COMP8410 Data Mining 2018

Assignment 2

Maximum marks	**100**
Weight	**20% of the total marks for the course**
Length	**Maximum of 10 pages and maximum of 3,000 words, in both cases excluding cover sheet, bibliography and appendices.**
Layout	**A4 margin, at least 11 point type size, use of typeface, margins and headings consistent with a professional style.**
Submission deadline	**5:00pm, Friday, 18 May**
Submission mode	**Electronic, via Wattle**
Estimated time	**15 hours**
Penalty for lateness	**100% after the deadline has passed**
First posted:	**16th April, 9am**
Last modified:	**16th April, 9am**
Questions to:	**Wattle Discussion Forum**

This assignment specification may be updated to reflect clarifications and modifications after it is first issued.

It is strongly suggested that you start working on the assignment right away. You can submit as many times as you like. Only the most recent submission at the due date will be assessed.

In this assignment, you are required to submit a single **report** in the form of a PDF file. You may also attach supporting information (appendices) as one or more identified sections at the end of the same PDF file. Appendices will not be marked but may be treated as supporting information to your report. Please use a **cover sheet** at the front that identifies you as author of the work using your u-number and name, states the report word count, and identifies this as your submission for COMP8410 Assignment 2. The cover sheet and appendices do not contribute to the page limit or word count.

You are expected to write in a style appropriate to a professional report. You may refer to http://www.anu.edu.au/students/learning-development/writing-assessment/report-writing for some useful stylistic advice. You are expected to have both an introduction and a conclusion in your report.

No particular layout is specified, but you should follow a professional style and use no smaller than 11 point typeface and stay within the maximum specified page count and the maximum specified word count. Page margins, heading sizes, paragraph breaks and so forth are not specified but a

professional style must be maintained. Text beyond the page limit or word count limit will be treated as non-existent.

This is a single-person assignment and should be completed **on your own**. Make certain you carefully reference all the material that you use, although the nature of this assignment suggests few references will be needed. It is unacceptable to cut and paste another author's work and pass it off as your own. Anyone found doing this, from whatever source, will get a mark of zero for the assignment and, in addition, CECS procedures for plagiarism will apply.

No particular referencing style is required. However, you are expected to reference conventionally, conveniently, and consistently. References are not included in the word count and page limit. Due to the context in which this assignment is placed, you may refer to the course notes or course software where appropriate (e.g. "*For this experiment Rattle was used*"), without formal reference to original sources, unless you **copy text** which always requires a formal reference to the source.

An assessment rubric is provided. The rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work.

Your assignment submission will be treated confidentially. It will be available to ANU staff involved in the course for the purposes of marking.

Task

You are to study the supplied data set and to apply data mining processes and techniques to discover interesting things about the data. You are to write a short report that justifies and explains your methods in detail, presents your results, and evaluates and interprets the results you find. In the following, the task is described in terms of what your report should contain, not in terms of the steps you should take to carry out the assignment. In your report, similarly, you should describe the methods used in terms of the language of data mining, not in the terms of commands you typed or buttons you selected.

1. Introduce the problem

You must provide some context to the data mining project you are working on. You should properly refer to the purpose of learning and assessment for COMP8410, but in addition you should set some goals for the exercise – what do you expect to learn from the data? What are you looking for? It is possible that you may not achieve the goals you set here, but it should be possible to trace the results you present back to the goals as motivating questions. Furthermore, you should review the goals you state here in your conclusion.

2. Describe your data

You must

- identify the source of the data,

- broadly describe the attributes in the data,
- identify the population over which the data is sampled,
- offer a cursory assessment of data quality, and
- include a basic statistical summary of the data you have.

This should comprise a brief description of the data necessary to explain the context for the work presented here in a self-contained way, although for more detail it might refer to information provided with this assignment specification or elsewhere.

3. Describe your methods

You are encouraged to use **Rattle or R** for this assignment. You may use external tools instead for part or all of the work (e.g. you might prefer to use python for data pre-processing). Use of alternative tools may make your explanations of methods more wordy your methods more difficult to reproduce, and your assignment harder to mark, so take this into account. You will not be awarded marks for methods where your method cannot be understood.

