Lesson 1 Interactive Modeling With SAS® Visual Statistics

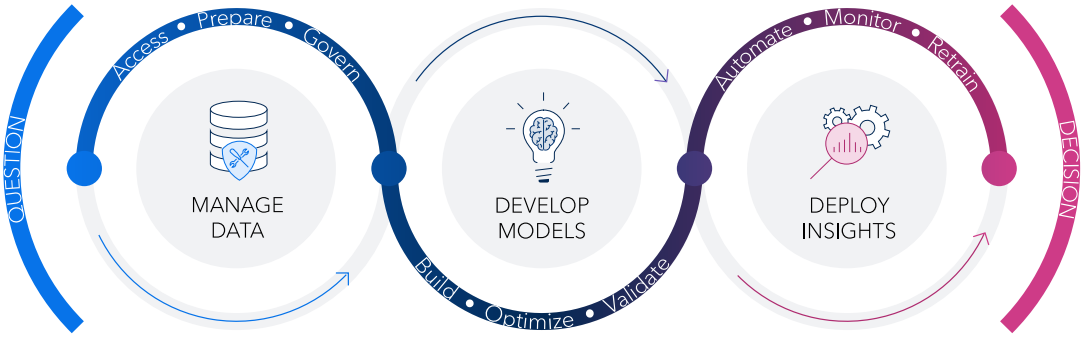
1.1 Introduction

Objectives

- Introduce SAS Viya.
- Introduce SAS Visual Statistics in SAS Viya.
- Review logistic regression basics.
- Demonstrate logistic regression model.

sas innovate 2025

SAS Viya Connects All Aspects of the AI and Analytics Life Cycle



sas innovate 2025

SAS Viya is a cloud-enabled, in-memory analytics engine that provides quick, accurate, and reliable analytical insights. In SAS Viya, the SAS High-Performance Architecture enables the high-performance analytics engine. The CAS In-Memory Engine continues the ability to perform processing in memory and the ability to distribute processing across nodes in a cluster. The CAS In-Memory Engine adds highly efficient node-to-node communication and uses an algorithm to determine the optimal number of nodes for a given job.

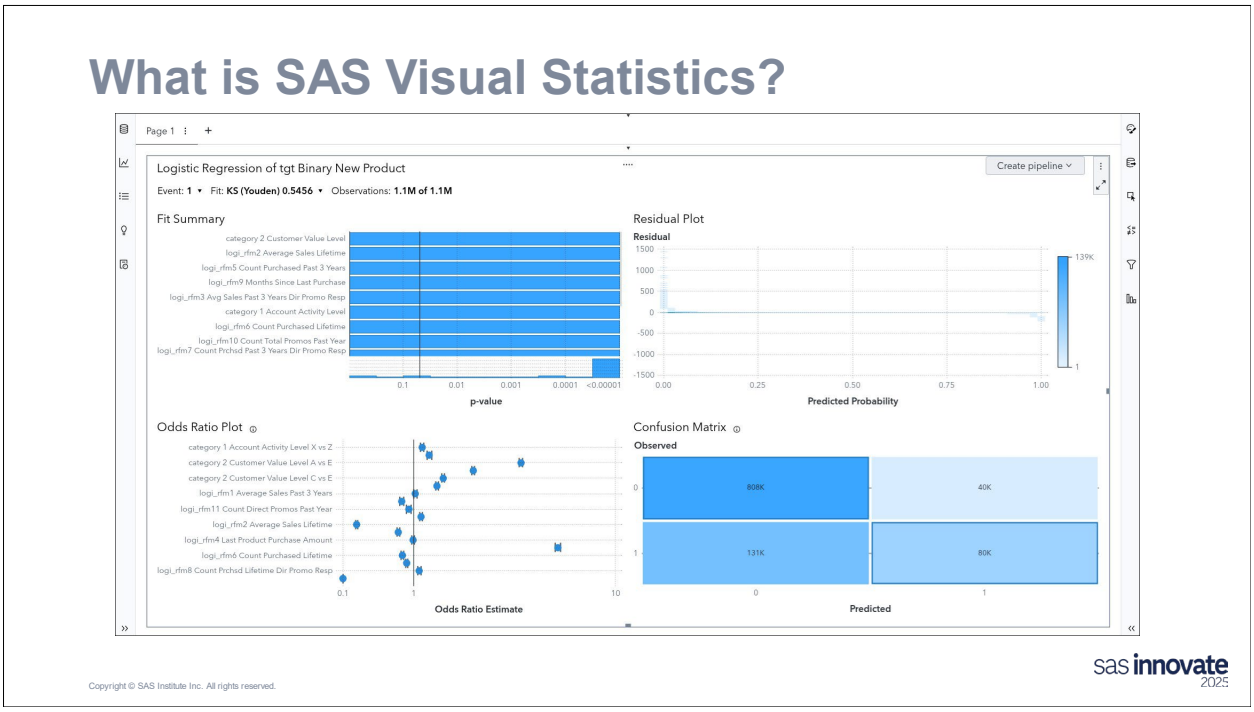
SAS Cloud Analytic Services, or CAS, is a server that provides the cloud-based run-time environment for data management and analytics with SAS. By *run-time environment*, we refer to the combination of hardware and software where data management and analytics take place.

The server can run on a single machine or as a distributed server on multiple machines. The distributed server consists of one controller and one or more workers. This architecture is often referred to as a *massively parallel processing architecture*. For both modes, the server is multi-threaded for high-performance analytics.

The distributed server has a communication layer that supports fault tolerance. A distributed server can continue processing requests even after losing connectivity to some nodes. The communication layer also enables you to remove or add nodes from a server while it is running.

One of the design principles of the server is to handle large problems and to work with tables that exceed the memory capacity of the environment. In order to address this principle, data in the server is managed in blocks. Whenever needed, the server caches the blocks on disk. It is this feature that enables the server to manage memory efficiently, handle large data volumes, and remain responsive to requests.

You can use a variety of interfaces to interact with the CAS In-Memory Engine. These interfaces include SAS Studio, which is a browser-based interface for writing SAS code. You can also use programming interfaces for R, Python, Java, and Lua to access this CAS functionality. In addition, you can continue to submit SAS code in batch mode.

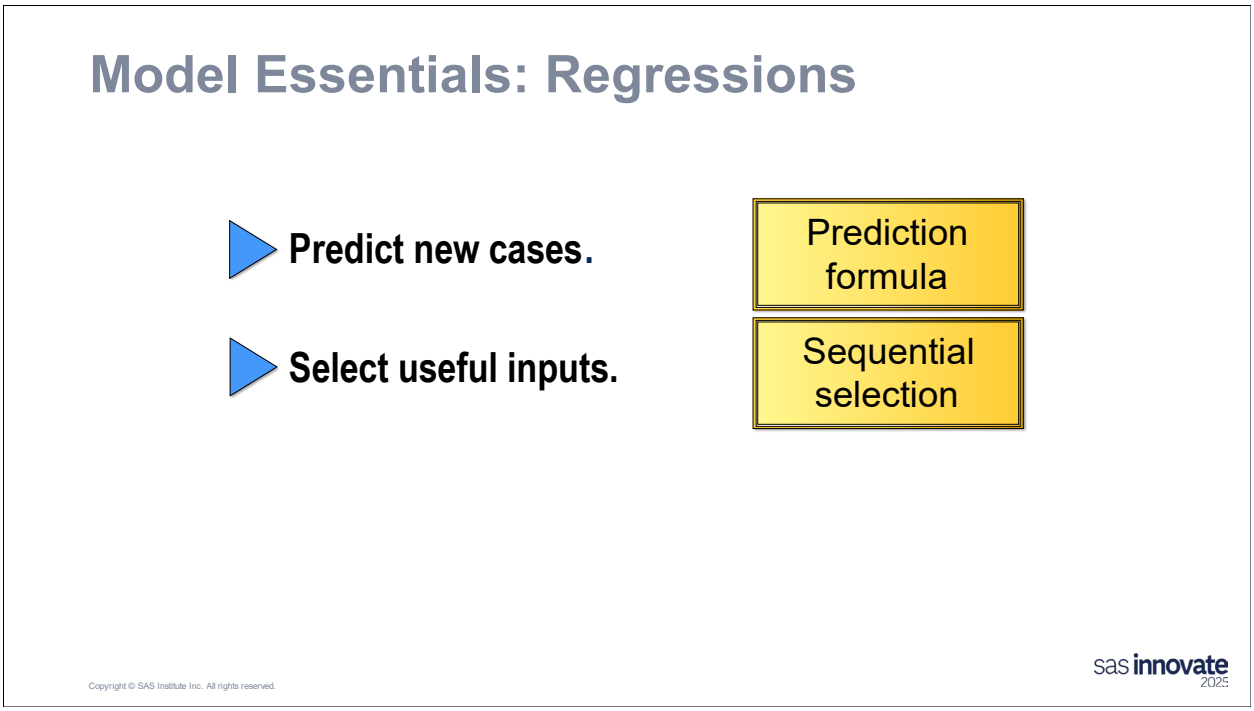


SAS Visual Statistics on SAS Viya is an add-on to SAS Visual Analytics that enables you to develop and test models using a scalable in-memory engine in a common SAS Viya environment. SAS Visual Analytics enables you to explore, investigate, and visualize data sources to uncover relevant patterns. SAS Visual Statistics extends these capabilities by creating, testing, and comparing models based on the patterns discovered in the analytics and visualizations. SAS Visual Statistics can export the score code, before or after performing model comparison, for use with other SAS products and to put the model into production.

