Decision Tree Lab

BA706 Fall 2024

Submitted by

Grace Adaji
(301373339)

1. Initial Data Exploration

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.

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 over 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels:

Name	Model Role	Measurement Level	Description
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
DemClusterGroup	Input	Nominal	Neighborhood group
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

Although two target variables are listed, these exercises concentrate on the binary variable **TargetBuy**.

a. Create a new diagram named **Organics**.

b. Define the data set **AAEM.ORGANICS** as a data source for the project.

1) Set the model roles for the analysis variables as shown above.
2) Examine the distribution of the target variable. What is the proportion of individuals who purchased organic products?
3) 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**.

4) 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?

Finish the **Organics** data source definition.

c. Add the **AAEM.ORGANICS** data source to the Organics diagram workspace.

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.

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

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

1) How many leaves are in the optimal tree?

   Which variable was used for the first split?

2) What were the competing splits for this first split?

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

4) In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to **3** to enable three-way splits.

5) Create a decision tree model. Use average square error as the model assessment statistic.

6) How many leaves are in the optimal tree?

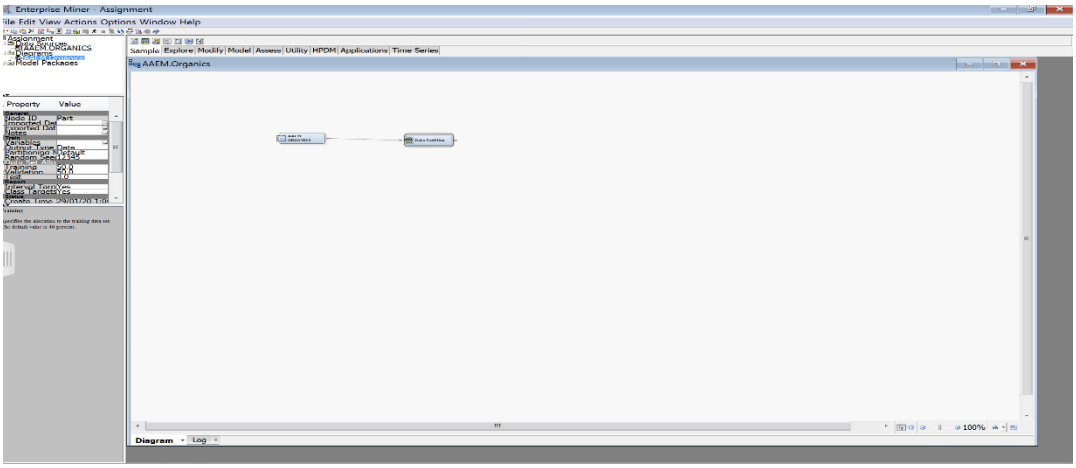
f. Based on average square error, which of the decision tree models appears to be better?

g. Based on your analysis write a report to the supermarket manager how to determine the potential customers for the new line of organic products. Specify the variables (and their various levels) in relation to customer's decision to purchase these products. Do you think the loyalty coupons worked?

