Bank Customers Market Segmentation

POAi Lab Semester Project Report

Participants:

Muhammad Haris	F2022376040
Zaiba Saeed	F2022376008
Sarah Abbas	F2022376026

Project Description:

This project involves data cleaning, data analysis, and Machine learning. In this project, we are working with a bank to segment their customers and create at least three distinct groups of customers for marketing purposes. This will give them insights into their customers and credit card spending and help them realign their marketing strategies.

The Project is divided into several steps:

- **UNDERSTAND THE PROBLEM STATEMENT AND BUSINESS CASE**
- **LIBRARIES**
- **CLEAN, VISUALIZE AND EXPLORE DATASET**
- **UNDERSTAND THE THEORY AND INTUITION BEHIND K-MEANS**
- **LEARN HOW TO OBTAIN THE OPTIMAL NUMBER OF CLUSTERS (ELBOW METHOD)**
- **FIND THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD**
- **APPLY K-MEANS METHOD**
- **APPLY PRINCIPAL COMPONENT ANALYSIS AND VISUALIZE THE RESULTS**
- **HIERARCHICAL CLUSTERING, PRINCIPAL COMPONENT ANALYSIS AND VISUALIZE THE RESULTS**

Understand the Problem Statement and Business Case:

Problem Statement:

We need to **segment our customer base** to deliver targeted marketing campaigns. Machine learning algorithms can analyze and clean customer data to create these segments, leading to improved marketing ROI. This ill help them to launch targeted marketing campaign tailored to specific group of customers.

Business Case:

The bank marketing team would like to leverage Ai/ML to launch a targeted marketing ad campaign that is tailored to specific group of customers. In order for this campaign to be successful, the bank has to divide its customers into atleast 4 distinctive groups. This process is known as “marketing segmentation” and it is crucial for maximizing the marketing campaign conversion rate.

Dataset Selection:

The dataset used for this purpose has 8950 instances of customer credit card details. The dataset has following columns:

```
CUSTID: Identification of Credit Card holder

BALANCE: Balance amount left in customer's account to make purchases
```

BALANCE_FREQUENCY: 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

ONEOFFPURCHASES: Maximum purchase amount done in one-go

INSTALLMENTS_PURCHASES: Amount of purchase done in installment

CASH_ADVANCE: Cash in advance given by the user

PURCHASES_FREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)

ONEOFF_PURCHASES_FREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)

PURCHASES_INSTALLMENTS_FREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)

CASH_ADVANCE_FREQUENCY: How frequently the cash in advance being paid

CASH_ADVANCE_TRX: Number of Transactions made with "Cash in Advance"

PURCHASES_TRX: Number of purchase transactions made

CREDIT_LIMIT: Limit of Credit Card for user

PAYMENTS: Amount of Payment done by user

MINIMUM_PAYMENTS: Minimum amount of payments made by user

PRC_FULL_PAYMENT: Percent of full payment paid by user

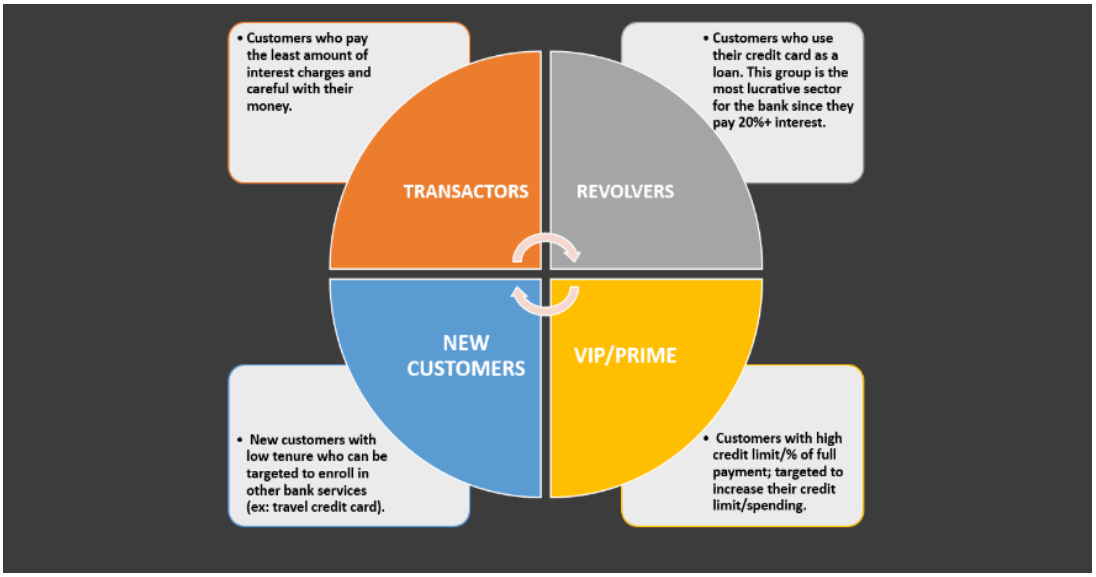
TENURE: Tenure of credit card service for user

