Developing Credit Scorecards Using Credit Scoring for SAS® Enterprise Miner™ 13.1

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *Developing Credit Scorecards Using Credit Scoring for SAS® Enterprise Miner™ 13.1*. Cary, NC: SAS Institute Inc.

Developing Credit Scorecards Using Credit Scoring for SAS® Enterprise Miner™ 13.1

Copyright © 2013, SAS Institute Inc., Cary, NC, USA

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

January 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.

13.1-P2:emcsgs

Contents

Using This Book . . . . . *v*

Chapter 1 • Introduction to Credit Scoring for SAS Enterprise Miner **1**
What Is Credit Scoring for SAS Enterprise Miner? 1
Getting to Know the Graphical User Interface 2

Chapter 2 • Set Up the Project **5**
About the Tasks That You Will Perform 5
Create a New Project 5
Create a Data Source 6
Create a Diagram 7
Partition the Data 8

Chapter 3 • Explore the Data and Create the Scorecard **9**
About the Tasks That You Will Perform 9
Group the Characteristic Variables into Attributes 9
Create a Scorecard with a Logistic Regression Model 14
Perform Reject Inference on the Model 21
Create the Final Scorecard 25

Appendix 1 • References **29**

Glossary **31**
Index **35**

Using This Book

Audience

This tutorial covers how to use Credit Scoring for SAS Enterprise Miner to build a consumer credit scorecard. The tutorial assumes that you are familiar with the process of credit scoring. It is focused on reviewing the features and functionality of the core set of credit scoring nodes, and should not be considered a complete review of the capabilities of SAS Enterprise Miner. The analysis typically would include other important steps, such as exploratory data analysis, variable selection, model comparison, and scoring.

Requirements

Credit Scoring for SAS Enterprise Miner is not included with the base version of SAS Enterprise Miner 13.1. If your site has not licensed Credit Scoring for SAS Enterprise Miner, the credit scoring node tools will not appear in your SAS Enterprise Miner 13.1 software.

Chapter 1

Introduction to Credit Scoring for SAS Enterprise Miner

What Is Credit Scoring for SAS Enterprise Miner? 1

Getting to Know the Graphical User Interface 2

What Is Credit Scoring for SAS Enterprise Miner?

Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques describe who should get credit, how much credit they should receive, and which operational strategies will enhance the profitability of the borrowers to the lenders (Thomas, Edelman, and Crook 2002).

Credit Scoring, as defined by SAS, includes the following:

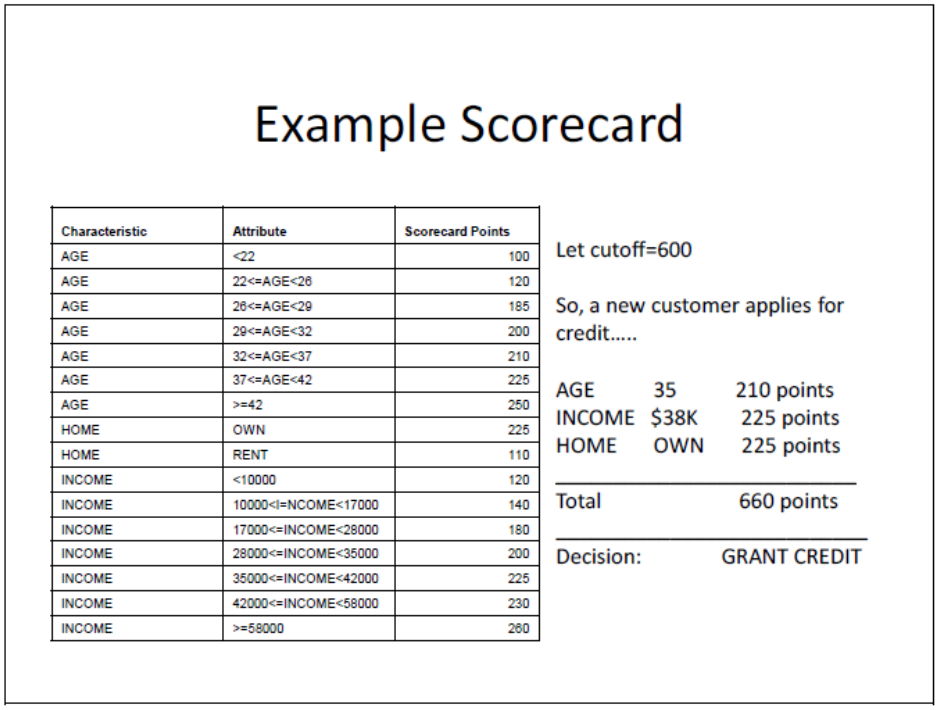
- applying a statistical model to assign a risk score to a credit application or an existing credit account
- building the statistical model
- monitoring the accuracy of one or more statistical models
- monitoring the effect that score-based decisions have on key business performance indicators

Although credit scoring is not as glamorous as pricing exotic financial derivatives, it is one of the most successful applications of statistical and operations research techniques in finance and banking. Without an accurate and automated risk assessment tool, the phenomenal growth of consumer credit would not have been possible over the past 40 years (Thomas, Edelman, and Crook 2002).

In its simplest form, a scorecard is built from a number of characteristics (that is, input or predictor variables). Each characteristic includes a number of attributes. For example, age is a characteristic, and “25-33” is an attribute. Each attribute is associated with a number of scorecard points. These scorecard points are statistically assigned to differentiate risk, based on the predictive power of the characteristic variables, correlation between the variables, and business considerations.

For example, using the Example Scorecard in Figure 1.1, an applicant who is 35, makes $38,000, and is a homeowner would be accepted for credit by this financial institution’s scorecard. The total score of an applicant is the sum of the scores for each attribute that is present in the scorecard. Lower scores imply a higher risk of default, and higher scores imply a lower risk of default.

Figure 1.1 *Example Scorecard*



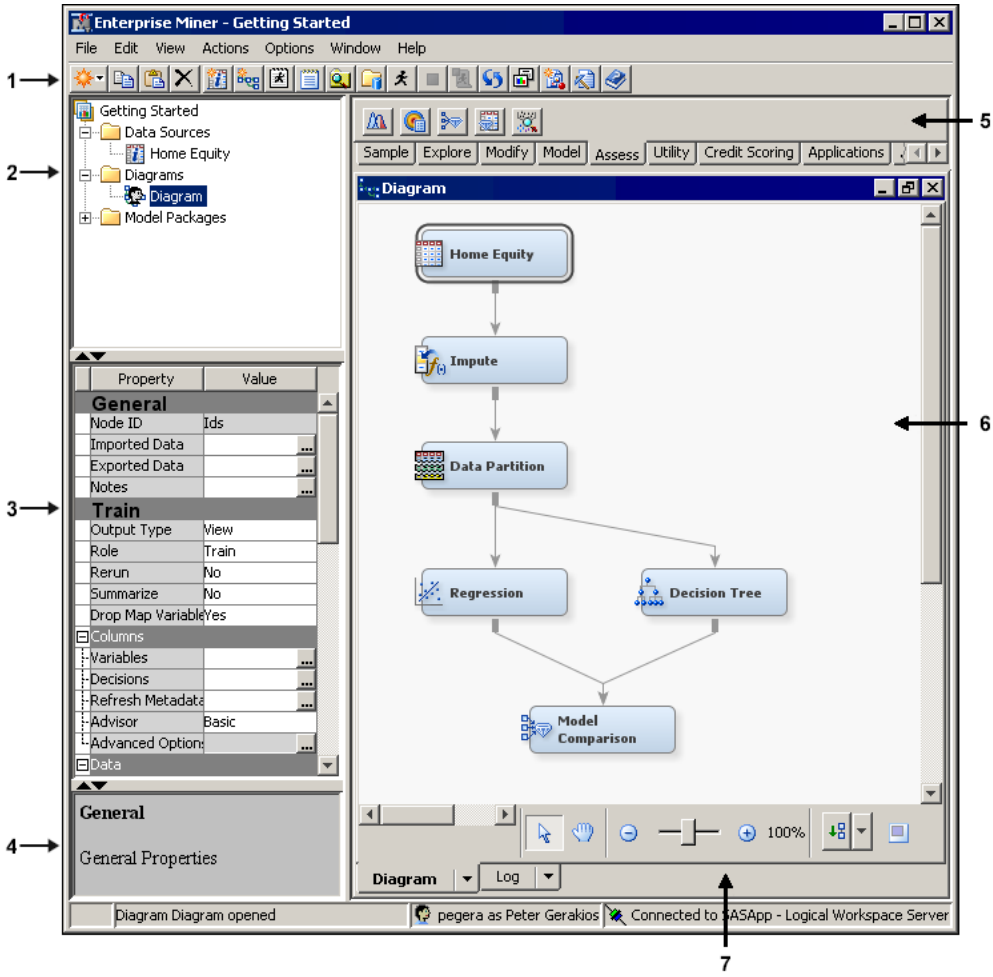
Credit Scoring for SAS Enterprise Miner contains the following nodes, which are added to your SAS Enterprise Miner toolbar to support scorecard development:

- **Interactive Grouping** — groups input variables into bins before the credit scorecard is built. An initial, automatic grouping can provide optimal splits, but this node enables you to regroup the variables through an interactive interface. It also has the capability to screen or select variables.
- **Scorecard** — uses the grouped variables as inputs in a logistic regression model and usually follows the Interactive Grouping node. In addition, it scales the regression parameters to compute score points and the resulting scorecard. Finally, the Scorecard node performs score and characteristic (variable) analysis that helps in understanding the scorecard, and aids in crafting a score-based strategy.
- **Reject Inference** — offers three standard, industry-accepted methods for inferring the performance of the rejected applicant data by the use of a model that is built on the accepted applicants.
- **Credit Exchange** — enables the use of scorecards in SAS Credit Risk for Banking. Because it plays no part in the development of the scorecard, coverage of this node is beyond the scope of this tutorial.

Getting to Know the Graphical User Interface

You use the SAS Enterprise Miner Graphical User Interface (GUI) to build a process flow diagram that controls your data mining project.

Figure 1.2 *The SAS Enterprise Miner GUI*



1. Toolbar Shortcut Icons — Use the toolbar shortcut icons to perform common computer functions and frequently used SAS Enterprise Miner operations. Move the mouse pointer over any shortcut button to see the text name. Click on a shortcut icon to use it.
2. Project Panel — Use the Project Panel to manage and view data sources, diagrams, results, and project users.
3. Properties Panel — Use the Properties Panel to view and edit the settings of data sources, diagrams, nodes, and users.
4. Property Help Panel — The Property Help Panel displays a short description of any property that you select in the Properties Panel. Extended help can be found on the Help main menu.
5. Toolbar — The Toolbar is a graphic set of node icons that you use to build process flow diagrams in the Diagram Workspace. Drag a node icon to the Diagram Workspace to use it. The icon remains in place on the Toolbar, and the node in the Diagram Workspace is ready to be connected and configured for use in the process flow diagram.
6. Diagram Workspace — Use the Diagram Workspace to build, edit, run, and save process flow diagrams. In this workspace, you graphically build, order, sequence, and connect the nodes that you use to mine your data and generate reports.

