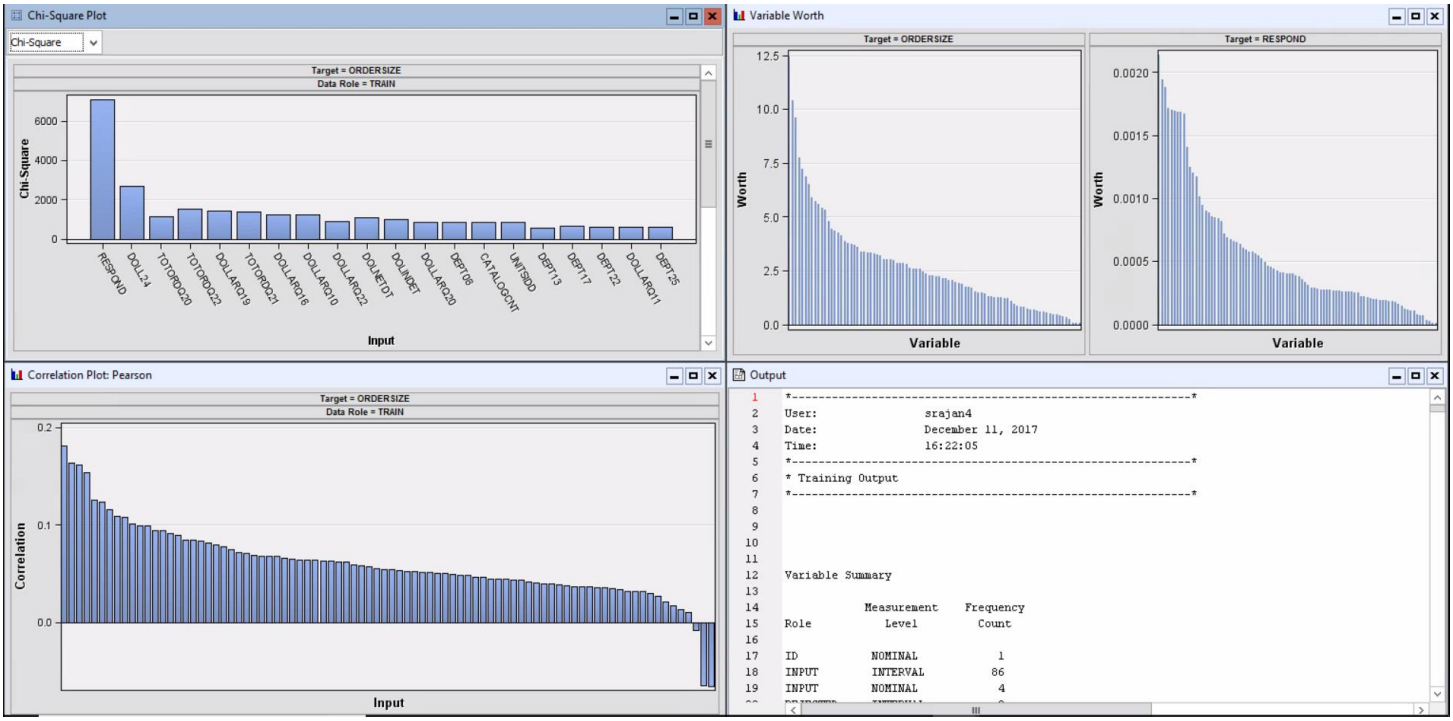
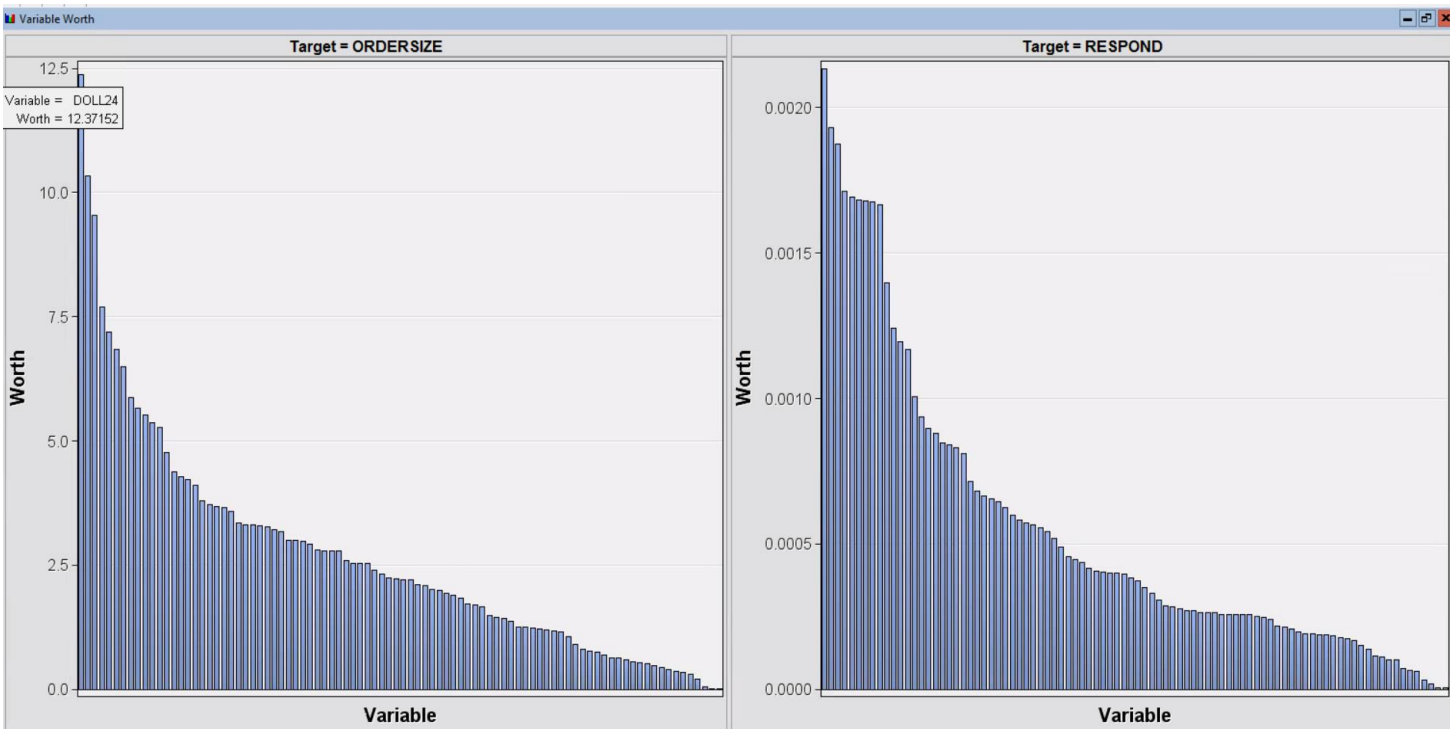
Submitted By: Harika Katragadda and Sonika Rajan

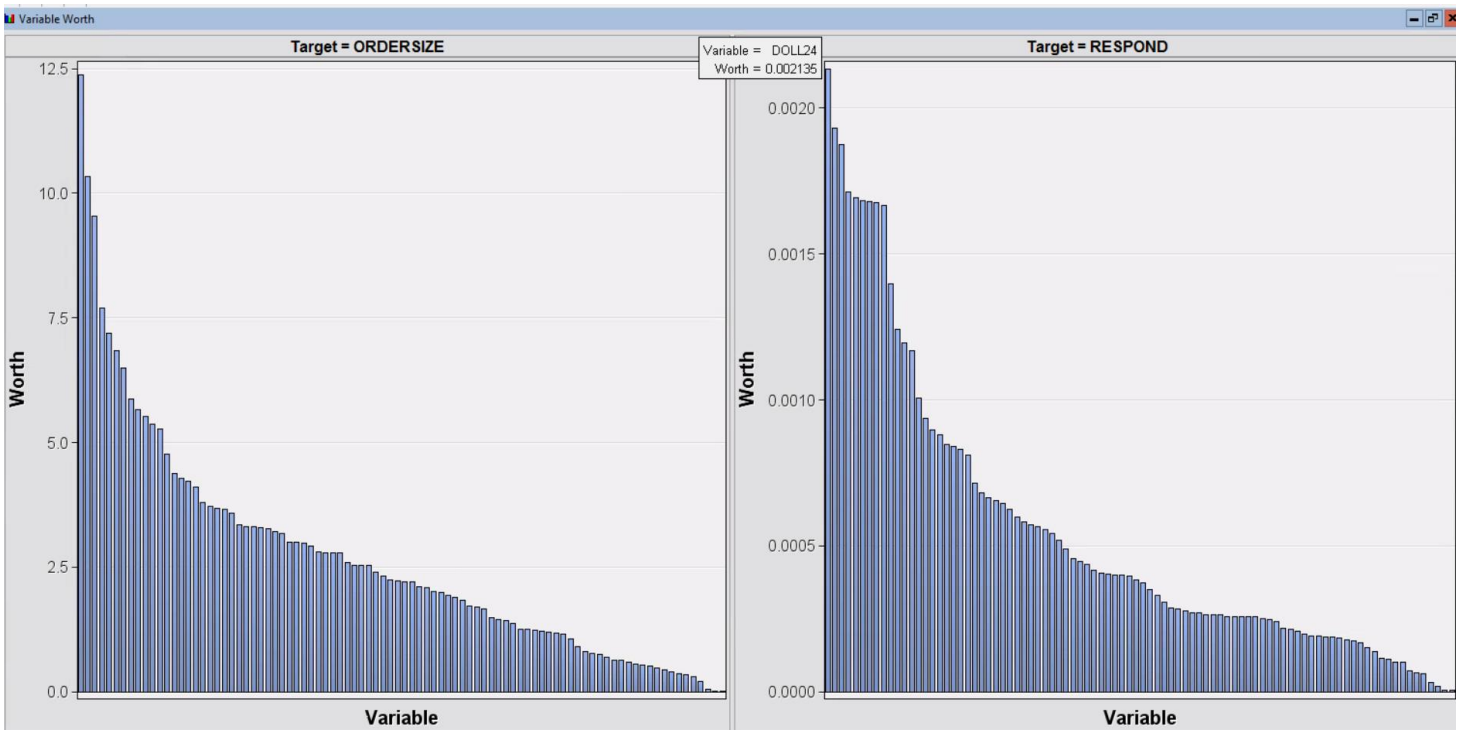
TASK 1

Node Stat Explore can be used to generate statistical summary of input data.



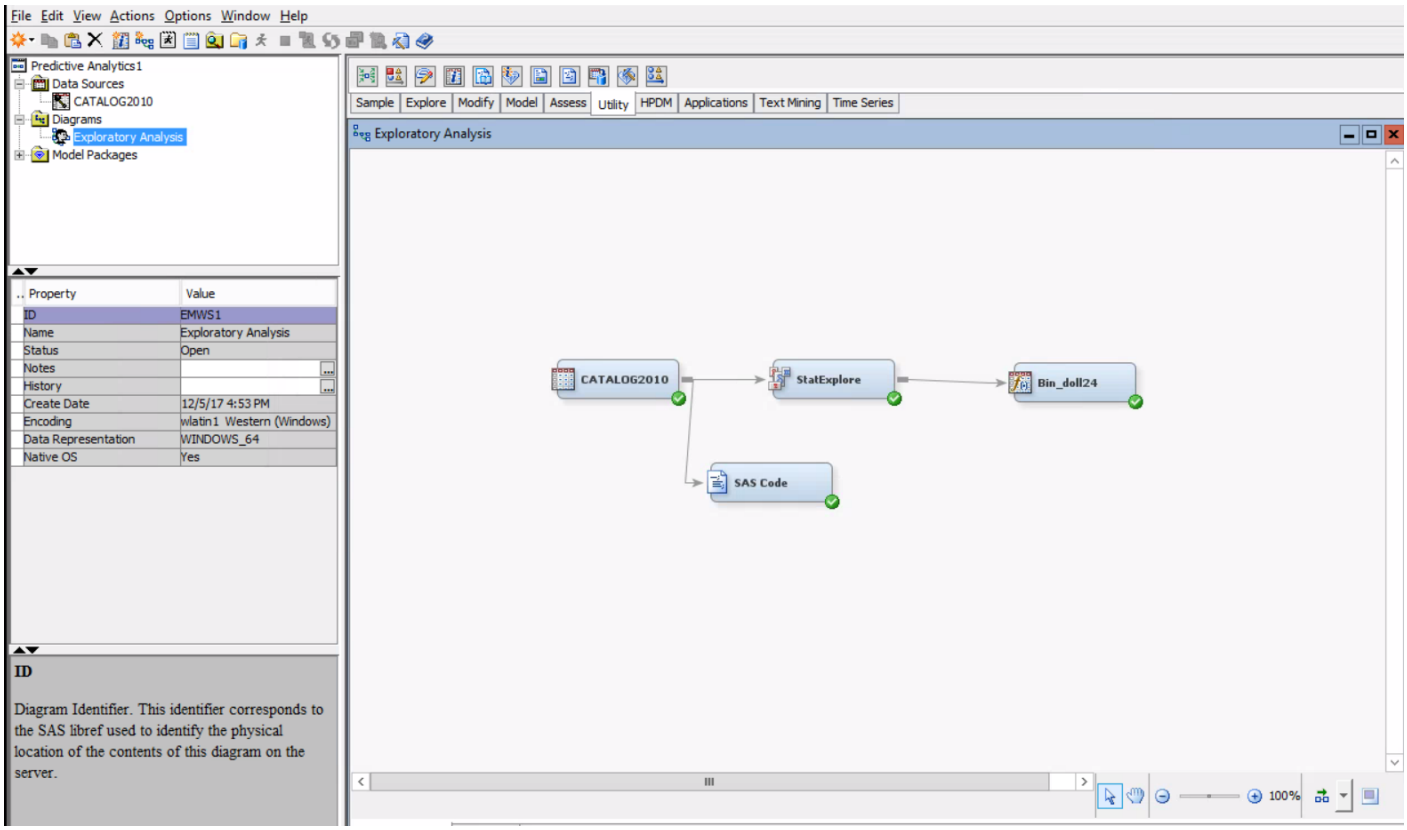
From the above graphs it can be noted that for both the target variables RESPOND and ORDERSIZE, the worth of input variable DOLL24 has the highest value.


Variable Worth
Target = ORDERSIZE
Target = RESPOND
Variable = DOLL24
Worth = 12.37152
Worth
Variable
12.5
10.0
7.5
5.0
2.5
0.0
0.0020
0.0015
0.0010
0.0005
0.0000

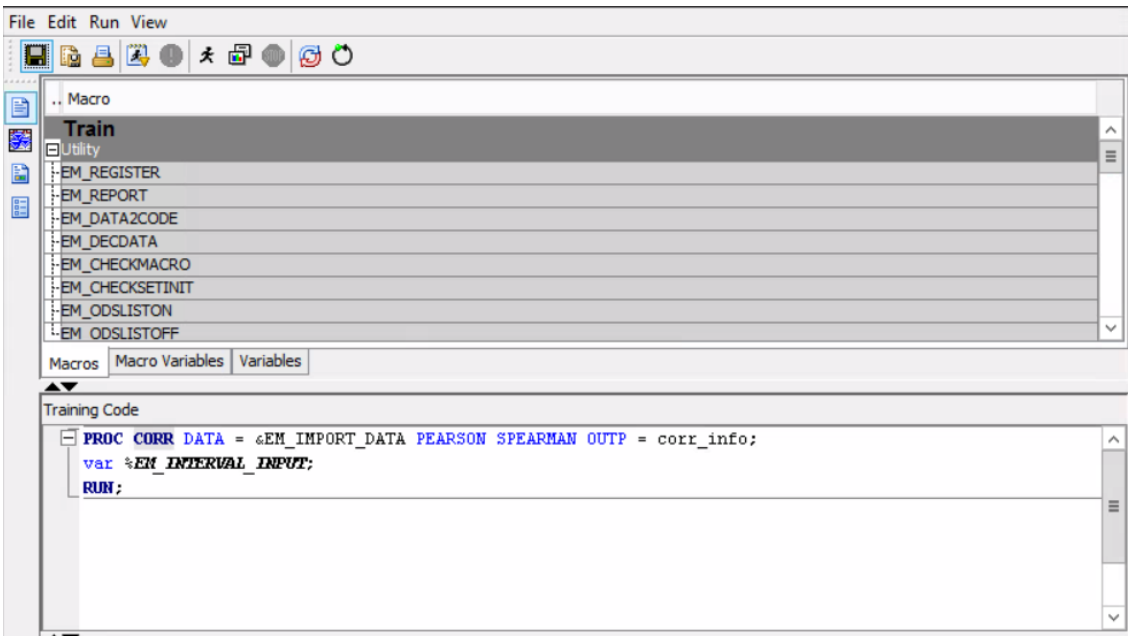

Variable Worth
Target = ORDERSIZE
Target = RESPOND
Variable = DOLL24
Worth = 0.002135
Worth
Variable
12.5
10.0
7.5
5.0
2.5
0.0
0.0020
0.0015
0.0010
0.0005
0.0000

Variables with high correlation

We can use SAS Code Node under Utility Class to determine correlation between variables



Below code is used to find correlation using SAS Code



Below screenshot gives the sample on correlations and other characteristics from output obtained while running the code.

Output

Prob > |r| under H0: Rho=0

	ACTBUY	BUYPROP	CATALOGCNT	DAYLAST	DEPT01	DEPT02	DEPT03	DEPT04	DEPT05	DEPT06	DEPT07	DEPT08	DEPT09	DEPT10	DE
ACTBUY	1.00000	0.84916	0.48067	-0.18504	0.16840	0.18164	0.19629	0.21167	0.18389	0.22534	0.06598	0.17186	0.12584	0.16569	0.1
num qrtrs w/buy		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
BUYPROP	0.84916	1.00000	0.29811	-0.27499	0.10262	0.12469	0.11938	0.14065	0.11340	0.14058	0.04222	0.12330	0.08541	0.11510	0.0
% quarters w/buy	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
CATALOGCNT	0.48067	0.29811	1.00000	-0.33437	0.31743	0.28115	0.39454	0.36075	0.34466	0.42988	0.09462	0.23841	0.20017	0.26286	0.2
number of catalogs received	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DAYLAST	-0.18504	-0.27499	-0.33437	1.00000	-0.13693	-0.10678	-0.17668	-0.14548	-0.13701	-0.15758	-0.02444	-0.11780	-0.09419	-0.12344	-0.0
days since last	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT01	0.16840	0.10262	0.31743	-0.13693	1.00000	0.18712	0.21293	0.17880	0.16937	0.19446	0.06844	0.09858	0.05228	0.07559	0.0
womens apparel	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT02	0.18164	0.12469	0.28115	-0.10678	0.18712	1.00000	0.17365	0.14848	0.15257	0.14815	0.04885	0.15248	0.04164	0.07861	0.0
womens sleepwear	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT03	0.19629	0.11938	0.39454	-0.17668	0.21293	0.17365	1.00000	0.25188	0.17493	0.19532	0.03203	0.08708	0.07917	0.09836	0.0
womens underwear	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT04	0.21167	0.14065	0.36075	-0.14548	0.17880	0.14848	0.25188	1.00000	0.16624	0.18542	0.04580	0.08225	0.06480	0.16989	0.0
womens hosiery	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT05	0.18389	0.11340	0.34466	-0.13701	0.16937	0.15257	0.17493	0.16624	1.00000	0.18806	0.03618	0.08455	0.04765	0.08415	0.1
womens footwear	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.(
DEPT06	0.22534	0.14058	0.42988	-0.15758	0.19446	0.14815	0.19532	0.18542	0.18806	1.00000	0.05332	0.10097	0.07391	0.11832	0.1
womens misc	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.(
DEPT07	0.06598	0.04222	0.09462	-0.02444	0.06844	0.04885	0.03203	0.04580	0.03618	0.05332	1.00000	0.05433	0.05793	0.04748	0.0
mens apparel	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.(
DEPT08	0.17186	0.12330	0.23841	-0.11780	0.09858	0.15248	0.08708	0.08225	0.08455	0.10097	0.05433	1.00000	0.13007	0.10383	0.1
mens sleepwear	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.(
DEPT09	0.12584	0.08541	0.20017	-0.09419	0.05228	0.04164	0.07917	0.06480	0.04765	0.07391	0.05793	0.13007	1.00000	0.14422	0.0
mens underwear	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.(

Correlation values for variables can range from -1 to +1. Closer the coefficients are to +1.0 and -1.0, greater is the strength of the relationship between the variables.

From above analysis, some variables which have high correlation

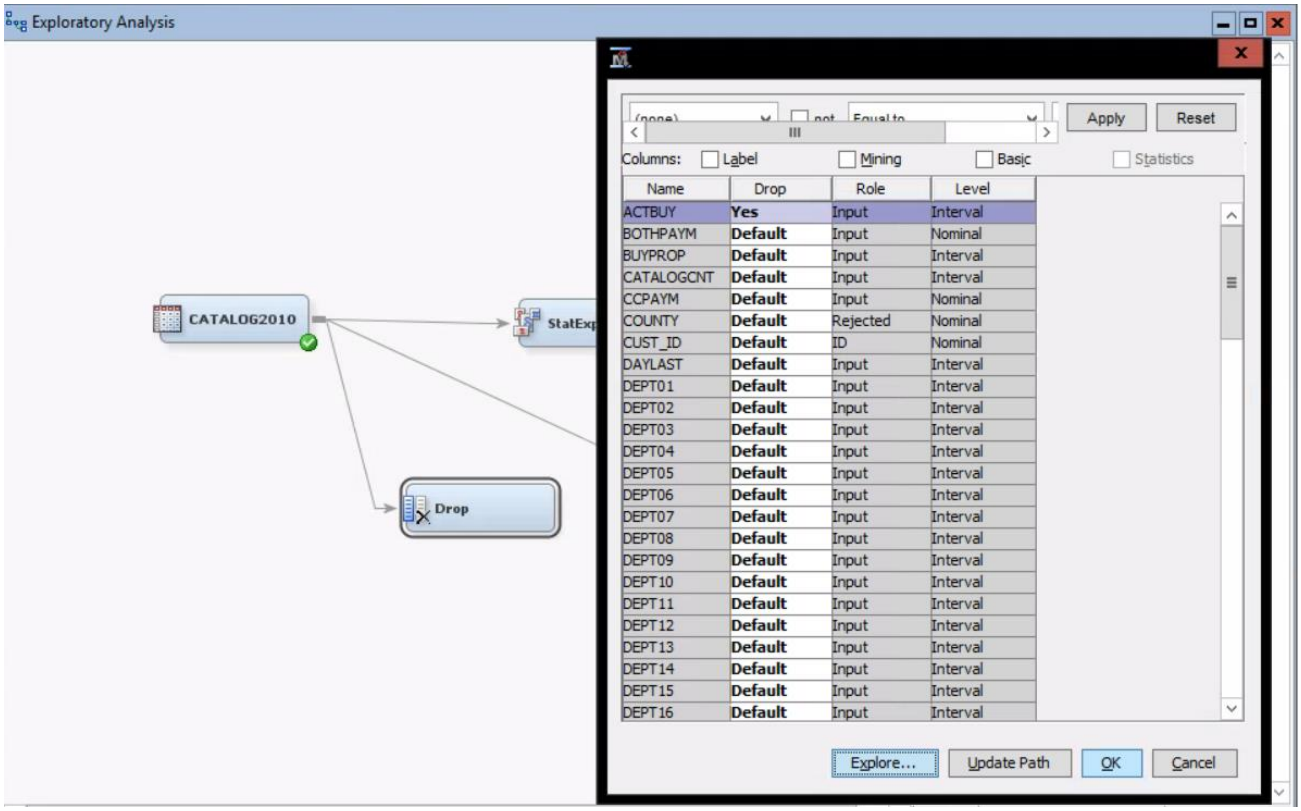
MONLAST (months since last) - 0.99998 – DAYLAST (days since last). Either of them can be calculated from the other.

