IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 4, April 2021

DOI: 10.17148/IARJSET.2021.8401

Customer Segmentation and Customer Life Time Value

Heena Gupta[1], Nazia Afsar Chand[2]

Assistant Professor, Department of Computer Science, Mount Carmel College, Bangalore, India[1]

Student, Department of Computer Science, Mount Carmel College, Bangalore, India[2]

Abstract: Customer segmentation permits for an efficient allocation of promoting resources and maximizes opportunities to cross-sell and up-sell clients so as to extend average order price & customer period price. In today's era of availability of wide variety of products and services, customer segmentation and identifying the potential customer aids in effective business growth. This paper shows the grouping of the customers, using the k-means algorithm, based on the RFM (Recency Frequency Monetary) analysis followed by the customer life time value prediction to identify the potential customers.

Keywords: RFM, Customer Life Time Value, Machine Learning, Customer Segmentation, Clustering

I. INTRODUCTION

Customer segmentation (also referred to as market segmentation) is grouping customers into specific promoting teams, maybe narrowing them down by gender, interests, shopping for habits or demographic. Customer segmentation can be done on the basis of demographic, geographic, behaviour and psychographic factors. It is regarding over atching customers with acceptable product offers. It conjointly means you communicate together with your customers upported with what you recognize regarding them. It's regarding differentiating your most profitable customers and craft your product and services to fulfil their specific wants. Ultimately, segmentation is all about making relevant looking experiences that build whole loyalty.

In several cases, you'll be able to divide customers dynamically supported pre-set conditions like things in a customer's handcart, things purchased, or recently viewed things. Such on-the-fly segmentation would possibly customise what banner adds a specific client sees, mechanically counsel product which may be of interest to client. The result will be a lot of partaking client expertise. By differentiating their client base, businesses with higher target people and maximize sales and supply a lot of tailored searching experiences. The characteristics your business chooses to specialize in are going to be specific to your product, moreover the kind, sector and size of your business. As an example, reckoning on what you'd prefer to determine regarding your customers, you will favour to cluster them in line with demographics, location, mode and temperament, or behaviour. A lot of powerful results may be achieved by segmenting customers supported their desires and specific promoting objectives. A bank would possibly section its customers supported age, legal status, and checking or bank account balances to spot young couples who could also be considering shopping for a house and therefore searching for a home equity credit.

Customer Life Time Value (LTV) is that the price a client contributes to your business over the whole lifespan at your company. it's an awfully necessary metric and is employed where ever there are creating necessary choices regarding sales, marketing, development, and client support. Customer period worth is what quantity a client can bring throughout their entire time as a paying client. At a look, LTV tells you ways a client adds value to your business and provides your insight into their overall worth. From there, you'll have a stronger understanding of what quantity you ought to investment in client and their retention going forward. Not solely that, however client period gives you clues into whether or not you'll expect bound customers to become repeat customers. If their client period worth is high, chances are high that they're fans of your services and can still obtain a lot of your product.

II. LITERATURE SURVEY

Many papers were reviewed associated with the subject of Client Retention, Client Segmentation and personalised offers.

Initially, reviewed a paper named "Classifying the segmentation of client price via RFM model and RS theory", done by Ching-Hsue Cheng, You-Shyang Chen [1]. This paper prompt a procedure not solely to extend and improve

DOI: 10.17148/IARJSET.2021.8401

classification accuracy however conjointly to derive out the classification rules to own a superb client management. moreover, whereas reading this paper the drawbacks of information mining tools and also the ways in which of developing them were mentioned. Authors conjointly propose the ways and techniques to simply and objectively cluster the client segmentation.

Another paper referred to as "K-modes cluster algorithmic rule for Categorical Data" by N. Sharma and N. Gaud [2], gave ample data relating to the formulas and mathematical approaches of K-modes algorithmic rule. Moreover, it explained K-modes being associate extension to the quality K-Means cluster algorithmic rule and shows the most modifications to K-Means additionally.

