

Department of Supply Chain and Business Technology Management

John Molson MBA Program

**BSTA 678**

**Data Mining Techniques**

Winter 2024

Assignment No. 2

Due April 1

Solve each question individually, discuss with the team and submit one common solution as a team. Don’t partition the work! Instead, everyone should work on every question, then discuss your work together and finalize the submission. Describe who did what in the below. Indicate % contribution for each member of the team. Your mark will be awarded accordingly. In case of disagreement, contact professor.

Write your answers in the Word document between the questions. Copy/paste as needed.

Submit to Moodle as a one zipped file containing EM project, Excel and Word document with your answers.

Assignment No. 1

Name / % contribution / Question you worked on

1. \_Shubham Todi\_\_ \_100%\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_all\_\_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_

In questions below use SAS Academic HUB course Applied Analytics Using SAS Enterprise Miner 15.1 (or 14.2)

Question 1.

Review Lesson 6. Model Assessment – complete and submit the Practice from 6.

1. Connect all models in the ORGANICS diagram to a Model Comparison node.

Solution:

A screenshot of a computer

Description automatically generated

1. Run the Model Comparison node and view the results.

Solution:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Which model was selected? Based on what criteria?

Solution:

A screenshot of a computer

Description automatically generated

Answer: Decision tree model was selected.

We choose from selection criterion- model with lowest validation misclassification rate is best model and hence decision tree was selected.

1. Which model has the best ROC curve?

Solution:

A graph showing a line

Description automatically generated with medium confidence

The Decision Tree has best ROC curve as it seems to have the highest/largest area under curve from the rest of the curves. i.e. most outer curve line on ROC plot.

Note: We use Validation data for all our comparisons

1. What is the corresponding ROC Index?

Solution:

The fit statistics in Model Comparison node results shows us “ROC Index” values for all models:

A grey rectangular object with many colored objects

Description automatically generated with medium confidence

The “ROC Index” can also be seen by selecting View 🡪 Model 🡪 Statistics Comparison in the Model Comparison Node Results window.

Look for the Data Role: Valid and Statistics Label: Valid: ROC Index.

A screenshot of a computer

Description automatically generated

Best ROC index is 0.824 shared by Decision Tree (2), Neural and Decision Tree.

3- Open the exported data from the Model Comparison node. Explore the RANK data set. What is the number of event cases for each model at a selection depth of 5%?

Solution:

We click on 3 dots beside exported data in Model Comparison node settings: A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Click on “Browse” while highlighting RANK row from the explorer.

A table appears with many columns, we go to depth column and sort the RANK data set created by the model comparison node by increasing depth.

The “number of events” column is what we are refereeing to in the question for top 5% depth all on Validation data.

A screenshot of a computer

Description automatically generated

Question 2.

Review Lesson 7. Model Implementation – complete and submit the Practice from 7.

1. Create a Score data source for the ScoreOrganics data set.

Solution:

* Select File > New > Data Source.
* Select SAS Table
* Select created library – Vle\_sas, which has imported data from the course
* Click on Scoreorganics table and press OK

A screenshot of a computer

Description automatically generated

* Proceed to final steps of the Data Source Wizard by clicking Next.

A screenshot of a computer

Description automatically generated

* Select “Score” as the role.
* Click Next > Finish.

1. Score the ScoreOrganics data set. Use the model that was selected with the Model Comparison node.

Solution:

* Bring “Score” Node from “Assess” tab to the “organics\_outage” diagram.
* Connect the “Model Comparison” node to the “Score” node.
* Bring and connect the “ScoreOrganics” data source to the “Score” node.
* Run the Score node.
* Browse the exported data from the Score node to confirm the scoring process.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Question 3.

Review Lesson 8.2. Introduction to Pattern Discovery – complete and submit the Practice from 8.2.

The DUNGAREE data set gives the number of pairs of four different types of dungarees that were sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans that were sold.

A table with text on it

Description automatically generated

1. Create a new diagram in your project. Name the diagram Jeans.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Define the data set DUNGAREE as a data source.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Select Ok and then next till Step 8 of Data wizard and then finish.

Bring the Data source as node to the diagram.

A screenshot of a computer

Description automatically generated

Select A screenshot of a computer

Description automatically generated Advanced in Advisor

1. Determine whether the model roles and measurement levels assigned to the variables are appropriate. Examine the distribution of the variables.

* Are there any unusual data values?
* Are there missing values that should be replaced?

