

Chapter 6 Model Assessment

0.1	Introduction.....	Error! Bookmark not defined.
0.2	A Section Title	Error! Bookmark not defined.
	Demonstration: <Type title of demo here.>	Error! Bookmark not defined.
	Exercises	Error! Bookmark not defined.
0.3	Chapter Summary.....	Error! Bookmark not defined.
0.4	Solutions	Error! Bookmark not defined.
	Solutions to Exercises	Error! Bookmark not defined.
	Solutions to Student Activities (Polls/Quizzes)	Error! Bookmark not defined.

6.1 Model Fit Statistics

Summary Statistics Summary	
Prediction Type	Statistic
▶ Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
▶ Rankings	ROC Index (concordance) Gini coefficient
▶ Estimates	Average squared error SBC/Likelihood
3	...

As introduced in Chapter 3, summary statistics can be grouped by prediction type.

For decision prediction, the Model Comparison tool rates model performance based on accuracy or misclassification, profit or loss, and by the Kolmogorov-Smirnov (KS) statistic. Accuracy and misclassification tally the correct or incorrect prediction decisions. Profit is detailed later in this chapter. The Kolmogorov-Smirnov statistic describes the ability of the model to separate the primary and secondary outcomes.

Summary Statistics Summary	
Prediction Type	Statistic
▶ Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
▶ Rankings	ROC Index (concordance) Gini coefficient
▶ Estimates	Average squared error SBC/Likelihood

For ranking predictions, the Model Comparison tool gives two closely related measures of model fit. The ROC index is similar to concordance (introduced in Chapter 3). The Gini coefficient (for binary prediction) equals $2 \times (\text{ROC Index} - 0.5)$.



The ROC index equals the percent of concordant cases plus one-half times the percent tied cases. Recall that a pair of cases, consisting of one primary outcome and one secondary outcome, is *concordant* if the primary outcome case has a higher rank than the secondary outcome case. By contrast, if the primary outcome case has a lower rank, that pair is *discordant*. If the two cases have the same rank, they are said to be tied.

Summary Statistics Summary	
Prediction Type	Statistic
▶ Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
▶ Rankings	ROC Index (concordance) Gini coefficient
▶ Estimates	Average squared error SBC/Likelihood

5

For estimate predictions, the Model Comparison tool provides two performance statistics. Average squared error was used to tune many of the models fit in earlier chapters. The Schwarz's Bayesian Criterion (SBC) is a penalized likelihood statistic. The likelihood statistic was used to estimate regression and neural network model parameters and can be thought of as a weighted average squared error.



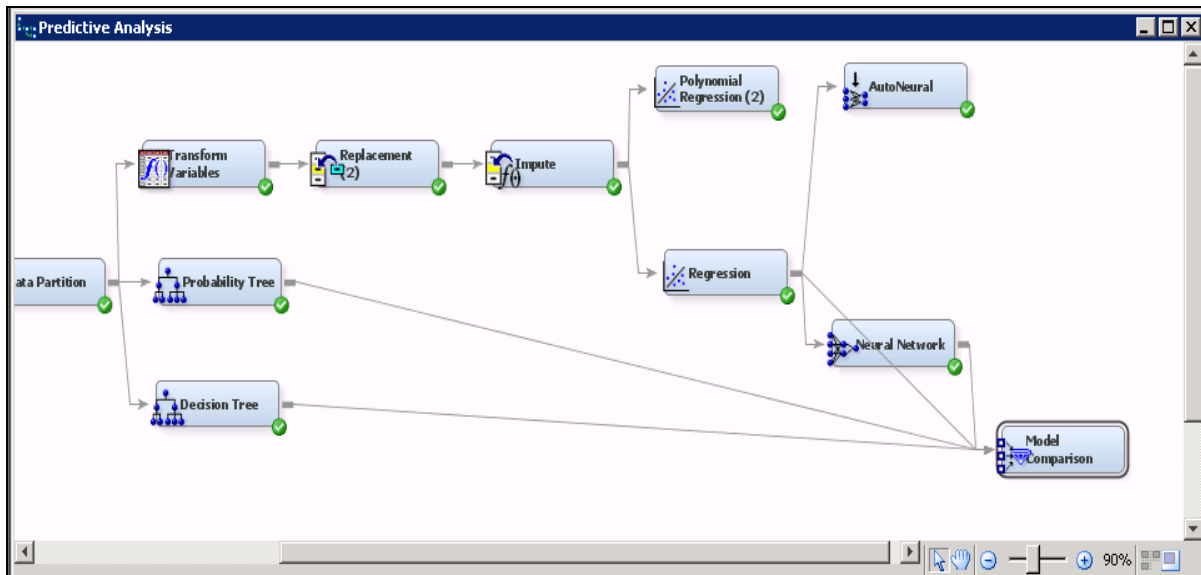
SBC is provided only for regression and neural network models and is calculated only on training data.



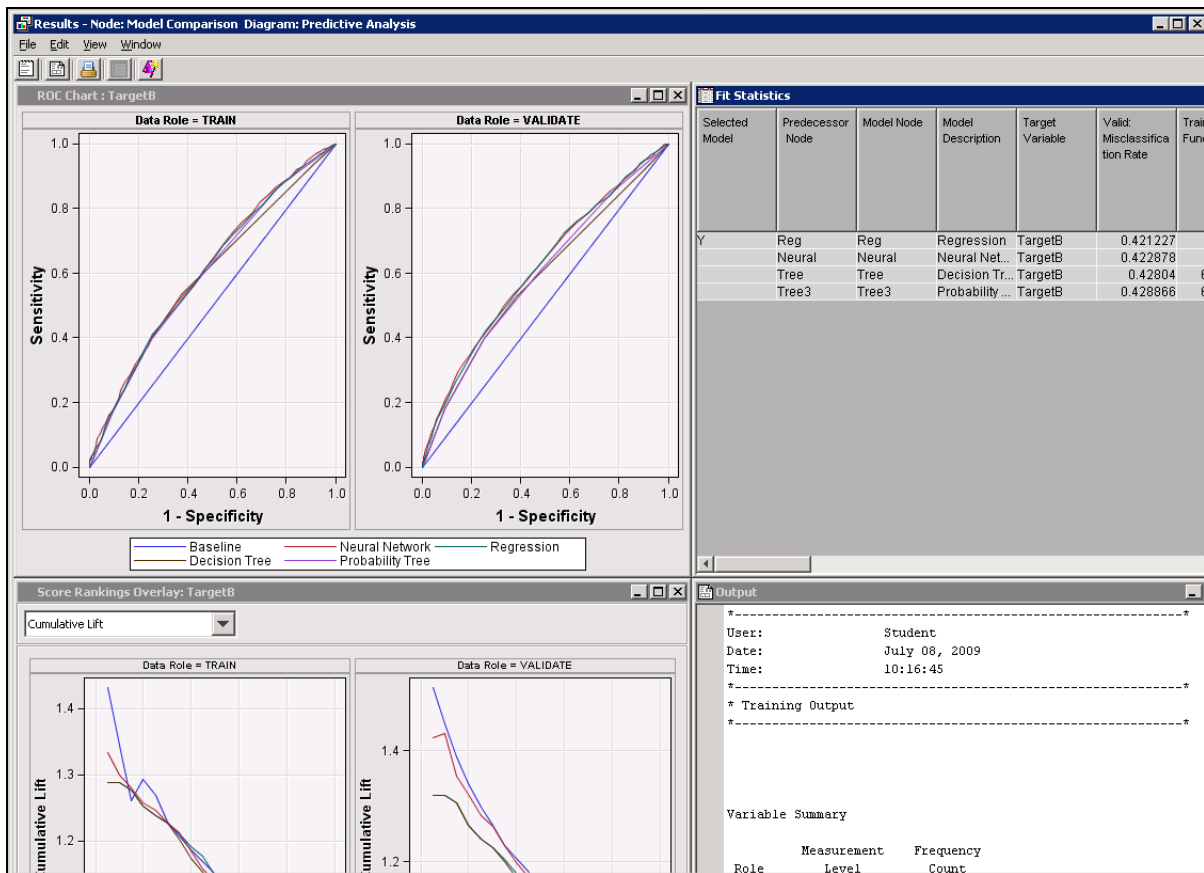
Comparing Models with Summary Statistics

After you build several models, it is desirable to compare their performance. The Model Comparison tool collects assessment information from attached modeling nodes and enables you to easily compare model performance measures.

1. Select the **Assess** tab.
2. Drag a **Model Comparison** tool into the diagram workspace.
3. Connect both Decision Trees, the Regression node, and the Neural Network node to the Model Comparison node as shown. (Self-study models are ignored here.)



4. Run the Model Comparison node and view the results. The Results window opens.



The Results window contains four sub-windows: ROC Chart, Score Rankings, Fit Statistics, and Output.

5. Maximize the Output window.

6. Go to line 90 in the Output window.

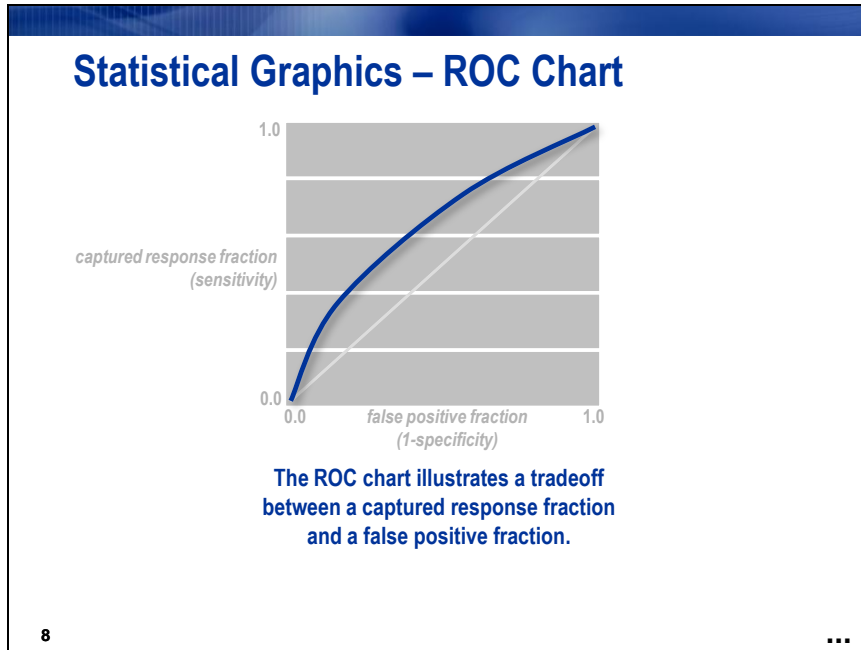
Data Role=Valid				
Statistics	Reg	Neural	Tree	Tree3
Valid: Kolmogorov-Smirnov Statistic	0.16	0.16	0.14	0.14
Valid: Average Squared Error	0.24	0.24	0.24	0.24
Valid: Roc Index	0.61	0.61	0.58	0.59
Valid: Average Error Function	0.67	0.67	0.68	0.68
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff
Valid: Cumulative Percent Captured Response	14.33	14.49	13.22	13.22
Valid: Percent Captured Response	7.18	6.90	6.60	6.60
Valid: Frequency of Classified Cases	4843.00	4843.00	4843.00	4843.00
Valid: Divisor for ASE	9686.00	9686.00	9686.00	9686.00
Valid: Error Function	6533.21	6512.78	6584.21	6582.17
Valid: Gain	43.06	44.71	32.00	32.00
Valid: Gini Coefficient	0.22	0.22	0.17	0.18
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.00	0.00	0.00	0.00
Valid: Kolmogorov-Smirnov Probability Cutoff	0.54	0.52	0.43	0.49
Valid: Cumulative Lift	1.43	1.45	1.32	1.32
Valid: Lift	1.44	1.38	1.32	1.32
Valid: Maximum Absolute Error	0.86	0.87	0.64	0.64
Valid: Misclassification Rate	0.42	0.42	0.43	0.43
Valid: Sum of Frequencies	4843.00	4843.00	4843.00	4843.00
Valid: Root Average Squared Error	0.49	0.49	0.49	0.49
Valid: Cumulative Percent Response	71.55	72.37	66.02	66.02
Valid: Percent Response	71.90	69.01	66.02	66.02
Valid: Sum of Squared Errors	2332.27	2323.48	2357.33	2356.28
Valid: Number of Wrong Classifications	2040.00	2048.00	2073.00	2077.00

The output shows various fit statistics for the selected models. It appears that the performance of each model, as gauged by fit statistics, is quite similar.

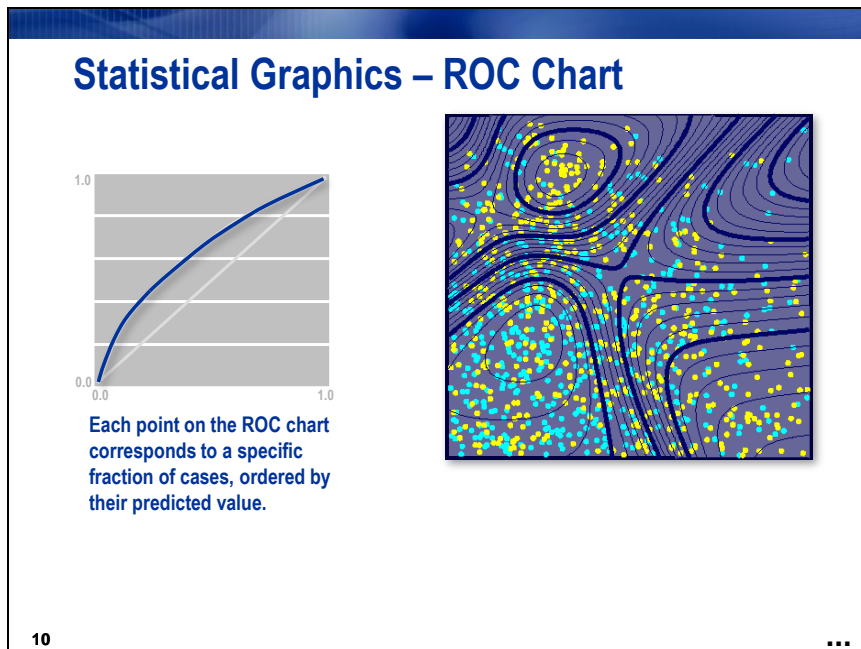
As discussed above, the choice of fit statistics best depends on the predictions of interest.

