Final Project Report

WEEK-13

Group Name: DataThor
Name: Mohammad Shafiqul Islam
Email: si_shuvo95@yahoo.com
Country: Bangladesh
University: TU Dortmund University
Specialization: Data Science

Agenda

Problem Statement

Approach

EDA & Summary

EDA Recommendations

Model Building

Model Selection

Performance Metrics

Final Recommendations

Problem Statement

- XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category).
- Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 group as this will be inefficient for their campaign.

Approach

EDA has been done using several steps:

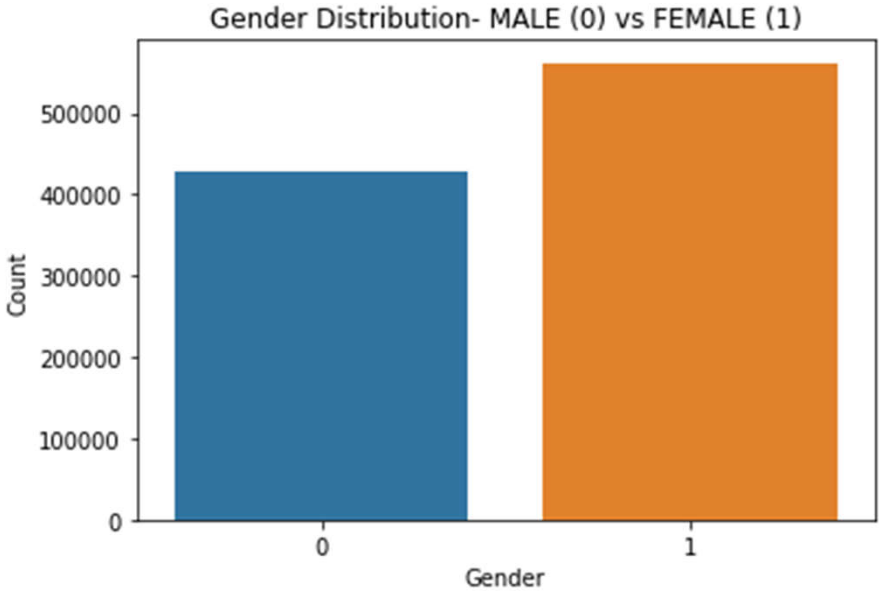
- Examine the Datasets to find the best possible dataframe.
- Visualizing different insights to find out the key factors.
- Proposed modelling techniques for customer segmentation.
- Recommendations.

Data Glacier
Your Deep Learning Partner

Data Exploration

- There was only one file: cust_seg.csv
- The dataset was in CSV format.
- Originally
  - Total Features: 48
  - Total Rows: 1000000
- The columns which has no impact are removed.
- There were some random unexpected values like NA, -999999, former customers (P). They are removed from the dataframe.
- After Data Cleansing:
  - Total Features: 35
  - Total Rows: 989173

