 Data Analytics Assignment Summer 2021

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 customers and the previous attempts to make marketing calls to them. The data also tells us if 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* provided for you. A data dictionary is provided describing the columns in this file.

# Submission Requirement:

You should hand in a report describing the modelling process you followed and your results. You should attempt to frame the problem in the form of CRISP-DM framework to better facilitate the discussion. Refer to the relevant CRISP-DM stages at each stage of your report. You do not need to submit code or data. The report is worth 100 marks in total and must cover the following:

# Business Understanding [10 marks]

Describe the task you were given: is it classification or regression?; describe the data you received and the requirements of the finished system, including why data mining is suitable for this task. 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 should be treated as categorical or numeric; nominal, ordinal, continuous or discrete; and whether or not it is likely to be of use in building the solution. Explain your decisions: if you rule out any variables at this stage, you can justify your choice using summary statistics, or a histogram plot of its distribution.

# Data Preparation [15 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. (you do not need to give histograms of all variables, just the ones that need some cleaning) If you corrected any mis-typed or corrupted entries in the data, report what you changed, such as any rules you used, or examples of specific data points that were cleaned.

# Modelling [40 marks]

You must use three different techniques and build models with each: these should include one tree-based model, one based on logistic regression, and one based on neural networks. Try to make each model perform as well as it can: if you varied the hyperparameters of a model, show which hyperparameters you varied and how this impacted on the results. Describe how you split the data for training, validation 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 store performance, but only select a single model for the final test.

Don't try to find a perfect or extremely accurate model - one does not exist! We are interested in the procedure you followed and the justification you give for choosing particular model types/parameters/features.

# Evaluation [15 marks]

Analyse and describe the level of accuracy the 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, and are there any types of error that StirBank would want to avoid more than others? Show a lift curve or a ROC curve for the decision as to whether or not a customer will make a deposit.

# Submission

The deadline for this assignment is 12 noon on Friday 30 July 2021. Please submit your report via Canvas at <https://canvas.stir.ac.uk/courses/9654/assignments/63418>. You should save it 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 you built, just the report. Your report should be around 3000 words, but this is not a strict limit and no penalties will be issued for longer/shorter reports. 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 avoid technical terms in your report, but where you use them please explain what they mean in plain language too. To maximise your mark, closely follow the instructions above and include everything that is asked for.

# Note on Avoiding Academic Misconduct:

**Work which is submitted for assessment must be your own work.** All students should note that the University has a formal policy on Academic Integrity and Academic Misconduct (including plagiarism) which can be found at <https://bit.ly/37fYxPw>

Plagiarism: We are aware that assignment solutions by previous students can sometimes be found posted on GitHub or other public repositories. Do not be tempted to include any such code in your submission. Using code that is not your own will be treated as “poor academic practice” or “plagiarism” and will be penalized.

To avoid the risk of your own work being plagiarised by others, do not share copies of your solution, and keep your work secure both during and after the assignment period.

Collusion: This is an individual assignment: working together with other students is not permitted. If students submit the same, or very similar work, this will be treated as "collusion" and all students involved will be penalized.

Contract cheating: Asking or paying someone else to do assignment work for you (contract cheating) is considered gross academic misconduct, and will result in termination of your studies with no award.

Late submissions: This assignment is subject to the usual grade penalties for a late submission. You can email questions about it to [sbr@cs.stir.ac.uk](mailto:sbr@cs.stir.ac.uk).