Target Marketing Campaign Using Different Data Mining Algorithms - PVA Donors

Ana Sofia Silva
m20200220@novaims.unl.com
Bussiness Analytics

José Francisco Alves
m20200653@novaims.unl.com
Business Analytics

Miguel Nunes
m20200615@novaims.unl.com
Business Analytics

Abstract

This project aims to implement unsupervised models in order to discover patterns and information that was previously undetected. That is, to intuitively fish out the potentially relevant attributes we need to go through the data dictionary to better understand how donors behave and to identify the different segments of potential donors within their database with the aim of developing Customer Segmentation.

The given dataframe has a high volume of donors information's (features). In order to create meaningful segments of the given dataframe, a set of steps were handled to extract, transform, load and analyse the sample. After the data cleaning and analysis, the selected variables were divided according to its nature, it was created two segments: Social and Value. After the data partition, it was implemented K-Means for metrical features and K-Modes for categorical. The structure of this report is divided in Introduction, Data Preparation, Clustering, Cluster Profiling and Marketing approach.

1. Introduction

The main goal of this project is to analyse a sample of the results, containing 95 412 donors, looking for a Customer Segmentation applied for the different segments of donors. It's expected to explain the selected clusters, approaches that has been taken and advantages/disadvantages of our decisions.

Our group assumed to select 43 features because of operational considerations, where we select intuitively the most critical ones to our project objective. This approach was also adopted due to the issue of creating Customer Segmentation.Then, after proceeding with the selection we begin the missing values treatment and the outliers analysis. Then, was decided to do some transformations, dealing with some features, and divide the attributes into ’Value’ and ’Social’ subsets. For each subset, we applied clustering models such as K-means and K-modes and hierarchical clustering.

The applied models were selected according to the variables input and output types. Most of the theoretical support for the used segmentation models was presented in the Data Mining class.

The produced python script for this project is organized in sections with the aim of support this report.

2. Data Preparation

Data preparation is the process of taking raw data and getting it ready for an analytical study. To achieve the final stage of preparation, the data must be cleaned, formatted, and transformed into something suitable by analytics tools.

2.1 Feature Selection

At the beginning, when we looked to our dataframe we thought that it would make sense to divide in ’Social’ and ’Values’ features. In order to reduce our initial huge amount of those features it was necessary to understand and select the crucial attributes. The selecting process was based on choose intuitively relevant variables with the aim of being propitious to our case study. All the attributes chosen are present in attachment 7.

2.2 Missing values treatment

Missing values treatment is a routine present in Data Cleaning, process that remove incomplete data and ensure quality on analysis.

In order to treat, we make the percentage of missing values in each column and select the ones with ’NaN’ above 40%. Therefore, features ’PVSTATE’, ’VETERANS’ and ’PCOWNERS’ were dropped due to more than 89% of missing data.

In respect to ’MAJOR’ and ’PEPSTRFL’ variable, we assume that ’_’ character and blank cell, represents that does not match with Major Donor and is not considered to be a PEP Star, respectively, so these values were assigned as 0.

In concern to 'WEALTH2' and 'TIMELAG', missing values were filled with median. The reason why we choose the median method it was due to the fact of these method being more robust to the presence of outliers and the distributions of these features being characterised by skewed. About 'HOMEOWNR', 'GENDER' and 'DOMAIN' features, those values were filled with mode method. This process was chosen just because these 3 variables are categorical. For 'INCOME' variable the remaining method chose was filling the missing values with the mean, since we were not able to identify outliers and the distribution doesn't seem to be skewed.

2.3 Outlier Analysis

We divided the outliers analysis into two subsections: univariate and multivariate analysis.

2.3.1 Univariate Analysis

A univariate outlier is a data point that consists of an extreme value on one variable. A column 'Outliers_Uni' was created and the analysis of the outliers with a threshold of 5 was done to locate themselves. In these column, the existence of an outlier was represented by '1'. Our group used a threshold of 5 to have a more robust analysis.

By looking at the boxplots (Figure 1), according to the outliers in the features 'ETH7','ETH10','ETH11','AFC1', 'AFC2', 'AFC3' due to the high number of outliers, the features will be eliminated from the dataset.

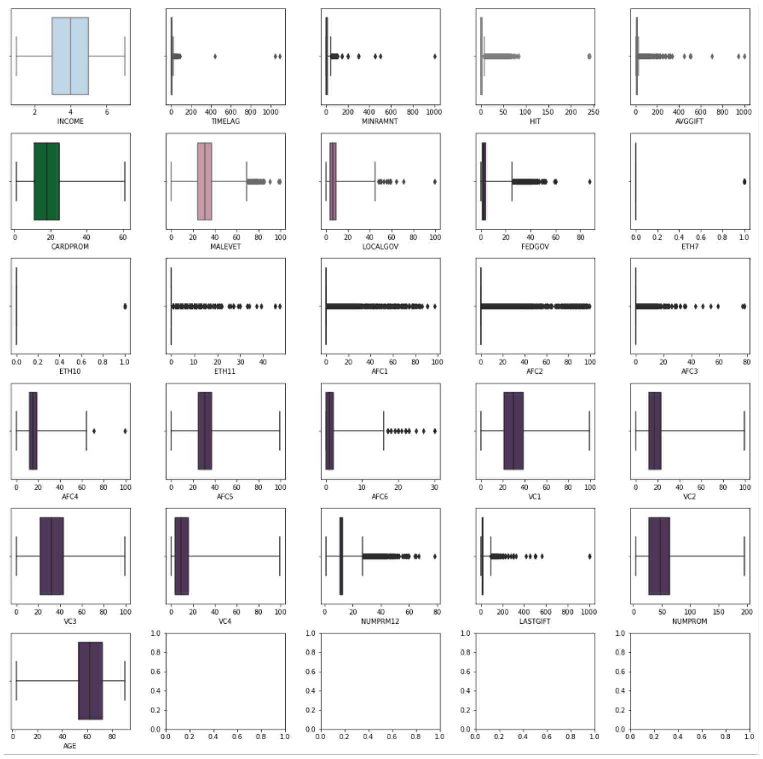


Figure 1: Matrix plot for Boxplot visualizations of each variable.

