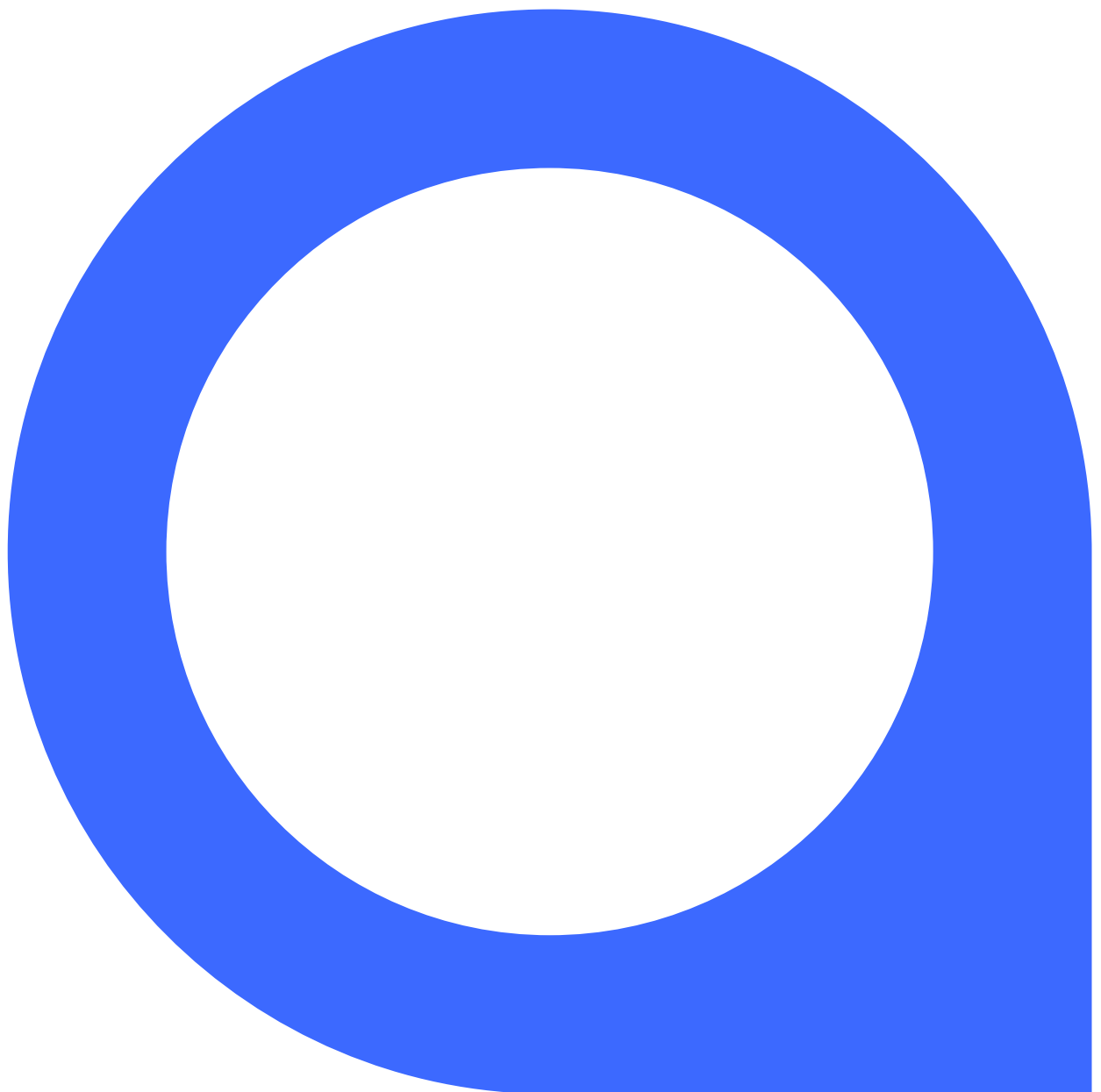


Data Science Applications

Assignment Semester 2 2024





Preamble

The main purpose of the assignment from your perspective is to help you to:

- consider the business environment in which a problem is to be solved;
- apply data science techniques to solve a business problem; and
- communicate the outcomes of your analysis to business stakeholders.

