Insurance Company Customers' Segmentation

Descriptive Methods for Data Mining

Final Project

2021/2022

1st Semester

Group 20

André Antunes Oliveira m20211253

Ana Francisca Dias m20211085

Joana Ferreira Marques m20211173

Index

Index of Figures

Index of tables

Introduction

Within the scope of the Data Mining course, we were proposed to carry out a project taking on the role of consultants specialized in Data Mining.

The basis of the project is a set of customer data from an insurance company, for which it is necessary to study to identify different customer segments to help the marketing department understand the profiles of its customers.

In this work, the CRISP-DM model will be adopted, where the different clusters of chosen customers will be identified, described, and explained with the respective approach of advantages or disadvantages of different decisions.

1. Business Understanding

1.1. Business objectives

As no means of communication was provided with the company in question, it was not possible to gather specific customer requirements or information about their situation in the initial stage of the project. In this way, we will take a general approach to any company in the insurance industry.

The main objective of any business is to increase sales, profits, and customers. In the insurance industry, customers pay a premium to insure themselves or an asset. However, some customers are more profitable to the company than others due to the size and number of their claims. In other words, the main objective is to know which are the most profitable customers for the company.

The identification of different customer profiles is important for various departments within an insurance company, however, the challenge proposed in this project focuses only on the perspective of the marketing department.

For the marketing department, information about customer profiles is a decisive factor from the perspective of defining strategies for attracting new customers and retaining them. The implementation of marketing campaigns, alternative communication channels, and product promotions are some of the activities in the sector that, when directed to target customers, make this process more efficient and productive for the company.

The main objective of the project is being able to identify the different client profiles in the data set provided from the insurance company to help the marketing department direct the promotions and campaign to target clients profiles.

The client data set provided assumes:

- That the current year is 2016;
- The premium negative values are premiums paid in previous years

2. Data Understanding

2.1. Description and exploration of data

According to the information received, there are 10 296 insurance different clients in the dataset. The type of variable and its quantity are shown in table 1:

Variable	Description	Type of variable	Quantity
CustID	Client ID	ID	10 296
FirstPolYear	First-year as a company customer	Continuous	10 266
BirthYear	Customers birth year	Continuous	10 279
EducDeg	Customers academic degree	Nominal	10 279
MonthSal	Customers gross monthly salary (€)	Continuous	10 260
GeoLivArea	Customers area of residence	Nominal	10 295
Children	Indication of the costumer's children	Binary Symmetric	10 275
CustMonVal	Customer monetary value (€)	Continuous	10 296
ClaimsRate	Customers claims ratio	Continuous	10 296
PremMotor	Monthly premium paid on car insurance (€)	Continuous	10 262
PremHousehold	Monthly premium paid in household insurance (€)	Continuous	10 296
PremHealth	Monthly premium paid in health insurance (€)	Continuous	10 253
PremLife	Monthly premium paid in life insurance (€)	Continuous	10 192
PremWork	Monthly premium paid in work insurance (€)	Continuous	10 210

Table 1 - Description, type, and quantity of each variable

Initially, the basic descriptive statistics were plotted to get an overall view of the dataset, measuring the central tendency and the distribution of each attribute

	CustID	FirstPolYear	BirthYear	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife	PremWork
Mean	5,148.5	1,991.1	1968	2,506.7	2.7	0.7	177.9	0.7	300.5	210.4	171.6	41.9	41.3
Standard Deviation	2,972.3	511.3	19.7	1,157.4	1.3	0.5	1,945.8	2.9	211.9	352.6	296.4	47.5	51.5

Table 2 - Mean and Standard Deviation of dataset's features.

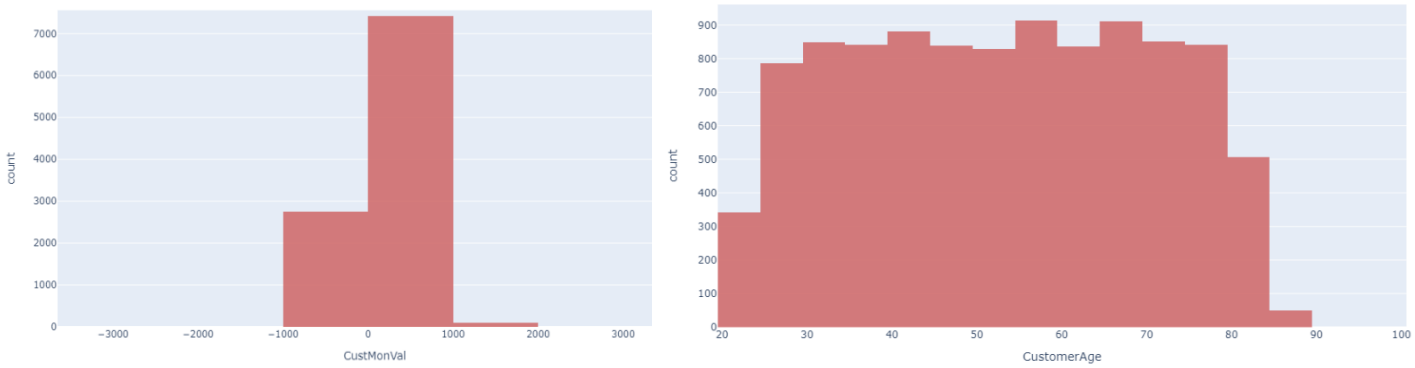

Table 3 - Customers monthly salary and customers age histograms

When performing a detailed analysis on the variables related to the customer profile (Figures 1 to 5), it is possible to understand that:

- The customer monetary value[1] (CustMonVal) is highly skewed with a mean relatively low. Its standard deviation suggests that the observations tend to be spread over a large range of values. This indicates that this variable is prone to the existence of outliers.
- The variable CustomerAge has been created using BirthYear, to ease the representation of the customer age. Customers have a mean of roughly 53 years old, with a standard deviation of 19.7. As expected, there are fewer customers below the 30s and above the 80s.
- EducDeg characterizes the customer's level of education in b'1, b'2, b'3, or b'4, which represent basic, secondary, master's, or doctoral education, respectively. As figure 3 shows 80% of customers have a high school or master educational level. Therefore, we can state that there is a class imbalance on this variable.
- Customers with one or more children represent 70% of the customers, as figure 5 shows. As a result, we can state that there is a class imbalance on this variable.
- GeoLivArea characterizes the customer living area in 1, 2, 3, and 4. As figure 4 shows, 80% of the customers live in areas 1 and 4. Therefore, we can state that there is a class imbalance on this variable.
- The feature FirstPolAge has been created using FirstPolYear, to ease the representation of the customer's connection with the company. Customers policies' have a mean of roughly 40 years old, with a standard deviation of 510.9. This indicates that this variable is highly skewed and, as a result, prone to the existence of outliers.
- The variable MonVal relates to the customer monthly gross salary. Its standard deviation is 1157.4, indicating that this variable is highly skewed. In addition, the distribution reveals that it can have potential outliers, as there is a great difference between the maximum value and the third quartile (Figure X).

- ClaimsRate refers to the percentage of the claims made by a customer in the last two years, calculated as the ratio between the amount paid by the insurance company and the premiums. When analyzing its distribution, it is visible that this continuous variable is very skewed and presents a great gap between the maximum value and the third quartile – indicating potential outliers.

[1] Customer Monetary Value = (annual profit from the customer) x (number of years since a customer) - (acquisition cost)

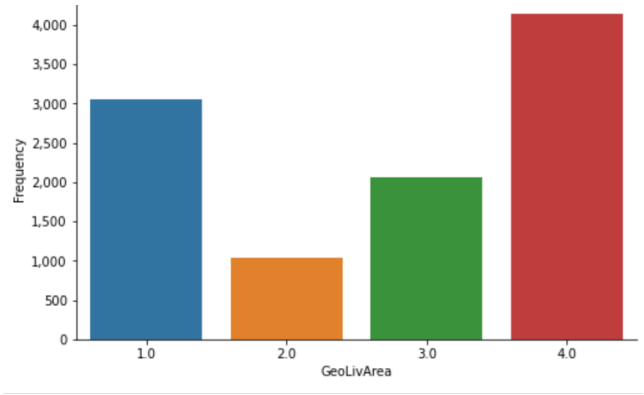

Figure 2 - GeoLivArea class distribution

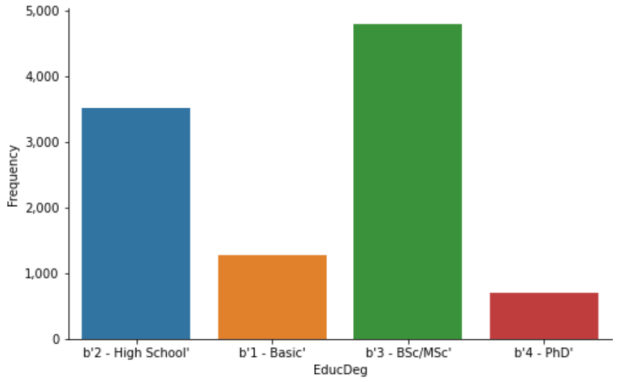

Figure 1 - Education level class distribution

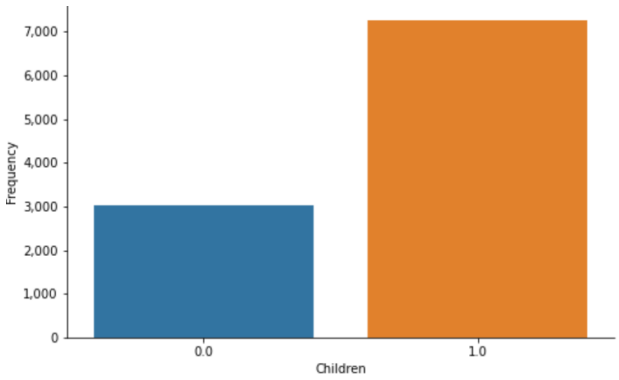

Figure 3 - Children's Class

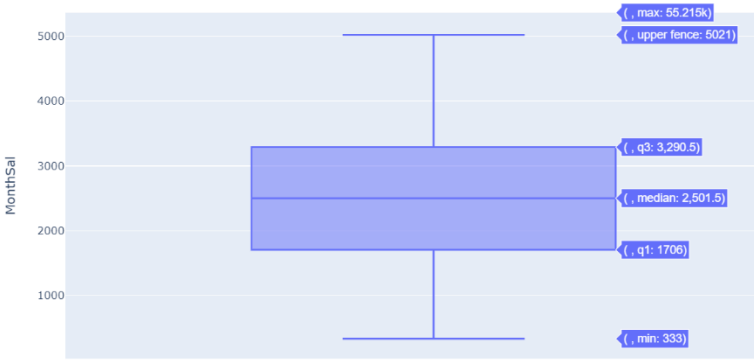

Figure 4 - Box plot of MonthVal.