Prediction Type	Validation Fit Statistic	Direction
Decisions	Misclassification	smallest
	Average Profit/Loss	largest/smallest
	Kolmogorov-Smirnov Statistic	largest
Rankings	ROC Index (concordance)	largest
	Gini Coefficient	largest
Estimates	Average Squared Error	smallest
	Schwarz's Bayesian Criterion	smallest
	Log-Likelihood	largest

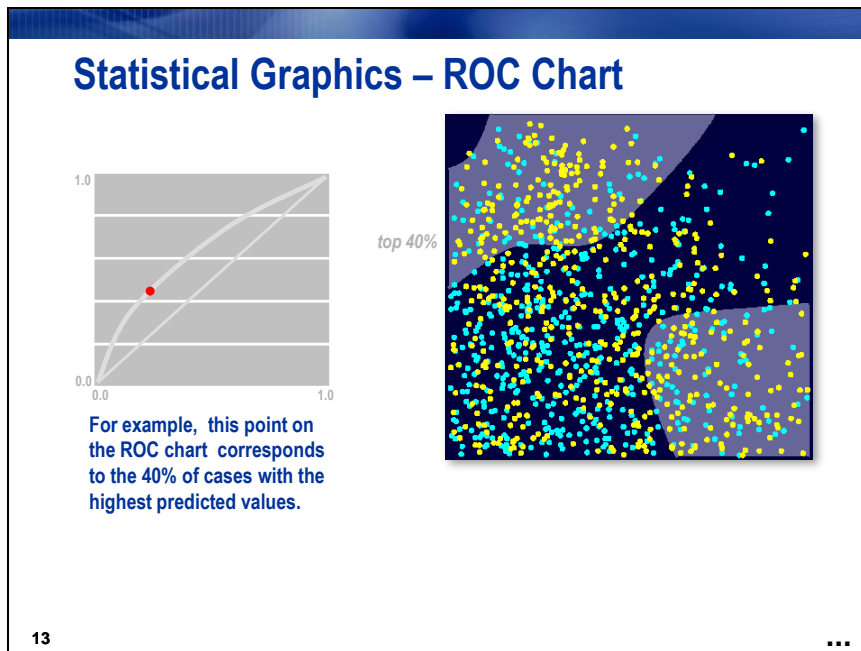
6.2 Statistical Graphics



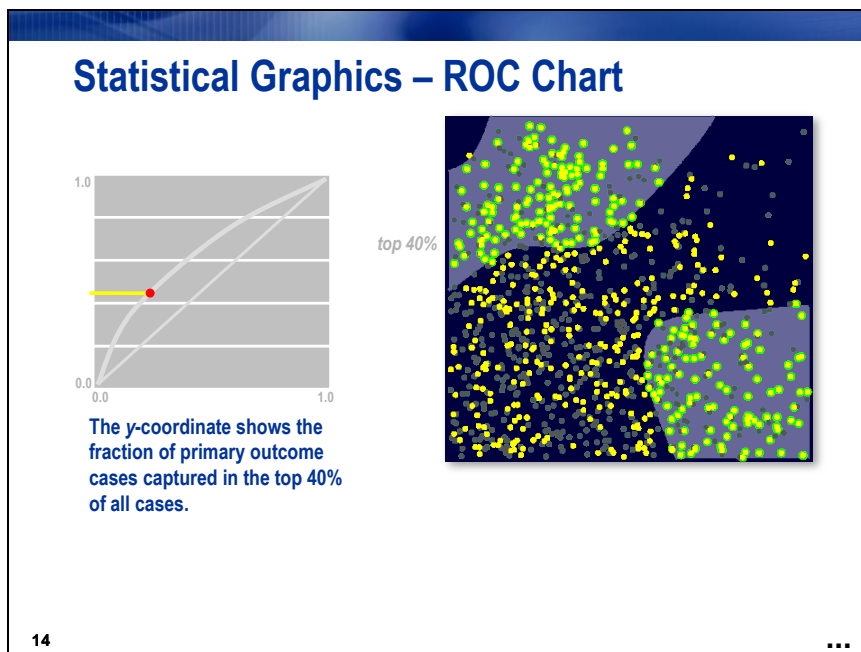
The Model Comparison tool features two charts to aid in model assessment: the ROC chart and the Score Rankings chart. Consider the ROC chart.



To create a ROC chart, predictions are generated for a set of validation data. For chart generation, the predictions must be rankings or estimates. The validation data is sorted from high to low (either scores or estimates). Each point on the ROC chart corresponds to a specific fraction of the sorted data.

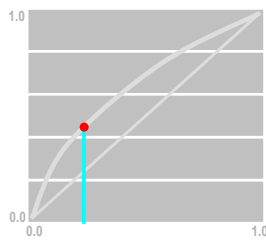


For example, the red point on the ROC chart corresponds to the indicated selection of 40% of the validation data. That is, the points in the gray region on the scatter plot are in the highest 40% of predicted probabilities.



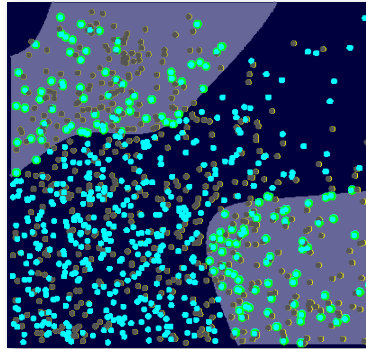
The vertical or y-coordinate of the red point indicates the fraction of primary outcome cases “captured” in the gray region (here about 45%).

Statistical Graphics – ROC Chart



The x-coordinate shows the fraction of *secondary outcome* cases captured in the top 40% of all cases.

top 40%

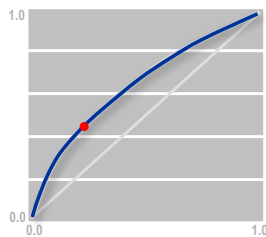


16

...

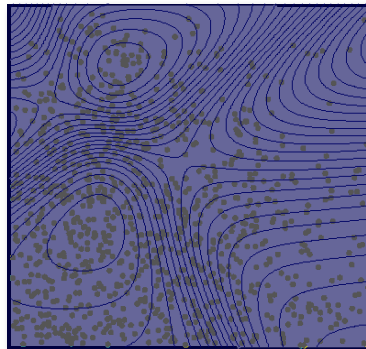
The horizontal or x -coordinate of the red point indicates the fraction of secondary outcome cases “captured” in the gray region (here about 25%).

Statistical Graphics – ROC Chart



Repeat for all selection fractions.

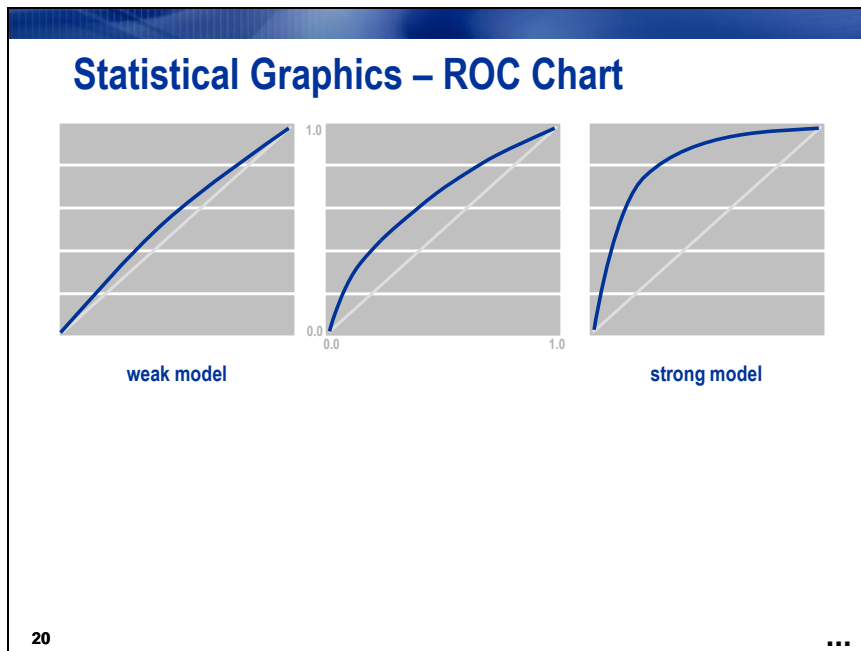
top 40%



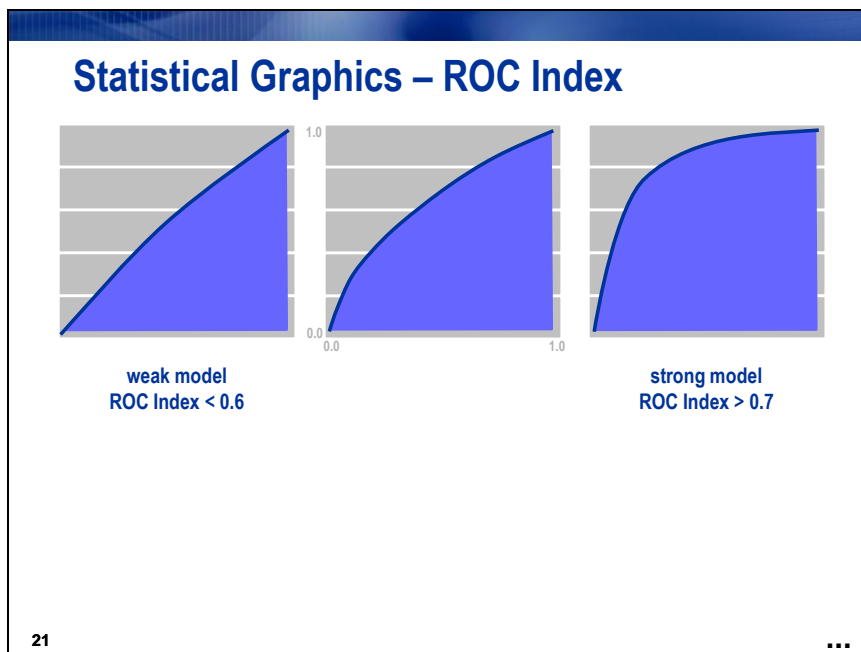
18

...

The ROC chart represents the union of similar calculations for all selection fractions.



The ROC chart provides a nearly universal diagnostic for predictive models. Models that capture primary and secondary outcome cases in a proportion approximately equal to the selection fraction are weak models (left). Models that capture mostly primary outcome cases without capturing secondary outcome cases are strong models (right).



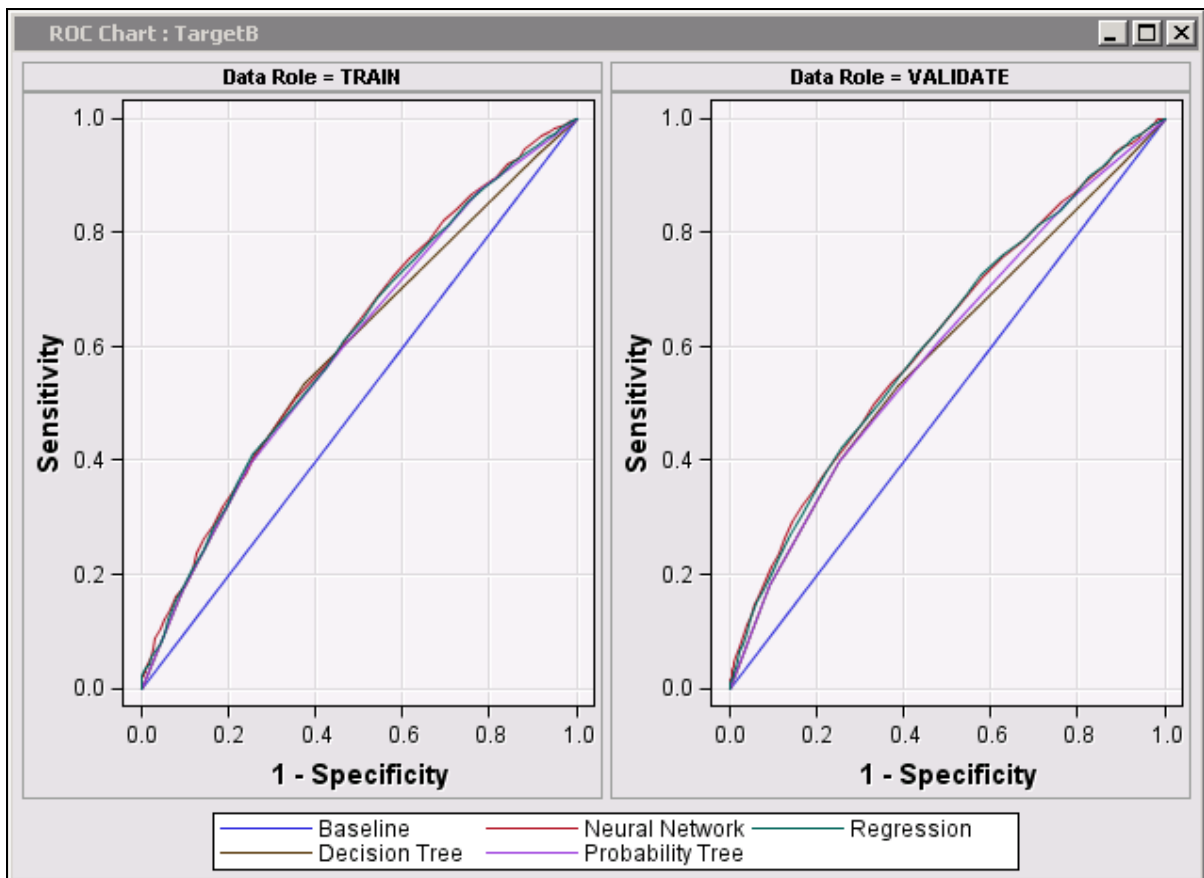
The tradeoff between primary and secondary case capture can be summarized by the area under the ROC curve. In SAS Enterprise Miner, this area is called the *ROC Index*. (In statistical literature, it is more commonly called the *c*-statistic.) Perhaps surprisingly, the ROC Index is closely related to concordance, the measure of correct case ordering.



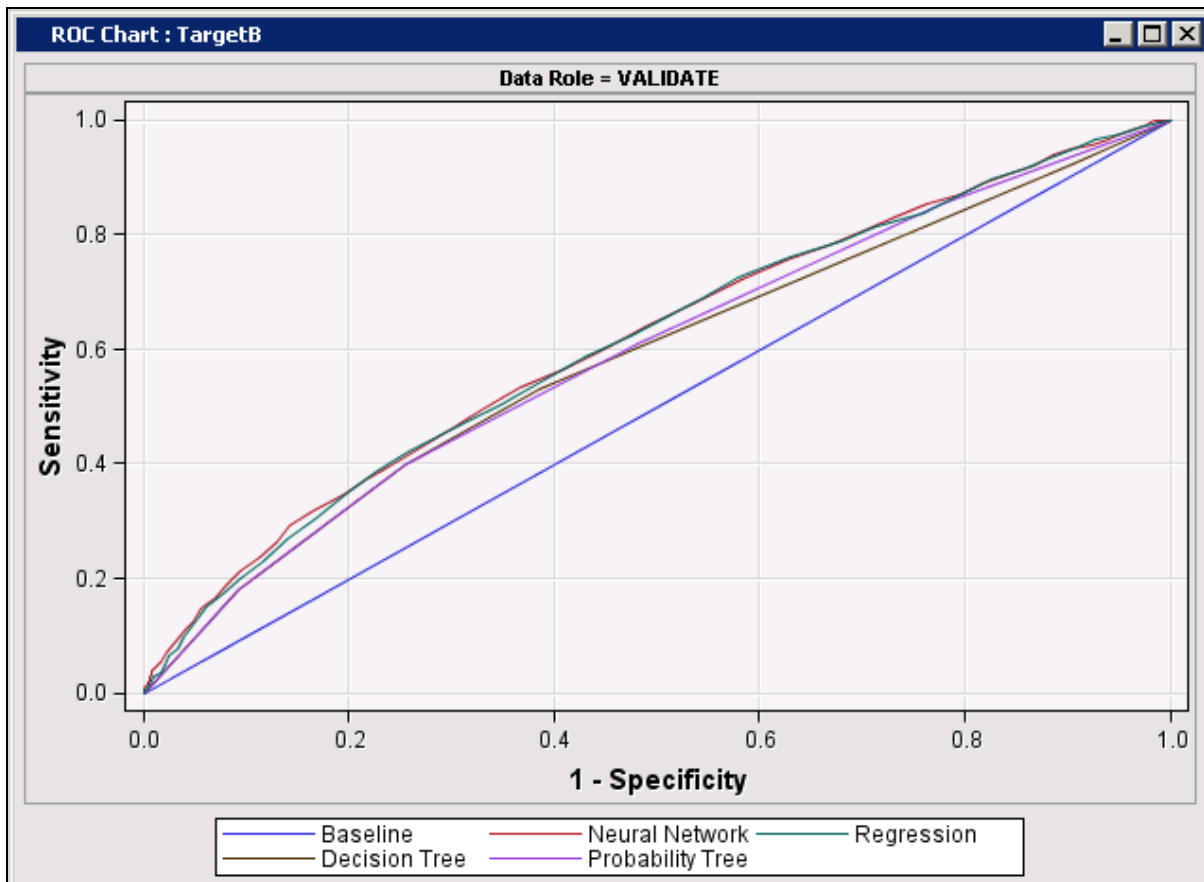
Comparing Models with ROC Charts

Use the following steps to compare models using ROC charts.

