.8. (Clicking once more sorts the rows in ascending order of **StDev**.)


Summary Statistics for Read Text File (0)
File Edit View Options Chart Help
Data View
Continuous
Categorical
String
Date
Other

#	Variable	Mean	Min	Max	▼ StDev	Missing
7	PaymentDiff	-55.17	-466.85	0.00	48.05	0
4	PercPastDue	1.65	0.00	80.00	5.02	0
3	Delinquency	0.67	0.00	14.00	1.71	0
5	MonthsPastDue	0.60	0.00	11.00	1.51	0
6	CurrentLTV	0.75	3.51E-3	1.62	0.21	0

Output 1

Total Number Columns: 7
Total Number Rows: 4979

Continuous columns: 5
Categorical columns: 1
String columns: 1
Date columns: 0
Other columns: 0

Figure 3.8: *Sorting a column in the node viewer by clicking the column header.*

Notice that a point-down triangle appears next to the variable name for which the data is currently sorted. (The triangle points up when the data is sorted in ascending order.)

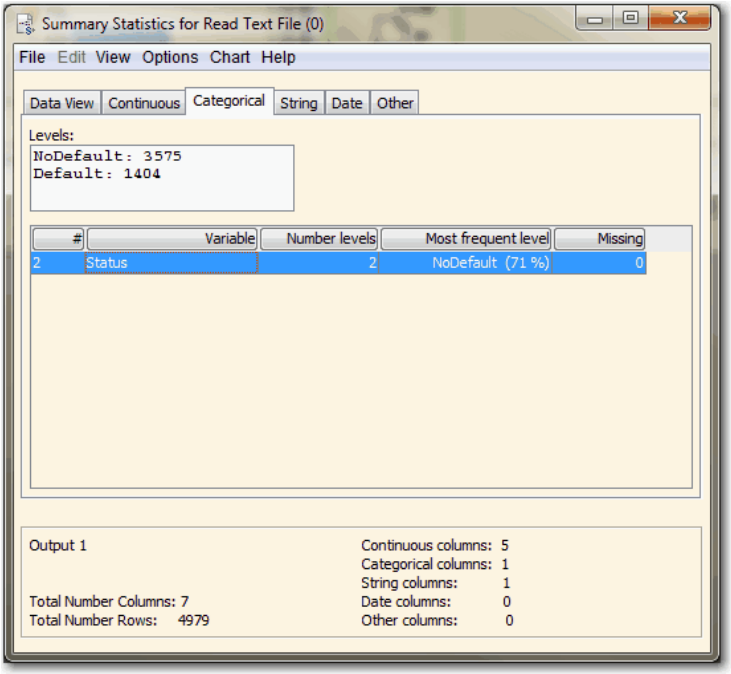
3. Click through the menus at the top to examine the variety of ways you can manipulate the viewer. For example:

   - From the **View** menu, you can view an HTML report of the data.

- From the **Edit** menu, you can copy the data to a clipboard.
- From the **Options** menu, you can change the number of displayed digits.
- From the **Chart** menu, you can selection options for plotting the data.

See the section Using Spotfire S+ Graphs on page 95 or the *Spotfire Miner User's Guide* for more information on creating Spotfire S+ plots and adding them to your worksheet as new nodes.

4. Click the **Categorical** tab of the node viewer. You have only one categorical variable, `Status`, as shown in Figure 3.9.


Summary Statistics for Read Text File (0)
File Edit View Options Chart Help
Data View
Continuous
Categorical
String
Date
Other

Levels:

```
NoDefault: 3575
Default: 1404
```

#	Variable	Number levels	Most frequent level	Missing
2	Status	2	NoDefault (71 %)	0

Output 1

Total Number Columns: 7
Total Number Rows: 4979

Continuous columns:	5
Categorical columns:	1
String columns:	1
Date columns:	0
Other columns:	0

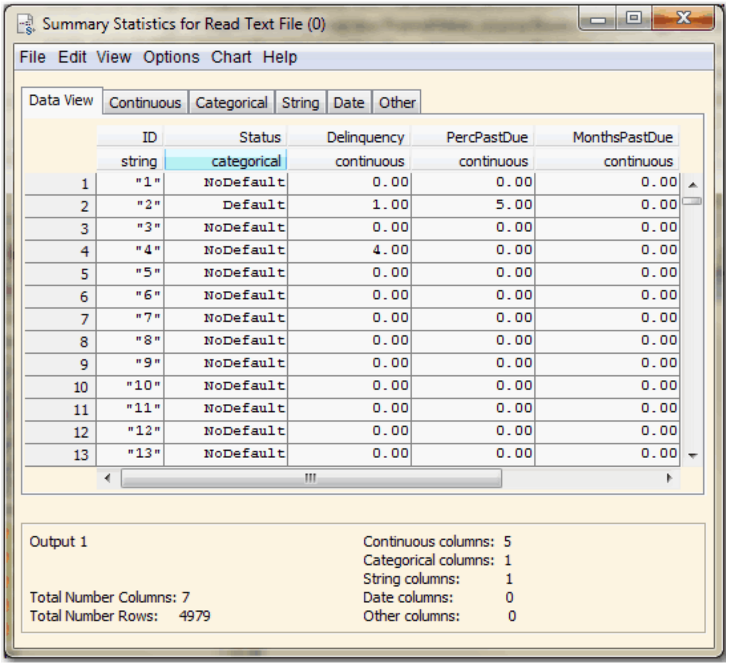
Figure 3.9: *The **Categorical** page of the node viewer.*

Summaries of categorical variables consist of the number of levels for each variable, the most frequent level observed, and the number of missing data.

5. Click anywhere in the `Status` variable row. The counts for each level of the variable appear in the box at the top right of the page.

Note for later that there are 3575 observations of `NoDefault` and 1404 observations of `Default`.

6. Click the **Data View** tab to display the entire data set, as shown in Figure 3.10. You can scroll through the data by using the scroll bar at the bottom and right of the data grid.


Summary Statistics for Read Text File (0)
File Edit View Options Chart Help
Data View
Continuous
Categorical
String
Date
Other

	ID	Status	Delinquency	PercPastDue	MonthsPastDue
	string	categorical	continuous	continuous	continuous
1	"1"	NoDefault	0.00	0.00	0.00
2	"2"	Default	1.00	5.00	0.00
3	"3"	NoDefault	0.00	0.00	0.00
4	"4"	NoDefault	4.00	0.00	0.00
5	"5"	NoDefault	0.00	0.00	0.00
6	"6"	NoDefault	0.00	0.00	0.00
7	"7"	NoDefault	0.00	0.00	0.00
8	"8"	NoDefault	0.00	0.00	0.00
9	"9"	NoDefault	0.00	0.00	0.00
10	"10"	NoDefault	0.00	0.00	0.00
11	"11"	NoDefault	0.00	0.00	0.00
12	"12"	NoDefault	0.00	0.00	0.00
13	"13"	NoDefault	0.00	0.00	0.00

Output 1

Total Number Columns: 7
Total Number Rows: 4979

Continuous columns: 5
Categorical columns: 1
String columns: 1
Date columns: 0
Other columns: 0

Figure 3.10: *An example of the* ***Data View*** *page of the node viewer.*

Preparing the Data

A customer's credit score can be significant in predicting whether that customer will default on a loan, so add this information to the other data by merging the two data sets.

To find the best model for the data, try comparing several models. This process is typically done in the following three stages:

- Train the model(s)
- Test the model(s)

- Validate the model(s).

After merging the data files, partition the new data set to create training and testing data sets. By partitioning the data and using different data to build and test the models, you get a better estimate of the modeling errors. If you want to have an unbiased estimate of the errors from final chosen model, you would create three data sets: one for training, one for testing, and one for validation. For expediency, partition the data into two data sets: one for training the model and one for comparing (testing) the models.

Note that the **Partition** node has three *output ports*, designated by black triangles, on the right side of the node. You can use these output ports to output the resulting partitioned data sets for different operations, such as writing files. (See Figure 3.11 for an illustration.)

After you create the two data sets, the completed network resembles Figure 3.11. A copy of the finished worksheet is provided in the file **examples/MortgageDefaultExample/ MortgageDefault.Explore.imw**. Continue this example by building onto the network in the current worksheet.

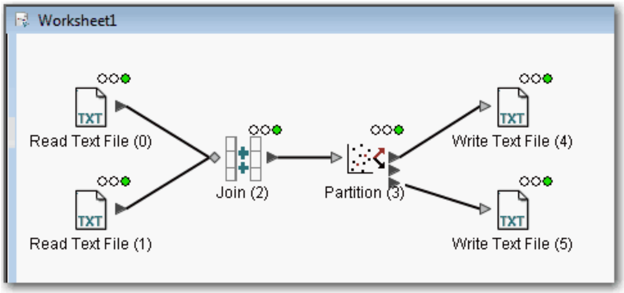


Figure 3.11: *Network for reading in data, stratifying it and partitioning it into two data sets which are then written to a text file.*

Merging the Data Sets

In this exercise, first join the data from two files, and then partition the data.

To create one data set from the two files, use a **Join** node. Both files have an `ID` column, so you can easily match the data rows. More complicated joining is possible; refer to information on the **Join** node in Chapter 6, Data Manipulation, of the *Spotfire Miner User's Guide.*

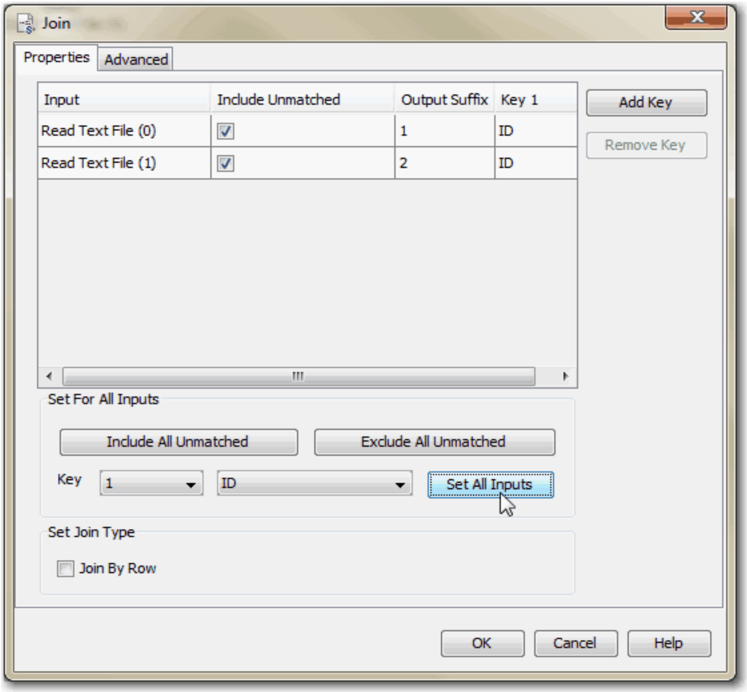
7. Double click the **Join** node under the **Data Manipulation/ Columns** folder. A new **Join** node appears in the worksheet. Move this node to the right of the **Read Text File** nodes.

