**Online Retail Customer Segmentation**

Pravin. M, Data Science Trainee, Almabetter, Bangalore

**Abstract:**

Customer Segmentation is one of the most prominent activity done by almost every organization across the world. Especially for those who are in B2C business customer segmentation becomes inevitable. Customer Segmentation has many advantages. But most important advantage was it enables the business to know better their customers. And able to sell right products at right price to the right customers. This enables us to build a model which is capable of identifying and clustering the customers into various segmentation.

**Keywords:**

1. Online Retail Customer Segmentation,
2. Customer Segmentation,
3. Retail Domain,
4. Descriptive Analytics,
5. Exploratory Data Analysis,
6. Univariate Analysis,
7. Bivariate Analysis,
8. Multivariate Analysis,
9. Statistical Analysis,
10. RFM Analysis,
11. RFM Modelling,
12. RFM based Segmentation,
13. Unsupervised Learning,
14. Distance based Clustering,
15. Hierarchical based Clustering,
16. Density based Clustering,
17. Elbow Curve,
18. Silhouette Score,
19. Dendogram and,
20. Hyper parameter tuning.

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **Sl.NO** | **CONTENTS** | **PG.NO** |
| 1 | Problem Statement | 2 |
| 2 | Business Goal | 3 |
| 3 | Introduction | 3 |
| 4 | Dataset Description | 4 |
| 5 | Data Wrangling | 4 |
| 6 | Descriptive Analytics – Exploratory Data Analysis | 5 |
| 7 | Statistical Analysis – RFM Analysis | 8 |
| 8 | Data Pre-processing | 9 |
| 9 | Unsupervised ML Analysis – Clustering and Segmentation | 9 |
| 9 | Results Comparison | 11 |
| 10 | Conclusion & Recommendations | 12 |
| 11 | References | 12 |

**Problem Statement**

In this project, our task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

**Business Goal**

The company mainly sells unique all occasion gifts. Many of the customers of the company are wholesalers. Our main goal is to segment all these wholesale customers into various segments. So that we can be able to know the taste of wholesalers. Also the demand of each products based on region, holidays, hours, years, months and on festive seasons.

**Introduction**

Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser or a mobile app. Consumers find a product of interest by visiting the website of the retailer directly or by searching among alternative vendors using a shopping search engine, which displays the same product's availability and pricing at different e-retailers. As of 2020, customers can shop online using a range of different computers and devices, including desktop computers, laptops, tablet computers and smartphones.

An online shop evokes the physical analogy of buying products or services at a regular "bricks-and-mortar" retailer or shopping centre, the process is called business-to-consumer (B2C) online shopping. When an online store is set up to enable businesses to buy from another businesses, the process is called business-to-business (B2B) online shopping. A typical online store enables the customer to browse the firm's range of products and services, view photos or images of the products, along with information about the product specifications, features and prices.

As there are many vendor out there in the internet. And internet penetrability across various population in globe enable us to sell various types of products to different population of customers across globe. Internet not only changes the nature of retail business but also the consumer interaction with the business. So to have high engagement with the customers. It is inevitable for the business to understand the taste of their customers. This enables us to build a clustering model which is capable of clustering the whole customers of the business and align each other based on their taste, frequency, country and so on. The ultimate goal of the business is to yield high profit margins from their products. So in this project the clustering will be done based on three prominent factors such as Recency, Frequency and Monetary.

**Dataset Description**

Attribute Information:

* **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
* **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* **Description**: Product (item) name. Nominal.
* **Quantity**: The quantities of each product (item) per transaction. Numeric.
* **InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
* **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
* **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
* **Country**: Country name. Nominal, the name of the country where each customer resides.

**Data Wrangling**

* **Importing & Loading Datasets**: Importing the online retail customer datasets from google drive and loading it into the colab notebook.
* **Unique Value Identification**: After loading the dataset. Unique value identification of all variables are done. This will help us in prior planning for data engineering.
* **Missing values Diagnostics**

This test is conducted to identify whether there is any missing values in the dataset. And if so it is treated in beforehand to avoid facing any issues during EDA and ML Modelling. From this diagnostics we found that Customer ID variable has some missing values.