7. Diagram Navigation Toolbar — Use the Diagram Navigation Toolbar to organize and navigate the process flow diagram.

Chapter 2

Set Up the Project

About the Tasks That You Will Perform 5
Create a New Project 5
Create a Data Source 6
Create a Diagram 7
Partition the Data 8

About the Tasks That You Will Perform

To set up the example project, you will perform the following tasks:

1. You will create a new SAS Enterprise Miner project.
2. You will define a new library that enables SAS Enterprise Miner to access the sample data.
3. You will define two new data sources in the project.
4. You will create a new diagram within the project.
5. You will partition the input data source.

The steps in this example are written as if you were completing them in their entirety during one SAS Enterprise Miner session. However, you can easily complete the steps over multiple sessions. To return to the example project after you have closed and reopened SAS Enterprise Miner, click **Open Project** in the Welcome to Enterprise Miner window, and navigate to the saved project.

Create a New Project

In SAS Enterprise Miner, you store your work in projects. A project can contain multiple process flow diagrams and information that pertains to them.

To create the project that you will use in this example:

1. Open SAS Enterprise Miner.

2. In the Welcome to Enterprise Miner window, click **New Project**. The Create New Project Wizard opens.
3. Proceed through the steps that are outlined in the wizard. Contact your system administrator if you need to be granted directory access or if you are unsure about the details of your site's configuration.
   a. Select the logical workspace server to use. Click **Next**.
   b. Enter `Getting Started with Credit Scoring` as the **Project Name**.

      The **SAS Server Directory** is the directory on the server machine in which SAS data sets and other files that are generated by the project will be stored. It is likely that your site is configured in such a way that the default path is appropriate for this example. Click **Next**.
   c. The **SAS Folder Location** is the directory on the server machine in which the project itself will be stored. It is likely that your site is configured in such a way that the default path is appropriate for the example project that you are about to create. Click **Next**.

      Note: If you complete this example over multiple sessions, then this is the location to which you should navigate after you select **Open Project** in the Welcome to Enterprise Miner window.
   d. Click **Finish**.

Create a Data Source

In this section, you define the CS_ACCEPTS SAS data set as a SAS Enterprise Miner data source. A SAS Enterprise Miner data source defines all the information about a SAS table or a view to another file type that is needed for data mining. This information includes the name and location of the data set, variable roles, measurement levels, and other attributes that inform the data mining process. After they are defined, the data sources can be used in any diagram within a project and can be copied from one project to another.

It is important to note that data sources are not the actual training data, but instead is the metadata that defines the source data. The source data itself must reside in an allocated library. This project uses data in the SAMPSIO library.

To create a new data source for the sample data:

1. On the **File** menu, select **New** ⇨ **Data Source**. The Data Source Wizard opens.
2. Proceed through the steps that are outlined in the wizard.
   a. **SAS Table** is automatically selected as the **Source**. Click **Next**.
   b. Enter `SAMPSIO.CS_ACCEPTS` as the two-level filename of the **Table**. Click **Next**.
   c. The Data Source Wizard — Table Information window appears. Metadata is data about data sets. Some metadata, such as field names, is stored with the data. Other metadata, such as how a particular variable in a data set should be used in a predictive model, must be manually specified. When you define modeling metadata, you are establishing relevant facts about the data set prior to model construction.

      Click **Next**.

d. Click **Advanced**. Use the Advanced option when you want SAS Enterprise Miner to automatically set the variable roles and measurement levels. Automatic initial roles and level values are based on the variable type, the variable format, and the number of distinct values contained in the variable.

   Click **Next**.

e. In the Data Source Wizard — Column Metadata window, change the value of **Role** for the variables to match the description below.

   - _freq_ should have the **Role** Frequency.
   - GB should have the **Role** Target.
   - All other variables should have the **Role** Input.

   To change an attribute, click on the value of that attribute and select from the drop-down menu that appears. Click **Next**.

   You can use the **Show code** option to write SAS code to conditionally assign variable attributes. This is especially useful when you want to apply a metadata rule to several variables.

f. In the Data Source Wizard — Decision Configuration window, click **Next**.

g. In the Data Source Wizard — Create Sample window, click **Next**.

h. The **Role** of the data source is automatically selected as **Raw**. Click **Next**.

i. Click **Finish**.

The CS_ACCEPTS data source has been added to your project.

To add the CS_REJECTS data, complete the following steps:

1. On the **File** menu, select **New** ⇨ **Data Source**. The Data Source Wizard opens.
2. Proceed through the steps that are outlined in the wizard.
   a. **SAS Table** is automatically selected as the **Source**. Click **Next**.
   b. Enter **`SAMPSIO.CS_REJECTS`** as the two-level filename of the **Table**. Click **Next**.
   c. The Data Source Wizard — Table Information window appears. Click **Next**.
   d. Click **Advanced**. Click **Next**.
   e. In the Data Source Wizard — Column Metadata window, ensure that the value of **Role** for all variables is set to **Input**. Click **Next**.
   f. In the Data Source Wizard — Create Sample window, click **Next**.
   g. Change the **Role** of the data source to **Score**. Click **Next**.
   h. Click **Finish**.

Create a Diagram

Now that you have created a project and defined the data source, you are ready to begin building a process flow diagram.

To create a diagram and add the first node complete the following steps:

1. On the **File** menu, select **New** ⇨ **Diagram**.
2. Enter `CS Example` as the **Diagram Name**, and click **OK**. An empty diagram opens in the Diagram Workspace.
3. Select the CS_ACCEPTS data source in the Project Panel. Drag it to the Diagram Workspace; this action creates the input data node.

Although this example develops one process flow diagram, an advantage of SAS Enterprise Miner is that you can open multiple diagrams at the same time. You can also disconnect from and reconnect to a diagram as long as you have also configured the Web Infrastructure Platform as a middle-tier component when using a SAS Enterprise Miner solution. Other users can also access the same project. However, only one user can open a diagram at a time.

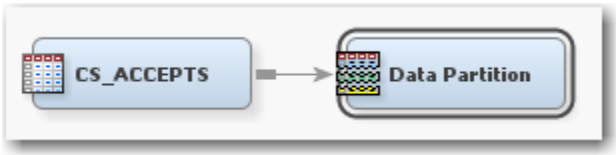
Partition the Data

In data mining, a strategy for assessing the quality of model generalization is to partition the data source. A portion of the data, called the *training data set*, is used for preliminary model fitting. The rest is reserved for empirical validation and is often split into two parts: validation data and test data. The *validation data set* is used to prevent a modeling node from overfitting the training data and to compare models. The *test data set* is used for a final assessment of the model.

Note: In SAS Enterprise Miner, the default data partitioning method for class target variables is to stratify the target variable or variables.

To use the Data Partition node to partition the input data into training and validation sets:

1. Select the **Sample** tab on the Toolbar.
2. Select the **Data Partition** node icon. Drag the node to the Diagram Workspace.
3. Connect the **CS_ACCEPTS** node to the **Data Partition** node.



4. Select the **Data Partition** node. In the Properties Panel, scroll down to view the data set allocations in the Train properties.
   - Click on the value of **Training**, and enter `70.0`
   - Click on the value of **Validation**, and enter `30.0`
   - Click on the value of **Test**, and enter `0.0`

   These properties define the percentage of input data that is used in each type of mining data set. In this example, you use a training data set and a validation data set, but you do not use a test data set.
5. In the Diagram Workspace, right-click the **Data Partition** node, and select **Run** from the resulting menu. Click **Yes** in the confirmation window that opens.
6. In the Run Status window, click **OK**.

Chapter 3

Explore the Data and Create the Scorecard

About the Tasks That You Will Perform 9
Group the Characteristic Variables into Attributes 9
Create a Scorecard with a Logistic Regression Model 14
Perform Reject Inference on the Model 21
Create the Final Scorecard 25

About the Tasks That You Will Perform

You have already set up the project and partitioned the input data. In this chapter, you will create the credit scorecard by performing the following tasks:

1. You will group the characteristic variables into attributes.
2. You will use a logistic regression model to create an initial scorecard.
3. You will perform reject inference on the logistic regression model.
4. You will create the final scorecard using the information that was obtained in the previous steps.

Note that the above steps are only part of the scorecard development process. Other tasks (such as exploratory data analysis, variable selection, and model comparison), while important, are not included in this tutorial.

Group the Characteristic Variables into Attributes

You will now use the Interactive Grouping node to perform variable grouping, which is a binning transformation performed on the input variables. Variable grouping is also referred to as classing.

1. From the **Credit Scoring** tab, drag an **Interactive Grouping** node to the Diagram Workspace. Connect the **Data Partition** node to the **Interactive Grouping** node.

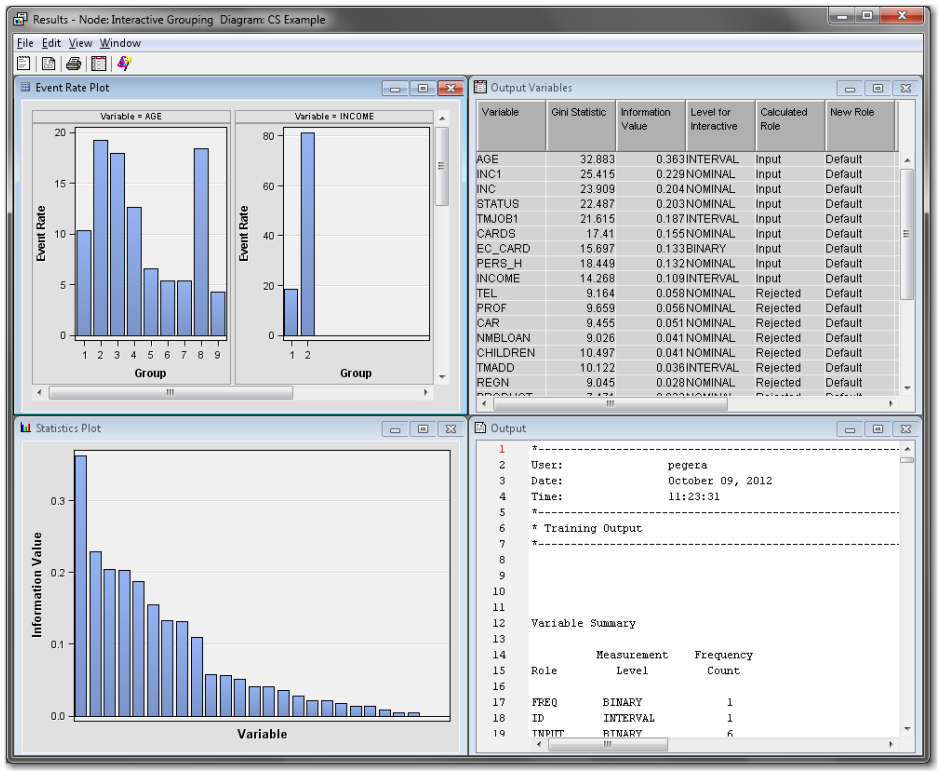
   The Interactive Grouping node performs the initial grouping automatically. You can use these initial groupings as a starting point to modify the classes interactively. By default, the unbinned interval variables are grouped into 20 quantiles (also called

bins), which are then grouped based on a decision tree. The Interactive Grouping node enables you to specify the properties of this decision tree.

