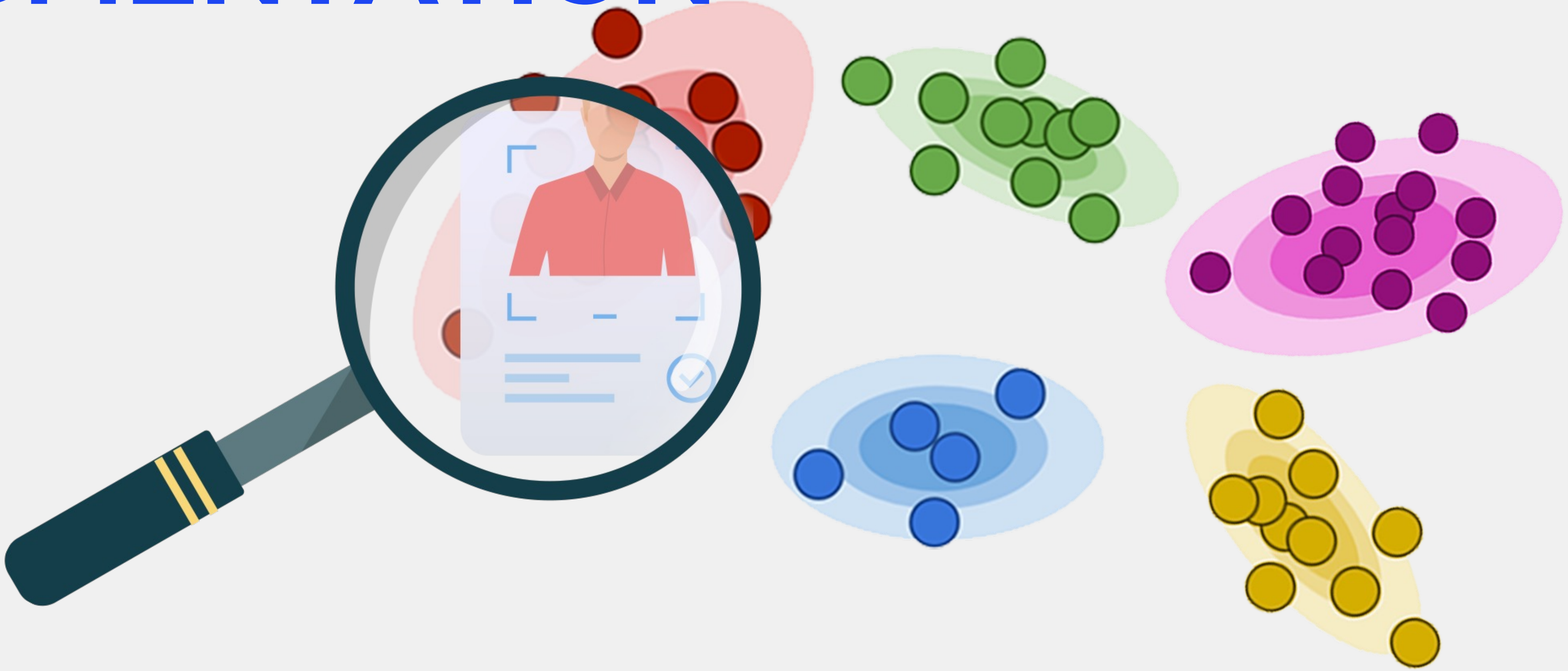
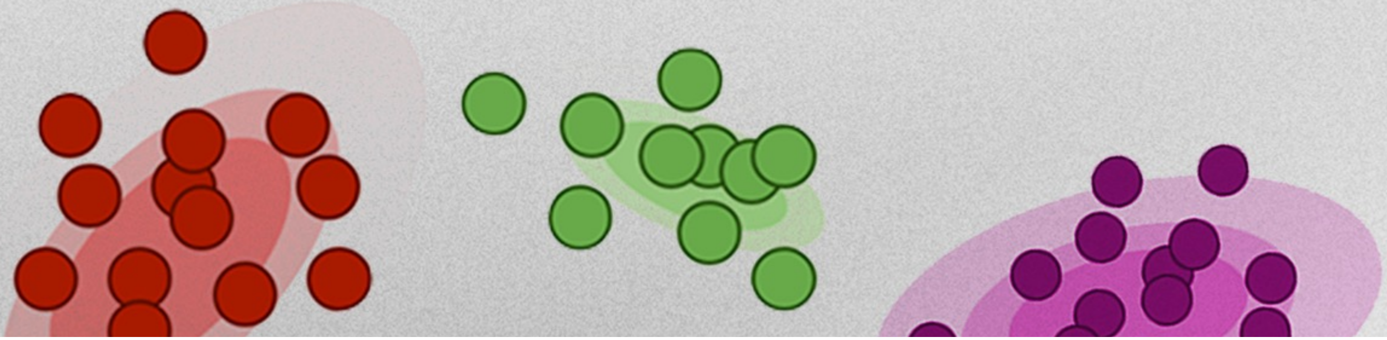
CLIENT SEGMENTATION

POLITECNICO MILANO 1863



Bressan Manuel - Maddaloni Federica - Malighetti Matteo Giuseppe - Manzoni Claudio - Recalcati Andrea

Fintech – A.Y. 21/22



Problem Description

From a dataset describing different features of **5000** people we want to create meaningful **financial personas** through methods of **clustering**.

Outline

1. Dataset Exploration
2. Measure Definition
3. Cluster Analysis
4. Cluster Interpretation

STEP 1 – DATASET EXPLORATION

5000 people identified by **12 numerical** features and **5 categorical** features

NUMERICAL + ORDINAL

Age	[19 – 95]	integer
FamilySize	[1 – 6]	integer
Income	[0,1]	continuous
Wealth	[0,1]	continuous
Debt	[0,1]	continuous
FinEdu	[0,1]	continuous
ESG	[0,1]	continuous
Digital	[0,1]	continuous
BankFriend	[0,1]	continuous
LifeStyle	[0,1]	continuous
Luxury	[0,1]	continuous
Saving	[0,1]	continuous

CATEGORICAL

Gender	{0, 1}
Job	{1, 2, 3, 4, 5}
Area	{1, 2, 3}
CitySize	{1, 2, 3}
Investments	{1, 2, 3}

STEP 1 – DATASET EXPLORATION

5000 people identified by **12 numerical** features and **5 categorical** features

NUMERICAL + ORDINAL

MINMAX SCALER

Age	**[0,1]**	**continuous**	**ESG**	[0,1]	continuous
FamilySize	**[0,1]**	**continuous**	**Digital**	[0,1]	continuous
Income	[0,1]	continuous	**BankFriend**	[0,1]	continuous
Wealth	[0,1]	continuous	**LifeStyle**	[0,1]	continuous
Debt	[0,1]	continuous	**Luxury**	[0,1]	continuous
FinEdu	[0,1]	continuous	**Saving**	[0,1]	continuous

CATEGORICAL – we don't drop the first column because we need it in the distance

ONE HOT ENCODING

Gender	**Gender_0**	**Gender_1**			
Job	**Job_1**	**Job_2**	**Job_3**	**Job_4**	**Job_5**
Area	**Area_1**	**Area_2**	**Area_2**		
CitySize	**CitySize_1**	**CitySize_2**	**CitySize_3**		
Investments	**Investments_1**	**Investments_2**	**Investments_3**		

STEP 1 – DATASET EXPLORATION

5000 people identified by **12 numerical** features and **5 categorical** features

NUMERICAL

	Age	FamilySize	Income	Wealth	Debt	FinEdu	ESG	Digital	BankFriend	LifeStyle	Luxury	Saving
0	0.065789	0.6	0.679599	0.705895	0.268264	0.770735	0.465122	0.718914	0.581720	0.612604	0.901051	0.293334
1	0.368421	0.0	0.873299	0.919090	0.747693	0.892883	0.521675	0.986877	0.778748	0.868977	0.917477	0.850925
2	0.250000	0.2	0.942846	0.902289	0.451701	0.504873	0.640388	0.772055	0.677446	0.761279	0.768338	0.521778
3	0.631579	0.4	0.548115	0.425051	0.614591	0.512343	0.518146	0.607305	0.648808	0.337033	0.519331	0.715921
4	0.184211	0.0	0.820609	0.734639	0.851100	0.889625	0.783674	0.730646	0.746853	0.915946	0.614119	0.637907

CATEGORICAL

	Job_1	Job_2	Job_3	Job_4	Job_5	Area_1	Area_2	Area_3	CitySize_1	CitySize_2	CitySize_3	Investments_1	Investments_2	Investments_3
0	1	0	0	0	0	0	1	0	0	1	0	1	0	0
1	0	1	0	0	0	0	1	0	0	0	1	0	0	1
2	0	1	0	0	0	1	0	0	0	1	0	0	0	1
3	0	1	0	0	0	1	0	0	0	1	0	0	1	0
4	0	1	0	0	0	1	0	0	0	0	1	0	1	0

STEP 2 – MEASURE

There are various clustering methods available:

- **partitioning** methods like k-means and k-medoids
- **density-based** clustering like DBScan
- **hierarchical** clustering

All of these methods rely on the definition of a proper **similarity function** to assess whether two elements are near or far from each other.