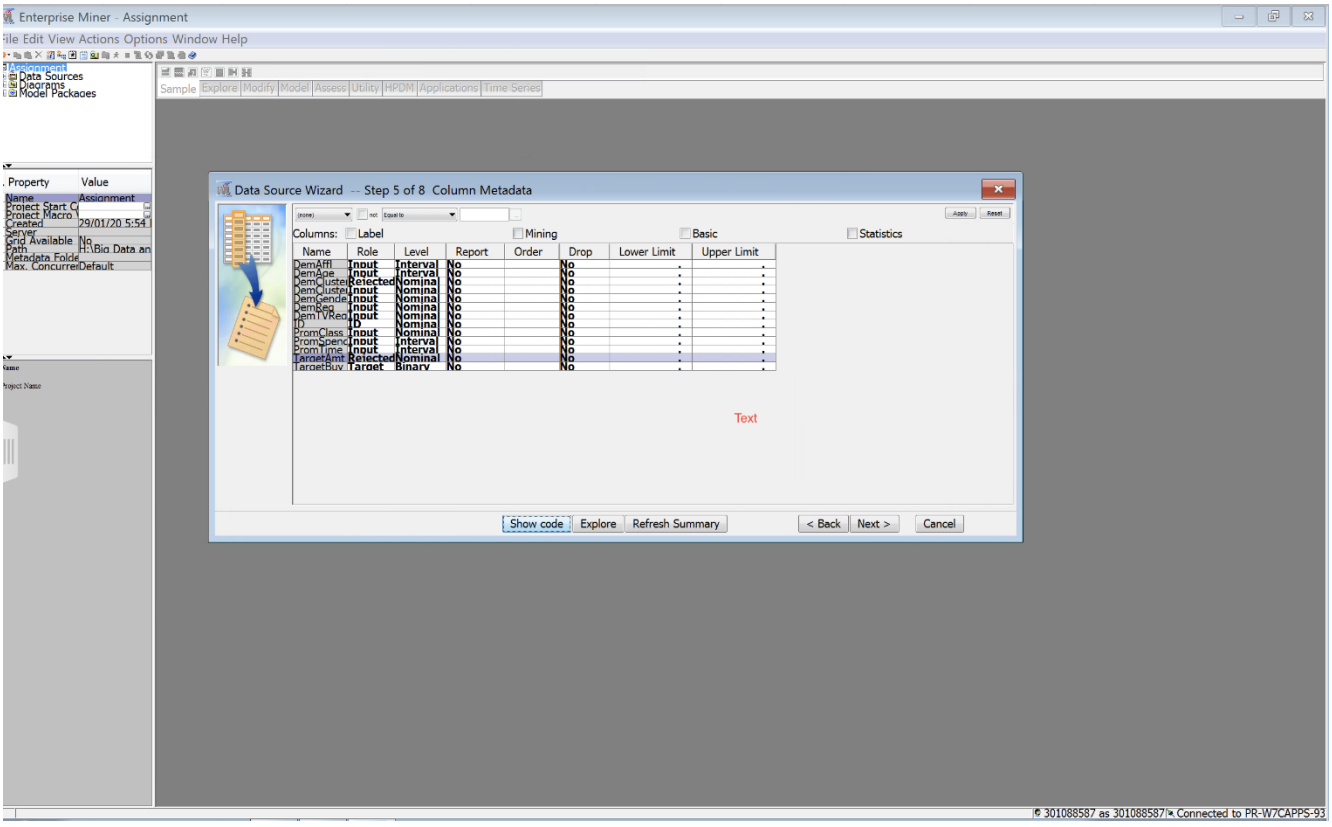
Answers:

a. Create a new diagram named **Organics**.

b. Define the data set **AAEM.ORGANICS** as a data source for the project.



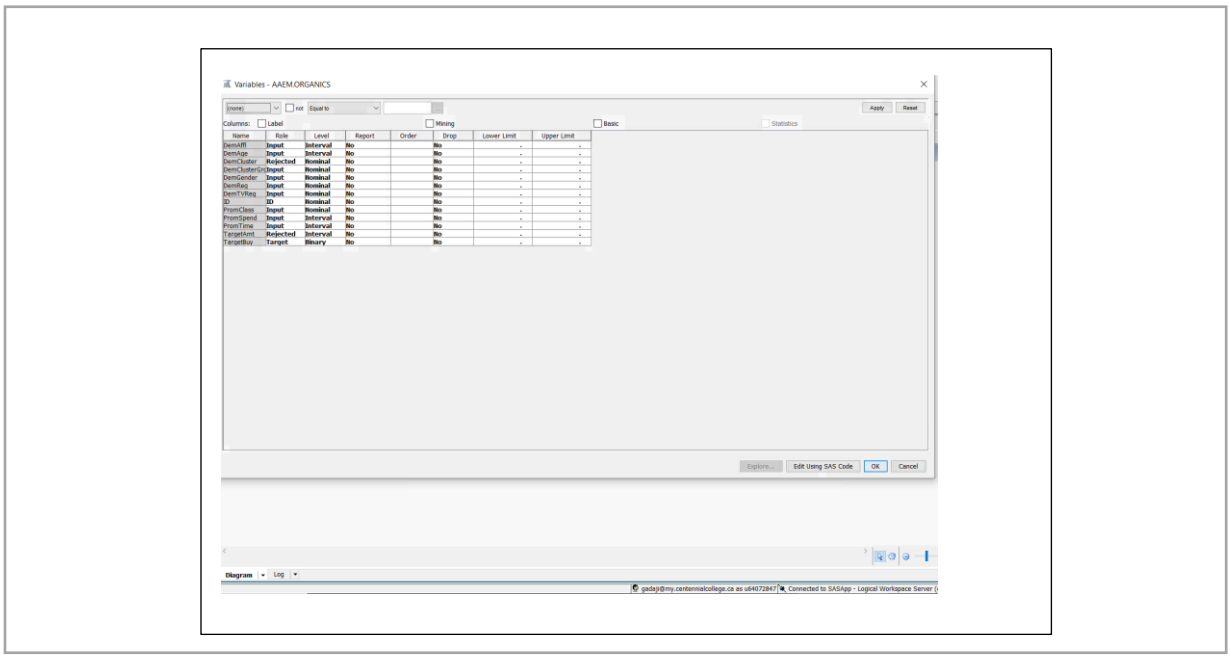
1) Set the model roles for the analysis variables as shown.