SAS Visual Statistics enables you to rapidly create powerful statistical models in an easy-to-use, web-based interface. After you have created two or more competing models for your data, SAS Visual Statistics provides a model comparison tool. The model comparison tool enables you to evaluate the relative performance of two or more models against each other and to choose a champion model. A wide variety of model selection criteria is available. Regardless of whether you choose to perform a model comparison, you can export model score code for any model that you create. With exported model score code, you can easily apply your model to new data.

The following models are available in SAS Visual Statistics:

- ***Linear regression*** attempts to predict the value of an interval response as a linear function of one or more effect variables.
- ***Logistic regression*** attempts to predict the probability that a binary or ordinal response will acquire the event of interest as a function of one or more effects.
- ***Nonparametric logistic regression*** is an extension of the logistic regression model that allows spline terms to predict a binary response.
- ***Generalized linear model*** is an extension of a traditional linear model that allows the population mean to depend on a linear predictor through a nonlinear link function.
- **Generalized *additive model*** is an extension of the generalized linear model that allows spline terms to predict an interval response.
- ***Decision tree*** creates a hierarchical segmentation of the input data based on a series of rules applied to each observation.
- ***Cluster*** segments the input data into groups that share similar features.



Regression models enable you to characterize the relationship between a response variable and one or more predictor variables. With linear regression, the response variable is continuous. With logistic regression, the response variable is categorical. When the response variable is limited to only two categories (dichotomous), the appropriate model is binary logistic regression.

Binary Logistic Models

- **Credit Scoring:** Can credit score and home ownership predict loan default?

- **Predictor Variables:**
- Credit Score: 300-850
- Home Ownership: Yes/No/Rent

- **Response Variable:**
- Loan Default: Yes/No

sas innovate 2025

One example of a binary logistic regression model can be found in credit scoring. Can credit score and home ownership help predict the likelihood of a customer defaulting on a loan? In this scenario, one of the predictor variables is continuous (credit score) and the other happens to be categorical (home ownership) with three distinct levels. The response variable is also categorical and coded as character values.

1.2 Hands-On Workshop

In this workshop, you use the data set **vs_bank**. This data set consists of observations taken from account holders at a large financial services firm. The accounts represent consumers of home equity lines of credit, automobile loans, and other short- to medium-term credit instruments. Appropriate data cleansing has already been applied, so we can begin with statistical modeling. The target variables relate to whether that account holder purchased a new product from the bank in the past year. The data sets contain more than 1 million rows and 24 columns. A list of variables and their labels is shown below.

Target Variable

B_TGT	New Product (Binary)

Categorical Inputs

CAT_INPUT1	Account Activity Level
CAT_INPUT2	Customer Value Level

Interval Inputs

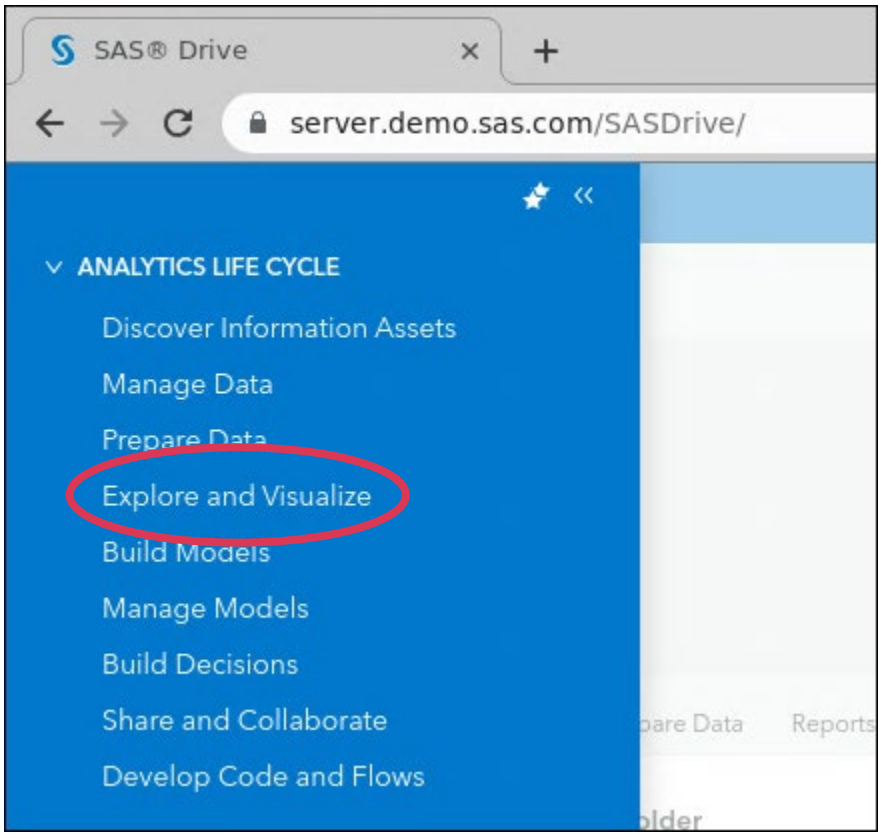
RFM1	Average Sales Past Three Years
RFM2	Average Sales Lifetime
RFM3	Avg Sales Past Three Years Dir Promo Resp
RFM4	Last Product Purchase Amount
RFM5	Count Purchased Past Three Years
RFM6	Count Purchased Lifetime
RFM7	Count Prchsd Past Three Years Dir Promo Resp
RFM8	Count Prchsd Lifetime Dir Promo Resp
RFM9	Months Since Last Purchase
RFM10	Count Total Promos Past Year
RFM11	Count Direct Promos Past Year
RFM12	Customer Tenure

Note: Other variables, not listed here, are also included in the data set. Variables with the prefix **I_** are imputed. Variables with the prefix **RI_** are imputed and replaced. Variables with the prefix **LOGI_** are imputed and log transformed. Variables with the prefix **DEMOG_** are demographic inputs.

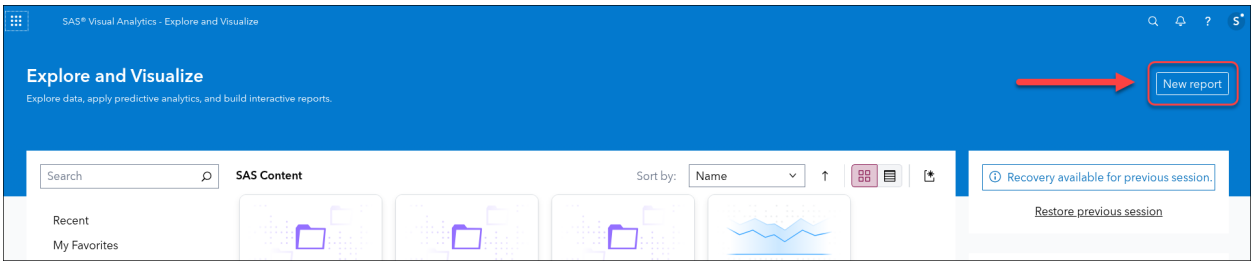
Building a Logistic Regression Model

This demonstration illustrates how to build a logistic regression model in SAS Visual Analytics. 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. You then perform model validation and variable selection.

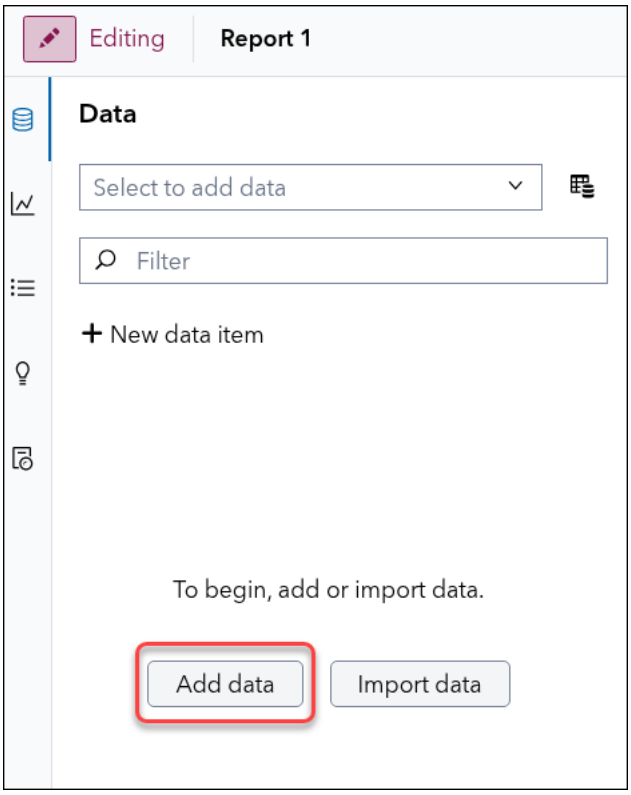
1. From the desktop, open Google Chrome.
2. From the Google Chrome toolbar, select **SAS Drive**.
3. Enter **student** in the **User ID** field, if necessary.
4. Enter **Metadata0** in the **Password** field, if necessary.
5. Select **Sign in**.
6. If requested to save the password, select **Save**.
7. Select **YES** when asked about assumable groups.
8. Access the applications menu in the upper left of the window and select **Explore and Visualize**.



