Market Department Segmentation Using K-Means Clustering and PCA

S. Abishek, E.Dhanush ,B.Nikhila

G.Aditya

ABSTRACT:

This report addresses the use of clustering algorithms in the market(customer) segmentation to define a marketing strategy of a credit card company. Customer segmentation splits customers into groups based on common traits, which helps banks, corporations, and companies enhance their products and services.Principal Component Analysis (PCA) is a technique for identifying a company's client categories based on their credit card transaction history.

Introduction:

Marketing is critical to a company's growth and long-term viability. Marketers can aid in the development of a company's brand, customer engagement, revenue growth, and sales. Marketers help businesses expand by bringing in new customers. Marketers engage customers and understand their needs. Knowing and identifying clients' wants is one of the most difficult tasks for marketers. By understanding the customer, marketers can launch a targeted marketing campaign that is tailored for specific needsIn order for this campaign to be successful, the bank has to divide its customers into at least 3 distinctive groups. This process is known as "marketing segmentation" and is crucial for maximizing marketing campaign conversion rate. Credit Cards are a big business for any Bank. Issuing a Credit card is not only a Brand awareness strategy but a revenue-generating exercise too.

A revolving customer is one who just pays the minimum amount as their due so that he is not considered to have missed his Credit Card Payment. As a result, it is a significant opportunity for a bank to make money, but it must collaborate closely with its Risk and Strategy teams, which then collaborate with the Marketing team.

Problem survey:

Segmentation is an important aspect of developing marketing goals and strategies, and establishing those goals.

(a) an analysis of how products should be sold or manufactured, based on current client categories.

(b) the discovery of new segments as potential targets for existing products or new product development.

Because a company's resources are limited, segmentation is essential in determining how to best identify and service its customers. Effective segmentation allows a business to figure out which customer groups it should target and how to best position its products and services for each category.

DATASET(Dataset description):

You've been employed as a consultant by a bank in New York City in this case study. The bank has a lot of information on their customers over the previous six

months.We have been provided with extensive data on the bank's customers for the past 6 months. Data includes transactions frequency, amount, tenure... etc. The bank's marketing department wants to start a targeted ad campaign by segmenting their clients into at least 3 categories. If data about the customers is available, data mining algorithms can be applied to perform market segmentation.

In order to design a marketing plan, this instance requires the creation of a customer segmentation. The sample Dataset covers the usage patterns of around 9000 active credit card users over the previous six months. The file contains 18 behavioral variables at the customer level.

Following is the Data Dictionary for Credit Card dataset:-

CUST*ID* : Identification of Credit Card holder (Categorical)
BALANCE : Balance amount left in their account to make purchases (
BALANCEFREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
PURCHASES : Amount of purchases made from account
ONEOFF*PURCHASES* : Maximum purchase amount done in one-go
INSTALLMENTSPURCHASES : Amount of purchase done in instalment
CASH*ADVANCE* : Cash in advance given by the user
PURCHASESFREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in instalments are being done (1 = frequently done, 0 = not frequently done)
CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
CASHADVANCETRX : Number of Transactions made with "Cash in Advance"
PURCHASES*TRX* : Number of purchase transactions made
CREDITLIMIT : Limit of Credit Card for user
PAYMENTS : Amount of Payment done by user
MINIMUM_PAYMENTS : Minimum amount of payments made by user
PRCFULLPAYMENT : Percent of full payment paid by user
TENURE : Tenure of credit card service for user

PREPROCESSING:

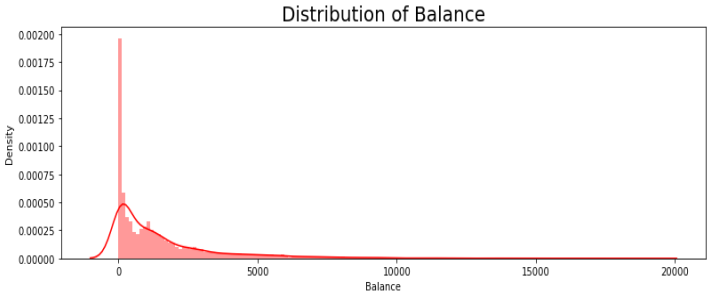
Importing necessary libraries such as pandas, NumPy, seaborn, matplotlib, scikit learn etc. Import the dataset into a dataframe using pandas and check out the head of the dataframe and find out how variables are related based on mean and standard deviation. Mean BALANCE is $1,564, BALANCE_FREQUENCY is frequently updated on average ~0.9, Average PURCHASES_FREQUENCY is around 0.5, Average ONEOFF PURCHASES FREQUENCY, Average CREDIT_LIMIT is ~$4,500, Average TENURE is 11.5 years.

The main purpose of data pre-processing is to find any null values in the dataset. We have calculated the percentage of data that is missing. There are two variables with missing data, namely CREDIT_LIMIT and MINIMUM_PAYMENTS. The missing values in these columns make up a insignificant percentage of the data set and can be safely deleted without risking a loss

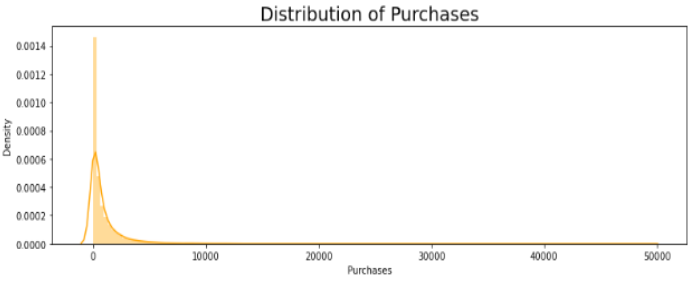
in data. The missing data in CREDIT_LIMIT make up less than 1% of the data and in MINIMUM_PAYMENTS only around 3%.\So the missing values are filled up with the mean of the MINIUM_PAYMENTS and mean of the CREDIT_LIMIT.

EDA:

From the Initial distribution plots of every variable in the dataframe we see that these distributions are very left skewed and there are a lot of zero values. From the distribution plot of Balance (total amount of money that you owe to your credit card company) range from 0 to 20,000 dollars. This could be owing to the fact that many people have "zero balance cards."The basic idea is that by keeping your balance low (in this case zero) but your credit limit high, this would increase your credit utilization ratio and in turn increase your credit score.



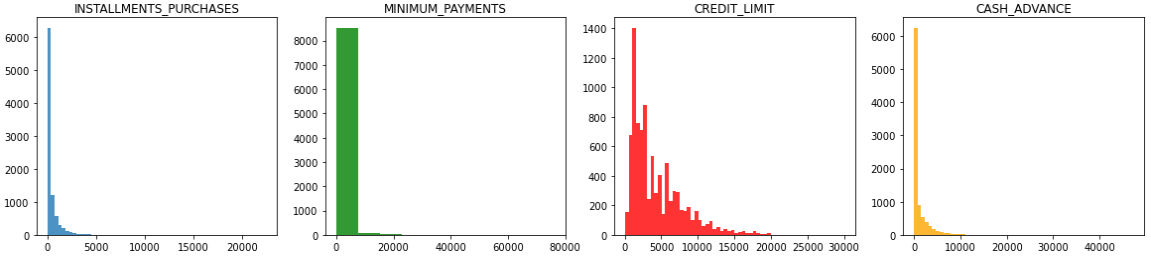
Balance frequency is the score, 0-1, given to accounts based on how frequent the account's balances are updated. One being the highest score. In a word, it's a measurement of how often consumers use their credit card. In the count plot of balance frequency most of the accounts have the score of one, the best score, meaning that most people do use credit cards frequently and only a small number of people keep their cards relatively inactive.



From the distribution plot of Purchases, A lot of people have the purchase amounts of 0 which make sense since earlier we also see that a lot of people are holding zero balance cards.

Starting with different purchase types. When the credit limit is increased, we may expect purchases to increase as well. There does not appear to be a correlation between credit limit and these variables.In fact, the amount tends to stay at zero while the credit limit increases. This again, shows that credit card users really do want to keep their balance low to utilize their credit. This might signify that, for most people, credit cards are tools for credit utilization rather than a spending device.

