MEGI

Master Degree Program in

Statistics and Information Management

A REFRESHED VISION OF NON-LIFE INSURANCE PRICING

A Generalized Linear Model and Machine Learning Approach

Carina de Miranda Clemente

Project Work

presented as partial requirement for obtaining the Master Degree Program 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

A REFRESHED VISION OF NON-LIFE INSURANCE PRICING

By

Carina de Miranda Clemente

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Risk Analysis and Management.

Supervisor: Gracinda Rita Diogo Guerreiro

Co-Supervisor*:* Jorge Miguel Ventura Bravo

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Catarina Clemente

Lisboa, 30/11/2022

DEDICATION

This project is dedicated to my grandmother Maria, who had always been one of the strongest supporters of my academic journey.

One of her wishes was the successful completion of my studies, which she witnessed with the beginning of the development of this project.

Unfortunately, she is not here today to see me deliver it.

Avó, I did it.

ACKNOWLEDGEMENTS

To my coworkers Mariana, Filipa, Rodrigo and Nuno, it would have been impossible to develop this project without your help. You were always there to answer my questions and to guide me through this long path.

To my mother Natalina and my boyfriend Bruno, who supported me all along despite my absent times.

To my aunt Elisabete and my cousin Catarina, who would always make sure I was being able to hold down the fort.

To Catarina and Sofia, who did not get tired of my thousand questions and always listened to my (repetitive) concerns.

To my supervisor professor Gracinda and co-supervisor professor Jorge, that despite their tight availability were able to share some of their knowledge with me.

ABSTRACT

Insurance companies are faced with a constantly changing world, in a daily basis. There are a number of known risks that are quantifiable, alongside with many more that remain unknown or difficult to measure. It is only natural that the pricing of such risks must evolve side by side with the state-of-the-art technology that is available. In the late years there has been a rise in the number of studies that conclude on the better fitting of models based on machine learning technology, when it comes to estimate the prices charged by insurance companies in order to hedge the risks they endure. This project work aims to provide a refreshment of the pricing methods applied by a given insurance company operating in Portugal, by developing GLM-based frequency and severity modelling on a subset of the motor portfolio of the company, backed up by a machine learning model: Gradient Boosting. In doing so, there is an expectation of improvement of the accuracy of the model, providing better fitting estimates that could translate into a fairer tariff for both the insurance company and its clients. In fact, it was concluded that the Gradient Boosting approach outputted the lowest total deviance associated with the frequency model. In terms of severity, it was the GLM that produced to the lowest value. With the development of this project, there is now an open path in my company for the inclusion of machine learning methods on the development of insurance tariffs, being here proven that with little required input, this approach can in fact lead to very good results and thereby add value to the classic methodology.

KEYWORDS

Non-life Insurance; GLM; GBM; Tariff; Machine Learning; Frequency; Severity

RESUMO

As companhias de seguros deparam-se diariamente com um mundo em constante mudança. Diversos riscos são conhecidos e quantificáveis, mas muitos outros permanecem desconhecidos ou são difíceis de mensurar. É só natural que o *pricing* destes riscos evolua lado a lado com a tecnologia mais recente. Nos últimos anos tem havido um acréscimo do número de estudos que chegam à conclusão de que os modelos baseados em tecnologia de *machine learning* devolvem melhores resultados no que toca à estimação dos prémios a cobrar pelas companhias, por forma a cobrir os riscos que seguram. Este trabalho de projeto tem como objetivo a apresentação de uma nova visão dos métodos de *pricing* aplicados por uma dada seguradora a atuar em Portugal, ao desenvolver os modelos de frequência e severidade numa amostra do portfólio automóvel, através das metodologias de Modelos Lineares Generalizados e de *Gradient Boosting*. É expectável uma melhoria na precisão dos modelos, que resulta em estimativas mais corretas que se traduzem numa tarifa mais justa, não só para a seguradora como para o cliente. De facto, foi concluído que o modelo de frequência pela abordagem de *Gradient Boosting* retornou o desvio total mais baixo. Em termos de severidade, foi a abordagem de Modelos Lineares Generalizados que chegou ao desvio mais baixo. Com o desenvolvimento deste projeto, há agora uma porta aberta na companhia para a implementação de modelos de *machine learning* no desenvolvimento das tarifas, ficando aqui provado que com pouco input, esta abordagem pode se facto levar a bons resultados e assim adicionar valor à metodologia clássica.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ML	Machine Learning
GLM	Generalized Linear Model
MTPL	Motor Third Party Liability
GBM	Gradient Boosting Model
CV	Cross-validation
PDP	Partial Dependence Plot
ICEP	Individual Conditional Expectation Plot

1. INTRODUCTION

When it comes to the estimation of tariffs and the definition of pricing structures, Generalized Linear Models (GLM) are the most widely used tool by insurers. However, there have been developed several Machine Learning (ML) models in the past decade, that might threaten the position of the GLM as the most suitable model to be used by insurance companies.

Although research in the actuarial field is mainly focused in GLM, there has been registered a rise in published papers that have studied the implementation of ML models in non-life insurance, many of which conclude that this new approach can lead to faster outputs, with a higher predictive value.

One of the reasons behind the lack of employment of machine learning methods lies behind the fact that these models are harder to interpretate, when compared to the classical models, like GLM. Despite this fact and acknowledging that ML models are in fact an important tool for actuaries, the Institute and Faculty of Actuaries has decided in 2019 to include ML related questions in their exams, with the purpose of ensuring that the curriculum stays relevant and up to date, reflecting the skills required for an actuary in a constantly changing world (Hetherington, 2020).

1.1. MOTIVATION

This project comes as a refreshment of the current pricing methods applied by a Portuguese insurance company, that has agreed to provide a subset of data on a third-party liability motor portfolio. The pricing team applies the classical GLM (through the use of the EMBLEM software, by Willis Towers Watson) in order to obtain the best fit of the prices adequate to each risk profile.

It is believed that in order to be able to provide a pure premium that is a direct consequence of the risk that the policy imposes, the models applied should be the ones that provide the best estimations possible. Even though GLM has a solid worldwide approval, it is crucial that insurance companies keep up with the state-of-the-art findings, having proof that these are effective. It is in this line-of-thought that the non-life actuaries of the company believe that this project could add value to their current methodology, that does not include any type of machine learning techniques or back-testing.

1.2. OBJETIVES

The main objective of this project is to propose a refreshed new view of the current tariff applied to the third-party liability motor coverage of a Portuguese insurance company, by predicting the claim frequency and severity.

A GLM will also be developed, to be used as a benchmark. This model will be built upon the assumptions made by the pricing team in the past, facilitating the pre-processing of the variables to be chosen as part of the model. This will ease the expected workflow, allowing for an extra analysis of the machine learning model in study.

The main research question underlying this project is: can ML models improve the accuracy of the current severity and frequency models that are being implemented in this insurance company?

Supported by the available literature, it is believed that the main research question could be true, whose veracity will be investigated throughout this project. Considering that event, the model that will be developed would have the power to be fastest, better fitting and more efficient than the benchmark.

This project will allow the company to be up to date with the new findings in the frequency and severity prediction modelling, as it is it is very important that the prices that are practiced are fairly linked to the risk that the client imposes. The relevance of this work can be considered high, as it brings a new light into what is presently being done. It is of special importance because the company has never backtested its present model against a ML approach. As so, it presents as a great synergy among the two parts, by allowing the consolidation of knowledge acquired during this master and an innovation opportunity for the company to stay ahead of its competitors.

1.3. Structure

This paper will be structured as follows:

Chapter 1 is an introductory Chapter and Chapter 2 exhibits the literature review, presenting the current state-of-the-art of the applications of machine learning in the actuarial field.

Chapter 3 describes the methodology applied throughout this project, namely the theory behind the concepts of severity and frequency and the models to be applied: Generalized Linear Model and Gradient Boosting. On Chapter 4, it is included a descriptive and exploratory statistics of the data, which allows for a better understanding of the dataset that is being studied.

Chapter 5 presents the fitting of the distributions and the frequency and severity models, according to the GLM and GBM approaches. There is also a comparison between the models previously obtained, through the analysis of the significant variables, the total deviance and the residuals.

