

**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

Initially, we extracted the dataset content from the a2z\_insurance.sas7bdat file into the notebook, obtaining a dataset of 10290 customers with 14 variables, as can be observed in Figure 1.

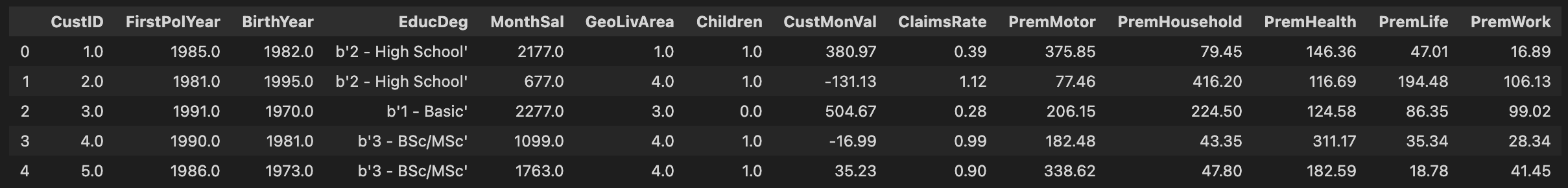


Figure 1 – Dataset head

Then, it was 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.

In fact, there are also variables that refer to the service provided by the company, such as "ClaimsRate", "PremMotor", "PremHousehold", "PremHealth", "PremLife" and "PremWork", and variables that refer to specific characteristics of each customer, such as "FirstPolYear", "BirthYear", "EducDeg", "MonthSal", "GeoLivArea", "Children" and "CustMonVal". Therefore, in the future doing a separate analysis of customer vs products could be a useful process.

The next step was to assess the quality of the dataset therefore boxplots were produced. As can be seen in Figure 2, 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 2 – Metric Variables’ Box Plot

Chart

Description automatically generatedSample 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.

Figure 3 – Metric Correlations Matrix

# Data Preparation

Data preparation is a crucial first task in this Project. Since the team has come across with 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 negative impact in the latter analysis of the clusters, therefore they needed to be rectified.

# 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