

Advanced (Business) Data Analytics

ASSIGNMENT 2

Summary

- Type: Project report
- Learning Objectives Assessed: 1, 2, 3, 4, 5
- Due Date: 21st May 2020 11 AM
- Deliverables:
 - Written report submitted via Turnitin,
 - RapidMiner process, and
 - Oral presentation and Q&A
- Weight: 70%

This assignment is an individual assignment. The aim is to provide experience in the steps involved with text processing and creating, evaluating, improving models, and finally presenting and interpreting the model in a business report. You are strongly encouraged to commence this assignment by the end of week 7 of the semester, and you should progress thoughtfully through the steps. Hasty decisions made early in the design process may result in much more work later.

Feel free to discuss concepts and ideas with peers, but remember your submission must be your work. Be careful not to allow anyone to copy your work.

Specification

Online reviews help consumers reduce uncertainty and risks faced in purchase decision-making by providing information about products and services. However, the overwhelming amount of data continually being produced in online review platforms introduces a challenge for customers to read and judge the reviews. Reviews may appear on the company's websites, social media, or review platforms. Companies are aware of the impacts of online reviews on consumer attitudes and behaviors. Consumers also expect reviews to be **HELPFUL** to assist them in making more informed purchasing decisions.

In Amazon.com, reviews on products are available, and customers can read the reviews before their purchase (see Figure 1).

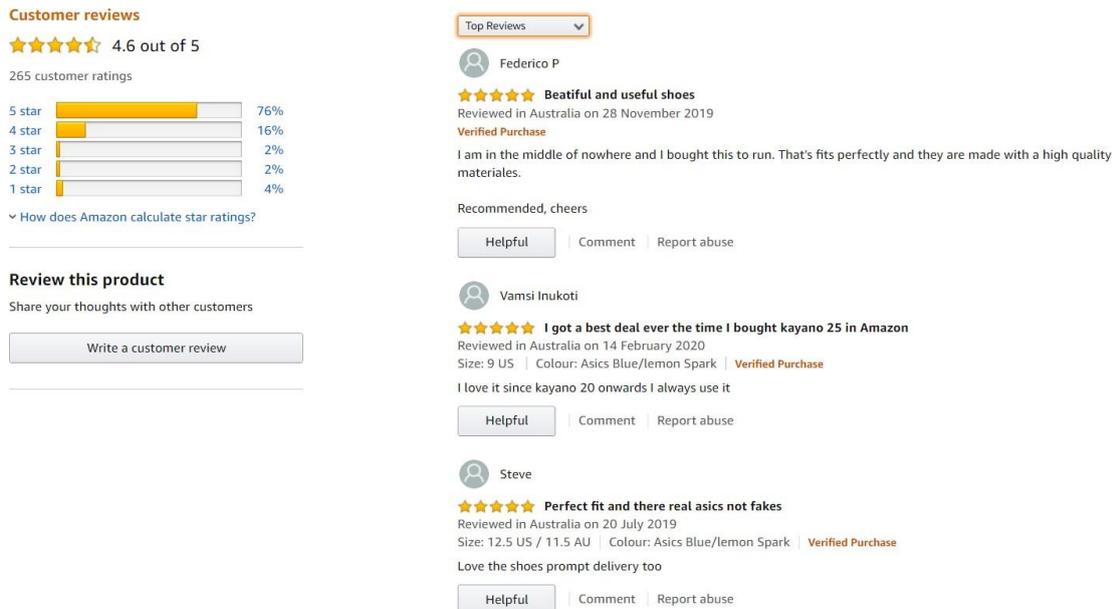


Figure 1 Sample reviews on a product in Amazon.com

When customers read a review, they can mention if they have been **HELPFUL** (see Figure 2).



Figure 2 Voting if a review is HELPFUL in Amazon.com

Before starting the assignment and going through the rest of the assignment specification, you need to read some reviews on Amazon.com and familiarize yourself with how the review platform works.

Business problem: Some online reviews are not read by any customers. Amazon.com sort reviews either based on recency or top score (see Figure 3).



Figure 3 Sorting reviews in Amazon.com

Top reviews are those that are read by the customers and have received more **HELPFUL** votes.

Based on this sorting mechanism in Amazon.com, if there is a fair and in-depth review that is not recent and has not been scored by the readers, the review won't be very visible on the platform. And if the product has hundreds of reviews, customers likely miss this fair and in-depth review, simply because they do not have enough time to read all the reviews, and the review platform sorts the reviews based on either recency or top score.

