Journal of Engineering Sciences
Vol 13, Issue 05, MAY/2022
ISSN NO: 0377-9254

Customer Segmentation Using RFM Model and Different Clustering Algorithms

Vamsi Krishna T[1], Venkata Joshi Y[2], Venkata Yeswanth Kumar B[3], Sameer Sd[4]
[1]Asst Professor, Department of Computer Science and Engineering
[2,3,4,] Student, Department of Computer Science and Engineering
[1,2,3,4,] Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, AP,
[*] yeswanthbolisetty@gmail.com

ABSTRACT

Customer segmentation is the process of classifying the customers based on characteristics such as behavioural, demographic, geographic, and psychographic models. The segmenting is mainly used in supermarkets, shopping malls, and other private limited enterprises. Customers are classified into groups which totally depend on their purchasing. For this, an RFM model is used for the segmentation and also for the value analysis which was gained by using the sales data online. In addition to that, the clustering algorithms like K-means clustering, Mini Batch K-means, Agglomerative clustering, are performed. The virtue and the potency of our method which has been proposed in this paper are reinforced by boosting the results of some indices, such as widening the number of customers who are active and increasing the consumption amount. To gain a high and good level of customer satisfaction, customer relationship management (CRM) techniques are brought forward.

Key Words:

Customer Relationship Management (CRM), RFM,, Davies Bouldin Index, Principal Component Analysis (PCA), K-means clustering, Mini Batch K-means clustering, Agglomerative clustering.

Introduction

In the present technology, online shopping is playing a major role in the trading pattern all over the world. The statistics has shown that the online retail sales reached 127.5 trillion upto 2020. In an online environment, the behaviour of the purchase of the customers changes vigorously. For predicting the online behaviours based on the data mining an excellent customer-oriented marketing strategy is therefore much needed by selling the enterprises.

Data mining, the process of finding correlations and patterns, can discover hidden knowledge of great relevance from vast amounts of online transaction data, is the most inclined method for the customers purchase behaviour analysis. In the present span of big data, data mining is counted to have wide application prospects over the industry. In the past two decades, there have been more magnificent theories about data mining with ample industrial applications.

RFM stands Recency, Frequency, Monetary, and refers to the purchase done most recently, frequency of purchase, and monetary value of purchase, respectively. Recency is the difference between the last purchase of the customer and the last date of the statistical period.

The greater value of R, the shorter the interval. Frequency is the number of purchases made by the customer around the statistical period. If the value of the F is greater, the customer is more likely to buy the products again and again. Monetary denotes the complete amount spent on consumer purchases over the statistical period. The larger the summation of the purchase value, the more loyal the consumer is in general.It can be used as a direct indicator of a company's production capability.

To break down the customer data within the customer relationship management framework, data mining tools have been used. Data mining can be used to uncover important information about a customer's behaviour and qualities. As a result, it is critical for businesses looking to recruit and keep consumers, as it assists them in maximising value for customers and supporting their customer management and market strategy decisions. In the age of big data, the use of data mining in the CRM area is undeniably a growing trend. Clustering or segmentation, which splits customers into broad groups based on similarity, is one of the most extensively used data mining models.

In this paper, our analysis is based on real-world data from a company in the UK. By combining RFM model and K-means approaches, we achieve customer segmentation and offer management options. We generate a standardised dataset for further analysis using online transaction data collected from December 2010 to December 2011. As to this foundation, we serve the customer segmentation and value analysis using an RFM model and K-means algorithm. After that, a PCA technique is used to determine the weight of RFM indicators. Customers are split into four groups based on their purchase behaviours[9].

Relevant Works

RFM Model: Hughes of the American Database Institute proposed the RFM model in 1994 [15]. As a common tool of customer value analysis, it has been widely used for calculating customer lifetime value, customer segmentation, and behaviour analysis[10].

According to research, the higher the R or F value, the more likely the matching client will do a new transaction with the supplier. Furthermore, the more the value of M, the more likely the products or services are purchased by the client from the supplier endlessly. While Hughes regarded all three variables as equally important[15], Stone argued that the importance of the three variables varies per industry due to differences in their features[10], implying differing weights for these variables.

