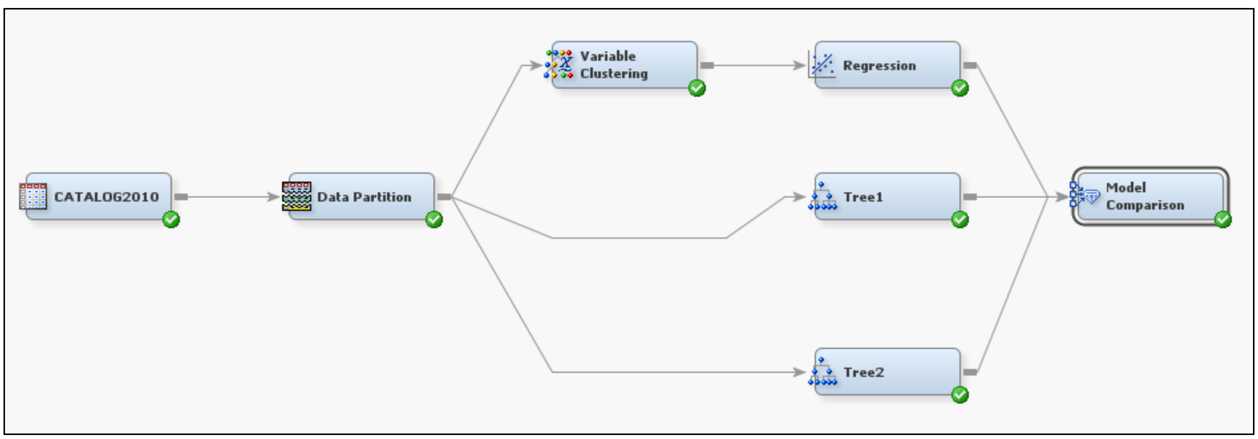
PART 3 - PREDICTIVE MODELING USING LOGISTIC REGRESSION

The steps to fit a logistic regression model are the same as the steps to build a decision tree model except that the Variable Clustering node is used to reduce redundancy and the Regression node is used to select relevant inputs.

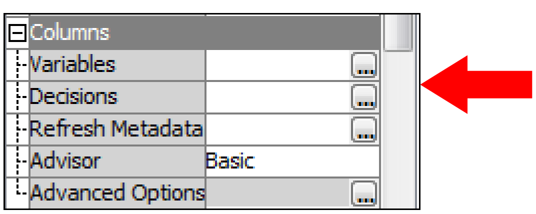
CATALOG CASE STUDY: FITTING A LOGISTIC REGRESSION MODEL

A mail-order catalog retailer wants to increase revenue by targeting customers who are most likely to purchase a product from the catalog. Use the **CATALOG2010** data from the previous chapter and fit a logistic regression model in SAS Enterprise Miner by simply adding the Variable Clustering node and the Regression node to the decision tree diagram. The steps that are added to the model-building process include eliminating redundant variables using the Variable Clustering node, eliminating irrelevant variables using the Regression node, and generating model assessment statistics and plots using the Model Comparison node.



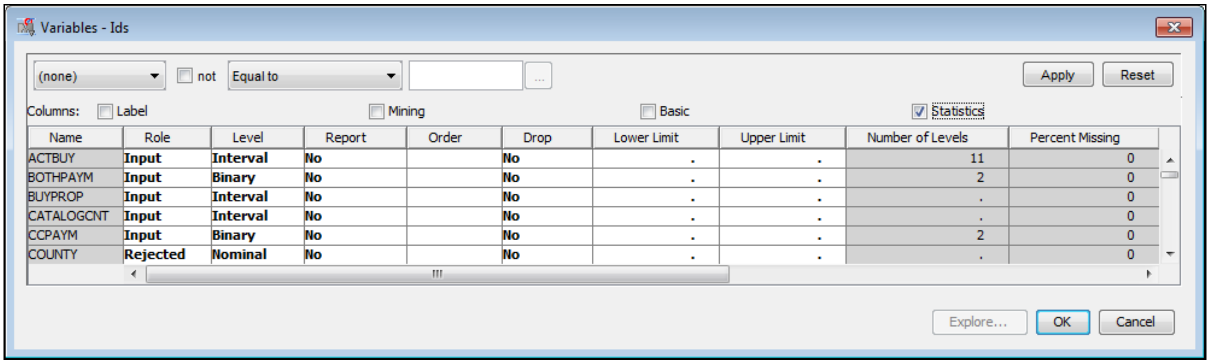
1. Review the **CATALOG2010** data set.