You must use **at least two distinct data mining algorithms** as taught in this course. You may additionally use multiple other methods taught in this course. Further, you may choose to use some methods not addressed in this course. You must **justify** your choice of methods with reference to the data types involved, the questions you are looking to answer, the benefit of application to practice, computational feasibility, experimentation experience, or other reasons.

Application of some methods, or addressing particular questions, may require you to **pre-process** the data in some way. For example, if you are looking to predict outcomes independently of years you could consider removing year identification from the dataset. You must include either a statement that no such processing was performed or else brief information on any

- removal of provided data from consideration,
- imputation or other transformation, or
- differences in the basic data summary from that you prepared from the original data.

Data pre-processing can be a never-ending task. Be careful to exercise your judgement on how much you do here, taking account of the marking rubric.

Your description must be sufficient for a reasonably competent professional in the field to **reproduce** your major results. You may choose to attach detailed specifications or configuration parameters as an appendix (which does not contribute to the word count). If you are using methods that were not taught in the course it would normally be necessary to provide extra detail over that that can be assumed for methods taught in the course. Extended technical detail may be included in an appendix or by well-chosen references that contain enough information to implement the technique.

4. Present your results.

You must explain what you found. This should not be a complete listing of everything you found. You should select results that are interesting, surprising, explanatory, answer your initial questions, or are otherwise **meaningful,** and explain why they are meaningful. Your selected results must be supported by appropriate **formal quality measures** and must be **interpreted** within the context of the problem context you gave. Your interpretation must be pitched towards an expert in the field related to the data source but who may not be an expert in data mining. You might consider using diagrams to assist but use your judgement about any added value of diagrams.

5. Conclude with opportunities for application of your results and identification of further work.

Here you should write about the significance of your results and the challenge (or not) of using the results to make changes in the practice for which your data was collected. This analysis should be made in the context of the goals you set in your introduction, and you can afford to speculate about possible impacts of what you found.

You are not expected to be an expert in the area of application, nor to solve challenges you might raise with putting your results into practice. Identifying further work may include identifying additional data that could be used to refine the results you found, or alternative methods that should be tried with additional resources.

Assessment Rubric

This rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work. Your assignment will be marked out of 100, and marks will be scaled back to contribute to the defined weighting for assessment of the course.

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
Overall holistic evaluation of the report	10	9-10 Highly original and very interesting. Excellent, detailed and relevant discussion that develops and enhances the reader's understanding of the topic. Very clear key message and closely associated conclusion.	7-8 Interesting with some originality. Relevant discussion of sufficient detail to allow the reader to develop a clear understanding of the topic. Identifiable key message and related conclusion.	6 Interesting but lacking originality. Although mostly relevant, discussion sometimes lacks sufficient detail to allow the reader to develop a consistent understanding of the topic. Apparent key message and associated conclusion.	5 Not very interesting or original. Discussion is not always relevant nor sufficiently detailed to enable the reader to develop an understanding of the topic. Difficult to be certain what the key message is and how the conclusion relates to it	0-4 Boring and mundane. Discussion lacks detail, is mostly irrelevant and doesn't help the reader to develop an understanding of the topic. No discernible key message or conclusion.
Communication, Structure and Presentation	10	8-10 Exemplary use of language enhancing the quality of the submission. Very well ordered with logical and clear structure supported by appropriate headings and sub headings. All use of others' ideas and materials acknowledged. References are all included and are formatted	7 Very good use of language. Well-ordered and logical. Headings and sub-headings help to clarify text. All use of others' ideas and material is acknowledged. All references are included, though some minor inconsistency of in-text citation or formatting.	6 Reasonable but needs some revision. Mostly well-ordered and logical, most supported by headings and sub-headings All use of others' ideas and material is acknowledged. Some references are missing and occasional inconsistencies of in-text citation and formatting.	5 Poor, needs significant revision. Order is not always logical and is sometimes confusing. Headings are largely those suggested by the assignment specification and the questions posed. All use of other's ideas and material is acknowledged, though sometimes	0-4 Very difficult to understand. Order is confusing and not always logical. Headings and sub-headings do little to help clarify the text Not all use of other's ideas and material is acknowledged. Missing in-text citations, i.e. plagiarism. References in the bibliography not used in the