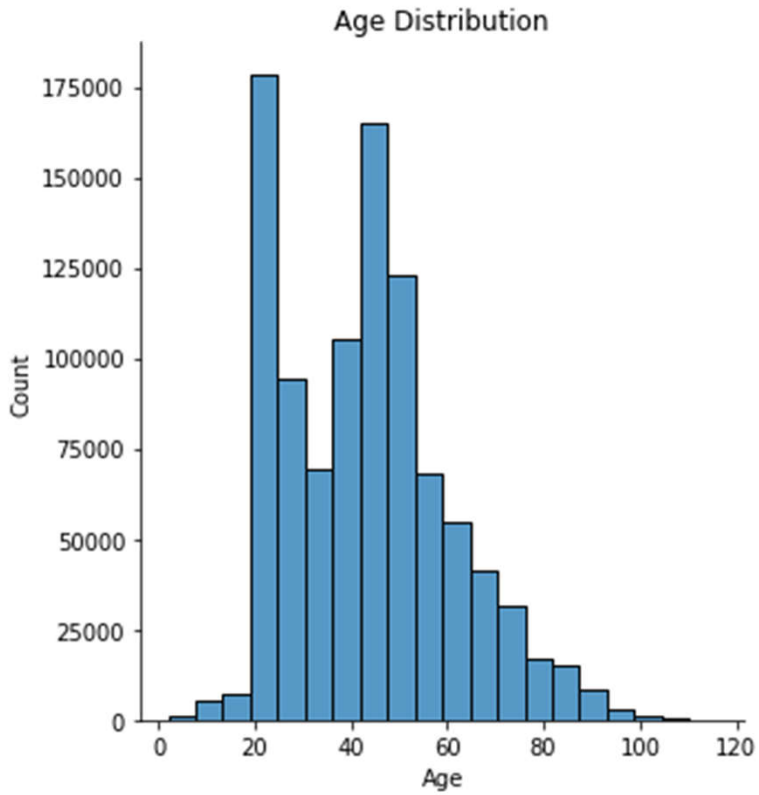
Distribution of Different Gender



Female customers have the higher ratio when it comes to the number of bank customer.

Data Glacier
Your Deep Learning Partner

Distribution of Different Age



Most of the Customers are middle aged (between 20-50).

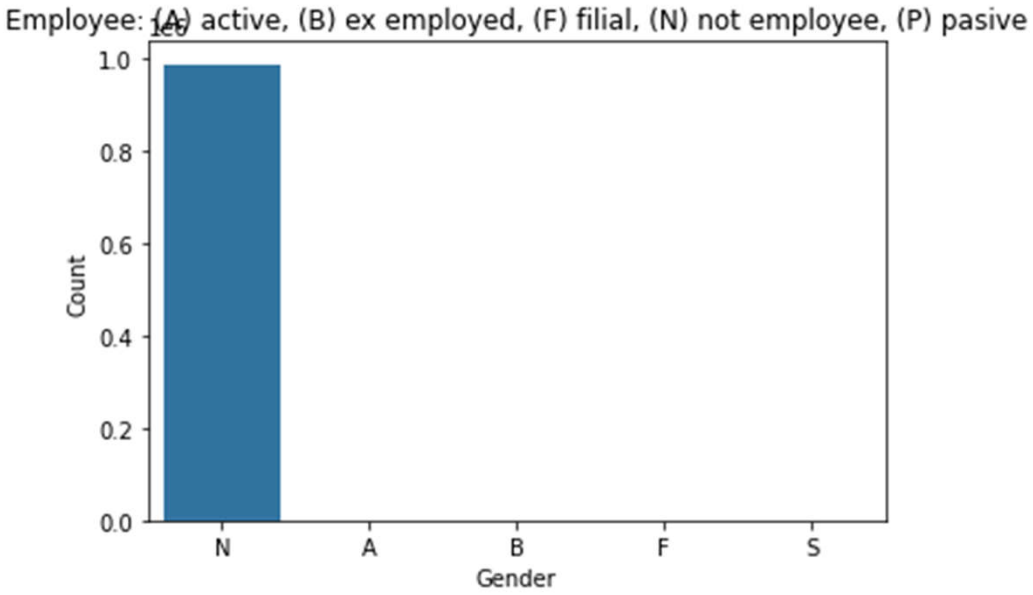
Data Glacier
Your Deep Learning Partner

Location Based Distribution

Location	Customer Count
ES	982260
FR	546
AR	542
DE	487
GB	480
...	...
SL	2
TG	2
CD	2
GE	2
AL	1