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



2. Compute some basic statistics.

   Select **Statistics** in the upper right corner.



3. 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.

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

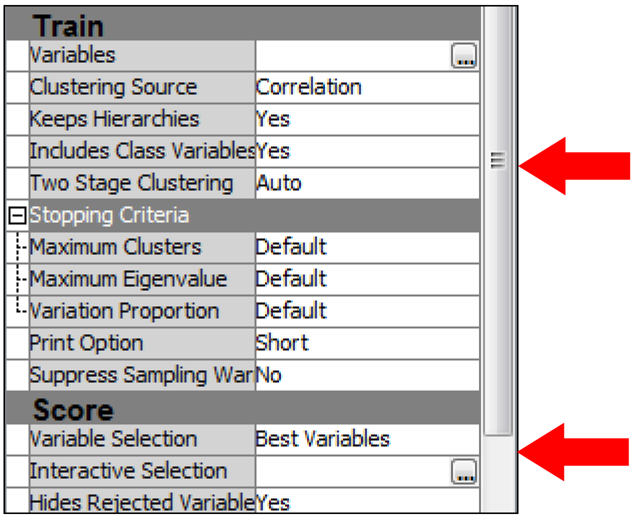
4. Cluster variables according to their similarities.

   a. Click the **Explore** tab and drag the **Variable Clustering** node into the diagram.

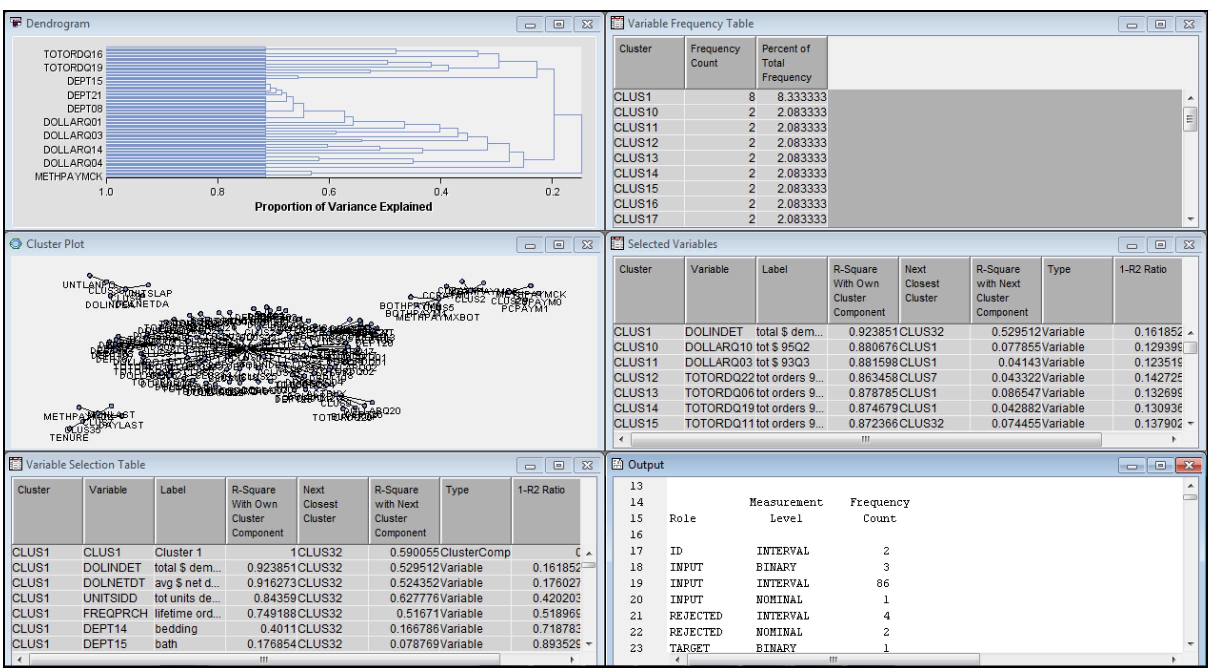
   b. Connect the **Data Partition** node to the **Variable Clustering** node.

