

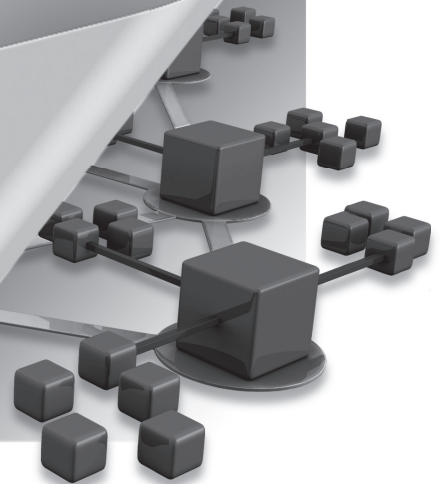
# **Predictive Modeling with SAS<sup>®</sup> Enterprise Miner<sup>™</sup>**

**Practical Solutions for Business Applications**

*Third Edition*

Kattamuri S. Sarma, PhD

## **Solutions to Exercises**



This set of Solutions to Exercises is a companion piece to the following SAS Press book: Sarma, Kattamuri S., Ph.D. 2017. *Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition*. Cary, NC: SAS Institute Inc.

**Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition**

Copyright © 2017, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-62960-264-6 (Hard copy)

ISBN 978-1-63526-038-0 (EPUB)

ISBN 978-1-63526-039-7 (MOBI)

ISBN 978-1-63526-040-3 (PDF)

All Rights Reserved. Produced in the United States of America.

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

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

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

**U.S. Government License Rights; Restricted Rights:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

July 2017

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

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

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

# Solutions to Exercises

**Kattamuri S. Sarma, PhD**

## Chapter 2.

### 2.12 Exercises

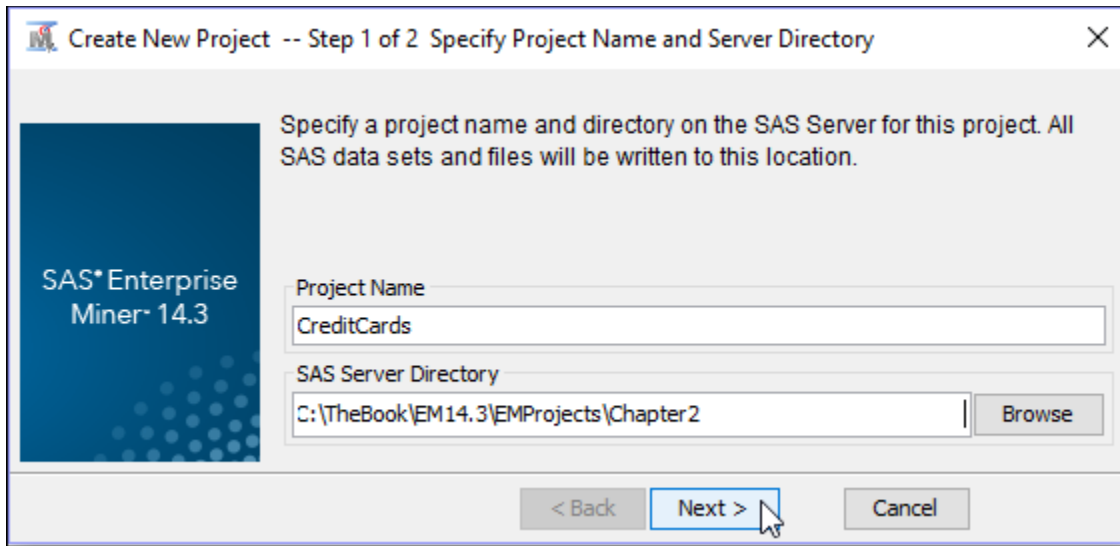
The exercises in this section use the data set Ch2\_Clus\_Data2. This is credit card holder data. The target variable is “Cancel” which takes the value 1 if the card is cancelled and 0 if it is not cancelled during a given period of observation. The data set has 16 variables including the target. The exercises highlight the differences in measurement scales which are based on Meta Data Advisor Options selected (Basic vs. Advanced)

Exercise 1: Create a new project called “Credit Cards”

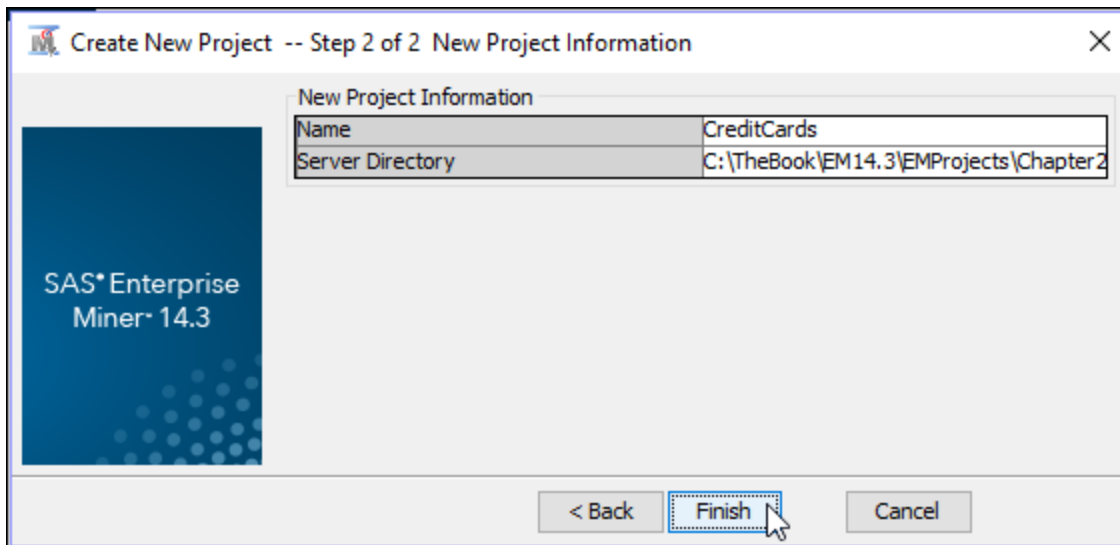
Display 2.1



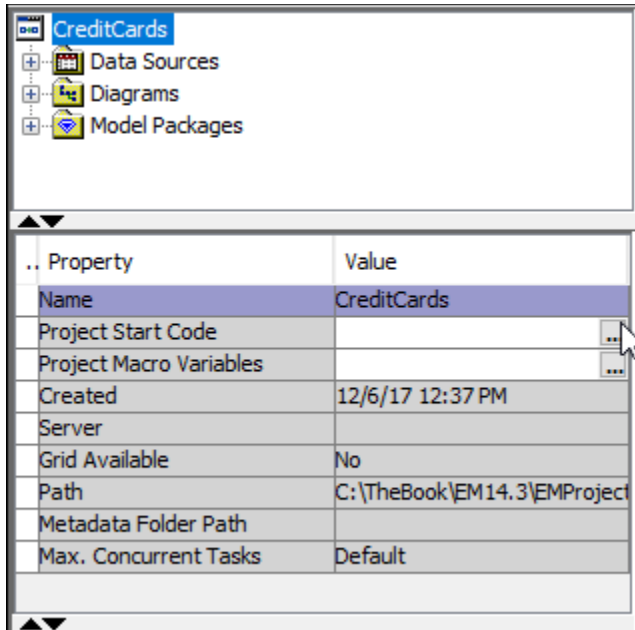
Display 2.2



Display 2.3

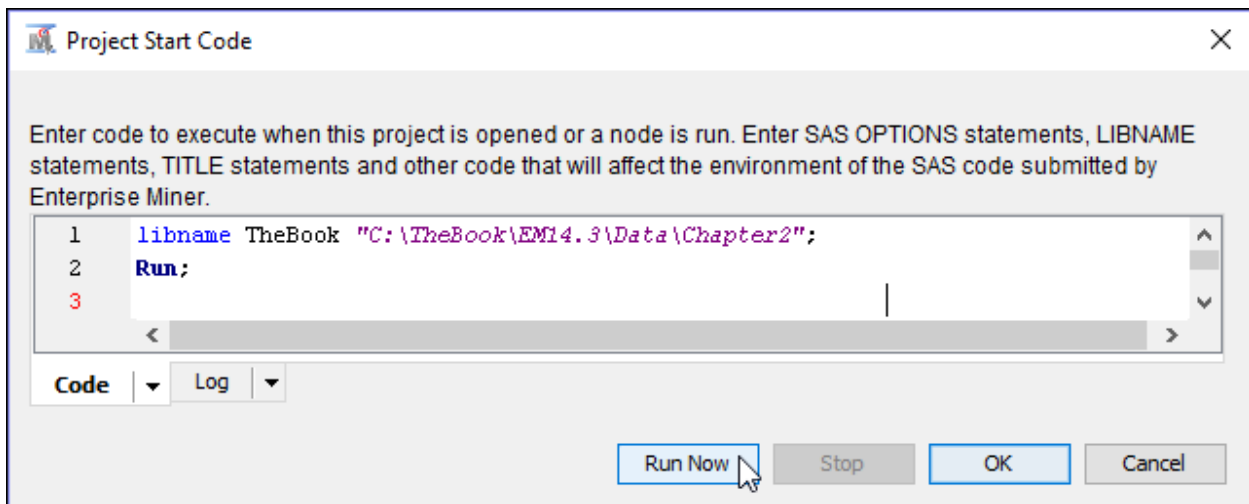


Display 2.4



In order to create a library reference for the data set, we must enter the “libname” statement in the Project Start Code window and click on “Run Now.”

Display 2.5

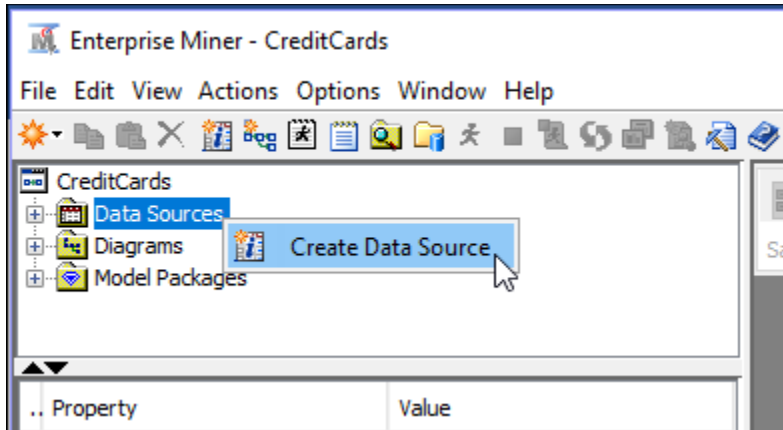


Check the log window to verify that the library is successfully created. Click “OK”.

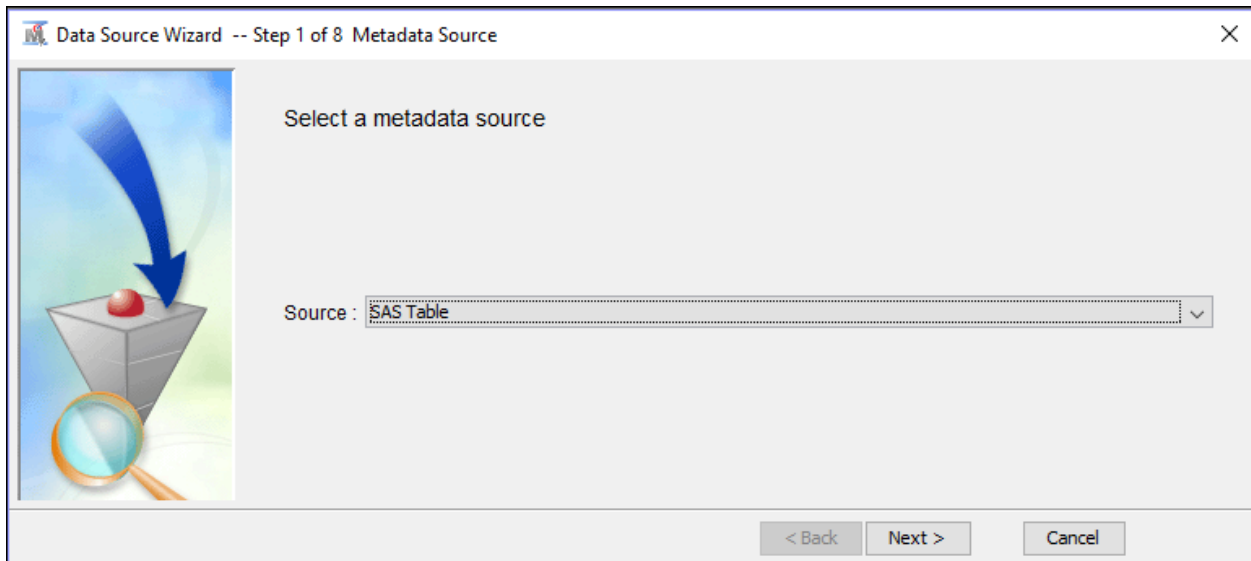
Exercise 2: Create a data source using the sas data set Ch2\_Clus\_Data2. This data set is located in the directory library “TheBook” which we defined earlier (See Display 2.5 and Display 2.6).

Open the data source wizard as shown in Display 2.6.

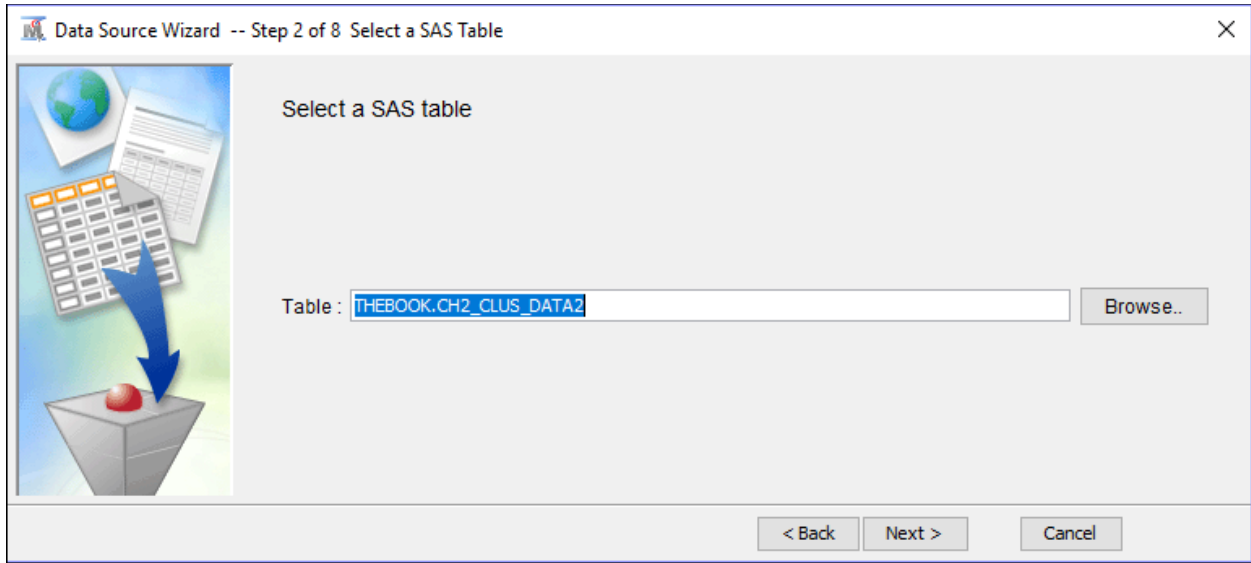
Display 2.6



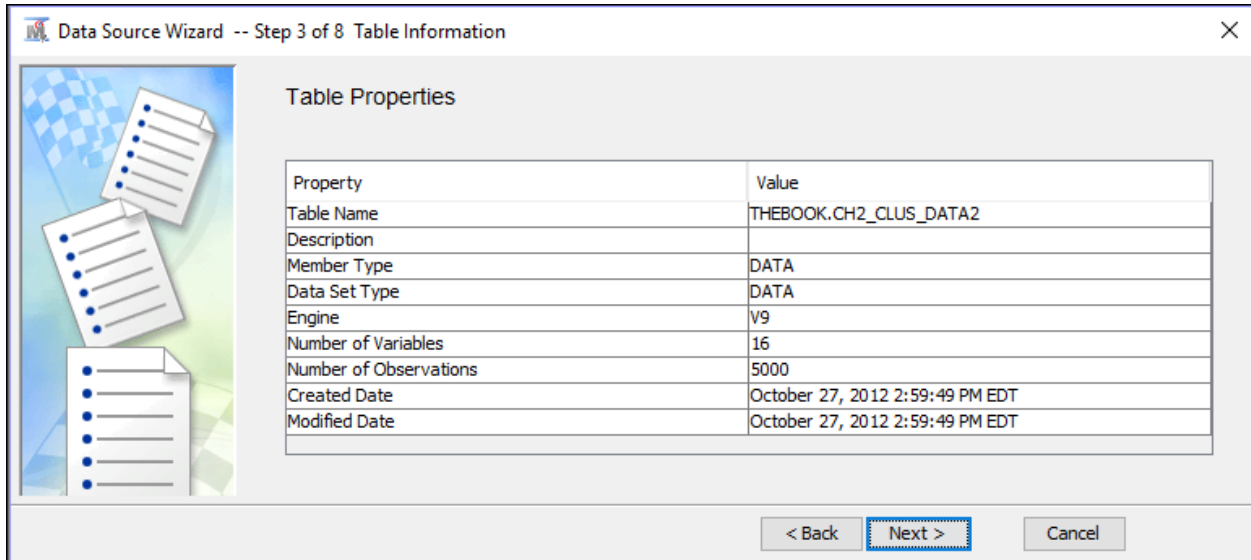
Display 2.7



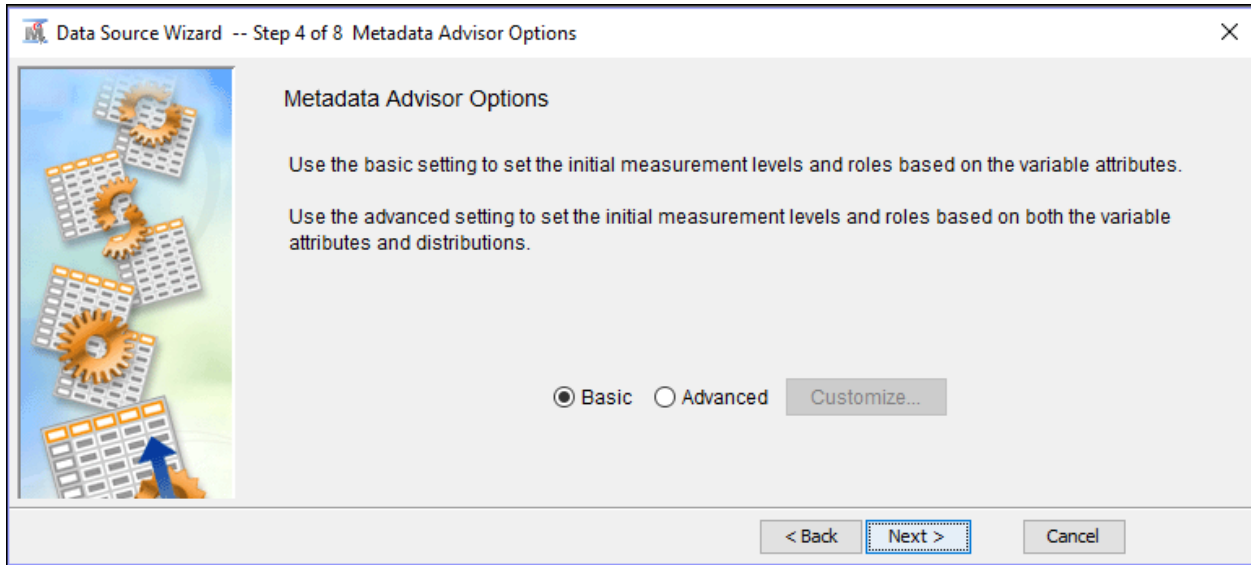
Display 2.8



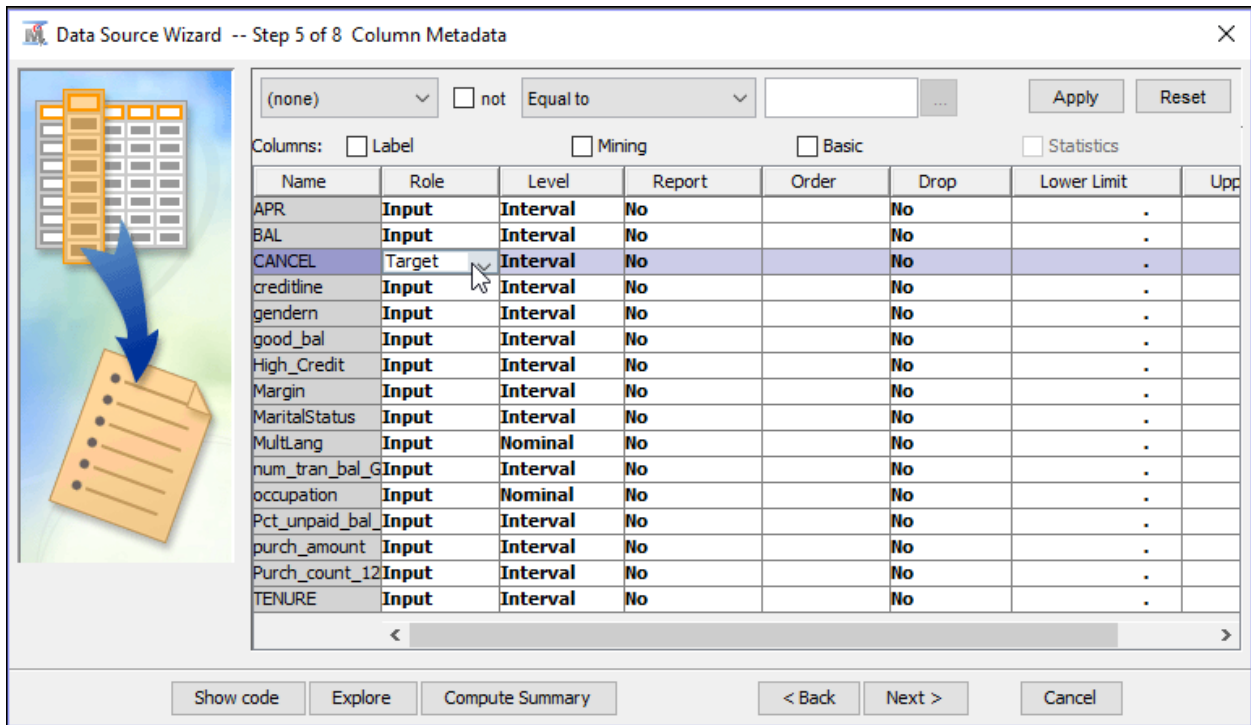
Display 2.9



## Display 2.10

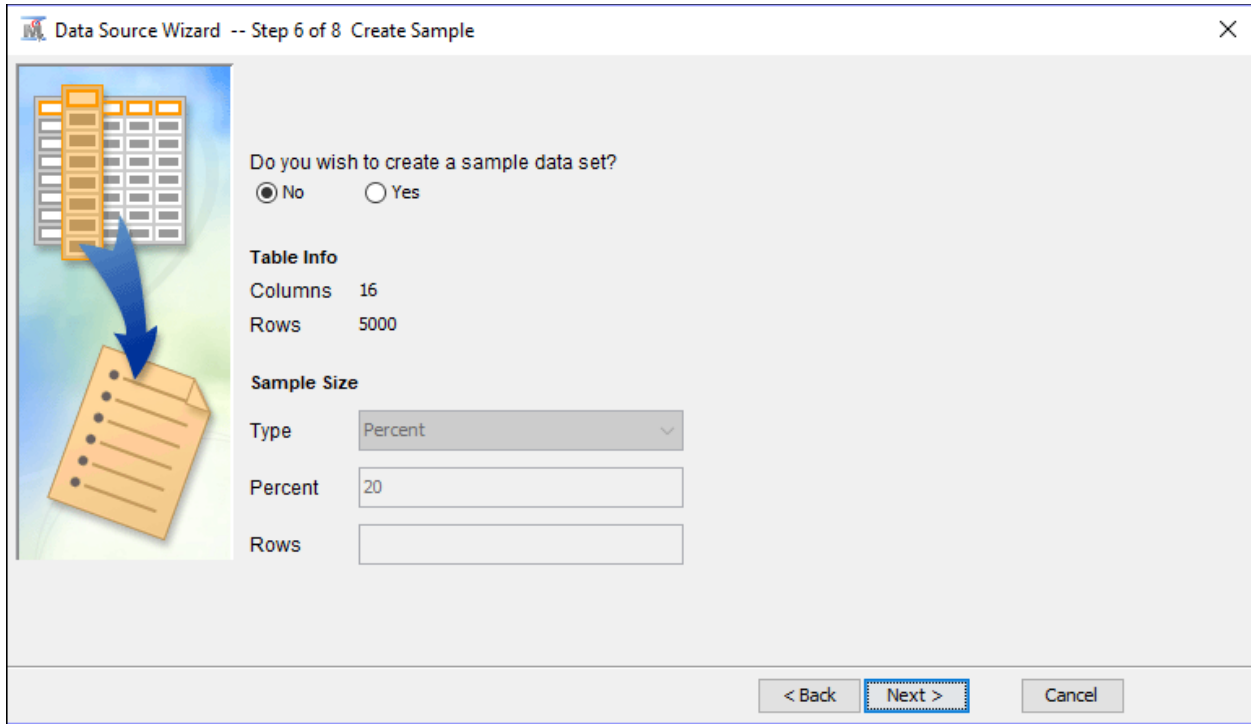


## Display 2.11

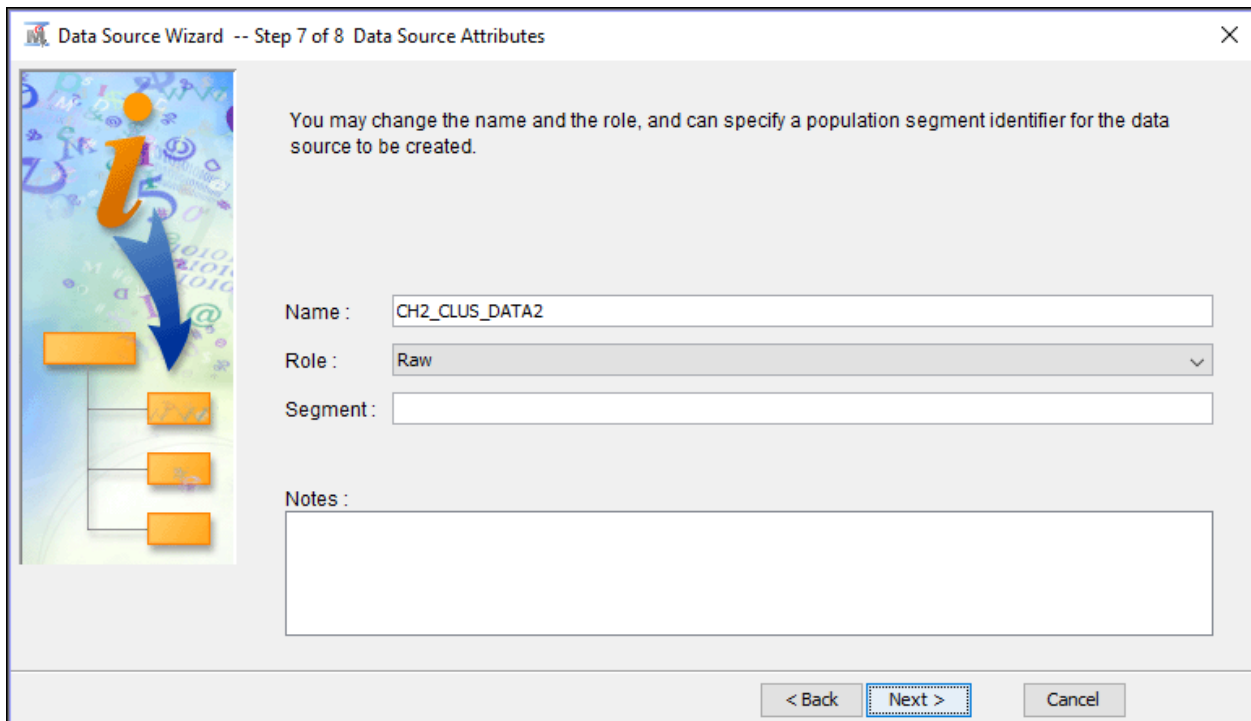




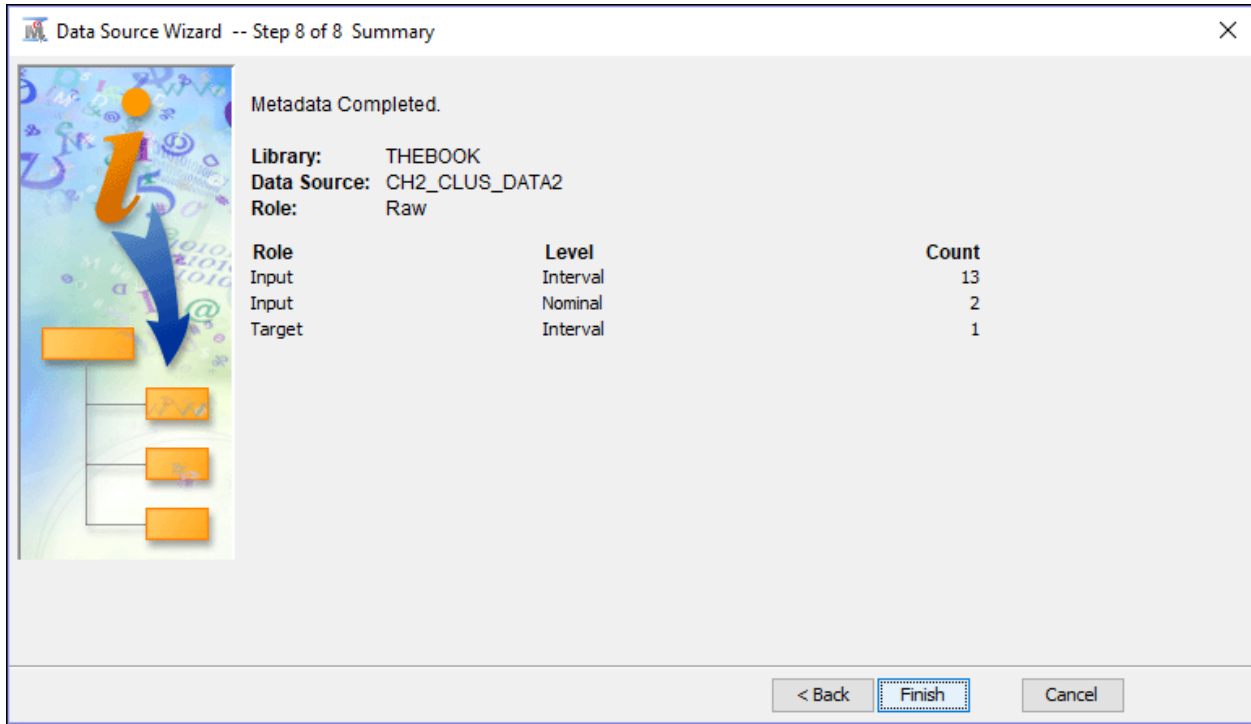
Display 2.12



Display 2.13



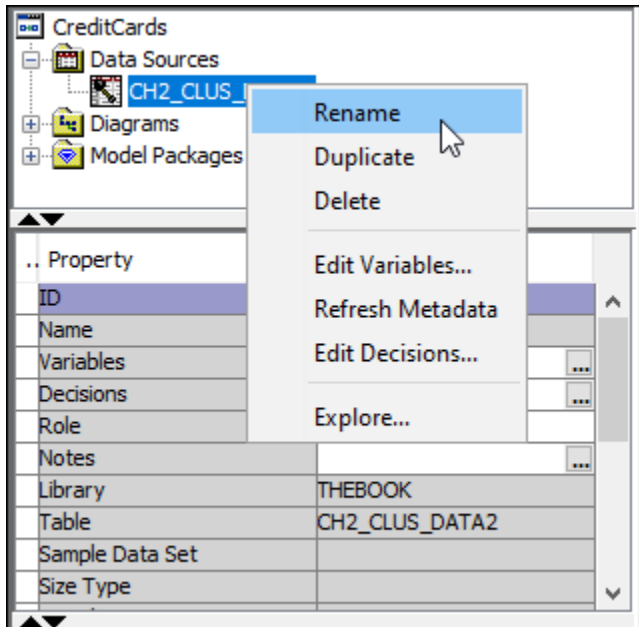
Display 2.14



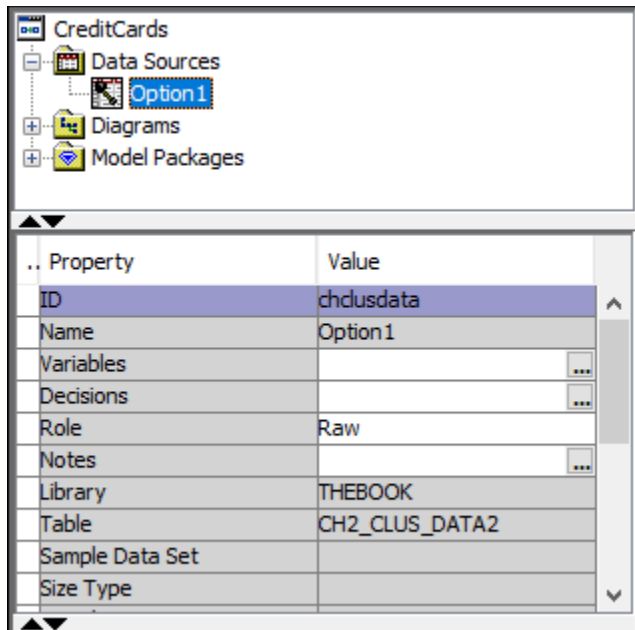
(2d) There are 14 interval scaled variables and 2 nominal scaled variables (see Display 2.14).

(2e) Displays 2.15 and 2.16 show how to rename the data source.

Display 2.15



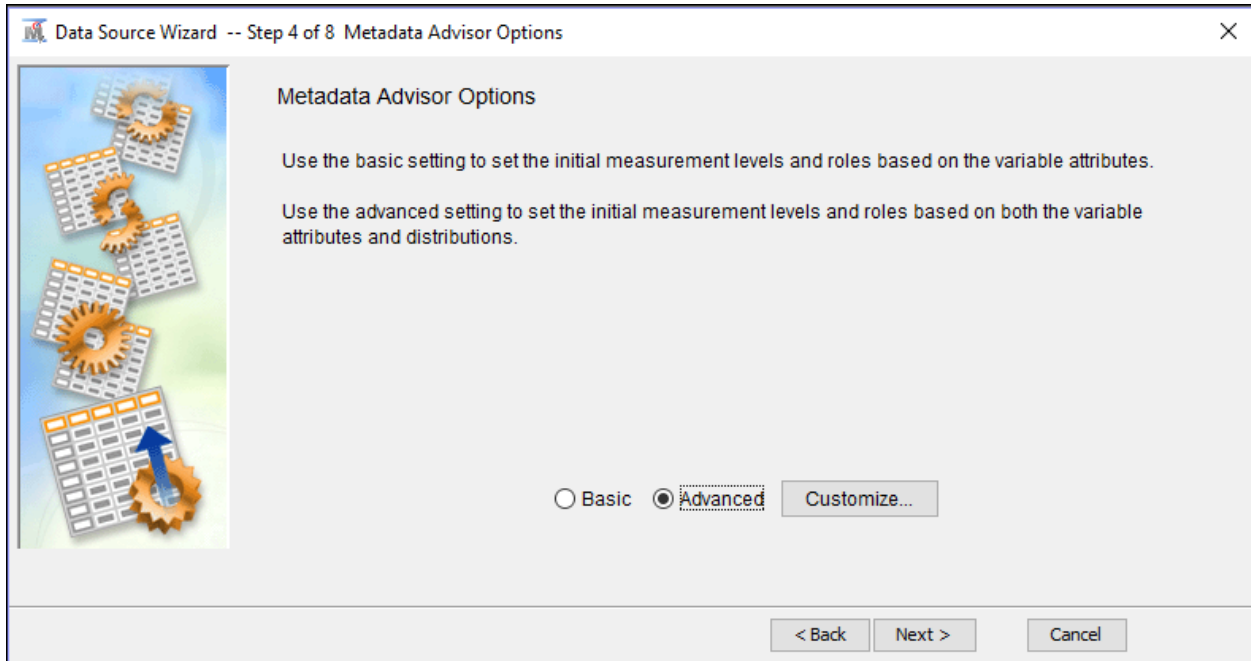
Display 2.16



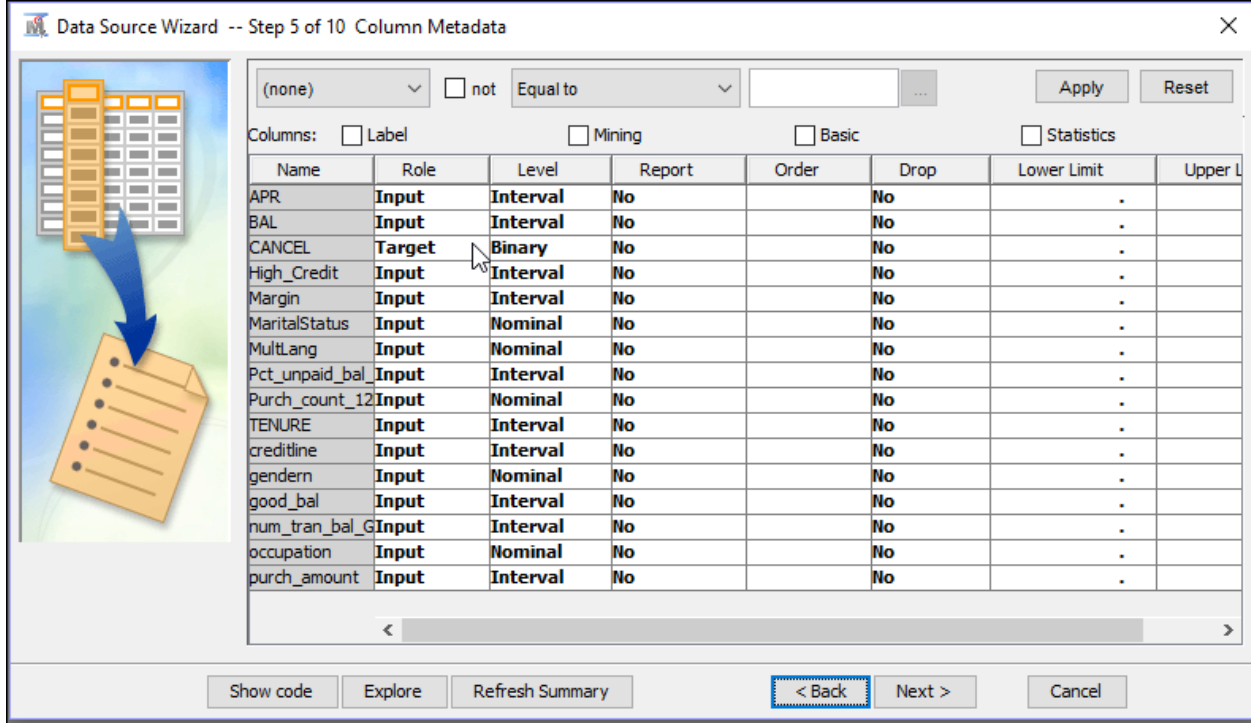
Exercise 3:

In Data Source Wizard, at Step 4 of the Metadata Advisor Options, select “Advanced” as shown in Display 2.17.

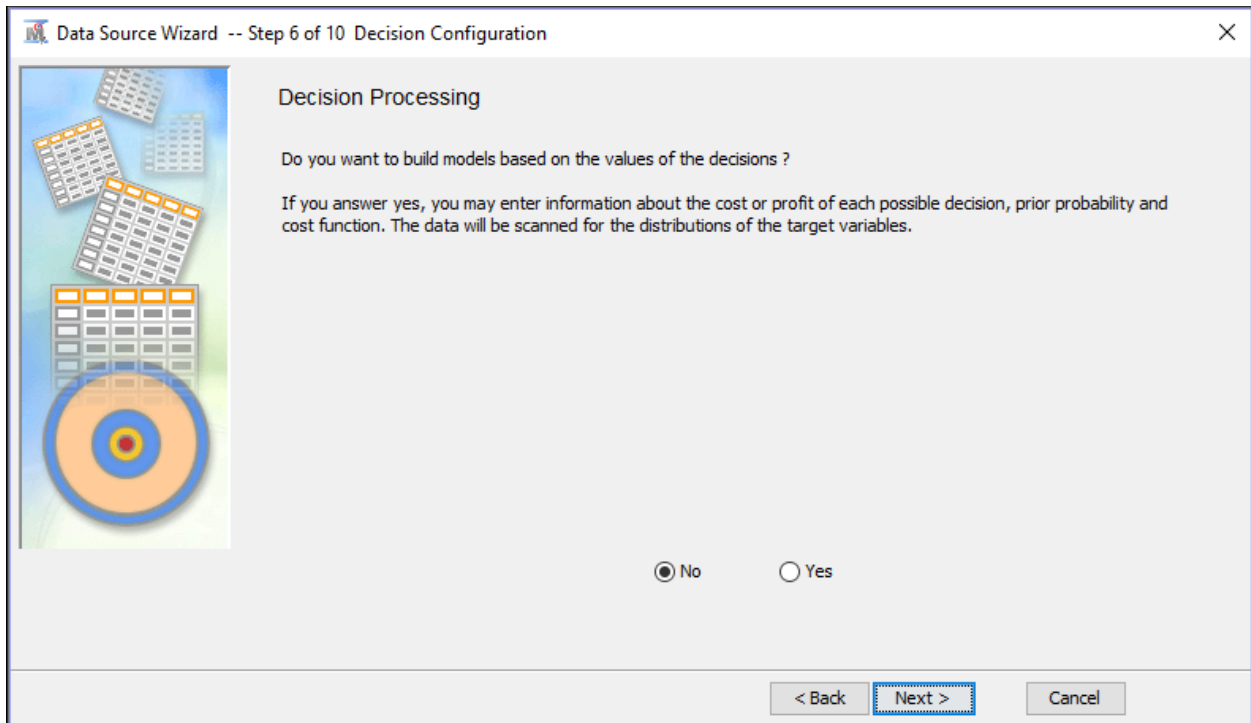
Display 2.17



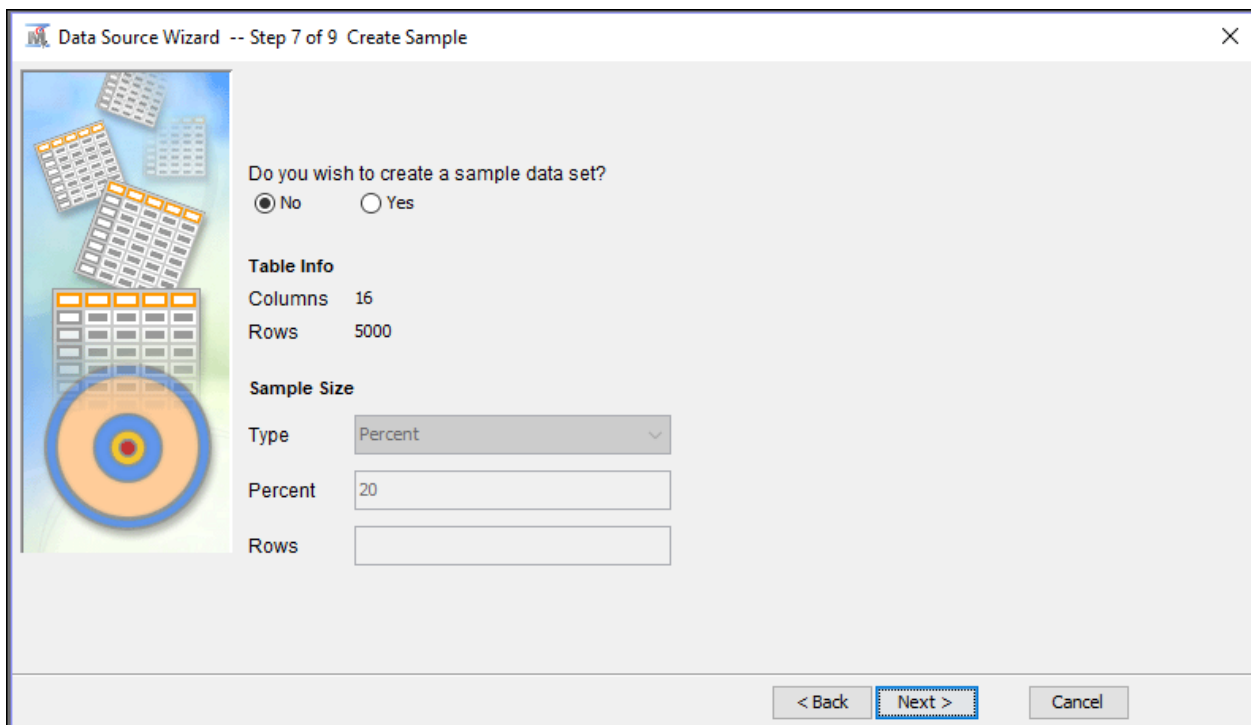
Display 2.17A



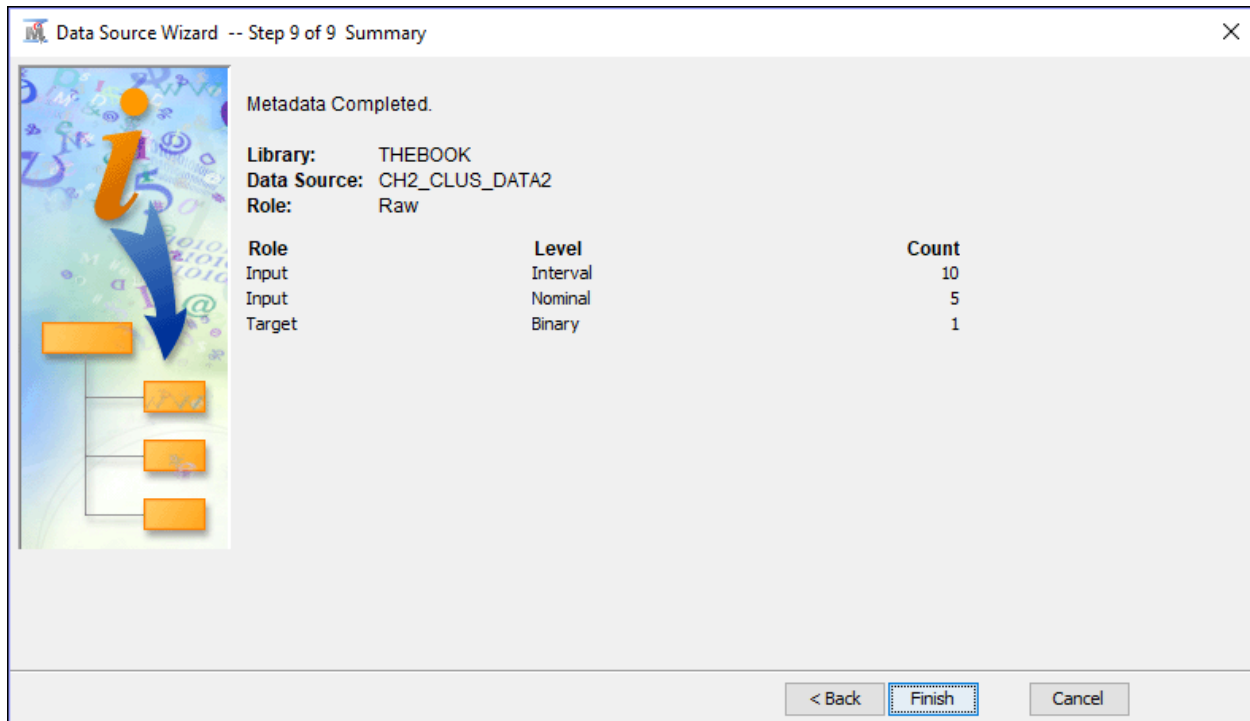
Display 2.18



Display 2.19



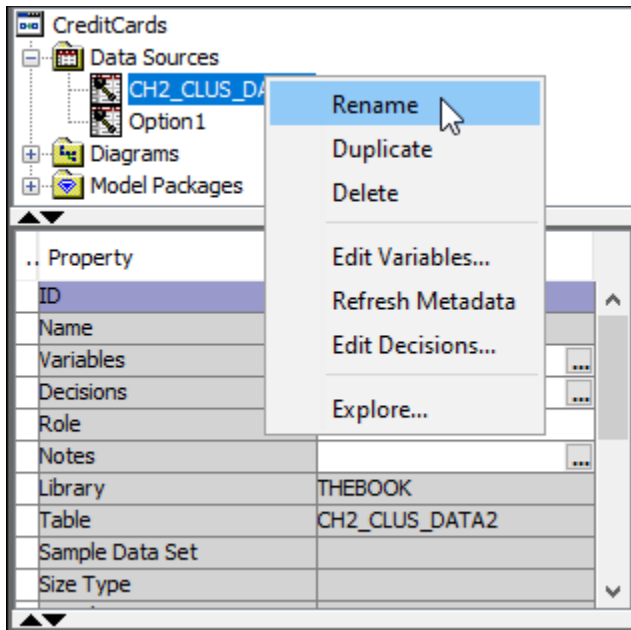
Display 2.20



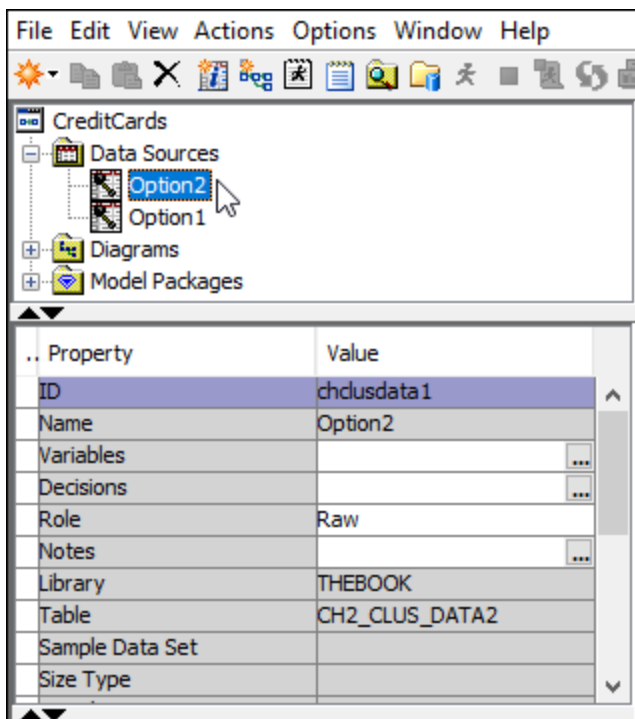
The measurement levels of the variables are shown in Display 2.20. When we select the “Advanced” Option, the Enterprise Miner detects that the Target variable is Binary. When the target is recognized as binary, the regression node fits a logistic regression by default.

When we select the “Basic” option, the Target Variable is treated as “interval” as shown in Display 2.14 and the regression node uses the ordinary least squares method by default.