Data Source: https://www.kaggle.com/arjunbhasin2013/ccdata

Model Selection:

We will be using clustering algorithms to cluster customers among distinct categories.

Our goal is to divide customers among 4 distinct groups as shown in image.



LIBRARIES:

Libraries used in this project include:

- Pandas
- Numpy
- Matplotlib
- Scikit-learn
- Seaborn
- Jupyterthemes

CLEAN, VISUALIZE AND EXPLORE DATASET:

We will read CSV(Dataset) from drive, providing colab access to Google Drive. Then we will go through some steps given below;

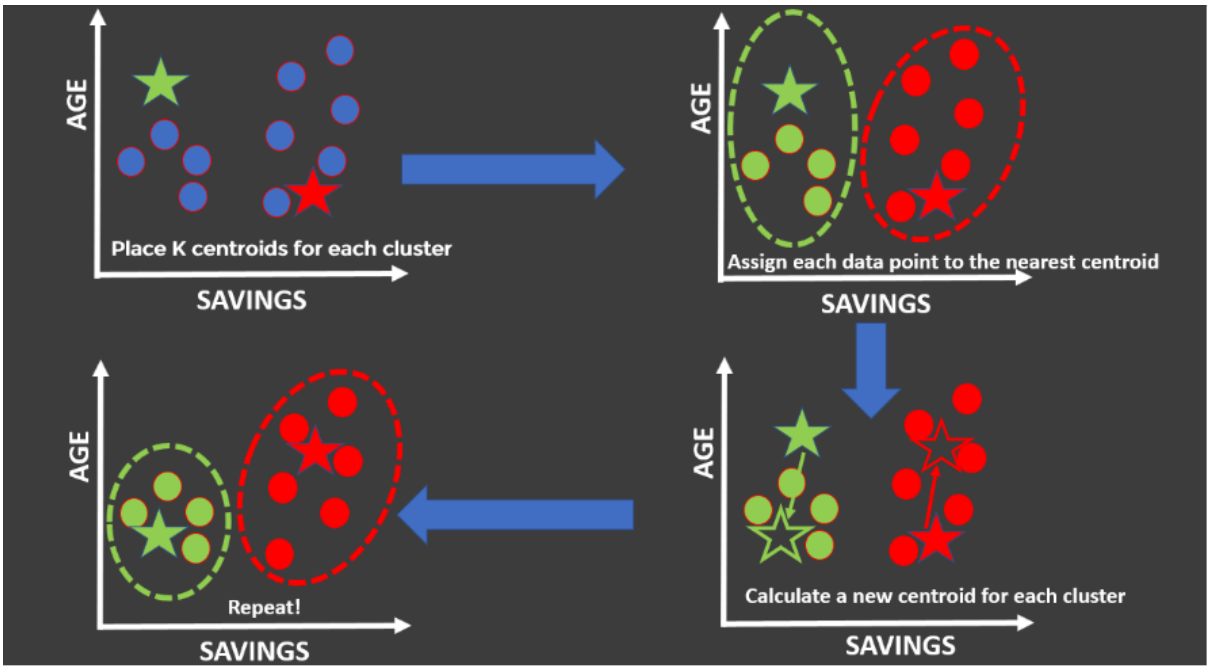
- getting information about columns of the dataset and using (.describe()) and (.info()) command
- Then check if the dataset has any missing values using Seaborn and pandas library.
- If there are any missing values, we will fill those values with the mean value. These mean values will not affect the data interpretation.
- Make sure there are no duplicate values.
- Drop the customer ID column as well because it is of no use.
- Visualize the kernel density estimation plot. Through KDE, we can visualize the distribution of observations in a dataset analogous to the histogram using Seaborn and Matplotlib library. KDE demonstrates probability density at different values in continuous variables.
- Visualize a heatmap of correlations between features in the credit card dataset. This helps us understand strong positive and negative relationships between features in the credit card dataset.