   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)

The Variable Selection property specifies whether you would like the rotated cluster components or the lowest 1-R^2 ratio variables to be the cluster representatives. ***Best Variables indicates the 1-R^2 ratio.***



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



The output features several interpretations of the same idea: representing variable clusters.

- **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 $ demand	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 98Q2	0.863458	CLUS7	0.043322	Variable	0.142725	YES
CLUS13	TOTORDQ06	tot orders 94Q2	0.878785	CLUS1	0.086547	Variable	0.132699	YES
CLUS14	TOTORDQ19	tot orders 97Q3	0.874679	CLUS1	0.042882	Variable	0.130936	YES
CLUS15	TOTORDQ11	tot orders 95Q3	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 94Q1	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 97Q2	0.879799	CLUS1	0.058933	Variable	0.127729	YES
CLUS2	CCPAYM0	CCPAYM=0	1	CLUS28	0.3108	Variable	0	YES
CLUS20	TOTORDQ14	tot orders 96Q2	0.84481	CLUS32	0.057243	Variable	0.164613	YES
CLUS21	TOTORDQ21	tot orders 98Q1	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 93Q1	0.875695	CLUS1	0.105666	Variable	0.138991	YES
CLUS25	TOTORDQ07	tot orders 94Q3	0.869525	CLUS1	0.082185	Variable	0.142158	YES
CLUS26	TOTORDQ13	tot orders 96Q1	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=CK	1	CLUS2	0.3108	Variable	0	YES
CLUS29	DEPT03	womens underwear	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/order	1	CLUS4	0.123804	Variable	0	YES
CLUS31	DEPT12	mens misc	0.367295	CLUS1	0.090206	Variable	0.695438	YES
CLUS32	CATALOGC...	number of catalogs received	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 since 1st	1	CLUS7	0.192899	Variable	0	YES
CLUS4	DOLINDEA	avg $ demand	0.912853	CLUS30	0.262021	Variable	0.118089	YES
CLUS5	BOTHPAYM0	BOTHPAYM=0	1	CLUS2	0.171641	Variable	5.36E-16	YES
CLUS6	TOTORDQ12	tot orders 95Q4	0.732581	CLUS32	0.077932	Variable	0.290021	YES
CLUS7	MONLAST	months since last	0.95065	CLUS35	0.197535	Variable	0.061498	YES
CLUS8	TOTORDQ15	tot orders 96Q3	0.872407	CLUS1	0.064934	Variable	0.136453	YES
CLUS9	TOTORDQ20	tot orders 97Q4	0.821266	CLUS7	0.054505	Variable	0.189037	YES

The labels are expanded for easier interpretation. ***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.

35 Clusters		R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio	Variable Label
Cluster	Variable				
Cluster 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
	DOLINDET	0.9239	0.5295	0.1619	total $ demand
	DOLNETDT	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
Cluster 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
Cluster 3	DOLL24	0.4767	0.3047	0.7526	$ last 24 months
	DOLLARQ17	0.8124	0.0787	0.2037	tot $ 97Q1
	TOTORDQ17	0.7392	0.0625	0.2782	tot orders 97Q1
Cluster 4	DOLINDEA	0.9129	0.2620	0.1181	avg $ demand
	DOLNETDA	0.9039	0.2546	0.1290	tot $ net demand
	UNITSLAP	0.4720	0.1137	0.5957	avg price/unit
Cluster 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
Cluster 6	ACTBUY	0.4676	0.3380	0.8042	num qrtrs w/buy
	DEPT25	0.3441	0.2087	0.8289	food
	DOLLARQ12	0.6580	0.0974	0.3789	tot $ 95Q4
	TOTORDQ12	0.7326	0.0779	0.2900	tot orders 95Q4
Cluster 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
Cluster 8	DOLLARQ15	0.8724	0.0862	0.1396	tot $ 96Q3
	TOTORDQ15	0.8724	0.0649	0.1365	tot orders 96Q3
Cluster 9	BUYPROP	0.5388	0.1214	0.5250	% quarters w/buy
	DEPT26	0.1761	0.0493	0.8666	gift
	DOLLARQ20	0.7222	0.0645	0.2969	tot $ 97Q4
	TOTORDQ20	0.8213	0.0545	0.1890	tot orders 97Q4
Cluster 10	DOLLARQ10	0.8807	0.0779	0.1294	tot $ 95Q2
	TOTORDQ10	0.8807	0.0810	0.1298	tot orders 95Q2
Cluster 11	DOLLARQ03	0.8816	0.0414	0.1235	tot $ 93Q3
	TOTORDQ03	0.8816	0.0436	0.1238	tot orders 93Q3

(Continued on the next page.)