On the other hand, by analysing Figure 2, we realised that 'TIMELAG', 'MINRAMNT', 'HIT' and 'AVGGIFT' showed us a highly narrow distribution, and 'AFC5', 'VC1', 'VC2' and 'VC3' represented a more dispersed distribution and it's close to the Gaussian.

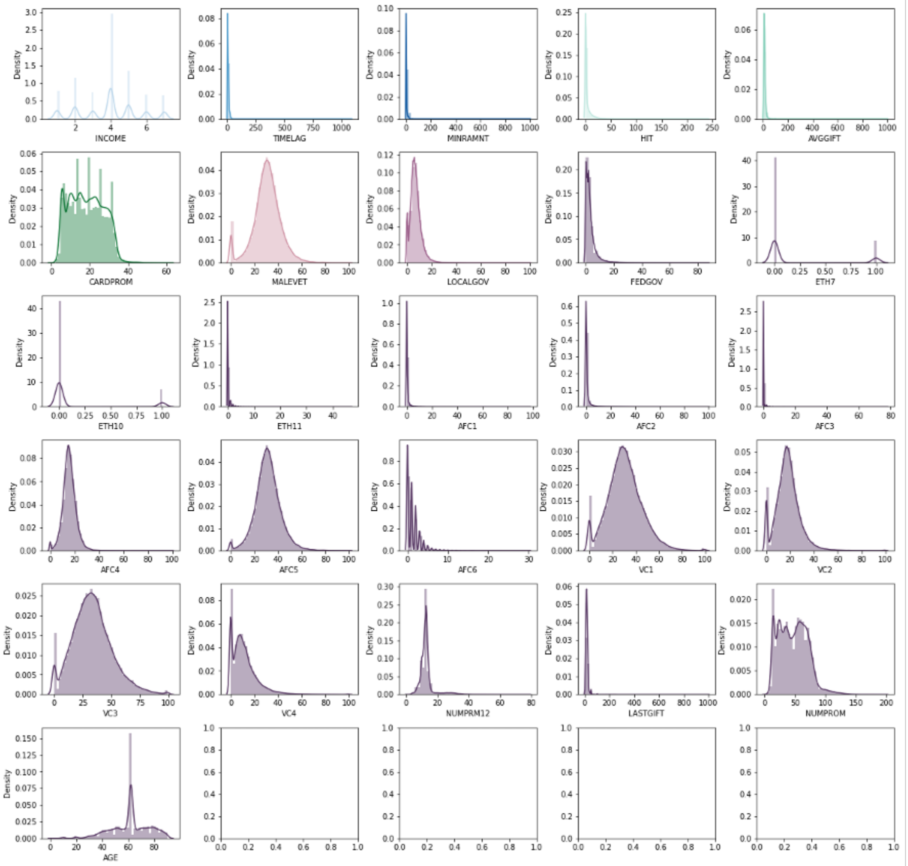


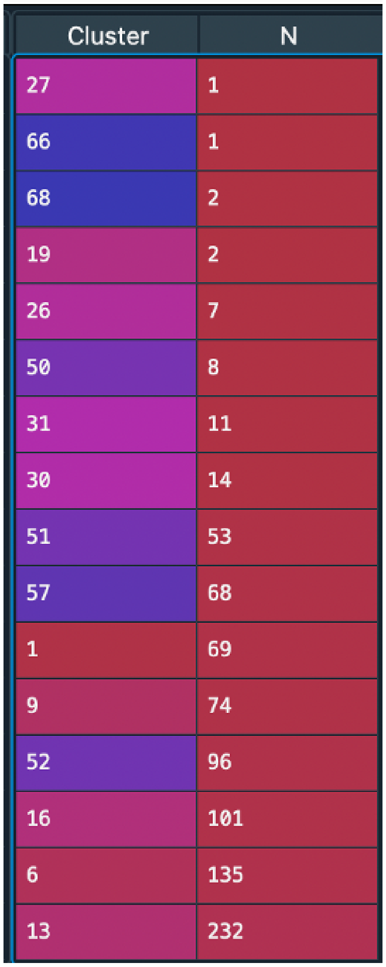
Figure 2: Matrix plot for Histogram visualizations.

2.3.2 Multivariate Analysis

A multivariate outlier is a combination of unusual scores on at least two variables. Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters“. Intuitively, values that fall outside of the set of clusters may be considered outliers.

In order to have a more robust analysis, we performed outliers identification through k-means with the same group of numerical variables but in this case standardized, so they can be comparable. With this approach, the main goal is to produce k-means with a larger number of clusters (k=70) so we can identify small clusters in terms of size which may contain outliers.

We consider multivariate outliers, the clusters with less than 200 observations.



Cluster	N
27	1
66	1
68	2
19	2
26	7
50	8
31	11
30	14
51	53
57	68
1	69
9	74
52	96
16	101
6	135
13	232

Figure 3: Number of observations per cluster.

