##### Building a Logistic Regression

This demonstration illustrates how to build a logistic regression model in SAS Visual Statistics. The demonstration uses the **VS\_BANK** 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.

**Adding Continuous and Classification Effects**

1. Start a new report.
2. Click the **Data** pane icon.
3. Click **Select to add data**.
4. Type **VS** in the filter search bar.
5. Select **VS\_Bank**.
6. Select **OK**.
7. Click the **Objects** pane icon.
8. Under SAS Visual Statistics, either double-click or drag and drop **Logistic Regression** onto the canvas.
9. In the Data pane, in the Measure column, select **Edit properties** on **tgt Binary New Product**. Change the classification to **Category**.
10. On the menu bar, click  (**Menu**) and select **Interface options Disable auto-refresh**.
11. Click the **Options** pane if it is not selected. Set the Variable selection method to **Fast Backward** and leave the significance level at **.01**.
12. Click the **Roles** pane to begin assigning variables to specific roles in the model.
13. Add **tgt Binary New Product** as the response variable.

**Note:** You can also drag and drop **tgt Binary New Product** onto the Logistic Regression canvas.

1. Add the 12 variables that begin with **logi\_rfm** as continuous effects.

**Note:** Use the Shift key to select several variables at the same time.

1. Add **category 1 Account Activity Level** and **category 2 Customer Value Level** as classification effects.

12.

Create the logistic regression model by

clicking



(

**Menu**

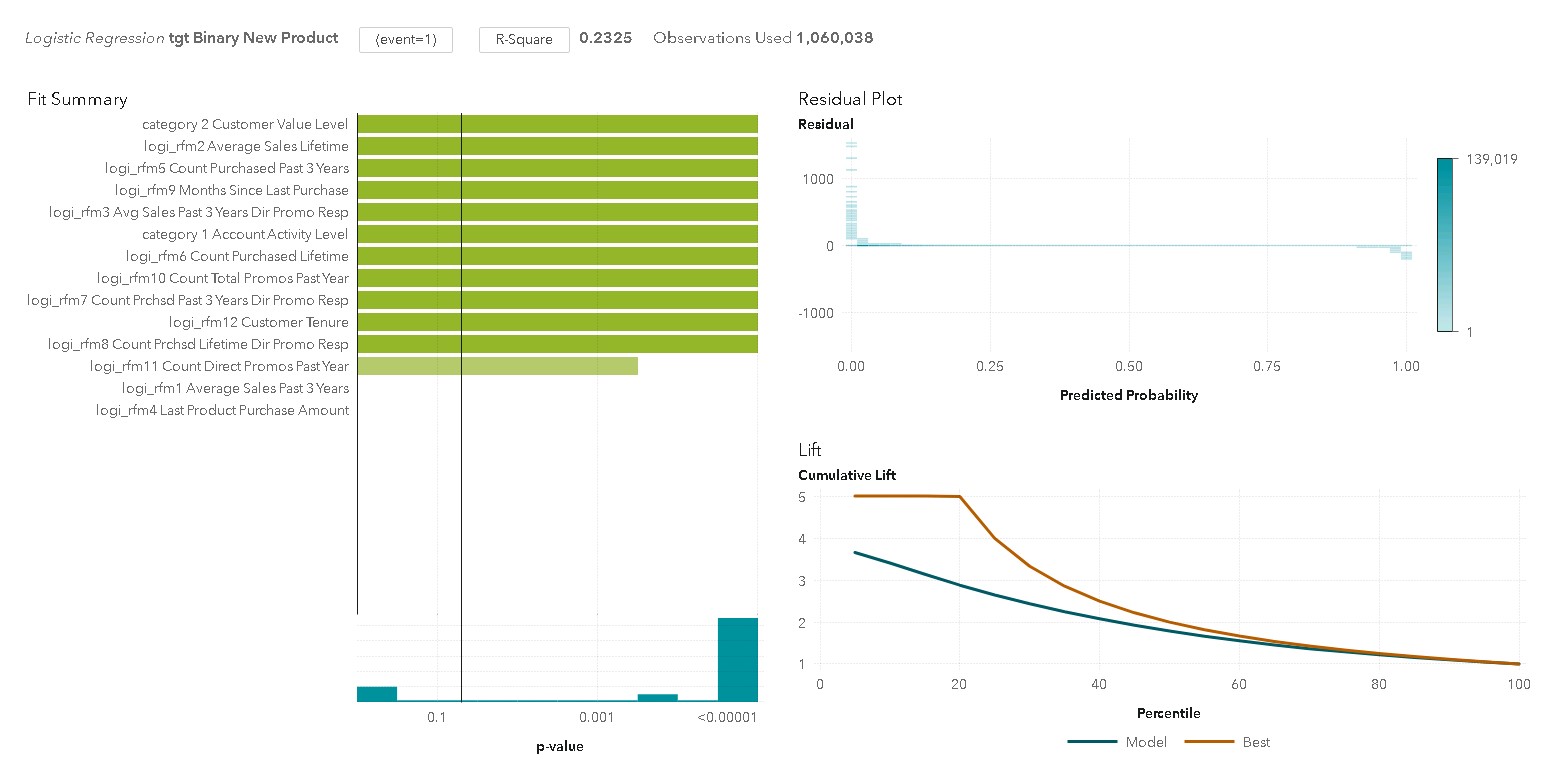
)

select **Interface options**

**Enable auto**

**refresh**

(Also, one time only refresh is available on the canvas.)

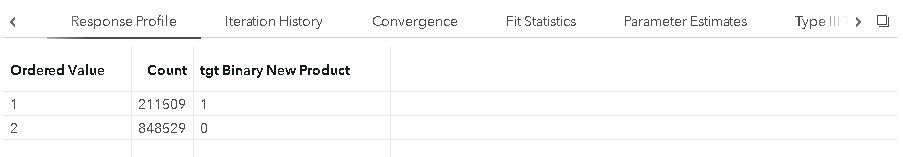


By default, the binary event level of highest sort order is selected. In the **tgt Binary New Product** variable, the event level of 1 is selected. This matches the interest in modeling the behavior of accounts with a purchase.