Display 2.21

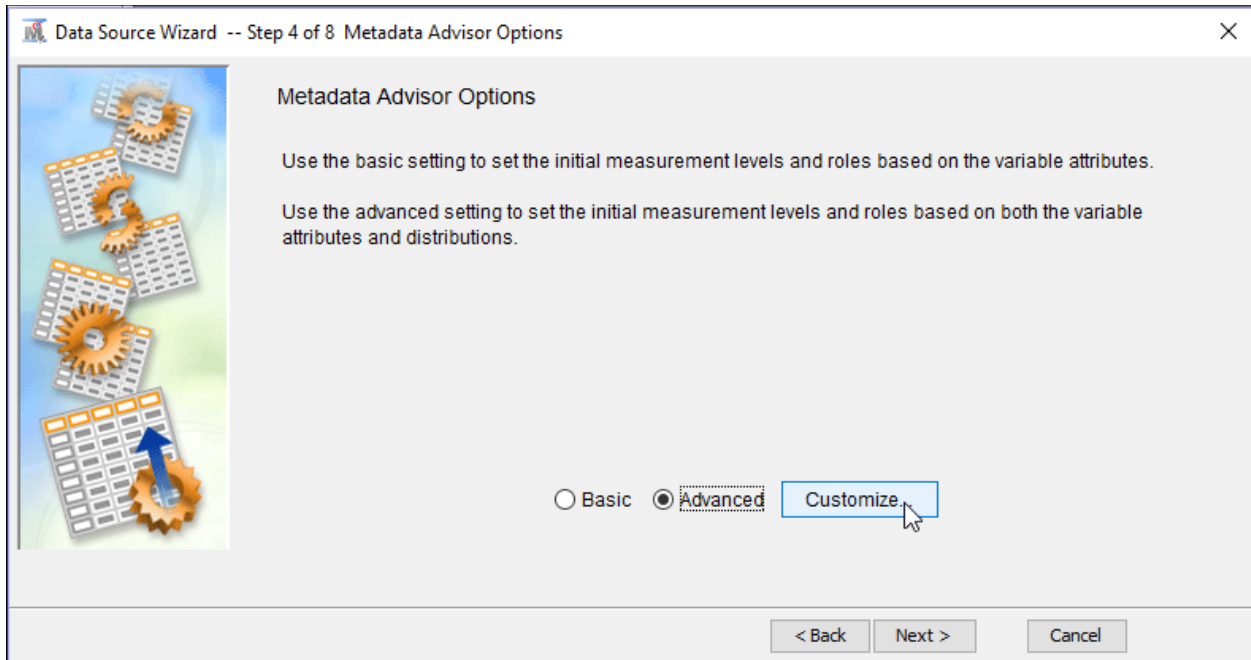


Display 2.22

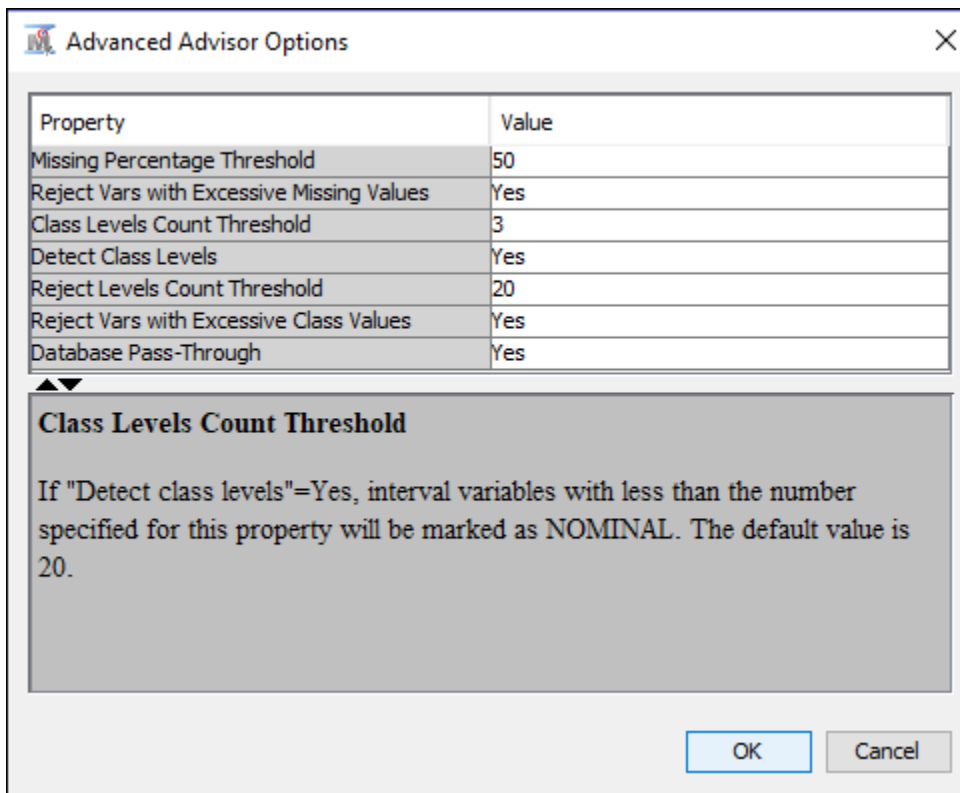


Exercise 4

Display 2.23



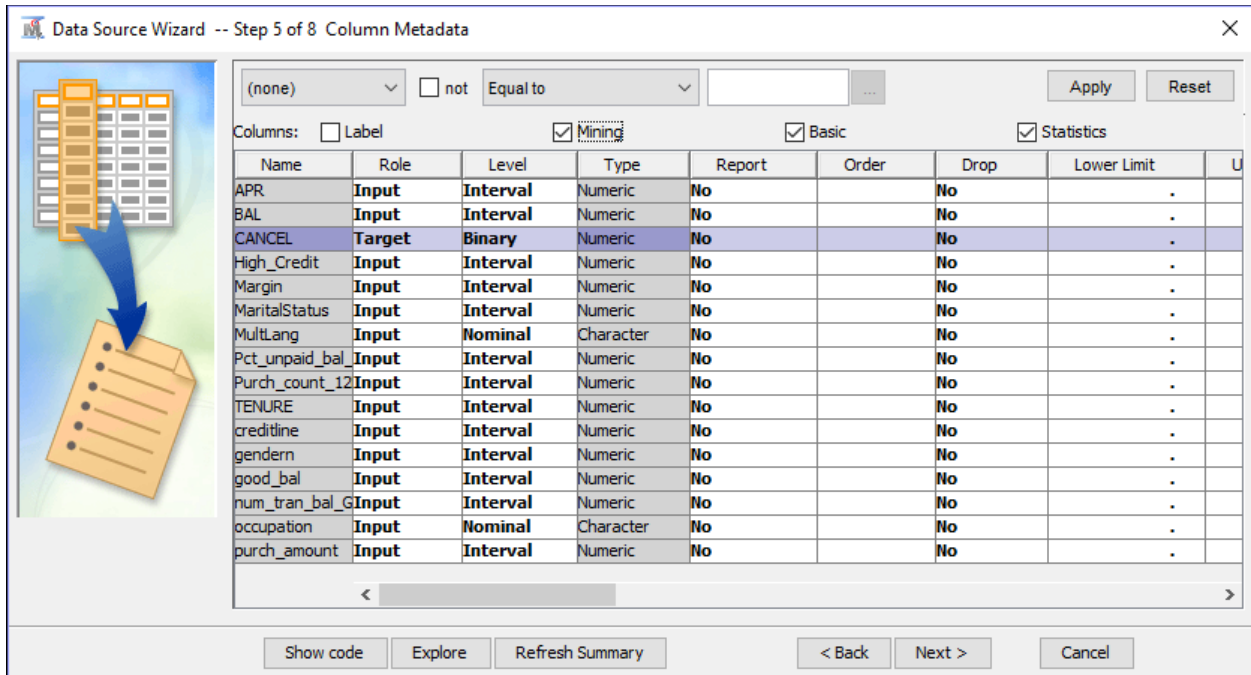
Display 2.24





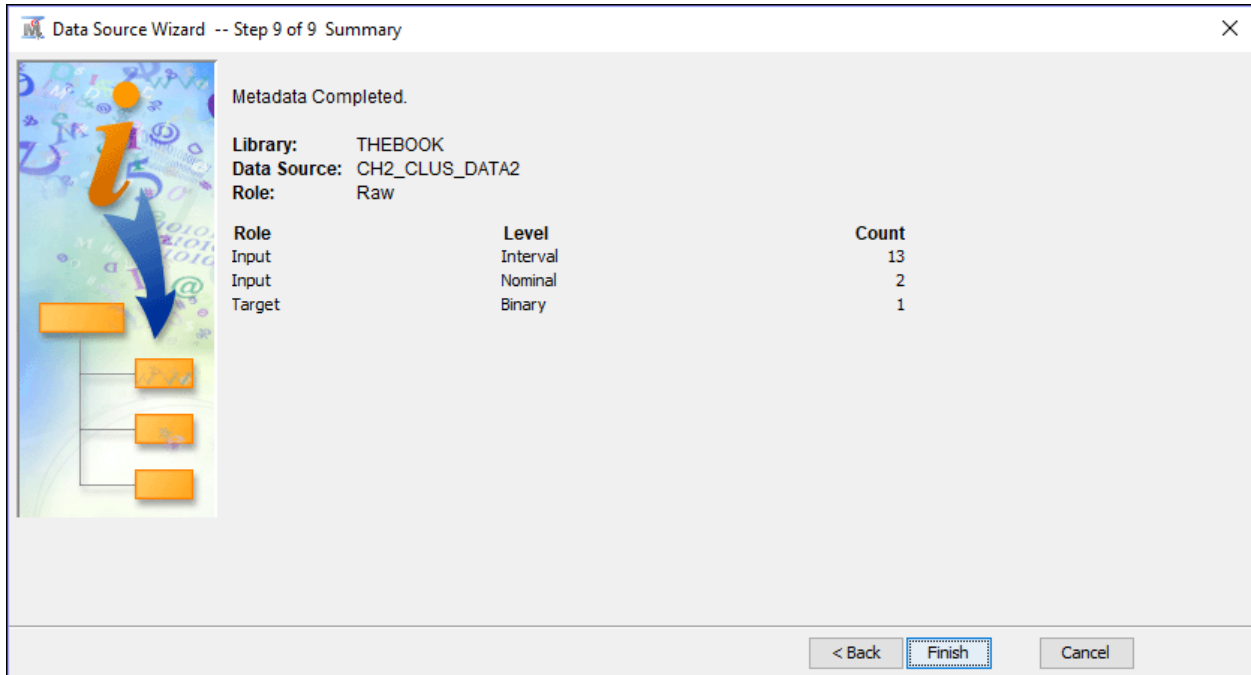
After you set the “Class Levels Count Threshold” property to “3”, you must enter and then click “OK”.

Display 2.24a



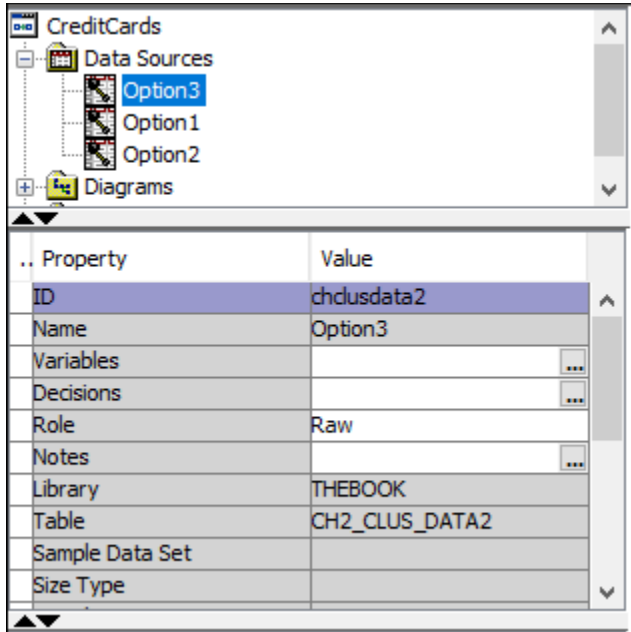
(4f)

Display 2.25



(4g) The data set is renamed as Option 3.

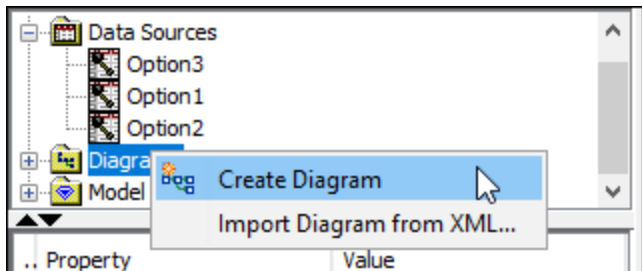
Display 2.25A



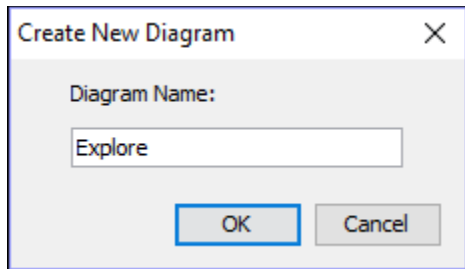
The data source "Option2" is used in exercises 5-8.

Exercise 5

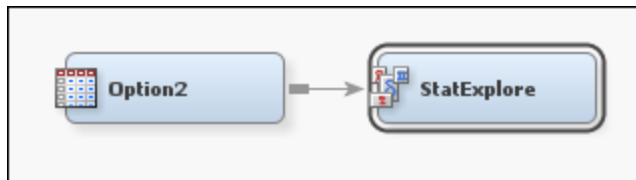
Display 2.26



Display 2.27



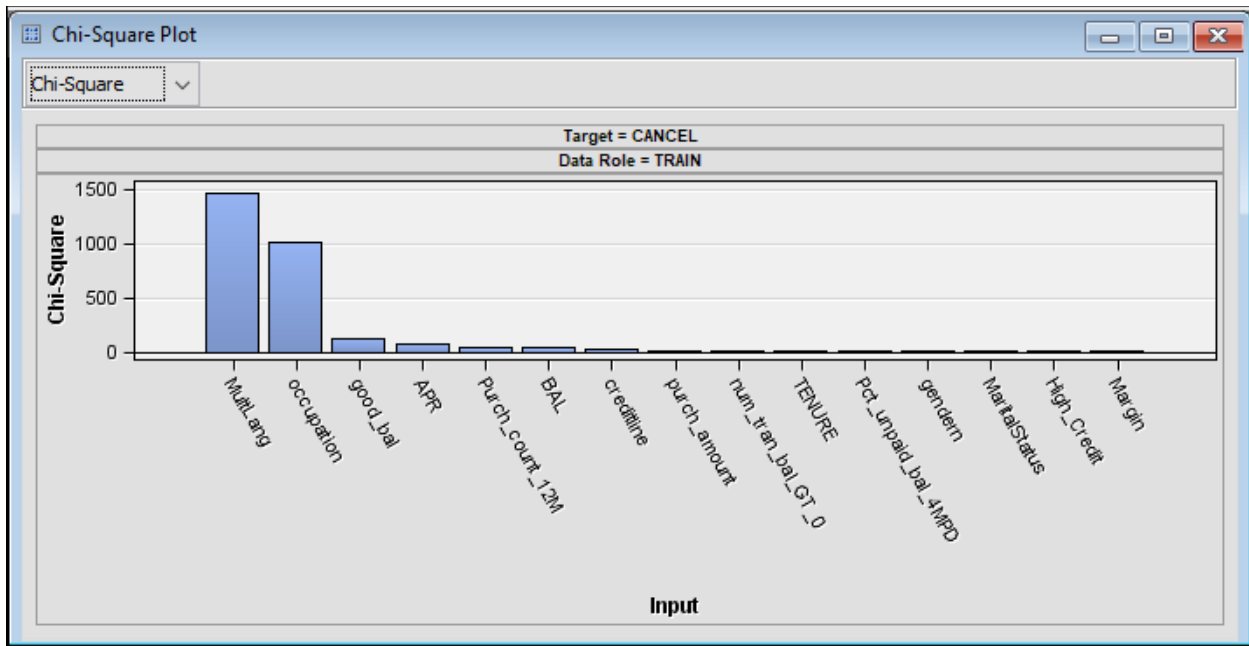
Display 2.28



Display 2.29

.. Property	Value
<b>General</b>	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
[-] Data	
Number of Observations	100000
Validation	No
Test	No
[-] Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	...
[-] Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
[-] Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	Yes
Number of Bins	5
[-] Correlation Statistics	
<b>Interval Variables</b>	
Generates Chi-Square statistics for interval variables by binning the variables.	

Display 2.30



Based on Chi-square value, the top three inputs are :

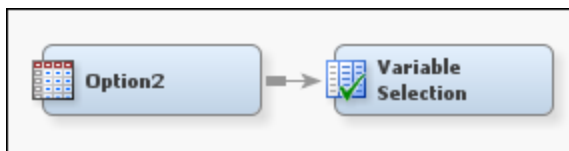
MultiLang: Indicates if the customer speaks multiple language

Occupation

Good\_bal (balance without delinquency). The amount of balance, which is not overdue.

### Exercise 6

Display 2.31



Display 2.32

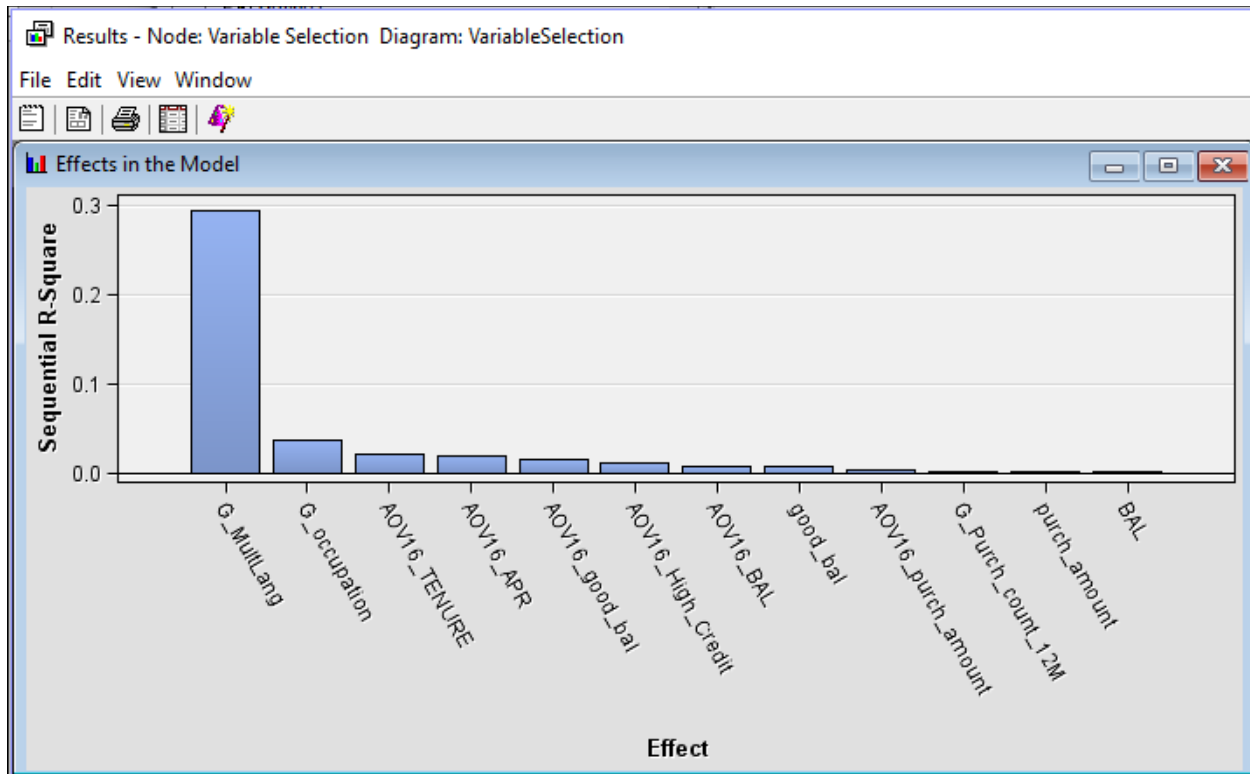
Property	Value
<b>General</b>	
Node ID	Varsel
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	R-Square
Manual Selector	...
Rejects Unused Input	Yes
<b>Bypass Options</b>	
Variable	None
Role	Input
<b>Chi-Square Options</b>	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
<b>R-Square Options</b>	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV16 Variables	Yes
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default
<b>Score</b>	
Hides Rejected Variables	Yes
Hides Unused Variables	Yes
<b>Status</b>	
Create Time	12/6/17 2:28 PM

**Use AOV16 Variables**

This option bins interval variables into 16 equally-spaced groups to help identify non-linear relationships with the target.

Display 2.33



Based on R-Square Criterion the top 5 variables are:

G\_MultiLang : Grouped form of the MultiLang variable, where some of the levels are combined.

G\_Occupation: Grouped form of the occupation variable , where some of the occupations combined

AOV16\_Tenure : The Tenure variable is binned into 16 groups. Tenure is the number of months a customer stayed with the credit card company.

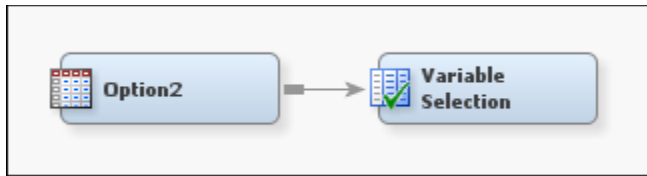
AOV16\_APR : APR is the annual percentage rate. This is the interest rate charged on the unpaid balance. The binned version of this variable is called AOV16\_APR

AOV16\_good\_bal : Balance outstanding without any part past due, is called good balance. This variable is binned and the binned version is called AOV16\_good\_bal.

Note: The AOV16 variables and Grouped (“G”) variables are discussed in the text. Please review the discussion.

Exercise 7

Display 2.34



Display 2.35

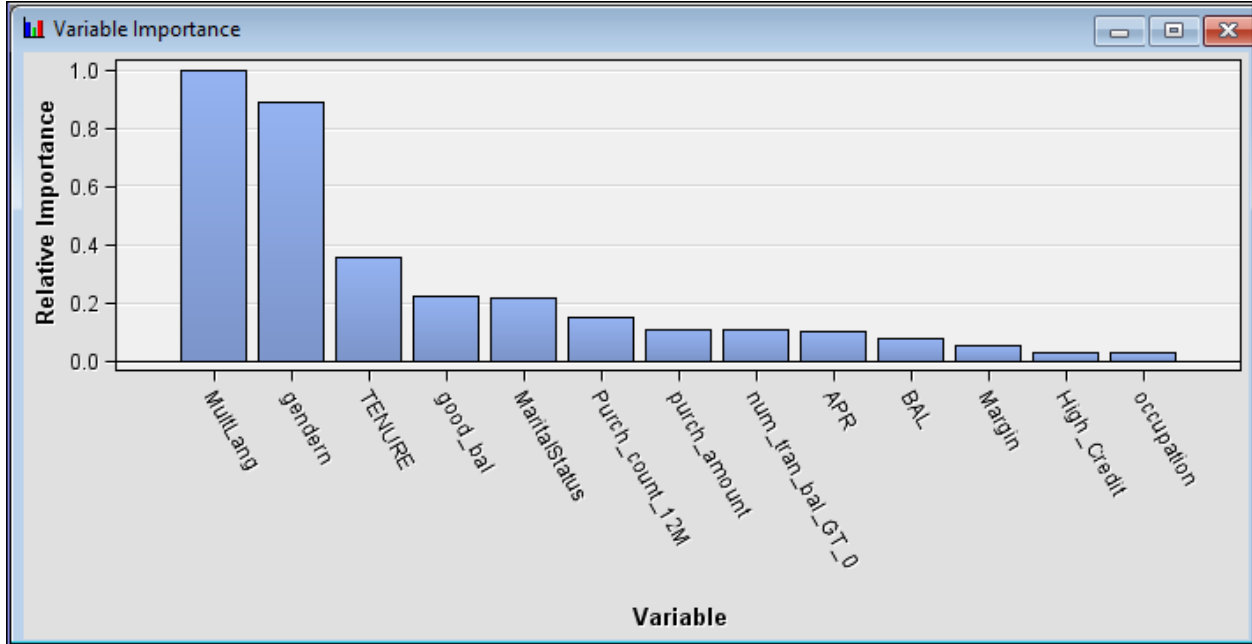
.. Property	Value
<b>General</b>	
Node ID	Varsel
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	Chi-Square
Manual Selector	...
Rejects Unused Input	Yes
<input type="checkbox"/> Bypass Options	
Variable	None
Role	Input
<input type="checkbox"/> Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
<input type="checkbox"/> R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV16 Variables	No
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default
<b>Score</b>	
Hides Rejected Variables	Yes
Hides Unused Variables	Yes
<b>Status</b>	
Create Time	12/6/17 2:40 PM

**Target Model**  
Specifies a variable selection model associated with targets.



When you set the Target Model Property to Chi-Square , some of the R-Square Options are not available. Hence, you cannot set the “Use AOV16 Variables” property to “Yes”, because it is set to “No” by default, when the Target Model Property is set to “Chi-Square”.

Display 2.36

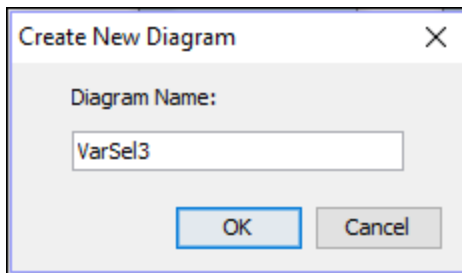


The top 5 variables are:

MultLang, Gendern (an indicator of gender), Tenure (explained above), good\_bal (explained above) and MaritalStatus.

Exercise 8

Display 2.37



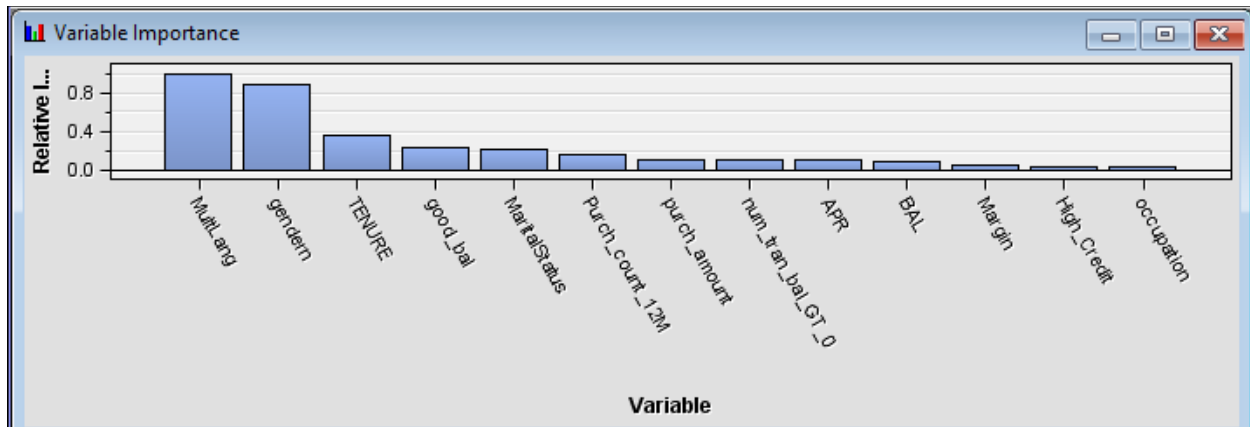
Display 2.38



Display 2.39

Property	Value
<b>General</b>	
Node ID	Varsel
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	R and Chi-square
Manual Selector	...
Rejects Unused Input	Yes
<input type="checkbox"/> Bypass Options	
Variable	None
Role	Input
<input type="checkbox"/> Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
<input type="checkbox"/> R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV 16 Variables	Yes
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default
<b>Score</b>	
Hides Rejected Variables	Yes
Hides Unused Variables	Yes
<b>Status</b>	
Create Time	12/6/17 2:51 PM

Display 2.40



The top 5 variables are:

MultLang, Gendern (an indicator of gender), Tenure (explained above), good\_bal (explained above) and MaritalStatus.

## Chapter 3

Create a new project called “Ch3\_Solutions”

Display 3.1

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name  
Ch3\_Solutions

SAS Server Directory  
C:\TheBook\EM14.3\EMProjects\Chapter3

< Back    Next >    Cancel

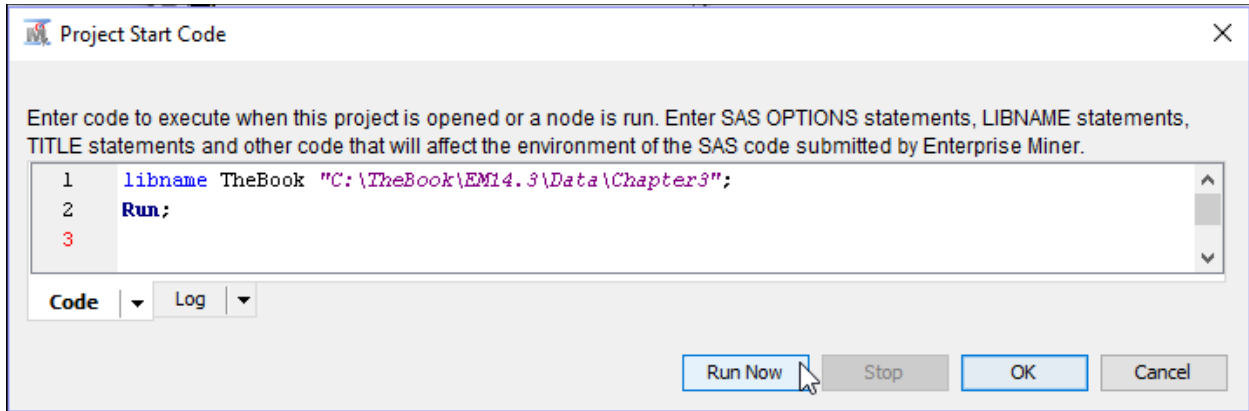
Display 3.2

New Project Information	
Name	Ch3_Solutions
Server Directory	C:\TheBook\EM14.3\EMProjects\Chapter3

< Back    Finish    Cancel

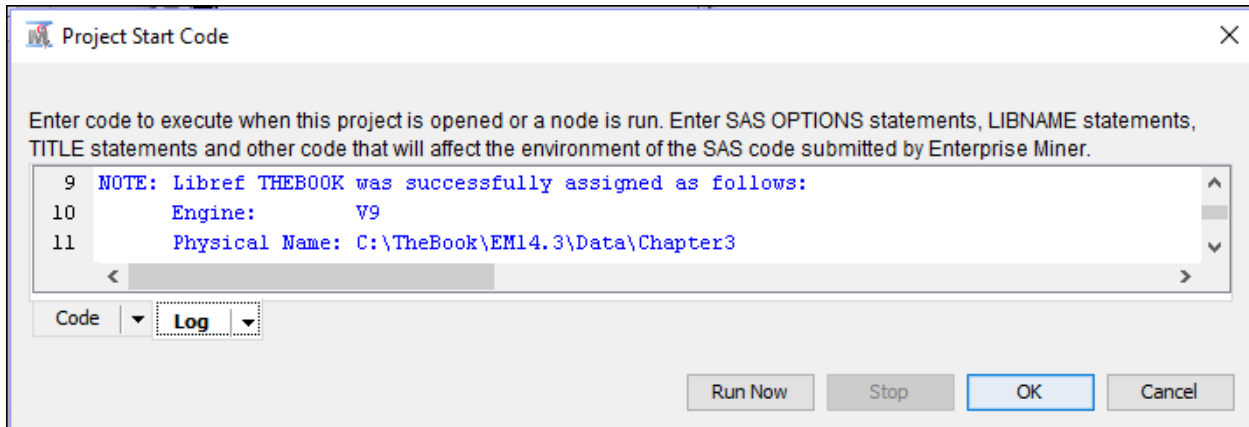
Create a libref by entering the libname statement in the Project Start Code window, and click “Run Now”.

Display 3.3



Open the log window and make sure that the libref is successfully assigned. Then click on "OK".

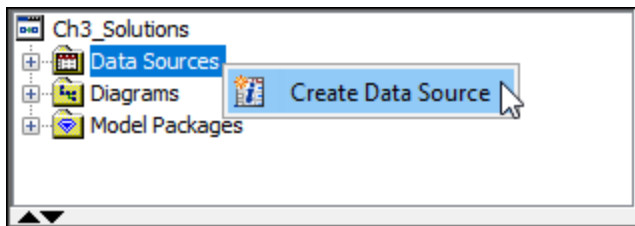
Display 3.4



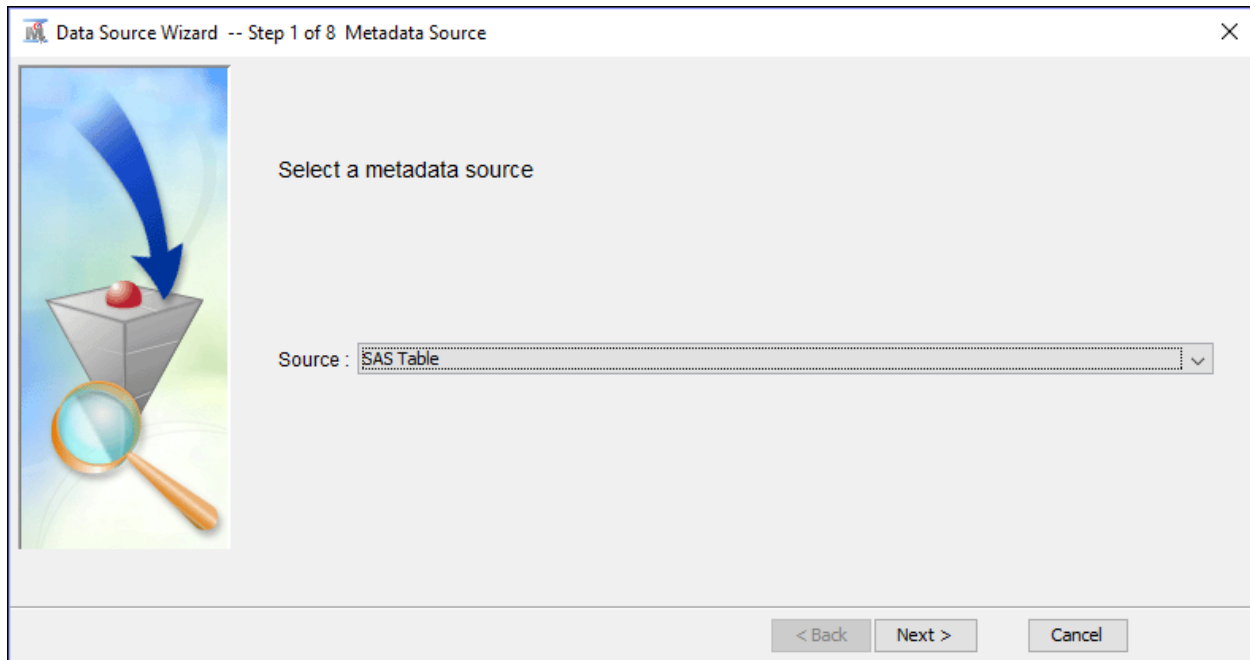
### Exercise 1

Create a data source using the data source wizard.

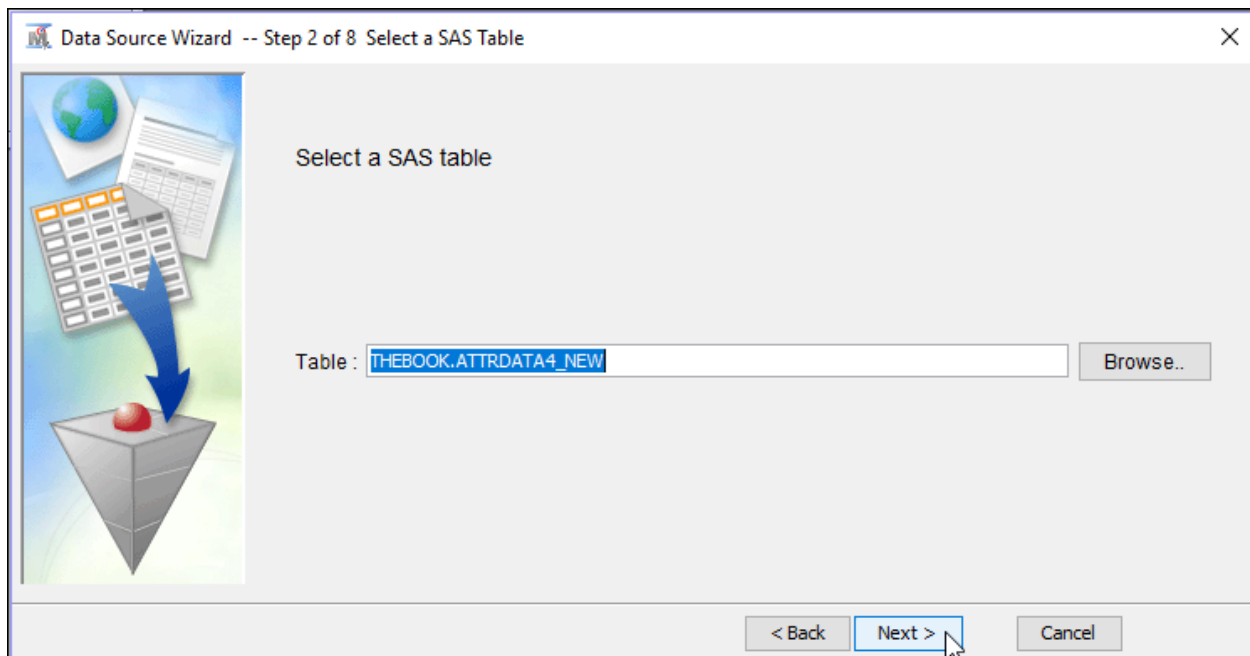
Display 3.5



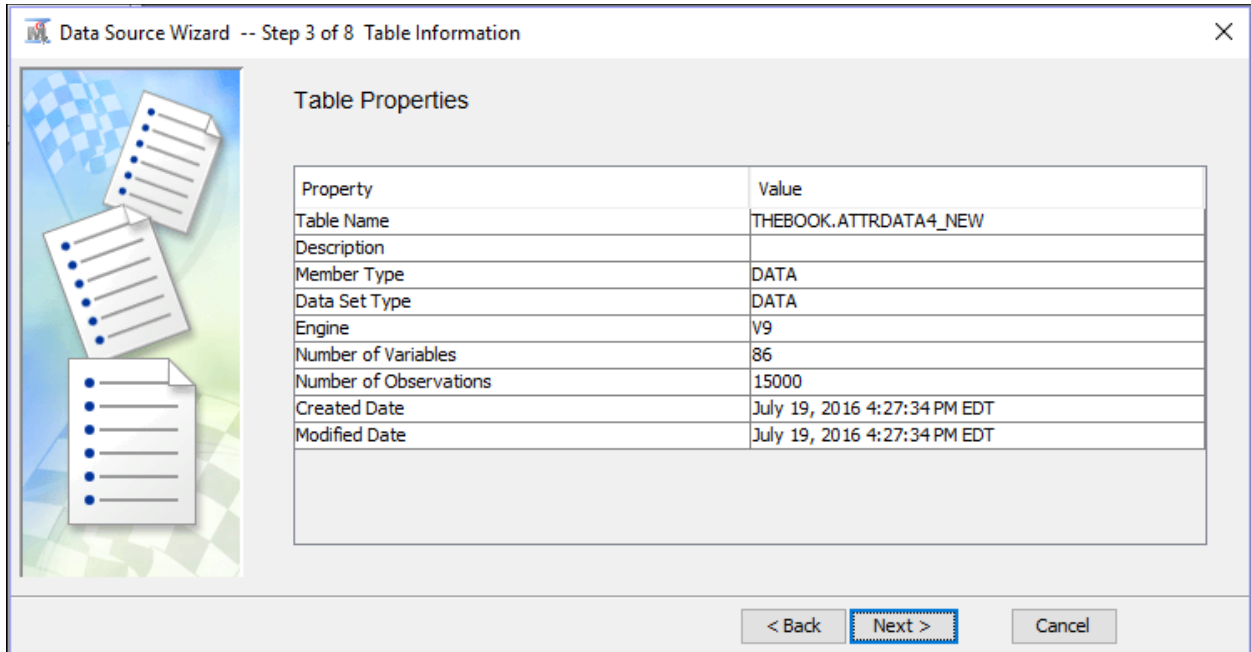
Display 3.6



Display 3.7

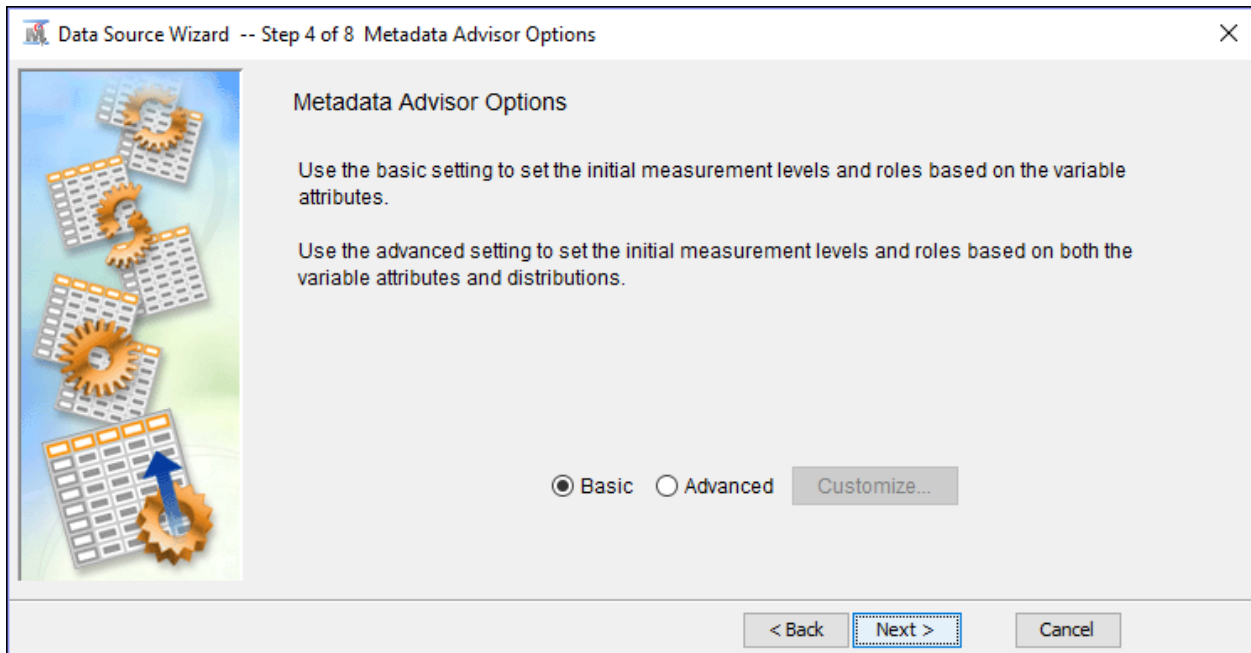


Display 3.8



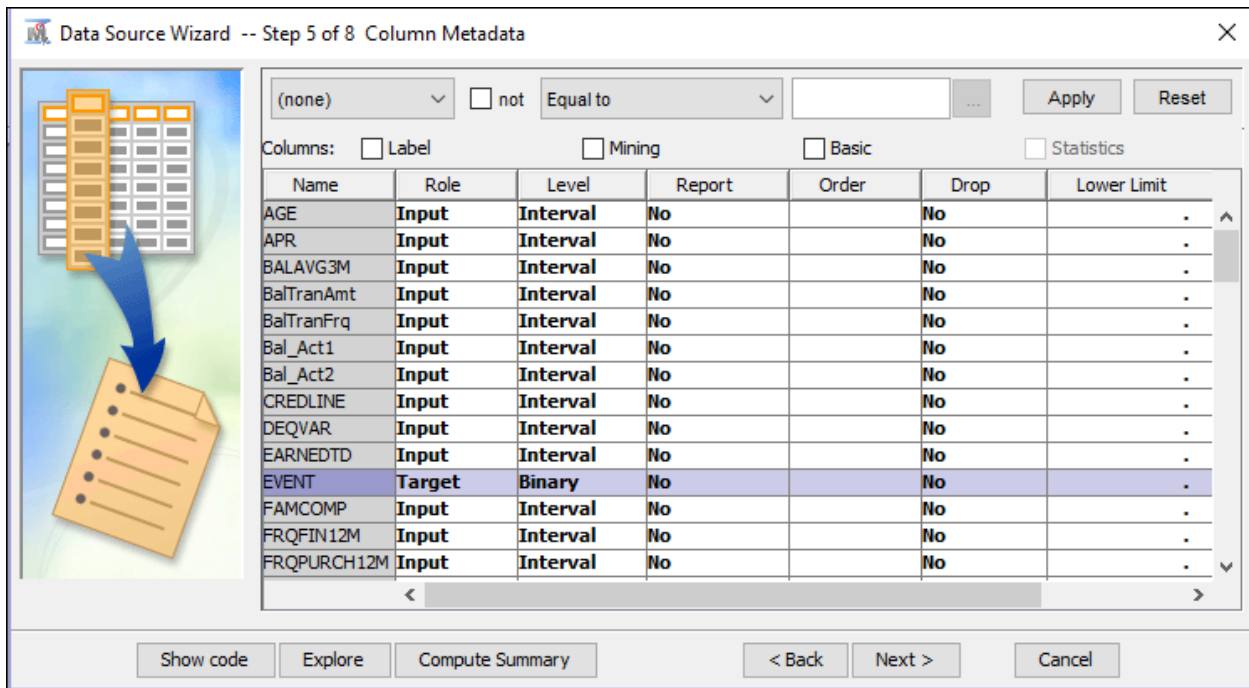
Exercise 2

Display 3.9



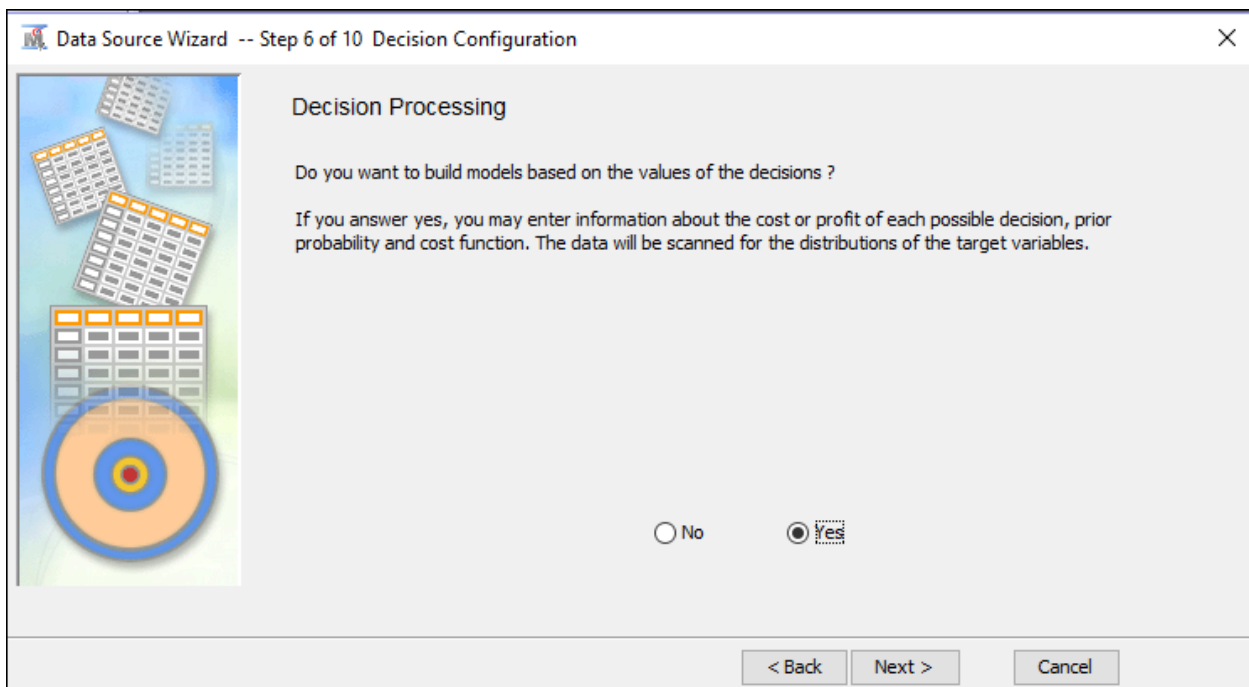
### Exercise 3

#### Display 3.10



The target variable “EVENT” can be considered a “default”. Its measurement level is set to “Binary” because it takes the values 1 and 0 : 1 if the customer defaults, and 0 if the customer does not default in the observation window.

