COMP5310 Project Stage 2

Develop and evaluate a predictive model

Due: 11:59PM on 15th of May 2025 (Week 11)

*This assignment is worth **25%** of the final mark of the unit of study.*

GROUPS

This stage is done with the ***same group members you worked with for Stage 1***. However, under exceptional circumstances, an alternative group may be created by the tutor when a group is reduced in size due to members discontinuing this unit. If this applies to you, please email the unit coordinator maryam.khaniannajafabadi@sydney.edu.au or the TAs: ssri4213@uni.sydney.edu.au or weiyi.wang@sydney.edu.au to discuss this.

Note: Each member of the group is required to complete individual tasks, but the project will be submitted as a combined effort. The final project will be marked as a whole, with both individual and group components contributing to the final grade. All assessments will be based on the single, submitted document.

Dispute resolution

If, during the course of the assignment work, there is a dispute among group members that you can't resolve or that will impact your group's capacity to complete the task well, you need to inform the unit coordinator maryam.khaniannajafabadi@sydney.edu.au or the TAs: ssri4213@uni.sydney.edu.au or weiyi.wang@sydney.edu.au. Make sure that your email specifies the group name and is explicit about the difficulty; also, make sure this email is copied to all group members (including anyone you are complaining about).

We need to know about problems in time to help fix them, so set early deadlines for group members, and deal with non-performance promptly (don't wait till a few days before the work is due to complain that someone is not delivering on their tasks). If necessary, the coordinator will split a group and leave anyone who didn't participate effectively in a group by themselves (they will need to achieve all the outcomes on their own). **This option is only available up until Thursday Week 9**, which is the last day with time to resolve the issue before the due date. For any group issues that arise after this time, you will need to try to resolve the problem on your own, and you will continue to be treated as a single group. If someone doesn't provide the material required for the report, or their material is not of the agreed standard, you should still have the report show what that person did. Their section of the report may be empty if they don't produce anything, or it may have material but not enough. In such cases, please put a "Note to marker" on the front page of the report, which describes the circumstances. That way, we can consider how best to apply the marking scheme. Note that it is not expected or sensible for other members to do the work that someone failed to deliver.

PROJECT

Overview

The objective of Stage 2 of the project is to build a robust predictive model using the clean dataset obtained in Stage 1. This stage will involve advanced predictive modelling techniques, as well as thorough model evaluation and optimisation processes.

Important Notes:

1. ***You MUST work in the same groups you worked on during Stage 1.***
2. ***Further cleaning of the dataset, addition of previously dropped columns, and or removal of columns are permitted if you wish.***
3. ***Each member must use a different modelling technique to develop their predictive model.***
4. ***Changing of target variable and research question is also permitted, if the group chooses to do so.***

DELIVERABLES

Report

The report must have a maximum of 3 pages for each individual section and maximum of 3 pages for the group section (including both group components 1 and 2) for a group of 2, and a maximum of 4 pages for the group section for a group of 3. You must use the high-level headings, as provided below, to indicate the different sections and sub-sections of the report.

COMP5310 Project Stage 2

Develop and evaluate a predictive model

You must use line spacing of at least 1.15pt, margins of at least 1.8cm, and body font size of at least 10pt. The goal is to convey the problem clearly and concisely.

The report should be in PDF format, named using the following convention: **"GroupX_A2_Report.pdf",** where X is your group number. **DO NOT SUBMIT A FOLDER THAT IS NAMED GroupX_A2_Report**. It must have a front page that gives the **group number**, and the **list of members** involved (giving their **SIDs** AND **unikeys**, **NOT their names**).

The body of the report must have a structure as follows:

Group Component 1

The report must begin with a group section including:

1. **Topic and research question:** Describe the research problem comprehensively, emphasizing its significance in the domain. All members must agree upon and aim to answer the same research question. Clearly articulate the research question and highlight its implications for various stakeholders. Discuss how addressing this question could lead to actionable insights or improvements in decision-making for the stakeholders.
2. **Dataset:** Provide a detailed overview of the dataset and discuss any challenges, class imbalances, and or biases present in the data and how they might impact the modelling process.

3. Setup

3.1. Modelling agreements: Identify an attribute that you will all make predictions about and agree on at least two measures of success for the predictive models you will be producing. These measures should go beyond standard accuracy metrics and may include areas under the receiver operating characteristic curve (AUC-ROC), F1-score, precision-recall curves, etc. Explain the rationale behind these measures and their suitability for the research question.

3.2. Data division: Describe the process of how you divided your data into training, validation (if applicable), and test sets. Explain the rationale behind the data division, considering strategies like temporal validation or stratified sampling.