Importance and motivation: Your developed model predicts if the reviews would be HELPFUL. To develop the model, you will use the HELPFUL values of existing reviews. This model can be used as an assistive tool in Amazon.com in several ways. For example, Amazon.com can add another metric for sorting reviews, named 'projected helpfulness'. The readers can then sort the reviews based on the predicted helpfulness value provided by your model. This way, Amazon.com assures that no valuable review is missed among hundreds of reviews on a product.

Dataset

A2 dataset consists of reviews on different products in Amazon.com. Reviews include product and user information, ratings, and a plain text review. Table 1 shows the regular attributes of the dataset.

Table 1 The regular attributes of review dataset from Amazon.com

Regular attribute	Description
Product_Score	Rating between 1 to 5 provided by the reviewer about the quality product on which the review has been written
Average_Product_Score	
Product_Category	The category of the product on which the review has been written
Reviewer_Helpfulness	An index showing how helpful the reviewer who has written the review has been
Reviewer_Activity	An index showing how active the reviewer who has written the review has been
Review_Order	The order or review among the written reviews on the product
Review_Count	Total number of reviews on the product on which the review is written
Review_Summary	A summary of the review written by the reviewer in the title of review
Review_Text	The body of the review

There are three ID attributes, namely, Review_ID, Reviewer_ID, and Product_ID. You don't need to use these attributes in model building. Table 2 shows the label attributes of the dataset.

Table 2 The target attributes of review dataset from Amazon.com

Special attribute	Description
Total_Reads	Total number of people who have read the review You need to use this attribute as a label one in developing the prediction model.
Helpfulness_Label	A label showing if the review has been helpful You need to use this attribute as a label one in developing the classification model.

Important note: You should not use the above special attributes as regular attributes. You won't get any marks for developing a model that uses either `Total_Reads` or `Helpfulness_Label` as a regular attribute.

Your task is to

- develop one model to predict the `Total_Reads` of the newly posted reviews and
- develop another model to assign a value to the `Helpfulness_Label` attribute of newly posted reviews.

Your approach should involve the following tasks:

- Data exploration and preparation
 - Structured data preparation. You need to discuss the basic statistics of the available data, whether or not it is balanced.
 - Unstructured data (text) preparation. It should at least include TF-IDF & SVD generation and sentiment analysis. You may also consider topic modeling.
- Model building (classification), improvement and evaluation
 - You can use all the classification techniques and ensemble methods to develop a classification model to assign a `Helpfulness_Label` attribute of newly posted reviews.
- Model building (prediction), and evaluation
 - You can use NN to develop a model for predicting `Total_Reads`. You can use the same input variable in both prediction and classification models.

Deliverables

You can use RapidMiner for this assignment. But if you would like to challenge yourself and improve your Python skills, you are more than welcome to use Python. **The choice of RapidMiner or Python won't have an impact on your mark.**

However, please note that **using Python can be challenging**, and you will need to do your assignment based on your research too. It would be an invaluable experience, which can be a great addition to your CV as a project done using Python.

Apart from your regular consultation, which is about the assignment, you can book **ONLY 30 MINUTES** for troubleshooting your Python issues if you decided to use Python. If you decided to use RapidMiner, you have a simpler task as you are already familiar with the platform, and you can focus more on model building than the software.

Your reports should include the following parts:

- Executive summary
- Data exploration and preparation

- Classification, improvement, and evaluation
- Prediction and its evaluation
- Conclusion

It is up to you to decide what proportion of your report goes to each part. You may include tables, charts or screenshots of your analysis and models. At the end of your analysis, your RapidMiner process (or Python script) should be uploaded via dropbox along with your report.

The consistency of your rmp file (or Python script) with the results in your report will be checked. You don't need to provide the screenshots of your RapidMiner process (or Python script), as they can be observed from your submitted file. Consider the following points for designing your process:

- You can create up to two .rmp files (or Python script) for the assignment.
 - The first process can be for text preparation, and
 - The second one can be for model building
- You should not modify your assignment data file before importing it in RapidMiner (or Python). **If you decided to submit two processes. The output of your first process should be the input of your second process.**
- All of your analysis should be done **after importing** your assignment data in RapidMiner (or Python). Otherwise, your marker can not run your process.

Important note: the nature of this assignment is different from assignment 1. Your developed model will assist in algorithmic decision-making in Amazon.com. While in your first assignment, the explainability of your model was also important, in this assignment, you need to achieve higher accuracy. As the use of text data would be at the heart of your model, it won't be possible to explain how the model classifies or predicts. Therefore, you need to try your best to improve your model in prediction and classification.

In algorithmic decision-making for these types of problems, we are not looking for a high level of accuracy. Even 60% of accuracy, for example, can be helpful to Amazon.com

Please see the rubric and data description in the appendices.

Formatting and professionalism

The project report is to be written to a professional standard. This requires a formal writing style – do not use dot points - and adopt a professional tone. Given the report's nature, you may choose to write this essay in the first person. The report must be consistent with the University's policies on academic integrity, plagiarism, and consequences, as noted below. The report should be typed (in Times Roman 12-point font or larger, single-spaced), and the Word Count should be 2000 words (+/- 10%) in total length. The Word Count excludes the title page, tables, footnotes, and references (if required). The word limit must be observed, or the assessment will be affected, as noted in the rubric. No appendices are to be provided.

Submission

To be done through Blackboard Assignment Submission and Turnitin, as indicated in Learn.UQ. Acceptable submission formats are Microsoft Word and PDF formats for the reports and .rmp (or Python script) for the process. The files **MUST** be named in the format of :

- AdvancedBDAnalytics_StudentLastName_StudentID.pdf (or a .docx or .doc extension).

If your ID is 41724593 and your surname is Mory, the name of your files would be:

- AdvancedBDAnalytics_Mory_41724593.pdf. The written assignment file should not be zipped.

Plagiarism

It is understandable that students talk with each other regularly and discuss problems and potential solutions. However, it is expected that the submitted assignment is a unique document – all parts of the assignment are to be completed solely by the individual student. In cases where an assignment is perceived to not be a unique work, a loss of marks and other implications can result. For further information about academic integrity, plagiarism, and consequences, please visit:

<http://ppl.app.uq.edu.au/content/3.60.04-student-integrity-and-misconduct>.

Administrative Requirements

Submission Date

11 AM 21st May 2020

For each calendar day (i.e., including Saturdays and Sundays) or part thereof after the submission deadline, a penalty of 5% of the total possible assignment marks will be deducted until the assignment is submitted.

Consultation sessions

To ensure that an equal and sufficient amount of time is allocated for every student who attends consultation sessions regarding the practical aspects of advanced data analytics, the average consultation time (during busy consultation times) will be limited to 15 minutes per student. The main aim of this restriction during busy periods is to ensure student equity and minimize waiting time. However, in circumstances where no other students are waiting, longer consultation times will be provided.

You can join the consultation sessions without booking, but if you would like to secure your spot, please make a booking in the 'Booking an appointment' page.

Deadline extensions

An extension to the assignment deadline will only be considered for legitimate reasons and with supporting documentation. A request for an extension is assessed by the Assessment, Examinations & Misconducts Coordinator. You may discuss your situation with your course coordinator, but you still need to make a formal extension request using the form identified on the Electronic Course Profile for this course. Extensions will not be granted where the School is not satisfied; you took reasonable measures to avoid the circumstances that contributed to you not submitting by the due date. The following are not grounds for an extension:

- holiday arrangements (including overseas travel)

- misreading a due date
- social and leisure events
- moving house
- the pressure of work/competing deadlines
- computer issues

Please refer to the Electronic Course Profile for this course for more detail.

Marking Rubric

Your Turnitin report and RapidMiner process are worth 40 marks.