- The column ’Cluster’, ’N’ is merged into the non_categorical dataframe and $'N' <= 14$.
- Creation of $'Outlier_Multi' = 0$ and 1, if it has outlier according column ‘N’
- For clusters with $N <= 200$–$'Cluster_Multi' = 1$

2.4 Transform features

After dealing with all these preparation, it’s time to transform some features in order to the data be wrought.

Initially, ’HOMEOWNR’, ’MAJOR’ and ’PEPSTRFL’ were transformed to binary features, due to the fact that they enable us to use a single regression equation to represent multiple groups. This means that we don’t need to write out separate equation models for each subgroup. On the other hand, 'GENDER' variable had 4 different values ('U', 'J', 'A', 'C') where the meaning were quite similar - 'Unknown Gender' - therefore was created a separated value as 'Others' ('O') to replace them.

Once 'LASTDATE' and 'MAXADATE' were 'object' type features it was crucial convert them to 'date time' to establish a meaning that makes sense for our analysis.

Before doing the fill missing values step, was crucial to create a new variable 'Age' from 'DOB'. Instead of having the date of birthday (DOB) of the donor it was better transform it into age giving a more directly answer for our analysis.

Our group found it appropriate to create dummy variables only for 1 byte of 'MDMAUD' and 'DOMAIN'. The reason of these action was because it was a categorical variables that could take several values.

'MDMAUD' was transformed into 11 new features according to the 'MDMAUD_R', 'MDMAUD_F' and 'MDMAUD_A'. These new features are represented by ordinal values because of significance order.

'DOMAIN' was reduced to only the first byte (urbanicity level of the donor's neighborhood) and each of the features, 'DOMAIN_U', 'DOMAIN_S', 'DOMAIN_T', 'DOMAIN_C' and 'DOMAIN_R', are represented by a column. All columns has values '1' due to absence of significance order.

On 'GENDER' feature was created only 2 columns for 'MALE' and 'OTHERS' where female ('F') is assumed to be part of 'MALE' column. This column is represented by binary values.

After all these transformations, variable 'MDMAUD', 'DOMAIN' and 'GENDER' are dropped. Finally we proceed to a concatenate between non categorical and categorical dataframes and dropped all the unnecessary attributes.

2.5 Data Partition and Standardization

Before starting the segmentation, it was fundamental to divide the dataframe into two partitions of variables. The variables were divided in the 'Value' variables which are described by the ones that gives value to the dataframe, having all the same target (Income), and the 'Social' variables which describes every characteristic of every observation of the dataframe.

This type of division can prevent the loss of information caused by the overlap of variables, this is, when one set of variables take over the remaining ones in the analysis. Therefore, this approach will provide us a more meticulous segmentation and consequently a well define marketing strategy.

After doing the partition phase, it was important to start with the correlation analysis. Our group applied Spearman correlation matrix for each partition (Figure 10 and

11) to evaluate the correlation between variables. Spearman method was chosen due to the fact that variables are not normally distributed and the relationship between the variables is not linear.

Firstly, taking into account Spearman correlation coefficients, irrelevant variables were excluded due to high and low correlations. Secondly, in order to understand if the sample division by clusters was significant and to build suitable clusters for segmentation analysis, attempts were made to find the variables that best discriminate the sample. For the 'Value' partition, we decided to remove NUMPROM and MDMAUD_A_C due to higher correlations with other features. Features HIT and TIMELAG were removed because they didn't have any significant correlation with other variables.

In the 'Social' partition, there were features also highly correlated, for example 'AFC5' with 'MALEVET' and 'AFC5' with 'AFC4'.

Correlation matrices were created, not only to visualise the existence of relationships between variables, but also to understand which features were more appropriate to doing the clustering.

To finish the Data Preparation step, we finally conclude with the data standardization. This process was made with StandardScaler function in order to rescaling distribution of values so that the mean of observed values is 0 and the standard deviation is 1.

3. Clustering

In this section we will apply some clustering techniques, such as k-means followed by hierarchical clustering, Self-Organizing Maps (SOM) succeeded by k-means or hierarchical clustering, k-modes and lastly density-based clustering (DBSCAN).

3.1 K-Means clustering followed by hierarchical clustering

Taking into account that our main goal is customer segmentation, we started to perform the k-means clustering. K-means objective is to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is, the objective aims for high intracluster similarity and low intercluster similarity.

Considering k-means algorithm weaknesses, defining the initial number of clusters is one of them. Therefore, in order to find the appropriate number of clusters we perform the partitioning algorithm (k-means) with a large number of clusters (k=100) and then we apply a hierarchical clustering based on the dendrogram.

To decide the number of clusters to retain through dendrogram, our group chose 'ward' method. This method tends to result in, approximately, equal groups due to its minimisation of internal variation.

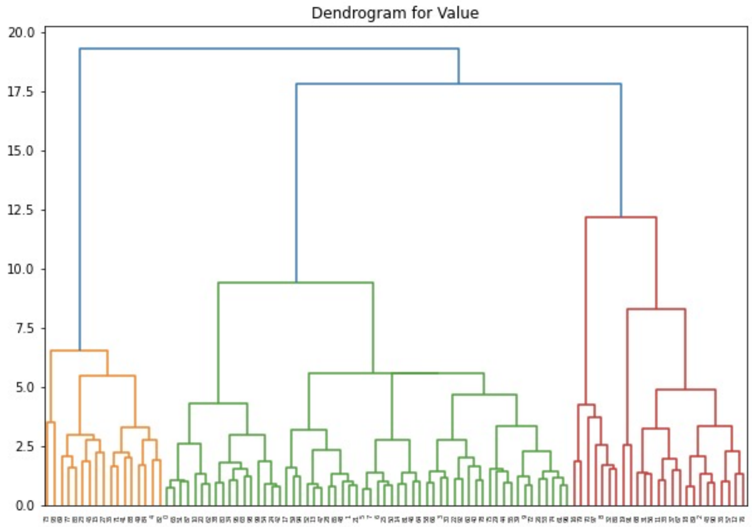


Figure 4: 'Value' dendrogram.

