INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1645)
Available online at: https://www.ijariit.com

Customer segmentation and prediction analytics in ERP for jewelry domain

Mohammed Mafaz
mafazazhaar@gmail.com
Loyola College, Chennai, Tamil Nadu

ABSTRACT

Customers are the link to a business success. Any organization must focus on a huge number of customer, for this customer satisfaction and loyalty should be incorporated along with the long-term goals of the organization. As the backbone for all marketing activities, customer analytics comprises techniques like predictive modelling, data visualization, information management and segmentation. Customer analytics is becoming critical. Customers have access to information anywhere, any time – where to buy , what to shop for , what proportion to pay, etc. The deeper the understanding of customers' buying habits and lifestyle preferences, the more accurate your predictions of future buying behaviours are going to be – and therefore the more successful you'll be at delivering relevant offers that attract rather than alienate customers. Generally, organizations spend a lot of money to acquire new customers but they do not realize that the majority of the sales and profits come from their existing customers. In this thesis, the existing customers are analyzed and given a customer lifetime score and based on this score, the customers are nurtured by the sales team to increase the profits. In order to maximize sales and conversions, customer segmentation and product recommendation engine is used effectively. Customer segmentation is the process of dividing the customers into many groups that supported common characteristics so companies can market to every group effectively and appropriately. Segmentation is performed using k means clustering. Segmentation allows marketers to raised tailor their marketing efforts to varied audience subsets. It is important to predict the customer segment for any new customer which can be done using supervised classification algorithms such as Logistic Regression, Naïve Bayes, Random Forest, K Nearest Neighbour and Support Vector Classifier.

Keywords— *Customer Analytics, Customer Value Analysis, Customer Lifetime Score, Customer Value and Satisfaction, Customer Segmentation, Customer Loyalty, K Means Clustering, Logistic Regression, Naïve Bayes, Random Forest, K Nearest Neighbour and Support Vector Classifier.*

1. INTRODUCTION

In any organization, Customer Value Analysis plays a vital role. Generally, organizations spend a lot of money to acquire new customers but they do not realize that the majority of the sales and profits come from their existing customers. Hence, it is vital to maximize the profits from existing customers. To retain and maximize the existing customers, customer analysis, nurturing the customers, maximizing the conversions are the three major steps.

1.1 CUSTOMER ANALYTICS

Customer analysis has three main stages: customer profiling, monitoring customer satisfaction and customer value for the organization.

1.1.1 Customer Profiling:

Organization's customer profiling is completed so by examining issues such as:
a) buying behaviour: types of products purchased by customer, quantity ordered, frequency of orders, payment behaviour (method of payment), settlement range.
b) lifestyle (for individual customers): activities, interests and opinions.
c) customer value to the organization: the annual value of purchases made by customer, customer share, net present value of estimated profit generated during the course of the customer relationship.

1.1.2 Customer Value and Satisfaction:

The value offered to the client represents the difference between the entire value offered to the client and therefore the total cost of the client. The size of the benefit offered to the client depends on the following elements: the value of the services, the value of the staff and the value of the image. Consumer perceived value is that the difference between what the buyer receives and what the organization offers, things of various ways of meeting the necessity. Consumers choose a specific product due to the worth offered by the merchandise, so experts use this analysis of customer value offered revealing strengths and weaknesses of the corporate against its competitors. The main steps of this analysis are:

(a) Identify the major attributes and benefits that consumers appreciate - consumers are surveyed about the attributes, benefits and the performance they are looking for in choosing a product or service;
(b) quantitative assessment of the importance of various attributes or benefits - consumers are asked to determine the importance of each attribute and advantage of the product;
(c) assessment of company performance and competitors depending on the attributes and advantages mentioned in relation to their importance;
(d) analyse how consumers like the company's offer in a given segment according to the main competitors - if the company's offer is better than all the attributes mentioned competitors, the company may choose to increase or maintain its price at the level of competitors, and will be able to increase market share;
(e) ongoing monitoring of value for consumers - the company has to regularly carry out research to determine consumer perceived value;

Evaluation of an organization cannot be achieved without knowing the level of customer satisfaction on products, services, staff organization, and communication with the organization. Measuring customer satisfaction are often achieved using various methods:

(a) periodic survey - customer satisfaction can be measured directly and also can be used to determine respondents' intention to buy or wish to recommend the company or brand to other consumers;
(b) mystery shopping, which provides information on the behaviour of sales staff and customer satisfaction thereon;
(c) analysis of consumer complaints - can be used as a method of consumer dissatisfaction, but with many reservations because studies have shown that although consumers may be dissatisfied, only a minority of consumers do complain, while others consider it not worth effort, don't know where or whom to call.

1.1.3 Analysis of the customer to the Organization:

The aim of this step is to give a score out of 100 for each customer based on his or her value to the organization. After the calculation of the score, the organization will categorize them into group such as great customer ($80 \leq$ score ≤ 100), good customer ($60 \leq$ score < 80), average customer ($40 \leq$ score < 60) and poor rated customer (score < 40).

The score can be calculated using Customer Lifetime Value method. This method involves calculating the present value of the entire stream of profits it generates client relationship, assuming a medium term time horizon. Practitioners often considered a period of two to 5 years to estimate the longer term value date. The method provides a quantitative framework for investment planning with customers and helps marketers to adopt a long-term strategy, but the tactic requires accurate estimates of revenue from each customer and therefore the costs involved them (visits, promotion , discounts).

$$\mathrm{CLV} = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} \quad \text{------ (Equation 1)}$$

t (from 1 to T) is that the unit of your time (month, year etc..)
T - estimated timeframe
pt - the customer purchases at time t (revenue from customer at time t)
ct - the direct costs associated with customer
rt - the probability of purchase recurrence by the client or the existence of at time
i - rate of interest (cost of capital that's wont to calculate internet present value) the corresponding unit of time t
After analysing the customer value, the organization needs to nurture the great and the good customers to personalize the customer-organization relationship.

1.2 NURTURING CUSTOMERS

Nurturing customers is an overlooked aspect of marketing. It is really a low-risk, high reward activity in terms of long-term revenue as it can increase your customer lifetime value. It improves the strength of your brand over time, creating important ambassadors which will help reinforce the strength of your brand against the competition.

It is important to remember that we cannot recoup all existing customers. Unfortunately, some customers are truly “one-time” customers. This is where the art of nurturing, a low-risk high reward tactic, can start yielding profitable results.