8. Left-click and hold the mouse button over the output port of the **Read Text File (0)** node. Drag the mouse until it is over the top input port of the **Join** node and release the mouse button. A link appears to connect the two nodes.

9. Repeat the previous step, connecting the **Read Text File (1)** node to the lower input port of the **Join** node.

Now that the input to the **Join** node is specified, set the node properties. In these data files, there is a one-to-one correspondence in the customer ID, so you do not need to worry about unmatched rows. The completed properties page is shown in Figure 3.12.Right-click the **Join** node icon and select **Properties** from the menu.

10. In the **Set for All Inputs** group, the **Key** number drop-down list box should show 1. In the **Key** value drop-down list box, select `ID`. Click **Set All Inputs**.



Input	Include Unmatched	Output Suffix	Key 1
Read Text File (0)	☑	1	ID
Read Text File (1)	☑	2	ID

Set For All Inputs

Key 1 ID

Set Join Type

□ Join By Row

Figure 3.12: *Completed properties page for the **Join** node.*

11. Click **OK** to close the dialog.

12. Click the Run to Here button (▶|) on the toolbar.

To be sure the data is merged correctly, open the viewer and check the column names and types.

13. Click the Viewer button (👁)on the Spotfire Miner toolbar. Open the **Data View** page and use the horizontal scroll to check that the last column is now `CreditScore`.

You are ready to create the train and test data sets.

Partitioning the Data

14. Under the **Data Manipulation/Rows** folder in the explorer pane, double-click the **Partition** node. Position the new **Partition** node to the right of the **Join** node.

15. Left-click the output port of the **Join** node and drag a link to the **Partition** node.

16. Right-click the **Partition** node and select **Properties**.

17. In the **Train** box, type `70`. In the **Test** box, type `30`.

For this example to be repeatable, set a seed for the random sampling in the **Partition** node.

18. Click the **Advanced** tab, and then click **Enter Seed**. Use the default value of 5. Click **OK**.

19. From the **Toolbar**, click **Run** (▶).

Notice in the message pane that only the **Partition** node is executed. Nodes that have a green status do not rerun.

The top output port of the partition node passes 70% of the randomly-sampled data. The remaining 30% is output from the lower port.

Next, write the two data sets to text files to use later.

Writing Data to Text Files

20. Scroll to the bottom of the **Explorer** pane to the **Data Output/File** folder. Find the **Write Text File** component under this folder.

21. Add two **Write Text File** nodes to your worksheet, positioning them beside the **Partition** node as shown in Figure 3.11.

22. Link the **Partition** node to the **Write Text File** nodes.

Hint

To delete a link between nodes, right-click the link and select **Delete Link**.

You can change the shape of the links from straight lines to orthogonal lines by right-clicking the link and clearing **Diagonal Link**. Alternatively, you can change all links to orthogonal lines by clicking **Edit** ▶ **Select All**, and then clicking **View** ▶ **Toggle Diagonal Links**.

23. Double-click the upper **Write Text File** node to open its **Properties** page.
24. Click **Browse**, and then click the **Examples** icon.
25. Open the **MortgageDefaultExample** folder, and then, in the **File name** box, type **Mymortdef.train.txt**. Click **Open**. This creates a new file, if it does not exist, or overwrites an existing file. (Use a different file name if you do not want to create a new file or overwrite an existing file.)
26. Change the **Delimiter** selection to `single space delimited` and click **OK**.
27. Double-click the lower **Write Text File** node to open its **Properties** page.
28. Click **Browse**, and then open the **MortgageDefaultExample** folder. In the **File name** box, type **Mymortdef.test.txt**. Click **Open**.
29. Click **OK**.
30. Click **Run** (▶)on the **Toolbar**.

When you provide only a file name, the default path is your Spotfire Miner working directory, as specified by your operating system. The training and testing data sets have been written in that directory.

The completed worksheet in the examples directory (**examples/MortgageDefaultExample/MortgageDefault.Explore.imw**) shows that you can combine or collect the **Read Text File (1), Join**, and **Partition** nodes to create one *collection* node that joins the data files and partitions the data into the two data sets. You can add this node to your User library for future use. Refer to the *User's Guide* for how to create and use collection nodes and add them to the User library.

Saving a Worksheet

To save this worksheet:

31. From the main menu, select **File ▶ Save As** and browse to a location to save the file.
32. Type a file name into the **File name** box and click **Save**. The file name you type is appended with the extension **.imw** automatically.

Next, model the data.

CREATE A MODEL

Modeling in the present context means *predictive* modeling. When you use the training data set, which contains a known target variable, you can apply supervised learning to generate a predictive model. Then you can use this model to make predictions, called *scores*, about the target variable. In this example, you are predicting the probability that a customer defaults on a loan.

First, train the model to the data, and then predict using various models and compare their predictions to the observed data.

You could continue the example by adding to the previous worksheet and network, as shown in Figure 3.2 (**examples/ MortgageDefaultExample/MortgageDefault.complete.imw**). Instead, begin a new worksheet that reads the train and test data files that you exported in the previous section Explore Data. The final network is shown in Figure 3.13 and a copy of the finished worksheet can be found at **examples/MortgageDefaultExample/ MortgageDefault.model.imw**.

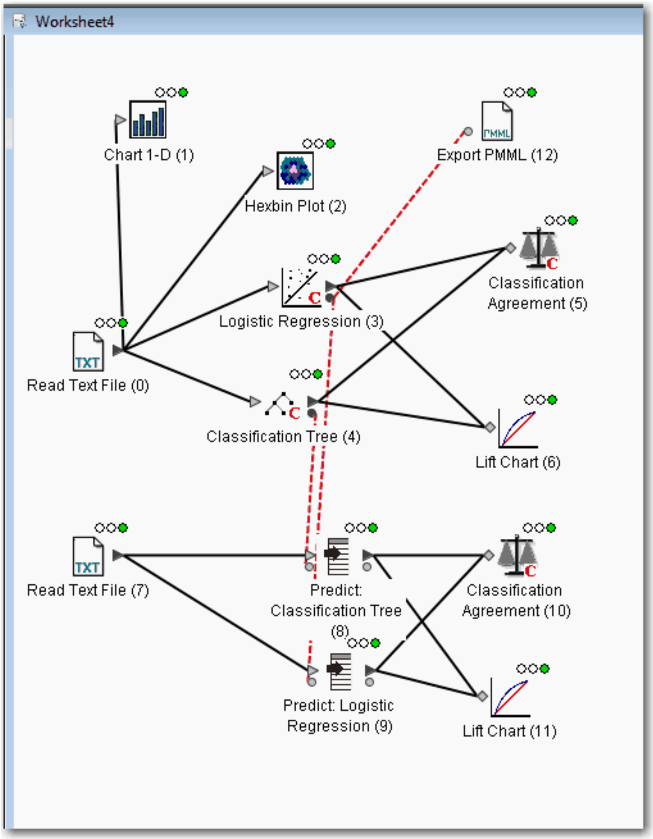


Figure 3.13: *Completed worksheet for the modeling phase, **MortgageDefault.model.imw**.*

Inputting The Training Data

To begin, open a new worksheet and read in the training data that was created by partitioning the merged data set in the section Explore Data.

1. On the main menu, click **File** ► **New** to create the new worksheet.
2. From the **Data Input/File** folder of the explorer pane, double-click a **Read Text File** component to add its node to the worksheet. Drag the node about an inch down the left side of the worksheet using the mouse.

3. Double click the **Read Text File** node to open the **Properties** page.
4. Click **Browse**, and then open the folder **examples/MortgageDefaultExample**. Select the **mortdef.train.txt** file, and then click **Open**.

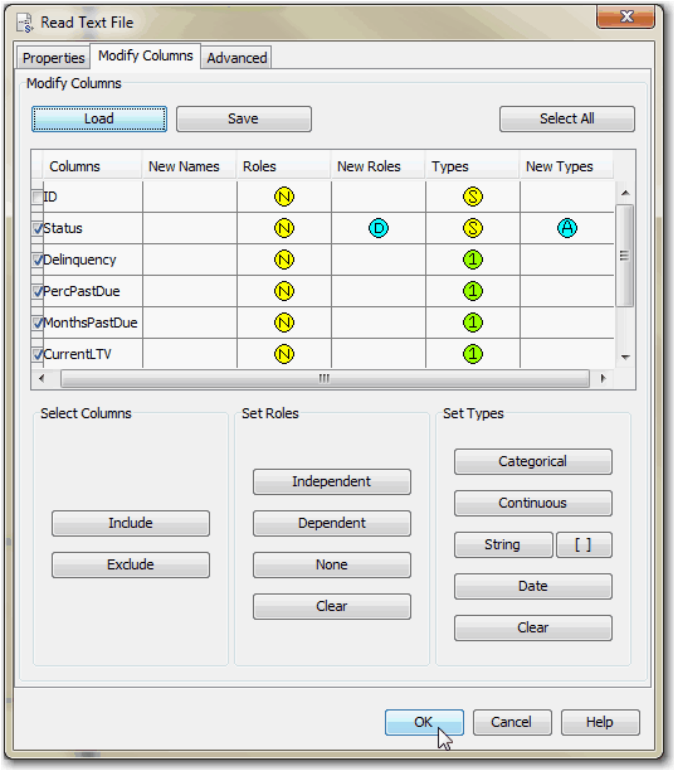
Set the properties.

5. To see of the first ten rows of data in the data file, click **Update Preview**.
6. Click the **Modify Columns** tab.
7. Select the variable **Status** by clicking anywhere in its row.
8. In the **Set Roles** group, click **Dependent**. In the **Set Types** group, click **Categorical**. (Specifying **Set Roles** as **Dependent** is for convenience in completing a later step in this exercise.)

Exclude `ID` from the modeling process.

9. Click the column name `ID`, and then, in the **Select Columns** group, click **Exclude**.
10. Click **OK** to close the dialog.

Now, the `ID` column is not read in from the data files. For greater detail on importing data files, see the section Access Data on page 57 or the *TIBCO Spotfire Miner User's Guide*. The completed dialog page for this section is shown in Figure 3.14. Click **OK** to accept the changes.



