Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation

Dr. Huma Lone[1], Dr. Prajakta Warale[2]

[1]Freelance trainer, Pune, India
[2] Associate Professor, Rajgad Institute of Management Research and Development, Dhankawadi, Pune-43, India
[1]humalone@gmail.com,[2]prajaktawarale@gmail.com

Abstract

Customer segmentation is the division of a business customersinto categories called customer segments such that each customer segment exhibits similar characteristics. This division of customers is built on factors that can directly or indirectly affect the market or business such as product preferences or expectations, locations, behaviours etc. Customer segmentation can be implemented through clustering, which is one of the highlyrecognized machine learning techniques. Cluster analysis is applied in many business applications, from customized marketing to industry analysis. It is an unsupervised learning technique that divides a dataset into a set of meaningful sub-classes, called clusters. It helps to comprehend the natural grouping in a dataset andcreate clusters of similar records which depends on several measurements made in the form of attributes.

This research paper has focused on creating customers clusters by applying K-means and Agglomerative clustering algorithms on a dataset consisting of 200 customers. Various machine learning libraries were used in Python programming language to implement and visualize the results.

Keywords— Agglomerative clustering, Elbow method, Dendrogram, Clustering, Customer Segmentation, K-means Clustering

INTRODUCTION

Guided by the fact that customers are usually distinct, and this distinctiveness is characterised by their behaviour, customer segmentation is an essential aspect to be explored by organizations. It is the process of separating an organization’s customer bases into distinct clusters or groups based on several customer dimensions or features. These dimensions or features can be their geographic info, their buying behaviour, their demographic information, their psychographic attributes etc. [1]. There are several reasons to go for customer segmentation [2], [3]. These reasons are elaborated below:

- C**ustomer Understanding:** For any business to be successful and profitable, the first and foremost task is to know their customers. It is important for an organisation to understand its customers and know their needs and demands.
- **Target Marketing:** The capability to focus on marketing efforts well is the most compelling factor for customer segmentation. If a company recognizes the distinct groups of its customer then it can propose better marketing campaigns which can be intended differently for different segment.
- **Optimal Product Placemen**t: A well designed and useful customer segmentation approach can also help business companies with developing or offering new products.
- **Higher Revenue:** This is the surest and apparent requirement of any customer segmentation process. Customer segmentation can have an advantage to reveal better revenue due to the collective

effects of all the above mentionedadvantages.

Customer segmentation can be implemented through one of the core Machine Learning tasks. The most obvious and widely used task to perform customer segmentation is clustering which is an unsupervised machine learning task that considers input vectors or features without referring to known, or labelled outcomes. Clustering categorizes objects into various strata built on similarity or distance measure [4], [5]. The main objective is to distinguish the clusters in ways that would be useful for giving useful insights. It is a procedure of segregating a set of instances into subsets. Each subset is a cluster. Data points present in a cluster are comparable to each other and contradictory to data points present in other clusters. There are many clustering methods in the literature. Partitioning and Hieratical are two such methods[1], [6], [7].

PARTITIONING CLUSTERING METHOD

It constructs *k* partitions of the data, where each separation represents a cluster and $k \leq n$ where n represents the number of data points. Most partitioning methods are distance-based. This method generates an initial partitioning for given *k* and then uses an iterative relocation method that tries to improve the partitioning by changing data points from one cluster to another. K-means clustering is one of the partitioning algorithms based on centroid calculation [6], [7].k-means algorithm is the most known and widely used partitioning methods. K-means algorithm takes the input parameter, k, and partitions a set of n data points into k clusters or groupssuch that the inter-cluster similarity is low and intra cluster similarity is high. Cluster similarity is measured to the mean value of the objects in a cluster, which can be viewed as the cluster’s centroid or center of gravity [1], [8],[9], [10]. Figure 1 describes the steps involved in K-means algorithm.