These skills will also help you pass the end of semester assessment and perform well in the workplace.

The specific skills that are being developed and assessed in the assignment are the ability to:¹

- evaluate how well data describes business activity;
- develop solutions to a range of classification problems using GLMs, tree-based models, ensembling and neural networks;
- evaluate solutions produced by classification models;
- explain how clustering techniques can be used to gain business insight;
- perform k-means and hierarchical clustering;
- evaluate a clustering algorithm using internal, external, and manual validation;
- apply each step in the natural language processing pipeline to solve a variety of business problems;
- evaluate the outcomes of natural language processing models;
- implement strategies for gaining stakeholder support for data science projects;
- communicate relevant points in language appropriate to the audience, in a logical and coherent manner; and
- meet business standards for presentation of work, both modelling and written materials.

This assignment provides an opportunity for you to think deeply, spend time preparing a detailed answer and self-reflect on your writing skills. Whilst there is ample time to write your assignment answers, you should ask yourself if you need to spend more time improving your writing skills to help you pass time-limited examinations.

¹ The skills listed here are learning objectives from the subject's syllabus, apart from the last two skills on the list which are assessable in every subject. This assignment does not cover every component of the learning objectives listed above.



The assignment requires you to build models and create a set of sensible assumptions or parameters for those models. Consequently, there is no single right answer meaning you are assessed on your reasoning and process. You therefore need to demonstrate *how* you derived your assumptions or model parameters and your answers. It is important that you describe what you did as the marker(s) will want to understand if you are able to apply knowledge to the specific situation described in this assignment. We are also looking for you to demonstrate that you can deal with uncertainty in a reasonable way.

A key actuarial skill is to obtain a grasp of the qualitative nature of outputs from models and describe them. This assignment is designed to test your ability to explain your model(s) and their outputs to a non-technical audience.

Marking Guide

This assignment represents 50% of the available marks for the Data Science Applications subject². Your assignment mark will be combined with your exam mark to determine your overall result for the subject.

It is anticipated that Fellowship students will spend at least 50 hours to complete the assignment. In past semesters, some students have spent significantly more time than this, particularly those students who aim for a grade of Above Pass Level or Significantly Above Pass Level.

A detailed rubric is provided with the assignment question and will be used by the markers to assess your performance. The rubric has been posted on the Assignments page of Canvas to guide you as to what is required to achieve full marks for each part of the assignment. You should check that the components of your answer cover the items in the rubric.

You should also use clear structure in your written, coded, and video answers to make it easy for markers to find where you have responded to each of the rubric criteria.

² For students completing the subject as a microcredential Certificate path, the assignment represents 100% of the available marks for the microcredential.



Submission

Deadline

The deadline for submission is **12:00 midday AEST on 4th October 2024**.

Submit your assignment via the Assignments page in Canvas. If you experience technological issues when submitting your assignment, please send a copy of your assignment by email to education@actuaries.asn.au.

Penalties apply for late submissions (see section on 'Penalties'). You should anticipate potential delays by preparing and submitting your work in advance of the deadline.

Should circumstances arise that mean you cannot submit your assignment on time, you should contact education@actuaries.asn.au in advance of the deadline and apply for special consideration.

File format

The submitted documents must consist of one pdf file and one Jupyter notebook. Files in other formats will not be marked. The naming convention for files is:

DSA 2024 S2 Assignment member ID.(file extension as appropriate)

Please note that if you resubmit an assessment, Canvas automatically adds a suffix to the file name (such as '-1' for the first resubmission). You do not have to make any adjustment for this.

Coversheet

A coversheet for the assignment is provided on the Assignments page in Canvas. Complete and attach this coversheet as the front page of your pdf file.



