SAS Predictive Analysis

Part 1: RMF Analysis

Task One: Catalog 2010 Dataset

1. Correlation between variables

   To find out the correlations between predictor variables, first, we need to select appropriate predictor variables, whose data type are interval, instead of nominal or ordinal. Among the 98 variables in the dataset, variables that satisfy this criterion are MONLAST, TENURE, UNITSIDD, UNITSLAP, UNTLANPO, FREQPRCH, DOLINDET, DOLNETDT, DOLINDEA, DOLNETDA, DOLL24, DEPT01 to DEPT27, TOTORDQ01 to TOTORDQ22, etc.

   Second, we use the following SAS code to find out the correlation matrix of variables. We selected variable MONLAST, TENURE, UNITSIDD, UNITSLAP, UNTLANPO, FREQPRCH, DOLINDET, DOLNETDT, DOLINDEA, DOLNETDA, and DOLL24, to demonstrate the process.

```
1.  ods select Cov PearsonCorr;
2.  proc corr data='C:\Users\hpang\Documents\My SAS Files\Association Rule Mining\ABA1\SAS_DATA_BA\CATALOG20
    10' noprob outp=OutCorr /** store results **/
3.          nomiss /** listwise deletion of missing values **/
4.  var MONLAST TENURE UNITSIDD UNITSLAP
    UNTLANPO FREQPRCH DOLINDET DOLNETDT DOLINDEA DOLNETDA DOLL24;
5.  run;
6.
7.  proc print data=OutCorr; run; /** print output **/
```

   From the correlation matrix, we find that variable DOLNETDT is highly correlated with DOLINDET with a correlation coefficient of .99, and DOLNETDA is highly correlated with DOLINDEA with a correlation coefficient of .95 (see highlights as below). According to the description of the database, we learn that DOLINDET and DOLNETDT are gross total dollar demand and net total dollar demand respectively. Similarly, DOLINDEA and DOLNETDA are gross average dollar demand and net average dollar demand respectively. Therefore, it is not surprising that they have high correlation with each other.

Obs	_TYPE_	_NAME_	MONLAST	TENURE	UNITSIDD	UNITSLAP	UNTLANPO	FREQPRCH	DOLINDET	DOLNETDT	DOLINDEA	DOLNETDA	DOLL24
1	MEAN		38.68	83.26	10.92	22.05	2.67	4.17	196.67	187.86	47.75	45.30	45.69
2	STD		40.34	60.10	17.15	20.43	2.32	5.29	314.09	302.35	37.75	36.41	94.26
3	N		48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00
4	CORR	MONLAST	1.00	0.45	-0.24	0.29	-0.17	-0.21	-0.19	-0.19	-0.03	-0.01	-0.36
5	CORR	TENURE	0.45	1.00	0.28	0.14	-0.19	0.47	0.33	0.34	-0.07	-0.04	-0.05
6	CORR	UNITSIDD	-0.24	0.28	1.00	-0.13	0.34	0.80	0.88	0.88	0.21	0.21	0.54
7	CORR	UNITSLAP	0.29	0.14	-0.13	1.00	-0.23	-0.06	0.07	0.07	0.50	0.48	0.00
8	CORR	UNTLANPO	-0.17	-0.19	0.34	-0.23	1.00	-0.02	0.18	0.18	0.51	0.50	0.24
9	CORR	FREQPRCH	-0.21	0.47	0.80	-0.06	-0.02	1.00	0.82	0.81	-0.01	-0.00	0.40
10	CORR	DOLINDET	-0.19	0.33	0.88	0.07	0.18	0.82	1.00	0.99	0.33	0.32	0.58
11	CORR	DOLNETDT	-0.19	0.34	0.88	0.07	0.18	0.81	0.99	1.00	0.32	0.34	0.57
12	CORR	DOLINDEA	-0.03	-0.07	0.21	0.50	0.51	-0.01	0.33	0.32	1.00	0.95	0.35
13	CORR	DOLNETDA	-0.01	-0.04	0.21	0.48	0.50	-0.00	0.32	0.34	0.95	1.00	0.33
14	CORR	DOLL24	-0.36	-0.05	0.54	0.00	0.24	0.40	0.58	0.57	0.35	0.33	1.00

We consider net demand is a more important indicator of profitability, thus, we choose to remove DOLINDET and DOLINDEA from the analysis.
After dropping the two variables, the correlation matrix looks good as below.

Obs	_TYPE_	_NAME_	MONLAST	TENURE	UNITSIDD	UNITSLAP	UNTLANPO	FREQPRCH	DOLNETDT	DOLNETDA	DOLL24
1	MEAN		38.68	83.26	10.92	22.05	2.67	4.17	187.86	45.30	45.69
2	STD		40.34	60.10	17.15	20.43	2.32	5.29	302.35	36.41	94.26
3	N		48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00	48356.00
4	CORR	MONLAST	1.00	0.45	-0.24	0.29	-0.17	-0.21	-0.19	-0.01	-0.36
5	CORR	TENURE	0.45	1.00	0.28	0.14	-0.19	0.47	0.34	-0.04	-0.05
6	CORR	UNITSIDD	-0.24	0.28	1.00	-0.13	0.34	0.80	0.88	0.21	0.54
7	CORR	UNITSLAP	0.29	0.14	-0.13	1.00	-0.23	-0.06	0.07	0.48	0.00
8	CORR	UNTLANPO	-0.17	-0.19	0.34	-0.23	1.00	-0.02	0.18	0.50	0.24
9	CORR	FREQPRCH	-0.21	0.47	0.80	-0.06	-0.02	1.00	0.81	-0.00	0.40
10	CORR	DOLNETDT	-0.19	0.34	0.88	0.07	0.18	0.81	1.00	0.34	0.57
11	CORR	DOLNETDA	-0.01	-0.04	0.21	0.48	0.50	-0.00	0.34	1.00	0.33
12	CORR	DOLL24	-0.36	-0.05	0.54	0.00	0.24	0.40	0.57	0.33	1.00

In addition, correlation matrix can also be generated by using Variable Clustering node.

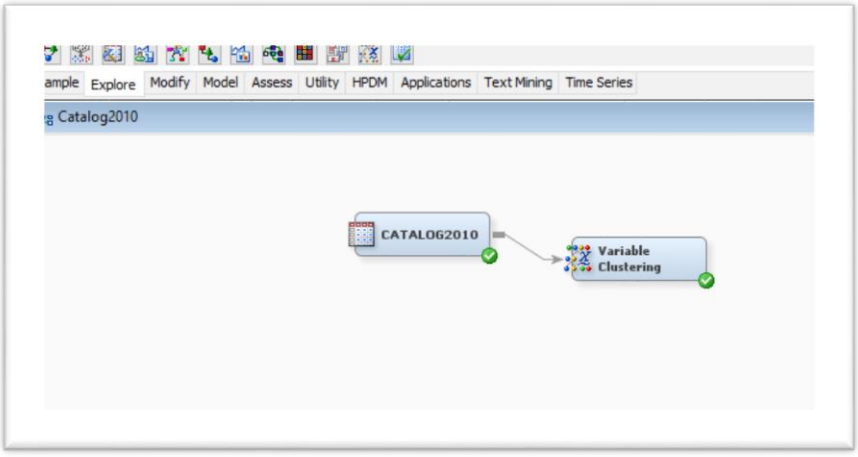
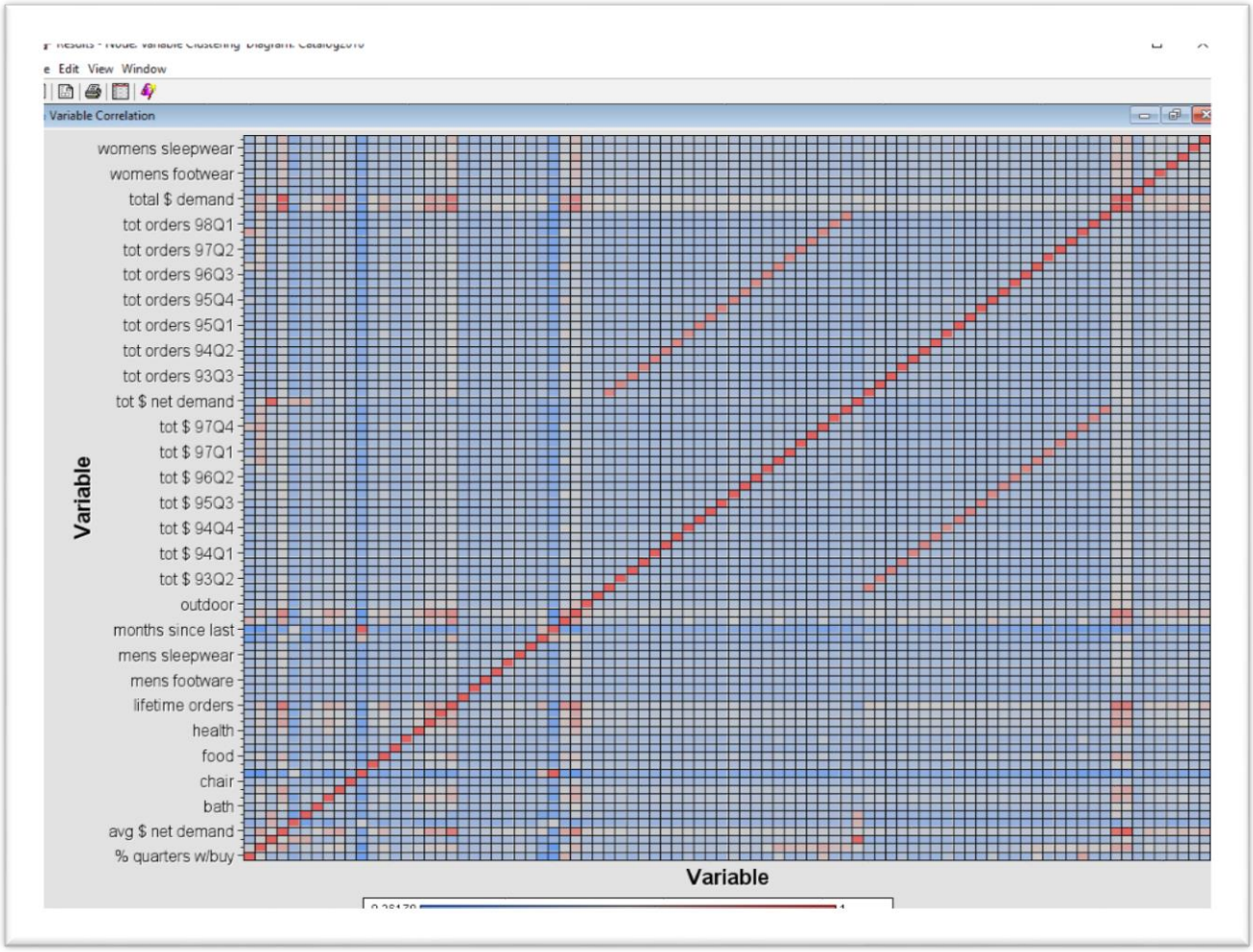
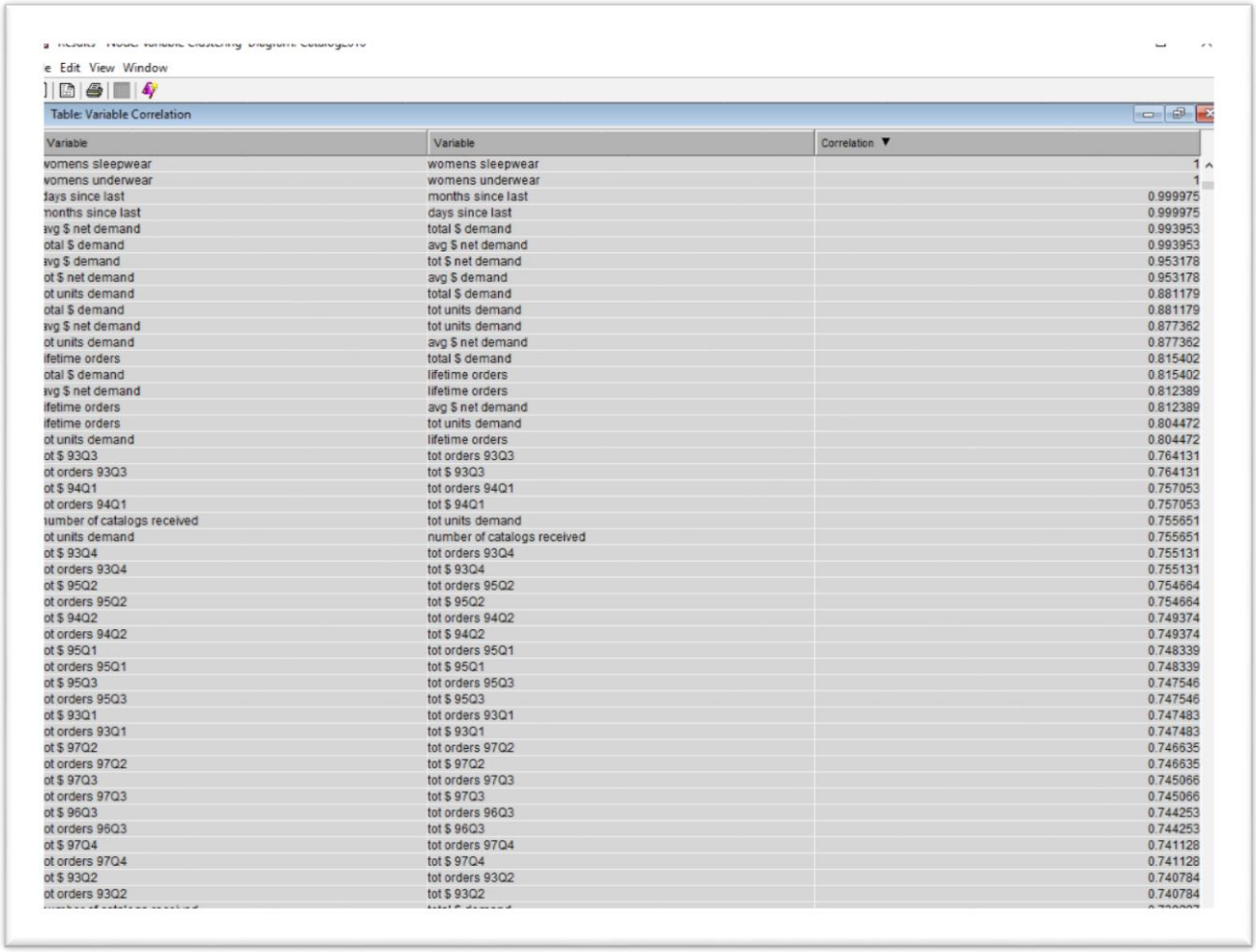


Diagram view of the correlation matrix. The red cells indicate highly correlated variables.



Because there are so many variables, it is easier to find out the most correlated variables from the table view of the correlation matrix as shown below.



The table lists all the correlations between variables. The correlation coefficient column can be sorted in descending order, and it becomes more efficient to look up variables that are highly correlated. The only downside of using this table is that it displays label names instead of variable names, and this cannot be changed by update View setups. Thus, it makes it very difficult to read and trace back to the dataset. However, after we spent some time in linking the labels back to the variable names, we were able to verify our previous results generated by SAS code are correct.

2. Outliers

There are multiple ways to find out outliers for each variable. One way is to generate a box plot for each variable. Another way is to use SAS code for extreme values. We prefer to use below code to find out extreme values of variables as it is more explicit. We use variables MONLAST, TENURE, UNITSIDD, UNITSLAP, UNTLANPO, FREQPRCH, DOLNETDT, DOLNETDA, and DOLL24 as an example.

```
1. ods select ExtremeObs;
2. proc univariate data='C:\Users\hpang\Documents\My SAS Files\Association Rule Mining\ABA1\SAS_DATA_BA\CATALOG2010' outtable=OutExtreme;
3. var MONLAST TENURE UNITSIDD UNITSLAP UNTLANPO FREQPRCH DOLNETDT DOLNETDA DOLL24;
4. run;
5. proc print data=OutExtreme; run;
```

Below are the outputs of the code. Extreme Observations shows the extreme values of a variable. For example, for variable MONLAST, the extreme values are as below. Zero values could represent missing or unknown values in the dataset as mentioned in the description of the dataset.

Program Editor - Output

```
1
2  1
3
4  The UNIVARIATE Procedure
5  Variable:  MONLAST  (months since last)
6
7          Extreme Observations
8
9  ----Lowest----        ----Highest---
10
11 Value      Obs        Value      Obs
12
13     0    48179          258    15743
14     0    48089          258    42042
15     0    48081          258    46543
16     0    48042          259    11772
17     0    47956          271    17431
18
19 2
20
21 The UNIVARIATE Procedure
22 Variable:  TENURE  (months since 1st)
23
24         Extreme Observations
25
26 ----Lowest----        ----Highest---
27
28 Value      Obs        Value      Obs
29
30     0    47709          276    40953
31     0    44638          276    41335
32     0    41896          276    45977
33     0    39192          276    47279
34     0    39012          276    48118
35
36 3
```

```
The UNIVARIATE Procedure
Variable:  UNITSIDD  (tot units demand)

        Extreme Observations

----Lowest----        ----Highest---

Value      Obs        Value      Obs

    1    48352          325     8871
    1    48350          326    19052
    1    48314          423     4484
    1    48310          429     8065
    1    48309          432    31650

4

The UNIVARIATE Procedure
Variable:  UNITSLAP  (avg price/unit)

          Extreme Observations

------Lowest-----        ----Highest----

   Value      Obs         Value      Obs

 1.00000    26907         500.0    22934
 1.00000    22944         502.0     7255
 1.00400    18980         502.0    32305
 1.35000    18813         618.0    26056
 1.38462    13198         768.5    43274

5
```

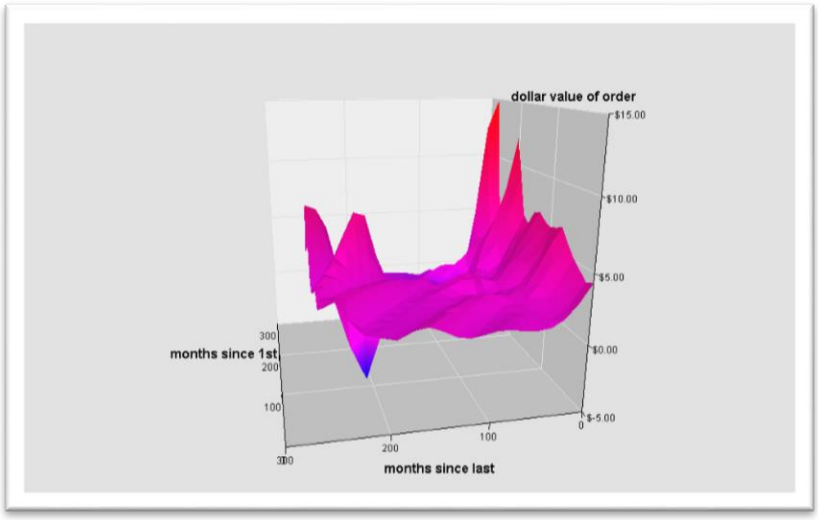
The output also includes statistics of all the variables as shown below. We can calculate outliers for each variable using formula 1.5*(Q3-Q1). For example, for variable MONLAST, the cutoff value for outliers is 1.5*(52-9)= 64.5. Values that are greater than 64.5 are consider as outliers. Along with the high number of extreme values we have from above output, we can tell that there are many low value customers who haven't made any purchase in 5.3 years in the database.

