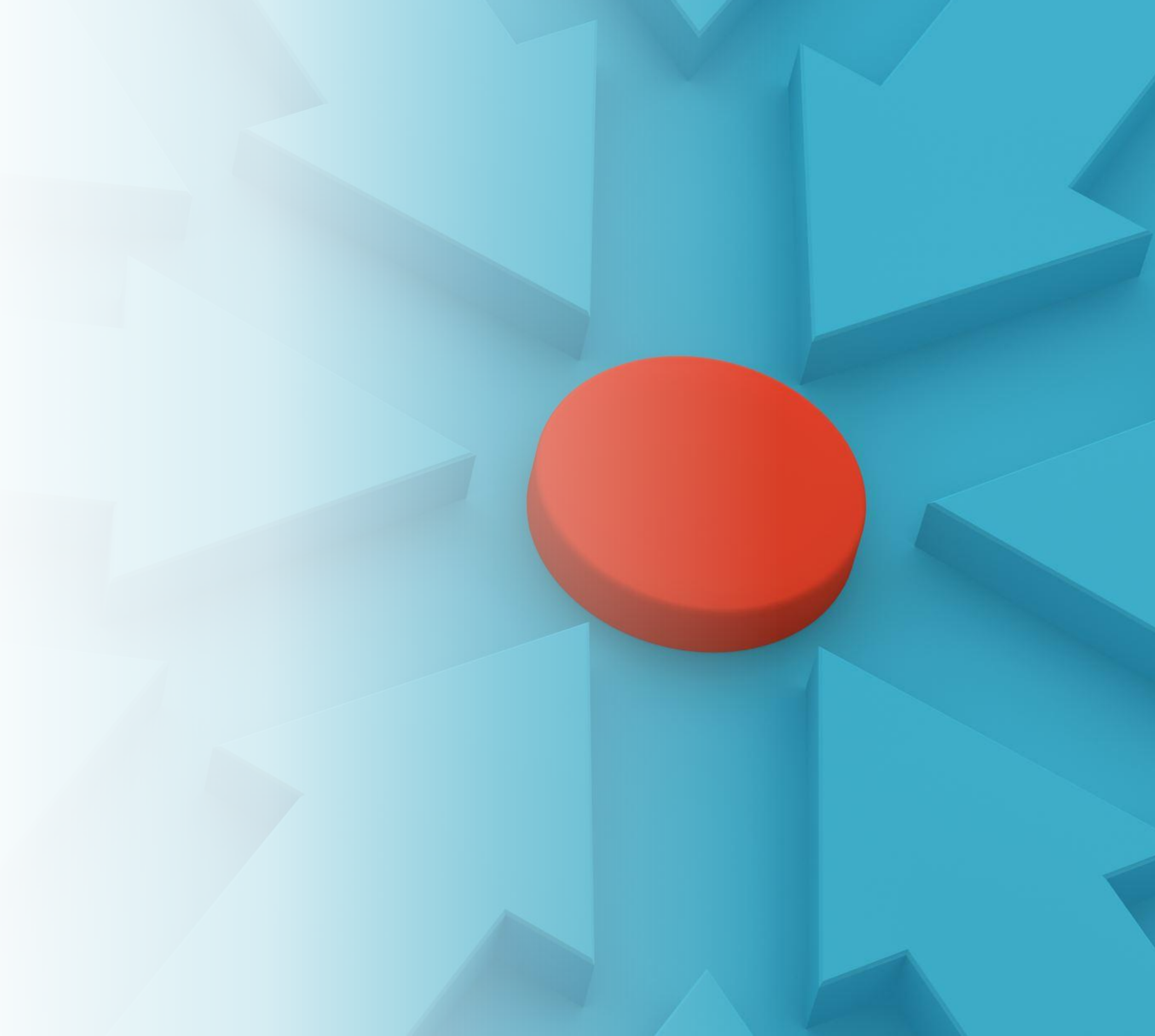
Data Glacier

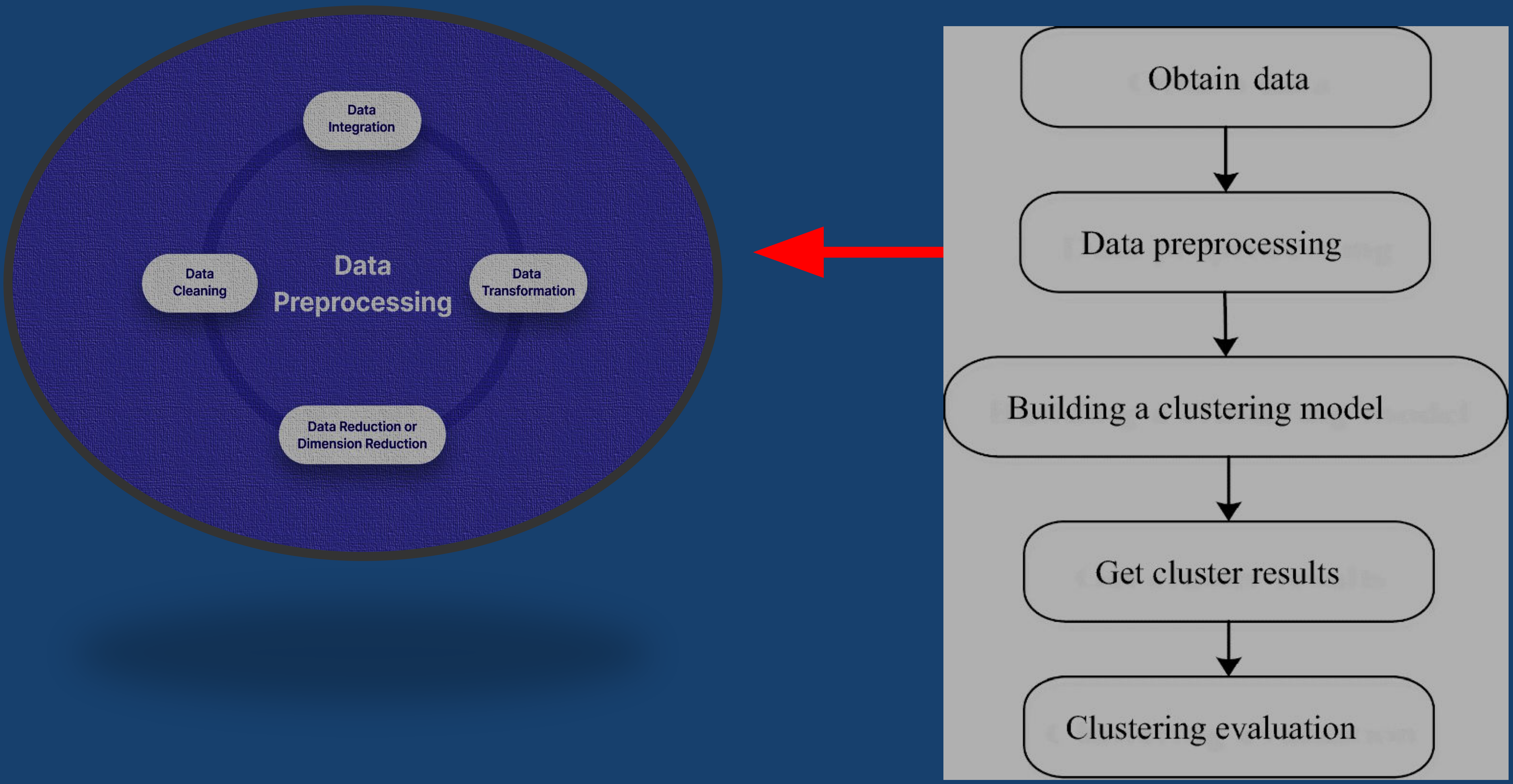
Your Deep Learning Partner

Customer Segmentation

11/30/2022



Segmentation Process



Data Preprocessing

Data Integration

Data Transformation

Data Reduction or Dimension Reduction

Data Cleaning

Obtain data

Data preprocessing

Building a clustering model

Get cluster results

Clustering evaluation

Problem Statement

Motivation:

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.

Objective:

Clustering the XYZ Bank customers into 5 groups by finding the pattern which group certain kind of customer in one category.

Data Collection

Total number of observations	1000000
Total number of files	1
Total number of features	47
Base format of the file	csv
Size of the data	366.2+MB

Data types:

Float type features:9

Int type features:24

Object type features:15

Memory usage: 366.2+ MB

Data Preprocessing

- Features names are inappropriate: all the names of the columns are in spanish, we convert into english language with rename methods in pandas.
- Some features data types are inappropriate: For instance, age and seniority months columns have float values, those columns are converted into appropriate data types using pandas. Some Data quality issues.