Columns	New Names	Roles	New Roles	Types	New Types
☐ ID		N		S	
☑ Status		N	D	S	A
☑ Delinquency		N		1	
☑ PercPastDue		N		1	
☑ MonthsPastDue		N		1	
☑ CurrentLTV		N		1	

Figure 3.14: *The completed* ***Modify Columns*** *page for reading in the training data.*

Plotting the Data

If you were using this data set to review the loans in your portfolio to identify those at risk of default, you can use a predictive model to examine patterns in the data. This model uses the information about loans (percent and months past due, current loan-to-value and payment differential, and credit and delinquency scores) to predict the probability of default.

To get an initial overview, first use a **Chart 1-D** node to examine histograms of the data and determine which columns can give you the information to determine default and no-default likelihood.

Setting Hexbin Plot Node Properties

11. Drag a **Chart 1-D** node to the worksheet and link it to your **Read Text File** node.

12. Double-click the **Chart 1-D** node to open its **Properties** dialog.

13. In the **Properties** page, in the **Available Columns** list, highlight `Status` and add it to the **Group By** list box using its double-right arrow (>>).

14. Select the remaining items in the **Available Columns** list, and then add them to the **Display** list box using its double-right arrow (>>).

15. Click **OK** to close the dialog and then right-click and select **Run to Here** to run the network. Right-click and select **Viewer** to display the histograms in Figure 3.15.

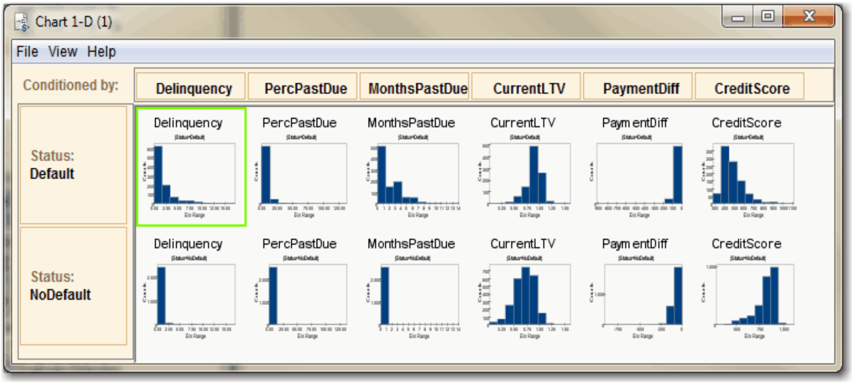


Figure 3.15: *Histogram showing credit scores and current loan-to-value variables.*

These histograms indicate that credit scores and current loan-to-value variables show some interesting variations with the `Status` value, particularly `CreditScore` and `CurrentLTV`.

Next, use a hexbin plot to investigate these two columns further. Because hexbin plot bins the data, rather than plotting individual points for each row of data, you can use it with large data sets and still display readable charts. Also, you can specify handling **All Rows** with a hexbin plot, implementing the Big Data Trellis feature.

In this example, set the hexbin plot x and y axes to `CreditScore` and `CurrentLTV`, respectively. Use the `Status` variable to condition the data.

Setting Hexbin Plot Node Properties

16. In the explorer pane, click the **Spotfire S+** tab.
17. In the **Two Columns - Continuous** folder, double-click **Hexbin Plot** to add a **Hexbin Plot** node to the worksheet.
18. Position the **Hexbin Plot** node above and to the right of the **Read Text File** node, and then link the nodes.
19. Double-click the **Hexbin Plot** node to display its properties dialog.
20. In the **Data** page, set the **x Axis Value** to `CurrentLTV`. Set the **y Axis Value** to `CreditScore`.
21. In the **Conditioning** box, select `Status`.
22. In the **Row Handling** group, select **All Rows**. (Selecting **All Rows** uses the Big Data library Trellis function.)
23. You can examine the options in the other tabs of the dialog, but accept the default options. For more information about using Hexbin Plot options and other Spotfire S+ charts, see the Spotfire S+ Library chapter in the *Spotfire Miner User's Guide*.

24. On the **Data** page, click **Apply** to display the **Hexbin Plot** Trellis graph. The graph, using standard colors, appears in Figure 3.16.

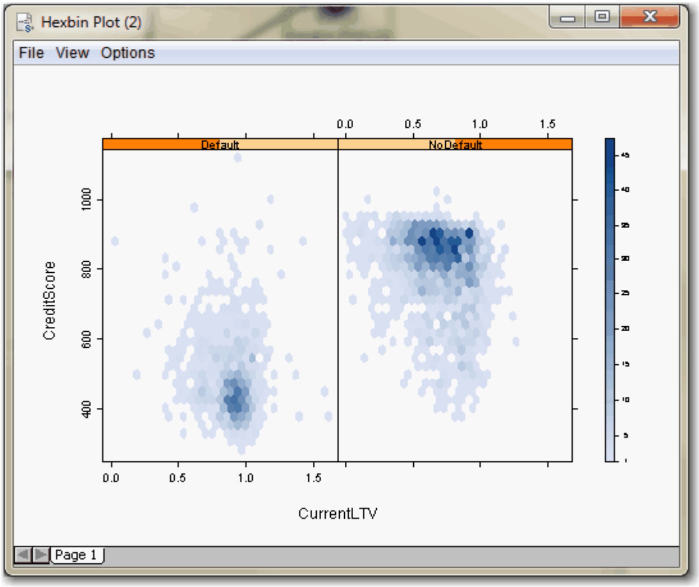


Figure 3.16: *Mortgage default example hexbin plot.*

Note that in the data example, customers who have lower credit scores and relatively higher current loan-to-value ratings tend to be at higher risk for defaulting on their loans.

Changing the Chart Color Display

If the display does not show the color scheme you want, you can change the hexbin plot colors. To change the colors from **Default** to **Standard**, as shown in Figure 3.16, do the following:

25. In the chart window, on the menu, click **Options ▶ Set Graph Colors**.

26. Click **Standard**.

Optionally, you can edit the colors by clicking **Edit Colors**, and in the **Edit Graph Colors** dialog, selecting a new scheme, changing individual colors, or blending a range of colors. See the *Spotfire Miner User's Guide* for more information.

Training the Models

The dependent variable is a categorical variable, which creates a need for a classification test. In this example, you use logistic regression and classification trees. Both models are appropriate for a data set with a categorical predictor.

27. In the Explorer pane, click the **Main** tab. Expand the **Model/ Classification** folder, and then double-click a **Logistic Regression** component to add its node to the worksheet. Link it to the **Read Text File** node.

28. Place the model node below the **Logistic Regression** node as shown in Figure 3.13.

Setting Model Node Properties

29. Double-click the **Logistic Regression** node to open its properties page.

Notice that the `Status` variable is marked with Ⓐ and Ⓓ because when you read in the data, you specified that the `Status` variable was the dependent variable.

30. Click `Auto` to move the `Status` variable into the **Dependent Column** box.

Because you did not set the independent variables when you read in the data, you must move them manually.

31. Click to select `Delinquency`, and then hold the SHIFT key and click on the last variable, `CreditScore`. The whole column should now be highlighted. Click the double right arrow button >> to move these variables into the **Independent Columns** box.

Hint

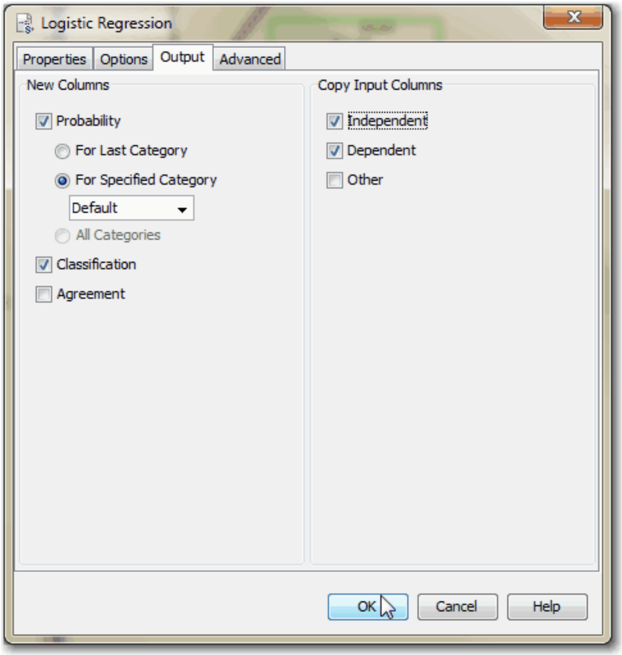
You can add interaction terms to the **Independent Columns** by selecting the interacting variables in the **Independent Columns** box, and then clicking **Interactions**. Selecting more than two variables and clicking **Interactions** adds all combinations of interactions. You can remove an interaction by selecting its term, and then clicking the double left arrow button.

Check to make sure that the other properties of the model are set correctly. By default, Spotfire Miner returns the computed probabilities for the last level in the dependent variable. Change the default settings to get more meaningful results for your prediction.

32. Click the **Output** tab. In the **New Columns** group, select **For Specified Category**. For this exercise, in the drop-down box, accept **Default**.

33. To create plots later that include independent variables, in the **Copy Input Columns** group, select **Independent**.

34. The completed dialog box is displayed in Figure 3.17. Click **OK** to set the properties for this node.



New Columns

☑ Probability

○ For Last Category

● For Specified Category

Default

○ All Categories

☑ Classification

☐ Agreement

Copy Input Columns

☑ Independent

☑ Dependent

☐ Other

Figure 3.17: *The completed* ***Logistic Regression Output*** *dialog.*

Add a **Classification Tree** node and set its properties to the same values as the **Logistic Regression** node.

35. In the Explorer pane, click the **Main** tab. Expand the **Model/ Classification** folder, and then double-click a **Classification Tree** component to add its node to the worksheet. Link it to the **Read Text File** node.

36. Double-click the **Classification Tree** node to open its properties dialog.

Click each tab to see the default settings. Most of these settings are fine for this example; however, you must designate the independent variables and save them to the output for later use.