Lastly, Chapter 6 presents the conclusion and final remarks of this project, followed by Chapter 7 that states the limitations that were encountered throughout the development of this project, as well as recommendations for future works.

2. LITERATURE REVIEW

The basis of insurance pricing lies on one simple assumption: in order to best price the risk, there must be an accurate prediction of the future losses. This literature review summarizes the evolution of analytical models used to price non-life policies, from the current state-of-the-art to the most recent breakthrough findings in this field.

One of the first analytical models developed to be used as a tool to predict the desired response variables was introduced in 1972, the famous Generalized Linear Model (Nelder & Wedderburn, 1972). This model generalizes a linear regression by allowing a non-linear relationship among the predictive variables and the response via a link function. It is based on this model that worldwide insurance companies predict the claim severity and frequency associated with their portfolio of policies. The reason for this global application lies on the fact that the GLM can be very robust and easy to interpret, as the predictors' value can be directly depicted from the model output, making this model an easy and valuable tool to explain to the stakeholders the rationale behind the pricing of such policy.

With the beginning of the 21st century, there has been an increase of data available to the companies, followed by a natural desire of the researchers to employ this data in the most efficient way possible. With this, came the development of artificial intelligence and machine learning, namely in the actuarial field. The fact that these models can be harder to interpretate (when compared to the GLM) has led to a discouragement to apply them in real-life situations, with only a few papers being published in the 2000s, such as the work developed by Smith et al. (2000) that applies data mining models, specifically decision trees and neural networks, to uncover the customer retention and claim patterns. However, as some recent events have emerged, like the usage of bigdata, especially telematics data and datasets composed of numerous images, and the evolution of the computer's GPUs (Lecun et al., 2015) the popularity of ML has spiked.

Over the last five years, there has been registered an exponential growth of the global Artificial intelligence/ML market, which is expected to keep growing in the five years that follow, with a special magnitude in North America, surpassing 5 billion US$ as of the present year prediction, expecting around 27 billion US$ of value in 2026 (Malhotra & Sharma, 2018). In a survey conducted in 2020 by Willis Tower Watson, 26% of American insurers claim to use ML and AI to build risk models for decision making and 22% to reduce required manual input, values that have doubled when compared to the previous two years. Around 60% of the companies inquired have shown the wish to implement these capabilities by 2021 (Willis Tower Watson, 2020).

There are several recent articles that study the application of tree-based models, such as the works developed by Quan & Valdez (2018) that compared the usage of univariate and multivariate response variables when predicting frequency in several non-auto coverages, utilizing the CART, random forest and Gradient Boosting models. Based on a Swedish home insurance portfolio, there have also been developed tree-based models to predict frequency, such as simple decision trees, random forest and Gradient Boosting, with the overall conclusion that the latest two outperform the first, which is natural, given the growing increase in complexity between these models (Tober, 2020).

When predicting the claim frequency in auto insurance, a study has shown that among several tree-based models, XGBoost has a better accuracy in terms of normalized Gini than other models (Fauzan

& Murfi, 2018). With the same goal of predicting claim occurrence, and with the differentiating factor of utilizing telematics data, an article published by Pesantez-Narvaez et al. (2019) has shown evidence that XGBoost requires more to match the predictive performance of the logistic regression, only increasing the predictive performance slightly. There were also difficulties in interpretating the coefficients.

With the objective of developing a full tariff plan for a Belgian TPL motor cover, there have been compared the performance of simple regression trees, random forest and boosted trees, using the GLM as a benchmark. It was reached the conclusion that boosted trees have the capability of outperforming the classical model (Henckaerts et al., 2020). With a similar approach, Noll et al. (2018) predicted claim frequency of a French motor TPL dataset, using regression trees, Gradient Boosting and neural networks, again using GLM as a term of comparison. The authors concluded that Gradient Boosting and neural networks outperformed the GLM, but also stated that the development of the benchmark model could have been improved. In the same year Su & Bai, (2020) predicted the frequency and severity for the TPL motor cover, combining the stochastic gradient boost and a profile likelihood approach to estimate the parameters of the distributions. This work differs by introducing a dependence between claim frequency and claim average cost, using the claim frequency as a predictor in the regression model for the severity. It was concluded that dependent models have a better performance, being superior to other state-of-the-art models.

There are also studies that focus on other covers with great exposure, such as collision. In 2021 it was published a study that developed frequency prediction on a Swiss motor portfolio, using GLM and GAM as reference models and two random forest models, one for claim severity and other for the log-transformed claim severity. The usage of the log-normal transformation of severity did not lead to any performance gains when the random forest was applied, however it was still the favorite choice for explaining the right-skewed claims. Globally, GAM has a better performance (Staudt & Wagner, 2021).

Another studied application of ML models in non-life insurance is the ability to compare the online information provided by several direct insurers. Grize et al. (2020) collected data from 20 competitors' websites, using web-crawlers, in order to obtain a dynamic pricing system for online motor vehicle liability insurance. This work used a commercial product that contained a very large number (over 50) of standard ML algorithms.

Even though ML models are not the standard of the industry to be used as a base to predict frequency and severity, one could always take a more secure approach and use ML as a way to facilitate the choice of the variables to be employed in the classical GLM. Extreme gradient boost has been utilized to detect the interaction between the variables and LASSO and Ridge to select the variables to be used in the model (Zhifeng, 2020).

Artificial intelligence can have a great impact in the insurance field, as the results of several papers enlighten that there can be cost efficiency and new revenue streams, transitioning the insurance business model from loss compensation to loss prediction and prevention (Eling et al. (2021). Despite the new studies that have been developed, the insurance sector is still behind in the global artificial intelligence movements.

It should be noted that this project will not study the application of AI in insurance pricing, instead focusing on the development of a ML model based on a conventional motor dataset.

3. METHODOLOGY

To best understand how to answer the main question of this dissertation, that is if machine learning methods can provide an accuracy improvement of the coefficients used to obtain the tariff structure, the methodology section overviews the fundamental theory behind the main topics discussed: the basic principles of insurance and pricing, the generalized linear model and the Gradient Boosting model. There is also conferred a descriptive and exploratory analysis of the dataset in study, which allows for a better interpretation of the characteristics of the MTPL cover portfolio.

3.1. INSURANCE FUNDAMENTALS

This section presents a global view of the insurance pricing industry and the core principles applied on this field.

3.1.1. Basic Principles of Insurance Pricing

The underlying assumption behind insurance contracts lies on the trade of protection against an uncertain risk of future financial loss for a fair and adequate premium.

Under the insurance policy coverages, the insured person can require a compensation for the suffered losses, by filing a claim. In the event of an accident, assuming the possibility of quantifying and measuring the losses, the insurance company is obligated to issue a compensation, according to the terms of the contract agreed by both parties.

The premium practiced by the insurance company must reflect the most significative risk factors represented by the portfolio of policies. The balance in the insurance pricing system lies in the premium practiced, which must be adjusted to each type of risk. It should be the closest possible to the predicted number of incurred losses for each risk profile, making it fair, but also consider the rentability of the insurance company, making it profitable. It is due to the importance of the premium to be charged to the customers that insurance companies take special precaution in developing the most accurate models that can lead to the tariff structure to be applied.

In Portugal, the MTPL cover has been compulsory since 1980, following the emission of the decree law No. 408/79, series 1 of 1979-09-25. This cover ensures that losses suffered by third parties, are covered in the event of a claim. In case the responsible of the accident does not have an active insurance contract, the interests of the injured parties are still protected. The minimum insurance capitals are reviewed every five years and the latest update was released in March 2022, with effects in June of the same year. The capital is composed of 6 450 000 € minimum for personal injuries and 1 300 000 € for material damage, summing up a total of 7 750 000 € (*Circular n.o 2/2022, de 15 de Março*, 2022). Given the mandatory nature of this cover, it is only natural that it represents the largest portion of the gross earned premium of most non-life insurance companies, highlighting the importance of the accuracy of the tariff structure to be applied.

3.1.2. Premium Estimation

The premium that is charged to the policyholder is usually decomposed in three components: Pure Premium, representing the expected value of the losses associated to the risk in question; Security Margin, established to hedge against randomness and variance of the risk; Other charges, related to the management of the policy and taxes.

