**Part 1: PREDICTIVE ANALYSIS RFM**

**TASK 2**

*a. Define PVA97NK as a data source in SAS Enterprise Miner. Use the Advanced Metadata Advisor*

*options to customize the following:*

• Change the Class Levels Count Threshold from 20 to 5.

• Change the Reject Levels Count Threshold from 20 to 80.

Graphical user interface, text, application

Description automatically generated

• Reject the variable TargetD.

Graphical user interface, table

Description automatically generated

*b. Create a new diagram and transform the R, F, and M variables as described previously to create four*

*bins of each variable. Concatenate them to create an RFM variable.*

Diagram

Description automatically generated

Variable GiftTimeLast\_Rev = -1\*GiftTimeLast

Table

Description automatically generated

Create a new variable for Monetary value. Variable TotalGiftAmount = GiftAvgAll\*GiftCntAll

Graphical user interface, table

Description automatically generated with medium confidence

BinRFM variables

Table

Description automatically generated

RFM

Graphical user interface

Description automatically generated

*c. Explore the data and perform graphical RFM analysis using a grouped pie chart and a stacked bar*

*chart.*

Chart, pie chart

Description automatically generated

The inner ring of the pie represents non-responders to target gift flag , and the outer ring shows responders to target gift flag. The area of each slice demonstrates the relative proportion of responders and non-responders to gift flag in each RFM group.

Chart, bar chart

Description automatically generated

The grouped bar chart shows the number of observations in each RFM group as the height of the bar,

with the upper portion representing the portion of response. The lower portion of the bar is

the proportion of non-responders to the target gift flag.

The 26th bar from the left has a relatively high proportion of responders and this makes sense because it’s in the group 04 (Recency)04(Frequency)04(Monetary Value).

*d) Calculate response rate for 040404 and 030303 group?*

Chart, bar chart

Description automatically generated

There are 359 responders in group 040404.

Chart, bar chart

Description automatically generated

There are 188 non- responders in group 040404.

The response rate for the 040404 group is 359 / (188+359) = 65.63%.

Chart, bar chart

Description automatically generated

There are 224 responders in group 030303.

Chart, bar chart

Description automatically generated

There are 219 non- responders in group 030303.

The response rate for the 030303 group is 224 / (219+224) = 50.56%.

*e. Each promotional mailing (request for a gift) costs $2.3, and the average donation is about $21 What*

*is the break-even response rate for this promotion? Do any RFM cells exceed this response rate?*

*Remember to account for the fact that in the population, 95% of mailings are not responded to, while*

*this sample is oversampled to 50% responders and 50% non-responders.*

The break-even response rate for this promotion is 10.95%. Many RFM cells exceed this response rate such as 030303 and 040404.

**Part 2: PREDICTIVE MODLING**

**TASK 3: ORGANICS DATASET**

*1. Set the roles for the analysis variables as shown above.*

Table

Description automatically generated

*2. Examine the distribution of the target variable.*

Table

Description automatically generatedThe proportion of individuals who purchased organic products is 24.7716%

*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.*

Graphical user interface, application, table, Excel

Description automatically generated

*4. As noted above, only* ***TargetBuy*** *is used for this analysis, and it should have a role of* ***Target****.*

*Can* ***TargetAmt*** *be used as an input for a model that is used to predict* ***TargetBuy****? Why or*

*why not?*

No, because **TargetAmt** implies a buy, therefore, it cannot be used to predict **TargetBuy.** Adding this variable to the model will cause a perfect collinearity between **TargetAmt** variable and **TargetBuy** target variable.

