

DATA ANALYTICS APPLICATIONS

ASSIGNMENT SEMESTER 2 2021



PREAMBLE

The main purpose of this assignment is to help you to:

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

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

- select appropriate techniques for acquiring domain knowledge;
- outline security and privacy laws and regulations that apply to data analytics work;
- develop and evaluate solutions to a classification problem using neural networks;
- implement principal component analysis;
- perform k-means clustering;
- evaluate a clustering algorithm using internal and manual validation;
- apply each step in the natural language processing pipeline to solve a business problem;
- evaluate the outcomes of natural language processing models;
- explain the challenges that apply when using natural language processing models;
- choose a suitable form of model deployment and architecture pattern that meets the needs of users; and
- discuss options for testing a model on a subset of the population.

You will be required to apply knowledge to specific situations in the time-constrained end of semester examination. This assignment provides an opportunity for you to think more deeply and spend significantly more time preparing a detailed answer. This assignment will also help you self-reflect on your writing skills. Whilst there is ample time to write your answers for the assignment, you should ask yourself if you need to spend more time improving your writing skills to help you pass the examination.

The assignment requires you to build models and select appropriate parameters for those models. Consequently, there is no single right answer meaning you will be assessed on your reasoning and process more so than the actual answer you arrive at. 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 can 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 in a non-technical manner. This assignment is designed to test how well you can explain your model and outputs in a straight-forward way to a non-technical audience.



ASSIGNMENT WEIGHTING

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

If you choose not to submit an assignment then you are still entitled to sit the examination, but it is unlikely that you will pass the subject.

It is anticipated that you will spend around 40 to 50 hours to complete the assignment. This is a guide as some students will spend more time than this and some students will spend less.

MARKING RUBRIC

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

SUBMISSION

The deadline for submission is **9:00 a.m. (AEST) on Monday, 27 September 2021**.

Should circumstances arise that mean you cannot submit your assignment on time, you should contact the Chief Examiner in advance of the deadline. If you experience technological issues when submitting your assignment, please attach a copy of your assignment in your email to the Chief Examiner. Penalties will be applied to late submissions without prior approval. These penalties are outlined in the Frequently Asked Questions document on Canvas. We therefore suggest you anticipate potential delays by preparing and submitting your work in advance of the deadline.

The submitted documents must consist of one pdf file and one Jupyter notebook. The Jupyter notebook must be capable of running successfully in Google Colab as markers will use this platform to view and access the notebooks. Files in other formats will not be marked. The naming convention for both files is:

DAA_2021_S2_Assignment_candidate number.

Marks may be deducted if the file name does not follow the required naming convention as this creates additional work for the markers.¹ Penalties for incorrectly named files are outlined in the Frequently Asked Questions document on Canvas.

¹ 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.



A coversheet for the assignment is provided in Canvas. Please attach this coversheet to the front of your pdf file.

Some questions in the assignment may have a specific word or time limit. Markers will not read or watch any part of your answer that exceeds these limits. Please remember to stay within any word or time limits that are specified.

As part of this assignment, you are required to record a 5-minute video summary of your analysis and findings. Advice about how to record an effective video summary is provided in Appendix 1. You should submit your video by following these steps:

- create a video recording using the naming convention 'DAA_2021_S2_Assignment_candidate number.';
- use your video recording to create an 'unlisted' YouTube video (see instructions in Appendix 2);² and
- insert your YouTube video URL as a hyperlink in your assignment pdf file.

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.

ASSIGNMENT CONTEXT

You are an equity analyst working for an active fund manager. One of the fund's investment managers has approached you after listening to an interesting [podcast](#).³ The podcast discussed an experiment in which shares were automatically bought or sold based on the sentiment of Donald Trump's tweets.⁴ Thus, the investment manager would like to explore some ideas with you about using tweets to predict share price movements and boost the fund's investment performance.

After discussing this idea with the investment manager, you agree to investigate the relationship between tweets and the Australian Stock Exchange (ASX) share price movements in the television industry. You have decided to focus on the television industry because you expect that there is a strong relationship between television networks and celebrities who frequently post to Twitter.

² Appendix 2 provides advice for students who do not have access to YouTube due to their location.

³ Goldstein, J. (Host). (2019, October 9). BOTUS. *Planet Money* [Audio podcast]. NPR.
<https://www.npr.org/transcripts/768370374>

⁴ A tweet is a message sent on the social media application Twitter.