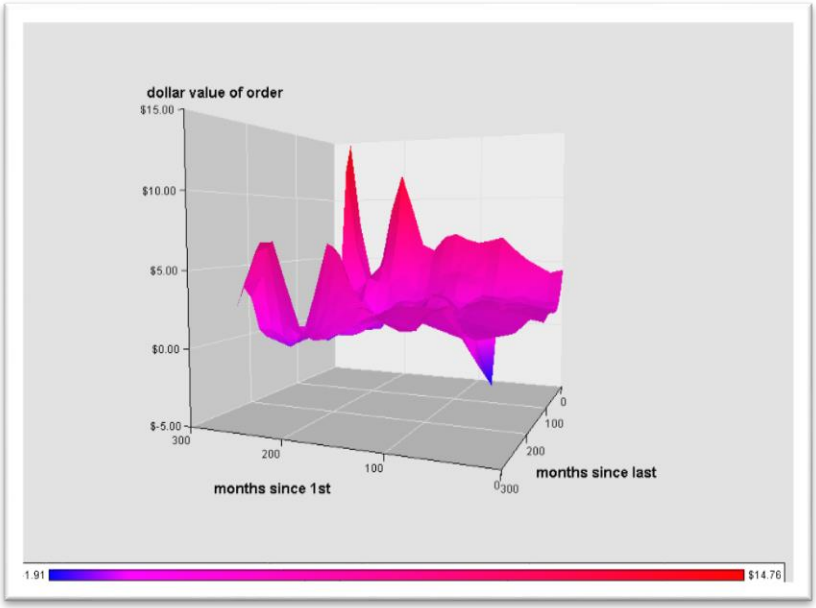
Obs	_VAR_	_LABEL_	_NOBS_	_NMISS_	_SUMWGT_	_SUM_	_MEAN_	_STD_	_VARI_	_SKEW_	_KURT_	_MIN_	_P1_	_P5_	_P10_	_Q1_	_MEDIAN_	_Q3_	_P90_	_P95_	_P99_	_MAX
1	MONLAST	months since last	48356	0	48356	1870330.00	38.678	40.345	1627.69	1.77074	3.310	0	0	2.0	4.00	9.0000	25.000	52.000	97.000	128.000	175.00	271.
2	TENURE	months since 1st	48356	0	48356	4025975.00	83.257	60.105	3612.60	0.61870	-0.428	0	3	8.0	13.00	31.0000	70.000	128.000	164.000	188.000	240.00	276.
3	UNITSIDD	tot units demand	48356	0	48356	528258.00	10.924	17.155	294.29	6.05298	69.555	1	1	1.0	1.00	2.0000	5.000	13.000	25.000	38.000	80.00	432.
4	UNITSLAP	avg price/unit	48356	0	48356	1066248.88	22.050	20.435	417.58	6.98677	113.763	1	5	7.5	8.95	11.9500	16.875	25.002	38.438	52.063	102.00	768.
5	UNTLANPO	avg units/order	48356	0	48356	129051.66	2.669	2.317	5.37	7.88094	217.401	1	1	1.0	1.00	1.2457	2.000	3.114	5.000	6.333	11.00	121.
6	FREQPRCH	lifetime orders	48356	0	48356	201499.00	4.167	5.291	27.99	4.59092	42.577	1	1	1.0	1.00	1.0000	2.000	5.000	9.000	14.000	26.00	150.
7	DOLNETDT	avg $ net demand	48356	0	48356	9084117.87	187.859	302.354	91417.72	6.14970	69.606	0	0	14.0	20.85	41.0000	92.950	212.725	431.990	656.500	1435.05	8028.
8	DOLNETDA	tot $ net demand	48356	0	48356	2190579.81	45.301	36.409	1325.64	3.65926	30.395	0	0	11.7	15.90	23.5113	35.900	55.723	84.307	109.717	183.65	768.
9	DOLL24	$ last 24 months	48356	0	48356	2209432.12	45.691	94.260	8885.03	5.89095	67.841	0	0	0.0	0.00	0.0000	0.000	55.650	130.850	197.650	428.60	2433.

3. 3-D graph

Below is a 3-D graph generated by SAS. We used ORDERSIZE (dollar value of order) as the dependent variable, as the other target variable RESPONSE is binary. MONLAST (months since last) and TENURE (months since 1st) are selected as two independent variables.

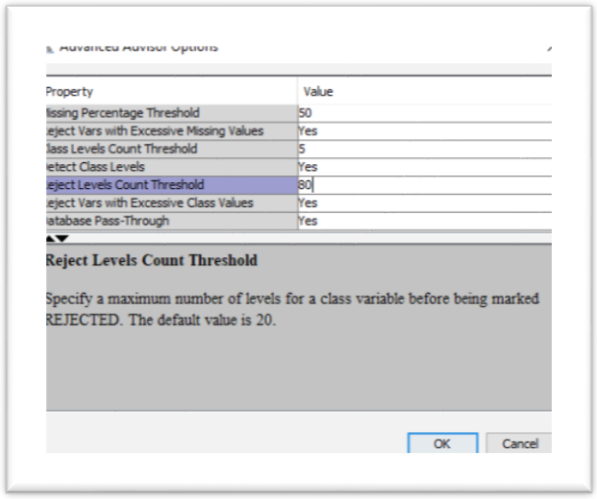
The graph indicates that older customers who also made recent purchase tend to spend more on the orders (the red peaks). On the other hand, customers who are relatively new to the catalog and didn't make any recent purchase tend not to spend on the orders (the blue peak).



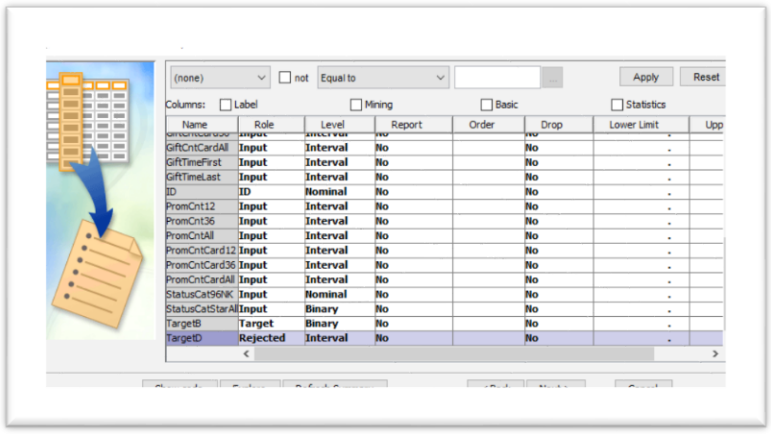


Task two: RFM Analysis on PVA97NK

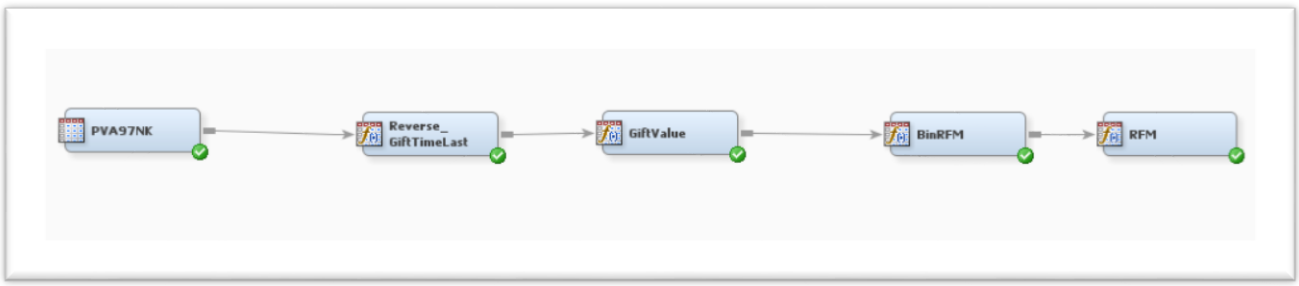
1. Customizing the dataset
   a) Change thresholds for the dataset



Reject variable target D

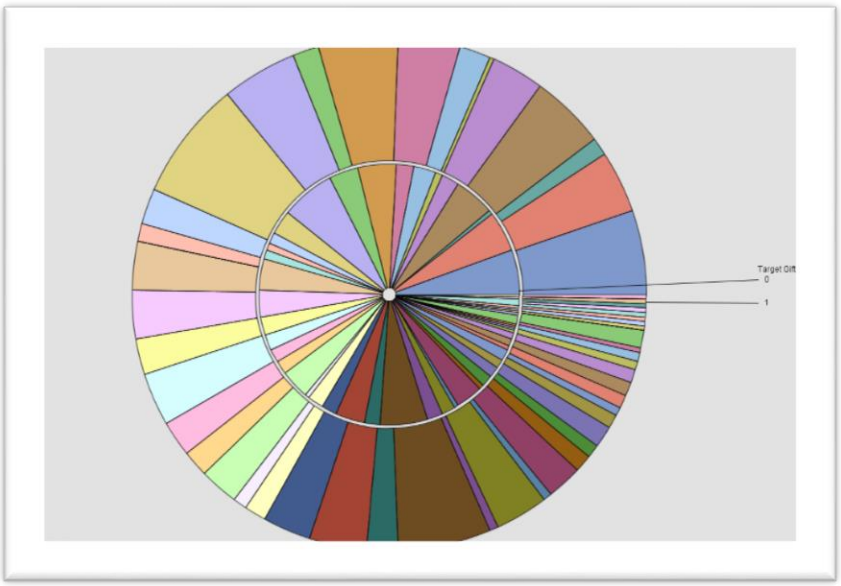


   b) Perform transformations on the dataset to arrive at the RFM table

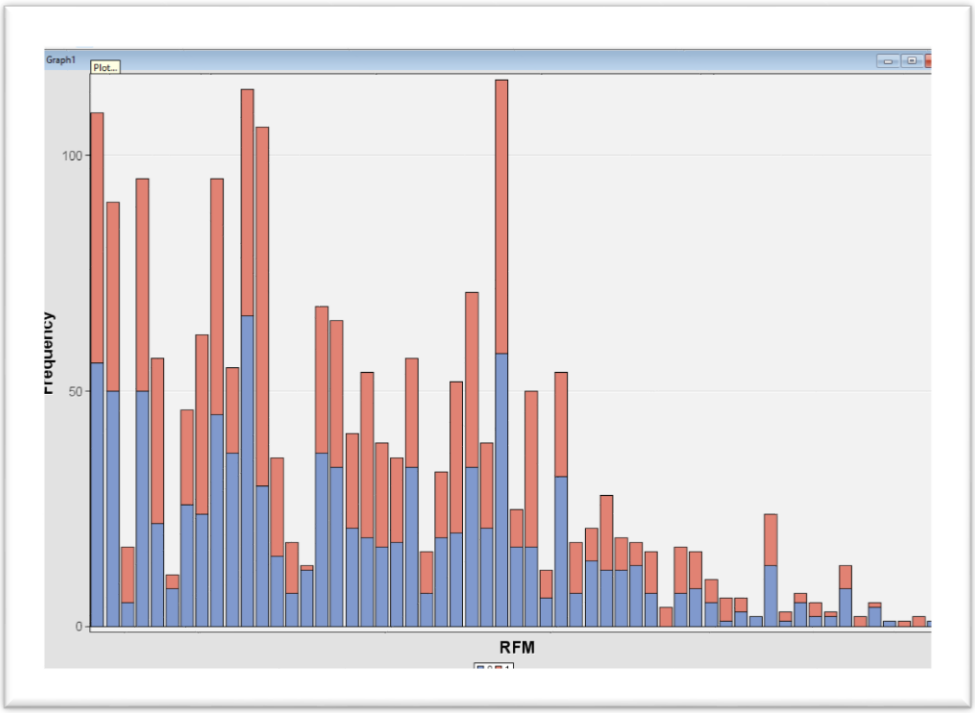


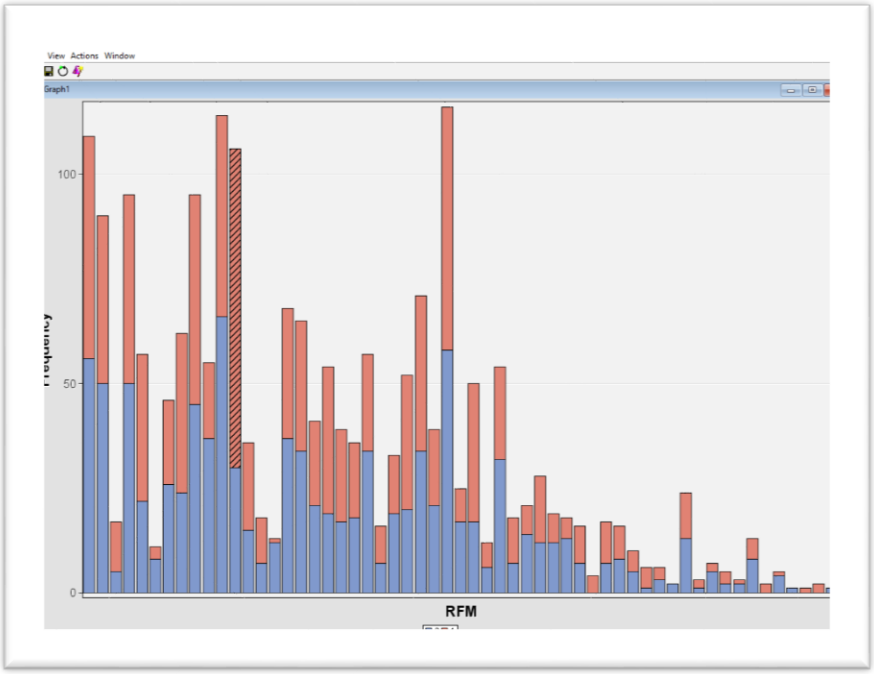
c) Pie chart and stacked bar chart generated form the RFM table

Below pie chart shows the detailed slices of each group. The inner ring of the pie shows response and the outer ring shows non-response.

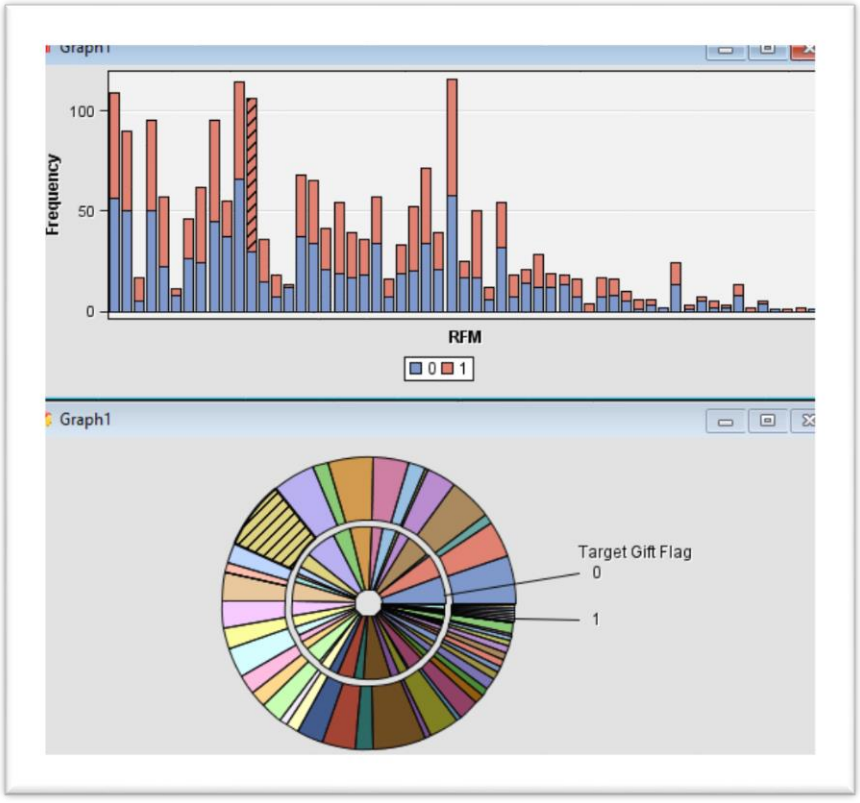


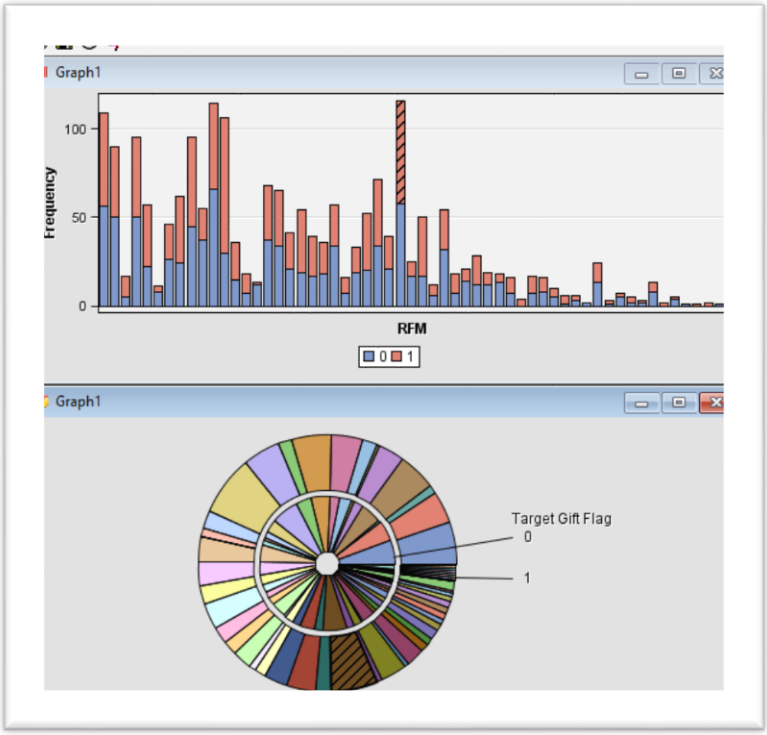
Below stacked bar chart shows proportion of response (red) and non-response (blue).