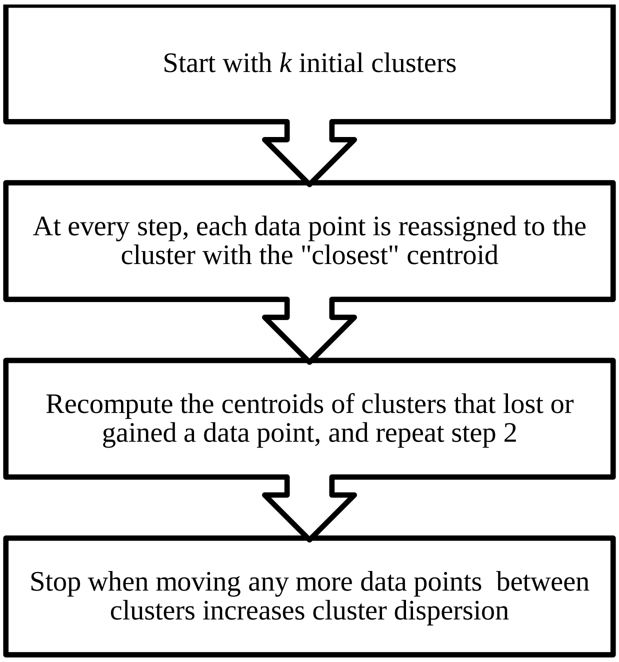
Start with *k* initial clusters
At every step, each data point is reassigned to the cluster with the "closest" centroid
Recompute the centroids of clusters that lost or gained a data point, and repeat step 2
Stop when moving any more data points between clusters increases cluster dispersion

Figure 1: K-Means Clustering Algorithm

HIERARCHICAL CLUSTERING METHOD

It operates by grouping data points into a hierarchy or “tree” of clusters which results in summarization and visualization of data points. Hierarchical clustering method may be an agglomerative method or divisive method, differing on whether the breakdown is formed in a bottom-up (merging) or top-down (splitting) manner [7].

The agglomerative method is also called the bottom-up approach, begins with each data points making a separate group. It merges the data instances in succession close to one another, until all the groups become one group (the topmost level of the hierarchy), or certain termination conditions are fulfilled. A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering. It shows how data instances are grouped together into one. Data points are displayed at the bottom and similar data points are joined by lines whose vertical length reflects the distance between the data points[6], [7]. Figure 2 describes the steps involved in agglomerative clustering algorithm.

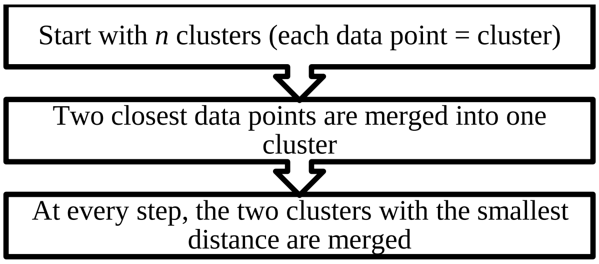
Start with n clusters (each data point = cluster)
Two closest data points are merged into one cluster
At every step, the two clusters with the smallest distance are merged

Figure 2: Agglomerative Clustering Algorithm

Many people are of the opinion that clustering,and customer segmentation can be used interchangeably. Even though, it is true that clustering is one of the highly appropriate and widely used methods for segmentation, it is not the only method. It is observed that a clustering-based segmentation will be better than other segmentation methods. This method involves collecting data about the customers in the form of attributes or featuresand then discovering different clusters from that data. Finally, label these clusters by analysing the characteristics of the clusters [11].

DATA COLLECTION AND RESEARCH METHODOLOGY

The data set named “Mall_customers.csv” was taken from Kaggle.com. Kaggle is an online platform of data scientists and machine learning experts[12]. It permits users to collect and publish datasets, explore, and construct models in a web-based data-science environment. The dataset contained basic information about customers of a supermarket mall like customer ID, age, gender, annual income, and spending score. There were five features (CustomerId, Gender, Age, Annual Income and Spending Score) with 200 customers. Two different clustering algorithms: K-means and Agglomerative clustering had been used for customer segmentation[7]. Since more than 2-dimensional data is difficult to visualize, only last two features (Annual Income and Spending Score) were considered as input to the two clustering algorithms. Python, is the most preferred and largely used programming language for machine learning applications had been used as a data analysis tool[11]. Execution of the code for both the clustering algorithms had been done in Jupiter Lab, which is an IDE (Interactive Development Environment) for Python.

