**Solution Document**

**Assignment 2**

**Business Analytics with SAS – F16**

**(MIS 6324.502)**

Submitted by

Group 8

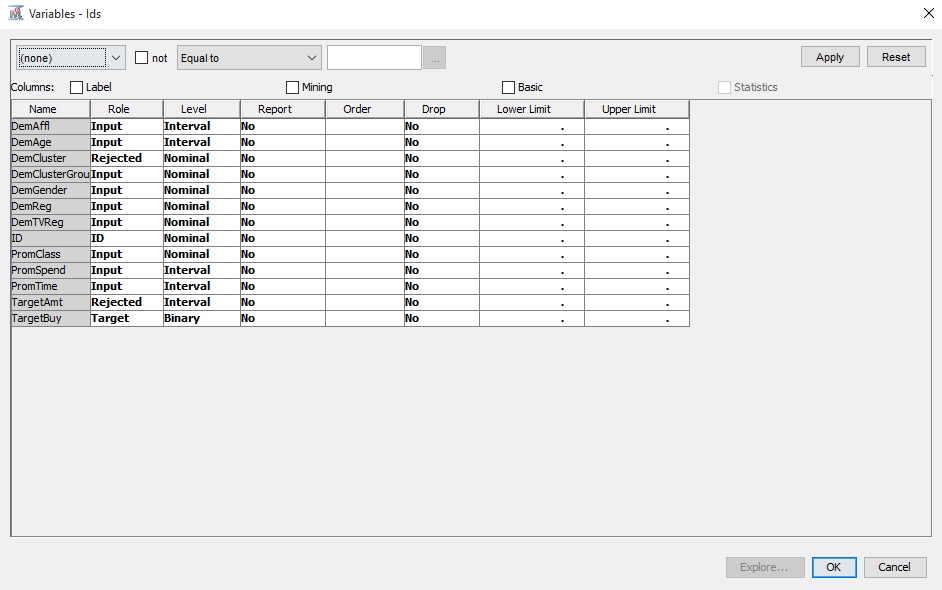
|  |  |
| --- | --- |
| Mayank Kothari | mkk160130 |
| Shamsundar Kulkarni | sxk162231 |
| Divya Jayaprakash | dxj160830 |
| Gautami Murugan | gxm161330 |
| Vijay Balaji | vxb161530 |

