Demo with Call Center Data

Using Numeric and Textual Data for Predictive Modeling

Case Study: Improving Predictive Model Using Textual Data

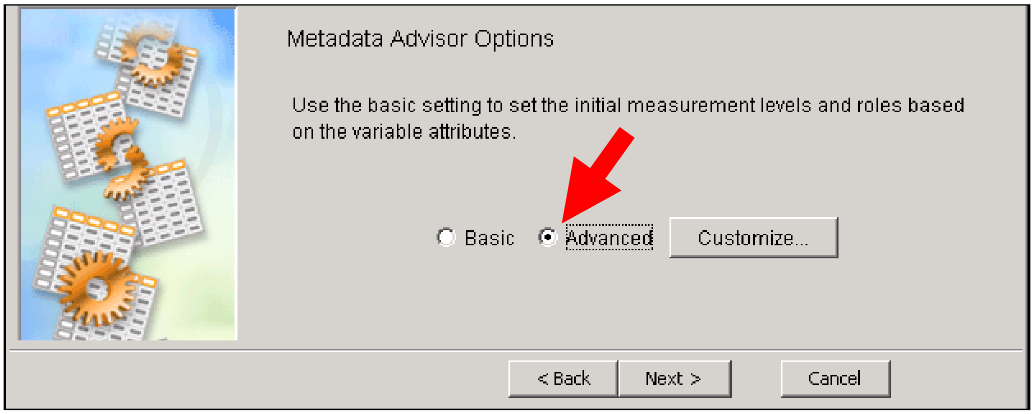
Case Description

The data used in the case study are created based on a real data set of a client company (Fuel Stop Company with over 300 gas stations in the United States). Some of the text comments, variable names, and descriptions have been disguised to protect the identity of the client company and the actual nature of the project. However, you can make out the general nature of the variables by their names. The case involves customers calling the fuel company's call center for many different reasons. Customers' comments via phone were captured by call center reps and typed in a form. These comments were later merged with numeric variables from the fuel company's database about these customers (by matching them via the company's loyalty card number).

The merged data set (**GAS_TEXT_NUMERIC_DATA**) is being used in this case study. The purpose of this case study is to demonstrate how the use of textual data in conjunction with numeric data in a predictive model improves the performance of the predictive model.

Note: In the steps below, we open an already created diagram ***gas_text_numeric_predmdel(nodata)*** and then create a library (name it Course) and add the data source to this digram.

1. Create a new project or start with an existing project.
2. Right-click diagrams in the project panel and select Import Diagram from XML. Select he diagram ***gas_text_numeric_predmdel(nodata)***
3. Create a library (name it as Course)to point to a folder where the data are located. Add the data source, **gas_text_numeric_data**, to the project (via your library).
4. In Step 4, ensure that you select **Advanced** under Metadata Advisor Options as shown below.