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

d*. Add a* ***Data Partition*** *node to the diagram and connect it to the* ***Data Source*** *node. Assign 65%*

*of the data for training and 35% for validation.*

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated*e. Add a* ***Decision Tree*** *node to the workspace and connect it to the* ***Data Partition*** *node.*

*f. Create a* ***decision tree*** *model autonomously. Use* ***Misclassification*** *as the model assessment*

*statistic.*

Graphical user interface, text

Description automatically generated

1. How many leaves are in the optimal tree? 14 leaves

Graphical user interface, application

Description automatically generated

*2. Which variables were used for the first split?*

Age variable is used for the first split. Age < 44.5 and Age >= 44.5 or Missing

Diagram

Description automatically generated with low confidence

*What were the competing splits for this first split?*

Affluence Grade (importance = 0.7806) and Gender (importance = 0.5184)

Table

Description automatically generated with medium confidence

*3. Which variables were used for the second split for all branches from first split?*

Affluence Grade was used for the second split for all branches from the first split. The cutoff value for affluence grade for the left node is affluence grade < 9.5 and >= 9.5 or Missing. The cutoff vale for the right node is affluence grade < 12.5 or Missing or >= 12.5.

Diagram

Description automatically generated

*4. Discuss the results and provide your insights*

Diagram, timeline

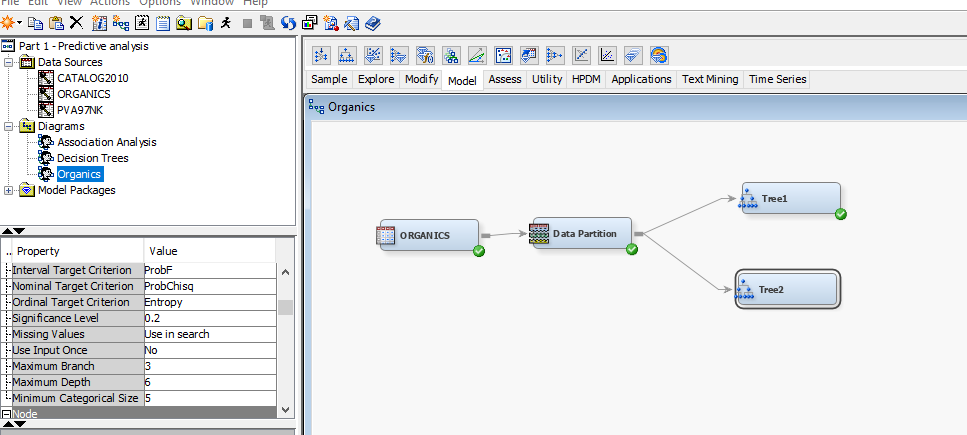
Description automatically generated

This is a complex tree with 51 nodes in total and nodes are reused to split as needed. Affluence grade variable also was also used multiple times (node 2,3,7,9,10,20), with different cutoff values to split the nodes.

***g.*** *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 allow for three-way splits.*



*2. Create a decision tree model using* ***Misclassification*** *as the model assessment statistic.*

Graphical user interface

Description automatically generated

*3. How many leaves are in the optimal tree?*

16 leaves in tree2

Graphical user interface

Description automatically generated

***h.*** *Based on* ***Misclassification rate****, which of the decision tree models appears to be better?*

Diagram

Description automatically generated with medium confidence

Tree2 has a slightly higher Misclassification rate than Tree1 0.18858 compared to 0.18380.

Graphical user interface, table

Description automatically generated

Graphical user interface

Description automatically generated

**Part 3: PREDICTIVE MODELING USING LOGISTIC REGRESSION**

Review the **CATALOG2010** data set.

A screenshot of a computer

Description automatically generated with medium confidence

Cluster variables according to their similarities.

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Run the **Variable Clustering** node. View the results.

Table

Description automatically generated with low confidence

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Summary of the final cluster solution

Table

Description automatically generated

Table

Description automatically generated

Estimate a logistic regression model.

A picture containing graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Standard output for Logistics regression

Table

Description automatically generated

The output below shows the results of the final model

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

The assessment statistics for the validation data set for the 5th percentile, the10th percentile, and so on.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

View model performance across the fitted models.

Graphical user interface, application

Description automatically generated

View the misclassification rate.

Graphical user interface, chart

Description automatically generated

Run the Model Comparison node and view the results.

A picture containing diagram

Description automatically generated

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Text, letter

Description automatically generated

Table

Description automatically generated