



Detailed Assignment Feedback

Question 1 – Domain knowledge

This question was generally answered well.

The most common mistake made by students was a failure to link all the newly acquired domain knowledge on the television industry back to the context of using tweets to predict share price movements. For example, some students included a lot of relevant information about the industry but did not explicitly link this back to the analysis to be conducted. Other students attempted to link industry features back to the analysis, but these links were not particularly convincing. A statement of how each industry feature was relevant to the analysis was a requirement of a passing grade for this question.

The best answers made it clear where they had sourced their television industry knowledge from, with thorough footnoting of almost every assertion. These answers made it clear that the students were able to acquire knowledge of a new industry. Weaker answers did not make it clear where they had gained their knowledge from, with a few students neglecting to reference their sources altogether. These weaker answers did not demonstrate an ability to acquire domain knowledge about a new industry. For example, some of these answers appeared to draw on the student's existing (limited) knowledge of the industry, rather than by using the domain knowledge acquisition skills taught in the subject.

The best answers also presented data to back up their assertions. Weaker answers made high level comments without sufficient detail to support their conclusions. For example, some answers stated that Channel 7 only has a small impact on SVW's share price but did not provide supporting evidence such as the percentage of SVW's profits that were linked to Channel 7.



Question 2 – Twitter rules

Many students discussed similar components of Twitter's rules in this question. For example, 'Off-Twitter matching' was a common rule that was explained.

The best answers briefly outlined relevant rules and provided a clear, specific explanation of how their analysis would comply with each rule. Some students summarised all the Twitter rules in a table and indicated whether they were relevant to the analysis. While this was not a requirement to gain full marks in this question, it helped to demonstrate a very clear understanding of the various Twitter rules and their relevance for this analysis.

Some weaker answers stated how they would comply with three or four rules but did not actually state what the rules were. Other weaker answers stated that rules were already being complied with because they were not relevant to the analysis.

Question 3 – NLP, PCA and clustering

Markers observed a wide range of Python coding proficiency in student answers to this question. Markers also recognised that different Python packages exist to perform similar tasks. Students could use any reasonable package for their analysis.

The strongest answers to this question:

- structured their notebooks according to the sections in the assignment, to make it clear to markers where they had addressed each rubric criterion;
- discussed each step that was being undertaken and why it was being performed;
- performed a range of checks on their code's output.

3a – Natural language processing

Strong answers to this part of the assignment:

- performed exploratory analysis relevant to the problem, such as looking at the number of unique tokens, n-gram frequencies and distinct characters, to get a better understanding of the data they were working with;



Data Analytics Applications

Semester 2 2021 Assignment Feedback

- performed robust checking after each step had been carried out, such as checking random samples of tweets to ensure that each step was carried out as expected and performing holistic checking such as checking the number of unique tokens produced after each step;
- added custom words to the spell-checking dictionary, to recognise the unique nature of the tweet text;
- considered multiple vectorisation methods such as TF-IDF and BERT;
- looked at token frequencies when deciding how to reduce the corpus size when fitting TF-IDF; and
- calculated summary statistics or manually calculated a few vector values to confirm that the vectorisation results were as expected.

Weaker responses:

- only checked items such as the dataset's dimension, column names and feature types;
- when checking outputs, only checked the first N tweets, rather than a random sample;
- applied text cleaning steps such as removing unwanted characters, case conversion, and removing stop words without proper consideration or comment on the potential loss of information when performing each of these steps;
- arbitrarily reduced the corpus size when fitting TF-IDF to meet memory requirements of Google Colab.

3b – Principal component analysis

To score a 4 or 5 on this question, students were expected to explain the advantages and disadvantages of PCA in the context of this analysis. Answers often discussed multiple reasonable advantages and disadvantages of PCA but did not explain why these were relevant given the data, modelling methods and problem to be solved.

Some students focused on advantages and disadvantages of PCA relative to other dimension reductions methods rather than advantages and disadvantages of dimensionality reduction in general. Credit was given for these answers if they were properly explained and justified given the problem context.