1. In the Options pane, under Model Display, select **General**. Change the plot layout to **Stack** to expand the Fit Summary on the canvas. Only 12 of the 14 input variables are used to build this model.
2. On the report, click  (**Maximize**) to see the details table. Click the **Selection Summary** 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.

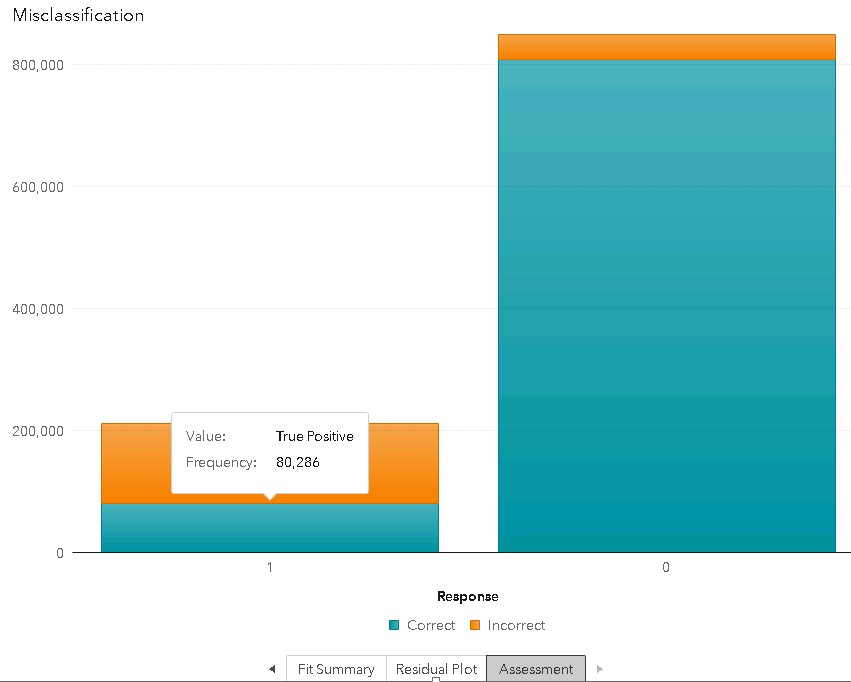


1. Click the **Response Profile** tab to review the original distribution of the target variable.



1. On the report, click **Restore** to close the details table.
2. Click the **Residual Plot** tab. Residuals for logistic regression can be useful in identifying observations that are not explained well by the model.
3. Click the **Assessment** tab. Right-click the chart and select **Misclassification**.

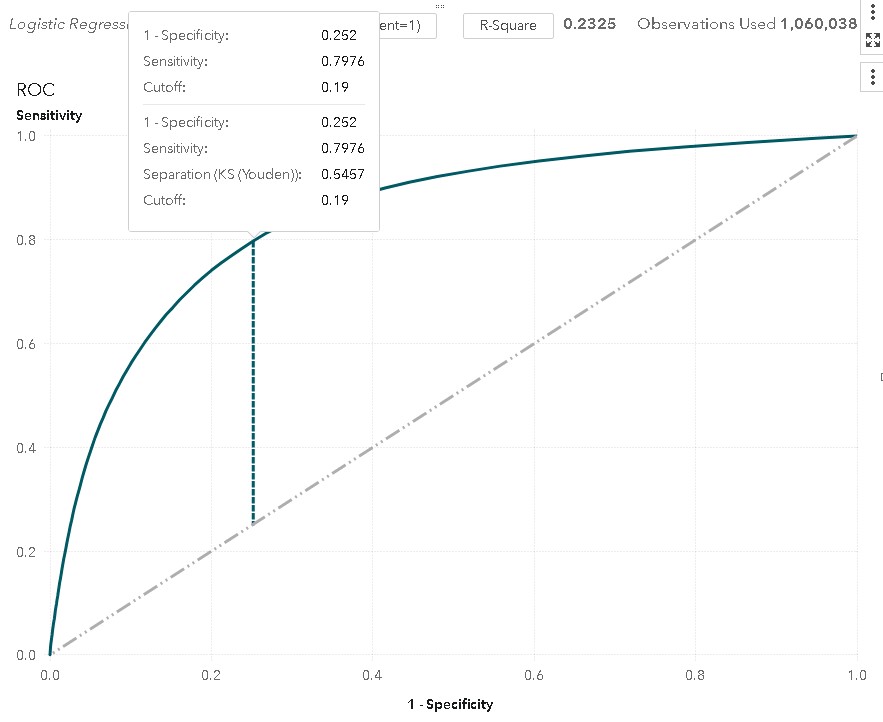
This model does a good job classifying the nonevent, but it does not do as well on the event itself.



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

Logistic Regression

1. Right-click the misclassification chart and select **ROC**. Find the tooltip at the very top of the Max Separation line where it intersects the ROC curve. It reveals an optimal cutoff value of 19%.



