Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Insurance Company

Group Members - AR

António Fonseca number: r20181154

João Carvalho, number: r20181122

01, 2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1 INTRODUCTION

An Insurance Company based in Portugal that provides insurance plans in 5 major categories: Motor, Household, Health, Life and Work Compensation pretends to better understand all their different Customers' Profiles.

Our group was asked to develop a Customer Segmentation Model over a dataset of 10296 customers. The model is intended to generate segments accordingly to the Premium Features and Customers Features. Therefore, the objective is to develop a model that identifies which segments of clients should this insurance company focus more.

2 DATA EXPLORATION

The first step before starting Data Preprocessing is to analyse the data. Regarding EducDeg, we can see that half of the population has at least a bachelor's degree and by analysing children we understand that the majority of the sample has kids (appendix figure 1). Regarding the metric features we can conclude that the insurance plans more valuable to the company are Motor, Household and Health since the amount paid by the customers is higher in these three categories (appendix table 1).

3 DATA PRE-PROCESSING

3.1 COHERENCE CHECKING

As for the coherence checking, we first started by removing observations where FirstPolYear (Year of the customer´s first policy) was bigger than the current year of the database: 2016. One observation was removed.

After analysing the variable BirthYear, we firstly started by removing one record where the customer's age was higher than 100. Then, we found that there were 1997 observations where the customer's birthday year was higher than its first policy. Since there was such a big number of incoherencies, we decided to remove one of the variables. Considering that the variable FirstPolYear is automatically filled by the computer when a client performs his first policy and the variable BirthYear is manually filled by the customer when creating an account, it seemed more reasonable to us to drop BirthYear since it is more prone to errors. We have also checked whether there was no Teenagers with children and people under sixteen years old earning a salary.

3.2 OUTLIERS

We decided to treat our data for outliers using four different methods for univariate and multi variate outliers:

3.2.1 Mannual Approach

In the manual approach we tried to identify outliers observing both the histograms and box plots of the numerical variables in the data. We looked for extreme values in the plots and set a threshold for each variable such that every observation that exceeds it is considered an outlier. Outliers identified: 11,44%.

3.2.2 Standard Deviation method

The standard deviation method identifies observation which values, once standardized, are too distant from the mean of that variable. And "too distant" means that they are more than a certain threshold of standard deviations away from the mean. Outliers identified: 3,88%.

3.2.3 K-means for outlier detection

We used k-means for outlier detection in order to identify multi variate outliers. Once the clusters were arranged by the algorithm, the observations identified as outliers are the ones that are further away from their centroid. Outliers identified: 4,88%..

3.2.4 Isolation Forest

Isolation forest is a decision tree-based method for outliers detection. The idea is that outliers should have some extreme attributes and should be infrequent in the data and then, regarding the nature of a decision tree, they should be identified in the root of the tree. Outliers identified: 4,88%.

3.2.5 Final Filter

In the end, we decided to remove from our analysis only the observations that were identified as outliers by the four methods. We concluded that this is a safe approach as we don´t want to remove too many observations from the data but we also do not want abnormal observation harming our analysis. Outliers removed: 1,2%.

3.3 MISSING VALUES

There are considerable amounts of missing values present in our data (appendix table 2). However, none of the variables has lots of missing values to be excluded from the analysis since none of them exceeds the 20% threshold. Before imputing the final values, we used a KNN imputer to impute auxiliar values in all the numerical features´ missing values since, as explained further, we used regressions to estimate the missing values and we could not impute a missing value to a observation´s feature that had a missing value on other feature.

3.3.1 Categorical Features imputation

The are two categorical features that have missing values: EducDeg and Children.

EducDeg only has two missing values and so we concluded that a simple imputation method would be enough. For that reason, we used a central tendency method: the mode. Children had thirteen missing values and, since it is a binary feature, we decided to impute them using a logistic regression using all the numerical features in the dataset for the estimation of the probability of the value of Children is 1 (or 0).

3.3.2 Numerical Features imputation

The variables with the most missing values belonged to the numerical features. These were: MonthSal, PremMotor, PremHealth, PremLife and PremWork.

We decided to impute these missing values using a OLS regression to estimate them. For each one of the variables, the regression was estimated using all the others numerical features and the variable Children (which is binary, therefore can be included in the estimation).

3.4 FEATURE ENGINEERING

3.4.1 Changing Data Types

Regarding the original variables we decided to change the data type in 4 of them:

- FirstPolYear to integer;
- Children to Boolean;
- EducDeg to string;
- GeoLiveArea to string.

3.4.2 Encode Categorical Variables

As for the encoding of categorical we decided to use One Hot Encoder which represents categorical variables as binary vectors. The non-metric features encoded were EducDeg and GeoLiveArea.

3.4.3 Creation of other Variables

The new variables created were:

- **YearsAsCustomer:** Number of years since the clients' first policy;
- **TotalPremium:** Sum of all the premium features;
- **SalarySpentRatio:** Ratio between the MonthSal and the TotalPremium;
- **Prem Ratios:** Ratios between each premium feature (PremMotor, PremHealth, PremLife, PremHousehold and PremWork) and the TotalPremium.

We decided to create these variables since it is known that it is a good practice to create Ratio Variables in order to reduce noise and amplify the signal.