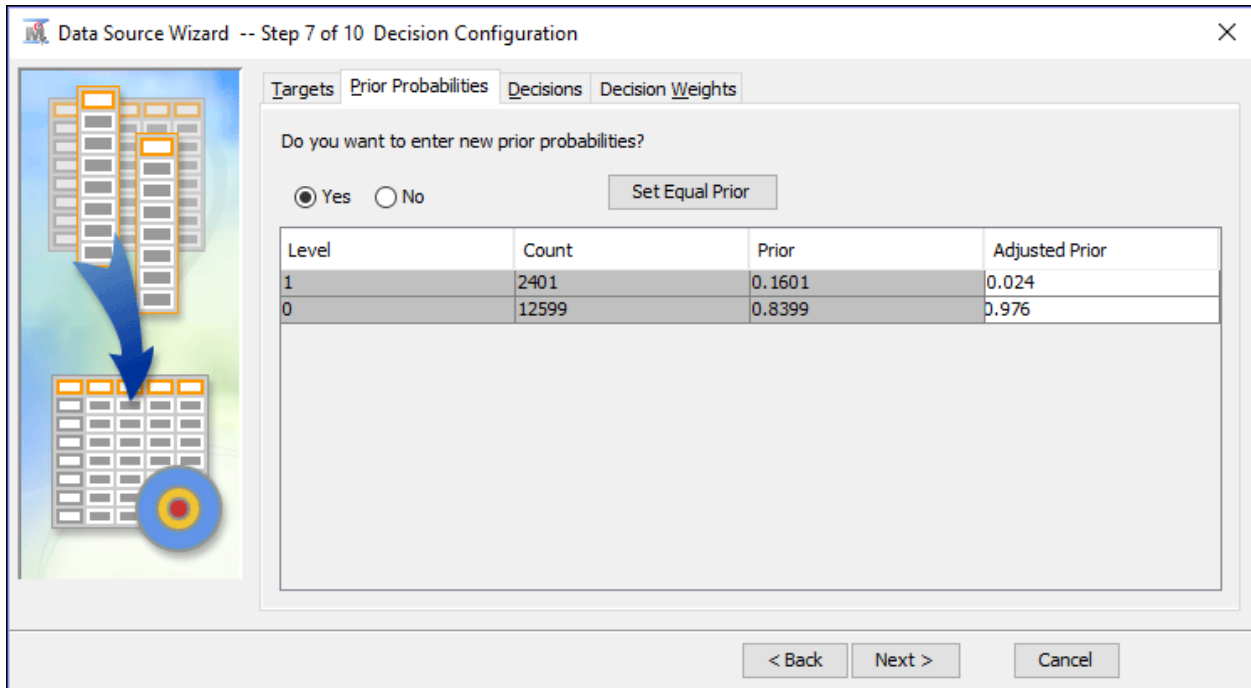
#### Display 3.11



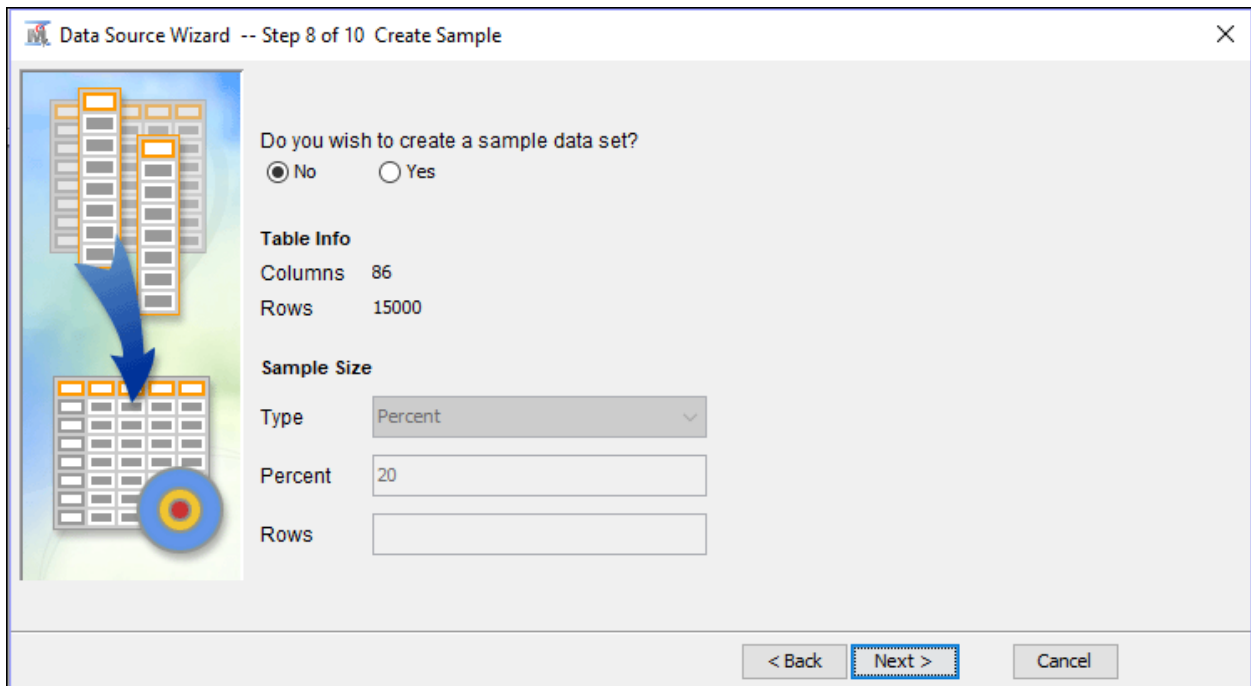


Exercise 4

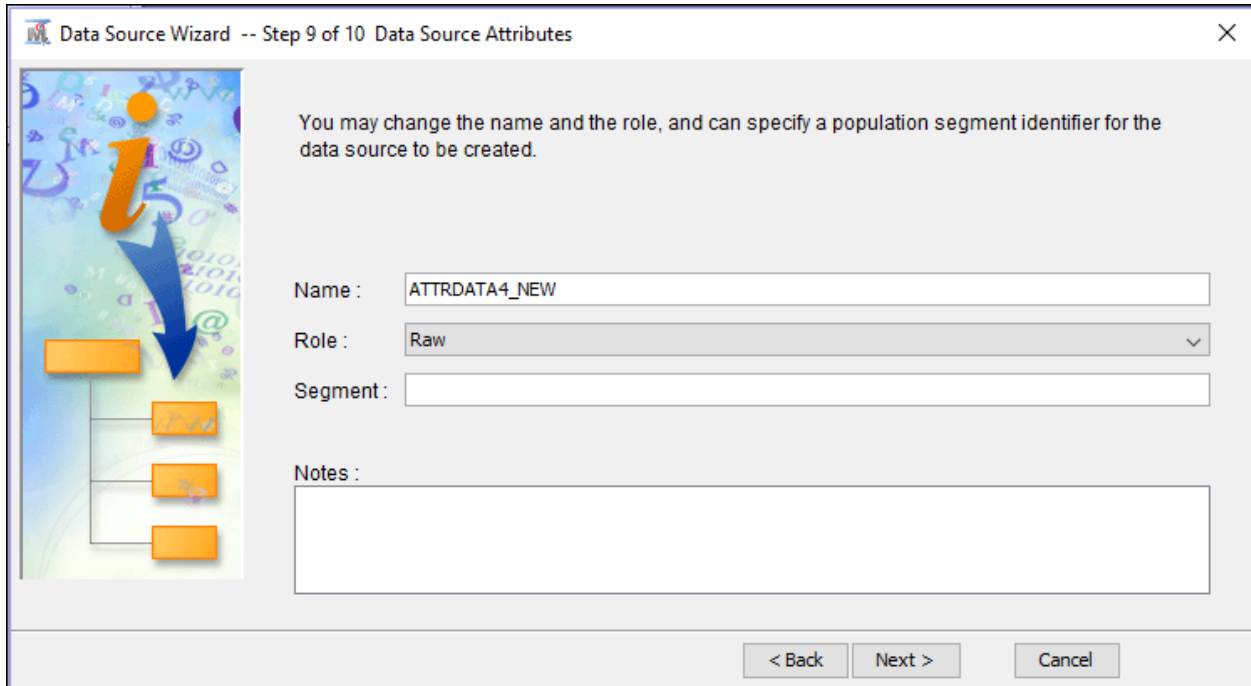
Display 3.12



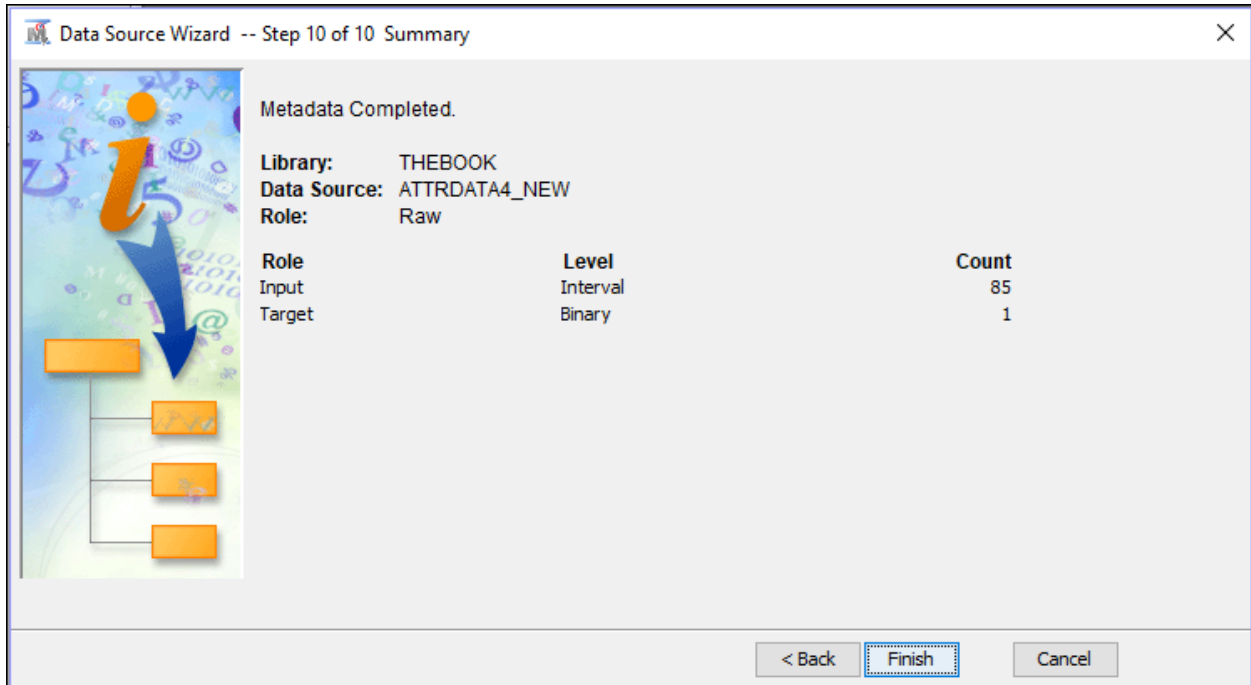
Display 3.13



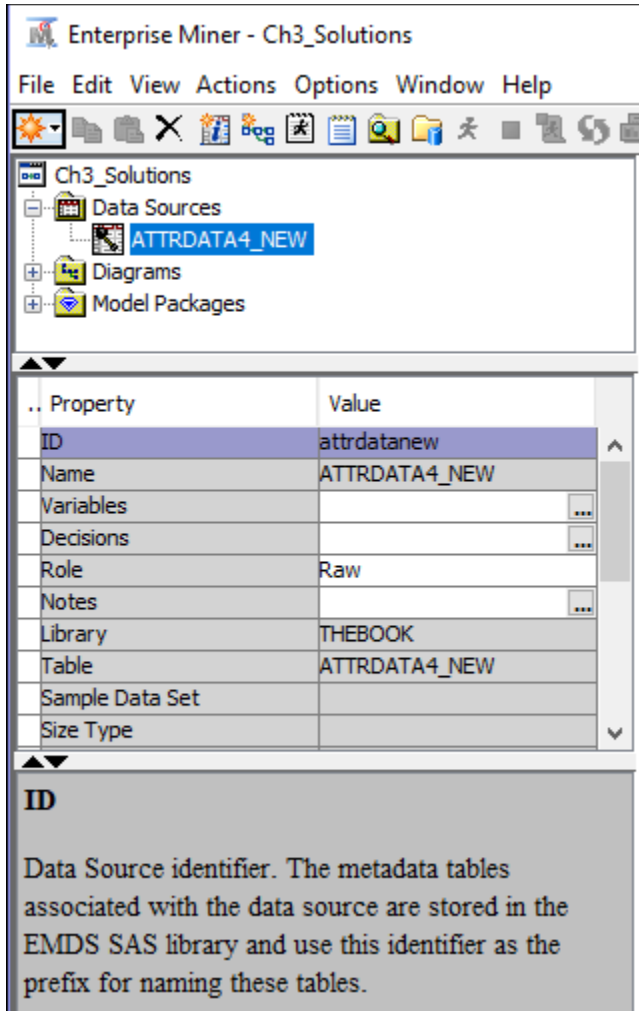
Display 3.14



Display 3.15

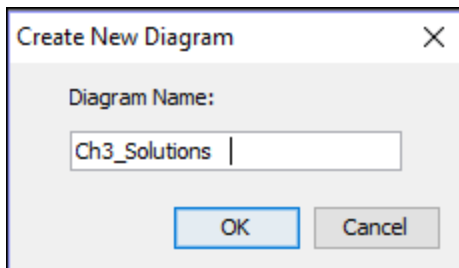


Display 3.16



Exercise 5

Display 3.17



Display 3.18

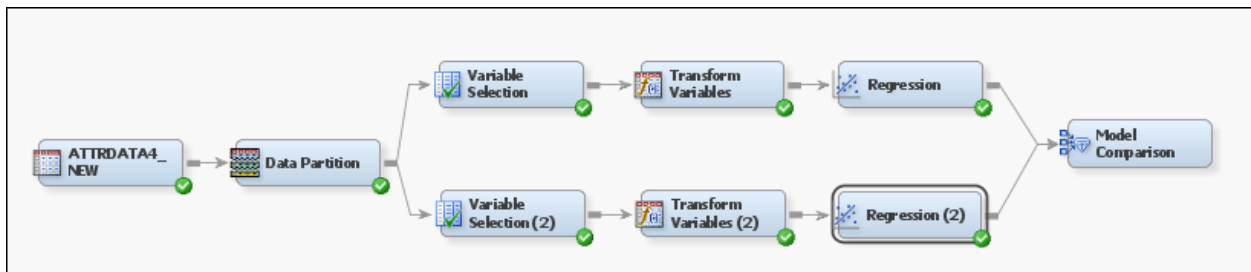
The screenshot shows the SAS Enterprise Miner interface. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. The main workspace is divided into two panes. The left pane shows a tree view with folders for Ch3\_Solutions, Data Sources (containing ATTRDATA4\_NEW), Diagrams (containing Ch3\_Solutions), and Model Packages. The right pane displays the properties of the selected diagram, Ch3\_Solutions.

Property	Value
ID	EMWS1
Name	Ch3_Solutions
Status	Open
Notes	
History	
Create Date	12/9/17 4:24 AM
Encoding	wlatin1 Western (Windows)
Data Representation	WINDOWS_32
Native OS	Yes

**ID**

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Display 3.19



## Data Partition

### Display 3.20

Property	Value
<b>General</b>	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<input type="checkbox"/> Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes

Logistic Regression Estimated from the upper segment of the Process Flow in Display 3.19

### Display 3.21

```

The selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:
Intercept OPT_GOODBAL OPT_L2BNKENQ OPT_MKTVAL OPT_NUMRPD6H OPT_OpenCardsPct OPT_PPSCORE8 OPT_SALESFRQ OPT_TOTBAL

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood      Likelihood
Intercept      Intercept &      Ratio
Only      Covariates      Chi-Square      DF      Pr > ChiSq

6598.713      4100.731      2497.9814      14      <.0001

Type 3 Analysis of Effects

Effect      DF      Wald
Chi-Square      Pr > ChiSq

OPT_GOODBAL      1      173.7686      <.0001
OPT_L2BNKENQ      3      66.1808      <.0001
OPT_MKTVAL      3      726.5914      <.0001
OPT_NUMRPD6H      1      25.9232      <.0001
OPT_OpenCardsPct      2      54.3740      <.0001
OPT_PPSCORE8      2      76.3478      <.0001
OPT_SALESFRQ      1      22.7529      <.0001
OPT_TOTBAL      1      14.7030      0.0001
    
```

Display 3.22 shows the logistic regression estimated from the lower segment of the Process Flow in Display 3.19.

### Display 3.22

The selected model, based on the error rate for the validation data, is the model trained in Step 13. It consists of the following effects:

Intercept OPT\_MKTVAL TI\_AOV16\_APRS TI\_AOV16\_GOODBALL1 TI\_AOV16\_MAXCRED3 TI\_AOV16\_NUMTOTOPN2 TI\_AOV16\_PCTOPNTTOTR1 TI\_AOV16\_PPSCORE813 TI\_AOV16\_PPSCORE814

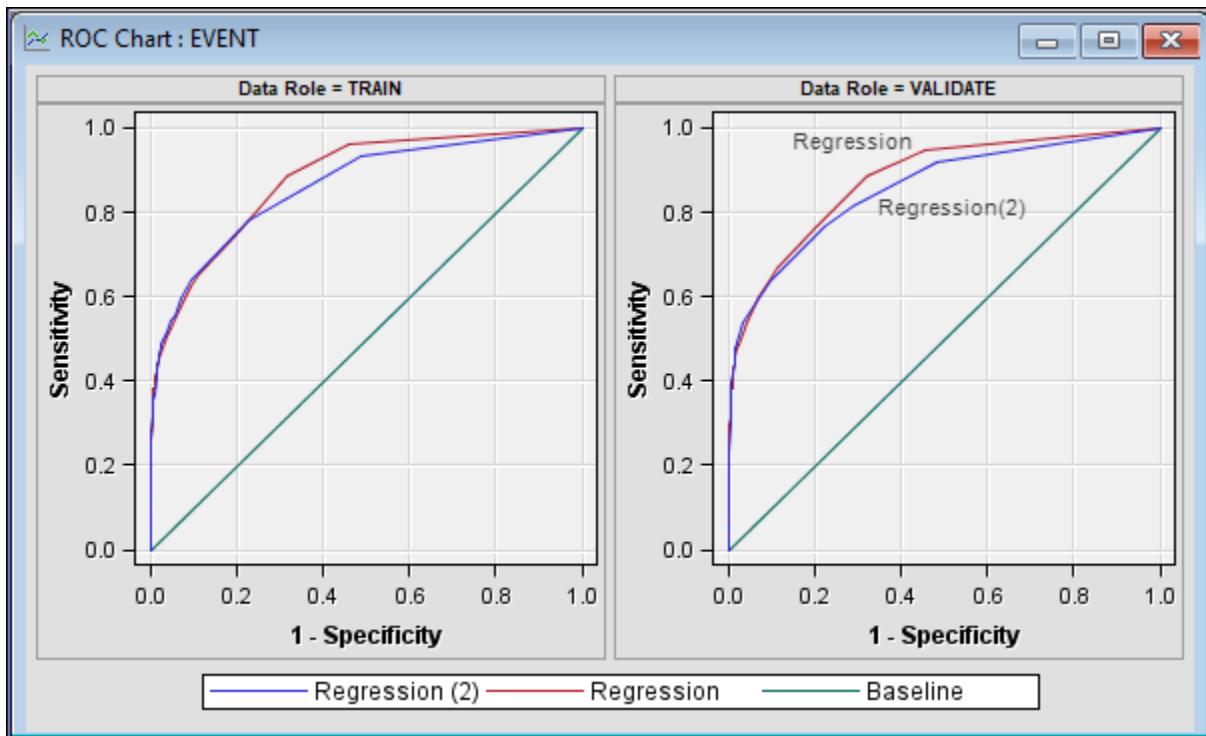
Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
6598.713	4218.476	2380.2369	15	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
OPT_MKTVAL	3	793.9265	<.0001
TI_AOV16_APRS	1	111.9575	<.0001
TI_AOV16_GOODBALL1	1	116.2403	<.0001
TI_AOV16_MAXCRED3	1	22.4711	<.0001
TI_AOV16_NUMTOTOPN2	1	26.2170	<.0001
TI_AOV16_PCTOPNTTOTR1	1	42.4707	<.0001
TI_AOV16_PPSCORE813	1	96.6912	<.0001
TI_AOV16_PPSCORE814	1	62.9732	<.0001
TI_AOV16_SALESFRQ1	1	84.4505	<.0001
TI_AOV16_TENURE16	1	18.3166	<.0001
TI_AOV16_TENURE2	1	68.1567	<.0001
TI_AOV16_TOTBAL3	1	30.6351	<.0001
TI_AOV16_TOTFC1	1	50.2309	<.0001

### Display 3.23



From the ROC charts based on the validation data set in Display 3.23, the Regression from the upper segment of the process flow performed better than the Regression (2) in the lower segment of the

process flow. This shows that the optimal binning (by transformation node in the upper segment) yields better equation than the AOV16 variables used in the lower segment.

Bin	Percentile	Number of responders	Number of non-responders	Number of observations	%Response	Cumulative %Response	Cumulative responses	Cumulative %captured response	Lift	Cumulative Lift	Cum Resp	Cum Obs	
1	5	95	280	375	25.3%	25.3%	95	53.4%	10.7	10.7	95	375	25.3%
2	10	14	361	375	3.7%	14.5%	109	61.2%	1.6	6.1	109	750	14.5%
3	15	12	363	375	3.2%	10.8%	121	68.0%	1.3	4.5	121	1125	10.8%
4	20	9	366	375	2.4%	8.7%	130	73.0%	1.0	3.7	130	1500	8.7%
5	25	9	366	375	2.4%	7.4%	139	78.1%	1.0	3.1	139	1875	7.4%
6	30	8	367	375	2.1%	6.5%	147	82.6%	0.9	2.8	147	2250	6.5%
7	35	8	367	375	2.1%	5.9%	155	87.1%	0.9	2.5	155	2625	5.9%
8	40	6	369	375	1.6%	5.4%	161	90.4%	0.7	2.3	161	3000	5.4%
9	45	5	370	375	1.3%	4.9%	166	93.3%	0.6	2.1	166	3375	4.9%
10	50	5	370	375	1.3%	4.6%	171	96.1%	0.6	1.9	171	3750	4.6%
11	55	3	372	375	0.8%	4.2%	174	97.8%	0.3	1.8	174	4125	4.2%
12	60	1	374	375	0.3%	3.9%	175	98.3%	0.1	1.6	175	4500	3.9%
13	65	0	375	375	0.0%	3.6%	175	98.3%	0.0	1.5	175	4875	3.6%
14	70	1	374	375	0.3%	3.4%	176	98.9%	0.1	1.4	176	5250	3.4%
15	75	1	374	375	0.3%	3.1%	177	99.4%	0.1	1.3	177	5625	3.1%
16	80	0	375	375	0.0%	3.0%	177	99.4%	0.0	1.2	177	6000	3.0%
17	85	1	374	375	0.3%	2.8%	178	100.0%	0.1	1.2	178	6375	2.8%
18	90	0	375	375	0.0%	2.6%	178	100.0%	0.0	1.1	178	6750	2.6%
19	95	0	375	375	0.0%	2.5%	178	100.0%	0.0	1.1	178	7125	2.5%
20	100	0	375	375	0.0%	2.4%	178	100.0%	0.0	1.0	178	7500	2.4%
Total		178											
Average Response		2.4%											

# Chapter 4

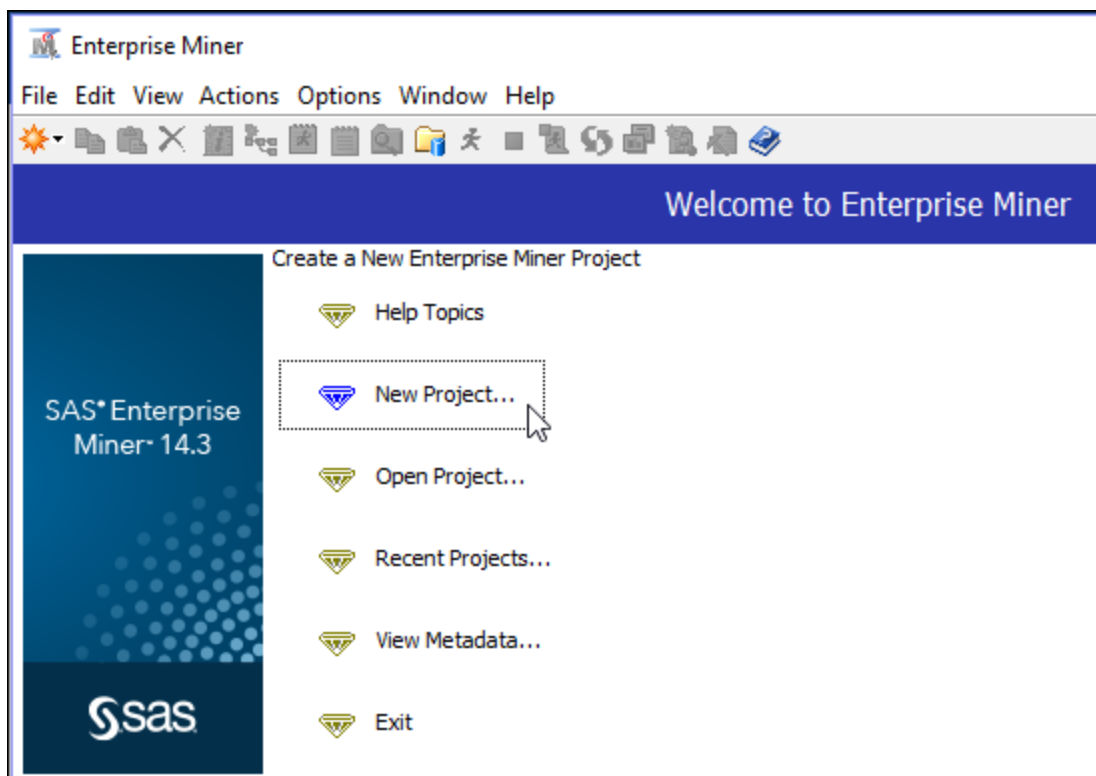
## Section 4.9

Exercise 1: In this exercise we use the data set Ch4\_BookData1.

Let us create the project “Ch4\_Solutions” in the directory “C:\TheBook\EM14.3\EMProjects” to illustrate the solution to Exercise 1. (You can create the project in any directory, using any name for the project. )

1. Open Enterprise Miner by clicking on the Enterprise Miner icon. I am using Enterprise Miner 14.3 for the solutions. You can use any version 13.1 or later. Enterprise Miner window opens as shown in Display 4.1 below:

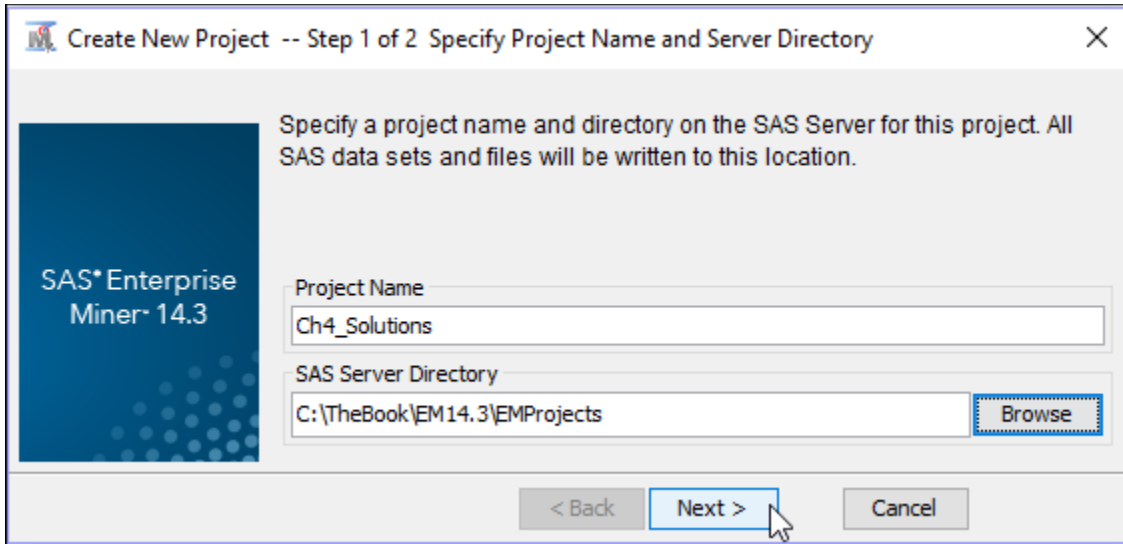
Display 4.1



2. Create a new project by clicking on “New Project” as shown by the mouse-pointer in Display 4.1
3. In the first window that appears you enter the name of the project and the directory where you want to save your project, and click on “Next” as shown in the next directory.

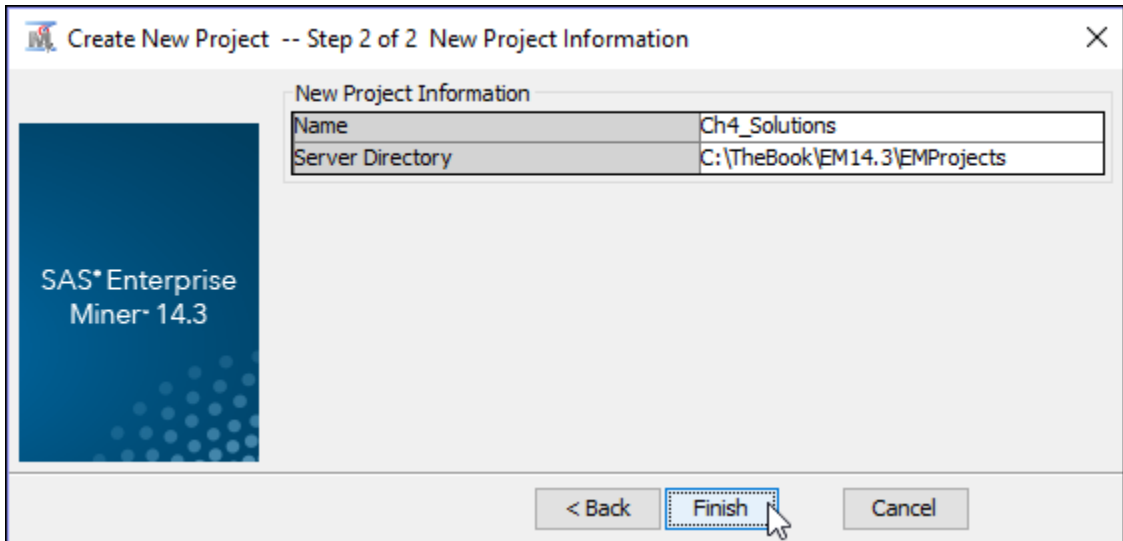


Display 4.2



4. The next window shows what you entered in the previous window- the project name and the directory where the project will be saved.

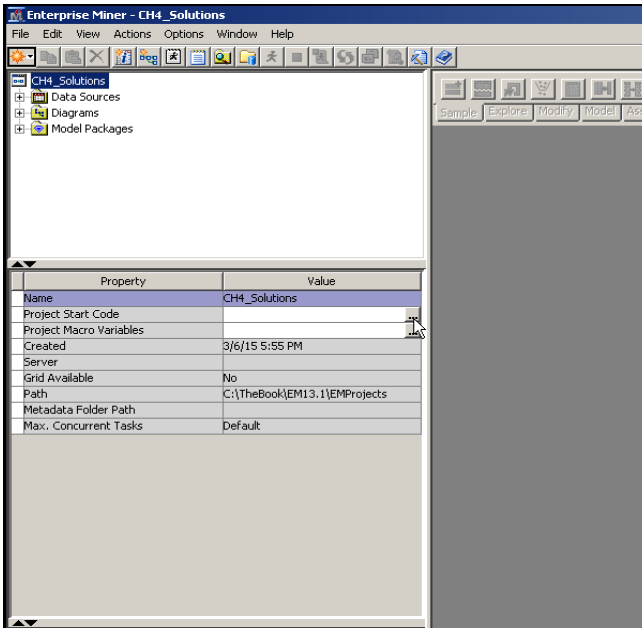
Display 4.3



When you click on the "Finish" button, the project is created and the project window opens.

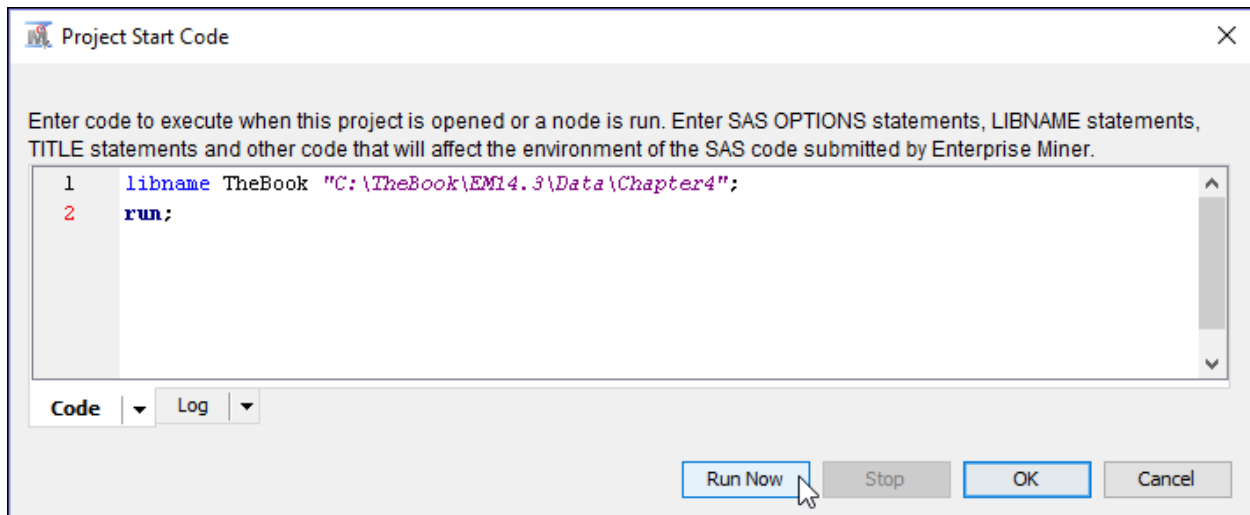
5. Now we have to tell the Enterprise Miner where our data is located. You can do this by creating a libref in Project Start Code.

Display 4.4



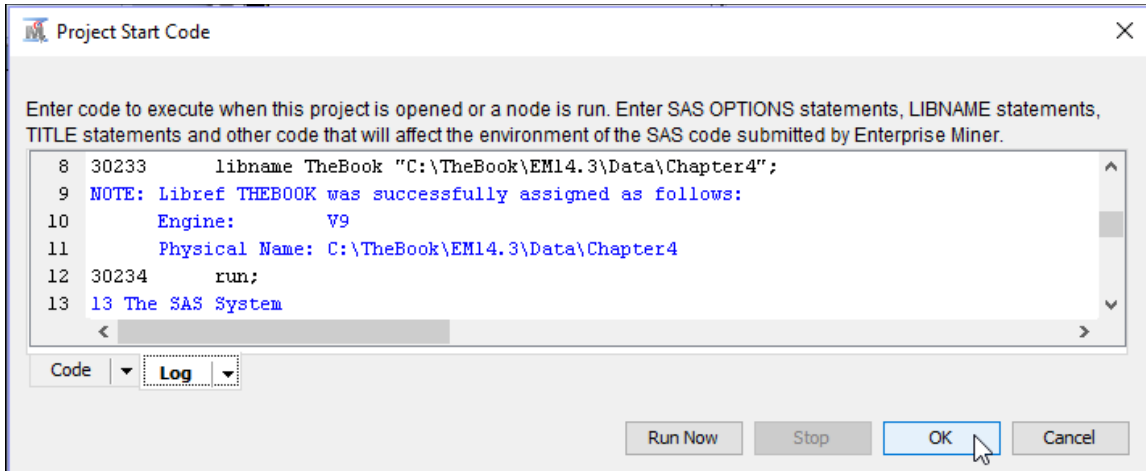
6. Click on the '...' in the value column to the start of "Project Start Code" property as shown in Display 4.4. The Project Start Code window opens.
7. Type the library statement as shown below (Display 4.5)

Display 4.5



Click on "Run Now" button and click on "log" under tab to verify that the libref is created successfully.

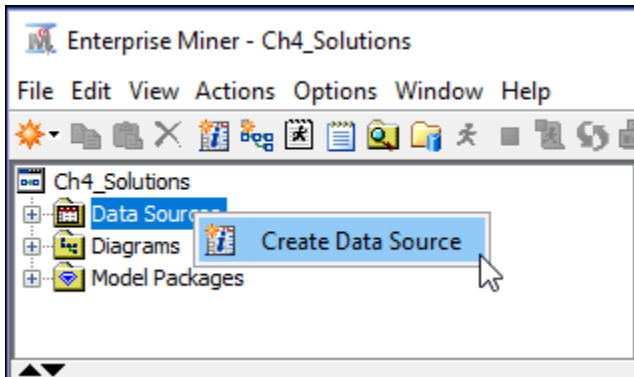
Display 4.6



Click on "OK" to close the window.

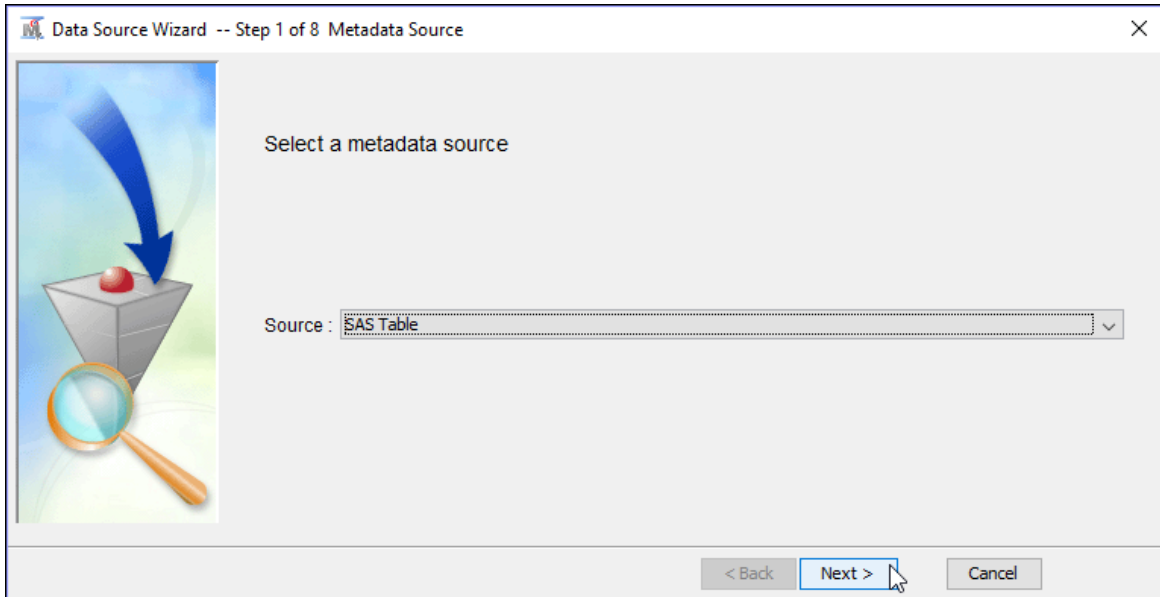
8. The next task is to create a data source. Right-click on "Data Sources" and click on "Create Data Source"

Display 4.7



9. By clicking on "Create Data Source", the Data Source Wizard opens as follows:

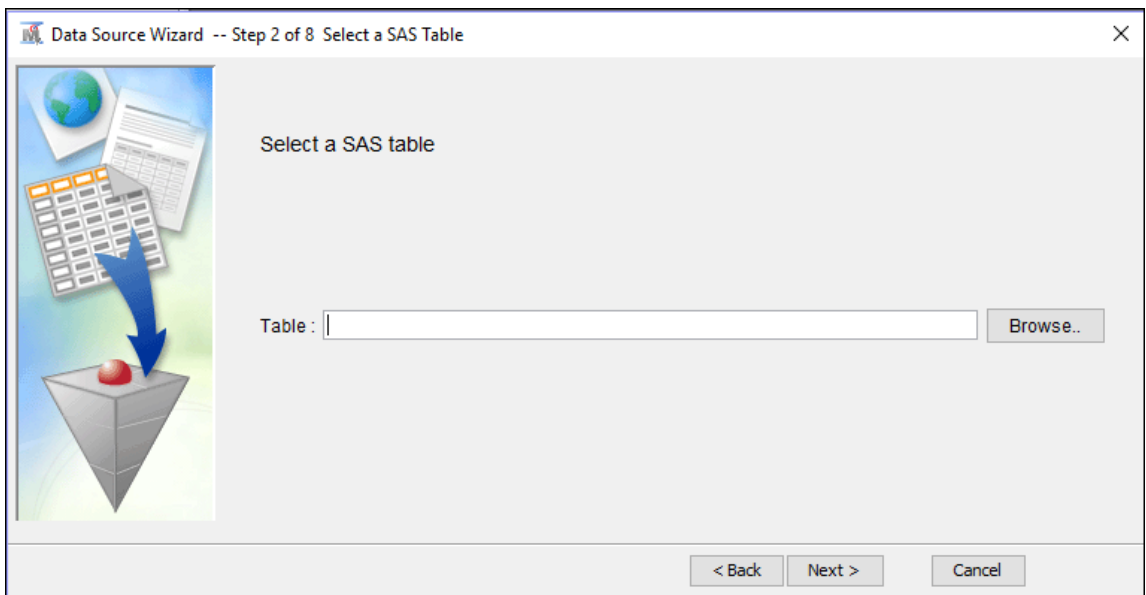
Display 4.8



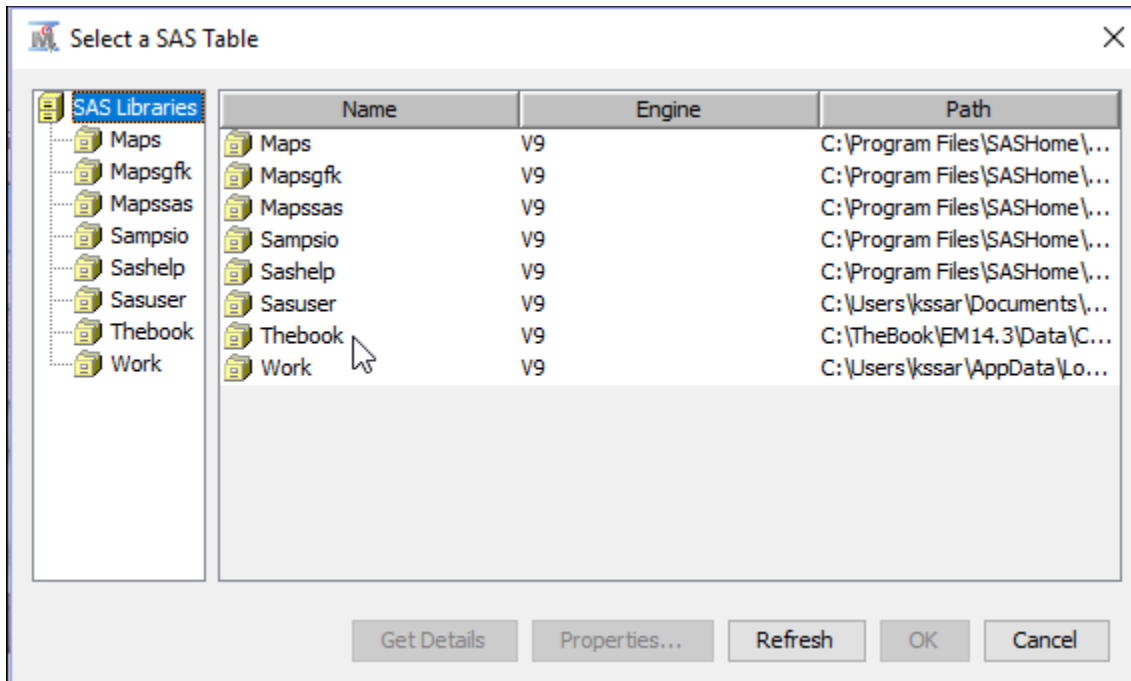
If you are using a SAS data set for this project then the metadata source is a SAS Table. This is already selected by the Enterprise Miner. So click on “Next”. From the display above, you can see that the Metadata Source has 8 steps.

10. In the second step of the Meta Data Source, you select the table. Earlier we created a “libref” for the directory where the SAS table for this project is stored (See display 4.5). To point to the library where the data is stored, we click on the “Browse” button as shown in Display 4.9.

Display 4.9

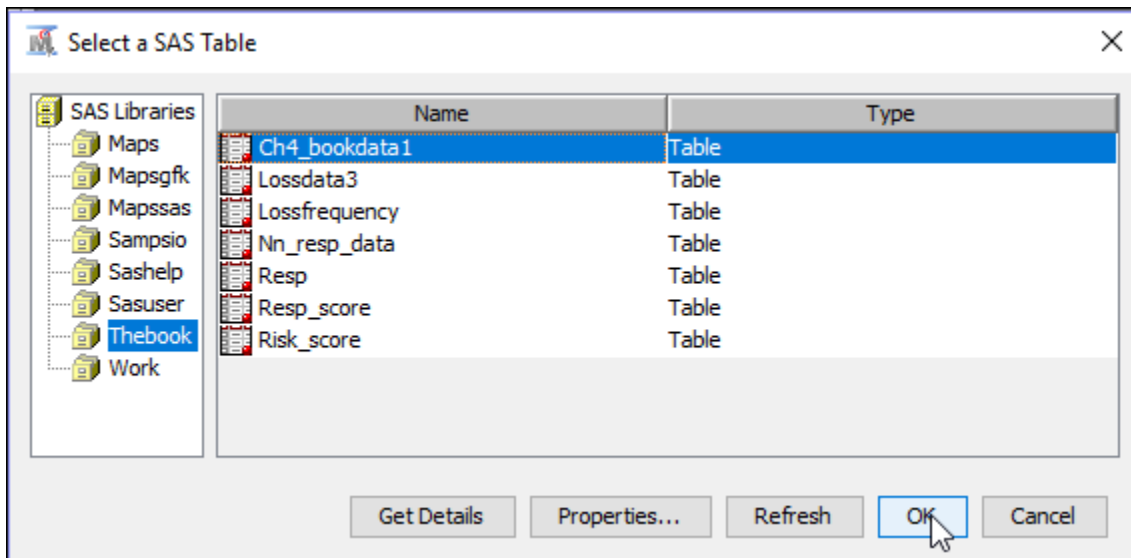


Display 4.10



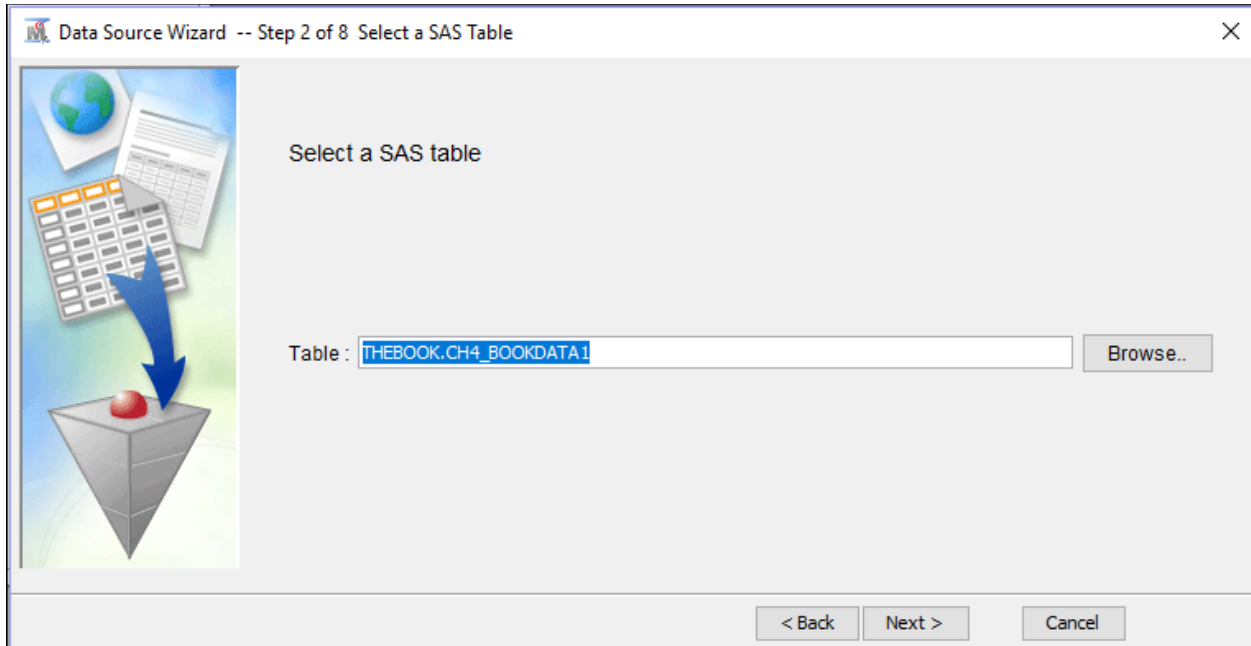
Double click on "TheBook" shown under the names in the "Select a SAS Table" window as shown in the above display. You will get a list of tables in the selected directory (TheBook, in this example). Select the table and click OK. The table we selected is "CH4\_bookdata1".

Display 4.11



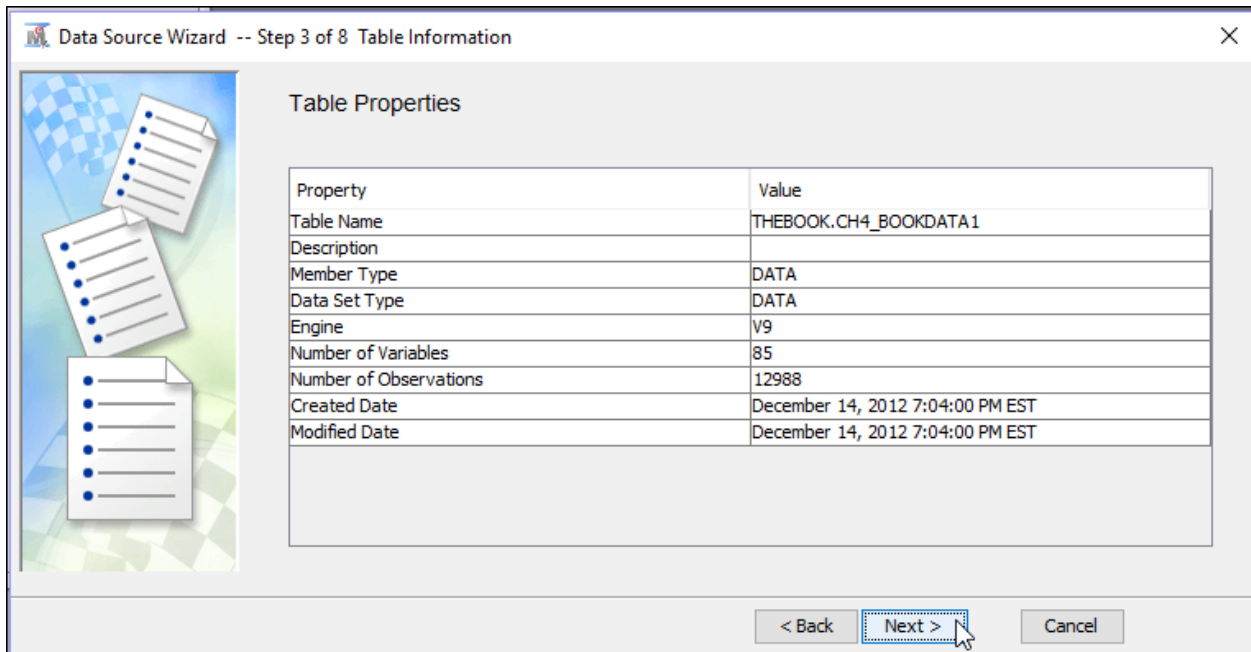
The Data Source Wizard shows the table selected.

Display 4.12



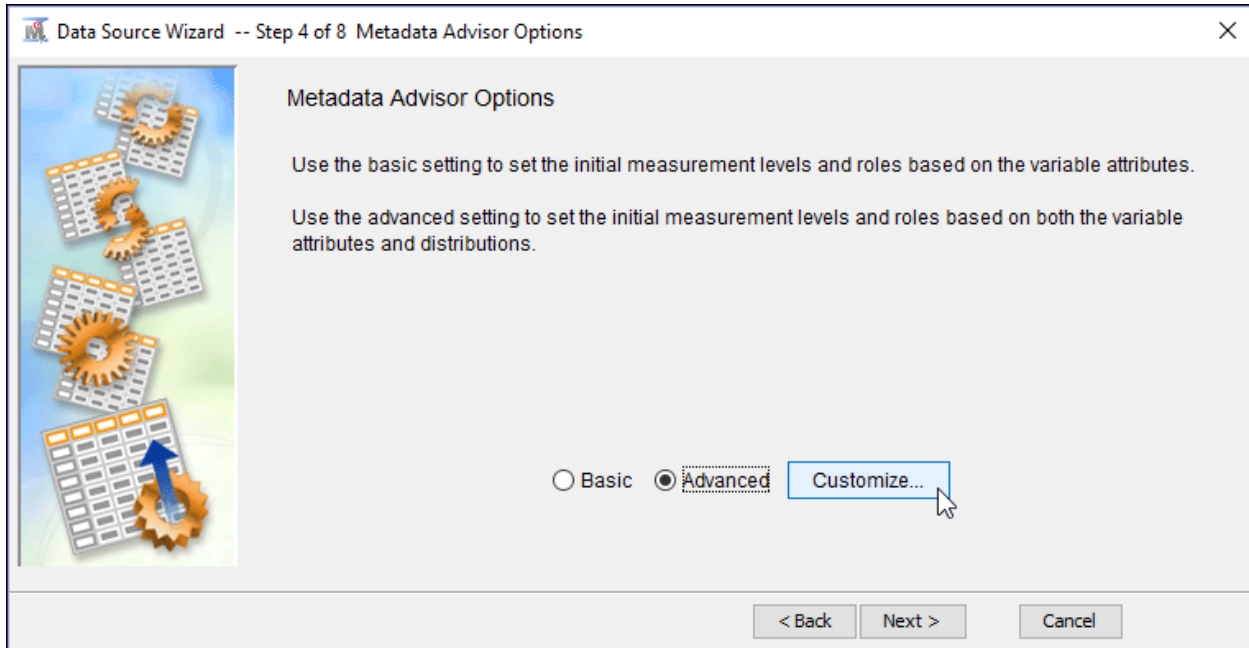
Click on “Next” button. You will see a window showing the properties of the table (SAS data set) that you selected for this data source.

Display 4.13



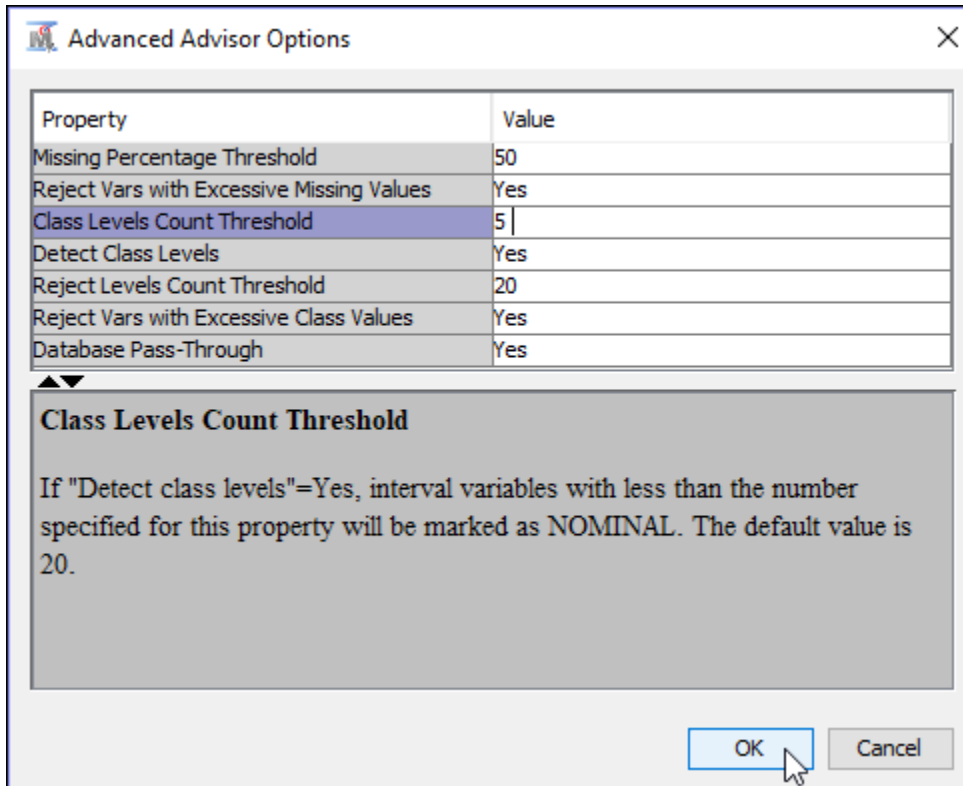
Click “Next”

Display 4.14



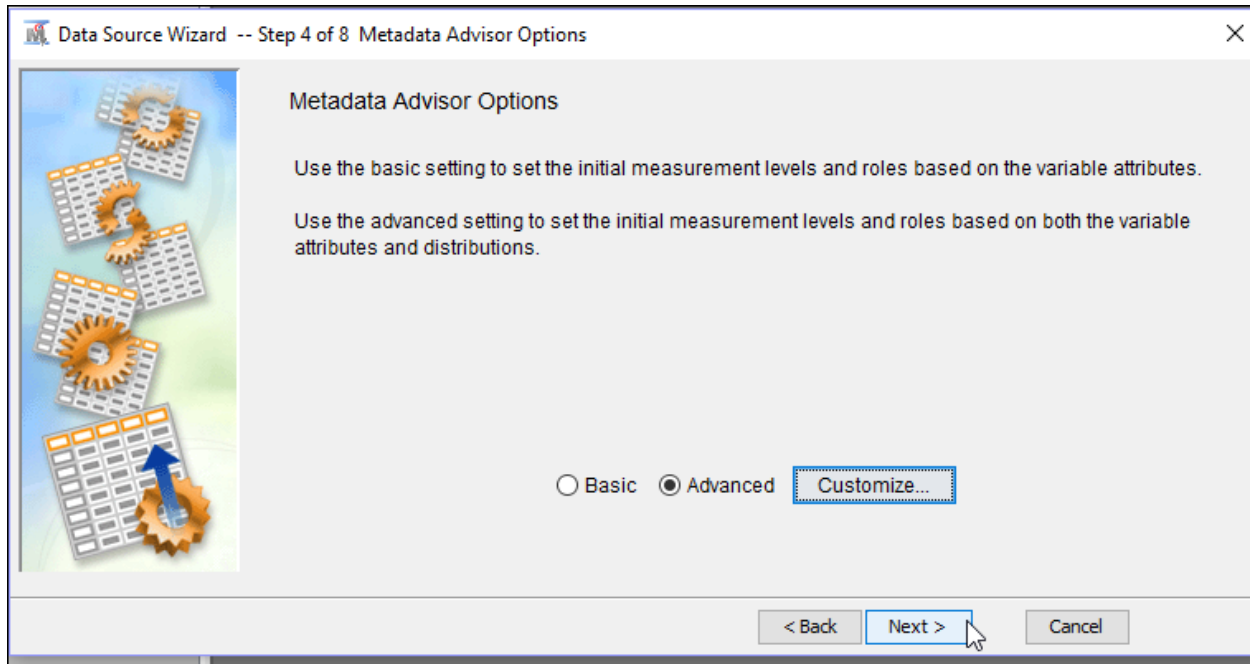
Select “Advanced” and “Customize” option as above. The window for customization opens as shown in Display 4.15.

