CIS-445 DATA MINING

Project 2

Due: See Blackboard

Worth 100 points

Objectives

Learn about **Commercial Data Mining Tool: SAS Enterprise Miner (EM) 14.1**

Task: Classification/Prediction of the Binary Outcome, Scoring
Models: Neural Network, Logistic Regression, Decision Tree
Nodes: File Import, Neural Network, Decision Tree, Regression, Model Comparison, Score, Reporter, SAS code (extra credit)
Software: SAS 9.4, SAS Enterprise Miner 14.1

Important Notes

Use user-friendly names for the project name and the diagram name. Save the project on the J drive.

Project

1. You will be using the same small data set that we used in class when we discussed decision trees. (See Fig. 10.1 on p. 147 in ClassPack.) The data set contains 7 variables. There are 6 independent/input variables in columns 1 through 6. Columns 1 and 3 represent variables Age and Residence, respectively. Column 6 represents variable Income taking two nominal values "low" and "high" only. The remaining variables X2, X4, and X5 in columns 2, 4, and 5, respectively, are not meaningful. The dependent/target variable WidgBuy is in column 7 and it takes two values only: "Yes" or "No". Your task is to build 3 simple models: the decision tree model, neural network model, and logistic regression model. Then you need to evaluate and compare the performance of the three models using the correct classification accuracy rates, confusion matrices as well as receiver operating characteristics (ROC) charts and lift charts. Finally, you need to use the best model to score each of the 9 cases in the new data set for which the class label for variable WidgBuy is unknown.

2. Using Excel 2013 or 2016 create two data sets named: *WidgBuyTrain.xlsx* and *WidgBuyScore.xlsx*. See the next page for the two data sets. Save the data sets on the J drive in the project directory. Make sure that all data below is typed correctly! Also, type the first row that contains the names of the 7 attributes of the first data set and 6 attributes of the second data set. The latter file has no attribute WidgBuy. The names of these attributes will be transferred to SAS EM as the variable names. The first data set *WidgBuyTrain.xlsx* will be used to train/build the three models, and the best model of the three will score each of the 9 cases/customers in the second data set *WidgBuyScore.xlsx* to determine if the customer is a widget buyer or not.

Use the File Import node to import *WidgBuyTrain.xlsx*. Use the File Import option on the properties panel to locate the file. Follow the wizard. All variables will initially have the Role Input by default. Change the Role of variable WidgBuy to Target and the Level to Binary. Explore the distribution of all variables. On the File Import window click on the Preview tab to see the file. After you run the node, the Excel file has been read and converted to SAS data set EMWS1.FIMPORT_DATA, where EMWS1 is the name of the library on which the data set resides and FIMPORT_DATA is the SAS data set name. SAS named the data set and the library because you did not have an option to provide your own user-defined names. One can see the name if one clicks on Imported Data in the Property Panel. The Role of the data set will be Train by default. See the Property Panel of the File Import node.

WidgBuyTrain.xlsx

Age	X2	Residence	X4	X5	Income	WidgBuy
37	2	NY	0.8	4	low	Yes
32	1.3	LA	0.7	5	high	No
44	1.5	NY	0.9	8	high	No
37	0.3	NY	0.8	9	high	No
28	0.6	CHI	0.6	2	low	No
29	1.1	NY	0.6	3	high	Yes
33	2	NY	0.8	2	low	Yes
45	1.5	LA	0.4	7	high	No
22	0.3	CHI	0.5	5	high	No
33	2	NY	0.8	4	low	Yes
28	0.6	LA	0.6	7	high	No
34	0.9	NY	0.3	6	low	Yes
27	0.3	NY	0.9	2	high	Yes
33	0.6	NY	0.1	4	low	Yes
45	0.9	CHI	0.7	5	high	No
33	2	LA	0.6	8	low	Yes
22	1.8	NY	0.3	2	high	Yes
36	0.3	NY	0.1	5	low	Yes
44	0.1	NY	0.9	4	high	No
33	0.3	NY	0.6	2	low	Yes

WidgBuyScore.xlsx

Age	X2	Residence	X4	X5	Income
35	2	LA	0.8	4	Low
30	1.3	NY	0.7	5	Low
45	1.5	NY	0.9	8	High
40	0.3	NY	0.8	9	High
25	0.6	LA	0.6	2	Low
34	1.1	CHI	0.6	3	High
37	2	CHI	0.8	2	High
40	1.5	LA	0.4	7	High
27	0.3	CHI	0.5	5	Low