PYTHON AS DATA ANALYSIS TOOL

Widely used open-source programming languages such as Python, Rand tools such as, Tableau, XLMiner, RapidMiner, MATLAB, SAS, Stata etc. are extensively used for data analysis, interactive computing, and data visualization. Python is a general-purpose programming language which is designed to be simple and easy to understand. Artificial learning, machine learning and deep learning applications are developed easily with Python’s libraries such as numpy, pandas, matplotlib and scikit-learn. These support libraries have made python as an incredible option as a primary language for data science application [11]. Figure 3 shows the essential python libraries for Machine Learning Applications [14].

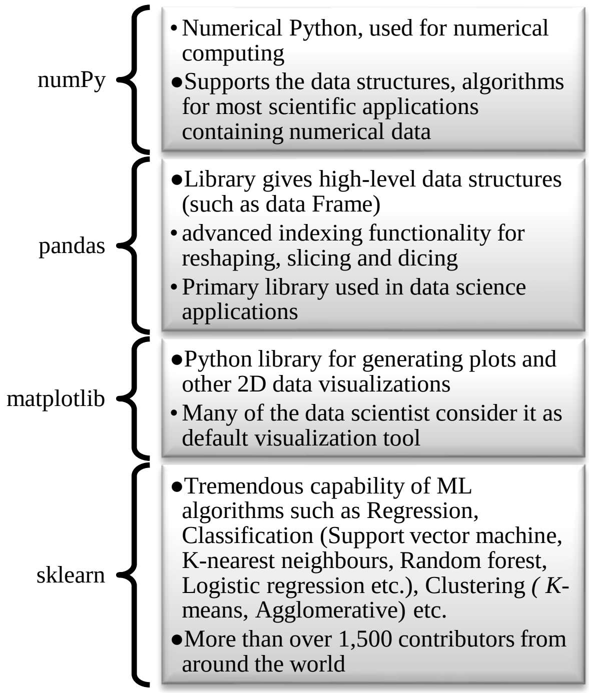
numPy
• Numerical Python, used for numerical computing
●Supports the data structures, algorithms for most scientific applications containing numerical data
pandas
●Library gives high-level data structures (such as data Frame)
• advanced indexing functionality for reshaping, slicing and dicing
• Primary library used in data science applications
matplotlib
●Python library for generating plots and other 2D data visualizations
• Many of the data scientist consider it as default visualization tool
sklearn
●Tremendous capability of ML algorithms such as Regression, Classification (Support vector machine, K-nearest neighbours, Random forest, Logistic regression etc.), Clustering (K-means, Agglomerative) etc.
●More than over 1,500 contributors from around the world

Figure 3: Essential Python Libraries for Machine Learning Applications

DATA ANALYSIS AND INTERPRETATION

Figure 4 demonstrates the process for implementing K-means and Agglomerative clustering on the selected dataset in python. Import libraries, import dataset, finding the optimal number of clusters, training the model on dataset and cluster visualisations are the five essential steps for applying clustering algorithms for customer segmentation.

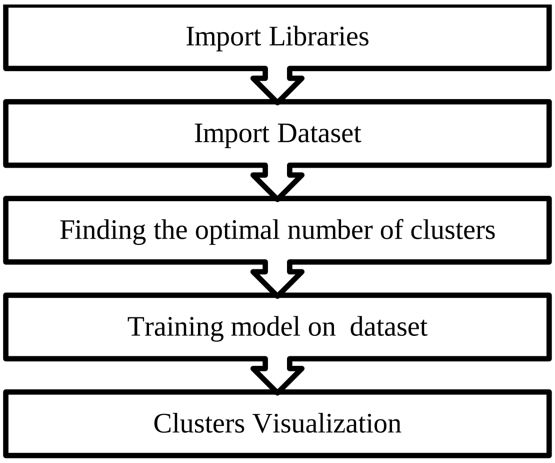
Import Libraries
Import Dataset
Finding the optimal number of clusters
Training model on dataset
Clusters Visualization

Figure 4: Process for implementing K-means and Agglomerative clustering on dataset

IMPORT LIBRARIES AND DATASET

Numpy, sciPy, matplotlib, pandas and scikit learn libraries were used for cluster analysis. The dataset named Mall_customers.csvwas imported in Jupiter notebook using read.csv function. For each step mentioned in figure 4, the details of the libraries, their modules along with the function, class and attribute names are illustrated in Figure 5 and 6.

