NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

The use of business attributes in Motor Insurance Pricing

Case study of a Portuguese Insurance Company

Catarina Canelas Félix

Dissertation presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

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

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

THE USE OF BUSINESS ATTRIBUTES IN MOTOR INSURANCE PRICING

Case study of a Portuguese Insurance Company

by

Catarina Félix

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Risk Analysis Management

Advisor: Prof. Rui Alexandre Henriques Gonçalves

February 2019

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Prof. Rui Gonçalves, for all the help and support, his patience and friendship.

To my co-workers in the Insurance Company that made this work possible, thank you, your opinion and good mood were very important.

A very special gratitude to my family, specially my mother and brother for all the care and understanding. All of those long days and nights working on this dissertation wouldn't be possible without you.

I would also like to thank all my friends, in particular to Alina Teles, Bárbara Tavares, Breno Gonçalves, Carla Tavares and Flávia Moreira, for all the friendship and companionship they provided during my academic path.

To all, my deepest thank you.

Lastly, I would like to dedicate this dissertation to my loved grandfather, which I am sure is very proud and wishing to be here celebrating this moment with me.

ABSTRACT

The insurance activity offers their clients the opportunity to transfer risk in exchange of a fixed insurance premium. This premium should be enough to assure that the company will be able to respond to its possible future liabilities. These liabilities are, obviously, unknown when the premium is calculated, what means that they should be estimated a priori. However, different people belong to different risk classes, which leads to one of the big challenges of the insurance activity: the definition of a technically balanced tariff, or rate.

This dissertation has the objective to develop a pricing analytical model for Motor insurance using business factors and insured environment variables. In order to do so we need to treat our data and do exploratory analysis. After these preliminary steps we will construct two different models, one for claims' frequency and another for claims' severity using linear regressions in data mining. At the end of this work we pretend to indicate which variables explain better our data. The data we are using in this dissertation was provided by a Portuguese insurance company.

KEYWORDS

Insurance; Pricing; Premium; Analytical Models

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. BACKGROUND AND THEORETICAL FRAMEWORK

An insurance company has as principal objective to satisfy its clients for a future need in response to a fair premium. The insurance sector is characterized by its social role, his prudency and stability, which offers proper protection to companies and families correspondents' risks (ASF, 2016).

There are two main groups inside the insurance activity, the Life Insurance and the Non-life Insurance. This work will be related with Motor insurance, especially Motor Own Damage, which belongs to the Non-life group.

Nowadays the insurance sector is growing and recovering the economic activity in Portugal, with a positive impact in non-life group, especially in worker's compensation, motor and health, according with Insurance Portuguese Society (APS, 2017).

According with "Jornal Económico (2017)" the current challenges of the insurance sector are the insurance industry evolution into a new level of digital applications and machine learning techniques utilization. One of the main discussed themes is the use of analytical techniques to evaluate claims and prevent frauds.

Also, the insureds will have a new behaviour from now on, with the presence of Millennials into the market, what leads to more requirements, and make them to be more conscient of the risk and with more awareness of saving needs– exactly what characterize the Insurance sector (Económico, 2017).

The use of analytical models in this dissertation is due to the fact that the technology is advancing, and the companies need to go along to be actualized and modernized. The usual approach in insurance companies is to develop models using generalized linear models (GLM), that are a number of key ratios as dependent on a set of rating factors. (Gangam & Engelhardt, n.d.)

1.2. Study Relevance and Objectives

The number of cars in Portugal is increasing year by year (7,7% comparing 2016 with 2017) which leads to an increase of motor insurance policies (Caravela, 2017). Particularly, the policies with motor own damage insurance have increased by 9,7% their premium in 2017, when compared with 2016 (Companhia de Seguros Allianz Portugal, 2017). These facts turn motor own damage insurance pricing studies even more important and relevant nowadays.

Insurance companies would like to have a fairer premium price. Usually the premium is calculated with the data divided by groups with the same characteristics, however, this approach leads us to a portfolio where a huge quantity of policies will obtain the same level of risk, so it would be better if we calculate risk for smaller groups (Filler, 2012). More and more the companies want to have a personalized insurance premium, adequate to each person, and it owns characteristics.

This kind of premium is very important in regulatory and business level. Regulatory level because it is the best way to not have injustice and to have proper rates, and business level because it gives stability, simplicity and accuracy in the calculations, allowing to respond to any situation and to control any loss.

To develop this type of insurance pricing we will use in the present dissertation predictive analytics, where the outcome is a score for each individual policy (Filler, 2012).

McKinsey & Company (Columbus, 2017) performed a recent study relatively to artificial intelligence and companies profit margins. The study shows that companies who fully supported artificial intelligence initiatives have achieved 3 to 15 percentage point higher profit margin. Most of the business leaders who were interviewed for this survey expect margins to increase up to 5% points in the next year.

This dissertation aims to create a Motor Own Damage pricing analytical model with business attributes and insured environment variables that explain our portfolio behavior. We will develop this study with two different datasets: a dataset by policy and a dataset by policyholder (merging the policies for the same policyholder into only one line). We will also develop an exploratory analysis of various variables and develop linear regression models to predict claims' frequency and severity. In the end we pretend to combine the results of analytical models with the exploratory analysis we will perform.

2. LITERATURE REVIEW

2.1. ANALYTICS IN RATE MAKING

Rate making is the process where insurance companies try to predict the likelihood that a claim will be made and the total amount of claims in a policy period with the objective of pricing the products accordingly. In order to predict the above, the company must quantify the risks it is willing to assume and the premiums it will charge to assume them, having in consideration the overall goals of the company and the government regulations (Khopkar & King, 2007).

In insurance industry there is the need of setting the price of the product before knowing the cost of it. This fact makes the rate making process even more important and relevant. Insurance companies need to try to estimate how many claims will occur and how large these claims will be. The need to compete on price with the industry is also an encouragement to the rate making development. Rate making is an area where the entire structure of organisations should leverage it for competitive advantage (Khopkar & King, 2007).

Usually the ratemaking models are developed using Generalized Linear Models (GLMs), which will be explained in next sub-chapter, that are the traditional statistical approaches using statistical distributions. There are other approaches that can be used, such as the use of analytical models, which is a relatively new approach.

The recent regulation and international competition have made possible for insurance providers to offer innovative rate structures geared toward attracting and rewarding good customers and avoiding at the same time the bad customers as much as possible. these mentioned rate structures differ from the traditional ones because they consider more risk factors than the ones considered by the traditional structure (SAS Institute, 2003). According with SAS (2003), a study found out that drivers who spend time working on their own cars tend to have less accident costs. These personal hobbies are not being used traditionally in the rate setting. Starting to use these kind of variables opens up the number of potential descriptors that could be used in defining risk groups. Regulatory requirements and data privacy may not enable the use of some of these variables, however, more descriptors will be used in rate making than before.

Data mining is the chosen method for managing the complexity introduced by using the additional variables. More significative variables for the accident behaviour can be found using predictive modelling and produce smaller and more homogeneous subgroups of drivers. “Rate setting will involve the determination of many niches of good and bad customers and result in rules that can characterize these various groups. This shifts the emphasis of the industry from the problem of “determining the optimal premium” to the issue of “improving customer relationship”: identify low risk customers, adjust their premiums in order to win their loyalty, improve customer retention and increase market share.” (SAS Institute, 2003) Data mining is also a good choice to monitor the effectiveness of these goals and the performance of the several rules developed for setting the insurance premiums.

The usual steps to construct the rate making process are:

1. Collect all the relevant data
    - Check and clean it

2. Explore data
    - Determine relationships and trends
    - Transform values and define derived variables
3. Compute several models
    - For example, decision trees, regressions, neural networks.
4. Assess results
    - Analyze the results from technical and business perspective

The advantage of using Data Mining into rate making (or pricing) processes is the possibility to use many more initial variables than in the traditional approaches, the data determines which variables are significant to use, computes intervals and category groups, it is faster than the traditional ones, and can also do the same as them (Drews, 2000).

2.2. GENERALIZED LINEAR MODELS

The Generalized linear Models (GLMs) have become a standard approach for non-life insurance pricing. The GLMs are an extension of the Gaussian linear models' framework that is derived from the exponential family. The objective of these models is to estimate an interest variable (Y) depending on a certain number of explanatory variables (X_i) (David, 2015).

The variable Y can be a binary variable, a countable variable or a real positive variable. In the case of being a binary variable it can only have the value zero or one. If it is a count variable the values will belong to the set of natural numbers. On the other hand, if it is a real positive variable its values belong to the set of positive real numbers.

"Conditioned on the explanatory variables (X_i), the random variables ($Y_1, Y_2, \dots, Y_n$) are considered to be independently, but not identically distributed, that have the probability density generated by the expression:

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right), y_i \epsilon S$$

where S represents a subassembly that belongs to $\mathbb{N}$ or $\mathbb{R}$ set, θ_i is the natural parameter and ϕ is the scale parameter. In binomial and poisson distributions, the scale parameter has the value 1, and for Gamma distribution it is unknown and has to be estimated" (David, 2015).

The probability density function of the variables ($Y_1, Y_2, \dots, Y_n$) can be defined using the following:

$$f(y|\theta, \phi) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi) = \exp\left(\frac{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi)\right)$$

The objective of this model is to obtain the expected values of the dependent variables over conditional means, given independent observations. Here, there are the parameters $\beta_1, \beta_2, \dots, \beta_n$, through a function ($g$) of the dependent variable mean (μ_i), written as a linear combination of the variables (X_i):

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{n} \beta_i x_{ij} = x_i^t \beta = \eta_i$$

The monotonous and differentiable function g is known as a link function since it connects the linear predictor η_i with the mean μ_i.

As the objective of this dissertation is to calculate the premium, there is the need to develop the frequency estimation model and the claims' cost estimation model (David, 2015).

According with Antonio & Valdez (2012) the frequency of claims has a good modelling if it is modelled by the Poisson distribution. The frequency of the claims is a discrete variable, what satisfies the Poisson assumptions. On other hand, the claims' cost is usually modelled by the Gamma distribution (David, 2015).

The pure premium stands for the expected cost of all claims occurred in the insured period, which is calculated by statistical methods incorporating all the available information. The calculation of the pure premium is obtained by multiplying the claims' frequency and the claims' cost since the pure premium is the mathematical expectation of the annual cost claims declared by the policyholders. The following formula refers to the pure premium calculation:

$$E\left[\sum_{i=1}^{N} C_i\right] = E[Y] \times E[C_i]$$

For the claims' amount ($C_1, C_2, \ldots$) independent of their number (Y).

Calculating the claims' frequency and the claims' cost is very relevant because the risk factors that influence the two components of the pure premium are not the same for both. The separate analysis allows us to have a clearer perspective on which and how the risk factors are manipulating the premium (David, 2015).

2.3. ANALYTICAL MODELS

Advanced analytics applied in insurance allows us to find new growth opportunities and protect and optimize companies' enterprise (Ernst & Young, 2013).

According with Ernst & Young (2013), the three questions insurers asks are:

1. What more can our own data tell us?
2. What else could we learn if added external data to our model?
3. How can we build the power of analytics into day-to-day decisions?

With this work we pretend to answer the first and second questions.

"In personal auto insurance, big data is making a big difference. Traditionally, underwriters have developed auto insurance prices based on smaller data — such as the car's make, model, and manufacturer's suggested retail price (MSRP). But "bigger data" is now available, providing far more information and allowing insurers to price policies with a better understanding of the vehicle's safety. From manufacturers and third-party vendors, insurers can learn about a car's horsepower, weight, bumper height, crash test ratings, and safety features. That big data helps insurers create sophisticated predictive models and more accurate vehicle-based rate segmentation." (Cummings, n.d.)

The analytics performed by actuaries are extremely significant to insurance companies existence and profitability (Clarke & Libarikian, 2014). Motor insurers started to add behavior-based credit scores into their analysis instead of only have in consideration historical data. Insurers also have become aware (by empirical evidence) that people who pay their bills on time are also safer drivers (Clarke & Libarikian, 2014). All these facts complement our objective of having in consideration external data.

In order to make statistical predictions, insurance companies rely on "The Law of Large Numbers". This law says that **"**as the number of identically distributed, randomly generated variables increases, their sample mean (average) approaches their theoretical mean"(Routledge, n.d.). According with this law insurance companies are not able to predict individual accidents, but they also don't need to, they only need to know how the behavior will be in general. This approach is good for the company but not so good for the policyholder, because a good driver might end up in a portfolio with a lot of bad drivers and all the premiums will be the same, which is unfair (Agababa, n.d.). Therefore insurers developed a unique way to extract actionable insights from Data Analytics to track individual policyholder behavior and price policies accordingly. (Agababa, n.d.)

Insurance companies could use predictive modeling and analytics to predict the probability of a policyholder having an accident or having their car stolen, for example. The use of this predictive models and the information obtained with that give to insurers an extra knowledge about their portfolio. Insurers can gain a lot by monitoring policyholders driving habits, behaviors and routines, and then can compare them against other policyholders in their portfolio. For motor insurance is easier to get this information, the insurers only need a small box installed inside vehicles or even an app downloaded into the policyholder smartphone. Portugal is starting to embrace this technological advance and ideas, and there is already one Portuguese insurance company with an app that evaluates policyholders driving behavior.

Data analytics has a several number of models and approaches, and one of them is the linear regression model, that will be explained in the next sub-chapter.

2.4. LINEAR REGRESSION MODELS

In linear regression models is assumed that the dependence of Y on X_1, X_2, … , X_p is linear. Linear regression is extremely useful conceptually and practically nowadays. Due to linearity characteristic, the following model is assumed:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 and β_1 are two unknown constants that represent the intercept and the parameters, respectively, and ϵ is the error term (Stanford, n.d.).

In linear regression models an estimation for the target variable is formed from a simple linear combination of inputs, as we can see in the following formula:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

where $\widehat{\beta_0}$ corresponds to the intercept estimate, and $\widehat{\beta_1}$ to the parameters estimate. The intercept centers the range of predictions, and the remaining parameter estimates determine the trend strength (or slope) between each input and the target (SAS Institute, 2017).

The objective is to minimize the squared error function:

$$\sum (y_i - \hat{y}_i)^2 = \sum (\epsilon_i)^2$$

The intercept and parameter estimates are chosen in order to minimize the squared error between the predicted and the observed target values. The prediction estimates can be viewed as a linear approximation to the expected value of a target conditioned on observed input values (SAS Institute, 2017).

In this dissertation the models that will be used are the linear regression models in data analytics due to their simplicity and easier comprehension by external parties.

