**MBAD / DSBA 6201 – Business Intelligence & Analytics**

UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE of ENGINEERING

# Project : Predictive Modelling-Decision Trees

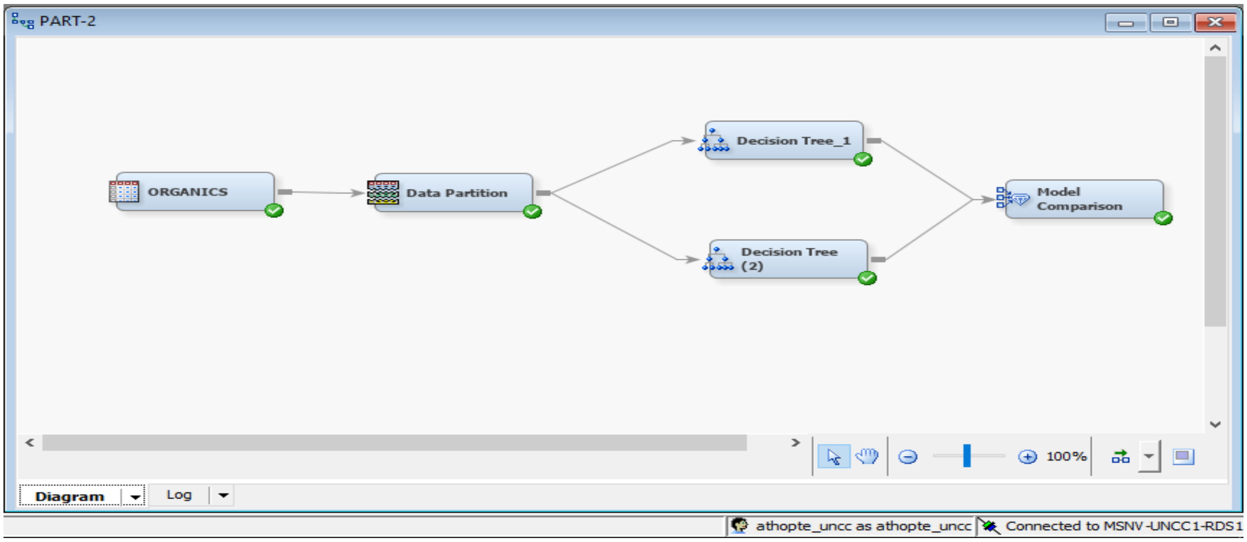
*Project by:-*

**Aniket Amar Thopte**

<u>***TASK: Predictive Modelling Using Decision Trees on Organics Dataset***</u>

**CONSTRUCTING A DECISION TREE PREDICTIVE MODEL:**

Use the ORGANICS data and fit two decision tree models in SAS Enterprise Miner. The diagram below shows the nodes that are needed to fit decision tree models. The steps include splitting the data into training and validation data sets using the Data Partition node, selecting useful inputs using Decision Tree nodes, and generating model assessment statistics and plots using the Model Comparison node.



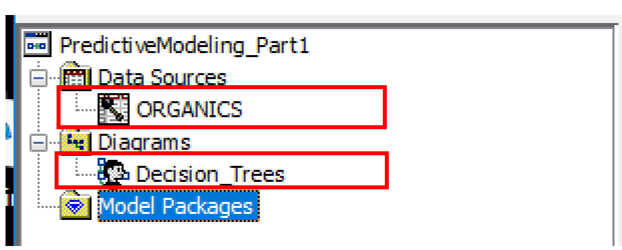

## Initial Data Exploration

A supermarket offers a new line of organic products. The supermarket's management wants to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

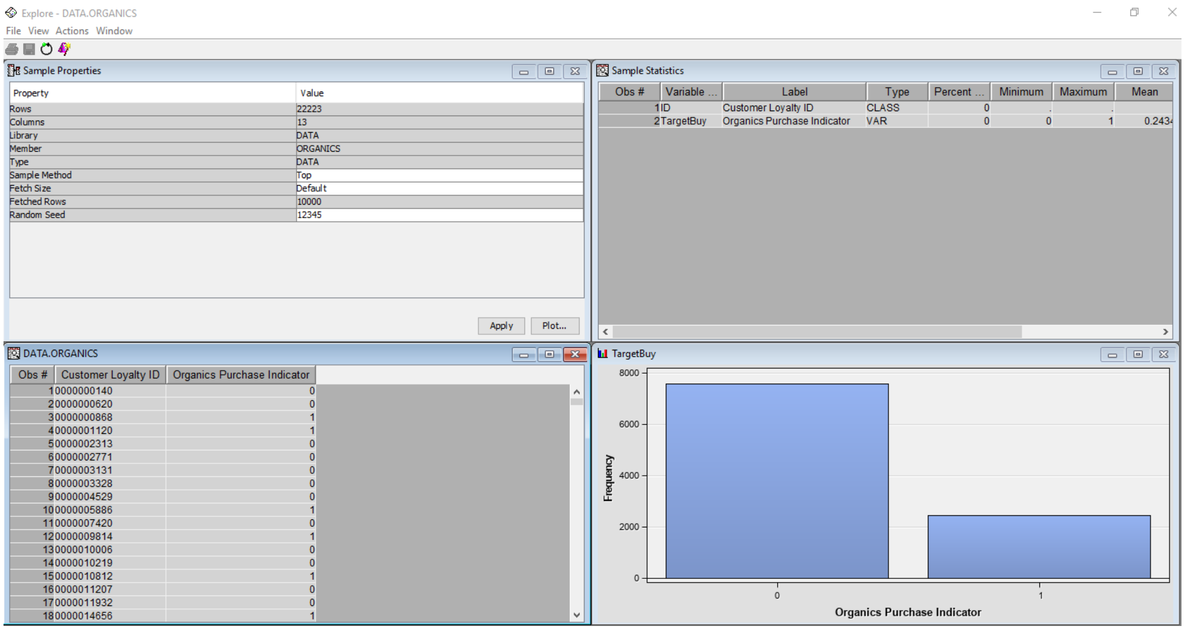
The **ORGANICS** data set contains 13 variables and more than 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels.