Display 4.15



Set the Class Levels Count Threshold to 5 and click “OK”. Then the following (Display 4.16) window opens:

Display 4.16

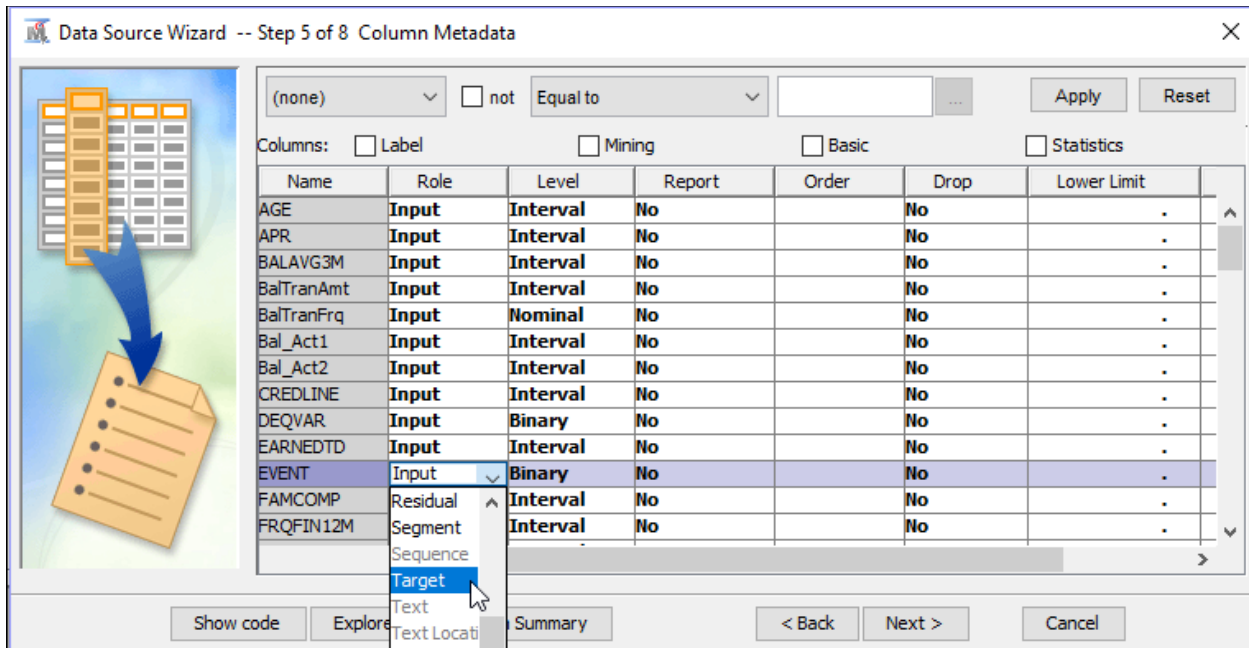


Click “Next”. In the next window, I changed the role of the variable “Event” to “Target”. This is done by clicking on the role column for the variable “Event”. The variable “Event” indicates default. It takes the value 1 if the customer defaults and 0 if he/she does not default in the observation period.

Changing the role of the variable “Event” is shown in Displays 4.17 and 4.18.

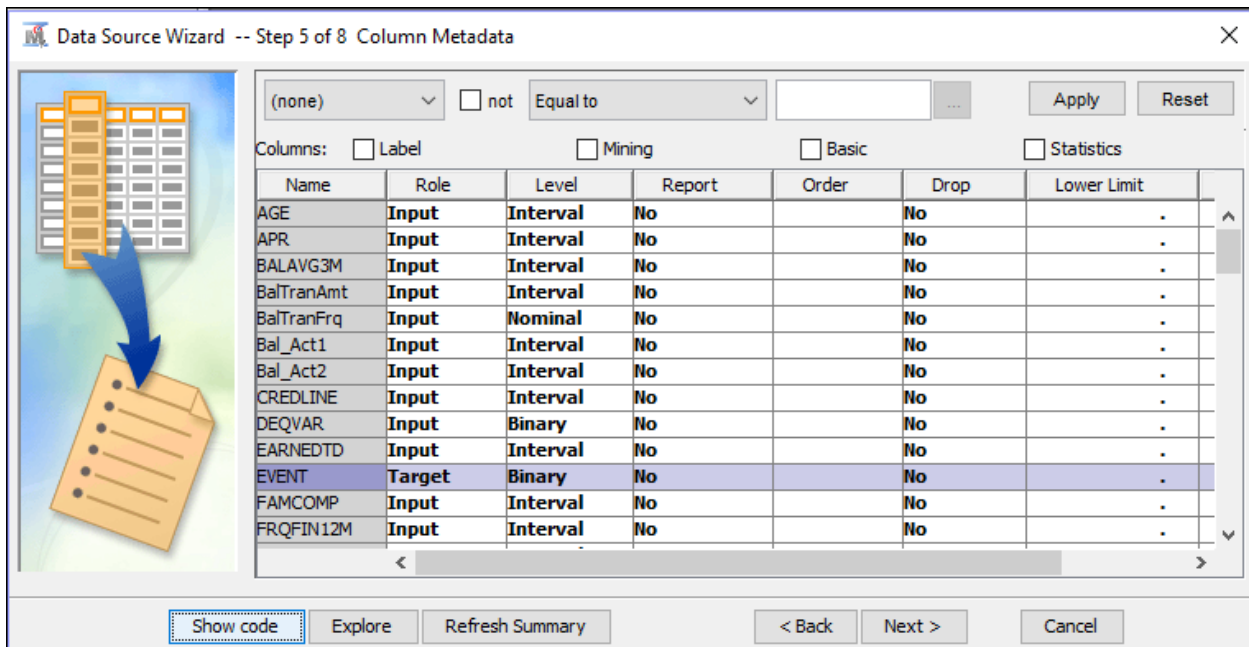


Display 4.17



The role of the variable “Event” is changed to “Target”

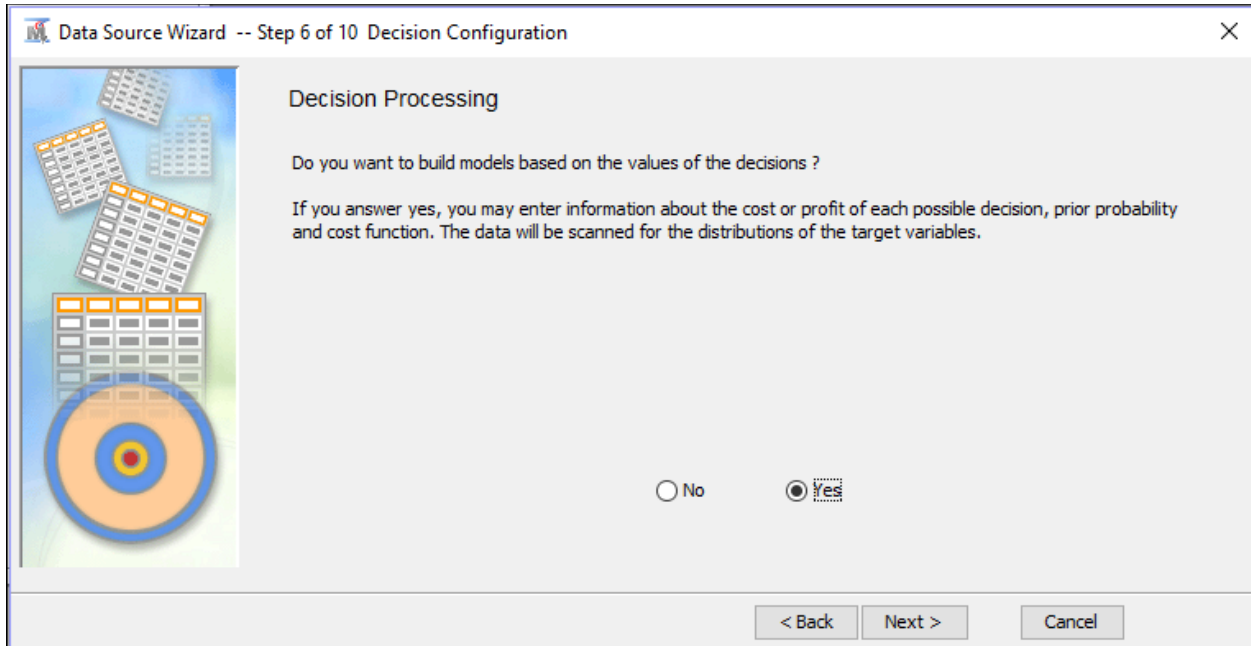
Display 4.18



Click “Next”

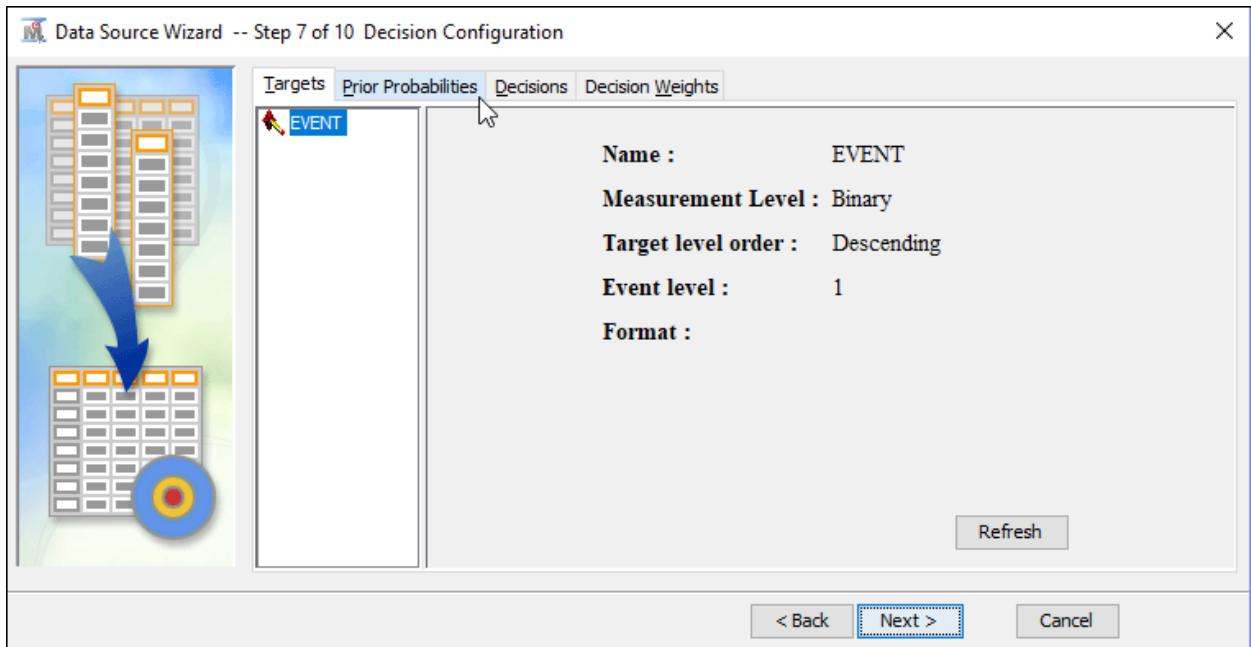
Select “Yes” to Decision Processing and click on “Next” (Display 4.19).

Display 4.19



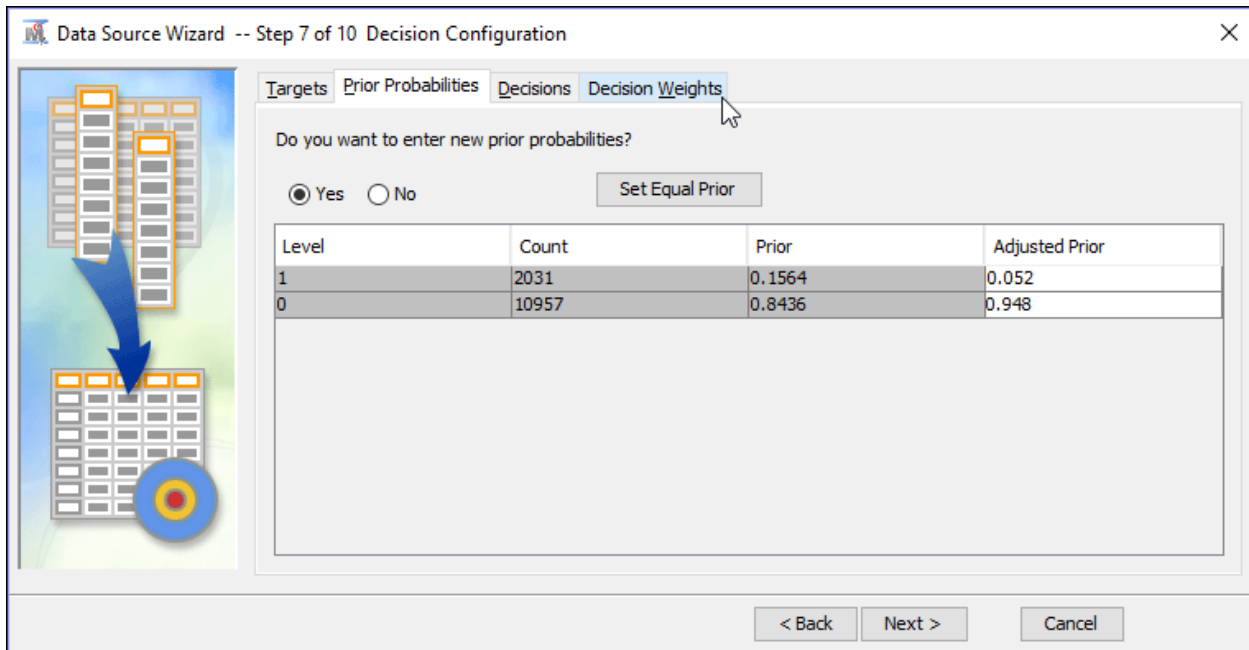
Select the “Prior Probabilities” Tab in Display 4.20

Display 4.20



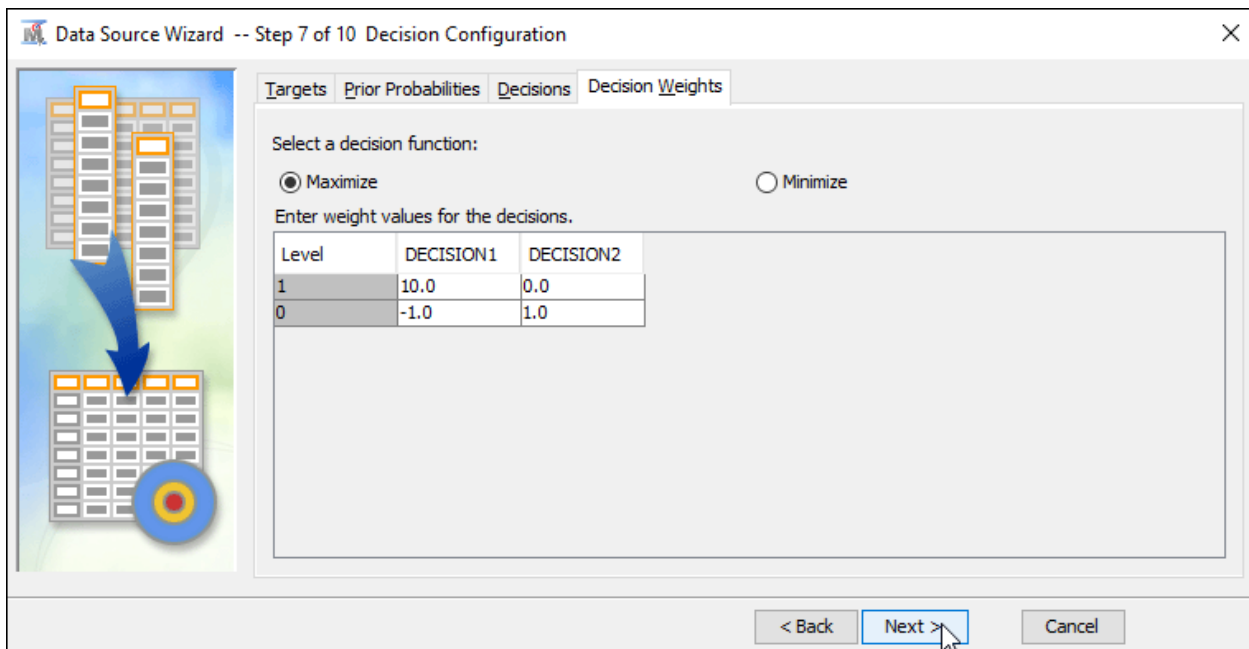
Select “Yes” to the question “Do you want to enter new prior probabilities?” and enter the new prior probabilities in the Adjusted Prior column as shown in the above display.

Display 4.21



Then select the tab “Decision Weights”. Enter the decision weights from Display 4.1 (Page 174 of the book).

Display 4.22



Click Next.

In the next window I answered “No” to the question “Do you wish to create a sample data set?” and clicked next (This window is not shown here)

The data source is created and the following window is displayed.

You can write notes and click next (Display 4.23). You get a summary.

Display 4.23

The screenshot shows the 'Data Source Wizard -- Step 9 of 10 Data Source Attributes' window. On the left is a vertical sidebar with a large orange 'i' icon and a blue arrow pointing down, along with a flow diagram of orange boxes. The main area contains the following fields:

- Name :** CH4\_BOOKDATA1
- Role :** Raw (dropdown menu)
- Segment :** (empty text box)
- Notes :** (empty text area)

At the bottom right, there are three buttons: '< Back' (highlighted with a dashed border), 'Next >', and 'Cancel'.

Display 4.24

The screenshot shows the 'Data Source Wizard -- Step 10 of 10 Summary' window. On the left is the same sidebar as in Display 4.23. The main area displays the following information:

Metadata Completed.

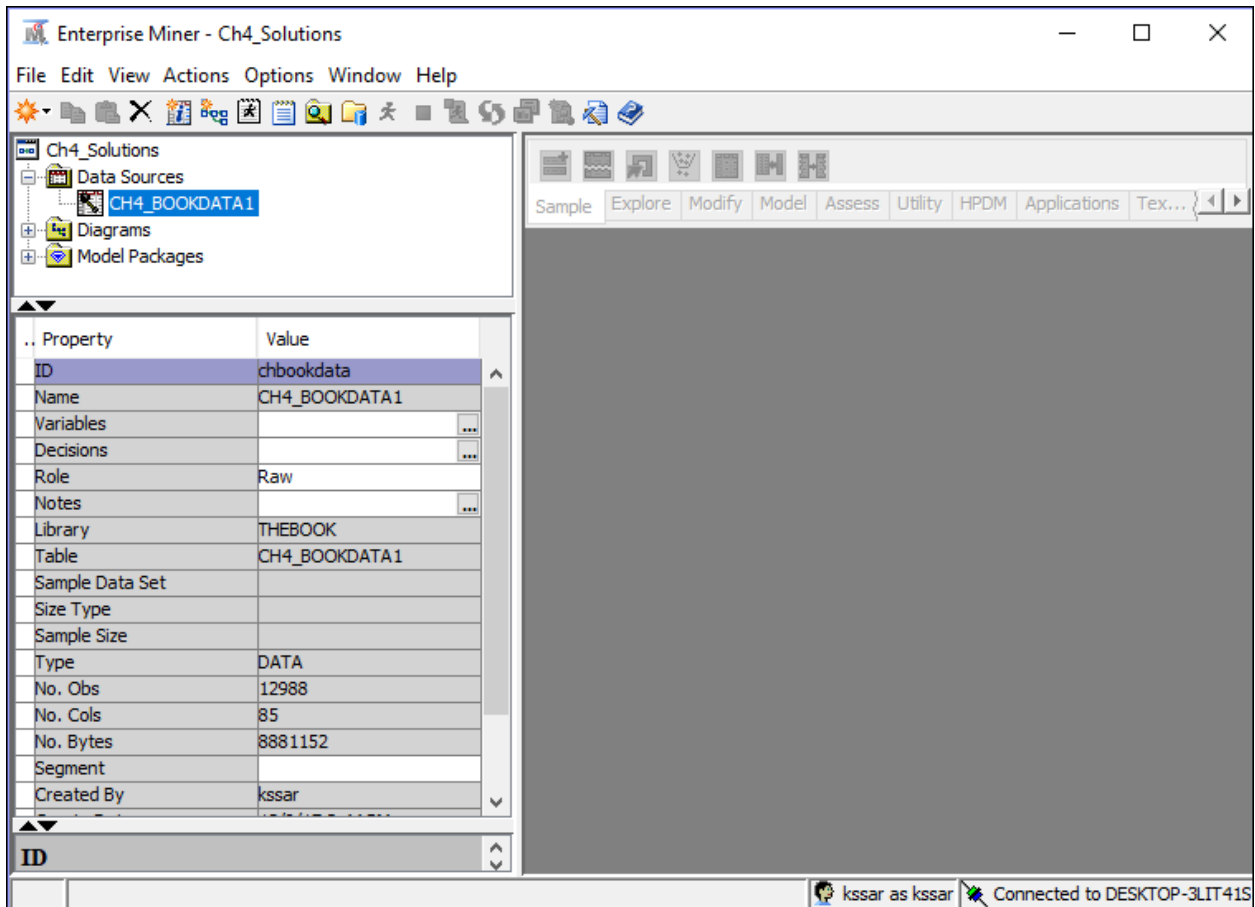
**Library:** THEBOOK  
**Data Source:** CH4\_BOOKDATA1  
**Role:** Raw

Role	Level	Count
Input	Binary	4
Input	Interval	75
Input	Nominal	3
Rejected	Interval	2
Target	Binary	1

At the bottom right, there are three buttons: '< Back', 'Finish' (highlighted with a dashed border), and 'Cancel'.

Click “Finish” and the data source is created.

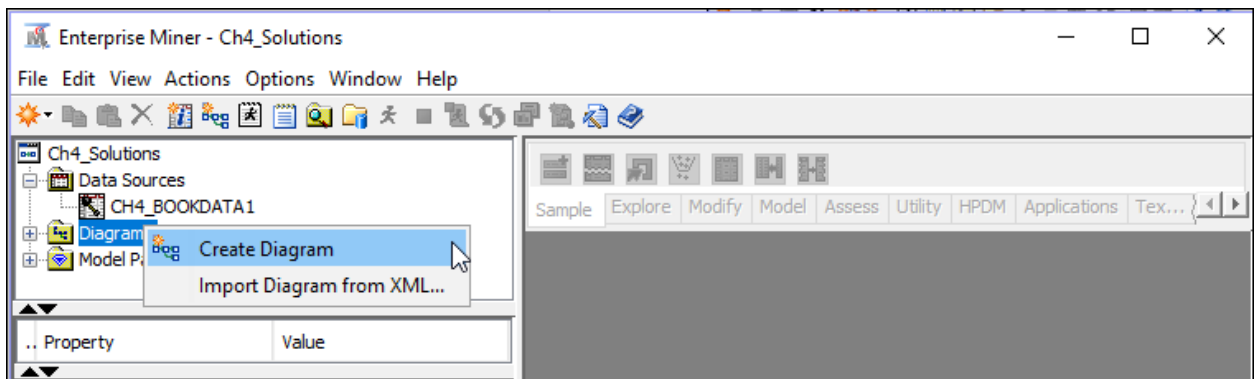
Display 4.25



Create Diagram:

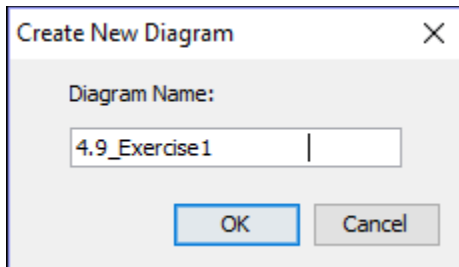
Open the project, right-click on diagrams, and select “Create Diagram

Display 4.26



Type a name for the Diagram (such as 4.9\_Exercise 1) as shown in Display 4.27.

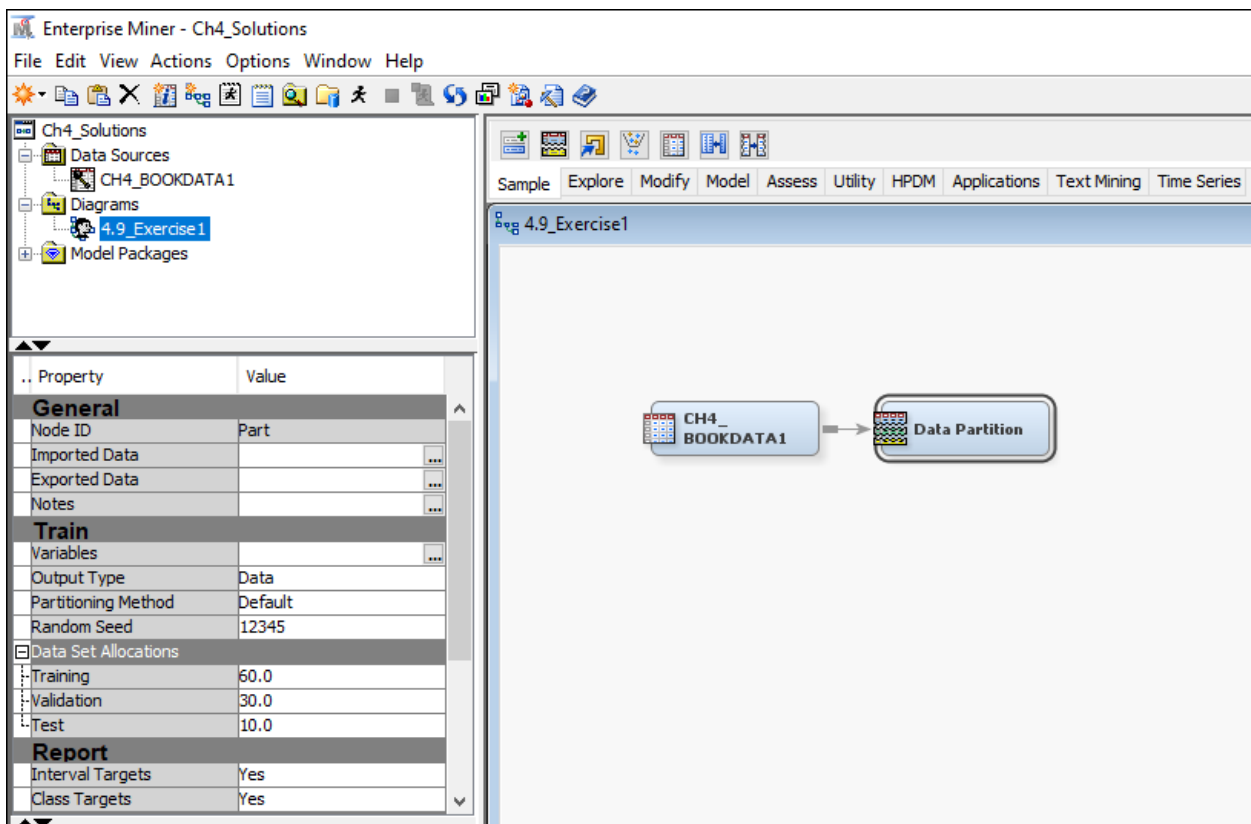
Display 4.27



Click OK to create the diagram.

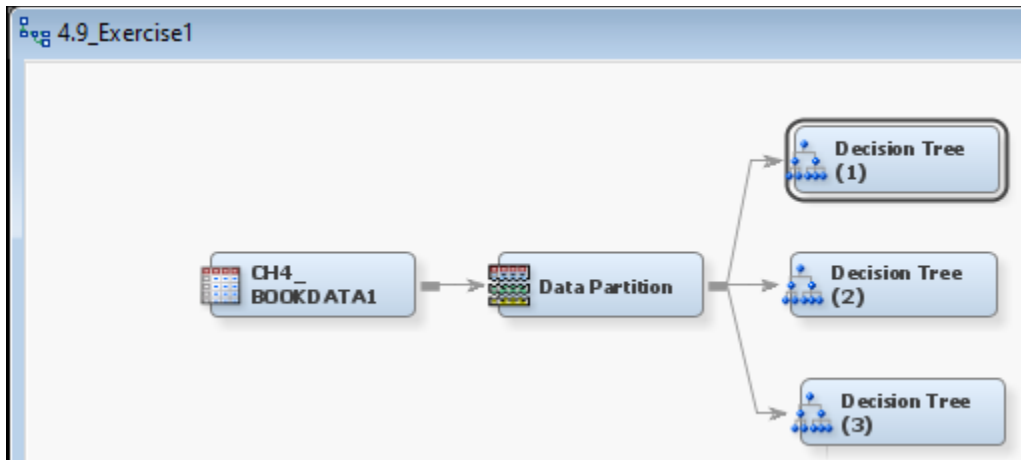
Drag the data source into the work space diagram and attach the data partition node. Set the data set allocation as shown in Display 4.28.

Display 4.28



Add the three Decision Tree nodes as shown in Display 4.29. The first decision Tree node is called “Decision Tree” by default. But I renamed it as Decision Tree (1).

Display 4.29



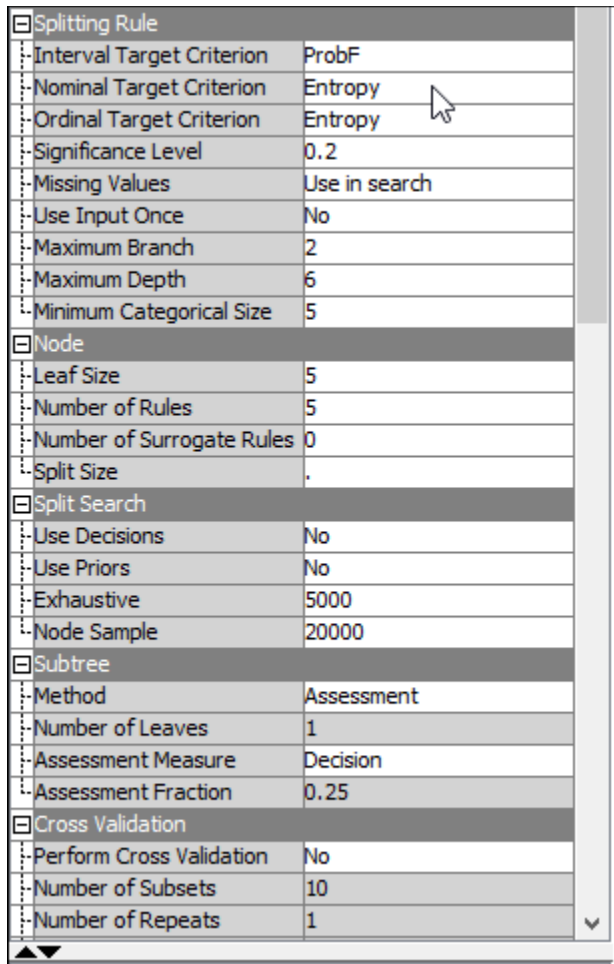
Properties of Decision Tree (1) are shown in Display 4.30.

Display 4.30

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1

Property settings of Decision Tree (2) are shown in Display 4.31.

Display 4.31



Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	Entropy
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1

Property settings for the Decision Tree (3) are shown in Display 4.32.

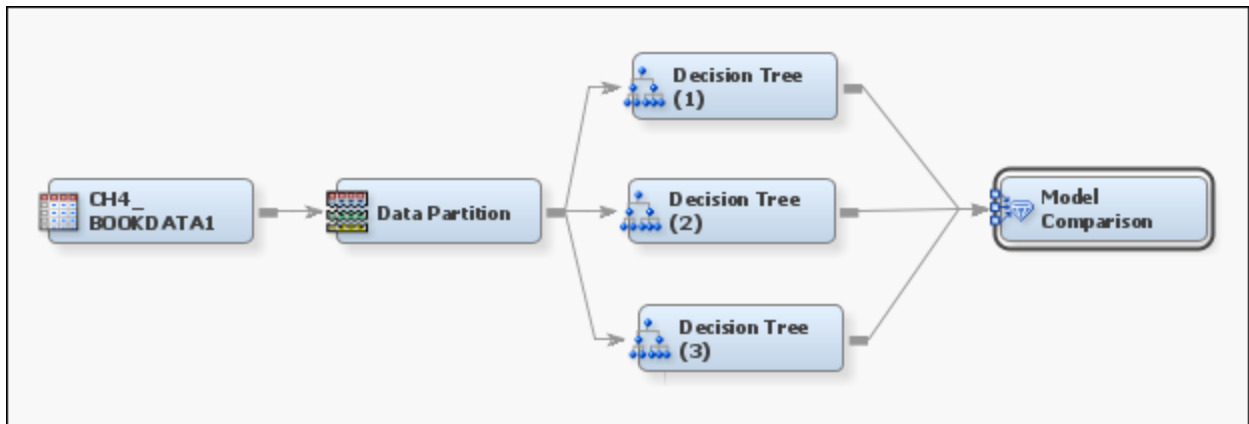


Display 4.32

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	Gini
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1

Add a model comparison node as shown in Display 4.33

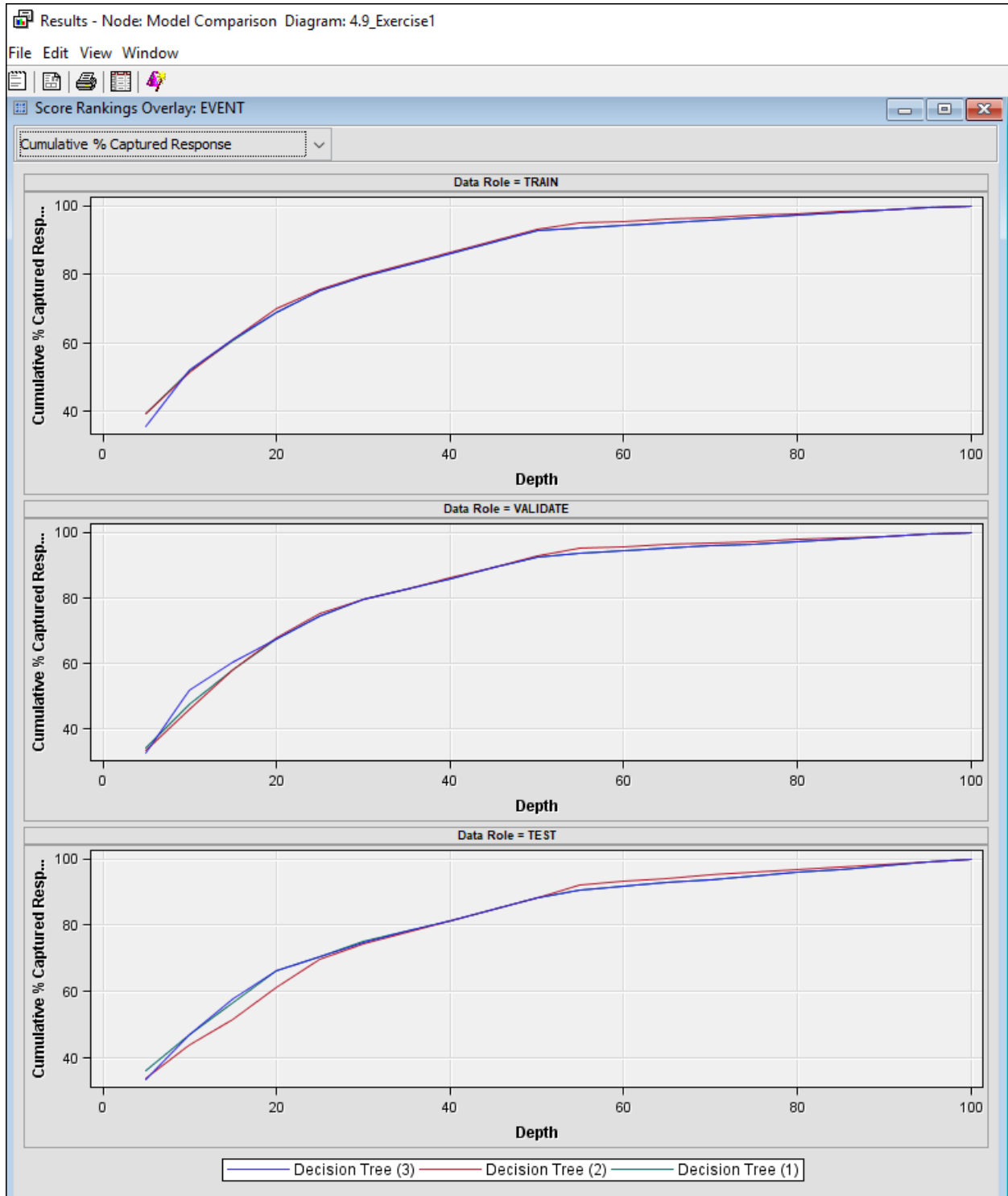
Display 4.33



Run the decision tree nodes and then the Model comparison node. Open the results window of the Model comparison node. I selected the cumulate capture rates graphs as shown in Display 4.34.

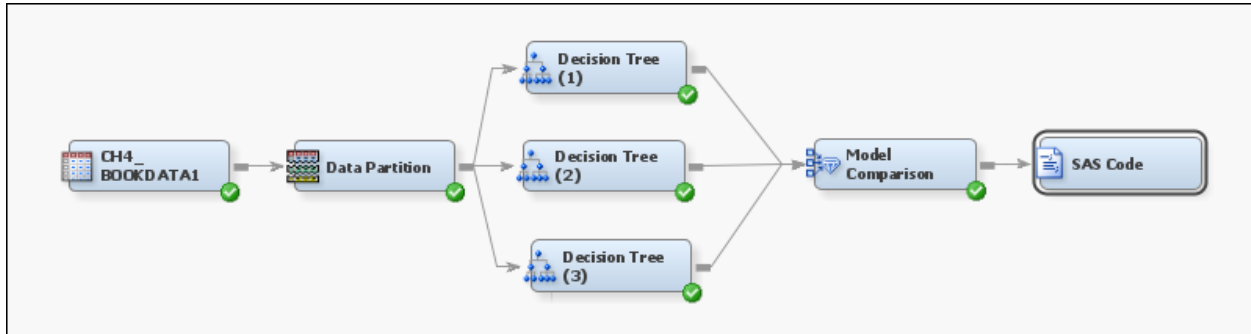
Display 4.34 shows the cumulate capture rates for the three models in training, validation and test data sets. We use the test data set for an independent evaluation of the models.

Display 4.34



From the cumulative capture rates for the Test data sets shown in the bottom frame of Display 4.34, it appears Decision Tree (1) is slightly better than the other two models. We can see the differences by printing the tables underlying these graphs. To enter the SAS code needed for these tables we attach a SAS code node to the process flow as shown in Display 4.35

Display 4.35



Enter the code shown in Display 4.36 into the code editor window of the SAS Code node.

Display 4.36

```

Title "TEST";
Data Test;
  set &EM_LIB..mdlcomp_emrank;
  keep model decile cap capc ;
  if datarole = "TEST";
run;
DATA TestM1(rename = (model=Model1 CAP=CAP1 CAPC = CAPC1))
  TestM2(rename = (model=Model2 CAP=CAP2 CAPC = CAPC2))
  TestM3(rename = (model=Model3 CAP=CAP3 CAPC = CAPC3));
set Test;
if upcase(MODEL) = "TREE" then output TestM1; else
if upcase(MODEL) = "TREE2" then output TestM2; else
if upcase(MODEL) = "TREE3" then output TestM3;
run ;
Data Test_all;
  merge TestM1 TestM2 TestM3 ;
  by decile ;
run;
proc print data=test_all ;
run;
  
```

After you run the above code, open the output window of the SAS Code node. You will see the capture rates (CAP1, CAP2 and CAP3) and the cumulative capture rates (CAPC1, CAPC2 and CAPC3) for the three models as shown in Display 4.37.

Display 4.37

Obs	Model1	CAP1	CAPC1	DECILE	Model2	CAP2	CAPC2	Model3	CAP3	CAPC3
1	Tree	36.0127	36.013	5	Tree2	33.8235	33.824	Tree3	33.5186	33.519
2	Tree	10.7334	46.746	10	Tree2	9.9815	43.805	Tree3	13.4114	46.930
3	Tree	9.8923	56.638	15	Tree2	7.5835	51.389	Tree3	10.8603	57.790
4	Tree	9.5102	66.149	20	Tree2	9.6597	61.048	Tree3	8.2907	66.081
5	Tree	4.3919	70.541	25	Tree2	8.5526	69.601	Tree3	4.3919	70.473
6	Tree	4.3919	74.932	30	Tree2	4.5577	74.159	Tree3	4.3919	74.865
7	Tree	3.1417	78.074	35	Tree2	3.6942	77.853	Tree3	3.2093	78.074
8	Tree	3.3387	81.413	40	Tree2	3.5062	81.359	Tree3	3.3387	81.413
9	Tree	3.3387	84.751	45	Tree2	3.5062	84.865	Tree3	3.3387	84.751
10	Tree	3.3387	88.090	50	Tree2	3.5062	88.371	Tree3	3.3387	88.090
11	Tree	2.4962	90.586	55	Tree2	3.5397	91.911	Tree3	2.4962	90.586
12	Tree	1.0460	91.632	60	Tree2	1.3984	93.310	Tree3	1.0460	91.632
13	Tree	1.0460	92.678	65	Tree2	0.8444	94.154	Tree3	1.0460	92.678
14	Tree	1.0460	93.724	70	Tree2	0.8444	94.998	Tree3	1.0460	93.724
15	Tree	1.0460	94.770	75	Tree2	0.8444	95.843	Tree3	1.0460	94.770
16	Tree	1.0460	95.816	80	Tree2	0.8444	96.687	Tree3	1.0460	95.816
17	Tree	1.0460	96.862	85	Tree2	0.8444	97.532	Tree3	1.0460	96.862
18	Tree	1.0460	97.908	90	Tree2	0.8444	98.376	Tree3	1.0460	97.908
19	Tree	1.0460	98.954	95	Tree2	0.8444	99.221	Tree3	1.0460	98.954
20	Tree	1.0460	100.000	100	Tree2	0.7795	100.000	Tree3	1.0460	100.000

By comparing the cumulative capture rates, we can decide which model is better. For example at the 6<sup>th</sup> decile (30<sup>th</sup> percentile) the cumulative capture rate is 74.932 for Decision Tree 1 (Model 1) , 74.159 for Model 2 and 74.865 for Model 2. This means that, if you target the top 30% of customers selected model 1, you will capture 74.932% of all defaults, If you target the top 30% of customers selected by Model 2, you will capture 74.159% of all defaults and if you target the top 30% of customers selected by Model 3, you will capture 74.865% of all defaults. Hence Model 1 (Decision Tree 1) is most powerful in identifying the defaults. Model 3 is the next most powerful model and Model 2 is the least powerful. But, the differences are not that large.

Exercise 2

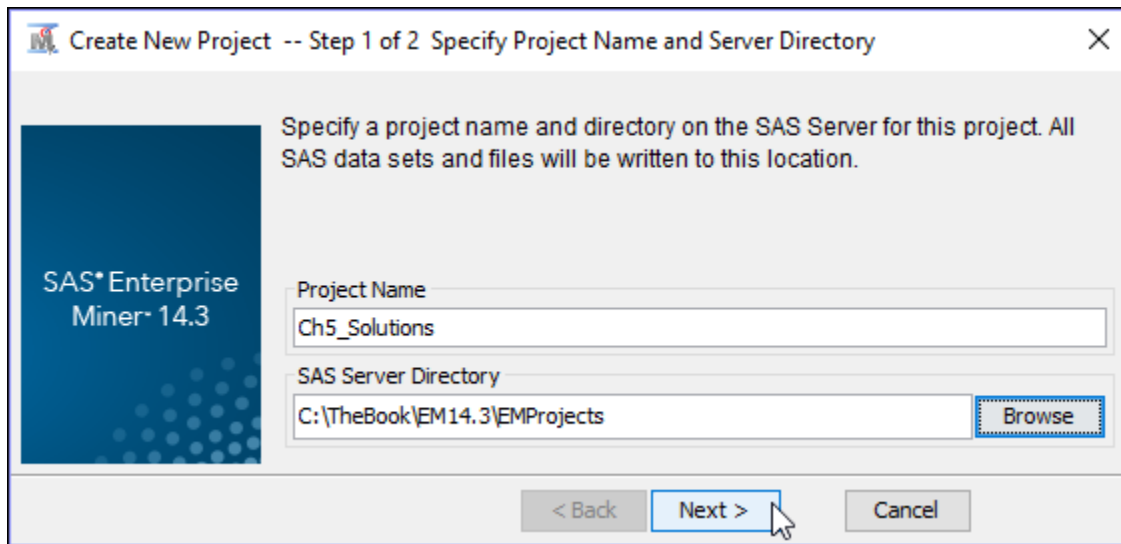
- (4) If a customer is assigned to Leaf Node 1, then the predicted probability of response = 0.02.
- (5) Using the Profit matrix given in display 4.8, we calculate the expected profits under Decision 1 (responder) and Decision 2 (non-responder).

	Decision 1	Decision 2		Posterior probability
1	5	0		0.02
0	-1	0		0.98
	Expected Profit Under Decision 1=			-0.88
	Expected Profit Under Decision 2=			0

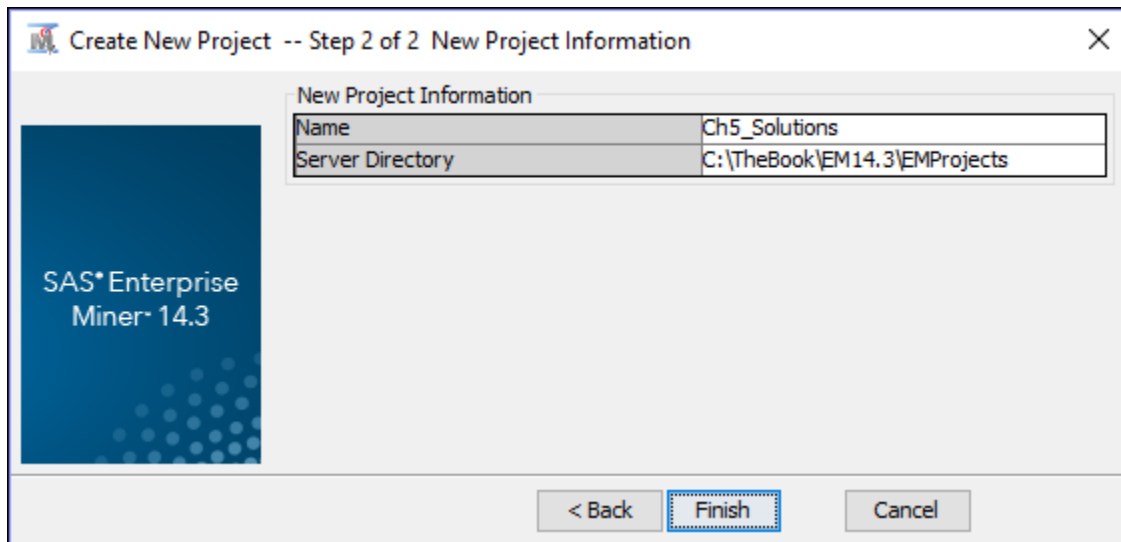
Since the expected profit under Decision 2 is greater than the expected profit under Decision 1, we label the leaf node as “non-responder” node. Since the selected customer is assigned to Leaf Node 1, we should not send invitation to the customer.

