CCT College Dublin

Assessment Cover Page

Module Title:	Strategic Thinking
Assessment Title:	Repeat CA
Lecturer Name:	James Garza
Student Full Name:	Natalia Gomes Fernandes
Student Number:	*2020350*
Assessment Due Date:	28/07/2024
Date of Submission:	28/07/2024

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

Customer Segmentation in Online Retail: A Machine Learning Approach

Introduction

In the competitive world of e-commerce and retail, understanding customer behaviour is essential for providing a better customer experience. One effective method to achieve this is through customer segmentation, which involves dividing customers into groups based on their behaviour patterns. By implementing this strategy, businesses can develop targeted strategies and personalised customer experiences leading to increased customer satisfaction and loyalty (The importance of customer segmentation for businesses, 2023)

This report aims to demonstrate the application of machine learning techniques for customer segmentation in an online retail business. The analysis utilises the "Online Retail" dataset from UCI (Online Retail, n.d.), which contains all transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retailer. The primary objectives of this study are:

- To group similar customers based on their purchasing habits, and
- To predict the group to which new customers might belong.

In this report, we will explain the methodologies applied, including the specific algorithms, models, and libraries used in our analysis. Subsequently, we will present our findings derived from the data exploration and modelling process. The report will conclude with a comprehensive summary of our discoveries and a discussion of their implications for the rental business.

Business Description

Hypothesis:

Our hypothesis is that effective customer segmentation can be achieved by analysing customers' buying habits and that these segments can be used to predict the classification of new customers based on their initial purchasing patterns.

General Goal:

The general goal of this project is to create distinct customer segments using the Online Retail database, focusing on existing customers with established purchasing histories. Furthermore, we aim to develop a predictive model capable of assigning new customers to these predefined clusters based on their early purchasing behaviour. This segmentation and prediction strategy will enable the business to tailor its marketing messages more effectively, ensuring that each customer receives communications that are relevant to their specific needs and preferences.

Technologies Used

Models and machine learning algorithms

Our analysis employs a combination of unsupervised and supervised machine learning techniques:

1. Unsupervised Learning: K-means clustering algorithm for grouping similar customers based on their purchasing habits.

2. Supervised Learning: Decision Tree and Random Forest algorithms for predicting the cluster assignment of new customers based on their behaviour.

Libraries

The analysis leverages several Python libraries for data preparation, visualisation, and machine learning modelling:

- NumPy: For efficient numerical computations

- Pandas: For data manipulation and analysis

- Matplotlib: For creating static, animated, and interactive visualisations

- Seaborn: For statistical data visualisation

- SciPy: For scientific and technical computing

- Scikit-learn: For machine learning algorithms and model evaluation

Accomplishment

Data

The analysis utilises the Online Retail dataset from UCI (Online Retail, n.d.) .), which provides comprehensive information about customer purchases from an online store. The dataset comprises 8 features (columns) and 541,909 observations (rows).

The 8 features are:

1. InvoiceNo: A unique 6-digit integral number assigned to each transaction

2. StockCode: A unique 5-digit integral number assigned to each distinct product

3. Description: The name of the product

4. Quantity: The number of units of each product per transaction

5. InvoiceDate: The date and time when each transaction was generated

6. UnitPrice: The price per unit of the product in sterling

7. CustomerID: A unique 5-digit integral number assigned to each customer

8. Country: The name of the country where each customer resides

In the figure below we can visualise the first 10 rows of the Online Retail dataset.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

Figure 1- First 10 rows from the Dataset

Attributes

Data attributes are the columns in the dataset used to build, test, or score a model (Oracle Machine Learning, n.d.). In this project, the attributes used for training the machine learning models are derived from the original features: InvoiceNo, Total (calculated), and Recency (calculated).

Data Preparation and pre-processing

Data preparation is crucial for ensuring the accuracy and consistency of the analysis (Stedman, n.d.). Before applying descriptive statistics and data visualisation techniques, several preprocessing steps were taken:

1. Checking for missing values: 1,454 missing values were identified in the Description column, and 135,080 in the CustomerID column.

2. Identifying duplicates: 5,225 duplicated entries were found in the entire dataset.

3. Handling missing data: Records with missing values were removed to ensure data integrity.

After the initial cleaning process, basic statistics were computed for the dataset:

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	401604.000000	401604	401604.000000	401604.000000
mean	12.183273	2011-07-10 12:08:23.848567552	3.474064	15281.160818
min	-80995.000000	2010-12-01 08:26:00	0.000000	12346.000000
25%	2.000000	2011-04-06 15:02:00	1.250000	13939.000000
50%	5.000000	2011-07-29 15:40:00	1.950000	15145.000000
75%	12.000000	2011-10-20 11:58:30	3.750000	16784.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	250.283037	NaN	69.764035	1714.006089

Figure 2- first descriptive statistics

The initial analysis revealed potential data quality issues:

- Minimum quantity order of -80,995, indicating the presence of return transactions or data errors

- Maximum quantity and UnitPrice values suggesting the presence of outliers

- Minimum UnitPrice of 0.00, potentially representing cancelled orders or data entry errors

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
222681	C556445	M	Manual	-1	2011-06-10 15:31:00	38970.0	15098.0	United Kingdom

Figure 3- Negative quantity UnitPrice

Further investigation revealed 8,872 entries with negative quantities and unit prices very high, confirming the presence of outliers in the dataset.

