International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 02, 2021

Customer Segmentation Analysis

P. Lisasri[1] , N Rajeswari[2]

[1]MCA Student, Department of Computer Applications, PSG College of Technology Coimbatore, India
[2]Assistant Professor(Sr.Gr), Department of Computer Applications, PSG College of Technology Coimbatore, India

Eamil: [1]lisasri25@gmail.com, [2]nrj.mca@psgtech.ac.in

Abstract: We experience a daily reality such that huge measures of information are gathered each day. Examination of this information is a significant necessity. In the advanced time of development, there is a huge scope contest that is superior to all, and business systems should be completed by present-day conditions. The present business is based on imaginative thoughts, in light of the fact that a lot of potential customers are befuddled about what to purchase and what not to purchase. Organizations working together can't analyze target leads. This is the place where machine learning comes in, applying different calculations to recognize stowed away examples in information to settle on better choices. The idea of target customer group is refined through the customer segmentation using clustering technology.

Keywords : Segmentation, K-Means, Principle Component Analysis, Logistic Regression, Linear Regression

1. INTRODUCTION

Today, one of the fundamental objectives of the business is to comprehend customer conduct and arrange it into proper gatherings dependent on the outcomes got. Organizations are searching for answers to questions, for example, who are the most mainstream customers and what items can draw in the most customers on the racks of retail locations to expand deals and income. Along these lines, it very well may be underlined that the issue of this project is customer segmentation, which permits organizations to all the more likely comprehend the conduct of customers, in order to all the more successfully address their issues. Accordingly, the motivation behind this project is to assess the relationship of unsupervised AI calculations to track down the optimal number of customer groups.

S
T
P
Segmentation, Targeting, and Positioning
FRAMEWORK

Fig 1.1 Customer Segmentation Model

In Customer Analytics, STP frameworks play a major role to improve the business as show in fig 1.1. Customer Segmentation helps to divide customers into groups or clusters based on common characteristics. This project segment customer using customers' demographic

)

characteristics like occupation, gender, age and marital status. Psycho graphic characteristics like spending, consumption patterns, products and previously purchased products. After identifying each segment from the data set, analyze the performance of the segments to the purchases. Scope of this project is to identify the most potential customers by the results of Purchase probability, Brand choice of customers helps managers to easily communicate with a targeted group of the audience, provides opportunities for up-selling and cross selling and able to identify new products that are of interest to customers. Interpreting each cluster segments and identifying the purchase behavior helps in selecting the best medium for serving the business and increase the revenue.

2. RESEARCH METHODLOGY

Our research methodologies involves Agglomerative Clustering, K-Means Clustering and K-Means based on PCA Clustering. The various phases of customer segmentation workflow was depicted in fig 1.2 and explained in detail in the upcoming sections of the paper.

