DATA MINING GROUP PROJECT: PVA

Paralyzed Veterans
of America

Course: Data Mining

Semester: Fall 2020

Lecturer Professor: Prof. Fernando Lucas Bação

Assistant professors: David Silva, João Fonseca

Group T:
Ana Amaro (20200598@novaims.unl.pt

Gonçalo Almeida (20200594@novaims.unl.pt)

NOVA
IMS
Information
Management
School

INDEX

1. Introduction

The present report describes a clustering attempt and analysis of a list of donors from the PVA organization, in the pursuit of the development of a marketing strategy for each segment of donors found. All the donors in study are lapsed, meaning they have made their previous donation in the last 13 to 24 months, and PVA is concerned the more time it passes, the less likely they are to donate again. In the following sections, we will explore how we processed the data provided by the organization, which clusters did we end up creating and why, and lastly, our brief recommendations of a marketing plan to reach those donors and lead them to donate again.

2. Data Processing

The data provided presented significant discrepancies, duplications of variables and sometimes a too exhaustive focus on some aspects of the donors' demographic environment, among other problems, like missing values. Furthermore, the metadata was insufficient in what concerned the meaning of some of the variables in study. Despite the aforementioned challenges, in this section, we will present our approach to face these issues and create a workable dataset for clustering purposes.

2.1. Feature analysis

The first step towards clustering is understanding the variables we are working with. In this case, given the high dimensionality, for most variables, a simple visual analysis using Pandas Profiling was the best option to quickly summarize information, while going through the metadata.

After this initial step, we proceeded to go through the variables in a one-by-one basis, looking for inconsistencies, ways to improve variables, fix redundancies, treat easy to detect outliers and facing other situations we might find. All the steps taken and the reasoning behind are provided in an exhaustive way in the notebook associated with this report, nevertheless, we decided to cover here some of the situations we had to deal with, for illustrative purposes.

Starting by the format of the variables, we decided to convert all the "dates" we had to a continuous spectrum (being years of age in the case of the DOB, Date of Birth, or months since an event, in the case of the dates for specific donations), and in the reverse process, we discretized the variable TIMELAG (time passed between first and second donations) as a way to circumvent the problem with donors that only donated once: in this particular case, those donors would not be included in any interval of the discretization (that were exhaustive), and therefore would be treated as a separated category. Note that the variables created along the report and the reasoning behind are summarized on Table 1 (appendix).

One specific variable that due to the unclear metadata needed to be investigated further, was the MAILCODE. It was suggested that some donors had a "bad" address, which lead us to think that the PVA association may have detected that some mailings were not reaching the donor. In order to confirm if this was the case -and therefore these donors should be excluded from analysis, as their behaviour would be erratic (e.g., they should arbitrarily have had stop donating)-, or if the meaning of this variable was something else -that we do not have enough information to judge on-, we compared the distribution of the months that have passed since each donor has made his last donation, when the MAILCODE indicated a "bad" address and otherwise. Given the distributions were fairly similar, we did not find evidence to exclude any donors based on this variable.

In this one-by-one analysis, and as mentioned, we also decided to deal with some clear outliers. In this regard, for instance, the variables AGE (previously obtained from DOB) and HIT presented what looked like atypical behaviours, which were managed in different ways. Firstly, regarding the AGE variable, we noticed that there were donors younger than 18 years, and although this may be possible, it looked like an extremely improbable case, that even if true, would most likely result in a segment of donors with a behaviour completely different from the remaining and therefore with no generalization power. Furthermore, after checking in the bivariate space, we discovered that some of those donors had made donations even before they were born, which also supported our decision to just drop these donors from our dataset, resulting in a loss of around 1% of our data. In a completely different scenario, we had the case of the HIT variable, where some values seemed to be too extreme when compared with the remaining ones, but considering there was no way to have a reasonable degree of certainty that those entries were incorrect, we just decided to define a ceiling value, instead of disregarding those entries, and considered all values above that threshold as having the ceiling value.

Another important situation faced had to do with the Census data. Some of the variables from there had clearly impossible values (zeros), as it happened with the variable POP901 (that represented the number of people in a neighbourhood). Moreover, the donors that had this as 0, also had the vast majority of the remaining Census data filled with zeros. Based on this, we decided to consider that all Census data from donors that contained 0 in the POP901 (and consequently in the vast majority of the remaining) as being missing values. Regarding the remaining zeros in the Census data, and considering the difficulty to separate the true ones from the ones that had an erratic behaviour, we decided to interfere as little with the data as possible, only considering as missing values the entries in variables where it was really impossible to a 0 to exist (e.g., in variables like the number of households, POP903, we considered for a 0 to be impossible, while for the variable POP902, that contains the number of families, we allowed for zeros to happen, considering that depending on the way the neighbourhood is constructed, and how a family has to have at least two persons, this might be possible).

Lastly, we also decided, in this phase, to remove some variables that we found to be repeated, to have a too erratic behaviour and/or to contain irrelevant information for the clustering problem at hand. The reason, not only for dropping these variables, but also the remaining numeric ones along the project, is summarized in Table 2 (appendix).

2.2. Imputation of missing values

Usually, it is a good practice to do outlier detection before inputting the missing values, mainly when using statistical methods to fill the missing values (e.g., fill with the mean/median). Nevertheless, and not disregarding the initial outlier detection done while analysing the data, we decided to start by filling the missing values, before going deeper to the outlier detection techniques, as part of our process of outlier detection will need to be performed in a dataset with no missing values.

The first step we took was to remove from our analysis all variables that contained more than 40% of missing values and donors that contained more than 50% of missing entries in variables (with this, we lost less than 1% of our original donors). The reasoning behind this was to assure the reliability of the imputation mechanisms, that when dealing with such large percentage of missing data become very limited in their “predictive” power.

