DATA MINING

PVA'S CUSTOMER SEGMENTATION

FOR MARKETING ANALYSIS

Paralyzed Veterans of America

Carolina Pina r20170790@novaims.unl.pt
Mariana Camarneiro r20170744 @novaims.unl.pt
Matilde Pires r20170783@novaims.unl.pt

INDEX

1. Introduction

This project aims to produce a Customer Segmentation for the Paralyzed Veterans of America (PVA)- a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease.
From the 13 million donors of this association database, we were provided with information on 95412 donors, having 476 specified features about each – either information about themselves, the type of donor they are and the neighbourhood they live in or.
From these customers, the majority are Lapsed donors - those who made the last donation to PVA 13 to 24 months ago. These are our main focus since they are less likely to donate again, but it is still possible to recapture them, so it is worth investing. On the other hand, we will still consider the non-lapsed (mainly those who made their last donation even longer ago), but since they are more unlikely to be recovered, few resources must be spent doing that.
Therefore, after cleaning and transforming the data, we resorted to different Data Mining algorithms in order to segment those clients into 4 different groups/clusters, who reveal a similar behaviour, in order to latter on define a logical marketing strategy for each.

2. Data Pre-processing

2.1. First Step in Removing Variables

Since this dataset has many variables (476), we decided to begin our analysis by running a pandas profiling, to be able to see the distribution of all variables, while simultaneously looking at their meaning and interpretability, to be able to do a selection at first. With this in mind, we developed five initial elimination criteria.
First, we removed all the variables which, considering the context of this study, did not seem relevant at all for our goal. Next, we removed all variables that were not interpretable or had a misleading or wrong characterization, leaving us unable to use them properly in a future analysis. Thirdly, analysing the pandas profiling, we eliminated all variables whose values or distribution did not add relevant nor useful information. Examples of this are variables with more than 45% of missing values or binary variables where one of its values would have an extremely low presence (less than 5%), making it not worthy for analysis. Also, as an initial criterion we took into consideration the variables associated with each other (more in terms of meaning and from the same "set" of variables) and analysed their correlations. From the variables that were highly correlated, we decided to keep only one, selecting the ones with the best interpretability to represent all of them who shared similar information. This analysis was clearly more prominent in the neighbourhood section of the data. Lastly, we eliminated the variables that offered repeated, and therefore not useful or new, information given all the ones we had already decided to keep moving forwards.
The criteria that led to eliminating the greatest number of variables were the ones related to the lack of importance, relevance and usability of the variables themselves. By the end of this first big removal of variables, we had, instead of 476, 286 variables.

It is worth mentioning that we are aware that many of the variables we initially decided to keep are still, at this stage, possibly correlated with others, which is why we intend to have a second round of feature selection, looking more deeply into these possible correlations. A second note relies on the fact that we kept some variables we did not intend to take into our final analysis only to be able

to generate new features from them in our Feature Engineering phase or use them for some type of future data assessment. With this said, all these variables will then be eliminated at the proper stage.
During this first analysis, it was visible that some variables had incorrect values - such as nulls signed with an empty space or an empty space as a value in a binary. So, it was imperative to harmonize the dataset by correcting those mistakes.

2.2. Missing Values

During our analysis, we discovered missing values in a few variables, and we decided to deal with most of them by imputing those null values, instead of deleting the rows, since they were too many. Those variables are:

RAMNT_ X (varying) *TIMELAG (9973)* *NUMCHILD (83026)* *GENDER (2957)* *FISTDATE (2)*

DOB (23883) *INCOME (21286)* *HOMEOWNR (22228)* *DOMAIN (2316)*

Filling in with 0:

All the variables *RAMNT_x*, contain nulls for the records where a donor did not send a gift, which is equivalent to having donated 0. In *TIMELAG*, which represents the number of months between the first and second donation, when a donor does not have a record of the second donation, this value is null. After looking at these observations, it is clear these are one-time donors, so we set it to 0. As for the variable NUMCHILD, considering there is no value 0, we believe the missing values represent donors who do not have children, so once more that was our solution. A comment worth being made is that the variables *ADATE_X* and *RDATE_X* reveal null values that are also not truly missing since they represent donors to whom the respective promotion was not sent. However, these variables will be turned into binary later on, so they are not dealt with in this step.

For numeric variables:

Since we intend to use *DOB* (Date of Birth) to compute the donors' age, and this, by being numeric allows us to fill in its missing values with a method more accurate for each observation, than by filling *DOB* in with a date. At this point, we created the variable *AGE* and dropped DOB.
Therefore, from the variables that ought to be fixed, *INCOME* and newly create *AGE*, being numerical, will be filled in using K-Nearest Neighbours (KNN Imputer). This method fills in missing values using the average of the non-null neighbours' values found in the training set. This way, it enables for a more accurate imputing than by using a constant such as the mean or the median to replace the nulls, since it takes the values in the remaining variables into consideration, instead of looking only at the feature with nulls. However, as there are so many variables in the dataset, we decided to choose, for each, only a smaller subset of the variables more correlated with them, to take as a reference.

