MODUL 13
SAS® VDMML: DATA PARTITION & NEURAL NETWORK MODELS

THEME DESCRIPTION

Students understand and are able to implement the SAS Visual Data Mining and Machine Learning using neural Network algorithm to provides procedures for sampling, data exploration, dimension reduction, and model assessment.

WEEKLY LEARNING OUTCOMES (SUB-LESSONS)

CLO-1-Sub-CLO-13:
Understand how to implement machine learning and its applications, build data partition and provide honest assessment using SAS VDMML with Neural Network Algorithm– C5.

Through the following learning steps:

1. Partitioning Data
2. Build Neural Network model
3. Honest Assessment and Exploration of the Neural Network Model

PRACTICUM SUPPORT

a. Windows Operating System
b. (any) Browser Application

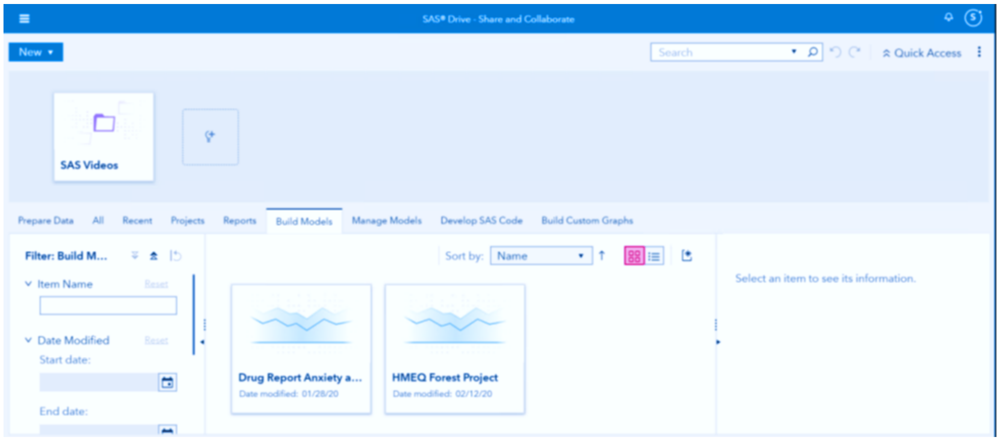
PRACTICUM STEPS

1. Partitioning data

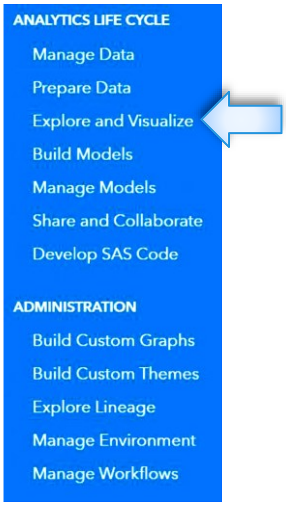
- In predictive modeling, the standard strategy for honest assessment of model performance is data splitting. A portion is used for fitting the model.
- That portion is the training data set.
- The remaining data are separated for empirical validation. The validation data set is used for monitoring and tuning the model to improve its generalization.
- The tuning process usually involves selecting among models of different types and complexities. The tuning process optimizes the selected model on the validation data.

a. Start, https://welcome.oda.sas.com

b. Open SAS®Studio, then open new tab for SAS® Drive to go to SAS Drive page.



Click ≡ (Show list of applications) the upper left corner of the SAS Drive page. **Select Explore and Visualize**.



ANALYTICS LIFE CYCLE

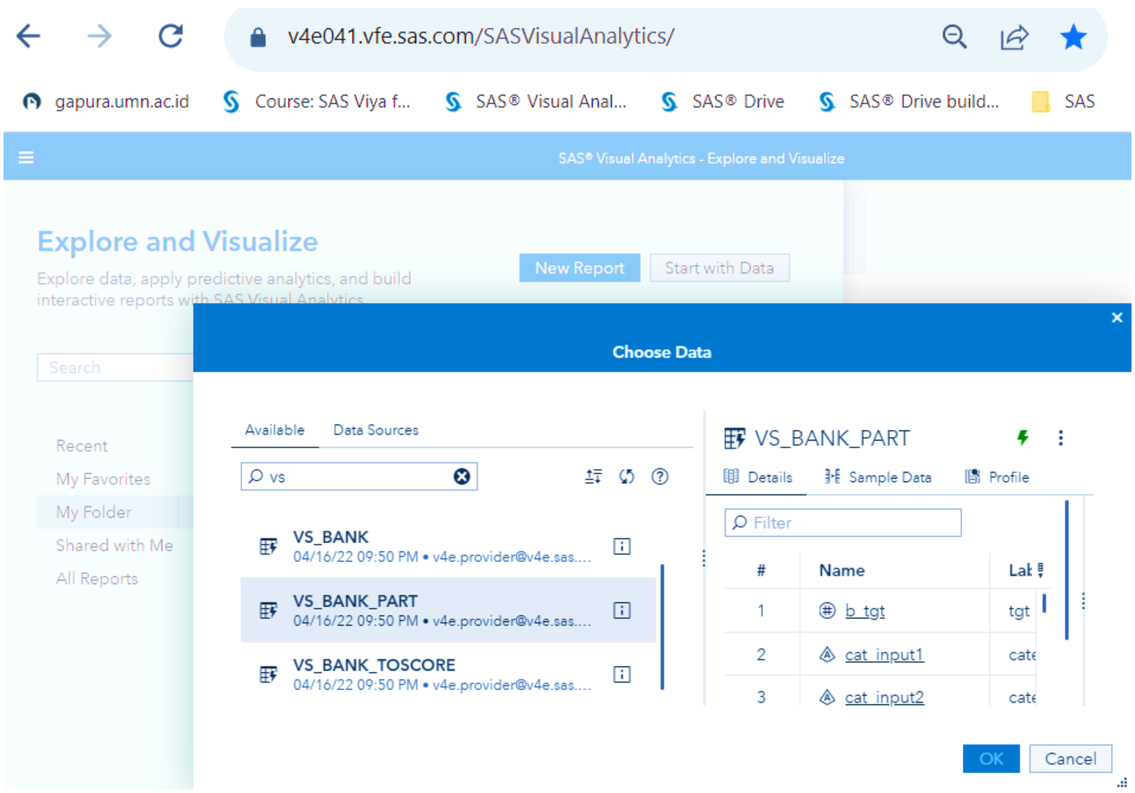
- Manage Data
- Prepare Data
- Explore and Visualize
- Build Models
- Manage Models
- Share and Collaborate
- Develop SAS Code

ADMINISTRATION

- Build Custom Graphs
- Build Custom Themes
- Explore Lineage
- Manage Environment
- Manage Workflows

c. Click Start with Data.

d. In the Choose Data window, choose **vs_bank_part** table.



v4e041.vfe.sas.com/SASVisualAnalytics/

gapura.umn.ac.id | Course: SAS Viya f... | SAS® Visual Anal... | SAS® Drive | SAS® Drive build... | SAS

SAS® Visual Analytics - Explore and Visualize

Explore and Visualize

Explore data, apply predictive analytics, and build interactive reports with SAS Visual Analytics

New Report | Start with Data

Search

Recent
My Favorites
My Folder
Shared with Me
All Reports

Choose Data

Available | Data Sources

vs

VS_BANK
04/16/22 09:50 PM • v4e.provider@v4e.sas....

VS_BANK_PART
04/16/22 09:50 PM • v4e.provider@v4e.sas....

VS_BANK_TOSCORE
04/16/22 09:50 PM • v4e.provider@v4e.sas....

