**INTERNSHIP PROJECT REPORT**

On

**“Customer Segmentation Using Data Science”**

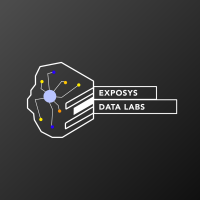
Submitted By

Mokshada Rajendra Godase

Shubham Ramesh Yadav

Duration: 02/09/2020 To 02/10/2020

Submitted To



Exposys Data Labs

ABSTRACT

The emergence of many business competitors has engendered severe rivalries among competing businesses in gaining new customers and retaining old ones. Due to the preceding, the need for exceptional customer services becomes pertinent, notwithstanding the size of the business. Furthermore, the ability of any business to understand each of its customers’ needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation. Each segment comprises customers who share similar market characteristics. The ideas of Big data and machine learning have fuelled a terrific adoption of an automated approach to customer segmentation in preference to traditional market analyses that are often inefficient especially when the number of customers is too large. In this project, the k-Means clustering algorithm is applied for this purpose . A python program has been developed and the program is been trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. The features are the average amount of goods purchased by customer per month and the average number of customer visits per month. From the dataset, four customer clusters or segments were identified with 95% accuracy, and they were labeled: High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular-Visitors (LBIV) .

TABLE OF CONTENTS

Chapter 1. Introduction

Chapter 2. Existing Method

Chapter 3. Proposed method with Architecture

Chapter 4. Methodology

Chapter 5. Implementation

Chapter 6. Conclusion and References

1. INTRODUCTION

As more and more business being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat. Change or die have become the simple rule of marketing in today’s world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company’s database. Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualise what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them. Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabelled dataset. There are a number of clustering algorithm over which like k-means, hierarchical clustering, DBSCAN clustering etc. In this paper, three different clustering algorithms have been implemented over a dataset with two features with 200 records.

2. EXISTING METHOD

Jayant et al states that in customer segmentation, the customers are divided into different groups where customers of the same group are similar to each other in terms of marketing. Customers are divided into different clusters based on various attributes such as age ,interests, age, spending habits etc.

Sulekha et al provides the four popular bases for segmentation

1) Geographic Segmentation: segmentation on the basis geographic region, population density or climate.

2) Demographic Segmentation: market segment on the basis of age, size and family type, etc.

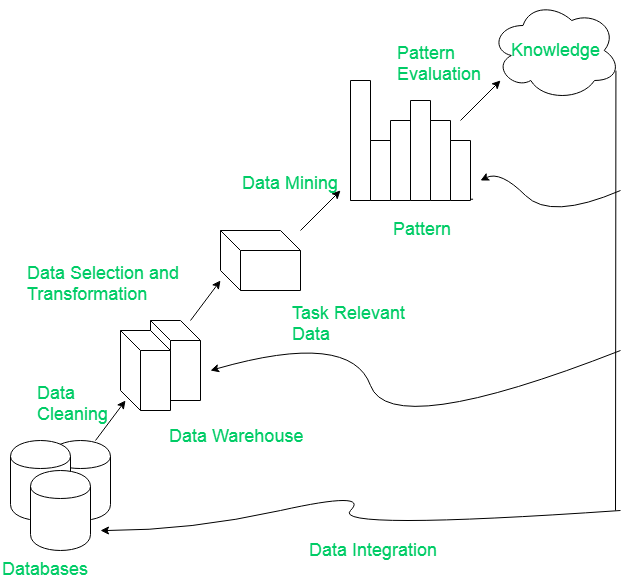
3) Psychographic Segmentation: segmentation based on customer’s life style variables like interests, opinions, attitudes etc.

4) Behavioural Segmentation: segmentation is based on actual customer behaviourtowards products like brand loyalty, user status, readiness to buy etc. Customer segmentation is based on based on the strategy called divide and conquer by utilising the advantage of segmentation the marketers can gain advantage over a particular segment and slowly can prevail over other marketers. Using market segmentation the marketers can focus more on customer relationship management which was not earlier possible with existing mass marketing tactics.

3. PROPOSED METHOD WITH ARCHITECTURE

Workflow :

The basic workflow for customer segmentation consists of only 3 steps: Data Cleaning, Data selection or transformation, Data Mining (k-Means clustering) and Pattern Evaluation.



Programming Language :

The programming language in question was going to be python . The language is very suitable for data mining, and with libraries explained below it has become the most popular programming language for Data Science .