A blue box with black text

Description automatically generated

Data roles and levels may be views by this.

A screenshot of a computer

Description automatically generated

We select and highlight the variables rows and click on explore to know more:

A screenshot of a computer

Description automatically generated

Graphical representation of data source:

A screenshot of a computer

Description automatically generated

We can basically see all the variables histogram plot normally distributed.

The “Store ID” seems odd as it has no normality and constant distribution over entire range as it is an ID column which we will later adjust in roles.

Even “total pairs of dungarees sold” column, it is basically a dependent variable or in terms of Data mining - a target variable, its plot is somewhat skewed.

There do not seem to be any missing values, but we can check exact % missing values from “Sample statistics” window.

Statistical representation of data source:

A screenshot of a computer

Description automatically generated

0% missing values for all variables present in the data source.

1. Assign the variable STOREID the model role ID and the variable SALESTOT the model role Rejected. Make sure that the remaining variables have the Input model role and the Interval measurement level. Why should the variable SALESTOT be rejected?

We revisit the data source node – edit variables Ids and select following roles:

A screenshot of a computer

Description automatically generated

As clear in previous parts “salestot” is the “total pairs of dungarees sold” and it is not an input variable rather a dependent variables and it should be rejected because it is the sum of the other input variables in the data set and is not to be considered as an independent input value for cluster analysis.

1. Add an Input Data Source node to the diagram workspace and select the DUNGAREE data table as the data source.

To create the data source as data table, drag the DUNGAREE data source onto the diagram workspace.

A screenshot of a computer

Description automatically generated

(As done in previous part already to show analysis of variables)

1. Add a Cluster node to the diagram workspace and connect it to the Input Data node.

A screenshot of a computer

Description automatically generated

Cluster node is in Explore.

1. Select the Cluster node. Leave the default setting as Internal Standardization > Standardization. What would happen if inputs were not standardized?

A screenshot of a computer

Description automatically generated

The cluster basically forms group based on shortest Euclidian distance between different points(records). The distance is measured in terms of distance between 2 points in space taking all input variables values. If we do not standardize the data, the clustering is dominated by the inputs with the largest range.

1. Run the diagram from the Cluster node and examine the results. Does the number of clusters created seem reasonable?

A screenshot of a computer

Description automatically generated

Normally, 4-5 or even max 6-7 clusters seem sufficient. The Cluster node’s Automatic number of cluster specification method seems to generate an excessive number of clusters which is in the tune of 20.

1. Specify a maximum of six clusters and rerun the Cluster node. How does the number and quality of clusters compare to that obtained previously?

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Apparently, all but one of the segments is well populated. All other 5 clusters are fairly distributed in large groups. It is a better model compared to last over categorised clusters which might be of no use to determine anything of use.

1. Use the Segment Profile node to summarize the nature of the clusters.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Segment 6 contains stores selling a higher-than-average number of original jeans, but lower-than-average number of stretch and fashion, with equitable in leisure.

A graph of two people

Description automatically generated with medium confidence

Segment 4 contains stores selling a higher-than-average number of leisure jeans and lower-than-average in other styles of jeans.

A graph of a person and person

Description automatically generated with medium confidence

Segment 2 contains stores selling a higher-than-average number of stretch jeans and lesser-than-average number of leisure jeans of all.

A graph of a bar graph

Description automatically generated with medium confidence

Segment 5 contains stores selling a higher-than-average number of fashion jeans and less than-average number of original jeans rest are equal number distributed.

A graph of a bar graph

Description automatically generated with medium confidence

Segment 1 contains stores selling a higher-than-average number of original jeans except for leisure type jeans style.

A graph with red and blue bars

Description automatically generated

Segment 3 contains stores selling small numbers of all jeans’ styles with no stretch type.

Question 4.

Reconstruct and submit the Banking Segmentation Case Study (A1).

Case Study Description

A consumer bank sought to segment its customers based on historic usage patterns. Segmentation was to be used for improving contact strategies in the Marketing Department.

A sample of 100,000 active consumer customers was selected. An active consumer customer was defined as an individual or household with at least one checking account and at least one transaction on the account during a three-month study period. All transactions during the three-month study period were recorded and classified into one of four activity categories:

* traditional banking methods (TBM)
* automatic teller machine (ATM)
* point of sale (POS)
* customer service (CSC)

A three-month activity profile for each customer was developed by combining historic activity averages with observed activity during the study period. Historically, for one CSC transaction, an average customer would conduct two POS transactions, three ATM transactions, and 10 TBM transactions. Each customer was assigned this initial profile at the beginning of the study period. The initial profile was updated by adding the total number of transactions in each activity category over the entire three-month study period.