VS_BANK_PART

Details | Sample Data | Profile

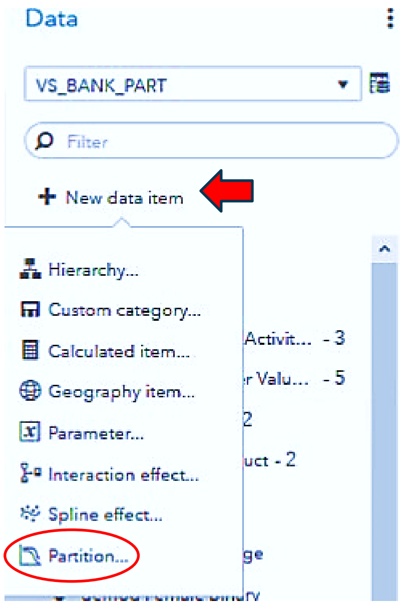
Filter

#	Name	Lal
1	b_tgt	tgt
2	cat_input1	cate
3	cat_input2	cate

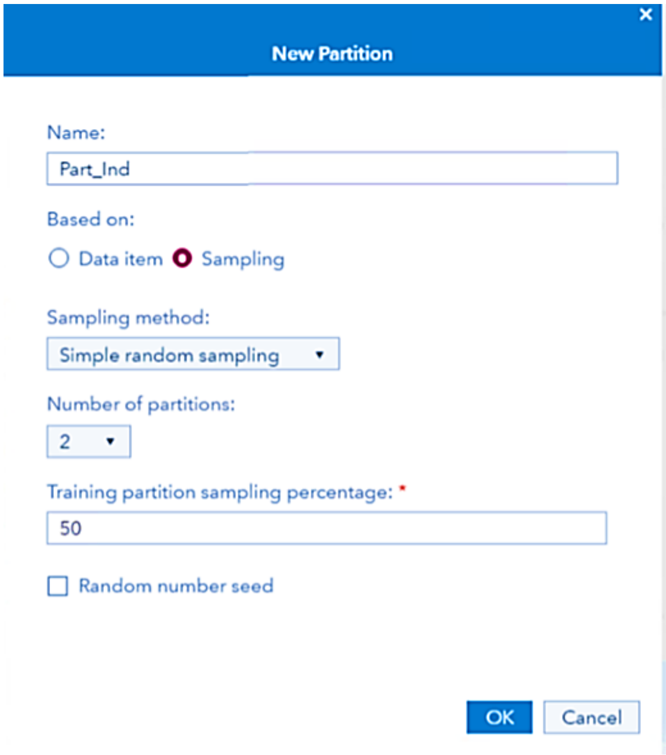
OK | Cancel

- The Data pane enables you to work with data sources, create new data items (hierarchy, calculated item, custom category, interaction effect, spline effect, partition), add a data source filter, and view and modify properties for data items.

e. In the Data pane, click New data item and select Partition



f. In the New Partition window, enter Part_Ind in the Name field and 50 in the Training partition sampling percentage field. Your New Partition window properties should appear as follows:



New Partition

Name:
Part_Ind

Based on:
○ Data item ● Sampling

Sampling method:
Simple random sampling ▾

Number of partitions:
2 ▾

Training partition sampling percentage: *
50

□ Random number seed

- The Number of partitions option indirectly specifies whether a test partition is included. When you select 3 for this option, a training partition, validation partition, and testing partition are created.
- When you select 2, only a training partition and validation partition are created. At least 1% of the data must be used for the validation data.
- Therefore, the sum of the training partition and testing partition must be strictly less than 100.

g. Click OK. The Partition variable is added to the Category list.

h. In the upper left corner of the report, click + (New page) next to Page as "Partition Data".

i. In the left pane, click the Data icon. Drag Part_Ind to the canvas
  - The variable Part_Ind contains values 0 and 1 for each partition created.
  - The bar chart shown as Output figure A exhibits that the two partitions have nearly equal frequency.

j. Save your report. Click (Menu) and select Save as "**IS429 Lab Week#13 Exploring a NN model NIM-Name**"

k. Screenshotted your report as shown on Output figure A and save as "IS-429 Lab Week#13A Partition yourname.jpg"

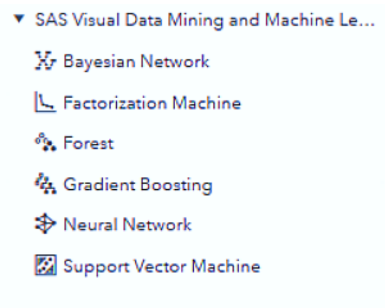
2. Build Neural Network model

- This demonstration illustrates the training and exploration of a neural network model in SAS Visual Data Mining and Machine Learning on the VS_BANK_PART data.
- Still using the report, you created earlier on "**IS429 Lab Week#13 Exploring a NN model NIM-Name**".

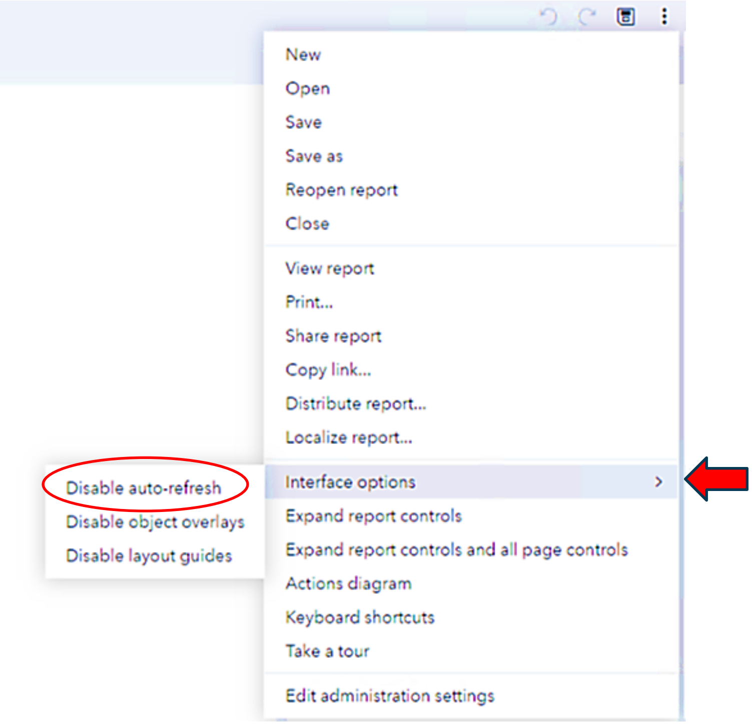
a. In the upper left corner of the report, click + (New page) next to Page as "Neural Network model".

b. Select Objects and scroll down to the SAS Visual Data Mining and Machine Learning objects.

c. Highlight Neural Network, and either double-click it or drag it into the main field.



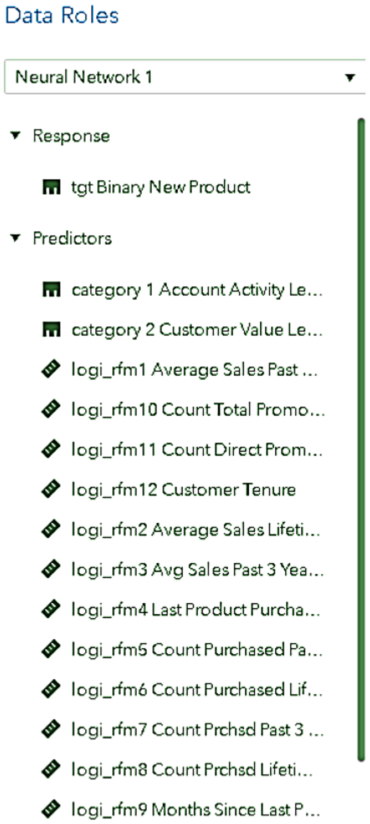
d. Disable the auto-refresh option by clicking Menu at the top right of your screen and selecting Interface options ⍰ Disable auto-refresh.



e. Select the Roles pane. Add tgt Binary New Product as the Response variable.