Some students struggled to justify the number of components selected beyond recognising that fewer components would explain less variance. The best answers considered the potential impact of their component number choice on the clustering exercise, and ultimately the impact on their share price prediction.

3c&d - Clustering

The strongest answers considered different features that could be used in the clustering such as the retweet flag and username.

Students who did not reduce the dimension of the text embedding sometimes had technical challenges with this question, since it took a long time to run K-means for a range of cluster sizes on high dimensional data.

Weaker answers fit K-means on only a small number of components, sometimes without considering or evaluating the loss in information of this approach. Weaker answers also made dubious cluster size selections, identifying minor inflections in the elbow plots as kinks or elbows. In contrast, the best answers considered a wider range of cluster sizes if no clear elbow was observable or at least called out this limitation and used some other heuristic to choose a sensible number of clusters.

Most students included detailed manual validation by creating word clouds and/or word frequency lists for each cluster and attempting to profile each cluster. However, interpretations of these clusters seemed tenuous in some cases. The best answers linked the manual validations to the industry context and considered the problem context (share price prediction) in their evaluation.

Internal validation was less consistently included with many students stopping at the WCSS by cluster size created in 3c. A single internal validation statistic without any kind of reference or comparison was not given any credit.



Question 4 - Classification

Markers were generally impressed with student's modelling ability given many of them had only developed Python coding skills while studying this subject.

In defining a suitable response variable, the best answers considered various options and then recommended the best option, given the problem context. In contrast, weaker answers either just calculated a response variable without discussing the reasoning for it or discussed their reasoning but did not consider alternative options.

The best answers to this question explained each step that was being undertaken, including explaining which model diagnostics were being used and why. These answers used an iterative approach to model building, whereby each iteration had a purpose, being to improve the model based on model diagnostics produced in the previous iteration. For example, a good answer might have identified that model overfitting was occurring and then implemented measures such as early stopping or tree pruning/complexity reduction, to reduce overfitting. In contrast, weaker answers mechanically tried different combinations of model architectures and hyperparameter selections, without being guided by model outputs in the previous step.

The best answers also compared their chosen model's output to that from one or more benchmark models, such as a random guess, linear regression, or market-level investment returns over the relevant period.

Many students understood how to produce training, validation, and test metrics but weaker answers did not demonstrate an ability to interpret these metrics. Weak answers focussed on training metrics, not understanding the importance of looking at validation and test metrics to avoid overfitting to the training data. The weakest answers provided little or no commentary on the output of any calculations.

Some students focussed only on accuracy as an optimisation metric, not recognising that this is particularly problematic for data with unbalanced classes.

Many students did not look at the confusion matrix output to see that their models were almost always predicting the same class, or never predicted one of the classes.



Question 5 – Risks and implementation consideration

In general, 5b (implementation considerations) was answered better than 5a (risks).

The best answers to this question were very specific to the problem context. These answers considered risks that may not be identified prior to implementation, such as Tweet text changing over time. In contrast, weaker answers either discussed generic risks such as 'model error', 'data error', and 'governance error' and/or identified risks that could easily be tested prior to implementation, such as how long the model takes to run.

The best answers to this question explained how each risk might manifest and the potential adverse impacts of each risk within this problem context. Quite a few students described concepts such as model drift and spurious learning, but only the better responses described how these issues might come about in the context of this exercise. Also, the best answers talked about what the impact of these issues would be, such as negative impacts on the fund's profits or reputation.

Most students reflected on different forms of implementation, such as decision support tools compared to automated decision engines.

Question 6 – Video executive summary

Most videos were of good quality with an appropriate level of detail.

The best videos:

- had clear structure so that the audience could follow what they were talking about;
- provided audience-appropriate descriptions of methodology and/or modelling outcomes that were free of too much detail or technical jargon; and
- presented recommendations or next steps that made sense in the context of their conclusion from the analysis.

In contrast, some students presented modelling outcomes that showed limited predictive power of their models relative to a random guess, but still went on to discuss a detailed road map of further developing the model and rolling it out across the business.



Data Analytics Applications

Semester 2 2021 Assignment Feedback

Several students spoke very quickly to be able to cover more detail. This sometimes made their communication difficult to follow.