`

DATA MINING GROUP PROJECT

Marketing Plan for PVA Donors

Master Degree Program in Data Science and Advanced Analytics

2020/2021

Group AK

Henrique Eduardo Espadinha Renda – m20200610

Lorenzo Pigozzi – m20200745

Nguyen Huy Phuc – m20200566

`

Contents

`

1. Introduction

The dataset was provided by the Paralyzed Veterans of America (PVA). PVA is a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. With an in-house database of 13 million donors, PVA is also one of the largest direct mail fundraisers in the United States of America.

The dataset contains 478 columns of 95412 donors. A Customer Segmentation will be conducted using in such a way that it will be possible for the PVA to better understand how their donors behave and identify the different segments of donors/potential donors within their database.

2. Data Preprocessing

The preprocess of the data is a really important step in a Data Mining project, especially in this case where we are dealing with a really big number of features and observations.

Due to the complexity of the high-dimensional space of this dataset, we only used 129 variables, based on our interpretation and point of view. Among the dropped features, some of them had a lot missing values or non-coherent values; other ones were not easy to interpret or not useful for this analysis.

Thus, after this previous selection, we identified 5 categories of features, that correspond to subset of variables: Characteristics, Interests, Mail Responses, Neighborhood, Promotions and Gifts.

Each category explains a different aspect of the donor.

2.1. Characteristics

The first category that we analyzed is Characteristics. In this category are included all the variables that refer to the personal information of the donor such as geographic, demographic, income status, etc.

2.1.1. Feature engineering

First of all, we used the variable DOB (Date of Birth) to create a new variable: computing the difference between the date of birth and the today's date in years, we obtained the variable Age.

The second transformation was the conversion of blank spaces, that we had in almost all the variables, in NaNs values.

Then, we converted some values for the variable Gender: we had 2 values for this variable ('C', 'A') that it wasn't specified in the metadata, thus we replace these 2 values with 'U', Unknown.

The last conversion computed was the transformation of the ZIP code in a uniform way, we deleted the symbol '-' that was inserted in some of the observations.

2.1.2. Unique values check

We also checked the unique values of this subset of variables. The main purpose of this check is to identify if the variables have a really preponderant value in a significant percentage of the donors; if this is the case, the variable is not interesting for our purpose because not useful to identify differences among the donors.

We realized that the variables MAILCODE, MDMAUD and MAJOR have a percentage of a unique value higher than the 98 % of data.

2.1.3. Correlation check

The purpose of this check is the detection of the high-correlated variables. Indeed, if two variables are really high-correlated among each other, we need only one of them for our purpose of clustering: we don't want to repeat the same information many times.

For this category, we realized that the only two numerical variables high-correlated are WEALTH1 and WEALTH2, but we also decided that we will not use these variables to cluster but only for the interpretation, so we didn't drop them

2.1.4. Missing values check

The last check that we realized for the subset of variables was about the missing values. As this dataset is a real one and not created appositely for the project, a lot of the variables have a high percentage of NaNs. Often it happens in real cases.

So, realizing this check for the Characteristics variables, we identified a high percentage of NaNs for the following variables: CHILD03, CHILD07, CHILD12, CHILD18.

Considering all the checks and the meaning of the variables analyzed, we ended up dropping the following variables: MAILCODE, 'MDMAUD, CHILD03, CHILD07, CHILD12, CHILD18, MAJOR.

2.2. Interests

The Interests variables are dummy variables that define the interests of the donors. Each one of them represents an interest, and the possible values are 'Yes' or 'No'.

2.2.1. Feature engineering

After a brief visualization of these variables, we noticed that the only value registered in the data is 'Y', that stays for 'Yes'. For the negative response we only have a blank space, meaning that the donor doesn't have this king of interest.

Thus, we compute the same transformation for all these variables, replacing the blank space with 'N'. Then, we encoded the variables with the positive response = 1 and the negative response = 0.

2.2.2. Unique values check

Due to the nature of this subset of variables, in this case the unique values check is not really meaningful. By the way at least we concluded that we don't have an interest without any donor assigned to.

2.2.3. Correlation check and Missing values check

We realized that we don't have really high-correlated variables for the Interests. Probably the reason is that the interests considered for the analysis are well defined, and not really similar among each other.

Furthermore, we already know from the unique values check that we don't have missing values for these variables, but only 0 and 1.

`

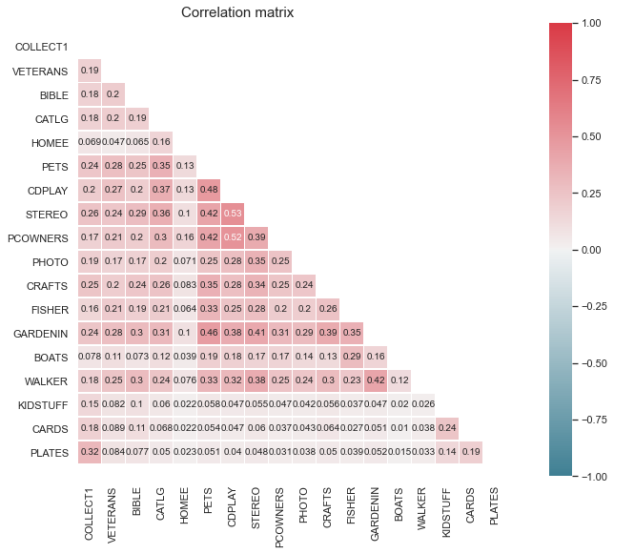

Figure 1: Correlation matrix

2.3. Mail Responses

The Mail Responses variables summarizes the information of the donor's responses for other mail order offers.

2.3.1. Feature engineering

Starting the analysis of this category, we noticed that the variables have value that corresponds to the number of responses registered for different typology of mailing lists. Otherwise, we concluded that the features don't need an engineering transformation.

2.3.2. Unique values check

Checking the unique value for each variable, we concluded that the most common value for all the variables is 0 response. By the way, in our opinion this subset of features is still interesting.

