Customer Segmentation Analysis in R

Gayathri Kurapati

15/06/2022

Abstract

With the technological advancements and increase in the usage of digital devices like smartphones and computers, a vast amount of data is collected daily. Analysis of this data is very important. This comes in as a result of increased competition, especially for businesses where one wants to be better than everyone. Therefore, there needs to have business strategies that fit in with the modern era of innovation. Businesses are successful depending on the number of potential customers to acquire goods or services. However, there is a challenge that many businesses have been experiencing when it comes to identifying their potential target customers and what products or services they acquire at any given time. This, therefore, brings in the need to take and utilize data in making proper decisions. In terms of purchasing goods or services, every person is different from one another, and it is not easy to just look and identify what one person buys or likes. However, with the help of machine learning methods, a business can identify any hidden patterns in data that are essential for proper decision-making. Therefore, customer segmentation through the K-Means clustering algorithm is implemented in the project that helps group available customers with the same attributes. The dataset is obtained from the Kaggle.com website, and the customers are segmented according to their similar characteristics in terms of goods purchases and what they like as a group. The elbow method is utilized to determine the optimal clusters.

Introduction

Over the years, there has been an increase in competition among businesses. The availability of the vast amount of data has called for using data mining techniques to help extract strategic and meaningful insights. Data mining involves using different techniques to extract patterns from data to ensure the data has been presented in a human-readable format and can be utilized in decision-making (Pal, 2011). With clustering, the data objects are partitioned into clusters to ensure similar objects are in a cluster. Therefore, customer segmentation involves the process through which there is a division of the customer base into different groups identified as customer segments in such a way that each customer segment is made up of customers who have the same attributes (Camilleri, 2018). Interests, gender, age, and spending habits are some of the different ways that segmentation is done concerning marketing for businesses. Customer segmentation is essential as it helps businesses identify the products that are related to each customer segment and ensure proper management of the product's supply and demand, targeting and identifying the target potential customer base, modification of market programs, and have predictions on the customer's defection (Kumar, 2010). Therefore, this project utilizes the K-means clustering algorithm to identify customer segments. With this customer segmentation, it will be possible for businesses to gain a competitive advantage, increase marketing efficiency, identify customer retention, determine new market opportunities, and demonstrate better knowledge of their customer's needs. ## Literature Review With the increased competition among businesses over the years, there is a need to ensure that businesses find better strategies that help in satisfying their customer's demands and also, as per the customer's needs, they attract new ones to enhance their profits. Manually it is tedious to identify customers and satisfy each customer's demand (Dominici & Palumbo, 2013). This is because customers may be very different in terms of preferences, demands, and tastes.

With the variation in the different types of customers of a retail distributor, this project needs to understand how to best structure the delivery of services and ensure the availability of products to meet their customer's needs to ensure satisfaction. Therefore, to avoid the phrase one-size-fits-all, it is recommended to perform customer segmentation to ensure that customers are clustered into groups with the same behavioral characteristics or attributes. In a study by Baghel and Parashar (2020), customer segmentation is identified as a strategy involving dividing the market into homogenous groups. The authors used economic, demographical, and geographical data conditions for the customer segmentation technique (Baghel & Parashar, 2020). With clustering, it involves the process through which information is grouped into a dataset based on some similar attributes or characteristics (Ceri et al., 2013). Several algorithms are available that can be applied to perform clustering on datasets. Therefore, the project will consider whether, with the available dataset, it is possible to segment customers into segment groups by considering their homogenous attributes and purchasing preferences.

Data

The data for this project has been downloaded from

Having the data imported into the R environment. The next step includes identifying its dimensions, the names, and a summary to understand the variables we will work on within the project.

dim(customerData)

## [1] 200 5

str(customerData)

## 'data.frame': 200 obs. of 5 variables:

## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...

## $ Gender : chr "Male" "Male" "Female" "Female" ...

## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...

## $ Annual.Income..k.. : int 15 15 16 16 17 17 18 18 19 19 ...

## $ Spending.Score..1.100.: int 39 81 6 77 40 76 6 94 3 72 ...

names(customerData)

## [1] "CustomerID" "Gender" "Age"

## [4] "Annual.Income..k.." "Spending.Score..1.100."

The data is made up of 200 records and 5 variables. Next step, to ensure accurate results, we check for any missing values in the dataset.

colSums(is.na(customerData))

## CustomerID Gender Age

## 0 0 0

## Annual.Income..k.. Spending.Score..1.100.

## 0 0

The dataset does not have any missing values in its columns. This means that not much data cleaning was needed for this dataset.

Methodology

Age will be the main variable for the analysis in this project, so we need to have more information about it. First, using the head() function, the first few rows of the age variable are displayed, and a summary of the variable is also displayed using the summary() function.

summary(customerData$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 18.00 28.75 36.00 38.85 49.00 70.00

sd(customerData$Age)

## [1] 13.96901

Next, we need to check the correlation between the variables in the dataset.

#install.packages("GGally")

library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':

## method from

## +.gg ggplot2

ggcorr(customerData, label = TRUE, label\_alpha = TRUE)

## Warning in ggcorr(customerData, label = TRUE, label\_alpha = TRUE): data in

## column(s) 'Gender' are not numeric and were ignored



Next step will involve clustering. In this project, the gap statistic method have been used to help in finding the optimal number of clusters.