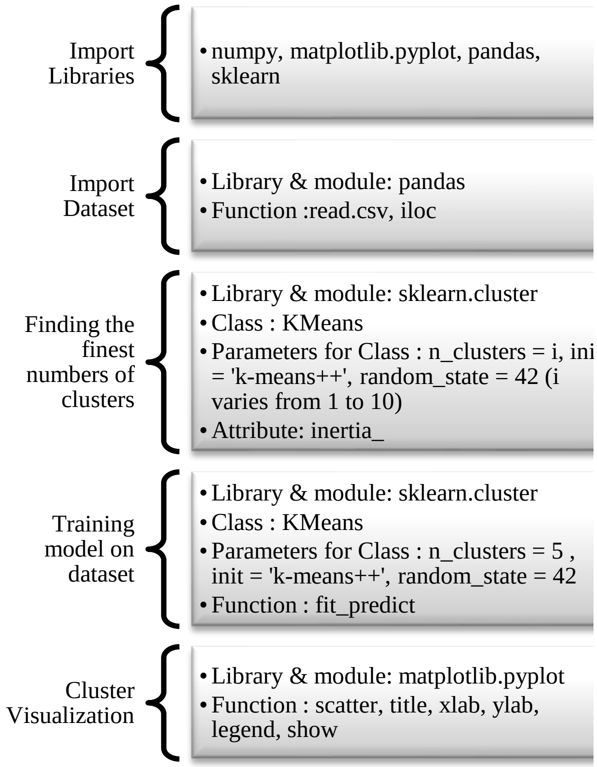
Import Libraries
• numpy, matplotlib.pyplot, pandas, sklearn
Import Dataset
• Library & module: pandas
• Function :read.csv, iloc
Finding the finest numbers of clusters
• Library & module: sklearn.cluster
• Class : KMeans
• Parameters for Class : n_clusters = i, ini = 'k-means++', random_state = 42 (i varies from 1 to 10)
• Attribute: inertia_
Training model on dataset
• Library & module: sklearn.cluster
• Class : KMeans
• Parameters for Class : n_clusters = 5 , init = 'k-means++', random_state = 42
• Function : fit_predict
Cluster Visualization
• Library & module: matplotlib.pyplot
• Function : scatter, title, xlab, ylab, legend, show

Figure 5: Python libraries for implementing K-means clustering

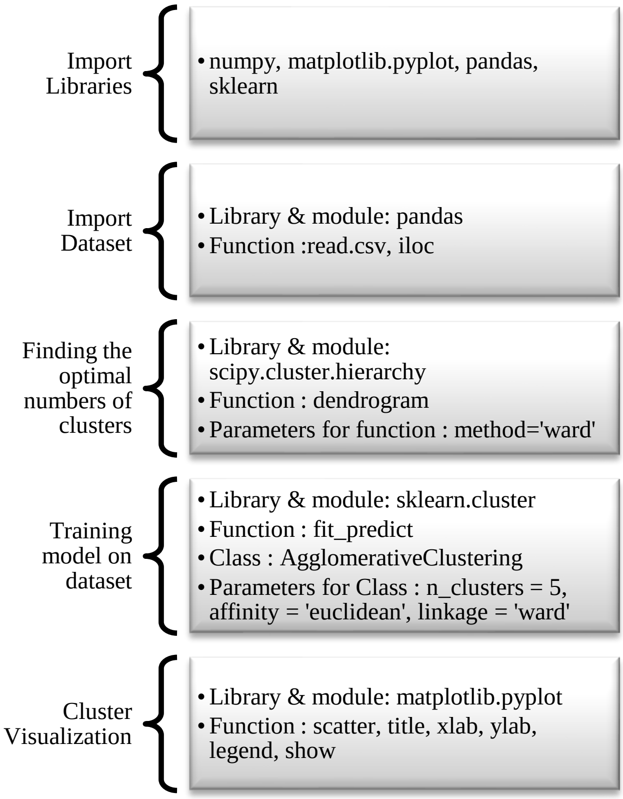
Import Libraries
• numpy, matplotlib.pyplot, pandas, sklearn
Import Dataset
• Library & module: pandas
• Function :read.csv, iloc
Finding the optimal numbers of clusters
• Library & module: scipy.cluster.hierarchy
• Function : dendrogram
• Parameters for function : method='ward'
Training model on dataset
• Library & module: sklearn.cluster
• Function : fit_predict
• Class : AgglomerativeClustering
• Parameters for Class : n_clusters = 5, affinity = 'euclidean', linkage = 'ward'
Cluster Visualization
• Library & module: matplotlib.pyplot
• Function : scatter, title, xlab, ylab, legend, show