Use the Decision Tree, Neural Network, and Regression nodes to build the models, and then the Model Comparison node to assess the quality of the models. Unlike in Tutorials 1 and 2, you will not use the Data Partition node here because we will use all 20 cases for the training set to build the models. In other words, the data set is too small and there are no cases left for the validation set, which would help to build (and often test) the models; and for the test set, which would provide the unbiased real-world performance of the models for cases not seen/used during training. There is no need for any kind of variable transformation, binning, or imputing missing values in this simple project. Accept all the default settings provided by the regression node. SAS EM examined the Level of the target variable as Binary and automatically determined that Logistic Regression (LR) is needed here. Recall from the class notes that LR calculates the probability that the customer is a widget buyer or not. Use the Entropy technique (the same technique that was used in the classroom and homework #5 examples) for the decision tree for the Nominal and Ordinal Target Criterions on the Property Panel. Also, use a neural network with 1 neuron in a hidden layer (the default is 3 neurons).

Import Excel data set *WidgBuyScore.xlsx* and use the best model to score it by the Score node. On the Property Panel of the File Import node remember to change the Role of the data set to Score. For extra credit add and run the SAS Code node to see exactly which of the 9 cases are classified as Widget Buyers, and which are not, and with what probability. The standard cutoff=0.5 is used to classify the cases. At the end add the Reporter node (mandatory), run it, and view/scan a multipage pdf file which contains the results from each node of your workflow.

For each case/customer, each of the 3 models generates the probability that the customer is a widget buyer or not. Different operating cut-off points can be selected for analysis by the Cut-off node (not used here). They affect the correct and incorrect classification rates and the number of false positive and false negative errors. Assuming that the target event is detecting customers who are widget buyers, the standard 0.50 cut-off point means that if the tool generated the value greater than or equal to 0.50, a customer is classified as a widget buyer; otherwise he/she is not.

In the Receiver Operating Characteristic (ROC) charts, the strength of the models is demonstrated by the degree to which the curves push upward and to the left. Curves located in the upper right corner and the lower left corner represent lower and higher cut-offs, respectively. To be able to interpret these charts in a more subtle way, one would have to understand how these charts were constructed. ROC charts are also briefly discussed on p. 67 in your Course Pack. If time allows, we will briefly talk about them in class and the example of the ROC chart is posted on BB in the Course Document and Assignments/Projects folders.

The Lift and ROC charts give more insight into the global strength and performance of the models for a continuum of cut-off points within the range [0, 1]. The charts are available in the Model Comparison node. On each of the two charts, you should have 3 models plotted, i.e., the decision tree, neural network, and logistic regression. When you have 3 curves, one for each model, plotted on a single chart, it is easier to compare the performance of the models. Note that in this project, however, the neural network and logistic regression perform equally well for all cut-off points. Therefore, on the charts you will see two curves only as the curves for the neural network and logistic regression exactly overlap themselves. The curves look somewhat ragged because of the too few data points used to construct them.

My hints stop here as I have explained the project thoroughly in class.

Submit via e-mail

1. A 3-page single spaced report describing the project assumptions and the results. Use the 1" top and bottom margins and 1" left and right margins on the page, and font 12 Times Roman. From the Results and Output windows for each node create and discuss the confusion matrices for each model. Include these matrices in your written part of the report. These confusion matrices are created for the 0.5 cut-off. Interpret the results: the confusion matrices as well as the ROC and Lift charts, the rules generated by the decision tree, the importance of variables, and the effects generated for the logistic regression coefficients. Which variables have the most predictive power? Has the logistic regression model identified the same variables as the decision tree in terms of their predictive power? Examine the final weights of the neural network. Can you make any sense of them? How many of the 9 cases from the WidgBuyScore.xlsx data set were classified as Widget Buyers and Non-Widget Buyers? For extra credit describe the results from the SAS Code node in terms of the probability with which each of the 9 cases was classified. Attach the following items to your report.
    a. Workflow/diagram with all nodes you used.
    b. The tree diagram generated by the decision tree which shows the variables, branches, and nodes. (From the diagram one can easily develop the classification rules.)
    c. The window with the rules generated by the decision tree.
    d. The table with the relative importance of the variables used in the decision tree.
    e. Lift and ROC charts for the 3 models.
    f. The window with the final weights for the neural network.
    g. The chart with the effects for the regression model.
    h. The output with the probabilities from the SAS Code node (for extra credit)
2. Merge all items from point 1 and 1.a-1.h in a single file named *Project2_YourLastName_YourFirstName.**pdf*** and submit it via e-mail to jozef.zurada@louisville.edu on or before the due date.