Moving on, in the case of the numeric variables where decimal places were allowed, we used the method KNNImputer to fill the missing values, using as estimators the variables that were more strongly correlated with the one to be filled. On the other hand, when decimal places were not possible (e.g., POP903), or when although theoretically possible, it was clear that the variable was constructed with only integer/rounded numbers (e.g., AGE) -obviously in some variables we could still have used the KNNImputer and round the values afterwards, but not knowing the criteria for rounding initially used, we felt this approach would be safer, as it is still a valid technique that instead of using the mean of the nearest neighbours, in practice, it fills the missing value with the "mode" of the closest neighbours-, we used the same method that when filling the ordinal variables, this is, the method KNeighborsClassifier to make the imputation, again, using as estimators the variables more strongly correlated with the target one. Note that for the ordinal variables, we recurred to the Spearman correlation (between the ordinal variable we wanted to fill and the numeric variables that would be used to fill it), instead of the Pearson, which was used for the numeric ones.

Finally, when approaching the categorical variables, we calculated the normalized mutual information score, and in the cases the variables with missing values had at least a moderate value of shared entropy, we filled their missing values using a conditional mode, based on the variables it shared more information with. Otherwise, when there was no significant shared entropy, we imputed the missing values based on the own mode of the variable, with the exception for the variables from the RFA's, where we felt a best estimator would be for each missing entry to just be filled with the next most recent RFA (e.g., if RFA_22 had a missing value for Recency, we would fill it with the value of the Recency from RFA_21), as at least this way we were able to base the imputation directly on the behaviour of the same donor.

Having now a complete dataset, the next step was to proceed to the outlier detection. In this process, we used two methods based on the normal distribution of data -one for the univariate space (IQR method) and the other for the multivariate (Mahalanobis distance)-, and one based on the density of the data (Local Outlier Factor method). For the methods based on the normal distribution, we had to set very soft thresholds (high multiplier for the IQR method, and high percentage for the Mahalanobis distance) for what could be considered an outlier - while at same time, only considering entries as outliers regarding the normal distribution in the intersection of the outliers for both methods-, as otherwise we would end up considering a too high percentage of our donors as outliers. Note, lastly, that the outliers found here were stored separately to be later added again to our dataset, contrary to the ones detected in section 2.1, where we just dropped the outliers as the entries found there were most likely incorrect data.

2.3. Feature selection

Considering the high dimensional space we are working with, we recurred to several techniques in an attempt to reduce it. The majority of the methods used were based on the correlation between variables: for the numeric data, we used the Pearson correlation with a maximum threshold of 0.85 between each pair of variables to detect redundancies, while also using a minimum threshold of 0.15 to remove variables that are not related with any others; for the ordinal variables, we used the same maximum threshold, 0.85, but this time using the Spearman correlation, between not only themselves, but also between them and numeric variables; finally, for categorical variables, we used the normalized mutual shared information, using 0.85 as the limit of shared information allowed between variables.

In addition to this approach, we also used a Variance Threshold to remove variables that, due to the lack of variability, using 2.5% as the minimum needed, would have a very reduced discriminatory power.

Before proceeding, we decided to reinspect the outliers in our dataset, as now its lower dimensionality may have changed what should be considered an outlier. In this process, we used the same methods as before, and around 4.5% of our original dataset was considered as outliers. Note that we could have just made the outlier detection after the dimensionality reduction, but that could have yield less trustworthy results, especially considering the methods used (correlations and variance).

After all the previously mentioned steps, we One Hot Encoded the categorical variables and used the Standard Scaler method to standardize our numeric features.

3. Clustering

In this section, and with a processed dataset, we will explore and develop our clustering techniques. In order to do this, we will first cluster by two different perspectives and then merge them. The first one will be based on the neighbourhood data, where we will attempt to get a broad view of the social, economic, and demographic environment in which each donor is inserted, while the second one will look at the individual data we have from each donor (e.g., donation history, personal interests or even some demographic insight, like the age), in an attempt to understand their behaviours at a deeper level.

3.1. Neighbourhood perspective

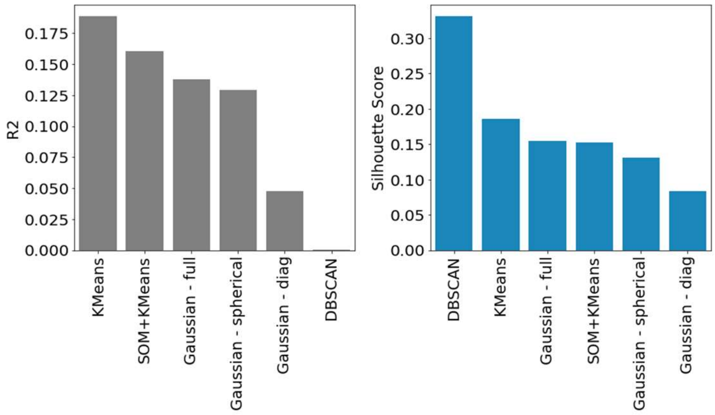
Considering the significant number of variables provided in the Census data that characterize each neighbourhood, we decided to use a PCA to reduce the dimensionality of the input space. As we only want a broad view for this clustering perspective, we decided to only keep 15 principal components, that accounted for around 60% of the total variability of our input space.

After this, we used several methods to try to figure out the ideal number of clusters: the Inertia (for the K-Means), the Bayesian and Akaike Information Criteria (for the Gaussian clustering technique) and the Silhouette average score, being that all suggested 2 being the "best" number of clusters. With this knowledge, we proceeded to try different clustering algorithms: K-Means, DBSCAN, Gaussian and Emergent Self-Organizing Maps combined with the K-Means.

Comparing the different approaches through the R^2 and the average Silhouette score, the one achieving the best results in the R^2 was the K-Means, while the DBSCAN and the Gaussian (with diagonal shape), achieved the best scores in what concerned the average Silhouette score. Nevertheless, after analysing the clusters formed and the plot of the Silhouette, it became clear that the average Silhouette score was not "uniformly distributed" across the clusters formed for these last two algorithms, which leads to a severe reduction of interpretability for the clusters formed.

Next, taking advantage of the ability of the DBSCAN to identify outliers, we decided to extract those from our dataset and re-attempt all the previously mentioned steps (considering around 1% of our original dataset as outliers by this reason). With this, we were able to improve the results in some of our algorithms. The overall best method for clustering was the K-Means, when considering not only being the best one in terms of R^2 and the second best in average

Silhouette score (now, only surpassed by the DBSCAN, that has the problems mentioned above), but also by looking at the split done between the two clusters and how interpretable the results would be.

