Data Mining Project

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Insurance Company Customer Segmentation

Group BG

Guilherme Simões, number: 20211003

Manuel Borges, number: 20200596

Rodrigo Joaquim, number: 20211024

January 2022

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

INDEX

1. Introduction iii
2. Data and Variables description (SAMPLE) iii
3. Exploratory Data Analysis (EXPLORE) iv
   3.1. Descriptive Statistics iv
   3.2. Coherence checking v
   3.3. Correlations v
   3.4. Outliers v
4. Data Pre-processing (MODIFY) vi
   4.1. Missing Values Assessment vi
   4.2. Transforming variables vi
      4.2.1. Outliers for new variables vii
   4.3. Feature Selection vii
5. Modelling (MODEL) vii
   5.1. Cluster algorithms vii
      5.1.1. K-Means viii
      5.1.2. Self-Organizing Maps (SOM) viii
      5.1.3. Self-Organizing Maps (SOM) + K-means viii
      5.1.4. Self-Organizing Maps (SOM) + HC viii
      5.1.5. Mini Batch K-means viii
      5.1.6. Gaussian Mixture Model (GMM) viii
      5.1.7. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) ix
      5.1.8. Hierarchical clustering (HC) ix
      5.1.9. DBSCAN ix
   5.2. Chosen clustering algorithm ix
6. Result and Discussion (ASSESS) ix
   6.1. Customer segmentation ix
   6.2. Customer Profile (Personas) ix
   6.3. Marketing approaches propose xi
7. Conclusion xii
8. References xiii
9. Appendix xiv
      9.1.1. Data Visualization xiv
      9.1.2. Clusters xvi

1. Introduction

This project was developed with the purpose of analysing the insurance company´s dataset to cluster the given data and give some marketing approaches.

While doing so, the data was processed, to create new variables, delete the ones that have no value, and correct inconsistencies and missing inputs. Our mission as data analytics consultants is to provide the Portuguese-based insurance company the best customer segmentation and profiling based on their characteristics and choices, while tailoring the ideal marketing approach. This will bring multiple benefits to the organization like cost reduction and more satisfied customers. All actions performed and explained in this report will help and lead us to a better understanding of the insurance company client's behaviour, identifying the different segments and their characteristics.

To support us and lead the project in the most correctly and accurate way possible we ended up following a pre-defined methodology. According to the needs and objectives of the project we decided to choose the SEMMA methodology, having also in consideration using the CRISP-DM method.

The SEMMA methodology is a tool created by SAS based on 5 principles. We start with a **Sample** of the data, then **Explore** it with basic description, visualization, and statistics of the data – Exploratory Data Analysis. After, **Modify** the data – here we divide the data pre-processing into missing values assessment, dealing with the outliers and feature selection. Next is the **Model** step, where we implement and try different clustering algorithms and finally the **Assess** step – where the analysis of the results and a have been done and a marketing segmentation is suggested to the client.

To initialize the process, and considering 2016 as the current year, we will first clarify the business understanding and objectives and the company's data mining goals. To succeed in this process, starting with a data preparation and a data pre-processing is essential to improve future analysis.

2. Data and Variables description (SAMPLE)

The data provided by the insurance company throughout the *Analytic Based Table* platform contains information regarding 10296 customers that are characterized by 13 different variables. There are 5 sociodemographic variables: *Birthday*, *Education*, *Area*, *Children*, and gross *Salary* in euros.

Regarding the variables towards the insurance company, we have:

- the *Customer Monetary Value (CMV)* that represents the annual profit from the customer times the number of years that they are a customer minus the acquisition cost of the customer.
- The *Claims Rate* is a variable that calculates the ratio between the amount paid by the insurance company.
- The *Year of the customer's first policy*.
- And finally, there is data for the amount, in euros, of five premiums in the LOB. The premium for *Motor*, *Household*, *Health*, *Life*, and *Work Compensation*.

It is important to refer that the variables regarding the premiums can have negative values that express reversals occurred in 2016, paid in the previous year. We could assess if a client cancelled an insurance policy or not, based on the existence of negative values.

3. Exploratory Data Analysis (EXPLORE)

To have a reliable data set and apply it in the best possible way to obtain trustworthy results, we must start by cleaning the data. First, we examine the data - the set of missing values, duplicate observations, and data types. Then, after a deeper exploration of the data set, we proceed with specific and more detailed methods, such as checking the coherence and correlation of the data. Data preparation is a very relevant part on the entire pipeline of the project. It can take up to sixty to eighty percent of the entire analysis process and is critical to getting the data into the most useful form and quality for the next steps of the analysis. The goal is to transform the data so that its content can be best appreciated, and better results can be achieved - good data is a prerequisite for good models.

3.1. Descriptive Statistics

To gain insight into the data, we next created a table of descriptive statistics for the most important variables. A high dispersion of the data can be seen when comparing the minimum, maximum, and median of the variables, indicating possible outliers. In the first policy year, the biggest observation is 53.784 - a value that does not make sense. Moreover, using the year of birth as an example, we can easily understand that a person could not have been born in 1028 and still be alive.

Analysing the table below, we can also see that almost all variables have missing values. Depending on the case and the variable, we looked at different situations - We assumed the missing values for the premiums variables to be premiums that could not be paid, not missing information. If we subtract the total number of entries (10.296) from the number of observations counted, we know that 30 values are missing for the first insurance year, for example.

