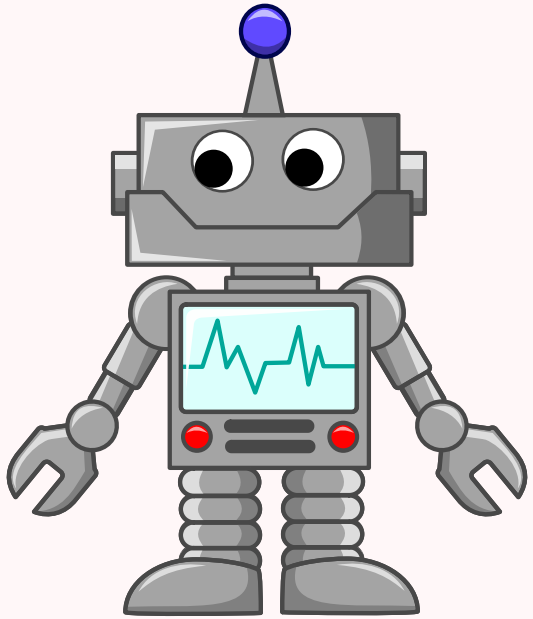
The Future of Customer Segmentation, Artificial Intelligence



What's in store for us?

Project Description

What is the problem?

Products are not tailored and advertised in a way attractive to different types of demographic, this causes a reduction in sales due to the lack of attention by the customer.

The cause

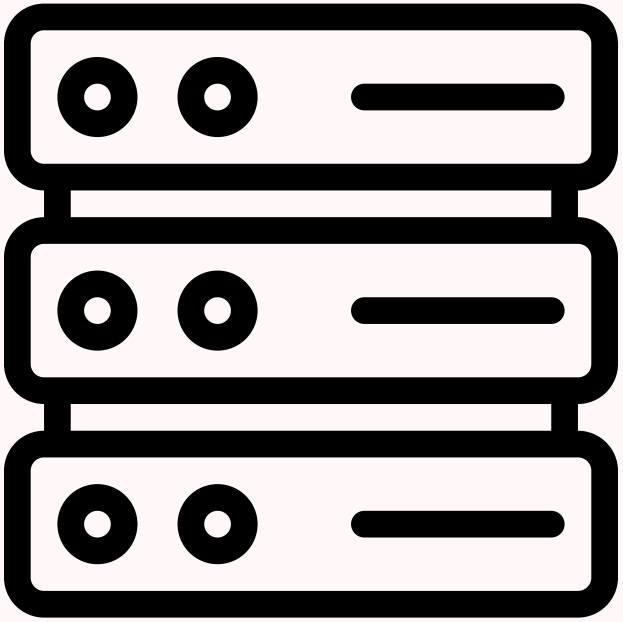
It is caused due to the lack of market segmentation and resources to segment the market, often it is expensive to conduct primary research this is as,
for example, data packages advertised are too general instead of specific rendering the advertising useless. There has to be a more targeted advertising
and marketing to different segments of people who use telecommunication, for example, younger people will see more advertisements for Etisalat and Du on social media
while older people more on billboards and calls to upgrade their current package

The solution

Customer segmentation.
To have a more specific way of advertising and marketing the customers need to be separated based on similar characteristics, this is so advertising can be more targeted
to a certain demographic to be more appealing, for higher sales

Data Description

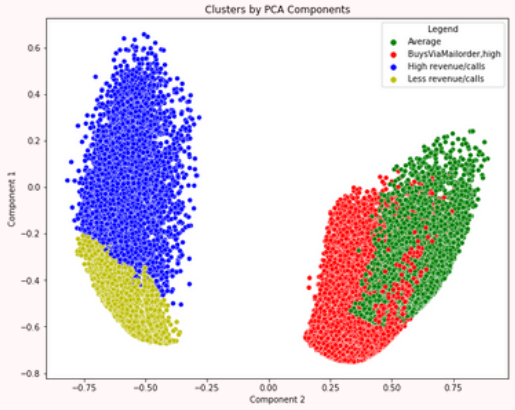
The dataset contains data about usage of customers, it contains the amount charged to the customers as well as the the monthly revenue and the monthly minutes per customer, whether they own computers and if they have credit cards.



Dataset

https://drive.google.com/file/d/1id8I-Tw0XYedJiKk4KeJuHq5mdDevG58/view?usp=sharing

