

**Data Mining Project**

**MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**Customer Segmentation for Fictional Insurance Company in Portugal**

Group D

Guzel Bayazitova, m20210699

Tongjiuzhou Liu, m20211012

Sergio Corral, m20211011

December 2021

INDEX

[1. Introduction iii](#_Toc91539078)

[1.1. Data Observation iii](#_Toc91539079)

[1.1.1. Initial Data iii](#_Toc91539080)

[1.1.2. Coherence Check iii](#_Toc91539081)

[1.1.3. Data Preparation iv](#_Toc91539082)

[1.1.4. Feature Engineering iv](#_Toc91539083)

[1.1.5. Outliers Removal iv](#_Toc91539084)

[2. Title 2 vi](#_Toc91539085)

[2.1. Title 2.1 vi](#_Toc91539086)

[2.1.1. Title 2.1.1 vii](#_Toc91539087)

[3. References viii](#_Toc91539088)

[4. Appendix 1 - Tables. ix](#_Toc91539089)

# Introduction

An insurance company (fictional), based in Portugal, requested a data science department to carry a Customer Segmentation analysis, in order to understand people’s needs and demands according to their position in each segment of clientele. The objective of the report is to define clusters based on client’s past preferences and behaviours, that would characterise the Customer’s Profiles, and provide recommendations to Marketing Department for the best approach for each cluster.

## Data Observation

### Initial Data

The ABT (Analytic Based Table) contains the original data with 10.296 observations and 14 variables. There are no duplicated values, however, there are missing ones, which needs to be analysed and imputed for further clustering modelling. The descriptive statistics table shows (categorical variables are highlighted) that there are outliers present, like in *FirstPolYear* variable, where max value is 53784, or in *BirthYear*, with min value equal to 1028. The premiums have some negative values, which could be interpreted as reversals in the current (2016) year, paid in previous one. It could either because there were errors in financial transactions, and insurance company reimbursed the amounts, or these customers canceled their memberships and requested partial refund.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Std** | **Min** | **0.25** | **0.5** | **0.75** | **Max** |
| FirstPolYear | 511.267913 | 1974 | 1980 | 1986 | 1992 | 53784 |
| BirthYear | 19.709476 | 1028 | 1953 | 1968 | 1983 | 2001 |
| EducDeg | NaN | NaN | NaN | NaN | NaN | NaN |
| MonthSal | 1157.44963 | 333 | 1706 | 2501.5 | 3290.25 | 55215 |
| GeoLivArea | 1.266291 | 1 | 1 | 3 | 4 | 4 |
| Children | 0.455268 | 0 | 0 | 1 | 1 | 1 |
| CustMonVal | 1945.81151 | -165680.4 | -9.44 | 186.87 | 399.778 | 11875.9 |
| ClaimsRate | 2.916964 | 0 | 0.39 | 0.72 | 0.98 | 256.2 |
| PremMotor | 211.914997 | -4.11 | 190.59 | 298.61 | 408.3 | 11604.4 |
| PremHousehold | 352.595984 | -75 | 49.45 | 132.8 | 290.05 | 25048.8 |
| PremHealth | 296.405976 | -2.11 | 111.8 | 162.81 | 219.82 | 28272 |
| PremLife | 47.480632 | -7 | 9.89 | 25.56 | 57.79 | 398.3 |
| PremWork | 51.513572 | -12 | 10.67 | 25.67 | 56.79 | 1988.7 |

Table 1.1 – Descriptive statistics before imputation

### Coherence Check

The initial check shows, that there is a certain discrepancy in *FirstPolYear* and *BirthYear* variables: the first year, when the customer purchased his membership, is less than his birth year. There are 1997 of such observations, which is about 19.4% of overall data. Since it can’t be just erroneous data entry, an assumption is made that it is a case of transferring ownership of a policy (from parents to their children or from individuals donating their life insurance to charities, etc). The beneficiaries still pay the premiums, so that policies stay active, as result they have higher age, than actual owner’s age. We consider these data as valid and proceed without altering.

### Data Preparation

The variables are divided into non-metric, such as *EducDeg, GeoLivArea* and *Children*, the rest, excluding *CustID*, are defined as metric. The original dataset contains missing values, which we must to fill up in order to apply further data processing, clustering methods and tools. Since there is the biggest amount of missing data among premiums values, we start the data imputation with the 4 variables: *PremMotor*, *PremHealth*, *PremLife,* and *PremWork* (*PremHousehold* doesn’t have missing values*)*. Customers from these observations might have bought a car insurance, but didn’t buy life and house insurances, so the assumption made that they spend zero amount on this particular packages. As such, for customers, who have missing values in some types of insurance, its decided that they didn’t purchase them, and all the premiums missing values are replaced with 0.

The missing values of *MonthSalary* are replaced by the median, based on the assumption that an average of the population monthly salaries is a good representation in this instance. The *CustMonVal* and *ClaimsRate* don’t have any missing values.

The rest of the numerical features were imputed using KNN method, with 5 closest neighbours and their “uniform” weights, which would resolve all the missing values in *FirstPolYear* and *BirthYear.*

Lastly, all the missing categorical values were replaced with their corresponding modes, making an assumption, that the most frequent value is a meaningful approximation.

Shall we talk here about KNN method?

### Feature Engineering

The *FirstPolYear* and *BirthYear* are converted into *member\_age* and *age* respectively, by subtracting the values from 2016. The variables in the dataset are renamed accordingly, and metric and non\_metric features are updated.

### Outliers Removal

### 

# Title 2

## Title 2.1

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 2.1 – Illustrative figure

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.

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

Table 2.1 – Illustrative table

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.

### Title 2.1.1

Example of unnumbered list:

* Item 1
* Item 2
* Item 3

#### Title 2.1.1.1

Example of numbered list:

1. Item 1
2. Item 2
3. Item 3

# References

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

# Appendix 1 - Tables.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Additional Information** |
| ID | ID |  |
| First Policy | Year of the customer’s first policy | May be considered as the first year as a customer |
| Birthday | Customer’s first policy | The current year of the database is 2016 |
| Education | Academic Degree |  |
| Salary | Gross monthly salary (€) |  |
| Area | Living area | No further information provided about the meaning of the area codes |
| Children | Binary variable (Y=1) |  |
| CMV | Customer Monetary Value | Lifetime value = (annual profit from the customer) X (number of years that they are a customer) - (acquisition cost) |
| Claims | Claims Rate | Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years |
| Motor | Premiums (€) in LOB: Motor | Annual Premiums (2016). Negative  premiums may manifest reversals occurred in the current year, paid in previous one(s) |
| Household | Premiums (€) in LOB: Household |
| Health | Premiums (€) in LOB: Health |
| Life | Premiums (€) in LOB: Life |
| Work Compensation | Premiums (€) in LOB: Work Compensations |

Figure 1.1 – ABT table with variables description