**Note:** The threshold indicated above is where the sum of sensitivity and specificity are maximized.

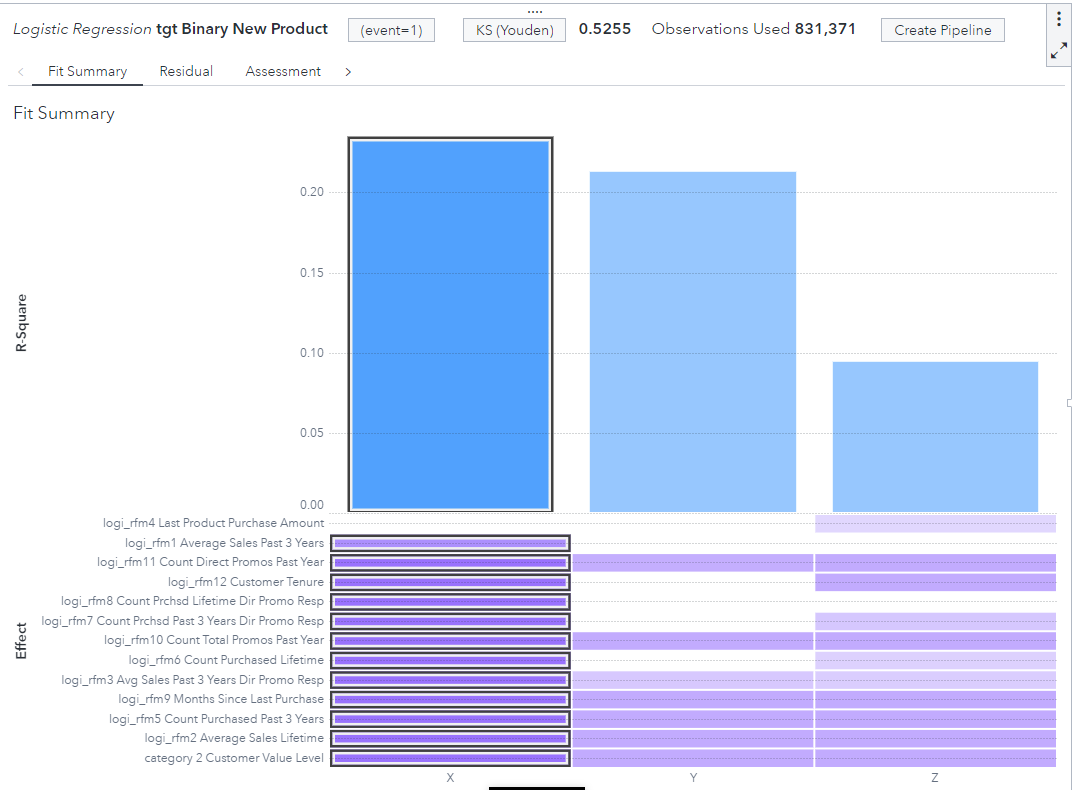
1. Click the **Options** pane. Under Assessment, change the Prediction cutoff slider to **.2** and press Enter. Go back to the misclassifications plot to see that the true positive slightly more than doubles to 165,564.

**Note:** To change the prediction cutoff value, you can either use the slider or click the pointer and enter a value.

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 lesson about clustering.

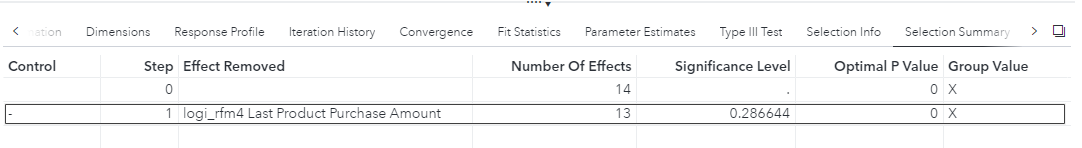
1. Select the logistic regression object.
2. Click the **Roles** pane if it is not selected. (Select the logistic regression canvas if it is not the active object.) Drag the **category 2 Customer Value Level** variable from the Classification effects to the Group by role.
3. If the Fit Summary is not expanded across the full canvas, then click the **Options** pane and change the plot layout to **Stack**.
4. Select the **category 2 Customer Value Level X** bar in the Fit Summary plot. Level **X** has the highest R-square (0.2348) of the three customer levels.



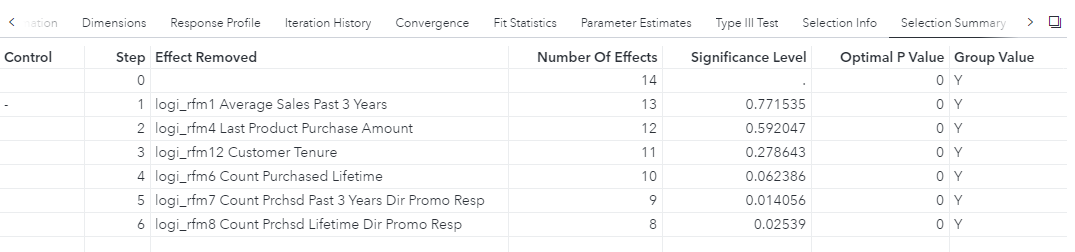
Modeling with Group

**Note:** With group-by processing, the Fit Summary panel enables you to quickly visually assess which variables are important to a BY group. The purple bars indicate which BY groups are of significance to the selected variable.

1. On the report, click  (**Maximize**) to see the details table. Click the **Parameter Estimates** tab and then the **Selection Summary** tab. Examine the summary. Verify that one term was dropped for the logistic regression model that was built for the customer value level X BY group: **logi\_rfm4 Last Product Purchase Amount**



1. Select the **level Y** bar in the goodness-of-fit plot. Notice that a new logistic regression model is created for this BY group. Click the **Selection Summary** tab. Examine the summary. Verify that six terms were dropped for the logistic regression model that was built for this BY group:



1. On the report, click **Restore** to close the details table.

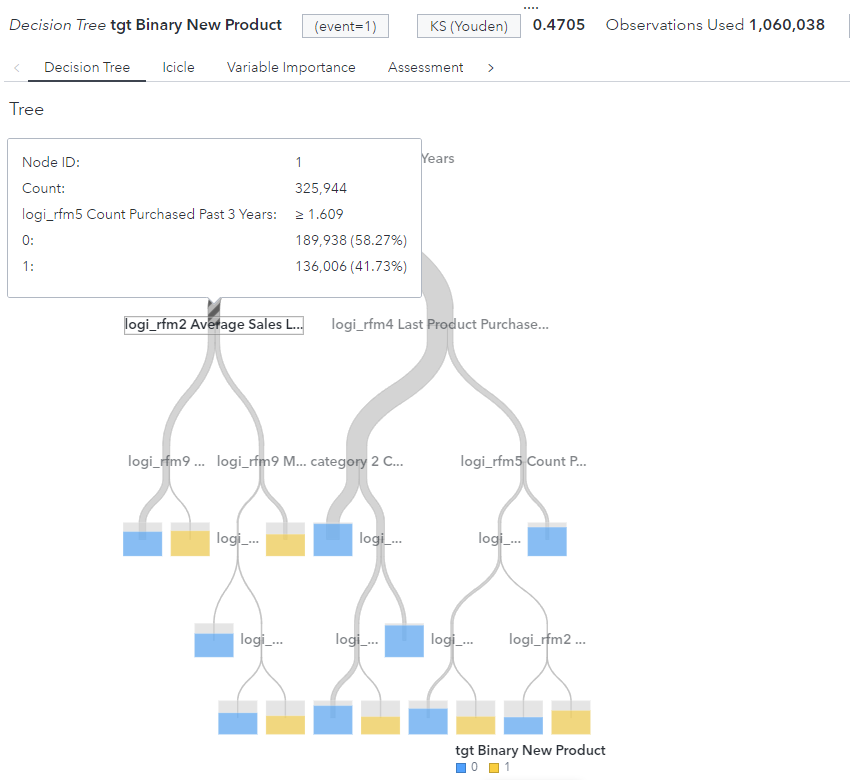
#### Creating a Decision Tree Analysis in SAS Visual Statistics

