# Customer Segmentation Using Unsupervised Machine Learning Techniques DRAFT DISSERTATION RESEARCH PAPER SUBMISSION 2021-22 (TANISHQ VERMA 210642458)

***Abstract* – This paper contains finding and accurate application for Customer Segmentation Using Unsupervised Machine Learning Techniques. Different types of Unsupervised machine learning techniques have been applied on the selected dataset to cluster the data into different clusters. The particularly selected dataset contains 8 attributes and features which helps in appropriate applying of these clustering techniques. This paper assesses the techniques and explains why the selected techniques have been applied to it and will further explain the result of each of the techniques. In this report, data reprocessing techniques have been applied to understand the dataset and visualize the dataset. A literature review has also been specified as part of all the references that have been taken while creating the model. The literature review will further help in discussions for the reader to help explore more domains. We have evaluated the dataset on different types of clustering techniques such as RFM, Hierarchical Clustering, K-Means and DBSCAN to identify the pros and cons of the model and give which techniques is the best of them.**

***Keywords* – Customer Segmentation, Data Cleaning, Data pre-processing, Clustering Techniques, Agglomerate Clustering, K-Means, DBSCAN, Hierarchical Clustering**

## I. INTRODUCTION

Customer Segmentation is an effective process through which large business can uses effective strategies and tactics to better target customers. Each and every customer for each industry are unique from one another therefore one single strategy cannot be one fit for all. This is where customer segmentation comes into place to handle the different customers and fit or classify each customer according to their behavior and habits.

Customer Segmentation is the process in which customers are segregated into different groups based on common characteristics such as demographics, spending habits, targeting specific market etc. Customer Segmentation help business understand the customers and how they can further target their customers by branding like brand messaging, positioning and will further help their business to improve.

When a business understands the segmented customers, this helps them to identify the need for each of the segmented customer. The clustered customers then have a similar behavior or habit which can help business to find a common approach and target them all since they can apply one fit for all strategies for the whole segment.

Customer Segmentation can help business to increase their profit by understanding what their customers want and in addition, the customers also get benefit by getting what they want from the companies. This approach works well and is a win-win situation for both the customers as well as for organizations.

### *A. Types of Customer Segmentation*

Customer Segmentation are done based on different factors. Each of these factors are unique i.e. they are not one fit for all and specific to the business. Each factor can further be sub-classified into two parts. First, the type of the customer which include their Age, Gender, Demographic, Income, Family, Relationship Status, Job type and other. For the second, which includes the behavior of the customer are spending habits, recency, money spent, buying products on sale or fresh arrival, membership, or loyalty to the business and other.

### *B. OBJECTIVES*

- Selecting the dataset which contains appropriate numbers of data samples in it.
- Applying Data Cleaning and Data Pre-processing techniques to it which will further help in visualizing the data.
- Visualizing the data and understand the data and all the features in the data set.
- Selecting the right unsupervised machine learning techniques like RFM, DBSCAN, K-means and get the appropriate output for each one.
- Getting the result for each of the technique and help explain why the technique generated different result.

### *C. Selecting the dataset*

The dataset that we selected is from UCI Machine Learning Repository called Online Retail Data Set. This dataset contains the transaction record from 01/12/2010 to 9/12/2011 on UK based and non-store online retail. This dataset is especially targeted towards clustering and classification techniques. The dataset contains 541909 number of instances with 8 number of attributes. Each of the attributes can be further defined as –

1. Invoice no – This is a 6-digit integer combination which is unique to each transaction.
2. Stock Code – This is a 5-digit integer combination which is unique to each product.
3. Description – This is a description of the product.
4. Quantity – This is a numerical value which defines the quantity of product in the transaction.
5. Invoice Date – This attribute contains the numerical data and time when the transaction was generated on.
6. Unit Price – This contains the per unit price of the product in pounds sterling.
7. Customer ID – This is a 5-digit integer combination which is unique to each customer.
8. Country – This is the name of the

country to which the customer belongs.

The data type for each attribute in the dataset is of Object type except Quantity which is of integer type and Unit Price and Customer ID which both are of float type.

## II. LITERATURE REVIEW

This section of the report contains the different types of reports, online discussion, journals, literature reviews and references. The outcome and the assumptions for the same have been explained below.

The first report is titled 'Customer segmentation using unsupervised clustering algorithm' by Sharma Sheelesh Kumar, Professor, Pathak Garvit, MCA Student, Chandralo and MCA Student, Kumari Sweta [1] published under the Department of IT, Institute of Management Studies, Ghaziabad, Uttar Pradesh, India by Indian Journals.com. The aim of the paper was to apply the different types of clustering techniques and further explain if the clustering techniques where efficient to calculate the cluster or clustering the given dataset into different types of clusters. These paper applies the K-means clustering technique along with strategic analytic and also did the different types of data cleaning and data pre-processing techniques as well before doing the data visualizing to understand the dataset.

The second report that was taken part literature review is titled 'Customer Segmentation using K-means Clustering' by Tushar Kansal, Suraj Bahuguna, Vishal Singh and Tanupriya Choudhury [2] by Dept. of Informatics, School of Computer Science, Dehradun. This research paper which is published under IEEE uses a dataset to apply customer segmentation using unsupervised machine learning technique. This paper has taken into account the three different types of clustering techniques which includes K-means, Agglomerative, and Meanshift and further explains the clustering done by each of these techniques. Further, it includes that a python program was build which contains the output for each techniques and further explains that it used standard scalar to help easy and efficiently apply the data to the clustering to get the best result on their 200 training data.

