Data Mining Project Report

Segmentation of An Insurance Company Client Dataset

Authors:

David Silva 2016730
Susana Paco 20190821

Introduction

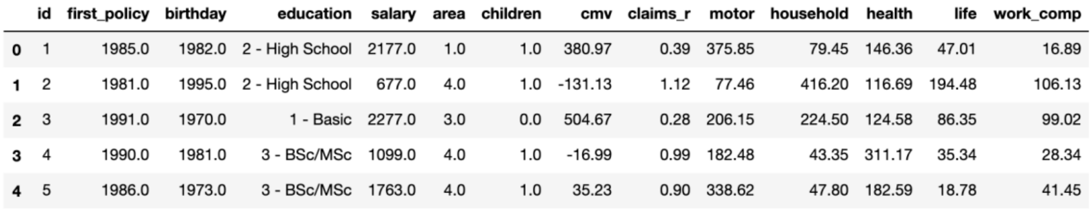
The client, an insurance company, wishes to better understand the scope of its clients, in order to better serve them and increase their ROI (Return On Investment).
The group was given an ABT (Analytic Based Table), consisting of 10.290 customers and given the task of analyzing the table for evident groups of clusters, extracting the behaviour of said clusters and provide insights on how to better serve them.

The project is contained within a GitHub repository which can be accessed through the following link: https://github.com/theinsilicobiology/DataMiningFinalProject. Inside the repository, there is a Jupyter Notebook with all the relevant analysis. All the theoretical references used to justify every decision in this process are contained within the notebook, appended to the relevant code section that utilizes these references.

Data Understanding

As a first step, grasping the contents of the dataset is a crucial task.
After uploading the dataset to the notebook, we realized that we are dealing with a dataset of 10290 customers with 14 variables (Figure 1).

These 14 variables can be divided into two groups, being that "education", "area" and "children" are non-metric variables and "first_policy", "birthday", "salary", "cmv", "claims_r", "motor", "household", "health", "life" and "work_comp" are metric variables. These variables should be dealt with separately from now forward.



	id	first_policy	birthday	education	salary	area	children	cmv	claims_r	motor	household	health	life	work_comp
0	1	1985.0	1982.0	2 - High School	2177.0	1.0	1.0	380.97	0.39	375.85	79.45	146.36	47.01	16.89
1	2	1981.0	1995.0	2 - High School	677.0	4.0	1.0	-131.13	1.12	77.46	416.20	116.69	194.48	106.13
2	3	1991.0	1970.0	1 - Basic	2277.0	3.0	0.0	504.67	0.28	206.15	224.50	124.58	86.35	99.02
3	4	1990.0	1981.0	3 - BSc/MSc	1099.0	4.0	1.0	-16.99	0.99	182.48	43.35	311.17	35.34	28.34
4	5	1986.0	1973.0	3 - BSc/MSc	1763.0	4.0	1.0	35.23	0.90	338.62	47.80	182.59	18.78	41.45

Figure 1 Dataset head

The variables "first_policy", "birthday", "education", "salary", "area", “children” and “cmv” directly refer to the clients, being characteristics specific to each customer while "claims_r", "motor", "household", "health", "life" and "work_comp" refer to the products provided by the company. Doing a separate analysis of customer vs products might be a handy process in the future.

Knowing the influence of outliers in a dataset, producing the box plots allows us to have a first grasp on the quality of the dataset. As we can see from Figure 2, the dataset is highly influenced by outliers, specially in the product variables. The existence of these outliers is a hindrance to the quality of the results produced by any model and will be dealt with.

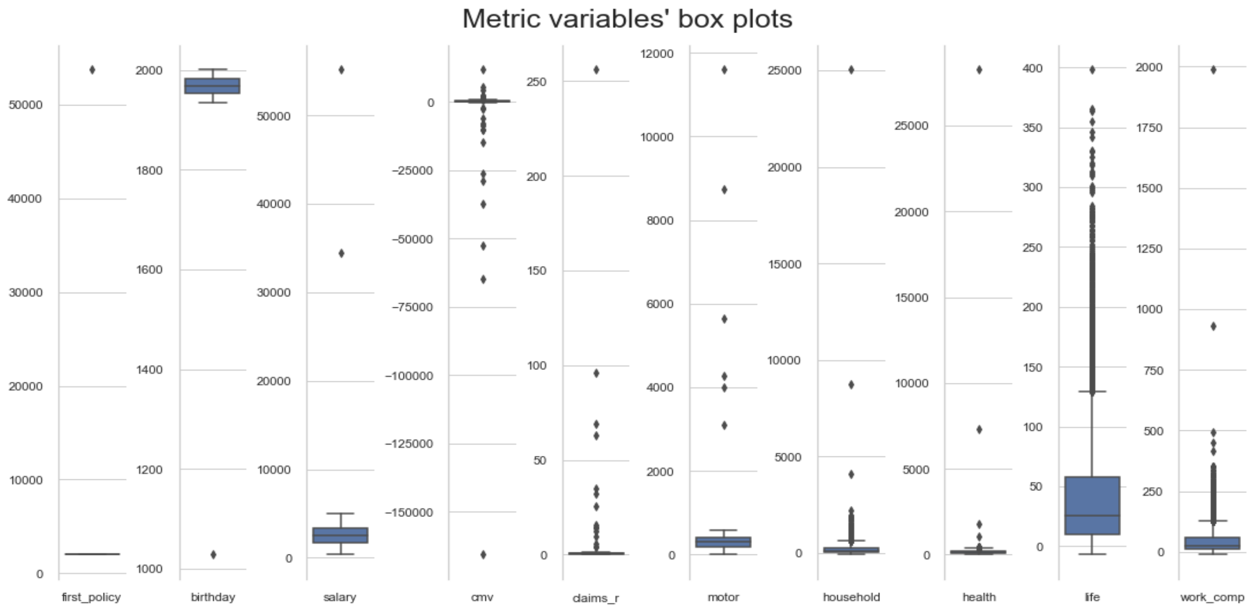


Figure 2 Metric variables' box plots

A good first visualization policy, the correlation matrix was plotted (Figure 3) which indicates us which variables have the potential of becoming useful in posterior modelling. This first correlation matrix is quite disheartening, showing not very good values. This dataset demands a serious and deep process of data preparation in order to improve data quality which will reflect on improvement of this matrix.

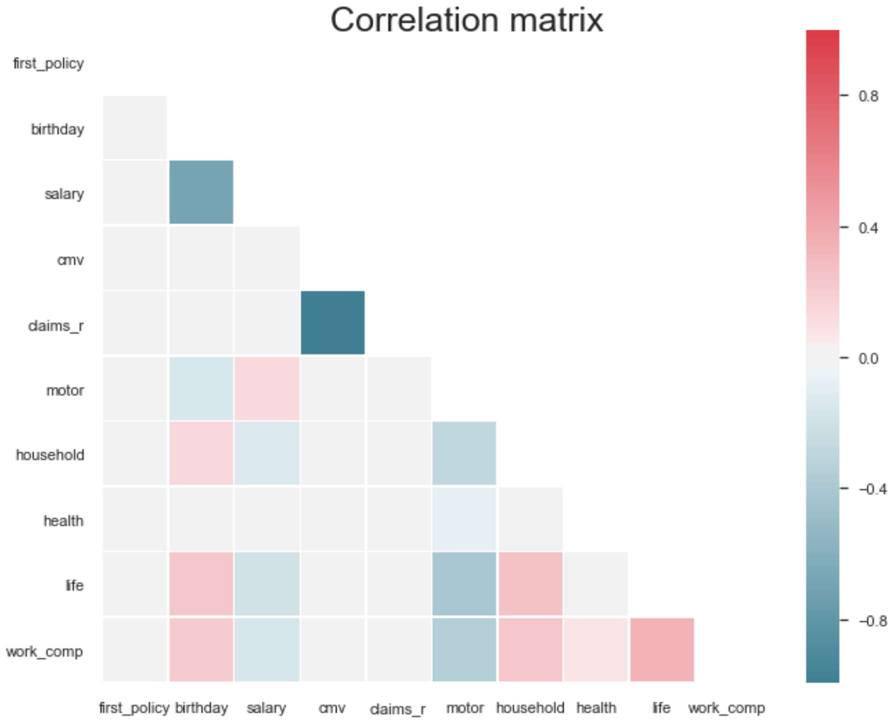


Figure 3 Correlations matrix

With this, the tasks for the data preparation were set. The dataset has errors to be dealt with, several outliers, missing values and requires some feature engineering.