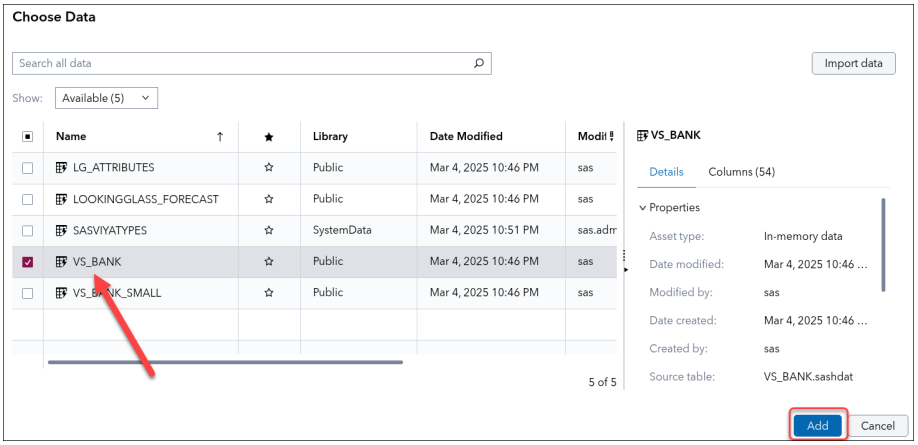
9. Load the SAS data set into CAS.
   a. In the Explore and Visualize window, click **New report**.



   a. On the Data tab on the left of the screen, select **Add data** to load an in-memory table to SAS Visual Analytics.



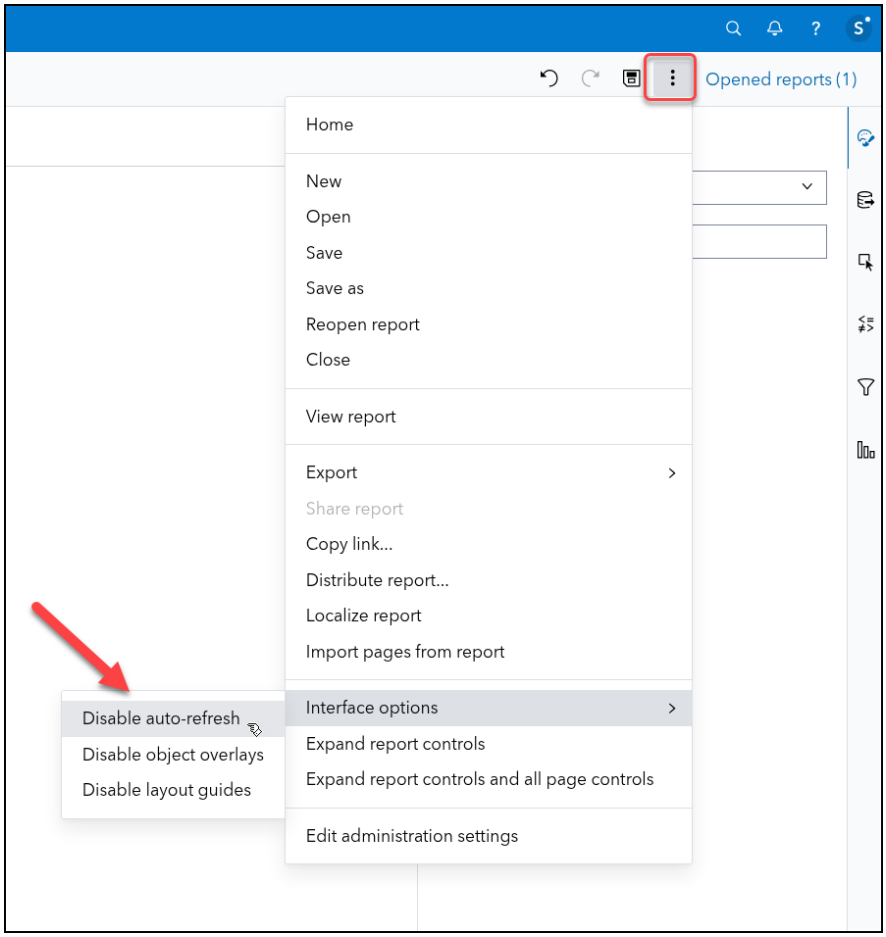
b. In the Choose Data window, select **VS_Bank > Add**.



c. The table is now available to the newly created report.

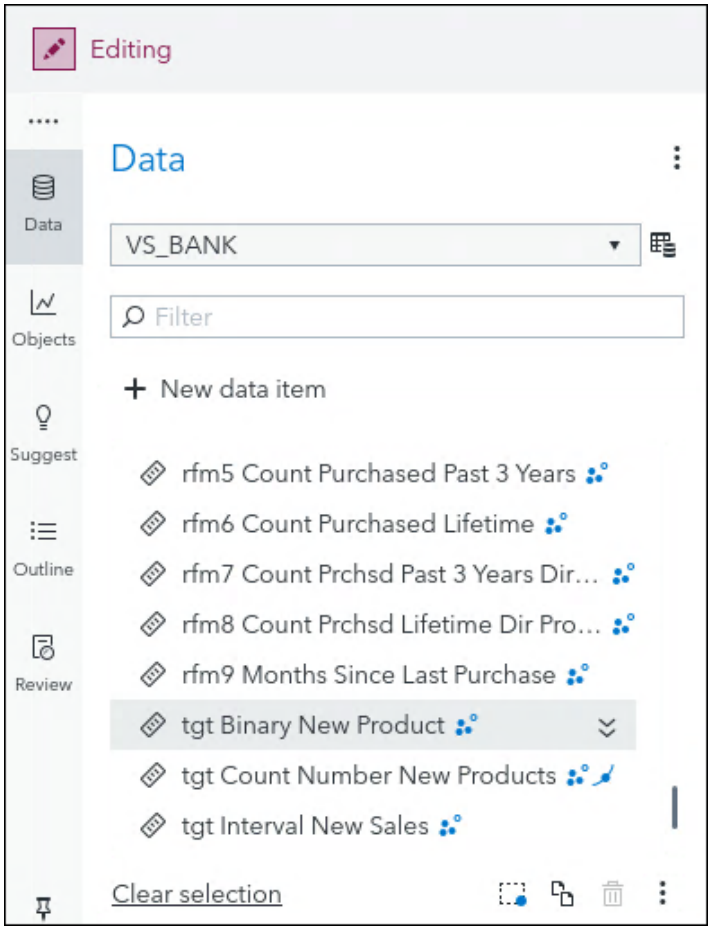
10. From the menu bar, click ⋮ (**More**) and select **Interface options** > **Disable auto-refresh**.

Disabling auto-refresh enables you to set up several roles in the model before it is created. Otherwise, the model is updated anytime that a change is made.

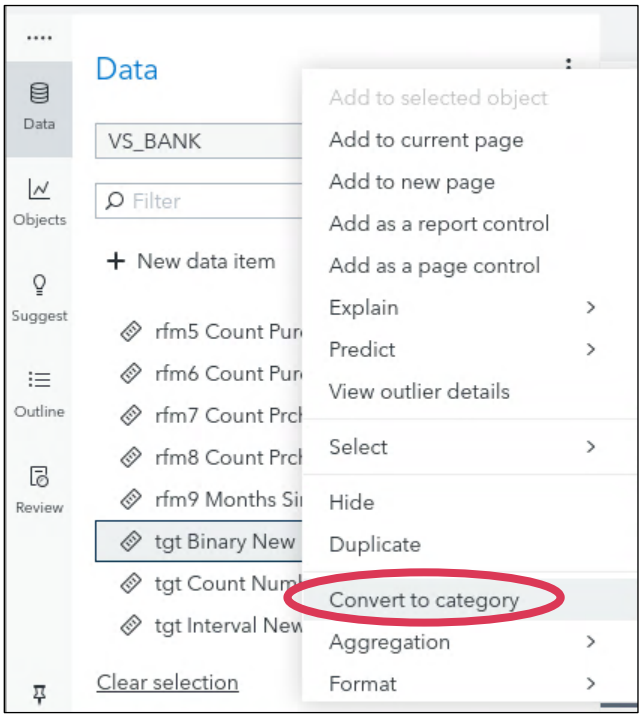


11. On the Data tab on the left of the screen, scroll down and find the measure **tgt Binary New Product**.

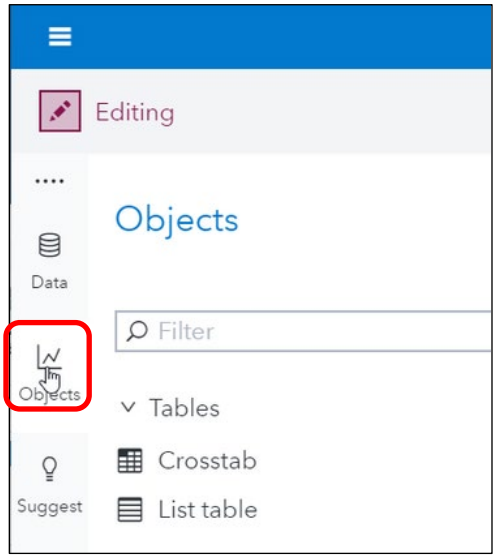
    The variable **tgt Binary New Product** (**b_tgt**) is the primary dependent variable for categorical response modeling in this workshop. It is a binary flag that codes responders with 1 and non-responders with 0. Because it is numeric, it is treated as interval valued by default.



