INFS4203/7203 Project Phase II (20 marks)

Semester 2, 2024

Due date:

16:00 on 25th October 2024 (Brisbane Time) (Phase II, 20%)

All assignments should be submitted to the UQ Blackboard. If any assignment fails to be submitted appropriately before due, a penalty will be applied according to the ECP. Please take the responsibility to ensure your submission is successful before due time. Email submission will not be accepted.

Overview

In Phase II, you will implement your proposal submitted in Phase I, with necessary **adjustment** according to the empirical performance and the feedback from the proposal. This is an individual assignment. The completion of the assignment should be based on your own design and feedback from the proposal.

Track 1: Data-oriented project

In Phase II, you will be provided with the test data named *test_data.csv*. The first row describes features' names. Except the first row, each row in the data file corresponds to one data point. There are 817 test data points in this file, and each column represents the same feature as the training data *DM_project_24.csv*. Note that the test data only has 105 columns, without labels, i.e., without the final column "Target (Column 106)" in training data *DM_project_24.csv*. Labels for the test data will not be released and will be used by the teaching team for marking only.

In this phase, you will need to implement the ideas in your proposal and classify the test data. In the marking phase, "F1" of the test data will be used for making. When calculating F1, "1" is counted as positive label, and "0" as negative label. **You need to submit:**

- A *result report* on
  - Test result: the prediction on test data (in integer type) and
  - Evaluation result: the evaluated accuracy and F1 on the training data using cross-validation (in float type).
- Code and *readme* file, which include
  - Readme file, which should include
    - **Final Choices:** The final pre-processing methods, classification model, and hyperparameters you used to achieve your reported test results.

        - **Environment Description:** A clear and thorough description of your coding environment (operating system, programming language and version, additional installed packages, etc.).
        - **Reproduction Instructions:** Detailed instructions on how to run the code so that your reported results for pre-processing, model selection, hyperparameter tuning, testing, and evaluation can be reproduced.
        - **Additional Justifications:** Any additional justifications or references for the methods you implemented.
        - The readme file can be in text format, such as .md, .docx, .pdf, or .txt.
    - Training, Evaluation, and Testing Code:
        - Training Procedures: All code related to pre-processing, training on the training data, prediction on the test data, and generation of the result report.
        - The code must include a main function in a main file (for example, main.py) to execute the overall process.
        - Please fix the random seeds to ensure your results are reproducible.
    - Pre-processing Selection, Model Selection, and Hyperparameter Tuning Procedures – how you made the final choice of the selected pre-processing, model and hyperparameters:
        - Include the code for the detailed procedures related to pre-processing selection, model selection, and hyperparameter tuning.
        - If you need further explanation of the overall selection and tuning procedure, you may put them in the Readme file
        - Please fix the random seeds to ensure your results are reproducible.
- Additional requirements
    - Please include the provided training and test files into your submitted *.zip* file for reproducing your results in the marking phase. The generated result report file (same as submitted) should be in the root directory.
    - Any programming language can be used. However, if you use Python, please submit *.py* files instead of any other formatted files. If you use the Jupyter Notebook or Colab, please submit *.py* file instead of *.ipynb* file.
    - Please submit your best prediction according to your cross-validation results. Multiple test results submitted will not be marked

Format for the Result Report

- The result report should be named as *sxxxxxxx.infs4203 (sxxxxxxx, an s followed by **seven** digits is your student username)* with the same **Submission Title** when submitting through the "Report Submission" Turnitin link provided. For example, if your student username is *s1234567*, then the result report should be named as *s1234567.infs4203* and submitted with the same Submission Title.
- The result report should be composed of 818 rows. For the first 817 rows, the ith row gives the prediction of the ith test instance, either 1 or 0 (in integer type). The last row (row 818) gives the accuracy (first column, rounded to the nearest **3rd decimal place**) and F1 (second column,

rounded to the nearest **3rd decimal place**) evaluated by yourself through cross-validation on the training data, both in float type.

- Please separate the values in each row and column with **commas**, and ensure each row ends with a **comma**.
- You could refer to *result_report_example.infs4203,* which provides an example (Note: This is NOT the groundtruth) of the result report.

Note that result report submitted in other forms or names will not be accepted or marked.

Format for the Code and Readme File

- Together with the result report, you need to submit a *readme* file and all your *codes*.
- The *readme* file and your *codes* should be compressed into **one** zip file named *sxxxxxxx.zip (sxxxxxxx is your student username)* with the same **Submission Title** when submitting through the "Readme and code submission" Turnitin link provided.

Note that code and readme file submitted in other forms or names will not be accepted or marked.

We recommend you follow the Google Style Guides (https://google.github.io/styleguide/) for the programming style. Following such style is not mandatory for this assignment but using it may benefit your future career as a data scientist!