- More than 80% of records are missing particular features: we handled null values using two approaches Simple Imputer , and IterativeImpute which both approaches works well with big data size.

Data Preprocessing

Feature engineering:

- Same subsets of the group have different columns (one hot encoding format). For instance, term deposit and account types.
- problems are difficult to identify the customers which have no term deposit and number of accounts types to solve this problem we merge different columns into single columns which belong to the same categorical group.

Building Clustering models

Before staring building models, we installed the Rapid Library that have cuml models.

Cuml models works with GPU and parasitism computations by segmenting the main data into baches.

We applied two approaches :

Approach 1:

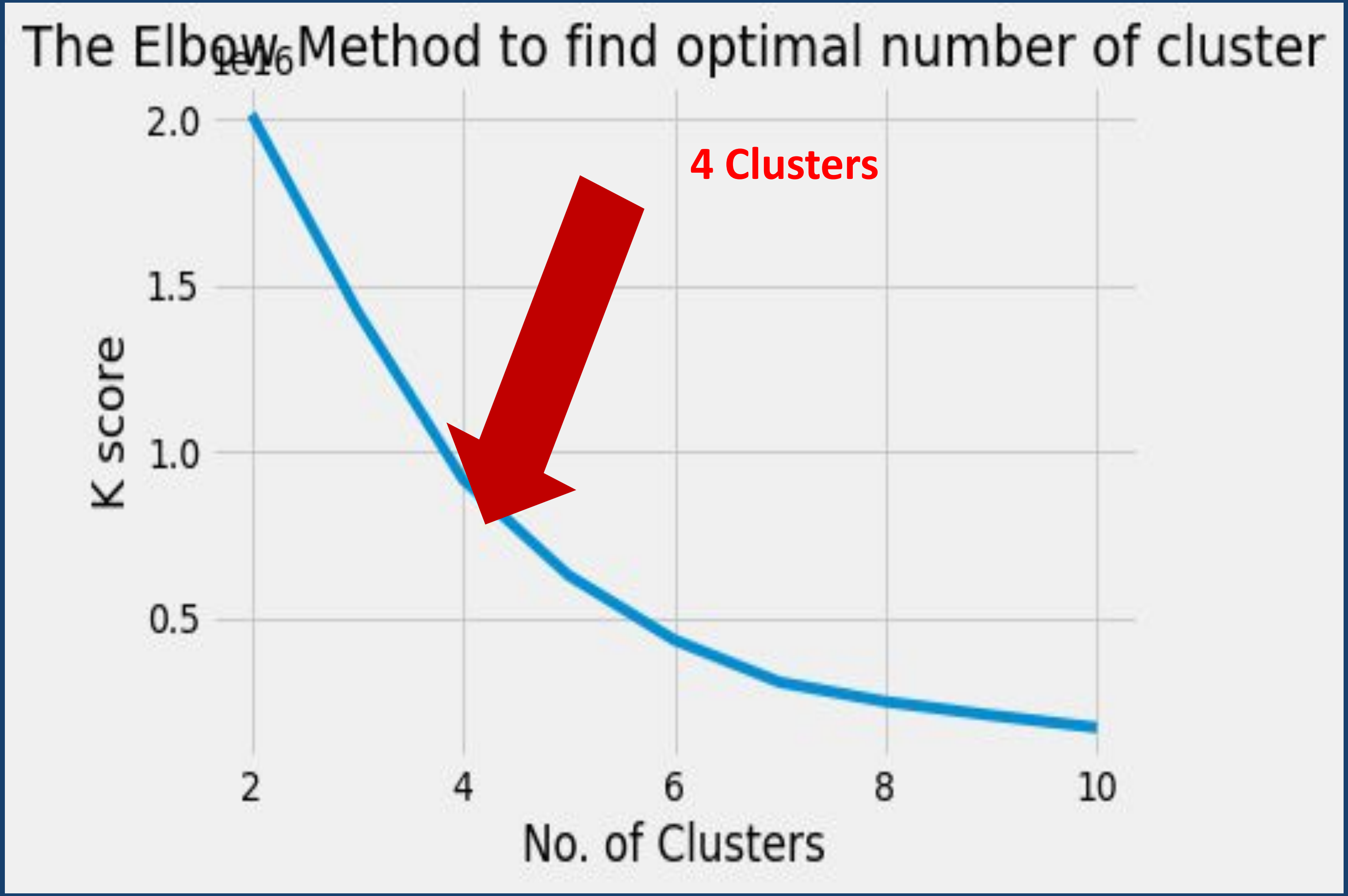
Feature selection: by choosing three features;["Customer seniority","Gross income of the household","Age"], then applying min_max scaler, Pca transform. Lastly, two clustering models were applied; Kmeans clustering and Agglomerative Clustering

