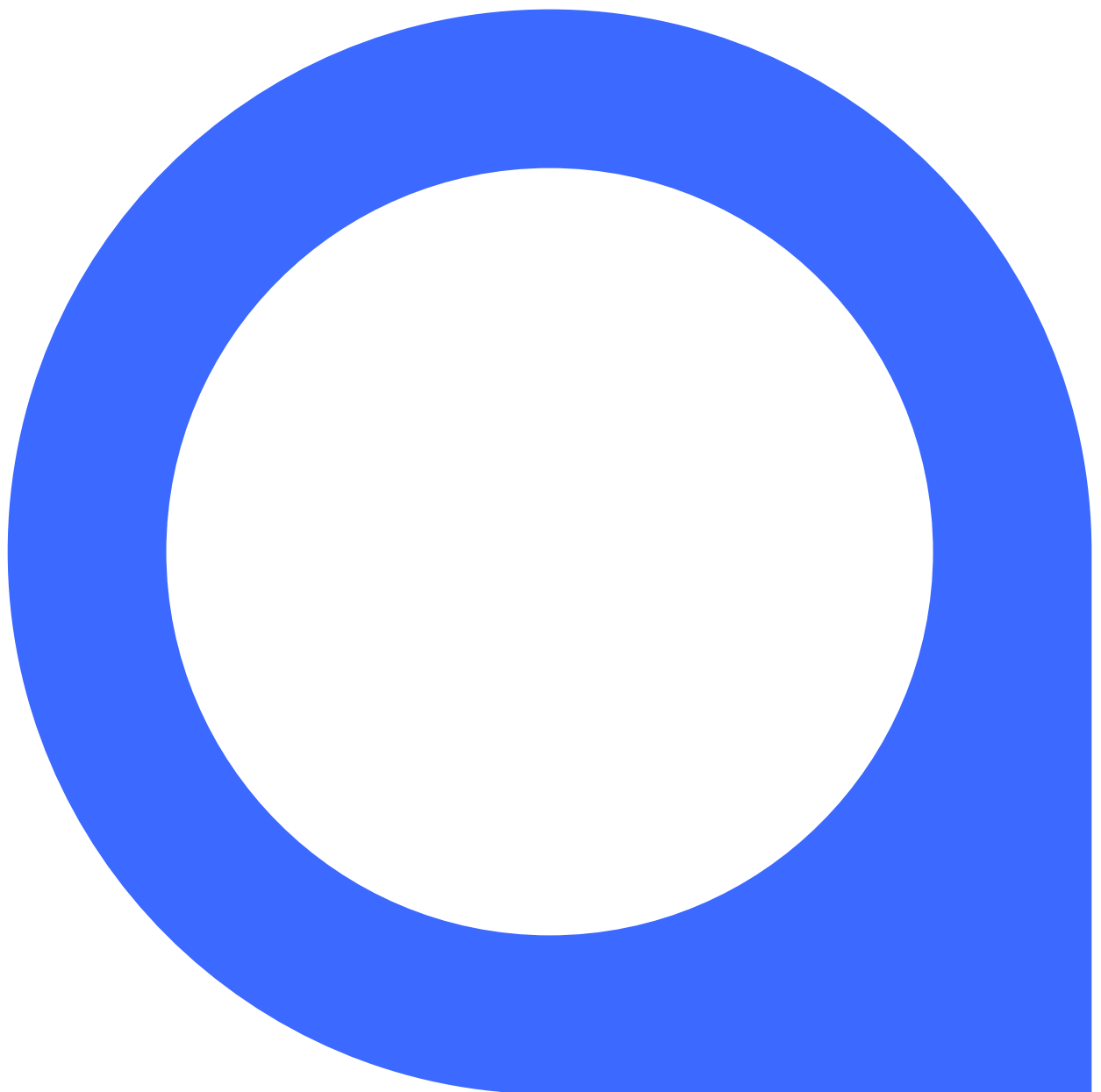


Data Science Applications

Assignment Semester 1 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

- apply appropriate techniques to acquire domain knowledge;
- 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;
- 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;
- 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 chose parameters for your models and derived your answers. It is important that you describe what you did as the marker 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 a minimum of 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 pm (midday, AEST) on 19th April 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 S1 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 three-minute video summary of your analysis and/or findings. Advice about how to record an effective video summary is provided in a separate document on the Assignment page in Canvas. You should submit your video by following these steps:

- create a video recording using the naming convention 'DSA 2024 S1 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 being viewed and running successfully in Google Colab as markers will use this platform to 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 while there is no word limit for the comments in your notebook, 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 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

Deadline

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 24 hours of due date and time: 10% x maximum mark available;
- 1 to 2 days late: 20% x maximum mark available;
- 2 to 3 days late: 30% x maximum mark available;
- 3 to 4 days late: 40% x maximum mark available;
- 4 to 5 days late: 50% x maximum mark available;
- >5 days 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
- a sample assessment graded as 'Significantly above pass level'.



Assignment Context

You are a data science actuary working for a movie production company. The company would like to use data science to improve the scripts for their movies.

The company would firstly like your help in exploring characteristics of famous movie quotes, which will help them write scripts that are memorable, sell box office tickets, and win them Academy Awards.

In addition, many of the actors that the company works with have stipulations in their contracts setting out things they will and will not do when acting in a movie. For example, some actors may stipulate that they will not:

- swear;
- refer to taboo topics like sex, religion, or politics; or
- otherwise say something that might offend members of the public.