After numerous attempts with different similarity functions and different methods, we resorted to **K-medoids** using a personalized similarity function that we called **MixDistance_eucl.**

MixDistance_Eucl

The distances are computed separately: using the **Hamming distance** for the categorical features (which computes the number of elements that differ), and using the **Euclidean distance** for the numerical features. Then, the weights *wNum* and *wCat* are computed as the percentage of numerical and categorical variables respectively. The final customized distance is the linear combination of the two.

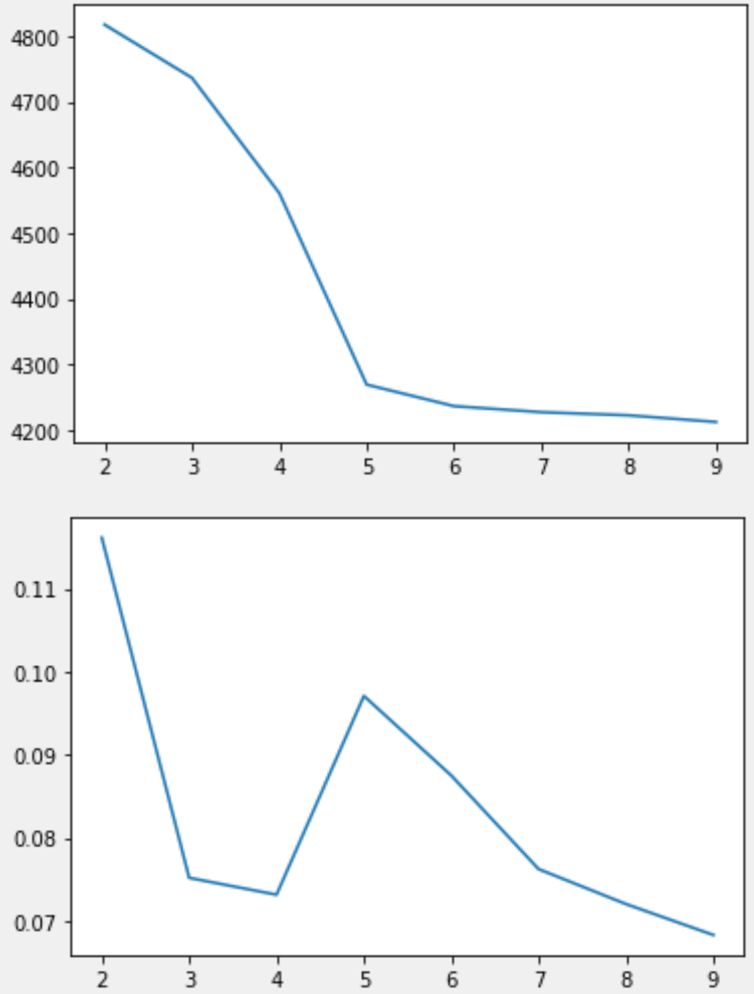
Numerical Features	OHE of Categorical Features

wNum * Euclidean distance $\in [0, \sqrt{12}]$ + wCat * Hamming distance $\in [0, 4]$

Elbow Analysis

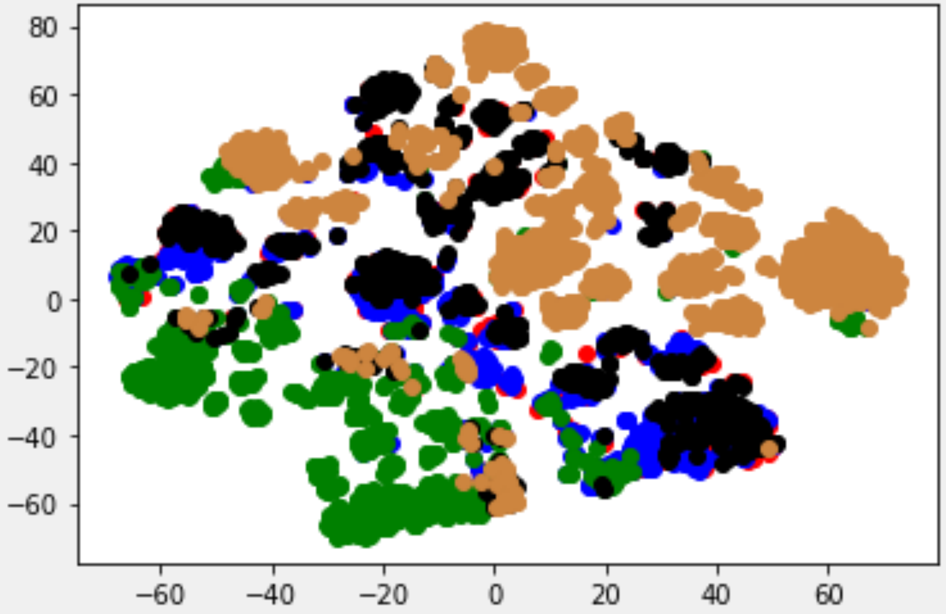
After some attempts, we decided to drop the feature *Gender* for the clustering, since in some of them it tended to create the clusters giving too much importance to this feature.

We performed the clustering thorugh **K-medoids** with k from 2 to 9, and obtained these graphs for the inertia (up) and the silhouette (down), which clearly suggest to use **k = 5** as the number of clusters.

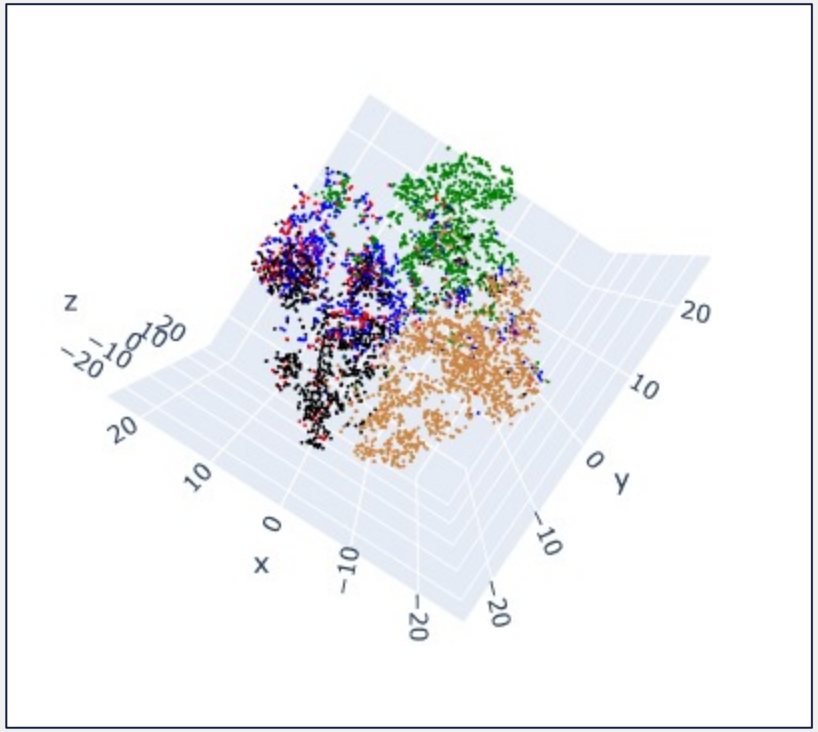


Cluster Visualization

t-SNE 2D



t-SNE 3D



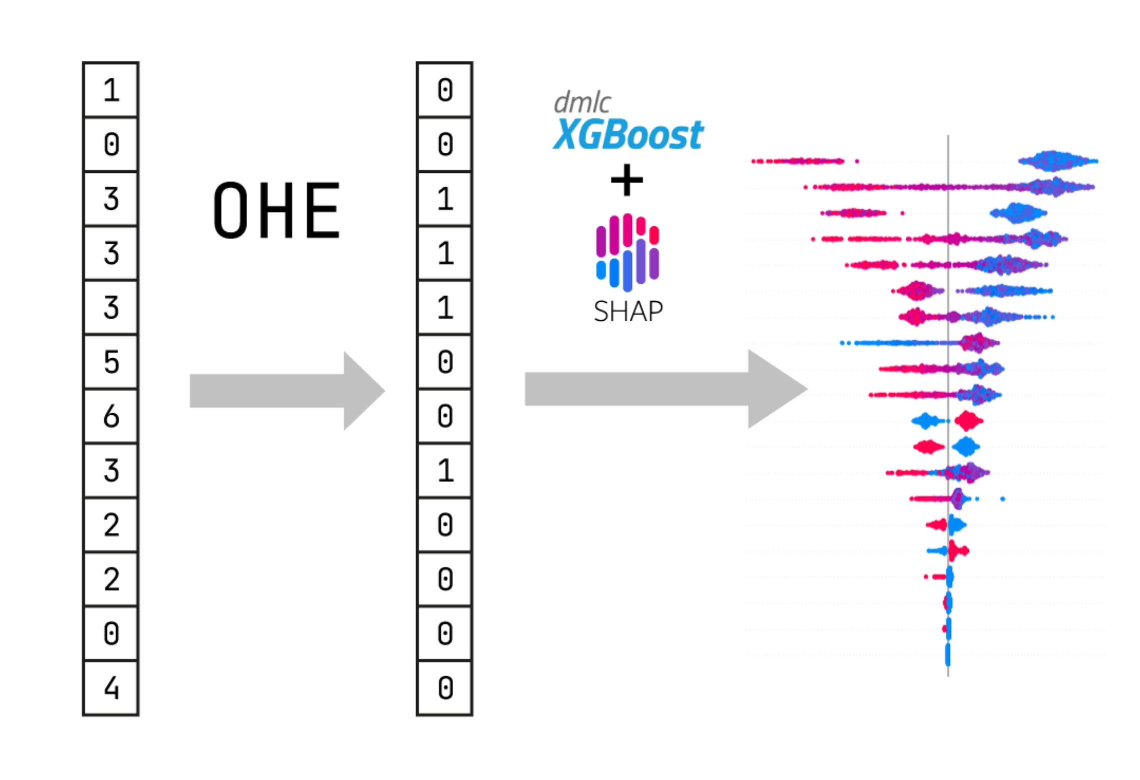
(3D plot on the code to appreciate better)

