CUSTOMER SEGMENTATION

Sudhanshu Chouhan, Hritik Sharma
Data science trainees,
AlmaBetter, Bangalore

Abstract

The goal of "serving all" is similar to "serving none". Marketers are constantly looking for ways to refine the way they segment markets. Segmentation involves diving markets into smaller portions (segments) of consumers with similar needs for a given good or service. This tutorial-like study explores the application of various algorithms and analytical techniques that are used to segment markets. These techniques include regression, cross-tabulation, hierarchical clustering, and K-Means clustering performed through analytical tools such as R-Studio and MS Excel. The analyses drew upon the "Customer Data" dataset from "Kaggle", which contained eight variables: age, income, marital status, ownership status, household size, family total sales and family total visit. The findings demonstrate how such statistics could help the businesses understand the customers and target the specific customer persona with unique campaigns and offerings.

Problem Statement

Online Retail Customer Segmentation Online-Retail-Customer-Segmentation In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

. Introduction

Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviours, so you can market to those customers more effectively.

These customer segmentation groups can also be used to begin discussions of building a market persona. This is because customer segmentation is typically used to inform a brand's messaging, positioning and to improve how a business sells – so marketing personas need to be closely aligned to those customer segments in order to be effective.

The marketing "persona" is by definition a personification of a customer segment, and it is not uncommon for businesses to create several personas to match their different customer segments.

But for that to happen, a business needs a robust set of customer segments off of which to base it. Which leads us to the next section, distinguishing the difference between customer segmentation and market segmentation, so that your segmentation is as accurate as possible.

Steps Involved:

Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

Exploratory data analysis

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

K-means clustering

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

The term ‘K’ is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

For a better understanding of k-means, let's take an example from cricket. Imagine you received data on a lot of cricket players from all over the world, which gives information on the runs scored by the player and the wickets taken by them in the last ten matches. Based on this information, we need to group the data into two clusters, namely batsman and bowlers.

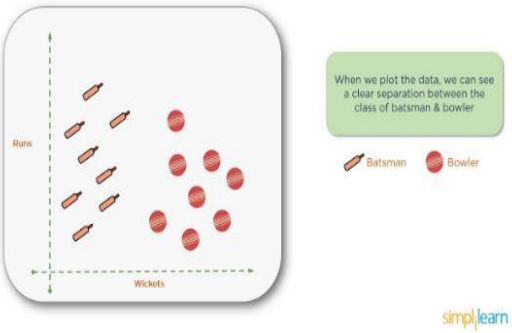
Let's take a look at the steps to create these clusters.

Solution:

Assign data points

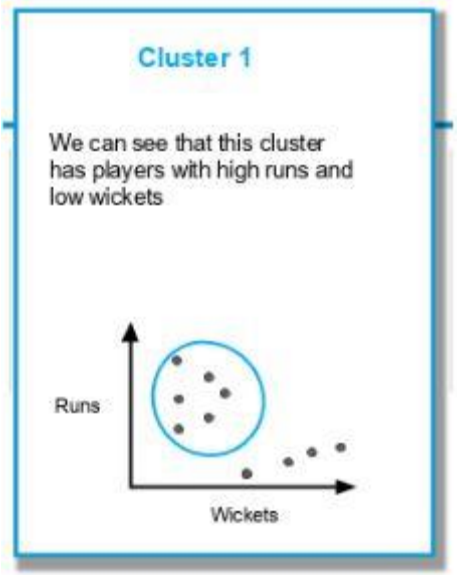
Here, we have our data set plotted on ‘x’ and ‘y’ coordinates. The information on the y-axis is about the runs scored, and on the x-axis about the wickets taken by the players.

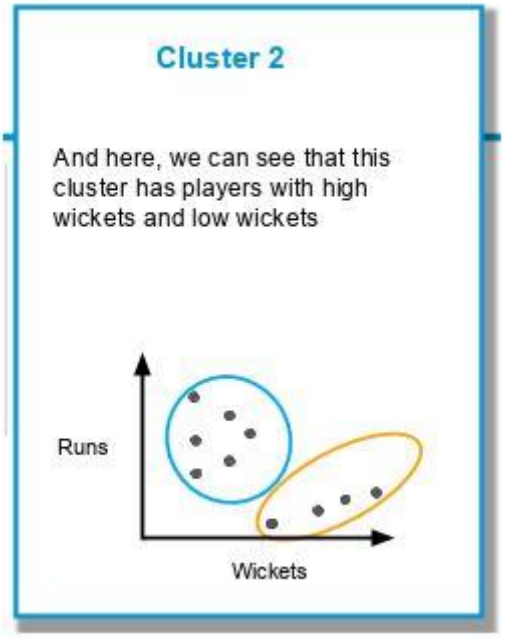
If we plot the data, this is how it would look:



Perform Clustering

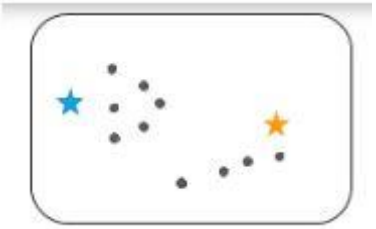
We need to create the clusters, as shown below:





Considering the same data set, let us solve the problem using K-Means clustering (taking K = 2).

The first step in k-means clustering is the allocation of two centroids randomly (as K=2). Two points are assigned as centroids. Note that the points can be anywhere, as they are random points. They are called centroids, but initially, they are not the central point of a given data set.



The next step is to determine the distance between each of the randomly assigned centroids' data points. For every point, the distance is measured from both the centroids, and whichever distance is less, that point is assigned to that centroid. You can see the data points attached to the centroids and represented here in blue and yellow.

The next step is to determine the actual centroid for these two clusters. The original randomly allocated centroid is to be repositioned to the actual centroid of the clusters.

This process of calculating the distance and repositioning the centroid continues until we obtain our final cluster. Then the centroid repositioning stops.

As seen above, the centroid doesn't need anymore repositioning, and it means the algorithm has converged, and we have the two clusters with a centroid.