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10266	1991,1	511,3	1974	1980	1986	1992	53784
BirthYear	10279	1968	19,7	1028	1953	1968	1983	2001
MonthSal	10260	2506,7	1157,4	333	1706	2501,5	3290,25	55215
CustMonVal	10296	177,9	1945,8	-165680,42	-9,44	186,87	399,78	11875,89
ClaimsRate	10296	0,7	2,9	0	0,39	0,72	0,98	256,2
PremMotor	10262	300,5	211,9	-4,11	190,59	298,61	408,3	11604,42
PremHousehold	10296	210,4	352,6	-75	49,45	132,8	290,05	25048,8
PremHealth	10253	171,6	296,4	-2,11	111,8	162,81	219,82	28272
PremLife	10192	41,9	47,5	-7	9,89	25,56	57,79	398,3
PremWork	10210	41,3	51,5	-12	10,67	25,67	56,79	1988,7

Table 1 – Variables Descriptive Statistics

The percentiles can also be very useful in presenting some understandings from the data. For example, for Claims Rate, we can see that the 75th percentile is 0.98, while the maximum value is 256.2 - meaning that we may have upper outliers.

3.2. Coherence checking

After carefully exploring the dataset, it is necessary to better understand each variable. We began by confirming the birthday variable by narrowing the age from 1926 to the current year (2016) - so there was no one in the dataset older than ninety or whose data was entered incorrectly after 2016. After this process, we understood that a customer with incorrect data regarding their birthday year could not possibly have been born in 1028. It was also checked to see if there was a customer with a first registration before the birth year or after the current year. Initially a strange value was found with a first policy year of 53784, so we can easily rule out this observation based on their obviously incorrect information.
With this rule, we ended up with 1997 observations where the first policy year was before birth, and since this is a large percentage of the dataset, we could not delete all the incoherent rows. Since the first policy year was calculated by the company and the birthday was likely provided by the customers, we decided that it was more likely that customers had provided incorrect information when filling out the forms. By this logic, many customers could have provided incorrect information, but nothing can guarantee us that others did not. Therefore, we decided to delete the Birthday column.

3.3. Correlations

With the intention of understanding the relationships between the variables, we calculated the correlations between them using Spearman's correlation. Although we used the variable GeoLivArea to check the correlations, we know that no values are shown because it is a nominal variable. We found that there is only a strong correlation between the variables. This correlation is between CustMonVal and ClaimsRate, which have a negative correlation of approximately -1.0. (Fig 1)

3.4. Outliers

An outlier is a data point that deviates significantly from other observations. This can be due to an incorrect entry of values, a sampling error, or an unusual but true value in the data set.
In this type of project, such values can cause serious problems, affecting the variance and standard deviation of the data distribution and leading to skewed distributions that make analysis of the data difficult. So, the best we could do to solve this problem was to remove them, and to do this we used several approaches. First, we did it "by hand", that is, we plotted histograms and boxplots, and after checking where the outliers were, we removed them accordingly.

We then used 2 algorithms to help us identify the remaining outliers. The algorithms we used were Isolation Forest and Minimum Covariance Determinant. Isolation Forest gave a percentage of outlier removal of 1% and Minimum Covariance Determinant gave a percentage of outlier removal of almost 2% (1.99999%). We then used PCA to visualize the identified outliers before removing them.

As we can see in Fig. 2, there were still some outliers left. In numbers, with all the approaches we used, we ended up removing a total of 345 observations.

4. Data Pre-processing (MODIFY)

4.1. Missing Values Assessment

Feature	Missing Values
FirstPolYear	30
EducDeg	17
MonthSal	36
GeoLivArea	1
Children	21
CustMonVal	0
ClaimsRate	0
PremMotor	34
PremHousehold	0
PremHealth	43
PremLife	104
PremWork	86

Table 2 – Missing Values

After observing the table..., we saw that *GeoLivArea* had only one missing value. As we have 10293 rows and for this variable, we only had one missing value, we decided to delete the row.

We also saw that FirstPolYear contained 30 missing values, and we thought of running a regression on them, but the result was 0% of the explained variance, so we deleted the 29 rows, since one already corresponded to the previously deleted row.

For the missing values on the Premiums, we decided that it was easier and better to replace all the missing values with 0, as this means that a customer has not spent any money on that specific premium, hence the 0, as in money spent for 0. From this point of view, it is not surprising that we get a total of 267 missing values. This number comes about because there is no requirement or rule that a customer must have more than one premium, many of them may have only one premium.

For the last three variables EducDeg with 17 missing values, Children with 21 missing values, and MonthSal with 36 missing values, we decided to impute the missing values using K nearest neighbours. To do this, we decided to use 7 neighbours because we tried other k values and found that k=7 gave the best results. The weight used was uniform, so all points in the neighbourhood were weighted the same.

4.2. Transforming variables

With the aim of creating a better, more precise, and meaningful segmentation of insurance customers, we considered it necessary to create new variables in addition to those already present in the dataset. Thus, 3 new variables were created:

- *Years_As_Client* measures how long clients have been with the insurance company. This is the difference between the current year (2016, year of record) and the first contract year of each client.
- *YearSal* measures the annual salary of clients. It is calculated by taking the monthly salary times 14 to account for Christmas and holiday bonuses. This feature was created because we felt that it would be useful to know the annual salary, since the amount spent on each premium is expressed as a value for a year.

- *Total_Premiums* sums up the amount of the money spent by each customer in each premium category, except the cancelled ones.
- *Cancelled* is a binary variable that indicates whether the client has cancelled an insurance policy.

4.2.1. Outliers for new variables

When we introduced these new variables, we again faced a familiar problem of dealing with outliers. We followed the same approach as before and started to deal with them "by hand", again creating boxplots and histograms. We saw that there were not so many outliers, so we applied two thresholds to the Total_Premiums variable and omitted the outliers. In the end, using the outlier removal from a previous point and this point, we removed about 3.5% total outliers from our data set.

