SAS Assignment #3

Prepared By: Group #5

Joel Camacho

Mohamed Mazarul

Bradley John

The purpose of this project was to continue the knowledge enhancement of SAS Enterprise Miner. The project required us to apply decision tree nodes, partition the data and use the average square error as the model assessment statistic to answer a business problem for a supermarket with a new line of organic products. The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all the loyalty program participants and collected data that includes whether these customers purchased any of the organic products. Detail of these attributes is as follows.

Business Problem

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

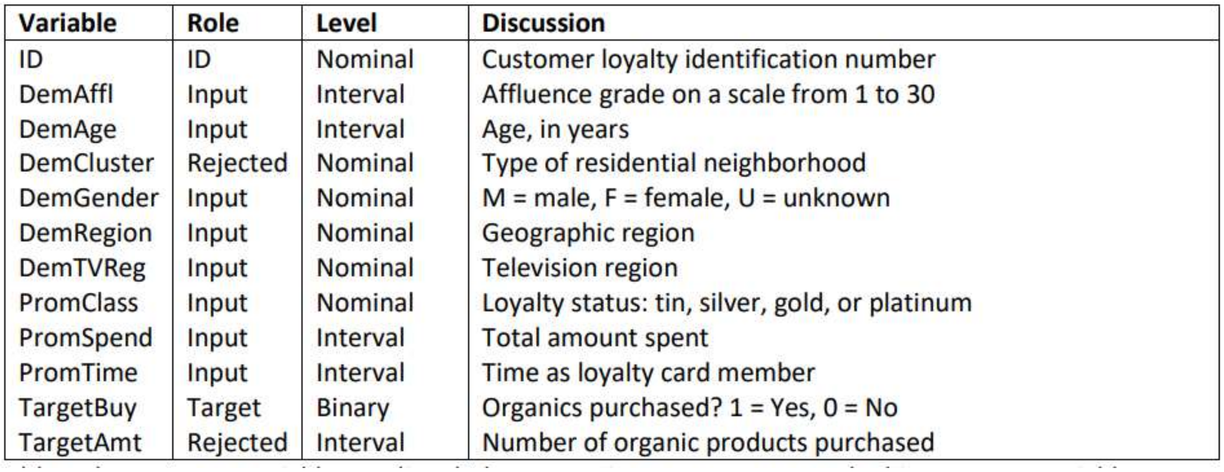
Questions & Building Model

1. How do the purchase orders vary by gender?
2. Does age play a role in purchasing orders?

Data description

The ORGANICS data set contains:

- 13 variables and
- over 22,000 observations

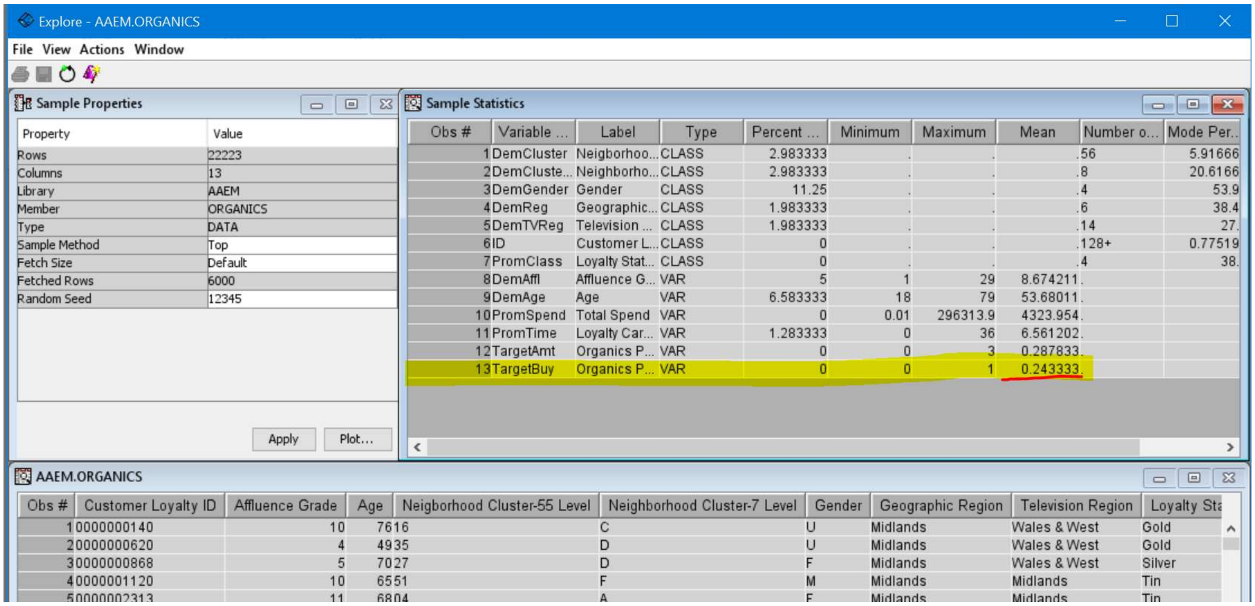


Variable	Role	Level	Discussion
ID	ID	Nominal	Customer loyalty identification number
DemAffl	Input	Interval	Affluence grade on a scale from 1 to 30
DemAge	Input	Interval	Age, in years
DemCluster	Rejected	Nominal	Type of residential neighborhood
DemGender	Input	Nominal	M = male, F = female, U = unknown
DemRegion	Input	Nominal	Geographic region
DemTVReg	Input	Nominal	Television region
PromClass	Input	Nominal	Loyalty status: tin, silver, gold, or platinum
PromSpend	Input	Interval	Total amount spent
PromTime	Input	Interval	Time as loyalty card member
TargetBuy	Target	Binary	Organics purchased? 1 = Yes, 0 = No
TargetAmt	Rejected	Interval	Number of organic products purchased

Our first step was to import the data in EM using all the variables and set the roles as listed above then applied the required parameters to answer the business problem. The following steps were completed (see Appendices for screenshots)

1. Decision Tree nodes were created
2. A Data Partition node was added by assigning 50% of the data for training and 50% for validation

Data Exploration



Property	Value
Rows	22223
Columns	13
Library	AAEM
Member	ORGANICS
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	6000
Random Seed	12345

Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...
1	DemCluster	Neigborhoo...	CLASS	2.983333	.	.	.	56	5.91666
2	DemCluste...	Neighborho...	CLASS	2.983333	.	.	.	8	20.6166
3	DemGender	Gender	CLASS	11.25	.	.	.	4	53.9
4	DemReg	Geographic...	CLASS	1.983333	.	.	.	6	38.4
5	DemTVReg	Television ...	CLASS	1.983333	.	.	.	14	27.
6	ID	Customer L...	CLASS	0	.	.	.	128+	0.77519
7	PromClass	Loyalty Stat...	CLASS	0	.	.	.	4	38.
8	DemAffl	Affluence G...	VAR	5	1	29	8.674211	.	
9	DemAge	Age	VAR	6.583333	18	79	53.68011	.	
10	PromSpend	Total Spend	VAR	0	0.01	296313.9	4323.954	.	
11	PromTime	Loyalty Car...	VAR	1.283333	0	36	6.561202	.	
12	TargetAmt	Organics P...	VAR	0	0	3	0.287833	.	
13	TargetBuy	Organics P...	VAR	0	0	1	0.243333	.	

Obs #	Customer Loyalty ID	Affluence Grade	Age	Neigborhood Cluster-55 Level	Neighborhood Cluster-7 Level	Gender	Geographic Region	Television Region	Loyalty Sta
1	0000000140	10	76	16	C	U	Midlands	Wales & West	Gold
2	0000000620	4	49	35	D	U	Midlands	Wales & West	Gold
3	0000000868	5	70	27	D	F	Midlands	Wales & West	Silver
4	0000001120	10	65	51	F	M	Midlands	Midlands	Tin
5	0000002313	11	68	04	A	F	Midlands	Midlands	Tin

The proportion of individuals who purchased organic products appears to be 24.3%.

Implementations and Findings

We created a decision tree model. We used average square error as the model assessment statistic. We also used Subtree Assessment Plot to assess the model.