The company would like you to build a model to automatically identify quotes in movie scripts that their leading actors may have a problem with, based on the bullet list above.

Your aim in this assignment is therefore to help the movie producer improve their scripts by:

- exploring characteristics of famous movie quotes (Question 2); and
- identifying quotes in movie scripts that actors may have a problem with (Question 3 and 4).

To help you complete your analysis, you have sourced a dataset ('DSA 2024 S1 assignment data.xlsx') containing almost 1,000 famous movie quotes. The data dictionary for this dataset is provided in Table 1.



Table 1: Data dictionary for the assignment dataset

Column name	Data type	Values	Description
Movie Quote	string	various	A famous quote from the movie.
Movie Name	string	various	The movie's name.
Actor	string	various	The actor who said the famous quote.
Character	string	various	The movie character who said the famous quote.
Year	integer	1939 - 2019	The year the movie was released.
Genre	string	various	The genre(s) of the movie.
Director	string	various	The director of the movie.
Writer	string	various	The writer of the movie's script.
Production Company	string	various	The producer of the movie.
Sentiment_VADER	number	[-1, 1]	A sentiment score produced by the Python library 'VADER' where: <ul style="list-style-type: none">-1 is extremely negative sentiment;0 is neutral sentiment; and1 is extremely positive sentiment.
Sentiment_TextBlob	number	[-1, 1]	A sentiment score produced by the Python library 'TextBlob' where: <ul style="list-style-type: none">-1 is extremely negative sentiment;0 is neutral sentiment; and1 is extremely positive sentiment.



Column name	Data type	Values	Description
label_ChatGPT	string	1. sex 2. religion 3. politics 4. rude 5. insensitive 6. none	<p>ChatGPT's response to the following prompt for each movie:</p> <p><i>'For the movie quote 'abc' from the movie 'def', please indicate whether the quote is about:</i></p> <p>1. sex 2. religion 3. politics 4. rude 5. insensitive 6. none of the above</p> <p><i>The answer should just be the number (1 to 6) and the appropriate label from the list above.</i></p> <p><i>For example, if the quote was 'May the Force be with you' from the movie 'Star Wars: Episode VII - The Force Awakens' then the answer would be '2. religion' because the quote could be perceived to be about a greater force or religion.</i></p> <p><i>As another example, if the quote was 'When you're looking for a cancer kid, he should be hopeless. He should have a wheelchair, he should have trouble talking, he should have a little pet goldfish in a Zip-lock bag, hopeless.' from the movie 'Thank You for Smoking' then the answer would be '5. insensitive' because the quote is insensitive towards children with cancer.</i></p>



Assignment Questions (Total 100 marks)

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

Answer Question 4 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.

Prepare the data

Answer Question 1 in your Jupyter notebook.

1.

- a. Examine the dataset to gain an understanding of what you have to work with. Your examination should include a summary of key characteristics of the dataset. *Note you are required to split the data into training, validation, and test sets within this question. You should apply your judgement in deciding when to perform that split.*

(5 marks)

- b. Apply cleaning steps to improve the quality of the data for modelling purposes. *If you choose to undertake some of the data cleaning in Question 1a, you should list these steps in your answer to Question 1b so that markers of Question 1b are aware of all the cleaning steps you have undertaken. Note that cleaning of 'label_ChartGPT' should be done in Question 1d.*

(5 marks)

- c. Calculate vectorised features that represent the cleaned 'Movie Quote', for use in Question 3 (classification) and potentially also in Question 2 (clustering). *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)

- d. Construct a response variable that indicates whether a movie quote is 'problematic' or 'not problematic', to be used in the classifier in Question 3.

(5 marks)



Explore characteristics of famous movie quotes

Answer Question 2 in your Jupyter notebook. Note that the tasks in Question 2 are independent of the tasks in Question 3—the exploration of famous movie quotes in Question 2 is not intended as a lead in exercise to identifying quotes that actors may have a problem with in Question 3.

2.

- a. Apply a clustering algorithm to the dataset to identify different types of famous movie quotes. *Note that you may choose which features to include in the clustering algorithm: you may choose to use the Movie Quote itself, and/or other features in the dataset.*
(5 marks)
- b. Describe the key characteristics of different types of famous movie quotes, based on your clustering outcomes. *Your description should be provided in a text cell following any manual validation that you undertake, be less than 1,000 words, and be presented in a way that is easy for executives of the movie production company to follow.*
(15 marks)
- c. Suggest ways that the company can use your findings from 2a and 2b. *Your suggestion should be less than 500 words and presented in a way that is easy for executives of the movie production company to follow.*
(5 marks)

Identify quotes that may be problematic

Answer Question 3 in your Jupyter notebook.

3.

- a. Suggest strategies you will undertake to address class imbalance in the response variable.
(5 marks)
- b. Construct a neural network to classify quotes as being 'problematic' or 'not problematic'.
(20 marks)
- c. Calculate relevant measures of success for your final selected model. *You will interpret the outcomes of these measures of success in your answer to Question 4a.*
(5 marks)



Answer Question 4 in your pdf document, with the video (4c) included as a YouTube link.

4. Prepare an executive summary of your findings from the classification model, to be presented to executives of the movie production company. Your executive summary should include:

a. a summary of how good your classifier is at meeting the production company's objective (500-word limit);

(5 marks)

b. the limitations of your modelling and actions that could be taken to overcome the limitations (1,000-word limit); and

(10 marks)

c. a three-minute video that explains, to script writers, how your classifier works, to make them more open to using such a model either now or in the future.

(10 marks)

END OF ASSIGNMENT