BTMA 531 Assignment 2

Due Feb 14, 2024 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 [30] Use the attached "online_shoppers_intention2.csv" dataset for this question. This data is a slightly modified version of the data from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.

a) [10] Create a logistic regression model on a random set of 10,000 observations in the dataset that classifies the revenue status of a shopper based on all inputs. Use *set.seed(99)* before sampling the training set for consistency. Based on the regression results (and ignoring any possible statistical issues with the model), how much more or less likely is it for a returning visitor to make a purchase compared to a new visitor?

b) [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) [7.5] Calculate the prediction accuracy of your classifier for the test set and draw the confusion matrix. Calculate the specificity and sensitivity of the predictive model.

d) [7.5] What do you think about the performance of your classifier? How would you modify the classifier to improve its performance when predicting purchasers (revenue = True)? Try your approach to achieve such a model and derive the Type I and Type II errors.

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(99)* 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 [20] 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) [2.5] Create another tree using only a 1000 observations from the dataset, selected randomly. Use *set.seed(99)* for consistency. Predict the classes for the rest of the observations (719 observations).

c) [2.5] 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? Use *set.seed(99)* for consistency.

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] Consider the attached "BankMarketingSample.csv" data sample from direct phone marketing to encourage clients to subscribe to term deposit service. Here is the data dictionary of the dataset:

bank client and call data:

1 - age (numeric)

2 - job : type of job (categorical)

3 - marital : marital status (categorical)

4 - education (categorical)

5 - default: has credit in default? (categorical: “no”,“yes”,“unknown”)

6 - housing: has housing loan? (categorical: “no”,“yes”,“unknown”)

7 - loan: has personal loan? (categorical: “no”,“yes”,“unknown”)

8 - duration: contact duration, in seconds (numeric)

9 - voicemail: did the call go to voicemail? (categorical: “no”,“yes”,“unknown”)

social and economic context attributes

10 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

11 - cons.price.idx: consumer price index - monthly indicator (numeric)

12 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

13 - euribor3m: euribor 3 month rate - daily indicator (numeric)

14 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

15 - y - has the client subscribed a term deposit? (binary: “yes”,“no”)

a) [5] Consider that you want to build a classification model in order to aid you with which clients to target for the direct marketing. What classification model(s) would be appropriate to create such a predictive model? Explain.

b) [7.5] Consider the type I and type II errors in targeting clients. Describe these errors in this context. What do you think about the costs of each error? Which is more costly? How would you reduce the error type with the higher cost?

c) [7.5] In creating a realistic predictive model for classification to use for the purpose of deciding who to target for phone calls, is there any variable(s) in this dataset that would not be useful for prediction? Explain.