4.3. Feature Selection

With the intention of achieving the best performance of our models in the next phase, it was of paramount importance that we select the best set of variables for the application of clustering algorithms. Therefore, we decided to use the Random Forest algorithm, an algorithm that can quickly and easily provide us with the importance of each feature, to help us select the best features for the clustering algorithms. We also thought about dividing the variables into two or three different categories, but after careful consideration, we concluded that due to the small number of features and data, we would get many clusters, some of which would not contain that many records, and some records could not be assigned to any cluster. For this reason, we selected 9 features that we considered the best solution considering the results of the random forest graph gave.

After doing this and checking the boxplots, we found that GeoLivArea was irrelevant. Since we omitted this irrelevant feature, all other features were up for discussion. To evaluate their importance, we decided to apply the random forest algorithm to help us decide which ones to use.

The selected features are the following: *PremWork*, *PremMotor*, *PremHealth*, *PremLife*, *Total_Premiums*, *YearSal*, *Years_As_Client*, *CustMonVal*, and *Cancelled*.

5. Modelling (MODEL)

5.1. Cluster algorithms

After determining what features and how many clusters we would use, we began applying the clustering algorithms to see which algorithm would produce the better results. We applied the following algorithms: k-means, SOM + k-means, SOM+ HC, Mini Batch K-means, GMM, Birch, HC, and DBCAN.

5.1.1. K-Means

It is an algorithm that aims to divide observations into a number of k clusters, where each observation belongs to the cluster with the closest mean. By using k-means, we obtained good results because the clusters behaved very differently in almost every category, with clusters 0 and 1 having a similar size and cluster 3 representing the largest number of individuals. (Fig 4)

5.1.2. Self-Organizing Maps (SOM)

Is one of the most popular clustering algorithms because it provides dimensionality reduction by transforming the high-dimensional input space into a typically two-dimensional one.
In addition, one of the best features of SOM is that it does not make any assumptions about the distribution of the variables, and therefore it can handle nonlinear relationships and skewed distributions very well. With this algorithm, we sometimes get a not so clear representation of the data. To overcome this, we applied K-means and Hierarchical Clustering to SOM to see how the results would differ

5.1.3. Self-Organizing Maps (SOM) + K-means

With this method, some reasonable results can be obtained because the differences between variables are not so large and the size of clusters decreases from cluster to cluster. (Fig. 5)

5.1.4. Self-Organizing Maps (SOM) + HC

This method gave much better results because the clusters are similar and different in different areas at the same time, which means that we have a good segmentation of customers. The clusters have a good size because we have two clusters that are larger than the others, although one is slightly larger than the other, and the third is slightly smaller than the other two. (Fig 6) and (Fig 7)

5.1.5. Mini Batch K-means

This algorithm uses small random stacks of data with a fixed size so that they can be stored in memory. At each iteration, a random sample is taken from the data set and used to update the clusters. This process is repeated until convergence. Using mini-batch k-means can result in significant savings in computation time, but the clusters may lose some quality. Since this algorithm is based on k-means, it was only normal that it would also give a good result even though the observations are not so well distributed among the clusters. (Fig 8)

5.1.6. Gaussian Mixture Model (GMM)

This is a probabilistic model that assumes that all data points are generated from a mixture of Gaussian distributions with unknown parameters. This model is very similar to k-means, but the difference is that GMM incorporates information about the covariance structure of the data as well as the centre of the latent Gaussian distributions. In this algorithm, the distribution of clusters was not so good, because we have one cluster with a very large number of observations and another with a very small number, and therefore this algorithm is not so good. (Fig 9)

5.1.7. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Clusters large datasets by creating a smaller and more compact dataset that retains as much information as possible from the original dataset. Then, the clustering is applied to this smaller dataset instead of the original, larger dataset. The distribution of observations was not very good, as clusters 0 and 2 are very similar. (Fig 10)

5.1.8. Hierarchical clustering (HC)

Connect the most similar examples and gradually add the most similar clusters together until all the connections are found and the result can be represented as a dendrogram. The observations are more or less well distributed and do not behave too similarly in terms of features (Figs. 11, 12)

5.1.9. DBSCAN

Summarizes points that are close to each other based on a distance measure (Euclidean distance is most commonly used) and a minimum number of points. This algorithm also has the distinction of flagging points that are in low-density regions as outliers. DBSCAN did not perform as well with the data provided, as the cluster with the largest number of observations almost did not change its behaviour when interacting with the features. (Figs. 13 and 14)

5.2. Chosen clustering algorithm

After applying the optimal features to all 9 clustering algorithms we chose, the algorithm with the best performance when using the r-squared metric for measurement was SOM combined with Hierarchical Clustering, by some margin, as we can see in Fig. 15. After selecting the final algorithm, we performed Multiple Correspondence Analysis (MCA) to link the categorical variables to the clusters created.

6. Result and Discussion (ASSESS)

6.1. Customer segmentation

One of the most critical aspects of this project was to visibly identify different clusters with apparent differences in its profile of clients and its characteristics. This was exactly what was desired to be possible to create multiple marketing campaigns accordingly with each type of client and their characteristics and behaviours. The optimal number of clusters chosen for defining the most relevant characteristics keeping their identity were three. In the Figure below we can observe the main characteristics of each cluster and the number of clients inserted in each segment.

6.2. Customer Profile (Personas)

A persona is common marketing approach tool that allows us to better understand and visualize the target customers, a lot like centroids. Although is not a “mean” persona, is a representation of a cluster´s key characteristics. It´s a a way of better understanding the type of client and their behaviour in a more practical and interesting manner.

Fig 7: Simple Profiling using Self-Organizing Maps (SOM) + HC

Persona 1 - cluster 0 – Michael Scoot – Old loyal client with high income but low spending's

Mike is an adult businessman and owner of a paper company. Although he has a very good salary, he does not spend much on insurance premiums. Most of his expenses are on car insurance for his cars. Michael is considered a very loyal customer, as he has been a customer of the insurance company for more than twenty years. However, he knows that he is not a good customer because of his weak and cheap insurance decisions.

