Report on Customer Segmentation to Define Marketing Strategy for a Bank

About the Project

In this project, I will be performing an unsupervised clustering of data on the behaviour of about 9000 active credit card holders in a bank during the last 6 months. Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster. I will divide customers into segments to develop a targeted marketing strategy to customers with similar characteristics.

Data Description

The Dataset used summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.

Source of Data: Kaggle.com

Following is the Data Dictionary for Credit Card dataset:-

CUSTID: 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

ONEOFFPURCHASES: Maximum purchase amount done in one-go

INSTALLMENTSPURCHASES: Amount of purchase done in installment

CASHADVANCE: 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 installments 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 Advanced"

PURCHASESTRX: 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

Objective

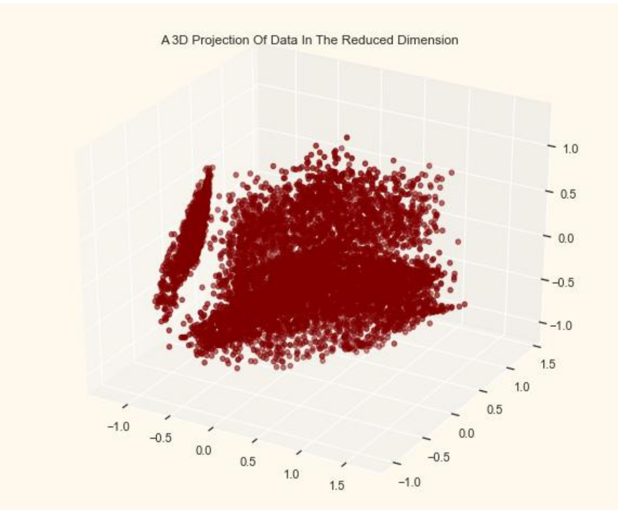
The main objective of this project is to develop a customer segmentation to define targeted marketing strategy for Credit card users.

Stages in Project

Below are the stages taken in this project:

1. Data Acquizition
2. Data Cleaning and Exploration.
3. Feature Engineering (
4. Dimensionality Reduction
5. Clustering (Kmeans and Agglomerative Clustering model)
6. Evaluating Models
7. Key Findings
8. Conclusion.

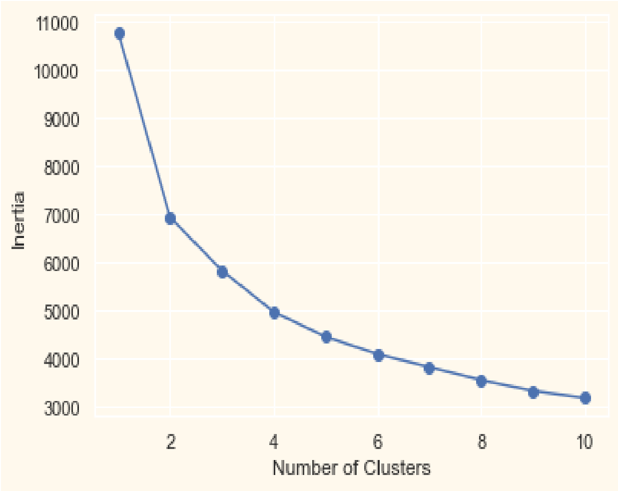
Dimensionality Reduction



Principle component analysis (PCA) was used to reduce the data to 3 features and the 3D projection of the Data in reduced dimension was visualized as shown.

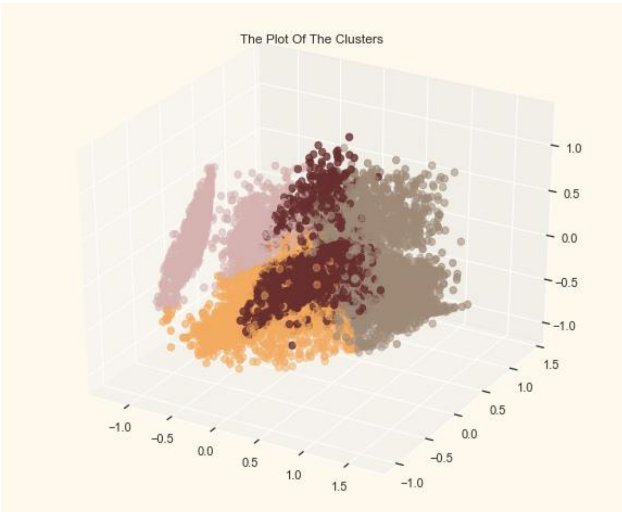
3D Projection of Data

Number of Clusters Using Elbow Method



Elbow method was used to determine the best number of clusters for unsupervised learning models. Inertia was plotted against Number of clusters and 4 was determined to be an optimal number of clusters for the data.