2. Select the **Interactive Grouping** node in the Diagram Workspace. Set the value of the **Interval Grouping Method** property and the **Ordinal Grouping Method** property to **Monotonic Event Rate**.

   Set the value of the **Maximum Number of Groups** property to `10`.

3. Right-click the **Interactive Grouping** node and click **Run**. In the Confirmation window that appears, click **Yes**. In the Run Status window that appears, click **Results**.



The Output Variables window displays each variable's Gini Statistic and information value (IV). Note that a variable receives an **Exported Role** of **Rejected** if the variable's IV is less than 0.10. Recall that IV is used to evaluate a characteristic's overall predictive power (that is, the characteristic's ability to differentiate between good and bad loans). Information value is calculated as follows:

$$IV = \sum_{i=1}^{L} \left(DistrGood_i - DistrBad_i \right) \cdot \ln\left(\frac{DistrGood_i}{DistrBad_i} \right)$$

Here L is the number of attributes for the characteristic variable. In general an IV less than 0.02 is unpredictive, a value between 0.02 and 0.10 is weakly predictive, a value between 0.10 and 0.30 is moderately predictive, and a value greater than 0.30 is strongly predictive.

The Gini statistic is used as an alternative to the IV. The formula for the Gini statistic is more complicated than the IV and can be found in the SAS Enterprise Miner help documentation.

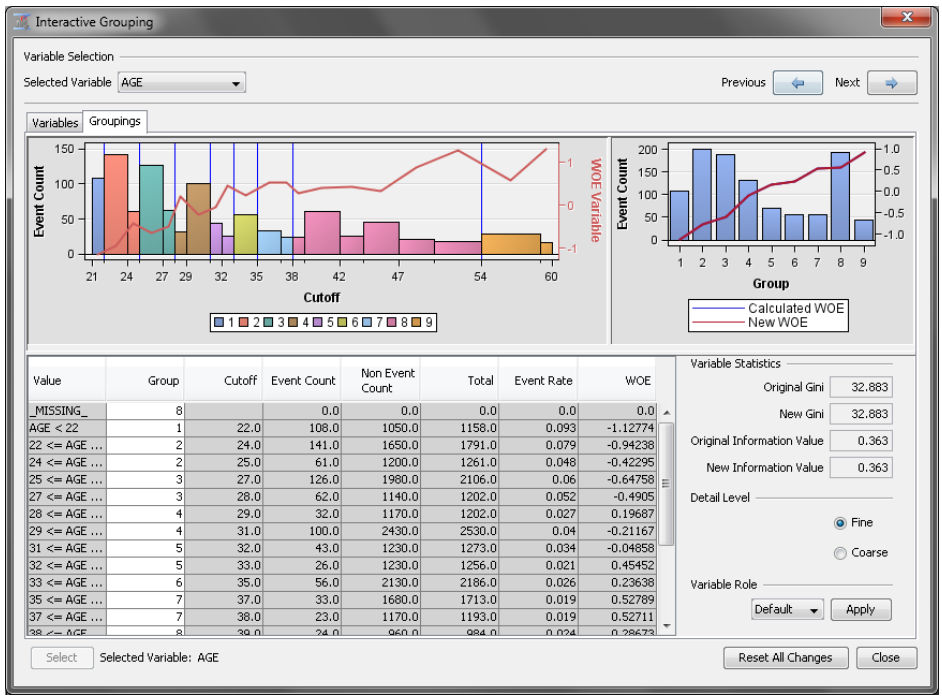
In the Properties Panel of the **Interactive Grouping** node, you can specify the cutoff values for the Gini and IV statistics. For example, the default IV cutoff of 0.10 for rejecting a characteristic can be changed to another value using the **Information Cutoff Value** property.

The Statistics Plot window shows a bar chart of each variable against its information value. You can move your cursor over a bar to display tooltip information, which includes the variable name, IV, and variable role.

Based on the IV, the variables AGE, INC1, INC, STATUS, TMJOB1, CARDS, EC_CARD, PERS_H, and INCOME are considered the candidate inputs to build the final scorecard in the regression step. The IV and Gini statistics might change if the groupings of the attributes are changed. The initial, automatic binning provides a good starting point to create the groupings for the scorecard, but the groupings can be fine-tuned.

Close the Results window.

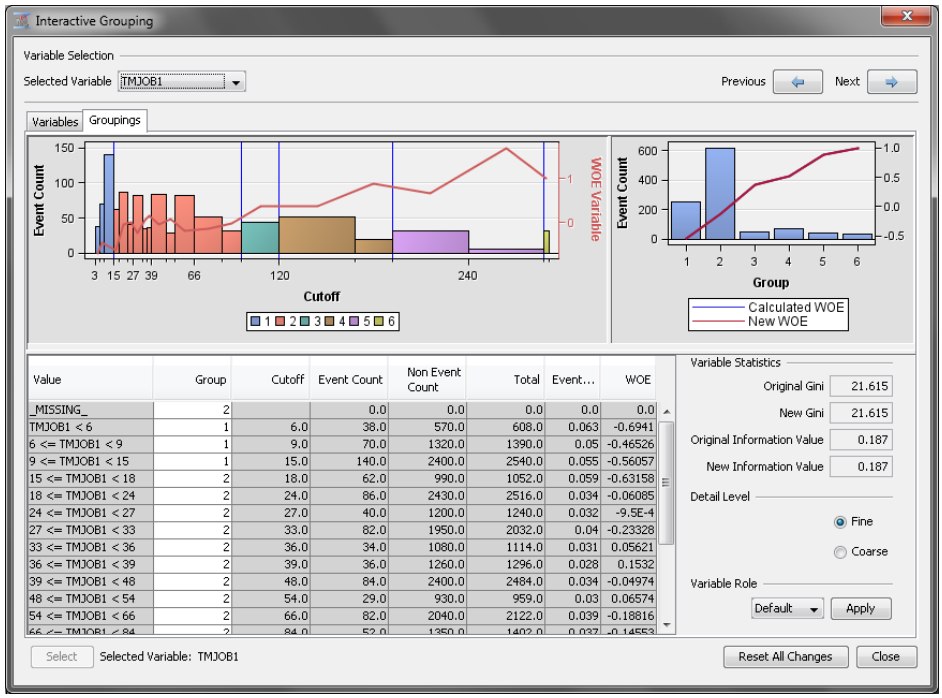
4. Now that you have run the node, you can open the Interactive Grouping application. Click the button in the **Interactive Grouping** property. This opens the Interactive Grouping window.



By default, the variables are sorted by their information value, given in the **Original Information Value** column. Also, the variable that is selected by default is the variable with the greatest IV. In this example, that variable is AGE.

Use the drop-down menu in the upper left corner of the Interactive Grouping window to select the variable **TMJOB1**. The variable TMJOB1 represents the applicant's time at their current job.

Select the **Groupings** tab in the Interactive Grouping window.



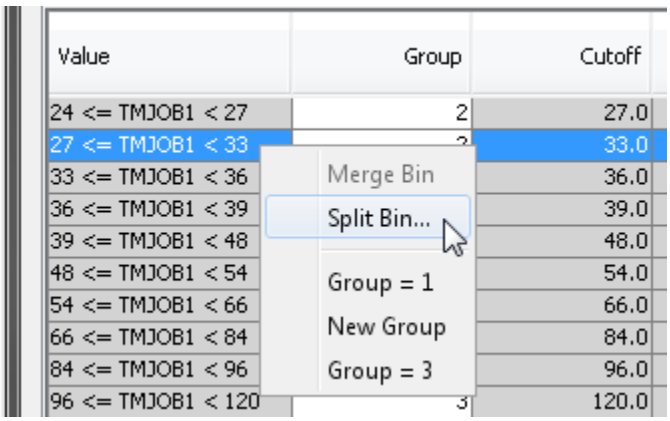
The plot on the right shows the weights of evidence for each group of the variable INC. Recall that the weight of evidence (WOE) measures the strength of an attribute of a characteristic in differentiating good and bad accounts. Weight of evidence is based on the proportion of good applicants to bad applicants at each group level. For each group i of a characteristic, WOE is calculated as follows:

$$WOE = \ln\left(\frac{DistrGood_i}{DistrBad_i}\right)$$

Negative values indicate that a particular grouping is isolating a higher proportion of bad applicants than good applicants. That is, negative WOE values are worse in the sense that applicants in that group present a greater credit risk. By default, missing values are assigned to their own group. The shape of the WOE curve is representative of how the points in the scorecard are assigned. As you can see on the **Groupings** tab, as time on the job increases, so does WOE.

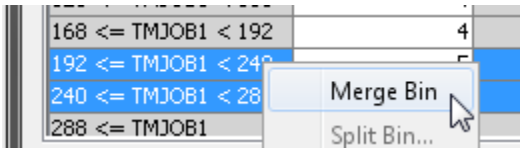
The plot on the left shows the details of each group for the selected variable. It shows the distribution of the bad loans within each group.

You can use the table to manually specify cutoff values. Suppose that you want to make 30 a cutoff value in the scorecard. Select the row that contains 30 in the score range, as shown below.



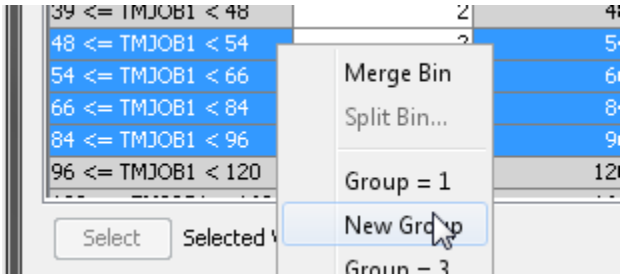
In the Split Bin window, enter `30` in the **Enter New Cutoff Value** dialog box. Click **OK**. Note that Group 2 now contains another bin that has a cutoff value of 30.

You can also use the **Groupings** tab to combine multiple bins within a group. Suppose that you want to combine the two bins in Group 5 into a single bin. Select the rows that correspond to Group 5, right-click one of the rows, and select **Merge Bin**.



Note that Group 5 now contains just one bin.