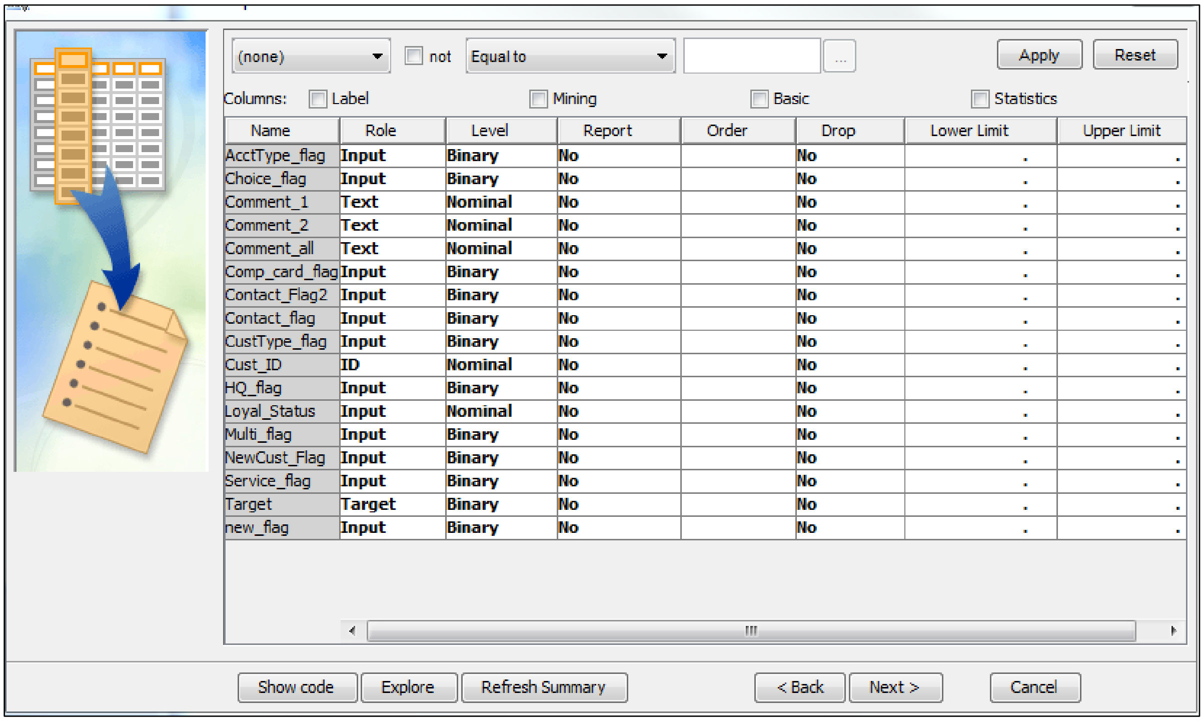


Metadata Advisor Options

Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

○ Basic ● Advanced

5. The variable roles and levels are shown below.



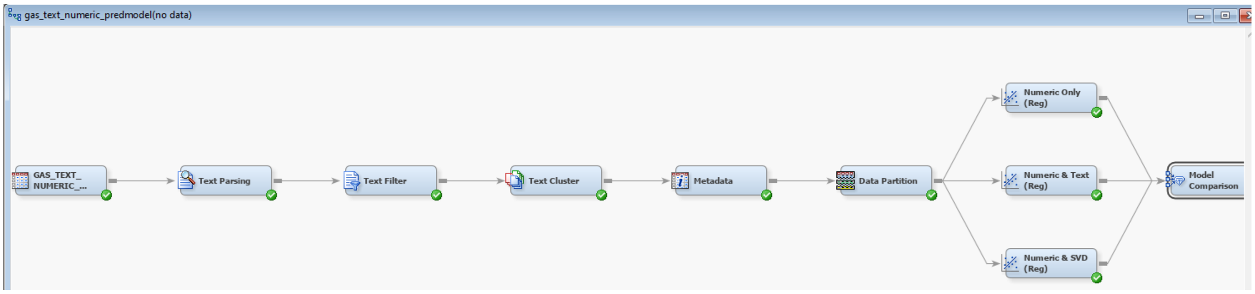
Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AcctType_flag	Input	Binary	No		No	.	.
Choice_flag	Input	Binary	No		No	.	.
Comment_1	Text	Nominal	No		No	.	.
Comment_2	Text	Nominal	No		No	.	.
Comment_all	Text	Nominal	No		No	.	.
Comp_card_flag	Input	Binary	No		No	.	.
Contact_Flag2	Input	Binary	No		No	.	.
Contact_flag	Input	Binary	No		No	.	.
CustType_flag	Input	Binary	No		No	.	.
Cust_ID	ID	Nominal	No		No	.	.
HQ_flag	Input	Binary	No		No	.	.
Loyal_Status	Input	Nominal	No		No	.	.
Multi_flag	Input	Binary	No		No	.	.
NewCust_Flag	Input	Binary	No		No	.	.
Service_flag	Input	Binary	No		No	.	.
Target	Target	Binary	No		No	.	.
new_flag	Input	Binary	No		No	.	.