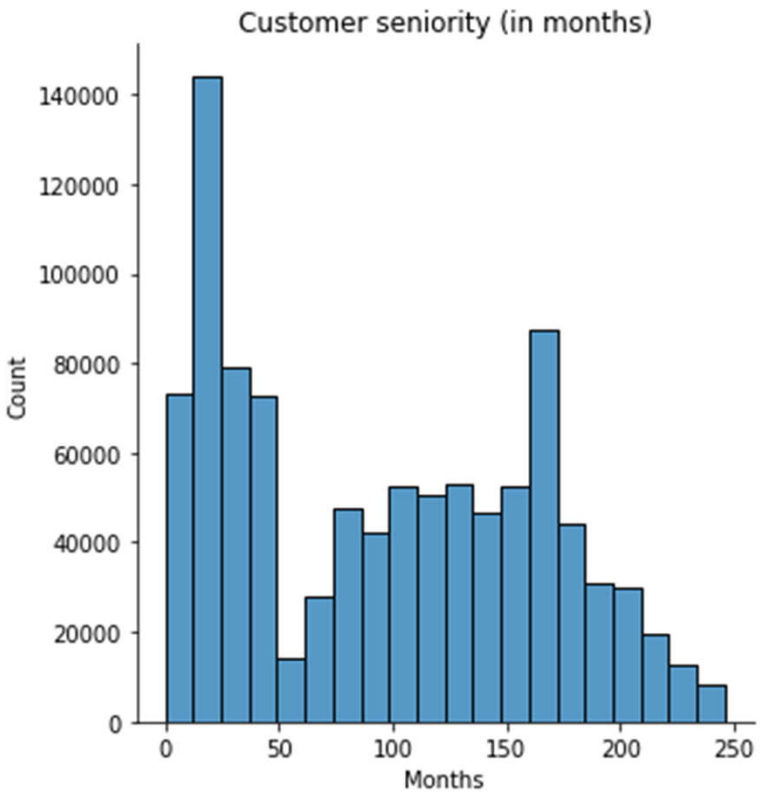
Clearly, location has not any effect as most of the customers are from ES. So, this feature is dropped.

Employee Status Count



All the customers' status are "not employee (N)". So, there is no relation between the customers who also working in the bank. For this reason, this feature is dropped.

Distribution of Customer Seniority



Most of the customers are new (0-50 months).

Distribution of Being Primary Customer

Status (1 = Primary or 99 = Primary during the month but not at end of the month)	Count
1	988113
99	1101

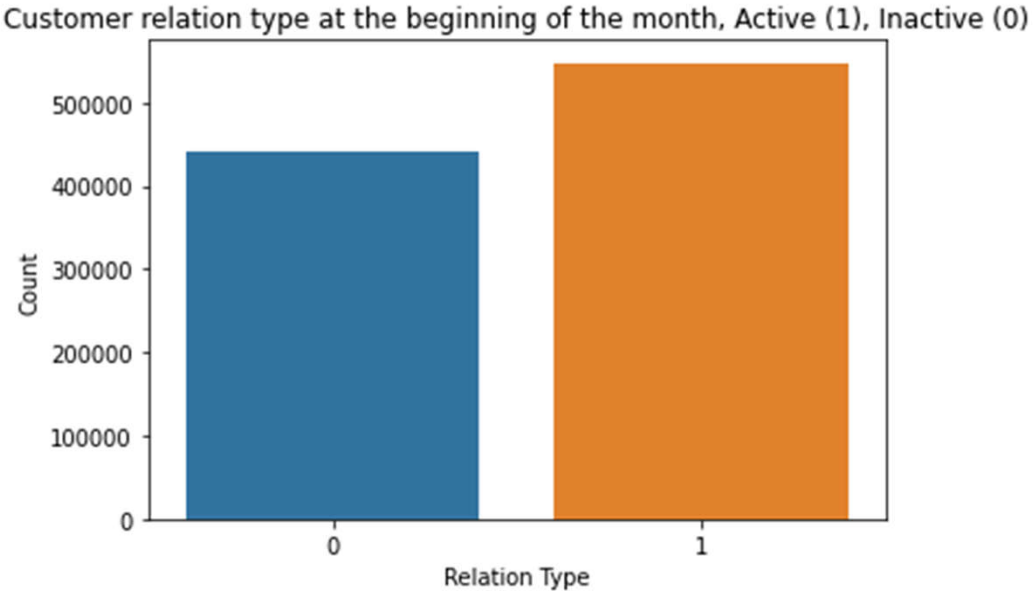
Most of customers are the primary customers.

