FTEC4003 Data Mining for FinTech Course Project Assignment

1 BASIC RULES

1.1 Due Date

Submission of the report: **23:59:59 on December 17th (Sunday), 2023**.

1.2 Reminder

- This is a group project including two tasks. The number of team members in each team is up to 3.
- You are **NOT** allowed to **COPY** code/report from the Internet or others (unless specified for some exceptional cases). Any plagiarism case will be seriously punished.
- The assessment will be based on your results, submitted files, and report.
- Pleases send your group information to **ftec4003@se.cuhk.edu.hk** before **23:59:59 on November 12th (Sunday), 2023**.
- For late submission, a penalty per day will be applied after the deadline (30%, 30%, and 40% for the following days, respectively). You won't get any marks for more than a **3-day delay**. Please submit your assignment before the deadline.
- Language: **Python 3.x**. You can use any package you like.
- Operating System Platform: Windows / Linux / macOS.
- You are strongly encouraged to read the tutorial materials on the **Blackboard**.

1.3 Marking Scheme (Total: 19 marks)

- Task 1: 8 marks
- Task 2: 11 marks

2 TASKS

2.1 Task 1: Travel Cancellation

The first task is to conduct a classification task with Python 3.x and compare the performance of several standard methods learned from class. The detailed requirements are described as follows.

- Run the classification task using all methods among Decision Tree, k-Nearest Neighbor, Naive Bayes, SVM, and Ensemble Methods. As for the Ensemble Method, choose one from the three learned methods, i.e. bagging, AdaBoost, and random forest. Compare the performance of the two best methods in your report. Please show how you have tuned the basic parameters (those covered in the lecture) and justify your final choice of the parameters according to your experimental analysis.
- Description of datasets: Please refer to the file **Task-1-Travel-Cancellation.pdf** under the directory **Task-1-Travel-Cancellation** for details.

- Output: For each item in **test.csv**, you need to predict its class (1/0). Please store your result in a file named **GID_submission_1_method.csv** (replace "GID" with your group ID and "method" with the best two method names. e.g., 1_submission_1_svm.csv). The format should be the same as **sample_submission_1.csv**. It would help if you were careful about **the number of lines** and the predicted result, which should be **1 or 0**.
- In your report, record the performance of the classification task. Please use the command line tool named **evaluate_1** (tool names may get a little different depending on the platforms) under the directory **Task-1-Travel-Cancellation** to get the performance of your result. We will use the F1-score of the "1" class to measure your submission.

2.2 Task 2: Credit Evaluation

The second task is a competitive classification task. Please achieve as high a score as you can. Methods are unlimited in this task (i.e., you can use the techniques not covered in this class). The detailed requirements are described as follows.

- The champion and runner-up for **this task** will get an award certificate
- Descriptions of the datasets can be found in the file **Task-2-Credit-Evaluation.pdf** under the directory **Task-2-Credit-Evaluation**.
- Output & report: The output is similar to task 1 except that you should store your result in the **GID_submission_2.csv** and evaluate the result via **evaluate_2** (tool names may get slightly different depending on the platforms). The format should be the same as **sample_submission_2.csv**. It would help if you were careful about **the number of lines** and the predicted result, which should be **1 or 0**. We will use the F1-score of the "1" class to measure your submission.

3 SUBMISSION GUIDELINES

3.1 What to submit

- A README file. Please name it **README.txt** or **README.md**. This file should include the following sections
  - Student numbers and names of all team members.
  - A brief description of all files.
- Code files (i.e., *.py). Please make sure your code can be **reproduced** so that we can verify your result.
- Output files (i.e., **submission_1.csv** and **submission_2.csv**).
- A file named **FTEC4003_report_GID.pdf**, where GID denotes your group ID. The file should include a brief description of the platform, the method, experimental evaluations, results, and conclusions of the two tasks. Please show your names and student numbers on the cover page of your report.

3.2 Submission instructions

- Please package all your files (including the **README.txt** or **README.md**, code files, output files, and report **FTEC4003_report_GID.pdf** into a **ZIP** file named **FTEC4003_project_GID.zip**), where GID is your group ID.
- Submit the package file with the Subject **FTEC4003 SUBMISSION GID** to the course mail **ftec4003@se.cuhk.edu.hk**, where GID is your group ID. (Please do use upper case in the Subject to ease the submission process).