

KRITHIKA VELLORE PRABHAKAR

SMITAKSHI GUPTA

DSBA 6201- SAS PREDICTIVE ANALYTICS

PROJECT 2- PART 1- PREDICTIVE ANALYTICS EDA&RFM

Task one: You need to perform additional analysis on various variables and make a report.

- o You might want to study which variables are highly correlated. If you find such variables

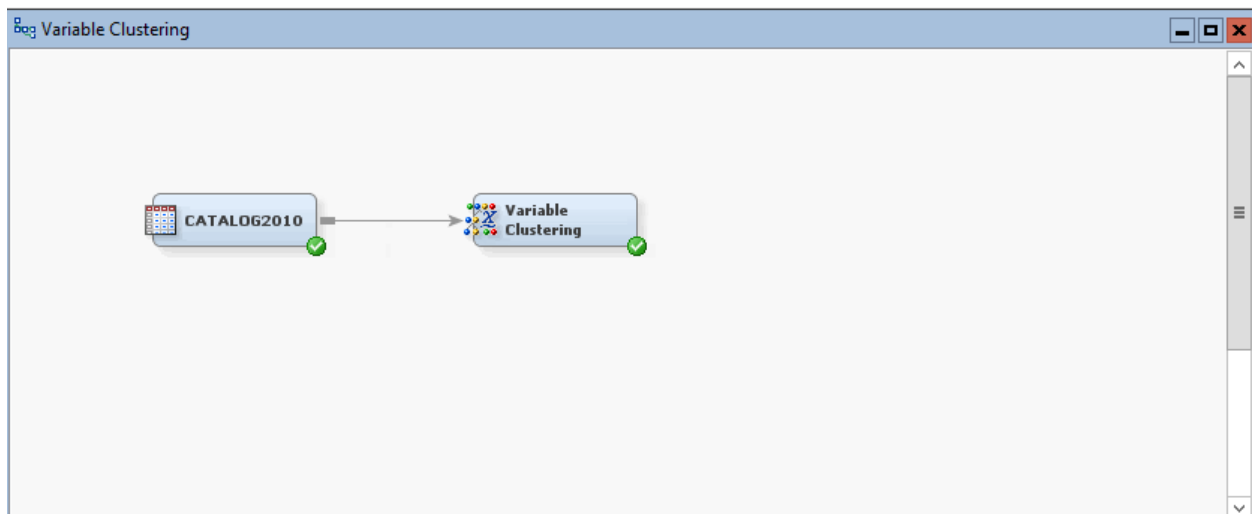
you can suggest dimension reduction by dropping one of the variables.

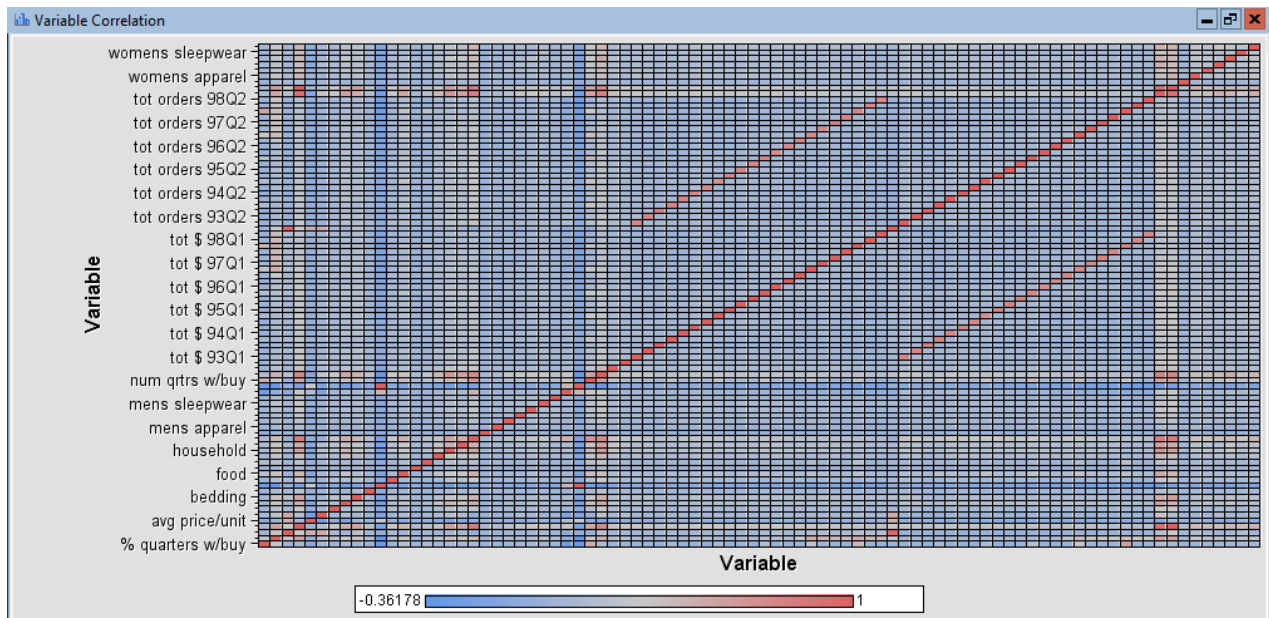
- o You can study in if there are outliers in your variables

- o You can make 3-D plots to get a better sense of how independent variables affect the dependent variables.

To find correlation

To study the variables which are highly correlated, 'Variable Clustering' node can be used. It is used for choosing the best variables for analysis, removes collinearity, and removes redundancy.





Variable	Variable	Correlation ▼
window	window	1
womens apparel	womens apparel	1
womens footwear	womens footwear	1
womens hosiery	womens hosiery	1
womens misc	womens misc	1
womens sleepwear	womens sleepwear	1
womens underwear	womens underwear	1
days since last	months since last	0.999975
months since last	days since last	0.999975
avg \$ net demand	total \$ demand	0.993953
total \$ demand	avg \$ net demand	0.993953
avg \$ demand	tot \$ net demand	0.953178
tot \$ net demand	avg \$ demand	0.953178
tot units demand	total \$ demand	0.881179
total \$ demand	tot units demand	0.881179
avg \$ net demand	tot units demand	0.877362
tot units demand	avg \$ net demand	0.877362
lifetime orders	total \$ demand	0.815402
total \$ demand	lifetime orders	0.815402
avg \$ net demand	lifetime orders	0.812389
lifetime orders	avg \$ net demand	0.812389
lifetime orders	tot units demand	0.804472
tot units demand	lifetime orders	0.804472
tot \$ 93Q3	tot orders 93Q3	0.764131
tot orders 93Q3	tot \$ 93Q3	0.764131
tot \$ 94Q1	tot orders 94Q1	0.757053
tot orders 94Q1	tot \$ 94Q1	0.757053

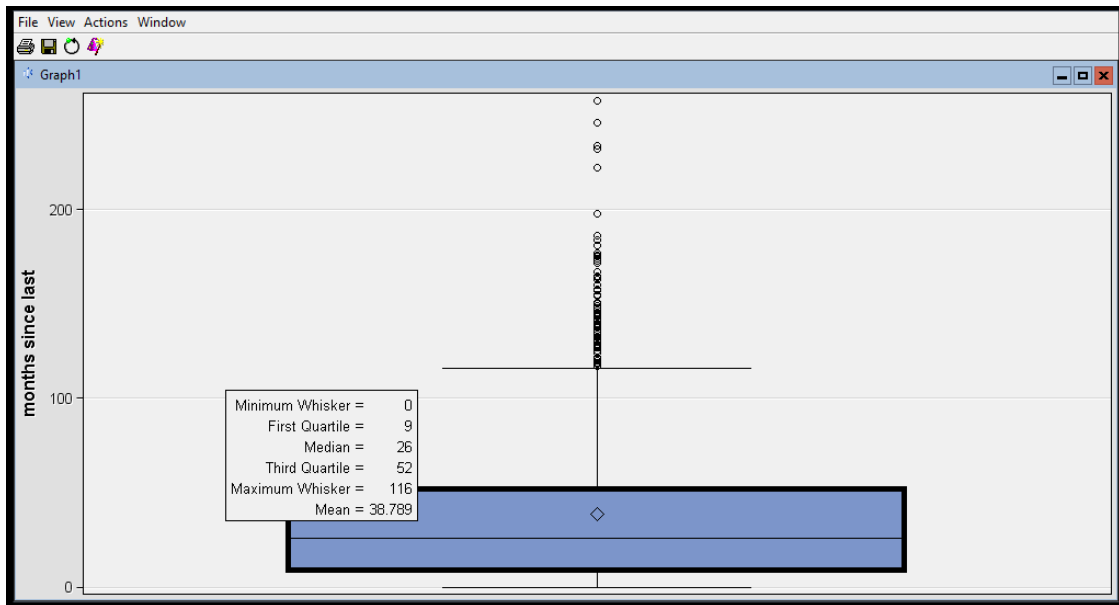
From the Correlation matrix, it is seen that avg \$ net demand have a high correlation of 0.9939 with total \$ demand. This means the variable DOLNETDA and DOLINDET are correlated with each other. Also, the avg \$ demand and tot \$net demand are correlated with a value of 0.9531 which means that the variable DOLINDEA and DOLNETDT.

Considering net demand is a more important indicator for the business than the gross demand, we choose to remove DOLINDET and DOLINDEA from the analysis for dimension reduction.

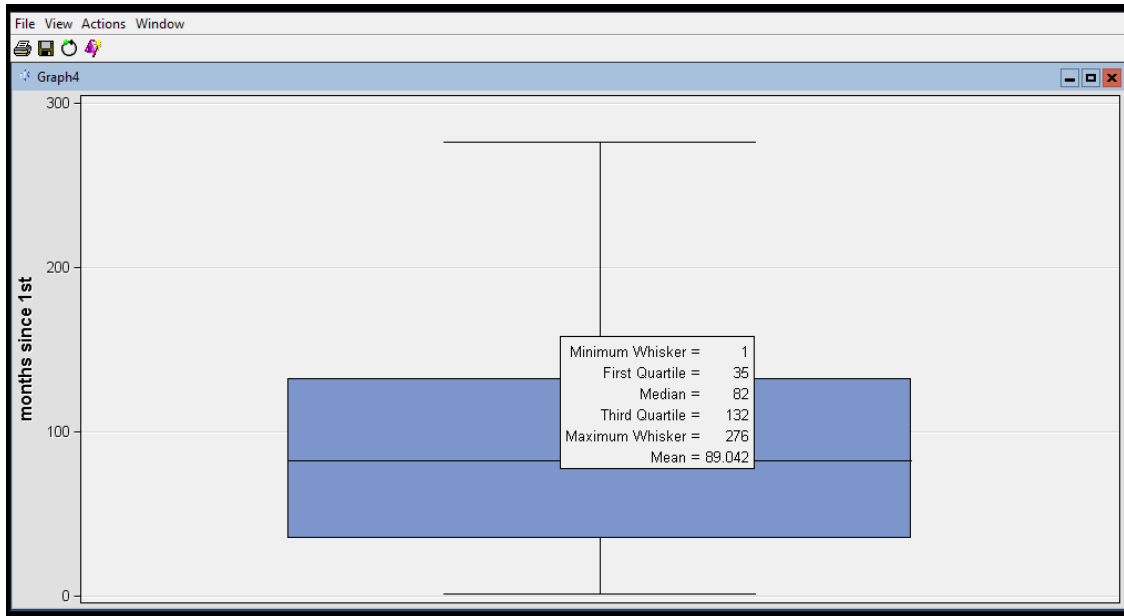
To find outliers

In order to find the outliers, we plot box and whisker plots. Considering MONLAST (number of purchases since last purchase), the plot will be as follows.

From the plot, we can calculate outliers for each variable using formula $1.5*(Q3-Q1)$. For MONLAST, the cutoff value for outliers is $1.5*(52-9)= 64.5$. Values that are greater than 64.5 are consider as outliers.

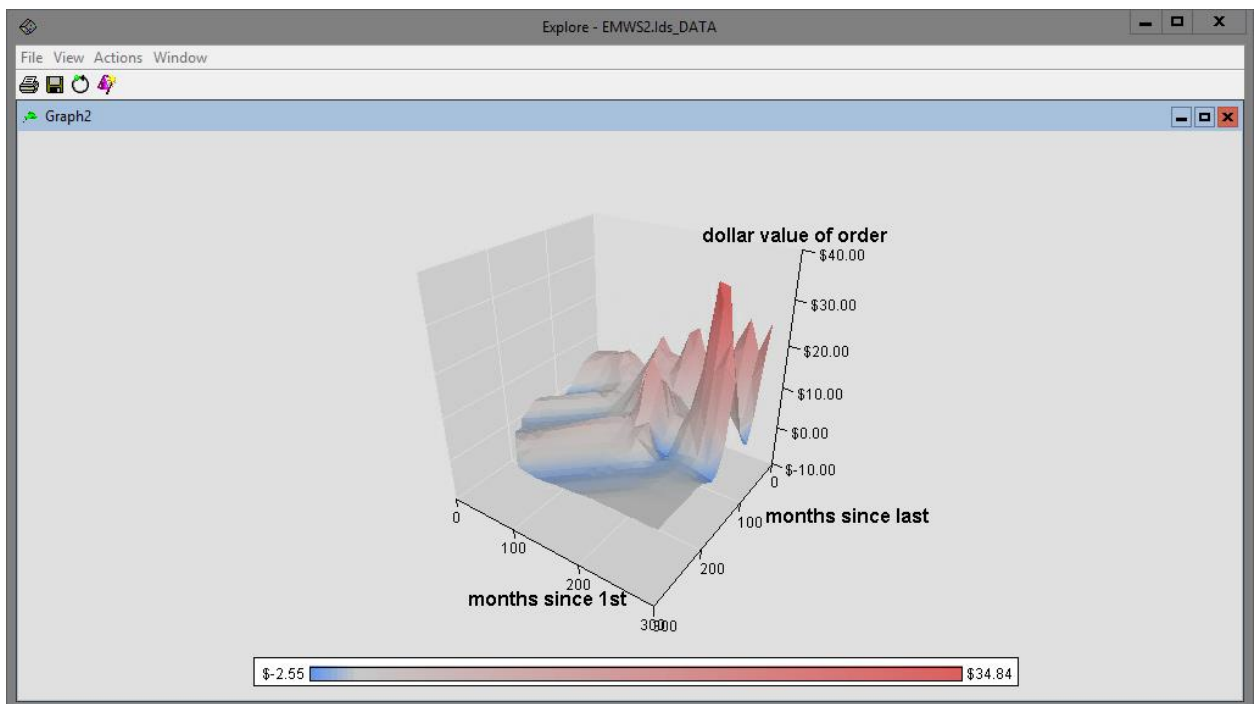


For Tenure(Month since 1st), the cutoff value for outliers is $1.5*(132-35)= 145.5$. Values that are greater than 145.5 are consider as outliers.



3-D Plots

We used ORDERSIZE (dollar value of order) as the dependent variable, as the other target variable RESPONSE is binary, and the independent variables are MONLAST (months since last) and TENURE (months since 1st). The plotted 3D chart is as follows.



Task two:

A national veterans' organization seeks to better target its solicitations for donation. By soliciting only the most likely donors, less money is spent on solicitation efforts and more money is available for charitable concerns. Solicitations involve sending a small gift to an individual and including a request for a donation. Gifts to donors include mailing labels and greeting cards. The organization has more than 3.5 million individuals in its mailing database. These individuals are classified by their response behaviors to previous solicitation efforts. Of particular interest is the class of individuals identified as *lapsing donors*. These individuals made their most recent donation between 12 and 24 months ago. The organization seeks to rank its lapsing donors based on their responses to a greeting card mailing sent in June of 1997. (The charity calls this the 97NK Campaign.) With this ranking, a decision can be made to either solicit or ignore a lapsing individual in the June 1998 campaign.

- a) Define **PVA97NK** as a data source in SAS Enterprise Miner. Use the Advanced Metadata Advisor options to customize the following:
 - Change the Class Levels Count Threshold from 20 to 5.
 - Change the Reject Levels Count Threshold from 20 to 80.

The screenshot shows the 'Advanced Metadata Advisor' dialog box in SAS Enterprise Miner. The 'Property' table is visible, with the 'Reject Levels Count Threshold' property highlighted and its value set to 80. Below the table, the 'Reject Levels Count Threshold' section provides a description: 'Specify a maximum number of levels for a class variable before being marked REJECTED. The default value is 20.' The 'OK' and 'Cancel' buttons are at the bottom right.

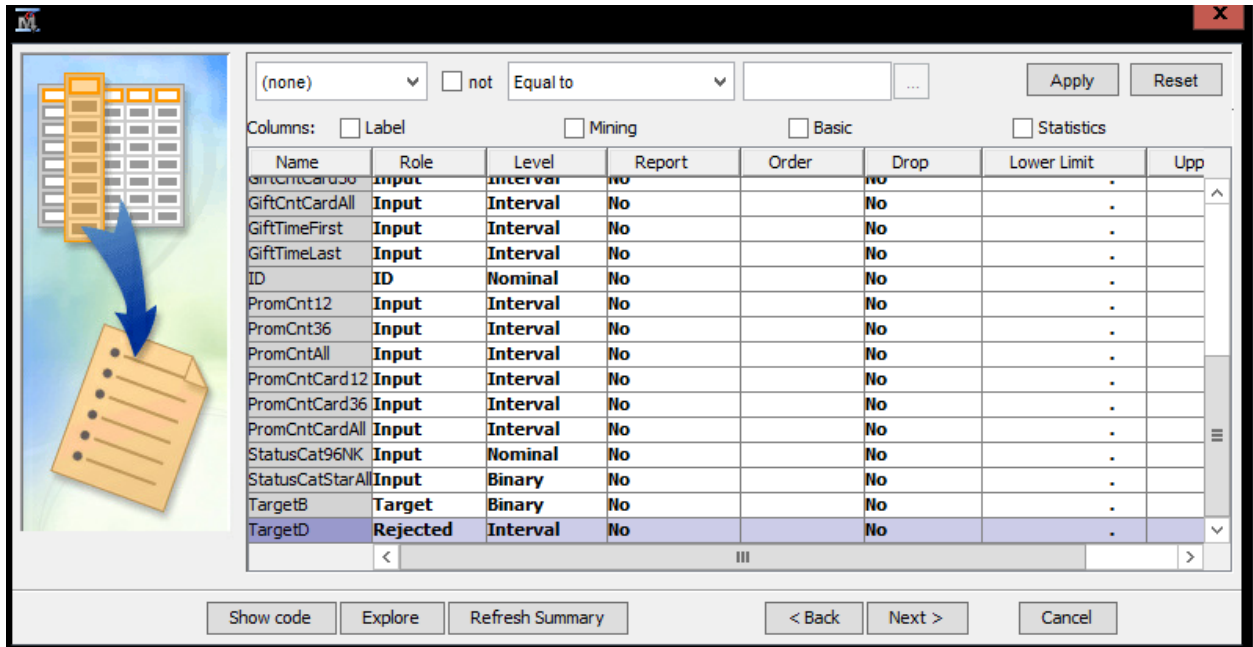
Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Values	Yes
Class Levels Count Threshold	5
Detect Class Levels	Yes
Reject Levels Count Threshold	80
Reject Vars with Excessive Class Values	Yes
Database Pass-Through	Yes

Reject Levels Count Threshold

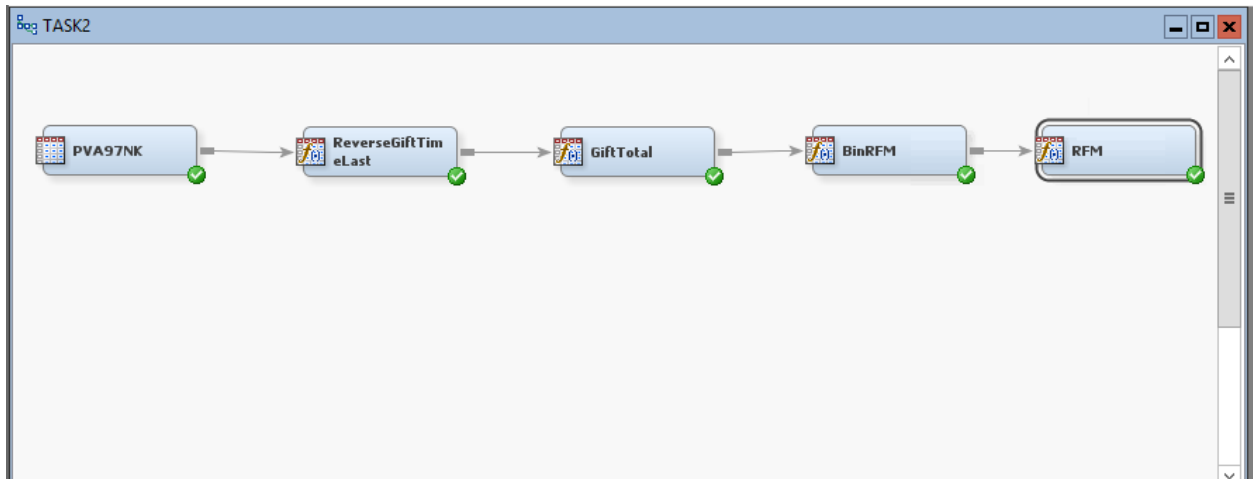
Specify a maximum number of levels for a class variable before being marked REJECTED. The default value is 20.

OK Cancel

Reject the variable **TargetD**.

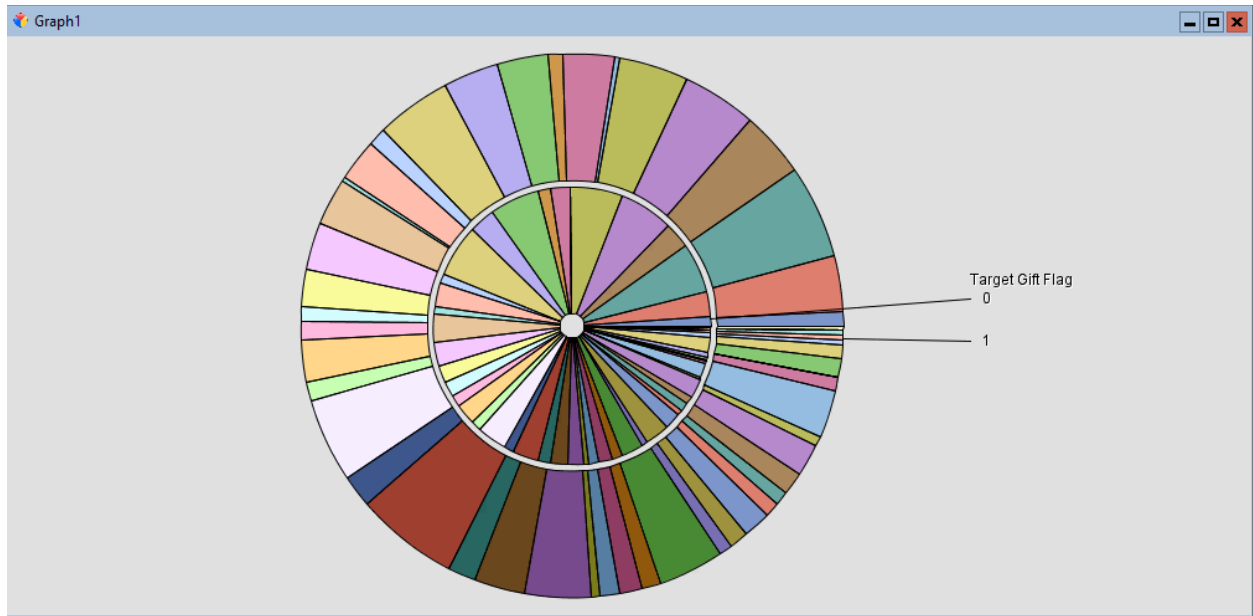


b) Create a new diagram and transform the R, F, and M variables as described previously to create four bins of each variable. Concatenate them to create an RFM variable.

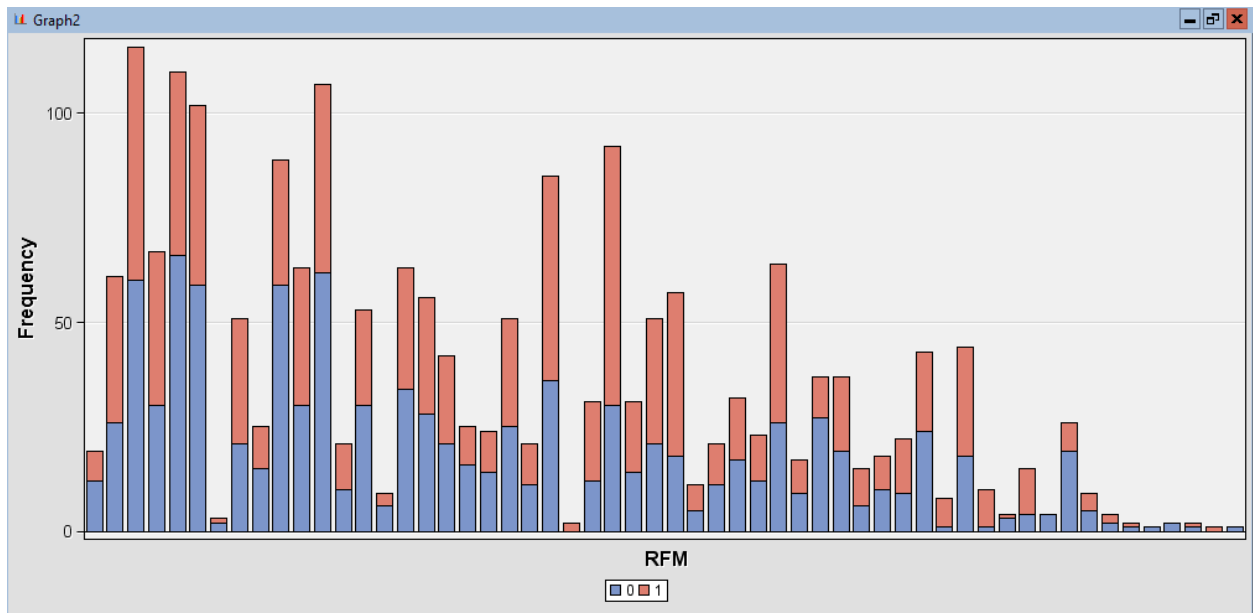


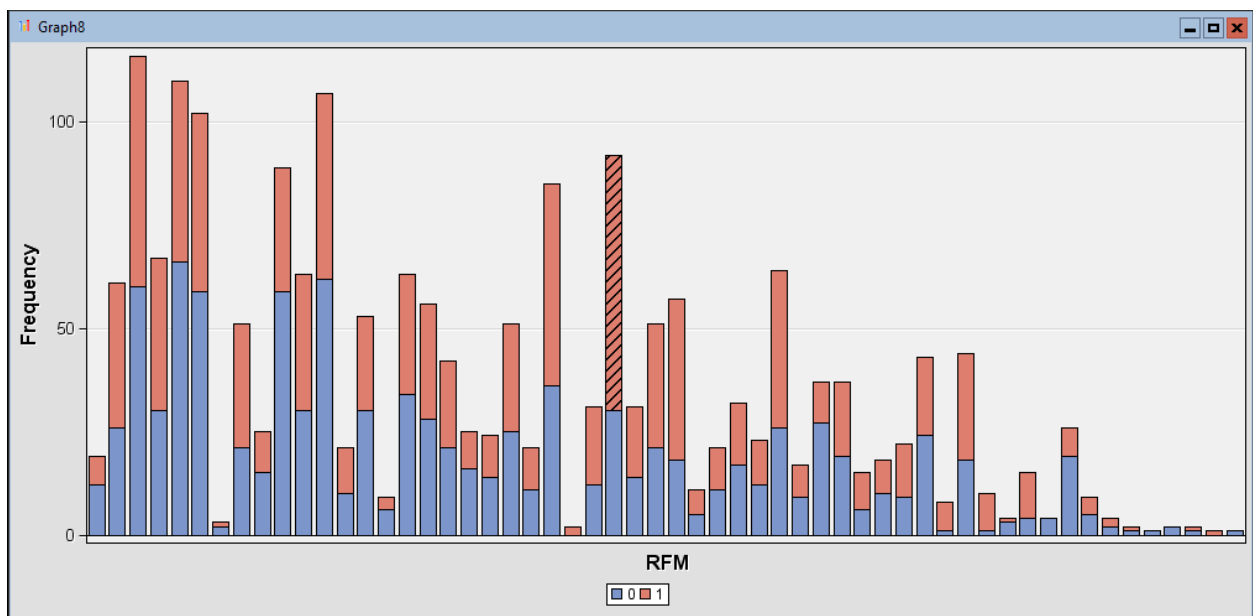
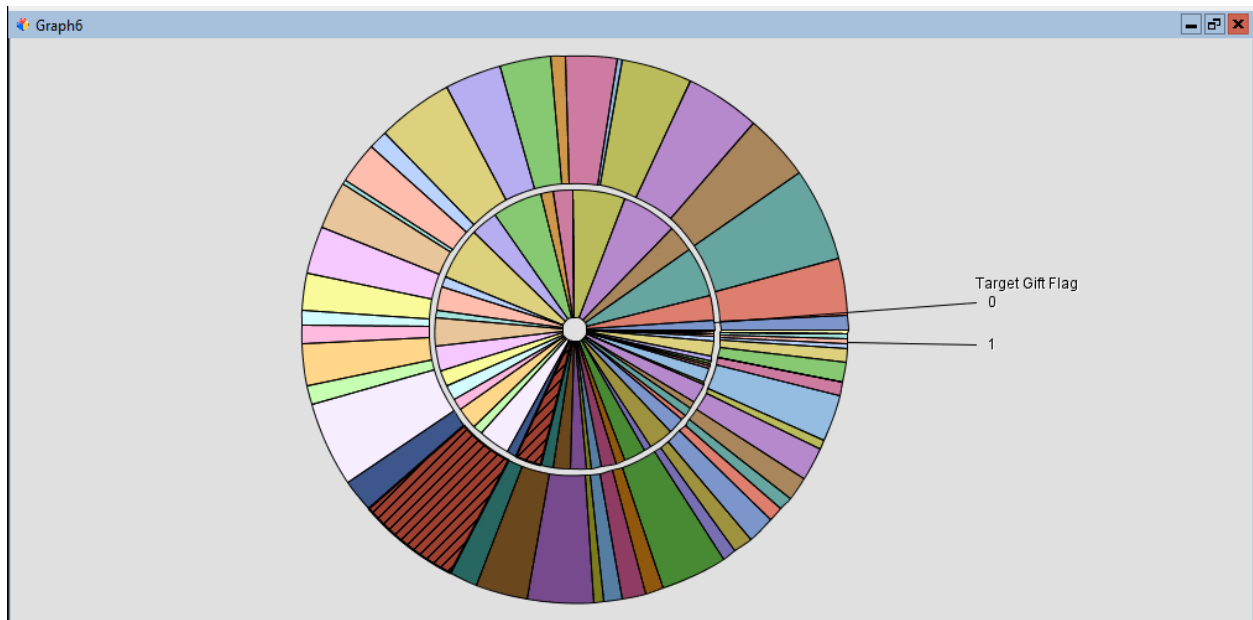
c) Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar chart.