Another necessary material relating to K-Means technique was named “Implementing & Improvisation of K-Means cluster Algorithm” by Unnati R. Raval and Chaita Jani [3]. during this work, the authors have highlighted the cluster techniques because the most significant a part of the information analysis and sit down with K-Means in concert of the oldest and widespread cluster techniques. Also, the benefits and downsides of K-Means algorithmic rule and also the ways in which in conjunction with techniques to boost the present algorithmic rule for higher accuracy and performance were mentioned.

Shreya Tripathi, Aditya Bhardwaj and Poovammal E in their paper referred to as "Approaches to cluster in client Segmentation" [4], have examined some cluster algorithms as well as K-Means and ranked cluster from numerous sides, and indicate the ultimate results of scrutiny these techniques. Also, it showed an ideal correspondence in time, place and circumstance for exploitation every of the algorithms. throughout the analysis method, the authors have known the benefits and downsides of those cluster algorithms.

As per the paper “Revised DBSCAN algorithmic rule to cluster information with dense adjacent clusters” [5], the algorithmic rule has been wide utilized in several areas of science due to its plain structure and also the ability to notice clusters of varied sizes. With the assistance of DBSCAN algorithmic rule, the tight square measures are found and recursively dilated to search out dense indiscriminately formed clusters. However, once detective work border objects of adjacent clusters, the algorithmic rule is unbalanced. the last word cluster result obtained from DBSCAN depends on the order within which objects square measure processed throughout the algorithmic rule run.

“Developing a model for activity client loyalty and price with RFM technique and cluster algorithms” [6] is another work reviewed. Nowadays, the client relationship management ways play a substantial role in business areas. per this paper, knowing their customers behaviour, priorities, and wishes supported information permits banking sector to own a selected cluster of shoppers to counsel them personalised offers and increase their usage time of existing product. Also, the authors have delineated RFM technique and totally different cluster algorithms for client segmentation. Ultimately, the construct of client loyalty and retention through behavioural and demographic options was thought of one among the necessary aspects within the paper.

According to the paper “Review on client Segmentation Technique on Ecommerce” [7], There square measure many parts to try to client segmentation, which states client segmentation is associate activity to divide customers or item into teams that have constant characteristics. Information that required for client segmentation square measure internal information and external information. the inner information embraces demographic information and information purchase history, whereas the external information embrace cookies and server logs. Internal information is obtained once client does registration or transactions and external information is obtained from net server or different supply. Ways of client Segmentation is classified into straightforward technique, RFM (Recency, Frequency, financial Value) technique, Target technique, and unattended technique. on the right track technique, scientist specialize in one variable, it is product or purchase. unattended technique was used once cluster method scientist have several variables. Method of client Segmentation is simplified into shaping business objective, grouping information, information preparation, analysing variable, processing, and performance analysis.

“Customer’s life-time price exploitation the RFM model within the banking industry: a case study [8]” by Esmaeil Nikumanesh and Ameer Albadvi expressed that the semipermanent customer’s price is considered as the concert of the essential metrics for the monetary consequences of the customer-firm relationships. This price is associate acceptable benchmark for activity the performance of the firm and its monetary markets. It targeted on customers instead of product. forward that convenience of information rises at the client level, the customer’s life-time price will play an important role in promoting and company strategy within the future.

Another fascinating paper was “Performance sweetening of client Segmentation employing a Distributed Python Framework, Ray [9]” by Debajit Datta, Rishav Agarwal, Preetha Evangeline David, that expressed that client

DOI: 10.17148/IARJSET.2021.8401

segmentation is that the core of any recommender system used. Recommender Systems, on the opposite hand, became the crux of on-line searching and streaming platforms. Each on-line streaming platform bases its profits on the success of its recommender system, whereas a recommender system is merely nearly as good as its client segmentation algorithms. They numerous algorithms they used for Classification square measure were call trees, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), AdaBoost Classifier, Gradient Boost Classifier and provision Regression.

