**EXPOSYS DATA LABS**

**PROJECT 1**

**CUSTOMER SEGMENTATION USING**

**K-MEANS CLUSTERING**

**DONE BY**

**SYED ALEEM**

**Abstract**

Customer Relationship Management(CRM) has always played a crucial role as a market strategy for providing organizations with the quintessential business intelligence for building, managing and developing valuable long-term customer relationships. A number of business enterprises have come to realize the significance of CRM and the application of technical expertise to achieve competitive advantage. This study explores the importance of Customer Segmentation as a core function of CRM as well as the various models for segmenting customers using clustering techniques. The available clustering models for customer segmentation, in general, and the major models of K-Means and Hierarchical Clustering, in particular, are studied and the virtues and vices of the techniques are pointed out. Finally, the possibility of developing a hybrid solution by the combination of the above two techniques, having the ability to outperform the individual models, is discussed.

As the market is widening, the rate of competition between all business entities is rapidly growing. Hence, these business enterprises are increasing their expenditure on their marketing strategies to achieve competitive advantage. In this context, the significance of employing Information Technology(IT) solutions to marketing campaigns emerges as a pivotal step in a modern approach to business. Customer Segmentation is a popular technique of partitioning the customer base into externally distinct and internally uniform groups in order to create varied marketing strategies for targeting each group according to its characteristics. Generally speaking, it is defined as the process whereby the consumers of a business enterprise are divided into groups according to their preferences, characteristics and purchasing behaviour. By studying and analysing large volumes of collected customer data, businesses can improve their marketing decisions based on the customers‟ preferences. Maximum profits can be generated for any business entity if the resources are utilized judiciously in order to cultivate the most loyal and useful group of customers once customer segmentation and clustering have enabled the allocation of customers to such groups. The total customer set can be divided and grouped into clusters based on their buying behaviour, frequency, demographics etc. Hence, instead of studying each customer individually, firms can group similar customers together so that their needs can be better understood.

**TABLE OF CONTENTS**

1. **INTRODUCTION 1**
2. **EXISTING SYSTEM 2**
3. **PROPOSED SYSTEM 3**
   1. **K-MEANS CLUSTERING 3**
4. **METHODOLOGY 5**
5. **IMPLEMENTATION 6**
   1. **THE ELBOW METHOD 6**
   2. **CODE 6**
   3. **OUTPUT 8**
   4. **LIBRARIES 8**
6. **CONCLUSION 9**
7. **Introduction**

In the contemporary day and age, the importance of treating customers as the principal asset of an organization is increasing in value. Organizations are rapidly investing in developing strategies for better customer acquisition, maintenance and development. The concept of business intelligence has a crucial role to play in making it possible for organizations to use technical expertise for acquiring better customer insight for outreach programs. In this scenario, the concept of CRM garners much attention since it is a comprehensive process of acquiring and retaining customers, using business intelligence, to maximize the customer value for a business enterprise. One of the two most important objectives of CRM is customer development through customer insight. This objective of CRM entails the usage of an analytical approach in order to correctly assess customer information and analysis of the value of customers for better customer insight. Keeping up with the changing times, organizations are modifying their business flow models by employing systems engineering as well as change management and designing information technology(IT) solutions that aid them in acquiring new customers, help retain the present customer base and boost the customer’s lifelong value. Due to the diverse range of products and services available in the market as well as the intense competition among organizations, customer relationship management has come to play a significant role in the identification and analysis of a company’s best customers and the adoption of best marketing strategies to achieve and sustain competitive advantage. One of the most useful techniques in business analytics for the analysis of consumer behaviour and categorization is customer segmentation. By using clustering techniques, customers with similar means, end and behaviour are grouped together in to homogenous clusters. Customer Segmentation helps organizations in identifying or revealing distinct groups of customers who think and function differently and follow varied approaches in their spending and purchasing habits. Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. This study aims to explore the avenues of using customer segmentation, as a business intelligence tool within the CRM framework as well as the use of clustering techniques for helping organizations redeem a clearer picture of the valuable customer base. The concepts of customer relationship management, customer segmentation as a core function of CRM as well as the approach of segmenting customers using clustering techniques are discussed. The available clustering models for business analysis in the context of customer segmentation, the advantages and disadvantages of the two main models chosen for our study- K - Means and Hierarchical Clustering, as well as the possibility of developing a hybrid model which can outperform the individual models is surveyed