Classification

One-Hot-Encoding for every cluster label to have a **Binary Classification Problem** ($label = 1$ to the elements belonging to that cluster, $label = 0$ for all the others).

Then we apply **XGBoost** to this problem and analyze the interpretation of the model through:

- **Shap Values**,
- **feature importance**,
- **permutation tests** to state if a cluster has a different mean for a certain feature with respect to the whole dataset
- **histograms**.



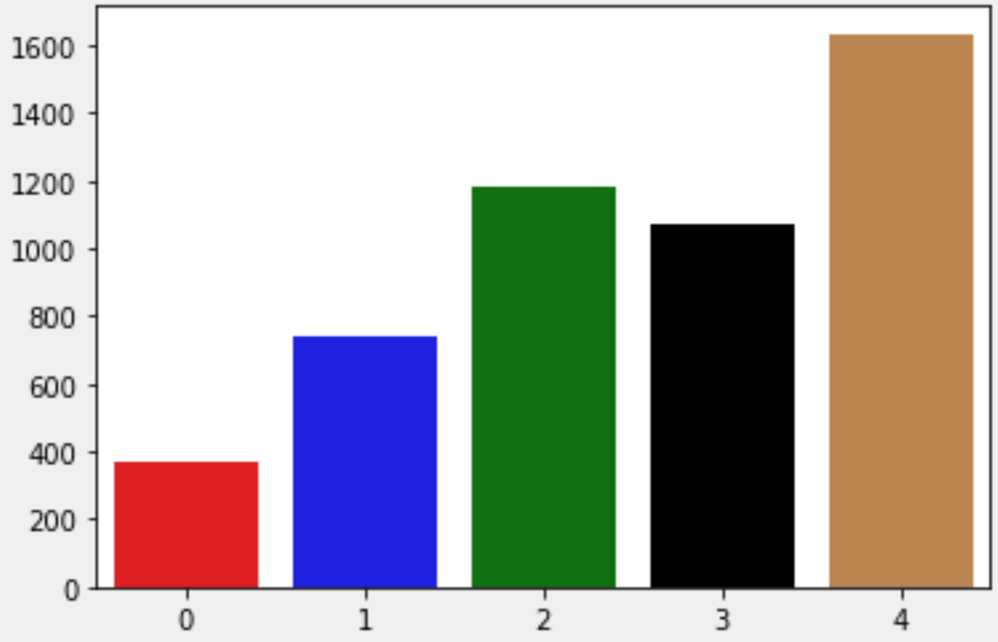
Permutation Test

A permutation test is a **statistical test** that requires no parametric assumption and relies on **permutations of the dataset** in order to test a hypothesis.

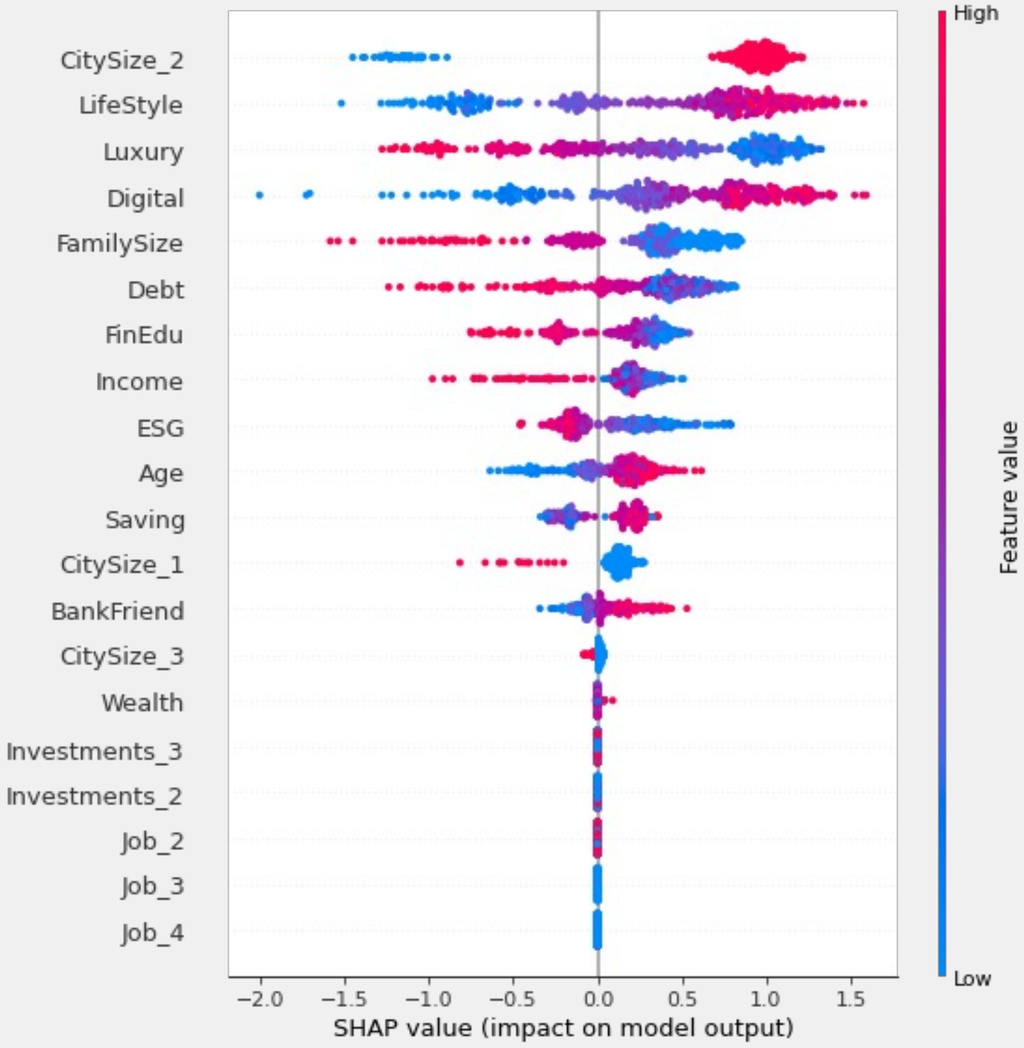
In our case, we structured a test to check if two groups have the same mean ($\boldsymbol{H_0} : \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$).

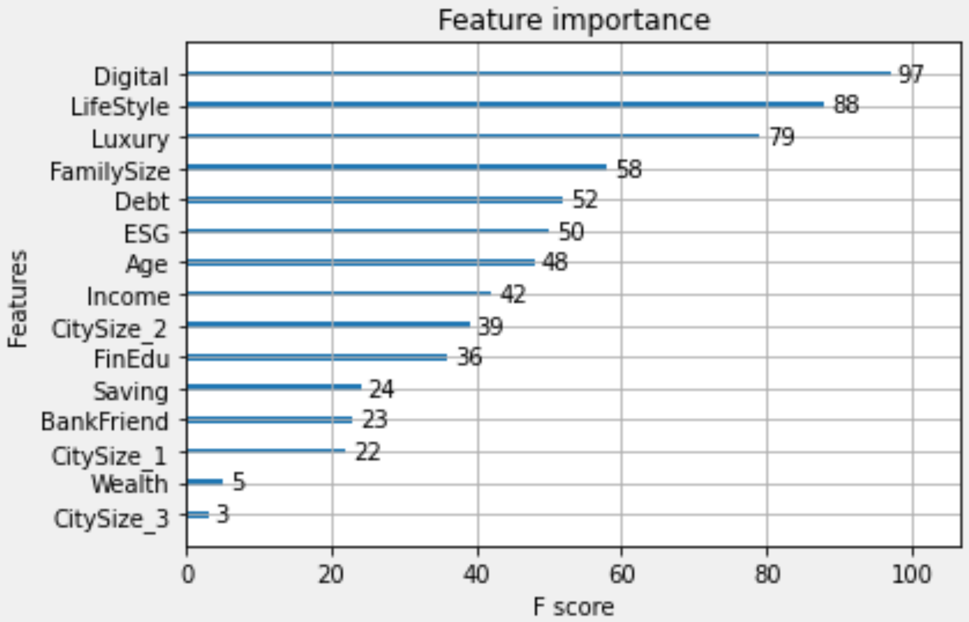
To do so, we merge the two groups and then sample again two groups with the same cardinalities randomly. After a fixed number of iterations (here **1000**), the ***p-value*** is computed as the ratio of times that we observed more extreme configurations of the dataset with respect to the one we actually observed. In this case, we consider it acceptable to reject the null hypothesis for p-values below **1%**.