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

Type your answer here:24.33%

3) 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**.



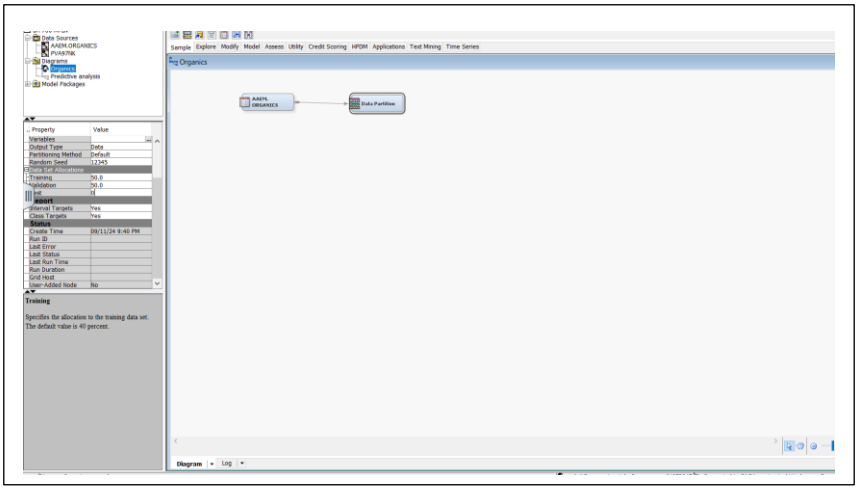
4) As noted above, only **TargetBuy** will be used for this analysis and should have a role of **Target**. Can **TargetAmt** be used as an input for a model used to predict **TargetBuy**? Why or why not?

Type your answer here: No. Only one type of variable can be predicted at a time. In this case since we are interested in predicting which customers are likely to purchase, we will use only Target buy, which indicates weather or not Organic products were purchased(yes or no) rather than Target Amt, which shows the number of Organics products purchased

Finish the **Organics** data source definition

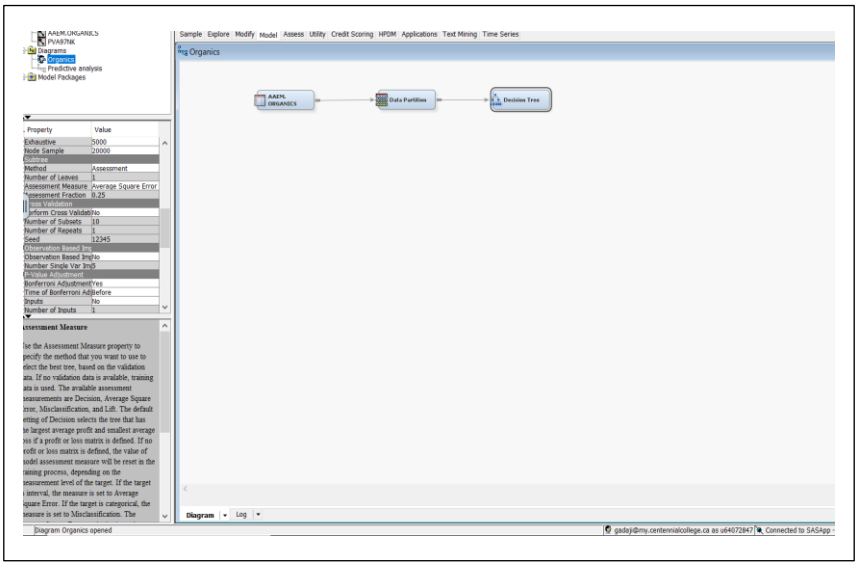
h. Add the **AAEM.ORGANICS** data source to the Organics diagram workspace.