1. Maximize the ROC chart.

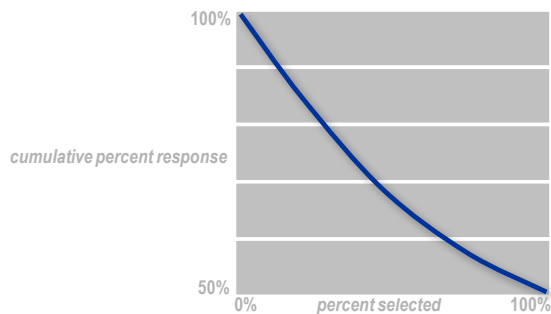


2. Double-click the **Data Role = VALIDATE** plot.



The ROC chart shows little difference between the non-tree models. This is consistent with the values of the ROC Index, which equals the area under the ROC curves.

Statistical Graphics – Response Chart



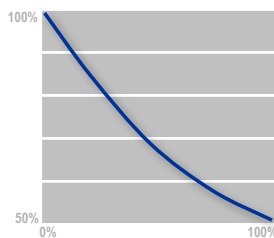
The response chart shows the expected response rate for various selection percentages.

23

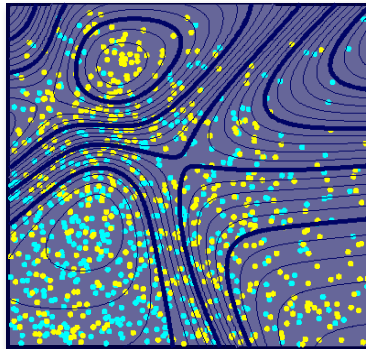
...

A second category of assessment charts examines response rate. It is the prototype of the so-called Score Rankings charts found in every model Results window.

Statistical Graphics – Response Chart



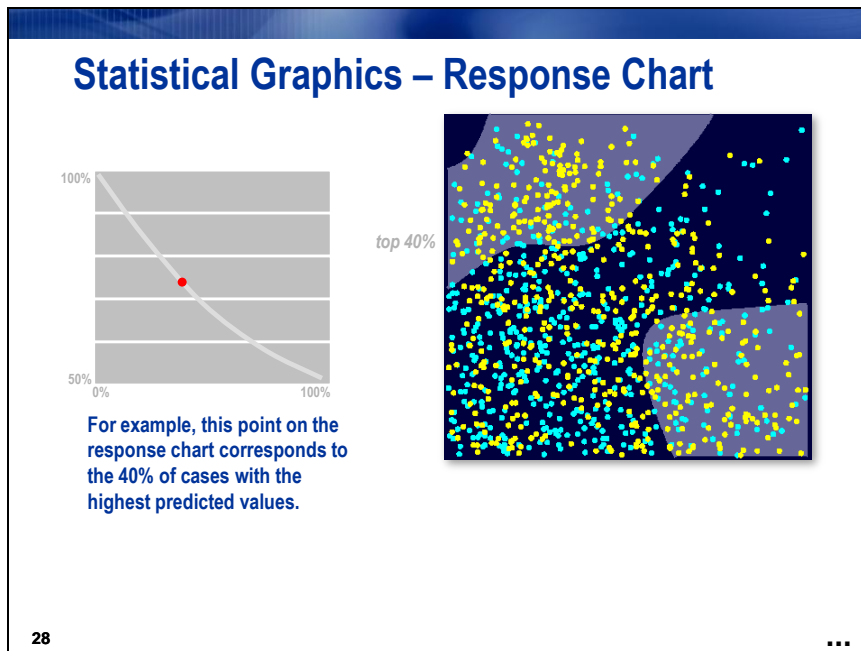
Each point on the response chart corresponds to a specific fraction of cases, ordered by their predicted values.



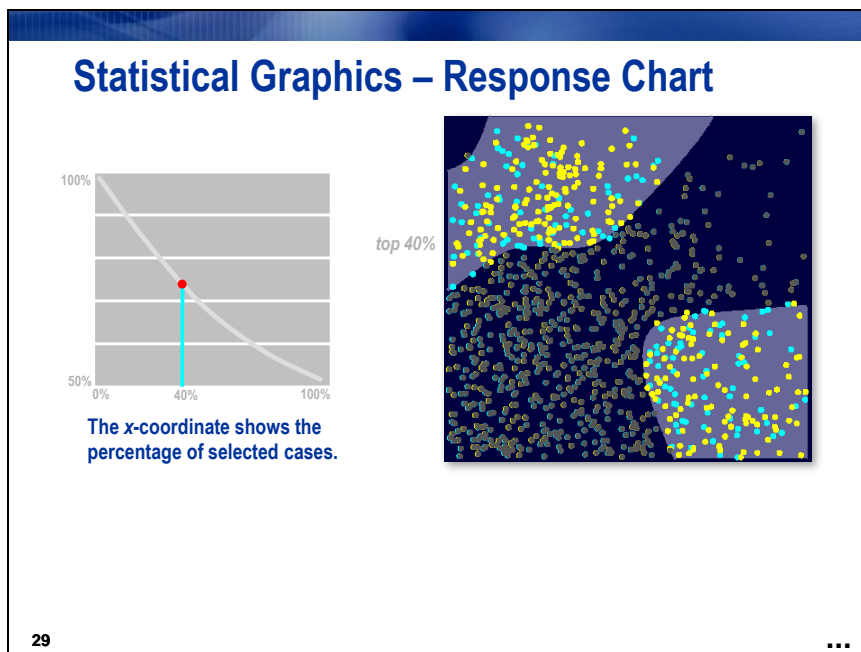
25

...

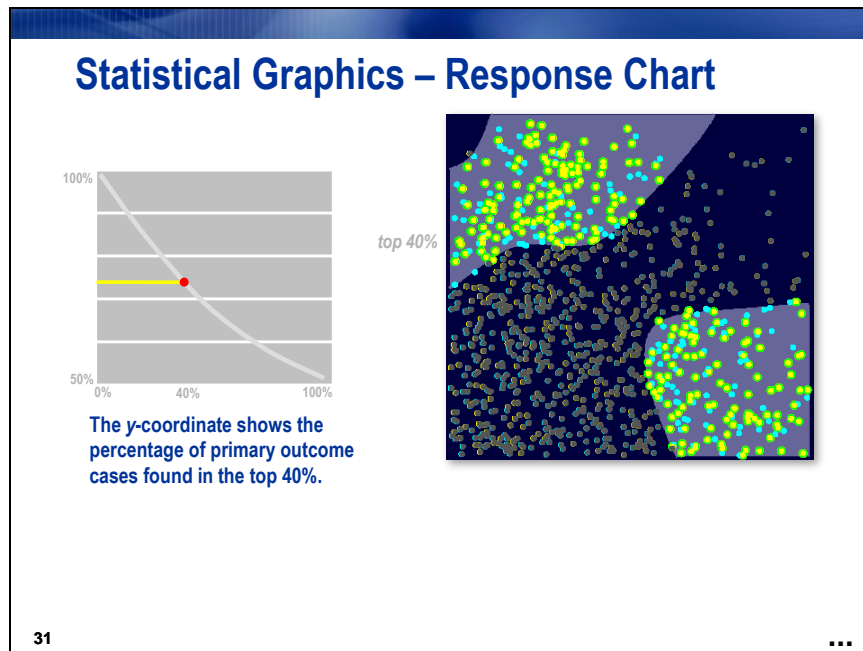
As with ROC charts, a model is applied to validation data to sort the cases from highest to lowest (again, by prediction rankings or estimates). Each point on the response chart corresponds to a selected fraction of cases.



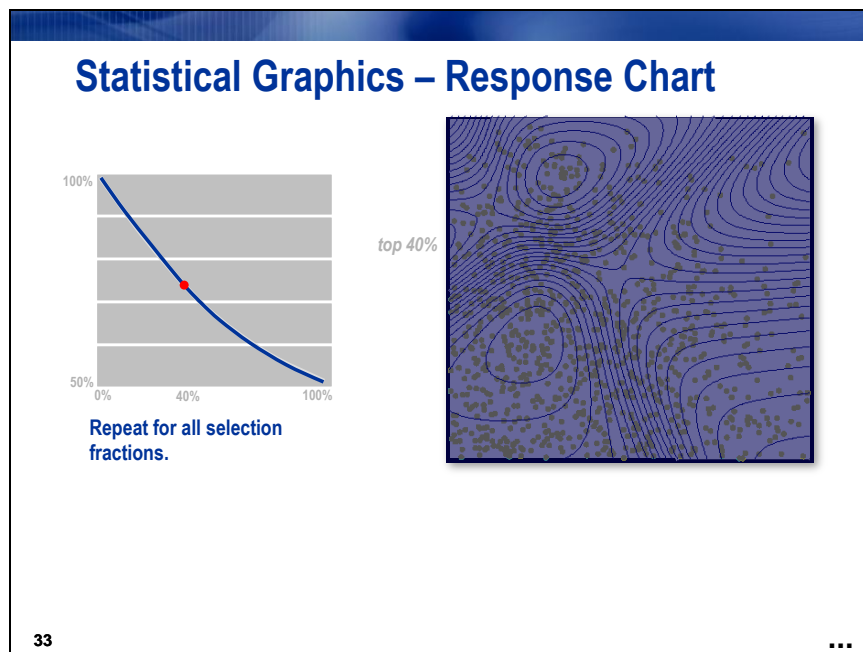
For example, the red point on the Response chart corresponds to the indicated selection of 40% of the validation data. That is, the points in the gray region on the scatter plot are in the highest 40% of predicted probabilities.



The x-coordinate of the red point is simply the selection fraction (in this case, 40%).



The vertical coordinate for a point on the response chart is the proportion of primary outcome cases in the selected fraction. It is called the *cumulative percent response* in the SAS Enterprise Miner interface and is more widely known as *cumulative gain* in the predictive modeling literature. Dividing cumulative gain by the primary outcome proportion yields a quantity named *lift*.



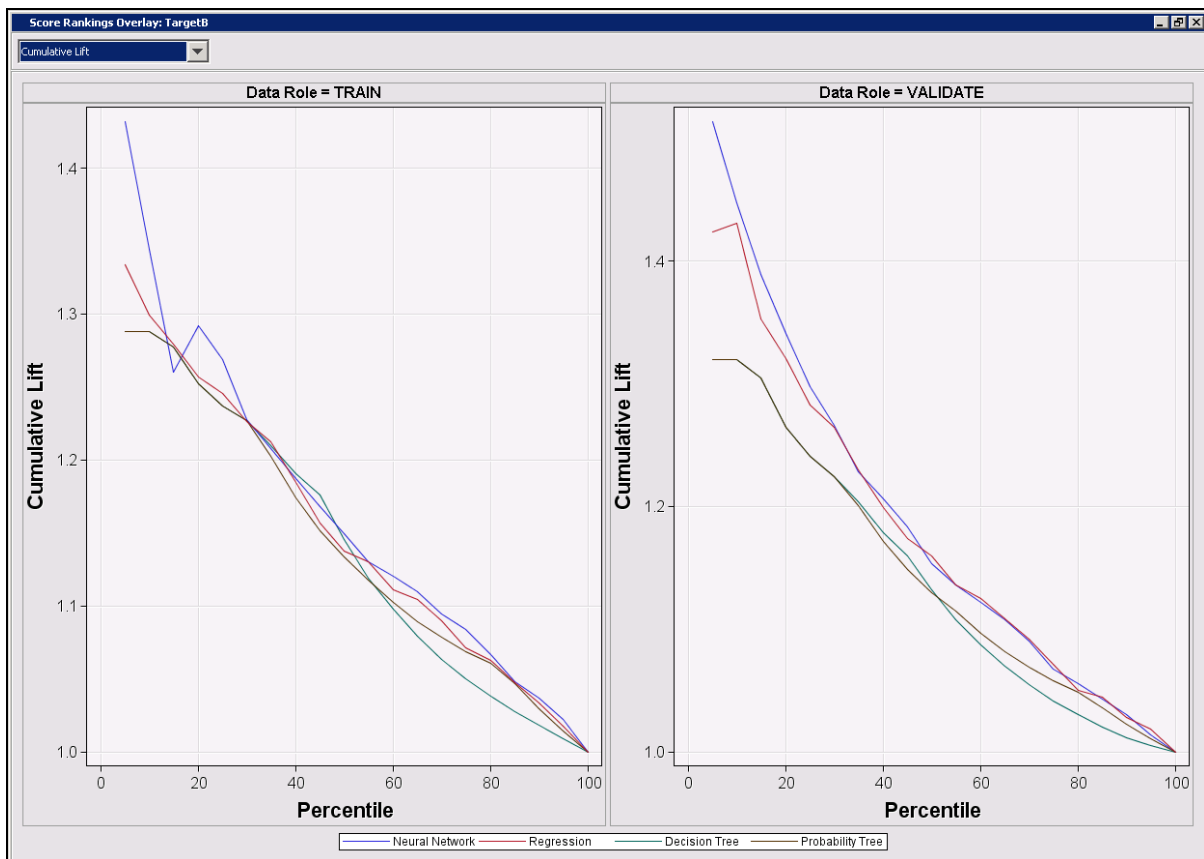
Plotting cumulative gain for all selection fractions yields a *gains chart*. Notice that when all cases are selected, the cumulative gain equals the overall primary outcome proportion.