1.2.1 Benefits of Nurturing Customers:

Create happy and loyal customers that can become your biggest brand ambassadors. We all know that brand ambassadors can create a positive influence for your brand.
b) Keeps your products on top of mind by frequent sales call. They’ll remember you when they are ready to buy.
c) It’s a more effective way for profitability as retained customers cost less than new customer acquisition.
d) Helps strengthen your brand. Who wouldn’t want to stay buying from a robust brand?

1.2.2 Steps involved in nurturing a customer

To nurture a customer within the jewellery domain, their purchase behaviour will play a serious role. Most of the nurturing during this domain is completed through messaging and calling. Few samples of nurturing are given below:

(a) Reminding the customer, that you simply haven't visited the buy 272 days will make the customer feel happy and privileged that the shop is keeping track of everything.

(b) Giving birthday/anniversary wishes to the customer. As people in India generally buy jewellery during anniversary period.

(c) Wishing Merry Christmas to Christians, Eid wishes to Muslims, Diwali wishes to Hindus and lots of more will create a customized experience for the customer and usually , people buy jewellery during the festival week. Such nurturing models are often built as an algorithm where the marketing team will have a dashboard and that they will know which customer to call during specific period of time .

1.3 MAXIMIZING CONVERSIONS

To maximize conversions, customer segmentation and merchandise recommendation should be used effectively.

1.3.1 Customer Segmentation:

Customer segmentation is that the process of dividing customers into groups supported common characteristics so companies can market to every group effectively and appropriately. Segmentation allows marketers to raised tailor their marketing efforts to varied audience subsets. Those efforts can relate to both communications and merchandise development. Specifically, segmentation helps a company:

a) Create and communicate targeted marketing messages which will resonate with specific groups of consumers , but not with others (who will receive messages tailored to their needs and interests, instead).

b) Select the simplest channel for the segment, which could be email, social media posts, radio advertising, or another approach, counting on the segment.

c) Identify ways to enhance products or new product or service opportunities.

d) Establish better customer relationships.

e) Test pricing options.

f) Focus on the foremost profitable customers.

g) Improve customer service.

h) Upsell and cross-sell other products and services.

Common characteristics in customer segments can guide how a corporation markets to individual segments and what products or services it promotes to them. alittle business selling hand-made guitars, for instance , might plan to promote lower-priced products to younger guitarists and higher-priced premium guitars to older musicians supported segment knowledge that tells them that younger musicians have less income than their older counterparts. Similarly, a meals-by-mail service might emphasize convenience to millennial customers and “tastes-like-mother-used-to-make” benefits to baby boomers.

Customer segmentation are often practiced by all businesses no matter size or industry and whether or not they sell online or face to face. It begins with gathering and analysing data and ends with working on the knowledge gathered during a way that's appropriate and effective.

2. METHODOLOGY

2.1 ABOUT THE DATASET

The dataset used for study is a secondary data recorded by ABC Ventures which consists of the information of each customer who have purchased any jewellery in Chennai and Bangalore in the recent 5 years in the ABC Ventures client’s shop.

This dataset is composed by the following features:

CustomerID: Unique ID assigned to the customer
Gender: Gender of the customer
Age: Age of the customer
Marital Status: Single/Married
Annual Income ('0000): Annual Income of the custome***Amount Spent***: Amount Spent by the customer in the purchasing jewellery in the last 5 years.

Method of sampling used in the analysis is Stratified sampling. At first, the whole dataset was divided into 3 sub parts (high amount spent, medium amount spent and low amount spent) and then from each subpart, using simple random sampling, 100 samples are chosen. Stratified sampling improves the accuracy and representativeness of the results by reducing sampling bias.

2.2 TOOLS

- Jupyter Notebook
- Google Colab
- Google Drive
- Microsoft Office
- Programming Language: Python

Figure 2. Logo of the tools used in the analysis

2.3.1 K MEANS CLUSTERING:

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different.

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
   - Compute the sum of the squared distance between data points and all centroids.
   - Assign each data point to the closest cluster (centroid).
   - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The objective function is

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

------ (Equation 2)

2.3.2 LOGISTIC REGRESSION

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

------ (Equation 3)

2.3.3 SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

SVM can be of two types

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

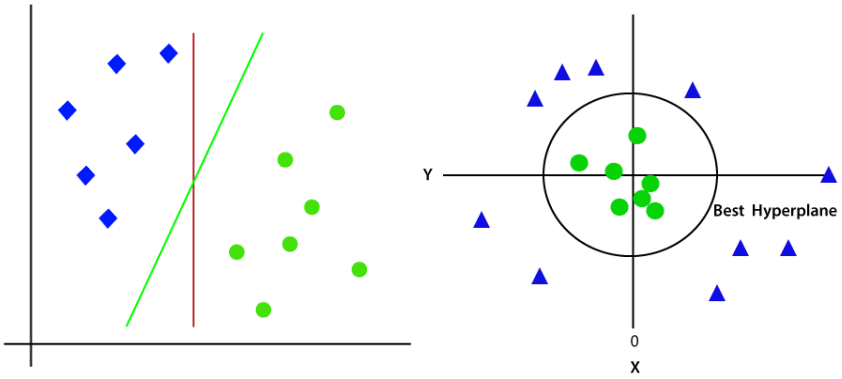

Figure 3. Linear and Non Linear SVM

2.3.4 K NEAREST NEIGHBOURS:

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute. Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set.

2.3.5 GAUSSIAN NAIVE BAYES:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve, which is symmetric about the mean of the feature values. The likelihood of the features is assumed Gaussian; hence, conditional probability is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

------ (Equation 4)

2.3.6 RANDOM FOREST:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

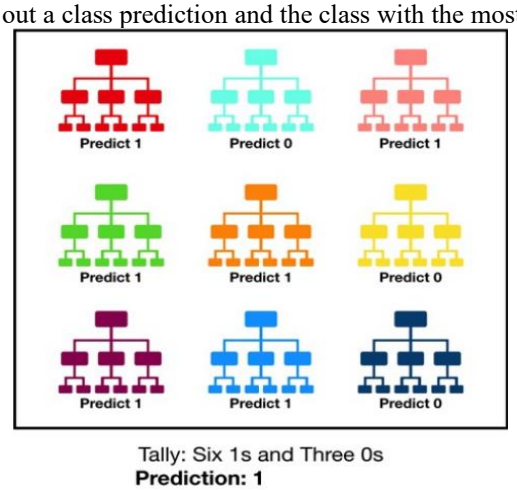

Figure 4.Random Forest Algorithm

2.4 METRICS OF EVALUATION

2.4.1 SUM OF SQUARED ERROR:

SSE is used to find optimal number of clusters. SSE is defined as the sum of the squared distance between centroid and each member of the cluster. Then plot a K against SSE graph. We will observe that as K increases SSE decreases as change will be small. So the idea of this algorithm is to choose the value of K at which the graph decrease abruptly. This sort of produces a "elbow effect" in the graph. The point where there is an elbow effect is considered as ideal number of clusters (k).

