ECM3420 Coursework Assessment

Module convenor: Dr. Edward Chuah

This CA will account for 40% of your final grade for ECM3420.

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 techniques you learned throughout this module to train few machine learning models on a dataset that you feel passionate about, select the techniques that best suits your needs, and communicate insights you found from your modeling exercise.

After going through some guided steps, you will have insights that either explain or predict your outcome variable. As a main deliverable, you will submit a report that helps you focus on highlighting your analytical skills and thought process.

You are expected to leverage a wide variety of tools such as Jupyter notebook, Python and the relevant machine learning libraries (Keras, Tensorflow, Pytorch, etc.), but your report should focus on present findings, insights and next steps. Before you begin, you will need to choose a data set that you feel passionate about. This can be a data set similar to the data you have available at work or data you have always wanted to analyse. For some people this will be sports data sets, while some other folks prefer to focus on data from a datathon or data for good. Data for Good, inspired by DataKind.org, brings together leading data scientists with high impact social organizations through a comprehensive, collaborative approach that leads to shared insights, greater understanding, and positive action through "data in the service of humanity". Below are the links to 5 data sets:

1. Fortune 500. URL: https://data.world/aurielle/fortune-500-2017
2. AT&T stock price data. URL: https://www.kaggle.com/konstantinparfenov/att-sbc-stock-price-data/version/1
3. COVID-19 variants. URL: https://www.kaggle.com/gpreda/covid19-variants
4. Leukemia gene expression. URL: https://www.kaggle.com/brunogrisci/leukemia-gene-expression-cumida
5. Stock exchange data. URL: https://www.kaggle.com/mattiuzc/stock-exchange-data

Required

Once you have selected a data set, you will produce the deliverables listed below:

A. Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.
B. Brief description of the data set you chose and a summary of its attributes.
C. Brief summary of data exploration and actions taken for data cleaning and feature engineering.

D. Summary of training two machine learning models. For regression, the model will be multiple linear regression, polynomial regression, LASSO regression and ridge regression. For classification, the model will be multilayer perceptron, convolution neural network and variants of multilayer perceptron and convolution neural network such as ResMLP and GoogLeNet. For clustering, the model will be K-means, hierarchical clustering, DBSCAN and OPTICS.
E. A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy and explainability.
F. Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your models.
G. Suggestions for next steps in analysing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation, a better prediction, etc.

Compulsory: Please submit a PDF file containing the deliverables A to G. 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 5 main points:

1. Does the report include a paragraph detailing the main objective(s) of this analysis *[10 marks]*?
    a. This report is missing a planning section for the data analysis. *[0 marks]*
    b. Yes. This plan includes a detailed subtask section or a good vision of what is possible to do with this data set. *[5 marks]*
    c. This plan exceeds expectations. In addition to plan out subtasks and vision for this analysis, it also anticipates possible snags that might be incorporated into preliminary hypothesis of the data. *[10 marks]*

2. Does the report include a section describing the data *[10 marks]*?
    a. There is no summary or it is hard to put together what variables are available or how they might be used. *[0 marks]*
    b. There is a basic summary, like a data dictionary. *[5 marks]*
    c. The summary of the data is presented with graphs of distributions and plots that show the relation between features and the outcome variable. *[10 marks]*

3. Does the report include a section with variations of machine learning models and specifies which one is the model that best suits the main objective(s) of this analysis *[10 marks]*?
    a. No. It is not clear if a machine learning model was used in this analysis, or the machine learning model is missing. *[0 marks]*
    b. Yes. Two machine learning models are included and it discusses findings and results appropriately. *[5 marks]*
    c. Yes. There are two machine learning models. One of them is presented as the better alternative and some findings are presented. The findings should include what variations of a machine learning model should be considered (testing splits, cross validation, polynomial features, regularized regressions, cluster selection, etc.). *[10 marks]*

4. Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis [10 marks]?
    a. No. There are no takeaways, insights or findings about this problem. *[0 marks]*
    b. Yes. Some takeaways and findings derived from the model are presented. *[5 marks]*
    c. Yes. Takeaways and findings derived from the model are well presented. *[10 marks]*

5. Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques [10 marks]?
    a. No. There is no mention of possible flaws or plans to revisit the analysis. *[0 marks]*
    b. Yes. There is some discussion presented on possible flaws of this model and a plan to revisit this with additional data or different predictive modeling techniques. *[5 marks]*
    c. Yes. There is a comprehensive list of possible flaws of this model and a detailed plan to revisit this with additional data or different predictive modeling techniques. The quality of this section gives it full marks. *[10 marks]*

FAQs

Q1: Do I have to come up with my own data set?

Ans: You are highly encouraged to find a data set you feel really passionate about. This will help you showcase analytical work that truly matches your skills. But if you prefer, you can use some of the data sets from this module.

Q2: Is it OK to choose the same data set as someone else?

Ans: Yes, more than one person can analyse the same data set. Most likely your insights will be different from your peers and you will still be able to showcase your own talent as a unique solution.

Q3: Do I have to train more than two machine learning models?

Ans: You are required to train two machine learning models to highlight which model improved your prediction or interpretation.

Q4: Is this an individual assignment?

Ans: This is an individual assignment. You can ask for help or assistance on technical issues and general direction of your analysis, but the interpretation of the analytical output and the writing of the report should be your own.