2.5. INPUT SELECTION

Input selection is a task that all predictive models should perform. A way to find the best set of inputs for a regression model is to try every single combination of inputs. However, the number of models that we would obtain having this approach increases exponentially in the number of available inputs, and this strategy is impractical for real prediction problems.

Alternatively, we can restrict the variables search to a sequence of improving models. This method is usually used to find models with good predictive performance, even when it may not find the single best model. There are three main sequential selection methods: forward, backward and stepwise methods (SAS Institute, 2017).

2.5.1. Forward selection

Forward selection method creates a sequence of models that increase complexity. On other words, the sequence starts with the baseline model (which is a model predicting the overall average target value for all cases), and then the algorithm searches the set of one-input models and selects the model that most improves on the baseline model. Then, it searches the set of two-input models that contain the input selected in the previous step and selects the model showing the most significant improvement. A sequence of increasingly complex models is generated when we add a new input to those selected in the previous step. The sequence ends when no significant improvement can be made.

The p-value, the usual statistic measure of significance, quantifies that improvement. When terms are added it always increases the model's overall fit statistic. The p-value can be calculated when the change in the fit statistic is calculated and assuming that this change conforms to a chi-squared distribution. A large fit statistic change, which corresponds to a large chi-squared value, is unlikely. This way, a small p-value indicates a significant improvement, and when no p-value is below a predetermined entry cutoff, the forward selection procedure finishes (SAS Institute, 2017).

2.5.2. Backward selection

In opposition to forward selection, the backward selection creates a sequence of models of decreasing complexity. It starts with all the available inputs, and therefore, has the highest possible fit statistic. Inputs are sequentially removed from the model, and, at each step, the input with the highest p-value is removed. The sequence ends when all the inputs have a p-value that is less than the predetermined stay cutoff (SAS Institute, 2017).

2.5.3. Stepwise selection

The stepwise selection combines elements from the forward and the backward selection procedures. The method begins as in the forward method and adds sequentially the inputs with the smallest p-value below the entry cutoff. However, after each input addition, the algorithm revaluates the statistical significance of all the included inputs. If the p-value of any of the included inputs exceeds the stay cutoff, the input is removed from the model and re-entered into the set of inputs available for inclusion in a subsequent step. The process ends when all the inputs available for inclusion in the model have p-values in excess of the entry cutoff and all the inputs included in the model have p-values below the stay cutoff (SAS Institute, 2017).

3. METHODOLOGY

In this work we intend to develop an insurance pricing model. For that we will construct two models, one regarding to the frequency of the claims and other regarding their severity (average cost). However, before that, we need to collect the data that will be used into this dissertation and treating it.

The difficulty of data collection process is sometimes underestimated, but it is a hard process. With security, privacy and cost issues in getting the data, it is also a time-consuming process, so, in order to solve this problem, most researchers generate synthetic data. Our intention with this dissertation was to work with real data and have the possibility to analyze Portuguese data.

When the raw data is collected it is necessary to be treated, see if more variables are required to use in our study and conduct an exploratory analysis. After this we are able to go through the construction of the two different models we will build: claims' frequency model and claims' severity model.

3.1. DATA TREATMENT

The data used in this work was provided by a Portuguese Insurance Company and the period under consideration is between 01/01/2015 and 31/12/2017 (three years). Our data corresponds to Motor Own Damage policies that were in force in the considered period.

The original data was composed by four data bases, corresponding each one to the claims occurred in 2015, 2016, 2017 and the policies that were in force at least one day since the beginning of the company. It had the following composition:

Data	Number of Rows	Number of Variables/Columns
2017 Policies	326.399	100
2015 Claims	1.280	83
2016 Claims	2.089	87
2017 Claims	3.255	91

Table 1 – Datasets provided by the Company

The policies data has more than one row for policy, because it is subdivided in policies and their correspondent risk. It was necessary to clean our data, as cleaning the variables that were not relevant (for example payment modality, tariff, birthday date, etc.) and merge all the rows of one policy into one row. Then, there was also the need to aggregate the claims to the correspondent policy. While we were building up the final dataset it was clear that the policies address and the claims' address were missing, such as the annulation date and the replacement date. These variables are necessary for the calculation of risk exposure (annulation date and replacement date) and to study the relation between the claims' location and the policy address in the system. To get a complete dataset, we requested the new variables to the company.

In order to make a good analysis to our data we created eleven variables:

Variables
Risk exposure
Number of claims in the residence area in 2015
Number of claims outside the residence area in 2015
Number of claims in the residence area in 2016
Number of claims outside the residence area in 2016
Number of claims in the residence area in 2017
Number of claims outside the residence area in 2017
Number of claims occurred in the 3 years of the study
Claims' cost in the 3 years of study
Claims' frequency
Claims' severity (cost per claim)

Table 2 – Calculated variables

The risk exposure was tricky to calculate because the data was not very clear, and we needed to have in consideration various scenarios (with or without annulation date and replacement date). The following 10 scenarios represent the method of risk exposure calculation.

Scenario 1

For the policies with annulation date before 01/01/2015 which were not replaced

$$risk\ exposure = 0$$

Scenario 2

For the policies with annulation date before 01/01/2015 which were replaced but with replacement date before the annulation date

$$risk\ exposure = 0$$

Scenario 3

For the policies with annulation date before 01/01/2015 which were replaced before 01/01/2015 and after the annulation date

$$risk\ exposure = 1$$

Scenario 4

For the policies with annulation date before 01/01/2015 and were replaced

$$risk\ exposure = \frac{31\ Dec\ 2017 - Replacement\ Date + 1}{31\ Dec\ 2017\ -\ 01\ Jan\ 2015 + 1}$$

Scenario 5

For the policies with beginning date before 01/01/2015 with no annulation date

$$risk\ exposure = 1$$

Scenario 6

For the policies with beginning date before 01/01/2015, annulation date before 31/12/2017 and no replacement date or replacement date after 31/12/2017

$$risk\ exposure = \frac{Annulation\ Date - 01\ Jan\ 2015 + 1}{31\ Dec\ 2017\ -\ 01\ Jan\ 2015 + 1}$$

Scenario 7

For the policies with beginning date before 01/01/2015, annulation date before 31/12/2017 and replacement date before 31/12/2017 and greater than the annulation date

$$risk\ exposure = \frac{Annulation\ Date - 01\ Jan\ 2015 + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1} + \frac{Replacement\ Date - Annulation\ Date + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1}$$

Scenario 8

For the policies with beginning date after 01/01/2015, annulation date before 31/12/2017 and replacement date before 31/12/2017 and greater than the annulation date

$$risk\ exposure = \frac{Annulation\ Date - Beginning\ Date + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1} + \frac{Replacement\ Date - Annulation\ Date + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1}$$

Scenario 9

For the policies with beginning date after 01/01/2015, annulation date before 31/12/2017 and replacement date before 31/12/2017 and greater than the annulation date

$$risk\ exposure = \frac{Annulation\ Date - Beginning\ Date + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1} + \frac{Replacement\ Date - Annulation\ Date + 1}{31\ Dec\ 2017 - 01\ Jan\ 2015 + 1}$$

Scenario 10

For the policies with beginning date after 01/01/2015 and no annulation date or annulation date after 31/12/2017

$$risk\ exposure = \frac{31\ Dec\ 2017 - Beginning\ Date + 1}{31\ Dec\ 2017\ -\ 01\ Jan\ 2015 + 1}$$

The frequency and the severity of the claims were calculated with the following formulas:

$$frequency = \frac{Number\ of\ claims\ occured\ in\ the\ 3\ years}{Risk\ exposure * 3}$$

$$severity = \frac{Claims\ cost\ in\ the\ 3\ years}{Number\ of\ claims\ occured\ in\ the\ 3\ years}$$

The frequency variable will be used to develop some graphs to perform an exploratory analysis that will take place in the sub-chapter 3.2. The severity variables, besides it is going to be used in same analysis as the frequency, it will also be the target variable for the severity model we will perform in the sub-chapter 3.3.

Then, to meet this dissertation objective of construct an insurance premium with the help of variables external to the policyholder or policy object we also need external data. This data was extracted from INE (Portuguese Statistical Institute) and is composed by the medium salary from year 2015 and 2016 by home area (locality) in one dataset and male and female present population also by home area in another dataset.

The next step was to aggregate the external data, aggregated by locality, into our dataset. However, these variables did not have an immediate aggregation since they were by locality and the dataset did not have the locality variable corrected and filled in some cases. In order to make a proper aggregation of the data we needed to use the postcode because it was the most accurate geographic variable we had in the dataset, and then we needed to find out of which locality the postcode refers to. Because of that, we were forced to extract data from CTT (Portugal Post Office Institution) to build the connection between the postcodes and the localities.

The CTT data was composed by 3 datasets. The first only had the district code (DD) and the correspondent description. The second had the district code (DD), county code (CC), and the correspondent county descriptive. The third dataset was more complex and had seventeen variables, which we only have used six, the district code (DD), county code (CC), locality code (LLLL), locality description, four digits postcode (CP4), 3 digits postcode (CP3).

Our Dataset	CTT first Dataset	CTT second Dataset	CTT third Dataset	INE Dataset
•CP4 •CP3	•District code •District description	•District code •County code •County description	•CP4 •CP3 •Locality code •Locality description •County code •District code	•2015 Gains •2014 Gains •Present population (male) •Present population (female)

Figure 1 – External datasets composition

At this point we were able to match the datasets from the insurance company, CTT and INE. We used the postcode (CP4 and CP3) to match the CTT dataset to ours, and then the county code and district code to have in our dataset the county and district descriptions. The following scheme illustrates how did we match the datasets:

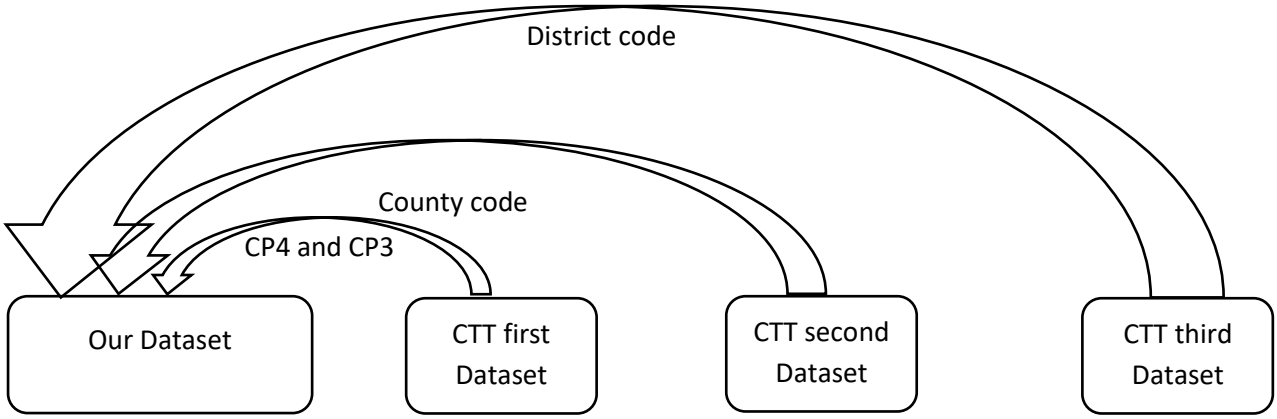

Figure 2 – External data aggregation

After the match between our data and CTT data, we stayed with the following datasets:

Our Dataset
•CP4 •CP3 •District code •District description •County code •County description •Locality code •Locality description

INE Dataset
•2015 Gains •2014 Gains •Present population (male) •Present population (female)

Table 3 – Datasets composition

After matching the CTT datasets with our we needed to add also INE dataset into it, using the variable locality. But there was a factor to consider: the localities we obtained through the CTT data could not exist in the INE data. Having this possibility, we couldn't use the locality as a matching variable for all the policies. We decided that the locality to use would be the CTT locality if it was in INE data, if not, the locality to use would be the CTT county if the county was in INE data, and if not, the locality to use would be the CTT district, if it was in INE data, that was in INE data in the most of the cases, as we can see in the scheme that follows:

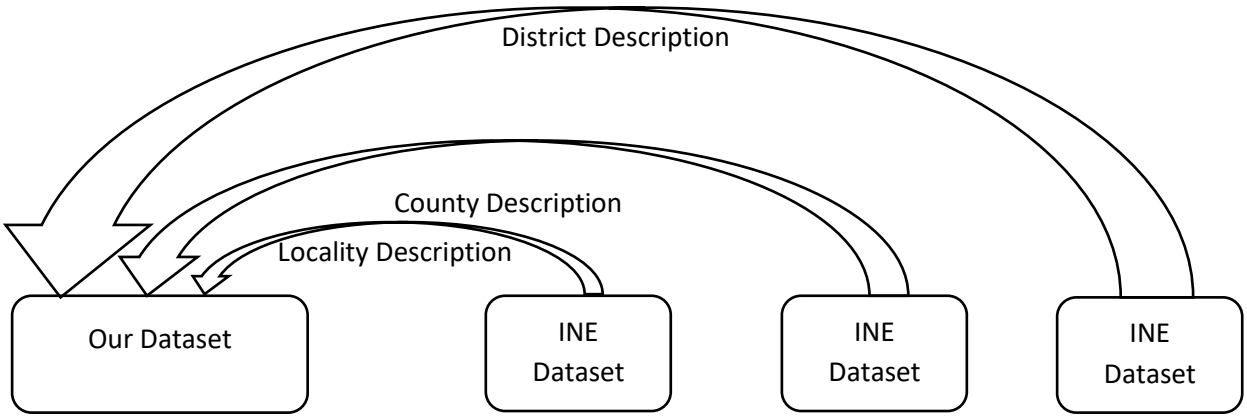

Figure 3 – Descriptions agregation

However, there were some cases that even with this approach we couldn't get a locality to use because of the postcode in company dataset (it was incomplete, or missing, and even sometimes the same postcode is associated to different localities). In those cases, around 200, we searched for the correct locality manually using then the locality to aggregate INE data to them.

After having all this information aggregated, we decided to not have in consideration the company policy, foreign policies, and the policies that were not from particulars. We made this decision because INE data is only related with Portuguese people, and not companies or foreign people. We also deleted from the data the polices with the variable sex equals to 0 (does not have a proper meaning), policies with beginning date after 31/12/2017 that were already in the system, and the policies where the driver was younger than 18 years old and older than 99 years old. We only stayed with the policies with car seats equal to 5, vehicle use being particulars, and claims' costs greater or equal to 0 (since we don't want to analyze negative costs). With the manipulation mentioned before we ended up with 27 variables and 12983 rows.

The final data has 27 variables divided in five types: the vehicle variables, the policyholder characteristics variables, calculated variables, claims' variables and the external variables.