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
		consistently and appropriately. Diagrams and/or images are ideally suited to the points where they are used.	Diagrams and/or images are used effectively.	Diagrams and/or images improve readability.	inconsistently. Missing references and inconsistent in-text citation and formatting. Diagrams and/or images are not well selected.	text. Poorly and inconsistently formatted. Diagrams and/or images detract from the key messages.
Problem Description	10	9-10 Goals are clear, challenging, and suitable for the data used. Wider context of goals is discussed (e.g. expected impact or importance). The problem description provides context for the data mining that is connected and used in the rest of the work.	7-8 Problem description is clear and suitable for the data used. The problem description provides context for the data mining that is connected and used in the rest of the work.	6 The problem description provides adequate context for the mining work although some key elements could be expanded to support richer analytical work.	5 Problem description is barely adequate for the purpose. Problem description does not connect tightly with the work performed.	0-4 Key elements of the problem description are missing or insufficiently explained.
Data Description	10	9-10 Source, attributes, population, quality assessment and basic statistical summary provided. Description demonstrates deep understanding of the data.	7-8 Source, attributes, population, quality, and basic statistical summary provided.	12-13 Source, attributes, population, quality, and basic statistical summary provided. Data description should be clearer.	10-11 Most of the required information provided and correct.	0-4 Required information not provided and/or incorrect or misleading, demonstrating lack of engagement with the data.
Method description	30	24-30 At least 2 course methods applied. R, Rattle and other tools have been properly	21-23 At least 2 course methods applied. R, Rattle and other tools have been properly	18-20 At least 2 course methods applied. R, Rattle and other tools have been properly	15-17 At least 2 course methods applied. Not always clear what software tools were used.	0-14 Less than 2 course methods applied. Not clear what software tools were used

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
		identified and used appropriately. Reproduction of major results is possible from description of methods. Data pre-processing is well-suited to the methods used and the mining goals, with justification (or no-pre-processing, with convincing justification). Careful parameter setting and tuning explained and justified by experimentation or theory or both. Justification for methods chosen demonstrates careful attention to the applicability and limitations of the methods to the problem goals.	identified and used appropriately. Reproduction of major results is possible from description of methods. Extensive, directed, experimentation with data preprocessing or tuning parameters explained. Justification for methods chosen is clear and linked to problem goals.	identified and used appropriately. Reproduction of major results is possible from description of methods. Some experimentation with data preprocessing or tuning parameters evident. Justification for methods chosen is clear.	Unclear that reproduction of results is possible from description of methods. Experimentation with data pre-processing or tuning parameters barely evident, suggesting a simplistic approach to the problem. Weak justification for methods chosen.	Methods not described in adequate detail for reproduction. Justification for methods chosen absent or unconvincing.
Results	20	17-20 Outstandingly useful and potentially actionable results found. Results are clearly interpreted for domain expert.	14-16 Results presented are well selected for significance to the mining goals. Results are interpreted for domain expert. Results are well supported by selected quality measures	12-13 Major results are clearly presented, with quality measures present but overall interpretation for domain experts could be sharper.	10-11 Results are clearly presented, with typical quality measures present.	0-9 Scant attention to evaluation appropriate to the methods used.

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
		Results supported by well-selected quality measures, explained in terms of impact for domain expert.	that are explained for domain expert.			
Conclusion and further work	10	9-10 Thoughtful analysis of how results could be applied, including identifying challenges. Ideas for further work are creative, relevant and exciting, and tied to application context.	7-8 Analysis of potential application of results ties to problem goals and recognises the application context. Ideas for further work are significant and realistic.	6 Statement of how results could be applied to the domain is realistic in the context of the problem goals. Ideas for further work are present but could be better tied to the problem and application domain.	5 Statement of how results could be applied to the domain is given. Possible further work identified.	0-4 Missing analysis of application or extension.