Video summary

As part of this assignment, you are required to record a five-minute video summary of your findings. Advice about how to record an effective video summary is provided in an Appendix. You should submit your video by following these steps:

- create a video recording using the naming convention 'DSA 2024 S2 Assignment member ID';
- use your video recording to create an 'unlisted' YouTube video (see instructions in the Appendix)³; and
- insert your YouTube video URL as a hyperlink in your assignment pdf file.

Jupyter notebook

The Jupyter notebook should use the assignment notebook template provided. The notebook must be capable of running successfully in Google Colab as markers will use this platform to view and access the notebooks. Within the notebook you should:

- explain each step taken in your analysis in a text cell above your code; and
- evaluate and comment on the output from each step in a text cell below the output.

Please note that, unless specified, there is no word limit for the comments in your notebook. However, markers will look more favourably on students who provide clear and succinct commentary, compared to those who provide no commentary or those who provide too much commentary, including those who repeat large sections of the subject materials in their comments. This latter approach makes it very difficult for a marker to assess your understanding of the step being taken or the output being produced.

Word or time limit

Some questions in the assignment have a specific word or time limit. Markers will not read or watch any part of your answer that exceeds this limit. Keep your word count or presentation timing within any limits that are specified. The word count includes any text within tables, text boxes or images consisting primarily of text. The word count does not include:

- contents table or index; and
- references to sources used.

³ The Appendix also provides advice for students who do not have access to YouTube due to their location.



Keep in mind one of the key principles taught in the Communication, Modelling and Professionalism subject: always write as clearly and succinctly as possible, while still including enough information that will be useful for your audience. With that in mind, consider whether each word, sentence, or paragraph you include in your assignment adds to or detracts from the message you are trying to convey. Importantly, know that 'more' is usually not 'best'.

Plagiarism

By submitting your assignment, you are implicitly stating that the work is your own.

Remember that an important aspect of being a professional actuary is to always act with integrity. Committing plagiarism by copying another person's work or not properly referencing other sources used in your assignment is a breach of the Integrity principle under the Actuaries Institute's Code of Conduct.

Any suspected plagiarism will be referred to the Institute's Executive General Manager, Education for review. Depending on findings, a complaint regarding the member may be made to the Institute's Conduct Committee. Subject marks may not be released until the matter is resolved.

Penalties

Late submissions

Penalties will be applied to late submissions without prior approval.

If you submit an assessment after the due date (whether that is the original due date or any extended due date you have been granted), the following penalties apply:

- within one day (24 hours) of due date and time: 20% x maximum mark available;
- more than one day late: 100% x maximum mark available (i.e. assessment score = 0).

Please note that 'days' above refers to calendar days, not working days.

Incorrectly formatted submissions

There is no direct penalty if an assessment is submitted in a format with an incorrect file name or an incorrect format (e.g. submitted as a word document when a pdf document was required).



If a submission does not include a relevant identifier (member ID) in the file name, or an incorrect identifier is used, then it may take time to identify you as the student and you may be asked to resubmit your work with an appropriate identifier.

If you fail to submit in the file format that was required, then you may be required to resubmit your work with the correct file format, particularly relevant to modelling or coding assignments.

If either situation arises then this will probably cause you to submit late and hence incur the late submission penalties outlined above. Students should therefore follow all assessment instructions provided.

Feedback

Our approach to feedback is for students to receive general feedback and a sample assessment marked as 'Significantly above pass level'.

You should review the general feedback that is provided to all students as well as the sample assessment. After reviewing the general feedback, you should use the rubric to grade the sample assessment and your submission. This will help you to compare the assessments and identify areas where your submission could have been improved.

Our belief is that this active approach to studying will provide you with a deeper understanding of where you need to improve. This is the best way for you to learn about your areas of strength and weakness. We do not provide students with individual feedback on their assessments.

