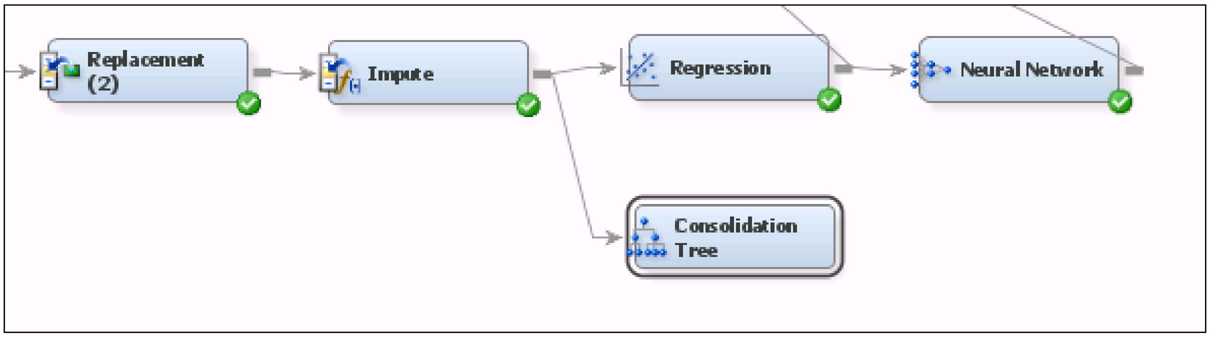
Consolidating Categorical Inputs

Follow these steps to use a tree model to group categorical input levels and create useful inputs for regression and neural network models.