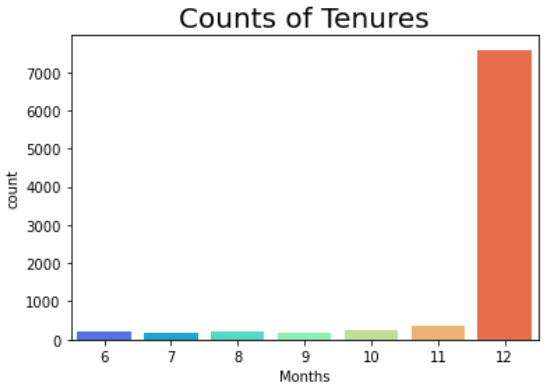
We have purchase frequency or how often do people make purchases. There seems to be the same number of people scoring low and high. Then again, we saw earlier, because of the high balance frequency, that



maybe people are just really good at paying off their credit cards.

Tenure:

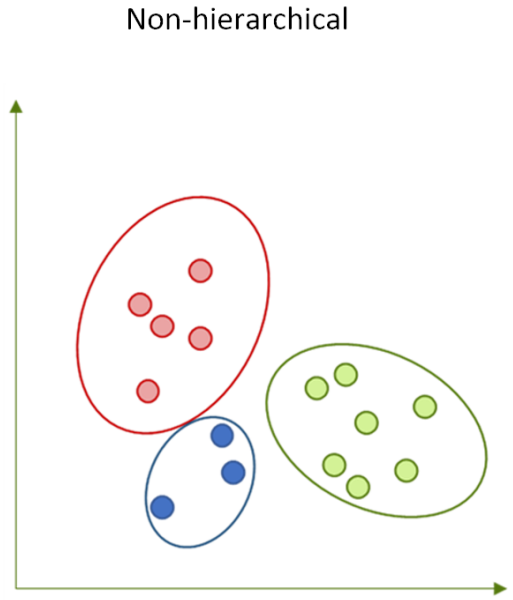
Tenure is the repayment period of the cards, ranging from 6-12 months. The majority of the cards are for a 12-month period. The longer your term, the higher the interest rate you pay, but the longer you have to pay it back, and this is the choice most people choose.



Model Building:

K-MEANS INTUITION

As we know that K-means is an unsupervised learning algorithm-means works by grouping some data points together (clustering) in an unsupervised fashion. The algorithm groups observations with similar attribute values together by measuring the Euclidean distance between points. Clustering is a technique in which similar objects are grouped to form clusters. These groups are called clusters as shown in figure 3. Many data mining projects use clustering methods for data prediction. Clustering is used in computer graphics, pattern recognition, data compression and image analysis. There are different clustering techniques present which are computed by different formulas for finding distance to create clusters. These are density-based clustering, centroid-based and distribution-based clustering.



In this research K-means clustering algorithm is used for building and testing Users credit card data. There are many distance measures, which are used to find clusters. In this research Euclidean distance measure is used. Formula for Euclidean distance measure as in the equation below.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Steps involved in the in K-Means:

1) n is taken as input from the user, which is used to create n number of clusters.

2) A random number of points is chosen as the Centroid in the dataset.

3) Then from points and centroid Euclidean distance is measured to place it in the nearest cluster.

4) The centroid for each cluster is re-positioned to get the correct centroid.

5) These steps are followed in an iterative manner until the centroid gets static. These steps ensure that all clusters which are formed are balanced.

Here we decided to use StandardScaler and not MinMaxscaler since our data tend to be skewed to the left and only a few

data points on the right and StandardScaler can handle outliers better.

THE "ELBOW METHOD" FOR DETERMINING THE APPROPRIATE NUMBER OF CLUSTERS (K)

We must identify the optimal number of clusters using the Elbow Method (K).

The elbow method is a heuristic method for interpreting and validating consistency in cluster analysis, with the goal of determining the optimal number of clusters in a dataset.

If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.