Customer Relation Type (at the beginning of the Month)

Status (1 = Primary customer, 2 = co-owner, P = Potential, 3 = former primary, 4 = former co-owner)	Count
1	989171
2	2
3	41

Most of customers are the primary customers at the beginning of the month. So, this also has no effect.

Customer Relation Distribution



Both active and inactive customers are there, although the number of active customer is higher.

Data Glacier
Your Deep Learning Partner

Residence Status

Status (S = Same as Bank, N = Different to the Bank)	Count
S	982219
N	6954

Most of customers' residence status is same as the bank.

Birth Country Status

Status (S = Foreigner, N = Native)	Count
S	946283
N	42890

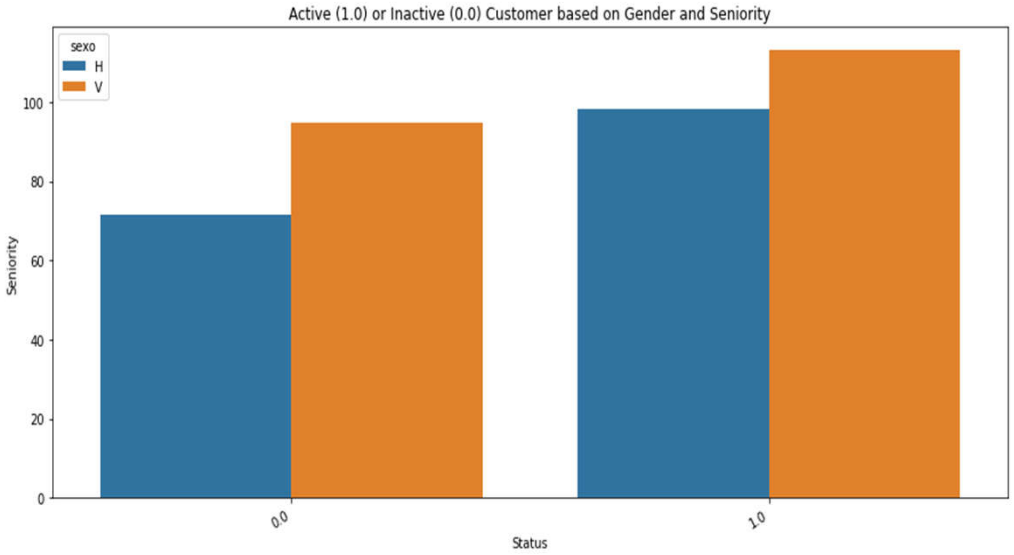
Most of customers are native.

Spouse Status

Status (S = Employee of the Bank, N = Not an employee)	Count
S	2
N	989171

Almost all the customers' spouse are not an employee of the bank. So this feature is removed.