Finally, you can use the **Groupings** tab to create a new group from the defined bins. Suppose that you want Group 2 to contain fewer observations. Select the last four rows of Group 2, where the value of TMJOB1 is between 48 and 96, and then right-click the node and select **New Group**.



Note that there are now 7 groups, but this change did not have a significant impact on the WOE curve.

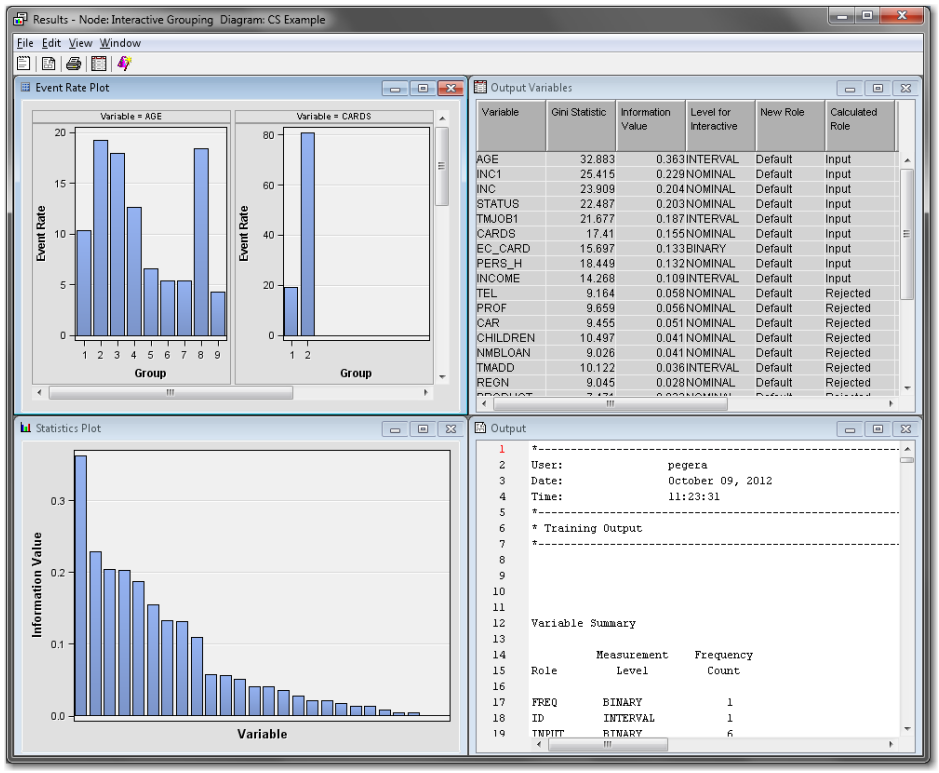
In general, changes to your grouping or binning will affect the WOE graph. For example, there might be a characteristic input that should have increasing, monotonic WOE values for each group. If the auto-binning of the Interactive Grouping node does not find these groupings, then the ability to fine-tune the groupings to achieve a monotonic WOE graph can be quite powerful.

The Interactive Grouping node can create groups that are based on several binning techniques, including statistically optimal ones. The user can change these bins based on business knowledge and known bias in the data to make the WOE trends logical. The changes previously made are suggested only if the analyst has expert knowledge and has a specific reason for changing the bins of a characteristic variable.

Also, after changes are made in the Interactive Grouping node as shown above, it is possible that the statistics for WOE, IV, and Gini can change. Some of the variables that were examined in the Results window might now not be candidates for input into the scorecard based on the IV and Gini statistics.

Close the Interactive Grouping window. Click **Yes** in the Save Changes window.

5. In the Diagram Workspace, right-click the **Interactive Grouping** node and click **Results**. Compare the new results to those observed earlier.



Notice that the original candidate variables are still candidate variables after the changes that you made in the Interactive Grouping window. Close the Results window.

Create a Scorecard with a Logistic Regression Model

You are now ready to use the grouped variables in a logistic regression model to create a scorecard.

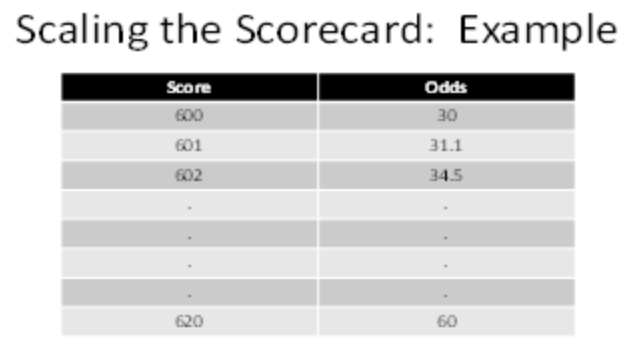
1. From the **Credit Scoring** tab, drag a **Scorecard** node to the Diagram Workspace. Connect the **Interactive Grouping** node to the **Scorecard** node.

The Scorecard node is used to develop a preliminary scorecard with logistic regression and scaling. In SAS Enterprise Miner, there are three types of logistic regression selection methods to choose: forward, backward, and stepwise. There is also a selection in the Properties Panel of the Scorecard node for no selection method, so that all variable inputs enter the model.

After the selection method is chosen, the regression coefficients are used to scale the scorecard. Scaling a scorecard refers to making the scorecard conform to a particular range of scores. Some reasons for scaling the scorecard are to enhance ease of interpretation, to meet legal requirements, and to have a transparent methodology that is widely understood among all users of the scorecard.

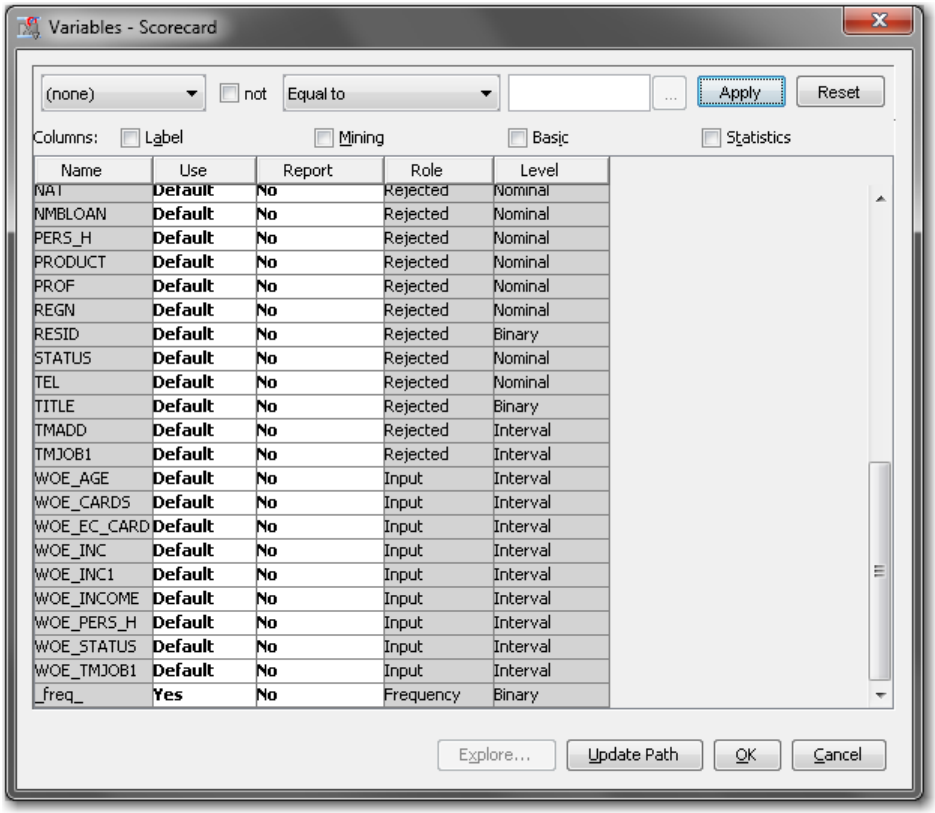
The two main elements of scaling a scorecard are to determine the odds at a certain score and to determine the points required to double the odds. Scorecard points are associated with both of these elements. Consider the scorecard shown in the figure below. The scorecard points are scaled so that a total score of 600 corresponds to good:bad odds of 30:1 and that an increase of 20 points corresponds to a doubling of the good:bad odds.

Figure 3.1 *Scaling the Scorecard: Points to Double the Odds*



The Scorecard node controls how a scorecard is scaled with the **Odds**, **Scorecard Points**, and **Points to Double the Odds** properties. The details of how SAS Enterprise Miner scales the scorecard are beyond the scope of this tutorial. See Siddiqi (2006) and Anderson and Hardin (2009) for a more detailed discussion about scorecard scaling.

2. In the Diagram Workspace, select the **Scorecard** node. Click the button in the **Variables** property.

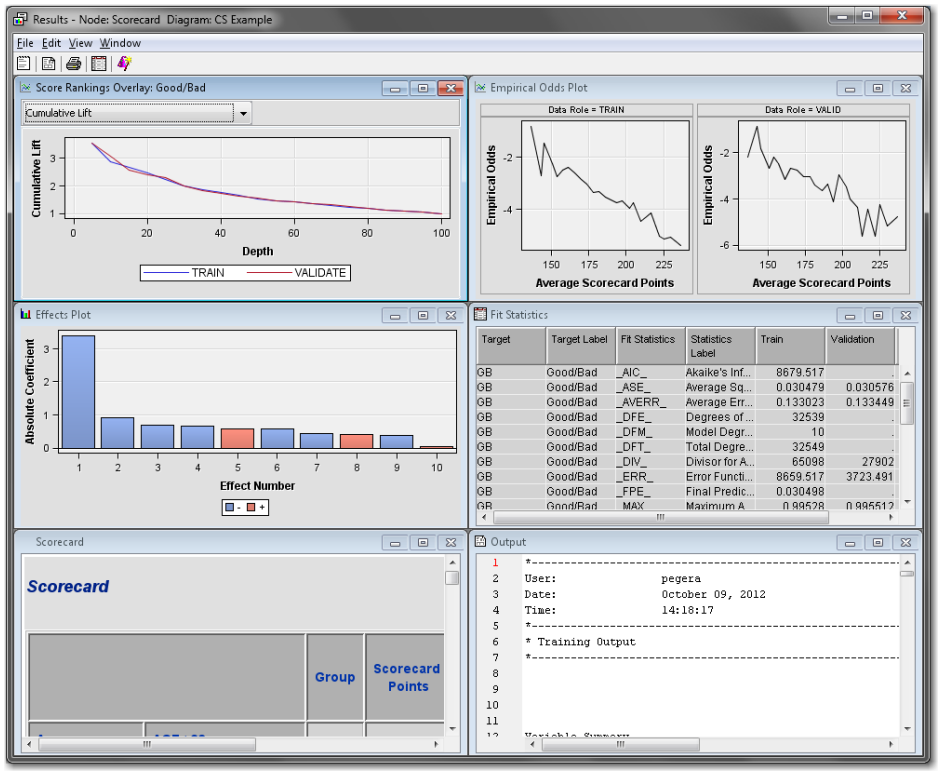


Note that for each original input variable there are now corresponding WOE_ and GRP_ variables. These were created by the Interactive Grouping node. Only the variables that exceed the Gini or IV cutoff set in the Interactive Grouping node are set to **Input**. All original inputs are set to **Rejected**.