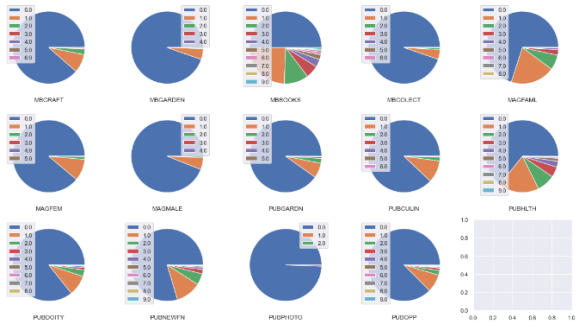

Figure 2: Pie-plots

2.3.3. Correlation check

Analyzing the correlation matrix, there are some variables correlated among each other, but the correlations are not significant to consider to drop any of them.

2.3.4. Missing values check

This subset of variables presents a really high-percentage of missing values: all the variables have the same number of NaNs which is 52854 (55.4% of the dataset). It means that for 52854 donors, we don't have the information of the mail responses to other mailing lists.

2.4. Neighborhood

The subset of Neighborhood variables, as the name suggests, describes the neighborhoods of the donor using a very large number of variables; indeed, we have a lot of information about this aspect of donors, exactly more than 285 features. The majority of them gives us information using percentages of a specific characteristic of the neighborhood. Due to the large number of variables, we selected only a part of them, the most significative ones for our purpose. In fact, many features available are not useful to identify groups of similar neighborhoods for the donors. Furthermore, many similar variables were highly correlated. Thus, we considered in the following analysis 27 variables.

2.4.1. Feature engineering

We quickly concluded that this subset of variables does not require a further engineering.

2.4.2. Unique values check

Checking the unique values, we didn't find variables with a unique value recurring more the 90 % of cases, and only 2 variables, ETH6 and AFC1, more than 80 %.

By the way, we decided to don't drop them.

2.4.3. Correlation check

Looking at the correlation matrix, we noticed that 2 variables are highly correlated, IC2 (Median family income in hundreds) and IC5 (per capita Income). We decided to keep the second variable as more generical and probably more useful than the second one.

2.4.4. Missing values check

Furthermore, the Neighborhood variables don't contain missing values.

2.5. Promotion & Gift

'Promotion & Gift' is a feature subset containing promotion history from PVA. The subset contains specific information about donors' type, promotions date and donation amount of 23 promotions by PVA. The subset also contains summary variables about the total history of donation of the donors.

2.5.1. Feature engineering

Using the summary features provided by PVA, we are able to engineer new features that is useful for customer segmentation analysis (RFM)

RFA_USEDTO_R: The most recent type of code of each donor before they are considered to become Lapsed Donors.

'RECENCY': Time in days since the last promotion and the last gift from the donor: = 'MAXADATE' - 'LASTDATE'

'LIFETIME': Time in days since the last promotion and the first gift from the donor: = 'MAXADATE' - 'FISTDATE'

2.5.2. Unique values check

Features that have more than 99% of unique value are dropped since they will not be useful for clustering analysis: RFA_2R, MDMAUD_R, MDMAUD_F, MDMAUD_A

`

2.5.3. Correlation check

Features that are highly correlated (more than 85%) and having same interpretation meaning with other features were dropped

3. Data Cleaning

The third step of the project is focused on solving problems related with data quality. In a data mining project this reveals to be a crucial phase to discover inaccurate or incomplete data and then improve the quality through correcting of detected errors and omissions.

Clean data by identifying errors or corruptions, correcting or deleting them can be time-consuming but without this stage your final analysis will suffer in accuracy or you could potentially arrive at the wrong conclusion.

3.1 Non-sense values

An analysis was initially carried out on all metric features belonging to the Neighborhood subgroup in order to find anomalous values at first sight.

The variables AGE and AGE901 related to the age of the participants and the median age of the population that surrounds them were filtered out so that there were no records of subjects with negative values.

According to the census the county of Kalawao is the less populated with 86 habitants so we will drop all the records with "Number of habitants" (POP901) less than that number.

Next, we dropped all the rows that presented NaNs amount of “Average Person Per Household” (HHP2) equals to zero because we don’t believe that would be possible.

Finally, we notice that the column EC1 that characterizes the 'Median Years of School Completed by Adults 25+' was certainly not filled with valid values presenting a maximum indicator of 170 years and a minimum one of 89 years so we dropped it in order to improve the data quality.

3.2 Outliers

In statistics, an outlier is a data point that differs significantly from other observations. This can occur due to the incorrect entry, sampling error or because an exceptional but true value present in the dataset.

In this type of project, this kind of values can cause serious problems affecting the variance and standard deviation of the data distribution resulting in skewed distributions what makes difficult to analyze the data.

In Figure 1 we have a representation of the outliers of each feature by using boxplots that identify an outlier as a point outside the quantiles. With this investigation we can state that we are facing a dataset with scattered values.

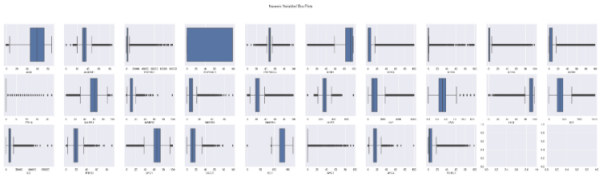

Figure 3: Box-plots

Knowing that probably the quality of the data can be influent by multivariate outliers, we found to be useful to have a general perspective of this kind of outliers that are combinations of unusual scores on at least two variables.

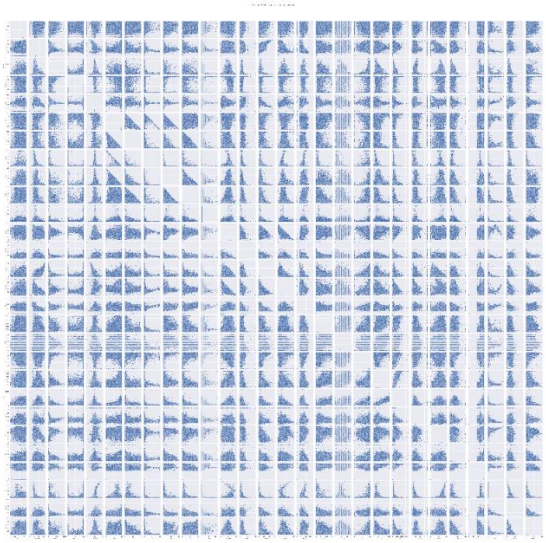

Figure 4: Pairwise relationship of numerical variables

To clear these anomalies from the data we start to use the Local Outlier Factor (LOF) based in the local density formed by the distance between k nearest neighbors. In this unsupervised anomaly detector, every point is assigned with a score that are compared with the others to find outliers, the more the value of any observation more the chance that it will be an anomaly.

