SAS Workshop: SAS Visual Statistics 7.1

Course Notes

SAS Workshop: SAS Visual Statistics 7.1 Course Notes was developed by Bob Lucas, Andy Ravenna, Catherine Truxillo, Chip Wells, and Terry Woodfield. Editing and production support was provided by the Curriculum Development and Support Department.

SAS Workshop: SAS Visual Statistics 7.1 Course Notes

Book code E2745, course code SGF15VSW, prepared date 09Mar2015. SGF15VSW_001

Table of Contents

To learn more…

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web at http://support.sas.com/training/ as well as in the Training Course Catalog.

For a list of other SAS books that relate to the topics covered in this course notes, USA customers can contact the SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the SAS Bookstore on the web at http://support.sas.com/publishing/ for a complete list of books and a convenient order form.

Chapter 1 Workshop Demonstrations

1.1 Workshop Demonstrations

Starting a Project in SAS Visual Statistics and Creating Segments

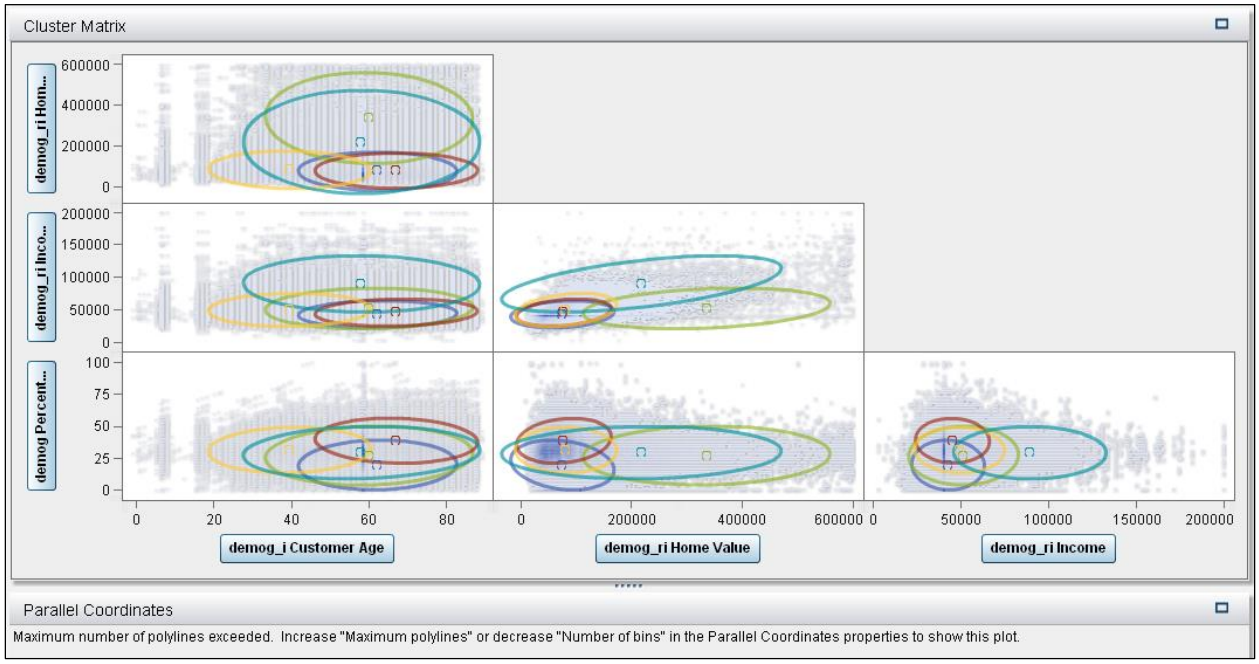
This demonstration illustrates how to start a new analysis in SAS Visual Statistics and create segments based on demographic input variables using the **Model_Bank13** data set.

1. From the computer desktop, open Google Chrome. Select the favorites link for the **SAS Visual Analytics Home Page**.
2. From the SAS Visual Analytics Home Page, select **Create Analytical Model**. Alternatively, select the link for SAS Visual Statistics in the Google Chrome favorites.
3. Click **Select a Data Source**.
4. Click **MODEL_BANK13**. Click **Open**.

In Visual Statistics, the Data pane shows the 3 category variables and the 51 measure variables, including a number of imputed and transformed input variables, which will be used later. Imputed RFM variables have the prefix **i_**, and log-transformed inputs have the prefix **logi_**. Demographic variables have been recoded (**_r**) and imputed (**_i**).

SAS Visual Statistics opens to a linear regression analysis by default.

5. From the toolbar, click the **New Cluster analysis** button  .
6. Select the following demographics variables by pressing Ctrl and clicking them in the Data pane: **demog_i Customer Age**, **demog_ri Home Value**, **demog_ri Income**, and **demog Percentage Retired**.
7. Drag the selected variables into the work area.



The results show the Cluster Matrix window and the Parallel Coordinates window. The Parallel Coordinates window does not show a graph. You will remedy this shortly.

8. Click the **Properties** tab in the right pane and investigate the settings.

The default number of clusters is five. This is arbitrary. For the purposes of modeling the banking data, four clusters are convenient. This is also arbitrary.

9. Change the **Number of clusters** property to **4**. The results are updated to show a solution that has four clusters.

The initialization of the cluster centers is determined jointly by the Seed and Initial Assignment settings. By default, Forgy is used to select the initial centers, with a random number seed of 1234. You can change the seed to use a different randomization.

The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the Update step. This computes the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, whereas Random Partition places all of them close to the center of the data set.

Variable standardization is performed by default.

The Parallel Coordinates graph is useful for profiling your clusters. You can modify the properties to see the graph by increasing the maximum number of polylines, the number of bins, or both.

The number of visible roles selections control how many variables are used for profiling using either the scatter plot matrix or the parallel coordinates plot.

The Cluster Matrix window shows prediction ellipses and cluster centroids on each scatter plot.