1. To access the alternative menu, in the upper right corner of the logistic regression, hold down the Alt key. Click  (**More**) and select **Duplicate on new page as**  **Decision Tree**. (You must continue to hold down the Alt key.)
2. The entire logistic regression model is now duplicated as a decision tree. Recall the distribution of the values in **tgt Binary New Target**. Approximately 20% of the data consists of 1s (responders). This is the root node of the decision tree.

**Note:** For **Binary Target**, the number of observations is greater than one million and the event target level (event of interest) to be modeled is 1. The event of interest is consistent with previous models.

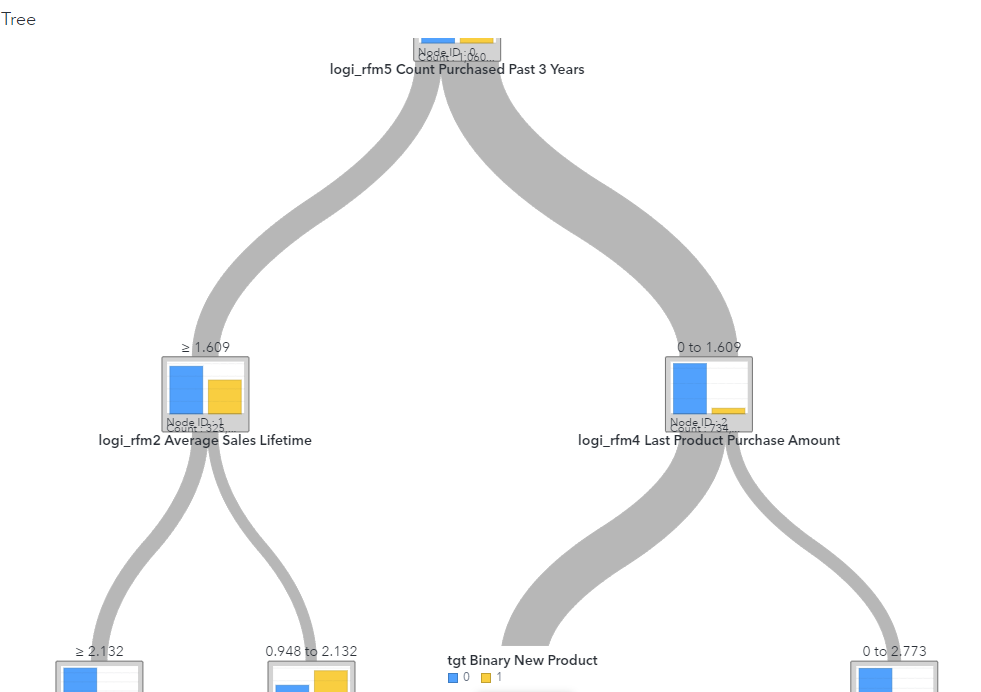


1. Click the left node after the first split of the top (root) node of the decision tree.



1. Scroll to zoom in and view the characteristics of observations in this partition of the data.

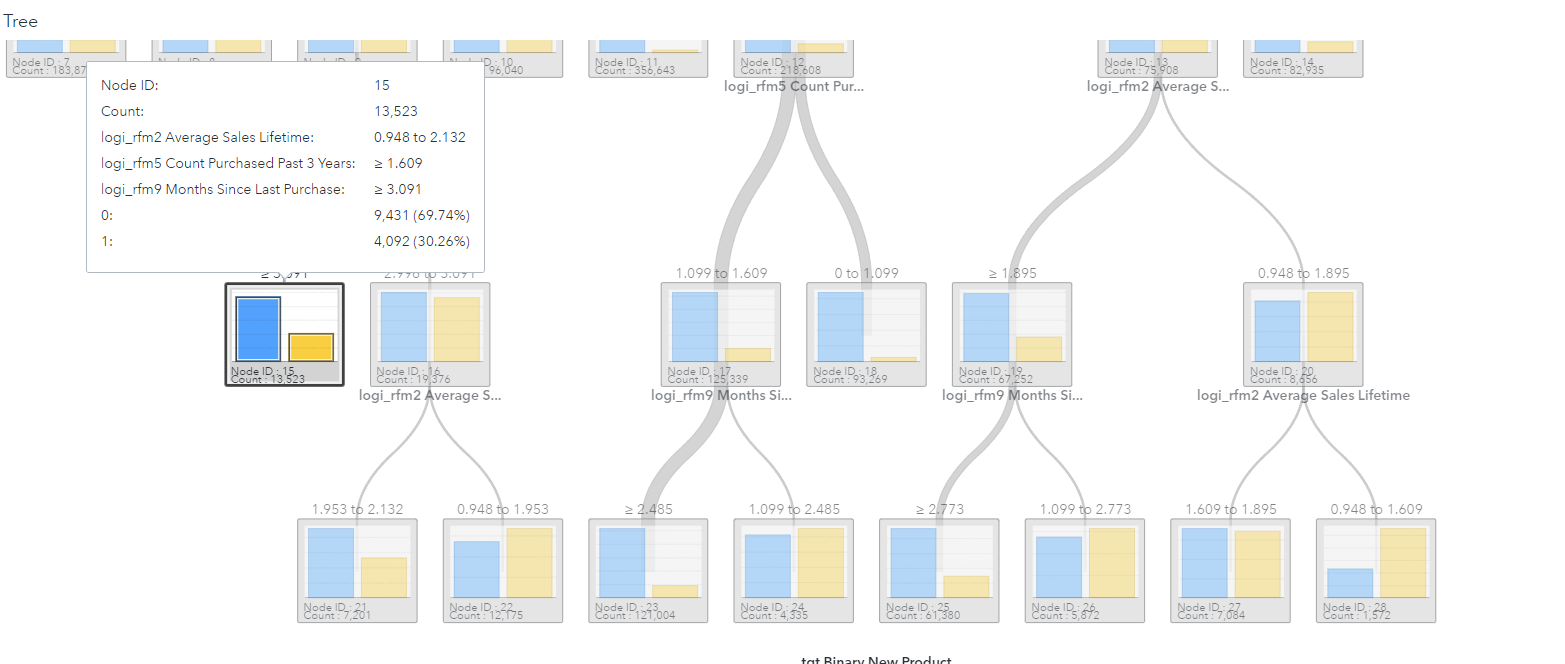
**Note:** The Zoom functionality centers the diagram on your cursor. A good tip is to position the mouse pointer on the part of the tree that you are interested in investigating before zooming.