3.5 FEATURE SELECTION

Firstly, we started by normalizing the data using the StandardScaler. Then, we removed again outliers using DBSCAN. Although DBSCAN is not an Outlier Detection Algorithm, we can use it as Outlier removal since when it tries to allocate the points to a cluster, the records that are considered noise points are given their own class, -1. Therefore, we chose to remove all noise points, 122 records. We concluded that another outlier analysis was needed since observing the component planes of a SOM implementation while clustering (which is explained further ahead in this paper), there were some units that really stood out from the others which is indicative of the presence of outliers (appendix figure 2). Adding this removal with the others performed in Incoherence Checking and Outliers, we removed 2.83% of the original dataset which is a small and healthy percentage of observations.

As for the proper Feature selection, we analysed the correlation between the variables and decided to keep, when correlated, the ratio features over the original ones (appendix figure 3). Therefore, the variables that were not used were: ClaimsRate, FirstPolYear, MonthSal, TotalPremium and all of the original premium features which were too correlated with at least one of the variables that were used in the analysis. We chose to split the variables into two groups of features (Premium Features and Customer Features) which they will then undergo Clustering separately.

This division´s purpose is to understand the different kind of customers from different points of view: the premium features represent the clients´ main interests and needs regarding insurance packages, while the customer features represent the social conditions of the clients. The final features that compose the two groups are:

Premium Features

- PremMotorRatio;
- PremHealthRatio;
- PremLifeRatio;
- PremHouseholdRatio;
- PremWorkRatio;

Custom Features

- YearsAsCustomer;
- SalarySpentRatio;

It is important to mention that the first time we chose the variables for the Premium Features we also included 'CustMonVal'. However, after assessing the feature importance of a cluster solution we identified that this variable was not discriminative at all (appendix table 4).

4 CLUSTERING

After the Pre-processing part, we started by clustering the data in two different perspectives as explained in the Feature Selection Part (Premium and Customer).

Nevertheless, the approach was the same for each one. We performed four different cluster algorithms: K-means, K-means on SOM units, Hierarchical clustering on SOM units and Mean Shift algorithm. In the end, in order to conclude which method was better we used performance measures such as R^2 and silhouette and analysed which cluster interpretation would best fit our data.

4.1 MODELING

4.1.1 K-means

For the k-means clustering, we decided to initialize the centroids with the default option in python 'kmeans++'. To decide on the number of clusters to be initialized we evaluated the silhouette and the inertia of the correspondent solutions (elbow graphic and silhouette visualization). If any doubt would arise on the number of clusters to be initialized, we would perform both options and evaluate the results in the end along with the other algorithms.

4.1.2 K-means on SOM units

For the k-means on SOM units clustering, we first performed SOM on the initial observations to rearrange them in a large number of clusters (SOM units). More specifically, 400 SOM units since the map size used was 20x20. We would then visualize the component planes, the U-map and the Hits map to get some insight on correlations between features, the data distribution, feature importance and even outliers.

Then the k-means clustering is applied on the SOM units using the same techniques to determine the number of clusters. Once again, if any doubt would arise on the number of clusters to be initialized we would perform both options and evaluate the results in the end along with the other algorithms.

4.1.3 Hierarchical Clustering on SOM units

Similar to the K-means on SOM units, we first started by performing SOM with the same specifications as before. Then, we plot the R2 for different clusters using different hierarchical techniques. To select the number of clusters we used the silhouette analysis.

4.1.4 Mean Shift

We decided that it was important to perform a density clustering algorithm, such as mean shift, on our data. Regarding the implementation of the algorithm: the main initial parameter to be estimated is the bandwidth and we chose one that would give us an interpretable and desired number of clusters. Nevertheless, we used the R^2 to visualize the impact of different bandwidth values in the algorithm´s performances. In the end, the mean shift algorithm´s performance measures were poor when compared with the other algorithms, even for different bandwidth values.

4.2 PREMIUM FEATURES SOLUTION

After performing the 4 algorithms mentioned for the premium features, we concluded that k-means on SOM units was the best algorithm to segment the data. We decided that a 3 clusters solution was the best fit for our dataset in this intermediate step. We then proceed to analyse the clusters' profiling (appendix figure 4) as well as the distribution of the categorical variables (appendix figure 5).

The three clusters have similar sizes, and each one´s description follows:

Cluster 0 – Workers: These cluster is formed by the customers who pay significantly more premiums regarding Work, Life and Household insurance plans and below average in Motors plans. They are not so educated as the other two clusters, the majority has only the high school diploma.

Cluster 1 – Health-centric people: It is formed by clients whose main insurance related expenditures are in health. There other expenditures are average. Regarding the categorical variables they are average in both aspects (Education and Children).

Cluster 2 – Parents: these customers are significantly above the average regarding what they pay in Motor plans. Regarding the other plans, they score far below average. These are the clients who are more educated and are more likely to have children.

This cluster profiling is only to be used as guideline for the final merged clustering technique.

4.3 CUSTOM FEAT SOLUTION

After performing the 4 algorithms mentioned for the custom features, we concluded that k-means on SOM units was the best algorithm to segment the data. We decided that a 4 clusters solution was the best fit for our dataset in this intermediate step. We then proceed to analyse the clusters' profiling (appendix figure 6) as well as the distribution of the categorical variables (appendix figure 7).

The three clusters have similar sizes, and each one´s description follows:

Cluster 0 – Promising New Clients: These cluster is formed by the customers who are below average in both features ('SalarySpentRatio' and 'YearAsCustomer'). They are considered promising since they spend a relatively low fraction of their salary, therefore they might be willing to spend more money on premiums.