10. Change the Number of bins property from **16** to **8**. This enables the Parallel Coordinates graph to display.

The Parallel Coordinates graph shows a mosaic plot of the clusters on the left. It also bins each input into *k* equal-interval bins, where *k* is the value specified in the Number of bins property.

Because these bins are equal interval rather than equal proportion, some bins might have no data.

Each input is labeled with the minimum and maximum data values, and polylines extend from each cluster through the bins on the inputs. The thickness of the polyline is proportional to the number of cases in that polyline.

You can use interactive graph actions to filter clusters, ranges of values, and observations for profiling. Your instructor will walk you through examples of profiling these clusters.

11. Select **Show Summary Table** to see summary statistics for the clusters. Notice that the clusters are numbered 0 through 3.
12. Right-click the cluster matrix and select **Derive a Cluster ID Variable**.

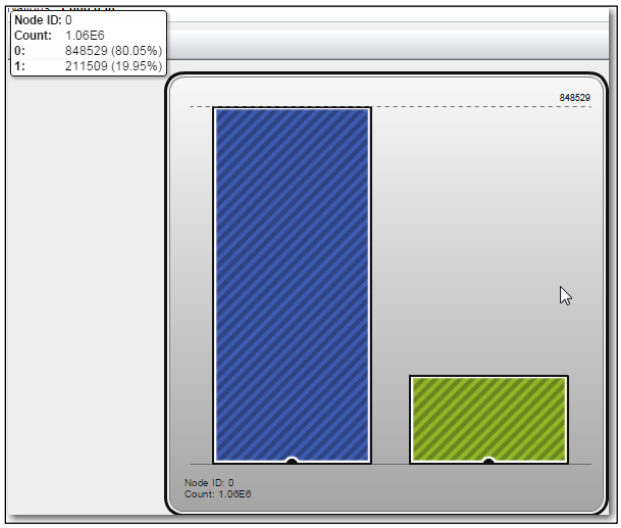
You can also save the score code to an external file, enabling you to assign cluster membership to new data. To do so, select **File** ⇨ **Export** ⇨ **Model Score Code**.

Creating and Cultivating a Decision Tree in SAS Visual Statistics

This demonstration illustrates how to create a decision tree in SAS Visual Statistics and cultivate the tree autonomously.

Creating a Decision Tree Analysis in SAS Visual Statistics

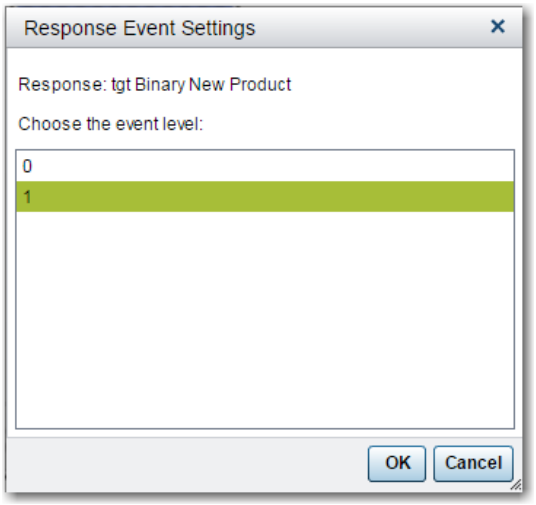
1. Click the **New Decision Tree** button. The dependent variable for analysis is the **tgt Binary New Product** field in the **MODEL_BANK13** data set.
2. Recall that the binary target is a numeric (1/0) flag that indicates response/non-response. Make sure that its measurement level is **Category**, as shown above.
3. Click the **Roles** tab on the right side of the Decision Tree window. Select **tgt Binary New Target** as the response variable.



The distribution of the values in **Binary Target** is shown in the middle window. About 20% of the data consists of ones (responders). This is the root node of the decision tree.

For **Binary Target**, the number of observations is over a million and the event target level (event of interest) to be modeled is 0. The event of interest will be changed to 1 to be consistent with subsequent models.

4. Select the **Advanced** property next to **Response**. Change the event level from **0** to **1**, and click **OK**.



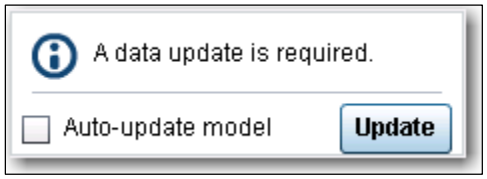
Cultivating a Decision Tree Autonomously

1. Clear the **Auto-update model** box at the bottom of the Decision Tree Properties portion of the window.

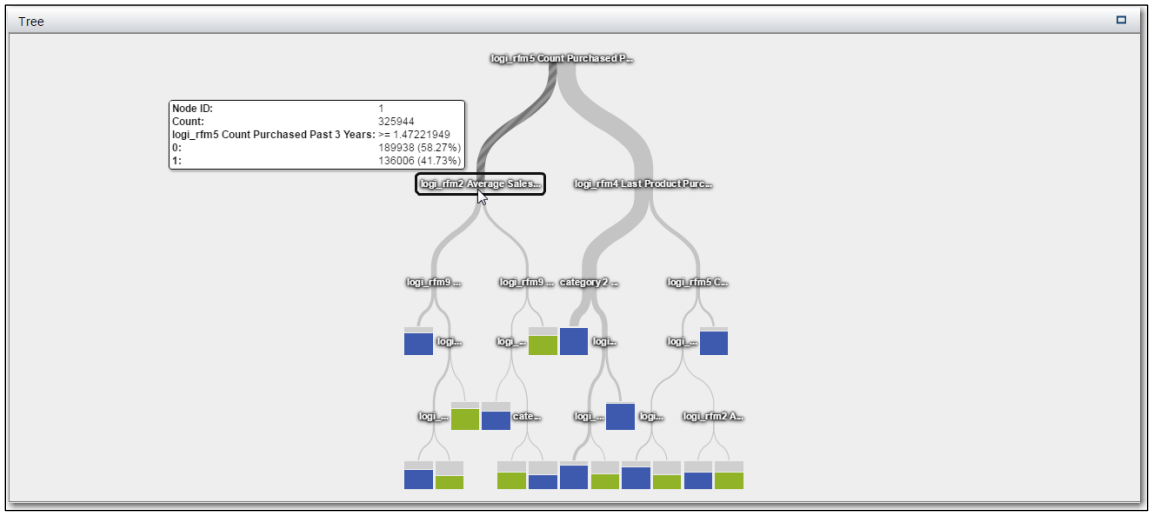
2. Select the input variables.