RFM is frequently one used in customer value analysis, and the scholars have expanded it to cover a wide range of issues. To evaluate the weight of the variables of RFM, Liu and Shih engaged an analytic hierarchy process (AHP), a clustering method for categorising customers, as well as an association rule method for recommending products to different groups of clients. To develop customer classification rules, Cheng and Chen used RFM research with a crude set theory. Chiang gave an idea on RFMDR model which is an improved form of RFM analysis, for identifying major online purchasing customers and generating fuzzy association rules for the industry. For the postal services Kolarovszki et al.

evolved a multidimensional segmentation-based modelling system. In the postal industry, this CRM design is useful.

For evaluating potential users, Song et al. recommended using a time series-based statistical approach. This method can be used to segment RFM time intervals in a large-scale dataset. Heldt et al. introduced an RFM per product (RFM/P) model in light of the fact that most RFM models are built from a consumer perspective rather than a product perspective. Customers' values for all items are calculated separately first, then summed together to generate an overall customer value in this approach. This can be used to do empirical investigation of financial institutions and supermarkets [11]. Adnan Amin et al. used rough set, classification, and data transformation approaches to study the prediction of customer attrition in the telecom business under various scenarios [1-4].

Clustering Algorithms: Clustering is the process of combining similar objects from a collection of physical or abstract objects into groups. Macqueen originally employed the K-means technique in 1967, and it has since been widely used in a variety of industries, which includes mining of data, statistical data analysis, and some other business applications.

Customer segmentation is one of the most common uses of K-means, according to the literature. The K-means algorithm is commonly used to identify and design relevant marketing strategies for valuable customers. To execute customer relationship management, Cheng & Chen used an RFM model and K-means, and experimental findings demonstrate that the model they constructed is an effective way to analyse customer value [10]. In [14], Khalili-Damghani et al. proposed a hybrid soft computing strategy based on clustering, rule extraction, and decision tree methodology to anticipate segmentation of new customers of customer-centric enterprises.The K-means approach is not only faster in calculation than other clustering algorithms, but it may significantly lower the rate of data misclassification. As a result, we cluster based on R-F-M qualities using the K-means technique. The number of clusters and the startup settings determine the accuracy of this method. The well-known elbow method is often used to calculate K's value.

Because of its fast efficiency, K-means is one of the most common clustering methods. Because K-means requires the entire dataset to be stored in main memory, its computation time increases as the size of the datasets being analysed grows. As a result, numerous strategies for reducing the algorithm's time and spatial cost have been presented. The Mini batch K-means algorithm is a distinct technique.

The empirical findings reveal that it can save a significant amount of time while sacrificing some cluster quality, but no comprehensive study of the algorithm has been conducted to determine how the characteristics of the datasets, such as the number of clusters or their size, directly affect the separation quality. In [19], the intrusion detection system over Big Data employs mini batch k-means clustering. Using mini batch k-means with PCA has given a good result in their experiment.

There is some research about agglomerative hierarchical clustering. In [17], Li et al. proposes a Q-criterion based hierarchical clustering algorithm, named HACNJ. With a

computational complexity of O(n3) and a space complexity of O(n2), HACNJ seems to have the same complex nature as basic hierarchical clustering, which is incredibly expensive when dealing with large datasets. They proved that HACNJ is beneficial through studies on the Iris dataset. In [18], Yogita Rani and Dr. Harish Rohil discussed various enhanced principal component analysis algorithms in depth. These techniques are designed to address the drawbacks of traditional hierarchical clustering protocols.

Methodology

The proposed customer value analysis process is discussed in this section. The procedure is divided into four sections, as indicated in Figure 1: (1) Data preparation and preprocessing; (2) Normalisation of RFM model indices; (3) Index weight analysis; and (4) Customer clustering by the K-means algorithm. To identify target customers, the RFM model and the K-means algorithm are applied to every dimension of customer data..