**Note:** Position your mouse pointer on the gray path to open pop-up information about that path.

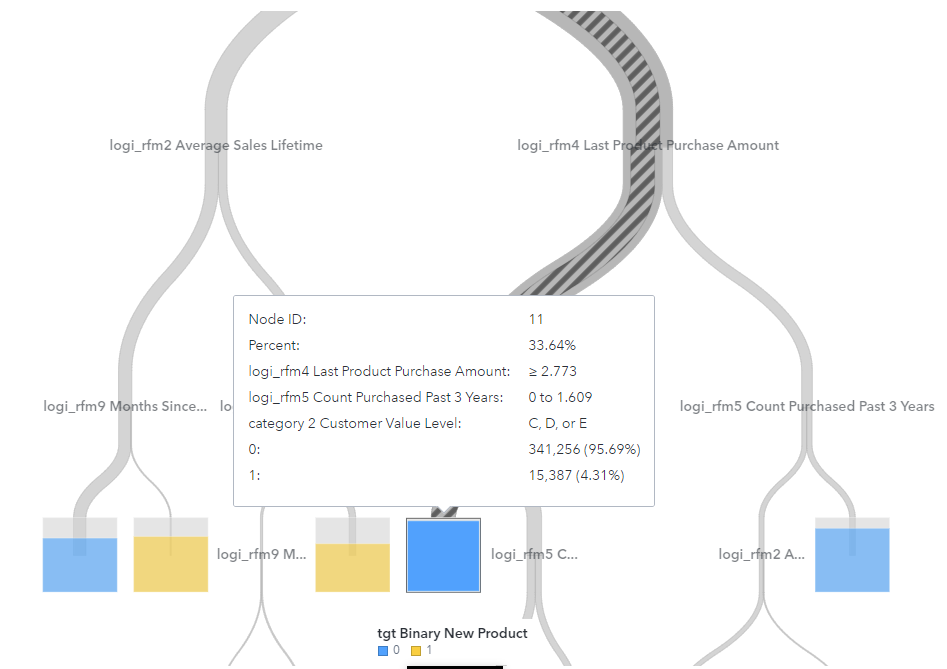
The split point is the number on top of the node. Cases in this partition of the data have a value of **logi\_rfm5 Count Purchased Past 3 Years** that is greater or equal to 1.609. There are 325,944 cases in this node, and 136,006 of them are responders. The proportion of responders is approximately 42%.

1. Navigate to the second to last row of the tree. Then navigate to and select the terminal leaf shown below.



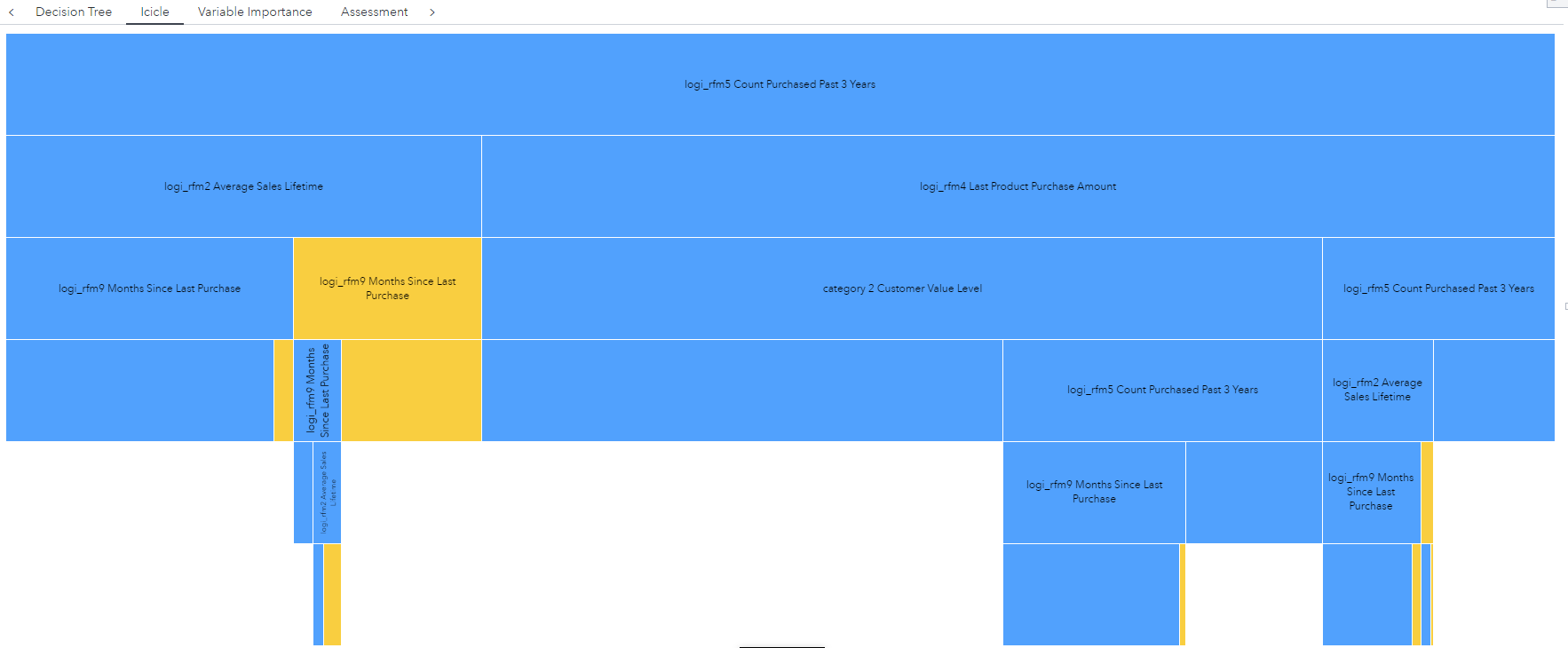
The listed predictors and corresponding split points summarize the characteristics of the 13,523 cases in the node. The node contains about 30% responders. That is, according to this tree, cases with predictors in the ranges listed by the input split points have a probability of response equal to 30%.