Grouped Pie Chart

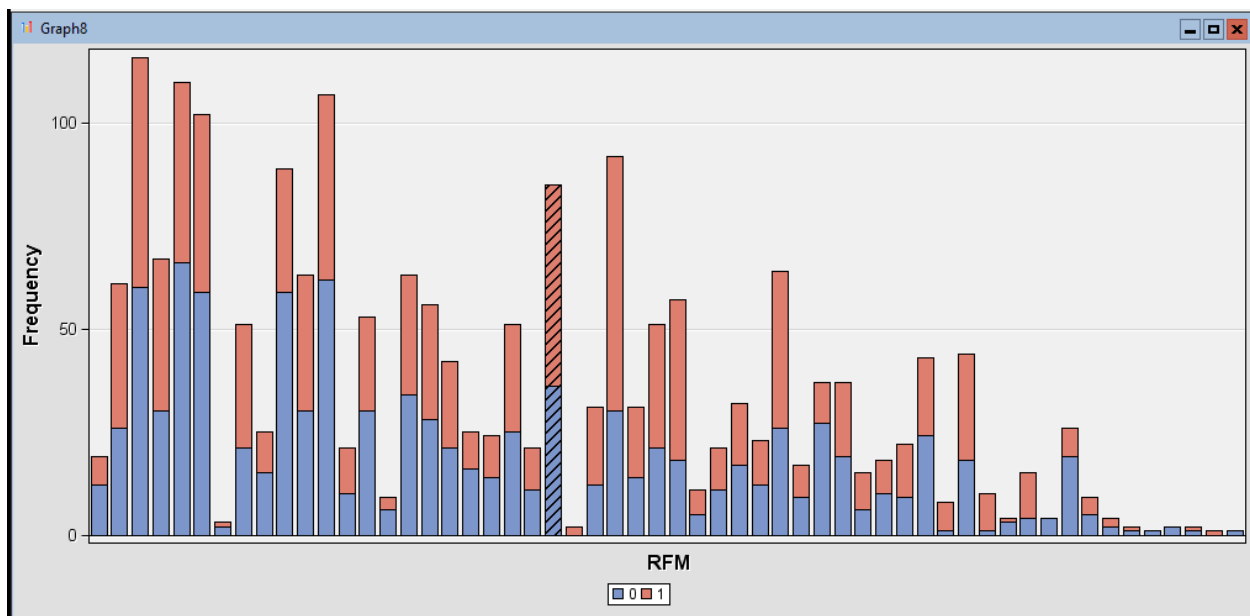
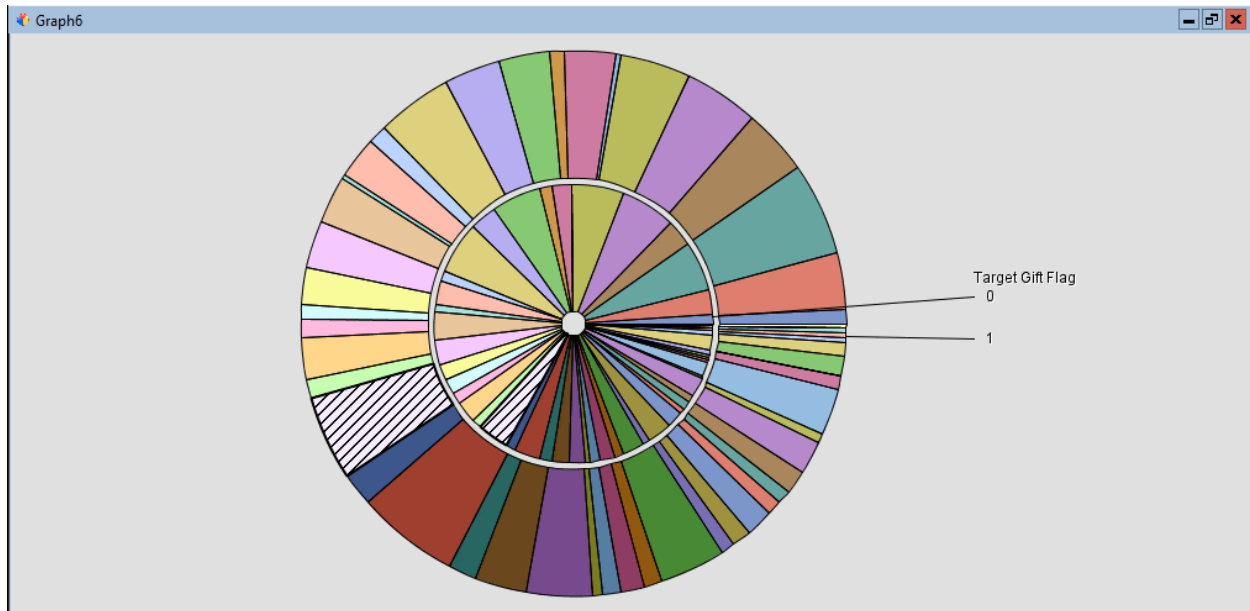


Stacked Bar Chart





Highlighted for 030303 , which has a response rate of 49. The outer ring of the pie chart, which is highlighted as yellow, shows the proportion of the responders of the group to the population of the dataset.



d) Calculate response rate for 040404 and 030303 group?

$$\text{Group 040404} = 62 / (62 + 30) = 67.3\%$$

$$\text{Group 030303} = 49 / (49 + 36) = 57.76\%$$

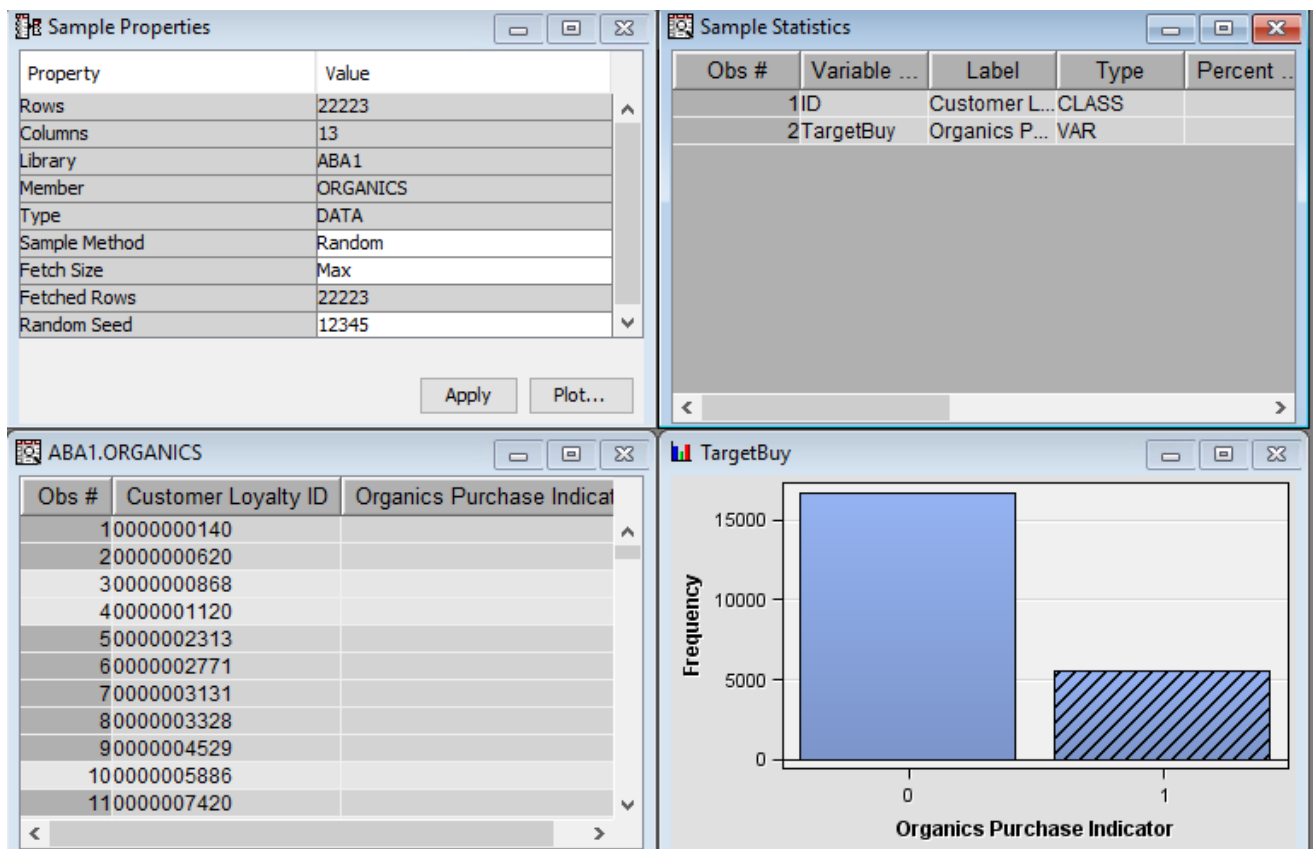
e) Each promotional mailing (request for a gift) costs \$2.3, and the average donation is about \$21. What is the break-even response rate for this promotion? Do any RFM cells exceed this response rate? Remember to account for the fact that in

the population, 95% of mailings are not responded to, while this sample is oversampled to 50% responders and 50% non-responders.

Break-even response rate for this promotion: $2.3/21=10.95\%$. The dataset oversampled the responder to 50% compared to 5% in normal cases. The best response group 040404 only has a response rate of 71.7%. Given a normal condition of 5% response rate, the 040404 group would only have 7.17% ($71.7\%/10$) response rate. Therefore, no group satisfies the break-even threshold.

PREDICTIVE MODELLING - DECISION TREE IN SAS E-MINER - ORGANICS

- a. Create a new diagram named **Organics**.
- b. Define the data set **ORGANICS** as a data source for the project.
 - 1) Set the roles for the analysis variables as shown above.
 - 2) Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?



Among the customers, only 5505 individuals purchased the Organic products which makes approximately 24% individuals.

- 3) The variable **DemClusterGroup** contains collapsed levels of the variable **DemCluster**. Presume that, based on previous experience, you believe that

DemClusterGroup is sufficient for this type of modeling effort. Set the model role for **DemCluster** to Rejected.

4) As noted above, only **TargetBuy** is used for this analysis, and it should have a role of **Target**. Can **TargetAmt** be used as an input for a model that is used to predict **TargetBuy**? Why or why not?

The Decision Tree is a classification predictive modelling technique and usually relates to a categorical target variable. The variable TargetAmt is a numerical variable and needs to be changed into categories in order to be used as a decision variable as an input for the model. Further, if a numerical variable is used as a target then it can be an issue during model deployment and model validation assessments.

5) Finish the **ORGANICS** data source definition.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DemAmt	Input	Interval	No	No	No	-	-
DemAge	Input	Interval	No	No	No	-	-
DemCluster	Rejected	Nominal	No	No	No	-	-
DemClusterGroup	Rejected	Nominal	No	No	No	-	-
DemGender	Input	Nominal	No	No	No	-	-
DemReg	Input	Nominal	No	No	No	-	-
DemTVReg	Input	Nominal	No	No	No	-	-
ID	ID	Nominal	No	No	No	-	-
PromClass	Input	Nominal	No	No	No	-	-
PromSpend	Input	Interval	No	No	No	-	-
PromTime	Input	Interval	No	No	No	-	-
TargetAmt	Rejected	Interval	No	No	No	-	-
TargetBuy	Target	Binary	No	No	No	-	-

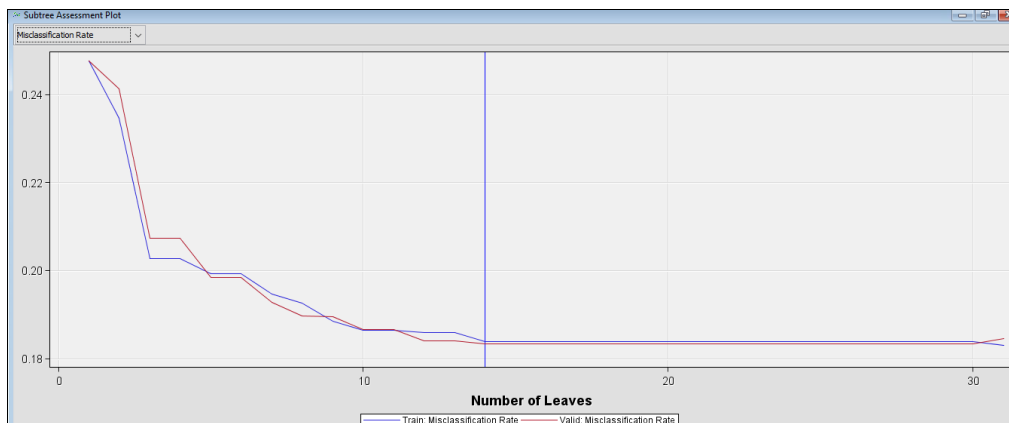
c. Add the **ORGANICS** data source to the Organics diagram workspace.

d. Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 65% of the data for training and 35% for validation.