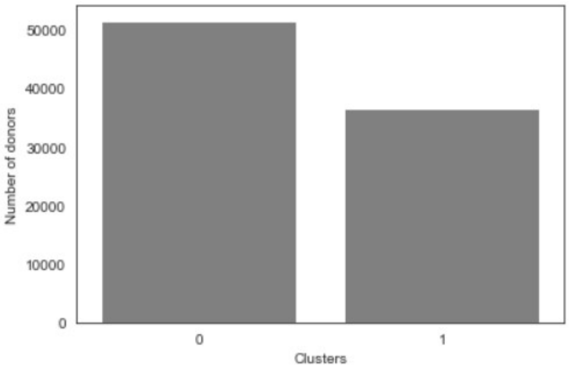


Graphs 1, 2 - On the left, the different R^2 values for each cluster solution; on the right, the different values of average Silhouette score for each cluster solution. Note that both graphs reflect the results after the outliers from DBSCAN have been removed.

Having the best algorithm selected, we used a decision tree and the R^2 of each variable to understand which were the most important variables regarding the separation of the clusters. Based on this, we decided to keep the variables that had an importance, in terms of entropy, higher than the mean, or that had a R^2 of at least 5%. Note that to do this, we used the original neighbourhood variables before applying the PCA, as doing it directly on the PCA's, would not be as significant.

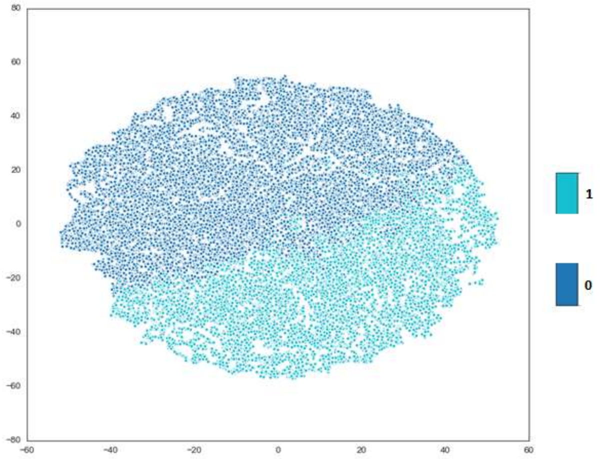
With this new selection of variables, we applied again a PCA, and keeping, once more, around 60% of the variability (of our new reduced input space), we proceeded with only 6 principal components. On those, we applied a K-Means with 2 clusters, and we achieved an average Silhouette score of close to 0.32 and a R^2 score of 0.35.

The distribution of donors per cluster is summarized in the graph below, where we can see that the majority of the donors belongs to cluster 0.



Graph 3 - The distribution of donors per each cluster.

Furthermore, and although the average Silhouette score of our clustering solution is not very impressive, which can suggest only little to moderate differentiation between clusters, when applying a t-SNE visualization, we can easily see points from the same cluster close to each other, despite the "high" concentration of points even in the separation between clusters.



Graph 4 - The t-SNE visualization for the clusters formed.

One of the major drawbacks of using principal components instead of the original variables is the difficulty in interpretability, but considering this clustering perspective is focused on only getting a general picture of the environment of the donor, instead of a fully detailed one, we can still use the correlations between each principal component with the original variables to get meaning out of them, in a not as exhaustive way. Acknowledging this, we will mainly focus on the principal components that have a cleaner explicability of our original variables, even more so, because those were where the major differences were visible and that allowed for a better segmentation between clusters (nevertheless, all the 6 PCA's were used for optimization purposes of our clustering solutions):

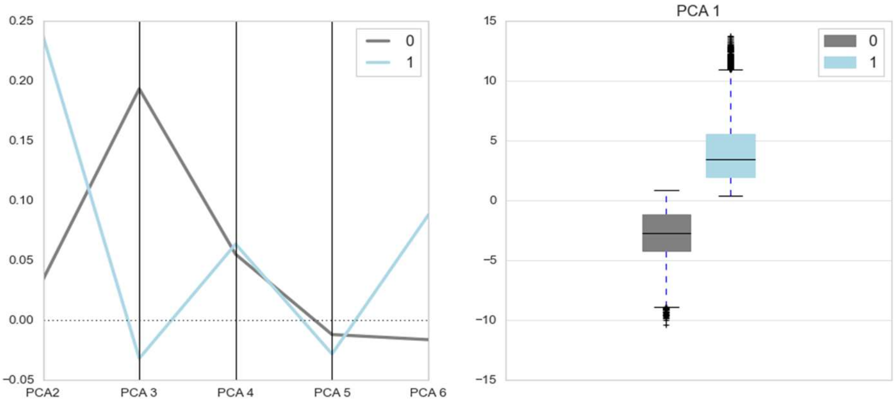
- PCA 1: This is by far the principal component that varies the most between the two clusters formed, and it is also the one where the interpretative power is stronger. It has a high positive correlation with variables where is stored the percentage of people in the neighbourhood that receive the higher amounts of income and income per capita. Furthermore, and with some obvious indirect link with income, this principal component is also strongly positively correlated with high levels of education, with the high values of house evaluations and renting prices, and strongly negative correlated with lower levels of education, people receiving subsidies from government and living below poverty level. In a nutshell, this principal component contains in majority the information regarding the economic environment in which the donors are inserted.

- PCA 2: This variable captures some strong positive relation with the percentage of foreign-born people in the neighbourhoods (with also a strong-moderate correlation with the percentage of Filipinos and Pacific Asians neighbours). Moreover, there is also some positive correlation with living in urban areas and with single parents in the neighbourhood.

- PCA 3: In this principal component, there is a strong positive correlation with older people living in the neighbourhood and a negative correlation with adults in the labour force and younger age groups. Overall, this principal component seems to be related with an, on average, elderly neighbourhood.

- PCA 4, PCA 5 and PCA 6: In these last 3 principal components, the intensity of the correlations with the original variables is relatively much weaker, as it makes sense given how the technique of PCA works. The number 4, similarly to the number 3, also indicates some positive correlation with an older neighbourhood (and will obviously differ in the other variables, but those do not have correlations with values significant enough to be reported), although not nearly as strong, while the number 5 is, in general terms, positively correlated with the number of people living in a rural area and having as occupation the agriculture. Lastly, the number 6 is negatively correlated with the housing in the neighbourhood being mainly residential and not for vacations, while positively correlated with people living in the state in which they were born and with a high area deprivation index (which can indicate the lack of some conditions of livelihood or simply neighbourhoods that by being more isolated have less schools, hospitals and other services nearby).