Cluster 12	DOLLARQ22	0.8635	0.0485	0.1435	tot $ 98Q2
	TOTORDQ22	0.8635	0.0433	0.1427	tot orders 98Q2
Cluster 13	DOLLARQ06	0.8788	0.0925	0.1336	tot $ 94Q2

	TOTORDQ06	0.8788	0.0865	0.1327	tot orders 94Q2
Cluster 14	DOLLARQ19	0.8747	0.0500	0.1319	tot $ 97Q3
	TOTORDQ19	0.8747	0.0429	0.1309	tot orders 97Q3
Cluster 15	DOLLARQ11	0.8724	0.0754	0.1380	tot $ 95Q3
	TOTORDQ11	0.8724	0.0745	0.1379	tot orders 95Q3
Cluster 16	DOLLARQ04	0.8768	0.0545	0.1303	tot $ 93Q4
	TOTORDQ04	0.8768	0.0673	0.1321	tot orders 93Q4
Cluster 17	DOLLARQ05	0.8719	0.0831	0.1398	tot $ 94Q1
	TOTORDQ05	0.8719	0.0820	0.1396	tot orders 94Q1
Cluster 18	DOLLARQ16	0.8665	0.0843	0.1458	tot $ 96Q4
	TOTORDQ16	0.8665	0.0874	0.1463	tot orders 96Q4
Cluster 19	DOLLARQ18	0.8798	0.0601	0.1279	tot $ 97Q2
	TOTORDQ18	0.8798	0.0589	0.1277	tot orders 97Q2
Cluster 20	DOLLARQ14	0.8448	0.0663	0.1662	tot $ 96Q2
	TOTORDQ14	0.8448	0.0572	0.1646	tot orders 96Q2
Cluster 21	DOLLARQ21	0.8603	0.0475	0.1467	tot $ 98Q1
	TOTORDQ21	0.8603	0.0435	0.1461	tot orders 98Q1
Cluster 22	DOLLARQ09	0.8736	0.0746	0.1366	tot $ 95Q1
	TOTORDQ09	0.8736	0.0761	0.1368	tot orders 95Q1
Cluster 23	DOLLARQ02	0.8690	0.0925	0.1443	tot $ 93Q2
	TOTORDQ02	0.8690	0.0978	0.1452	tot orders 93Q2
Cluster 24	DOLLARQ01	0.8757	0.1086	0.1395	tot $ 93Q1
	TOTORDQ01	0.8757	0.1057	0.1390	tot orders 93Q1
Cluster 25	DOLLARQ07	0.8695	0.0824	0.1422	tot $ 94Q3
	TOTORDQ07	0.8695	0.0822	0.1422	tot orders 94Q3
Cluster 26	DOLLARQ13	0.8453	0.0930	0.1706	tot $ 96Q1
	TOTORDQ13	0.8453	0.0701	0.1664	tot orders 96Q1
Cluster 27	DOLLARQ08	0.8535	0.0946	0.1618	tot $ 94Q4
	TOTORDQ08	0.8535	0.1106	0.1647	tot orders 94Q4
Cluster 28	PCPAYM0	1.0000	0.3108	0.0000	PCPAYM=0
	PCPAYM1	1.0000	0.3108	0.0000	PCPAYM=1
	METHPAYMCK	1.0000	0.3108	0.0000	METHPAYM=CK
Cluster 29	DEPT01	0.3956	0.1594	0.7189	womens apparel
	DEPT02	0.3687	0.1327	0.7279	womens sleepwear
	DEPT03	0.4738	0.2009	0.6586	womens underwear
	DEPT04	0.3702	0.1596	0.7494	womens hosiery
	DEPT05	0.3208	0.1479	0.7971	womens footwear

(Continued on the next page.)