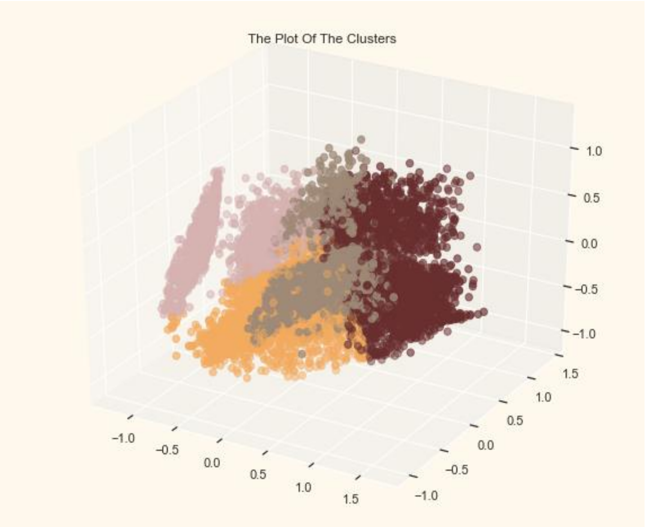
Kmeans Model



The PCA reduced data was modeled using the KMeans algorithm with 4 clusters and the clusters were visualized in 3D as shown.

Distribution of Cluster for Kmeans Model

Agglomerative Clustering model



The PCA reduced data was modeled using the Agglomerative clustering algorithm with 4 clusters and the clusters were visualized in 3D as shown.

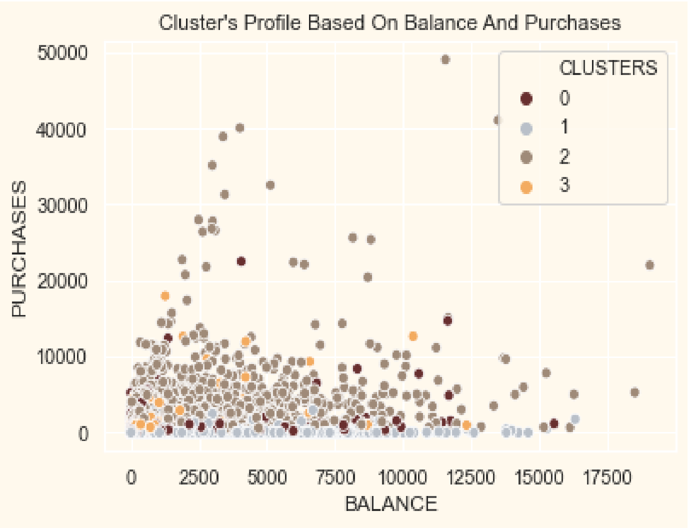
Distribution of clusters for Agglomerative clustering

The Kmeans model was selected to be a better model and used for further analysis.

Model Evaluation

The clusters seems to be fairly clustered with Cluster 1 having more data in it

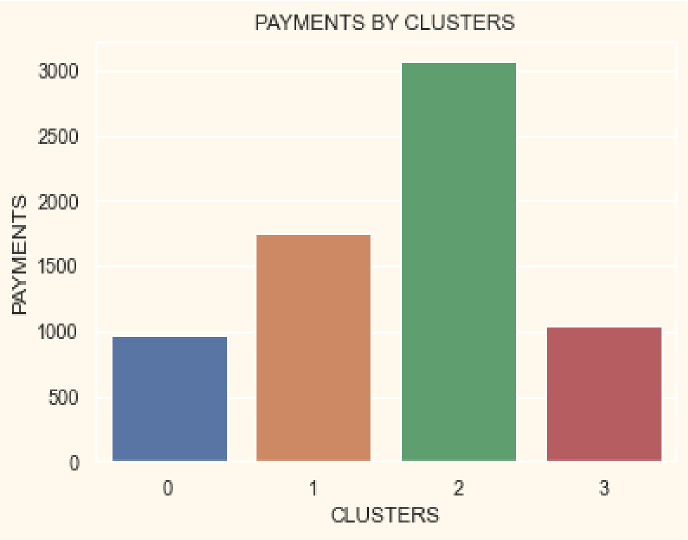
Cluster's Profile Based on Balance and Purchases



Purchases vs Balance plot shows the following

- ❖ Group 0: Low/Moderate Balance and low purchase
- ❖ Group 1: Low Balance and Low Purchase
- ❖ Group 2: High Balance and High Purchase
- ❖ Group 3: Low Balance and Low Purchase

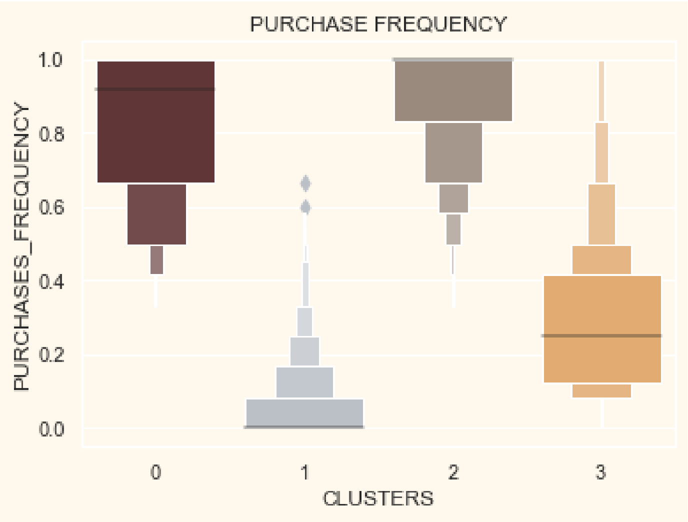
Average Payments by Clusters



This shows the average credit card payments made by customers in each cluster. Clearly customers in Cluster B made the most payment with the credit card.