1. Right-click anywhere in the canvas and select **Percent**. Select the large leaf in the center of the tree.



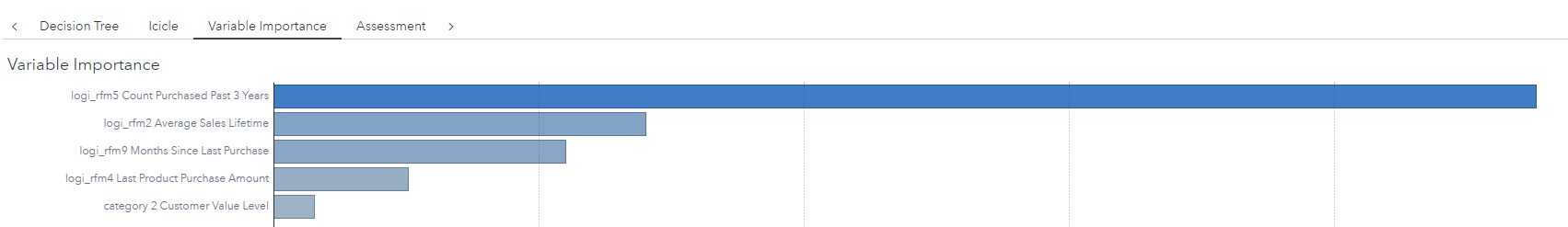
The nodes now contain overall percentage information instead of counts. This terminal node contains 33.64% of the cases in the tree.

1. Click the **Icicle Plot** tab.



This plot summarizes the leaves of the decision tree. The area of each rectangle corresponds to the percent of cases in each leaf. The color of each rectangle indicates the majority level of the binary target in each leaf.

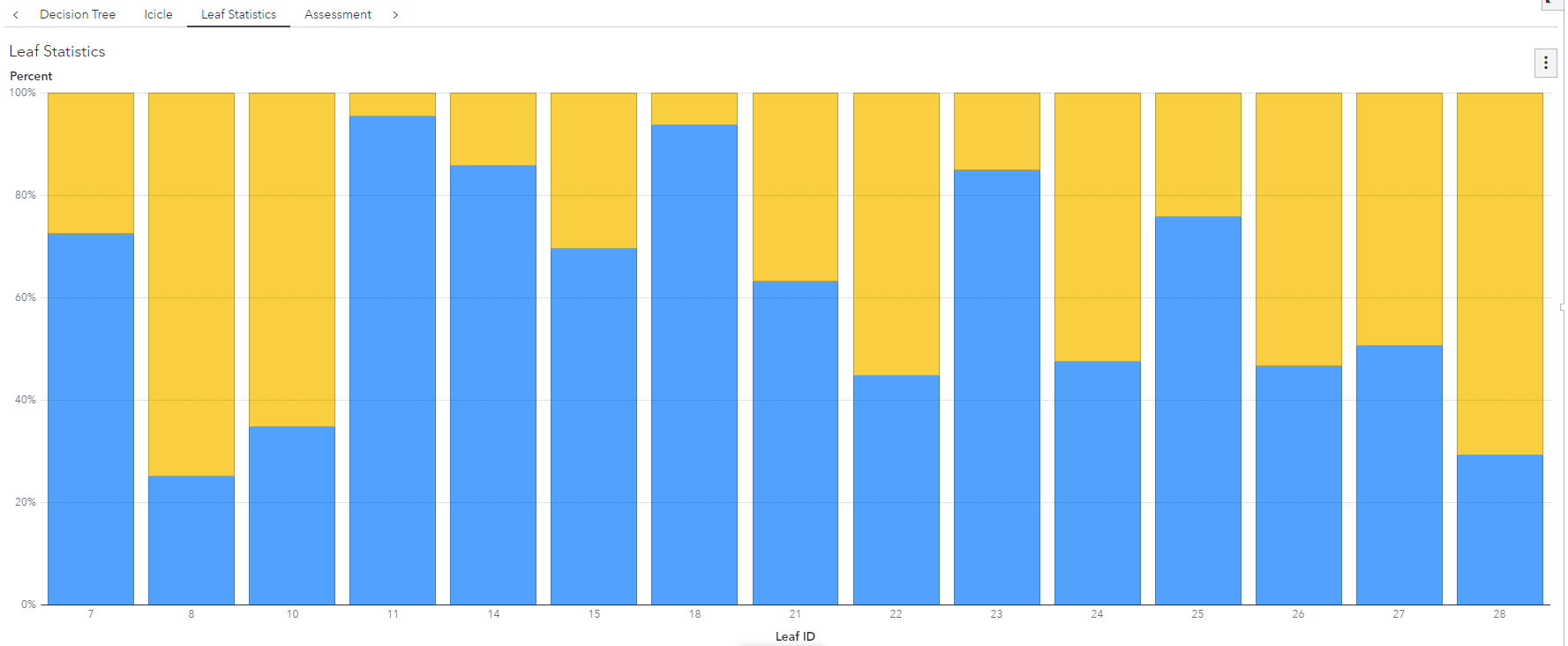
**Note:** When you select a node in either the decision tree or the icicle plot, the corresponding node is selected in the other location.