Cluster 30	UNTLANPO	1.0000	0.1238	0.0000	avg units/order
Cluster 31	DEPT07	0.1185	0.0231	0.9024	mens apparel
	DEPT08	0.2880	0.0951	0.7868	mens sleepwear
	DEPT09	0.2913	0.0482	0.7446	mens underwear
	DEPT10	0.3128	0.0804	0.7473	mens hosiery
	DEPT11	0.2010	0.0474	0.8388	mens footware
	DEPT12	0.3673	0.0902	0.6954	mens misc

Cluster 32	CATALOGCNT	0.7897	0.6117	0.5416	number of catalogs received
	DEPT06	0.3359	0.1902	0.8201	womens misc
	DEPT13	0.4801	0.2840	0.7261	kitchen
	DEPT20	0.0326	0.0169	0.9841	furniture
	DEPT22	0.5862	0.3539	0.6406	household
	DEPT23	0.4425	0.2303	0.7243	beauty
	DEPT24	0.3272	0.1334	0.7764	health
	DEPT27	0.2906	0.1364	0.8214	outdoor
Cluster 33	DEPT21	1.0000	0.0102	0.0000	light
Cluster 34	DEPT18	0.5326	0.0455	0.4897	chair
	DEPT19	0.5326	0.0265	0.4802	window
Cluster 35	TENURE	1.0000	0.1929	0.0000	months since 1st

There are 35 clusters and, therefore, 35 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
34	67.792590	0.7062	0.2631	1.014065	0.0326	1.0077
35	68.641746	0.7150	0.2631	0.984193	0.0326	0.9841

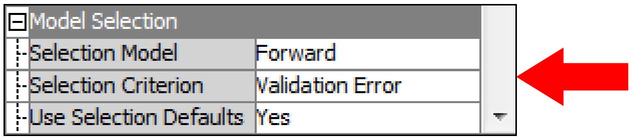
The clusters explained 71.5% of the variation in the data.

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

6. Estimate a logistic regression model.

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

b. Select **Selection Model** ⇨ **Forward** and select **Selection Criterion** ⇨ **Validation Error** from the Regression node Properties panel.

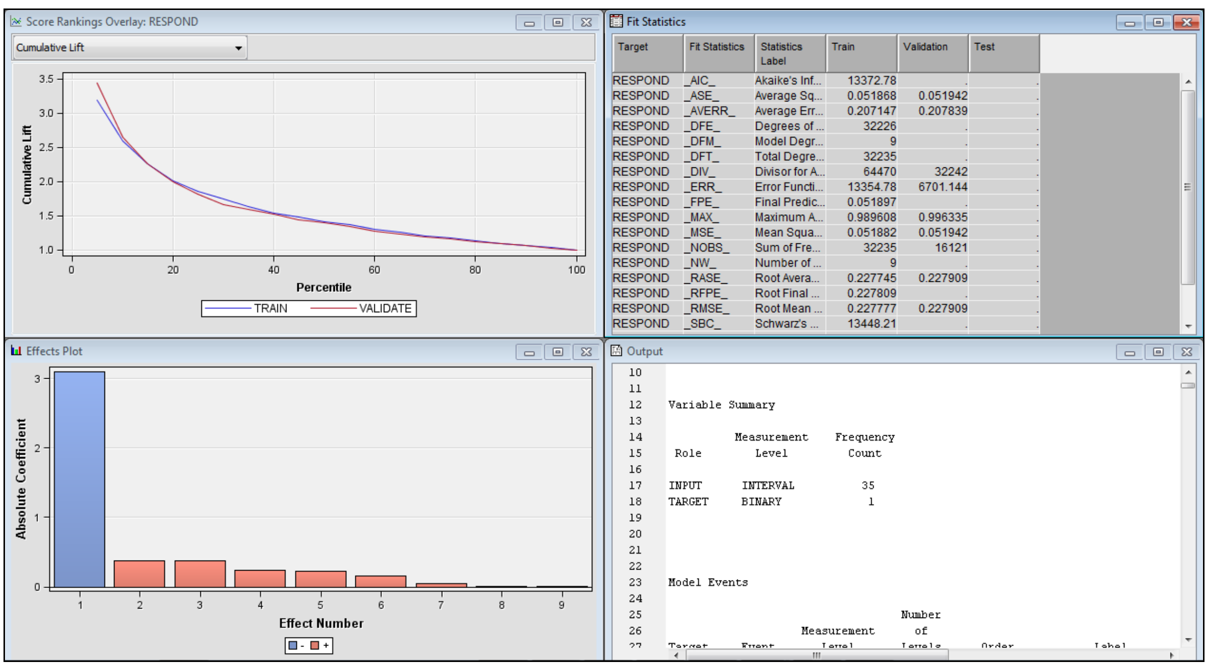


The choices for the selection criterion have several model fit statistics that are useful for model selection.

