International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

Data Severance Using Machine Learning for Marketing Strategies

Atul Suryawanshi[1], Shantanu Sapkal[2], Aniket Patil[3], Satyajeet Vhanbatte[4], Shatakshi Kokate[5]

[1234]*Dept of Computer Science & Engineering, DYPCET, Kolhapur.*
[5]*Professor, Dept. of Computer Science & Engineering, DYPCET, Kolhapur.*

Abstract- *In Worldwide markets, there was a need to segment and analyze the customers and product data that was being generated over the period which was in ample amount. In CMO 2019 Survey, the 41% of marketing leaders said that their teams use artificial intelligence to improve their customer acquisition and segmentation efforts.*

The idea behind this model is to improve the marketing strategies among the firms by using Data Severance Technology. The implemented parts are basically like customer segmentation and product segmentation using multiple algorithms such as RFM (Recency, Frequency, Monetary), k-means, ABC and Pareto Analysis. The model demonstrates the benefits using these algorithms for analyzing customer purchasing and product selling data in trade sector with formation of meaningful clusters. In this paper, The system proposed two clustering models to segment 642,234 customers data and 20,870 products data by considering their RFM values and Class A, Class B, Class C values. In earlier processes, functioning was not efficient to segment and analyze the data. Hence, The System developed this model to overcome traditional segmentation techniques and scale-up the processing and analyzing raw data with the help of well-structured algorithms and marketing strategies.

***Key Words*: Segmentation, Marketing Strategy, Elbow Method, k-means, RFM (Recency, Frequency, Monetary), Pareto, ABC analysis, Clustering, Machine Learning, etc.**

1. INTRODUCTION

In today's environment, where the world is viewed as a global village, marketing has become a critical component of any company's success. Because competition is cutthroat, it is becoming increasingly difficult for any competitor to stay in the market for an extended period. Marketing's underlying belief is that if you don't change, you'll die. As a result, it is necessary to establish an appropriate marketing plan over time. The right marketing strategy aids businesses in achieving their marketing objectives. Marketing objectives aid in the achievement of company objectives, which attempt to gain a competitive advantage over competitors.[1]

Data Severance refers to using a variety of data analysis techniques and algorithms to discover previously unknown, valid patterns and relationship in large dataset [2]. Data Severance techniques like clustering and segmentation can be used to find meaningful patterns for future predictions. This model introduced Data Severance technique using Machine Learning into two segments. First is Customer Segmentation and Second is Product Segmentation. In Customer Segmentation, the system uses RFM (Recency, Frequency, Monetary) and K-means clustering algorithm to mine data of customer into the meaningful clusters. Secondly, In Product Segmentation ABC Analysis implemented with Pareto chart to analyse the product data for the future marketing strategies.

Companies must gain a better understanding of their clients' data in all aspects. Customer-company engagement has become increasingly dependent on detecting similarities and variances among customers, forecasting their behaviours, and offering better options and possibilities to customers. In this case, segmenting customers based on their data became critical.

For many years, RFM (recency, frequency, and monetary) values have been used to determine which consumers are important to the organisation, which customers require promotional efforts, and so on. Organisations and individuals have widely employed data-mining tools and strategies to analyse their stored data. Clustering is a data mining task that has been used to group people, objects, and other things [3]. Customers, as marketers are aware, have a wide range of demands and desires. Companies have utilised a variety of segmentation criteria and approaches to better identify and understand client groups and supply them with more appropriate products and services to meet their various needs and desires.

A Clustering algorithm is a method for grouping a physical or abstract entity into groups of related things. A cluster is a collection of data objects that are similar to one another but distinct from those in other clusters [4]. Cluster analysis is based on unsupervised learning, which is based on the similarity in clustering data sets. The K-means algorithm has numerous advantages in the partition-based clustering technique, including basic mathematical ideas, fast convergence, and ease of implementation [5].

ABC analysis is one method that the organisation should explore. ABC has a significant impact on the performance of retail businesses. Companies must ensure that products required by customers are continually available [6]. ABC analysis is a method of inventory classification that divides items into three groups (A, B, and C), with A being the most valued and C being the least valuable. The goal of

this strategy is to focus managers' attention on the critical few (A-items) rather than the insignificant many (C-items).