For categorical variables:

As for the categorical variables *DOMAIN*, *HOMEOWNER* and *GENDER*, as there is no way to impute the nulls using a predictive model as KNN, the best option would be to use the mode. However, taking into consideration the specific variables and considering the existing categories, we decided it would be best to take a different approach. So, since the variable *HOMEOWNER* provides information regarding the donors that are known to be homeowners, we impute the missing values with the class 'Unknown'. As for *GENDER*, we tried to understand if it would make sense to fill them

in based on the donor's title code (*TCODE*). However, we concluded that these were very incoherent with their gender, so it is best to simply replace the nulls with 'Unknown', as we do not have information to specify the gender.
Regarding *DOMAIN*, it was a more difficult decision since this is a categorical variable composed of two parts that we believe to be extremely useful for the analysis: the urbanicity level of the donor's neighbourhood and its socio-economic status. At first, we split this variable into two containing that information separately, and we decided to impute the missing values in two steps. Start by filling in the *DOMAIN* that refers to the urbanicity level, based on three variables from the neighbourhood that contain similar information: *POP90C1* (percent population in urbanized area), *POP90C2* (percent population outside urbanized area) and *POP90C2* (percent population inside rural area). Here we defined that donors whose neighbourhood has a higher percentage of population living in urban areas, are assigned the value "U", those where the percentage is higher for population living in rural areas are assigned the value "R" and for those where the percentage is higher for population living outside urban areas are assigned "T". Then, for the second part of *DOMAIN* we decided to impute them with the most frequent value (mode) for each category of *DOMAIN_1*. So, for every urbanicity level we calculate the mode of the socio-economic status and then fill in the missing values with the mode for each corresponding urbanicity level. This way, we make sure that the imputation is more accurate than simply filling in with the overall mode and are still able to keep those records.
Lastly, the situation in the variable FISTDATE is different from all others because having two donors that do not have a record of their first donation is problematic, since all people in the database are supposed to have done at least one donation. Looking at the records, it is possible to see that they seem to be very incoherent: there is no first date, but the variable TIMELAG is not null and has values of 1044 and 1088 months between the first and second date, which is around 87 and 90 years. Therefore, we decided to delete these rows.

2.3. Outlier Detection

<u>Boxplot and Countplot Analysis:</u>
When dealing with a huge dataset with several variables, such as the one in this project, methods of dealing with outliers are more prone to errors. Therefore, the first approach we adopted to remove outliers from numeric variables was performing a Boxplot analysis.
Through pandas profiling exploration, we were able to observe graphically the distribution of each variable, from which we selected the ones that did not follow a normal distribution and seemed to have extreme values to plot. We divided those numeric variables into three subgroups to ease the analysis - separating the amounts of money given by a donor (2) and the data relative to the neighbourhood (3) from the rest (1).
All the points that seem to deviate from the majority were manually removed from both (1) and (3) subsets. While performing this analysis, it was verified that different donors presented a significant number of common outliers throughout their variables, meaning the donors who were considered to have outliers in one particular variable were majority the same with outliers in the other badly distributed ones.
However, in the amount section (2), since the scale of values did not differ much, we did not consider the points that were far from most observations to be outliers. Instead, we believe that a donor is free to give more money to promotions if he desires, without a limit imposed.
When dealing with categorical variables we opted to observe their respective count plots graphics and remove the values that were illogical and did not make sense in the context of the variables. By

the end, both analyses translated in the removal of 1.09% (1812) observations from the entire dataset.

DBSCAN:

Our second approach when dealing with the outliers was DBSCAN. After defining logical parameters and applying this method, 1.31% (1254) of the total points were considered as outliers. However, when comparing the points defined as outliers using the Boxplots and the Countplots with the ones using DBSCAN, only 21 were correspondent.

Since this method might lead to errors in such a huge dataset and it is usually not recommended in such scenarios, we decided not to consider these as outliers and, instead, focus and use the ones obtained through the first approach.

2.4. Feature Engineering

Replacing variables with created binaries:

There were 4 variables stating the number of children of each donor, each one relative to a different range of ages. Since this specification is not relevant for our analysis, we decided to replace them with 2 binaries stating if the donor has a child (up to 12 years old) and if the donor has a teenager (13 to 18), *Childhome* and *Teenhome*.

Using the code 00 from the variables *SOLP3* and *SOLIH*, we instead created a binary variable (*RefusePromo*), specifying whether a donor specifically solicited not to be mailed "P3" or "in-house" programs. A single binary variable (*ProgramDonor)* was created, specifying donors who have given to PVA's "P3" or "in-house" programs, using *RECP3* and *RECINHSE*, that were 2 dummy variables, where again, most donors had never donated to.

From the variables of interests, we created one for *Hobbies,* that identifies donors who have an interest in at least one of the following: fishing, gardening, walking, crafts and collectables. Also, one for *Gadgets,* for donors with an interest in CDs, stereos or photographs. Lastly, we created a variable *Republicans* that reveals whether a donor lives in a republican dominated state (according to the 2016 elections).

Ratios and combinations of variables:

We transformed the variables *ADATE_x* and *RDATE_x* into binaries to know whether a donor received or donated for a promotion of a specific type, and using them we created a new variable calculating the total of mailings received for each type of promotion (*Num_x*), and another one with the proportion of gifts given considering the promotios received (*Promo_x*). Similarly, we created the average amount donated by each donor for each kind of promotion (*Amnt_Promo_x*), using *RAMNT_X* variables,.

The *Inactive_time* variable was created considering the date of the most recent promotion we have information (*ADATE_2)* and the most recent gift given (*LASTDATE*). The variable *Donation_activity* was generated considering the donor's first (*FISTDATE*) and last gift (*LASTDATE*). The *Response Rate* variable was calculated using the number of lifetime gifts to date (*NGIFTALL*) and the number of promotions received (*NUMPROM*).

From *SEC1* and *SEC2* (percentage of people in private and public school) and the total population, we calculated the number of students in that neighbourhood (*n_students*) and the percentage of students in private and public schools (*E_Pub* and *E_Priv).*

TOPS Variables:

There were some variables that represented categories we found too specific to keep separated – such as the 16 existing foe the percentage of each type of Ancestry in each neighbourhood. So we

create 1 single categorical variable representing the most common Ancestry (1st *ANCx*) for each donor's neighbourhood (*1st ANC*). We acted the same way for the variables regarding ethnicities (*ETHx*), jobs (*OOCx*) and employment sectors (*EICx*), although for these we left the 3 most relevant ones.

Grouping Variables:

By grouping some of the variables *AGECx* (percentage of adults in each range of ages), we were able to reduce the number of classes of ages into 3 considered relevant to our study: *Younger adults, Adults* and *Elders*.