To help you complete your analysis, your data team has gathered the following for you:

- a file called 'DAA 2021 S2 Assignment – data SVW prices.csv' containing ASX share price data in 10 minute intervals in the period 1 July 2020 to 23 June 2021 for Channel 7 (ASX code 'SVW')⁵
- a file called 'DAA 2021 S2 Assignment – data NEC prices.csv' containing ASX share price data in 10 minute intervals in the period 1 July 2020 to 23 June 2021 for Channel 9 (ASX code 'NEC')⁵
- a file called 'DAA 2021 S2 Assignment – data tweets.csv' containing a history of tweets to 29 June 2021 for a range of Twitter users linked to the Australian television industry; and
- a file called 'DAA 2021 S2 Assignment – data dictionary.xlsx' that describes the data contained in each of the above three data files.

The investment manager has given you 7 weeks to complete your analysis and prepare a 5-minute video executive summary of your findings.

ASSIGNMENT QUESTIONS

(Total 100 marks)

Questions 1, 2, 5 and 6 below must be answered in your pdf document. These do not need to be provided in a report format but should be written using language suitable for the investment manager.

Questions 3 and 4 must be answered in your Jupyter notebook. Different markers may be reviewing your pdf document and Jupyter notebook so no marks will be awarded for answers to Question 3 and 4 that are contained in your pdf document or answers to Questions 1, 2, 5 or 6 that are contained in your Jupyter notebook.

1. Explain, in 1,000 words or less, some of the key characteristics of the television industry that are relevant to this share price analysis. You should include the source(s) of your information. Your explanation should demonstrate your ability to apply skills in acquiring domain knowledge. Answer this question in your assignment pdf file. **(10 marks)**
2. Explain, in 750 words or less, how your analysis will comply with Twitter's security, privacy, or other relevant data use rules as set out in Twitter's [developer policy](#). Answer this question in your assignment pdf file. **(5 marks)**

The next question requires you to apply natural language processing (NLP) and unsupervised learning to explore the Twitter data contained in 'DAA 2021 S2 Assignment – data tweets.csv'. Answer this question in your assignment Jupyter notebook. Within the notebook, you should explain each of the steps taken in your analysis and evaluate the output from each step.

⁵ This share price data was sourced from [Iress](#), a technology company providing software to the financial services industry in Asia-Pacific, North America, Africa and Europe.



3.

a. Calculate vectorised features that represent the feature 'tweet_text' by applying the following NLP steps:

- i. extract and tokenise; (3 marks)
- ii. clean; (3 marks)
- iii. stem and/or lemmatise; and (3 marks)
- iv. vectorise. (3 marks)

Note that you may choose to 'ii. clean' at different stages of the vectorisation process. You should clearly indicate in your notebook which cleaning steps you are undertaking at different stages.

b.

- i. Explain the advantages and disadvantages of applying principal components analysis (PCA) to reduce the dimension of the vectorised features. (3 marks)
- ii. Apply principal components analysis to reduce the dimension of the vectorised features. (2 marks)

c. Apply k-means clustering to group the tweets into a small number of distinct clusters. (3 marks)

d. Evaluate the clusters created within the context of using tweets to predict share price movements. (5 marks)

The next question investigates the relationship between tweets and the share price movements of Channel 7 (SVW) or Channel 9 (NEC). You can choose to explore the share price of either channel and do not have to investigate both channels. Answer this question in your assignment Jupyter notebook. Within the notebook, you should explain each of the steps taken in your analysis and evaluate the output from each step.

4.

a. Calculate a suitable response variable to indicate whether your chosen TV channel's share price increased, decreased, or remained unchanged following each tweet. This response variable will be used in parts b and c. (10 marks)

b. Construct a neural network to predict whether your chosen TV channel's share price will increase, decrease, or remain unchanged following a tweet from the Twitter users included in this analysis. You should experiment with different network architectures and hyperparameters and different features to use in the model. (10 marks)

c. Construct a tree-based model to predict whether your chosen TV channel's share price will increase, decrease, or remain unchanged following a tweet from the Twitter users included in this analysis. You should experiment with different tree designs and hyperparameters and different features to use in the model. (10 marks)

d. Evaluate how good your selected neural network and tree-based models' predictions are in meeting the objective of this analysis. (10 marks)