At the end of the semester, you will receive:

- a letter to indicate whether you have passed or failed the subject;
- if you have failed the subject, a breakdown of your grade for each assessment;
- general feedback to all students about assessment performance; and
- sample assessment(s) that were graded as 'Significantly above pass level'.



Assignment Context

You are a data science actuary who has been engaged by the Membership team (who you can assume are not actuaries or data scientists) at the Actuaries Institute ('the Institute') to help improve their members' experience in using the Institute's CPD Knowledge Hub ('the CPD Hub'). The CPD Hub was built in 2018 as a way to collate thousands of actuarial resources ('assets') which are shared with members as CPD resources.

The Membership team is aware that membership engagement with the CPD Hub has declined in recent years. The Institute would like to entice more members to engage with the CPD Hub by:

- choosing a simpler, more effective way to group the CPD Hub's assets;
- placing assets that are more likely to be popular on the CPD Hub's front page; and
- using Generative AI (GenAI) to produce engaging summaries of each asset.

To help you complete this task, the Institute has provided you with a file ('DSA 2024 S2 assignment data.xlsx' or 'the assignment dataset') containing meta data for all the assets contained on the CPD Hub as at 9 April 2024. The data dictionary for this dataset is set out in Table 1.

Table 1: Data dictionary for the assignment dataset

Column name	Data type	Values	Description
Columns in 'Docs' tab of spreadsheet			
DMSId	integer	various	Unique identifier for each asset in the document management system (DMS)
FileName	string	various	File name of the asset
FileType	string	13 categories such as html, pdf, mp3 and ppt	File type of the asset
FileSize	integer	various	File size of the asset (in bytes)
Title	string	various	Title of the asset
StartPublishing	datetime	Dec 2001 to Mar 2024	Date the asset was published to The CPD Hub



Column name	Data type	Values	Description
StopPublishing	datetime	Nov 2023 to Nov 3023	Date the asset will no longer be published to The CPD Hub (if 0, no end date for publishing specified)
Cpd	integer	0 to 10	Number of CPD points available for reading, listening to or watching the asset.
Format	string	11 categories such as Article, Video, Workshops	Format of the asset
Level	string	3 categories (Introductory, Advanced, All)	Knowledge level required for the asset
Description	string	various	Description of the knowledge contained in the asset
Region	string	3 categories (International, Domestic and International, Domestic)	Region that the knowledge contained in the asset is relevant to
Copyright	string	4 categories (Institute, Other, Public Domain, Speaker/Author)	Who holds the copyright to the asset
EventAgeGroup	string	3 categories (All, Retired Actuaries, YAP)	The intended audience group, based on age, for the event from which the asset originated (YAP stands for Young Actuaries Program)
EventType	string	6 categories including Insights/conferences, Major, Other	The type of event from which the asset originated
Link	string	various	html weblink to the asset
Tags	string	various	tags applied to the asset to indicate the relevant practice area (or sub-topic within a practice area) to which the asset's contents relate
Columns in 'Downloads' tab of spreadsheet			
DMSId	integer	various	Unique identifier for each asset in the document management system (DMS)
DownloadDate	datetime	Aug 2018 to Apr 2024	Date the document was downloaded by a user from the DMS



Assignment Questions (Total 100 marks)

Answer Questions 1, 2, and 3 in your Jupyter notebook using the assignment template provided.

Answer Questions 4 and 5 in your pdf document.

Different questions in this assignment may be reviewed by different markers, so your answer to each question should be self-contained. No marks will be awarded for answers to a question that are only contained in your answers to other questions.

1. Explore and prepare the data

Your work with the Institute will start by providing them with a summary of the assets currently contained in the CPD Hub.

Answer Question 1 in your Jupyter notebook.

