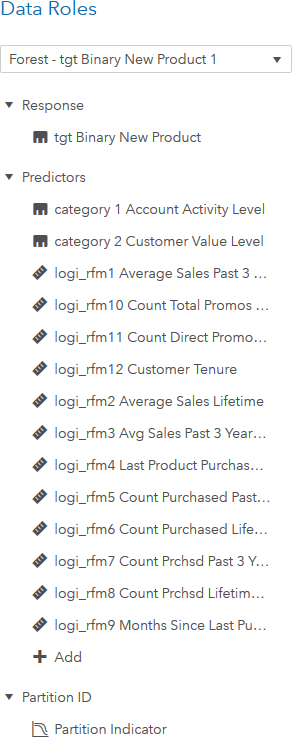
Demonstração 08 – Florestas Aleatórias

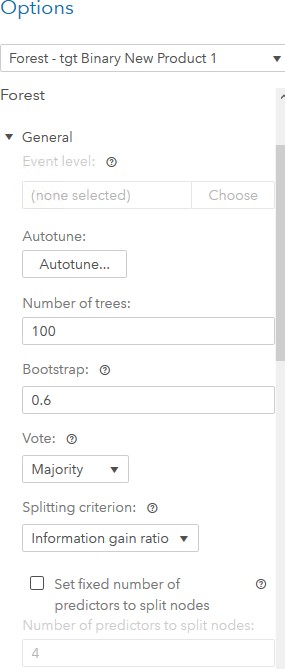
This demonstration illustrates the training and exploration of a forest model in SAS Visual Data Mining and Machine Learning.

1. Select the new **Forest** on the canvas of new page to make it the active object.
2. For convenience, rename the newly created page to **Forest**.
3. Disable the **auto-refresh** option, if not done already.

The settings for the Data Roles pane are reproduced below for your convenience.