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	65.0
Validation	35.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/6/18 7:33 PM
Run ID	5215d4d2-3db6-4747-9fce-a7f72ae

e. Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.

f. Create a **decision tree** model autonomously. Use **Misclassification** as the model assessment statistic.

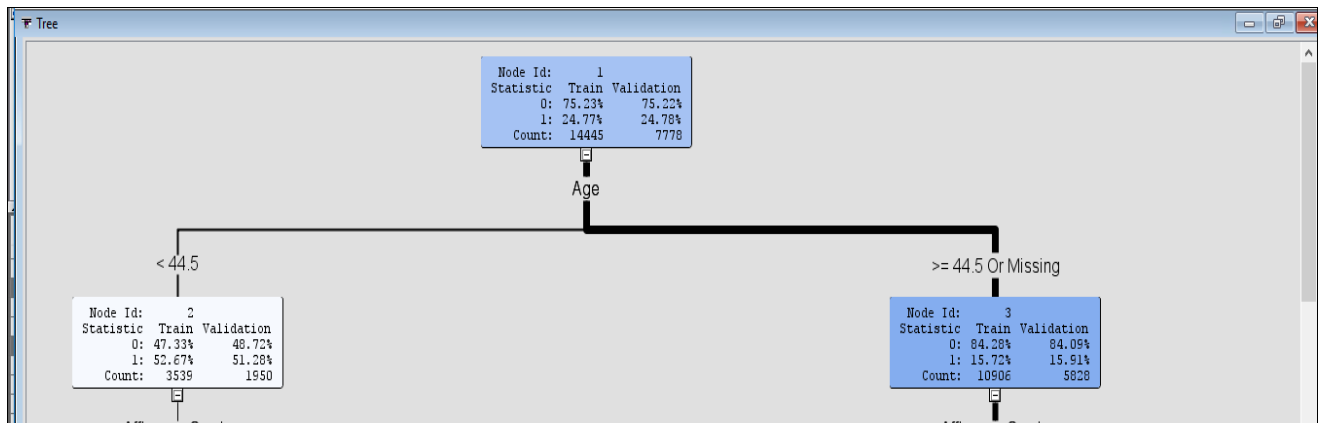


1) How many leaves are in the optimal tree?

The validation performance under Misclassification Rate shows that the optimal tree appears to have fourteen leaves.

- 2) Which variables were used for the first split? What were the competing splits for this first split?

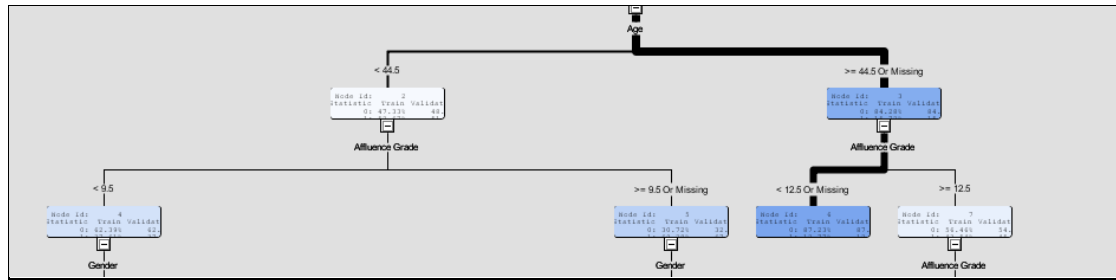
The first split decision is based on the DemAge variable. The criteria were **DemAge < 44.5** and **DemAge => 44.5 or missing**



The competing splits for this split could have been any input variable like DemAffl, DemGender, PromClass and PromSpend

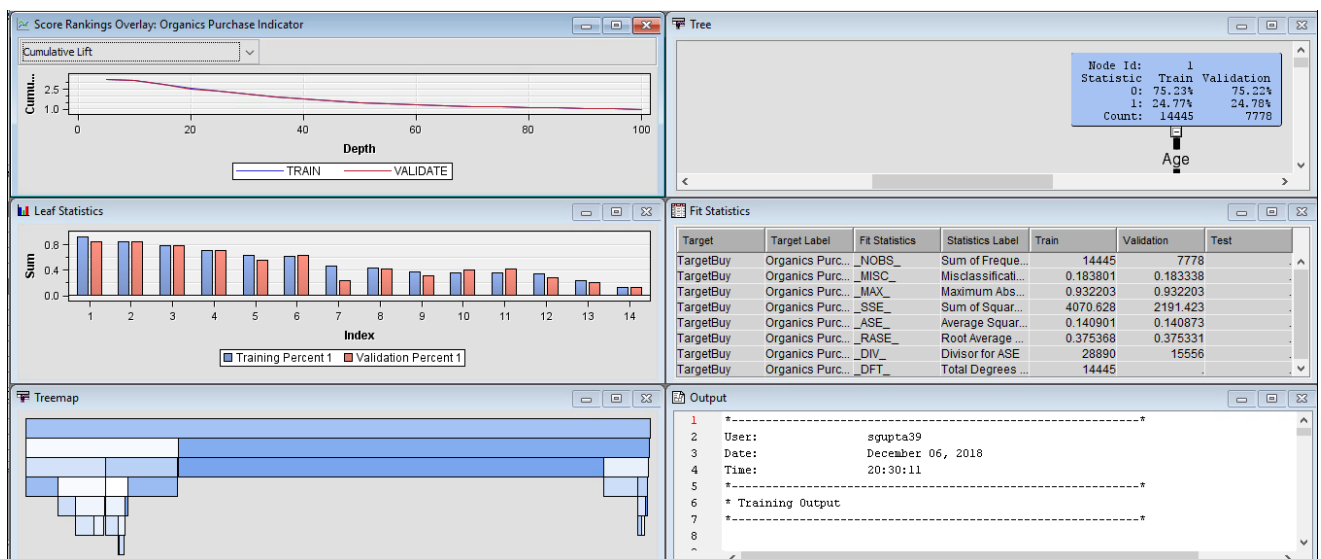
- 3) Which variables were used for the second split for all branches from first split?

The variable DemAffl is used for the second split for all branches from the first split.



4) Discuss the results and provide your insights

The Results window contains a variety of diagnostic plots and tables, including a cumulative lift chart, a tree map, and a table of fit statistics. The diagnostic tools shown in the results vary with the measurement level of the target variable.



The decisions are based on the two factors – accuracy and misclassification rate.

The misclassification rate shows that the optimal tree has 14 leaves and then it leads to a lower misclassification rate for both training and validation datasets as the number of leaves increases.

Since the gains chart is a good indicator of how deep within the marketers are willing to go with respect to promoting a product.



Model Fit Statistics

Since we have provided approximately 65% to the training data sets , it will result in a stable model however it can be a comparatively stable predictive model however the model assessments might be less stable.

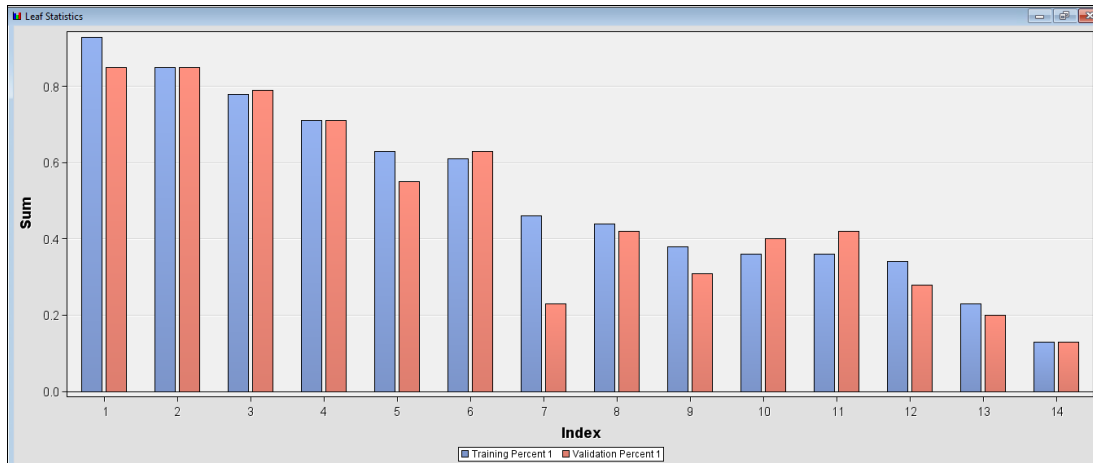
Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetBuy	Organics Purc...	_NOBS_	Sum of Freque...	14445	7778	.
TargetBuy	Organics Purc...	_MISC_	Misclassificati...	0.183801	0.183338	.
TargetBuy	Organics Purc...	_MAX_	Maximum Abs...	0.932203	0.932203	.
TargetBuy	Organics Purc...	_SSE_	Sum of Square...	4070.628	2191.423	.
TargetBuy	Organics Purc...	_ASE_	Average Squar...	0.140901	0.140873	.
TargetBuy	Organics Purc...	_RASE_	Root Average ...	0.375368	0.375331	.
TargetBuy	Organics Purc...	_DIV_	Divisor for ASE	28890	15556	.
TargetBuy	Organics Purc...	_DFT_	Total Degrees ...	14445	.	.

Data Role=TRAIN Target Variable=TargetBuy Target Label=Organics Purchase Indicator							
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	223.013	3.23013	3.23013	80.0098	80.0098	723	0.80010
10	214.901	3.06777	3.14901	75.9880	78.0003	722	0.75988
15	185.699	2.27255	2.85699	56.2906	70.7671	722	0.56291
20	152.885	1.54397	2.52885	38.2439	62.6391	722	0.38244
25	131.393	1.45513	2.31393	36.0433	57.3155	723	0.36043
30	110.746	1.07456	2.10746	26.6167	52.2014	722	0.26617
35	89.982	0.65340	1.89982	16.1847	47.0582	722	0.16185
40	72.683	0.51544	1.72683	12.7675	42.7733	722	0.12767
45	59.211	0.51544	1.59211	12.7675	39.4363	723	0.12767
50	48.449	0.51544	1.48449	12.7675	36.7705	722	0.12767
55	39.643	0.51544	1.39643	12.7675	34.5892	722	0.12767
60	32.304	0.51544	1.32304	12.7675	32.7714	722	0.12767
65	26.085	0.51544	1.26085	12.7675	31.2311	723	0.12767
70	20.763	0.51544	1.20763	12.7675	29.9128	722	0.12767
75	16.150	0.51544	1.16150	12.7675	28.7702	722	0.12767
80	12.114	0.51544	1.12114	12.7675	27.7704	722	0.12767
85	8.547	0.51544	1.08547	12.7675	26.8870	723	0.12767
90	5.382	0.51544	1.05382	12.7675	26.1029	722	0.12767
95	2.549	0.51544	1.02549	12.7675	25.4013	722	0.12767
100	0.000	0.51544	1.00000	12.7675	24.7698	722	0.12767

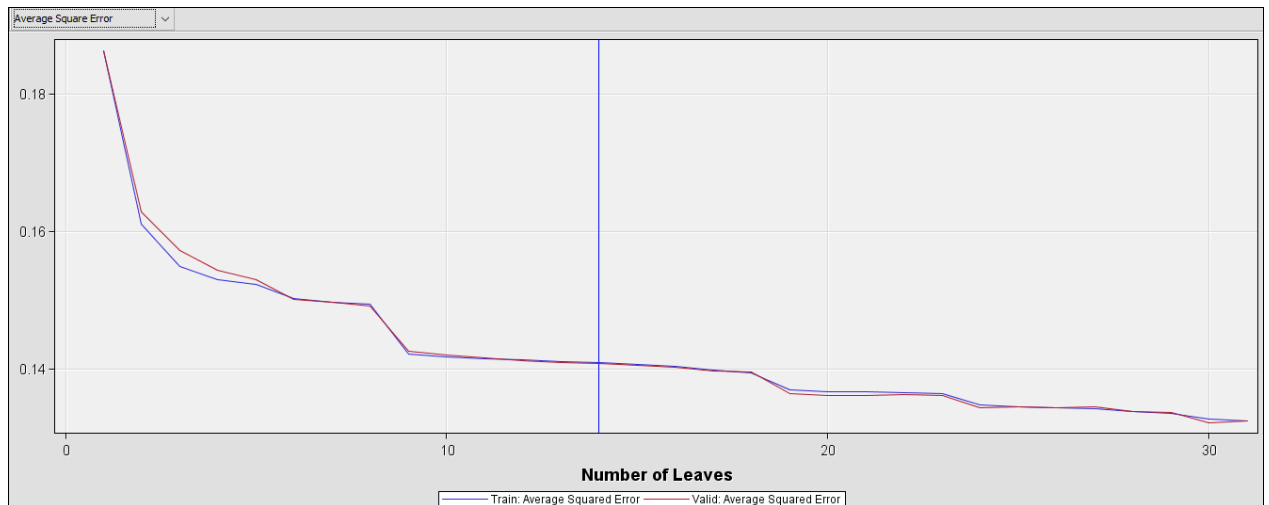
For Validation data

Data Role=VALIDATE Target Variable=TargetBuy Target Label=Organics Purchase Indicator							
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	225.133	3.25133	3.25133	80.5516	80.5516	389	0.80297
10	213.905	3.02677	3.13905	74.9882	77.7699	389	0.75188
15	182.510	2.19721	2.82510	54.4359	69.9919	389	0.55375
20	148.210	1.45310	2.48210	36.0005	61.4941	389	0.38063
25	132.064	1.67481	2.32064	41.4934	57.4939	389	0.36044
30	110.907	1.05122	2.10907	26.0440	52.2523	389	0.26608
35	90.006	0.64598	1.90006	16.0041	47.0739	389	0.17106
40	72.695	0.51516	1.72695	12.7631	42.7851	389	0.12767
45	59.230	0.51516	1.59230	12.7631	39.4493	389	0.12767
50	48.484	0.51516	1.48484	12.7631	36.7869	388	0.12767
55	39.667	0.51516	1.39667	12.7631	34.6024	389	0.12767
60	32.319	0.51516	1.32319	12.7631	32.7821	389	0.12767
65	26.102	0.51516	1.26102	12.7631	31.2418	389	0.12767
70	20.774	0.51516	1.20774	12.7631	29.9217	389	0.12767
75	16.156	0.51516	1.16156	12.7631	28.7776	389	0.12767
80	12.115	0.51516	1.12115	12.7631	27.7765	389	0.12767
85	8.550	0.51516	1.08550	12.7631	26.8933	389	0.12767
90	5.381	0.51516	1.05381	12.7631	26.1081	389	0.12767
95	2.546	0.51516	1.02546	12.7631	25.4057	389	0.12767
100	0.000	0.51516	1.00000	12.7631	24.7750	388	0.12767

Leaf Statistics



The model subtree assessment plot based on average squared error plot shows the average square error corresponding to each tree in the sequence as the data is sequentially split. The plot indicates that with more complexity, neither the training data set performance nor the validation data set performance is better. Both the training and validation datasets show evidence of overfitting.



g. Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.