   Select **category 1 Account Activity Level** in the data portion of the Decision Tree window. Press and hold the Ctrl key, and then select **category2 Customer Value Level** and all of the **logi_RFM** variables. Drag and drop the selected variables to the **Predictors** role, or right-click and select **Assign ⇨Predictors**.

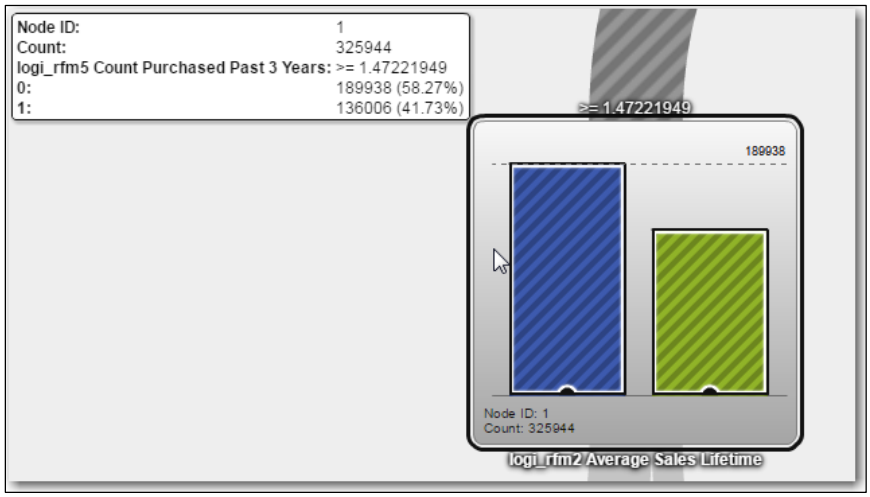
3. Click **Update**.



4. Click the left node after the first split of the top (root) node of the decision tree.

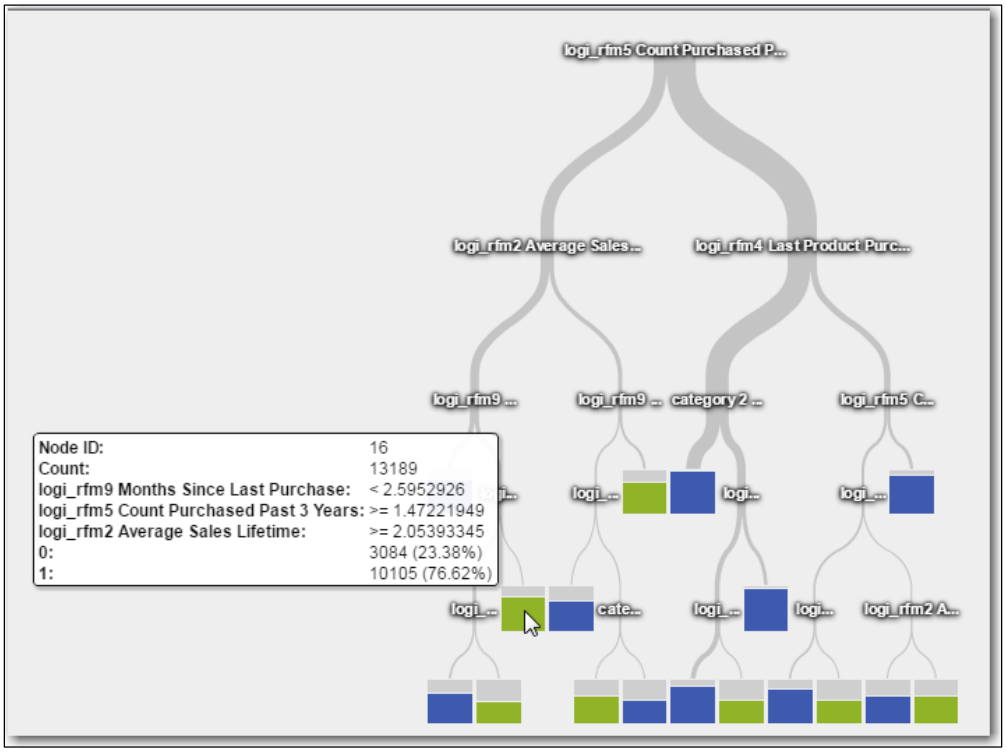


5. Use the wheel on your mouse to zoom in and view the characteristics of observations in this partition of the data.



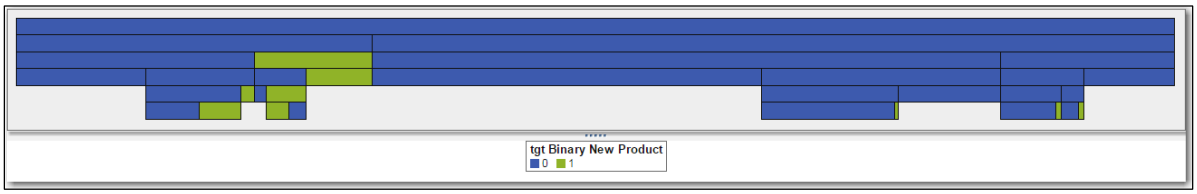
The split point is the number on top of the node. Cases in this partition of the data have a value of **logi RFM 5 Count Purchased Past 3 Years** that is greater or equal to 1.472. There are 325,944 cases in this node, and 136,006 of them are responders. The proportion of responders is about 42%. Recall that the proportion of responders in the training data is about 20%.