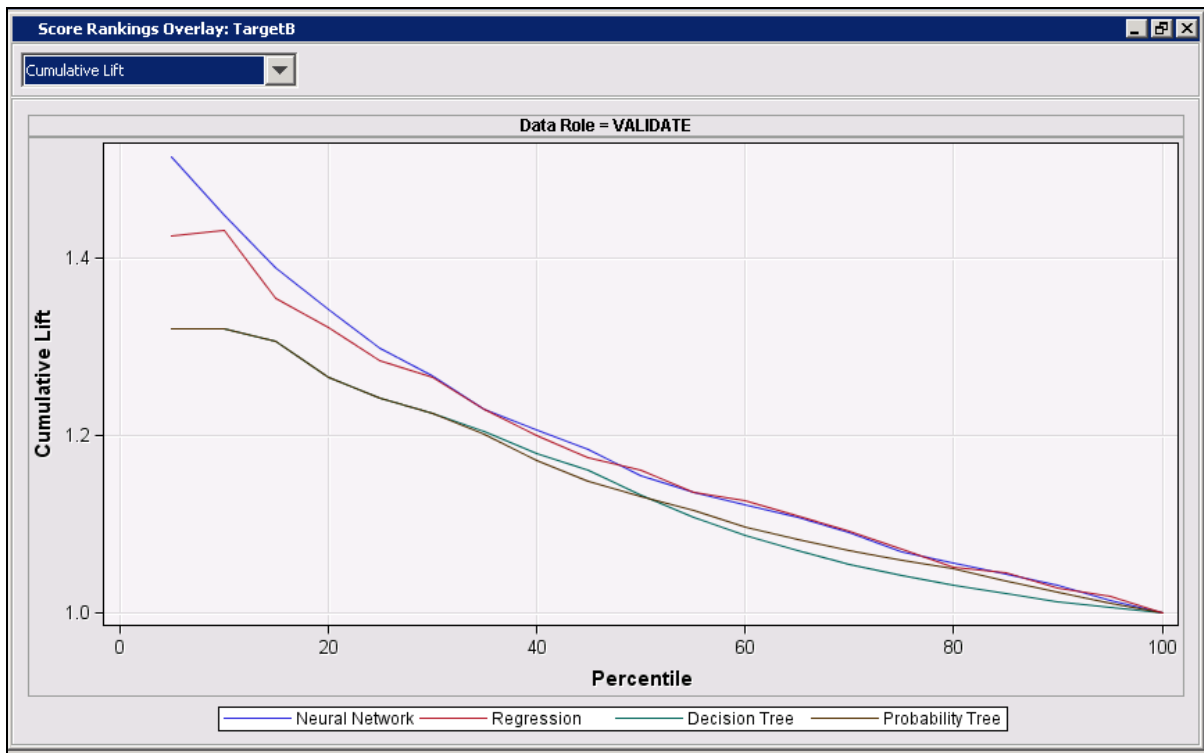
Comparing Models with Score Rankings Plots

Use the following steps to compare models with Score Rankings plots:

1. Maximize the Score Rankings Overlay window.



2. Double-click the **Data Role = VALIDATE** plot.

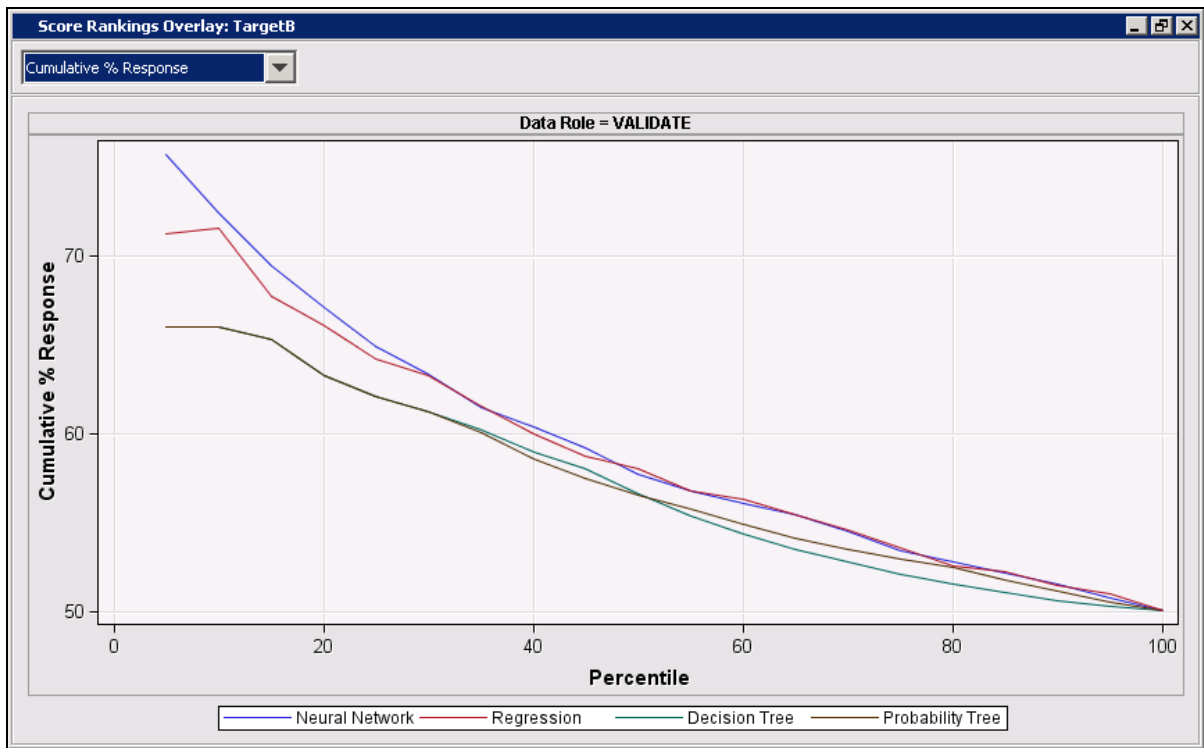


The Score Rankings Overlay plot presents what is commonly called a *cumulative lift chart*. Cases in the training and validation data are ranked based on decreasing predicted target values. A fraction of the ranked data is selected. This fraction, or *decile*, corresponds to the horizontal axis of the chart. The ratio, (proportion of cases with the primary outcome in the selected fraction) to (proportion of cases with the primary outcome overall), is defined as *cumulative lift*. Cumulative lift corresponds to the vertical axis. High values of cumulative lift suggest that the model is doing a good job separating the primary and secondary cases.

As can be seen, the model with the highest cumulative lift depends on the decile; however, none of the models has a cumulative lift over 1.5.

It is instructive to view the actual proportion of cases with the primary outcome (called *gain* or *cumulative percent response*) at each decile.

3. Select **Chart** ⇒ **Cumulative % Response**. The Score Rankings Overlay plot is updated to show Cumulative Percent Response on the vertical axis.



This plot should show the response rate for soliciting the indicated fraction of individuals. Unfortunately, the proportion of responders in the **training data** does not equal the true proportion of responders for the **97NK campaign**. The training data under-represents nonresponders by almost a factor of 20!

This under-representation was not an accident. It is a rather standard predictive modeling practice known as *separate sampling*. (*Oversampling, balanced sampling, choice-based sampling, case-control sampling*, and other names are also used.)



Adjusting for Separate Sampling

If you do not adjust for separate sampling, the following occurs:

- Prediction estimates reflect target proportions in the training sample, not the population from which the sample was drawn.
- Score Rankings plots are inaccurate and misleading.
- Decision-based statistics related to misclassification or accuracy misrepresent the model performance on the population.

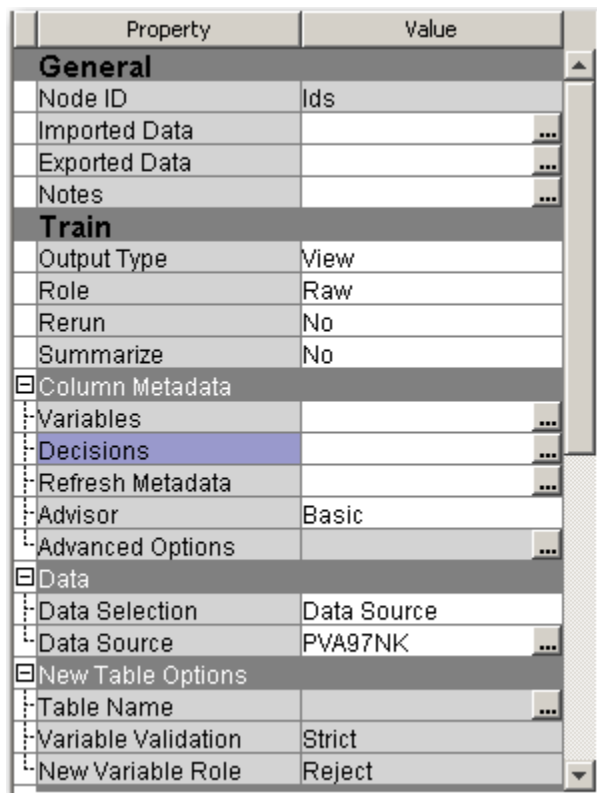
Fortunately, it is easy to adjust for separate sampling in SAS Enterprise Miner. However, you must rerun the models that you created.



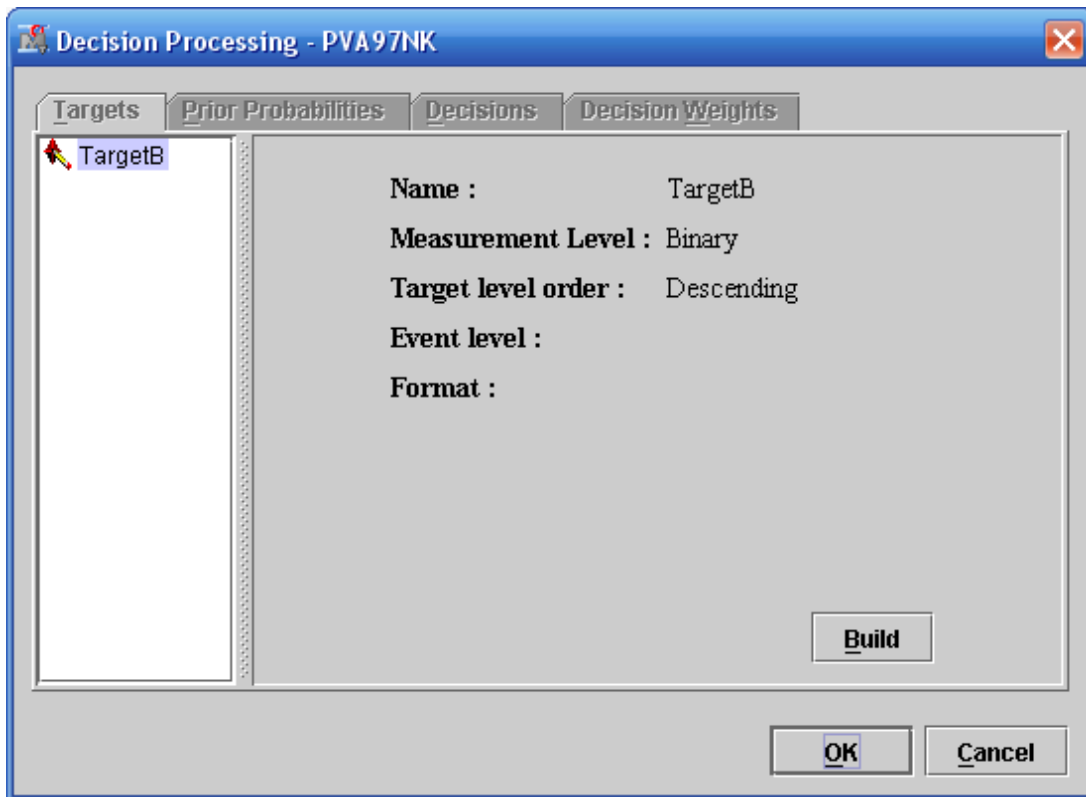
Because this can take some time, it is recommended that you run this demonstration during the discussion about the benefits and consequences of separate sampling.

Follow these steps to integrate sampling information into your analysis.

1. Close the Results - Model Comparison window.
2. Select **Decisions** ⇒ ... in the **PVA97NK** node's Properties panel.



The Decision Processing - PVA97NK dialog box opens.



The Decision Processing dialog box enables you to inform SAS Enterprise Miner about the extent of separate sampling in the training data. This is done by defining *prior probabilities*.

3. Select **Build** in the Decision Processing dialog box. SAS Enterprise Miner scans the raw data to determine the proportion of primary and secondary outcomes in the raw data.

4. Select the **Prior Probabilities** tab.

Decision Processing - PVA97NK

Targets **Prior Probabilities** **Decisions** **Decision Weights**

Do you want to enter new prior probabilities?

☐ Yes ☒ No Set Equal Prior

Level	Count	Prior
1	4843	0.5
0	4843	0.5

OK Cancel

5. Select **Yes**. The dialog box is updated to show the Adjusted Prior column.

Decision Processing - PVA97NK

Targets **Prior Probabilities** **Decisions** **Decision Weights**

Do you want to enter new prior probabilities?

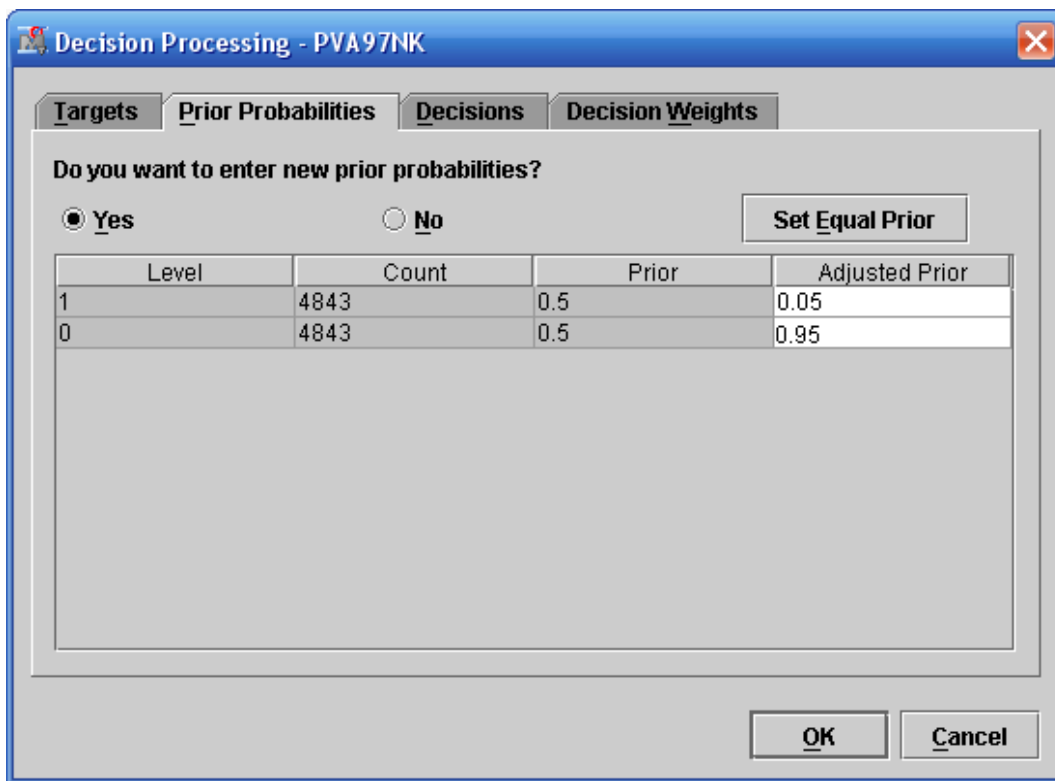
☒ Yes ☐ No Set Equal Prior

Level	Count	Prior	Adjusted Prior
1	4843	0.5	0.0
0	4843	0.5	0.0

OK Cancel

The Adjusted Prior column enables you to specify the proportion of primary and secondary outcomes in the original **PVA97NK** population. When the analysis problem was introduced in Chapter 3, the primary outcome (response) proportion was claimed to be 5%.

6. Type **0.05** as the Adjusted Prior value for the primary outcome, Level 1.
7. Type **0.95** in the as the Adjusted Prior value for the secondary outcome, Level 0.



The dialog box titled "Decision Processing - PVA97NK" has four tabs: "Targets", "Prior Probabilities", "Decisions", and "Decision Weights". The "Prior Probabilities" tab is selected. It contains a question "Do you want to enter new prior probabilities?" with two radio buttons: "Yes" (selected) and "No". To the right of the radio buttons is a button labeled "Set Equal Prior". Below this is a table with four columns: "Level", "Count", "Prior", and "Adjusted Prior".

Level	Count	Prior	Adjusted Prior
1	4843	0.5	0.05
0	4843	0.5	0.95

At the bottom right of the dialog box are two buttons: "OK" and "Cancel".