37. Repeat Steps 30.-34. above for this node, and then click **OK** to set the properties for this node.

Running the Model Nodes

38. Click the **Run** button ▶ on the toolbar to run the network.

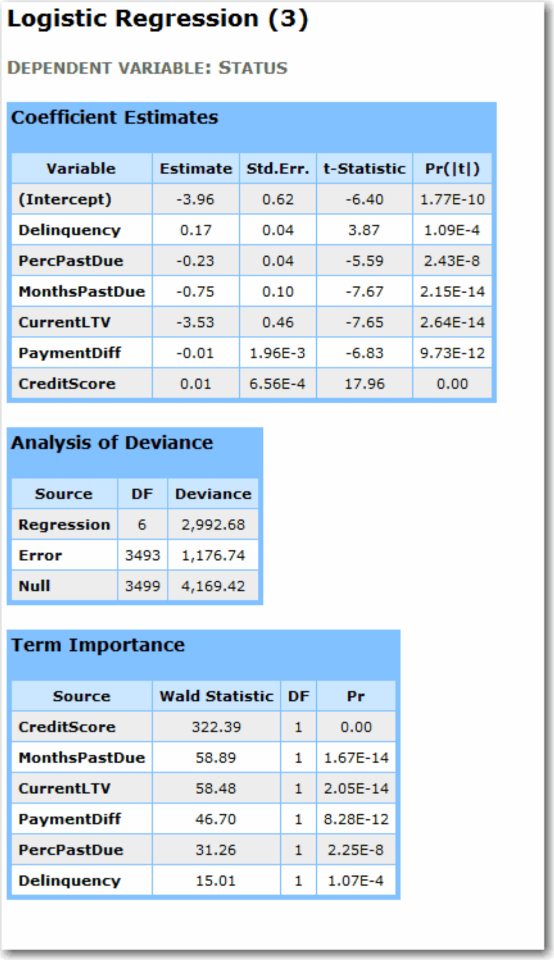
Viewing the Models

You now have two models for the data. Examine each to understand the differences between them.

39. Right click the **Logistic Regression** node and click **Viewer**.

The viewer for this node is an HTML report, shown in Figure 3.18.

> **Note**
>
> Spotfire Miner opens **.html** files with the application associated with **.html** files (for example, Internet Explorer® or Mozilla Firefox).



Logistic Regression (3)

DEPENDENT VARIABLE: STATUS

Coefficient Estimates

Variable	Estimate	Std.Err.	t-Statistic	Pr(|t|)
(Intercept)	-3.96	0.62	-6.40	1.77E-10
Delinquency	0.17	0.04	3.87	1.09E-4
PercPastDue	-0.23	0.04	-5.59	2.43E-8
MonthsPastDue	-0.75	0.10	-7.67	2.15E-14
CurrentLTV	-3.53	0.46	-7.65	2.64E-14
PaymentDiff	-0.01	1.96E-3	-6.83	9.73E-12
CreditScore	0.01	6.56E-4	17.96	0.00

Analysis of Deviance

Source	DF	Deviance
Regression	6	2,992.68
Error	3493	1,176.74
Null	3499	4,169.42

Term Importance

Source	Wald Statistic	DF	Pr
CreditScore	322.39	1	0.00
MonthsPastDue	58.89	1	1.67E-14
CurrentLTV	58.48	1	2.05E-14
PaymentDiff	46.70	1	8.28E-12
PercPastDue	31.26	1	2.25E-8
Delinquency	15.01	1	1.07E-4

Figure 3.18: *Viewer for Logistic Regression node.*

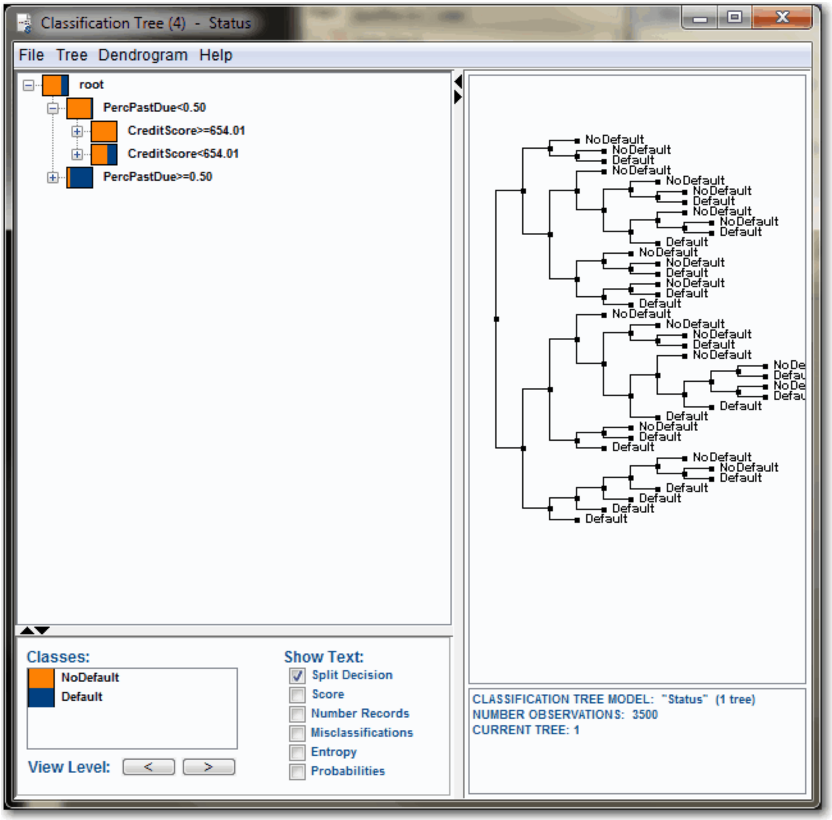
The **Coefficient Estimates** and **Term Importance** tables indicate that all independent variables are significant in this model. The top three variables for this node are `CreditScore`, `MonthsPastDue`, and `CurrentLTV`. Note that, while `CreditScore` and `PaymentDiff` are significant, their coefficients are very small. You could investigate adding interaction terms between predictors to see if you can improve the model; however, that exercise is not part of this example.

Spotfire Miner's **Classification Tree** node stores information about all of the variables in the tree, including the relative importance of each split. Open the viewer for the Classification Tree node (shown in Figure 3.19) and examine the **Relative Term Importance** chart.

40. Right-click the **Classification Tree** node and click **Viewer**.

41. In the upper left window, to expand the node, click the plus sign next to `PercPastDue < 0.50`.

Notice that you can also expand the hierarchal view by clicking the dendrogram in the right window pane .



⊟ root
⊟ PercPastDue<0.50
⊞ CreditScore>=654.01
⊞ CreditScore<654.01
⊞ PercPastDue>=0.50

Classes:
NoDefault
Default

Show Text:
☑ Split Decision
☐ Score
☐ Number Records
☐ Misclassifications
☐ Entropy
☐ Probabilities

CLASSIFICATION TREE MODEL: "Status" (1 tree)
NUMBER OBSERVATIONS: 3500
CURRENT TREE: 1

Figure 3.19: *The viewer for the* ***Classification Tree*** *node with the first node expanded in the hierarchical view pane.*

42. Click **Tree ► View Column Importance** to show a boxplot of the relative column importance.

The bar chart in Figure 3.20 shows the relative change in deviance for each column in the model. At each split, you know the column (variable) split, and the change in deviance due to the split. You get the change in deviance for the column by adding the changes in deviance for all splits in which it was used. Those columns with large changes in deviance are very important in the model and appear at the top of the chart. The most predictive variables are those with values greater than zero in their respective column importance plot. You could use this information to filter out the columns in the data set where the deviance was very close to zero.

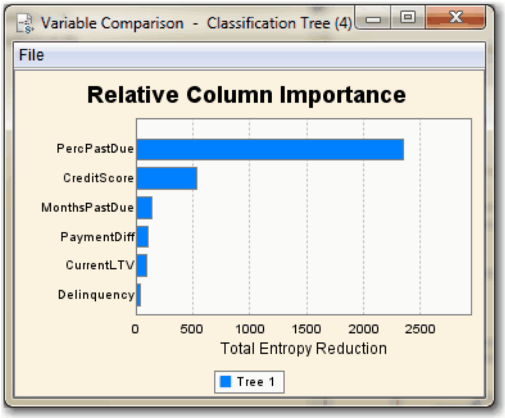


Figure 3.20: *Variable Comparison chart for the* ***Classification Tree*** *model.*

43. Close the viewers.

Selecting a Model

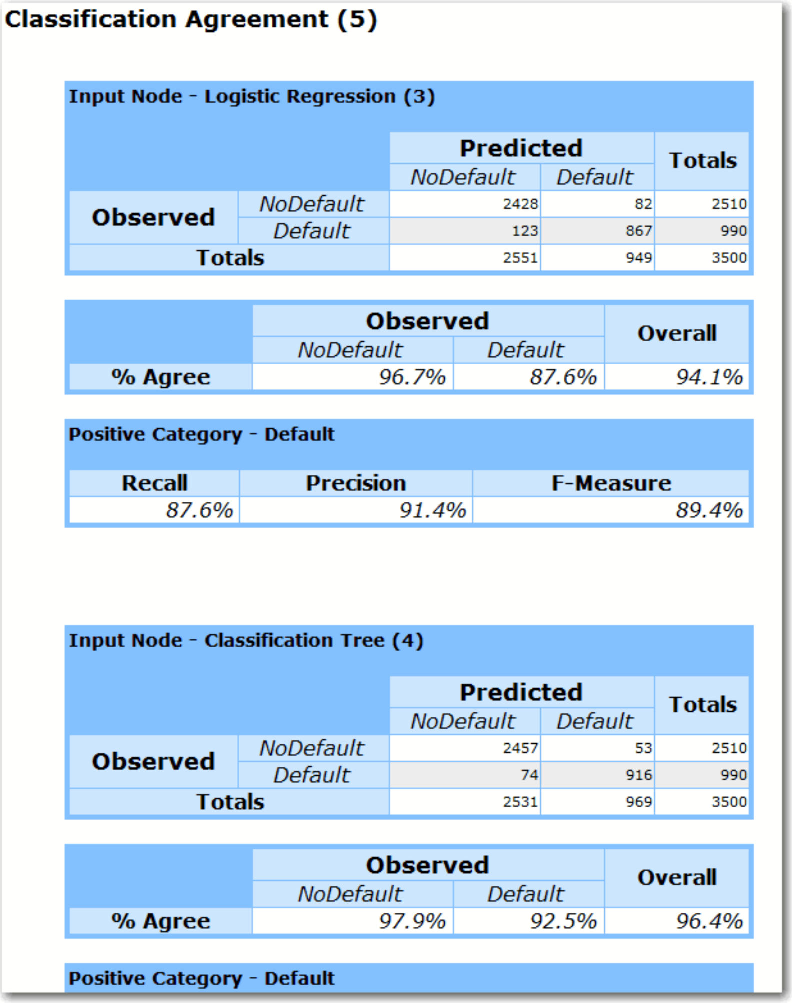
You want to learn how well the models predict an individual defaulting on a loan. This information is formatted in several confusion matrices (one for each model) by the **Classification Agreement** component. In addition, the **Lift Chart** component provides a graphical comparison to help you evaluate the models.

44. In the Explorer pane, expand the **Assess/Classification** folder, and then add to the worksheet a **Classification Agreement** node and a **Lift Chart** node. Link each of these nodes to the outputs of each of the modeling nodes.

Notice that the input port of these new nodes is a diamond instead of the more common triangle. This indicates that this type of node accepts more than one input.