To address these issues, the Interquartile Range (IQR) method was employed to identify outliers on both the Quantity and UnitPrice columns. This process identified 26,646 outliers in Quantity and 35,802 in UnitPrice.

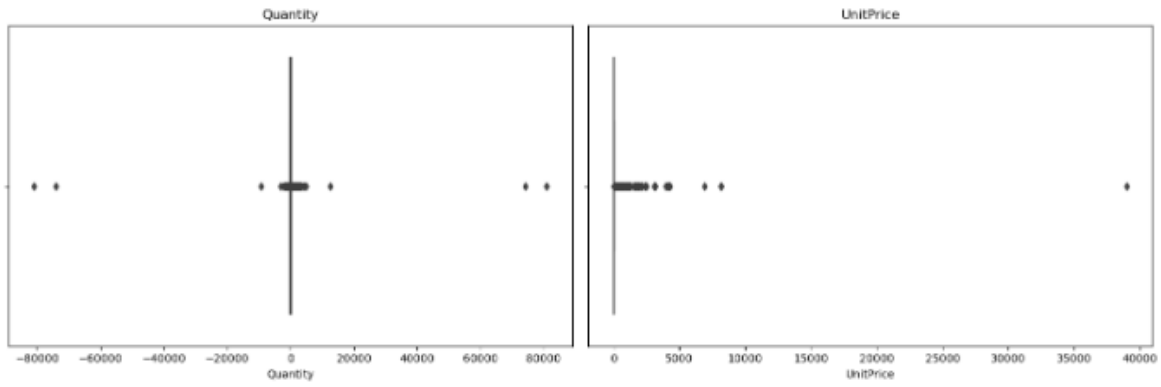

Figure 4- Outliers in Quantity and UnitPrice

After careful consideration of the product prices and purchase quantities, thresholds were established to remove extreme outliers:

- Quantity: Values between 0 and 1,000 were deleted

- UnitPrice: Maximum value set to 6,000

After outlier removal, the dataset was reduced to 241,387 rows.

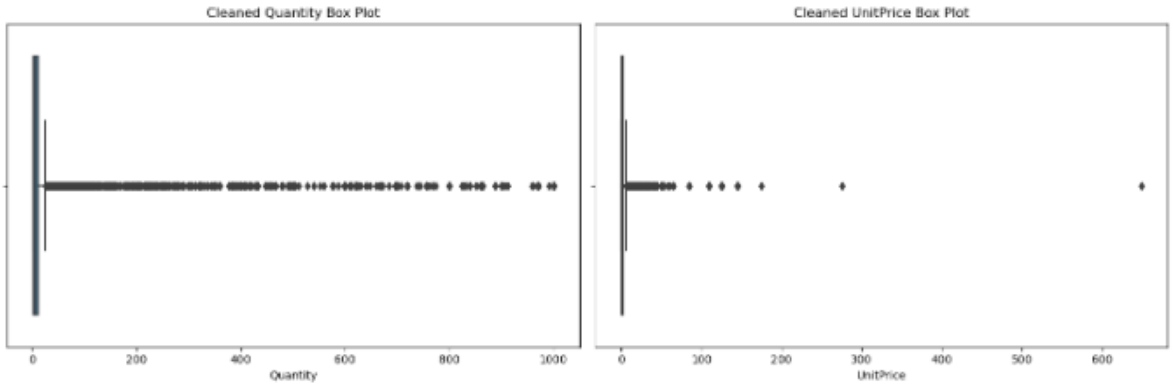

Figure 5- Cleaned Outliers - Quantity and UnitPrice

Descriptive statistics and Data Visualisation

Descriptive statistics provide a concise summary of the data, typically including measures such as mean, median, standard deviation, minimum, and maximum values (What is Descriptive Statistics: Definition, Types, Applications, and Examples, 2024).