According to the ABC approach, a corporation should rank things on a scale of A to C, based on the following rules: A-items are goods with the highest yearly consumption value; typically, the top 70-80% of a company's annual consumption value accounts for only 10-20% of total inventory items. B-items are interclassed items with a medium consumption value; they typically account for 30 percent of total inventory items and account for 15-25 percent of annual consumption value. C-items, on the other hand, are the things with the lowest consumption value; typically, the bottom 5% of annual consumption value accounts for 50% of total inventory items.

For a number of brands in the CPG (Consumer packaged goods) business, the Pareto Principle, sometimes known as the 80/20 rule, is used. Pareto analysis is a statistical technique for selecting the small number of activities that have the most important overall effect in decision-making.[7] It is based on the principle of finding the top 20% of reasons that must be addressed in order to solve 80% of the problems. The Pareto Principle describes markets dominated by a limited number of best-selling products, with a small group of buyers accounting for most of the purchasing volume.

The goal is to determine what observable customer and product segmentation algorithm are the need-based market segmentation actionable.

2. METHODOLOGY

For Customer Segmentation, data was taken from a retail store, while for Product Segmentation that data was used. In this work, customer segmentation algorithms and product segmentation analyses were used to get an accurate result.

2.1.1 RFM (Recency, Frequency, Monetary) analysis

Total data is loaded, and total prices are determined in such a way that Quantity is multiplied by price for monetary calculations in RFM analysis. After that, invoice dates for frequency are separated and CustomerIDs for recency are extracted from the datasets, and other columns such as frequency, recency, and monetary value are created.

Then, using a different dataset, Logarithmic Transformation is used to obtain minimum values for speedier computations. Following that, RFM Processing start.

RFM Dataset

	Frequency	Recency	Monetary
12,362.0000	1	1	130.0000
12,435.0000	0	1	1,008.0000
12,490.0000	1	21	603.9400
12,533.0000	1	44	929.9200
12,636.0000	1	1	141.0000

Figure - 1: To begin, the data set is imported and classified according to Invoice, Description, Quantity, Invoice, Date, Price,

CustomerID, Country, Customer City, Age, and Gender. So, while sorting this out, it's broken down into three categories: recency, frequency, and monetary value.

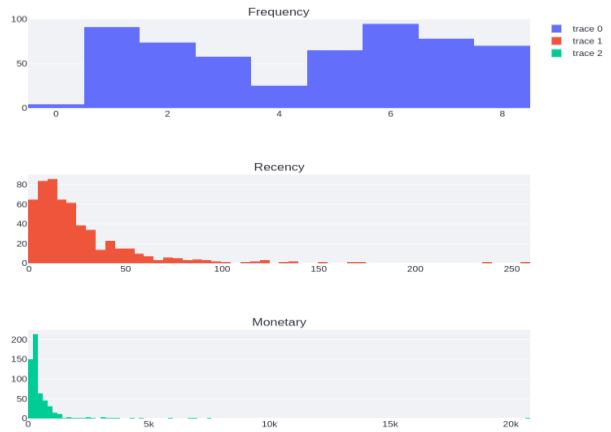

Figure - 2.1: Distribution of the Features

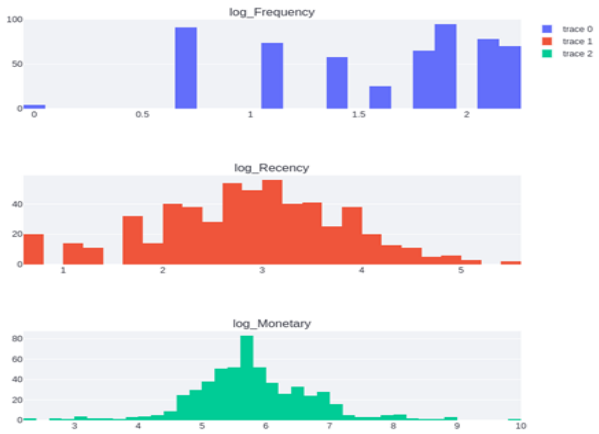

Figure - 2.2: Distribution of Features (Logarithmic)

2.1.2 ELBOW method

How is the cluster formed?

Scikit-learn is a machine learning package that can be used via Python. It may be used to produce accurate forecasts using machine learning techniques. Using the code below, the K-means clustering model was imported into Scikit-learn.

from sklearn.cluster import KMeans

A cluster is a collection of data points that have been grouped together due to particular commonalities.