1. Connect a **Decision Tree** node to the **Impute** node, and rename the Decision Tree node **Consolidation Tree**.

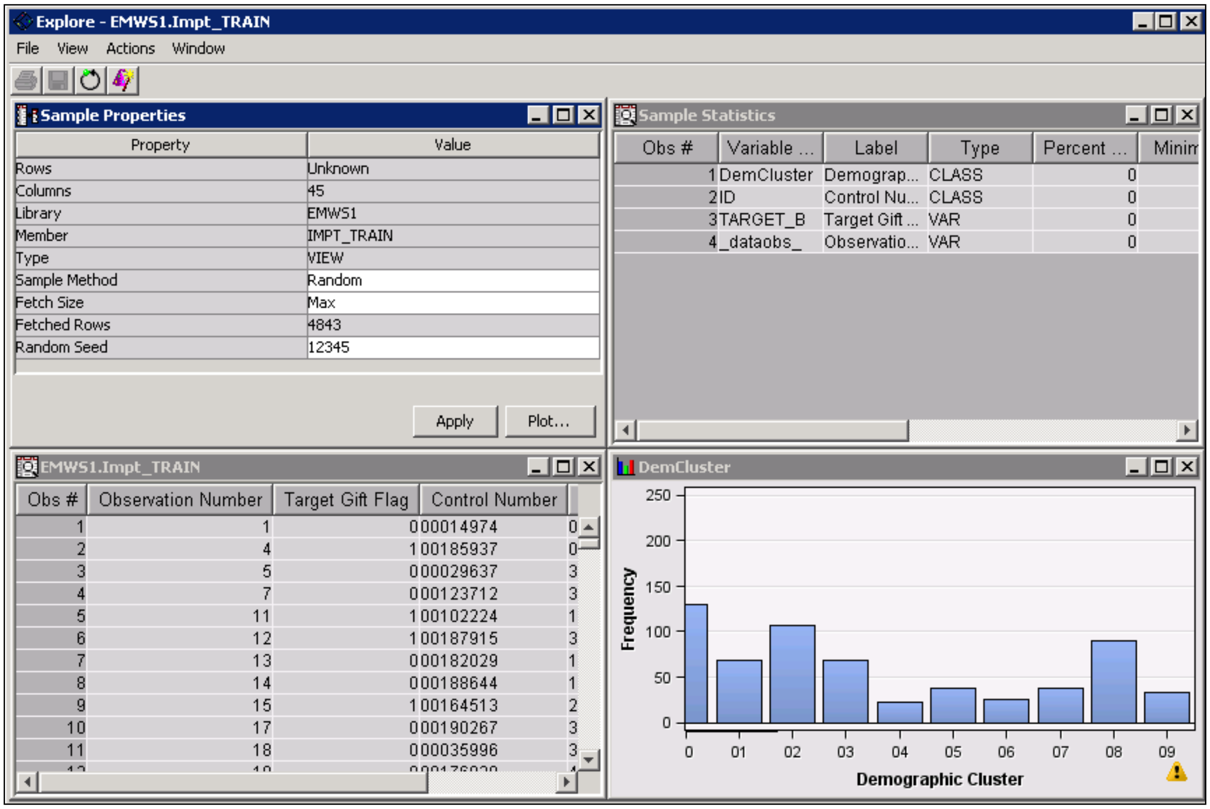


The Consolidation Tree node is used to group the levels of **DemCluster**, a categorical input with more than 50 levels.

You use a tree model to group these levels based on their associations with **TARGET_B**. From this grouping, a new modeling input is created. You can use this input in place of **DemCluster** in a regression or other model. In this way, the predictive prowess of **DemCluster** is incorporated into a model without the plethora of parameters needed to encode the original.

The grouping can be done autonomously by simply running the Decision Tree node, or interactively by using the node's interactive training features. You use the automatic method here.

2. Select **Variables** from the Consolidation Tree Properties panel.
3. Select **DemCluster** ⇨ **Explore**. The Explore window appears.