a) Create a new diagram named Organics ( In our case Data Source created is named Organics and Diagram created is named ***Decision_Trees*** for our convenience)




b) We have changed the variables for this data analysis . Set the target variables as TargetBuy and reject TargetAmt is rejected other input values are set with corresponding roles for interval and nominal.

To examine the distribution of the target variable we have explored the data. Using the exploration analysis, the proportion of individuals who purchased organic products are as follows:




**The frequency of products bought is 5,505 and the frequency of products not bought is 16,718. The ratio of people who purchased the products is 24.7%.**

The variable DemClusterGroup contains collapsed levels of the variable DemCluster. Presume that, based on previous experience, you believe that DemClusterGroup is sufficient for this type of modeling effort. Set the model role for DemCluster to Rejected.

The data clearly gives information that that variable TargetBuy is a binary variable and takes values 1 or 0 which indicates if the product was bought or not. TargetAmt gives information on the total amount of organic products which have been sold and is a dependent variable. Hence, we cannot use this to predict the target.

c) We have added the ORGANICS data source to the Organics diagram workspace.

d) Partition the data set into training and validation data sets.

Use the Properties panel to select the fraction of data devoted to the training, validation, and test partitions. By default, less than half the available data is used for generating the predictive models.

We have assigned 65% of the data for training and 35% for validation.

| Data Set Allocations | |
|---|---|
| Training | 65.0 |
| Validation | 35.0 |
| Test | 0.0 |

e) We have added a Decision Tree node to the workspace and connected it to the Data Partition node.

```
*------------------------------------------------------------
*
User:                    athopte_uncc
Date:                    December 17, 2020
Time:                    18:34:19
*------------------------------------------------------------
*
* Training Output
*------------------------------------------------------------
*




Variable Summary

```

```
Variable Summary

              Measurement    Frequency
Role             Level          Count

ID            NOMINAL             1
INPUT         INTERVAL            4
INPUT         NOMINAL             5
REJECTED      INTERVAL            1
REJECTED      NOMINAL             1
TARGET        BINARY              1




Partition Summary

                                        Number of
Type              Data Set            Observations

DATA        EMWS5.Ids_DATA               22223
TRAIN       EMWS5.Part_TRAIN             14445
```