6. Zoom out and select the terminal leaf shown below.



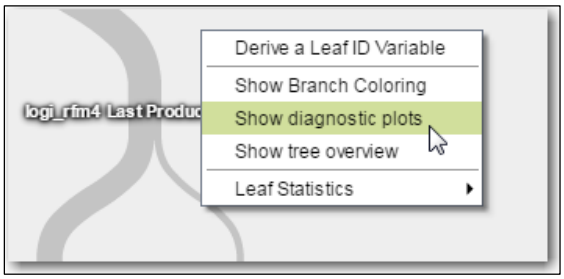
The listed predictors and corresponding split points summarize the characteristics of the 13,189 cases in the node. The node contains about 77% responders. That is, according to this tree, cases with predictors in the ranges listed by the input split points have a probability of response equal to 77%.

7. Scroll down to the icicle plot.

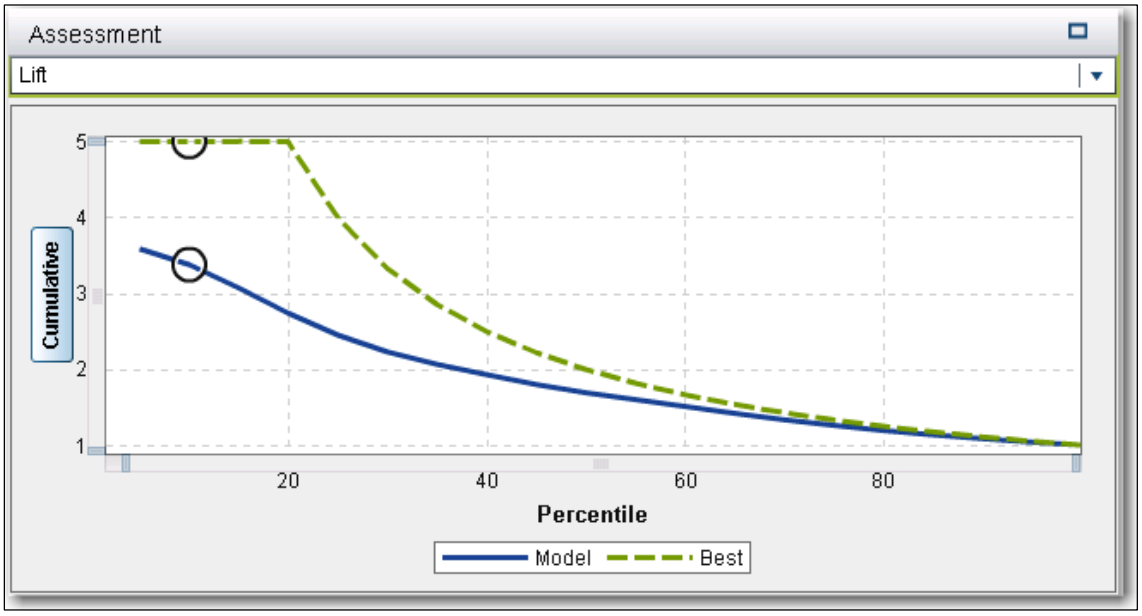


This plot summarizes the leaf structure of the decision tree. The area of each rectangle corresponds to the count of cases in each node. The color of each of the rectangle indicates the majority level of the categorical target in each leaf.

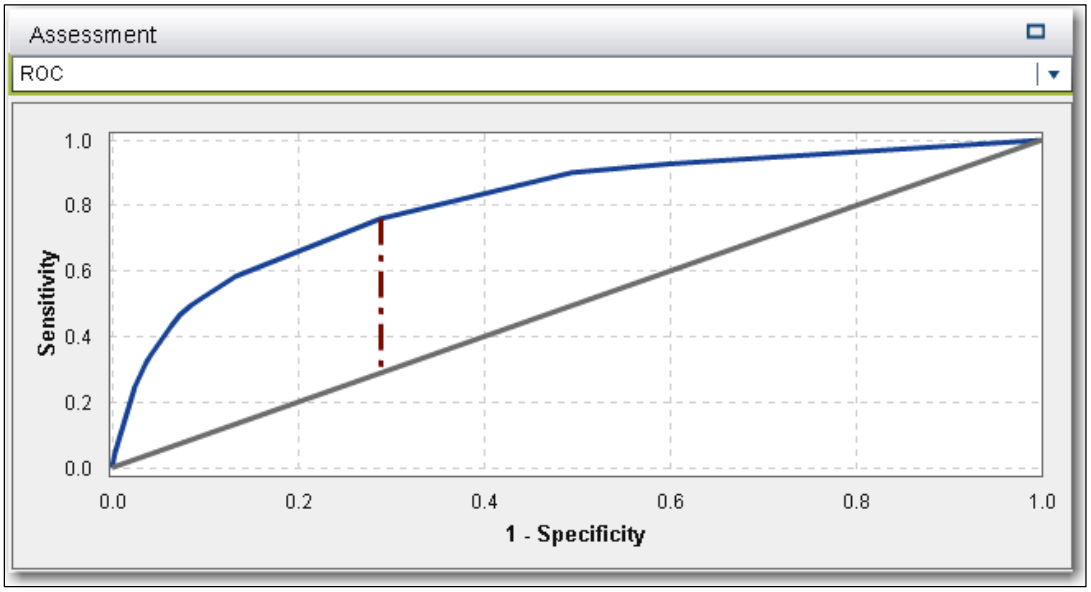
8. Right-click in the (non-Tree) field of the Tree diagram and select **Show Diagnostic Plots**.



The lift plot summarizes the tree model's ability to rank order cases (blue line) relative to a naïve model (a horizontal line with an intercept of 1) and a best model (green line). For example, when rank ordered by the model, the top 10% (percentile) of the data has about 3.3 times as many responders as a random ordering of the data, and 1.7 times fewer responders than a perfect ordering of the training data.