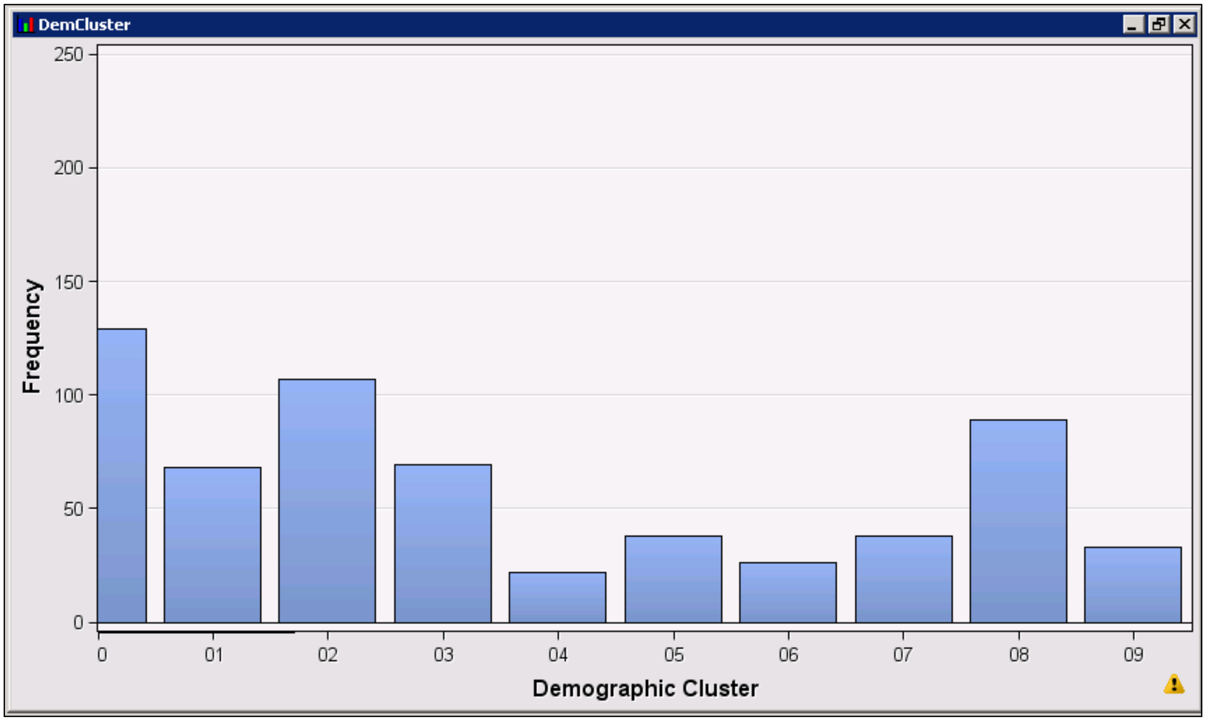


Property	Value
Rows	Unknown
Columns	45
Library	EMWS1
Member	IMPT_TRAIN
Type	VIEW
Sample Method	Random
Fetch Size	Max
Fetched Rows	4843
Random Seed	12345

Obs #	Variable ...	Label	Type	Percent ...	Minim
1	DemCluster	Demograp...	CLASS	0	
2	ID	Control Nu...	CLASS	0	
3	TARGET_B	Target Gift ...	VAR	0	
4	_dataobs_	Observatio...	VAR	0	

Obs #	Observation Number	Target Gift Flag	Control Number	
1	1	0	00014974	0
2	4	1	00185937	0
3	5	0	00029637	3
4	7	0	00123712	3
5	11	1	00102224	1
6	12	1	00187915	3
7	13	0	00182029	1
8	14	0	00188644	1
9	15	1	00164513	2
10	17	0	00190267	3
11	18	0	00035996	3

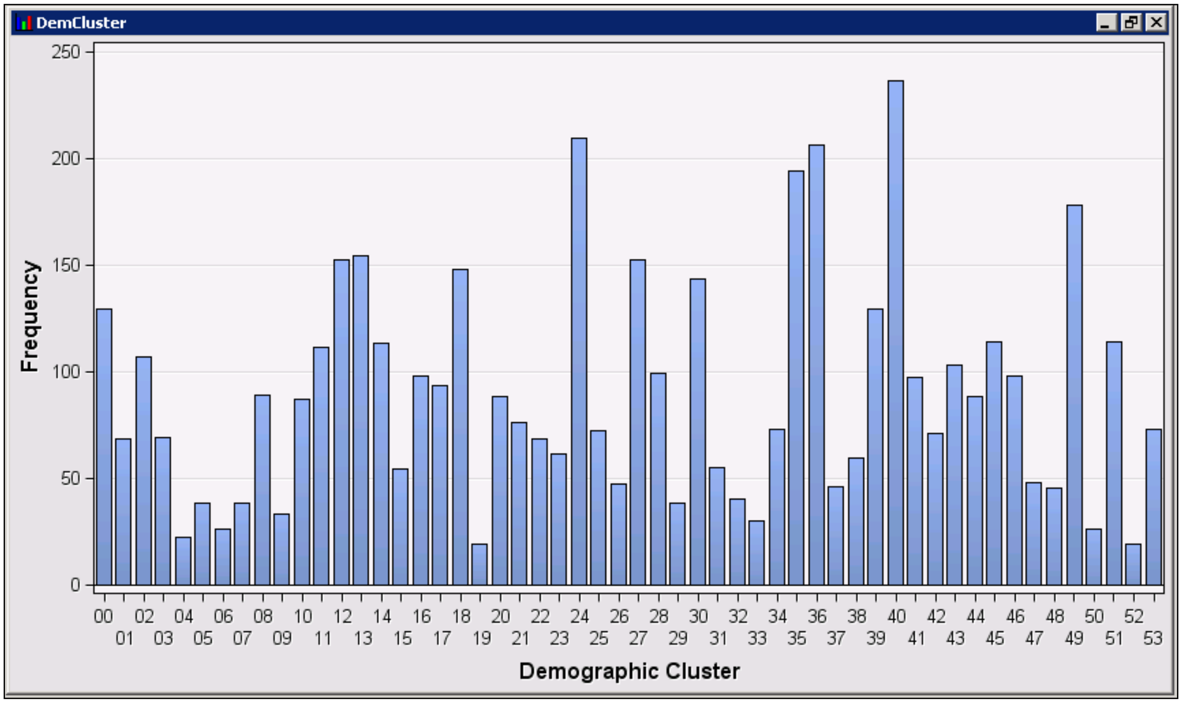
4. Maximize the **DemCluster** histogram.