* **Missing Values Imputation**

From the previous test we found that Customer ID has missing values and it is imputed by means of replacing all the NaN values by most frequent datas.

* **Data Imputation**: It involves modification and correction in certain variables as well as values.
* Invoice No variable is first converted into string datatype and then values starting with “C” are removed.
* In Quantity variables there are some negative values. It is then removed from the dataset.
* Similarly in Unit Price some variable has negative values it is also then removed from the dataset.

**Descriptive Analytics – Exploratory Data Analysis**

Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question “What happened?” (or What is happening?), characterized by traditional business intelligence and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives. After data cleaning up. The dataset is ready for exploratory data analysis. Exploratory Data Analysis for the credit card default prediction consists of 3 sections as follows.

**Univariate Analysis**

It is the simplest form of analysing data. Uni means one, so in other words your data has only one variable. It doesn’t deal with causes or relationships and its major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

Here the univariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. Variables involved in analysing are as follows

* **Quantity**
* High Quantity Purchase – Identifying top 15 high quantity purchase done by the customers.
* **Unit Price**
* High Unit Price – Identifying top 15 high unit price value of purchase done by the customers.
* **Country**
* High Customer Population Countries – Identifying top 15 countries with high customer population.
* Medium Customer Population Countries – Identifying countries with medium customer population.
* Low Customer Population Countries – Identifying countries with low customer population.
* **Description**
* High Demand Products – Identifying top 15 high demanded products from the customers.
* Low Demand Products – Identifying bottom 15 low demanded products from the customers.
* **Invoice Date**
* Day-wise Purchases – Identifying peak purchase days.
* Month-wise Purchases – Identifying peak purchase months.
* Year-wise Purchases – Identifying peak purchase years.
* Hour-wise Purchases – Identifying peak purchase hours.
* **Stock Code**
* Most Frequent Stock Code – Identifying top 15 most frequent stock code.
* Less Frequent Stock Code – Identifying bottom 15 less frequent stock code.

**Bivariate Analysis**

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

Here the Bivariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. The main focus of this analysis is to find significant relationship between the variables. The categories or segments of analysis involves

* **Customer Base Analysis**
* High Customer base Countries – Identifying high quantity purchasing customers.
* Medium Customer base Countries – Identifying moderate quantity purchasing customers.
* Low Customer base Countries – Identifying low quantity purchasing customers.
* **United Kingdom Customer Analysis**
* Peak Purchase Day in UK – Identifying peak purchase day in UK.
* Peak Purchase Months in UK – Identifying peak purchase months in UK.
* Peak Purchase Hour in UK – Identifying peak purchase hour in UK.
* **Saudi Arabia Customer Analysis**
* Peak Purchase Day in Saudi Arabia – Identifying peak purchase day in Saudi Arabia.
* Peak Purchase Months in Saudi Arabia – Identifying peak purchase months in Saudi Arabia.
* Peak Purchase Hour in Saudi Arabia – Identifying peak purchase hour in Saudi Arabia.

**Multivariate Analysis**

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle. This type of analysis is almost always performed with software as working with even the smallest of data sets can be overwhelming by hand.

Here the multivariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. This analysis is focused on finding correlation between variables like

* **Demand Analysis**
* Number of Holidays – Identifying number of holidays.
* Peak Hours on Holiday – Identifying peak hours on holiday.
* Most Wanted Products on Holiday – Identifying most wanted products on holiday.
* Least Wanted Products on Holiday – Identifying least wanted products on holiday.
* **Festive Season Analysis**
* Most Wanted Products on Festive Season – Identifying most wanted products on festive season.
* Least Wanted Products on Festive Season – Identifying least wanted products on festive season.
* **Sales Analysis**
* Average Sales of each Products – Identifying average sales of each products.
* Top Country-wise Sales – Identifying top sales based countries.
* Bottom Country-wise Sales – Identifying bottom sales based countries.
* Most Demanded Products – Identifying most demanded products.
* Least Demanded Products – Identifying least demanded products.
* Daily Sales – Identifying daily sales.
* Monthly Sales – Identifying monthly sales.
* Yearly Sales – Identifying yearly sales.
* Hourly Sales – Identifying hourly sales.