Considering each policy in the portfolio, the total cost of the claims, S, is given by the following expression:

$$S = \sum_{i=1}^{N} X_i \tag{3.1}$$

where N represents the number of claims filed (in a one-year period) and X_i represents the cost of the i^{th} claim, for *i=1, …, N*.

Assuming that the number of claims per year (frequency) and correspondent cost (severity) are independent, the expected value of the total cost is:

$$E[S] = E\left[\sum_{i=1}^{N} X_i\right] = E\left[E\left[\sum_{i=1}^{N} X_i\right] \middle| \, N = n\right] = E[N]E[X] \tag{3.2}$$

The pure premium can also be translated into:

$$PP = frequency \times severity \tag{3.3}$$

It is commonly known and accepted that the charging of the pure premium alone would be inappropriate. According to ruin theory, studied by Filip Lundberg, the pure premium alone is insufficient because in the long run the ruin is known to be inevitable, regardless of the existence of initial reserves (Gudmundarson et al., 2021).

The last step to obtain the pure premium with a loading is the choice of the premium principle to be applied in order to obtain the final value.

Expected Value Principle

The premium is an increasing linear function of α, the safety security loading. It is equal to the pure premium when $\alpha = 0$.

$$P = (1 + \alpha)\, E[S], \quad \alpha \geq 0 \tag{3.4}$$

Variance Principle

$$P = E[S] + \alpha\, Var[S], \quad \alpha \geq 0 \tag{3.5}$$

Overcomes the fragility of the previous principle, by taking into consideration the fluctuations of S. Here, the premium depends not only on the expected value but also the variance of the losses.

Standard Deviation Principle

$$P = E[S] + \alpha \sqrt{Var[S]}, \quad \alpha \geq 0 \tag{3.6}$$

This last principle takes into account the expected value of the losses alongside with the standard deviation. The premium given by (3.4) is an increasing linear function of α.

3.1.3. Frequency

One of the factors that influences the premium of a policy is the frequency, which represents the expected number of claims per exposure time, usually one year. Exposure is a measure used to evaluate the risk present in a given portfolio held by the insurance company.

Given its count nature, the number of claims filled by the policyholder can sometimes follow a Poisson distribution.

A discrete random variable N is said to follow a Poisson distribution with parameters $\lambda \in \mathbb{R}^+$ and $x \in \mathbb{N}_0$, with the correspondent probability mass function:

$$f(x, \lambda) = P(N = x) = e^{-\lambda} \frac{\lambda^x}{x!} \tag{3.7}$$

where $E[X] = Var[X] = \lambda$.

The claims $N_1, \ \dots \ , N_n$ are identified as a family of independent distributed variables, with the same parameter λ for each claim. Thanks to this assumption, the problem is reduced to the estimation of this same parameter.

Having this into consideration, the Poisson distribution is considered very adequate to model counting exercises in which the probability of success is reduced, very much like the number of claims.

Other very common distribution applied when modelling the number of claims is the Negative Binomial. It is usually the case when the variance is above average, a phenomenon known as over-dispersion, making it less compatible with the Poisson distribution assumptions.

3.1.4. Severity

The average cost of a claim is commonly known as severity. Given the nature of this variable, it is common to test the fitting of the Gamma distribution to estimate it.

Considering *Y* a continuous variable, it is said to follow a Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$ and its probability density function is defined as follows:

$$f(x) = \frac{\left(\frac{y\alpha}{\theta}\right)^{\alpha}}{y\,\Gamma(\alpha)} e^{-y\frac{\alpha}{\theta}}, \quad y \in \mathbb{R}^{+} \tag{3.8}$$

With $E[Y] = \theta$ and $V[Y] = \frac{\theta^2}{\alpha}$.

Because this distribution is right skewed and allows for large values in the right tail, it is most of the times suited as a good fit for the distribution of the cost of the claims.

Different alternatives to model the individual claims could be the Lognormal, Pareto or Weibull distributions, among others.

3.1.5. A Priori Tariff Structure

The explanatory variables used as rating factors represent a priori measurable information regarding the policyholder. Meaning that this information is collected prior to any claim taking place, being solely based upon certain measurable characteristics, such as the brand of the vehicle or the age of the driver. Therefore, it is impossible to collect information regarding the driving behaviour of the policyholders, as it does not have a measurable or visible nature.

3.2. Generalized Linear Models

3.2.1. Introduction

Developed and introduced by Nelder and Weddenburn in 1972, the GLM are the current state-of-the-art when it comes to the modelling and explanation of claim frequency and severity. This type of model is a generalization of the multiple linear regression model and is usually applied in situations where there is the aim of modelling the relation between the variables and their influence on each other. According to this methodology, the response variable is distributed following a distribution belonging to the exponential family, for example a Gamma or Negative Binomial. Distributions of the exponential family are well suited to model real life problems, namely motor claims.

3.2.2. GLM Definition

The classical regression model, GLM, establishes and studies the relation between the response or dependent variable, *Y,* which follows a certain probability function $f(x)$ and the explanatory or independent variables, $X_1, X_2, \dots X_n$ existent in a sample of n observations. It can be stated that the linear regression models follow this given structure:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_n X_{in} + \varepsilon_i\ , \quad i = 1, \dots, n \tag{3.9}$$

with ε_i representing the normally distributed vector of random errors.

By assuming that the distribution of the variable in study is a member of the exponential family of distributions, there can be stated several advantages when comparing to the linear regression, such as the possibility of the response variable following other distributions other than the Normal and the ability to overpass the linear structural form $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, achieving a model in which the goal is to model a transformation of the average value $h(\boldsymbol{X}\beta)$, and the presence of heteroscedasticity.

The GLM is essentially composed of three elements: a random component, a systematic component (the linear predictor) and a link function.

Random Component

Specifies the probability distribution of the response variable ***Y***, which must follow a distribution belonging to the exponential family (e.g. Gamma, Poisson, Negative Binomial, etc.).

Systematic Component

The explanatory independent variables are related to the response variable as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in}\ , \quad i = 1, \dots, n \tag{3.10}$$

These variables can be re-written as a linear combination, also known as linear predictor, η_i:

$$\eta_i = \sum_{j=0}^{n} x_{ij}\beta_j \; = \boldsymbol{X}_i^T \boldsymbol{\beta} \;\; , \quad i = 1, \dots, n \tag{3.11}$$

With $\boldsymbol{X}_i$ as the matrix of the explanatory variables and $\boldsymbol{\beta}$ the correspondent model coefficients.

Link Function

This component establishes a non-linear relation between the linear predictor η_i and the expected value of the response variable, $E[Y_i] = \mu_i$. The link function, which is monotonous and differentiable, is defined as follows:

$$g(\mu_i) = \eta_i, \quad i = 1, \dots, n \tag{3.12}$$

Its inverse function is given by:

$$\mu_i = g^{-1}(\eta_i) \tag{3.13}$$

3.2.3. The Exponential Family

The exponential family comprises many of the well-known probability distributions. This following Section is based upon the works developed by Choi (2017).

A random variable Y is said to follow a distribution belonging to the exponential family of distributions (in natural or canonical form) if its density function is of the following structure:

$$P_\theta(y) = exp\left[\frac{\theta\, T(y) - A(\theta)}{\phi} + C(y, \phi)\right] \;\; , y \geq 0 \; and \; \theta, \phi > 0 \tag{3.14}$$

where:

- $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ is the canonical (natural) parameter, related to the expected value of *Y*;
- ϕ is the dispersion parameter, sometimes introduced to control the shape of $P_\theta(y)$ and is usually known;
- $T(y)$ is a map, $A(\theta)$ is the log partition (cumulant) function and $C(y, \phi)$ is the element independent from the canonic parameter.

Following the structure above, it is stated that $E[Y] = \mu = A'^{(\theta)}$ and $V[Y] = A''^{(\theta)}\phi$.

The next two assumptions are always taken into consideration when it comes to the study of the exponential family of distributions:

- When $P_\theta(y)$ is said to be a probability density function, it is assumed to be continuous as a function of y, meaning there is no singularity in the probability measure $P_\theta(y)$).
- When $P_\theta(y)$ is said to be a probability mass function, it exists a range of discrete values of $P_\theta(y)$ that are the same for all θ and all y.

In the event of $P_\theta(y)$ satisfying either of the conditions, it is said to be regular.