TOTORDQ03(tot orders 93 q1) - 0.99951 - DOLLARQ03 (tot $ 93 q3). As orders is directly related to dollars, it is obvious that these two variables are correlated.

ACTBUY (number qtrs. w/buy) - 0.84916 – BUYPROP (%qtrs w/buy) – Since one is number and other is percentage on same data, one of them can be dropped.

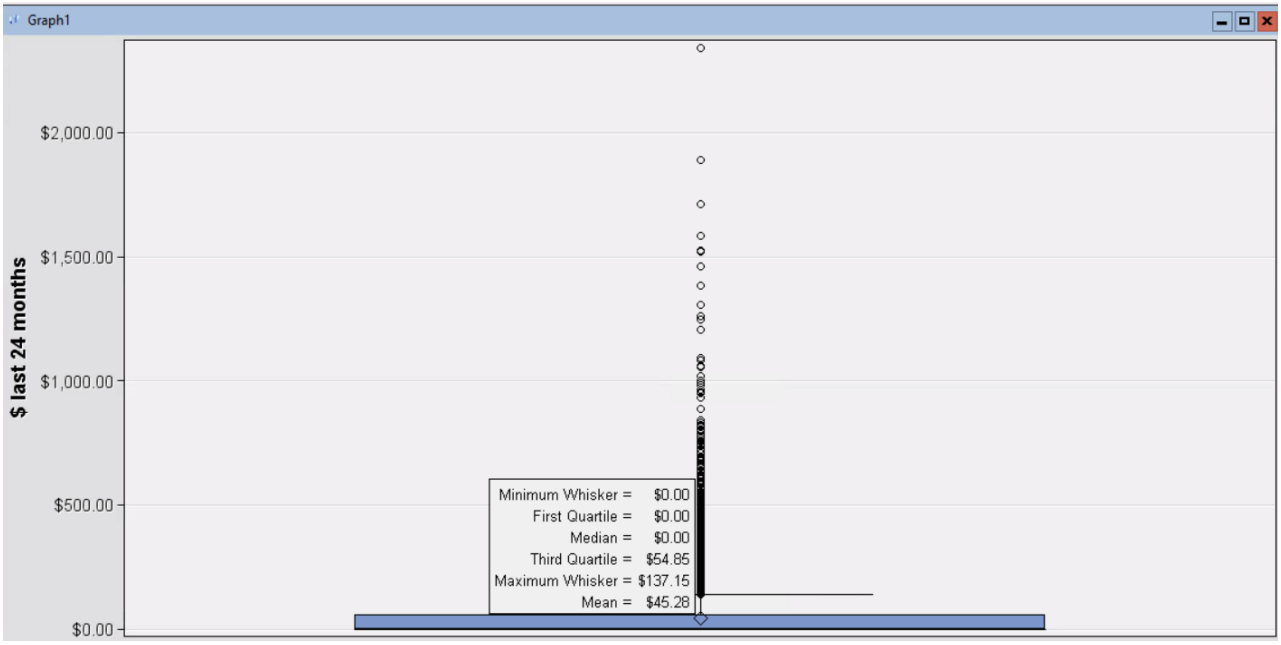
In each case, we can multiple the two to form interaction term to reduce correlation/We can drop either of them

Variables can be dropped using the DROP Node.



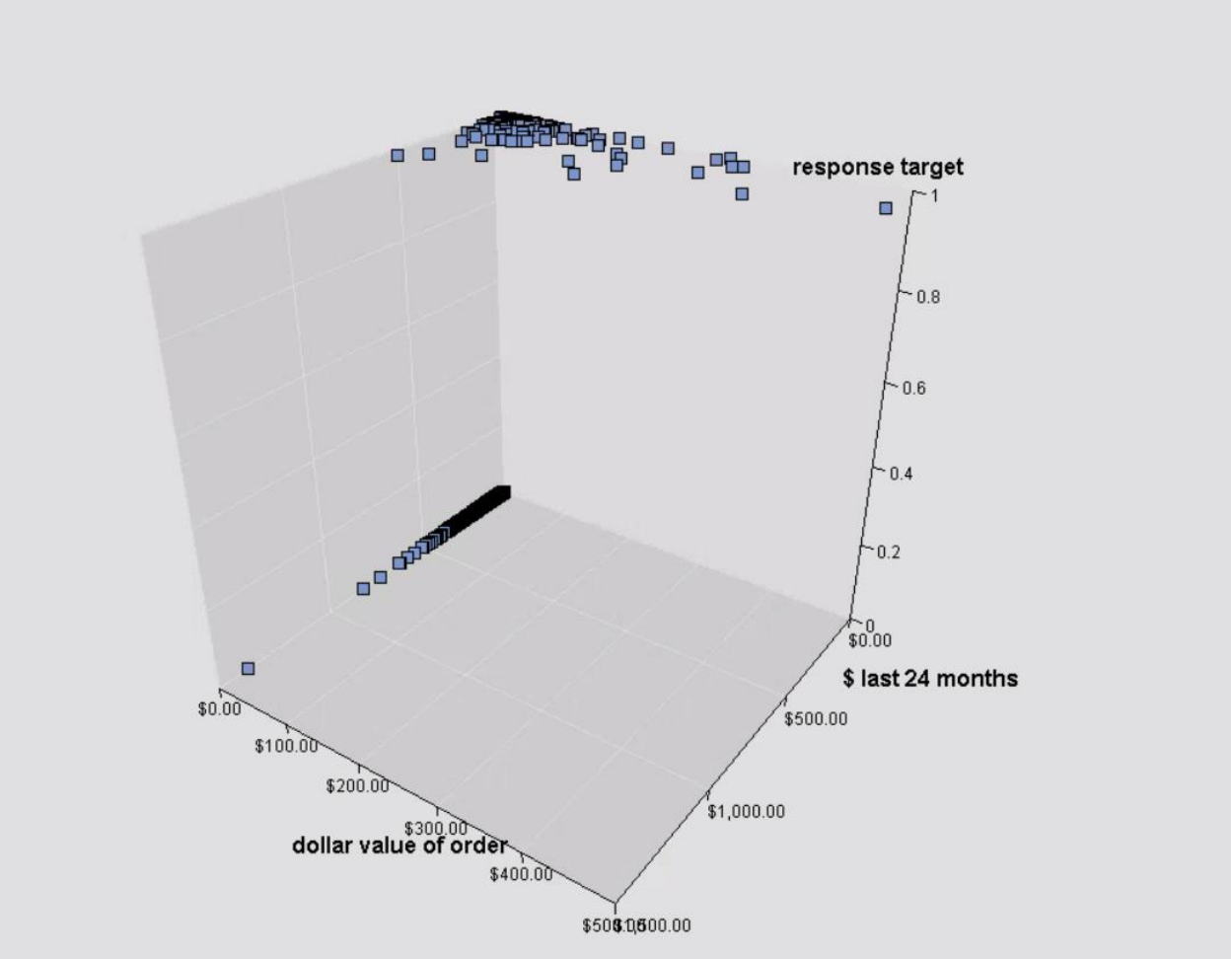
Outliers:

We have used box plot to identify outliers. Below is the analysis on DOLL24 variable. Upper fence is determined as 137.15.



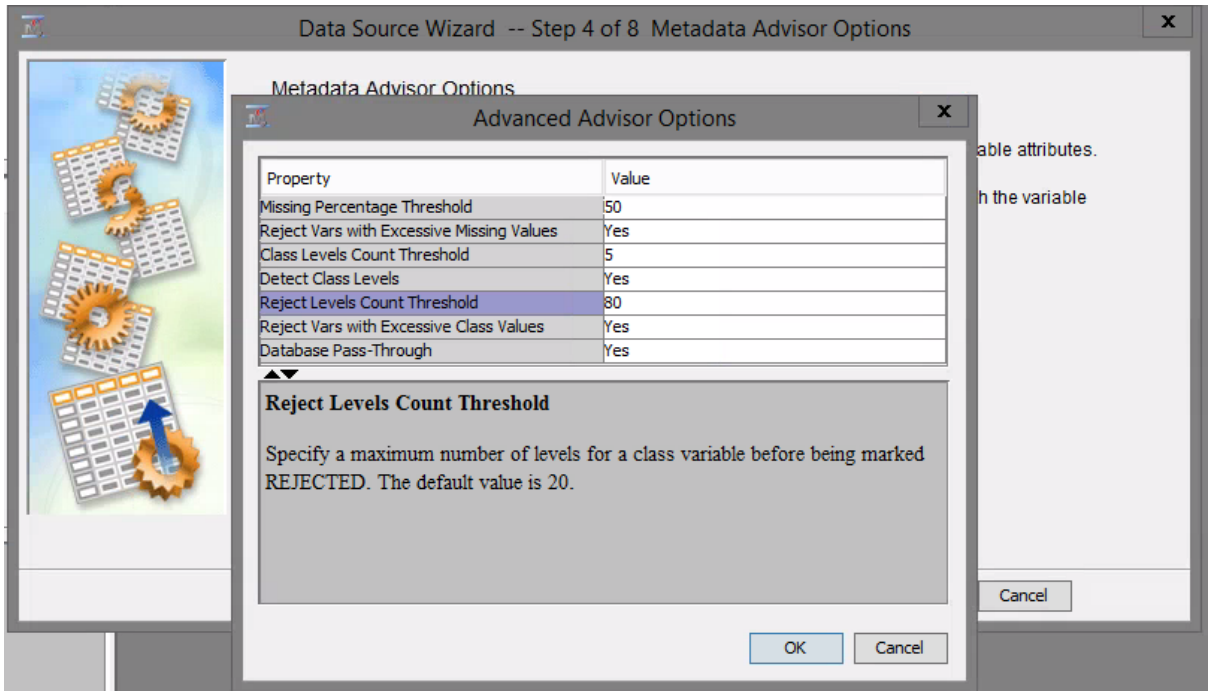
3-D PLOT to sense how independent variables affect dependent variables.

We have two dependent variables – response and Order size. Independent variable chosen in doll24 as it has high worth as seen in statistical plot previously. With high value in $value of last 24 months and dollar value of order, response is 1 as seen below.

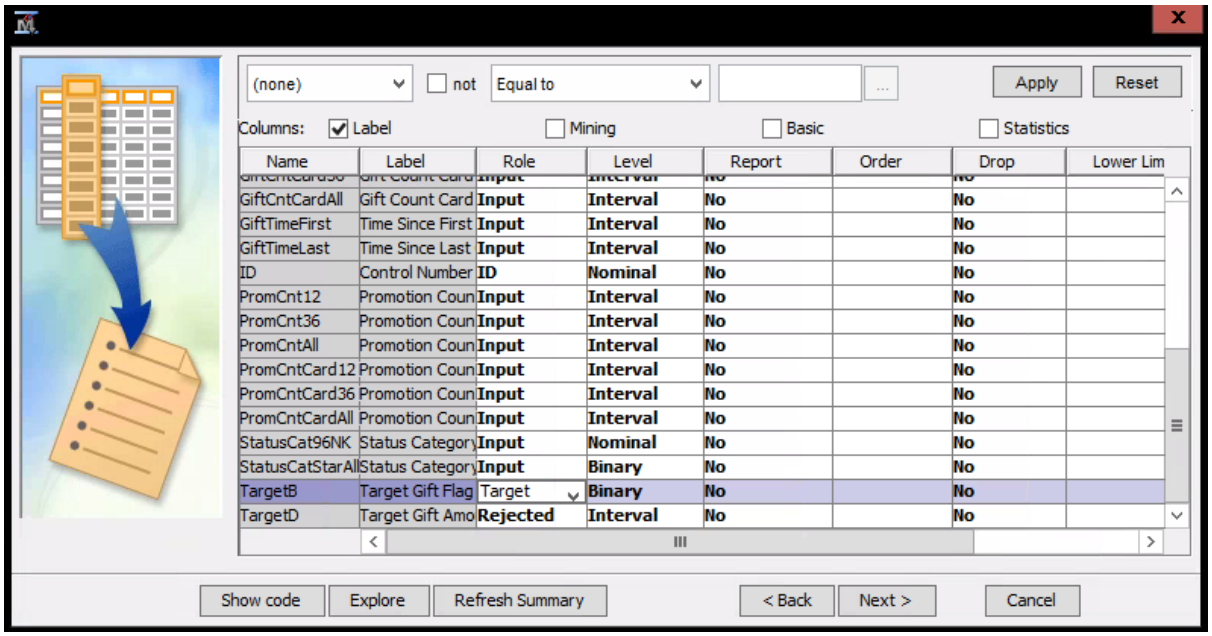


TASK 2

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.

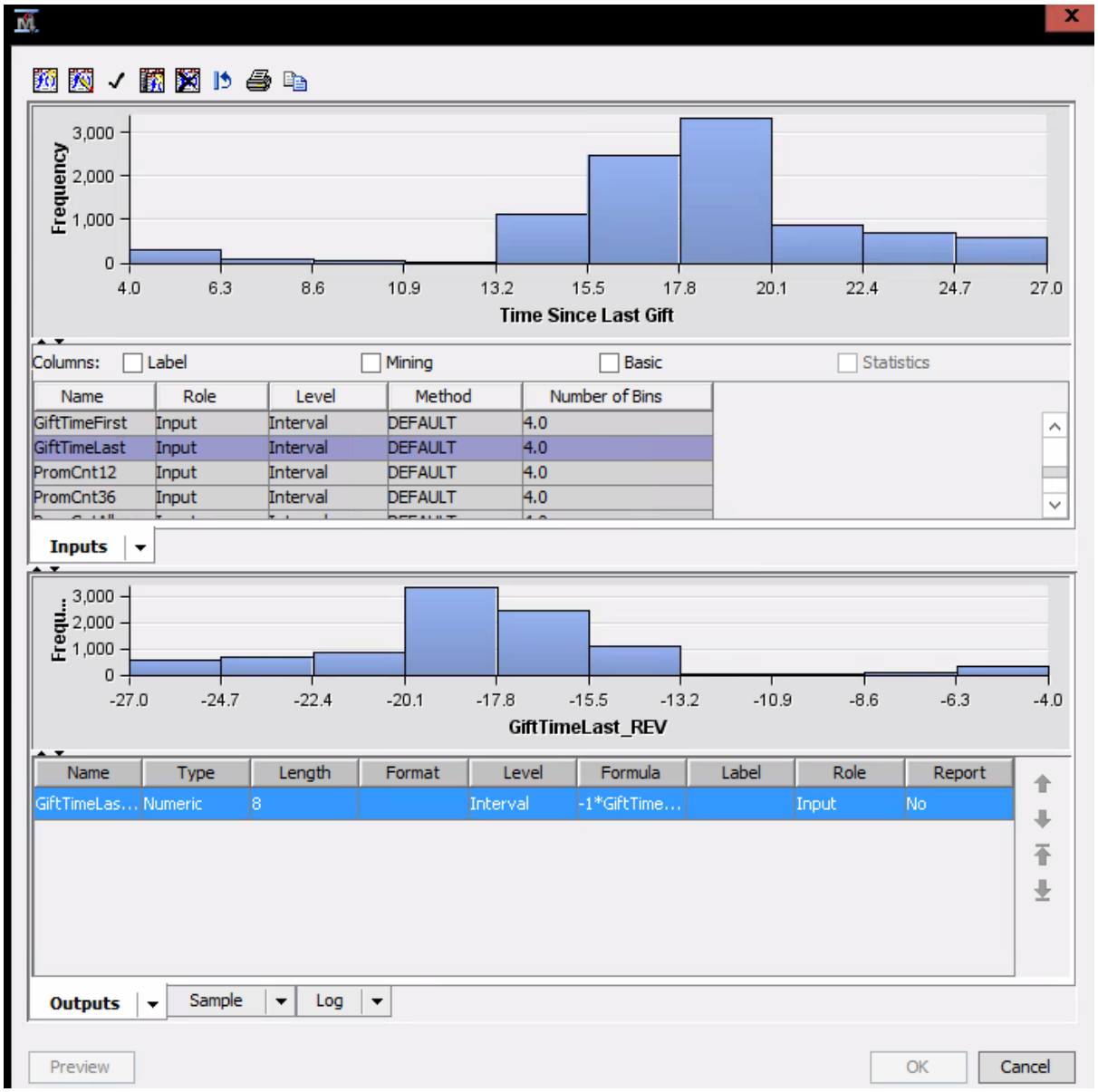


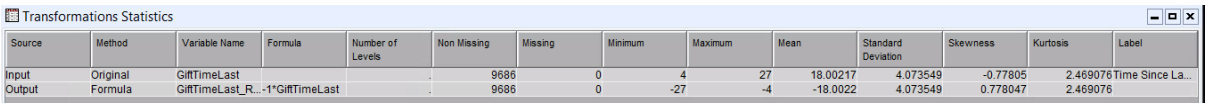
   • Reject the variable **TargetD**.



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.

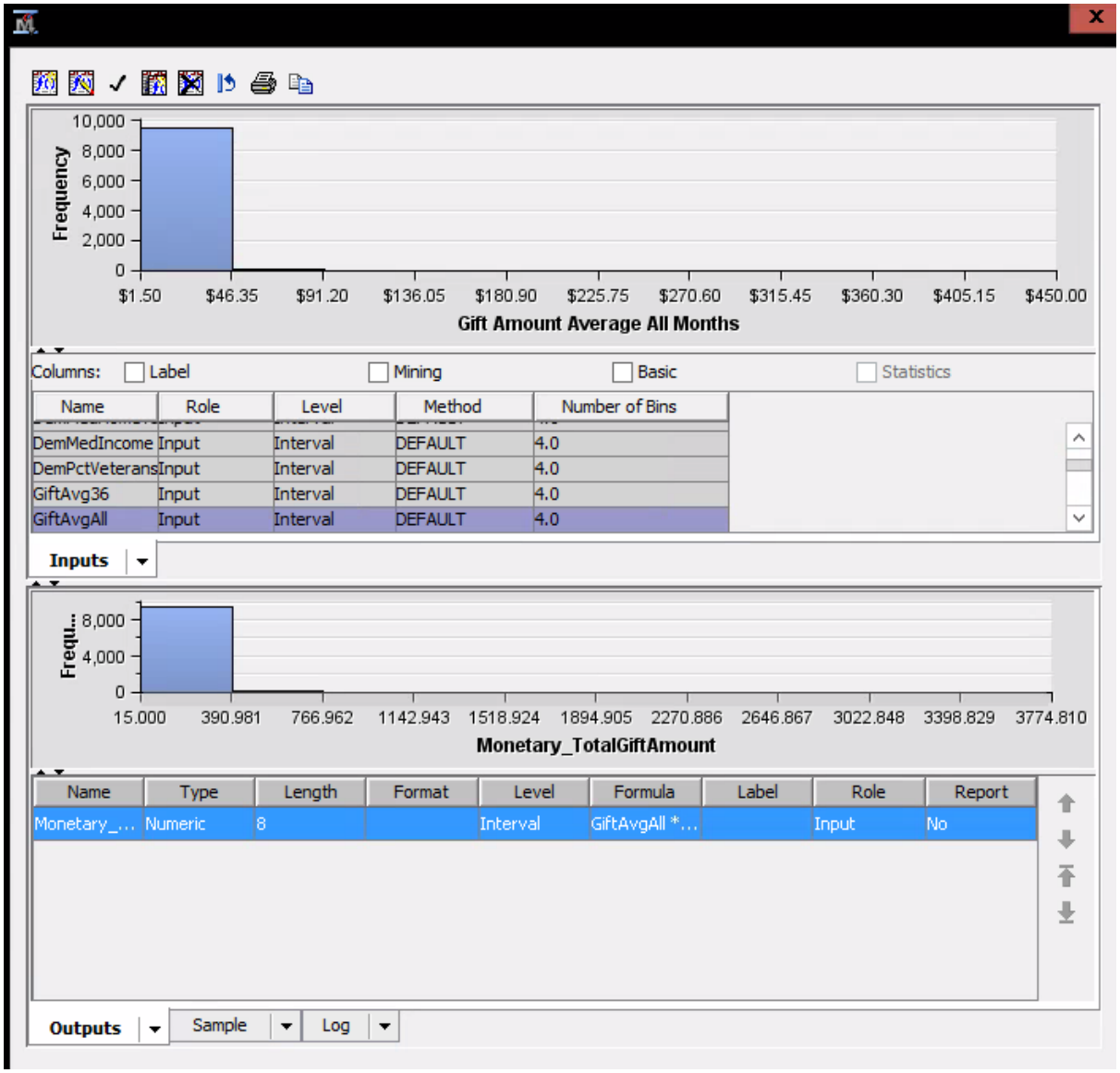
Forming RFM:
R – Recency (-1*GiftTimeLast)
F – GiftCntAll(Gift Count over all months)





Second Transformation is called Monetary_TotalGiftAmount,
Creating formula

Monetary_TotalGiftAmount= GiftAvgAll * GiftCntAll



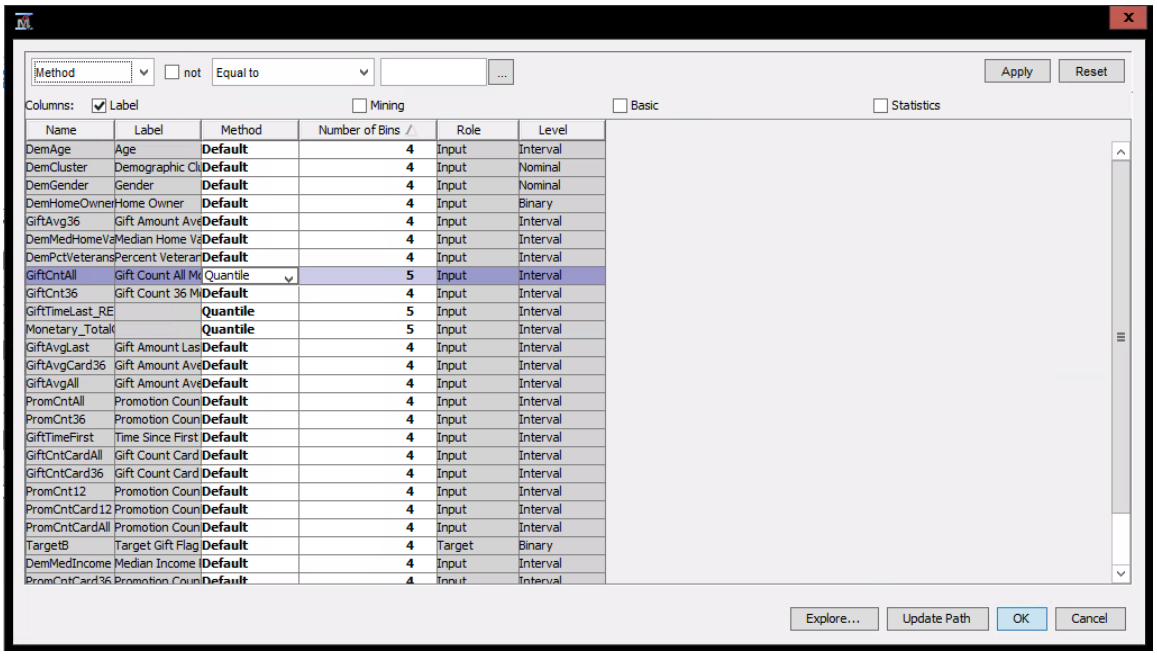
RESULTS:

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	GiftAvgAll		.	9686	0	1.5	450	12.48932	9.209297	14.48649	561.7552	Gift Amount Ave...
Input	Original	GiftCntAll		.	9686	0	1	91	10.50764	8.993401	1.863109	6.047766	Gift Count All M...
Output	Formula	Monetary_Total...	GiftAvgAll * Gift...	.	9686	0	15	3774.81	107.0642	112.0174	7.838657	160.0225	

Third Transformation is called **BinRFM**, we change bin number from 4 to 5 and Method from Default to Quantile.