Finally, the paper by Balmeet Kaur, Pankaj Kumar Sharma on “Implementation of client Segmentation exploitation Integrated Approach [10]”. This paper projected a study on integrated novel approach supported cluster exploitation K-means and associative mining exploitation Apriori technique. per the authors, after identification of targeted customers and their associative shopping for pattern, the business managers take the strategic profitable choices consequently. the matter of characteristic potential client is gaining a lot of and a lot of attention is highlighted during this paper.

III.METHODOLOGY

How much does one pay to draw in new customers, as compared to the expenses on holding the existing? To sustain and expand business, one ought to understand having the ability to retain existing customers is as necessary as exploring new customers. If the speed of shopper’s feat is larger than rate of recent customers coming into, our customers information is really shrinking.

Not each deal is profitable, not all the shoppers square measure financially enticing to the business. It's crucial to confirm resources allotted or deployed square measure in line with profit or worth a client carries. Selling goal is to maximise influence of customised plans on targeted customers.

A. Proposed Model

The proposed model is divided into two modules:

- Exploring Customers Segmentation with RFM Analysis through K-Means Clustering with Python.
- Predicting the Customer Life Time Value (LTV) through Linear Regression.

B. Data Collection

The dataset “Online Retail” is taken from the UCI Machine Learning Repository.

The dataset is a transactional data that contains transactions from December 1st 2010 until December 9th 2011 for a UK-based online retail.

The different attributes of the dataset are product name, quantity, price, and other columns that represents ID

There are 8 columns:

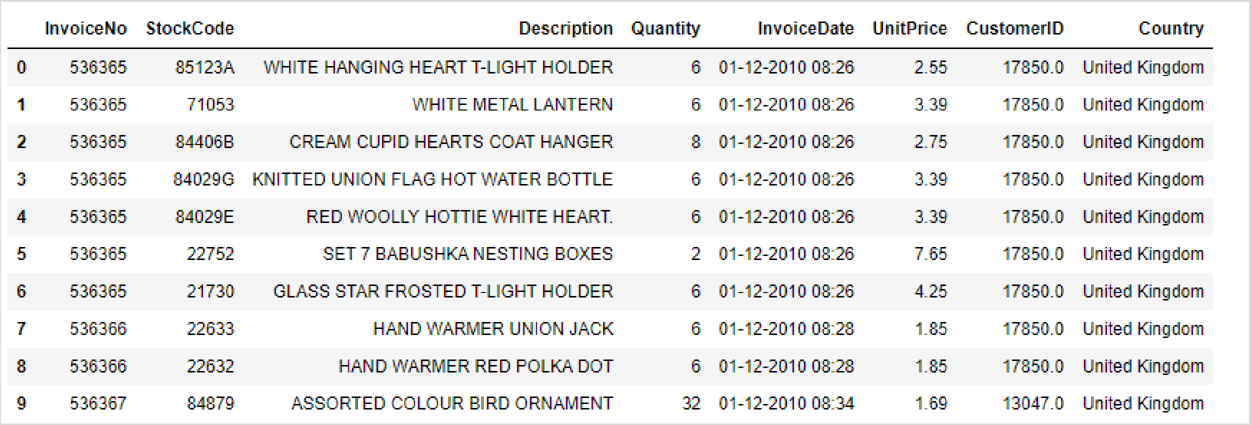
- Invoice No — an integer value
- Stock Code — a string of alphanumeric characters
- Description — details of the product
- Quantity — the number of items purchased
- Invoice Date — the date of invoice
- Unit Price — the price of each product per piece
- CustomerID — a unique integer value for every customer
- Country — the country to which the customer belongs to

C. Data Cleaning and Pre-Processing

The gathered data must be cleaned and pre-processed and after improving the data, it is read to run on the algorithms. The duplicate values are removed, dropping of irrelevant columns, data is arranged with numerical values by pre-processing and by this model building and selecting the features becomes easier. Pre-processing plays the vital role for the whole dataset.