f. Add category 1 Account Activity Level, category 2 Customer Value Level and

g. the twelve variables that begin with logi_rfm as Predictors



Data Roles

Neural Network 1

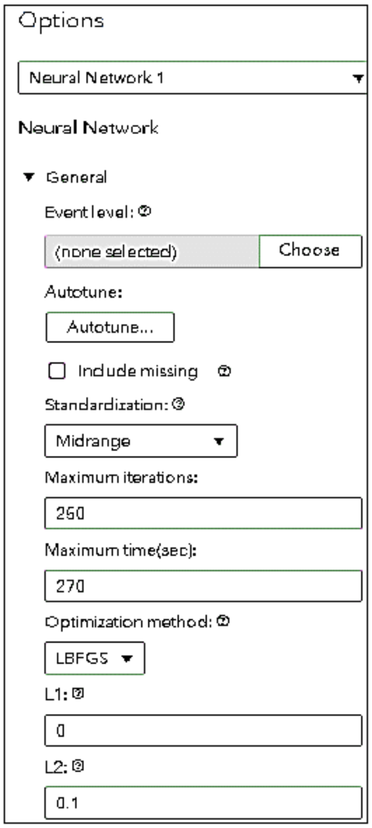
▼ Response

- tgt Binary New Product

▼ Predictors

- category 1 Account Activity Le...
- category 2 Customer Value Le...
- logi_rfm1 Average Sales Past ...
- logi_rfm10 Count Total Promo...
- logi_rfm11 Count Direct Prom...
- logi_rfm12 Customer Tenure
- logi_rfm2 Average Sales Lifeti...
- logi_rfm3 Avg Sales Past 3 Yea...
- logi_rfm4 Last Product Purcha...
- logi_rfm5 Count Purchased Pa...
- logi_rfm6 Count Purchased Lif...
- logi_rfm7 Count Prchsd Past 3 ...
- logi_rfm8 Count Prchsd Lifeti...
- logi_rfm9 Months Since Last P...

- Scrutiny of the neural network options reveals that the default architecture consists of one hidden layer and 10 hidden units (neurons). Input variable standardization is performed by default on interval (measure) valued inputs, and no direct connections are allowed between the input and target layers



Options

Neural Network 1

Neural Network

▼ General

Event level: (none selected) Choose

Autotune: Autotune...

□ Include missing

Standardization: Midrange

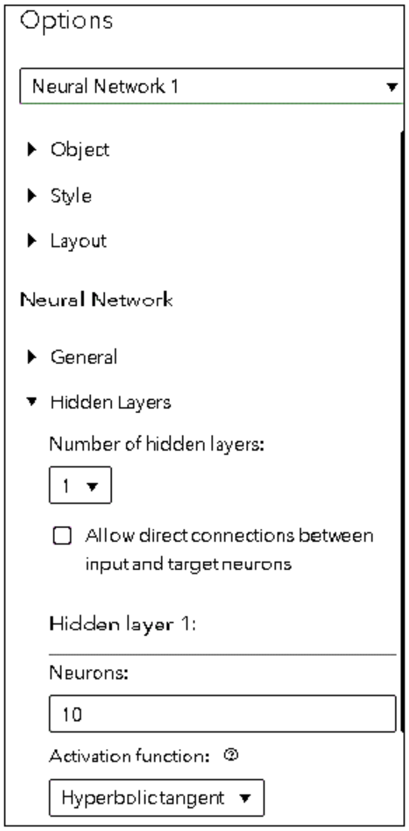
Maximum iterations: 250

Maximum time(sec): 270

Optimization method: LBFGS

L1: 0

L2: 0.1



Options

Neural Network 1

▶ Object

▶ Style

▶ Layout

Neural Network

▶ General

▼ Hidden Layers

Number of hidden layers: 1

□ Allow direct connections between input and target neurons

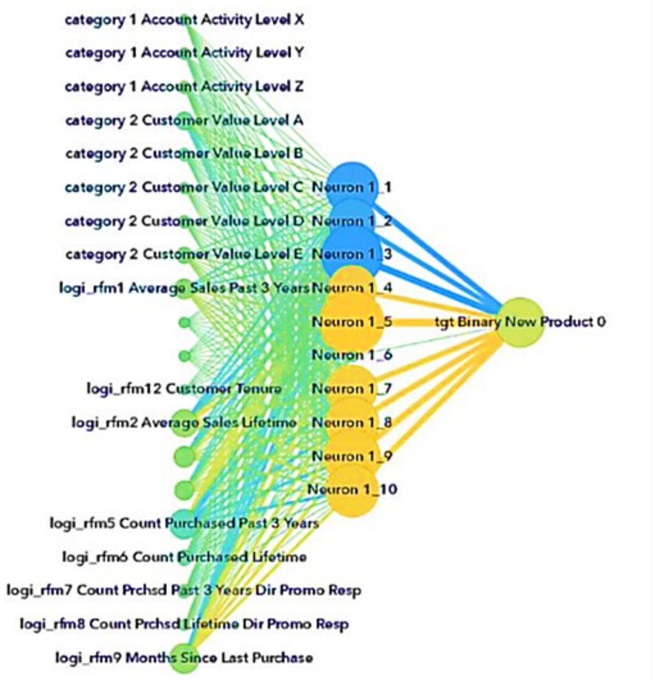
Hidden layer 1:

Neurons: 10

Activation function: Hyperbolic tangent

h. In the options pane under the Model Display section, select the Network Diagram property and select the box corresponding to the Neuron labels property.
i. Click (Refresh) to fit the neural network model.

- You can investigate assessment statistics from the summary pane. By default, the KS (Youden) value is displayed. Click KS (Youden) and select Misclassification Rate.
- The model uses 1,060,038 observations, and the model is built to predict event=1.
- The Network diagram illustrates the default architecture and consists of an input layer, one hidden layer with 10 neurons and an output layer.
- The size of each neuron is a function of the magnitude of the estimated weights (estimated parameter values) contained in it.


category 1 Account Activity Level X
category 1 Account Activity Level Y
category 1 Account Activity Level Z
category 2 Customer Value Level A
category 2 Customer Value Level B
category 2 Customer Value Level C
category 2 Customer Value Level D
category 2 Customer Value Level E
logi_rfm1 Average Sales Past 3 Years
logi_rfm12 Customer Tenure
logi_rfm2 Average Sales Lifetime
logi_rfm5 Count Purchased Past 3 Years
logi_rfm6 Count Purchased Lifetime
logi_rfm7 Count Prchsd Past 3 Years Dir Promo Resp
logi_rfm8 Count Prchsd Lifetime Dir Promo Resp
logi_rfm9 Months Since Last Purchase
Neuron 1_1
Neuron 1_2
Neuron 1_3
Neuron 1_4
Neuron 1_5
Neuron 1_6
Neuron 1_7
Neuron 1_8
Neuron 1_9
Neuron 1_10
tgt Binary New Product 0

- You can tune the model further by changing the number of hidden layers, the number of neurons in each hidden layer, activation function for each layer, or other options.
- Screenshotted your Neural Network model as shown on Output figure B and save it into "IS-429 Lab Week#13B NN model yourName.jpg"

3. Honest Assessment and Exploration of the Neural Network Model

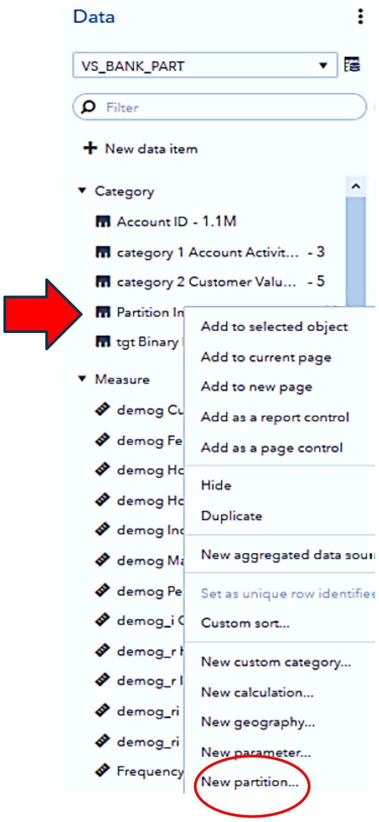
- To perform honest assessment, you select the model that performs best on the validation data set. In the previous section, you created a partition indicator variable that can be used here to monitor the performance of model using validation data.
- However, it is noted that each time a report is generated within visual analytics, the partition is re-created. Therefore, it is suggested to add the partition indicator variable to the data table prior to pulling it into visual analytics.
- This would bypass the partition variable being re-created at the report level. The VS_BANK_PART data set also has a partition indicator variable already created just to prevent the recreation of partition variable at report level. In this course, by now you have two partition indicator variables.
- Therefore, for convenience, you might want to hide the Part_Ind variable created in a previous demonstration and use the Partition Indicator variable (which already exists in the data set) throughout.

a. To hide the variable, click the Data tab and locate the Part_Ind variable.