This way to clear values that don’t follow a normal distribution is highly influent by the K-distance (distance between a point and the nearest neighbors), by the Nearest Neighbor, by the Reachability-Distance (distance between a point and the maximum k-distance) and by the Local Reachability Density lrd (inverse of the average reachability distance of a point).

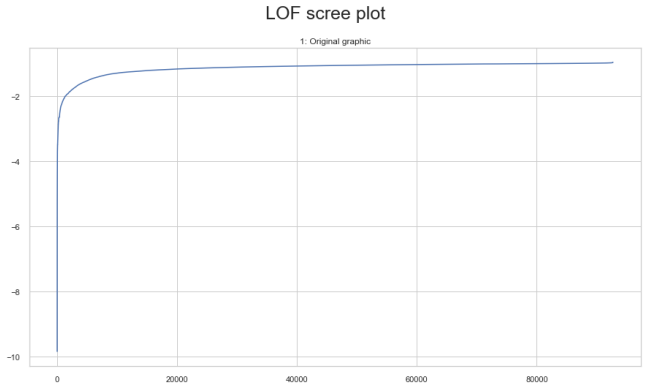

Figure 5: LOF scree Plot

From the LOF scree plot, we are able to say that the outliers can be the observations that have the LOF negative score below -2.

3.3 Missing Values

Missing values in a dataset can be a problem very present before starting modelling related, for example, with the fact that some machine learning algorithms demand those null values be imputed before proceeding further. A popular approach to missing data imputation is to use a model to

predict the missing values. This requires a model to be created for each input variable that has missing values.

From the first analysis to the features, we can identify that the subcategory that have more missing values is the characteristics INCOME, WEALTH1, WEALTH2, AGE and more which cannot be imputed with the danger of creating bias information that will affect the final result. Thus, we will use these features only for interpretation

3.4. Data Transformation:

Promotion & Gift features

The distributions of these features are highly skewed and maybe contain outliers. To improve the normality of the data, we will consider multiple data transformation techniques including Logarithm, Box-Cox [8], Yeo Johnson [9], Quantile [10] to transform the dataset in order to making it more Gaussian-like. Measuring R2 of Kmeans clustering showed that the most suitable transformation method for this data subset is the Logarithm of 10. We also calculate the skewness and kurtosis parameter of each features to evaluate the transformation. For some specific features (highlighted in red) where the transformation does not generate better result, they will not be transformed.

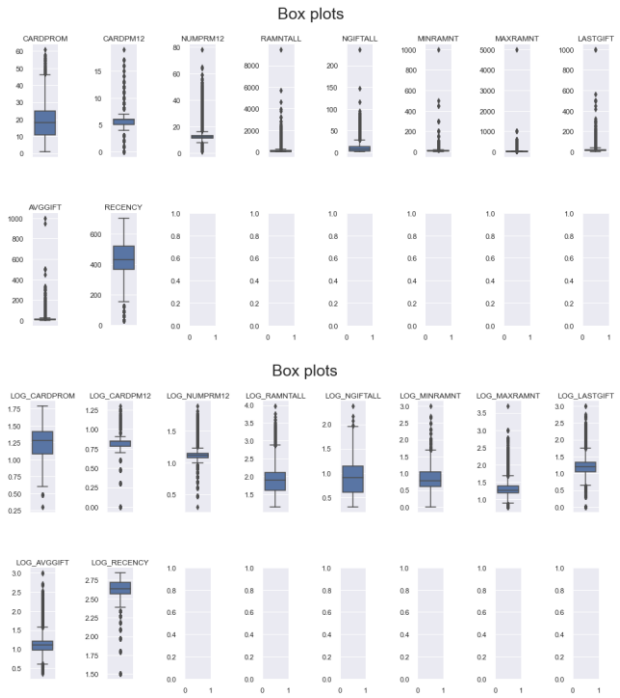

Figure 6: Before and after transformation

4. Clustering

4.1. Interests features:

The Interests sub-group is composed by 18 dummy variables, that can be used to identify some common aspects among all the donor's interests.

First of all, it's useful to apply the Self Organizing Maps algorithm, in order to visualize and have a better idea of the distribution of the donors in the dimensional space.

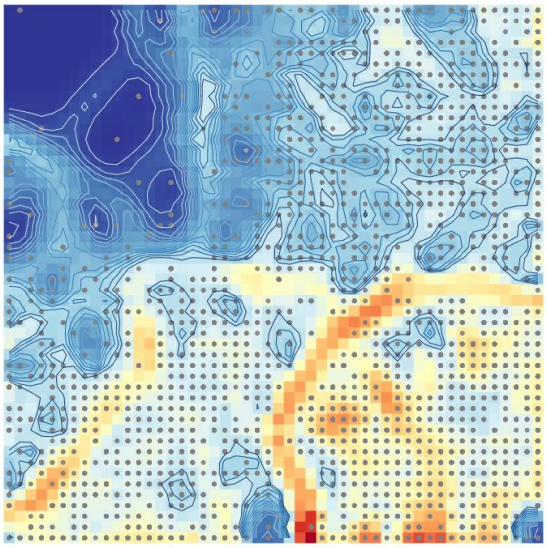

Figure 7: U-Matrix

Analyzing the plots, it's clearly defined a very high populated area on the top-left part of the plot. The reason is simple: the 66 % of the donors doesn't have information for the Interests variables, and by default the values for all the interest variables are set = 0. Thus, we consider all these observations as a cluster, with no interests defined. But for the other observations, it's still possible to group the donor's interests and have an aspect more to consider for the final clusters.

Considering that the Interests variables are dummies, we concluded that the clustering algorithms based on density aren't the most appropriate. The k-means algorithm seemed to be the best approach, as it gave us the best result. Using the inertia plot, we chose 3 clusters; it's possible to visualize the result in 2 principal components dimensional space thanks to the 2 first components.