- Using average square error as the assessment measure results in a tree with 29 leaves in Model 1
- Age is used for the first split
- Competing splits are Affluence Grade and Gender

For comparison purposes we created a second model that allowed an additional split at each node. In this model

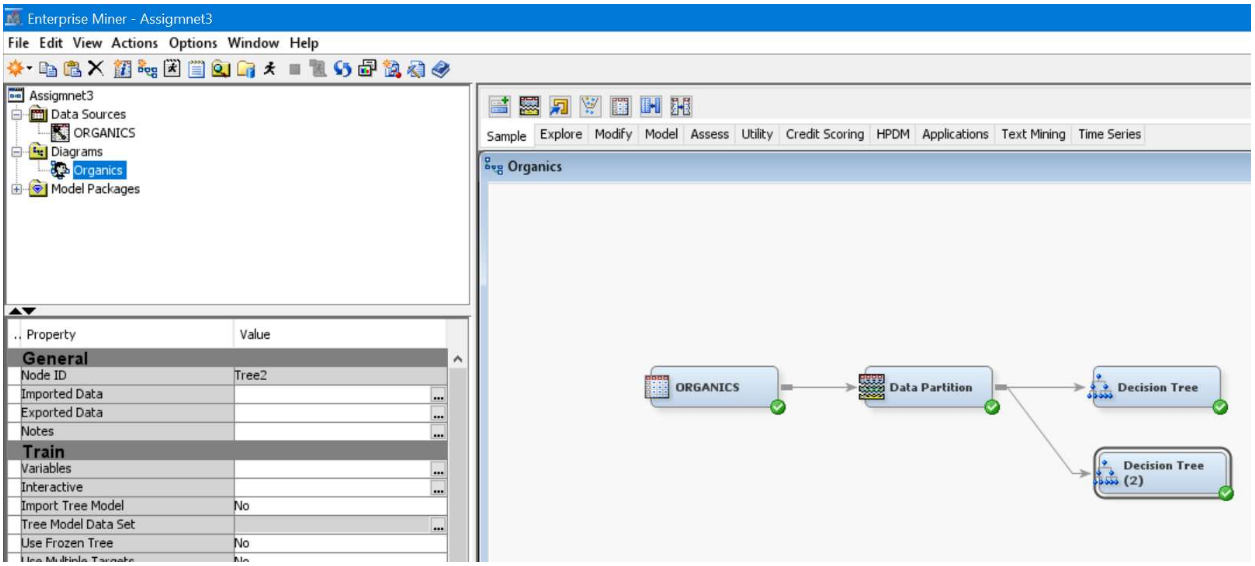
- Using average square error as the assessment measure results in a tree with 32 leaves in Model 2
- Age is used for the first split
- The competing splits is Affluence Grade

Conclusion

Our data showed us that the solution to acquiring more customer purchases is to focus on two areas of the business. The first is to focus on women since the data shows they have a higher purchase rate. The second is to market towards middle aged customers. These customers are the top buyers and those are the ones that need to be targeted. Each of the models gave a similar final answer and could be used. We performed a model comparison (Appendix D) and the misclassification rates were nearly the same for each. Each model also performed similarly for a variety of other measures.

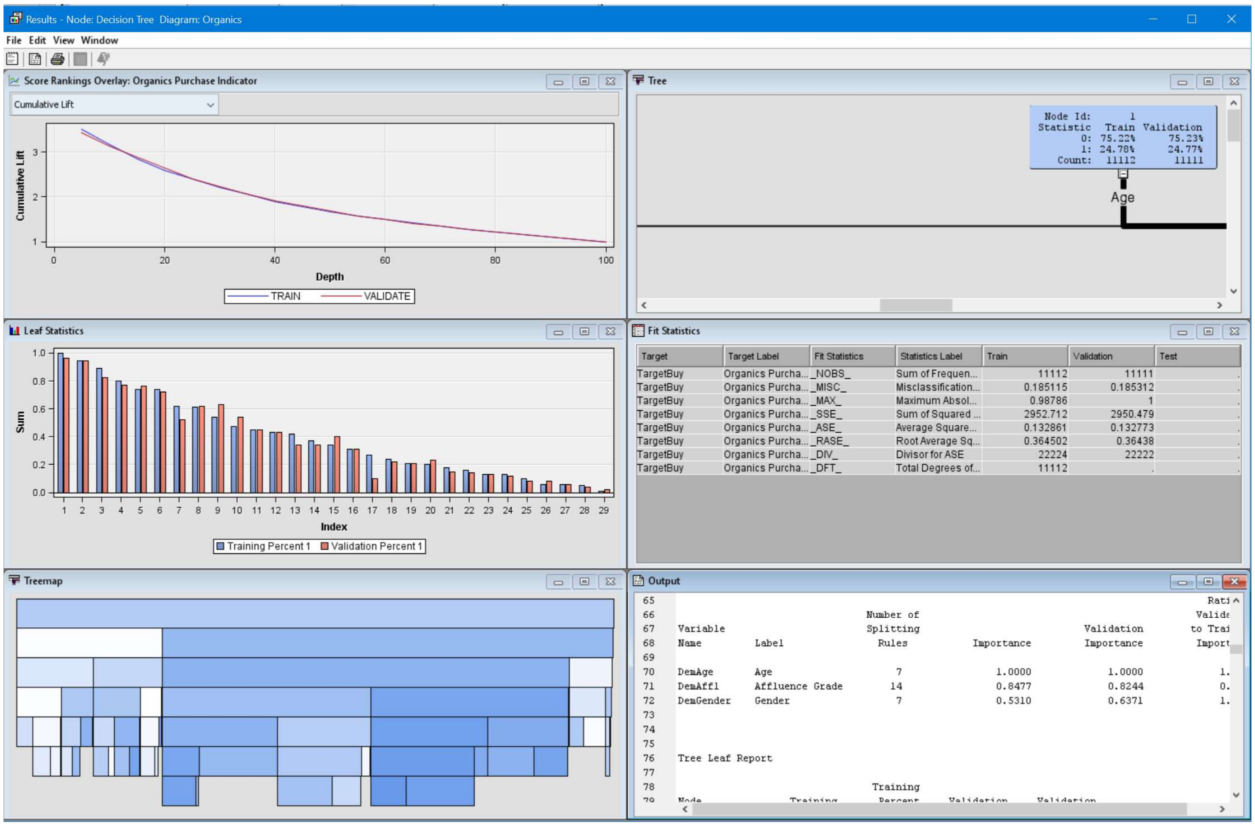
Appendices:

A: Overall SaS Diagram



Property	Value
General	
Node ID	Tree2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No

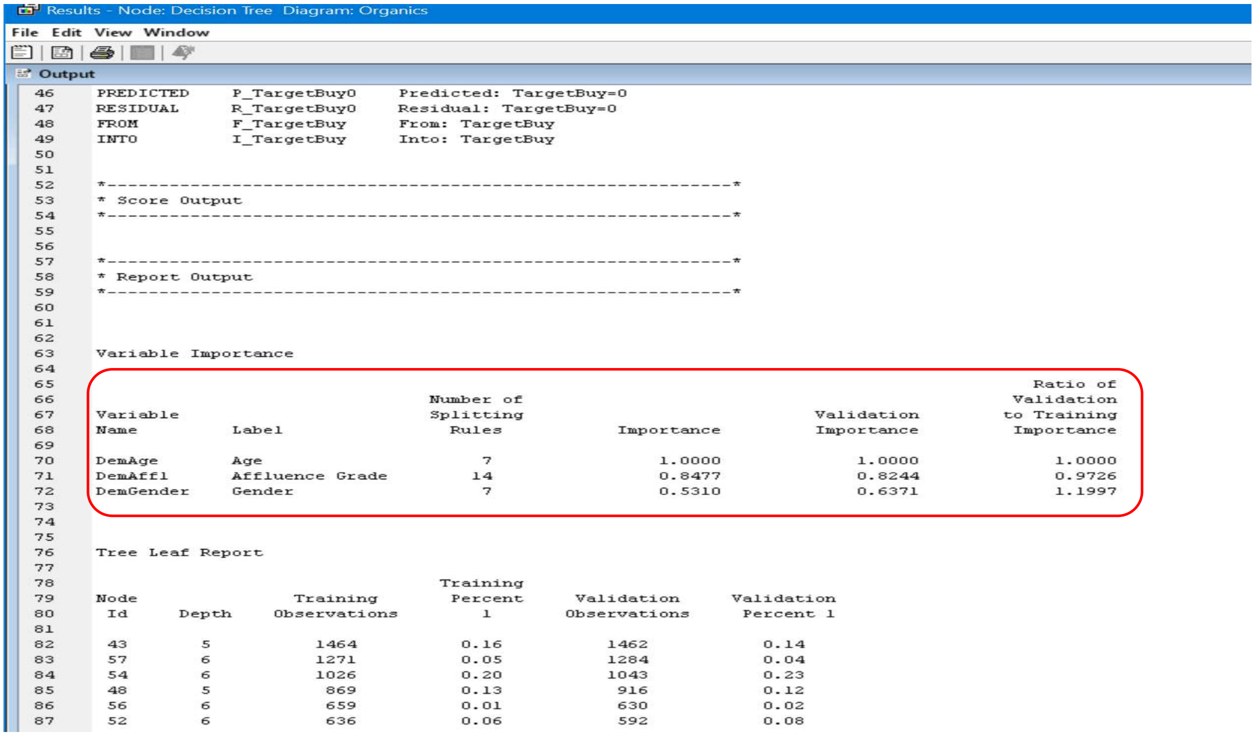
B. Decision Tree1 Results:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetBuy	Organics Purcha...	_NOBS_	Sum of Frequen...	11112	11111	.
TargetBuy	Organics Purcha...	_MISC_	Misclassification...	0.185115	0.185312	.
TargetBuy	Organics Purcha...	_MAX_	Maximum Absol...	0.98786	1	.
TargetBuy	Organics Purcha...	_SSE_	Sum of Squared ...	2952.712	2950.479	.
TargetBuy	Organics Purcha...	_ASE_	Average Square...	0.132861	0.132773	.
TargetBuy	Organics Purcha...	_RASE_	Root Average Sq...	0.364502	0.36438	.
TargetBuy	Organics Purcha...	_DIV_	Divisor for ASE	22224	22222	.
TargetBuy	Organics Purcha...	_DFT_	Total Degrees of...	11112	.	.