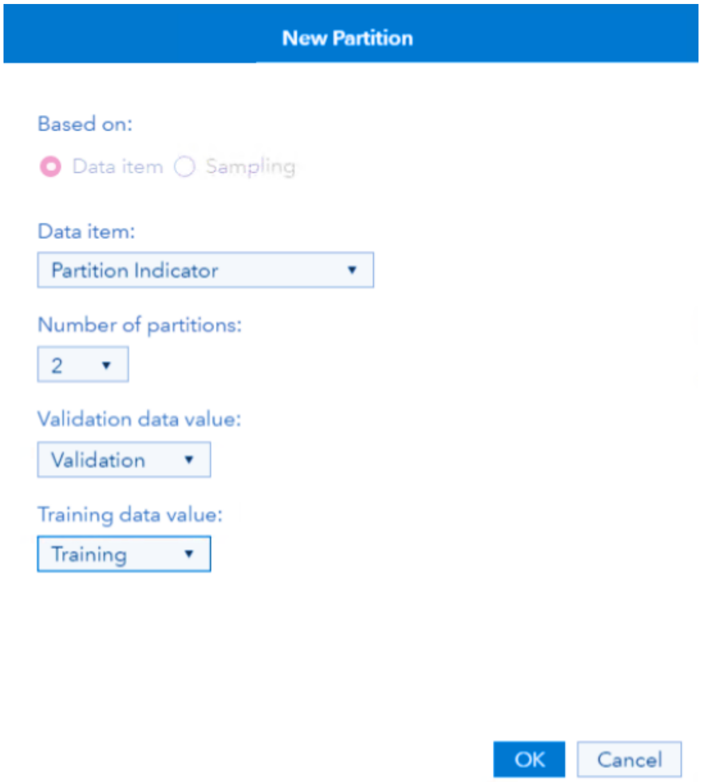
b. Right-click Part_Ind variable and select Hide

c. To perform honest assessment, click the Data tab and then find the partition indicator variable under Category.

d. Right-click Partition Indicator and select New partition



e. The Partition Indicator variable was created such that the validation data values are flagged as Validation and training data values are flagged as Training. Ensure that the New Partition window has settings as shown below:


New Partition

Based on: ● Data item ○ Sampling

Data item: Partition Indicator

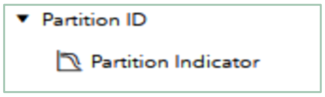
Number of partitions: 2

Validation data value: Validation

Training data value: Training

f. Click OK to close New Partition window.

g. When the partition indicator variable is set as the partition column, it is ready to use for honest assessment. Click the Roles tab and assign Partition Indicator under Partition ID.



▾ Partition ID

Partition Indicator

h. Click Refresh to update the model.

i. Results are updated to reflect the change in the underlying training sample. Note that the validation misclassification rate is 0.1698.

j. Click the Options tab at the right of your screen to explore the neural network model's properties and options.

k. Click the Hidden Layers property and change Number of hidden layers to 2.

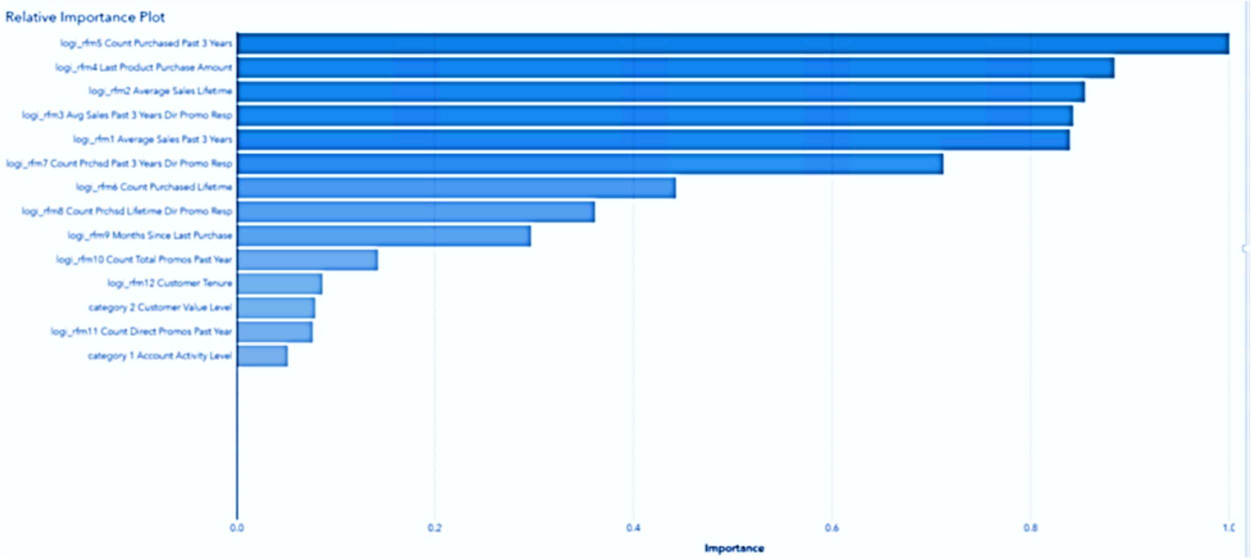
l. Under the Hidden layer 1 and Hidden layer 2 properties, change the Neurons value from the default value of 10 to 5.

m. Click the Refresh button to update the model.

n. The updated model produces a validation misclassification rate of 0.1655, which indicates a slight improvement over the previous result.

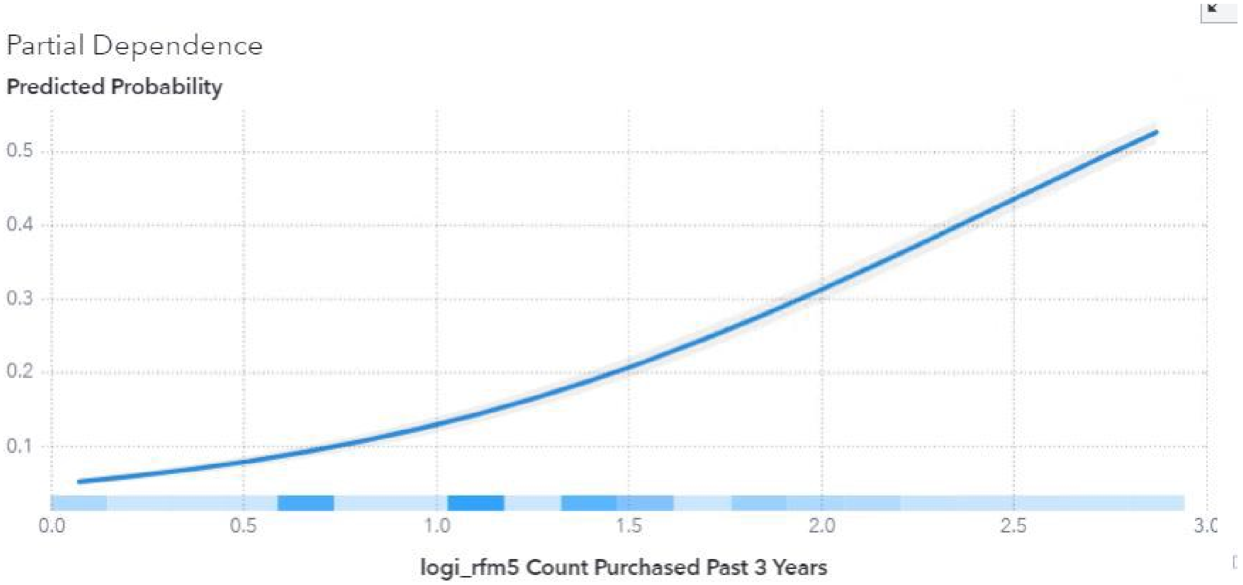
o. Click the Options pane. In the Model Display section, select General and change Plot layout to Stack.

p. Click the Iteration plot tab to see how the objective function changes at each iteration as the network is grown. To switch to a relative importance plot, right-click the Iteration Plot window and select Relative Importance plot from list of choices:



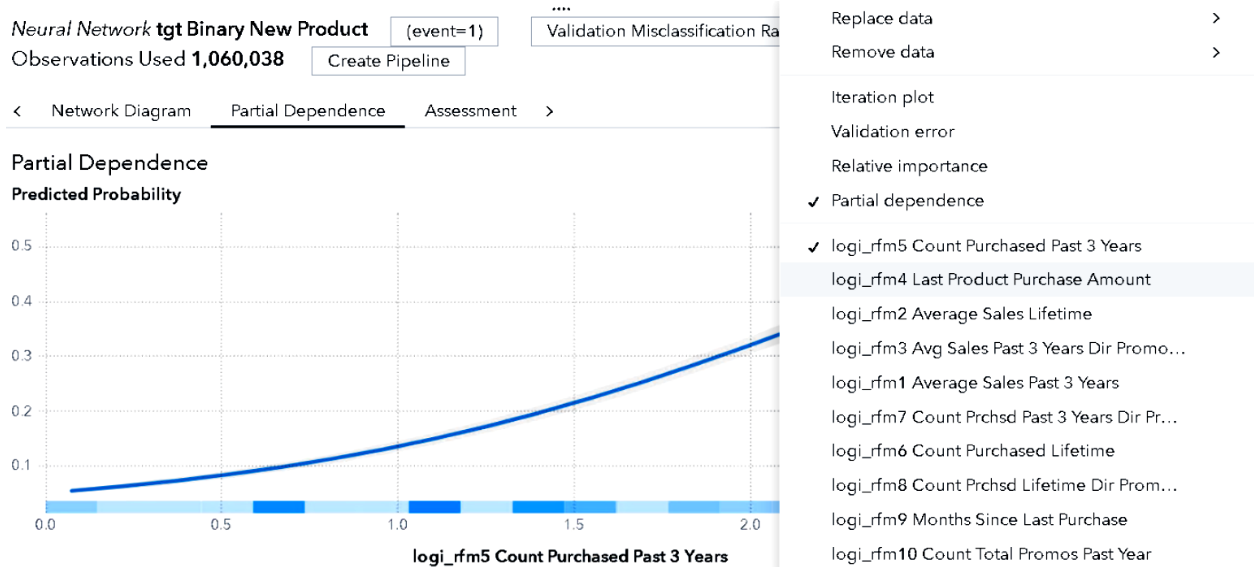
- The variables in the Relative Importance plot are ranked using their first-split log worth in a decision tree when applied to the scored training data.
- When validation data is available you can also switch to validation error plot, which displays the change in validation error at each iteration of the training process.

q. Right-click on the Relative Importance Plot window, select Partial dependence, and click Refresh:

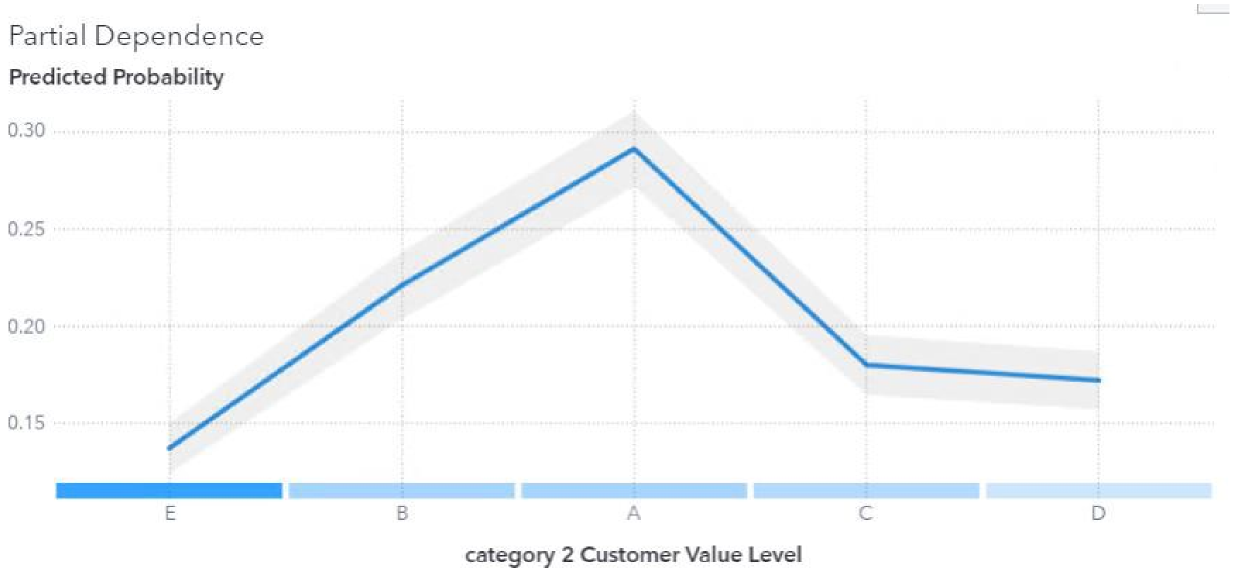


- The PD plot can show if the relationship between the target and a predictor is linear, monotonic, or more complex. According to the model that was fit previously, the above plot shows the relationship between the transformed version of number of products purchased in past three years and the model's prediction.
- Here, the model's predicted probability tends to increase when the log transformed number of products purchased in past three years increases, although this trend is not linear.

r. In order to understand the relationship between the model's predictions and other inputs, right-click on the current PD plot and select a variable from the list of available inputs:


Neural Network tgt Binary New Product
(event=1)
Validation Misclassification Ra
Observations Used 1,060,038
Create Pipeline
Network Diagram
Partial Dependence
Assessment
Partial Dependence
Predicted Probability
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
1.5
2.0
logi_rfm5 Count Purchased Past 3 Years
Replace data
Remove data
Iteration plot
Validation error
Relative importance
✓ Partial dependence
✓ logi_rfm5 Count Purchased Past 3 Years
logi_rfm4 Last Product Purchase Amount
logi_rfm2 Average Sales Lifetime
logi_rfm3 Avg Sales Past 3 Years Dir Promo…
logi_rfm1 Average Sales Past 3 Years
logi_rfm7 Count Prchsd Past 3 Years Dir Pr…
logi_rfm6 Count Purchased Lifetime
logi_rfm8 Count Prchsd Lifetime Dir Prom…
logi_rfm9 Months Since Last Purchase
logi_rfm10 Count Total Promos Past Year

s. Select the category2 Customer Value Level variable and click Refresh.



- As you might expect, the predicted probability of purchasing a product is minimum for the customers who are least profitable and creditworthy (that is, customers that belong to level E), whereas it is highest for the customers who are most profitable and creditworthy (that is, level A customers).
- This can be an important insight for the business. Screenshotted your Partial Dependence as shown on Output figure C and save it into "IS-429 Lab Week#13C PD yourName.jpg"

t. Click the Assessment tab. By default, it displays a confusion matrix as shown on Output figure B, which helps analyze the performance of a classifier. Below output summarizes the counts of true positives, false positives, true negatives and false negatives on training and validation data. The higher the values along the diagonal the better.

u. The 2nd row and 2nd column cell for the validation data gives the number of true positive cases, which means around 36799 cases are correctly predicted as 1 (event) when they are known to be 1 (event). Similarly, 1st row and 2nd column cell value on the validation data shows the count of false positive cases, that means approximately 18775 cases are misclassified as events (1), when actually they are known non-events (0).

v. Screenshotted your confusion matrix as shown on Output figure D and save it into "IS-429 Lab Week#13D Confuse yourName.jpg"

w. Right-click on the confusion matrix plot and select Lift chart. The lift plot summarizes the 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 15% (percentile) of the data has about 3.09 times as many responders as a random ordering of the data, and 1.91 times fewer responders than a perfect ordering of the training data.
- Screenshotted your lift chart as shown on Output figure E and save it into "IS-429 Lab Week#13E Lift yourName.jpg"

x. Right-click in the center of cumulative lift chart and select ROC to switch to a ROC curve. The ROC curve 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 dashed line corresponds to an optimal threshold in the data.

