International Journal of Recent Advances in Multidisciplinary Topics
Volume 4, Issue 4, April 2023
https://www.ijramt.com | ISSN (Online): 2582-7839

Unsupervised Recency Frequency and Monetary based Customer Segmentation

Lakshit Chauhan[1*], Kshitij Mittal[2], Garv Pratap Singh[3], Santu Mahapatra[4], Nidhi Chandra[5]

[1,2,3,4]*Student, Amity School of Engineering and Technology, Amity University, Noida, India*

[5]*Faculty, Amity School of Engineering and Technology, Amity University, Noida, India*

***Abstract*: The domain of this project is Data Science. Basically, we are given a real-life dataset of an Ecommerce company from which we are expected to draw meaningful insights related to customer purchasing behavior and recommend focused strategies to improve revenue and enhance customer retention based on the results provided by the customer segmentation model. The company majorly categorizes their clients based on one-of-a-kind business metrics which include how recently they spend or visited (recency), how often they spend (frequency), how much they spend (monetary). And this is why we use RFM approach to know the accuracy of our clustering model by comparing the silhouette score of clustering model with the RFM score which makes our model better to perform customer segmentation for the companies by categorizing their customers into Loyal Passive and Critical ones with better accuracy so that they can build the best targeted strategies to retain their customers and increase revenues.**

***Keywords*: Unsupervised, Recency, Frequency, Monetary, Customer segmentation.**

1. Problem Definition

To build an unsupervised learning model which will analyze customers via RFM approach for the given data. We are expected to draw meaningful insights from the data by customer segmentation and recommend focused strategy for each segment to improve revenue and enhance customer retention.

2. Introduction

The domain of this project is Data Science. We have taken data from an e-customer company where they required to know their loyal customer on the basis of different locations. Provided dataset contains customer-id, item-code, invoice number, date of purchase, quantity, time, price per unit, price, shipping location, canceled status, number of returns, sold as set etc, which are having different data types which can be treated performing different tests. In a broader way this data contains the date-time of sale, customer shipping location, and price of a single unit of a year. We are expected to draw meaningful insights and recommend focused strategies to improve revenue and enhance customer retention and for that we will initially start with the thorough study of data which requires understanding of each and every column and what type of value they add in calculating the result. This includes data cleaning, treatment of missing values, deletion of unnecessary fields, removal of duplicate values, through which we achieve the target of EDA (Exploratory Data Analysis) process. By this we will have a meaningful data set on which we have analyzed and performed graphical visualization for better understanding of data. On a longer run we tend to analyze data and calculate the RFM score and Silhouette score from the clustering model to get the accuracy of the model which lets us know who our loyal passive and critical customers are. This will help the organization to work upon its customers especially on the passive customers by applying different strategies and providing additional offers to them. The organization can further use this analysis to do improvements to avoid cancellation on the basis of specific locations.in Retail and E-Commerce B2C, and extra extensively in B2B, one of the key factors shaping the commercial enterprise method of a company could be information of client behavior. More especially, know-how are they new or present purchasers, what are their preferred products, and so on. Such information would in turn assist direct marketing, sales, account control and product teams to aid this client and improve the product presenting.

3. Literature Review

The basic three stages of a data science model include conceptual, which is the first stage in which the data is conceptually enforced, made into relationships, and fed into the data format. The data is fed by key points, data entities, and relationships in a logical structure in the second logical stage, and in the third physical stage, the data model idea is used. When faced with categorization issues, data scientists usually start by asking themselves, "What category does this data belong to?”. There are several justifications for categorizing data. Maybe the information is a scanned image of a written document, and you want to determine what alphabet or numeric sequence the image depicts. If the information relates to a cancer detection strategy, you could be wondering if it belongs in the "positive" or "negative" category. Other categories could be concerned with figuring out the crop's health or if a tweet is true or false. There are various algorithms and techniques that are used to categorize the data. There are several types of models called RFM that are employed here after we analyze the raw data and use algorithms to get the results (Recency

[*] Corresponding author: lakshit072001@gmail.com

Frequency Monetary Value). The whole definition of RFM is as follows: R denotes recency, the freshness of customer activity be it purchases of visits, F denotes frequency, frequency of customer transactions or visits, and M denotes monetary, the intention of customer to spend or purchasing power of customer. Data wrangling assists in dataset unification and enhances their usefulness by converting datasets into a format that is suitable with the target system. Finding and updating data that is incomplete, incorrect, redundant, or superfluous is known as data cleaning. The cohesiveness of an object with its own cluster in relation to other clusters is measured by the silhouette value (separation). An item is thought to be well matched to its own cluster and poorly matched to surrounding clusters if the silhouette has a high number. The range of the silhouette is 1 to +1.

Approach used:

RFM Customer Segmentation Approach with the K-means clustering algorithm.

As per the research, we got to know some points by which the RFM approach can be considered better one including:

RFM is much of the time liked over different models for the accompanying reasons:

Simplicity: RFM is a clear model that is straightforward and executed. It requires just fundamental information (recency, recurrence, and financial worth) and doesn't include complex factual procedures.