	Very poor	Below Expectations	Meets Expectations	Good	Very Good	Outstanding
Executive summary – 2 marks	Provides no evidence of summarizing the project at a managerial level.	Provides little evidence of summarizing the project at a managerial level.	Demonstrates satisfactory summary of the project at a managerial level.	Demonstrates good summary of the project at a managerial level.	Demonstrates a very good summary of the project at a managerial level.	Demonstrates exemplary summary of the project at a managerial level.
Structure data exploration and preparation - 4 marks	Provides no evidence of preparing, cleaning & transforming structured data.	Provides little evidence of preparing, cleaning & transforming structured data.	Demonstrates satisfactory evidence of preparing, cleaning & transforming structured data.	Demonstrates good evidence of preparing, cleaning & transforming structured data.	Demonstrates a very good evidence of preparing, cleaning & transforming structured data.	Demonstrates exemplary evidence of preparing, cleaning & transforming structured data.
Unstructured data preparing - 8 marks	Provides no evidence of preparing, cleaning & transforming unstructured data.	Provides little evidence of preparing, cleaning & transforming unstructured data.	Demonstrates satisfactory evidence of preparing, cleaning & transforming unstructured data.	Demonstrates good evidence of preparing, cleaning & transforming unstructured data.	Demonstrates a very good evidence of preparing, cleaning & transforming unstructured data.	Demonstrates exemplary evidence of preparing, cleaning & transforming unstructured data.
Developing, improving and evaluating a classification model - 15 marks	Develops no or inadequate model that indicate an obvious lack of comprehension of available classification & ensemble techniques and their evaluation.	Develops few models that indicate very little comprehension of available classification & ensemble techniques and their evaluation.	Develops one or more acceptable creative model that indicates satisfactory comprehension of available classification & ensemble techniques and their evaluation.	Develops one or more good creative models that indicate comprehension of available classification & ensemble techniques and their evaluation.	Develops one or more very good creative models that indicate a deep comprehension of available classification & ensemble techniques and their evaluation.	Develops one or more exemplary creative models that indicate a deep comprehension of available classification & ensemble techniques and their evaluation.
Developing and evaluating a prediction model - 8 marks	Develops no or inadequate model that indicates an obvious lack of comprehension of prediction & its evaluation.	Develops a model that indicates very little comprehension of prediction & its evaluation.	Develops an acceptable creative model that indicates satisfactory comprehension of prediction & its evaluation.	Develops a good creative model that indicates comprehension of prediction & its evaluation.	Develops a very good creative model that indicates a deep comprehension of prediction & its evaluation.	Develops an exemplary creative model that indicates a deep comprehension of prediction & its evaluation.
Conclusion - 3 marks	The conclusion of the project is superficial , lacking any consideration of the strength and limitations of the project.	The conclusion of the project is partial , lacking consideration of the strength and limitations of the project.	The conclusion of the project is satisfactorily and considers the strength and limitations of the project.	The conclusion of the project fairly considers the strength and limitations of the project and provides fair recommendations for future works.	The conclusion of the project is thorough and includes reasonable consideration of the strength and limitations of the project and provides thorough recommendations for future works.	The conclusion of the project is insightful and includes quite thorough consideration of the strength and limitations of the project and provides exemplary recommendations for future works.

Oral presentation

Your presentation and the Q&A session are worth 30 marks. You can book a time for oral presentation from 25 May to 2 Jun though booking an appointment page on the blackboard. Your marker will set a Zoom meeting with you for the oral presentation. It will take a maximum of 15 minutes. You need to make sure you have your camera on when you join the meeting via Zoom for the oral presentation. The marker will record the meeting.

You don't need to prepare any slides for the presentation. You only need to open the process that you submit for the assignment in your RapidMiner and share your desktop with your marker. Then your maker will ask you 4 sets of questions shown in the below table.

Table 3 The rubric for the oral presentation

	A poor explanation of how the process works	An acceptable explanation of how the process works	An good explanation of how the process works and why they are chosen	An accurate explanation of why the operators are chosen and the other possibilities
Structure data exploration and preparation - 4 marks	A poor understanding of the operators used for structured data preparation.	A thorough understanding of the operators used for structured data preparation.	A thorough understanding of the operators used for structured data preparation, and why they are chosen.	Could explain why the operators are used for structured data preparation, and also a sound understanding of some other operators that could have possibly been used for structured data preparation.
Unstructured data preparing - 7 marks	A poor understanding of the operators used for unstructured data preparation.	A thorough understanding of the operators used for unstructured data preparation.	A thorough understanding of the operators used for unstructured data preparation, and why they are chosen.	Could explain why the operators used for unstructured data preparation, and also a sound understanding of some other operators that could have possibly been used for unstructured data preparation.
Developing, improving and evaluating a classification model – 12 marks	A poor understanding of the operators used for classification, its improvement & evaluation.	A thorough understanding of the operators used for classification, its improvement & evaluation.	A thorough understanding of the operators used for classification, its improvement & evaluation, and why they are chosen.	Could explain why the operators used for classification, its improvement & evaluation, and also a sound understanding of some other operators that could have possibly been used for classification, its improvement & evaluation.
Developing and evaluating a prediction model - 7 marks	A poor understanding of the operators used for prediction and its evaluation.	A thorough understanding of the operators used for prediction and its evaluation.	A thorough understanding of the operators used for prediction and its evaluation, and why they are chosen.	Could explain why the operators used for prediction and its evaluation, and also a sound understanding of some other operators that could have possibly been used for prediction and its evaluation

Bonus points

You can use cluster analysis before model building. Using this approach, you can then develop different models in each cluster (5 bonus marks). Please note that if you chose to apply the above approach to receive bonus points, you need the research on your own. Also, you need to demonstrate that using cluster analysis has helped in improving the accuracy of your prediction and classification.