Numerosity of the clusters



CLUSTER 0 – MIDDLE CLASS





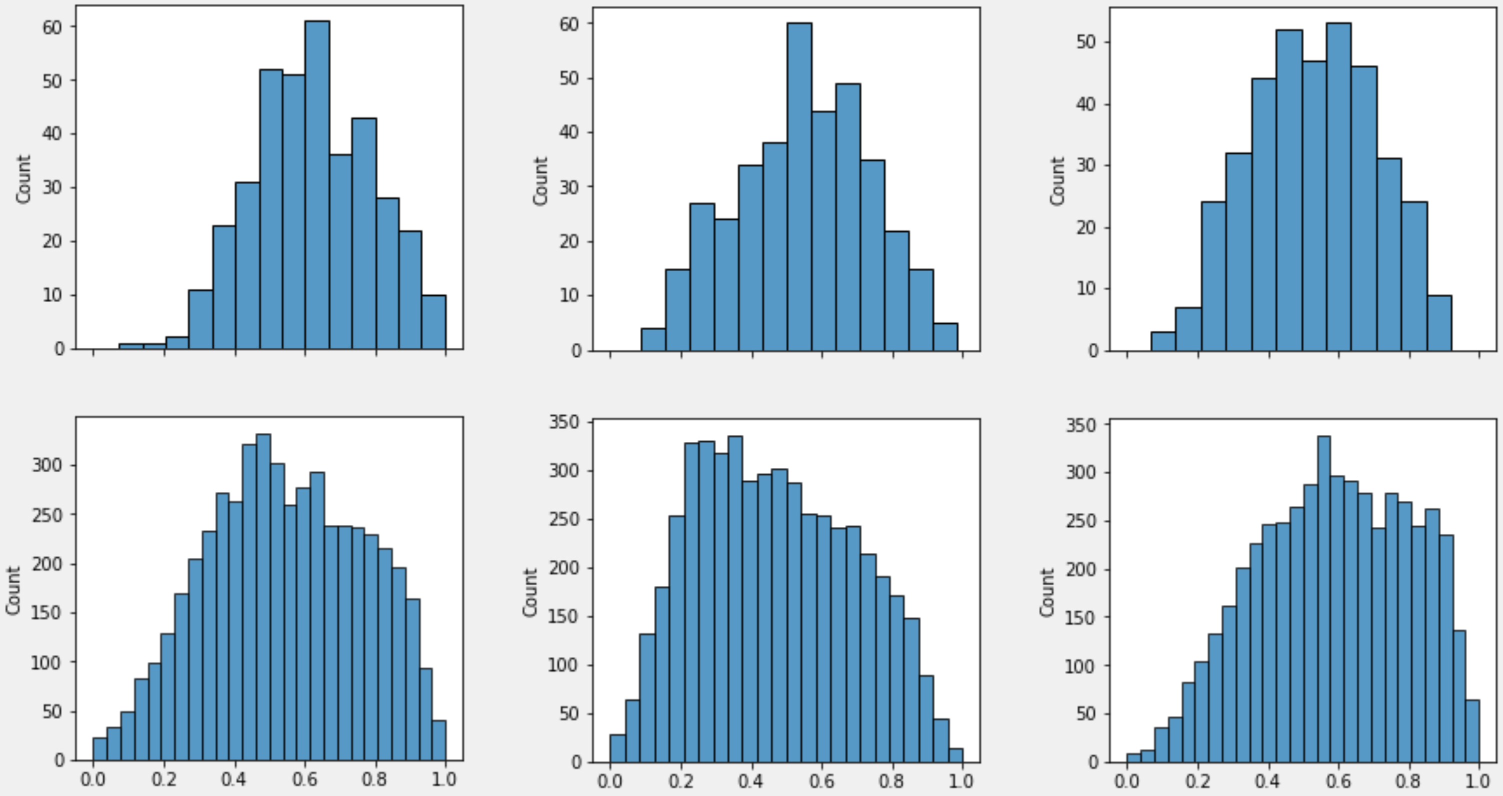
Both from the feature importance from the XGBoost and from the Shap values, we investigated more in depth the features *CitySize_2, LifeStyle, Luxury, Digital, Age.*

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 0 – MIDDLE CLASS

CLUSTER 0

DATASET



DIGITAL p_val = 0.0

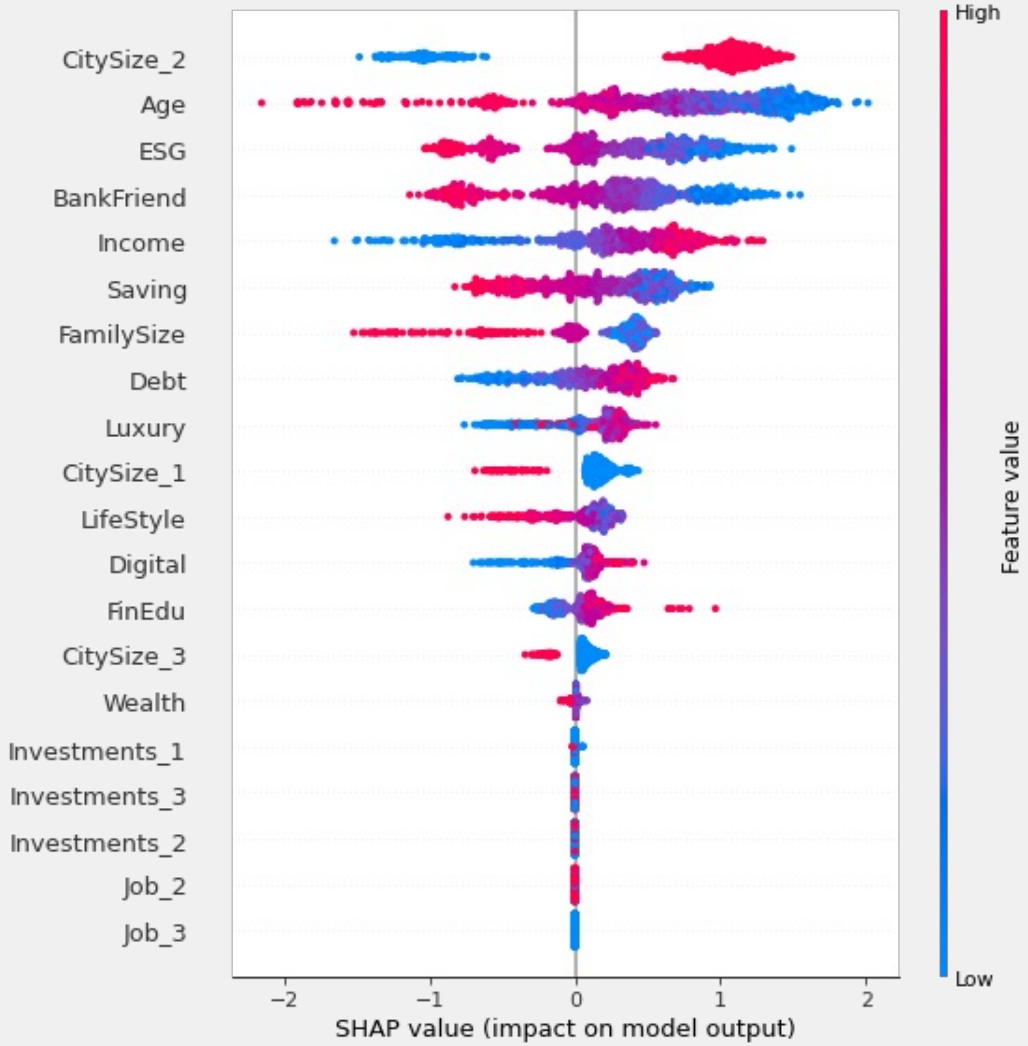
LIFESTYLE p_val = 0.0

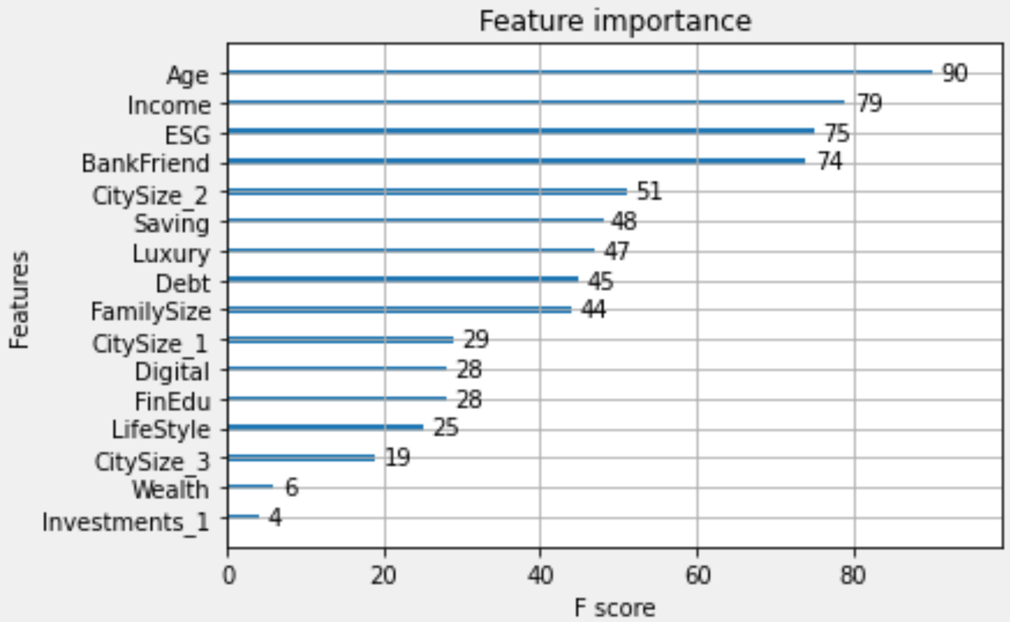
INCOME p_val = 0.0

This cluster can be represented by **middle-aged** people from **middle-sized cities**, with **medium incomes** (a bit below the average, but not low), who stay up-to-date being **digitalized** and more interested in their **lifestyles** than in luxury.

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 1 – YOUNG SPENDERS

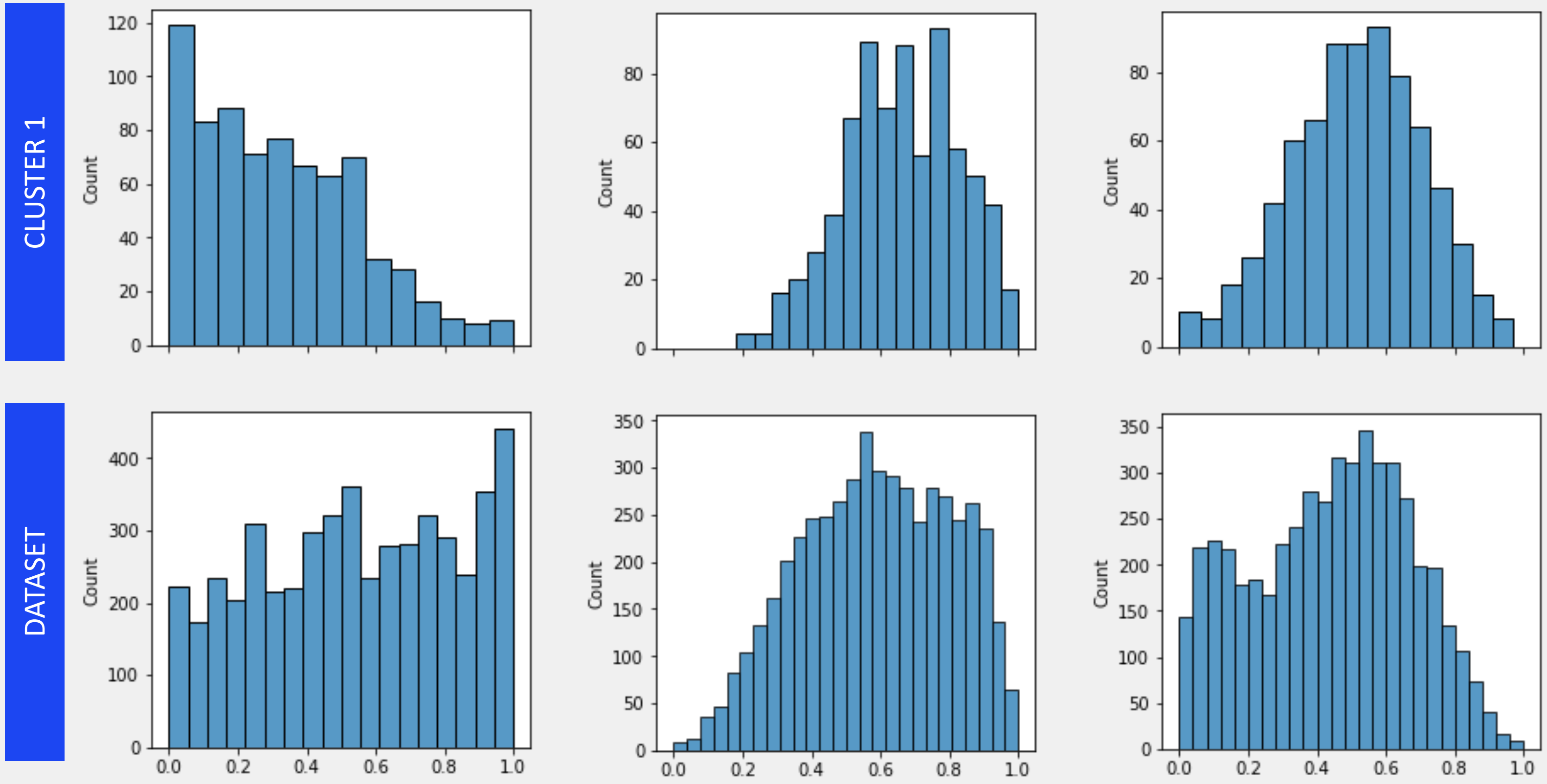




Both from the feature importance from the XGBoost and from the Shap values, we investigated more in depth the features *CitySize_2, Age, BankFriend, Income, Debt, FamilySize.*

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 1 – YOUNG SPENDERS



AGE

p_val = 0.0

INCOME

p_val = 0.0

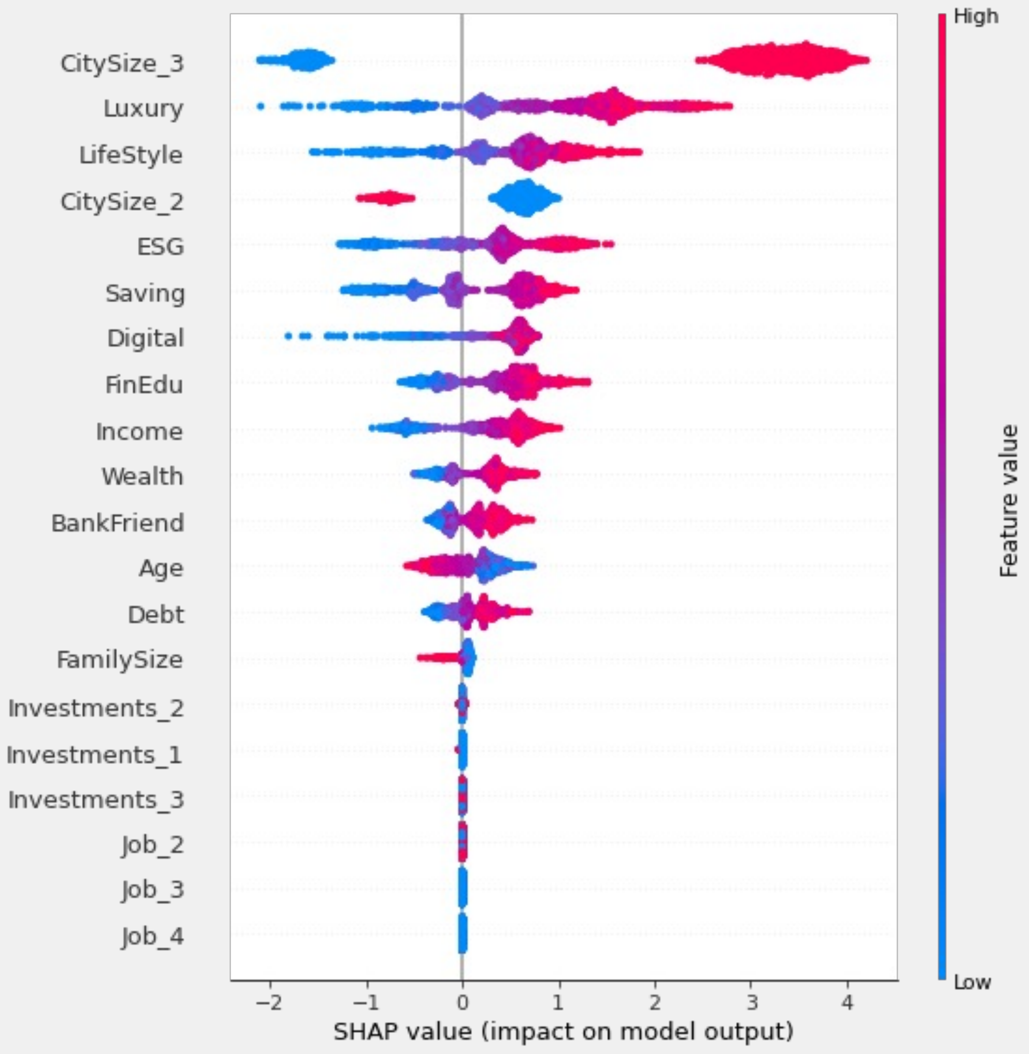
DEBT

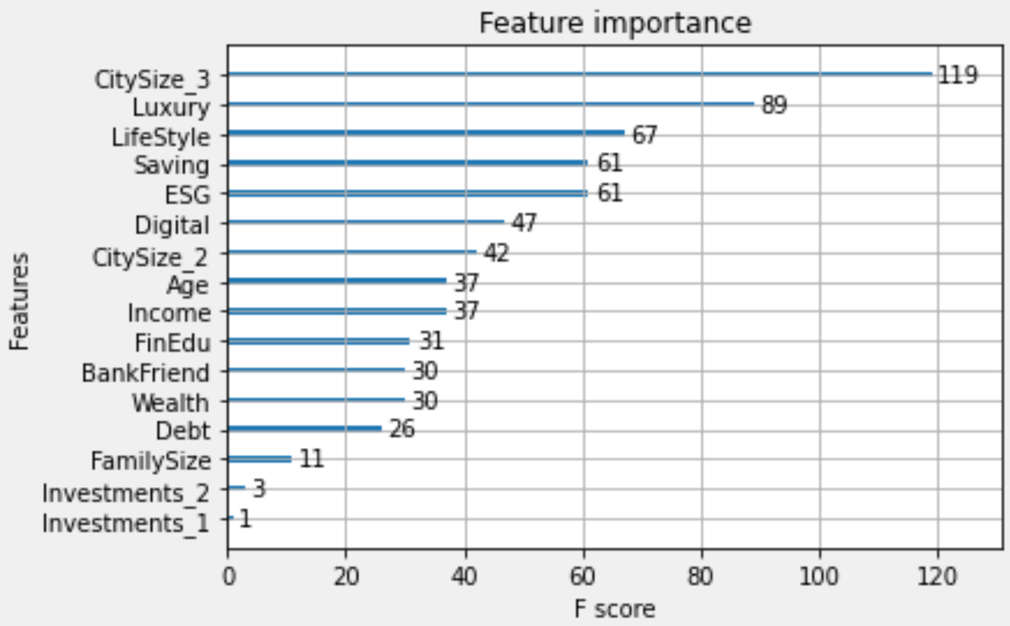
p_val = 0.0

This cluster can be represented by **young** people from **middle-sized cities**, with pretty **high incomes**, they don’t tend to save and are prone to **debt** (probably interested in leasings, not in luxury), still **without families**.

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 2 – ELITE

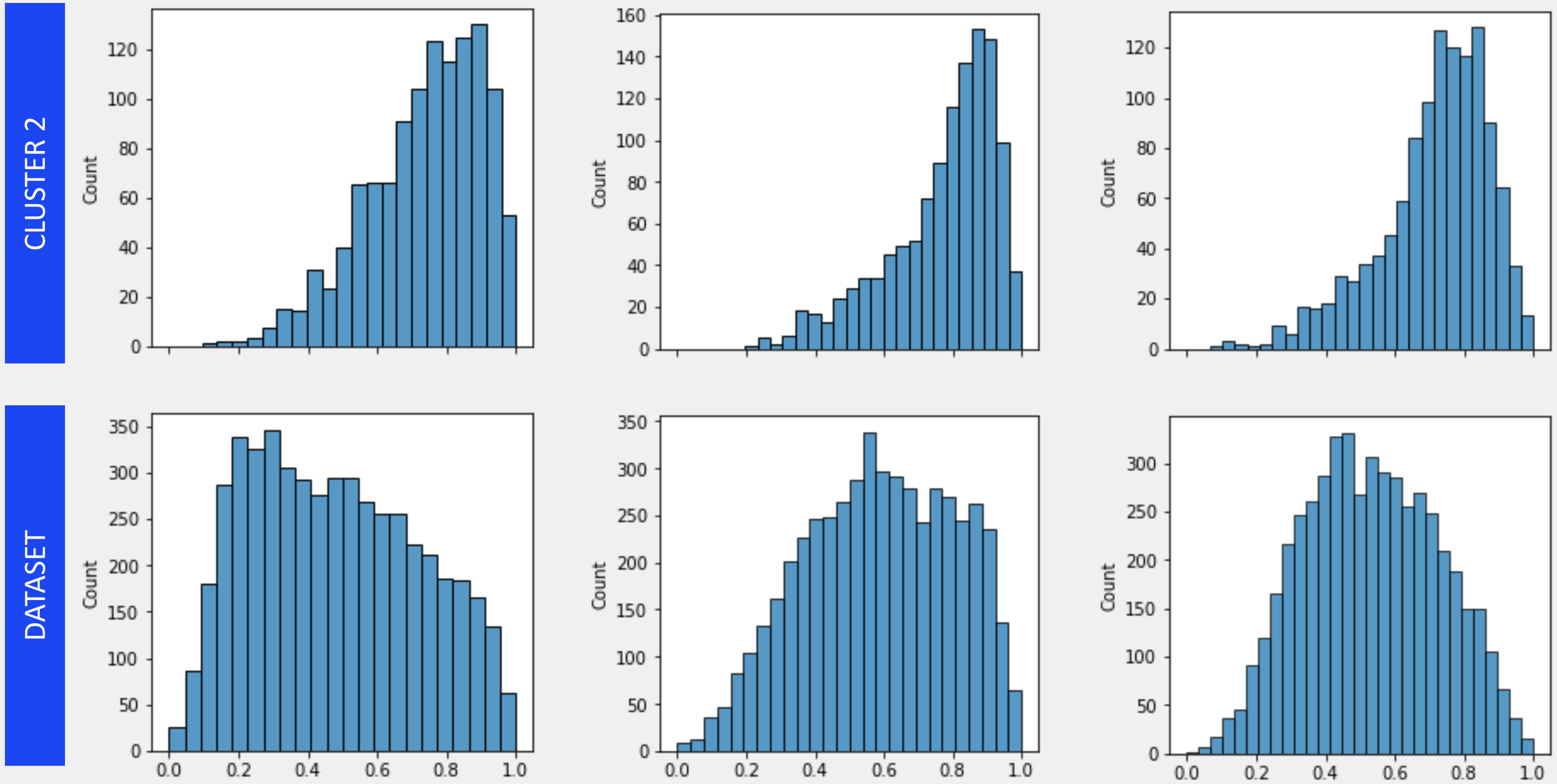




Both from the feature importance from the XGBoost and from the Shap values, we investigated more in depth the features *CitySize_3, Luxury, LifeStyle, ESG, FinEdu, Income.*

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 2 – ELITE



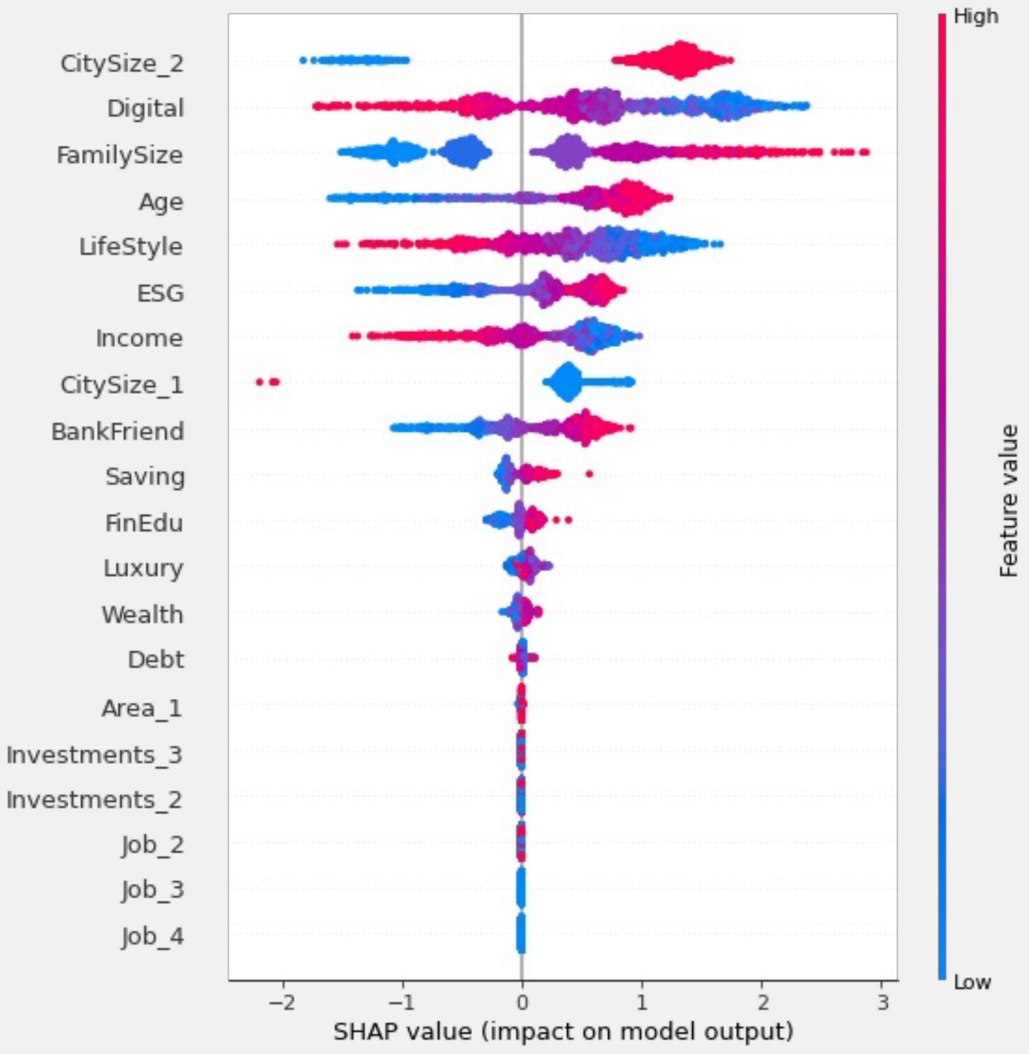
LUXURY INCOME FINEDU

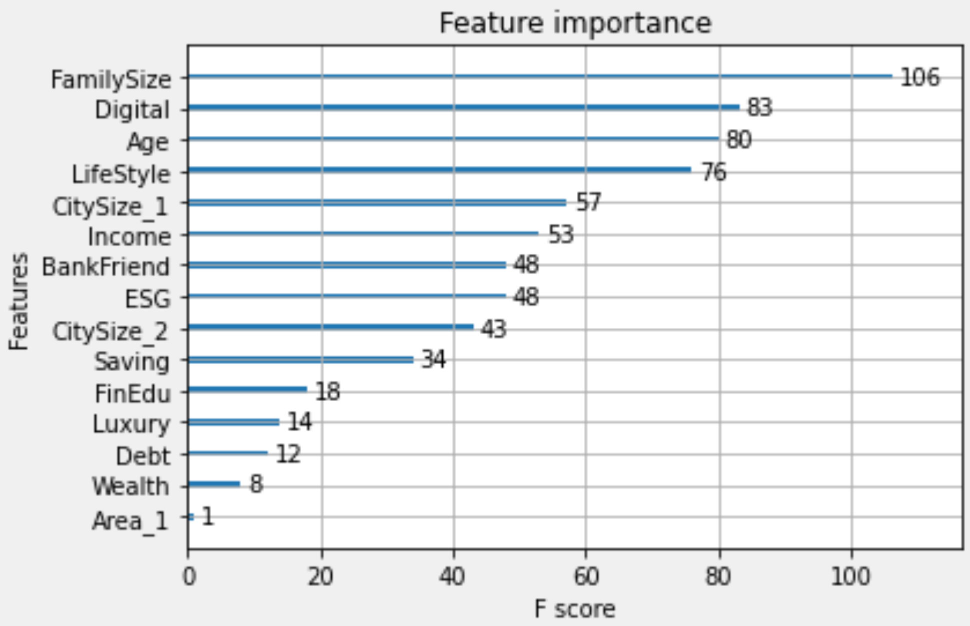
p_val = 0.0 p_val = 0.0 p_val = 0.0

This cluster can be represented by **rich** people from the **big cities**, with almost all the features way above the average: **higher income** and **wealth**, higher values of **lifestyle** and **luxury**, of **financial education** and **digitalization**, but also higher interest in **ESG** (Environment, Social and Governance)

STEP 4 – CLUSTER INTERPRETATION

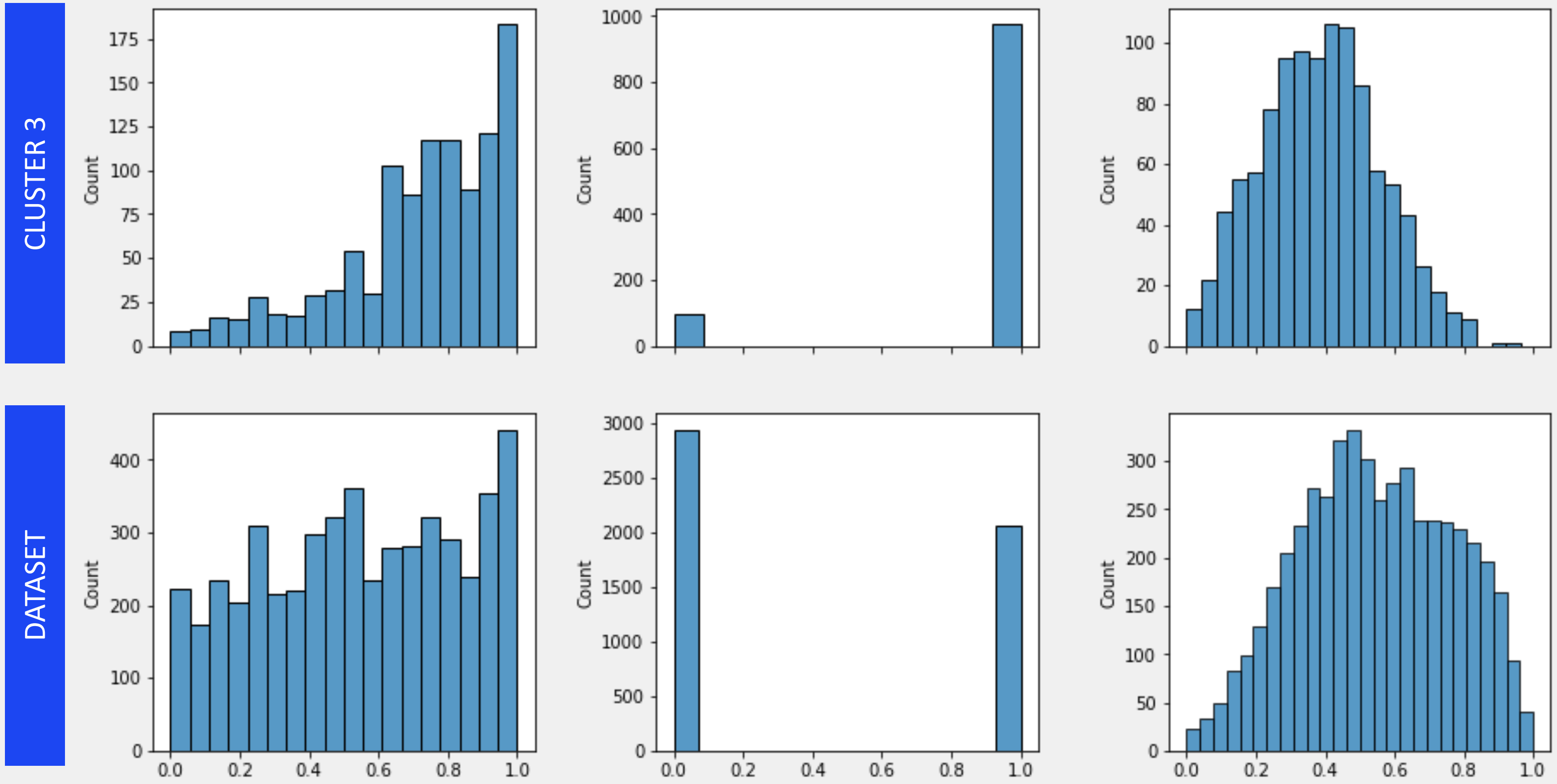
CLUSTER 3 – OLD PEOPLE FROM MEDIUM CITIES





Both from the feature importance from the XGBoost and from the Shap values, we investigated more in depth the features *CitySize_2, Digital, FamilySize, Age, Income.*

