BTMA 531 Assignment 2

Due Feb 15, 2023 by noon on D2L

Instructions

- You should create a single R script (or R Markdown if you wish) called {firstname}_{lastname}_Asgn2.r, which has the required code for all parts of the assignment.
- Make sure to use commenting (#) so that your R script file can be understood by someone else. Yo do not need to comment on what you are trying to do in each line, but it should be clear where the answer for each question is. Provide the written answers in comments (#) within the R script file.
- Make sure that your R script is executable from top to bottom on another computer. Make sure all requirements of questions are done using R (e.g. do not use calculator, excel, . . . to calculate things).
- The assignment submission should be done through D2L dropbox. Upload the R file to the assigned dropbox folder in D2L.
- The purpose of the assignments is to help you learn through practice. You may want to use R help or search online for answers. However, note that this is an individual assignment. The work you submit should be 100% yours. Do not copy, share, or ask for files, chunks of code, or answers. Refer to the course outline for some examples of what to do and what not to do, and to Code of Student Conduct for more information on cheating and plagiarism. If you are note sure about a behavior, please ask.
- The focus is on the response to the question, and not on the approach you use to get there. However, unless otherwise stated, you should use only code to get to the response (that is, you cannot use a calculator for example to do some intermediate calculations, everything needs to be done in R).
- You are allowed to use online resources to find answers or chunks of code. However, make sure that you reference the source in your script using commenting (#). This is generally a good practice in coding, and for the purpose of this course, removes any confusion when several students use the same online source and therefore have similar codes in their scripts.

Questions

1 [25] Use the attached “transaction_dataset.csv” for this problem. This is a modified version of the dataset from https://www.kaggle.com/vagifa/ethereum-frauddetection-dataset. The goal is to make a model that predicts whether a transaction made on the Ethereum platform (a type of cryptocurrency and a platform for smart contracts) is fraudulent. The variable that shows whether a transaction is fradulent is called *FLAG* in the dataset. Read the data to an object called “ether”.

a) [7.5] Create a logistic regression model on a random set of 2,000 observations in the dataset that classifies the fraudulent status of a transaction based on all inputs except the index. Use *set.seed(1)* before sampling the training set for consistency.

b) [7.5] Use the logistic regression classifier you created to predict the classes for the remaining (test) observations in the dataset. Do the proper transformation so that your predicted results show the predicted class using the Bayes boundary.

c) [5] Calculate the prediction accuracy of your classifier for the test set and draw the confusion matrix. Calculate the false positive rate (Type I error) and false negative rate (Type II errror).

d) [5] What do you think about the performance of your classifier? How would you modify the classifier to improve its performance when predicting fraudulent transactions? [Note: this approach may not actually work in this case given our logistic regression, but that is fine!]

2 [30] Use the attached "accent-mfcc-data-1.csv" dataset for this question. This dataset includes 11 different Mel-frequency cepstrum (MFC) attributes on soundtracks of different people reading words. The MFC data is often used in sound processing. In this application, the goal is to predict the native language of the speaker from six European languages (ES, FR, GE, IT, UK, US) based on the MFCs. This dataset is based on https://archive.ics.uci.edu/ml/datasets/Speaker+Accent+Recognition.

a) [5] Create an LDA classifier on a random set of 250 observations in the dataset that classifies the *language* based on all MFC inputs. Use *set.seed(1)* before sampling the training set for consistency.

b) [2.5] Use the LDA classifier you created to predict the classes for the remaining 79 (test) observations in the dataset.

c) [2.5] Calculate the prediction accuracy of your classifier for the test set and draw the confusion matrix.

d) [2.5] Now create a QDA classifier on the same random set of 250 observations as before that classifies the *language* based on all MFC inputs.

e) [2.5] Use the QDA classifier you created to predict the classes for the remaining 79 (test) observations in the dataset.

f) [5] What is the accuracy of predictions for the QDA for the test set? How does QDA accuracy compare to the LDA model you created before? What does this comparison say about the structure of the data? What additional steps could you take to make sure your answer is correct?

g) [10] Use KNN to predict the language using all MFC inputs. Use the same random set of 250 observations as before for training. Create two models, one with k=5 and one with k=10. What are the accuracies for these two models? Explain what you see when comparing K=10 and K=5.

3 [25] The attached "CarEvals.csv" dataset includes data on conditions and evaluations of second hand cars. There are 6 input variables, and an outcome variable called "Class" (This is a modified dataset based on https://archive.ics.uci.edu/ml/datasets/Car+Evaluation).

a) [5] Create a classification tree for classifying the "class" variable based on the other variables. Plot the tree.

b) [5] Create another tree using only a 1000 observations from the dataset, selected randomly. Use *set.seed(1)* for consistency. Predict the classes for the rest of the observations (719 observations).

c) [5] Calculate the accuracy of the predictions and draw the confusion matrix.

d) [5] Use cross-validation to find the best size of the tree. Plot the cross-validation error and the tuning parameter versus tree size. What is the best tree size?

e) [5] Prune the tree to the best size found in part d. Calculate the accuracy of the newly created model with the rest of the observations (719 observations).

4 [20] Using the attached "AirQualityUCI" dataset from https://archive.ics.uci.edu/ml/datasets/Air+Quality:

a) [5] Add the data to R as an object called "AQdata". Then create a new object from the time series where observations that have an *NA* in any of the variables have been removed. Call this new object $AQdataFULL$.

b) [5] Given that this data is the hourly data for air quality, are there any issues with removing observations with missing values? Why? What would you do in this case?

c) [5] Create a sequence plot, a lag plot, and a histogram of the True hourly averaged concentration CO data from $AQdata$ (use $AQdataFULL$ only if the plot cannot be created with missing values). What are your observations based on these plots?

d) [5] Create a Q-Q plot and Q-Q line for comparison of True hourly averaged concentration CO data with the normal distribution from *AQdata*. What do you think about the distribution of data based on this plot?