1) In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to 3 to allow for three-way splits.

2) Create a decision tree model using **Misclassification** as the model assessment statistic.

3) How many leaves are in the optimal tree?

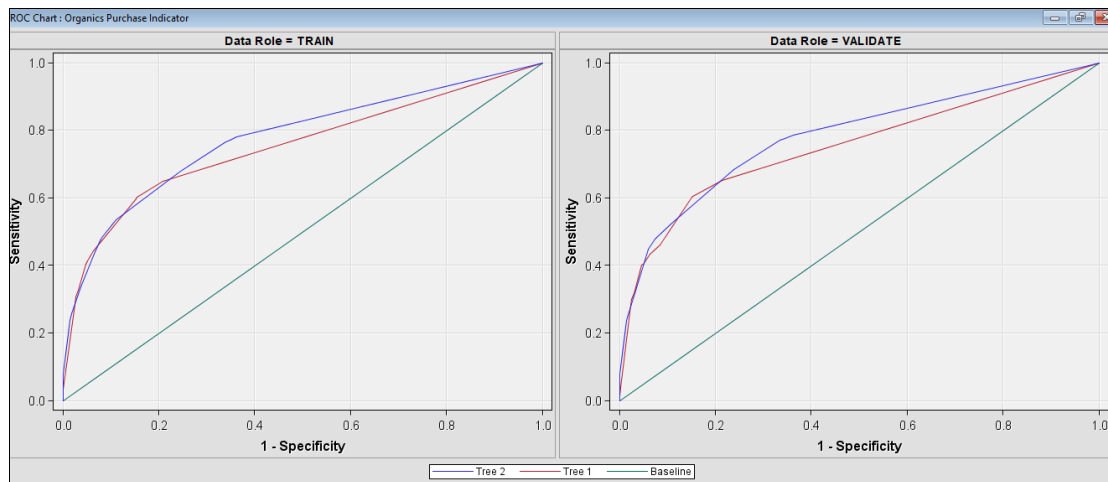
There are 16 leaves in the optimal tree.

h. Based on **Misclassification rate**, which of the decision tree models appears to be better?

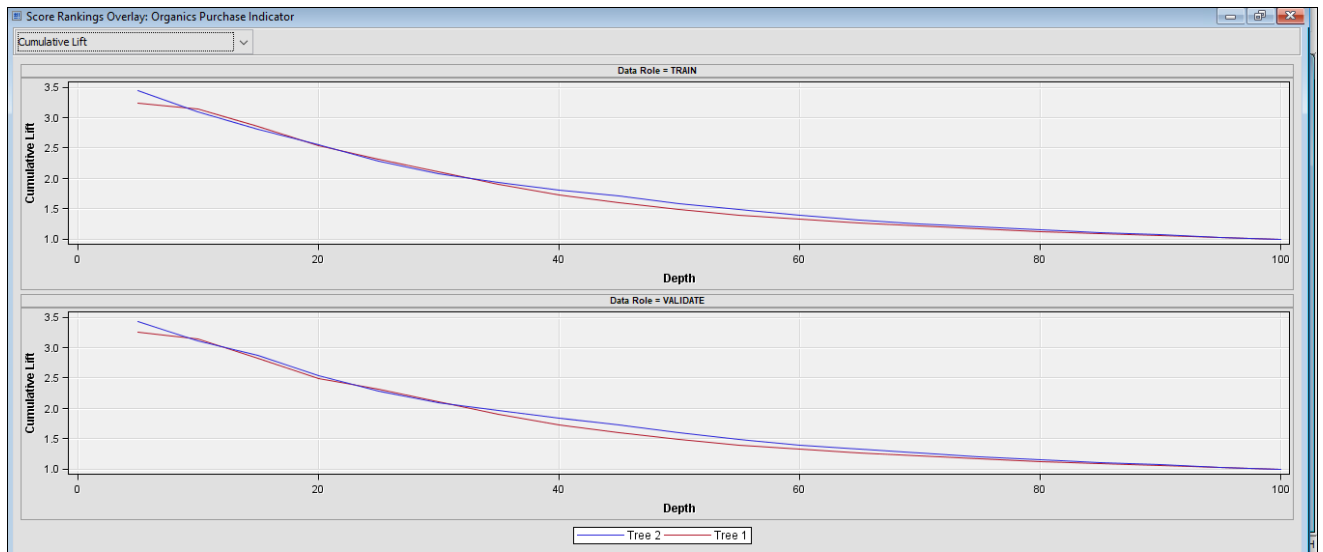
The 2nd decision tree appears to be better because it has a lower misclassification rate than the 1st one.

Output							
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	A	S
Y	Tree2	Tree 2	0.18282	0.13934	0.18858	0	0
	Tree	Tree 1	0.18334	0.14090	0.18380	0	0

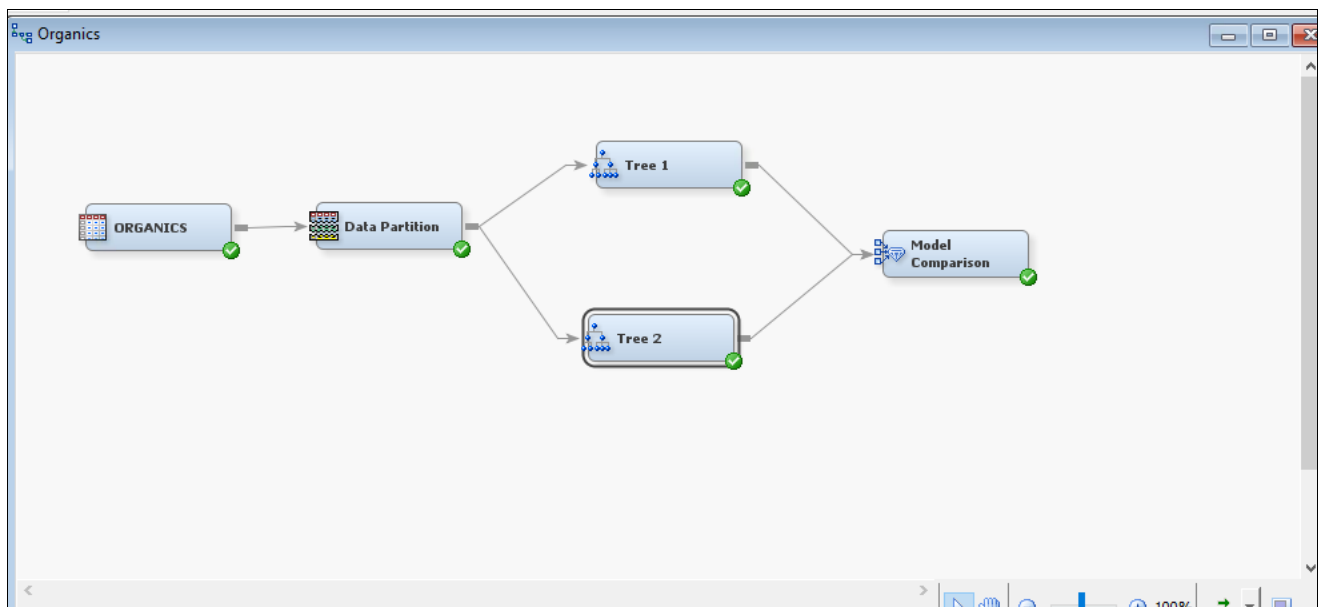
The 2nd decision tree appears to be better because it has a higher ROC statistic as well as a lower average squared error on both training and the validation data set.



The lift statistics of both the trees seemed similar however, the 2nd decision tree has a slightly better lift than the first.

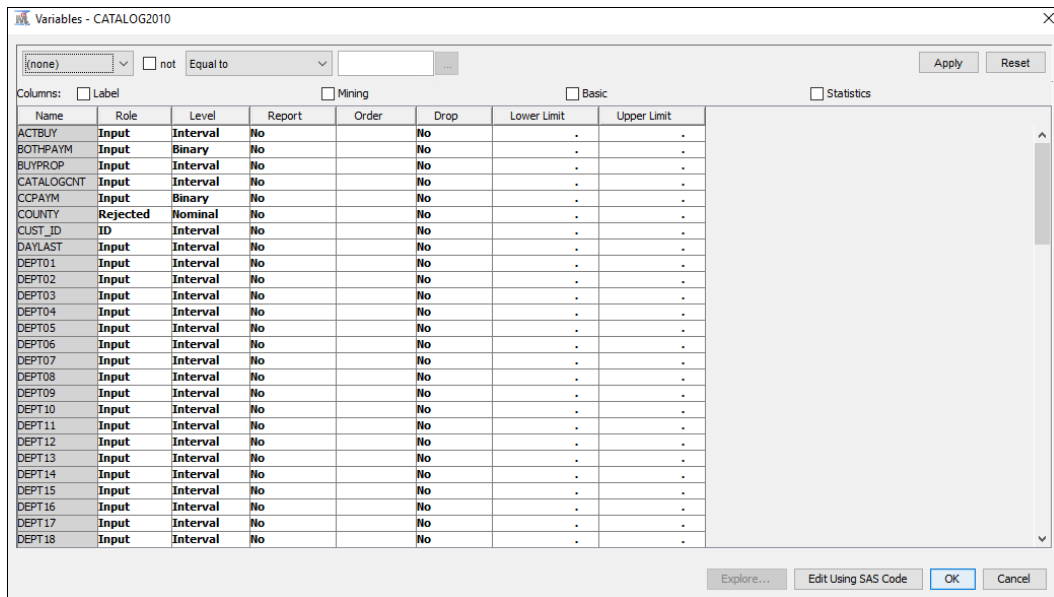


The Decision Tree Diagram for Organics Dataset



PREDICTIVE MODELING USING LOGISTIC REGRESSION

In this exercise we use the CATALOG2010 and fit a logistic regression model in SAS E Miner by adding the Variable Clustering node and the Regression node to the decision tree diagram. The steps that are added to the model-building process include eliminating redundant variables using the Variable Clustering node, eliminating irrelevant variables using the Regression node, and generating model assessment statistics and plots using the Model Comparison node.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ACTBUY	Input	Interval	No		No	-	-
BOTHPAYM	Input	Binary	No		No	-	-
BUYPROP	Input	Interval	No		No	-	-
CATALOGCNT	Input	Interval	No		No	-	-
CCPAYM	Input	Binary	No		No	-	-
COUNTY	Rejected	Nominal	No		No	-	-
CUST_ID	ID	Interval	No		No	-	-
DAYLAST	Input	Interval	No		No	-	-
DEPT01	Input	Interval	No		No	-	-
DEPT02	Input	Interval	No		No	-	-
DEPT03	Input	Interval	No		No	-	-
DEPT04	Input	Interval	No		No	-	-
DEPT05	Input	Interval	No		No	-	-
DEPT06	Input	Interval	No		No	-	-
DEPT07	Input	Interval	No		No	-	-
DEPT08	Input	Interval	No		No	-	-
DEPT09	Input	Interval	No		No	-	-
DEPT10	Input	Interval	No		No	-	-
DEPT11	Input	Interval	No		No	-	-
DEPT12	Input	Interval	No		No	-	-
DEPT13	Input	Interval	No		No	-	-
DEPT14	Input	Interval	No		No	-	-
DEPT15	Input	Interval	No		No	-	-
DEPT16	Input	Interval	No		No	-	-
DEPT17	Input	Interval	No		No	-	-
DEPT18	Input	Interval	No		No	-	-

After checking the variable statistics, we found out that none of the variables have missing values, so an imputation node is not necessary.

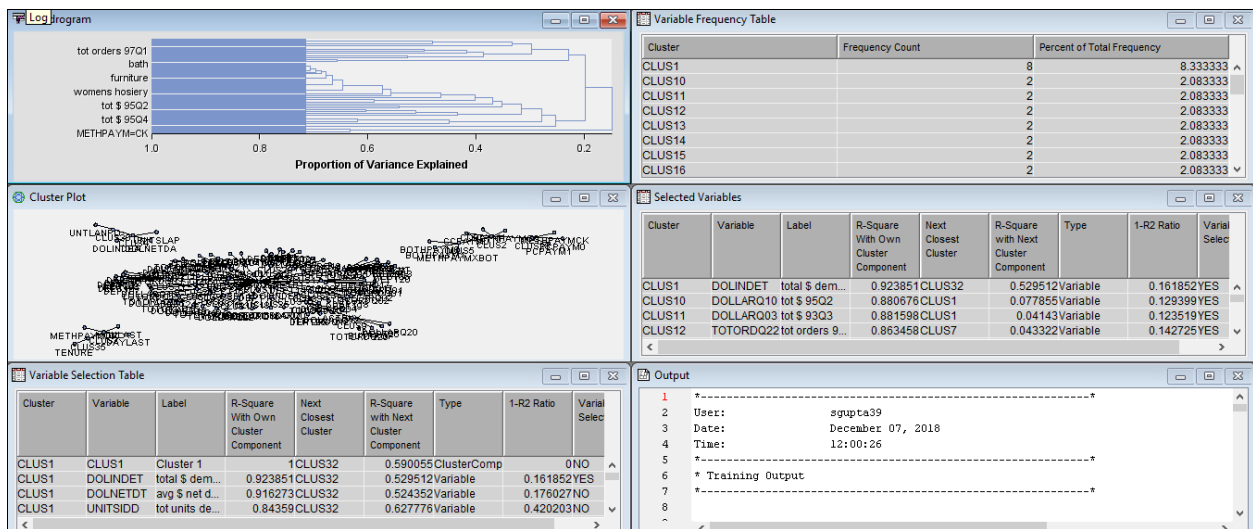
The next step of the exercise is variable clustering. The Variable Clustering node enables us to group variables according to their similarity. The cluster representatives can be automatically selected (by default) or interactively chosen by the user.