The number of centroids in the dataset is referred to as the goal number k. A centroid is a fictional or actual place that represents the cluster's center.

By lowering the in-cluster sum of squares, each data point is assigned to one of the clusters.

The number of clusters (K) are altered in the Elbow approach from 1 to n. To calculate WCSS for each value of K. (Within-Cluster Sum of Square). In a cluster, WCSS is the sum of squared distances between each point and the centroid. The plot appears like an Elbow when plotting of the WCSS with the K value. The WCSS value will begin to fall as the number of clusters grows. When K = n, the WCSS value is the highest. When examined graph will shift rapidly at a point, forming an elbow shape. The graph begins to travel practically parallel to the X-axis at this point. The ideal K value, or the optimal number of clusters, corresponds to this point.

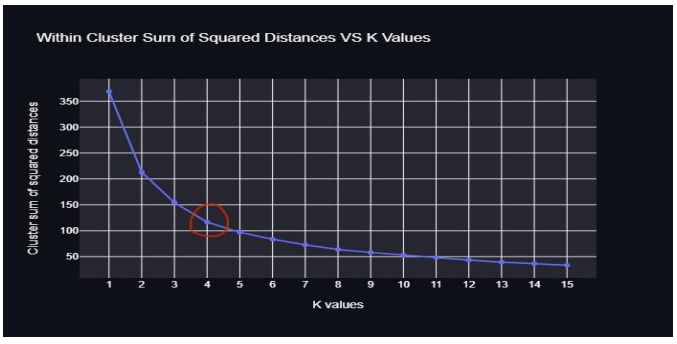

Figure – 3: Within Cluster Sum of Squared Distances VS K Values

The K value, or ideal number of clusters, is x, which is the point at which the elbow shape is formed.

2.1.3 K-MEANS CLUSTERING ALGORITHM:

Step 1: Using the Elbow approach, calculate K and provide the number of clusters K.

Step 2: Assign each data point to a cluster at random.

Step 3: Calculate the coordinates of the cluster's centroid.

Step 4: Calculate the distances between each data point and the centroids, then reassign each point to the cluster centroid with the shortest distance.

Step 5: Determine cluster centroids once more.

Step 6: Repeat steps 4 and 5 until reached a global optimum, when no more improvements are conceivable and data points cannot be switched from one cluster to the next.

2.1.4 Discussion

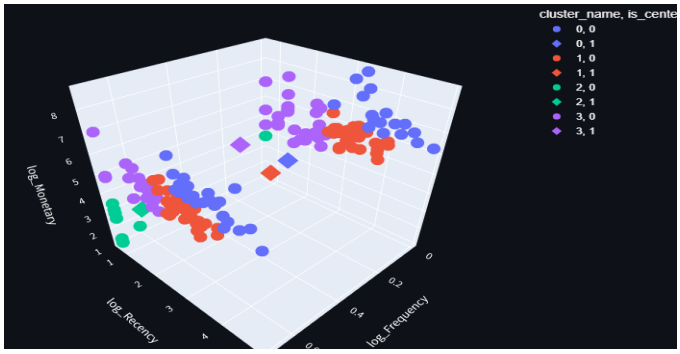

Figure - 4: K-Means Clustering (3D Graph).

As shown in Figure 4, Created a 3D graph of a Clusters utilizing K-Means Algorithmic Data as well as the Python Library matplotlib.pyplot

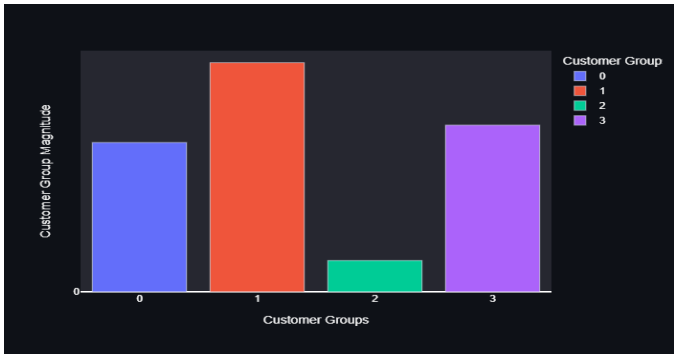

Figure – 5: Cluster Graph as Customer Group

CLUSTER 0

The conclusion from these figures that buyers are purchasing things in a reasonable manner. These clients must be regulars.