- a. **Examine** and then **clean** the assignment dataset to gain an understanding of the assets currently contained in the CPD Hub. Do not examine or clean the 'Description' feature, as that will be done in the next section of this question. *Note you are required to split the data into training, validation, and test sets within this question. When splitting the data, you should do so in preparation for use in Question 3. You should apply your judgement in deciding when to perform that split.*

(5 marks)

- b. **Clean** the 'Description' feature then **calculate** vectorised features that represent that feature, for use in either Question 2 (clustering) and/or Question 3 (classification). *You should perform this vectorisation on the training, validation, and test datasets, making sure you take steps to avoid leakage when applying the vectorisation method to the validation and test datasets.*

(5 marks)

- c. Summarise, in 500 words or less, key characteristics of the dataset. *Your answer should be communicated using language suitable for sharing with the Membership team at the Institute*

(5 marks)



2. Choose a simpler, more effective way to group the CPD Hub's assets

Each asset in the CPD Hub is currently tagged with multiple labels to indicate the practice area and sub-area that the asset is relevant to. This tagging system has become quite complicated over the years and may be making it harder for members to find material that is of interest to them. Your next task is therefore to help the Institute create a simpler way of grouping assets in the CPD Hub.

Answer Question 2 in your Jupyter notebook.

- a. **Suggest**, in 500 words or less, which features should be included in a clustering algorithm to find a simpler, more effective way to group the CPD Hub's assets.

(5 marks)

- b. **Apply** a clustering algorithm using the features suggested in Question 2a.

(5 marks)

- c. **Examine** the clustering outputs using manual validation.

(10 marks)

- d. **Suggest**, in 500 words or less, a simpler, more effective way to group the CPD Hub's assets, based on your examination of the clustering outputs. *Your answer should be communicated using language suitable for sharing with the Membership team of the Institute.*

(5 marks)



3. Predict the future popularity of assets

As outlined in the Question Context, the Institute would like to make more strategic decisions about where to promote assets in the CPD Hub, based on how popular each asset is expected to be. These decisions will apply to both existing assets and new assets that are added to the CPD Hub over time. Your next task is to build a neural network classification model to predict whether each asset will have 'low', 'medium', or 'high' popularity amongst the Institute's members, using the total number of downloads of an asset as a proxy for 'popularity'.

Answer Question 3 in your Jupyter notebook.

- a. **Construct** a response variable for your classification model. *You should allow for different assets having different amounts of time available for downloading.*

(10 marks)

- b. **Suggest** four metrics you will use to evaluate the success of your classifier.

(5 marks)

- c. **Construct** your neural network classifier.

(20 marks)

- d. **Interpret** the performance of your chosen neural network classifier, using the metrics suggested in Question 3b. *Your answer should be communicated using language suitable for sharing with the Membership team at the Institute.*

(5 marks)

4. Use Generative AI (GenAI) to produce engaging summaries of the CPD Hub's assets

Your next task is to investigate the use of GenAI tools to create engaging summaries of assets in the CPD Hub. To answer this question, you should:

- select one asset in the CPD Hub that is of interest to you;
- read/watch/listen to the asset;



- use a GenAI tool to create a summary of the asset using the prompt 'summarise this article' with an attached copy of the article; and
- use a GenAI tool to create a more engaging summary of the asset using a more tailored prompt.

Note that you may use any GenAI tool to create the summaries of the asset.

Answer Question 4 in your pdf file.

- a. **Critique**, in 500 words or less, the initial summary of the asset produced by GenAI. *You should include the summary created by GenAI in your answer – this summary will not count towards the word limit. Your answer should be communicated using language suitable for sharing with the Institute.*

(5 marks)

- b. **Explain**, in 500 words or less, the process you followed to produce a more engaging summary of your chosen asset, which addresses the issues uncovered in your answer to Question 4a. *You should include your revised prompt and the revised summary created by GenAI in your answer – this revised prompt and summary will not count towards the word limit.*

(5 marks)

5. Video summary of findings

Answer Question 5 in your pdf file.

Prepare a five-minute video, for presentation to the Membership team at the Institute, to summarise your findings from Questions 2, 3, and 4. *You should structure your video to have a clear start, middle, and end, with clear transitions between all sections.*

(10 marks)

END OF ASSIGNMENT