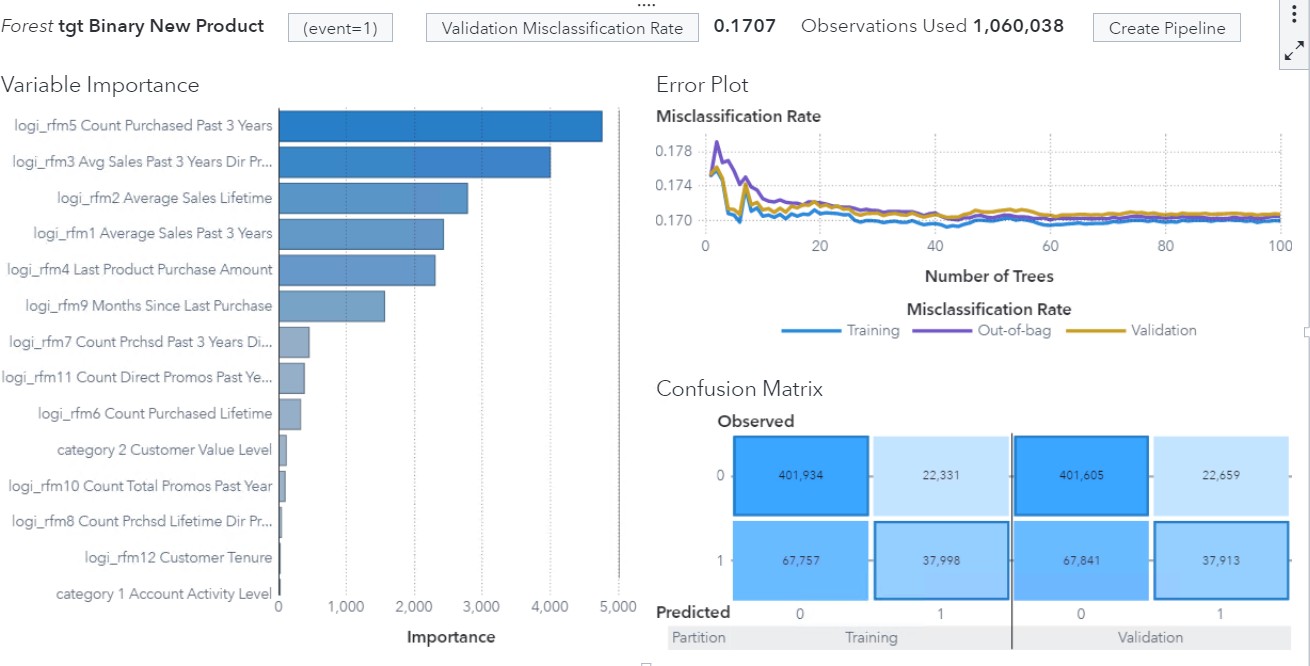


A quick review of the General properties under the Forest options reveals that the default forest has 100 trees. A sample of 60% of the data rows is randomly drawn to grow each tree, splitting is done using information gain ratio criteria, and predictions are created from the tree ensemble via a majority vote.

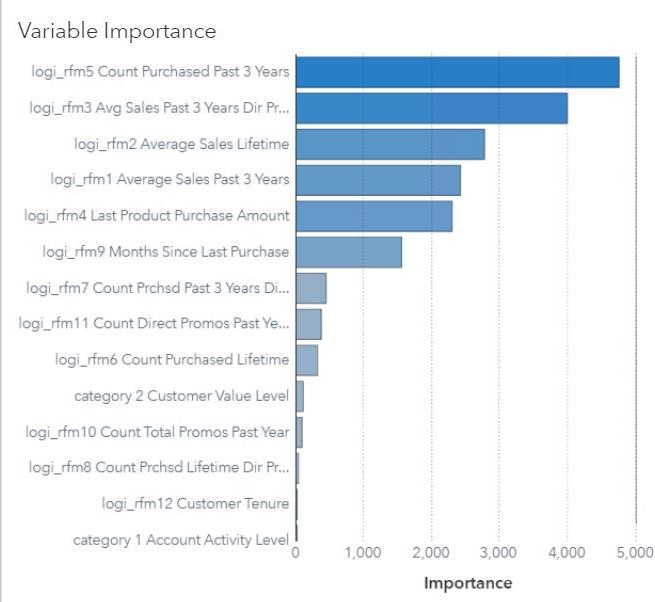


1. Click the **Refresh** button to create a default forest model.
2. The summary bar in the results window displays general model information. The model

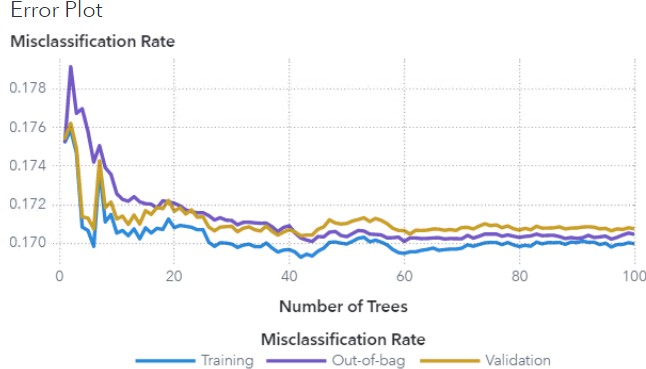
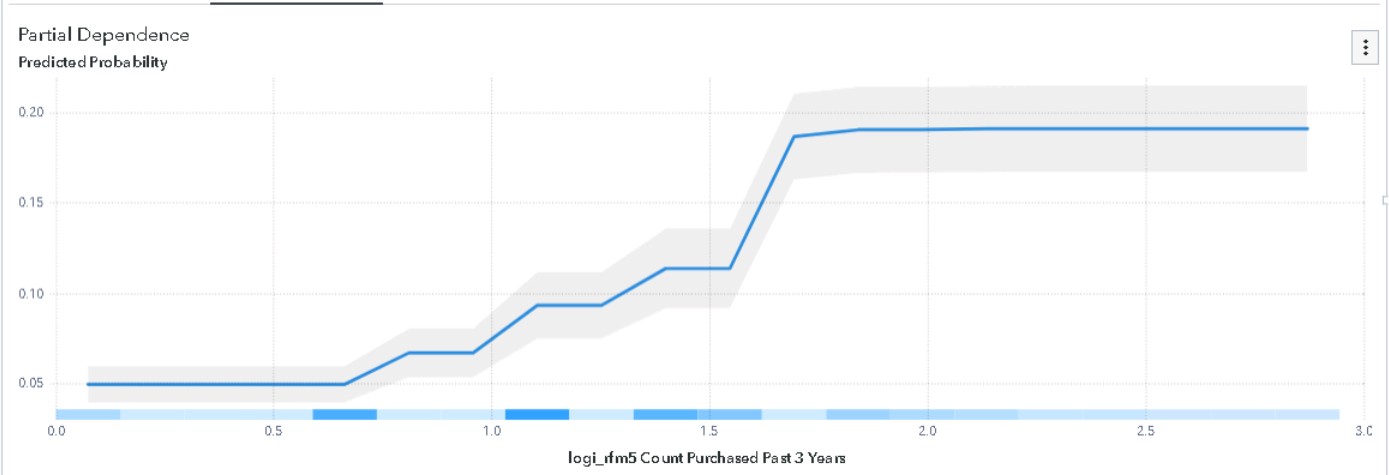
is trained for level **event=1** of the target variable. By default, validation KS (Youden) statistics are displayed. Click **Validation KS (Youden)** and select **Validation Misclassification Rate**.