The highlighted bar in the above picture is for group 040404, which has the highest response rate (76). Its corresponding pie chart is in the below picture. 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.





The highlighted bar chart above shows group 030404 with the second highest response rate (58). It corresponds to the brown slice of the outer ring of the pie chart.

d) Calculation of the response rates for 040404 and 030303 group:

Group 040404 = 76/(76+30) = 71.7%
Group 030303 = 50/(50+45) = 52.63%

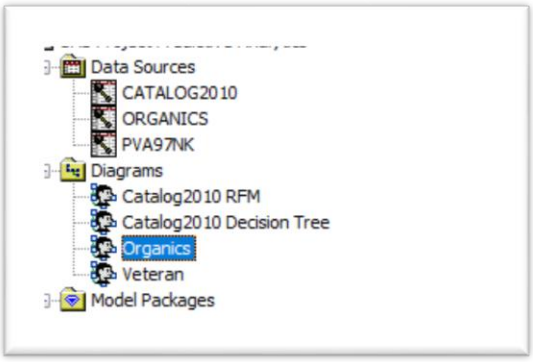
e) Break-even response rate for this promotion: 2.3/21=10.95%

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 threshhold.

Part 2: Decision Tree

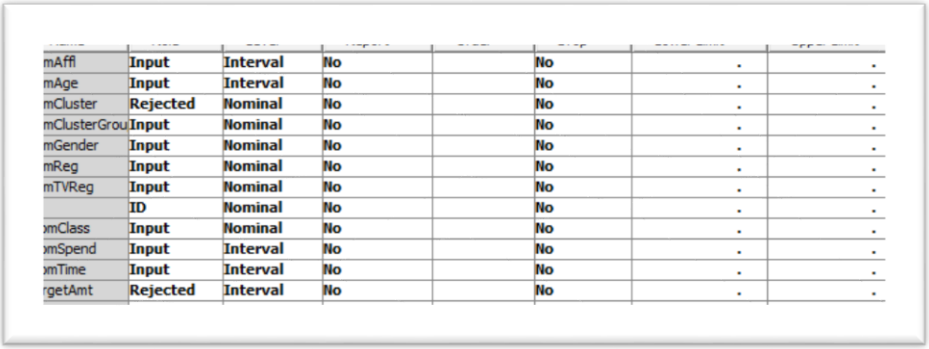
Task Three: Decision Tree Analysis on Organics

1. Create a new diagram named Organics



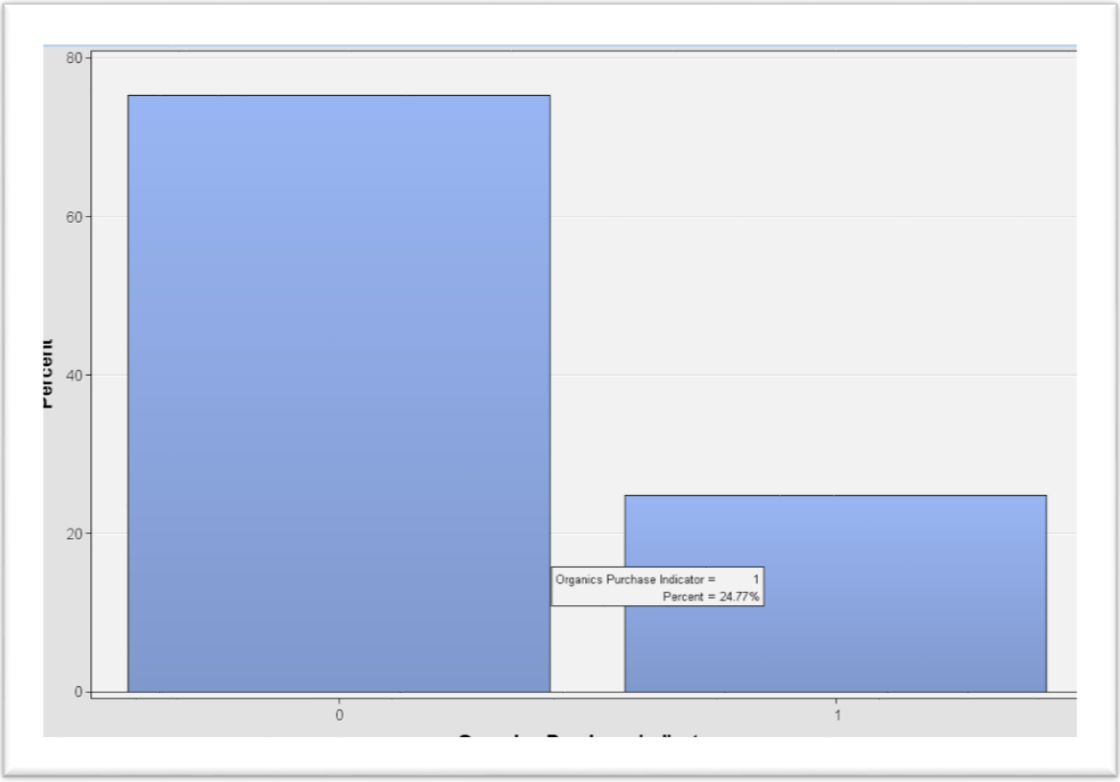
2. Define the data set Organics as a data source for the project

   a. Set the roles for the analysis variables



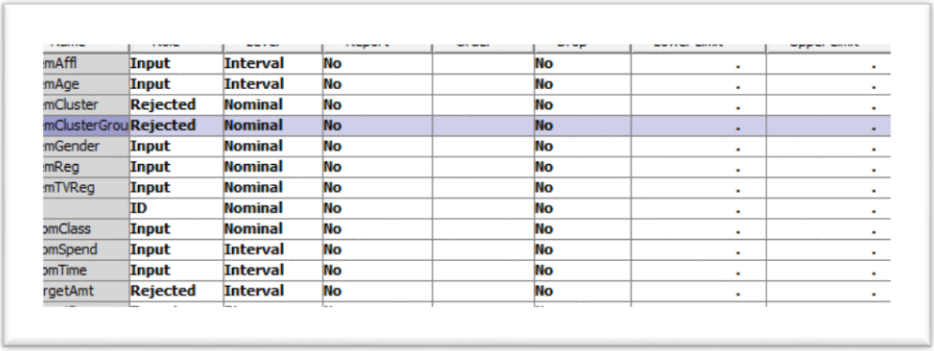
mAffl	Input	Interval	No		No	.	.
mAge	Input	Interval	No		No	.	.
mCluster	Rejected	Nominal	No		No	.	.
mClusterGrou	Input	Nominal	No		No	.	.
mGender	Input	Nominal	No		No	.	.
mReg	Input	Nominal	No		No	.	.
mTVReg	Input	Nominal	No		No	.	.
	ID	Nominal	No		No	.	.
mClass	Input	Nominal	No		No	.	.
mSpend	Input	Interval	No		No	.	.
mTime	Input	Interval	No		No	.	.
rgetAmt	Rejected	Interval	No		No	.	.

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



From the bar chart above, we can see the percentage of individuals who purchased organic products is 24.77%.

c. Set DemClusterGroup tp Rejected



mAffl	Input	Interval	No		No	.	.
mAge	Input	Interval	No		No	.	.
mCluster	Rejected	Nominal	No		No	.	.
mClusterGrou	Rejected	Nominal	No		No	.	.
mGender	Input	Nominal	No		No	.	.
mReg	Input	Nominal	No		No	.	.
mTVReg	Input	Nominal	No		No	.	.
	ID	Nominal	No		No	.	.
mClass	Input	Nominal	No		No	.	.
mSpend	Input	Interval	No		No	.	.
mTime	Input	Interval	No		No	.	.
rgetAmt	Rejected	Interval	No		No	.	.

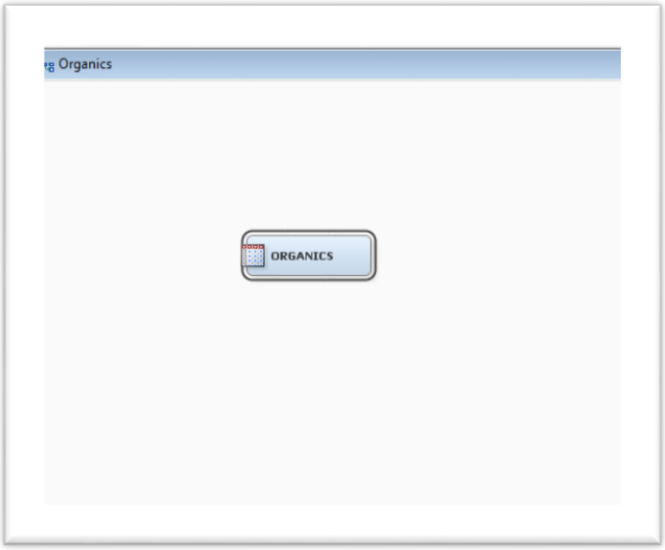
d. Can **TargetAmt** be used as an input for a model that is used to predict **TargetBuy**? Why or why not?

No. **TargetAmt** is highly correlated with **TargetBuy**. For any none zero value in **TargetAmt,** it indicates **TargetBuy** = 1. Therefore, **TargetAmt** should not be used as an input for the model.

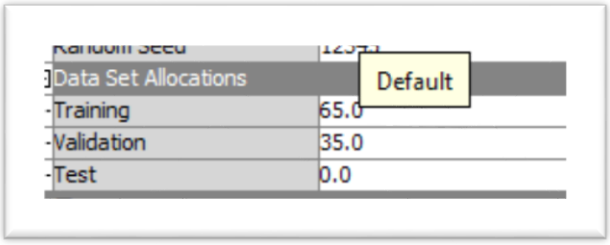
e. Finish the **ORGANICS** data source definition

Clicked **OK.**

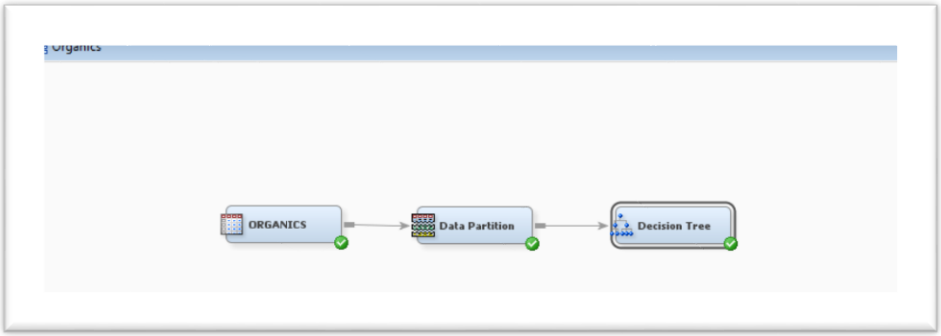
3. Add the **ORGANICS** data source to the Organics diagram workspace



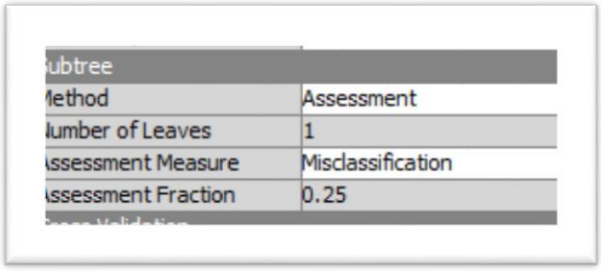
4. Add a Data Partition node to the diagram and connect it to the Data Source node. Assign 65% of data for training and 35% for validation.



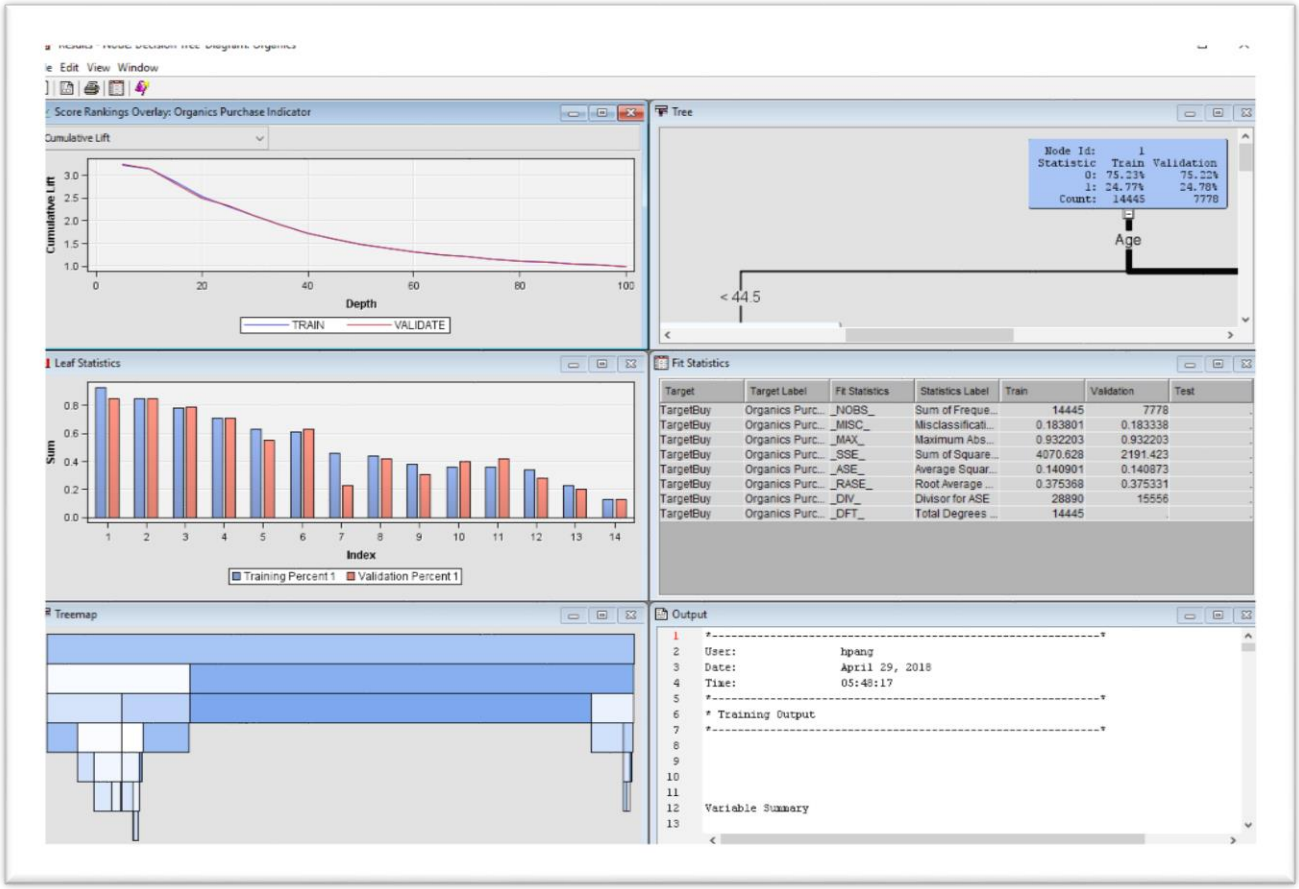
5. Add a Decision Tree node to the workspace and connect it to the Data Partition node.

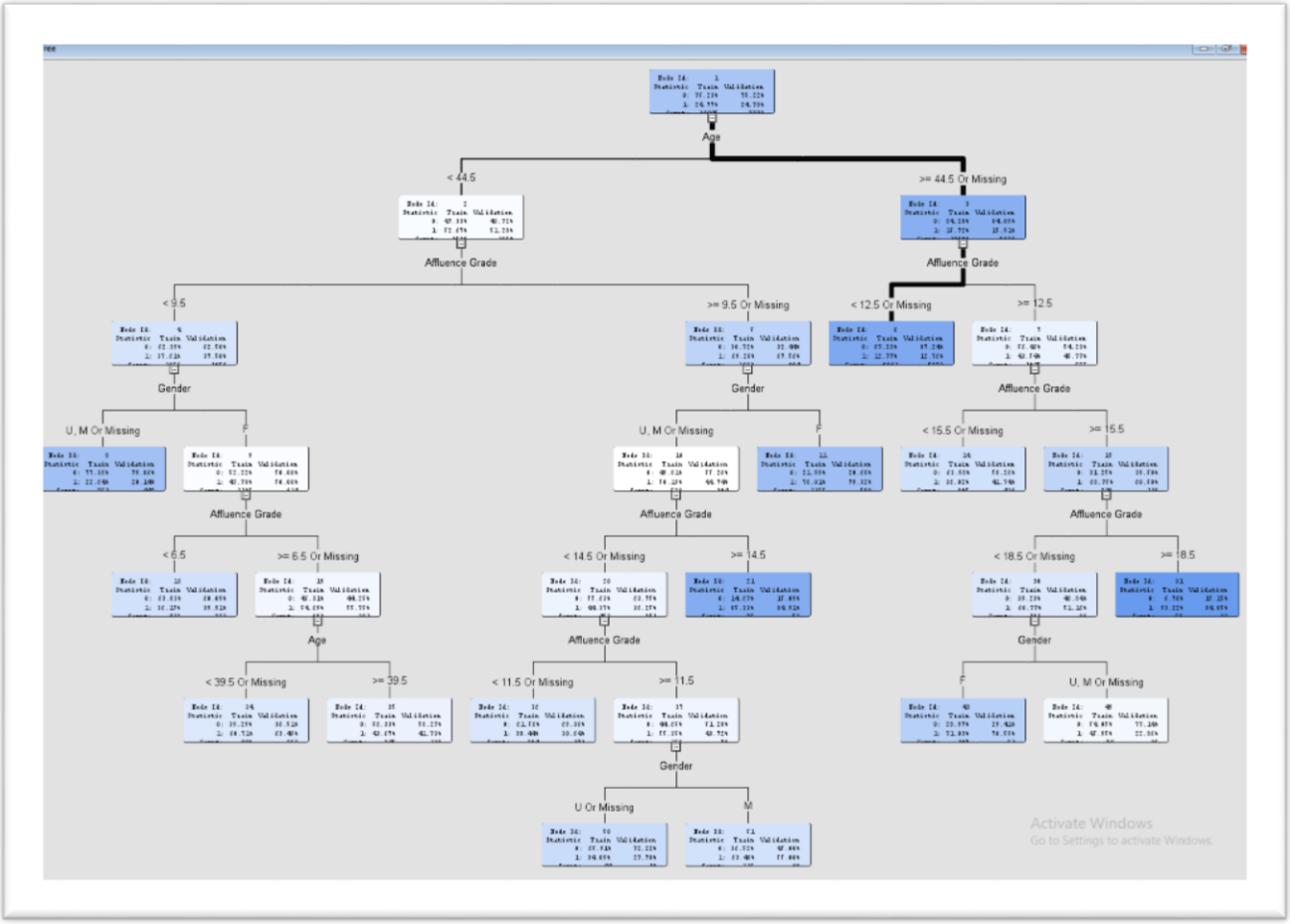


6. Create a decision tree model autonomously. Use **Misclassification** as the model assessment statistic.



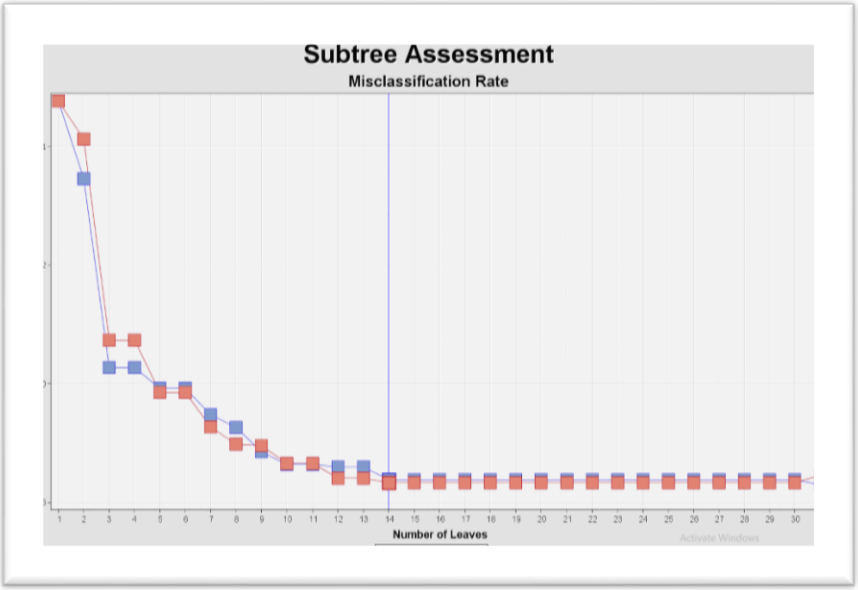
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25