45. Run the network.

46. Open the viewer for the **Classification Agreement** node (shown in Figure 3.21) and scroll through the window.



Classification Agreement (5)

Input Node - Logistic Regression (3)

		Predicted		Totals
		NoDefault	Default	
Observed	NoDefault	2428	82	2510
	Default	123	867	990
Totals		2551	949	3500

	Observed		Overall
	NoDefault	Default	
% Agree	96.7%	87.6%	94.1%

Positive Category - Default

Recall	Precision	F-Measure
87.6%	91.4%	89.4%

Input Node - Classification Tree (4)

		Predicted		Totals
		NoDefault	Default	
Observed	NoDefault	2457	53	2510
	Default	74	916	990
Totals		2531	969	3500

	Observed		Overall
	NoDefault	Default	
% Agree	97.9%	92.5%	96.4%

Positive Category - Default

Figure 3.21: *The viewer for the* ***Classification Agreement*** *node.*

The **Classification Agreement** component produces *confusion matrices*, which indicate the number and proportion of observations that are classified correctly by the models.

The classification tree has the highest overall success (largest **Overall % Agree**) at 96.4%. The prediction rate of the logistic regression model is 94.1%.

47. Close the viewer for the **Classification Agreement** node.
48. Open the viewer for the **Lift Chart** (see Figure 3.22).

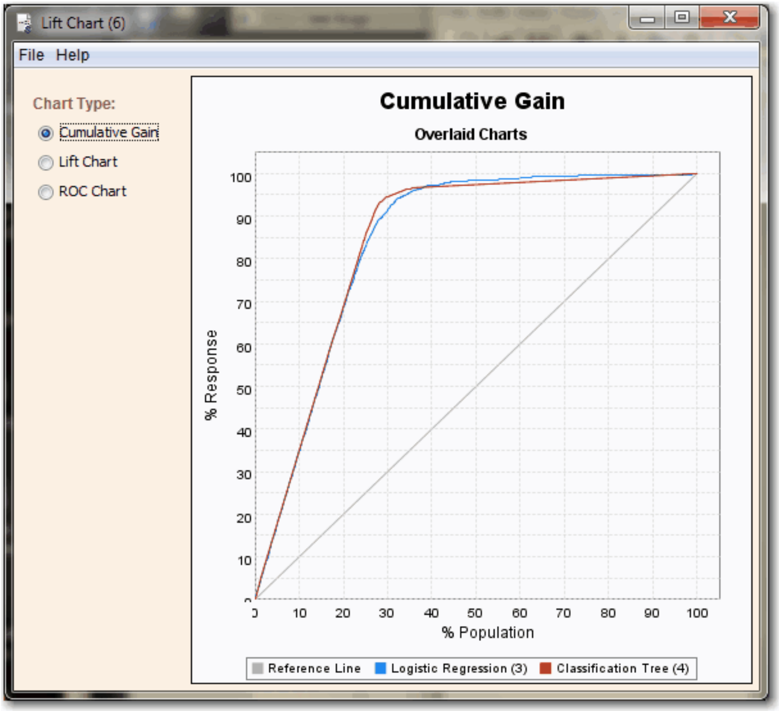


Figure 3.22: *The viewer for the* ***Lift Chart*** *node.*

The viewer for the **Lift Chart** node provides a graphical comparison of the models. This component computes and displays three different charts: *lift, cumulative gain,* and *ROC.* The charts consist of a lift curve and a baseline. The greater the area between the lift curve and the baseline, the better the model.

As you can see in Figure 3.22, the two curves intertwine, so essentially they are equivalent .

49. Close the viewer for the **Lift Chart** node.

Testing the Models

Now that you have created the models and done a preliminary comparison, you can test the models using a new data set. First, read in the test data that you created in the explore worksheet. Then create predictor nodes from the two models, and then apply those predictors to the testing data.

Read The Test Data

50. Add a **Read Text File** node to the worksheet below the existing one.
51. Open the properties dialog, click **Browse**, and select **mortdef.test.txt**.
52. Click the **Modify Columns** tab.
53. Click anywhere in the row containing the variable name **Status** to select it.
54. In the **Set Roles** group, click **Dependent**, and in the **Set Types** group, click **Categorical**.
55. Click Column name `ID`, and in the **Select Columns** group, click **Exclude**.
56. Click **OK** to close the dialog.

Create a Predictor Node

57. Right-click the **Classification Tree** node, and from the menu, click **Create Predictor**.
58. Note that a **Predict: Classification Tree** node appears on the worksheet with a red dashed line connecting it to the model. Reposition the new node, if necessary, and link it to the output of the **Read Text File** node.

Note

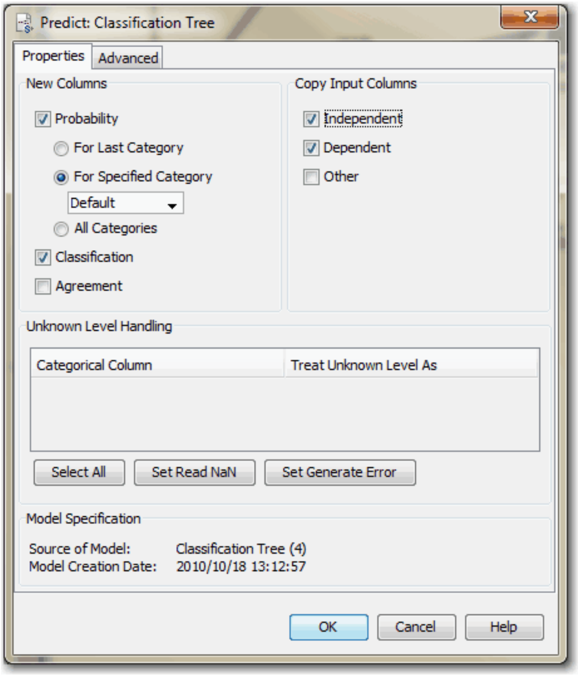
The model ports are circular to distinguished them from other input and output ports, which have triangular or diamond shapes.

Deleting the model link creates a static predict node, meaning that the predictive model does not change even if the model from which it was generated changes.

59. Open the properties dialog of the **Predict: Classification Tree** node.

60. To add the independent variables to the output data, in the **Copy Input Columns** group, select **Independent**.

    The completed dialog page is shown in Figure 3.23.



New Columns

☑ Probability

○ For Last Category

● For Specified Category

Default

○ All Categories

☑ Classification

☐ Agreement

Copy Input Columns

☑ Independent

☑ Dependent

☐ Other

Unknown Level Handling

Categorical Column	Treat Unknown Level As

Model Specification

Source of Model: Classification Tree (4)
Model Creation Date: 2010/10/18 13:12:57

Figure 3.23: *The* ***Properties*** *page of the* ***Predict: Classification Tree*** *dialog.*

61. Click **OK** to close the dialog.

62. Using the Logistic Regression node (rather than Classification Tree), repeat Steps 57-60.

63. Run the network.

Comparing Models

Next, view the results of the models on the testing data. Good variable selection means that the percentage of correct classifications on the testing data is very similar to the percentage on the training data. Typically, it is slightly lower. Do not expect large differences in the cumulative gain or lift of the models. If you see large differences, you should adjust the models.

64. Add a **Classification Agreement** node and a **Lift Chart** node to the worksheet.
65. Link each of these nodes to the outputs of the **Predict** nodes.
66. Run the network.
67. Open the viewers for both assessment nodes.

The prediction rates of these two models are very close–94.2% for the classification tree compared to 94.8% for the logistic regression model. These results make the model choice a bit arbitrary. For demonstration purposes, select the logistic regression model and export it to a PMML model file. By exporting the model, you can use it later in another worksheet to score the data.

Exporting a Model

68. From the **Model/Files** folder in the explorer pane, create an **Export PMML** node and position it above the top **Classification Agreement** node, as shown in Figure 3.13.
69. Connect the model port on the right hand side of the **Logistic Regression** node to the input model port of the **Export PMML node**.
70. Open the properties dialog for the **Export PMML** node and set the **PMML File Name** to be **logisticRegModelMortgage.xml**.
71. Click **OK** to close the dialog.
72. On toolbar click **Run To Here** (▶|).

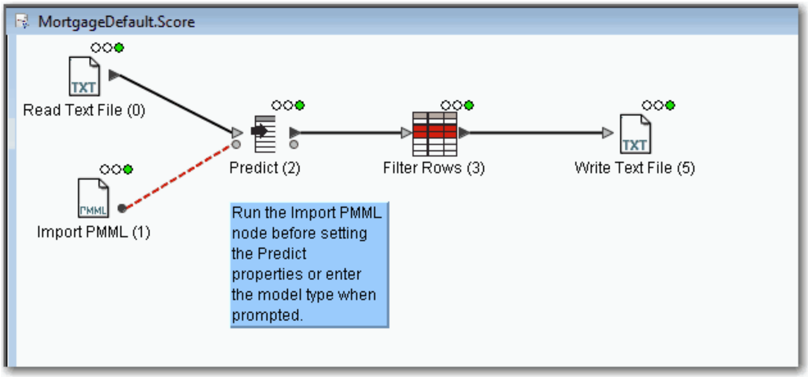
In this example, the viewer for the **Export PMML** node looks similar to the viewer for the **Logistic Regression** node; however, this is not always the case.

Often, the final step in the modeling process is to use validation data to assess the generalization error of the final selected model. Because you used the training data to construct the models and the testing data to select a model, the error measures based on these data sets are biased towards being higher than the error you expect to see on new data. Looking at a new data set (the validation data) can provide unbiased estimates of the error. This exercise does not demonstrate this step.

DEPLOY MODEL

The last step in the TIBCO Spotfire Miner approach to data mining is the scoring/ deployment step. In this step, using a new data set, your model predicts the probability that each customer will not default (`NoDefault`).

To mimic a real, production situation, create a new worksheet, and then import the model and the new data. Then predict the customers who will not default on their home mortgage loans with a probability of > .98 and write these results to a text file for delivery. The completed worksheet is shown in Figure 3.24.



Run the Import PMML node before setting the Predict properties or enter the model type when prompted.

Figure 3.24: *Completed worksheet for scoring new data using an imported model.*

Add the first three network nodes to the worksheet, and then set the properties for each node. The completed properties dialog for the **Read Text File** node is shown in Figure 3.25.

Importing the Scoring Data

1. From the **Data Input/File** folder in the explorer pane, double-click the **Read Text File** component to add a **Read Text File** node to the worksheet.
2. From the **Model/File** folder in the explorer pane, Double-click the **Import PMML** component to add an **Import PMML** node to the worksheet. Position this node below the **Read Text File** node.