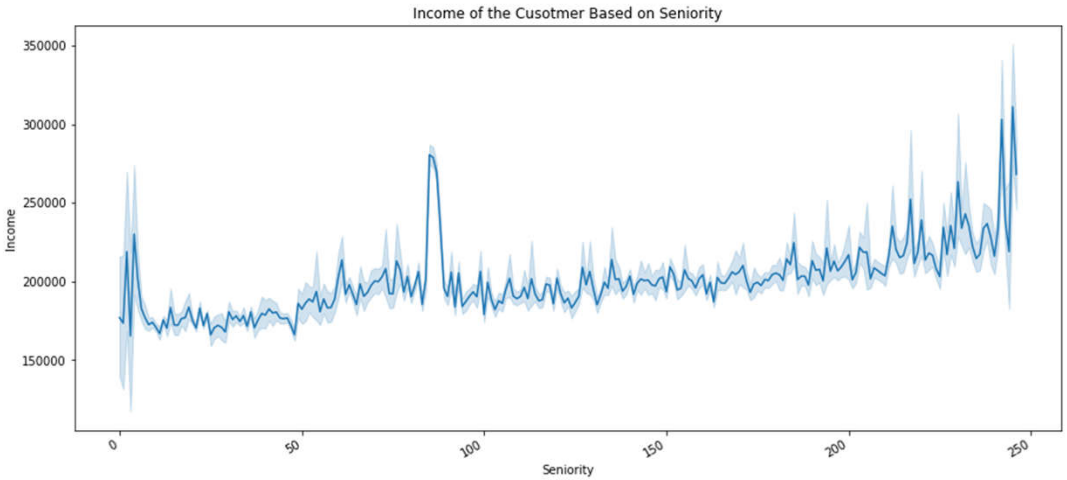
Gender and Seniority Effect on Customer Relation



Mostly, senior customers are active. Also, female customer has the higher ration in terms of gender.

Data Glacier
Your Deep Learning Partner

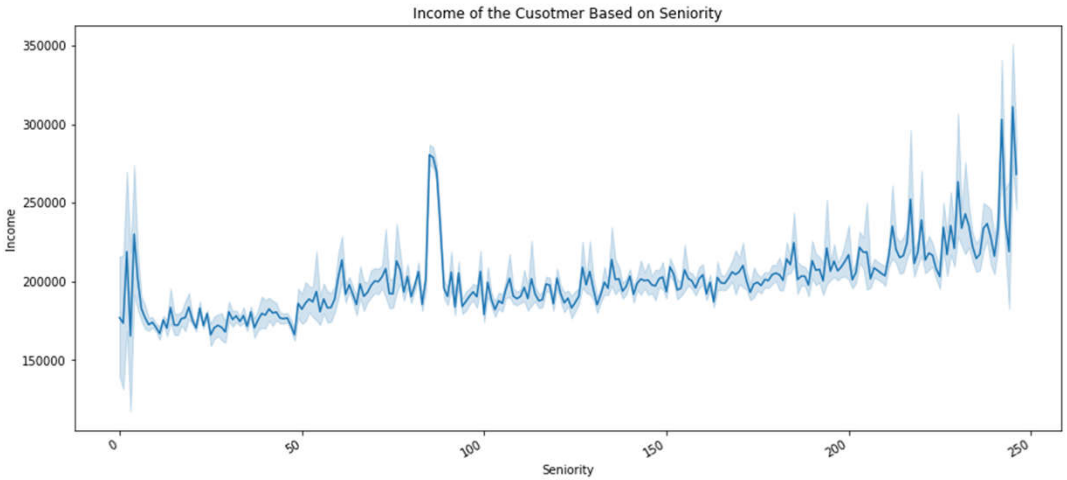
Income Distribution of the Customers Based on Seniority



Senior customers has the higher income ratio.

Data Glacier
Your Deep Learning Partner

Income Distribution of the Customers Based on Seniority



Senior customers has the higher income ratio.

Data Glacier
Your Deep Learning Partner

Other Factors

- There are some other Yes/No factors such as
  - Account type
  - Deposit type
  - Funds
  - Mortgage
  - Pensions
  - Loans
  - Taxes
  - Card

Using a statistical method (KDE: Kernel Density Estimation), we found most of columns have the ‘No’ value.

Correlations

- Small positive correlation (0.57) between age and seniority.
- Very strong positive correlation (0.81) between the customers who are active at the beginning of the months and current active customers.
- Very strong positive correlation between Payroll account holders who uses payrolls (0.76), pensions (0.81), and small correlation who uses debit cards (0.56).