The **DemCluster** input has more than 50 levels, but you can see the distribution of only a few of these levels.

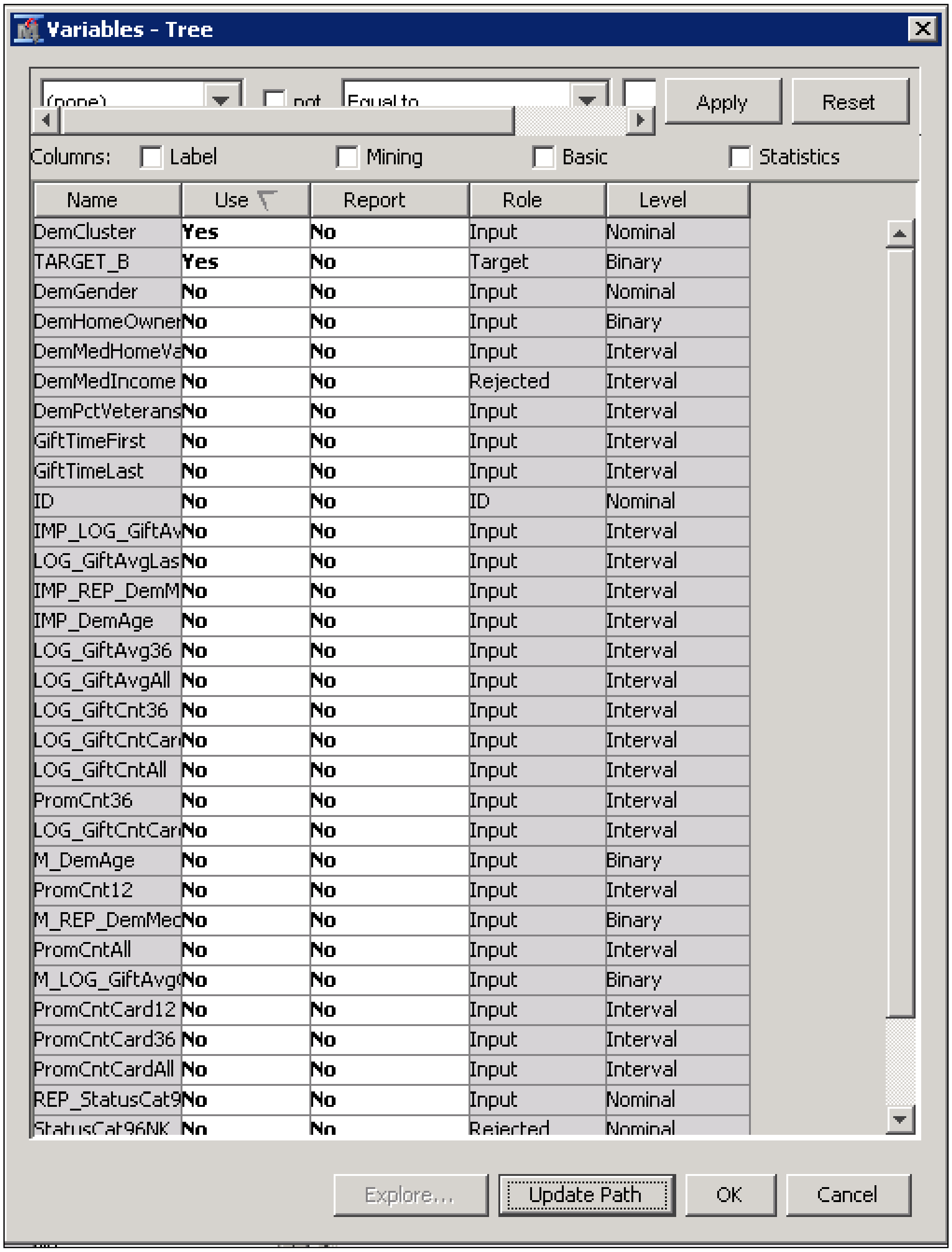
5. Click ⚠ in the lower right corner of the DemCluster window.

