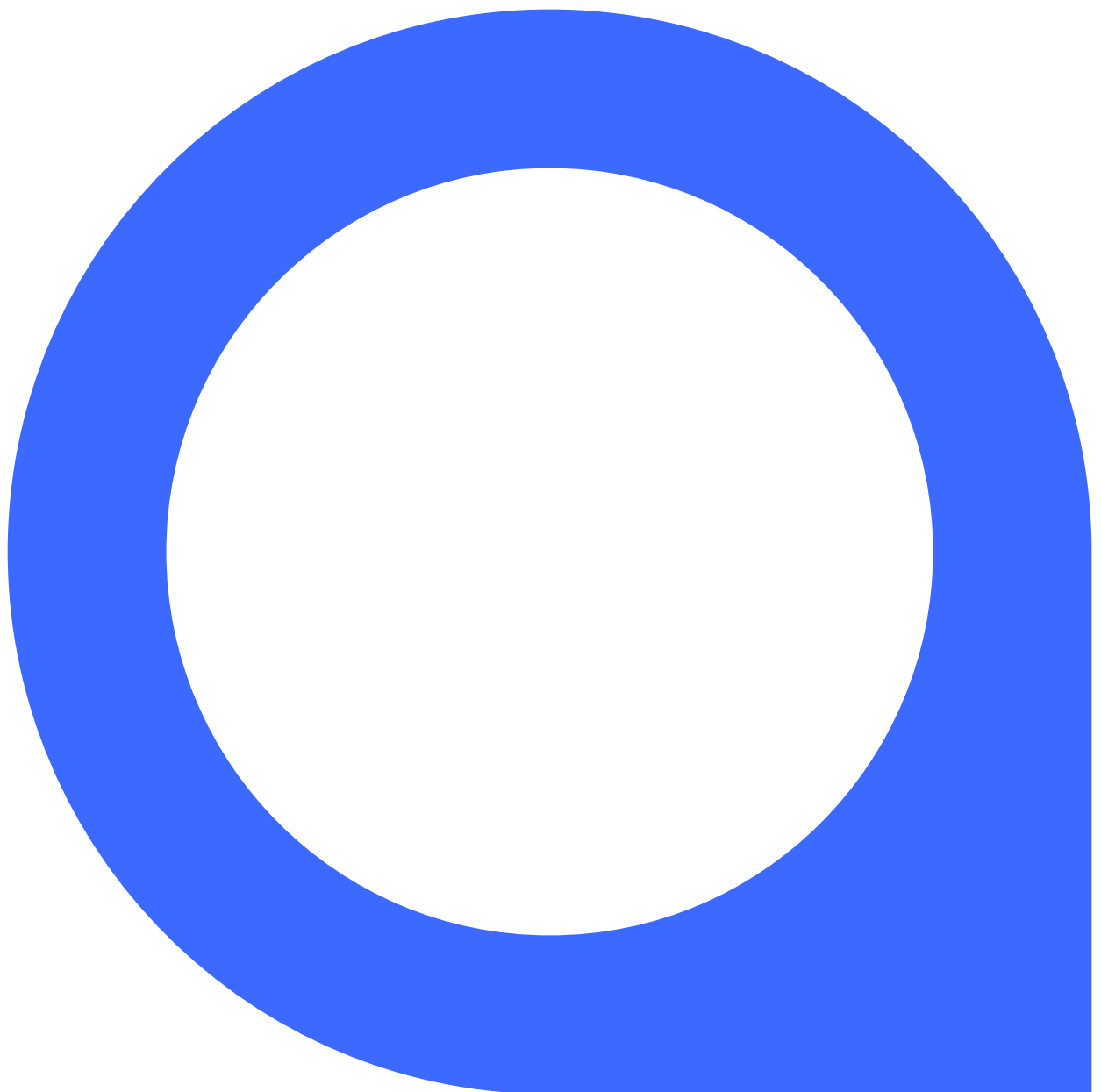


# Data Science Applications

Assignment Semester 1 2025





## 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:<sup>1</sup>

- 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.

---

<sup>1</sup> 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 subject<sup>2</sup>. 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.

---

<sup>2</sup> 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 17 April 2025**.

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](mailto: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](mailto: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 2025 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 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 2025 S1 Assignment member ID';
- use your video recording to create an 'unlisted' YouTube video (see instructions in the Appendix)<sup>3</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 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.

---

<sup>3</sup> 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 consulting actuary engaged by Bigtel, an Australian mobile network retailer specialising in selling to consumers (not to businesses). The National Consumer Association recently published a report comparing mobile network retailers, placing Bigtel poorly against their main competitors. This has impacted the Bigtel share price, and the Board of Directors is demanding management take corrective action immediately. The directors' concerns include:

- ongoing problems with subscription pricing caused by a bug-ridden pricing system,
- customer service call centre costs that exceed industry averages, and
- customer churn rates that exceed industry averages.

