EAI 6010 22894 - Applications of Artificial Intelligence

Module 6: Final Assignment – Segmentation and Profiling Project

Instructor: Prof. Abhijit Sanyal

Date : 04/08/2021

Author

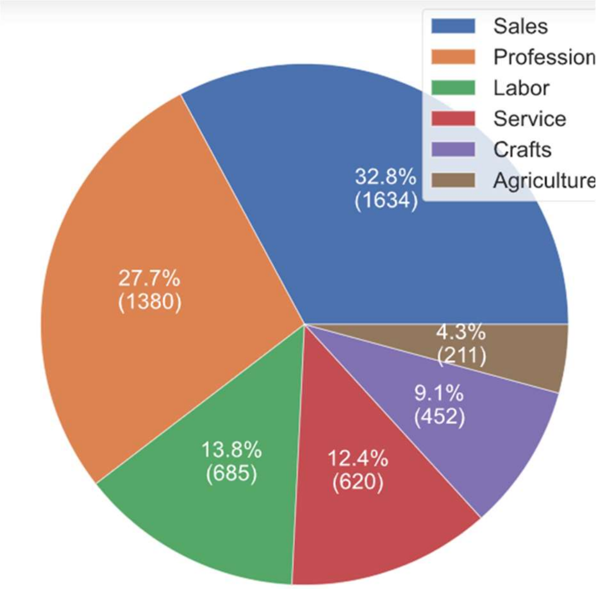
Anubhav Saha

Executive Summary

In this report, we are examining a customer dataset provided by a Telecommunications company. Our objective is to perform segmentation of this customer dataset based on certain metrics that we need to identify. These metrics or the segmentation drivers can be identified after performing a thorough exploratory analysis of our dataset. The segmentation of our customer dataset would allow us to profile our customers based on those metrics which would then help us to work on strategies that would ensure better customer retention and lesser churn rate. Now to identify these metrics that would drive our segmentation project, we start with performing an exploratory analysis of our customer dataset.

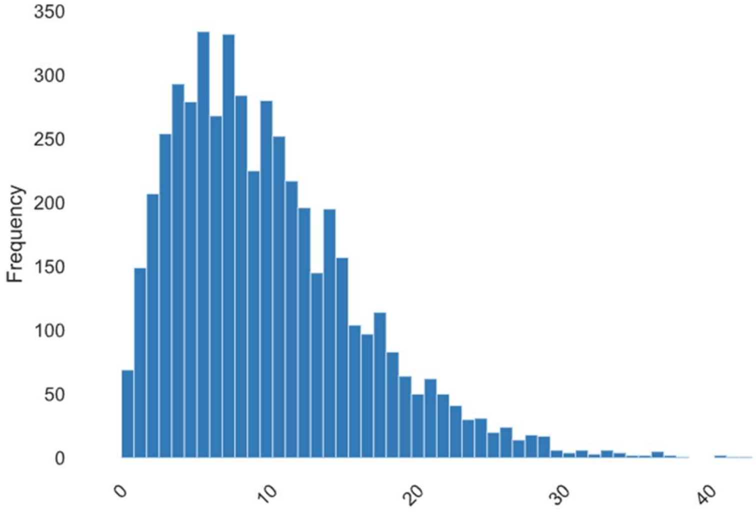
A quick look at our dataset tells us that the dataset consists of predominantly customer demographics such as the age of the customer, gender, region they belong to etc. Some of these attributes can be more important that others, such as whether the customer is retired or not, whether the customer is a News subscriber or owns certain types of devices, etc. We need to identify which of these attributes can be key drivers in profiling the customers so that strategies can be developed to ensure their retention of services. Some of the key attributes such as Debt to income ratio and the household size can be key indicators on whether the customers are likely to retain the services they have opt for or are possible to churn.

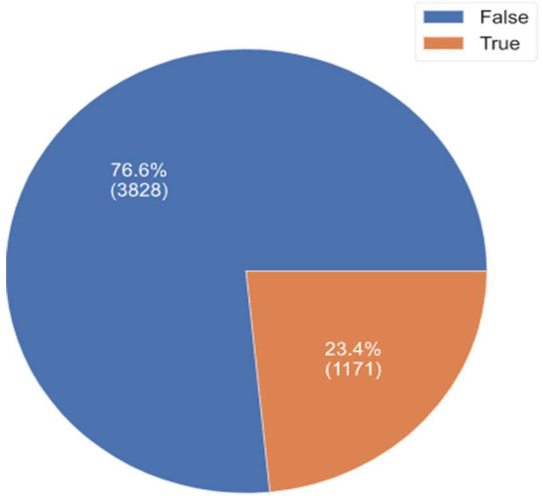
It is useful to get a sense of the employment background of the customers, that would help in profiling the customers together. The graph below shows the proportions of customers based on their Job Category.