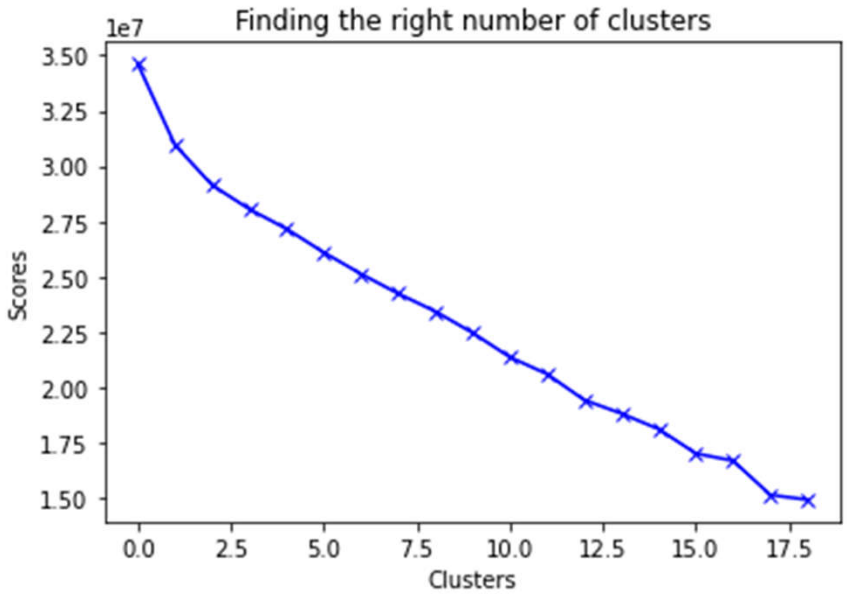
EDA Recommendations

- From the analysis, it is found that female customers are more active. So, the bank can invest offer in them.
- Most of the customers are middle-aged (20-50 years). So the bank can roll out the offer to young customers.
- Location has no effect. So, segmentation based on graphical location will not be right.
- Most of the customers' residence is the same as the bank's and most of the customers' are native. So banks should prioritize them.
- Senior customers are more active. So, the bank can prioritize the customer based on seniority.
- There is an increasing trend regarding the seniority and the income of senior customers. So, the bank should offer good value to senior customers.

Proposed Modelling Technique

- As we need to segment the customers into different cluster, we will use K-means clustering.
- K-means is an unsupervised clustering algorithm (Base).
- It works by grouping some data points together by using Euclidian distance between the points.
- K = Number of clusters/groups
- Auto Encoder for Dimensionality Reduction (Secondary Model).

K-Means Clustering



At first, we need to find the optimal number of cluster. Here, the elbow method is used and the number is 3 with a range of value from 1 to 20.

Data Glacier
Your Deep Learning Partner

Clusters (Base Model)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 0

- - Most Seniority (~133 months)
- - Active Customer
- - Low Income (~189000)
- - No Current Accounts
- - Payroll Account
- - Payrolls
- - Pensions
- - Direct Debit

Clusters (Base Model)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 1

- - Medium Seniority (~107 months)
- - Active Customer
- - High income (~195000)
- - Current Accounts
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Clusters (Base Model)