Figure 6: Python libraries for implementing Agglomerative clustering

DECIDING OPTIMAL NUMBERS OF CLUSTERS

One of the challenges in both the algorithms was to decide the number of clusters before implementing these algorithms. To find the best possible number of clusters in K-means Clustering,Elbow method was used [15],[16]and dendrogram was used in Agglomerative clustering.

The basic intention behind k-means clustering, is to define clusters such that total within-cluster sum of square (WCSS) is minimized. The total WCSS calculates the compactness of the clustering. The Elbow method considers the total WCSS that depends on the number of clusters and can be used to find the optimal number of clusters [15],[16]. A curve of WSS for different values of K was plotted using the selected dataset. The plotted curve looked like an elbow as shown in the figure 7. The location of a bend in the plot is normallycounted as a marker of the optimal numbers of clusters where adding another cluster does not further enhance the total WSS much. From the curve, the optimal number of clusters was chosen as five.

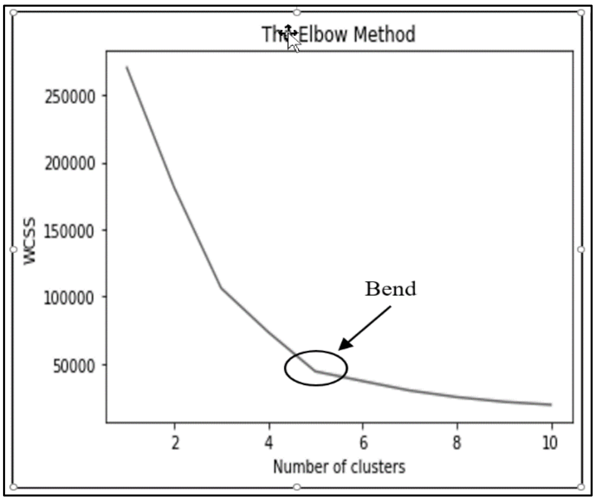

Figure 7: Elbow method for finding the optimal number of clusters for K-means clustering

Dendrogram is a treelike figure that summarizes the procedure of clustering. Data points are shown at the bottom. Similar data points are joined by lines whose vertical length reflects the distance between the data points[6].If the dendrogram tree is cut with a horizontal line at a height where the line can traverse the maximum distance up and down without intersecting the merging point,appropriate number of clusters can be found [17]as shown in figure 8. From the dendrogram, the appropriate or optimal number of clusters was chosen as five, similar in elbow method.

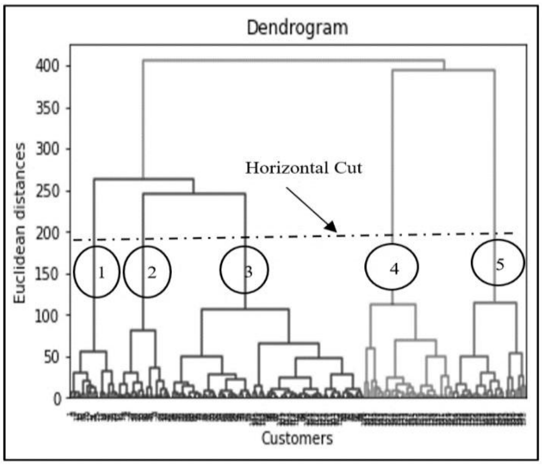

Figure 8: Dendrogram for deciding the number of clusters for agglomerative algorithm

TRAINING MODEL ON DATASET AND CLUSTER VISULIZATION

Considering the best number of clusters as 5, the model was trained on selected dataset by using sklearn library. Matplotlib library was used for visualization. Figure 9 and10 clearly shows five clusters as an output of K-means and Agglomerative clustering algorithm based on two features (Annual Income and Spending Score). Each small circle in figure 9 and 10 represents a customer in the dataset. The five bigger circles in figure 9 are the centroids of each cluster. Despite having their own advantages and disadvantages the output of both the algorithms were similar for selected dataset. Hence both the methods can be appropriately applied on the selected dataset.

