A picture containing icon

Description automatically generated

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Machine Learning for Business* |
| **Assessment Title:** | *CA1* |
| **Lecturer Name:** | *Sam Weiss* |
| **Student Full Name:** | [*Riccardo Possieri*](https://moodle.cct.ie/user/view.php?id=13536&course=2625) |
| **Student Number:** | *sba23439* |
| **Assessment Due Date:** | *03/04/2024* |
| **Date of Submission:** | *05/04/2024* |

**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. |

CA 1

1. Introduction

The purpose of this report is to analyse customer transaction data from an online retailer using machine learning techniques as per CA. The dataset provided includes details such as invoice numbers, product descriptions, quantities, prices, and countries. Our objective is to apply clustering algorithms to segment customers based on their purchasing behavior and perform market basket analysis to uncover patterns in item co-occurrences. We will explain each step we did and the reason why and finally our conclusions and interpretation.

2. Data Exploration and Preparation

As we could see when we printed out the information about the two sheets of the dataset, we realized that we had to work with a big dataset of almost a million observations: (525461, 8) (541910, 8). After that, we conducted exploratory analysis and prepared the dataset for analysis, merging the two sheets in a unique dataset of 1067371 rows and 8 columns. We addressed missing values by dropping rows with missing descriptions and retaining those with missing customer IDs. Additionally, we removed transactions with negative prices to ensure data integrity. By cleaning and refining the dataset, we ensured its suitability for subsequent analysis. Also, in the preprocessing step, we converted InvoiceDate column to datatime format to able us to work with that.   
  
We plot the distribution of Countries and we realized that most of the transactions were from UK. We also plotted the price distribution with a boxplot and we noticed we had negative prices on the dataset: 5 negative transactions. Since we had only 5 transactions with negative prices, we could drop them. Another important step we have done was to enumerate them so we could use it in the clustering. We also added a column containing the total price for a stock code and quantity, so we multiplied the price for the quantity.

We preprocessed the dataset by grouping the data by InvoiceNo so that every row would correspond to a single transaction. we would add the number of items purchased in each transaction as a new column as well as the total price of the transaction.

Invoice TotalPrice Quantity  
0 489434 505.30 166  
1 489435 145.80 60  
2 489436 630.33 193  
3 489437 310.75 145  
4 489438 2286.24 826

3. Clustering

3.1 Feature Selection and Scaling

To prepare for clustering, we selected relevant features such as TotalPrice and Quantity and scaled the data using the StandardScaler to mitigate scale-related biases. This step is essential for algorithms like K-means, which are sensitive to feature scales.

3.2 Clustering Techniques

We applied multiple clustering techniques, including K-means, K-medoids, and DBSCAN, to identify customer segments based on their purchasing behavior. Each algorithm offered us unique advantages, and we evaluated their performance using metrics such as silhouette score and then we compared them.

“Clustering is the task of dividing the unlabeled data or data points into different clusters such that similar data points fall in the same cluster than those which differ from the others. In simple words, the aim of the clustering process is to segregate groups with similar traits and assign them into clusters.” Clustering: Different Methods, and Applications (Updated 2024). Let’s understand this with an example. Suppose you are the head of a rental store and wish to understand the preferences of your customers to scale up your business. Is it possible for you to look at the details of each customer and devise a unique business strategy for each one of them? Definitely not. But, what you can do is cluster all of your customers into, say 10 groups based on their purchasing habits and use a separate strategy for customers in each of these 10 groups. And this is what we call clustering Thanks to the silhouette score we could measure of how well defined the clusters were within a the dataset with values closer to 1. In our analysis, the silhouette score for K-means clustering (0.993) was significantly higher than that of DBSCAN (0.925). This suggested that the clusters generated by K-means were more distinct and well-separated compared to those produced by DBSCAN.

After the comparison, we noticed that K means clustering was more suitable and DBSCAN value told us that it was more spread out in the feature space. Therefore, for the purpose of customer segmentation and deriving actionable insights from the data, K-means clustering appears to be the preferred choice. Thanks to that and for academical goal, we could say that K means is better in order to identify distinct customer segments. In summary we could say that is very important to select the appropriate clustering algorithm based on the dataset.

4. Market Basket Analysis

Market basket analysis is a strategic data mining technique that we can use to recognize patterns after we analyse the datasets. In our CA, we used a Descriptive MBA (Market Basket Analysis). Descriptive MBA is useful when we want to identify patterns and associations among items purchased together. In the code that we made, the focus is on understanding the relationship of the products purchased by customers in the past transactions.

4.1 Data Preparation

For market basket analysis, we just needed invoice and StockCode, we could then retrieve the description of the product from the original dataset. For a single stockCode there were multiple description for 1232 times. This implied we must use description for the transaction encoder input. We needed to map every unique description to a unique integer to use the TransactionEncoder, we created a dictionary to do that and use it with map on the dataframe and then we replaced descriptions with corresponding integers. We then fit the TransactionEncoder to the transactions and transform the data into one-hot encoded format.

4.2 Association Rule Mining

We utilized the Apriori and FP-growth algorithms to mine association rules from the transaction data. These algorithms differ in their approach to candidate generation and database scans, but both offer insights into item relationships and co-occurrences.

4.3 Results and Interpretation

The association rule mining revealed interesting patterns and relationships among items purchased by customers. These rules provide actionable insights into product bundling, cross-selling, and targeted promotions, with the potential to enhance customer engagement and optimize marketing strategies.

Apriori Algorithm:  
Support: 0.02  
Threshold: 0.6  
Best Rules:  
Antecedents: (RED HANGING HEART T-LIGHT HOLDER),   
Consequents: (WHITE HANGING HEART T-LIGHT HOLDER)Support: 0.025,   
Confidence: 0.699,   
Lift: 6.163

FP-growth Algorithm:  
Support: 0.02  
Threshold: 0.6  
Best Rules:  
Antecedents: (RED HANGING HEART T-LIGHT HOLDER),   
Consequents: (WHITE HANGING HEART T-LIGHT HOLDER)  
Support: 0.025,   
Confidence: 0.699,   
Lift: 6.163

Based on the provided information, both algorithms yield the same best rule, which is the association between purchasing "RED HANGING HEART T-LIGHT HOLDER" and "WHITE HANGING HEART T-LIGHT HOLDER". This rule has a support of 0.025, a confidence of 0.699, and a lift of 6.163 for both algorithms. Therefore, this association rule can be considered the best for both Apriori and FP-growth algorithms under the given conditions.

5. Conclusion

To resume the CA and giving an overall conclusion, we must say that k means clustering was found to be the preferred choice for customer segmentation. Association rule mining highlighted some patterns to optimize the strategies of the marketing. The work gave us a clear understanding of customer behavior and some recommendations that could be take.

6. References

Sklearn.preprocessing.StandardScaler. (n.d.). scikit. Retrieved April 11, 2024, from <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

Algoritmi di clustering  |  machine learning  |  google for developers. (n.d.). Google. Google. Retrieved April 11, 2024, from <https://developers.google.com/machine-learning/clustering/clustering-algorithms?hl=it>

Kaushik, S. (2024). Clustering: Different methods, and applications (updated 2024). Analytics Vidhya. Retrieved April 11, 2024, from <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/> `

Kadlaskar, A. (2024). Market basket analysis: A comprehensive guide for businesses. Analytics Vidhya. Retrieved April 11, 2024, from <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/>

Mitchell, T.M. (2021) Machine learning. New York: McGraw-Hill.