SEIS 632 Data Analytics and Visualization

Assignment 1

Due: On canvas by midnight 9/23/21

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, the target variable of interest to us is the binary variable **TargetBuy**.
- Create a new diagram named **Organics**.
- Define the data set **AAEM.ORGANICS** as a data source for the project.
- Set the model roles for the analysis variables as shown above.
- Examine the distribution of the target variable.

Question 1: What is the proportion of individuals who purchased organic products?

- 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**.

- As noted above, only **TargetBuy** will be used for this analysis and should have a role of Target. Set the role for **TargetAmt** to Rejected.
- Finish the **Organics** data source definition.
- 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.
- Add a **Decision Tree** node to the workspace and connect it to the **Data Partition** node.
- Create a decision tree model autonomously. Use **Decision** as the model assessment statistic.

Question 2: How many leaves are in the optimal tree?

Question 3: Which variable was used for the first split?

Question 4: What were the competing splits for this first split? That is, based on the logworth what were the other top attributes at the second, third, and fourth place.

- Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.
- Create a decision tree model autonomously. Use **average square error** as the model assessment statistic.

Question 5: How many leaves are in the optimal tree?

Question 6: Which variable was used for the first split?

Question 7: What were the competing splits for this first split?

Submission:

1) Answers to the questions asked above.

2) In the diagram, as a last node, add a Reporter node from the Utility tab. Change the Nodes property of the Reporter node to All. Now right click on the Reporter node and select Run. Open the results and then click View. You should then able to open a pdf and save.

You should submit the above 2 items on canvas.