DATA7202 Assignment 1 (Weight: 25%)

Due: 28/3/2018.

Analysis of the UCI bank telemarketing dataset: exploratory data analysis, prediction, evaluation and inference with generalized linear models.

The dataset and associated information can be found at
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

We will focus on the expanded: bank-additional.zip dataset.

The main aim of this analysis is to predict whether or not the client will agree to make a long-term deposit with the bank. A related, although less important outcome, is to predict the length of the phone call. These two tasks will be attempted using different forms of regression analysis.

The dataset has around 41,188 observations. I would like you to also create a random subsample of the data containing 2000 observations (sampled without replacement), and attempt to build predictive models for the same two variables.

Manipulate the data as required to make the analysis possible. There are many missing values, which will have to be handled thoughtfully.

The data has already been analysed by Moro et al (2014), who obtained the data from a Portuguese retail bank and uploaded it to the UCI database.

1. Give your interpretation of the bank's aims in asking for this dataset to be analysed. What statistical tasks would these map to?

2. Perform exploratory data analysis as relevant to the construction of predictive models and the statistical tasks in your answer above. Investigate and highlight any apparent structure in the data. Also investigate the assumptions of the regression models and determine whether any transformations would be useful, and whether there are any outliers or influential points of note. You should do this only for the whole dataset and assume that much of what is learned applies to the subsample.

3. Use a general linear model to attempt to predict call duration based on the other variables which it is reasonable to include. Make an argument as to why you do not use some variables.as explanatory variables. You should attempt to check all assumptions of the model and report on this. You are free to attempt to transform the data or re-represent it in any way desired. However, be aware that repeated attempts to do this count as model selection and are thus likely to incur some selection bias with respect to any performance measure, unless efforts are made to quarantine data for evaluation. See e.g. Wood et al. 2007 for some discussion.

4. Use a logistic regression model to attempt to predict whether or not the client takes up a term deposit, based on the other variables which it is reasonable to include. You should

attempt to check all assumptions of the model and report on this. You are free to transform or re-represent the data as in the previous question.

5. Give relevant equations and assumptions to describe the general linear model and the logistic regression model used above. Describe the methods used to fit these models. Include derivation of the maximum likelihood estimate of the beta vector in the general linear model, and any assumptions.

6. Evaluate the predictive performance of each of the above two models using three appropriate metrics. Include details of each metric and its advantages and disadvantages. Where relevant, compare to the results of Moro *et al*. (2014).

7. Given the results of your analysis, and leaving the external variables at their sample averages, what would be the characteristics of (plausible explanatory variable values for) a client who has the highest predicted probability of taking up a term deposit Also – what would be the best method and time to contact them? Create a corresponding ideal client (represented by a set of explanatory variable values) and use the models to predict the call length and probability of success for this client.

8. Give confidence intervals and the results of statistical testing for all parameters or contrasts which seem relevant given your answer to question 1. Explain these results as relevant to the problem, including a section aimed purely at executives in the company. Feel free to revise your answer to question 1 while answering this. Where relevant, compare to the results of Moro *et al*. (2014).

Notes:

Where possible, give reasons for your answers. Please store all the R commands you use in a separate file and submit that also. Please include your name in the filename for all files submitted. You should not generally give R commands in your main report and should not include any raw output – i.e. just include figures from R (each with a title, axis labels and caption below) and put any relevant numerical output in a table or within the text.

As per http://www.uq.edu.au/myadvisor/academic-integrity-and-plagiarism, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Equations are either correct or not, but you should use consistent notation throughout your assignment, define all of it and ensure that your report flows logically.

You are asked to use the R software environment for this assignment. This is available on all computers in the Maths Department and is also free to install on any of your own computers. Information and downloads are available from http://www.r-project.org/ . Rstudio https://www.rstudio.com/ is a quality free interface for R.

Submit your assignment via TurnItIn on Blackboard. In this include the report (with graphs and tables included) and any R programs or scripts that you write to answer the assignment.

Acceptable formats for the report include a Word document or a pdf file. I plan to print each report out in greyscale to mark it, so please make sure that the colours in plots will be distinguishable from each other even in greyscale format.

Some References:

W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth Edition, Springer, 2002.

J. Maindonald and J. Braun, *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd edition, Cambridge University Press, 2010 (available online via UQ library).

S. Moro, P. Cortez. and P. Rita, A data-driven approach to predict the success of bank telemarketing, *Decision Support Systems*, 62, 22-31, 2014.

I.A. Wood, P.M. Visscher, K.L. Mengersen, Classification based upon gene expression data: bias and precision of error rates, *Bioinformatics* 23 (11), 1363-1370, 2007.