FTEC4003 Data Mining of FinTech Course Project

Due Date

- Submission of the report: 23:59:59 on December 3 (Thu.), 2020

Reminder

- This project consists of **TWO** group tasks. The number of team members in each team is up to 3.
- You are **NOT** allowed to **COPY** code/report from Internet or others (unless specified for some special cases). Any plagiarism case will be seriously punished.
- The assessment will be based on your results, submitted files and report.
- For late submission, a penalty per day will be applied after the deadline (10% for both the first two days, 20%, 30% and 30% for the following days respectively). You won't get any marks for more that **5-day delay**. Please do submit your project before the deadline.
- Language: **Python 3**. You can use any package you like.
- Operating System Platform: Windows / Linux / Max OS.
- You are strongly encouraged to read the tutorial materials on blackboard website.
- The teaching team reserves the right to withdraw the eligiblity of the team for the project awards (Ref. Task 2 for more details) if plagiarism is discovered.

Marking Scheme (Total: 20 marks)

- TASK 1: 8 marks
- TASK 2: 12 marks

TASK 1: Insurance Selling

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

1. Run the classification task using all these methods (DecisionTree, k-Nearest Neighbor, Bayes, SVM and Emsemble Method). As for the Emsemble Method, choose one from the three learned methods, i.e., bagging, adaboost, and random forest. Compare the performance of the best 2 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 for details. Datasets of task 1 are under directory Task-1-Insurance-dataset.
3. Output: For each item in insurance-test.csv, you need to predict the class (1/0) it belongs to. Please store your result in a file named submission_1_method.csv (replace **"method"** with the best 2 method name. e.g. submission_1_svm.csv). The format should be exactly same as samplesubmission.csv. You should be careful about the number of lines and the predicted result (1 or 0).
4. In your report, record the performance of the classification task. Please use the executable file (evaluate_1.exe / evaluate_1.linux / evaluate_1.macOS) from command line tools like cmd (in windows), terminal (in linux), terminal (in macOS) under directory Task-1-Insurance-dataset to get the performance of your result. We will use F1-measure to measure your submission.

TASK 2: How many Customers Stay

1. This is a competitive classification task. Please achieve 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 (instead of task 1) will get an award (Airpods Pro for the champion while BOSE SoundLink Micro for the runner-up).
3. Description of the datasets can be find in a README file under Task-2 How many Customers Stay.
4. Output & report: The output is similiar to task 1 except that you should store your result in file submission_2.csv and evaluate the result via evaluate_2. The format should be exactly same as samplesubmission.csv.We will use F1-measure to measure your submission. You should be careful about the number of lines and the predicted result (1 or 0).

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_method.csv and submission_2.csv).
3. A file named FTEC4003_report_XX.pdf, where **XX** denotes your group ID. The file should include brief description of the platform, the method, experimental evaluations, results and conclusions of the two tasks. Please show your names and student numbers in the cover page of your report.

4. All of your program. Please add some necessary comments into your program. Note that we will run your code.

Submission Instructions

1. Please package all of your files (including the README.txt (or README.md), the output files, your report FTEC4003_report_XX.pdf) and your program into a **ZIP** file, named as 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)