In the end we dropped all the original variables, except for *ETH1* and *ETH2*, referring to the percentage of whites and blacks in the neighbourhood.

2.5. Second Step of Removing Variables

As mentioned in our initial variable removal, we will now analyse the correlations between the remaining numeric variables, in order to remove any redundant or irrelevant information.

On the one hand, redundant information was identified amongst highly correlated variables. A threshold of an absolute correlation of higher than 0.75% was defined, and we obtained 45 different pairs. For each single case we decided which to keep and, considering some had a high correlation with several of the remaining variables, in the end 31 were eliminated.

On the other hand, irrelevant information was identified in variables correlated with few to no other variables. For this, we evaluated the correlation matrix with the 89 remaining numeric variables and searched for the variables who were not correlated with more than 5 variables, with at least an absolute value of 0.1. However, this was analysed more cautiously since some variables that fell into this category were still considered to be relevant. From this last step, 9 additional numeric variables were excluded, leaving us with a total of 139 total variables.

Note that not only on these specific steps were eliminated variables, but also on the ones mentioned and described above.

2.6. Coherence Check

Throughout the deep analysis conducted until this step, some incoherent values were clear and dealt with immediately. However, it is still important to make sure that the variables that remain in the dataset at this point are coherent and without evident mistakes or noise. For this reason, we explored a few tests:

By looking into the inactive time of each donor (the number of days since they last donated, until the last promotion was received), we realized that not all donors are lapsed, since some have not donated for longer than 730 (roughly around 24 months). Therefore, the variable *RFA_2R* is incorrect, because classifies all as lapsed, thus being dropped and a new binary one was created, simply stating whether the donor is lapsed or not, based in our own calculations.

Another important aspect to check is the donors' age, since to donate to the cause the person cannot be a minor. For this reason, we decided to eliminate from this analysis all donors who are younger than 16 years old (on the date of the last record: 2017). We defined this age as a threshold because it is used in the database as a measure of adulthood and, considering the American context, can also be seen as the first step in maturity.

Besides, there were no duplicates which was confirmed by checking that the identifier (*CONTROLN*) is unique.

2.7. Feature Selection

For our study and cluster analysis we decided to divide the remaining variables into three subcategories/segments that provide different insights of the donors, their neighbourhood and the military presence in it (since the organization analysed is related to veterans).
The views' composition depended on the cluster method used, given that not all can analyse the same type of variables. With this said, most methods we decided to apply could only use numeric variables, leaving us no option but to not include categorical variables in our segments. Also, we would like to mention that, at this stage, using all 139 variables to create clusters would be unrealistic, leaving us with a slightly more subjective selection. Moreover, all the variables we decided not to include in the creation of the clusters were later considered to be associated with the clusters in hand, thus visualizing their behaviour.

Personal View

Initially composed of 8 variables about the donors' personal characteristics: **income**, **number of children**, **age**, the **average amount given per donation**, the **total number of donations given**, **response rate**, the **time the donor spent (in days) without donating** until the last promotion was received and finally the **period of activity**.
As mentioned before, we had to develop a more subjective selection, not including some of the features created by us, in particular the less relevant ones: the total number of every type of promotion received, the ratio of every promotion responded and also the money spent in every type of promotion. Knowing we would not use them, we made sure to keep the global variables related to the same topic – not for each promotion type, but for all.
It is important to note that this was the view with the greatest number of categorical variables and with the ones we believed to be most relevant for the study. Particularly for this reason, we will perform a special clustering method – K-Prototypes, capable of using categorical variables along with numeric ones - ahead, with the following extra variables: **major donor**, **lapsed donor**, **interest in veterans**, **interest in the bible**, **interest in pets**, **shop by catalogue**, **owns a pc**, and **refuse promotions flag**. Again, the selection based on our goal for the analysis and what we considered most useful and impactful to create clusters.

Military View

Here, we grouped variables indicating the presence of military and veterans in the donors' neighbourhoods, all in percentages: **men active in the military**, **male veterans**, **female veterans**, **veterans from the Vietnam War**, **from World War II** and from **the Korean War** and **veterans serving after May of 1995**.

Neighbourhood View

Where we kept interesting and useful information about each donor's neighbourhood. Here, selecting the most appropriate variables to work with was the hardest task, given the fact that there were a lot more to choose from. For this reason, once more, we looked carefully into the remaining variables and chose immediately some that were clearly of interest for our goal: **average age of the neighbourhood population**, **percentage of females** and **percentage of men and women employed**. Then, in order to further reduce dimensionality, we decided to perform PCA to the remaining variables, that we hoped would provide a broader understanding of the types of neighbourhood when combined, while concatenating the knowledge into only one variable.

Principal Components Analysis:

For this, we developed three associations of variables, related to specific characteristics: (a) wealth and level of development, (b) social and educational level, and (c) family's constitution.
From these three analyses we obtained a total of 8 principal components, two from (a) and three from each (b) and (c). With this said, not being all 8 PC useful for our analysis and not having all a crystal-clear interpretation, we decided to only use four of them: (PC0 from (a)) Richer and more developed regions, (PC1 from (a)) Poorer and less developed regions, (PC1 from (b)) Neighbourhoods with more educated individuals and higher incomes and (PC2 from (c)) Neighbourhoods with fewer families and more non-family living arrangements.
Not wanting to disregard the variables that went into the creation of each principal component analysis but ended up not being a part of the final PC chosen, we decided to keep some of them in order not to lose that information: **number of individuals in the neighbourhood**, **percentage of white people**, **percentage of foreign born**, **percentage of widows** and lastly **percentage of two-parent earner families**.
With this view completed, it was further tested with K-means (explained ahead). This algorithm using the 13 selected variables led to disappointing results, since it was difficult to attribute a meaning to the PC's. For this reason, we decided to remove the four principal components due to their complex and complicated characterization, but chose to replace them with the original variables used in their construction that provided their biggest meaning, so as to still include the relevant information we were looking for in the principal components.
With this exchange completed we had less four variables but five new ones: **percentage of people below poverty level**, **per capita income**, **percentage of adults with only completed high-school** or equivalent, **percentage adults with a bachelor's degree** and finally **percentage of households with families**.