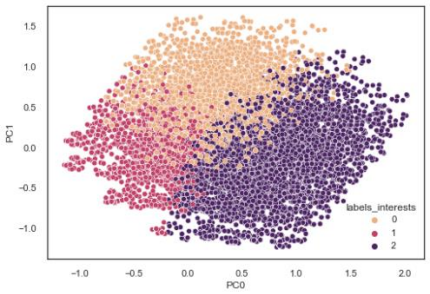

Figure 8: Principal Components plane

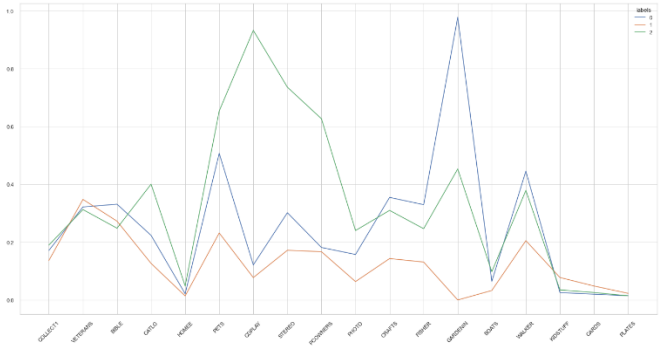

Figure 1: Mean-plot

`

Cluster 0: The "Indifferent" donor

- There isn't a particular interest that really affects the Indifferent Donor. Data shows that he is a bit interest in all of the topics, but nothing in particular.

Cluster 1: The "Green Thumb" Donor

- The Green Thumb Donor is a nature lover: his interests are crafts, fishes and walk outside. But his true hobby is mainly one: gardening. Interests variables: GARDENIN, CRAFTS, FISHER, WALKER

Cluster 2: The "Informatic" Donor

- The Informatic Donor is the man of the 21st century. He is an expert of music and technologies, such as pc, stereo and cd-players. Interests variables: CDPLAY, STEREO, PCOWNERS, PHOTO, PETS

Cluster 3: No interests information

- No information for these donors. They have all the interest variables equal 0.

4.2. Mail Respond features

The Mail Respond variables indicate the number of known times the donor has responded to other types of mail order including emails about different type of books and magazine offers.

4.2.1. Association rules

From the information about the features, it would be interesting to investigate about the association rules regarding to the offers by mail. The result of this analysis can be useful when designing the marketing campaign.

It is discovered that donors responded the most to offer for **Books** (hold **69%** of the donors who responded to at least one offers), following by **Health publications (49%)** and **General Family Magazines (40%).** There are certain association rules that donors who responded to health publications also responded to with books offers with confidence of 0.81 and lift of 1.17.

4.2.2. Clustering

One notice when transform this type of data is that they have a true zero. Therefore, we used MinMaxScaler instead of StandardScaler and the result is better for MinMaxScaler. For clustering algorithm, we evaluated 3 different solution and compared the result in term of R-squared score and class distribution.

Algorithm	Nº Clusters	R-Squared	Cluster Distribution
Hierarchical	3	0.3693	Unbalanced and mixed
Gaussian Mixture Model	3	0.2411	Unbalanced
K-Means	3	0.4238	Good

Since K-means has outperformed other clustering solution regarding both R2 and clusters' size, we will choose K-means as our final clustering model for the Donor behavior cluster analysis.

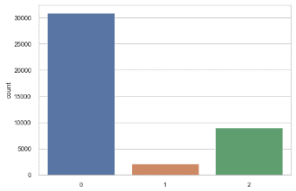

Figure 10: Clusters distribution

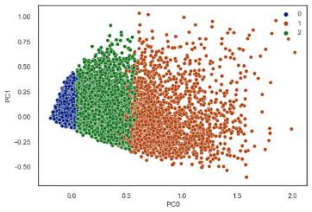

Figure 11: Principal Component plane

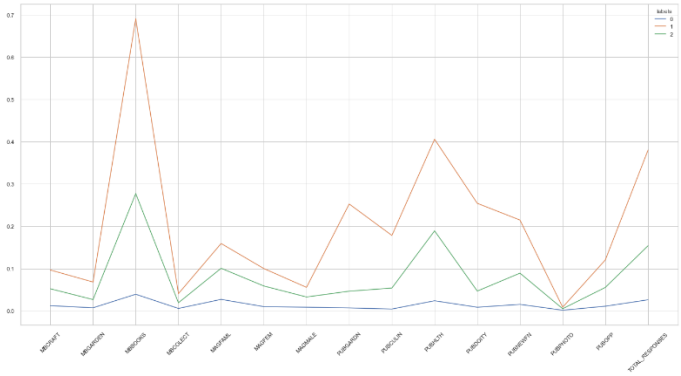

Figure 12: Mean value of each clusters

Cluster 0: The Not-interested

- Sometimes buying books and health publications and family magazine but not usually. Beside those readings, they rarely respond to other offer emails. One assumption is that maybe these are younger people who have more way to access to information so they tend to not buying from email.

Cluster 1: The Knowledge Seekers

- Buy a lot of books, along with publications about health, family magazine and also gardening. These donors could be the older ones, and they have been buying these types of books and magazine by email for a long time

Cluster 2: The Balanced Buyers

- Responds to a medium number of offers comparing to the knowledge seekers. However, they have less interest in DIY and gardening publications

Cluster 3: No mail responds information

- No information for these donors regarding to email offer responds

4.3. Neighborhood features

4.3.1. Dimensionality Reduction for Neighborhood's Metric Features

To reduce the number of numerical variables presented in the Neighborhood subgroup, we adopted the mathematical procedure of principal component analysis. With the help of the PCA () method present in the scikit-learn library, we convert a set of possibly correlated observations into a set of uncorrelated values called principal components.

In order to have an explanatory number of components, we have prepared a table containing the cumulative sum of eigenvalues that indicates the sum of the variances present in each component. In other words, a larger eigenvalue means that that principal component explains a large amount of the variance in the data.