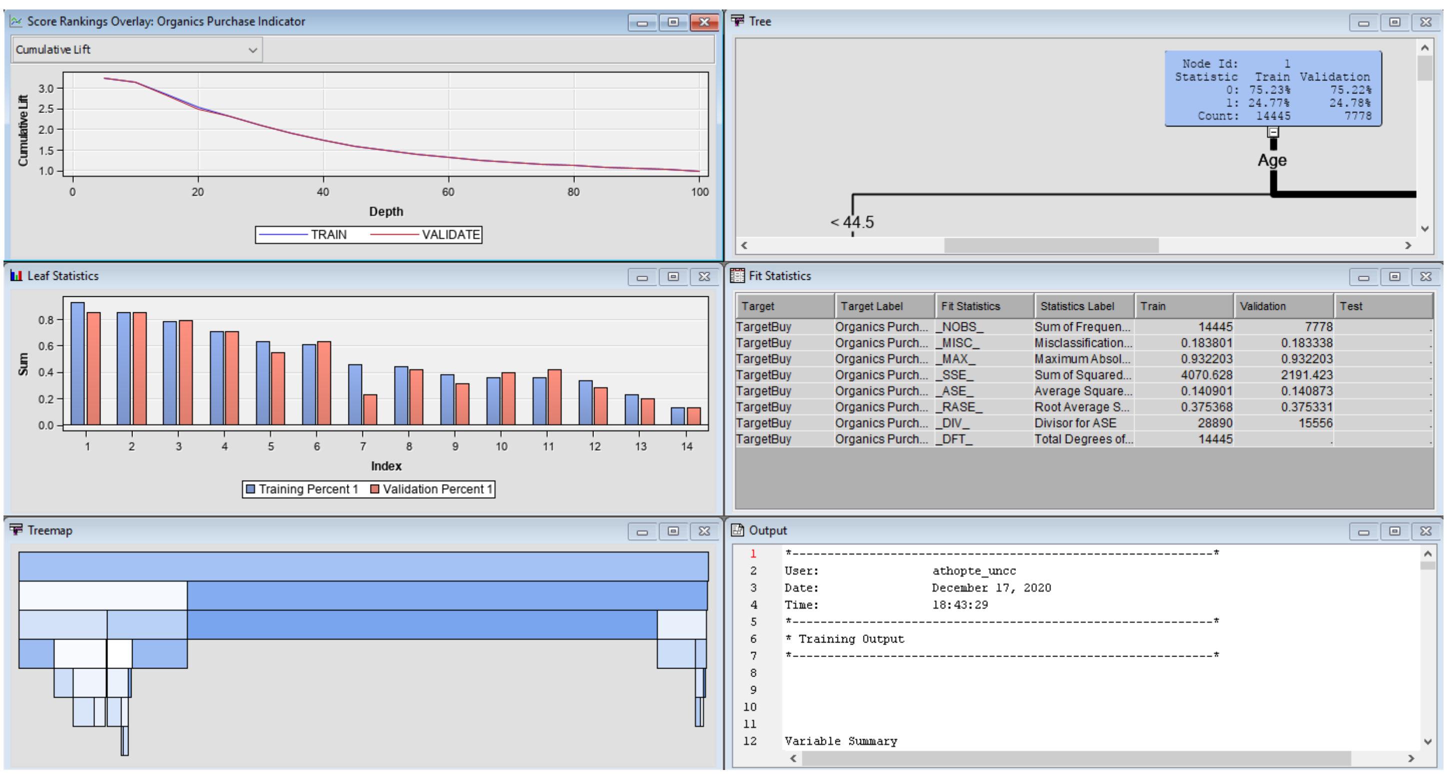
```
VALIDATE    EMWS5.Part_VALIDATE        7778


*------------------------------------------------------------
*
* Score Output
*------------------------------------------------------------
*


*------------------------------------------------------------
*
* Report Output
*------------------------------------------------------------
*




Summary Statistics for Class Targets

```

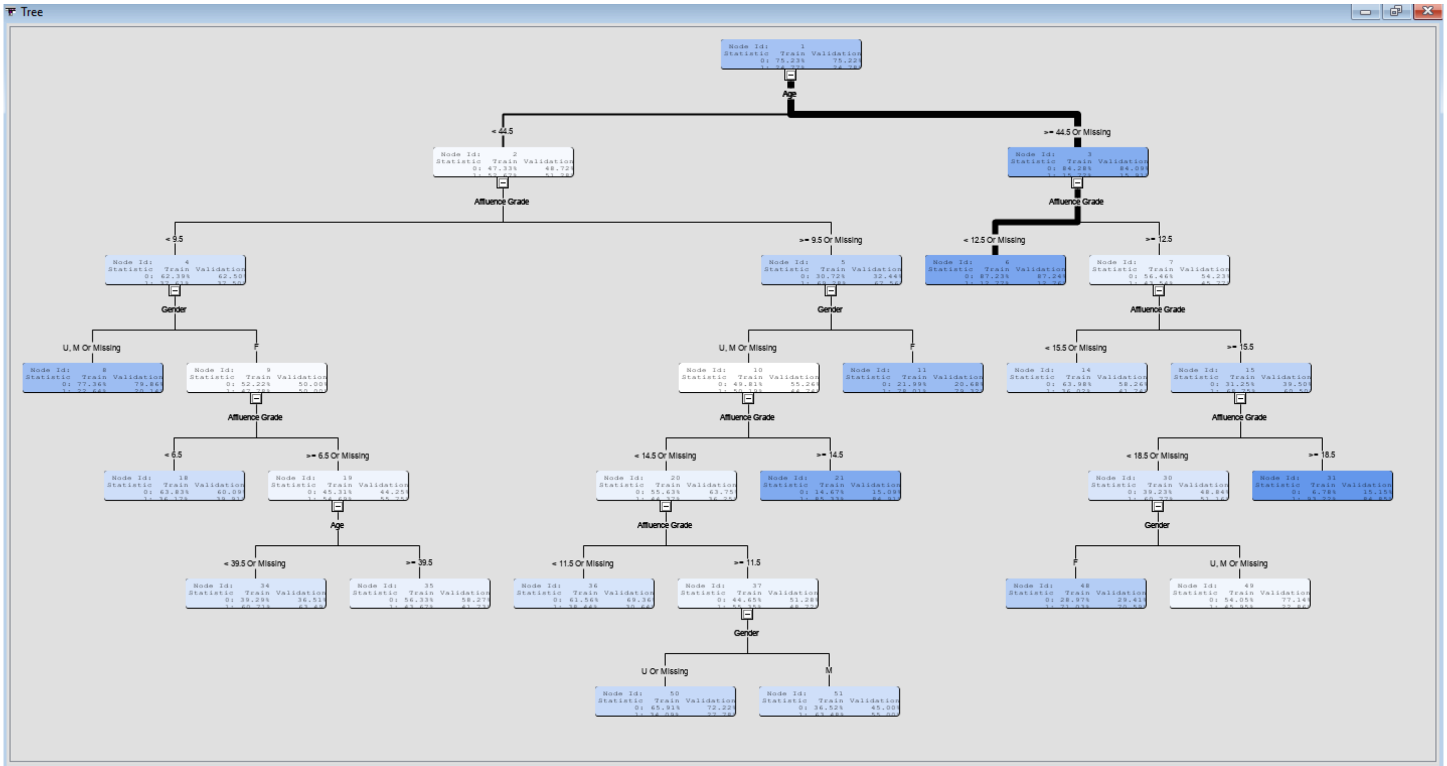
```
Data=DATA

             Numeric    Formatted    Frequency
Variable      Value       Value        Count       Percent
           Label

TargetBuy       0           0          16718       75.2284
Organics Purchase Indicator
TargetBuy       1           1           5505       24.7716
Organics Purchase Indicator


Data=TRAIN

             Numeric    Formatted    Frequency
```