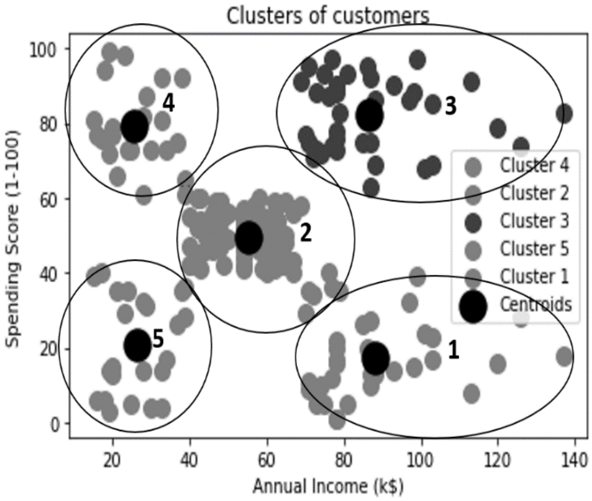

Figure 9: Cluster Analysis by K-means Algorithm

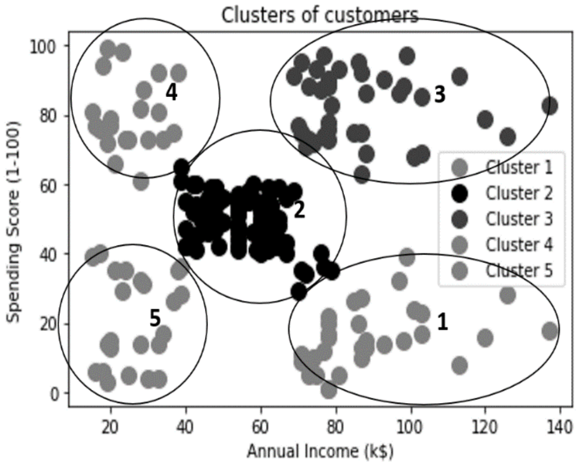

Figure 10: Cluster Analysis by Agglomerative Algorithm

This Clustering Analysis provided with a clear insight about the different segments of the customers in the Mall. These five segments of Customers were named as Saver, Vigilant, Target, Extravagant and Impecunious based on their spending score and annual income. Figure 11 shows the cluster characteristics of five clusters.

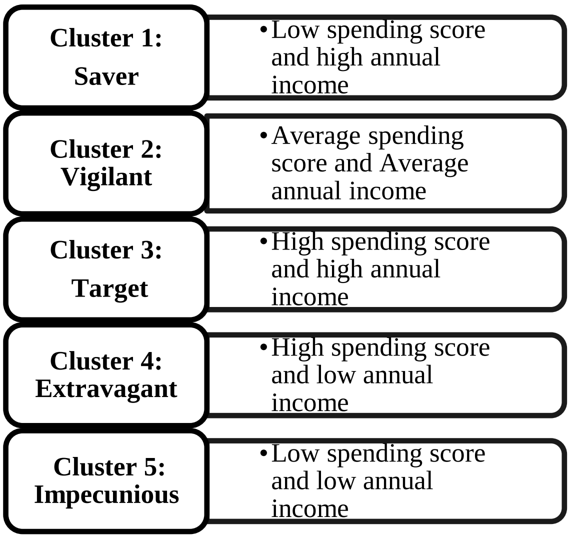
Cluster 1: Saver
•Low spending score and high annual income
Cluster 2: Vigilant
•Average spending score and Average annual income
Cluster 3: Target
•High spending score and high annual income
Cluster 4: Extravagant
•High spending score and low annual income
Cluster 5: Impecunious
•Low spending score and low annual income

Figure 11: Five Clusters with their characteristics

Saver (Cluster 1): The customers in this category have low spending score and high annual income. Despite having high income, this cluster of customers spend less. One of the reasons could be that these customers are unsatisfied or disappointed with the mall's services. The Mall marketing team should focus on this cluster as customers in this cluster have potential to spend.