CLUSTER 3 – OLD PEOPLE FROM MEDIUM CITIES

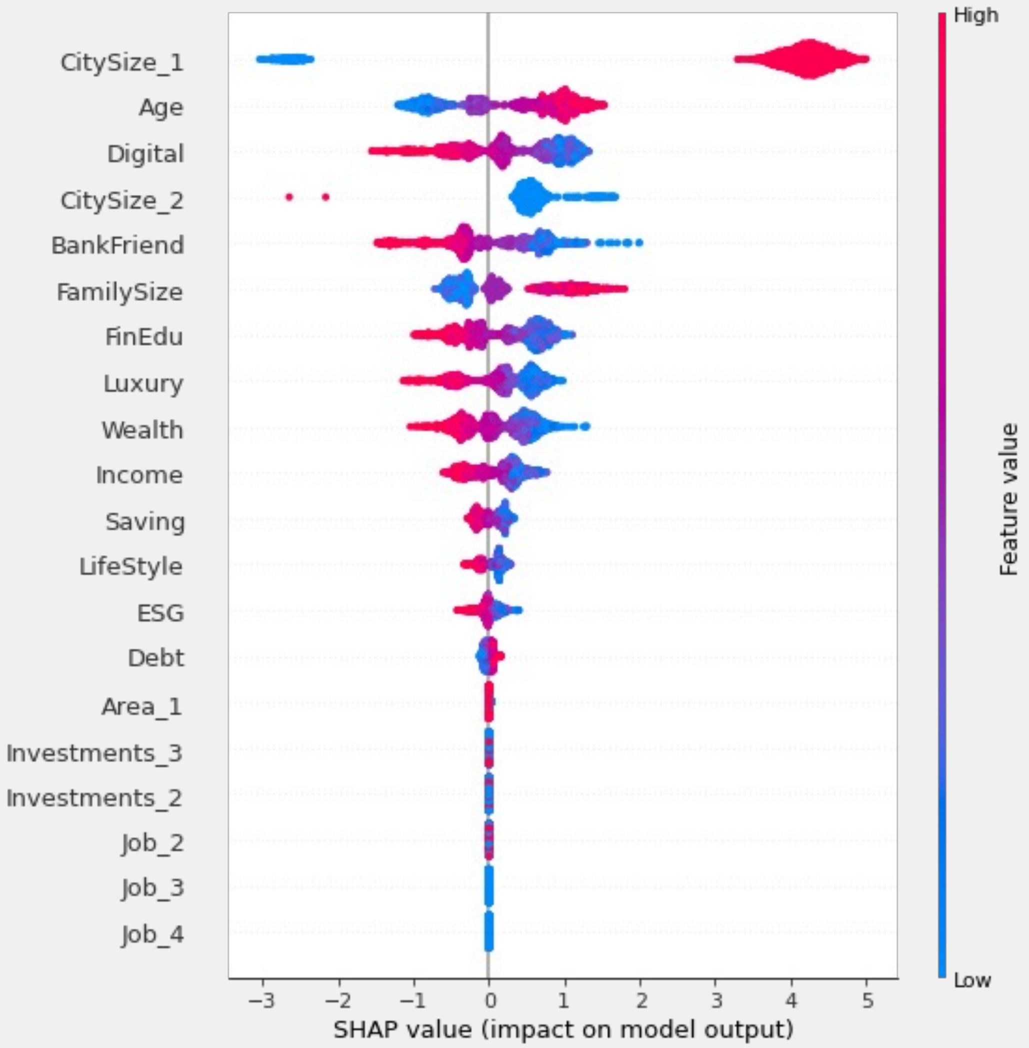


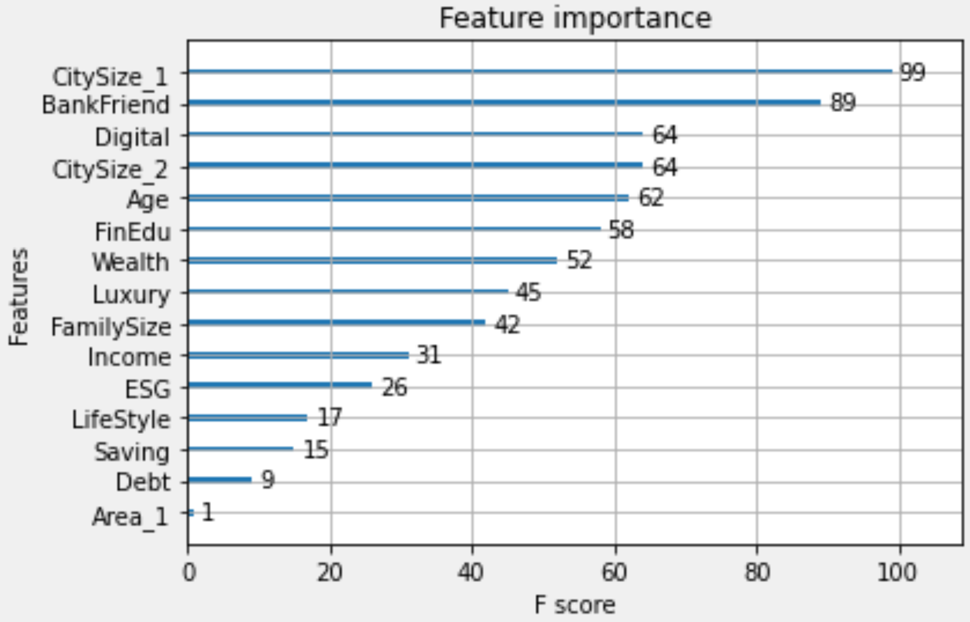
AGE	CITYSIZE_2	DIGITAL
p_val = 0.0	p_val = 0.0	p_val = 0.0

This cluster can be represented by **older** people from the **middle-sized cities**, typically with **families, not very digital** and with **lower incomes** with respect to the average and **lower** values of **lifestyle** and **luxury**.

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 4 – SMALL TOWNS

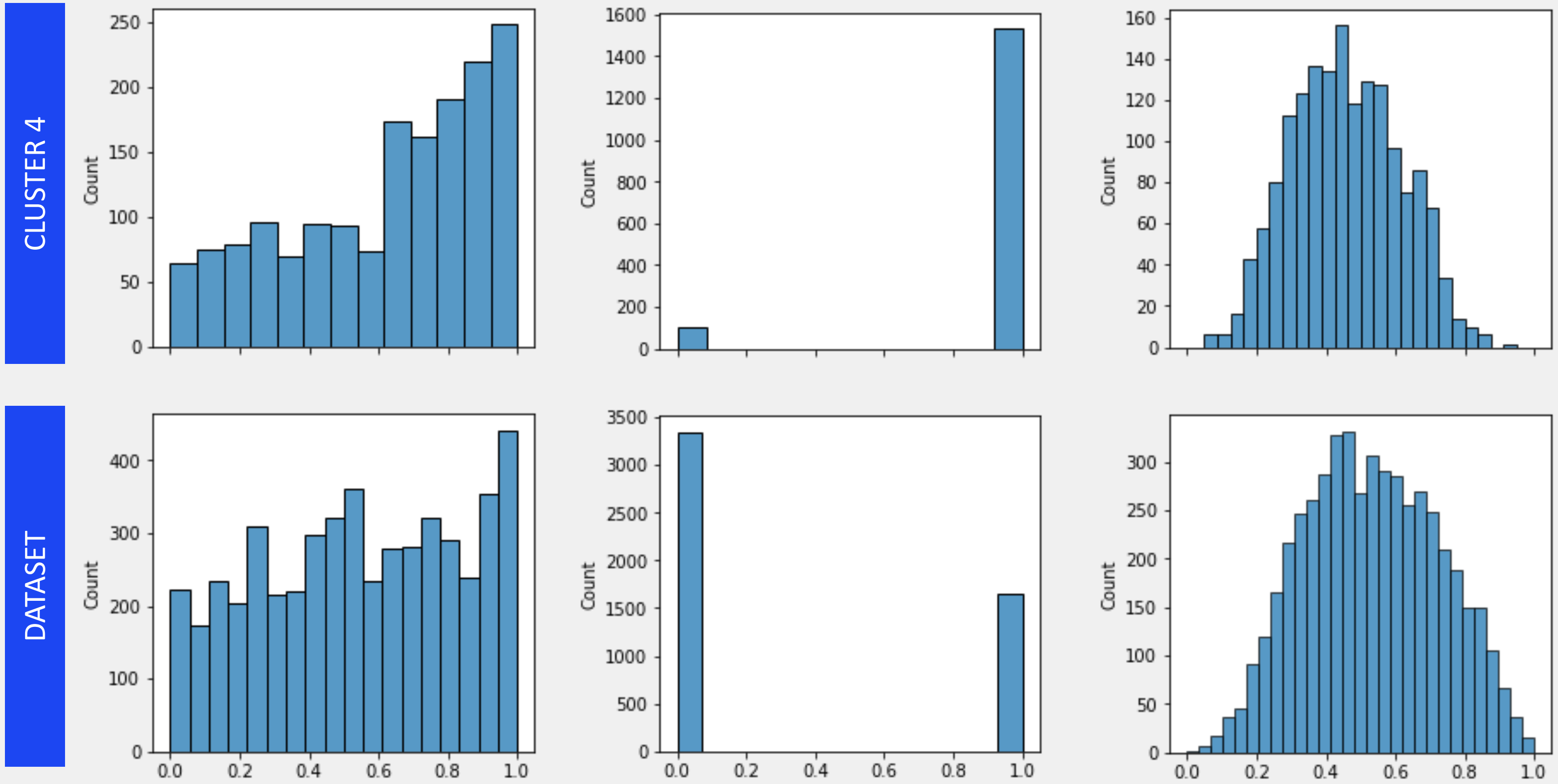




Both from the feature importance from the XGBoost and from the Shap values, we investigated more in depth the features *CitySize_1, Age, BankFriend, Digital, FinEdu.*

STEP 4 – CLUSTER INTERPRETATION

CLUSTER 4 – SMALL TOWNS



AGE p_val = 0.0

CITYSIZE_1 p_val = 0.0

FINEDU p_val = 0.0

This cluster can be represented by **older** people from **small towns**, **not very digital** (in particular, few people with high values of this feature, the majority with medium values) and **not very financially educated**. **Not** very interested in **luxury** and with **lower incomes** than the average. It is the **largest** cluster.