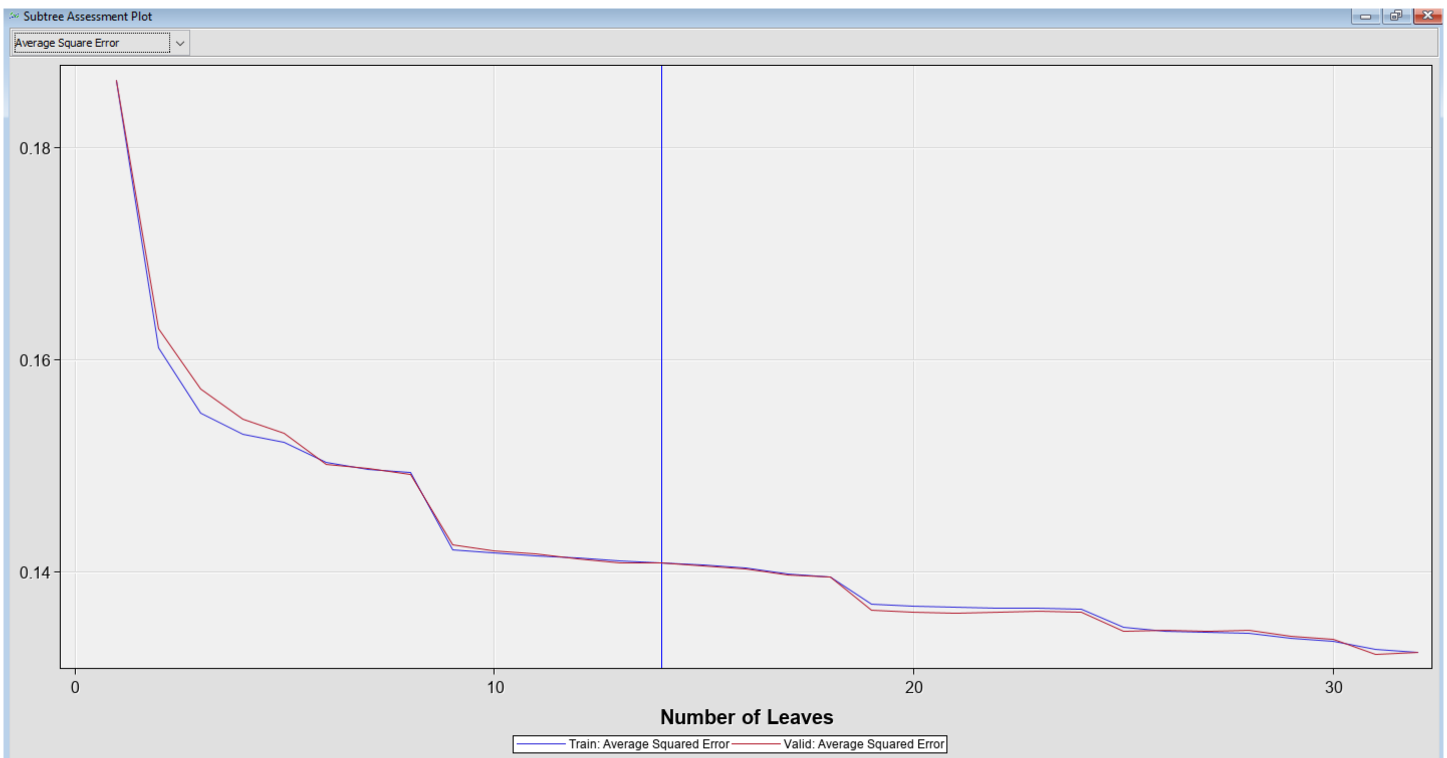
```
Variable       Value        Value        Count       Percent
          Label

TargetBuy        0            0          10867       75.2302
    Organics Purchase Indicator
TargetBuy        1            1           3578       24.7698
    Organics Purchase Indicator


Data=VALIDATE

               Numeric     Formatted    Frequency
Variable       Value        Value        Count       Percent
          Label

TargetBuy        0            0           5851       75.2250
    Organics Purchase Indicator
TargetBuy        1            1           1927       24.7750
    Organics Purchase Indicator
```

f) We create our first decision tree and add it to the node partition.