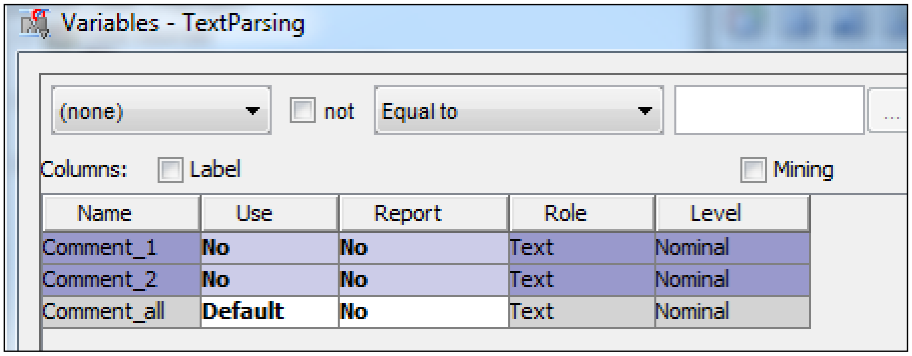
6. Click through and finish the next data creation steps by accepting the default options.

7. Drag the **gas_text_numeric_data** data to the diagram space.

8. **Add** the data source to Text Parsing node.



9. Right-click the **Text Parsing** node and select **Edit Variables**. Note that the Use role for **Comment_1** and **Comment_2** has been changed to **No**.



Variables - TextParsing

Columns: □ Label □ Mining

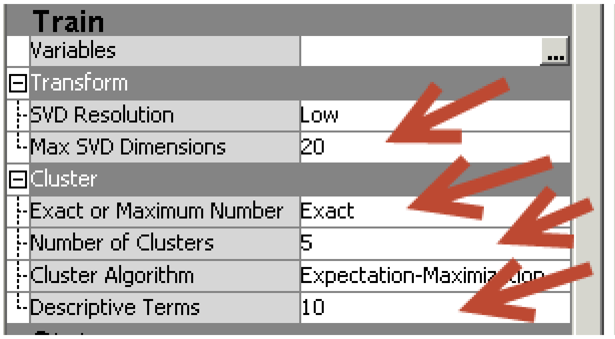
Name	Use	Report	Role	Level
Comment_1	No	No	Text	Nominal
Comment_2	No	No	Text	Nominal
Comment_all	Default	No	Text	Nominal

In this case study, you are using all of the comments together to create text clusters. It is, however, possible to create clusters separately for **Comment_1** and **Comment_2** and perhaps you should explore that on your own as a self-study.

10. Right-click the **Text Cluster** node and examine the results.

    You will find that there are many small clusters with few observations when the Text Cluster node is run with its default settings. This is not surprising given the small data set.

11. The following highlighted changes have been made in the properties panel of the Text Cluster node. Given small data size, for demonstration, we will ask SAS Text Miner to create a maximum of 20 SVD dimensions and exactly 5 clusters and describe those clusters using 10 terms