Age
< 44.5
>= 44.5 Or Missing
Affluence Grade
Affluence Grade
< 9.5
>= 9.5 Or Missing
< 12.5 Or Missing
>= 12.5
Gender
Gender
Affluence Grade
U, M Or Missing
F
U, M Or Missing
F
< 15.5 Or Missing
>= 15.5
Affluence Grade
Affluence Grade
Affluence Grade
< 6.5
>= 6.5 Or Missing
< 14.5 Or Missing
>= 14.5
< 18.5 Or Missing
>= 18.5
Age
Affluence Grade
Gender
< 39.5 Or Missing
>= 39.5
< 11.5 Or Missing
>= 11.5
F
U, M Or Missing
Gender
U Or Missing
M
Activate Windows
Go to Settings to activate Windows.

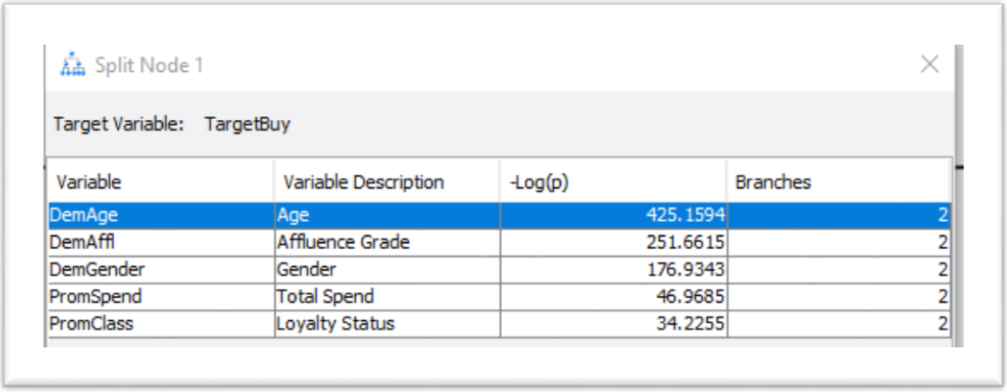
a. How many leaves are in the optimal tree?

The optimal tree has 14 leaves as shown by the Subtree Assessment Plot below, as the misclassification rate does not change significantly for either training or validation model when the three grows into more than 14 leaves.



b. Which variables were used for the first split? What were the competing splits for this first split?

DemAge (Age) were the first split. From the interactive Decision Tree Diagram below, we learned that **DemAffl** (Affuence Grade), **DemGender** (Gender) have the second and third highest -Log(p) values (information). Therefore, **DemAffl, DemGender** were the competing splits for the first split.

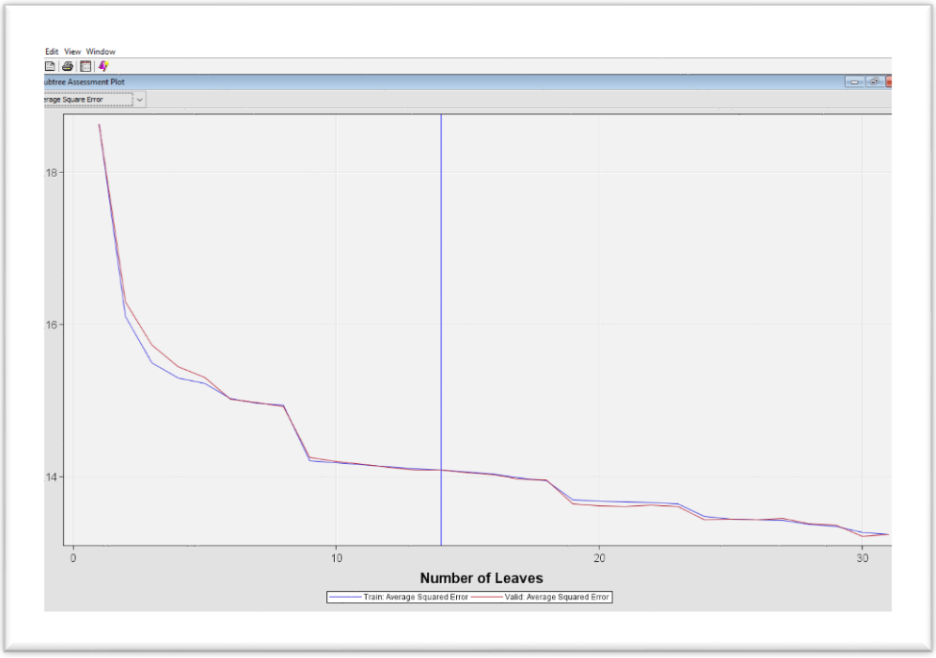


Variable	Variable Description	-Log(p)	Branches
DemAge	Age	425.1594	2
DemAffl	Affluence Grade	251.6615	2
DemGender	Gender	176.9343	2
PromSpend	Total Spend	46.9685	2
PromClass	Loyalty Status	34.2255	2

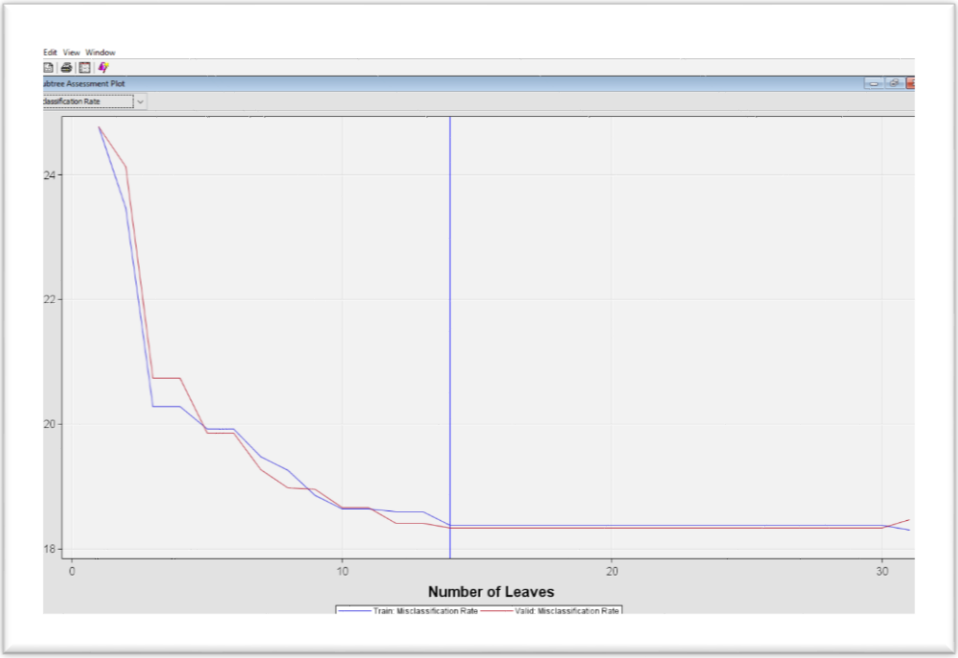
c. Which variables were used for the second slit for all branches from the first split?

DemAffl (Affluence Grade)

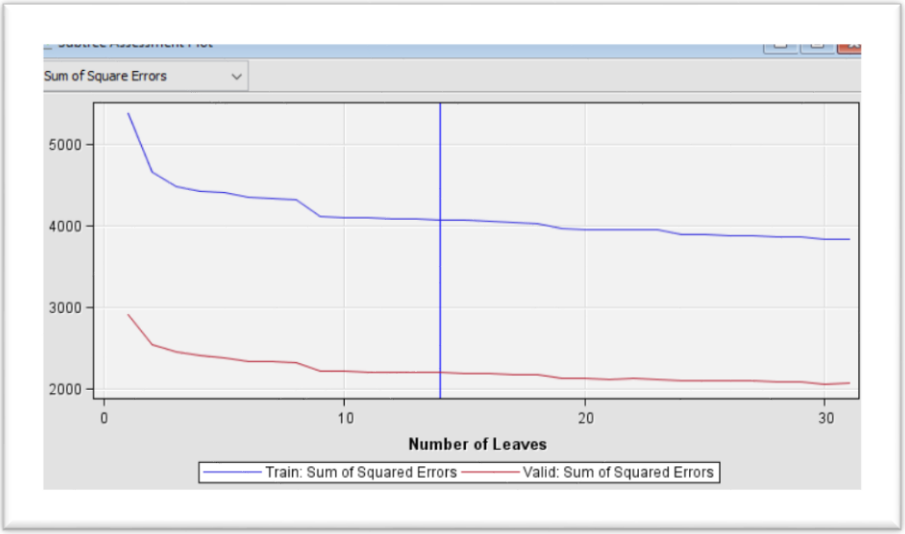
d. Discuss the results and provide your insights



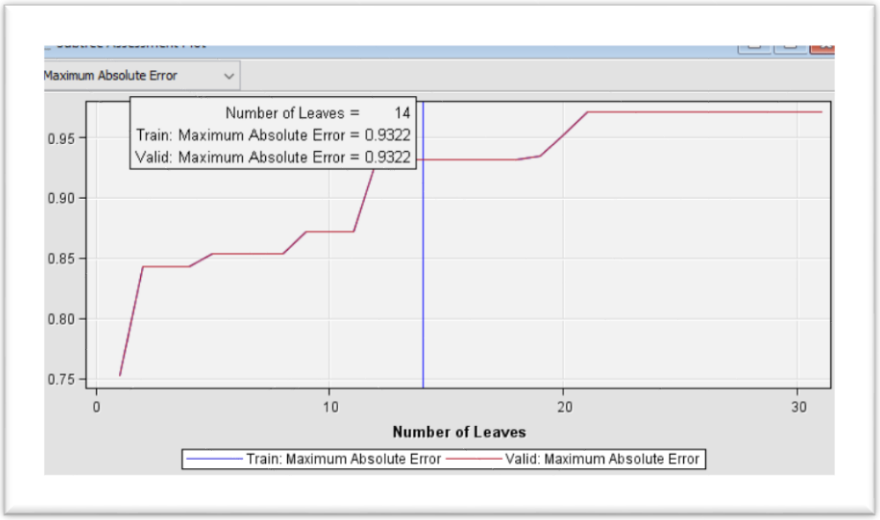
The Subtree Assessment Plot of Average Square Error shows that the model fits both training and validation datasets well, as the average square error (the deviation between prediction and actual outcome is squared and averaged across each outcome category) diminishes as the complexity of the tree increases. The optimal tree appears to have approximately 14 leaves. For validation performance, over the range of 30 to 31 leaves, the precision of the model diminishes slightly.



The Misclassification Rate Plot confirms the observation under Average Square Error that the optimal tree appears to have approximately 14 leaves. The misclassification rate diminishes as the tree grows more complex for both training and validation datasets. However, after fourteen leaves, the decrease of the misclassification rate is not noticeable. Performance of validation dataset over the range of proximate 30 leaves becomes worse as the tree become more complex.

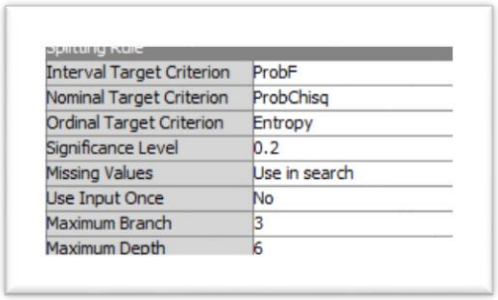


The Sum of Square Errors shows the same trend as Misclassification Rate. Both validation and training dataset do better as the tree grows more complex. The optimal tree appears to have 14 leaves as the rest of the decrease in sum of square errors is not significant.



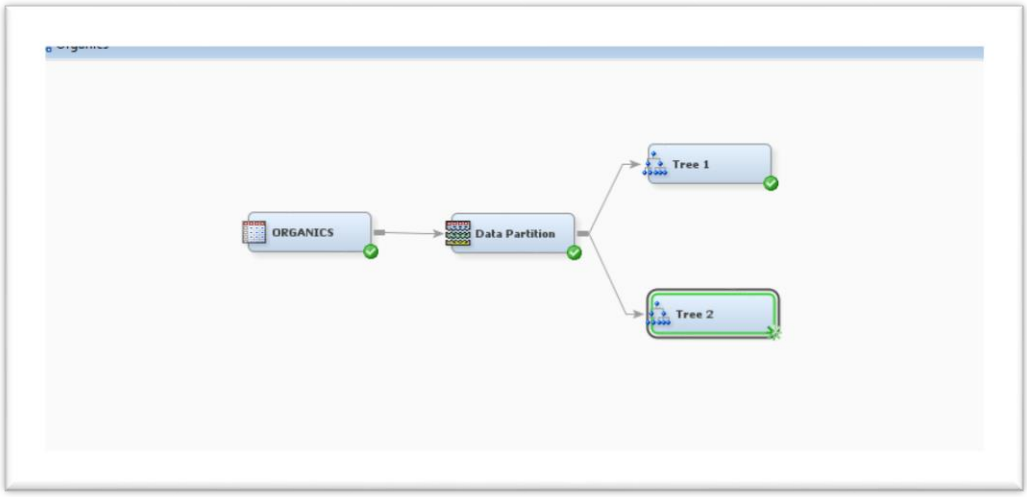
The Maximum Absolute Error shows that the maximum error in a tree with 14 leaves is .9322 for both training and validation.

7. Add a second Decision Tree node to the diagram and connect it to the Data Partition node.
   a. Change the maximum number of branches from a node to 3 to allow for three-way splits

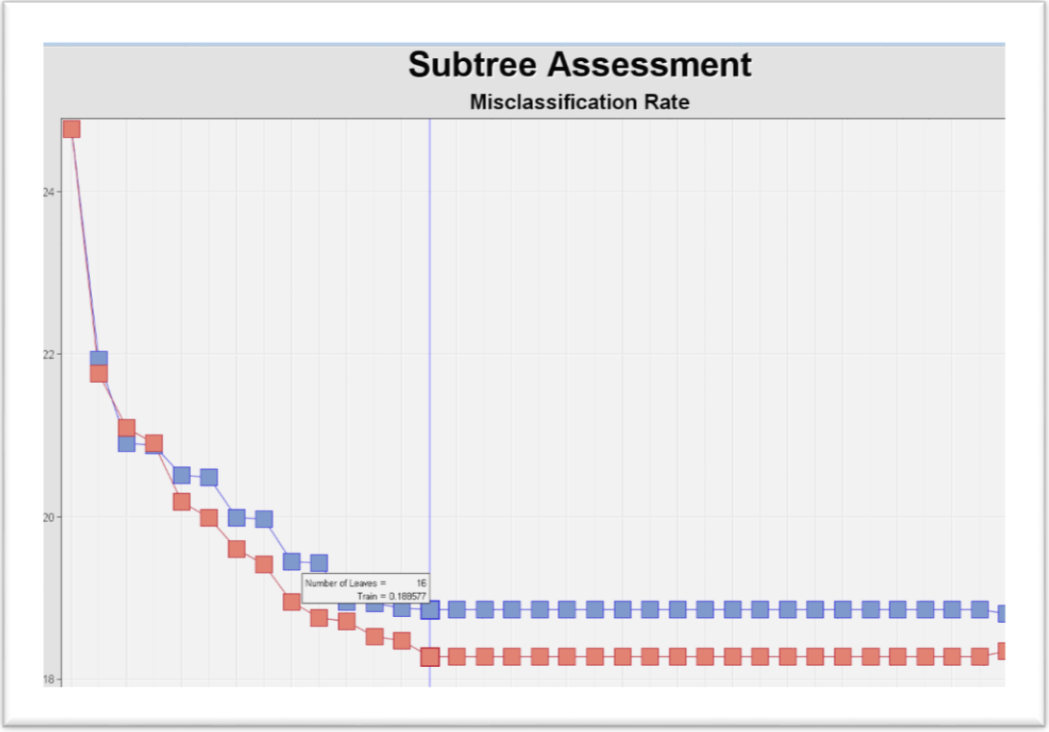


Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6

   b. Create a decision tree using Misclassification as the model assessment statistic