As mentioned above, the PCA 1 is the one presenting the higher discriminatory power between clusters. In fact, in order for the differences between other principal components to be visible, in the graphs below, we had to separate the PCA 1 from the remaining due to the difference in scale.



Graphs 5, 6 - The differences in mean (and distribution/dispersion in the PCA 1's case) per cluster for the different Principal Components. Note that for a better visualization of the differences between clusters, the representation chosen for PCA 1 was different from the remaining.

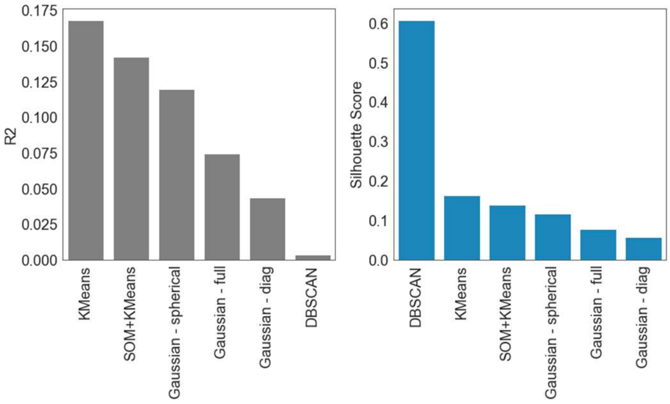
In the PCA 1, we can clearly see that the cluster 0 -that, as mentioned above, is also the one with more donors-, has a significant lower income neighbourhood (and all the consequences/causes from that mentioned above, like the education level and the need for government aid) than the cluster 1.

Regarding the other principal components, they did not show the same differentiated behaviour between clusters, at least not to a point where it was as significant, given our interest in a simple understanding of the environment in which the donors are inserted. Building on this, the PCA 4 and the PCA 5 did not show any major difference in mean terms between the two clusters, while the PCA 2, PCA 3 and PCA 6 showed some variability. This variability seems to suggest that the wealthiest neighbourhoods are, one average, also the ones where the population is younger and where there is a higher percentage of foreign-born citizens (particularly from Philippines and Pacific Asia) and single parents, while at same time where the percentage of people born in the state of residence is also higher and where the area deprivation index seems to be higher, that again can be -and considering the other characteristics of this cluster is more likely than any other reason- due to the geographic location of the neighbourhoods and not necessarily because there is lack of conditions (among other less clear details possible to extract from the PCA's, but that don't help in the task of getting a general picture of the environment each donor belongs to).

3.2. Individual/Personal perspective

Regarding the personal/individual clustering perspective, and considering we want to keep a higher amount of detail in this perspective, we will use a PCA technique, but in a different way than in the neighbourhood perspective. Here, we will use the PCA only as a mean to be able to understand which are the most important features, without incurring in the curse of dimensionality. Based on this, we will keep 14 principal components, which allow us to retain around 80% of the total variability.

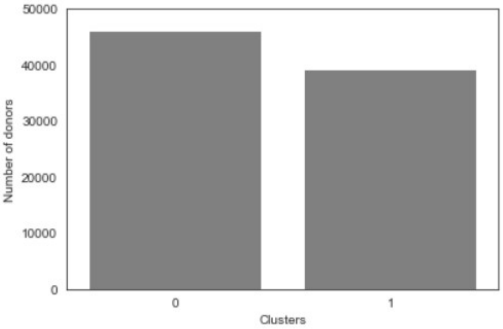
For these principal components, and after applying the same steps as for the neighbourhood perspective, we also found that the ideal number of clusters is 2, with the best clustering solution being the K-Means, after removing the outliers provided by the DBSCAN (which accounted for around 4% of our original dataset).



Graphs 7, 8 - On the left, the different R^2 values for each cluster solution; on the right, the different values of average Silhouette score for each cluster solution. Note that both graphs reflect the results after the outliers from DBSCAN have been removed.

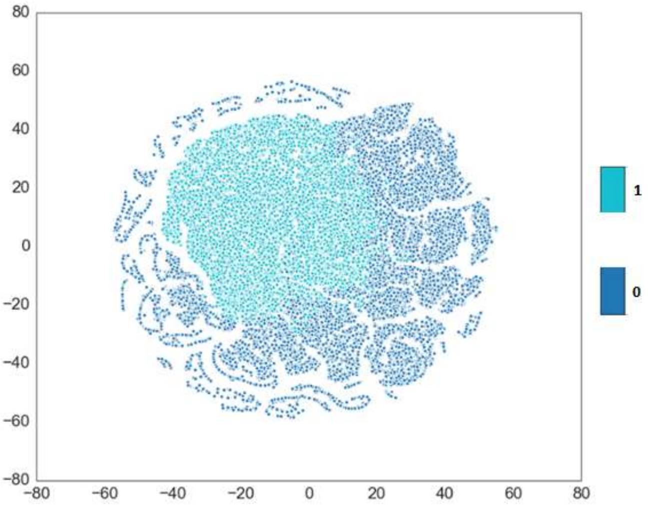
As with the neighbourhood perspective, we performed an evaluation of feature importance, keeping the same threshold for the R^2, but allowing the selection based on entropy to keep all variables that have an importance equal to half of the mean importance. This process allows us to limit the personal/individual perspective to contain only the 9 most important features, that we will use without the need for a new PCA. Our final clustering solution, based on K-Means with 2 clusters, obtained an average Silhouette score of close to 0.30 and a R^2 of around 0.33.

In this perspective, we obtained clusters with a much more similar distribution of donors per cluster, nevertheless, we still do not have an equal distribution, with more donors belonging to cluster 0, which, as we will discuss below, represents the donors who have been with PVA for less time and made less donations.