`

	Eigenvalue	Difference	Proportion	Cumulative
1	0.253429	0.000000	0.497093	0.497093
2	0.078540	-0.174889	0.154054	0.651147
3	0.044457	-0.034083	0.087201	0.738348
4	0.030900	-0.013558	0.060609	0.798957
5	0.025343	-0.005556	0.049710	0.848668
6	0.018902	-0.006442	0.037075	0.885743
7	0.010239	-0.008662	0.020084	0.905827
8	0.006958	-0.003281	0.013648	0.919475
9	0.006249	-0.000709	0.012257	0.931732

Figure 13: Eigenvalues variance

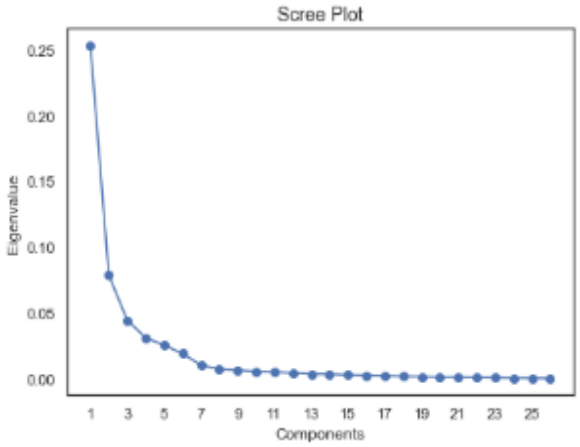

Figure 14: Eigenvalues Scree Plot

It is generally described that to describe a data set using main components we must use at least 80% of the sum of cumulative eigenvalues , and according to the cumulative R2 we chose 5 main components to describe the numerical values of this group of the dataset.

We can also use a scree plot to choose components and through the elbow method to evaluate how many main components we need.

4.3.2. DBSCAN for removing outliers

The unsupervised machine learning algorithm DBSCAN will be our first step to divide our data in clusters. We will not use this method to create different groups of observations, but to remove the outliers that can influence the final model performance.

To tune the hyperparameters of this algorithm we plot a k-distance graph to find out the best epsilon value that define how close points should be to each other to be considered part of the cluster. Using the elbow method, we were able assign eps = 0.15.

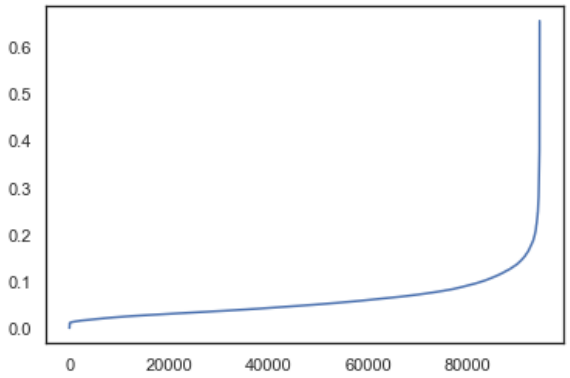

Figure 15: K-distance graph for epsilon value tuning

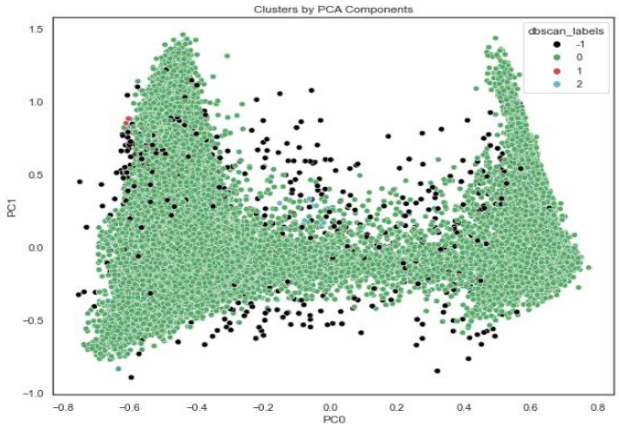

Figure 16: DBSCAN Outlier Detection

For min_samples parameter we stand with 20 points like we did in the practical classes.

Fitting our data frame with the principal component to the model we get 4 different clusters (3 clusters + outliers) and plotting the data we can see that the model could identify with success the most disperse points in this distribution (black points).

4.3.3. Applying K-means Algorithm

K-means is a cluster algorithm that tries to make K groups of data points as similar as possible assigned to a specific centroid. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

To process the learning data, the K-means algorithm in data mining starts by forming a first group of randomly selected centroids that will be used as the beginning points for every cluster, and then performs iterative calculations to optimize the positions of the centroids.

One of the main keys of this algorithm stands with the fact that we need to input a specific number of centroids in the beginning and for that we used an inertia plot so we can identify the number of clusters by the elbow method like you can check in the image. We used the n_clusters parameter with the value 3 in this case.

After that we fit our data to the K-Means model provided by Scikit-learn library, and we get a R^2 score of 0.81 which was our best performance comparing to the other models and we were able to get a good representation of our clusters.

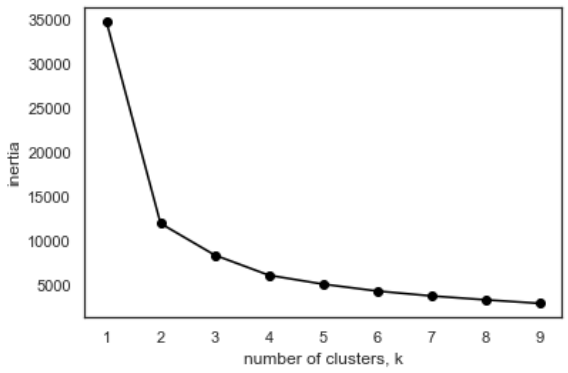

Figure 17: Inertia Plot

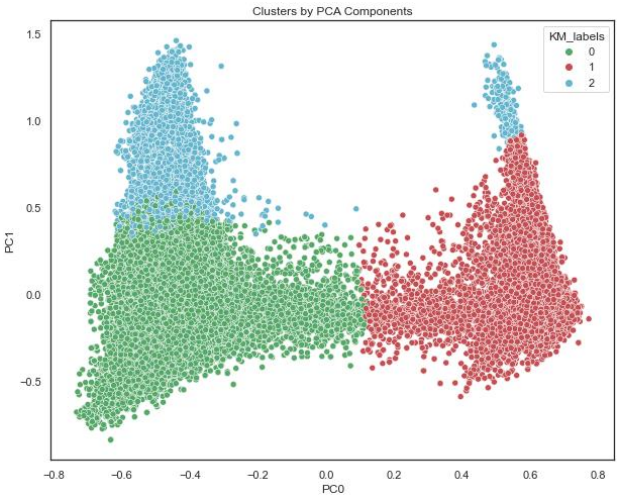

Figure 18: K-means clustering

4.3.4. Re-inserting the outliers: semi-supervised Classifier

In order not to lose too much information from the dataset, we decided to re-insert the outliers that had been removed before applying the K-Means algorithm using the K-Neighbors Classifier from the Scikit-Learn library. This method analyzed the eleven points closest to each outlier (n_neighbors = 11) through its distance and assigned which cluster it belonged to.