UNDERSTAND THE THEORY AND INTUITION BEHIND K-MEANS:

K-means is an unsupervised learning algorithm(clustering). K-means works by grouping some data points in an unsupervised fashion. The algorithm groups observations with similar attribute values together by measuring the Euclidian distance between points.

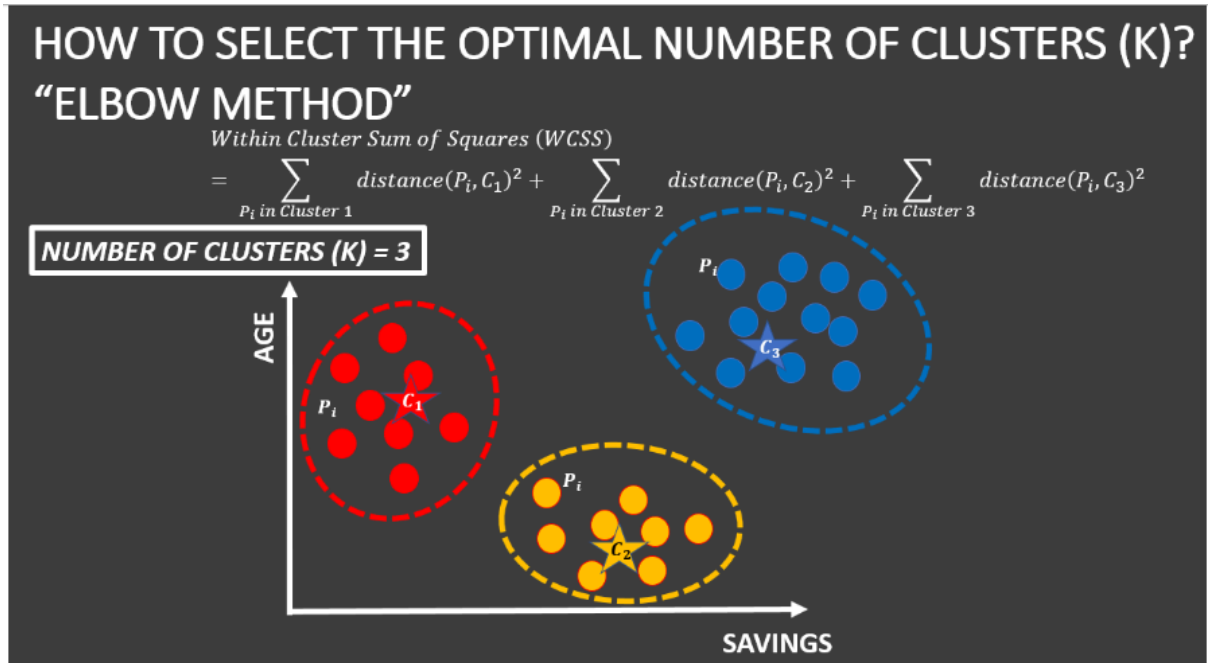
K-mean Algorithm Steps:

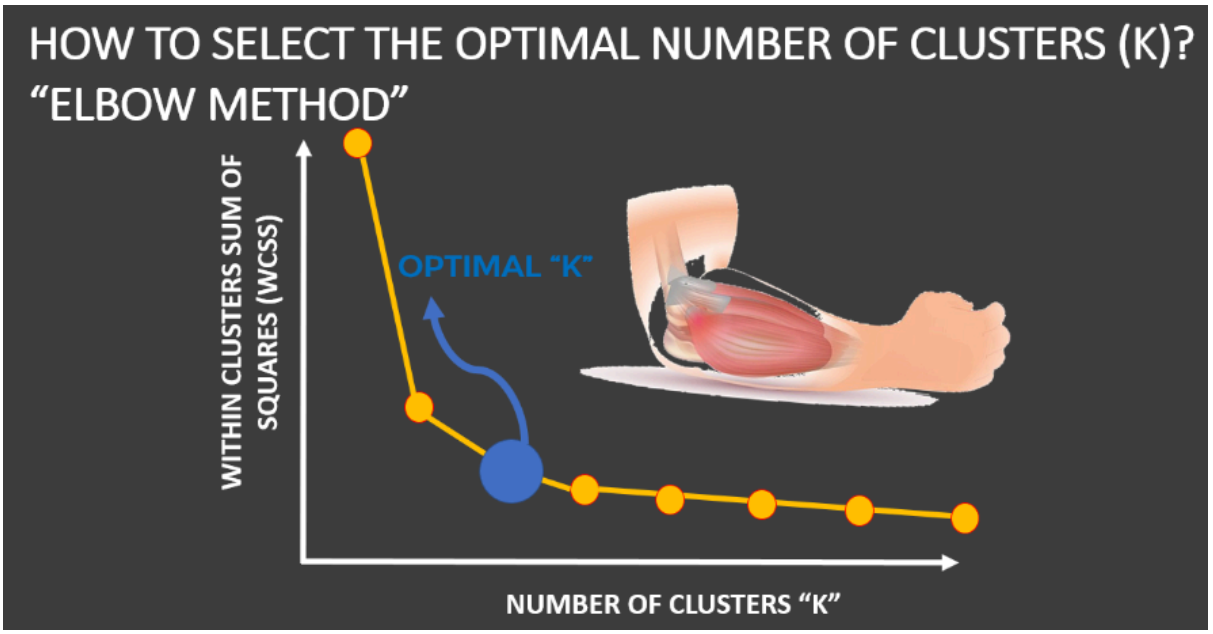
- Choose number of clusters ‘K’
- Select random K points that are going to be the centroids for each cluster
- Assign each data point to the nearest centroid, doing so will enable us to create ‘K’ number of cluster
- Calculate a new centroid for each cluster
- Reassign each data point to the new closest centroid
- Go to step 4 and repeat



FIND THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD:

The elbow method is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help find an appropriate number of clusters in a dataset. It is not always the most accurate method, especially for complex datasets or when the underlying cluster structure is not well-defined. while visualizing, If the line chart looks like an arm, then the ‘elbow’ on the arm is a value of k that is the best.





Steps:

- scaling data
- Then analyze the number of columns we want to use to visualize(This process is important as it will reduce the biases and increase accuracy)
- visualizing scaled data to analyze the optimal number of cluster

APPLY K-MEANS METHOD:

Although we intended for 4 groups of customers as explained in the data set. However, through the elbow method, we found that we could have 5 distinct groups of customers.

1. First Customer Cluster (Transactors): Those are customers who pay the least amount of interest charges and are careful with their money.
2. Second Customer Cluster(Revolvers): who use credit card as a loan (most lucrative sector)
3. Third Customer Clustere (VIP/Prime): High credit limit of $16k and highest percentage of full payments.
4. Fourth Customer Cluster (Low tenure): these are customers with low tenure, low balance
5. Fifth Customer Cluster (Dangerous): These are customers with high credit limit but less full payments

Steps:

- performing K-means clustering based on 5 value for clusters, obtained through the elbow method
- Interpret numbers through inverse transformation
- concatenating cluster labels into the original data frame
- visualize clusters

APPLY PRINCIPAL COMPONENT ANALYSIS AND VISUALIZE THE RESULTS:

Principal Component Analysis is a dimensionality reduction technique commonly used in machine learning and data analysis. It helps simplify complex datasets by identifying the most important underlying features that capture the majority of the data's variance.

It works by trying to find a new set of features called components. Components are composites of the uncorrelated given input features

Steps:

- Obtain principal components
- Creating a dataframe with the two components
- Concatenate the cluster labels to the dataframe
- Visualize

HIERARCHICAL CLUSTERING, PRINCIPAL COMPONENT ANALYSIS AND VISUALIZE THE RESULTS:

Hierarchical Clustering:

Hierarchical Clustering creates a hierarchy of clusters, where similar data points are grouped together at the lower levels of the hierarchy, and these groups are then further grouped together at higher levels.

Difference to K-Means:

K-Means pre-specify the number of clusters before running the algorithm. We used the elbow method to get the optimal number of clusters. Hierarchical Clustering uses dendrogram for optimal number of clusters.

Two Main Approaches:

Hierarchical Clustering uses two main approaches; agglomerative(bottom-up) and Division(top-down). We used agglomerative approach.

Steps:

Steps are almost the same, as we performed for K-means except we use dendrogram.

- Importing required libraries
- Plotting Dendrogram
- Defining clustering object with desired linkage method
- Adding Cluster labels to the dataframe
- Inverse transform clusters to visualize
- Plotting histogram of clusters
- Apply PCA and visualize the results

_____End_____
__