$Within\ Cluster\ Sum\ of\ Squares\ (WCSS) =$

Number of clusters (K) = 4

Here is the graph depicting the elbow method used to find the optimum number clusters using k mean analysis

The elbow graph is a useful tool for determining the number of clusters to utilise while separating data.The goal is to minimize inertia, the measure that determines the distance between each point and a central one during a clustering process.

We tried different numbers of clusters from 1-10 and then we graph inertia or wcss (within clusters sum square) against the cluster number. Inertia is basically how close the data points in the clusters are to the centers, which means the lower it is the more fitting the points are to their respective clusters. Here, we are trying to find the place where the wcss is as low as possible while still keeping the number of clusters as low as possible. Here the optimum number of clusters is 4 clusters since it is the place where the graph starts to flatten out meaning that having a higher number of clusters will not yield a much more fitting machine.

After applying K-means clustering algorithm using elbow method,

- First customers cluster (Transactors): Those are customers who pay least amount of interest charges and are careful with their money. Cluster with lowest balance, lowest cash advance, and percentage of full payment = 23%.
- Second customers cluster (Revolvers): who use credit cards as a loan (most lucrative sector): highest balance and cash advance, low purchase frequency, high cash advance frequency (0.5),a high number of cash advances (16) and a low proportion of complete payment (3 percent).
- Third customers cluster (VIP/Prime): high credit limit $16,000, and highest percentage of full payment, target for increased credit limit and increased spending habits.
- Fourth customers cluster (Rare): these are customers with low tenure (7 years), low balance.

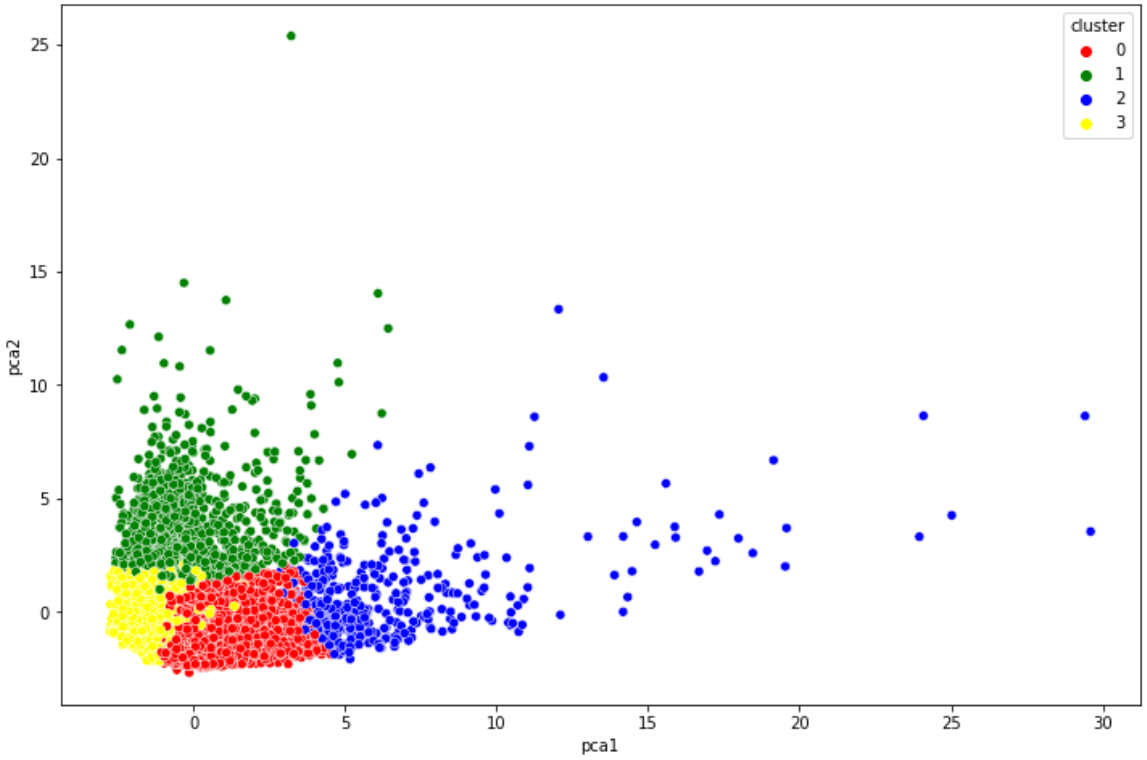
With four clusters, K-means is an effective tool for showing unique features of each cluster or customer group.