## Chapter 5

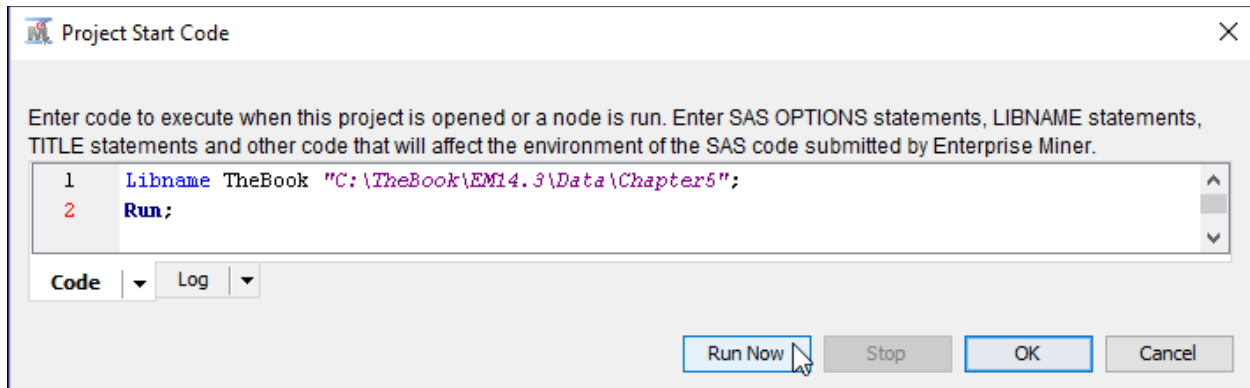
### Display 5.1



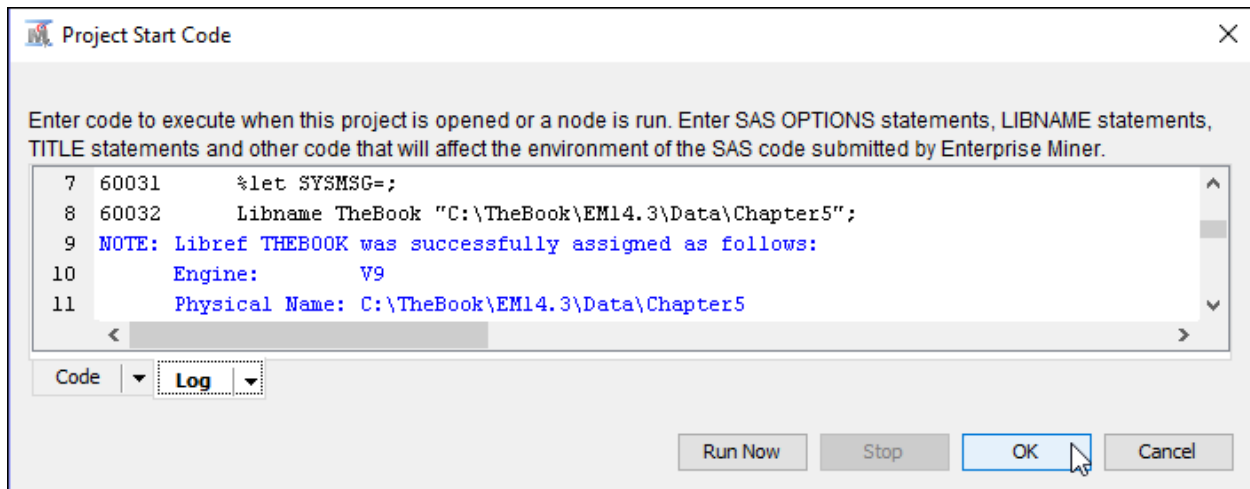
### Display 5.2



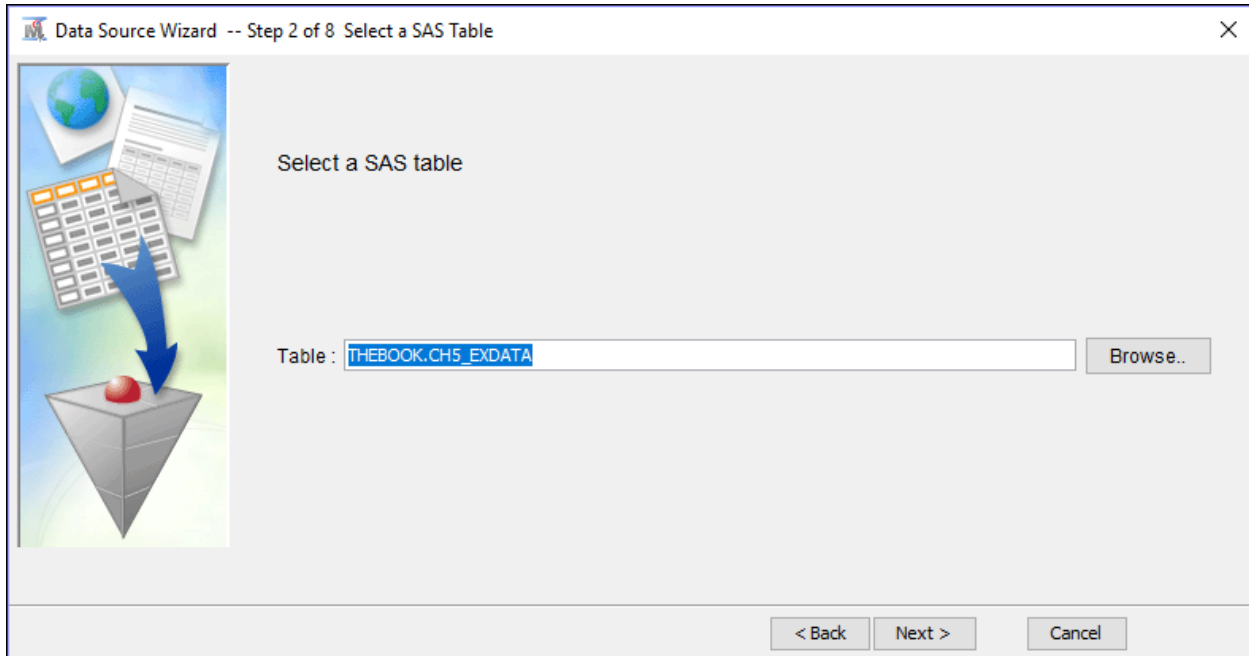
Display 5.3



Display 5.4



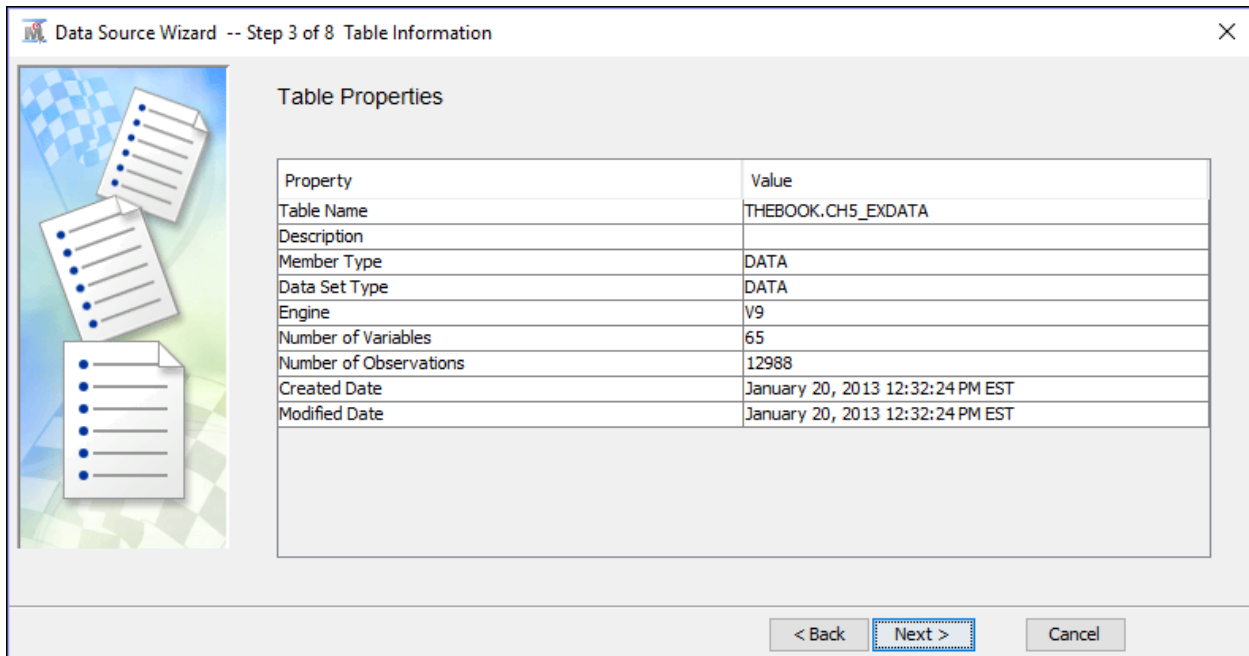
## Display 5.5



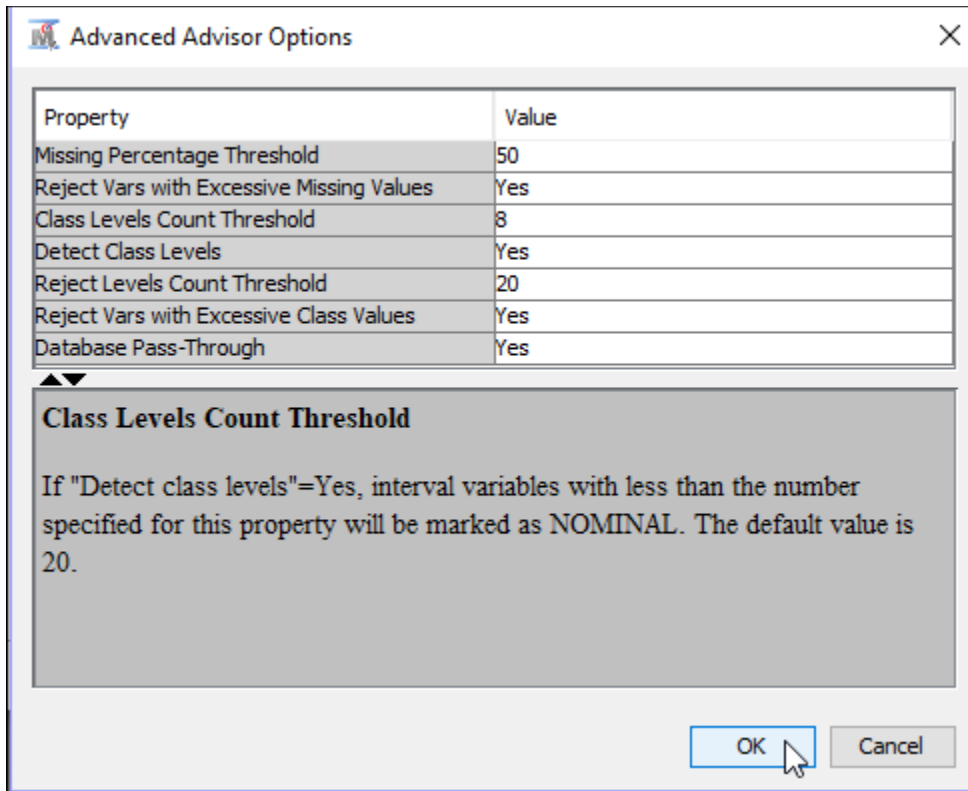
### Exercise 1:

Steps 1 and 2 of the Data Source Wizard are not displayed here, because they were discussed in the previous exercises. Step 3 of the Data Source Wizard is shown in Display 5.6.

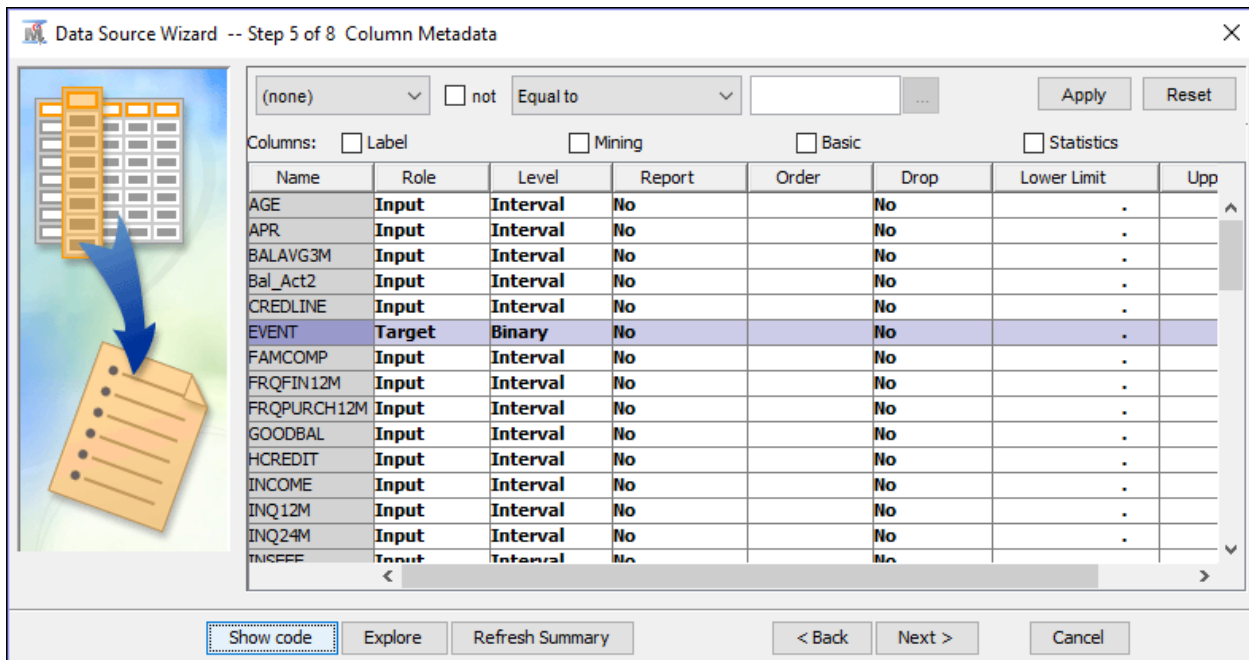
## Display 5.6



Display 5.7

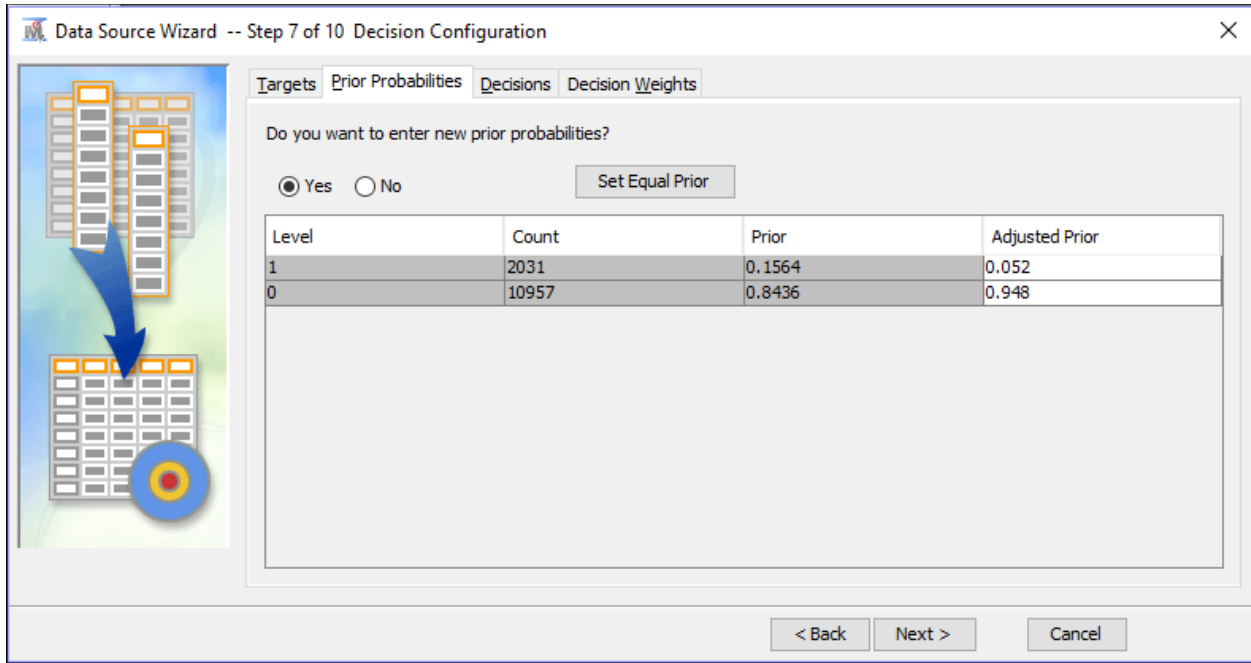


Display 5.8



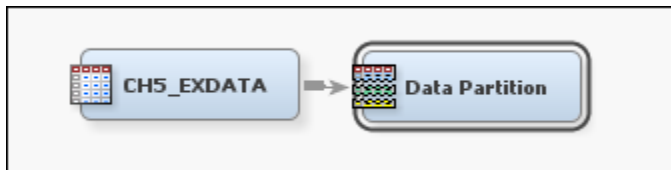


Display 5.9



Exercise 2

Display 5.10A

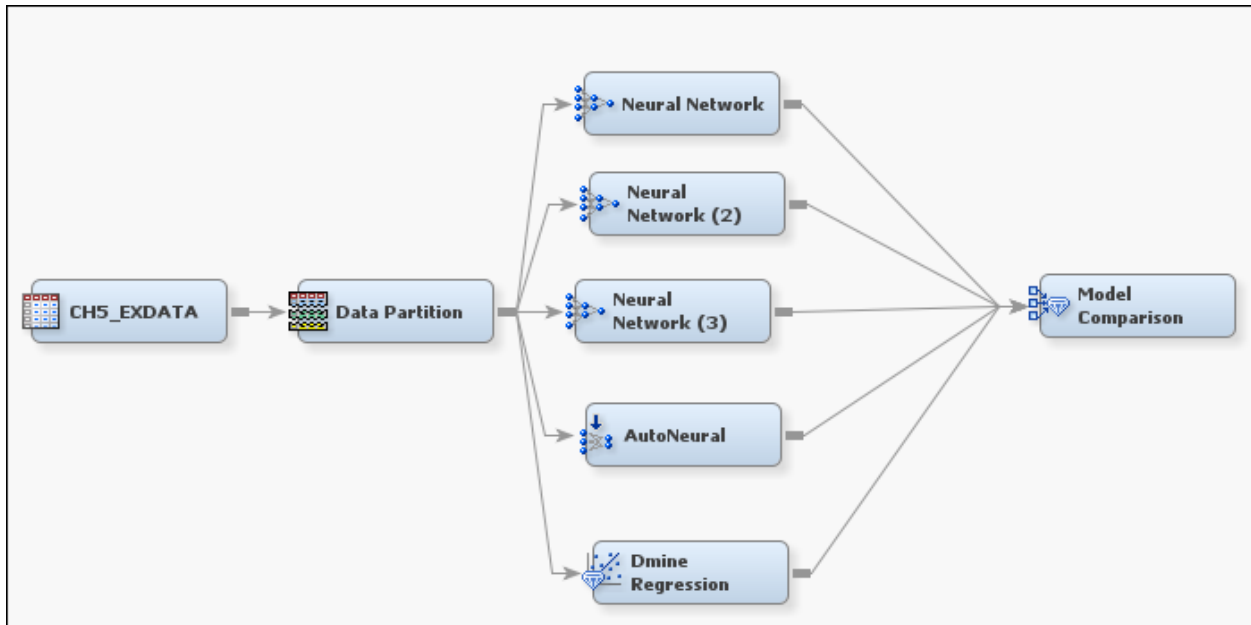


Display 5.10B

.. Property	Value
<b>General</b>	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	60.0
Validation	30.0
Test	10.0

Exercise 3.

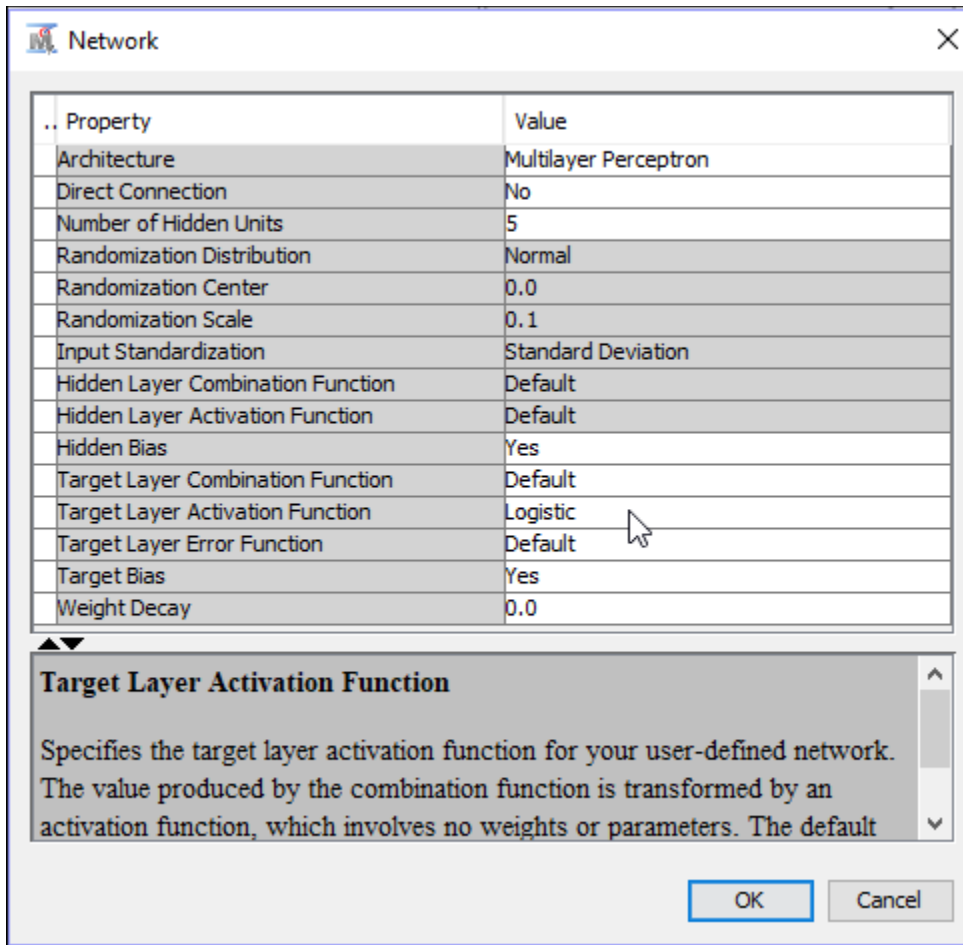
Display 5.11



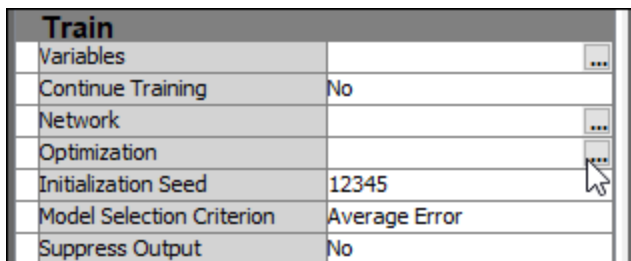
Display 5.12

Property	Value
<b>General</b>	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No

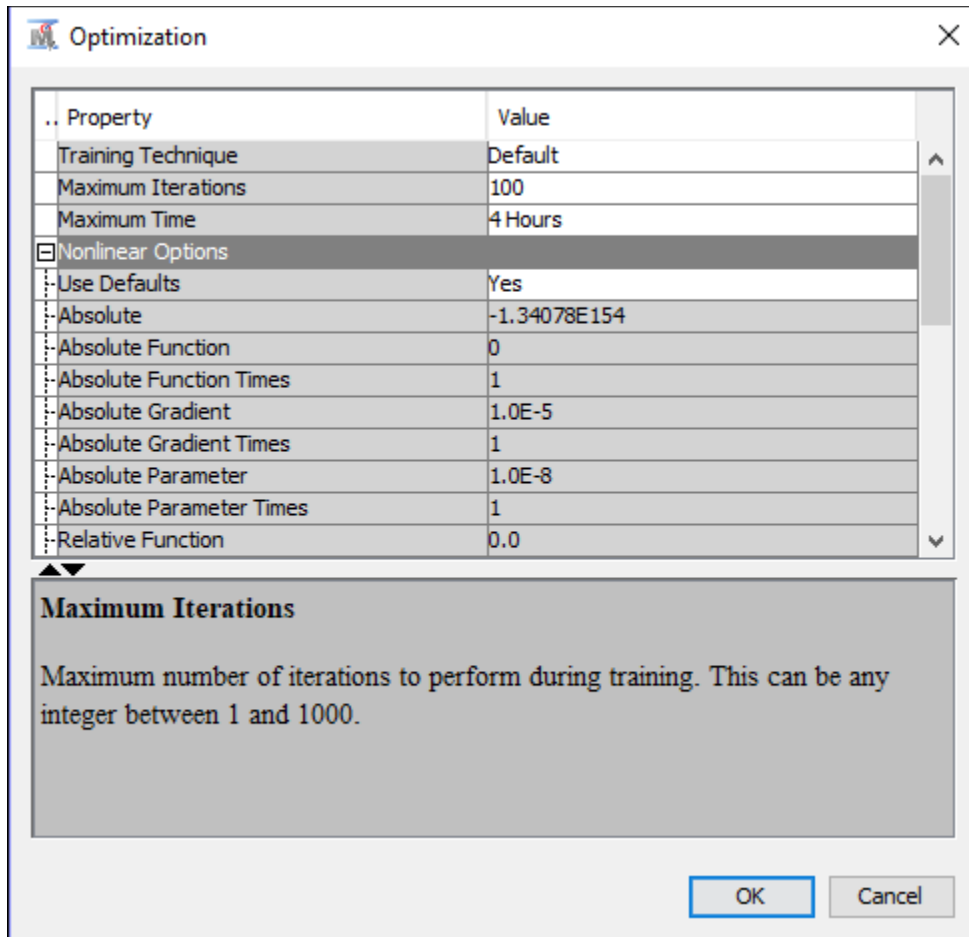
Display 5.13



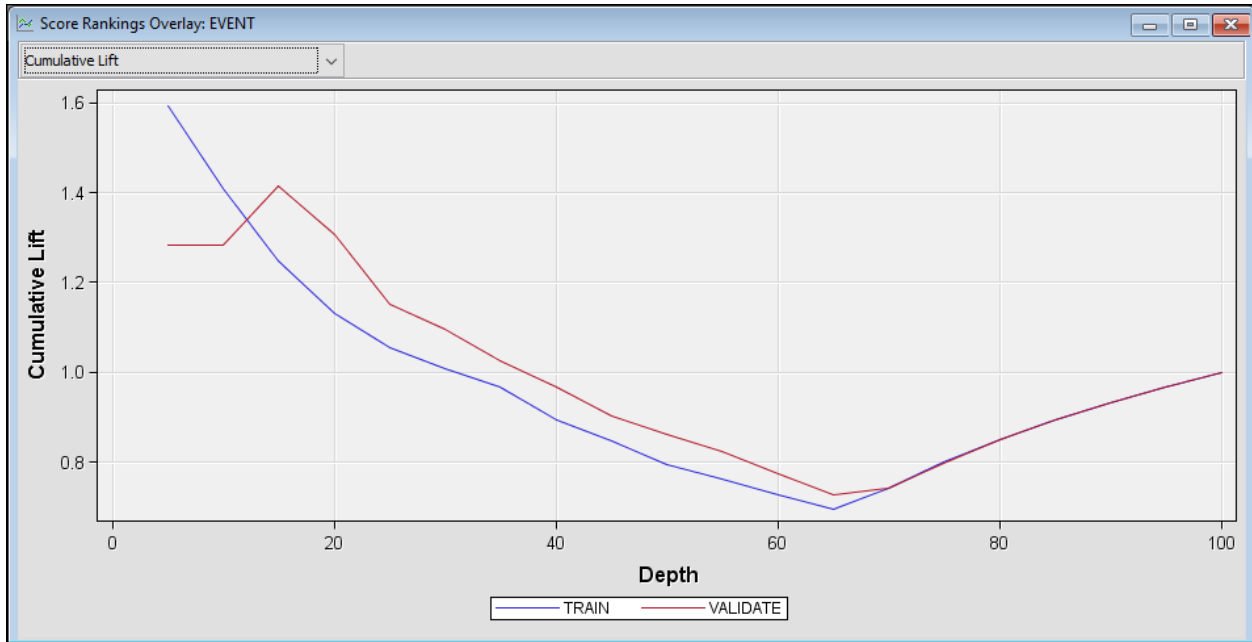
Display 5.15



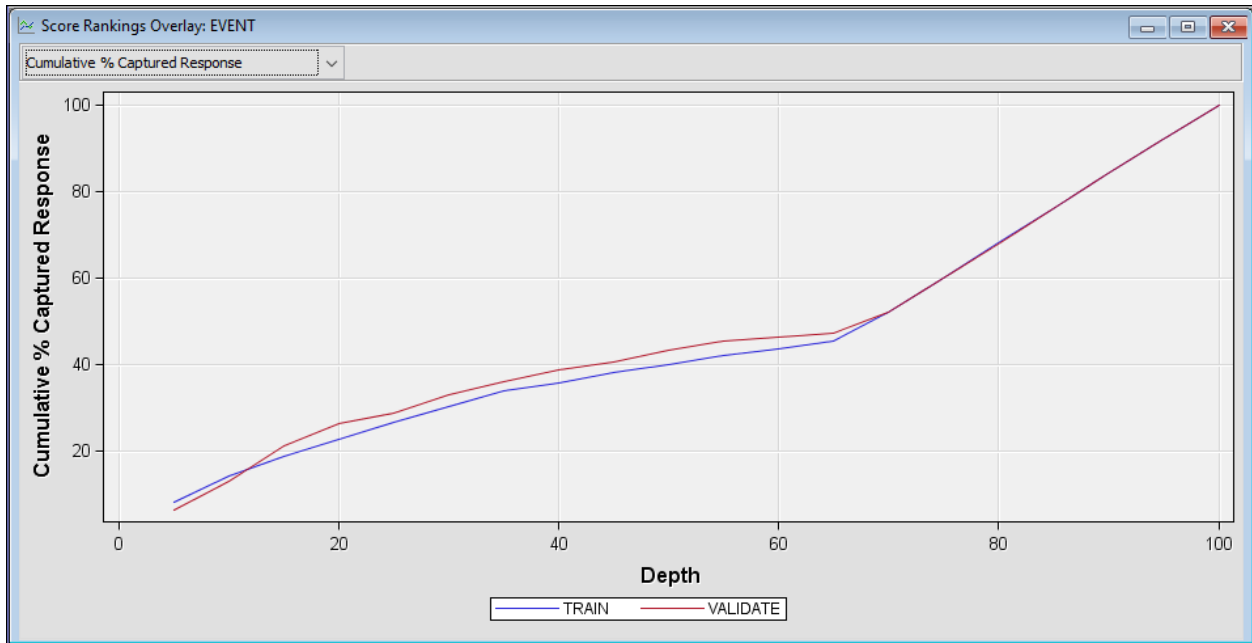
Display 5.16



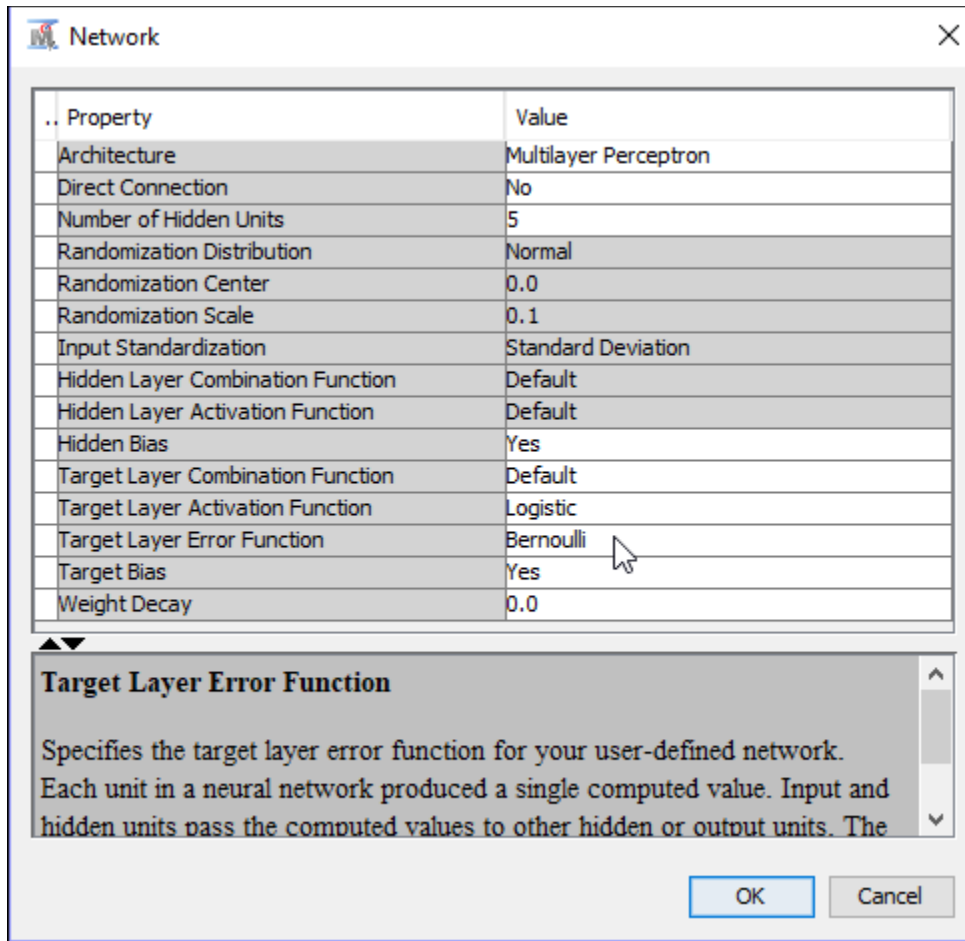
Display 5.17 (Cumulative Lift)



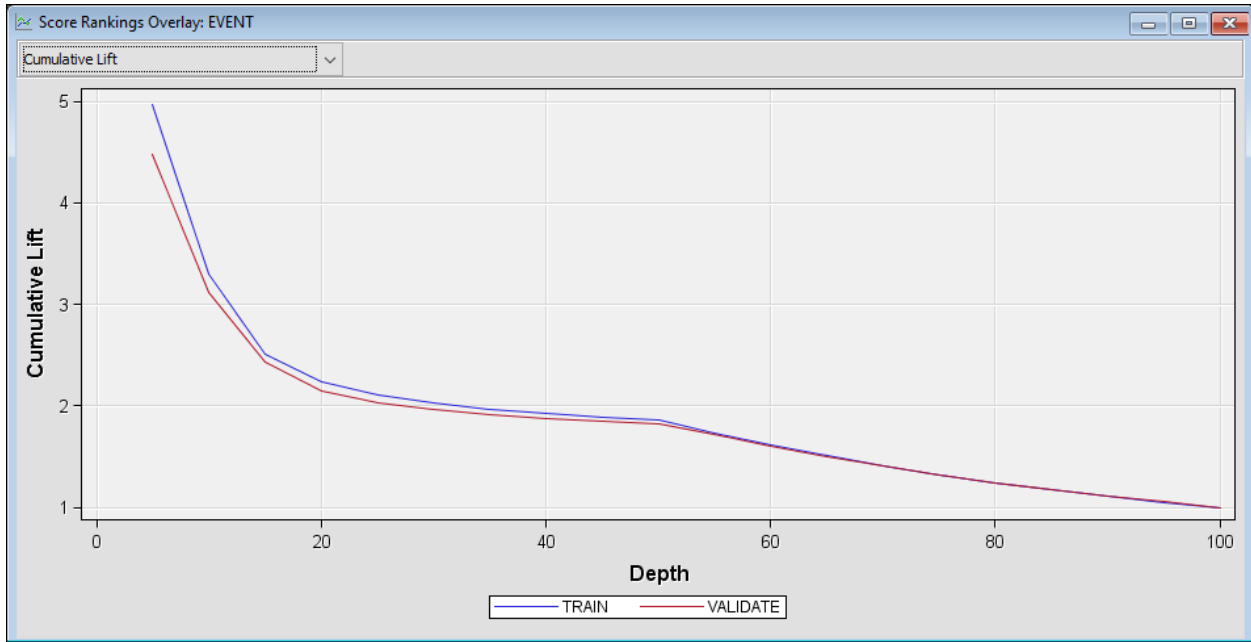
Display 5.18



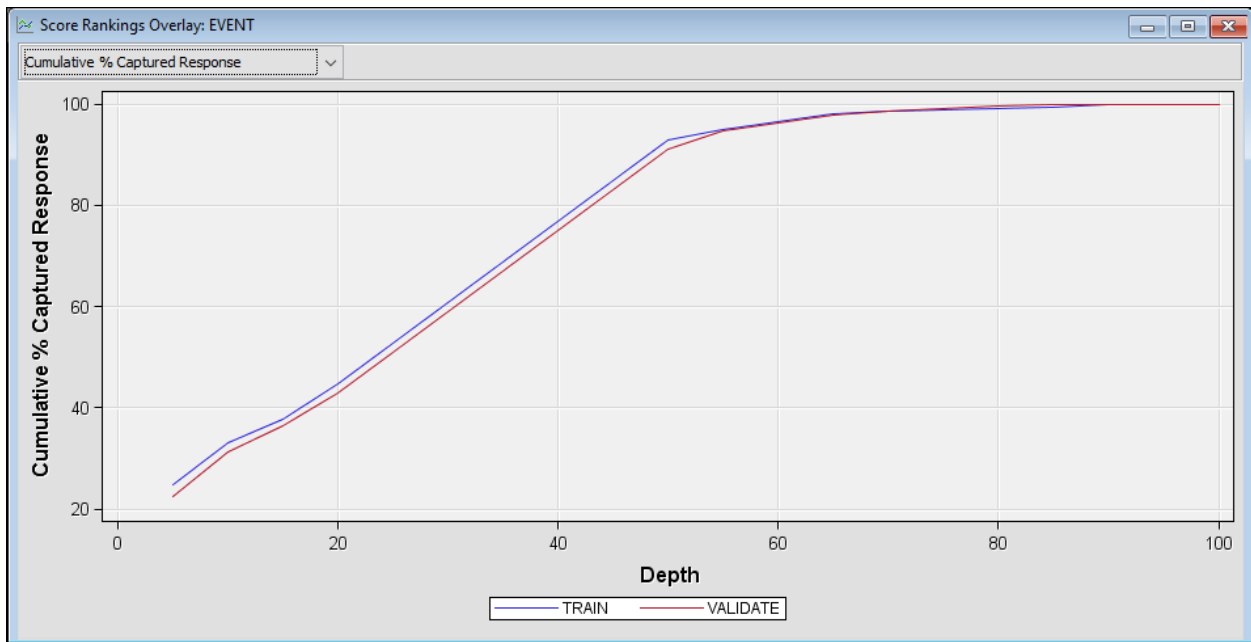
Display 5.19



Display 5.20



Display 5.21



Answer to Question 3d:

Yes, the model improved significantly when we changed the Target Layer Error Function Property to Bernoulli. This indicates that a Bernoulli error function is appropriate when the target layer activation function is logistic. The improvement can be seen by comparing the cumulative lift charts in Displays 5.17 and 5.20. A comparison of the cumulative capture rates in Displays 5.18 and 5.21 also confirms

that a Bernoulli error function is more appropriate than the default error function when the target layer activation function is logistic.

To see the SAS Code you do View → Scoring → SAS Code. Displays 5.22 and 5.23 give a partial view of the SAS code. Display 5.22 shows that the Hidden Unit Activation Function is tanh.

Display 5.22 (Partial list of the sas code)

```
H11 = 1.79279303430874 + H11 ;
H12 = -0.28844556732052 + H12 ;
H13 = -1.41348112144533 + H13 ;
H14 = 1.41030460683973 + H14 ;
H15 = 0.3011311937617 + H15 ;
H11 = TANH(H11 ) ;
H12 = TANH(H12 ) ;
H13 = TANH(H13 ) ;
H14 = TANH(H14 ) ;
H15 = TANH(H15 ) ;
```

Display 5.23 shows the calculation of probability of the Event (default in this example)

Display 5.23

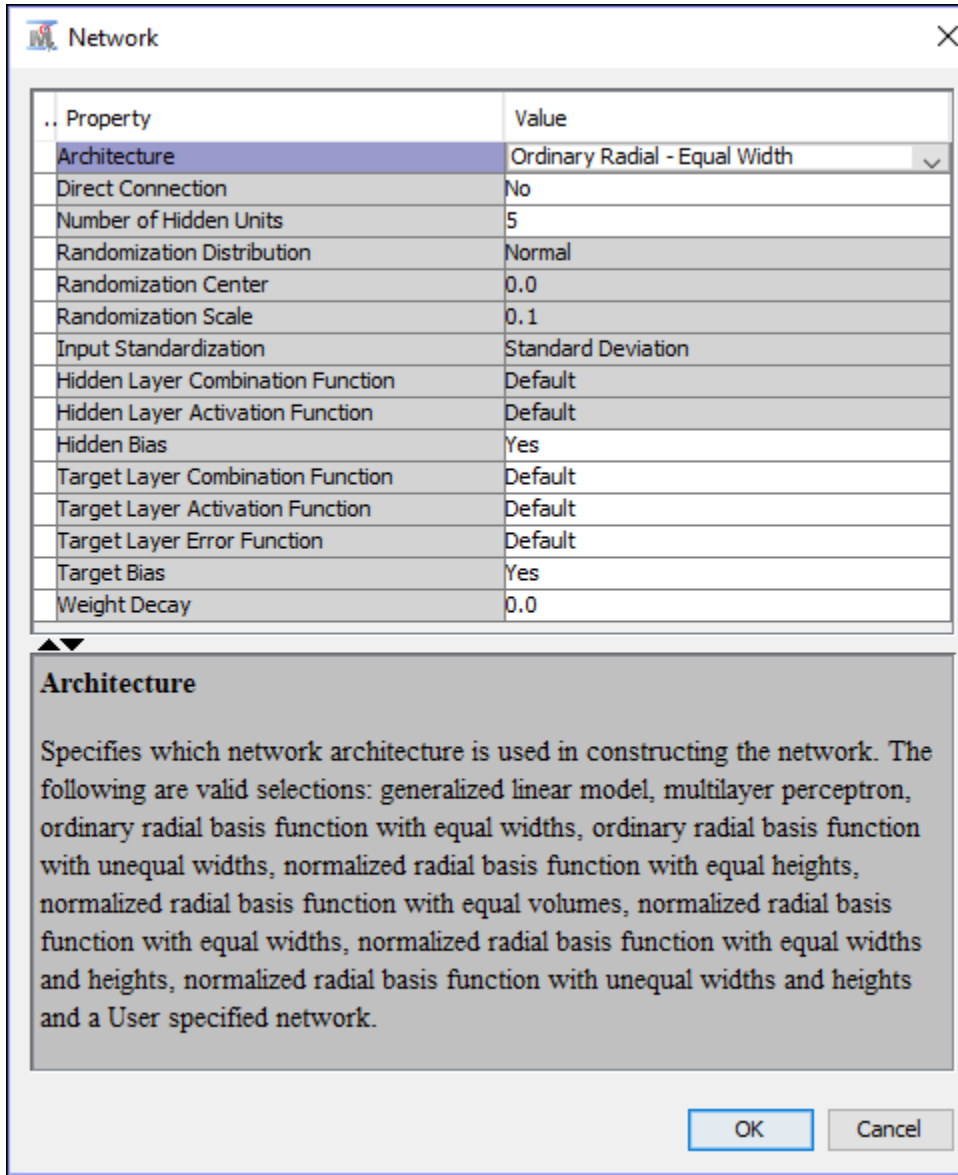
```
IF _DM_BAD EQ 0 THEN DO;
  P_EVENT1 = -0.53152571676659 * H11 + 1.10654049440834 * H12
    + 1.05579863331732 * H13 + -0.7523055031816 * H14
    + 0.7802227266483 * H15 ;
  P_EVENT0 = 0.53152571676659 * H11 + -1.10654049440834 * H12
    + -1.05579863331732 * H13 + 0.7523055031816 * H14
    + -0.7802227266483 * H15 ;
  P_EVENT1 = -2.09442074800226 + P_EVENT1 ;
  P_EVENT0 = 2.09442074800226 + P_EVENT0 ;
  DROP _EXP_BAR;
  _EXP_BAR=50;
  P_EVENT1 = 1.0 / (1.0 + EXP(MIN( - P_EVENT1 , _EXP_BAR)));
  P_EVENT0 = 1.0 / (1.0 + EXP(MIN( - P_EVENT0 , _EXP_BAR)));
END;
ELSE DO;
  P_EVENT1 = . ;
  P_EVENT0 = . ;
END;
IF _DM_BAD EQ 1 THEN DO;
  P_EVENT1 = 0.11149193548387;
  P_EVENT0 = 0.88850806451612;
END;
```

The code given in Display 5.23 shows that the formulae used are as expected. They are similar to the formulae one uses for calculating probabilities of default (Event) when one uses a logistic regression.



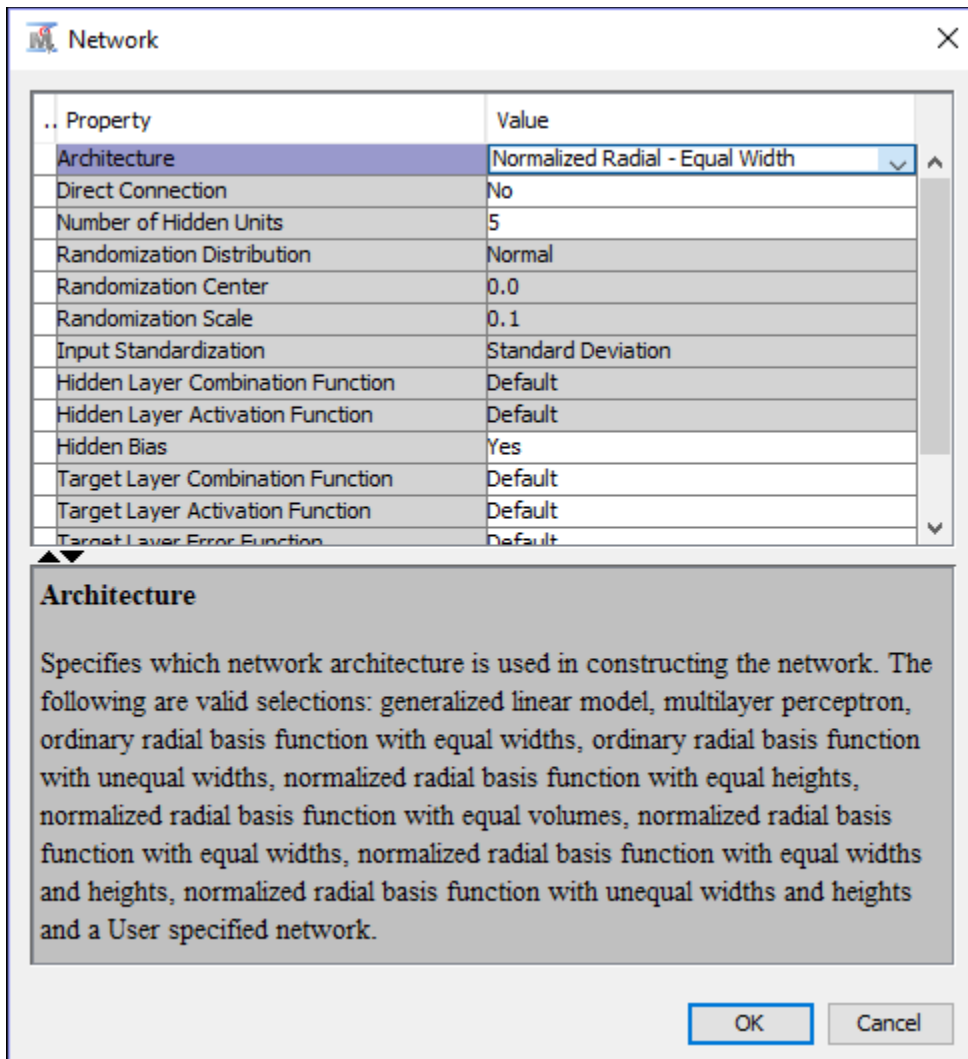
### Exercise 3f

Display 5.24

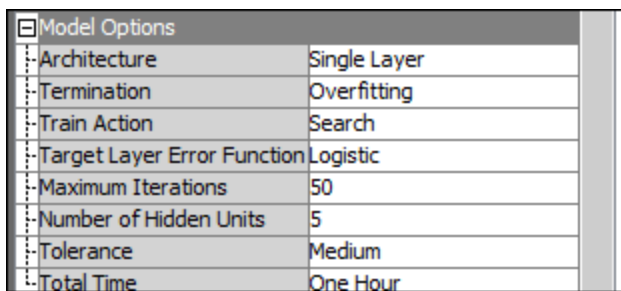


### Exercise 3g

Display 5.25

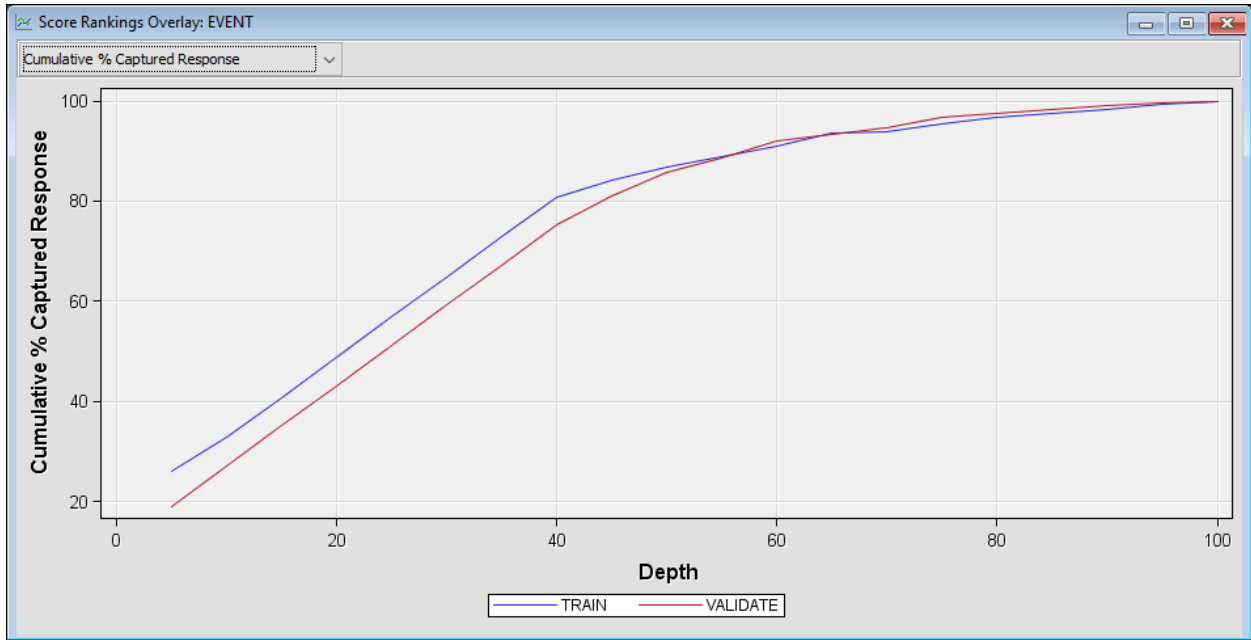


Display 5.26 (Auto Neural)

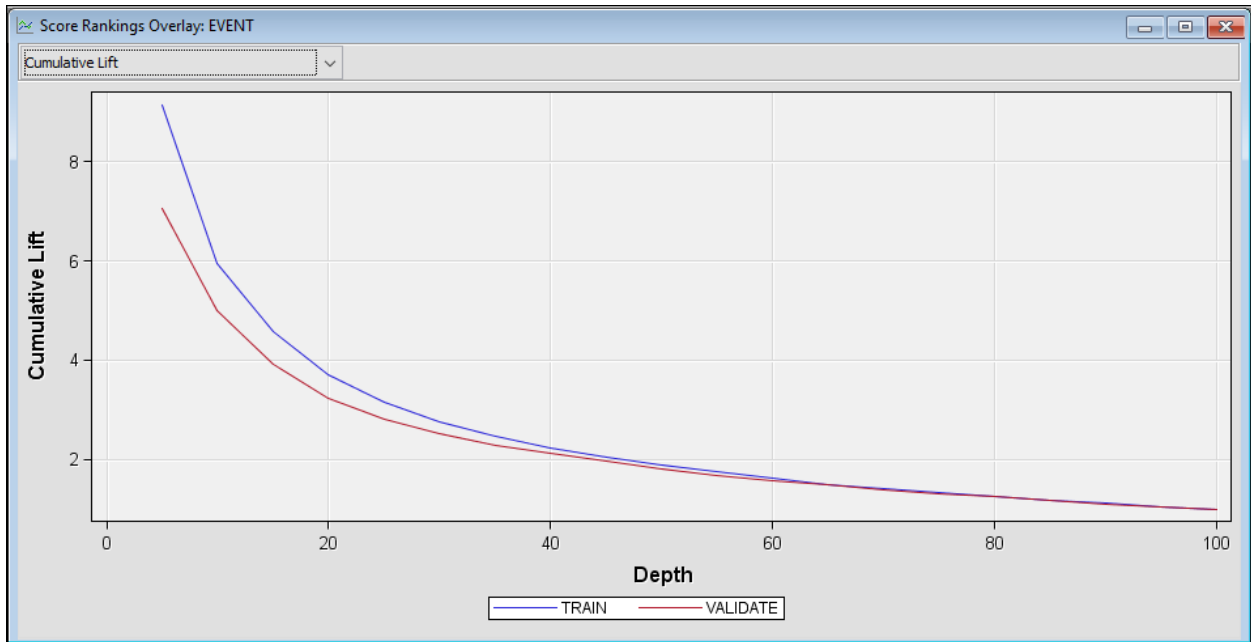


Display 5.27 shows the cumulative Capture rates for the model created by AutoNeural .