After plotting the dendrogram (Figure 4) for 'Value' partition, we thought that would be wisely to choose 3 clusters to apply hierarchical clustering. On the other hand, for 'Social' analysis, we also choose 3 clusters (Figure 6).

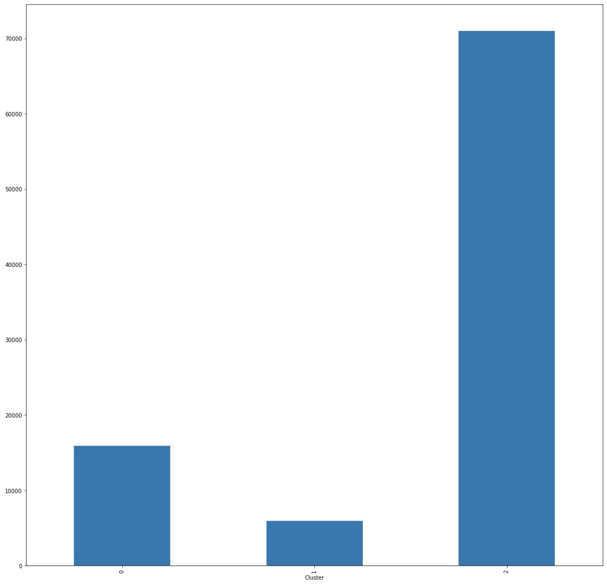


Figure 5: Number of observations per cluster in 'Value' partition.

After getting the number of observations per cluster, we observed that cluster have a huge discrepancy of observations (Figure 5). The principle reason of this behaviour can be seen on the dendrongram, where we have cluster with very few centroids comparing with others. These cluster will have less donors for customer segmentation.

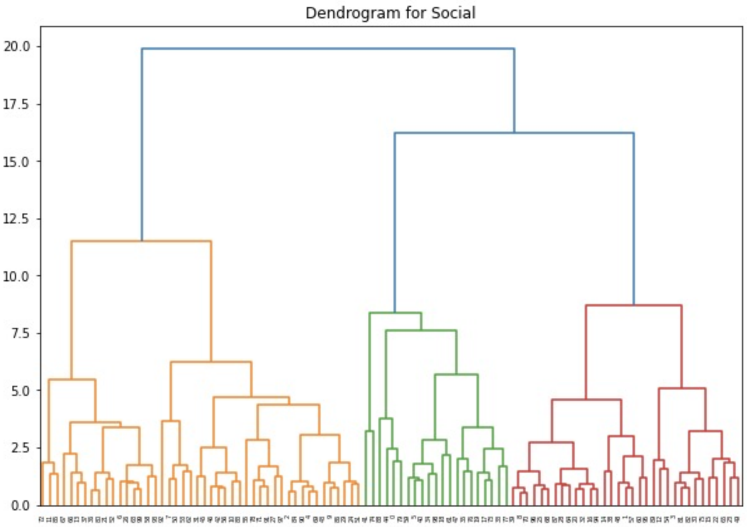


Figure 6: 'Social' dendrogram.

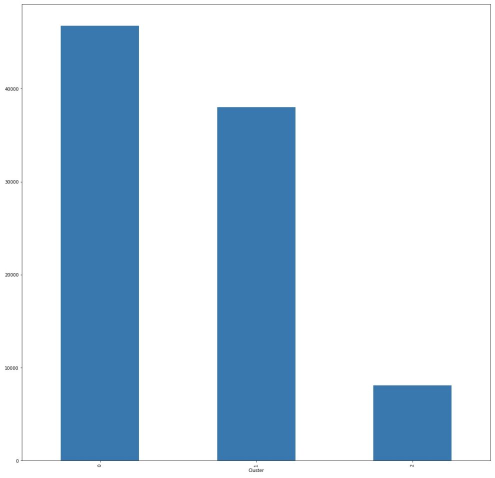


Figure 7: Number of observations per cluster in 'Social' partition.

In this partition we observed that clusters have more distributed observations comparing with 'Value' partition (Figure 7). The frequency table and the dendrogram show this.

3.2 K-Modes clustering

The k-modes clustering algorithm supports the categorical data objects. It selects the initial centroids randomly from the given data objects. Due its randomness in its selection of initial centroids, it provides the local optimum solution. Since, the k-Modes algorithm comes from the k-means algorithm, it can also be treated as an optimisation problem.

In this section we will include only the 'Social' variables because the analysis of the categorical variables for 'Value' was not conclusive.