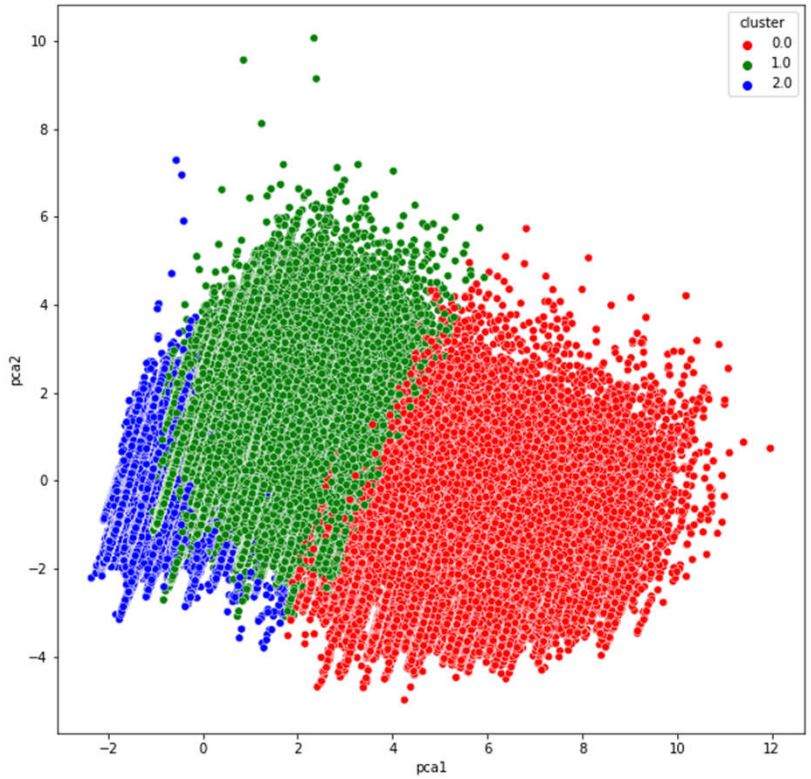
Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 2

- - Low Seniority (~78 months)
- - Inactive Customer
- - Low Income (~187000)
- - Current Accounts
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Plotting the Clusters (Base Model)



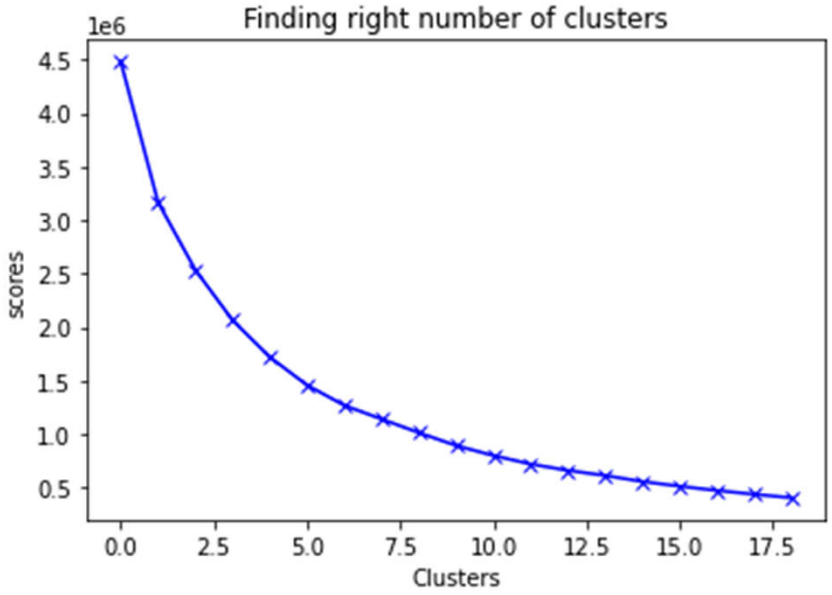
We applied PCA to find the clusters.

Data Glacier
Your Deep Learning Partner

Auto Encoder (Secondary Model)

- To perform better dimensionality reduction, we applied Auto Encoder.
- We reduced the number of column to 10.
- There are total 25 epochs which took almost 3 hours to train.
- Then, we find the optimal number of clusters with the range of 1 to 20.

K-Means Clustering (Auto Encoder)



This time the optimal number of clusters is 5.