**Exercise 1: Customer Classification**

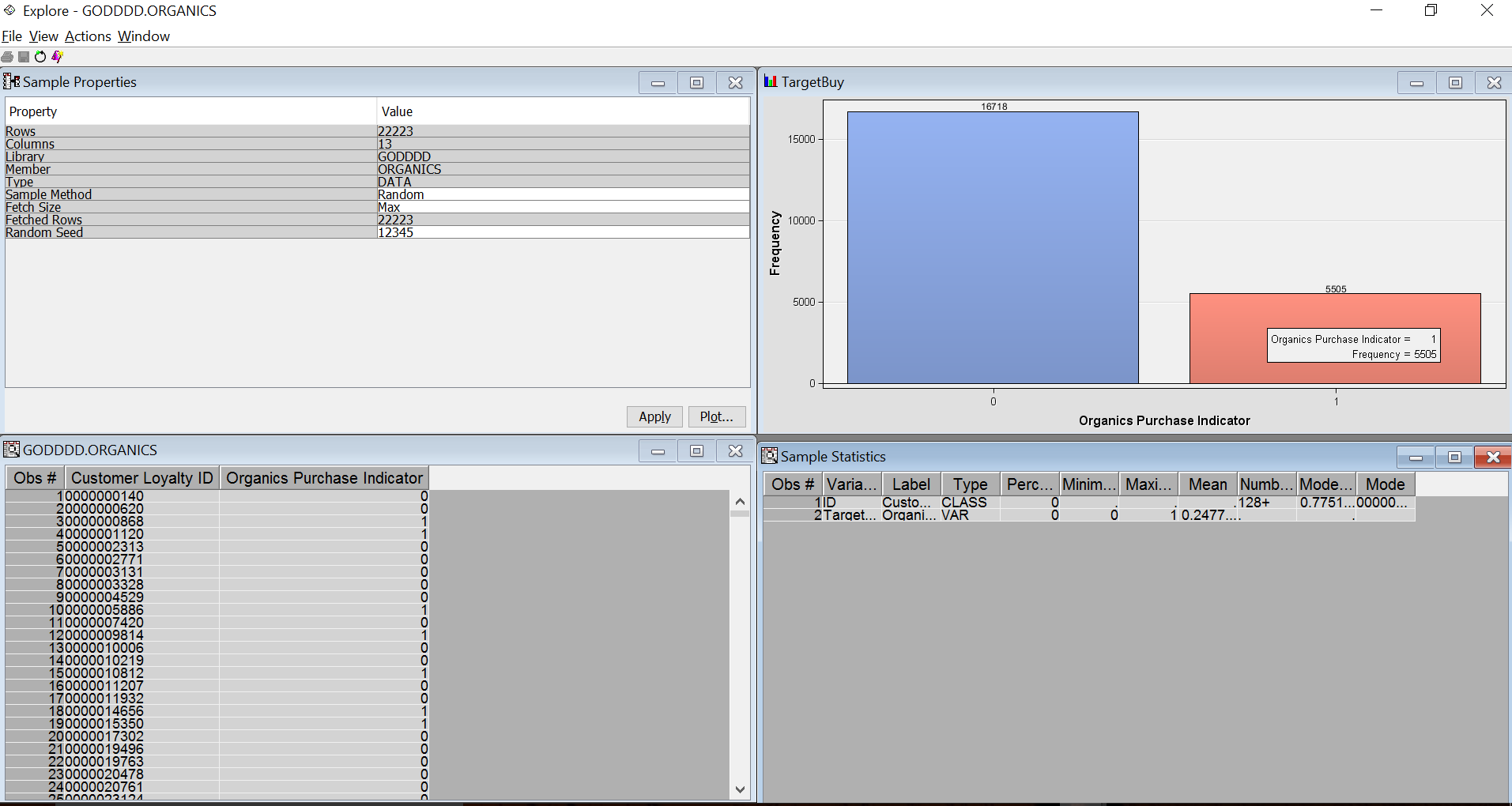
1. Create a new diagram. Name it Organics.

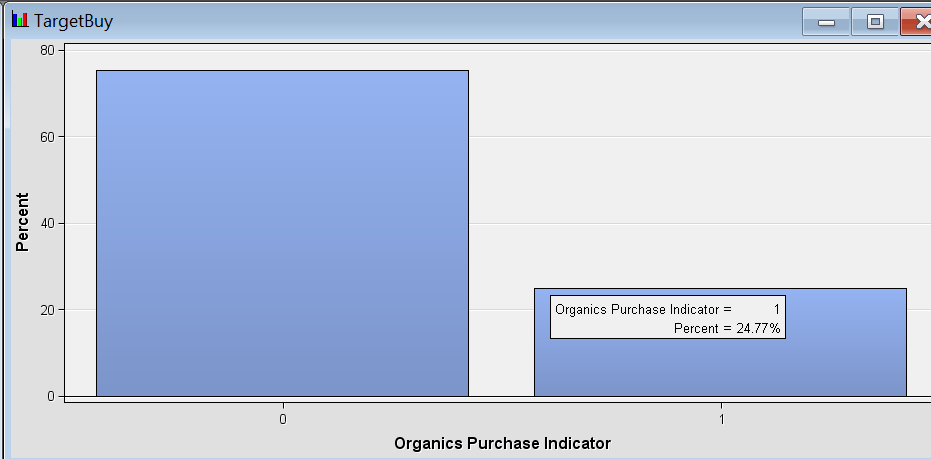
2. Define the data set ORGANICS as a data source.