- Screenshotted your lift chart as shown on Output figure F and save it into "IS-429 Lab Week#13F ROC yourName.jpg"

y. Place your cursor at the top of the Max Separation line where it intersects the ROC curve. A tooltip reveals an optimal cutoff value of 0.20.

z. Again, right-click the ROC curve to switch to Misclassification plot and the output shown as the Output figure G, screenshotted it and save into "IS-429 Lab Week#13G Assessment yourName .jpg".

- The true positive frequency (validation data) is 36,799 at the default prediction cutoff value of 0.50.
- Click the Options tab. In the Neural Network section, click the Assessment property to change the Prediction cutoff value from the default of 0.5 to 0.2.
- Click Refresh. Examine the Misclassification plot again, and you find that the number of true positives has more than doubled, to 83,429.

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

bb. The Model Information provides summary details about the model's architecture.

cc. There are 125 weights estimated in the chosen architecture for the 14 variables assigned as predictors

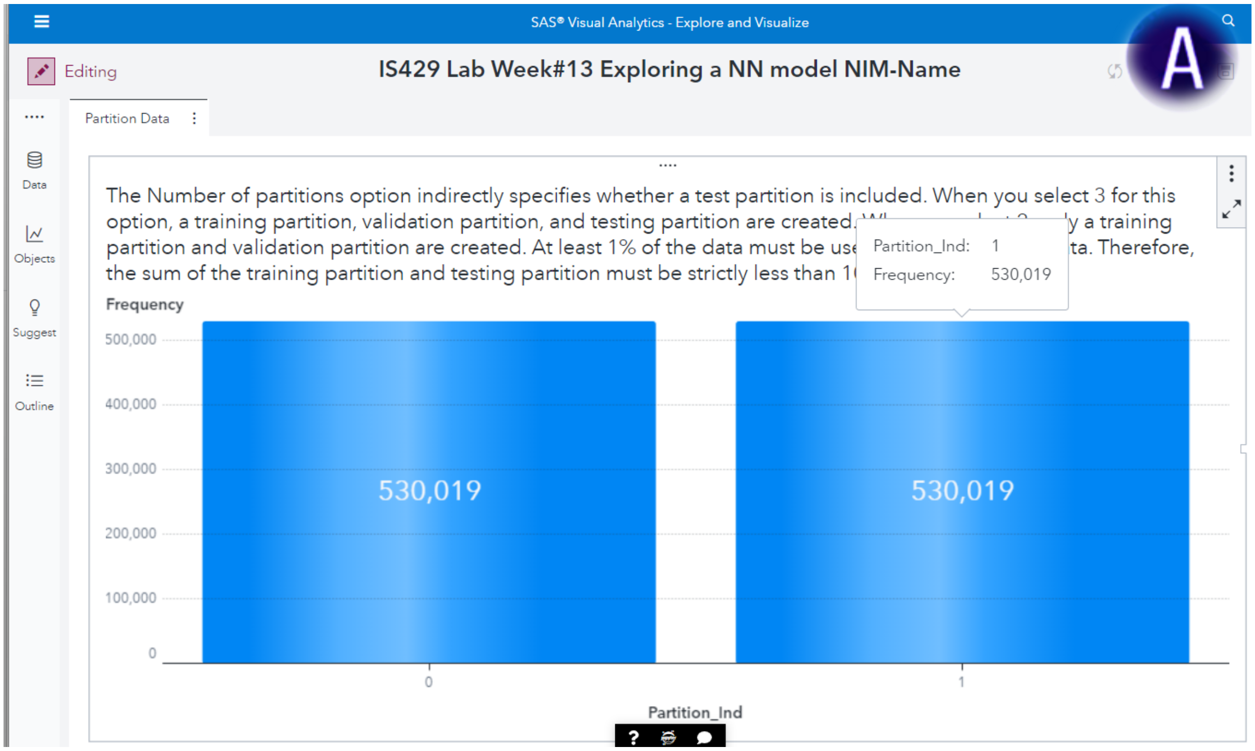
dd. Misclassification details give confusion matrix information about both the training and validation data sets. It clearly indicates that the model does fairly well on predicting training and validation cases.

➲ Finally, today's practicum is over, collect all the screenshots you produce into a word format file as "IS-429 Lab Week#13 Screenshots NIM.doc/docx"

➲ **Zipped your word file and SAS report, immediately submit immediately today to e-Learning IS-429 Practicum Week#13 with the naming format IS-429 VDMML NN model Week#13 NIM yourName.zip**.

OUTPUT/ RESULTS

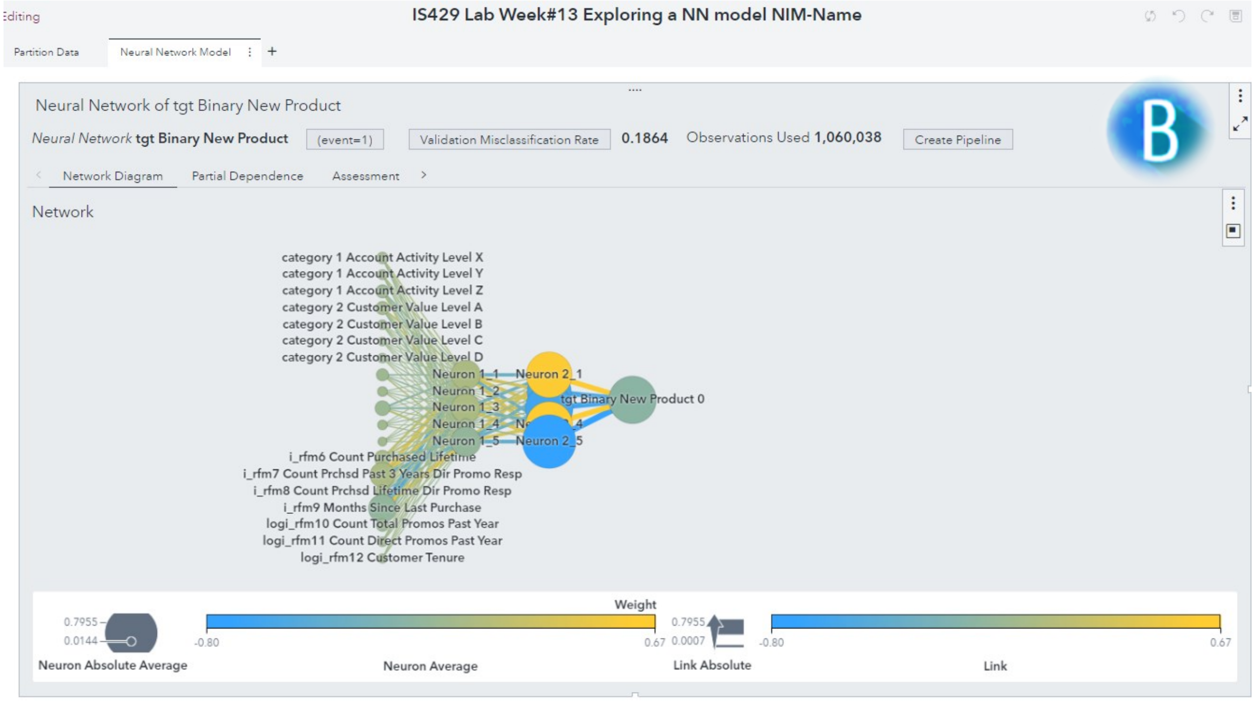
A. Partitioning Data


SAS® Visual Analytics - Explore and Visualize
Editing
IS429 Lab Week#13 Exploring a NN model NIM-Name
Partition Data
Data
Objects
Suggest
Outline
A

The Number of partitions option indirectly specifies whether a test partition is included. When you select 3 for this option, a training partition, validation partition, and testing partition are created. Wh y a training partition and validation partition are created. At least 1% of the data must be use ta. Therefore, the sum of the training partition and testing partition must be strictly less than 1(

