CSCU9M5 Machine Learning Assignment 2024

University of Stirling

This assignment is designed to reproduce a commercial machine learning 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 the owner of a chain of shops, and the project aim is to build a classifier capable of saying whether or not a new store will be profitable. The data provides a description of existing stores and a variable that classifies the store performance as good or bad. You can download the file from Canvas.

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.

Your report should follow the CRISP-DM methodology and should also document the correct methodology for training a machine learning model – particularly the use of train, validate and test data and the method of searching hyperparameters. It should have the following sections:

Business Understanding

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 machine learning approach. Define any terminology that you will use in the report (for example, model, variable, task, etc.). Describe the project methodology you will use. Discuss how the model might be used to improve the business and what impact different types of error might have on the effectiveness of the model.

Data Understanding

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

Describe your test data separation process. 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 or recoding you performed.

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

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

Submission details are on Canvas. Upload your report to Canvas by the deadline. Your report should not be more than 3000 words long. Marks will be allocated in line with the University common marking scheme. That means that to gain a first, you must show creative thinking and insight. Reflect on the decisions you make, and justify them, ideally with references to the literature (or other sources). Simply following the correct procedure is sufficient for a 2:1 grade at most.

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 machine learning 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.

This assignment is worth 50% of the overall grade for the course and is subject to the usual grade penalties for late submission.