****

**MISM 6212 – Data Mining and Machine Learning for Business**

**Final Project Report**

Customer Segmentation Analysis for Online Marketing Retail Strategy using Data Mining and Machine Learning Methods

**Group 5**

Vasatika Ghadiyaram

Atithi Prasai

Riya Reji

Yiyang Zhang

Faizal Rizvi

**Industry Overview**

The online retail industry, which involves the buying and selling of goods and services through the internet, has grown rapidly due to technological advancements and convenience. According to Statista, the global e-commerce market is expected to reach $6.5 trillion by 2023. Major players in the industry include Amazon, Alibaba, JD.com, Walmart, and eBay, with popular product categories such as consumer electronics, apparel and accessories, and beauty and personal care. Despite its growth and success, the industry also faces challenges such as fierce competition, customer segmentation, and logistics issues.

To gain a competitive advantage, online retail companies need to gain a deeper understanding of their customers' needs and preferences by grouping them based on shared characteristics such as purchasing history, website behavior, device/channel usage, demographics, and psychographics. By doing so, they can tailor their marketing strategies and offerings to better meet their customers' needs, which can lead to increased customer loyalty, higher customer satisfaction, and ultimately increased sales and profits. Additionally, online retail companies need to ensure a seamless and convenient customer experience, efficient delivery and logistics, and competitive pricing to stay ahead in the highly competitive market.

**Business Problem**

In the fiercely competitive online retail industry, companies need to understand their customers' needs and preferences to gain a competitive advantage. Customer segmentation is one way to achieve this, but current methods may be time-consuming and fail to capture complex patterns and trends. This limits a company's ability to develop effective personalized marketing campaigns and offerings, which can lead to reduced customer loyalty and lower profits. To stay ahead of the competition, online retail companies must explore new and innovative ways of customer segmentation that accurately capture insights into their customers' behavior and preferences.

​

**Objectives and Solution Techniques Employed**

We seek to leverage data mining techniques to automate the process of customer segmentation, identify more complex patterns and trends in the data, and optimize business processes to potentially increase sales and gain a competitive advantage in the market.​

To achieve our project objective, we will employ two solution techniques: hierarchical clustering and K-means clustering. Hierarchical clustering and K-means clustering are both unsupervised machine learning techniques that can be used to group customers based on shared characteristics. These two techniques can help us identify customer segments that may not be apparent through traditional segmentation methods and inform our personalized marketing campaigns.

Compared to traditional segmentation methods, data mining techniques offer several advantages. Traditional segmentation methods often rely on manual data analysis, which can be time-consuming and may not identify complex patterns and trends in the data. In contrast, data mining techniques can automate the segmentation process and identify more nuanced patterns and trends in the data. This allows us to identify customer segments that may not be apparent through traditional segmentation methods and tailor our marketing campaigns accordingly.

Overall, the use of data mining techniques can provide significant benefits to online retail companies seeking to gain a deeper understanding of their customers and increase their competitiveness in the market.

**Data Collection and Description**

For our problem we collected the data from the Online Retail dataset available on the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/online+retail), as we are applying the data mining and machine learning techniques to the overall industry.

This is a transnational data set from a UK-based and registered non-store online retail which contains all the transactions occurring between 2010 and 2011. ​The company mainly sells unique all-occasion gifts to wholesalers and other customers.​

**Description:**

Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Attribute​** | Data Type​ | Description​ |
| InvoiceNo​ | object​ | 6-digit integral number assigned to each transaction. If it starts with letter 'c', it indicates a cancellation​ |
| StockCode​ | object​ | 5-digit integral number assigned to each distinct product.​ |
| Description​ | object​ | Name of the product.​ |
| Quantity​ | int64​ | Quantity of each product per transaction.​ |
| InvoiceDate​ | Datetime64​ | Date and time when each transaction was generated​ |
| UnitPrice​ | float64​ | Product price per unit in sterling​ |
| CustomerID​ | float64​ | 5-digit integral number assigned to each customer.​ |
| Country​ | object​ | Name of the country where each customer resides​ |

**Data Cleaning and Transformation**

**Data Cleaning:**

First, we imported the necessary libraries and then read the online retail dataset into a panda’s data frame. We performed the following data cleaning steps:

1. Imported and read in the "Online Retail.xlsx" dataset using the pd.read\_excel() function and stored it in the variable retail.
2. Removed all rows where the quantity of items purchased is less than or equal to zero.
3. Removed all rows where the CustomerID column has missing values. This is because the CustomerID column is a crucial identifier, and any row without this information cannot be used for further analysis.
4. Removed all rows where the Description column has missing values. This is because each product is unique, and any row without this information cannot be used for further analysis.

