NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

Data Mining

PVA Donor Segmentation

Fábio Lopes (20200597)

Felipe Costa (20201041)

Jorge Pereira (20201085)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ABSTRACT

Clustering and profiling are keys elements to understand and interpret a database. Clustering is a powerful data mining technique that allows us to interpret, understand, segment and characterized data. Marketing strategies can be created with results of a clustering analysis. Several algorithms and methods can be used to produce a good clustering solution. This study aims to segment the PVA donor database and define different marketing strategies for the resultant groups. The k-means algorithm will be applied to identify the clusters. The Python programming language was used to implement the solutions. Pandas and Sickit-Learn are the most used Python libraries on this work. The study identified 4 different clusters. To conclude, the Cluster Profiling was performed, and the marketing strategy was defined.

KEYWORDS

Data Mining; Clustering; Marketing; k-Means; Python; Sickit-Learn; Pandas; Cluster Profiling.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

PVA	Paralyzed Veterans of America
KDD	Knowledge Discovery in Databases
t-SNE	t-Distributed Stochastic Neighbor Embedding
PCA	Principal Component Analysis
KNN	k Nearest Neighbors

1. INTRODUCTION

Paralyzed Veterans of America (PVA) is an organization found in 1946 by World War II veterans who returned home with spinal cord injuries and found that the world had few solutions to the major challenges that they faced. Since the start, they have provided a better future for injured veterans by aiding in the medical research, advocacy, and civil rights for all people with disabilities [1].

Being one of the largest direct mail fundraisers of the United States of America, the organization relies on the donations of individuals to maintain and expand their work so, retaining existing donors is just as import as acquiring new ones. To that extent, the organization has previously employed segmentation strategies on their current donors to target specific donors with premium gifts to capture more donations.

From the 13 million donor base, PVA has extracted a sample of donors which are considered Lapsed, with no donation in the last 13 to 24 months ago. This group of donors is of special importance since the longer a donor remains without donating, the less likely it is for them to re-donate.

This study aims to use Clustering Algorithms and Profiling (2 of the canonical tasks of Data Mining), to cluster the donors into clusters of similar profiles, investigate how these clusters behave and define a marketing strategy to be used by PVA to regain the donors. For Han, Jiawei et al. [2], clustering can be used to assign class labels for a group of data. The process is defined so objects within a cluster have high similarity in comparison to one another but are rather dissimilar to objects in other clusters.

The k-means algorithm will be used to identify the clusters. The Sickit-Learn implementation of this algorithm will be preferred. The k-means algorithm clusters data by trying to separate samples in groups of equals variance, minimizing a criterion known as the inertia, or within-cluster sum-of-squares. K-means requires a pre-defined number of clusters and scale well to large number of samples [3].

At the end, the Clusters Profiling will be performed, and the clusters will be characterized. On this step the data used to cluster the donor will be compared across the clusters to identifies similarities and patterns that can be used to understands the groups.

Finally, a marketing strategy will be define based on the resultant clusters. The goal is identifying the donor's behavior on each cluster and design strategies to achieve better results.

All the code used to perform this study is available on a GitHub repository. The link is provided in the appendix section 5.2.

2. METHODOLOGY AND RESULTS

This works uses the Knowledge Discovery Process [4] as a process template to better extract knowledge from the dataset. KDD provides a reference to the Data Mining process by splitting the tasks in 9 consolidated steps.

1. Develop an understanding of the data and identify the goal of the KDD process.
2. Create a target dataset focusing on a subset of features and/or data samples.
3. Data cleaning and processing of the dataset to remove noise and/or outliers.
4. Data reduction and projection where the most useful features for the knowledge extraction will be identified.
5. Identification of the type of the task to be performed according to the step 1. This can vary between classification, regression, clustering, etc.
6. Exploratory analysis and hypothesis selection where the data mining algorithm will be chosen.
7. Data Mining, where the algorithm chosen will be applied to the dataset.
8. Interpretation of the mined patterns, where the results of the mining will be visualized and interpreted.
9. Acting on the discovered knowledge, where the knowledge acquired will lead to an action.

For this work, steps 1 and 2 are already concluded by the faculty, where the goal of the KDD is already defined, and the dataset provided is already a representative sample of population that we want to analyze. The conclusion of this work, as in step 9, will be to provide a simple marketing strategy for the PVA to target its customers.

The structure of this work is split in 6 different categories:

1. Data Exploration: an initial look into the dataset will be done to better understand the data at hand.
2. Data Preparation: clean the noise and outliers from the dataset and perform feature engineering in some of the variables.
3. Initial Clustering: create a first clustering solution.
4. Initial Profiling: gain a better understanding of each cluster originated in step 3 and evaluate new features to use in a new clustering approach.
5. Clustering: Using the knowledge gained in step 4, re-cluster the data.
6. Profiling: Extract the knowledge from the dataset using as guides the clusters computed in the previous step.

2.1. DATA EXPLORATION

The first analyzed column is the gender of donors which shows that most donors are Female, with over 51000 women donating. This represents almost 54% of the values. For Males, the percentage is 41%, being this around 39000 male donations. 5% of the dataset is split by Unknow or Joint Account.

Donations per State can help visualize the geographical distribution of the donors that are more active in donating. California, Florida and Texas are the states with the highest donation numbers.

Household Income and Wealth [Figure 1] are two variables that might be able to explain the behavior of the donors due to a higher financial capability of donating money when these values increase. From the dataset, regarding Household Income, there is an equilibrium of low and high income donors and, the most represented segment is the middle class. As for the Wealth Rating, the dataset is unbalanced to the higher wealth rating, having almost double the donors of the highest rating, compared to the lowest.