Output:



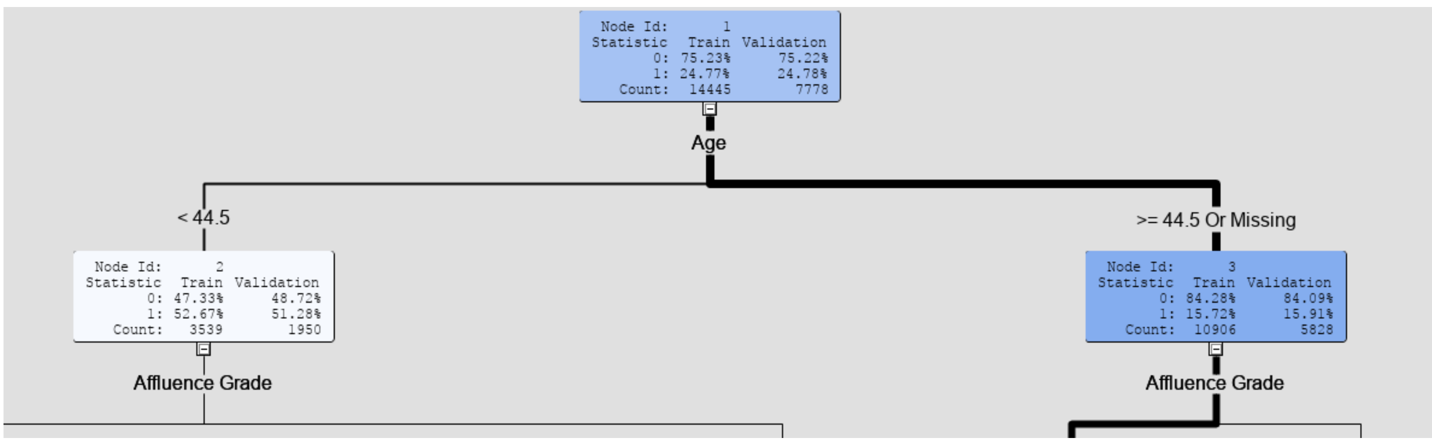


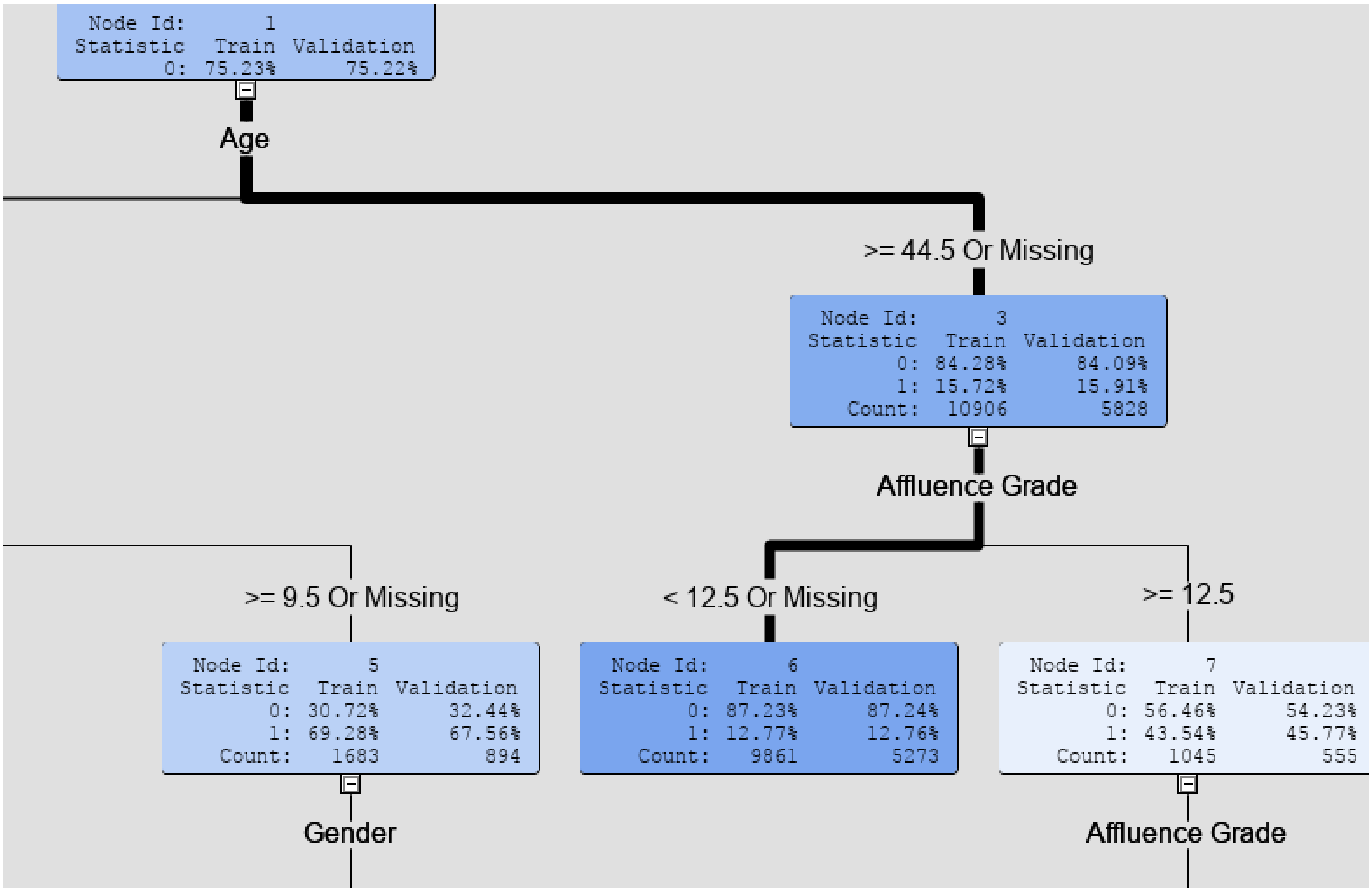

**Optimal tree has 16 number of leaves.**

Variable Importance

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