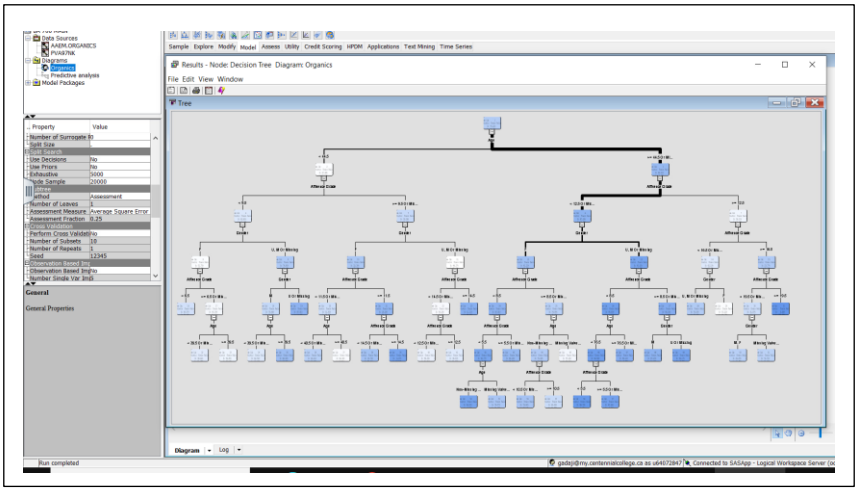
Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 50% of the data for training and 50% for validation.



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

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





1) How many leaves are in the optimal tree? Type your answer here:29

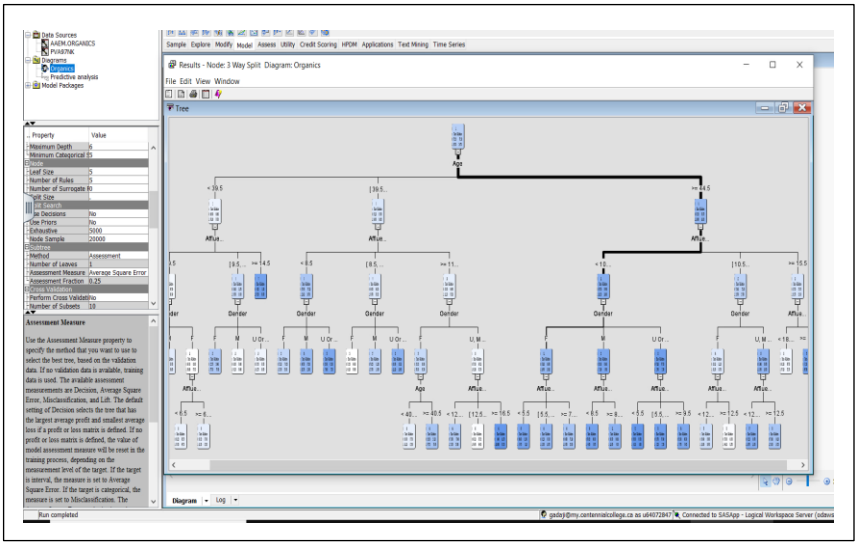
2) Which variable was used for the first split? Type your answer here:Age

3) What were the competing splits for this first split? Type your answer here: : Dem Affl, and then Dem Gender

k. Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.

1) In the Properties panel of the new Decision Tree node, change the maximum number of branches from a node to **3** to enable three-way splits.