Data Analytics Applications

Assignment Semester 2 2021

5. Answer this question in your assignment pdf file. The word limit for this question is 1,000 words.

If the fund were to use the neural network or tree-based model to influence their share trading activities:

- a. explain some of the key risks involved; and (5 marks)
- b. explain some of the key implementation considerations. (5 marks)

6. Prepare a 5-minute video executive summary for the investment manager. The investment manager has requested this form for the executive summary because they find it easier to digest information that is presented verbally. Answer this question in your assignment pdf file as a YouTube hyperlink to your video recording. Your summary should:

- explain the purpose and context of your analysis;
- summarise your findings;
- draw a conclusion about the usefulness of your findings for the business; and
- recommend next steps that should be taken.

(10 marks)

END OF ASSIGNMENT



APPENDIX 1 – VIDEO ADVICE

The following advice is provided to help make your video summary effective and easy for markers to find and understand your key points.

Your video should:

- feature a full or upper body shot of you to help you engage with your audience;
- have an appropriate volume and be free of background noise such that the marker can clearly hear what you are saying;
- not exceed the time limit; and
- not be sped up to fit within the required time limit - if your video is too long then you should consider removing some content.

To create an effective video, you should also remember to:

- plan the video to suit its intended audience and aim;
- apply structure to your presentation, with a clear start, middle and end;
- use transition statements to indicate movement between each of your key topics;
- make speaking notes to remind you of what to say on each key point;
- use visual aids to support your key messages;
- practise;
- engage your audience with your body language and voice; and
- be confident when delivering your message.



APPENDIX 2 – YOUTUBE INSTRUCTIONS

Students who do not have access to YouTube

Some students may not have access to YouTube due to their location. FOR THESE STUDENTS ONLY, please upload your video files directly to Canvas (preferably in an mp4 or mov format). We will then create an unlisted YouTube video for you.

In this case, your submitted video file should use the same naming convention as outlined in the submission section of this document.

Creating an unlisted YouTube video

An unlisted YouTube video is one that will not show up in YouTube search results and can only be seen by people you give the link to.⁶

To create an unlisted YouTube video, you need a Google account. If you don't already have a Google account, the following link provides instructions for setting one up:

<https://support.google.com/youtube/answer/161805?co=GENIE.Platform%3DDesktop&hl=en>

Once you have access to YouTube via a Google account, you are ready to create an unlisted YouTube video. The following YouTube video upload guide provides information about the basic steps required to upload a video to YouTube from either your computer or mobile device:

<https://support.google.com/youtube/answer/57407?co=GENIE.Platform%3DDesktop&hl=en>

When uploading your video, please choose the settings shown in the screen shots below.

⁶ Information about YouTube's privacy settings can be found at:
<https://support.google.com/youtube/answer/157177?co=GENIE.Platform%3DDesktop&hl=en>.



Data Analytics Applications

Assignment Semester 2 2021

EXAMPLESaved as draft

1 Details

2 Video elements

3 Visibility

Details

Title (required)
EXAMPLE

1. TITLE HERE

Description ⓘ
Tell viewers about your video

Thumbnail

Select or upload a picture that shows what's in your video. A good thumbnail stands out and draws viewers' attention. [Learn more](#)

Upload thumbnail

Playlists

Add your video to one or more playlists. Playlists can help viewers discover your content faster. [Learn more](#)

Playlists

Select

Audience

Is this video made for kids? (required)

6% uploaded 14 minutes left

Uploading video...

Video link
<https://youtu.be/smRMxAPgVEU>

Filename
EXAMPLE.mov

VIDEO URL

2. SCROLL DOWN FOR MORE OPTIONS

NEXT



EXAMPLESaved as draft

1 Details2 Video elements3 Visibility

Audience

Is this video made for kids? (required)

Regardless of your location, you're legally required to comply with the Children's Online Privacy Protection Act (COPPA) and/or other laws. You're required to tell us whether your videos are made for kids. [What's content made for kids?](#)

☐ Yes, it's made for kids

☒ No, it's not made for kids

3. SELECT NOT MADE FOR KIDS

☒ Age restriction (advanced)

Uploading video...

