Implementing Customer Segmentation using K-means clustering Algorithm

Internship Report

By

Anshu Kumari

A thesis submitted to

Exposys Data Labs

Sep, 2020

Abstract

Customer Segmentation is an important aspect in marketing campaigns, in identifying potentially profitable customers, and in developing customer loyalty. It is the process of grouping customers together based on common characteristics. In this paper, different types of clustering algorithms have been discussed to segment the customers. Specifically, I have implemented K-means clustering algorithms for this purpose. There are a total five feature columns with 200 different rows on a dataset which has been taken from a supermarket mall. By applying clustering, 5 segments of cluster have been formed labelled as sensible, Good, Target, Cautious, and careless customers. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

Contents

Chapter 1

Introduction

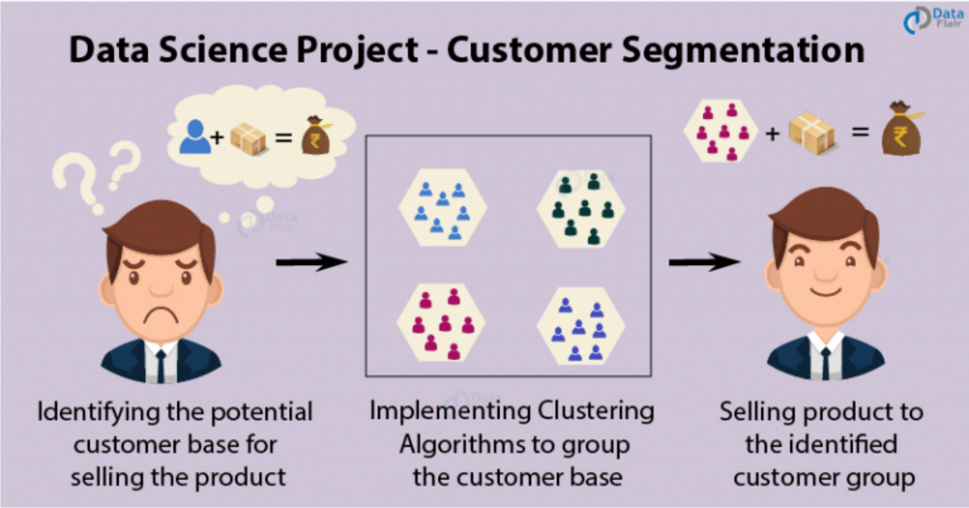
This chapter presents the introduction of the thesis that includes the brief description of Customer segmentation and their applications. This chapter also presents the scope of this thesis and the contributions of the thesis.
With the evolution of new technologies and increasing growth of e-commerce it is important for every business to adapt new strategies which help them to win the competitive environment. E-commerce transactions are no longer a new thing. Many people shop with e-commerce and many companies use e-commerce to promote and to sell their products. Because of that, overloading information appears on the customers' side. Overloading information occurs when customers get too much information about a product then feel confused. Personalization will become a solution to overloading problems. Here, machine learning comes into the play to solve the problem.

1.1 Customer Segmentation

In marketing, one way to increase profits is to communicate with customers to determine customer wishes. Communication is built according to the characteristics of the customer. Communication is very difficult to create using personal approaches. So it is necessary to divide customers into groups that have the same characteristics, and this is called customer segmentation.
Customer segmentation has several benefits:

- It enables us to match between the customer and an offer of similar products.
- It changes the way we communicate with the customer based on customer data.
- It identifies the most profitable customers.
- It enables us to update the products and services to meet customer needs

This chapter will discuss the customer data for customer segmentation, customer segmentation methods, and customer segmentation process, and then the methods will be classified based on data processing.


Data Science Project - Customer Segmentation
Data Flair
Identifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product to the identified customer group

1.2 Data for Customer Segmentation:

This dataset is composed by the following five features:

- *CustomerID:* Unique ID assigned to the customer
- *Gender:* Gender of the customer
- *Age:* Age of the customer
- *Annual Income (k$):* Annual Income of the customer
- *Spending Score (1-100):* Score assigned by the mall based on customer behavior and spending nature.

The data contains 200 records with 5 features- Customer Id, Gender, Age, Annual Income (k$) and Spending Score (1–100). Here Spending Score refers to Score assigned by the mall based on customer behavior and spending nature.