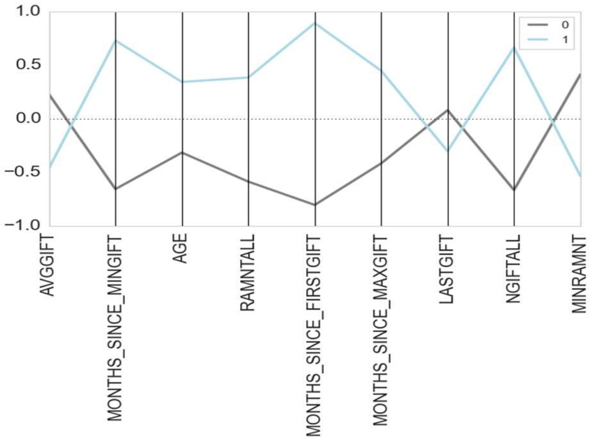
Graph 9 - The distribution of donors per each cluster.

In an attempt to visualize our multidimensional cluster solution, we used the t-SNE technique and, similarly to the neighbourhood perspective, although the separation between clusters is not well defined (in the sense of existing low density on the borders), the points from the same cluster tend to be grouped together, with some exceptions.



Graph 10 - The t-SNE visualization for the clusters formed.

Regarding the characterization of each cluster, inside this perspective, we ended up with the variables that contain the total number of gifts given, the minimum, the total, the average and the last amount donated, the time since the minimum, the maximum and the first gift, and also the age of each donor. In this sense, all the variables seem to behave differently between the two clusters, as it is summarized in the graph below. Note that all variables are standardized to have a mean of 0 and a standard deviation of 1, nevertheless, and as already mentioned, there are significant outliers according to the normal distribution in the majority of the variables we worked with, being that only the most extreme cases were removed when doing outlier detection.



Graph 11 - The differences in mean per cluster for the different variables in study.

The cluster 1 is clearly composed by donors who have made their first donation significantly longer ago than the donors from cluster 0, and that have also donated a larger number of gifts in volume and in amount.

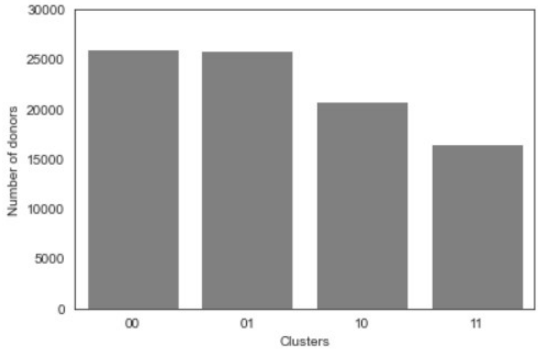
On the other hand, the donors from the cluster 0 have donated more on average, which may indicate that they donate more for each gift, but less frequently than the other cluster's donors, or that simply as they have been donating for less time than the other donors, there may exist a decreasing curve of the amount donated over time. This decreasing curve over time of amount donated is also backed by how the cluster 1 contains the donors who have made not only their first donation, but also their maximum donation, longer ago. Moreover, the last donation was also lower in monetary value for those from cluster 1.

Lastly, there also seems to exist some age separation, with the donors from cluster 0 being younger, on average, than the ones from the other cluster.

3.3. Merging the perspectives

With the two different perspectives already built, the next logical step is to merge them into one, but before this, we need to join the outliers of each perspective to the cluster they more closely resemble (as the clusters for each view are already constructed without the interference of their own outliers). In order to do this, we constructed a Decision Tree Classifier, that based on the parameters of the outliers, was able to attribute them to a particular cluster inside each clustering perspective.

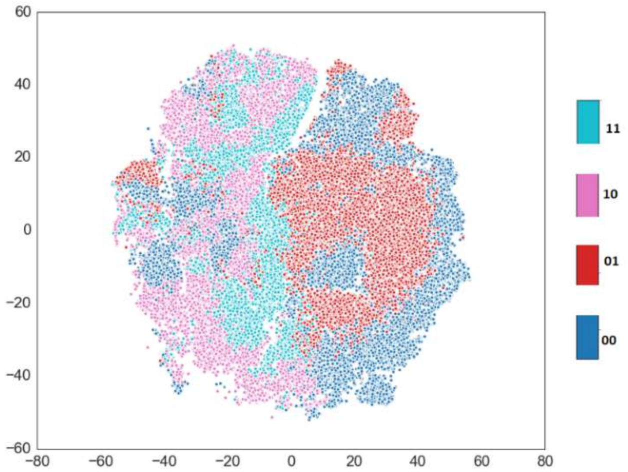
Our final clustering solution with four clusters presented two large clusters with more than 25 thousand donors in each, one with around 20 thousand donors, and a fourth cluster with a significantly lower number of donors, as we can see in the graph below. Note that this number of donors per cluster is before adding the donors considered as general outliers in the phase of outlier detection.



Graph 12- The distribution of donors per each cluster.

Regarding the cluster evaluation, the final R^2 was 0.334, while the average Silhouette score was close to 0.09. This very low value of silhouette score may be explained by how our final clusters were constructed from two merged perspectives, which may create some overlap in the final clusters.

Furthermore, this tendency for not as clear separation between clusters is also visible when looking at the t-SNE representation, where -although some agglomerations of points from the same cluster being visible- there seems to be a significantly higher erratic behaviour than when looking only to the perspectives separately.