The variable importance plot lists the inputs that contribute most to the classification model. The variable **logi\_rfm5 Count Purchased Past 3 Years** appears to be the most important predictor for this model.



The error plot displays the misclassification rate calculated on training, validation, and out-of-bag samples across the number of tress.

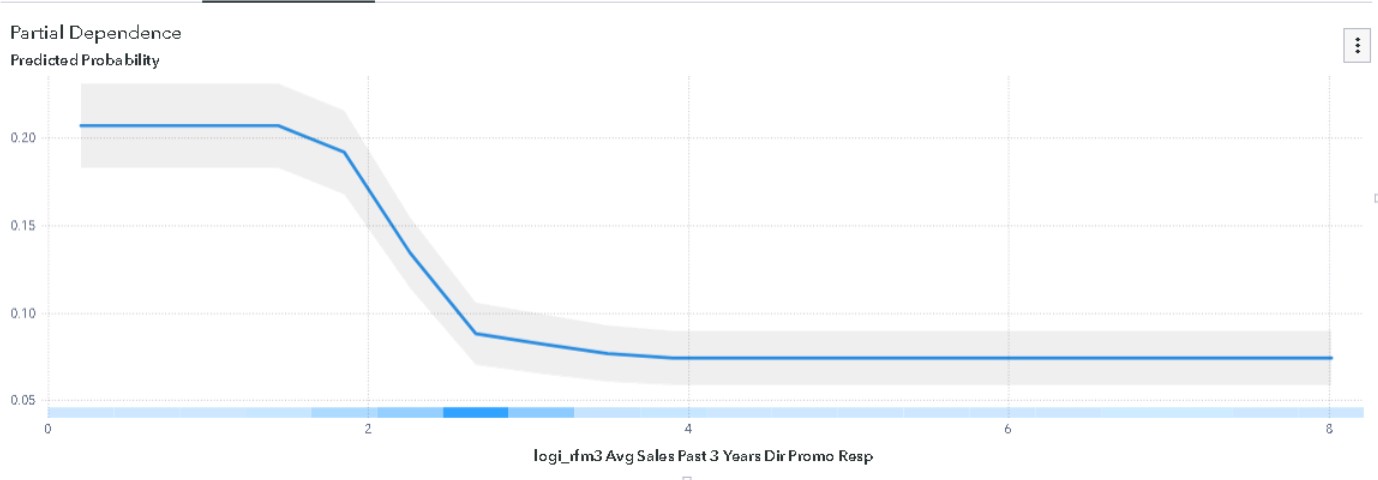


**Note:** Some SAS Visual Data Mining and Machine Learning models are created with a nondeterministic process. This means that you might experience different results when opening the report again later

1. Right-click on the Error Plot window and select **Partial dependence**. Click **Refresh.**

You get a PD plot that displays the relationship between the target and the predictor that had the highest variable importance value (that is, logi\_rfm5 Count Purchased Past 3 Years). The plot indicates that the predicted probability increases (though not linearly) as the log transformed number of products purchased in past three years increases from 0.6 to 1.7 (approximately). The values outside of this range have little additional impact on the predicted probability.