Approach 2:

Feature reduction: by using all the features, minmax scaler and then applying PCA to find the two main components. And then applying Agglomerative Clustering

Approach 1:

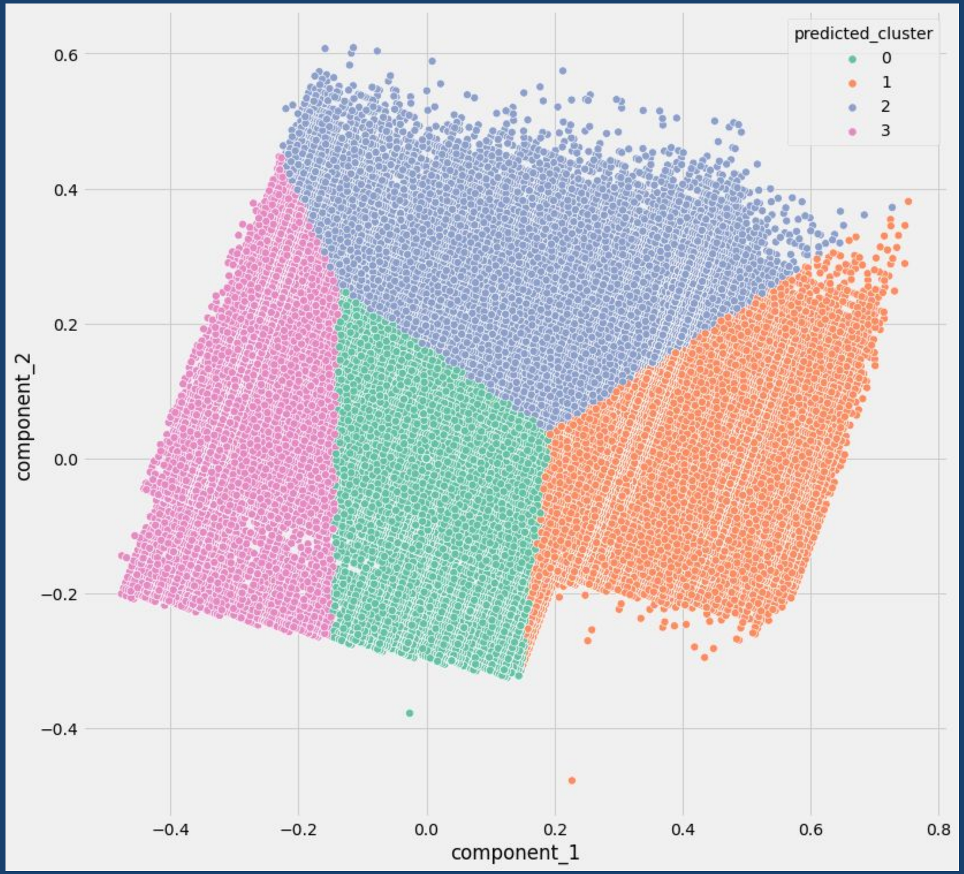
We applyied the Elbow before applying the approach 1. The elbow method shows that the number of the clusters is 4.

