CSCU9T6 Data Mining Assignment 2021

University of Stirling

This assignment is designed to reproduce a commercial data mining consultancy project following the CRISP-DM methodology. You are provided with a file of data and required to build a series of machine learning models and then report on your findings.

The client is a bank, and the project aim is to build a classifier capable of saying whether or not a potential loan customer will pay back the loan. The data provides a description of previous loans customers and a field that says whether or not they repaid the loan.

You can use any software of your choice (for example, Orange or scikit learn in Python) and you will not be required to submit any code, just a report. You should employ best practice for both the project management and the machine learning aspects of the project. The data you need for the project is also available on the course Canvas page.

Your report should follow the CRISP-DM methodology and should also document the correct methodology for training a machine learning model. It should have the following sections:

Business Understanding 10 Marks

Describe the task you were given, the data you received and the requirements of the finished system. Explain why it is a suitable task for a data mining approach. Define any terminology that you will use in the report (for example, model, variable, task, etc.). Describe the project methodology you will use.

Data Understanding 10 Marks

List the variables that you found in the file. For each one, say whether it should be treated as nominal or numeric, continuous or discrete and whether or not it should be considered for building the solution. Identify the inputs and outputs to the model. Explain your decisions.

Data Preparation 10 Marks

Describe what you did with the data prior to the modelling process. Show histograms of one example variable before and after any pre-processing that you carried out. If you corrected any mis-typed entries in the data, report what you changed. Describe what scaling you performed and how much test data you separated out.

Modelling 50 Marks

Now you must build some models from the data. Pick three suitable techniques and build a number of models using each one. For each technique, explore different values for hyperparameters using an appropriate validation technique. Describe how you used that technique.

Describe what hyperparameters were explored and what effect this had. A summary table is the best way to present these results. 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 chosen the best model, train a final solution using all the training data and report the appropriate performance metrics.

Results and Errors 20 Marks

Now test the final model on your test data. Analyse and describe the level of accuracy the model achieves and the errors your model makes. Show a confusion matrix for the model. Are there any areas of the data where it performs worse than in others?

Submission

The deadline for submission is Friday, April 9th. Upload your report via canvas by the deadline. Your report should not be more than 3000 words long.

You do not need to submit the models that you built, just the report.

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.

Feedback

You will receive feedback and a grade by April 30th.

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 Kevin Swingler. You can email questions about it to kms@cs.stir.ac.uk.