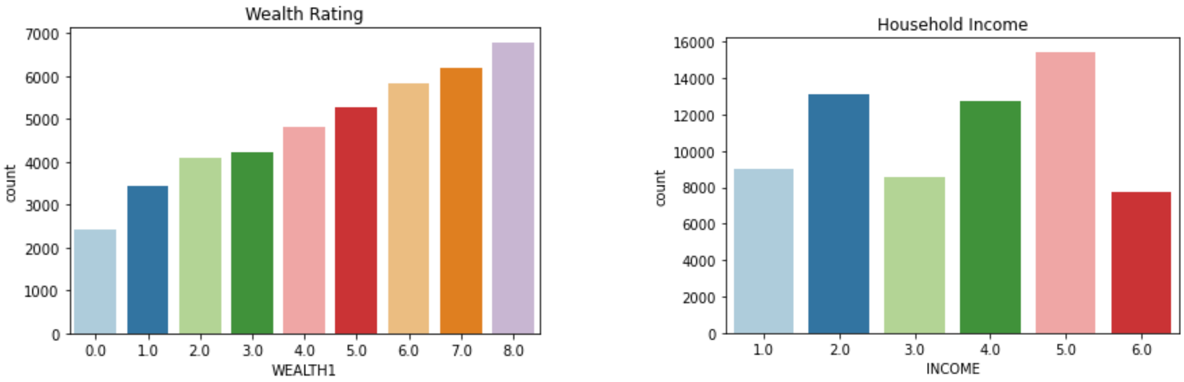

Figure 1 - Count of donors per Wealth Rating and Household Income

Analyzing the RFA_2F variable can give an understanding of the typical behavior of the sample of donors in the last period of recency. Close to half of the donors consider the act of donation as a one-time event and, as expected the least represented segment is of donors with 4 or more donations.

By analyzing the RFA_2R feature, all the donors in the sample are Lapsing donors.

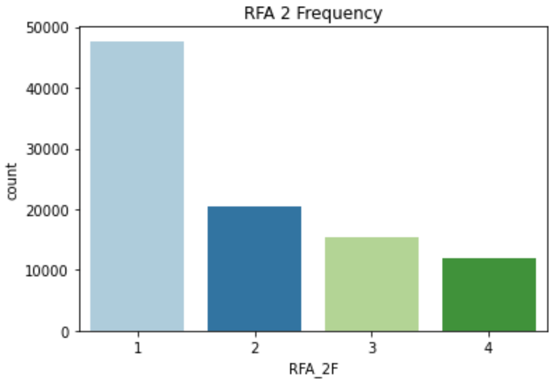

Figure 2 - Count of donors per Recency Frequency

In the dataset, there are a total of 475 features, where 350 are numerical and the remaining 125 are categorical.

A close inspection on each column is required to verify the data and by inspecting the columns, there are 92 features with missing values and, 62 features have more than 50% of missing values.

2.2. DATA PREPARATION

The original dataset has 475 features, which is a very high number to work with so, to start, a manual inspection was performed on the dataset in conjunction with the analysis of the metadata. In this inspection the features were investigated and were dropped based on their irrelevance. For example, there are sets of variables which are summarized in another set of variables, like ADATE_x, RAMNT_x, etc. After this process, 341 features remain in the dataset.

To prepare the data, the following operations were performed:

- Convert all the datetime objects into elapsed number of days until the current timestamp. For the Date of Birth, the Age of the donor in years was calculated.
- All Categorical Variables which represent a binary variable with a letter were replaced to proper 0 and 1 binary representation.
- Cleanup the ZIP variable which contains dashes improperly placed.
- For the SOLIH and SOLP3 variables, the assumption was made that if the user has selected the possibility of mailing, it is possible to send email up to a maximum of 12 occasions per year.
- From the DOMAIN variable, 2 ordinal variables were created: URBANICITY which states the type of neighborhood where the donor resides and SOCIOECON which states the Socio-Economic status of the of the donor's neighborhood.
- RFA_2R was dropped since its variance is 0. All donors are lapsing donors.
- Creation of ordinal variables from the RFA_2A, MDMAUD_R, MDMAUD_F, MDMAUD_A categorical variables.

After the data was prepared, the MinMax scaler was used to normalize all the numerical features of the dataset to the range $[-1,1]$.

The next step in the preparation is reduce the dimensionality of the dataset to be able to use clustering algorithms without incurring in the curse of dimensionality, where more features will lead to a sparser dataset in a high dimensionality space. The follow techniques/methods were used to reduce the dataset:

- Drop any variable in which the missing values represent more than 40% of the observations.
- Analyze the Pearson's correlation matrix and:
    - If 2 variables have an absolute correlation greater than 0.9, drop one of them since both can explain the same behavior in the data.
    - If the maximum absolute correlation between 1 variable and all the other variables is less than 0.3, drop the variable since it can't explain the behavior of the data.
- All the Neighborhood features, URBANICITY and SOCIONECON were dropped since the focus is to segment the behavior of the donors.
- Flag and remaining categorical features were dropped.

After this process, the dataset contains 20 numeric features for 95412 observations.