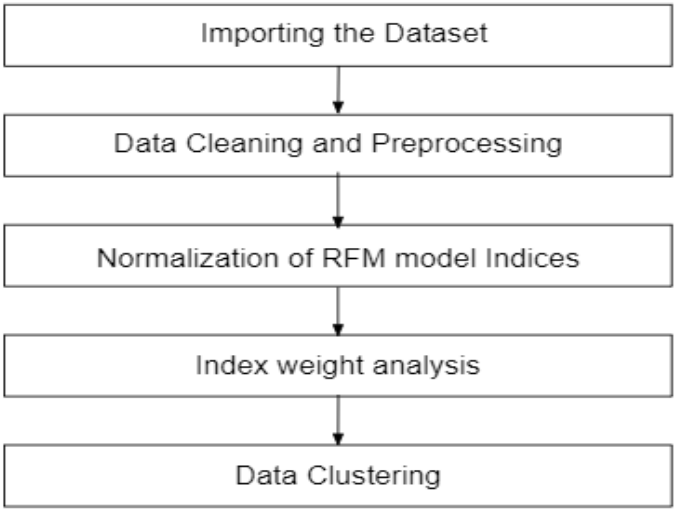


Fig.1: Step-by-step process

The research analysis process is described in detail in the following steps :

Data Preprocessing

First, an original dataset based on RFM model parameters is chosen for the empirical case study. The first dataset is created by cleaning the original dataset to remove outliers and erroneous numbers. The data is then translated into a format that is easier and more efficient to process for customer value analysis by removing unnecessary attributes.

Normalisation of RFM model indices

To eliminate the impact of numerical values on the classification results, the min-max normalisation method is used to standardise the data and obtain the initial standardised dataset (see formula (1)): Given the large differences in the value ranges of the three indicators of the RFM model, i.e., time since last purchase, purchase frequency, and total

Vol 13, Issue 05, MAY/2022
ISSN NO:0377-9254

purchase amount, the min-max normalisation method is used to standardise the data and obtain the initial standardised dataset. The formula is given by

$$x`_{ij} = \frac{x_{ij} - \min\{x_j\}}{\max\{x_j\} - \min\{x_j\}} \quad (1)$$

where x_{ij} represents j^{th} index of i^{th} sample.

Indicator weight analysis

The value and relative relevance of each inspection index of a measured object is referred to as index weight. A principal component analysis method is used to assign weights to the RFM model because the research target of this particular paper is distinguished by more clients and vast consumption data. Principal component analysis is a statistical analytic method that reduces multiple indicators into a few comprehensive ones using a dimensionality reduction methodology. The weight of each indicator is equal to the variance contribution rate of the principal component. As the variance contribution rate rises, the importance of the major component grows.

The following is a description of the computation procedure:

Step 3(a): The following is the equation for creating a principal component analysis model:

$$\begin{cases} F_1 = X_1U_{11} + X_2U_{12} + \ldots + X_mU_{1m}, \\ F_2 = X_1U_{21} + X_2U_{22} + \ldots + X_mU_{2m}, \\ F_3 = X_1U_{31} + X_2U_{32} + \ldots + X_mU_{3m}, \end{cases} \quad (2)$$

$$U_i = (A_{ii}), \quad (3)$$

where U_{ij} is the linear combination of the original variables i(1,2,…,p), and j(1,2,…,m) contains the proportional coefficients of a principal component. In most circumstances, m and p denote the composite principal component score, whereas W is the weight showing the component's variance contribution rate. The below formula is used to carry off the weight normalisation:

$$F = F_1(W_1(W_1+W_2+W_3)) + F_2(W_2(W_1+W_2+W_3)) + F_3(W_3(W_1+W_2+W_3)), \quad (4)$$

where F is the dataset of subsequent clustering.

Step 3(b): The primary component load matrix U, the factor load matrix A, and the eigen value are calculated using the following formula:

Following the selection of K initial cluster centres Ci (1 I k) at random from the dataset, the Euclidean distance between the remaining data objects and the cluster centre Ci is calculated. The target data object is assigned to the cluster with the nearest cluster centre Ci. To start the next iteration, the new cluster centre is calculated as the average of all data objects in each cluster. This procedure is repeated until the cluster centre no longer changes or until the maximum number of iterations is reached.