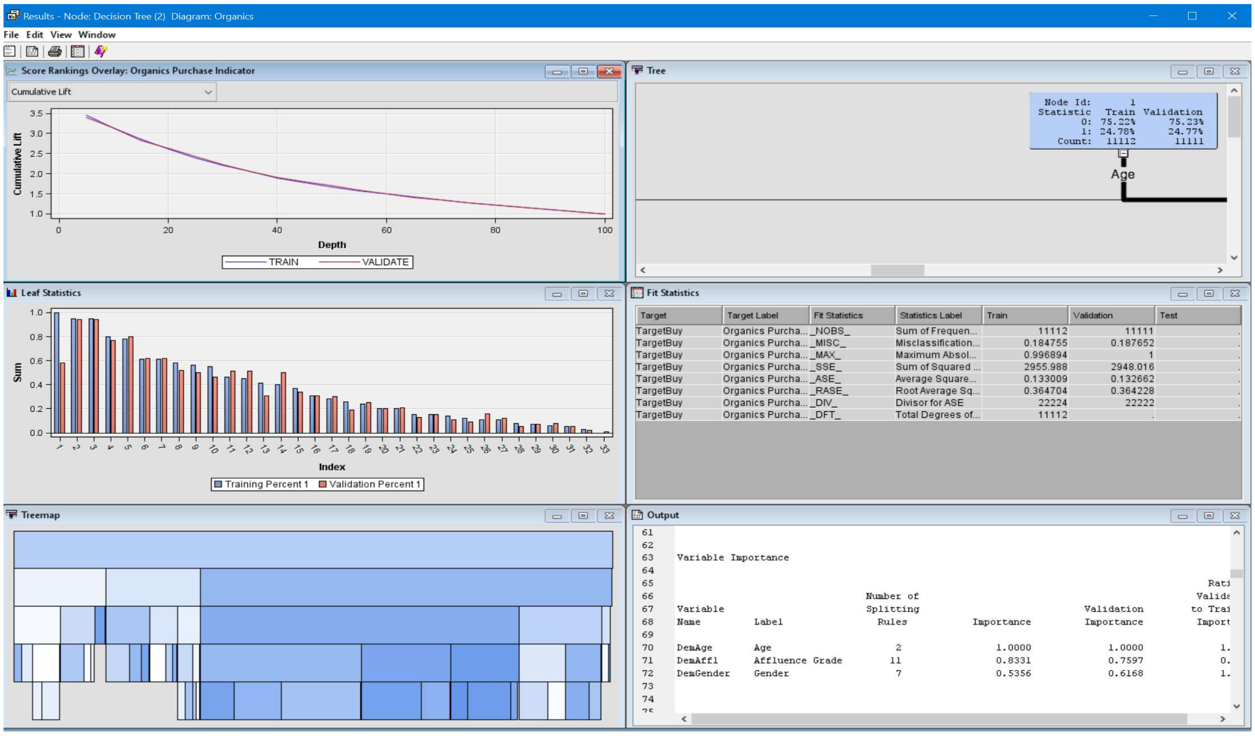
```
                                                                                   Rati
                                     Number of                                     Valida
Variable                             Splitting                      Validation     to Trai
Name         Label                   Rules         Importance       Importance     Import

DemAge       Age                       7             1.0000           1.0000         1.
DemAffl      Affluence Grade          14             0.8477           0.8244         0.
DemGender    Gender                    7             0.5310           0.6371         1.



Tree Leaf Report

                                      Training
Node                 Training         Percent      Validation     Validation
```