Data Preparation

Data preparation is a crucial first task in a data mining project. It allows us to remove most errors in the dataset, dealing with missing values and overall improve the dataset conditions before moving to modelling. Categorical variables should be dealt with separately in order to assess their importance for the final analysis and finally, as the dataset has not many variables, some feature engineering might come in handy.

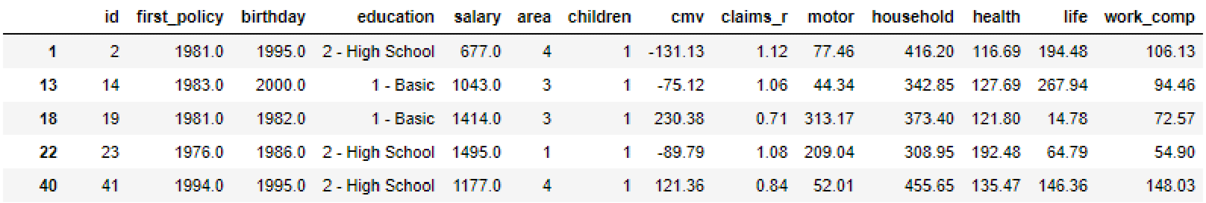
Dealing With Errors

From the data understanding, we realized that this dataset requires quite some work before modelling. First, we'll begin by dealing with the errors in the dataset, classified into 3 groups: Coherence Checks, Outliers and Missing Values.

Coherence Checks

As this dataset supposedly comes from a real-life situation, a first check should be to verify that the data is coherent with reality. For this, a few sanity checks were performed in accordance to standard questions of the area.

"Do we have clients where the first policy date predates their birthday?" We realized that in this dataset (Figure 4) we have 1997 clients where the first policy is older than their own birthday. While not impossible (family policies and company policies that predate the birth of a client are possible) it is an uncommon event.



	id	first_policy	birthday	education	salary	area	children	cmv	claims_r	motor	household	health	life	work_comp
1	2	1981.0	1995.0	2 - High School	677.0	4	1	-131.13	1.12	77.46	416.20	116.69	194.48	106.13
13	14	1983.0	2000.0	1 - Basic	1043.0	3	1	-75.12	1.06	44.34	342.85	127.69	267.94	94.46
18	19	1981.0	1982.0	1 - Basic	1414.0	3	1	230.38	0.71	313.17	373.40	121.80	14.78	72.57
22	23	1976.0	1986.0	2 - High School	1495.0	1	1	-89.79	1.08	209.04	308.95	192.48	64.79	54.90
40	41	1994.0	1995.0	2 - High School	1177.0	4	1	121.36	0.84	52.01	455.65	135.47	146.36	148.03

Figure 4 Incoherent individuals

"Do we have minors with children?" This is a given, clients that are reported as being younger than 18 should not have children. While not impossible, it should be a rare event. However, we found that 87 clients are reported as minors and having children.

"Do we have minors with car policies?" As a final sanity check, this should be an impossible behaviour as minors cannot drive in Portugal. We also found that 116 clients are minors with car insurance policies.

With these 3 sanity checks of the birthday variable failing by a significant margin, we've asked the client for feedback and realized that the amount of errors in this variable is too significant to continue using it. We then move forward removing the birthday variable from the analysis.

Outliers

The outlier removal process, as the dataset presents some serious challenges, was done in a several step protocol.

The analysis was divided into parametric and non-parametric methods, in the parametric methods, it was assumed that the multivariate distribution is close to a normal and therefore demanded that we transformed the variables to approximate this distribution.

Our protocol went as follows:

- Unreasonable values - eliminating values that do not make sense with reality;

- Univariate Distribution conversion to Normal - assuming that, in most cases, several univariate normal distributions approximate a multivariate normal distribution, we needed to convert every univariate distribution to a normal. The quality of these conversions was assessed using QQ plots which demonstrated a good approximation quality for some of the transformations. We used the best ones;
- Univariate Outliers - the first outlier removal techniques used were the parametric ones namely standard deviation test and interquantile range;
- Multivariate Outliers - we then moved to multivariate outlier removal techniques namely using the mahalanobis distance (Figure 5) within the Chi-Square test and using the Local Outlier Factor (LOF) method which is a density-based method for outlier detection (Figure 6);
- We stored all the results of these analyses in a table that describes the outcome of each combination of several outlier removal methods (Figure 7);
- Comparing the results, we concluded that the best outcome (the one that provides the best box plot quality while removing the least number of outliers) is a combination of the standard deviation, interquantile range, mahalanobis distance and the QQ plots.

With this we move forward by using the 4th combination of the table, removing a total of 48 outliers from the dataset. We also stored the outliers so that we could predict their cluster membership after the modeling process.

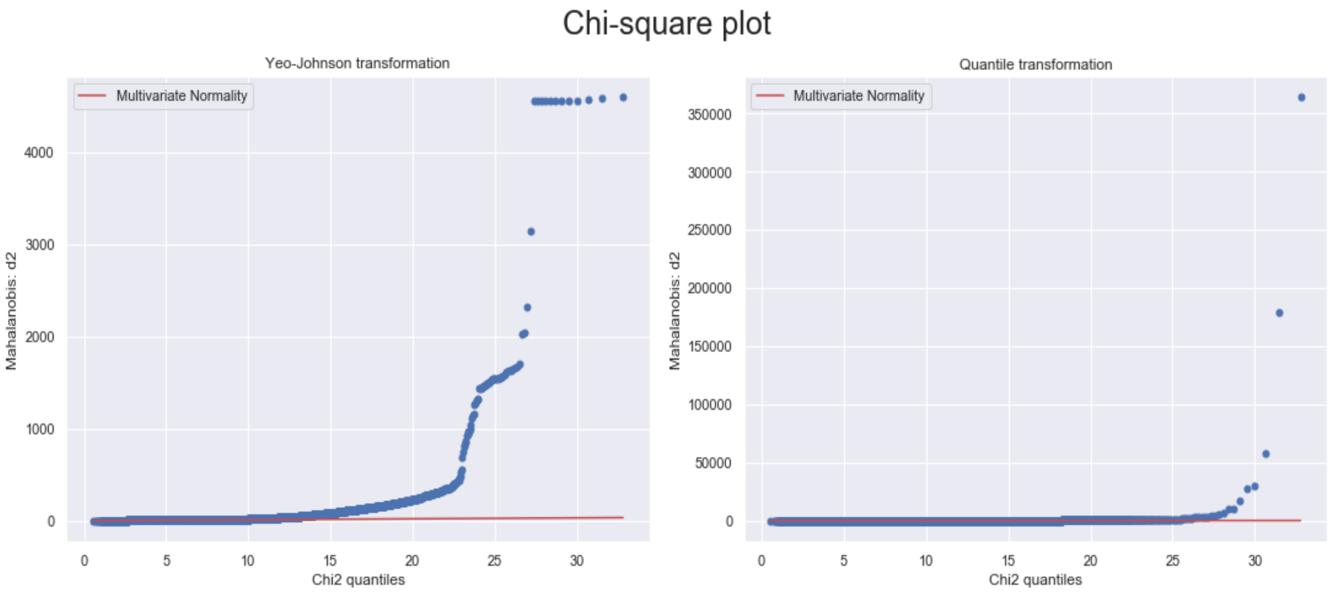
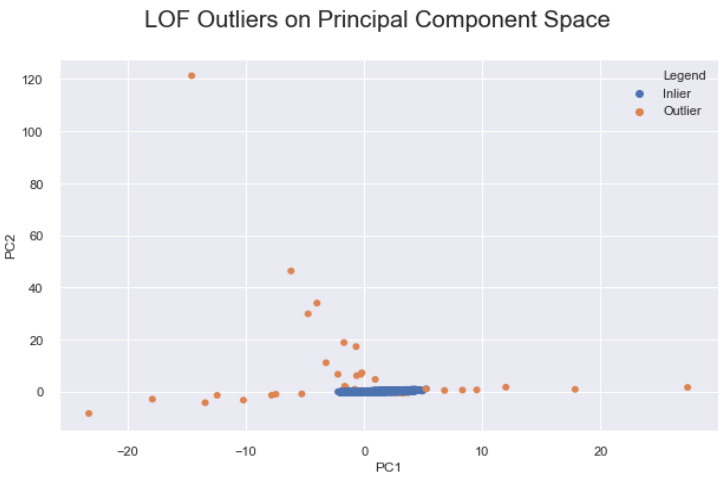
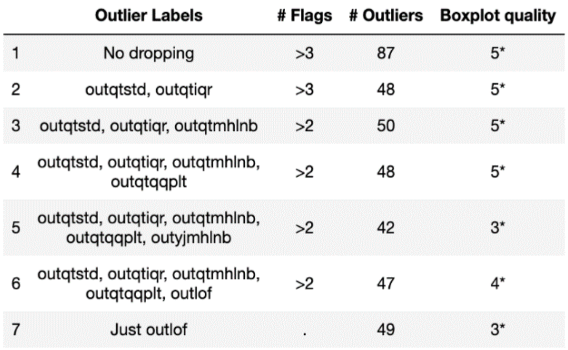


Figure 5 Multivariate outlier detection





	Outlier Labels	# Flags	# Outliers	Boxplot quality
1	No dropping	>3	87	5*
2	outqtstd, outqtiqr	>3	48	5*
3	outqtstd, outqtiqr, outqtmhlnb	>2	50	5*
4	outqtstd, outqtiqr, outqtmhlnb, outqtqqplt	>2	48	5*
5	outqtstd, outqtiqr, outqtmhlnb, outqtqqplt, outyjmhlnb	>2	42	3*
6	outqtstd, outqtiqr, outqtmhlnb, outqtqqplt, outlof	>2	47	4*
7	Just outlof	.	49	3*

Figure 6 PCA space with LOF outliers

Figure 7 Outlier labels combinations

Missing Values

The dataset possesses missing values that need to be dealt with. First, let’s check the dimension of the problem:

From Figure 8 one can see that most of the missing values are concentrated on the variables life and work_comp and that household, claims_r and cmv don't have missing values. Besides, we can also see that at most, rows will have 4 missing values and that most of them contain one missing value. Regarding missing values, it is strange that almost no one has "0" in one or more premiums making this a dataset that almost everybody has almost all products. This dataset is not in accordance with the general knowledge of the area of business, namely, insurance. The only premium column that has zeros is the household and even this column only has 60 individuals with zero, which makes for around 0.6% of the client database, which is an extremely low percentage compared to what would be expected.

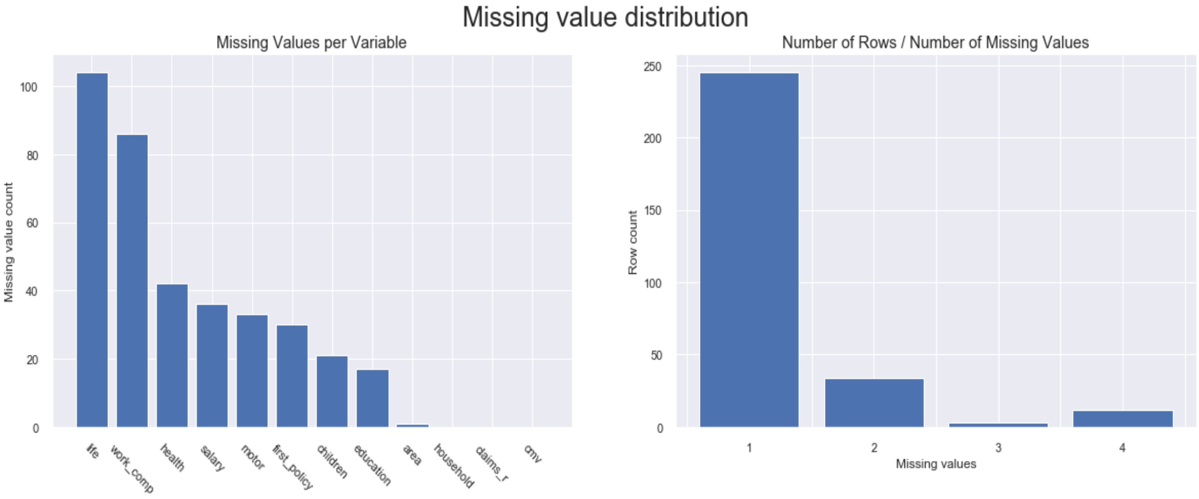
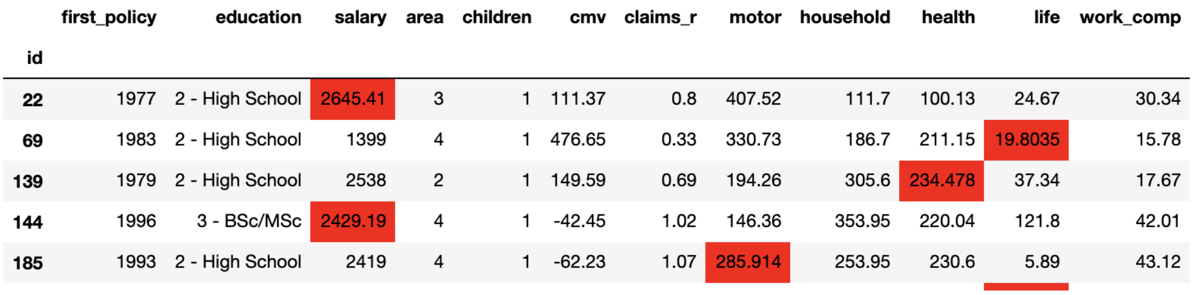


Figure 8 Missing value assessment

Not knowing if the NaNs values in the premiums are "0"s or a proper missing value, and knowing that the database already has errors in other variables (i.e. birthday), we decided to go forward on an approach that still remains valid even after the data quality is improved, by inputting the missing values using the K-Nearest Neighbors method as it is independent of which scenario is the correct one.

Figure 9 shows the imputation of some missing values. We used the K-Nearest Neighbors imputer provided in the scikit-learn package to impute missing values in metric variables using the mean of the neighbors, however the same couldn't be used for non-metric variables. For this reason, we had to resort to the Simple imputer, which replaces the NaNs by the mode of the neighbors. It is also possible to visualize the imputed values above since they are highlighted.



	first_policy	education	salary	area	children	cmv	claims_r	motor	household	health	life	work_comp
id												
22	1977	2 - High School	2645.41	3	1	111.37	0.8	407.52	111.7	100.13	24.67	30.34
69	1983	2 - High School	1399	4	1	476.65	0.33	330.73	186.7	211.15	19.8035	15.78
139	1979	2 - High School	2538	2	1	149.59	0.69	194.26	305.6	234.478	37.34	17.67
144	1996	3 - BSc/MSc	2429.19	4	1	-42.45	1.02	146.36	353.95	220.04	121.8	42.01
185	1993	2 - High School	2419	4	1	-62.23	1.07	285.914	253.95	230.6	5.89	43.12

Figure 9 Missing value imputation

Non-Metric Variables

Variables such as Education, Children and Area are non-metric, thus need to be dealt with separately in order to properly add them to the analysis to their full extent.

From Figure 10 we can compare the mean estimate (points) along with its confidence interval (width) across different levels of area (x axis), children (color) and education (columns) for each metric variable (rows). This allows one to partially see whether the categorical variables impact or not the distribution of the metric variables (i.e. if the average changes for different levels of the categorical variables) and in which way. The concept of this graphic is very similar to the one behind a three-way ANOVA test and can be considered almost as an illustration of this test.

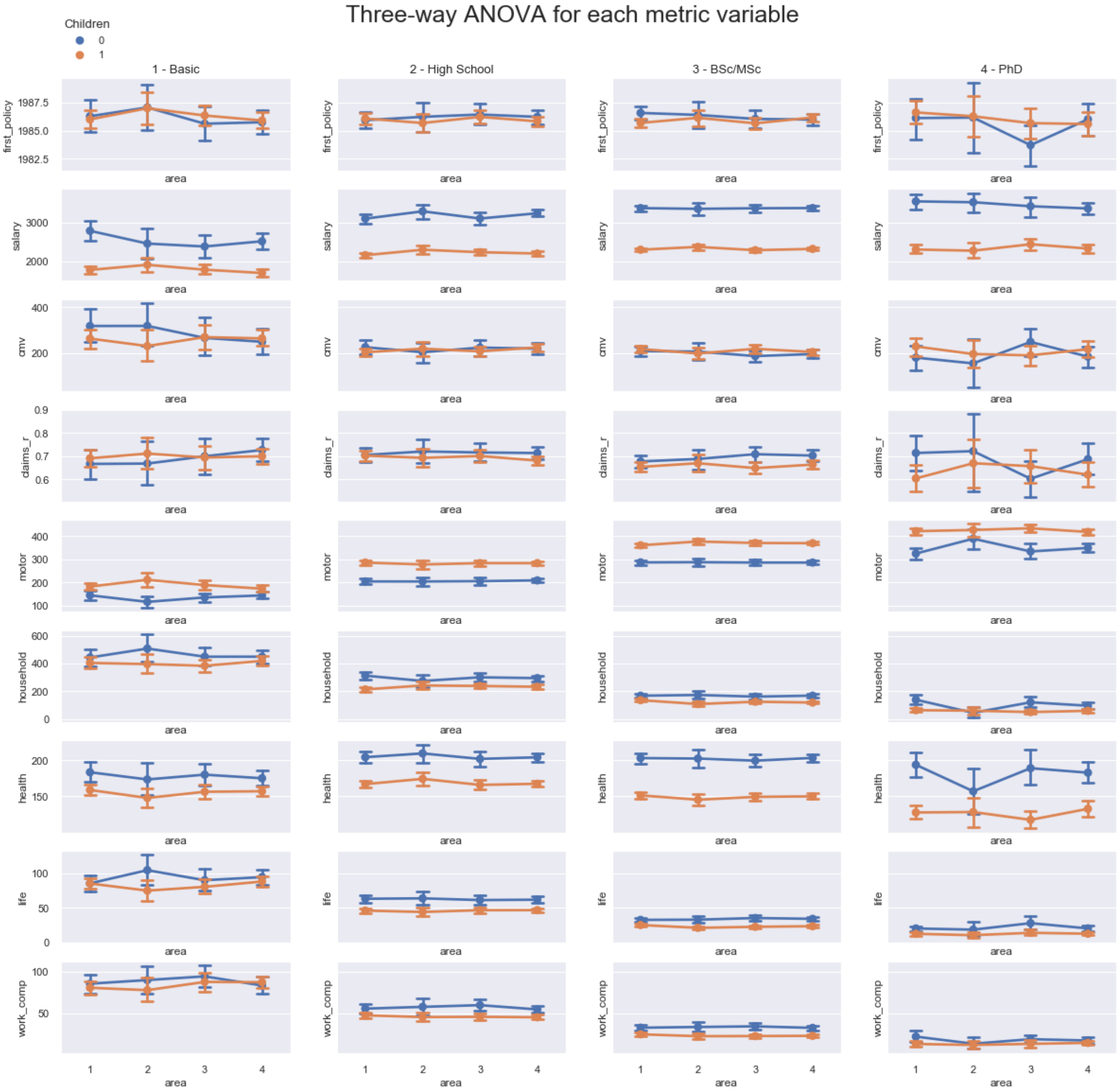


Figure 10 Categorical variable analysis

We can easily observe that customers with **children** have lower salaries, higher motor and lower health. It isn't so noticeable but children also slightly impacts the remaining LOBs, so we concluded that children is, in fact, a meaningful variable that could provide useful information for the cluster analysis. About **education**, we can also see some clear patterns: people with basic education receive fairly less salary than the other groups and the cmv for this group is also higher, relative to the LOBs, the higher the education, the higher the motor will be, even though household, life, and work_comp will decrease. The only LOB that remains sort of constant across education is health. Finally, there doesn't seem to exist significant deviations in the metric variables' distribution according to **area**, thus we decided to exclude this variable from further analysis.

Feature Engineering

This dataset has a very limited number of variables which could severely limit our analysis. In order to improve the comprehension of the data and improve the correlations between variables a few different extra variables were thought off and experimented with the dataset.

Below are the new features we decided to create together with a brief explanation of each:

- **Years as a customer**: Instead of using first_policy we will use years_customer;
- **Annual profit proxy**: This is a proxy variable since it doesn't consider the acquisition cost of the customer (however we think that this should be relatively the same for each customer);
- **Total premiums**: This variable reflects the amount of money each customer spent with the insurance company in the last year (monetary variable);
- **Yearly salary**: Instead of using salary we will use year_salary because other variables are measured yearly;
- **Proportion of salary spent in insurance:** This variable relativizes the amount of money spent by each customer by taking into consideration their salary;
- **Each premium LOB/ Total premiums:** This variable reflects the same information as the previous premium LOBs but as a proportion of the spending of each customer;
- **Canceled**: This variable expresses which customers canceled a contract (i.e. reversals in premiums).

Final Validation

With all the steps above performed a final check is in place to assess if the data preparation process actually resulted in a better performing dataset, ready for modelling. We produced the correlation matrix again and now (Figure 11) we can see that the values have greatly improved and that we can move on to modelling. The boxplots of the prepared dataset (that we can check in the same section of the notebook) reveal that the outliers have in fact been removed.

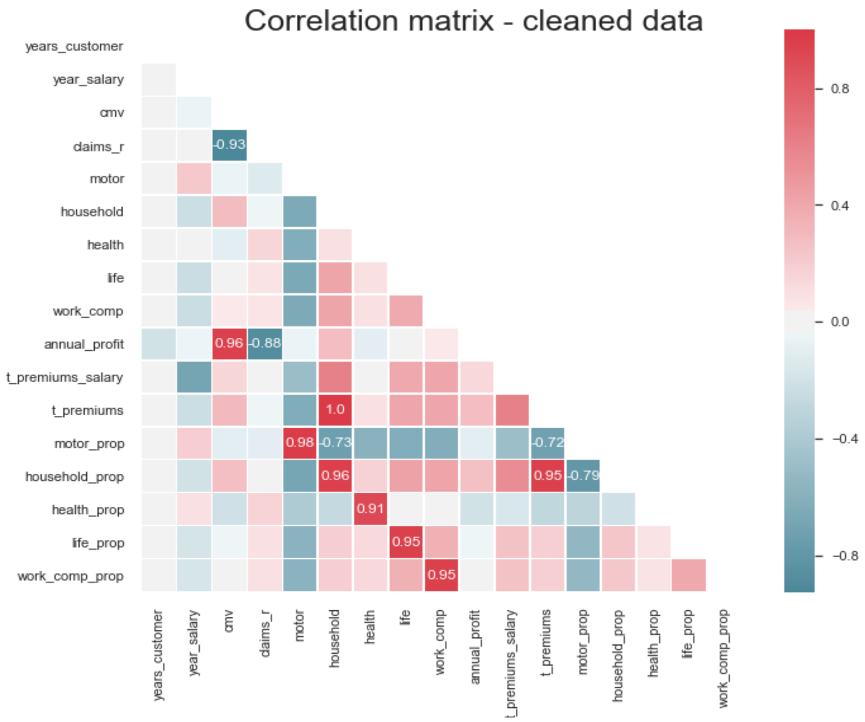


Figure 11 Correlation matrix of cleaned data

We must not forget that the features related to relative proportions of the products and the absolute values of the products should not be mixed as they are direct derivatives from the absolute values and their high correlation might invalidate the analysis.

Modelling

Finally, with the dataset in appropriate conditions, we moved forward to modelling.

We decided to use 3 main models: K-Means for its simplicity and reliability, Gaussian Mixture Modelling (GMM) as opposition to K-Means, being a soft clustering method and Self Organizing Maps (SOM) as a mostly unbiased method to obtain clusters. As an aid for both K-Means and SOM, Hierarchical Clustering (HC) was applied to check if we could obtain better results and to support the number of clusters decision.

We applied the methods to 3 “views” of the data to better assess results and understand the customer vs product segmentations. First, as a generalized method, we assess the complete data. Then we went for a product segmentation, analyzing both absolute product values and relative product values (their proportion relative to the total spent by the customer within the company). We finally went for a combined metric and non-metric data clustering analysis focusing on the customer segmentation, by using K-Prototypes, a method that accepts mixed data types and by attempting a K-Means approach with the non-metric variables transformed through one hot encoding (as a comparison method, as this approach can be reliable sometimes). We also employed some approaches to select the best cluster solution among many, however in the following section, we will just present the selected solutions since these approaches can be consulted in detail in the notebook.

To sum up the cluster profiles we used simple correspondence analysis to assess the association between clusters and the non-metric variables. For the metric variables, we employed the parallel coordinates plot to assess the average behaviour for each cluster.

We decided to define a single set of clusters that could be used to propose marketing strategies. We utilized a simple Hierarchical Clustering to group the cluster from both the Product and Customer segmentations. Finally, we proposed a simple set of rules that can be applied to classify customers into the set of defined clusters by using a Decision Tree.

Results/Discussion

Complete Data Segmentation

In the first view, analyzing the complete data, we used the K-Means algorithm and the GMM as a comparison. The results go as seen in Figure 12.

From the silhouette score results (which show a worse behaviour for the GMM - and can be seen in the notebook) and after comparing the 2 cluster solutions, it is evident that the GMM produces worse solutions since we obtain clusters with more similar average behaviour. For example, in terms of life and work_comp the 3 clusters lines are closer to each other in the GMM than in the K-Means solution. Besides, we can see that cluster 1 in GMM doesn't appear to have any particular behaviour, since the cluster line overs around the mean of each variable.

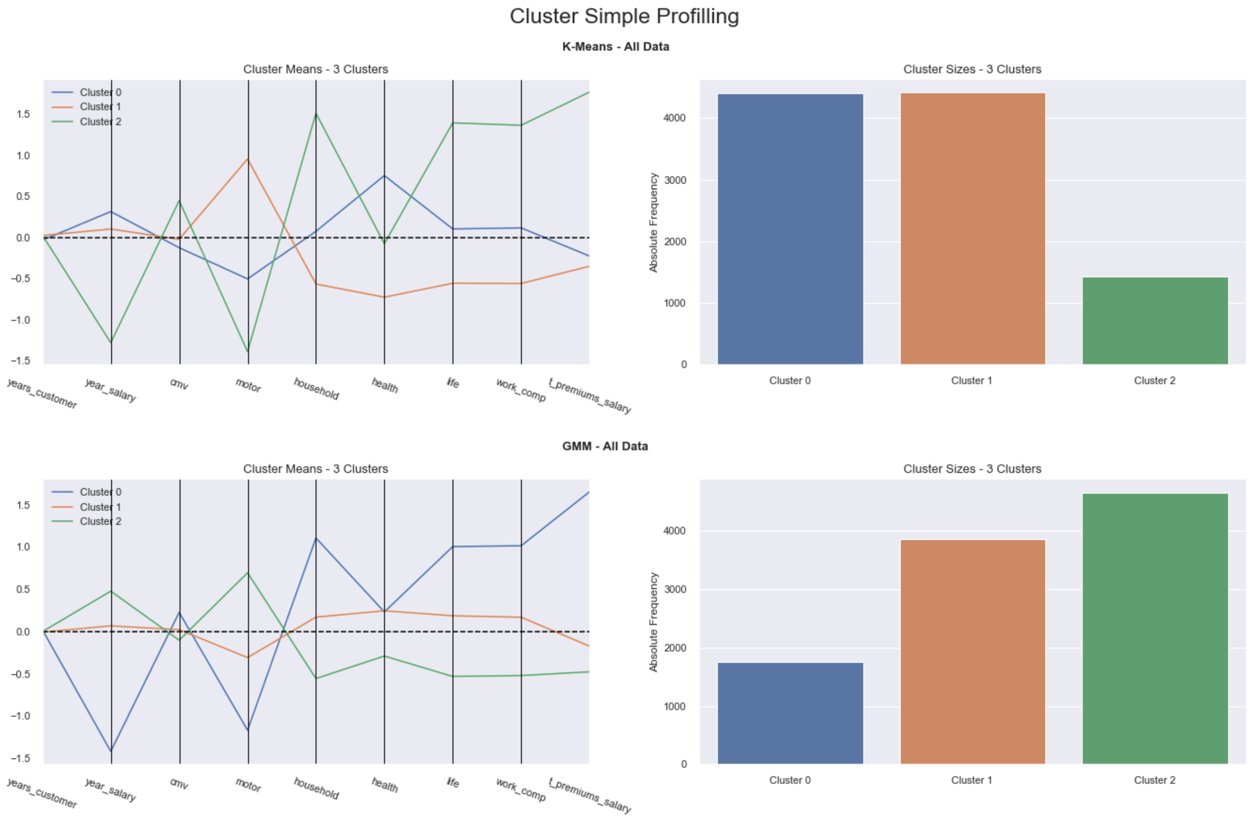


Figure 12 Comparison between K-Means and GMM for complete data

By analyzing the K-means solution, we can see that the variables that provide more discriminant behaviours across the clusters are motor, household and life, suggesting that the customers differ mainly according to the product behaviour. Furthermore, we can see (cluster 2) that there's a clear relationship between the year_salary and the product behaviour: people that earn less, invest mainly in the LOBs household, life and work_comp (and don't spend in motor). Besides, these customers are the ones that spend the highest proportion of their salary in premiums (t_premiums_salary) and the ones that have higher customer monetary value (cmv), making them a very important segment, despite their small size.

We can also identify a segment (cluster 1) of customers that spend a lot in motor and in nothing else. Finally, cluster 0 are the customers that spend the most in health, have the highest salary and are averse to spending money in motor.

Product Segmentation

For product segmentation, we first produced a direct comparison using K-Means between absolute and relative values in order to comprehend which view would be more productive to understand the company’s product. It was revealed that the relative values are more useful as they define a small but significant cluster, which spends a high amount on life and work compensation, in comparison to the other products (Figure 13).

This extra cluster can be attributed to the use of proportions, which points out al cluster of clients that even though it doesn't spend an absurd amount in life and work compensation, the proportion of the total amount spent in these two LOBs is very large. We believe this smaller cluster is important and that the use of proportions is beneficial for the analysis since it provides a purer view of the customers' behaviour regarding the products. Also, the silhouette score for this clustering is higher than the one using the absolute data. As a final decision, we opted to perform the Product Segmentation based on the proportion data.

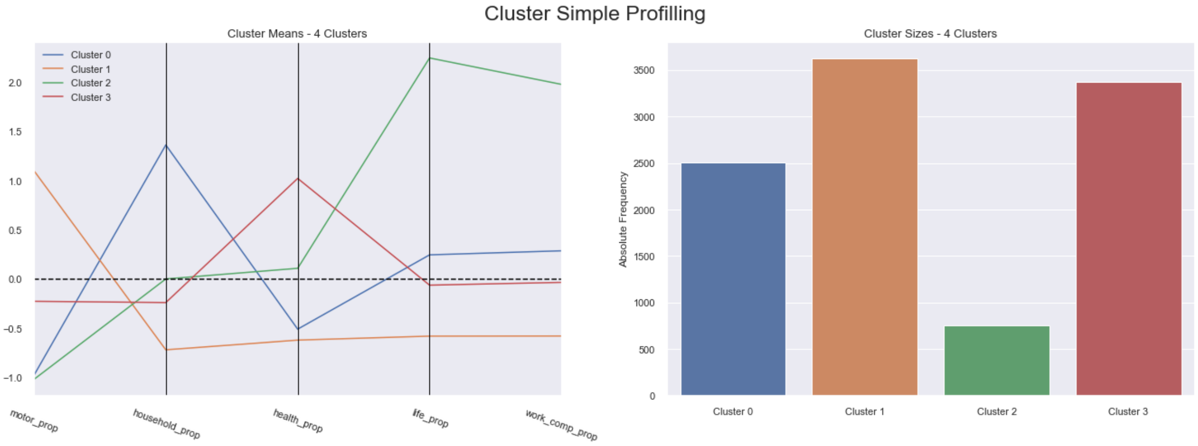


Figure 13 Simple profiling of relative product K-Means

Moving forward with using the relative proportions of the products, we then performed the full protocol (GMM, K-Means + HC, SOM + HC, SOM + K-Means) to this “view”.

In order to compare the clustering approaches, we used the silhouette score as well as some other quality measures and the interpretability of the clusters. The best results came from SOM+ K-Means and can be visualized in Figures 14 and 15.

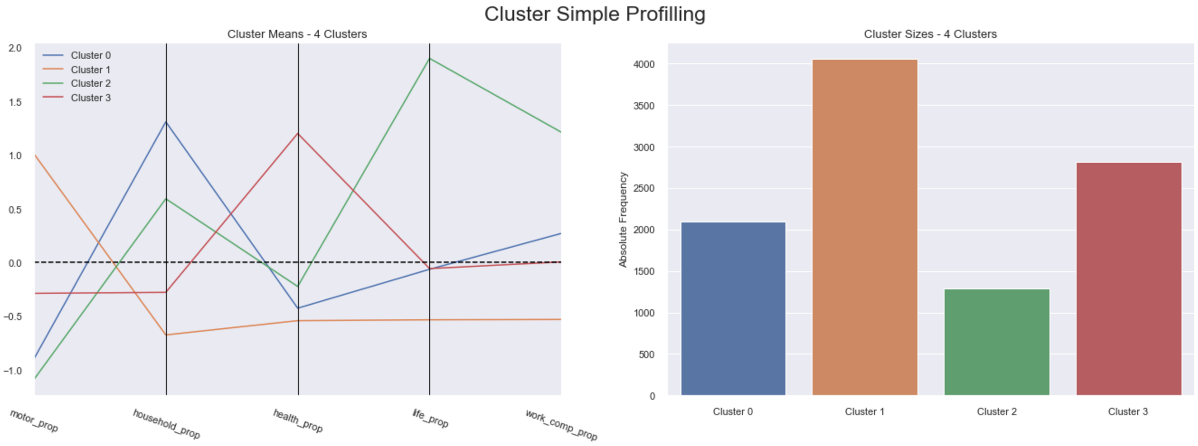


Figure 14 Simple profiling of relative product SOM + K-Means

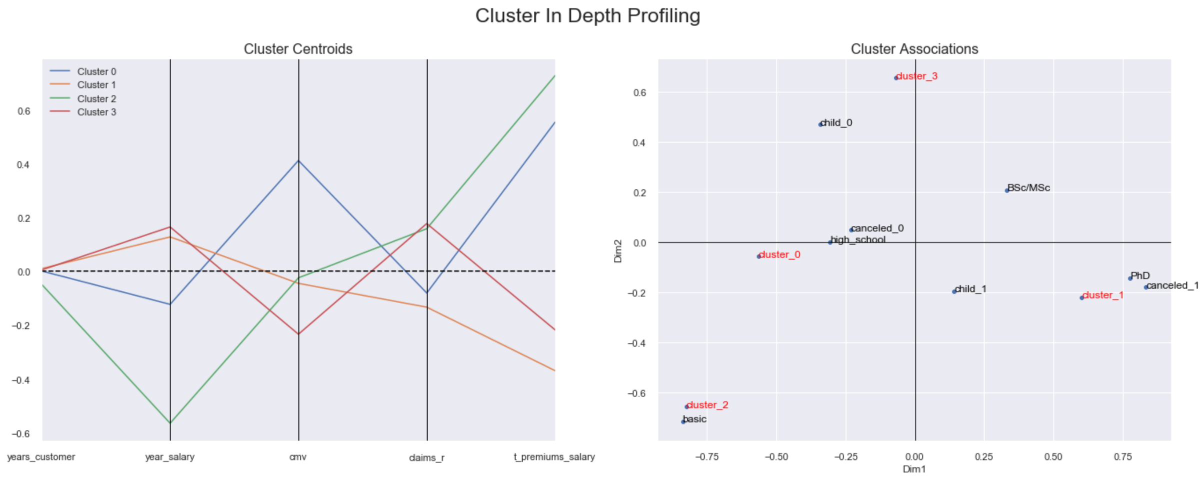


Figure 15 In depth profiling of relative product SOM + K-Means

From the results, we characterize and attribute a label to each cluster. The Correspondence Analysis perceptual map (Cluster Associations) can be interpreted by looking at the angles between vectors of each modality. If the angle is acute as for example in cluster_1 and basic, then there's a positive association. An obtuse angle indicates a negative association and a right angle indicates no association.

Cluster_0 - High Household and Low Motor Spenders: The main characteristics of this cluster is their high spend rate on household policies and their low spend rate on motor policies. They reveal the highest customer monetary value (cmv) and the second highest proportion of salary spent on the company’s products. They seem associated with non-cancelation and with a high school level of education. Their relative size is around 2000 customers;

Cluster_1 - High Motor Spenders: This cluster’s main characteristic is that they don't buy anything else, but motor insurance and they are the biggest spenders on it. If we analyze the other variables, we discover that their salary is the second highest on average, have the lowest claims rate and are the clients that spend the least proportion of their salary in the company. From the correspondence analysis we see a high association rate of this cluster with PhD level of education and a high level of cancellations. This is the biggest segment with around 4000 clients;

Cluster_2 - High Life and Work Compensation Spenders: The main characteristics of this cluster is their high spend rate on life and work compensation and they are the lowest spenders on motor policies. Further analysis reveals that they receive the lowest average salary and their claims rate is fairly high. They are the ones spending the highest proportion of their salary in the company. Regarding nonmetric variables they reveal a high association with basic education. This is the smallest cluster with roughly 1300 clients;

Cluster_3 - High Health Spenders: The main characteristics of this cluster is their high spend rate on health policies. They receive the highest average salary, reveal the lowest customer monetary value (cmv), from the highest claims rate coupled with the second lowest percentage of income spent in the company. They seem to be associated with being childless, being a segment of roughly 2800 customers.

Customer Segmentation

For the combined metric and non-metric data clustering analysis, significant of a customer-based segmentation process, we performed a K-prototypes based approach to analyze both types of data combined:

From Figures 16 and 17, we can extract some relevant information regarding each cluster. Following, we characterize and attribute a label to each cluster from the customer segmentation:

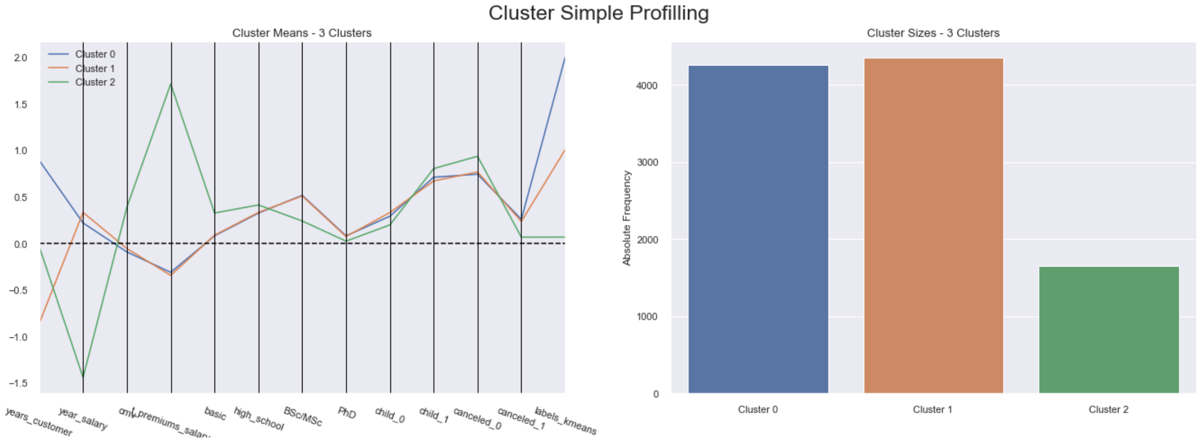


Figure 16 Simple profiling of customer K-Prototypes

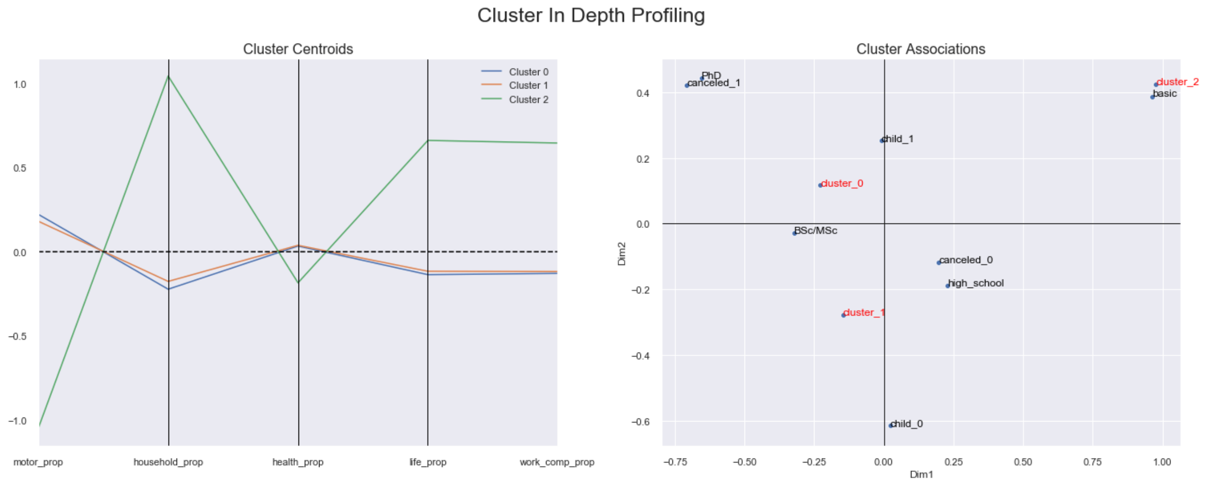


Figure 17 In depth profiling of customer K-Prototypes

Cluster_0 - Old Customers: The main characteristic of this cluster is their high years_customer, which indicates that they are with the company for many years. Besides, they reveal an average behaviour in every product variable and a high positive association with higher education and cancelation rate. Their relative size is around 4000 customers;

Cluster_1 - New Customers: This cluster's main characteristic is their low years_customer, indicating that they are relatively recent customers. This segment is very similar in terms of product consumption to the Old Customers cluster, having an average behaviour in every LOB. Besides, they also have a slightly higher income than the Old Customers cluster. From the correspondence analysis, we see a high association with having no children. This is the biggest segment with around 4500 clients;

Cluster_2 - Valuable Customers: The main characteristics of this cluster is their low yearly salary and a high proportion of salary spent in the company. This makes them the cluster with the highest customer monetary value. Further analysis reveals that they don't invest in motor and spend their money mainly on household, life and work_comp. Regarding the categorical variables, they are positively associated with basic education. This is the smallest cluster with roughly 1700 clients.

Cluster Concatenation

After these 3 segmentations, we wanted to define a single set of clusters that we could use to provide the marketing strategies. For this task, we used a simple Hierarchical Clustering. The results are shown in Figures 18 and 19. Following, we characterize and attribute a label to each cluster:

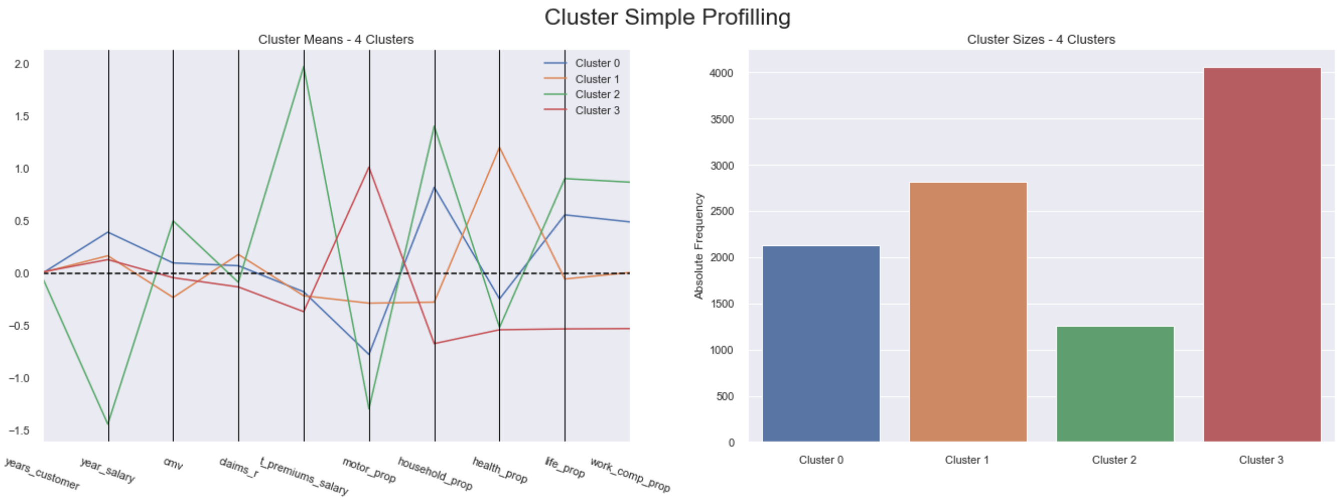


Figure 18 Simple profiling of concatenation HC

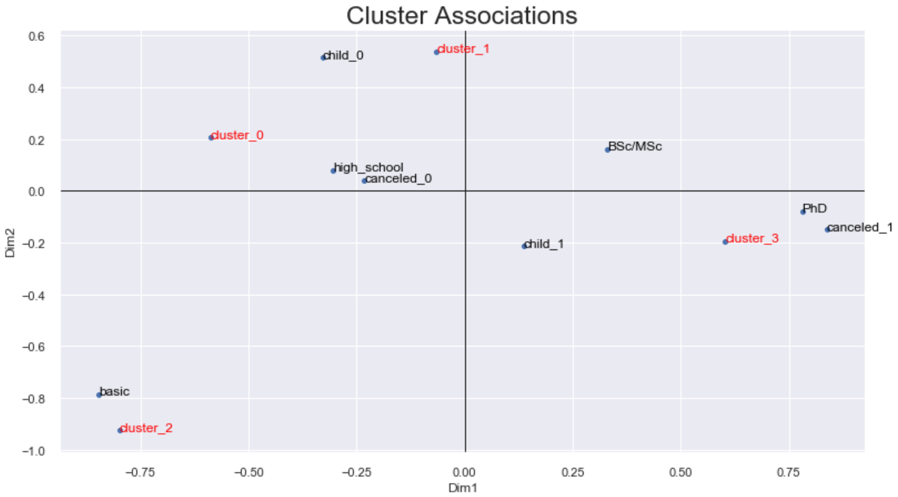


Figure 19 In depth profiling of concatenation HC

Cluster_0 - Diverse and Potential: This cluster is characterized by its higher consumption of household, life and work_comp, while having a low consumption of motor. They are also the cluster with the highest income and have an above average customer monetary value. From the correspondence analysis, we conclude that these customers are negatively associated with cancelation and higher levels of education. This cluster is composed of around 2000 customers;

Cluster_1 - Health: This cluster's main characteristic is the high proportion of spending on health. The remaining behaviours don't stand out much, however, they are the cluster with the lowest customer monetary value and the highest claims rate. They also seem to be positively associated with not having children. Finally, this cluster is composed of roughly 2800 customers.

Cluster_2 - Diverse and Valuable: This is the smallest cluster with around 1250 customers. Nevertheless, they are the most important in terms of customer monetary value with an average of around 0.5 standard deviations above the mean. They are also the customers that earn less on average, however, they are the ones that spend the highest proportion of their salary in the company. Besides, they are characterized by having an extremely low proportion of spending in motor and an extremely high in household. Also, they have a fairly high proportion of spending on life and work compensation. It can be said that this cluster has a similar product consumption behaviour to Cluster_0. Finally, they are highly positively associated with a basic level of education.

Cluster_3 - Motor: The main characteristic of this cluster is the high proportion of spending in motor and low proportion of spending in every single other LOB. Besides they are the customer with the lowest claims rate and proportion of salary spent in the company. They are the biggest cluster with roughly 4000 customers, and they are positively associated with high levels of education and cancelation of contracts.

Decision Tree Classifier

We built a decision tree classifier to classify the outlier observations according to the final concatenated cluster. This classifier also provides a useful tool for the future, as a set of rules can be extracted that helps understand why customers belong to a certain cluster. With this information, we can easily classify future customers by making simple queries to the database. This will allow the application of the marketing strategies to the company customers in a dynamic way.

It turns out that with a simple decision tree with a depth of 3 levels, we can predict the test observations with an accuracy of around 86%. The decision tree can be visualized in the notebook at the Classification of Outliers section. Below we can see what the rules are to classify customers into each cluster:

- **Diverse and Potential**: If motor_prop ≤ 0.49 and health_prop ≤ 0.28 and t_premiums_salary ≤ 0.05, then the customer belongs to this cluster with a 73% probability;

- **Diverse and Valuable**: If motor_prop ≤ 0.49 and health_prop ≤ 0.28 and t_premiums_salary ≥ 0.05, then the customer belongs to this cluster with a 94% probability;
- **Health**: motor_prop ≤ 0.49 and health_prop ≥ 0.28 and motor_prop ≥ 0.2, then the customer belongs to this cluster with a 94% probability;
- **Motor**: motor_prop ≥ 0.49 and health_prop ≤ 0.3 and health_prop ≤ 0.27, then the customer belongs to this cluster with a 99% probability.

Marketing Recommendations and Next Steps

From the results discussed above, we decided to move forward to create a marketing strategy relevant to the final results (i.e. the concatenated clusters)**.** The following suggestions will be useful to tackle these segments from a marketing perspective.

The **Cluster_0 - Diverse and Potential** seems to be the cluster with the highest average salary so their potential to become more profitable customers is in place. This leads us to think that developing accessory products for their biggest expenditures (household, life and workers' compensation) is a good strategy. It's crucial to keep them satisfied with premium care as their cancellation rate is low;

The **Cluster_1 - Health** are the highest risk customers, as their claims rate is the highest one. As they have a high average salary and the second lowest proportion of income spent on the company there is a margin to increase their cmv, as it is the lowest. This low cmv reveals that the health policies of the company need to be readjusted as they do not bring a lot of return to the company. Thus, a redefinition of these policies to better serve these customers (lowering the claims rate) and increase the return on investment for the company is crucial.

The **Cluster_2 - Diverse and Valuable** are the best clients of the company, as they have the highest cmv. They should be well accompanied by the company by providing them with premium valued customer care related to their highest preoccupation, their household. The company is in a good place to develop accessory products to this policy such as security products for the household, for example.

Finally, the **Cluster_3 - Motor**: they seem to be a wealthy and low risk group (revealing the lowest claims rate) so promoting combined products at a discount can be a good approach with this group. Also, their high salary and interest in motor vehicles can be used to create premium versions of the motor policies, that can become of interest for this segment, characterized by a high association with PhD education (premium products can become a way to decrease their cancellations) A further analysis to understand the high rate of cancellations in this group is paramount.

Besides the Marketing Recommendations it is important to assess the problems of the dataset that should be dealt with and important Next Steps in this market segmentation in order to improve further analysis:

- The dataset reveals highly problematic variables such as birthday which demonstrates a high rate of errors. Such variables need to be evaluated in depth on the origin of these errors and measures should be put in place to significantly decrease them;
- The dataset only possesses customers that seem to buy all the company's products, which doesn't seem to correlate with the Portuguese reality. A check on if this sample is representative of the company should be in place;
- More information about how the variables were collected by the company is important in order to provide the data science team with valuable insights about the error rates of each variable;
- More information about the living area is important to assess the relationship between the living area and the customer's behaviour.

With these insights is paramount that improving data quality and information should be prioritized before any next steps regarding customer segmentation should be attempted.

Conclusion

In this work, a full end-to-end data mining project was performed. The data wasn't provided in the best condition and therefore the first thing we had to do was to prepare the data for clustering. Coherence checks, outlier detection and missing value imputation were performed to make the data usable. In total, 48 observations were removed due to being outliers and around 400 missing values were imputed. Before proceeding to the cluster analysis, we also analyzed the importance of the non-metric variables as these are an obstacle to the common clustering techniques. We ended up excluding the area categorical variable, however, we also created 11 new features to help us discriminate between clusters.

We divided the clustering process into 3 segments: Overall, Customer and Product. Our intention was to use each segment to understand a specific customer behaviour. In these segments we utilize a wide range of techniques such as K-Means, SOM, HC, K-Protypes and GMM. Often, we used combinations of these techniques as well. To analyze each cluster, (this kind of analysis can be consulted in the notebook) we used methods such as Silhouette analysis, K-elbow plots, R^2 plots, Dendrograms, Correspondence Analysis, etc.

In order to reduce the number of clusters to consider, we proposed a final set of clusters by grouping the ones we had from Customer and Product segmentation with HC. We also employed a Decision Tree classifier to predict to which cluster the outliers belong, as well as to extract the set of rules that helps us understand why each customer belongs to each cluster.

Finally, we propose a set of possible marketing strategies to each identified cluster that can be utilized by the marketing department of the company to develop more complex and personalized marketing approaches.