Assignment 3: Customer Segmentation Project using Machine Learning in R

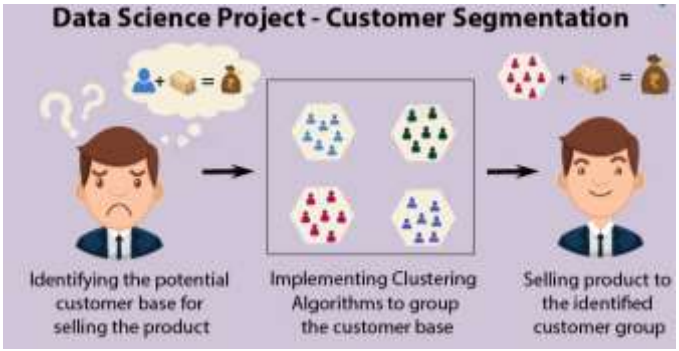
Objective: Data Science R Project - Customer Segmentation. Implement customer segmentation in R with a customer data set provided. Share the R script with all relevant comments and present the code and output in a word document with commentary.

ZIP the two files in a file name: ID_Name_Section_Assignment3.zip

Submission date: 17th December 2020

In this machine learning project, we will explore the customer data, upon which we will be building our customer segmentation model. Also, in this data science project, we will see the descriptive analysis of our data and then implement several versions of the K-means algorithm.

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of ***K-means clustering*** which is the essential algorithm for clustering unlabeled dataset.



What is Customer Segmentation?

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

Implement Customer Segmentation in R?

Suggested steps:

1. Use Mall_Customers.csv customer data file for analysis. Perform data exploration. Import the essential packages required for this role and then read customer data. Go through the input data to gain necessary insights about it.
2. Now display the first six rows of our customer dataset and use the summary() function to output summary of it.
3. Create a barplot and a piechart to show the gender distribution across our customer_data dataset.
4. From the barplot, observe that the number of females is higher than the males and visualize a pie chart to observe the ratio of male and female distribution. From the above graph, we conclude that the percentage of females is **56%**, whereas the percentage of male in the customer dataset is **44%**.
5. Plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.
6. From the above two visualizations, conclude that the maximum customer ages are between 30 and 35. Identify the minimum and maximum age of customers.
7. Create visualizations to analyze the annual income of the customers. Plot a histogram and then proceed to examine this data using a density plot.
8. From the above descriptive analysis, conclude that the minimum annual income of the customers is 15 and the maximum income is 137. People earning an average income of 70 have the highest frequency count in our histogram distribution. The average salary of all the customers is 60.56. Also observe that the annual income has a ***normal distribution***.
9. The minimum spending score is 1, maximum is 99 and the average is 50.20. We can see Descriptive Analysis of Spending Score is that Min is 1, Max is 99 and avg. is 50.20. From the histogram, show that customers between class 40 and 50 have the highest spending score among all the classes.
10. Describe the k-means clustering algorithm. Summing up the K-means clustering:

- specify the number of clusters that we need to create.
- The algorithm selects k objects at random from the dataset. This object is the initial cluster or mean.
- The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid.
- k clusters in the data points update the centroid through calculation of the new mean values present in all the data points of the cluster. The kth cluster's centroid has a length of p that contains means of all variables for observations in the k-th cluster. We denote the number of variables with p.
- Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment

stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

11. Determining Optimal Clusters. While working with clusters, you need to specify the number of clusters to use. You would like to utilize the optimal number of clusters. To help you in determining the optimal clusters, there are three popular methods –

    - Elbow method
    - Silhouette method
    - Gap statistic

    Calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. Then calculate the total intra-cluster sum of square (iss). Then, proceed to plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters. Implement this in R. From the above graph, we conclude that 4 is the appropriate number of clusters since it seems to be appearing at the bend in the elbow plot.

12. With the help of the average silhouette method, measure the quality of our clustering operation. With this, determine how well within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k clusters. Using the silhouette function in the cluster package, we can compute the average silhouette width using the kmean function. Here, the optimal cluster will possess highest average.
13. **Gap Statistic Method** - We can use this method to any of the clustering method like K-means, hierarchical clustering etc. Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. With the help of **Monte Carlo simulations**, one can produce the sample dataset. For each variable in the dataset, we can calculate the range between min(xi) and max (xj) through which we can produce values uniformly from interval lower bound to upper bound. For computing the gap statistics method utilize the clusGap function for providing gap statistic as well as standard error. In the output of our kmeans operation, we observe a list with several key information. From this, we conclude the useful information being –

- **cluster –** This is a vector of several integers that denote the cluster which has an allocation of each point.
- **totss –** This represents the total sum of squares.

- **centers –** Matrix comprising of several cluster centers
- **withinss –** This is a vector representing the intra-cluster sum of squares having one component per cluster.
- **tot.withinss –** This denotes the total intra-cluster sum of squares.
- **betweenss –** This is the sum of between-cluster squares.
- **size –** The total number of points that each cluster holds.

14. Visualizing the Clustering Results using the First Two Principle Components. From the above visualization, observe that there is a distribution of 6 clusters as follows –

    Cluster 6 and 4 – These clusters represent the customer_data with the medium income salary as well as the medium annual spend of salary.

    Cluster 1 – This cluster represents the customer_data having a high annual income as well as a high annual spend.

    Cluster 3 – This cluster denotes the customer_data with low annual income as well as low yearly spend of income.

    Cluster 2 – This cluster denotes a high annual income and low yearly spend.

    Cluster 5 – This cluster represents a low annual income but its high yearly expenditure.

15. Summarize your learning from this customer segmentation project of machine learning using R. With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation.