Individual Component

The report must include a dedicated section for **each group member**. Each section should clearly state the **member's Unikey** to identify their individual contribution / component:

The report must include a dedicated individual section for each group member. Each section should clearly state the member's Unikey to identify their individual sections in the report (THIS IS A UNIKEY: ABCD1234). DO NOT PROVIDE STUDENT ID OR STUDENT NAME TO IDENTIFY ANY OF THE SECTIONS.

1. Predictive model

Note: Each member must choose a different predictive modelling technique.

1.1. Model Description: Name and describe your technique, discuss the assumptions underlying this technique, and critically evaluate their validity in the context of the dataset. Highlight the strengths and limitations of the chosen technique and justify its suitability for the research question and dataset characteristics. Modelling techniques **not covered** in the tutorial sessions, such as neural networks (CNNs, LSTMs, RNNs, GANs, etc) or including bagging or gradient boosting techniques (GBM, XGBoost, LightGBM, CatBoost, AdaBoost, etc.) are preferred.

1.2. Model Algorithm: Provide a detailed explanation of the algorithm powering your chosen technique, including its underlying principles, such as (but not limited to) mathematical equations, hyperparameters, and potential variations. Using pseudocode or flowchart diagrams, provide the step-by-step execution of the algorithm. ***(You can type the pseudocode in Jupyter Notebook and put the screenshot of the pseudocode here. You cannot put the screenshot of the pseudocode in the appendix. If you do, it will not be marked). If you choose to draw a flowchart, you can create it on any online tool or software and attach its screenshot here. You must put the screenshot of the flowchart diagram here, in the main report. If you put it in the appendix, it will not be marked.***

1.3. Model Development: Describe the process of building the predictive model, including advanced data preprocessing techniques such as feature scaling, dimensionality reduction (e.g., Principal Component Analysis), or feature engineering. Discuss the selection of model-specific functions and hyperparameters, providing theoretical justification and empirical validation. Also, you will identify the Python functions and chosen parameters you selected and what they mean.

Note: You don't have to include the code in the report, as you will submit it separately.

2. Model Evaluation and Optimization

2.1. Model Evaluation: Perform a comprehensive evaluation of your model's performance using the agreed-upon measures of success. Interpret the results in the context of the research question and dataset characteristics, considering factors such as class imbalance, noise, and interpretability. Discuss the implications of the evaluation metrics and identify potential areas for improvement.

2.2. Model Optimisation: Explore advanced optimisation techniques to further enhance your model's performance, explaining your choices clearly. This may involve hyperparameter tuning using techniques like grid search.

<u>Group Component 2</u>

Finally, a second group section at the end of the report should include:

1. **Discussion:** Engage in a critical discussion on the strengths and limitations of each modelling technique employed by group members. Compare and contrast the performance of various models quantitatively and qualitatively. Reflect on the broader implications of model selection for addressing the research question effectively.

2. **Conclusion:** Synthesize the findings from individual model evaluations and provide a recommendation on the most effective predictive model for answering the research question. Justify your recommendation based on empirical evidence, theoretical considerations, and domain knowledge. Propose potential avenues for future research, including data collection strategies, model refinement techniques, and interdisciplinary collaborations.

Appendix Section in Report

Your report **MAY** have an appendix section. The appendix section must be placed at the very end of the full report, that is after all the individual and group components. The appendix section can also include your references, if you have used any. Any information in the appendix will not be counted towards marking. All the essential figures that you choose to mention must be provided in their respective sections of the main report.

Code and Dataset

Along with the report, you must also submit the Python code used in this assignment as a ***single zip or tar.gz folder and that folder must be named using the following convention: "GroupX_A2",*** **where X is your group number. DO NOT SUBMIT A FOLDER THAT IS NAMED GroupX_A2.** This compressed folder must contain the following items:

1. The Jupyter notebook (named "***GroupX_A2_Code.ipynb***") with all the code mentioned below. Therefore, each group will only have 1 code file that will contain all the code for each of the members. **Each member's code must be identified by that member's unikey (DO NOT PROVIDE STUDENT ID OR STUDENT NAME)**. (You can create a markdown cell to mentioning the student's unikey to indicate their respective section).

2. The final cleaned version of the dataset in CSV format, named "***GroupX_FinalCleanData.csv***".

The Jupyter Notebook Python file must contain the following:

- **Group Component 1:**
  - **The beginning of this section must be clearly marked and labelled. We recommend using a Markdown Cell that says "Group Component 1".**
  - **Preliminary Changes to Data**: all steps used to further process the data, such as adding previously dropped features, removal of existing attributes, and conversion of data type of attributes to a more suitable data type.