Actionability: RFM gives clear and noteworthy experiences into client conduct, permitting organizations to distinguish and target explicit client sections with customized showcasing systems.

Flexibility: RFM can be applied to different ventures and plans of action, making it a flexible instrument for client division and investigation.

Proven effectiveness: RFM has been broadly utilized in the advertising business for a really long time and has been displayed to create positive outcomes in further developing client maintenance, expanding deals, and helping productivity.

RFM (Recency, Frequency, Monetary) is a generally involved strategy in client division and examination, and there are a few other comparative models that can be utilized for this reason.

Here are some of them:

RFV (Recency, Frequency, Value): This model is like RFM however considers the complete financial worth of the client's exchanges rather than only the normal money related esteem.

RFE (Recency, Frequency, Engagement): This model considers the degree of client commitment, for example, the quantity of site visits, email opens, and web-based entertainment and communications, notwithstanding recency and recurrence.

CLV (Customer Lifetime Value): This model spotlights on foreseeing the complete worth a client will bring to the business over their whole lifetime, considering variables like buy history, dedication, and reference conduct.

CHAID (Chi-squared Automatic Interaction Detection): This model is a choice tree-based division procedure that distinguishes the main drivers of client conduct and fragments clients in light of these drivers.

These models can be utilized related to or as an option in contrast to RFM, contingent upon the particular business needs and objectives.

Also, we have used the k- means approach in this project for clustering the data. So why is this clustering approach used in this project and why not other approaches like fuzzy as k-means clustering is the most basic clustering approach and it is quite an old approach?

K-means clustering and fuzzy clustering are both generally involved unaided learning methods for bunching information. While the two methods enjoy their benefits and drawbacks, there are a few circumstances where k-means might be liked over fluffy bunching:

K-means is computationally less expensive: K-means is a quicker calculation contrasted with fluffy bunching as it requires less calculation time because of its easier execution. In circumstances where the dataset is exceptionally huge or the quantity of factors is high, k-means might be more plausible.

K-means has clear boundaries between clusters: K-means produces particular groups with obviously characterized limits between them, which can be valuable in applications where there is a requirement for fresh, distinct bunches.

K-means is more intuitive and easier to interpret: K-means produces bunches that are more straightforward to decipher, as every information point is doled out to a solitary group, while in fluffy bunching, information focuses are relegated participation scores showing their level of having a place with each bunch, which can be more hard to decipher.

4. Module Description

The framework of this project is kept basic. The modules or libraries used in this project are highlighted below which includes description about them and how they have been integrated in the project.

A. Python

Python is an interpreted, item-oriented, high-level programming language with dynamic semantics. Its high-stage constructed in records systems, blended with dynamic typing and dynamic binding, make it very appealing for Rapid Application Development, as well as to be used as a scripting or glue language to connect present additives collectively.

B. Pandas

Pandas is a software library used for manipulation and analysis of data. It is written for the Python programming language. Many DS (data structures) and processes for using and manipulating numerical tables and time series are provided by this library. Pandas is a free software.

C. Numpy

For working with arrays, we use a Python library called NumPy. It also has different functions, Paul working with linear algebra, Fourier transform and matrices. It is a free and open-source software. NumPy stands for numerical Python.

D. MatPlot seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

E. Matplotlib

Matplotlib is a data visualization library which is built in Python for NumPy arrays. It also works with SciPy stack. It is used for 2D plots of arrays.

F. Sklearn

Sklearn is also called or referred to as Scikit-learn. This library is used for machine learning. It consists of lots of tools for statistical modeling and machine learning techniques. It also includes classification, regression, clustering, and dimensionality reduction.

5. Methodology

A. Data Collection [1]

We searched for various datasets on every. This data contains the date-time of sale, customer shipping location, and price of a single unit from 2016 to 2017 from KPMG India.

B. Data Cleaning & Wrangling [2] [3]

The data cleaning and wrangling process is as follows:

1. Checking the Shape of Data
2. Checking for Missing Values
3. Selecting duplicate rows except first
   - 3.1. Occurrence based on all columns
   - 3.2. Print the resultant Data frame
   - 3.3. Dropping Duplicates from dataset - 8 rows were dropped
4. Fixing datatype
5. Price Column analysis
   - 5.1. Plot distribution of Price
   - 5.2. Deleting Customer ID where sum of Price is <= 0
6. Date of Purchase Analysis
   - 6.1. Range of Days in the Data Set
   - 6.2. Set Max Date to a variable
   - 6.3. Extracting Month & Year to a new variable
7. Drop Columns that are not going to be used

C. Exploratory Data Analysis [4] [5]

1. Bar Plot to Show Profile of Data - Count of Transaction, Invoice, Customer ID
   - 1.1. Analysis of Invoice Generation
2. Sales & Returns Analysis
   - 2.1. Sales by month
   - 2.2. Checking for returned products
3. Analysis of Items Sold
   - 3.1. Most productive Item codes
4. Bar plot of Activity by time intervals
5. Bar plot of Shipping Location