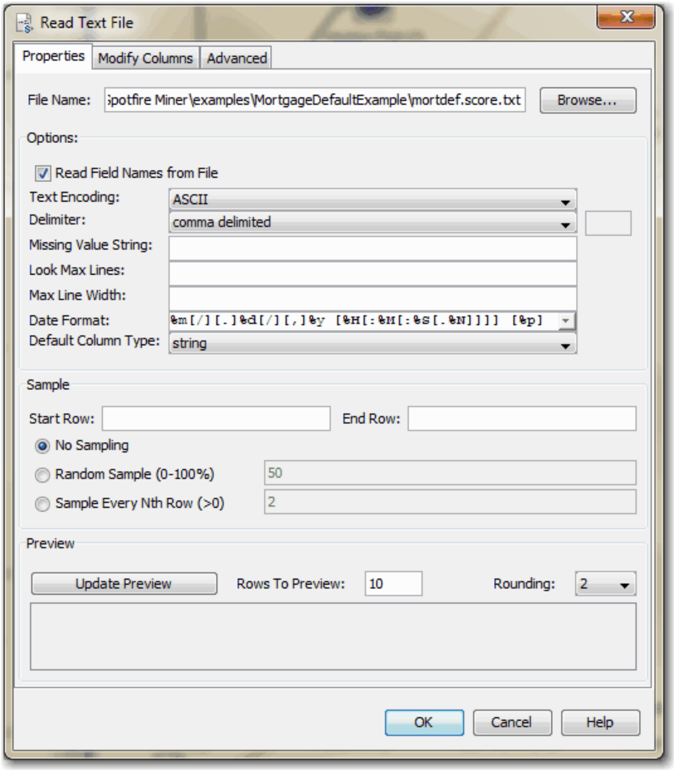
3. From the **Model/Prediction** folder in the explorer pane, Double-click the **Predict** component to add a **Predict** node to the worksheet. Position this node to the right of the other two nodes.

4. Link the output port of the **Read Text File** node to the input port of the **Predict** node.

5. Link the output model port of the **Import PMML** node to the input model port of the **Predict** node.

6. Double-click the **Read Text File** node to open its properties dialog.

7. Click **Browse**, select **mortdef.score.txt** and click **Open**.



File Name: ;potfire Miner\examples\MortgageDefaultExample\mortdef.score.txt

☑ Read Field Names from File

Text Encoding: ASCII

Delimiter: comma delimited

Missing Value String:

Look Max Lines:

Max Line Width:

Date Format: %m[/][.]%d[/][,]%y [%H[:%M[:%S[.%N]]]] [%p]

Default Column Type: string

Start Row: End Row:

● No Sampling

○ Random Sample (0-100%) 50

○ Sample Every Nth Row (>0) 2

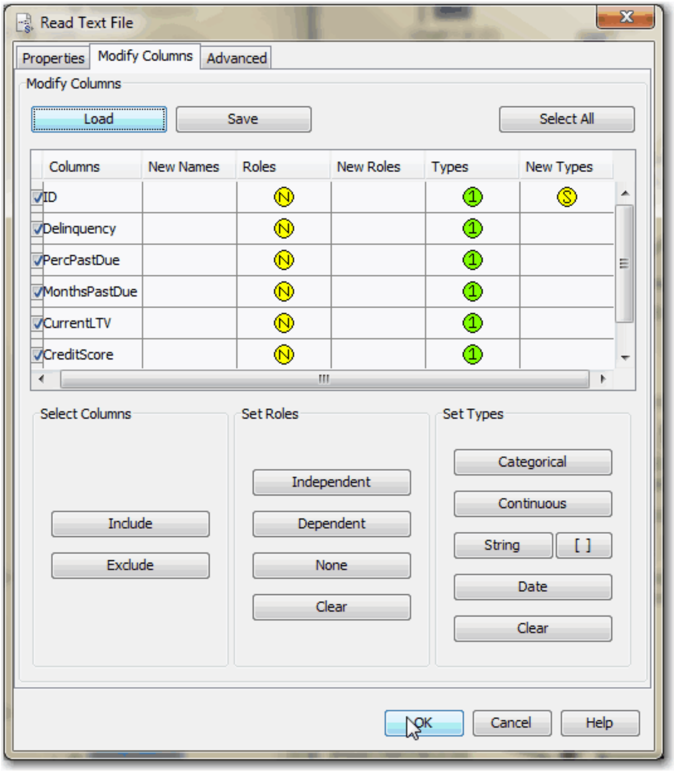
Rows To Preview: 10 Rounding: 2

Figure 3.25: *Completed* ***Read Text File*** *dialog for the scoring data set.*

8. Click the **Modify Columns** tab. (The completed dialog is shown in Figure 3.26.

9. Click anywhere in the `ID` variable row.

10. In the **Set Types** group, click **String**.)



Columns	New Names	Roles	New Roles	Types	New Types
☑ID		N		1	S
☑Delinquency		N		1	
☑PercPastDue		N		1	
☑MonthsPastDue		N		1	
☑CurrentLTV		N		1	
☑CreditScore		N		1	

Figure 3.26: *The completed **Modify Columns** dialog for the **Read Text File** node which imports the scoring data.*

11. Click **OK** to close the dialog.

Importing the Model

12. Double-click the **Import PMML** node to open its properties dialog.

13. Type **logisticRegModelMortgage.xml** into the **PMML File Name** box or browse for the file in this **MortgageDefaultExample** folder, and then click **OK**.

14. While this node is still selected, execute it by clicking **Run To Here** . This imports the model so that the properties of the **Predict** node can be set.

Predicting

15. Double-click the **Predict** node to open its properties dialog. On the **Properties** page of the predict node, in the **Copy Input Columns** group, clear the **Dependent** variable box and select **Independent** and **Other**. Click **OK**.

Setting the **Predict** properties as described in step 15 prevents the model from looking for `Status`, and outputs the customer `ID` column with the predicted data.

Before running the network, add a **Filter Rows** node to filter for only those customers whose probability of `NoDefault` is greater than .98. Then you can export this list to a text file for delivery.

16. From the **Data Manipulation/Rows** folder in the explorer pane, double-click the **Filter Rows** component to add a **Filter Rows** node. Position the new node to the right of the **Predict** node, and link it to the **Predict** node.
17. Double-click the node to open its properties dialog.
18. On the **Properties** page, in the **Qualifier** box, type `get("Pr(Default)") > 0.98`. Click **OK** to close the dialog.
19. From the **Data Output/File** folder in the explorer pane, double-click the **Write Text File** component to add a **Write Text File** node. Position the new node to the right of the **Filter Rows** node, and link it to the **Filter Rows** node.
20. Double-click the node to open its properties dialog.
21. On the **Properties** page, click **Browse** and navigate to the **examples/MortgageDefaultExample** folder.
22. In the **File Name** box, type **CustomerNoDefault.txt**, and then click **Open**. In the **Delimiter** list, select **tab delimited**. Click **OK**.
23. Run the network.

You have created a tab-delimited text file containing information about which customers are most likely to not default on their loans. You can now deliver this file to a bank's loan department to use in loan risk assessment.

EXPLORE THE SPOTFIRE S+ LIBRARY

You can add to your exploratory and predictive capabilities using Spotfire S+ graphs, and you can create complex models using the **S-PLUS Script** node.

The S language engine from Spotfire S+ is part of the basic TIBCO Spotfire Miner™ system and does not need to be installed explicitly. The **Spotfire S+** page appears in the explorer pane. We do not describe the S language engine in detail in this book. Consult the printed or online documentation for Spotfire S+ for more detailed information.

> **Note**
>
> Spotfire Miner works only with the included S-PLUS libraries and S language engine. You cannot use an externally-installed version of Spotfire S+ with Spotfire Miner. If you plan to work with Spotfire S+ or the S language extensively, consider using TIBCO Spotfire S+.
>
> Spotfire S+ provides features that are not included in Spotfire Miner, such as the Spotfire S+ GUI, Spotfire S+ Workbench integrated developer environment, Spotfire S+ console application (sqpe), plus support for Automation and for other interfaces (including OLE, DDE, and COM).

Using Spotfire S+ Graphs

Click the **Spotfire S+** tab in the explorer pane to show the S-PLUS nodes. Using Spotfire S+ provides several exploratory graphing options that can provide more information about the data. For example, look at a histogram of the predicted probabilities conditioned on the dependent variable, `Status`.

1. Open the supplied example worksheet, **examples/MortgageDefaultExample/MortgageDefault.Model.imw**.
2. Run the network.
3. From the **Spotfire S+** page, open the **Explore/One Column - Continuous** folder, drag and drop the **Histogram** component, placing it close to the **Predict: Logistic Regression** node.

The completed worksheet (**MortgageDefault.Model-SPlus**) is shown in Figure 3.27.