CLUSTER 1

The conclusion from these figures that people are buying fewer things, but they are buying more than normal.

CLUSTER 2

The conclusion from these figures that consumers are rookies who have just recently begun using the store, which explains why their frequency and monetary values are low. These beginner clients might be quite valuable if it can entice them with discounts and special offers.

CLUSTER 3

The conclusion from these figures that the clients are rookies who are buying things in a hurry. These are the clients who are most likely to patronize the shop.

With the processing of the preceding algorithms, arrived at the ultimate result of Customer Segmentation, as illustrated in Figure 5.

2.2.1. ABC USING PARETO ANALYSIS:

As illustrated in Figure 6, the Dataset is first put into the model and then segmented into the following parameters: SKU (Stock Keeping Unit), Item, Family, Category, Store, Day, and Quantity. Following that, the ABC analysis, using Pareto as a guide, begins the process of classifying these qualities into three groups: Class A, Class B, and Class C.

	SKU	ITEM	FAMILY	CATEGORY	STORE	DAY	QTY
2304	692	692	2	0	0	1	33
2601	989	989	2	0	0	1	20
2810	1198	1198	2	0	0	1	42
5353	3741	692	2	0	1	1	21
5650	4038	989	2	0	1	1	11

Figure – 6: Loaded Dataset segmented into above parameters.

The Pareto Principle states that in every system, the majority of the results originate from only 20% of the efforts or causes. ABC Analysis revealed the 20% of the goods that offer around 80% of the value based on Pareto's 80/20 Rule.

Pareto figure formation

1. Compiled a list of the categories that were used to organize things.
2. Appropriate measurements were determined. The terms Frequency, Recency, and Monetary are all used interchangeably.
3. Determined the period covered by the Pareto figure: Is it true that there is only one work cycle? Is it truly a full day? Is it true that it's been a week?
4. Gathered data, marking the category each time, or compiled data that was previously available.
5. Added the totals for each of the categories.
6. For the measurements used a scale that was appropriate. The maximum value will be the highest subtotal from step 5. On the figure's left side, took a note of the scale.
7. Each category's bar was created and labelled. The tallest was placed on the far left, followed by the next tallest on the right, and so on. Small-measurement categories can be put together as "other" if there are a lot of them.
8. By dividing the subtotal for each category by the total for all categories, the percentage for each category was calculated. Created a right-hand vertical axis with percentage labels. Double-checked that the two scales are in sync. On the right scale, for example, the left measurement for one-half should be precisely opposite 50 percent.
9. By combining the subtotals for the first and second categories and placing a dot above the second bar to denote the total, Calculated and drew cumulative sums. To symbolize the new total, Added the subtotal for the third category to that amount and placed a dot above the third bar. The technique was continued for the remaining bars. The dots at the top of the first bar were connected. The last dot on the right scale should be at 100 percent.

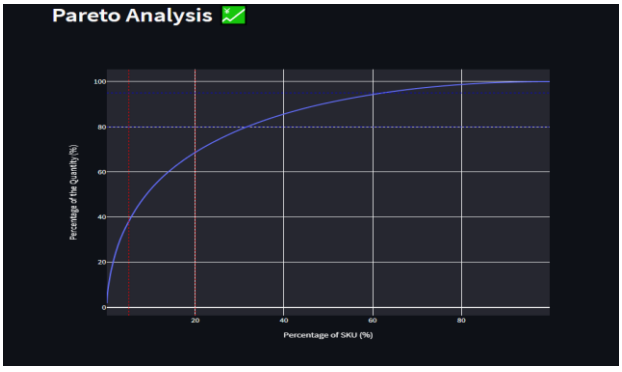

Figure – 7: Pareto Analysis

2.2.2 ABC ANALYSIS WITH DEMAND VARIABILITY

$$CV = (SD/\bar{x}) * 100.$$

CV= Coefficient of Variation

SD= Standard Deviation

$\bar{x}$ = Mean

Demand Linearity, also known as Process Linearity, is a measure of the volatility or dispersion of a time series around the average, such as client orders. The standard deviation to mean ratio is known as the coefficient of variation (CV). Even when the means are different, CV is useful for assessing the degree of variance across distinct time series. The bigger the relative variance or dispersion, the higher the Cv.

In Excel, Calculated the series' standard deviation (s, or) =STDEV.P (cell range).

