Customer segmentation using RFM Analysis and CLV Estimation

presented by
Shashank Gupta, Jay Trivedi, Nikhil Chavan, Charu Joshi

Project Guide: Prof. Spiros Papadimitriou
Course: Algorithmic Machine Learning

Table of Contents

1. Abstract

This project paper describes how machine learning algorithms can be useful in predicting the customer lifetime value (CLV) and analyse customer behaviour based on Recency-Frequency-Monitory value(RFM) parameters. Customer lifetime value (CLV) is a measure of the total value a customer will bring to a business over the course of their relationship with the company. While RFM analysis can help businesses understand which customers are the most valuable and identify opportunities for upselling and cross-selling.

Using machine learning techniques, it is possible to predict CLV and perform RFM analysis in a more efficient and accurate manner. This can be done by training a model on historical customer data, including information about their purchases and interactions with the business. The model can then be used to make predictions about the CLV and RFM score of new customers, allowing the business to prioritize their efforts and resources.

Overall, a project utilizes Kmeans clustering and NBD & Gamma Gamma unsupervised machine learning techniques to combines CLV and RFM analysis which provide valuable insights to businesses and help them better understand and serve their customers.

2. Introduction

2.1 Problem Statement

An ecommerce company has collected the transaction data of its customer over years and company wants to utilize the data for increasing the revenue and segmenting the customer for targeted marketing and application testing (beta testing)

2.2 Objective

In this project we will define customer segmentation and marketing matrix by analysing the customer transactional data using machine learning algorithms. CLV and RFM are two widely used matrices in the industry for this.

2.3 What is Customer Lifetime value (CLV) and why it is important?

Customer Lifetime Value (CLTV or CLV) can be defined as the estimated net profit the customer will contribute during their future relationship with the company. It represents the total amount of money a customer is expected to spend in business, or on products, during their lifetime. CLTV provides a picture of the business long-term and its financial viability.

Knowing each customer’s customer lifetime value helps you know how much you should be spending on customer acquisition. A customer’s acquisition cost could be more than what they spend on their purchase, but if you nurture that relationship, their CLV may grow to an amount that’s well worth the investment. That’s just one of the many reasons why success in the customer-centered economy means understanding the importance of customer lifetime value.

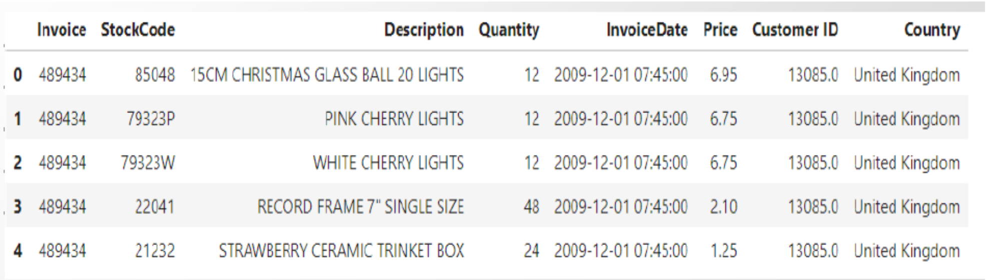
CLV-based Customer Segmentation helps with more effective segmentation. Clusters of customers can be identified based on the long-term revenue potential of a customer. Customized offers & products can be designed to unlock maximum customer value. CLV is one of the key metrics to track the performance of the company as compared to the competitors.

A high value of CLTV indicates product-market fit, the brand’s goodwill, and expected recurring revenue from the existing customers. Estimating CLTV requires RFM analysis, where RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait.

- **Recency:** Refers to the freshness of customers, be it visits & purchases.
- **Frequency:** Refers to the frequency of customer transactions or visits.
- **Monetary:** Refers to the intention of customer to spend.

3. Data source

We have used the customer segmentation data set available at UCI repository and Kaggle for this project. The dataset includes transaction details of an e-commerce retail store for the period 2009 and 2010. Cursory look at the data shows that it holds the transactional information of the items purchased by each customer from different geographic regions. After performing basic statistical operations, we found the dataset has around 1 million rows including null values for some of the transactions.