These are the libraries used in the project :

1. Scikit-learn(0.22.1v) : Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN.
2. Seaborn (0.10.1v) : Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
3. Numpy(1.19.1v) : Numerical Python, a library that adds support for large multidimensional matrices and a collection of high-level mathematical functions.
4. Pandas( 1.1.1v) : Builds upon NumPy and provides high-performance table-like data structures well suited for handling and manipulating large sets of data.
5. Matplotlib (3.3.1v) : An extensive plotting library that allows for easy visualization of any kind of data analysis.

DATASET :

Dataset in this project used are gender , age , annual income and spending score 200 training samples have been used .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
| 1 | Male | 19 | 15 | 39 |
| 2 | Male | 21 | 15 | 81 |
| 3 | Female | 20 | 16 | 6 |
| 4 | Female | 23 | 16 | 77 |
| 5 | Female | 31 | 17 | 40 |
| 6 | Female | 22 | 17 | 76 |
| 7 | Female | 35 | 18 | 6 |
| 8 | Female | 23 | 18 | 94 |
| 9 | Male | 64 | 19 | 3 |
| 10 | Female | 30 | 19 | 72 |

4. METHODOLOGY

**K-means Clustering**: It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre’s are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.

**Elbow Method**: Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

Methodology:

Step 1 : Data Collection

The dataset has been taken from a local retail shop consisting of features such as , Gender , age , Annual Income and spending score.

Step 2 : Data Cleaning

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_(database)), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the data.

The Customer Id Column has been dropped as it does not seem relevant to the context .

Step 3 : Data Selection

The attributes that were selected are Gender , Age , Annual Income and Spending Score .

Step 4 : Data Mining

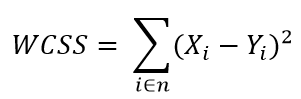
K means Clustering:

Choosing the optimal number of clusters:

The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).

WCSS is calculated by :



Where *Yi* is centroid for observation *Xi*

1. Plot the curve of WCSS vs the number of clusters K.
2. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Step 5 : Pattern Evaluation

Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.

Find interestingness score of each pattern.

Uses summarization and Visualization to make data understandable by user.

5. IMPLEMENTATION

Step 1 : Data Collection

import numpy as np

import pandas as pd

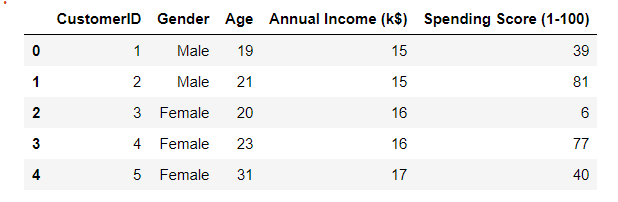
import matplotlib.pyplot as plt

import seaborn as sns

import os

df = pd.read\_csv("Mall\_Customers.csv")

df.head()



Step 2 : Data Cleaning

df.drop(["CustomerID"], axis = 1, inplace=True)

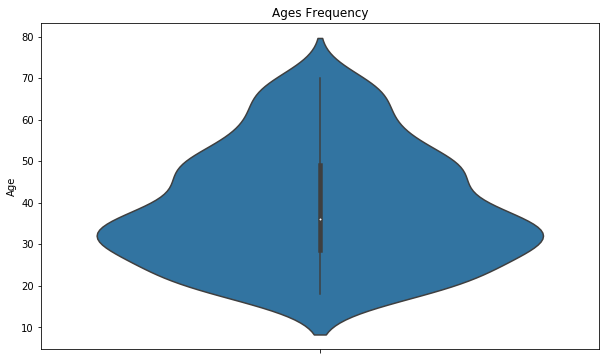
plt.figure(figsize=(10,6))

plt.title("Ages Frequency")

sns.axes\_style("dark")

sns.violinplot(y=df["Age"])

plt.show()



Step 3 : Data Selection

plt.figure(figsize=(15,6))