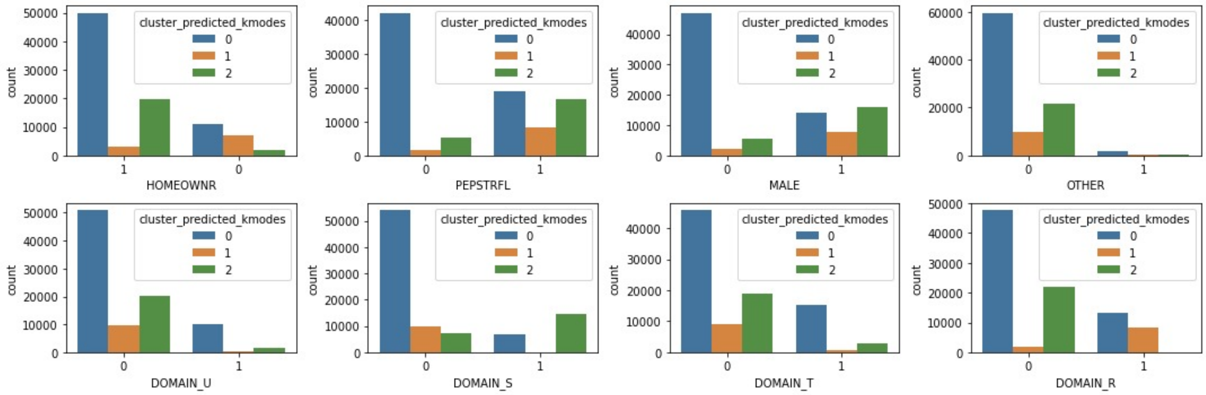


Figure 8: K-modes for 'Social'.

4. Cluster Profiling

Cluster profiling section is crucial to be able to create oriented marketing approaches. Here we will differentiate each cluster according to its characteristics and, consequently, name them.

Below its represented the polar graphic of K-Means clustering of each partition. This type of visualisation is important to measure the distribution in the different variables. One simple way to center an image is to use the adjustbox package with the export option. It provides the center=¡width¿ key to which centers the image around the given width. It defaults to the so use:

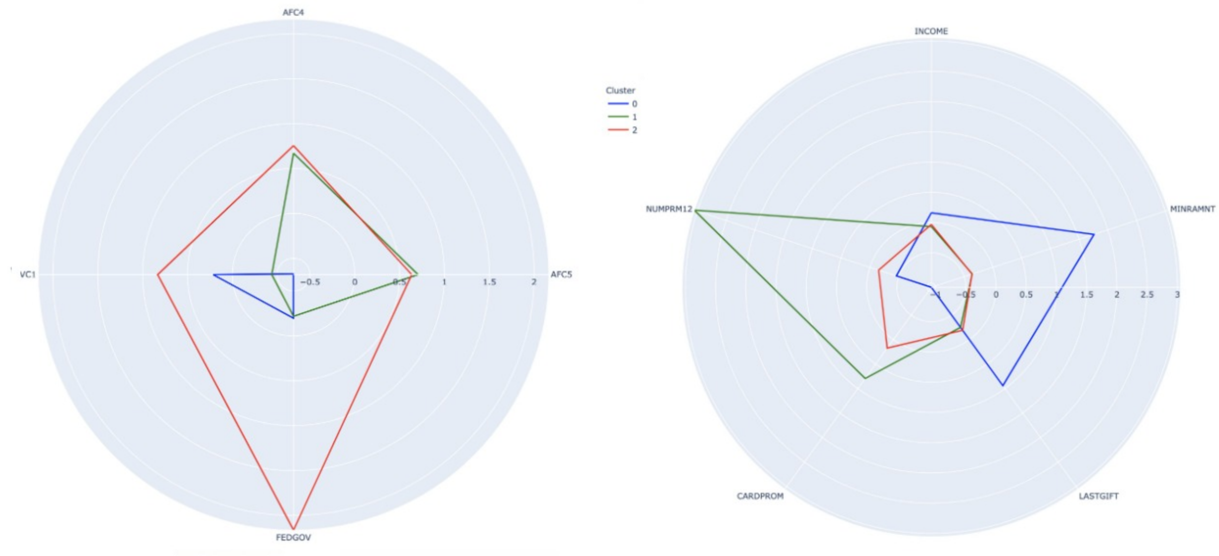


Figure 9: Centroids distribution for 'Social' (left) and 'Value' (right).

4.1 'Value' partition

Cluster 1 is mostly composed by donors who received promotions in the last 12 months but only few of them give a recent gift. Considering this, is named as **unfriendly donors**.

Cluster 2 corresponds to the most balanced one between all the variables. There is a slightly pick on CARDPROM which is related to the possession of promotions cards until the date. This factor can be related to a small percentage of the sample whose has a medium income. Furthermore, there's a balanced distribution of the number of samples in this cluster. Is named as **balanced donors**

Lastly, for **Cluster 0** we can say that gives small gifts however, some of them are recent. Besides that, doesn't have any card promotion and very few promotions received in the last 12 months. With that being said, is named as **small donation donors**

4.2 'Social' partition

Cluster 2 is donors employed by federal government and quite them Vietnam males adults veterans. For 'VC1', 'AFC4' and 'AFC5' cluster is equally distributed, that is, this 3 variables aren't appropriate to describe the sample. Also, this is the largest cluster. Is named as **Federal gov. donors**.

Cluster 1 is mostly represented by male and adult veterans (age +16). Besides that, this cluster doesn't have Vietnam donors and it's the second large cluster in 'Social' partition. Thus is named as **Male adult donors**.

Finally, **Cluster 0** is the one that has very few centroids and are composed by Vietnam veterans donors employed by federal government. Curiously doesn't have any adult and any male veteran. This is named as **Federal gov. Vietnam veterans donors**

5. Marketing Approach

After proceeding with the cluster section, it's fundamental define the best marketing approach for each cluster, that is, for each of our donors. With these approaches we can apply ideas that best fit each of the segments.

5.1 'Value' partition

In general, 2 of the 3 clusters have an opposite distribution given that there's 2 distinct picks in 'MINRAMNT' and 'NUMPRM12' with less pronounced in ' LASTGIFT' and 'CARDPROM'. 'INCOME' is the less covered variable by the 3 clusters.