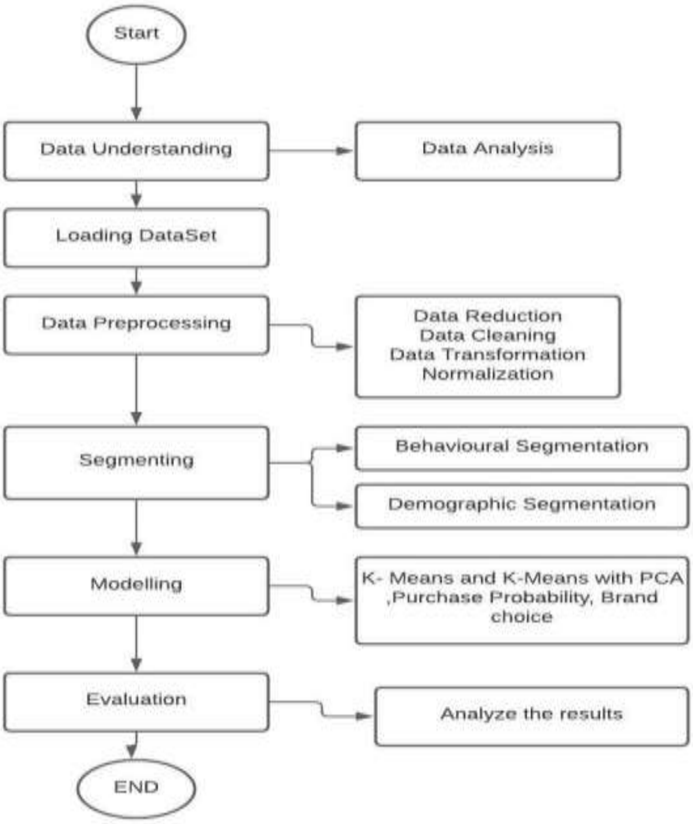

Fig 1.2 Workflow of Customer Segmentation

DATA PREPROCESSING

Individual customers have the features that characterize them. Quantify these similarities and differences. Users consider the features and transform them so that they fall within same numerical range. Able to compare the difference between the values is called as standardization.

A) HIERARICHAL CLUSTERING

Hierarchical clustering is of two types

1. Agglomerative
2. Divisive

In this model, agglomerative clustering[2] is used to compute a predetermined number "n" of clusters by first processing each point in the dataset .As a separate cluster and then on each iteration it merges the clusters together based on Euclidean distance until a predetermined number of "n" clusters is obtained. Divisive clustering is the opposite of agglomerative clustering. [1]To determine the number of clusters "n", a dendrogram for data set was calculated as shown in fig 1.3.

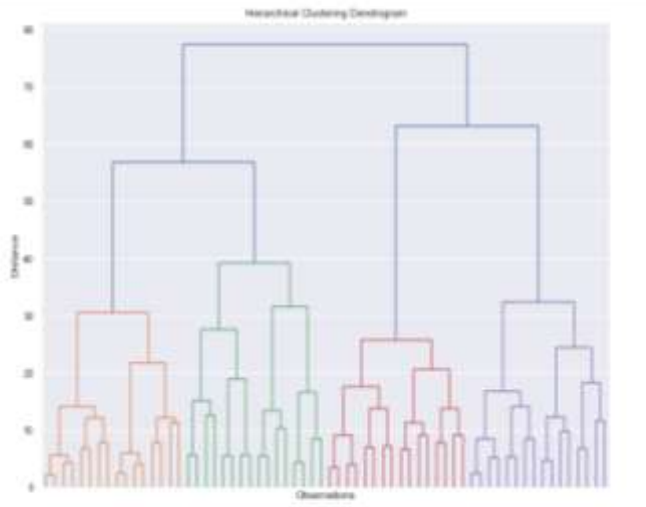

Fig 1.3 Hierarchical Clustering Dendrogram

B) K-MEANS CLUSTERING

Clustering algorithms create such groups in groups that are comparable in specific attributes. Likeness is characterized as far as how close items are in space.

K-means[3] calculation in perhaps the most well known centroid based calculation. Assume that the dataset , D, contains n objects in space. Apportioning strategies partition the items in D into k groups, C1,...,Ck , that is, Ci ⊂ D and Ci ∩j = ∅for (1 ≤i, j ≤k). Centroid-based apportioning utilizes the centroid of a cluster, Ci , to address that cluster. Adroitly, the centroid of cluster is its middle point. The difference between an item p ∈ Ci and ci , representative of the group, is estimated in dist(p,ci), where dist(x,y) is the Euclidean distance between two focuses x and y**.**

Algorithm: The k-means[4] algorithm is an iterative algorithm. The aim of k-means clustering is to find the optimal k clusters and their centers and aids to reduce the total error.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) Arbitrarily choose k objects from D as the initial cluster centers (2) repeating (3) reassigning each object to the cluster to which the object is most similar, based on the

average of the objects in the cluster (4) update the cluster . Calculate the average values of objects for each cluster (5) until unchanged.