The below formula(5) is used to determine the Euclidean distance between the data objects in space and the cluster centre :

$$D(x, Ci) = \sqrt{\sum^{m} (x_i - C_{ij})}$$

Where, x is the data object, C_i is the i^{th} cluster centre, and m is the dimension of the data object, and the other two variables present in the square root are the x and C_i 's j^{th} attribute values respectively.

Clustering the customers by K-means algorithm

The number of clusters chosen, k, has a considerable impact on the clustering results. The elbow approach is commonly used in practice to determine the appropriate k value. The SSE-k interaction curve is structured like an elbow, and the value of k that corresponds towards this elbow is the true cluster frequency of the data. The SSE (sum of squared errors) is the main indicator of the elbow method, as shown in the following formula:

$$SSE = \sum_{f=1}^{k} \sum_{x \in Ci} |d(x_i Ci)|^2$$

Where Ci is the ith cluster and SSE is the total clustering error, which is a measure of clustering quality.

Clustering the customers by Mini Batch K-means algorithm

The primary idea behind the Mini Batch K-means algorithm is to employ relatively small batches of data with a fixed dimension so that they may be kept in ram. Each iteration obtains a fresh random selection from the dataset and uses it to update the clusters, and the process is continued until convergence. Each mini lot modifies those clusters using only a curved composition of the prototypical values and the data, with a diminishing learning algorithm as the number of iterations advances. The inverse of the quantity of observations allocated to a cluster during the procedure is this learning rate. Because the effect of incoming data diminishes as the number of observations increases, convergence can be observed since no adjustments in the clusters exist for multiple iterations in a row.

Clustering the customers by Agglomerative clustering algorithm

To execute agglomerative hierarchical clustering, we'll follow the methods below.

1. Preparing the data
2. Based on statistical information provided in step 1, use the linkage function to organise the items into hierarchical cluster trees. The linkage function is used to connect objects/clusters that are close in proximity.
3. Choosing where the hierarchical tree should be divided into clusters. This divides the data into sections.

About Data Set

The dataset covers all transactions for a UK-based and registered, non-store online retail between 01/12/2010 and 09/12/2011, involving 4373 customers, and contains 5.4 lakh instances. The company primarily sells one-of-a-kind all-occasion gifts. Many customers of the company are wholesalers. This company sells 4,071 types of products which mainly consist of gift articles.

InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country are the eight attributes in the dataset. InvoiceNo is a transaction number of the products that are purchased by the customer. StockCode is a unique code of a particular product, different products have different Stock codes. Description field contains the name of a particular product. Quantity means the number of particular products that are taken by the customer. InvoiceDate is the date and time of a particular transaction. UnitPrice is the price of the particular product. CustomerID is a 5- digit unique number that is given to each and every customer. Country field contains the name of the country in which the customer purchased. In this dataset, there are some missing values in the CustomerID field and some outlier records in Quantity.

Prerequisite steps:

In the dataset, the CustomerID field is filtered in such a way that there should be no missing values. The Quantity field must have a value that is greater than zero but there are some negative values in the Quantity field so, the Quantity field must be filtered which contains only positive values. After filtering the initial dataset, all the outliers are removed to form a dataset. To make it a Systematic and standardised dataset, the range technique must be utilised. Then, principal component analysis is implemented to weight RFM indicators to get the final Standardised dataset.

User Classification Results:

Mini Batch K-means algorithm: This algorithm is very similar to the K-means algorithm but overcomes some problems of k-means algorithm. Here, we considered 3 different number of clusters i.e., 3, 4, 5 clusters. This algorithm gave the near result as the K-means algorithm. The efficiency can be measured by using Davies Bouldin Score and Silhouette Score.

Agglomerative clustering algorithm: Agglomerative clustering is just a "bottom-up" methodology to clustering. To put it another way, each item is first thought of as a single group (leaf). The connected components are often the most comparable and are joined into a new larger cluster at each phase of the method (nodes). This method is repeated until all points belong to a single large cluster (root). Agglomerative clustering is a technique for segmenting clients based on a set of factors. The clusters that are taken into account are comparable to the clusters that have been taken into account below. Davies Bouldin Score and Silhouette Score can be used to assess the algorithm's efficiency. The efficiency is only for this particular dataset.

