**Assignment 7 - SAS #2 – Decision Trees**

**Guidelines**

* Follow the steps below, using the in-class exercise as a guide.
* You must submit your answers electronically in a single Word document. Fill in the answer sheet at the end of this document.
* You must include your name at the top of the document.
* Your answers should be emailed, as an attachment, to your instructor with the subject:
**MIS2502: Decision Trees**
* The email must be sent by the start of class the day the assignment is due.

***If you do not follow these instructions, your assignment will be counted late.***

***Instructions: Follow the steps and answer the questions below. Then email this document to your instructor.***

You’ll be working on the project you created in the previous assignment. Remember, this project used the “Organics” data set. When you open SAS Enterprise Miner, you should be able to find your work under the File/Recent Projects. If you can’t find it there, go to File/Open Projects… and search for your project.

*Make sure that the variable roles and levels in the Organics data set are still properly set up from the last assignment. You can verify this by right-clicking the Organics node in your diagram and selecting “Edit Variables.” Check those settings against page 1 of “Introduction to Working with SAS.”*

**Create a Decision Tree based on the Organics Data Set**

1. Add a **Data Partition** node to the diagram and connect it to the **Data Source** node. Assign 60% of the data for training and 40% for validation (and 0% for test). Run it.

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

Create a decision tree model autonomously (i.e., just run the Decision Tree node).

How many leaves are in the optimal tree?
*(HINT: In your subtree assessment plot, you can click near the “Number of leaves” label and drag right to zoom in*

*Waypoint: The average squared error of the plot should be .1403)***ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

**Include a screen shot showing where you found your answer (place on the answer sheet)**

1. Which variables were used for the first and second splits?  **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

**Include a screen shot showing where you found your answer (place on the answer sheet)**
2. Add a second **Decision Tree** node to the diagram and connect it to the **Data Partition** node.

In the Properties panel of the new Decision Tree node, change the maximum number of branches (Maximum Branch) from a node to 3 to allow for three-way splits.

How many leaves are in the optimal tree?

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

**Include a screen shot showing where you found your answer (place on the answer sheet)**

1. Based on misclassification rate, which of the decision two tree models appears to be better (the first [autonomous] one or the second one)?  **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

**Include a screen shot showing where you found your answer (place on the answer sheet)**
2. Explain your answer in #4 (why did you say that model was the better one?).

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

**<<GO TO THE NEXT PAGE!!>>**

1. Start at the top of the second decision tree you created and work your way down to answer the following questions. You don’t need to include screen shots for these questions – just provide the answer:
***(Remember, use the validation set to find the probabilities, and a “1” means they buy.)***

|  |  |  |
| --- | --- | --- |
| Item | Question | Answer |
| 6a | What is the probability that a 33 year old man with affluence grade 10 buys Organics? | (use the answer sheet) |
| 6b | What is the probability that a 23 year old woman with affluence grade 3 buys Organics? | (use the answer sheet) |
| 6c | What is the probability that a 55 year old man with affluence grade 6 buys Organics? | (use the answer sheet) |
| 6d | What is the probability that a 64 year old woman with affluence grade 4 buys Organics? | (use the answer sheet) |
| 6e | What is the probability that a 40 year old woman with affluence grade 13 buys Organics? | (use the answer sheet) |

**Compute and Evaluate Chi-Squared Statistics**

Consider the following based on a different data set than what you have done so far in this assignment.

1. Compute the Chi-Squared statistic for the following potential split variables:
(Note that you’ll need to construct the “expected” distributions for each variable to come up with the Chi-Squared statistic!)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Observed for PromSpend (total dollars spent at store) |  |  | Observed for PromTime (months as loyalty card member) |
|  | <50 | >=50 |  |  |  | <6 | >=6 |  |
| Buy | 520 | 730 | 1250 |  | Buy | 370 | 880 | 1250 |
| No Buy | 480 | 770 | 1250 |  | No Buy | 630 | 620 | 1250 |
|  | 1000 | 1500 | 2500 |  |  | 1000 | 1500 | 2500 |

**ANSWER:** PromSpend\_\_\_\_\_\_\_\_\_\_\_ PromTime \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

1. Which variable is a stronger differentiator (PromSpend or PromTime) with regard to whether a consumer buys organics?

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (use the answer sheet)

Answer Sheet for Assignment 7

Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 *Note that the item numbers correspond to the assignment instructions.*

|  |  |
| --- | --- |
| Item Number | Answer |
| 1 (answer) |  |
| 1(screen shot) |  |
| 2 (answer) |  |
| 2 (screen shot) |  |
| 3 (answer) |  |
| 3 (screen shot) |  |
| 4 (answer) |  |
| 5 (answer) |  |
| 6a (answer) |  |
| 6b (answer) |  |
| 6c (answer) |  |
| 6d (answer) |  |
| 6e (answer) |  |
| 7 (answer) |  |
|  |  |
| 8 (answer) |  |