Elbow Method:

Elbow method[1] is used in K-means to determine the similar customers falls within the same segment. The basic idea behind K-means is to obtain a minimum total within-Cluster variance. More clusters better capture the group of data objects with high similarity

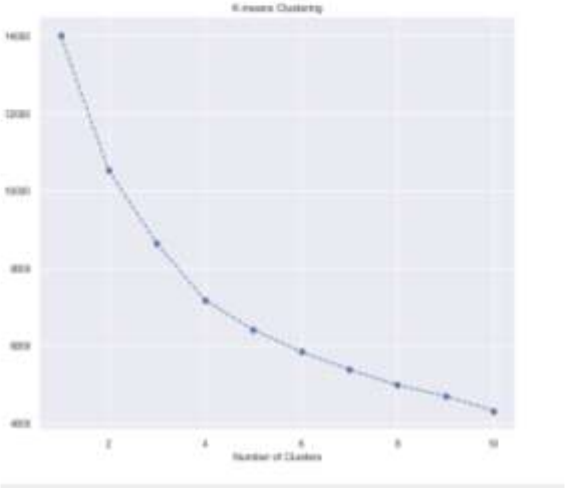

Fig 1.4 K-Means Clustering

Clustering algorithm used with different values of k ranging from 1 to 10 to obtain the optimal clusters as shown in fig 1.4. Total intra-cluster sum of square was calculated. Plot the sum of the squares between the clusters against the number of clusters. Optimal clusters can be pointed in the graph indicated with a bend in plot. Customer's falls under various segments was depicted in fig 1.5

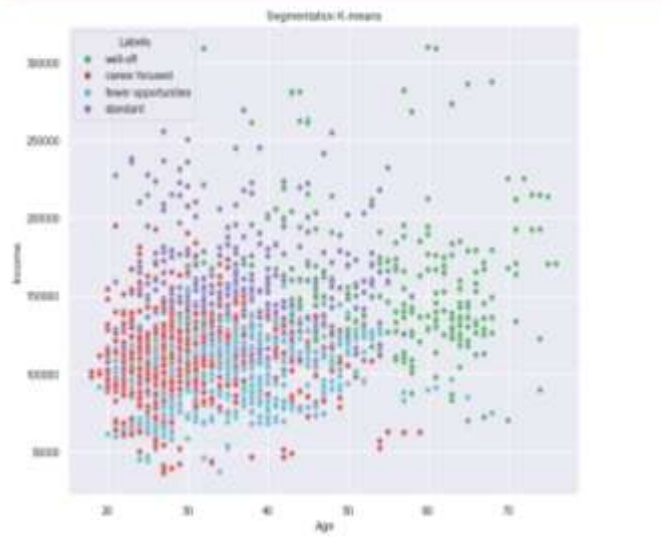

Fig 1.5 K-Means Segmentation

C) K-MEANS BASED ON PCA CLUSTERING

Using Principle Component Analysis, Linearly reduce the components of variance would reduce the features significantly while preserving most of the information. PCA class has a built in method that transforms the data in the desired way. The result of this in new array where each observation is described by the components will be considered as PCA scores. Now segmenting again with K-Means clustering for these PCA scores as shown in fig 1.6.

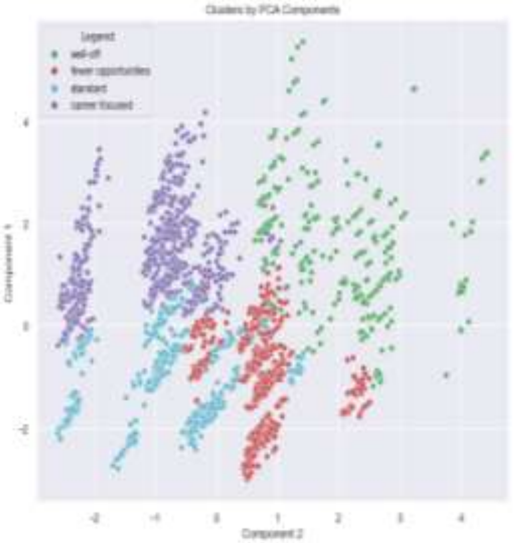

Fig 1.6 K-Means Clustering with PCA

D) PURCHASE ANALYTICS

Purchase Analytics is the main approach of the Positioning. In identifying the purchase behaviour, three particular questions arises, one will the customer buy a product from a particular product category when they enter the shop to which brand is the customer is going to choose and another will be how many units is the customer going to purchase in this part.

For these purchase probability, Brand choice probability used the linear regression and logistic regression to interpret their ability.

3. TECHNOLOGY USED

Language: Python

Packages: Pandas, Scipy, Mat-plot, sklearn, sea-born, pickle

Front End Requirements: Flask

Flask is a web framework that provides libraries to build lightweight web applications in python.