Cluster 1 –Older Clients: These cluster is formed by the customers who are above average in both features ('SalarySpentRatio' and 'YearAsCustomer').

Cluster 2 – Promising Older Clients: these customers are below average in the salary spent ratio and above average in the years as customer. They are considered promising since they spend a relatively low fraction of their salary, therefore they might be willing to spend more money on premiums.

Cluster 3 –New Clients: It is formed by clients who are above average in the salary spent ratio and below average the number of years they have been customers.

This cluster profiling is only to be used as guideline for the final merged clustering technique.

4.4 FINAL SOLUTION

Finally, to deliver the final cluster solution, the two previous analysis must combine. Every customer is now labelled for his premium and customer characteristics, which would lead us to 12 final clusters. Since it is too much, we decided to merge them using a hierarchical clustering technique (appendix figure 8). We then proceed to analyse the clusters' profiling (appendix figure 9) as well as the distribution of the categorical variables (appendix figure 10).

Cluster 0 - Workers

It´s characterized by the clients who pay the most life, work, and household related premiums. Note that the PremHousehold feature is highly positively correlated with Total Premium, which means that these are the customers who pay the most premiums. Regarding the Health plans they are below average, and they have also stood out from the other groups since the clients pay significantly below average premiums regarding Motor insurance plans. They also spend, compared to the other clients, a relatively high fraction of their salary in the company which is expected since their monthly salaries are the lowest. Another important aspect to refer is that among all clusters, they have the lowest levels of education.

Marketing Approach

These clients seem to value the company´s insurance plans since they are the biggest spenders in some of the features. We recommend that, in order to attract these clients to the motor and health plans, marketing campaigns should be developed.

As we concluded, these customers seem to have a lower purchase power and, therefore, it could be a good strategy to develop packages that give discounts to the clients. For example, create a specific package where these clients by buying a motor or health insurance plan receive a discount in one of other plans that they already have (5-10% as an example). Other possibility is to develop a package where they buy both plans (Motor and Health) and receive a discount in the first months.

Although, these customers appear to be less valuable to the company, Loyalty plans should be developed in order to maintain these customers.

Cluster 1 – Parents

It´s characterized by the clients who pay the most Motor related premiums. Regarding the other insurance plans, they score below average in all of them, including in the household category which is extremely correlated with the feature total Premium. This means that they are, in fact, paying relatively low total values of premiums. However, their salary spent ratio is the lowest relatively to the others. It is important to notice that these customers have a monthly income above average. As expected, they are the most educated clients. One important thing to note is that they have a children ratio much higher than the other two clusters.

Marketing Approach

These customers seem to value less non-mandatory insurance plans (Motor plans are mandatory) even though they appear to have the money to spend and, therefore, the company should develop marketing strategies to engage and attract with them more efficiently.

One possible approach for this cluster is to promote family health insurance plans for them. It would be a good idea if the insurance company promoted a plan where all the family would get a health insurance plan, since they appear to have more children than the other two clusters and it is very expensive to afford health insurances for every family member.

Regarding the other 3 plans, Life and Work tend to be cheaper than the Household. Furthermore, these customers seem to have a more conservative lifestyle and, therefore, the company should focus on marketing approaches that include household plans. One possible idea would be to offer a discount (5%) on their Motor insurance plans (which they pay by far the greater values) if they were to purchase a household insurance.

Cluster 2 – Health Centric Customers

It´s characterized by the clients who pay the most Health related premiums. Regarding the other insurance plans, they score average in all of them. One important thing to note is that their salary spent ratio is slightly above average.

Marketing Approach

Even though these customers score average in all of the insurance plans apart from the health one, these are the clients who earn the most and their salary spent ratio is higher than the usual, which is expected.

One possible marketing approach is to promote Life plans by applying a monthly discount in the first year (5%) since these are the customers who worry the most with health issues. Another approach would be to offer an increase ceiling in their health expenditures in exchange for a purchase of a Motor plan as they are the cluster who spends the least on this category which is the most profitable one for the insurance company.

4.5 CLUSTER VISUALIZATION

To better visualize the clusters, we performed a UMAP visualization on the final solution (appendix figure 11), where we can clearly see the boundaries for the three clusters which is a good indicator that the clusters are well formed and heterogeneous between each other.

4.6 OUTLIERS' LABELS

In the end of the cluster analysis, we decided it would be important to predict the outliers´ labels. They were excluded from the analysis since its extreme values would create a bias but they should be assigned to a cluster so that they can also be target of the most appropriate marketing campaigns. We decided to use a decision tree classifier to predict their labels which was trained using the already labeled data.

5 LIMITATIONS

As for limitations on the project we consider that the lack of variables, variables´ descriptions and reliability of some of them were a considerable difficulty that we had to overcome. One of the variables that was not reliable and, therefore, we had to remove it was BirthYear which had 1997 records that did not make sense. This attribute would be important to segment the dataset on the customer characteristics.

Moreover, to produce a more concrete and valuable marketing campaign it would have been important to know more about the company, more specifically, about the current campaigns and offers available for the clients as we could have been more easily able to identify strengths and weaknesses of every campaign and their suitability for the different clusters. As an example, if we know what the current most successful or unsuccessful packages are and how the customers react to them, we could be more precise on which variables or kind of customers to focus on.