The final report for literature review is titled 'E-commerce Customer Segmentation via Unsupervised Machine Learning' by Boyu Shen [3] in ACM Digital Library. This research work is recent and was published on 17 May 2021. Being recent and having a little different approach, this paper uses different types of clustering techniques and it uses of the clustering technique to segment the customers into different clusters. To illustrate, using the K-means to clusters the customers, Apriori Algorithm used to analyse product and TF-IDF is used to category the product into different groups.

## III. DATA MANAGEMENT

### *A. Data source and description*

The dataset for the project is selected from an open source website called UCI Machine Learning Repository.

This dataset contains the transaction record from 01/12/2010 to 9/12/2011 on UK based and non-store online retail. This dataset is especially targeted towards clustering and classification techniques. The dataset contains 541909 number of instances with 8 number of attributes.

*1) Data Preprocessing:* For the data reprocessing part and data cleaning we applied different types of techniques. These includes removing duplicate values, removing null value, removing values which have no meaning. We further visualised data on graph according to countries. The result was as expected that most of the data was form UK as it is an UK based online retail dataset. Further, we changed the data types of the attributes so that appropriate techniques shall be applied. Further, we also standard scaled the data set for easily applying the techniques.

*2) Feature Selections:* As part of Feature Selection, we selected all of the features values except for product description. All of the features were important however we used the existing features to create new features as well. To exemplify, we used the invoice data to calculate the recency for the model attribute.

*3) External Libraries:* Some of the python external libraries that have been used for this project have been described below. These includes-

NumPy – A library helps and backing the extremely enormous multiple layer cluster and frameworks. NumPy is utilized to give admittance to enormous number related libraries and tackle troublesome numerical capabilities for working on these exhibits and grids.

Pandas – Pandas is an extremely famous library with regards to Machine Learning models. A system library in Python can assist with handling information. A df or DataFrame is characterized in 2 Dimension grid that upholds large numbers of the procedure on csv record and helps calculations simple.

Scikit-learn – A Machine Learning library utilized in Python that is normally utilized in Data Science and Artificial Intelligence. It is an extremely clear and simple to utilize library utilized for carrying out calculations.

Mathplotlib- A library that is utilized for representation purposes that comprise of various kinds of charts and plots. The legend and pivot alongside scales can be modified by the need alongside certain progressions in the variety also.

## IV. METHODOLOGY

Some of the techniques that have been used in the project so far have been described below. These techniques are some and still in process and some other techniques shall also be included along with this too better understand the clustering techniques using unsupervised machine learning techniques.

### *A. RFM or Recency Frequency and Monetary*

RFM technique is one of the unsupervised clustering machine learning techniques where it calculates and cluster each of the data based on the three factors that is recency, frequency and monetary. Recency means how recent a customer bought the product, frequency means how frequently a customer is buying a products and monetary means how much money is the customer is spending. Based on these three factors customers or data is clustered into different types of classification.

### *B. Hierarchical Clustering*

Hierarchical clustering is one of the unsupervised machine learning techniques to cluster data into similar types of clusters based on the factors. The basic working of the clustering technique is as follows. In the beginning, each data point is assumed to be a separate cluster and after each level of clustering the closest clusters or data point are clustered into one bigger cluster until a single cluster is formed. Dendrogram is used to visualize the cluster made at each of the clustering stages. Hierarchical clustering is of different types of single linkage, complete linkage and average linkage.

### *C. K – Means Clustering*

K- Means clustering is another type of unsupervised machine learning clustering technique. It uses some other techniques which is used to calculate the optimal number of cluster to be made for the dataset. Some of these include Elbow method, Silhouette Score, and others to calculate the optimal numbers of clusters. K-Means uses the a number of observations into k clusters where of the observation belongs to the nearest cluster.

### *D. DBSCAN (Density Based Spatial Clustering of Application with Noise)*

DBSCAN is another is another unsupervised machine learning technique used for clustering. As the name of the techniques suggest, it is a density-based clustering technique where a group of datasets in some space, it group together that are close to each other making an outline which separate the low-density data point from high data points.

## V. EXPLORATORY DATA ANALYSIS

In the above section, we will be analysing the result for each of the different types of unsupervised machine learning techniques that we have applied to the dataset. We will further also give the time taken by the machine to run and generate the output for it. Further, we will be comparing and analysing the result for each of them. Each technique will have a special relation and value to the model and each technique will be unique to its own prospective.

## VI. TESTING AND RESULTS

After completing the visualising for the different types of techniques that we applied to the dataset we will compare the result of them and further give the conclusion to each of the technique.

## VII. CONCLUSIONS

This section will contain the conclusion made for each of the technique and how accurate was the technique into segmenting customers into different clusters. Further, it will also contain time taken as well as how each technique differs from one another into making counter evidence and corroborating each others result.

## REFERENCES


[1] Customer segmentation using unsupervised clustering algorithm Sharma Sheelesh Kumar, Professor, Pathak Garvit, MCA Student, Chandralo and MCA Student, Kumari Sweta Department of IT, Institute of Management Studies, Ghaziabad, Uttar Pradesh, India
URL: `https://researchgate.net/publication/338471736_Prediction_of_Phone_Prices_Using_Machine_Learning_Techniques`

[2] Customer Segmentation using K-means Clustering Tushar Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury Dept. of Informatics, School of Computer Science, Dehradun.
URL: `https://ieeexplore.ieee.org/abstract/document/8769171`

[3] E-commerce Customer Segmentation via Unsupervised Machine Learning Boyu Shen ACM Digital Library
URL: `https://dl.acm.org/doi/abs/10.1145/3448734.3450775`