The PROFILE data set contains all 100,000 three-month activity profiles. This case study describes the creation of customer activity segments based on the PROFILE data set.

Note: The diagram containing this analysis is stored as an XML file which is included with the course data. You can open this file by right-clicking Diagrams and selecting Import Diagram from XML in SAS Enterprise Miner. All nodes in the opened file, except the data node, contain the property settings outlined in this case study. If you want to run the diagram, you need to re-create the case study data set using the metadata settings indicated below.

Case Study Data

A table with text on it

Description automatically generated

Data source named: “PROFILE” was defined and imported as new node in a new diagram named: “a1” with above mentioned metadata variable roles and levels.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

We explore the inputs:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A plot of the input distributions showed highly skewed distributions for all inputs. The statistics showed no missing values.

It would be difficult to develop meaningful segments from such highly skewed inputs. Instead of focusing on the transaction counts, it was decided to develop segments based on the relative proportions of transactions across the four categories. This required a transformation of the raw data.

A Transform Variables node was connected to the PROFILE node.

A diagram of a phone with a cable connected to it

Description automatically generated with medium confidence

The Transform Variables node was used to create category logit scores for each transaction category:

category logit score = log(transaction countin category / transaction countout of category)

Transformations

The transformations were created using these steps:

* Run the node to see interactive mode tools, Select … beside “Formulas” in the Transform Variable node's Properties panel. The Formulas window appears.

A drawing of a sign and a line

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Click the Create icon as indicated above. The Add Transformation dialog box appears.

A screenshot of a computer

Description automatically generated

Build Formula for log transformation of CNT\_TBM category

A screenshot of a computer

Description automatically generated

- Click OK to add the transformation. The Add Transformation dialog box closes and you return to the Formula Builder window.

A screenshot of a computer

Description automatically generated

* For each transaction category such as CNT\_ATM, CNT\_CSC, CNT\_POS create & enter the name and formula for same log transformations.

A screenshot of a computer

Description automatically generated

And hit enter.

Run the Transform Variables node.

A screenshot of a computer

Description automatically generated

Segmentation was to be based on the newly created category logit scores. Before proceeding, it was deemed reasonable to examine the joint distribution of the cases using these derived inputs. A scatter plot using any three of the four derived inputs would represent the joint distribution without significant loss of information.

A three-dimensional scatter plot was produced using the following steps.

1. Select Exported Data from the Properties panel of the Transform Variables node. The Exported Data window appears.

2. Select the TRAIN data and select Explore the Explore window appears.

3. Select Actions  Plot or click (the Plot Wizard icon). The Plot Wizard appears.

4. Select a three-dimensional scatter plot.

5. Select Role X, Y, and Z for LGT\_ATM, LGT\_CSC, and LGT\_POS,

6. Click Finish to generate the scatter plot.

A graph showing a number of blue squares

Description automatically generated

The scatter plot showed a single clump of cases, which makes this analysis a segmentation (rather than a clustering) of the customers. There were a few outlying cases with apparently low proportions on the three plotted inputs. Given that the proportions in the four original categories must sum to 1, it followed that these outlying cases must have a high proportion of transactions in the non-plotted category, TBM.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Three changes to the Cluster node default properties were made, as indicated in the following image. One change is related to standardization of inputs and two are related to limiting the number of clusters created to 5.

Note: Because the inputs were all on the same measurement scale (category logit score), it was decided to not standardize the inputs. Thus, the property Internal Standardization was changed to None.

Only the four LGT inputs defined in the Transform Variables node were set to Default in the Cluster node.

A screenshot of a computer

Description automatically generated

Running the Cluster node and viewing the Results window confirmed the creation of five nearly equally sized clusters.

A screenshot of a computer

Description automatically generated

A Segment Profile node attached to the Cluster node helped interpret the contents of the generated segments.

Only the LGT inputs were set to Yes in the Segment Profile node. Rest was set to no or default as selected. Settings of segment profile node was also left at default values:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The following profiles were created for the generated segments:

A graph of a graph

Description automatically generated with medium confidence

Segment 5 had a higher than average rate of customer service contacts and point-of-sale transactions. This segment was labeled Service.