Table 1 - List of features to perform the initial clustering and their description

Feature	*Description*
INCOME	Household Income of the donor
WEALTH1	Wealth Rating of the donor
WWIIVETS	% of WWII Vets
FEDGOV	% of Employed by the Federal Government
WEALTH2	State specific Wealth Rating
CARDPROM	Lifetime number of card promotions received to date
CARDPM12	Card promotions received in the past 12 months
NUMPRM12	Lifetime number of promotions received to date
RAMNTALL	Dollar amount of lifetime gifts to date
NGIFTALL	Number of lifetime gifts to date
MINRAMNT	Dollar amount of smallest gift to date
MAXRAMNT	Dollar amount of largest gift to date
LASTGIFT	Dollar amount of most recent gift
AVGGIFT	Average dollar amount of gifts to date
RFA_2F	Number of gifts in the period of recency
elapsed_MINRDATE	Number of days elapsed since the smallest gift to date
elapsed_MAXRDATE	Number of days elapsed since the largest gift to date
elapsed_LASTDATE	Number of days elapsed since the last gift
LASTGIFTAMOUNTCATEGORY	Donation category for the last gift
MDONOR_GIVING_RECENCY	Recency category for the Major donors

By reducing the feature space, it is now possible to use the KNN imputer to replace all the missing values with the mean of the 5 nearest neighbors.

To perform the detection and removal of outliers 2 methods were tested:

- Inter Quartile Range Outlier detection
    - This method provided poor results in the dataset since it classifies about 70% of the dataset as outliers. By analyzing the process variable by variable, it was possible to

identify that most of the variables were classifying about 2/3% of the dataset as outliers, while other were close to 40%.

- Isolation Forest
    - With the knowledge gained by using IQR, the final process to remove the outliers was an Isolation Forest with a 0.1 contamination factor.

2.3. Initial Clustering

On this step, the goal was to perform an initial clustering approach to segment the donors based on the 20 variables remaining from the last step. This number of variables is still sub-optimal, but it will be a starting point to further analyze the donors, and later improve the approach by selecting the variables which better segment them.

The algorithm chosen to do this first approach was the k-Means. The reasoning behind this decision considers that, the goal of the segmentation is to find donors which resemble each other in the database, as in, finding donors which are the closest to each other taking into reference a middle point which unites them. In this case, k-Means is the appropriate algorithm to use since, other algorithms, like DBSCAN, will try to segment the customers by finding shapes of similar donors, which would not produce great results.

Since k-Means requires the number of clusters to be set à priori, the Elbow method was used to find a good solution. The Elbow method is an heuristic which helps determining the numbers of clusters to use in a specific dataset by computing the inertia for a range of number of clusters (inertia is the criterion that the k-Means algorithm tries to minimize and, can be defined as the within cluster sum of square distances) and choosing the one where an increase of the number of clusters does not significantly reduces the inertia.

In this work, the range of cluster numbers selected as $[1,11]$ and, in Figure 3 - Inertia for a range of number of clusters, we can see the plot of the inertia for the range specified. Cluster Number 4 is when there ceases to exist a significant drop in the inertia by increasing the number of clusters so, this was the number of clusters selected.

With the clusters defined, the distance of the outlier samples to each of the clusters was computed and the centroid with the smallest distance to each sample was identified. As such, all the outlier samples were assigned to each computed cluster and returned to the dataset.

Since the clustering occurred with 20 variables, to visualize the results a dimensionality reduction algorithm had to be used. In this work, both t-SNE and PCA were utilized.

t-SNE [5] is an algorithm used to visualize high-dimensionality data by converting similarities between data points to joint probabilities and minimizing the Kullback-Leibler divergence between the joint probabilities of the low and high dimensional data.

PCA [6] is an algorithm used to reduce the dimensionality of a dataset to its Principal Components by using Singular Value Decomposition of the data to project it to a lower dimensional space.

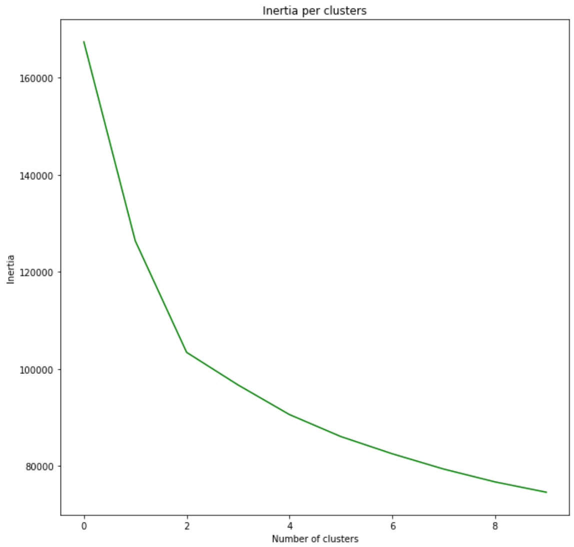

Figure 3 - Inertia for a range of number of clusters

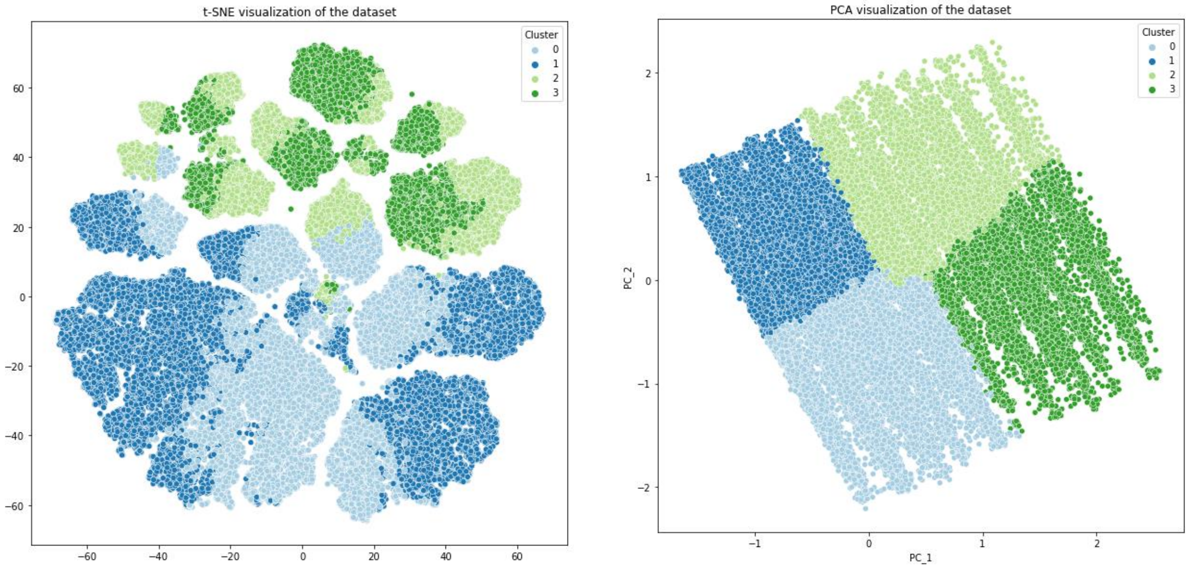

Figure 4 - t-SNE (left) and PCA (right) representation of the high dimensional dataset in a 2d space

In both algorithms, the goal is to reduce the number of features from 20, to a 2d representation of the dataset where each new 2d point will labelled with the cluster assignment and plotted in a scatter plot where each cluster will be colored differently.

In Figure 4, we can see the resulting visualization for both t-SNE and PCA. In both, we can see that there is some homogeneity in the datapoints associated with each cluster and, in the case of PCA, we can clearly see the separation of the clusters.

2.4. Initial Profiling

On this step the clusters were analyzed, and the features were compared to find the characteristics that distinguish them.

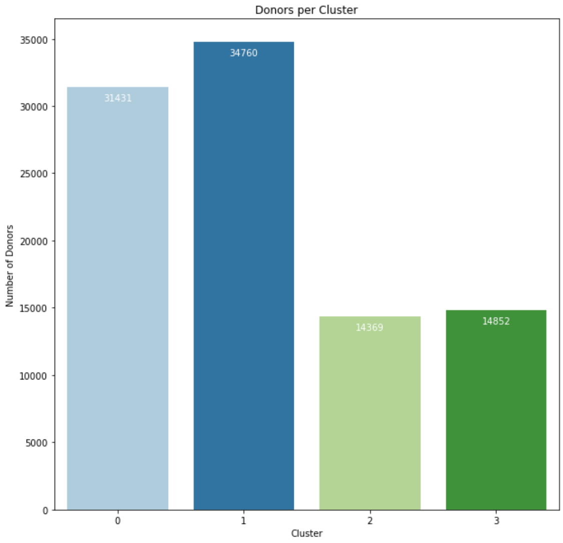

Figure 5 - Donors distribution per Cluster

The donor distribution by cluster [Figure 5], where the Cluster 1 is the one with the highest number of donors, and the Cluster 2 has the smallest number. Around 69% of the donors were included in Cluster 0 and 1, while Clusters 2 and 3 contain about 31% of the donors.

To assess and identify the most important feature a Decision Tree Classifier [Figure 6], was created using the features that were considered on the Clustering. The response variable was the Cluster Label. The classifier was able to predict 82.79% of the donors correctly. The features there are more discriminant for the classifier were: RFA_2F, WEALTH1, INCOME.

All the other features were not relevant to the classifier, and as such, we decide to exclude them and run the clustering algorithm again only with these three features.

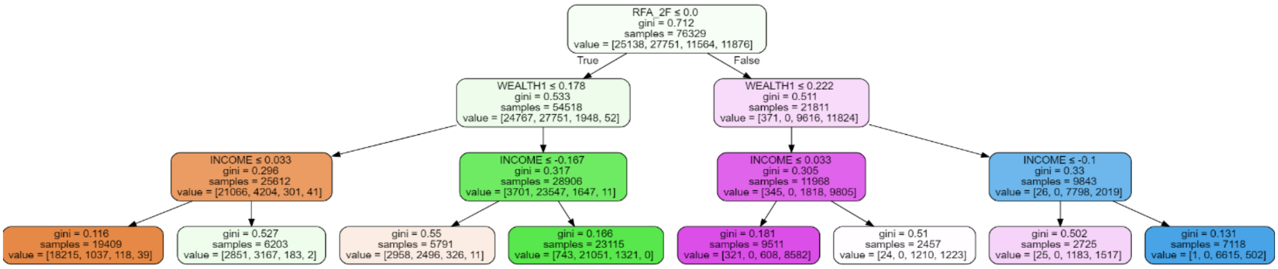

Figure 6 - Classification Tree for the first Cluster Solution

2.5. CLUSTERING

Considering the knowledge gained in the previous step, the clustering was performed once again only using the 3 most import variables.

The same methodology was used as in Section 2.3. The process began by computing the Inertia for a range of cluster numbers and analyzing the Elbow Method [Figure 7]. And, as before, the number of clusters which will be used is 4.

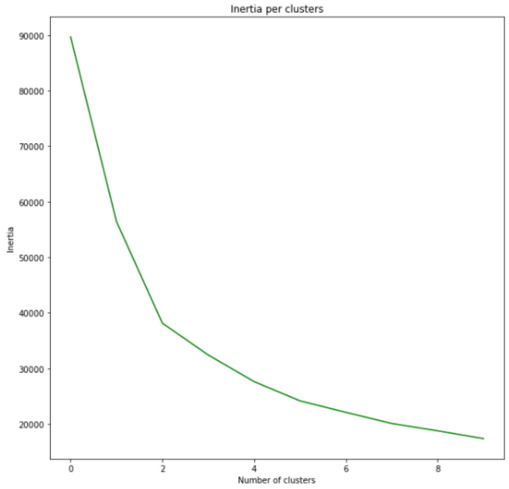

Figure 7 - Inertia for a range of number of clusters after feature selection

The outliers were classified using the minimum distance from each sample to the centroids and returned to the dataset.

2.6. PROFILING

To understand and characterize the Cluster, several features and relationships were explored. The goal was to find the important features that could be used to distinguish the clusters.

The donor distribution by cluster [Figure 8], where the Cluster 2 is the one with the highest number of donors, and the Cluster 1 has the smallest number. Around 71% of the donors were included in Cluster 0 and 3, while clusters 1 and 3 contain about 29% of the donors.

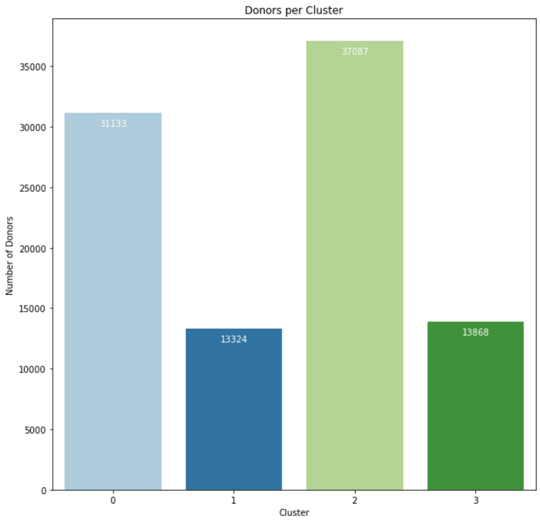

Figure 8 - Donor Distribution between the clusters

To analyze the geographical distribution of the clusters within the USA [Figure 9] the mode of each cluster was computed by State and we can see that Cluster 1 is not the most common cluster in any state. Cluster 3 is the most common cluster in the states on New Hampshire and Vermont, while we have a good balance in the distribution of Cluster 0 and 2 throughout the country.

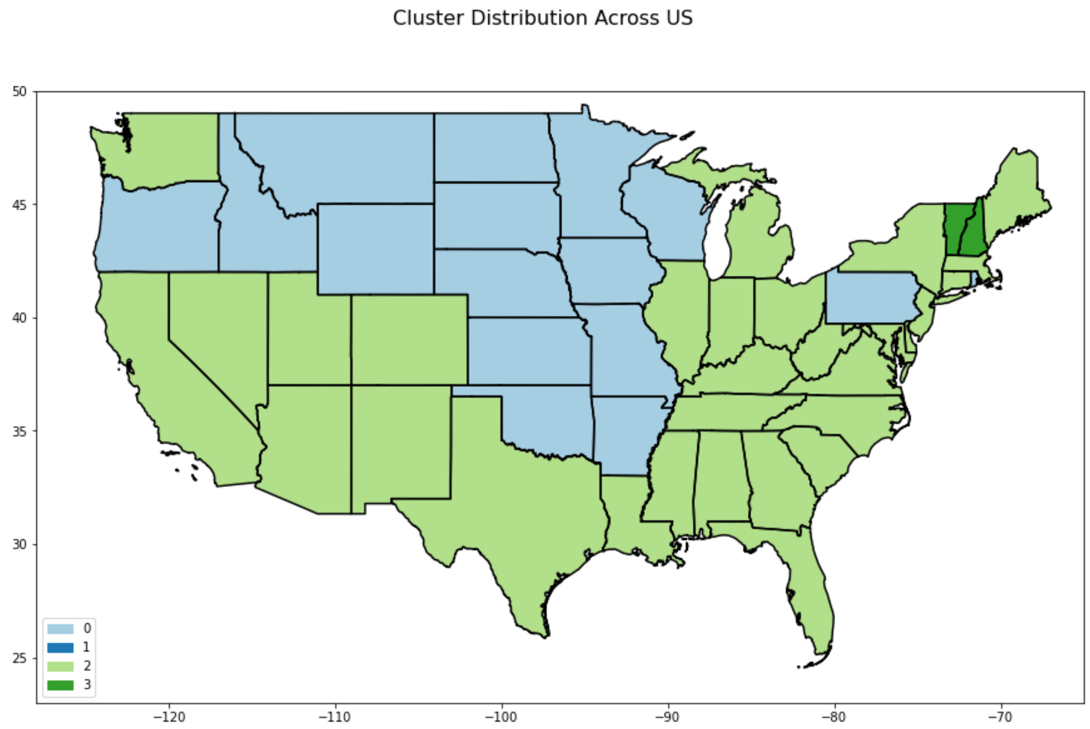

Figure 9 - Donors distribution Across the US

The comparison between INCOME, WEALTH, RFA_2F and LASTAMOUNTGIFCATEGORY and the database average [Figure 10], paints a clearer picture on the profile of each cluster. Although the latter 2 variables are ordinal, one can still take conclusions from the average order of the clusters.

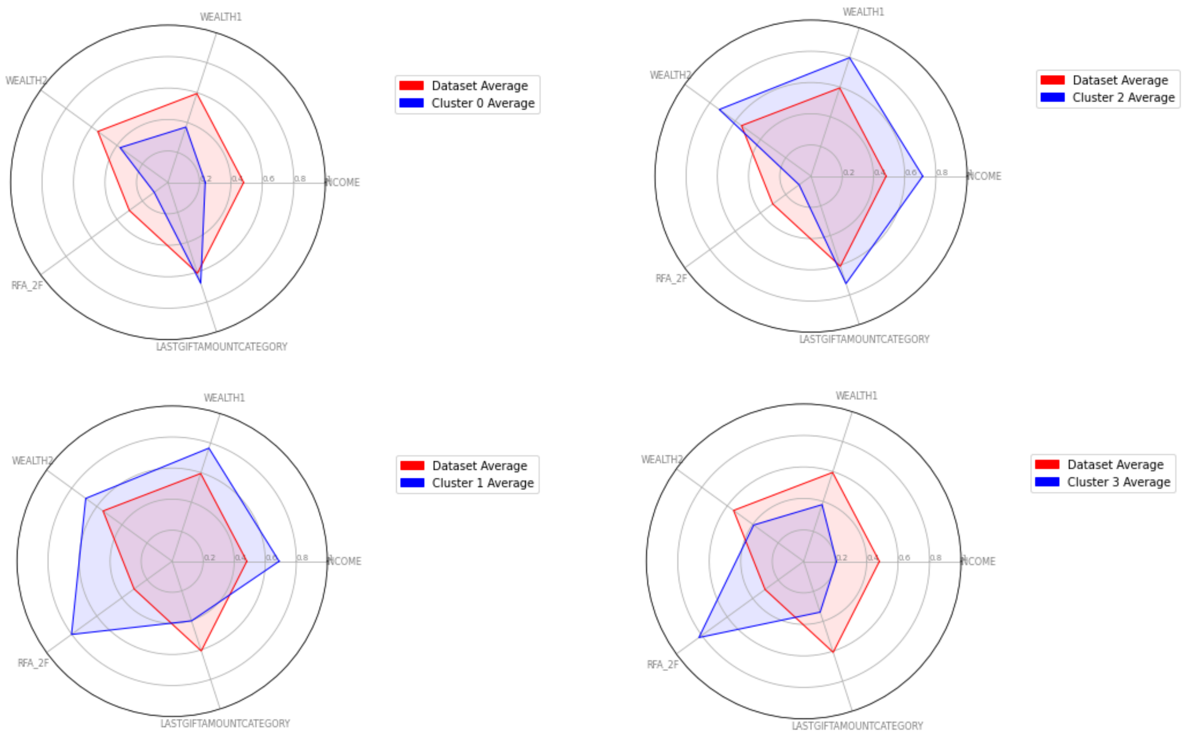

Figure 10 - Scaled Average of cluster on selected features versus the dataset average

Cluster 0 donors are neither donors with a high Income nor does they have a high frequency of donations but, the last donation given was of a high value. Cluster 1 donors have a high Income and the frequency of donations is high but, the last donation was below average. Cluster 2 donors have just as high as Income as Cluster 1, but unlike the latter, they have a very low frequency of donating coupled with a high last donation made. Cluster 3 donors have a low income and the last donation was below average, but they make more donations than all other clusters.

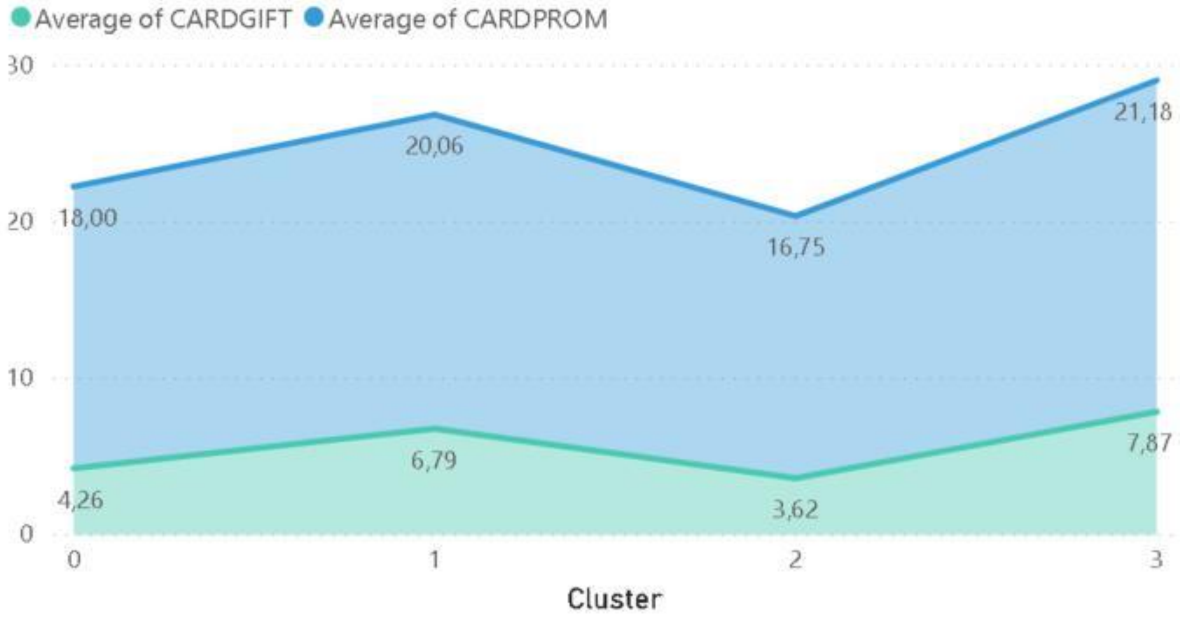

Figure 11 - Susceptibility to Card Promotions

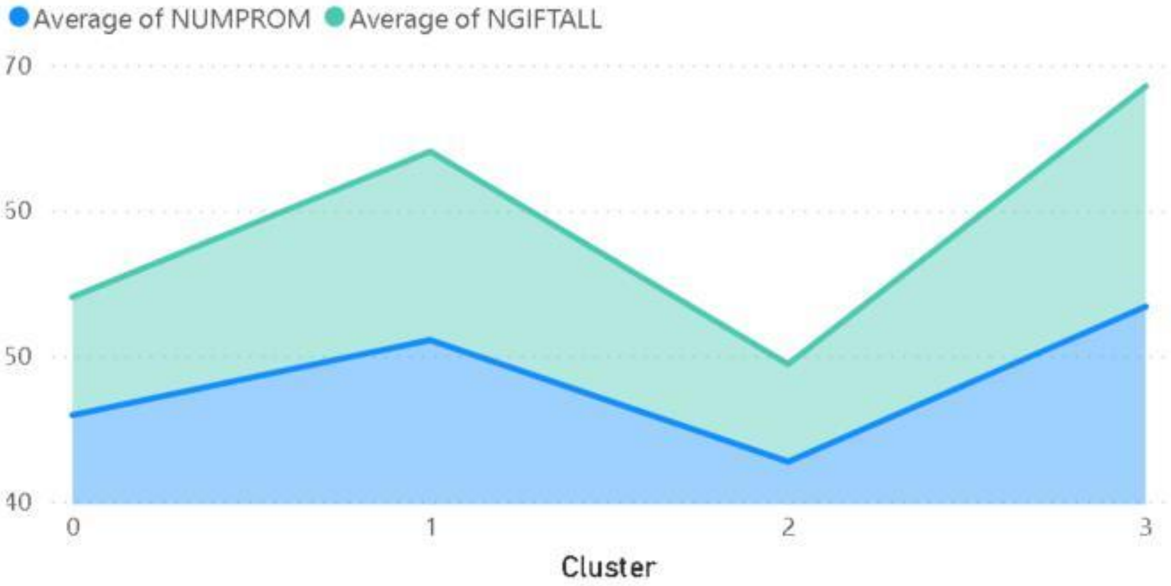

Figure 12 - Susceptibility to Promotions

On Figure 12 is possible to observe the distribution of the features NUMPROM (Lifetime number of promotions received to date) and NIGIFTALL (Number of lifetime gifts to date), also that Clusters 1 and 3 are the ones that gave more donations per promotions. It's also possible to conclude that sending card promotions has no positive effect on the number of donations [Figure 13] since the response to card promotions is very similar to the response to overall promotions.

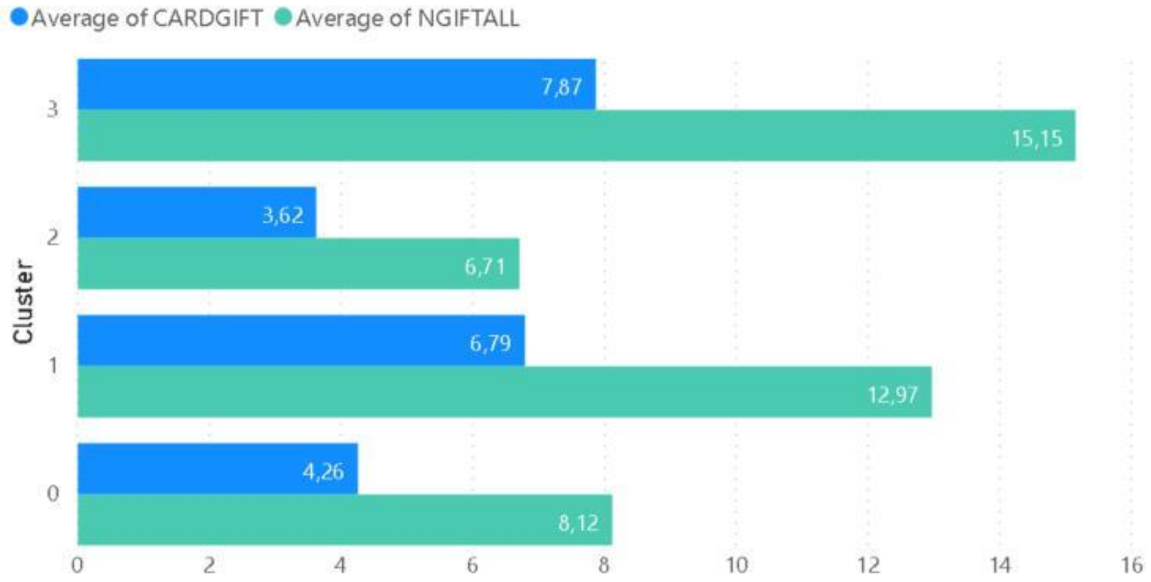

Figure 13 - Comparison between Response to Card Gift vs All gifts

The feature RAMTALL is the dollar amount of lifetime gifts to date. On Figure 14 is possible to observe that Cluster 3 contains the donors with the largest donations. As observed on previous charts this Cluster also receives more promotions and with the better response but, although Cluster 0 and 2 don't response well to promotions, their average donation value is close to the others.

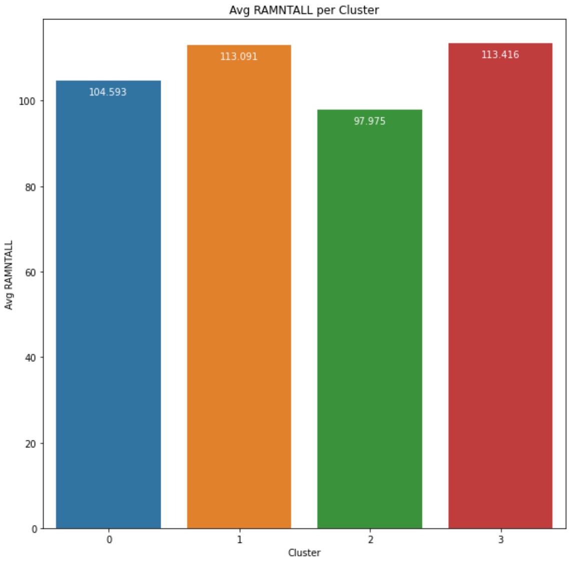

Figure 14 – Average Lifetime Dollar Amount Donated

3. CONCLUSIONS

In this work, several Data Mining techniques were applied to segment donors on a sample database from the PVA association.

With the clusters defined the marketing strategy recommended is:

1. Rework the Gift Card sent, since the response of all the clusters to these card gifts do not seem to increase the donation frequency or amount in comparison with the normal promotional material.
2. Clusters 1 and 2 seem to have a large unused potential since they have the Highest Income of all the clusters and:
    a. Cluster 1 has a high frequency of donations but a lower donation amount per gift. The strategy for this cluster, should be to increase the number of times a donor donates, even in smaller amounts. This cluster has a high response to promotions sent so a redesign and customized card that encourages to give a recurring amount by appealing and raising awareness about how this money can change association members life and by sharing PVA stories. This may increase the value of this cluster in a long run.
    b. Cluster 2 has a low frequency but a high donation amount per gift. The strategy here should be to drive the frequency up, but this cluster has a low reception to promotions sent so, a new approach must be devised by the association. Is expected that offering multi ways to give, so donors can put their preferences first and personalized messages and thank you notes can be a great start to engage and make this cluster to give more often.
3. Clusters 0 and 3 are the segments where the approach is less clear due to the low Income compared to the other segments. From the data we can gather that:
    a. Cluster 3 has the highest response to promotional material sent and a very high frequency of donations. Due to the low income, it might not be possible to increase the amount per donation, but a focus on increasing the promotions sent might help drive the total donation amount of this segment.
    b. Cluster 0 has the lowest donation frequency of all and has a low response to promotional material while maintaining a high total donation amount, which leads to believe that these donors give what they can, when they can and are not susceptible to a marketing campaign targeted to increase their frequency or donation amount. Despite of that, is recommended more emotional appealing communications to create more opportunities for this cluster to relate to the association and consequently donate.

Overall, the biggest challenge on the study was the high dimensionality of the dataset. Right at the start of the work, the focus shifted from applying different clustering algorithms and experimenting with different solutions to, reducing the dimensionality of the dataset without losing valuable information. This occupied a significant chunk of the project timeline and, as such, the experimentation with different clustering algorithms and solutions was reduced to two attempts. The culprit of this was the initial reliance on PCA to reduce the number of features and then, performing the clustering on top of the Principal Components but, as time passed, it was soon realized that profiling these clusters created from synthetic variables would be troublesome and a restart on the analysis was required.

In the end, this study reflected a near real world scenario and it greatly helped understanding the struggles of a real-world dataset and to have a better comprehension on how to deal with them.

4. BIBLIOGRAPHY

[1] "Paralyzed Veterans of America," [Online]. Available: https://pva.org/.

[2] J. Han, M. Kamber and J. Pei, in *Data Minning: Concepts and Techniques*, Walthman, Morgan Kauffman, 2012.

[3] "Scikit Learn," [Online]. Available: https://scikit-learn.org/.

[4] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, AAAI Press / The MIT Press, 1996, pp. 1-34.

[5] L.J.P. van der Maaten and G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research,* pp. 2579-2605, 2008.

[6] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* pp. 611-622, 1999.

5. APPENDIX

5.1. PROJECT DELIVERABLES

- Jupyter Notebooks
  - 1-DataExploration: Initial overview of the data
  - 2-DataPreparation: Preparation of the input data for the clustering algorithm
  - 3-Clustering: Initial Clustering attempt
  - 4-Profiling: Initial Profiling to detect relevant variables
  - 5-ReClustering: Clustering of the data using the variables chosen in the Profiling
  - 6-ReProfiling: Profiling the Cluster solution obtained in the previous step
- Configuration Files
  - ColumnsToKeep: List of columns to drop from the source dataset which were selected to be dropped in the first analysis
  - NeighborhoodColumns: List of the Neighborhood columns
  - US_Name2Code: Mapping between ISO State Code to the State Name to be used in the Geography Visualization
  - US-ShapeFile (directory): Location of the Shape File to draw the map of the United States
- Excel file
  - Initial Analysis of the variables to Keep/Drop

5.2. GITHUB REPOSITORY

https://github.com/JorgeMPereira/DMProj-2020-BR