9. Change the assessment plot to **ROC**. The ROC chart summarizes the true positive rate (Sensitivity) and false negative rate (1 – Specificity) across thresholds or cutoffs in the data. The 45 degree line represents the performance of the naïve model, and the vertical red line corresponds to an optimal threshold in the data.



Creating a Logistic Regression in SAS Visual Statistics

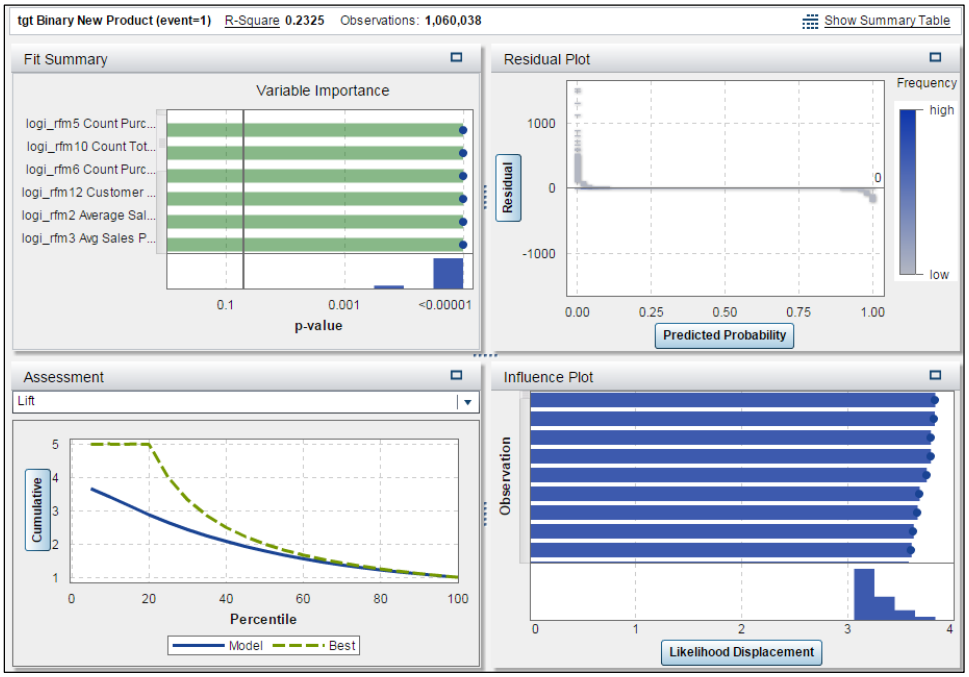
This demonstration illustrates how to build a logistic regression model in SAS Visual Statistics. The demonstration uses the **Model_Bank13** data to model whether a customer contracted for at least one product in the previous campaign season. You create a binary logistic regression with both categorical and continuous explanatory variables.

1. From the toolbar, click  to begin modeling a logistic regression.
2. In the Measure column, right-click **tgt Binary New Product** and select **Category**.
3. Click the **Properties** tab if it is not selected. Click the **Use Variable selection** check box and set the significance level to **.05**.
4. Clear the **Auto-update model** check box at the bottom of the Properties tab.

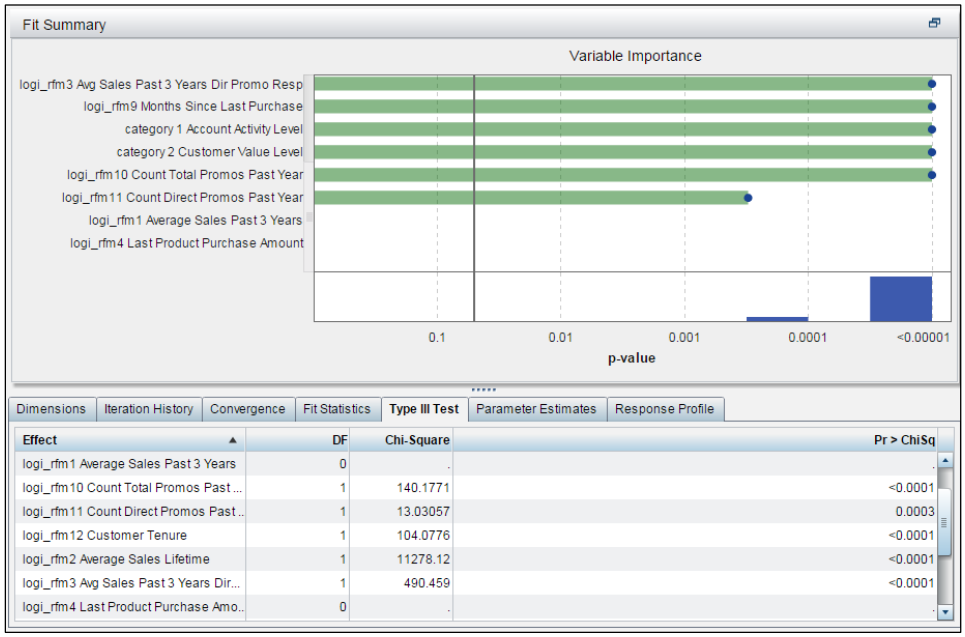
   Turning off auto-update enables you to set up several roles in the model before it is created. Otherwise, the model is updated anytime a change is made.
5. Click the **Roles** tab to begin assigning variables to specific roles in the model.
6. From the Measure column, drag **tgt Binary New Product** into the work area. Select **Advanced** on the Roles tab. In the Response event settings window, select **1** as the event level and click **OK**.

   The first categorical variable dropped into the work area defaults to the Response role unless otherwise specified. Under the Advanced options, you can change the response variable to any other categorical variable.
7. From the Category column, click **category 1 Account Activity Level** and **category 2 Customer Value Level** while holding down the Ctrl key. Assign both of these variables to Classification Effects by dragging them into the work area.
8. From the Measure column, use the Shift key to select all of the **logi_rfm** variables (**logi_rfm1 Average Sales Past 3 Years** through **logi_rfm9 Months Since Last Purchase)**. Assign all 12 of these variables to Continuous Effects by dragging them into the work area.