5.1.1 Unfriendly donors

To reach a new target audience, awareness actions can be promoted door-to-door, in order to make the association problems known. For this target it would be wise to increase the promotion to enforce the donors to give bigger gifts more recently. Extend promotions plan is a good approach, in order to influence the donation and draw the attention.

5.1.2 Balanced donors

For this cluster, the campaigns applied in Unfriendly donors and Small donation donors can be also applied in this segment due to its equal distribution in the variables.

5.1.3 Small donation donors

If this donors are characterised by give recent gifts We should create a marketing approach heavily based on increasing the promotions and give cards to the donors in order to improve recency donations.

5.2 'Social' partition

5.2.1 Federal gov. donors

For federal government employees could be provide grocery discounts if they provide a regular donation per month . If they continue to donate each month for 12 months, are able to win a gift basket. Also, federal government employees would benefit from medical taxes.

5.2.2 Male adult donors

Given that this cluster is represented by veterans and assuming that they usually frequent medical services, we promote awareness events among the population in hospitals, pharmacies and health centers.

5.2.3 Federal gov. Vietnam veterans donors

Applied to Vietnam veterans donors they would received a distinction for their military services during the through a lunch hosted by PVA donors for veterans. This lunch has a associated cost that will be reverted to Paralysed Veterans of America.

6. Conclusion

From the beginning of this project until the very end, it is possible to say that all the procedures that was needed to be done, in order to extract the best segmentation for all the donors that we have in our dataset, was very challenging.

The first step and probably the most challenging was to understand our problem, analyse it and prepare it. This phase was crucial for all the project because that defined our goal and all the outcome. After the analyse of the data, we decided to divide all the features in two different groups (Social and Value). After that, we had special attention with all the missing value per feature in order to remove incomplete data and ensure quality on analysis. Furthermore, there was an outlier treatment with univariate and multivariate analysis. At last, we needed to do a partition and data standardization to determine the correlation matrices for each groups that we had (Social and Value).

Subsequently we used the algorithms of K-means, Hierarchical Clustering and K-modes to choose the best possible segmentation, that had the most interpretable and coherent information to be analysed in discussed on our next step.

Finally, the section of cluster profiling was crucial to make a good and oriented marketing approach. With the representation of the centroids of each partition was possible to understand better each feature per partition (Social and Value).

With all the analyse that have been done, there was proposed some promotions or initiatives in order to increase substantially the number and the value of donations received as soon as possible. It is also very important to identify new target audiences to preserve the future of the association.

7. Attachment

7.1

The most important features selected were:

INCOME – Household income

TIMELAG – Number of months between first and second gift

MINRAMNT – Dollar amount of smallest gift to date

HIT - Indicates total number of known times the donor responded to a mail order offer other than PVA's

AVGGIFT – Average dollar amount of gifts to date

CARDPROM – Lifetime number of card promotions received to date

CARDGIFT - Number of lifetime gifts to card promotions to date

PVASTATE – Indicates whether the donor lives in a state serviced by the organization's EPVA chapter

MALEVET - % Male Veterans

LOCALGOV - % Employed by Local Gov

STATEGOV - % Employed by State Gov

FEDGOV - % Employed by Fed Gov

HOMEOWNR – Homeowner flag

WEALTH2 – Wealth rating

GENDER

MAJOR - Major ($$) Donor Flag _ = Not a Major Donor X = Major Donor

VETERANS - VETERANS (Y/N)

PEPSTRFL - Indicates PEP Star RFA Status blank = Not considered to be a PEP Star 'X' = Has PEP Star RFA Status

ETH7 - Percent Japanese

ETH10 - Percent Korean

ETH11 - Percent Vietnamese

Neighborhood:

AFC1 - Percent Adults in Active Military Service

AFC2 - Percent Males in Active Military Service

AFC3 - Percent Females in Active Military Service

AFC4 - Percent Adult Veterans Age 16+

AFC5 - Percent Male Veterans Age 16+

AFC6 - Percent Female Veterans Age 16+

VC1 - Percent Vietnam Veterans Age 16+

VC2 - Percent Korean Veterans Age 16+

VC3 - Percent WW2 Veterans Age 16+

VC4 - Percent Veterans Serving After May 1995 Only

NUMPRM12 - Number of promotions received in the last 12 months (in terms of calendar months translates into 9603-9702)