- Select ellipsis next to the **Variables** in the properties panel.Select the variables **GiftLastTime_Rev, Monetary_TotalGiftAmount** and **GiftCntAll**, change the method from **Default** to **Quantile** and change the number of bins **4** to **5** and click ok, Run the node



Name	Label	Method	Number of Bins	Role	Level
DemAge	Age	Default	4	Input	Interval
DemCluster	Demographic Cl	Default	4	Input	Nominal
DemGender	Gender	Default	4	Input	Nominal
DemHomeOwner	Home Owner	Default	4	Input	Binary
GiftAvg36	Gift Amount Ave	Default	4	Input	Interval
DemMedHomeVa	Median Home Va	Default	4	Input	Interval
DemPctVeterans	Percent Veteran	Default	4	Input	Interval
GiftCntAll	Gift Count All M	Quantile	5	Input	Interval
GiftCnt36	Gift Count 36 M	Default	4	Input	Interval
GiftTimeLast_RE		Quantile	5	Input	Interval
Monetary_Total(		Quantile	5	Input	Interval
GiftAvgLast	Gift Amount Las	Default	4	Input	Interval
GiftAvgCard36	Gift Amount Ave	Default	4	Input	Interval
GiftAvgAll	Gift Amount Ave	Default	4	Input	Interval
PromCntAll	Promotion Coun	Default	4	Input	Interval
PromCnt36	Promotion Coun	Default	4	Input	Interval
GiftTimeFirst	Time Since First	Default	4	Input	Interval
GiftCntCardAll	Gift Count Card	Default	4	Input	Interval
GiftCntCard36	Gift Count Card	Default	4	Input	Interval
PromCnt12	Promotion Coun	Default	4	Input	Interval
PromCntCard12	Promotion Coun	Default	4	Input	Interval
PromCntCardAll	Promotion Coun	Default	4	Input	Interval
TargetB	Target Gift Flag	Default	4	Target	Binary
DemMedIncome	Median Income I	Default	4	Input	Interval

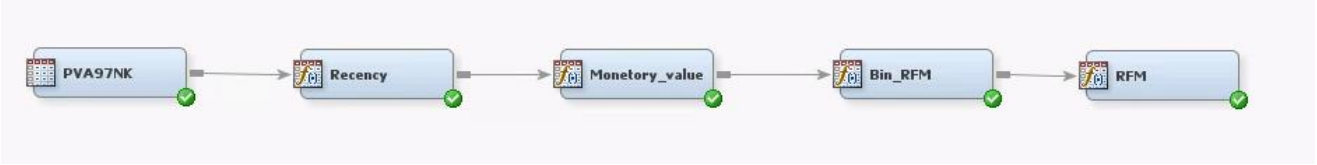
Results:

Transformations Statistics

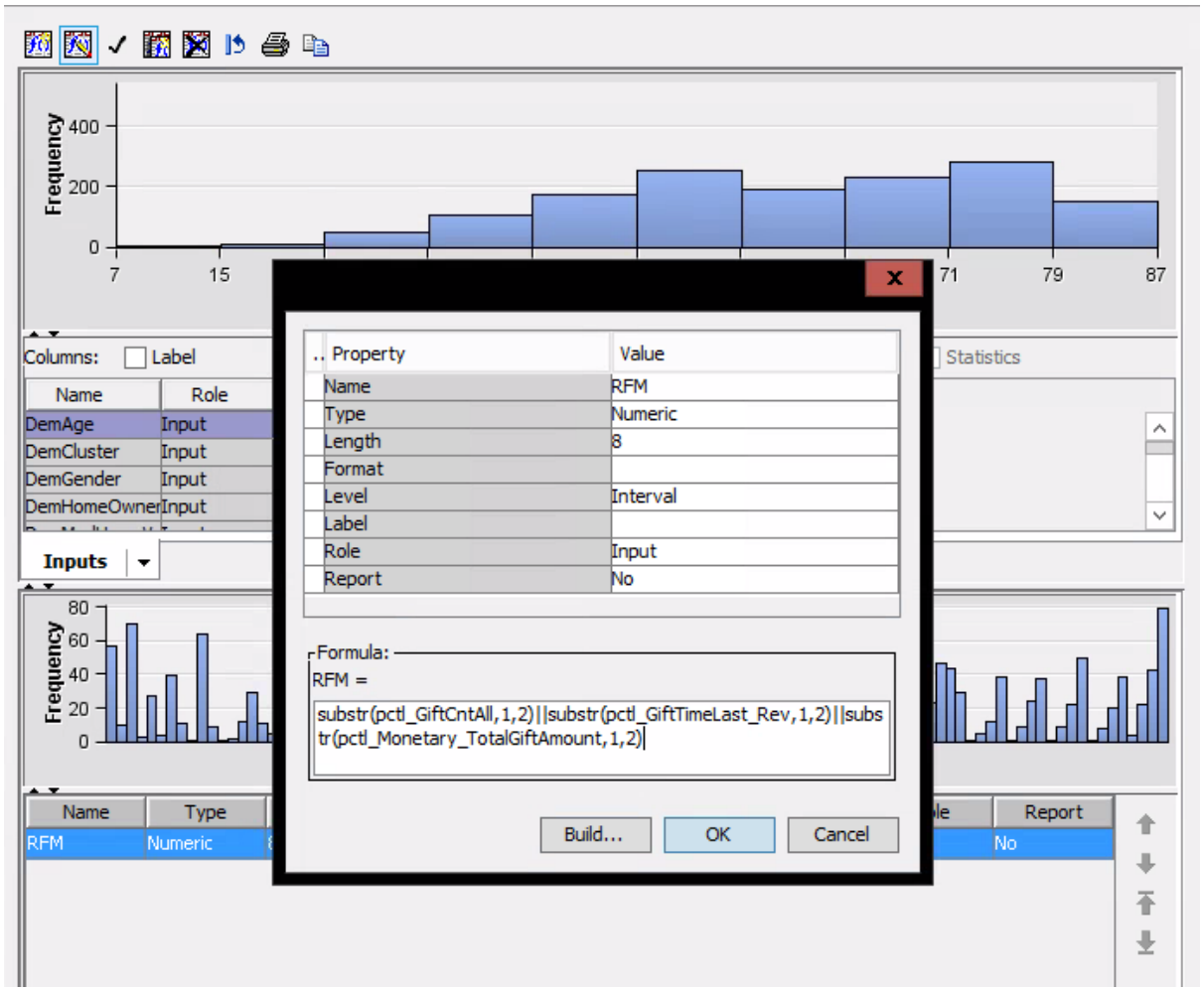
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	GiftCntAll			9686	0	1	91	10.50764	8.993401	1.863109	6.047766	Gift Count All M...
Input	Original	GiftTimeLast_R...			9686	0	-27	-4	-18.0022	4.073549	0.778047	2.469076	
Input	Original	Monetary_Total...			9686	0	15	3774.81	107.0642	112.0174	7.838657	160.0225	
Output	Computed	PCTL_GiftCntAll	Quantile(5)	5		0							Transformed: ...
Output	Computed	PCTL_GiftTime...	Quantile(5)	5		0							Transformed G...
Output	Computed	PCTL_Monetar...	Quantile(5)	5		0							Transformed M...

Fourth Transformation is called RFM defined by creating formula using substring function.

- Add another **Transform Variables** node from Modify tab into the diagram node. Connect **BinRFM node** to the **Transform Variables** node. Rename **Transform Variables** node to **RFM.**



Results:



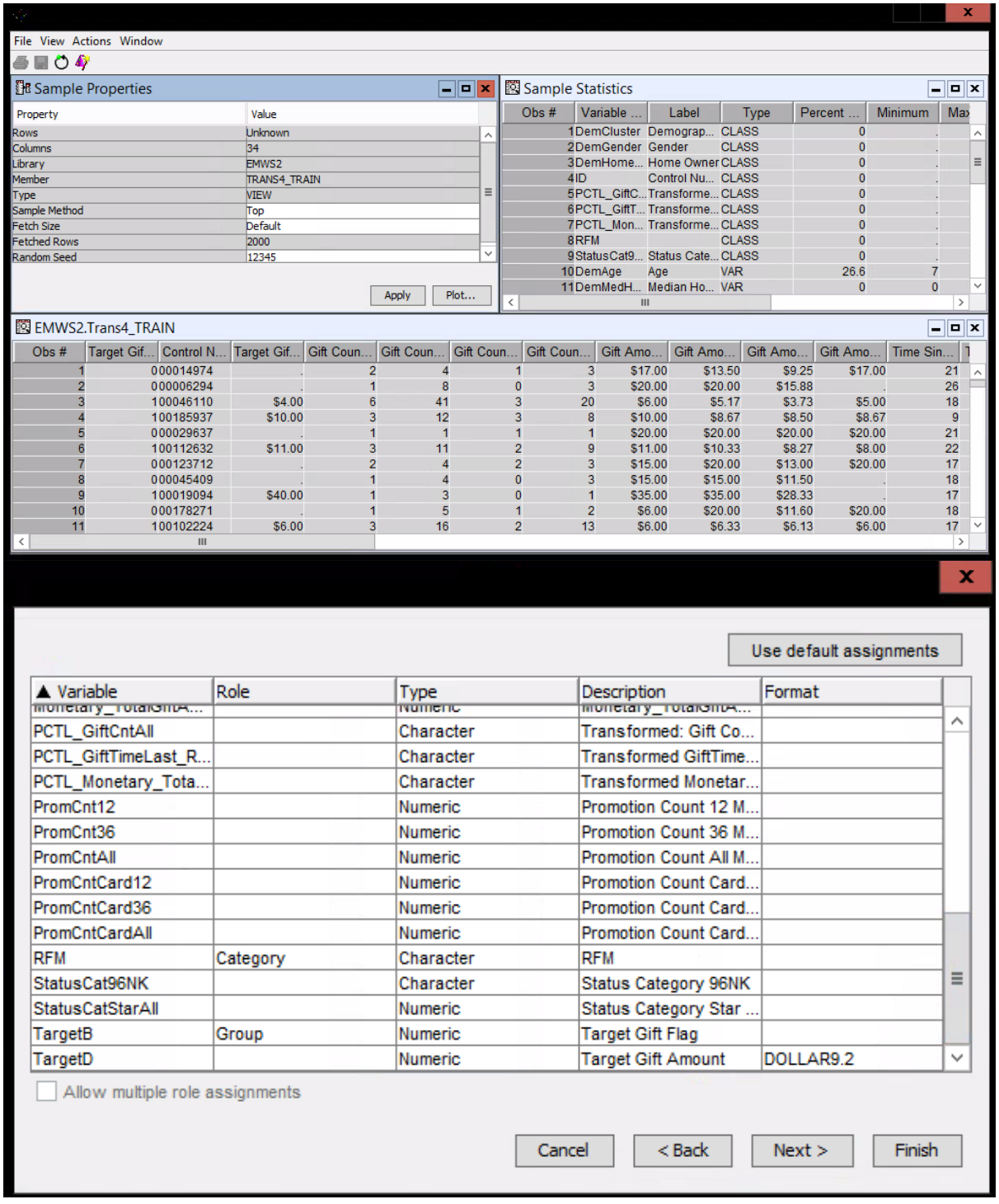
Run the **RFM** node and view the results.

- Total number of observations = 116

Transformations Statistics

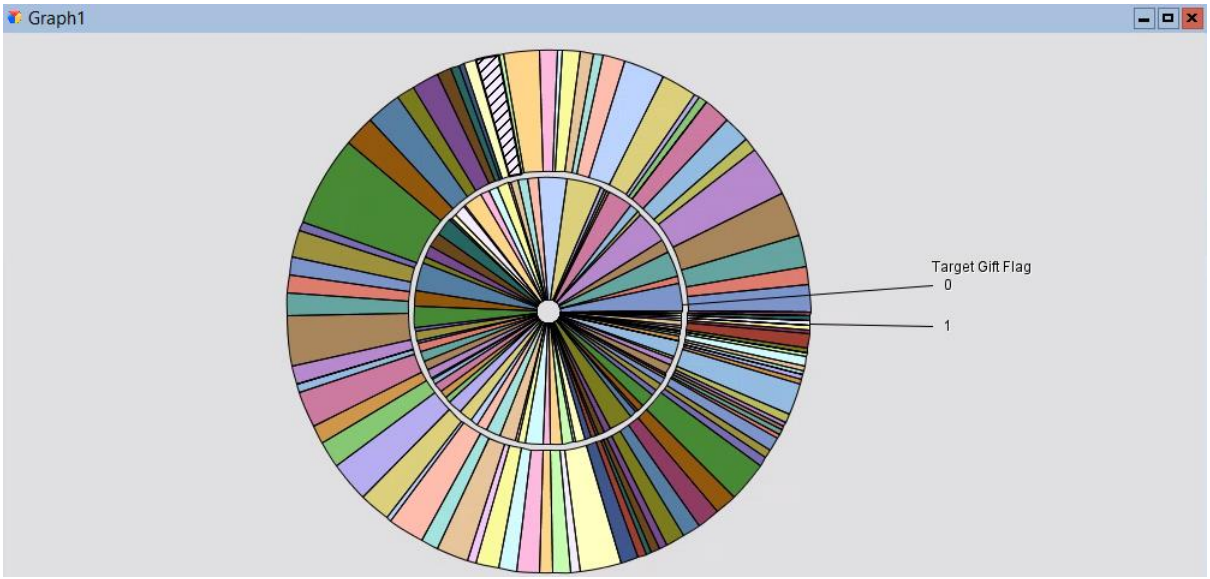
Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	PCTL_GiftCntAll		5	.	0	.	.	.	.	.	.	Transformed: ...
Input	Original	PCTL_GiftTime...		5	.	0	.	.	.	.	.	.	Transformed G...
Input	Original	PCTL_Monetar...		5	.	0	.	.	.	.	.	.	Transformed M...
Output	Formula	RFM	substr(pctl_Gift...	116	.	0	.	.	.	.	.	.	.

Transformation Results of substring function:



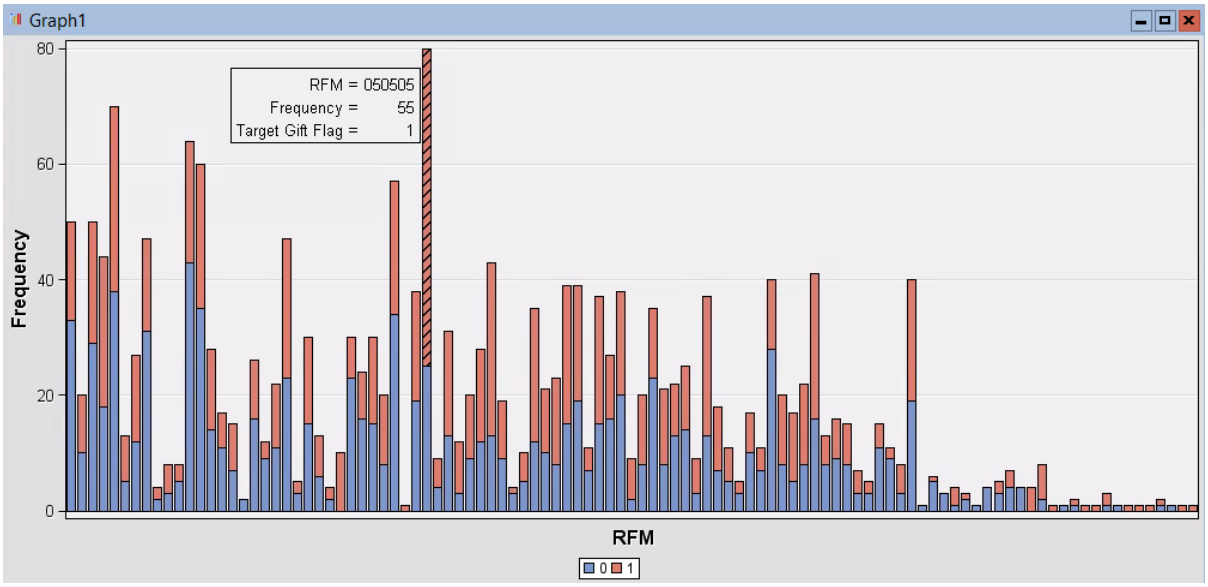
C) Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar chart.

Grouped pie chart – Since we are limited by the maximum number of slices. Pie chart groups certain values of RFM and displays it as others.



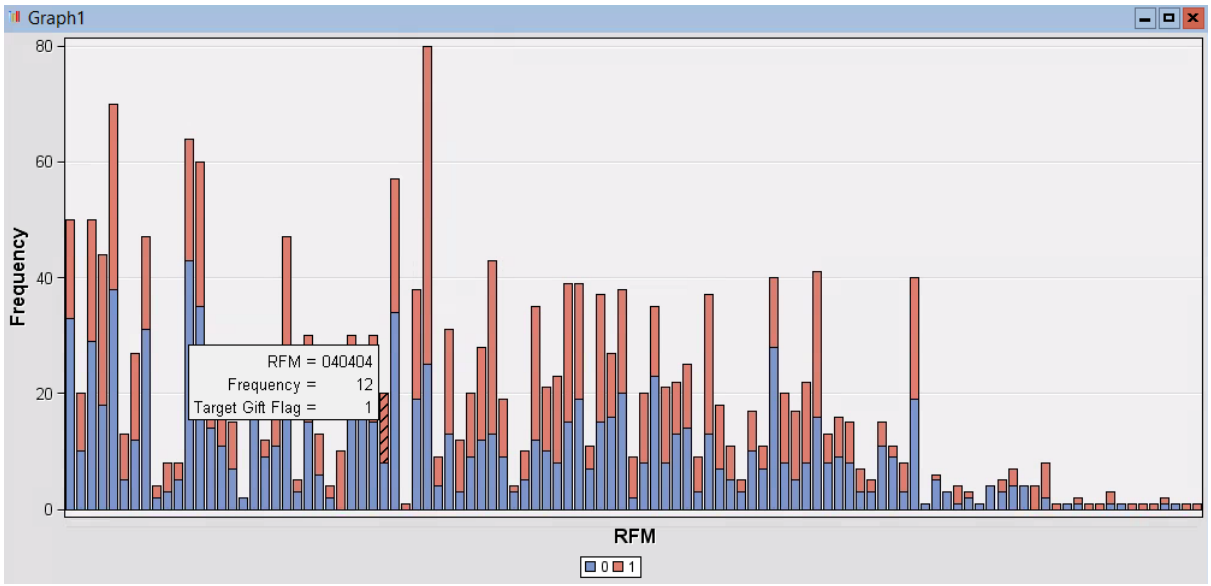
We are not able to interpret any conclusions, so we go for Bar charts.
We can see that people falling in bin 050101 have donated the most. There is less number of people not donating

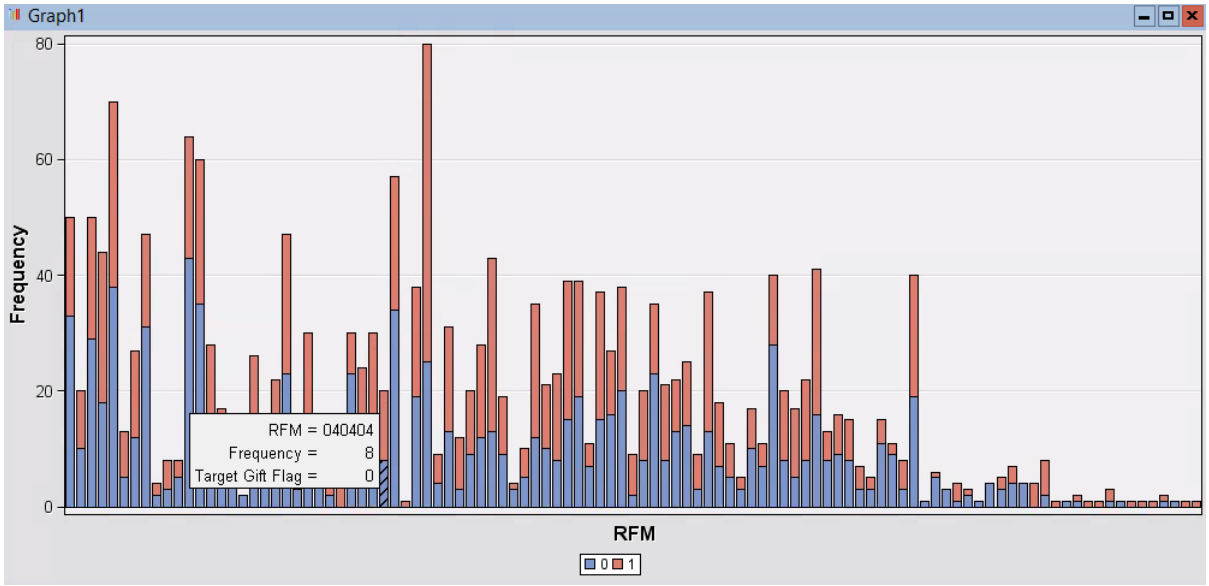
(ii) Bar Chart:



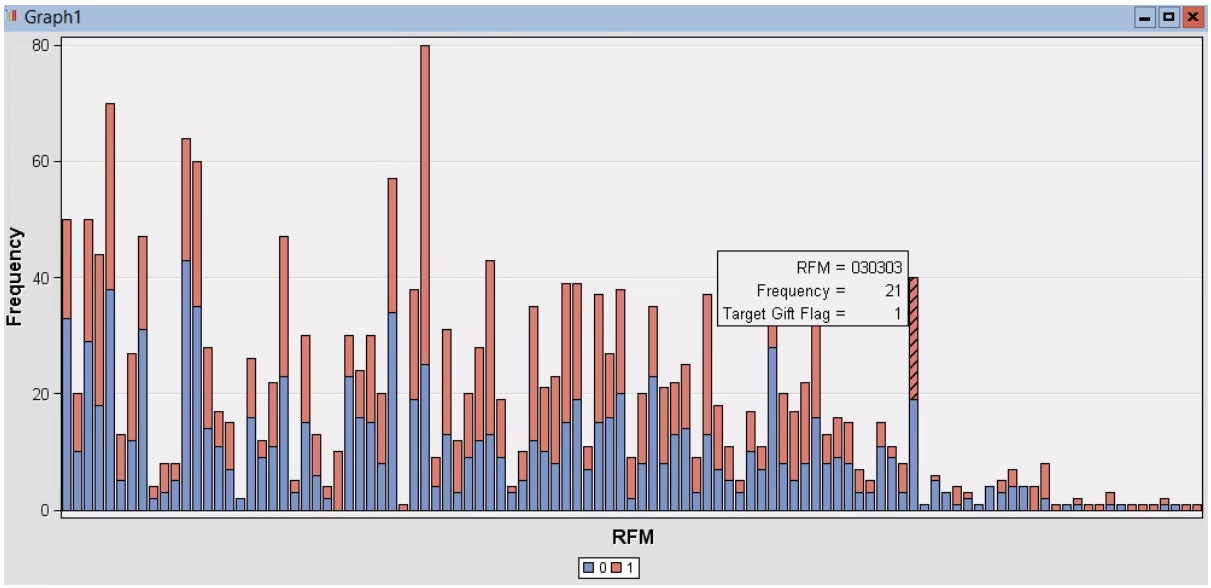
- The red color in bar chart indicates the people who had received gifts and the blue color indicates the people who have not received gifts.The total bar chart indicates the total donations.
- The optimal solution is that red color part of chart is less than the blue color part.