Vehicle variables	Policyholder variables	Calculated variables	Claims' variables	External variables
•Capital •Type of vehicle •Cubic capacity •Vehicle years •Fuel	•Policy ID number •Policyholder code •County •District •Region •Driver's age •Driving license years •Sex •Beginning date •Annulation date •Replacement date	•Risk Exposure •Frequency •Severity	•Number of claims in 2015, 2016 and 2017 •Number of claims in home area in 2015, 2016 and 2017 •Number of claims outside of home area in 2015, 2016 and 2017 •Claims cost in 2015, 2016 and 2017	•2015's average monthly salary •2014's average monthly salary •Male present population •Female present population

Table 4 – Treated dataset variables

Our data have the following structure:

Policies number	Claims' Frequency	Claims' Severity
12983	12,6%	1647,5€

Table 5 – Dataset structure

We will create two different models in this dissertation. The first to estimate the absolute frequency of the claims and the second to estimate the claims' severity. In order to do so, there was the need of construct class variables for the following variables: driver's age, driving license years, vehicle's age,

cubic capacity, capital, average 2015 mensal gains, present population male, present population female, frequency and severity.

We created a new variable called Frequency_Class which represents the classes where we divided the absolute frequency variable (number of claims). The following table explains how we made the division:

Frequency_class	Frequency interval
0	$[0]$
1	$[1]$
2	$[2]$
3	≥ 3

Table 6 – Frequency classes

The same principle occurred to the severity variable, and we created also another variable (Severity_Class) to categorize it.

Severity_class	Severity interval
0	$[0]$
1	$]0, 1500]$
2	$]1500, 3000]$
3	$]3000, 6000]$
4	> 6000

Table 7 – Severity classes

The partition in classes performed here was discussed with the experts in the insurance company that provided the data. Together we develop these partitions created the new classes variables, not only for the frequency and severity variables already seen, but also for the other variables we will see in the sub-chapter 3.2.

3.2. EXPLORATORY ANALYSIS

In order to analyse how is the behaviour of our portfolio, it is essential to analyse each relevant variable in our data set comparing with the frequency and severity.

Driver's age

The variable of driver's age, called "N_years_driver" presented values between 18 and 90, because we made this filter to not have in consideration the outliers and we made the following adjustment to create the new variable "N_years_driver_class" of driver's age class:

N_years_driver_class	Driver's age interval
1	≤ 25
2	$]25, 35]$
3	$]35, 50]$
4	$]50, 65]$
5	> 65

Table 8 – Driver's age classes

The first graph in the figure 4 represents the claims' frequency for each class:

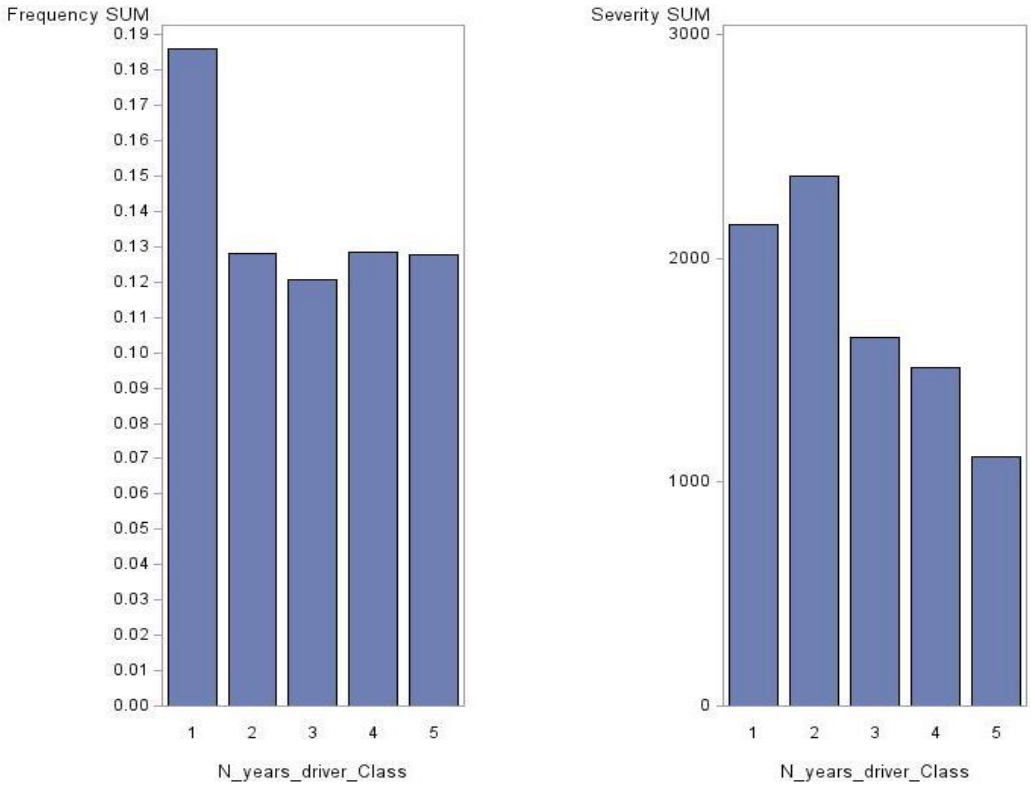

Figure 4 – Frequency and severity by driver's age classes

Analysing this graph, it is possible to see the significative differences between the first class and the others. The first class, that corresponds to the drivers with 25 years old or less, is the class with the highest frequency, achieving the value of 18,6%.

Regarding the severity, it is possible to see in the second graph of figure 4 that the first and second class of driver's age (the youngest drivers) are the ones that have the more expensive claims.

Driving license years

The variable of driving license years, called "N_years_driving_license" took values between 0 and 69, and such as the driver's age variable, it needed to be divided in classes. The class partition was the following:

N_years_driving_license_class	Driving license's years interval
1	≤ 5
2	$]5, 15]$
3	$]15, 30]$
4	$]30, 45]$
5	> 45

Table 9 – Driving license classes

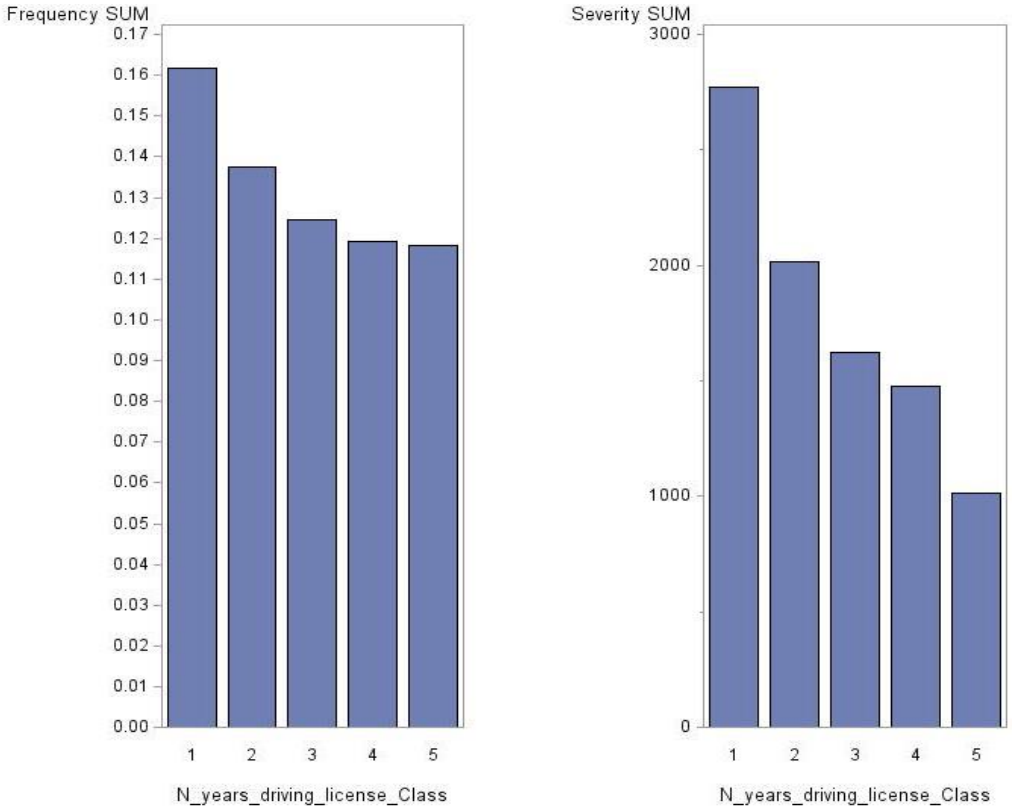

Figure 5 – Frequency and severity by driving license years classes

According with the frequency graph on figure 5 above, the class 1 (driving license years less or equal to 5) is the one with more frequency (claims by risk exposure), and the severity is decreasing while the drinving license years are increasing. The same occurs with the severity (visible in the figure 5 also), having the first class the value of 2771,91€ and the fifth class the value of 1012,02€.

Vehicle's age

The variable of vehicle's years, called "N_years_vehicle", had values between 0 and 44 years. A categorization was also needed for this variable, and was performed as follows:

N_years_vehicle_ class	Vehicle's age interval
1	≤ 5
2	$]5, 10]$
3	$]10, 15]$
4	> 15

Table 10 – Vehicle years classes

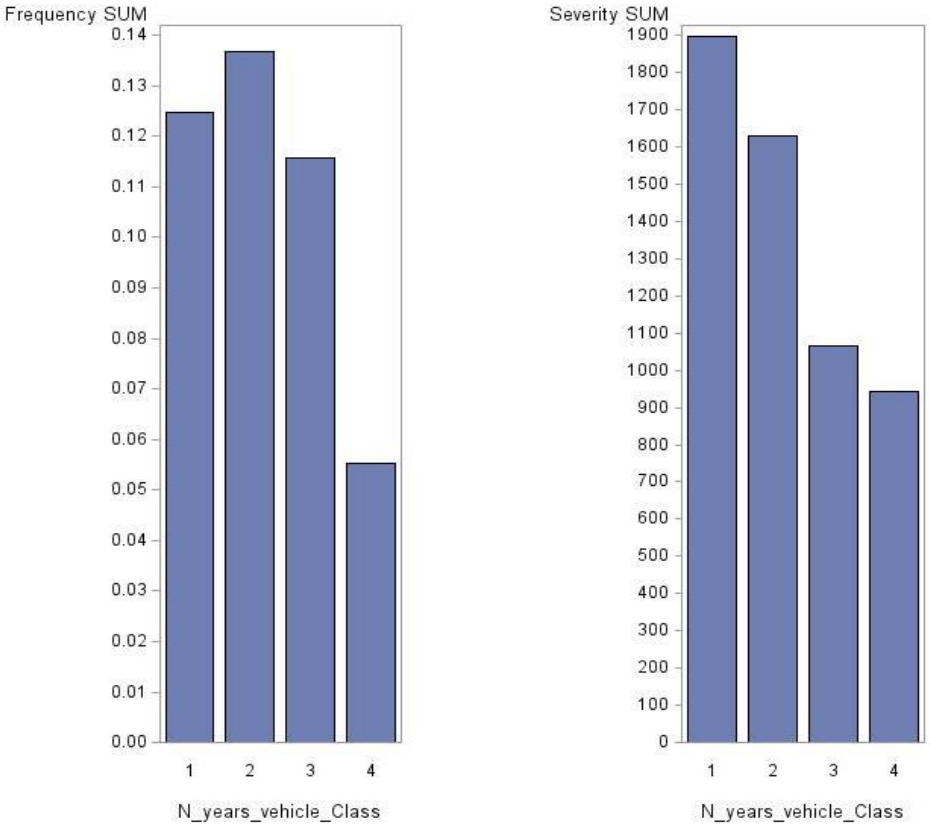

Figure 6 – Frequency and severity by vehicle years classes

The frequency of the claims is the highest for the class 2, vehicle years between 6 and 10 years, followed by class 1, vehicle years less or equal than 5 years. The claims for class 1 and 2 are also the more expensive ones, according with the graph below, and it decreases as long as the vehicle years increase.

Fuel

The GPL/GNC and Hybrid fuel policies are the ones with the highest claims' frequency, as we can see in the figure 8, however these policies are a small part of our portfolio, so these values are overestimated. Between the diesel and gasoline policies, diesel policies have more claims' frequency than gasoline policies. For the severity, also the diesel policies are more expensive than the gasoline policies (figure 9).

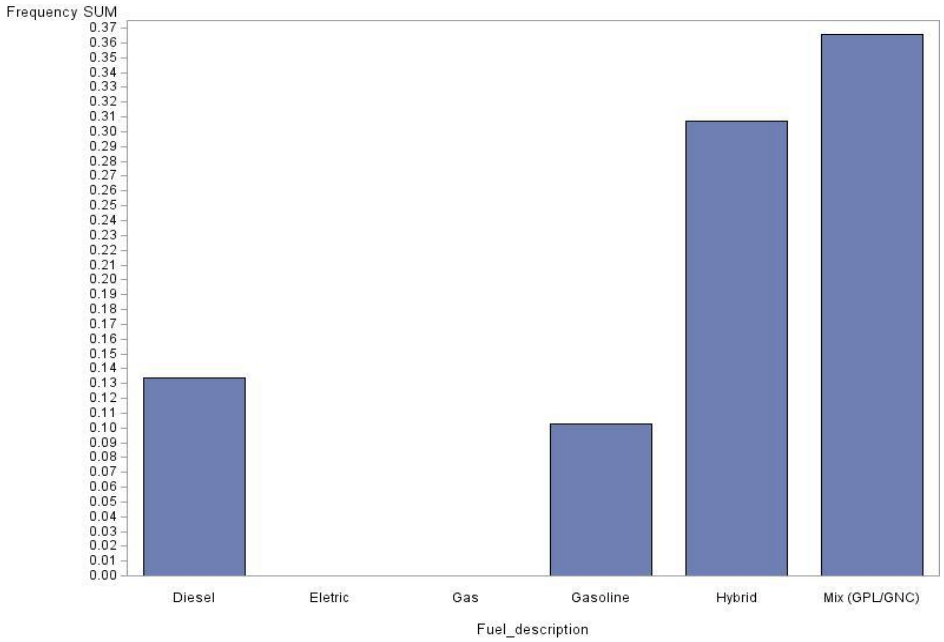

Figure 7 – Frequency by type of fuel

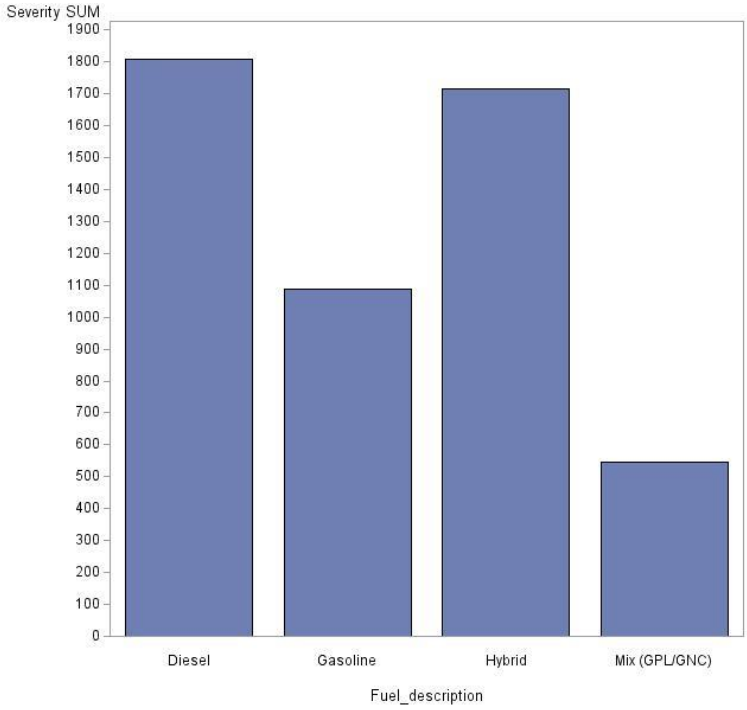

Figure 8 – Severity by type of fuel

Sex

Analysing the variable "Sex" it is possible to see that the difference between male and female drivers is not relevant, although male drivers have less claims than female but more expensive, as demonstrated in the graphs below.

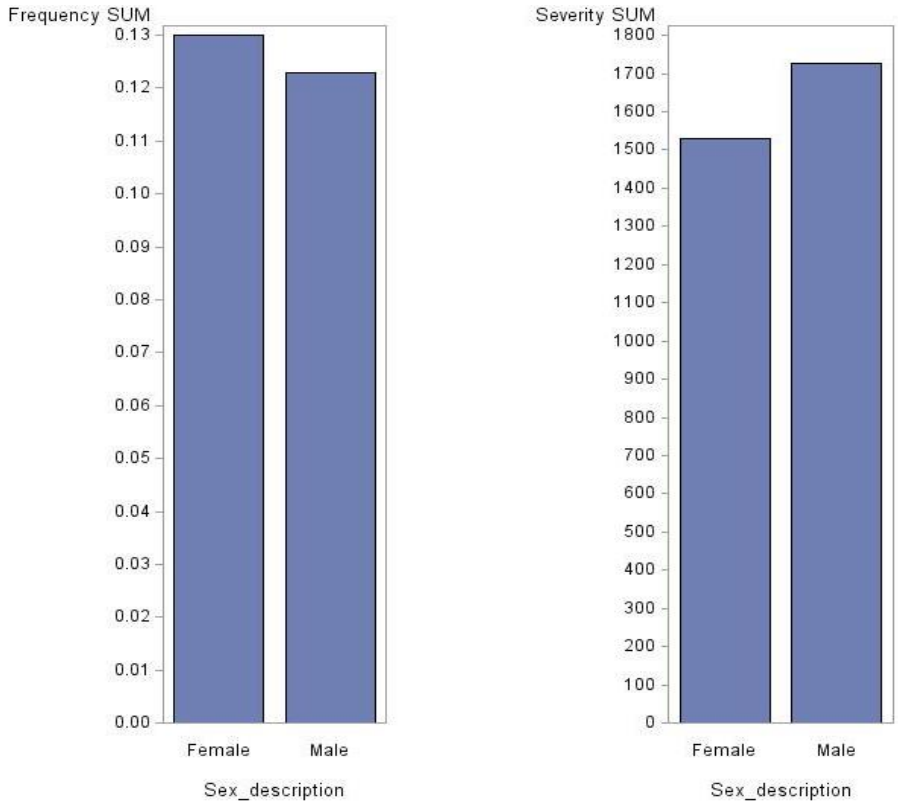

Figure 9 – Frequency and severity by sex

Capital

The Capital variable is important when the portfolio is related with motor own damage, since the claims are directly related into it, such as the premium. To have a proper analyse using this variable we divided it into classes as well:

Capital_class	Capital interval
1	≤ 10000
2	$]10000, 20000]$
3	$]20000, 30000]$
4	$]30000, 40000]$
5	$]40000, 50000]$
6	> 50000

Table 11 – Capital classes

According with the claims' frequency graph in figure 10, policies with more capital means policies with higher frequency rates.

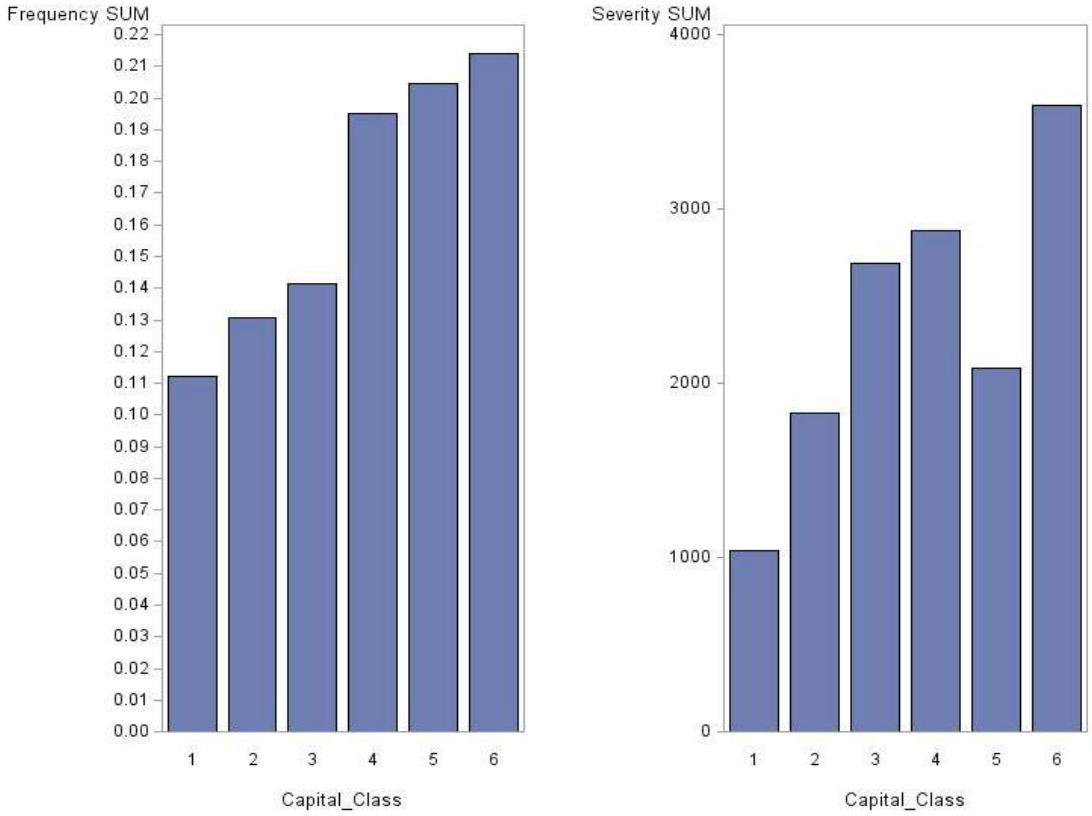

Figure 10 – Frequency and severity by capital classes

The same happens for the severity, in exception of class 5, policies with capital between 40.000€ and 50.000€.

Cubic Capacity

The variable of vehicle cubic capacity is relevant to be an input to the model, so we also divided it into classes:

Cubic_Capacity_class	Cubic Capacity interval
1	≤ 1000
2	$]1000, 1400]$
3	$]1400, 1600]$
4	$]1600, 2000]$
5	> 2000

Table 12 – Cubic capacity classes

The claims' frequency graph below shows us that as long the cubic capacity is growing, the frequency is also growing. For the severity analyse we have the same conclusion, in exception of class 2, that is less than the first class.

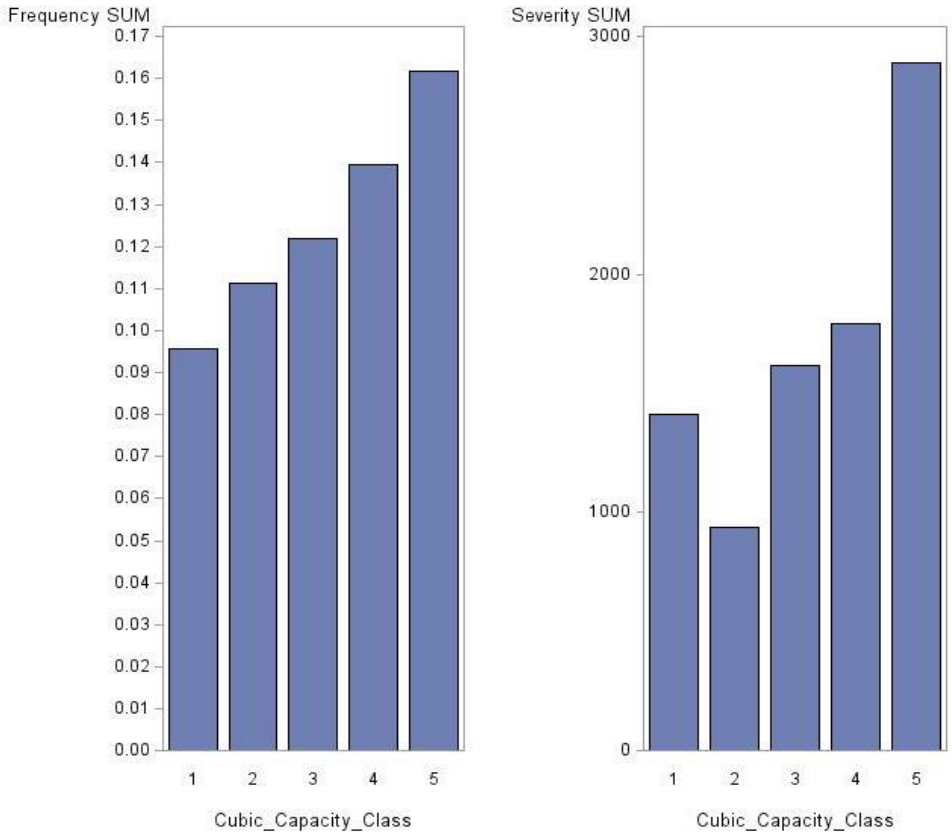

Figure 11 – Frequency and severity by cubic capacity classes

2015's average monthly salary

The variable of 2015's average monthly salary called "Gains_2015" that was extracted from INE data had the following class division:

Gains_2015_class	Gains 2015 interval
1	≤ 800
2	$]800, 1000]$
3	$]1000, 1200]$
4	$]1200, 1400]$
5	$]1400, 1600]$
6	$]1600, 1800]$
7	> 1800

Table 13 – 2015's average monthly salary classes

Observing the claims' frequency graph on figure 12 does not look that the average monthly salary has any impact on the claims' frequency of the policies.

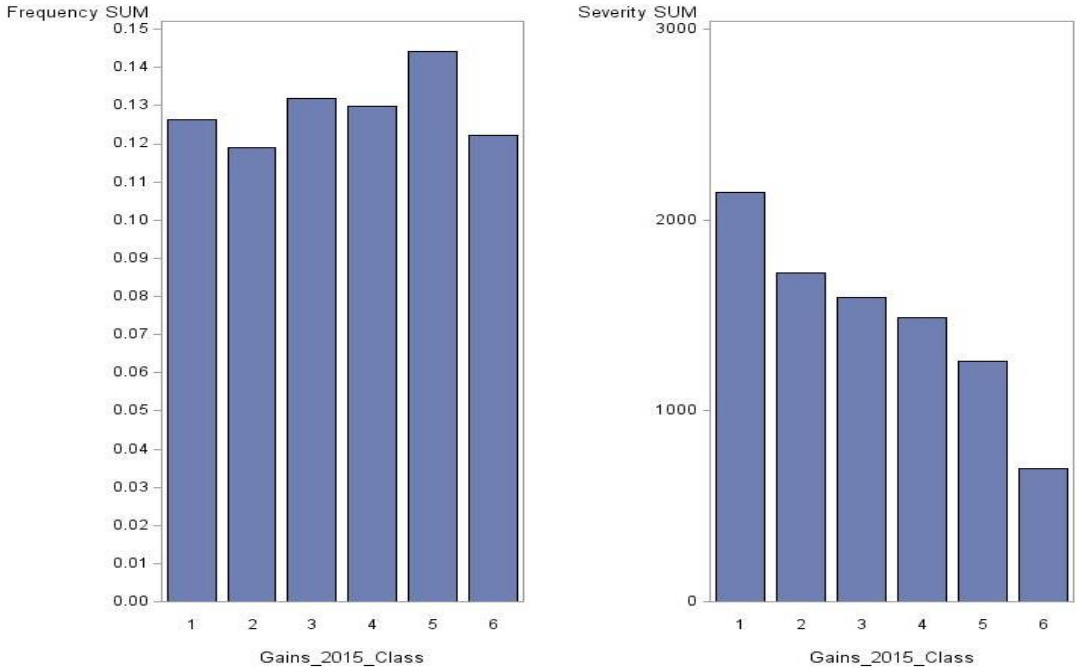

Figure 12 – Frequency and severity by average monthly 2015 salary classes

On other hand, for the severity, it is possible to see, in the severity graph of figure 12 above, that as long as the salary classes are increasing the severity is decreasing, being the class 6, when the average monthly salary is greater than 1.600€ and below or equal 1.800€, the lowest severity.

Female and male present population

The variables of female and male present population called respectively "Present_Population_female" and "Present_Population_male" were divided into the following classes:

Present_Population_male_ class	Number of present population (male) interval
1	≤ 8000
2	$]8000, 16000]$
3	$]16000, 30000]$
4	$]30000, 50000]$
5	$]50000, 100000]$
6	$]100000, 150000]$
7	> 150000

Table 14 – Male present population classes

Present_Population_female_ class	Number of present population (female) interval
1	≤ 8000
2	$]8000, 16000]$
3	$]16000, 30000]$
4	$]30000, 50000]$
5	$]50000, 100000]$
6	$]100000, 150000]$
7	> 150000

Table 15 – Female present population classes

Observing the graphs in figure 13, there is a relation between the male present population and the claims' frequency. We can see that the frequency decreases from class 1 to class 3 and then it increases from class 3 to class 6, and the class 3, population between 16.000 and 30.000 male people, has the lowest frequency. For the severity there isn't a relation between those variables.

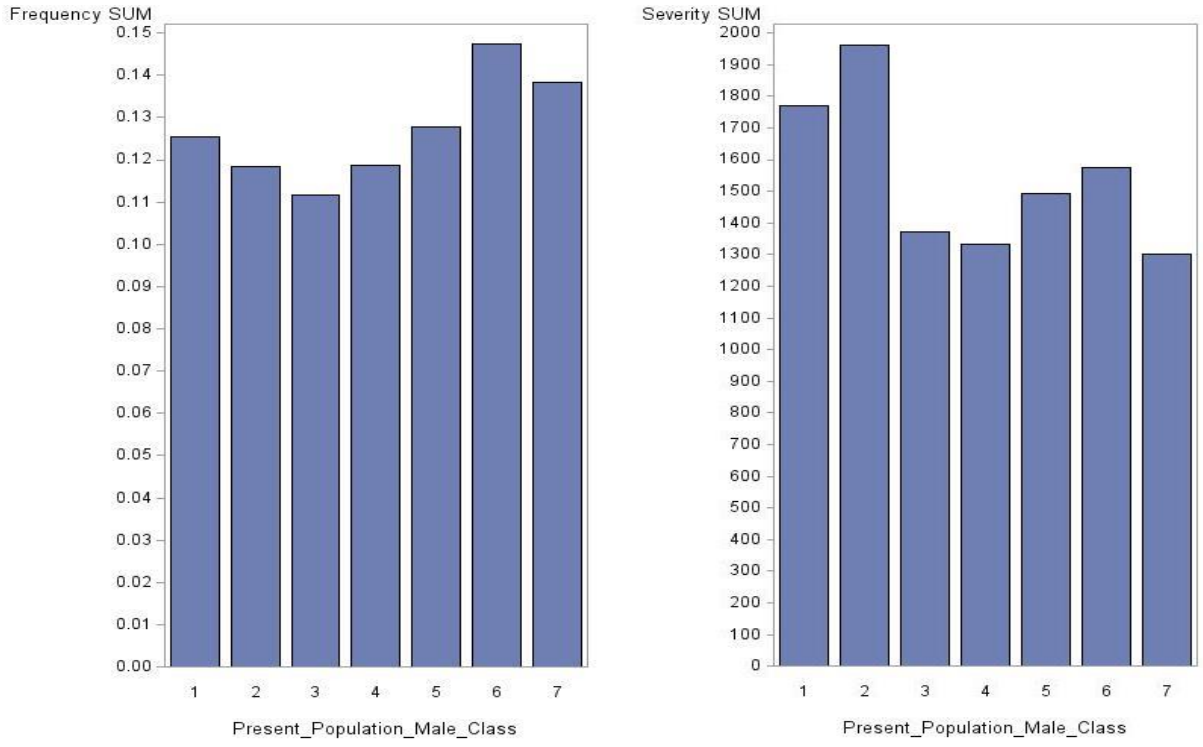

Figure 13 – Frequency and severity by male present population classes

For the female approach the relationship isn't so obvious, but we can say that the frequency decreases until class 4 and then it increases till class 6 too. For the severity there isn't a relation between those variables.

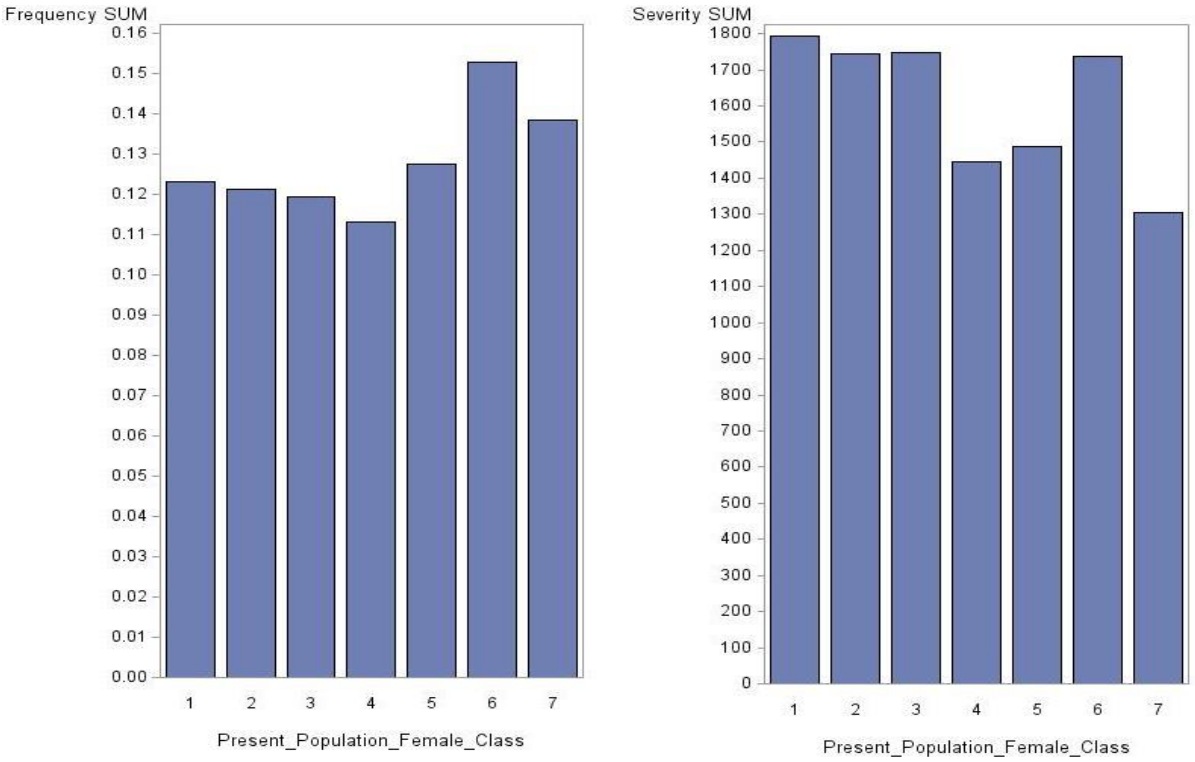

Figure 14 – Frequency and severity by female present population classes

Number of claims in or out the policyholder home area by district

The claims occurred for each policy could be in or out the policyholder home area. Analysing the distribution by district distinguishing the claims occurred in or out the home area we ended up with the following graph:

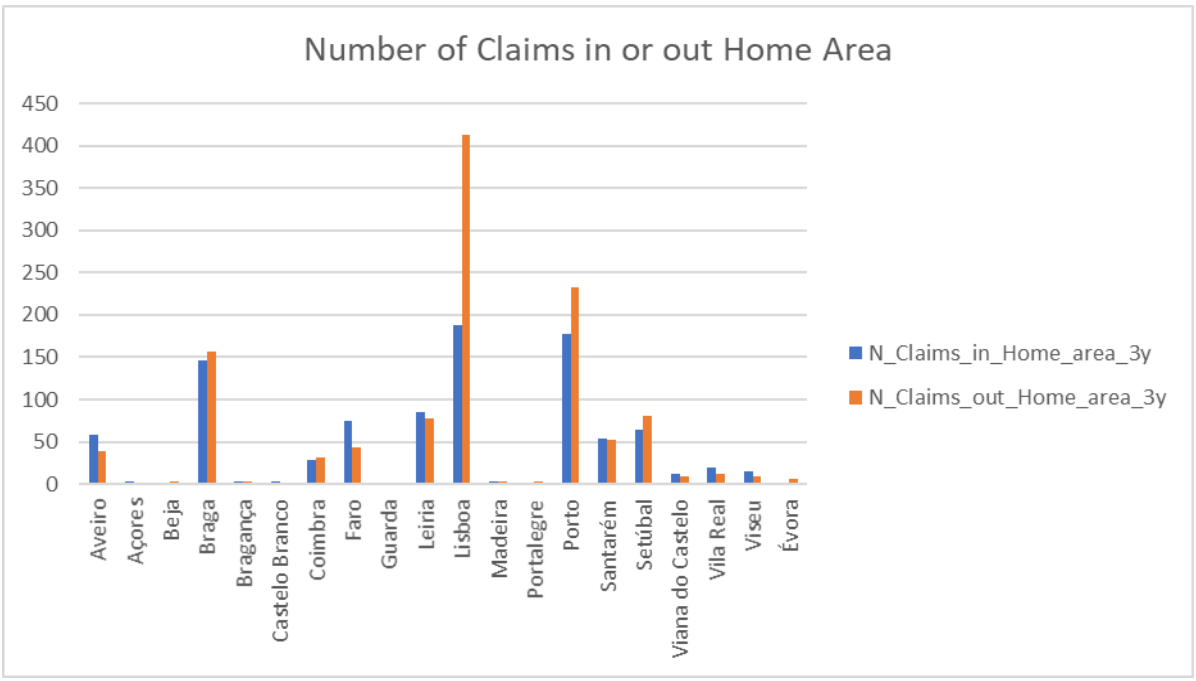

Figure 15 – Comparison between the number of claims in and out driver's home area

As we can see, there are more claims occurred out of the policyholder home area for almost all the districts in study. Most of the companies pricing models use the driver's home area variable as an important variable and make decisions based on it, however, this analyse shows that the claims don't occur usually in the driver's home area.

Variables correlation

In order to analyse if some variables were correlated with our target variables (claims' frequency and severity) we performed a correlation between them and driver's age variable, vehicle's years variable and driving license years variable. This analyse was performed with the continuous variables and not the categorical ones.

The table obtained was the following:

Pearson Correlation Coefficients, N = 12983 Prob > |r| under H0: Rho=0		
	Frequency	Severity
N_years_driver_Class	-0.00155	-0.02293
	0.8602	0.0090
N_years_driving_license_Class	-0.00155	-0.02376
	0.8602	0.0068
N_years_vehicle_Class	0.00358	-0.01264
	0.6834	0.1499

Table 16 – Correlation table

Analysing the table, it is possible to see that any of the variables have a correlation with claims' frequency or claims' severity, exactly what we were expecting to see.

3.3. MODEL DEVELOPMENT

After the variables' treatment, and their exploratory analysis, we were ready to start our models. According with data characteristics and models' details mentioned in the "Literature Review" we must choose the best suited model to the frequency and the best to the severity model possible.

To build this part of this dissertation we will use the SEMMA methodology. This methodology is applied by SAS Enterprise Miner, the software we will use in this work. SEMMA stands for Sample, Explore, Modify, Model, Access, and is a methodology usually used to perform analytical models.

3.3.1. Frequency model

We entered into the model phase with a dataset with the following variables only: Frequency_Class, Capital_Class, Category_of_Vehicle, Cubic_Capacity_Class, District, Fuel, N_years_driver_Class, N_years_driving_license_Class, N_years_vehicle_Class, Sex, Type_of_Vehicle, Gains_2015_Class,

Present_Population_Male and Present_Population_Female, where the Frequency_Class is the target variable.

The variables are classified as nominal (categorical) or ordinal (categorical where the order is important) in the following way:

Variable	Role	Classification
Frequency_Class	Target	Nominal
Capital_Class	Input	Nominal
Category_of_Vehicle	Input	Nominal
Cubic_Capacity_Class	Input	Nominal
District	Input	Nominal
Fuel	Input	Nominal
N_years_driver_Class	Input	Ordinal
N_years_driving_license_Class	Input	Ordinal
N_years_vehicle_Class	Input	Ordinal
Sex	Input	Nominal
Type_of_Vehicle	Input	Nominal
Gains_2015_Class	Input	Nominal
Present_Population_Male	Input	Nominal
Present_Population_Female	Input	Nominal

Table 17 – Models' variables

We created a node in SAS Miner called "Data Partition" to separate our data in train dataset and validation dataset – we stayed with 70% into the train part and the other 30% into the validation one.

After this node creation, we created the three regression nodes, called "Regression_Freq_Forward", "Regression_Freq_Backward" and "Regression_Freq_Stepwise", with the relevant characteristics as follow:

Regression node name	Regression_Freq_Forward	Regression_Freq_Backward	Regression_Freq_Stepwise
Regression Type	Linear Regression	Linear Regression	Linear Regression
Link Function	Logit	Logit	Logit
Input Coding	GLM	GLM	GLM
Selection Model	Forward	Backward	Stepwise

Table 18 – Frequency regression models' characteristics

3.3.2. Severity Model

For the severity model we couldn't use the same data as for the frequency. Here we needed to use only the policies that had costs, which equivales to 1301 policies (10% of the dataset used for the frequency model).

The input variables for this model were the same as for the frequency model, with the exception of the target variable that now was the variable "severity_class".

Using the same methodology, we also create a node of data partition and the regression models' nodes with the same characteristics as for the frequency model.

Regression node name	Regression_Sev_Forward	Regression_Sev_Backward	Regression_Sev_Stepwise
Regression Type	Linear Regression	Linear Regression	Linear Regression
Link Function	Logit	Logit	Logit
Input Coding	GLM	GLM	GLM
Selection Model	Forward	Backward	Stepwise

Table 19 – Severity regression models' characteristics

The results of these models are analysed in the chapter 4. It was also developed decision trees and neural networks, but the best results were obtained by the linear regression models.

3.4. POLICYHOLDER APPROACH

We also would like to study the behaviour of the claims by policyholder instead of only study it by policy. With the dataset by policy and organised with the 27 interest variables we made some arrangements to develop the dataset and to have in consideration the policies that a policyholder has, his claims, and his own characteristics, which should be the same between policies, but they are not

for some of them. We also dropped some variables that are not equal for all the policies such as the policy fractionation, number of car seats, Bonus Malus, Business Management, vehicle type, etc. For the capital, we decided to calculate their sum. We also decided that for the variables that should be equal, but they are not (for example the county, district, years of driving license, years of the driver, etc) we kept the values that appeared in the first row. The following table explains the data manipulation made:

Variables in data set by policyholder	**Manipulation made from the policies dataset**
Policy_Holder_code	
Count_of_Policy	Number of policies for each policyholder
Min_of_Beginning_Date	The oldest beginning date
SUM_of_Capital	Sum of the Capitals of the policies of each policyholder
County	If not equal kept the first row
District	If not equal kept the first row
Region	If not equal kept the first row
N_years_driving_license	If not equal kept the first row
N_years_driver	If not equal kept the first row
N_years_vehicle_mean	Mean of the policies' vehicles' age of each policyholder
Sex	If not equal kept the first row
Sum_of_N_Claims_3y	Sum of all the claims occurred in the 3 years of each policyholder
Sum_of_Claims_Cost_3y	Sum of all the claims' cost in the 3 years of each policyholder
Max_of_Annulation_Date	Most recent annulation date
Max_of_Replacement_Date	Most recent replacement date
2015_Gains	If not equal kept the first row
2014_Gains	If not equal kept the first row
Present_Population(Male)	If not equal kept the first row
Present_Population(Female)	If not equal kept the first row
Risk_Exposure	Calculate as in policies dataset
Frequency	Calculate as in policies dataset
Severity	Calculate as in policies dataset

Table 20 – Policyholder variables

Having this dataset, and to develop two models as in the policies dataset, there was also the need to create the same class' variables.

3.4.1. Frequency model

Similarly to policies dataset we started the model with the following variables and with the following characteristics:

Variable	Role	Classification
Frequency_Class	Target	Nominal
District	Input	Nominal
SUM_of_Capital_Class	Input	Nominal
N_years_driver_Class	Input	Ordinal
N_years_driving_license_Class	Input	Ordinal
N_years_vehicle_Class	Input	Ordinal
Sex	Input	Nominal
Gains_2015_Class	Input	Nominal
Present_Population_Male	Input	Nominal
Present_Population_Female	Input	Nominal

Table 21 – Policyholder models' variables

Having the same approach as for the policies dataset, we created a node in SAS Miner called "Data Partition" to separate our data in train dataset and validation dataset (with 70% vs 30% respectively).

After that we created the three regression nodes as well, called "Regression_Freq_Forward", "Regression_Freq_Backward" and "Regression_Freq_Stepwise", with the relevant characteristics as follows:

Regression node name	Regression_Freq_Forward	Regression_Freq_Backward	Regression_Freq_Stepwise
Regression Type	Linear Regression	Linear Regression	Linear Regression
Link Function	Logit	Logit	Logit
Input Coding	GLM	GLM	GLM
Selection Model	Forward	Backward	Stepwise

Table 22 – Frequency regression models' characteristics (policyholder dataset)

3.4.2. Severity Model

For the severity model we had to use only the policyholders that had some costs, such as in the model for policies dataset.

The input variables for this model were the same as for the frequency model, with the exception of the target variable that was the variable “severity_class”.

Using the same methodology, we also created a node of data partition and the regression models’ nodes with the same characteristics as for the frequency model.

Regression node name	Regression_Sev_Forward	Regression_Sev_Backward	Regression_Sev_Stepwise
Regression Type	Linear Regression	Linear Regression	Linear Regression
Link Function	Logit	Logit	Logit
Input Coding	GLM	GLM	GLM
Selection Model	Forward	Backward	Stepwise

Table 23 – Severity regression models’ characteristics (policyholder dataset)

The results of these models are analysed in the chapter 4.

4. RESULTS AND DISCUSSION

In this chapter we will analyze the frequency and severity models' results for the policies dataset and the frequency and severity models for the policyholder dataset.

4.1. FREQUENCY MODEL FOR POLICIES DATASET

As referred previously, in the sub-chapter 3.3.1., it was developed three different frequency models, with forward, backward and stepwise approach.

The regression model with the forward selection model concluded that the significative variables to predict the target variable (frequency_class) were:

Variables	Pr > ChiSq
Fuel	<.0001
N_years_driver_Class	0.0071
N_years_vehicle_Class	0.0007
Present_Population_Female_Class	<.1144

Table 24 – Significative variables of frequency model with forward selection

On the other hand, with the backward selection model, the variables were:

Variables	Pr > ChiSq
N_years_driver_Class	0.0108
N_years_vehicle_Class	0.0013

Table 25 – Significative variables of frequency model with backward selection

And, with stepwise selection model, we had:

Variables	Pr > ChiSq
N_years_driver_Class	0.0108
N_years_vehicle_Class	0.0006

Table 26 – Significative variables of frequency model with stepwise selection

According with SAS methodology, to see if a model is good, we should have in consideration the misclassification rate (MISC) and the average squared errors (ASE) statistics, and see if they are quite the same in both train and validation data sets.

	Forward		Backward		Stepwise	
	ASE	MISC	ASE	MISC	ASE	MISC
Train	5,3794%	11,953%	5,3932%	11,953%	5,3942%	11,953%
Validation	5,5624%	12, 289%	5,5380%	12, 289%	5,5407%	12,289%

Table 27 – Frequency models' summary

To compare the three models and choose the best one, we pick the model with the lowest misclassification rate, and if there are equal rates, we choose the one with the lowest average squared errors, always in the validation set. So, the chosen model is the one calculated with the Backward approach and with the vehicle years and driver's age having a significative impact.

4.2. SEVERITY MODEL FOR POLICIES DATASET

Now we will analyse the severity models developed earlier for the policies dataset.

The regression model with the forward selection model concluded that the significative variables to predict the target variable (severity_class) were:

Variables	Pr > ChiSq
Cubic_Capacity_Class	0.0219

Table 28 – Significative variables of severity model with forward selection

With the backward selection model, the variables were exactly the same:

Variables	Pr > ChiSq
Cubic_Capacity_Class	0.0219

Table 29 – Significative variables of severity model with backward selection

And, with stepwise selection model, we had:

Variables	Pr > ChiSq
Cubic_Capacity_Class	0.0219

Table 30 – Significative variables of severity model with stepwise selection

Now, we used the same approach as in the frequency model to choose the best model, and, as referred, we should have in consideration the misclassification rate (MISC) and the average squared errors (ASE) statistics and see if they are quite the same in both train and validation data sets.

	Forward		Backward		Stepwise	
	ASE	MISC	ASE	MISC	ASE	MISC
Train	13,399%	35,194%	13,399%	35,194%	13,399%	35,194%
Validation	13,568%	36,224%	13,568%	36,224%	13,568%	36,224%

Table 31 – Severity models' summary

Assuming the same approach, and looking at the misclassification rate, all the models have exactly the same values and the same significative variable, the Cubic_Capacity_Class, which means the vehicle cubic capacity is the most significative in relation to claims' severity.

4.3. FREQUENCY MODEL FOR POLICYHOLDER'S DATASET

The regression model with the forward selection model concluded that the significative variables to predict the target variable (frequency_class) were:

Variables	Pr > ChiSq
N_years_driver_Class	0.0798
N_years_vehicle_Class	0.0013
SUM_of_Capital_Classes	0.0383

Table 32 – Significative variables of frequency model with forward selection (policyholder dataset)

On the other hand, with the backward selection model, the variables were:

Variables	Pr > ChiSq
N_years_vehicle_Class	0.0017
Present_Population_Female_Class	<.0001
Present_Population_Male_Class	<.0001
SUM_of_Capital_Classes	0.0361

Table 33 – Significative variables of frequency model with backward selection (policyholder dataset)

And, with stepwise selection model, we had:

Variables	Pr > ChiSq
N_years_vehicle_Class	0.0108
SUM_of_Capital_Classes	0.0006

Table 34 – Significative variables of frequency model with stepwise selection (policyholder dataset)

Having, again, the same approach to choose the best model we analysed the misclassification rate (MISC) and the average squared errors (ASE) statistics:

	Forward		Backward		Stepwise	
	ASE	MISC	ASE	MISC	ASE	MISC
Train	5,7547%	12,863%	5,7490%	12,863%	5,7591%	12,863%
Validation	5,7831%	12, 944%	5,7902%	12, 944%	5,7845%	12,944%

Table 35 – Frequency models' summary (policyholder dataset)

To compare the three models and choose the best one, we pick the model with the lowest misclassification rate, and if there are equal rates, we choose the one with the lowest average squared errors, always in the validation set. So, the chosen model is the one calculated with the Forward approach and with the vehicle years, driver's age and sum of capital having a significative impact.

4.4. SEVERITY MODEL FOR POLICYHOLDER'S DATASET

The regression model with the forward selection model concluded that the significative variables to predict the target variable (severity_class) were:

Variables	Pr > ChiSq
District	<.0001
SUM_of_Capital_Classes	0.0590

Table 36 – Significative variables of severity model with forward selection (policyholder dataset)

With the backward selection model, the variables were:

Variables	Pr > ChiSq
District	<.0001

Table 37 – Significative variables of severity model with backward selection (policyholder dataset)

And, with stepwise selection model, we had the same as in the Backward:

Variables	Pr > ChiSq
District	<.0001

Table 38 – Significative variables of severity model with stepwise selection (policyholder dataset)

Now, we used the same approach as in the frequency model to choose the best model and, as referred, we should have in consideration the misclassification rate (MISC) and the average squared errors (ASE) statistics and see if they were quite the same in both train and validation data sets.

	Forward		Backward		Stepwise	
	ASE	MISC	ASE	MISC	ASE	MISC
Train	12,932%	34,994%	13,061%	35,327%	13,462%	35,327%
Validation	13,306%	36,387%	13,462%	36,387%	13,681%	37,387%

Table 39 – Severity models' summary (policyholder dataset)

Assuming the same approach, the best model is the forward model, with the District and sum of capital being the most significative variables, however the sum of capital variable significative level is a higher than the 5%.

5. CONCLUSIONS

To create a Motor Own Damage pricing analytical model with business attributes and insured environment variables we developed exploratory analysis and linear regression models to predict claims' frequency and severity. We built these linear regression models for the policies' dataset and for the policyholders' dataset.

For the frequency model of policies dataset, we have, as we see in the sub-chapter 4.1, the driver's age in classes and the vehicle's years in classes as the significative variables. We have analyzed the significative level for each class of this variables and see what impact they have in the target variable. The class 2 of the driver's age in classes variable has a significative impact for the class 2 and 3 of frequency, and the class 3 of the driver's age in classes variable has a significative impact on class 1 and 2 of frequency. Also, the class 2 of the vehicle's years in classes has a significative impact on class 2 of the target variable frequency. Analyzing the classes of cubic capacity variable with significative level we see that only the class 2 have significative impact for the class 3 and 4 of the severity classes.

Besides this analysis with the regression results and in order to combine both claims' frequency and severity, we went through the exploratory analysis performed in the sub-chapter 3.2 and merged all the information in the following graphs. The first one concerns to the cubic capacity analysis, then the second to the driver's age and the third to the vehicle's years. The squares in blue correspond to the information obtained through the regression models, and the squares in green correspond to the exploratory analysis performed.

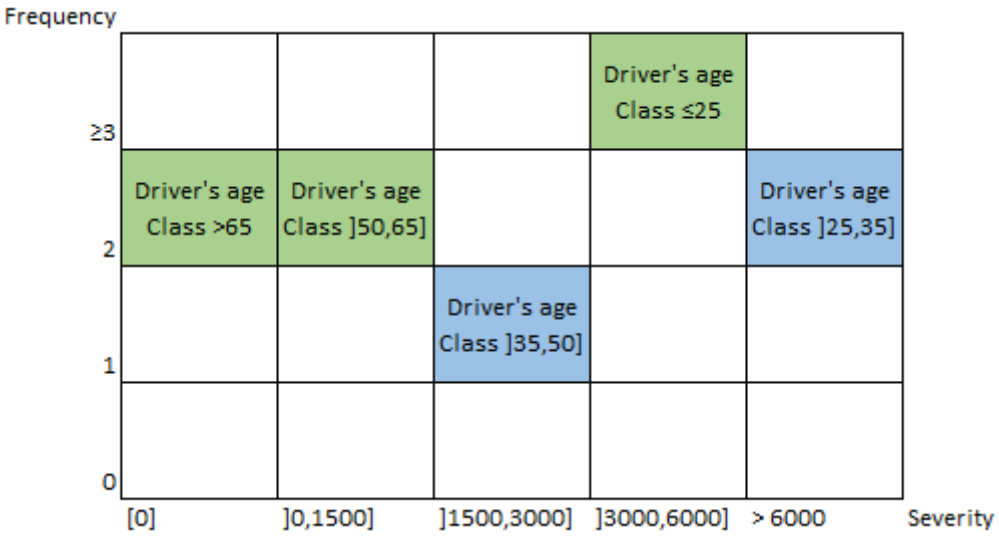

Figure 16 – Driver's age claims' frequency and severity analysis

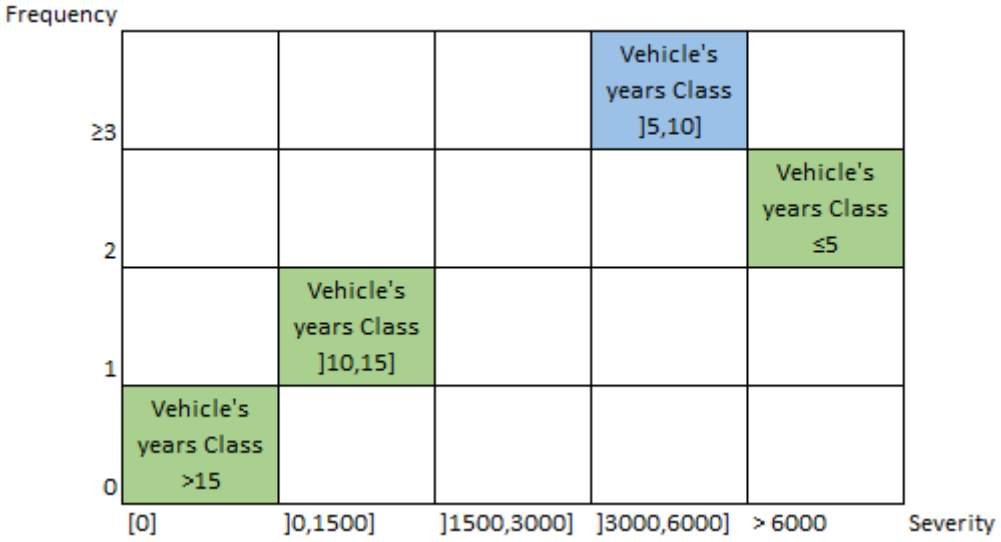

Figure 17 – Vehicle years claims' frequency and severity analysis

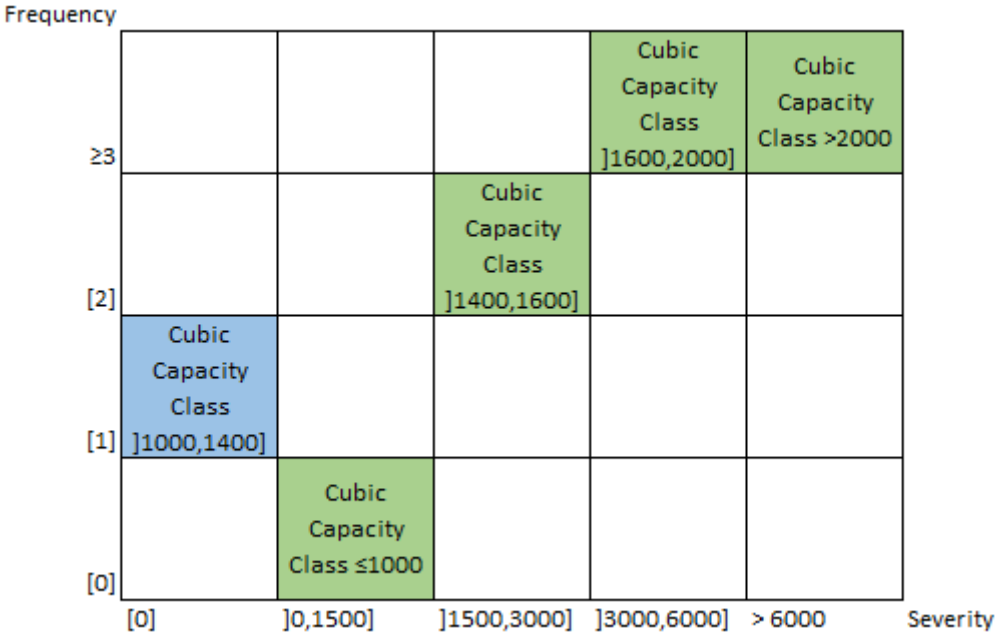

Figure 18 – Cubic Capacity claims' frequency and severity analysis

The following graph corresponds to the merge of these three variables in relation to the frequency and severity. The squares in yellow matches with the mix of class variables that came from regression models and exploratory analysis.

