



Detailed Assignment Feedback

Question 1 – Domain knowledge

In this question, students had to research the book publishing industry and explain the key characteristics of the industry that would be relevant to their analysis.

Most students were able to find relevant sources relating to the book publishing industry and use these to identify key book industry characteristics such as the different publishers, genres, authors in the market and trends in sales over time.

The best answers provided a clear explanation of the relevance of each of these characteristics for the analysis to be undertaken. These answers often contained a sub-heading to indicate where this discussion of relevance for each characteristic could be found in their answer.

Weaker answers outlined several characteristics of the industry but did not explain the relevance of these characteristics for their analysis. These weaker answers also lacked structure, making it difficult for markers to identify each key characteristic and/or its relevance. Some students provided more detail than was necessary about each characteristic, leaving too few words to explain each characteristic's relevance.

Some students spent time analysing the GoodReads website as opposed to investigating the book publishing industry as-a-whole, as required by the question and rubric.

Question 2 – Exploratory data analysis

In this question students had to describe the book dataset using exploratory data analysis.

Most students were able to identify missing values and duplicate observations of books, as well as proposing suitable fixes to these issues.



Better answers commented on the fact that books published more recently had less developed book review data than books published several years ago. Better answers also used a good structure in their notebook, making it easier for markers to understand what was being explored in each section. Some students listed, for each feature, its usefulness, issues, and proposed features with sub-headings for each. The best answers also examined the correlation between features and commented on why some features that are heavily correlated with the response variable were not included in the analysis to avoid feature leakage (i.e. features that would not be available at the time of publishing, such as 'number of 5-star ratings').

Weaker answers failed to discuss the usefulness of each feature within the problem context. For example, they might have examined, say, six features without commenting on why they were being examined in the context of their later analysis. Weaker answers also failed to provide sufficient commentary on the visualisations/plots they produced for each feature. Weaker answers also provided commentary within the code cells, rather than as separate text cells above and below their code. This made it harder for markers to follow their analysis.

Question 3 – Vectorising features

This question required students to calculate vectorised features to represent each book's title and description numerically.

Most students implemented at least four reasonable cleaning steps to meet the marking rubric requirements. Few students explained each of these steps and why they had been undertaken. Stronger answers identified the implications each cleaning step would have on the output of their vectorisation. Explanations that were also linked to relevant data checks showed the stronger students' understanding of a natural language process.

Some students used the 'clean_text' function verbatim from Case Study 2 in Module 7 to tokenise and lemmatise the features. While this itself did not reflect negatively on a student's result, very few students who took this approach offered any explanation of each step in the function and why it was important to perform.



While most students used existing Python packages for vectorisation of text (especially for TF-IDF), others built these functions from scratch. Neither option was preferred, provided students checked and validated the output from either method. However, stronger answers considered multiple vectorisation methods (e.g. TF-IDF and BERT-based) and were able to compare these methods and justify their final choice.

Question 4 – Clustering

This question required students to examine book titles by applying clustering algorithms to the vectorised book titles.

Overall, markers were impressed with most students' ability to execute a clustering model on complicated data. Most students were also able to perform checks on the clustering and interpret the results using various validation methods.

Better answers had each piece of their analysis contributing to telling a story about the book titles. These answers discussed why the student had implemented various methods, including the pros and cons of the method within the problem context, performed their analysis, then discussed the outputs of the clustering. The strongest answers interpreted the output, outlined key insights from it and related it strongly to the problem context and next steps to be performed. Better answers also supplemented their discussion with external research with references to the source of that information.

Weaker answers didn't provide any recommendations from the clustering exercise. They gave textbook answers rather than critically evaluating the output from their clustering in relation to the book sales prediction analysis. For example, some students suggested using 50 clusters because there was an elbow in the elbow curve at this point. These students did not demonstrate that they had thought through the implications of this suggestion for their book sales predictions.

Weaker answers also had issues in their output, such as class imbalance, no elbow in the elbow curve, or had clusters that didn't make sense, and did not comment on these issues. If the output had issues, typically due to not reducing the dimension of the vectorised feature, students still scored quite well if they were able to identify and discuss these issues.



Question 5 – Classification

This question required students to build a classification model to predict book sales.

5a – Response variable

The strongest answers recognised that the number of book reviews depends heavily on how long a book has been published for and made an appropriate adjustment to allow for unequal development times of different books. Stronger answers also justified their choice of book sales bands based on the needs of the publisher and their research from Question 1.

Weaker answers did not recognise that the number of book reviews was under-developed for more recently published books. Weaker answers also mechanically created different bands for predicted book sales without considering the problem context and what these bands might mean for the book publisher.

Most students incorrectly believed that they were adding value by doing a log-transformation of the response variable.

5b – Classification model

Most students were able to iterate through several classifiers, but only some students provided strong logic between their iterations, demonstrating that their iterations were directed at improving the model's outcomes.

The strongest answers took steps to reduce overfitting such as optimising the number of epochs and hyperparameter optimisation. These answers also tended to experiment with feature engineering, such as grouping together highly ordinal features. Strong answers also looked to understand their model such as by comparing actual and predicted book sales/book reviews or by producing variable importance or partial dependence plots.



