Assignment

(Data Analytics and Business Intelligence /PG – 8696/8697)

This assessment item contributes 35% of the total marks of this unit and is due in **Week 12 on Friday at 23:59.** You should plan and complete your assignment before the deadline. The deadline is the local time of Canberra, Australia. It is your responsibility to correctly adjust your clocks, including the one in Canvas personalized user interface.

Submissions (single MS Word or pdf document) are through Canvas (http://uclearn.canberra.edu.au/). Please familiarize yourself with Canvas submissions before the deadline. After you upload your assignment report, **please DO NOT FORGET to click the submit button**; otherwise, your submitted assignment may remain as a draft.

No other forms of submission will be accepted.

Extensions

Students can apply for an extension to the submission due date for an assessment item through extenuating, evidenced circumstances (specific details are found through the Assessment Policy and Procedures. Section 9.12). Extensions must be applied for before the due date. Documentary evidence (e.g., medical certificate) will be expected for an extension to be granted, however this will not guarantee that the application will be successful. The Unit Convener or relevant Discipline Convener will decide whether to grant an extension and the length of the extension.

An Assignment Extension form is available from the Student Forms page.

Late submission of assignments without an approved extension will result in a penalty of 5% reduced marks from the total available, per calendar day late. An assignment submitted over 7 days late will not be accepted.

If a student chooses to submit his/her assignment via the Internet off the campus, it is the responsibility of the student to guarantee the accessibility of the Internet. Not being able to access to the Internet at a location which is outside of the campus is not an excuse for extension.

This assessment is a group task and only one submission per group is allowed. Ensure all resources used are appropriately referenced in a referencing style of your choice; but you need to maintain consistency.

Please submit your draft report to URKUND Student Text-matching Checker for a similarity/plagiarism check before you submit your final report.

Overview

In this assignment you will be working with real-world data and problems. You will be required to analyse a climate/weather data discussed on page 4. You are expected to perform five of the six steps in the data mining process on your chosen dataset. These five steps are:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation

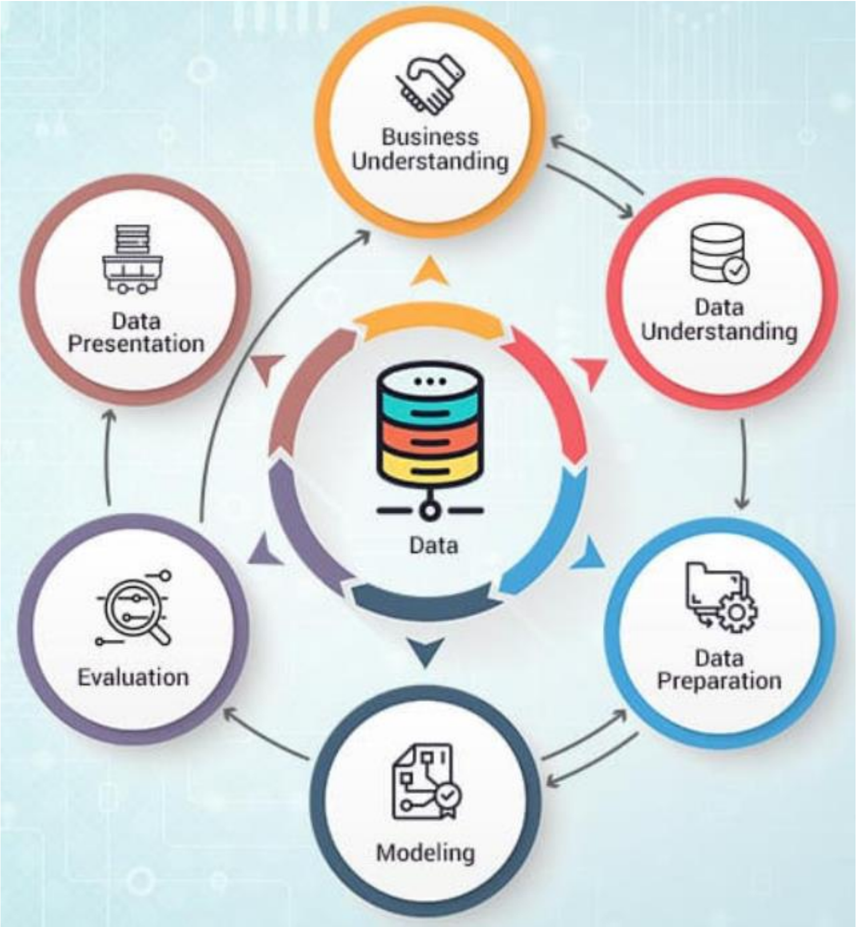

Image: http://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process

As it is the case for most real-world data mining projects, each of the datasets requires some pre-processing to get the data into a form suitable for mining.

Problem formulation

In this assignment, you are going to work on applying data mining to solve a real-world business problem. This involves identifying a business in Australia that is impacted by rainfall. The impact of rainfall on a business could be positive or negative.

- If the impact of rainfall on the business is positive, it creates an opportunity. The aim is to make recommendations that help the business capitalise on the opportunity created by rainfall.
- If the impact of rainfall on the business if negative it poses a risk, and the business may incur loss. The aim is to make recommendations that help the business minimise the loss incurred due to rainfall.

The first step will be to identify a business that is likely to be impacted (positively or negatively) by rainfall and obtain data on their business operations such as revenue over a period of time. Once the revenue (or information that will help us quantify the impact of rainfall on the business) information is obtained, the weather data for the business location (closest) can be downloaded from the Bureau of Meteorology website to explore if there is correlation between revenue of the business and rainfall.

If a correlation (positive or negative) exists between business revenue and rainfall, data mining may be used to build predictive models to forecast rainfall and make recommendations to improve its business processes.

Alternative problems and tools

You are allowed and encouraged to formulate your own project problem in other areas of applications of data mining. You can obtain and use datasets other than what is provided in this assignment as long as you can formulate a data mining problem that requires you to apply the CRISP-DM steps.

You are also allowed to perform all tasks using R or Python if you are proficient in using those programming languages. You will need to submit your fully commented scripts along with your final report.

Presentation of the progress report

As shown in the figure above, a data mining process involves six main steps. It is estimated Data Preparation (understanding and cleaning data) takes 50% – 80% of the time spent on a data mining project. ***It is strongly recommended that understanding of business and data is completed and cleaning of data has commenced at this milestone***.

Presentation of the progress report (PowerPoint presentation in Week 7) provides an opportunity to discuss any challenges you are facing as a group in undertaking the project. It also allows the unit convenor to assess the progress that has been made and provide direction to help complete the project successfully and timely.

Deliverables

The deliverable for this assignment will be a report with associated datasets in csv file format and fully commented scripts (R, Python), if any. The report will be in the style of a professional document of **no more than 12 pages**[1] single-spaced A4, Times New Roman (12), including relevant graphs and output to support your reporting. A *suggested* breakdown of the 12 pages is given in the sections below. However, you may apply a different breakdown as needed. The report will detail your methods and the results of your modelling. The report will contain the following sections:

1. Executive Summary (1 page/5 marks)
   - Summarise the data, methods and results of your report.
2. Introduction (1 page/5 marks)
   - Describe the context of the problem for a geographical area of your choice.
   - Why is this problem important for a geographical area of your choice?
   - What are the aims of the modelling for the business of your choice?
3. Data (2 pages/5 marks)
   - Describe your data visually and numerically.
   - Provide details of the pre-processing undertaken in order to meet the conditions of your chosen modelling techniques.
4. Analyses (2 pages/5 marks)
   - Describe the modelling procedure with enough detail to allow someone else to repeat the modelling. At least two (**for UG level students**) and three (**for PG level students**) different modelling techniques (e.g., decision trees, ensemble decision trees) should be applied to the problem.
5. Results (2 pages/5 marks)
   - Describe the performance of your models. Use appropriate charts and visualisation to convey the results to your audience.
6. Discussion (2 pages/5 marks)
   - Interpret your results.
   - Which model performed the best? Why?
   - What do the results mean in light of your aims?
   - Which were the most important variables in your model?
7. Conclusion (1 page/5 marks)
   - Summarise your findings.
8. Citation and references (1 page/no marks but required)
   - Use Harvard or APA author-date style e.g., Richardson (2012, p. 217) suggested that …

Note: Results and discussions may also be presented as a combined section.

[1] **Prior approval from the convenor is required to exceed the page limit. Exceeding page limit without approval may result in a penalty of up to 5%.**

Datasets

Weather dataset

This data can be retrieved online for a geographical area of your choice, from
http://www.bom.gov.au/climate/data/index.shtml?bookmark=200

Download each of the last 14 months of data (you may have to download these in 14 separate files).

Once the data has been downloaded, create a single spreadsheet by combining the 14 separate files you have downloaded. You will then need to use Excel to add a derived target variable – RainTomorrow (**Hint**: its values may be determined by applying a threshold on the Rainfall variable).

This historic dataset can be used to build a predictive model to predict whether it will rain tomorrow. Decision tree is one modelling technique that may be appropriate for this problem; you will need to choose and apply at least two different (three for PG students) modelling techniques.