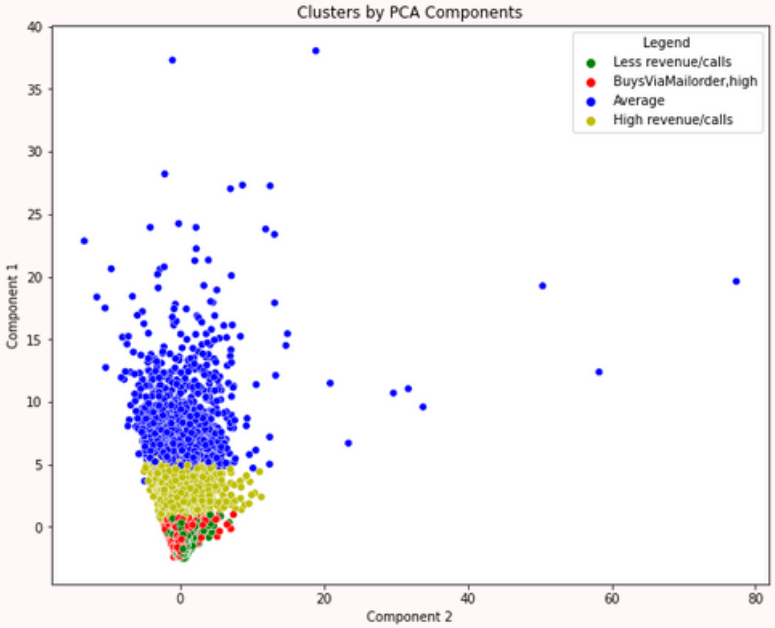
Gaussian Mixture



1. Import libraries
2. Import the dataset and read it
3. Check the data
4. Clean the data
5. Average it out
6. Check the outlier data (data which is far outside the average), and remove it
7. Then we need to check the factorability, which is can we find factors in the dataset? (this is similar to correlation)
8. We perfrom a Bartlett test to see the p value and if 0 the data is statistically significant, indicating that the observed correlation matrix is not an identity matrix
9. Then we do a KMO test to see the qualitive range of data we have
10. Then we choose the amount of factors using eigen values using a graph
11. We also see the variance of factors
12. Then we use a factor loading graph The factor loading is a matrix that shows the relationship of each variable to the underlying factor..
13. Then we see the variance of each factor
14. Then we create a data frame with the factors we found earlier (eg:10)
15. Then we use the GMM model but first we need to choose the amount of clusters using our data frame factors and "bic"
16. Then we output the result after clustering
17. Finally we label each cluster based off their standout feature

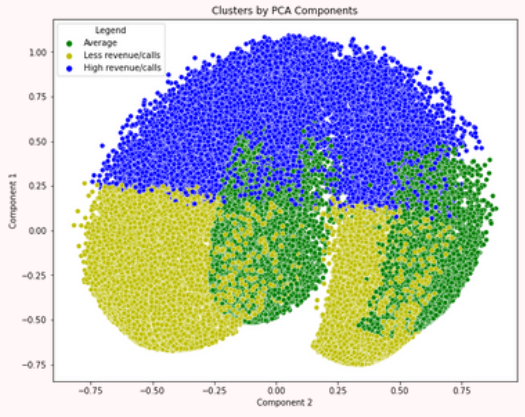
(TIP: To recheck the amount of clusters needed you can check your previous algorithms)

K-Means



1. Import libraries
2. Read the data set, and clean the data to see if there's any null dat
3. For further verification we make a correlation heatmap to see if th data is correlated enough to separate it
4. Then we have to standardize the data
5. Then we import PCA and fit PCA with our dataset
6. Now, Let's see the explained variance ratio by each component.
7. We now can plot the cumulative sum of explained variance.
8. From the plotted graph we see how many components would keep least 80% of the explained variance
9. Then the data is separated into thee components
10. Then we make a heatmap to see if the parameters in the component are related to eachtoher (ie: Income, Occupation)
11. Then we transform our data and store it into a variable
12. Then we have to implement the k means algorithm, for that then stored to each within clusters sum of squared value to WCSS list. The we plot a graph which shows the amount of clusters for the data, the we take the most optimal clusters using the "elbow" method, we us the optimal cluster for k means clustering
13. Then we see the old data set with new comps and labels
14. Then take the average of each of the clusters
15. We can also convert segment numbers to the label and see the observation number and proportions of each segment by total observation.
16. Then we plot the segmented data using PCA clustering with labe

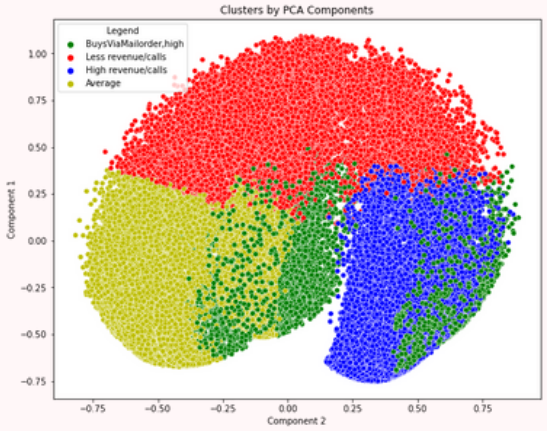
Mean Shift



1.Import libraries
2.Import the dataset and read
3.Check the data - standardize and normalize the data
4.Clean the data
5.Average it out
6.Use estimate bandwidth to find out the estimated number of clusters
7.Get the silhouette score and Davies Bouldin score for the number of clusters given by estimate bandwidth function
8.Fit the data to PCA to find the primary axes which will be used to graph and segment it
9.Finally label each cluster

Mini Batch K-Means

1.Import libraries
2Import the dataset and read it
3.Check the data - standardize and normalize the data
4.Clean the data
5.Average it out
6.Test the silhouette and Davies Bouldin score for several clusters and pick the best one
7.Fit the data to PCA to find the primary axes which will be used to graph and segment it



Results

Sillhouette and Davies Bouldin score

meanshift: silhouette = 0.229,Davies Bouldin = 1.629

gaussian mixture: silhouette =0.21 ,Davies Bouldin = 1.52

kmeans: silhouette = 0.25,Davies Bouldin = 1.425

minibatchkmeans: silhouette =0.21 ,Davies Bouldin = 1.9

Conclusion K Means is the best algorithm

Report

We were assigned to choose an algorithm to segment customers based on similar characteristics. After the dataset was selected (we chose the telecom industry), we tested this data into four different algorithms: K means, MiniBatchKmeans, Gaussian mixture, mean shift, the results were then plotted on a graph. In the code, we had included two scores silhouette and Davies Bouldin the most successful algorithm will have a higher silhouette and lower Davies Bouldin score. Based on our results we have concluded that K Means is the best algorithm for our selected dataset