This plot provides the variable importance information for the variables that are used in the tree.

The importance value is determined by the total Gini reduction. The variable **logi\_rfm5 Count Purchased Past 3 Years** seems to be the most important effect in this decision tree. Examine the tree to recall that this variable is used in the first split.

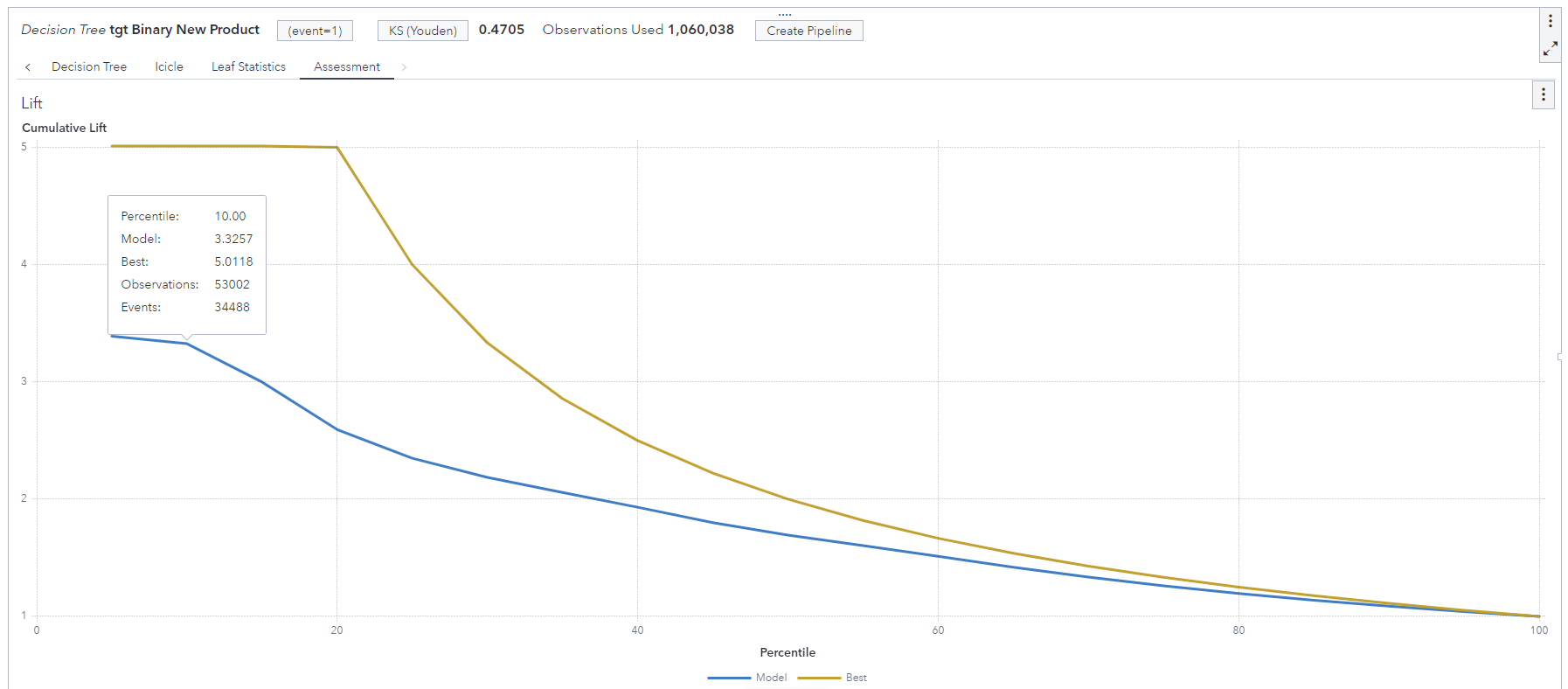
1. Right-click in the variable importance plot and select **Leaf statistics**.



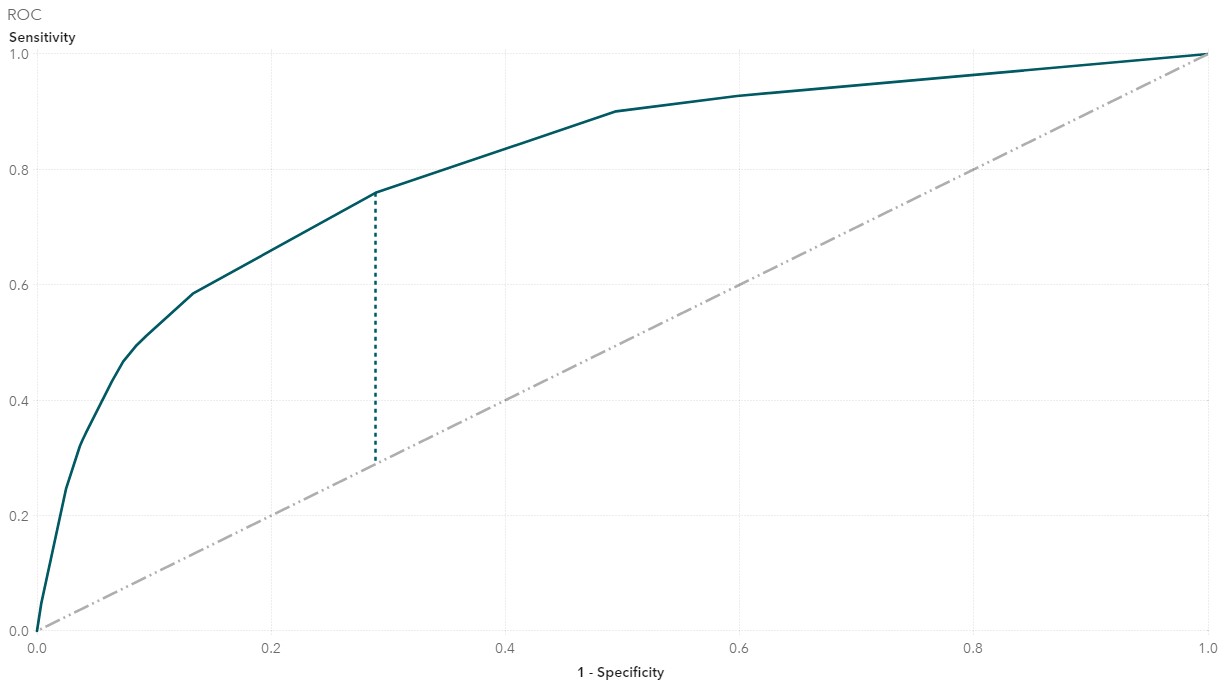
The Leaf Statistics panel displays a bar chart of simple statistics for each leaf. This enables you to do a quick visual comparison of leaf purity. Node 8 contains the highest percentage of responders at nearly 75%.