After selecting the variable clustering from the Explore panel, we made a few changes in the Properties Panel.

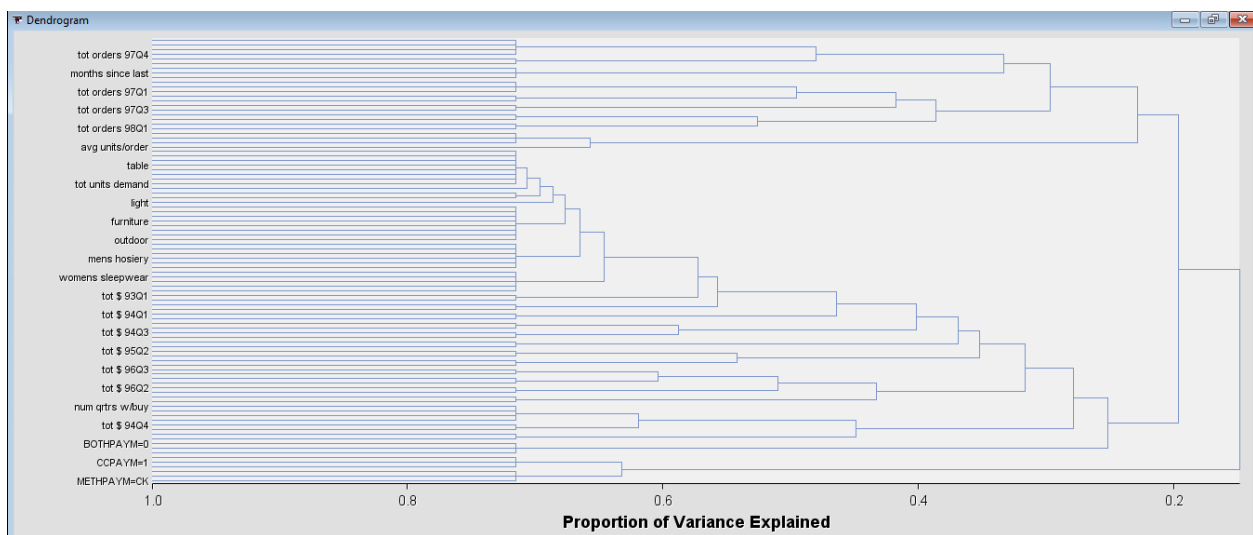
We changed the Includes Class Variables property to **Yes** and the Variable Selection property to **Best Variables**. Best Variables indicates the 1-R2 ratio.

Property	Value
Notes	
Train	
Variables	
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	Yes
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warning	No
Score	
Variable Selection	Best Variables
Interactive Selection	
Hides Rejected Variables	Yes
Status	
Create Time	12/7/18 11:51 AM
Run ID	
Last Error	
Last Status	

After running the Variable Clustering, we get the following results.



The Dendrogram window shows the hierarchical nature of the variable clusters.



The Selected Variables window shows one input for each cluster, chosen according to the 1-R2 ratio. These variables are the candidates for Logistic Regression.

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS10	CLUS10	Cluster 10		1CLUS1	0.086789	ClusterComp	0	NO
CLUS10	DOLLARQ10	tot \$ 95Q2	0.880676	CLUS1	0.077855	Variable	0.129399	YES
CLUS10	TOTORDQ10	tot orders 9...	0.880676	CLUS32	0.080972	Variable	0.129837	NO
CLUS11	CLUS11	Cluster 11		1CLUS1	0.047316	ClusterComp	0	NO
CLUS11	DOLLARQ03	tot \$ 93Q3	0.881598	CLUS1	0.04143	Variable	0.123519	YES
CLUS11	TOTORDQ03	tot orders 9...	0.881598	CLUS32	0.043563	Variable	0.123794	NO
CLUS12	CLUS12	Cluster 12		1CLUS1	0.052825	ClusterComp	0	NO
CLUS12	TOTORDQ22	tot orders 9...	0.863458	CLUS7	0.043322	Variable	0.142725	YES
CLUS12	DOLLARQ22	tot \$ 98Q2	0.863458	CLUS1	0.048496	Variable	0.143501	NO
CLUS13	CLUS13	Cluster 13		1CLUS1	0.101836	ClusterComp	0	NO
CLUS13	TOTORDQ08	tot orders 9...	0.878785	CLUS1	0.086547	Variable	0.132699	YES
CLUS13	DOLLARQ06	tot \$ 94Q2	0.878785	CLUS1	0.092487	Variable	0.133568	NO
CLUS14	CLUS14	Cluster 14		1CLUS1	0.053019	ClusterComp	0	NO
CLUS14	TOTORDQ19	tot orders 9...	0.874679	CLUS1	0.042882	Variable	0.130936	YES
CLUS14	DOLLARQ19	tot \$ 97Q3	0.874679	CLUS1	0.050004	Variable	0.131918	NO
CLUS15	CLUS15	Cluster 15		1CLUS1	0.082236	ClusterComp	0	NO
CLUS15	TOTORDQ11	tot orders 9...	0.872366	CLUS32	0.074455	Variable	0.137902	YES
CLUS15	DOLLARQ11	tot \$ 95Q3	0.872366	CLUS1	0.075391	Variable	0.138041	NO
CLUS16	CLUS16	Cluster 16		1CLUS1	0.063901	ClusterComp	0	NO
CLUS16	DOLLARQ04	tot \$ 93Q4	0.876755	CLUS1	0.054508	Variable	0.13035	YES
CLUS16	TOTORDQ04	tot orders 9...	0.876755	CLUS6	0.067273	Variable	0.132134	NO
CLUS17	CLUS17	Cluster 17		1CLUS1	0.094639	ClusterComp	0	NO
CLUS17	TOTORDQ05	tot orders 9...	0.871852	CLUS32	0.082009	Variable	0.139596	YES
CLUS17	DOLLARQ05	tot \$ 94Q1	0.871852	CLUS1	0.083149	Variable	0.13977	NO
CLUS18	CLUS18	Cluster 18		1CLUS1	0.087871	ClusterComp	0	NO
CLUS18	DOLLARQ16	tot \$ 96Q4	0.866462	CLUS1	0.084269	Variable	0.145827	YES
CLUS18	TOTORDQ16	tot orders 9...	0.866462	CLUS6	0.087387	Variable	0.146325	NO
CLUS19	CLUS19	Cluster 19		1CLUS1	0.06765	ClusterComp	0	NO
CLUS19	TOTORDQ18	tot orders 9...	0.879799	CLUS1	0.058933	Variable	0.127729	YES
CLUS19	DOLLARQ18	tot \$ 97Q3	0.879799	CLUS1	0.068496	Variable	0.127729	NO

The Variable Frequency Table window reports how many inputs fall in each cluster.

Cluster	Frequency Count	Percent of Total Frequency
CLUS1	8	8.333333
CLUS10	2	2.083333
CLUS11	2	2.083333
CLUS12	2	2.083333
CLUS13	2	2.083333
CLUS14	2	2.083333
CLUS15	2	2.083333
CLUS16	2	2.083333
CLUS17	2	2.083333
CLUS18	2	2.083333
CLUS19	2	2.083333
CLUS2	3	3.125
CLUS20	2	2.083333
CLUS21	2	2.083333
CLUS22	2	2.083333
CLUS23	2	2.083333
CLUS24	2	2.083333
CLUS25	2	2.083333
CLUS26	2	2.083333
CLUS27	2	2.083333
CLUS28	3	3.125
CLUS29	5	5.208333
CLUS3	3	3.125
CLUS30	1	1.041667
CLUS31	6	6.25
CLUS32	8	8.333333
CLUS33	1	1.041667
CLUS34	2	2.083333
CLUS35	1	1.041667
CLUS4	3	3.125
CLUS5	3	3.125
CLUS6	4	4.166667

The bottom of the results in the Output window shows the complete list of which variables were in each cluster. Variables in the same cluster were similar in the analysis. The procedure selects the variable with the lowest 1-R2 ratio as the cluster representative.

Output					
5165	35	1	1	1	1.0000
5166					
5167	Total variation explained = 68.64175 Proportion = 0.7150				
5168					
5169					
5170					
5171	35 Clusters	R-squared with			
5172			Own	Next	1-R**2
5173	Cluster	Variable	Cluster	Closest	Ratio
5174					Variable
5175	Cluster 1	DEPT14	0.4011	0.1668	0.7188
5176		DEPT15	0.1769	0.0788	0.8935
5177		DEPT16	0.1757	0.0805	0.8965
5178		DEPT17	0.1476	0.0563	0.9032
5179		DOLINDET	0.9239	0.5295	0.1619
5180		DOLNETDT	0.9163	0.5244	0.1760
5181		FREQPRCH	0.7492	0.5167	0.5190
5182		UNITSIDD	0.8436	0.6278	0.4202
5183					
5184	Cluster 2	CCPAYMO	1.0000	0.3108	0.0000
5185		CCPAYM1	1.0000	0.3108	0.0000
5186		METHPAYMCC	1.0000	0.3108	0.0000
5187					
5188	Cluster 3	DOLL24	0.4767	0.3047	0.7526
5189		DOLLARQ17	0.8124	0.0787	0.2037
5190		TOTORDQ17	0.7392	0.0625	0.2782
5191					
5192	Cluster 4	DOLINDEA	0.9129	0.2620	0.1181
5193		DOLNETDA	0.9039	0.2546	0.1290
5194		UNITSLAP	0.4720	0.1137	0.5957
5195					
5196	Cluster 5	BOTHPAYMO	1.0000	0.1716	0.0000

There are 35 clusters and, therefore, 35 variables selected.

The last table in the Output window shows a summary of the final cluster solution.

Output							
	Number	Explained	Variation	Explained	Eigenvalue	R-squared	Ratio
	of	by	Explained	by a	in a	for a	for a
	Clusters	Clusters	by Clusters	Cluster	Cluster	Variable	Variable
5311							
5312							
5313							
5314							
5315	1	14.269655	0.1486	0.1486	5.075727	0.0043	
5316	2	18.878088	0.1966	0.1578	3.699895	0.0106	0.9916
5317	3	21.903078	0.2282	0.1877	2.748968	0.0111	0.9918
5318	4	24.167157	0.2517	0.1877	2.623513	0.0111	0.9916
5319	5	26.705701	0.2782	0.1893	2.073863	0.0113	0.9914
5320	6	28.466963	0.2965	0.1958	2.023150	0.0113	0.9927
5321	7	30.338115	0.3160	0.1958	1.906582	0.0113	1.0351
5322	8	31.993414	0.3333	0.2145	1.829092	0.0117	1.0203
5323	9	33.791416	0.3520	0.2145	1.734839	0.0117	1.1004
5324	10	35.400976	0.3688	0.2252	1.661294	0.0119	1.0955
5325	11	37.045861	0.3859	0.2333	1.645777	0.0120	1.0950
5326	12	38.514796	0.4012	0.2333	1.641035	0.0120	1.0950
5327	13	40.075066	0.4174	0.2462	1.578818	0.0122	1.0903
5328	14	41.568625	0.4330	0.2462	1.555217	0.0122	1.0903
5329	15	43.111311	0.4491	0.2462	1.548926	0.0122	1.0903
5330	16	44.584612	0.4644	0.2462	1.545298	0.0122	1.0903
5331	17	46.100224	0.4802	0.2538	1.537361	0.0124	1.0892
5332	18	47.580873	0.4956	0.2538	1.522834	0.0124	1.0892
5333	19	48.978020	0.5102	0.2538	1.518141	0.0124	1.0892
5334	20	50.489323	0.5259	0.2538	1.507744	0.0124	1.0892
5335	21	51.997062	0.5416	0.2538	1.501807	0.0124	1.0892
5336	22	53.498395	0.5573	0.2538	1.499339	0.0124	1.0892
5337	23	54.982973	0.5727	0.2615	1.482932	0.0126	1.0875
5338	24	56.453832	0.5881	0.2693	1.480621	0.0125	1.0875
5339	25	57.934102	0.6035	0.2693	1.474760	0.0125	1.0875
5340	26	59.408841	0.6188	0.2693	1.400954	0.0125	1.0875
5341	27	60.681938	0.6321	0.2693	1.327516	0.0125	1.0875

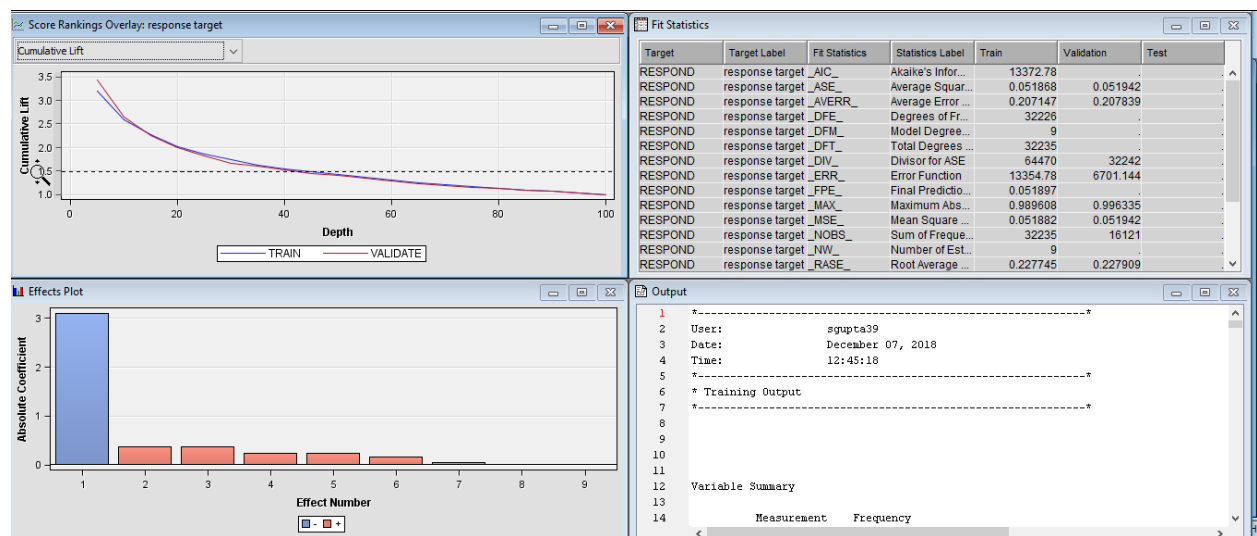
The clusters explained 71.5% of the variation in the data.