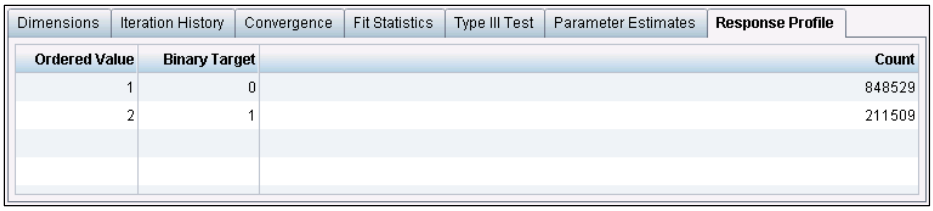
9. Create the logistic regression model by selecting either of the **Auto-Update model** check boxes.



10. Maximize the Fit Summary panel and scroll down until you see the removed variables. Select **Show Summary Table** and click the **Type III Test** tab to verify that both the **logi_rfm1 Average Sales Past 3 Years** and **logi_rfm4 Last Product Purchase Amount** variables were eliminated during the backward selection.



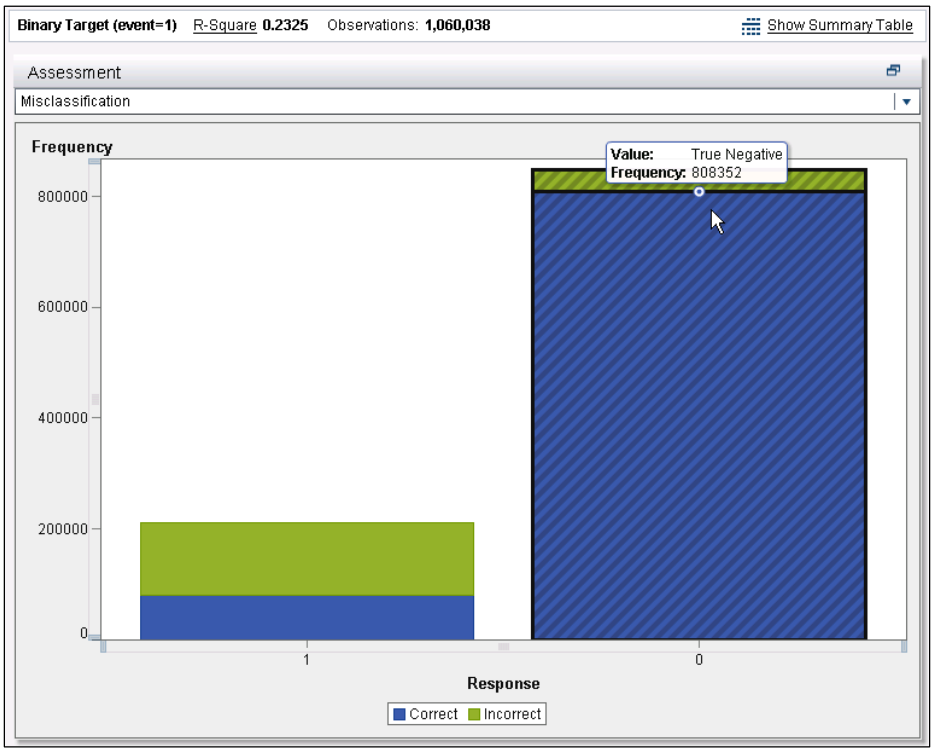
11. Click the **Response Profile** tab to review the original distribution of the target variable.



Ordered Value	Binary Target	Count
1	0	848529
2	1	211509

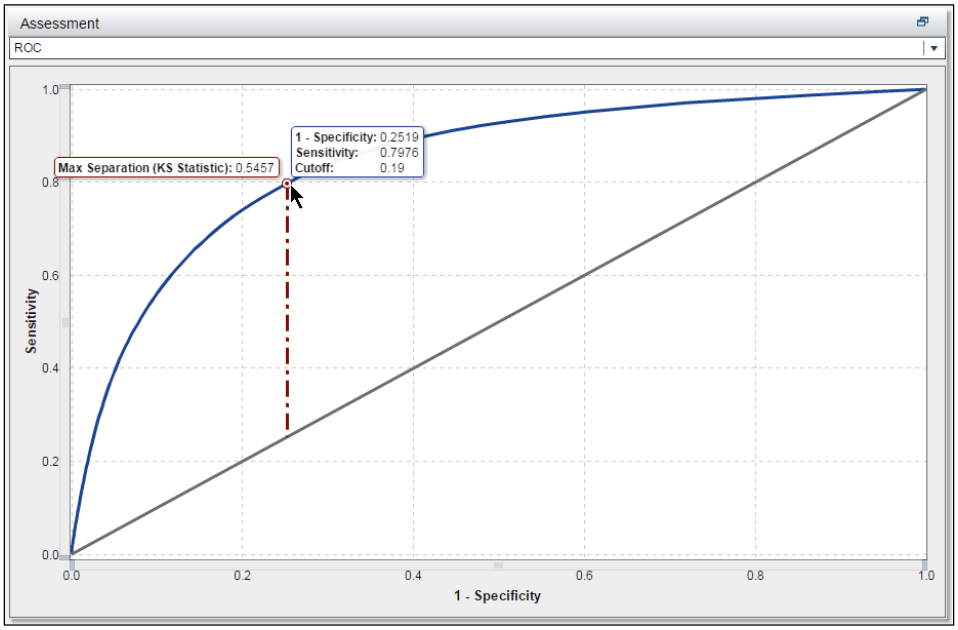
12. Select **Hide Summary Table** to close the logistic regression summary table and then restore the Fit Summary panel to its original size. Now maximize the Assessment panel and select the misclassification plot.