3.2.4. Model Estimation and Fitting

As explained by Piet de Jong and Gillian Z. Heller (2008) constructing a GLM, based upon a response variable Y, can be achieved following six steps:

1. Choose a response distribution $f(y)$ and $A(\theta)$. The response distribution must be chosen accordingly to the problem and data in hands.
2. Choose a link function $g(\mu)$, as represented on (3.12).
3. Choose the explanatory variables $\boldsymbol{x}$ in terms of which $g(\mu)$ is to be modelled.
4. Collect several observations $y_1, \ldots, y_n$on the dependent variable and correspondent $x_{11}, \ldots, x_{1n}$ on the independent variables. This sample is assumed as a random sample from the entire population.
5. Fit the model by estimating $\boldsymbol{\beta}$ and, if unknown, ϕ. The fitting is performed using a statistical software such as SAS, R or EMBLEM.
6. Considering the estimations of $\boldsymbol{\beta}$, generate predictions (or fitted values) of y for different combinations of *x* while at the same time examine the performance of the model, by analysing the residuals, as well as other diagnosis tools. The estimates values of β are also very useful to determine whether a given explanatory variable *x* is important or not in determining μ.

3.2.4.1. Parameter Estimation

The part of the process of developing a generalized linear model that holds the most interest is the estimation of the regression parameters $\boldsymbol{\beta}$, as shown on equation (3.10). These parameters are obtained by maximizing the log-likelihood function (de Jong & Heller, 2008). Because in the scope of the GLM these equations do not have an analytical solution, it is common practice to apply numeric methods to obtain a solution (Guerreiro, 2016).

Having this problem into consideration, Nelder & Wedderburn (1972) developed an algorithm to reach a solution for these equations, known as Iterative Weighted Least Squares Estimation, based upon the Fisher scoring method. There is also a good explanation of this method in Chen & Shao (1993).

The maximum likelihood estimation is settled upon choosing the parameter estimates which maximize the likelihood of observing the sample $\mathrm{y}_1, \dots, \mathrm{y}_\mathrm{n}$. Each of the y_i has probability function $f(y_i)$, which therefore depends of θ and ϕ if applicable. Considering that the y_i are independent, their joint probability function is as follows:

$$f(y, \theta, \phi) = \prod_{i=1}^{n} f(y_i, \theta, \phi) \tag{3.15}$$

The log-likelihood can be obtained as the logarithm of the likelihood function:

$$l(\theta, \phi) \equiv \sum_{i=1}^{n} ln\, f(y_i, \theta, \phi) \tag{3.16}$$

It is equivalent to say that the parameters θ and ϕ can also be obtained by maximizing the log-likelihood function. It is preferred to maximize the later instead of the likelihood because it is easier to work analytically (de Jong & Heller, 2008).

3.2.4.2. Complex Components

The EMBLEM and R softwares enable the possibility of capturing special characteristics in the model, such as the orthogonal polynomials and the interactions between independent variables, that otherwise would be hard to manage.

Orthogonal Polynomials

Belonging to the class of polynomials, orthogonal polynomials obey to an orthogonality relation, such as any two different polynomials in the sequence are orthogonal to each other under a certain inner product. According to equation (3.10) given the linear predictor structure, each explanatory variable has a beta assigned to it. Regardless of that, there is often a trend with curvature when comparing the observed values versus explanatory variable, which indicates that it would be best to fit a polynomial.

Considering p the order of the polynomial, the polynomial component can be expressed as:

$$\sum_{n=1}^{p} \beta_{ip} x_i^p \tag{3.17}$$

Despite not being linear in terms of the explanatory variable, this component is linear in terms of beta, meaning it can be estimated using the same algorithm as the linear predictor. The new coefficients to be estimated are then coefficients of the transformed variate powers of order t, $P_t(x_i)$, assuming the underlying relation associated with orthogonality:

$$\sum_{i=1}^{n} P_r(x_i)P_s(x_i) = 0, \qquad r \neq s, \qquad r, s\ \in 1, 2, \ldots, p \tag{3.18}$$

Meaning the orthogonal polynomial predictor of variable x_i can be estimated as:

$$\eta_{x_i} = \alpha_0 + \alpha_1 P_1(x_i) + \cdots + \alpha_p P_p(x_i) \tag{3.19}$$

There are not any problems of correlation after the orthogonal transformation is applied, but the order t of the polynomial should not be too high, to avoid overfitting and simplify the model (Zhifeng, 2020).

Interactions

An interaction takes place when an independent variable has a different effect on the outcome depending on the values of other independent variable.

The simplest interaction is the two-way interaction, taking two variables into account. There could be three scenarios:

1. **Two numerical continuous variables:** The interaction consists in the product of the variables, adding one more coefficient to be estimated, the interaction term:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{3.20}$$

   This equation can be reformulated as:

$$\eta = \beta_0 + \beta_1 x_1 + (\beta_2 + \beta_{12} x_1) x_2 \tag{3.21}$$

   The interpretation of the interaction is clear: the relative increase (decrease) with respect to each unit of x_2 is translated as $(\beta_2 + \beta_{12} x_1)$.
2. **Two categorical variables:** Variable 1 has n_1 levels and variable 2 has n_2 levels. For each variable, the base level is separated from the rest, which are grouped together, resulting in two levels each. This results in the addition of $(n_1 - 1)\ (n_2 - 1)$ factors in the original linear predictor.
3. **One numerical and one categorical:** Assuming the categorical variable has n levels, this will add $(n - 1)$ new coefficients to the model.

3.2.4.3. Model Selection

The aim of this step is to obtain a model with the minimum amount of variables, each and every one necessary to better explain the response variable *Y*, by approximating the predicted curve to the observed curve, avoiding over and under fitting. In the beginning there is usually a large number of variables available to choose from, and this process can be quite difficult.

While using the EMBLEM software, the go-to procedure of variable selection is forward stepwise: Initially, the only parameter considered in the model is µ, the average value of all y_i observations. The variables are added to the model sequentially, usually starting with those that were considered significant in previous models developed by the company. There are two measures used to assess if a variable should be included in the model or not: the deviance and the Likelihood Ratio test.

When facing the possibility of adding a new variable, we are in the presence of two nested models: the model which includes the variable is a sub model of the initial one. The difference between both models' deviance, assuming that they have p_1 and p_2 parameters each, should asymptotically follow a Chi-Square distribution:

$$\Delta\, Deviance = -2\sum_{i=1}^{n} loglik\left(y_i, \eta_{ip_2}\right) - \left(-2\sum_{i=1}^{n} loglik\left(y_i, \eta_{ip_1}\right)\right) \sim \mathcal{X}^2_{(p_1-p_2)} \tag{3.22}$$

The chi-square *p-value* is interpreted as the probability of a random variable that is chi-squared distributed with $p_1 - p_2$ degrees of freedom being greater than the difference of deviance between the nested models. If the *p-value* is lower than the significance level, then the model with p_2 parameters is significantly better than the model with less parameters, meaning that the new variable should be included in the model. Otherwise, the initial model should remain as it was, because the variable does not add value to the model in question.

Another important tool to choose a model amongst a variety of models is the Akaike Information Criterion, applied when the models are not nested. It was developed by Hirotsugu Akaike in 1971. It is a criterion that measures the balance between the quality of fitting and the amount of parameters included in the model in study. When comparing the AIC between several models, it serves as a measure of the loss in information when choosing a certain model over another.

This method is based upon the log-likelihood function $loglik(\beta)$, rewarding the quality of the fitting, alongside with a correction element associated with the amount of parameters in the model, p, which penalizes models with a higher number of variables. In summary, it is the deviance plus 2 times the number of parameters. It is given by:

$$AIC = -2loglik(\beta) + 2p \tag{3.23}$$

The best model is the one with the lowest AIC.

3.2.4.4. Quality of Fitting

This step comes after the choice of the variables of which the coefficients are the most significant, that is, after finding the model that best fits the data. Afterwards, it is necessary to evaluate the quality of that fitting. This evaluation is proceeded through the analysis of the deviance and residuals.

Residuals

Residuals are a measure used to evaluate the choice of a certain response distribution and to outlier values. The residuals are defined as:

$$\widehat{e_i} = y_i - \widehat{\mu_i} \tag{3.24}$$

representing the difference between the observed value y_i and the fitted value $\widehat{y_i}$. These residuals can be evaluated in a number of ways, the most frequently used being the Deviance residuals, described bellow.

- **Deviance Residuals**
  The deviance residuals are obtained as the direction of the difference between each fitted and observed value:

$$r_D = \; sign(y_i - \widehat{\mu_i})\sqrt{d_i} \tag{3.25}$$

where $d_i = \; 2\, y_i \, log\left(\frac{y_i}{\widehat{\mu_i}}\right) - (y_i - \widehat{\mu_i})$.
This measure is often chosen, because it can be preferable to other types of residuals in the diagnosis of GLM, such as Person's residuals, or even response residuals or working residuals, depending on the data that is being modelled.

3.3. Gradient Boosting

3.3.1. Introduction

Gradient Boosting originated from the idea if whether or not a weak learner, defined as one whose performance is at least slightly better than random chance, could be modified (through the calculation of a form of residuals) to achieve a better value.

Introduced by Friedman (2001) Gradient Boosting is formally defined as an ensemble model of decision trees, in which multiple weak models are aggregated into a more powerful predictor.

A decision tree is a supervised learning approach that is used to solve classification and regression problems, also known as regression tree, when the latter is addressed. This method was first developed and published by Breiman et al. (1984). It is built on a tree structure, where the internal nodes are the data variables, the branches the decision rules and the two nodes the binary output.

One of those nodes represents a decision node used for decision-making and the other one is a leaf node, representing the outcome of those decisions. This type of model is usually outperformed by more complex algorithms, although decision trees can be used as a combination in ensemble algorithms, such as Gradient Boosting (Hanafy & Ming, 2021). Figure 3.1 below represents the structure of a simple decision tree.

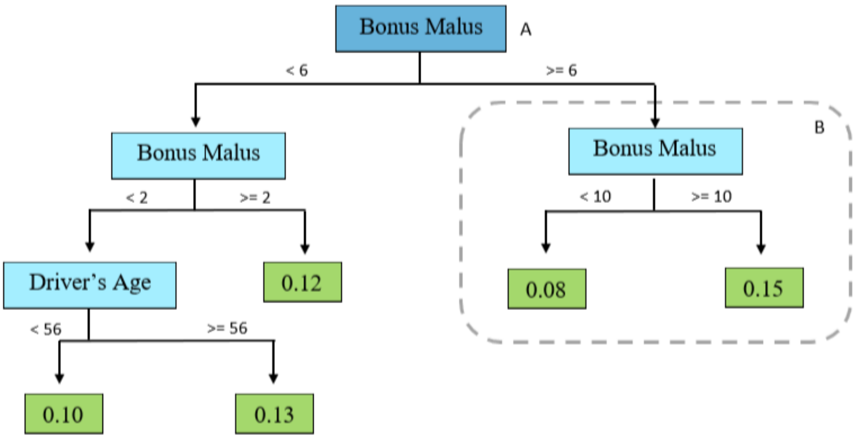


Figure 3.1 - Structure of a simple decision tree. A: Root node. B: Sub-Tree

Source: Example adapted from Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2020). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. North American Actuarial Journal, 1–31

As a part of this ensemble model, each tree improves the current model fit, thereby using information from previously developed trees.

3.3.2. Modelling Steps

In previous Sections of this paper, it has been stated that the problem at hands can be translated into finding a function $f(x)$ to predict a response variable y from a set of variables x, in order to minimize a certain loss function $L(f(x), y)$. This also applies in the GBM, of which minimization process relies on the iterative tuning of parameters, making it a gradient descending algorithm.

In boosting, $f(x)$ is estimated by the following sum:

$$\hat{f}(x) = \sum_{m=0}^{M} \beta_m h(x, a_m) \tag{3.26}$$

where $h(x, a_m)$ are simple functions known as base or weak learners. Both a and β are fitted to the training data in a step-wise manner, starting with an initial guess $\widehat{f_0}(x)$ and then, for each m, obtaining the following elements:

$$(\beta_m, \boldsymbol{a}_m) = \underset{\beta,\alpha}{argmin} \sum_{i=0}^{N} L\left(y_i, \hat{f}_{m-1}(\boldsymbol{x}_i) + \beta\ h(\boldsymbol{x_i}, \boldsymbol{a})\right) \tag{3.27}$$

$$\hat{f}_m(\boldsymbol{x}) = \hat{f}_{m-1}(\boldsymbol{x}) + \beta_m\ h(\boldsymbol{x}, \boldsymbol{a}_m)$$

Equation (3.26) is approximately solved by GBM through a two-step process:

1. Fit $h(\boldsymbol{x}, \boldsymbol{a})$ by minimizing the following least-squares sum:

$$\boldsymbol{a}_m = \underset{\alpha,\rho}{argmin} \sum_{i=0}^{N} [\tilde{y}_{im} - \rho h(\boldsymbol{x}_i, \boldsymbol{a})]^2 \tag{3.28}$$

in which,

$$\tilde{y}_{im} = -\left[\frac{\partial L\left(y_i, \hat{f}(\boldsymbol{x}_i)\right)}{\partial \hat{f}(\boldsymbol{x_i})}\right] \tag{3.29}$$

$$\hat{f}(\boldsymbol{x}) = \hat{f}_{m-1}(\boldsymbol{x})$$

with each interaction, the pseudo-residuals $\tilde{y}_{im}$ for observation i in iteration m are used to assess the regions of the predictor space for which the model does not have a good performance, and therefore improve the fit in a direction of better overall performance. This approach is known as stepwise gradient descent and ensures that a lower loss is obtain at the following iteration until convergence.

2. Calculate β_m:

$$\beta_m = \underset{\beta}{argmin} \sum_{i=0}^{N} L\left(y_i, \hat{f}_{m-1}(x_i) + \beta\ h(\boldsymbol{x_i}, \boldsymbol{a}_m)\right) \tag{3.30}$$

Considering that the base learners $h(.,.)$ are decision trees, the parameters a_m are the splitting variables and splitting points that define the tree. This means that the base learner is of the following form:

$$h(x_i, \{R_{lm}\}_1^L) = \sum_{l=1}^{L} \bar{y}_{lm}\, \mathbb{I}(\boldsymbol{x} \in R_{lm}) \tag{3.31}$$

where $\bar{y}_{lm}$ is the mean of $\tilde{y}_{im}$ on the region R_{lm}.

Because the value of the base learners $h(.,.)$ is constant for each region of the tree, $\beta\ h(\boldsymbol{x_i}, \boldsymbol{a_m})$ could be simplified to γ, calculated as the constant that has to be added to the previous model fit to minimize the loss function. Meaning that (3.30) can be re-written as:

$$\gamma_{lm} = \underset{\gamma}{argmin} \sum_{i=0}^{N} L\big(y_i, \hat{f}_{m-1}(x_i) + \gamma\big) \tag{3.32}$$

where γ minimizes (3.30) over that region. Having the previous into consideration, the current approximation $\hat{f}(x)_{m-1}$ is updated for each region R_{lm}, using γ_{lm}:

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \lambda\gamma_{lm}\mathbb{I}(\boldsymbol{x} \in R_{lm}) \quad , \quad 0 < \lambda \leq 1 \tag{3.33}$$

Considering λ the learning rate (also known as shrinking parameter) that determines the learning pace of the algorithm, by shrinking updates for $\boldsymbol{x} \in R_{lm}$. A lower value of λ outputs a better performance, reducing overfitting, but also increases the computational power required, because more trees are necessary to converge to a good solution. Usually, λ is fixed at the lowest value possible within the computational restraints (Henckaerts et al., 2020).

As the previous steps emphasize, GBM follows the same approach as GLM, minimizing a given loss function. This loss function could be the Poisson deviance (3.34) or Gamma deviance (3.35), according to the variable in study.

$$D\big(y, \hat{\lambda}\big) = 2\sum_{i=1}^{b}\left[y_i\, log\left(\frac{y_i}{\hat{y}_i v_i}\right) - \big(y_i - \hat{\lambda}_i v_i\big)\right] \tag{3.34}$$

$$D(y, \hat{\mu}) = 2\sum_{i=1}^{b} w_i \left[log\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right] \tag{3.35}$$

3.3.3. Tuning Approach

Machine learning usually relies on training data to construct a model, validation data to tune the parameters to be applied and test data to evaluate the out-of-sample performance of such model.