Display 5.27

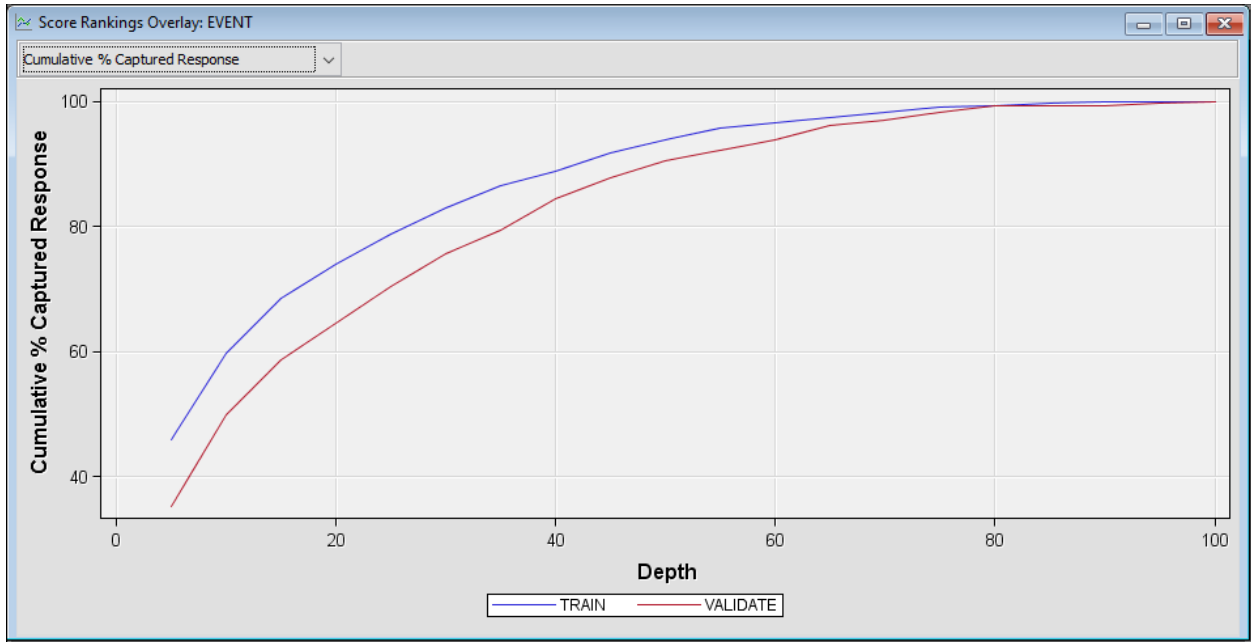


Display 5.28 (Lift charts for DMine Regression)



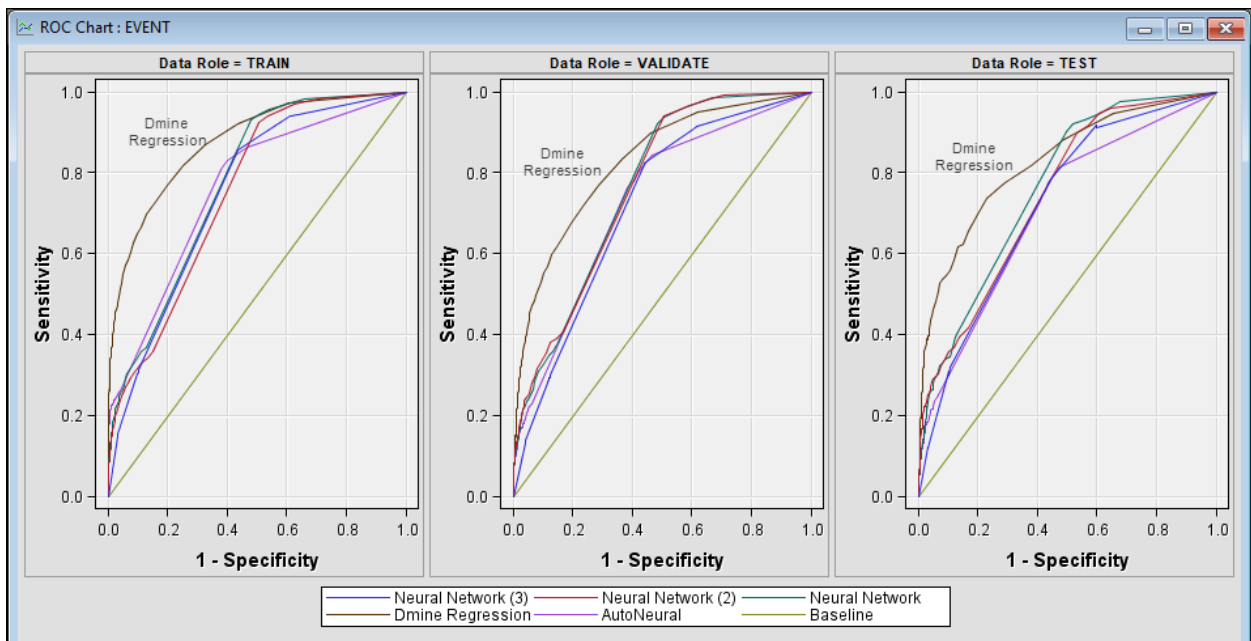
Display 5.29 shows Cumulative capture rate for the Dmine Regression

Display 5.29



Run the Model Comparison node, open the Results Window. Display 5.30 shows the ROC charts for all the five models compared.

Display 5.30



From Display 5.30, the best model is produced by the DMine Regression.

## Chapter 6

Create a project and give it a name.

Display 6.1

The screenshot shows a dialog box titled "Create New Project -- Step 1 of 2 Specify Project Name and Server Directory". On the left is a blue sidebar with the text "SAS\* Enterprise Miner 14.3". The main area contains the instruction: "Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location." Below this are two input fields: "Project Name" with the value "Ch6\_Solutions" and "SAS Server Directory" with the value "C:\TheBook\EM14.3\EMProjects". A "Browse" button is next to the directory field. At the bottom are buttons for "< Back", "Next >", and "Cancel".

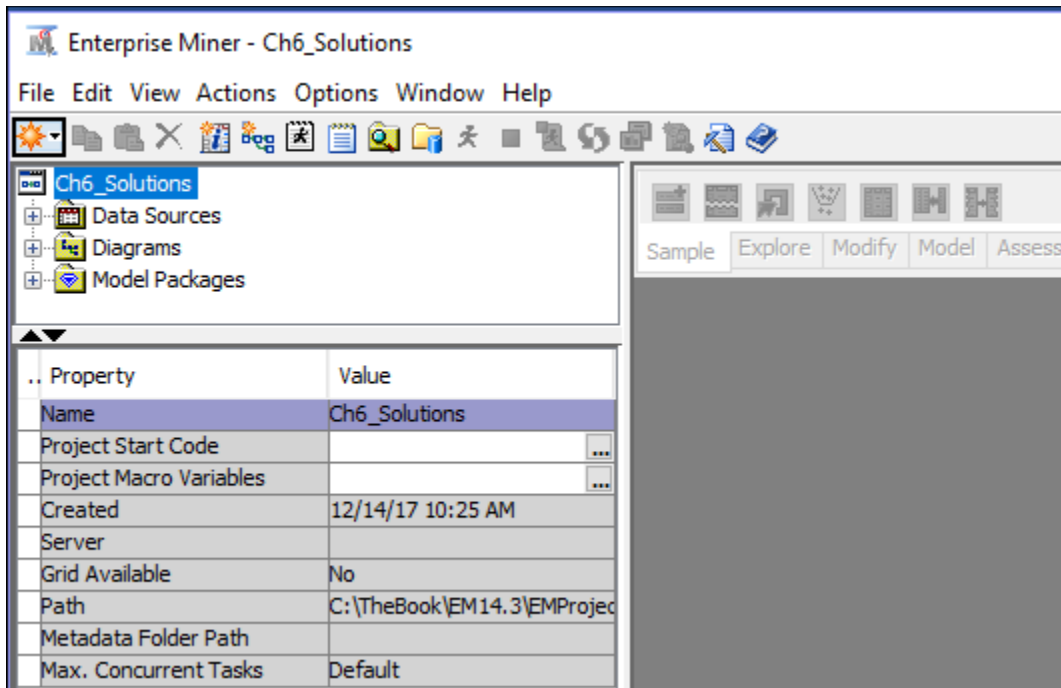
Display 6.2

The screenshot shows a dialog box titled "Create New Project -- Step 2 of 2 New Project Information". On the left is a blue sidebar with the text "SAS\* Enterprise Miner 14.3". The main area contains a table titled "New Project Information":

New Project Information	
Name	Ch6_Solutions
Server Directory	C:\TheBook\EM14.3\EMProjects

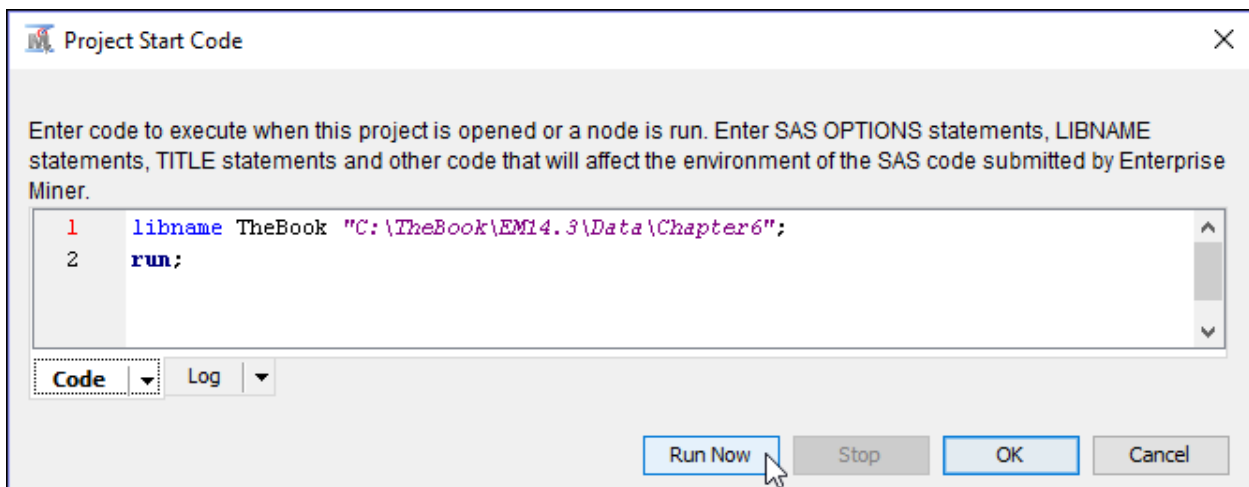
At the bottom are buttons for "< Back", "Finish", and "Cancel".

Display 6.3



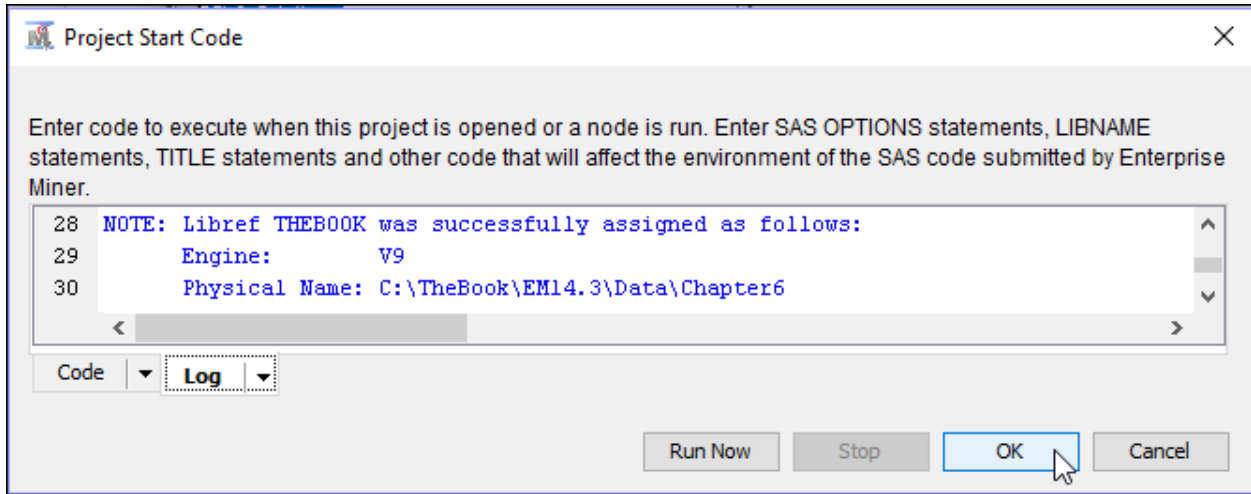
Create a libref in the Project Start code window

Display 6.4



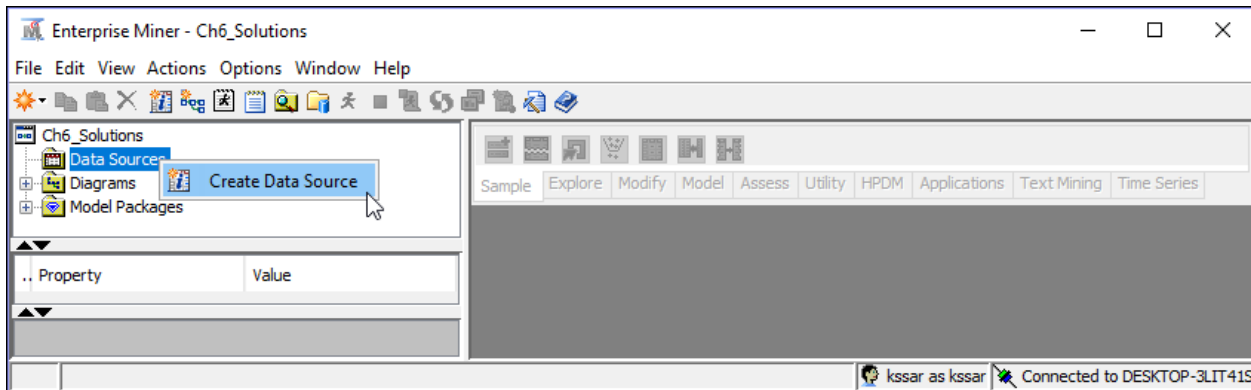
Check the log window and click "OK"

Display 6.5

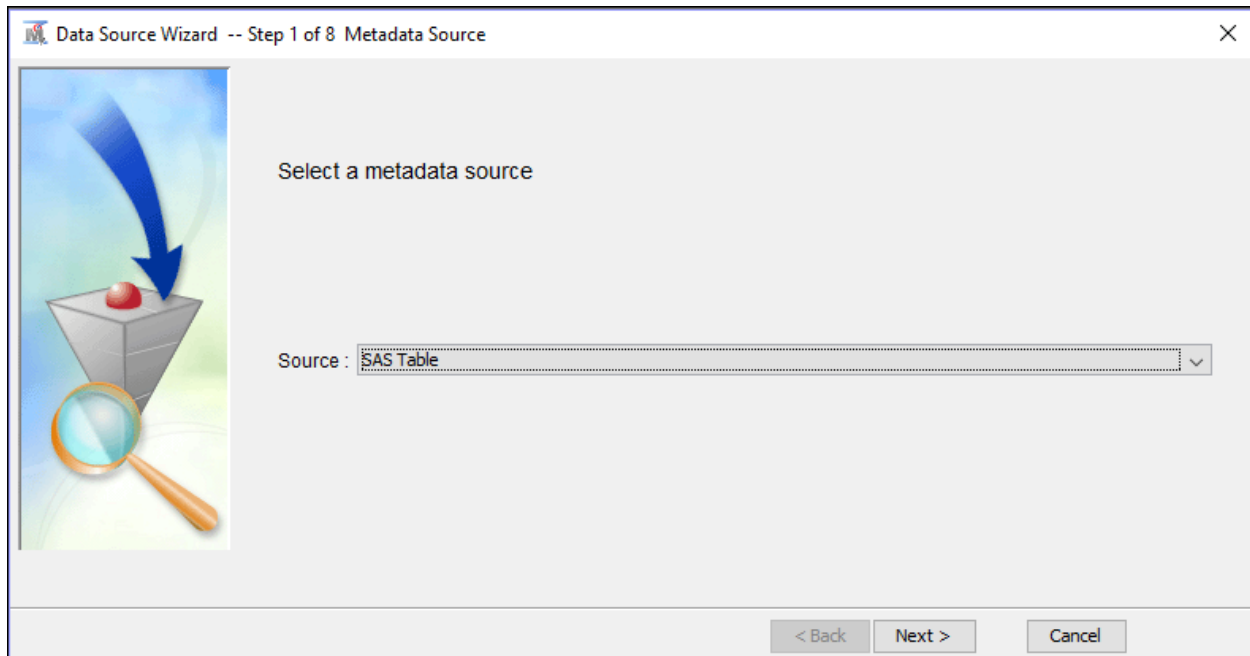


Create a data source using the data source wizard as illustrated in Displays 6.6 –6.21

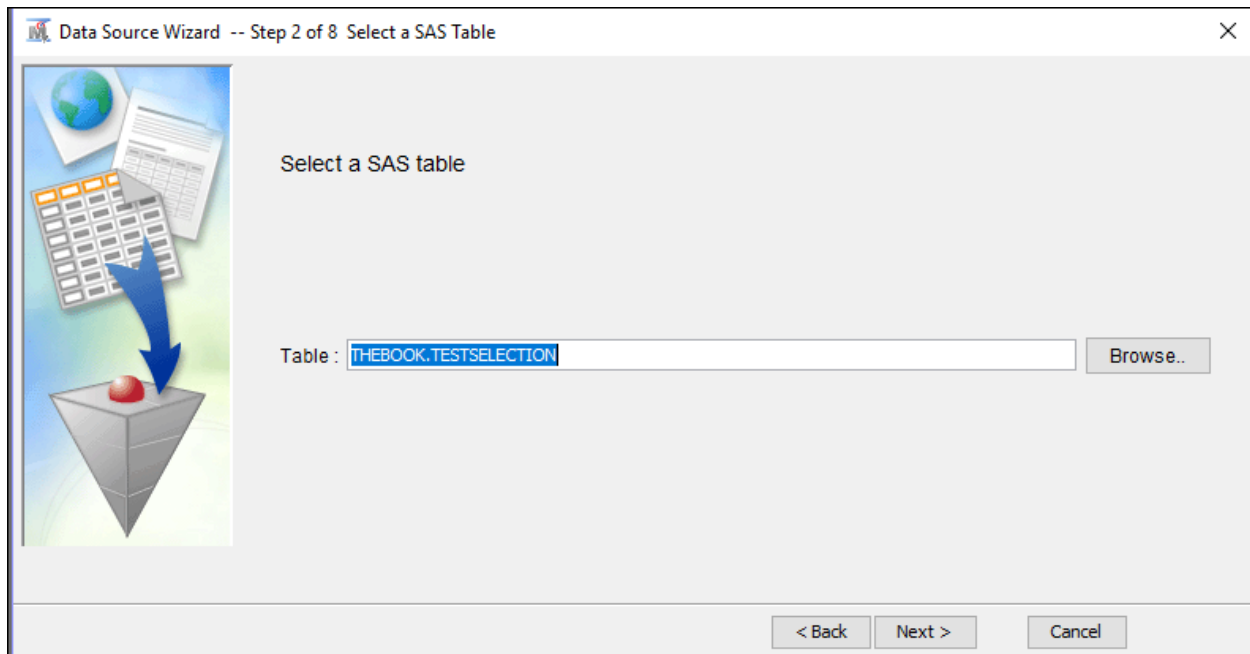
Display 6.6



Display 6.7

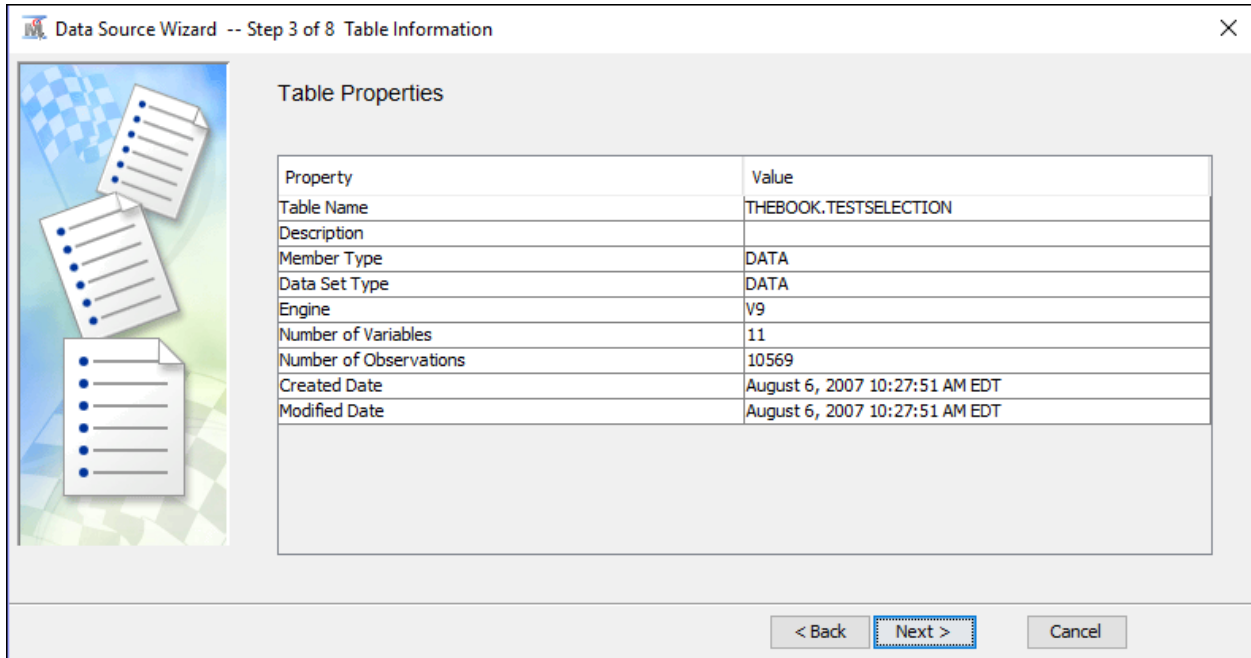


Display 6.8

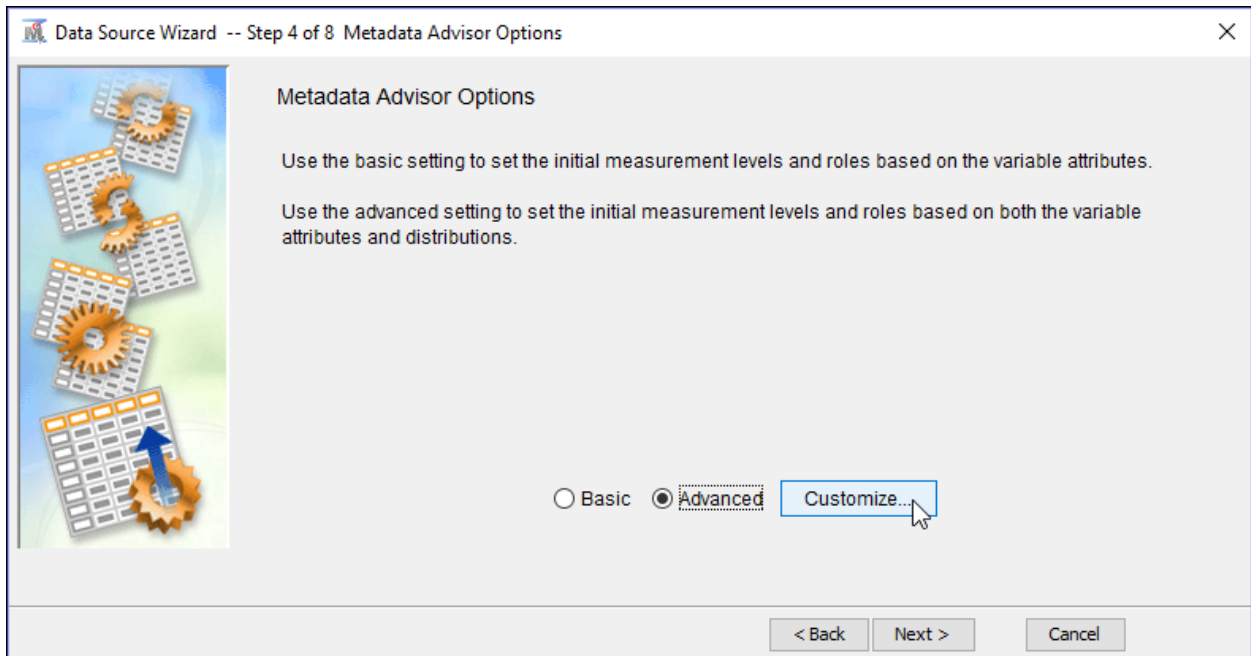




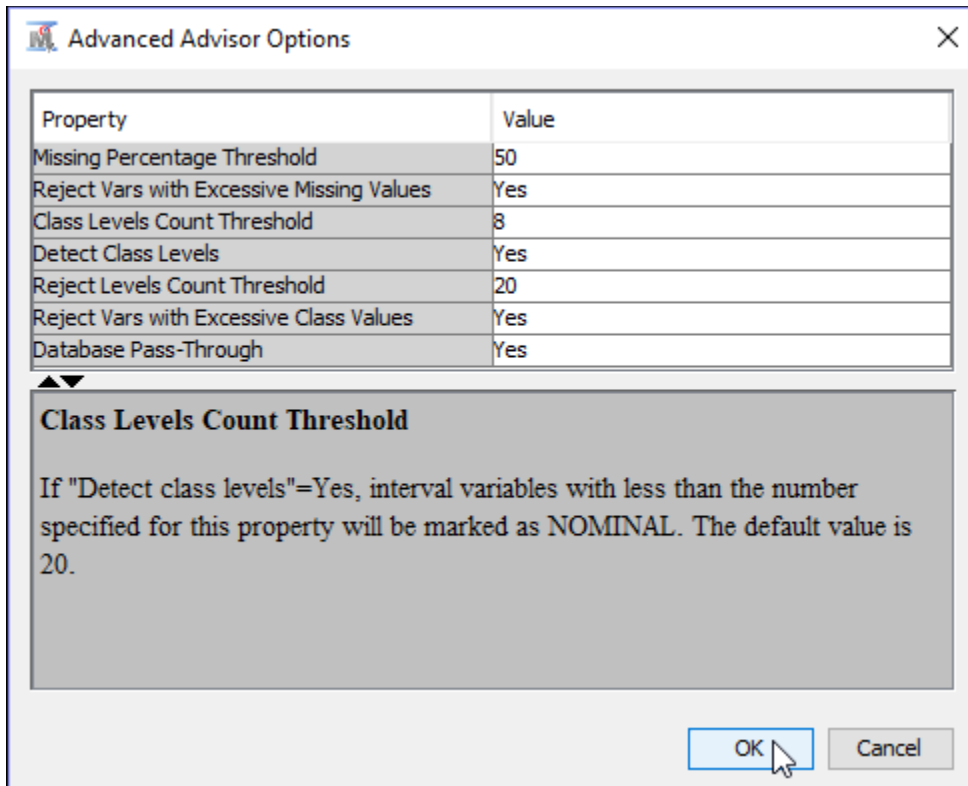
Display 6.9



Display 6.10

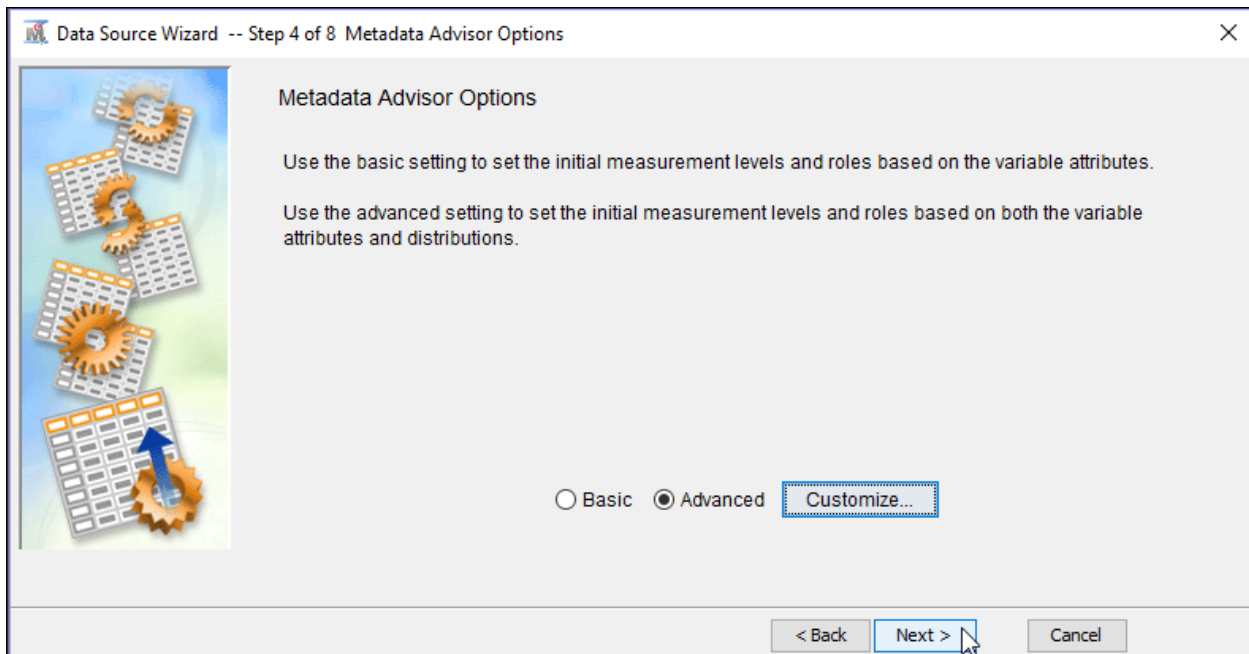


Display 6.11

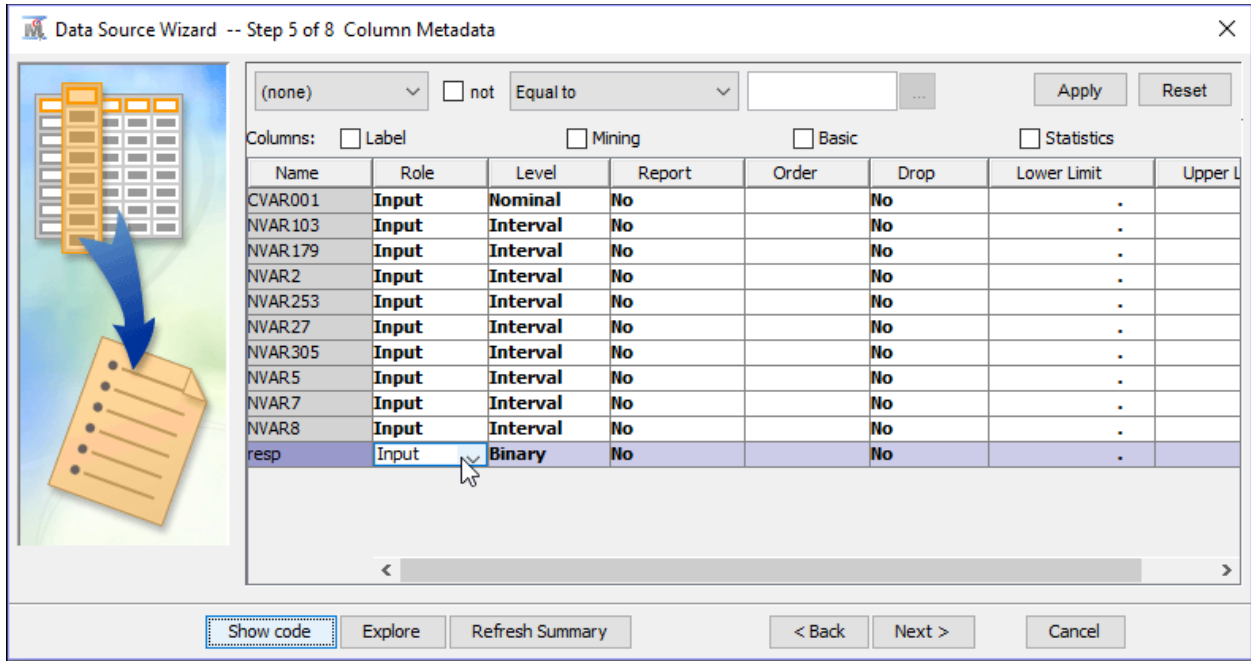


After you type in the value for Class Levels Count Threshold property, you must enter, then click "OK".

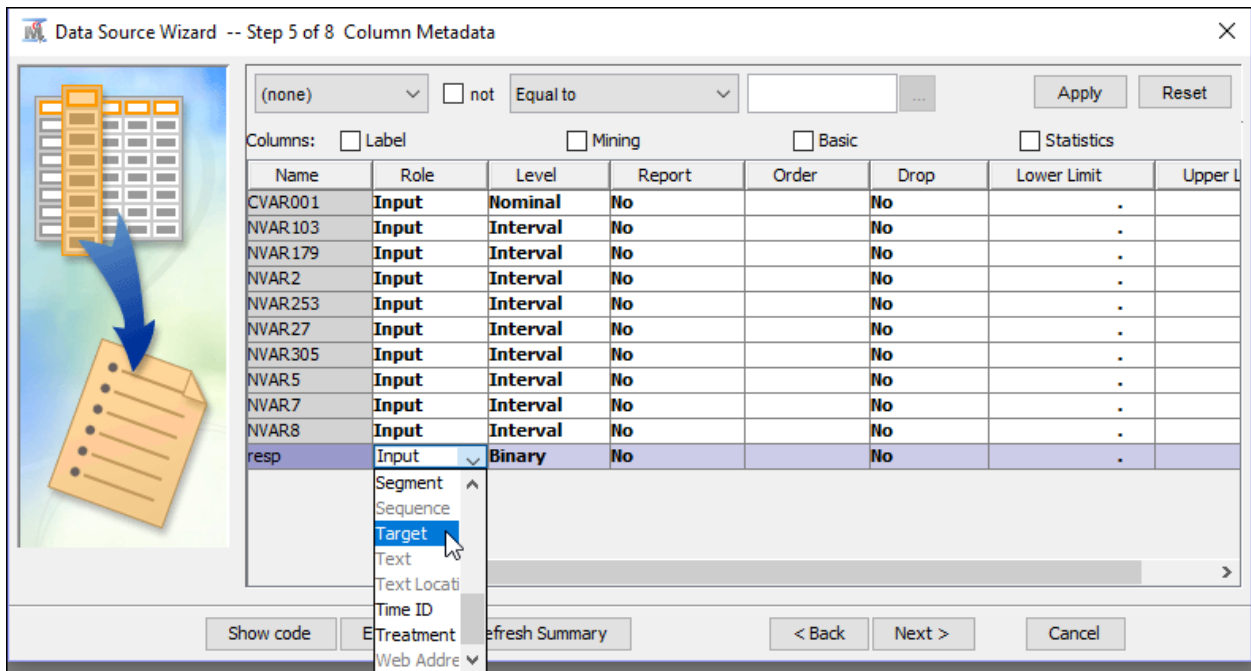
Display 6.12



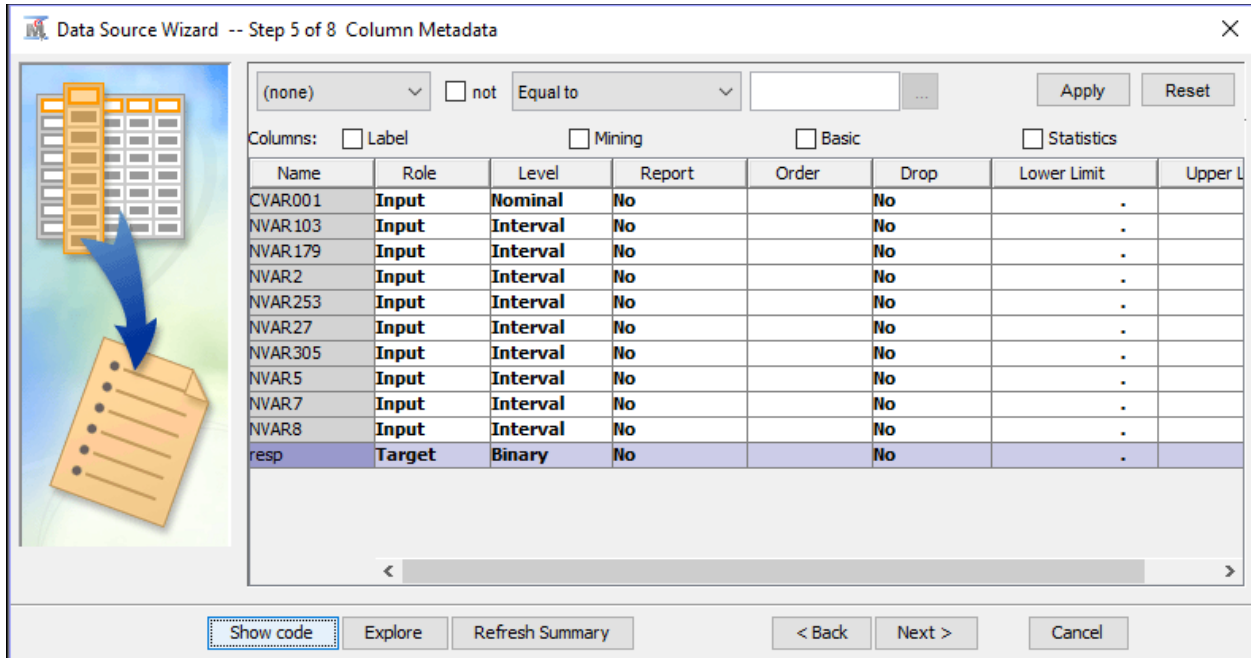
Display 6.13



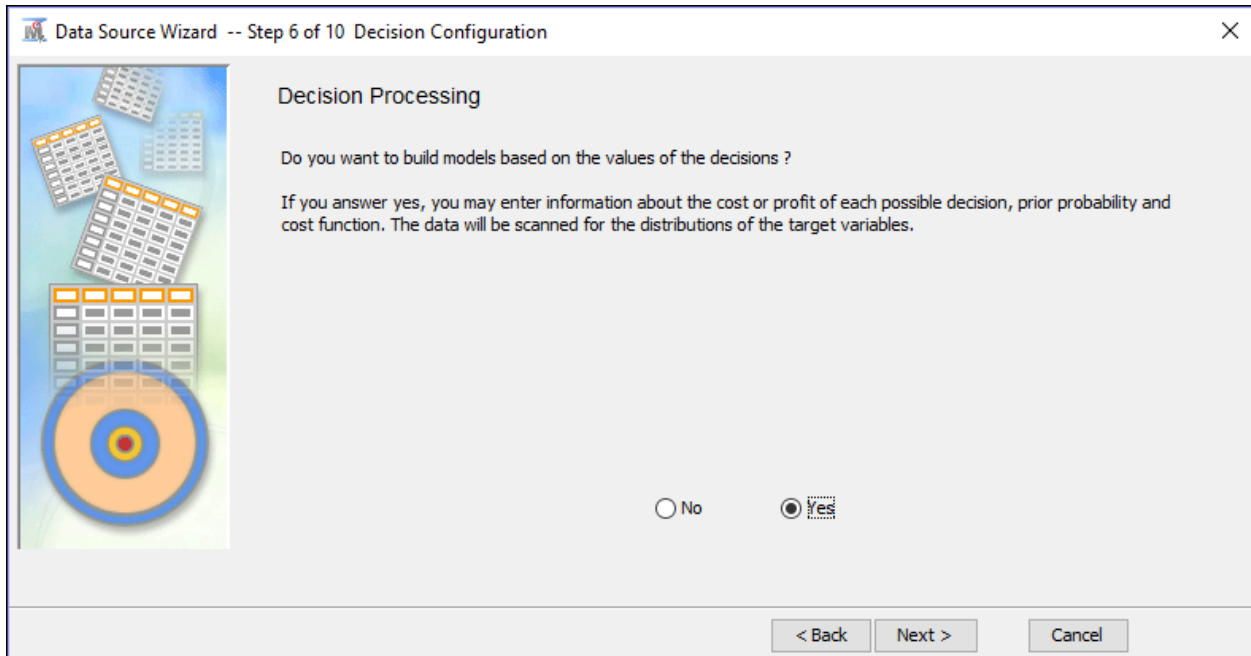
Display 6.14



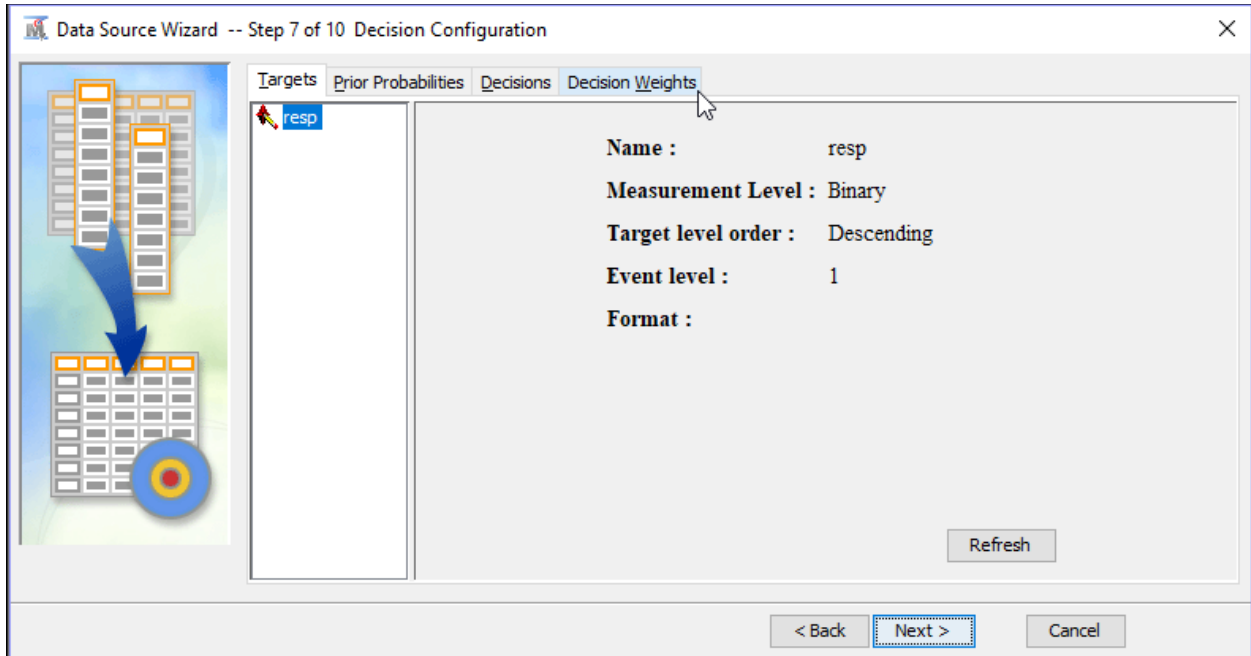
Display 6.15



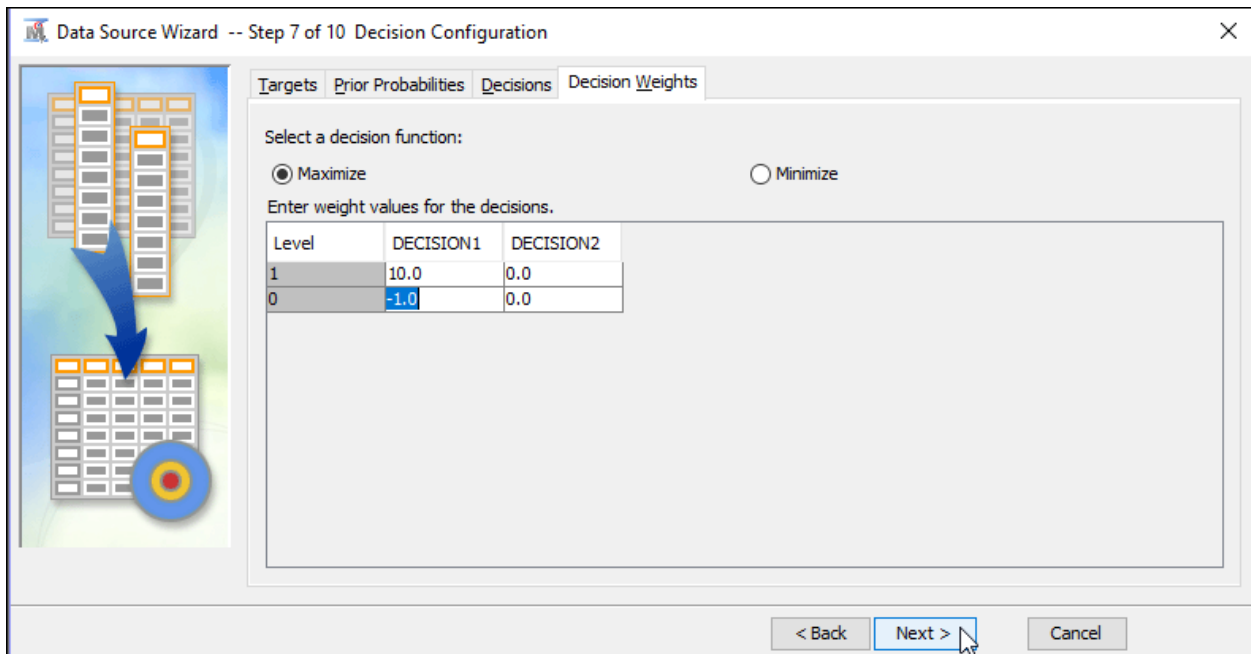
Display 6.16



Display 6.17



Display 6.18



Display 6.19

Data Source Wizard -- Step 8 of 10 Create Sample

Do you wish to create a sample data set?

No  Yes

**Table Info**  
Columns 11  
Rows 10569

**Sample Size**  
Type: Percent  
Percent: 20  
Rows:

< Back Next > Cancel

Display 6.20

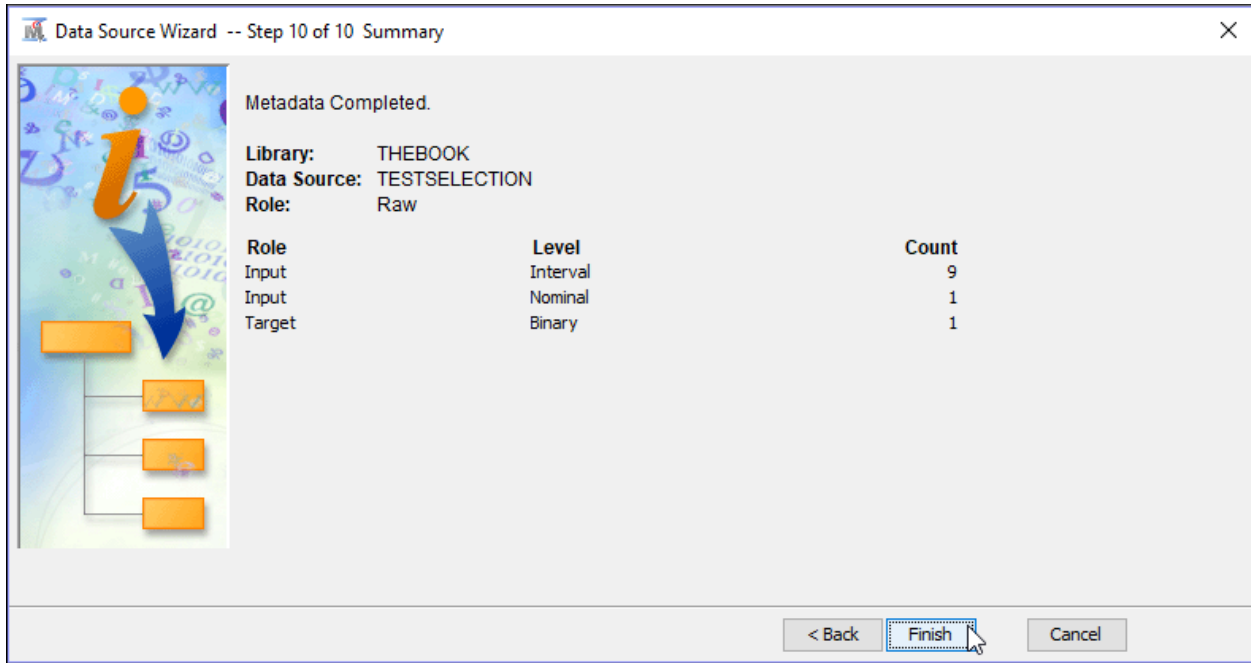
Data Source Wizard -- Step 9 of 10 Data Source Attributes

You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : TESTSELECTION  
Role : Raw  
Segment :  
Notes :

< Back Next > Cancel

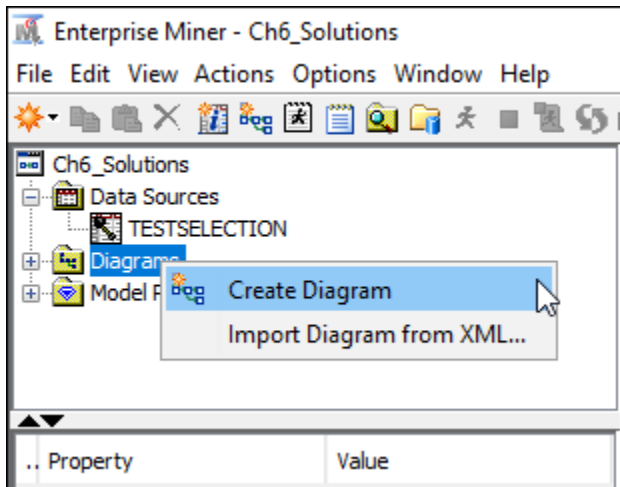
Display 6.21



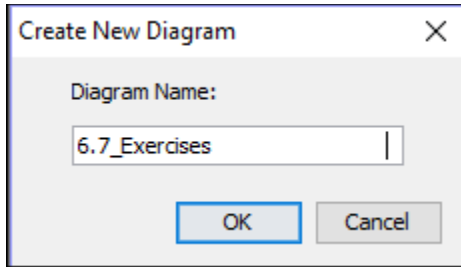
When you click “Finish”, the data source creation will be completed.

Create a Process Flow Diagram.

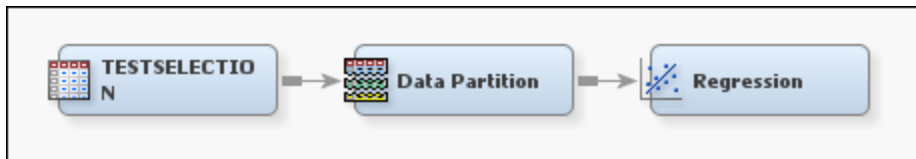
Display 6.22



Display 6.23



Display 6.24



Display 6.25 (Data Parturition Node Properties)

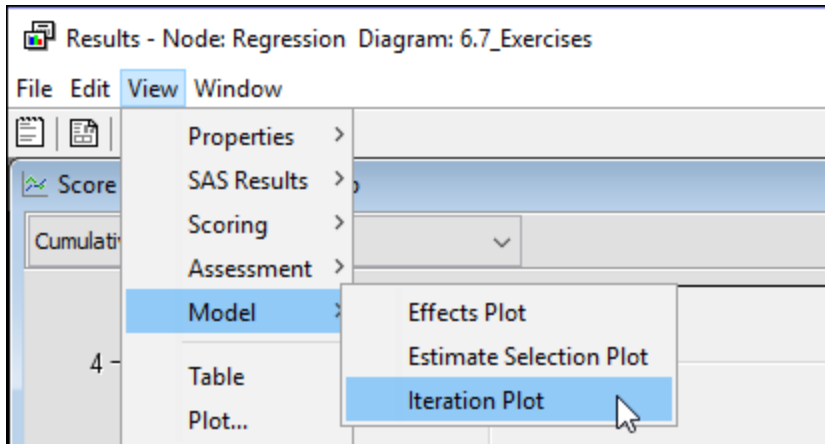
Property	Value
<b>General</b>	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	40.0
Validation	30.0
Test	30.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	12/14/17 4:12 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No



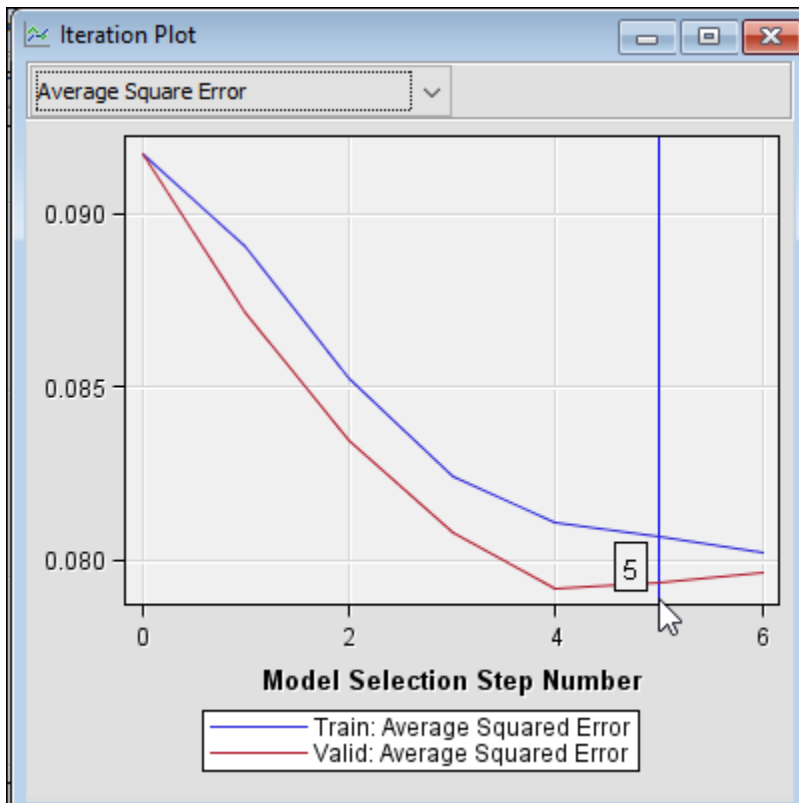
Display 6.26 (Regression Node Properties – a partial view Q3 a, b and c)

Model Selection	
Selection Model	Forward
Selection Criterion	Schwarz Bayesian Criterion (SBC)
Use Selection Defaults	Yes
Selection Options	...

