Honor code

By submitting this assignment, I affirm the following:

1. All work presented in this assignment is my own. I have not collaborated with others or copied work from any unauthorized source.
2. If I used AI tools like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
3. I understand the submitted code and can explain my work if asked.

I declare that I have read, understood, and agree to abide this honor code.

Name: Kevin Vu

Student number: 713837tv

Date: 02/12/2024

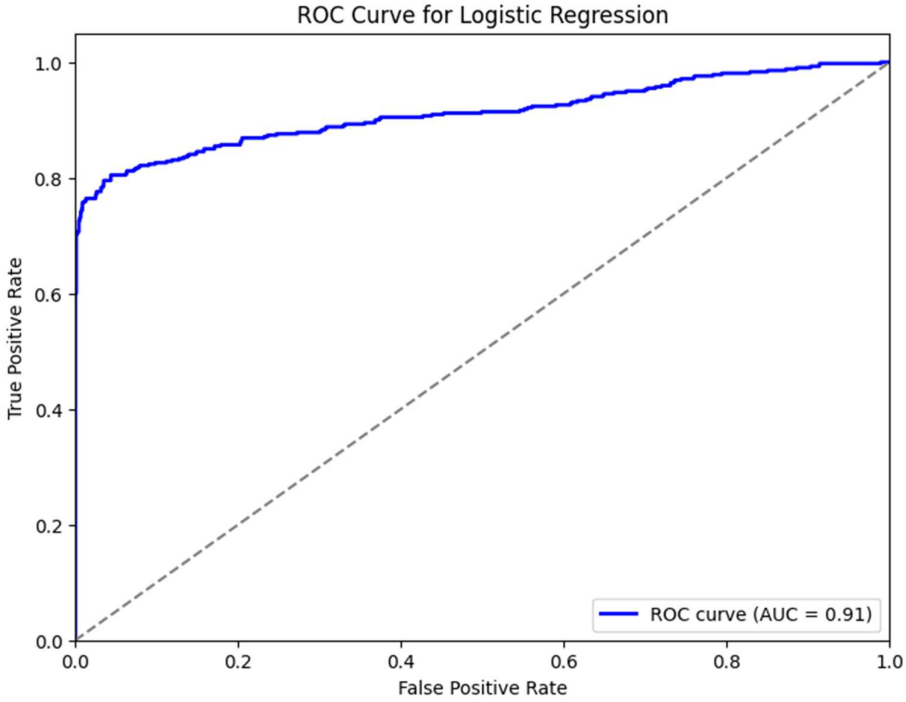
1a. Confusion matrix for threshold 0.2:

		Predicted	
Actual		1	0
	1	365	55
	0	195	723

1b. Confusion matrix for threshold 0.8:

		Predicted	
Actual		1	0
	1	273	147
	0	1	917

1c. Explanation of AUC:



An **AUC of 0.91** means that the model has a **91% chance** of correctly distinguishing between a randomly chosen positive instance and a randomly chosen negative instance.

2a. Confusion matrix for threshold 0.5:

		Predicted	
Actual		1	0
	1	276	144
	0	1	917

2b. Accuracy with explanation: 0.8916
Interpretation: 89.16% of the predictions made by the model are correct, which means the model performs well overall.

2c. Precision with explanation: 0.9964
Interpretation: When the model predicts the positive class (1), 99.64% of the time it is correct. The model has a very high precision, which means it rarely makes false positive errors.

2d. Sensitivity with explanation: 0.6571
Interpretation: The model correctly identifies 65.71% of all actual positive cases. While this is a good rate, there's still room for improvement in capturing more positive cases.

2e. Specificity with explanation: 0.9989
Interpretation: The model correctly identifies 99.89% of the actual negative cases. This is a great specificity rate, meaning the model does a good job of avoiding false positives.

2f. True positive rate with explanation: 0.6571
Interpretation: The true positive rate is the same as sensitivity, which means the model identifies 65.71% of all actual positive cases.

2g. False positive rate with explanation: 0.0011
Interpretation: The model incorrectly classifies 0.1% of the actual negative cases as positive. This is a very low false positive rate, indicating that the model is very accurate at rejecting negatives.

3a. Euclidean distance between customers 245 and 431:0.0000

3b. Manhattan distance between customers 82 and 197: 0.0000

3c. Centroid of the first 50 customers:

fiction	0.94
non_fiction	0.94
childrens_book	0.80
self_help	0.90
mystery	0.92

4a. Categories with highest co-occurrence: ('fiction', 'mystery') with a value of 461

4b. Categories with lowest co-occurrence: ('non_fiction', 'childrens_book') with a value of 388

5. Size of each cluster:
Cluster Sizes:

cluster	
0	102
1	100
4	100
3	100
2	98

6a. Support of {fiction}: 0.9400

6b. Support of {non_fiction}: 0.8940

6c. Support of {fiction, self_help}: 0.8400

7a. Confidence of {fiction} → {mystery}: 0.9809

7b. Confidence of {non_fiction} → {self_help}: 0.9687

7c. Confidence of {fiction, self_help} → {childrens_books}: 0.9119

8a. Lift of {fiction, self_help} → {childrens_books}: 1.1399

8b. Lift of {fiction} → {non_fiction}: 1.0020

8c. Lift of {non_fiction} → {self_help}: 1.0884

9a. Explanation of support of {fiction, self_help}: **0.8400**
This means that **84%** of the transactions contain both **fiction** and **self_help** books. In other words, these two types of books are commonly purchased together.

9b. Explanation of confidence of {fiction, self_help} → {childrens_books}: 0.9119
This means that in **91.19%** of the transactions where both **fiction** and **self_help** are purchased, **childrens_books** are also bought. It shows a strong likelihood of buying **childrens_books** when **fiction** and **self_help** are bought together.

9c. Explanation of lift of fiction, self_help} → {childrens_books}: **1.1399**
This indicates that customers who buy both **fiction** and **self_help** books are approximately **1.14 times more likely** to also buy **childrens_books** than if there were no relationship between the items. Since the lift is greater than 1, it shows a **positive association** between these items.