6 CONCLUSIONS

In these project we were asked to develop a cluster analysis on the customers of an insurance company. Initially we explored and pre-processed the data with the objective of reducing as much as possible the noise and amplify the signal, several techniques were used in this part regarding Coherence Checking, Outliers Detection, Missing Values and Feature Engineering and Selection.

Afterwards, we decided to perform the cluster analysis from two different perspectives: Premium and Customer features, in order to clarify how the clusters are formed. Numerous clustering algorithms were developed and applied. Subsequently we merged both perspectives with the goal of obtaining a valuable and feasible solution.

The final output of the project are detailed marketing campaigns developed for each cluster with the objective of maximizing the company's profits.

7 REFERENCES

Outlier Detection with Isolation Forest

Available at: https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Retrieved on 20/12/2021

Mean Shift

Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

Retrieved on 14/12/2021

DBSCAN

Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Retrieved on 29/12/2021

DBSCAN

Available at: https://www.novaims.unl.pt/fbacao/papers/the%20self-organizing%20ma%20the%20geo-som.pdf

Retrieved on 29/12/2021

8 APPENDIX (DOESN'T COUNT FOR THE 10PAGE LIMIT)

:

	count	mean	std	min	25%	50%	75%	max
MonthSal	10093.0	2518.731398	1106.884344	333.00	1729.000	2516.00	3297.000	55215.00
CustMonVal	10127.0	198.052106	1667.703029	-165680.42	-9.165	185.82	397.415	1448.28
ClaimsRate	10127.0	0.707176	2.559871	0.00	0.390	0.72	0.980	256.20
PremMotor	10093.0	299.662798	139.269244	1.78	195.260	301.28	408.520	3106.62
PremHousehold	10127.0	199.892678	223.676536	-75.00	48.900	131.70	284.500	4130.70
PremHealth	10084.0	168.671378	74.070055	-2.11	112.800	163.81	219.930	442.86
PremLife	10023.0	40.895695	45.820357	-7.00	9.890	25.45	56.900	398.30
PremWork	10042.0	40.171713	45.306080	-12.00	10.670	25.56	55.900	451.53
FirstPolYear	10127.0	1986.028538	6.603995	1974.00	1980.000	1986.00	1992.000	1998.00

Table 1 -Descriptive Statistics of the original variables

	Frequency	Frequency %
FirstPolYear	0	0.000000
EducDeg	2	0.019512
MonthSal	34	0.331707
GeoLivArea	0	0.000000
Children	13	0.126829
CustMonVal	0	0.000000
ClaimsRate	0	0.000000
PremMotor	34	0.331707
PremHousehold	0	0.000000
PremHealth	43	0.419512
PremLife	104	1.014634
PremWork	85	0.829268

Table 2- Missing Values

	Method	Clusters	R*2	Shilhouette
0	K-means	3	0.572	0.312
1	K-means	4	0.694	0.317
2	K-means on SOM Units	3	0.555	0.342
3	K-means on SOM Units	4	0.619	0.33
4	K-means on SOM Units	5	0.737	0.354
5	Hierarchical on SOM Units	3	0.542	0.322
6	Mean Shift	3	0.265	0.308

Table 3- Results Premium Features

	R2 Feature importance
PremMotorRatio	0.723654
PremHealthRatio	0.537581
PremLifeRatio	0.239249
PremWorkRatio	0.210213
PremHouseholdRatio	0.626950
CustMonVal	0.042784