Weaker answers paid too much attention to training set results when choosing a final model, rather than validation metrics. These answers also tended to focus only on metrics such as accuracy, without paying enough attention to other confusion matrix metrics such as precision and recall and what these meant for classifying books into different predicted book sales buckets. Few students understood that accuracy is not a well-behaved metric to train data analytics models on, particularly when the response variable has imbalanced classes. Optimising on a metric such as AUC or log-loss would have reduced the number of problems students had with their models.

Weaker answers also had feature leakage, which resulted in their models seeming to fit the data with high accuracy. For example, some students used features such as 'Number of 5-star ratings' which is not known at the time of publishing. As another example, some students created an author popularity feature but based this on books in the training dataset.

Several students didn't understand that you don't need to one-hot-encode numeric features and that doing so will lose information and degrade the performance of a model.

5c – Model evaluation

The best answers to this question linked their evaluation of the model to the publisher's needs, rather than just restating basic definitions of confusion matrix metrics. Better answers also thought outside the box and applied metrics such as weighted costs, again with strong linkage to the operating model of the publisher and how the model would be used in practice.

Weaker answers to this question did not understand how to choose a suitable benchmark. Those answers that had imbalanced classes typically misapplied the random benchmark, allocating books to each class with the same probability. The majority class classifier (where all observations are assigned to the most common class) would have been a better benchmark in this instance.



Question 6 – Limitations

In this question students were required to explain limitations of their analysis and steps that could be taken to overcome these limitations.

Most students explained at least five limitations and explained steps that could be taken to overcome each limitation. The limitations explained included:

- the dataset not containing actual sales;
- the limited number of features available;
- the dataset only including the 'best books ever';
- the use of a classification rather than regression model;
- the dataset containing books published over a large time span; and
- data quality issues in the dataset.

Strong answers typically had a clear structure. These answers tended to identify a broad range of key limitations and they expressed themselves clearly.

Weak answers focussed on only one or two key limitations. They might have listed four or five limitations but many of these were the same issue. Weaker answers also included generic limitations that were not context specific, such as memory and processing speed, model interpretability, and maintenance requirements of the model. They also covered limitations that had already been addressed in the modelling, such as data cleansing requirements which had already been performed.

Weak answers were also difficult to understand or did not explain steps that could be taken to overcome the limitations (e.g. the steps were only outlined or were unacceptable solutions).

Question 7 – Executive summary

In this question students were required to provide a five-minute executive summary of their findings for the publisher.



Data Analytics Applications

Semester 1 2022 Assignment Feedback

Most students were able to structure their presentations using segments such as context, data, methodology, modelling, and conclusions or next steps. It should be noted that the question did not ask for all these areas to be covered but asked students to focus on the context, findings, and conclusions or next steps. As a result, some students wasted much of their time discussing the data, methodology, and modelling which was likely of lesser interest to the book publisher.

In presenting the evaluation of the model, most students presented a range of metrics such as accuracy, precision, and recall. However, only better presentations identified what these meant for the small book publisher. For example, these presentations might have recommended that for a small publisher with limited resources, precision should have been maximised over recall to publish a small number of books with the highest probability of success.

Weaker presentations provided a step-through of each assignment question, as opposed to an engaging presentation for the publisher. These presentations typically did not link their recommendations to the problem context nor explain technical terms in language suitable for the target audience (the publisher).

Weaker presentations also tended to include students reading off a script rather than reading from brief speaking notes after having rehearsed the presentation multiple times prior to finalising it.

Sample assignment graded as 'Significantly above pass level'

A sample assignment is provided as an example of one that was graded as 'Significantly above pass level'. Students should use this example, along with the assignment rubric, to help them self-assess their own assignment attempts.

The answers in the sample assignment were outstanding. The student's answers to all questions had good structure and clearly addressed each rubric criterion. Specific comments on each of their assignment answers is provided below:

- **Question 1:** They clearly outlined five key characteristics with an excellent presentation of the relevance of each characteristic for the analysis to be undertaken.
- **Question 2:** Their analysis was rigorous with good summaries of the output.



Data Analytics Applications

Semester 1 2022 Assignment Feedback

- **Question 3:** They provided a well-structured, concise explanation of why each step was performed and completed multiple reasonableness checks for each step. They used a variety of vectorisation techniques and compared these before choosing the approach to use going forward.
- **Question 4:** They performed internal, external, and manual validation and provided a context-related justification for choosing three clusters.
- **Question 5:** They made very strong adjustments for lack of development in book ratings. They justified the number of bins selected for the response variable and checked their output throughout the code. They made excellent use of feature engineering to enhance the data. Their chosen model metrics were strongly linked to the business context.
- **Question 6:** They covered a wide range of limitations and split each limitation into an explanation and steps to overcome the limitation, making it easy for the markers to see they had met all rubric criteria.
- **Question 7:** They explained technical concepts in plain English and explained all key findings very well.

Please note that this assignment is not 'perfect' and there were other ways to answer each of the questions and still achieve very high marks.