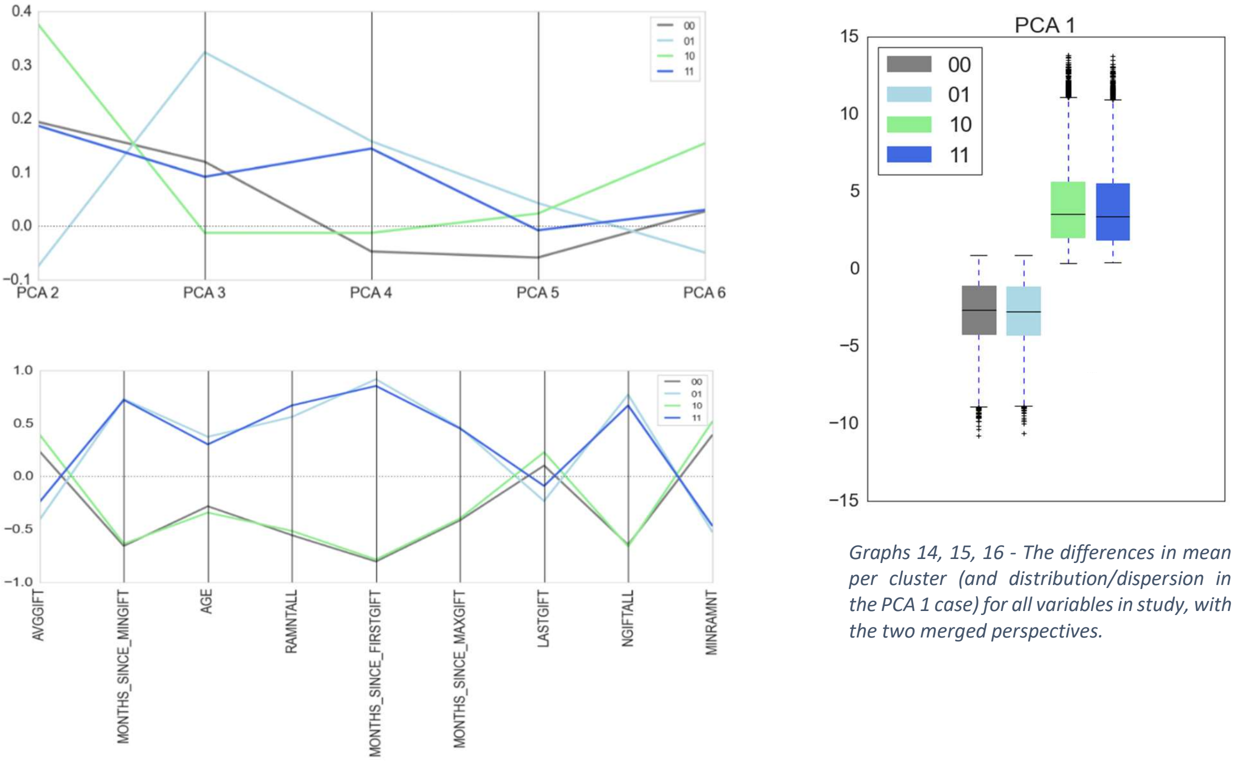
Graph 13 - The t-SNE visualization for the clusters formed.

Considering the clusters formed, the cluster 01 contains the intersection of the donors that we previously classified as living, on average, in a less economic prosper environment (and with lower education levels), with the donors who have been with the PVA organization for longer and donated more in total. In this same cluster, we seem to have the lower value for PCA 2 from all clusters -which implies the lowest percentage of foreign persons living in the neighbourhood and single parents-, while at same time, the higher value for PCA 3 -that implies an older neighbourhood, with less people in the labour force (furthermore, it also presents the highest value for PCA 4, that in some ways captures the same information as the PCA 3 and the lowest value for PCA 6, which among other things may suggest we have less people from the same state they were born in, in relation to the other clusters' donors). Additionally, when looking for the personal clustering variables, and although it presents a similar behaviour to the one from the other cluster built from the Personal cluster 1, it seems to constantly follow it in a diminished way when it comes to the amounts donated (being the total amount, the minimum, the average or the last gift), which is most likely linked with these donors living in a less favourable economic environment, than the other part of the Personal cluster 1. Moreover, and in the similar tendency suggested by the neighbourhood variable PCA 3, the donors from this cluster seem to be the elderly, on average.

On the other side of the Personal clustering perspective, we have the donors from the cluster 00. In terms of PCA 1, the behaviour of these donors is fairly close to the one from the cluster 01 (considering the big gap between those two and the other two clusters), while regarding the remaining variables from the neighbourhood perspective, there do not seem to exist extreme movements that are worth mentioning like in the previous cluster (it presents the lowest value for PCA 4 and PCA 5, but the difference is not very significant, even more considering these variables don't have a very interesting interpretability for the problem at hand). When looking at the Personal perspective, we can see the same behaviour pattern as exhibited by the donors from cluster 01, in the sense that all the variables regarding amounts donated seem to follow in a lower form the pattern from the donations of the donors from the same Personal Cluster (in this case 0), but from a more economic developed neighbourhood.

Moving now to the two clusters where it is known, from the previous analysis of the Neighbourhood perspective, that we have the donors who live, on average, in a more economic developed environment. The cluster 11 is the one with lower number of donors belonging to it (16484 donors), while the behaviour of the PCA 1 is clearly higher when compared with the two clusters mentioned above. Moreover, as happened for cluster 00, the behaviour of the remaining principal components does not present anything worth mentioning for the problem being discussed in this report. The behaviour of the variables in the Personal side has already been discussed above, following the same pattern as the other cluster 1 from the Personal perspective, but with more expressive movements regarding the donation amounts when compared with the cluster 00.

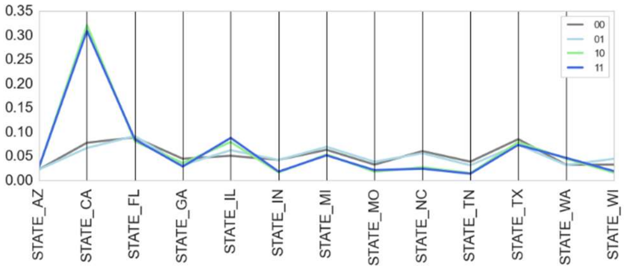
In the same way, the cluster 10, in the Personal variables, also presented the same pattern values associated with the cluster 0 from the Personal perspective, this is, similar to the ones from cluster 00, but making the movements more positively expressive in what concerned the amounts donated. The PCA 1 behaviour is also fairly close to the behaviour from cluster 11, but in the remaining principal components there are some points worth mentioning. Here, the behaviour is relatively the opposite to the cluster 01, being where the younger neighbourhood is found as the value of PCA 3 is the lowest, while the percentage of foreign born seems that may also be the highest, on average, given the high value for the PCA 2 (moreover, it also presents the highest value for PCA 6, that once more, among other interpretations, it is related with more people from the state they were born in).



Graphs 14, 15, 16 - The differences in mean per cluster (and distribution/dispersion in the PCA 1 case) for all variables in study, with the two merged perspectives.

