RMIT Vietnam University

School of Science and Technology

COSC2789 - Practical Data Science Assignment 3: Group Project

Due: 23:59, Thursday (21st, January 2021) (Week 12)

This assignment is worth 30% of your overall mark.

0. Assignment Teams

This assignment should be carried out in groups of *maximum 4*. It is up to you to form a team. Once you have formed your team, you should register your team on Canvas.

Important: you must register your team within 1 week at the latest. Anyone without a team by 15th December 2020 will be randomly assigned to a team.

If you have strong reasons for needing to complete the assignment with less than 4 members, you may apply to do so by sending an email to the lecturer, explaining your reasons. However, bear in mind that the requirements and available marks will be the same as for a team of 4.

In addition, please submit what percentage each member contributed to the assignment and include this in your report. The contributions of your group should add up to 100%. The ones with too little contribution (e.g. less than 15% contribution) will have their marks reduced.

1. Introduction

This assignment is intended to give you practical experience with all steps of the data science process, from data retrieving, data wrangling to modelling and model deployment.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis - it is your responsibility to stay informed with regards to any announcements or changes.

2. Where to Develop Your Code

You are encouraged to develop and test your code in two environments: Jupyter Notebook (or Jupyter Lab) on Lab PCs or your laptop. You are encouraged to use GitHub for project management and code sharing.

3. Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. More information on Academic Integrity is available at https://www.rmit.edu.vn/students/my-studies/assessment-and-exams/academic-integrity

4. General Requirements

This section contains information about the general requirements that your assignment must meet. Please read all requirements carefully before you start.
_ You must do the analysis and modelling in Python Jupyter Notebook/Jupyter Lab.
_ You must include all the python and other files for your model deployment, including the trained models.
_ Parts of this assignment will include a written report, this *must* be in ***PDF* format**.
_ You must include all the python and other files required for the visualization dashboard.
_ You must include the "requirements.txt" with all the required python libraries for this project.
_ You must include a plain text file called "readme.txt" with your submission. This "readme.txt" file should include your name(s) and student ID(s), and instructions for how to execute your submitted script files. This is important as *model deployment and automation* is part of the 6th step of data science process, and will be assessed strictly.

Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is gryphon, then that is exactly the file name you should submit; Gryphon, GRYPHON, griffin, and anything else but gryphon will be rejected.

Task 0: Choosing your project topic

Identify the data science problem that you want your project to solve. The data science problem has to be solvable using Classification, Time Series or Clustering approaches. Please choose carefully as you have to list measurable project goals, tangible deliverables and work on the project will full data pipeline and model deployment to solve that problem.

Task 1.1: Data Retrieving and Cleaning (2%)

For this assignment, you need to select at least **one** suitable dataset for your data science problem, in a domain that is of interest to you. This dataset MUST have at least 10,000 data instances. You will need to: include a detailed description of the data in your report, and describe each attribute of it, including the type, the range of possible values, whether it contains any missing values/errors, etc. You MUST include a copy of the dataset, to allow the assessment of your modelling result.

NOTE: Please do not use any dataset that we used in the teaching (including lectures, tutorials and other assignments) of this course.

Being a data scientist, you know that it is vital to set **the goal of the project**, then **thoroughly pre-process** any available data (each attribute) before starting to analyse and model it. In your report, you need to clearly state the goal of your project, and the design/steps of pre-processing your data. Please ensure you understand the data you selected, including the meaning of each attribute.

Task 1.2: Data Exploration (3%)

Explore the selected data, carrying out the following tasks:

- Explore at least 3 columns using appropriate descriptive statistics and graphs (if appropriate), e.g. the distribution of a numerical attribute, the proportion of each value of a categorical attribute. For each explored column, please think carefully and report in your report.
  1. the way you used to explore a column (e.g. statistics or the graph);
  2. what you can observe from the way you used to explore it and write your detailed analysis.