   The histogram expands to show the relative frequencies of each level of **DemCluster**.



   The histogram reveals input levels with low case counts. This can detrimentally affect the performance of most models.

6. Close the Explore window.
7. Select **Use** ⇨ **No** for all variables in the Variables window.
8. Select **Use** ⇨ **Yes** for **DemCluster** and **TARGET_B**.



Name	Use	Report	Role	Level
DemCluster	Yes	No	Input	Nominal
TARGET_B	Yes	No	Target	Binary
DemGender	No	No	Input	Nominal
DemHomeOwner	No	No	Input	Binary
DemMedHomeVa	No	No	Input	Interval
DemMedIncome	No	No	Rejected	Interval
DemPctVeterans	No	No	Input	Interval
GiftTimeFirst	No	No	Input	Interval
GiftTimeLast	No	No	Input	Interval
ID	No	No	ID	Nominal
IMP_LOG_GiftAv	No	No	Input	Interval
LOG_GiftAvgLas	No	No	Input	Interval
IMP_REP_DemM	No	No	Input	Interval
IMP_DemAge	No	No	Input	Interval
LOG_GiftAvg36	No	No	Input	Interval
LOG_GiftAvgAll	No	No	Input	Interval
LOG_GiftCnt36	No	No	Input	Interval
LOG_GiftCntCar	No	No	Input	Interval
LOG_GiftCntAll	No	No	Input	Interval
PromCnt36	No	No	Input	Interval
LOG_GiftCntCar	No	No	Input	Interval
M_DemAge	No	No	Input	Binary
PromCnt12	No	No	Input	Interval
M_REP_DemMec	No	No	Input	Binary
PromCntAll	No	No	Input	Interval
M_LOG_GiftAvg	No	No	Input	Binary
PromCntCard12	No	No	Input	Interval
PromCntCard36	No	No	Input	Interval
PromCntCardAll	No	No	Input	Interval
REP_StatusCat9	No	No	Input	Nominal
StatusCat96NK	No	No	Rejected	Nominal

After you sort on the column **Use**, the Variables window should appear as shown above.

9. Select **OK** to close the Variables window.

10. Make these changes in the Train property group.

    a. Under the Subtree section, select **Assessment Measure** ⇨ **Average Squared Error**. This optimizes the tree for prediction estimates.

    b. Under the P-Value Adjustment section, select **Bonferroni Adjustment** ⇨ **No**.

When you evaluate a potential split, the Decision Tree tool applies, by default, a Bonferroni adjustment to the splits logworth. The adjustment penalizes the logworth of potential **DemCluster** splits. The penalty is calculated as the log of the number of partitions of **DemCluster** levels split into two groups, or $log_{10}(2^{L-1} - 1)$. With 54 distinct levels, the penalty is quite large. It is also, in this case, quite unnecessary. The penalty avoids favoring inputs with many possible splits. Here you are building a tree with only one input. It is impossible to favor this input over others because there are no other inputs.

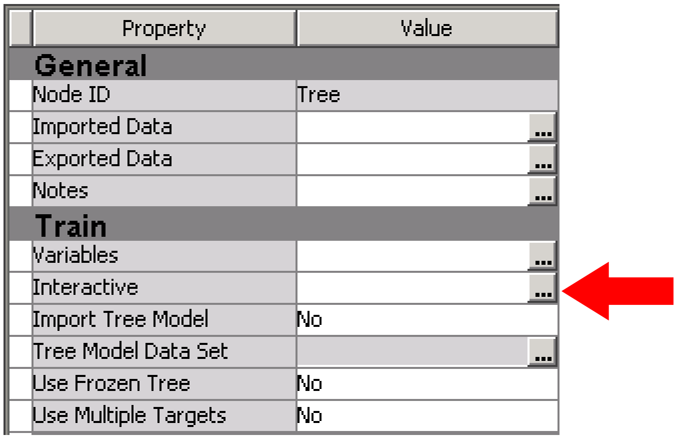
11. Make these changes in the Score property group.

    a. Select **Variable Selection** ⇨ **No**. This prevents the decision tree from rejecting inputs in subsequent nodes.

    b. Select **Leaf Role** ⇨ **Input**. This adds a new input (**_NODE_**) to the training data.