Persona 2 – Cluster 1 – Pam Beesly - Recent and reasonable spending´s client with low value despite his average income.

Pam is a young secretary in a paper company. She is very concerned about her health because she is the vast majority of the insurance company's customers, and so almost all of her expenditures on insurance are for the best health insurance. Miss Beesly represents the vast majority of customers with average salaries and normal insurance premiums. Because Pam has only recently purchased her insurance and has not chosen the best options, she is considered a low-value customer by the insurance company.

Persona 3 – Cluster 2 - Darryl Philbin – Despite low income, recently gain a high concerned with his insurance.

Darryl is a middle-aged warehouse worker who works with machinery and heavy boxes. He knows that he has a dangerous job and therefore considers it very important to have a decent insurance policy, despite his low income. Precisely because of his job, he spends mainly on life and work insurance. Despite his young age as a customer, Mr. Philbin is very valuable to the company because he goes out of his way to buy such good insurance. Darryl represents a small portion of the high cash value customers.

6.3. Marketing approaches propose

After some clustering techniques have been assembled, tested, and properly understood is finally possible to accurately propose and plan the best marketing approach with each group of clients, according obviously to the conjugation of several aspects – it will be taken into consideration the gap of new clients, the value of each segment to the company and the profiles of each individual characteristics.

There are several and differentiated approaches for each segment and their own identities. Our team supported with research in marketing principals in the Business to consumer (B2C) approach, and constantly having into consideration what the data "tells" us, have defined what might be the best approach for each segment of customers:

- For the oldest clients with an extended insurance policy and high income but that doesn´t like to spend a lot – cluster 0 – would be good option for the insurance company to highlight the possibility of **memberships** with higher quality insurances in multiple marketing emails and messages to their clients to let them know better options and the advantages of activating a membership – it could benefit the organization by making this type of clients expend more money in their membership to upgrade their insurance
- Regarding the segment from cluster 1 and having into account their characteristics - their main concern is the health insurance and reasonable salary - it is crucial to offer them **packaging solutions** with the possibility of having quantity discounts in case of multiple insurance subscriptions. This would allow the company to convince this type of client to increase the variety of protections and covers.
- As we can see in the segment from cluster 2 – This type of client reveals a big effort to pay all their insurances and so it is very important to host **creative and engaging content** related not only to the company but also their insurance options to motivate them to keep them or even upgrade their possibilities.

7. Conclusion

To conclude, our main goal was to provide the marketing department of the Portuguese-based insurance company the best understanding of theirs's clients' profile and characteristics by analyzing all the dataset provided and developed a marketing plan.

Even though some crucial information regarding the costumers were missing in the dataset which made it harder for the clustering process, we truly believe we have come up with very good results – our team has prepared this report to summarize, explain and propose a marketing segmentation an approach. To help us come up with these results we used python programming language.

The lack of information from clients is one of the points of improvement for the insurance company in the future – Improve their client's data collect to obtain better exploratory analysis and segments.

We started by exploring the dataset and deal with all the missing values, duplicated values, incoherence's in the data, and outliers. After that we reduced the size of our features by dropping the irrelevant ones and creating new ones, that we tough were necessary.

The next step was to select the number of clusters, and which features to use. To do that we used the elbow method to get the number of clusters and concluded using 3 and used the random forest algorithm to get the best features.

Since we had already everything ready to do the analysis, we applied the best 9 features to all the clustering algorithms we chose to use.

At the end the one we ended up choosing was the SOM with HC, because it was the one that gave the best performance when we used the r-squared metric the evaluate the performance of each one of them.

Finally, the marketing approach was carefully constructed not only based on what the data "tells" us but also having into consideration budget, priorities, and importance.

8. References

Petra Perner (Ed.). (2010). Advances in Data Mining: Applications and Theoretical Aspects. Berlin; New York: Springer, ©2010.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert systems with applications, 40(1), 200-210.

Prado, K. S. D. (2019, June 3). How DBSCAN works and why should we use it? - Towards Data Science. Medium. https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

Stardardization in Cluster Analysis. (2018, September 24). Alteryx Community. https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/Standardization-in-Cluster-Analysis/ta-p/302296

GeeksforGeeks. (2019, May 13). ML | Mini Batch K-means clustering algorithm. https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/

Learn, S. C. I. K. I. T. (2021). 2.1. Gaussian mixture models. Scikit-Learn. https://scikit-learn.org/stable/modules/mixture.html

GeeksforGeeks. (2020, July 7). ML | BIRCH Clustering. https://www.geeksforgeeks.org/ml-birch-clustering/

SEMMA. (2021, October 11). Data Science Process Alliance. https://www.datascience-pm.com/semma/

Stark, A. (2021, December 13). Data Preprocessing & Exploratory Data Analysis (EDA) for Data Science: Tackling the Taarifa Challenge. Medium. https://towardsdatascience.com/data-preprocessing-and-eda-for-data-science-50ba6ea65c0a