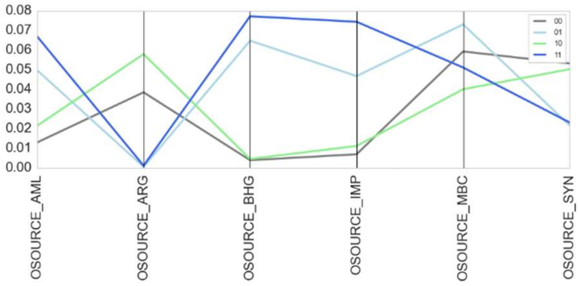
Although the analysis of the numeric variables has already given us enough information to develop a simple marketing report, we also tried to look for patterns in the proportions of the categorical variables, in order to see if there was a significant difference among clusters. In this regard and considering the large number of variables, we will select which variables to study based on our personal intuition and on the interpretability we believe we can extract from the possible findings. Building on this, we attempted to compare how the proportions vary across the clusters for the variables referring to the state where the donors live, the source from where they were retrieved and their interests. Note that we only proceeded with the variables that were not removed up until now, as we previously used feature selection in the categorical features.

Regarding the State variable, the most interesting finding was the fact that we have a significant concentration of our donors inserted in the clusters 10 and 11 in California (32.1% and 30.9%, respectively). This means that a significant part of our donors from wealthiest neighbourhoods tends to be located on California.



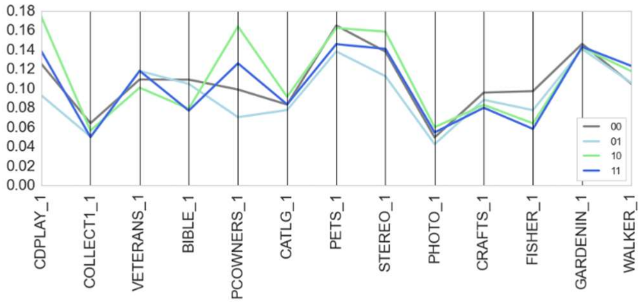
Graph 17 – The differences in proportion per cluster for the variable STATE.

From the analysis of the OSOURCE variable, there seems to be an identifiable pattern in the source a donor is acquired from, and the personal perspective cluster he belongs to. Theoretically, we could focus on the sources that provided significantly more donors from the cluster 1 of the Personal perspective, as those are the ones that historically have donated a higher amount of gifts and value, but it would be important to verify first with the PVA if this different pattern across sources doesn't just have to do with some sources being used for a longer time (which could justify the major amount of donors from cluster 1 of the Personal perspective, as those are also the ones who have been with the organization for longer, and not necessarily because the donors from those particular sources tend to donate more over time).



Graph 18 – The differences in proportion per cluster for the variable OSOURCE.

The analysis of the interests of our donors also have yield some interesting conclusions, particularly, how the interests tend to vary across the two clusters from the original Neighbourhood perspective, as we can see, for instance, with the variable PCOWNERS, that tends to appeal more to donors from neighbourhoods that are wealthier on average (cluster 10 and 11), while the BIBLE one tends to be more an interest from the other cluster's donors (clusters 00 and 01).



Graph 19 – The differences in proportion per cluster for the variables related with the donors' interests.

Lastly, regarding our ordinal variables, we only found worth to explore the variable Income (stores the income category a donor belongs to), that helped us to confirm a trend we were already expecting: the income of the neighbourhood is a good proxy for the income of the donor. We verify this by seeing that the medium value of the variable Income is higher for the donors belonging to the clusters with higher values of PCA 1, and lower, otherwise.

Before proceeding to our marketing recommendations, we added back the outliers found in section 2.3. (in order for PVA to be able to also apply those recommendations to the "outlier donors"), using a Decision Tree Classifier to predict to which cluster each outlier is most likely to belong to.

4. Marketing Recommendations

Although in the process of clustering we have collected several information about PVA donors, from a marketing perspective, we are mainly interested in a few, but actionable, points found.

The first thing that we figured out, and that is relevant from a marketing standpoint, is that we have more donors inserted in a less favourable economic environment than otherwise. This is a very important information, as we discovered that donors from less economic developed neighbourhoods tend to donate lower amounts when compared with donors that belong to the same Personal perspective cluster but from richer neighbourhoods.

Additionally, and using intuition, there seems to exist a decreasing curve of the amount donated over time, with the longer-time donors, starting to donate less after some time, which makes their donations less expressive, on average, when compared with more recent donors.

Another concerning point to PVA is how on average the donors from the Personal Cluster that have higher number and amounts of gift donated, tend to be older and live in an older neighbourhood, which may signal that PVA should start to consider how to engage younger masses.

Based on these major points, we develop some small marketing recommendation for how we believe PVA should proceed.

Firstly, we believe the donors that should concern more the PVA organization, regarding being Lapsed, are the ones from cluster 00 and 10, as those are the ones who have been with PVA for a lower amount of time, and therefore where the risk of not being "loyal" is higher. In order to face this, PVA should attempt to send promotional items that educate towards the importance of the organization and the service it is paying to the veterans of America. This way, PVA would most likely raise awareness that could make those donors re-donate.

Furthermore, and considering that the donors from the cluster 00 are inserted in a less developed economic environment, PVA could also include little messages like "Every dollar counts in this fight.", which may lead the donors from these clusters to keep donating even if lower amounts.

On the case of the donors from the cluster 10, it may also be a good strategy to include messages in the promotions but considering the environment in which the donors are inserted (on average, they are inserted in younger neighbourhoods with more economic means and presence of more foreign-born people), those messages should have a different nature. For those donors, a good approach might be to try to appeal to their own pride, instead of patriotism (as in our research we have found that on average younger people tend to be less patriotic [1], and considering the younger and more foreign, in comparation, environment they are inserted in, it may even have an exponential effect on this lack of patriotism), by including quotes like "The average donors in this neighbourhood donates "x" amount of dollars.".

Regarding the donors from 01 and 11, and with the risk mentioned above of existing a decreasing curve of donations as time passes, the PVA should focus on stopping this trend. One good step towards this, could be when sending the promotions to the donors, PVA personalizing a "thank you flyer" with the amounts a specific donor has donated per campaign and the number of veterans each donation reached. The reasoning behind, would be for each donor to be impacted by how the reductions in their donations have real impacts in the number of people helped and reached.