2.4.2 CONFUSION MATRIX:

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. Using the four values, important metrics such as Recall, Precision, Specificity, and Accuracy are calculated. It is extremely useful for measuring AUC-ROC Curve.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 5. Confusion Matrix

2.4.3 ACCURACY SCORE:

Accuracy represents the number of correctly classified data instances over the total number of data instances.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$ ------ (Equation 5)

3. RESULTS AND DISCUSSION

3.1 DATA EXPLORATION

The dataset used for analysis has 200 observations and 6 variables with reference to Figure 4.1.

```
(200, 6)
```

Figure 6. The number of observations in the dataset

seaborn.countplot() method is used to show the counts of observations in each categorical bin using bars. From figure 7, it is noted that the dataset has more number of females when compared to males.

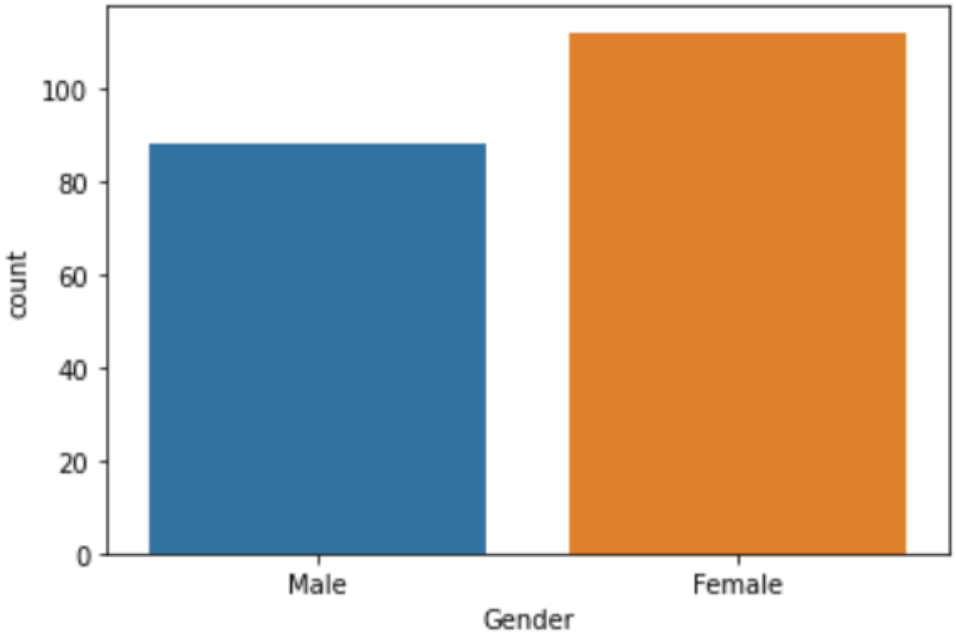

Figure 7. Countplot of Gender variable

From figure 8 below, it is clearly seen that the dataset is very imbalanced regarding marital status and this is expected as more married people purchase jewellery when compared to the singles.

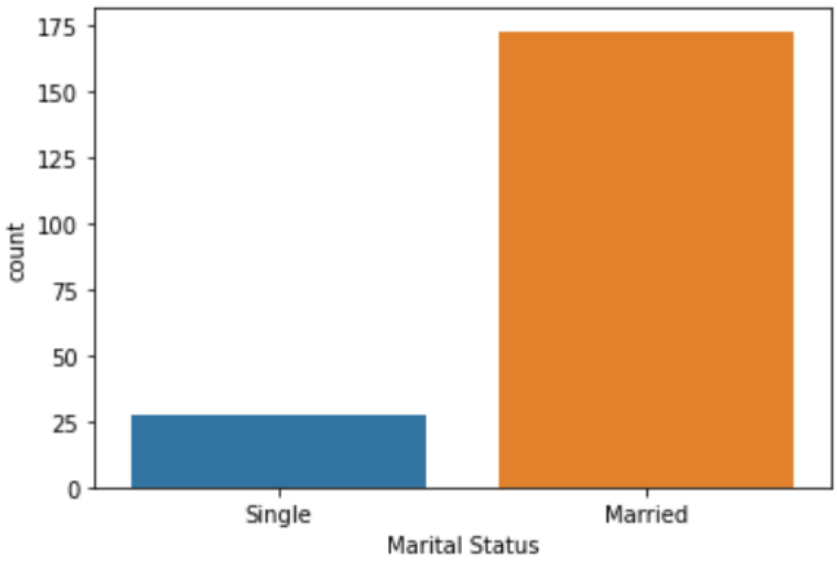

Figure 8. Countplot of Marital Status variable

From figure 9, it is inferred that most of the jewellery purchasing customers in the sample dataset have an age between 26-35

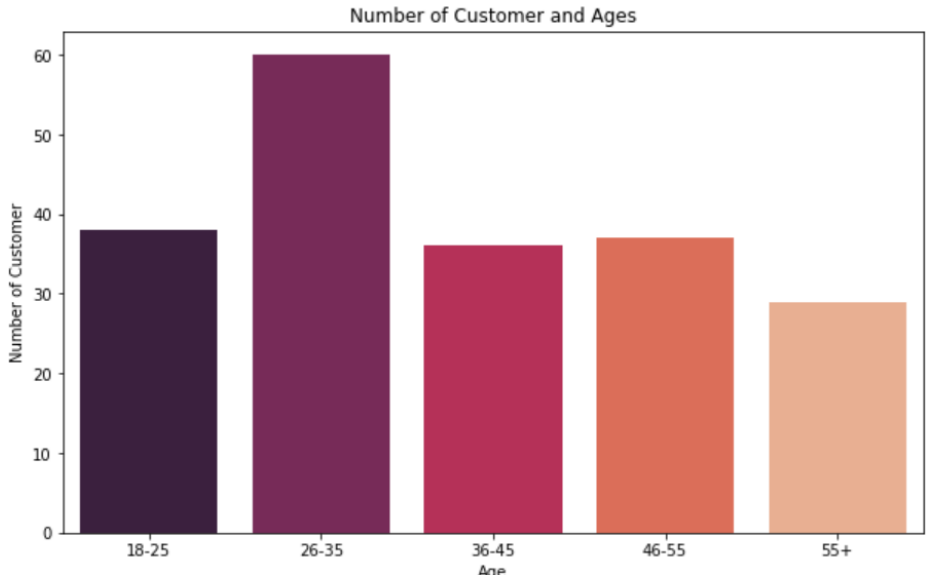

Figure 9. Age wise analysis

From figure 10, it is inferred that most of the jewellery purchasing customers in the sample dataset have an age between 26-35.