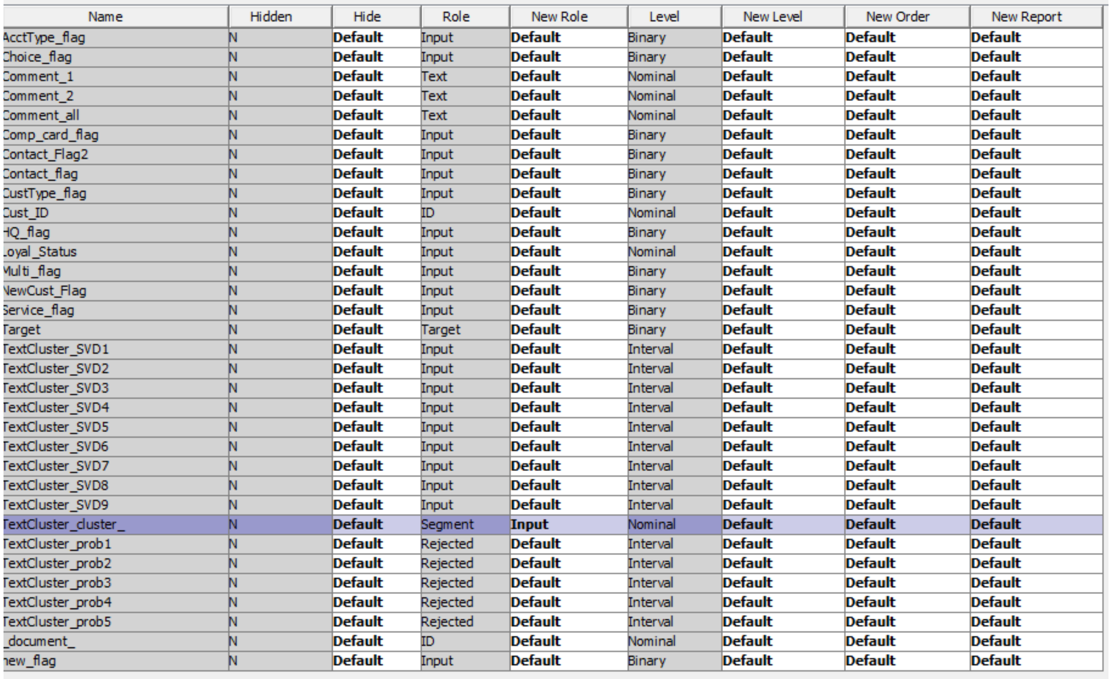


Train	
Variables	...
⊟ Transform	
SVD Resolution	Low
Max SVD Dimensions	20
⊟ Cluster	
Exact or Maximum Number	Exact
Number of Clusters	5
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	10

12. Right-click the **Text Cluster** node and examine the results.

    You should explore the cluster solution to get a feel for what these clusters might represent. You can use a Segment Profile node to profile these clusters using the numeric variables in the data.

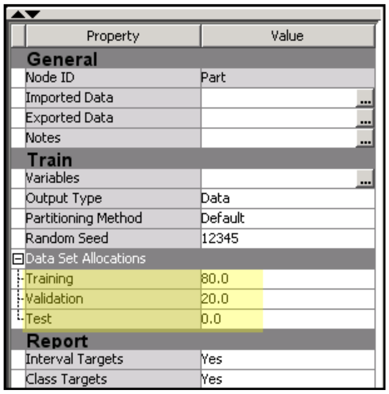
13. In the **Metadata** node, click the ellipsis button next to **Train** in the Variables section of the properties panel of the metadata. Then note the following changes as shown below.



Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
AcctType_flag	N	Default	Input	Default	Binary	Default	Default	Default
Choice_flag	N	Default	Input	Default	Binary	Default	Default	Default
Comment_1	N	Default	Text	Default	Nominal	Default	Default	Default
Comment_2	N	Default	Text	Default	Nominal	Default	Default	Default
Comment_all	N	Default	Text	Default	Nominal	Default	Default	Default
Comp_card_flag	N	Default	Input	Default	Binary	Default	Default	Default
Contact_Flag2	N	Default	Input	Default	Binary	Default	Default	Default
Contact_flag	N	Default	Input	Default	Binary	Default	Default	Default
CustType_flag	N	Default	Input	Default	Binary	Default	Default	Default
Cust_ID	N	Default	ID	Default	Nominal	Default	Default	Default
HQ_flag	N	Default	Input	Default	Binary	Default	Default	Default
Loyal_Status	N	Default	Input	Default	Nominal	Default	Default	Default
Multi_flag	N	Default	Input	Default	Binary	Default	Default	Default
NewCust_Flag	N	Default	Input	Default	Binary	Default	Default	Default
Service_flag	N	Default	Input	Default	Binary	Default	Default	Default
Target	N	Default	Target	Default	Binary	Default	Default	Default
TextCluster_SVD1	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD2	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD3	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD4	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD5	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD6	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD7	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD8	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_SVD9	N	Default	Input	Default	Interval	Default	Default	Default
TextCluster_cluster_	N	Default	Segment	Input	Nominal	Default	Default	Default
TextCluster_prob1	N	Default	Rejected	Default	Interval	Default	Default	Default
TextCluster_prob2	N	Default	Rejected	Default	Interval	Default	Default	Default
TextCluster_prob3	N	Default	Rejected	Default	Interval	Default	Default	Default
TextCluster_prob4	N	Default	Rejected	Default	Interval	Default	Default	Default
TextCluster_prob5	N	Default	Rejected	Default	Interval	Default	Default	Default
document	N	Default	ID	Default	Nominal	Default	Default	Default
new_flag	N	Default	Input	Default	Binary	Default	Default	Default