D. RFM Calculations & Statistical Segmentation [6]

Here we manually calculated the RFM scores and the steps involved:

1) Making column "Customer ID" as the index and dropping unwanted columns
2) RFM & Model Data - Aggregating Data by Customer ID
3) Feature Engineering - Calculating RFM values
4) Plot distribution of Recency, Frequency, Monetary
5) Statistical RFM Ranking, RFM Grouping, RFM Scoring & Segmentation
6) Plotting the RFM Groups
7) Plotting the RFM Score
8) Plotting the Customer Groups basis RFM Score
9) 3D Plot based on 9 RFM Scores

E. RFM based Clustering Model & Validation [7]

The steps are:

1. Fix Skewness - Updating outliers
2. 3D Data Plot post Outlier correction
3. Testing & Correcting Skewness
   - 3.1. Testing for skewness of the data 4. Plot distribution of Recency, Frequency and Monetary Value
4. Performing Log Transformation to correct skewness
5. Re-Testing for skewness of the data
6. Applying the Elbow Method
7. Running K-Means to our desired number of optimal clusters (k = 3)
   - 7.1. Making clusters in the backend not fitted to dataset
   - 7.2. Fitting the cluster to the dataset 8.3. cluster allocation
8. Overall Silhouette score
   - 8.1. Silhouette score of each data point
9. Cluster Membership
   - 9.1. Cluster wise mean r/f/m/sil values
   - 9.2. Pattern of sil scores across clusters

F. Modeled Cluster Profiling & Analysis

Based on above segmentation some of the derived strategies are:

1. Loyal
   - a. Provide dedicated and superior Customer Support
   - b. Do loyalty programs with enhance membership benefit-points convert to cash
2. Passive
   - a. Send them personalized emails with offers and encourage them to shop more
   - b. Offer discounts and roll out campaigns during 10 AM to 4 PM
   - c. Enroll to base tier of Loyalty Program
3. Critical
   - a. Communicate & stay in touch using personalized emails
   - b. If showing interest, roll-out cash backs
   - c. Special promotions to groups that are on the fringes

d. Roll out Satisfaction Survey

G. Recommendations and Strategy

Now as the Clustering Model suggests 3 Customer Groups which have been renamed as Loyal, Critical and Passive. So, we or the client can perform the analysis:

1) Key Metrics by Customer Group
2) Customer Engagement & Retention Strategies

6. Application

We could find the Key Metrics by Customer Grouping which will help the company or the organization for Customer Engagement & can make the Retention Strategies for the targeted customers.

7. Result and Discussion

In this, we worked towards a real-world problem statement and drew meaningful insights from data and recommended focused strategies to improve revenue and enhance customer retention. Here we derived the approach and steps that are required to build a better customer segmentation model using RFM Technique which results in better accuracy to segment customers into loyal passive and critical and hence help company to build focused business strategies.

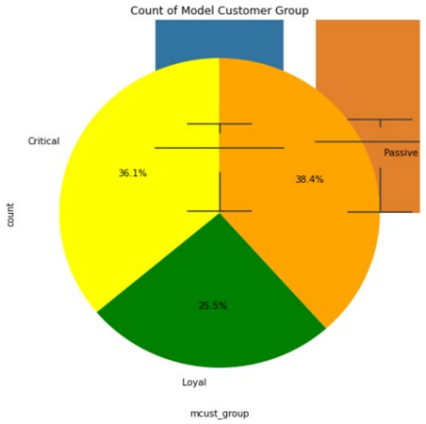

Fig. 1. Count of model customer group

8. Conclusion

For many years companies are maintaining an unstructured data of customer information and the products they purchased which is adding no value. But after doing data cleaning and wrangling, the data can be used by applying clustering models with proper analysis approach like RFM which has comparatively better accuracy and also helps in customer segmentation (Loyal, Passive and Critical) which is then used to enhance customer retention and improve revenues. The E-commerce Industry (like- Amazon, Blink-it etc.) can get the best benefit out of this project for analyzing their customers to know that who are their Loyal, Passive and Critical customers and what are the ways they can enhance customer retention and improve revenue. This can increase the success rate of any organization.

References

[1] Osborne, Jason. (2013). Best practices in data cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.

[2] Deshmukh, Ratnadeep & Wangikar, Vaishali. (2011). Data Cleaning: Current Approaches and Issues.

[3] Patil, Malini & Hiremath, Basavaraj. (2018). A Systematic Study of Data Wrangling. International Journal of Information Technology and Computer Science. 10. 32-39.

[4] Komorowski, Matthieu & Marshall, Dominic & Salciccioli, Justin & Crutain, Yves. (2016). Exploratory Data Analysis.

[5] Pramanik, Jitendra & Samal, Abhaya Kumar & Sahoo, Kabita & Pani, Dr. Subhendu. (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering. 8. 4727-4735.

[6] Kabasakal, İnanç. (2020). Customer Segmentation Based on Recency Frequency Monetary Model: A Case Study in E-Retailing. 13. 47-56.

[7] Li, Youguo & Wu, Haiyan. (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia. 25. 1104-1109.