Table 4 - CustMonVal feature importance

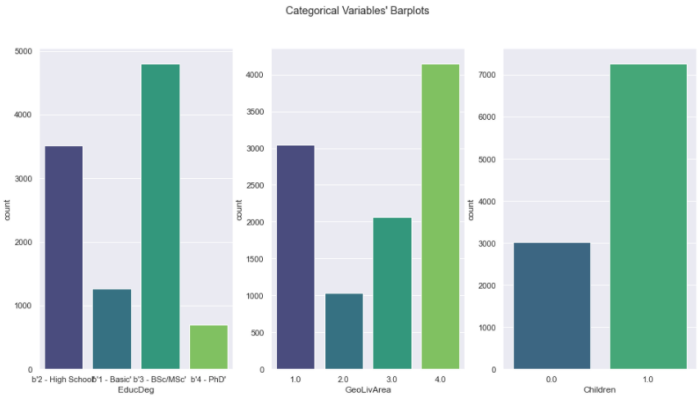

Figure 1- Non-Metric Variables

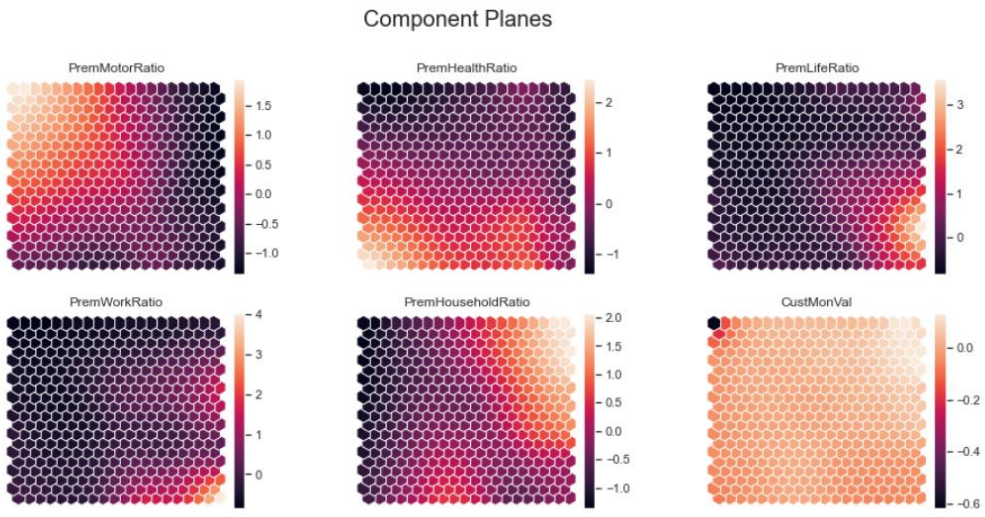

Figure 2 - Component Planes

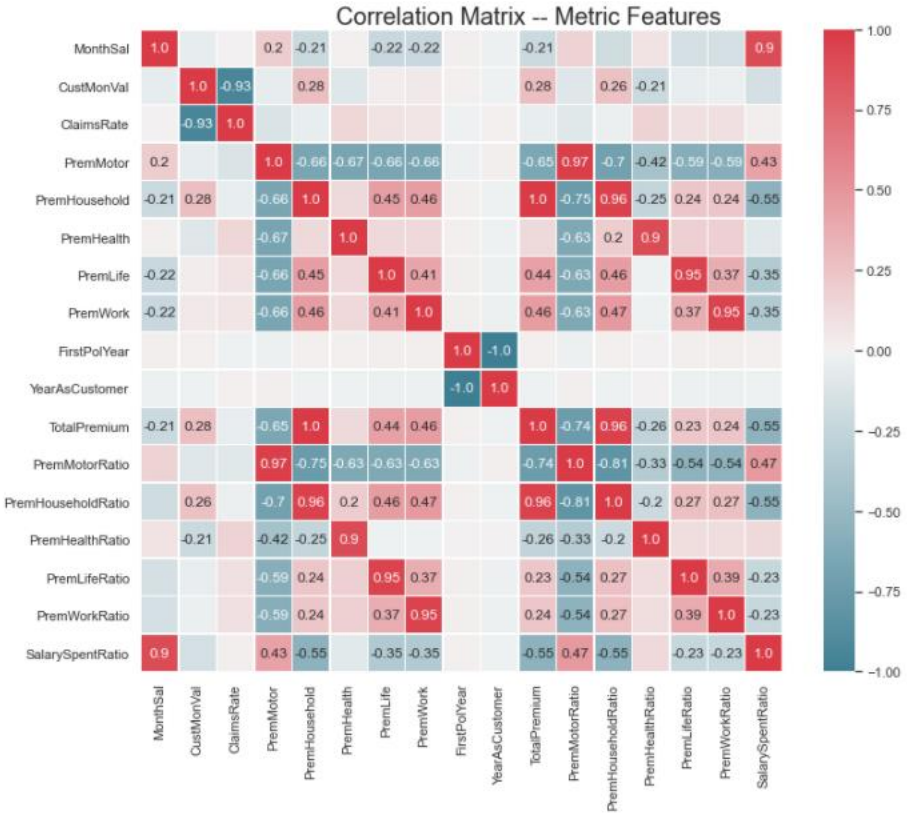

Figure 3- Correlation Matrix

label	PremMotorRatio	PremHouseholdRatio	PremHealthRatio	PremLifeRatio	PremWorkRatio
0	-0.988019	1.135179	-0.330707	0.600833	0.473172
1	-0.213116	-0.299174	0.990824	-0.014903	0.078393
2	1.071807	-0.692353	-0.597112	-0.596600	-0.588847