Moreover, and considering how the cluster 01 is inserted in the neighbourhoods that are older on average -and the donors themselves are also older-, and that also have, on average, the lowest number of foreign citizens, this may be the perfect group to appeal to the patriotism. In this regard, cards customized with the U.S. flag, and with quotes associated historically to the concept of mission for the country may produce good results here.

Lastly, and considering two aspects mentioned above -the donors who donate more in total and amount tend to be older, and the fact that the majority of our donors lives in neighbourhoods not as wealthy as the ones clustered as part of the Neighbourhood perspective 1-, the PVA could focus on, not only attempting to make the lapsed donors donate again, but also to try to find more donors, and we would suggest this to be made particularly in California, where there seems to be a significant amount of neighbours with higher incomes, on average. Furthermore, and as mentioned above, it could be interesting to find out if the sources have changed with time, or if they have been used for reasonability the same amount of time, because if that is the case, there seem to be a tendency for donors provided by the BHG and IMP lists to belong to the clusters 01 and 11, that are associated with higher total amounts and gift donated.

5. Conclusion

After the analysis of the data provided by the PVA organization, we were able to divide the donors in 4 distinct clusters, being their separation made in pairs of two different perspectives: Individual/Personal and Neighbourhood. This perspective separation was a clear attempt of not only understanding the donors themselves, but also the environment in which they are inserted in, as this obviously will impact their behaviour.

Moreover, we not only suggested some marketing recommendations that PVA could implement to attempt to make their lapsed donors donate again (by capturing their interests in different ways), but also, to help the organization to grow their donors' base in a sustainable way, and in an attempt to reach the potential donors' segments that may be of higher interest to the organization, given their behaviour.

Being this said, and although there is some room for a further improvement if the project was done in direct conversation with the PVA organization (as some information was lacking), we believe to have been able to spot some key differences between the segments of donors, that will help PVA understand their donors and improve their operations.

[1] Pew Research Center. *The Generation Gap and the 2012 Election*, 2011. Consulted here.

6. Appendix

Note: These tables do not count towards the page limit, like it was discussed, as they are not part of the report, neither are necessary for the full reading of the same, as they are just auxiliar tables.

Variable Created	Explanation
MONTHS_SINCE_FIRSTGIFT	Created from variable ODATEDW, converting it into months since the first gift, instead of keeping it in date format
AGE	Created from variable DOB, transforming it into years (age), instead of keeping it in date format
DOMAIN_URBANICITY	Created from the first digit of variable DOMAIN
DOMAIN_SOCIAL_ECONOMIC	Created from the second digit of variable DOMAIN. Note that we inverted the order of the "ranking", in order for it to be in line with the remaining ordinal variables
ANCEU	Created from the sum of variables ANC1 to ANC15, in order to capture all the European ancestries in just one variable
MONTHS_SINCE_MINGIFT, MONTHS_SINCE_MAXGIFT, MONTHS_SINCE_LASTGIFT	Converted the variables MINRDATE, MAXRDATE and LASTDATE to months since a specific date, instead of being kept in a date format
RFA_X_Recency, RFA_X_Frequency, RFA_X_Amount	Created from the first, second and third digits from variables RFA_X (with X from 3 to 24), respectively
TIMELAG_0-4, TIMELAG_5-9, TIMELAG_10-14, TIMELAG_15-19, TIMELAG_20-30, TIMELAG_>30	Variables created through the discretization of variable TIMELAG

Table 1 - The variables that were created during the project, and the reasoning behind them.

Variable (s)	Explanation
MALEMILI, MALEVET, VIETVETS, WWIIVETS, LOCALGOV, STATEGOV, FEDGOV	Repetitive information
AGE903, AGE906	Concern the average and median above the 25 years threshold. No relevant information for our goal, given we have a variable with a more natural threshold: 18 years old
CHILC1 to CHILC5	Considering the goal of a simple neighbourhood analysis, variables CHIL1-CHIL3 capture enough information
MC1 to MC3, TPE1 to TPE13, DW1 to DW9, HUR1, HUR2, HUPA1 to HUPA7, ETHC1 to ETHC6, HC1 to HC21, VOC1 to VOC3	For diverse reasons, considering the simple analysis intended, these variables were considered to either have no useful information and/or repeated information (with other variables capturing the relevant information in a more direct way)
OCC1, OCC2, OCC3, OCC4, OCC5, OCC6, OCC7, OCC8, OCC9, OCC10, OCC11, OCC12, OCC13	We decided to either only keep the occupation of the neighbour or the industry in which they work. Based on this, we believed the industry to have a more discriminatory power, so we kept it instead
AC1, AC2	Considering the goal of a simple neighbourhood analysis, it doesn't make sense to keep these variables, having the variables AGEC1 to AGEC7
ANC1, ANC2, ANC3, ANC4, ANC5, ANC6, ANC7, ANC8, ANC9, ANC10, ANC11, ANC12, ANC13, ANC14, ANC15	Concern European ancestry, which was condensed in a new single variable, ANCEU
RHP1, RHP2, RHP3, RHP4, HHP1, HHP2	The scale of these variables was inconclusive, which takes interpretability out of them
NUMCHLD, MBCRAFT to PUBOPP	These variables presented a percentage of NaN values higher than

POP902, POP903, POP90C5, ETH2, ETH13, AGE902, AGE904, AGE905, HHAGE1 to HHAGE3, HHN4, HHN6, MARR3, HV2, HU2, HU4, HHD1 to HHD6, HHD9, HHD11, HVP1, HVP2, HVP4, HVP6, RP2, RP4, IC2 to IC4, IC15 to IC23, HHAS1, LFC2 to LFC5, EC7, AFC2, AFC5, LSC1 to LSC3, CARDPROM,'NUMPROM, CARDGIFT	These variables presented a correlation with other variables, in absolute terms, higher than 0.85
EIC6, RAMNT_3 to RAMNT_6, RAMNT_10 to RAMNT_14, RAMNT_17	These variables did not present any correlation with other variables, in absolute terms, lower than 0.4.
TIMELAG	We discretised this variable, creating new ones, so there is no need to keep it in our analysis.

Table 2 - The reason(s) which lead to each numeric variable to be dropped along the project.