Vigilant (Cluster 2): The customers in this category have average spending score and average annual income. This cluster of customers may not be the prime target of the mall, but marketing team should target their marketing efforts to retain them to increase their spending score.

Target (Cluster 3): The customers in this category have high spending score and high annual income. These customers can be considered as prime source of mall's profit. These customers are satisfied from mall's service and are regular customers. The marketing team should target them with attractive offers to gain more profits.

Extravagant (Custer4): The customers in this category have high spending score and low annual income. Despite having low income, they tend to buy products. This could be because these customers are satisfied with mall's service. The marketing team may not want to focus on these customers as they have low income, so beyond certain limit they cannot be attracted to offers from the mall. The focus could be to that extent that the mall should not lose them.

Impecunious (Cluster 5): The customers in this category have low income and hence they prefer to spend less. The Mall marketing team will not be interesting to focus on this cluster of customers

DISCUSSION AND CONCLUSION

Understanding a business's customer base is extremely important for any business organisation. Customer segmentation is one of the ways to gain deeper understanding of customer behaviour. It is one of the important applications of cluster analysis amongst many applications spread across different domains.Sales and marketing efforts can be well designed for these clusters of customers to achieve high return on investment. Unsupervised machine learning algorithms such as K-means and Agglomerative clustering algorithms can be easily applied using python support libraries to summarize and visualize the clusters. The current research applied K-means and Agglomerative clustering algorithms on Mall_Customers dataset and discovered different clusters from that data. Finally, these clusters were labelledas Saver, Vigilant, Target, Extravagant and Impecunious by analysing the characteristics of the clusters.These clusters can help marketing team of the Mall to focus on these segments of customers differently and achieve maximum profit.

REFERENCES

1. H. H. Ali and L. E. Kadhum, 'K-Means Clustering Algorithm Applications in Data Mining and Pattern Recognition', *Int. J. Sci. Res.*, vol. 6, no. 8, pp. 1577–1584, 2017.
2. T. A. Nguyen, 'Customer Segmentation: A Step By Step Guide For B2B', 2018. https://openviewpartners.com/blog/cust

omer-segmentation/#.X-ay_NgzY2w (accessed Oct. 02, 2020).

3. ‘10 Reasons for Customer Segmentation in Customer Service’, 2019. https://helprace.com/blog/10-reasons-for-customer-segmentation (accessed Oct. 05, 2020).
4. S. Mukhopadhyay, *Advanced Data Analytics Using Python: With Machine Learning, Deep Learning and NLP Examples*. Apress, 2018.
5. S. Kumar, ‘Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization’, *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 10, pp. 3161–3166, 2014.
6. G. Shmueli, N. R. Patel, and P. C. Bruce, *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley and Sons, 2011.
7. J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
8. R. Raval Unnati and J. Chaita, ‘Implementing & Improvisation of K-means Clustering Algorithm’, *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, pp. 191–203, 2016.
9. Y. Li and H. Wu, ‘A clustering method based on K-means algorithm’, *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012.
10. S. Singh and N. S. Gill, ‘Analysis and study of k-means clustering algorithm’, *Int. J. Eng. Res. Technol.*, vol. 2, no. 7, 2013.
11. D. Sarkar, R. Bali, and T. Sharma, ‘Practical machine learning with Python’, *A Probl. Guid. to Build. real-world Intell. Syst. Apress, Berkely*, 2018.
12. ‘Mall_Customers’. https://www.kaggle.com/akram24/mall-customers (accessed Sep. 12, 2020).
13. A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. ‘ O’Reilly Media, Inc.’, 2016.
14. M. Wes, *Python for data analysis*. O’Reilly Media, Inc., 2012.
15. ‘CLUSTER VALIDATION ESSENTIALS’. https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#ptimal-number-of-clusters-3-must-know-methods/ (accessed Sep. 25, 2020).
16. T. M. Kodinariya and P. R. Makwana, ‘Review on determining number of Cluster in K-Means Clustering’, *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.
17. ‘An Introduction to Clustering and different methods of clustering’. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ (accessed Sep. 27, 2020).