K-means algorithm: The K-means clustering algorithm is regarded top among all the clustering algorithms since its purpose is to organise data points into separate non-overlapping groupings. The K-means clustering algorithm is used in Customer Segmentation to gain a better understanding of them, which may then be used to boost the company's income.

Using the K-Means clustering algorithm, the data is organised into groups. Depending on the K value, the clusters are constructed, which ranges from 3 to 5, and the Davies Bouldin Index is used to find the ideal value of K. The lower the Davies Bouldin Score for K, the more optimum the clusters.

To implement K-means clustering, there is a module in Python language known as sklearn. In addition to that some other modules like feature engine, matplotlib are also used here.

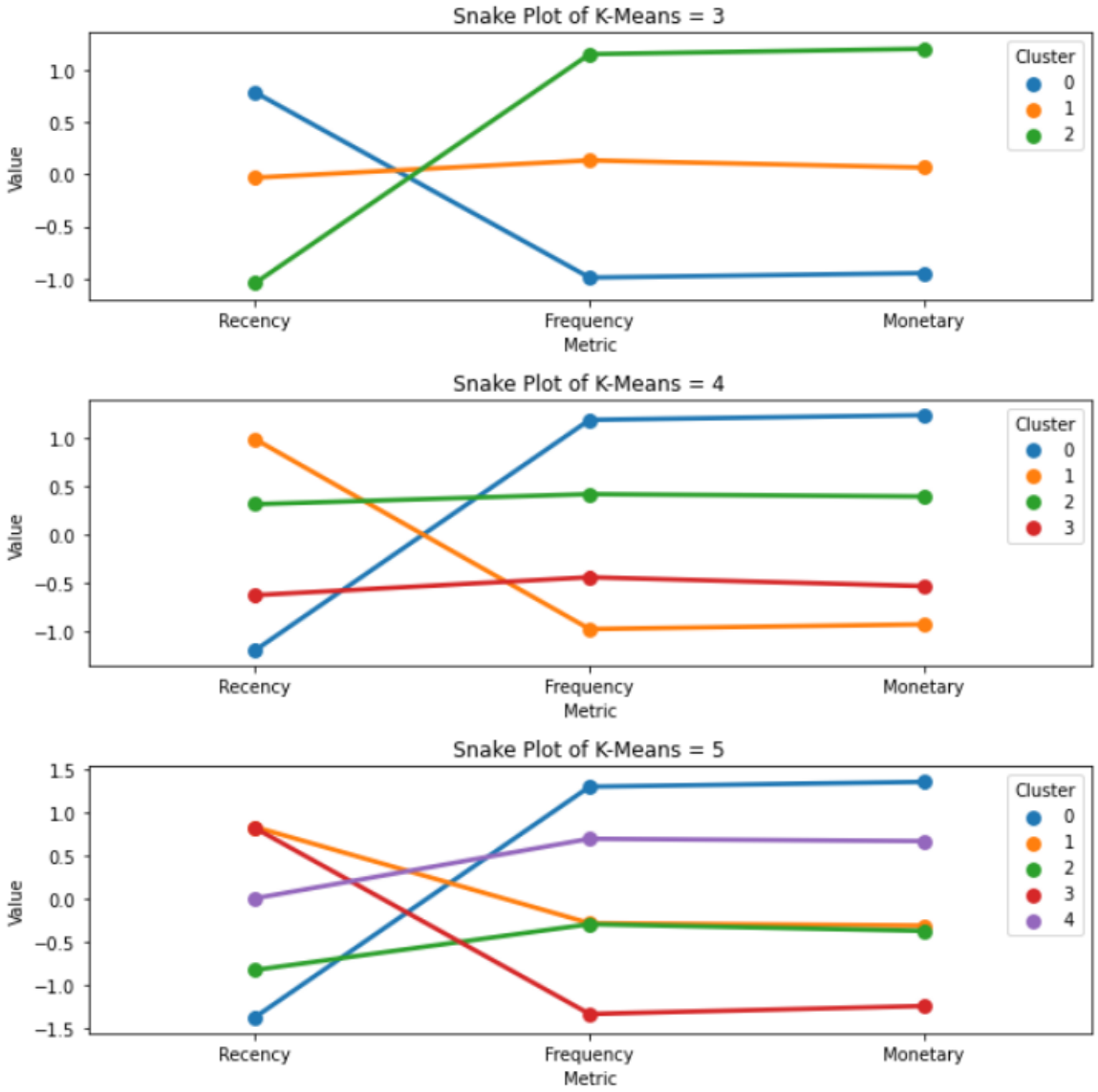
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGIN(	6	12/1/10 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LA	6	12/1/10 8:26	3.39	17850	United Kingdom
536366	22633	HAND WARMER	6	12/1/10 8:28	1.85	17850	United Kingdom
536367	22748	POPPY'S PLAYHC	6	12/1/10 8:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRIN	8	12/1/10 8:34	3.75	13047	United Kingdom