Bigtel's management has asked you to use your data science skills to address these issues.

To help you complete this task, Bigtel has provided you with a file ('DSA 2025 S1 assignment data.xlsx' or 'the assignment dataset') containing a sample of data for all customer activity up to 31-Jan-2025. The data dictionary for this dataset is set out in Table 2 and 2.

**Table 1: Tables in the assignment dataset**

Table name	Description
Customer	Customer details, including their name, address, date of birth, and payment details. Note: this table can contain multiple rows per customer, one row for each version of the details for that customer.
Invoice	Monthly invoice events showing the total invoice amount.
Items	Invoice items, the monthly usage of phone calls and data, one row per usage type, no row if usage was zero.
Call Centre	Transcripts from inbound phone calls from customers asking billing and technical support questions.

**Table 2: Data dictionary for the assignment dataset**

Table name	Column name	Data Type	Description
customer	RowId	string	Unique identifier of each row in the table, in GUID format. Uniquely identifies each customer and version combination.



Table name	Column name	Data Type	Description
customer	CustomerId	categorical	The unique identifier of each customer, in GUID format.
customer	ValidFrom	datetime	UTC timestamp of when this version of a customer's attributes becomes valid or live
customer	Active	numeric	Whether this customer has any active cell phone service subscriptions. True for has active contracts, false if no active cell phone service subscriptions.
customer	Gender	categorical	The customer's gender. Can only have values Male or Female.
customer	Contract	categorical	The subscription contract type. This can only take the values 'Month-to-month', 'Two year', or 'One year'. This value is ordinal categorical, as it represents a period of time.
customer	PaperlessBilling	categorical	Whether the invoice is sent digitally ('Yes') or a paper invoice is mailed to the customer ('No').
customer	PaymentMethod	categorical	The method of invoice payment. Can only take the values 'Electronic check', 'Mailed check', or 'Bank transfer (automatic)'
customer	DateOfBirth	date	The customer's date of birth.
customer	DependentCount	numeric	The customer's number of dependent children aged less than 18.
customer	MaritalStatus	categorical	The customer's marital status. Can only take the values 'Single' or 'Married'
customer	Title	categorical	The customer's title. Can only have values Mr. Mrs. or Ms.
customer	GivenName	categorical	The customer's given name
customer	MiddleInitial	categorical	The customer's middle initial.
customer	Surname	categorical	The customer's family or surname.
customer	StreetAddress	categorical	The customer's residential address' building number and street name.



Table name	Column name	Data Type	Description
customer	City	categorical	The city name of the customer's residential address
customer	State	categorical	The 2-character state code of the customer's residential address. For example, TX is the code for Texas.
customer	ZipCode	categorical	The zip code of the customer's residential address. Contains only digits but is a categorical variable.
customer	EmailAddress	categorical	The customer's email address.
customer	TelephoneNumber	categorical	The customer's phone number.
customer	CCType	categorical	The customer's credit card type. Can only take the values 'Visa' or 'MasterCard'
customer	NationalID	categorical	The customer's national identification number.
customer	Latitude	numeric	The latitude of the customer's residential address.
customer	Longitude	numeric	The longitude of the customer's residential address.
customer	record_available_at	datetime	A UTC timestamp for when this row was added to the database.
invoice	InvoiceEventId	string	Unique identifier of each row in the table, in GUID format. Uniquely identifies each monthly invoice sent to the customer.
invoice	CustomerId	categorical	The unique identifier of each customer, in GUID format.
invoice	InvoiceTimestamp	datetime	The UTC timestamp of when this invoice event occurred.
invoice	InvoiceAmount	numeric	The total amount of the invoice, to be paid by the customer.
invoice	tz_offset	categorical	The local timezone offset of the invoice event.
invoice	record_available_at	datetime	A UTC timestamp for when this row was added to the database.



Table name	Column name	Data Type	Description
invoice items	InvoiceItemId	string	Unique identifier of each row in the table, in GUID format. Uniquely identifies each cell phone service usage item within an invoice.
invoice items	InvoiceEventId	categorical	Uniquely identifies the invoice event to which this cell phone usage value belongs.
invoice items	UsageType	categorical	The cell phone usage type. Can only take the values 'Data Usage (GB)', 'International Calls (minutes)', 'Local Calls (minutes)', 'SMS', 'Roaming Data Usage (GB)' or 'Roaming Calls (minutes)'
invoice items	Usage	numeric	The monthly amount of cell phone usage. This will be in units of gigabyte for data, and minutes for phone calls.
invoice items	record_available_at	datetime	A UTC timestamp for when this row was added to the database.
call center	CallCentreEventId	string	Unique identifier of each row in the table, in GUID format. Uniquely identifies each service center phone call transcript.
call center	CustomerId	categorical	The unique identifier of each customer, in GUID format.
call center	Timestamp	datetime	The UTC timestamp of when this customer service call center phone call occurred.
call center	Transcript	string	The full text transcript of the customer service call.
call center	tz_offset	categorical	The local timezone offset of the invoice event.
call center	record_available_at	datetime	A UTC timestamp for when this row was added to the database.
customer	State	categorical	The 2-character state code of the customer's residential address. For example, TX is the code for Texas.
customer	ZipCode	categorical	The zip code of the customer's residential address. Contains only digits but is a categorical variable.
customer	EmailAddress	categorical	The customer's email address.