3. Segmentation

3.1. Clustering algorithms

In order to find relevant patterns in the donors' behaviours, characteristics and environment, we performed several cluster analyses. It was our decision to test different algorithms for our segments/perspectives, to be able to choose the one that provides the best results for our data and goal. For all the clustering solutions tested, we computed three metrics to help compare the results and guide us in the decision of which to choose:
- *Silhouette coefficient:* measure of how similar each point in one cluster is to those of other clusters (negative values indicate wrong classification; values close to 0 indicate proximity to other cluster and closer to 1 are optimal);
- *Calinski-Harabasz index:* ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (the higher, the denser and well separated, hence the better);
- *Davies-Bouldin index*: measure that compares the distance between clusters with the size of the clusters themselves and this index is its average (the closer to 0, the better the partition).
Note: for a deeper and more exhaustive and detailed analysis, consult the Jupyter notebook provided, as we will only highlight the final solutions here.

- **K-means**

Our first attempt was with k-means algorithm, in which all three segments yielded good results. For each segment we tried different number of clusters and the best solutions were: 2 clusters for the personal view and 3 clusters for the military and neighbourhood view.

- **Self-Organizing Maps (SOM) - with k-means**

Aside from k-means we decided to try using self-organizing maps since this algorithm is known to be useful when dealing with large datasets, even if mostly for the visualizations. In fact, when applying SOM on its own, the algorithm was not particularly useful for any of our views/segments, even the U-matrix yielded unclear agglomerations of observations. Consequently, to be able to achieve clearer outcomes, we applied k-means on a 10-per-10 grid (100 units), so that the partitioning algorithm could find more similar groups from this reduced input space. With it, similarly as with k-means, for the neighbourhood and military views, the best result was with 3 clusters, whereas in the personal view, 2 clusters still proved better. Nonetheless, in all three, the results were not better than with k-means although similar, except for the neighbourhood view that was significantly worse.

- **Hierarchical Clustering - with k-means and SOM**

Considering the large dimension of our dataset, even when subdivided into three segments, we decided it best not to use hierarchical clustering alone, since it would most certainly be too computationally intensive and lead to a confusing result. Instead, we combined it with k-means and Self-Organizing Maps, allowing for the respective algorithms to generate a smaller input space - for SOM we opted for a 10-per-10 grid, resulting in 100 units and for k-means for 100 centroids-, on top of which we applied this agglomerative method to further group the already grouped observations and further reduce the number of clusters.

As for the results with k-means, they proved to be considerably worse than simply k-means, and there were no clusters with better and more interesting interpretations.

Regarding the solutions combined with SOM, for the military and personal view the solution was fairly similar to the one with SOM + k-means. However, for the remaining one, this approach proved to be quite unsuccessful.

- **DBSCAN**

Although we understand that density-based algorithms do not tend to yield good results when applied to large datasets and also for the problem in hands - of grouping donors to build marketing strategies-, we still wanted to try it in our data to understand if there was an underlying distribution this clustering approach could identify more clearly. Yet, as expected, this approach was not successful, despite several tries to improve the algorithm.

- **K-Prototypes**

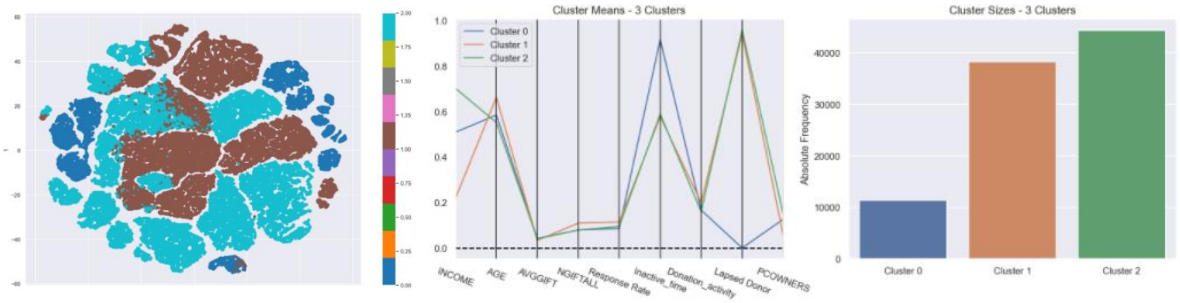
Lastly, as mentioned, we only included numeric features for the creation of the clustering solutions for each segment, since the typical algorithms do not support the junction of categorical variables. Nonetheless, for the personal view we realized there were a few characteristics of the donors that are potentially useful to distinguish them, that were not being taken into account simply because of their data type. For this reason, we used k-prototypes – an algorithm that allows for the usage of both numeric variables (as k-means) and categorical ones (as k-modes) and is particularly useful in large datasets with mixed data. As stated, we only used this method for this view, since we believe the others do not have categorical variables that could be paramount to their creation. In the end we tried a few alternatives for the number of clusters and, curiously, when comparing the solution with 2 clusters to the k-means' solution (also with 2 clusters), it was clear to see that the values were very similar, even of the categorical variables that we later joined to visualize the k-means clustering solution.

However, the solution that brought the most relevant information was k-prototypes with three clusters, since it generated a cluster entirely composed of non-lapsed donors, in which we have less interest, so we can use to separate it and attribute it less importance.

3.2. Final Clustering Solutions per segment

Personal View

As mentioned above, the algorithm selected to create this view's clusters was K-Prototypes, duo to its particular cluster creation and, in our case, to its relevant solution. Below you can find the respective clusters means, clusters size and clusters two-dimensional representation:



First Cluster: Non-Lapsed donors

Th is cluster was the smallest and completely composed by non-Lapsed donors. These showcase the lowest response rate, the highest inactive time and also the least gift giving frequency while active. Interestingly, they were the donors who gave the highest average amount per gift, however, this difference among clusters is not extremely high. Represent the cluster with the least number of major donors.

