NOVA IMS Information Management School

DATA MINING PROJECT REPORT

Clustering of PVA's lapsed donors

2020 - 2021

Paralyzed Veterans of America

MISSION: ABLE

Authors :
Filipe Dias (mm20200655@novaims.unl.pt)
Edgardo Juarez (mm20200749@novaims.unl.pt)
Lilou Dang-Thai (m20200743@novaims.unl.pt)

INTRODUCTION

The non-profit organization, Paralyzed Veterans of America, is fundraising campaigns to help US veterans with injuries and diseases. Their wish is to understand the behavior of their lapsed donor — people who donated to 13 to 24 months — and to find an approach to recapture their interest in order to have new incoming gift. The team was given a dataset of 476 features on 95 412 donors. More than half of the features are regarding the neighborhood of the donors. Neighborhood has an impact on every individual in their daily life and therefore has a significant influence on their tendency to donate. Our purpose is to understand:

How neighborhoods are influencing our lapsed donors on their donations? What type of areas should PVA target, and what marketing approach should they use to reactivate their lapsed donors?

DATASET EXPLORATION

Our dataset is composed of 95 412 donors and 476 variables. The variables are divided into 3 parts, around 15% of them are about the donors individually, the second part and most important is the data regarding the neighborhood of the donors provided by the most recent census and the last part is about the gifts' donors.

FEATURE SELECTION

After going through the features of the dataset and analyzing some of the social variables such as marital status, the ethnicity, the military veterans and the active military in neighborhoods we noticed the either really high or low correlation between them.We then decided to focus on economic features of the neighborhoods' donors and the gifts related features.

NEIGHBORHOOD FEATURES

From all the dataset we picked:

	IC1	HHAS4	VOC2	LFC1	EC1	EIC9	HV2	MHUC1
0	307	1	77	56	120	6	635	6
1	1088	3	92	70	160	11	5218	20
2	251	11	65	65	120	3	546	9
3	386	20	43	69	120	3	1263	16
4	240	14	45	61	120	2	594	6

Figure 1 : Neighborhood dataset head

Neighborhood features description

- IC1: Median Household Income in hundreds per year.
- HHAS4: Poverty level
- EC1: Years of education
- HV2: Average home value in hundreds
- EIC9: Percentage employed in Finance, Real Estate, Insurance.
- MHUC1: Median Homeowner Cost w/ Mortgage per Month dollars.
- LFC1: Percentage of people in labour force.
- VOC2: Percentage of households with more than vehicles.

These features describe the economic lifestyle of the neighborhood, they will permit us access to the behavior of the lapsed donors due to their environment development. Our features give us an idea of the donor's property, if they consume and what kind of expenses they have.

GIFT FEATURES

For the ones related to the donor's gift we picked up:

	RAMNTALL	NGIFTALL	FISTDATE
0	240.0	31	2009-11-01
1	47.0	3	2013-10-01
2	202.0	27	2010-01-01
3	109.0	16	2007-02-01
4	254.0	37	1999-03-01

Figure 2 : Gift dataset head

Gift features description

- RAMNTALL: Dollar amount of lifetime gifts to date
- NGIFTALL: Number of lifetime gifts to date
- FISTDATE: Date of first gift

```
RangeIndex: 95412 entries, 0 to 95411
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   RAMNTALL  95412 non-null  float64
 1   NGIFTALL  95412 non-null  int64
 2   FISTDATE  95410 non-null  object
 3   IC1       95412 non-null  int64
 4   HHAS4     95412 non-null  int64
 5   VOC2      95412 non-null  int64
 6   LFC1      95412 non-null  int64
 7   EC1       95412 non-null  int64
 8   EIC9      95412 non-null  int64
 9   HV2       95412 non-null  int64
 10  MHUC1     95412 non-null  int64
```

Figure 3 : Dataset types

These features describe the frequency and the monetary values of donations to the PVA. We are taking FISTDATE variable as a measure of how much time they have been supporting the foundation.

From the figure below we can see that most of our data is integer type, apart from RAMNTALL and FISTDATE. The FISTDATE had 2 missing values, which were eventually dropped.

DATA PREPROCESSING

OUTLIER CHECK

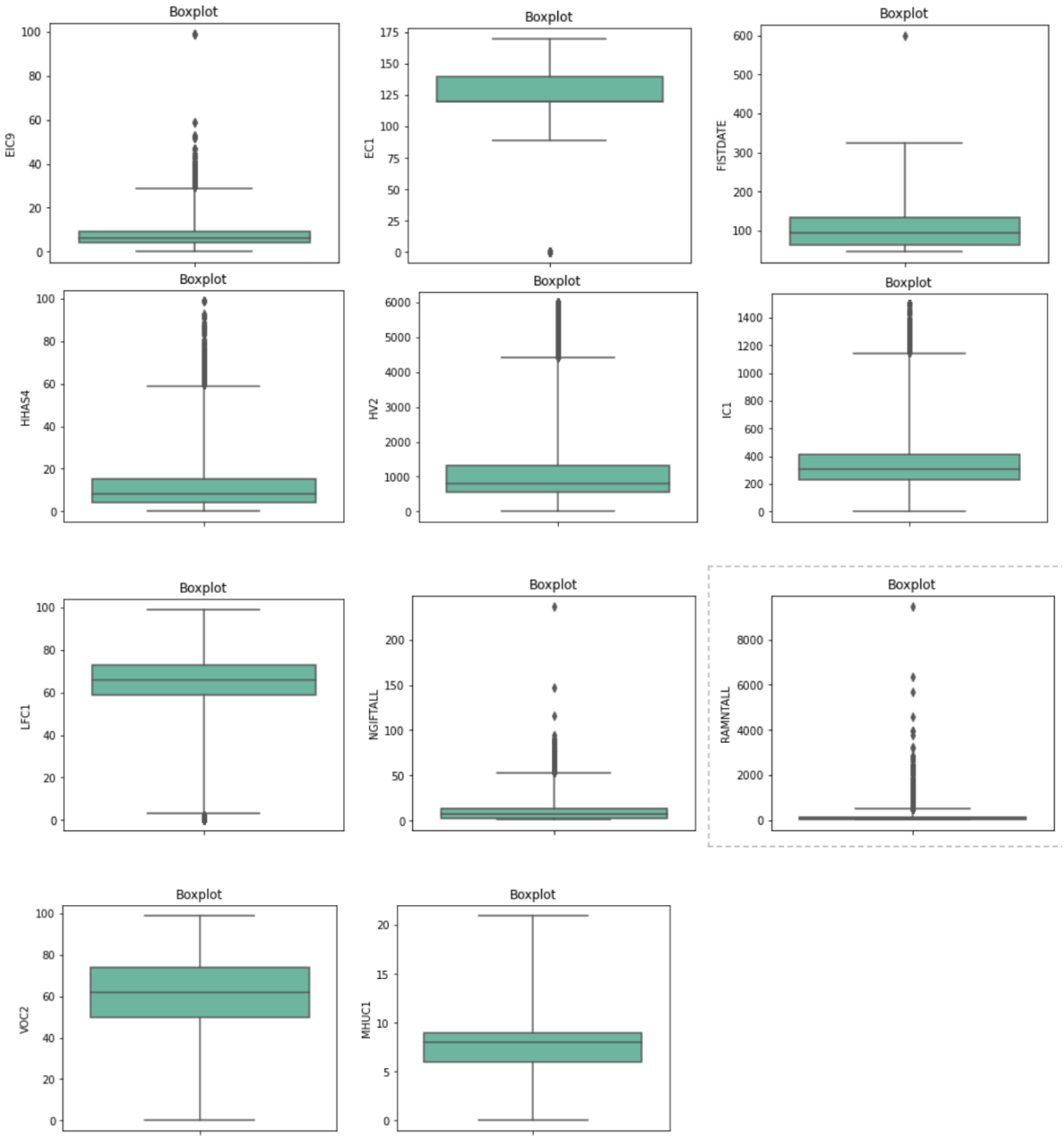

Figure 4 : Box-plot of our variables outliers

In this box-plot we can see that most of the features have outliers but VOC2 and MHUC1. RAMNTALL and NGIFTALL show many outliers. These box-plots had the whiskers set to 4xIQR.

OUTLIER HANDLING

As a multivariate outlier detection method, we tried two approaches:

- **LOF —** Local Outlier Factor is a semi-supervised density-based algorithm to detect outliers. This method computes the distances based on the number of "**n_neighbors"** we define.

- **Isolation Forest** – It's a tree-based unsupervised algorithm, it is build based on decision trees and random forests.

As univariate method we used the 'Tukey's method which is based on the IQR.

Tukey's box plot method: *'Next to its visual benefits, the box plot provides useful statistics to identify individual observations as outliers. Tukey distinguishes between **possible** and **probable outliers**. A **possible outlier** is located between the **inner** and the **outer fence**, whereas a **probable outlier** is located **outside the outer fence**.'*

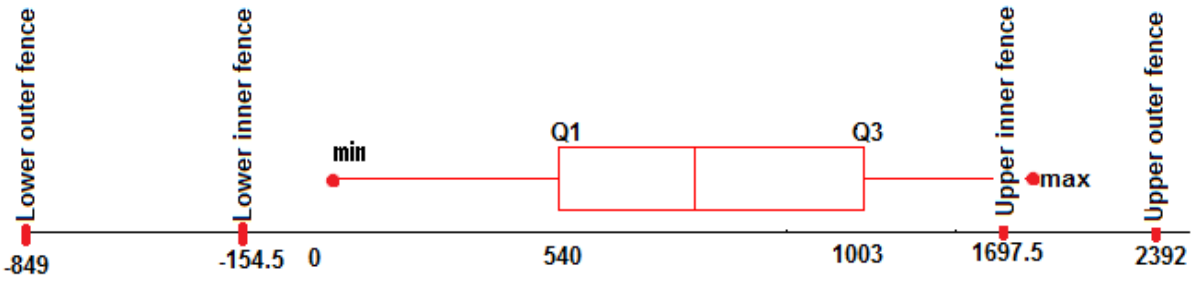

Figure 6 : Tukey's box-plot

CLEAR OUTLIER CASES

In the case of **EC1:**

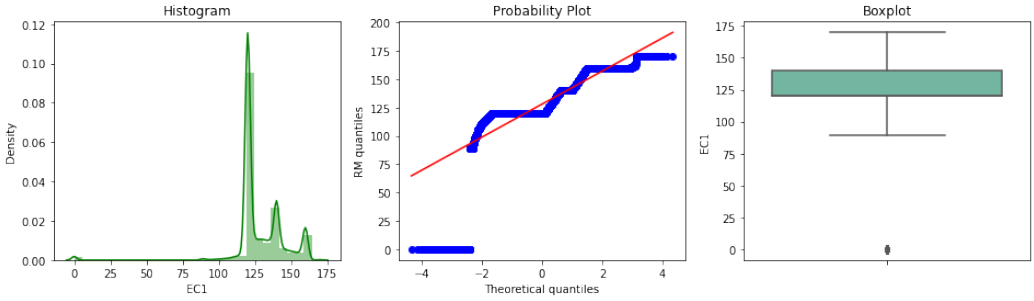

Figure 7 : EC1 after outlier removal

In the case of **EIC9**:

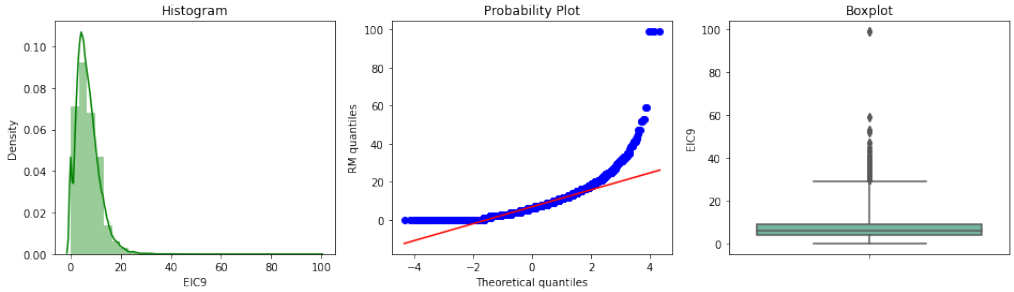

Figure 8 : EIC9 after outlier removal

In the case of **FISTDATE:**

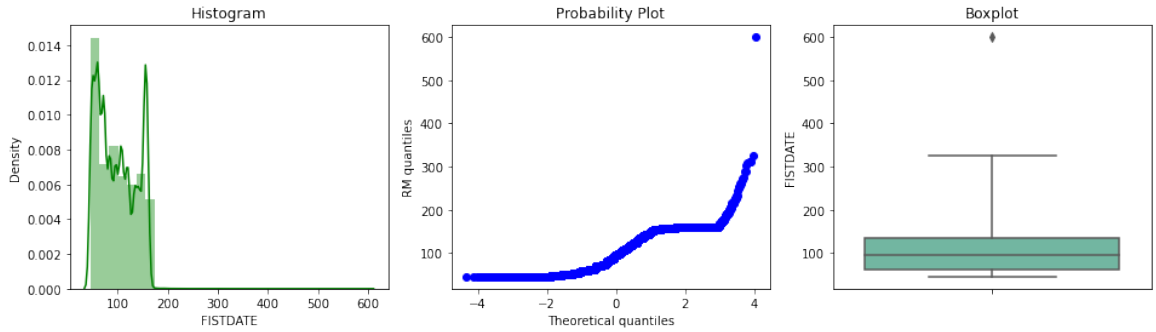

Figure 9 : FISTDATE after outlier removal

In the end we applied Tukey's method with the outer fence set to 4 times the IQR, this way the removal of data was not so aggressive. Because 'Tukeys' based on IQR, we removed the global outliers, to remove the local outliers, on top of 'Tukey' we used the **Sklearn.neighbors-LOF** as a multivariate outlier removal, with the following parameters: "n_neighbours = 20, contamination = 0.03, metric = 'Manhattan'. After some trial and error, we found out that these parameters were the best for our data, we had to set the threshold to 0.03 so it did not remove too much data on top of the univariate.

The method was used on all variables. After clearing the probable outliers, we had a data loss of 4.6%, after applying LOF with a threshold of 0.03 we had a total data loss of 7.6%.

Feature IC1, here we have an example of the difference in the data distribution before and after outlier removal, so we take that outlier removal is important, especially for some algorithms like K-means.

Before:

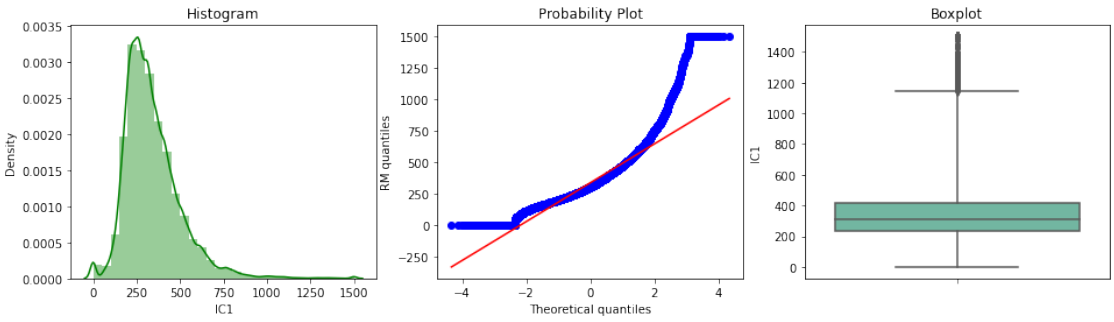

Figure 10 : IC1 before outlier removal

After:

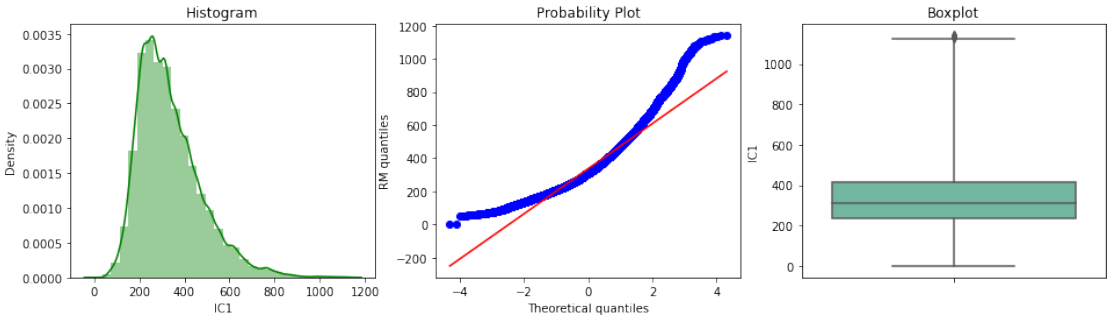

Figure 11 : IC1 after outlier removal

FEATURE ENGINEERING

First date: We did transform the dates into months considering the year was 2020, with this we know how long ago the first donation was made and it also tells us if that donor is an old or recent donor.

EC1: After reading the description and exploring EC1, we noticed that the values were not possible for education years, after some research we found out that one possible cause is the fact that storing floats in a database is expensive, so after dividing the feature by 10, the numbers make sense.

CORRELATION MATRIX

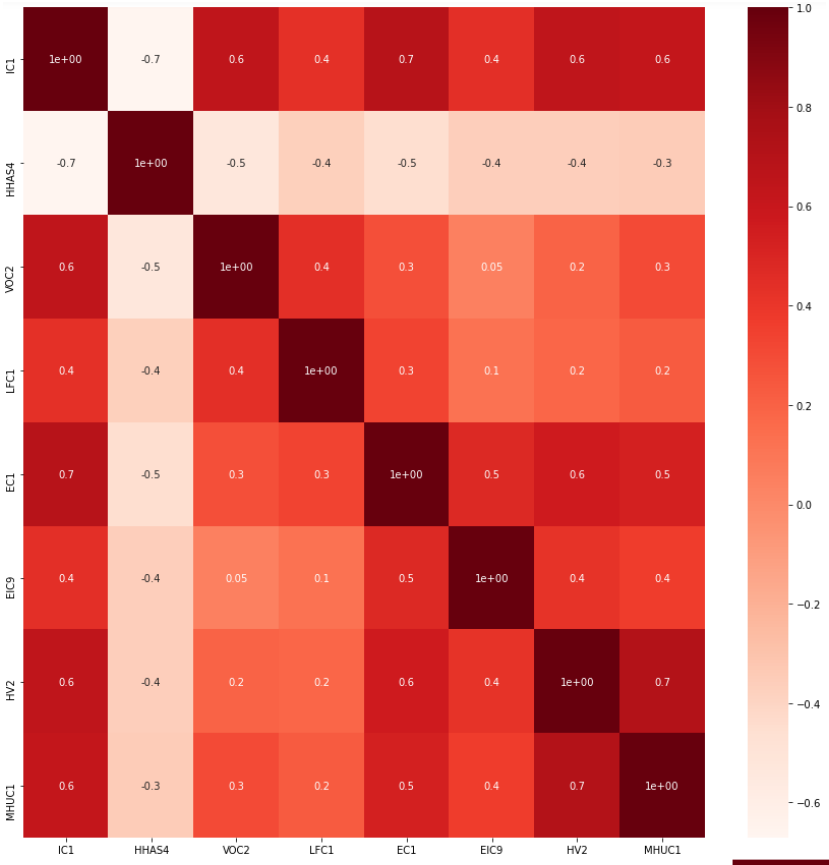

Figure 13 : Neighborhood correlation matrix

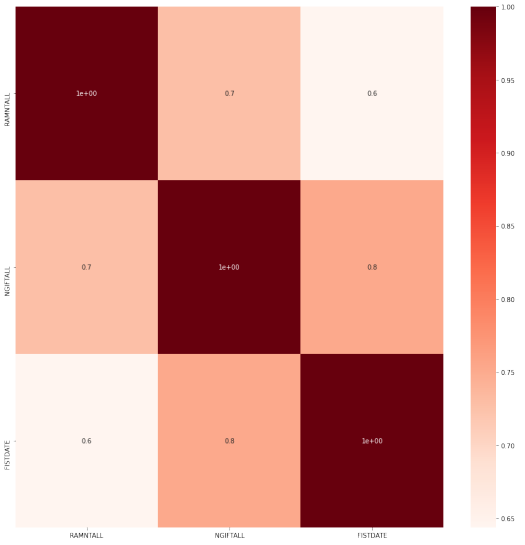

Figure 12 : Gift correlation matrix (Pearson's method)

After checking both correlations matrix, we think that our features have a good correlation between each other.

There are no features with very high correlation, that would give our clusters collinearity problem. We set a maximum threshold of +/-0.8 for the max and +/-0.4 for the minimum

DATA NORMALIZATION

Before any kind of modeling, we must make our data standardized due to the different feature's scales. There are different tools to do that, Sklearn.preprocess library provides us a lot of them. We tried many different scales like: “Normalizer()”, “RobustScaler()”, “StandardScaler()”, but we decided to use the “MinMaxScaler”.

CLUSTERING

For the clustering we decided to go with K-means, to support our decisions on Hierarchical Clustering, the silhouettes and the inertia K-elbow plot.

PROCESS

1st - Calculate the inertia by running k-means in a range from 2 to 10, so we could use the elbow method with the support of the function called “KneeLocator”, this function returns the biggest slope change in the line and returns the suggested number of clusters.

2nd - Run k-means with the suggested number of clusters from the inertia plot and use Hierarchical to verify the number of clusters and check our solution and if we can reduce the number of clusters.

3rd - We analyze the silhouettes score, this is another tool that will help us understand how well samples are clustered, meaning, the quality of the cluster.

4th - Run k-means with the number of clusters decided after the analysis.

GIFT SOLUTION CLUSTERING

1st Step:

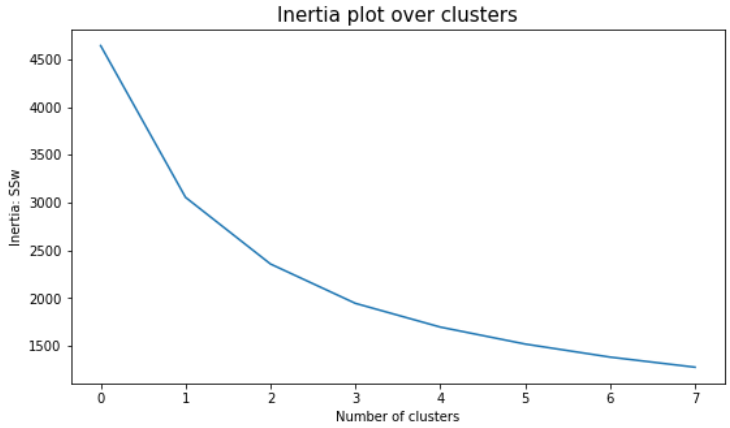

Figure 14 : Inertia plot to determine our number of gift clusters

Looking at the inertia line we could decide from 2 to 5 clusters, the KneeLocator function returns 4 for the highest slope change in the line, so we are going with 4 for the next steps.

2nd Step:

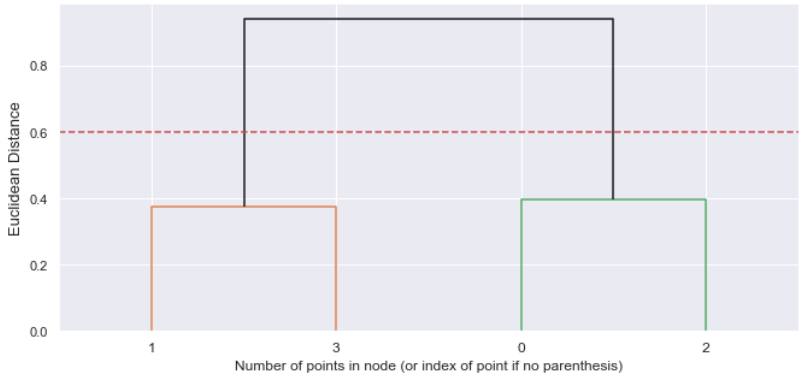

Figure 15 : Hierarchical gift clustering

After checking out cluster solution on hierarchical clustering, we decided that we should take 2 clusters.

3rd Step:

Below we have the scores for the ranges from 2 to the 4:

For n_clusters = 2, the average silhouette_score is: 0.5580689200492257

For n_clusters = 3, the average silhouette_score is: 0.47728482311103426

For n_clusters = 4, the average silhouette_score is: 0.47149764959644763

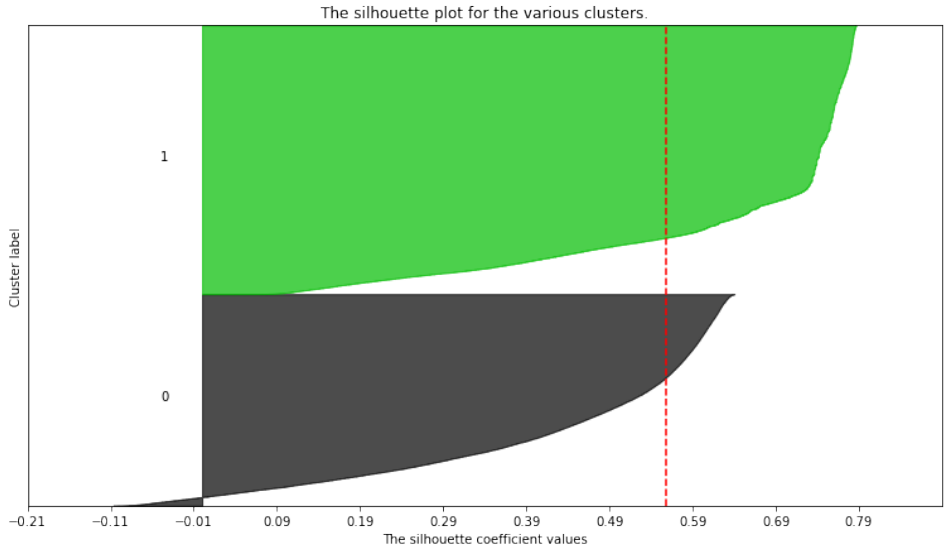

Figure 16 : Silhouette of the gift clustering

The best score was for 2 clusters, which confirms the previous results. Our final solution will be done with 2 clusters.

NEIGHBORHOOD SOLUTION CLUSTERING

1st Step:

Figure 17 : Inertia plot to determine our number of neighborhood clusters

Looking at the inertia line we could decide from 2 to 6 clusters, the KneeLocator function returns 5 for the highest slope change in the line, so we are going with 5 for the next steps.

2nd Step:

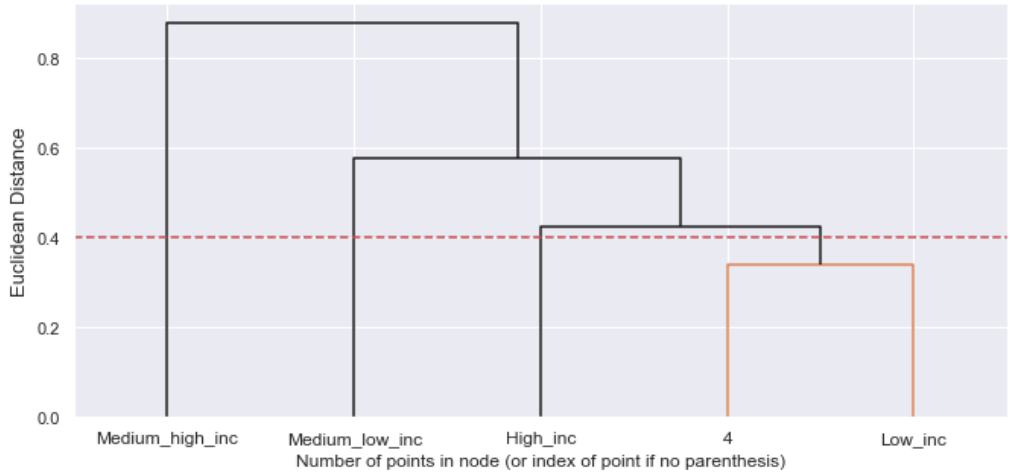

Figure 18 : Hierarchical neighborhood clustering

After checking out cluster solution on hierarchical clustering, we decided that we should take 4 clusters.

3rd Step:

Below we have the scores for the ranges from 2 to the 4:

For n_clusters = 2, the average silhouette_score is: 0.31963065329058615

For n_clusters = 3, the average silhouette_score is: 0.21966210727477212

For n_clusters = 4, the average silhouette_score is: 0.19866515651054895

For n_clusters = 5, the average silhouette_score is: 0.20781065165276816

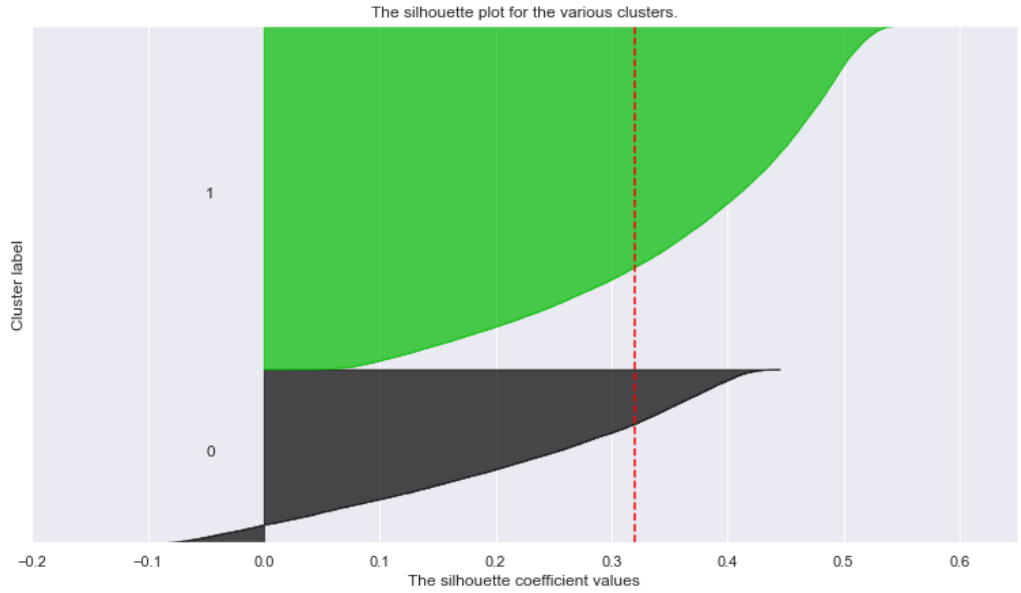

Figure 19 : Silhouette for 2 neighborhood clusters

Although the highest score in this case is 2, by looking at this solution we can conclude that cluster 1 has a lot more samples that cluster 0.

If we look at the silhouette for the 4-cluster solution we still have a big group, but more data split in the other clusters. This might not be the best cluster solution due to those negative values, but we rather go with this better divided cluster solution.

So, even so the 4 clusters solution is the one with lowest score and a not so good silhouette, we decided to rely on the hierarchical clustering and the inertia plot for our final decision.

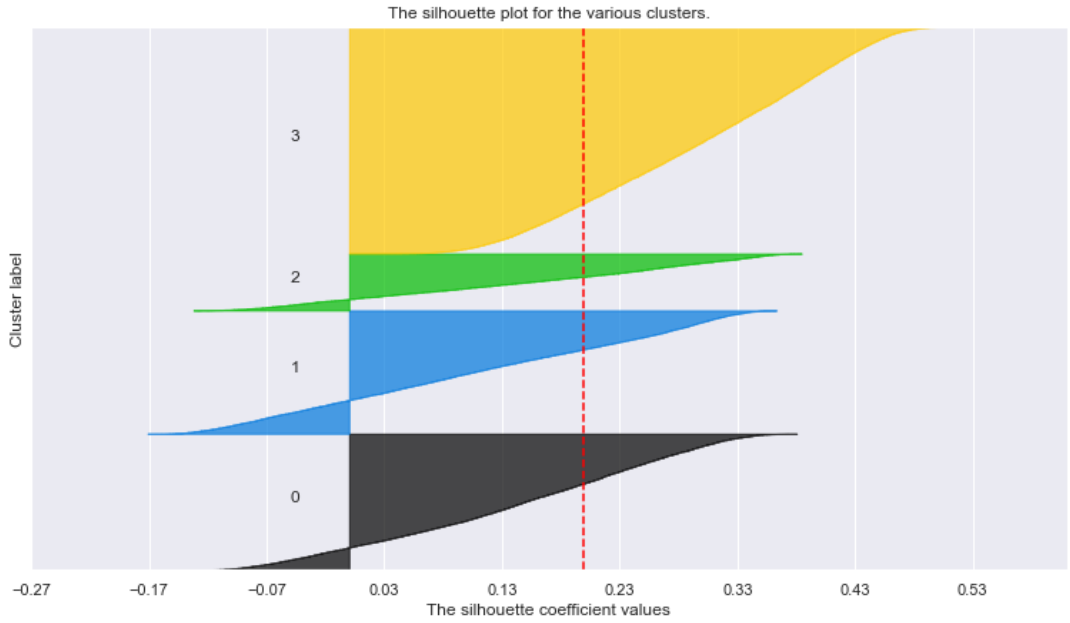

Figure 20 : Silhouette for 4 neighborhood clusters

GIFT DONORS SEGMENTATION

Cluster_0 - Older donors (blue): The principal characteristic of this cluster is that they are long time donators of PVA's. Their amount of donation is proportional with the number of times they gave. Between the two clusters, cluster_0 size is the smallest with around less than 40 000 donors.

Cluster_1 - Recent donors (orange): On the contrary to cluster_0, it concerns donors that started donated a short time ago. The number of times they donated is positively correlated with the total amount of donation. It includes more than half of the total donors: 50 000.

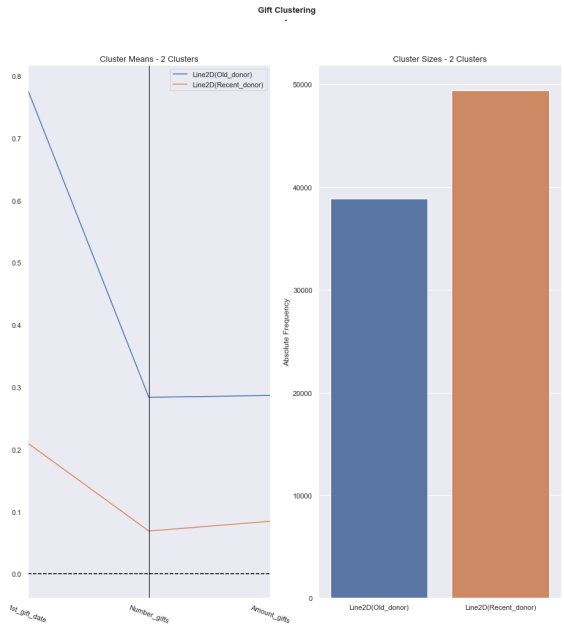

Figure 21 : Gift clustering

NEIGHBOR SEGMENTATION

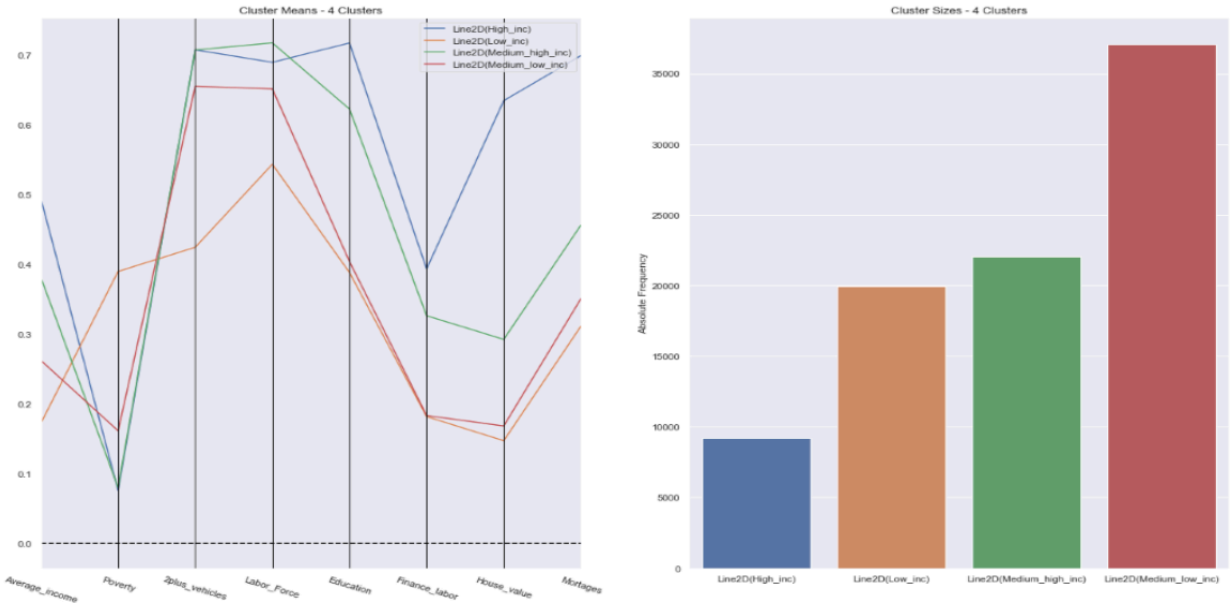

Figure 22 : Neighborhood clustering

Cluster_0 - High Income: The principal characteristic of this cluster is the high median households' income of the neighborhood. We can see that almost 3/4 of the donors own more than 2 cars, and that the average home value is significantly high, it infers that donors from this cluster are from the most wealthy and fancy neighborhoods. Roughly half of the population is employed in either finance, real estate or Insurance sectors, we can interpret it as neighborhood from more urbanized cities. They also have the most years of education. This cluster is the smallest from the 4 with a size of less than 10 000 donors.

Cluster_1 - Low Income: This cluster represents the low-income neighborhoods; it is showed with the median households' income but also the high percentage of individuals below poverty level. Compare to the other 3 clusters, the most important difference is average of households in neighborhood having at least 2 cars. At least 3/5 of the population in all clusters, except cluster_1 have at least 2 cars, against only a little more than 2/5 for cluster_1. They are also less working neighborhood, with lower education level. With a lower average of home value, it makes them the less wealthy neighborhoods. Comparing to the high-income cluster, the proportionality between average home value and the median homeowner cost with mortgage, is higher regarding the low-income cluster. Their size is around 20 000 donors.

Cluster_2 - Medium high income: This cluster follows mainly the pattern of the high-income cluster, it has a high median income household, a low percentage of individuals below poverty level and as cluster_0 3/4 of the neighborhood have at least 2 cars. However, they are the most working neighborhoods even though their years of education is a little bit lower. The difference between the average home value of cluster_1 and cluster_2 is significant which shows that it represents less fancy neighborhoods.

Cluster_3 - Medium low income: It is the bigger cluster, with more than 35 000 donors. It has a medium low median households' income, but represents also working neighborhoods, with more than 3/5 of their population owning at least 2 cars. It shows a hard-working cluster from less fancy neighborhoods. This segment is a bit similar to the low-Income cluster_1 regarding the home value, and education of the neighborhoods.

CLUSTER MERGING

After using hierarchical clustering to check our clustering solution, we decided to go with 3 clusters.

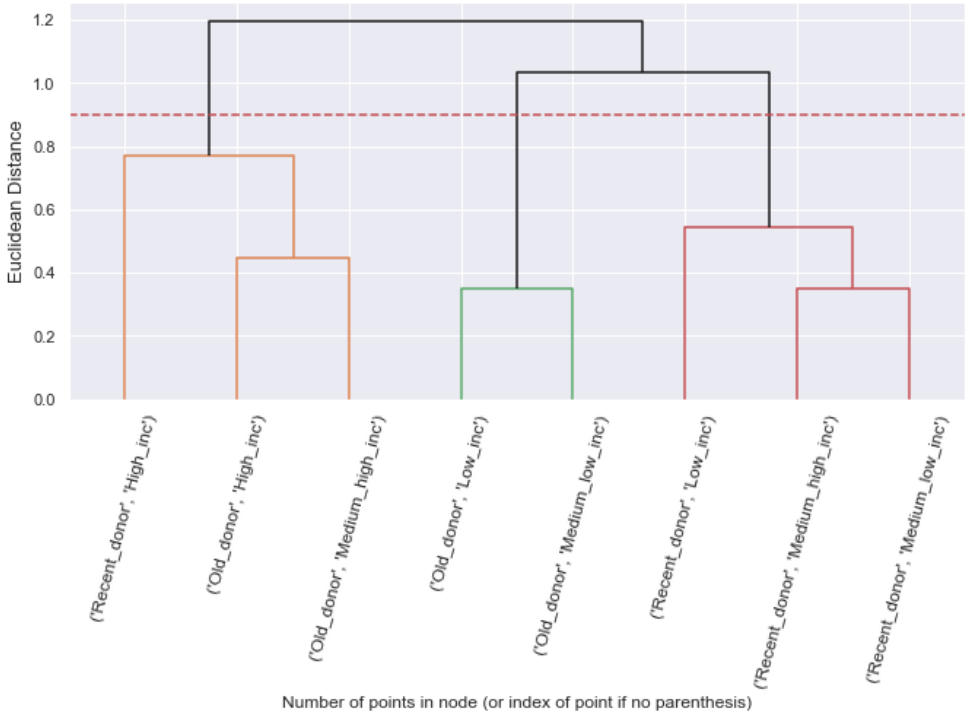

Figure 23 : Hierarchical clustering of our merge clusters

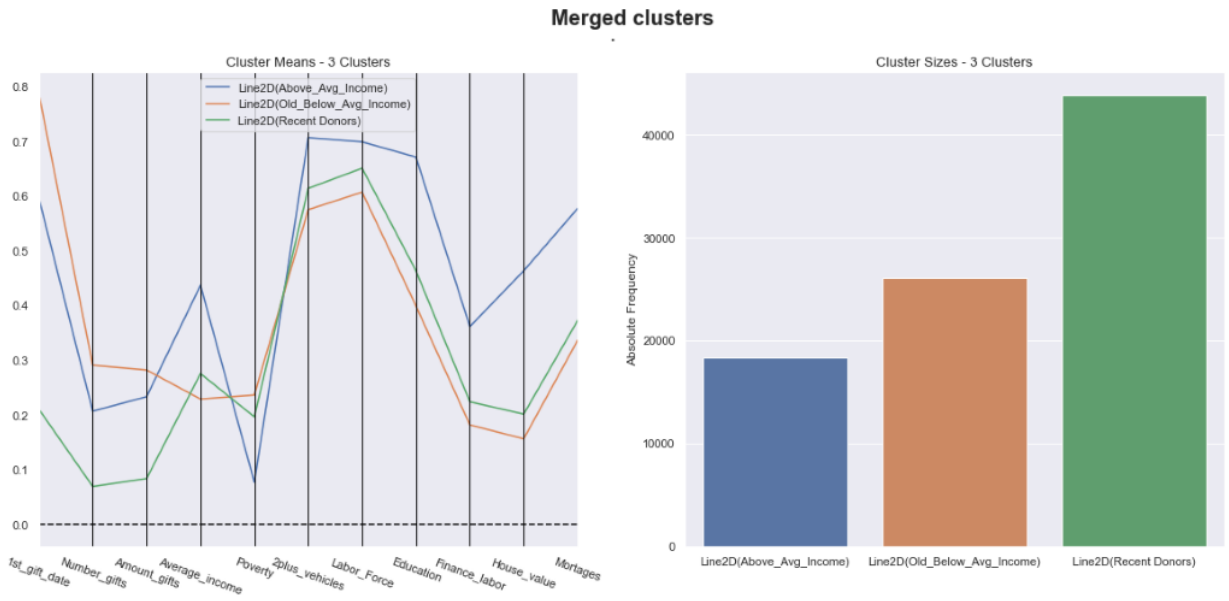

Figure 24 : Merge of our clusters

Cluster_Blue - Donors from High Household Income Neighborhoods: It is our smaller cluster with less than 20 000 donors, it represents neighborhoods high income and low percent of people living below poverty level. They represent the wealthy and fancy neighborhoods, with the most cars, the expensive homes, and the higher cost of mortgage. They are from the most educated area, and 40% are employed in finance, real estate or insurance. As stated before, it can be inferred that those neighborhoods are from more urbanized cities. They appeared to give less times but more important value gift.

Cluster_Orange - Oldest donors from Low Household Income Neighborhoods: This cluster shows that the oldest and most loyal donators of PVA are mainly from low-income neighborhood with the highest percentage of individual living below the poverty level.

Cluster_Green - Most recent donors from Medium Income Neighborhoods: This is the bigger cluster with approximatively 43 000 donors. PVA's most recent donors is following the patterns of our oldest donors, meaning they are from medium low income working neighborhoods with a lower education level, and less wealthy with a low average home value.

MARKETING STRATEGY - TARGETING

Even though lapsed donors are donors that stopped giving, considering that they at least donated once, they are most likely to give again, usually regained lapsed donors give higher amount of gift than newcomers. From the result of the clusters, we suggest 3 possible marketing strategy in order to recapture the interest of our lapsed donors. We are relying on our merge clusters, of the donors and neighborhood segmentation. From the donor's segmentation, we saw that it separated into long term and recent

donors, depending on the date of their first gift. From the first and second cluster, we are dealing with mostly long-term donators.

Firstly, the **Cluster_Orange Oldest Donors from Low Income Neighborhood** is the loyal group, they have been donating multiple times for a long time. Even though they have mortgages on their houses, and are not part of the wealthiest neighborhood, they are donating to help veterans. We assume that they were giving more frequently in the past but not a high amount of donations, and then they stopped donating. One of the goals should be to recapture their interest into this cause for veterans, to regained frequent donors. Since they have been long term donors, they will most likely come back with the right approach. We suggest that the next promotion should be customized with names, and it should re-engage them to the cause, such as a meaningful story of a veterans that needs help, or to show them how grateful the foundation is of their support.

Secondly, we have the **Cluster_Blue Donors from High Income Neighborhoods**, on the figure 24 we can see that the total amount of donation is positively correlated with the number of times they gave. This is our smaller cluster — around 19 000 donors — but from our analysis of the features, we realize that donors might donate less times, but donate a more important amount than cluster_orange. We suggest a more VIP approach concerning this cluster and concentrate on lapsed donor that gave a high amount of gift. In order to reactivate the VIP lapsed donor, a special strategy should be put in place. Donors will most likely donate if they feel like they are needed, that their donations are meaningful and is helping directly veterans. Communication can be time consuming and expensive, but it is the more effective way to regain the interest of a donors. From this 19 000 donors cluster, the donors that gave the higher amount and the numbers of time could be find, and the top — number of PVA's choosing — of those could be phoned. Reminding the cause of the non-profit to people is important, it shows that it isn't just about giving money but helping veterans to fight injuries and diseases. Relation between donors and veterans could be made. Emails, promotions any kind of communication with VIP donors should be personalized. The goal would be to make them frequent donors as cluster_orange, but with a higher amount of money per gift.

Finally, our third clusters showed that an important part of our donators made their first donation recently. **The Cluster_Green Most Recent Donors from Medium Neighborhoods** follows the same pattern as cluster_0, the donors come from the same kind of neighborhood. Considering that they have been donators for a shorter amount of time and made fewer gifts, their interest might be harder to recapture. Since it is composed with more than 40 000 of our lapsed donors, the approach of recapturing them should be done through mass advertising, meaning in addition to customized promotions, this cluster should also be targeted with advertisement through Facebook or Google.

The marketing strategy assessed was made with the decision with put in the feature selection of the dataset given. 476 different features are composing the data, meaning a lot of different approach can be made depending on the variables chosen.

CONCLUSION

In this project, our purpose was to understand the behavior of PVA's donors, and to present a possible strategy to reactivate their lapsed donors. We decided to focus on the economical aspect of the neighborhoods' donors. After analyzing the dataset, and the features, we reduced our data into two categories, 11 features about neighborhood and gift's donors. Before targeting clusters, we cleaned the data of missing values, dealt with the outliers, and changed the type of values when necessary. We then proceeded in normalizing our data using the MinMaxScaler.

We decided to cluster the two groups separately to understand their behavior, and then progressed by merging the results.

Finally, after the cluster where made, it showed us that donors are mainly from two types of neighborhood, and that their behaviors of donation matched with the types of area they are from. People from high-income areas are most likely to give bigger donations a less amount of time, against people from medium-low-income areas that tend on giving a higher amount of time but with lower amounts. Out of the 3 clusters, 3 marketing strategies were proposed in order to recapture the interest of their lapsed donors.

You can find our project on our Github Repository : https://github.com/Behemot6/DM-project-team-super-cool.git

REFERENCES

https://www.theshelbyreport.com : logo cover page PVA

A. (2020, December 16). *Detecting And Treating Outliers In Python — Part 1* - Towards Data Science. Medium. https://towardsdatascience.com/detecting-and-treating-outliers-in-python-part-1-4ece5098b755

Mahto, P. (2020, September 21). Local Outlier Factor: A way to Detect Outliers - MLpoint. Medium. https://medium.com/mlpoint/local-outlier-factor-a-way-to-detect-outliers-dde335d77e1a

Kumar, A. (2020, September 17). Kmeans Silhouette Score Explained With Python Example. dzone.com. https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-examm#:%7E:text=Silhouette%20score%20is%20used%20to,each%20sample%20of%20different%20clusters

kneed. (2020, August 13). PyPI. https://pypi.org/project/kneed/

Alam, M. (2020, November 20). Isolation Forest: A Tree-based Algorithm for Anomaly Detection. Medium. https://towardsdatascience.com/isolation-forest-a-tree-based-algorithm-for-anomaly-detection-4a1669f9b782

Rees, S. (2020, January 21). How to Renew Lapsed Donors. Get Fully Funded. https://getfullyfunded.com/how-to-renew-lapsed-donors/