1.3 The importance of customer segmentation

For a growing business, segmentation is necessary to know our customers and our market, and share this understanding across teams. Beyond a certain size, it's impossible to do without.
At Intercom, we have benefitted from customer segmentation in these ways:

- Describing types of customers in a common way across go-to-market, product, and engineering. For example, our Sales team is now able to give segmented customer feedback to our product leaders to influence our roadmap.
- Understanding our most and least engaged customers at a granular level. For example, our Analytics team might find that one customer segment tends to use our product weekly, while another just monthly.

- Surfacing promising or untapped business opportunities. For example, imagine this scenario: our Marketing team discovers a new segment that's already converting well without having been explicitly targeted before.

- Enabling us to make tactical decisions with a holistic view of our customers. For example, our Product team could decide to build a data export API after learning our fastest growing segment exports their conversation data far more often than other segments.

- Informing our approach to the market. For example, our leadership team might decide to focus our company strategy on targeting the segments with the best revenue retention.

- Assessing progress on our marketing strategy. For example, our Finance team is now able to confirm whether new customer growth is up in our target segments.

1.4 Roadmap of the Thesis

The structure of the thesis is as follows:

1. The ***Chapter 1*** is an introductory part which discusses the scope of the thesis, about the contribution of this thesis and the motivation for writing it.

2. The ***Chapter 2*** discusses the existing methods used to clustering the customers.

3. The ***Chapter 3*** discusses in details about the proposed method that is, K-means Algorithm with architecture.

4. The ***Chapter 4*** discusses methods to optimize the number of clusters in the dataset.

5. The ***Chapter 5*** discusses the implementation and their analysis to segment the customers in the dataset.

6. The ***Chapter 6*** comprises of the conclusion and further work of the thesis.

Chapter 2

Existing Methods

2.1 K-Means Algorithm

K-Means is one of the most famous algorithm for clustering.sum of squared distances of samples to their closest cluster center scores of number of cluster to select number of groups.Calculate the Within Cluster Sum of Squared Errors (WSS) 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 optimal K value is found to be 5 using the elbow method.

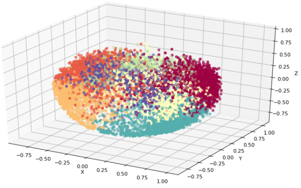


Figure 2.1: K-Means Clustering Algorithm

2.2 MiniBatch K-Means

The MiniBatch K-Means is faster than K-Means. However, sometimes it gives a slight different result and after n-cluster is determined 6 cluster according to the metrics. The main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence.

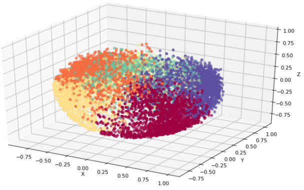


Figure 2.2: MiniBatch K-Means

2.3 Hierarchical Clustering

An algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Hierarchical clustering is a clustering technique that aims to create a tree like clustering hierarchy within the data.We pick some k to be the branching factor. This defines the number of clusters.

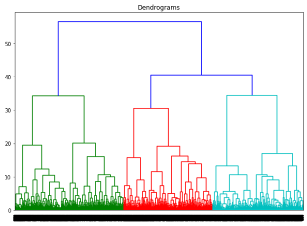


Figure 2.3: Hierarchical Clustering

2.4 DBSCAN

DBSCAN, as the name implies, is a density-based clustering algorithm. Density refers to the proximity of data points in a cluster and it is good for data which contains clusters of a similar density.Divides the dataset into n dimensions. DBSCAN counts this shape as a cluster. DBSCAN iteratively expands the cluster, by going through each individual point within the cluster, and counting the number of other data points nearby.

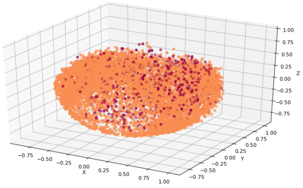


Figure 2.4: DBSCAN Clustering

2.5 GMM Algorithm

Gaussian Mixture Models (GMMs) assume there are a number of Gaussian distributions, and each of them represents a cluster. Therefore a Gaussian Mixture Model tends to group together the data points that belong to a single distribution. Here rather than identifying clusters by “nearest” centroids, we fit a set of k gaussians to the data. And we estimate gaussian distribution parameters such as mean and Variance for each cluster and weight of a cluster. After learning the parameters for each data point we can calculate the probabilities of it belonging to each of the clusters.