Table.1: Some data entries from the initial dataset

From the customer purchase data, there are different types of customers like high amount spenders, and recently visited customers, most often visiting customers and combination of all these which can appear as outliers.

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4596	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34

Table.2: RFM values for each customer



Graph.1: Snake plot of clusters when K is 3,4,5

Some of the RFM values for customers are shown in the table above. Recency, Frequency, and Monetary Values differ for each consumer. As said earlier, there are different customers for a Retail market. Customers can be grouped using Recency, Frequency, and Monetary

values, as well as various mathematical requirements. Here, we will consider three cases, and they are as follows:

1. When K=3

   In this scenario, we divide all of the clients into three groups. Customers in group1 have only a short time since their last purchase but have spent a smaller amount of money. Customers in group 2 had higher Recency, Frequency, and Monetary indicators than the overall average. Customers in group3 have had a long time since their last purchase but they spend a huge amount of money.

2. When K=4

   In this case, there will be four clusters in which all the customers are categorised into one of those. In Group1, Customers had a long time in visiting, but frequency and monetary metrics have higher value. In Group2, Customers were most recently visited but spent a smaller amount. In Group3, Customers have overall values greater than the average values. In Group4,Customers have overall values less than the average values.

3. When K=5

   In this example, all of the consumers are divided into one of the five categories below. In Group1, Customers had a long time in visiting, but frequency and monetary metrics have higher value.In Group2, Customers were most recently visited but spent an average amount. In Group3, Customers have overall values less than the average values. In Group4, Customers were most recently visited but spent a smaller amount. In Group5, Customers have total values that are higher than the average.

The above algorithms are used in this paper and have given the results based on the data in the dataset. The efficiency of the algorithms are measured by using Davies Bouldin Score and Silhouette Score. The scores are show in the below tables

Davies Bouldin Score

	No. of clusters = 3	No. of clusters = 4	No. of clusters = 5
K-means clustering algorithm	1.119152952	1.065050395	1.07449785
Mini Batch K-means algorithm	1.11265606	1.071551092	1.075152686
Agglomerative clustering algorithm	1.064495215	1.261423259	1.26252073

Table 3: Davies Bouldin Scores of the each algorithm

Silhouette Score

	No. of clusters = 3	No. of clusters = 4	No. of clusters = 5
K-means clustering algorithm	0.306856147	0.312955130	0.28423120
Mini Batch K-means algorithm	0.30317712	0.302479362	0.284414287
Agglomerative clustering algorithm	0.284130635	0.230499287	0.226291935

Table 4: Silhouette Score of each algorithm

Conclusions

In this paper, Customer purchase patterns are analysed systematically by using RFM model and K-means clustering algorithm. We used three values of K and generated clusters appropriately to find the optimal number of clusters. Using the Davies Bouldin Index, we discovered that four clusters is the ideal amount. Customers are classified into four groups based on their purchase behaviour. In addition to that, we have analysed the customer purchase behaviour within a given time period like giving start date and end date to get the purchase behaviour in that time period. This result will be helpful for the persons who apply marketing strategies accordingly.

In the future, the new algorithms may be used to find out the different outcomes and new strategies can be used to attract the customers. Now-a-days, only customer purchase behaviour cannot be the same all the time, it is very difficult to identify each customer behaviour.