Table name	Column name	Data Type	Description
customer	TelephoneNumber	categorical	The customer's phone number.
customer	CCType	categorical	The customer's credit card type. Can only take the values 'Visa' or 'MasterCard'
customer	NationalID	categorical	The customer's national identification number.
customer	Latitude	numeric	The latitude of the customer's residential address.
customer	Longitude	numeric	The longitude of the customer's residential address.
customer	record_available_at	datetime	A timestamp for when this row was added to the database.



## Assignment Questions (Total 100 marks)

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

Answer Questions 2 and 4 in your pdf document.

Different markers will review different questions in this assignment, so your answer to each question and part of that 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 examine the call transcript data

Your work with Bigtel will start by providing them with a summary of the transcript data they provided.

Answer Question 1 in your Jupyter notebook.

- a. Clean the 'Transcript' data in the call centre event table, then calculate vectorised features representing that column. Use both word embeddings and TF-IDF vectorisation. *You should not partition the data before doing this cluster analysis.* (5 marks)
- b. Apply a clustering algorithm using the vector features calculated in Question 1a to provide insights into the primary natures and outcomes of those calls. *You should justify your choice of clustering algorithm, distance measure, and number of clusters, comparing your choices to alternatives. Use internal validation and the business context to support your justifications.* (5 marks)
- c. Apply discriminator modelling that predicts your clusters (from Q1b) using only the TF-IDF vectorisations, to understand the top 10 keywords that distinguish each cluster from another. *You should use a random forest algorithm and variable importance. Use the default hyperparameters for a random forest algorithm. You are not required to assess the performance of the model.* (5 marks)
- d. Examine the clustering outputs using manual validation. (10 marks)



- e. Summarise, in 500 words or less, the results of your cluster analysis, and how they relate to the three issues management have asked you to address. *Your answer should be communicated using language suitable for sharing with the management team at Bigtel.*

**(5 marks)**

## 2. Use generative AI (GenAI) to categorise call transcripts

Your next task is to investigate the use of GenAI tools to assess the sentiment, topic, and outcome of Bigtel's call transcripts. Note that you may use any GenAI tool to create the summaries of the asset.

Answer Question 2 in your pdf file.

- a. Construct an LLM prompt that scores the sentiment (from -1 to 1), categorises the reason for the call (into technical support, change subscription plan, billing, or other), identifies whether this is a repeat call by the customer for this specific issue, and estimates the call outcome (into a net promoter score). The output should be in JSON format. Apply your LLM prompt to ten randomly selected call transcripts.

**(5 marks)**

- b. Summarise, in 500 words or less, the results of your LLM prompt on the ten transcripts and how they relate to the three issues management have asked you to address. *Your answer should be communicated using language suitable for sharing with the management team at Bigtel.*

**(5 marks)**

## 3. Predict customer churn

Your next task is to build a neural network model that predicts customer churn. Answer Question 3 in your Jupyter notebook.

- a. Examine and then clean the assignment dataset to understand the data. Do not examine or clean the 'Transcript' data column, as that was done earlier in this assignment. *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. Propose a unit of analysis, stating the entity and timestamp(s) that identify each unique row in your training and scoring data. *As part of your answer, you should propose whether to use regularly spaced timestamps, event-driven timestamps, or a combination of both.*  
**(5 marks)**
- c. Construct a response variable for your classification model. *You must explain your choices in the design of the response variable you constructed.*  
**(5 marks)**
- d. Suggest four metrics you will use to evaluate the success of your classifier. **(5 marks)**
- e. Construct your neural network classifier. *You must demonstrate an ability to fine-tune model architecture, hyperparameters, feature selection, regularisation, and optimisation algorithm.*  
**(20 marks)**
- f. Interpret the performance of your chosen neural network classifier using the metrics suggested in Question 3d and a comparison to benchmarks. *Your answer should be communicated using language suitable for sharing with the management team at Bigtel.*  
**(5 marks)**
- g. Interpret the behaviours of your chosen neural network classifier using feature importance, partial dependence (on the top 5 most important features), and SHAP explanations (on three validation examples). *Your answer should be communicated using language suitable for sharing with the management team at Bigtel.*  
**(5 marks)**

## 4. Video summary of findings

Answer Question 4 in your pdf file.

Prepare a five-minute video, for presentation to the management team at Bigtel, to summarise your findings from Questions 1, 2, and 3. *You should structure your video to have a clear start, middle, and end, with clear transitions between all sections.*  
**(10 marks)**

**END OF ASSIGNMENT**