    This model does a good job classifying the non-event, but it does not do as well on the event itself.



    The true positive frequency is 80286 at the default prediction cutoff value of .50.

13. In the Assessment panel, select the ROC chart. Find the tooltip at the very top of the Max Separation line where it intersects the ROC curve. It reveals a cutoff of 19%.



14. Click the **Properties** tab. Change the prediction cutoff to .**19** and press Enter. Check back on the Misclassifications plot to see that the true positive slightly more than doubles to 168691.

15. Ensure that the Auto-update box is checked.

Adding a Group-By Variable to a Logistic Regression

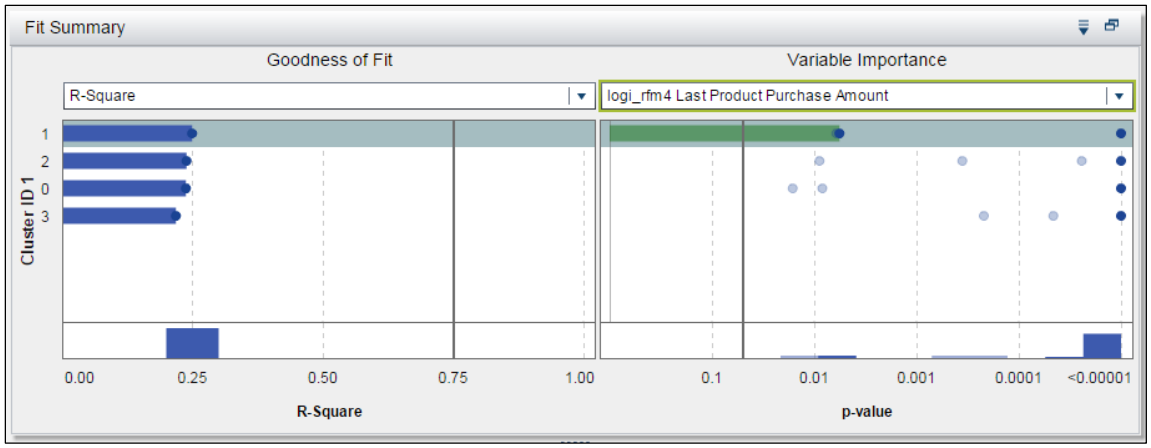
This demonstration illustrates how to interactively group a logistic regression model by a variable that was created in the cluster analysis.

1. Create a copy of the Logistic 1 model by clicking the down-triangle on the tab and selecting **Duplicate**. This creates a tab called **Copy of Logistic 1**.

2. In Logistic 2, click the **Roles** tab if it is not selected and assign the **Cluster ID 1** variable to the Group By role.

3. Maximize the Fit Summary panel and verify that cluster **1** is selected under Goodness of Fit. Select **logi_rfm4 Last Product Purchase Amount** under Variable Importance to validate that it is significant only to that BY group.

   This might suggest that the amount spent on the last product that a customer purchases can be useful in only those models with a particular demographic. Examine the cluster created in the previous chapter to obtain the attributes of this group of customers.



With group-by processing, the Fit Summary panel enables you to quickly visually assess which variables are important to a BY group. Select the BY group first on the left side of the panel. Then select a variable of interest on the right side of the panel. The green bars indicate which BY groups are of significance to the selected variable.

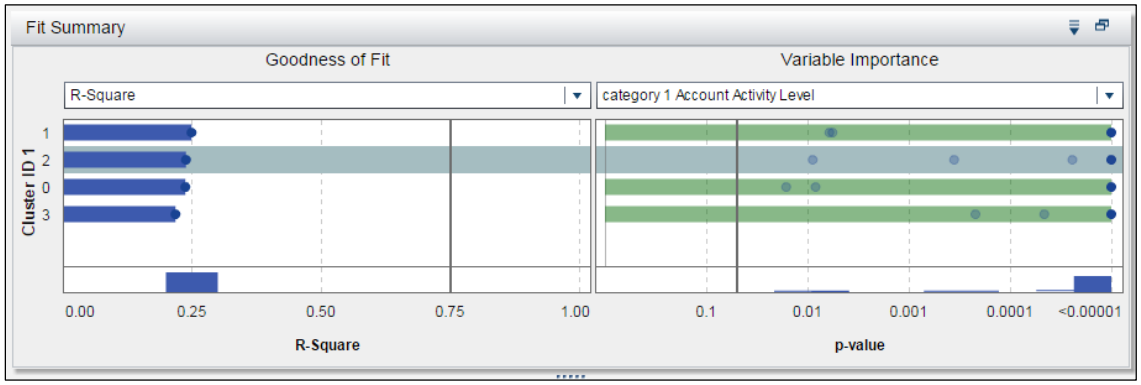
4. Select **Show Summary Table** and click the **Parameter Estimates** tab. Select the **Estimate** column heading to sort the column. Examine the estimates to verify that there were four terms dropped for the logistic regression model built for this BY group: **category 2 Customer Value Level E**, **logi_rfm8 Count Prchsd Lifetime Dir Promo Resp**, **logi_rfm10 Count Total Promos Past Year,** and **category 1 Account Activity Level Z**. The categorical levels are not associated with hypothesis tests because they are reference levels and have fixed parameter estimates of 0.

