MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2024: Assignment 2

Your task	• The objective of this assignment is to gain familiarity with classification models using R. • This is an individual assignment.
Value	• This assignment is worth **30%** of your total marks for the unit. • It has 35 marks in total.
Suggested Length	• 8 – 10 A4 pages (for your report) + extra pages as appendix for your R script. • Font size 11 or 12pt, single spacing.
Due Date	**11.55pm Monday 20th May 2024**
Submission	• Submit a single PDF file and single video presentation file on Moodle. • Use the naming convention: *FirstnameSecondnameID.{pdf, mp4, mov etc.}* • Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	• In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.
Late Penalties	• 10% (3 mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Instructions and data

The objective of this assignment is to gain familiarity with classification models using R. We want to create models that may be used to predict whether or not a website will be legitimate or designed for phishing – that is, stealing personal data from users.

You will be using a modified version of the PhiUSIIL Phishing data, hosted by the UCI Machine Learning Archive https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset. A research paper based on this data is available here https://doi.org/10.1016/j.cose.2023.103545.

There are two options for compiling your written report:
(1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
(2) As an R Markdown document that contains the R code with the discussion/text interleaved. Render this as an HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

Creating your data set

Clear your workspace, set the number of significant digits to a sensible value, and use '**Phish**' as the default data frame name for the whole data set. Read your data into R and create your individual data using the following code:

```
rm(list = ls())
Phish <- read.csv("PhishingData.csv")
set.seed(XXXXXXXX) # Your Student ID is the random seed
L <- as.data.frame(c(1:50))
L <- L[sample(nrow(L), 10, replace = FALSE),]
Phish <- Phish[(Phish$A01 %in% L),]
PD <- Phish[sample(nrow(Phish), 2000, replace = FALSE),] # sample of 2000 rows
```

Questions (10 Marks)

1. Explore the data: What is the proportion of phishing sites to legitimate sites? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-valued attributes. Is there anything noteworthy in the data? Are there any attributes you need to consider omitting from your analysis? **(1 Mark)**

2. Document any pre-processing required to make the data set suitable for the model fitting that follows. **(1 Mark)**

3. Divide your data into a 70% training and 30% test set by adapting the following code (written for the iris data). Use your student ID as the random seed.

```
set.seed(XXXXXXXX) #Student ID as random seed
train.row = sample(1:nrow(iris), 0.7*nrow(iris))
iris.train = iris[train.row,]
iris.test = iris[-train.row,]
```

4. Implement a classification model using each of the following techniques. For this question you may use each of the R functions at their default settings if suitable. **(5 Marks)**

- Decision Tree
- Naïve Bayes
- Bagging
- Boosting
- Random Forest

5. Using the test data, classify each of the test cases as 'phishing (1)' or 'legitimate (0)'. Create a confusion matrix and report the accuracy of each model. **(1 Mark)**

6. Using the test data, calculate the confidence of predicting 'phishing' for each case and construct an ROC curve for each classifier. You should be able to plot all the curves on the same axis. Use a different colour for each classifier. Calculate the AUC for each classifier. **(1 Mark)**

7. Create a table comparing the results in Questions 5 and 6 for all classifiers. Is there a single "best" classifier? **(1 Mark)**

Investigative Tasks (18 Marks)

8. Examining each of the models, determine the most important variables in predicting whether a web site will be phishing or legitimate. Which variables could be omitted from the data with very little effect on performance? Give reasons. **(2 Marks)**

9. Starting with one of the classifiers you created in Question 4, create a classifier that is simple enough for a person to be able to classify whether a site is phishing or legitimate by hand. Describe your model with either a diagram or written explanation. What factors were important in your decision? State why you chose the attributes you used. Using the test data created in Question 3, evaluate model performance using the measures you calculated for Questions 5 and 6. How does it compare to those in Question 4? (**4 Marks**)

10. Create the best tree-based classifier you can. You may do this by adjusting the parameters, and/or cross-validation of the basic models in Question 4. Show that your model is better than the others using the measures you calculated for Questions 5 and 6. Describe how you created your improved model, and why you chose that model. What factors were important in your decision? State why you chose the attributes you used. **(4 Marks)**

11. Using the insights from your analysis so far, implement an Artificial Neural Network classifier and report its performance. Comment on attributes used and your data pre-processing required. How does this classifier compare with the others? Can you give any reasons? **(4 Marks)**

12. Fit a new classifier to the data, test and report its performance in the same way as for previous models. You can choose a new type of classifier not covered in the course, or a new version of any of the classifiers we have studied. Either way, you will be implementing a new R package. As a starting point, you might refer to James et al. (2021), or look online. When writing up, state the new classifier and package used. Include a web link to the package details. Give a brief description of the model type and how it works. Comment on the performance of your new model. **(4 Marks)**

Report and Video Presentation (7 Marks)

Write a brief report (suggested length 8 – 10 pages) summarizing your results. Use commenting in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown. **(3 Marks)**

Record a short presentation using your smart phone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings, as well as describing how you conducted your research and any assumptions made. Pay particular emphasis to your results for the investigative tasks. **(Submission Hurdle and 4 Marks)**

Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but please document these in your report and include in your R code.

Description of the data

Attributes A01:A25	Website Attribues
Class	The target attribute. Whether the web site is phishing or legitimate.

References

An Introduction to Statistical Learning with applications in R, 2nd Ed, 2021. (Springer Texts in Statistics), James, Witten, Hastie and Tibshirani. (Available on-line from Monash Library.)