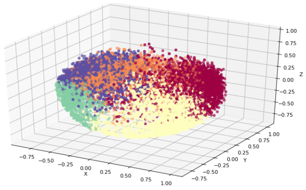


Figure 2.5: GMM Algorithm

2.6 MeanShift

Meanshift is falling under the category of a clustering algorithm in contrast of Unsupervised learning that assigns the data points to the clusters iteratively by shifting points towards the mode (mode is the highest density of data points in the region, in the context of the Meanshift). Unlike the popular K-Means cluster algorithm, mean-shift does not require specifying the number of clusters in advance. The number of clusters is determined by the algorithm with respect to the data.

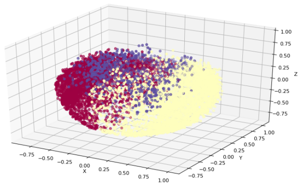


Figure 2.6: MeanShift Algorithm

Chapter 3

K-means Algorithm with Architecture

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It 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.

3.1 K-means Clustering Method:

The way kmeans algorithm works is as follows:

1. Specify number of clusters K.
2. Partition of objects into k non-empty subsets
3. Identifying the cluster centroids (mean point) of the current partition.
4. Assigning each point to a specific cluster
5. Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
6. After re-allotting the points, find the centroid of the new cluster formed.

- 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.

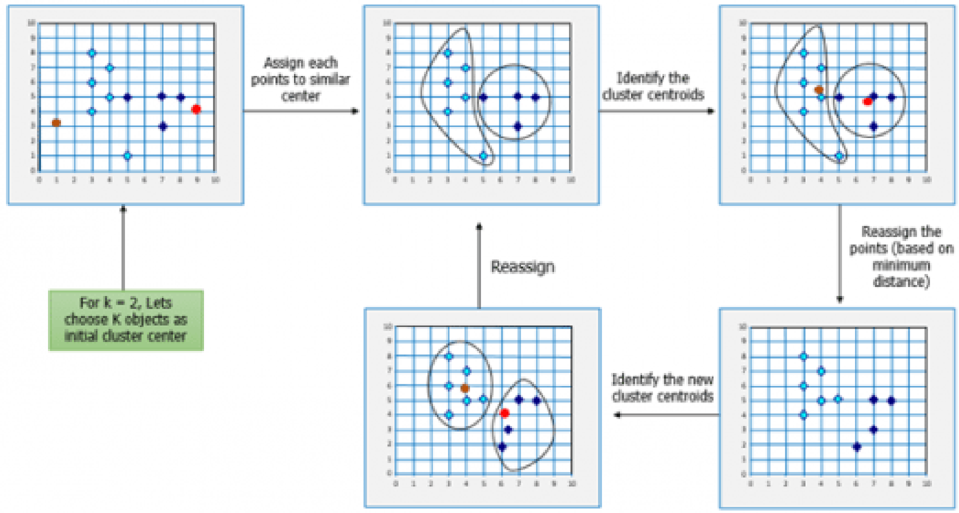

Assign each points to similar center
Identify the cluster centroids
Reassign the points (based on minimum distance)
Identify the new cluster centroids
Reassign
For k = 2, Lets choose K objects as initial cluster center

Figure 3.1: K-Means Clustering Algorithm

3.2 Applications of K-Means Clustering

K-Means clustering is used in a variety of examples or business cases in real life, like:

- ***Academic Performance:*** Based on the scores, students are categorized into grades like A, B, or C.
- ***Diagnostic systems:*** Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.
- ***Search engines:*** Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.
- ***Document clustering***
- ***Identifying crime-prone areas***
- ***Insurance fraud detection***
- ***Public transport data analysis***
- ***clustering of IT alerts***
- ***Behavioural Segmentation***
- ***Anomaly Detection***
- ***Social Network Analysis***
- ***Customer Segmentation***

There are just a few examples where clustering algorithm like K-means is applied. The clustering algorithm plays the role of finding the cluster heads, which collects all the data in its respective cluster.

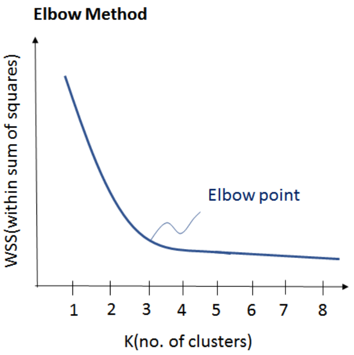
Chapter 4

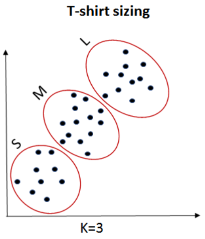
Methodology: Optimal Number of Clusters

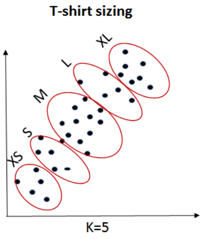
4.1 Methods to choose optimal number of clusters

Now, the important question is how should we choose the optimum number of clusters? There are two possible ways for choosing the number of clusters.