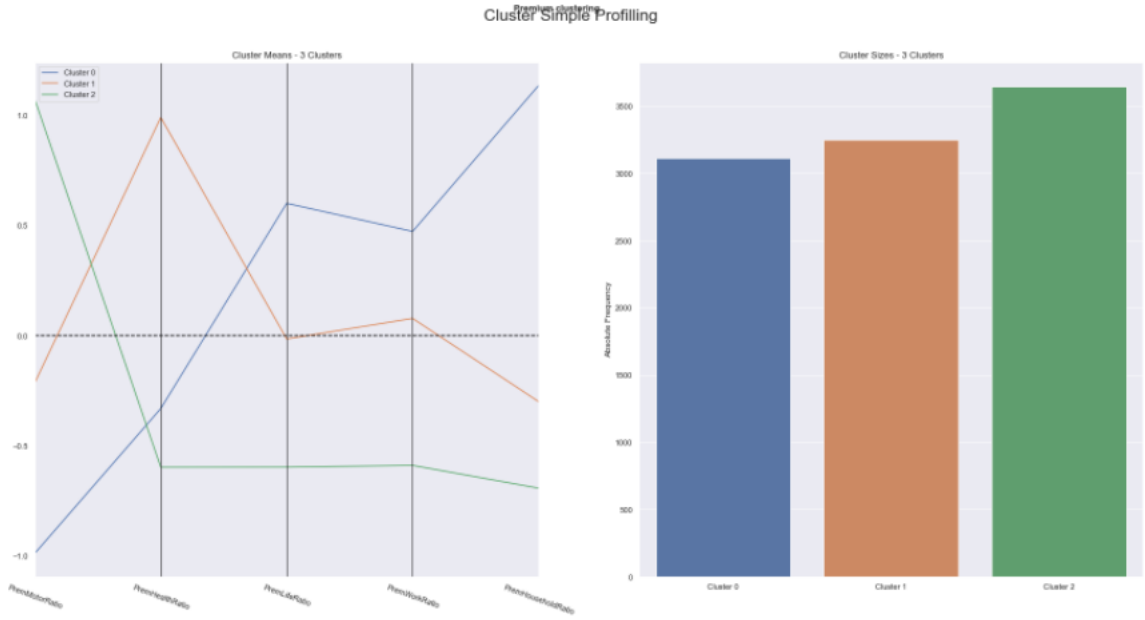

Figure 4- Clusters' Profilling Premium Features

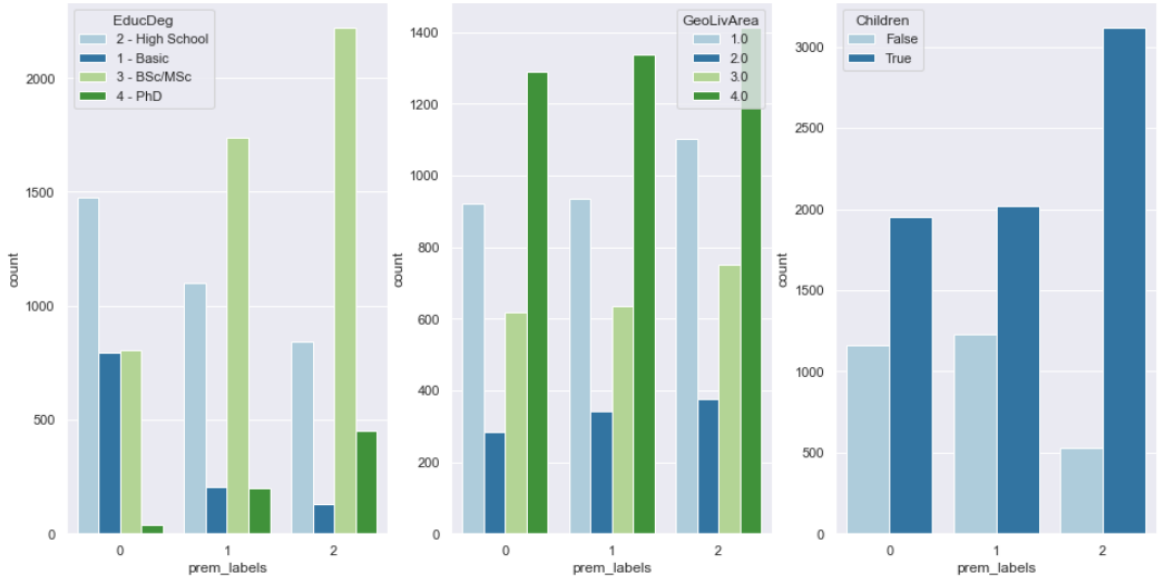

Figure 5- Clusters' Profiling Premium Features (categorical variables)

label	YearAsCustomer	SalarySpentRatio
0	-0.860763	0.745563
1	0.894772	-0.728555
2	0.860765	0.735049
3	-0.829269	-0.751689