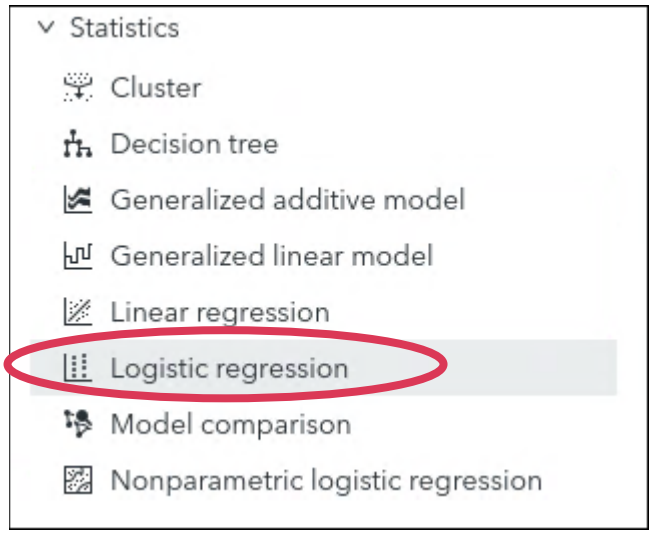
12. Right-click **tgt Binary New Product** and select **Convert to category**.



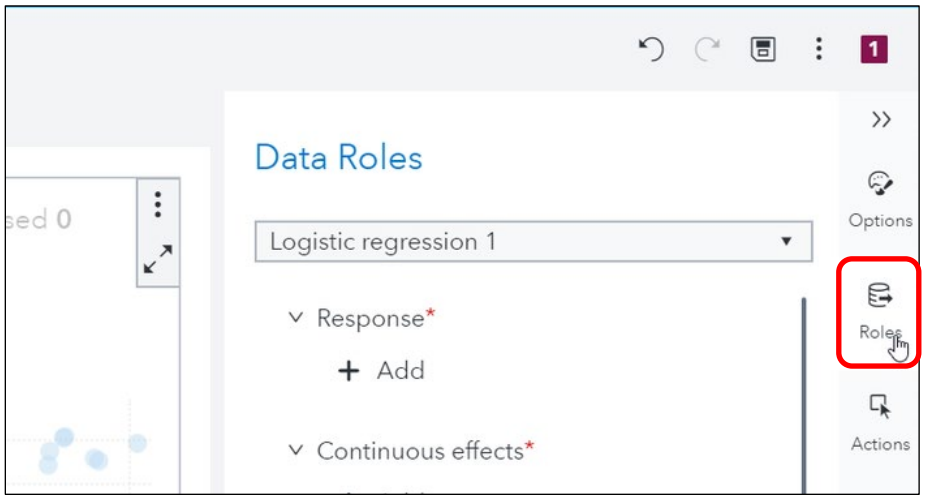
13. Click the **Objects** tab below the Data tab**.**



14. Scroll down and find the list of statistics.
15. Double-click **Logistic regression** or drag and drop it onto the canvas on the page.

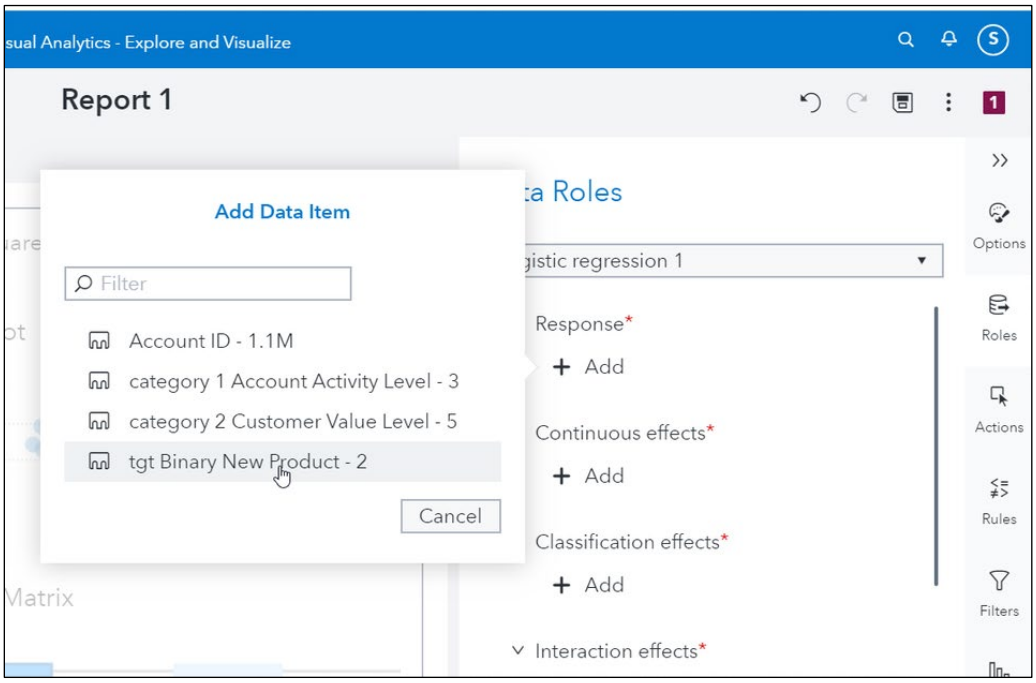


16. On the rightmost side of the screen, click the **Roles** tab.



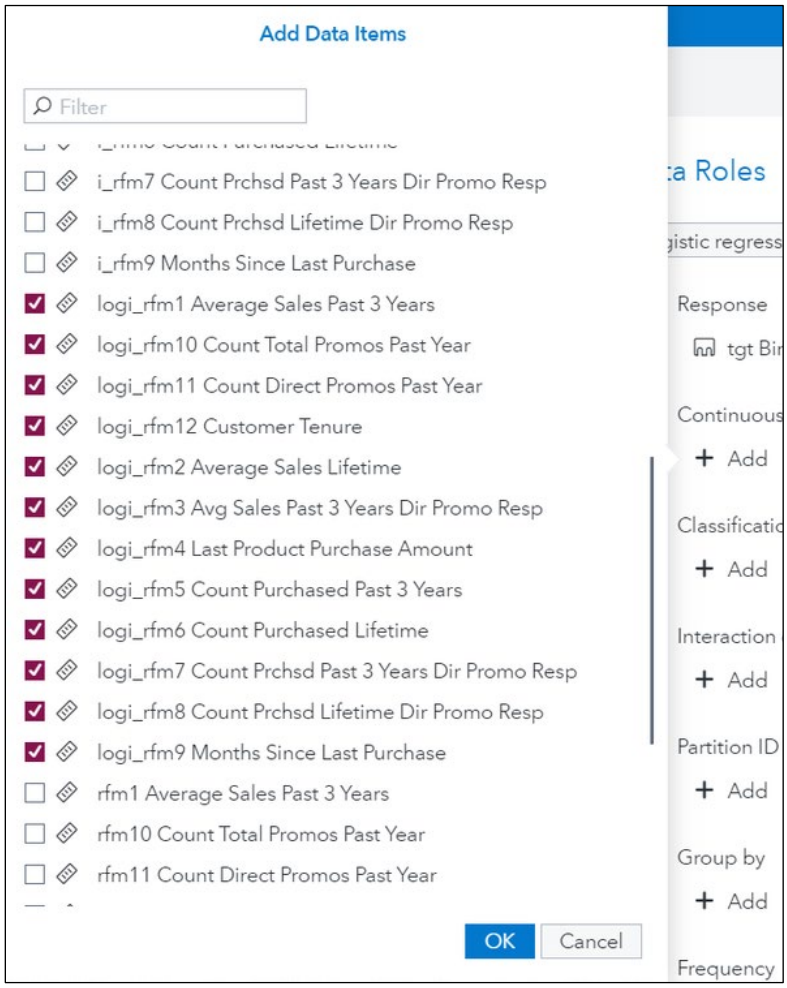
17. Under **Response**, click **Add**.

18. Select **tgt Binary New Product** from the Add Data Item window.

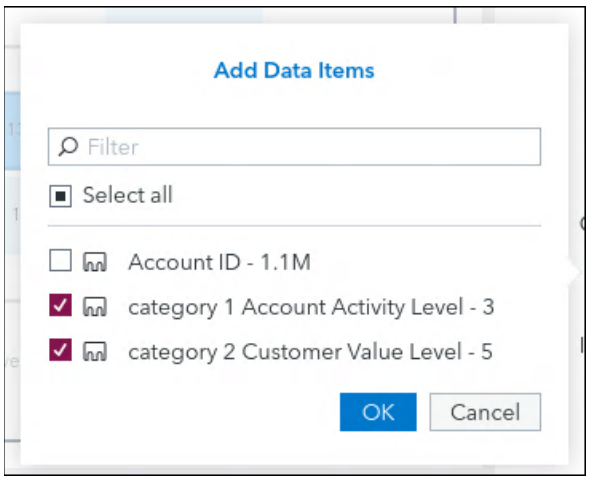


19. Under **Continuous effects**, select **Add**.

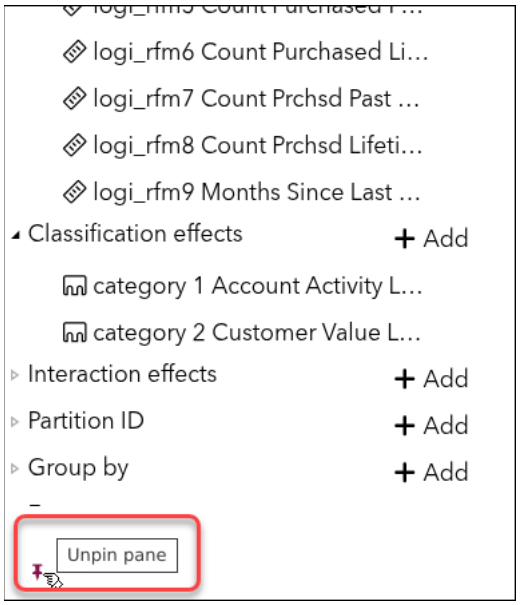
20. Use the Shift key to select all items with a **logi_rfm** prefix.