9. Appendix

9.1.1. Data Visualization

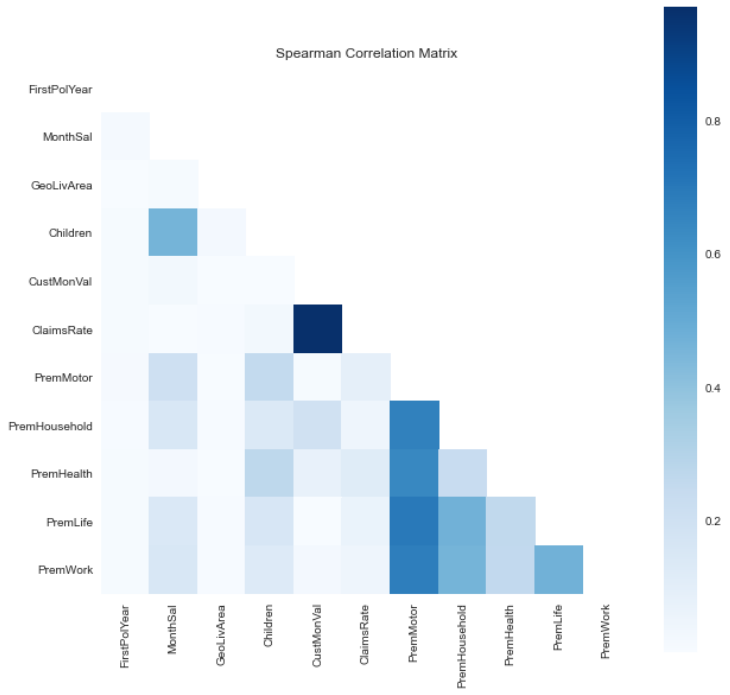

Fig 1: Spearman Correlation matrix

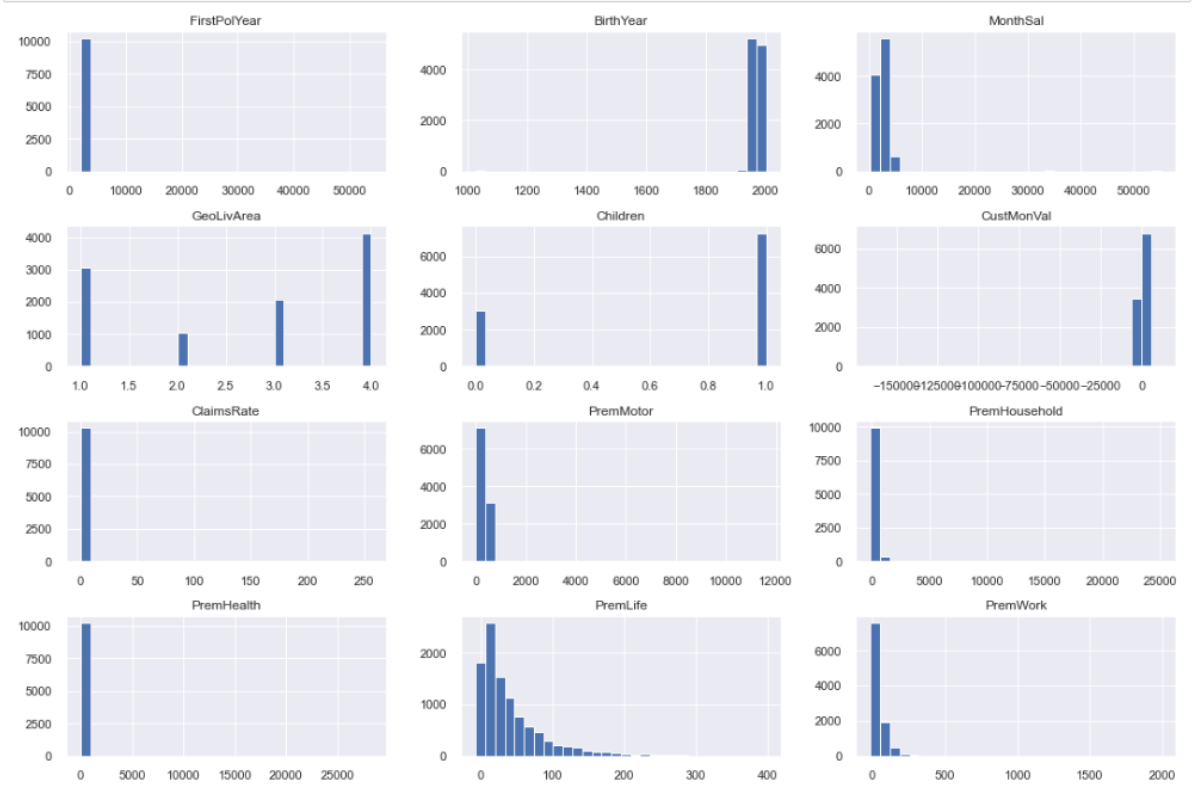

Fig 17: All variables' histograms

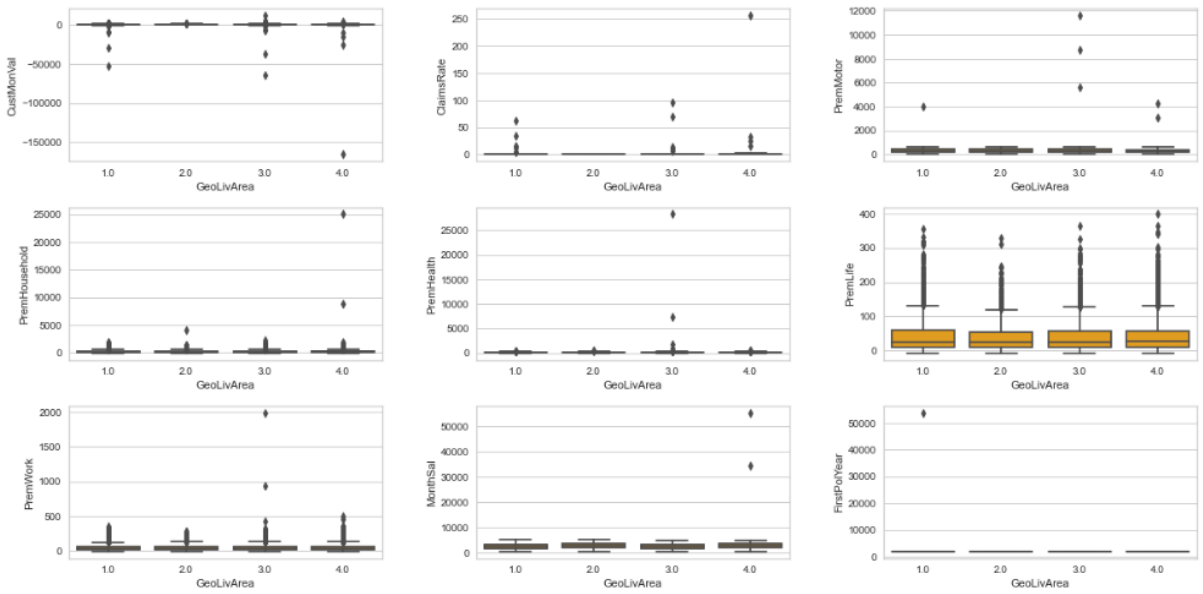

Fig 18: Boxplots to check relevance of GeoLivArea

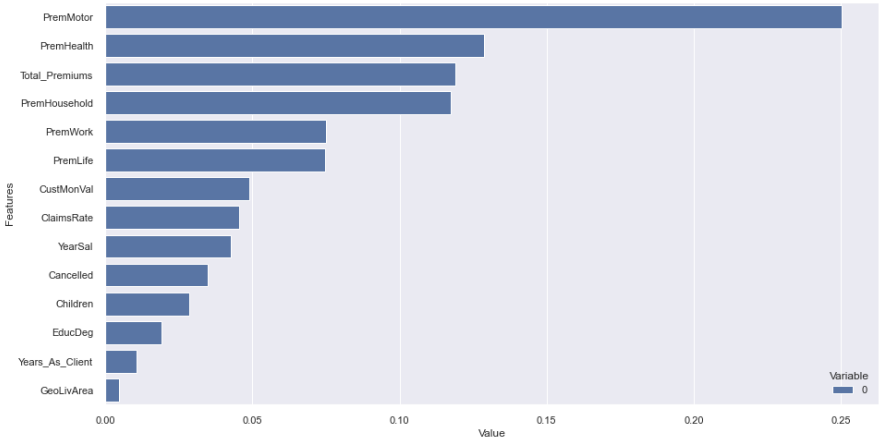

Fig 3: Random Forest algorithm

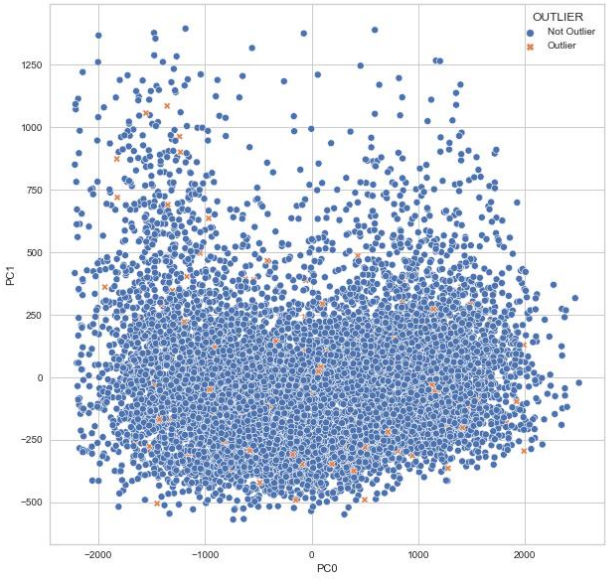

Fig 2: PCA outliers' visualization

9.1.2. Clusters

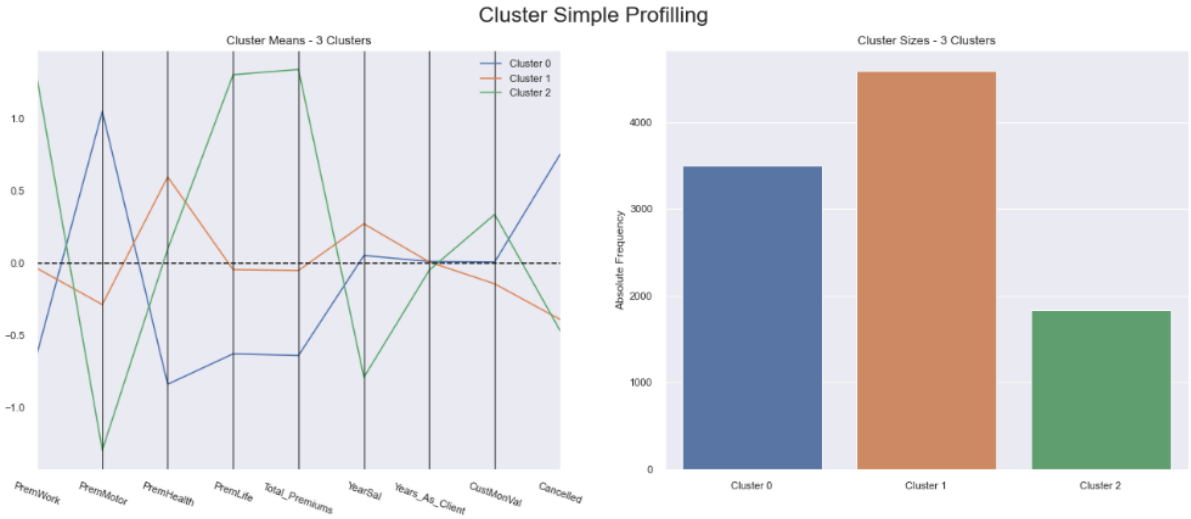

Fig 4: Simple Profiling Value using K-means

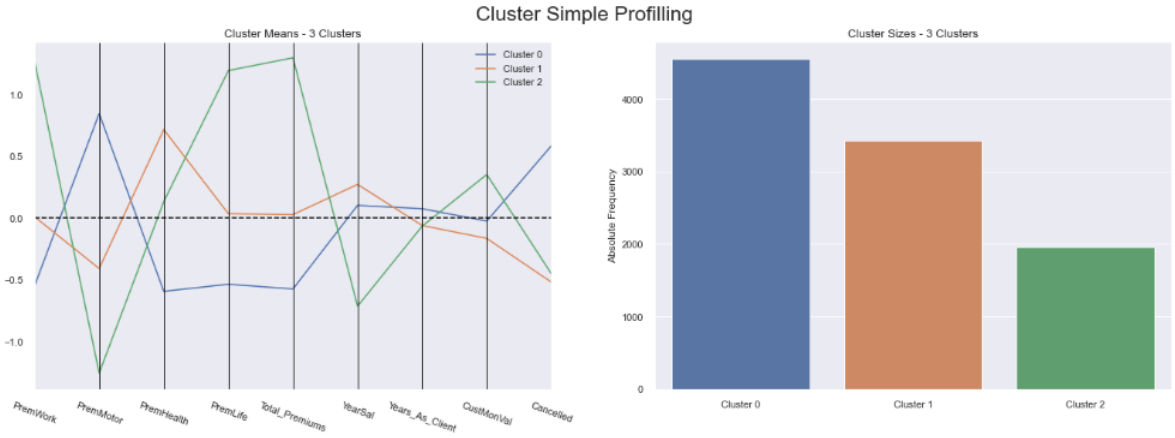

Fig 5: Simple Profiling Value using SOM + K-means

Fig 6: SOM + HC R2 plot to decide HC method