After cleaning the dataset, the descriptive statistics for the numerical features were as follows:

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	241387.000000	241387	241387.000000	241387.000000
mean	11.777726	2011-07-07 08:36:02.917638912	3.337553	15191.445331
min	2.000000	2010-12-01 08:26:00	1.000000	12347.000000
25%	3.000000	2011-04-04 11:18:00	1.650000	13767.000000
50%	6.000000	2011-07-25 12:22:00	2.100000	15039.000000
75%	12.000000	2011-10-17 10:43:00	3.950000	16713.000000
max	1000.000000	2011-12-09 12:50:00	649.500000	18287.000000
std	28.390535	NaN	3.665832	1715.195064

Figure 6- Descriptive statistics after clean data

The cleaned dataset now presents more reasonable minimum and maximum values for Quantity and UnitPrice, allowing for a more reliable analysis.

Data visualisation techniques, such as histograms and scatterplots, were employed to gain deeper insights into complex data relationships (What is data visualisation?, n.d.).

A histogram was used to visualise customer purchase frequencies:

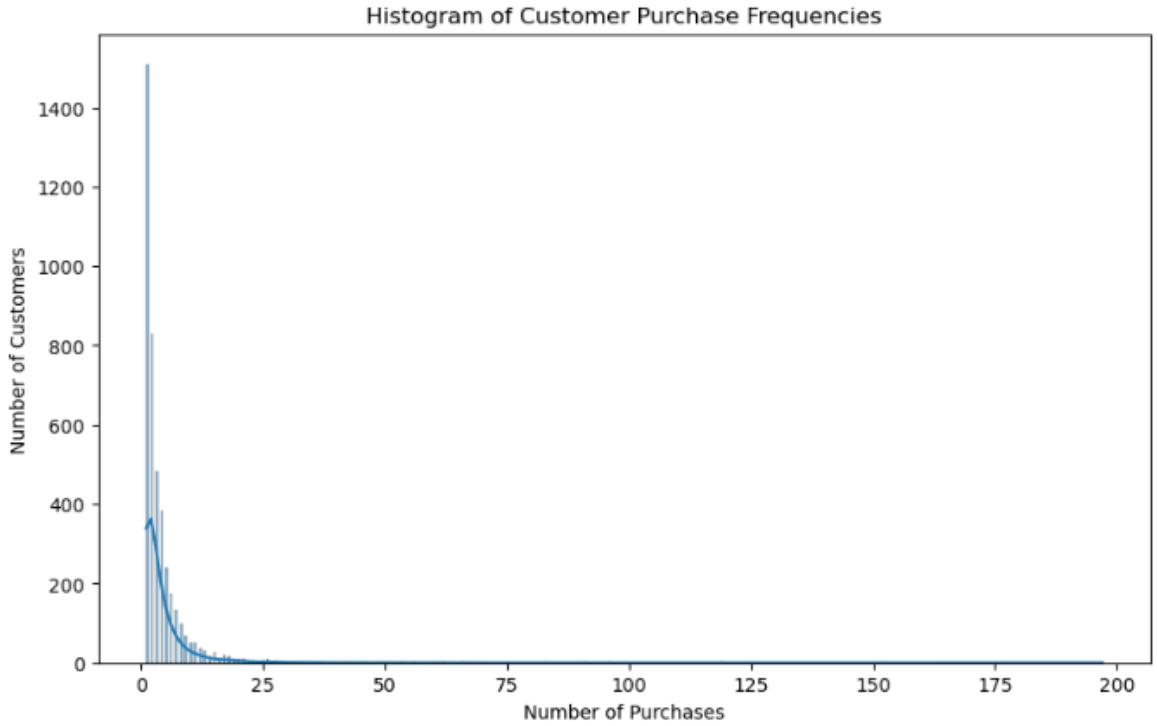

Figure 7- Customer Purchase Frequencies

Key insights from the histogram:

- Approximately 70% of customers made between 0 and 25 purchases

- Around 30% of customers made between 25 and 200 purchases

A scatterplot was created to examine the relationship between the customer's total amount spent and the number of orders, providing insights into customer lifetime value:

Figure 8 - Scatterplot Customer Total Amount vs Number of Orders

The scatterplot reveals:

- A positive correlation between the customer's total amount spent and the number of orders

- Some customers place a large number of orders but spend relatively little in total

- Other customers place few orders but spend substantial amounts

Finally, the relationship between Quantity and Unit Price was analysed:

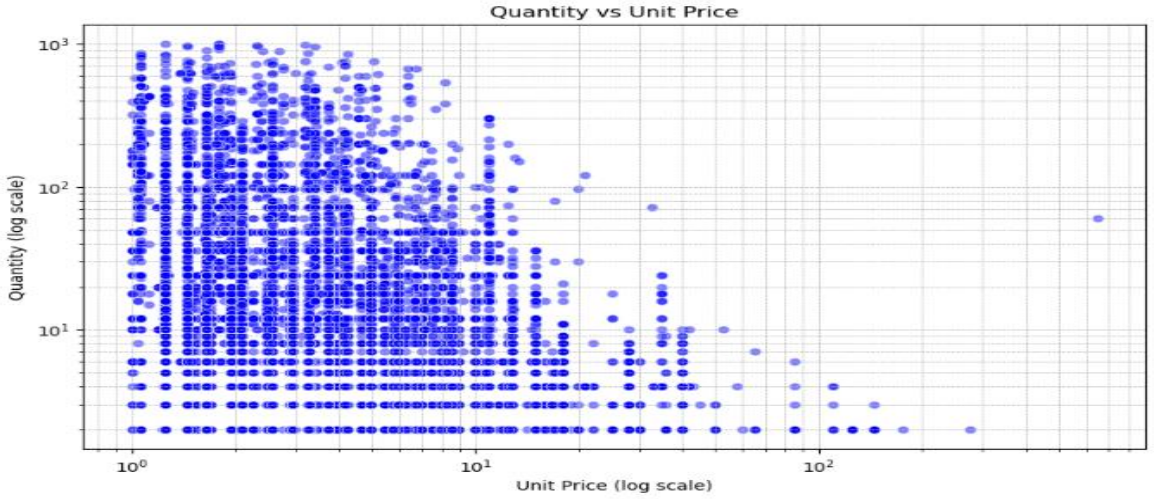

Figure 9 - Scatterplot Quantity vs Unit Price (log scale)

The Quantity vs Unit Price scatterplot on a logarithmic scale illustrates:

- An inverse relationship between unit price and quantity, with higher quantities generally associated with lower prices

- Concentration of data points in the lower unit price range (10^0 to 10^1) and higher quantity range (10^1 to 10^3)

- Customers tend to purchase larger quantities of cheaper items more frequently

Machine Learning

Three distinct machine learning algorithms were applied in this analysis: K-Means clustering, Decision Tree, and Random Forest.

K-Means focused on customer segmentation, dividing customers into clusters based on their purchase behaviour, specifically considering the number of orders, total amount spent, and recency. Decision Tree and Random Forest algorithms were used to predict the cluster assignment of new customers, trained using the data frame with clusters created after application of the K-Means clustering algorithm.

K-Means

K-Means is an unsupervised learning algorithm that groups data points into sets based on their degree of similarity (K-means Clustering Algorithm: Applications, Types, & How Does It Work?, 2024).

In this project, K-Means was employed for customer segmentation, allocating customers to different clusters based on quantity purchased, total amount spent, and recency. A new data frame was created containing:

- CustomerID: Unique customer identifier

- Total: Total amount spent by each customer

- InvoiceNo: Total number of orders made by each customer

- Recency: Time since the customer's last order

After applying standardisation to these features, an elbow graph was created to determine the optimal number of clusters:

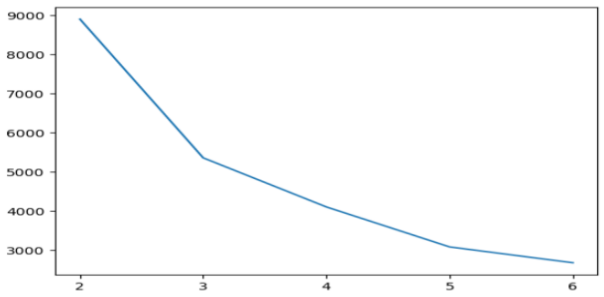

Figure 10- Elbow Graph K-Means

Additionally, the silhouette score technique was used to evaluate how similar an object is to its own cluster compared to other clusters (Samina, n.d.):

```
For N Cluster2 the silhouette scores is 0.5596699247727642
For N Cluster3 the silhouette scores is 0.5827685790230728
For N Cluster4 the silhouette scores is 0.6104238385151366
For N Cluster5 the silhouette scores is 0.5934469670175383
For N Cluster6 the silhouette scores is 0.589541264926804
```

Figure 11- Silhouette score

Based on the results of the elbow graph and silhouette score, the optimal number of clusters for the K-Means model was determined to be 3. Below we can visualise a cluster boxplot based on customer last seen (Recency):