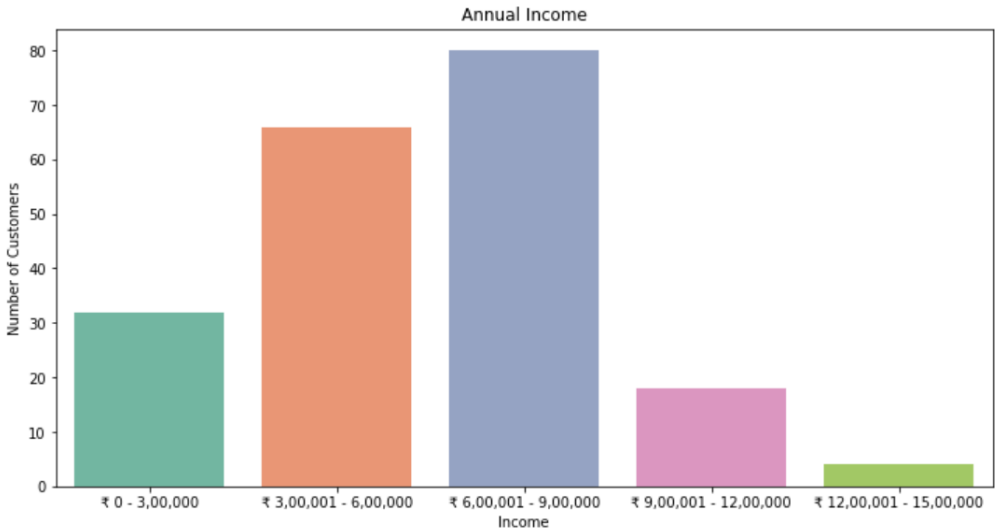

Figure 10. Income Analysis

3.2 CALCULATION OF CUSTOMER LIFETIME VALUE (CLV SCORE)

CLV SCORE FORMULA

$$\boldsymbol{CLV\ SCORE} = (\boldsymbol{A.S} \times \mathbf{100}) \div (\boldsymbol{Age} \times \boldsymbol{A.I} \times \boldsymbol{M.S.P})\text{-----(Equation 6)}$$

where,

Table 2. Abbreviation used in CLV formula

Abbreviations	**Expansion**
CLV	Customer Lifetime Value
A.S	Amount Spent
A.I	Annual Income
M.S.P	Marital Status Points*

*marital status points = 50 if married; marital status points = 20 if single.

From figure 11, it is inferred that most of the jewellery purchasing customers in the sample dataset have a CLV Score between 41-60. Customers with CLV Score > 60 are considered as good existing customers.

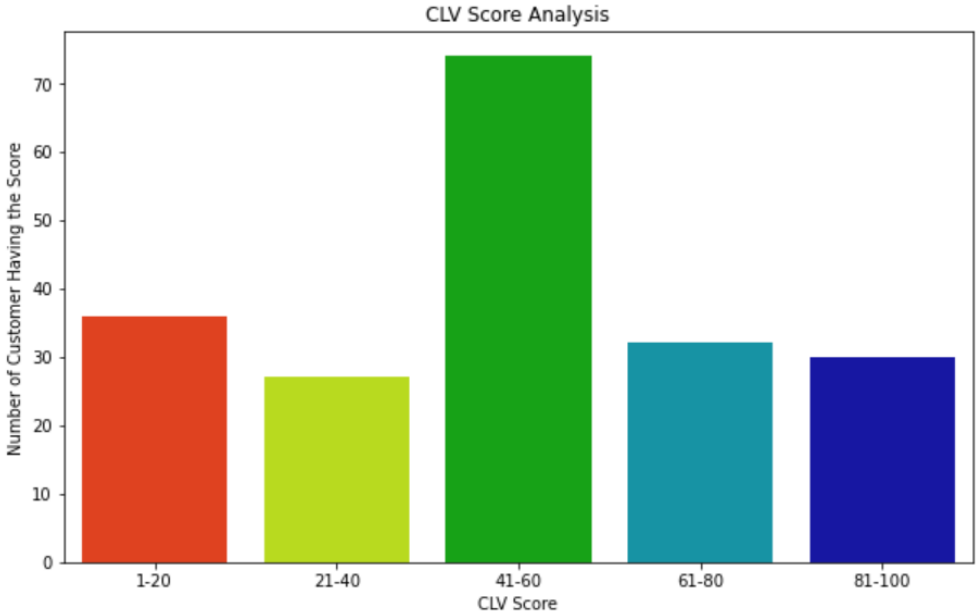

Figure 11. CLV Score Analysis

3.3 K MEANS CLUSTERING

In cluster analysis, the elbow method is commonly used in determining the number of clusters in a data set. A plot, which is a line chart of the SSE for each value of *k*. If the line chart looks like an arm, then the "elbow" on the arm is the value of *k* that is the best. It is clear from the figure that we should take the number of clusters equal to 5, as the slope of the curve is not steep enough after it.

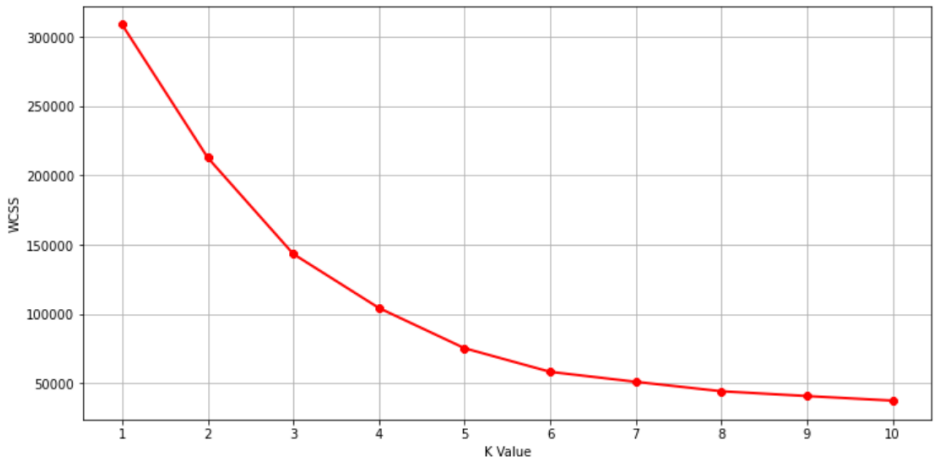

Figure 12. Elbow point Identification

From figure 13, it is clearly seen that the five cluster points are distinctly plotted. It is seen that the customers can be broadly grouped into 5 groups based on their CLV Score. Age and Annual Income.