Marking rubric for the report

Marking Rubric				
	Pass (50% – 64%)	**Credit (65% – 74%)**	**Distinction (75% – 84%)**	**HD (85% – 100%)**
Executive summary	*Basic* Goal/hypothesis incomplete. Little context provided, or context unclear or confusing. Summary of results and conclusion unclear.	*Competent* Goal/hypothesis mentioned but unclear or confusing. Context is provided, but links to the goal/hypothesis unclear.	*Proficient* Briefly describes goal/hypothesis of modelling. Provides (in a few sentences) the context of problem being explored. Provides brief summary of results and conclusion.	*Advanced* Briefly describes goal/hypothesis of modelling. Provides (in a few sentences) the context of problem being explored. Clearly identifies why the reader should care about the problem. Provides brief summary of methods employed, results obtained, and conclusion reached.
Introduction	*Basic* Goal/hypothesis	*Competent* Goal/hypothesis	*Proficient* Describes goal/hypothesis	*Advanced* Describes goal/hypothesis of

	incomplete.	mentioned but unclear or confusing.	of experiment.	experiment.
	Background information includes unrelated information and/or is insufficient to support goal/hypothesis.	Background information included but not sufficient to support goal/hypothesis.	Background information provides context for experimental question.	Background information giving context to goal included.
	Little organisation of information leading into the goal/hypothesis.	Organised from broad to specific with respect to the topic.	Organised from broad to specific with respect to the topic.	Organised from broad to specific with respect to the topic.
				Finishes section with brief summary of approach.
Methods (Data and Analysis)	*Basic*	*Competent*	*Proficient*	*Advanced*
	Describes modelling procedure.	Describes modelling procedure.	Describes modelling procedure.	Describes modelling procedure.
	Provides a general description of the experiment with little detail, making it difficult for a reader to repeat the experiment.	Provides enough information for someone to infer how to do the experiment, but details may be unclear.	Provides enough detail for someone to repeat the experiment from the instructions.	Provides enough detail for someone to repeat the experiment from the instructions.
	Inclusion of too much result.	Inclusion of some results.	Written in narrative (paragraph) form.	Does not include unnecessary detail - written to an audience familiar with the topic.

	Not written in narrative (paragraph) form	Not written in narrative (paragraph) form		Written in narrative (paragraph) form.
Results	*Basic* Results incompletely presented in narrative form. Tables and figures present, but not referenced or cited properly. Data mixed with much interpretation or conclusions.	*Competent* Results are presented in narrative form. Tables and figures cited, but not necessarily in correct order. Data mostly free of interpretation or conclusions.	*Proficient* Results are presented in narrative form. Tables and figures are cited appropriately and numbered by the order addressed in the text. Data is free of interpretation or conclusions.	*Advanced* Results are presented in narrative form. Tables and figures are cited appropriately and numbered by the order addressed in the text. Apparent trends in data are identified. Data is free of interpretation or conclusions.
Discussion	*Basic* Explanation loosely relates to results collected.	*Competent* Results interpreted and explanations offered for trends or patterns as well as anomalies in the data. Original goal/hypothesis mentioned, but unclear how results relate.	*Proficient* Results interpreted and explanations offered for trends or patterns as well as anomalies in the data. Addresses results with respect to original goal/hypothesis.	*Advanced* Results interpreted and explanations offered for trends or patterns as well as anomalies in the data. Addresses results with respect to original goal/hypothesis.

				Poses further questions to continue research or address limitations with current project.
Conclusion	*Basic* Results summarised. Conclusion not made with respect to goal/hypothesis.	*Competent* Results summarised. Conclusion relates to the goal/hypothesis but is lengthy or unclear.	*Proficient* Results summarised. Conclusion is clearly related to the goal/hypothesis. Includes summary of the implications of the results.	*Advanced* Results summarised. Conclusion is succinct and clearly answers the goal/hypothesis. Includes a succinct summary of the key implications of the results.
References	*Basic* Few references cited from text listed. May be listed in correct format as per APA. No references other than the textbook used.	*Competent* References cited include most sources cited in text. May be listed in correct format as per APA. No references other than the textbook used.	*Proficient* References cited include all sources in text but listed in an incorrect format as per APA. Only those references cited in the text included in reference list Additional references used.	*Advanced* References cited include all sources in text, listed in a single and consistent format. Only references cited in the text included in reference list. References other than the textbook used.
	Basic	*Competent*	*Proficient*	*Advanced*

Format and Style	Some/all of narrative written in present or future tense. Frequent use of 1st person context. Written to an audience with limited knowledge of the material. Many grammar and spelling errors present.	All narrative written in past tense. Some inconsistencies between 1st and 3rd person context. Written to an audience with limited knowledge of the material. Several grammar and spelling errors present.	All narrative written in past tense. Consistent use of 1st or 3rd person. Written to an audience with limited knowledge of the topic. Only a few grammar and spelling errors.	All narrative written in past tense. Limited use of 1st person. Written to an informed audience generally familiar with the material. Minimal to no errors in grammar and spelling.