Display 6.27



Display 6.28



### Display 6.29 (from the Output Window)

The selected model, based on the Schwarz Bayesian criterion, is the model trained in Step 5. It consists of the following effects:

```
Intercept  CVAR001  NVAR103  NVAR253  NVAR27  NVAR305
```

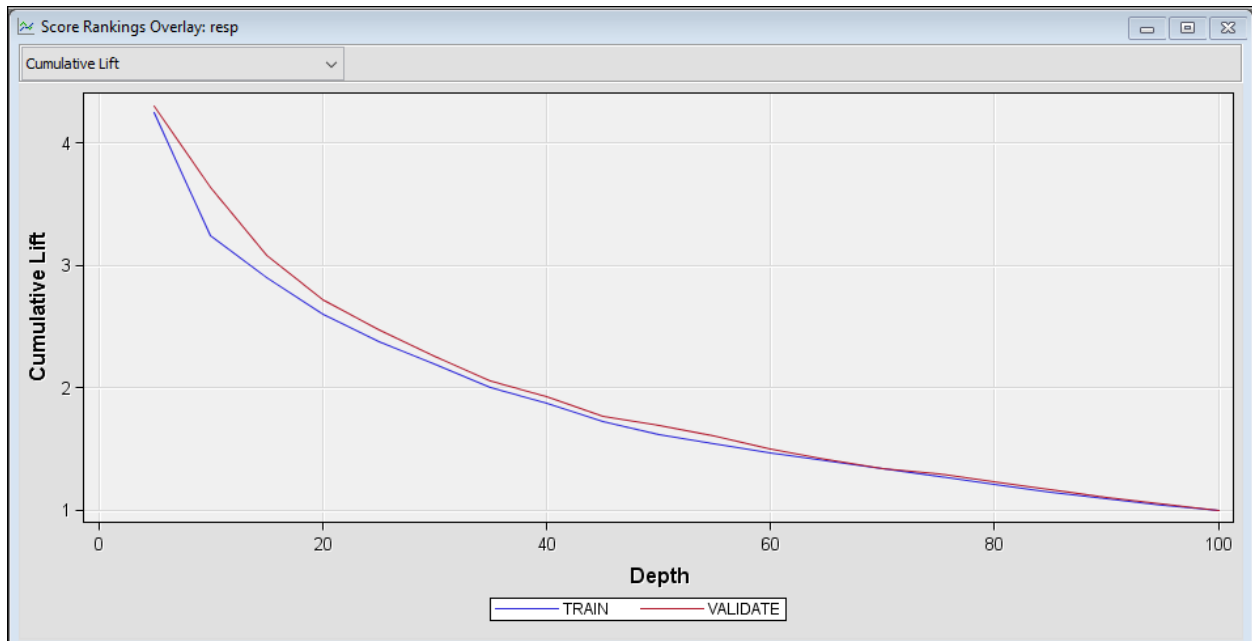
Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood	Likelihood			
Intercept Only	Intercept & Covariates	Ratio	Chi-Square	DF
2163.746	1728.164	435.5818	13	Pr > ChiSq <.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
CVAR001	9	91.9091	<.0001
NVAR103	1	178.4902	<.0001
NVAR253	1	49.8683	<.0001
NVAR27	1	10.6516	0.0011
NVAR305	1	22.3969	<.0001

### Display 6.30

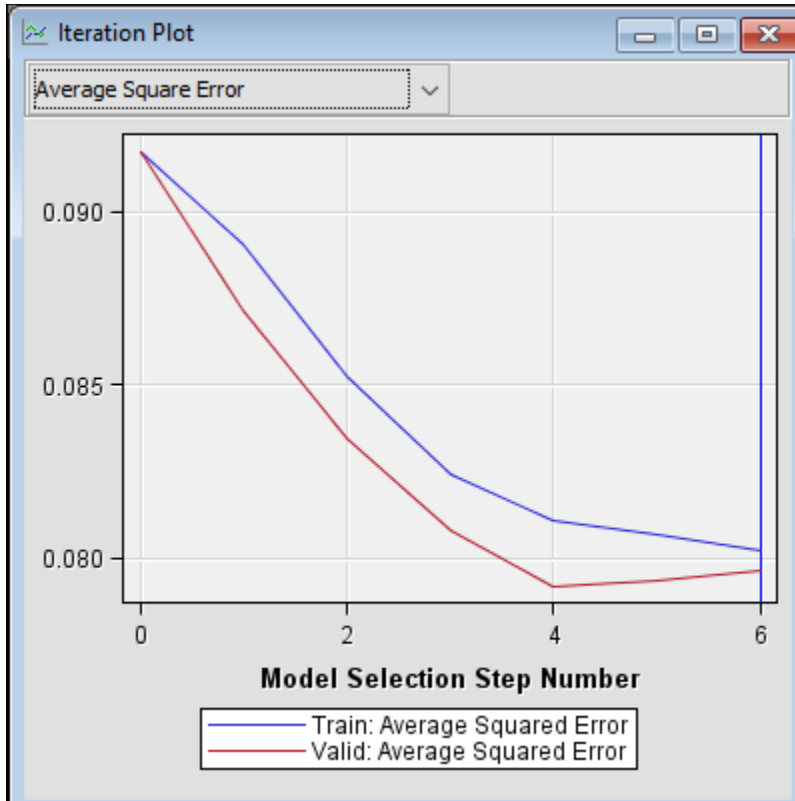


### Question 6

### Display 6.31

Model Selection	
Selection Model	Forward
Selection Criterion	Akaike Information Criterion (AIC)
Use Selection Defaults	Yes

Display 6.32



Display 6.33

The selected model, based on the Akaike information criterion, is the model trained in Step 6. It consists of the following effects:

Intercept CVAR001 NVAR103 NVAR253 NVAR27 NVAR305 NVAR7

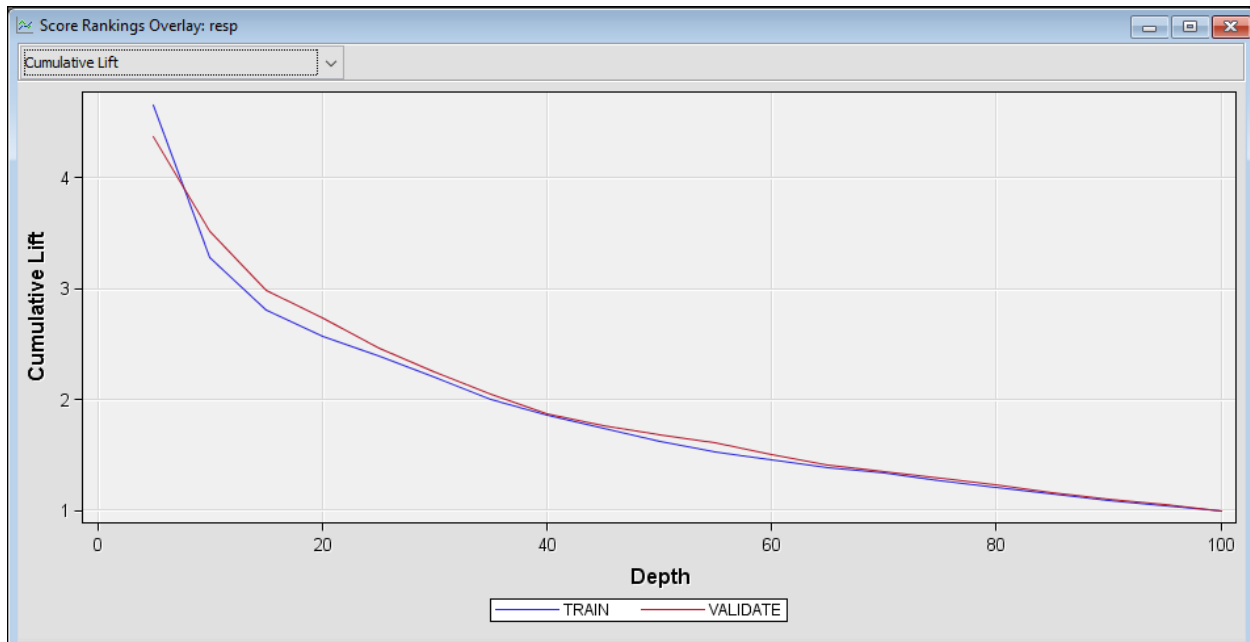
Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood	Likelihood			
Intercept Only	Intercept & Covariates	Ratio	Chi-Square	DF Pr > ChiSq
2163.746	1719.165	444.5805	14	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
CVAR001	9	91.5441	<.0001
NVAR103	1	176.8076	<.0001
NVAR253	1	50.4588	<.0001
NVAR27	1	13.3377	0.0003
NVAR305	1	18.5836	<.0001
NVAR7	1	9.3743	0.0022

Display 6.34



From the iteration plot in Display 6.28 you can see that the Schwarz Bayesian Criterion (SBC) selected 5 explanatory variables selected at the 5<sup>th</sup> iteration. The selected variables are: CVAR001, NVAR103, NVAR253, NVAR27 and NVAR305. By the Akaike Information Criterion (AIC) 6 explanatory variables are selected (See Display 6.32)– the 5 variables selected by the SBC criterion plus an additional variables NVAR7. The reason for this the SBC criterion imposes a larger penalty for each additional explanatory variable than the AIC criterion does. This can be seen from equations (6.29) and (6.30) given in pages 406 and 408 in the third edition of the book. You should also look at the output window and scroll down to see which variable is added at each iteration. Displays 6.29 and 6.33 show the variables selected at the final iteration for SBC and AIC.

Displays 6.30 and 6.34 show lift charts for the models generated by the SBC and AIC. There is no significant difference between the lift. But the model produced by SBC is parsimonious in the sense that it has fewer explanatory variables.

## Chapter 7

Display 7.1

The screenshot shows a dialog box titled "Create New Project -- Step 1 of 2 Specify Project Name and Server Directory". On the left is a blue sidebar with the SAS Enterprise Miner 14.3 logo. The main area contains the instruction: "Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location." Below this are two input fields: "Project Name" with the value "Ch7\_Solutions" and "SAS Server Directory" with the value "C:\TheBook\EM14.3\EMProjects". A "Browse" button is next to the directory field. At the bottom are buttons for "< Back", "Next >", and "Cancel".

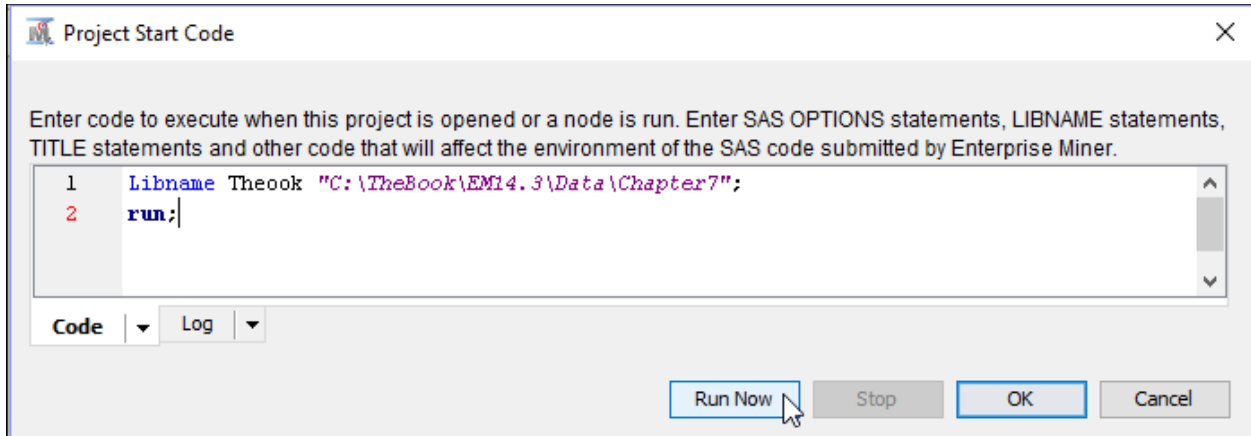
Display 7.2

The screenshot shows a dialog box titled "Create New Project -- Step 2 of 2 New Project Information". On the left is a blue sidebar with the SAS Enterprise Miner 14.3 logo. The main area contains a table titled "New Project Information" with the following data:

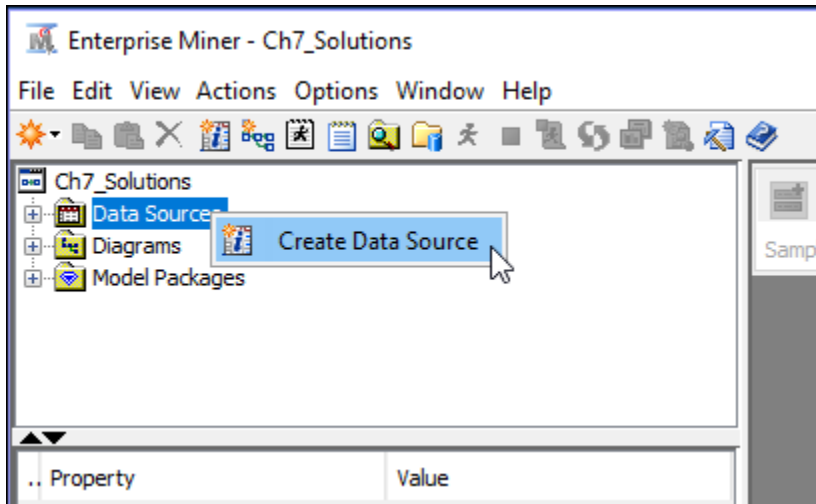
New Project Information	
Name	Ch7_Solutions
Server Directory	C:\TheBook\EM14.3\EMProjects

At the bottom are buttons for "< Back", "Finish", and "Cancel".

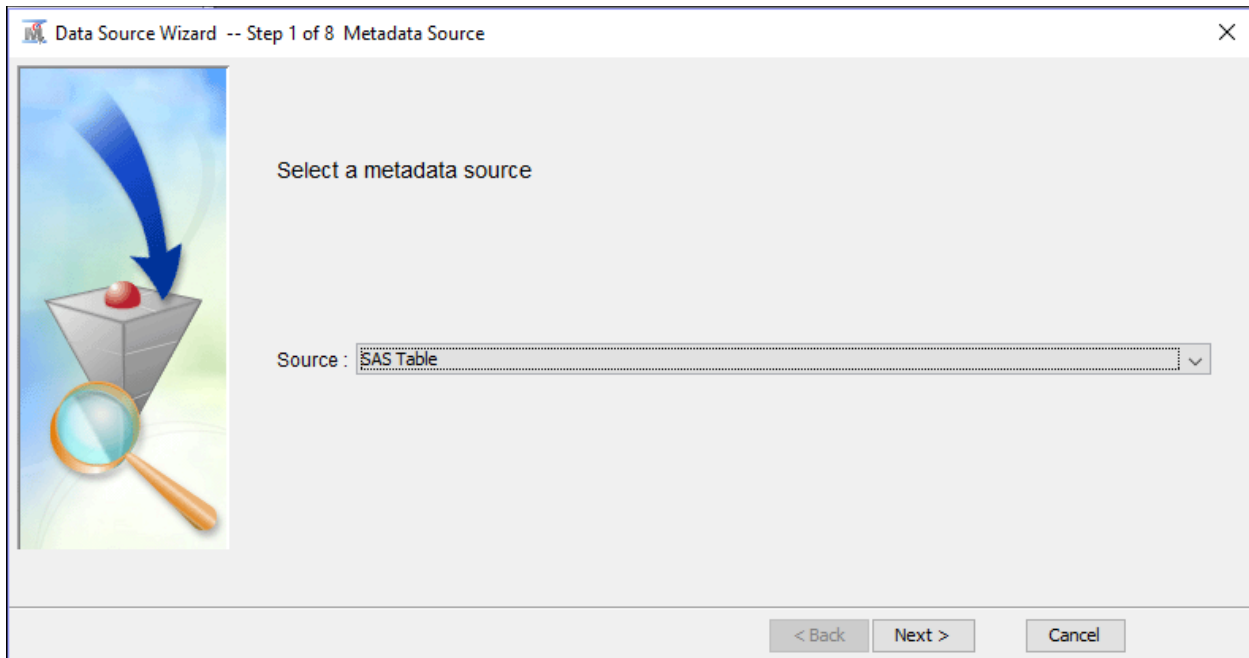
Display 7.3



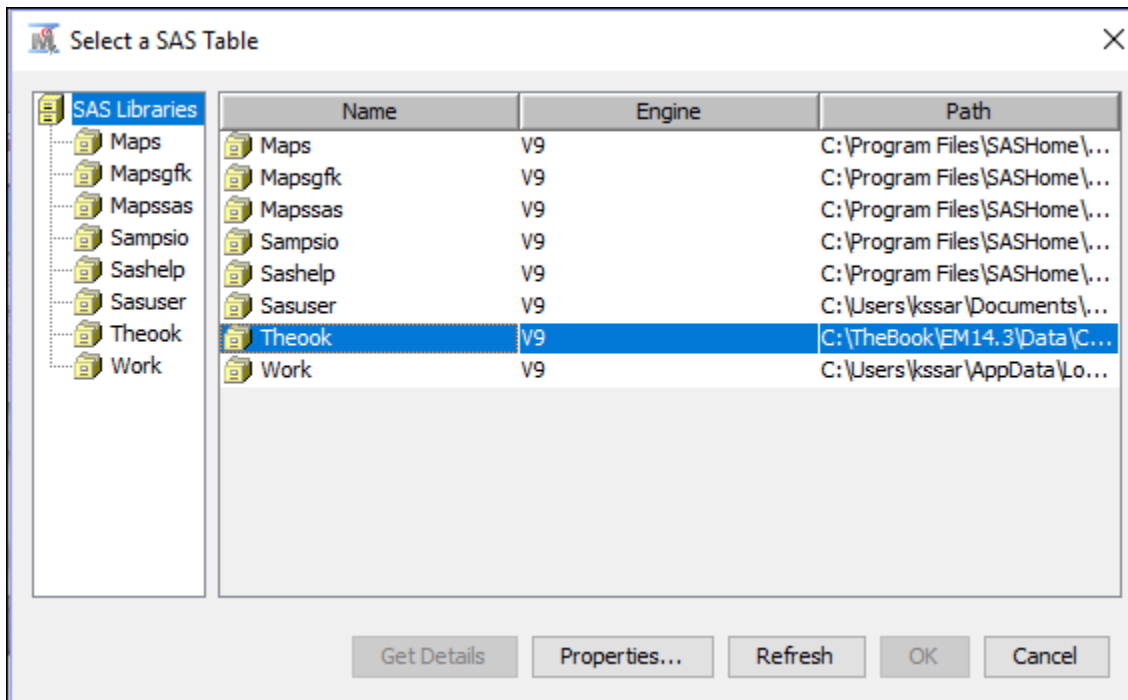
Display 7.4



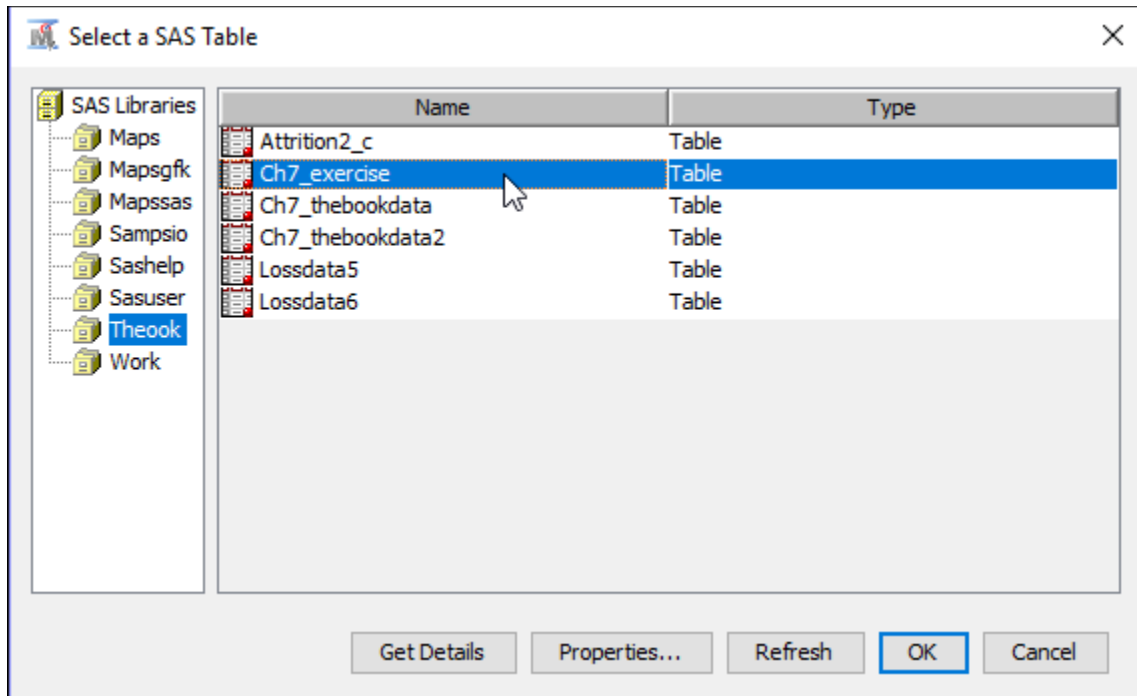
Display 7.5



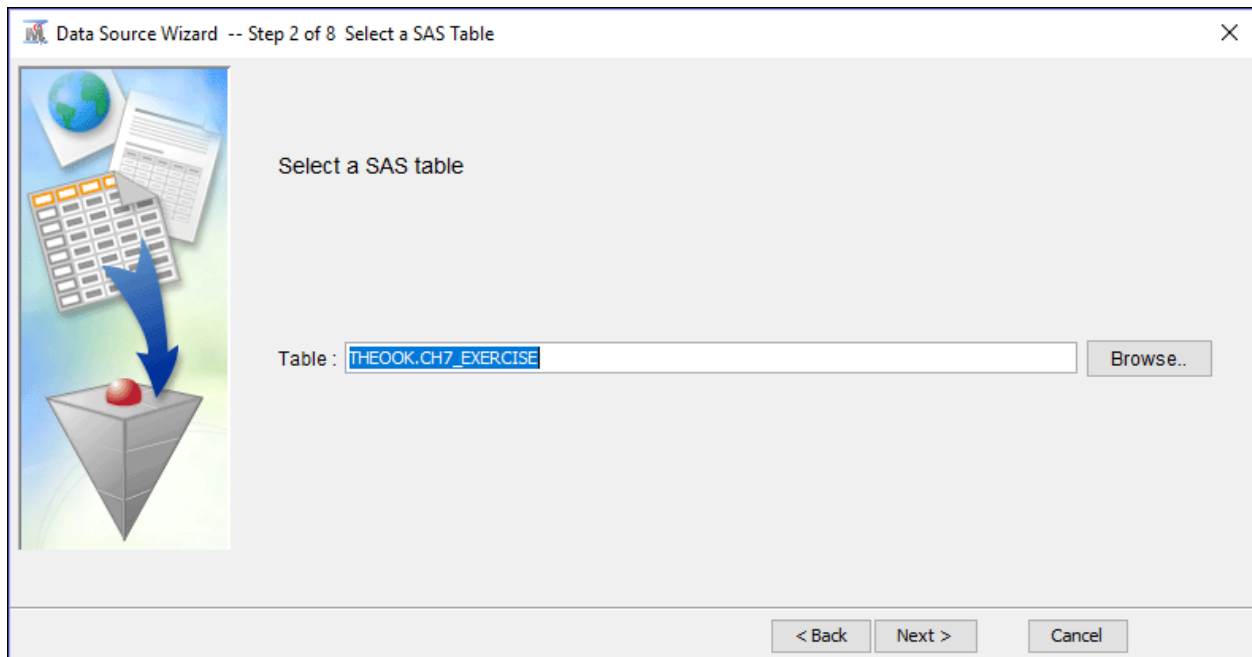
Display 7.6



Display 7.7

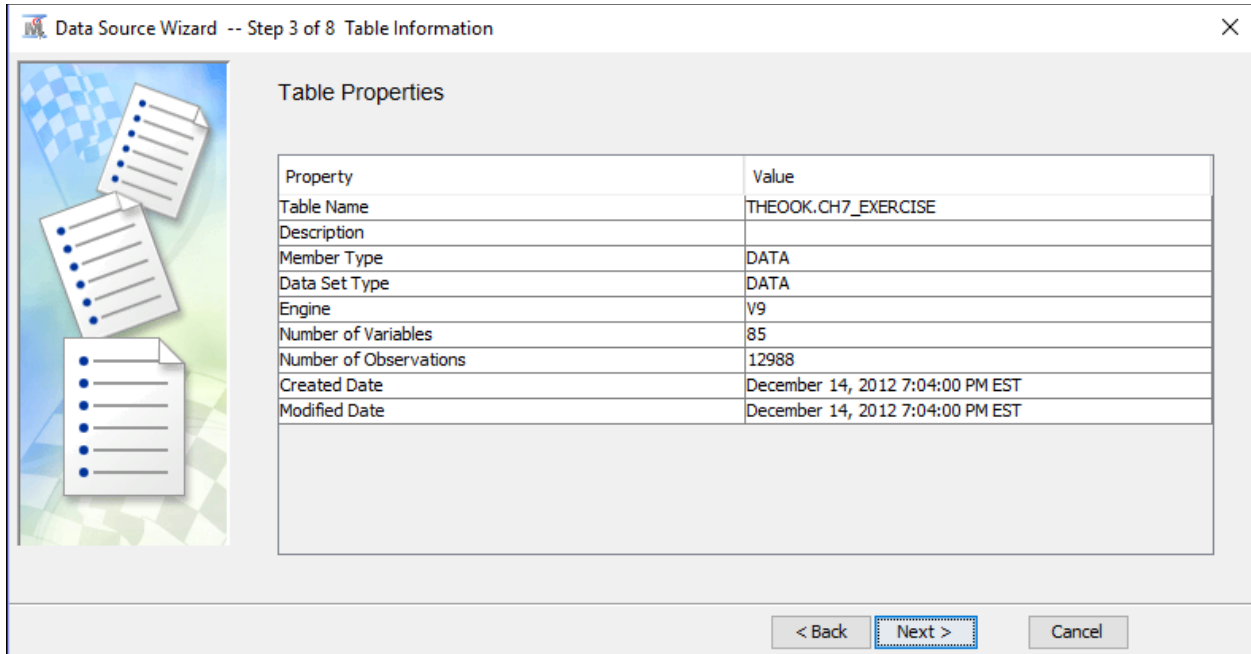


Display 7.8

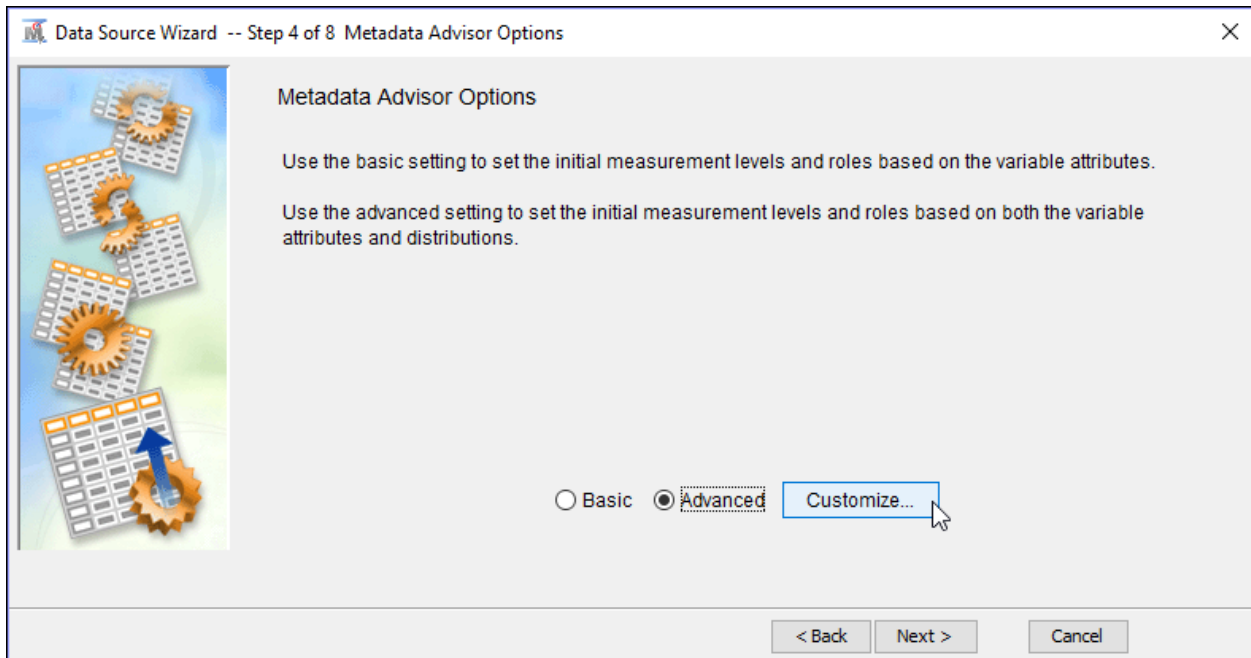




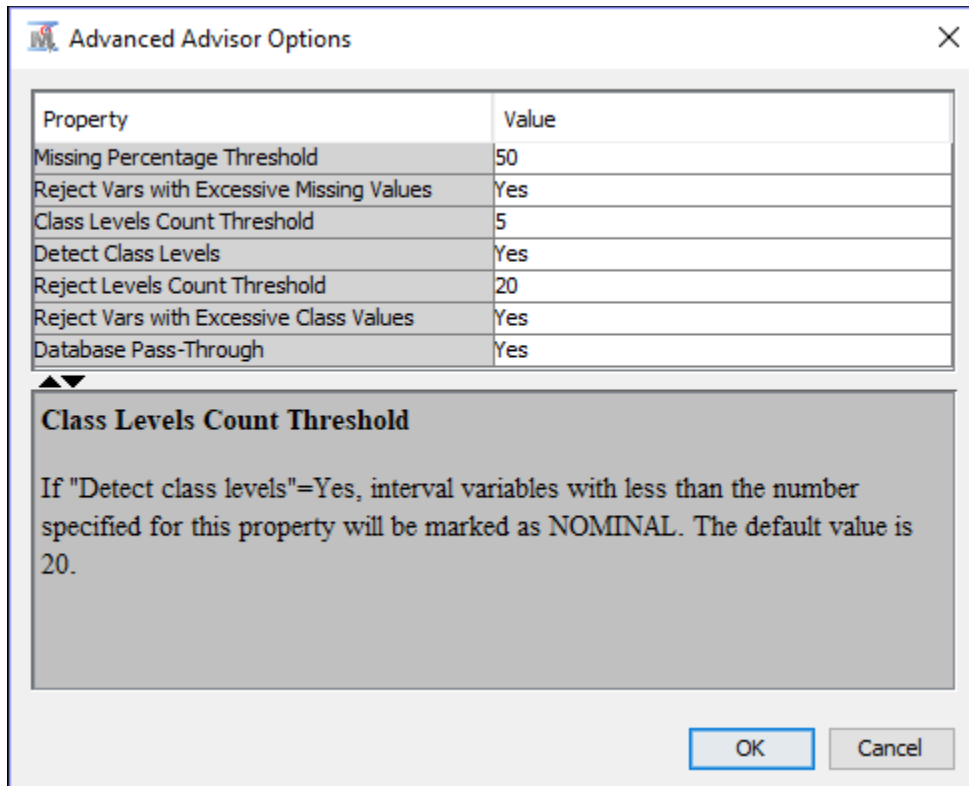
Display 7.9



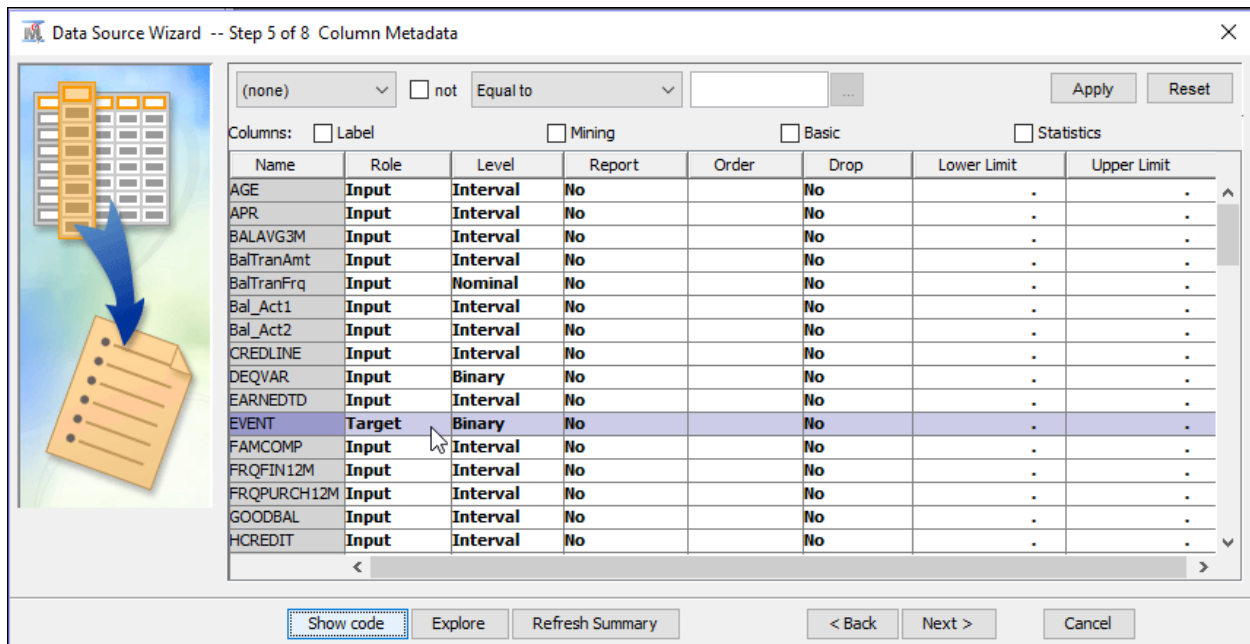
Display 7.10



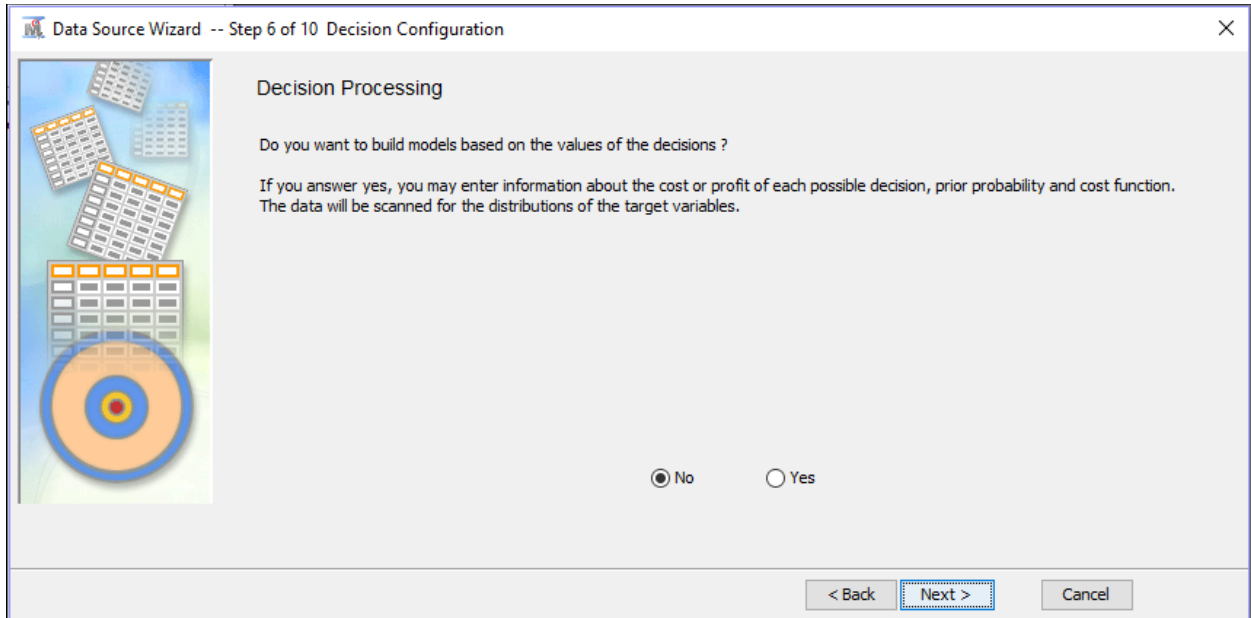
Display 7.11



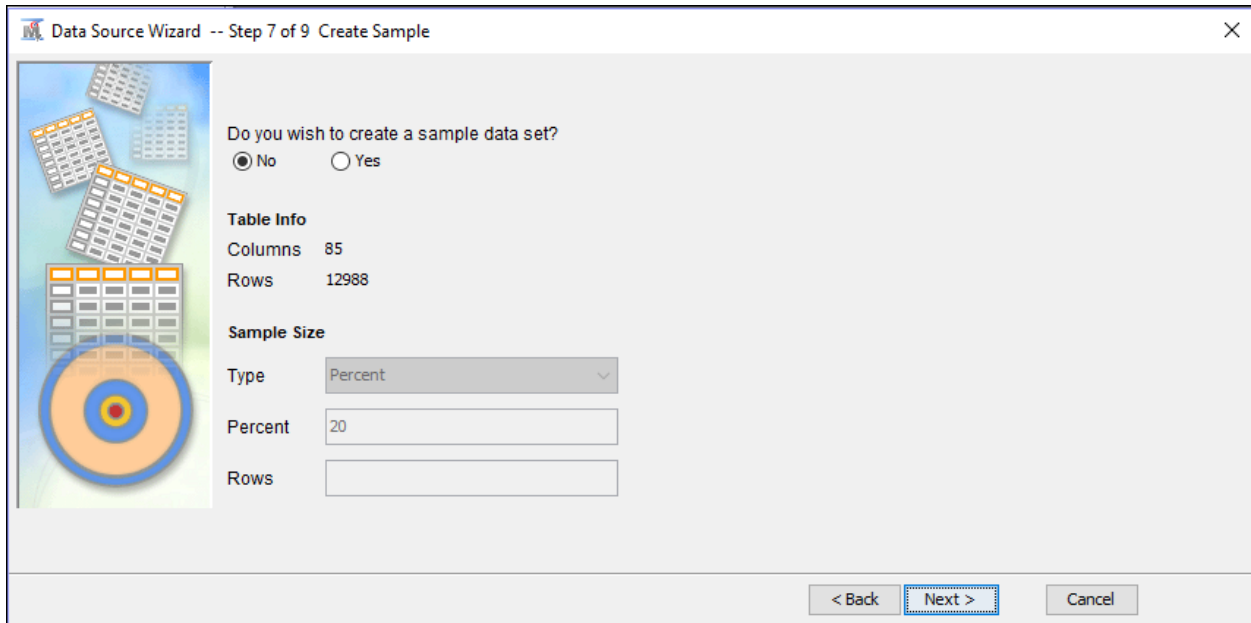
Display 7.12



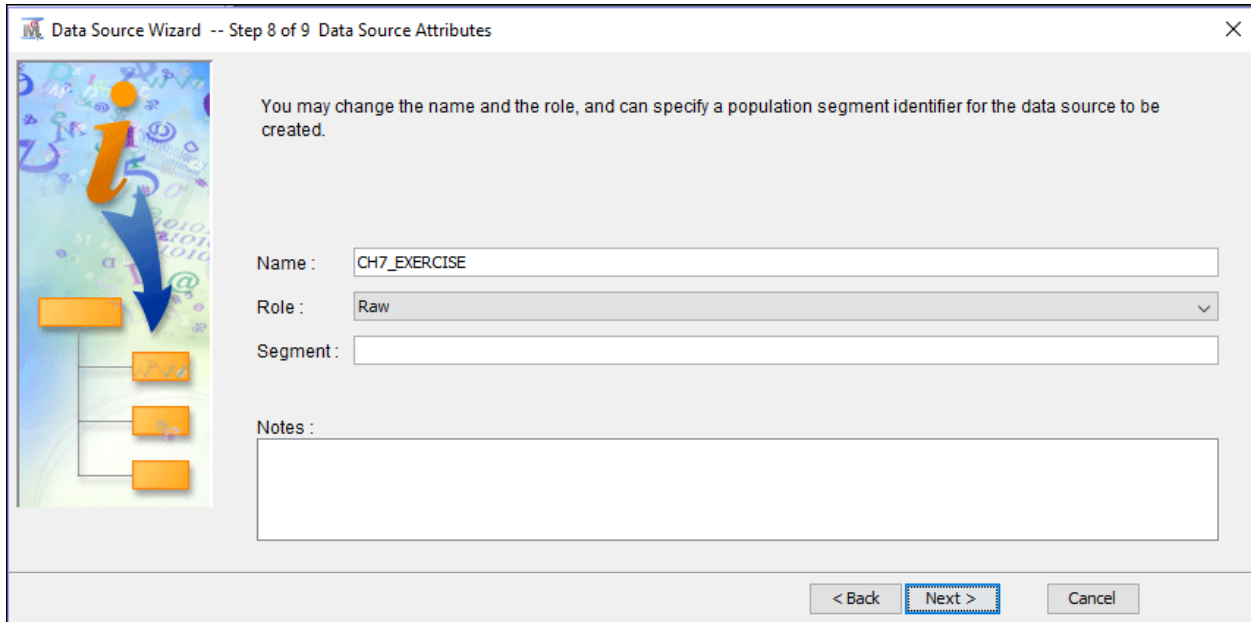
Display 7.13



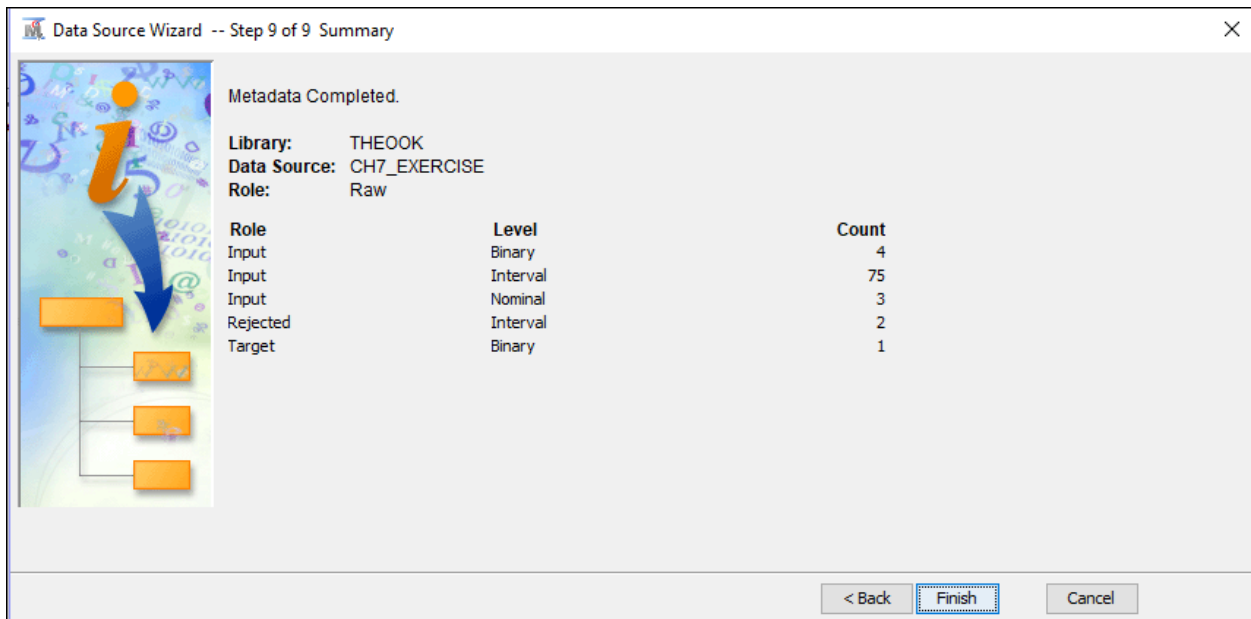
Display 7.14



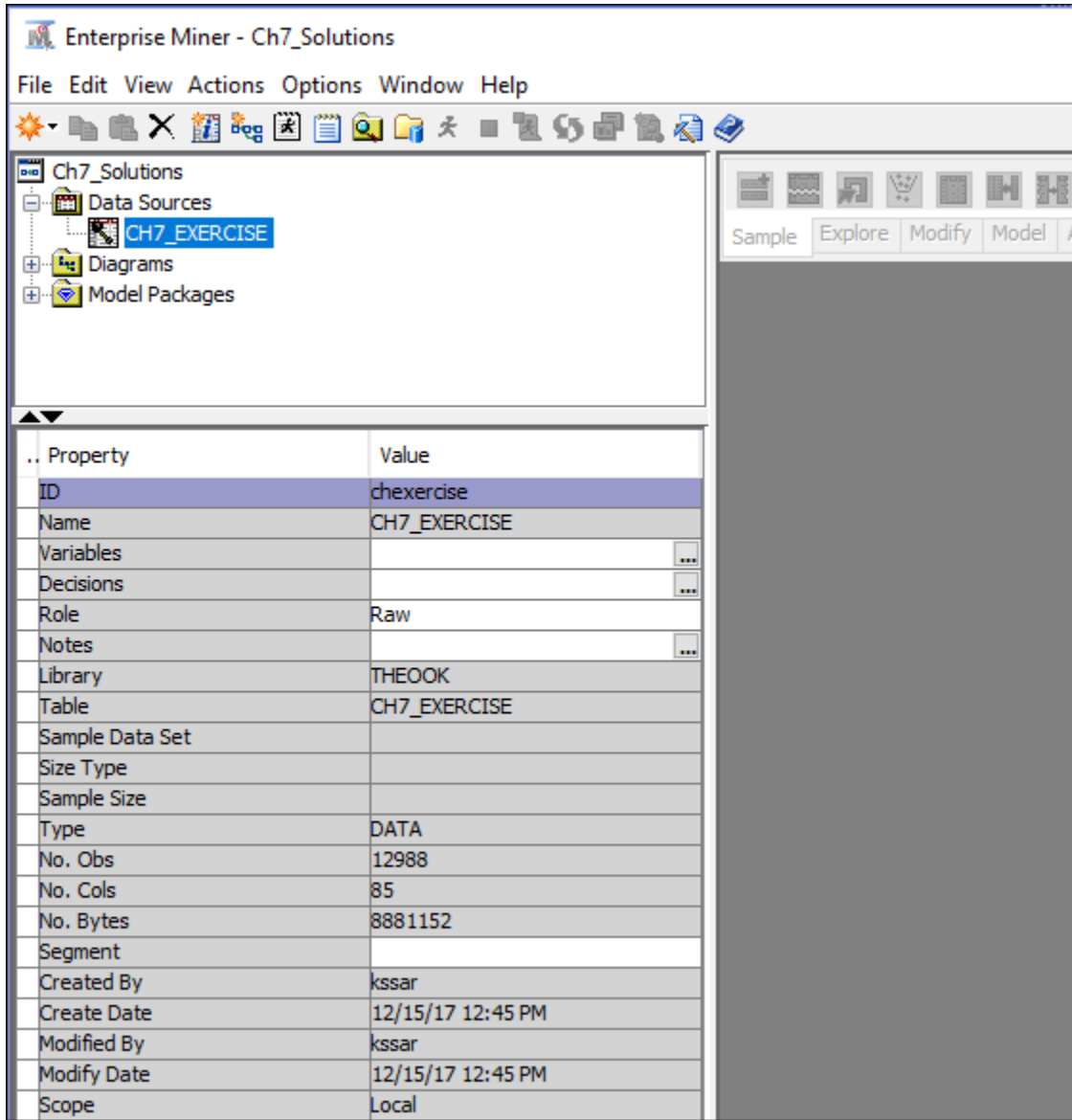
Display 7.15



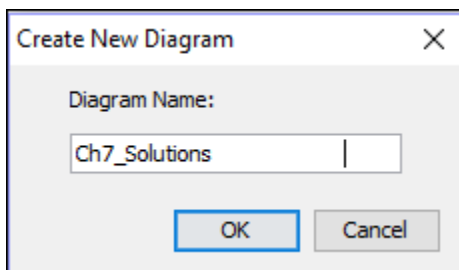
Display 7.16



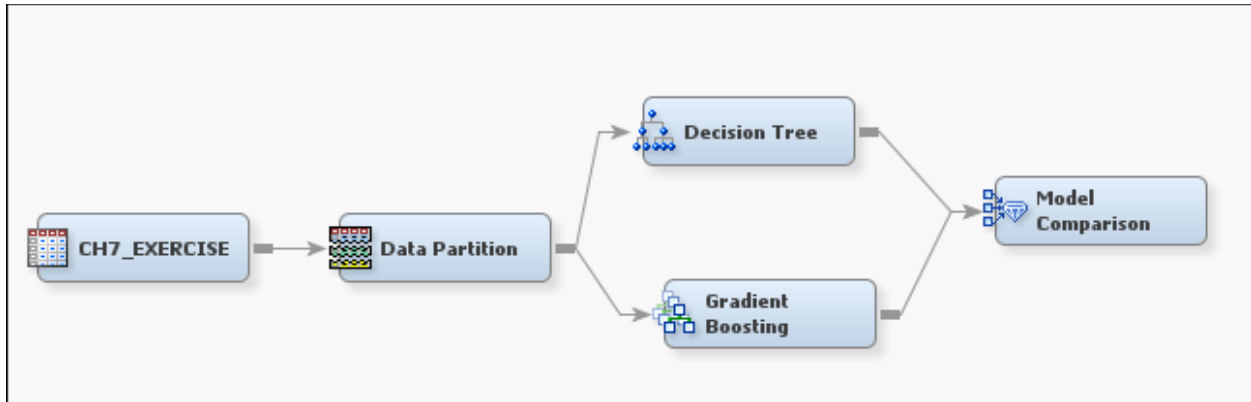
Display 7.17



Display 7.18



Display 7.19



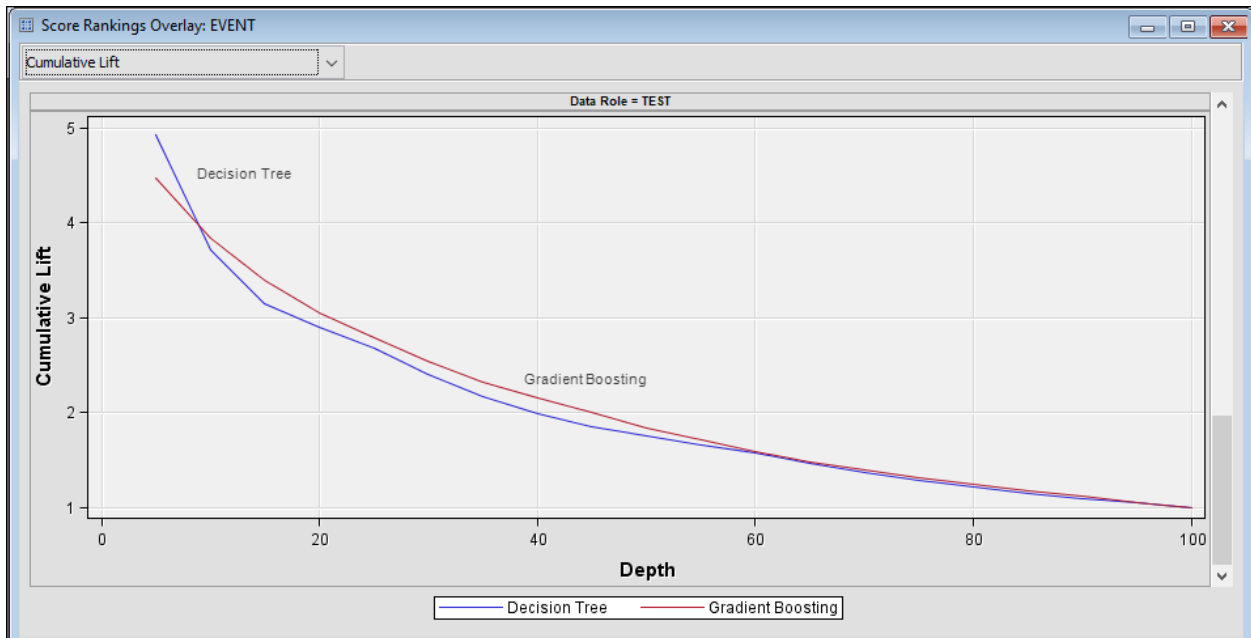
Display 7.20 (Properties of the data partition node)

Property	Value
<b>General</b>	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	50.0
Validation	30.0
Test	20.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	12/15/17 6:02 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

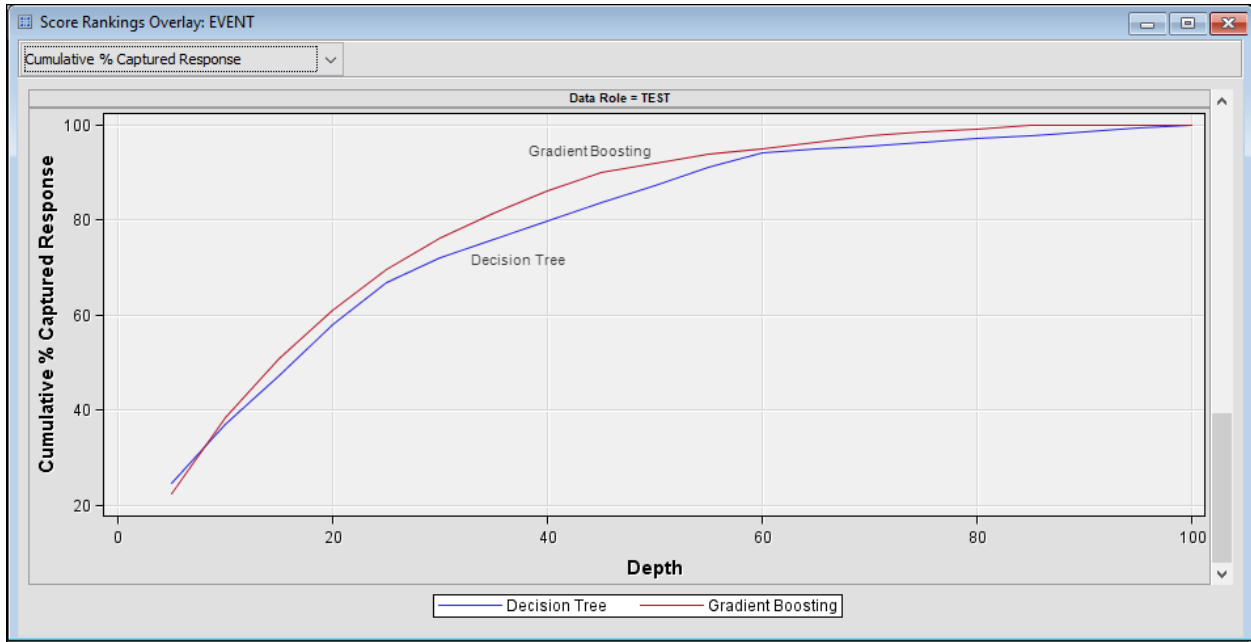
Display 7.21

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Display 7.22



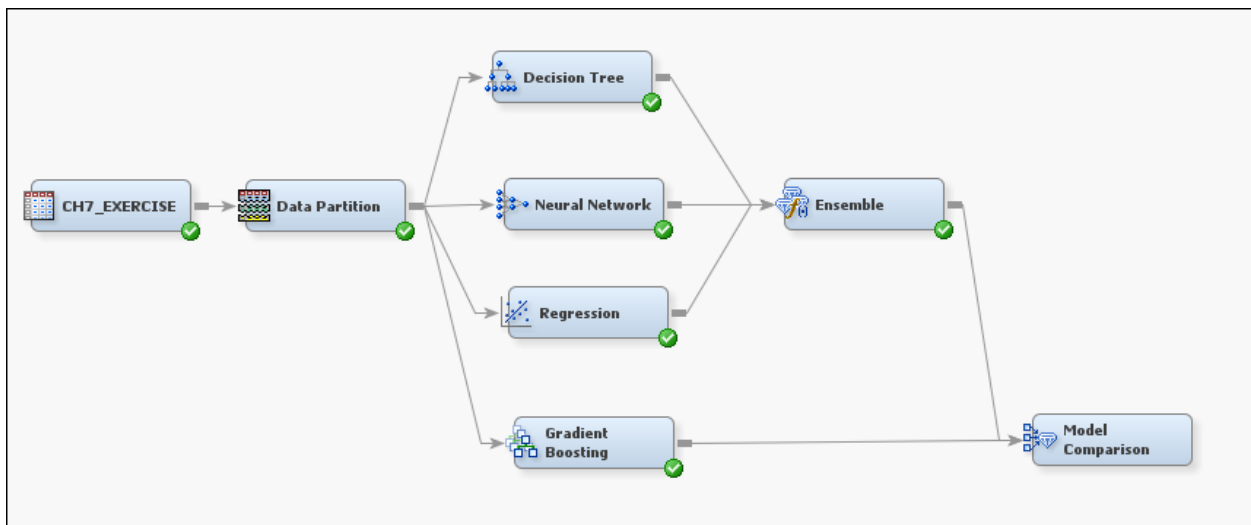
Display 7.23



Comparing the cumulative % captured response of the two models shown in Display 7.23, Gradient Boosting seems to be slightly better than Decision Tree.

