CUSTOMER SEGMENTATION USING AI

SHRERAM S P
Computer Science and Engineering (CSE)
Bannari Amman Institute of Technology (BIT)
Erode, INDIA
shreram.cs19@bitsathy.ac.in

CHANDRU A
Electronics and Communication Engineering (ESE)
Bannari Amman Institute of Technology (BIT)
Erode, INDIA
chandru.ec19@bitsathy.ac.in

SUBASH R
Information Science and Engineering (ISE)
Bannari Amman Institute of Technology (BIT)
Erode, INDIA
subash.ig19@bitsathy.ac.in

***Abstract*—The emergence of many business rivals creates serious competition among competing companies to win new customers and retain existing ones. All of this makes the need for excellent customer service critical, regardless of the size of your business. In addition, a company's ability to understand its customers' needs is more powerful in providing targeted customer service and developing customized marketing programs for customers. This understanding is made possible by systematic segmentation of customer data. The development of big data and machine learning has led to the successful adoption of automated approaches to customer segmentation in place of traditional market analysis, which is often inefficient, especially when there are too many customers this paper, using Machine learning K means Clustering Algorithm, the dataset containing the customers' data is segmented into different segments on the basis of the three parameters Recency, Frequency, and Monetary (also called RFM analysis).**

1. INTRODUCTION

The availability of customer data on a large scale and the escalating level of corporate competitiveness have led to an expanded usage of knowledge-mining techniques to extract critical and strategic information from organizational datasets. Data filtering is the process of extracting logical information from a dataset and presenting it for decision support in a manner that is understandable to humans. Areas including statistics, machine learning, artificial intelligence, and data mining systems are differentiated by data processing methodologies. Bioinformatics, meteorology, fraud detection, financial analysis, and customer segmentation are just a few examples of data processing applications.

Customer segmentation is the process of categorizing customers into marketing segments based on their shared characteristics. To be more precise, it indicates creating consumer segments based on shared traits is the best marketing strategy. Client segmentation involves gathering data on each customer and analyzing it to find various patterns that can be used to create segments. Face-to-face interviews, telephone interviews, surveys, and research using data from published sources linked to market categories are some of the finest ways to collect information. Demographic, RFM (Recency, Frequency, Monetary) analysis, HVCs (High-Value Customers), customer status, behavioral, psychographic, etc., are examples of segmentation categories that are frequently used. Marketing strategy, promotion strategy, budget efficiency, product development, and others are some of the primary advantages of client segmentation. In this paper, we used the fundamental analytics functionality to give the decision-makers—in this case, company investors—the data they needed to make the best choice.

So, we propose a strategy for lowering risk variables and offer input on choosing new company investments. We suggested using the K-means approach to segment customers. Using information analytics to segment the client base is our solution. Groups of consumers can be created based on the behaviors they exhibit frequently. These actions reflect their understanding of, attitude toward, use of, spending level, or reaction to a product. For this consumer segmentation, we employed the K-Means clustering approach from machine learning. Unsupervised Learning is the category in which this type of learning falls. The k-Means algorithm, the k-nearest algorithm, the Sorting Map (SOM), and other algorithms are examples of integration algorithms. These algorithms are prepared to discover clusters in them without any prior knowledge of the information by repeatedly comparing input patterns until stable qualifications within the training instances are produced by relying on the subject matter or the approach.

2. METHODOLOGY

There are three major divisions:

- Data Pre-processing
- RFM Analysis Calculation
- Clusters Creation

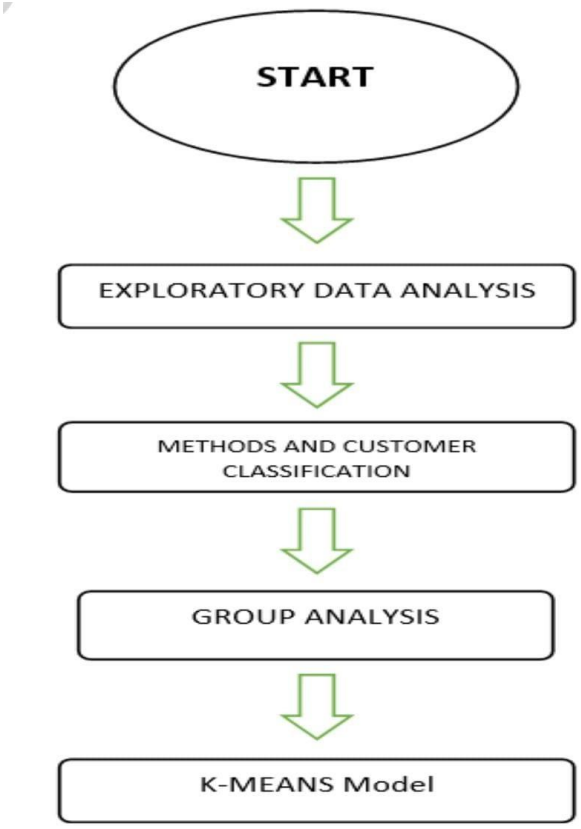


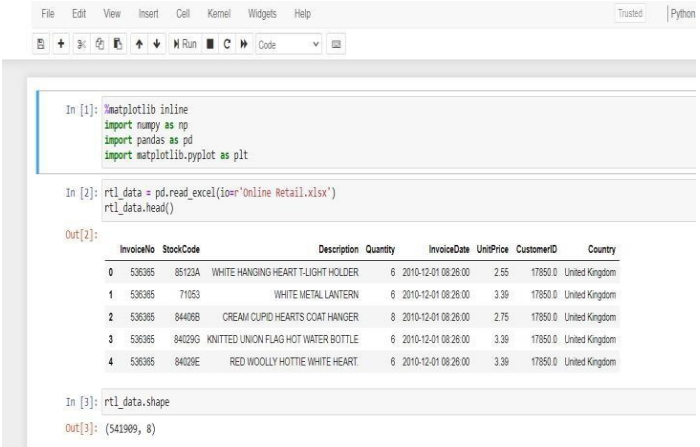
Fig 1: *Steps in Building K means Model*

The execution of a developed Python programme in Jupyter Notebook requires importing the following required packages:

- **Pandas** – It's an open-source Python toolkit for data science, data analysis, and machine learning. Its foundation is NumPy, which supports multidimensional arrays in a library. One of the most popular tools for data manipulation, Pandas, performs well. Including numerous additional Python data science modules.
- **Numpy** – It is a collection of multidimensional array objects and the tools needed to work with them. Using the Python package NumPy, we can execute arrays that can perform logical and mathematical operations. Scientific Python (NumPy) is frequently used in conjunction with SciPy and Matplotlib (plotting library). The popular technical computing platform MATLAB is commonly replaced by this combination. On the other hand, Python is now thought of as a more up-to-date and complete programming language than MATLAB.
- **seaborn** – A matplotlib-based open-source Python library is called Seaborn. It is utilized for data visualization and exploratory data analysis. Using data frames and the Pandas library is a breeze with Seaborn. The resulting graphs can also be easily modified.
- **matplotlib** –Matplotlib is a fantastic Python visualization package for 2D array charts. The complete SciPy stack is compatible with Matplotlib, a multi-platform data visualization package built on NumPy arrays. One of the most important benefits of visualization is the capacity to display enormous amounts of data in straightforward graphics. In addition to other graph types, Matplotlib offers line, bar, scatter, and histogram graphs.
- **Scikit** – It is a free machine learning programme written in Python frequently referred to as sklearn. It includes support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques. It is designed to work with the Python numerical and scientific libraries NumPy and SciPy.

2.1. DATA PRE-PROCESSING

Pandas is used to load the dataset, which is an excel file.



```
In [1]: %matplotlib inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

In [2]: rtl_data = pd.read_excel(io=r'Online Retail.xlsx')
        rtl_data.head()
```

Out[2]:

	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

```
In [3]: rtl_data.shape
```

Out[3]: (541909, 8)

Fig 2: *Dataset loaded using Pandas*

We remove all of the Data's duplicate entries. To display the greatest number of customers' purchases, the count of customers is calculated and arranged in descending order.

Exploratory Data Analysis (EDA) is done to the dataset. Anomalies or repeated data are deleted. Clues and Patterns findings are searched in the dataset.

2.2. RFM ANALYSIS

Since our dataset only contains sales records, we can only use an RFM-based model to locate segments where R is Recency (how recently a purchase occurred), F is Frequency (how frequently transactions are performed), and M is Monetary value (Value of all transactions). Each customer's financial, frequency, and recency scores are computed. In order to determine recent purchases, the most recent date is used as a placeholder. Using CustomerID, all transactions are pooled, and then aggregate lambda operations are carried out. This operation will yield numbers that show the recentness, frequency, and total amount spent by a particular consumer to date. These are all kept in a brand-new data frame called RFM scores. The distribution for recency is right-biased, it should be noted.