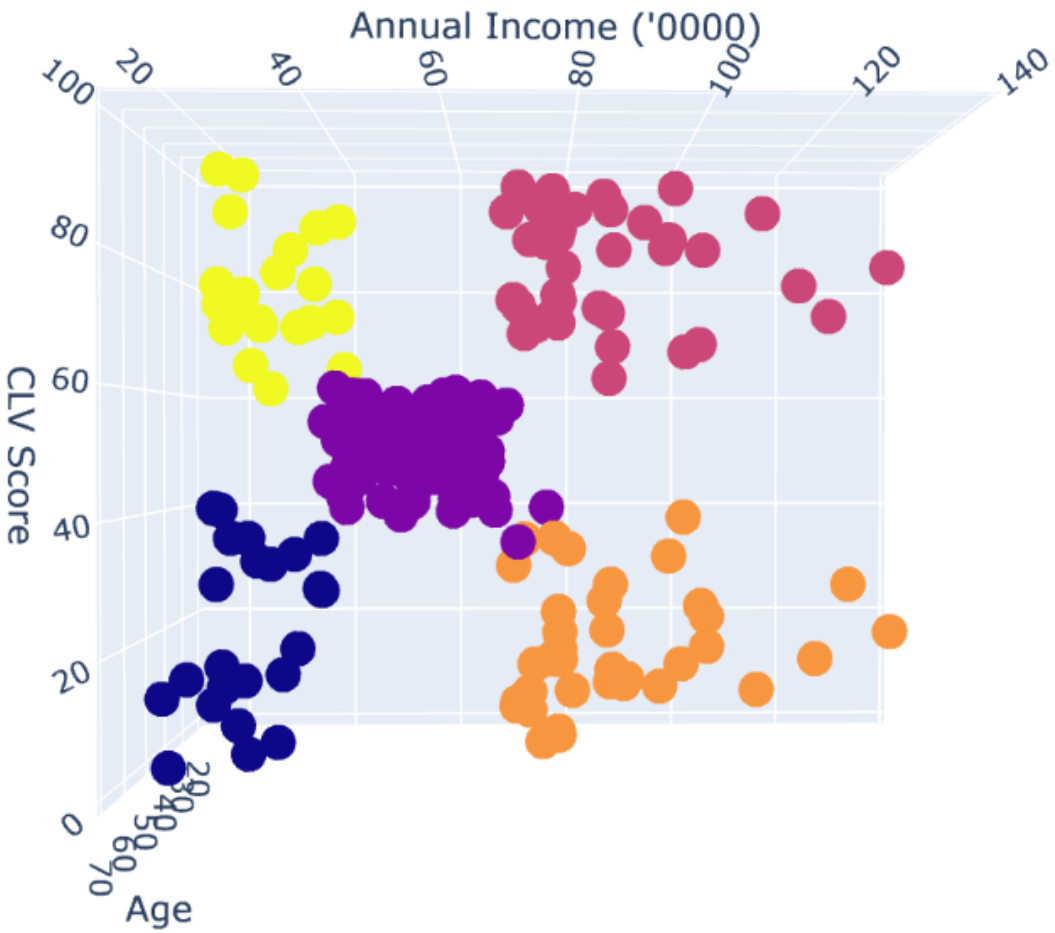

Figure 13. K Means Clustering

From figure 14, it is inferred that most of the jewellery purchasing customers in the sample dataset fall under cluster 1.

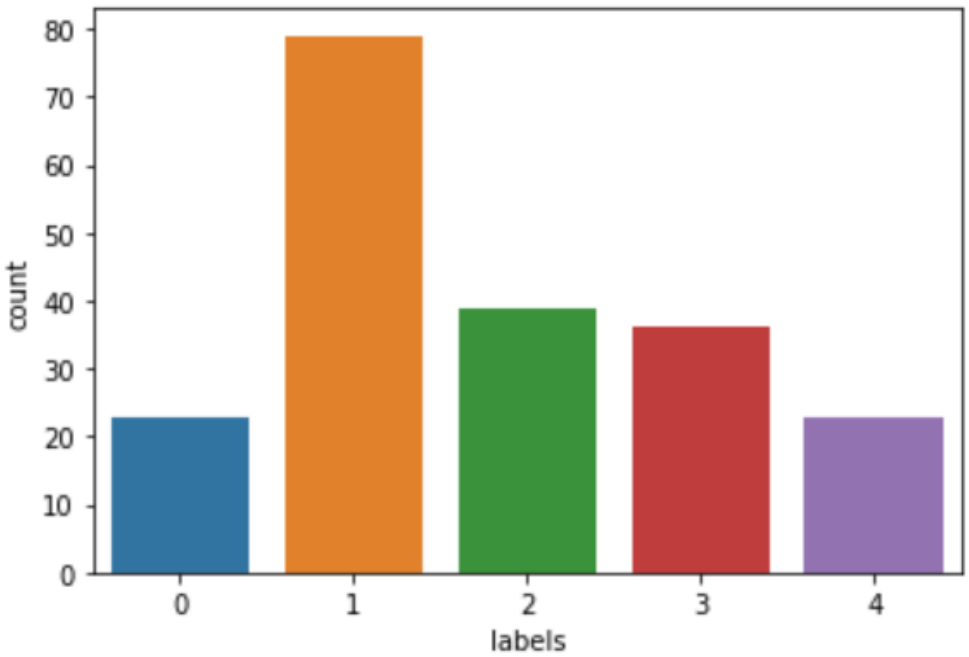

Figure 14. Cluster Analysis

3.4 CLASSIFICATION ALGORITHMS

From figure 15, it is inferred that the overall accuracy of the Logistic Regression model is 92%. The accuracy of cluster 1 is 82%, accuracy of cluster 2 is 100%, accuracy of cluster 3 is 100%, accuracy of cluster 4 is 77%, and accuracy of cluster 5 is 100%.


```
Overall accuracy score is  0.9166666666666666
```

