COMP5310 Project Stage 2

Develop and evaluate a predictive model

THE UNIVERSITY OF SYDNEY

Due: 11:59PM on 6th of May 2024 (Week 11)

*This assignment is worth **20%** of the final mark of the unit of study.*

GROUPS

This stage is usually 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 TA daniela.rivasromero@sydney.edu.au with copy to your tutor to discuss this.

***Note:** there is work required from each member separately, but the project is handed in as a combined effort, and it is marked as a whole: there will be individual and group components to the marks, all 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 TA daniela.rivasromero@sydney.edu.au. Make sure that your email specifies the lab session and 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) and your lab tutor.

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 Monday 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 dataset chosen by the group in stage 1. This stage will involve advanced predictive modelling techniques and thorough model evaluation and optimization processes.

Note: *If you decide to change the dataset and/or research question, this needs to be notified and approved by your tutor.*

DELIVERABLES

Report

The report should have a maximum of 3 pages for each individual section and maximum 2 pages for the group section. It should use high-level headings, as provided below, to indicate the different sections and sub-sections of the report and use line spacing of at least 1.15 and body font size of at least 10pt. The goal is to convey the problem clearly and concisely.

The report should be targeted at a tutor whose goal is to see what you did, so they can allocate a mark. It should have a front page that gives the group name and lists the members involved (giving their SIDs and unikeys, **NOT their names**), and then the body of the report has a structure as follows (this corresponds to the marking scheme):

Group Component 1

The report should begin with a group section including:

1. **Setup**

1.1. **Topic and research question:** Describe the research problem comprehensively, emphasizing its significance in the domain. 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.

1.2. **Dataset:** Provide a detailed overview of the dataset, including its source, size, and complexity. Discuss any challenges or biases present in the data and how they might impact the modeling process.

1.3. **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 area 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.

Individual Component

The report should follow with a section per group member (state the member's unikey), with:

1. **Predictive model**

   Note: Each member needs to 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.

   1.2. **Model algorithm:** Provide a detailed explanation of the algorithm powering your chosen technique, including its underlying principles, hyperparameters, and potential variations. Illustrate the step-by-step execution of the algorithm and discuss any recent advancements or adaptations relevant to your application.

   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. Explain the rationale behind the division of data into training, validation, and test sets, considering strategies like temporal validation or stratified sampling. 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.

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

Group Component 2

There should be a final group section including:

1. **Discussion:** Engage in a critical discussion on the strengths and limitations of each modeling 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.

COMP5310 Project Stage 2

Develop and evaluate a predictive model

Code and Dataset

You must also submit a copy of the dataset the group used for stage 2 of the project, alongside the **Python code with well-structured organisation and a clear explanation** each member used for developing and evaluating their predictive models. This should be submitted as a **single zip or tar.gz folder**. This compressed folder should contain a subfolder for each member of the group, using their unikey as name of the folder. Then, each subfolder should contain the **clean dataset** from stage 1 and the **Python code** in a Jupyter Notebook used for developing and evaluating their predictive model.

COMP5310 Project Stage 2

Develop and evaluate a predictive model

MARKING

Marking Criteria	Marks
Group Component 1	
Setup	2
Individual Component	
Model description	2
Model algorithm	2
Model development	3
Model evaluation	2
Model optimization	2
Model complexity	1
Code quality	1
Group Component 2	
Discussion	3
Conclusion	2
TOTAL	**20**

Deductions

- 5% of the maximum awardable mark will be deducted if your 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.
- 5% of the maximum awardable mark will be deducted per day of late submission. After ten calendar days late, a mark of zero will be awarded.
- Deductions will apply if failure to follow instructions provided.