21. With the 12 data items selected, click **Apply**.
22. Under **Classification effects**, select **Add**.
23. Click **category 1 Account Activity Level** and **category 2 Customer Value Level**. Then click **Apply**.



24. Create the logistic regression model by clicking ⋮ (**More**) and selecting **Interface options** > **Enable auto-refresh**.
25. Collapse (unpin) the right pane to maximize the display of the logistic regression model.



General model information appears along the top of the model, including the name of the response variable, the event of interest, the model evaluation criteria, and the number of observations used to build the model.

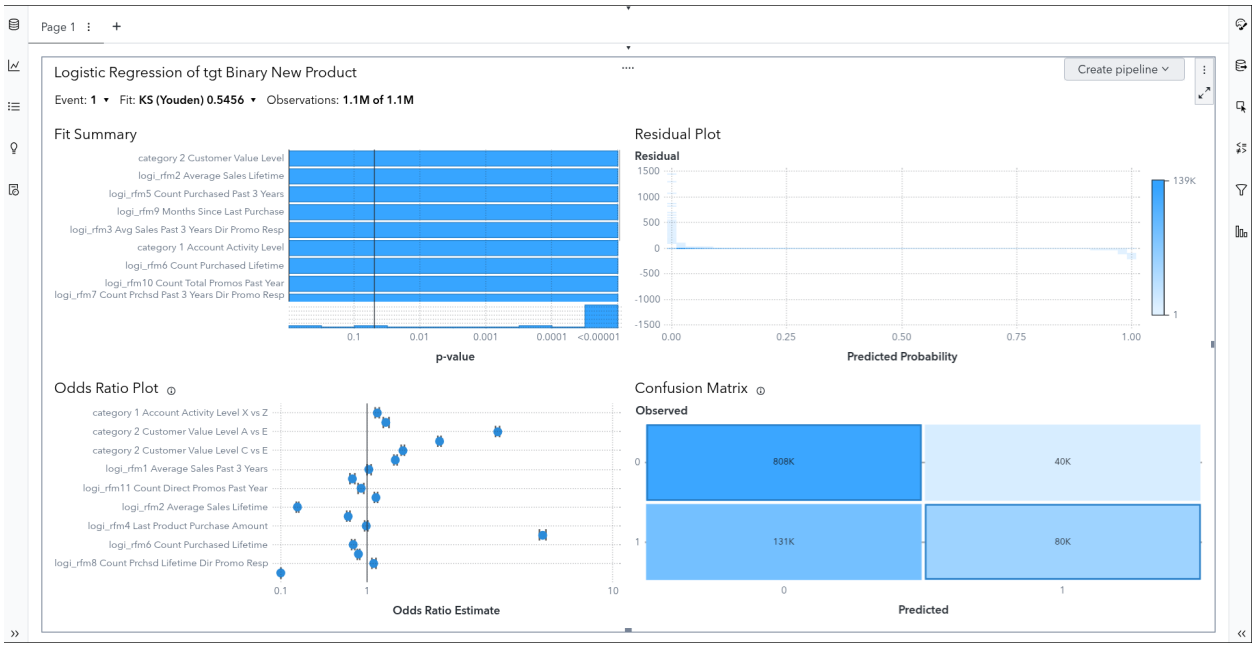
For our binary target variable, a value of 1 represents an account that did contract for a product during the campaign (in other words, a sale). As seen in the summary bar, the logistic regression models the event of interest (making a sale) with **event=1**.

Odds ratio estimates compare the odds of an event in one group to the odds of the event occurring in another group. The odds ratio estimates are provided for each effect in the model

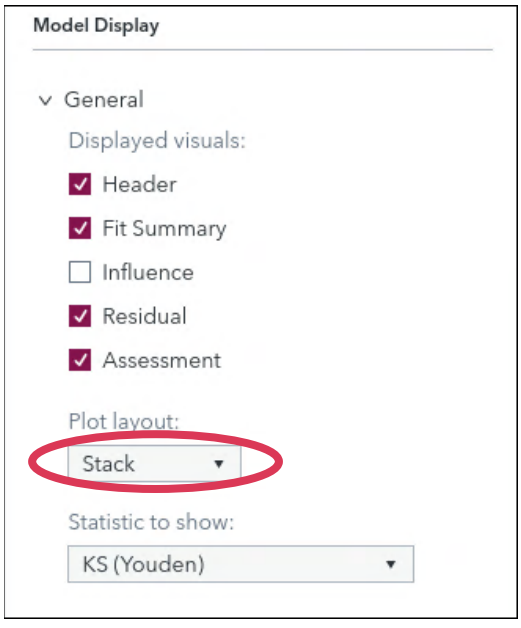
The Fit Summary pane is used to determine the most significant predictor variables that affect the response variable. The variable importance plot displays the effects on the Y axis and the *p*-values on the X axis. The variable importance is based on the negative log of the *p*-value (-log(*p*-value)). A larger –log(*p*-value) indicates a more important variable.

The residual plot is used to access the quality of the model and to identify outlier observations. The plot appears as either a scatter plot for smaller data or as a heat map when used with larger data.

The confusion matrix describes the performance of a classifier. It contains frequency counts of both the correct and incorrect predictions broken down by class.

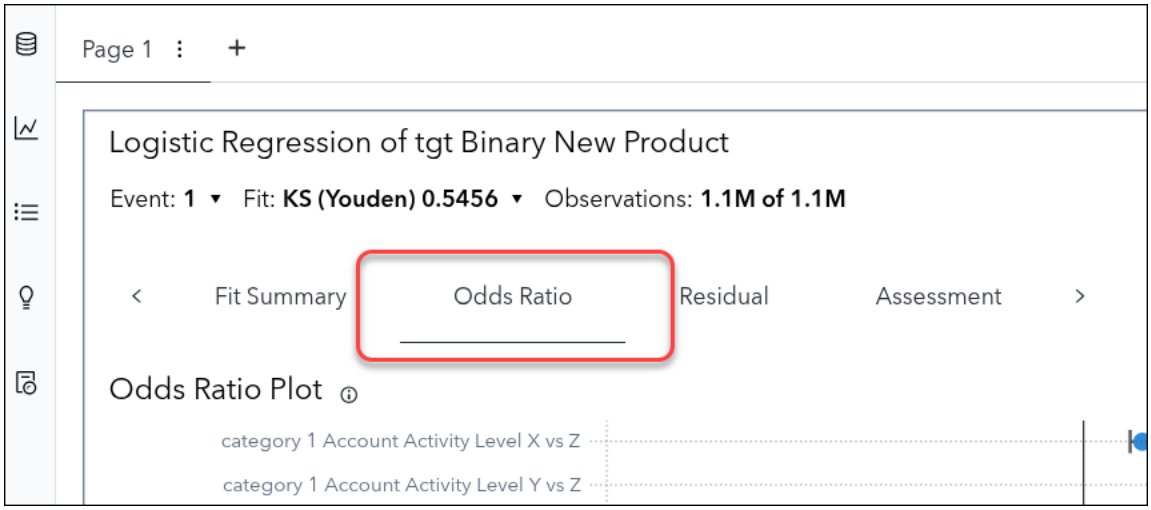


26. On the Options tab on the right of the screen, under **Model Display**, expand **General** and change **Plot layout** to **Stack**. The model canvas appears, and **Fit Summary** is the default tab selected.



In the Fit Summary pane, one of the variables is not significant at 5%. One of the variables could be considered on the border of significance with a *p*-value of .0393.

27. Click the **Odds Ratio** tab above the plot on the canvas.

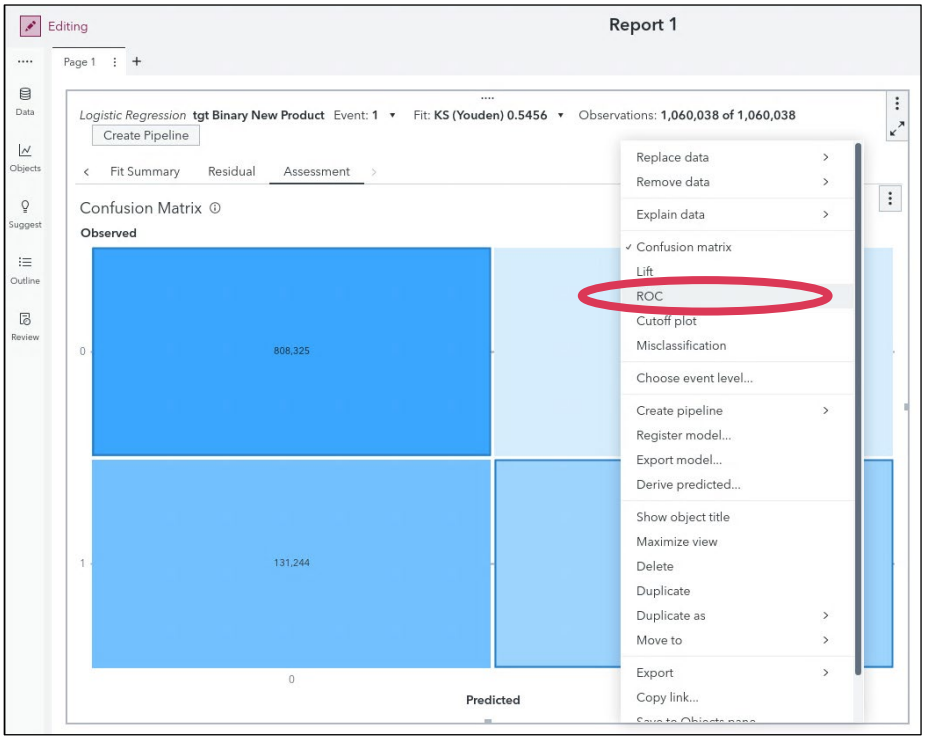


Odds ratios are particularly useful for interpreting the effects of both categorical and continuous effects on the target of a logistic regression model. This is because the odds ratio quantifies the change in odds of the outcome for a one-unit change in the predictor variable. The largest positive odds ratio estimate in this plot is 5.8 for the variable **logi_rfm5 Count Purchased Past 3 Years**. In layman's terms for each additional product purchased in the last 3 years, we are almost 6 times more likely to find a purchaser.