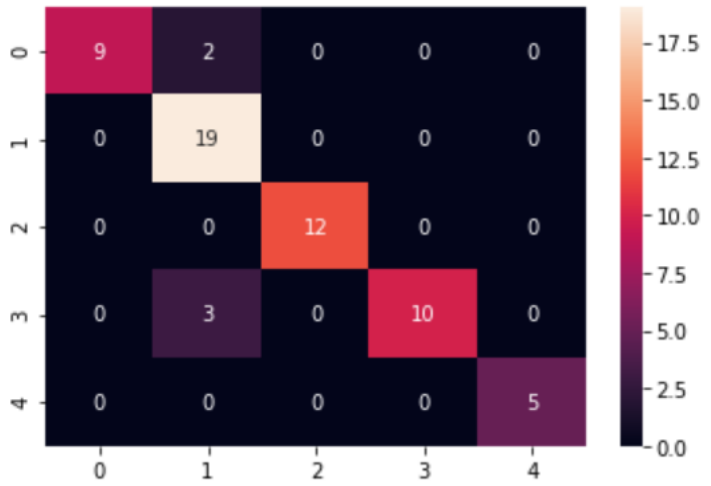

Figure 15. Results of Logistic Regression

From figure 16, it is inferred that the overall accuracy of the Support Vector Machine model is 92%. The accuracy of cluster 1 is 91%, accuracy of cluster 2 is 100%, accuracy of cluster 3 is 100%, accuracy of cluster 4 is 77%, and accuracy of cluster 5 is 80%.

```
Overall accuracy score is  0.9166666666666666
```

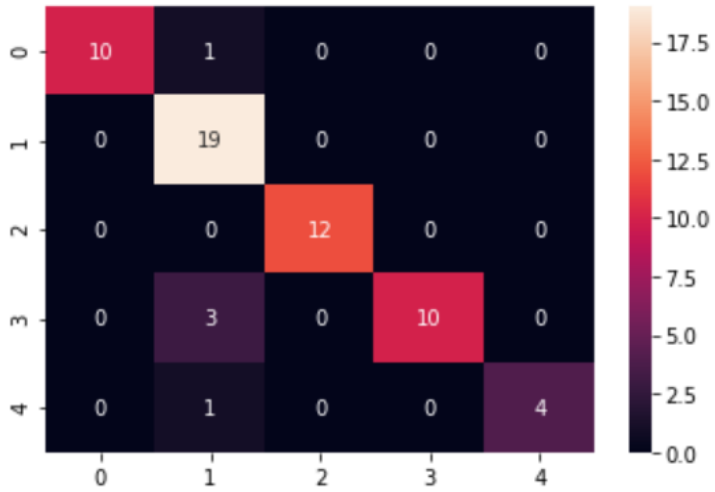

Figure 16. Results of Support Vector Machine

From figure 17, it is inferred that the overall accuracy of the K Nearest Neighbour model is 75%. The accuracy of cluster 1 is 27%, accuracy of cluster 2 is 100%, accuracy of cluster 3 is 100%, accuracy of cluster 4 is 62%, and accuracy of cluster 5 is 60%.

```
Overall accuracy score is  0.75
```

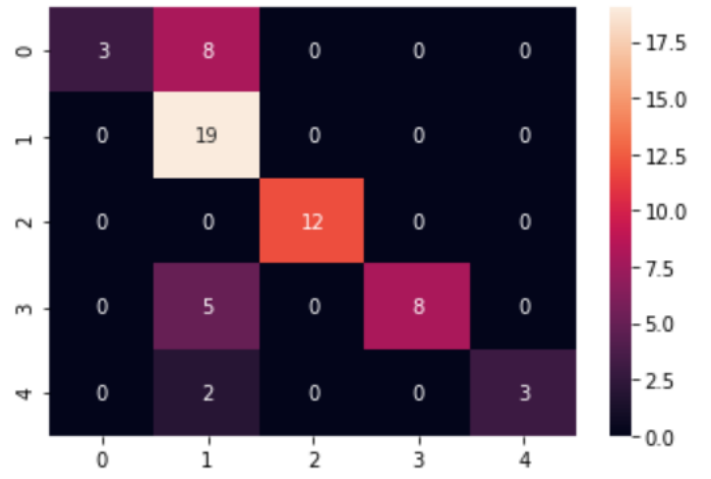

Figure 17. Results of K Nearest Neighbour

From figure 18, it is inferred that the overall accuracy of the Gaussian Naïve Bayes model is 88%. The accuracy of cluster 1 is 91%, accuracy of cluster 2 is 95%, accuracy of cluster 3 is 100%, accuracy of cluster 4 is 77%, and accuracy of cluster 5 is 60%.

```
Overall accuracy score is  0.8833333333333333
```

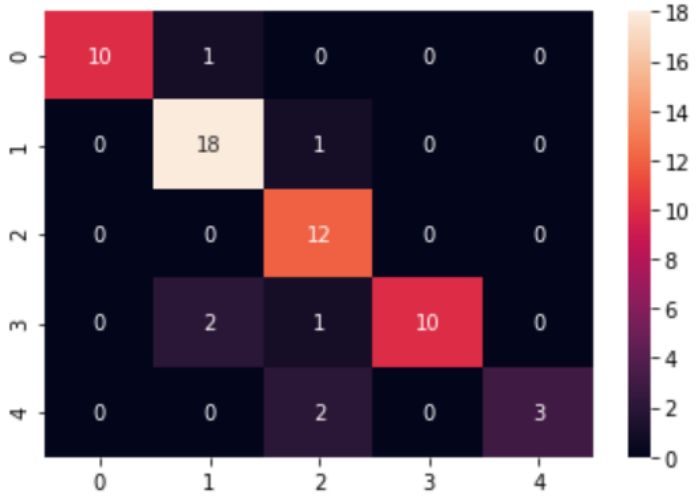

Figure 18. Results of Gaussian Naïve Bayes

From figure 19, it is inferred that the overall accuracy of the Random Forest model is 97%. The accuracy of cluster 1 is 91%, accuracy of cluster 2 is 100%, accuracy of cluster 3 is 92%, accuracy of cluster 4 is 100%, and accuracy of cluster 5 is 100%.

```
Overall accuracy score is  0.9666666666666667
```