Using the Scorecard node, you can use either the WOE_ variables, which contain the weight of evidence of each binned variable, or the GRP_ variables, which contain the group ID. The **Analysis Variables** property of the Scorecard node is used to specify whether regression is using WOE_ variables or GRP_ variables. The default is to use WOE_ variables.

Close the Variables window.

3. Change the value of the **Scorecard Type** field to **Detailed.**
4. Right-click the **Scorecard** node and click **Run**. In the Confirmation window, click **Yes**. In the Run Status window, click **Results**.



Maximize the Fit Statistics window. The Fit Statistics window displays fit statistics such as the average square error (ASE), the area under the receiver operating characteristic curve (AUR), and the Kolmogorov-Smirnov (KS) statistic, among others. Notice that the AUR is 0.71203 for the Validation data set.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
GB	Good/Bad	_AIC_	Akaike's Inf...	8679.517	.	.
GB	Good/Bad	_ASE_	Average Sq...	0.030479	0.030576	.
GB	Good/Bad	_AVERR_	Average Err...	0.133023	0.133449	.
GB	Good/Bad	_DFE_	Degrees of ...	32539	.	.
GB	Good/Bad	_DFM_	Model Degr...	10	.	.
GB	Good/Bad	_DFT_	Total Degre...	32549	.	.
GB	Good/Bad	_DIV_	Divisor for A...	65098	27902	.
GB	Good/Bad	_ERR_	Error Functi...	8659.517	3723.491	.
GB	Good/Bad	_FPE_	Final Predic...	0.030498	.	.
GB	Good/Bad	_MAX_	Maximum A...	0.99528	0.995512	.
GB	Good/Bad	_MSE_	Mean Squa...	0.030489	0.030576	.
GB	Good/Bad	_NOBS_	Sum of Fre...	32549	13951	.
GB	Good/Bad	_NW_	Number of ...	10	.	.
GB	Good/Bad	_RASE_	Root Avera...	0.174583	0.174859	.
GB	Good/Bad	_RFPE_	Root Final ...	0.174637	.	.
GB	Good/Bad	_RMSE_	Root Mean ...	0.17461	0.174859	.
GB	Good/Bad	_SBC_	Schwarz's ...	8763.422	.	.
GB	Good/Bad	_SSE_	Sum of Squ...	1984.147	853.1198	.
GB	Good/Bad	_SUMW_	Sum of Cas...	65098	27902	.
GB	Good/Bad	_MISC_	Misclassific...	0.032228	0.032327	.
GB	Good/Bad	_KS_	Kolmogoro...	0.319602	0.332062	.
GB	Good/Bad	_AUR_	Area Under ...	0.711516	0.71203	.
GB	Good/Bad	_Gini_	Gini Coeffic...	0.423032	0.42406	.
GB	Good/Bad	_ARATIO_	Accuracy R...	0.423032	0.42406	.

When you finish analyzing the fit statistics, minimize the Fit Statistics window.

Maximize the Scorecard window. The detailed Initial Scorecard displays information such as the scorecard points for each attribute, WOE, event rate (percentage of bad applicants in that score range), percentage of population, and the regression coefficient for each attribute. The **Percentage of Population** is the percentage of bad applicants who have a score higher than the lower limit of the score range.

Scorecard

Scorecard

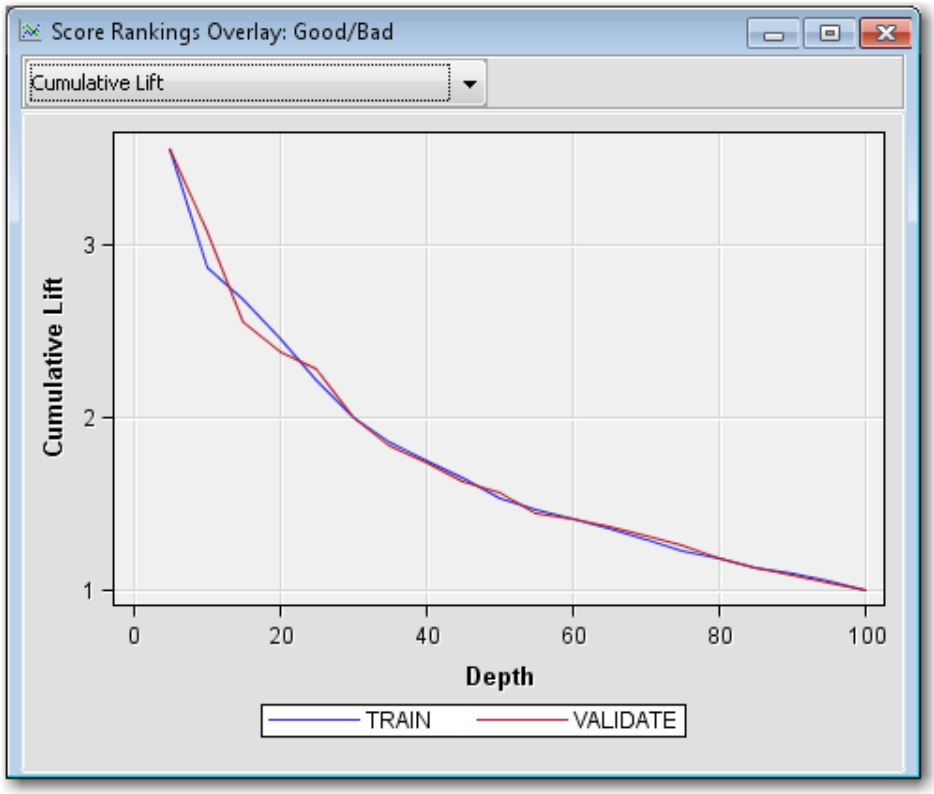
		Group	Scorecard Points	Weight of Evidence	Event Rate GB = 1	Percentage of Population	Coefficient
Age	AGE< 22	1.00	-1	-1.13	9.33	3.56	-0.65
	22<= AGE< 25	2.00	6	-0.76	6.62	9.38	-0.65
	25<= AGE< 28	3.00	9	-0.59	5.68	10.16	-0.65
	28<= AGE< 31	4.00	19	-0.10	3.54	11.47	-0.65
	31<= AGE< 33	5.00	24	0.17	2.73	7.77	-0.65
	33<= AGE< 35	6.00	25	0.24	2.56	6.72	-0.65
	35<= AGE< 38	7.00	31	0.53	1.93	8.93	-0.65
	38<= AGE< 54,						

When you finish analyzing the Initial Scorecard, minimize the Scorecard window.

Maximize the Score Rankings Overlay window. By default, the Score Rankings Overlay window plots the Cumulative Lift chart. Recall that lift is the ratio of the percent of targets (that is, bad loans) in each decile to the percent of targets in the

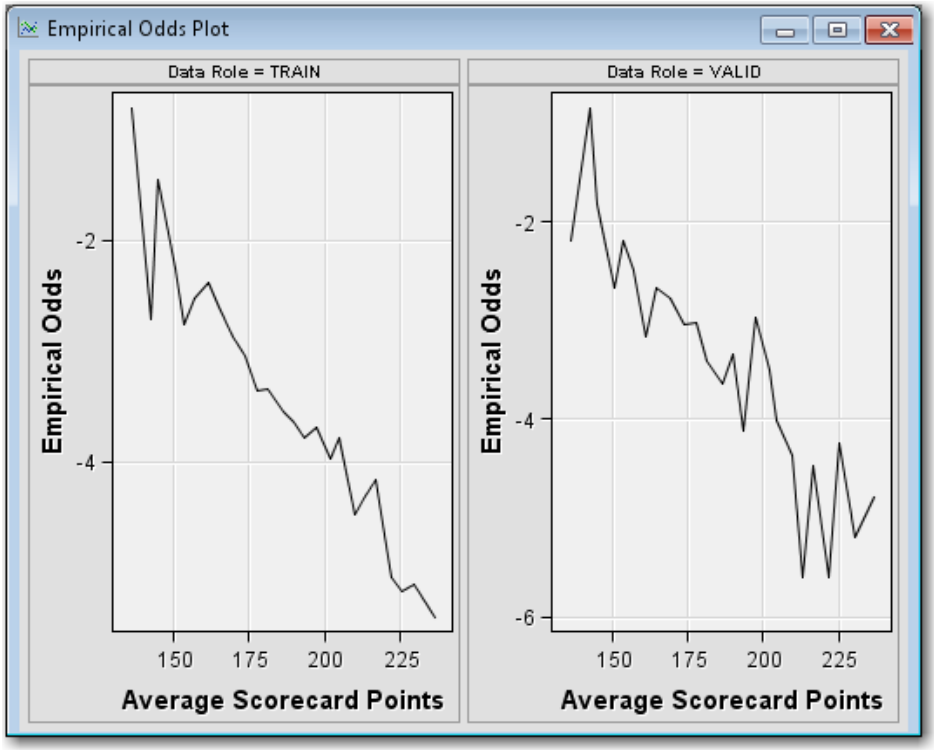
entire data set. Cumulative lift is the cumulative ratio of the percent of targets up to the decile of interest to the percent of targets in the entire data set.

For lift and cumulative lift, the higher value in the lower deciles indicates a predictive scorecard model. Notice that both Lift and Cumulative Lift for this scorecard have high lift values in the lower deciles.



When you finish analyzing the Score Rankings Overlay, minimize the Score Rankings Overlay window.

