CUSTOMER SEGMENTATION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

Ms. ANITHA L(ASSISTANT PROFESSOR,CSE)

CHITHRA.K (LRCE18CS066)

MEGHA.C (LRCE18CS070)

RASHID.T.E (LRCE18CS072)

NASLI SHAMSUDHEEN(RCE18CS036)

ABSTRACT

In carrying out successful E-Commerce, the most important things are innovation and understanding what customers want. Now-a-days the ease of using e-commerce encourages the customers to buy using e-commerce. It runs on the basis of innovation having the ability to enthrall the customers with the products, but with such a large raft of products leave the customers confused of what to buy and what not to. According to business, a company may create three segments like High (Group who buys often, spends more and visits the platform recently), Medium (Group which spends less than high group and is not that much frequent to visit the platform) and Low (Group which is on the verge of churning out). This is where Machine Learning provides a crucial solution, several algorithms are applied for revealing the hidden patterns in data for better decision making. In this paper we proposed a customer segmentation concept in which the customer bases of an establishment are divided into segments based on the customers' characteristics and attributes. This idea can be used by the B2C companies to outperform the competition by developing uniquely appealing products and services and making it reach potential customers. This approach is implemented using "K-Means" and "C-Means", which are unsupervised clustering machine learning algorithms.

OBJECTIVE

In this Project we applied the basic analytics functionality to provide the decision makers. with the required information to make the right decision. Here we define a solution for reducing risk factors and also contribute to decision making for new business investments. We proposed to use the K-Means and C -Means technique for customer segmentation. Our solution is to segment the customers based on information analytics.

FUNCTIONAL REQUIRMENTS

Hardware Requirements

Processor : IntelCorei5orbetter

CacheMemory : 6 MB or more

Memory : 4 GBRAM or Above

HardDisk : 80GB or above

DisplayType : SVGA Color Monitor

Keyboard : Enhanced 104 Standard

Mouse : USB 2.0,2 Button

Wi-FiAdapter : Broadcom 4313 GN802.11b/g/n 1x1

Software requirements

- OperatingSystem : Any OS
- IDE : Visual studio
- Front end : Python
- Back end : Python,SQL
- Database : MySQL5.5

SYSTEM MODULES

Data Pre-processing

Data pre-processing involves transforming raw data to well-formed datasets. So that data mining analytic can be applied. It refers to manipulation or dropping of data before it is used to ensure or enhance the performance. In this project the dataset which is an excel file is loaded using pandas and the duplicate values in the entity are dropped. We dropped all the duplicate entries in CustomerID and Country column. Count of customers for each country is calculated and sorted in descending order to see from which country the maximum number of customers purchases. From the first few rows we observed that the maximum number of customers are from the UK in Country. So we grouped the customers based on country. So we filtered out other countries using the “query” method.In the consolidated dataset, only the Description and CustomerID columns had null values. Those entries are dropped off. Also the negative entries in Quantity as it can not be negative. Invoice_Data which is in string format and that is converted to date and time format as it is necessary while calculating Recency. A new column Total_Amount is added as a product of Quantity and Unit_Price for each customer. This is useful in Monetary calculation.

RFM-Score Calculation

Our dataset is limited to sales records, we can use a RFM based model for finding

segments where R is Recency (how recently a purchase happened), F is Frequency (how frequent transactions are made), M is Monetary value(Value of all transactions).Recency, Frequency and Monetary score for each customer is calculated. The latest date is assigned as a placeholder to calculate recent purchases. All the transactions are grouped using CustomerID and then aggregate lambda operations are performed. As a result of this operation numbers will be obtained which depicts the recency. Frequency and how much a specific customer spent till date. All these are stored in a new dataframe RFMscores. To note, the for recency is right skewed. The individual recency , frequency and monetary values are concatenated and converted to string using map function. This is done to easily check which group the customer belongs to.

This RFM score column shows the loyalty of engagement of the customer. In our case,the lower the value of RFM score, the more loyal the customer will be as well as more engaged he/she would be. Based on this in the next step

Loyalty_Level like Platinum,Gold, Silver and Bronze levels are assigned to each customer. From this we could derive a conclusion that if the customer is in the platinum group we can say that they are the best customers whereas in the bronze group, the customer hasn't purchased for a longer time. With this a company can decide to provide special attention, offers and priority access to newly launched products to their platinum customers. On the other hand, if the customer falls into the bronze group, the company can give some.

Clustering

K-Means

K-Means is an unsupervised learning algorithm used for clustering tasks which works really well with complex datasets. It is an iterative algorithm that partitions the dataset into “K” pre-defined non overlapping subgroups (clusters) where each data point belongs to only one group.