AIC is the Akaike information criterion, and SBC is the Schwarz’s Bayesian Criterion. These are goodness-of-fit measures that you can use to compare one model to another, and they are useful when you do not have validation data to use for selecting the best model. These measures adjust the -2 log likelihood statistic for the number of terms in the model and the number of observations. The difference between the two measures is that SBC uses a bigger penalty for extra variables. Therefore, SBC favors more parsimonious models. For both measures, lower values indicate a more desirable model.

Other choices in the selection criterion include validation error, validation misclassification, profit/loss, and validation profit/loss. When validation data is available, you should select the model based on the validation performance. In this example, you use validation error.

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



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

The Score Rankings Overlay window shows a cumulative lift chart where, for a given percentile, 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 3.43 on the validation data set. This means that if the catalog company mailed to the top 5 percent of its customers based on the predicted probabilities, then you would obtain 3.43 times more responders compared to a 5-percent random sample of the customers.

Other graphs that are available in the Score Rankings Overlay window include the Gains chart, the %Response chart, and the Cumulative %Response chart.

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.

The Output window gives the standard output for logistic regression.

```
Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          35
TARGET     BINARY             1




Model Events

                                          Number
                          Measurement       of
Target      Event            Level        Levels       Order           Label

RESPOND       1             BINARY           2       Descending    response target
```

The initial lines of the Output window summarize the roles of the variables used (or not) by the Regression node. The model has 35 inputs that predict a binary target.

```
The DMREG Procedure

                   Model Information

Training Data Set               EMWS1.VARCLUS_TRAIN.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                    Total
   Value     RESPOND     Frequency

       1     1                1825
       2     0               30410
```

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 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.

```
                      Summary of Forward Selection

        Effect              Number       Score                 Validation
  Step  Entered         DF      In  Chi-Square   Pr > ChiSq    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  CCPAYM1          1      11      9.3003       0.0023        6709.1
    12  TOTORDQ05        1      12      6.4600       0.0110        6716.5
    13  DOLLARQ09        1      13      5.3211       0.0211        6717.9


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

Intercept  CATALOGCNT  DOLINDET  MONLAST  TOTORDQ12  TOTORDQ18  TOTORDQ20  TOTORDQ21  TOTORDQ22
```

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

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

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.

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 parameter estimates measure the rate of change in the logit (log of the odds) corresponding to a one-unit change in the predictor variable, adjusted for the effects of the other predictors. For example, a one-unit change in **CATALOGCNT** (number of catalogs received) corresponds to a .054 increase in the log odds of purchasing a product from the catalog, adjusted for the other predictor variables. The Wald chi-square and its associated *p*-value test whether the parameter estimate is significantly different from 0.

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

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

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

Data Role=VALIDATE Target Variable=RESPOND

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	1	1	0.98951	0.0124
0.90-0.95	1	0	0.91630	0.0062
0.85-0.90	0	1	0.86650	0.0062
0.80-0.85	0	0	.	0.0000
0.75-0.80	0	0	.	0.0000
0.70-0.75	0	1	0.71097	0.0062
0.65-0.70	1	2	0.67365	0.0186
0.60-0.65	1	0	0.62780	0.0062
0.55-0.60	2	1	0.59372	0.0186
0.50-0.55	2	3	0.51951	0.0310
0.45-0.50	2	2	0.48270	0.0248
0.40-0.45	2	1	0.42371	0.0186
0.35-0.40	4	9	0.36608	0.0806
0.30-0.35	6	15	0.31939	0.1303
0.25-0.30	15	24	0.27050	0.2419
0.20-0.25	14	69	0.22243	0.5149
0.15-0.20	41	163	0.17055	1.2654
0.10-0.15	115	748	0.11918	5.3533
0.05-0.10	383	5483	0.06734	36.3873
0.00-0.05	324	8684	0.03636	55.8774

