Orindary Degree in Computing: Data Mining.

Data Mining Assessment:
Mine a dataset

Submitted by: Name, Student number

Submission date

Declaration

I herby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ordinary Degree in Computing in the Institute of Technology Blanchardstown, is entirely my own work except where otherwise stated.

Author: _____ Dated: _____

Table of Contents

. . . . .

Business Understanding (Max 1 page)

Give a brief background to the dataset itself (1 paragraph).

In a single bullet point: state what to you want to accomplish when mining this data.

Data Mining objective
In one or more bullet points: State your technical objectives for mining the data.

Data Understanding (max 6 pages including diagrams).

Describe the data
For each attribute, give its description and data type. For numeric attributes, give mean, min, max and st dev; for nominal attributes with a few values, list the values.

This could be laid out in a table.

Comment on the data, rather than just putting in a screen shot from Rapidminer, to illustrate your understanding of meta data.

Explore the data
Discuss the results of an initial exploration of the data using graphs and exploratory statistics. You do NOT need to report on ALL attributes in this section, but comment on anything you found of interest, such as attributes or groups of attributes that seem predictive; correlated attributes; attributes with limit value because of too much or too little variability, attributes with unusual distributions etc. Your discussion should be with respect to your initial business and data mining objectives.

Verify data quality
Does the dataset have many missing values?

Is the presence of noise, bias or outliers likely to be an issue?

Are there sufficient attributes and examples to achieve your mining objectives?

Data Preparation (1 page max per technique tried).

Before you complete any data pre-processing, train a model on your dataset to get a baseline accuracy. For every preparation technique done, run the model again to see it's effect on model accuracy, then report on what you tried, why you tried it, whether or not it improved the model, and why you think it did / did not make a difference.

Select Data

If you need to reduce the number of rows or columns in the dataset, discuss the approaches you tried, and what worked best.

Clean Data

If data quality was an issue, discuss the approaches you tried, and what worked best.

Construct Data

Detail data transformations you tried, why you thought it would be useful, and how well they work. The report should get across the iterative nature of this phase. It should also get across that you used the results of data exploration to inform this phase, rather than randomly trying different techniques in the hope that something would work.

Modelling

Select modelling technique (1 paragraph)

Discuss which algorithms are most appropriate for the dataset and mining objectives, justify your selection.

Generate Test Design

Explain how you will generate training and test data and how you will evaluate your results. (1 paragraph)

Build and Assess the model

For each algorithm:

Detail the parameter values tried, the model generated (did you learn anything from the model itself, e.g. decision tree nodes). Discuss and interpret the

model accuracy, and if relevant, how the accuracy might be improved. Include diagrams where relevant. (1 page max per algorithm)

Evaluation (1 page max)

The purpose of this section is to document, in business, non-technical terms, what information you have learnt from the dataset. This discussion should focus on your original business objective(s), but can also include other things you have learnt along the way.