1. ***Elbow Method:*** Here, I draw a curve between WSS (within sum of squares) and the number of clusters. It is called elbow method because the curve looks like a human arm and the elbow point gives us the optimum number of clusters. As we can see that after the elbow point, there is a very slow change in the value of WSS, so we should take the elbow point value as the final number of clusters.







2. ***Purpose Based:*** We can run k-means clustering algorithm to get different clusters based on a variety of purposes. We can partition the data on different metrics and see how well it performs for that particular case. Let's take an example of marketing T-shirts of different sizes. We can partition the dataset into different number of clusters depending upon the purpose that we want to meet. In the following example, I have taken two different criteria, price and comfort.
   Let's see these two possibilities as shown in the image below.

   (a) ***K=3:*** If we want to provide only 3 sizes(S, M, L) so that prices are cheaper, we will divide the data set into 3 clusters.

   (b) ***K=5:*** Now, if we want to provide more comfort and variety to our customers with more sizes (XS, S, M, L, XL), then we will divide the data set into 5 clusters.

4.2 Steps to implement K-means Algorithm

Now, once we have the value of k with us, let's understand it's execution.

- ***Initialisation:***
  Firstly, we need to randomly initialise two points called the cluster centroids. Here, we need to make sure that our cluster centroids depicted by an orange and blue cross as shown in the image are less than the training data points depicted by navy blue dots. k-means clustering algorithm is an iterative algorithm and it follows next two steps iteratively. Once we are done with the initialization, let's move on to the next step.

- ***Cluster assignment:***
  In this step, it will go through all the navy blue data points to compute the distance between the data points and the cluster centroid initialised in the previous step. Now, depending upon the minimum distance from the orange cluster centroid or blue cluster centroid, it will group itself into that particular group. So, data points are divided into two groups, one represented by orange color and the other one in blue color as shown in the graph. Since these cluster formations are not the optimised clusters, so let's move ahead and see how to get final clusters.

- ***Move Centroid:***
  Now, we will take the above two cluster centroids and iteratively reposition them for optimization. we will take all blue dots, compute their average and move current cluster centroid to this new location. Similarly, we will move orange cluster centroid to the

average of orange data points. Therefore, the new cluster centroids will look as shown in the graph. Moving forward, let's see how can we optimize clusters which will give us better insight.

- ***Optimization:***
  We need to repeat above two steps iteratively till the cluster centroids stop changing their positions and become static. Once the clusters become static, then k-means clustering algorithm is said to be converged.

- ***Convergence:***
  Finally, k-means clustering algorithm converges and divides the data points into two clusters clearly visible in orange and blue. It can happen that k-means may end up converging with different solutions depending on how the clusters were initialised.

Chapter 5

Implementation

5.1 Environment and tools

1. scikit-learn - 0.23.2
2. seaborn - 0.11
3. numpy -1.19
4. pandas-1.1.2
5. matplotlib-3.3.2

5.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA for short) is an important aspect to analyze any datasets. It is generally used to insight datasets by visual representation and most importantly looking for patterns and relations within feature columns of the dataset.