In order to have some perspective about the client role inside the insurance company, the variables FirstPolYear, CustMonVal, ClaimsRate, PremMotor, PremHousehold, PremHealth, PremLife e PremWork were analyzed.

Represented in figure 6 is the analysis of the variables PremMotor, PremHousehold, PremHealth,

PremLife, and PremWork. Figure 6 shows that the life line of business has a higher quantity of premiums paid which indicates that is the most sold product. The life insurance is followed by work and motor line of business. The household and health insurances are the least sold products of the insurance company.

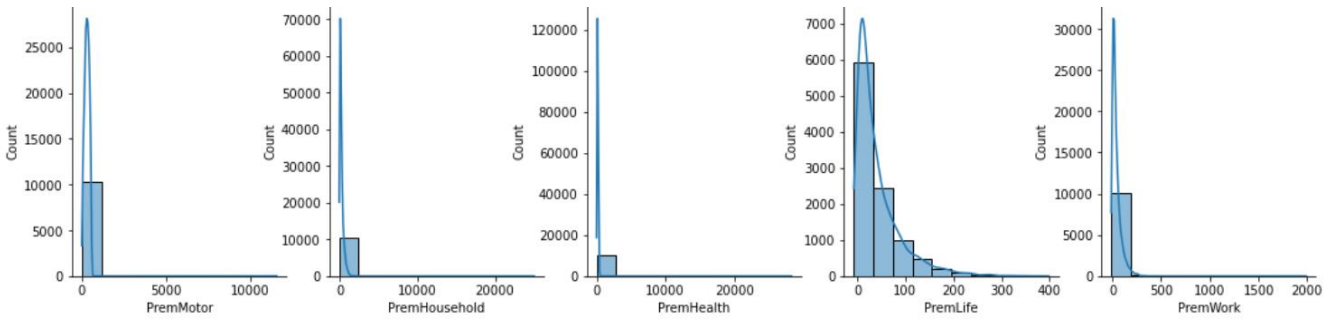

Figure 5 - PremMotor, PremHousehold, PremHealth, PremLife and PremWork distributions.

From figure 6 it is possible to understand the correlations between the variables. Regarding customers related variables, the MonthSal and CostumerAge have a correlation of 0.93, Birthyear and Children have a correlation of 0.49.

Regarding company variables, the ClaimsRate and CustMoneyVal have a correlation of -0,97 which indicates that the fewer the customer's claims are, the higher the value of the customer to the company. The variable ClaimsRate is also related to all the premium variables which can imply that surcharges or discounts can be applied to the insurance premium if the claim rate decreased or increases for a certain customer. It is also important to mention that the variables Customer ID and FirstPolYear have no relevant correlations

There are no strong correlations between the lines of business premiums, however premium motor has a correlation of 0.47 and 0.48 with premium life and premium work, respectively and premium health has a correlation of 0.48 with premium work. These relations can imply that these products are usually sold together.

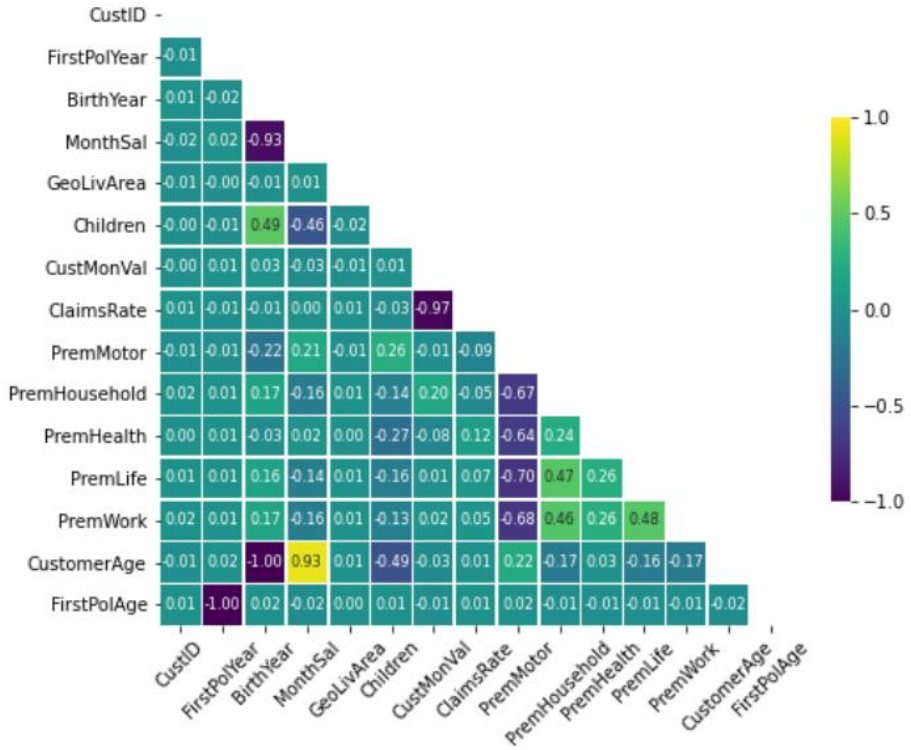

Figure 6 - Correlations between variables

2.2 Data Quality

2.1.1. Missing Values

The preliminary assessment (Table 8) has shown that there are missing values on some of the features of the dataset. When analyzing in detail the variables with missing values, the full picture can be summarized as follows:

Variable	Missing values
FirstPolYear	30
BirthYear	17

EducDeg	17
MonthSal	36
GeoLivArea	1
Children	21
PremMotor	34
PremHealth	43
PremLife	104
PremWork	86

Table 4 - Missing values by feature.

.

3. Data Preparation

Following the CRISP-DM methodology, the data preparation stage is a crucial part of the process, as data quality directly impacts the final output. It involves cleaning, filtering, and integrating data, in order to produce a high-quality dataset to be used on the modeling stage.

3.1. Data Types & Duplicate Values

In order to prepare data for the modeling stage, one of the fundamental steps is to ensure that the features have the correct data type. As presented below, variables CustID, GeoLiveArea and Children required data type transformations.

Variable	Raw Data Type	Transformed Data Type
CustID	float64	object
EducDeg	object	object
MonthSal	float64	float64
GeoLivArea	float64	int64
Children	float64	int64
CustMonVal	float64	float64
ClaimsRate	float64	float64
PremMotor	float64	float64
PremHousehold	float64	float64
PremHealth	float64	float64
PremLife	float64	float64
PremWork	float64	float64

Table 5 - Variable data type and transformed data type

In addition, it was verified that the dataset provided didn't contain any duplicate records.

3.2. Missing Values

According to the information provided in section 3.1, there were missing values on some of the features of the dataset that needed to be carefully addressed:

- The rows with missing records on variable BirthYear were dropped, as age is an important trait to understand the customer profile.
- The customer's loyalty can be derived from feature FirstPolYear, therefore all rows with missing instances of this variable were deleted.
- Missing values on binary feature Children were deemed to be zero, an approximation considering that such field might have been left blank on the source system, in order to state that the customer has no children.
- The rows with missing instances on variable GeoLivArea were dropped, as the geographic location is an important trait to understand the customer profile.
- Missing records on feature MonthSal were considered to be zero. This is an approximation, taking into account that such a field might have been left blank on the source system in order to state that the customer had no income.
- As the categorical variable EducDeg is fundamental to understanding the customer's profile, the missing values were populated with a new category to classify no education level.
- The missing instances on premium features (PremMotor, PremHealth, PremLife, PremWork) were considered as zero. An insurance premium is the amount of money an individual or business pays for an insurance policy that covers healthcare, auto, work, and/or life insurance. Once earned, the premium is income for the insurance company. If there is no income reported, then one can assume that not all customers are underwritten all insurance products available.

It should be noted that the features identified in the previous section as variables with class imbalance (EducDeg, GeoLivArea, and Children) kept roughly the same class proportions after the treatment of the missing value.

3.3. Miscoded Values

The educational background is relevant to understanding which products a customer subscribes. This is even more important as there are work, health, and life policy coverages – the individual educational track impacts directly its work type (e.g. most unqualified jobs can involve a high level of danger).

As there was some class imbalance detected on the feature EducDeg, it was required further analysis of its cardinality. There were no miscoded values on this variable, however, for customer modeling purposes, it can be helpful to use a code instead of a category. As a result, the column EducDeg_CD was created, storing all EducDeg codes for later use.

EducDeg Class	EducDeg Class Code	Cardinality
None	0	2
b'1 - Basic'	1	1268
b'2 - High School'	2	3497
b'3 - BSc/MSc'	3	4791
b'4 - PhD'	4	694

Table 6 - EducDeg class code and cardinality

3.4. Outliers

An outlier is an extremely large or small observation relative to the rest of the dataset. It may represent a data entry error or may be a genuine observation, however it needs to be properly addressed before the modeling stage.

As previously referred on the data exploration stage, there are several variables that highlight the potential existence of outliers. In order to detect and filter the anomalies among the observations, the following actions were taken:

Feature	Min	Max	Filtering Method	Upper Limit	Lower Limit	Number of Outliers	Rationale
FirstPolYear	1974	53784	Default	2016	-	1	No observation can be greater than the current year of the database (2016).
BirthYear	1028	2001	Default	-	1900	1	Given that the greatest age any human has ever lived is 122 years, no observation can surpass that value.
CustomerAge	20	993	None	-	18	0	Considering that this column is generated from BirthYear, after removing its outliers, the check needed to be done to spot any underage customers.
FirstPolAge	23	51763	None	-	CustomerAge	1998	Considering that this column is generated from FirstPolYear, the check needed to be done to spot any policies starting before the BirthYear. However, no filter was applied because policies can be transferred to any dependents, being previously associated to the caretakers.
ClaimsRate	0	256.2	None	-	-	-	No observation can have a negative rate.
CustMonVal	-165680	11875.89	Z-score	Z-score = 3	Z-score = - 3	13	When analyzing the distribution, we can assume that the data should be approximately normal. Therefore, we applied a z-score approach to remove the outliers higher and lower than the upper and lower limits.
MonthSal	333	55215	Z-score	Z-score = 3	Z-score = - 3	2	When analyzing the distribution, we can assume that the data should be approximately normal. Therefore, we applied a z-score approach to remove the outliers higher and lower than the upper and lower limits.
PremMotor	-4.11	11604.42	Z-score	Z-score = 3	Z-score = - 3	6	When analyzing the distribution, we can assume that the data should be approximately normal. Therefore, we applied a z-score approach to remove the outliers higher and lower than the upper and lower limits.
PremHousehold	-75	25048.8	Z-score	Z-score = 3	Z-score = - 3	36	When analyzing the distribution, we can assume that the data should be approximately normal. Therefore, we applied a z-score approach to remove the outliers higher and lower than the upper and lower limits.
PremHealth	-2.11	28272	Z-score	Z-score = 3	Z-score = - 3	3	When analyzing the distribution, we can assume that the data should be approximately normal. Therefore, we applied a z-score approach to remove the outliers higher and lower than the upper and lower limits.
PremLife	-7	398.3	None	-	-	-	After studying the distribution, no filters were applied.
PremWork	-12	1988.7	None	-	-	-	After studying the distribution, no filters were applied.

Table 7 - Detected outliers and actions taken

It is important to describe the z-score approach to remove outliers, shown in Table 6. Z-scores can quantify the unusualness of an observation when the data follows a normal distribution, being the number of standard deviations above and below the mean that each value falls:

$$Z = \frac{X - \mu}{\sigma}$$

Assuming that the data should resemble a normal distribution, any z-score greater than 3 or less than -3 is considered to be an outlier. This rule of thumb is based on the empirical rule. From this rule we see that almost all the data (99.7%) should be within three standard deviations from the mean, as shown in Figure 8.

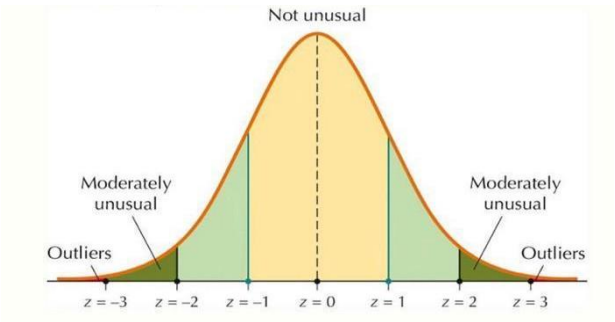

Figure 7 - Normal distribution

3.5. Feature Generation & Label Encoding

Considering the initial dataset provided, the following variables were created:

- **CustomerAge** – age of the customer, based on the BirthYear.
- **FirstPolAge** – first policy age, based on the FirstPolYear.
- **CustomerAge_bins** – age of the customer binned into 6 segments (<20, 20-29, 30-39, 40-49, 50-59, >=60), based on the CustomerAge. This feature was then one hot encoded[2].
- **FirstPolAge_bins** – first policy age binned into 6 segments (<20, 20-29, 30-39, 40-49, 50-59, >=60), based on the FirstPolAge. This feature was then one hot encoded.
- **TotalPremium** – total premium, based on the sum of all premium features (PremMotor, PremHousehold, PremHealth, PremLife and PremWork).
- **TotalPremium_bins** – total premium binned into 4 segments[3], based on the distribution of TotalPremium shown in Figure 9. This feature was then one hot encoded.
- **TotalCost** – Amount paid by the insurance company, based on TotalPremium divided by ClaimsRate[4]. This feature was then one hot encoded.
- **TotalCost_bins** – total premium binned into 4 segments[5], based on the distribution of TotalCost shown in Figure 10. This feature was then one hot encoded.
- **MonthSal_bins** – customer gross monthly salary binned into 7 segments[6], based on the distribution of MonthSal. This feature was then one hot encoded.

[2] One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.

[3] <626=low, 626-697=medium, 698-822=high, >=823=very high

[4] ClaimsRate = Amount paid by the insurance company (€)/Premiums(€) (in the last two years)

[5] <292=low, 292-544=medium, 545-646=high, >=647=very high

[6] 0=*unemployed*, 1-333=*low income* ,333-1706=*medium low income*,1707-2501= *medium high income*, 2502-3290=*high income*, 3290-5021=*very high income*, >=5021=*extremely high income*

- **CustMonVal_bins** – customer monetary value binned into 4 segments[7], based on the distribution of CustMonVal. This feature was then one hot encoded.

The features GeoLivArea, Children and EducDeg were additionally one hot encoded in order to prepare the modeling dataframe.

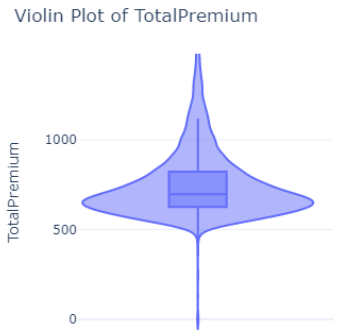

Figure 8 - Distribution of TotalPremium

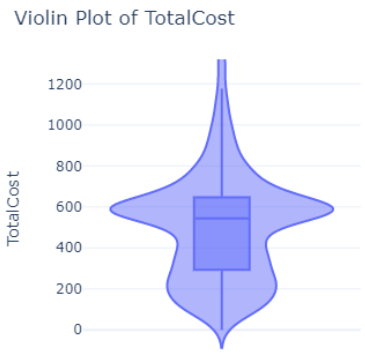

Figure 9 - Distribution of TotalCost

3.6. Feature Scaling

Before moving into the modeling stage, it is important to address the normalization of features, as most of them present different scales. All distance-based algorithms (such as k-Means) are affected by the scale of the variables. An algorithm impacted by the higher/lower magnitude of variables is already biased by default.

In order to overcome this problem, all continuous variables were brought to the same scale using the min-max scaler. As the formula suggests, it compresses the features into a range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset. This method is completely independent of any assumption that the feature follows a normal distribution or not.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4. Modeling

4.1. Technique

In the case of customer segmentation analysis, Principal Components Analysis (PCA) and K-Means Clustering were the two selected techniques. As the transformed dataset contains a high number of dimensions, by applying PCA we are not only retaining all the important information, but also preparing data for the clustering, as the clusters will be likely more visible.

[7] *<=0=negative, 1-187=low, 188-398=medium,>=399=high*

4.1.1. Principal Component Analysis

Principal component analysis is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables, the principal components, that successively maximize variance.

4.1.2. K-Means

K-means is a centroid-based algorithm, or a distance-based algorithm, where the distances are calculated to assign a point to a cluster. Therefore, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

4.2. Development

After applying the PCA, the number of principal components to be clustered was selected based on retaining those with the highest variance explained. When analyzing Figure 11, it is clear that the explained variance stabilizes after the 20th principal components, therefore the first 20 dimensions of PCA were used on the clustering analysis.

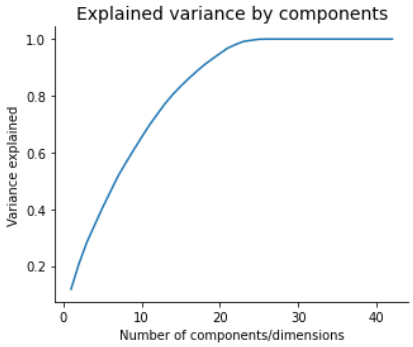

Figure 10 - Variance by componentes

In order to determine the most suitable number of clusters, two tests were used: the elbow method and the silhouette score:

- The elbow method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid and represent it in a plot. Afterwards the value of k is picked where the average distance falls suddenly.
- The silhouette score is a measure based on the separation distance between the resulting clusters, indicating how close each point in one cluster is to points in the neighboring clusters. Silhouette coefficients have a range of range of [-1, 1]. Values near +1 indicate that the sample is far away from the neighboring clusters, while 0 states that the sample is on or very close to the decision boundary between two neighboring clusters, and negative values indicate that those samples might have been assigned to the wrong cluster.

The results of both methods are presented on Figure 12 and 13. The elbow method seems to be inconclusive, as there is no clear sudden drop. However, when analyzing the silhouette scores, it is possible to see that the most suitable number of clusters for the dataset is 6.

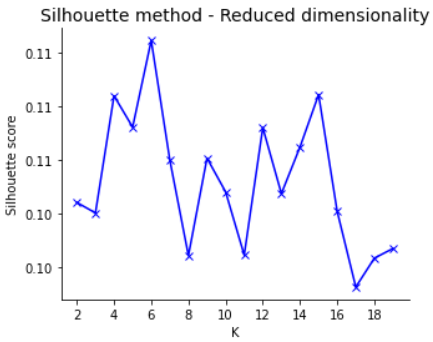

Figure 12 – Silhouette Method

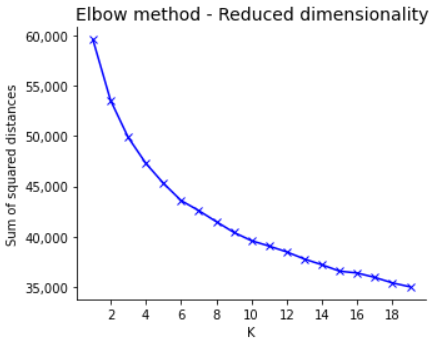

Figure 11 – Elbow Method

4.3. Assessment

In order to assess the model results, we must analyze the clusters cardinality (number of objects per cluster) and magnitude (sum of distances from all objects to the centroid of the cluster). When analyzing Figures 14 and 15, it is visible that clusters 0 and 2 seem to be outliers due to lower cardinalities and magnitudes. However, as a higher cardinality tends to result in a higher magnitude, clusters are only deemed as anomalous when cardinality does not correlate with magnitude. Therefore, Figure 17 clearly shows that in general cardinality is correlated to magnitude for all clusters, as a result we can legitimately state that no major anomalies seem to exist in the clusters.

In addition, Figure 16 shows a representation of the first two principal components by cluster, which explain roughly 20% of the variance combined. It is visible that there is already a clear pattern between clusters, especially segregating cluster 1 from the others.