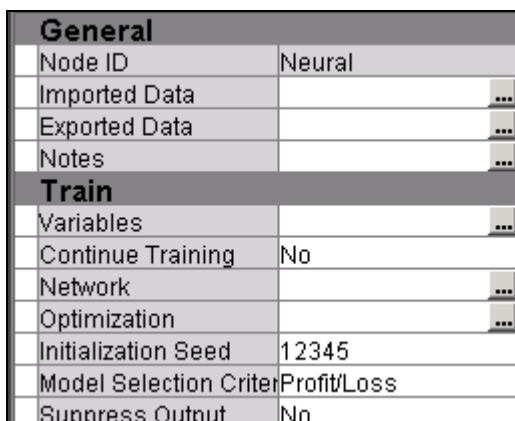
8. Select **OK** to close the Decision Processing dialog box.

Decision Processing and the Neural Network Node

If your diagram was developed according to the instructions in Chapters 3 through 5, one modeling node (the Neural Network) needs a property change to correctly use the decision processing information.

Follow these steps to adjust the Neural Network node settings.

1. Examine the Properties panel for the Neural Network node. The default model selection criterion is Profit/Loss.



The Properties panel for the Neural Network node is shown with the "General" tab selected. It contains the following properties:

General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No

2. Select **Model Selection Criterion** ⇒ **Average Error**.

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



When no decision data is defined, the neural network optimizes complexity using average squared error, even though the default says Profit/Loss. Now that decision data is defined, you must manually change Model Selection Criterion to Average Error.

Decision Processing and the AutoNeural Node

Decision processing data (prior probabilities and decision weights) is ignored by the AutoNeural node. Predictions from the node are adjusted for priors, but that actual model selection process is based strictly on misclassification (without prior adjustment). This fact can lead to unexpected prediction results when the primary and secondary outcome proportions are not equal. Fortunately, the **PVA97NK** data has an equal proportion of primary and secondary outcomes and gives reasonable results with the AutoNeural node.



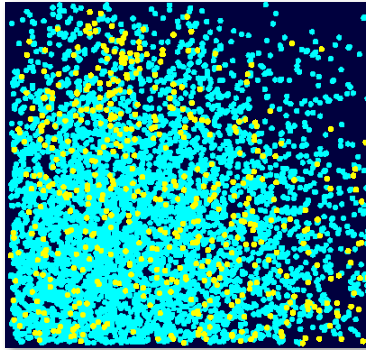
Using the AutoNeural node with training data that does not have equal outcome proportions is reasonable only if your data is not separately sampled **and** your prediction goal is classification.

Run the diagram from the Model Comparison node. This reruns the entire analysis.

6.3 Adjusting for Separate Sampling

Outcome Overrepresentation

A common predictive modeling practice is to build models from a sample with a primary outcome proportion different from the original population.

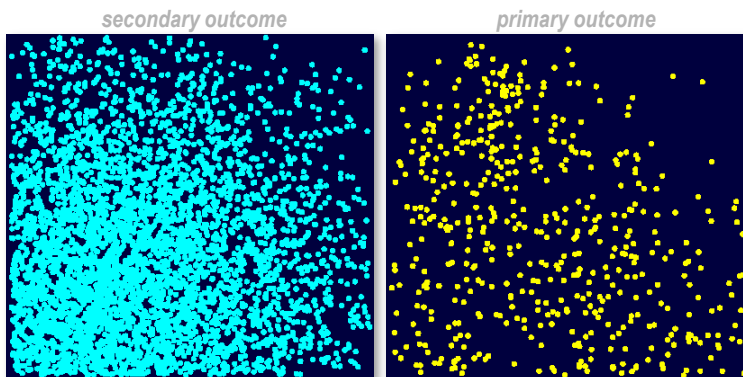


37

...

A common predictive modeling practice is to build models from a sample with a primary outcome proportion different from the true population proportion. This is typically done when the ratio of primary to secondary outcome cases is small.

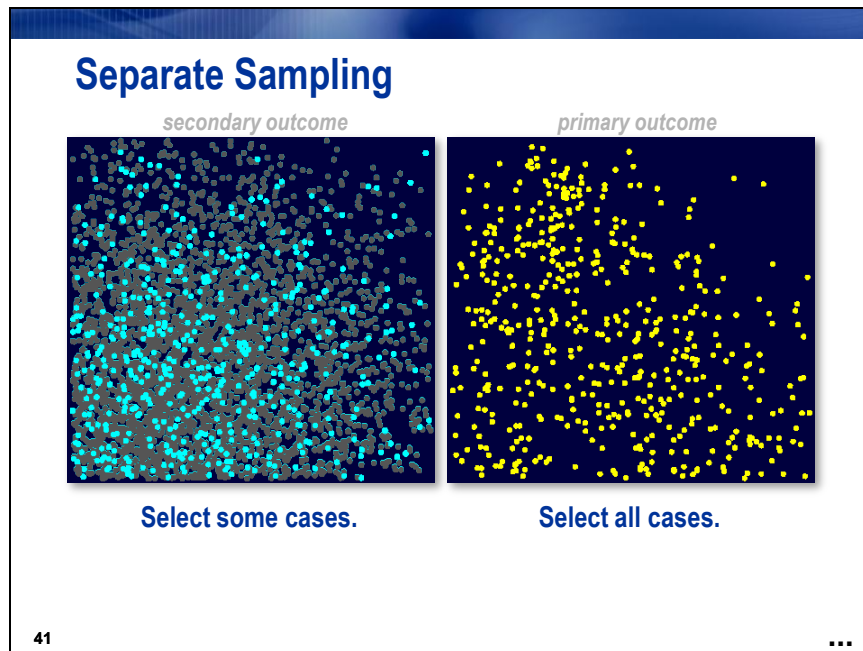
Separate Sampling



Target-based samples are created by considering the primary outcome cases separately from the secondary outcome cases.

40

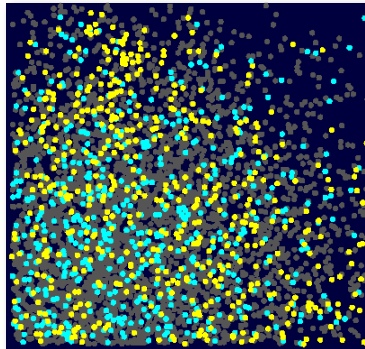
...



Separate sampling gets its name from the technique used to generate the modeling data, that is, samples are drawn separately based on the target outcome. In the case of a rare primary outcome, usually all primary outcome cases are selected. Then, each primary outcome case is matched by one or (optimally) more secondary outcome cases.

The Modeling Sample

- + Similar predictive power with smaller case count
- Must adjust assessment statistics and graphics
- Must adjust prediction estimates for bias



43

...

The advantage of separate sampling is that you are able to obtain (on the average) a model of similar predictive power with a smaller overall case count. This is in concordance with the idea that the amount of information in a data set with a categorical outcome is determined not by the total number of cases in the data set itself, but instead by the number of cases in the rarest outcome category. (For binary target data sets, this is usually the primary outcome.) (Harrell 2006)

This advantage might seem of minimal importance in the age of extremely fast computers. (A model might fit 10 times faster with a reduced data set, but a 10-second model fit versus a 1-second model fit is probably not relevant.) However, the model-fitting process occurs only after the completion of a long, tedious, and error-prone data preparation process. Smaller sample sizes for data preparation are usually welcome.

While it reduces analysis time, separate sampling also introduces some analysis complications.

- Most model fit statistics (especially those related to prediction decisions) and most of the assessment plots are closely tied to the outcome proportions in the training samples. If the outcome proportions in the training and validation samples do not match the outcome proportions in the scoring population, model performance can be greatly misestimated.
- If the outcome proportions in the training sample and scoring populations do not match, model prediction estimates are biased.

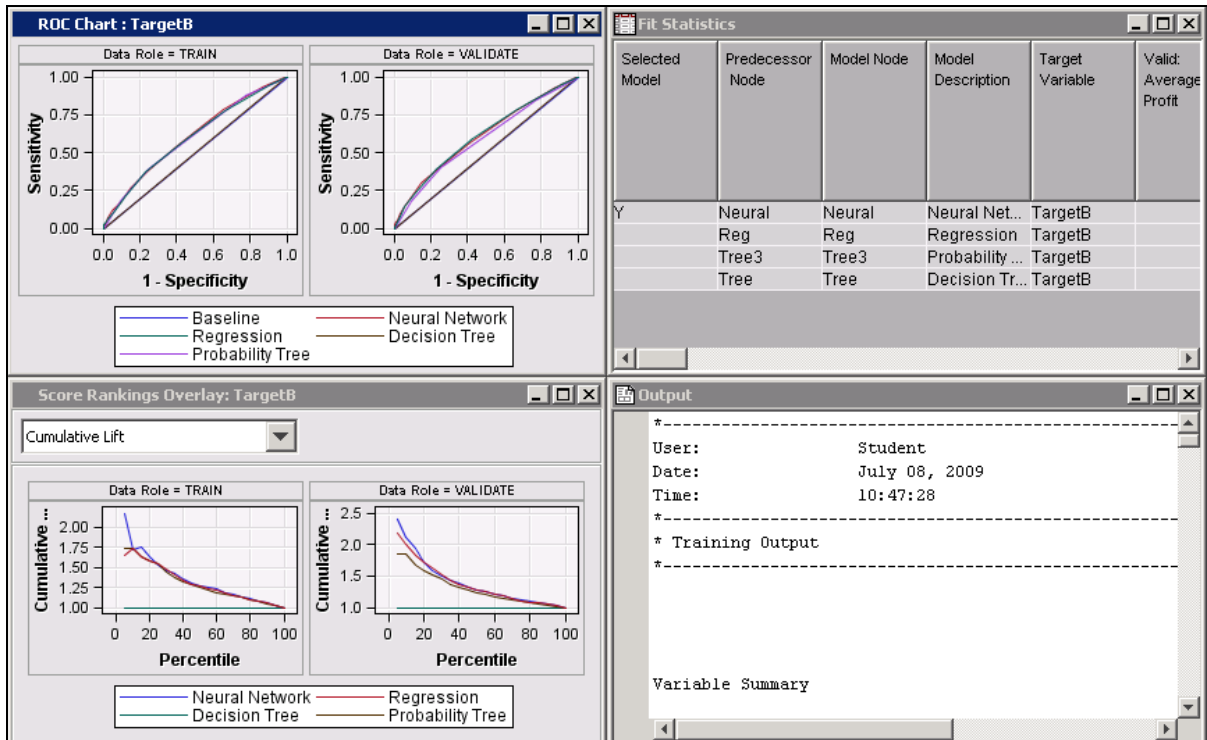
Fortunately, SAS Enterprise Miner can adjust assessments and prediction estimates to match the scoring population if you specify *prior probabilities*, the scoring population outcome proportions. This is precisely what was done using the Decisions option in the demonstration.



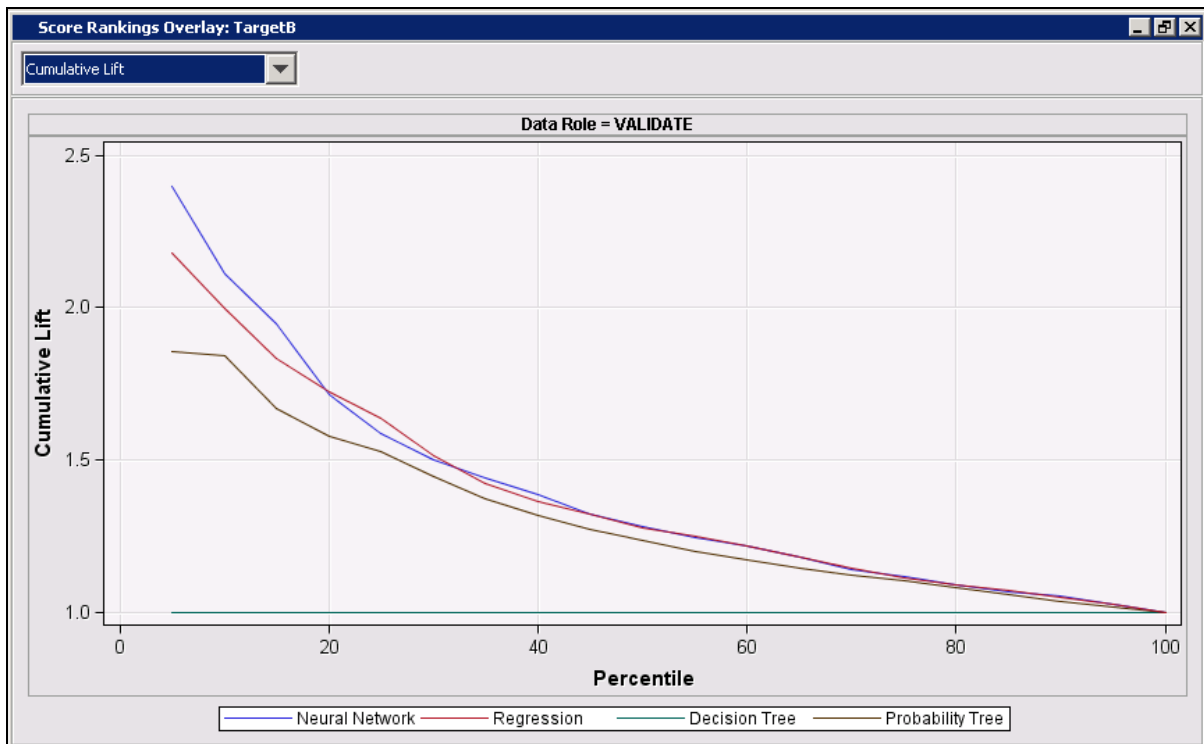
Adjusting for Separate Sampling (continued)

The consequences of incorporating prior probabilities in the analysis can be viewed in the Model Comparison node.

1. Select **Results...** in the Run Status dialog box. The Results - Model Comparison window opens.



2. Maximize the Score Rankings Overlay window and focus on the Data Role = VALIDATE chart.

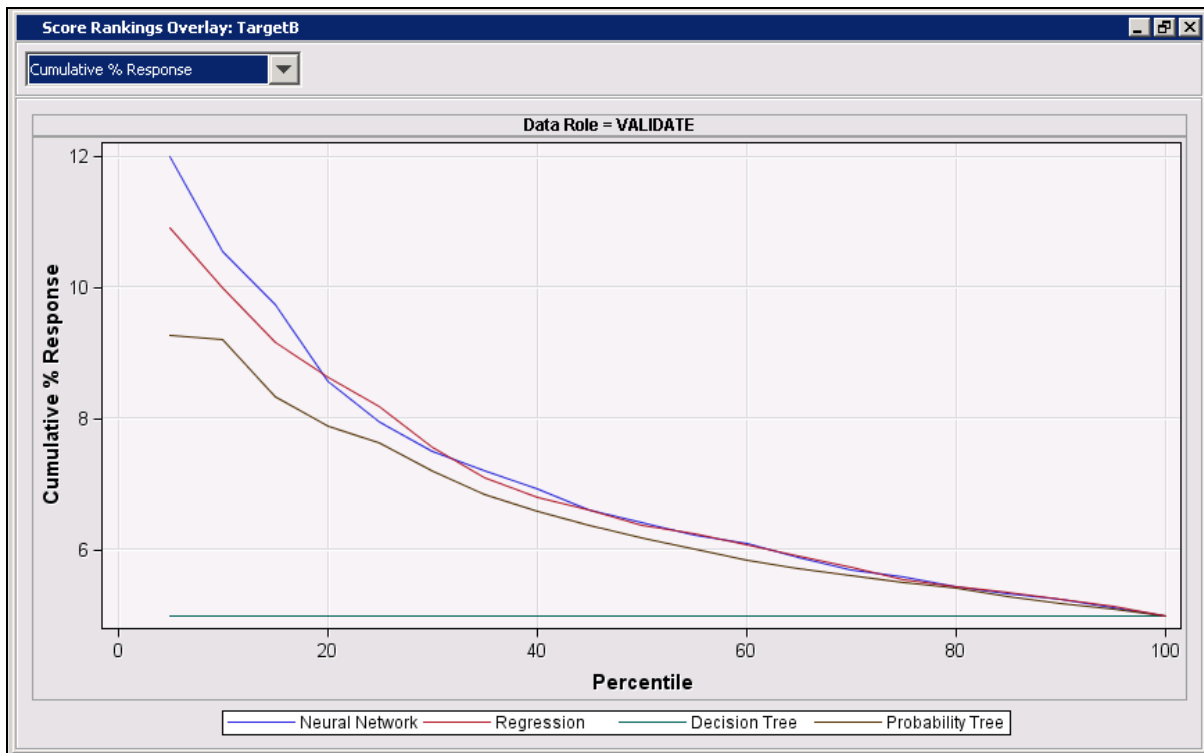


Before you adjusted for priors, none of the models had a cumulative lift over 1.5. Now most of the models have a cumulative lift in excess of 1.5 (at least for some deciles).