DOI: 10.17148/IARJSET.2021.8401

Dataset in Fig. 1 shows the information which is needed for the analysing the data. Pre-processing is done on the column 'Description'. It uses regular expressions to remove punctuation and lowers capital letters. Also, missing values are removed from the 'CustomerID' column



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047.0	United Kingdom

Fig. 1 Cleaned and pre-processed dataset

IV. Machine Learning Approaches

The overall idea here is to perform the RFM (Recency Frequency Monetary) analysis on the dataset used for our segmentation and obtain the RFM score for each customer. Using the RFM score obtained then perform clustering using the k-means algorithm and finally predict the Customer Life Time Value (LTV) to find out the potential customers.

A. RFM Analysis

RFM analysis is used for the customer data at aggregate level and is used to segment customers into similar groups. It has been used in businesses since long time, especially as part of marketing technique.

Three main variables of analysis**,** **R**-recency, **F**-frequency, and **M**-monetary, are defined and computed. These three values are quite important and indicate customer's involvement and satisfaction. These are easy to obtain from the basic information for each purchasing history of the customer.

RFM Analysis Steps as below:

- Aggregating and computing RFM variables for each ID.
- Assigning RFM score.
- Segmenting customers according to scorings.
- Analysing characteristics/trait of targeted clusters members.

B. KMeans Algorithm

KMeans rule is repetitive rule that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters), where every information belongs to just one cluster. It tries to form the intra-cluster information points as similar as potential whereas conjointly keeping the clusters as completely different (far) as potential.

It assigns information points to a cluster such the total of the square distance between the information points and therefore the cluster's centre of mass (arithmetic mean of all the information points that belong to it cluster) is at the minimum. The less variation we get at intervals clusters, the additional homogenous (similar) the information points at intervals an equivalent cluster.

C. Linear Regression

Linear regression is one of the simplest and most famous Machine Learning algorithms. It is basically a method of statistics that is used for prediction and analysis of the data. Linear regression generates predictions for continuous, real or numeric variables. Example: Age, Salary

Linear regression aims to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

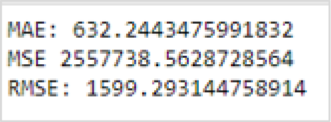
The algorithm shows a linear relationship between a dependent variable and one or more independent variables; therefore, it is called as linear regression. Since linear regression gives the linear relationship, which means it finds how the value of the dependent variable is changing in accordance to the value of the independent variable.

V. Experimental Results and Analysis

Python is a high-level object-oriented scripting language, designed to be readable it uses English keywords more and uses indentation, whereas other languages use punctuation. Functions gives best modularity for our application and a high amount of reusability of the code. In python classes and objects are easily used.

Python libraries for data analysis by making use of NumPy, Pandas and Scipy for the selected dataset. Open-source library pandas is used to manipulate, analyse, load, and visualize the selected datasets. The other open-source library scikit-learn builds smart models and make cool predictions and is used in machine learning algorithms.

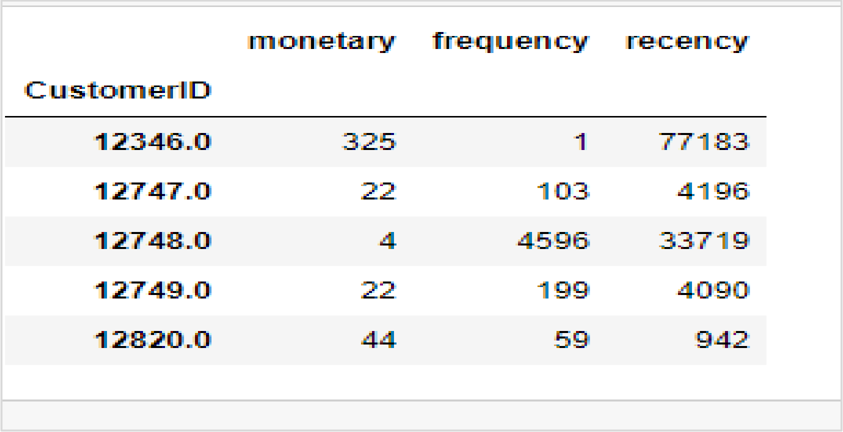
Output of the model is visualized graphically across the dataset for chosen test dataset. The visualization represents study of real value, also the prediction of results. RFM analysis, KMeans clustering and Linear regression tells the results gained by the analysis. Also, gives the customer life time value for each customer. It gives the highest R-Square value with maximum accuracy in the regression analysis. The analysis also gives the values for R-square, MSE and MAE values in Fig 2.



```
MAE: 632.2443475991832
MSE 2557738.5628728564
RMSE: 1599.293144758914
```