Video link
<https://youtu.be/smRMxAPgVEU>

Filename
EXAMPLE.mov

Paid promotion

If another party paid to show a product or service in your video, let us know. Paid promotions need to follow our ad policies and any applicable laws. [Learn more](#)

☐ My video contains paid promotion like a product placement or endorsement

4. SCROLL DOWN FOR MORE OPTIONS

☐ Add a message to my video to inform viewers of paid promotion ?

Tags

Tags can be useful if content in your video is commonly misspelled. Otherwise, tags play a minimal role in helping viewers find your video. [Learn more](#)

Add tag

Enter a comma after each tag 0/500

Language, subtitles, and closed captions (CC)

Select your video's language and, if needed, a caption certification

Video languageCaption certification ?

10% uploaded12 minutes left

NEXT



EXAMPLE Saved as draft

1 Details 2 Video elements 3 Visibility

Select This content has never aired...

UPLOAD SUBTITLES/CC

Recording date and location
Add when and where your video was recorded. Viewers can search for videos by location.

Recording date None Video location None

License and distribution
Learn about [license types](#) and [distribution](#).

License Standard YouTube License

☒ Allow embedding

☒ Publish to subscriptions feed and notify subscribers

Category **"PUBLISH TO SUBSCRIPTION FEED AND NOTIFY SUBSCRIBERS"**
Add your video to a category so viewers can find it more easily

Education

Comments and ratings
Choose if and how you want to show comments

Comment visibility Disable comments Sort by Top

☒ Show how many viewers like and dislike this video

18% uploaded 10 minutes left NEXT

5.UNCHECK "ALLOW EMBEDDING"

6.UNCHECK "PUBLISH TO SUBSCRIPTION FEED AND NOTIFY SUBSCRIBERS"

7.SELECT DISABLE COMMENTS

8. SELECT NEXT



EXAMPLESaved as draft

✓ Details

2 Video elements

3 Visibility

Video elements

Use cards and an end screen to show viewers related videos, websites, and calls to action. [Learn more](#)

You can complete this step after the standard definition (SD) version of your video has been processed. While you wait, you can close this screen or go to the next step.

Add an end screen
Promote related content at the end of your video

ADD

Add cards
Promote related content during your video

ADD

9. SELECT NEXT

21% uploaded 10 minutes left

BACK**NEXT**



EXAMPLE Saved as draft

Details **2 Video elements** **3 Visibility**

Visibility

Choose when to publish and who can see your video

10. SELECT UNLISTED

☒ **Save or publish**
Make your video public, unlisted, or private

☐ **Public**
Everyone can see your video

☐ Set as instant Premiere

☒ **Unlisted**
Anyone with the video link can see your video

☐ **Private**
Only you and people you choose can see your video

☐ **Schedule**
Select a date to make your video public

Before you publish, check the following:

Do kids appear in this video?
Make sure you follow our policies to protect minors from harm, exploitation, bullying, and violations of labor law. [Learn more](#)

Looking for overall content guidance?
Our Community Guidelines can help you avoid trouble and ensure that YouTube remains a safe and vibrant community. [Learn more](#)

11. TAKE NOTE OF VIDEO URL

12. SELECT SAVE

Uploading video...

EXAMPLE

Video link
<https://youtu.be/smRMxAPgVEU>

24% uploaded 10 minutes left

BACK **SAVE**



Video uploading

Your video is still uploading. Keep this browser tab open until uploading completes. Your video will be **unlisted** once uploading and processing finishes.

EXAMPLE

**13. KEEP BROWSER WINDOW OPEN
UNTIL VIDEO UPLOAD COMPLETE**

 29% uploaded 9 minutes left

CLOSE

Once your video has finished uploading, you should copy the video URL (see step 11 in the diagrams above) and paste this into your presentation coversheet.

Optional step: using a 'brand channel' to hide your name

You will not be anonymous in your video as your face will be visible.⁷ However, it is preferable that your name does not appear in your video or in the YouTube channel that you upload your video to. The following link provides information about how to create a new channel in YouTube using a brand name rather than your personal name:

<https://support.google.com/youtube/answer/1646861?hl=en>.

Please use these instructions to create a new channel that does not include your personal name. The actual name you choose does not matter.⁸

⁷ There is a process in place to ensure that markers do not mark videos for students that they know.

⁸ You will not be penalised if you do not follow this optional step when uploading your video to YouTube.