1. Now right-click on the current PD plot and select the variable **logi\_rfm3 Avg Sales Past 3 Years Dir Promo Resp** from the list of available inputs. Click **Refresh**.



The PD plot between target and transformed RFM3 variable suggests that the probability of purchasing a new product decreases (though not linearly) as the value of the transformed average sales amount in past three years in response to a direct campaign variable increases from an initial nonzero value to 2.6 and becomes constant for any further increase in the predictors value.

**Exploration of the Forest Model**

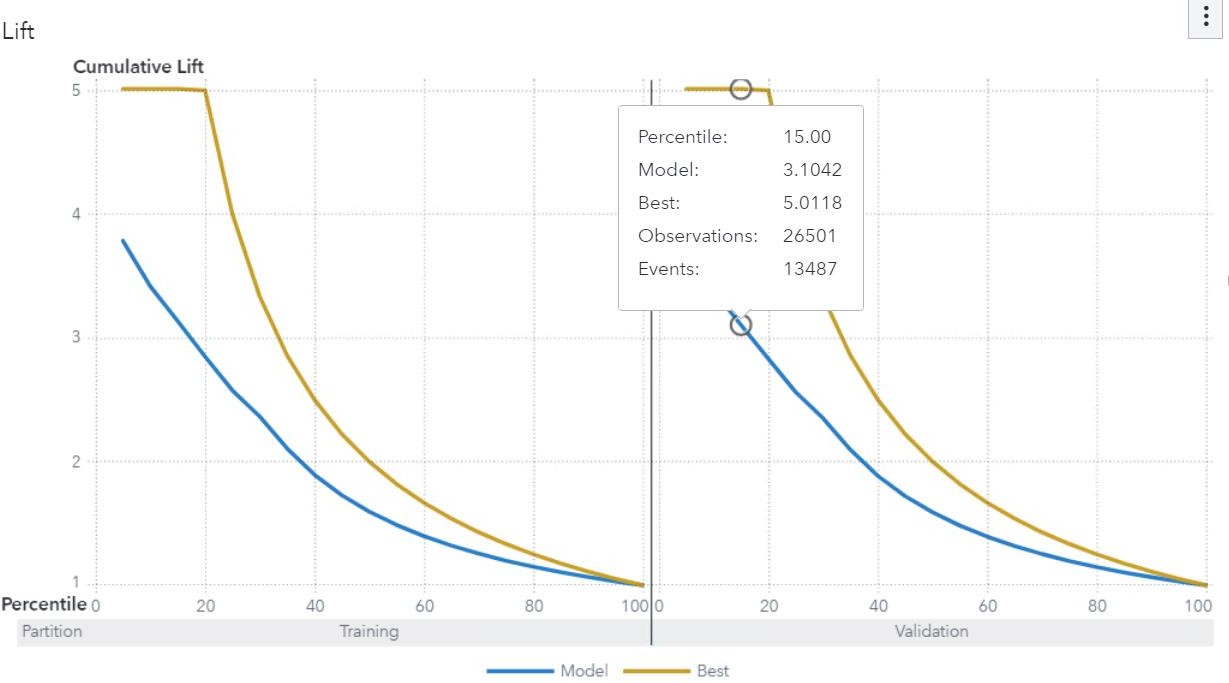
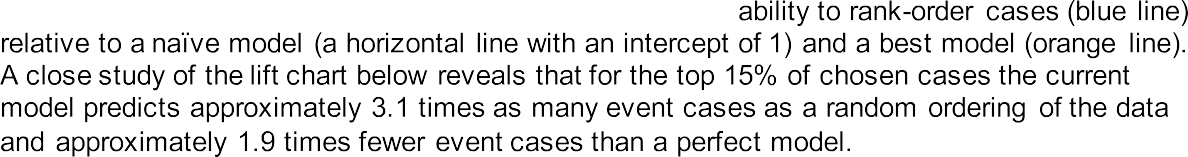
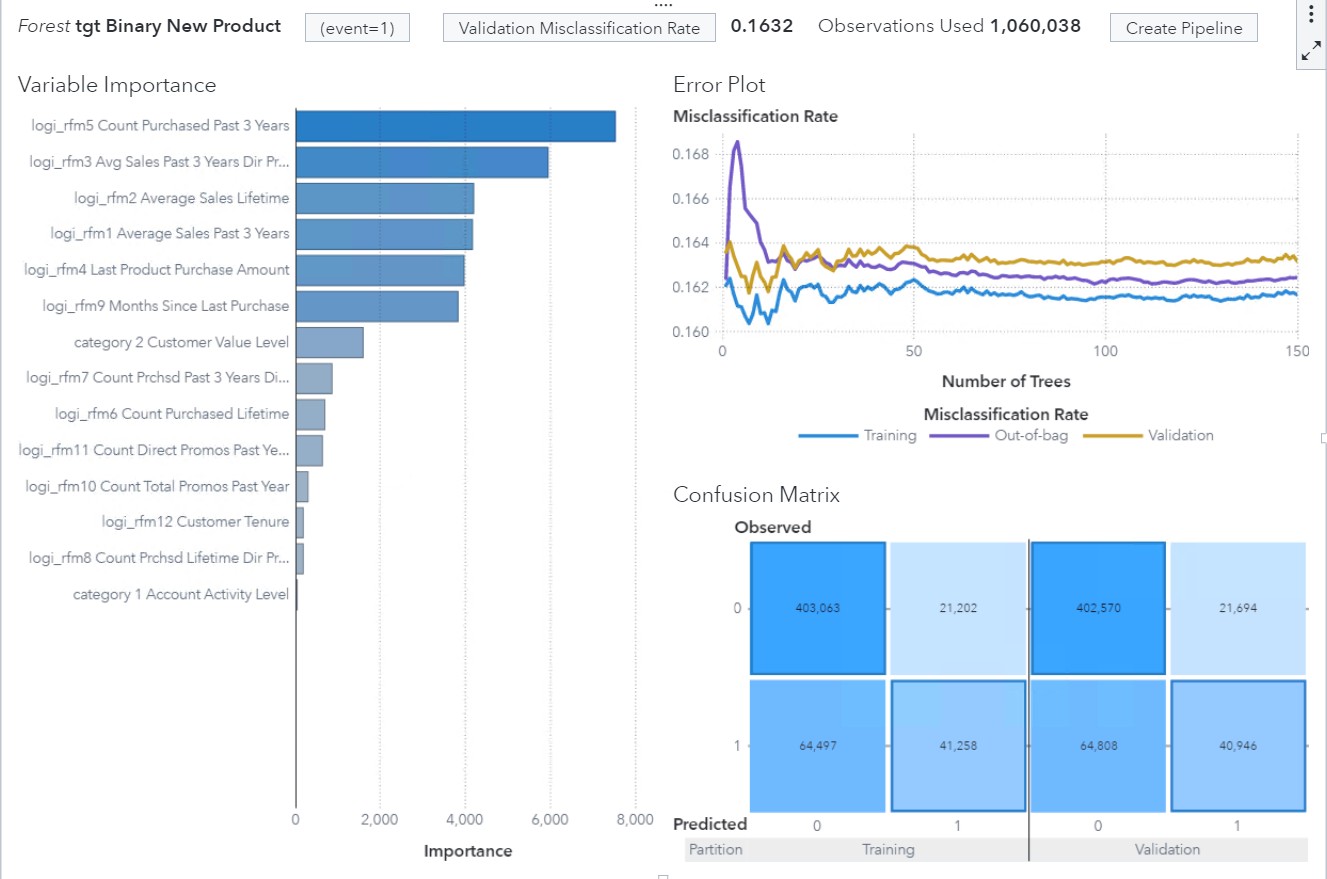
1. In the right pane, click **Options**. Under Forest, change the **Number of Trees** option to **150** and the **Bootstrap** option to **0.8**. Under Decision Tree, change **Maximum levels** to **12** and change **Leaf size** to **20**.