Fig. 2 Model Evaluation metrics

The RFM analysis is done for every customer based on how recently, how frequently and how much he spends on the product, as shown in Fig. 3

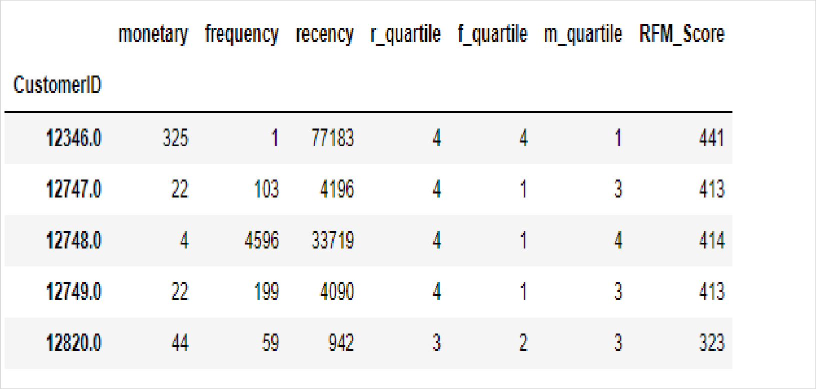


	monetary	frequency	recency
CustomerID			
12346.0	325	1	77183
12747.0	22	103	4196
12748.0	4	4596	33719
12749.0	22	199	4090
12820.0	44	59	942

Fig. 3 The result of the RFM analysis

The RFM score for each customer is generated after getting the RFM per quartile, which will be used to form clusters using KMeans algorithm, as shown in Fig. 4

DOI: 10.17148/IARJSET.2021.8401



	monetary	frequency	recency	r_quartile	f_quartile	m_quartile	RFM_Score
CustomerID							
12346.0	325	1	77183	4	4	1	441
12747.0	22	103	4196	4	1	3	413
12748.0	4	4596	33719	4	1	4	414
12749.0	22	199	4090	4	1	3	413
12820.0	44	59	942	3	2	3	323

Fig. 4 The RFM score

The results of the RFM analysis in the form of heat map, as shown in Fig. 5

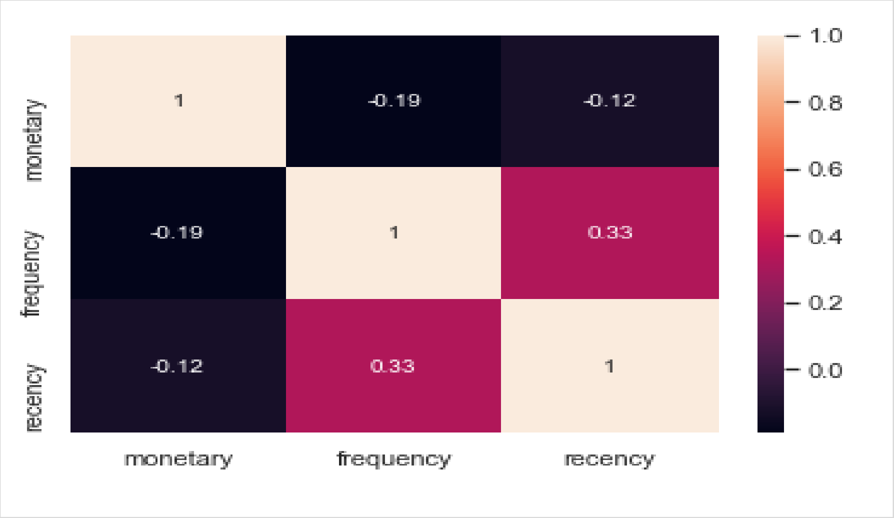


Fig. 5 Heat Map showing the RFM analysis

We, now get the clusters of the customers based on their RFM score, as shown in Fig. 6 and Fig. 7. We can see that the majority of the customers belong to the red cluster, followed by green and the least being the blue cluster. Using this information, the businesses can identify their valuable and consistent customers and plan their market strategies accordingly.