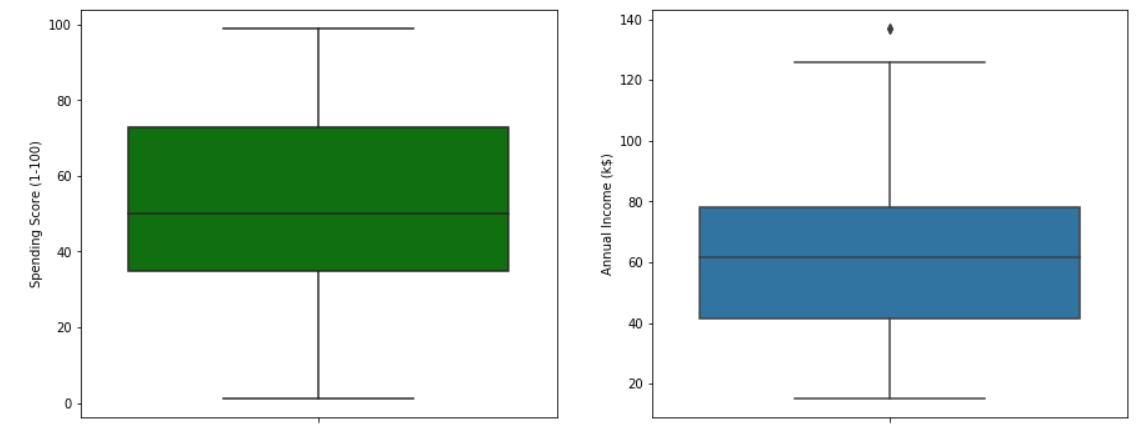
plt.subplot(1,2,1)

sns.boxplot(y=df["Spending Score (1-100)"], color="green")

plt.subplot(1,2,2)

sns.boxplot(y=df["Annual Income (k$)"])

plt.show()



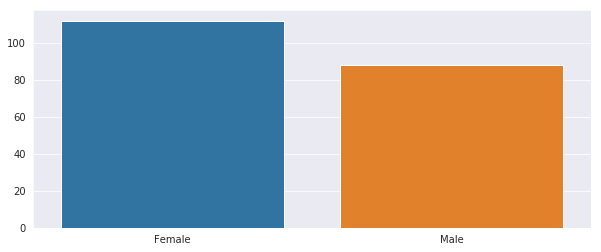
genders = df.Gender.value\_counts()

sns.set\_style("darkgrid")

plt.figure(figsize=(10,4))

sns.barplot(x=genders.index, y=genders.values)

plt.show()

****

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=(15,6))

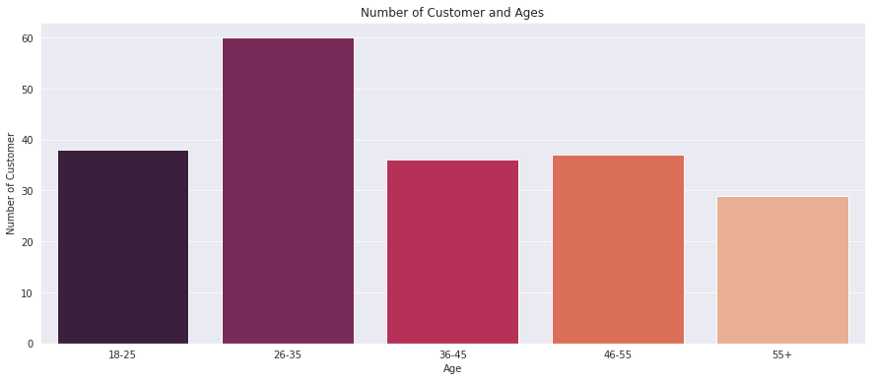
sns.barplot(x=x, y=y, palette="rocket")

plt.title("Number of Customer and Ages")

plt.xlabel("Age")

plt.ylabel("Number of Customer")

plt.show()

****

ss1\_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 20)]

ss21\_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) & (df["Spending Score (1-100)"] <= 40)]

ss41\_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) & (df["Spending Score (1-100)"] <= 60)]

ss61\_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 80)]

ss81\_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) & (df["Spending Score (1-100)"] <= 100)]

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=(15,6))