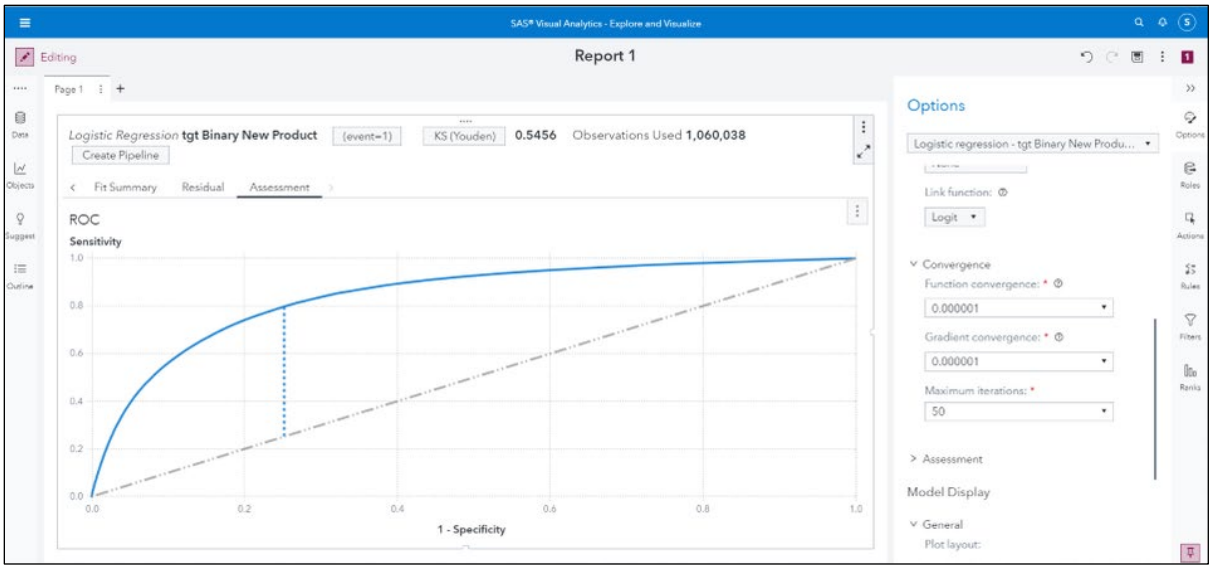
28. Click the **Assessment** tab above the plot on the canvas.

The confusion matrix shows that 80,265 customers who made a purchase were predicted to make a purchase. These are known as the *true positives*. It also shows that 40,204 customers were incorrectly classified to have made a purchase (*false positives*).

29. Right-click on the confusion matrix and select **ROC**.

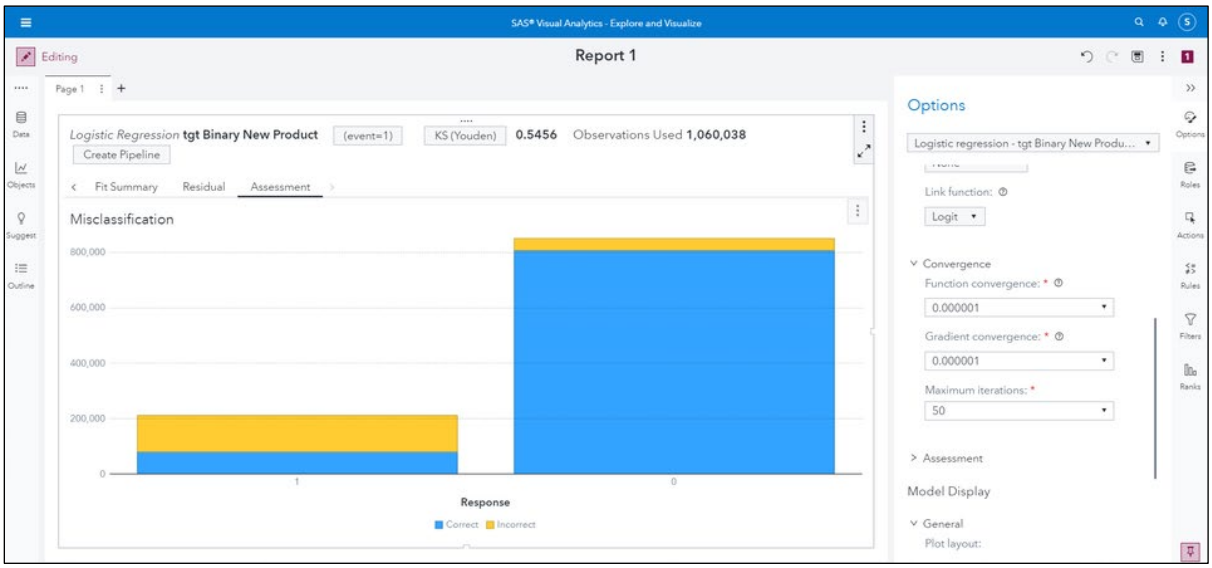


The ROC (receiver operating characteristic) chart is a graphical display that gives a measure of the predictive accuracy of a logistic model. The classification accuracy of a model is demonstrated by the area under the curve or the degree that the ROC curve pushes upward and to the left.



30. Right-click the ROC chart and select **Misclassification**.

    A misclassification plot displays how many observations were correctly and incorrectly classified for each value of the response variable. This is a graphical representation of the confusion matrix.

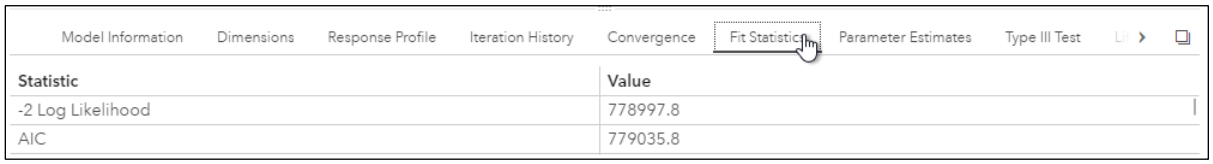


31. On the report, click  (**Maximize**) to see the Details table.

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

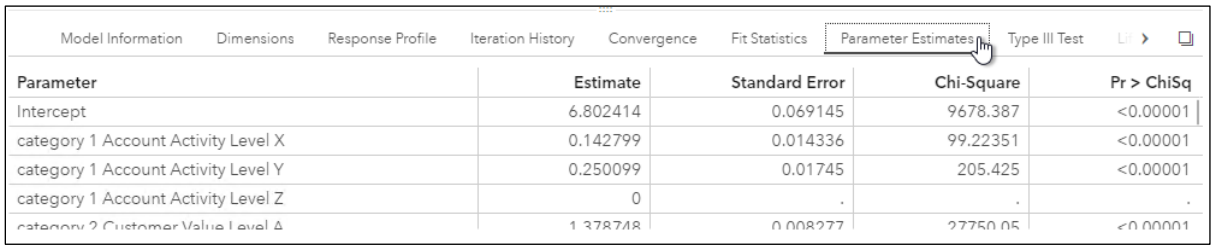
Ordered Value	Count	tgt Binary New Product
1	211509	1
2	848529	0

33. Click the **Fit Statistics** tab. The Fit Statistics table displays statistics about the estimated model.



Statistic	Value
-2 Log Likelihood	778997.8
AIC	779035.8

34. Click the **Parameter Estimates** tab. The Parameter Estimates table displays the parameter estimates or coefficients of each model effect and their associated statistics.