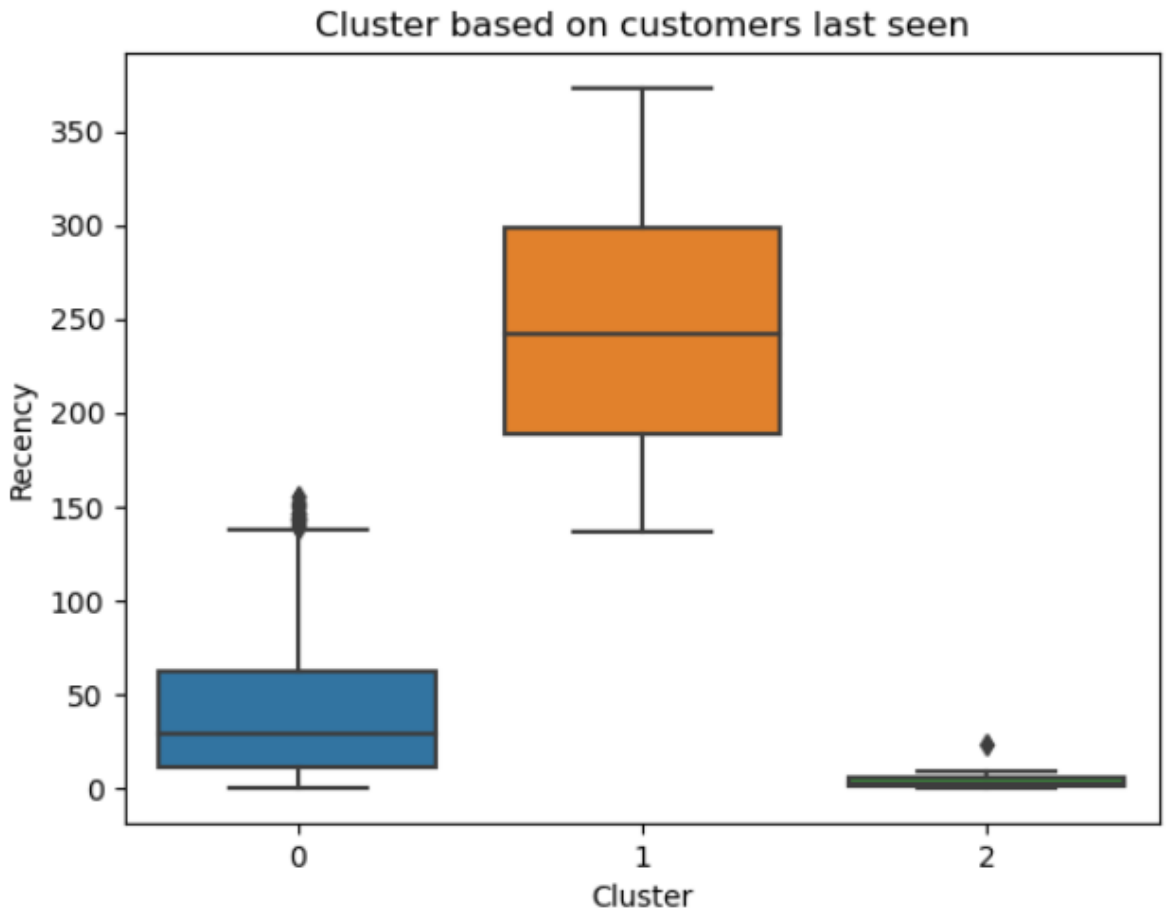

Figure 12- Cluster-based on Recency

The resulting clusters can be characterised as follows:

- **Cluster 0:** Customers with a recency of around 50 days

- **Cluster 1**: Customers with the least recent purchases

- **Cluster 2**: The most active customers

An overview of the three clusters based on CustomerID, InvoiceNo, Total, and Recency can be visualised on the figure below:

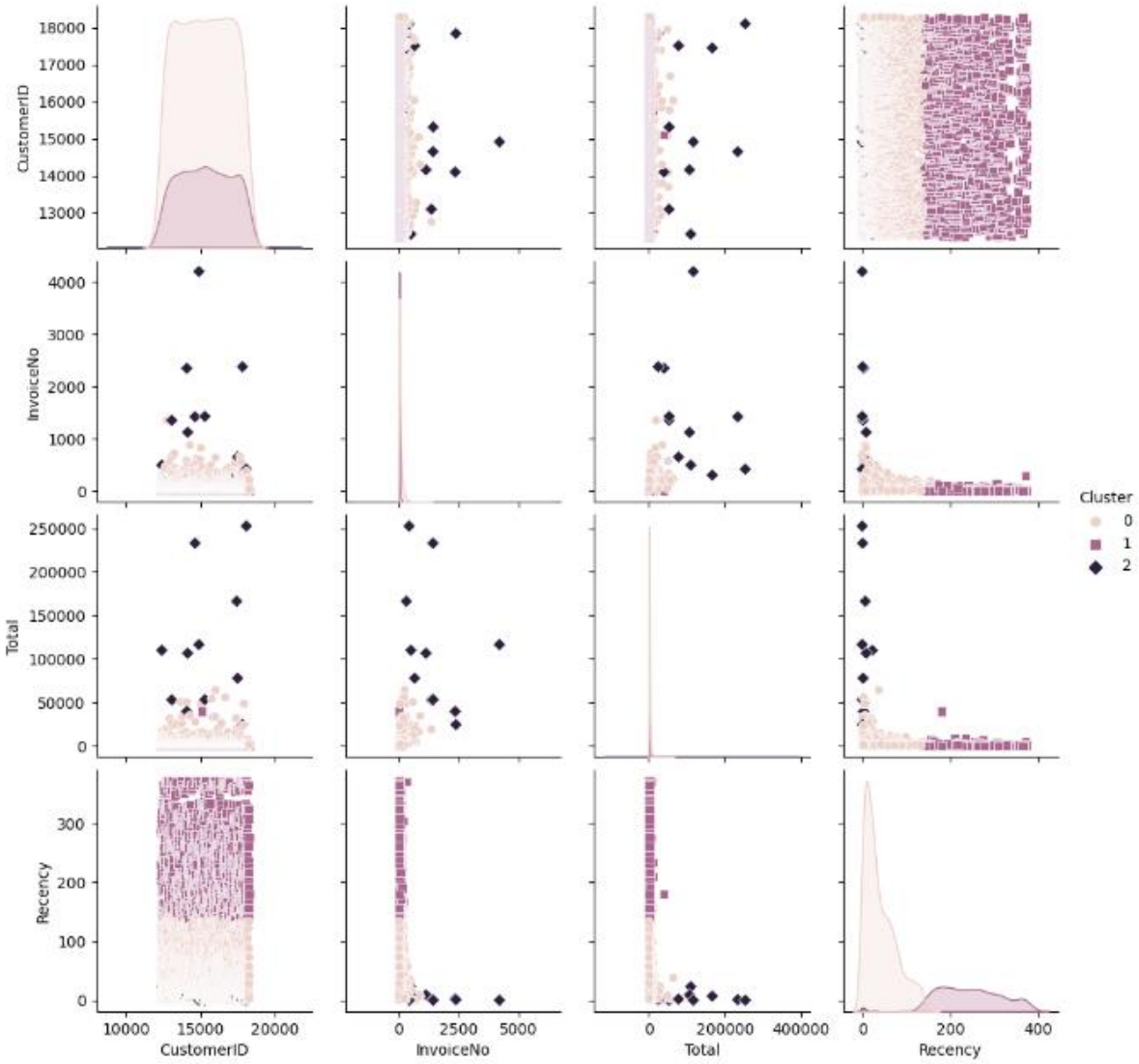

Figure 13- Cluster Overview

A 3D scatter plot clearly illustrates the segmentation when comparing the number of orders and the total amount spent:

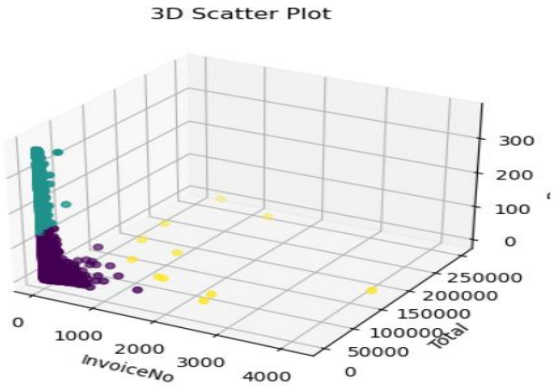

Figure 14- 3D Scatterplot Cluster

Decision Tree

A Decision Tree is a supervised machine-learning algorithm that predicts outcomes based on input data. It has a tree-like structure where each internal node tests an attribute, each branch corresponds to an attribute value, and each leaf node represents the final decision or prediction (Decision Tree in Machine Learning, 2024).

The primary objective of using a Decision Tree in this analysis was to train a model capable of predicting the cluster assignment for new customers based on the clustering performed by K-Means algorithm.

After splitting the data frame into training (75%) and testing sets (25%), the model achieved an accuracy of 99.81%.

The graphical representation of the trained decision tree model is shown below:

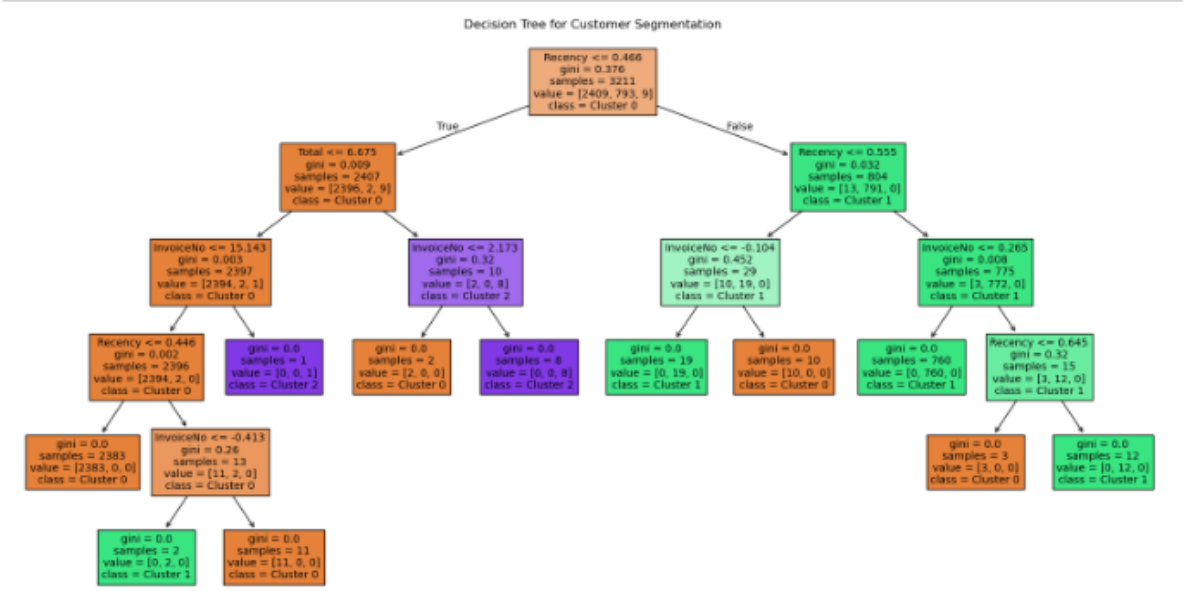

Figure 15- Decision Tree

To better understand the model's predictions, we can analyse the confusion matrix:

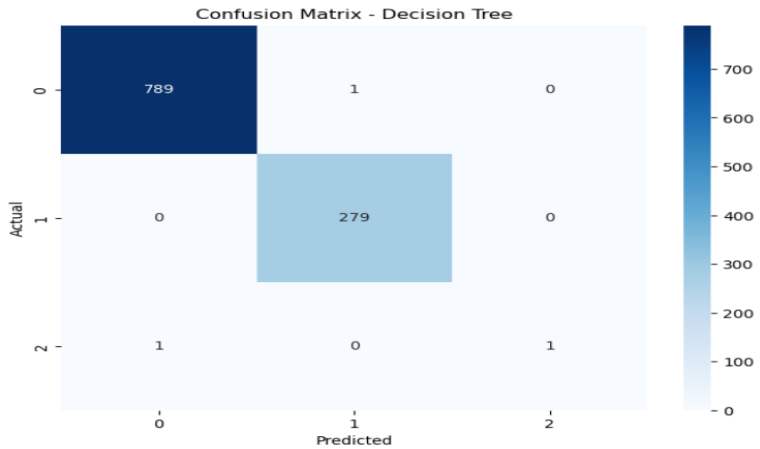

Figure 16- Confusion Matrix - Decision Tree

The confusion matrix reveals:

- The model correctly predicted 789 customers in cluster 0, with only one misclassification (predicted as cluster 1)

- 100% accuracy in predicting customers in Cluster 1

- 50% accuracy for cluster 2 predictions

Random Forest

Random Forest is another supervised machine learning algorithm that creates multiple Decision Trees during the training phase. Each tree is constructed using a random subset of the dataset and considers a random subset of features at each partition (Random Forest Algorithm in Machine Learning, 2024).

The Random Forest model was trained using the following parameters:

- Number of decision trees: 10

- Criterion for splitting nodes: "entropy"

The model achieved the same overall accuracy as the Decision Tree (99.81%). However, analysis of the confusion matrix reveals some differences:

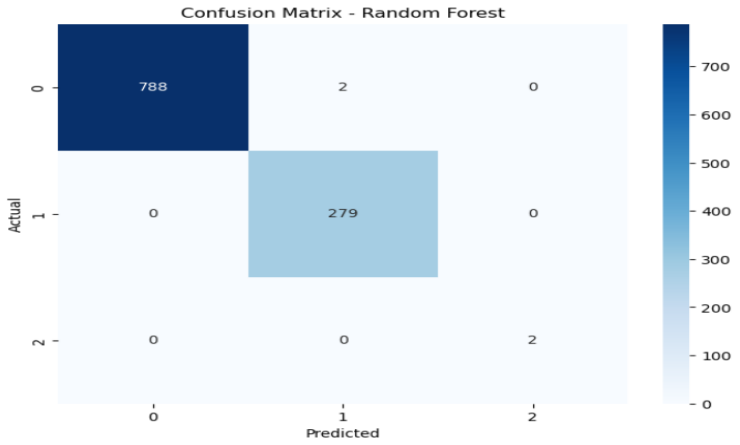

Figure 17- Confusion Matrix Random Forest

The Random Forest confusion matrix shows:

- 2 false positives for cluster 0

- 100% correct predictions for Cluster 1 and Cluster 2

Hyperparameter tuning and cross-validation

Due to the high accuracy of both models, hyperparameter tuning is not necessary. The well-separated nature of the clusters likely contributed to the models' ability to distinguish between them effectively, regardless of hyperparameter settings.