**Variables Age was used for the first split.**



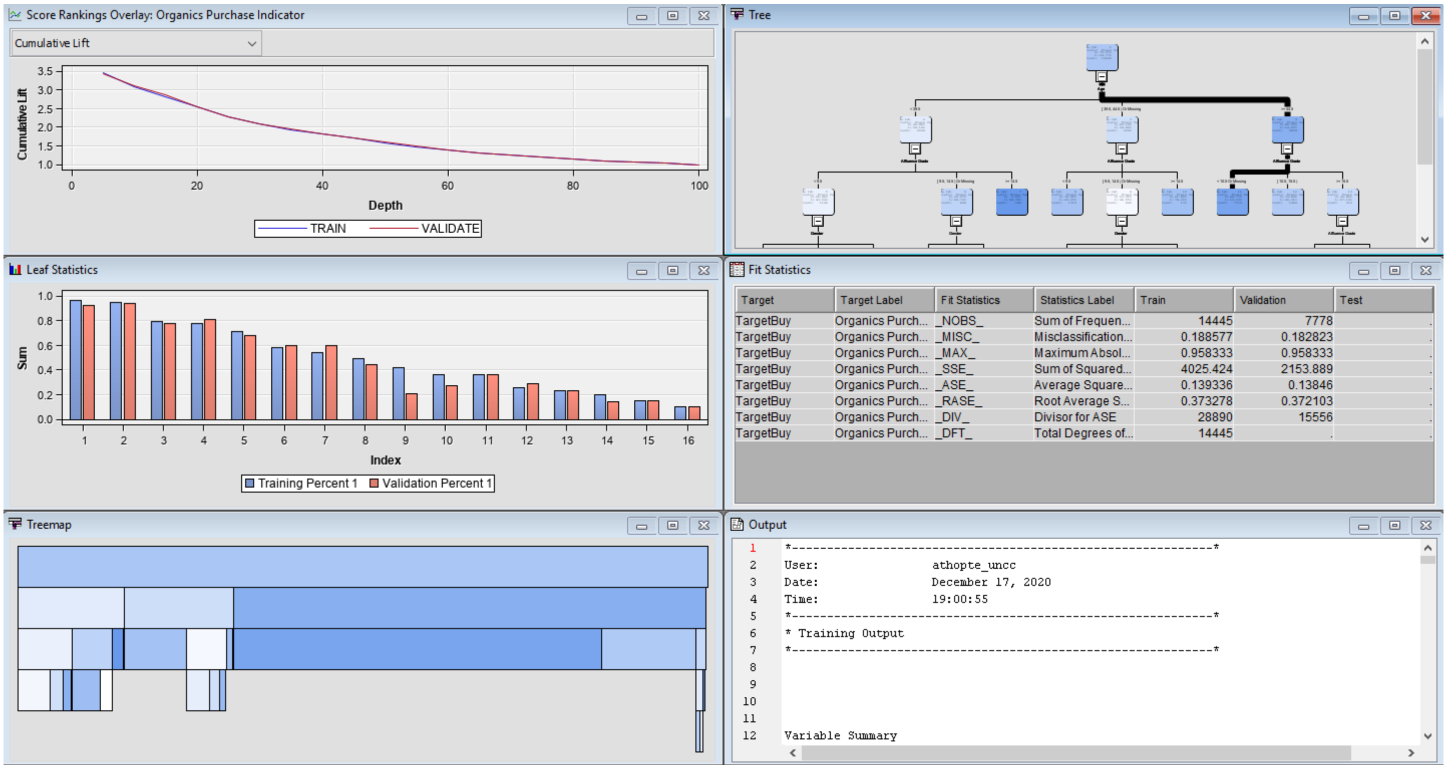

**Variable Gender & Affluence Grade was used for second split.**

The decision tree gives us an optimal result hence we can conclude that the model is good.

