ECMM445 Learning from Data
Continuous Assessment

Date set: 27th October, 2022
Hand-in date: **11:59am 7th December, 2022**

This continuous assessment (CA) comprises 40% of the module assessment.

Note that electronic submission is required and instructions are provided at the end of the specification. The submission is a written report with word count 1,500 words +/-10%.

Specification

One of the main objectives of this module is to help you gain hands-on experience in communicating insightful and impactful findings to stakeholders. In this coursework, you will use the tools and knowledge you learned throughout this module to select the techniques that best suit your needs, train models on a dataset, and communicate insights you found from your modeling exercise. After detailing and implementing your methodology, you will write-up the insights that you have gathered to explain the patterns in your data and any relations to the outcome variable.

You are expected to leverage a variety of computational tools within a Jupyter notebook using Python and relevant machine learning libraries and produce results that are reproducible, robust and the product of good experimental design based on a sound understanding of the methods used.

Before you begin, you will need to choose a dataset from the selection below. Below are the links to 5 data sources:

1. HESA Higher Education Student Data. URL: https://www.hesa.ac.uk/data-and-analysis/students
2. Stock exchange data. URL: https://www.kaggle.com/mattiuzc/stock-exchange-data
3. Fortune 500. URL: https://data.world/aurielle/fortune-500-2017
4. AT&T stock price data. URL: https://www.kaggle.com/konstantinparfenov/att-sbc-stock-price-data/version/1
5. Security and vulnerabilities. URL: https://www.cvedetails.com/

1 Required

Once you have selected a data set, you will produce a report containing the sections listed below:

1. Introduction
    - A. Outline the main objectives of the analysis and research question.
    - B. Detail the context of the research question and previous findings.
2. Methodology and Dataset
    - C. Describe the technical features of the dataset you chose and a summary of its attributes.
    - D. Document the methodology for data exploration, data cleaning, feature engineering, modelling and analysis.

3. Results
    - E. Summarise training or modelling outcomes from at least two machine learning models. For regression, the modelling approaches should include multiple linear regression, polynomial regression, LASSO regression and ridge regression. For classification, the modelling approaches should include multilayer perceptron, convolution neural network and variants of multilayer perceptron or convolution neural network. For clustering, the modelling approaches should include K-means, hierarchical clustering, DBSCAN and OPTICS.
    - F. Explain the comparison of your models and the outcomes. Conclude with a recommendation about the accuracy and explainability of the data and modelling.
4. Discussion
    - G. Summarise key findings and insights referring back to the original research question and objectives and provide supporting evidence from the results data derived from your models.
    - H. Suggest next steps in analysing this data. The explanation should identify the limitations of the modelling approaches used and may include suggestions for revisiting this model and adding specific data features or datsets to achieve a better explanation or a better prediction.

Compulsory: Please submit a PDF file containing the deliverables A to H. You should include the visuals from your code output, but this report is intended as a summary of your findings, not as a code review.

Optional: You may submit your code as a python notebook (.ipynb file) or as a print out in the appendix of your main PDF report.

Grading

The grading will be based on four sections. Note, each section has a rubric to breakdown the expectations of three main classification boundaries for a pass (40%+), second-class (50%+) and first-class (70%+) submissions. You should aim to exceed these expectations to achieve over 80% in any section:

1. Introduction: Does the report include accurate background context for the research question and detail the main objective(s) of this analysis [10 marks]?
    - This report makes a basic statement about the data source.
    - The introduction includes a detailed subtask section and a good vision of what is possible and interesting to do with this dataset.
    - In addition to clear subtasks and vision for this analysis and understanding of the dataset context, it also anticipates possible snags that might be incorporated into preliminary hypothesis of the data based on literture.
2. Methodology: Does the report describe the data and the methods used accurately [10 marks]?
    - There is a basic summary of the analysis steps and variables are available.
    - There is a technical description of the data and a reproducible methodlogy. Typically this is linked to a JuPyter notebook.
    - The summary of the data is presented with graphs of distributions and plots that show the relation between features and the outcome variable. The methodlogy is repeatable, reproducible and replicable through. This will be linked to a well structured and professional Jupyter notebook.
3. Results: Does the report include a section with variations of machine learning models and specify which one is the model that best suits the main objective(s) of this analysis [10 marks]?
    - The results provide a basic a machine learning model output.
    - Two or more machine learning models are included and the results are clearly discussed and compared.
    - Two or more machine learning approaches are evaluated. Each approach is evaluated with varying hyper parameters and pre-processing steps. The alternative appoaches and findings from adjusting hyper parameters are then presented for comparison.
4. Discussion: Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis [10 marks]?
    - There conclusions are basic and provide only high level insights or findings about this problem.
    - A series of conclusions are derived from the modelling results. The interpretation of the results is extended to consider the research question and objectives.
    - An evidence based set of conclusions are draw from the results. The modelling approach results are interpreted to address the research question and highlight flaws in the data and analysis techniques that limit the ability to understand further the research question. A plan of action to extend the analysisis proposed.

Submitting your work

The CA requires electronic submission to the BART online submission platform.

Electronic You should submit your pdf and jupyter notebook code via the electronic submission system at `https://bart.exeter.ac.uk/` under CEMPS and Harrison. Use the category containing the module code and 2022-23 Continuous Assessment 1. If you are uploading both files, create and upload a compressed version of your files as a single file using the `zip` compression format.