Data Glacier
Your Deep Learning Partner

Clusters (Auto Encoder)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 0

- - Seniority (~78 months)
- - Inactive Customer
- - Low Income (~187000)
- - Current Accounts
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Clusters (Auto Encoder)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 1

- - Seniority (~107 months)
- - Active Customer
- - Medium income (~195000)
- - Current Accounts
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Clusters (Auto Encoder)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 2

- - Seniority (~133 months)
- - Active Customer
- - Low Income (~189000)
- - Current Accounts
- - Payroll Account
- - Payrolls
- - Pensions
- - Direct Debit

Clusters (Auto Encoder)

Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 3

- - Seniority (~4 months)
- - Active Customer
- - High Income (~201000)
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Clusters (Auto Encoder)

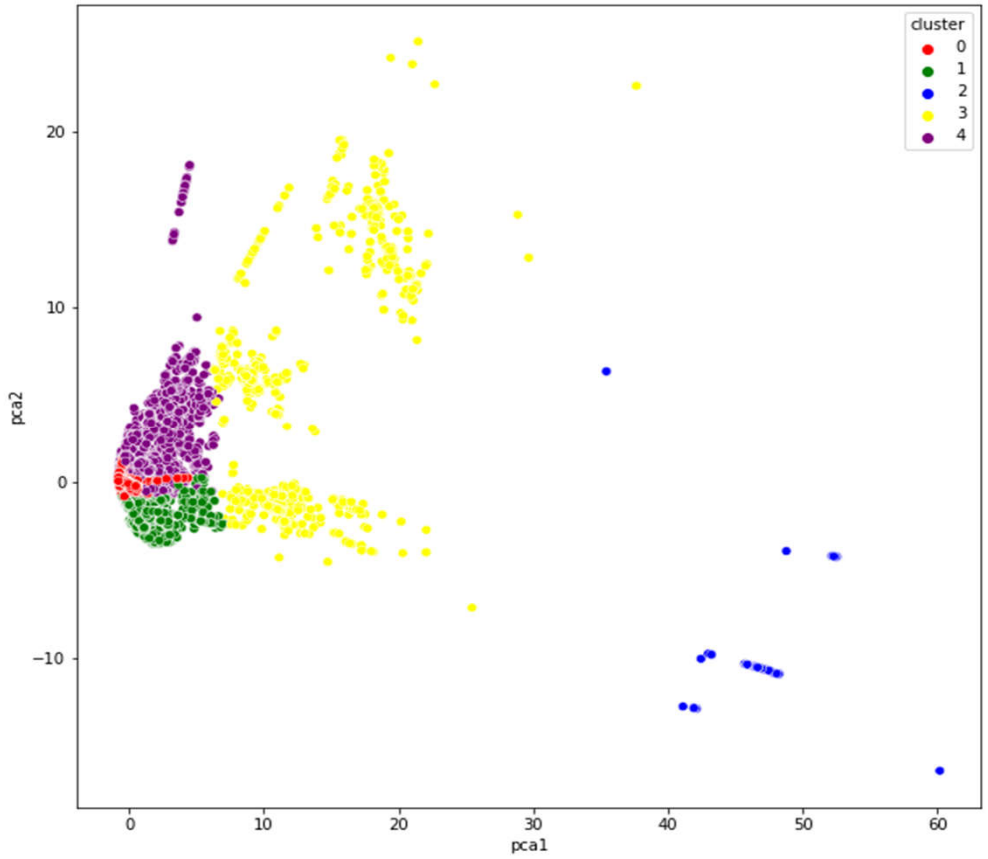
Data Glacier
Your Deep Learning Partner

After applying the K-Means clustering we have got the following attributes:

Cluster 4

- - Seniority (~161 months)
- - Active Customer
- - High Income (~203000)
- - No Payroll Account
- - No Payrolls
- - No Pensions
- - No Direct Debit

Plotting the Clusters (Auto Encoder)



We again applied PCA to find the clusters.

Data Glacier
Your Deep Learning Partner

Recommendations

- From the base model, Cluster 0 can be considered as the most efficient customers where Cluster 2 is less efficient.
- In every clusters, the main factor was the seniority.
- From the model with Auto Encoder, Cluster 4 is most efficient customers with highest income and seniority.
- The main factors are seniority and incomes.

Data Glacier
Your Deep Learning Partner

Thank You