  - **Data split for train/validation/test sets:** all the steps required to split the data into training, validation (if required), and testing sets, including (but not limited to) strategies like temporal validation or stratified sampling.
  - **The end of this group component section must be clearly marked and labelled.**

- **Individual Component:**
  - ***Title and Unikey Markdown: Before starting the code for the individual component, you must create a markdown cell and provide the title, which states the model name and the Unikey of the student who will work on that model.***
  - **Data Pre-processing:** Provide all relevant data pre-processing techniques implemented, such as PCA, feature scaling, etc.
  - **Initial Model Development and Evaluation:** All the model building steps required to successfully construct and train the predictive model.
  - **Model Optimisation:** All the hyperparameter tuning steps, the results of each hyperparameter test run, and the decision of the optimal hyperparameter configuration that will be used by the student.
  - **Model Results:** All the results, including relevant graphs (curves), plots, accuracy percentages, and other relevant metrics of model performance that show the performance behaviour for the optimal model, must clearly be depicted.
  - **The end of each of the individual sections must be clearly marked and labelled.**

- **Group Component 2:**
  - **The beginning of this group component section must be clearly marked and labelled. We recommend using a Markdown Cell that says "Group Component 2".**
  - **Optimal Model Comparison:** All the optimal models must be compared and contrasted with each other.
  - **Final Model Recommendation:** Once all the models have been compared, the final recommendation of the most appropriate model must be mentioned, including its performance metrics and scores obtained after testing.
  - **The end of this group section must be clearly marked and labelled.**

COMP5310 Project Stage 2

Develop and evaluate a predictive model

Submission Portals

1. Please upload the report in PDF format, named "***GroupX_A2_Report.pdf***", in the Report submission portal.

2. Please upload the main group folder, named "***GroupX_A2***" containing the code and final clean datasets in the Code and Dataset submission portal.

KEY POINTS OF INFORMATION

1. ***In both the report and the code file, each student must designate their own section / component using their Unikey (THIS IS A UNIKEY: ABCD1234). DO NOT provide Student ID or the name of the student.***

2. ***Your report MUST HAVE a front page. The front page of the report MUST contain Assignment Title which is Project Stage 2, Group Number, the Student ID (SID) of each member and the UNIKEY of each member. DO NOT provide names of any group member. DO NOT PROVIDE any additional information.***

3. ***DO NOT include a Content's Page at any part of your report. It is not required. Adding a content's page will be counted towards the total page count and marks will be deducted if any report section goes beyond the permissible page count limit set forth in the assignment guide.***

4. ***ONLY 1 SUBMISSION PER GROUP is required. In other words, only 1 member from the group must submit the assignment on Canvas.***

5. ***Each report may have an appendix section. The appendix section must be placed at the very end of the full report, that is after all the individual and group components. Any information in the appendix, apart from the references, will not be counted towards marking.***

6. ***The different report sections and sub-sections are aligned with the marking rubric. Therefore, please INCLUDE ONLY the requested contents and DO NOT MIX OR MERGE THE SECTIONS, as this will interfere with the marking process. If you fail to do so, this won't be considered for marking.***

7. ***You MUST ONLY USE Jupyter Notebook for your code. You are NOT PERMITTED to use any other IDE, such as Google Colab, Spyder, etc for your code file. MARKS WILL BE DEDUCTED if students require the markers to run the code file in any IDE other than Jupyter Notebook.***

8. ***Students must follow the report format exactly as given in the assignment guide. DO NOT add your own sections or sub-sections to the report. DO NOT RENAME ANY SECTION HEADINGS. Simply follow the report format mentioned in this assignment guide. Providing your own section headings or following a format different than what is mentioned in the assignment guide will lead to those sections being ignored by the marker. No further appeals will be entertained.***

9. ***In the case of misnamed sections, sections being absent, or sections not following the right order as mentioned in the assignment guide, that specific section will be completely ignored by the marker.***

10. ***Ethical AI usage, critical thinking, and originality are fundamental expectations in all assignments. Ensure that your work is independently produced and reflects your own understanding.***

MARKING

Marking Criteria		Marks
Group Component 1		
	Topic and research question	1
	Dataset	1
	Setup	2
Individual Component		
	Predictive model	6
	Model evaluation and optimisation	6
	Model complexity	1
	Code quality	1
Group Component 2		
	Discussion	4
	Conclusion	2
	Report format and presentation	1
	TOTAL	**25**

Deductions

- 1 mark will be deducted from a specific student if their individual section of the report exceeds the maximum number of pages. If the group section exceeds the maximum number of pages, the deduction will apply to all group members.
- 1 mark will be deducted from the team member whose unikey is not mentioned at the commencement of his or her respective code section and report section.
- 5% of the maximum awardable mark will be deducted per day of late submission. Zero marks will be awarded after 5 calendar days from the due date.

References

You can refer to the link provided below to understand what a Flow Chart and Pseudocode are and how to create them using appropriate standards and norms:
https://www.youtube.com/watch?v=O8vPR3zh5go