It is based on WSGI toolkit and jinja2 template engine. Flask is considered as a micro framework. WSGI is an acronym for web server gateway interface which is a standard for python web application development. It is considered as the specification for the universal interface between the web server and web application.

Jinja2 is a web template engine which combines a template with a certain data source to render the dynamic web pages.

IDE : Jupyter Notebook

Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document

4. RESULTS

Hierarchical clustering is not efficient because its computational complexity **is** O **(n ^ 3),** which means that it cannot be used on the large datasets.. But K-Means has computational complexity of O(nkd) in which 'n' is the number of points, 'K' is the clusters count, and 'd' is the count of characteristics, making it fastest among the other algorithms. As shown in fig 1.7 segmented the data set into four groups and classified them as well-off, carrier focus, standard, and fewer opportunities. Identified each segment purchase behaviour to the purchase data.

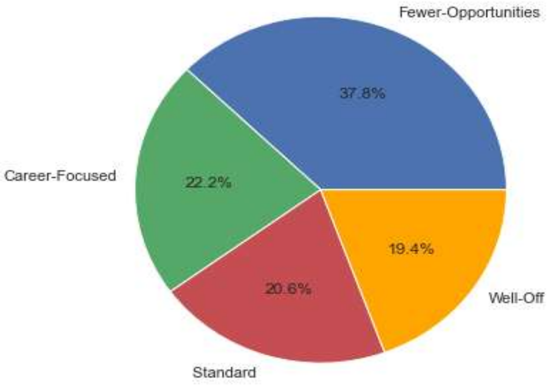

Fig 1.7 Segment proportions of store visits by each segments

Interpret the Purchase Probability, Brand choice Probability and Purchase quantity using Linear and Logistic Regression as shown in fig 1.8

Out[[illegible]]:

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Component 1	Component 2	Component 3	N Obs	Prop Obs
Segment K-means PCA												
standard	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
career focused	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
fewer opportunities	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
well-off	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

```
In [[illegible]]: df_segm_pca_kmeans['Legend'] = df_segm_pca_kmeans['Segment K-means PCA'].map({0:'standard',
                                                                  1:'career focused',
                                                                  2:'fewer opportunities',
                                                                  3:'well-off'})
```

Fig 1.8 K-Means Segmentation results with PCA

5. CONCLUSION

In this paper, customer purchase behaviour are analyzed systematically based on the customer data and the purchase data of a mall using K-Means, K-Means based on PCA, and find the probability using Linear and Logistic Regression. Segmentation and the probability will help the managers to launch offers for targeted customers, to encourage them to buy more products and also helps in selecting the best medium for communicating with the targeted segment.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my faculty guide Mrs N Rajeswari Assistant Professor, Department of Computer Applications for her vision and tireless support to me in submitting this research paper. I would also like to thank Dr A Chitra, Professor and Head, Department of Computer Applications, and all faculties of the Department of Computer Applications, PSG college of Technology, for their encouragement and intuitive comments.

6. REFERENCES

[1] Musadig Aliyev,Elien Ahmadov, Habil Gadirli, Arzu Mammadova and Emin lasgarov,“ Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning”, August 2020

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

[3] Yash Kushwaha, Deepak Prajapati,“Customer Segmentation Using K-Means Algorithm”, International Journal of Creative Research Thoughts ,September 2008

[4] T Kanungo, D M Mount, N S Netanyahu, C.D. Piatko, R Silverman and A Y Wu, “An Efficient K-Means Clustering Algorithm,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 24, Issue 7, July 2002

[5] Bhattacharjee, S. and Raheja, S. “Using Marketing Analytics to Understand Consumer Lifestyle for Hair Salons in Delhi and Kolkata”, IARS’ International Research Journal. Vic. Australia, 10(2) 2020. doi: 10.51611/iars.irj.v10i2.2020.133.

[6] Rosette, P. O. “Effect of Global Recession on Indian Realty Sector and Its Future Developments”, IARS’ International Research Journal. Vic. Australia, 5(2) 2015. doi: 10.51611/iars.irj.v5i2.2015.49.