**Note:** Generally, increasing the number of trees tend to improve the performance of the forest.

However, the improvement decreases as the number of trees increases. At the same time, more trees require more computation and beyond a certain point the tradeoff might not be worth it.

1. Click **Refresh** to update the model.

The updated performance of the model can be seen at the top of the workspace. The final validation misclassification rate of the new model is approximately 0.1632. The validation performance of the forest is shown below. Model comparison is explored in the next lesson.

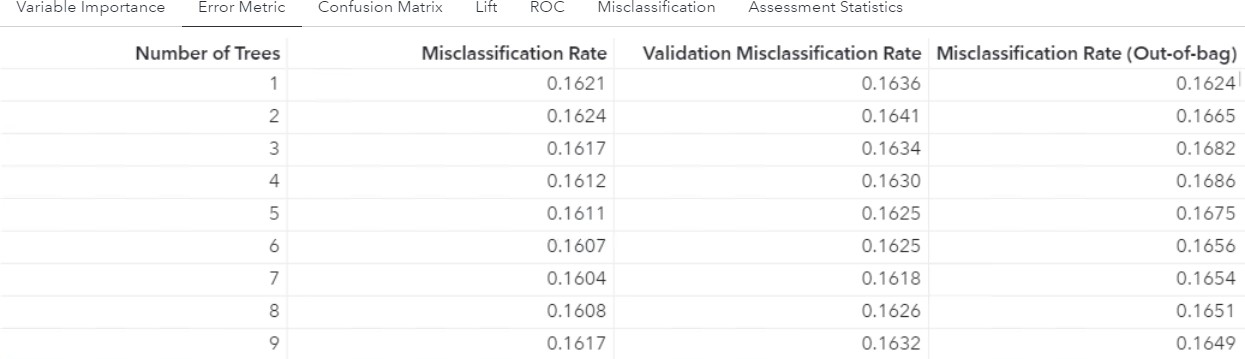


Similarly, the ROC curve, Confusion matrix and Misclassification plot can be examined to assess the performance of the model.

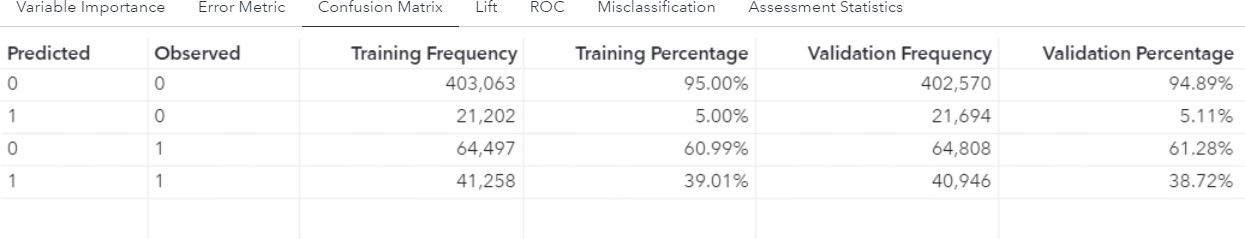
1. Click **Maximize** on the top right of your page for additional details about the fitted model.



Detail information includes Variable Importance diagnostics and Error Metric information. The Error Metric reports the misclassification rates, based on out-of-bag and validation data, as the model iterates from minimum to maximum complexity (from 1 to 150 trees). Lift and ROC information provide tables that can be used to create Lift and ROC plots based on training and validation data. The Assessment statistics tab provides values of assessment statistics calculated for the model on training as well as validation data.



Confusion matrix details table provides the number of correct and incorrect cases classified by the model on training and validation dat sets.



1. Save this report as Demo\_08\_Floresta in the My Folder location.