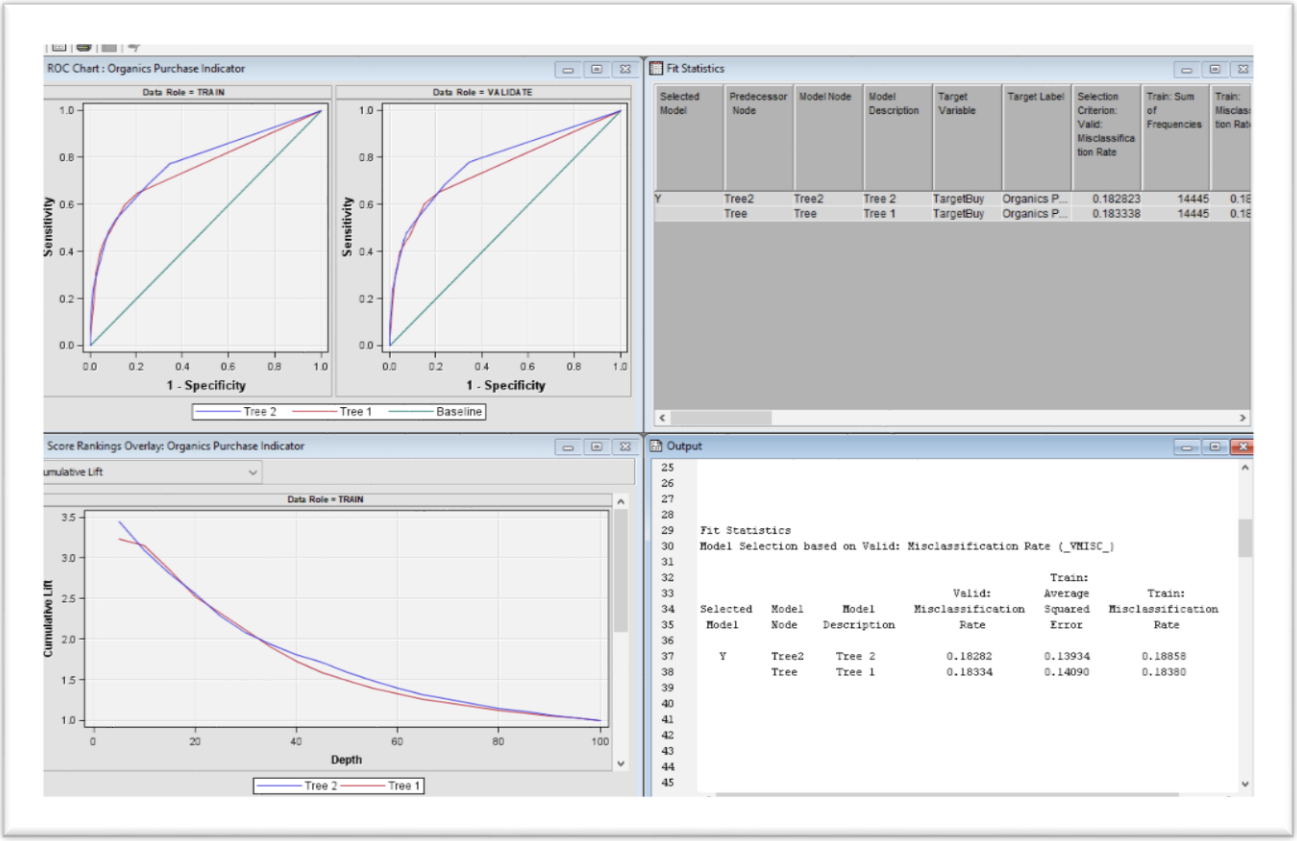


c. How many leaves are in the optimal tree?



There are 16 leaves in the optimal tree as shown in the above Subtree Assessment plot, as there is not any major decrease in misclassification rate after 16 leaves for validation datas.

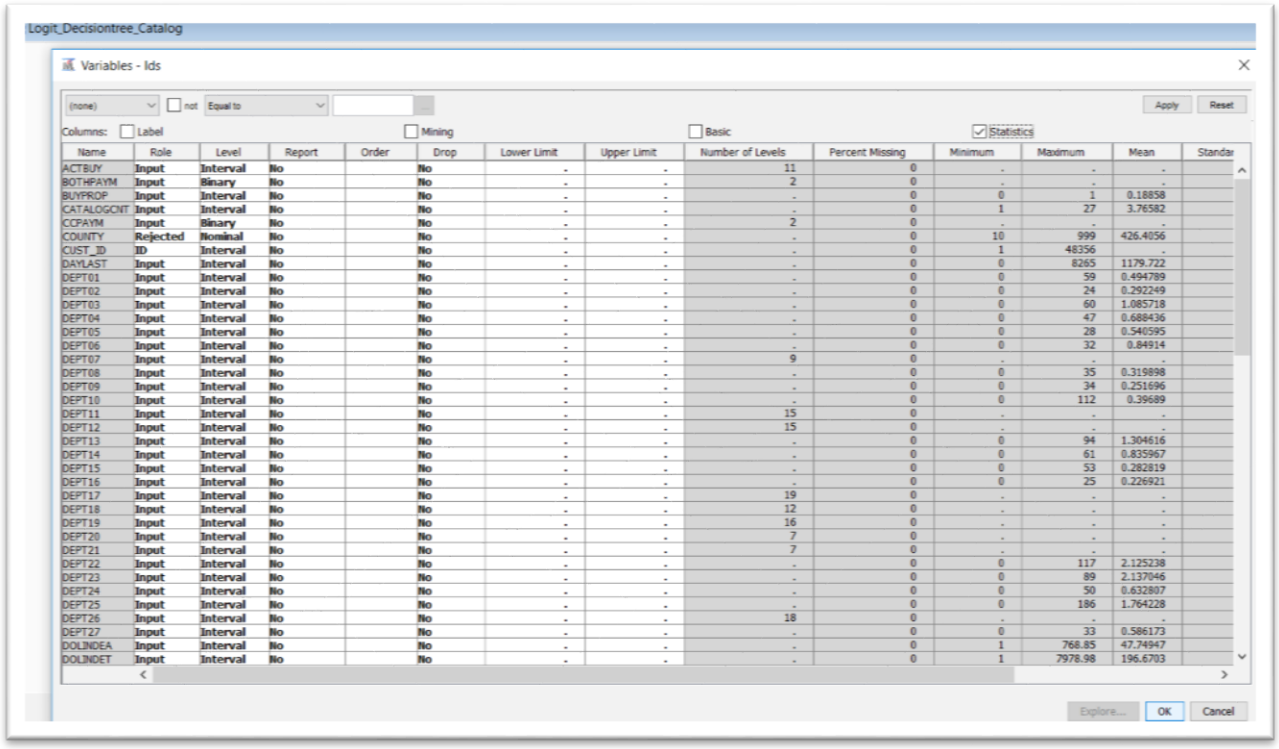
8. Based on Misclassification rate, which of the decision tree models appears to be better?



Using Model Comparison node to compare these two trees, we found that Tree 2 had a higher lift and a higher ROC statistic along with a lower average misclassification rate. Therefore, Tree 2 is better.

Part 3: Logit Regression Catalog2010

1. Set data source as CATALOG2010, and selected **Statistics** in the upper right corner.

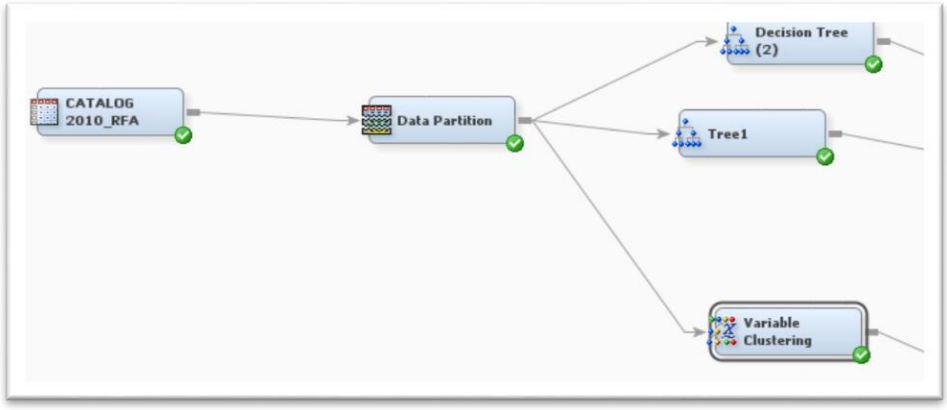


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standar
ACTBUY	Input	Interval	No		No	.	.	11	0	.	.	.	
BOTHPAYM	Input	Binary	No		No	.	.	2	0	.	.	.	
BUYPROP	Input	Interval	No		No	.	.	.	0	0	1	0.18858	
CATALOGCNT	Input	Interval	No		No	.	.	.	0	1	27	3.76582	
CCPAYM	Input	Binary	No		No	.	.	2	0	.	.	.	
COUNTY	Rejected	Nominal	No		No	.	.	.	0	10	999	426.4056	
CUST_ID	ID	Interval	No		No	.	.	.	0	1	48356	.	
DAYLAST	Input	Interval	No		No	.	.	.	0	0	8265	1179.722	
DEPT01	Input	Interval	No		No	.	.	.	0	0	59	0.494789	
DEPT02	Input	Interval	No		No	.	.	.	0	0	24	0.292249	
DEPT03	Input	Interval	No		No	.	.	.	0	0	60	1.085718	
DEPT04	Input	Interval	No		No	.	.	.	0	0	47	0.688436	
DEPT05	Input	Interval	No		No	.	.	.	0	0	28	0.540595	
DEPT06	Input	Interval	No		No	.	.	.	0	0	32	0.84914	
DEPT07	Input	Interval	No		No	.	.	9	0	.	.	.	
DEPT08	Input	Interval	No		No	.	.	.	0	0	35	0.319898	
DEPT09	Input	Interval	No		No	.	.	.	0	0	34	0.251696	
DEPT10	Input	Interval	No		No	.	.	.	0	0	112	0.39689	
DEPT11	Input	Interval	No		No	.	.	15	0	.	.	.	
DEPT12	Input	Interval	No		No	.	.	15	0	.	.	.	
DEPT13	Input	Interval	No		No	.	.	.	0	0	94	1.304616	
DEPT14	Input	Interval	No		No	.	.	.	0	0	61	0.835967	
DEPT15	Input	Interval	No		No	.	.	.	0	0	53	0.282819	
DEPT16	Input	Interval	No		No	.	.	.	0	0	25	0.226921	
DEPT17	Input	Interval	No		No	.	.	19	0	.	.	.	
DEPT18	Input	Interval	No		No	.	.	12	0	.	.	.	
DEPT19	Input	Interval	No		No	.	.	16	0	.	.	.	
DEPT20	Input	Interval	No		No	.	.	7	0	.	.	.	
DEPT21	Input	Interval	No		No	.	.	7	0	.	.	.	
DEPT22	Input	Interval	No		No	.	.	.	0	0	117	2.125238	
DEPT23	Input	Interval	No		No	.	.	.	0	0	89	2.137046	
DEPT24	Input	Interval	No		No	.	.	.	0	0	50	0.632807	
DEPT25	Input	Interval	No		No	.	.	.	0	0	186	1.764228	
DEPT26	Input	Interval	No		No	.	.	18	0	.	.	.	
DEPT27	Input	Interval	No		No	.	.	.	0	0	33	0.586173	
DOLINDEA	Input	Interval	No		No	.	.	.	0	1	768.85	47.74947	
DOLINDET	Input	Interval	No		No	.	.	.	0	1	7978.98	196.6703	

2. **State** variable Role set to Rejected.

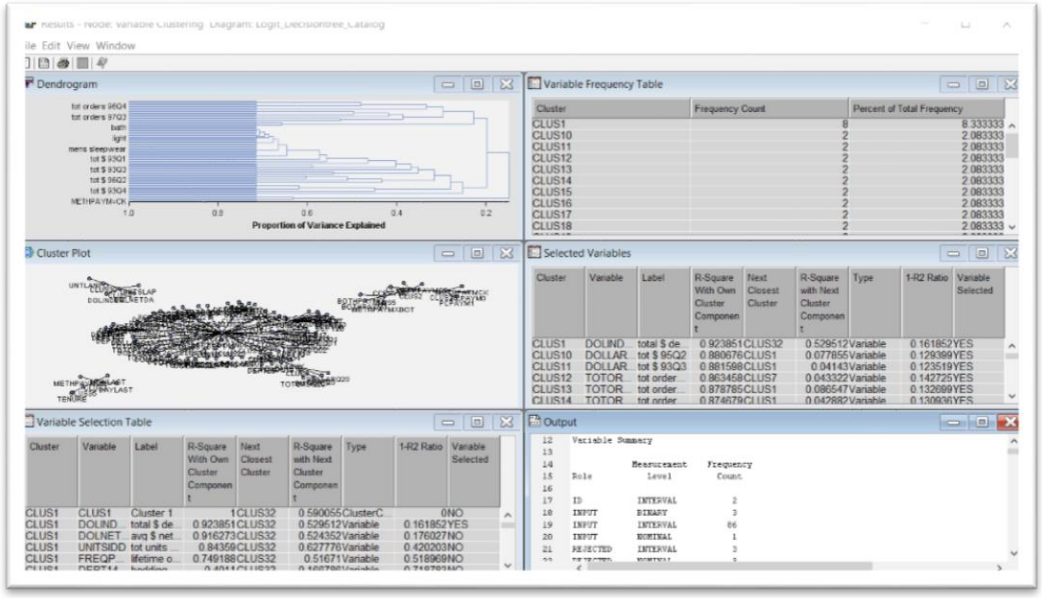
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing
OLNETDT	Input	Interval	No		No	.	.	.	0
TBUYLST	Rejected	Interval	No		No	.	.	.	0
TBUYORG	Rejected	Interval	No		No	.	.	.	0
REQPRCH	Input	Interval	No		No	.	.	.	0
ETHPAYM	Input	Nominal	No		No	.	.	4	0
ONLAST	Input	Interval	No		No	.	.	.	0
RDERSIZE	Rejected	Interval	No		No	.	.	.	0
CPAYM	Input	Binary	No		No	.	.	2	0
ESPOND	Target	Binary	No		No	.	.	2	0
TATE	Rejected	Nominal	No		No	.	.	21	0
ENURE	Input	Interval	No		No	.	.	.	0
OTORDQ01	Input	Interval	No		No	.	.	8	0
OTORDQ02	Input	Interval	No		No	.	.	7	0
OTORDQ03	Input	Interval	No		No	.	.	6	0
OTORDQ04	Input	Interval	No		No	.	.	11	0
OTORDQ05	Input	Interval	No		No	.	.	6	0
OTORDQ06	Input	Interval	No		No	.	.	7	0
OTORDQ07	Input	Interval	No		No	.	.	7	0
OTORDQ08	Input	Interval	No		No	.	.	10	0

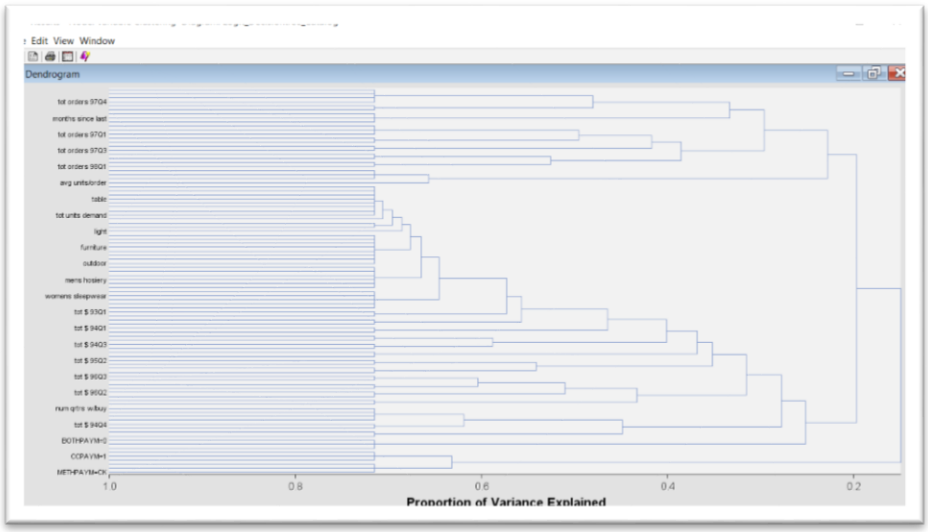
3. Added **Variable Clustering** node and changed **Includes Class Variables** property to Yes and Variable Selection property as Best Variables.



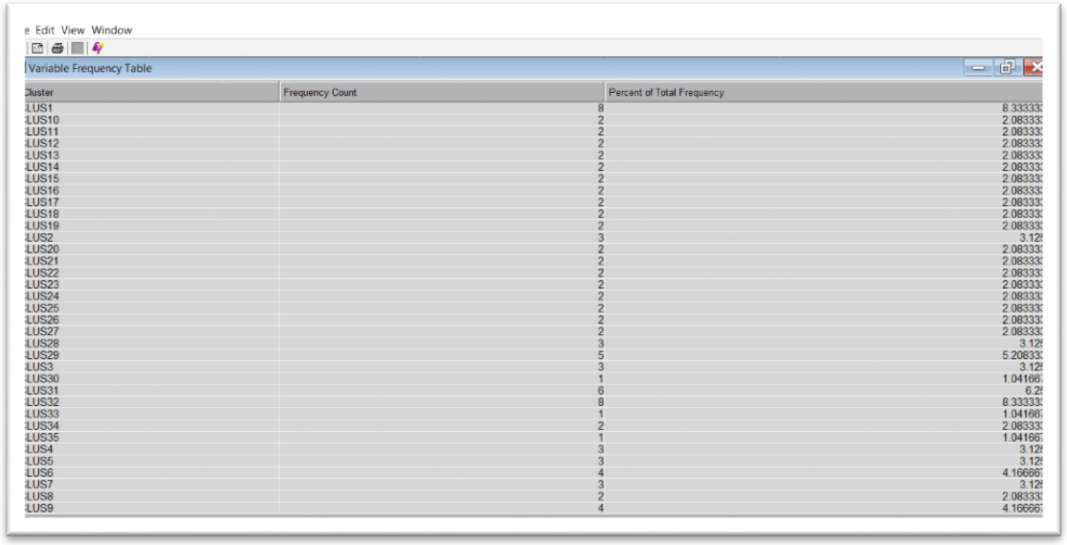
Property	Value
General	
Node ID	VarClus
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	Yes
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	.
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Wari	No
Score	
Variable Selection	Best Variables
Interactive Selection	...
Hides Rejected Variables	Yes

4. Results of Variable Clustering node.



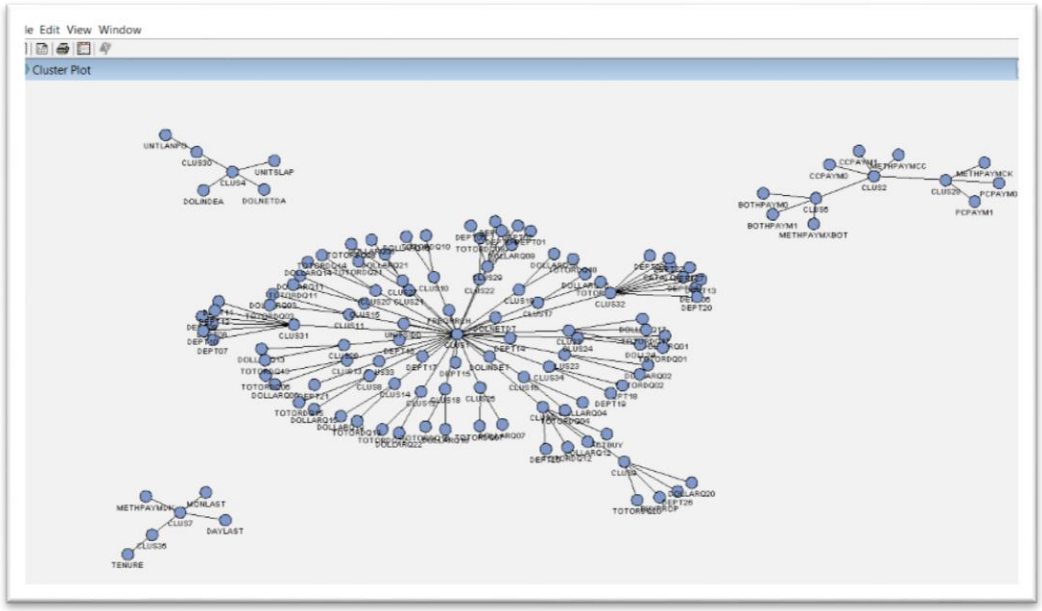


The Dendrogram graph shows the hierarchical structure of the clusters.



