ITNPBD6 Data Analytics Assignment 2020

Computing and Maths
University of Stirling

The banks have had a difficult few years and have been finding that people no longer trust them with their money. *StirBank* are keen to get people saving with them again, so they have been running a marketing programme, but this involves calling many customers and is both costly and risks annoying people. This is where you come in. We have got some data from the bank describing 8000 of its customers and the previous attempts to make marketing calls to them. The data also tells us whether or not each customer responded to the marketing by setting up a regular savings deposit "made_deposit".

The question is simple – is there any way to predict which customers are more likely to respond positively to a marketing call? Your assignment is to answer that question using data mining techniques and produce a system that would be able to tell the bank which customers it should target the marketing at.

You can use Orange, Python, R, or any data mining package of your choice. The data for the assignment is in a file bank-tr.*csv* included with this document.

You should hand in a report covering the following:

Introduction [10 marks]
Describe the task you were given, the data you received and the requirements of the finished system, including why data mining is suitable for this task rather than a more conventional . Define any terminology that you will use in the report (for example, model, variable, task, etc.).

Data Summary [10 marks]
List the variables that you found in the file provided by the company. For each one, say whether it is nominal or numeric, continuous or discrete and whether or not it is of use in building the solution. Explain your decisions.

Data Preparation [10 marks]
Describe what you did with the data prior to the modelling process. Show histograms of the data before and after any pre-processing that you carried out. If you corrected any mis-typed entries in the data, report what you changed (either specific changes or what rules you used to correct the data).

Background [30 marks]
The bank would like to better understand the different approaches to data mining. In particular, they have heard of three model types: decision trees, multi-layer perceptrons and logistic regression. Give a detailed technical description of each technique and the way the models are represented, how they learn and make predictions, and how easy it would be for the bank to understand the model and the

reasons behind each prediction it makes. Also describe what parameters may be changed and what effect this has. Include one diagram showing the structure of each type of model.

Modelling [25 marks]

You must build models using the three different techniques above, and choose one to recommend to the bank for supporting its targeted marketing. If you varied the parameters of a model, show how this impacted on the results. Describe how you split the data for training and testing purposes. Be methodical and record each result. This stage is a little like scientific research – you are carrying out experiments in your search for the best solution. Once you have a solution, show how you verified its robustness. For the three different techniques report on their comparative ability to predict a positive response from a customer.

Results and Errors [15 marks]

Analyse and describe the level of accuracy your chosen model achieves and the errors your model makes. Show a confusion matrix for your model. Are there any areas of the data where it performs worse than in others? Show a lift curve or a ROC curve for the decision as to whether or not a customer will respond to a marketing call.

Submission

The deadline for this assignment is 4pm on Friday 3 April 2020. Please submit your report via the Assignments space on Canvas as a doc or pdf file bearing your university username (3 letters + 5 digits, e.g., xyz00001.pdf).

You do not need to submit the models that you built, just the report. As a guideline, your report should be around 3000-5000 words. However, this is not a strict limit and no penalties will be issued for reports outside this range – just write what you need to provide the required information clearly and concisely. You can assume that the client has a good technical understanding of data mining and statistics, so do not shy away from technical terms in your report. Where you use them, however, explain what they mean in plain language too. To maximise your mark, make sure you follow the instructions above and include everything that is asked for in the report.

Plagiarism

Work which is submitted for assessment must be your own work. All students should note that the University has a formal policy on plagiarism which can be found at http://www.quality.stir.ac.uk/ac-policy/assessment.php. You can test your report prior to submission using the “Similarity Checking Space” on Canvas.

This assignment is subject to the usual grade penalties for late submission. You can email questions about it to sbr@cs.stir.ac.uk .