B.1 – Output Only



```
PREDICTED      P_TargetBuy0     Predicted: TargetBuy=0
RESIDUAL       R_TargetBuy0     Residual: TargetBuy=0
FROM           F_TargetBuy      From: TargetBuy
INTO           I_TargetBuy      Into: TargetBuy


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


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



Variable Importance

                                                                                   Ratio of
                                Number of                                         Validation
Variable                        Splitting                          Validation     to Training
Name          Label               Rules          Importance        Importance     Importance

DemAge        Age                   7              1.0000            1.0000         1.0000
DemAffl       Affluence Grade      14              0.8477            0.8244         0.9726
DemGender     Gender                7              0.5310            0.6371         1.1997



Tree Leaf Report

                                        Training
Node                    Training         Percent       Validation       Validation
 Id       Depth       Observations          1          Observations     Percent 1

 43         5             1464            0.16           1462             0.14
 57         6             1271            0.05           1284             0.04
 54         6             1026            0.20           1043             0.23
 48         5              869            0.13            916             0.12
 56         6              659            0.01            630             0.02
 52         6              636            0.06            592             0.08
```

C: Decision Tree 2

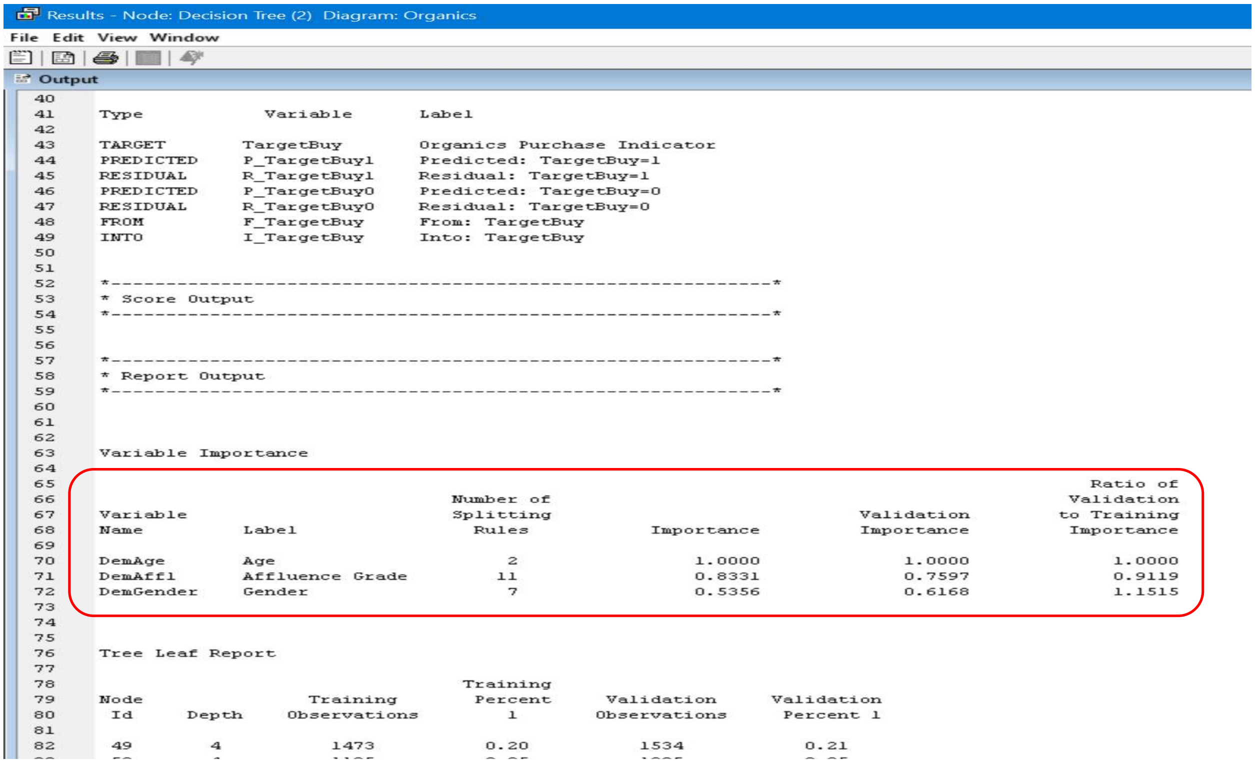


```
Node Id:     1
Statistic   Train  Validation
   0:  75.22%      75.23%
   1:  24.78%      24.77%
Count:  11112       11111
```

