FFTEC4003 Data Mining for FinTech Course Project Assignment

Due Date

- Submission of the report: 23:59:59 on December 18 (Sun.), 2022.

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.
- You are not allowed to use any information from the test data (e.g., the output in our evaluation).
- Please send your group information to ftec4003@se.cuhk.edu.hk before 23:59:59 on November 13 (Sun.), 2022.
- 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**. 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 website.

Marking Scheme (Total: 19 marks)

- TASK 1: 8 marks
- TASK 2: 11 marks

TASK 1: Debt Default Prediction.

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

1. Run the classification task using all methods among DecisionTree, k-Nearest Neighbor, Naive Bayes, SVM, and Ensemble Methods. As for the Emsemble 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.
2. Description of datasets: Please refer to the file README.md under the directory Task-1-Debt-Default-Prediction for details.
3. Output: For each item in assignment-test.csv, you need to predict its class (1/0). Please store your result in a file named submission_1_method.csv (replace "method" with the best two method name. e.g., submission_1_svm.csv). The format should be the same as samplesubmission.csv. It would help if you were careful about the number of lines and the predicted result, which should

be 1 or 0.

4. 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-Debt-Default-Prediction to get the performance of your result. We will use the F1-score of the "1" class to measure your submission.

TASK 2: Startup Failure Prediction.

1. This is a competitive classification task. Please achieve an as high score as you can. Methods are unlimited in this task (i.e., you can use techniques not covered in this class).
2. The champion and runner-up for this task will get an award certificate.
3. Descriptions of the datasets can be found in a README file under the directory Task-2-Startup-Failure-Prediction.
4. Output & report: The output is similar to task 1 except that you should store your result in file 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 samplesubmission.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.

Submission Guidelines

What to Submit

1. A README file. Please name it README.txt or README.md (the latter is recommended). This file should include the following sections:
    - Student numbers and names of all team members.
    - A brief description of all files.
2. Output files (i.e., submission_1.csv and submission_2.csv).
3. A file named FTEC4003_report_XX.pdf, where **XX** 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.

Submission Instructions

1. Please package all your code files (including the README.txt (or README.md), the output files, and your report FTEC4003_report_XX.pdf into a **ZIP** file named FTEC4003_project_XX.zip, where **XX** is your group ID.
2. Submit the package file with the Subject **FTEC4003 SUBMISSION XX** to the course mail, **ftec4003@se.cuhk.edu.hk**, where **XX** is your group ID. (Please do use upper case in the Subject to ease the submission process)