#install.packages("NbClust")

#install.packages("factoextra")

library(cluster)

library(gridExtra)

library(grid)

library(NbClust)

library(factoextra)

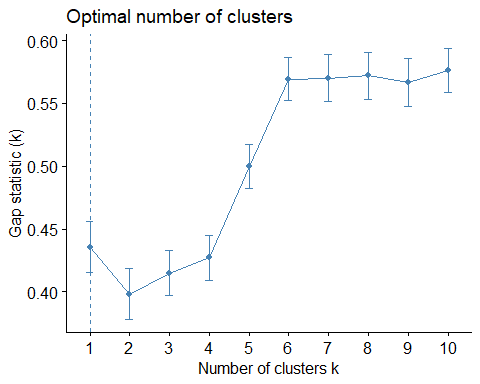
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

set.seed(125)

stat\_gap <- clusGap(customerData[,3:5], FUN = kmeans, nstart = 25,

K.max = 10, B = 50)

fviz\_gap\_stat(stat\_gap)



The optimal cluster is k=6.

Results

With the optimal clusters identified, the segmentation process was performed. The 6 clusters were plotted.

k6<-kmeans(customerData[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")

k6

## K-means clustering with 6 clusters of sizes 45, 21, 35, 39, 38, 22

##

## Cluster means:

## Age Annual.Income..k.. Spending.Score..1.100.

## 1 56.15556 53.37778 49.08889

## 2 44.14286 25.14286 19.52381

## 3 41.68571 88.22857 17.28571

## 4 32.69231 86.53846 82.12821

## 5 27.00000 56.65789 49.13158

## 6 25.27273 25.72727 79.36364

##

## Clustering vector:

## [1] 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2

## [38] 6 2 6 1 6 1 5 2 6 1 5 5 5 1 5 5 1 1 1 1 1 5 1 1 5 1 1 1 5 1 1 5 5 1 1 1 1

## [75] 1 5 1 5 5 1 1 5 1 1 5 1 1 5 5 1 1 5 1 5 5 5 1 5 1 5 5 1 1 5 1 5 1 1 1 1 1

## [112] 5 5 5 5 5 1 1 1 1 5 5 5 4 5 4 3 4 3 4 3 4 5 4 3 4 3 4 3 4 3 4 5 4 3 4 3 4

## [149] 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3

## [186] 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4

##

## Within cluster sum of squares by cluster:

## [1] 8062.133 7732.381 16690.857 13972.359 7742.895 4099.818

## (between\_SS / total\_SS = 81.1 %)

##

## Available components:

##

## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"

## [6] "betweenss" "size" "iter" "ifault"

pcclust= prcomp(customerData[,3:5],scale=FALSE)

summary(pcclust)

## Importance of components:

## PC1 PC2 PC3

## Standard deviation 26.4625 26.1597 12.9317

## Proportion of Variance 0.4512 0.4410 0.1078

## Cumulative Proportion 0.4512 0.8922 1.0000

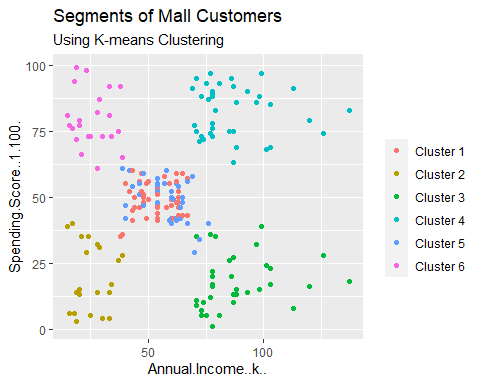
pcclust$rotation[,1:2]

## PC1 PC2

## Age 0.1889742 -0.1309652

## Annual.Income..k.. -0.5886410 -0.8083757

## Spending.Score..1.100. -0.7859965 0.5739136

Visual illustration of the distribution of the 6 clusters. 

Cluster 1 is the representation of the customerData with high annual spending and annual income. Cluster 2 is a representation of low yearly spending and high annual income. Cluster 3 represents the customerData with low yearly and low annual income. Clusters 4 and 6 involve the customerData for medium annual salary spend and medium-income salary. Cluster 5 includes the high yearly expenditure as well as low annual income.

Conclusion

Having performed customer segmentation for the customerData, it is possible to effectively understand variables, which is essential in making careful decisions. This is because it is possible to identify customers, which is helpful for businesses in releasing services and products that target the potential customers concerning different parameters such as spending patterns, age, income, and many others. With the project, for better segmentation, it has been noted that there needs to put into consideration patterns such as product reviews.

**References**

Baghel, M. T. S. K. S., & Parashar, L. (2020). Customer Segmentation Using Clustering and Data Mining Techniques.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In Travel marketing, tourism economics and the airline product (pp. 69-83). Springer, Cham.

Ceri, S., Bozzon, A., Brambilla, M., Valle, E. D., Fraternali, P., & Quarteroni, S. (2013). Classification and Clustering. In Web Information Retrieval (pp. 39-56). Springer, Berlin, Heidelberg.

Dominici, G., & Palumbo, F. (2013). How to build an e-learning product: Factors for student/customer satisfaction. Business Horizons, 56(1), 87-96.

Kumar, V. (2010). A customer lifetime value-based approach to marketing in the multichannel, multimedia retailing environment. Journal of Interactive Marketing, 24(2), 71-85.

Pal, J. K. (2011). Usefulness and applications of data mining in extracting information from different perspectives. Annals of Library and Information Studies (ISSN: 0972-5423), 58(1), 7-16.