

**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

Index

[1. Introduction iii](#_Toc90566665)

[2. Data Understanding iii](#_Toc90566666)

[References v](#_Toc90566667)

[Appendix v](#_Toc90566668)

# 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 analyzing the table for evident groups of clusters, extracting the behavior 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.

# Data Understanding

The company’s ABT from 2016 has 10 296 observations and 14 variablesthat are described in the following table (Table 1).

|  |  |  |
| --- | --- | --- |
| 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

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.

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 assess the quality of the dataset therefore boxplots were produced. As can be seen in Figure 1, the dataset is highly influenced by outliers, especially in the product variables, which affect the quality of the results obtained and hence needs to be addressed.

A screenshot of a computer

Description automatically generated with low confidenceFigure 1 – Metric Variables’ Box Plot

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

Chart

Description automatically generated

Figure 3 – Metric Correlations Matrix

# 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. As a consequence, these can have a negative impact on the latter analysis of the clusters, therefore they needed 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.

## 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 to 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 First\_Year. 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.

## 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 1 and as we already said, most of the variables have outliers. Therefore, we will go step by step throughout the variables, analyzing the histograms and boxplotsfor each one that raises a red flag.

## Filling in the Missing Values

# References

1. scikit-learn.org (n.d.). YelowBrick: Machine Learning Visualization Retrieved from: <https://www.scikit-yb.org/en/latest/>
2. What Are the Advantages of Decision Trees?. Retrieved from: <https://smallbusiness.chron.com/advantages-decision-trees-75226.html>

# Appendix