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

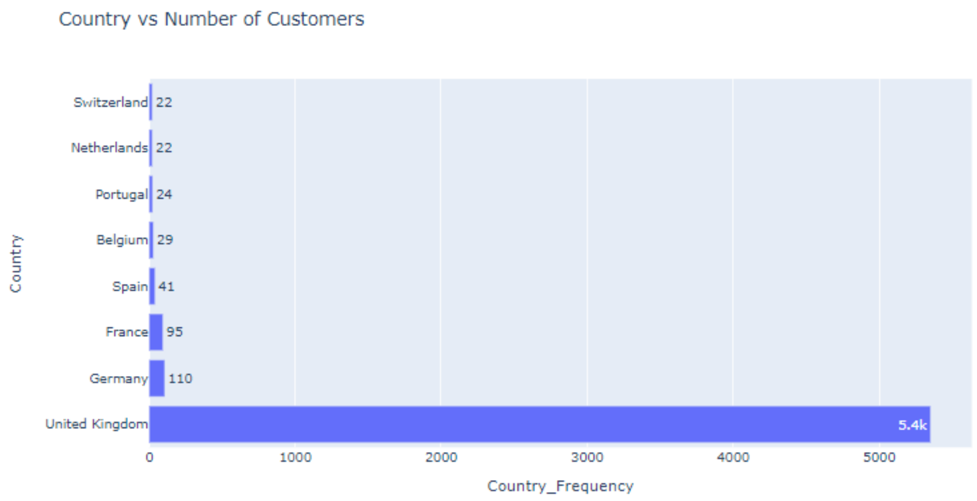
- **Invoice No**: Invoice number. Nominal. A 6-digit integral number is uniquely assigned to each transaction. If this code starts with the letter ’c,’ it indicates a cancellation.
- **Stock Code**: Product (item) code. Nominal. A 5-digit integral number is uniquely assigned to each different product.
- **Description**: Product (item) name. Nominal.
- **Quantity**: The quantities of each product (item) per transaction. Numeric.
- **Invoice Date**: Invoice date and time. Numeric. The day and time when a transaction was generated.
- **Unit Price**: Unit price. Numeric. Product price per unit in sterling.
- Customer ID: Customer number. Nominal. A 5-digit integral number is uniquely assigned to each customer.
- **Country**: Country name. The name of the country where a customer resides. Nominal.

4. Approach

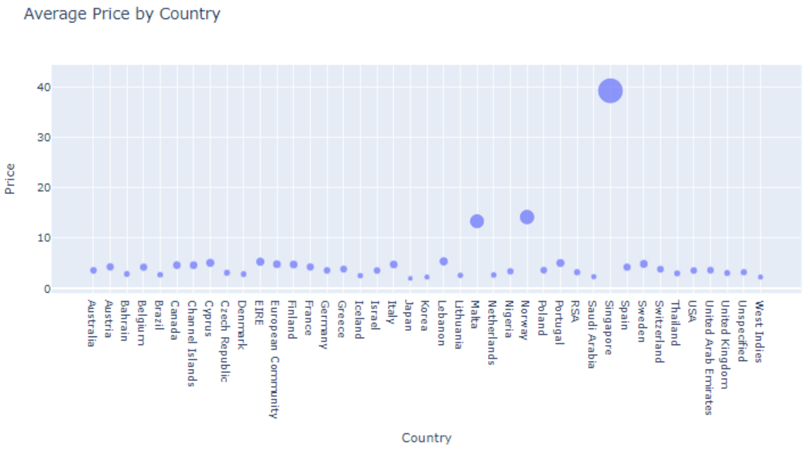
4.1 Exploratory Data Analysis

Exploratory Data Analysis revealed significant trends in customers of different countries. The findings are as follows:

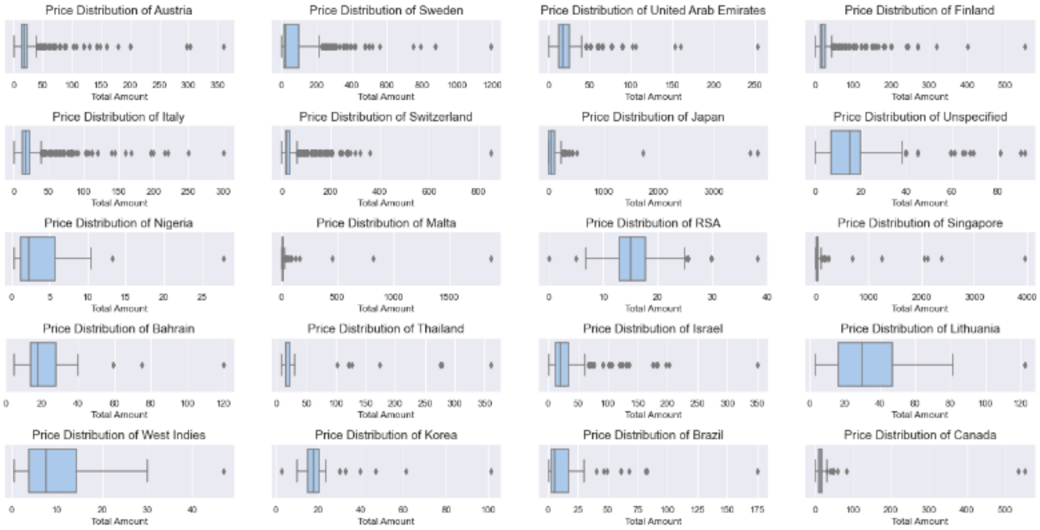
A. **Customer by Countries**: We were interested in finding out the demographics of the customers. As per the below chart, the majority of the customers are situated in United Kingdom. Minorities are nowhere near UK’s customer base.



B. **Customer by Countries**: Now, we are interested in finding out the average price of orders by the customers. Following is a bar chart explaining the same:



C. **Price Distribution**: Finally, we looked at the total amount spent on customers' distribution in each country.



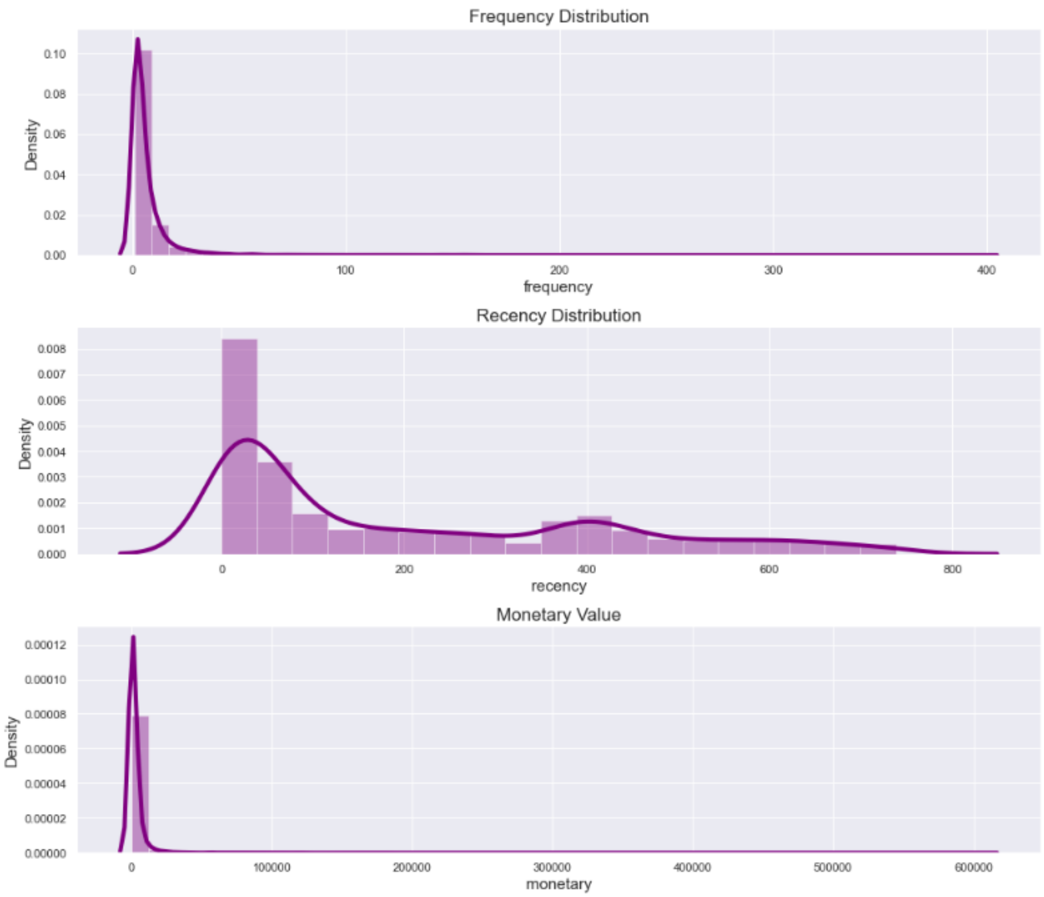
EDA observation

- Majority of the customers are situated in United Kingdom.
- Singapore, Norway, and Malta have highest average price distribution.
- Price Distribution data looks highly skewed, indicating both small and big purchases.

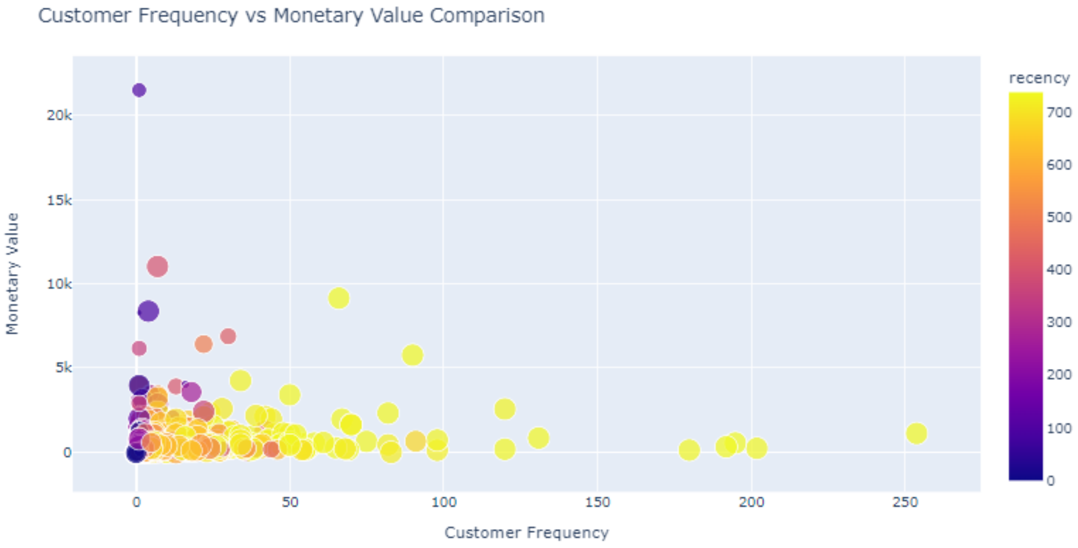
4.2 RFM Analysis & Feature Engineering

After completing EDA, we proceed with the RFM analysis. In this section, we extracted the key features - Recency, Frequency, and Monetary Value via feature engineering.

The distributions of extracted RFM attributes are shown as follows:

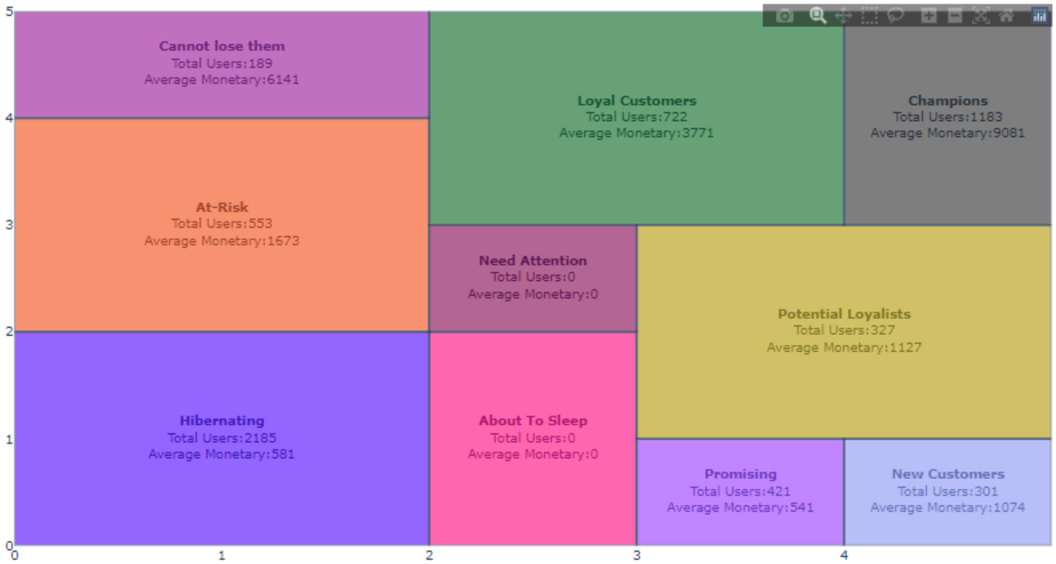


Frequency and Monetary values are following the pareto distribution which was expected. Next, we prepared a scatter plot between Customer Frequency and Customer Monetary value, keeping recency on heatmap and overall tenure on size.



4.2.1 Customer Segmentation using Recency & Frequency

We performed attribute discretization on the RFM features and created a score matrix based on the distribution. Further, we performed the customer segmentation using the Recency and Frequency attributes. Following is the plot we obtained:



4.2.2 Key takeaways of RFM Analysis

- Some customers have High Monetary value but low frequency indicating bulk orders made by the customers.
- Some customers have High Frequency but low Monetary value indicating small but frequent orders made by the customers.
- Only a few customers have balanced frequency, monetary, and recency values.
- Champions and Hibernating customers dominate the segmentation plot.
- Champions have high average monetary scores while Hibernating customers have low monetary value.

4.3 Customer Segmentation using Customer Lifetime Value (CLTV)

In this section, we estimate the CLTV first using the RFM features, and BetaGeoFitter and NBD pareto model imported from the lifetimes package of Python. These models help in estimating the CLTV based on the following equation:

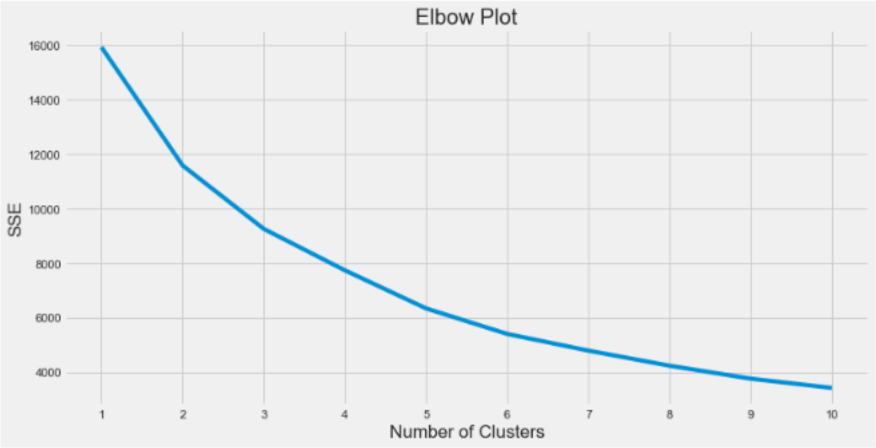
$$\text{CLTV} = \begin{matrix}\text{Average Order}\\ \text{Value}\end{matrix} \times \begin{matrix}\text{Average Repeat}\\ \text{Purchases}\end{matrix} \times \begin{matrix}\text{Average Time}\\ \text{Customers Stay}\end{matrix}$$

BetaGeoFitter and NBD/Pareto Models are the most used probabilistic models for predicting the customer's lifetime value. The NBD Pareto model predicts the future transactions of every customer. The output of the NBD Pareto model is combined with the Gamma-Gamma model, which adds the monetary value aspect of the customer transactions and returns the Customer Lifetime Value (CLV).