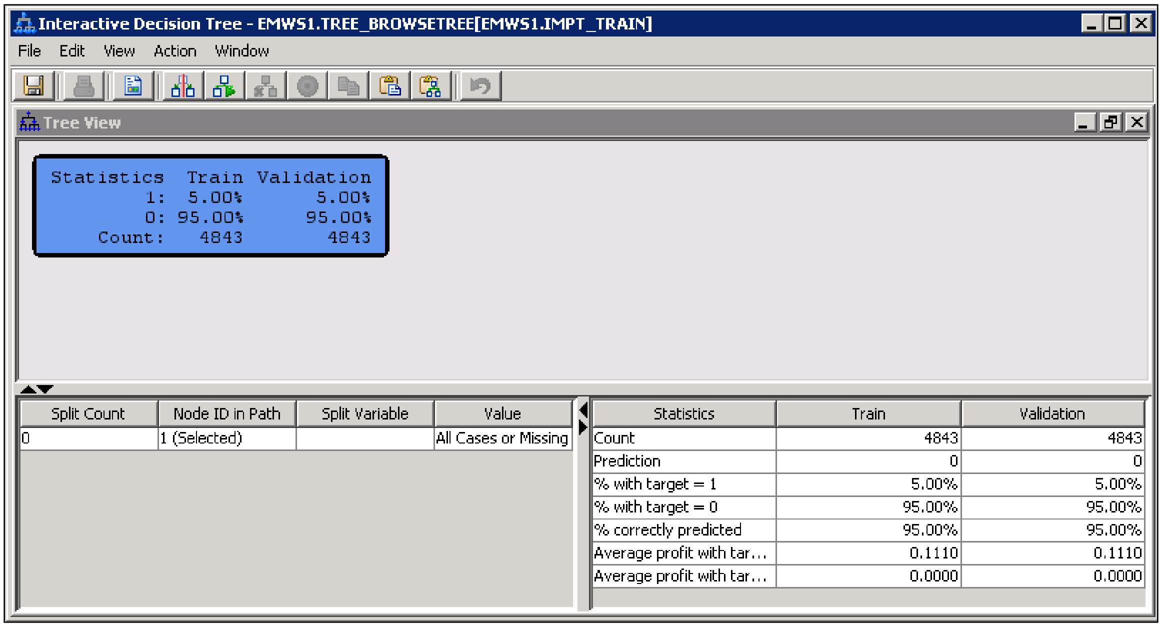
12. Now use the Interactive Tree tool to cluster **DemCluster** values into related groups.

    a. Select **Interactive** ⇨ ... from the Decision Tree Properties panel.



Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No

The SAS Enterprise Miner Tree Desktop Application opens.