Partition_Ind: 1
Frequency: 530,019

B. Neural Network model

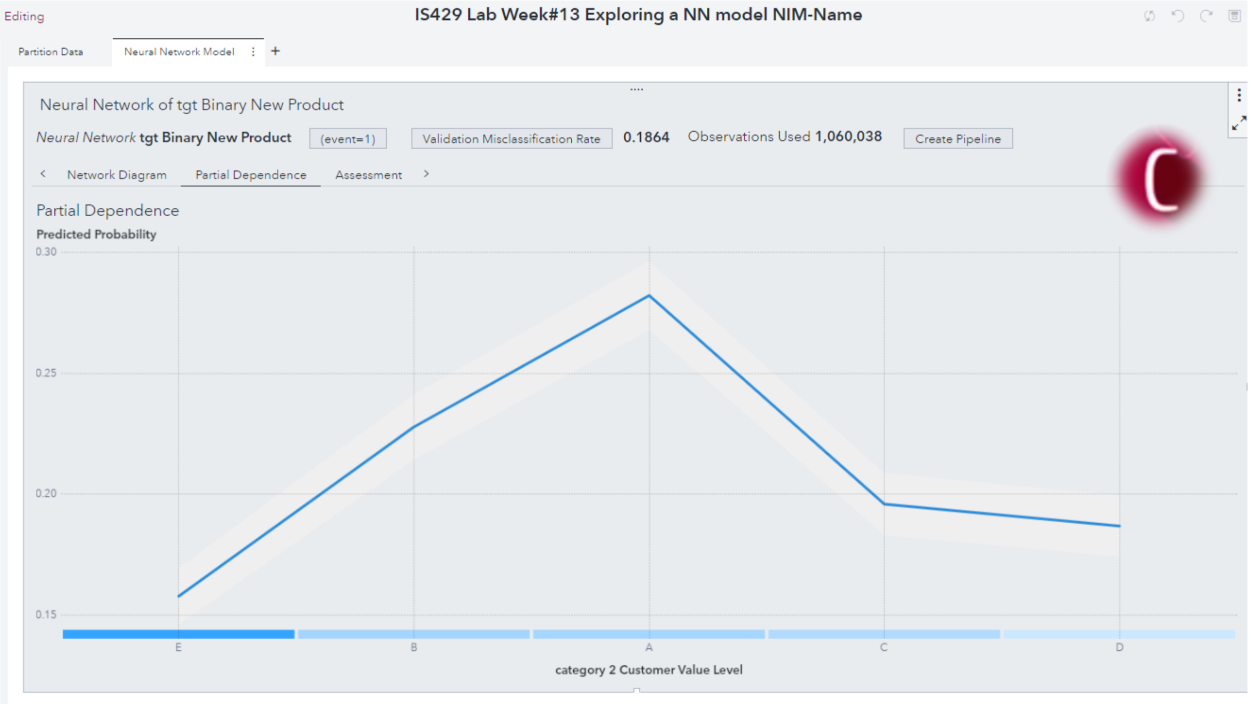

Editing
IS429 Lab Week#13 Exploring a NN model NIM-Name
Partition Data
Neural Network Model
B

Neural Network of tgt Binary New Product

Neural Network **tgt Binary New Product** (event=1) Validation Misclassification Rate **0.1864** Observations Used **1,060,038** Create Pipeline

Network Diagram
Partial Dependence
Assessment
Network
category 1 Account Activity Level X
category 1 Account Activity Level Y
category 1 Account Activity Level Z
category 2 Customer Value Level A
category 2 Customer Value Level B
category 2 Customer Value Level C
category 2 Customer Value Level D
Neuron 1_1
Neuron 1_2
Neuron 1_3
Neuron 1_4
Neuron 1_5
Neuron 2_1
Neuron 2_5
tgt Binary New Product 0
i_rfm6 Count Purchased Lifetime
i_rfm7 Count Prchsd Past 3 Years Dir Promo Resp
i_rfm8 Count Prchsd Lifetime Dir Promo Resp
i_rfm9 Months Since Last Purchase
logi_rfm10 Count Total Promos Past Year
logi_rfm11 Count Direct Promos Past Year
logi_rfm12 Customer Tenure
Weight
0.7955
0.0144
Neuron Absolute Average
-0.80
0.67
Neuron Average
0.7955
0.0007
Link Absolute
-0.80
0.67
Link

C. Partial Dependence



Editing

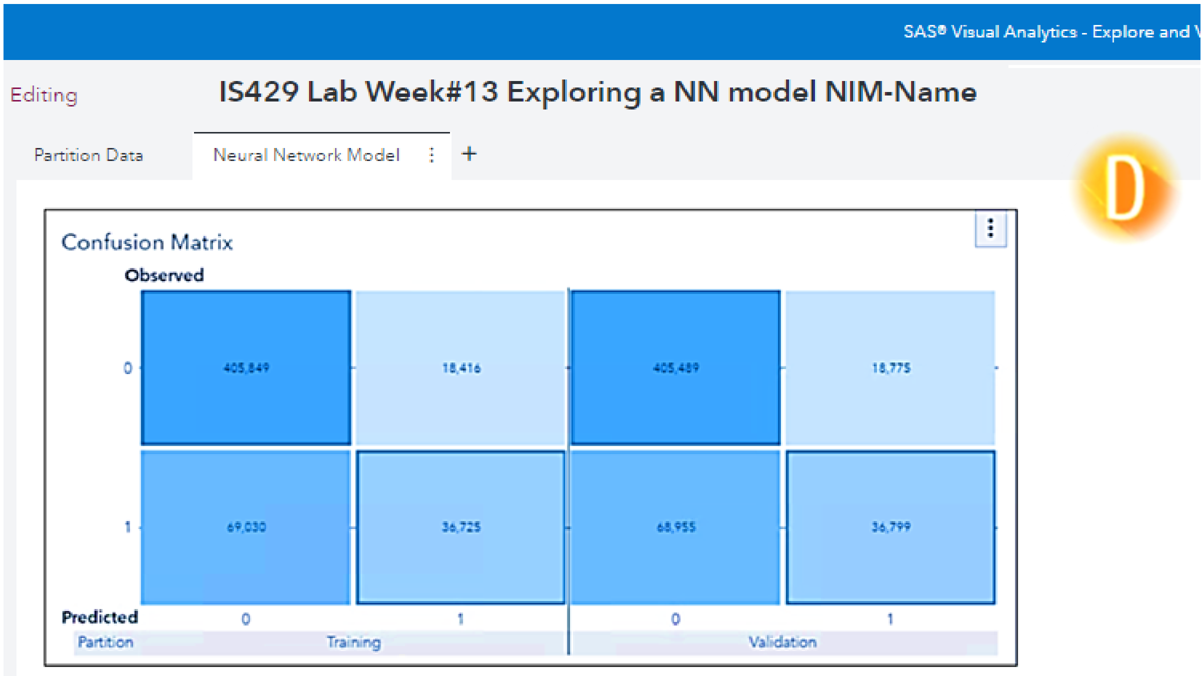
IS429 Lab Week#13 Exploring a NN model NIM-Name

Partition Data | Neural Network Model

Neural Network of tgt Binary New Product

Neural Network **tgt Binary New Product** (event=1) Validation Misclassification Rate **0.1864** Observations Used **1,060,038** Create Pipeline