1. **Existing system**

Clustering, which is also called as cluster analysis, is an important subject in data mining. The data set is partitioned into some groups and the data points in each group, that is, the cluster are more related to each other than to those in other clusters. These data points are clubbed together by detecting correspondences according to the attributes found in raw data, however the main purpose is to find the suitable number of clusters which are relevant as well as insightful for analysis purposes. This process is a repeatable and iterative task where vast amounts of raw data are scanned for similarities and patterns. The unorganized data is searched for knowledge that is important and then data points are assigned. For getting favourable results, a specific clustering algorithm along with certain parameters may be best suited in a market domain. Clustering is a type of data mining technique used in a number of applications, involving areas such as machine learning, classification and pattern recognition. There are various clustering algorithms which differ from each other in terms of the approach they follow in order to do the grouping of the objects according to their characteristics. In Partitioned Based Clustering all the data points are taken as a single cluster in the beginning. These data points are then separated into clusters by iteratively positioning these objects between the clusters. Some of the partitioning algorithms are K-Means, K-Medoids and K-Modes. For Hierarchical Clustering, one of the two present approaches can be followed for implementation. One is Agglomerative (bottom-up) approach and another one is Divisive (top-down) approach. In the agglomerative process, each observation begins in its own cluster or segment and then pairs of these formed segments are combined when moving up the hierarchy. In the divisive approach, all the observations begin in one cluster and then are repeatedly split into different clusters. The results are in the form of dendrograms. In Density Based Clustering, the clusters are defined as regions of higher density than the rest of the dataset. Objects are differentiated as core, noise and border points. Grid Based Clustering algorithms partition the data set into grid structures containing a number of cells. Grid algorithm uses subspace and hierarchical clustering techniques. STING, CLIQUE are some of the grid-based clustering techniques. For implementing Model Based Clustering, the data points are grouped together on the 804 International Journal of Engineering & Technology basis of various techniques like statistical methods, conceptual and robust clustering methods. Some of the well-known algorithms are SOM, SLINK and COBWEB.

1. **Proposed system**

A. Clustering for Segmentation Purposes Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. Both, hierarchical and non-hierarchical clustering algorithms are widely used in customer segmentation, most prominent among them being K-Means and Agglomerative Hierarchical Clustering. K-Means has been used as part of their clustering approach. Also implemented K-Means for customer segmentation on their dataset. Although, hierarchical clustering algorithm seems unsuitable to many. Many have used it for intelligent customer segmentation for their research and have made use of it for applying clustering algorithms on the transaction data from a supermarket. K-means and Hierarchical Clustering algorithms are useful for clustering data and find extensive usage in customer segmentation. Hence, they will be our main focus of interest.

* 1. **K-Means Clustering**

K-Means is one of the most widely used clustering algorithms, and is simple and efficient. The aim of K-Means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori. These centroids should be placed in a wise fashion so that the results are optimal which otherwise can differ if locations of the centroids change. So, they should be placed as far as possible from each other. Each data point is then taken and associated with the nearest centroid until no data points are pending. This way an early grouping is done and at this point, k new centroids have to be recalculated as these will be the centers of the clusters formed earlier. After having calculated these centroids, the data points are then allocated to the clusters to the nearest centroids. In this iteration, the centroids change their position stepwise until no further modifications have to be done and the location of the centroids remain intact. The K-Means algorithm is relatively simple. The „K‟ cluster points, which will be the centroids, are placed in the space among the data points. Each data point is assigned to the centroid for which the distance is the least. After each data object has been assigned, centroids of the new groups are re-calculated. The above two steps are repeated until the movement of the centroid ceases. This means that the objective function of having the least squared error is completed and it cannot be improved further. Hence, we get K clusters as a result. K-Means algorithm aims at minimizing an objective function, which here, is the squared-error. It is an indicator of the distance of the data points from their respective cluster centers. The process in this algorithm always terminates but the relevance or the optimal configuration cannot be guaranteed even when the condition on the objective function is met. The algorithm is also sensitive to the selection of the initial random cluster centers. That is why it runs multiple times to reduce this effect but for a large number of data points, it tends to perform very well even though it is iterative. A variety of cluster validity indices are used, major of them are Dunn, Davies Bouldin, Silhouette, Sum of Squares within Cluster (SSWC), C, Calinski-Harabasz. SSWC is simple and the most widely used criterion to gauge the validity of the clusters. Smaller values of SSWC mean better clusters. Here we apply K-Means Clustering algorithm on a relatively small dataset and the results are depicted. The dataset is based on customer information for a mall and has 5 attributes named Customer, Genre, Age, Annual Income and Spending Score. It consists of 200 observations, each of which refers to a unique customer and the spending scores are decided and calculated by the company, based on their spending habits. Hence annual income and spending scores are the key indicators in this data. The age attribute of the customers can also be experimented with, to analyse which age group works best for a business. Any business would always keep the monetary values of any customer as top indicators. Thus, the annual income and spending scores of the customers will be best suited for clustering. As K-Means algorithm requires the number of clusters as input, below we will use the elbow method to get the optimal number of clusters which can be formed. It works on the principal that after a certain number of „K‟ clusters, the difference in SSE (Sum of Squared Errors) starts to decrease and diminishes gradually. Here, the WCSS(Within-Cluster-Sum-of-Squared-errors) metric is used as an indicator of the same. Hence, the „K‟ value, specifies the number of clusters.

In Figure 1, it can be observed that an elbow point occurs at K=5. After K=5, the difference in WCSS is not so visible. Hence, we will choose to have 5 clusters and provide the same as input to the K-Means algorithm.