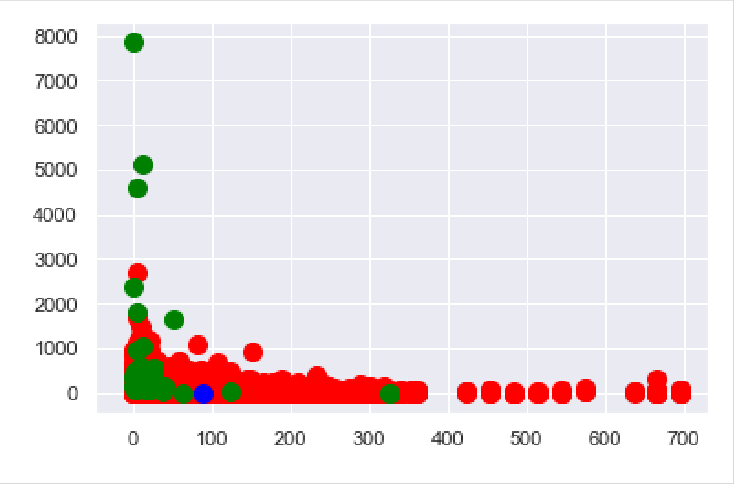
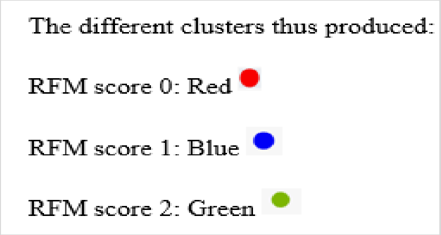


Fig. 6 Clusters of the customers based on the RFM score



The different clusters thus produced:

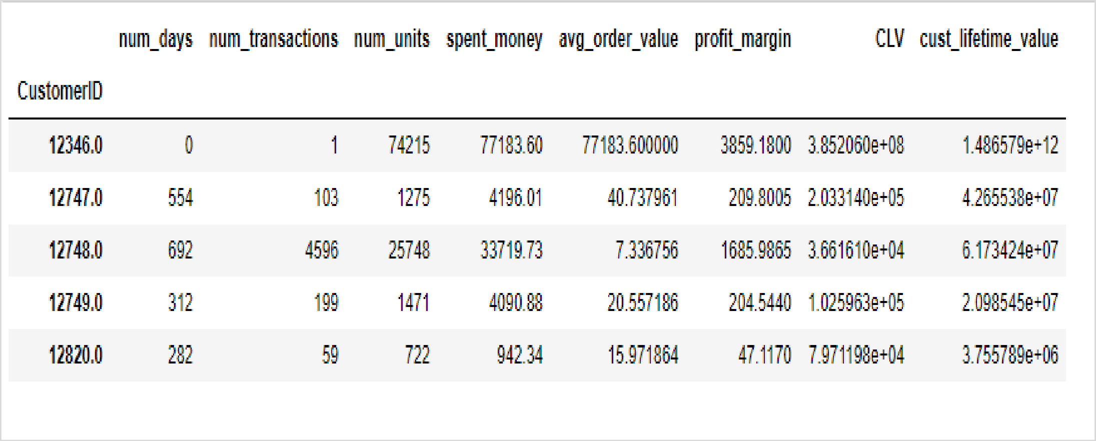
RFM score 0: Red 🔴

RFM score 1: Blue 🔵

RFM score 2: Green 🟢

Fig. 7. The cluster information

The Customer Life Time Value for every customer is predicted using the Linear Regression algorithm to identify the potential customers to the retail chain, as shown in Fig. 8