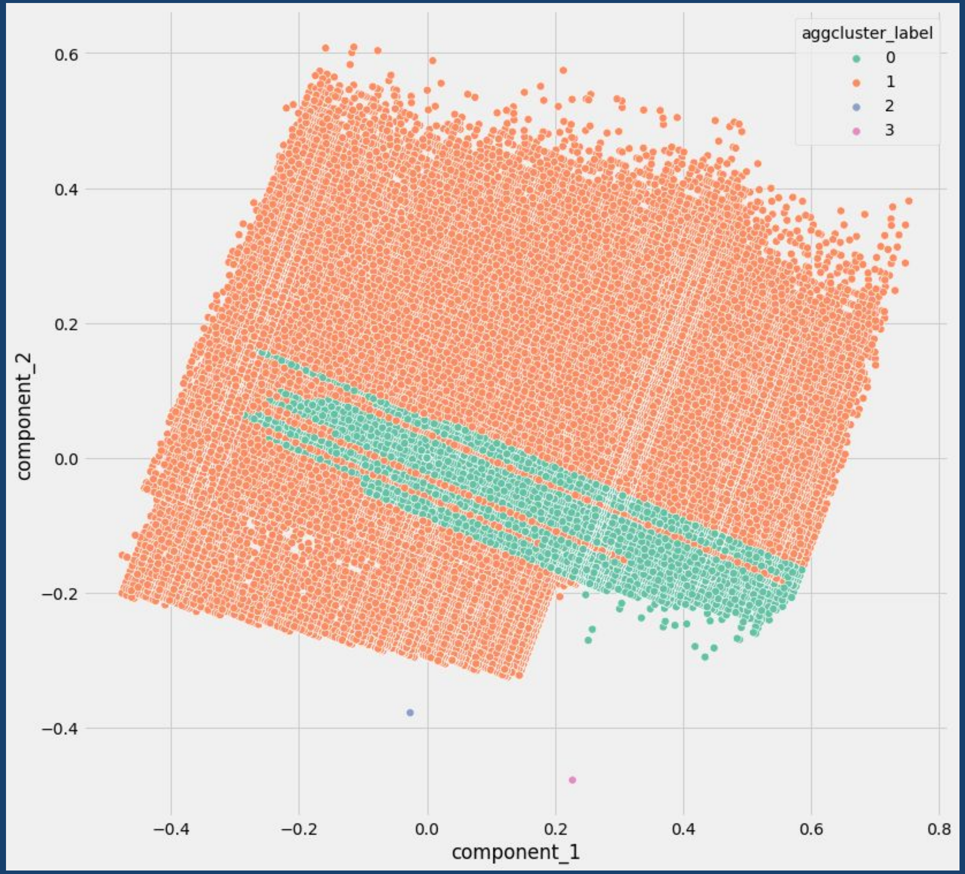


Approach 1: Clustering models

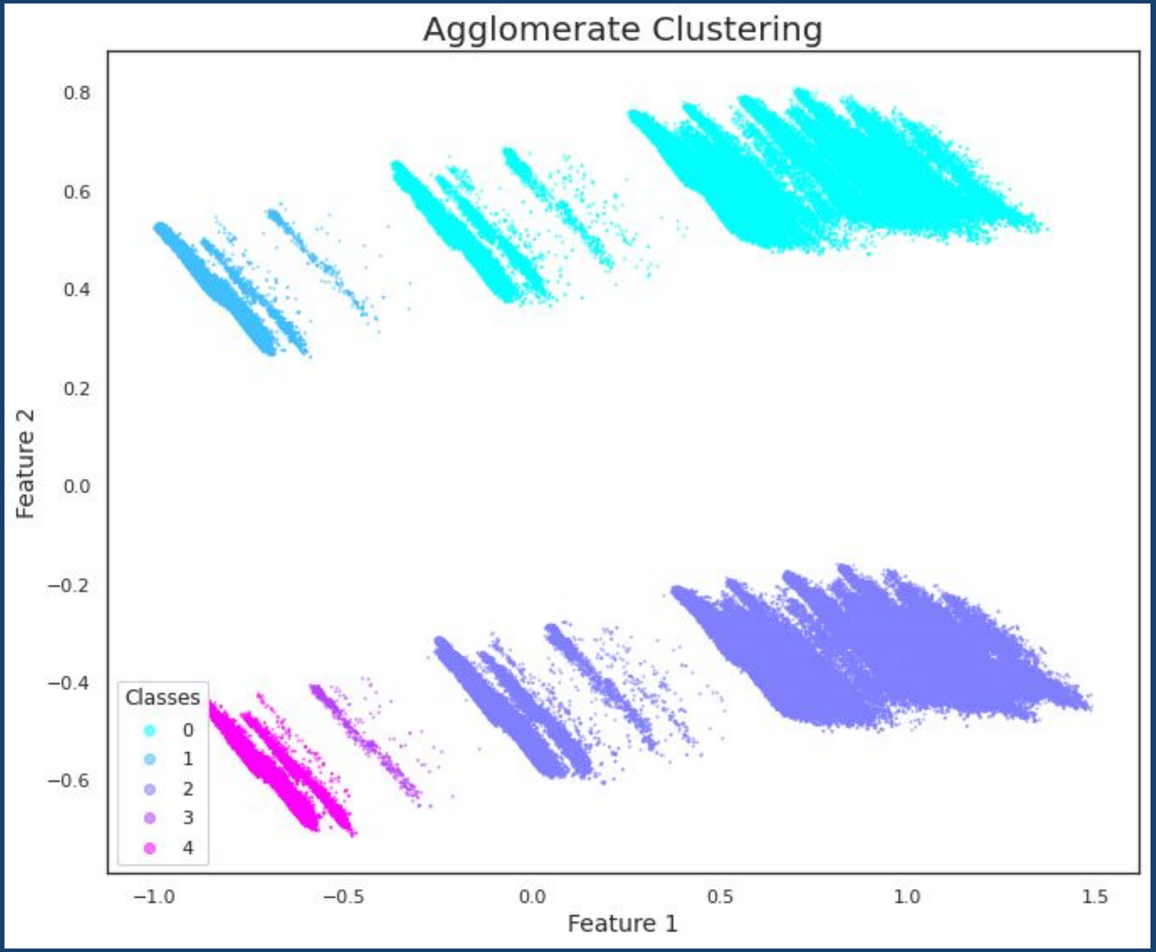
Winner

Kmeans Clustering and Agglomerative Clustering from RAPIDS CuML was applied. The results shows that Kmeans clustering is more efficient in segmenting the data.





Approach 2:
Clustering models



AgglomerativeClustering from RAPIDS CuML was applied. The results shows that agglomerate clustering is efficient in segmenting the data.