The exception is the Decision Tree model. Its lift is exactly 1 for all percentiles. The reason for this is the inclusion of disparate prior probabilities for a model tuned on misclassification. On the average, cases have a primary outcome probability of 0.05. A low misclassification rate model might be built by simply calling everyone a non-responder.

3. Select **Chart:** ⇒ **Cumulative % Response**.



The cumulative percent responses are now adjusted for separate sampling. You now have an accurate representation of response proportions by selection fraction.

With this plot, a business analyst could rate the relative performance of each model for different selection fractions. The best selection fraction is usually determined by financial considerations. For example, the charity might have a budget that allows contact with 40% of the available population. Thus, it intuitively makes sense to contact the 40% of the population with the highest chances of responding (as predicted by one of the available models). Another financial consideration, however, is also important: the profit (and loss) associated with a response (and non-response) to a solicitation. To correctly rank the value of a case, response probability estimates must be combined with profit and loss information.



Creating a Profit Matrix

To determine reasonable values for profit and loss information, consider the outcomes and the actions you would take given knowledge of these outcomes. In this case, there are two outcomes (response and non-response) and two corresponding actions (solicit and ignore). Knowing that someone is a responder, you would naturally want to solicit that person; knowing that someone is a non-responder, you would naturally want to ignore that person. (Organizations **really do not** want to send junk mail.) On the other hand, knowledge of an individual's actual behavior is rarely perfect, so mistakes are made, for example, soliciting non-responders (false positives) and ignoring responders (false negatives).

Taken together, there are four outcome/action combinations:

	Solicit	Ignore
Response		
Non-response		

Each of these outcome/action combinations has a profit consequence (where a loss is called, somewhat euphemistically, *a negative profit*). Some of the profit consequences are obvious. For example, if you do not solicit, you do not make any profit. So for this analysis, the second column can be immediately set to zero.

	Solicit	Ignore
Response		0
Non-response		0

From the description of the analysis problem, you find that it costs about \$0.68 to send a solicitation. Also, the variable **TargetD** gives that amount of response (when a donation occurs). The completed profit consequence matrix can be written as shown.

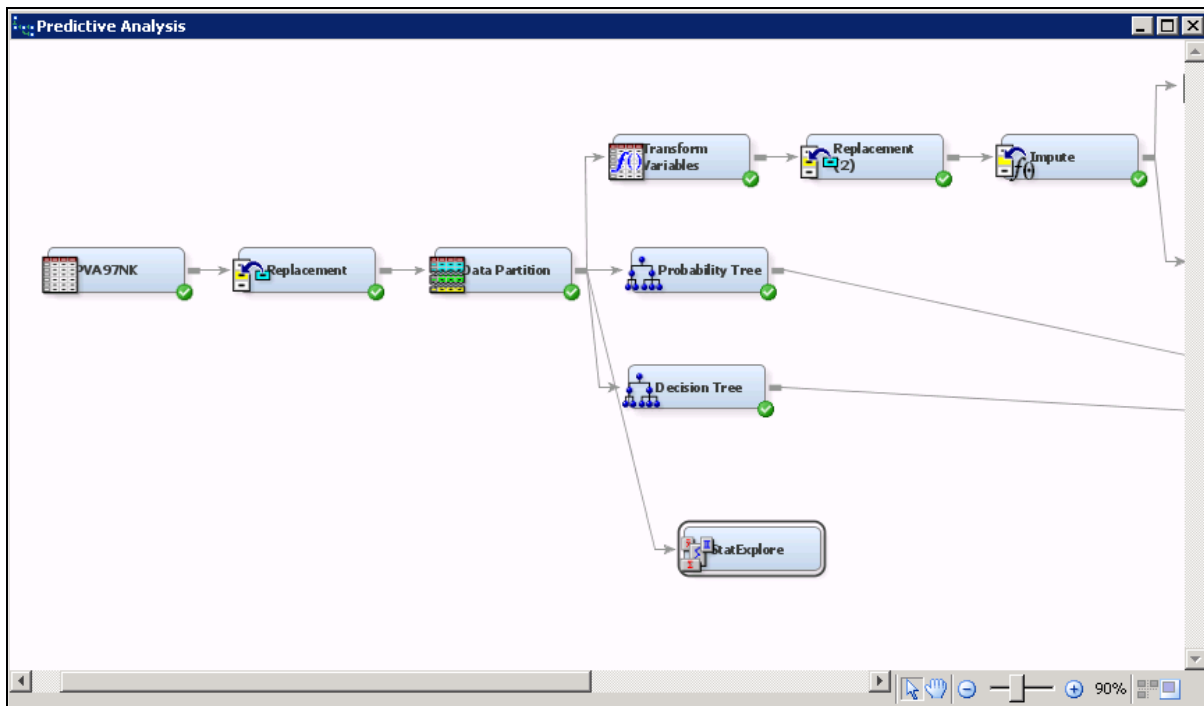
	Solicit	Ignore
Response	TargetD -0.68	0
Non-response	-0.68	0

From a statistical perspective, **TargetD** is a random variable. Individuals who are identical on every input measurement might give different donation amounts. To simplify the analysis, a summary statistic for **TargetD** is plugged into the profit consequence matrix. In general, this value can vary on a case-by-case basis. However, for this course, the overall average of **TargetD** is used.

You can obtain the average of **TargetD** using the StatExplore node.

1. Select the **Explore** tab.
2. Drag the **StatExplore** tool into the diagram workspace.

3. Connect the **StatExplore** node to the **Data Partition** node.



4. Select **Variables...** from the Properties panel of the StatExplore node.

The Variables - Stat window opens.

5. Select **Use** ⇒ **Yes** for the **TargetD** variable.

Variables - Stat

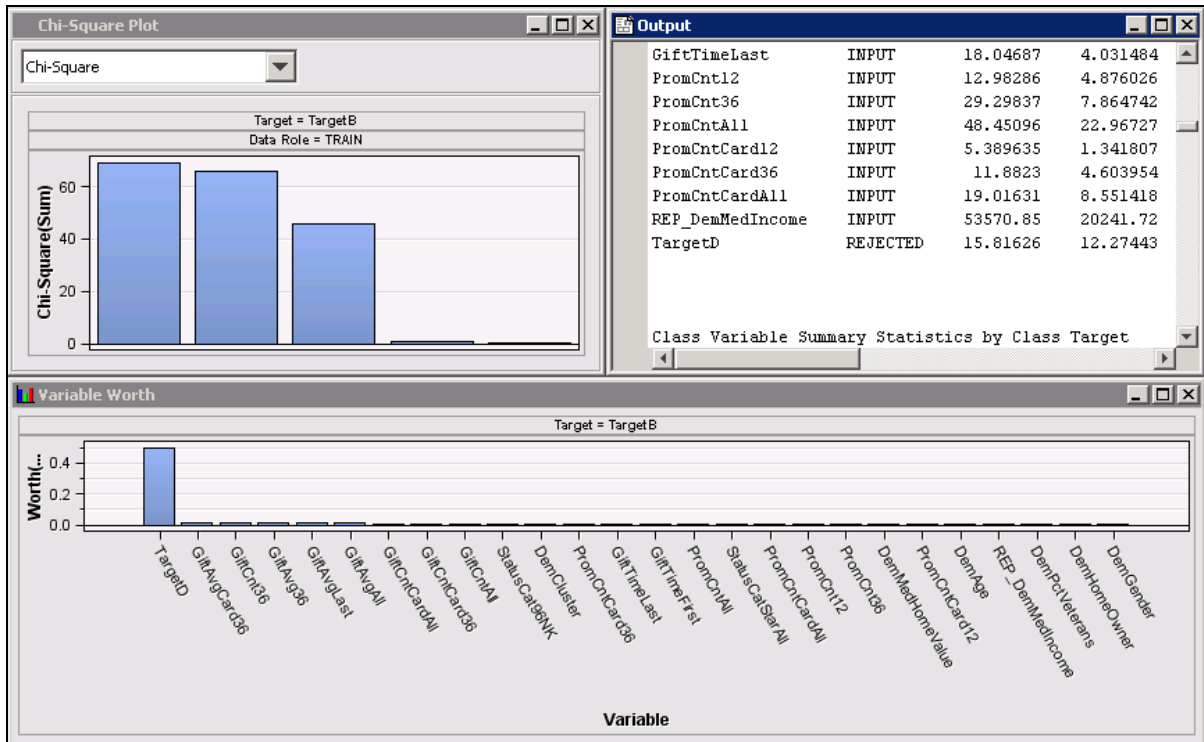
(none) ☐ not Equal to ☐

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Use	Report	Role	Level
DemAge	Default	No	Input	Interval
DemCluster	Default	No	Input	Nominal
DemGender	Default	No	Input	Nominal
DemHomeOw	Default	No	Input	Binary
DemMedHom	Default	No	Input	Interval
DemMedIncor	Default	No	Rejected	Interval
DemPctVetere	Default	No	Input	Interval
GiftAvg36	Default	No	Input	Interval
GiftAvgAll	Default	No	Input	Interval
GiftAvgCard36	Default	No	Input	Interval
GiftAvgLast	Default	No	Input	Interval
GiftCnt36	Default	No	Input	Interval
GiftCntAll	Default	No	Input	Interval
GiftCntCard36	Default	No	Input	Interval
GiftCntCardAll	Default	No	Input	Interval
GiftTimeFirst	Default	No	Input	Interval
GiftTimeLast	Default	No	Input	Interval
ID	Default	No	ID	Nominal
PromCnt12	Default	No	Input	Interval
PromCnt36	Default	No	Input	Interval
PromCntAll	Default	No	Input	Interval
PromCntCard	Default	No	Input	Interval
PromCntCard	Default	No	Input	Interval
PromCntCard	Default	No	Input	Interval
REP_DemMed	Default	No	Input	Interval
StatusCat96N	Default	No	Input	Nominal
StatusCatStar	Default	No	Input	Binary
TargetB	Default	No	Target	Binary
TargetD	Yes	No	Rejected	Interval
dataobs_	Default	No	ID	Interval

6. Select **OK** to close the Variables dialog box.

7. Run the StatExplore node and view the results. The Results - StatExplore window opens.



Scrolling the Output window shows the average of **TargetD** as \$15.82.

8. Close the Results - Stat window.
9. Select the **PVA97NK** node.
10. Select the **Decisions...** property. The Decision Processing - PVA97NK dialog box opens.

The screenshot shows the Decision Processing - PVA97NK dialog box. It has four tabs: "Targets", "Prior Probabilities", "Decisions", and "Decision Weights". The "Targets" tab is selected, showing a list of targets with "TargetB" selected. The main area displays the following properties for TargetB:

- Name : TargetB
- Measurement Level : Binary
- Target level order : Descending
- Event level : 1
- Format :

At the bottom right, there is a "Refresh" button. At the very bottom, there are "OK" and "Cancel" buttons.

11. Select the **Decisions** tab.

Decision Processing - PVA97NK

Targets **Prior Probabilities** **Decisions** **Decision Weights**

Do you want to use the decisions?

☒ **Yes** ☐ **No** **Default with Inverse Prior Weights**

Decision Name	Label	Cost Variable	Constant
DECISION1	1	< None >	0.0
DECISION2	0	< None >	0.0

Add **Delete** **Delete All** **Reset** **Default**

OK **Cancel**

12. Type **solicit** (in place of the word **DECISION1**) in first row of the Decision Name column.
13. Type **ignore** (in place of the word **DECISION2**) in the second row of the Decision Name column.

Decision Processing - PVA97NK

Targets **Prior Probabilities** **Decisions** **Decision Weights**

Do you want to use the decisions?

☒ **Yes** ☐ **No** **Default with Inverse Prior Weights**

Decision Name	Label	Cost Variable	Constant
solicit	1	< None >	0.0
ignore	0	< None >	0.0

Add **Delete** **Delete All** **Reset** **Default**

OK **Cancel**

14. Select the **Decision Weights** tab.

Decision Processing - PVA97NK

Targets Prior Probabilities Decisions **Decision Weights**

Select a decision function:

☒ **Maximize** ☐ **Minimize**

Enter weight values for the decisions.

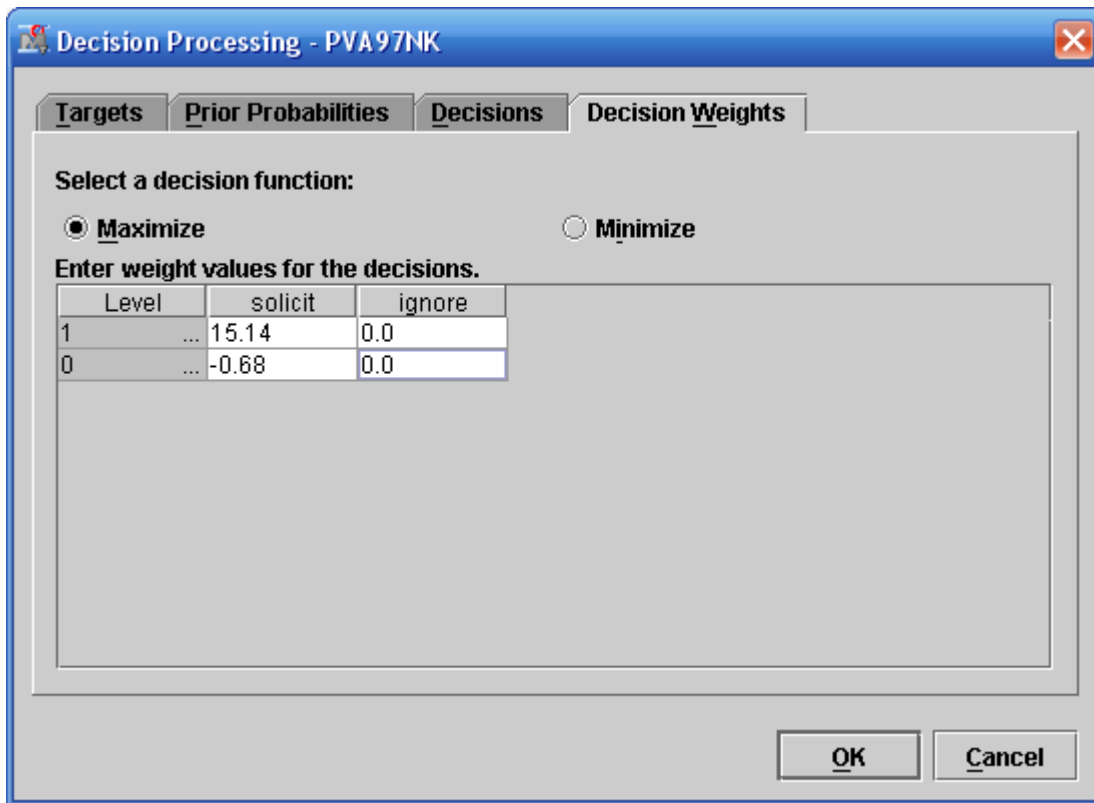
Level	solicit	ignore
1 ...	1.0	0.0
0 ...	0.0	1.0

OK Cancel

The completed profit consequence matrix for this example is shown below.