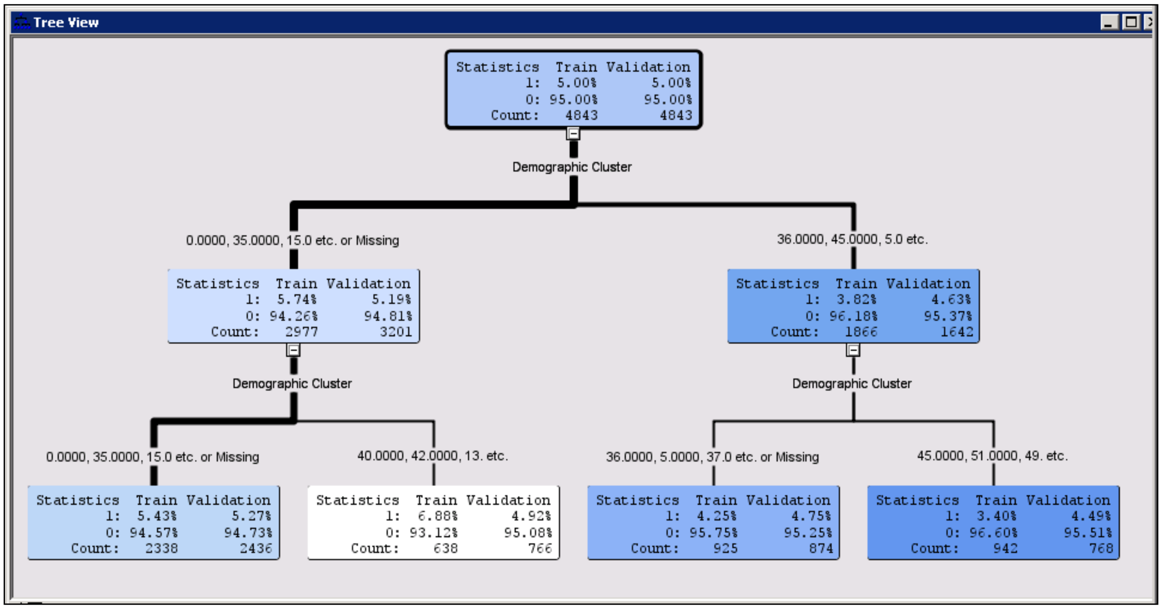

Interactive Decision Tree - EMWS1.TREE_BROWSETREE[EMWS1.IMPT_TRAIN]
File
Edit
View
Action
Window
Tree View

```
Statistics  Train  Validation
        1:  5.00%       5.00%
        0: 95.00%      95.00%
    Count:   4843        4843
```

Split Count	Node ID in Path	Split Variable	Value
0	1 (Selected)		All Cases or Missing

Statistics	Train	Validation
Count	4843	4843
Prediction	0	0
% with target = 1	5.00%	5.00%
% with target = 0	95.00%	95.00%
% correctly predicted	95.00%	95.00%
Average profit with tar…	0.1110	0.1110
Average profit with tar…	0.0000	0.0000

    b. Right-click the root node and select **Train Node** from the option menu.


Tree View
Demographic Cluster
0.0000, 35.0000, 15.0 etc. or Missing
36.0000, 45.0000, 5.0 etc.
Demographic Cluster
Demographic Cluster
0.0000, 35.0000, 15.0 etc. or Missing
40.0000, 42.0000, 13. etc.
36.0000, 5.0000, 37.0 etc. or Missing
45.0000, 51.0000, 49. etc.

```
Statistics  Train  Validation
        1:   5.00%       5.00%
        0:  95.00%      95.00%
    Count:    4843        4843
```

```
Statistics  Train  Validation
        1:   5.74%       5.19%
        0:  94.26%      94.81%
    Count:    2977        3201
```

```
Statistics  Train  Validation
        1:   3.82%       4.63%
        0:  96.18%      95.37%
    Count:    1866        1642
```

```
Statistics  Train  Validation
        1:   5.43%       5.27%
        0:  94.57%      94.73%
    Count:    2338        2436
```

```
Statistics  Train  Validation
        1:   6.88%       4.92%
        0:  93.12%      95.08%
    Count:     638         766
```

```
Statistics  Train  Validation
        1:   4.25%       4.75%
        0:  95.75%      95.25%
    Count:     925         874
```

```
Statistics  Train  Validation
        1:   3.40%       4.49%
        0:  96.60%      95.51%
    Count:     942         768
```

The levels of **DemCluster** are partitioned into four groups corresponding to the four leaves of the tree.

An input named _**NODE**_ is added to the training data. You can use the Transform Variables tool to rename _**NODE**_ to a more descriptive value. You can use the Replacement tool to change the level names.