In Excel, Calculated the mean (x, or) (or average) = AVERAGE (cell range).

The standard deviation to mean ratio was then computed.

Low variability is often defined as a CV less than 1.0, while highly stable demand is defined as a CV less than 0, as shown in graph 8.

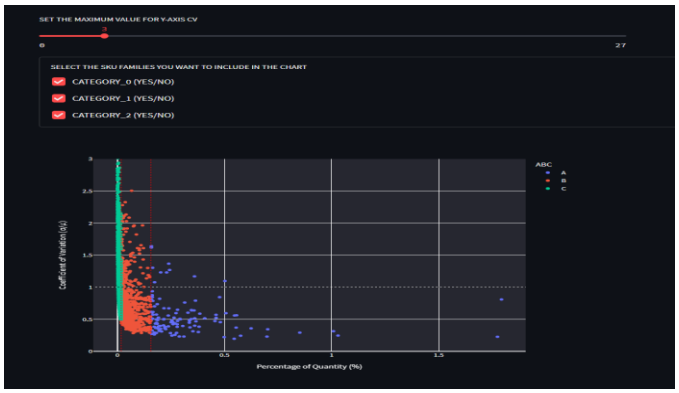

Figure - 8: ABC Analysis with Demand Variability

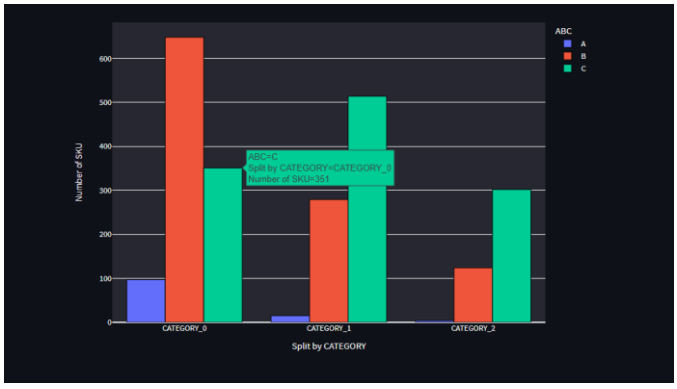

Figure 9. Demand Variability Split by Category

As shown in Figure 9, Plotted three categories with Number of SKUs in ABC analysis with Demand Variability. Each of the three categories (Category 0, Category 1, and Category 2) has three classes: Class A, Class B, and Class C.

2.2.3. THE SHAPIRO-WILK TEST FOR NORMALITY:

Normality refers to normal distribution, often known as the Gaussian distribution or the bell-shaped curve, is a type of statistical distribution. The mean and standard deviation of the data define the normal distribution, which is a symmetrical continuous distribution

The Shapiro-Wilk test generates a W statistic that determines if a random sample, x1, x2, …., xn, is drawn from a normal distribution. Small W values indicate a deviation from normality, as well as percentage points for the W statistic acquired by Monte Carlo simulations.

The following formula is used to obtain the W statistic:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where the ai are constants formed from the means, variances, and covariances of the order statistics of a sample of size n from a normal distribution, and the x(i) are ordered sample values (x (1) being the smallest).

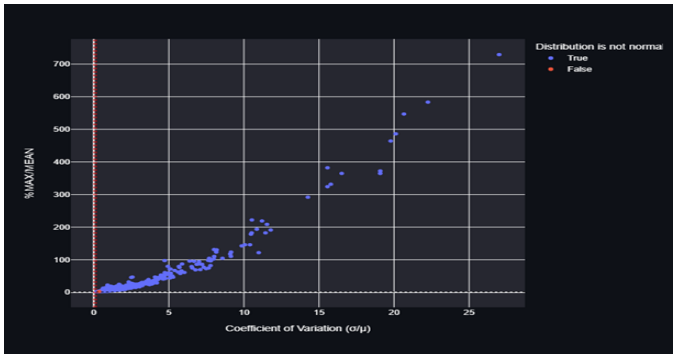

Figure - 10: Normality Test

2.2.4. EXAMPLE OF DISTRIBUTION WITH LOW CV (COEFFICIENT OF VARIATION)

Plotted a graph of daily sales volume with a count to get a figure where the normality of the distribution is less than 0.5, indicating that the distribution is normal.

