CS 4331 – Data Mining

Spring 2020

Assignment 2 - 8 Points

March 2, 2020

Full Name: _____ Class Section:_____

Acknowledge your collaborators or source of solutions, if any. **Submission by 3/10/2020 is required.** This exercise is for getting familiar with *Preparing to Model the Data and Decision Trees*. For coding, use Python 3 or R. A subset of your answers will be graded. Include your .py or .r file with helpful comments included and no hard-coded file paths (only the file name should be present when reading or writing the file). All files should be submitted in a zip file and named CS4331_00?_eraiderlogin_Vn.zip where ? is replaced with your class section number, eraiderlogin is your eraiderlogin, and n is replaced with the version. You may submit up to 3 times and only the last version will be graded.

1. Using the churn data set, determine how many records need to be resampled in order to have 20% of the rebalanced data set have true C*hurn* variable values. Create the rebalanced data set and confirm that the new data set has 20% true C*hurn* variable values.

2. Partition the rebalanced data set so that 67% of the records are included in the training data set and 33% are included in the test data set. Use a bar graph to confirm the proportions. Validate the training and test data sets by testing for the difference in the training and test means using *Day.Mins* (t-test) and the Z-test on the *Churn* variable. Try forming new training and test sets if there is enough evidence to reject the null hypothesis.

3. Create a CART model using the training set with the *Churn* target variable and whatever predictor variables you think appropriate. Try at least 3 different models. Compare the confusion tables and accuracies of the 3 different models.

4. Create a C5.0 model using the training set with the *Churn* target variable and whatever predictor variables you think appropriate. Try at least 3 different models. Compare the confusion tables and accuracies of the 3 different models. How does C5.0 compare in performance to CART?

Source: Chantal D. Larose, Daniel T. Larose, Data Science Using Python and R

All assignments will be checked for academic misconduct (cheating, plagiarism, collusion, falsifying academic records, misrepresenting facts, violations of published professional ethics/standards, and any act or attempted act designed to give unfair academic advantage to oneself or another student) defined by "OP 34.12: Grading Procedures, Including Academic Integrity" of TTU.