It is worth mentioning that, in our project, if a donor is non-Lapsed, he can either have stop donating for 25 months or more or he can have made a donation in the last 11 months. In this specific cluster, all the donors have not made a donation for more than two years. For this reason, we decided to separate all the individuals that belonged to this cluster. This is because the most relevant donors for this project are the lapsed ones, being the ones who need to be reminded to donate again, while for the non-lapsed, who have all made their last donation more than 2 years before the last promotion was sent, the efforts to reach them might be more unsuccessful. However, we still believe it is important to include them in the analysis, see how they behave and, considering they formed a whole cluster, address them, although with less importance given.

Second Cluster: Older donors, donating smaller amounts but for the longest time

Donors with the lowest household income and the highest age amongst all. Is the cluster with the greatest number of female donors. These tend to donate less money per gift, however they are the ones with the highest response rate, who have been donating for longer and with the highest frequency, and therefore have made more total donations. The donors in this cluster tend to live in rural areas and cities. Probably associated with their age, less donors own a PC and share interest in gadgets.

Third Cluster: Younger and more recent donors with a higher household income

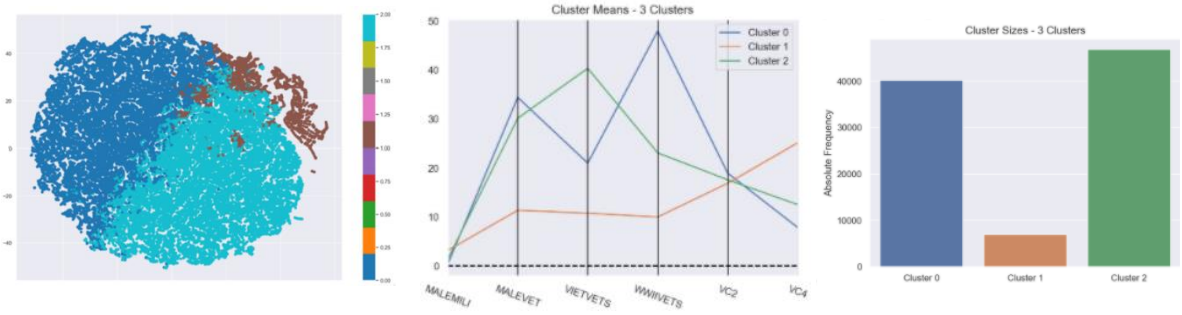
Cluster constituted by the youngest donors (around 59 years old) with the highest household income. As donors these tend to have a typical profile, while showcasing the lowest donation activity along with a smaller inactive time, something that could mean that these have been donors for a shorter amount of time. It has the highest presence of major donors Is the cluster with most male donors, most PC owners, and most interest showed in veterans, gadgets, overall hobbies and pets.

Regarding this particular segmentation, it is worth noting that, due to the characteristics of specific variables, the mean values that they represent for each cluster can be extremely low and therefor

very similar among them. Some examples of this are: “MAJOR”, a binary variable that in the entire dataset only has 277 observations equal to 1, therefor biasing the results at this phase; equal to this is “*RefusePromo*” with only 349 records equal to 1; and lastly a similar situation is observed in the several binary variables related to the donors’ interests.

Military View

For this particular solution, the best algorithm was the K-Means providing us with 3 distinct clusters regarding the military and veterans presence in the donors' neighbourhoods. Below you can find the respective clusters means, clusters size and clusters two-dimensional representation:



First Cluster: Neighbourhood with less active men in military and prominently older veterans

These donors live in the neighbourhoods with the least presence of active men in the military, but contrarily in ones with the greatest number of veterans. This is the cluster with the highest percentage of World War II and Korean War veterans, having also a high percentage of Vietnam war veterans. Finally, it represents the neighbourhood with the lower number of veterans serving after May 1995. Given that the wars with highest presence of veterans in this cluster are from 1939-1945 (W.W. II), 1950-1953 (Korean War) and 1955-1975 (Vietnam War) we can conclude that these neighbourhoods are composed mainly by older veterans. Something also confirmed by the low percentage of younger veterans (serving after 1995).

Second Cluster: Neighbourhood with the most active men in military and the least veteran's presence

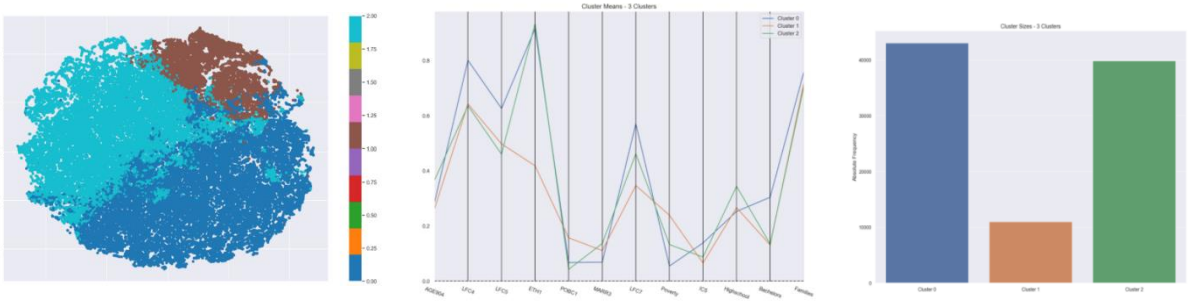
Neighbourhoods that represent the most active men in the military, note that this number is still considerably low, something we believe to be associated with the decrease of need for such services. In addition, the cluster has the lowest percentage of all veterans except the ones serving after 1995, confirming the idea of these neighbourhoods having more younger veterans. Curiously this is, significantly, the smallest cluster, indicating that such scenario is not the most common amongst the donors' neighbourhoods.