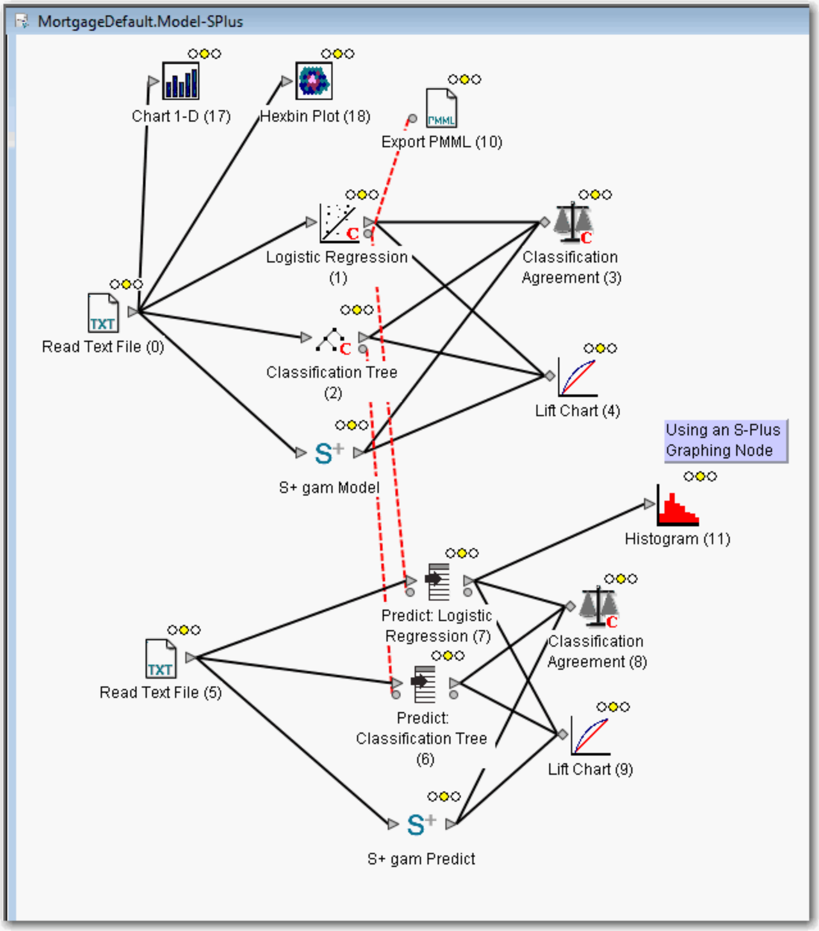
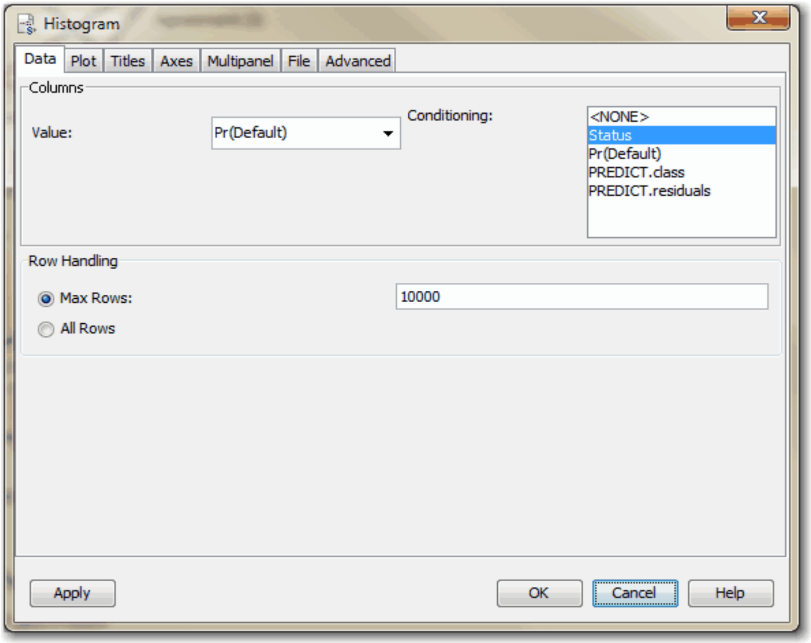


Figure 3.27: *The completed worksheet (**MortgageDefault.Model-SPlus.imw**) which calls S-PLUS nodes to graph results and compare models, including a Spotfire S+ model for the mortgage default data.*

4. Connect the **Predict Logistic Regression** node to the **Histogram** node, and then double-click the **Histogram** node to open its properties dialog. The completed dialog is shown in Figure 3.28.



Columns

Value: Pr(Default)

Conditioning:
<NONE>
Status
Pr(Default)
PREDICT.class
PREDICT.residuals

Row Handling

● Max Rows: 10000

○ All Rows

Figure 3.28: *The completed properties page for the* ***Histogram*** *node.*

5. In the **Columns** group, in the **Value** box, select `Pr(Default)`, and in the **Conditioning** box, select `Status`.

6. Click **Apply** to run the node and display the histogram of the plot `Percent of Total` vs. `Pr(Default)` (Figure 3.29).

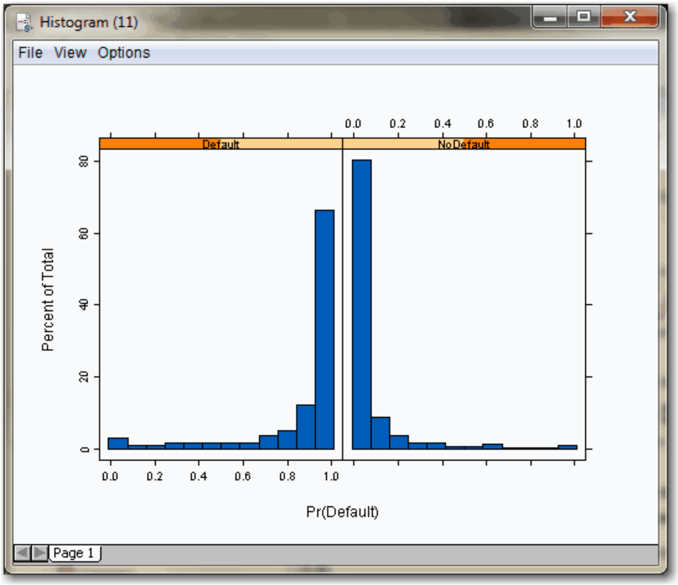


Figure 3.29: *A histogram of* ***Percent of Total*** *vs.* ***Pr(Default)*** *for the logistic regression model.*

This plot shows that the example has done a good job of predicting the probabilities; however, a logistic regression model cannot capture non-linear relationships between the predictors and the response: situations where there are thresholds. Would a non-linear model yield better predictions? An alternative model could be a GAM model (that is, a generalized additive model). A binomial GAM model is similar to a logistic regression model, but instead of having the predictor variable affect the response in a linear fashion, a GAM model allows the relationship to be an arbitrarily smooth function. For more details, see Hastie and Tibshirani (1990), or see the *Spotfire S+ Guide to Statistics, Volume 1*. You can create this model using an **S-PLUS Script** node, which is described in the next section.

7. If you do not want to modify the original worksheet, save this worksheet with a different file name.

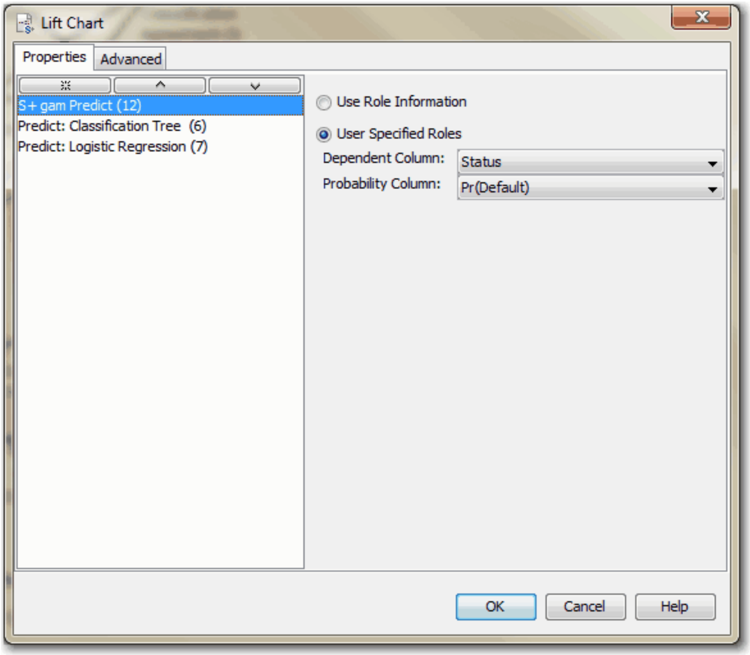
Modeling and Predicting Using S-PLUS Script Nodes

Compare an S-PLUS GAM model to the two previous models: the classification tree and logistic regression. The completed worksheet is shown in Figure 3.27. This worksheet file is **examples/ MortgageDefaultExample/MortgageDefault.Model-SPlus.imw**. You can find more detailed information about creating and using **S-PLUS Script** nodes in the *Spotfire Miner User's Guide.*

Creating a GAM Model Using an S-PLUS Script Node

1. Using the browser, open the **MortgageDefault.Model.imw** worksheet from the **examples/MortgageDefaultExample** folder.
2. From the **Utilities** folder, add an **S-PLUS Script** node to the worksheet. Connect this node to the top **Read Text File (0)** node.
3. Double-click the **S-PLUS Script** node to open its properties dialog.
4. On the **Script** page, click **Load**. Navigate to select the file **examples/MortgageDefaultExample/ MortgageDefault.gamModel.ssc**. Click **Open**.
5. Click **OK** to close the properties dialog.
6. To rename the node, either right-click the node and select **Rename**, or left-click the node name and type **splus gam Model**.
7. Connect the **splus gam Model** node to the **Classification Agreement** and **Lift Chart** nodes.

8. Open the **Lift Chart** node, and in the **Properties** list box, click **S-PLUS Script**. Select **User Specified Roles**, and in the drop-down, ensure that the **Dependent Column** is set to `Status` and the **Probability Column** is set to `Pr(Default)`.



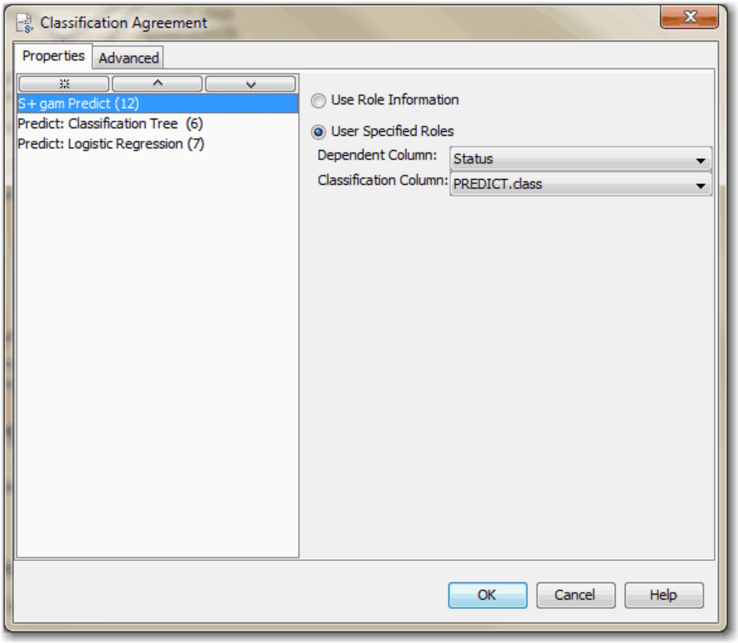
Lift Chart

S+ gam Predict (12)
Predict: Classification Tree (6)
Predict: Logistic Regression (7)

○ Use Role Information
● User Specified Roles
Dependent Column: Status
Probability Column: Pr(Default)

Figure 3.30: ***Lift Chart*** *properties for GAM model.*

9. Open **Classification Agreement** node, and repeat Step 8, except for **Classification Column**, select `PREDICT.class`.



S+ gam Predict (12)
Predict: Classification Tree (6)
Predict: Logistic Regression (7)

○ Use Role Information
● User Specified Roles
Dependent Column: Status
Classification Column: PREDICT.class

Figure 3.31: ***Classification Agreement*** *properties for GAM model.*

10. Run the network.

The GAM model outputs several graphs. The default property outputs these graphs as the node viewer. You can also set the option of viewing the graphs as the node runs using the node properties.

To compare how the new model did compared to the other models:

11. Open the **Classification Agreement** node viewer for the testing data.

The overall percent agreement for the GAM model falls between the logistic regression and classification tree agreement percents.