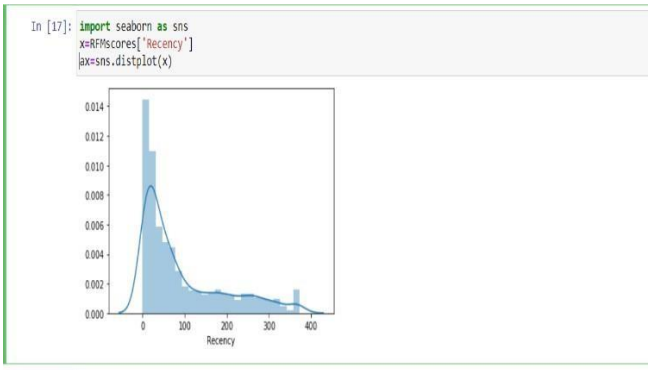


Fig 3: Recency Plot

Similar to the above recency plot, the frequency and Monetary plots and calculated.

2.3. CLUSTERING

K-Means is a clustering algorithm that uses unsupervised learning and performs exceptionally well with complex datasets. The dataset is divided into "k" pre-specified, non-overlapping subgroups (clusters) by an iterative process, and each data point only belongs to one cluster.

This is how the algorithm operates:

Step 1: Entering the k value to specify the number of clusters.

Step 2: Centroids are initialized by randomly choosing k data points for the centroids without replacement after shuffling the dataset.

Step 3: Continue iterating until the centroids remain unchanged. In other words, the clusters' assignments to the data points remain the same.

Before beginning the clustering procedure, the data is standardized and the scales for Recency, Frequency, and Monetary are adjusted. Finding the ideal number of clusters, or the "k value," is crucial. We employed the elbow method in this. It entails repeatedly iterating the method over a loop with an increasing number of cluster options before plotting a score as a function of the cluster count. The centroids are nearer cluster centroids as "k" rises. The reason this strategy is known as an elbow is that the improvement will eventually begin to fall quickly and create the appearance of an elbow on the graph. At the bend in this elbow, we take the cluster count and k-value.

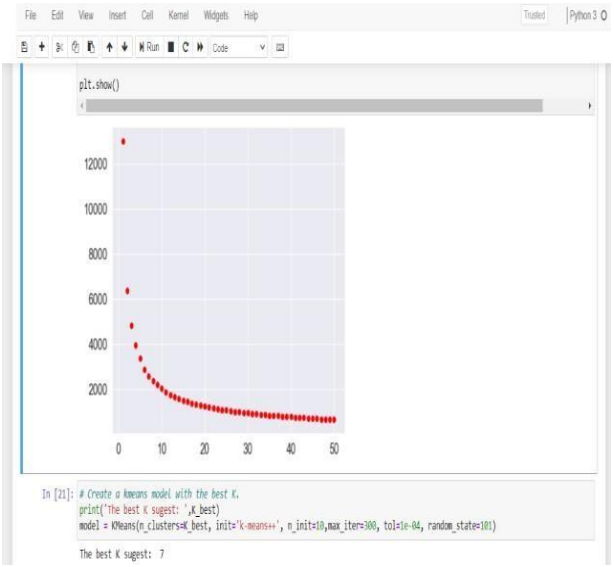


Fig 4: *K value using Elbow method*

3. FUTURE WORK

Given the scarcity of information about changes in customer behaviour, the recommended basic cluster model. To uncover the clients' hidden patterns and shopping tendencies, additional criteria and approaches are therefore required. The discovery of clusters of potential clients was aided by RFM and K-means. Additionally, cross-selling and market basket analysis tools can be used to analyse consumer behaviour and suggest new products in the hopes that they will buy them, which would be advantageous for both the customer and the retail institution.

4. CONCLUSION

This study used information gathered from an online retailer to deploy the k-Means clustering method for client segmentation. In our example, our model divided customers into three clusters that were mutually exclusive. This will be important for implementing additional data mining techniques, and the conclusions drawn from it are beneficial in helping the business wings make decisions.

5. REFERENCES

[1] "Customer segmentation based on survival character," IEEE Paper, Jul. 2003.

[2] "Customer Segmentation Using K Means Clustering," Towards Data Science, Apr. 2019.

[3] Juni & Nugroho, Nurma Sari, Ridi & Santosa, Paulus Lukito & Ferdiana,. Review on the title "Customer Segmentation Technique on Ecommerce". Advanced Science Letters. 2016.

[4] F.Daniel, Customer Segmentation using Machine Learning: classification, clustering , marketing . www.kaggle.com

[5] Rachel Blasucci. Title "Event triggered Customer Segmentation". DZone, July 23, 2018.