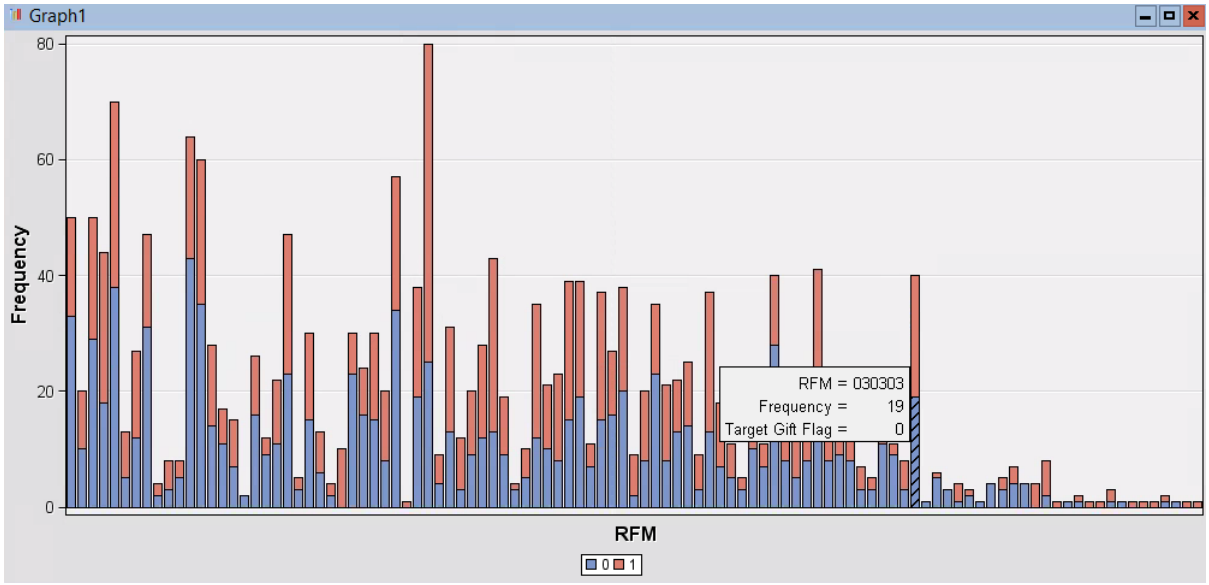
c) Calculate response rate for 040404 and 030303 group?





040404 – Response rate – 8/12 = 66.66%
030303 – Response rate – 19/21 = 90.4%





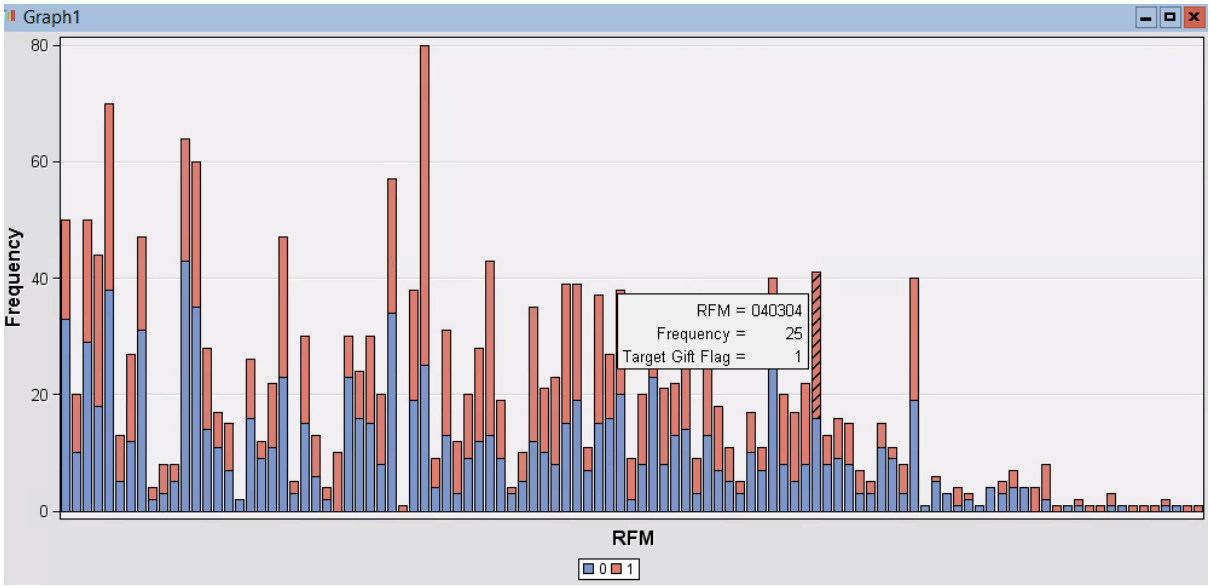
030303 has higher response rate.

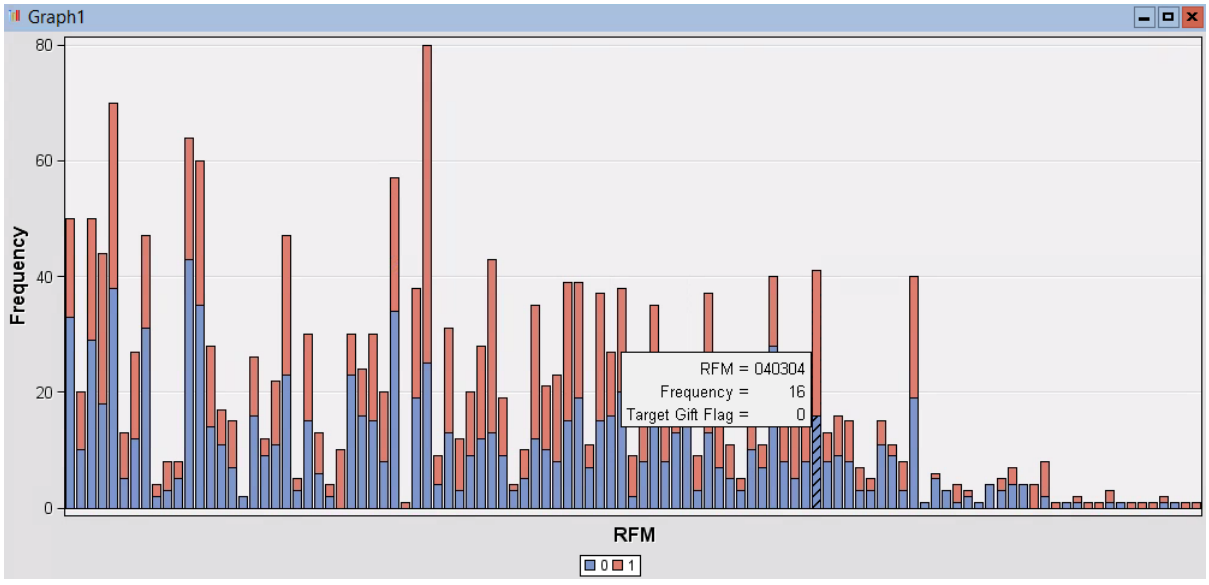
d) Each promotional mailing (request for a gift) costs $2.3, and the average donation is about $21 What is the break-even response rate for this promotion? Do any RFM cells exceed this response rate? Remember to account for the fact that in the population, 95% of mailings are not responded to, while this sample is oversampled to 50% responders and 50% non-responders.

Break even response rate = Each promotional mailing costs / average donation
= (2.30)/21
= 10%
Therefore, any RFM cells exceeding this response rate of 10% would be profitable to mail.
Given, that fact in the population, 95% of mails are not responded.

Therefore, responded mails are 5%.
Ratio = 5/95 = 0.0526

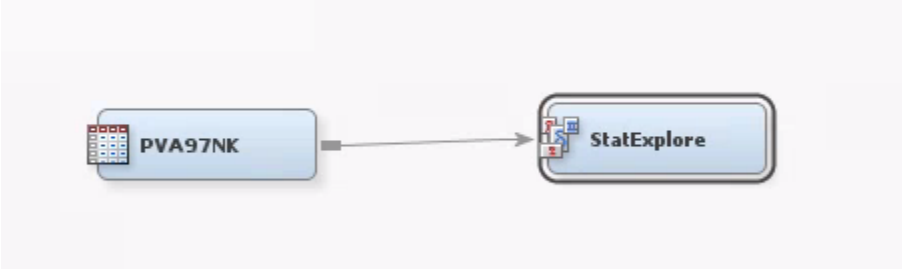
Consider Bin 040304, The Response value is 25 and no response is 16 so
[25*0.0526]/[(25*0.0526)/16] = 16.4%
Thus, the Bin 020503 can be considered as the response rate is more than 11%.



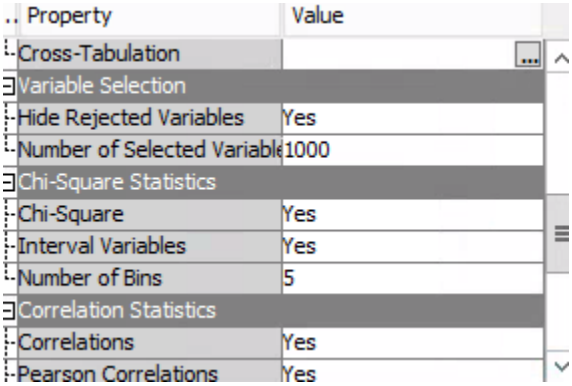


Additional Analysis on PVA97NK on Missing Values:

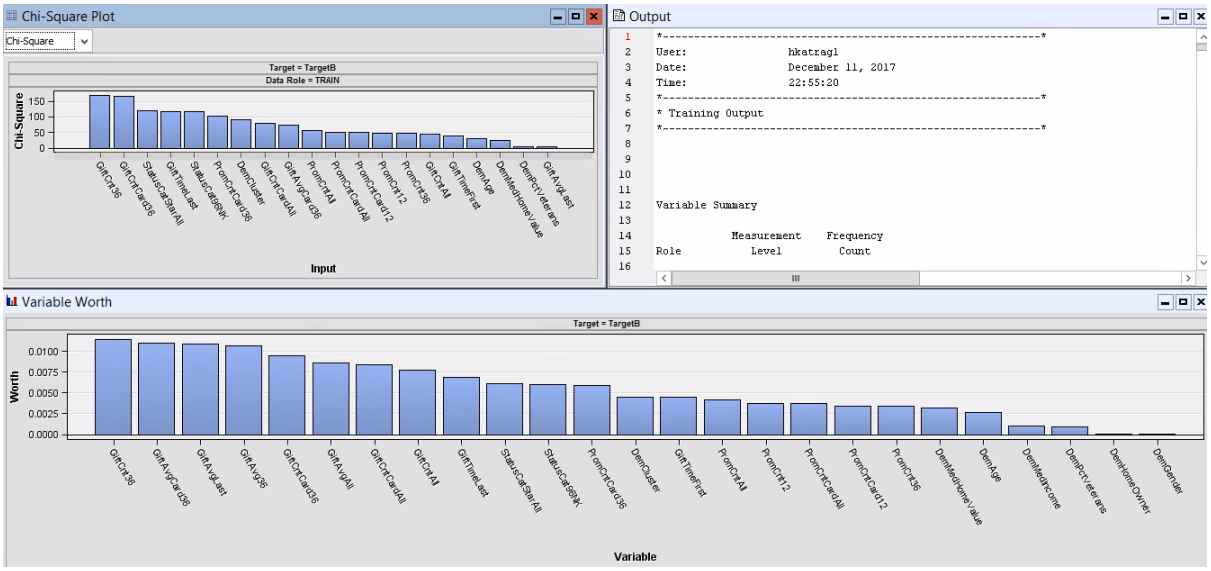
- Before building any model, it is important to check for any missing values in the input data. We know that regression ignores the missing values, this would decrease the amount of data that we use to build the model and lower the predictive power. Therefore, checking missing values and replace these missing values with some replacements are important before building model.



In the properties panel of StatExplore node, select the value of Interval Variables to Yes. We have selected Yes to distribute interval variables into 5 bins.



- Run the StatExplore node and observe results.



Open the Class Variable Summary Statistics and the interval variable summary statistics sections.

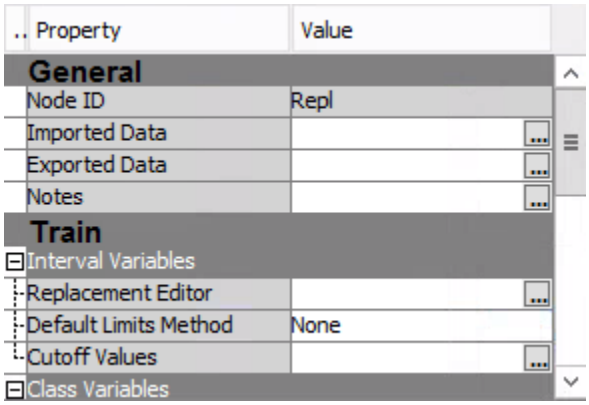
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	DemCluster	INPUT	54	0	40	4.46	24	4.14
TRAIN	DemGender	INPUT	3	0	F	53.92	M	40.52
TRAIN	DemHomeOwner	INPUT	2	0	H	55.51	U	44.49
TRAIN	StatusCat96NK	INPUT	6	0	A	60.15	S	24.42
TRAIN	StatusCatStarAll	INPUT	2	0	1	54.06	0	45.94
TRAIN	TargetB	TARGET	2	0	0	50.00	1	50.00

From the below results, we can observe that the two variables **DemAge** and **GiftAgeCard36** have missing values.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
DemAge	INPUT	59.15084	16.5164	7279	2407	0	60	87	-0.38791	-0.47761
DemMedHomeValue	INPUT	110986.3	98670.86	9686	0	0	76900	600000	2.378211	6.451365
DemMedIncome	INPUT	40491.44	28707.49	9686	0	0	43100	200001	0.310025	0.636848
DemPctVeterans	INPUT	30.60427	11.39499	9686	0	0	31	85	-0.20706	1.27441
GiftAvg36	INPUT	14.8762	10.05701	9686	0	0	13.5	260	5.627792	77.09997
GiftAvgAll	INPUT	12.48932	9.209297	9686	0	1.5	10.71	450	14.48649	561.7552
GiftAvgCard36	INPUT	14.22443	10.02271	7906	1780	1.33	12.5	260	6.051455	87.12627
GiftAvgLast	INPUT	16.01774	12.0418	9686	0	0	15	450	9.918893	246.0504
GiftCnt36	INPUT	3.205451	2.133421	9686	0	0	3	16	1.288353	2.047415
GiftCntAll	INPUT	10.50764	8.993401	9686	0	1	8	91	1.863109	6.047766
GiftCntCard36	INPUT	1.856597	1.595419	9686	0	0	1	9	1.172452	1.494867
GiftCntCardAll	INPUT	5.58249	4.736894	9686	0	0	4	41	1.331353	2.024864
GiftTimeFirst	INPUT	71.10035	37.69198	9686	0	15	68	260	0.195399	-1.24787
GiftTimeLast	INPUT	18.00217	4.073549	9686	0	4	18	27	-0.77805	2.469076
PromCnt12	INPUT	12.98885	4.823458	9686	0	2	12	59	2.873723	11.99538
PromCnt36	INPUT	29.34823	7.809743	9686	0	4	31	78	0.261958	2.174341
PromCntAll	INPUT	48.48348	23.06148	9686	0	5	48	174	0.460765	0.216596
PromCntCard12	INPUT	5.392009	1.323648	9686	0	0	6	17	0.684994	5.798685
PromCntCard36	INPUT	11.95468	4.571568	9686	0	2	13	28	-0.4266	-0.98685
PromCntCardAll	INPUT	19.00712	8.562193	9686	0	2	19	56	0.142856	-0.78032

Add **Replacement** node:

- In the properties panel, change the value of Default Limits Method to **None** from drop down. Because, to avoid the replacements of interval variables.
- Click on the ellipsis next to Replacement Editor



- We observe that the value for **DemGender** is U other than F or M. Change the replacement value as _UNKNOWN_.

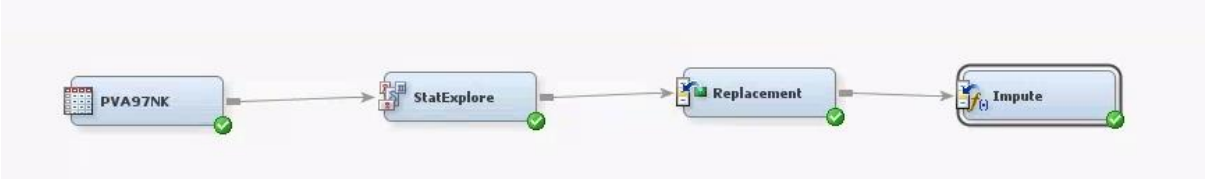
DemCluster	_UNKNOWN_	_DEFAULT_	.
DemGender	F		
DemGender	M		
DemGender	U	_UNKNOWN_	
DemGender	_UNKNOWN_	_DEFAULT_	.

- Run the Replacement node.

Total Replacement Counts

Variable	Role	Label	Train
DemGender	INPUT	Gender	538

Add **Impute** node.



- In properties panel, for **class variables** select **Tree surrogate** in **Default Input Method** from drop down.
- For **Interval Variables**, select **Median** in **Default Input Method** from drop down.
- Default Input Method specifies about which default statistic to be used to impute missing values. The missing values for interval variables are replaced by the median of non-missing values. The missing values for class variables are imputed using the predicted values from a decision tree.

Property	Value
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Median
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.

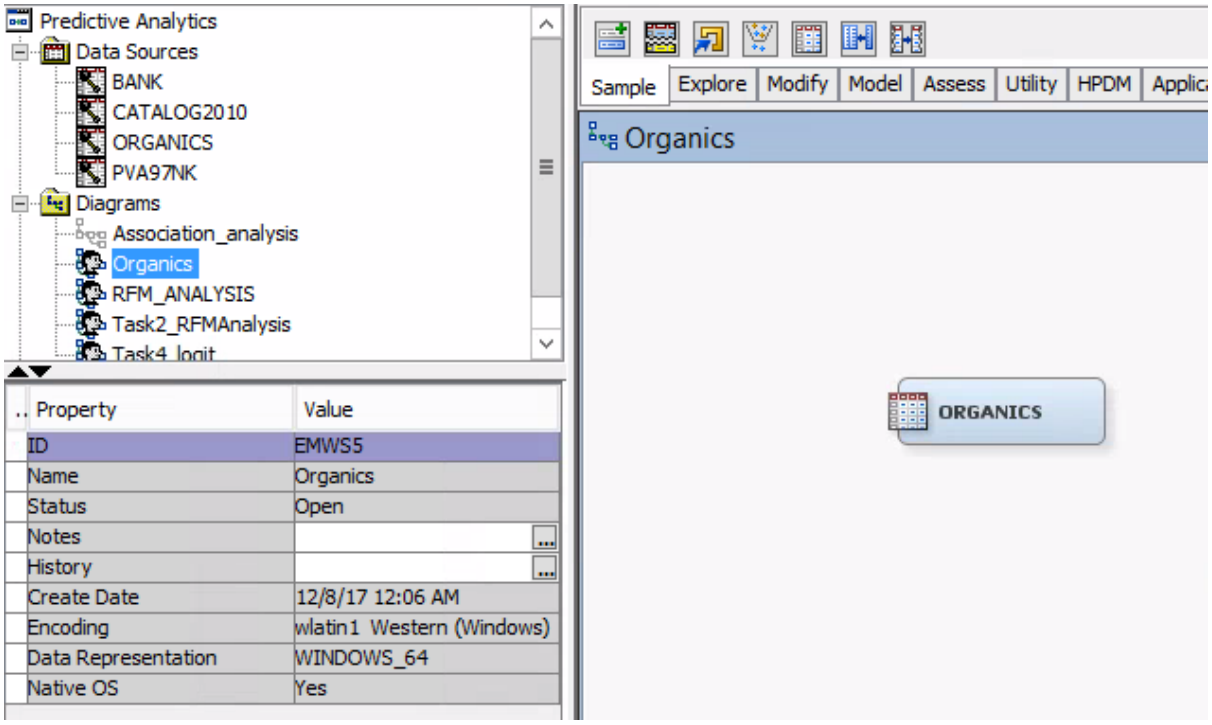
Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
DemAge	MEDIAN	IMP_DemAge	60	INPUT	INTERVAL	Age	2407
GiftAvgCard36	MEDIAN	IMP_GiftAvgCard36	12.5	INPUT	INTERVAL	Gift Amount Average Card 36...	1780

- From the above result, we can observe that, a new variable is created for each variable of missing values are imputed. The original variable is not overwritten and the new variable has the same name as original variable but prefaced with IMP_. In this way we can handle missing values in regression that ignores missing records before building the model.

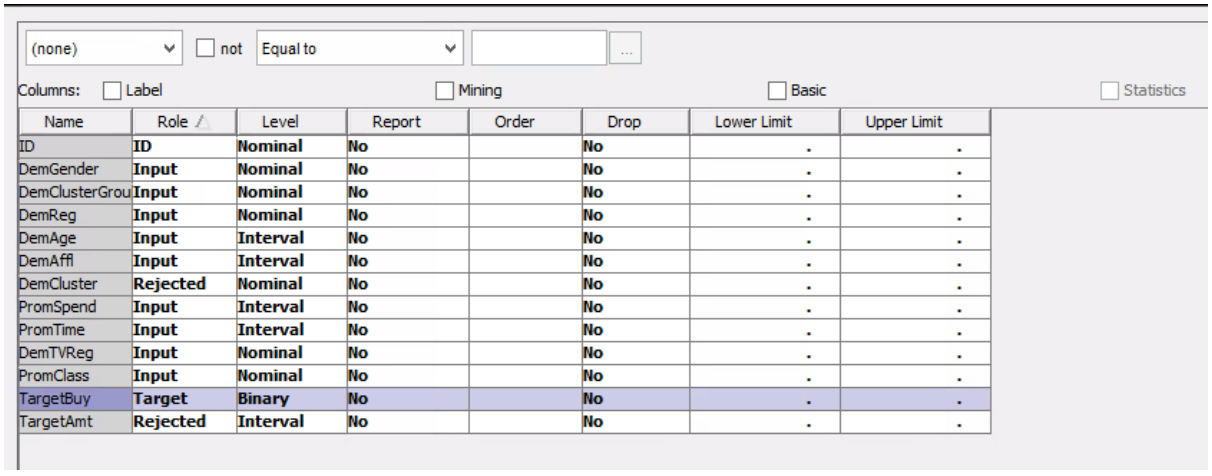
TASK 3

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.

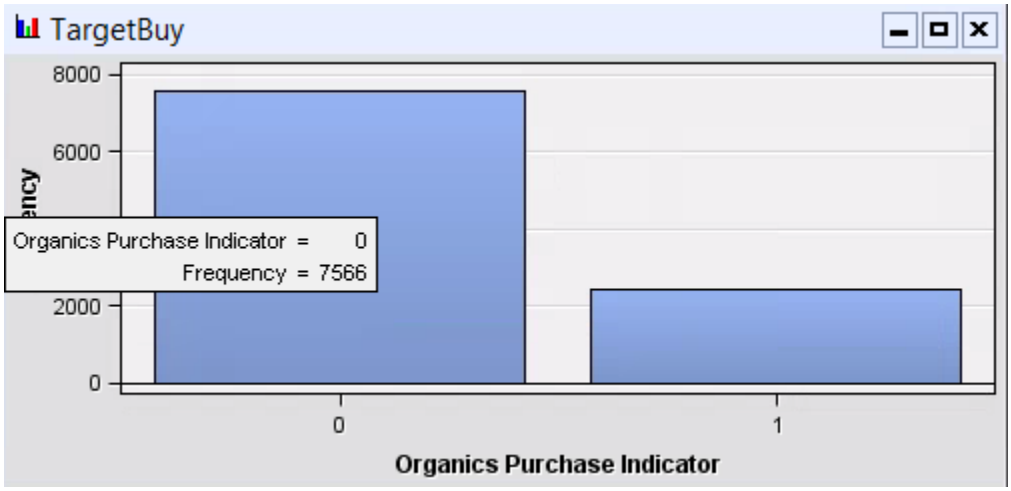


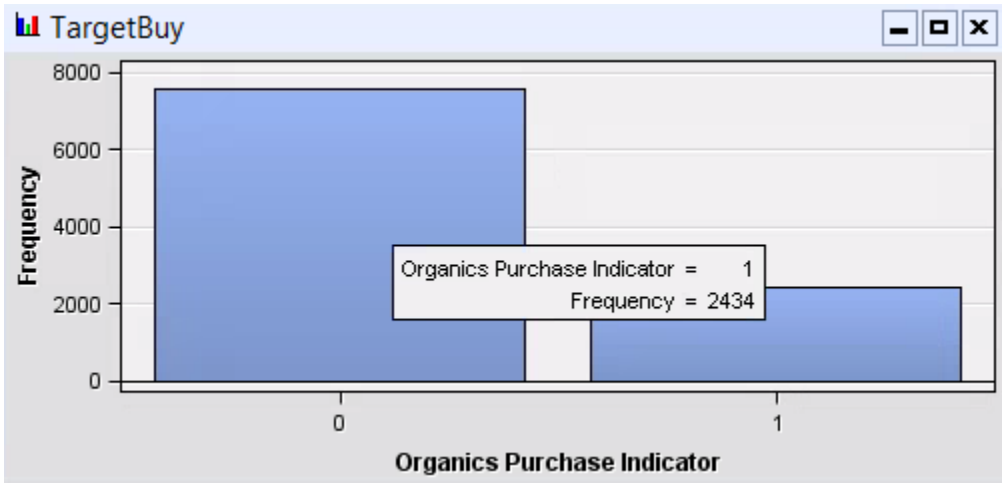
(none) not Equal to

Columns: Label Mining Basic Statistics

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

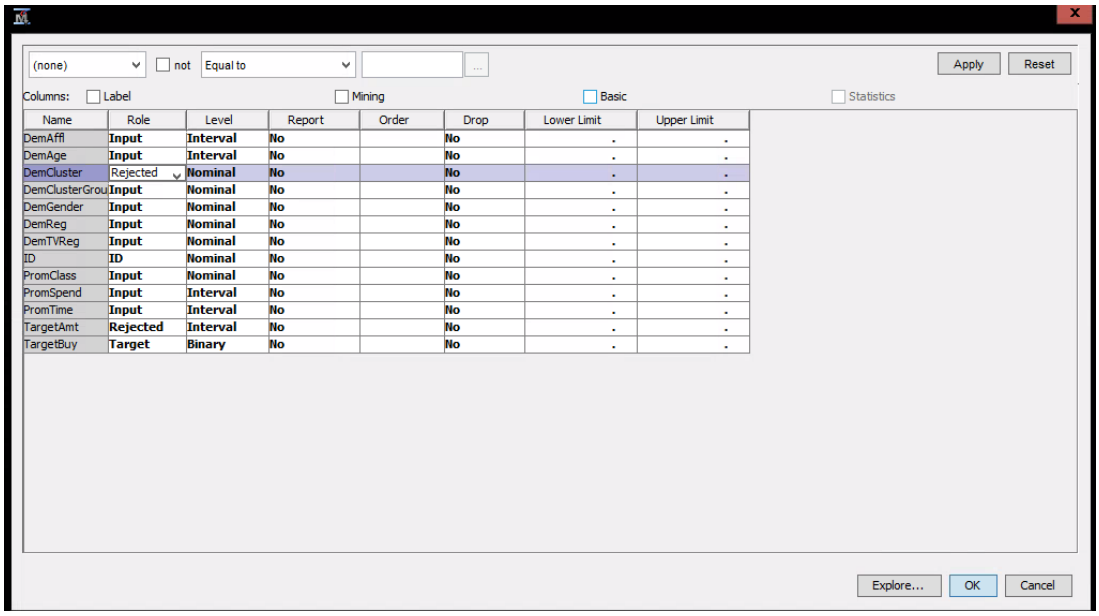
2) Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?





Number of individuals who did not purchased organic products= 7566
Proportion of individuals who purchased organic products= (2434) / (2434+7566) = 24.34 %

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.

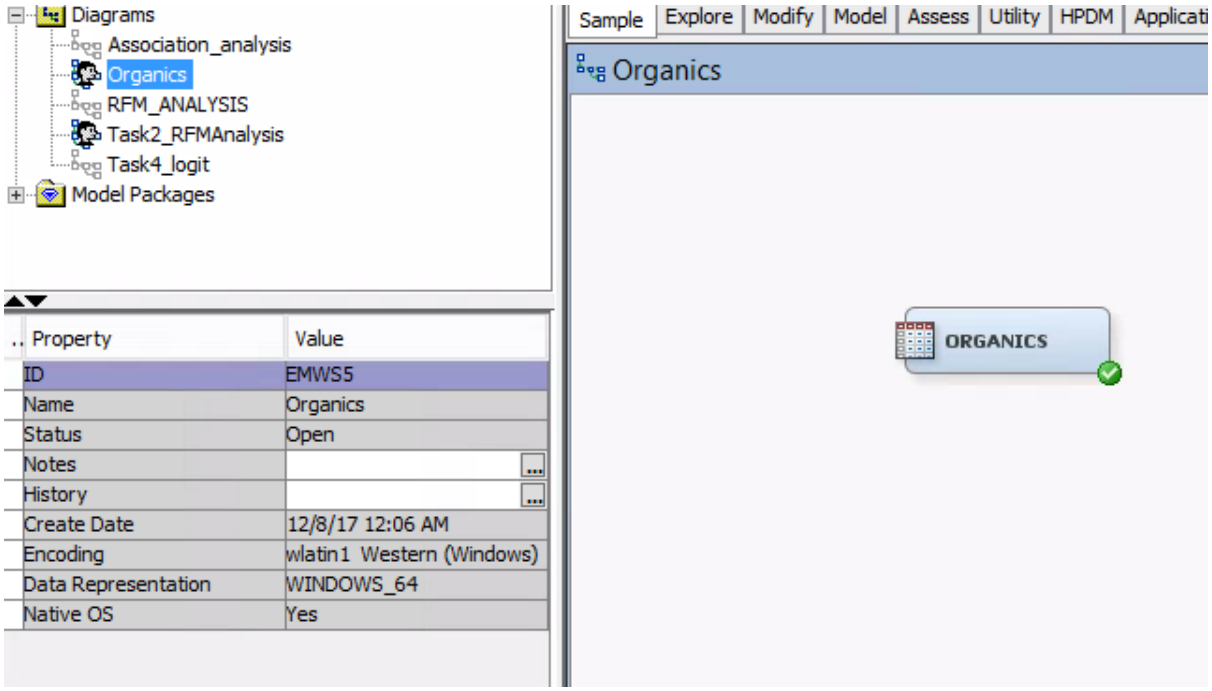


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DemAffl	Input	Interval	No		No	.	.
DemAge	Input	Interval	No		No	.	.
DemCluster	Rejected	Nominal	No		No	.	.
DemClusterGrou	Input	Nominal	No		No	.	.
DemGender	Input	Nominal	No		No	.	.
DemReg	Input	Nominal	No		No	.	.
DemTVReg	Input	Nominal	No		No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

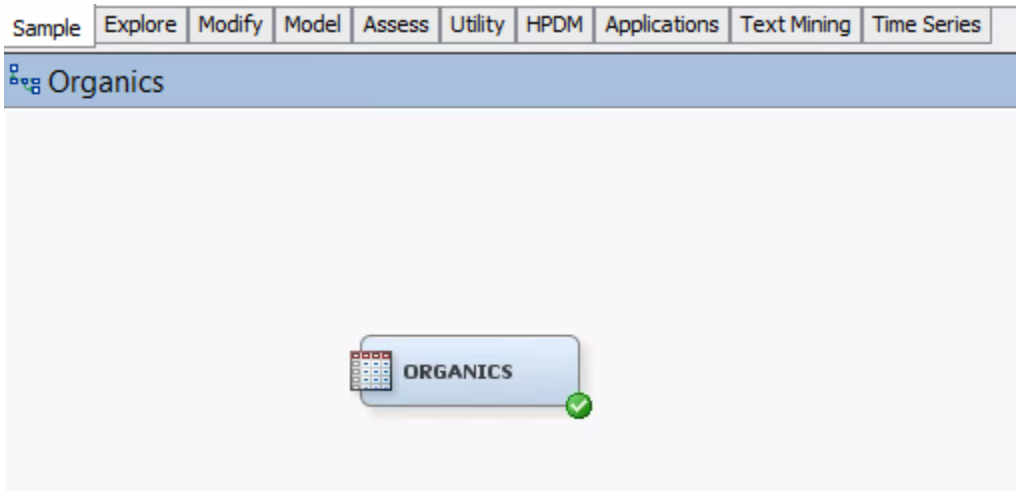
4) As noted above, only **TargetBuy** is used for this analysis, and it should have a role of **Target**. Can **TargetAmt** be used as an input for a model that is used to predict **TargetBuy**? Why or why not?

TargetAmt is the number of organic products purchased which depends on the **TargetBuy**. **TargetBuy** is response variable which indicates whether an individual has purchased an organic product or not. **TargetAmt** is dependent on the **TargetBuy**, only when a **TargetBuy** =1 that is when customer response is 1 he/she will have an amount. Since, input variable should be independent we cannot use **TargetAmt** as an input for a model that predict **TargetBuy** as, it is a dependent variable.

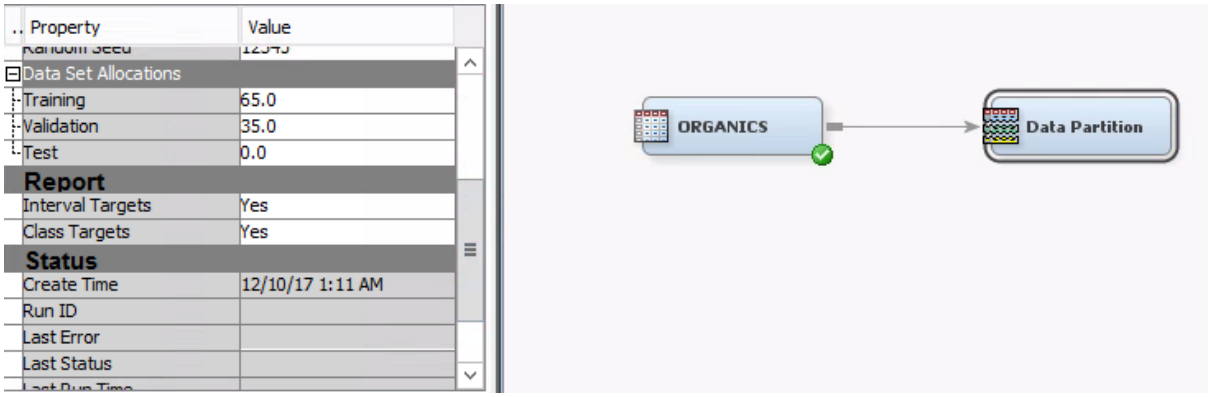
5) Finish the **ORGANICS** data source definition.



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



c. 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.



Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
TargetBuy	0	0	16718	75.2284	Organics Purchase Indicator
TargetBuy	1	1	5505	24.7716	Organics Purchase Indicator

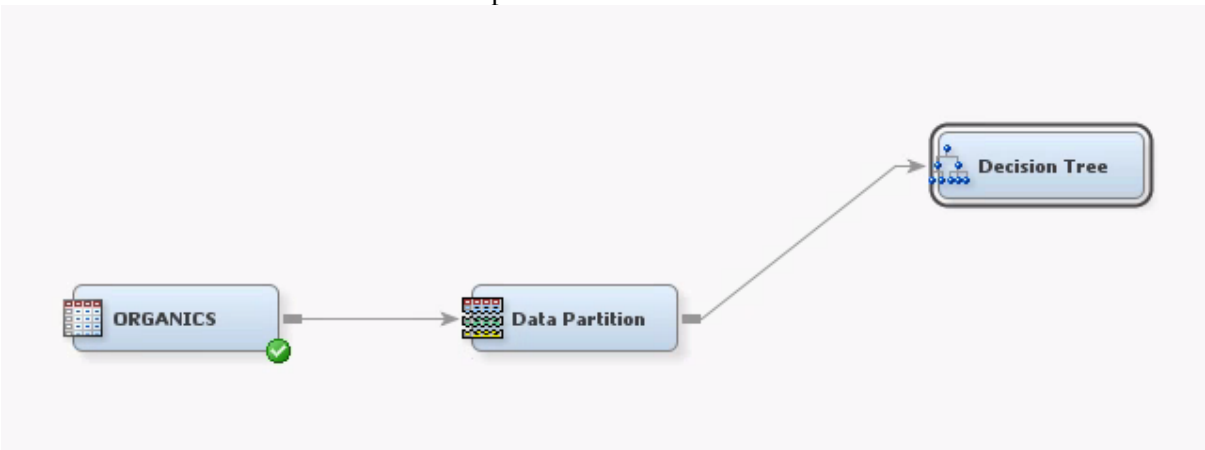
Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
TargetBuy	0	0	8359	75.2250	Organics Purchase Indicator
TargetBuy	1	1	2753	24.7750	Organics Purchase Indicator

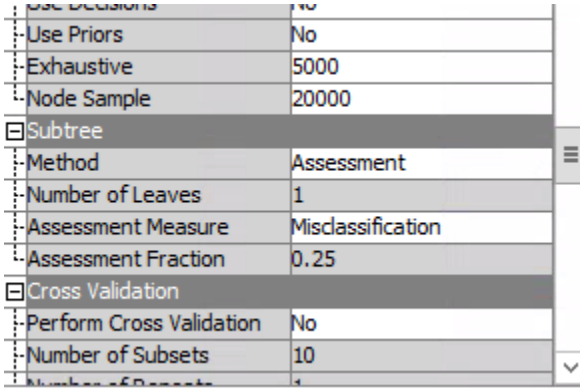
Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
TargetBuy	0	0	8359	75.2318	Organics Purchase Indicator
TargetBuy	1	1	2752	24.7682	Organics Purchase Indicator

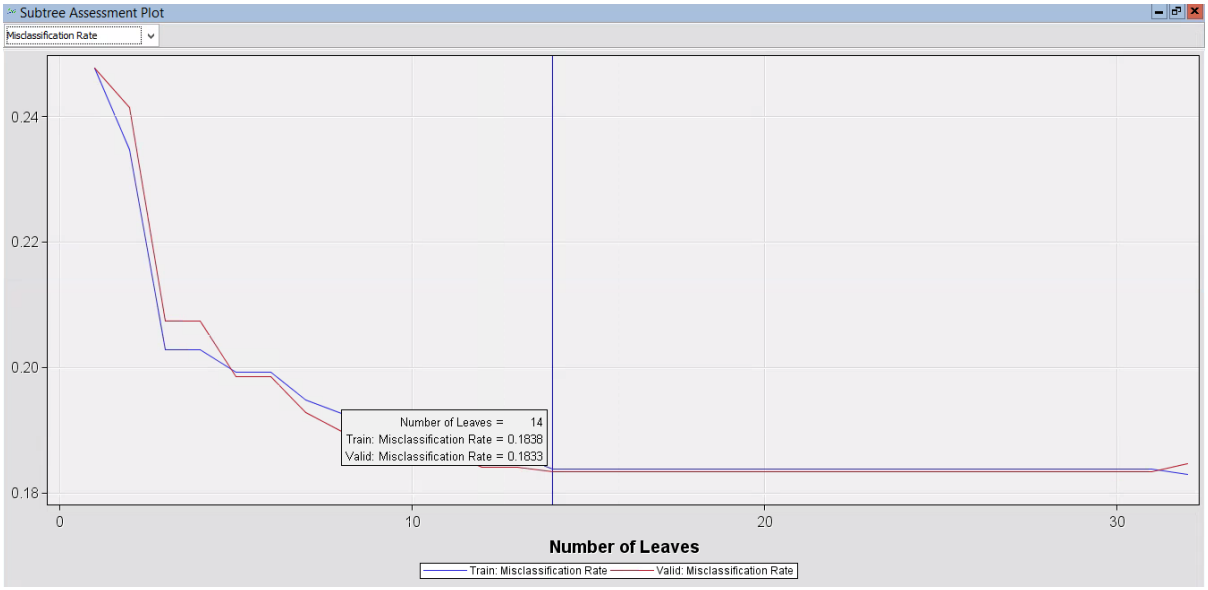
e. Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.



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



1) How many leaves are in the optimal tree?



The optimal tree has 14 leaves.

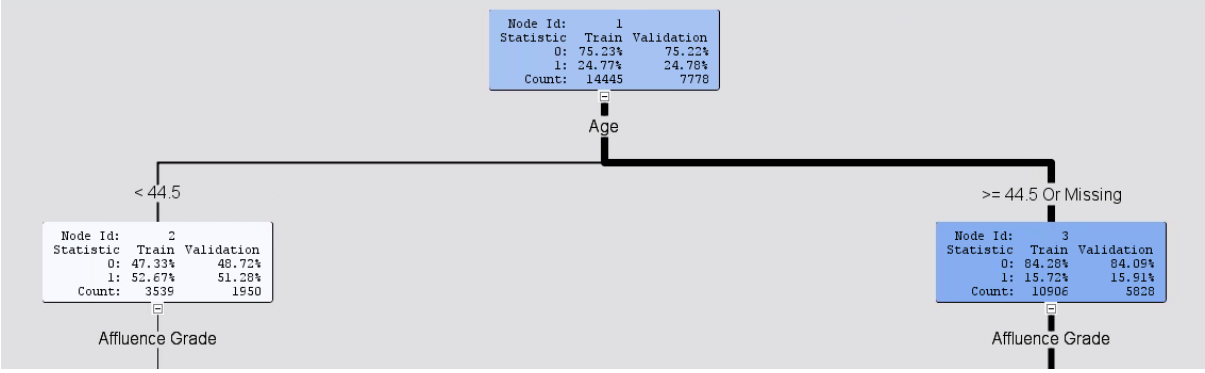
2) Which variables were used for the first split? What were the competing splits for this first split?

Variable Importance

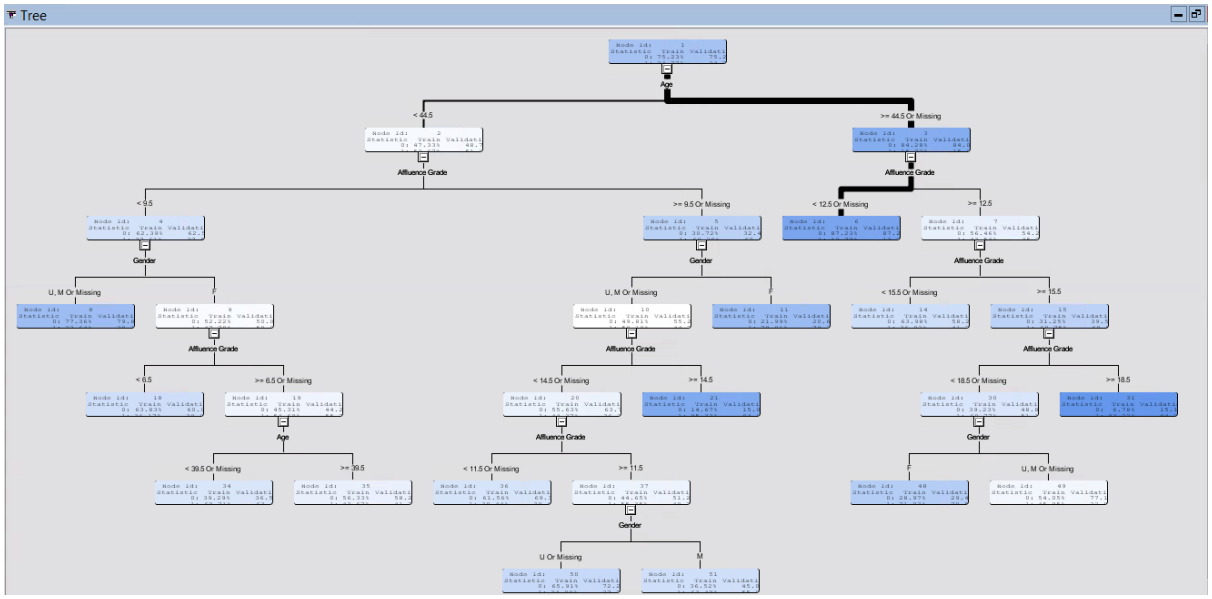
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
DemAge	Age	2	1.0000	1.0000	1.0000
DemAffl	Affluence G...	7	0.7806	0.7929	1.0158
DemGender	Gender	4	0.4090	0.5184	1.2674
PromSpend	Total Spend	0	0.0000	0.0000	.
DemCluste...	Neighborho...	0	0.0000	0.0000	.
DemReg	Geographic...	0	0.0000	0.0000	.
PromTime	Loyalty Car...	0	0.0000	0.0000	.
PromClass	Loyalty Stat...	0	0.0000	0.0000	.
DemTVReg	Television ...	0	0.0000	0.0000	.

From the above chart of Variable importance we can see that the variable Age is of high importance and selected as the first node.
DemAffl and DemGender can be considered as competing nodes to Age as seen in the

3) Which variables were used for the second split for all branches from first split?
DemAffl was used for the second split for all brances from first split.



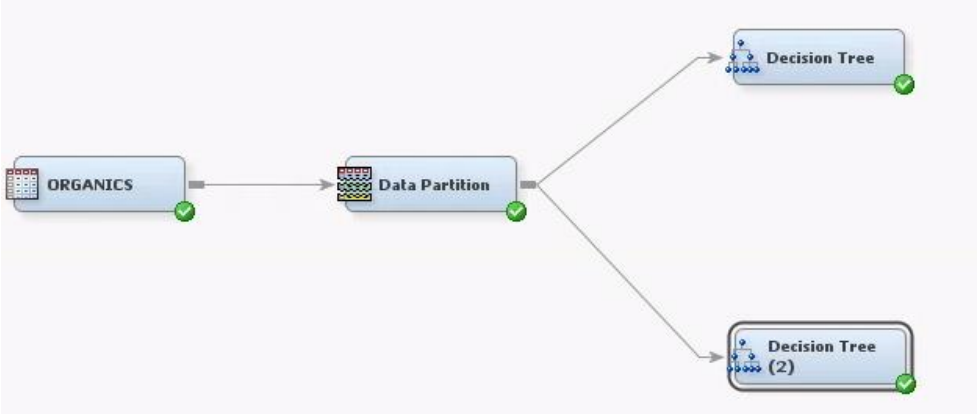
4) Discuss the results and provide your insights



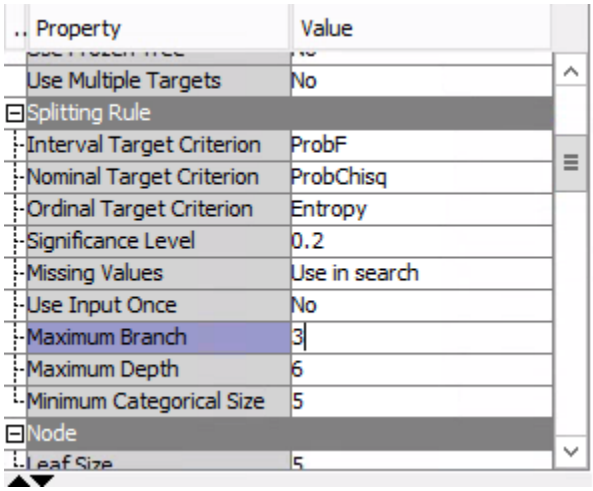
As per the tree constructed, the most influential attribute for classification is Age. In the training set out of the total records, 75.23% were classified as 0 and 24.77% were classified as 1.
Thus, all records are classified. Validation data is also classified in the same way. Training and validation makes the total observations in the dataset. This proves the importance of Age variable in this tree.

Out of 9 variables, only 3 variables are considered for the split. This has helped the compactness of the tree which is good.
However, the actual tree has 32 leaves but the optimal tree has 14 leaves which is quite a lot of difference. The misclassification rate for both training and validation is almost same. Hence the model is valid and is not having any overfitting.

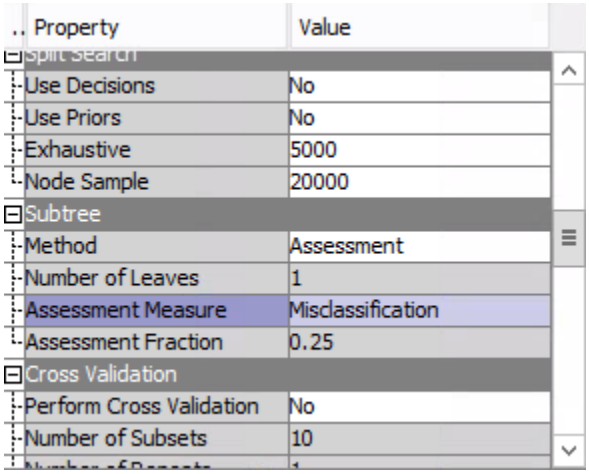
g. Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.



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.

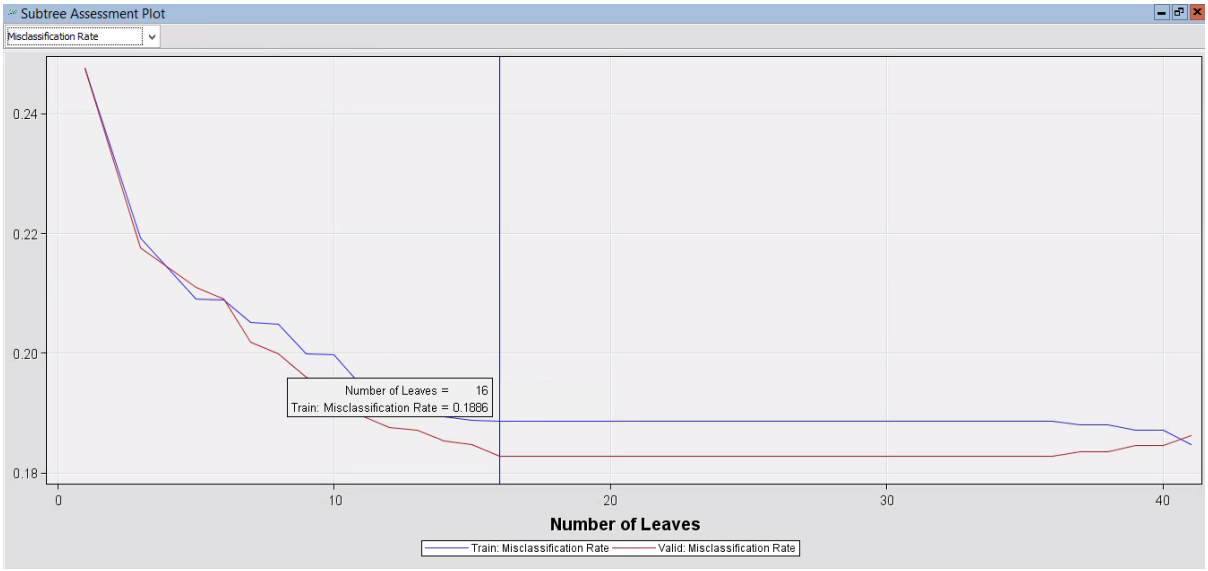


2) Create a decision tree model using **Misclassification** as the model assessment statistic.



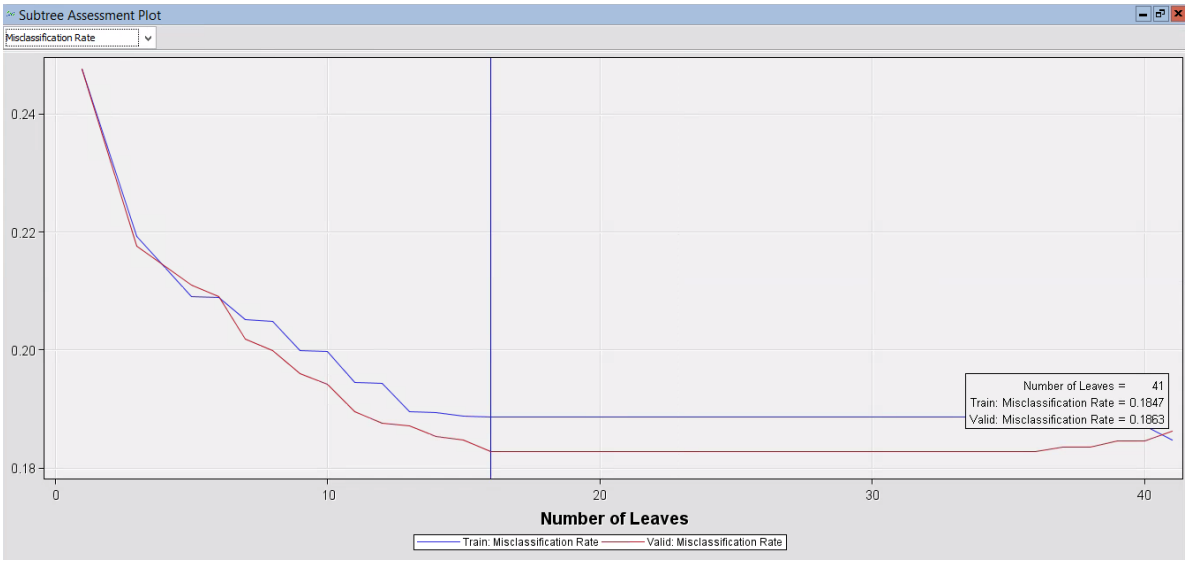
3) How many leaves are in the optimal tree?

As seen below, there are 16 leaves in a optimal tree in this case.

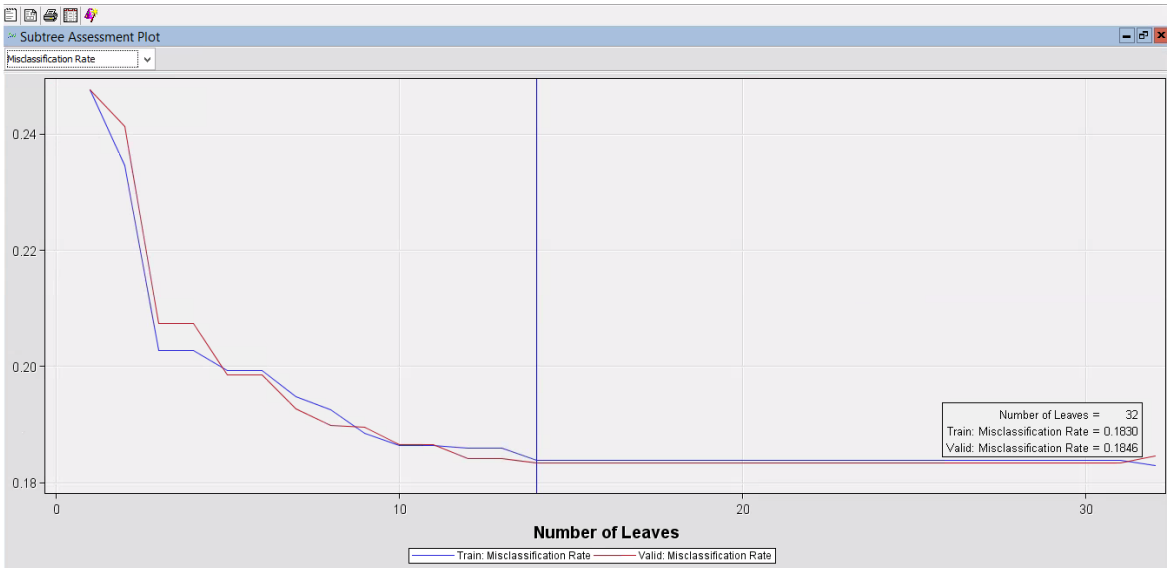


h. Based on **Misclassification rate**, which of the decision tree models appears to be better?

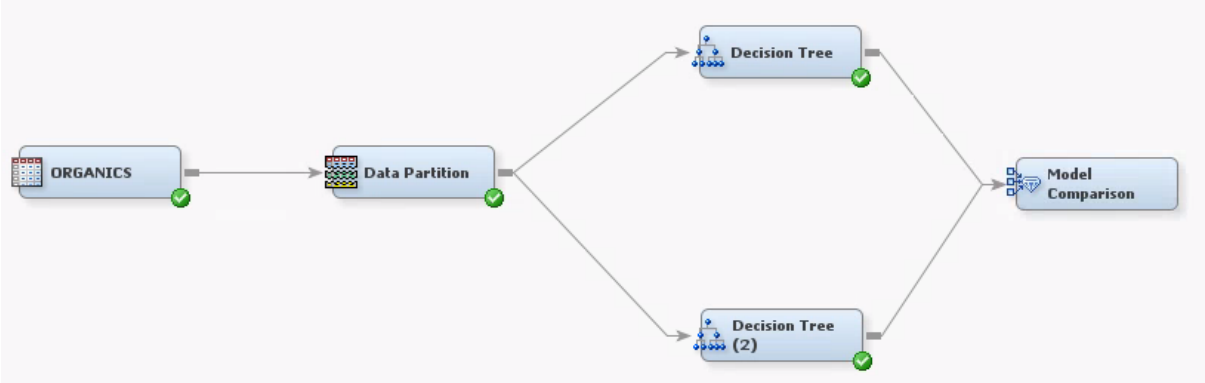
Tree 2:



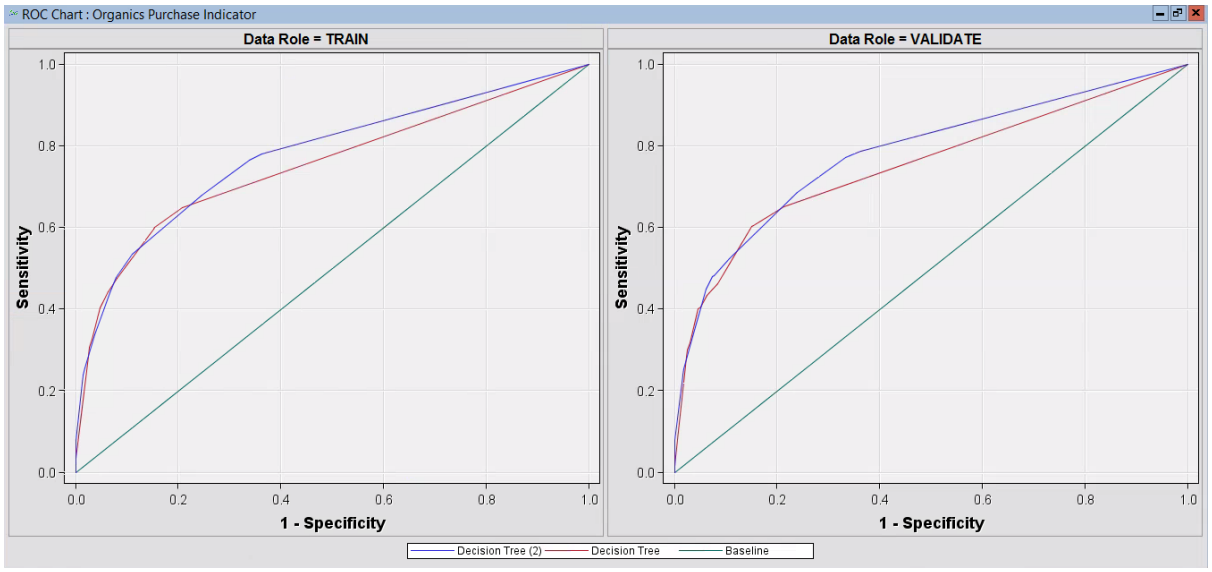
Tree 1:



Both the trees almost have the same misclassification rate. Hence we cannot make any conclusions from this.
We can use the model comparison node to compare both trees.



The area for the second decision tree is more when compared to first. Hence it is said to be better than the first decision tree.

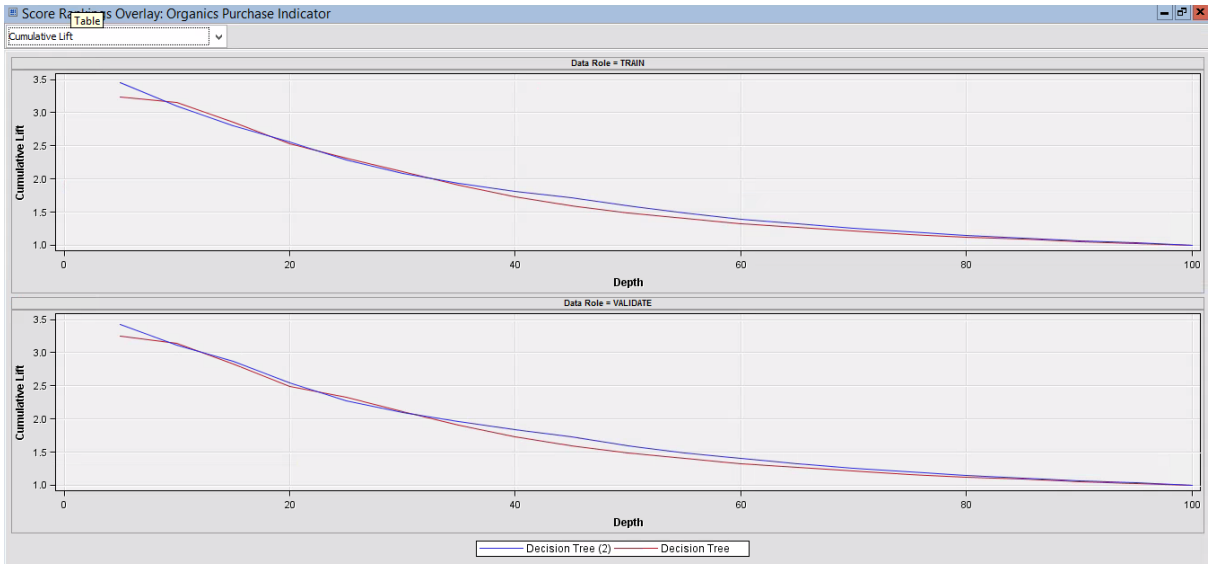


Also, average squared error for tree 1 is slightly more than tree 2 thus tree 2 is better. Also, for Tree2 the Average squared error is less, thus it is helpful in predicting customers who are interested in purchase of organic products.

Log tatistics

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	Decision Tr...	TargetBuy	Organics P...	0.182823	14445	0.188577	0.958333	4025.424	0.139336	0.373278	28890	14445	7778	0.182823	0.9583
	Tree	Tree	Decision Tr...	TargetBuy	Organics P...	0.183338	14445	0.183801	0.932203	4070.628	0.140901	0.375368	28890	14445	7778	0.183338	0.9322

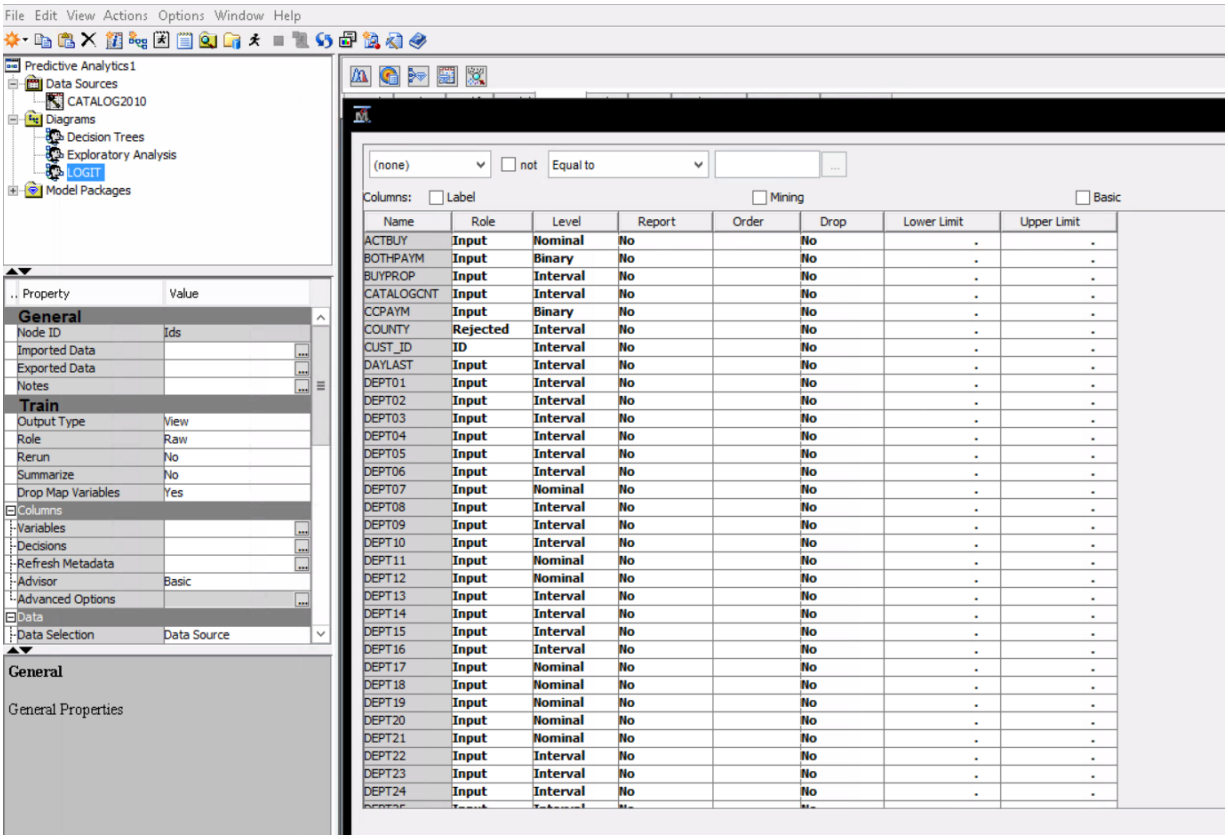
Cumulative Lift comparison for Tree1 and Tree2:



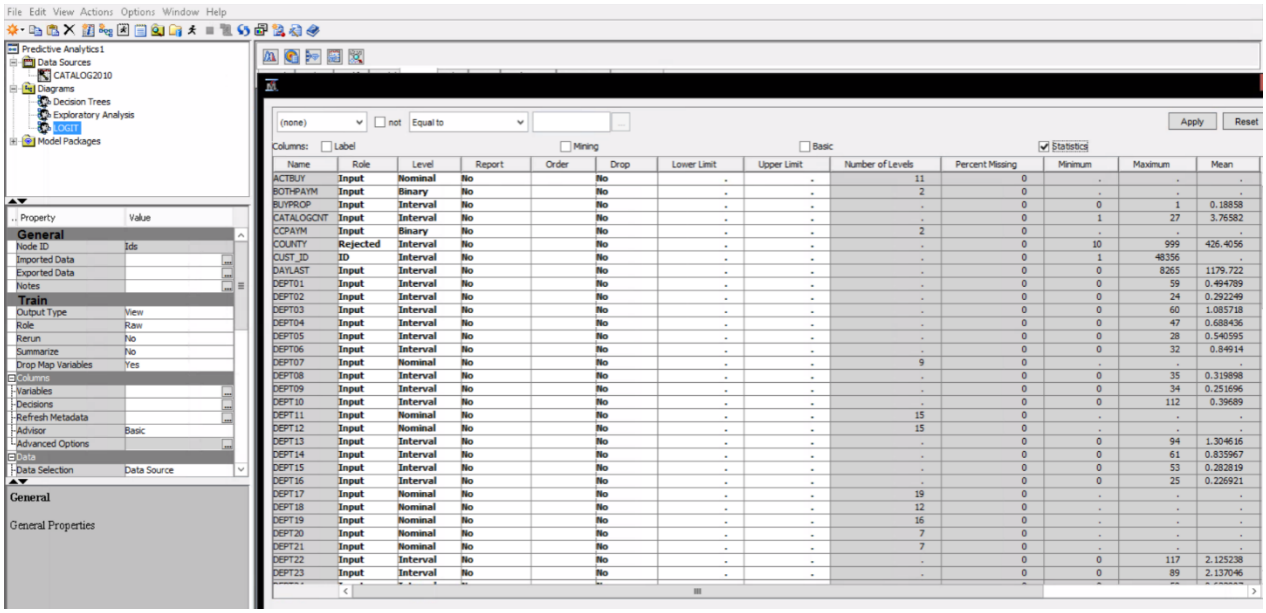
- From the above figure, we can see that both Lift and cumulative Lift values for Tree2 are higher than Tree1.Thus, we can say that Tree2 is better in predicting the customers who are interested in purchase of organic products.

TASK 4

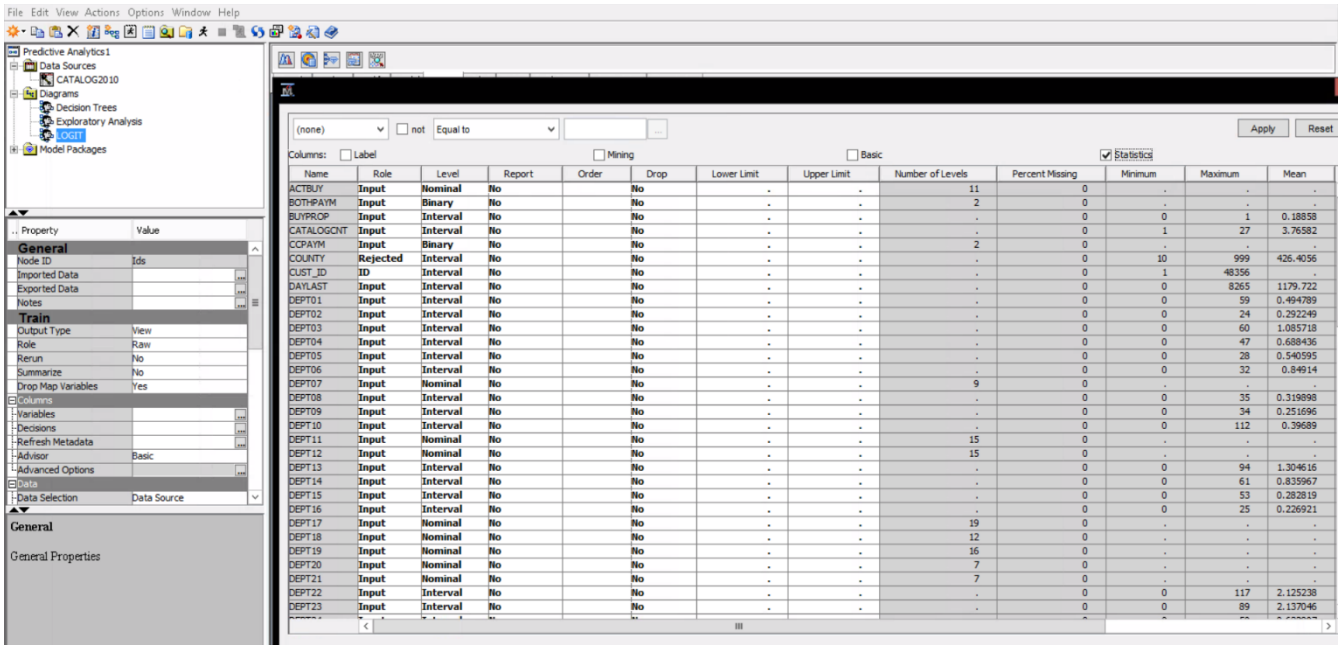
Review the CATALOG2010 data set. Select the CATALOG2010 data source in the decision tree diagram. From the panel on the left, click the ellipsis next to Variables.