**Unsupervised ML Analysis – Clustering and Segmentation**

**Statistical Analysis – RFM Analysis and Segmentation**

RFM (Recency, Frequency, and Monetary) analysis is a customer segmentation technique that uses past purchase behaviour to divide customers into groups. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

* **Recency (R)**: Days since last purchase
* **Frequency (F)**: Total number of purchases
* **Monetary (M)**: Total money this customer spent. We will create those three customer attributes for each customer.

**RFM Modelling**

It involves creating recent date for calculation of the R, F, M scores. And applying lambda function on Invoice Date, Invoice Number and Total Amount variable to calculate the R, F, M scores. And make them into a data frame. After modelling part done. It is visualized to understand the nature of R, F and M values. From visualization we came to know that there is some sort of right skewness is present in the data. And also we had performed correlation analysis between R, F and M. To find any correlation between themselves.

**RFM Segments Modelling**

The wholesome data frame is split into 4 by segregating them based on quantiles. And each quantiles are allocated with some ranks ranging from 1 to 4. These ranks are allocated based on the correlation analysis. From correlation analysis, we found that both F & M are positively correlated to each other. And R & F are negatively correlated to each other. So for calculating R score the ranks should be like 0.25:1, 0.5:2, 0.75:3 and else: 4. The ranks for F & M should of inverse in nature to that of R Scores. And then RFM Groups and RFM Scores are calculated. RFM Scores are calculated by summing all the values. For RFM Groups it is calculated by grouping all together.

**Customer Segmentation based on RFM Analysis**

After calculation of RFM Scores and RFM Groups. Based on the available data we can broadly classify the customers into 4 segments / clusters. They are as follows:

* **Best Customers**

These are the customers that bought recently, buy often and spend a lot. The customer segmentation is achieved by grouping up all the customers who has the RFM Group = 111.

* **Loyal Customers**

These are the customers. Whom has a high frequency purchase history with us are considered to be loyal. The customer segmentation is achieved by grouping up all the customers who has the R = 1.

* **New Customers**

These are the customers. Whom purchased from us recently and those purchase are of high amounts are considered to be new customers. The customer segmentation is achieved by grouping up all the customers who has the RFM Group = 141.

* **Lost Customers**

These are the customers. Whom used to buy frequently from us, and at one point they spent a lot, but they’ve stopped. The customer segmentation is achieved by grouping up all the customers who has the RFM Group = 411.

**Data Pre-processing**

From statistical analysis the R, F and M values are right skewed in nature. It is solved by means of log transformation. After log transformation the dataset is standardized by means of standard scaler.

**Machine Learning Analysis – Clustering and Segmentation**

* **K-Means Clustering**

K-means clustering is a method of vector quantization, originally from signal processing that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. K-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

As a part unsupervised ml analysis implementation. I have initiated the k-means clustering model with random parameter and fitted the dataset to the initiated model. And then it is trained on the dataset. After training the model is allowed to make clusters the whole dataset based on its training knowledge acumen. After clustering completed it is allowed to validate its quality of cluster by means of silhouette score. The results of validation scores are not optimal. So for further improvements hyper-parameter tuning is implemented for finding optimal number of clusters by means of **Elbow Curve.** Based on the results obtained from the elbow curve. The model is tuned. And once again the dataset is fitted in the final model and then allowed to make clusters. And it is allowed to validate its quality of clusters by means of silhouette score.

* **Agglometric Clustering**

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.