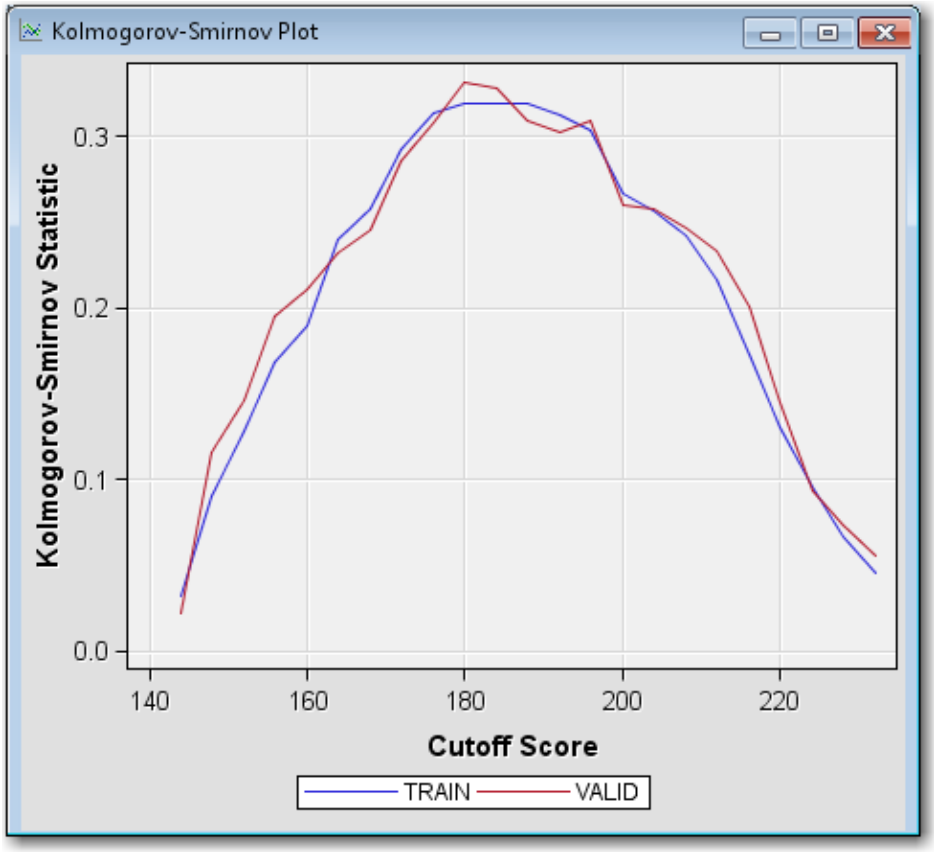
Maximize the Empirical Odds Plot window. An empirical odds plot is used to evaluate the calibration of the scorecard. The chart plots the observed odds in a score bucket against the average score value in each bucket. The plot can help determine where the scorecard is or is not sufficiently accurate. The odds are calculated as the logarithm of the number of bad loans divided by the number of good loans for each scorecard bucket range. Thus, a steep negative slope implies that the good applicants tend to get higher scores than the bad applicants. As would be expected with the previous plot, as the scorecard points increase, so does the number of good loans in each score bucket.



When you finish analyzing the Empirical Odds Plot, minimize the Empirical Odds Plot window.

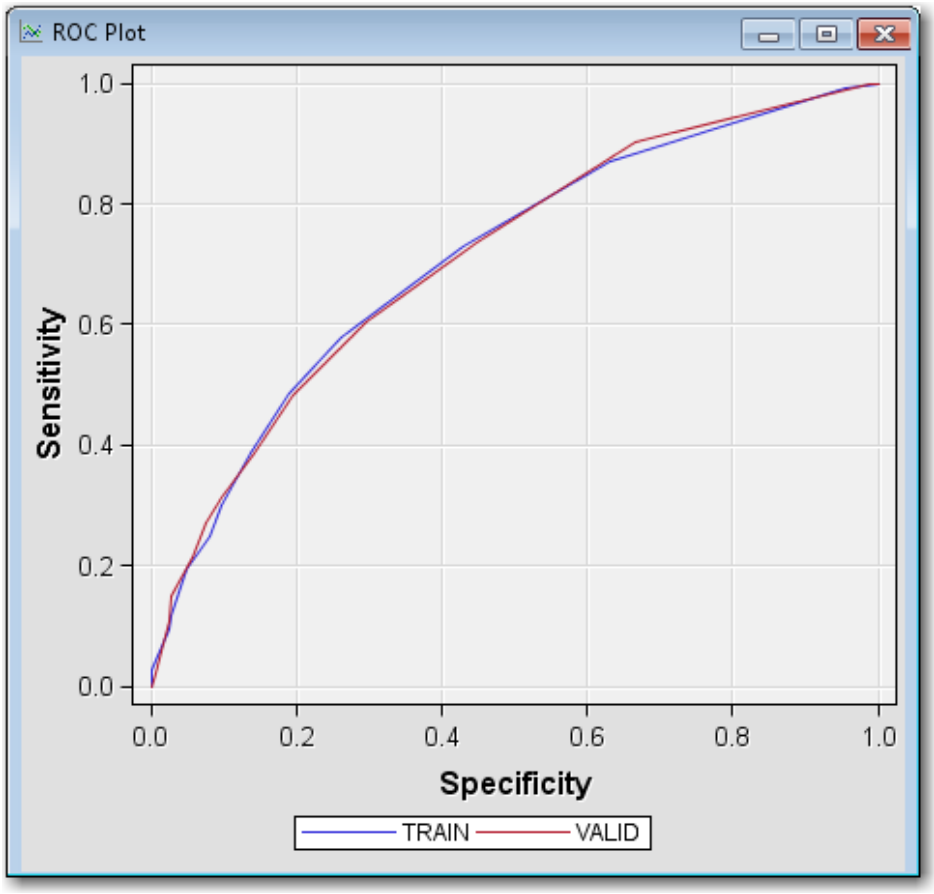
From the main menu, select **View** ⇨ **Strength Statistics** ⇨ **Kolmogorov-Smirnov Plot**. The Kolmogorov-Smirnov Plot shows the Kolmogorov-Smirnov statistics plotted against scorecard cutoff values. Recall that the Kolmogorov-Smirnov statistic is the maximum distance between the empirical distribution functions for the good applicants and the bad applicants. The difference is plotted, for all cutoffs, in the Kolmogorov-Smirnov Plot.

The weakness of reporting only the maximum difference between the curves is that it provides only a measure of vertical separation at one cutoff value, but not overall cutoff values. According to the plot above, the best cutoff is approximately 180 (where the Kolmogorov-Smirnov score is at a maximum). At a cutoff value of 180, the scorecard best distinguishes between good and bad loans.



When you finish analyzing the Kolmogorov-Smirnov Plot, close the Kolmogorov-Smirnov Plot window.

From the main menu, select **View** ⇨ **Strength Statistics** ⇨ **ROC Plot**. The ROC plot is a graphical measure of sensitivity versus 1–specificity. The AUR (which is close to 0.71 for the validation data from the previous Fit Statistics table) measures the area below each of the curves that you see drawn in the plot. The AUR is generally regarded as providing a much better measure of the scorecard strength than the Kolmogorov-Smirnov statistic because the area being calculated encompasses all cutoff values. A scorecard that is no better than random selection has an AUR value equal to 0.50. The maximum value of the AUR is 1.0.

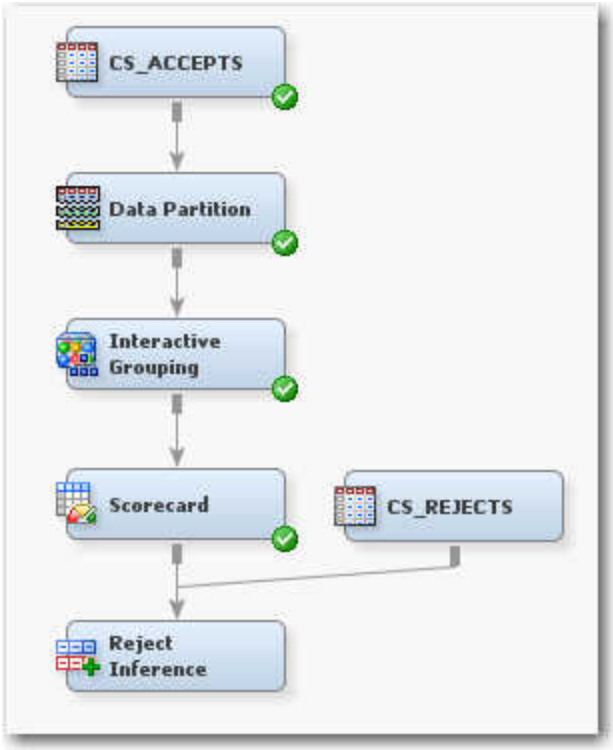


When you finish analyzing the ROC Plot, close the Results window.

Perform Reject Inference on the Model

The preliminary scorecard that was built in the previous section used known good and bad loans from only the accepted applicants. The scorecard modeler needs to apply the scorecard to all applicants, both accepted and rejected. The scorecard needs to generalize the "through the door" population. In this section, you perform reject inference to solve the sample bias problem so that the developmental sample will be similar to the population to which the scorecard will be applied.

1. From the **Credit Scoring** tab, drag a **Reject Inference** node to the Diagram Workspace. Connect the **Scorecard** node to the **Reject Inference** node.
2. From the Project Panel, drag the **CS_REJECTS** data source to the Diagram Workspace. Connect the **CS_REJECTS** data set to the **Reject Inference** node.



The Reject Inference node attempts to infer the behavior (good or bad), or performance, of the rejected applicants using three industry-accepted inference methods. You can set the inference method using the **Inference Method** property.

The following inference methods are supported in SAS Enterprise Miner:

- **Fuzzy** — **Fuzzy** classification uses partial classifications of "good" and "bad" to classify the rejected applicants in the augmented data set. Instead of classifying observations as "good" and "bad," fuzzy classification allocates weight to observations in the augmented data set. The weight reflects the observation's tendency to be "good" or "bad."

  The partial classification information is based on the probability of being good or bad. This probability is based on the model built with the CS_ACCEPTS data set that is applied to the CS_REJECTS data set. Fuzzy classification multiplies these probabilities by the user-specified Reject Rate parameter to form frequency variables. This results in two observations for each observation in the Rejects data. Let p(good) be the probability that an observation represents a good applicant and p(bad) be the probability that an observation represents a bad applicant. The first observation has a frequency variable that is defined as (Reject Rate)*p(good) and a target variable of 0. The second observation has a frequency variable defined as (Reject Rate)*p(bad)and a target value of 1.

  Fuzzy is the default inference method.

- **Hard Cutoff** — **Hard Cutoff** classification classifies observations as either good or bad based on a cutoff score. If you choose **Hard Cutoff** as your inference method, you must specify a **Cutoff Score** in the **Hard Cutoff** properties. Any score below the hard cutoff value is allocated a status of bad. You must also specify the **Rejection Rate** in **General** properties. The **Rejection Rate** is applied to the CS_REJECTS data set as a frequency variable.

- **Parceling** — **Parceling** distributes binned, scored rejected applicants into either a good bin or a bad bin. Distribution is based on the expected bad rates that are calculated from the scores from the logistic regression model. The parameters

that must be defined for parceling vary according to the **Score Range** method that you select in the **Parceling** properties group. All parceling classifications require that you specify the **Rejection Rate**, **Score Range Method**, **Min Score**, **Max Score**, and **Score Buckets** properties.

You must specify a value for the **Rejection Rate** property when you use either the **Hard Cutoff** or **Parceling** inference method. The **Rejection Rate** is used as a frequency variable. The rate of bad applicants is defined as the number of bad applicants divided by the total number of applicants. The value for the **Rejection Rate** property must be a real number between 0.0001 and 1. The default value is 0.3.