NUMPROM – Lifetime number of promotions received to date

PCOWNERS - Home PC owners/users

LASTGIFT - Dollar amount of most recent gift

MDMAUD_R - Recency code for MDMAUD

MDMAUD_F - Frequency code for MDMAUD

MDMAUD_A - Donation Amount code for MDMAUD

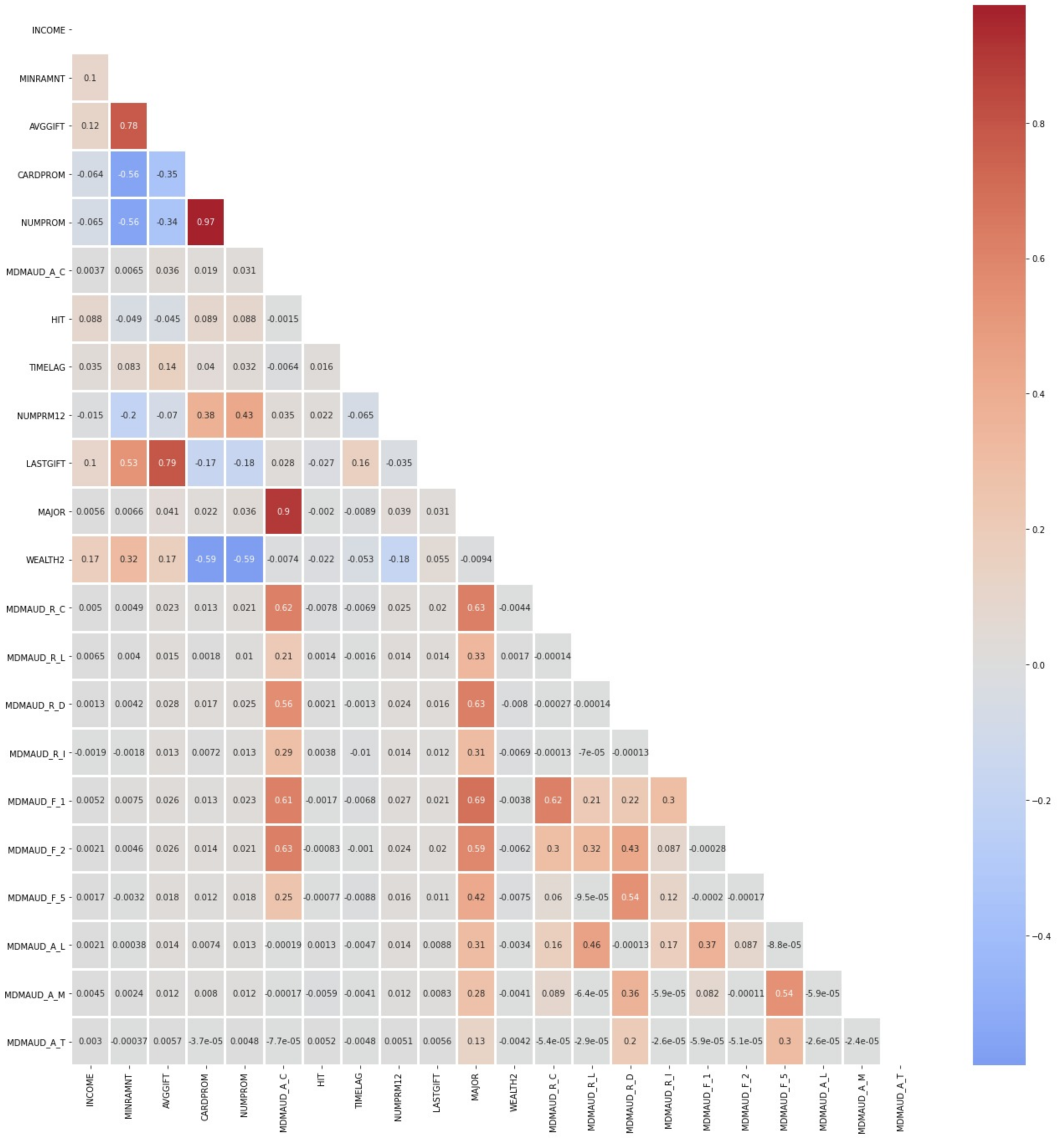
MDMAUD – Major donor matrix

LASTDATE – Date associated with the most recent gift

MAXADATE – Date of the most recent promotion received

DOB – Birthday year

7.2



	INCOME	MINRAMNT	AVGGIFT	CARDPROM	NUMPROM	MDMAUD_A_C	HIT	TIMELAG	NUMPRM12	LASTGIFT	MAJOR	WEALTH2	MDMAUD_R_C	MDMAUD_R_L	MDMAUD_R_D	MDMAUD_R_I	MDMAUD_F_1	MDMAUD_F_2	MDMAUD_F_5	MDMAUD_A_L	MDMAUD_A_M	MDMAUD_A_T
INCOME																						
MINRAMNT	0.1																					
AVGGIFT	0.12	0.78																				
CARDPROM	-0.064	-0.56	-0.35																			
NUMPROM	-0.065	-0.56	-0.34	0.97																		
MDMAUD_A_C	0.0037	0.0065	0.036	0.019	0.031																	
HIT	0.088	-0.049	-0.045	0.089	0.088	-0.0015																
TIMELAG	0.035	0.083	0.14	0.04	0.032	-0.0064	0.016															
NUMPRM12	-0.015	-0.2	-0.07	0.38	0.43	0.035	0.022	-0.065														
LASTGIFT	0.1	0.53	0.79	-0.17	-0.18	0.028	-0.027	0.16	-0.035													
MAJOR	0.0056	0.0066	0.041	0.022	0.036	0.9	-0.002	-0.0089	0.039	0.031												
WEALTH2	0.17	0.32	0.17	-0.59	-0.59	-0.0074	-0.022	-0.053	-0.18	0.055	-0.0094											
MDMAUD_R_C	0.005	0.0049	0.023	0.013	0.021	0.62	-0.0078	-0.0069	0.025	0.02	0.63	-0.0044										
MDMAUD_R_L	0.0065	0.004	0.015	0.0018	0.01	0.21	0.0014	-0.0016	0.014	0.014	0.33	0.0017	-0.00014									
MDMAUD_R_D	0.0013	0.0042	0.028	0.017	0.025	0.56	0.0021	-0.0013	0.024	0.016	0.63	-0.008	-0.00027	-0.00014								
MDMAUD_R_I	-0.0019	-0.0018	0.013	0.0072	0.013	0.29	0.0038	-0.01	0.014	0.012	0.31	-0.0069	-0.00013	-7e-05	-0.00013							
MDMAUD_F_1	0.0052	0.0075	0.026	0.013	0.023	0.61	-0.0017	-0.0068	0.027	0.021	0.69	-0.0038	0.62	0.21	0.22	0.3						
MDMAUD_F_2	0.0021	0.0046	0.026	0.014	0.021	0.63	-0.00083	-0.001	0.024	0.02	0.59	-0.0062	0.3	0.32	0.43	0.087	-0.00028					
MDMAUD_F_5	0.0017	-0.0032	0.018	0.012	0.018	0.25	-0.00077	-0.0088	0.016	0.011	0.42	-0.0075	0.06	-9.5e-05	0.54	0.12	-0.0002	-0.00017				
MDMAUD_A_L	0.0021	0.00038	0.014	0.0074	0.013	-0.00019	0.0013	-0.0047	0.014	0.0088	0.31	-0.0034	0.16	0.46	-0.00013	0.17	0.37	0.087	-8.8e-05			
MDMAUD_A_M	0.0045	0.0024	0.012	0.008	0.012	-0.00017	-0.0059	-0.0041	0.012	0.0083	0.28	-0.0041	0.089	-6.4e-05	0.36	-5.9e-05	0.082	-0.00011	0.54	-5.9e-05		
MDMAUD_A_T	0.003	-0.00037	0.0057	-3.7e-05	0.0048	-7.7e-05	0.0052	-0.0048	0.0051	0.0056	0.13	-0.0042	-5.4e-05	-2.9e-05	0.2	-2.6e-05	-5.9e-05	-5.1e-05	0.3	-2.6e-05	-2.4e-05	