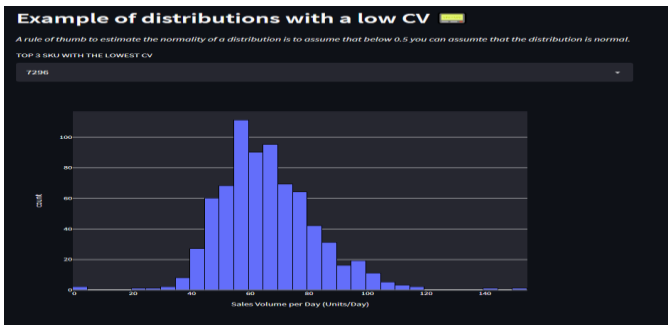

Figure - 11: Distribution with Low CV (Coefficient of Variation)

2.2.5 Discussion

While running all the algorithms on a dataset, concluded that the product demand is stable or not stable, and that the product belongs to whatever class in terms of volatility, volume, and costing. We can determine which marketing strategies should be used in the future to increase the profitability of our firm and product sales.

3. CONCLUSION

Firms must understand their consumer bases and product sales to generate more money. Consequently, developed this prototype to obtain exact and rapid results in order to strengthen their marketing tactics.

Algorithms used like k-means, RFM analysis, ABC analysis with Pareto and Applied the Shapiro-Wilk test for normality in this product. Obtained information on customers' acquisition, purchasing behaviors, and the money they spend on a product by executing this project model.

Products Demand, Volatility and Volume are key parameters of this research. Knowing this allows us to develop or adjust various marketing strategies which benefits in products and customers segmentation and future scope in order to create more income.

4. REFERENCES

[1]. Sulekha Goyat. "The basis of market segmentation: a critical review of literature" in European Journal of Business and Management, 2011

[2]. Wang Shunye, Cui Yeqin, Jin Zuotao and Liu Xinyuan "K-means algorithm in the optimal initial centroids based on dissimilarity" in Journal of Chemical and Pharmaceutical Research, 2013

[3]. Onur DOĞAN, Ejder AYÇİN and Zeki Atıl BULUT. "CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY" in International Journal of Contemporary Economics and Administrative Sciences, 2018

[4]. Sun, J.; Liu, J.; Zhao, L. "Clustering algorithm research." In Journal of Software 2008 [CrossRef]

[5]. Li, X.; Yu, L.; Hang, L.; Tang, X. "The parallel implementation and application of an improved k-means algorithm." in The Journal of Electronic Science and Technology (JEST), 2017

[6]. I M D P Asana, M L Radhitya, K K Widiartha*, P P Santika and I K A G Wiguna "Inventory control using ABC and min-max analysis on retail management information system" in International Conference on Innovation in Research, 2022

[7]. Daniel M. McCarthy and Russell S. Winer "The Pareto rule in marketing revisited: is it 80/ 20 or 70/20?" in Springer, 2019

BIOGRAPHIES

Atul Suryawanshi: Receiving his bachelor's degree in computer science and Engineering from D Y Patil College of Engineering & Technology, Kolhapur, India. It is affiliated to Shivaji University, Kolhapur. His research interests are in Artificial Intelligence, clustering techniques and marketing strategies.

Shantanu Sapkal: Receiving his bachelor's degree in computer science and Engineering from D Y Patil College of Engineering & Technology, Kolhapur, India. It is affiliated to Shivaji University, Kolhapur. His research interests are in Machine Learning, algorithms, and fuzzy sets.

Aniket Patil: Receiving his bachelor's degree in computer science and Engineering from D Y Patil College of Engineering & Technology, Kolhapur, India. It is affiliated to Shivaji University, Kolhapur. His research interests are in Data mining, clustering techniques and marketing strategies.

Satyajeet Vhanbatte: Receiving his bachelor's degree in computer science and Engineering from D Y Patil College of Engineering & Technology, Kolhapur, India. It is affiliated to Shivaji University, Kolhapur. His research interests are in sales analysis, fuzzy sets, and marketing techniques.

Mrs. Shatakshi Kokate: Received her bachelor's and master's degree in computer science and engineering from Shivaji University, Kolhapur. She is a PhD candidate in Rashtrasant Tukdoji Maharaj Nagpur University and faculty member in the Department off Computer Science and Engineering, D Y Patil College of Engineering and Technology, Kolhapur. Her research interests are in cloud computing, blockchain, IoT and computer networks.