The next step in the model-building process is to eliminate the irrelevant predictor variables. This can be accomplished using the Regression node in SAS Enterprise Miner. The Regression node can create several types of regression models, including linear and logistic. The type of default regression type is determined by the target's measurement level.

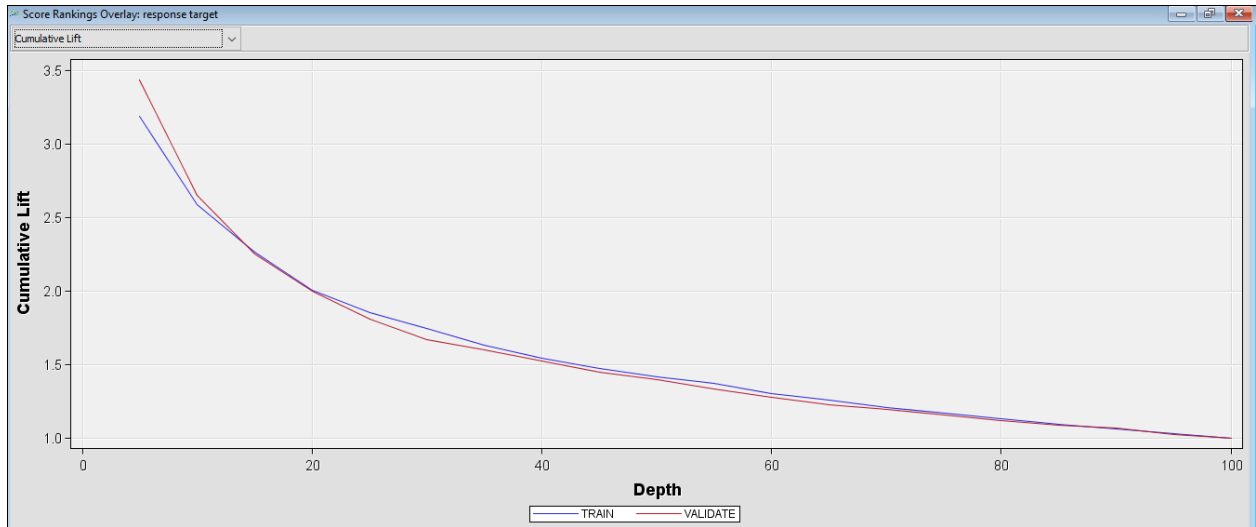
After placing the regression model in the diagram, we made a few changes to the model selection attribute.

Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	

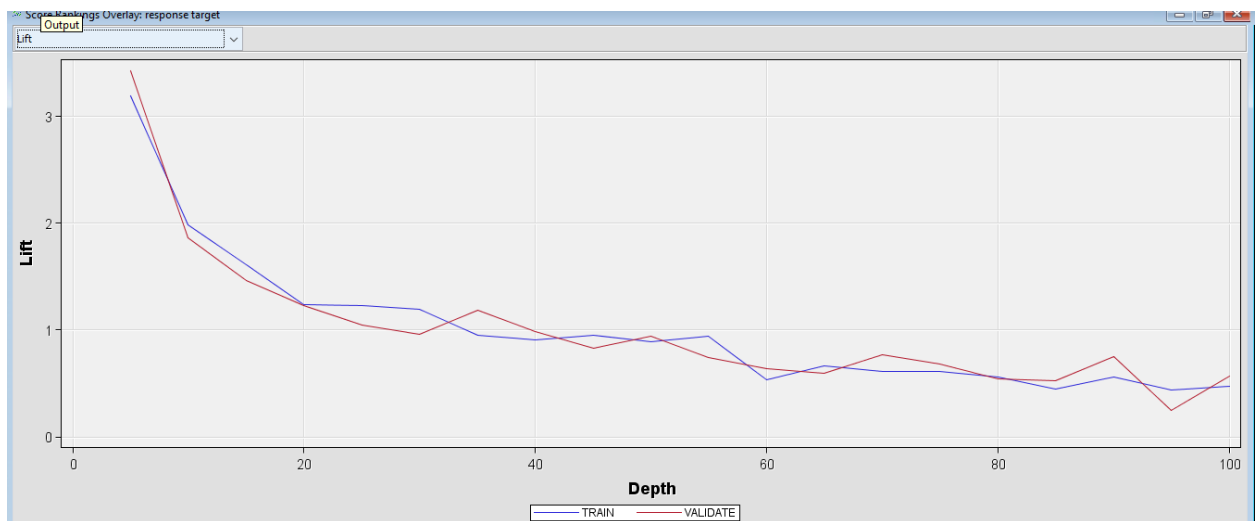
The Results window contains four sub-windows: Score Rankings Overlay, Fit Statistics, Effects Plot, and Output.



The Score Rankings Overlay window shows a cumulative lift chart where, for a given percentile, see the lift of the model can be seen.

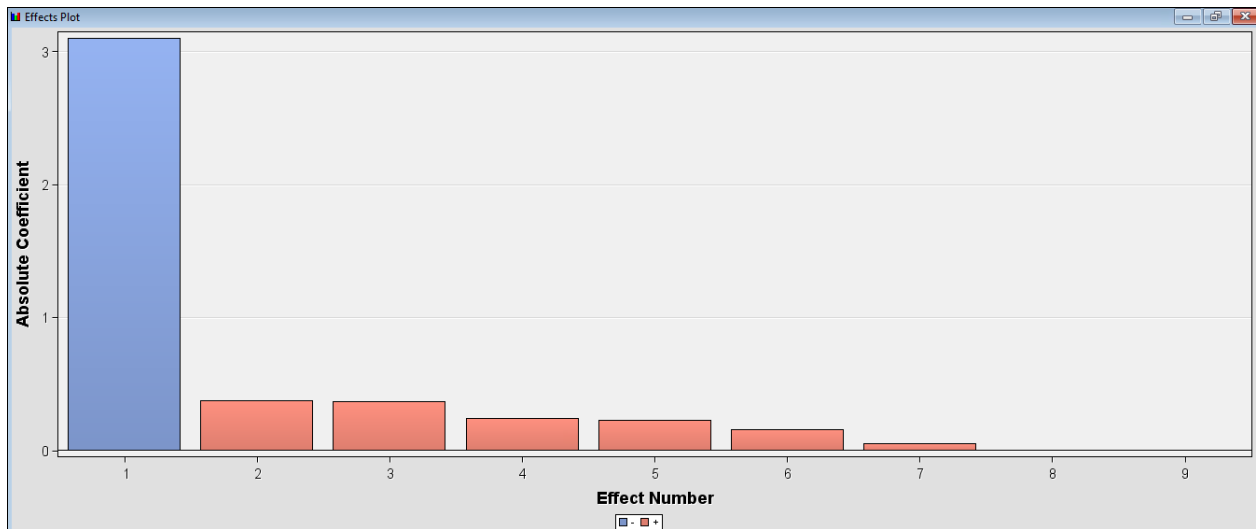


The lift of the model can also be seen from the results. The **lift** is a measure of the performance of a targeting **model** (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting **model**. When evaluating machine learning models there is a plethora of possible metrics to assess performance. There are things like accuracy, precision-recall, ROC curve and so on. All of them can be useful, but they can also be misleading or don't answer the question at hand very well. Lift, on the other hand, is an ideal measure of evaluating the performance and quality of a machine learning model.



By positioning the mouse cursor over a point along the lift curve for the validation data, we can see a pop-up flag with information about the percentile and lift. For example, at the 5th percentile, the lift is 3.43 on the validation data set. This means that if the catalog company mailed to the top 5 percent of its customers based on the predicted probabilities, then we would obtain 3.43 times more responders compared to a 5-percent random sample of the customers.

The Effects Plot window shows a bar chart of the absolute values of the coefficients in the final model. The bars are color-coded to indicate the algebraic signs of the coefficients.



The Fit Statistics window shows a table of model fit statistics. If the decision predictions are of interest, model fit can be judged by misclassification. If estimate predictions are the focus, model fit can be assessed by average squared error. If there is a large discrepancy between the values of these two statistics on the training and validation data sets, then there is evidence of overfitting the model.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RESPOND	response target	_AIC_	Akaike's Information Criterion	13372.78		
RESPOND	response target	_ASE_	Average Squared Error	0.051868		0.051942
RESPOND	response target	_AVERR_	Average Error Function	0.207147		0.207839
RESPOND	response target	_DFE_	Degrees of Freedom for Error	32226		
RESPOND	response target	_DFM_	Model Degrees of Freedom	9		
RESPOND	response target	_DFT_	Total Degrees of Freedom	32235		
RESPOND	response target	_DIV_	Divisor for ASE	64470		32242
RESPOND	response target	_ERR_	Error Function	13354.78		6701.144
RESPOND	response target	_FPE_	Final Prediction Error	0.051897		
RESPOND	response target	_MAX_	Maximum Absolute Error	0.989608		0.986335
RESPOND	response target	_MSE_	Mean Square Error	0.051882		0.051942
RESPOND	response target	_NOBS_	Sum of Frequencies	32235		16121
RESPOND	response target	_NW_	Number of Estimate Weights	9		
RESPOND	response target	_RASE_	Root Average Sum of Squares	0.227745		0.227909
RESPOND	response target	_RFPE_	Root Final Prediction Error	0.227809		
RESPOND	response target	_RMSE_	Root Mean Squared Error	0.227777		0.227909
RESPOND	response target	_SBC_	Schwarz's Bayesian Criterion	13448.21		
RESPOND	response target	_SSE_	Sum of Squared Errors	3343.919		1674.728
RESPOND	response target	_SUMW_	Sum of Case Weights Times Freq	64470		32242
RESPOND	response target	_MISC_	Misclassification Rate	0.056678		0.056758

The Output window gives the standard output for logistic regression.

34	Predicted and decision variables		
35			
36	Type	Variable	Label
37			
38	TARGET	RESPOND	response target
39	PREDICTED	P_RESPOND1	Predicted: RESPOND=1
40	RESIDUAL	R_RESPOND1	Residual: RESPOND=1
41	PREDICTED	P_RESPOND0	Predicted: RESPOND=0
42	RESIDUAL	R_RESPOND0	Residual: RESPOND=0
43	FROM	F_RESPOND	From: RESPOND
44	INTO	I_RESPOND	Into: RESPOND
45			
46			
47			
48			
49			
50	The DMREG Procedure		
51			
52	Model Information		
53			
54	Training Data Set	WORK.REG_DMREG.VIEW	
55	DMDB Catalog	WORK.REG_DMDB	
56	Target Variable	RESPOND (response target)	
57	Target Measurement Level	Ordinal	
58	Number of Target Categories	2	
59	Error	Bernoulli	
60	Link Function	Logit	
61	Number of Model Parameters	36	
62	Number of Observations	32235	
63			
64			
65	Target Profile		

12	Variable Summary		
13			
14		Measurement	Frequency
15	Role	Level	Count
16			
17	INPUT	INTERVAL	35
18	TARGET	BINARY	1
19			
20			

The Model Information table shows the training data set name, the target variable name, the number of target categories, the number of model parameters, and the number of observations. The Target Profile table shows the number of observations for each target category.

The DMREG Procedure

Model Information

Training Data Set	WORK.EM_DMREG.VIEW
DMDB Catalog	WORK.REG_DMDB
Target Variable	RESPOND (response target)
Target Measurement Level	Ordinal
Number of Target Categories	2
Error	MBernoulli
Link Function	Logit
Number of Model Parameters	36
Number of Observations	32235

The output of the forward selection method shows the results of each model fitted in each step. The Summary of Forward Selection table shows the variables that were selected in the forward selection method. This model has 13 inputs.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	DOLINDET	1	1	418.2287	<.0001	6878.3
2	TOTORDQ20	1	2	178.2565	<.0001	6847.2
3	MONLAST	1	3	113.3610	<.0001	6781.4
4	TOTORDQ22	1	4	47.4870	<.0001	6751.4
5	CATALOGCNT	1	5	36.9828	<.0001	6727.5
6	TOTORDQ18	1	6	19.9779	<.0001	6719.0
7	TOTORDQ21	1	7	14.9769	0.0001	6712.7
8	TOTORDQ12	1	8	13.5709	0.0002	6701.1
9	TOTORDQ19	1	9	11.8344	0.0006	6702.1
10	DEPT03	1	10	10.4403	0.0012	6701.4
11	CCPAYMO	1	11	9.3003	0.0023	6709.1
12	TOTORDQ05	1	12	6.4600	0.0110	6716.5
13	DOLLARQ09	1	13	5.3211	0.0211	6717.9

The likelihood ratio test tests the null hypothesis that all regression coefficients of the model are 0. A significant p -value for the likelihood ratio (for this example, the p -value is less than .0001) provides evidence that at least one of the regression

coefficients for an explanatory variable is nonzero. The final model contains 8 terms plus an intercept.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0				
-2 Log Likelihood Intercept Only	Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
14025.546	13354.783	670.7623	8	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-3.1029	0.0576	2903.87	<.0001		0.045
CATALOGCNT	1	0.0529	0.0101	27.45	<.0001	0.0912	1.054
DOLINDET	1	0.000109	0.000081	1.80	0.1800	0.0189	1.000
MONLAST	1	-0.00586	0.000931	39.59	<.0001	-0.1300	0.994
TOTORDQ12	1	0.1614	0.0461	12.28	0.0005	0.0363	1.175
TOTORDQ18	1	0.2438	0.0581	17.62	<.0001	0.0440	1.276
TOTORDQ20	1	0.3782	0.0427	78.56	<.0001	0.0963	1.460
TOTORDQ21	1	0.2291	0.0583	15.43	<.0001	0.0417	1.257
TOTORDQ22	1	0.3705	0.0580	40.79	<.0001	0.0642	1.448

The parameter estimates measure the rate of change in the logit (log of the odds) corresponding to a one-unit change in the predictor variable, adjusted for the effects of the other predictors. For example, a one-unit change in **CATALOGCNT** (number of catalogs received) corresponds to a .054 increase in the log odds of purchasing a product from the catalog, adjusted for the other predictor variables.