References

1. A. Amin, F. Al-Obeidat, B. Shah et al., "In the telecommunications business, just-in-time customer churn prediction," The Journal of Supercomputing is a publication dedicated to the study of supercomputers, vol. 76, no. 6, pg.no. 3920-3950, 2020.
   View at: Publisher Site | Google Scholar
2. A. Amin, B. Shah, A. M. Khattak et al., "A comparison of data transformation strategies for cross-company customer attrition prediction in telecommunication," Information Management's International Journal, vol. 46, pg.no. 305-320, 2019.
   View at: Publisher Site | Google Scholar
3. A. Amin, S. Anwar, A. Adnan et al., "Using a rough set method, Neurocomputing, vol.237, pg.no. 240-255, 2017.
   View at: Publisher Site | Google Scholar
4. A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar., "Using data certainty to anticipate customer attrition in the telecommunications business,". Journal of Business Research is a publication dedicated to the study of business, vol. 94, pg.no. 290-300, 2019.

View at: Publisher Site | Google Scholar

5. A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos., "Customer visit segmentation utilising market basket data," according to "Retail business analytics." Applications of Expert Systems, vol. 100, pg.no. 1-15, 2018.
View at: Publisher Site | Google Scholar
6. A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos., "Customer visit segmentation utilising market basket data," according to "Retail business analytics." Applications of Expert Systems, vol. 100, pg.no. 1-15, 2018.
View at: Publisher Site | Google Scholar
7. Y.-L. Chen, C.-L. Hsu, and S.-C. Chou., "Built multi-labeled decision tree,". Applications of Expert Systems, vol. 25, no. 2, pg.no. 200-210, 2003.
View at: Publisher Site | Google Scholar
8. M. S. Chen, J. Han, and P. S. Yu, "Data mining: a database perspective". Study of knowledge and data engineering, vol. 8, no. 6, pg.no. 865-885, 1996
View at: Google Scholar
9. E. W. T. Ngai, L. Xiu, and D. C. K. Chau., "Applications of the data mining techniques in CRM: a literature review and classification". Applications of Expert Systems, vol. 36, pg.no. 2530-2600, 2009.
View at: Publisher Site | Google Scholar
10. C.-H. Cheng and Y.-S. Chen., "Using the RFM model and RS theory to classify customer value segmentation,". Applications of Expert Systems, vol. 36, no. 3, pg.no. 4175-4185, 2009.
View at: Publisher Site | Google Scholar
11. R. Heldt, C. S. Silveira, and F. B. Luce., "From RFM to RFM/P: Predicting value per product of customers," The Journal of Business Research, vol. 3, 2019.
View at: Google Scholar
12. M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge., "Marketing knowledge management and data mining," Systems that aid in decision-making, vol. 31, no. 1, pg.no. 130-140, 2001.
View at: Publisher Site | Google Scholar
13. D. Chen, S. L. Sain, and K. Guo., "A case study of RFM model-based consumer segmentation utilising data mining for the online retail industry". Database Marketing & Customer Strategy Management, vol. 19, no. 3, pg.no. 195-210, 2012.
View at: Publisher Site | Google Scholar
14. K. Khalili-Damghani, F. Abdi, and S. Abol makarem., "For customer segmentation problems, hybrid soft computing technique, rule mining, and decision tree analysis". Soft Computing in Practice, vol. 73, pg.no. 815-830, 2018.
View at: Publisher Site | Google Scholar
15. A. M. Hughes., "The Strategic Database Marketing, Probus Publishers, Chicago, United States of America, 1994.
16. M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Kdd , Vol. 96, No. 34, pp. 226-231, 1996.
17. Jianfu, L., Jianshuang L., Huaiqing H. (2011). A Simple and Accurate Approach to Hierarchical Clustering. Journal of Computational Information Systems , 7(7), 2577–2584.
18. Rani, Y., Rohil, H. (2013). A Study of Hierarchical Clustering Algorithms. International Journal of Information and Computation Technology, 3(11), 1225–1232.
19. Kai Peng, Victor C.M. Leung, Qingjia Huang (2018). Clustering Approach Based on Mini Batch K-means for Intrusion Detection System over Big Data.
20. " Rotating Solar Trees ", Lecture Notes in Electrical Engineering 601, Springer Nature Singapore Pte Ltd. 2020, Page No: 482-487.