1. Click the **Assessment** tab. 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 (orange line). For example, when rank-ordered by the model, the top 10% (percentile) of the data has approximately 3.3 times as many responders as a random ordering of the data.



1. Right-click and select **ROC**. The ROC chart summarizes the true positive rate (Sensitivity) and false positive rate (1 - Specificity) across thresholds or cutoffs in the data. The 45-degree line represents the performance of the naïve model, and the vertical blue dashed line corresponds to an optimal threshold in the data.



14.

I

n the report,

c

lick



(

**Maximize**

)

on

the objec

t toolbar to enter maximize mode and see the

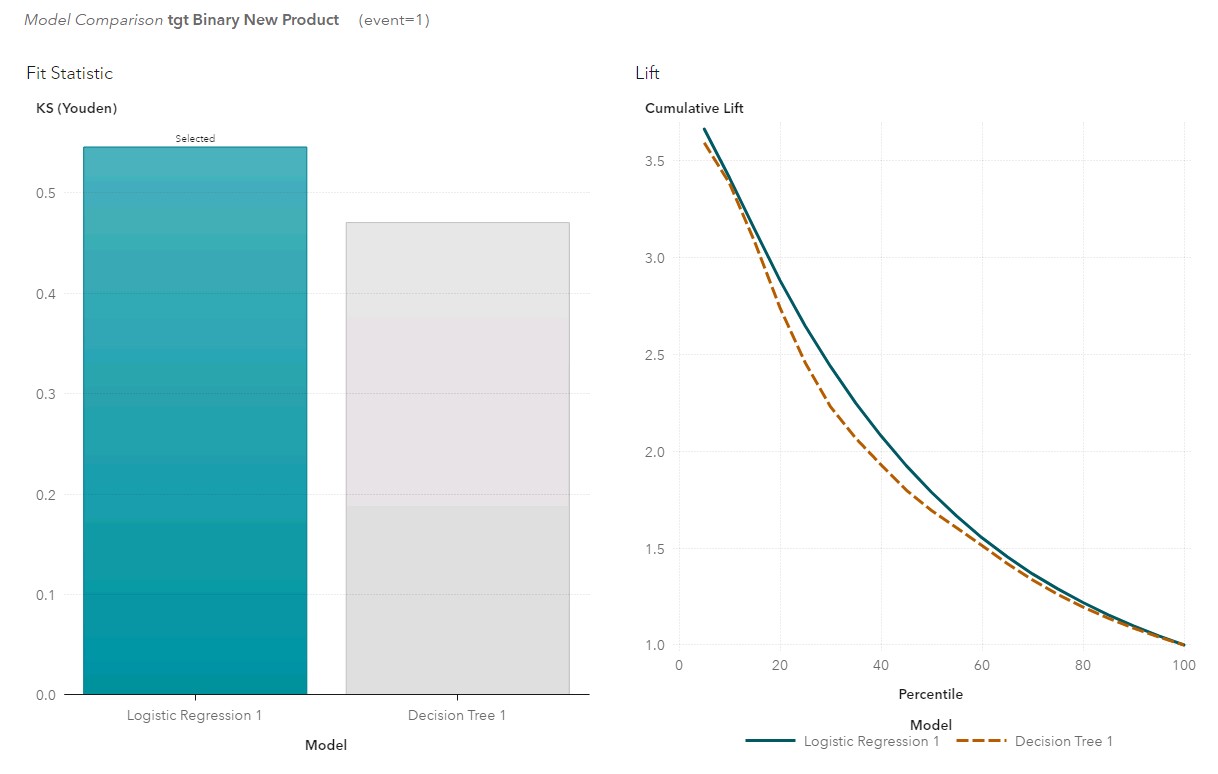
details table.

15. Click the **Variable Importance** tab. These statistics are used to build the variable importance plot. (Right-click in the canvas of the report and select **Export Data** to export results.) To close the details table and exit maximize mode, click  (**Restore**) on the object toolbar.

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

1. Start a new page to begin modeling a logistic regression. Click  (**New page**).
2. Click the **Objects** pane.
3. Under SAS Visual Statistics, either double-click or drag and drop **Model Comparison** onto the canvas. Select **Close** if you receive a warning about models not being available for comparison.
4. Click **Select all** under Available models. (There should be two models.) Click **OK**.

The logistic model is slightly preferred to the tree model that is based on the misclassification statistic. Also, observe the lift chart. Notice that the logistic model has a higher lift across all percentiles in the data.



6.

On the report, c

lick



(

**Maximize**

)

to see the

details table

.

Review the Statisti

cs table. N

ot

ic

e

that the logistic regression model has a higher C sta

tistic.