A fundamental part of successfully training a tree-based model is to control model complexity, maintaining a good balance between bias and variance which ultimately leads to a high prediction accuracy. One important concept that should always be taken into consideration is the bias-variance trade off: A large tree has low bias and high variance, whereas a small tree has high bias but low variance. Assuming pole examples, if there is a tree so deep that every input data has a corresponding terminal node, every data point in the training set would be well classified and the model would have a low bias. However, if this model is to be used to predict the response of new and unseen data inputs, the results would not be accurate, resulting in a very high variance, meaning that the tree is overfitted for the training set. By shortening the tree, there would be an increase in the bias, meaning that the model would have a higher prediction uncertainty, but a decrease on the variance.

In order to ensure that the GBM outputs the best and most accurate results, and therefore handling the bias-variance tradeoff, it is necessary to estimate the several parameters that compose this model. It can be achieved by applying the methodologies presented in the following two sections.

3.3.3.1. Cross Validation

Cross validation, commonly known as k-fold CV, is a resampling method that divides the training data into k random groups (or folds) of approximate same size, mutually exclusive and stratified, in order to assure that the resulting subsets' response variables follow a similar distribution (Hastie et al., 2008).

The given machine learning model is then fit on k-1 folds and the last fold is used to assess the model performance. This process is repeated k times, and each time a different fold is used as the validation set. This method outcomes k estimates of the generalization error, which are then averaged in order to obtain an approximation of the error that is to be expected in unseen data.

It is frequent to apply nested cross-validation when tunning machine learning hyperparameters. This process consists in a double loop of cross-validation: the inner loop serves for tunning the hyperparameters and the independent outer loop serves for assessing the quality of the model. Assuming the inner loop is composed of k_1 folds and the outer loop of k_2 folds, then the total number of trained models will be $k_1 k_2$.

For each iteration of the outer loop, it will be chosen one inner model (the one that minimizes the cross-validation error) which will be evaluated on the test set for the outer fold. In the end there will be k_2 estimates, which can be averaged to obtain the final model. This process is depicted bellow in Figure 3.2.

The typical choice of folds lies between 5 and 10, since there is no formal rule as to the size of k. The higher the amount of folds, the smaller the generalization error, however, there is also a large increase of the computational performance requirements (Boehmke & Greenwel, 2020).

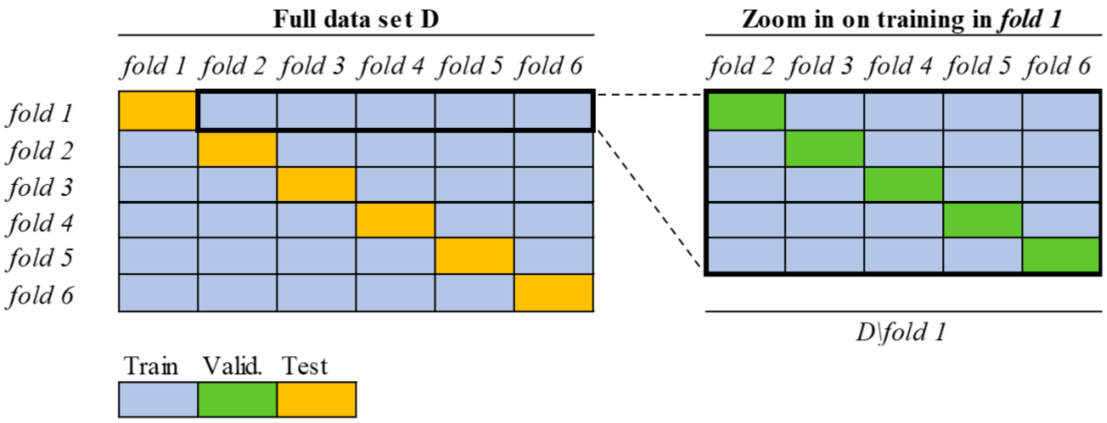

Full data set D
fold 1 fold 2 fold 3 fold 4 fold 5 fold 6
fold 1
fold 2
fold 3
fold 4
fold 5
fold 6
Zoom in on training in fold 1
fold 2 fold 3 fold 4 fold 5 fold 6
D\fold 1
Train Valid. Test

Figure 3.2 - Diagram of nested cross-validation, with inner loop of 5-fold cross-validation and outer loop of 6-fold cross validation. The hold-out test for data fold k is fold k, in orange. Considering fold k, the parameters are tuned on D\fold k, being the training portion in blue and validation in green. After the tuning, the model is trained on D\fold k using the optimal parameters estimated for the data in fold k

Source: Example adapted from Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2020). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. North American Actuarial Journal, 1–31

This method allows for the complete used of all data at hand, instead of dividing the dataset into a test and training datasets. By running through the train/test folds several times, there can be obtained a best estimate of the model performance.

3.3.3.2. Tree-specific and Hyper Parameters Choice

Machine learning models are characterized by the need of optimizing the tuning and hyperparameters, which are elements that define the model architecture. These parameters cannot be directly trained from the data, meaning that they are not model parameters. Because of this, the choice of these hyper parameters is frequently derived from experimentation. Four of the most commonly tunned parameters are described in Table 3.1.

The simplest and most commonly used method to obtain the optimal values for the previously mentioned hyper parameters is grid search (Su & Bai, 2020). This technique develops a model for each possible combination of all the hyperparameters that were provided, evaluating each of the models and therefore selecting the architecture which outputs the best results. Each of the models should be fitted to training and validating data. There are other techniques, such as random search (provides a statistical distribution for each of the hyperparameters, instead of a discrete set of values) or Bayesian optimization (sequential model-based optimization, uses the information from one experiment to improve the next, on the contrary of the previous techniques). The latter is usually more difficult to implement, as it is more complex.

Table 3.1 - Boosting and Tree-specific Hyper Parameters for the GBM

Boosting Hyper Parameters	Tree-specific Hyper Parameters
Number of trees: Total number of trees in the ensemble. As in GBM each tree is grown in sequence to fix the previous one, it is common that the total of trees will add up to a large value, such as thousands. As so, it is important to prevent overfitting, estimating this parameter through cross-validation.	**Tree depth:** Controls the depth of each individual tree. A smaller depth is associated with a higher computer efficiency, but a higher depth allows for the capturing of unique interactions in the algorithm, also increasing the possibility of overfitting. It is usually established between 1 to 5.
Learning Rate (λ): Represents the contribution of each tree to the final model and controls how quickly the algorithm learns. λ is usually chosen between 0.001 and 0.3.	**Minimum number of observations in terminal nodes:** Controls the complexity of each individual tree. A higher amount is associated with the prevention of overfitting and a smaller amount is best when there are imbalanced target classes, in classification problems. It usually ranges from 5 to 15 or could be set as a percentage of the training data.

Source: Adapted from Boehmke, B., & Greenwel, B. (2020). Hands-On Machine Learning with R. CRC Press

3.3.4. Model Interpretability and Evaluation

Interpretability is one of the most important characteristics of a model. Machine learning models are often known as "*black boxes*" given the lack of ease in the interpretation. However, one cannot simply trust the model and ignore the reason behind the decisions taken by the algorithm, which takes special importance when it comes to pricing models, as the worst nightmare of an insurance company is not being able to explain to a client why does he pay more or less than the standard.

Analogous to GLM, there are several tools that can facilitate the evaluation of the regression model, such as the mean squared error, root mean squared error, deviance and residuals (Boehmke & Greenwel, 2020).

The following subsections highlights some of the most used tools to enable model interpretability.

3.3.4.1. Variable Importance

Defined by Breiman (2001), variable importance is used to measure how important the several explanatory variables are in the prediction of the response variable. Having random forests as a base model, the author defined the importance of a variable in terms of the decrease in the loss function value when that variable is chosen as a feature to split a node on. This metric has special emphasis on unveiling the variables that actually matter for the prediction.