	Solicit	Ignore
Response	15.14	0
Non-response	-0.68	0

15. Type the profit values into the corresponding cell of the profit weight matrix.



The image shows a dialog box titled "Decision Processing - PVA97NK". It has four tabs: "Targets", "Prior Probabilities", "Decisions", and "Decision Weights". The "Decision Weights" tab is selected. Inside the dialog, there is a section "Select a decision function:" with two radio buttons: "Maximize" (selected) and "Minimize". Below this is a section "Enter weight values for the decisions." containing a table with three columns: "Level", "solicit", and "ignore". The table has two rows of data. At the bottom right of the dialog are "OK" and "Cancel" buttons.

Level	solicit	ignore
1	15.14	0.0
0	-0.68	0.0

16. Select **OK** to close the Decision Processing - PVA97NK dialog box.
17. Run the Model Comparison node.



It will take some time to run the analysis.

6.4 Profit Matrices



Statistical decision theory is an aid to making optimal decisions from predictive models. Using decision theory, each target outcome is matched to a particular decision or course of action. A *profit value* is assigned to both correct (and incorrect) outcome and decision combinations. The profit value can be random and vary between cases.

A vast simplification and common practice in prediction is to assume that the profit associated with each case, outcome, and decision is a constant. This is the default behavior of SAS Enterprise Miner.

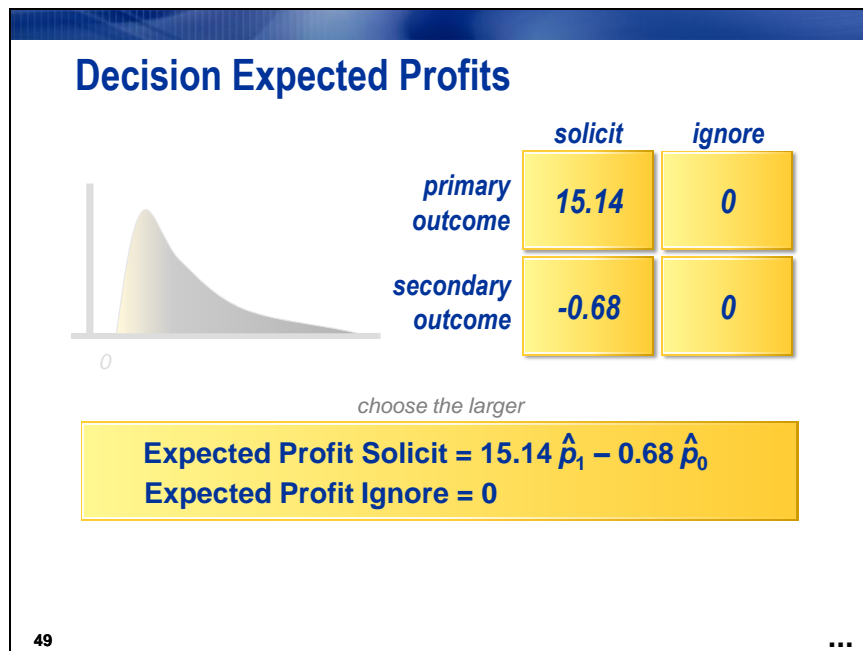


This simplifying assumption can lead to biased model assessment and incorrect prediction decisions.

For the demonstration, the overall donation average minus the solicitation cost is used as the (constant) profit associated with the primary outcome and the solicit decision.



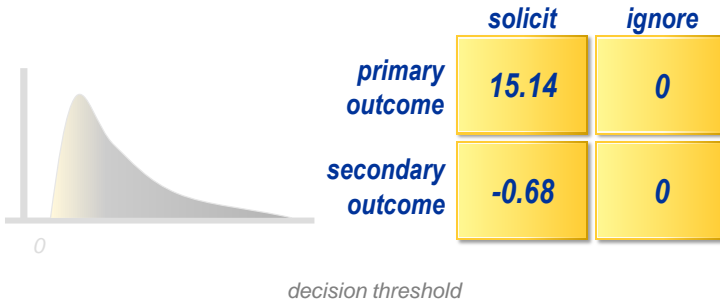
Similarly, the solicitation cost is taken as the profit associated with the secondary outcome and the solicit decision.



Making the reasonable assumption that there is no profit associated with the ignore decision, you can complete the profit matrix as shown.

With the completed profit consequence matrix, you can calculate the expected profit associated with each decision. This is equal to the sum of the outcome/action profits multiplied by the outcome probabilities. The best decision for a case is the one that maximizes the expected profit for that case.

Decision Threshold



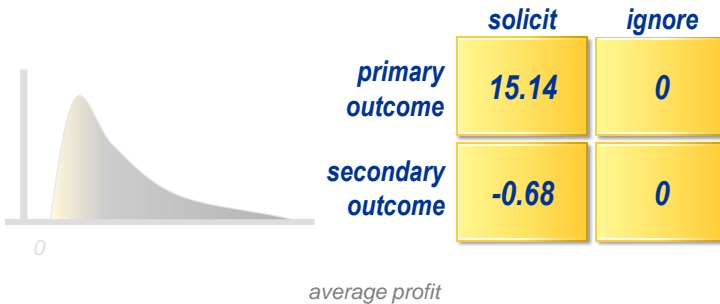
$$\hat{p}_1 \geq 0.68 / 15.82 \Rightarrow \text{Solicit}$$

$$\hat{p}_1 < 0.68 / 15.82 \Rightarrow \text{Ignore}$$

50

When the elements of the profit consequence matrix are constants, prediction decisions depend solely on the estimated probability of response and a constant decision threshold, as shown above.

Average Profit



$$\text{Average profit} = (15.14 \times N_{PS} - 0.68 \times N_{SS}) / N$$

N_{PS} = # solicited primary outcome cases

N_{SS} = # solicited secondary outcome cases

N = total number of assessment cases

51

A new fit statistic, *average profit*, can be used to summarize model performance. For the profit matrix shown, average profit is computed by multiplying the number of cases by the corresponding profit in each outcome/decision combination, adding across all outcome/decision combinations, and dividing by the total number of cases in the assessment data.

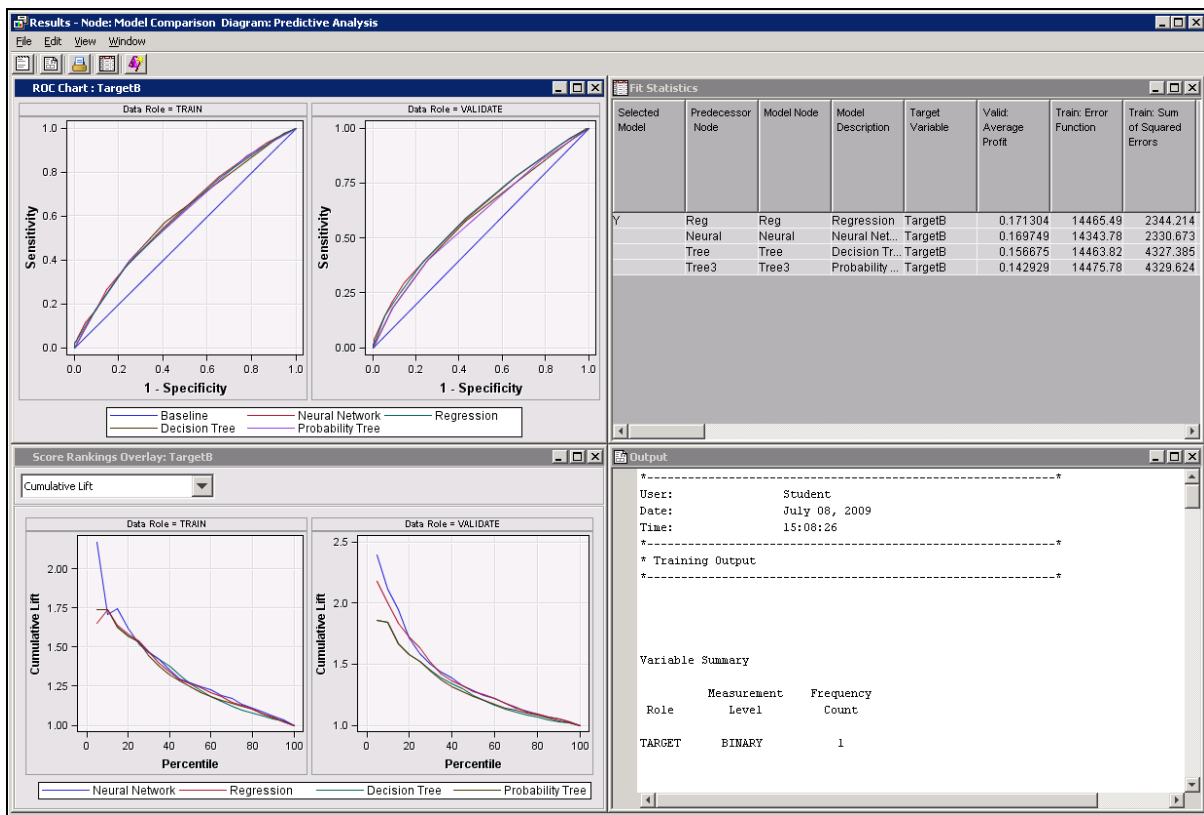


Evaluating Model Profit

The consequences of incorporating a profit matrix in the analysis can be viewed in the Model Comparison node.

Follow these steps to evaluate a model with average profit:

1. Select **Results...** in the Run Status dialog box. The Results - Node: Model Comparison window opens.



2. Maximize the Output window and go to line 30.

Fit Statistics							
Model Selection based on Valid: Average Profit (_VAPROF_)							
			Train:		Valid:		
			Valid: Average		Train:		Valid:
			Average Squared		Misclassification		Squared Misclassification
Selected Model	Model		Profit	Error	Rate	Error	Rate
Model	Node	Description					
Y	Reg	Regression	0.17130	0.24202	0.49990	0.24079	0.50010
	Neural	Neural Network	0.16975	0.24062	0.49990	0.23988	0.50010
	Tree	Decision Tree	0.15667	0.44677	0.49990	0.44689	0.50010
	Tree3	Probability Tree	0.14293	0.44700	0.49990	0.44720	0.50010

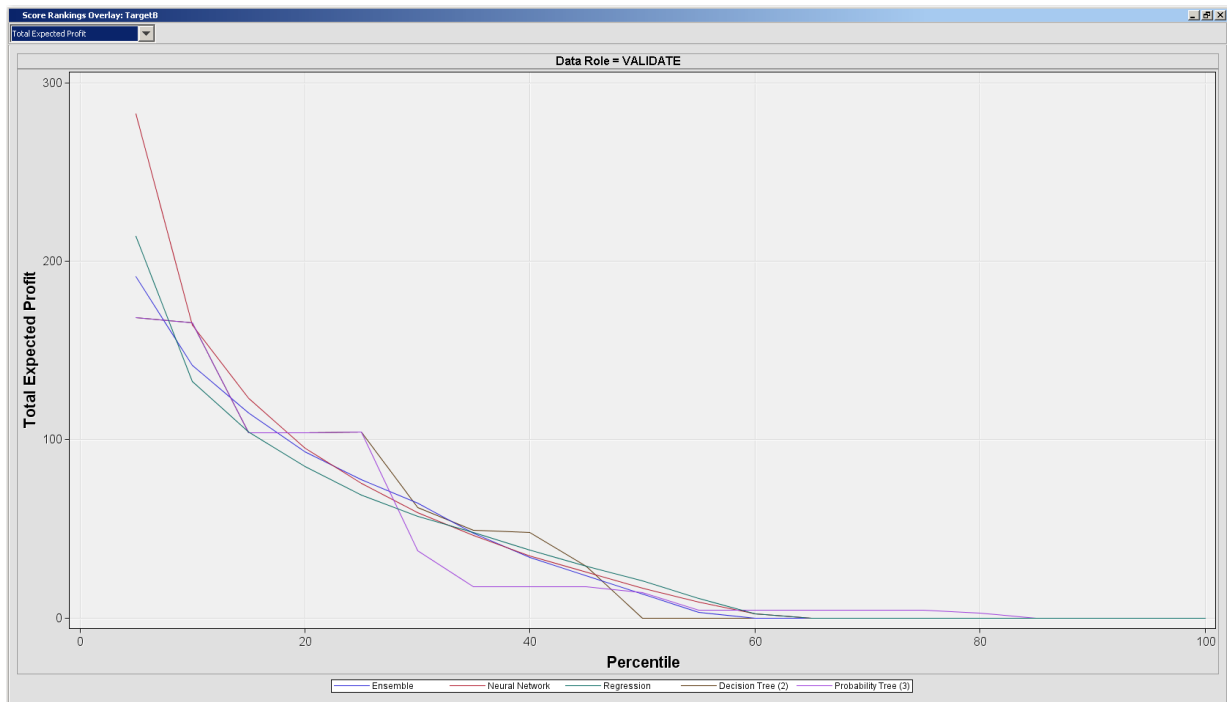
With a profit matrix defined, model selection in the Model Assessment node is performed on validation profit. Based on this criterion, the selected model is the Regression Model. The Neural Network is a close second.



Viewing Additional Assessments

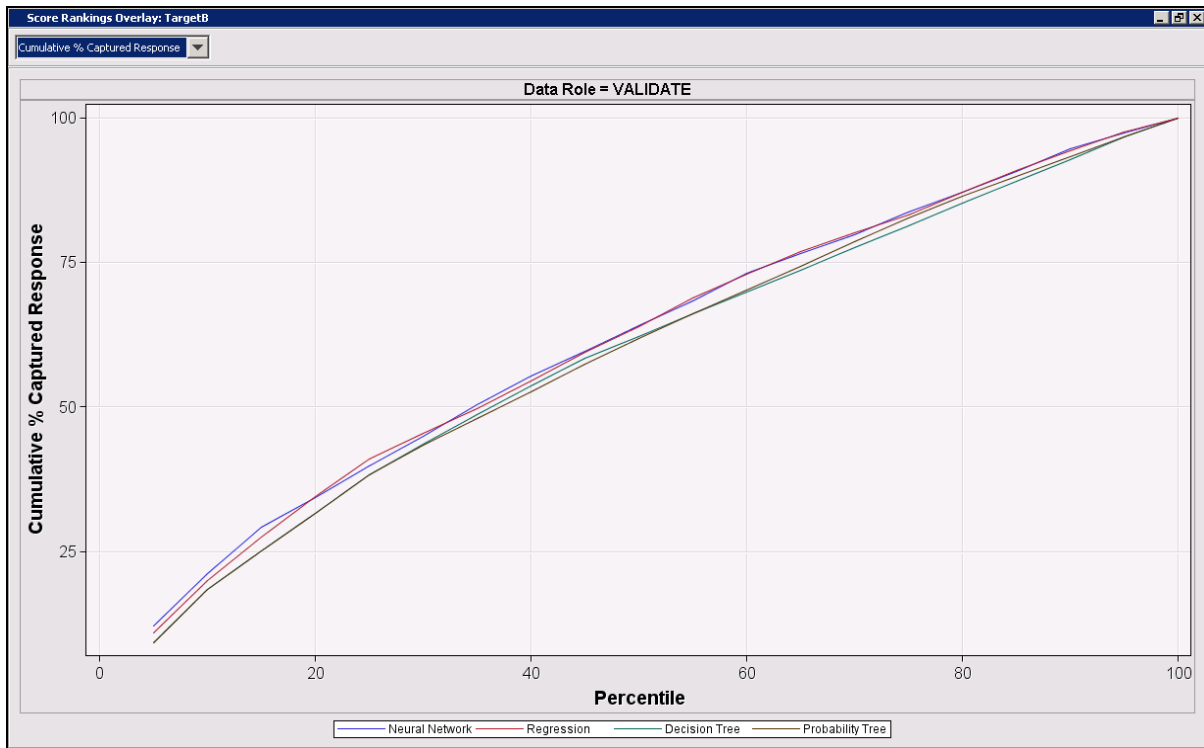
This demonstration shows several other assessments of possible interest.