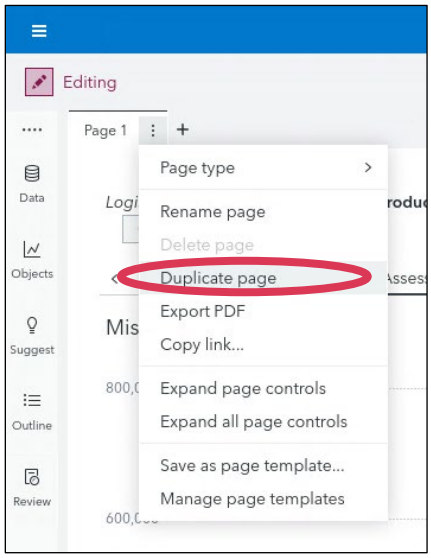


Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	6.802414	0.069145	9678.387	<0.00001
category 1 Account Activity Level X	0.142799	0.014336	99.22351	<0.00001
category 1 Account Activity Level Y	0.250099	0.01745	205.425	<0.00001
category 1 Account Activity Level Z	0	.	.	.
category 2 Customer Value Level A	1.378748	0.008277	27750.05	<0.00001

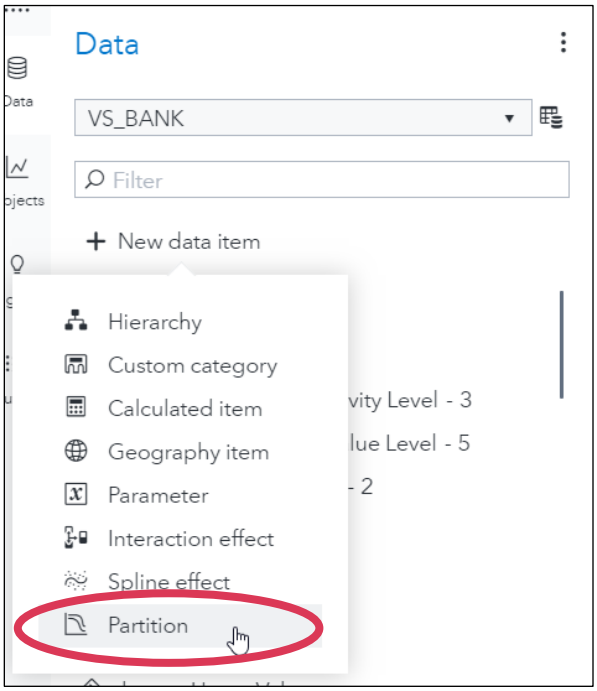
35. In the upper right corner of the report, click  (**Restore**) from the object toolbar to close the Details table and exit maximize mode.

Model Validation and Variable Selection

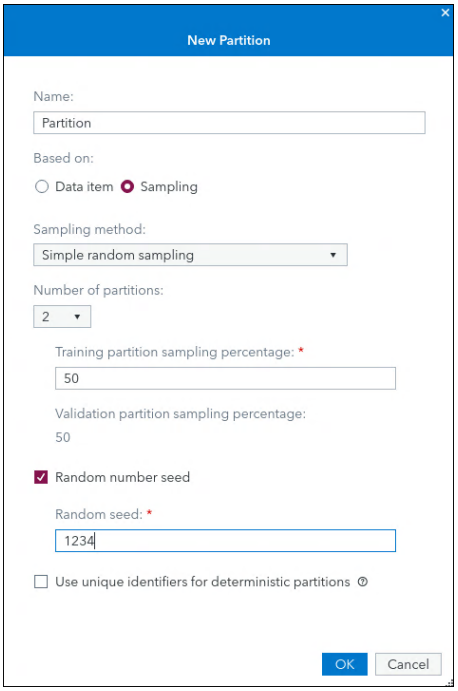
1. On the Page 1 tab of the Logistic Regression, click ⋮ (**Page menu**) and select **Duplicate Page**.



2. On the Data tab, click **+ New data item** and then select **Partition** from the drop-down list.



3. Create the partitions.
   a. In the New Partition window, enter **50** in the **Training partition sampling percentage** field.
   b. Select the box for **Random number seed** and enter **1234** in the input area.
   c. Click **OK**.

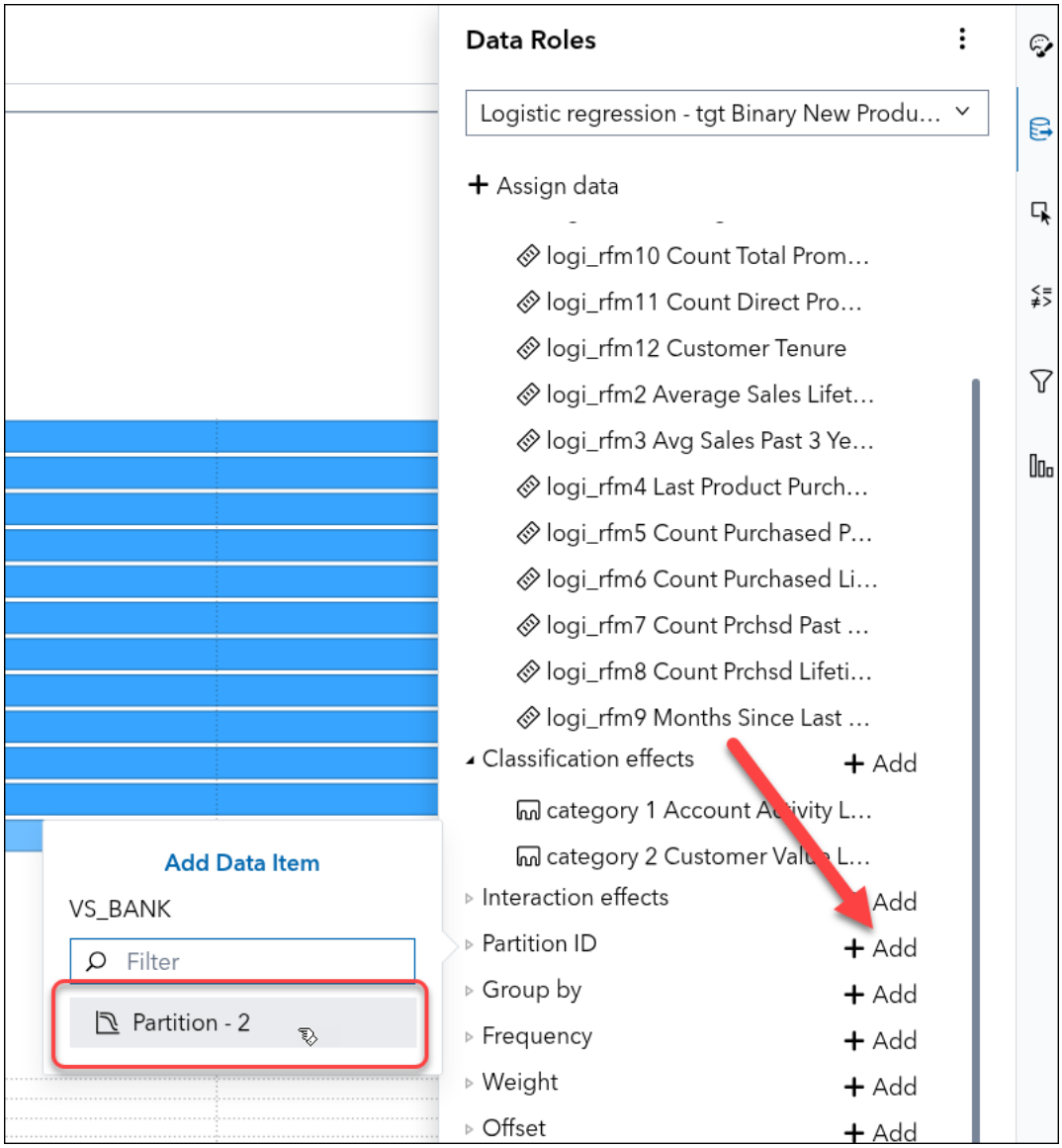


A standard strategy for honest assessment of model performance is data splitting. A portion of the data is reserved for fitting the model, known as the *training data set*. The remaining portion, known as the *validation data set*, is held out for empirical validation.

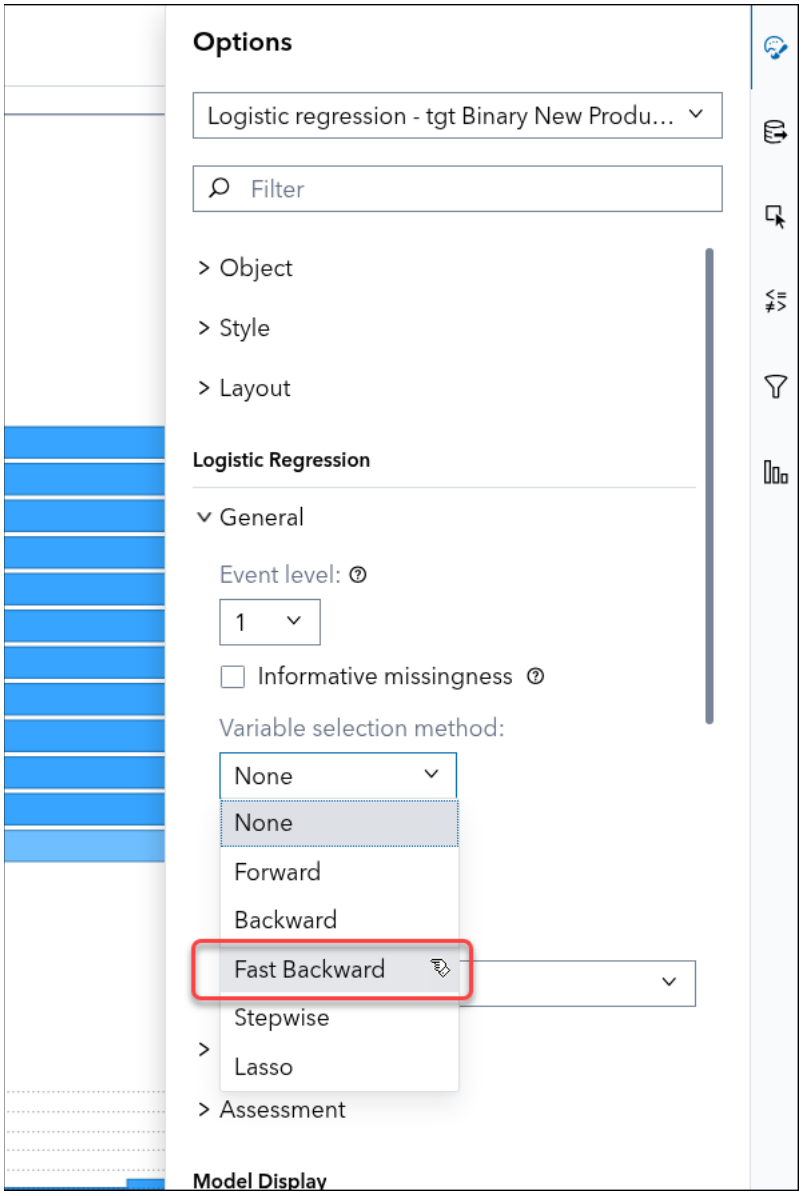
The new partition variable is added to the Category list.