- ❖ Group 0: Low Payment
- ❖ Group 1: Moderate payment
- ❖ Group 2: High Payment
- ❖ Group 3: Low Payment

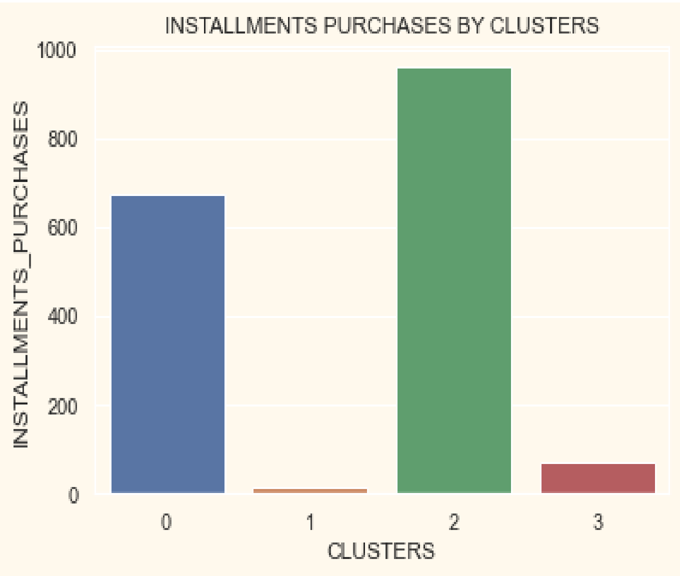
Purchase Frequency by Clusters



Cluster 2 has the highest purchase frequency averagely

❖ Group 0: High Purchase frequency
❖ Group 1: Low Purchase frequency
❖ Group 2: High Purchase frequency
❖ Group 3: Moderate Purchase frequency

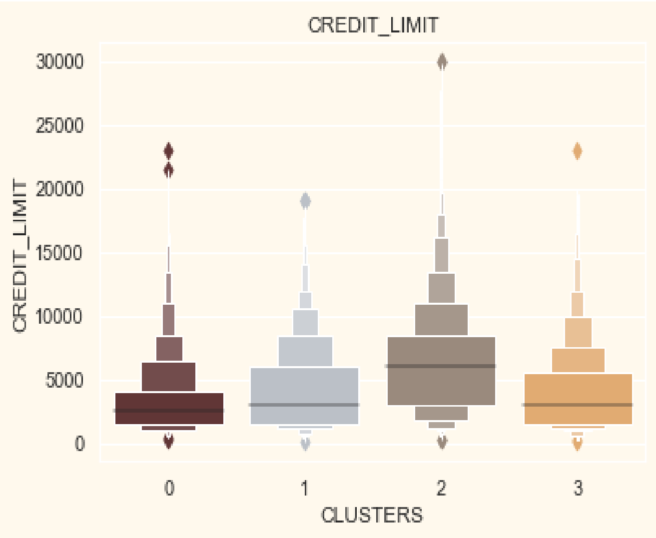
Installment Purchases by Clusters



Cluster 2 has made the highest Installment purchase averagely

❖ Group 0: Moderate Installment purchases
❖ Group 1: Very low Installment purchases
❖ Group 2: Very high Installment purchases
❖ Group 3: Low Installment purchases

Credit Limits by Clusters



Cluster 2 have the customers with highest credit limits

- Group 0: Low credit limit
- Group 1: Low credit limit
- Group 2: Moderate credit limit
- Group 3: Low credit

Summary Properties of Clusters/Key Findings

Group 0:

- Moderate balance and low purchase
- Low Payments
- High purchase frequency
- Moderate Installment purchases
- Low credit limit

Group 1

- Low Balance and Low Purchase
- Moderate payment
- Low Purchase frequency
- Very low Installment purchases
- Low credit limit

Group 2

- High Balance and High Purchase
- High Payment
- High Purchase frequency
- Very high Installment purchases

- Moderate credit limit

Group 3

- Low Balance and Low Purchase
- Low Payment
- Moderate Purchase frequency
- Low Installment purchases
- Low credit

Conclusion

With the help of clustering algorithms, credit card users have been clustered into 4 groups. Group 2 seems to be the most import cluster, however this depends on the purpose of the marketing and target audience.

A major flaw in the clustering model is that data were not distinctively separated as their clusters still having similar characteristics.

This can be improved by increasing the number of dimensionally reduced features. Also, there seems to be presence of a few outliers in the model dataset which could have affected the model results too. Further data cleaning could produce an improved model.