Customer-Segmentation

Auti Rupali
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Knowing the customer is the groundwork for starting or improving any product or service-related business. In this digital era activities of all the customers in every field are getting registered. Artificial intelligence has a great opportunity in the retail sector to read these experiences of customer activity and extract different patterns.

Our model for customer segmentation of a UK-based and registered non-store online retail was an attempt to extract the patterns of customers.

1. Problem Statement

We were provided with an unlabeled dataset on the transaction details of online retail customers. All the transactions were 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 were wholesalers.

Our task was to explore and analyze the data and to build a clustering model for customer segmentation.

2. Introduction

Machine learning is the only solution to learn from a huge no. of experiences in the shortest possible time. That is the reason, nowadays application of the ml model in the retail sector is at its peak.

In this project, we tried to create the best clustering model for online retail customer segmentation from the available transactional experiences of a UK-based and registered non-store online retail.

3. Steps Involved

Following steps were involved during this Unsupervised ML (Clustering) Project

3.1. **Connection with the Data**

The dataset was actually a collection of 5,41,909 experiences about transactions of online retail customers with 8 features or dimensions.

We needed to decode the set of experiences to build a model for customer segmentation.

At first, we imported the libraries or functions for making our journey easy and then got connected to the set of experiences.

3.2. **First Feelings of the Data**

When we saw the head of the data, we could understand what the set of experiences was all about. Then we tried to understand the features of the experiences.

3.3. **Deeper Understanding of the Data**

As there was a huge no. of experiences, we took the help of statistics to measure each and every feature in different dimensions, and thus step by step, we found the most important features or the exact way to decode the experiences.

“what gets measured gets done“.

3.4. **Cleaning the Data**

We removed 5,268 duplicate experiences from our dataset.

We dropped 1,35,037 null values in ‘CustomerID’ column as we could not impute them.

There were negative values in ‘Quantity’ column. We removed the experiences with negative values in ‘Quantity’.

There were zero values in ‘UnitPrice’ column. We removed the experiences with zero values in ‘UnitPrice’.

We created ‘Mean_UnitPrice’ and ‘Sum_Quantity’ features for each ‘CustomerID’ because, from the average unit price, we could understand what they buy, and from the sum of quantity, we could understand how much they buy.

We also checked the statistics several times on clean data to confirm the completion of processing.

3.5. **Treating Anomalies in the Data**

While we were finding out the general formula from the experiences, we were supposed to identify the true outliers or exceptional or abnormal experiences and keep them aside.

There were outliers for Mean_UnitPrice> 250 and Quantity> 10000 and thus we removed these experiences from our dataset.

3.6. **Final Feature Selection from the Data**

We needed to understand the distribution of the features and the relationship among the features for the decision of transformation, scaling, and final selection of features.

Here, the distribution of 'Mean_UnitPrice' and 'Sum_Quantity' was positively skewed. Thus we did log transformation on these features to normalize their distribution.

Here, all our input variables were truly independent as all the VIF values were below 10

3.7 **Preparation of Input Data**

Finally, we prepared the inputs (X) for our model in three steps:

i. Normalization (Log transformation),

ii. Train-Test Splitting and

iii. Scaling

3.8 **Building of Model-1**

In our final KMeans Model, the best no. of clusters was 3 according to silhouette score.

3.9 **Building of Model-2**

In our final AgglomerativeClustering Model, the best no. of clusters was 3 according to silhouette score.

3.10 **Building of Model-3**

In our final GaussianMixture Model, the best no. of clusters was 4 according to silhouette score, but we are selecting 3 for comparison with other models.

4. Challenges Faced

4.1 The first challenge was to find the relevant features.

4.2 The main challenge was to decide our expected clusters of customers to decide on the requirement of new features.

4.3 It was hard to remove 'InvoiceDate' column because the date and time stamp gives lots of information about the purchase pattern.

4.4 It was challenging to decide the cutoff margin for the outlier experiences.

4.5 The most challenging work was to create a dashboard for comparing model performance.

5. Approach Used

The performance of a machine learning model depends on three factors:

5.1 Quality of Data- cleaner experiences for better learning: We gave the highest importance to the exploration and pre-processing of data to produce quality data.

5.2 Quantity of Data-more experiences for better learning: We tried to minimize the loss of data to the extent it is possible.

5.3 Quality of Model-right model and right hyperparameters for better learning: We selected the model considering the volume of clean data and type of expected output. We also tuned the hyperparameters to produce the optimum model.

6. Conclusion

Conclusions drawn were as follows:

6.1 On the basis of the performance study of our three models, we selected KMeans model for online retail customer segmentation, as it was best fitting our expected customer segmentation with minimum overlap among all three models