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Data Transformation:**

After cleaning the data, we performed the following data transformation steps:

1. Created a new column called "Revenue" by multiplying the Quantity and UnitPrice columns.

2. Created a new column called "Revenue" by multiplying the Quantity and UnitPrice columns.

3. Grouped the dataset by CustomerID and calculated the total revenue for each customer.

4. Extracted the Revenue column from customer\_df and stored it in a new variable called X. This column will be used for clustering.

5. Normalized the X data using the StandardScaler() function. This is done to ensure that all features have the same scale, which is important for certain clustering algorithms such as KMeans.

Graphical user interface, text, application, email

Description automatically generated

**Data Models Used**

**K-Means Clustering:**

* K-means clustering is a machine learning algorithm that aims to partition a dataset into K clusters, where K is a user-defined parameter.​
* K-means is often used in unsupervised learning to group similar data points together based on their features.​
* The elbow plot is a technique used to determine the optimal value of K for K-means clustering by plotting the sum of squared distances between data points and their assigned cluster centers against different values of K.​
* The silhouette score is another metric that can be used to evaluate the quality of a K-means clustering solution by measuring the similarity of data points within clusters and the dissimilarity between different clusters. A higher silhouette score indicates a better clustering solution.​
* K-means clustering can be applied to the online retail dataset by using the customers' purchasing behavior as features and clustering them into groups based on their buying patterns. This can help retailers to identify customer segments and tailor their marketing strategies to better target each group.

**Steps Taken:**

1. First, we used Principal Component Analysis (PCA) for dimensionality reduction.
2. We used the elbow-curve/SSD method to determine the optimal number of clusters. We then iterated over a range of values for K and calculated the sum of squared distances for each value of K. We ploted the SSDs against the number of clusters to determine the optimal number of clusters.
3. We used silhouette analysis to evaluate the performance of the clustering. We iterated over a range of values for K and calculated the silhouette score for each value of K. We then printed the silhouette score for each value of K.
4. Considering the most ideal option wee performed k-means clustering with K=3.

Graphical user interface, text, application, email

Description automatically generated

**Chart, line chart

Description automatically generated**

**Hierarchical Clustering:**

Hierarchical clustering is a widely used unsupervised machine learning algorithm that groups similar data points together. Unlike k-means, which requires the number of clusters to be specified in advance, hierarchical clustering does not need this information.

The algorithm begins by considering each data point as its own cluster. The next step is to find the two closest clusters and merge them into a single cluster. The distance between clusters can be measured using various metrics, such as Euclidean distance, Manhattan distance, or cosine similarity. After the first merge, the algorithm proceeds to the next step, where it again finds the two closest clusters and merges them into a new cluster. This process continues until all data points are merged into a single cluster.

The result of hierarchical clustering can be visualized using a dendrogram, which is a tree-like diagram that shows the order of clustering and the distance between clusters. The dendrogram is constructed by plotting the distance between clusters on the y-axis and the data points on the x-axis. The height of each branch in the dendrogram represents the distance between the clusters being merged.

A picture containing graphical user interface

Description automatically generatedIn our case, we used single linkage and average linkage to plot the dendrogram. Code is as follows:

**Graphical user interface, text

Description automatically generated**Single linkage dendrogram is a type of agglomerative hierarchical clustering method that uses the shortest distance between two clusters to merge them into a larger cluster. This means that it looks for the smallest distance between any two data points in different clusters and considers that distance as the distance between the two clusters. Single linkage dendrogram tends to produce long, narrow clusters that may not be well-separated from each other.

On the other hand, average linkage dendrogram is a type of agglomerative hierarchical clustering method that uses the average distance between all pairs of data points in different clusters to merge them into a larger cluster. This means that it calculates the average distance between all pairs of data points in different clusters and considers that distance as the distance between the two clusters. Average linkage dendrogram tends to produce more balanced clusters that are better separated from each other than single linkage dendrogram.

With these plots, we decide the number of clusters is 3 clusters. However, the number of clusters obtained from hierarchical clustering methods is not always unique and may depend on the choice of distance metric, linkage method, and threshold distance. Therefore, it is recommended to try multiple methods and compare their results to determine the optimal number of clusters for a given dataset.