	num_days	num_transactions	num_units	spent_money	avg_order_value	profit_margin	CLV	cust_lifetime_value
CustomerID								
12346.0	0	1	74215	77183.60	77183.600000	3859.1800	3.852060e+08	1.486579e+12
12747.0	554	103	1275	4196.01	40.737961	209.8005	2.033140e+05	4.265538e+07
12748.0	692	4596	25748	33719.73	7.336756	1685.9865	3.661610e+04	6.173424e+07
12749.0	312	199	1471	4090.88	20.557186	204.5440	1.025963e+05	2.098545e+07
12820.0	282	59	722	942.34	15.971864	47.1170	7.971198e+04	3.755789e+06

Fig. 8 The Customer Life Time Value

VI. CONCLUSION AND FUTURE SCOPE

Customer segmentation is a way to improvise communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation is needed to get potential customers used to increase profits of the business. Potential customer data can be used to provide service and know the characteristics of customer. Customer segmentation can have a great and positive impact on business if done properly.

In competitive world of e-commerce marketing, the problem of identifying potential customer is gaining more and more attention. To address this problem, this paper proposes a study on integrated approach based on clustering using K-means and RFM (Recency, Frequency, Monetary) analysis. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly. We were able to determine which clusters have more customers and which are potential clusters or likely to be customers.

The results thus obtained can be used to target customers and try to find ways to increase your order count via email reminders or SMS notifications directed to other identification features.

Maybe you can give them a discount when they come back within 30 days. Ideally, you can provide a delayed coupon (which will be used at some point) at checkout. Similarly, you may want to try other sales and marketing strategies for the non-active customers.

REFERENCES

[1] Ching-Hsue Cheng, and You-Shyang Chen, “Classifying the Segmentation of Customer Value via RFM model and RS theory,” Expert Systems with Applications 36, 2009, pp. 4176–4184.

[2] N. Sharma, and N. Gaud, "K-modes Clustering Algorithm for Categorical Data," International Journal of Computer Applications, vol. 127, no. 17, 2015, pp. 1-6.

[3] Unnati R. Raval, and Chaita Jani, “Implementing & Improvisation of K-means Clustering Algorithm,” International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pp. 191-203.

[4] Shreya Tripathi, Aditya Bhardwaj, and Poovammal E “Approaches to Clustering in Customer Segmentation," International Journal of Engineering and Technology, 2018, pp. 802-807.

[5] Thanh N. Tran, Klaudia Drab, and Michal Daszykowski, “Revised DBSCAN algorithm to cluster data with dense adjacent clusters,” Chemometrics and Intelligent Laboratory Systems 120, 2013, pp. 92–96.

DOI: 10.17148/IARJSET.2021.8401

[6] Razieh qiasi, Malihe baqeri-Dehnavi, Behrouz Minaei-Bidgoli, and Golriz Amooee, "Developing a model for measuring customer loyalty and value with RFM technique and clustering algorithms," The Journal of Mathematics and Computer Science vol. 4 no.2, 2012, pp. 172 - 181.

[7] Juni Nurma Sari 1,2, Lukito Edi Nugroho1, Ridi Ferdiana 1, P. Insap Santosa 1," Review on Customer Segmentation Technique on Ecommerce", Advanced Science Letters All rights reserved Vol. 4, 400–407, 2011.

[8] Esmaeil Nikumanesh and Amir Albadvi," Customer's life-time value using the RFM model in the banking industry: a case study", Int. J. Electronic Customer Relationship Management, Vol. 8, Nos. 1/2/3, 2014.

[9] Debajit Datta, Rishav Agarwal, Preetha Evangeline David," Performance Enhancement of Customer Segmentation Using a Distributed Python Framework, Ray", International Journal of Scientific & Technology Research Volume 9, Issue 11, November 2020.

[10] Balmeet Kaur, Pankaj Kumar Sharma," Implementation of Customer Segmentation using Integrated Approach ", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6S, April 2019.