For cross-validation, a 5-fold cross-validation approach was employed. This method splits the data into 5 parts, trains the model in 4 parts, and tests it on the 5th part, repeating this process using a different part as the test set each time.

The cross-validation results were as follows:

- Decision Tree: 99.18% - 99.88%

- Random Forest: 99.30% - 99.88%

Comparing these results, we can conclude that the Random Forest model demonstrates more consistent performance compared to the Decision Tree model. The small score range of the Random Forest model indicates slightly higher stability and reliability in its predictions.

Conclusion

This comprehensive analysis of customer segmentation in online retail using machine learning techniques has given valuable insights and practical applications for business strategy. Through the implementation of K-means clustering, Decision Tree, and Random Forest algorithms, we have successfully segmented customers and developed predictive models for new customer classification.

The K-means clustering algorithm effectively divided the customer base into three distinct clusters based on their purchasing behaviour. Both the Decision Tree and Random Forest models demonstrated great accuracy in predicting the cluster assignment of the new customers, with a high 99.81% accuracy rate. The Random Forest model, in particular, showed slightly more consistent performance across different subsets of the data, as evidenced by the cross-validation results.

These findings can have several important implications for the online retail business:

1. Tailored Marketing: With customer segmentation, the retail business can develop more effective, personalised marketing campaigns that can reach each group's specific characteristics and preferences. (Kenton, 2024)

2. Inventory Management: Understanding the purchasing patterns of different customers can help inventory decisions by applying the method MRP (Materials Requirement Planning), ensuring that popular purchased items for each group are adequately stocked avoiding then the lack of products. (Hayes, 2024)

3. Customer Retention: By identifying the characteristics of the most valuable customer segment (Cluster 2), the retail shop can implement targeted strategies of customer retention to maintain and grow this group giving for example personalised service to improve their loyalty. (Hashemi-Pour, 2024)

4. Resource Allocation: The retail shop can optimise resource allocation by focusing more on high-value customer segments while also developing strategies to move customers from less active clusters (e.g., Cluster 1) to the most active (e.g., Cluster 2)

Further work could explore the addition of more features, such as seasonal purchasing patterns, to further refine customer segmentation.

In conclusion, this machine learning approach to customer segmentation offers a powerful tool for the retail business to enhance its understanding of its customer base and make data-driven decisions.

GitHub

The code for this project can be found at
https://github.com/nataliag248/NataliaGomes_StrategicThinkingHDip_RepeatCA

Presentation

The video for the presentation can be found at
https://drive.google.com/file/d/1SbYvMHmD_3N0Rhc2zCLYp5s4D_a3aaoY/view?usp=sharing

References

Decision Tree in Machine Learning. (2024, March 15). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/decision-tree-introduction-example/

Hashemi-Pour, C. (2024, June). *Customer Retention*. Retrieved from TechTarget: https://www.techtarget.com/searchcustomerexperience/definition/customer-retention

Hayes, A. (2024, June 27). *Inventory Management: Definition, How It Works, Methods & Examples*. Retrieved from Investopedia: https://www.investopedia.com/terms/i/inventory-management.asp

Kenton, W. (2024, July 10). *Tailored Advertising: Meaning, Effectiveness, Examples*. Retrieved from Investopedia: https://www.investopedia.com/terms/t/tailored-advertising.asp

K-means Clustering Algorithm: Applications, Types, & How Does It Work? (2024, July 23). Retrieved from Simplilearn: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm

Online Retail. (n.d.). Retrieved from UCI: https://archive.ics.uci.edu/dataset/352/online+retail

Oracle Machine Learning. (n.d.). Retrieved from Oracle: https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/21/dmprg/about-attributes.html#GUID-7AAB55D5-6711-4BE5-A0CE-B2A6B68ED689

Random Forest Algorithm in Machine Learning. (2024, July 12). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

Samina. (n.d.). *What is Silhouette Score?* Retrieved from Educative: https://www.educative.io/answers/what-is-silhouette-score

Stedman, C. (n.d.). *What is data preparation? An in-depth guide*. Retrieved from TechTarget: https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation

The importance of customer segmentation for businesses. (2023, November 17). Retrieved from ABMATIC AI: https://abmatic.ai/blog/importance-of-customer-segmentation-for-businesses

What is data visualization? (n.d.). Retrieved from IBM: https://www.ibm.com/topics/data-visualization

What is Descriptive Statistics: Definition, Types, Applications, and Examples. (2024, Jun 11). Retrieved from Simplilearn: https://www.simplilearn.com/what-is-descriptive-statistics-article