Third Cluster: Neighbourhood with a high veteran's presence, specially from Vietnam War

This last cluster refers to neighbourhoods with some presence of active men in the military, and a lot of presence of male veterans. Is also the cluster with the highest percentage of Vietnam War veterans. This cluster is the one with a bigger size, meaning it represents the neighbourhood of majority of PVA’s donors.
It is worth mentioning that all three clusters presented a very small and similar percentage of female veterans.

Neighbourhood View

Similarly, to the previous segmentation, the best solution was provided by K-means. Below you can find the respective clusters means, clusters size and clusters two-dimensional representation:



First Cluster: Richest neighbourhoods mainly composed of families

This first cluster represents the neighbourhoods with the highest income per capita, where the majority of men and women are employed, showcasing the lowest occurrence of individuals below the poverty level. It is also, almost completely, composed by white individuals and is the cluster with higher percentage of families and of two parent earner families. It has the highest percentage of adults with bachelor's degrees and the highest percentage of current students is private schools.

Second cluster: Poorer neighbourhood, highly mixed-racial and with less educated adults

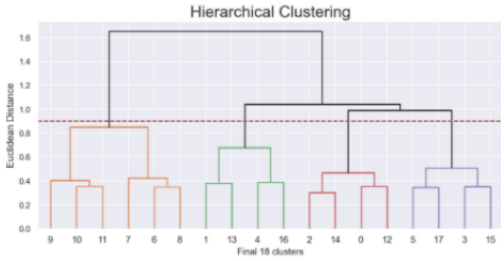
This was the smallest cluster from this solution. The cluster represents neighbourhoods with the younger average age of population. It has the lowest percentage of White people, while having the highest of Black and Hispanic individuals as well as foreign born. These neighbourhoods present the lowest income per capita, having a higher percentage of individuals living below the poverty rate. Have the highest percentage of students in public schools, being neighbourhoods located in less republican dominated states. The cluster represents the neighbourhoods where the percentage of completed education was the lowest (both high school and Batchelors completed). Maybe because of this, we can observe that the percentage of two parent earner families is the smallest.

Third cluster: Neighbourhoods with older people, almost completely white

Neighbourhood with the highest average age, composed with almost 95% of white individuals, having the least amount of foreign born and other ethnicities. Probably due to their age, the neighbourhoods have a higher percentage of widows, of male and female unemployed, as well as more adults with only high Scholl completed. Lastly these neighbourhoods are located in higher dominated states.

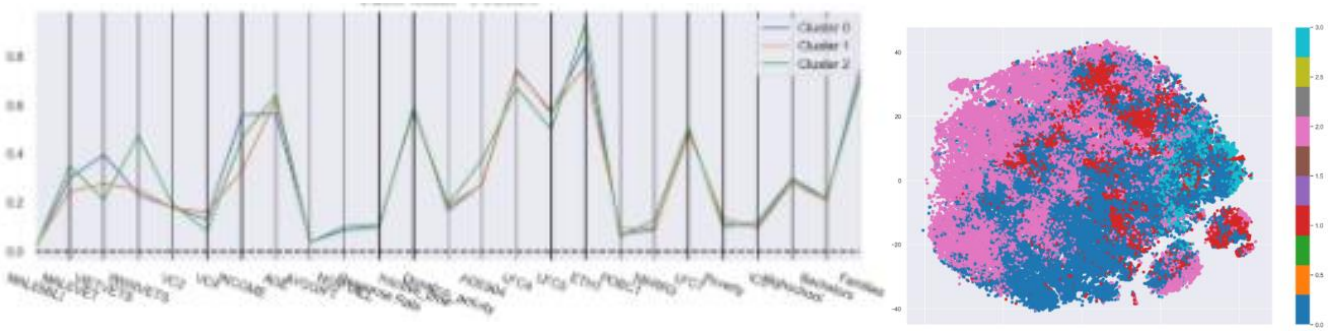
3.3. Joining Clustering Solutions and Marketing Strategies

In this last phase, due to our decision of separating the cluster constituted only by non-lapsed donors, we obtained a final all combined clustering solution of 18 clusters, by combining each view (2*3*3). Nonetheless, it was clear that many of these had fewer individuals and that the differences that make each cluster unique in most cases were not relevant enough to justify 18 marketing strategies being done specifically for each. Therefore, it became necessary to group slightly further from this baseline. To do so, we performed hierarchical clustering, in order to visualize more clearly through the dendrogram how they are grouped, and which are more similar.



On this stage we were less strict about the analysis of the dendrogram because our objective when applying hierarchical clustering was to be able to find the closest centroids to reduce the number of final clusters, not necessarily to have the absolute lowest number of clusters possible. Therefore, although a split would make more sense with 2 clusters, we decided to explore only the final solutions with 3, 4 and 5 clusters. Then, the decision relied heavily on which solution's interpretation made more sense and enabled us to have the most personalized and directed marketing strategies. This proved to happen with 4 clusters.

Before classifying each one, it is worth noting interesting global characteristics of the donors. Probably associated with their advanced average age (around 61 years old), these donors tend to not have children. Such reality does not allow us to obtain, otherwise, useful conclusions. In addition, the majority of donors are women, and in all clusters, it is not possible to obtain great differences regarding their gender. Lastly, it is curious the fact that most donors tend to donate the same average amount of dollars, as we were expecting a notorious difference in this regard.



First cluster: Richer and younger donors, highest amounts donated but a smaller number of donations and inactive for longer

These are the donors with the highest household income (around 4300$ per month), younger (around 59 years-old), who donated the highest amounts (an average of 13,4$ per donation). These are also the donors who have been inactive for longer and that have been donating for a shorter period, which could justify the smaller number of donations made overall. They tend to have more access to computers and shop slightly more through catalogues, and have a higher interest in pets, in hobbies and gadgets. As for their environment, these donors tend to live more in the suburbs or towns, in neighbourhoods with the highest percentage of households composed of families, with an average income per capita and poverty rate. In their neighbourhoods there is one of the highest presence of men active in the military service and around 29% of the men in the neighbourhood are veterans, more from the Vietnam War (more than in any other cluster).