a. Set the roles for the analysis variables as shown above



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





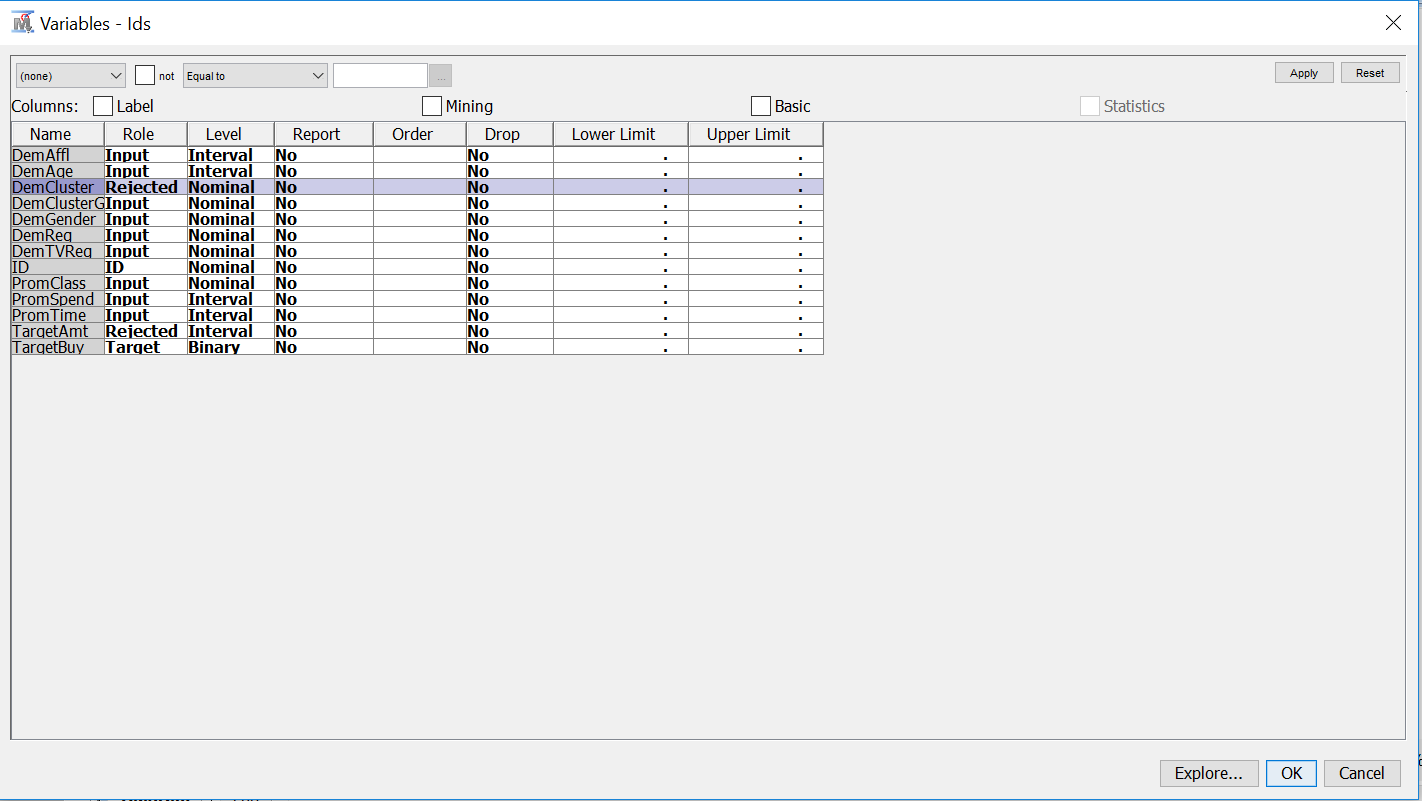
The organic purchase indicator with value 1 represents the proportion of the customer who purchased organic products.

The proportion of individuals who purchased organic products = 24.34%

5505 customers out of 22223 customers bought organic products.

c. 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.

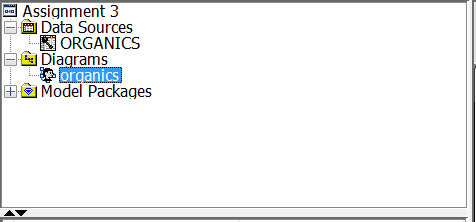


d. As noted above, only TargetBuy will be used for this analysis and should have a role of Target. Can TargetAmt be used as an input for a model used to predict TargetBuy? Why or why not?

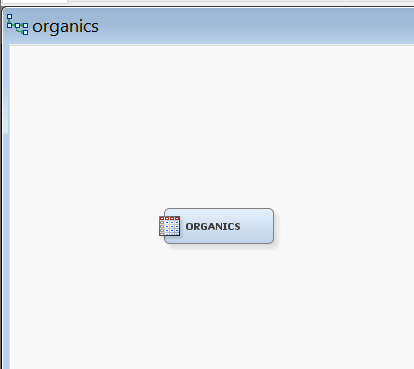
TargetAmt cannot be used as an input to predict TargetBuy because it contains collapsed levels of TargetBuy variable.i.e. it denotes the count of organic products that has been bought and is also measured at the same time when TargetBuy is obtained.

TargetAmt cannot be used as an input for our model to predict TargetBuy. TargetAmt is the total count of the organic products purchased by each customer. It can be measured only at the same time when TargetBuy is acquired. TargetAmt is just the detail of TargetBuy and not a good predictor of TargetBuy.

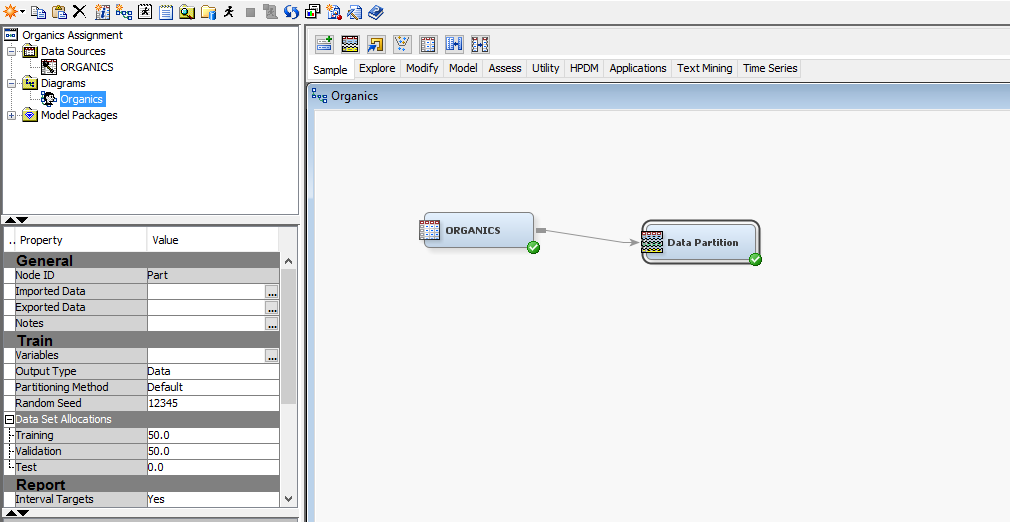
e. Finish the ORGANICS data source definition.

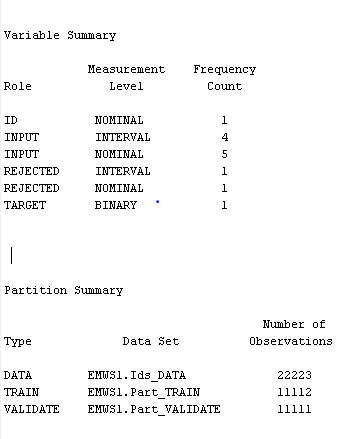


3. Add the ORGANICS data source to the Organics diagram workspace.

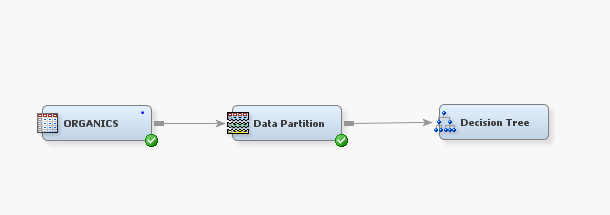


4. Add a Data Partition node to the diagram and connect it to the Data Source node. Assign 50% of the data for training and 50% for validation.

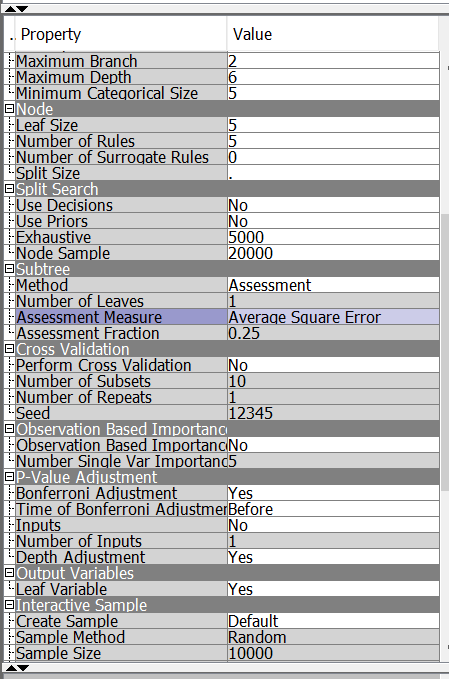




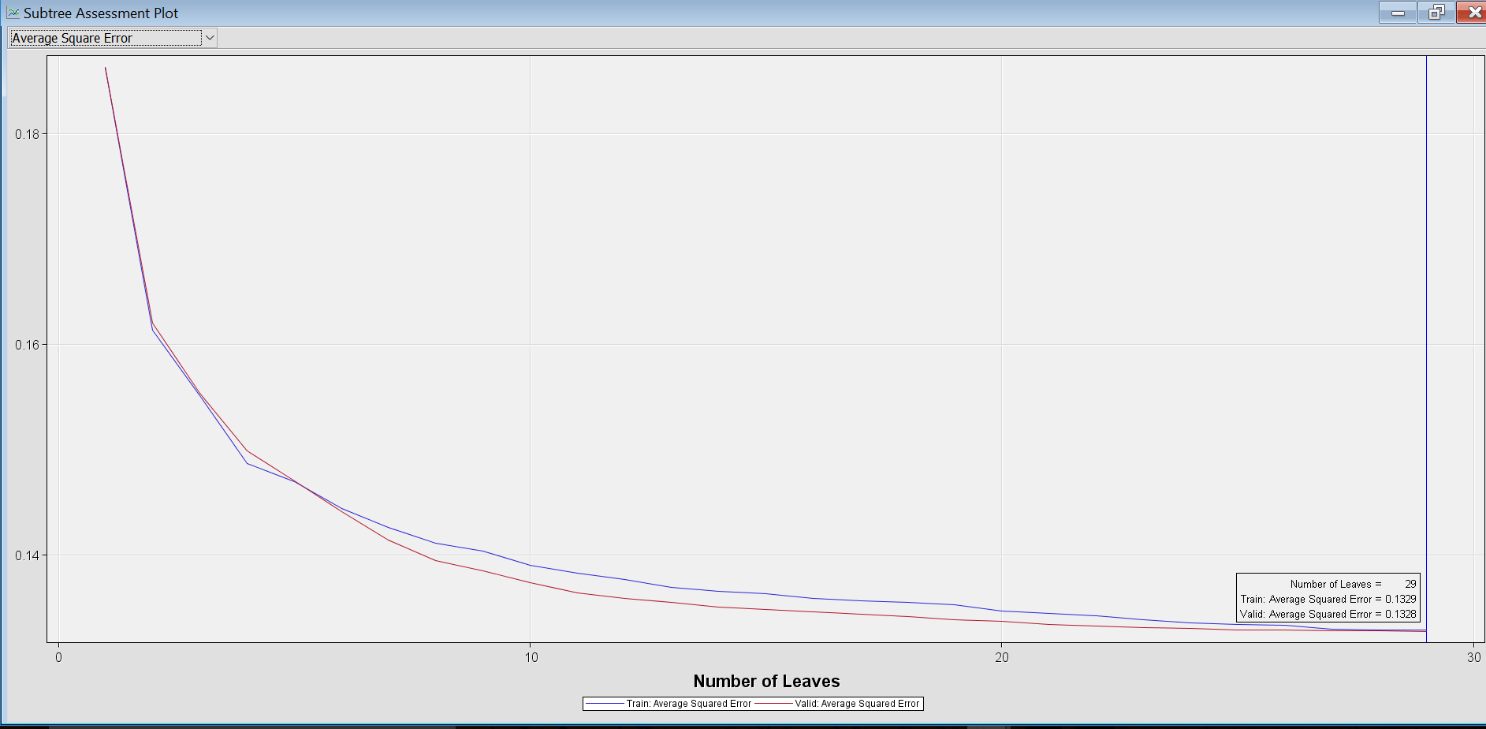
5. Add a Decision Tree node to the workspace and connect it to the Data Partition node.



6. Create a decision tree model autonomously. Use average square error as the model assessment statistic.



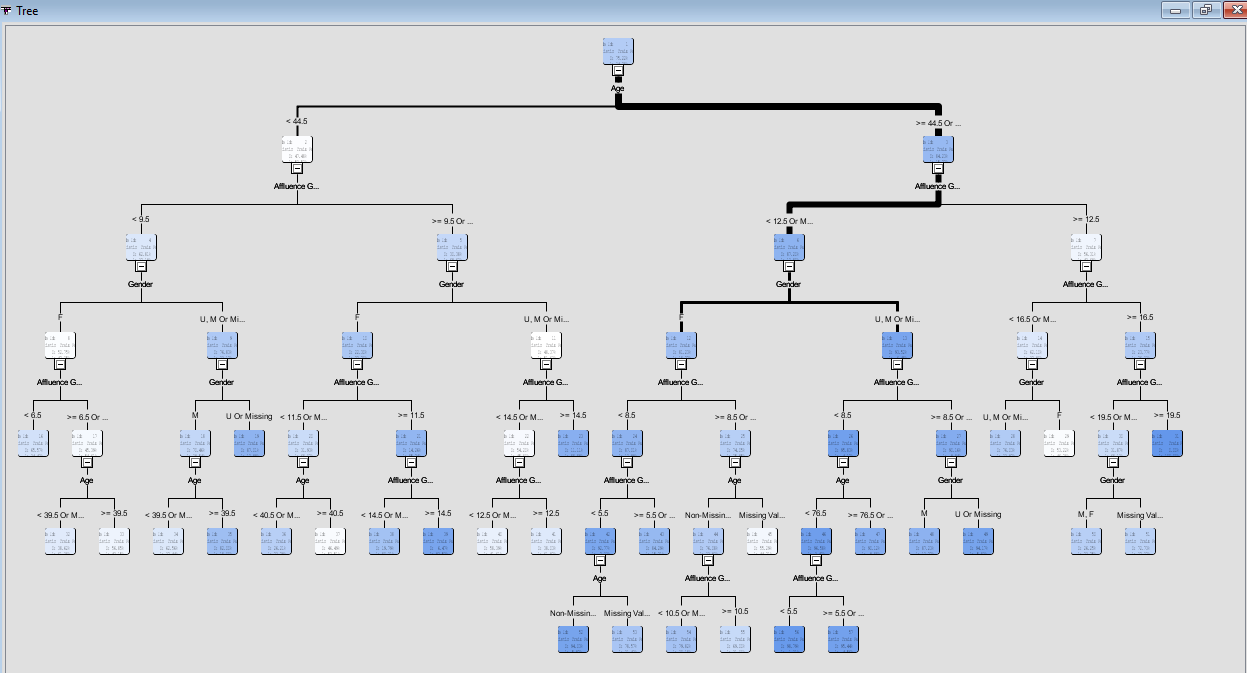
a. How many leaves are in the optimal tree?



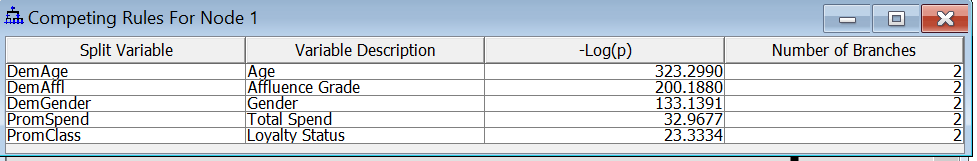
From the subtree assess plot we can see that using average square error as assessment measure, the number of leaves in optimal tree is 29.

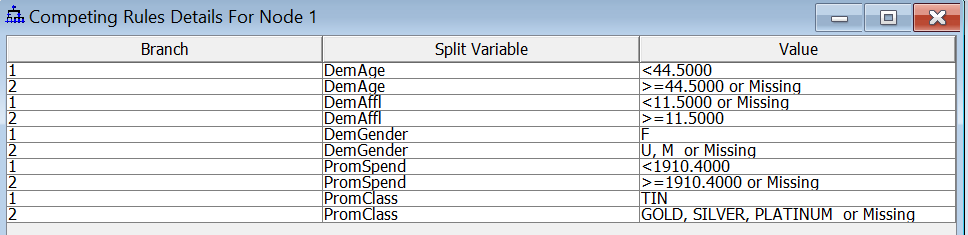
b. Which variable was used for the first split?

The variable used for the first split is AGE.



c. What were the competing splits for the first split?

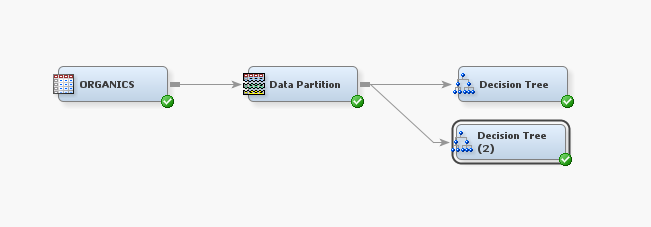


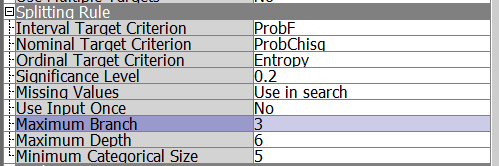


Affluence Grade, Gender, Total Spend, Loyalty Status are the variables competing with Age for the first split.

7. Add a second Decision tree node to the diagram and connect it to the Data Partition node.

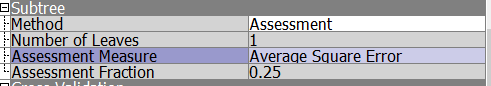
a. In the Properties panel of the new Decision Tree node, change the maximum number of branches to allow for three‐way splits.



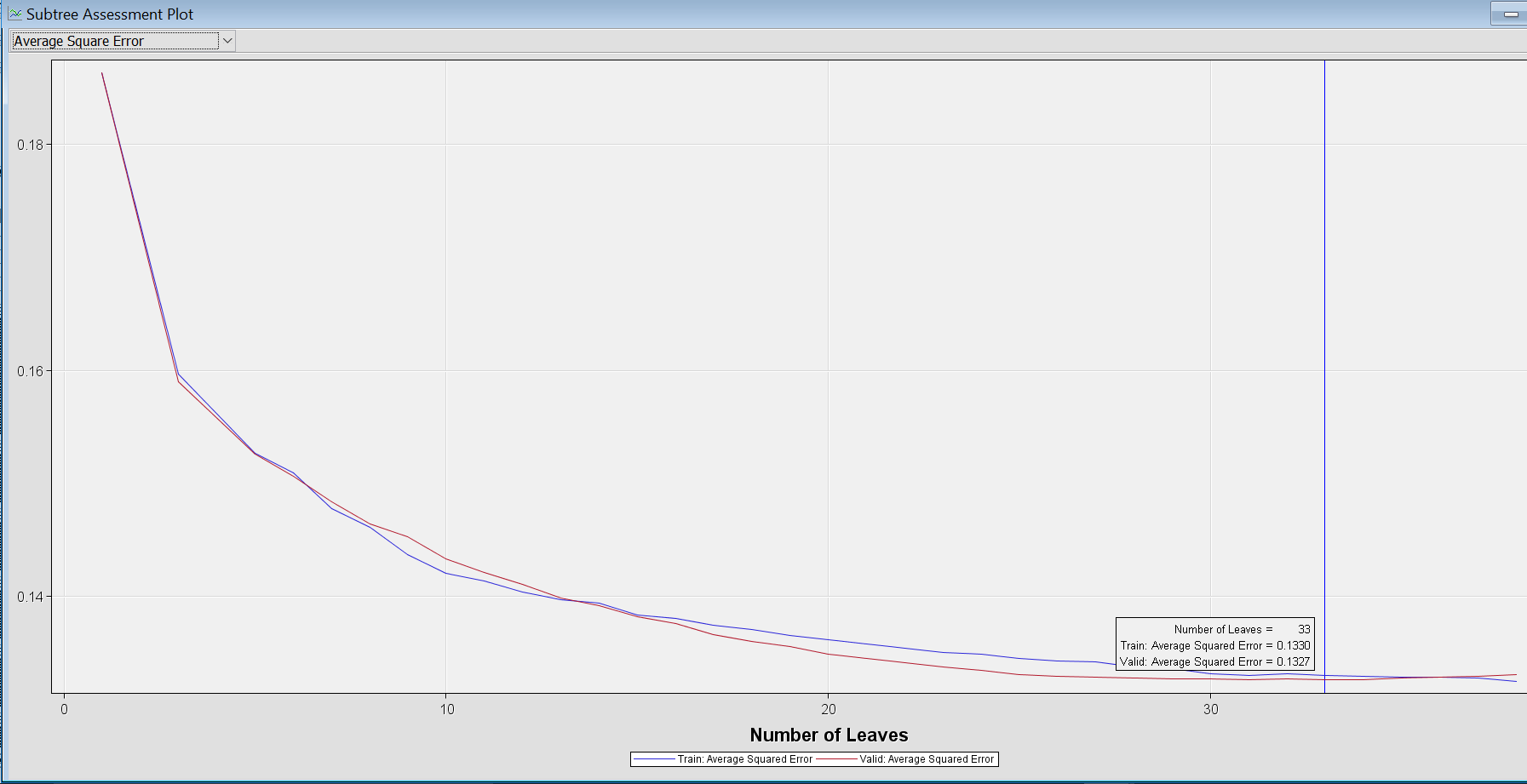


b. Create a decision tree model using average square error as the model assessment

statistic.



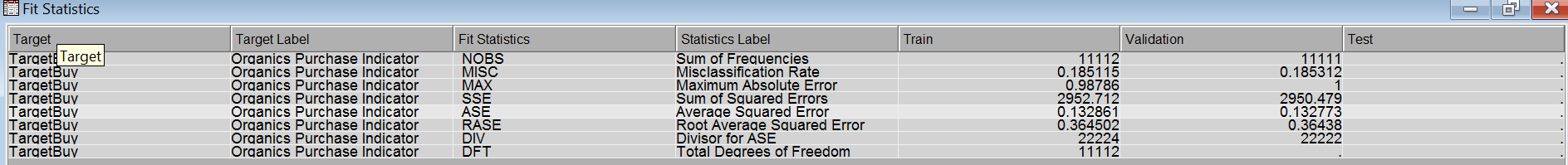
c. How many leaves are in the optimal tree?



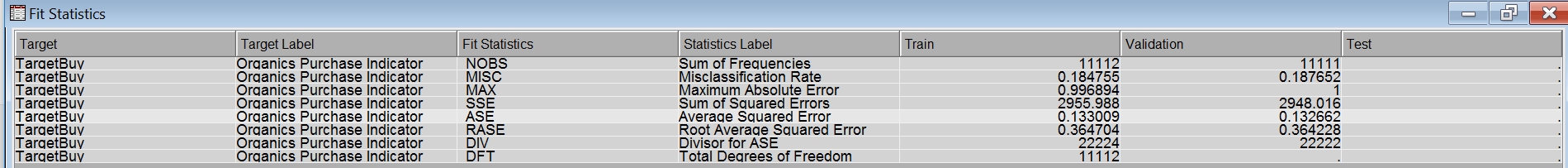
After changing the maximum number of branches to allow for three way splits we can see in the Subtree Assessment Plot that the number of leaves in optimal tree is 33.

8. Based on average square error, which of the decision tree models appears to be better?

Fit statistics for the first decision tree:



Fit statistics for the second decision tree:



From the fit statistics for both the decision trees we could see that:

Average squared error for second decision tree model with 3-way split is 0.1330. It is slightly higher than the average squared error for first decision tree with 2-way split, which is 0.1328. So the first decision tree with the low average squared error is a better model.