Fig 7: Simple Profiling Value using SOM + HC

Fig 8: Simple Profiling Value using MiniBatchKmeans

Fig 9: Simple Profiling Value using GMM

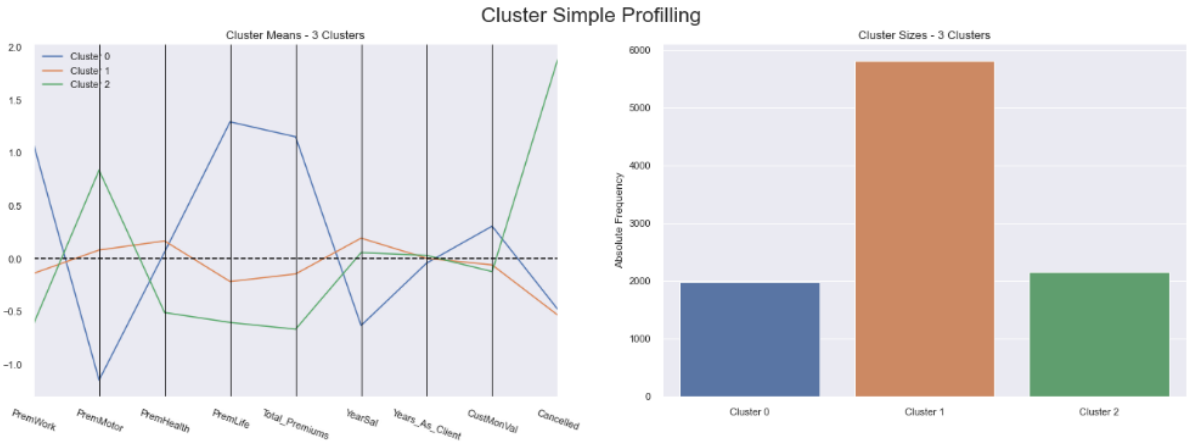

Fig 10: Simple Profiling Value using Birch

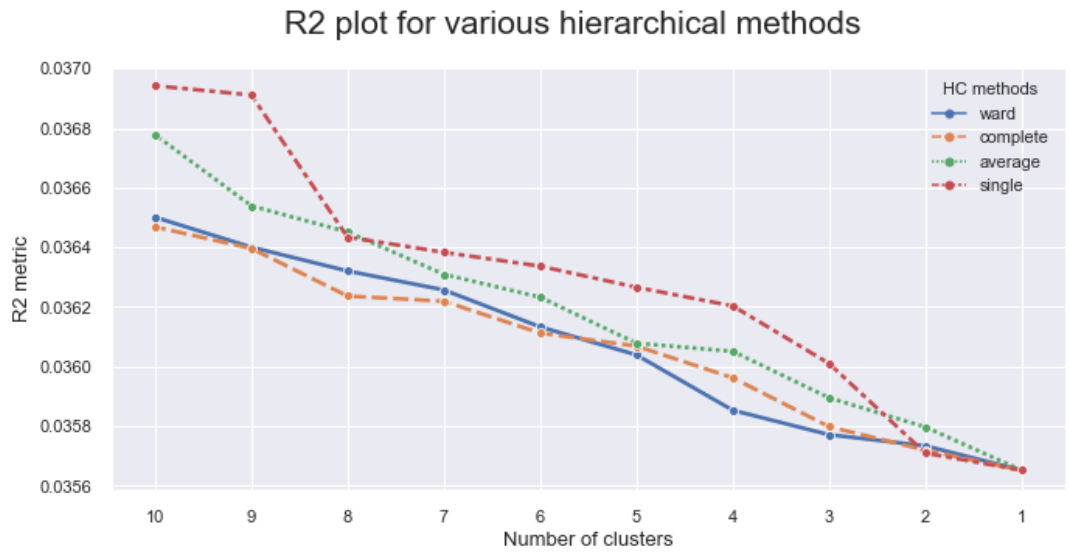

Fig 11: HC R2 plot to decide HC method