4.3.5. Interpretation

In order to interpret the clusters in an easiest way, we created components from further sub-groups of the neighborhood variables. We identified 5 sub-groups that define the neighborhood: popularity, origin, status, quality and veterans. Using the PCA for each group of variables, we got this result and the correspondent interpretation for each component:

Components explanation:

- neigh_rurality: negative correlated with percent population in urbanized areas
- neigh_origin_1: negative correlation with percent white, positive with percent black
- neigh_origin_2: positive correlation with percent Hispanic and percent foreign born
- neigh_status_1: negative correlation with percent married
- neigh_status_2: positive correlation with percent widowed
- neigh_quality_1: positive correlated with all (higher values = good neighborhood)
- neigh_quality_2: negative correlated with percent of houses occupied
- neigh_veterans_1: positive correlated with percent veterans

Using these components, we interpret the neighborhood clusters.

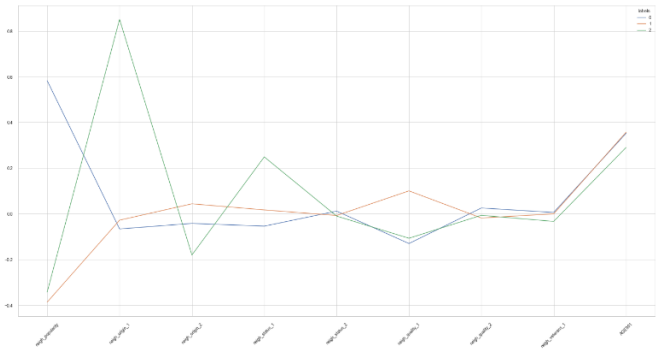

Figure 19: Mean values plot for each clusters

Cluster 0: The "Texas" Neighborhood

- Not urbanized areas
- Prevalence of white people
- Prevalence of married people
- Not very rich neighborhood

Supposition: Neighborhoods outside the big cities and populated areas (Examples: Texas, Albuquerque, and all the center of USA)

Cluster 1: The "Manhattan" Neighborhood

- Urbanized areas high populated
- More international neighborhoods (higher percentage of Hispanic and foreign born)
- Good quality of neighborhood (higher income)

Supposition: big neighborhood in the big cities (example: Manhattan in New York)

Cluster 2: The "Bronx" Neighborhood

- Urbanized areas high populated
- High percentage of black people
- Low percentage of foreign born
- Low percentage of married couples
- Poor neighborhoods (not high income)
- Low percentage of veterans
- Younger neighborhood

Supposition: the ghettos (Ex: Bronx in New York)

4.4. Promotion & Gift features

One of the most perplexing issues we face while trying to cluster donors is choosing the ideal number of segments. This is a key parameter for multiple clustering algorithms like K means, agglomerative clustering [11]. To decide the optimum number of clusters, we will use 4 different metrics: 'Elbow Method', 'Davies-Bouldin Index', 'Calinski-Harabasz Index' and 'Silhouette Coefficient'.

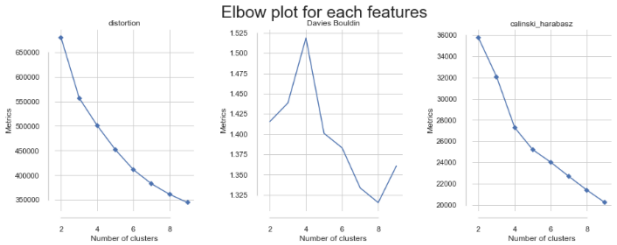

Figure 20: Elbow Plot for each feature

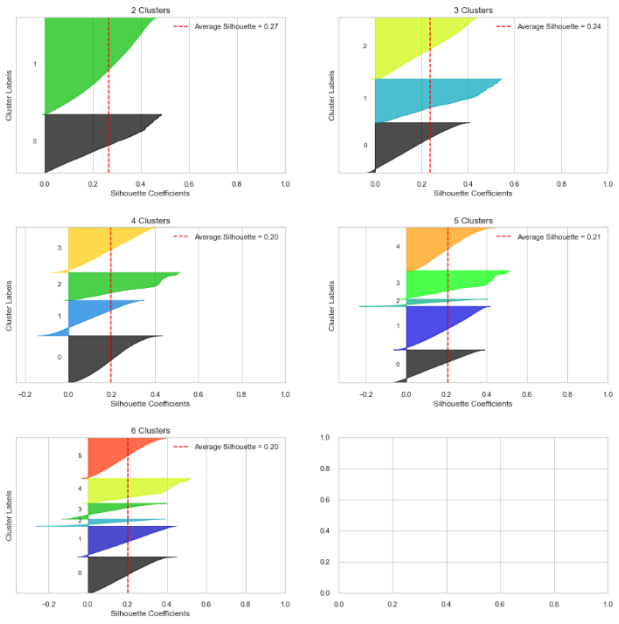

Figure 21: Silhouette coefficient plots

All three methods showing not completely clear result. However, there are points for evaluating:

- Observing the 'elbow' method, we can decide that the best number of clusters are between 3 and 6 clusters
- The DB index captures both the separation and compactness of the clusters. As DB index falls, the clustering improves. Therefore, the graph shows that 3 and 5 clusters can be considered good cluster solutions
- In the Calinski-Harabasz plots, higher the Calinski-Harabasz Index value, better the clustering model. As can be observed, from 3 clusters to 4 clusters showed a significant drop in the index, so it can be concluded that 3 is the optimum number of clusters regarding this index
- The solution of 3 clusters showed a good result in term of score and class distribution. The solution of 4 clusters is a little lower in Silhouette score; And despite having a quite good Silhouette score, solution of 5 clusters faces an unbalanced clusters' size.