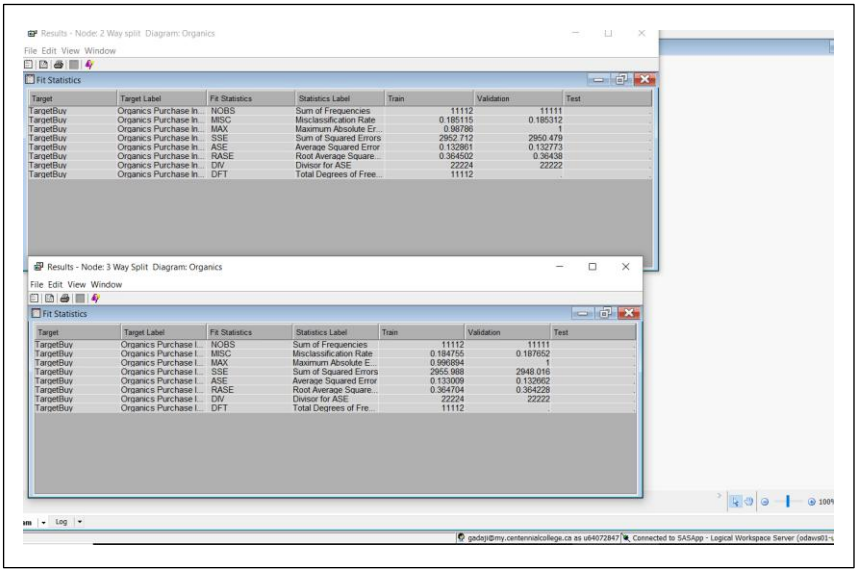
2) Create a decision tree model. Use average square error as the model assessment statistic.



3) How many leaves are in the optimal tree?

Type your answer here:33

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



Type your answer here: : Based on the Decision Tree models; the Three-way split is better than the 2-way split because the 3- way split has a slightly lower validation average square error(ASE) compared to the 2 way split . The validation ASE for 3-way split is 0.132662 3 while the Validation Average for 2-way split is 0.132773

m. Based on your analysis write a report to the supermarket manager on how to determine the potential customers for the new line of organic products. In the report, specify the variables (and their various levels) in relation to customer's decision to purchase these products and how would the loyalty coupons work?

Dear Manager,

Subject: **Analysis for Identifying Potential Customers for Organic Products**

As part of the supermarket marketing strategy to introduce a new line of organic products, an analysis was conducted to determine which customers are likely to purchase the organic products, utilizing data collected from the coupons distributed to customers that were in the supermarket loyalty program. Using a decision tree, a predictive model was developed to predict key factors that could influence the purchase and to make recommendations to the company for optimal segmentation.

Data Analysis Summary

The analysis utilized a data set of over 22,000 observations and 13 variables. Two decision tree models were created: one using two-way splits and another using three-way splits.

Description	**Two ways split tree**	**3 ways split tree**
Number of leaves	**29**	**33**
First split variable	**age**	**age**
Competing splits	**1.Affluence 2.Gender**	**1.Affluence2.Gender**
Average square error	**0.132773**	**0.132662**

Comparison

Both models had nearly the same performance, but the three-way split model was slightly better, achieving a lower validation Average Square Error (ASE) of 0.132662 compared to the two-way split model's ASE of 0.132773. This indicates that the three-way split model provided more accurate segmentation while maintaining strong predictive capabilities.

Key Insights Drawn.

The Decision tree revealed the following key predictors and segmentation levels:

1. **Age** was the most significant predictor of whether customers would purchase organic products, indicating that targeting specific age demographics greater than 44.5years could enhance the effectiveness of marketing efforts
2. After Age, the next level of segmentation should focus on **Affluence**. Higher affluence of customers will be more inclined to purchase organic products. **Gender** could be considered as specific gender may have different response to purchasing organic products
3. Loyalty coupons boosted initial purchases, and targeting segments by age, affluence, and maybe gender in that order could enhance marketing effectiveness.

Recommendation

Based on the analysis, my recommendation is to focus marketing on specific age groups, high-affluence customers, and possibly responsive gender segments using personalized offers and targeted coupons. There is need to continuously collect data and monitor customer behavior to adapt marketing strategies and enhance engagement. This will help in refining marketing efforts overtime.

Grace Adaji