A graph with a red line

Description automatically generated

Segment 2 customers had a higher than average use of traditional banking methods but were close to the distribution centers on the other transaction categories. This segment was labeled Transitionals because they seem to be transitioning from brick-and-mortar to other usage patterns.

A graph of a graph

Description automatically generated with medium confidence

Segment 3 customers eschewed traditional banking methods in favor of ATMs. This segment was labeled **ATMs**.

A graph of a graph

Description automatically generated with medium confidence

Segment 4 was characterized by a high prevalence of point-of-sale transactions and few traditional bank methods. This segment was labeled **Cashless**.

A graph of a bar graph

Description automatically generated with medium confidence

Segment 1 customers had a significantly higher than average use of traditional banking methods and lower than average use of all other transaction categories. This segment was labeled **Brick-and-Mortar.**

Deployment of the transaction segmentation can be facilitated by the Score node.

The Score node was attached to the Cluster node and run. The SAS Code window inside the Results window provided SAS code that will be capable of transforming raw transaction counts to cluster assignments by attaching the score data set to the score node and obtain results as and when required.

A close-up of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

Question 5.

Reconstruct and submit the Enrollment Management Case Study (A4).

In the fall of 2004, the administration of a large private university requested that the Office of Enrollment Management and the Office of Institutional Research work together to identify prospective students who would most likely enroll as new freshmen in the Fall 2005 semester. The administration stated several goals for this project:

* increase new freshmen enrollment
* increase diversity
* increase SAT scores of entering students

Historically, inquiries numbered approximately 90,000 or more students, and the university enrolled from 2400 to 2800 new freshmen each Fall semester.

Case Study Training Data

A table with text and images

Description automatically generated with medium confidence

A table of information

Description automatically generated

A screenshot of a calendar

Description automatically generated

The Office of Institutional Research assumed the task of building a predictive model, and the Office of Enrollment Management served as consultant to the project. The Office of Institutional Research built and maintained a data warehouse that contained information about enrollment for the past six years. It was decided that inquiries for Fall 2004 would be used to build the model to help shape the Fall 2005 freshman class. The data set Inq2005 was built over a period of a several months in consultation with Enrollment Management. The data set included variables that could be classified as demographic, financial, number of correspondences, student interests, and campus visits. Many variables were created using historical data and trends. For example, high school code was replaced by the percentage of inquirers from that high school over the past five years who enrolled. The resulting data set included more than 90,000 observations and more than 50 variables. For this case study, the number of variables was reduced. The data set Inq2005 is in the AAEM library, and the variables are described in the table above. Some of the variables were automatically rejected based on the number of missing values.

The nominal variables ACADEMIC\_INTEREST\_1, ACADEMIC\_INTEREST\_2, and IRSCHOOL were rejected because they were replaced by the interval variables INT1RAT, INT2RAT, and HSCRAT, respectively. For example, academic interest codes 1 and 2 were replaced by the percentage of inquirers over the past five years who indicated those interest codes and then enrolled. The variable IRSCHOOL is the high school code of the student, and it was replaced by the percentage of inquirers from that high school over the past who enrolled. The variables ETHNICITY and SEX were rejected because they cannot be used in admission decisions. Several variables count the various types of contacts the university has with the students.

Solution:

Data is defined and inserted into the diagram a4.

Some stats are studied: target variable counts have a high percent of zeros.

A screenshot of a computer

Description automatically generated

A table of text with numbers

Description automatically generated with medium confidence

A group of black text

Description automatically generated

A white paper with black text

Description automatically generated

Only 3.1 % of the target values are ones, making a value of 1 a rare event. Standard practice in this situation is to separately sample the zero s and ones. The Sample tool enables us to create a stratified sample in SAS Enterprise Miner.

* avg\_income and distance have missing values.

The Explore window was used to study the distribution of the interval variables.

The apparent skewness of all inputs suggests that some transformations might be needed for regression models.

A screenshot of a computer

Description automatically generated

Sample node settings:

A screenshot of a computer

Description automatically generated

Result:

A screenshot of a computer screen

Description automatically generated

The primary purpose of the predictions was decision optimization and, secondarily, ranking. An applicant was considered a good candidate if his or her probability of enrollment was higher than average.

Because of the Sample node, decision information consistent with the above objectives could not be entered in the data source node. To incorporate decision information, the Decisions tool was incorporated in the analysis.