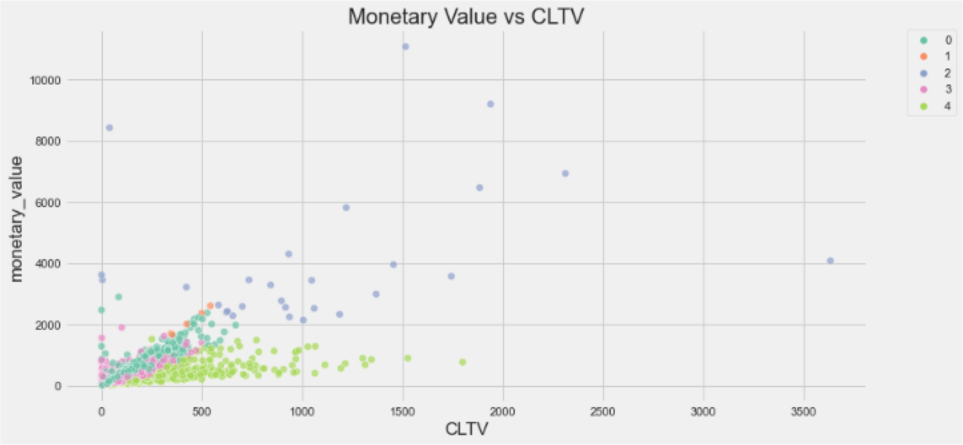
After calculating the CLV estimates, we performed customer segmentation, this time using the Recency, Frequency, Monetary Value, Customer Tenure, and estimated CLTV. We used the K-Means algorithm to segment clusters.

4.3.1 K-Means

The aim of using K-means is to group similar customers. We first performed normalization on data and then clustering using different K values in K-means and selected the optimum K using the Elbow Plot and knee locater from Python's sklearn and kneed package of Python. Following is the obtained Elbow plot:

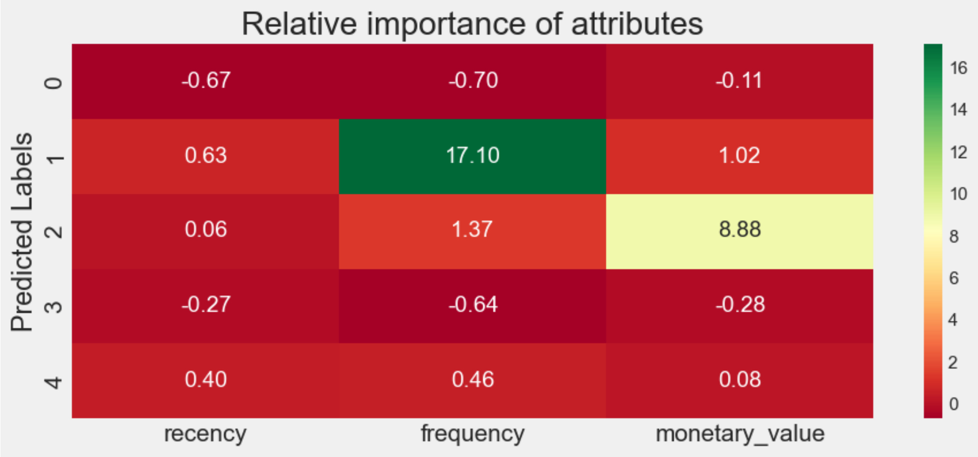


Following is the scatter plot obtained from K-means Clustering



4.3.2 Evaluation based on relative importance metrics

Based on the formed clusters, we calculated relative importance for all the RFM features for each cluster using the cluster average RFM divided into individual RFM values and achieved the following Relative Importance of attributes in a heatmap:



4.3.3 Key takeaways from Relative Importance Heatmap

- Cluster 1 has the highest Frequency score indicating room for improvement in Recency and Monetary value estimates.
- Cluster 2 has the highest relative Monetary Value score indicating room for improvements in recency and frequency estimates.
- Clusters 0, 3, and 4 do not have enough RFM scores and hence, require improvement in all RFM features.

5. Conclusion

Customer segmentation based on CLV and RFM values using K-Means clustering is more comprehensive approach as it eliminates the redundant clusters as we saw in manual RFM analysis. The relative importance heatmap easily helps us choose the right cluster based on the requirements of the use case the ecommerce company is trying to develop. Based on our analysis, we found the optimum number of clusters while the clustering parameters can further refine by performing hierarchical clustering techniques.