# Data Mining Assignment 2019

Division of Computer Science & Maths
University of Stirling

The banks are having a bit of trouble with debt at the moment. They have lent lots of money to people who promised to pay it back, and then didn't. In the future, they would like to avoid lending to the kind of person who won't pay back the loan, and that is where you come in. We have got some data from a bank describing 2000 of its loan customers. The data also tells us whether or not each customer repaid the loan.

The question is simple – Is there a difference between the people who repay the loans and those who don't? Your assignment is to answer that question using data mining techniques and produce a system that would be able to tell the bank how likely it is that a new customer would pay back a loan.

You should use the Weka data mining package, which is installed in the lab in 4X5 and is available to download from:
http://www.cs.waikato.ac.nz/~ml/weka/

Your report should cover the following:

### Introduction

Describe the task you were given, the data you received and the requirements of the finished system. Define any terminology that you will use in the report (for example, model, variable, task, etc.).

### Data Summary

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

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.

### Modelling

You must use two different techniques and build models with both: pick a suitable tree building algorithm and also use a multi-layer perceptron. Describe the different methods you used and the results that you got. Give a brief technical description of the techniques and the way the models are represented. Include one diagram showing the structure of each type of model that you build. Describe what parameters may be changed and what effect this has.

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 two different techniques report on their comparative ability to predict a defaulted loan, and also on how easy it would be for the insurance company to understand the model and the reasons behind each prediction it makes.

## Results and Errors

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

## Submission

Check the course web site for the submission deadline. You will need to submit your work 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. There is not a word limit on the report – 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.

This assignment is worth 50% of the overall grade for the course, and is subject to the usual grade penalties for late submission. This assignment is set by Dr Jingpeng Li. You can email questions about it to jli@cs.stir.ac.uk.