Therefore, having x_l as the variable of interest, $v(j)$ the split variable at index j and ΔL the difference in the loss function before and after the split on x_l (the improvements) for each tree t, it can be written as the sum over all splits where the variable of interest is included:

$$I_{x_l}(t) = \sum_{j=1}^{J-1} \mathbb{I}\,[v(j) = x_l]\Delta L \tag{3.36}$$

The main idea behind this approach is that important variables appear more often and higher in the decision tree, meaning that the sum grows faster for these variables. The values are then normalized, giving a sound idea about the relative contribution of each of the variables (Henckaerts et al., 2020). Equation (3.36) can be generalized to the ensemble techniques by averaging the importance of the variable of interest over the many trees T that compose the ensemble model:

$$I_{x_l} = \frac{1}{T}\sum_{t=1}^{T} I_{x_l}(t) \tag{3.37}$$

And therefore, comprising all trees in the GBM.

3.3.4.2. Partial Dependence Plots

On the other hand, partial dependence is characterized by the marginalization of a variable and capture of the effect that it holds on to the outcome predictions (Friedman, 2001). The plot of the partial dependence of the predicted variable and one of the independent variables can enlighten the relationship between the target and the feature, being it linear, monotonic or more complex.

Assuming $D \subset \{1,2,...,p\}$ and *V* its complement, $\boldsymbol{x}$ the training data and $\boldsymbol{x_D}$ the coordinates in *D* of $\boldsymbol{x}$. Considering the regression model, the partial dependence function can be depicted as:

$$f_D = E[f(\boldsymbol{x}_V, \boldsymbol{x}_D)] = \int f(\boldsymbol{x}_V, \boldsymbol{x}_D)\, dP(\boldsymbol{x}_V) \tag{3.38}$$

Given its complexity, the previous equation can be estimated by:

$$\hat{f}_D = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(\boldsymbol{x}_{V_i}, \boldsymbol{x}_D\right) \tag{3.39}$$

where $\boldsymbol{x}_{V_i}$ are the variables used to train the model, *n* the number of observations in the training data and $\hat{f}$ is the statistical model at use (Molnar, 2021). Overall, this method presents the average effect of the features of interest.

Taking special interest in the insurance field, a good application of this method would be to capture the relationship between the age of the driver and the amount of claims filled, which is known to be high amongst younger and middle-aged drivers.

3.3.4.3. Individual Conditional Expectation Plots

Similar to the previous topic, individual conditional expectation enlightens the dependence between the target function and a certain feature of interest, with the difference that it is presented for each sample separately, with one line per instance (Goldstein et al., 2015).

Having in consideration the elements applied in equation (3.39), for each instance in $\left\{\left(\boldsymbol{x}_{V_i}, \boldsymbol{x}_{D_i}\right)\right\}_{i\,=\,1}^{N}$, the curve $\hat{f}_D$ is plotted against $\boldsymbol{x}_{D_i}$ while $\boldsymbol{x}_{V_i}$ is fixed.

Overall, the PDP can be considered the average of the lines in an ICEP. It can be of favor to consider individual expectations over partial dependence in the event of weak interactions between the features for which the PDP is calculated and other features, because PDP could obscure a heterogeneous relationship between the variables, created by interactions. In that event, it would be wiser to analyse the ICEP, as it gives more insights and can be more intuitive to interpretate.

3.3.4.4. Friedman’s H-Statistic

The Friedman’s H-Statistic gives an estimation of the interaction strength between two feature variables by measuring how much of the prediction variance originates from the interaction effect between both variables (Friedman & Popescu, 2008). Considering $\hat{f}_D(\boldsymbol{x}_D)$ and $\hat{f}_E(\boldsymbol{x}_E)$ the one-dimensional partial dependence of the variables, as defined above in Section 3.3.4.2, and $\hat{f}_{DE}(\boldsymbol{x}_D, \boldsymbol{x}_E)$ the two-way partial dependence, the H-Statistic can be defined as:

$$H_{DE}^2 = \frac{\sum_{i=1}^{n}\left\{\hat{f}_{DE}\left(\boldsymbol{x}_D^{(i)}, \boldsymbol{x}_E^{(i)}\right) - \hat{f}_D\left(\boldsymbol{x}_D^{(i)}\right) - \hat{f}_E\left(\boldsymbol{x}_E^{(i)}\right)\right\}^2}{\sum_{i=1}^{n}\hat{f}_{DE}^2\left(\boldsymbol{x}_D^{(i)}, \boldsymbol{x}_E^{(i)}\right)} \tag{3.40}$$

4. DESCRIPTIVE AND EXPLORATORY ANALYSIS OF THE DATASET

4.1. Dataset Information

The dataset used on this project was composed of 2 464 181 observations of exposure and 75 263 observations of claims data for the material TPL cover, correspondent to 799 587 distinct policies, collected between 1 January 2016 and 31 December of 2019, for passenger private cars. The policies were grouped according to the several feature variables and years. A policy that did not register any change in its structure or characteristics would be depicted in four observations.

Besides the response variables, this dataset includes 36 feature variables, related to the client (such as the client's age, city or profession), the policy (such as the seniority) and the vehicle (such as the brand, age or horsepower). Table B.1 in Appendix B presents a summary of the independent variables used to develop the models.

4.2. Collection and Treatment of the Data

The raw data was collected from the insurance company's data warehouse, and afterwards treated using the Microsoft SQL Server Management Studio software.

To obtain a dataset of valuable and consistent information, there was the need to treat some of the variables, as there was missing information and obvious errors, very likely generated due to human-error. In most cases, missing information or atypical outliers were allocated to the level *999* (numeric variables) or *unknown* (textual variables). Only in some specific events of known human-error, being impossible to track the original value, the information was deleted, such as observations where the earned exposure was negative.

Similar to what had been done in previous years, the final dataset only included claims that originated incurred costs superior to 5 €.

Some of the variables were removed from this study prior to modelling, given the evident lack of quality or the very high similarity. These last were variables that were collected from both the clients' table and the policies' table. For example, the *Driver Age* originated from the clients table was removed, as only the *Driver Age* originated from the object table was used. It is common to have differences between variables that were expected to match, given that the client allocated to the policy is not always the regular driver, from which the information is collected in the objects table.

Figure C.1 in Appendix C demonstrates the correlation matrix, allowing for the study of the correlation among variables. Despite the last type of variables mentioned above, which clearly show an extremely high correlation, there seems to be a strong correlation between the location-related variables, such as *Concelho*, *Distrito*, *Delegation* and *Circulation Zone*, as well as between *Delegation* and *UEN* or *NBexe* and *Driver Age.* The *Own Damage* variable is correlated with several other. This information should be taken into account when it comes to modelling.

After the treatment of the variables, the dataset was left with 2 464 181 observations and 21 variables, depicted in Table B.2 in Appendix B.

Overall, considering the 799 587 distinct policies, this dataset presented 78 264 claims over the four years in study, leading to a total incurred cost of 97 908 920 €, as shown in Table 4.1.

Table 4.1 - Policies and claims count and cost from period 2016-2019

Number of Policies	Number of Claims	Total Incurred Cost
799 587	78 264	97 908 920 €

Source: Authors preparation

4.3. Univariate Descriptive Analysis

The frequency model response variable is the *Claim Count* (discrete variable), obtained through the original variable *Claim Number*. The severity model response variable is the cost of claims, *Claim Amount* (continuous variable), directly collected from the dataset.

In this Section it is presented a univariate descriptive analysis of the variables, with the goal of better describing and understanding them, which will be of great value and use when developing the models. Given the vast extension of feature variables taken into account, the descriptive analysis that is depicted in this paper will comprise the variables that will most likely reveal to be significant to the final models, taking into account previous models developed by the insurance company and the usual behaviour of the insurance market.

4.3.1. Response Variables

4.3.1.1. Claim Count

Used in the frequency model as response variable, *Claim Count* is a quantitative variable that represents the number of claims reported per policy, per year. The elementary statistics related to this variable are shown in Table 4.2.

Table 4.2 - Elementary descriptive statistics of claim count, weighted by exposure, from period 2016-2019

Total Number of Policy Records	Mean	Variance	Skewness	Kurtosis
2 464 181	0.04834	0.09289	16.40195	2 285.577

Source: Authors preparation.

According to the information displayed in the table above, this dataset presents 2 464 181 different policy records from year 2016 to 2019, leading to an average of 0.048 annual reported claims, per annuity and exposure.