As a part unsupervised ml analysis implementation. I have initiated the Agglometric Clustering model with random parameter and fitted the dataset to the initiated model. And then it is trained on the dataset. After training the model is allowed to make clusters the whole dataset based on its training knowledge acumen. After clustering completed it is allowed to validate its quality of cluster by means of silhouette score. The results of validation scores are not optimal. So for further improvements hyper-parameter tuning is implemented for finding optimal number of clusters by means of **Dendogram and Dendogram with Threshold.** Based on the results obtained from the dendogram with threshold. The model is tuned. And once again the dataset is fitted in the final model and then allowed to make clusters. And it is allowed to validate its quality of clusters by means of silhouette score.

* **DBSCAN Clustering**

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbours), marking as outliers points that lie alone in low-density regions (whose nearest neighbours are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

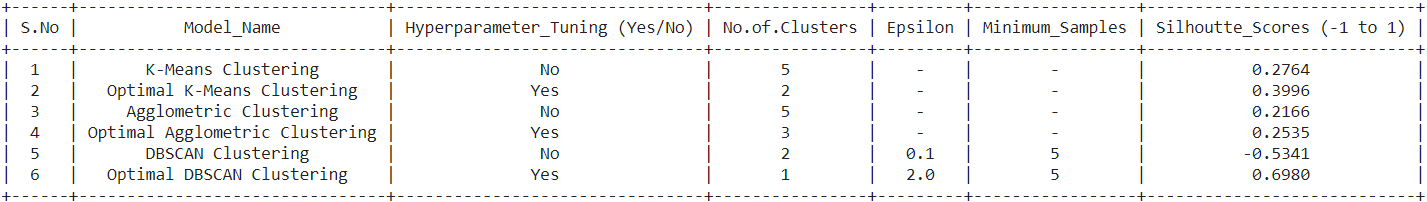
As a part unsupervised ml analysis implementation. I have initiated the DBSCAN Clustering model with random parameter and fitted the dataset to the initiated model. And then it is trained on the dataset. After training the model is allowed to make clusters the whole dataset based on its training knowledge acumen. After clustering completed it is allowed to validate its quality of cluster by means of silhouette score. The results of validation scores are not optimal. So for further improvements hyper-parameter tuning is implemented for finding optimal number of clusters by means of **Optimal Epsilon Identification and Optimal Minimum Samples Identification.** Based on the results obtained from the Optimal Epsilon Identification as well as Optimal Minimum Samples Identification. The model is tuned. And once again the dataset is fitted in the final model and then allowed to make clusters. And it is allowed to validate its quality of clusters by means of silhouette score.

**Results Comparison**

* **Silhouette Score**

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

Here silhouette scores are considered for accessing the quality of the clusters. Based on the silhouette scores of each models. It is compared with each other’s.

****

**Conclusion**

Conclusion & Recommendations are made from above result tabular column.

* For Online retail customer segmentation, I had implemented 3 clustering algorithms such as K-Means Clustering, Agglometric Clustering and DBSCAN Clustering.
* Each clustering algorithms are of unique types.
* K-Means Clustering algorithms are conventional type clustering and it works on distance based clustering.
* Agglometric clustering algorithms are hierarchal type clustering and it works on bottom - up approach.
* DBSCAN clustering algorithms are density type clustering and it works on density of object based clustering.
* Silhouette Score is used as accessing the quality of the clustering model implemented in this notebook.
* The Range for Silhoutte Score is between -1 to 1.
* From the above Result Comparison tabulation. We can able to notice that hyperparameter tuning had helped to model to perform better.
* The Model which has highest silhouette score was "Optimal DBSCAN Clustering" but it lacks in effective clustering.
* So the next highest value is for "Optimal K-Means Clustering".
* I would suggest that Optimal K-Means Clustering model would perform well on Online Retail Customer Segmentation.

**References**

* Wikipedia - <https://en.wikipedia.org/wiki/Online_shopping>