Figure 10: 'Value' correlation matrix.

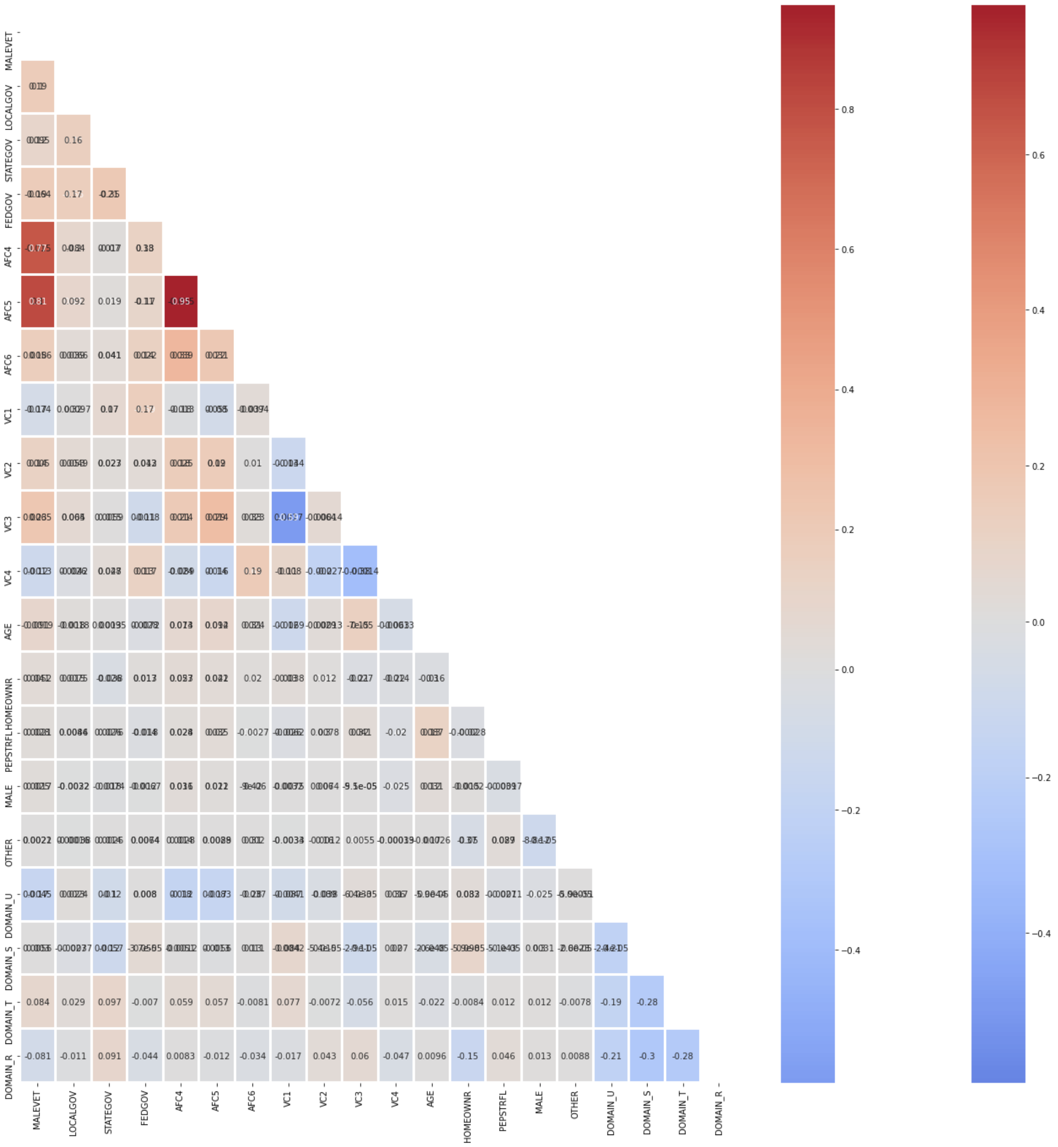


Figure 11: 'Social' correlation matrix.