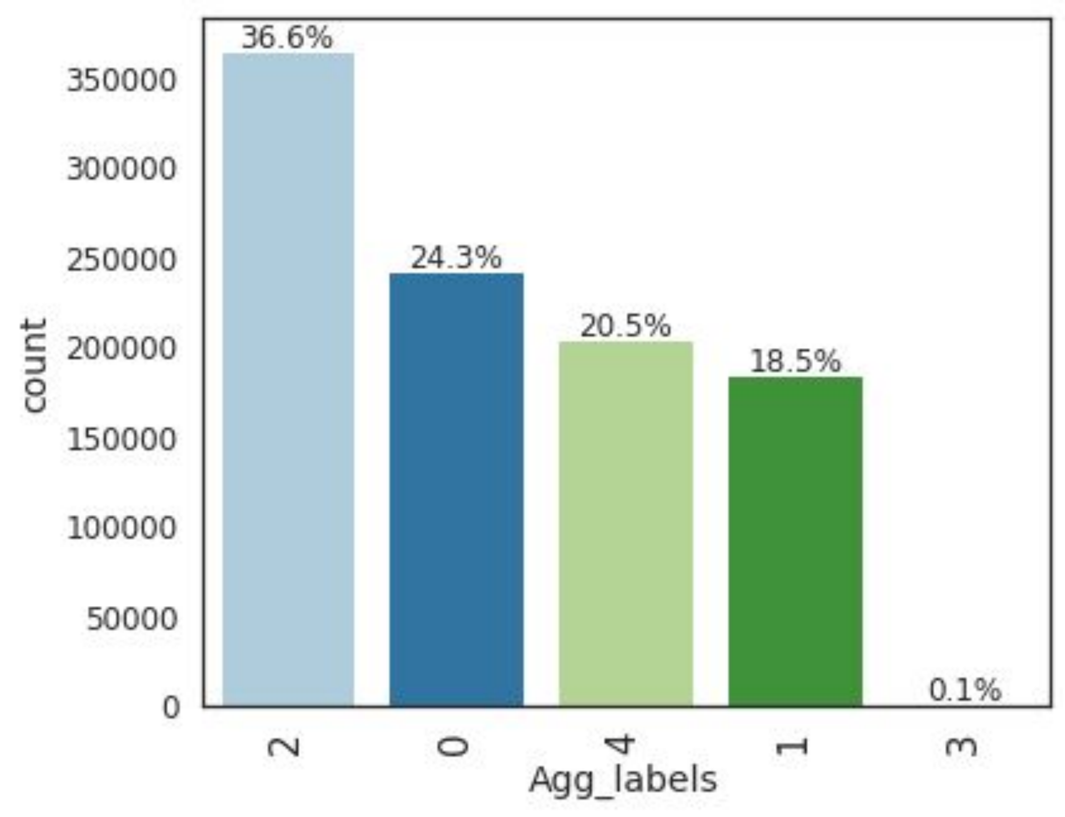
Silhoiett score for all the models

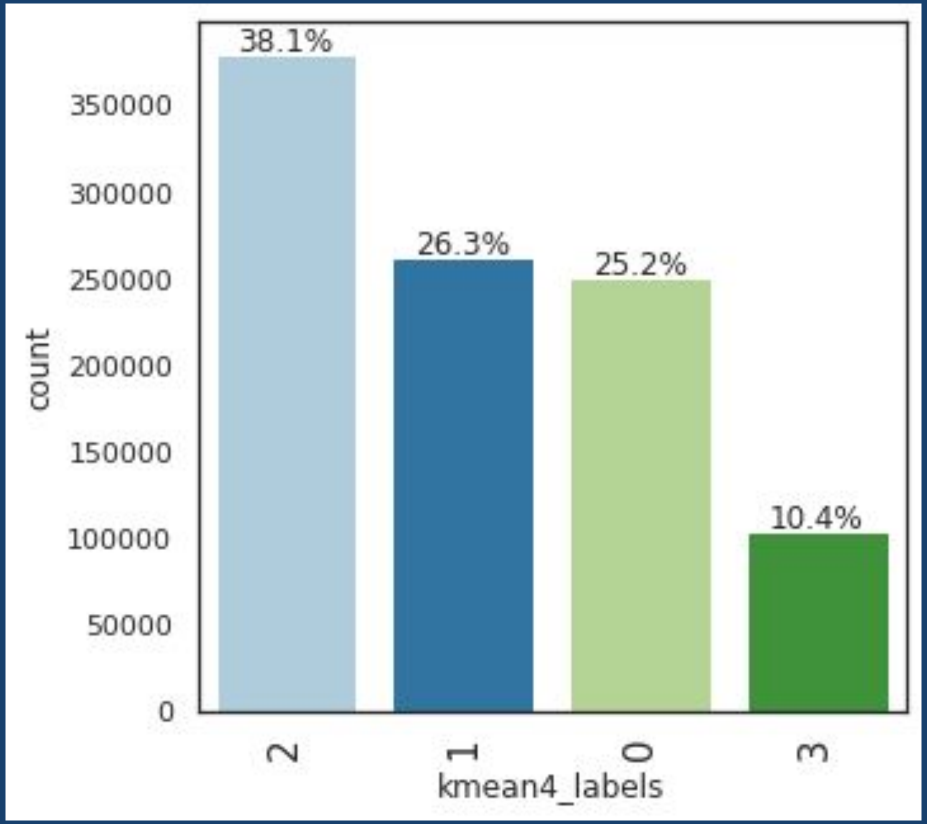
Cluster number=5	Cluster number=4
Silhoiett_score_AGG:0.596	Silhoiett_score_AGG: 0.54
	Silhoiett_score_Kmeans: 0.49

