**THE UNIVERSITY OF HONG KONG**

**DEPARTMENT OF COMPUTER SCIENCE**

**FITE7410 Financial Fraud Analytics**

**Second Semester, 2024-2025**

**Assignment 1 – Exploratory Data Analysis (EDA)**

**(Due Date: 13 Mar, 2025 (Thu) 23:59)**

**Assessment Criteria:**

* **Plagiarism**: Please follow the guidelines laid down by our MSc Programme office.
* You are allowed to discuss the assignment with your classmates, however, you should submit your individual work. Any direct copy and paste is PROHIBITED and would be considered as PLAGIARISM.
* **LATE PENALTIES**: 50% of assignment marks will be deducted for late submissions. 0 marks if the submission is later than 2 weeks.
* **NON-COMPLIANCE PENALTIES:** 10% deduction from assignment marks will apply for failure to follow instructions.
* Assignments would be marked based on the logic, presentation and understanding of the problem; not only on accuracy.

**Objectives of this assignment:**

* Perform data cleaning and preparation.
* Explore and visualize the data to identify patterns and trends.
* Engineer new features based on domain knowledge or insights from EDA.
* Prepare a report summarizing the findings from EDA.

**Instructions of this assignment:**

1. **(50%) Exploratory Data Analysis using R**

Use the provided dataset for the mini-case study. Download the dataset (A1\_data.csv) from Moodle.

[NOTE:

* This is a modified version of IEEE-CIS Fraud Dataset. The dataset description can be found at: https://www.kaggle.com/c/ieee-fraud-detection/data
* You MUST use the provided dataset on Moodle, NOT the one from Kaggle.]

Conduct exploratory analysis of the dataset downloaded.

**a. Data Handling (20 marks)**

* **Identification and Handling of Missing Values**
  + Use appropriate methods to identify and handle missing values in the dataset.
* **Outlier Detection and Management** 
  + To identify and handle outliers using appropriate methods.

**b. Feature Distribution Analysis (20 marks)**

* **Univariate Analysis** 
  + Use histograms, box plots, etc. to analyze individual features.
* **Bivariate/Multi-variant Analysis** 
  + Use scatter plots, correlation plots, etc. to explore relationships between features.

**c. Feature Engineering (10 marks)**

* **Creation of New Features**
  + Use appropriate methods to create new features for fraud detection.

[NOTE:

* An R script sample is provided, but you must complete the program.
* Alternatively, you can build the model independently and use any R library of your choice.]

1. **(50%) Write a short report on the following:**

Describe the dataset based on the EDA result, including:

* 1. **A description of the data cleaning and preparation process. (12 marks)**
  + Provide a detailed account of the data cleaning process.
  + Discuss any transformations or preprocessing steps taken.
  1. **Visualizations of the data, with clear labels and explanations. (12 marks)**
  + Include relevant visualizations with clear labels.
  + Ensure each visualization is accompanied by a concise explanation.
  1. **A description of the engineered features and their rationale. (12 marks)**
  + Clearly describe the engineered features.
  + Justify the relevance of these features for fraud detection.
  1. **A discussion of the key findings from EDA, including insights and potential hypotheses. (14 marks)**
  + Summarize the key findings from the EDA.
  + Discuss insights gained and potential hypotheses generated from the analysis.

[NOTE:

* The short report should consist of a main body of **maximum 2-3 pages**, focusing on your analysis and insights.
* Additional figures and diagrams can be included in a separate Appendix to support your report.]

1. Submission on Moodle:
   1. **R Language Script Submission**
      * Ensure the R script is well-commented
   2. **PDF Report Submission** 
      * The report should be clear, concise, and adhere to the 2-3 page guideline.
      * Include any additional figures or diagrams in an appendix.

Additional Notes

* Reports will be checked for similarity using Turnitin.
* Ensure adherence to the instruction to use the dataset provided on Moodle, not the Kaggle version.