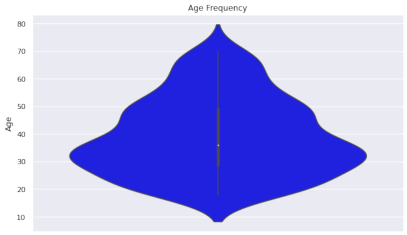


Figure 5.1: The distribution of age on the basis of frequency

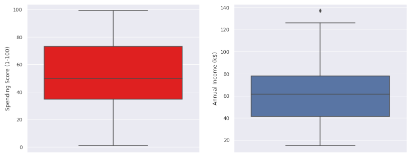


Figure 5.2: Distribution of Spending Score and Annual Income

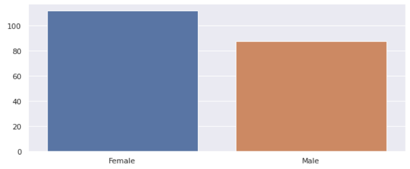


Figure 5.3: Bar plot of Gender Distribution

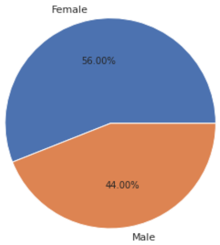


Figure 5.4: Pie plot of Gender Distribution

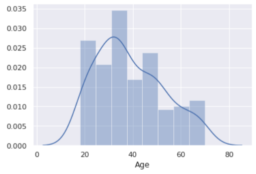


Figure 5.5: Variance of age

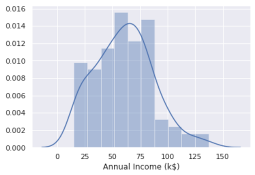


Figure 5.6: Variance of Annual Income

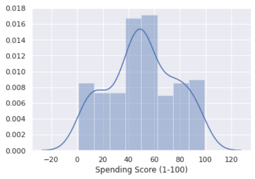


Figure 5.7: Variance of Spending Score

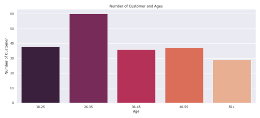


Figure 5.8: Number of Clusters vs Age Groups

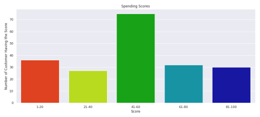


Figure 5.9: Number of Clusters vs Spending Score Groups

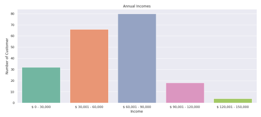


Figure 5.10: Number of Clusters vs Annual Income Groups

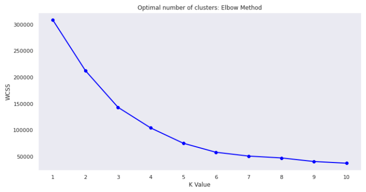


Figure 5.11: The optimal number of clusters by Elbow method

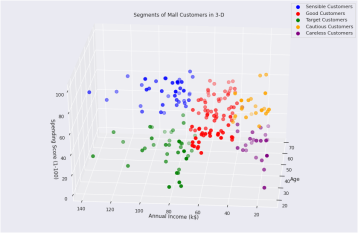


Figure 5.12: Customer Segmentation in 3-D view

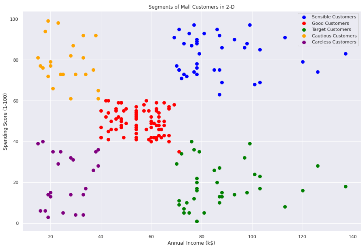


Figure 5.13: Customer Segmentation in 2-D view

5.3 Different types of clusters

After plotting the results obtained by K-means on this 3D graphic, it's our job now to identify and describe the five clusters that have been created:

1. ***Sensible customers:*** The blue cluster groups young people with moderate to low annual income who actually spend a lot.
2. ***Good customers:*** The red cluster groups reasonably young people with pretty decent salaries who spend a lot.
3. ***Target customers:*** The green cluster basically groups people of all ages whose salary isn't pretty high and their spending score is moderate.
4. ***Cautious customers:*** The orange cluster groups people who actually have pretty good salaries and barely spend money, their age usually lays between thirty and sixty years.
5. ***Careless customers:*** The purple cluster groups whose salary is pretty low and don't spend much money in stores, they are people of all ages.

Chapter 6

Conclusion

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation needed to get potential customers used to increase profits. Potential customer data can be used to provide service the characteristics of customer including E-commerce services as a media buying and selling online.

6.1 Summary

In this paper, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. It 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. We analyzed and visualized the data and then proceeded to implement our algorithm. we divided the data into seven clusters, because five clusters can be easily used to determine the behaviours of customers. However, each of the clusters have their own characteristics.

6.2 Some Observations

We can conclude the following observations from the dataset:

- There isn't much difference between the spending score of women and men, which leads us to think that our behaviour when it comes to shopping is pretty similar.
- Observing the clustering graphic, it can be clearly observed that the ones who spend more money in malls are young people. That is to say they are the main target when it comes to marketing, so doing deeper studies about what they are interested in may lead to higher profits.
- Although young people seem to be the ones spending the most, we can't forget there are more people we have to consider, like people who belong to the pink cluster, they are what we would commonly name after "middle class" and it seems to be the biggest cluster.

- Promoting discounts on some shops can be something of interest to those who don't actually spend a lot and they may end up spending more.

6.3 Future Work

Although K-Means Clustering is a powerful technique in order to achieve a decent customer segmentation. But it is a very simple algorithm.

It has some limitations:

- It fails for non-linear data set.
- It unables to handle noisy data and outliers.
- It is applicable only when mean is defined i.e. fails for categorical data.

These are the some limitations which we need to improve in the future.