Decision node setting:



A screenshot of a chat

Description automatically generated

A screenshot of a computer

Description automatically generated

The nonzero values used in the decision matrix are the inverse of the prior probabilities (1/. 8. and 1/0.875 1.142857). Such a decision matrix, sometimes referred to as the central decision rule, forces a primary decision when the estimated primary outcome probability for a case exceeds the primary outcome prior probability (0.125 in this case)

Models were built to best predict the outcome.

Two rounds of predictive modeling were performed. In the first round, all cases were considered for model building. From the Decision node, partitioning, imputation, modeling, and assessment were performed. The completed analysis appears as shown.

A screenshot of a computer

Description automatically generated

The Data Partition node used 60% for t raining and 40% for validation.

• The Impute node used the Tree method for both class and interval variables. Unique missing

indicator variables were al so selected and used as inputs.

• The stepwise regression model was used as a variable selection method for the Neural Network and second Regression nodes

• The Regression node labeled Instate Regression included the variables from the Stepwise

Regression node and the variable Instate. It was felt that prospective students behave differently based on whether they are in state or out of state.

Model Comparison:

Neural networks model wins marginally over Decision Tree model and the 2 regressions. Compared overall on average profit for enrolment. The fit statistics show no overfitting of data and very low error rates and misclassification rates for validation data.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

The validation ROC chart showed an extremely good performance for all models. The neural model seemed to have a slight edge over the other models. This was mirrored in the Fit Statistics table, where the neural network has a validation ROC Index value of 0.981.

It should be noted that an ROC Index of 0.98 needed careful consideration because it suggested a near-perfect separation of the primary and secondary outcomes. The decision tree model provides some insight into this apparently outstanding model fit. Self-initiated contacts are critical to enrollment. Fewer than three self-initiated contacts almost guarantee non-enrollment.

Creating Prediction Models (Instate Only Cases)

A second round of analysis was performed on instate only cases. The analysis sample was reduced using the Filter node. The Filter node was attached to the Decisions node, as shown below (Only a portion of the flow is shown to assist in viewing.)

A close-up of a computer screen

Description automatically generated

Filter settings:

A screenshot of a computer

Description automatically generated

Select Class Variables and click ... After the path is updated, the Interactive Class Filter window appears.

Select Generate Summary and then click Yes to generate summary statistics

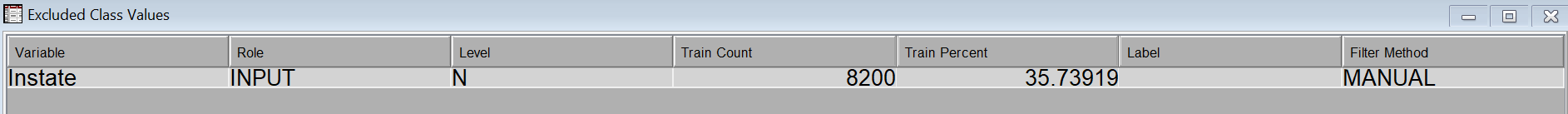
A screenshot of a computer

Description automatically generated

Select the N bar and select Apply Filter

Click OK to close the Interactive Class Filter window.

Run the Filter node and view the results.



After filtering, an analysis similar to the above was conducted with stepwise regression, neural network, and decision tree models.

The partial diagram (after the Filter node) is shown below:

A diagram of a tree

Description automatically generated

Assessing Prediction Models (Instate Only Cases)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

As before, model performance was gauged in the Model Comparison node

The ROC chart showed no clearly superior model, although all models had rather exceptional performance.

A white page with black text

Description automatically generated

The Fit Statistics table of the Output window showed a slight edge for the tree model in misclassification r ate T he validation ROC index and validation average profit favored the Stepwise Regression and Neural Network models. A gain, it should be noted that these were unusually high model performance statistics.

Deploying the Prediction Model

The Score node facilitated deployment of the prediction model, as shown in the diagram’s final form. (Only a portion of the flow is shown):

A screenshot of a computer screen

Description automatically generated

The best (instate) model was selected by the Instate Model Comparison node and passed on to the Score node. Another INQ2005 data source was assigned a role of Score and attached to the Score node. Columns from the scored INQ2005 were then passed into the Office of Enrollment Management’s data management system by the final SAS Code node.

A close-up of a sign

Description automatically generated



A white rectangular object with a black border

Description automatically generated

Run the last node.