Exercise 2

Display 7.24





Display 7.25 shows the properties of the data partition node. Display 7.26 shows the Subtree properties of the Decision Tree node.

Display 7.25

Property	Value
<b>General</b>	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	50.0
Validation	30.0
Test	20.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	12/16/17 7:32 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Display 7.26

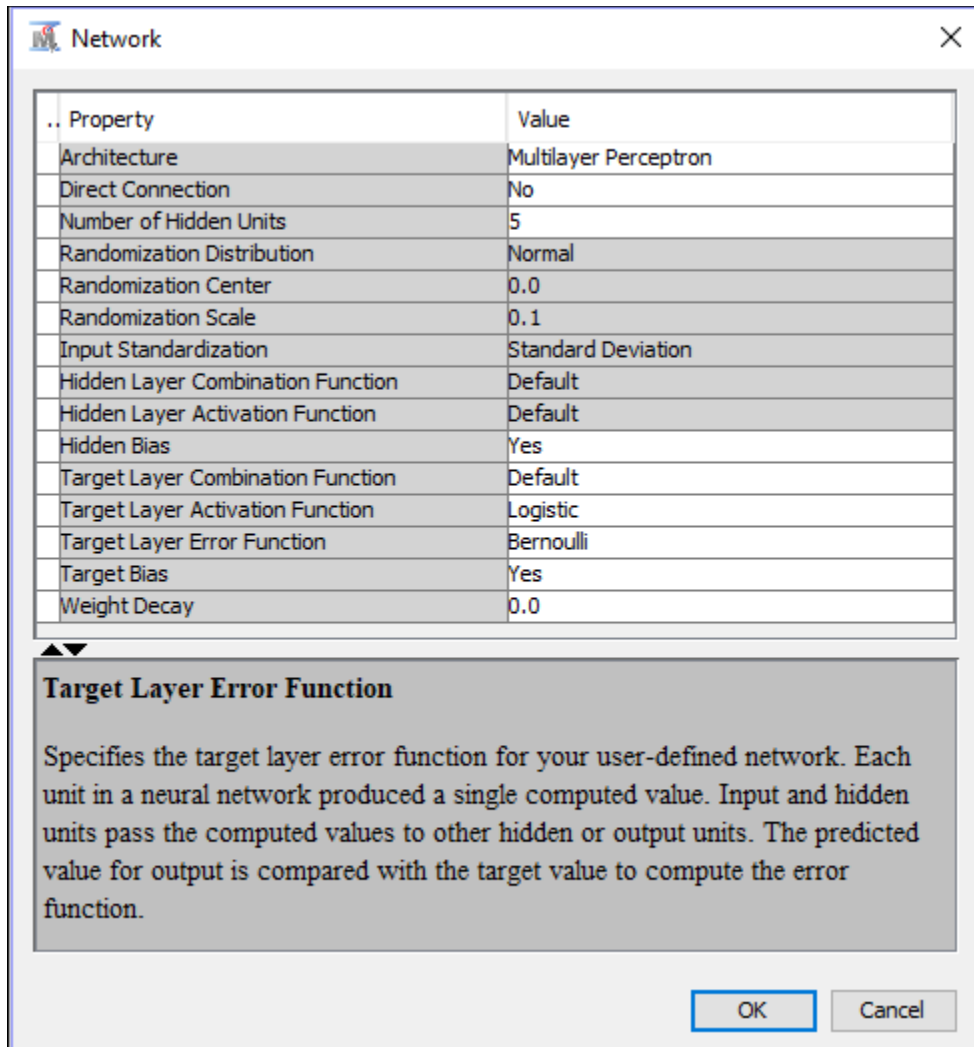
<b>Subtree</b>	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
<b>Cross Validation</b>	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
<b>Observation Based Importance</b>	
Observation Based Importance	No
Number Single Var Importance	5

Display 7.27 shows the properties of the Neural Network model and 7.28 shows the Network architecture. Display 7.29 shows the optimization parameters for the Neural Network.

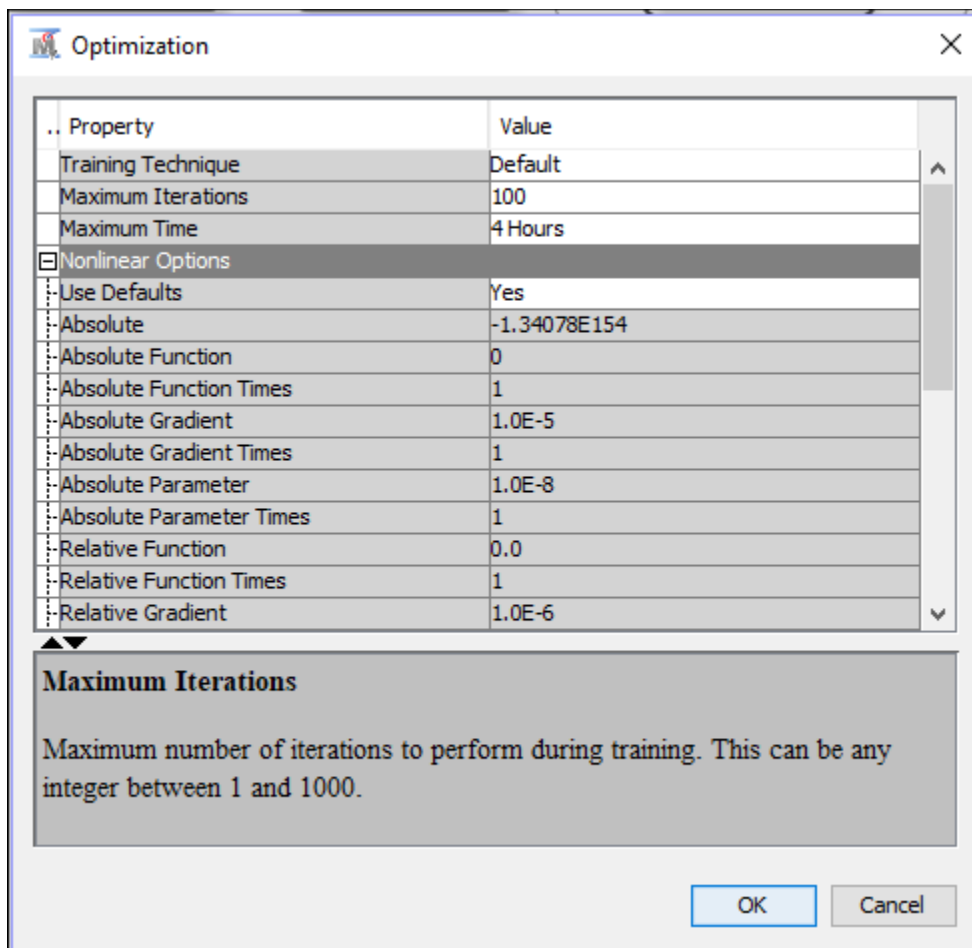
Display 7.27

.. Property	Value
<b>General</b>	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No
<b>Score</b>	
Hidden Units	No
Residuals	Yes
Standardization	No
<b>Status</b>	
Create Time	12/16/17 7:43 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Display 7.28

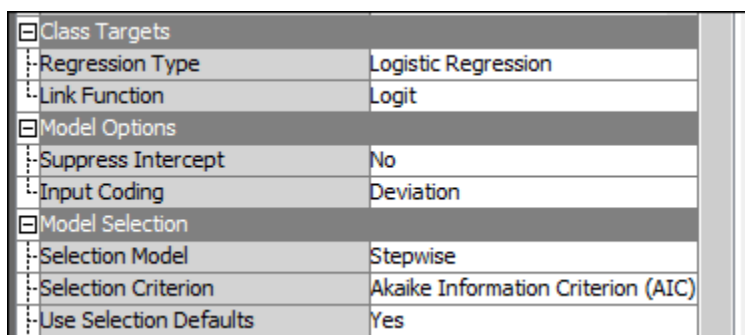


Display 7.29

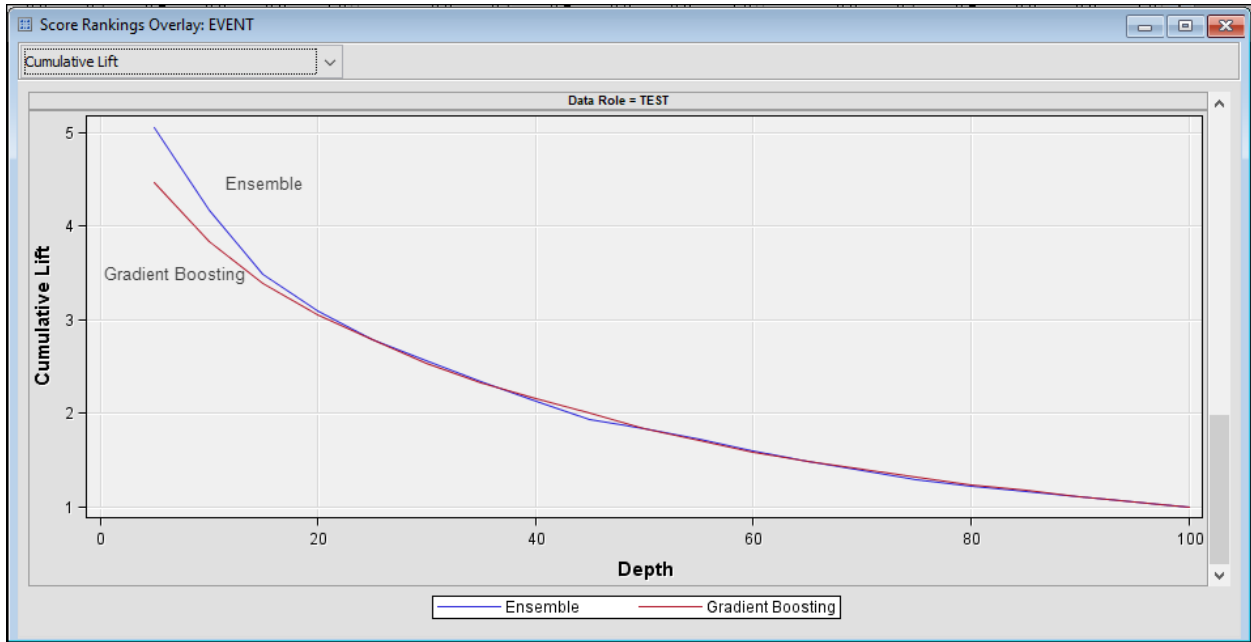


Display 7.30 shows the properties of the Regression node.

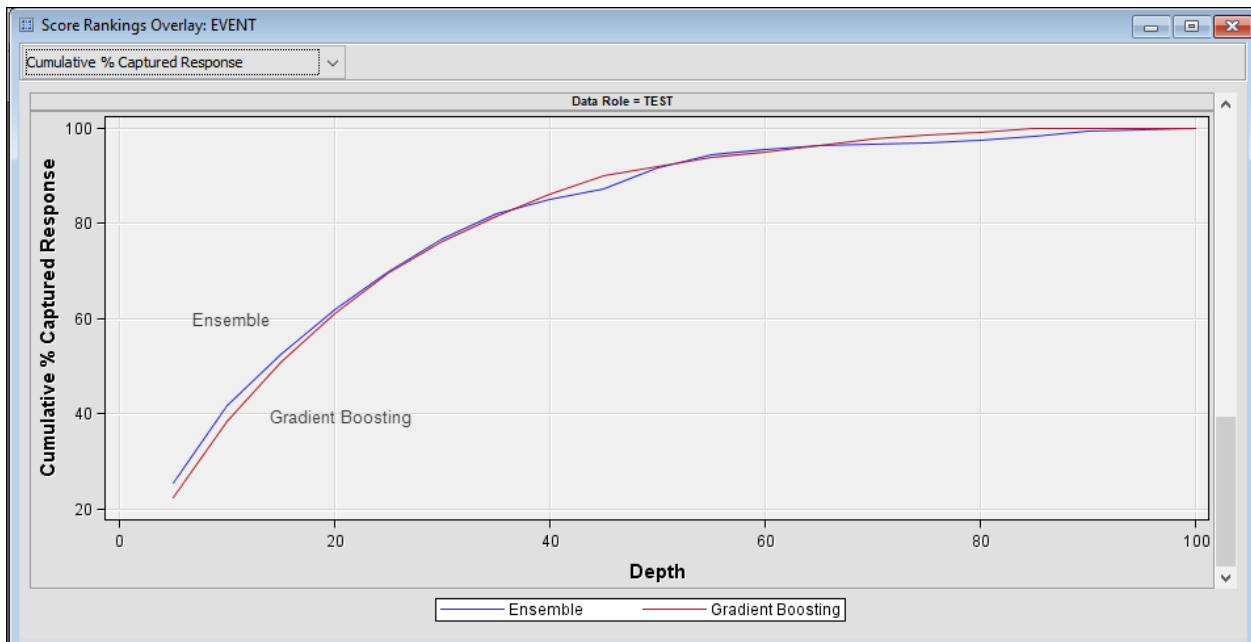
Display 7.30



Display 7.31

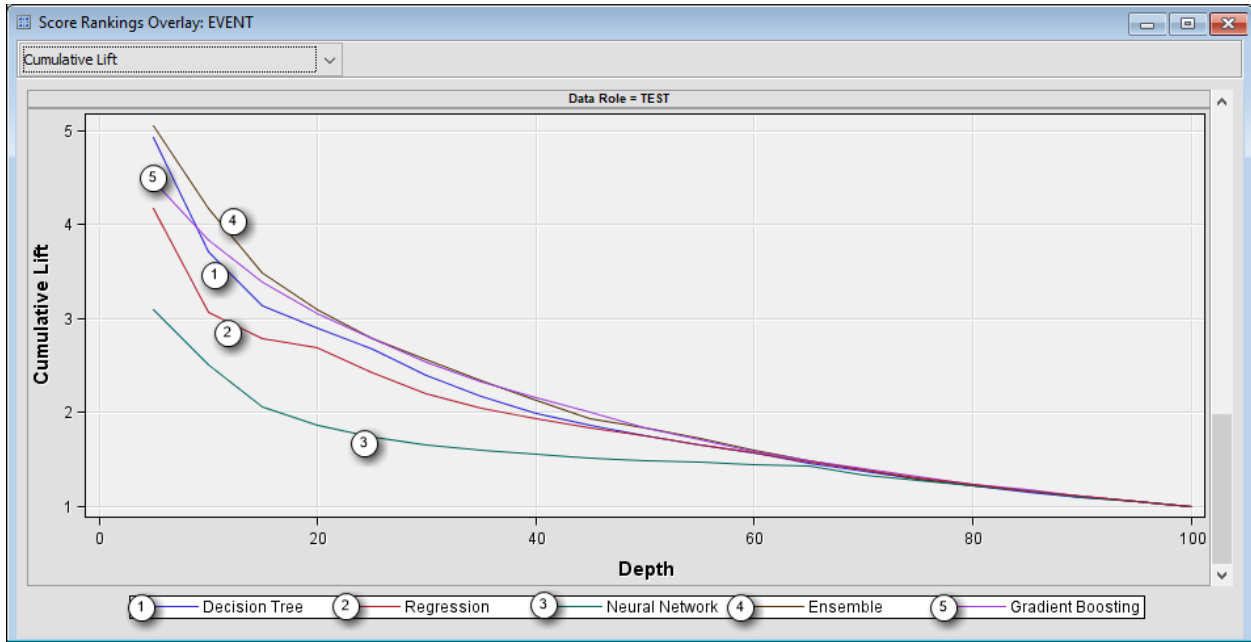


Display 7.32

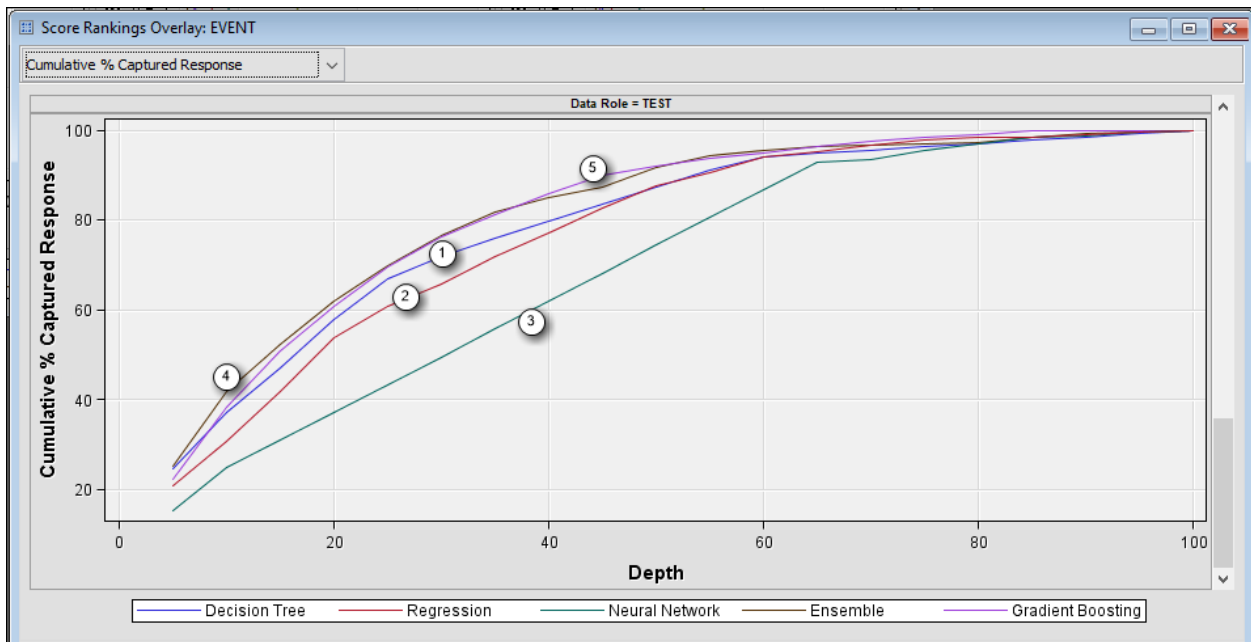


From displays 7.31 and 7.32, it appears that Ensemble is slightly better than Gradient Boosting in terms of model accuracy.

Display 7.33



Display 7.34



Based on the cumulative lift charts (Display 7.33) and Cumulative %Captured Response (Displays 7.34), Ensemble is the best model, although the difference between the Ensemble and Gradient Boosting is very small.

## Chapter 9

Using %tmfilter macro we downloaded web pages from

- (1) [www.wsj.com/news/economy](http://www.wsj.com/news/economy)
- (2) [www.webmd.com](http://www.webmd.com)
- (3) [www.nytimes.com/economy](http://www.nytimes.com/economy)

The %tmfilter macro is executed separately for each link shown in (1), (2) and (3) above. After running the %tmfilter macro three times, three sas data sets are created: (1) "tmlib.wsj", (2) "tmlib.webmd" and (3) tmlib.nytbusiness.

The links selected here are arbitrary. They are used for illustration only. You can do the same with links of your own interest.

Display 9.1

```
libname tmlib "C:\TheBook\EM13.1\TextMining\SASDATA";  
run;
```

Display 9.2

```
%tmfilter(url=http://www.wsj.com/news/economy,  
depth=1,  
dir=c:\TheBook\EM14.3\TextMining\dir1,  
destdir=c:\TheBook\EM14.3\TextMining\destdir1,  
norestrict=1,  
dataset=tmlib.wsj,numchars=32000,force=Y);
```

Display 9.3

The SAS System				
The FREQ Procedure				
Language				
LANGUAGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
English	114	100.00	114	100.00
Frequency Missing = 6				

Frequency Percent Row Pct Col Pct	Table of TRUNCATED by OMITTED			
	TRUNCATED(Truncated)	OMITTED(Omitted)		
		0	1	Total
0	111 92.50 96.52 95.69	4 3.33 3.48 100.00	115 95.83	
1	5 4.17 100.00 4.31	0 0.00 0.00 0.00	5 4.17	
<b>Total</b>	116 96.67	4 3.33	120 100.00	

Display 9.4

```
%tmfilter (url=http://www.webmd.com/,
depth=2,
dir=c:\TheBook\EM14.3\TextMining\dir2,
destdir=c:\TheBook\EM14.3\TextMining\destdir2,
norestrict=1,
dataset=tmlib.webmd,numchars=32000,force=Y);
```



Display 9.5

The SAS System				
The FREQ Procedure				
Language				
LANGUAGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
English	159	100.00	159	100.00
Frequency Missing = 15				

Frequency Percent Row Pct Col Pct	Table of TRUNCATED by OMITTED			
	TRUNCATED(Truncated)	OMITTED(Omitted)		
		0	1	Total
0	159 91.38 92.98 98.15	12 6.90 7.02 100.00	171 98.28	
1	3 1.72 100.00 1.85	0 0.00 0.00 0.00	3 1.72	
<b>Total</b>	162 93.10	12 6.90	174 100.00	

Display 9.6

```
%tmfilter (url=http://www.nytimes.com/section/business,
depth=1,
dir=c:\TheBook\EM14.3\TextMining\dir3,
destdir=c:\TheBook\EM14.3\TextMining\destdir3,
norestrict=1,
dataset=tmlib.nytbusiness,numchars=32000,force=Y);
```

Display 9.7

The SAS System				
The FREQ Procedure				
Language				
LANGUAGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
English	15	100.00	15	100.00
Frequency Missing = 1				
Frequency Percent Row Pct Col Pct	Table of TRUNCATED by OMITTED			
	TRUNCATED(Truncated)	OMITTED(Omitted)		
		0	1	Total
0	15 93.75 93.75 100.00	1 6.25 6.25 100.00	16 100.00	
<b>Total</b>	15 93.75	1 6.25	16 100.00	

Display 9.8 shows the creation of the target variable. If the web page is Economics related, then target = 1 and 0 otherwise.

Display 9.8

```
data wsj ;
  set tmlib.wsj ;
  keep text omitted target ;
  if omitted then delete ;
  TARGET = 1 ;
run;

data webmd ;
  set tmlib.webmd ;
  keep text omitted target ;
  if omitted then delete ;
  TARGET = 0 ;
run;

data nytbusiness ;
  set tmlib.nytbusiness ;
  keep text omitted target ;
  if omitted then delete ;
  TARGET = 1 ;
run;

data tmlib.combined ;
  set wsj webmd nytbusiness ;
  label TARGET = "Economics (1) / Health (0)" ;
run;
```

Display 9.9

Create New Project -- Step 1 of 2 Specify Project Name and Server Directory

SAS\* Enterprise Miner 14.3

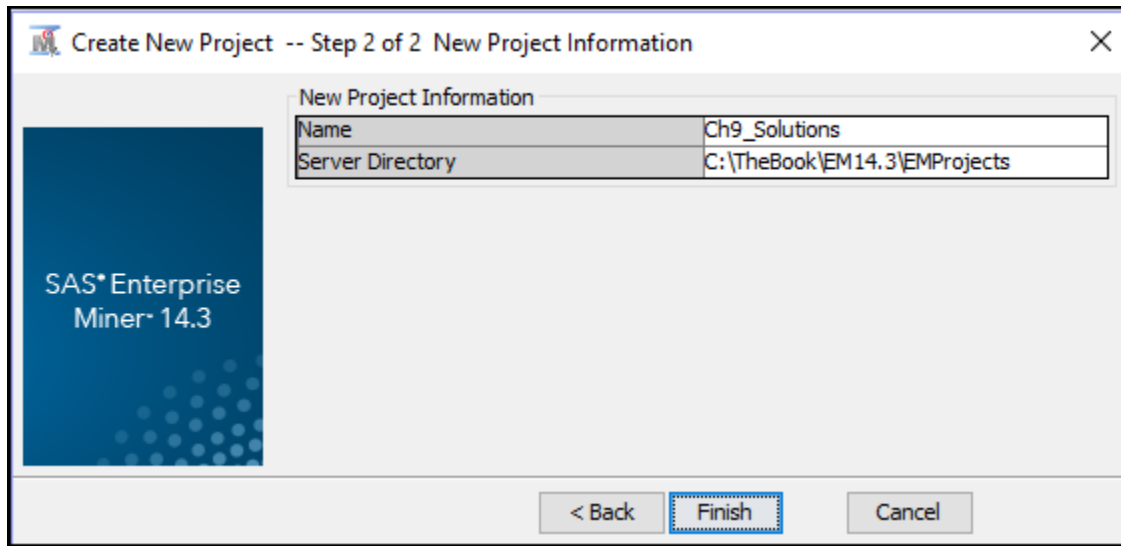
Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name  
Ch9\_Solutions

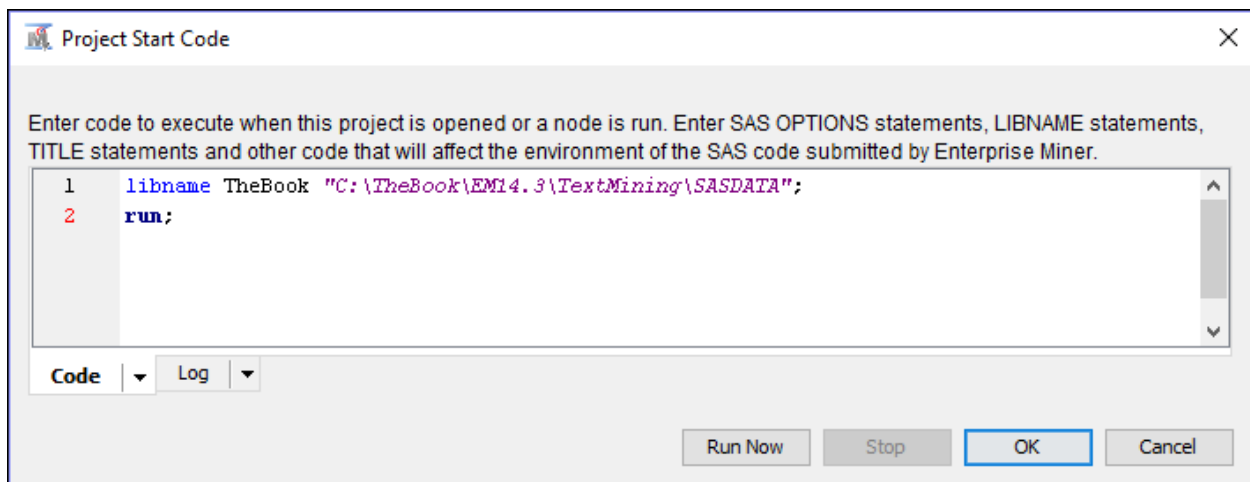
SAS Server Directory  
C:\TheBook\EM14.3\EMProjects

< Back   Next >   Cancel

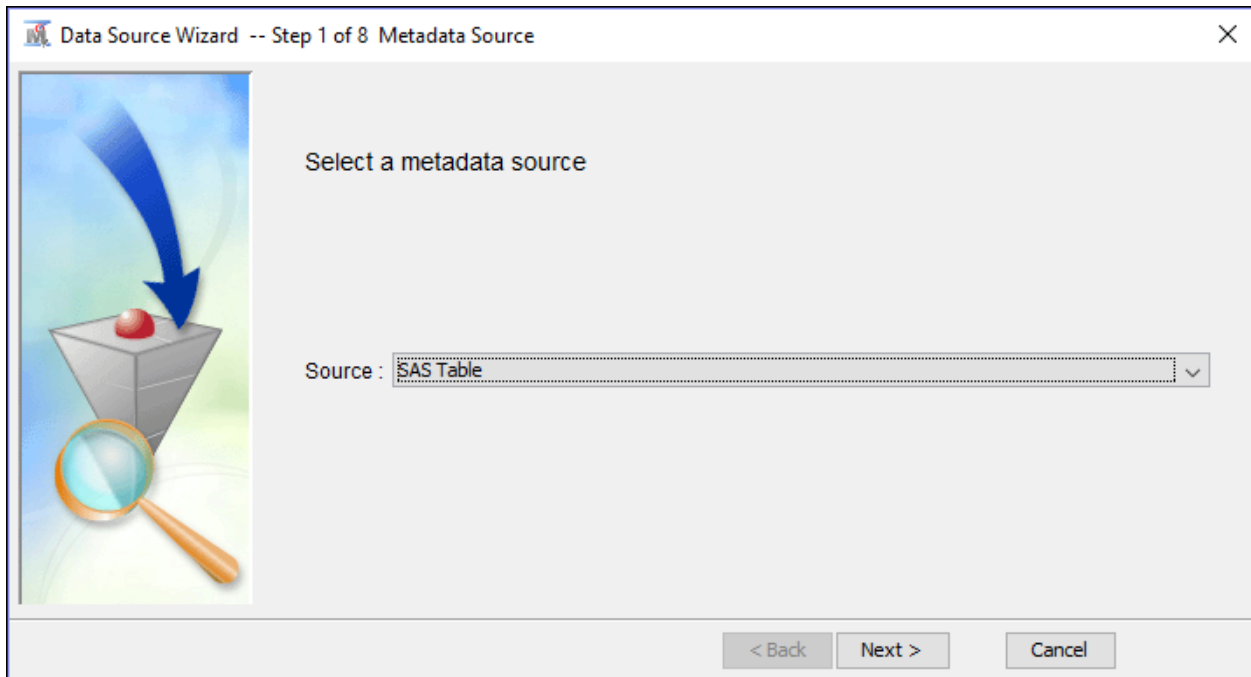
Display 9.10



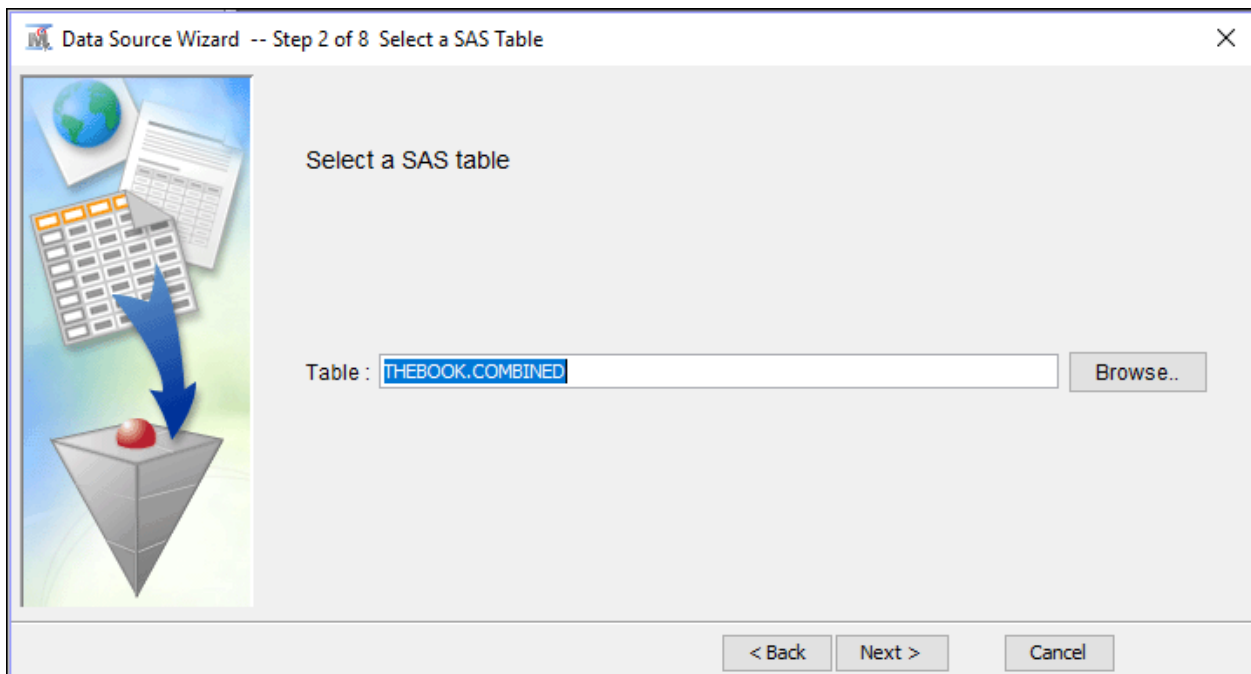
Display 9.11



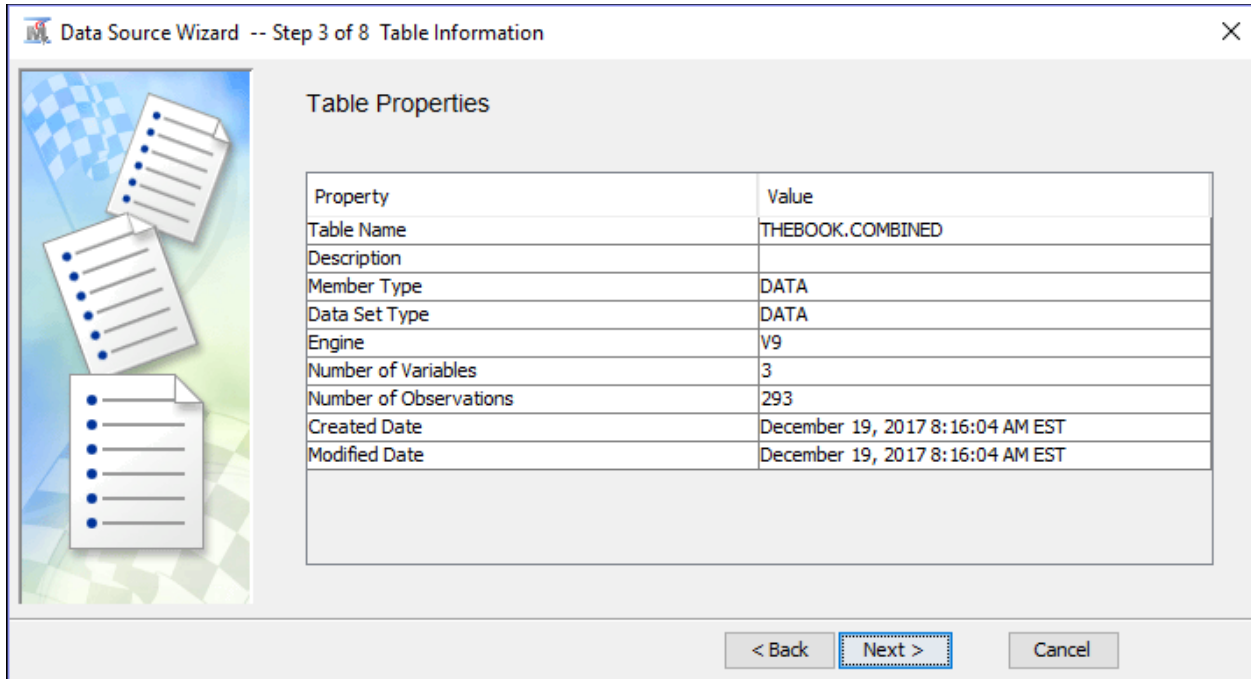
Display 9.12



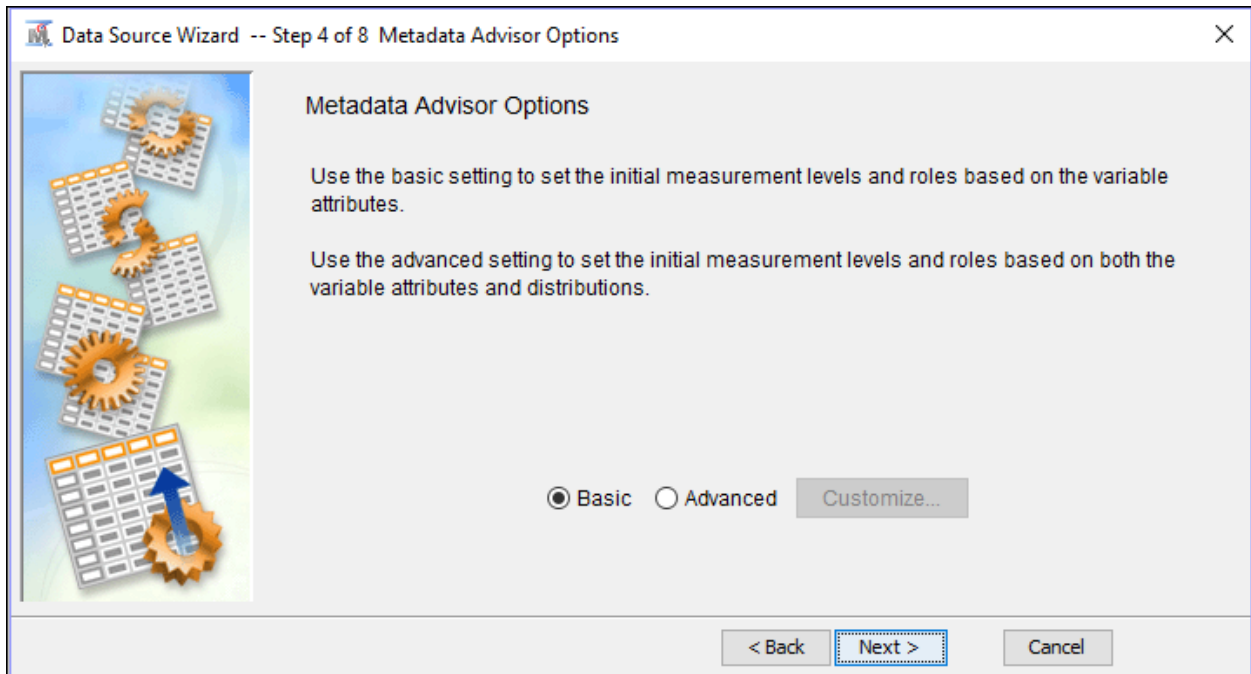
Display 9.13



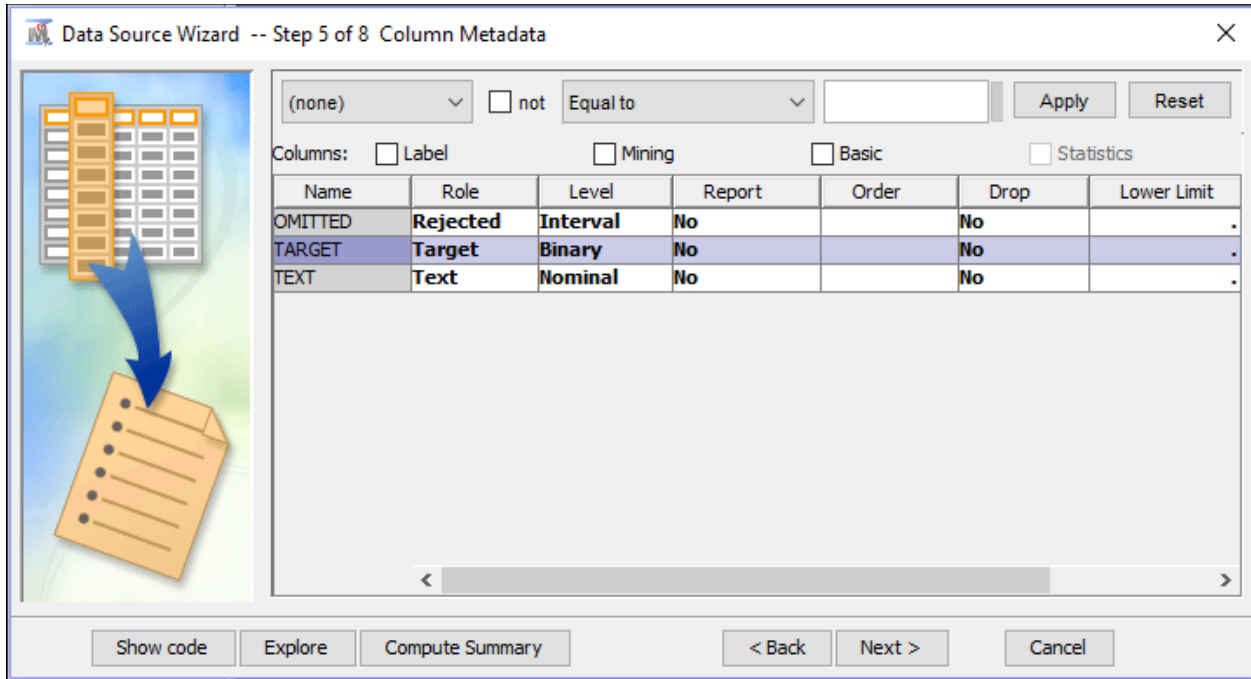
Display 9.14



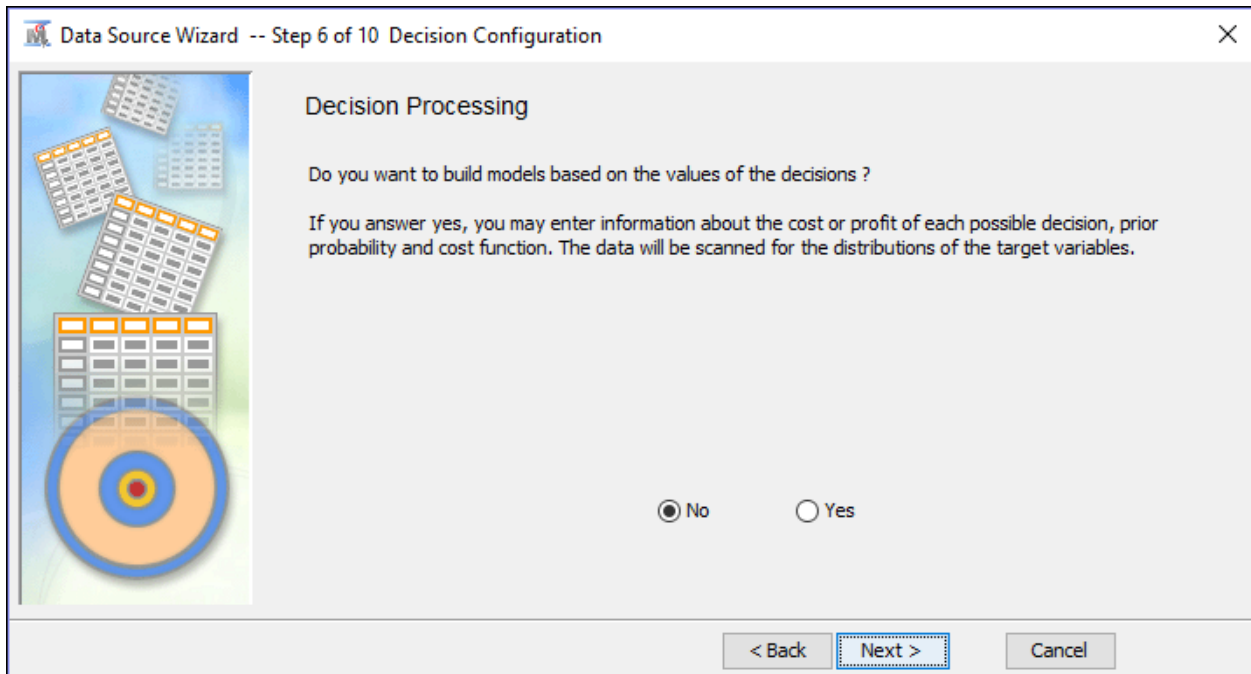
Display 9.15



Display 9.16



Display 9.17



Display 9.18

Data Source Wizard -- Step 7 of 9 Create Sample

Do you wish to create a sample data set?

No  Yes

**Table Info**

Columns 3  
Rows 293

**Sample Size**

Type Percent  
Percent 20  
Rows

< Back Next > Cancel

Display 9.19

Data Source Wizard -- Step 8 of 9 Data Source Attributes

You may change the name and the role, and can specify a population segment identifier for the data source to be created.

Name : COMBINED  
Role : Raw  
Segment :  
Notes :

< Back Next > Cancel



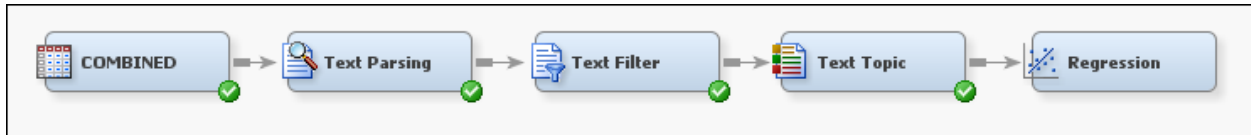
Display 9.20

Metadata Completed.

**Library:** THEBOOK  
**Data Source:** COMBINED  
**Role:** Raw

Role	Level	Count
Rejected	Interval	1
Target	Binary	1
Text	Nominal	1

Display 9.21



Display 9.22

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	7	0.172		0.016 dow,jones,+company,data,â	523	44
Multiple	8	0.357		0.016 mdscape,webmd,medicine,espaïol,français	275	84
Multiple	9	0.190		0.016 chg,preview,ytd,read,months	506	15
Multiple	10	0.182		0.016 dec,navigation,tech,search,subscriptions	495	12
Multiple	11	0.198		0.016 healthwise,poisoning,disease,syndrome,+source	566	26
Multiple	12	0.250		0.016 wsj,6:00am,boss,quirky,culture	322	43
Multiple	13	0.129		0.015 islands,republc,fr,fr,fr,fr,samoa	303	4
Multiple	14	0.174		0.016 stocks,+stock,barron,markets,pm	396	25
Multiple	15	0.249		0.016 medical,+cancer,moneyball,perspective,lewis	400	23
Multiple	16	0.178		0.015 kindle,amazon,mdscape,+cookie,+specialty	206	12
Multiple	17	0.153		0.015 +doctor,para,pacientes,sobre,de	176	8
Multiple	18	0.135		0.017 +user,services,third party,+site,+address	525	11
Multiple	19	0.164		0.016 +copy,embed,+account,+hour,close	366	11
Multiple	20	0.109		0.017 +disease,+symptom,medical,+treatment,pain	594	20
Multiple	21	0.118		0.017 â.â.+train,+kill,christmas	696	15
Multiple	22	0.160		0.016 wsj,+student,+membership,gift,complimentary	287	22

Display 9.23

```

The selected model is the model trained in the last step (Step 4). It consists of the following effects:
Intercept  TextTopic_raw17  TextTopic_raw19  TextTopic_raw7  TextTopic_raw8

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood      Likelihood
Intercept      Intercept &      Ratio
  Only      Covariates      Chi-Square      DF      Pr > ChiSq

402.898      15.330      387.5681      4      <.0001

Analysis of Maximum Likelihood Estimates

Parameter      DF      Estimate      Standard      Wald      Standardized
                DF      Estimate      Error      Chi-Square      Pr > ChiSq      Estimate      Exp(Est)

Intercept      1      -1.0130      0.8452      1.44      0.2307      -31.0306      0.363
TextTopic_raw17  1      -567.3      364.5      2.42      0.1196      2.8269      0.000
TextTopic_raw19  1      59.0116      34.3655      2.95      0.0859      13.4781      999.000
TextTopic_raw7   1      243.3      139.2      3.05      0.0805      -10.7847      999.000
TextTopic_raw8   1      -102.1      63.5545      2.58      0.1082

```

TextTopic\_Raw17 is characterized by terms “doctor”, “para” and “ patients” (See Display 9.22). From the logistic regression in Display 9.23 you can see that the coefficient of “texttopic\_raw17” is negative. It makes sense because the terms that describe TextTopic-raw17 are not related to economics.

Similarly TextTopic\_raw8 is characterized by terms “webmed”and “medicine” which are not related to economics . Hence the coefficient of TextTopic\_raw8 has negative sign.

TextTopic\_raw19 is characterized by terms “copy”, “account”, “hour” and “close” which may be related to economics or finance. Hence TextTopic\_raw19 has a positive sign.

TextTopic\_raw7 is characterized by terms “dowjones”, “company” and “data” which may be related to economics or finance. Hence TextTopic\_raw7 has a positive sign.

Therefore, the logistic regression presented in Display 9.23 makes sense, although some irrelevant terms appear in the text topic descriptions. This may be due to the fact that the pages included here are not cleaned.

You should experiment with more web pages of your interest and also clean the files, if possible.