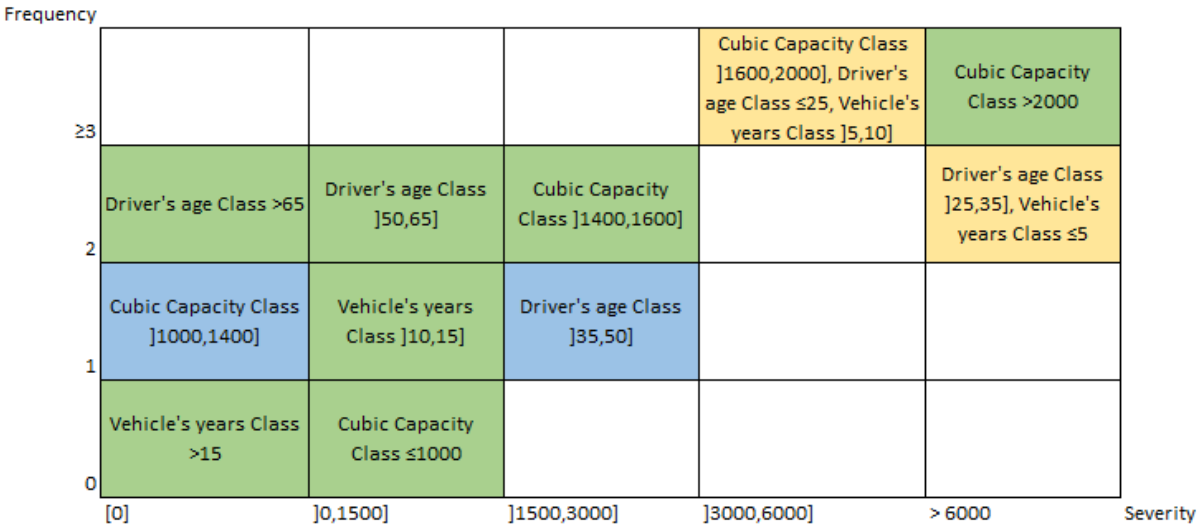

Figure 19 – Three variables in claims' frequency and severity analysis

From the figure 19 we can conclude that younger drivers have more claims' frequency and severity, in opposition to older drivers, that have less severity and less frequency. For the vehicle years, older vehicles have the lowest frequency and severity, and for the youngest the opposite. Also, as long as the cubic capacity classes are increasing the frequency and the severity increase as well. We can also realize that the riskiest cases are for the vehicles with the highest cubic capacity class, drivers less than 35 years old and cars with less than 10 years. The ones with the lowest risk are the vehicles with more than 15 years and lower cubic capacity.

The following graphs represent the number of riskiest policies in each district (figure 20) and in each Business Management (figure 21):

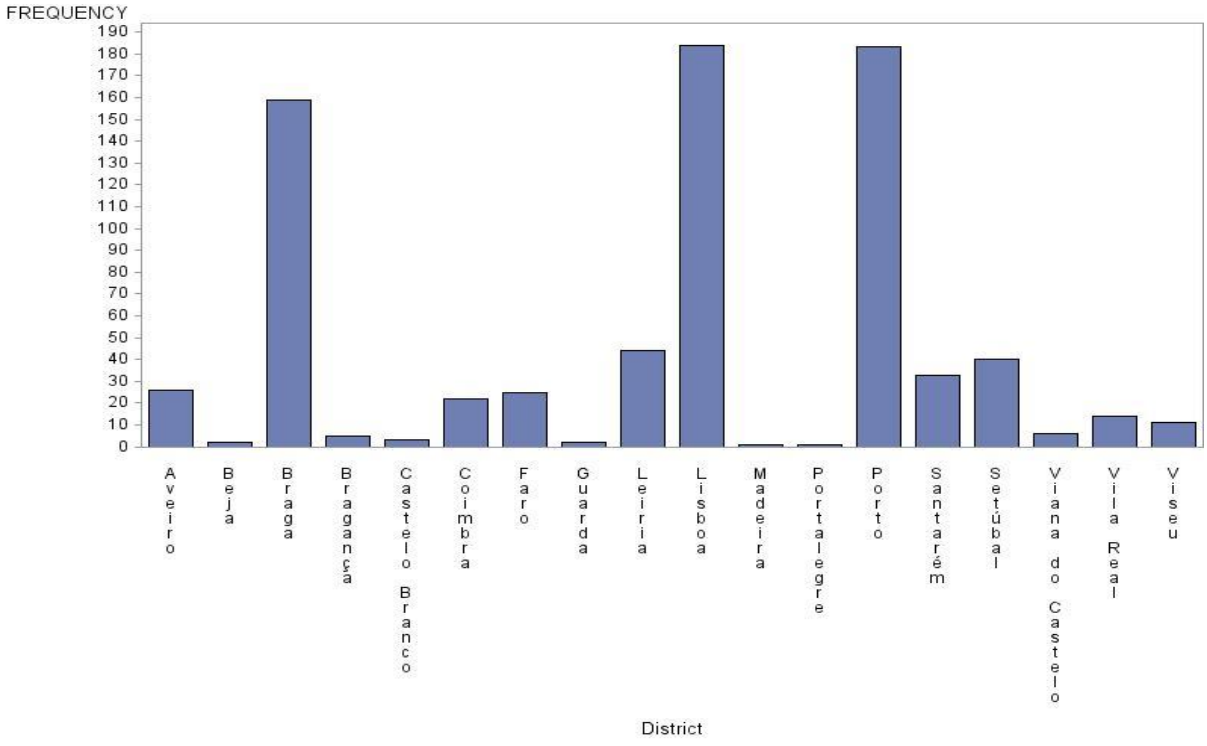

Figure 20 – Risky polices by district

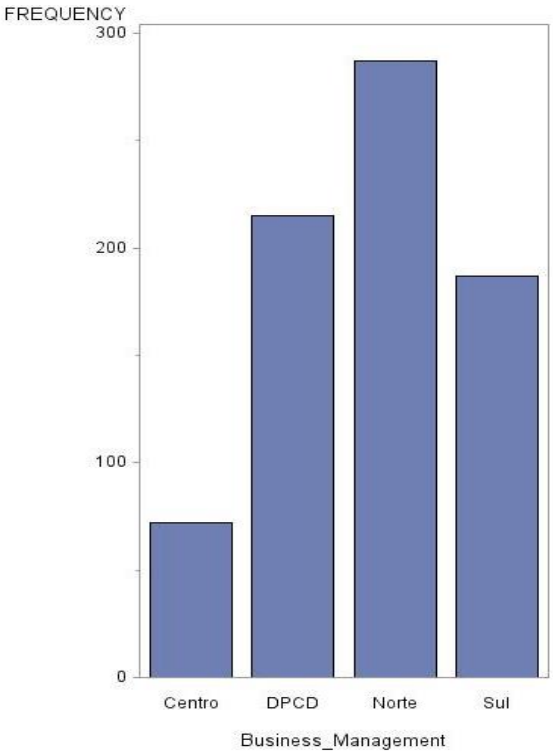

Figure 21 – Risky policies by Business Management

We can here conclude that Porto, Lisbon (Lisboa) and Braga are the districts with more risky policies and Beja, Madeira and Portalegre are the districts with the less risky policies. In relation to the business management departments, it is visible that North (Norte) has more risky policies and Center (Centro) has the least.

In relation to policyholder's dataset, and assuming the same approach for the conclusions, there was the need to build the frequency and severity graphs for this dataset. The graphs are as follows:

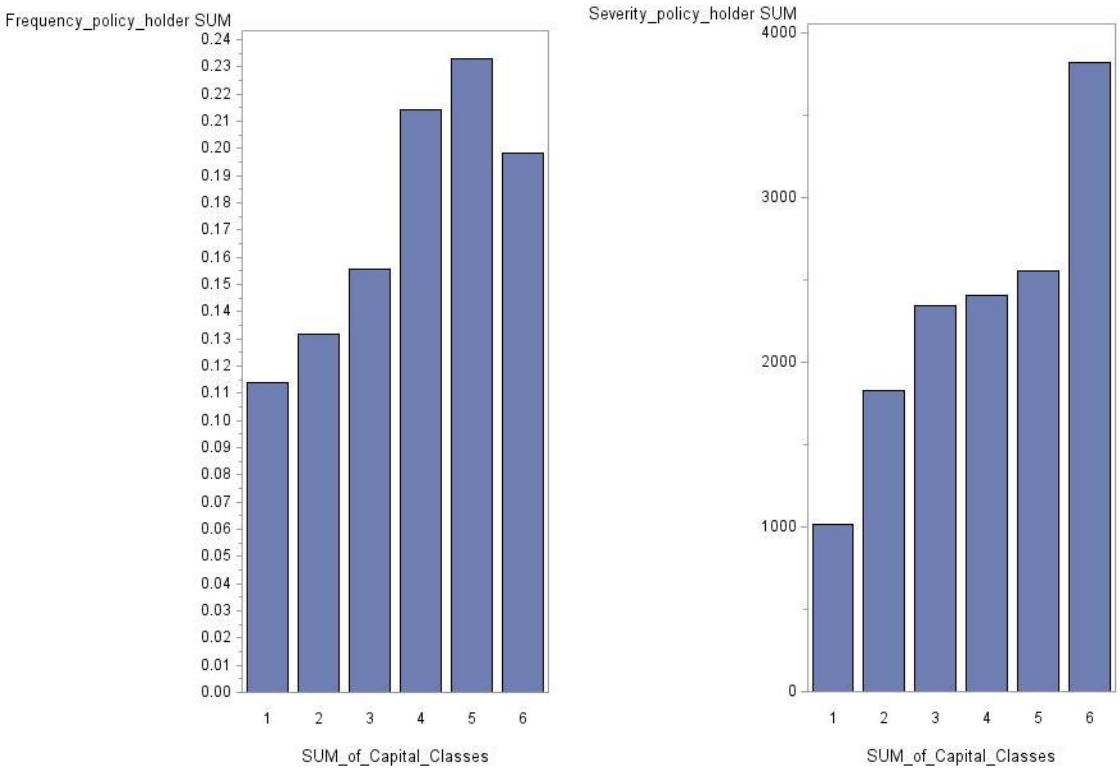

Figure 22 – Sum of capital classes claims' frequency and severity

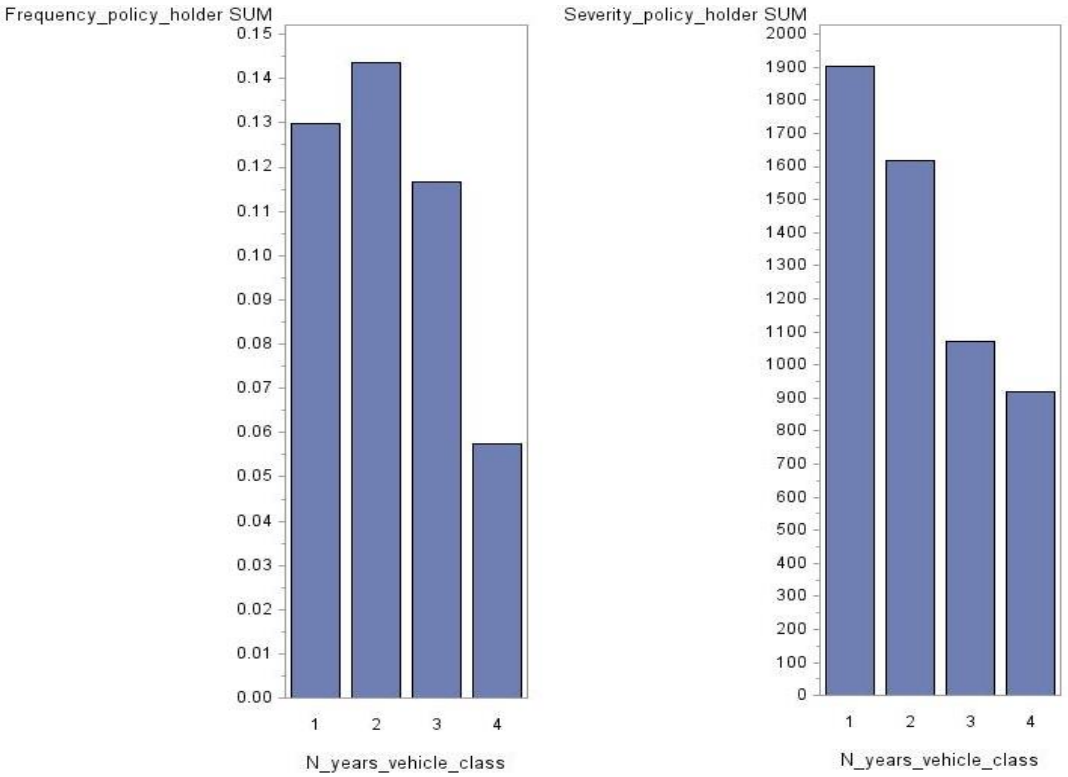

Figure 23 – Vehicle years classes claims' frequency and severity

Having the previous graphs, we were able to perform the following:

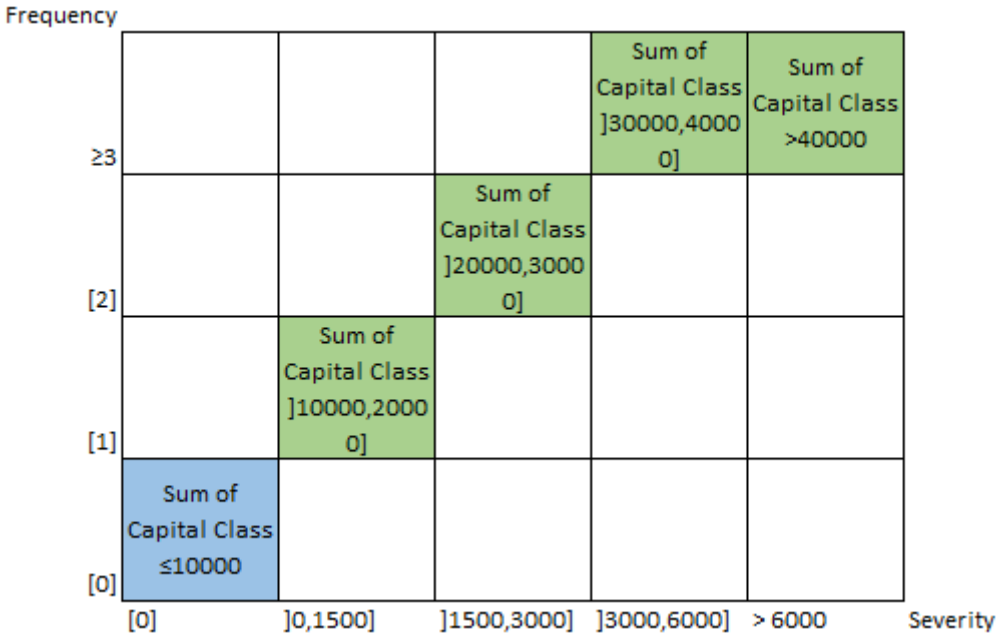

Figure 24 – Sum of capital classes claims' frequency and severity analysis

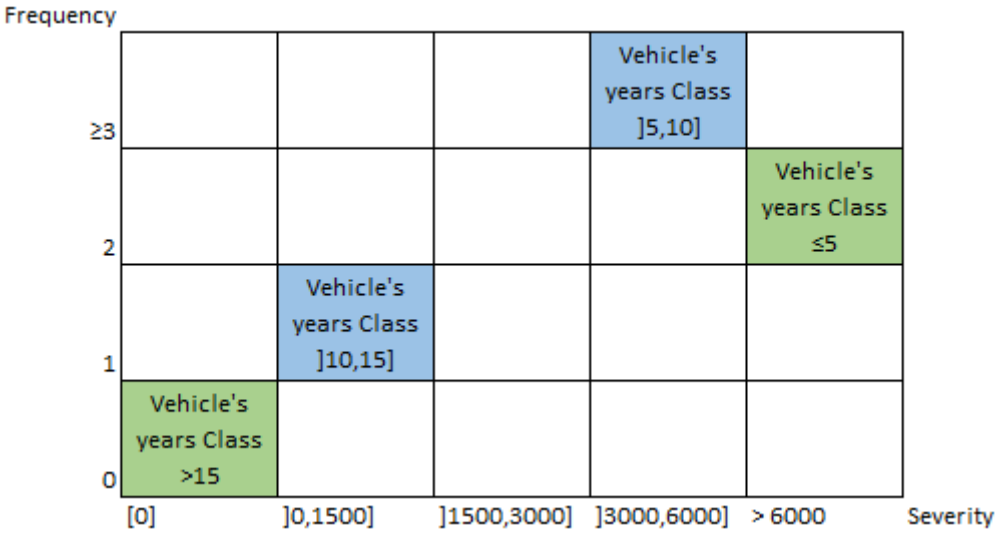

Figure 25 – Vehicle years classes claims' frequency and severity analysis

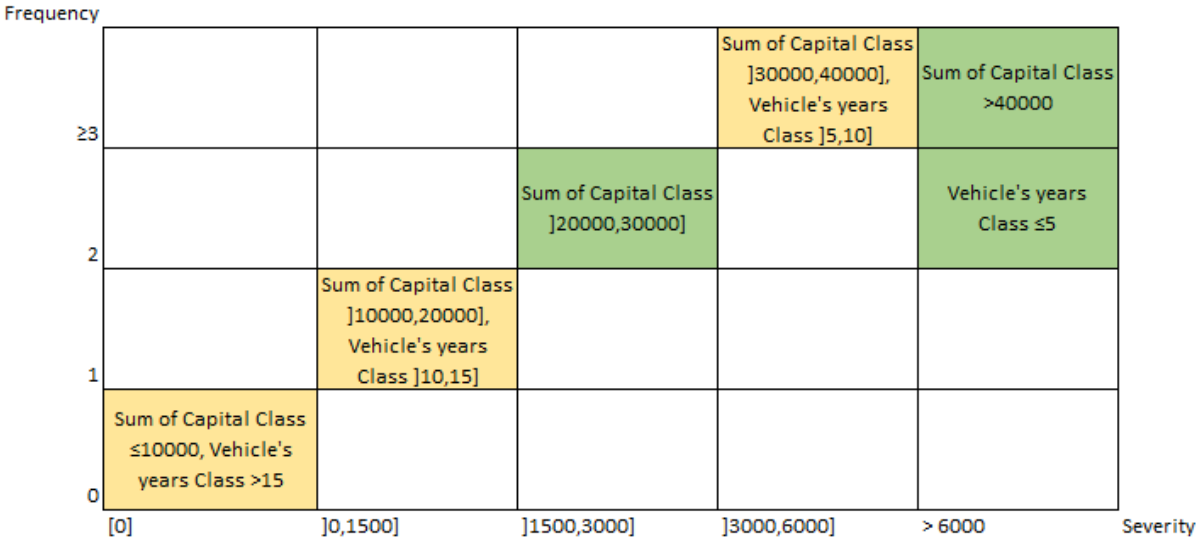

Figure 26 – Two variables in claims' frequency and severity analysis

In the figure 26 it is possible to observe that higher capital represents more risk and lower capital less risk. We can also analyze that for the vehicles years variable happens the same as in the policies dataset, the oldest vehicles have the lowest frequency and severity and the youngest have the highest frequency and severity. We can conclude that the riskiest policies are the ones with the sum of capital higher than 30.000€ and vehicles with less than 10 years. The policies with the least risk have the sum of capital less than 10.000€ and vehicle years more than 15 years.

The following graphs represent the number of riskiest policyholders in each district (figure 27) and for each driver's age class (figure 28):

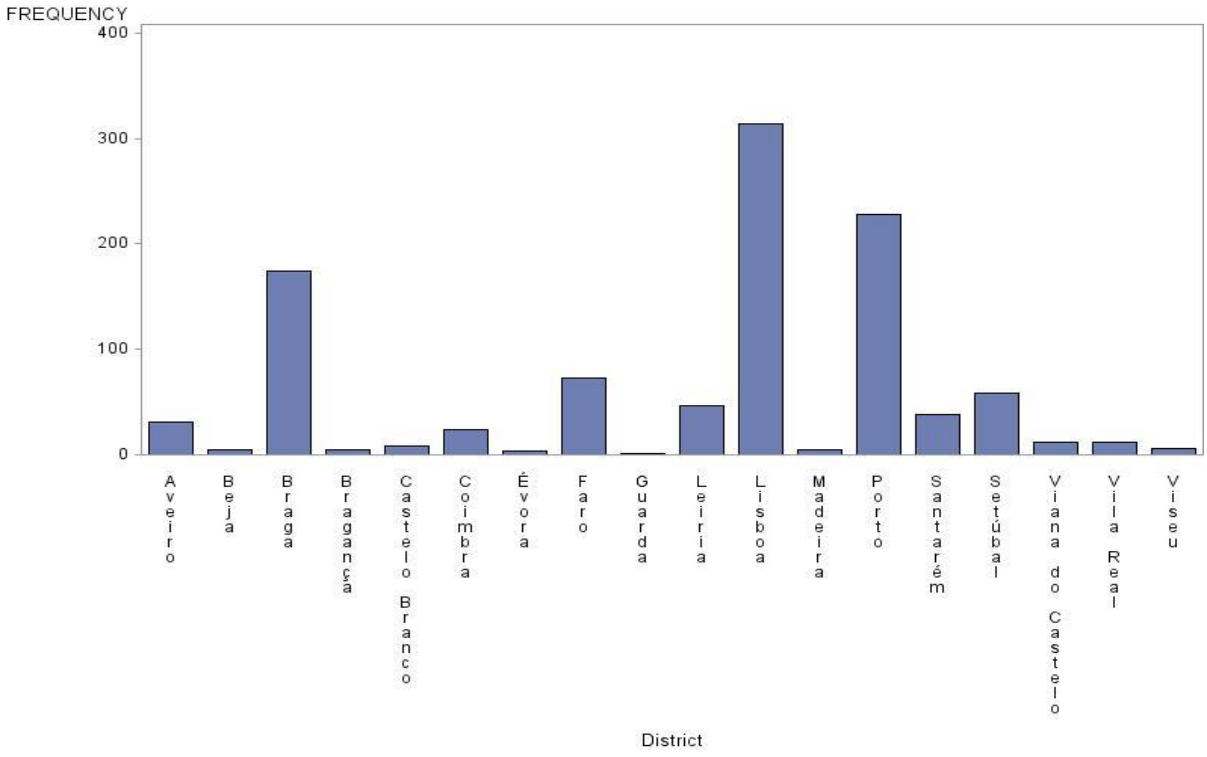

Figure 27 – Risky policyholders by district

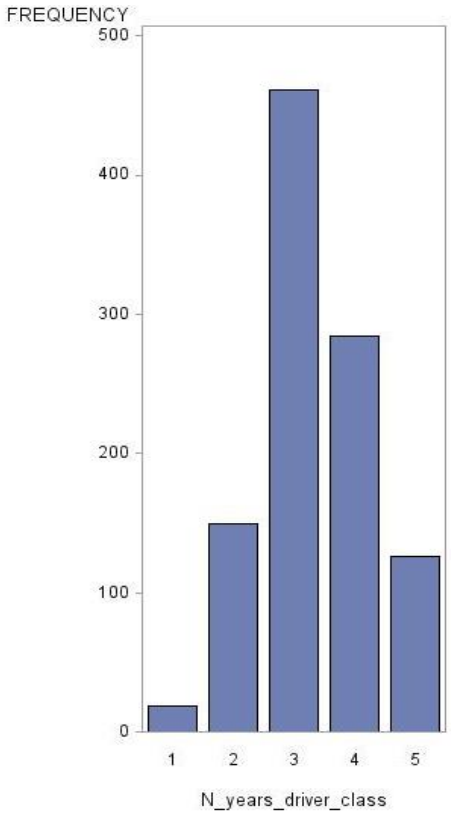

Figure 28 – Risky policyholders by driver's age classes

Such as in the policies approach, districts with more risky policyholders are Lisbon, Porto and Braga. For the driver's age classes, there are more risky policyholders in class 3, which represents the drivers with age between 36 and 50 years old. The first class, drivers' age less than 25 years old, is the class with less risky policyholders.

With this thesis the insurance company that has made available the data used here gets his data analyzed, can know better their portfolio and understand behaviors. Having this information, it is easier to make decisions and know which variables have a significative impact in the portfolio performance.

The approach of merging two different techniques, analytical models with exploratory analysis, helps the academy in developing new methods. For small insurance companies, with less data, this is a good method to work around the disadvantage.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

There are always a few limitations in relation to resources and time that constraints studies like the present one. Generally, companies are not worried about historical data quality which makes backoffice research much more complex and time-consuming.

A constraint that needs to be mentioned is the fact that we had not a significative number of insured environment variables in this study. In INE data there were not much environment data by home area (only the data we have extracted), and the company had not this kind of data as well. For example, it would be interesting to see if a person has a familiar relationship or be a co-workers or even friend of another person have any impact or relation in the way his claims occurred. Due to the GDPR (General Data Protection Regulation) there are some data we couldn't use too.

There is exhaustive literature about this dissertation scope, however most of it is related with motor third party liability pricing instead of motor own damage.

For future research, it would be interesting to develop more models that were not used in this dissertation. It would also be worth to perform this analysis with more data, being that external to policyholder or their characteristics.

7. BIBLIOGRAPHY

Agababa, R. (n.d.). Five Ways Data Analytics is Transforming the Insurance Industry - DATAVERSITY.

Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, *96*(2), 187–224. https://doi.org/10.1007/s10182-011-0152-7

APS. (2017). PANORAMA DO MERCADO SEGURADOR.

ASF. (2016). O presente e futuro da atividade seguradora em portugal _____, 1–8.

Caravela, C. de S. (2017). Relatório e Contas 2017. *Macroeconomics : A European Perspective /*. https://doi.org/10.1002/bit.26931.This

Clarke, R., & Libarikian, A. (2014). Unleashing the value of advanced analytics in insurance.

Columbus, L. (2017). McKinsey's State Of Machine Learning And AI, 2017.

Companhia de Seguros Allianz Portugal, S. A. (2017). Relatório e contas 2017.

Cummings, D. (n.d.). Big data can make a big difference in modern underwriting | Visualize | Verisk Analytics.

David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, *20*(15), 147–156. https://doi.org/10.1016/S2212-5671(15)00059-3

Drews, B. (2000). Insurance Rate Making Using Data Mining. *SAS EMEA*.

Económico, J. (2017). Quem é quem no Setor Segurador em Portugal. *Vogue Portugal*. https://doi.org/10.1007/s00423-012-1003-z

Ernst & Young. (2013). Advanced analytics for insurance, 24.

Filler, B. (2012). Insurance Pricing Models Using Predictive Analytics.

Gangam, S., & Engelhardt, A. (n.d.). Non-Life Insurance Pricing using R.

Khopkar, H., & King, J. (2007). SAS Global Forum 2007 Data Mining and Predictive Modeling Critical Success Factors for Rate Making with SAS ® Enterprise Miner TM SAS Global Forum 2007 Data Mining and Predictive Modeling.

Routledge, R. (n.d.). Law of large numbers | statistics | Britannica.com.

SAS Institute. (2003). Using Data Mining for Rate Making in the Insurance Industry, (June).

SAS Institute. (2017). *Applied Analytics Using SAS® Enterprise Miner™ Enterprise MinerCourse Notes*.

Stanford. (n.d.). Linear Regression.