The **Cutoff Score** property is used when you specify **Hard Cutoff** as the inference method. The **Cutoff Score** is the threshold score that is used to classify good and bad observations in the **Hard Cutoff** method. Scores below the threshold value are assigned a status of bad. All other observations are classified as good.

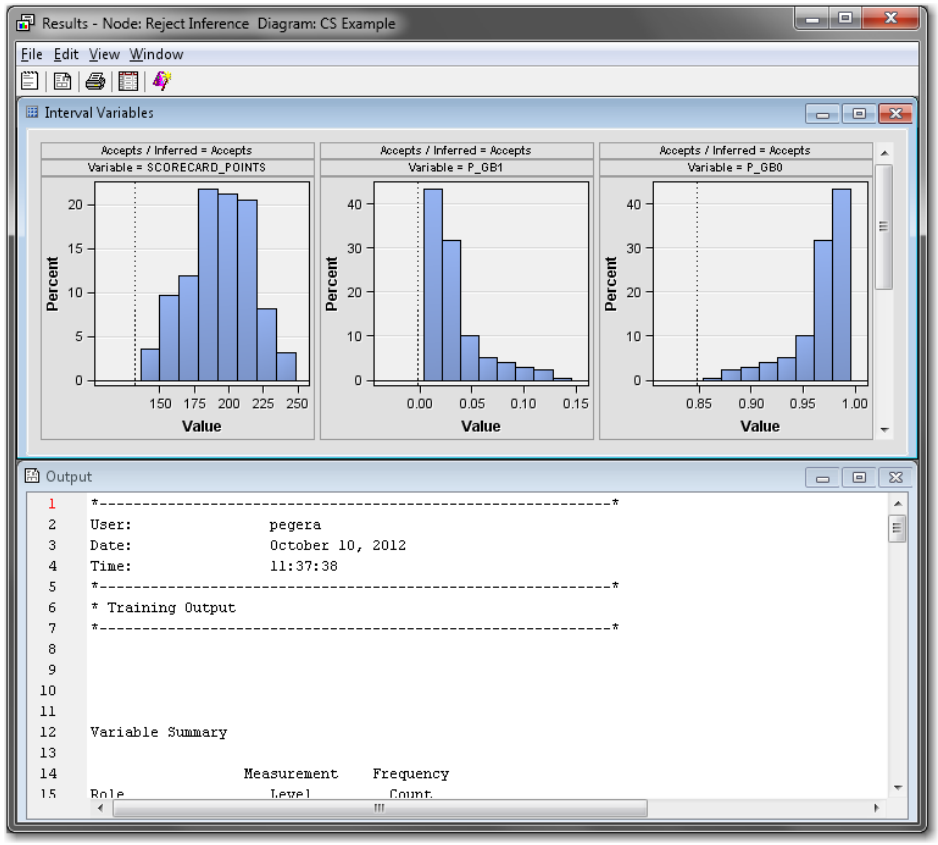
The **Parceling** properties group is available when you specify **Parceling** as the inference method.

The following properties are available in the **Parceling** properties group:

- **Score Range Method** — Use the **Score Range Method** property to specify how you want to define the range of scores to be bucketed. The available methods are as follows:
  - **Accepts** — The **Accepts** score range method distributes the rejected applicants into equal-sized buckets based on the score range of the CS_ACCEPTS data set.
  - **Rejects** — The **Rejects** score range method distributes the rejected applicants into equal-sized buckets based on the score range of the CS_REJECTS data set.
  - **Scorecard** — The **Scorecard** score range method distributes the rejected applicants into equal-sized buckets based on the score range that is output by the augmented data set.
  - **Manual** — The **Manual** score range method distributes the rejected applicants into equal-sized buckets based on the range that you input.
- **Score Buckets** — Use the **Score Buckets** property to specify the number of buckets that you want to use to parcel the data set into during attribute classification. Permissible **Score Buckets** property values are integers between 1 and 100. The default setting for the **Score Buckets** property is 25.

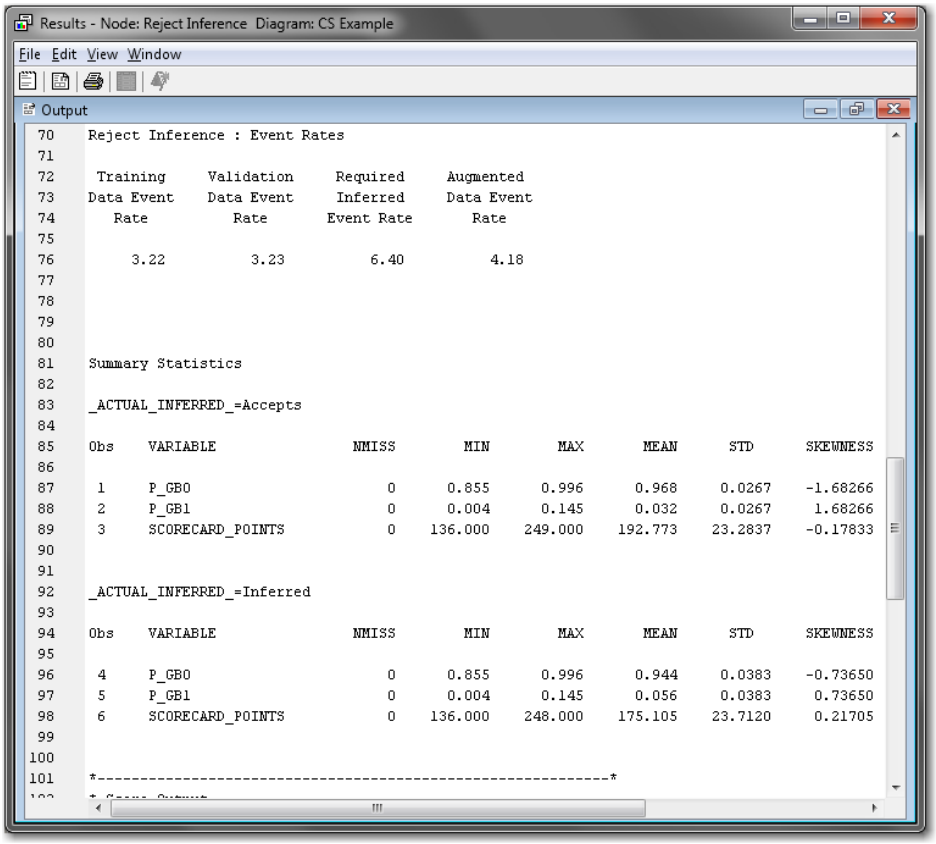
When you use the **Parceling** inference method, you must also specify the **Event Rate Increase** property. The proportion of bad and good observations in the CS_REJECTS data set is not expected to approximate the proportion of bad and good observations in the CS_ACCEPTS data set. Logically, the bad rate of the CS_REJECTS data set should be higher than that of the CS_ACCEPTS data set. It is appropriate to use some coefficient to classify a higher proportion of rejected applicants as bad. When you use **Parceling**, the observed event rate of the accepts data is multiplied by the value of the **Event Rate Increase** property to determine the event rate for the rejects data. To configure a 20% increase, set the Event Rate Increase property to 1.2.

3. For this example, you will use the default values for the **Reject Inference** node properties. Right-click the **Reject Inference** node and select **Run**. In the Confirmation window, click **Yes**.
4. In the Run Status window, click **Results**.



The Results window includes distribution plots for the score and predicted probabilities for the accepted observations and the inferred samples side-by-side.

Expand the Output window. Scroll down to the **Reject Inference: Event Rates** section to the event rates for various samples, including the augmented data.



The output from the **Reject Inference** node is the augmented data, with both CS_ACCEPTS and CS_REJECTS appended together. The **Training Data Event Rate** and the **Validation Data Event Rate** are the event rates (that is, the bad rates) for the accepted applicant's data set. The **Required Inferred Event Rate** is the event rate for the rejected applicant's data set. The **Augmented Data Event Rate** is the event rate for the augmented data set. The **Summary Statistics** sections display basic summary statistics for both the accepted and rejected applicants' data.

When you finish analyzing the **Reject Inference** node output, close the Results window.

Create the Final Scorecard

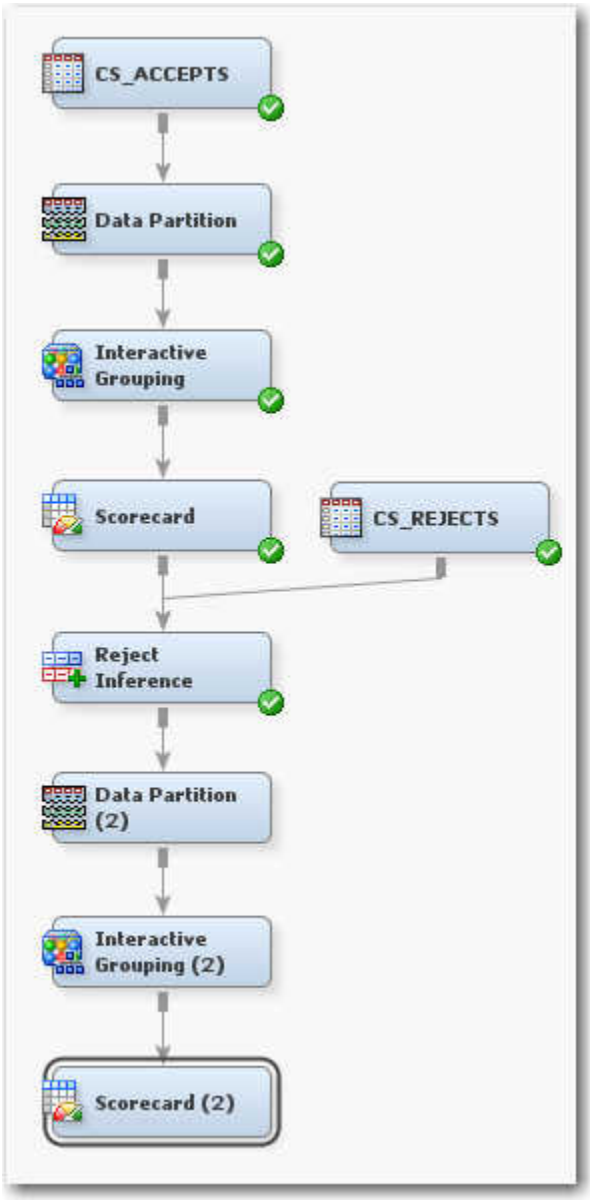
To build the final scorecard, you must repeat the data partitioning, grouping, and scorecard creation steps. These steps are now done on the augmented data set.

1. From the **Sample** tab, drag a **Data Partition** node to the Diagram Workspace. Connect the **Reject Inference** node to the **Data Partition (2)** node.