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

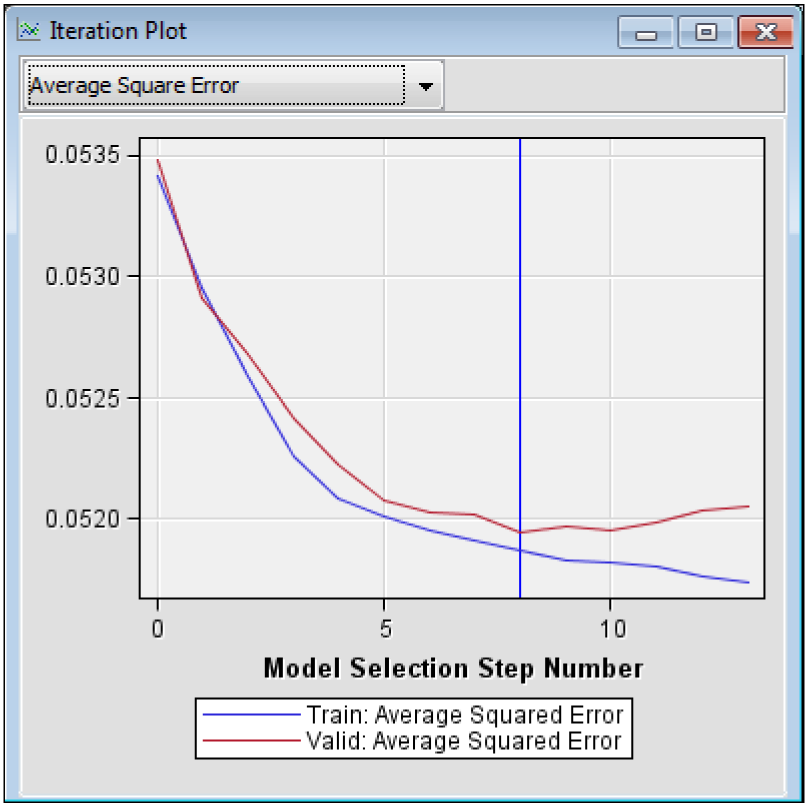
Data Role=VALIDATE Target Variable=RESPOND

Percentile	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

8. View model performance across the fitted models.

   Select **View** ⇨ **Model** ⇨ **Iteration Plot** in the Results window.

   The Iteration Plot window appears.

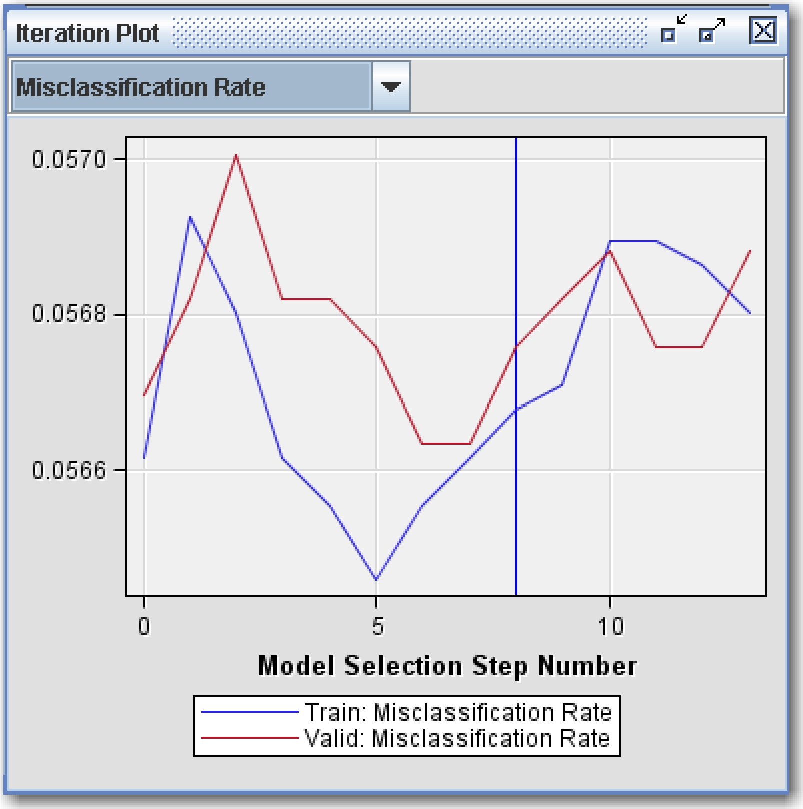


   The Iteration Plot window shows the average squared error (training and validation) from the model selected in each step of the backward selection process. The smallest average squared error occurs in model 8.