1. Maximize the Score Rankings Overlay window.
2. Select **Select Chart:** ⇒ **Total Expected Profit.**



The Total Expected Profit plot shows the amount of value with each demi-decile (5%) block of data. It turns out that all models select approximately 60% of the cases (although cases in one model's 60% might not be in another model's 60%).

3. Select **Select Chart:** ⇒ **Cumulative % Captured Response**.
4. Double-click the **Data Role = VALIDATE** chart.



This plot shows sensitivity (also known as *Cumulative % Captured Response*) versus selection fraction (Decile). By selecting 60% of the data, for example, you “capture” about 75% of the primary-outcome cases.

5. Close the Results window.



Optimizing with Profit (Self-Study)

The models fit in the previous demonstrations were optimized to minimize average error. Because it is the most general optimization criterion, the best model selected by this criterion can rank and decide cases. If the ultimate goal of your model is to create prediction decisions, it might make sense to optimize on that criterion.

After you define a profit matrix, it is possible to optimize your model strictly on profit. Instead of seeking the model with the best prediction estimates, you find the model with best prediction decisions (those that maximize expected profit).



The default model selection method in SAS Enterprise Miner is validation profit optimization, so these settings essentially restore the node defaults. Finding a meaningful profit matrix for most modeling scenarios, however, is difficult. Therefore, these notes recommend overriding the defaults and creating models with a general selection criterion such as validation average squared error.

Decision Tree Profit Optimization

One of the two tree models in the diagram is set up this way:

1. Select the original **Decision Tree** node.
2. Examine the Decision Tree node's Subtree property.

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

The Decision Tree node is pruned using the default assessment measure, Decision. The goal of this tree is to make the best decisions rather than the best estimates.

3. Examine the Probability Tree node's Subtree property.

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

The Probability Tree node is pruned using average squared error. The goal of this tree is to provide the best prediction estimates rather than the best decision.

Regression Profit Optimization

Use the following settings to optimize a regression model using profit:

1. Select the **Regression** node.
2. Select **Selection Criterion** ⇒ **Validation Profit/Loss**.

Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Profit/Loss
Use Selection Defaults	No
Selection Options	

3. Repeat steps 1 and 2 for the Polynomial Regression node.

Neural Network Profit Optimization

1. Select the **Neural Network** node.
2. Select **Model Selection Criterion** ⇒ **Profit/Loss**.

Train	
Variables	
Network	
Model Selection Criterion	Profit/Loss
Use Current Estimates	No



The AutoNeural node does not support profit matrices.

The next step refits the models using profit optimization. Be aware that the demonstrations in Chapters 8 and 9 assume that the models are fit using average squared error optimization.

3. Run the Model Comparison node and view the results.
4. Maximize the Output window and view lines 20-35.

Fit Statistics							
Model Selection based on Valid: Average Profit (_VAPROF_)							
			Train:		Valid:		
			Valid: Average		Train:		Valid:
			Average Squared		Average Squared		Misclassification
			Profit		Error		Rate
Selected Model	Model	Model Description	Profit	Error	Rate	Error	Rate
Y	Neural	Neural Network	0.17660	0.23645	0.49990	0.24096	0.50010
	Reg	Regression	0.15934	0.23711	0.49990	0.24419	0.50010
	Tree	Decision Tree	0.15667	0.44677	0.49990	0.44689	0.50010
	Tree3	Probability Tree	0.14293	0.44700	0.49990	0.44720	0.50010

The reported validation profits are now slightly higher, and the relative ordering of the models changed.



Exercises

1. Assessing Models

- a. Connect all models in the ORGANICS diagram to a Model Comparison node.
- b. Run the Model Comparison node and view the results.

Which model has the best ROC curve? _____

What is the corresponding ROC Index? _____

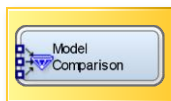
- c. What is the lift of each model at a selection depth of 40%? _____

6.5 Chapter Summary

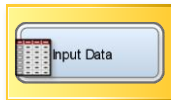
The Model Comparison node enables you to compare statistics and statistical graphics summarizing model performance. To make assessments applicable to the scoring population, you must account for differences in response proportions between the training, validation, and scoring data.

While you can choose to select a model based strictly on statistical measures, you can also consider profit as a measure of model performance. For each outcome, an appropriate decision must be defined. A profit matrix is constructed, characterizing the profit associated with each outcome and decision combination. The decision tools in SAS Enterprise Miner support only constant profit matrices.

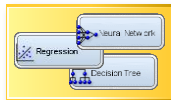
Assessment Tools Review



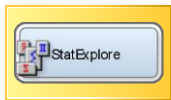
Compare model summary statistics and statistical graphics.



Create decision data; add prior probabilities and profit matrices.



Tune models with average squared error or appropriate profit matrix.



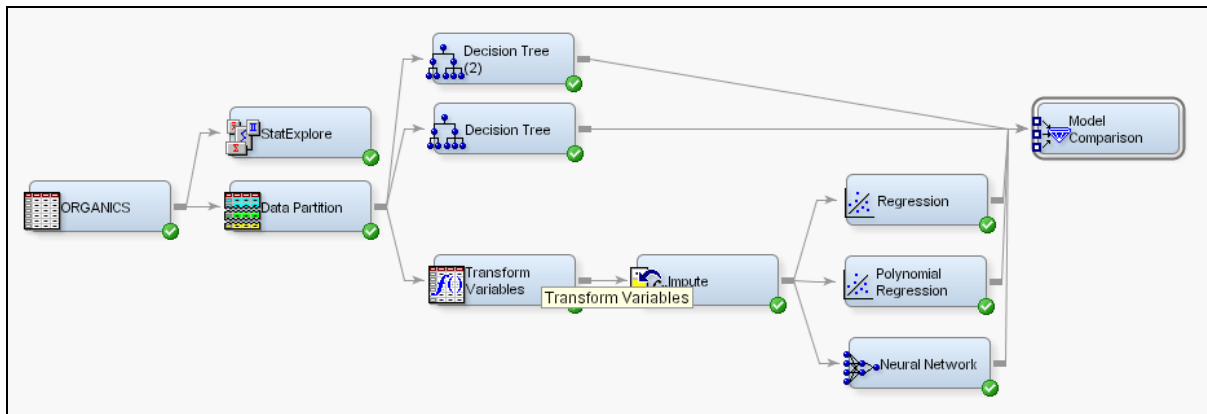
Obtain means and other statistics on data source variables.

6.6 Solutions

Solutions to Exercises

1. Assessing Models

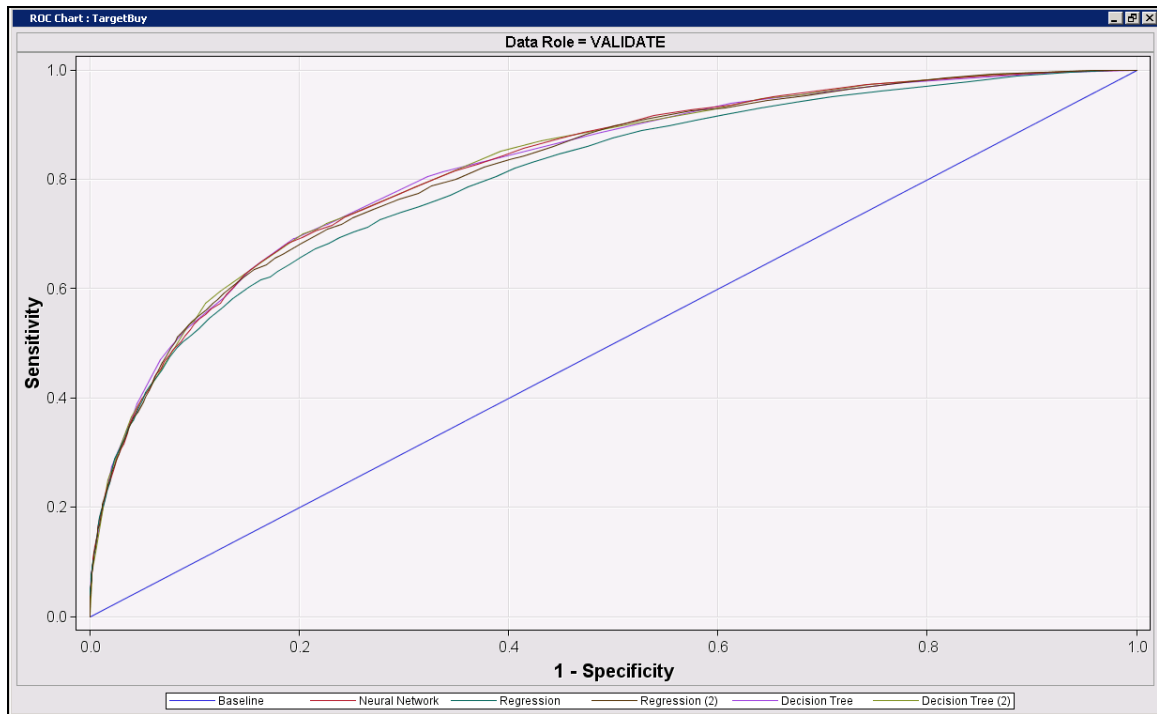
- a. Connect all models in the ORGANICS diagram to a Model Comparison node.



- b. Run the Model Comparison node and view the results.

Which model has the best ROC curve? **The Decision Tree seems to have the best ROC curve.**

What is the corresponding ROC index?



The ROC Index values are found in the Fit Statistics window.

Fit Statistics				
Selected Model	Predecessor Node	Model Node	Valid: Roc Index ▲	Model Description
Y	Reg	Reg	0.805062	Regression
	Reg2	Reg2	0.82007	Regression (2)
	Tree	Tree	0.823782	Decision Tree
	Neural	Neural	0.823818	Neural Network
	Tree2	Tree2	0.824339	Decision Tree (2)

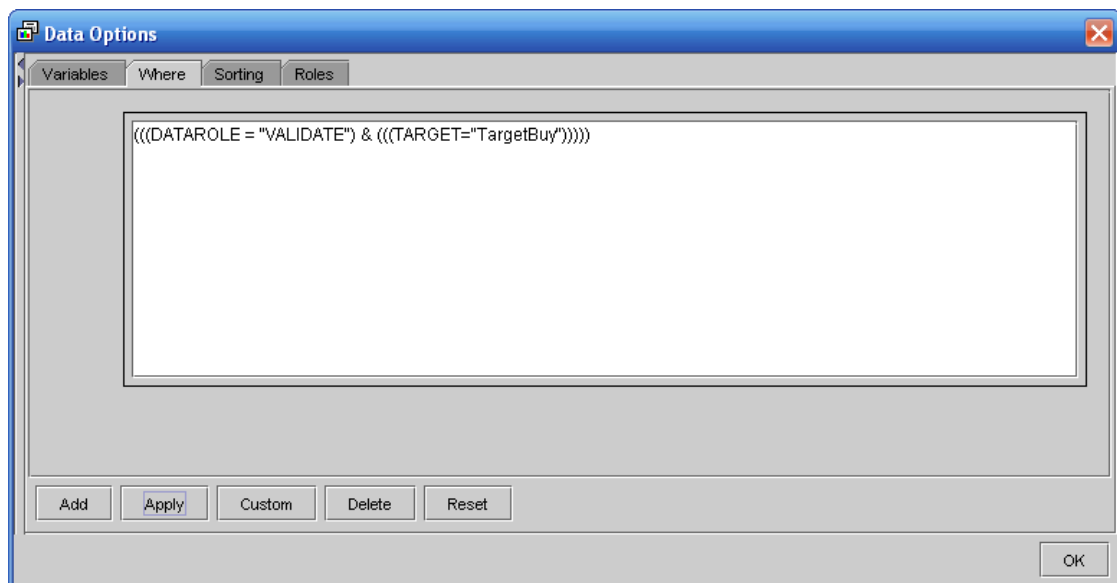
c. What is the lift of each model at a selection depth of 40%?

1) Select **Validation Score Ranking Chart**.

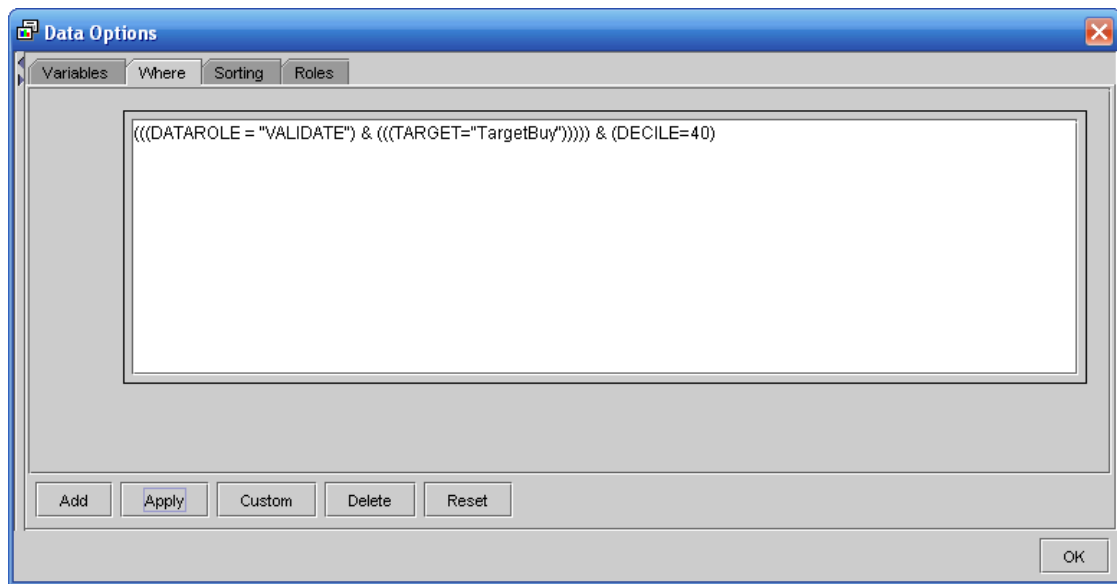


2) Right-click the chart and select **Data Options...** from the Option menu.

3) Select the **Where** tab.



- 4) Add the text **& (DECILE=40)** to the selection string.



- 5) Select **Apply**.

- 6) Select **OK**.

The Score Rankings plot shows the lift for the 40th percentile. You can read the values for each model from the plot.