D. Confusion Matrix



SAS® Visual Analytics - Explore and V

Editing

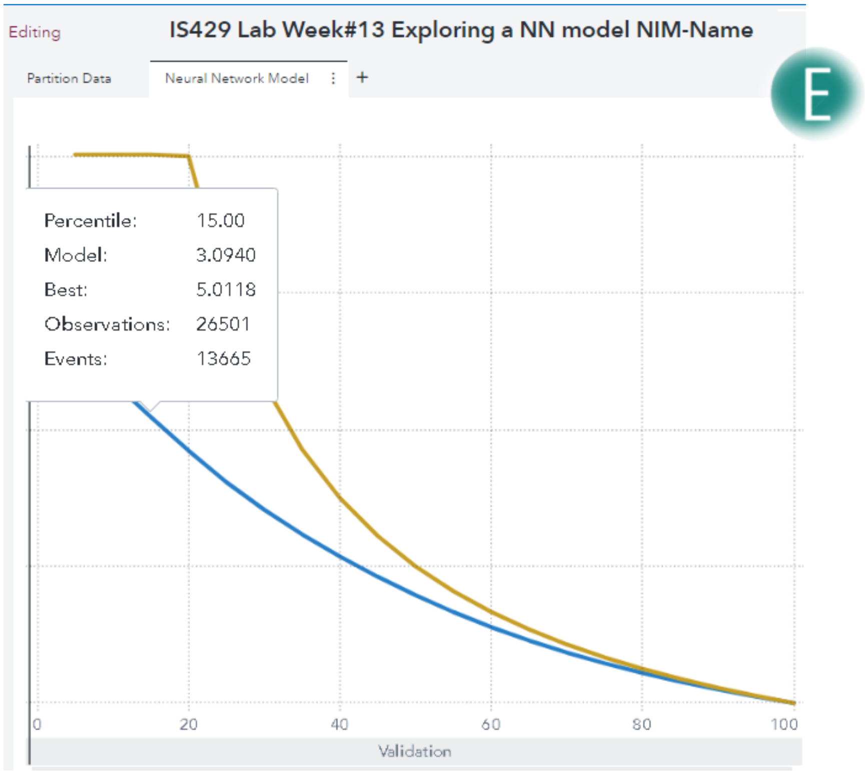
IS429 Lab Week#13 Exploring a NN model NIM-Name

Partition Data | Neural Network Model

Confusion Matrix

Observed	0	1	0	1
0	405,849	18,416	405,489	18,775
1	69,030	36,725	68,955	36,799
Predicted	0	1	0	1
Partition	Training		Validation	

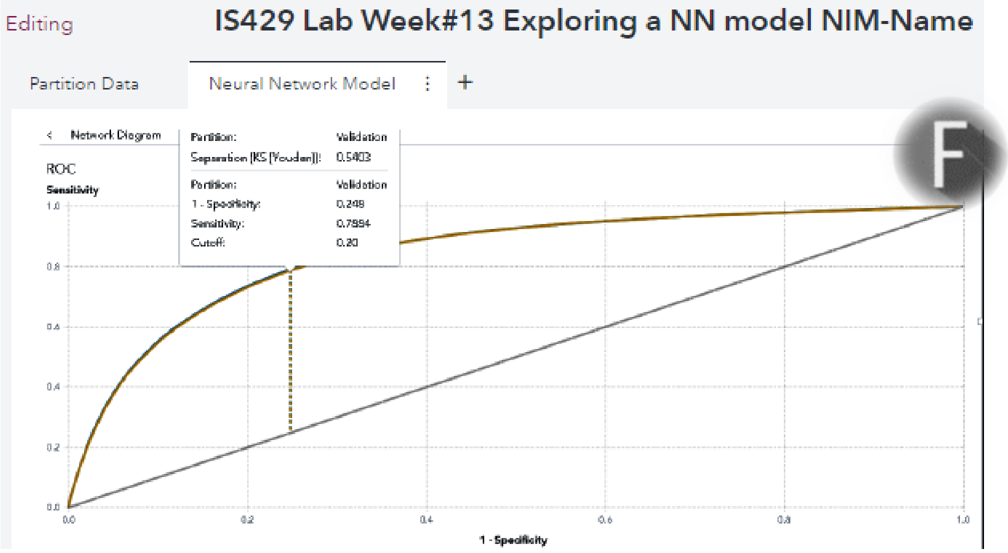
E. Lift Chart


Editing

IS429 Lab Week#13 Exploring a NN model NIM-Name

Percentile: 15.00
Model: 3.0940
Best: 5.0118
Observations: 26501
Events: 13665

F. ROC

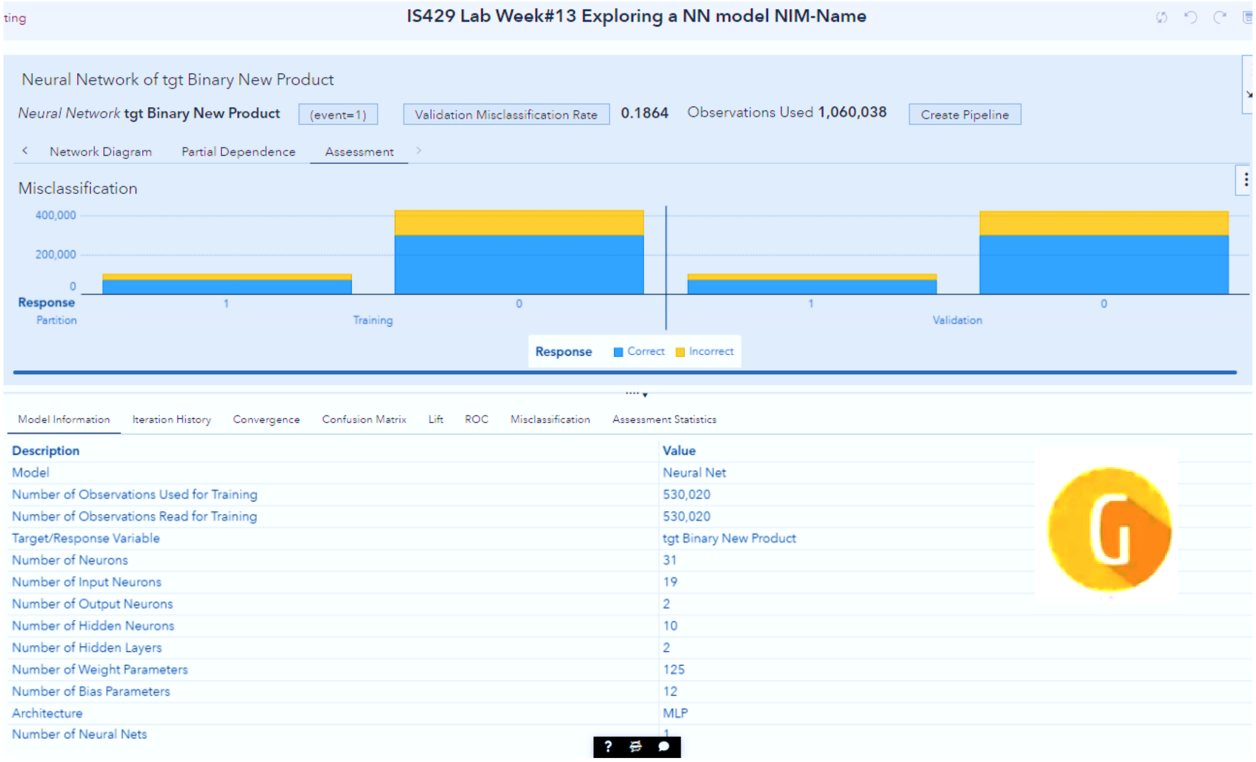

Editing

IS429 Lab Week#13 Exploring a NN model NIM-Name

Partition: Validation
Separation (KS (Youden)): 0.5403

Partition: Validation
1 - Specificity: 0.249
Sensitivity: 0.7994
Cutoff: 0.20

G. Assessment



IS429 Lab Week#13 Exploring a NN model NIM-Name

Neural Network of tgt Binary New Product

Neural Network **tgt Binary New Product** (event=1) Validation Misclassification Rate 0.1864 Observations Used **1,060,038** Create Pipeline

Network Diagram | Partial Dependence | Assessment

Model Information | Iteration History | Convergence | Confusion Matrix | Lift | ROC | Misclassification | Assessment Statistics

Description	Value
Model	Neural Net
Number of Observations Used for Training	530,020
Number of Observations Read for Training	530,020
Target/Response Variable	tgt Binary New Product
Number of Neurons	31
Number of Input Neurons	19
Number of Output Neurons	2
Number of Hidden Neurons	10
Number of Hidden Layers	2
Number of Weight Parameters	125
Number of Bias Parameters	12
Architecture	MLP
Number of Neural Nets	1

The End

REFERENCE

1. SAS Institute Inc. 2020. SAS® Visual Data Mining and Machine Learning in SAS® Viya®: Interactive Machine Learning. SAS Institute Inc. Cary, NC, USA.
2. SAS Institute Inc. 2020. SAS® Viya® Programming: Getting Started. SAS Institute Inc. Cary, NC, USA.
3. SAS® Support | Documentation
4. Other additional references are excerpts from various Online Learning/websites.