- Explore the relationship between at least 3 pairs of attributes, and show their relationship in an appropriate graph. You may choose which pairs of columns to focus on, but you need to generate a visualisation graph for each pair of attributes. Each of the attribute pair should address a **plausible hypothesis** for the data concerned. In your report, for each plot (pair of attributes), state the hypothesis that you are investigating. Then, discuss in details any interesting relationships (or lack of relationships) that you can observe from your visualisation.

(IMPORTANT: Please format each graph carefully, and use it in your final report. You need to include appropriate labels on the *x*-axis and *y*-axis, a title, and a legend. The fonts should be sized for good readability. Components of the graphs should be coloured appropriately.)

Task 2: Data Modelling (5%)

Model the data by treating it as **either** a *Classification, Time Series, or Clustering* Task, depending on which dataset you previously selected.

You must build 3 models within the particular Task category (i.e. 3 Classification models, 3 Time Series Models, or 3 Clustering models), and include the following steps for *each* model:

- Train/val/test split or k-fold cross validation (if appropriate)
- Feature Engineering
- Parameter Tuning
- Model training
- Model Evaluation and Selection
- Save the model for deployment

You need to include the analysis/explanation of each steps and the results, including a detailed discussion or all your models and the final model selection in your report.

Task 3: Model deployment and Automation (5%)

After you trained and saved your models, you now have to deploy the model to be used for inference. You can choose Dash (running on Flask), Flask, Django build the back-end for model deployment with API.

The models can be deployed locally on your local machine with 2 API functions:

- Evaluate: The API function will take the test set and output the evaluation metrics.
- Predict: The API function will take the test set and output the predicted values.

Task 4: Visualisation Dashboard (5%)

In this task, you need to build a visualization dashboard using Dash (or other preferred framework) and deploy it online together with the services where you deploy your model. The visualization dashboard must include at least 3 interactive plots.

- 2 plots must be about column pairs from Task 1 (similar to Assignment 1).
- 1 plot must be on model/prediction results from Task 2

NOTE: The design, including the layout, text description, colours, etc., will be taken into account this time. Therefore, you should try to build not only a functional dashboard, but a nice-looking one as well. No static image for this task.

Task 5: Report (5%)

Write your report and save it in a file called "report.pdf", and it must be in PDF format, and must be at most 30 pages (single column, including figures, not including the Cover page, Table of Content and References) with a font size between 10 and 12 points. Penalties will apply if the report does not satisfy the requirement. Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs.

Your report must have the following structure:

- Cover page, including: Title, Authors, Affiliations, Contact details, and Date
- Table of Content
- An abstract/executive summary
- Introduction
- Project Management, including: Project Goals, Deliverables, Program Stack, Timeline, and Member Contributions
- Methodology
- Results
- Discussion
- Conclusion
- References

Task 6: Presentation (5%)

You will be required to do a presentation in the last session of the course. The presentation should include, but not limit to:

- briefly describe your chosen problem and dataset(s).
- describe the data preparation steps.
- state the hypotheses/questions that you were investigating.
- explain what the modelling steps are, and what the results are.
- demo of the model deployment and visualization dashboard.
- show the final conclusion and recommendation.

The presentation should be at a maximum of 20 minutes per group, including 3-5 minutes for Q&A. Each group member has to present at least 2 slides in the presentation. Your presentation slides must be converted to PDF format and included in the submission before the presentation date.

What to Submit, When, and How

The assignment is due at 23:59, Thursday (21st, January 2021) (Week 12). Assignments submitted after this time will be subject to standard late submission penalties. There are four sets of file you need to submit:

- Notebook file containing your python commands, 'Assignment3.ipynb'.
    - For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:
        - Main menu → Kernel → Restart & Run All
        - Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.
- Your code files for Task 3 and Task 4. The "requirements.txt" and "readme.txt" with instructions for how to execute your submitted script files must be included as well.

NOTE: They must be submitted as ONE single zip file, named as your student numbers (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted on Canvas.

- Your report.pdf file in PDF with the correct format and length. Plagiarism will be checked for this and penalty will be applied.
- Your presentation slides in PDF format. Plagiarism will be checked for this and penalty will be applied.

NOTE: The two PDFs must be submitted in a separate section on Canvas for plagiarism check.

Please do NOT submit other unnecessary files.