14. Add a Data Partition node from the Sample tab to the Metadata node.

15. The following changes were made in the properties panel of the Data Partition node under Data Set Allocations: Training is set to **80**, Validation to **20** and Test to **0**.

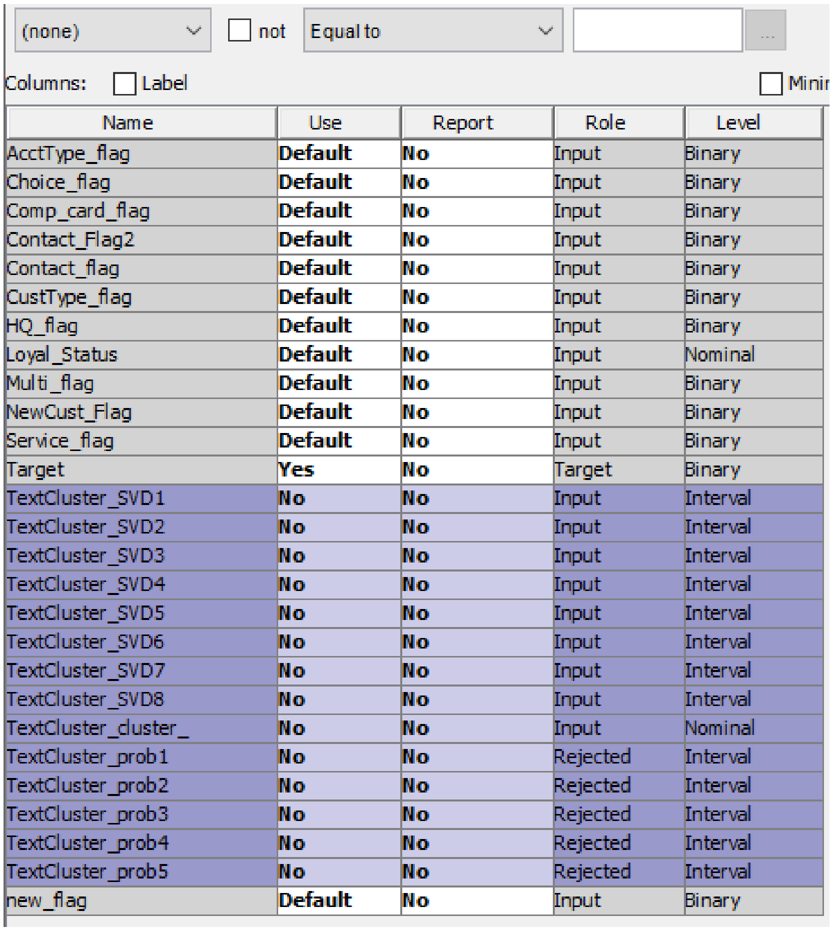


Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	80.0
Validation	20.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes

16. From the Model tab, a **Regression** node has been connected the **Data Partition** node. This node has been renamed as **Numeric Only (Reg)**.