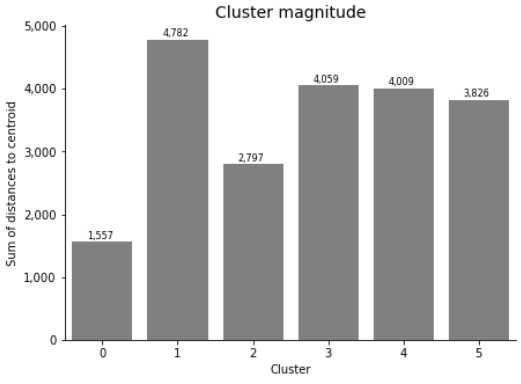

Figure 14 - Cluster magnitude

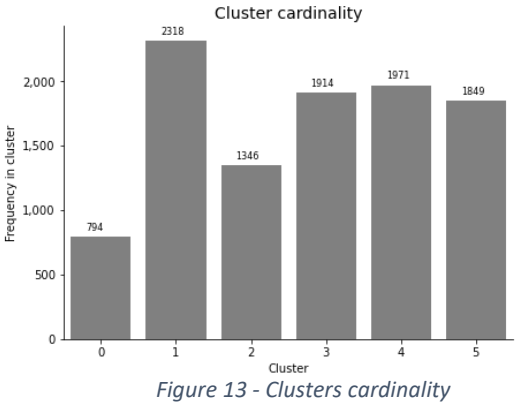

Figure 13 - Clusters cardinality

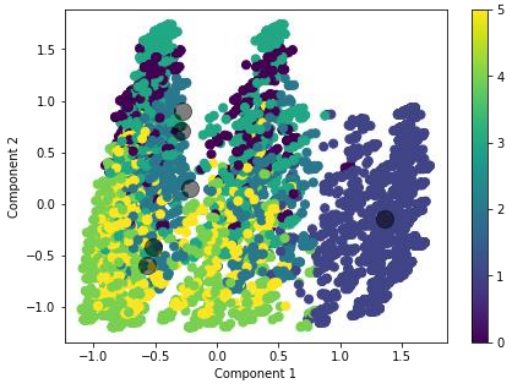

Figure 16 - Two principal components by cluster

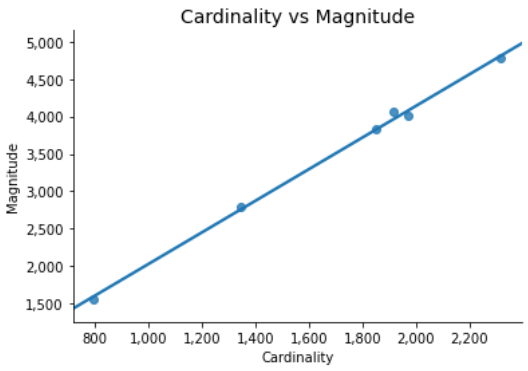

Figure 15 –Relation between cardinality and magnitude

5. Evaluation

5.1. Results

Considering the model results in the attachment (Appendix) we considered to this case the six clusters below:

Cluster 0: Younger customers with some acquisitive power

- Less educated customers (basic and high-school)
- Medium-low income gross monthly salary
- Higher total premium (premium features - PremMotor, PremHousehold, PremHealth, PremLife and PremWork).
- Customers with age above 20, but mostly between 20-29 years old
- 77% have 1 or more children
- Prone to direct marketing
- 70% of customers live in area 1 and 4

Cluster 1: Better educated, without family, customer with more that 60 years old

- Better educated customers (high-school and university - master degree)
- Very high income gross monthly salary
- Customers above 60 years old
- Without children
- Less likely to subscribe
- More likely to have a loan than customers of cluster 0 (probably for the house)
- 43% of customers live in area 4

Cluster 2: Better educated, with children and medium amount paid

- Better educated customers (high-school and university - master degree)
- Medium low and medium high income gross monthly salary
- 94% have 1 or more children
- Customers above 20, but mostly between 30-39 years old
- Medium amount paid by the insurance company
- Second cluster is less propense to direct marketing. But the first cluster is more propense
- 70% of customers live in area 1 and 4

Cluster 3: Medium income gross monthly salary and different levels of education

- Medium low and medium high income gross monthly salary
- Very high amount paid by the insurance company
- Low customer monetary value
- Customers above 20 years old
- 89% have 1 or more children
- Better educated customers (high-school and university - master degree)
- 41% of customers live in area 4

Cluster 4: Customers with different levels of education, children and low amount paid

- Customers above 20 years old
- Better educated customers (high-school and university - master and doctoral degree)
- 94% have 1 or more children
- Low amount paid by the insurance company
- High customer monetary value
- 70% of customers live in area 1 and 4

Cluster 5: Very high amount paid by the insurance company and different level of education

- Customers above 20, but mostly between 50-59 years old
- Better educated customers (high-school and university - master and doctoral degree)
- Very high amount paid by the insurance company
- 95% have 1 or more children
- Low customer monetary value
- No record of default
- Cluster more prone to direct marketing
- 89% of customers live in area 1, 3 and 4

5.2. Next Steps

As the main business objective of the project has been met, building a model capable of efficiently identifying the different client profiles, the next logical step is to proceed with the deployment in production. This allows the model to bring benefit to the organization, helping the marketing department directing the promotions and campaign to target clients profiles.

6. Deployment

In order to bring benefit to the company, the model needs to be deployed in a production environment. As a result, it is important to define a deployment strategy to help avoiding unnecessarily long periods of incorrect usage of data mining results:

- Considering that the customer segments don't need to be updated daily and no live access is required, it is recommendable to run the model once every month. This timeframe allows enough data to be gathered, so the output can bring consistent value to the business departments.
- The output of the model can be presented to the business stakeholders on a dashboard in Business Intelligence software like Tableau or Power BI, reaching as many business users as required according to the company data security policy.

- Finally, the model can be automated using ETL software like Airflow. It can connect automatically to the database to collect data, trigger the calculation runs on its own and present the

8. Appendix

	0	1	2	3	4	5
CustomerAge_<20	0	0	0	0	0	0
CustomerAge_20-29	0,5465994962	0	0,0750371471	0,2466039707	0,01217656012	0,03461330449
CustomerAge_30-39	0,2732997481	0	0,3187221397	0,2387669801	0,128361238	0,1746890211
CustomerAge_40-49	0,05541561713	0,0004314063848	0,2384843982	0,1666666667	0,2648401826	0,2736614386
CustomerAge_50-59	0,03400503778	0,002588438309	0,1129271917	0,1567398119	0,3505834602	0,3001622499
CustomerAge_>=60	0,09068010076	0,9969801553	0,2548291233	0,1912225705	0,2440385591	0,2168739859
FirstPolAge_<20	0	0	0	0	0	0
FirstPolAge_20-29	0,2531486146	0,2786885246	0,2295690936	0,2580982236	0,2612886859	0,2606814494
FirstPolAge_30-39	0,4420654912	0,4352890423	0,4546805349	0,4352142111	0,4424150178	0,4391563007
FirstPolAge_40-49	0,3047858942	0,2860224331	0,3157503715	0,3066875653	0,2962962963	0,3001622499
FirstPolAge_50-59	0	0	0	0	0	0
FirstPolAge_>=60	0	0	0	0	0	0
MonthSal_0	0,001259445844	0,003451251079	0,0044576 52303	0,002612330199	0,005073566717	0,002163331531
MonthSal_1-333	0	0	0	0	0	0
MonthSal_333-1706	0,7833753149	0	0,3350668648	0,4446185998	0,1395230847	0,1773931855
MonthSal_1707-2501	0,0919395466	0,006039689387	0,3246656761	0,263322884	0,3845763572	0,4088696593
MonthSal_2502-3290	0,08816120907	0,1919758412	0,2481426449	0,198014629	0,3587011669	0,3288263926
MonthSal_3290-5021	0,03526448363	0,7985332183	0,0876671 6196	0,09143155695	0,1121258245	0,08274743104
MonthSal_>=5021	0	0	0	0	0	0
Children_0	0,234256927	0,9918032787	0,0616641 9019	0,1175548589	0,06037544394	0,04651162791
Children_1	0,765743073	0,008196721311	0,9383358098	0,8824451411	0,9396245561	0,9534883721
EducDeg_0	0	0,0004314063848	0,0007429420505	0	0	0
EducDeg_1	0,3942065491	0,08714408973	0,0965824 6657	0,2053291536	0,04819888382	0,05840995133

EducDeg_2	0,4055415617	0,3451251079	0,3982169391	0,4258098224	0,2541856925	0,2742022715
EducDeg_3	0,1952141058	0,4935289042	0,4673105498	0,3385579937	0,5844748858	0,5662520281
EducDeg_4	0,005037783375	0,0737704918	0,03714710253	0,0303030303	0,1131405378	0,1011357491
GeoLivAre_1	0,3060453401	0,2812769629	0,3268945022	0,28892372	0,3003551497	0,2877230936
GeoLivAre_2	0,1095717884	0,1039689387	0,08766716196	0,08934169279	0,09944190766	0,1124932396
GeoLivAre_3	0,1889168766	0,1880931838	0,2154531947	0,2110762	0,2019279554	0,2001081666
GeoLivAre_4	0,395465995	0,4266609146	0,3699851412	0,4106583072	0,3982749873	0,3996755003
CustMonVal_<=0	0	0,2467644521	0,008915304606	0,6598746082	0,005580923389	0,4726879394
CustMonVal_1-187	0,01763224181	0,2579810181	0,220653789	0,2810867294	0,00304414003	0,4905354246
CustMonVal_188-398	0,137279597	0,3283002588	0,7325408618	0,05903866249	0,2567224759	0,03677663602
CustMonVal_>=399	0,8450881612	0,1669542709	0,03789004458	0	0,7346524607	0
TotalPremium_<626	0,008816120907	0,1867989646	0,1641901932	0,04545454545	0,4459665145	0,4716062737
TotalPremium_626-697	0,007556675063	0,2303710095	0,2734026746	0,2136886102	0,3399289701	0,3293672255
TotalPremium_698-822	0,04659949622	0,3062985332	0,4873699851	0,2716823406	0,1572805682	0,1665765279
TotalPremium_>=823	0,9370277078	0,2765314927	0,07503711471	0,4691745037	0,05682394723	0,03244997296
TotalCost_<292	0,2279596977	0,1626402071	0,00297176820	0,001044932079	1	0
TotalCost_292-544	0,6070528967	0,3274374461	0,9725111441	0	0	0,001081665765
TotalCost_545-646	0,1410579345	0,2415875755	0,02451708767	0	0	0,9989183342
TotalCost_>=647	0,02392947103	0,2683347714	0	0,9989550679	0	0

Table 8 - Model results