Dimensions | Iteration History | Convergence | Fit Statistics | Type III Test | **Parameter Estimates** | Response Profile

Parameter	Estimate ▲	Standard Error	z Value	Pr > |z|
logi_rfm9 Months Since Last Purchase	-1.49144	0.027348	-54.5354	<0.0001
logi_rfm2 Average Sales Lifetime	-1.25905	0.03608	-34.8961	<0.0001
logi_rfm11 Count Direct Promos Past Year	-0.45795	0.039362	-11.6344	<0.0001
category 2 Customer Value Level D	-0.24647	0.059252	-4.1597	<0.0001
logi_rfm3 Avg Sales Past 3 Years Dir Promo Resp	-0.17935	0.030555	-5.86978	<0.0001
logi_rfm7 Count Prchsd Past 3 Years Dir Promo Resp	-0.1266	0.020238	-6.25539	<0.0001
logi_rfm12 Customer Tenure	-0.08252	0.030084	-2.74286	0.0061
category 2 Customer Value Level C	-0.00944	0.054017	-0.17481	0.8612
category 2 Customer Value Level E	0	.	.	.
logi_rfm8 Count Prchsd Lifetime Dir Promo Resp	0	.	.	.
logi_rfm10 Count Total Promos Past Year	0	.	.	.
category 1 Account Activity Level Z	0	.	.	.
category 2 Customer Value Level B	0.054681	0.04827	1.132809	0.2573
logi_rfm4 Last Product Purchase Amount	0.06486	0.023434	2.767809	0.0056
logi_rfm6 Count Purchased Lifetime	0.180086	0.025661	7.018001	<0.0001
logi_rfm1 Average Sales Past 3 Years	0.26812	0.033875	7.915058	<0.0001
category 1 Account Activity Level X	0.274972	0.031024	8.863321	<0.0001
category 2 Customer Value Level A	0.454167	0.047312	9.599389	<0.0001
category 1 Account Activity Level Y	0.456628	0.03933	11.61015	<0.0001
logi_rfm5 Count Purchased Past 3 Years	1.877928	0.031077	60.42776	<0.0001
Intercept	4.657452	0.14507	32.10481	<0.0001

5. Select cluster **2** in the Fit Summary panel and notice that a new logistic regression model is created for this BY group. Examine the Parameter Estimates tab in the summary table to verify now that **logi_rfm4** is no longer part of the model, but **logi_rfm10** is.

6. Select **category 1 Account Activity Level** in the Fit Summary panel under Variable Importance. Verify that the Account Activity Level is not important to this cluster, but it is to all others.



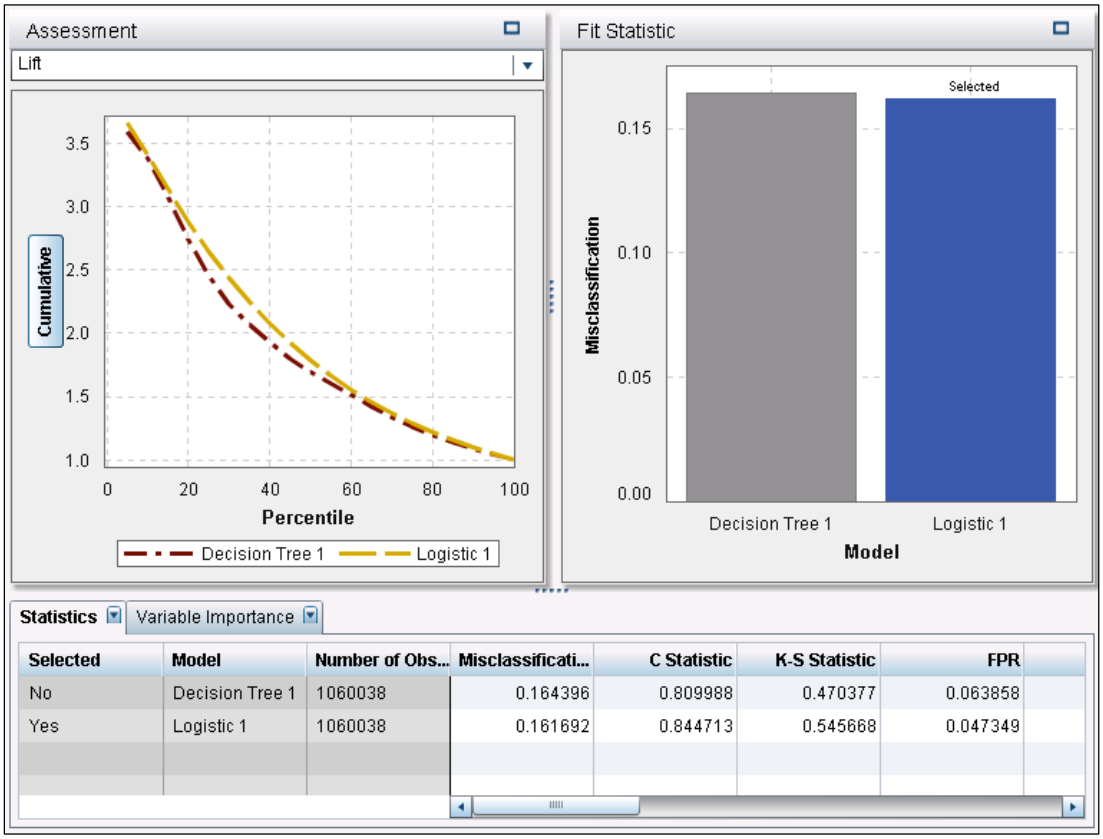
This lack of significance for Account Activity Level can be corroborated by examining the 0 estimates on the Parameter Estimates tab of the summary table.

Comparing Models in SAS Visual Statistics

This demonstration compares a logistic regression model with a decision tree model.

1. Click the **Decision Tree 1** tab to submit the decision tree model.
2. Click  to start up a model comparison. Select **Close** if you get a warning about models not being available for comparison.
3. Select **1** as the level.
4. Select both **Decision Tree 1** and **Logistic 1** in the Available Models pane and add them as selected models. Click **OK**.

   The Logistic 1 model is the selected model. Review the Statistics table to note that the logistic regression model has a higher C statistic. Observe the cumulative lift chart to note that the logistic model has a higher lift.



Selected	Model	Number of Obs...	Misclassificati...	C Statistic	K-S Statistic	FPR
No	Decision Tree 1	1060038	0.164396	0.809988	0.470377	0.063858
Yes	Logistic 1	1060038	0.161692	0.844713	0.545668	0.047349