17. Right-click the **Numeric Only (Reg)** node and select **Edit variables**.

18. Note the change in the Use role of all cluster variables to **No**. Click **OK**.

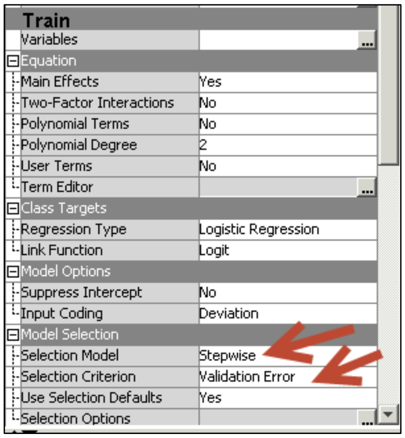


(none) □ not Equal to ...

Columns: □ Label □ Minir

Name	Use	Report	Role	Level
AcctType_flag	Default	No	Input	Binary
Choice_flag	Default	No	Input	Binary
Comp_card_flag	Default	No	Input	Binary
Contact_Flag2	Default	No	Input	Binary
Contact_flag	Default	No	Input	Binary
CustType_flag	Default	No	Input	Binary
HQ_flag	Default	No	Input	Binary
Loyal_Status	Default	No	Input	Nominal
Multi_flag	Default	No	Input	Binary
NewCust_Flag	Default	No	Input	Binary
Service_flag	Default	No	Input	Binary
Target	Yes	No	Target	Binary
TextCluster_SVD1	No	No	Input	Interval
TextCluster_SVD2	No	No	Input	Interval
TextCluster_SVD3	No	No	Input	Interval
TextCluster_SVD4	No	No	Input	Interval
TextCluster_SVD5	No	No	Input	Interval
TextCluster_SVD6	No	No	Input	Interval
TextCluster_SVD7	No	No	Input	Interval
TextCluster_SVD8	No	No	Input	Interval
TextCluster_cluster_	No	No	Input	Nominal
TextCluster_prob1	No	No	Rejected	Interval
TextCluster_prob2	No	No	Rejected	Interval
TextCluster_prob3	No	No	Rejected	Interval
TextCluster_prob4	No	No	Rejected	Interval
TextCluster_prob5	No	No	Rejected	Interval
new_flag	Default	No	Input	Binary

19. In the properties panel of the Numeric Only (Reg) node, the following changes have been made: the selection model is set to **Stepwise** and the selection criterion is set to **Validation Error**.

