Institiúid Teicneolaíochta Chorcaí
Cork Institute of Technology

Data8001

Lecturer : Aengus Daly

Assignment

Due - Monday 2nd Dec 2019, 11.59 pm

Please submit your work via Canvas ,((1) your .R script with comments, (2) your answer in .Rmd Markdown, (3) html, Word, pdf or similar format). Do not zip your files. You are also required to hand-up a hard copy of your answer to the lab on Tuesday 3rd Dec, 11am, C125, no binding or plastic cover is required.

Standard CIT penalties apply for work submitted after the due date.

Your submission should be your own work, plagiarism will be dealt with in accordance with CIT regulations.

Note this assignment is worth 40% of this module. Reference your work appropriately.

Annotate your code with comments especially for code that is complicated; marks will be given for these comments that display understanding of all the code you use, including code given in labs and class.

Marks are awarded for code that is succinct and neat and for the labelling of variables in a meaningful and clear manner. Marks are also awarded for answers that have a level of individual though and expression, so add these were possible in your comments.

Provide one .R and .Rmd file for part a), b), c) , d), e) and another one for parts f), g), h). Name these files FirstName_Surname_DA with appropriate extension.

Question 1

Kate, a manager at a financial institution has contacted you. She is asking you for assistance in assessing the credit worthiness of future potential customers. She has a data set of 793 past loan customer cases, with 14 attributes for each case, including attributes such as financial standing, reason for the loan, employment, demographic information, foreign national, years residence in the district and the outcome/label variable *Credit Standing* - classifying each case as either a good loan or bad loan.

The manager has 10 new customers, which she would like to know if she should consider them good or bad prospective loans.

The data sets are up on Canvas, in the assignment folder and called Credit_Risk6_final.xls, in 2 different sheets, Training_Data and Scoring_Data (The Scoring set is the 10 unlabelled potential loan customers.)

Data Details

Most of the attributes are self-explanatory; the name of some of the attributes are somewhat cumbersome but this is what you have been given; here are the further details of some of them:

Checking Acct - What level of regular checking account does the customer have –*No acct*, *0balance*, *low* (balance), *high* (balance)

Credit History – All paid – no credit taken or all credit paid back duly
Bank Paid – All credit at this bank paid back
Current – Existing loan/credit paid back duly till now
Critical – Risky account or other credits at other banks
Delay – Delay in paying back credit/loan in the past

Months Acct – The number of months the customer has an account with the bank.

Using R help Kate to answer the following questions.

a) Exploratory Data Analysis (EDA): - Carry out some EDA on the data set; carry out at least one trivariate analysis; do you notice anything unusual or any patterns with the data set? Detail these and outline any actions you propose to take before you start model building in part b).
Max word count 500 words.

10 marks

b) Build a decision tree model and give your decision tree, detailing its parameters. Explain how you decided on/fined tuned these parameters. (Include an image of your tree as well as a text output description.). Use set.seed(abc) where abc are the last 3 digits of your student no. Use this set.seed for all other model building below.

5 marks

c) Use the decision tree to predict results for the scoring set. Choose 5 different potential loan clients and explain to Kate in plain English how the decision tree works (15 marks) and how the accuracy/probabilities of these being a good/bad loan was calculated by the decision tree, outling your assumptions (5 marks).
Max word count 500 words.

20 marks

d) Now try and improve your model using 2 other approaches, e.g. ensemble technique, boosting or a different model. Explain your training/validation/testing methodology. Comment on your results and analyse why your model is giving better/worse results.

10 marks

e) Kate's company uses a process that is a mixture of a grading system and human input to grade each past loan as good or bad. Kate is suspicious that during a particular time that this process performed very poorly and produced inaccurate results. Develop a strategy so that you can you find a series of consecutive or nearly consecutive ID numbers of circa 10 or more, i.e. where these gradings show a suspiciously incorrect pattern. Detail how you go about your investigation and how you find this pattern.

10 marks

f) Develop a InfoGain algorithm that works on this dataset to calculate the variable for the first split. You may use the code developed in the labs as a starting point but make sure to annotate your code with comments explaining what it is doing. Note you can only used base R commands here no other packages are allowed. Comment on your results.

15 marks

g) Develop code in R that illustrates how boosting works using the formulae for adabag in the attached document. Use the Excel spreadsheet attached so that you have only ten data points. Use set.seed(abc) with abc being the last 3 digits of your student number to generate a random prediction (each time) for 4 iterations of boosting. Include a confusion matrix at the end for your final prediction and comment. Note you can only used base R commands here no other packages are allowed.

15 marks

h) Generate prediction probabilities obtained in your best model aboveand use R code to create and plot an ROC curve, note you can only used base R commands here no other packages are allowed. Comment on the ROC curve.

15 marks

[Total 100 marks]