Therefore, we will continue the clustering analysis using the number of clusters of 3

For clustering algorithm, we evaluated 3 different solution and compared the result in term of R-squared score, Silhouette Coefficient and class distribution. Another effective method used is to evaluate the visualization of the clustered solution using Principal Components plane and T-SNE.

Clustering Algorithm	Nº Clusters	R-Squared	Silhouette Coefficient	Cluster Distribution
K-MEANS	3	0.40620	0.2389	Normal
SOM AND K-MEANS	3	0.3862	0.2262	Normal
K-MEANS AND HIERARCHICAL	4	0.35358	0.19810	Unbalanced

Since K-means has outperformed other clustering solution regarding both R2, clusters' size and silhouette coefficient, we will choose K-means as our final clustering model for the Donor behavior cluster analysis.

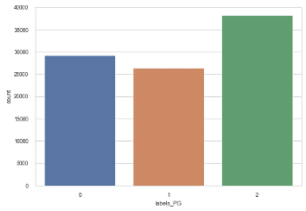

Figure 22: Cluster size

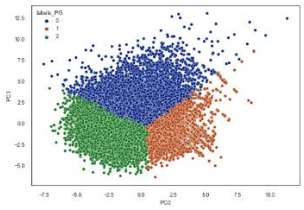

Figure 23: Principal Component plane

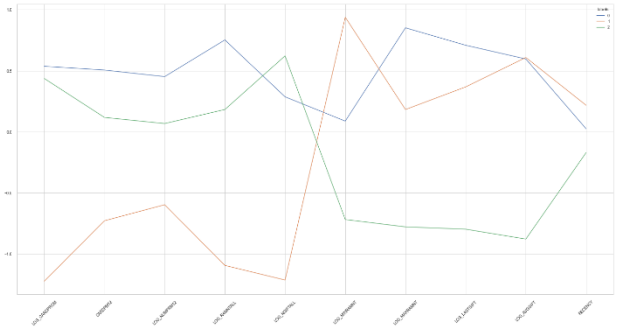

Figure 23: Mean value for each cluster

Cluster 0: The Loyal Donor

- Receive very high number of card promotion since the first date. They also receive the largest number of promotions and card promotions in the current/last year. They can be the most senior people who had donated since the very first date of the organization
- These donors have the highest value of donation without having the largest number of donating times. They also have the highest value regarding to the maximum donation amount. This also make sense with the value of their average gift which is also one of the highest.

Cluster 1: The Rising Donor

- Receive very small number of promotion cards since their first donation, this could be because of they had started donated recently. However, they also receive a relatively small number of promotions this year. This could be because they are not very active donor, so the association does not send them too much cards
- Despite the total amount of donation and total number of gifts is low. They have the highest score in term of average amount of gifts. This information also confirms our previous analysis that they started to donate much later than the others
- These donors also have the highest for the amount of minimum gift. An intuitive interpretation in this case is that because they do not donate a lot of time, so every time they donate, the amount is fairly higher than other people who donate more in term of frequency

Cluster 2: The Active Donor

- Having the highest number of donations. They are also very active in term of receiving promotion and respond to these promotions

`

- The value of each donation is relatively small but on the other hand, they donate very frequently

5. Merging the results

Merging the results is a crucial step to reduce the number of final clusters in order to manage good marketing campaigns for each cluster. In the previous steps, there are 4 subsets of clusters defined: **Mail responses, Interest, Neighborhood and Behavior**. However, since Mail responses and Interest in only informative for less than haft of the dataset, these clusters will only be considered as additional information for later combination in the marketing campaigns.

In this section, 2 methods of merging of Donor behavior and Neighborhood subsets are performed. The first is using hierarchical clustering of clusters centroid and the second is manual merging based on clusters' size. The first solution of hierarchical clustering did not generate a good result as the merged clusters is highly imbalanced and only able to retain the information of 1 cluster in the neighborhood clusters.

Continuing with the manual merging method, we firstly merged the 3 Donor behavior clusters with 2 Neighborhood clusters having the highest observations as 6 clusters. Secondly, the last cluster of neighborhoods having the lowest observation will be joined with all donor behavior as 1 cluster.

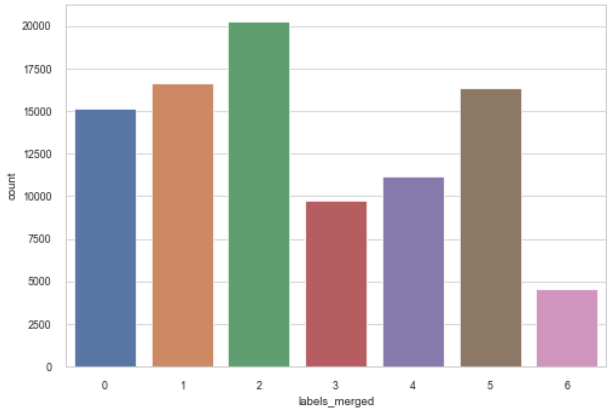

Figure 24: Mean value of each clusters

Cluster 0: 'Manhattan' + 'Rising donor`

Cluster 1: 'Manhattan' + 'Loyal donor`

Cluster 2: 'Manhattan' + 'Active donor`

Cluster 3: 'Texas + 'Rising donor`

Cluster 4: 'Texas + 'Loyal donor`

Cluster 5: 'Texas + 'Active donor`

Cluster 6: 'Bronx + 'Rising donor` + 'Loyal donor` + 'Active donor`

6. Cluster profiling

Clusters profiling is a really important phase for every Data Mining Project. In this step, it would be possible to reap the benefits of the work, and interpret the result that the algorithms provide in a meaningful way. It's performed using the characteristics features that has not been used in the previous clustering analysis such as AGE, INCOME, WEALTH1 & 2.

	Behavior	***Neighbor-hood***	***Interest***	***Age***	***Income***
Cluster 0	Rising donor	Manhattan	Technology	Younger	High
Cluster 1	Loyal donor	Manhattan	PC; not religious	Balanced	High
Cluster 2	Active donor	Manhattan	Walking	Older	Balanced
Cluster 3	Rising donor	Texas	Pets; Fisher; Boats	Younger	Balanced
Cluster 4	Loyal donor	Texas	None	Balanced	Balanced
Cluster 5	Active donor	Texas	Cards; Bible	Older	Low
Cluster 6	Mixed donor	Bronx	Bible	Balanced	Low

In the table above are summarized the main characteristics that define the clusters, based on the previous interpretations of the clusters by perspectives.