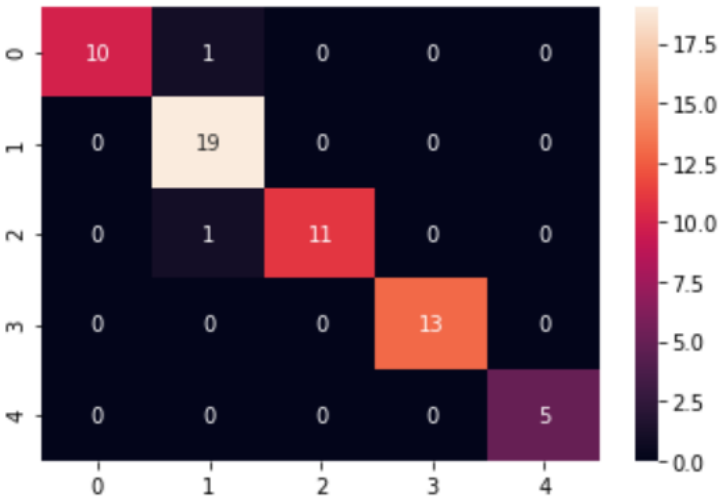

Figure 19. Results of Random Forest

4. CONCLUSIONS

1. Marketing successfully applied, requires the organization to possess the power to know customer value, create value and to supply value to the buyer . during this respect, it's necessary to possess a periodic evaluation of organization marketing performance in reference to objectives and resources consumed. Performance evaluation and control of the company's marketing are often achieved through marketing audit, a crucial component of strategic marketing planning. To be effective and serve the aim , marketing audit should cover all major activities of the enterprise. The audit indicates directions for future action, corrective plans incorporated within the short and future to extend the general efficiency of the organization. within the audit of selling , customer analysis plays a crucial role in three ways: in determining the right target market and consumer profile, the analysis of consumer value offered by the corporate through the strategies applied in product, price, distribution, promotion, and analysis of customer value for the corporate .
2. Cluster Analysis of the chosen sample of respondents explained tons about the possible segments, which existed within the target customer population. Once the amount of clusters was identified, a k-means clustering algorithm, which may be a non-hierarchical method, was used. For computing k-means clustering, the initial cluster centres were chosen and continuing number of iterations computed then final stable cluster centres until means had stopped further changing with next iterations. This convergent condition was also achieved by setting a threshold value for change within the mean. the ultimate cluster centres contained the mean values for every variable in each cluster
3. After computing the five clusters of segments, multiple supervised classification ML algorithm were wont to build a prediction model which can be utilized in future by the organizations to cluster the new customers.
4. In summary, a worth has been assigned to every customer supported the info , which is named CLV Score. Using CLV Score and other data points, segmentation of the purchasers is performed by the tactic of k means clustering. After

customer segmentation, many machine learning algorithms are built to predict the customer segment for any new customer; Random Forest algorithm gave the simplest accuracy (97%), hence this model is chosen for predictive analysis.

5. FUTURE ENHANCEMENTS

- The present study is limited to seven variables. There are many other variables such as DOB (Date of Birth), Marriage Anniversary, Number of Children and many more. These variables may be incorporated in developing future segmentation models.
- Certain modified sorts of Clustering, Supervised Classification ML Algorithms could also be considered with regard to Segmentation and Predictive modelling.
- The future work could involve more trials and automation of the market forecasting and planning.
- To maximize conversions, along side customer segmentation, product recommendation engine could even be built.

6. REFERENCES

[1] Aman Banduni (2017), "Customer Segmentation Using Machine Learning", International Journal of Engineering and Advanced Technology.
[2] Arora S. (2016), "Recommendation Engines: How Amazon And Netflix Are Winning The Personalization Battle"; Martechadvisor International.
[3] Baker.M, (2008), "The Strategic Marketing Plan Audit", Cambridge Strategy Publication Ltd.
[4] Balan,C, (2007), "Customer Relationship Management: Strategic, Operational And Analytical Aspects", Revista De Marketing.
[5] Blanchard et. al (2019), "Marketing Analytics Scientific Data: Achieve Your Marketing Objectives With Python's Data Analytics Capabilities"
[6] Ghalayini Dnam (2017), "Develop Insight Driven Customer Experiences Using Big Data And Advanced Analytics".
[7] Hitachi M (2015), "Customer 360-Degree View Reduce Customer Churn And Identify New Revenue Opportunities, Blueprint For Big Data Success".
[8] Jeffrey Spiess (2014), "Using Big Data To Improve Customer Experience And Business Performance", IEEE Explore Digital Library.
[9] Juni Nurma Sari, Lukito Nugroho (2016), "Review On Customer Segmentation Technique On Ecommerce", Journal Of Computational And Theoretical Nanoscience.
[10] Kabu Khadka, Soniya Maharjan (2017), 'Customer Satisfaction And Customer Loyalty', Centria University Of Applied Sciences.
[11] Kotler and Keller (2009), "Marketing Management", Pearson International Edition.
[12] Laika Satish (2017), "A Review: Big Data Analytics For Enhanced Customer Experiences With Crowd Sourcing".
[13] Mostafa Shabani (2015), "New Approach To Customer Segmentation Based On Changes In Customer Value".
[14] Parmerlee.D (2000), "Auditing Markets, Products, And Marketing Plans", NTC Business Books.
[15] Pelau.C (2008), "Marketing Controlling On Consumer Goods Market", Cambridge Strategy Publication Ltd.
[16] Radulescu D.M, Radulescu.V (2009), "Legal And Ethical Principles In Establishing The Prices Of The Goods And Services", Cetina Metalurgia International.
[17] Sunil Erevelles NF (2015), Big Data Consumer Analytics and The Transformation Of Marketing, Elseiver Journal.
[18] Tymoteusz Doligalski (2016) "Customer Analysis And Firm Performance In The Polish Insurance Market. Perspective Of Customer Profitability And Lifetime Value".
[19] Violeta Radulescu and Iuliana Cetina (2008),"Customer Analysis, Defining Component Of Marketing Audit", Procedia - Social And Behavioural Sciences.
[20] Yash Kushwaha And Deepak Prajapati (2018), "Customer Segmentation Using K-Means Algorithm", International Journal Of Creative Research Thoughts.