**Chart, scatter chart

Description automatically generated**

**Model Recommendation**

Based on the Silhouette scores, both K-means clustering, and hierarchical clustering performed well in this analysis. However, K-means clustering had a slightly higher Silhouette score of 0.96 compared to hierarchical clustering's score. This suggests that K-means clustering may be a slightly better choice for customer segmentation in this particular case. Additionally, K-means clustering has the added advantages of ease of implementation, speed and scalability, flexibility, and intuitive interpretation. However, it's important to note that the choice of clustering algorithm ultimately depends on the specific dataset and business problem at hand.

**Managerial and Practical Implications**

The project has several managerial and practical implications that can be useful for businesses.

* Firstly, the project demonstrates the importance of customer segmentation in marketing strategies. By segmenting customers based on their transactional behavior, businesses can tailor their marketing strategies to each segment and offer personalized promotions and incentives to encourage repeat purchases and foster brand loyalty. The project highlights the potential benefits of offering loyalty rewards and exclusive promotions to high-value customers, as well as targeted re-engagement campaigns for inactive customers.
* Secondly, the project highlights the value of data-driven decision making. By analyzing customer transaction data and clustering customers based on their behavior, businesses can gain insights into their customer base and make informed decisions about marketing strategies, product offerings, and operational improvements. The project also demonstrates the importance of regularly analyzing and updating customer segmentation to ensure that marketing strategies remain relevant and effective.
* Thirdly, the project shows the usefulness of machine learning algorithms such as K-means and hierarchical clustering in customer segmentation analysis. These algorithms can help businesses identify distinct groups of customers based on their behavior and provide insights into the unique characteristics and needs of each group. By using these algorithms, businesses can streamline their customer segmentation process and save time and resources compared to manual segmentation methods.
* Finally, the project emphasizes the importance of understanding customer behavior beyond just transactional data. By incorporating additional data sources such as demographic and psychographic information, businesses can gain a more holistic understanding of their customers and create more targeted marketing strategies. Overall, the project demonstrates the potential benefits of data-driven customer segmentation and the use of machine learning algorithms in marketing strategy development.

**Managerial Recommendations**

Based on the findings of the clustering analysis, here are some managerial recommendations:

1. Focus on high-value customers: Customers in Cluster Id 0  are the high-value customers with frequent transactions. The company should consider implementing a loyalty program or personalized promotions to incentivize repeat purchases and foster brand loyalty. By providing these customers with exclusive rewards and benefits, the company can encourage them to continue doing business with them and potentially increase their lifetime value.
2. Re-engage fewer active customers: Customers in Cluster Id 1 are not recent buyers and of least importance from a business perspective. The company should consider targeting them with personalized and targeted re-engagement campaigns. Offering exclusive discounts or promotions can incentivize them to make a purchase and re-engage with the brand. By targeting these customers with personalized messaging and incentives, the company may be able to recapture their attention and turn them into repeat buyers.
3. Chart, box and whisker chart

   Description automatically generatedTailor marketing strategies: Consider conducting further analysis to identify the specific characteristics and behaviors of customers in each cluster. By understanding what motivates customers in each segment, the company can tailor their marketing strategies and campaigns to better meet their needs and preferences. This could potentially lead to increased customer satisfaction and loyalty, as well as higher sales and revenue for the business.

**References**

* Online Retail Dataset. (n.d.). UCI Machine Learning Repository. Retrieved April 22, 2023, from <https://archive.ics.uci.edu/ml/datasets/online+retail>
* K-means clustering. (2022, March 28). In Wikipedia. Retrieved April 22, 2023, from <https://en.wikipedia.org/wiki/K-means_clustering>
* scikit-learn developers. (n.d.). sklearn.cluster.KMeans. scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation. Retrieved April 22, 2023, from <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
* Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7
* scikit-learn developers. (n.d.). sklearn.metrics.silhouette\_score. scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation. Retrieved April 22, 2023, from <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>
* Hierarchical clustering. (2022, March 23). In Wikipedia. Retrieved April 22, 2023, from <https://en.wikipedia.org/wiki/Hierarchical_clustering>
* scikit-learn developers. (n.d.). sklearn.cluster.AgglomerativeClustering. scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation. Retrieved April 22, 2023, from <https://scikitlearn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>
* Kumar, S. (n.d.). Online Retail K-Means & Hierarchical Clustering. Kaggle. Retrieved April 22, 2023, from <https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering>