University of Wollongong
School of Computing and Information Technology
CSCI946 Big Data Analytics
Autumn 2024

Assignment 3

(20 marks)

Due: 23:59 Beijing Time, 29 November 2024

Aim

This assignment aims to provide students with essential experience conducting big data analytics experiments with the R or the Python programming language. In this assignment, you should

- procedure big data analytics by following Big Data Analytics Lifecycle,
- appropriately choose, apply and evaluate core models/algorithms and analytics techniques to complete the analysis tasks,
- understand and integrate the knowledge and skills learned in this subject, including big data analytics lifecycle, data preparation, clustering, classification, regression, association rules, data/model evaluation, data visualization, and text/image/web data processing.

Group work: You are to work on this assignment as a group. Each group is to work independently from other groups on this assignment. Groups and group memberships are as specified on Moodle. You can form groups of your own accord. Each student can only join in one group. Each group should contain no more than 5 members and no less than 3 members. All group members are expected to contribute to this assignment. Please plan before starting the assignment, then keep a detail digital work log and timesheet for each group member. A justification and/or explanations must accompany all your answers to this assignment. One submission per group only.

Penalties: If a group member fails to make a minimum contribution, the member will be awarded zero marks. Claims of less or no contribution should provide evidence like a work log. Plagiarism of any part in this assignment will result in zero marks being awarded to the whole group.

Preliminaries

Read through the lecture slides, lab instructions and the recommended readings in Weeks 1 – 12. Conduct relevant background studies. You should use either R or Python for the tasks in this assignment. You can use any publicly accessible toolbox of library for R and Python. Your submission must include the source code file(s) which, when run, would re-create all your results.

About Dataset and Original Project

Assignment 3 uses a financial transaction dataset on Kaggle.(https://www.kaggle.com/datasets/ealaxi/paysim1?resource=download). The original dataset is in 493.53 MB, which may not be able to process on a personal computer. Instead, A3 provides a resized dataset – **A3dataset.csv**. In the emerging mobile money transactions domain, financial datasets are important to many researchers and in particular to us performing research in the domain of fraud detection. Kaggle presents a synthetic dataset generated using the simulator called PaySim as an approach to such a problem. PaySim uses aggregated data from the private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behavior to later evaluate the performance of fraud detection methods.

PaySim simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world.

This synthetic dataset is scaled down 1/4 of the original dataset and it is created just for Kaggle. We further scale down 85% for A3 with the purpose of practice.

NOTE: Assignment 3 is different to any public project. Copy from any public project will lead to zero mark for Assignment 3.

Essential Tasks of Assignment 3

Assignment 3 aims to detect fraudulent financial transactions. Your essential tasks include the following Tasks 1-4:

Task 1: Design a big data analytics project by following Big Data Analytics Lifecycle. (3 marks)
Task 2: Perform statistical analytics by following the article "The Banking Transactions Dataset and its Comparative Analysis with Scale-free Networks". Try programming and producing figures like Fig.1 – Fig.16 and tables like Table I – Table IV. Introductions and explanations of each step and each finding are required. Tables and Figures without introduction and explanation will result in a mark deduction. (10 marks)
Task 3: Based on findings from **Task 2**, apply at least two core models where clustering and classification methods are mandatory. An application of a neural network is preferred. A3 does not limit the selections of algorithms, and the number of algorithms applied. (5 marks)
Task 4: Study factors influencing fraud detection and suggest early financial warning from the perspective of the financial service provider. (2 marks)

A report is required to summarize Tasks 1-4 in a well-organized way and cite referred articles and programming resources in your writing. Tasks 1-4 may need R/Python programming to support your analysis.

Submission:

The submission link for Assignment 3 is on the subject's Moodle site. Only one submission per group. **The submission must be two files. One is the report (mandatory) named in "`A3.pdf`"; another is a zip file named "`A3.zip`", under 200 MB, and contains code (mandatory) and video presentation (optional).** Either following way is acceptable:

1. a report in `.pdf` format, and code files in `.R` or `.py`; or in `.ipynb`

A video presentation is optional in `.mp4` format or a shared link saved in a `.txt` file.

Important:

- The report must be in a single file and in `.pdf` or `.ipynb` format. The title page must list the full name and student ID of all members in the group. Clearly indicate members who did not make a minimum in contributions.
- The report does not have a page limit.
- The report will be checked by **Turnitin** system in Moodle site for **Plagiarism** test.
- Marks will be deducted for incomplete or vague descriptions.
- Sufficient, suitable, and legible annotation shall be provided in your code to make it easy to understand. Marks will be deducted for untidy code, code that is difficult to read, code that does not run, or code that does not reproduce the results in your report.

Note: Failure of your code to run may attract zero marks. Plagiarism of any part in your code, or any part in your report will attract zero marks for this assignment. It is the responsibility of the group to ensure that your submission does not contain plagiarized material. You may be requested to demonstrate and explain your program or explain your answer in the report. Marks are deducted if you are unable to offer an explanation. Marks will be awarded for correct design, implementation, style, completeness, and justification.

---------- ***END***----------