SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

ANL303
Fundamentals of Data Mining

Group-based Assignment

January 2024 Presentation

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for ANL303 Fundamentals of Data Mining.

The cut-off date for this assignment is **13 March 2024, 2355hrs.**

This is a group-based assignment. You should form a group of **4 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that the work distribution is inequitable to either yourself or your group mates, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective.

If your course involves programming, you are urged to read the following articles as well:
https://wiki.cs.astate.edu/index.php/Plagiarism_in_a_Programming_Context

https://www.turnitin.com/blog/plagiarism-and-programming-how-to-code-without-plagiarizing-2

Note to Students:

Compose your report using Microsoft Office Word, and save either as .doc or **.docx (preferred).**

You are to include the following particulars in your submission: Course Code, Title of the GBA, SUSS PI No., Your Name, and Submission Date.

For this GBA, it is mandatory that questions are not divided among group members. Each member must independently address and work on a question before engaging in group discussions of the question. In the event of a peer evaluation, each member is expected to submit their own original answers along with justifications for their individual contributions.

All peer evaluation requests must be submitted to the school at least three working days before the GBA due date. Late requests will not be considered.

Use of Generative AI Tools (Allowed)

The use of generative AI tools is allowed for this assignment.

- You are expected to provide proper attribution if you use generative AI tools while completing the assignment, including appropriate and discipline-specific citation, a table detailing the name of the AI tool used, the approach to using the tool (e.g. what prompts were used), the full output provided by the tool, and which part of the output was adapted for the assignment;

- To take note of section 3, paragraph 3.2 and section 5.2, paragraph 2A.1 (Viva Voce) of the Student Handbook;

- The University has the right to exercise the viva voce option to determine the authorship of a student's submission should there be reasonable grounds to suspect that the submission may not be fully the student's own work.

- For more details on academic integrity and guidance on responsible use of generative AI tools in assignments, please refer to the TLC website for more details;

- The University will continue to review the use of generative AI tools based on feedback and in light of developments in AI and related technologies.

Question 1

Credit card companies want to encourage their customers to make purchases using their credit cards because they earn fees and interest on those transactions. Individuals who have no card balance are not generating revenue for the credit card company since they are not using their cards for purchases. By sending promotional offerings to individuals with no card balance (i.e., zero-balance cardholders), the credit card company hopes to entice them to make a purchase. If they make a purchase, it will lead to the generation of revenue for the company.

Company ABC is a credit card company that often includes promotional offerings with their monthly credit card billings. The offers provide customers with an opportunity to purchase items, such as luggage, magazines, or jewellery. The dataset (credit_card_promotions.csv). contains customer information, such as age and income, obtained through their credit card applications with ABC. It also contains information about whether these customers have accepted various past promotional offerings sponsored by ABC. Details of the dataset are provided in Table 1.

Table 1. Details of the dataset

Field	Description
ID	Unique identifier for the customer
Gender	Gender of the customer
Age	Age of the customer
Income	Income of the customer
Luggage	Whether the customer accepted luggage promotion in the past (Yes/No)

Magazine	Whether the customer accepted magazine promotion in the past (Yes/No)
Jewellery	Whether the customer accepted jewellery promotion in the past (Yes/No)
Watch	Whether the customer accepted watch promotion in the past (Yes/No)
Life Insurance	Whether the customer accepted life insurance promotion in the past (Yes/No)
Credit Card Insurance	Whether the customer accepted credit card Insurance promotion in the past (Yes/No)
Dining	Whether the customer accepted dining promotion in the past (Yes/No)
Laptop	Whether the customer accepted laptop promotion in the past (Yes/No)
Theme Park	Whether the customer accepted theme park promotion in the past (Yes/No)
Zero-Balance Cardholder	Whether the customer is a zero-balance cardholder (Yes/No)

In this question, you are going to import the dataset to IBM SPSS Modeler and perform all data mining tasks (including data preparation, if any) using IBM SPSS Modeler.

(a) Identify the data quality issues in the dataset and perform data cleaning. In less than 100 words, briefly describe how the dataset is cleaned and becomes suitable for data mining using IBM SPSS Modeler. Provide necessary screenshot(s).

(15 marks)

(b) Using the dataset obtained from part (a), perform data visualisation that can answer the following enquiries:

- Enquiry 1: What are the top three most popular promotions among zero-balance cardholders?
- Enquiry 2: Is it true that the average income for zero-balance cardholders is higher than the average income for non-zero-balance cardholders in both gender groups?

Provide ***one (1)*** graphical display for each enquiry to support your answers.

(20 marks)

(c) Sending promotional materials to everyone is wasteful. ABC has decided to target their marketing efforts more effectively, specifically by aiming to encourage card usage and generate revenue from individuals with no card balance. To achieve this goal, they would need to identify the subgroup of customers who are likely to have no card balance.

(i) Construct a k-means model that can provide insights into the characteristics of customers likely to have no card balance. Describe how you identify your model as the final best model. For the model that you have chosen, provide the screenshot of the model's outputs (showing the cluster summary, and the cluster comparison) and describe the profile of each cluster.

(15 marks)

(ii) Discuss how you identify the target customers for ABC based on the clustering results.

(10 marks)

(d) Based on part (c), ABC has identified the target customer group for their promotional offerings. Only the target group would be sent promotional information. Construct an association rule mining model that can help ABC understand customer preferences among the promotional offerings within the target group.

Pick ***two (2)*** rules that you find interesting. Describe them in terms of support, rule support and confidence, and discuss insights that can help ABC generate revenue.

In your answer, provide screenshots that show clearly (i) which fields are selected for the analysis and their roles, (ii) the parameters used in the algorithm, and (iii) the association rules that you picked for discussion.

(20 marks)

(e) Assume that ABC has authorised a new life insurance promotion similar to the previous life insurance promotion. This new life insurance promotion is exclusive for new cardholders who have not had a chance to take advantage of a previous promotion. Based on the dataset described in Table 1, assess the suitability of association analysis for identifying the profile of new cardholders likely to accept the new life insurance promotion.

If it is suitable, discuss (i) which field(s) listed in Table 1 should be included as the antecedent and consequent; (ii) if any data preparation task should be done.

If it is not suitable, provide your justification.

You are not required to use IBM SPSS Modeler for part (e).

(10 marks)

Another 10 marks are allocated for your writing.
(Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective)

Your writing should be succinct but not at the expense of excluding relevant details. Highlight only the points that are relevant to your discussion. Use plain and simple language. Some questions may not come with absolutely right or wrong answers. For such questions, you have the liberty to express your views about the problem. However, your points have to be supported by evidence and good reasoning. It's the quality and not the length that counts. Make sure you follow the report guidelines and style specified in this assignment.

The topics in the main report should be presented in the order according to the sequence of the tasks/questions listed in the assignment; that is, in the order of (a), (b), ..., etc. You can have several sub-sections within a section if you deem appropriate. To avoid high Turnitin score, **do not** copy the assignment questions into the report.

The report must be self-contained. It is important to include all relevant tables and figures in the report as evidence to support the answers given.

The followings are some details of report format:

- Length: **should not exceed 7 pages** (including the relevant graphs, tables, references, screenshots and appendices (if any), but excluding the cover page)
- Font Style: Times New Roman
- Font size: 12
- Line spacing: 1.5
- Margins: 1" for the top, bottom, right and left
- Include the page number on each page

Some further suggestions:

- Ensure minimal grammatical and typographical errors
- Write clearly in plain English
- Write appropriately to the context
- Cite appropriate sources
- Provide a reference or bibliography at the end of the main report
- Include less relevant details in the Appendix
- Good overall presentation of the report

---- END OF ASSIGNMENT ----