So, these are important customers since they donate high amounts, but it would be ideal to increase the number and frequency of donations, while maintaining the amounts given. However, as they have generally a low response rate, it might mean they are not sufficiently interested in the promotions sent and perhaps donate slightly more arbitrarily, with less regularity. Taking this into account, with these specific donors, perhaps it would not be useful to keep sending them all the regular promotions, given they tend to not behave in a linear way. Instead, since these are younger individuals and several of them are computer owners, we believe initiating a more direct social media approach would be extremely beneficial – not in particular just for these but mainly. Such plan would involve a regular presence of PVA's social media accounts in the donor's platforms. The goal with this is to increase the donor's awareness for the cause, with more frequent adds and such, hoping to improve their frequency as donors.

Second cluster: Most valuable donors, poorer and with lower donation

This cluster is composed of donors who have the lowest income (2800$ per household), which makes them the poorest group amongst all our donors. Accordingly, they donate smaller monetary amounts, but still they prove to be the most valuable ones, considering they donated more recently and with a higher gift-giving frequency. Furthermore, they reveal relatively higher response rates, which means the promotions sent to them generally have a better outcome than to all others and it is also in this cluster than we find the highest percentage of major donors. These donors tend to live in neighbourhoods with the highest presence of men currently active in the military service and the smallest presence of veterans overall, although, from all the veterans, it is in this neighbourhood that the percentage of younger veterans (serving after 1995) is the highest. They also live in neighbourhoods with generally a younger population (average age of 33 years old), with the highest percentage of people employed and slightly higher education (higher percentage of people with bachelors).

Considering all this, we can conclude that these donors might be more prone to donating due to their empathy to the cause and perhaps slightly due to the fact that they live close to people connected to the military service and to young veterans who may engage them to the association. This would explain the fact that they donate frequently towards the cause, despite their low income. Such factor makes them one of the most valuable and reliable donors.

Furthermore, as they are the most recent donors, it is more likely they will donate again, so it is not necessary to take major steps to obtain bigger donations. In fact, the best approach is to keep reaching them consistently as has been done. Nonetheless, it could still be useful to send personalized promotions, explaining how the money donated has been spent and will be spent, so they realize the impact their donations have and are sensitized to keep donating. In these promotions there could be included a testimony of some of the veterans supported by the organization, to further build the donors' connection to the PVA. These promotions could be sent in the form of flyers as well, already with an envelope addressed to the organization where they could place their donation.

Third Cluster: Older donors who made many donations but not of the highest amounts given, live in whiter and richer neighbourhoods

This cluster is characterized by the oldest donors (around 65 years old), who made many donations, but with one of the lowest monetary amounts donated per gift. These donors tend to live in older neighbourhoods and the highest percentage of white people (less than 8% are not white) and lowest percentage of foreign-born people. These are the donors who tend to live in cities and rural areas, in the neighbourhoods with the highest income per capita and the lowest poverty rate. In addition, there is the highest presence of veterans, especially older ones, who were in the second World War and at the Korean war. This are also the donors who live in neighbourhoods more located in republican-dominated states.

In general, these are the donors who live in richer neighbourhoods with older population and the high presence of veterans might justify a bigger connection to the cause, which results in more donations. Therefore, what could be useful in this cluster is to maintain the typical approaches of sending reminder emails and keeping a strong presence in engaging them, since these are quite regular donors that have donated a relatively high number of times. Nonetheless, considering their surroundings and the fact that they live in smaller neighbourhoods in cities and rural areas, where there is a high chance that these donors know their neighbours and have a higher sense of belonging, we believe they would likely be interested in attending periodic social events. These events could be something symbolic organized by the PVA, like a simple barbecue to bring all the

people in the neighbourhood together, to socialize and further involve them with the cause, possible increasing the number of donors, and most importantly the money gifted as a donation, particularly in the days of the events themselves. These would have an entry fee which would be donated to the association.

Last Cluster: Donate high monetary amounts, but the least frequently and live in poorer, non-white and less educated neighbourhoods

These are the donors who donate one of the highest monetary amounts but have the lowest gift-giving frequency. These donors tend to live more in urban areas, in neighbourhoods mostly composed of non-white people (less than 40% are white), and with the highest percentage of people born outside of USA. They tend less to be men and have a lower interest in veterans and in gadgets, and slightly higher in bibles, which leads us to believe they are more religious. It is also a neighbourhood with more widows, smallest percentage of people employed and of households where both parents earn money. Although these are not the poorest donors, they are the ones who live in the poorest neighbourhoods (lowest income per capita and highest poverty rate), being also the neighbourhoods with the lowest education (lowest percentage of people with bachelors and high school) and higher percentage of students in public schools. From all, these are the donors who live in neighbourhoods less located in republic states, although they are still the majority. In these neighbourhoods the presence of men active in the military is the lowest and the presence of male veterans is also one of the lowest, and the bigger percentage of these is related to veterans from WWII (much older ones). It is worth mentioning that this is the smallest cluster.
So, considering these donors do not have much interest in veterans and have the lowest presence of all clusters of people related to the military service (both active and veterans), there is a possibility that these donors donate due to the spinal cord injuries or disease, and not so much due to the veteran cause. For this reason, maybe they should be appealed through a sentimental campaign that reflects the diseases of the veterans to be helped, and how PVA is performing that task of helping others. Furthermore, since these are the least frequent donors, PVA should try to connect with them on a deeper level, so they remember to donate more often, while maintaining the high monetary amounts given. These are also the donors of whom there is less information about their phone numbers, so campaigns should be sent through emails or through mail to their address. Additionally, since these donors live in the most diverse environments (in terms of culture and of the economy), the mailings sent should aim to focus on diverse veterans as well, to highlight and intensify the feeling of inclusion.