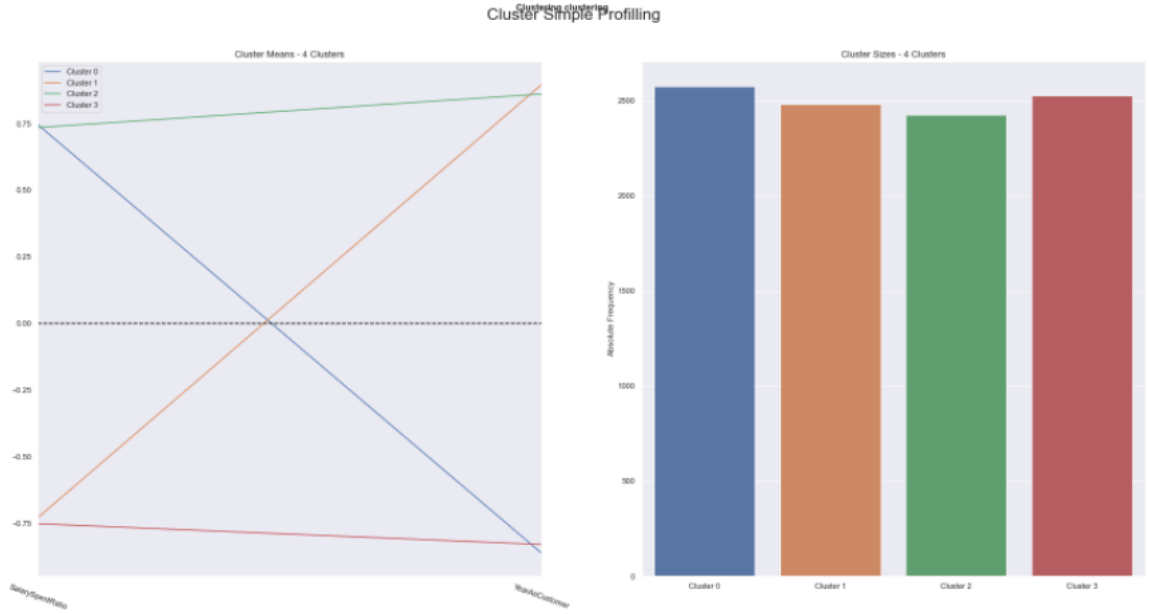

Figure 6- Clusters' Profiling Customer Features

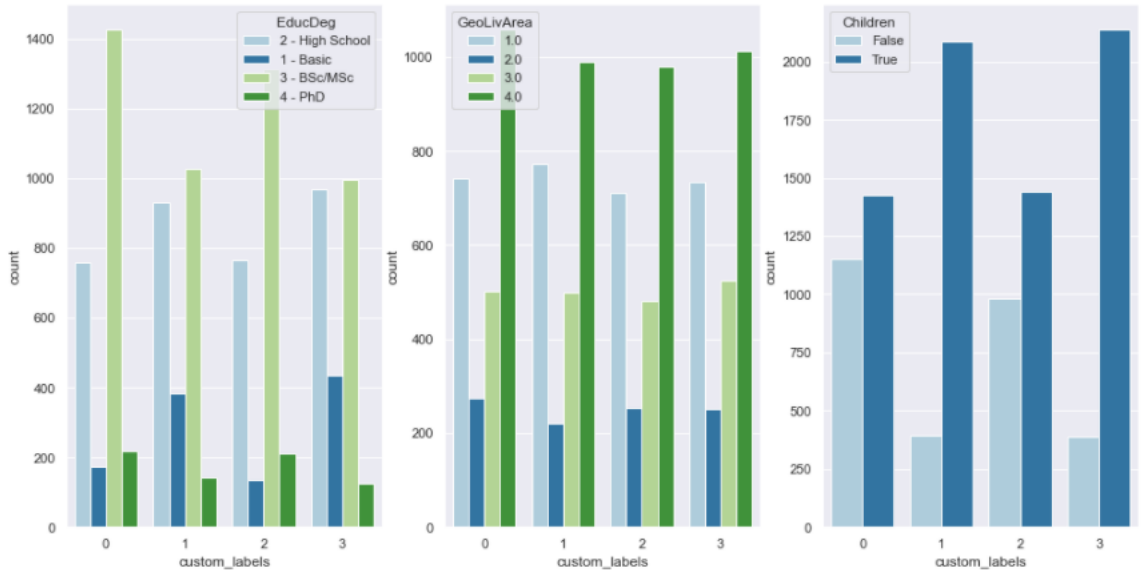

Figure 7 – Clusters' Profiling Customer Features (categorical variables)

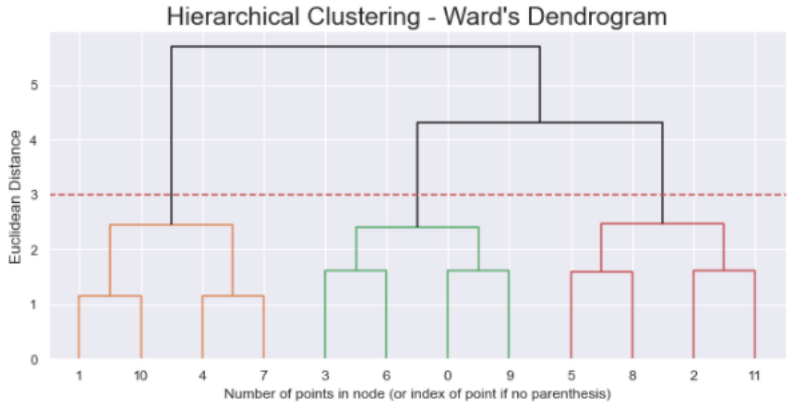

Figure 8 - Dendrogram (Selection of the final clusters)

	PremMotorRatio	PremHealthRatio	PremLifeRatio	PremWorkRatio	PremHouseholdRatio	SalarySpentRatio	YearAsCustomer	MonthSal
0	-0.988019	-0.330707	0.600833	0.473172	1.135179	-0.610560	-0.027259	-0.234550
1	1.071807	-0.597112	-0.596600	-0.588847	-0.692353	0.331639	0.009890	0.091571
2	-0.213116	0.990824	-0.014903	0.078393	-0.299174	0.212544	0.013783	0.124240