Odds Ratio Estimates	
Effect	Point Estimate
CATALOGCNT	1.054
DOLINDET	1.000
MONLAST	0.994
TOTORDQ12	1.175
TOTORDQ18	1.276
TOTORDQ20	1.460
TOTORDQ21	1.257
TOTORDQ22	1.448

The Wald chi-square and its associated p -value test whether the parameter estimate is significantly different from 0.

The parameter estimates cannot generally be compared across different variables because the coefficients depend directly on the units the variable was measured in. One solution is to use standardized estimates, which convert the parameter estimates into standard deviation units. The absolute value of the standardized estimates can be used to give an approximate ranking of the relative importance of the predictor variables. Therefore, **MONLAST** (months since last purchase) is the most important predictor variable followed by **TOTORDQ20** (total orders in the fourth quarter of 1997) and **CATALOGCNT** (number of catalogs received).

The odds ratio measures the effect of the predictor variable on the outcome, adjusted for the effects of the other predictor variables. For example, an increase of one month since the last order was placed yields a .6% decrease in the odds of purchasing a product from the catalog (calculated as $100(0.994 - 1)$). This might not be as meaningful on a month-by-month basis, so computed as years, it translates to a 7.2% decrease in the odds of responding for every year increase since the last purchase. Furthermore, a one-catalog increase in the number of catalogs received yields a 5.4% increase in the odds of purchasing a product.

The output also shows the assessment statistics for the validation data set for the 5th percentile, the 10th percentile, and so on.

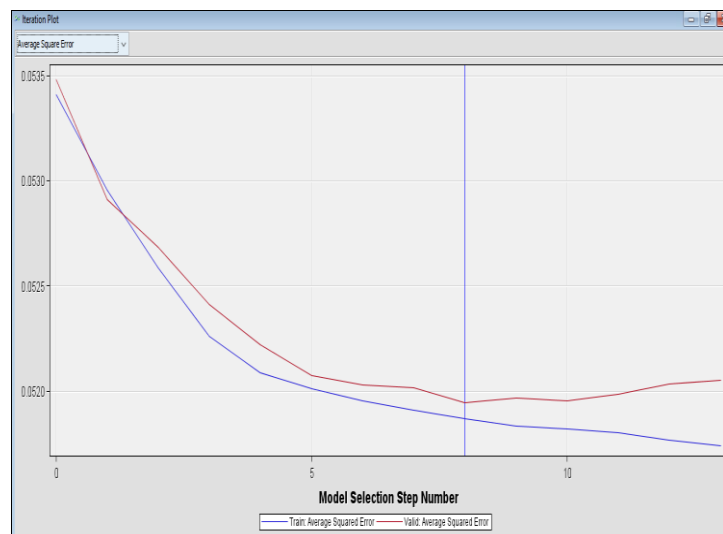
Data Role=VALIDATE Target Variable=RESPOND Target Label=response target							
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	243.140	3.43140	3.43140	19.4548	19.4548	807	0.18015
10	164.623	1.86007	2.64623	10.5459	15.0031	806	0.10221
15	125.304	1.46617	2.25304	8.3127	12.7739	806	0.08419
20	99.622	1.22546	1.99622	6.9479	11.3178	806	0.07390
25	80.710	1.05039	1.80710	5.9553	10.2456	806	0.06746
30	66.643	0.96286	1.66643	5.4591	9.4480	806	0.06347
35	59.719	1.18169	1.59719	6.6998	9.0555	806	0.05996
40	52.065	0.98474	1.52065	5.5831	8.6215	806	0.05518
45	44.409	0.83156	1.44409	4.7146	8.1875	806	0.05136
50	39.379	0.94098	1.39379	5.3350	7.9022	806	0.04757
55	33.472	0.74403	1.33472	4.2184	7.5674	806	0.04496
60	27.639	0.63461	1.27639	3.5980	7.2366	806	0.04298
65	22.366	0.59085	1.22366	3.3499	6.9377	806	0.04116
70	19.097	0.76591	1.19097	4.3424	6.7523	806	0.03973
75	15.680	0.67838	1.15680	3.8462	6.5586	806	0.03797
80	11.869	0.54708	1.11869	3.1017	6.3426	806	0.03586
85	8.378	0.52520	1.08378	2.9777	6.1446	806	0.03348
90	6.552	0.75497	1.06552	4.2804	6.0411	806	0.03016
95	2.268	0.25166	1.02268	1.4268	5.7982	806	0.02514
100	0.000	0.56896	1.00000	3.2258	5.6696	806	0.01863

Another useful table in the output shows the distribution of the posterior probabilities for the validation data set.

Assessment Score Distribution				
Data Role=TRAIN Target Variable=RESPOND Target Label=response targ				
Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	3	0	0.99536	0.0093
0.90-0.95	0	1	0.90364	0.0031
0.85-0.90	1	0	0.88983	0.0031
0.80-0.85	2	0	0.82147	0.0062
0.70-0.75	1	2	0.70692	0.0093
0.65-0.70	1	2	0.66852	0.0093
0.60-0.65	1	2	0.61125	0.0093
0.55-0.60	3	1	0.57665	0.0124
0.50-0.55	1	7	0.52110	0.0248
0.45-0.50	6	8	0.48451	0.0434
0.40-0.45	10	13	0.42256	0.0714
0.35-0.40	5	14	0.37218	0.0589
0.30-0.35	9	30	0.32593	0.1210
0.25-0.30	23	50	0.27418	0.2265
0.20-0.25	44	131	0.22092	0.5429
0.15-0.20	68	348	0.16967	1.2905
0.10-0.15	218	1465	0.11861	5.2210
0.05-0.10	793	10821	0.06718	36.0292
0.00-0.05	636	17515	0.03640	56.3084

Finally, we want to check the model performance across the fitted models.

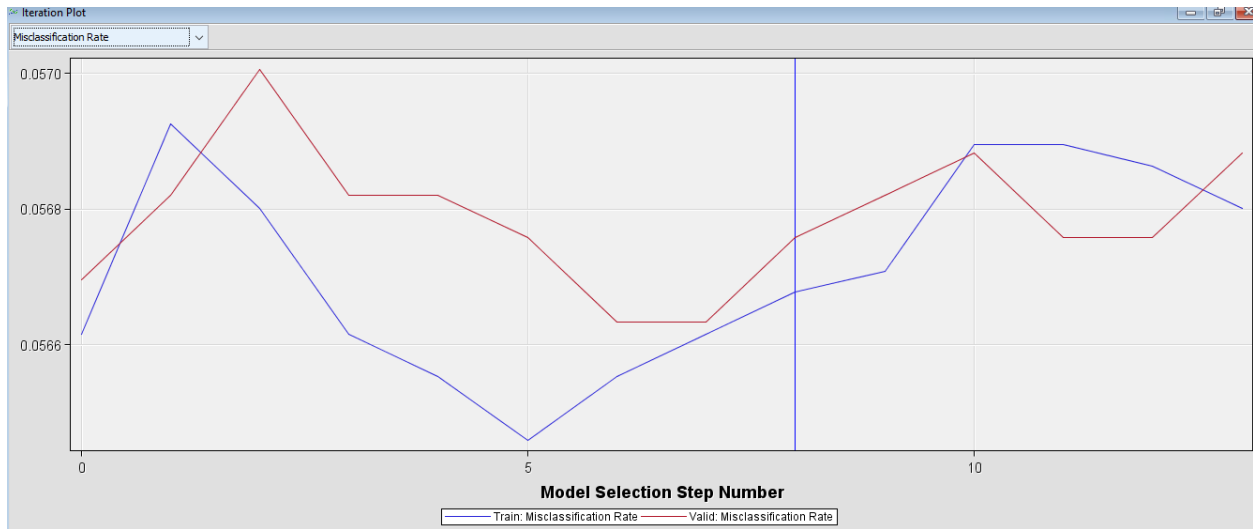
The Model Iteration Plot



The Iteration Plot window shows the average squared error (training and validation) from the model selected in each step of the backward selection process.

The smallest average squared error occurs in model 8.

Misclassification Rate

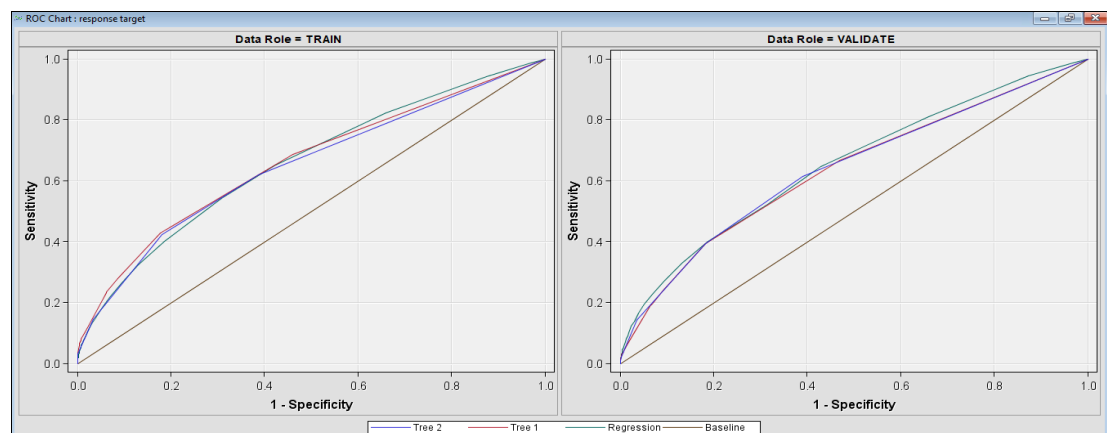
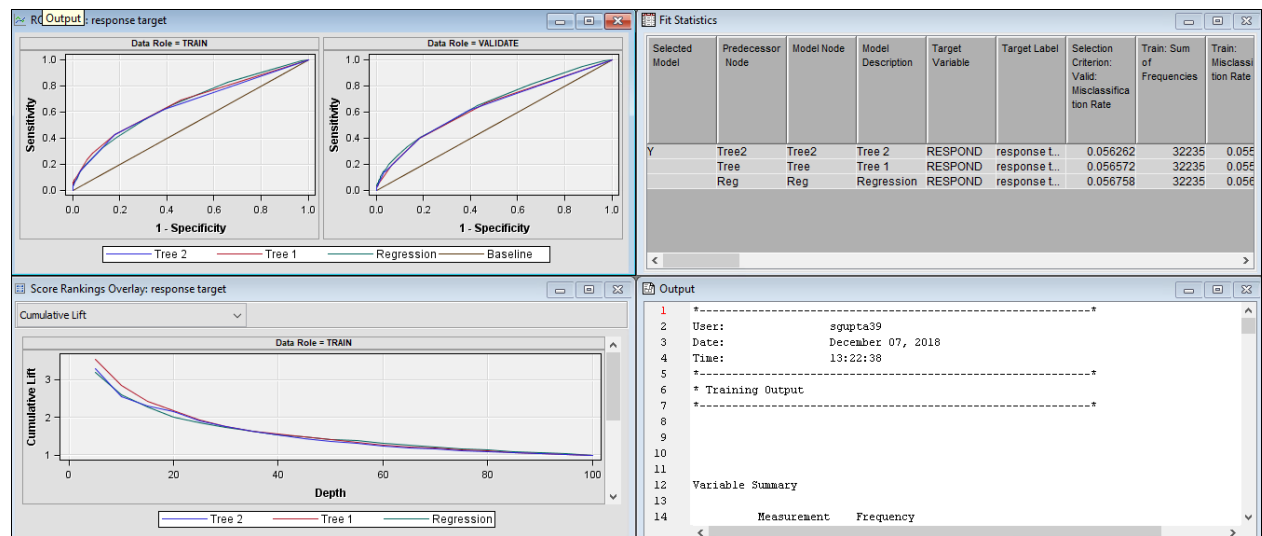


The iteration plot shows that the model with the smallest misclassification rate occurs in steps 6 and 7.

To compute a ROC curve, the Model Comparison node must be used. This node also is used later to collect assessment information from other modeling nodes and to compare model performance measures.

The Results window contains four sub-windows: ROC chart, Score Rankings Overlay, Fit Statistics, and Output.

The ROC chart window shows that two of the three models have good predictive accuracy as the ROC curves deviate from the 45% angle. The logistic regression and Tree 2 models perform similarly on the validation data set. The logistic regression performs slightly better. The Score Rankings Overlay window illustrates the cumulative lift chart for the training and validation data sets



The Fit Statistics window shows the model fit statistics for the training and validation data sets.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error
Y	Tree2	Tree2	Tree 2	RESPOND	response t...	0.056262	32235	0.055995	0.963901	3313.256	0.051392	0.226699	64470	32235	16121	0.056262	0.963
	Tree	Tree	Tree 1	RESPOND	response t...	0.056572	32235	0.055592	0.966231	3277.161	0.050832	0.22546	64470	32235	16121	0.056572	0.966
	Reg	Reg	Regression	RESPOND	response t...	0.056758	32235	0.056678	0.989608	3343.919	0.051868	0.227745	64470	32235	16121	0.056758	0.996

The Output window also shows various fit statistics for the selected models.

Target: RESPOND			
Data Role=Train			
Statistics	Tree2	Tree	Reg
Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.08	0.07	0.06
Train: Kolmogorov-Smirnov Statistic	0.24	0.25	0.23
Train: Akaike's Information Criterion	.	.	13372.78
Train: Average Squared Error	0.05	0.05	0.05
Train: Roc Index	0.65	0.66	0.66
Train: Average Error Function	.	.	0.21
Train: Cumulative Percent Captured Response	25.50	28.46	25.86
Train: Percent Captured Response	9.10	10.74	9.92
Selection Criterion: Valid: Misclassification Rate	0.06	0.06	0.06
Train: Degrees of Freedom for Error	.	.	32226.00
Train: Model Degrees of Freedom	.	.	9.00
Train: Total Degrees of Freedom	32235.00	32235.00	32235.00
Train: Divisor for ASE	64470.00	64470.00	64470.00
Train: Error Function	.	.	13354.78
Train: Final Prediction Error	.	.	0.05
Train: Gain	154.92	184.59	158.59
Train: Gini Coefficient	0.30	0.32	0.32
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.24	0.25	0.24
Train: Kolmogorov-Smirnov Probability Cutoff	0.06	0.06	0.06
Train: Cumulative Lift	2.55	2.85	2.59
Train: Lift	1.82	2.15	1.98
Train: Maximum Absolute Error	0.96	0.97	0.99
Train: Misclassification Rate	0.06	0.06	0.06
Train: Mean Square Error	.	.	0.05
Train: Sum of Frequencies	32235.00	32235.00	32235.00