It is only left to add that for the **non-lapsed donors**, as was accordingly described above, they are not the focus of our study, so we recommend that few resources shall be invested in recovering those clients. Therefore, more frequent emails should be sent to this group reminding them about how important those donations are for the institution and thanking them for the previous contribution.

Lastly, there are a few steps that we believe are important for the organization to be more engaged with all its donors globally, and that would most certainly impact greatly the donations, for donors of all clusters. As already mentioned, the first is to increase PVA's presence in social media, as it is obvious that these platforms are the easiest and cheapest way to engage customers and keep reminding them to donate. We were able to see that the organization already has an Instagram and Facebook account, but with few followers, especially considering the huge number of donors. Besides, by increasing its presence on social media, PVA would reach younger people, which would be useful considering that it is obvious that most donors are quite old (average age is above 60 years old). Another big advantage of doing so is that it would partly replace the need of sending so many

types of promotions to the donors' addresses (card ones, with and without labels, etc), that have proven to be quite inefficient. Even for the clusters of donors who donate more frequently, the response rate was never higher than 20%, which leads us to believe that too many promotions are being sent in vain, since they respond only to a few. By reinforcing the connection to the donors digitally, less cards and physical promotions would need to be sent, which would save resources and very likely still fulfil the end goal.

Reincluding Outliers

Previously, some donors were removed from the analysis because of their extreme values (outliers), so as the final step of this project we reincluded them in the clusters formed, through a semi-supervised learning approach, using a classification tree. In fact, the idea was to group them to the final cluster to which they are more similar. The same did not happen for the observations removed for being incoherent, since we believe the data from those donors is incorrect and should be confirmed, so they should not be addressed through a personalized marketing campaign until then. In the end, after completing the tree and applying the predictive model to the observations removed that were 1036, 35% were joined to the first cluster (richer and younger donors), 30% to the second cluster (poorer but most valuable donors), 16% to the fourth cluster (regular donors living in poorer and diverse neighbourhoods), 13% to the non-lapsed donors and 5% for the third cluster (older donors in richer neighbourhoods).

4. Conclusion

By the end of our project, there were a few conclusions worth mentioning. Firstly, with this work in particular, we were able to fully understand the importance of data pre-processing. Having received the raw dataset with so many observations, but especially with such a huge number of variables - many with incorrect or confusing data - we understand that producing a cluster analysis with useful and workable results would be impossible without treating the data. For this reason, this was the most extensive phase of the project. Baring this in mind, we were able to reduce from 476 to 139 in our pre-processing phase, including the creation of 40 additional ones, while also cleaning all the data presented in an unusable format. Regarding the missing values we used K-NN imputer for the numeric variables, and a particular mode imputation for categorical variables, fixing incorrect values all-the-while. When it came to the outlier and coherence analysis, there were not that many extreme values, nor non-logic ones that had to be removed, so in these steps we only removed around 2% of our records.

After this essential analysis, we generated three different views to form separated cluster analysis, based on different characteristics/information of PVA's donors: Personal view, Military view and Neighbourhood view. As for the clustering algorithms, we used a total of five different ones and compared them with three metrics, along with a thorough evaluation of the clusters' interpretation. By the end, we were able to conclude that, for our data, in general, K-Means was the algorithm that yielded the best results. However, we also decided to use K-Prototypes in one of our views, since it was more appropriate for dealing with both numeric and categorical variables. From this decision we obtained three different cluster solutions for each of our views.

An interesting conclusion at this stage was that non-lapsed donors formed a cluster on their own, for which reason we separated it from the rest, due to its lower importance. Therefore, after combining all views into a final cluster solution for all lapsed donors we obtained four main types of donors, represented by four clusters, plus the one for non-lapsed. From the three separated clusters for each view to the four final concatenated ones, we were able to observe permanent similarities and shared characteristics in all, confirming the quality of our

final solutions. Related to this, something worth noting on the data is the fact that several variables reveal near constant and similar values for most of the donors. Furthermore, it is crucial to note that, due to the huge amount of data, our final cluster solution revealed a few similarities across clusters, and it would be difficult to obtain more heterogeneous (and more homogeneous within themselves) as desirable. Despite this, we are confident in the solution reached, particularly since the clusters of the individual views that compose it are quite distinct and, in the end, we could still identify relevant behaviour patterns in each final cluster.

Finally, regarding the marketing approaches, we developed personalized ones for each final cluster. Globally we believe the methods used until this point were semi-effective, meaning that PVA still has room for growth in this department. This can be achieved not only with our specific suggestions for each cluster, but also with an intensification and resource allocation in social media, something PVA has started to do, but is still far from effective.

5. References

[1] Sklearn's documentation (2020), *Clustering: sklearn.cluster*
Available at https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
Accessed on December 2020

[2] ODSC - Open Data Science (2018), *Assessment Metrics for Clustering Algorithms*
Available at https://medium.com/@ODSC/assessment-metrics-for-clustering-algorithms-4a902e00d92d
Accessed on December 2020

[3] GeeksforGeeks (2019), *Dunn index and DB index – Cluster Validity indices*
Available at https://www.geeksforgeeks.org/dunn-index-and-db-index-cluster-validity-indices-set-1/
Accessed on December 2020

[4] Guru Prasad (2019), *Notes on K-prototype for clustering mixed typed data*
Available at https://medium.com/@guruprasad0o_o0/notes-on-k-prototype-for-clustering-mixed-typed-data-e80eb526b226
Accessed on December 2020

[5] Ilias Miraoui (2020), *Clustering Algorithms: A One-Stop-Shop*
Available at https://towardsdatascience.com/clustering-algorithms-a-one-stop-shop-6cd0959f9b8f
Accessed on December 2020

[6] *Paralyzed Veterans Of America*
Available at https://pva.org
Accessed on December 2020