SUSS
SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

ANL307
Predictive Modelling

Group-based Assignment

January 2025 Semester

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for Predictive Modelling.

The cut-off date for this assignment is **20 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.

Please keep a copy of your report before submission, and check whether your submission has been uploaded onto Canvas successfully (you can check if the Turnitin system can read your submission). **Failed or unreadable submissions will be marked as zero (0)**.

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.

Predictive Modelling is widely applied across various fields and industries. This group assignment is an open-ended Predictive Modelling project aimed at finding and tackling a challenge in the auction industry. You will work with public datasets comprising over 2,000 auction verification records. Your task is to explore the use of classification and/or regression models to analyse and predict auction outcomes effectively. Using the three core components of business analytics—business, data, and technique—or the CRISP-DM framework, structure your study and submit a comprehensive report.

This assignment includes two public datasets and provides three modelling options. You can choose Option 1, 2 or 3 for your GBA. Please indicate the option you have chosen in your report.

Datasets:

- Dataset 1: *UCI Auction Verification.csv*, contains information about the German 4G spectrum auction. This auction involved the allocation of licenses for specific radio frequency blocks by the German government to telecommunication companies for deploying and operating 4G mobile networks. Each radio frequency block (i.e., *process.b1, process.b2,* ...) was treated as a separate item, and telecommunication companies competed to secure the blocks they desired. The auction utilised a simultaneous multiple-round auction format, where participants can bid on multiple frequency blocks over several rounds. Additional data and background information can be found at the following links:

  https://archive.ics.uci.edu/dataset/713/auction+verification
  Reference paper: https://ieeexplore.ieee.org/abstract/document/9721192

- Dataset 2: *UCI Auction Verification Outliers.xlsx*, contains the learned anomaly scores of residual outliers, which indicates whether the corresponding auction design contains undesirable behaviors. The outlier labels were derived from a regression tree built using the auction verification information in the Dataset 1. Additional data and background information can be found at the following links:

  https://www.kaggle.com/datasets/jamestansc/auction-verification-regression-anomalies
  Reference paper: https://ojs.bonviewpress.com/index.php/jdsis/article/view/3861

The variable "*RecordNum"* in the two files can be used to merge the datasets. The two data files also can be downloaded from L01/LG01 Group → Modules → Assessment.

Modelling Options:

- Option 1: Use the Dataset 1 to build either a regression or classfication model for auction outcome verification (i.e., The target variable for the regression model is *verification.time*, while for the classification model, it is *verification.result*)
- Option 2: Combine the Dataset 1 and Dataset 2 to perform a classification task (i.e., detect anomalies, the target variable is *Outlier* in the Dataset 2).
- Option 3: Implement both Options 1 and 2. Compare the regression or classification results with and without including the anomalous data.

Note:

1. The modelling difficulty increases from Option 1 to Option 3, and the corresponding marks awarded will also differ based on the complexity of the task. Option 3 is the most challenging and you should understand the Dataset 2's reference paper well to do this properly.
2. If any aspects of the provided background or data description are unclear, you are encouraged to make reasonable assumptions to address or enrich them. Be sure to provide explanations for these assumptions, and cite references if applicable to support your reasoning.
3. The evaluation of your report will not solely depend on the predictive performance of your model. Instead, significant emphasis will be placed on the process you followed to identify and develop a suitable model.

Instruction

Your report should include (but not be limited to) the following key sections:

Part A. Introduction

Use up to **500 words** to provide basic information about your project. It should cover the following:

1. Introduce the background and the motivation of your project.

2. Discuss the prediction problem your team has identified to address, including a clear description of the business problem, business objective and data mining objective.

3. Review ***two (2)*** academic papers related to the identified business problem and briefly appraise the use of predictive modeling for addressing it. (*Note: At most one paper could be selected from the two provided references*).

(20 Marks)

Part B. Data Understanding and Preparation

Use up to **600 words** to provide a detailed overview of the dataset and describe the data preparation process, if applicable. It should include the following key points:

1. **Data Characteristics and Quality**: Provide a description of the original dataset, including its key characteristics and any observations about its quality issues.

2. **Data Exploration**: Perform and discuss data exploration to uncover patterns or trends. Use visualisations such as charts, figures, or tables to present your findings effectively.

3. **Data Preparation**: Referring to the proposed business problem in Part A, detail the data preparation steps necessary to make the dataset suitable for predictive modelling. This may include data sampling, merging, cleaning, transformation, feature selection,

feature generation, or incorporating additional data form publicly avaiable sources, etc. Please provide clear justifications for each step. If you determine that no data preparation is needed, please explain why the data is already adequate for predictive modelling.

(25 Marks)

Part C. Model Design and Interpretation

Use up to **1500 words** to cover the following key points about the modelling:

1. **Baseline Model**: Build a CART model with default settings. Report the modelling results and evaluate its performance. Discuss the possible limitations of the model in solving the proposed problem in Part A.

   It is acceptable to construct the model using the IBM SPSS Modeler or any open-source tools of your choice. If CART is not suitable for your case, you may select an alternative predictive model; ensure to provide a clear justification for your choice.

2. **New Model:** Referring to the performance and limitation discussed above, propose new/alternative model(s) that would be more suitable for your case. Construct the model(s) and interpret the modelling results. Clearly indicate the parameter settings and tunings to show more details of your modelling process.

3. **Modelling Comparison:** Compare the insights gained and prediction performance results of the proposed model(s) with those of the baseline model.

- *While not required, you are permitted to use/propose a predictive model(s) beyond the scope of ANL307.*
- *If you use software or tools beyond IBM SPSS Modeler, the relevant source code or modelling screenshots are needed to show your study.*
- *For different models to be compared fairly, model performance comparisons must be done in the same way.*
- *It is possible that your model does not perform better than the baseline model. It is more important that you clearly explain the reason(s) and account for the differences in predictive performance between the models.*

(45 Marks)

Part D. Conclusion

This section should use up to **300 words** to cover the following:

1. Summarise your work, which includes but is not limited to the important pattern(s) discovered from the results, and discuss how to deploy the implications.

2. Appraise the limitations of your proposal. Give a concluding remark and suggest possible ways to address the limitations.

3. Provide a screenshot of your IBM SPSS Modeler streams with proper annotations include any other support material (e.g., Python code file, *embed a copy of your Python code file (if any) into your Word document and display it as an icon*).

(10 Marks)

Report Writing

Your report should be written professionally such that the various aspects of predictive modelling are well distinguished. It should be succinct but not at the expense of excluding relevant details. You should provide enough details/descriptions of your study in the report so that your work can be properly assessed. Make sure you follow the report guidelines and style specified below. It is important to include all relevant tables and figures in the report as evidence to support your study and conclusion. You can refer to the structure of academic research papers. Use American Psychological Association (APA) referencing style to format your reference citation and reference list.

Further mark deductions will be imposed for poorly prepared reports. Pay attention to paraphrasing, and at all costs, avoid patchwriting. A significant penalty of **up-to-25-mark** deduction will be imposed if students are found to be sloppy in paraphrasing work from other sources. **If plagiarism is suspected, the case will be handed over to the Exam Department for further investigation**. **More information on effective paraphrasing strategies can be found on**

https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective

The followings are some details of the report format:

- Length: **should not exceed 16 pages** (including the relevant screenshots, graphs, tables, references, and appendices (if any), excluding the cover page). **Deduct 5 marks for each extra page. Deduct a maximum of 10 marks for excessive page count and other formatting errors.**
- **Figures and graphs are not included in the word count for each part.**
- 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
- Good overall presentation of the report

---- END OF ASSIGNMENT ----