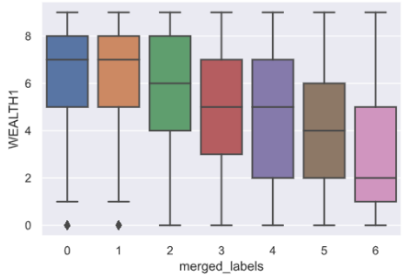

Figure 25: Boxplots of Wealth1 by clusters

Even if the variable WEALTH1 presents NaNs and it is not useful for clustering, we can use it for profiling. The plot above shows clearly that the clusters with the highest level of wealth are Cluster 0 and Cluster 1. Instead, Cluster 6 presents the lowest level of wealth. The result is in line with the previous interpretation done with the neighborhood's analysis.

7. Marketing Plan

7.1. Merged clusters

Cluster 0

Slogan: "A big impact for a life"

Increasing the awareness and the benefit that a single donation can do.

To encourage this donor, provide some example of a very difficult medical case: a veteran saved with a really expensive operation.

Additional gifts:

1) Making a subscription in such a way that they can donate every month, a smaller amount of money, but more times. In order to increase the frequency of the donations for these donors.
2) He can donate when shop online, in order to increase the accessibility to everyone.

Cluster 1

Slogan: "The road with your help"

Giving the idea that their loyalty on the organization is really precious for the veterans.

`

Additional tools:

1) Create coupon for expensive products in tech stores of the big cities
2) Provide newsletter with discoveries for the army-technologies.
3) Making advertises for scientific research in order to help the veterans.

Cluster 2

Slogan: "The work we're doing together, every day!"
They already are motivated for the donations; they are the more active donors. In order to continue on this way, give him info about the benefit that he is providing with the donations, so providing good news in the last period thank to the last donations, in which he is included.
Additional gift:

1) Creating a bridge between the donor and the veteran, providing the possibility to get in contact with the veterans that received help with his donation.

Cluster 3

Slogan: "They served USA: now they need your help"
Giving some examples of young veterans who has served in war and gets a handicap, and original from the same state as the donor.
Additional gifts:

1) Making a subscription in such a way that they can donate every month, a smaller amount of money, but more times. In order to increase the frequency of the donations for these donors.
2) Helping the veterans providing them outdoor activities.
3) Donations in order to provide to veterans the help of guide-dogs

Cluster 4

Slogan: "We will never forget your help!"
Giving the idea that their loyalty on the organization is really precious for the veterans.
Additional gifts:

1) Giving the donor the possibility to participate in reunion or testimonies of the veterans and meet them.
2) Offer advertisement of activities in order to encourage their interests.
3) Custom merchandising: advertise t-shirt with the personal name of the donor and a patriotic slogan.

Cluster 5

Slogan: "Remember the history, don't forget them"
In order to encourage him, providing an example of an old veteran who is alone. The goal is to create care homes for the alone older veterans.
Additional gifts:

1) Creating a bridge between the donor and the veteran, providing the possibility to get in contact with the veterans that received help with his donation.
2) Donate in order to provide to the veteran's religious events.

Cluster 6

Slogan: "Together we can make the difference!"
This donor is the one in the humbler condition. But the little help that he can provide can be crucial for one or more veteran's lives. Provide an example of a veteran who is saved with just a little effort.
Additional gifts:

1) Encourage the donor to make more donations that will help the neighborhood veterans.
2) Providing story of some black soldier in USA army.
3) he can donate in the local stores, in order to increase the accessibility to everyone. For instance, supermarkets, groceries, etc.

7.2. Mail responds clusters:

We can utilize the clusters of mail respond to tailor a personalized gift for the donors who is highly active in respond to the offers by email. Moreover, as these people are highly responsive to email offer, PVA can increase the frequency of sending promotions to them using email.

For example, about the gift, regarding the 'Knowledge seeker' cluster, these are usually the older people who love to read, PVA can send a book or magazine about family with special hand writing card to express the appreciation for these elders

8. Conclusion:

In order to help the Paralyzed Veterans of America (PVA), we analyzed the dataset provided by the non-profit organization with the objective of the creation of a marketing plan to increase the amount of donations based on historical data.

To begin with, we selected the most meaningful features for the extraction of knowledge, using preprocessing methods. From the 128 features, we organized them in 5 subgroups of variables to have different perspectives and analyze the behavior of our observations.

The analysis for each perspective was organized with the same structure: feature engineering and checks for unique values, non-sense values, correlation matrix and missing values; then we computed the DBSCAN and LOF for the outliers.

For the different perspectives, in order to create clusters, we used several unsupervised Machine Learning algorithms, such as K-means, Self-Organizing Maps, Hierarchical Clustering and Mean-shift. To evaluate each cluster solution, we used methods such as Silhouette, R-squared plots, K-elbow plots and Dendrograms.

Due to the big number of variables and for the different perspectives analyzed, we ended up with different clusters solution. The next step was merging the solutions obtained, in order to provide a final solution based on different aspects that characterize the donors. A manual merging turned out to be the best approach for a more robust final result.

Finally, we designed a marketing plan to fit personally each class, based on the interpretation of each group of donors derived by the profiling phase.

`

9. References

1 (https://www.trifacta.com/data-cleaning-in-data-mining/)

2 (http://www.lastnightstudy.com/Show?id=38/Data-Cleaning-in-Data-Mining)

3 (https://www.researchgate.net/publication/279363954_Data_mining_techniques_for_data_cleaning)

4 (https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e)

5 (https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/)

6 (https://towardsdatascience.com/local-outlier-factor-for-anomaly-detection-cc0c770d2ebe=)

7 (https://www.statisticssolutions.com/univariate-and-multivariate-outliers/)

8 (https://www.statisticshowto.com/box-cox-transformation/)

9 (https://www.stat.umn.edu/arc/yjpower.pdf)

10 (https://machinelearningmastery.com/quantile-transforms-for-machine-learning)

11 (https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad)