Student Name: Brandon Russell

Submission Date: 6/17/18

DBST 667 – Data Mining

Dr. Irene Tsapara

Week 4 Individual Exercise

Deliverables: Two Files: (1) Submit this lab report with answers to all questions including output screenshots into the ‘Individual Exercises Week 4’ assignment folder. (2) Submit an R script that contains all commands with comments that briefly describe each commands purpose.

Grading: This exercise is worth 2% of the course grade. All questions must be answered in your own words with any paraphrased references properly cited using in-text citations and a reference list as needed. In addition, grammatical and spelling errors may affect the grade.

Part 2 – Run an exercise on the credit approval dataset you used for the week 2 exercise last week, completing this report and providing the commands, output screenshots, and discussion/interpretation as requested. Ensure that all commands are saved in this report AND in an R script.

a. Introduction:

i. Identify the dependent variable and independent variables in the Credit Approval data set.

The dependent variable is "class". This is apparent for a few reasons, even with the attributes being masked from their original meaning. Being the last column is one indicator this column is dependent on the others. There are only two factors, "+" or "-", providing further indication this is a dependent yes/no

ii. Based on what you have learned this week about decision trees, provide a one-paragraph masters-level response describing what you anticipate that the ctree (conditional inference tree) method will accomplish for the Credit Approval data? Be specific about the behavior and structure of tree-based models.

b. **Data Pre-Processing: Load the credit approval data into R Studio using the read.csv command.**

   i. **What data pre-processing (if any) does the ctree method require for the credit approval data? Include the commands you ran and the output screenshot.**

   Command(s): >

   Output:

c. **Training and Test Data:**

   i. **Why do we split the dataset into training and test data? Provide a one-paragraph, masters-level response in your own words.**

ii. **Run the set.seed(1234) command and the commands from the tutorial to split the Credit Approval data into a training set containing 70% of the observations and a test set containing the remaining 30% of the observations. Include the commands and output screenshot.**

Commands: >

Output:

d. **CTree Method:**

i. **Run the command to build the ctree model and store the model in a variable called 'credit_ctree'. Include the command.**

Command: >

ii. **Run the print(credit_ctree) command to output the model in its textual format. Include the command and output screenshot.**

Command: >

Output:

iii. **Using the textual output of the 'credit_ctree' from above, interpret the structure of the model including the splitting and leaf nodes in your discussion. Provide a one-paragraph, masters-level response.**

iv. **Based on the results of your ctree model, which independent variables could be significant predictors of the dependent variable? Justify why you made this conclusion by supporting your answer with specific references to the ctree model itself. Provide a one-paragraph, masters-level response.**

e. **Visualize the CTree Model – Run plot(credit_ctree) to build the tree plot. Include the command, the plot output screenshot, and a one-paragraph, masters-level interpretation of the tree structure:**

Note: Use the available options from plot() to customize and enhance the appearance of your visualized tree.

Command: >

Output:

Interpretation:

f. **Confusion Matrix for the Training Data:**

   i. **Build the confusion matrix for the training data. Include the command and matrix output screenshot.**

   Command: >

   Output:

ii. **What is the classification accuracy for the training data? Provide the complete formula used (i.e. show your work) along with the final percentage (rounded to two decimals places).**

iii. **Other than classification accuracy, what other information can the confusion matrix provide? Hint: consider misclassification and the rates which your model actually predict correctly or incorrectly. To get you started: calculate the rate of true positive predictions by taking (True Positives)/(False Negatives + True Positives).**

g. **Evaluate the CTree Model on Test Data:**

i. **Build the confusion matrix for the test data. Include the command and matrix output screenshot.**

Command: >

Output:

ii. What is the classification accuracy for the test data? Provide the complete formula used (i.e. show your work) along with the final percentage (rounded to two decimals places).

iii. Compare the classification accuracy of the training data with the classification accuracy of the test data. Go beyond a simple "test data classification accuracy is greater than/less than training data classification accuracy" by comparing the actual false, actual true, predicted false, and predicted true totals.

h. New Instance – Now that your model is built, suppose that you now have a batch of new credit applications coming in for your model to process (newcredit.csv). You know all of the independent variables but you want your model to predict if these new applications will be approved or disapproved. Import newcredit.csv and provide the command you would use to predict the unknown dependent variable using your "credit_ctree" model along with the output screenshot showing the predictions. Hint: the function() you need was used several times in the tutorial:

Command: >

Output:

i. Summary

i. **What differences did you observe between the Apriori association rules method and the Conditional Inference ctree method? Compar and contrast at least two characteristics from each method. Provide a one-paragraph, masters-level response.**

ii. **Do missing values affect the results of the ctree model? Provide a one-paragraph, masters-level response.**

iii. **(Not graded) Which part of this exercise did you find the most challenging and what steps did you take to resolve the challenge?**

References