4. Iteration 4 BDAS (Steps 1 – 8)

Start Assignment

- Due 24 May by 11:59
- Points 100
- Submitting a file upload

1. Please make sure you submit your report file which is .docx or .pdf *(with your GitHub link on the first page of your report)***.**

2. Please do not submit a zip file of code or data. All of this needs to be uploaded and pushed on to your GitHub repository. You will share scripts/code projects by providing a web link to your GitHub repository (this needs to be on the first page of the report)

If you would like to keep your repository private and only share with the markers, please share your work to:

GitHub username: **TBA**
Alternatively, you can keep your repository public (everyone can view) and share the link with us in your iteration document.

See instructions on folk repository and pushing your work to GitHub in the below video:

If your repository is private please add the tutors as collaborators: Go to your repository on GitHub, go to Settings, and click Collaborators.

Also, when submitting your assignment, please append the following disclaimer at the end of the document:

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright.
(See: **https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html (https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html)**).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."

Iteration 4

Criteria	Ratings				Pts
1.1 - Business/situation objectives must be logical, and in line with data mining goals and the business success criteria. Must be conveyed in a clear manner.	**3 Pts Very Good**	**2.1 Pts Good**	**1.2 Pts Poor**	**0 Pts Very Poor**	3 pts
1.2 - Situation assessment must effectively describe the resources, requirements, assumptions, constraints, risks and contingencies of the project.	**2 Pts Very Good**	**1.4 Pts Good**	**0.8 Pts Poor**	**0 Pts Very Poor**	2 pts
1.3 - Data mining goals must be achievable and closely aligned with the business objectives/success criteria. Must be conveyed in a clear manner.	**3 Pts Very Good**	**2.1 Pts Good**	**1.2 Pts Poor**	**0 Pts Very Poor**	3 pts
1.4 - The project plan must address how each phase of the project will be carried out for the current iteration. A day-to-day timeline must be proposed within the project plan.	**2 Pts Very Good**	**1.4 Pts Good**	**0.8 Pts Poor**	**0 Pts Very Poor**	2 pts
2.1 - Collect initial data and describe where the data was collected from, how it was collected and any issues encountered during collection.	**2 Pts Very Good**	**1.4 Pts Good**	**0.8 Pts Poor**	**0 Pts Very Poor**	2 pts
2.2 – Data description must describe the format, quantity, fields and surface-level features of the data.	**2 Pts Very Good**	**1.4 Pts Good**	**0.8 Pts Poor**	**0 Pts Very Poor**	2 pts
2.3 - Data exploration must assist readers in understanding the data through the usage of strong visualisations (visualising the raw data). Must be communicated in a clear manner and explicitly linked to the rest of the steps.	**4 Pts Very Good**	**2.8 Pts Good**	**1.6 Pts Poor**	**0 Pts Very Poor**	4 pts
2.4 - Data quality must be verified by checking for errors, missing values, and data quality patterns explicitly.	**2 Pts Very Good**	**1.4 Pts Good**	**0.8 Pts Poor**	**0 Pts Very Poor**	2 pts
3. Data Preparation As a general guide: 3.1 - Data must be selected effectively. Goals, data quality, technical constraints, among other factors should be explicitly considered. 3.2 - To clean the data, issues must be made explicit, then explicitly resolved. 3.3 - Data must be appropriately constructed through the creation of new features/variables, and/or data repositories/tables. 3.4 -	**15 Pts Very Good**	**10.5 Pts Good**	**6 Pts Poor**	**0 Pts Very Poor**	15 pts

Criteria	Ratings				Pts
Integration must take place. This includes effectively merging data from various sources. 3.5 - Reformatting includes changing the formats of different data sources and trimming content, among other steps specific to the data.					
4.1 - Data must be reduced through the selection of features relevant to the predictor (horizontal reduction) and/or vertical reduction. This could be achieved through the use of feature selection/logical processes.	**2.5 Pts Very Good**	**1.75 Pts Good**	**1 Pts Poor**	**0 Pts Very Poor**	2.5 pts
4.2 - The data must be projected through the use of statistical transformations (such as taking the log of a distribution).	**2.5 Pts Very Good**	**1.75 Pts Good**	**1 Pts Poor**	**0 Pts Very Poor**	2.5 pts
5.1 - Match and discuss DM methods within the context of the DM objectives.	**5 Pts Very Good**	**3.5 Pts Good**	**2 Pts Poor**	**0 Pts Very Poor**	5 pts
5.2 - Select the appropriate DM method(s) in a logical manner. The selected DM method must be in line with the data mining goal/success criteria.	**5 Pts Very Good**	**3.5 Pts Good**	**2 Pts Poor**	**0 Pts Very Poor**	5 pts
6.1 Conduct exploratory analysis of DM algorithms within the context of the DM objectives. Then, discuss the analysis.	**2.5 Pts Very Good**	**1.75 Pts Good**	**1 Pts Poor**	**0 Pts Very Poor**	2.5 pts
6.2 - Select algorithm(s) in a logical manner based on the exploratory analysis and discussion.	**2.5 Pts Very Good**	**1.75 Pts Good**	**1 Pts Poor**	**0 Pts Very Poor**	2.5 pts
6.3 - Model(s) must be selected/built, and the appropriate algorithm/model parameter(s) must be selected.	**10 Pts Very Good**	**7 Pts Good**	**4 Pts Poor**	**0 Pts Very Poor**	10 pts
7.1 - Logical test designs must be created. Justify why a particular test design was used (for example, why was a 70/30 training/testing split used?).	**1 Pts Very Good**	**0.7 Pts Good**	**0.4 Pts Poor**	**0 Pts Very Poor**	1 pts
7.2 - Data mining must be conducted (the model must run).	**7 Pts Very Good**	**4.9 Pts Good**	**2.8 Pts Poor**	**0 Pts Very Poor**	7 pts

Criteria	Ratings				Pts
7.3 - Search for patterns and document the model's output.	**7 Pts** Very Good	**4.9 Pts** Good	**2.8 Pts** Poor	**0 Pts** Very Poor	7 pts
8.1 - Study and discuss the mined patterns. Carry out an in-depth discussion about the data, results, models and patterns.	**2.5 Pts** Very Good	**1.75 Pts** Good	**1 Pts** Poor	**0 Pts** Very Poor	2.5 pts
8.2 - Visualise the data, results, models and patterns in a clear and effective manner.	**5 Pts** Very Good	**3.5 Pts** Good	**2 Pts** Poor	**0 Pts** Very Poor	5 pts
8.3 - Interpret the results, models and patterns showing a clear understanding of the results.	**2.5 Pts** Very Good	**1.75 Pts** Good	**1 Pts** Poor	**0 Pts** Very Poor	2.5 pts
8.4 - Assess and evaluate the results, models and patterns using the appropriate methods/processes.	**2.5 Pts** Very Good	**1.75 Pts** Good	**1 Pts** Poor	**0 Pts** Very Poor	2.5 pts
8.5 - Multiple iterations must take place. Repeat various steps to ensure your model is effective/robust. Document the steps/processes/reasoning.	**7.5 Pts** Very Good	**5.25 Pts** Good	**3 Pts** Poor	**0 Pts** Very Poor	7.5 pts
					Total points: 100