The algorithm works as follows:

- Step-1 : Specifying the number of clusters – k value.
- Step-2 : Centroids are initialized by shuffling the dataset and then randomly selecting k data points for the centroids without replacement.
- Step-3: Repeat the iteration until there is no change to the centroids. i.e, assignment of data points to the clusters does not change.Recency, Frequency and Monetary are brought to the same scale and the data is normalized before clustering process. It is important to determine the optimum number of clusters i.e, “k value”. For this we used the Elbow Method. It involves running the algorithm multiple times over a loop with an increasing number of cluster choices and then plotting a score as a function of the number of clusters. When “k” increases, the centroids are closer to cluster centroids. The improvement will decline at some point rapidly creating an elbow-like shape ingraph and that is the whole reason this method is called as elbow. We take the count of clusters, k-value at the point where this elbow is bending.

C-Means

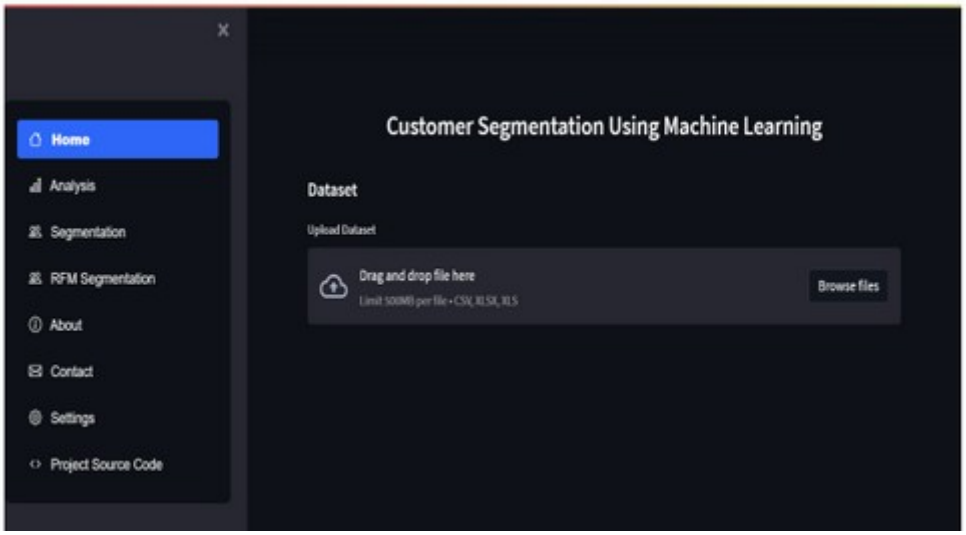
- Step 1: Set the number of clusters.
- Step 2: Initialize the fuzzy partition matrix.
- Step 3:Set the loop counter k=0.
- Step 4:Calculate the cluster centroid, calculate the objective value.

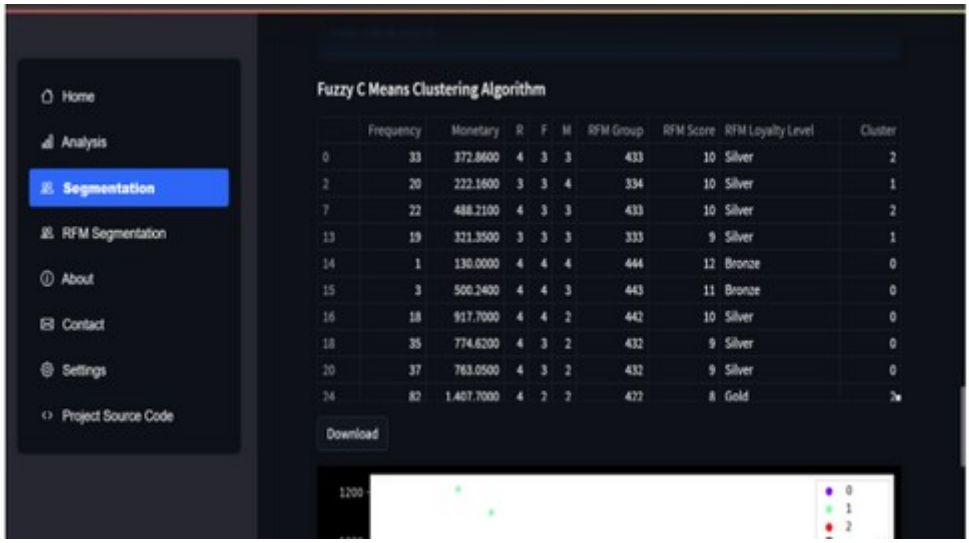
● Step 5:Compute the membership value in the matrix.

● Step 6:If the value of j between consecutive iterations is less than the stopping condition then stop; otherwise set k=k+1 and go to step 4.

● Step 7: Segmentation.

RESULTS





CONCLUSION

This work presented an implementation of the K-Means and C-Means clustering algorithm for customer segmentation using data collected from an online retail outfit.Our model has partitioned customers into mutually exclusive groups, three clusters in our case. And RFM score is also calculated and loyalty level is determined. This will be useful for applying further data mining strategies and the derived insights are helpful in decision making for the business wings.