APPENDIX

CODING

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, accuracy_score,classification_report
df = pd.read_csv("Customer_Data_New.csv")
df.head()
df.shape
sns.countplot(df.Gender)
sns.countplot(df['Marital Status'])
age18_25 = df.Age[(df.Age <= 25) & (df.Age >= 18)]
age26_35 = df.Age[(df.Age <= 35) & (df.Age >= 26)]
age36_45 = df.Age[(df.Age <= 45) & (df.Age >= 36)]
age46_55 = df.Age[(df.Age <= 55) & (df.Age >= 46)]
age55above = df.Age[df.Age >= 56]
x = ["18-25","26-35","36-45","46-55","55+"]
y=[len(age18_25.values),len(age26_35.values),len(age36_45.values),len(age46_55.values),len(age55above.values)]
plt.figure(figsize=(10,6))
```

```
sns.barplot(x=x, y=y, palette="rocket")
plt.title("Number of Customer and Ages")
plt.xlabel("Age")
plt.ylabel("Number of Customer")
plt.show()
ai0_30 = df["Annual Income ('0000)"][(df["Annual Income ('0000)"] >= 0) & (df["Annual Income ('0000)"] <= 30)]
ai31_60 = df["Annual Income ('0000)"][(df["Annual Income ('0000)"] >= 31) & (df["Annual Income ('0000)"] <= 60)]
ai61_90 = df["Annual Income ('0000)"][(df["Annual Income ('0000)"] >= 61) & (df["Annual Income ('0000)"] <= 90)]
ai91_120 = df["Annual Income ('0000)"][(df["Annual Income ('0000)"] >= 91) & (df["Annual Income ('0000)"] <= 120)]
ai121_150 = df["Annual Income ('0000)"][(df["Annual Income ('0000)"] >= 121) & (df["Annual Income ('0000)"] <= 150)]
aix = ["₹ 0 - 3,00,000", "₹ 3,00,001 - 6,00,000", "₹ 6,00,001 - 9,00,000", "₹ 9,00,001 - 12,00,000", "₹ 12,00,001 - 15,00,000"]
aiy = [len(ai0_30.values), len(ai31_60.values), len(ai61_90.values), len(ai91_120.values), len(ai121_150.values)]
plt.figure(figsize=(12,6))
sns.barplot(x=aix, y=aiy, palette="Set2")
plt.title("Annual Income")
plt.xlabel("Income")
plt.ylabel("Number of Customers")
plt.show()
df['Marital Status Points']='NaN'
a=list(range(200))
for i in a:
    if df['Marital Status'][i]=='Single':
        df['Marital Status Points'][i]=20
    else:
        df['Marital Status Points'][i]=50
df.drop(['Marital Status'],axis=1,inplace=True)
df['CLV Score']=(df['Amount Spent']/(df['Age']*df["Annual Income ('0000)"]*df['Marital Status Points']))*100
ss1_20 = df["CLV Score"][(df["CLV Score"] >= 1) & (df["CLV Score"] <= 20)]
ss21_40 = df["CLV Score"][(df["CLV Score"] >= 21) & (df["CLV Score"] <= 40)]
ss41_60 = df["CLV Score"][(df["CLV Score"] >= 41) & (df["CLV Score"] <= 60)]
ss61_80 = df["CLV Score"][(df["CLV Score"] >= 61) & (df["CLV Score"] <= 80)]
ssx = ["1-20", "21-40", "41-60", "61-80", "81-100"]
ssy = [len(ss1_20.values), len(ss21_40.values), len(ss41_60.values), len(ss61_80.values), len(ss81_100.values)]

plt.figure(figsize=(10,6))
sns.barplot(x=ssx, y=ssy, palette="nipy_spectral_r")
plt.title("CLV Score Analysis")
plt.xlabel("CLV Score")
plt.ylabel("Number of Customer Having the Score")
plt.show()
df['Gender']=pd.get_dummies(df['Gender'])
df
from sklearn.cluster import KMeans
wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(df.iloc[:,[3,4,6]])
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
X=df.iloc[:,[3,4,6]]
kmeans = KMeans(n_clusters = 5 , init = 'k-means++',random_state = 0)
y_kmeans = kmeans.fit(X)
print(kmeans.labels_)
X['labels']=kmeans.labels_
import plotly.express as px
ax=px.scatter_3d(data_frame=X,x='Age',y="Annual Income ('0000)",z='CLV Score',color='labels')
ax.show()
df['labels']=X['labels']
X=df.drop(['CustomerID','labels'],axis=1)
y=df['labels']
```



```
sns.countplot(y)
from sklearn.model_selection import train_test_split
Xtrain,Xtest,ytrain,ytest = train_test_split(X,y,test_size=0.3,random_state=0)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
Xtrain = sc.fit_transform(Xtrain)
Xtest = sc.transform(Xtest)
from sklearn.linear_model import LogisticRegression
LRG=LogisticRegression()
LRG.fit(Xtrain,ytrain)
ypred=LRG.predict(Xtest)
cm=confusion_matrix(ytest,ypred)
sns.heatmap(cm, annot=True)
print('Overall accuracy score is ',accuracy_score(ytest,ypred))
print(classification_report(ytest,ypred))
from sklearn.svm import SVC
SVC = SVC(kernel='rbf')
SVC.fit(Xtrain,ytrain)
ypred=SVC.predict(Xtest)
cm=confusion_matrix(ytest,ypred)
sns.heatmap(cm, annot=True)
print('Overall accuracy score is ',accuracy_score(ytest,ypred))
print(classification_report(ytest,ypred))
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier(n_neighbors = 10, metric = 'minkowski', p = 7)
KNN.fit(Xtrain,ytrain)
ypred=KNN.predict(Xtest)
cm=confusion_matrix(ytest,ypred)
sns.heatmap(cm, annot=True)
print('Overall accuracy score is ',accuracy_score(ytest,ypred))
print(classification_report(ytest,ypred))
from sklearn.naive_bayes import GaussianNB
GNB = GaussianNB()
GNB.fit(Xtrain,ytrain)
ypred=GNB.predict(Xtest)
cm=confusion_matrix(ytest,ypred)
sns.heatmap(cm, annot=True)
print('Overall accuracy score is ',accuracy_score(ytest,ypred))
print(classification_report(ytest,ypred))
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
classifier.fit(Xtrain,ytrain)
ypred=classifier.predict(Xtest)
cm=confusion_matrix(ytest,ypred)
sns.heatmap(cm, annot=True)
print('Overall accuracy score is ',accuracy_score(ytest,ypred))
print(classification_report(ytest,ypred))
```