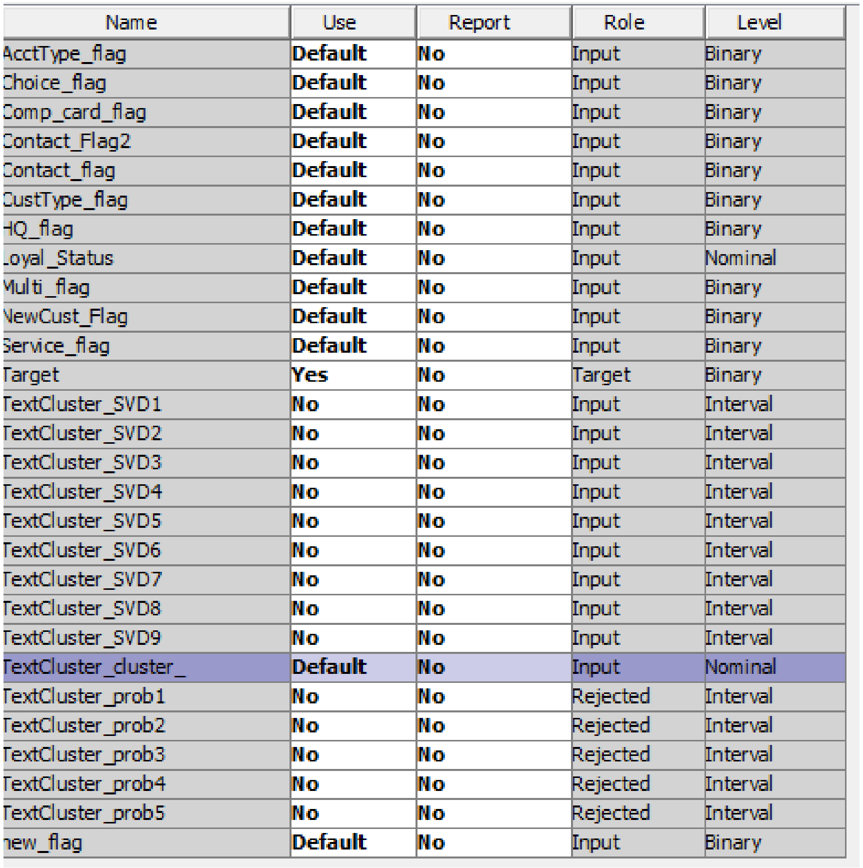


Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

20. In the diagram space, the Numeric Only (Reg) node has been copied and pasted. The name for the pasted node has been changed to **Numeric & Text (Reg)** and connected with the data partition node.

21. Right-click the **Numeric & Text (Reg)** node and select **Edit variables**.

22. Note the change to the Use role of the cluster membership variable from **No** to **Default** as shown below. Then click **OK**.

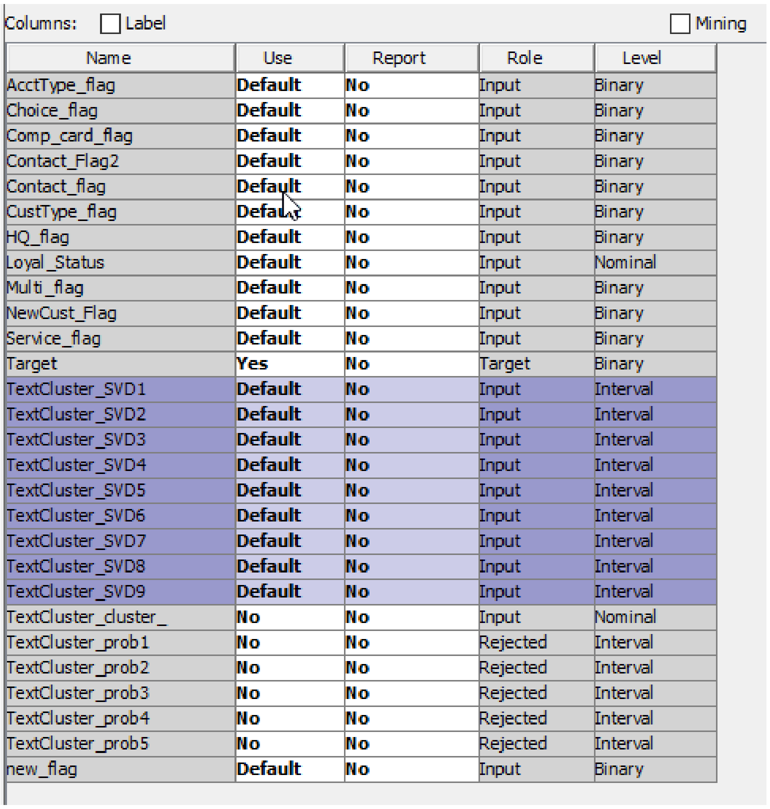


Name	Use	Report	Role	Level
AcctType_flag	Default	No	Input	Binary
Choice_flag	Default	No	Input	Binary
Comp_card_flag	Default	No	Input	Binary
Contact_Flag2	Default	No	Input	Binary
Contact_flag	Default	No	Input	Binary
CustType_flag	Default	No	Input	Binary
HQ_flag	Default	No	Input	Binary
Loyal_Status	Default	No	Input	Nominal
Multi_flag	Default	No	Input	Binary
NewCust_Flag	Default	No	Input	Binary
Service_flag	Default	No	Input	Binary
Target	Yes	No	Target	Binary
TextCluster_SVD1	No	No	Input	Interval
TextCluster_SVD2	No	No	Input	Interval
TextCluster_SVD3	No	No	Input	Interval
TextCluster_SVD4	No	No	Input	Interval
TextCluster_SVD5	No	No	Input	Interval
TextCluster_SVD6	No	No	Input	Interval
TextCluster_SVD7	No	No	Input	Interval
TextCluster_SVD8	No	No	Input	Interval
TextCluster_SVD9	No	No	Input	Interval
TextCluster_cluster_	Default	No	Input	Nominal
TextCluster_prob1	No	No	Rejected	Interval
TextCluster_prob2	No	No	Rejected	Interval
TextCluster_prob3	No	No	Rejected	Interval
TextCluster_prob4	No	No	Rejected	Interval
TextCluster_prob5	No	No	Rejected	Interval
new_flag	Default	No	Input	Binary