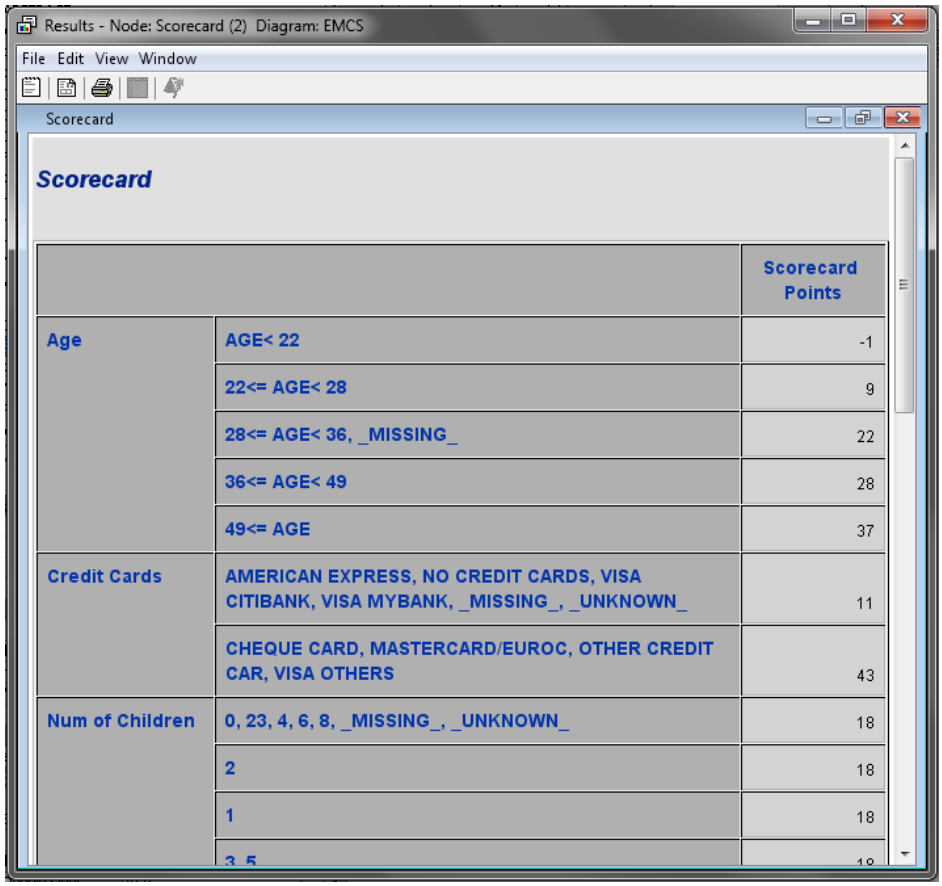
   In the **Data Set Allocations** property group, set the value of **Training** to `70`, **Validation** to `30`, and **Test** to `0`.

2. From the **Credit Scoring** tab, drag an **Interactive Grouping** node to the Diagram Workspace. Connect the **Data Partition (2)** node to the **Interactive Grouping (2)** node. This step uses the default values for the Interactive Grouping node.

3. From the **Credit Scoring** tab, drag a **Scorecard** node to the Diagram Workspace. Connect the **Interactive Grouping (2)** node to the **Scorecard (2)** node.



4. Right-click the **Scorecard (2)** node and click **Run**. In the Confirmation window, click **Yes**. In the Run Status window, click **Results**. Expand the Scorecard window. You now have a scorecard that is based on both accepted and rejected applicants, as shown below.



		Scorecard Points
Age	AGE< 22	-1
	22<= AGE< 28	9
	28<= AGE< 36, _MISSING_	22
	36<= AGE< 49	28
	49<= AGE	37
Credit Cards	AMERICAN EXPRESS, NO CREDIT CARDS, VISA CITIBANK, VISA MYBANK, _MISSING_, _UNKNOWN_	11
	CHEQUE CARD, MASTERCARD/EUROC, OTHER CREDIT CAR, VISA OTHERS	43
Num of Children	0, 23, 4, 6, 8, _MISSING_, _UNKNOWN_	18
	2	18
	1	18
	3, 5	18

Close the Results window.

Appendix 1

References

- Anderson, Billie, and J.M. Hardin. 2009. *Development of Credit Scoring Applications Using SAS Enterprise Miner*. SAS Institute Inc. Cary, NC.
- Siddiqi, Naeem. 2006. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Inc. Hoboken, NJ.
- Thomas, Lyn C., David B. Edelman, and Jonathan N. Crook. 2002. *Credit Scoring and Its Applications*. SIAM. Philadelphia, PA.

Glossary

data source
a data object that represents a SAS data set in the Java-based Enterprise Miner GUI. A data source contains all the metadata for a SAS data set that Enterprise Miner needs in order to use the data set in a data mining process flow diagram. The SAS data set metadata that is required to create an Enterprise Miner data source includes the name and location of the data set, the SAS code that is used to define its library path, and the variable roles, measurement levels, and associated attributes that are used in the data mining process.

Gini index
a measure of the total leaf impurity in a decision tree.

logistic regression
a form of regression analysis in which the target variable (response variable) represents a binary-level or ordinal-level response.

metadata
a description or definition of data or information.

model
a formula or algorithm that computes outputs from inputs. A data mining model includes information about the conditional distribution of the target variables, given the input variables.

node
(1) in the SAS Enterprise Miner user interface, a graphical object that represents a data mining task in a process flow diagram. The statistical tools that perform the data mining tasks are called nodes when they are placed on a data mining process flow diagram. Each node performs a mathematical or graphical operation as a component of an analytical and predictive data model. (2) in a neural network, a linear or nonlinear computing element that accepts one or more inputs, computes a function of the inputs, and can direct the result to one or more other neurons. Nodes are also known as neurons or units. (3) a leaf in a tree diagram. The terms leaf, node, and segment are closely related and sometimes refer to the same part of a tree. See also process flow diagram and internal node.

observation
a row in a SAS data set. All of the data values in an observation are associated with a single entity such as a customer or a state. Each observation contains either one data value or a missing-value indicator for each variable.

overfit
to train a model to the random variation in the sample data. Overfitted models contain too many parameters (weights), and they do not generalize well. See also underfit.

partition
to divide available data into training, validation, and test data sets.

PFD
See process flow diagram.

process flow diagram
a graphical representation of the various data mining tasks that are performed by individual Enterprise Miner nodes during a data mining analysis. A process flow diagram consists of two or more individual nodes that are connected in the order in which the data miner wants the corresponding statistical operations to be performed. Short form: PFD.

project
a user-created GUI entity that contains the related SAS Enterprise Miner components required for the data mining models. A project contains SAS Enterprise Miner data sources, process flow diagrams, and results data sets and model packages.

scorecard
a report that estimates the likelihood that a borrower will display a defined behavior such as payment default.

SAS data set
a file whose contents are in one of the native SAS file formats. There are two types of SAS data sets: SAS data files and SAS data views. SAS data files contain data values in addition to descriptor information that is associated with the data. SAS data views contain only the descriptor information plus other information that is required for retrieving data values from other SAS data sets or from files that are stored in other software vendors' file formats.

target variable
a variable whose values are known in one or more data sets that are available (in training data, for example) but whose values are unknown in one or more future data sets (in a score data set, for example). Data mining models use data from known variables to predict the values of target variables.

training
the process of computing good values for the weights in a model.

training data
currently available data that contains input values and target values that are used for model training.

underfit
to train a model to only part of the actual patterns in the sample data. Underfit models contain too few parameters (weights), and they do not generalize well. See also overfit.

validation data
data that is used to validate the suitability of a data model that was developed using training data. Both training data sets and validation data sets contain target variable

values. Target variable values in the training data are used to train the model. Target variable values in the validation data set are used to compare the training model's predictions to the known target values, assessing the model's fit before using the model to score new data.

variable
a column in a SAS data set or in a SAS data view. The data values for each variable describe a single characteristic for all observations. Each SAS variable can have the following attributes: name, data type (character or numeric), length, format, informat, and label.

Index

C
classing 9
cutoff score 23

D
Data Partition node 8, 25
data source 6
  create new 6
Data Source Wizard 6, 7

G
Gini Statistic 10, 11, 13, 16

I
inference methods 22
information value 10, 11, 13, 16
  calcuation 10
Interactive Grouping application 11
Interactive Grouping node 9, 16, 25

L
logistic regression
  model 14

M
metadata 6

P
parceling 23
process flow diagram
  create new 7
project
  create new 5

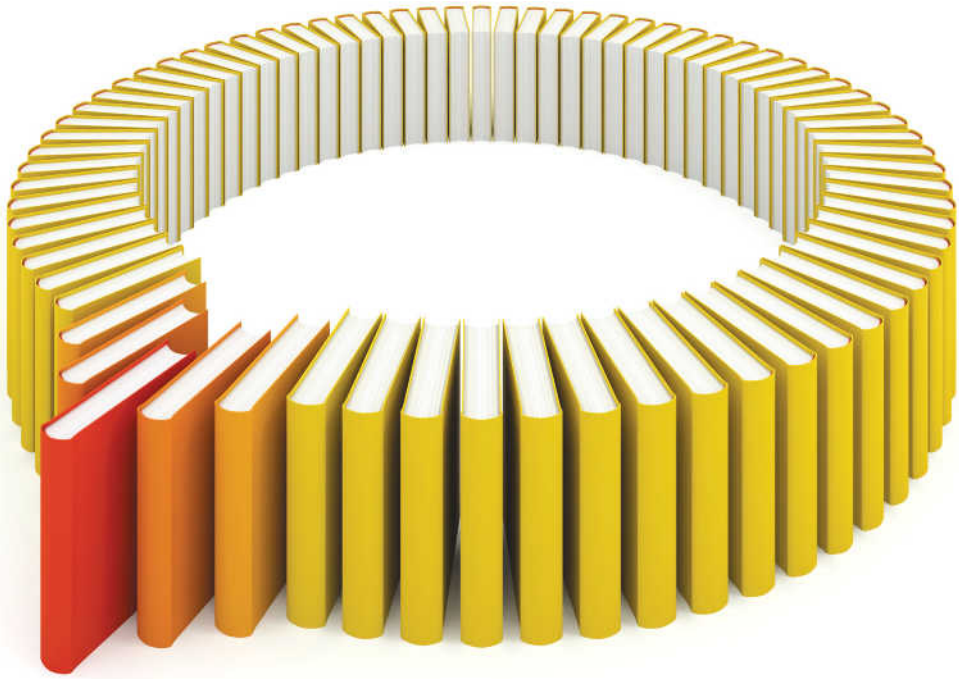
R
Reject Inference node 21, 22, 23, 25
rejection rate 23

S
Scorecard node 14, 15, 16, 25

T
test data 8
training data 8

V
validation data 8

W
weight of evidence 12, 13, 16, 17



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

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. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613