   ✎ If your iteration plot shows that the validation ASE is decreasing (so that the final model is selected), this suggests that you should change your *p*-values for forward selection to let more variables into the model. You can make this change in the Regression node's Properties panel.

9. View the misclassification rate.

From the Iteration Plot menu, select **Misclassification Rate**.

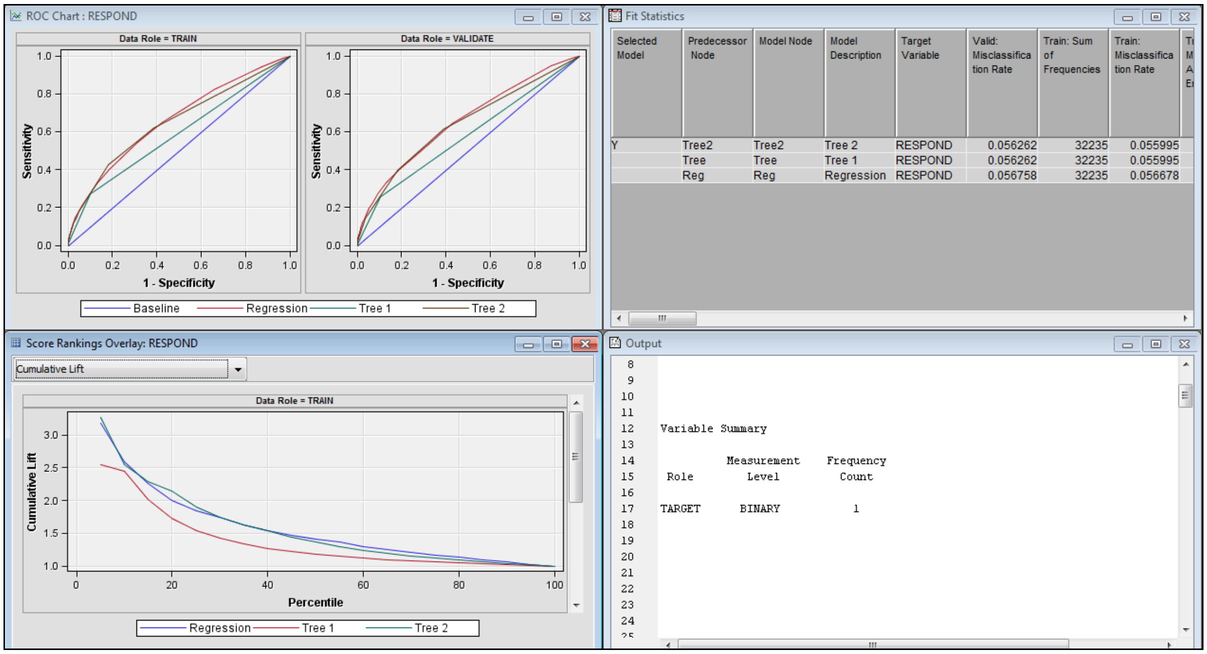


The iteration plot shows that the model with the smallest misclassification rate occurs in steps 6 and 7. If your analysis objective requires decision predictions, the predictions from the Step 6 model are as accurate as the predictions from the Step 7 model.

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.

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

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

Data Role=Valid

Statistics	Tree2	Tree	Reg
Valid: Kolmogorov-Smirnov Statistic	0.22	0.15	0.22
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	0.05	0.09	0.06
Valid: Cumulative Percent Captured Response	24.16	22.63	26.48
Valid: Percent Captured Response	8.53	10.97	9.30
Valid: Divisor for VASE	32242.00	32242.00	32242.00
Valid: Error Function	.	.	6701.14
Valid: Gain	141.49	126.21	164.62
Valid: Gini Coefficient	0.28	0.15	0.31
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.22	0.15	0.22
Valid: Kolmogorov-Smirnov Probability Cutoff	0.04	0.12	0.05
Valid: Cumulative Lift	2.41	2.26	2.65
Valid: Lift	1.71	2.20	1.86
Valid: Maximum Absolute Error	0.96	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	13.69	12.83	15.00
Valid: Percent Response	9.67	12.45	10.55
Valid: Root Mean Square Error	.	.	0.23
Valid: Sum of Squared Errors	1676.25	1693.19	1674.73
Valid: Sum of Case Weights Times Freq	32242.00	32242.00	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.