Comparing models

Agglomerative Clustering

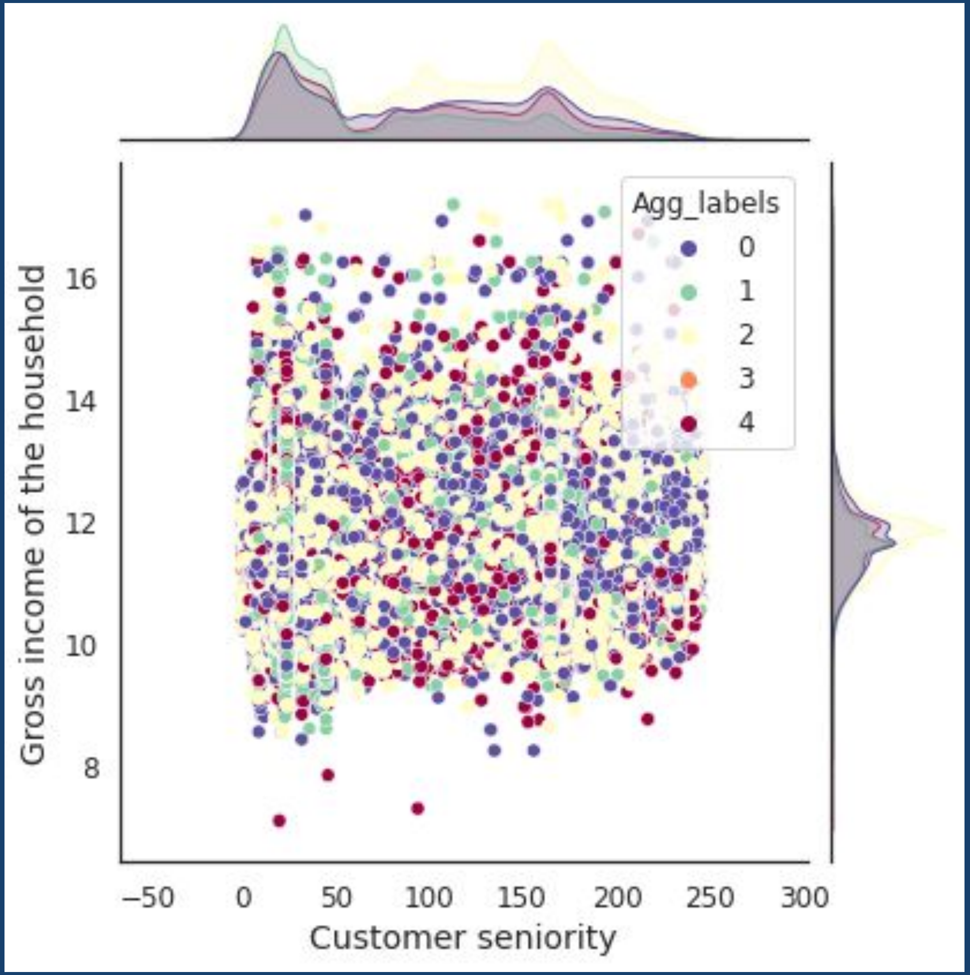
Kmeans clustering



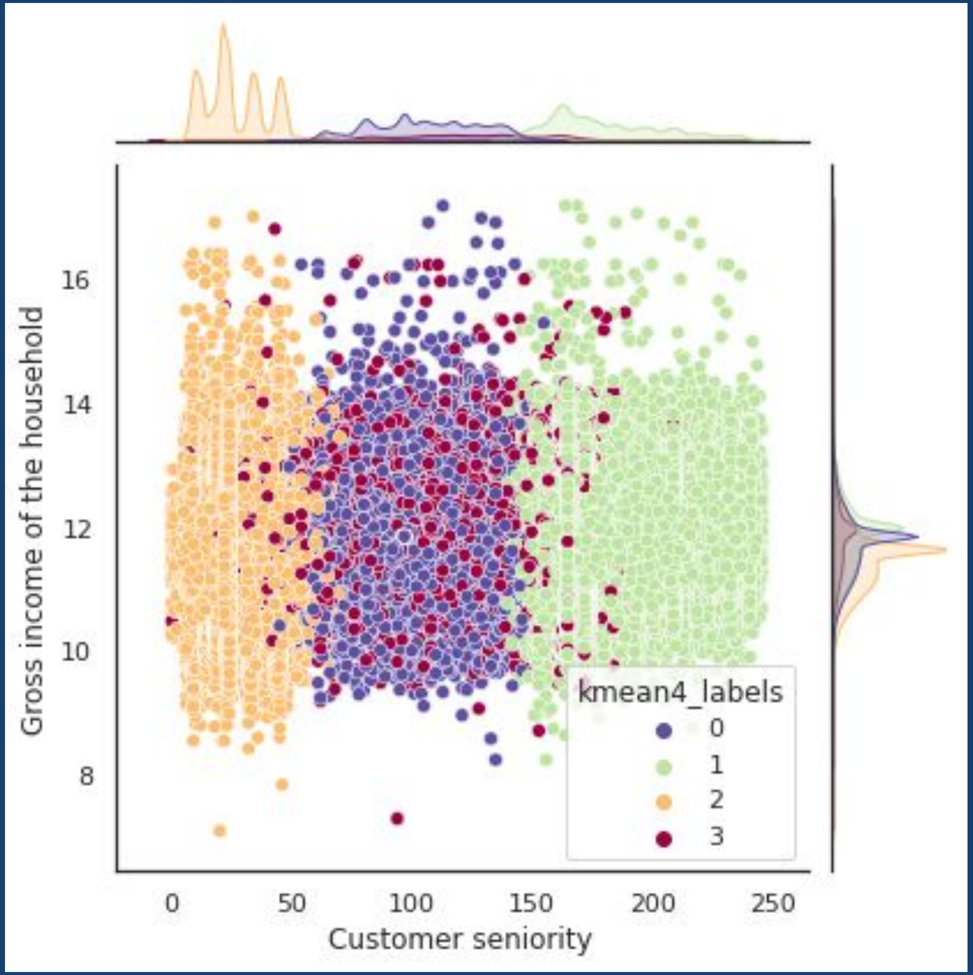


Comparing models

Agglomerative Clustering

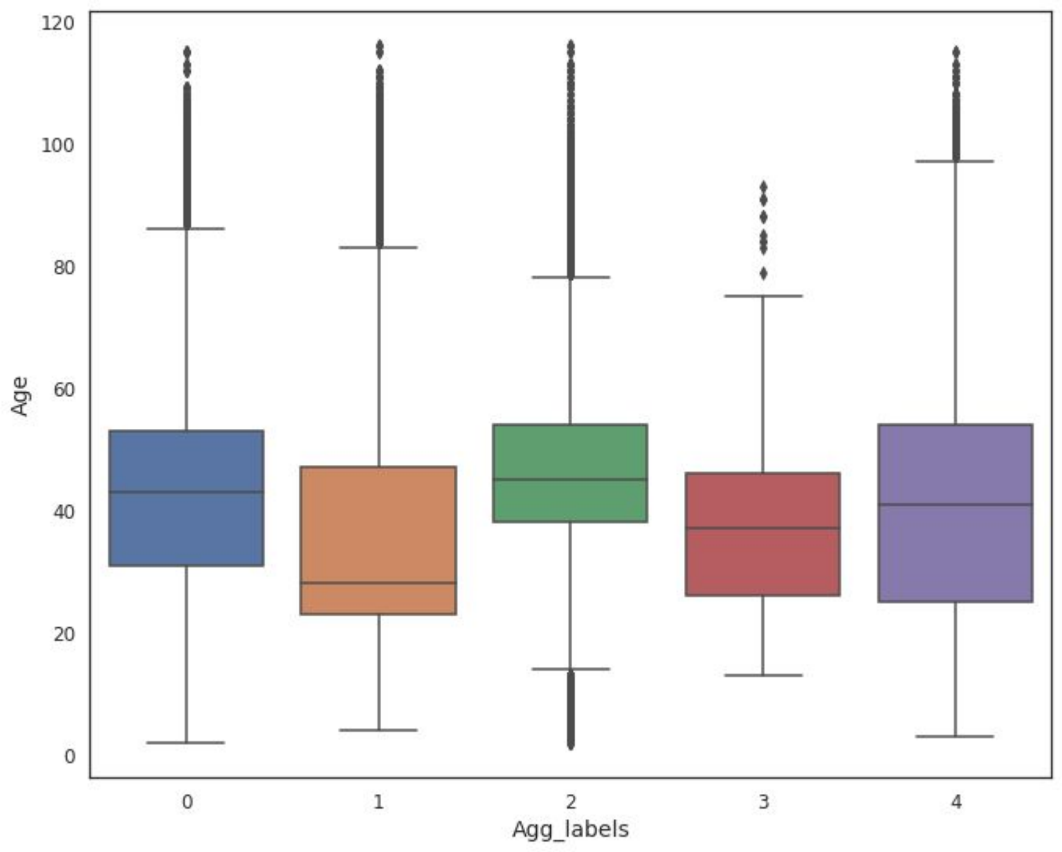


Kmeans clustering

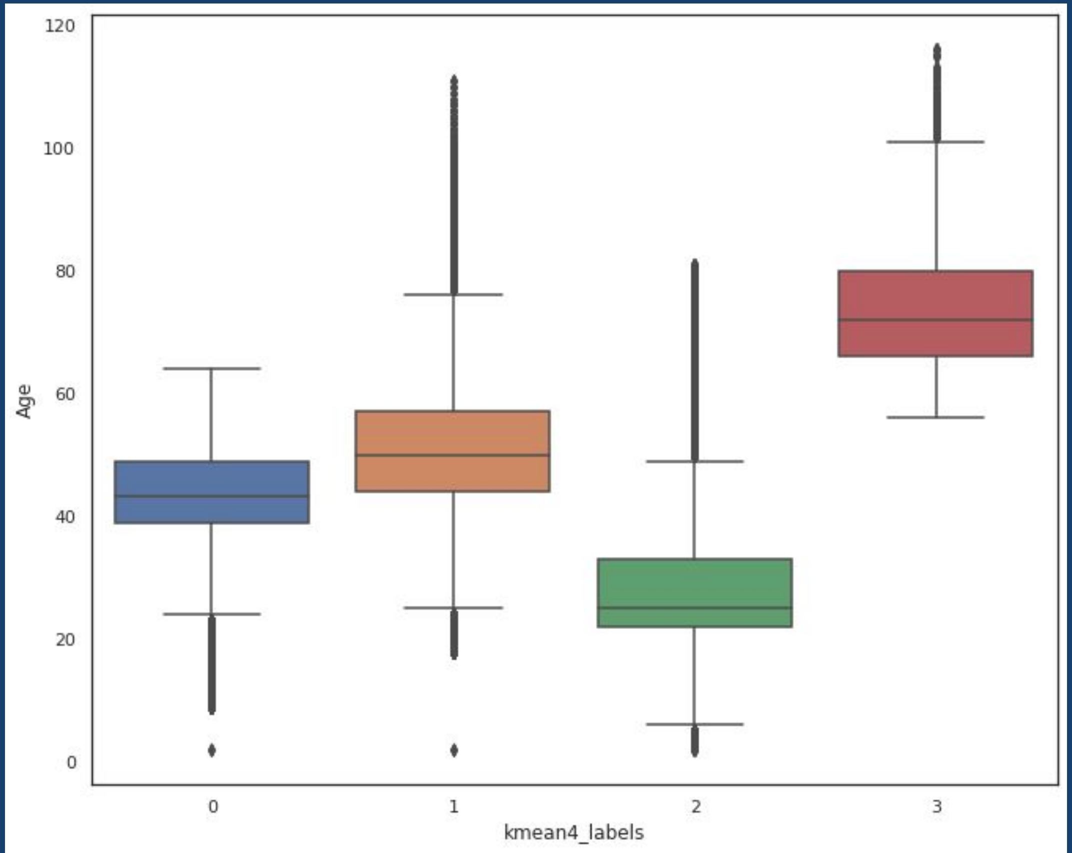


Comparing models

Agglomerative Clustering

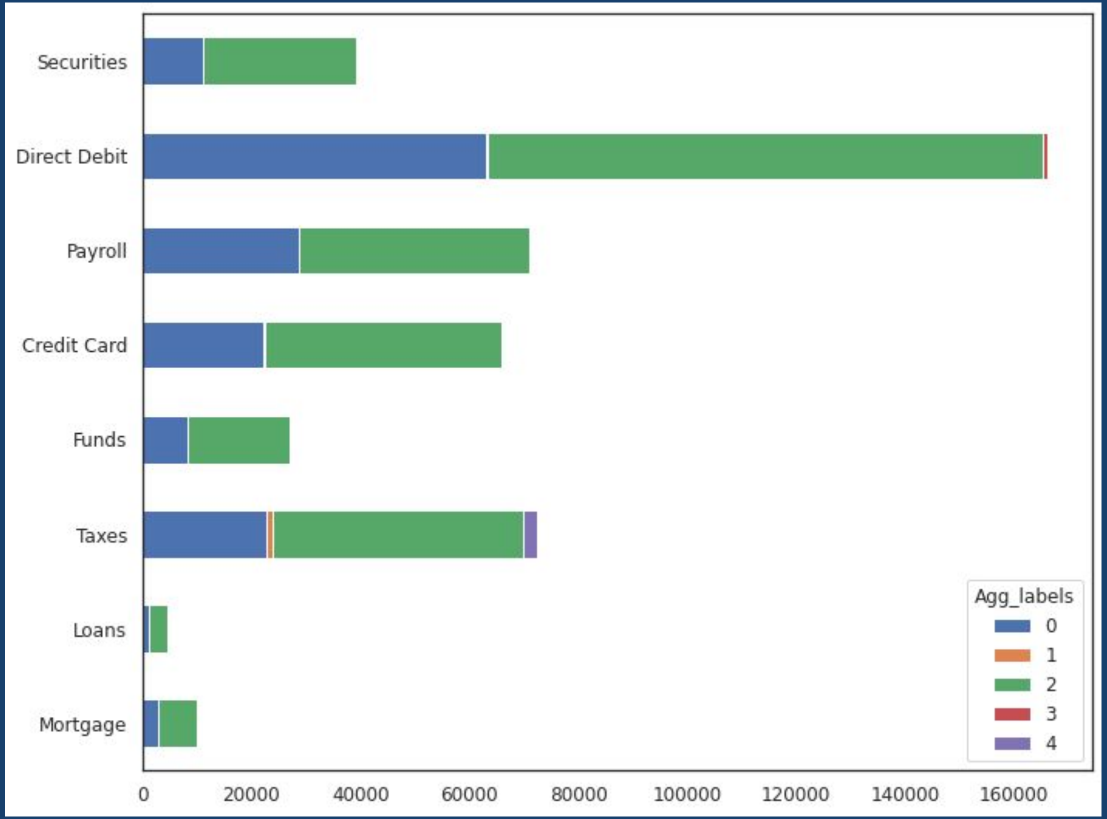


Kmeans clustering

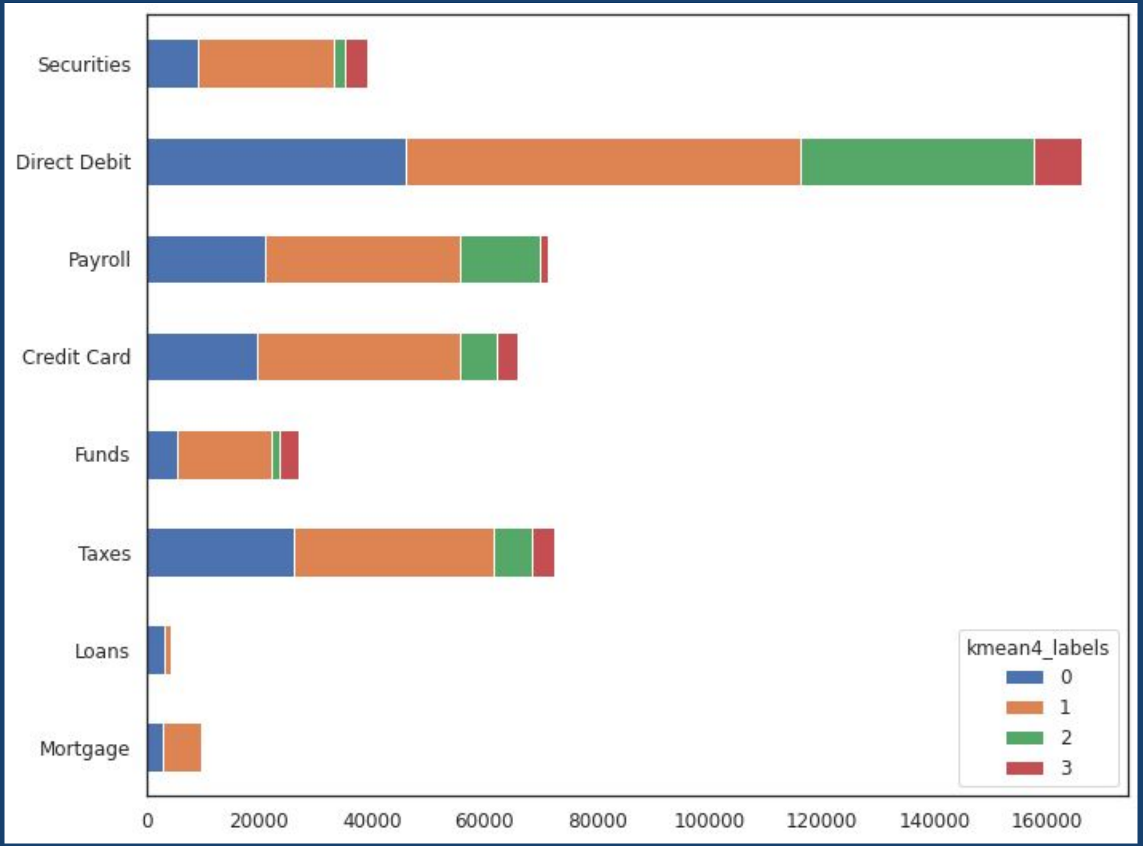


Comparing models

Agglomerative Clustering



Kmeans clustering



Results

We choose the first Approach, which uses Kmeans clustering with cluster number=4, and we applied it on selected numeric continuous features. Comparing the two models shows that this approach was more efficient in clustering the customers in meaningful groups.