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetBuy	Organics Purcha...	_NOBS_	Sum of Frequen...	11112	11111	.
TargetBuy	Organics Purcha...	_MISC_	Misclassification...	0.184755	0.187652	.
TargetBuy	Organics Purcha...	_MAX_	Maximum Absol...	0.996894	1	.
TargetBuy	Organics Purcha...	_SSE_	Sum of Squared ...	2955.988	2948.016	.
TargetBuy	Organics Purcha...	_ASE_	Average Square...	0.133009	0.132662	.
TargetBuy	Organics Purcha...	_RASE_	Root Average Sq...	0.364704	0.364228	.
TargetBuy	Organics Purcha...	_DIV_	Divisor for ASE	22224	22222	.
TargetBuy	Organics Purcha...	_DFT_	Total Degrees of...	11112	.	.

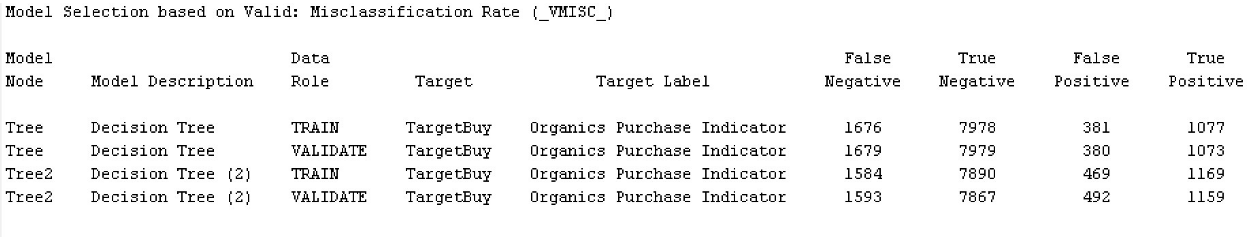
```


Variable Importance

                                                                           Rati
                               Number of                                  Valid
Variable                       Splitting                    Validation    to Trai
Name          Label              Rules         Importance    Importance    Import

DemAge        Age                  2             1.0000        1.0000        1.
DemAffl       Affluence Grade     11             0.8331        0.7597        0.
DemGender     Gender               7             0.5356        0.6168        1.