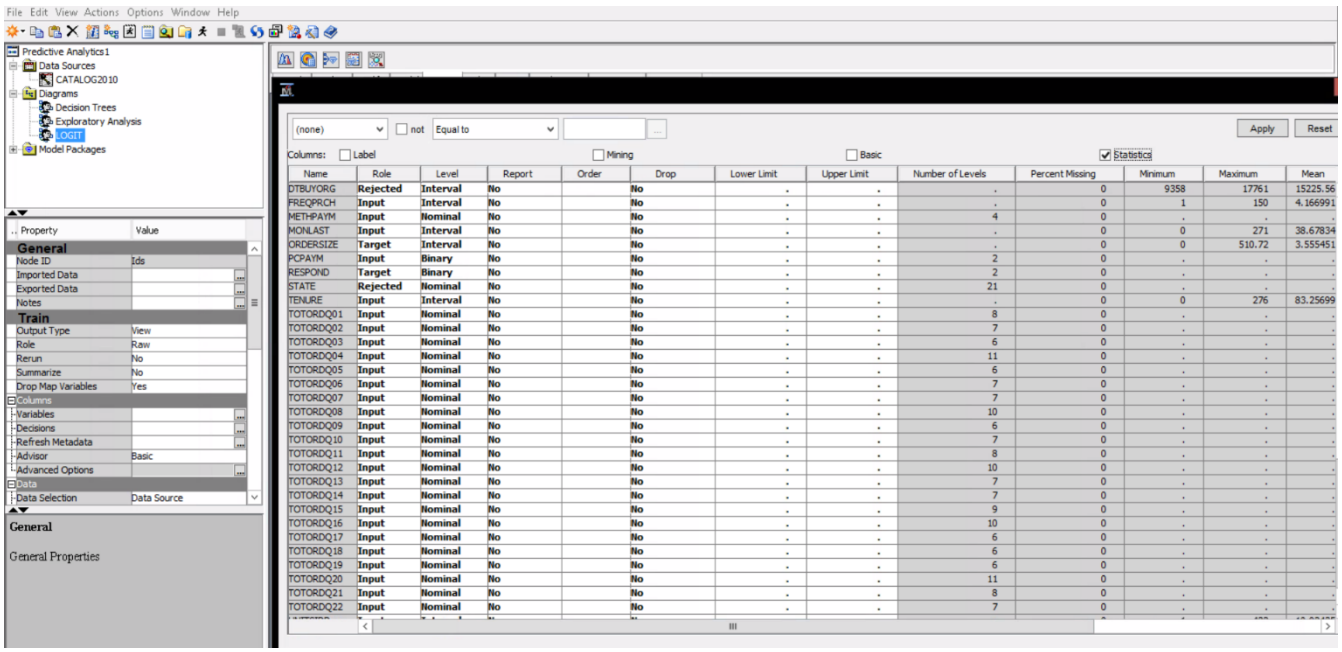


Compute some basic statistics. Select Statistics in the upper right corner



Inspect the results. a. Scroll down the list of variables. Notice that none of the variable have missing values. Therefore, the Imputation node is not necessary. b. Make sure that State has been assigned the role Rejected. c. Close the Variables window.

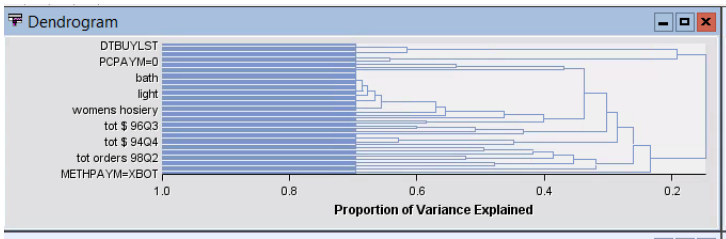




Cluster variables according to their similarities. a. Click the Explore tab and drag the Variable Clustering node into the diagram. b. Connect the Data Partition node to the Variable Clustering node. c. In the Properties panel, change the Includes Class Variables property to Yes and the Variable Selection property to Best Variables. (class variables are like categorical variables)

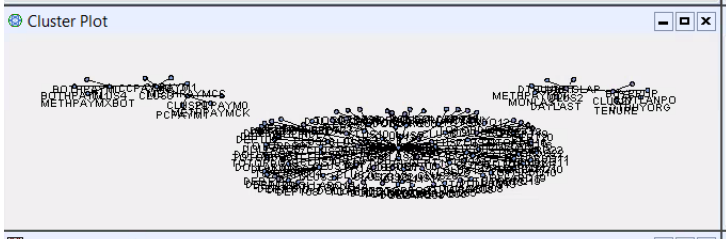
Property	Value
Variables	
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	Yes
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warning	No
Score	
Variable Selection	Best Variables
Interactive Selection	
Hides Rejected Variables	Yes
Status	
Create Time	12/12/17 3:21 AM
Run ID	
Last Error	

Right-click and run the Variable Clustering node. View the results.



Variable Frequency Table

Cluster	Frequency Count	Percent of Total Frequency
CLUS1	13	13.26531
CLUS10	3	3.061224
CLUS11	2	2.040816
CLUS12	2	2.040816
CLUS13	2	2.040816
CLUS14	2	2.040816
CLUS15	2	2.040816
CLUS16	2	2.040816
CLUS17	2	2.040816



Selected Variables

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	DOLINDET	total $ dem...	0.860923	CLUS30	0.445792	Variable	0.250947	YES
CLUS10	TOTORDQ20	tot orders 9...	0.818122	CLUS2	0.057868	Variable	0.19305	YES
CLUS11	DOLLARQ03	tot $ 93Q3	0.881598	CLUS1	0.044602	Variable	0.123929	YES
CLUS12	DOLLARQ22	tot $ 98Q2	0.863458	CLUS1	0.045071	Variable	0.142987	YES
CLUS13	DOLLARQ06	tot $ 94Q2	0.878785	CLUS1	0.090991	Variable	0.133348	YES

Variable Selection Table

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	Cluster 1	1	CLUS30	0.420462	ClusterComp	0	NO
CLUS1	DOLINDET	total $ dem...	0.860923	CLUS30	0.445792	Variable	0.250947	YES
CLUS1	DOLNETDT	avg $ net d...	0.852674	CLUS30	0.423187	Variable	0.255413	NO
CLUS1	UNITSIDD	tot units de...	0.842226	CLUS30	0.413054	Variable	0.268805	NO
CLUS1	CATALOGC...	number of c...	0.759258	CLUS30	0.409197	Variable	0.407482	NO

Output

```
1  *-------------------------------------------------------------*
2  User:                 hkatragl
3  Date:                 December 12, 2017
4  Time:                 12:42:53
5  *-------------------------------------------------------------*
6  * Training Output
7  *-------------------------------------------------------------*
8
9
10
```

The Selected Variables window shows one input for each cluster, chosen according to the 1-R 2 ratio. (If you want to override these decisions or add variables to the list of selected inputs, you can select the Interactive Selection property.)

- The Dendrogram window shows the hierarchical nature of the variable clusters.
- The Variable Frequency Table window reports how many inputs fall in each cluster.
- The Cluster Plot window offers an alternative to the tree diagram in the Dendrogram window.
- The Variable Selection window shows the variables that were (and were not) selected in each cluster.

Using the 1-R 2 ratio, the following variables were chosen from the clusters:

Selected Variables

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	DOLINDET	total $ dem...	0.860923	CLUS30	0.445792	Variable	0.250947	YES
CLUS10	TOTORDQ20	tot orders 9...	0.818122	CLUS2	0.057868	Variable	0.19305	YES
CLUS11	DOLLARQ03	tot $ 93Q3	0.881598	CLUS1	0.044602	Variable	0.123929	YES
CLUS12	DOLLARQ22	tot $ 98Q2	0.863458	CLUS1	0.045071	Variable	0.142987	YES
CLUS13	DOLLARQ06	tot $ 94Q2	0.878785	CLUS1	0.090991	Variable	0.133348	YES
CLUS14	TOTORDQ19	tot orders 9...	0.874679	CLUS1	0.046165	Variable	0.131387	YES
CLUS15	TOTORDQ11	tot orders 9...	0.872366	CLUS1	0.076733	Variable	0.138242	YES
CLUS16	DOLLARQ04	tot $ 93Q4	0.876755	CLUS1	0.054316	Variable	0.130323	YES
CLUS17	DOLLARQ05	tot $ 94Q1	0.871852	CLUS1	0.084252	Variable	0.139938	YES
CLUS18	DOLLARQ16	tot $ 96Q4	0.866462	CLUS1	0.08562	Variable	0.146043	YES
CLUS19	DOLLARQ18	tot $ 97Q2	0.879799	CLUS1	0.062006	Variable	0.128147	YES
CLUS2	DTBUYLST		0.957696	CLUS27	0.234074	Variable	0.055232	YES
CLUS20	TOTORDQ14	tot orders 9...	0.84481	CLUS1	0.061662	Variable	0.165388	YES
CLUS21	TOTORDQ21	tot orders 9...	0.860266	CLUS1	0.042824	Variable	0.145985	YES
CLUS22	DOLLARQ09	tot $ 95Q1	0.873589	CLUS1	0.076138	Variable	0.136829	YES
CLUS23	DOLLARQ02	tot $ 93Q2	0.869018	CLUS1	0.095522	Variable	0.144816	YES
CLUS24	DOLLARQ01	tot $ 93Q1	0.875695	CLUS1	0.107737	Variable	0.139314	YES
CLUS25	DOLLARQ07	tot $ 94Q3	0.869525	CLUS1	0.083166	Variable	0.14231	YES
CLUS26	TOTORDQ13	tot orders 9...	0.845289	CLUS1	0.072906	Variable	0.166877	YES
CLUS27	TENURE	months sin...	0.935694	CLUS2	0.193821	Variable	0.079766	YES
CLUS28	DOLLARQ08	tot $ 94Q4	0.853544	CLUS1	0.095589	Variable	0.161936	YES
CLUS29	METHPAYM...	METHPAYM...	1	CLUS3	0.3108	Variable	0	YES
CLUS3	CCPAYM0	CCPAYM=0	1	CLUS29	0.3108	Variable	0	YES
CLUS30	DEPT03	womens un...	0.473772	CLUS1	0.202037	Variable	0.659464	YES
CLUS31	DEPT12	mens misc	0.367295	CLUS1	0.100647	Variable	0.703512	YES
CLUS32	DEPT24	health	0.680643	CLUS1	0.164481	Variable	0.382226	YES
CLUS33	DEPT18	chair	0.385083	CLUS1	0.048191	Variable	0.646051	YES
CLUS4	BOTHPAYM0	BOTHPAYM...	1	CLUS3	0.171641	Variable	5.36E-16	YES
CLUS5	DOLLARQ17	tot $ 97Q1	0.81237	CLUS1	0.073334	Variable	0.202479	YES
CLUS6	TOTORDQ12	tot orders 9...	0.732581	CLUS1	0.080548	Variable	0.290846	YES
CLUS7	TOTORDQ15	tot orders 9...	0.872407	CLUS1	0.069808	Variable	0.137168	YES
CLUS8	DOLINDEA	avg $ dema...	0.976948	CLUS1	0.057633	Variable	0.024461	YES
CLUS9	DOLLARQ10	tot $ 95Q2	0.880676	CLUS1	0.080429	Variable	0.129761	YES

These variables are used as candidates in logistic regression models.

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-R 2 ratio as the cluster representative

Cluster 2	BUYPROP	0.1627	0.2604	1.1322	% quarters w/buy
	DAYLAST	0.8785	0.0960	0.1344	days since last
	DTBUYLST	0.8785	0.0960	0.1344	
	DTBUYORG	0.4575	0.0975	0.6011	
	MONLAST	0.8785	0.0960	0.1344	months since last
	TENURE	0.4574	0.0975	0.6013	months since lst
	UNITSLAP	0.1580	0.2208	1.0807	avg price/unit
	UNTLANPO	0.0792	0.2644	1.2517	avg units/order
	METHPAYMDK	0.6157	0.0410	0.4007	METHPAYM=DK
Cluster 3	CCPAYM0	0.7787	0.1716	0.2671	CCPAYM=0
	CCPAYM1	0.7787	0.1716	0.2671	CCPAYM=1
	METHPAYMCC	0.7787	0.1716	0.2671	METHPAYM=CC
	METHPAYMCK	0.7787	0.1043	0.2470	METHPAYM=CK
	PCPAYM0	0.7787	0.1043	0.2470	PCPAYM=0
	PCPAYM1	0.7787	0.1043	0.2470	PCPAYM=1
Cluster 4	BOTHPAYM0	1.0000	0.1174	0.0000	BOTHPAYM=0
	BOTHPAYM1	1.0000	0.1174	0.0000	BOTHPAYM=1
	METHPAYMXBOT	1.0000	0.1174	0.0000	METHPAYM=XBOT
Cluster 5	DOLL24	0.5745	0.3124	0.6188	$ last 24 months
	DOLLARQ17	0.4610	0.0695	0.5793	tot $ 97Q1
	DOLLARQ18	0.4704	0.0618	0.5645	tot $ 97Q2
	TOTORDQ17	0.4254	0.0657	0.6150	tot orders 97Q1
	TOTORDQ18	0.4595	0.0664	0.5790	tot orders 97Q2
Cluster 6	ACTBUY	0.5165	0.3532	0.7475	num qrtrs w/buy
	DEPT25	0.3382	0.1927	0.8198	food
	DOLLARQ08	0.3844	0.0979	0.6824	tot $ 94Q4
	DOLLARQ12	0.4244	0.1022	0.6412	tot $ 95Q4
	TOTORDQ08	0.4753	0.0923	0.5781	tot orders 94Q4
	TOTORDQ12	0.4976	0.0802	0.5462	tot orders 95Q4

Cluster 7	DOLLARQ13	0.3299	0.0872	0.7341	tot $ 96Q1
	DOLLARQ14	0.2869	0.0632	0.7611	tot $ 96Q2
	DOLLARQ15	0.4358	0.0807	0.6137	tot $ 96Q3
	TOTORDQ13	0.3231	0.0746	0.7314	tot orders 96Q1
	TOTORDQ14	0.3201	0.0636	0.7260	tot orders 96Q2
	TOTORDQ15	0.4431	0.0711	0.5995	tot orders 96Q3
Cluster 8	DOLINDEA	0.9769	0.0537	0.0244	avg $ demand
	DOLNETDA	0.9769	0.0543	0.0244	tot $ net demand
Cluster 9	DOLLARQ09	0.4780	0.0766	0.5653	tot $ 95Q1
	DOLLARQ10	0.5041	0.0835	0.5410	tot $ 95Q2
	TOTORDQ09	0.4976	0.0866	0.5500	tot orders 95Q1
	TOTORDQ10	0.5275	0.0872	0.5176	tot orders 95Q2
Cluster 10	DEPT26	0.2482	0.0459	0.7879	gift
	DOLLARQ20	0.8011	0.0651	0.2128	tot $ 97Q4
	TOTORDQ20	0.8181	0.0686	0.1953	tot orders 97Q4
Cluster 11	DOLLARQ03	0.8816	0.0461	0.1241	tot $ 93Q3
	TOTORDQ03	0.8816	0.0505	0.1247	tot orders 93Q3
Cluster 12	DOLLARQ21	0.4791	0.0491	0.5479	tot $ 98Q1
	DOLLARQ22	0.4929	0.0454	0.5312	tot $ 98Q2
	TOTORDQ21	0.4765	0.0439	0.5476	tot orders 98Q1
	TOTORDQ22	0.4913	0.0454	0.5329	tot orders 98Q2
Cluster 13	DOLLARQ06	0.5170	0.0949	0.5336	tot $ 94Q2
	DOLLARQ07	0.4743	0.0870	0.5758	tot $ 94Q3
	TOTORDQ06	0.5263	0.1024	0.5277	tot orders 94Q2
	TOTORDQ07	0.4987	0.0964	0.5548	tot orders 94Q3

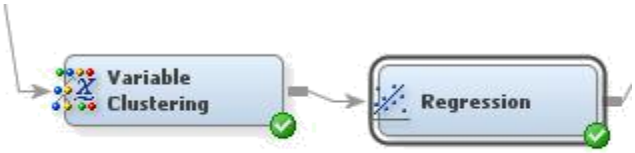
There are 33 clusters and, therefore, 33 variables selected.

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

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	14.269655	0.1486	0.1486	5.075727	0.0043	
2	18.878088	0.1966	0.1578	3.699895	0.0106	0.9916
3	21.903078	0.2282	0.1877	2.748968	0.0111	0.9918
4	24.167157	0.2517	0.1877	2.623513	0.0111	0.9916
5	26.705701	0.2782	0.1893	2.073863	0.0113	0.9914
6	28.466963	0.2965	0.1958	2.023150	0.0113	0.9927
7	30.338115	0.3160	0.1958	1.906582	0.0113	1.0351
8	31.993414	0.3333	0.2145	1.829092	0.0117	1.0203
9	33.791416	0.3520	0.2145	1.734839	0.0117	1.1004
10	35.400976	0.3688	0.2252	1.661294	0.0119	1.0955
11	37.045861	0.3859	0.2333	1.645777	0.0120	1.0950
12	38.514796	0.4012	0.2333	1.641035	0.0120	1.0950
13	40.075066	0.4174	0.2462	1.578818	0.0122	1.0903
14	41.568625	0.4330	0.2462	1.555217	0.0122	1.0903
15	43.111311	0.4491	0.2462	1.548926	0.0122	1.0903
16	44.584612	0.4644	0.2462	1.545298	0.0122	1.0903
17	46.100224	0.4802	0.2538	1.537361	0.0124	1.0892
18	47.580873	0.4956	0.2538	1.522834	0.0124	1.0892
19	48.978020	0.5102	0.2538	1.518141	0.0124	1.0892
20	50.489323	0.5259	0.2538	1.507744	0.0124	1.0892
21	51.997062	0.5416	0.2538	1.501807	0.0124	1.0892
22	53.498395	0.5573	0.2538	1.499339	0.0124	1.0892
23	54.982973	0.5727	0.2615	1.482932	0.0126	1.0875
24	56.453832	0.5881	0.2693	1.480621	0.0125	1.0875
25	57.934102	0.6035	0.2693	1.474760	0.0125	1.0875
26	59.408841	0.6188	0.2693	1.400954	0.0125	1.0875
27	60.681938	0.6321	0.2693	1.327516	0.0125	1.0875
28	62.009454	0.6459	0.2693	1.255397	0.0125	1.0875
29	63.032289	0.6566	0.2862	1.237242	0.0133	1.0759
30	63.821130	0.6648	0.2862	1.203431	0.0133	1.0759
31	64.942290	0.6765	0.2631	1.137188	0.0138	1.0682
32	65.829529	0.6857	0.2631	1.043253	0.0167	1.0147
33	66.816439	0.6960	0.2631	1.043149	0.0326	1.0143

The clusters explained 69.6% 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.

Click the Model tab and drag the Regression node into the diagram workspace. Connect the Variable Clustering node to the Regression node.



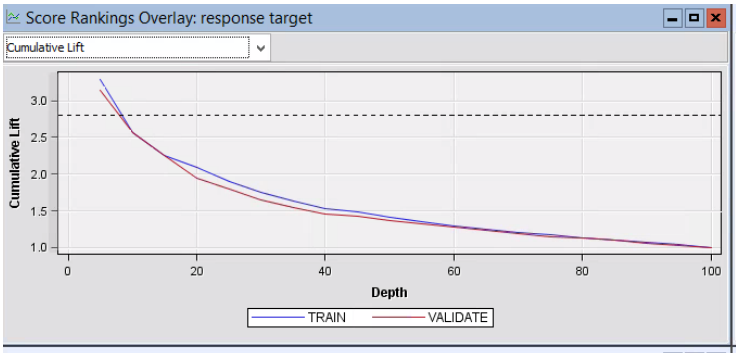
Select Selection Model

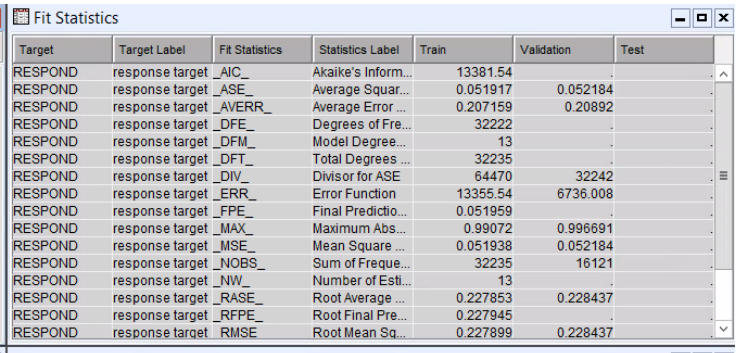
Forward and select Selection Criterion

Validation Error from the Regression node Properties panel.

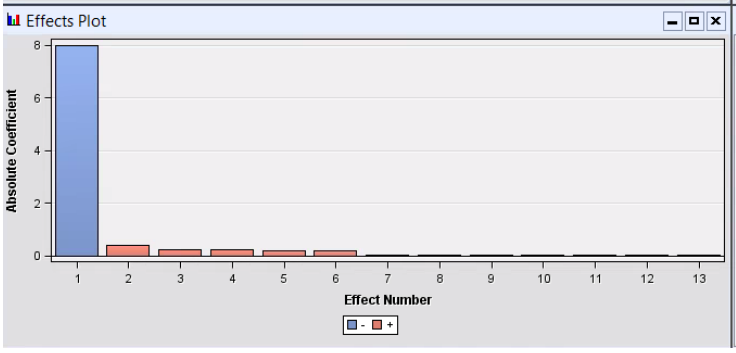
Property	Value
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes

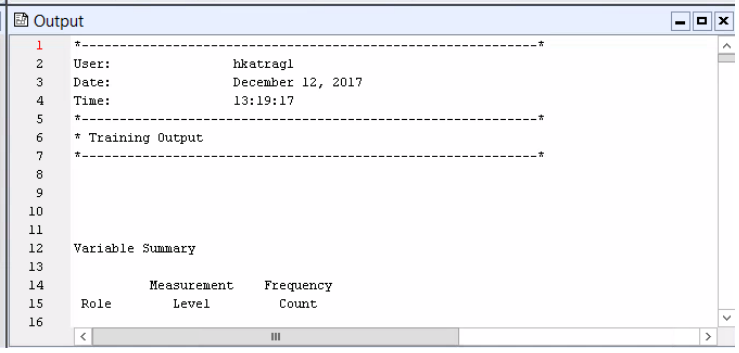
Right-click the Regression node and click Run. View the results



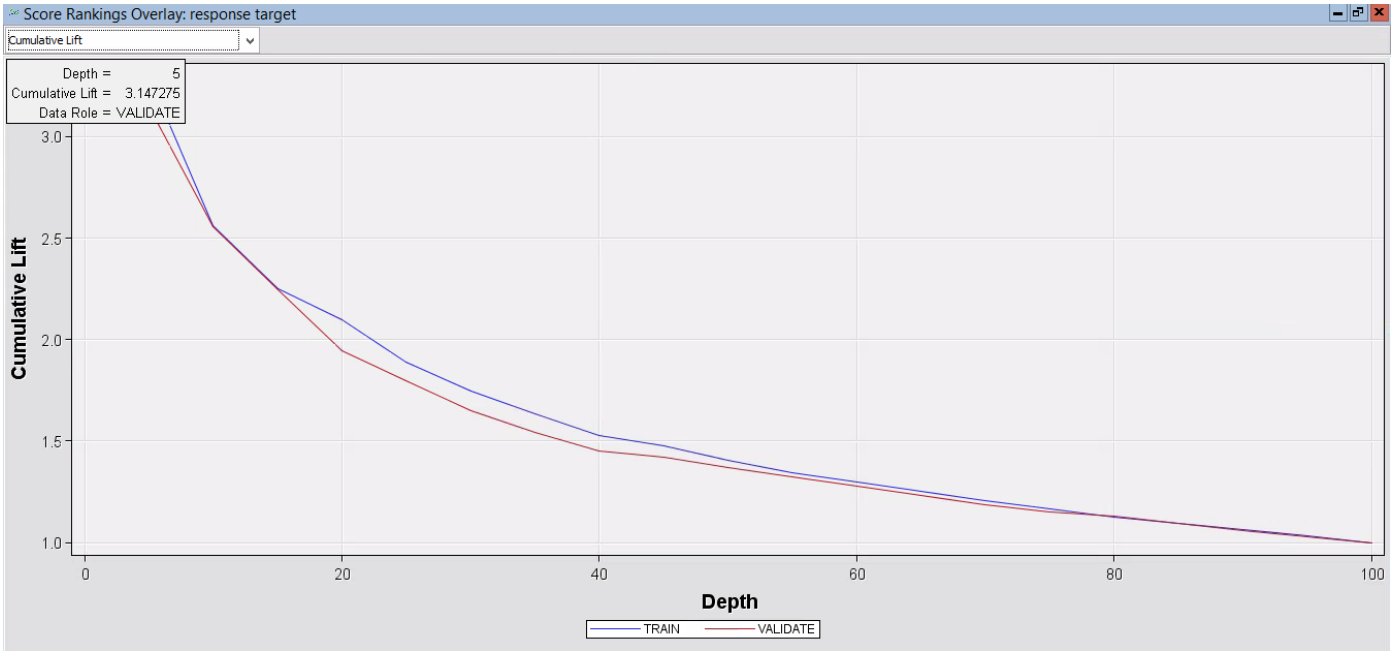


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RESPOND	response target	_AIC_	Akaike's Inform...	13381.54	.	.
RESPOND	response target	_ASE_	Average Squar...	0.051917	0.052184	.
RESPOND	response target	_AVERR_	Average Error ...	0.207159	0.20892	.
RESPOND	response target	_DFE_	Degrees of Fre...	32222	.	.
RESPOND	response target	_DFM_	Model Degree...	13	.	.
RESPOND	response target	_DFT_	Total Degrees ...	32235	.	.
RESPOND	response target	_DIV_	Divisor for ASE	64470	32242	.
RESPOND	response target	_ERR_	Error Function	13355.54	6736.008	.
RESPOND	response target	_FPE_	Final Predictio...	0.051959	.	.
RESPOND	response target	_MAX_	Maximum Abs...	0.99072	0.996691	.
RESPOND	response target	_MSE_	Mean Square ...	0.051938	0.052184	.
RESPOND	response target	_NOBS_	Sum of Freque...	32235	16121	.
RESPOND	response target	_NW_	Number of Esti...	13	.	.
RESPOND	response target	_RASE_	Root Average ...	0.227853	0.228437	.
RESPOND	response target	_RFPE_	Root Final Pre...	0.227945	.	.
RESPOND	response target	_RMSE	Root Mean Sq...	0.227899	0.228437	.