Cluster	Frequency Count	Percent of Total Frequency
CLUS1	8	8.33333
CLUS10	2	2.08333
CLUS11	2	2.08333
CLUS12	2	2.08333
CLUS13	2	2.08333
CLUS14	2	2.08333
CLUS15	2	2.08333
CLUS16	2	2.08333
CLUS17	2	2.08333
CLUS18	2	2.08333
CLUS19	2	2.08333
CLUS2	3	3.125
CLUS20	2	2.08333
CLUS21	2	2.08333
CLUS22	2	2.08333
CLUS23	2	2.08333
CLUS24	2	2.08333
CLUS25	2	2.08333
CLUS26	2	2.08333
CLUS27	2	2.08333
CLUS28	3	3.125
CLUS29	5	5.20833
CLUS3	3	3.125
CLUS30	1	1.04166
CLUS31	6	6.25
CLUS32	8	8.33333
CLUS33	1	1.04166
CLUS34	2	2.08333
CLUS35	1	1.04166
CLUS4	3	3.125
CLUS5	3	3.125
CLUS6	4	4.16666
CLUS7	3	3.125
CLUS8	2	2.08333
CLUS9	4	4.16666

The Variable Frequency Table shows the frequency of the input in each cluster

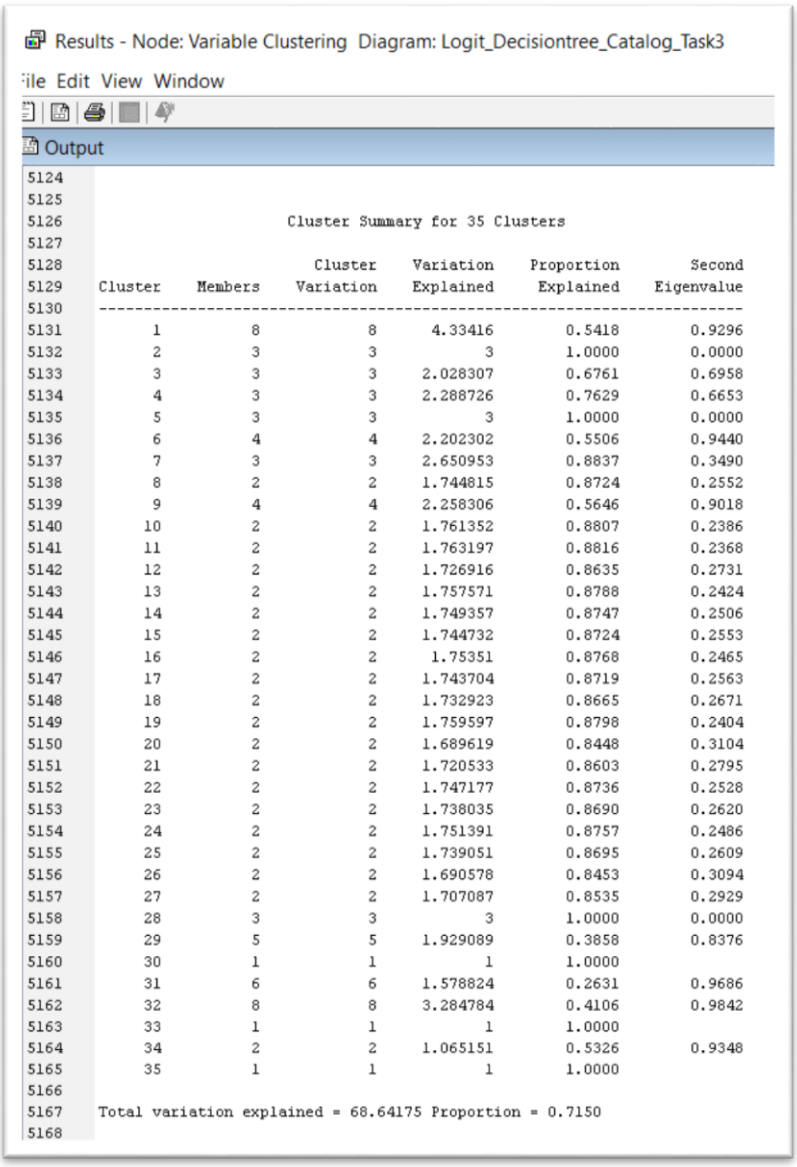


The Cluster Plot provide a tree diagram for the Dendrogram graph.

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	DOLINDET	total $ dem...	0.923851	CLUS32	0.529512	Variable	0.161852	YES
CLUS10	DOLLARQ10	tot $ 95Q2	0.880676	CLUS1	0.077855	Variable	0.129399	YES
CLUS11	DOLLARQ03	tot $ 93Q3	0.881598	CLUS1	0.04143	Variable	0.123519	YES
CLUS12	TOTORDQ22	tot orders 9...	0.863458	CLUS7	0.043322	Variable	0.142725	YES
CLUS13	TOTORDQ06	tot orders 9...	0.878785	CLUS1	0.086547	Variable	0.132699	YES
CLUS14	TOTORDQ19	tot orders 9...	0.874679	CLUS1	0.042882	Variable	0.130936	YES
CLUS15	TOTORDQ11	tot orders 9...	0.872366	CLUS32	0.074455	Variable	0.137902	YES
CLUS16	DOLLARQ04	tot $ 93Q4	0.876755	CLUS1	0.054508	Variable	0.13035	YES
CLUS17	TOTORDQ05	tot orders 9...	0.871852	CLUS32	0.082009	Variable	0.139596	YES
CLUS18	DOLLARQ16	tot $ 96Q4	0.866462	CLUS1	0.084269	Variable	0.145827	YES
CLUS19	TOTORDQ18	tot orders 9...	0.879799	CLUS1	0.058933	Variable	0.127729	YES
CLUS2	CCPAYM0	CCPAYM=0	1	CLUS28	0.3108	Variable	0	YES
CLUS20	TOTORDQ14	tot orders 9...	0.84481	CLUS32	0.057243	Variable	0.164613	YES
CLUS21	TOTORDQ21	tot orders 9...	0.860266	CLUS1	0.043468	Variable	0.146084	YES
CLUS22	DOLLARQ09	tot $ 95Q1	0.873589	CLUS1	0.074568	Variable	0.136597	YES
CLUS23	DOLLARQ02	tot $ 93Q2	0.869018	CLUS1	0.092496	Variable	0.144333	YES
CLUS24	TOTORDQ01	tot orders 9...	0.875695	CLUS1	0.105666	Variable	0.138991	YES
CLUS25	TOTORDQ07	tot orders 9...	0.869525	CLUS1	0.082185	Variable	0.142158	YES
CLUS26	TOTORDQ13	tot orders 9...	0.845289	CLUS32	0.070058	Variable	0.166366	YES
CLUS27	DOLLARQ08	tot $ 94Q4	0.853544	CLUS1	0.094627	Variable	0.161764	YES
CLUS28	METHPAYM...	METHPAYM...	1	CLUS2	0.3108	Variable	0	YES
CLUS29	DEPT03	womens un...	0.473772	CLUS1	0.200941	Variable	0.658559	YES
CLUS3	DOLLARQ17	tot $ 97Q1	0.81237	CLUS1	0.078702	Variable	0.203659	YES
CLUS30	UNTLANPO	avg units/or...	1	CLUS4	0.123804	Variable	0	YES
CLUS31	DEPT12	mens misc	0.367295	CLUS1	0.090206	Variable	0.695438	YES
CLUS32	CATALOGC...	number of c...	0.789691	CLUS1	0.61167	Variable	0.541574	YES
CLUS33	DEPT21	light	1	CLUS1	0.010229	Variable	0	YES
CLUS34	DEPT19	window	0.532576	CLUS1	0.026522	Variable	0.480159	YES
CLUS35	TENURE	months sin...	1	CLUS7	0.192899	Variable	0	YES
CLUS4	DOLINDEA	avg $ dema...	0.912853	CLUS30	0.262021	Variable	0.118089	YES
CLUS5	BOTHPAYM0	BOTHPAYM...	1	CLUS2	0.171641	Variable	5.36E-16	YES
CLUS6	TOTORDQ12	tot orders 9...	0.732581	CLUS32	0.077932	Variable	0.290021	YES
CLUS7	MONLAST	months sin...	0.95065	CLUS35	0.197535	Variable	0.061498	YES
CLUS8	TOTORDQ15	tot orders 9...	0.872407	CLUS1	0.064934	Variable	0.136453	YES
CLUS9	TOTORDQ20	tot orders 9...	0.821266	CLUS7	0.054505	Variable	0.189037	YES
CLUS1	CLUS1	Cluster 1	1	CLUS32	0.590055	ClusterComp	0	NO
CLUS1	DOLNETDT	avg $ net d...	0.916273	CLUS32	0.524352	Variable	0.176027	NO
CLUS1	UNITSIDD	tot units de...	0.84359	CLUS32	0.627776	Variable	0.420203	NO
CLUS1	FREQPRCH	lifetime ord...	0.749188	CLUS32	0.51671	Variable	0.518969	NO
CLUS1	DEPT14	bedding	0.4011	CLUS32	0.166786	Variable	0.718783	NO
CLUS1	DEPT15	bath	0.176854	CLUS32	0.078769	Variable	0.893529	NO

The Variable Selection Table above provides a list of selected variables base on 1-R^2 ratio in each cluster. The lower the 1-R^2 ratio, the more likely a variable is selected in a cluster. The selected variable is thought to be the best representative of the cluster. By using this process, independent variables to be used in the following analysis are reduced from 96 to 35. Some variables, such as METHPAYMCK, UNTLANP0, have zero as their 1-R^2 ratio, indicating that they are either categorical or they are the only one of the cluster.



Results - Node: Variable Clustering Diagram: Logit_Decisiontree_Catalog_Task3

File Edit View Window

Output

Cluster Summary for 35 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	8	8	4.33416	0.5418	0.9296
2	3	3	3	1.0000	0.0000
3	3	3	2.028307	0.6761	0.6958
4	3	3	2.288726	0.7629	0.6653
5	3	3	3	1.0000	0.0000
6	4	4	2.202302	0.5506	0.9440
7	3	3	2.650953	0.8837	0.3490
8	2	2	1.744815	0.8724	0.2552
9	4	4	2.258306	0.5646	0.9018
10	2	2	1.761352	0.8807	0.2386
11	2	2	1.763197	0.8816	0.2368
12	2	2	1.726916	0.8635	0.2731
13	2	2	1.757571	0.8788	0.2424
14	2	2	1.749357	0.8747	0.2506
15	2	2	1.744732	0.8724	0.2553
16	2	2	1.75351	0.8768	0.2465
17	2	2	1.743704	0.8719	0.2563
18	2	2	1.732923	0.8665	0.2671
19	2	2	1.759597	0.8798	0.2404
20	2	2	1.689619	0.8448	0.3104
21	2	2	1.720533	0.8603	0.2795
22	2	2	1.747177	0.8736	0.2528
23	2	2	1.738035	0.8690	0.2620
24	2	2	1.751391	0.8757	0.2486
25	2	2	1.739051	0.8695	0.2609
26	2	2	1.690578	0.8453	0.3094
27	2	2	1.707087	0.8535	0.2929
28	3	3	3	1.0000	0.0000
29	5	5	1.929089	0.3858	0.8376
30	1	1	1	1.0000	
31	6	6	1.578824	0.2631	0.9686
32	8	8	3.284784	0.4106	0.9842
33	1	1	1	1.0000	
34	2	2	1.065151	0.5326	0.9348
35	1	1	1	1.0000	

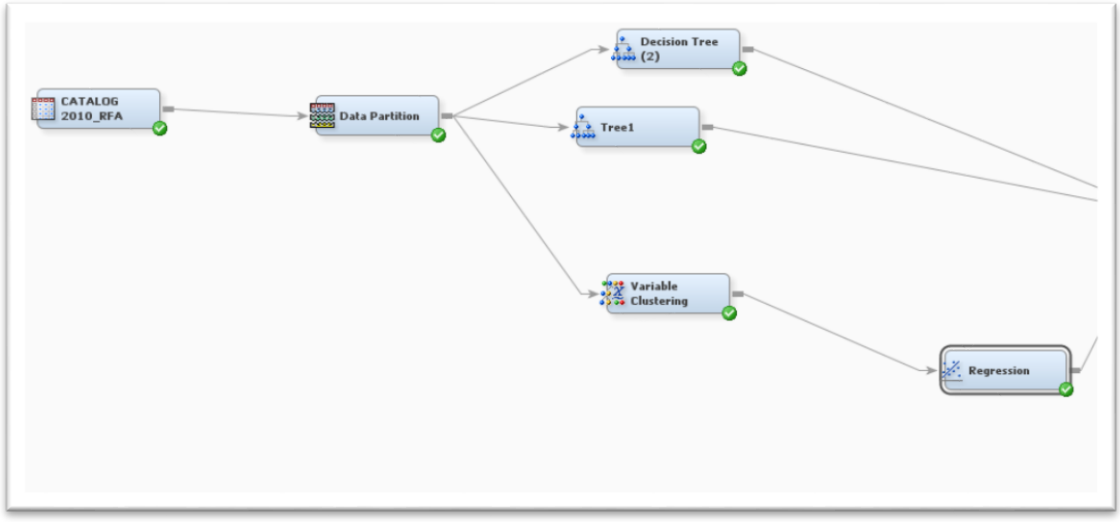
Total variation explained = 68.64175 Proportion = 0.7150

The Output window above shows total variation explained = 68.64175 and Proportion = 0.7150. "Proportion" represents the total explained variation divided by the sum of cluster variation. This value, 0.7150, indicates that about 71.5% of the total variation in the data can be accounted for by the thirty-five cluster components. The largest Second Eigenvalue is 0.9842. If the Second Eigenvalue is greater than one, the cluster needs to be split. Therefore, all the clusters are good clusters.

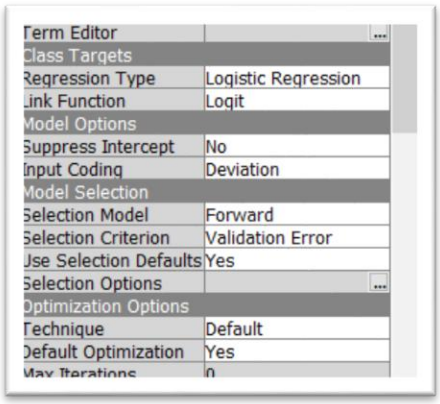
5 Clusters		Own	Next	1-R**2	Variable
luster	Variable	Cluster	Closest	Ratio	Label
luster 1	DEPT14	0.4011	0.1668	0.7188	bedding
	DEPT15	0.1769	0.0788	0.8935	bath
	DEPT16	0.1757	0.0805	0.8965	floor
	DEPT17	0.1476	0.0563	0.9032	table
	D0LINDET	0.9239	0.5295	0.1619	total $ demand
	D0LNETDT	0.9163	0.5244	0.1760	avg $ net demand
	FREQPRCH	0.7492	0.5167	0.5190	lifetime orders
	UNITSIDD	0.8436	0.6278	0.4202	tot units demand
luster 2	CCPAYM0	1.0000	0.3108	0.0000	CCPAYM=0
	CCPAYM1	1.0000	0.3108	0.0000	CCPAYM=1
	METHPAYMCC	1.0000	0.3108	0.0000	METHPAYM=CC
luster 3	D0LL24	0.4767	0.3047	0.7526	$ last 24 months
	D0LLARQ17	0.8124	0.0787	0.2037	tot $ 97Q1
	TOTORDQ17	0.7392	0.0625	0.2782	tot orders 97Q1
luster 4	D0LINDEA	0.9129	0.2620	0.1181	avg $ demand
	D0LNETDA	0.9039	0.2546	0.1290	tot $ net demand
	UNITSLAP	0.4720	0.1137	0.5957	avg price/unit
luster 5	BOTHPAYM0	1.0000	0.1716	0.0000	BOTHPAYM=0
	BOTHPAYM1	1.0000	0.1716	0.0000	BOTHPAYM=1
	METHPAYMXBOT	1.0000	0.1716	0.0000	METHPAYM=XBOT
luster 6	ACTBUY	0.4676	0.3380	0.8042	num qrtrs w/buy
	DEPT25	0.3441	0.2087	0.8289	food
	D0LLARQ12	0.6580	0.0974	0.3789	tot $ 95Q4
	TOTORDQ12	0.7326	0.0779	0.2900	tot orders 95Q4
luster 7	DAYLAST	0.9506	0.1976	0.0615	days since last
	MONLAST	0.9506	0.1975	0.0615	months since last
	METHPAYMDK	0.7497	0.1181	0.2838	METHPAYM=DK
luster 8	D0LLARQ15	0.8724	0.0862	0.1396	tot $ 96Q3
	TOTORDQ15	0.8724	0.0649	0.1365	tot orders 96Q3
luster 9	BUYPROP	0.5388	0.1214	0.5250	% quarters w/buy
	DEPT26	0.1761	0.0493	0.8666	gift
	D0LLARQ20	0.7222	0.0645	0.2969	tot $ 97Q4

Introducing Logit: Estimation of Logistic regression model. Using selected 35 variables from cluster output.