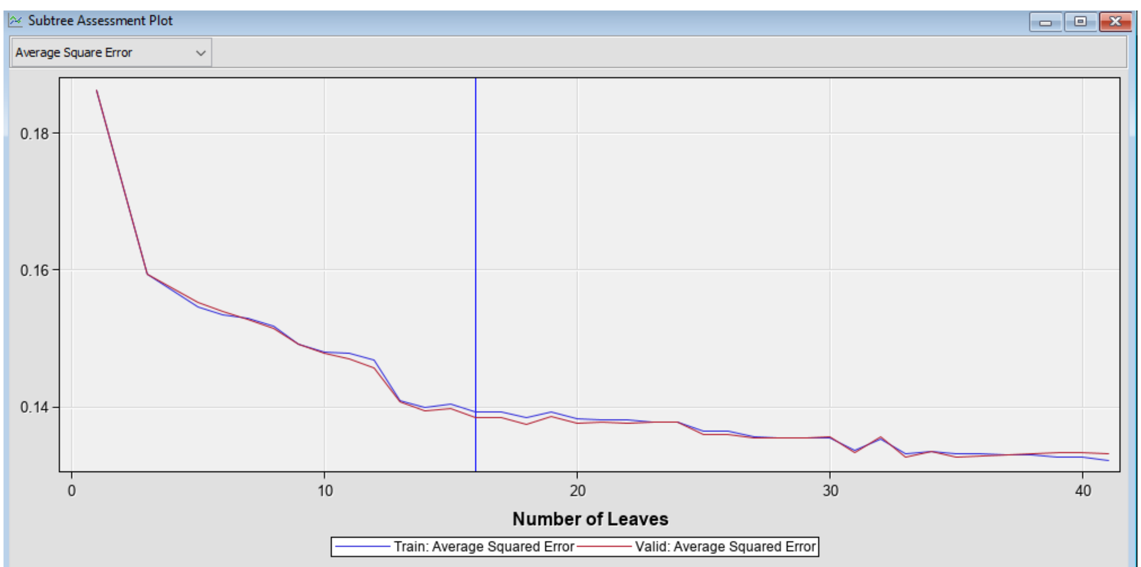
g) Add a second Decision Tree node to the diagram and connect it to the Data Partition node.

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

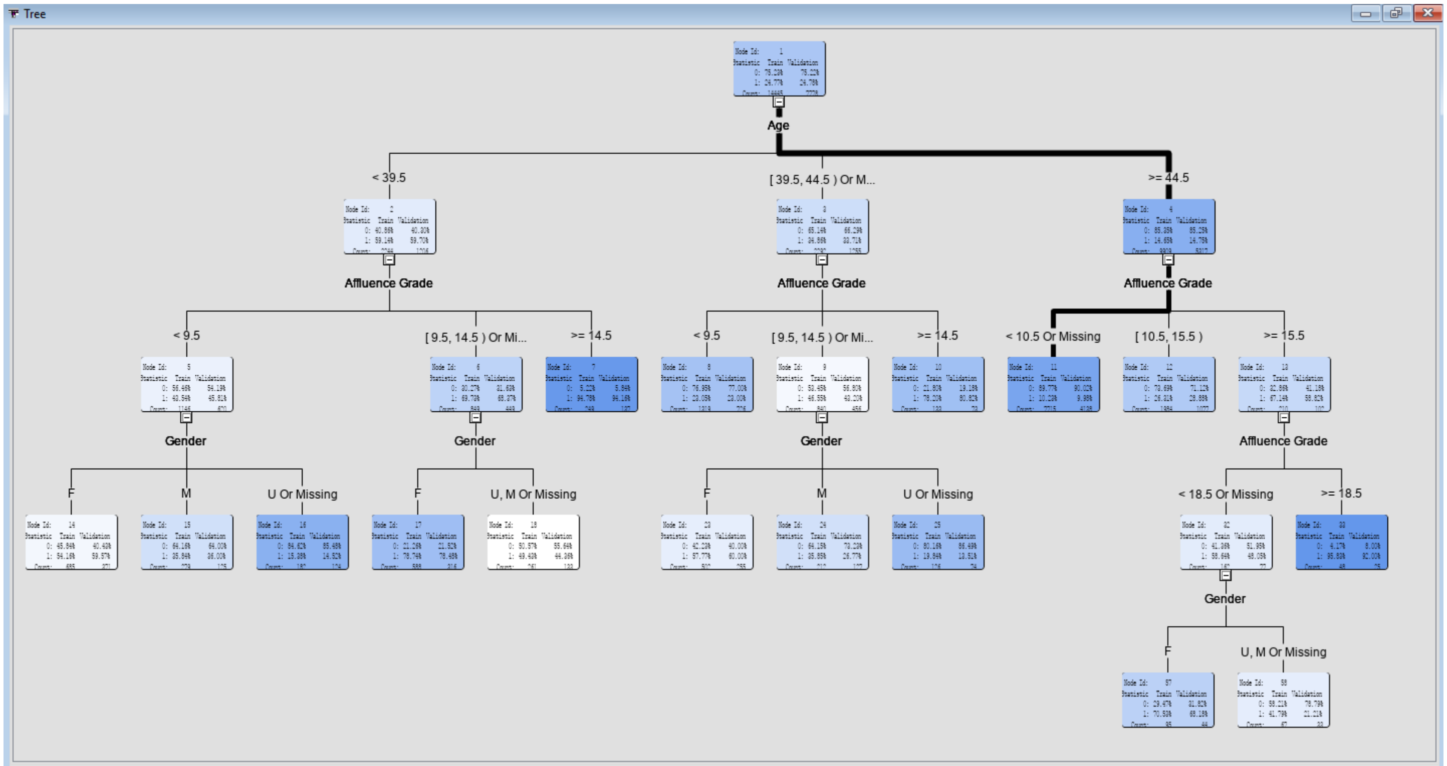
Output:

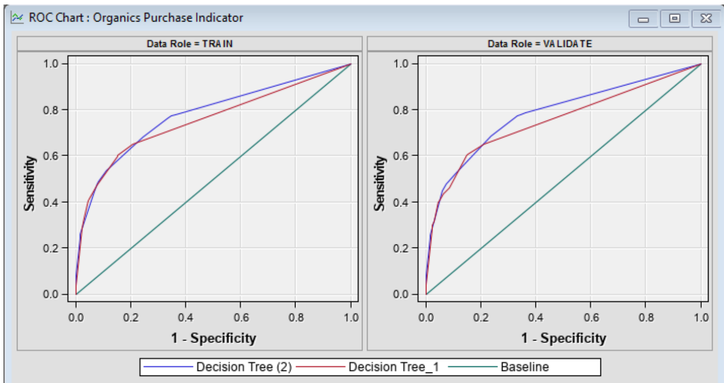



| Splitting Rule | |
|---|---|
| Interval Target Criterion | ProbF |
| Nominal Target Criterion | ProbChisq |
| Ordinal Target Criterion | Entropy |
| Significance Level | 0.2 |
| Missing Values | Use in search |
| Use Input Once | No |
| Maximum Branch | 3 |
| Maximum Depth | 6 |
| Minimum Categorical Size | 5 |




**Optimal tree has 14 number of leaves.**

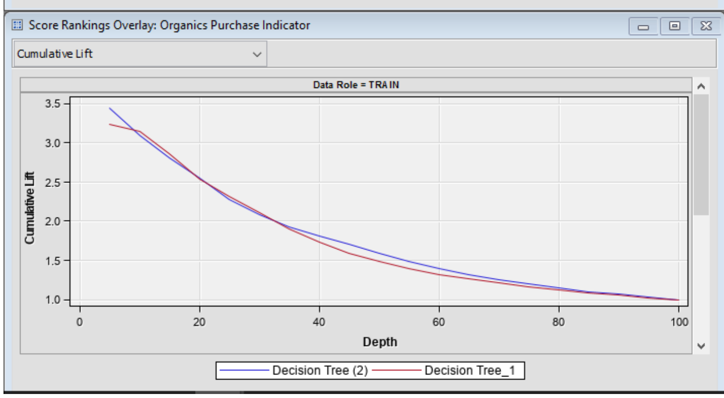


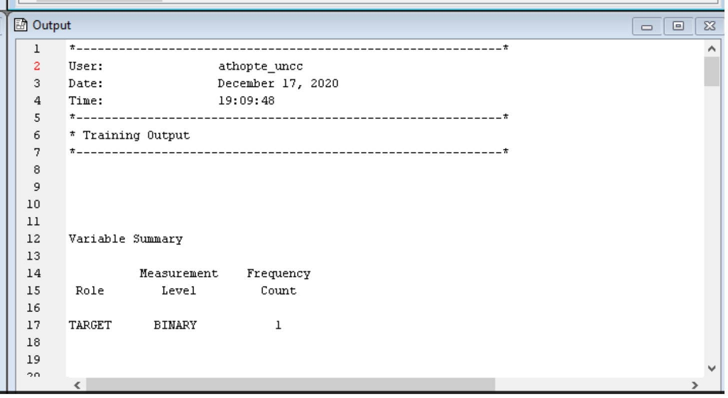





Fit Statistics

| Selected Model | Predecessor Node | Model Node | Model Description | Target Variable | Target Label | Selection Criterion: Valid: Misclassification Rate | Train: Sum of Frequencies | Train: Misclassification Rate |
|---|---|---|---|---|---|---|---|---|
| Y | Tree2 | Tree2 | Decision Tr... | TargetBuy | Organics P... | 0.182823 | 14445 | 0.188577 |
| | Tree | Tree | Decision Tr... | TargetBuy | Organics P... | 0.183338 | 14445 | 0.183801 |





Output

```
*------------------------------------------------------------*
User:                 athopte_uncc
Date:                 December 17, 2020
Time:                 19:09:48
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement    Frequency
 Role         Level         Count

TARGET      BINARY           1
```

From the above output, by observing the ROC Curve we can say that the Decision Tree-2 performs slightly better.