ANL310
Business Analytics Applications and Issues

Group-based Assignment

January 2025 Semester

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for ANL310 Business Analytics Applications and Issues.

The cut-off date for this assignment is **16 March 2025, 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.

Note to Students:

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

There are four (4) parts to this GBA:
1. **Powerpoint Presentation slides** (maximum 15 slides excluding title and sub section slides);
2. **15-minute oral presentation** (demonstration of the construction of streams/explanation of Python codes as well as insights gathered based on the presentation slides);
3. **10-minute questions and answers (Q&A); and**
4. **Team Dynamics.**

The presentation slides (PPT) should focus on the results and findings as well the recommendations, a brief description should be provided for each slide (except for title and sub section). The template for the slides had been uploaded to Canvas (L/LG group).

For Canvas submission, only the presentation slides (PPT) in PDF format should be submitted (refer to Appendix for the format). **Note that your PDF file must not be stored in an image format, doing so will cause Turnitin issues, and will result in a 20-mark deduction**. Each group is required to give a 15-minute oral presentation via Zoom or other video recording tools (link to be shared in the last presentation slide), **all members of the group must present**.

Based on the slides and the oral presentation submitted, there will be a 10-minute Question and Answer (Q&A) Zoom session which will conducted by your instructor. **All students are**

expected to know the entire GBA as questions can be directed at anyone during the Q&A. You are not allowed to redirect a question to another student.

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.

Every student is expected to complete the entire GBA on their own. Splitting of work is strictly not allowed. All members are expected to complete their own GBA work way before due date, thereafter they are to meet and synthesize their works to form the final submission. If we find students split the work, mark deduction will be imposed to all the members in the group.

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.

Background Information

BigHeart, a non-profit organisation, has been in operation for the last 20 years and like most non-profit organisations, its primary objectives include promoting the well-being of people and communities through their activities. BigHeart relies heavily on their volunteers to manage and run these activities.

In the recent two years, BigHeart has witnessed an increase in volunteer churn. The increase greatly impacted their productivity, performance and service quality. BigHeart decides to embark on the use of data mining to identify volunteers who are likely to churn. BigHeart hopes that early identification can help them to take appropriate and timely actions to prevent these volunteers from churning.

The following variables have been collected for a group of 1500 volunteers (dataset = BigHeart_Volunteers.xls) in the past two years for this purpose and are summarised in Table 1. There is one month of latency (information collected about the volunteer who churned was as at one month before he/she churned).

Table 1: Variables in BigHeart_Volunteers.xls

Variable	Definition	Value
VolID	Volunteer identification code	0001 to 1500
Age	Age of volunteer	21 to 60 years old
Gender	Gender of volunteer	Female Male
MStat	Marital status of volunteer: Single: Single Married: Married Others: Divorced/Separated/Widowed	Single Marr Others
AttendTrain	Whether the volunteer had attended any training in the past three months	Yes No
Rotation	Whether the volunteer had been rotated to other volunteering activities in the past three months within BigHeart	Yes No
PrevVolExp	Whether the volunteer had previous volunteering experience before joining BigHeart	Yes No
Conflict	Whether the volunteer had faced any conflict at work and/or home in the past three months	Yes No
PeerSupport	Whether the volunteer felt there was any support from their peer volunteers in the past three months	Yes No
SatRating	The satisfaction rating indicated by the volunteer for the last activity he/she participated	1 (lowest) to 7 (highest)
Feedback	No of feedback provided by the volunteer in the past three months	0 to 4
NoShow	Number of days of the volunteer did not turn up for volunteering activities in the past three months	1 to 10 days
NumVol	Number of years volunteering in BigHeart	1 to 24 years
Churn	Churn indicator	Yes No

Question 1
Association Analysis (21 marks)

BigHeart wants to find out if volunteer churn may be associated with volunteer's demographics (i.e, Gender, MStat) as well as their experience with BigHeart (i.e., AttendTrain, Rotation, PreVolExp, Conflict, PeerSupport).

Prepare the data accordingly. For this part, convert "Gender" and "MStat" to flag variables, although the Apriori node (for IBM-SPSS Modeler) can handle ordinal and nominal variables. (*Note: Convert "Gender" into 2 variables to capture both genders and "MStat" into 3 variables to capture the three different statuses. These flag variables should contain "Yes"/"True" or "No"/ "False".*)

Implement the necessary association analysis using Apriori and generate between 5 to 8 rules with the appropriate parameter settings using the 10 variables stated. (*Note: The minimum [antecedent] support level should be at least 10% and minimum confidence level should be at least 60% with maximum number of antecedents less than or equal to five.*)

You can use IBM-SPSS Modeler (Apriori node) OR Python (using Google Colab) for the data preparation and association analysis.

During the **oral presentation**:
For groups using IBM-SPSS Modeler (Apriori node), you should demonstrate the construction of the stream:
- Data preparation for the conversion of Gender and MStat;
- Types tab of the Data Source node showing all the appropriate settings of the variables;
- Node for association analysis, showing all the relevant parameter settings; and
- Output of the association analysis.

For groups using Python, you should explain the syntax used for:
- Data preparation for the conversion of Gender and MStat (along with the finalised dataset to be used for association analysis);
- Association analysis (Apriori); and
- Output of the association analysis.

* Bonus marks will be given to groups using Python with the correct syntax

Provide the following in the **presentation slides**:
- Identification of two (2) overlapping and two (2) subset rules;
- Interpretation of the rules with the use of three (3) rule sets; and
- Discussion of how the rules can help in making recommendation to reduce volunteers' churn.

Question 2
Clustering (22 marks)

BigHeart decided that to better understand the volunteers, segmentation and profiling (via clustering) should be attempted using their age, satisfaction rating (i.e., SatRating), no of feedback (i.e., Feedback), number of days of the volunteer did not turn up for activities (i.e., NoShow) and number of years volunteering in BigHeart (NumVol) in Table 1.

Using the original data (i.e., BigHeart_Volunteers.xls), implement the necessary clustering using the TwoStep clustering node (number of clusters should not be above 5) with the above-mentioned five variables.

During the **oral presentation**, you should demonstrate the construction of the stream for clustering, showing the following:
- Clustering node, showing the fields and relevant parameter settings;
- Outputs of the clustering node (including the average silhouette coefficient, relative distributions by cluster and cluster comparison); and
- Output of the node to analyse the incidence of volunteers with higher risk of churn among the clusters.

Provide the following in the **presentation slides:**
- Interpretation of the clustering results and evaluate the cluster profiles;
- Propose clusters labels (of 10 words or less) for the clusters based on the top 3 important inputs (or clustering criteria) in the clustering solution; and
- Discussion of two recommendations on how these insights can be used for deployment to reduce volunteer churn.

Question 3
Predictive Modelling (27 marks)

Using the original data (i.e., BigHeart_Volunteers.xls) with all variables as inputs (excluding VolID), construct three decision trees using CHAID, CART and C5.0, respectively, to predict who is more likely to churn. Use the following parameter settings:

CHAID
Minimum records in parent branch (%): 5%
Minimum records in child branch (%): 3%
No changes to the other settings

CART
Minimum records in parent branch (%): 5%
Minimum records in child branch (%): 3%
Overfit Prevention Set (%): 0%
No changes to the other settings

C5.0
Minimum records per child branch: 36
No changes to the other settings

Data Partition
Training the model: 80%
Testing the model: 20%
No changes to the other settings

Evaluate the models and select the champion model. Based on the champion model, re-construct the decision tree using all the data (i.e., 100% of the data) for the purpose of deployment.

During the **oral presentation**, you should demonstrate the construction of the stream for predictive modelling, showing the following:
- Types tab of the Data Source node, showing all the relevant settings of the variables;
- Partition node, showing the settings as stated;
- CHAID, CART and C5.0 nodes, showing the settings as stated;
- Output(s) of the node(s) used to evaluate the models; and
- Reconstructed "Tree" of the champion model (including the predictor importance).

Provide the following in the **presentation slides**:
- Evaluation of the models and selection of the champion model;
- Appraisal of the deployment of the model by interpretating the rules with high probability of churn from the re-constructed model; and
- Discussion of three (3) recommendations to help BigHeart reduce volunteer churn based on the insights generated.
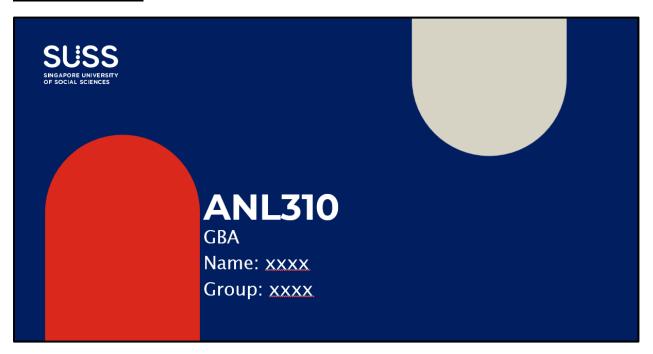
Oral Presentation and Team Dynamics including Questions and Answers (30 marks)

Every member of the group will also be graded individually based on the quality and effectiveness of his/her presentation, contribution to the assignment as well as responses to the questions raised.
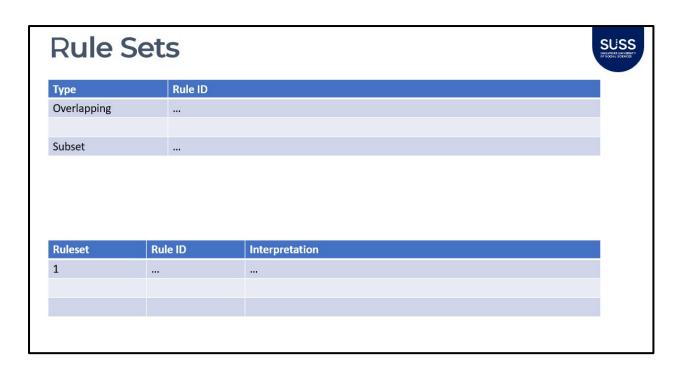
More information on the grading of the contribution to the assignment (Team Dynamics) will be shared during Seminar 3.
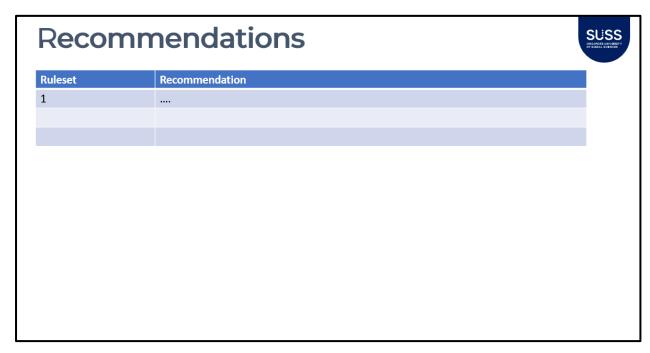
Appendix

Presentation Slides

Rule Sets

Type	Rule ID
Overlapping	...
Subset	...

Ruleset	Rule ID	Interpretation
1	...	...

Recommendations

Ruleset	Recommendation
1	....

Link for Oral Presentation

This is the link for our oral presentation:
XXXX

10

---- END OF ASSIGNMENT ----