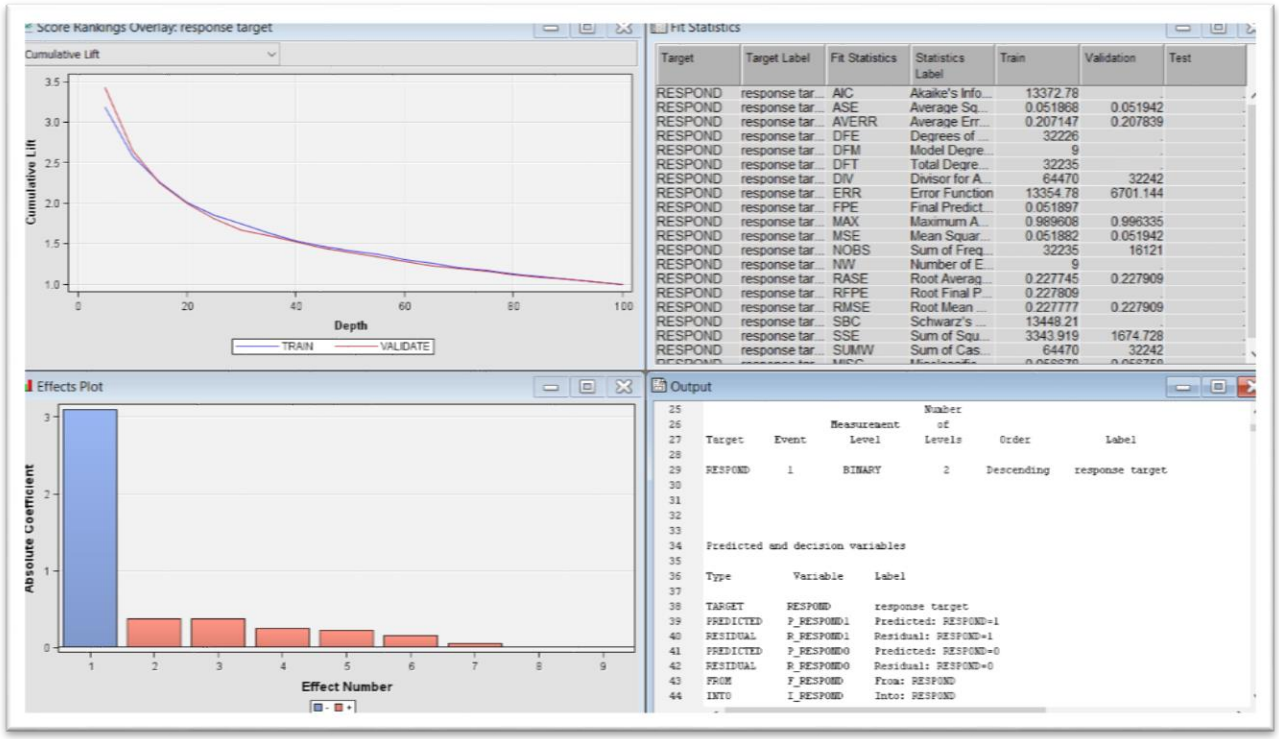
5. Connect regression node to variable clustering node.

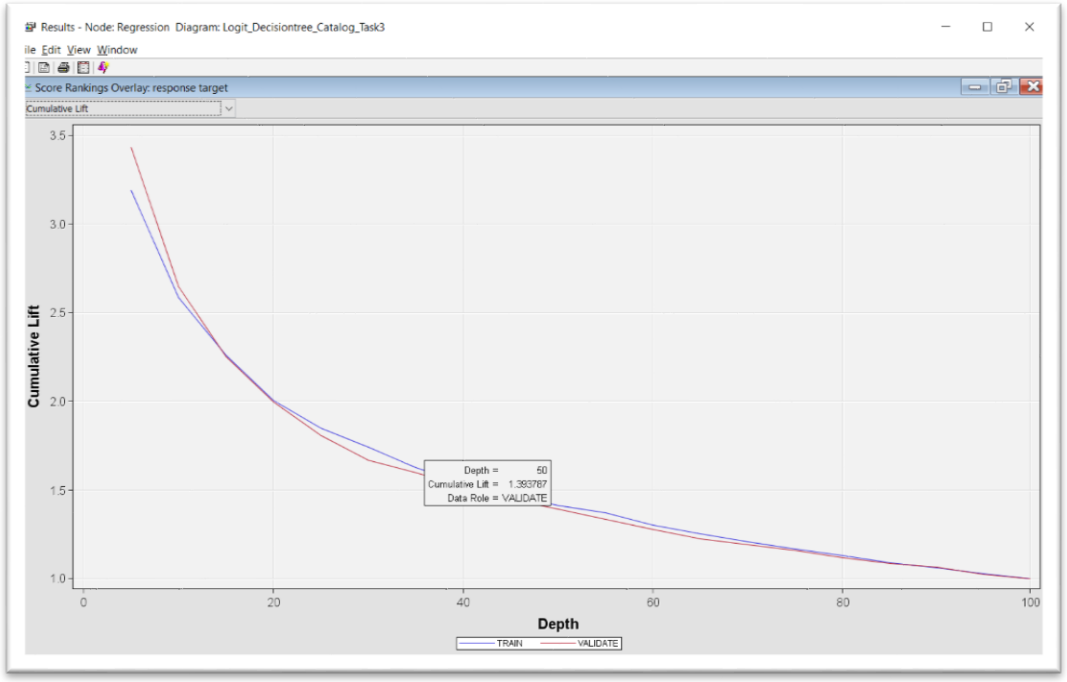


Select the model property as forward and selection criterion as Validation error.
Results of Regression node:

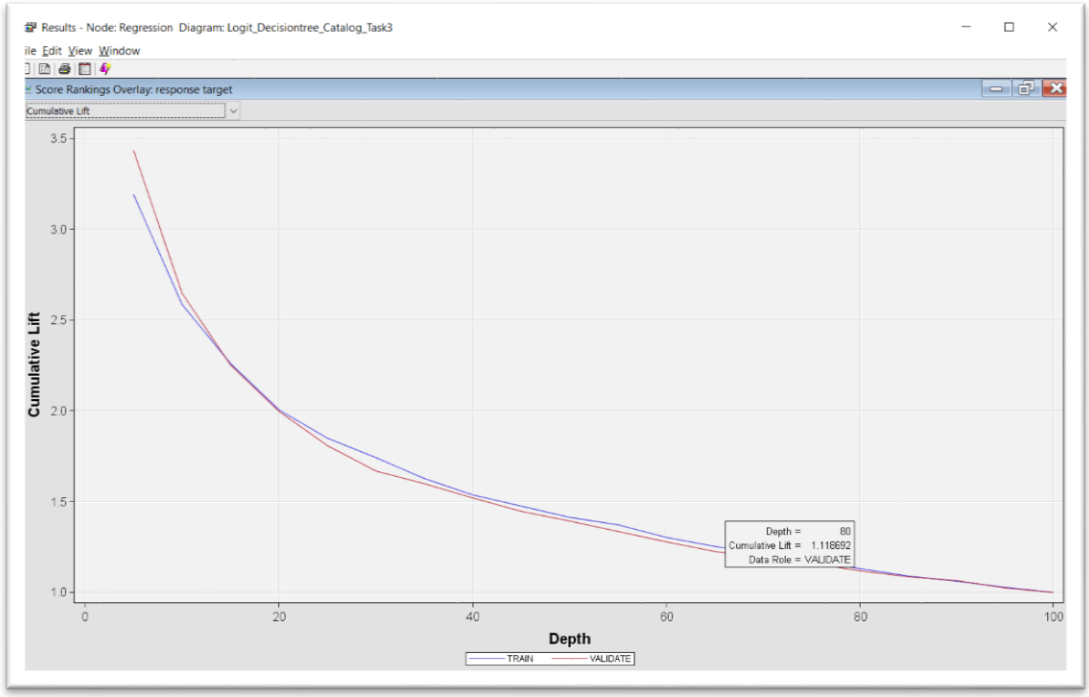


The Results window has four sub-windows, they are Score Rankings Overlay, Fit Statistics, Effects Plot, and Output as shown below.



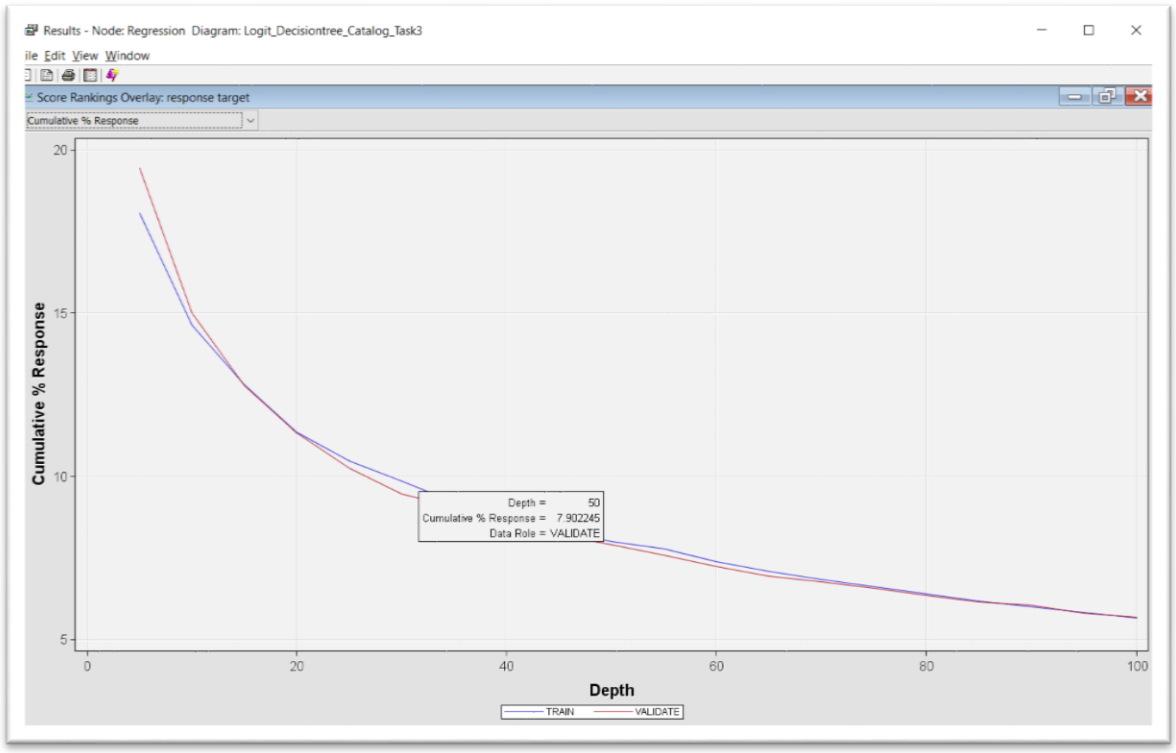


From above graph, lift is 1.39 at the 50th percentile on the validation dataset. This means that if the catalog company mailed to the top 50 percent of its customers based on the predicted probabilities, then they would obtain 1.39 times more responses compared to a 50 percent random sample of the customers.

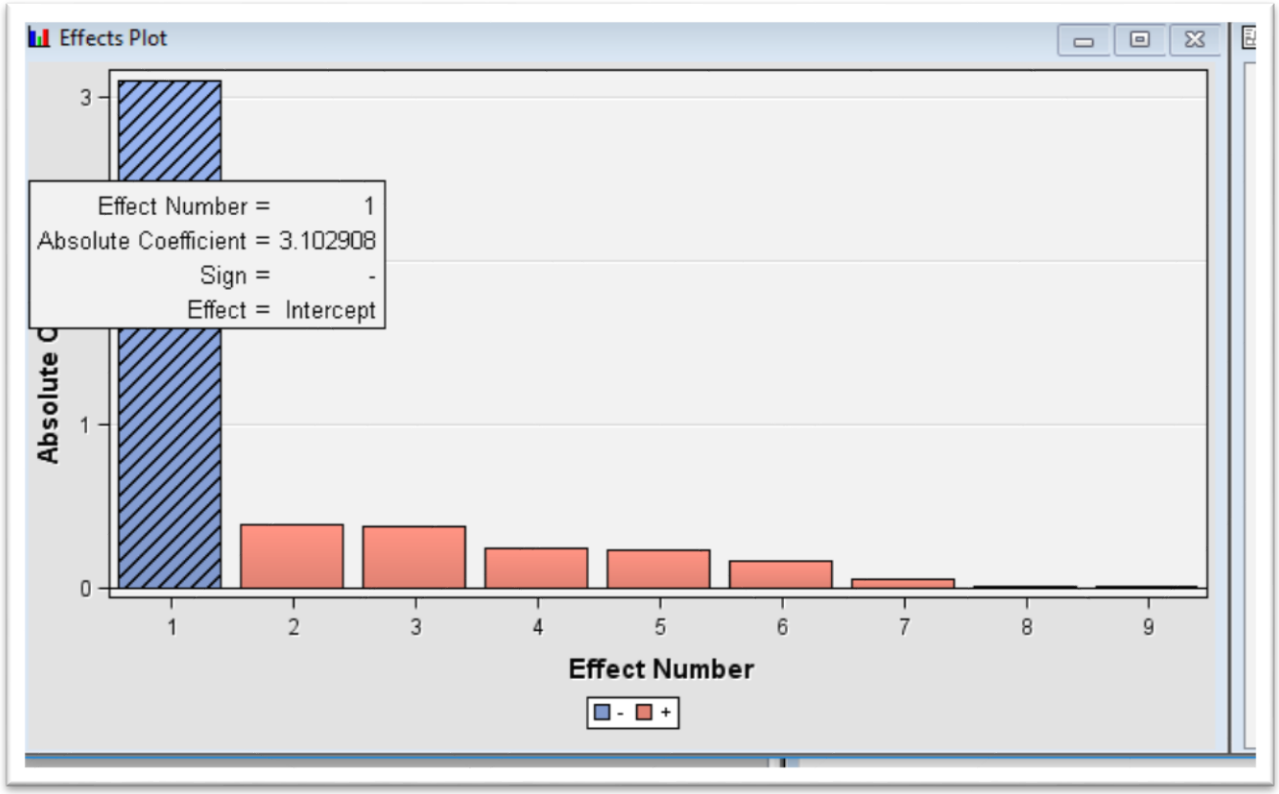


Similarly, we can infer from the above graph that at the 80th percentile, lift is 1.11 on the validation data set. This means that if the catalog company mailed to the top 80 percent of its customers based on the predicted probabilities, it would obtain 1.11 times more responses compared to 80 percent random sample of the customers.

From above two graphs we understood that Lift generally decreases as you choose larger and larger proportions of the data.



The Cumulative %Response curve shows the same trend as above graphs. The plotted values are cumulative actual probabilities of responders. From this graph, we see that the probability of response rate of top 50% of people who received catalogs is 7.9 times more compared to that of 50% random sample of customers.



The Effects Plot window above contains a bar chart of the absolute value of the model effects. The greater the absolute value, the more important that variable is to the regression model. Blue means a negative relationship while red means a positive relationship. Effect number indicates variables numbers. Effect 1 is corresponding to the Intercept of the logit regression model, which has the most negative effect on the model. This corresponds to the final output of the model as shown below.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-3.1773	0.0619	2634.34	<.0001		0.042
CATALOGCNT	1	0.0422	0.0105	16.18	<.0001	0.0728	1.043
CCPAYMO	1	0.1530	0.0514	8.87	0.0029	0.0416	1.165
DEPT03	1	0.0220	0.00712	9.55	0.0020	0.0342	1.022
DOLINDET	1	0.000046	0.000087	0.28	0.5970	0.00801	1.000
DOLLARQ09	1	-0.00266	0.00115	5.32	0.0211	-0.0305	0.997
MONLAST	1	-0.00596	0.000938	40.39	<.0001	-0.1324	0.994
TOTORDQ05	1	0.1597	0.0619	6.65	0.0099	0.0285	1.173
TOTORDQ12	1	0.1570	0.0460	11.65	0.0006	0.0353	1.170
TOTORDQ18	1	0.2115	0.0588	12.92	0.0003	0.0382	1.236
TOTORDQ19	1	0.2036	0.0625	10.60	0.0011	0.0344	1.226
TOTORDQ20	1	0.3796	0.0428	78.55	<.0001	0.0966	1.462
TOTORDQ21	1	0.2061	0.0588	12.29	0.0005	0.0375	1.229
TOTORDQ22	1	0.3543	0.0585	36.72	<.0001	0.0614	1.425

Fit Statistics:

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

Target Profile

Ordered Value	RESPOND	Total Frequency
1	1	1825
2	0	30410

Output

Fit Statistics

Target=RESPOND Target Label=response target

Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	13372.78	.
ASE	Average Squared Error	0.05	0.05
AVERR	Average Error Function	0.21	0.21
DFE	Degrees of Freedom for Error	32226.00	.
DFM	Model Degrees of Freedom	9.00	.
DFT	Total Degrees of Freedom	32235.00	.
DIV	Divisor for ASE	64470.00	32242.00
ERR	Error Function	13354.78	6701.14
FPE	Final Prediction Error	0.05	.
MAX	Maximum Absolute Error	0.99	1.00
MSE	Mean Square Error	0.05	0.05
NOBS	Sum of Frequencies	32235.00	16121.00
_NW_	Number of Estimate Weights	9.00	.
RASE	Root Average Sum of Squares	0.23	0.23
RFPE	Root Final Prediction Error	0.23	.
RMSE	Root Mean Squared Error	0.23	0.23
SBC	Schwarz's Bayesian Criterion	13448.21	.
SSE	Sum of Squared Errors	3343.92	1674.73
SUMW	Sum of Case Weights Times Freq	64470.00	32242.00
MISC	Misclassification Rate	0.06	0.06

Classification Table

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	94.3739	99.9507	30395	94.2919
1	0	5.6261	99.2877	1812	5.6212
0	1	53.5714	0.0493	15	0.0465
1	1	46.4286	0.7123	13	0.0403

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	CCPAYM0	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

he selected model, based on the error rate for the validation data, is the model trained in Step 8. It consists of the following effects:

ntercept CATALOGCNT DOLINDET MONLAST TOTORDQ12 TOTORDQ18 TOTORDQ20 TOTORDQ21 TOTORDQ22

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	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

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 picture above is the final regression model for the dataset. It includes 8 variables, and each variable is statistically significant as p value is < 0.01. Based on this model, we can say one-unit change in MONLAST (months since last purchase) variable corresponds to a .00586 decrease in the log of odds ratio of purchasing a product from the catalog.