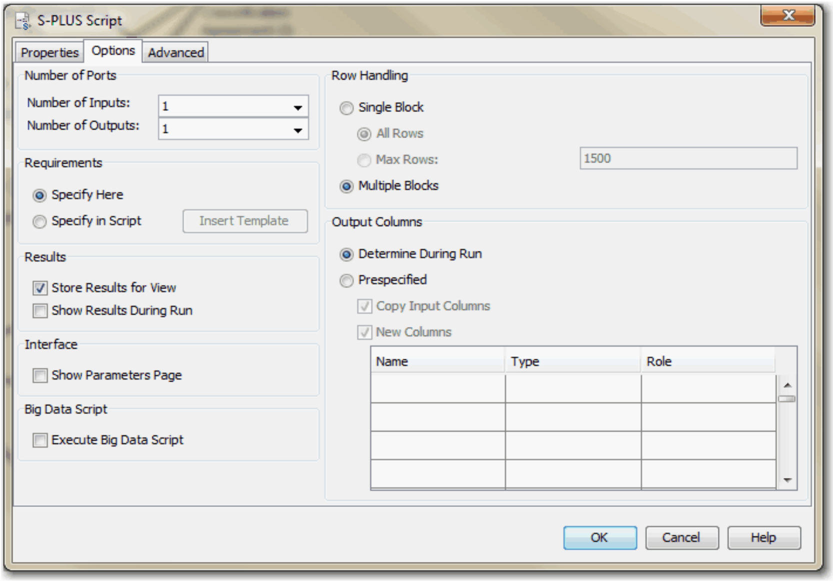
Predicting a GAM Model Using an S-PLUS Script Node

You can create a prediction node from the GAM model by using another **S-PLUS Script** node. In the code of the **splus gam Model** node you wrote out model information. This information can be accessed by another node and used for prediction.

12. Add an **S-PLUS Script** node below the other prediction nodes and connect it to the **Read Text File** node for the testing data.

13. Open the properties dialog to the Script page and load **examples/MortgageDefaultExample/ MortgageDefault.gamPredict.ssc**.

For predictions, especially for large data sets, use the **Multiple Blocks** option.



Number of Ports

Number of Inputs: 1

Number of Outputs: 1

Requirements

● Specify Here

○ Specify in Script

Results

☑ Store Results for View

□ Show Results During Run

Interface

□ Show Parameters Page

Big Data Script

□ Execute Big Data Script

Row Handling

○ Single Block

● All Rows

○ Max Rows: 1500

● Multiple Blocks

Output Columns

● Determine During Run

○ Prespecified

☑ Copy Input Columns

☑ New Columns

Name	Type	Role

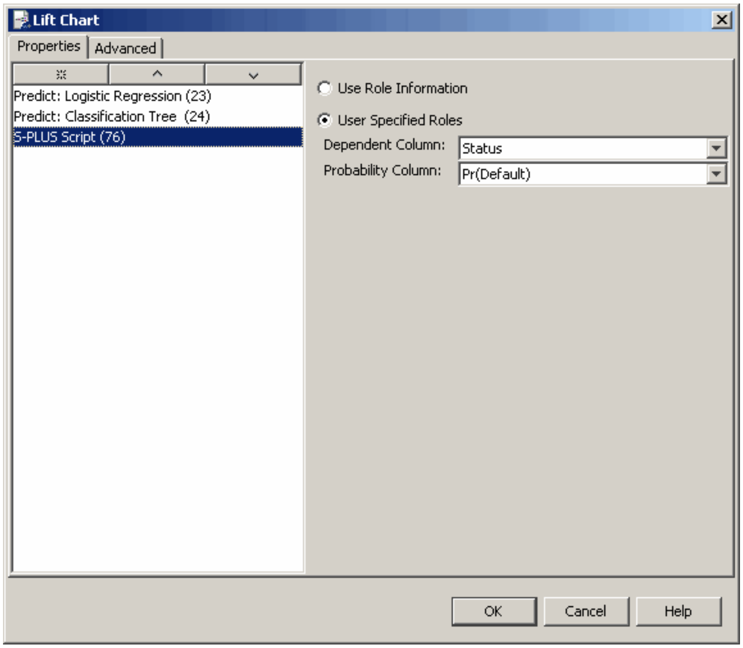
Figure 3.32: *The completed options property page for the* ***S-PLUS script*** *GAM predict node.*

14. Click to open the **Options** page. In the **Row Handling** group, select **Multiple Blocks**.

15. Rename the node to **splus gam Predict**.lo

16. Connect the **splus gam Predict** node to the **Classification Agreement** and **Lift Chart** nodes.

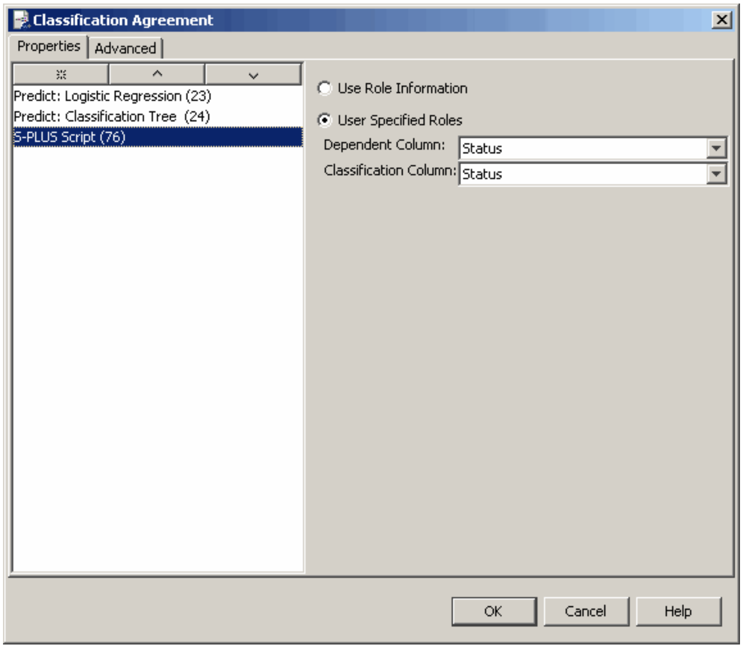
17. Specify the role information as you did in Steps 8 and 9, except in the **Classification Agreement** properties dialog, set **Classification Column** to `Status`. See Figure 3.33 and Figure 3.34 for an example.



Dependent Column: Status

Probability Column: Pr(Default)

Figure 3.33: ***Lift Chart*** *properties for prediction.*



Predict: Logistic Regression (23)
Predict: Classification Tree (24)
S-PLUS Script (76)

○ Use Role Information

● User Specified Roles

Dependent Column: Status

Classification Column: Status

Figure 3.34: ***Classification Agreement*** *properties for prediction.*

18. Run the network.

Copy this prediction node to the scoring worksheet to score the data. The completed worksheet is **examples/MortgageDefaultExample/MortgageDefault.Score-SPlus.imw**.

Add a scoring network to the previous scoring worksheet that uses the **splus gam Predict** node. The completed worksheet is shown in Figure 3.35.

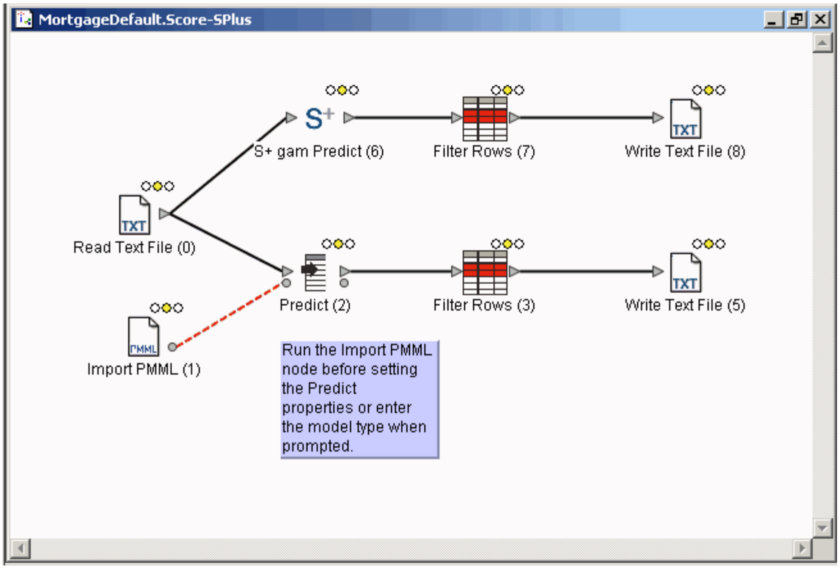


Figure 3.35: *Completed worksheet (**MortgageDefault.Score-SPlus.imw**) for scoring mortgage default data comparing logistic regression and Spotfire S+ GAM model.*

19. Select the **splus gam Predict** node on model worksheet and press CTRL-C to copy the node.

20. Open the scoring worksheet, **examples/ MortgageDefaultExample/MortgageDefault.Score.imw** and press CTRL-V to paste the predict node into this worksheet.

21. Drag the node to position it above the existing predict node and connect it to the **Read Text File** node.

22. Add a **Filter Rows** node and a **Write Text File** node as shown in Figure 3.35 and connect the nodes.

23. Double-click the **Filter Rows** node to open its properties page. In the **Qualifier** box, type `get(“Pr(Default)”) > 0.98`, and then click **OK**.

24. Double-click the **Write Text File** node to open its properties page. In the **File Name** box, type **CustomerNoDefault-gam.txt**.

25. Run the network.

You now have a list of loans that meet the criteria for having a low default probability.

In the example, all the models you created do a good job of predicting the default probabilities. Using the Spotfire S+ Library, you could explore the data in new ways and create more complex non-linear models for the data.

SUMMARY

In this example, you:

- Developed a model to predict the probability of customers defaulting on their home mortgage loans.
- Used all available customer data by merging data files and partitioning the data so that you could train and test the models.
- Created models that were more than 94% accurate in predicting the status of customer loans by using the logistic regression and classification tree models. (You used only the predictors in the model. To improve the models, you could try adding interaction terms.)
- Compared standard Spotfire Miner models to an additional one, a GAM model, provided with Spotfire S+. This model did as well as the other models. (Again, if you take more time to explore the data and variable interactions you could develop an even better model.)
- Finally, you met the objective of creating a list of customers whose risk of defaulting on their loans is within the acceptable risk range. The final list includes the predicted probabilities, which you can use to explore different risk scenarios. You could use this information to make a final decision on which loans to buy.

REFERENCES

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Spotfire S+ Guide to Statistics, Volume 1, TIBCO Software Inc.

INDEX