Even with a random seed specified, we might still see nondeterministic results due to the difference in data distribution and computational threads or the walker used to sample the partition column.

4. Select the duplicated **Logistic regression** on the canvas of Page 1 (1) to make it the active object.
5. On the Roles tab on the right of the screen, scroll down and add **Partition** as the partition ID.



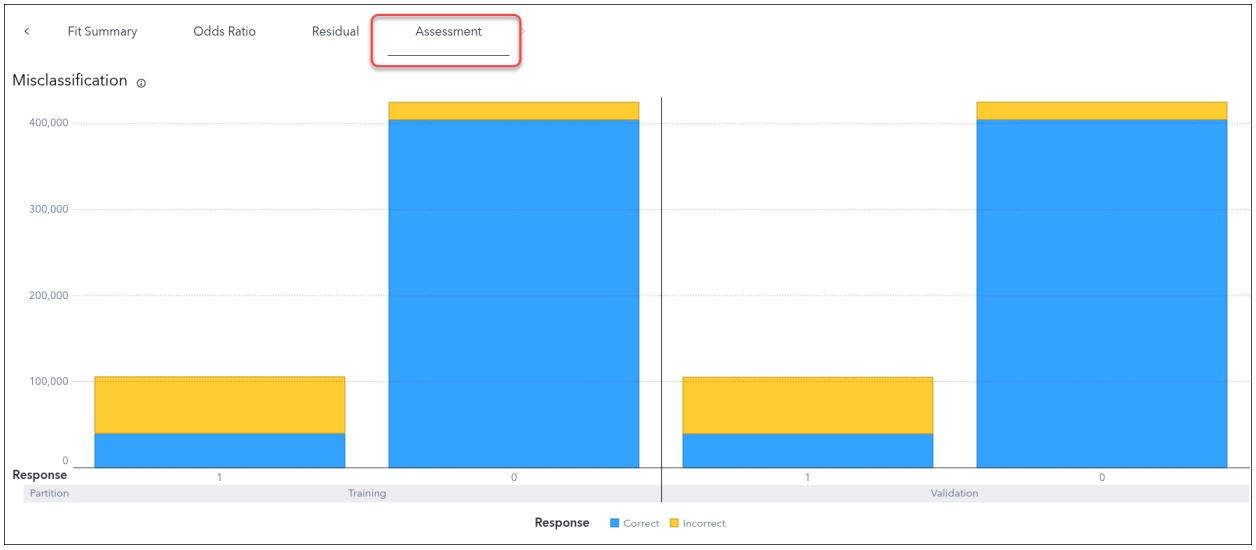
6. On the Options tab, select **Fast Backward** for the variable selection method.



7. Keep the significance level at .**01**.

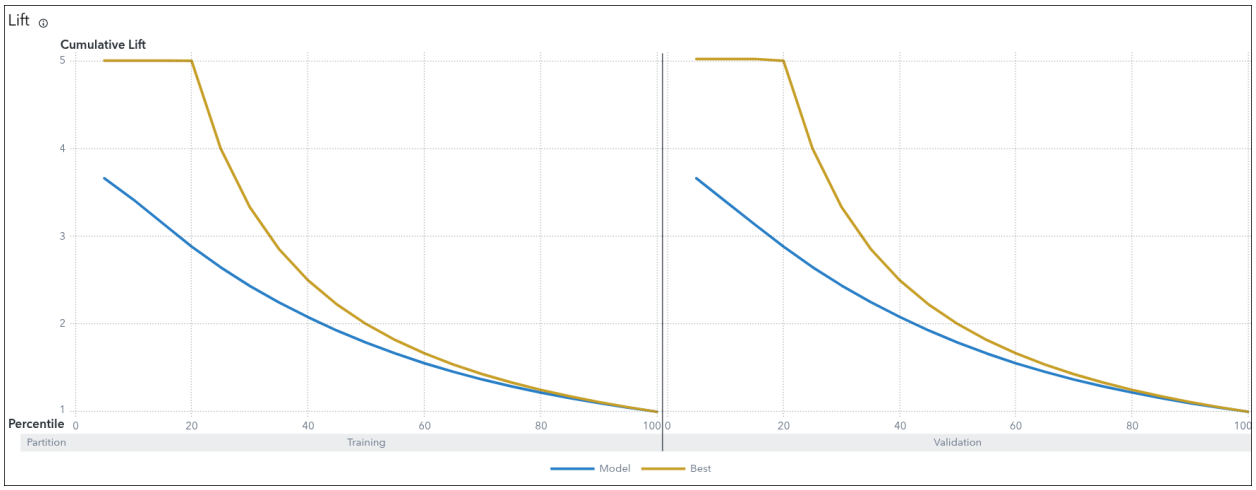
   This technique is available for logistic regression models, and it uses a numeric shortcut to compute the next selection iteration quicker than the backward selection method.

8. Click the **Assessment** tab on the canvas.



There are now assessment plots for both the training and the validation partitions.

9. Right-click on the misclassification display and select **Lift**.

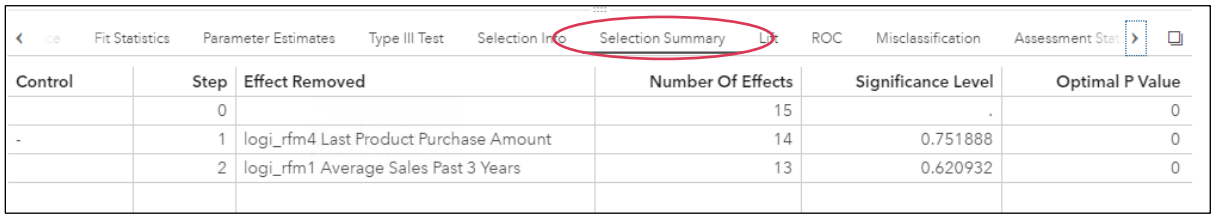


If we mouse over the blue line at the 5th percentile for the validation data, we get information about the model performance. The logistic regression model has a lift of approximately 3.66 at this percentile. Another way to think about this is as follows: if we were to contact the top 5 percent of our customers, we are almost 4 times as likely to reach a responder versus just picking customers at random. If we were to compare our logistic regression model against another model that had a higher lift at the 5th percentile, we would consider the other model to be the better performer. Since the data has been rank-ordered by likelihood of responding, the lower percentiles are more meaningful than the higher percentiles.

10. On the report, click  (**Maximize**) and collapse the panes on the right to see the Details table.

11. In the Details table, click the **Selection Summary** tab to verify that the variables were removed from the model during fast backward variable selection.

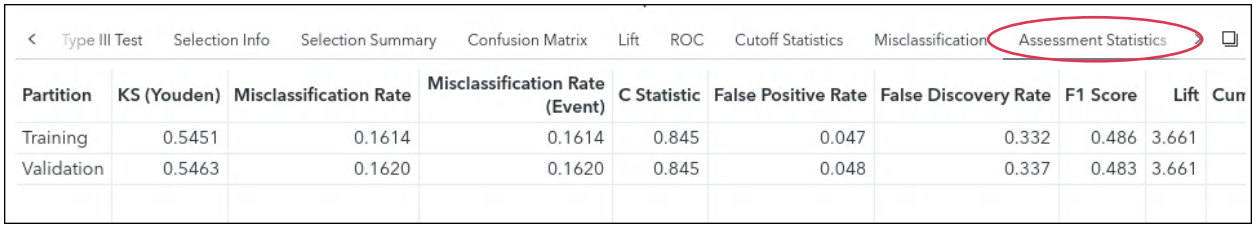
    Note: You can select  (**More Data Tables**) to navigate through the tables.



Fit Statistics | Parameter Estimates | Type III Test | Selection Info | Selection Summary | Lift | ROC | Misclassification | Assessment Stat

Control	Step	Effect Removed	Number Of Effects	Significance Level	Optimal P Value
	0		15	.	0
-	1	logi_rfm4 Last Product Purchase Amount	14	0.751888	0
	2	logi_rfm1 Average Sales Past 3 Years	13	0.620932	0

12. In the Details table, click the **Assessment Statistics** tab.



Type III Test | Selection Info | Selection Summary | Confusion Matrix | Lift | ROC | Cutoff Statistics | Misclassification | Assessment Statistics

Partition	KS (Youden)	Misclassification Rate	Misclassification Rate (Event)	C Statistic	False Positive Rate	False Discovery Rate	F1 Score	Lift	Cun
Training	0.5451	0.1614	0.1614	0.845	0.047	0.332	0.486	3.661	
Validation	0.5463	0.1620	0.1620	0.845	0.048	0.337	0.483	3.661	

    Now that we've included a partition, model fit statistics are available for both training and validation partiitons.

End of Demonstration