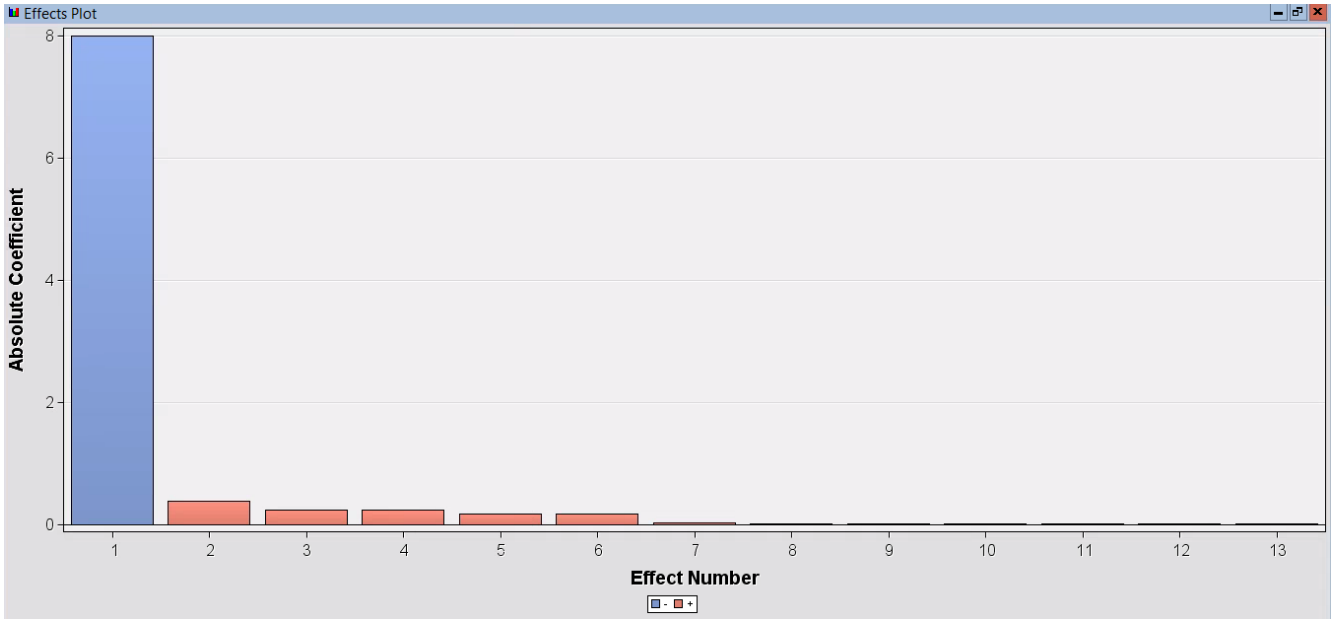


The Score Rankings Overlay window shows a cumulative lift chart where, for a given percentile, you can see the lift of the model.



By positioning the mouse cursor over a point along the lift curve for the validation data, you can see a pop-up flag with information about the percentile and lift. For example, at the 5th percentile, the lift is 3.15 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 you would obtain 3.15 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.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RESPOND	response target	_AIC_	Akaike's Information Criterion	13381.54	.	.
RESPOND	response target	_ASE_	Average Squared Error	0.051917	0.052184	.
RESPOND	response target	_AVERR_	Average Error Function	0.207159	0.20892	.
RESPOND	response target	_DFE_	Degrees of Freedom for Error	32222	.	.
RESPOND	response target	_DFM_	Model Degrees of Freedom	13	.	.
RESPOND	response target	_DFT_	Total Degrees of Freedom	32235	.	.
RESPOND	response target	_DIV_	Divisor for ASE	64470	32242	.
RESPOND	response target	_ERR_	Error Function	13355.54	6736.008	.
RESPOND	response target	_FPE_	Final Prediction Error	0.051959	.	.
RESPOND	response target	_MAX_	Maximum Absolute Error	0.99072	0.996691	.
RESPOND	response target	_MSE_	Mean Square Error	0.051938	0.052184	.
RESPOND	response target	_NOBS_	Sum of Frequencies	32235	16121	.
RESPOND	response target	_NW_	Number of Estimate Weights	13	.	.
RESPOND	response target	_RASE_	Root Average Sum of Squares	0.227853	0.228437	.
RESPOND	response target	_RFPE_	Root Final Prediction Error	0.227945	.	.
RESPOND	response target	_RMSE_	Root Mean Squared Error	0.227899	0.228437	.
RESPOND	response target	_SBC_	Schwarz's Bayesian Criterion	13490.49	.	.
RESPOND	response target	_SSE_	Sum of Squared Errors	3347.086	1682.505	.
RESPOND	response target	_SUMW_	Sum of Case Weights Times Freq	64470	32242	.
RESPOND	response target	_MISC_	Misclassification Rate	0.056957	0.05682	.

The Output window gives the standard output for logistic regression.

The initial lines of the Output window summarize the roles of the variables used (or not) by the Regression node. The model has 33 inputs that predict a binary target.

Output

```
*------------------------------------------------------------*
User:                  hkatragl
Date:                  December 12, 2017
Time:                  13:19:17
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
 Role           Level         Count

INPUT       INTERVAL           33
TARGET      BINARY              1



Model Events

                                    Number
                        Measurement    of
Target     Event          Level      Levels      Order            Label

RESPOND      1          BINARY         2      Descending    response target

```

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	34
Number of Observations	32235

Target Profile

Ordered Value	RESPOND	Total Frequency
1	1	1825
2	0	30410

The output of the forward selection method shows the results of each model fitted in each step. The output below shows the results of the final model.

The Summary of Forward Selection table shows the variables that were selected in the forward selection method. This model has 12 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	DTBUYLST	1	3	113.4062	<.0001	6781.3
4	DEPT03	1	4	26.5051	<.0001	6775.7
5	TOTORDQ21	1	5	18.1446	<.0001	6767.5
6	CCPAYMO	1	6	16.2387	<.0001	6774.0
7	TOTORDQ12	1	7	14.8674	0.0001	6761.4
8	DOLLARQ18	1	8	14.4541	0.0001	6758.8
9	DOLLARQ22	1	9	14.0805	0.0002	6746.0
10	TOTORDQ19	1	10	13.1080	0.0003	6745.5
11	TENURE	1	11	9.9074	0.0016	6737.5
12	DOLLARQ16	1	12	4.5804	0.0323	6736.0

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	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
14025.546	13355.537	670.0087	12	<.0001

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 DEPT03 corresponds to a .029 increase in the log odds of purchasing a product from the catalog, adjusted for the other predictor variables. 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, DTBUYLST is the most important predictor variable followed by TOTORDQ20 and TENURE.

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 in one unit of total orders placed yields a 1.469 times increase in the odds of purchasing a product from the catalog.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-8.0025	0.5399	219.73	<.0001		0.000
CCPAYMO	1	0.1739	0.0516	11.34	0.0008	0.0473	1.190
DEPT03	1	0.0288	0.00692	17.28	<.0001	0.0447	1.029
DOLINDET	1	0.000112	0.000078	2.06	0.1515	0.0193	1.000
DOLLARQ16	1	0.00130	0.000611	4.56	0.0328	0.0223	1.001
DOLLARQ18	1	0.00290	0.000770	14.21	0.0002	0.0358	1.003
DOLLARQ22	1	0.00262	0.000681	14.81	0.0001	0.0359	1.003
DTBUYLST	1	0.000278	0.000031	77.78	<.0001	0.1878	1.000
TENURE	1	0.00156	0.000472	10.95	0.0009	0.0517	1.002
TOTORDQ12	1	0.1727	0.0472	13.39	0.0003	0.0388	1.189
TOTORDQ19	1	0.2311	0.0621	13.84	0.0002	0.0390	1.260
TOTORDQ20	1	0.3845	0.0430	79.87	<.0001	0.0979	1.469
TOTORDQ21	1	0.2297	0.0585	15.42	<.0001	0.0418	1.258

Odds Ratio Estimates

Effect	Point Estimate
CCPAYMO	1.190
DEPTO3	1.029
DOLINDET	1.000
DOLLARQ16	1.001
DOLLARQ18	1.003
DOLLARQ22	1.003
DTBUYLST	1.000
TENURE	1.002
TOTORDQ12	1.189
TOTORDQ19	1.260
TOTORDQ20	1.469
TOTORDQ21	1.258

The output also shows the assessment statistics for the validation data set for the 5th percentile, the10th percentile, and so on.

Data Role=VALIDATE Target Variable=RESPOND Target Label=response target

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	2	1	0.98429	0.0186
0.80-0.85	1	1	0.81973	0.0124
0.75-0.80	1	1	0.76478	0.0124
0.70-0.75	0	2	0.72573	0.0124
0.65-0.70	0	3	0.68638	0.0186
0.60-0.65	1	1	0.62709	0.0124
0.55-0.60	1	0	0.59112	0.0062
0.50-0.55	2	1	0.52241	0.0186
0.45-0.50	0	2	0.46860	0.0124
0.40-0.45	3	2	0.41914	0.0310
0.35-0.40	8	6	0.37403	0.0868
0.30-0.35	7	14	0.32051	0.1303
0.25-0.30	4	26	0.26919	0.1861
0.20-0.25	18	60	0.21818	0.4838
0.15-0.20	32	162	0.17086	1.2034
0.10-0.15	111	713	0.11859	5.1113
0.05-0.10	410	5844	0.06730	38.7941
0.00-0.05	313	8368	0.03588	53.8490

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

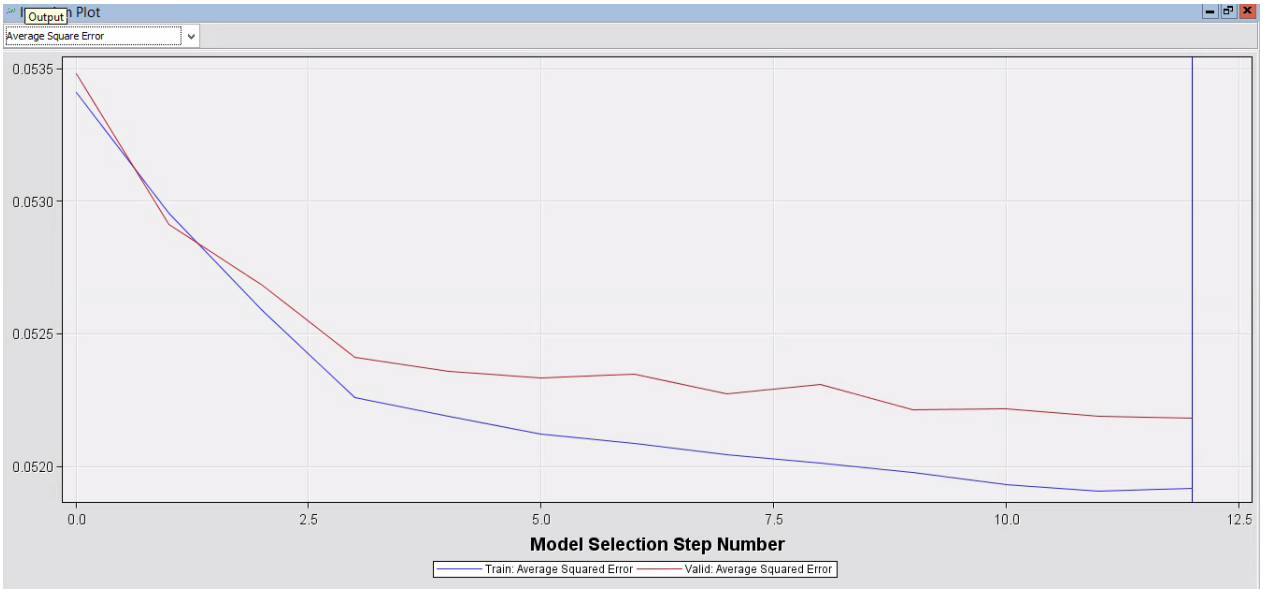
Data Role=VALIDATE Target Variable=RESPOND Target Label=response target

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	214.728	3.14728	3.14728	17.8439	17.8439	807	0.17814
10	155.875	1.96949	2.55875	11.1663	14.5071	806	0.10029
15	124.575	1.61936	2.24575	9.1811	12.7325	806	0.08483
20	94.700	1.05039	1.94700	5.9553	11.0388	806	0.07532
25	79.835	1.20358	1.79835	6.8238	10.1960	806	0.06895
30	65.184	0.91909	1.65184	5.2109	9.3653	806	0.06337
35	54.093	0.87533	1.54093	4.9628	8.7365	806	0.05990
40	45.227	0.83156	1.45227	4.7146	8.2338	806	0.05565
45	42.221	1.18169	1.42221	6.6998	8.0634	806	0.05230
50	36.972	0.89721	1.36972	5.0868	7.7658	806	0.04925
55	32.279	0.85344	1.32279	4.8387	7.4997	806	0.04650
60	27.639	0.76591	1.27639	4.3424	7.2366	806	0.04393
65	23.207	0.70026	1.23207	3.9702	6.9854	806	0.04169
70	18.784	0.61273	1.18784	3.4739	6.7346	806	0.03948
75	14.804	0.59085	1.14804	3.3499	6.5090	806	0.03740
80	12.826	0.83156	1.12826	4.7146	6.3968	806	0.03504
85	9.279	0.52520	1.09279	2.9777	6.1957	806	0.03227
90	5.640	0.43766	1.05640	2.4814	5.9894	806	0.02902
95	2.729	0.50331	1.02729	2.8536	5.8244	806	0.02470
100	0.000	0.48143	1.00000	2.7295	5.6696	806	0.01876

View model performance across the fitted models.

Select View -> Model -> Iteration Plot in the Results window. The Iteration Plot window appears.

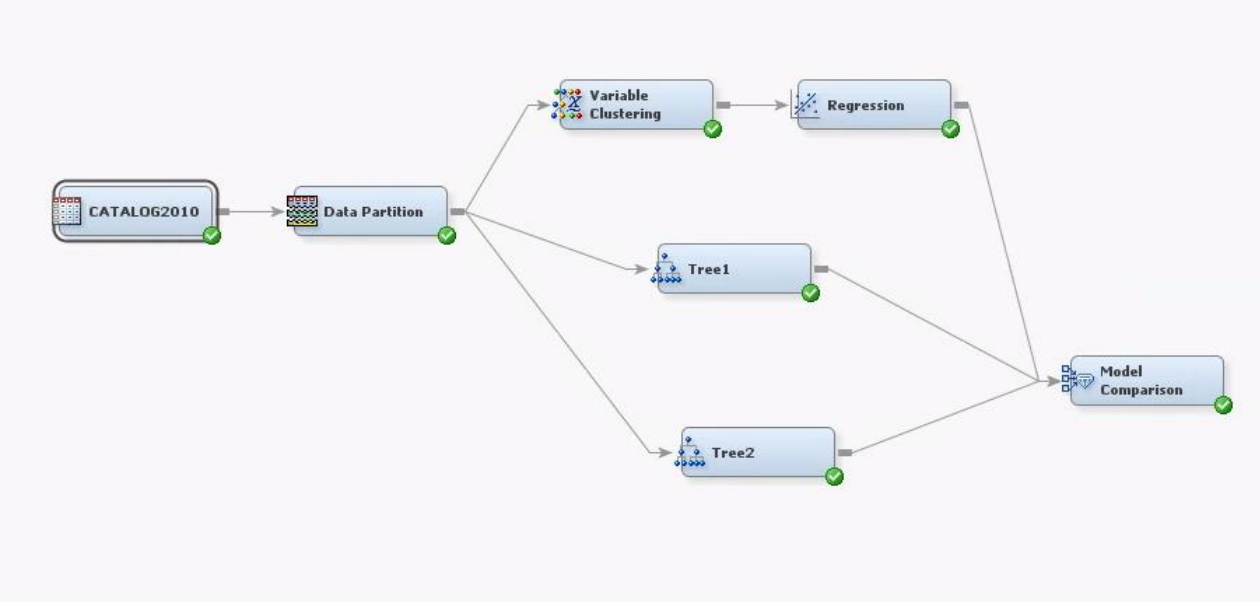
The Iteration Plot window shows the average squared error (training and validation) from the model selected in each step of the backward selection process. The smallest average squared error occurs in model 8.



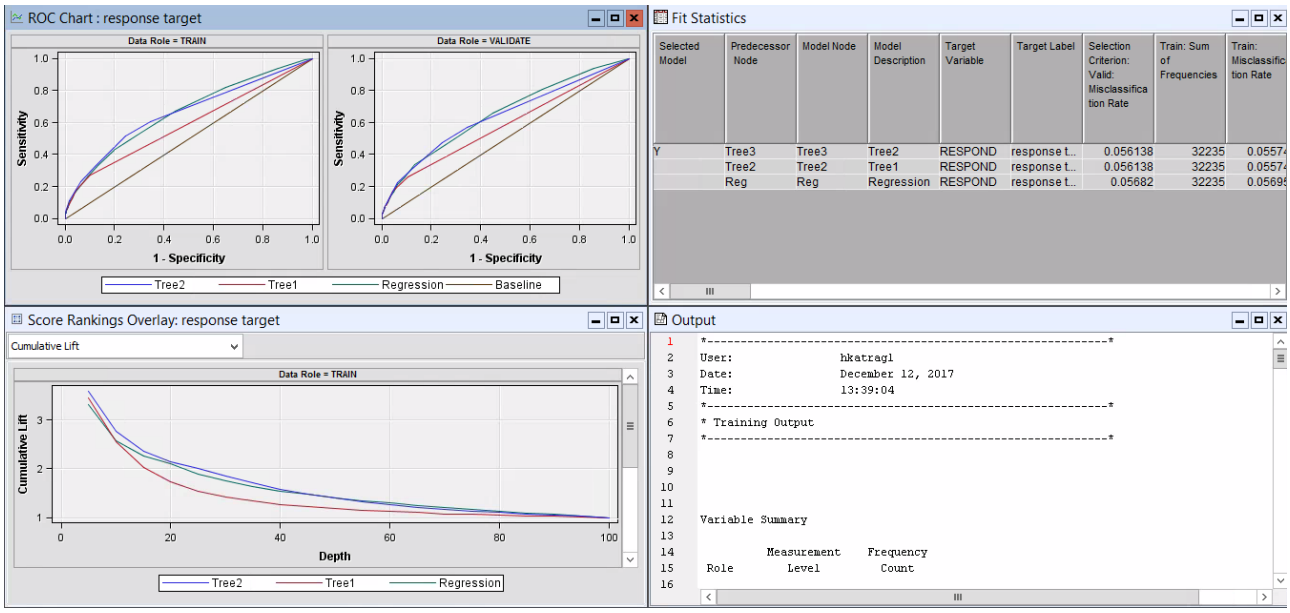
View the misclassification rate. From the Iteration Plot menu, select Misclassification Rate.

The iteration plot shows that the model with the smallest misclassification rate occurs in steps 6 and 7. If your analysis objective requires decision predictions, the predictions from the Step 6 model are as accurate as the predictions from the Step 7 model.

To compute an 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.



Run the Model Comparison node and view the results. a. Connect the Regression node to the Model Comparison node that you added earlier Right-click the Model Comparison node and click Run. View the results.



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

Data Role=Valid

Statistics	Tree3	Tree2	Reg
Valid: Kolmogorov-Smirnov Statistic	0.23	0.15	0.21
Valid: Average Squared Error	0.05	0.05	0.05
Valid: Roc Index	0.64	0.58	0.65
Valid: Average Error Function	.	.	0.21
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.08	0.07	0.07
Valid: Cumulative Percent Captured Response	26.20	23.76	25.60
Valid: Percent Captured Response	9.67	7.36	9.85
Valid: Divisor for VASE	32242.00	32242.00	32242.00
Valid: Error Function	.	.	6736.01
Valid: Gain	161.90	137.47	155.87
Valid: Gini Coefficient	0.29	0.16	0.30
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.22	0.15	0.21
Valid: Kolmogorov-Smirnov Probability Cutoff	0.06	0.05	0.05
Valid: Cumulative Lift	2.62	2.37	2.56
Valid: Lift	1.93	1.47	1.97
Valid: Maximum Absolute Error	0.97	0.95	1.00
Valid: Misclassification Rate	0.06	0.06	0.06
Valid: Mean Square Error	.	.	0.05
Valid: Sum of Frequencies	16121.00	16121.00	16121.00
Valid: Root Average Squared Error	0.23	0.23	0.23
Valid: Cumulative Percent Response	14.85	13.46	14.51
Valid: Percent Response	10.96	8.34	11.17
Valid: Root Mean Square Error	.	.	0.23
Valid: Sum of Squared Errors	1670.75	1681.35	1682.51
Valid: Sum of Case Weights Times Freq	.	.	32242.00

In general, if the type of prediction you want is a decision, then you want to minimize the misclassification rate and maximize the Kolmogorov-Smirov statistic. If the type of prediction is ranking, then you want to maximize the lift, the ROC index, and the Gini coefficient. If the type of prediction you want is estimates, then you want to minimize the average squared error (ASE). The three models are equal on ASE. The ROC and Gini favor the logistic regression model over the decision tree models. Lift is highest for the Tree model.