22. In the diagram space, the Numeric Only (Reg) node has been copied and pasted. The name of the pasted node has been changed to **Numeric & SVD (Reg)** and connected to the **Data Partition** node.

23. Right-click the **Numeric & SVD (Reg)** node and select **Edit variables**.

24. Note the change in the **Use** role of the SVD variables from **No** to **Default**. Click **OK**.

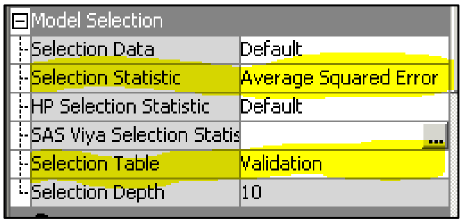


Columns: □ Label □ Mining

Name	Use	Report	Role	Level
AcctType_flag	Default	No	Input	Binary
Choice_flag	Default	No	Input	Binary
Comp_card_flag	Default	No	Input	Binary
Contact_Flag2	Default	No	Input	Binary
Contact_flag	Default	No	Input	Binary
CustType_flag	Default	No	Input	Binary
HQ_flag	Default	No	Input	Binary
Loyal_Status	Default	No	Input	Nominal
Multi_flag	Default	No	Input	Binary
NewCust_Flag	Default	No	Input	Binary
Service_flag	Default	No	Input	Binary
Target	Yes	No	Target	Binary
TextCluster_SVD1	Default	No	Input	Interval
TextCluster_SVD2	Default	No	Input	Interval
TextCluster_SVD3	Default	No	Input	Interval
TextCluster_SVD4	Default	No	Input	Interval
TextCluster_SVD5	Default	No	Input	Interval
TextCluster_SVD6	Default	No	Input	Interval
TextCluster_SVD7	Default	No	Input	Interval
TextCluster_SVD8	Default	No	Input	Interval
TextCluster_SVD9	Default	No	Input	Interval
TextCluster_cluster_	No	No	Input	Nominal
TextCluster_prob1	No	No	Rejected	Interval
TextCluster_prob2	No	No	Rejected	Interval
TextCluster_prob3	No	No	Rejected	Interval
TextCluster_prob4	No	No	Rejected	Interval
TextCluster_prob5	No	No	Rejected	Interval
new_flag	Default	No	Input	Binary

25. A Model Comparison node (from the Assess tab) has been connected with all Regression nodes.

26. Note the changes in the properties of the Model Comparison node.



Model Selection	
Selection Data	Default
Selection Statistic	Average Squared Error
HP Selection Statistic	Default
SAS Viya Selection Statis	...
Selection Table	Validation
Selection Depth	10

27. Right-click the **Model Comparison** node and examine the results.

Notice that for the validation data, the **Numeric & SVD (Reg)** model has clearly outperformed the **Numeric & Text (Reg) model,** which has outperformed the **Numeric Only (Reg)** model. Thus, additions of SVDS or text clusters (or both) have improved the predictive ability of the model over a model that has only numeric variables.

Self-Study:

Explore different panels in the Results window of the Model Comparison node and regression results on your own. Then do following:

- Attach a Text Topic node to the Text Cluster node. Use the **Text Topics** as additional input variables along with text cluster input variables in the predictive models. (Make sure that you change the ***Text_Topic variables roles to Input via a Metadata node*** as I demonstrated in the last example. Explore whether those changes improve the model.
- Try other predictive models such as decision tree and neural net on this data and see if that can improve model prediction.