COSC 2670/2738 Practical Data Science with Python

Project Assignment 2, Semester 2, 2022

Marks: This assignment is worth 45% of the overall assessment for this course.

Due Date : Friday, 14 October 2022, 11:59PM (End Week 10), via `canvas`. Late penalties apply. A penalty of 10% of the total project score will be deducted per day. No submissions will be accepted five days beyond the due date.

Objective

The key objectives of this assignment are to learn how to model data using python and scikit-learn. The assignment is broken into 4 separate problems, which you should solve in order as you may need functions you create from previous exercises. The problems are given to you as jupyter notebooks and include helper functions and clear instructions on what you need to do to solve each problem.

The data you will be working with is a dump from StackExchange, specifically the Stats stack exchange (a.k.a, Cross Validated). The data has been sampled and augmented for education purposes. You should be familiar with the data from A1. For this assignment, we will use only the questions from the Posts.feather file. The data is separated into two files, train and test sets, both are samples from the same dataset, and the only difference is that the train set has an additional column "Label".

The "Label" represents a popular topic that the question belongs to, and your goal is to explore and model the data in the train set to generate predictions successfully for the samples in the test set.

Provided files

The following files are provided:

- **SXXXX-A2-[DataExploration, Classification, Clustering, TestClassification].ipynb** : You should use the skeleton Jupyter notebook files to solve the problems and write your answers.
- **TrainQuestionsDF.feather.zstd** : The train (labeled) questions dataframe.
- **TestQuestionsDF.feather.zstd** : The test (missing labels) questions dataframe.
- **requirements.txt** The python packages you are expected to use for this assignment.
- **A2.pdf** : This file.

See the provided notebook files for further instructions.

Creating Your Workspace

You will notice that the four notebooks we provide have the prefix "SXXXXX-". Your very first task is to create a folder called SXXXXX, but replace the Xs with your actual student number. So, for example, if my student ID was S12345, I would name the

folder S12345, and make a copy of each notebook into this folder with the name S12345-A2-DataExploration.ipynb, and so on. If I did this correctly, I should see something like:

```
S12345/S12345-A2-DataExploration.ipynb
S12345/S12345-A2-Classification.ipynb
S12345/S12345-A2-Clustering.ipynb
S12345/S12345-A2-TestClassification.ipynb
S12345/TrainQuestionsDFfeather.zstd
S12345/TestQuestionsDFfeather.zstd
S12345/requirements.txt
```

Note that you should move the feather files into this working directory as well as you will need them in the same directory from where you run Jupyter notebook. The feather files will not be submitted as they are not tiny (they are real data!). More on what you need to submit later. Just remember, whenever I say SXXXXX henceforth, I mean your real student ID.

Important Note

In this assignment, you are expected to apply algorithms and functions that rely on pseudo-random processes, which will affect and vary the results. To ensure that the results are reproducible, you are expected to set the variable `RANDOM_SEED` to your student id at the top of each notebook. After properly setting the constant variable, make sure to pass it to the `random_state` argument in any class/function that has such an argument (can be verified through official documentation). This includes but not limited to `train_test_split`, `DecisionTreeClassifier` and `KMeans`. The random seed will affect your output and hence the validity of your answers. Make sure you set it in all notebooks at the beginning of your work.

Creating a New Anaconda Environment

Assuming you have already properly installed and setup Anaconda, you should create a new environment for A2 with jupyter and Python 3.8.x. Creating a new environment will ensure that you use the same package versions as your marker to avoid any unexpected discrepancies between your results and the marker's, which might again affect the validity of your answers. To create a new environment, you should run the following commands in the SXXXXX folder:

```
conda create --name [NAME] python=3.8
activate [NAME]
pip install -r requirements.txt
pip install notebook
```

Replace `[NAME]` with a name of your choice. The last command will install the packages in the specified versions you will need to use in this assignment. You should not install any additional packages without prior approval from the teaching team, as these may not be present during the assessment.

1 Marking Template

For clarity, I will briefly outline the marking policy for the assignment. There are generally four specific components you should consider:

- DataExploration - 8 points
- Classification - 14 points
- Clustering - 13 points
- TestClassification - 10 points

Each consists of coding challenges and related questions. You should follow the detailed instructions in the Jupyter notebooks for each of the four challenges.

Submission

All submissions must be made through MyRMIT. A link will be provided within canvas, under the Assignments tab for submissions. Assignments submitted through any other method will not be marked. To submit your files, create a top-level directory with your student number, as described above. This is a folder named SXXXXX, where SXXXXX is your student id. Inside that folder, the four Jupyter notebook files should be named as described above. A PDF version of each notebook (including the outputs) would be used to support your answers if the output during the check differs from the one you see. And up to four prediction files. You **should not submit the two feather data files**.

An example structure within the zip file is shown below for the files you should be submitting. If your student ID happened to be S12345, you would submit a file called `S12345.zip`, which contains the following, including the folder!

```
Folder: S12345/
Files within the folder:
  1. S12345/S12345-A2-DataExploration.ipynb
  2. S12345/S12345-A2-DataExploration.pdf
  3. S12345/S12345-A2-Classification.ipynb
  4. S12345/S12345-A2-Classification.pdf
  5. S12345/S12345-A2-Clustering.ipynb
  6. S12345/S12345-A2-Clustering.pdf
  7. S12345/S12345-A2-TestClassification.ipynb
  8. S12345/S12345-A2-TestClassification.pdf
  9-12. S12345/S12345-A2-predictions-[1,2,3,4].csv
```

A total of 9-12 files.

NOTE: There is a discussion group available in the course canvas. Please do not post code snippets in this forum. Do ask questions if you have them! We will do our best to answer them as quickly as we can. This is the best medium to ask a question, as there is a strong chance that if you are confused about something, your classmates are too.

Plagiarism Warning

University Policy on Academic Honesty and Plagiarism: It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student concerned. Please see the RMIT policy for more details: https://www.rmit.edu.au/students/my-course/assessment-results/academic-integrity.

THIS IS NOT A GROUP PROJECT. Students are reminded that this assignment is to be attempted individually. Plagiarism of any form will result in zero marks being given for this assessment, and can result in disciplinary action. We routinely use plagiarism software on projects! Please, please don't do it. Be aware that paying someone on a coding site to do it for you is a form of plagiarism. If you are submitting work that is not your own, regardless of how you got it – you are in breach of this policy.

Extension Policy

Individual extensions will are unlikely to be considered or granted by the PDS team. If you have an Equitable Learning Plan in place, please discuss issues you may have with me at least two days prior to the deadline. An ELP does not mean you are eligible for an extension, it just means that if you have a good reason to need one, we will work it out. We are not monsters, but doing things on time matters, and not just for this course but for all of your courses. Getting behind creates more stress. We consider timeliness as a core component of the assessment. If we decide to extend the deadline, the fairest way we can do this is to extend it for *everyone*, and that may not be possible given the the size of the course and university policies. We have set deadlines as late as we can in order to meet the university timelines we cannot control. So try to be early and not late. If you procrastinate and suddenly have other priories, you will be in trouble, and that is not a valid reason for an extension.

If you have suffered a personal tragedy or illness, there is a University process in place to grant extensions. Our preference is that you go this route if you must, as they have very clear criteria to grant exemptions. For more information about applying for Special Consideration, see the rules and regulations at https://www.rmit.edu.au/students/my-course/assessment-results/special-consideration-extensions/special-consideration.

Getting Help

Come talk to us during office hours. Email us. Use the discussion board. Ask a question in a lectorial or a practical. There is help available if you need it. We do strongly urge that you start with the discussion board because, as mentioned above, if you are confused, there is a very good chance someone else has exactly the same question, and it is not a great use of the team's time to keep answering the same question over and over.