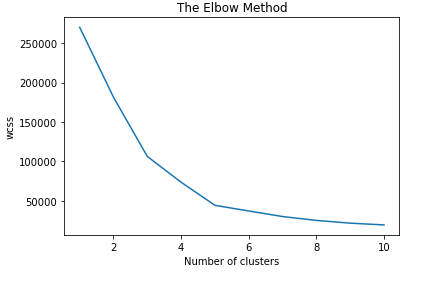


Figure 1

1. **Methodology**

Customer tiering is a method of segmentation based on how well the customer matches the goals of your business. For instance, you can use customer tiering to segment customers based on how much revenue you can expect them to bring to your business during the duration of your relationship, or by how closely that customer matches your own sales and marketing strategies.

This is a forward-thinking approach to segmentation because it ranks the importance of a customer or lead based on how much that customer can potentially bring in terms of value. Many businesses have taken tiered segmentation to a whole new level in the last few years in the form of account-based marketing, a strategy that focuses sales and marketing activities on a limited number of accounts believed to yield the highest potential value for your business. Rather than leveraging the power of big data and marketing automation to scale campaigns across a broad range of potential leads, account-based marketing turns the sights of both the sales and marketing teams toward a common goal of maximizing the potential return from a shortlist of accounts.

Demand generation marketers also recognize the potential value of tiered customer segmentation when it comes to working with your existing customer base. While marketing efforts have historically focused on lead generation activities, savvy teams leverage big data to uncover the potential value of the customers already buying from their business. Tiered segmentation allows demand generation marketers to divide existing customers based on their customer lifetime value.

Of course, like firmographic customer segmentation, the potential downside is that you cannot assume the needs of all the customers in a specific tier are the same. As a result, developing a marketing message to suit any particular tier may prove difficult.

1. **Implementation**
   1. **The Elbow Method**

Calculate the Within Cluster Sum of Squared Errors (WCSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

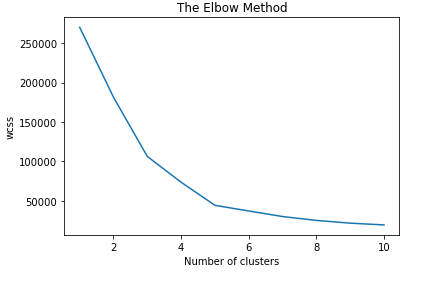
The steps can be summarized in the below steps:

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).

3. Plot the curve of WCSS vs the number of clusters K.

4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.



The optimal K value is found to be 5 using the elbow method.

* 1. **Code**

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

dataset = pd.read\_csv('Mall\_Customers.csv')

X = dataset.iloc[:, [3, 4]].values

from sklearn.cluster import KMeans

wcss = []

for i in range(1, 11):

kmeans = KMeans(n\_clusters = i, init='k-means++', random\_state = 42)

kmeans.fit(X)

wcss.append(kmeans.inertia\_)

plt.plot(range(1, 11), wcss)

plt.title('The Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel("wcss")

plt.show()

kmeans = KMeans(n\_clusters = 5, init='k-means++', random\_state=42)

y\_kmeans = kmeans.fit\_predict(X)

plt.scatter(X[y\_kmeans == 0, 0], X[y\_kmeans == 0, 1], s = 100, c = 'red', label = 'cluster 1')

plt.scatter(X[y\_kmeans == 1, 0], X[y\_kmeans == 1, 1], s = 100, c = 'blue', label = 'cluster 2')

plt.scatter(X[y\_kmeans == 2, 0], X[y\_kmeans == 2, 1], s = 100, c = 'green', label = 'cluster 3')

plt.scatter(X[y\_kmeans == 3, 0], X[y\_kmeans == 3, 1], s = 100, c = 'yellow', label = 'cluster 4')

plt.scatter(X[y\_kmeans == 4, 0], X[y\_kmeans == 4, 1], s = 100, c = 'purple', label = 'cluster 5')

plt.scatter(kmeans.cluster\_centers\_[:, 0], kmeans.cluster\_centers\_[:, 1], s = 300, c = 'black', label = 'centroids')

plt.title('Clusters of Customers')

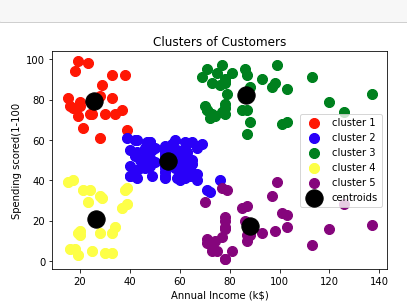
plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending scored(1-100')

plt.legend()

plt.show()

* 1. **Output**

****

* 1. **LIBRARIES**

Pandas - ver. **1.1.0**

Numpy - ver. **1.19.1**

Matplotlib - ver. **3.3.0**

Sklearn - ver. **0.23.2**

1. **Conclusions**

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.