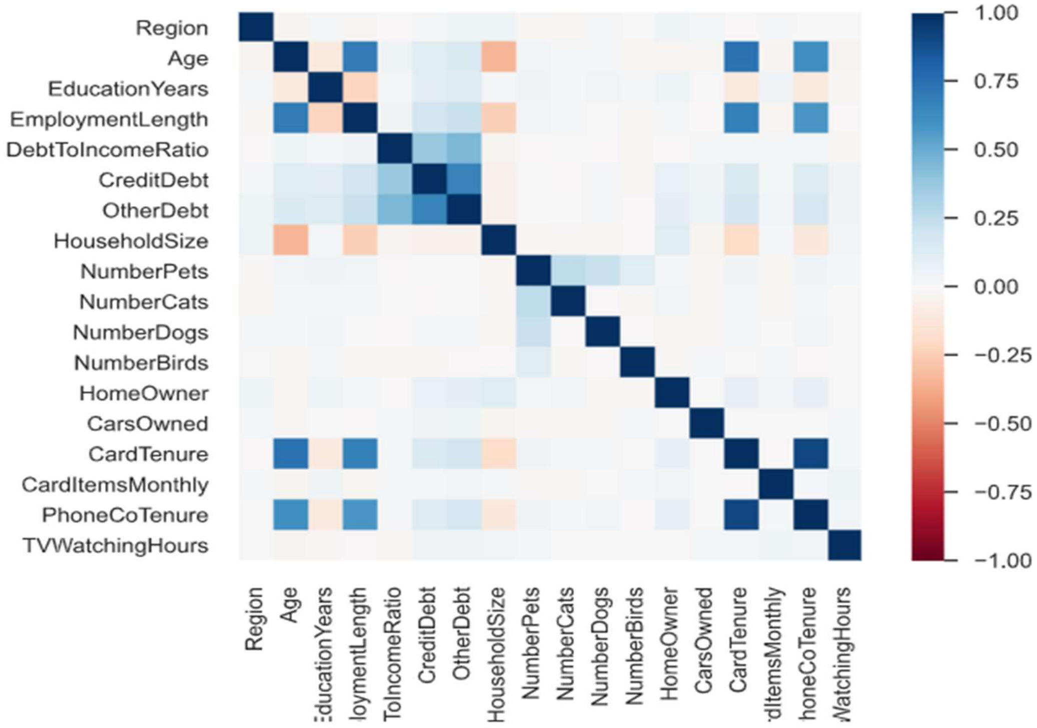
From the pie chart above, we can see that the majority of the customers belong to sales category comprising of 32.8%, followed by professional at 27.7%. These two categories cover majority of the customers. The Labor and Service job categories consists of about 13%. Very few of the customers belong to the Agriculture sector.

It will be useful information to observe the credit history information of the customers, that would include metrics like, have they ever defaulted on loan and what is their existing debt to income ratio shown as below.





We can see that most of the customers lie between 0-20 for the debt to income ratio and majority of them have never defaulted on loan. After seeing the individual metrics, it will be useful to check the correlation between these metrics in our dataset. The following plot shows the correlation between the features of our dataset.



The scale on the right shows the intensity of the correlation based on the shades of the colors blue and red which stand for direct and inverse correlation, respectively. There is some strong positive correlation between Age and Card Tenure, which logically makes sense. Same correlation can be seen with employment length and card and phone co tenure, which are all expected. An interesting correlation can be seen with the debt owners, those who have credit card debts also seem to have other types of debts, which can be useful information while creating segmentation hypothesis/approach.

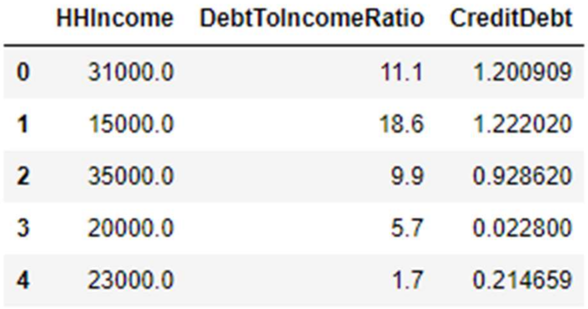
A popular strategy in market research for customer segmentation is the RFM segmentation method, which allows marketers to target the groups of customers with specific strategies to reduce the churn rate. So, we can use this approach and segment our customers based on the amount of revenue they are generating for the company. We can segment our customer database as, high value customers, mid value customers and low value customers. High value customers would be the segment of customers that would generate high amount of revenue to the company, hence retaining them would be the priority of the marketers. Similarly, low value customer segment would consist of customers who are not using a lot of the services and have minimum engagement with the company. The mid value segment will therefore consist of customers that lie in the middle of both ends, i.e. nominal engagement and revenue generation for the company.

Now the way to achieve these segments would be using k-means clustering, which is one of the most popular machine learning approach for clustering data points together based on their similarity. To go about applying this algorithm in our dataset, we have first preprocessed the dataset and standardized the dataset using the sklearn package in Jupyter Notebook. The recency in our RFM method tells us how much time has elapsed since a customer's last activity or transaction with the company. Our customer dataset does not have features showing the last engagement time; hence we move on to the frequency aspect which shows how often do the customers interact/engage with the company. We will identify features in our dataset that correspond to the frequency aspect. Finally, the monetary aspect would tell which customers spend the most to the company, hence we will identify attributes addressing to that.

Since our dataset does not have a feature directly corresponding to a customer's engagement, we move on to selecting the monetary attributes that predict the likely engagement of the customers, such as the household income of the customers, debt to income ratio and credit debt. These set of attributes together can give an idea about the likely future engagements of the customers. These features indicate a customer's ability to likely engage with the company's products or services in the future. For the frequency of engagement, we can consider combination of features such as TV watching hours, voice last month and household income. These features can indicate the amount of time the customers are engaging with the company's products/services. So, using the combination of these attributes, we will run k-means clustering to obtain our customer segments in an iterative way.

Analysis and detailed findings

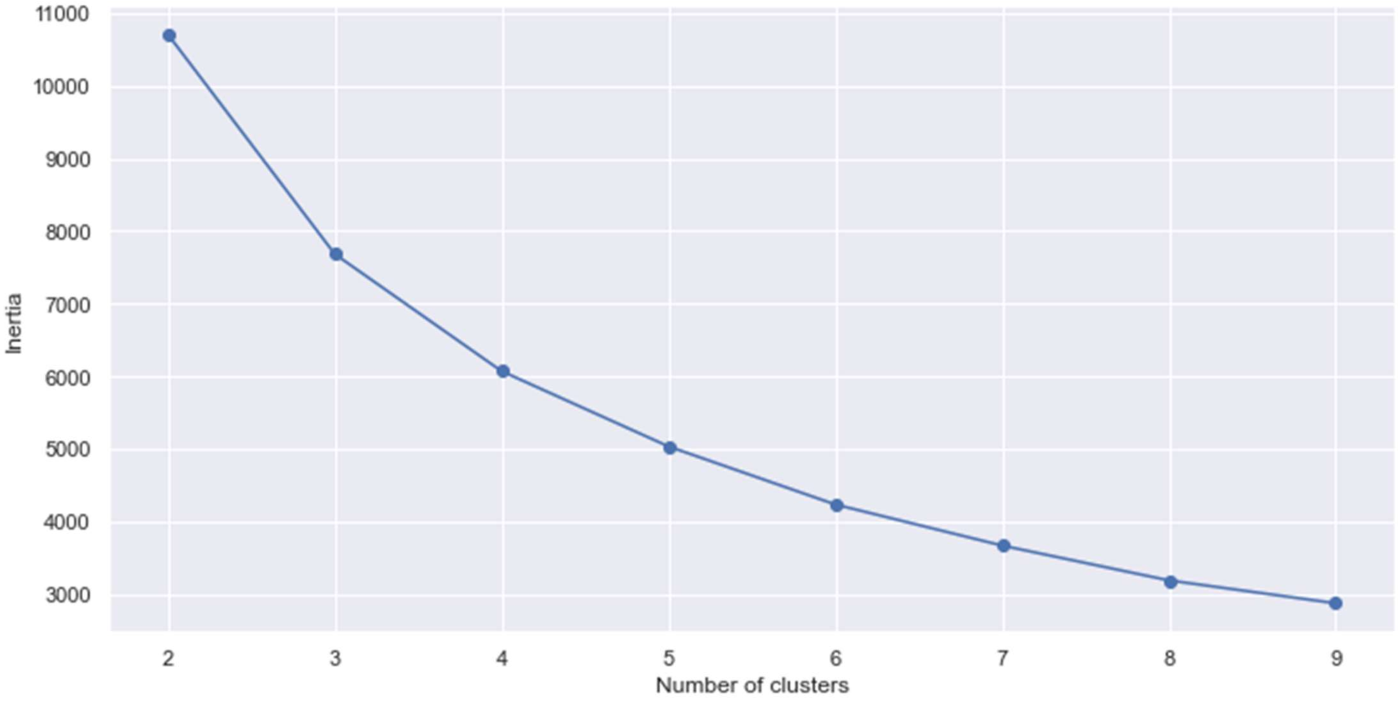
For the first iteration of k-means clustering, we use the attribute group of the monetary features that we identified, i.e., Household income, debt to income ratio and credit debt. Here is a glimpse of these features.



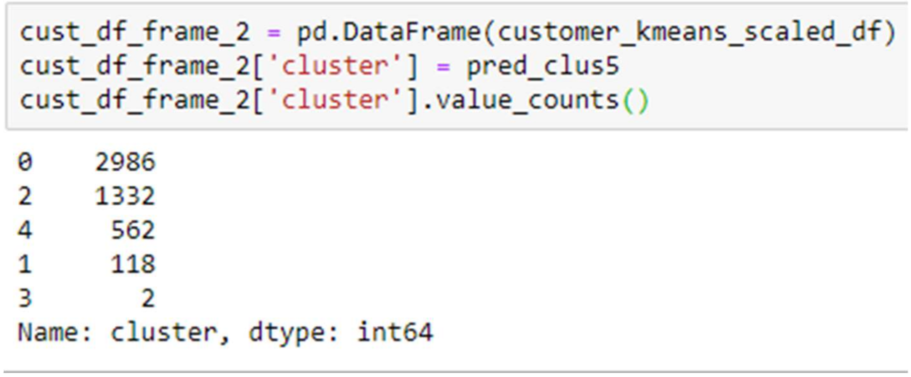
	HHIncome	DebtToIncomeRatio	CreditDebt
0	31000.0	11.1	1.200909
1	15000.0	18.6	1.222020
2	35000.0	9.9	0.928620
3	20000.0	5.7	0.022800
4	23000.0	1.7	0.214659

In order to use the k-means clustering algorithm, we standardize and scale this dataset using sklearn's StandardScalar method followed by running the k- means algorithm.

The k-means algorithm creates clusters of the data points by separating them in different groups of equal variances by minimizing the inertia. Inertia is a measure that indicates how internally coherent the clusters are. Therefore, we plot the inertia for our different cluster solutions against the number of clusters. This allows us to identify the right number of clusters that needs to be used while running the k-means clustering algorithm on our customer dataset.



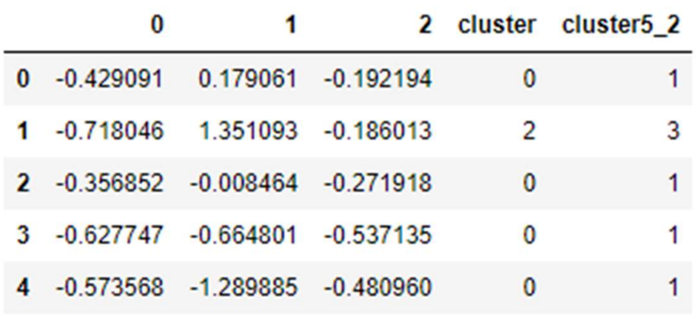
Based on the above plot, the optimal number of clusters can be somewhere around 4 and 5. Now, we will fit the k-means algorithm on our scaled dataset using number of clusters as 5 to create a 5 – segment solution.



```
cust_df_frame_2 = pd.DataFrame(customer_kmeans_scaled_df)
cust_df_frame_2['cluster'] = pred_clus5
cust_df_frame_2['cluster'].value_counts()
```

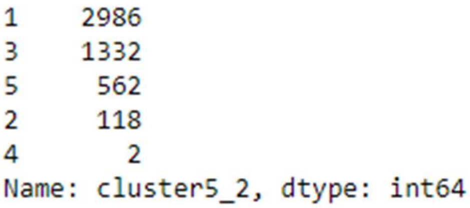
```
0    2986
2    1332
4     562
1     118
3       2
Name: cluster, dtype: int64
```

We get our 5 segment solution as above, indexed 0 to 4. Cluster 0 seems to have the highest number of customers of 2986, followed by cluster 2 with 1332 customers, cluster 4 with 562 customers, cluster 1 with 118 and cluster 3 with only 2 customers. We can fix the cluster number by adding 1 to it, so that we obtain clusters 1 to 5 as shown below:



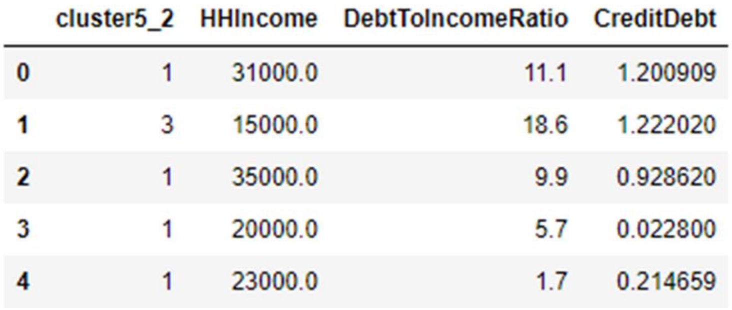
	0	1	2	cluster	cluster5_2
0	-0.429091	0.179061	-0.192194	0	1
1	-0.718046	1.351093	-0.186013	2	3
2	-0.356852	-0.008464	-0.271918	0	1
3	-0.627747	-0.664801	-0.537135	0	1
4	-0.573568	-1.289885	-0.480960	0	1

This column can now replace the original column to reference the cluster numbers. The counts would now look as below which is more intuitive to read.



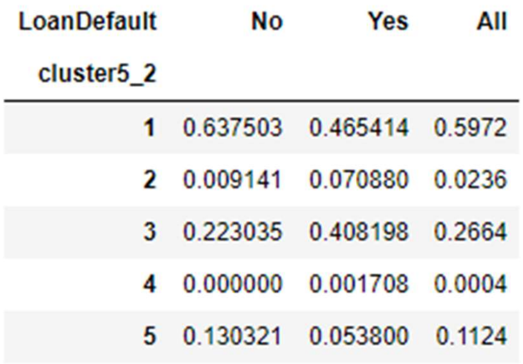
```
1    2986
3    1332
5     562
2     118
4       2
Name: cluster5_2, dtype: int64
```

We can now join these clusters to the original data so that all customers lie under clusters, the head would look like below:



	cluster5_2	HHIncome	DebtToIncomeRatio	CreditDebt
0	1	31000.0	11.1	1.200909
1	3	15000.0	18.6	1.222020
2	1	35000.0	9.9	0.928620
3	1	20000.0	5.7	0.022800
4	1	23000.0	1.7	0.214659

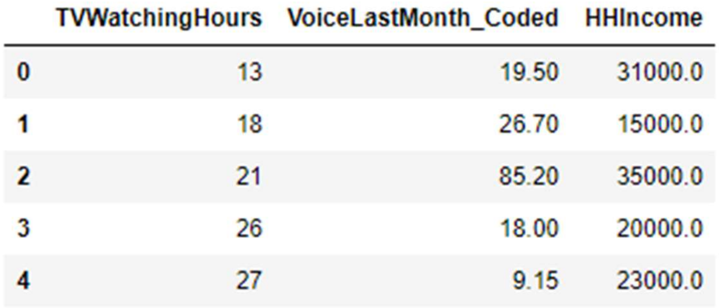
Now we can use cross tab with the feature of our choice to get a sense of how the customers in these clusters are doing with respect to the segment of our choice. Let us see how our customers in the 5 segments are performing with respect to Loan defaults as shown below which would give idea about monetary abilities of our customer segments and hence it will let us create specific strategies.



LoanDefault	No	Yes	All
cluster5_2			
1	0.637503	0.465414	0.5972
2	0.009141	0.070880	0.0236
3	0.223035	0.408198	0.2664
4	0.000000	0.001708	0.0004
5	0.130321	0.053800	0.1124

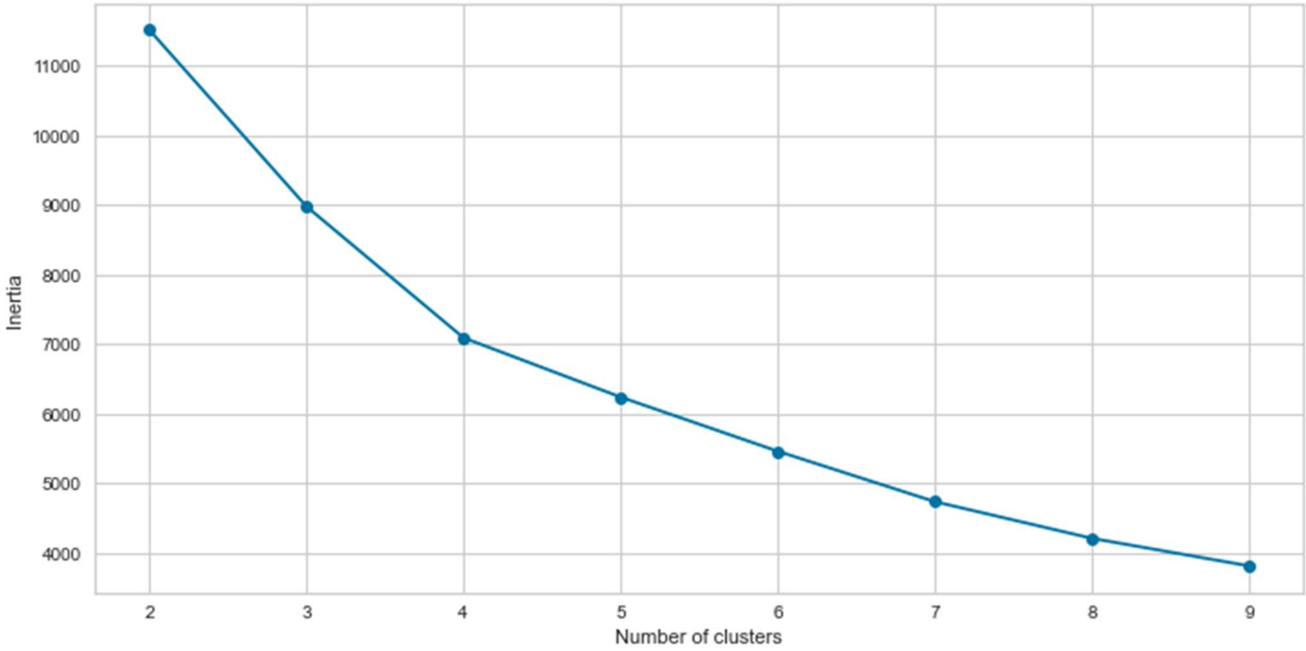
As we can see from the output above, our first segment of customers has maximum customers around 64% who have not defaulted on loan. This suggests that this cluster of customers can be put in high value segment of customers that we had derived earlier. Similarly, the customers in the cluster 3 seem to have a higher number of loan defaults, which can suggest that those cluster of customers are more suited for a low valued customer which supports the results obtained from the previous output.

Next, we can run our second iteration with the next set of attributes we had determined, that is the TV watching hours, Voice last month and the Household income. A glimpse of these attributes looks as below.



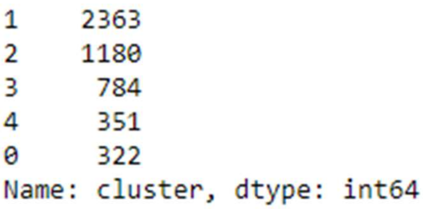
	TVWatchingHours	VoiceLastMonth_Coded	HHIncome
0	13	19.50	31000.0
1	18	26.70	15000.0
2	21	85.20	35000.0
3	26	18.00	20000.0
4	27	9.15	23000.0

We perform the similar set of activities of standardizing and scaling this data frame and obtain the plot of inertia versus the number of clusters to identify the optimal number of clusters.



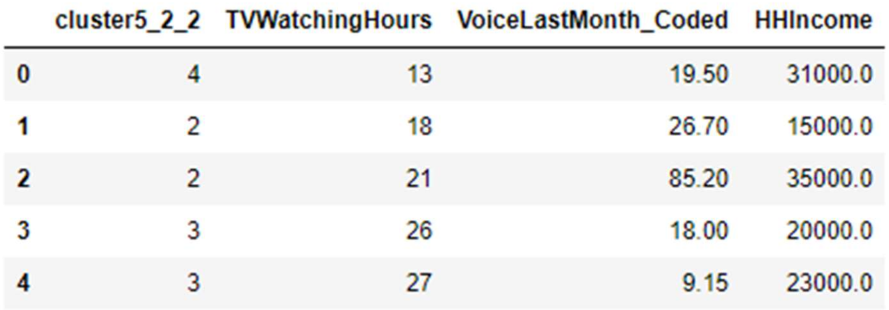
Similarly like the previous iteration, the optimal number of clusters looks to be around 4 and 5. Therefore we choose the number of clusters as 5 and proceed with the clustering algorithm.

We obtain the following clusters from the k-means clustering algorithm.



```
1    2363
2    1180
3     784
4     351
0     322
Name: cluster, dtype: int64
```

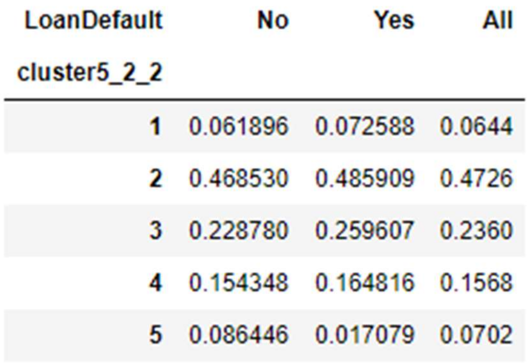
We can add 1 like before to make the cluster numbers more intuitive and then use that column for reference.



	cluster5_2_2	TVWatchingHours	VoiceLastMonth_Coded	HHIncome
0	4	13	19.50	31000.0
1	2	18	26.70	15000.0
2	2	21	85.20	35000.0
3	3	26	18.00	20000.0
4	3	27	9.15	23000.0

From the output above, it looks like clusters 4 and 2 consist of customers with higher household income. The customers in cluster 2 also appear to have a high voice last month value. Together it can suggest that these cluster of customers should belong to a high value customer segment who need to be retained in order to maintain profitability for the company.

This can be further checked by doing a cross tab with Loan Default like we did in previous iteration.



LoanDefault	No	Yes	All
cluster5_2_2			
1	0.061896	0.072588	0.0644
2	0.468530	0.485909	0.4726
3	0.228780	0.259607	0.2360
4	0.154348	0.164816	0.1568
5	0.086446	0.017079	0.0702

Here the number of loan defaults of yes and no appear to be in equal proportions, which makes sense because this time, our features were not purely monetary, but it had the frequency elements of number of hours of TV and the voice last month feature in them.

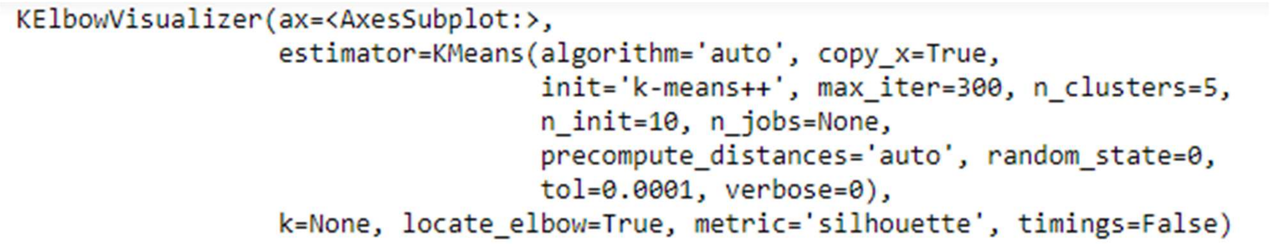
Conclusion:

Therefore, to conclude, we were able to identify certain features in our customer dataset that could act as drivers in our segmentation problem. We performed two sets of iterations, choosing a combination of three different features in each of these two iterations. The first iteration had features that were inspired by the monetary part of the RFM approach. Those monetary related features were identified to be Household income, debt to income ratio and credit card debt. Using these features, we identified the optimal number of clusters to be 5 by plotting the inertia vs clusters plot to look at the elbow. After finding the optimal number of clusters for our dataset, we ran the k-means clustering algorithm to cluster our customers into 5 clusters based on our 3 selected features representing the monetary aspect. Once we had the clusters of customers, we could observe which clusters of customers had higher income customers who have least number of loan default rates by cross tabbing it with that feature. These values helped us identify that cluster 1 consists of high value customers and therefore can be placed in the segment of high value customers. This profiling of customers would help the marketing team design custom promotions to these customers in order to reduce the churn rate and retain these customers as they are likely to provide a higher revenue to the company.

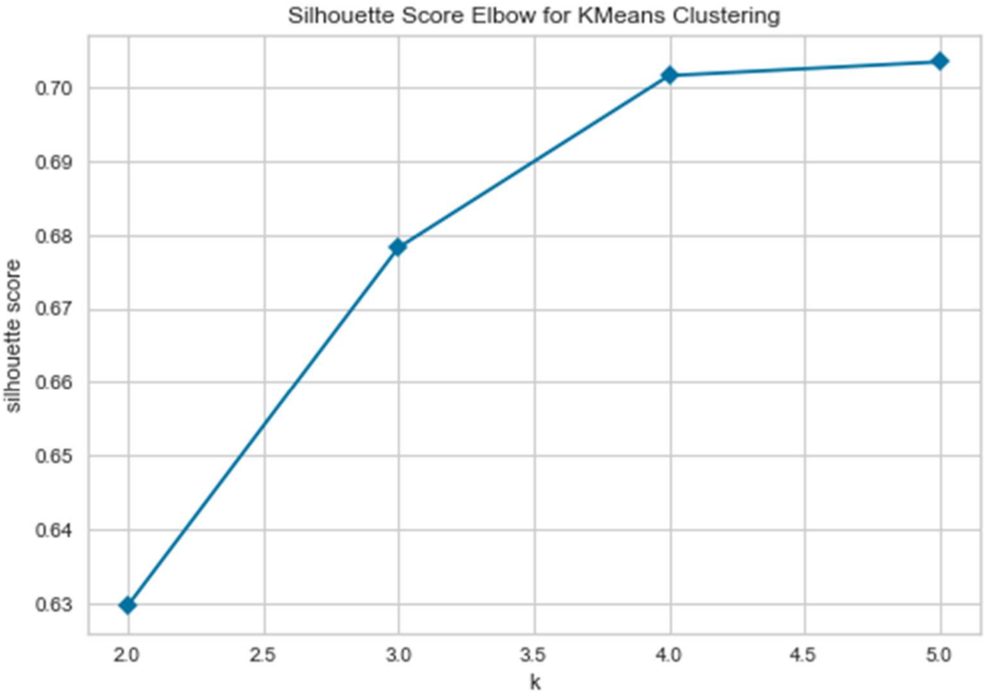
Similarly, in our second iteration, we used the combination of features inspired from the frequency aspect of the RFM methodology, which included the hours of TV watched, Voice last month and Household income. This iteration also provided 5 as the optimal number of clusters, after which we ran the k-means clustering algorithm and obtained the 5 clusters of customers. Here we found that the clusters 2 and 4 consisted of customers with higher household income, so these clusters could be placed in the segment of high value customers. Similarly, we can identify the clusters of customers that consist of customers with low household income and similar features that indicate lower monetary ability and lower engagement ability and then place them in the segment of low values customers. This was we will have a profile for low values customers and our marketing team can then derive strategies to provide promotions and discounts specifically targeting this profile of customers that could provide a higher chances of them increasing engagement and reducing the churn rate.

Thus, we observed that segmenting customers into different profiles can help us devise strategies to reduce customer churn rate and increase engagement and retainment. This can help grow the revenue of the company. The way to achieve these segments involve identifying the driver features from our dataset and then running unsupervised machine learning clustering algorithms that would cluster our dataset into similar data points based on the features that we provided and the number clusters that we wanted, the value of which was also derived technically. This way we can solve a company’s customer segmentation problem to better retain their customers by profiling them into segments to help the marketing department.

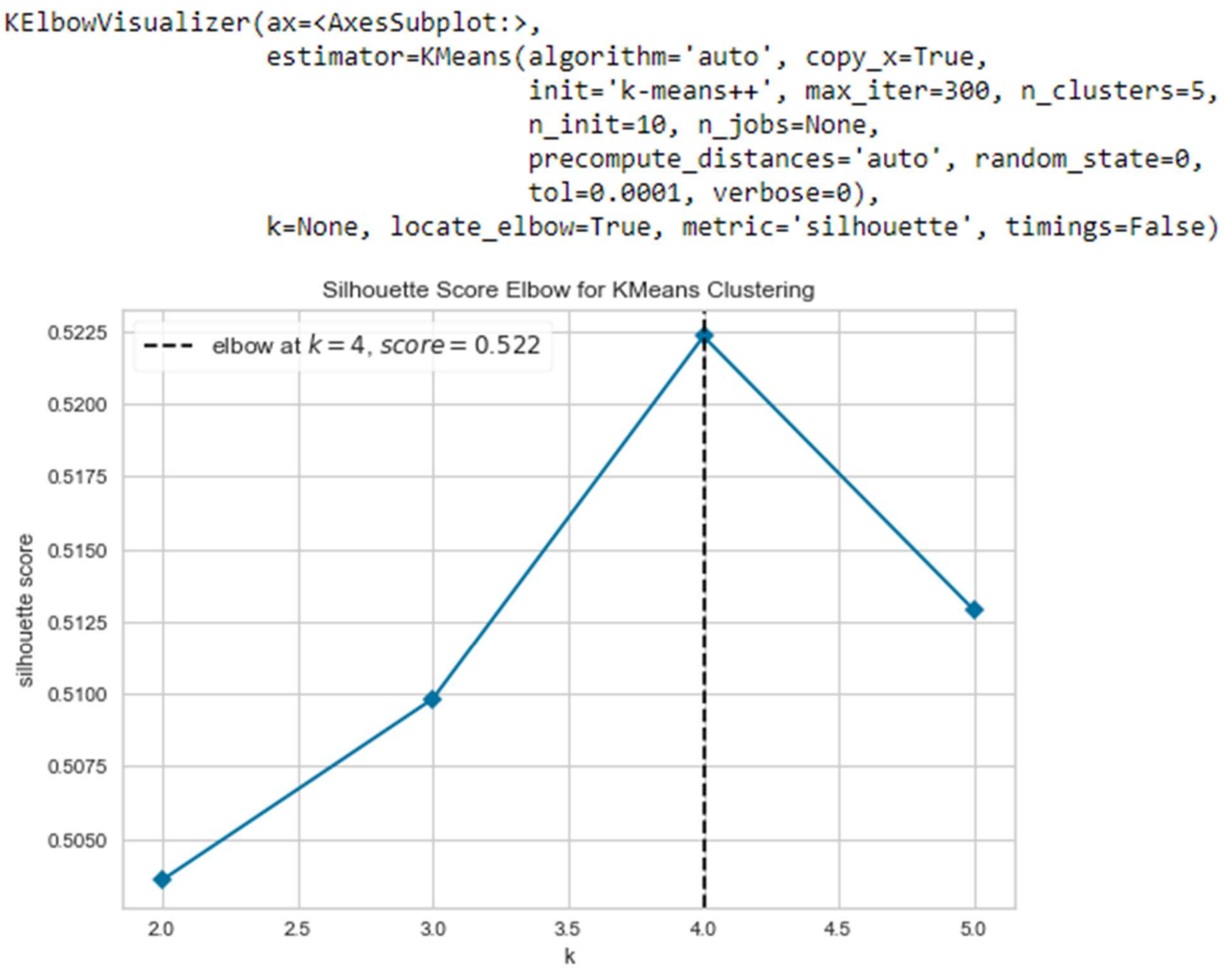
Appendices:



```
KElbowVisualizer(ax=<AxesSubplot:>,
                 estimator=KMeans(algorithm='auto', copy_x=True,
                                  init='k-means++', max_iter=300, n_clusters=5,
                                  n_init=10, n_jobs=None,
                                  precompute_distances='auto', random_state=0,
                                  tol=0.0001, verbose=0),
                 k=None, locate_elbow=True, metric='silhouette', timings=False)
```



Kelbow visualizer for the first iteration of monetary features.



```
KElbowVisualizer(ax=<AxesSubplot:>,
                 estimator=KMeans(algorithm='auto', copy_x=True,
                                  init='k-means++', max_iter=300, n_clusters=5,
                                  n_init=10, n_jobs=None,
                                  precompute_distances='auto', random_state=0,
                                  tol=0.0001, verbose=0),
                 k=None, locate_elbow=True, metric='silhouette', timings=False)
```

Kelbow visualizer for the second iteration of frequency related mixed features.

References

[1] Prof. Abhijit Sanyal (April 2021). EAI_6010_customer_data_exploratory_analysis_to_class.ipynb Retrieved from https://northeastern.instructure.com/courses/66677/modules

[2] Barış Karaman (towardsdatascienc.com) (04 May 2019). Customer Segmentation Retrieved from https://towardsdatascience.com/data-driven-growth-with-python-part-2-customer-segmentation-5c019d150444