# ITNPBD6 Data Analytics Assignment Spring 2021: Marking Scheme

Computing and Maths
University of Stirling

The assignment will be marked on the report, not the technical results. This document describes marks out of 100, which will be scaled appropriately to grades.

### 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 approach. Define any terminology that you will use in the report (for example, model, variable, task, etc.).

- Statement of the task: 3 marks. Give both marks for any reasonable description. Zero marks if this is missing
- Terminology: Up to 4 marks for defining at least two terms (model, learning …)
- Justification: 3 marks for explaining why this is a suitable task for data mining

### Data Summary 10 marks

List the variables that you found in the file provided by the company. For each one, say whether it is 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.

- 5 marks for correct summary of the variable types and data quantity.
- 2 marks for correctly identifying the variables that will not be of use for building a model.
- 3 marks for reasons supporting those choices such as they all have the same value or all have unique values.

### Data Preparation 15 marks

Records with missing or obviously erroneous data should either be corrected (with the assumptions for doing so clearly stated) or deleted.

- Test set removal – this should happen first. 4 marks for separating test data, 2 extra for doing it first.
- Data cleansing – 3 marks for listing changes made, 3 for a histogram illustrating how one particular problem was identified and fixed.
- Scaling – 3 marks

### 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 parameters of a model, show which hyperparameters you varied 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.

- 24 marks for applying the models; 8 each for trying three techniques and reporting the results for each
- 4 marks for clear description of validation method (training,test and validation data or CV)
- 8 for a systematic exploration of hyperparameters; table presenting results with method, hyperparameters and validation metric given
- 4 marks for picking a final solution, training on all the data and reporting the results

## 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 an ROC curve for the performance of the store.

- 4 marks for reporting the test results, not those of training or validation – this must be stated explicitly to get the marks
- 4 marks for showing and describing the confusion matrix properly
- 4 marks for a correct ROC or lift curve with correct explanation
- 3 marks for some reference to the application: are false positives or negatives more important here?

**Overall Approach 10 marks**

**10** marks for adopting a structured approach to the whole process and identifying the six CRISP-DM stages.

## Feedback

Note where marks are lost by annotating the document or adding comments. Include a final comment highlighting one or two main topics the student should re-visit to improve their understanding of the topic.