One way to compare parameter estimates to different variables is to use standardized estimates. This method converts parameter estimates into standard deviation units. The absolute value of a standardized estimate can be used for approximate ranking of a predictor variable by its relative importance. **MONLAST** is 0.13 as standard estimate and is the most important predictor variable followed by **TOTORDQ20** with a standard estimate of 0.096 and **CATALOGCNT** with a standard estimate of 0.091**.**

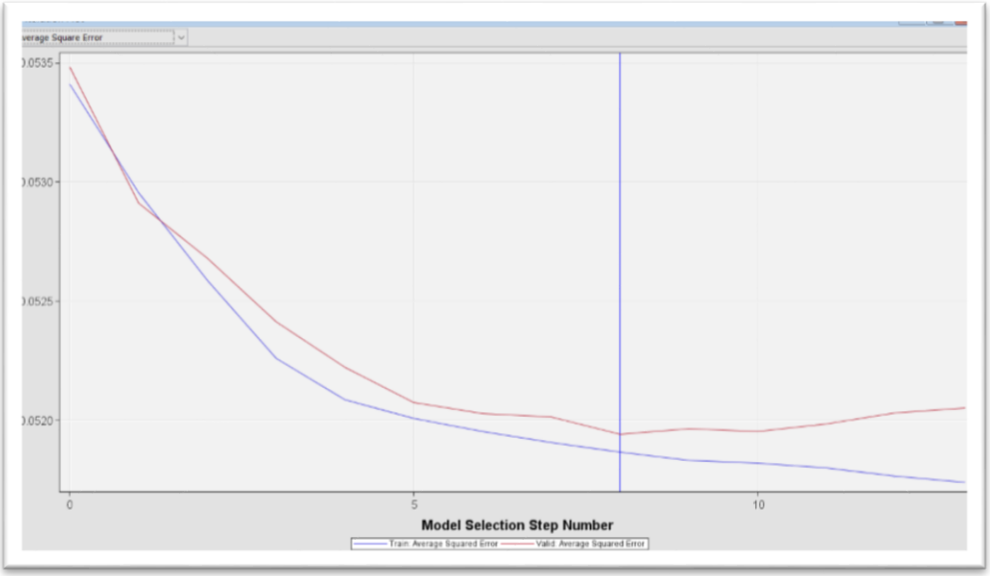
The odds ratio measures the effect of the predictor variable on the outcome. Here odds ratio for all variables is more than one except for MONLAST. TOTORDQ20 (total number of orders by calendar quarter) has highest odds ratio which indicates a positive effect in responders.

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

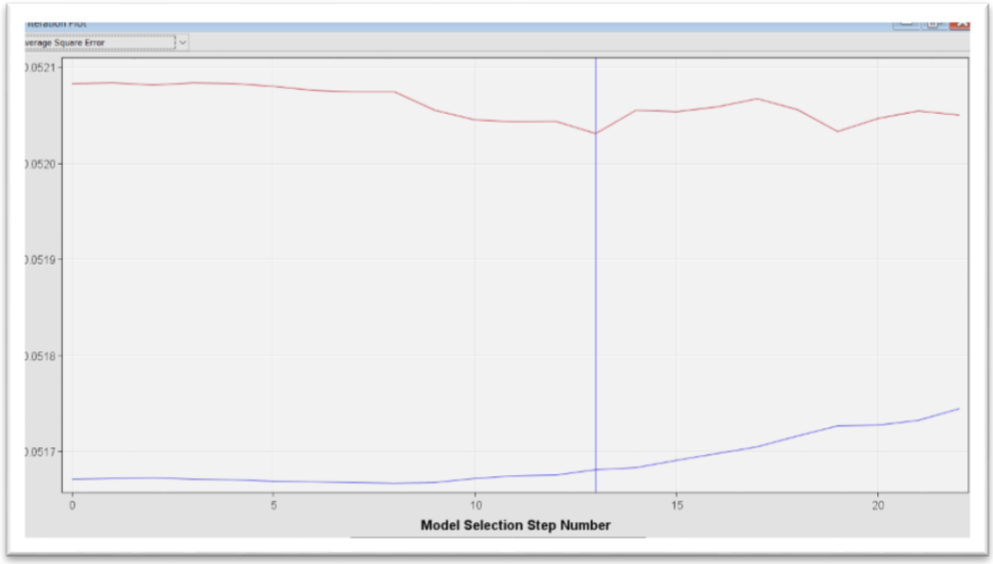
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	243.140	3.43140	3.43140	19.4548	19.4548	807	0.18015
10	164.623	1.86007	2.64623	10.5459	15.0031	806	0.10221
15	125.304	1.46617	2.25304	8.3127	12.7739	806	0.08419
20	99.622	1.22546	1.99622	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

Above data is statistical representation of Graphs in Score Rankings Overlay window.

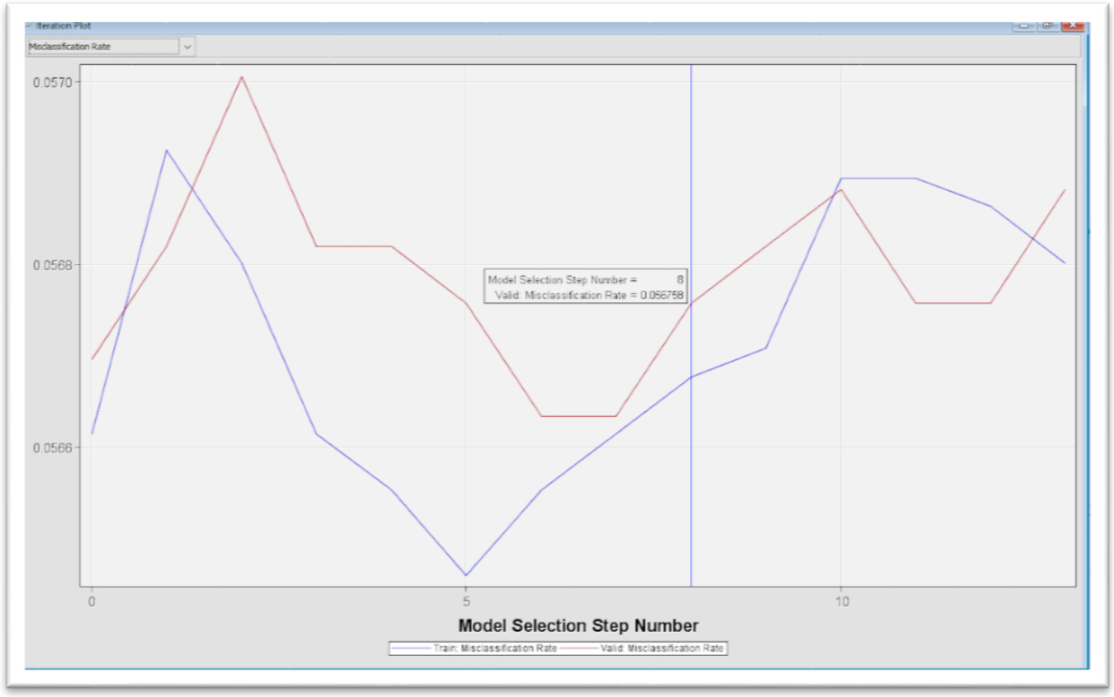
6. View model performance across the fitted models.

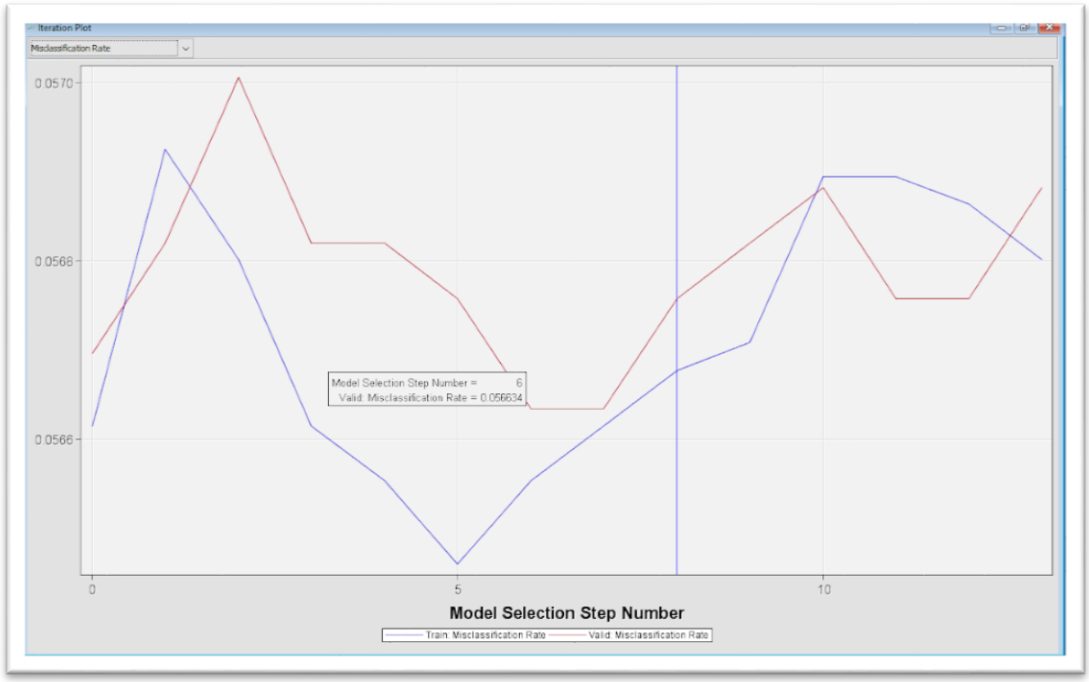


Above graph shows Average Square Error when Model Selection is forward. If the iteration plot shows that the validation ASE is decreasing, we should increase the p values for forward selection to let more variables into the model. As in this case, ASE is the smallest for validation data at Model 8, therefore, Model 8 is the final model.



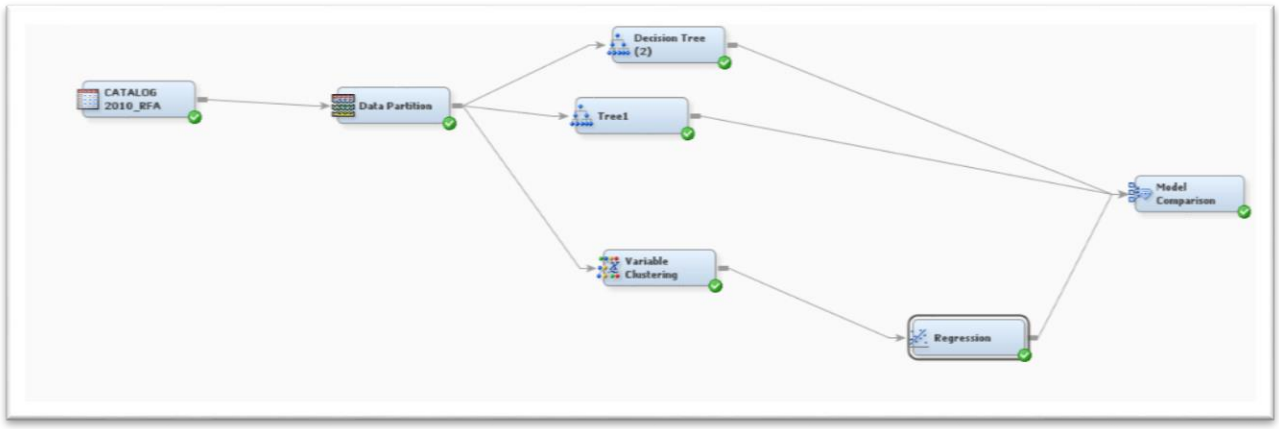
Above graph shows Average Square Error when Model Selection is backward. Compared to forward selection, ASE for backward is much higher for validation data, and there is a large discrepancy between the values of these two statistics on the training and validation data sets. The lowest ASE is at 14 step. Therefore, forward selection is better than backward selection in this case.



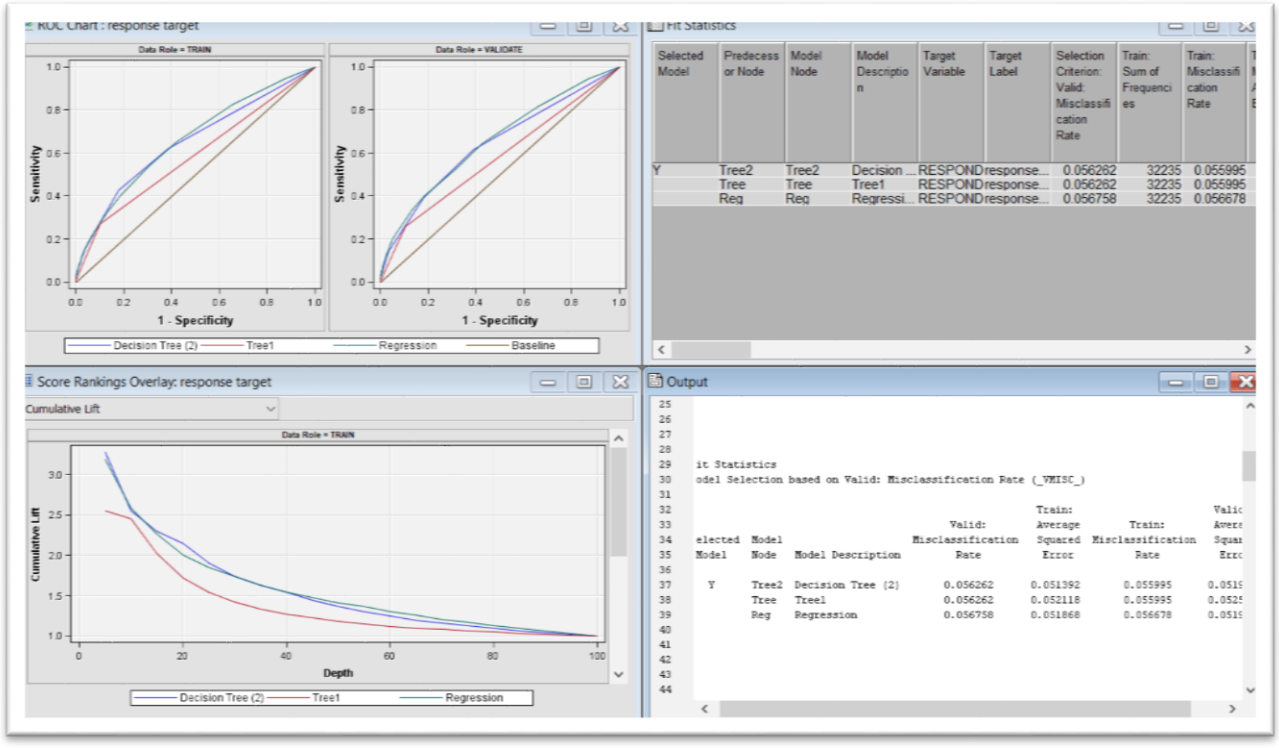


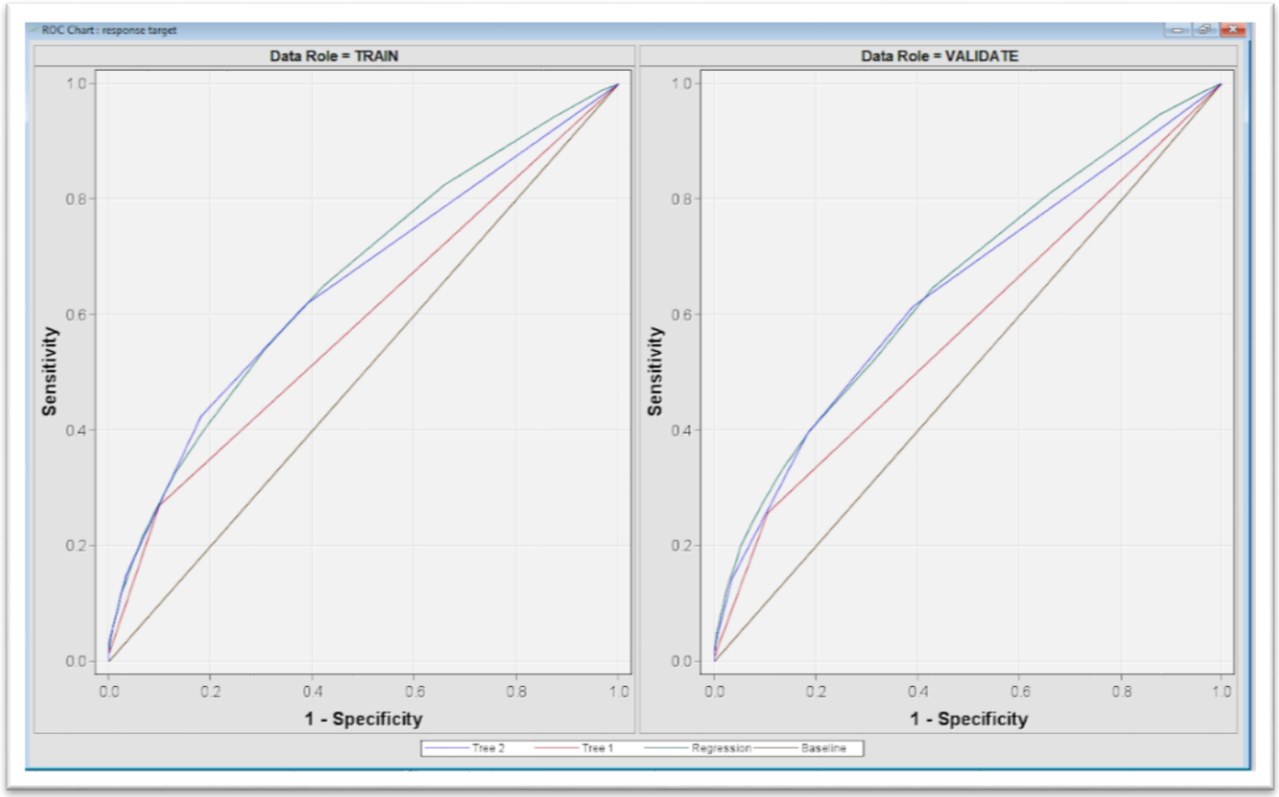
Above graphs shows Misclassification Rate for forward selection. The misclassification rate is the lowest at 6th step. Since the purpose of this model is for decision predictions. Thus, we will use Misclassification Rate as our selection criterion.

7. Run the Model Comparison node and view the results.



Result of Model Comparison:





The ROC chart window shows that the logistic regression and Tree 2 models perform similarly on the validation data set. The logistic regression is slightly better than Tree 2, as the area under the curve is bigger.

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Tree 2	0.056262	0.051392	0.055995	0.051990
	Tree	Tree 1	0.056262	0.052118	0.055995	0.052515
	Reg	Regression	0.056758	0.051868	0.056678	0.051942

Since the objective of our analysis is to predict decisions that would assist the company to increase its sales by sending catalogs to its target customers, our main focus should be on misclassification rate. For validation data set, the misclassification rate is lower in Tree 2

than the one in Logistic Regression model. Therefore, Tree 2 is the optimal model for this case.