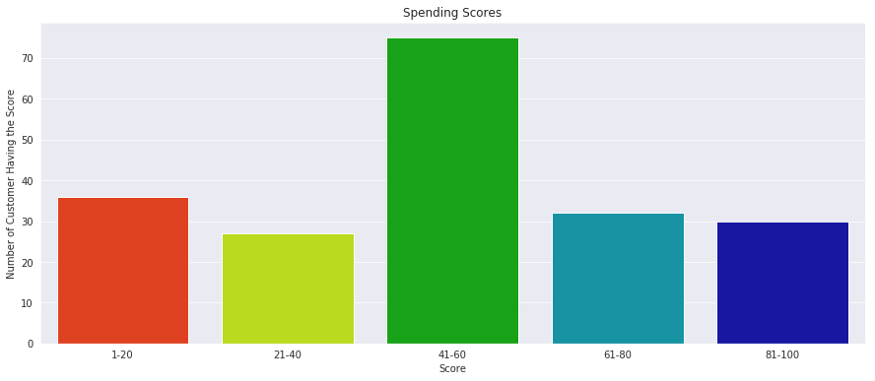
sns.barplot(x=ssx, y=ssy, palette="nipy\_spectral\_r")

plt.title("Spending Scores")

plt.xlabel("Score")

plt.ylabel("Number of Customer Having the Score")

plt.show()

****

ai0\_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 0) & (df["Annual Income (k$)"] <= 30)]

ai31\_60 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 31) & (df["Annual Income (k$)"] <= 60)]

ai61\_90 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 61) & (df["Annual Income (k$)"] <= 90)]

ai91\_120 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 91) & (df["Annual Income (k$)"] <= 120)]

ai121\_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 121) & (df["Annual Income (k$)"] <= 150)]

aix = ["$ 0 - 30,000", "$ 30,001 - 60,000", "$ 60,001 - 90,000", "$ 90,001 - 120,000", "$ 120,001 - 150,000"]

aiy = [len(ai0\_30.values), len(ai31\_60.values), len(ai61\_90.values), len(ai91\_120.values), len(ai121\_150.values)]

plt.figure(figsize=(15,6))

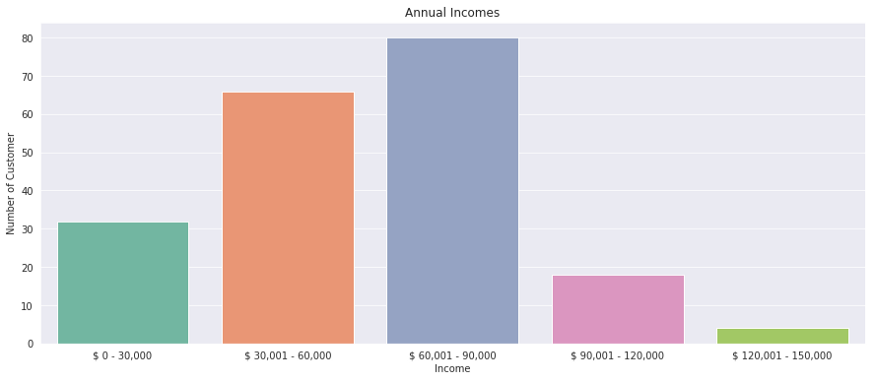
sns.barplot(x=aix, y=aiy, palette="Set2")

plt.title("Annual Incomes")

plt.xlabel("Income")

plt.ylabel("Number of Customer")

plt.show()

****

Step 4 : Data Mining

from sklearn.cluster import KMeans

wcss = []

for k in range(1,11):

kmeans = KMeans(n\_clusters=k, init="k-means++")

kmeans.fit(df.iloc[:,1:])

wcss.append(kmeans.inertia\_)

plt.figure(figsize=(12,6))

plt.grid()

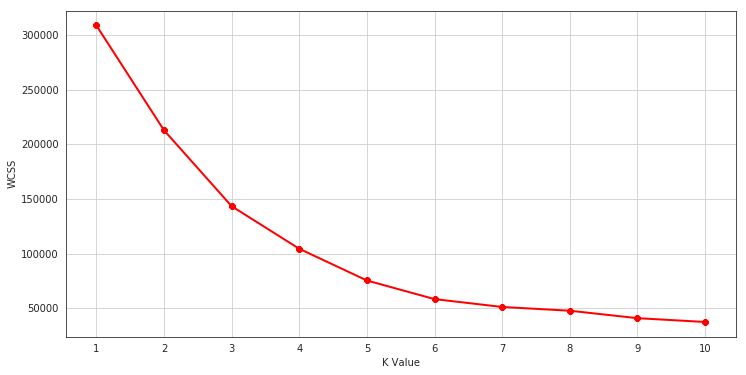
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")

plt.xlabel("K Value")

plt.xticks(np.arange(1,11,1))

plt.ylabel("WCSS")

plt.show()