```

C.1 – Model 2 Output Only

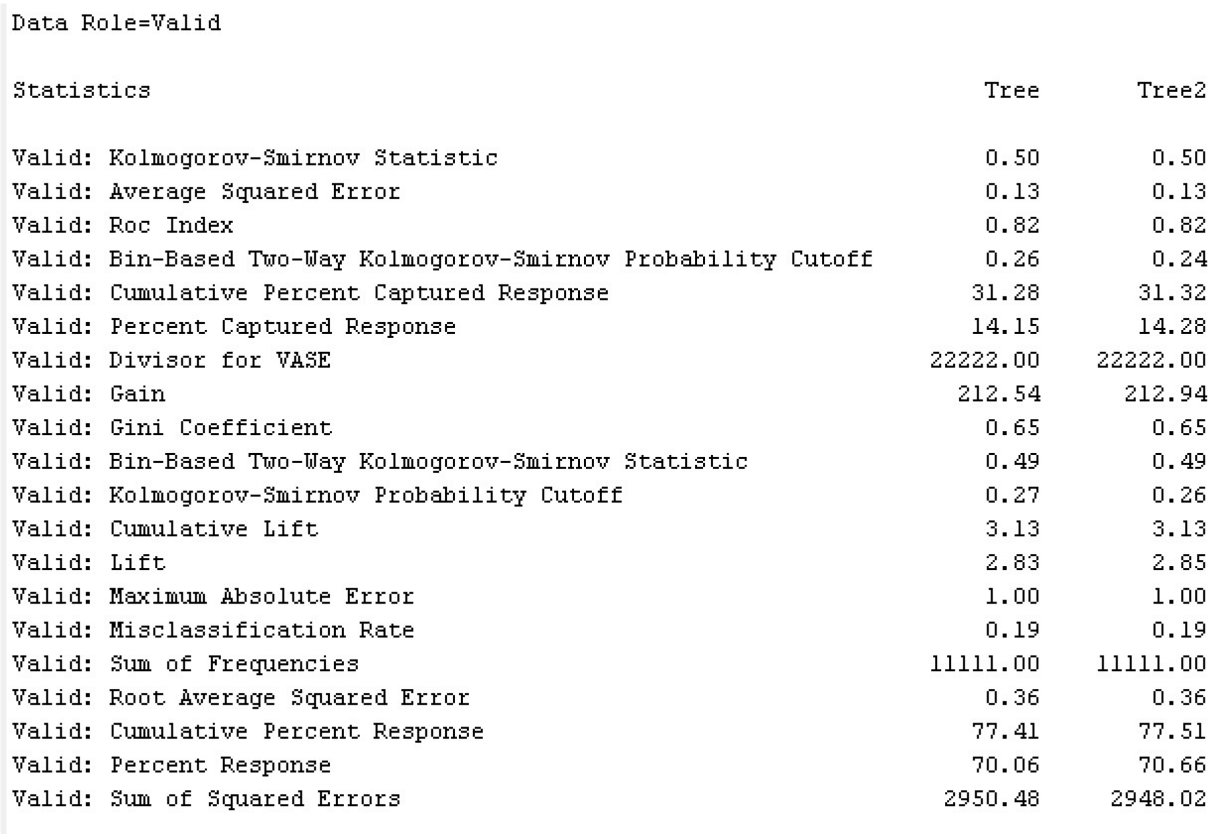


```
Type           Variable       Label

TARGET         TargetBuy      Organics Purchase Indicator
PREDICTED      P_TargetBuy1   Predicted: TargetBuy=1
RESIDUAL       R_TargetBuy1   Residual: TargetBuy=1
PREDICTED      P_TargetBuy0   Predicted: TargetBuy=0
RESIDUAL       R_TargetBuy0   Residual: TargetBuy=0
FROM           F_TargetBuy    From: TargetBuy
INTO           I_TargetBuy    Into: TargetBuy


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


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



Variable Importance

                                                                                  Ratio of
                                 Number of                                      Validation
Variable                         Splitting                       Validation    to Training
Name          Label                Rules          Importance     Importance     Importance

DemAge        Age                    2              1.0000         1.0000         1.0000
DemAffl       Affluence Grade       11              0.8331         0.7597         0.9119
DemGender     Gender                 7              0.5356         0.6168         1.1515



Tree Leaf Report

                                  Training
Node                Training      Percent     Validation    Validation
 Id      Depth    Observations       1       Observations    Percent 1

 49        4          1473          0.20         1534          0.21
```

D. Model Comparison



```
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model                         Data                                                  False      True      False      True
Node    Model Description     Role       Target      Target Label                  Negative   Negative   Positive   Positive

Tree    Decision Tree         TRAIN      TargetBuy   Organics Purchase Indicator    1676       7978        381       1077
Tree    Decision Tree         VALIDATE   TargetBuy   Organics Purchase Indicator    1679       7979        380       1073
Tree2   Decision Tree (2)     TRAIN      TargetBuy   Organics Purchase Indicator    1584       7890        469       1169
Tree2   Decision Tree (2)     VALIDATE   TargetBuy   Organics Purchase Indicator    1593       7867        492       1159
```



```
Data Role=Valid

Statistics                                                              Tree       Tree2

Valid: Kolmogorov-Smirnov Statistic                                     0.50        0.50
Valid: Average Squared Error                                            0.13        0.13
Valid: Roc Index                                                        0.82        0.82
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff          0.26        0.24
Valid: Cumulative Percent Captured Response                            31.28       31.32
Valid: Percent Captured Response                                       14.15       14.28
Valid: Divisor for VASE                                             22222.00    22222.00
Valid: Gain                                                           212.54      212.94
Valid: Gini Coefficient                                                 0.65        0.65
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic                   0.49        0.49
Valid: Kolmogorov-Smirnov Probability Cutoff                            0.27        0.26
Valid: Cumulative Lift                                                  3.13        3.13
Valid: Lift                                                             2.83        2.85
Valid: Maximum Absolute Error                                           1.00        1.00
Valid: Misclassification Rate                                           0.19        0.19
Valid: Sum of Frequencies                                           11111.00    11111.00
Valid: Root Average Squared Error                                       0.36        0.36
Valid: Cumulative Percent Response                                     77.41       77.51
Valid: Percent Response                                                70.06       70.66
Valid: Sum of Squared Errors                                         2950.48     2948.02
```