Principal Component Analysis (PCA):

PCA is an unsupervised machine learning algorithm. PCA performs dimensionality reductions while attempting to keep the original information unchanged. We took n_components = 2.PCA can analyse the

data to identify patterns in order to reduce the dimensions of the dataset with minimal loss of information.”High dimensionality" refers to a dataset with a large number of characteristics, which might lead to overfitting. The goal of dimensionality reduction is to lower the number of variables in our dataset as well as multicollinearity between them.This is often referred to as an elbow in the scree plot. PCA works by the following steps

1. Normalize the Data: If the features in the dataset are measured in different units, this is generally important.



2. Calculate the covariance matrix,

3. Compute the eigenvalues and eigenvectors,

4. Re-orient the data,

5. Biplots should be used to plot the data (PC1 against PC2).

Reducing the Correlation between Highly Correlated Variables

PCA has the advantage of reducing the strong correlation between correlated variables. Some variables in the original dataset have high correlations. For example, "PURCHASE and ONE-OFF PURCHASE" have a correlation of (0.92), while "PURCHASE FREQUENCY and PURCHASE_ INSTS FREQUENCY" have a correlation of (0.86). The correlation between all variables becomes smaller than after PCA is used (0.61)

Clustering Analysis:

Although PCA is mainly used for visualization and dimensionality reduction, it can also be employed in the clustering analysis, customer segmentation, and pattern recognition. PCA, unlike K-means, is not a direct solution for clustering, but it can be used to improve K-means clustering results by recognising more clusters than the K-means ideal number of clusters. As a result, we employ PCA to find additional clusters or groups in this work and compare the results to the K-means outputs. Instead of 17 variables, the PCA clustering technique uses five PCs. PCA was used to create four clusters from our data frame.

Since the procedure of PCA clustering produced 4 clusters, The PCA has

identified one more cluster or group of customers that was not discovered by the K-means clustering.As a result, we'll update the ideal K value and run another K-means clustering to check if the results improve with the new K value.

We looked into the use of clustering analysis in market segmentation and customer segmentation in this research.We identified active customers in order to apply proper marketing strategy towards them. We segmented the customers into four groups: Transactors, Revolvers, VIP/ Prime, Rare users.

CONCLUSION:

We conclude that for consumer segmentation, the K-means clustering approach is better. We also illustrated how Principal Component Analysis (PCA) can be utilised for dimension reduction and visualization of data as an unsupervised statistical technique. We demonstrated that PCA can analyse data to uncover patterns in order to minimise the dataset's dimensions with minimal information loss.In our dataset, we also demonstrated that PCA may successfully minimise multicollinearity between highly correlated variables.

A crucial step was properly pre-processing and scaling the data. In terms of modelling and clustering, perfect inspection and knowledge of the data are acknowledged to be quite crucial. If the features in the dataset are measured in various units, normalizing of the data should be considered in clustering analysis. In the case of the data applied in this research, K-means clustering seems to be the most appropriate fit.

More complicated models, on the other hand, could be researched in the future. As a credit card firm, we could spend more time on marketing efforts targeted at the proper people using the information from our cluster analysis. People in clusters 0 and 1 clearly have the financial means to spend, and because they already do, we can use their purchasing behaviours to improve our techniques for getting them to spend even more. The analysis also tells us that there is untapped potential in people from cluster 2 and 3. These people already have some balance but are not purchasing as much, with the right push we might be able to get them to use the card for spending and become important sources of revenue.

REFERENCES:

https://www.researchgate.net/publication/349094412_Incorporating_K-means_Hierarchical_Clustering_and_PCA_in_Customer_Segmentation

Elbow method (clustering) - Wikipedia

Elbow Method for optimal value of k in KMeans - GeeksforGeeks