RMIT Vietnam University
School of Science, Engineering and Technology

COSC2789 - Practical Data Science

Assignment 3: Group project

Due: 22:59, Thursday (the 19th, January 2024, Week 11)
This assignment is worth 30% of your overall mark.

Assignment Teams

This assignment is teamwork, each team with at most 3 students. 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 on Canvas. Anyone without a team **by 3rd January 2024** will be randomly assigned to a team. If you have strong reasons for needing to complete the assignment with less than 3 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 3. 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.

Introduction

This assignment covers core steps in the data science process. You will need to develop and implement appropriate steps, in Ipython (Jupyter Notebook), to complete the corresponding tasks. This assignment is intended to give you practical experience with the typical steps of the data science process.

The "Practical Data Science with Python" 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.

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.

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

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 in Python Jupyter Notebook/Jupyter Lab.
- 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 you project topic (1%)

This assignment covers the core steps of the data science process. You need to identify the data science problem that you want your project to solve. The data science problem must be solvable using Classification, Regression or Clustering approaches. Please choose carefully as you must list measurable project goals, tangible deliverables and work on the project will full data pipeline and model deployment to solve that problem.

Examples of **two types of problems** you may select to work on are as follows.

1. **Problem type 1: Focusing on Data Modelling**.

For this problem, you will model the data by treating it as a classification, a regression and/or clustering task, depending on your choice. You need to select **at least two tasks, one of which must be a clustering task**. For example, your choice can be c*lassification and clustering.*

You need to select one dataset from the following options, and then work on it:

1.1. Incident management process enriched event log Data Set. More details can be found from the following UCI webpage about this dataset:
https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log

1.2. Online Shoppers Purchasing Intention Dataset Data Set. More details can be found from the following UCI webpage about this dataset:
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

1.3. Online Retail II Dataset. More details can be found from the following UCI webpage about this dataset: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

2. **Problem type 2: Building a recommender system.**

For this problem, you will work on this dataset: Anonymous Microsoft Web Data Dataset. More details can be found from the following UCI webpage
https://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

You need to implement at least two approaches for building a recommender system, such as content-based recommendation and collaborative filtering-based recommendation. For each of the approaches, you need to use some of the data modeling techniques such as classification, regression and/or clustering.

Task 1: Retrieving and Preparing the Data (5%)

Being a careful 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 in Task 4, 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.

Task 2: Feature Engineering (3%)

Use suitable Python functions to extract potential features for model input. Conduct appropriate analysis to evaluate feature importance (e.g. correlation analysis), then use suitable method(s) to select the final features for the model. The feature choices must be explained via analysis.

Note: These steps must be performed consistently for train/val/test sets.

Task 3: Data Modelling (12%)

Model the data by treating it as either a *clustering, classification and/*or *regression* task, depending on your choice.

You must use at least **two different models** for each approach (i.e. two classification models and two clustering models), and when building each model, it must include the following steps:

- Select appropriate features.
- Select the appropriate model (e.g. *DecisionTree* for classification) from *sklearn*.
- Train and evaluate the model appropriately.
- Train and evaluate the model by selecting the appropriate values for each parameter in the model. You need to show how you choose these values and justify why you choose them.

After you have built two clustering models and two classification (or regression) models, on your data, the next step is to **compare** the performance of the models. You need to include the results of this comparison, including a recommendation of which model should be used, in your report (see Task 4).

For **Problem type 2**: The two tasks for data modeling (to be used in the two recommender systems) can be any of the three data modeling approaches (classification, clustering, regression).

Other Evaluation Criteria: Innovative Model (bonus 2%)

Out of the 4 selected models, there should be at least 1 innovative model (the other 3 models can be simple models). A simple model using only one algorithm for model training with some parameter tuning is not considered as an innovative model. For example, using a K-NN classifier from *scikit-learn* without any modification will be considered a simple model and won't have any point.

If you use a model from any research work, you must cite the reference correctly. An example of an innovative model is as below:

1 point: a linear stacking of multiple algorithms or an ensemble model.

2 points: a complex ensemble model or a complex combination of multiple algorithms. You can propose a new model (algorithm) here.

Give a short explanation about the classification results obtained from the innovative model.

Task 4: Report (4%)

Write your report and save it in a file called report.pdf, and it must be in PDF format, and must be **at most 16 (in single column format) pages for everything (including figures and references) with a font size between 10 and 12 points.** Penalties will apply if the report does not satisfy the requirements. Remember to clearly cite any sources (including books, research papers, course notes, source code, etc.) that you referred to while designing aspects of your programs.

Your report must have the following structure:

- A cover page, including:
  - Statement of the solution representing your own work as required.
  - Title
  - Author information
  - Affiliations
  - Contact details

  - Date of report
- Table of Content
- An abstract/executive summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Reference

Task 5: Presentation (5%)

You will be required to make 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.

– show the conclusion and recommendation.

You need to prepare 10-12 slides for the in-class presentation and demonstration.

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

5.1. Slide and presentation (2 points)

- The slides must follow RMIT University template.
- The slides and presentation must clearly present the research question(s), the used methods for solving the problem(s), the results, and recommendations.
- The presentation is scheduled on Tuesday, January 23rd, 2024 (week 12) during our regular class time).

5.2. Demo (1 point): The code runs without error, showing the exact results as presented in the report.

5.3. Q&A (2 points): Answer the questions by the lecturer and other students clearly and convincingly.

What to Submit, When, and How

The assignment is due at **20:59, Thursday the 19th, January 2024** (in Week 11). Assignments submitted after this time will be subject to standard late submission penalties. You need to submit the following files:

- 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).
- Your **report.pdf** file at most 16 (in single column format) pages (including figures and references) with a font size between 10 and 12 points.
- A "readme.txt" file (if needed) includes your name and student ID, and instructions for how to execute your submitted script files.

They must be submitted as ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas: Assignments/Assignment 2. Please do NOT submit other unnecessary files.