Submission

Only your last submitted version will be marked. All required files need to be submitted before due. Otherwise, penalty will be applied according to ECP, i.e.,

A penalty of 10% of the maximum possible mark will be deducted per 24 hours from time submission is due for up to 7 days. After 7 days, you will receive a mark of 0.

- Result report should be submitted through the "Report submission" Turnitin link provided on Blackboard -> Assessment -> Project Phase II -> Report submission before the deadline, with the Submission Title *sxxxxxxx.infs4203*.
- Compressed file of readme and codes should be submitted through the "Readme and code submission" Turnitin link provided on Blackboard -> Assessment -> Project Phase II -> Readme and code submission before the deadline, with the Submission Title *sxxxxxxx.zip*.

Marking standard

Submissions satisfying the following four conditions will be accepted and marked

1. The selected best pre-processing, model and hyperparameter can be reproduced from the submitted readme file and codes.
2. The classifiers used to do classification can be reproduced by the submitted readme file and codes.
3. The classifiers are generated by using only techniques delivered in INFS4203/7203 lectures.
4. The test and evaluation results can be reproduced by the submitted readme file and codes.
5. The test and evaluation results are generated by applying the learned classifiers to the data.

When the above five conditions are satisfied, the result report will be marked according to the F1 result on the test data in the following way (rounded to the nearest 1st decimal place)

- For F1 less than or equal to 0.3: Mark = F1÷0.3
- For F1 greater than 0.3 but less than 0.75: Mark = F1÷0.03- 9
- For F1 greater than or equal to 0.75 but less than 0.79: Mark = 100*F1-59
- For F1 greater than or less than 0.79: 20
- Please see the example below

F1	Mark
0.3	1
0.33	2
0.36	3
0.39	4
0.42	5
0.45	6
0.48	7
0.51	8
0.54	9
0.57	10
0.60	11
0.63	12
0.66	13
0.69	14
0.72	15
0.75	16
0.76	17
0.77	18
0.78	19
0.79	20

Training time or prediction time will not be counted into marking.

(End of Track 1. See the next page for Track 2 specifications.)

Track 2: Competition-oriented project

In this phase, you need to submit:

- A *result report* of the Public Leader Board results, including a screenshot and an URL of the Public Leader Board.
- A *readme* file with clear and thorough *description of your coding environment* (operation system, hardware requirement, programming language and its version, additional packages installed etc.) and instructions on how to run the code such that your final submission to Kaggle can be reproduced
- Your *implemented codes* including training and test codes which have a main function to generate the final submission to Kaggle.

Marking Standard

The following marking standard will apply, unless otherwise discussed with the teaching team for exceptionally challenging competitions.

You need to submit the evidence of your achievements in the public leading board by the end of the project deadline to earn your marks. **Your username in the public leading board must be your student username (sxxxxxxx, each x represents a digit)**.

If your targeted competition ends before the project deadline, you could show by cross-validation that you have achieved comparable performance to a particular competitor on the public leading board before the project deadline. Your project could then be assessed by the competitor's corresponding rank percentage on the public leading board.

You have to earn a public Leader Board top ranking index (your rank divided by the total number of competitors) by the project deadline

Earned marks = max (20 – max (public_LB_top_ranking_index – 0.4, 0)*30, 0)

That is, you earn 20 marks when having Public Leader Board top ranking to be within top 40% of all competitors.

Format

- The result report should be named as *sxxxxxxx.pdf or sxxxxxxx.doc/docx (sxxxxxxx is your student username)*. For example, if your student username is s1234567, then the result report should be named as *s1234567.pdf/doc/docx*.

Note that result report submitted in other forms or names will not be accepted or marked.

- Together with the report, you need to submit all your *code* and a *readme* file. The *readme* file and your *code* should be compressed into **one** zip file named *sxxxxxxx.zip (sxxxxxxx is your student username)*.

Note that code and readme file submitted in other forms or names will not be accepted or marked.

Submission

Only your submitted version will be marked. All required files need to be submitted before due. Otherwise, penalty will be applied according to ECP.

- Result report should be submitted through the "Report submission" Turnitin link provided at Blackboard -> Assessment -> Project Phase II -> Report submission before the deadline with the Submission Title *sxxxxxxx.pdf or sxxxxxxx.doc/docx*.
- Compressed readme file and code should be submitted through the "Readme and code submission" Turnitin link provided at Blackboard -> Assessment -> Project Phase II -> Readme and code submission before the deadline with the Submission Title *sxxxxxxx.zip*. Note that the zip file should be smaller than *100MB*. **If your file is larger than *100MB*, please contact infs4203@eecs.uq.edu.au before due time by email in case there is any penalty applied to later submission.**

End of Specification for Phase II