Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Segmentation of An Insurance Company Client Dataset

Group 8

Beatriz Gonçalves, number: 20210695
Diogo Hipólito, number: 20210633
Diogo Pereira, number: 20210657

01, 2022

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

INDEX

1. Introduction

The client, an insurance company, wishes to better understand the scope of its clients, in order to better serve them and increase their return on investment. The given ABT (Analytic Based Table) consists of 10.290 customers and the given task involves analysing the table for evident groups of clusters, extracting the behaviour of those clusters, and providing insights for the Marketing Department to better understand all the different Customers' Profiles.

The project can be found in a GitHub repository which can be accessed through the following link: https://github.com/beatrizctgoncalves/project_dm. The repository contains a Jupyter Notebook with all the relevant analyses. Note that all decisions made in this process are justified in the notebook with theoretical references, appended to the relevant code section that utilizes these references.

2. Data Understanding

The company's ABT from 2016 has 10 296 observations and 14 variables that are described in the following table (Table 1). We should note that in the premium variables we can have negative values that manifest reversals occurred in the current year (2016), paid in the previous one(s). This means that the clients with negative values cancelled the respective insurance.

Variable	Type	Description
CustID	float64	Customer ID
FirstPolYear	float64	First year as a customer
BirthYear	float64	Customer's Birthday Year
EducDeg	object	Customer's Education Level (1-Basic, 2-High school, 3-BSc/Msc, 4-PhD)
MonthSal	float64	Gross Monthly Salary (€)
GeoLivArea	float64	Categorical variable from 1 to 4 that identifies the living area (there is no information about the meaning of the area codes)
Children	float64	Binary variable that tells if the customer has children (1) or not (0)
CustMonVal	float64	Customer Monetary Value/Lifetime value = (annual profit from the customer) x (number of years that they are a customer) - (acquisition cost)
ClaimsRate	float64	Amount paid by the insurance company (€)/ Premiums (€) (Note: in the last 2 years)
PremMotor	float64	Annual premiums in Motor (€)
PremHousehold	float64	Annual premiums in Household (€)
PremHealth	float64	Annual premiums in Health (€)
PremLife	float64	Annual premiums in Life (€)
PremWork	float64	Annual premiums in Work Compensations (€)

Table 1 - Variable Description

Initially, we found that there are non-metric and metric variables, which requires that these variables be treated separately. The non-metric variables are *EducDeg*, *GeoLivArea* and *Children*, and the remaining variables are metrics.

The next step was to understand the relations between variables, we computed the correlations between them using the Pearson method. The correlation matrix, as can be seen in Figure 1, show us which variables have the most potential for future modelling.

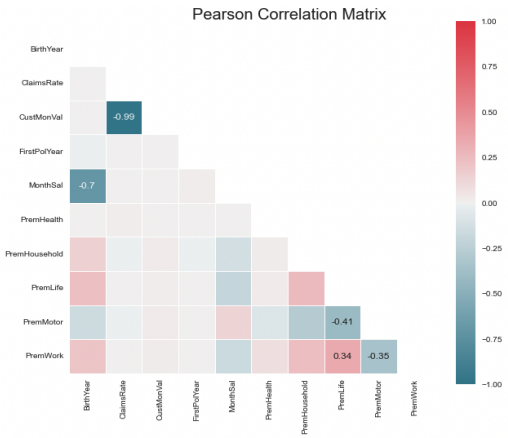

Figure 1 – Metric Correlation Matrix

After analysing it, we ended up noticing that there is a high correlation between *ClaimsRate* and *CustMonVal*, in fact it is almost a perfectly negative correlation, which would mean that with one variable we can predict the value of the other. However, looking at the problem description, it does not initially look like one variable was created using the other. According to the description, *CustMonVal = Total profit from customer * number of years as customer*, and *ClaimsRate = Amount paid by insurance / Premiums*. In the future we will decide if we keep both columns.

Other than that, there isn't any meaningful correlation between any of the variables, which is not a good start for our model. We will have to go through a deep data preparation process in order to have a workable dataset.

3. Data Preparation

Data preparation is a crucial first task in this project since the team has come across various issues with the variables in the dataset, such as missing values to different scales among the variables and outliers. Consequently, these can have a negative impact on the latter analysis of the clusters, therefore they need to be rectified. First, we'll begin by checking the coherence of our dataset, then identifying and removing outliers, and the last filling in the missing values.

3.1. Coherence Checking

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 with standard questions of the area.

The second rule is about the *FirstPolYear*. It does not make sense for the *FirstPolYear* to be smaller than 1143, since it is the year that Portugal was founded and we cannot have a record of something that did not happen yet, so it cannot be bigger than 2016. After applying this rule, we found out that there is also only one observation that does not obey it. This is clearly not right so we decided to delete this row.

The third rule is also about *FirstPolYear*. Since this variable represents the first contact of the client with the insurance firm, it does not make sense that this contact happened before the person was born - *FirstPolYear* cannot be smaller than *BirthYear*. After applying this rule, we found that 1997 observations did not comply with it. Since this is a huge number, we could not delete all the incoherent

rows. So, as *FirstPolYear* was calculated by the company and *BirthYear* was probably submitted by the customers, we decided that it is more likely that the customers gave wrong information. Assuming this, it would mean that 1997 customers filled the forms wrongly or that these customers inherited all this data from their parents. However, as we cannot be sure about that, we decided to delete the *BirthYear* column.

3.2. Identifying and Removing Outliers

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

Firstly, taking into consideration the Metric Variable's Box Plot on Figure 2, we concluded that most of the variables have outliers. Therefore, we will go step by step throughout the variables, analysing the histograms and boxplots for each one that raises a red flag.

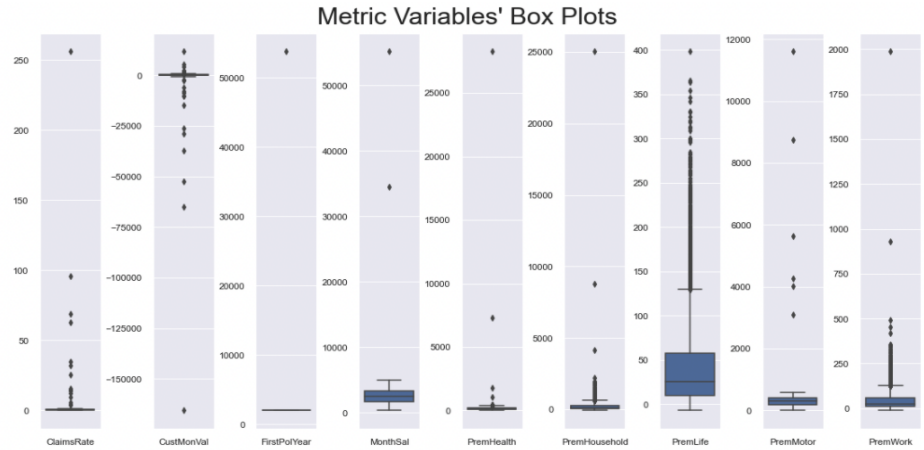

Figure 2 – Metric Variables' Box Plot

We first choose a simple approach, by deleting the values that were extremely out of context just by looking into Figure 2. For that, we dropped every row with the *ClaimRate* above 4, the rows with *CustMonVal* bellow -2000, *FirstPolYear* higher than 2017 and values of *MonthSal*, *PremHealth*, *PremHousehold*, *PremMotor* and *PremWork* above 30000, 5000, 4000, 2000 and 750 respectively. With this we conserved around 99,71% of the original data.

With this, we noticed that we still had dubious data, so we decided to try to filter even more using the IQR method for outliers' detection. Where we get the 0.25 quartile and the 0.75 quartile so that we could get the Interquartile Range (IQR), then, we defined a threshold and with this we defined a margin (*Threshold * IQR*) and for each variable every value that is higher than 0.75 quartile plus the margin or lower than 0.25 quartile minus the margin are considered outliers. With this method we kept 83,43% of the data from previous detection. With this method we conserved around 83,18% of the original data.

After applied de IQR method, we also did outlier removal using the Z-Score method which conserved around 93,31% of the original data.

Finally, we combined different outlier methods and obtained a percentage of data kept of 93,31% which is lower than the data kept after removing outliers manually whereby we choose to use only the manual filtering version.

3.3. Filling in the Missing Values

Most of the missing values are on the *PremLife* and *PremWork* variables as can be observed in Table 2. Therefore, we developed a bar plot of Premium variables as shown in Figure 3, in order to see how many zero values were on each of the premium variables.

Variable	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 of each Variable

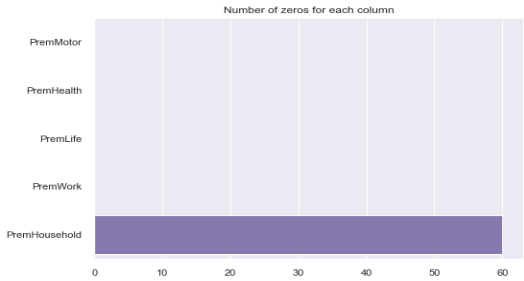

Figure 3 – Premium Variables' Bar Plot

The only Premium variable with zero values is *PremHousehold*, with 60 total zero values. Therefore, we considered two options. The first one was that zeros represent missing values. However, if so, we do not know why they were only present in *PremHousehold*. The second one was that missing values mean zero. Nevertheless, we could not justify why this was not applied in *PremHousehold* then.

The fact that we have already encountered errors previously, for instance the *BirthYear* variable, made us doubt about the integrity of the dataset and given all these factors, we decided not to replace the missing values with 0. Instead, we used an independent approach: imputing the missing values with the mean of its neighbours using KNNImputer. Note that for categorical features, we cannot use the mean as the imputer metric. Therefore, we used Simple Imputer for categorical variables, which replaced missing values with the mode of each variable.

Subsequently, we merged both data frames and concluded that the imputers were successful, since there are no missing values remaining in our data frame as can be seen in Figure 4.

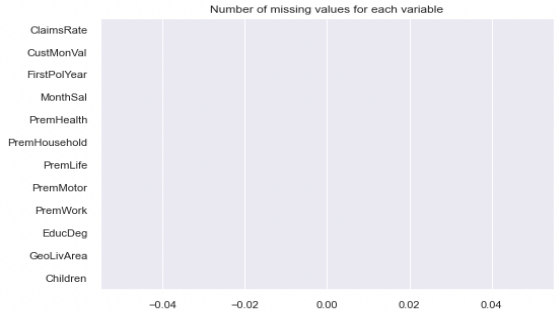

Figure 4 – Number of missing values by variable

3.4. Non-Metric Variables

As said before, *EducDeg*, *Children* and *GeoLivArea* are non-metric variables, thus need to be dealt with separately in order to properly add them to the analysis to their full extent. For that, we built the graphics present on Figure 5 composed by the mean estimate (points) along with its confidence interval (width) across different levels of area (x axis), children (colour) and education (columns) for each metric variable (rows). This allowed us to see whether the categorical variables impact or not the distribution of the metric variables and in which way.

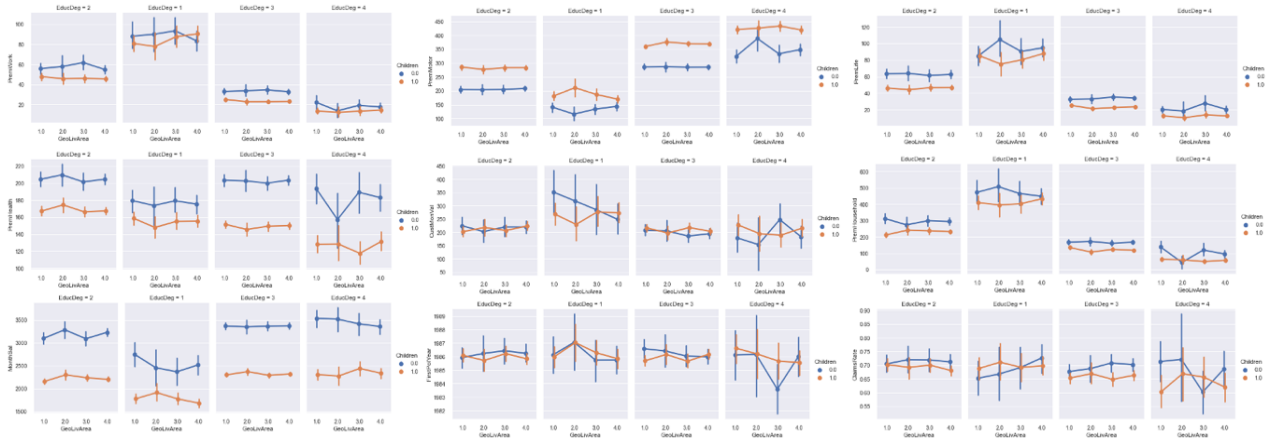

Figure 5 – Categorical variable analysis

First of all, we can conclude that *Children* is, in fact, a meaningful variable that could provide useful information for the cluster analysis since customers with children have less salary and minor but still noticeable effect on all the different premiums, that is higher motor and lower everything else.

Regarding *EducDeg*, we can see a clear impact on salary, particularly between basic education and all other kinds of education, this means that people with more education receive a higher salary. Also, there is a visible difference in premium costs since more education relates to a higher *PremMotor* and lower *PremLife* and *PremWork*.

Finally, we did not see any significant impact on any of the metric variables according to *GeoLivArea*, as such we decided to drop this variable.

4. Data Pre-processing

4.1. Feature Engineering

In order to try to gain explainability and partitioning power for a better and more precise customer segmentation, we decided to create new variables that are basically transformations of variables we already had in the data frame.

Firstly, we converted *MonthlySal* to a new variable, *YearlySalary*, since the premium costs are represented as yearly, we considered it would be better to use yearly instead of monthly salary for consistency's sake, this being *YearlySal = MonthSal * 12*.

Secondly, *FirstPolYear* was converted into a new variable named *ClientYears* because it measures the number of years since the first policy and this transformation makes the data simpler and easier to analyze. In other words, *ClientYears = 2016 - FirstPolYear*.

The next step was to create a variable named *TotalPremiums* that sums of all non-cancelled premiums categories, in order to know how much money each customer spent in our company in 2016. Using these last two new variables, we created the *PremiumsRate*, which measures the proportion of the salary spent in premiums divided by yearly salary in our company. This may be a good measure of a client's commitment to the company as it measures the effort of each customer to be a client.

Then, we created Premium Proportion for all premiums. These variables express how much a customer spent in one premium relative to the total spent (e.g. *PremWorkProp = PremWork/TotalPremiums*), if *PremWork* is less than 0, its value will be set 0.

Lastly, we created a variable called Cancelled that represents which customers have cancelled an insurance contract.

4.2. Correlations

At this phase, we find it important to go back to the correlations (Figure 6) and check how the new variables are related to each other and to the original ones, as we should not use high correlated variables when performing clustering analysis because it can inflate the importance of some variables, leading to wrong segmentation definitions.

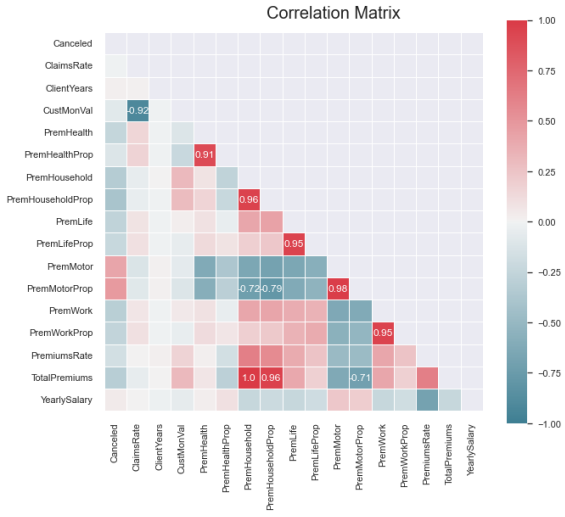

Figure 6 – Correlation Matrix of cleaned data

By looking to the Correlation matrix, we can see that many variables present perfect or almost perfect correlation, such as *CustMonVal* with *ClaimsRate* (-0.92), the newly created variables of proportion with their corresponding variable, as it was be expected and we have *PremMotorProp* and *TotalPremiums* with very high correlations with three other variables.

5. Clustering Algorithms

We decided to use three main algorithms: K-Means for its clarity and trustworthiness, this is a clustering algorithm that identifies k centroids and allocates each observation to the centroid closer to it; K-Prototype which is a clustering method based on partitioning, its algorithm is an improvement of

the K-Means and K-Mode clustering algorithm to handle clustering with the mixed data types; Hierarchical Clustering to check if we could obtain better results and to support the number of clusters decision; and Self Organizing Maps (SOM) as a mostly unbiased method to obtain clusters.

5.1. Product Segmentation

For the Product Segmentation, we wanted to study how customers were grouped according to what they bought. For this view, we used K-Means to compare between total and proportion values, to understand the company's product. As can be seem in Figure 7, there are 4 clusters. The 4-cluster implementation adds a small but distinguishable cluster, with high particularly Life and Work premium values. Although it represents a small number of customers, it is a very valuable and unique set of customers. Due to this, we selected the 4-cluster implementation as the best.

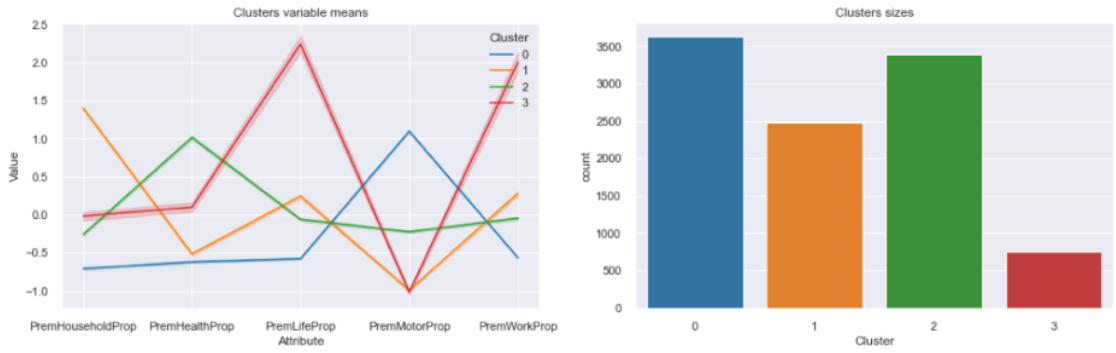

Figure 7 – Profiling of proportion values with K-Means

Subsequently, we decided to utilize the K-means clustering with proportional values with k=4, since in the total values cluster profiling, we didn't see the small set of customers that values Life and Work premiums highly. Also, there seem to be two major clusters whose values didn't really vary much between them, unlike what we saw in the proportion values option.

Consecutively, we decided to use Self-Organizing Map algorithm combined with Hierarchical Clustering since the first provides a powerful clustering algorithm which needs to be parameterized, namely when it comes to defining the grid of units that will cover the feature space. On the other hand, Hierarchical Clustering is a simple algorithm which does not need to be parameterized.

The result is present in the Figure below (Figure 8) and we can see that each cluster values more some premium then the others. For cluster 0 we have clients who use more the *PremHealth* with higher values in that feature. In turn, in cluster 1 we have clients who use more the *PremLife* and *PremWork*. On the other hand, for cluster 2 we have clients that use more the *PremHousehold*. Finally, in cluster 3 we have clients that use more the *PremMotor*.

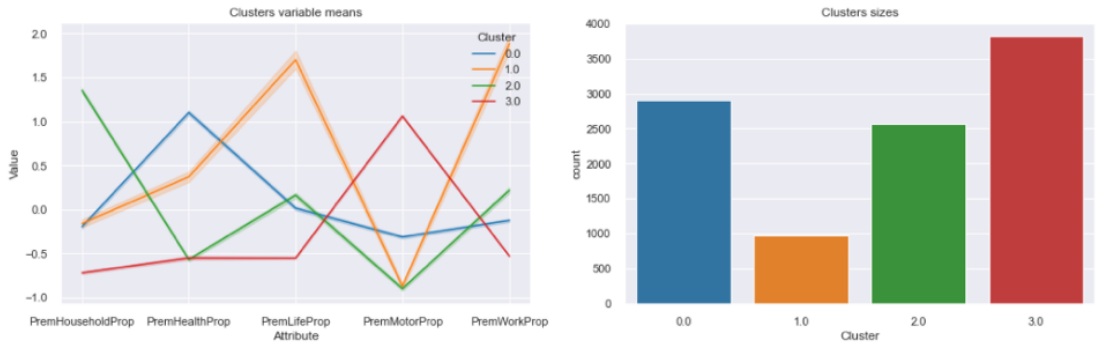

Figure 8 – Profiling of proportion product SOM + HC

Then, we performed another combination (SOM + K-Means) to this view. To compare the clustering approaches, we used the Elbow method, since it finds the average distance of each point in a cluster to its centroid and represents it in a plot, and the Silhouette method which computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. Finally, we concluded that the best results came from SOM+ K-Means and can be visualized in Figures 9 and 10.

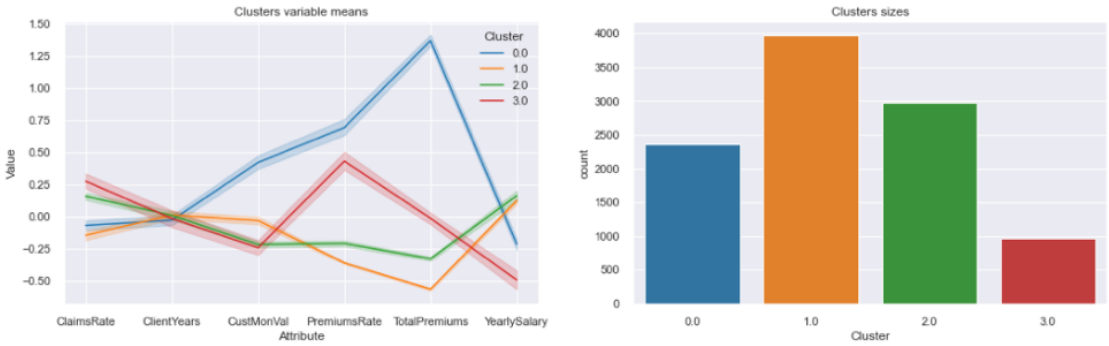

Figure 9 – Profiling of proportion product SOM + K-Means

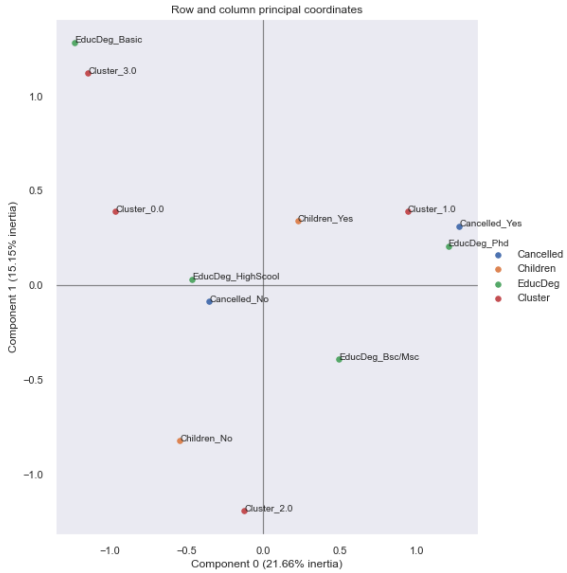

Figure 10 – Depth profiling of proportion product SOM + K-Means

From the results, we identified and categorized each cluster. The clusters are described as follows:

Cluster 0 - High Household and Low Motor Spenders: Spend the most on premiums and the highest percentage of salary on premiums, and have the highest customer value, however, they do not have the highest yearly salaries.

Cluster 1 - High Motor Spenders: Spend the least on total premiums, but the most on motor. Largest group of users, ~38% of the customers.

Cluster 2 - High Health Spenders: Spend the most on health and around the mean on all other premiums. Contrary to Cluster 1, these customers care about the other premiums. They also have the highest yearly salaries. Second largest group of customers.

Cluster 3 - High Life and High Work Spenders: These customers spend the most on life and work premiums and seemingly don't care about motor premiums. These customers have the lowest yearly salaries but are second in spending in premiums.

5.2. Customer Segmentation

For the Customer Segmentation and since we have metric and non-metric data clustering analysis, we decided to utilize the K-Prototype which is a clustering method based on partitioning. Its algorithm is an improvement of the K-Means and K-Mode clustering algorithm to handle clustering with mixed data types. From Figure 11, we can observe some relevant information regarding each cluster.

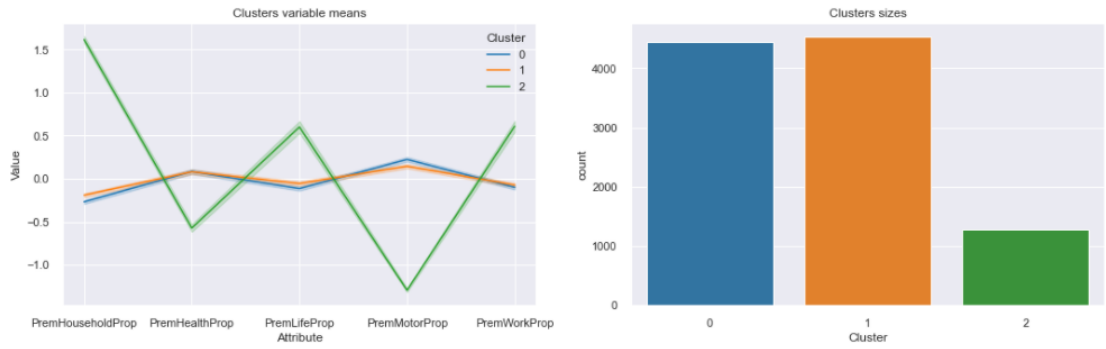

Figure 11 – Profiling of customer of K-Prototype

From the results, we identified and categorized each cluster. The clusters are described as follows:

Cluster 0 - Old customers: These customers are mainly characterized by the variable *ClientYears*, these are customers who have used the company's services for a long time. Other than that, they don't reveal any unusual behaviour.

Cluster 1 - New customers: Again, the *ClientYears* variable is the cluster's main characteristic, these are newer customers with average spending behaviours.

Cluster 2: This is a smaller subset of customers, containing around 12% of the observations. It is mainly characterized by their low yearly salaries, but high premium spending, having the highest *TotalPremiums* values. These customers spend mostly on household, life and work insurance and don't care about motor.

To conclude, the two main clusters make up around 88% of the dataset, and present almost identical variables in all variables except *ClientYears*. Also, the demographic variables, level of education and children do not vary significantly from each of the clusters, with every single cluster revealing around the same average values. Overall, the Product Segmentation analysis was much more indicative of the customers behaviour, and there was significantly higher variation in values between clusters.

5.3. Cluster Relationships

Observing the Figure below (Figure 12), we can see the relations between the product and customer clusters. We realized that the points present on the bottom right are the ones where the two product and customer clusters are characterized by their high yearly salary and high spending on Health. These customers aren't particularly associated to any level of education or cancellation rate.

On the other hand, in the bottom left, product clusters 0 and 1 are associated with customer cluster 0, these customers are characterized by their high customer value. On the top of the plot, the clusters are characterized by their low customer value, low salaries, but spend the highest percentage of their salary in insurance. These customers are associated with a basic level of education.

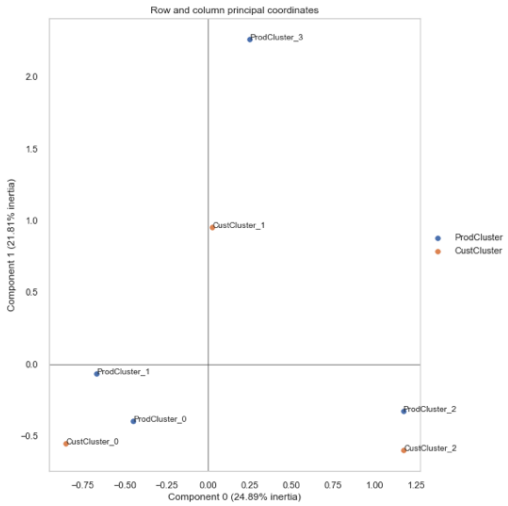

Figure 12 – Profiling of concatenation

5.4. Final Clusters

Since the customer data cluster results were disappointing as previously explained, we decided to use the clusters found in the product segmentation part to perform the final analysis. In the following Figures (Figures 13 and 14), we can observe the final clusters.

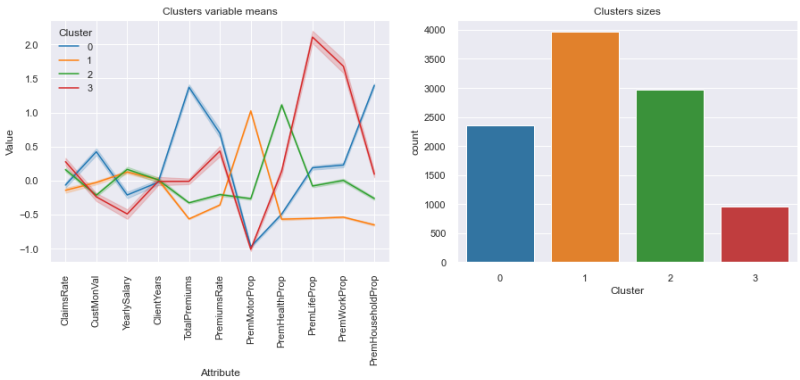

Figure 13 – Final profiling of concatenation of K-Prototype

Figure 14 – Final depth profiling of concatenation of K-Prototype

From the results, we identified and categorized each cluster. The final clusters are described as follows:

Cluster 0 - Most Valuable Customers: These customers spend the most on insurance, in total and relative to their salary, and have the highest customer monetary value. However, they have below average yearly salaries. Their main expenses are on Household, and they do not invest a lot on Motor.

Cluster 1 - Low Risk: This is the largest cluster, made of around 4000 observations. These customers spend the least on premiums, in total and relative to their salary. They spend a high proportion on motor and a low proportion on all other premiums. They have the lowest claims rate and high yearly salaries. According to the correspondence analysis, this cluster has a positive association with the PHD education and contract cancelations.

Cluster 2 - Untapped Potential: This is second largest cluster, made of around 3000 customers. These customers have the highest salaries, but the second lowest premiums rates. They spend a lot on health and around the average on all the other premiums. Their claims rate value is high, and they have a low customer monetary value.

Cluster 3 - Work and Life focus: This cluster is mostly characterized by its high life and work spending and positive association with Basic education. These customers have the second highest spending in total and relative to salary. Their yearly salary is the lowest and they also have the lowest customer monetary value and highest claims rate. This is by far the smallest cluster, made up of less than 1000 observations.

6. Marketing Approaches

From the results observed and discussed above, we created some marketing approaches. The following suggestions are useful when doing the marketing campaigns for each cluster.

The **Cluster 0 - Most Valuable Customers** seems to be the cluster where customers already spend a lot of their salary on insurance, so the company's main concern should be keeping these customers happy with premium customer support and then try to attract even more of these customers. Analyse the Household insurance competition, offer advantages to new customers like a 25% discount on household insurance for the first 6 months, or bundle other types of insurance for a discounted price.

The **Cluster 1 - Low Risk** are the stingy customers who are willing to cancel a service if they are not satisfied with it. Therefore, the key is not to try to push them other services they don't want, but instead offer premium services for services they are already interested in, in this case, motor vehicles. So, keep the low-risk income from these customers and offer additional optional premium services.

The **Cluster 2 - Untapped Potential** seems to be the group with high salaries, these costumers spend a balanced amount on all premiums and an above average amount on health, they don't have high cancelation rates. So why is their claims rate so high and customer value so low? The company should look to optimize their health insurance offers as these provide low income at a high risk. Perhaps the solution is increasing health prices, even if it means losing some customers.

The **Cluster 3 - Work and Life Focus** is the group of the customers that have spent a large portion of their salaries on the company's offered services, yet they have low customer values and high claims rates. These customers have an above average proportion spending for all types of insurance except for Motor with values below average, which is odd since they don't have a high cancelation rate. The company should explore budget options for motor vehicles insurance, and/or promote a budget bundle covering all types of insurance. Increasing motor spending would provide a low-risk source of income from these customers.

7. Conclusion

In this project, we were asked to develop a Customer Segmentation in order for the company's Marketing Specialists to understand their customers' Profiles. Before we started with the actual clustering, we had to prepare the data, starting with coherence checking, outlier removal, values imputation, and feature engineering, including the analysis of the non-metric variables as they were an impediment to the typical clustering techniques.

We decided to divide the clustering process into two phases: Product, which refers to how much the customer spent on each type of the company's provided insurance services; and Customer, where we looked at the customers' personal and spending data like total amount spent on premiums and customer monetary value. In these phases, we used various techniques such as K-Means, K-Prototype, Self-Organizing Maps, Hierarchical Clustering, and Gaussian Mixture Models (in some cases, we combined these techniques as well). In order to analyse each cluster, we utilized methods such as Elbow, Silhouette analysis, Dendrograms, among others.

To get the final clusters, we decided to use the ones found in the product segmentation part to perform the final analysis since the customer data cluster results were disappointing and the product segmentation analysis was much more indicative of the customer's behaviour, and there was significantly higher variation in values between clusters.

Finally, we suggested a set of possible marketing approaches to each identified cluster that can be helpful for the marketing department of the company to understand all the various Customers' Profiles and with that develop more elaborate and specific strategies.

8. References

1. KNN Imputation for Missing Values in Machine Learning – https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

2. The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical) – https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

3. GMM: Gaussian Mixture Models — How to Successfully Use It to Cluster Your Data? – https://towardsdatascience.com/gmm-gaussian-mixture-models-how-to-successfully-use-it-to-cluster-your-data-891dc8ac058f

4. Digital Tribes: customer clustering with K-Means – https://towardsdatascience.com/digital-tribes-customer-clustering-with-k-means-6db8580a7a60

5. Customer Segmentation in Python – https://towardsdatascience.com/customer-segmentation-in-python-9c15acf6f945

6. Silhouette Method — Better than Elbow Method to find Optimal Clusters – https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

7. Selecting Categorical Features in Customer Attrition Prediction Using Python – https://datascienceplus.com/selecting-categorical-features-in-customer-attrition-prediction-using-python/

8. K-Prototypes - Customer Clustering with Mixed Data Types – https://antonsruberts.github.io/kproto-audience/