Fig 12: Simple Profiling Value using HC

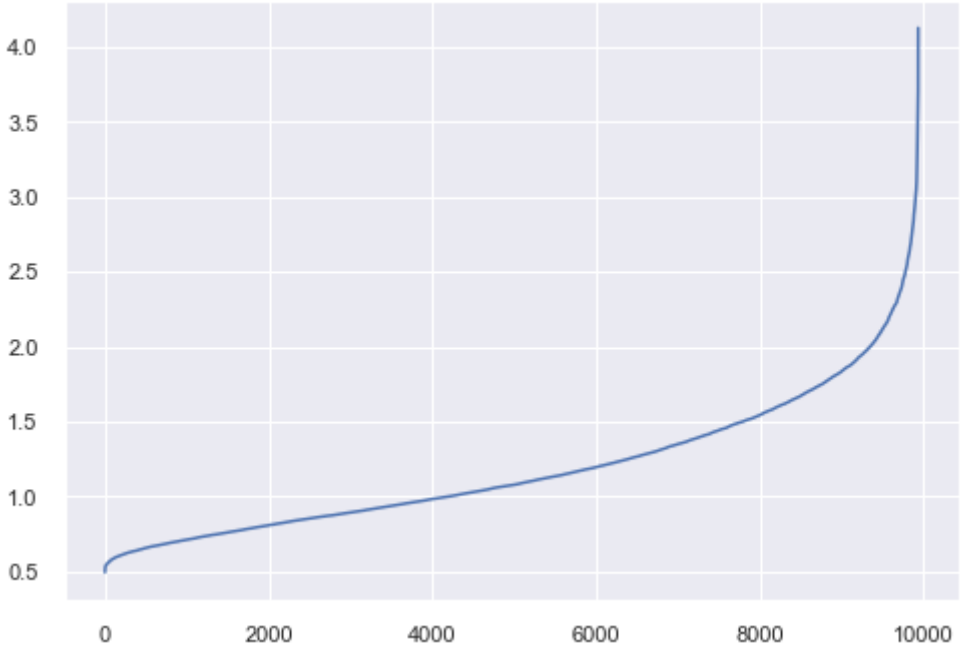

Fig 13: DBSCAN k-distance graph

Fig 14: Simple Profiling Value using DBSCAN

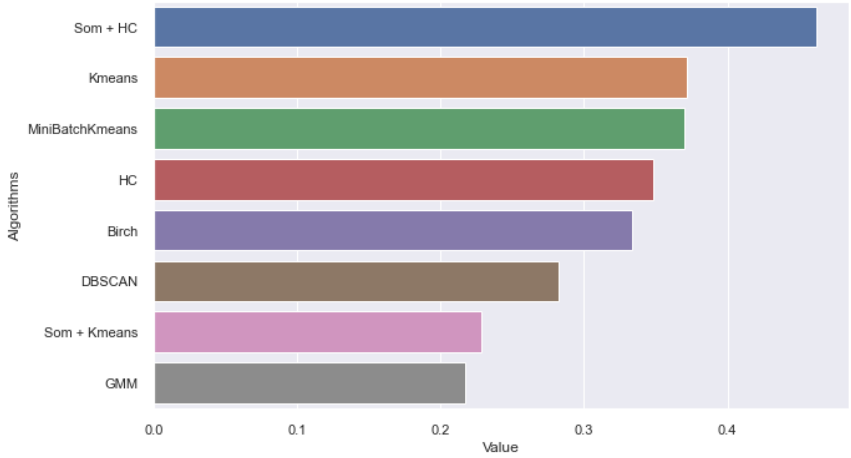

Fig 15: R2 metric comparison to decide best algorithm

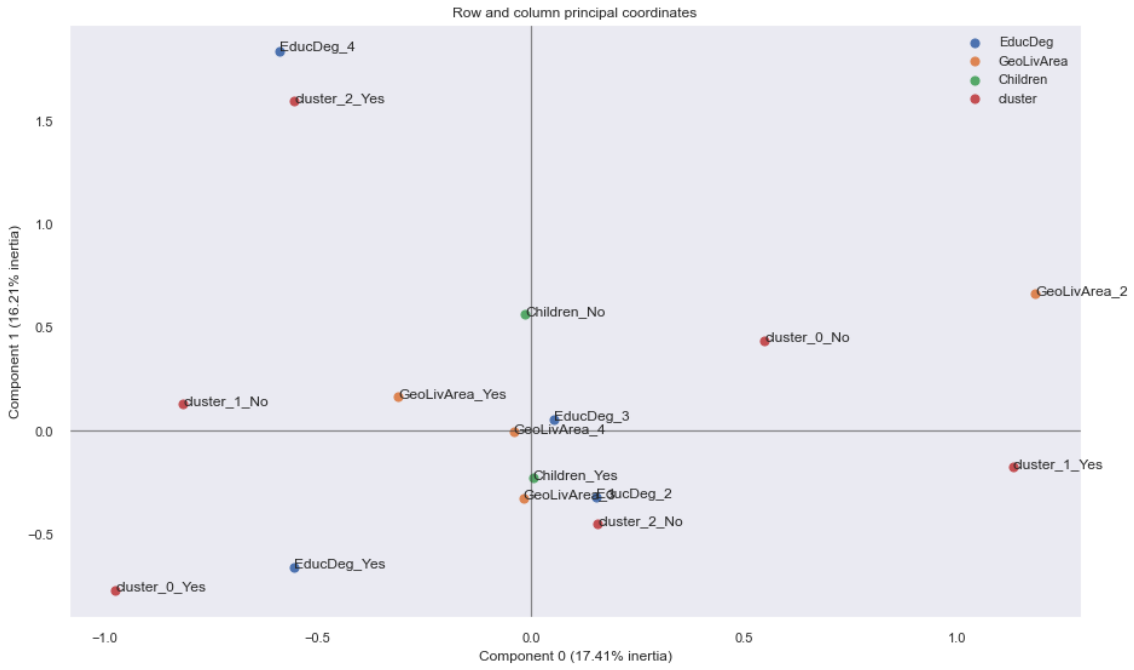

Fig 16: Multiple Correspondence Analysis graph