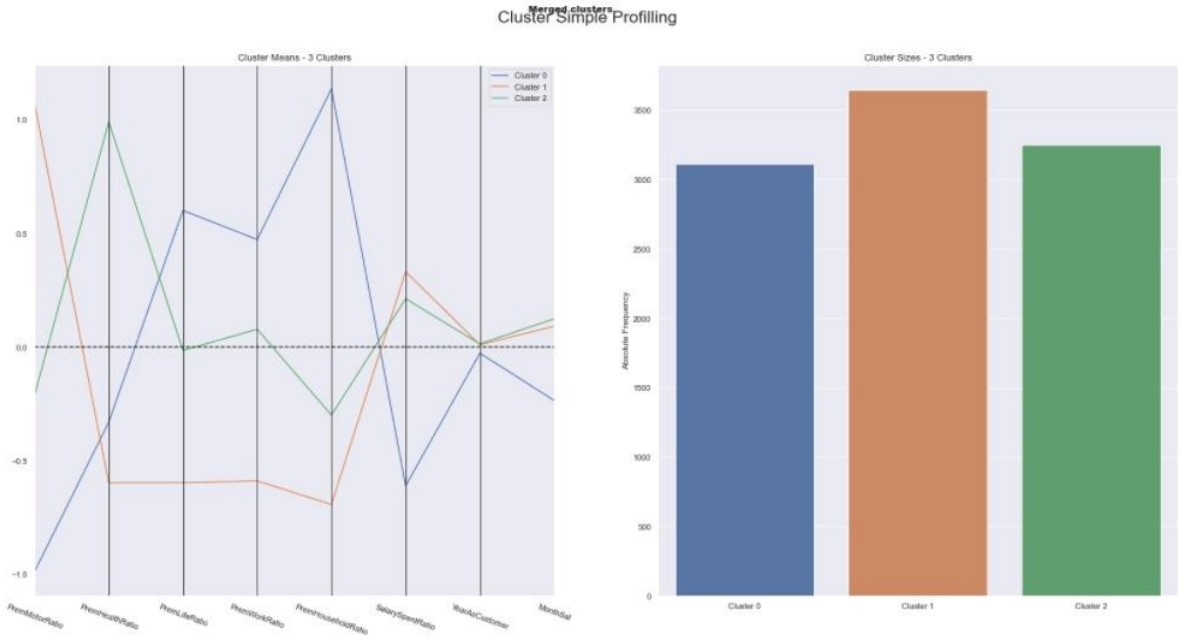

Figure 9- Final Clusters' Profiling

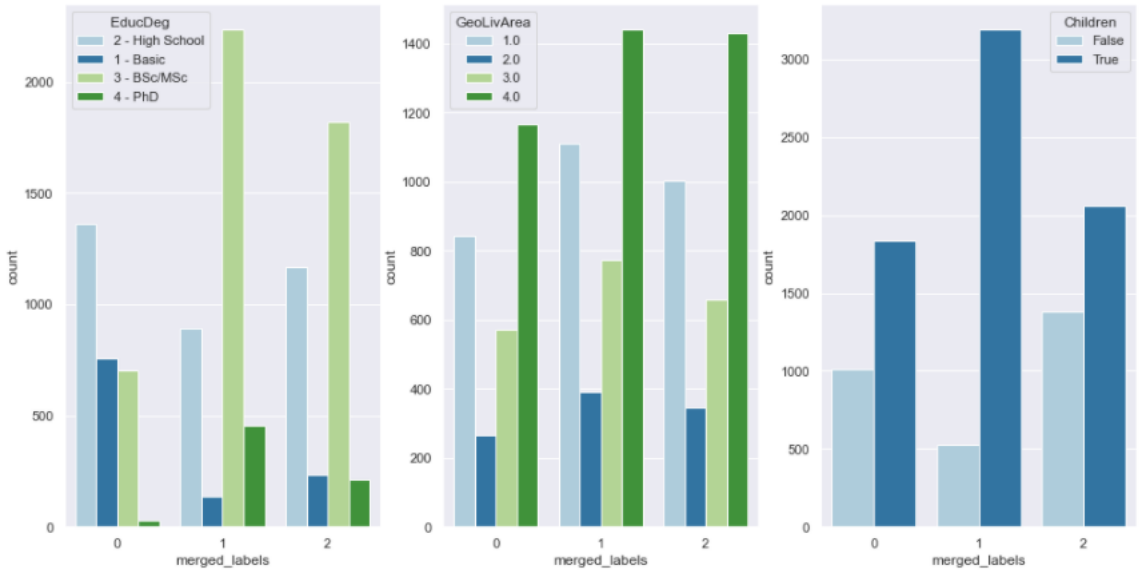

Figure 10 - Final Clusters' Profiling (categorical variables)

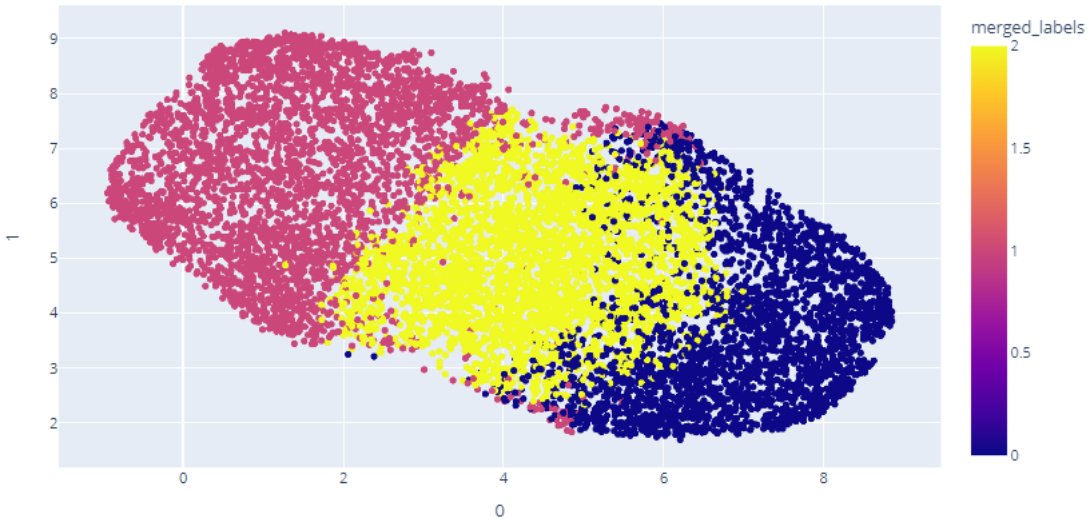

Figure 11 – UMAP