

# DATA ANALYTICS APPLICATIONS

ASSIGNMENT SEMESTER 1 2022



### 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;
- evaluate how well data describes business activity;
- develop and evaluate solutions to a classification problem using GLMs, tree-based models, ensembling and neural networks;
- 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 business problem; and
- implement strategies for gaining stakeholder support for data analytics projects.

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 and presentation 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.

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 **12:00 noon (AEST) on Monday 11 April 2022**.

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. Files in other formats will not be marked. The naming convention for both files is:

**DAA\_2022\_S1\_Assignment\_candidate number.**



If an assessment is submitted in a format with an incorrect file name or an incorrect format (e.g. the file name has no candidate number or the file is submitted as, say, a word document when a pdf document was required), you may be required to resubmit your assessment. This may cause you to submit late and hence incur a late submission penalty. You should therefore follow all assessment instructions provided.<sup>1</sup>

### PDF file

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\_2022\_S1\_Assignment\_candidate number';
- use your video recording to create an 'unlisted' YouTube video (see instructions in Appendix 2);<sup>2</sup> 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 of the steps 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.

---

<sup>1</sup> 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.

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



Please note that while there is no word limit for the comments that you include 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 very difficult for a marker to assess your understanding of the step being taken.

### 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 actuary who has been engaged by a small book publisher. They publish books from several different authors and across different genres. When deciding whether to accept a book proposal from an author, it is helpful for the publisher to have a rough estimate of the sales volumes that the book is likely to achieve. This can be influenced by a range of factors such as the author's popularity and previous successes, the book's title, and the book's description.

The publisher would like your help in building a model that will give them the capability to predict the sales volumes for books before they are published. Due to the publisher's small size, it has very limited data on past book sales, but has sourced a dataset called 'DAA 2022 S1 Assignment – book data.csv' from Goodreads' 'Best Ever Book List' [https://www.goodreads.com/list/show/1.Best\\_Books\\_Ever](https://www.goodreads.com/list/show/1.Best_Books_Ever) in 2020 (the book dataset). The data dictionary for the book dataset is provided in the table below. The publisher has made no attempt to clean this data before providing it to you and notes that there are some strange looking entries for some of the books.

The publisher has noted that actual sales data for books not published by them is very hard to come by and they were unable to obtain this information for the books in the book dataset. However, they have referred you to an article that discusses the relationship between Goodreads ratings and sales volumes (<https://bookriot.com/goodreads-ratings-and-sales/>). The publisher has suggested that you focus on predicting the number of Goodreads ratings that a book will receive and then convert this into an estimated number of sales using a rough 'rule-of-thumb' based on the information in the article or other research that you will carry out.



**Table 1: Data dictionary for the book dataset**

Column name	Data type	Values	Description
title	string	varied	The title of the book.
link	string	URLs	The URL address of the Goodreads listing of the book.
series	string	varied	The series that the book is from, if applicable.
author	string	varied	The author(s) of the book, separated by ','.
rating_count	number	non-negative integers	The number of Goodreads users that have rated the book.
review_count	number	non-negative integers	The number of Goodreads users that have written reviews of the book.
average_rating	number	0 to 5	The average rating (out of 5) given to the book by Goodreads users.
five_star_ratings	number	non-negative integers	The number of Goodreads users that have rated the book five stars.
four_star_ratings	number	non-negative integers	The number of Goodreads users that have rated the book four stars.
three_star_ratings	number	non-negative integers	The number of Goodreads users that have rated the book three stars.
two_star_ratings	number	non-negative integers	The number of Goodreads users that have rated the book two stars.
one_star_ratings	number	non-negative integers	The number of Goodreads users that have rated the book one star.
number_of_pages	number	non-negative integers	The number of pages in the book.
date_published	date/time	varied	The date the book was published.
publisher	category	varied	The publisher of the book.
genre_and_votes	category	varied	The genre of the book, followed by number of users voting for this genre. Multiple genres are separated by a comma.
isbn	number	9-13 digits	The unique International Standard Book Number (ISBN) of the book.
description	string	varied	A description of the book.



### ASSIGNMENT QUESTIONS

**(Total 100 marks)**

Questions 1, 6, and 7 below must be answered in your pdf file. These do not need to be provided in a report format but should be written or presented using language suitable for communication with the publisher.

Questions 2 to 5 must be answered in your Jupyter notebook using the assignment template provided.

Different questions 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 contained in your answers to other questions.

1. Explain, in 1,500 words or less, some of the key characteristics of the book publishing industry that are relevant to your 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. **(15 marks)**

Answer Questions 2 to 5 below in your assignment Jupyter notebook, using the notebook template provided.

2. Describe the book dataset using exploratory data analysis. This exploratory data analysis should give you a better understanding of the data that you have to work with when answering later questions in the assignment. You should focus on features that will help you answer Questions 3 and 5. **(10 marks)**
3. Calculate vectorised features that represent the title and description features in the book dataset. You should use tokenisation, cleaning, stemming or lemmatisation, and vectorisation to calculate these features which will be used in Questions 4 and 5b. **(13 marks)**
4. Examine the book titles in the dataset by applying clustering algorithms to the vectorised features representing each book's title. This step is designed to give you a better understanding of the different book titles that publishers use, as each book's title may be a feature that you decide to use in your answer to Question 5. **(7 marks)**





5.

- a. Calculate a response variable to indicate the estimated sales for each book based on the information available in the book dataset and the domain knowledge you have acquired in answering Question 1. Note that this response variable should be suitable for use in a classification model. **(10 marks)**

- b. Construct a classification model to predict, prior to publication, the total future sales of a book based on its title and/or description, or other features in the book dataset. You should experiment with different types of classifiers, model architectures, and hyperparameters. **(15 marks)**

- c. Evaluate how good your selected model's predictions are in meeting the objective of this analysis. **(10 marks)**

6. Explain, in 1,500 words or less, any limitations of the analysis you have completed, including steps you could take to overcome these limitations. These limitations might relate to the data and/or methodology used in your modelling. Answer this question in your assignment pdf file. **(10 marks)**

7. Prepare a 5-minute video executive summary of your findings for the publisher. 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; and
- draw a conclusion about the usefulness of your findings for the publisher, including any next steps that you would recommend.

**(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.

Please note that your video does not have to be 'perfect' to score full marks for it. The rubric provides more information about the exact criteria on which your video will be marked.



## 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.<sup>3</sup>

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.


---

<sup>3</sup> 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 1 2022

**EXAMPLE**Saved as draft×

1 Details

2 Video elements

3 Visibility

### Details

1. TITLE HERE

Title (required)  
EXAMPLE

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

**VIDEO URL**

2. SCROLL DOWN FOR MORE OPTIONS

Uploading video...

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

Filename  
EXAMPLE.mov

NEXT



# Data Analytics Applications

Assignment Semester 1 2022

**EXAMPLE**Saved as draft✕

1 Details

2 Video elements

3 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  
  
▼ **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  
☐ 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 language

Caption certification ?

10% uploaded 12 minutes left

NEXT

3. SELECT NOT MADE FOR KIDS

4. SCROLL DOWN FOR MORE OPTIONS



**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

Uploading video...

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

Filename  
EXAMPLE.mov

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



**EXAMPLE**Saved 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

☒ **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](#)

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 assignment file as instructed.

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

You will not be anonymous in your video as your face will be visible.<sup>4</sup> 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.<sup>5</sup>

<sup>4</sup> There is a process in place to ensure that markers do not mark videos for students that they know.

<sup>5</sup> You will not be penalised if you do not follow this optional step when uploading your video to YouTube.