****

km = KMeans(n\_clusters=5)

clusters = km.fit\_predict(df.iloc[:,1:])

df["label"] = clusters

from mpl\_toolkits.mplot3d import Axes3D

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

fig = plt.figure(figsize=(20,10))

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(df.Age[df.label == 0], df["Annual Income (k$)"][df.label == 0], df["Spending Score (1-100)"][df.label == 0], c='blue', s=60)

ax.scatter(df.Age[df.label == 1], df["Annual Income (k$)"][df.label == 1], df["Spending Score (1-100)"][df.label == 1], c='red', s=60)

ax.scatter(df.Age[df.label == 2], df["Annual Income (k$)"][df.label == 2], df["Spending Score (1-100)"][df.label == 2], c='green', s=60)

ax.scatter(df.Age[df.label == 3], df["Annual Income (k$)"][df.label == 3], df["Spending Score (1-100)"][df.label == 3], c='orange', s=60)

ax.scatter(df.Age[df.label == 4], df["Annual Income (k$)"][df.label == 4], df["Spending Score (1-100)"][df.label == 4], c='purple', s=60)

ax.view\_init(30, 185)

plt.xlabel("Age")

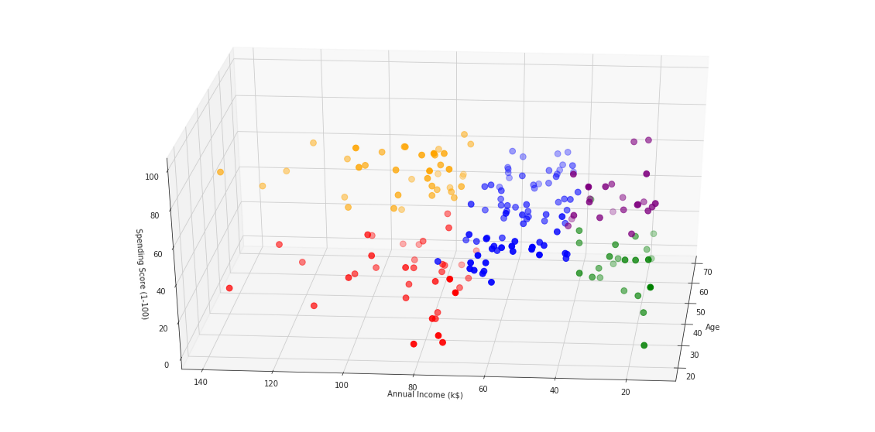
plt.ylabel("Annual Income (k$)")

ax.set\_zlabel('Spending Score (1-100)')

plt.show()

Step 5 : Pattern Evaluation

RESULT :



5. CONCLUSION AND REFERENCES

Conclusion :

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

References :

[1] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar “Telecom customer

segmentation based on cluster analysisAn Approach to Customer

Classification using k-means”, IJIRCCE,Year: 2015.

[2] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu

Electrical/Electronics & Computer Engineering Department, University of

Uyo, Uyo, Akwa Ibom State, Nigeria “Application of K-Means Algorithm

for Efficient Customer Segmentation: A Strategy for Targeted Customer

Services”, IJARAI,Year: 2015.

[3] T.NelsonGnanarajDr.K.Ramesh Kumar N.Monica“Survey on mining

clusters using new k-mean algorithm from structured and unstructured

data”, IJACST,Year: 2014.

[4] Yogita Rani and Dr. Harish Rohil“A Study of Hierarchical Clustering

Algorithm”, IJICT,Year: 2013.

[5] Omar Kettani, FaycalRamdani, BenaissaTadili“An Agglomerative

Clustering Method for Large Data Sets”, IJCA,Year: 2014.

[6] Snekha, ChetnaSachdeva, Rajesh Birok“Real Time Object Tracking

Using Different Mean Shift Techniques–a Review”, IJSCE,Year: 2013.

[7] SulekhaGoyat“The basis of market segmentation: a critical review of

literature”, EJBM,Year: 2011.

[8] Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and

Normalization Approach in Modified k-Means Clustering Algorithm”,

IJCSI,Year: 2011.

[9]Scikit-learn: https://scikit-learn.org

[10] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent

Classification & Clustering Of Lung & Oral Cancer through Decision Tree

& Genetic Algorithm, International Journal of Advanced Research in

Computer Science and Software Engineering,2015

[11] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative

and Automatic Lung and Oral Cancer Classification Using Soft Computing

Techniques, International Journal of Computer Science & Mobile

Computing,2015