Given that the variance is higher than the mean, there is proof that the distribution of claim count is over-dispersed. A positive and high value of skewness translates into very few observations on the right side of the distribution, which is very common for claim count. Lastly, 2 285.577 of kurtosis indicates that the distribution of claim count is very peaked, being a very narrow distribution with most of the responses in the left side, in this case for 0 claims, as Table 4.3 depicts.

Table 4.3 - Distribution of annual claims per policy record

Claims	**0**	**1**	**2**	**3**	**4**	**5**
Policy Records	2 388 567	73 131	2 334	135	10	4

Source: Authors preparation

On Figure 4.1 it is possible to confirm the assumptions derived from the elementary descriptive statistics mentioned above, being clear that the majority of observations fall on 0 claims, with only 3.07% of policy records having registered one or more claims in the period between 2016 and 2019.

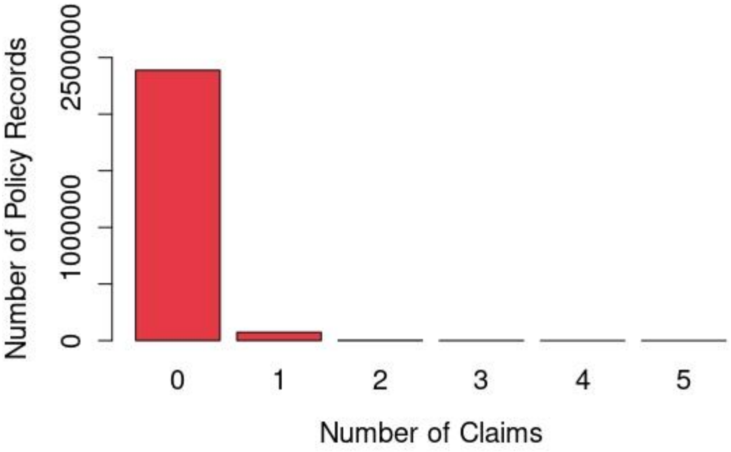


Figure 4.1 - Number of annual claims per Policy Record

Source: Authors preparation

4.3.1.2. Claim Amount

On the other hand, *Claim Amount* is the response variable related to the severity model. In Table 4.4 there are presented the elementary statistics related to this variable. The minimum amount filled corresponds to 5 € and the maximum to 108 603.70 €. Overall, the average cost per claim is 1 251.01 € and the average cost per policy is 122.45 €.

Table 4.4 - Elementary descriptive statistics of *Claim Amount*

Number of Policies	Number of Policies with associated costs	Average Cost per Claim	Standard Deviation of Cost per Claim	Average Cost per Policy	Standard Deviation of Cost per Policy
799 587	70 132	1 251.01 €	1 972.37 €	122.45 €	749.75 €

Source: Authors preparation

Table 4.5 depicts the right-skewness nature of the distribution of the claim costs, with 99% of claims having an associated cost under 8 724.04 €. Considering this threshold, it is possible to visualize the claims distribution in Figure 4.2.

Table 4.5 – Quantiles of *Claim Amount*

50%	90%	92.5%	95%	99%
836.60 €	3 116.90 €	3 428.23 €	3 952.60 €	8 724.04 €

Source: Authors preparation

As it is expected, the vast majority of claims have an associated cost situated around the observed average.

This distribution has two peaks: the first one, at around 300 €, is associated with IDS debtor and creditor claims, whose costs are average costs, monthly pre-defined by the convention with APS – Portuguese Association of Insurers (and therefore don't follow a theoretical distribution) and the second one is the natural peak associated with the average value of *Claim Amount*.

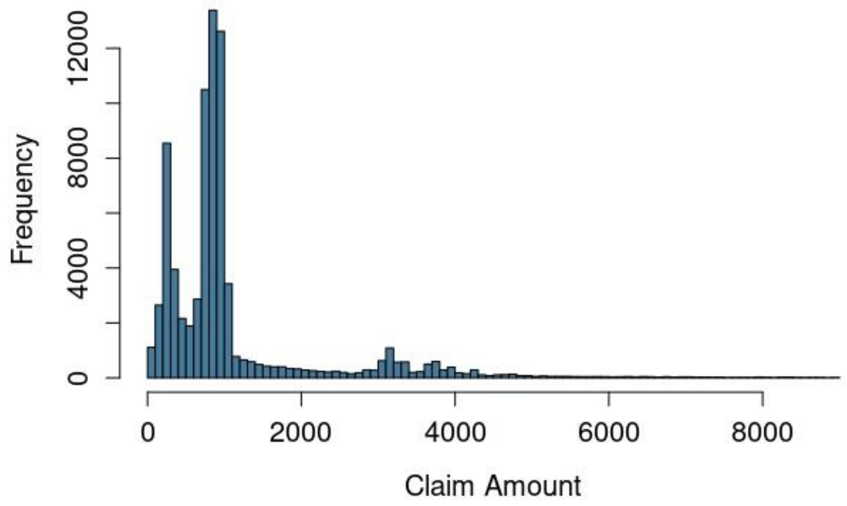


Figure 4.2 - Histogram of the *Claim Amount*

Source: Authors preparation

4.3.2. Offset Variable

4.3.2.1. Earned Exposure

This quantitative variable depicts the actual amount of exposure an insured vehicle has been exposed, measured as a fraction of the year. *Earned Exposure* is used as an offset in the frequency model, given that there must be different weights for claims suffered in policies with low exposure.

In this dataset, 40.3% of policy records have one full year of exposure, as Figure 4.3 depicts.

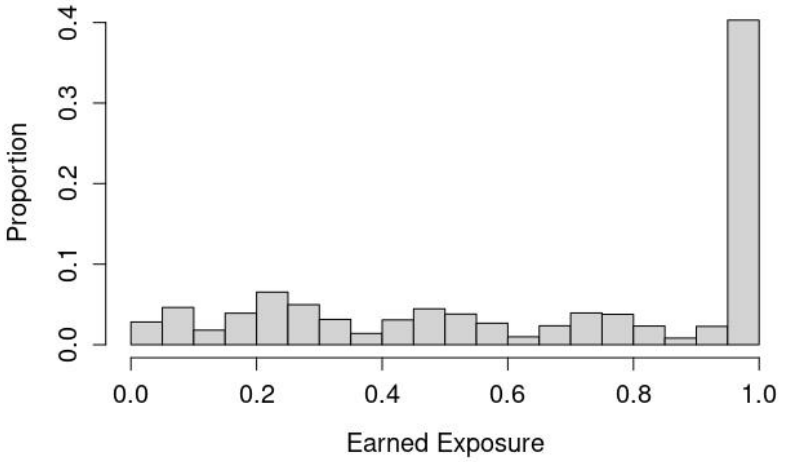


Figure 4.3 - Proportion of *Earned Exposure*

Source: Authors preparation

4.3.3. Feature Variables

In the following graphs, the y axis values (Average Cost and Annual Claim Frequency) have been removed, in order to comply with the insurance company privacy policy. For the numeric variables, the several levels were grouped sequentially according to the observed values, to facilitate the exploratory analysis. The depicted levels were not the final choice used to model.

The dashed lines represent the overall claim frequency of 0.0483 and average cost per claim of 1 251€. The y axis values have been replaced with two percentage values, representing the variation above (or bellow) average associated with the level with the highest (lowest) value, which gives a good indication on the behaviour of the data that is being represented. The vertical lines represent the confidence interval for the estimations.

4.3.3.1. Driver Age

The quantitative discrete variable *Driver Age* ranges from values between 18 and 93, being the average age in the portfolio 51 years old. In Figure 4.4 (left), there is a clear difference of frequency amongst the different age groups, with the peak of frequency at 21 years of age. There is then a decrease at 36 years of age, then again a very slight increase around the 50-years mark (most likely due to different drivers, usually descendants that have come to legal age to drive). It is followed by a second decrease, down to the age group of 70, when it is followed by the last upward wave, peaking at 92 years of age. The average cost follows a very similar pattern to the claim frequency, registering the peak at younger ages. The peak in frequency mentioned above is stated clear in Figure 4.4 (right), with ages [18, 30[depicting double the average frequency.

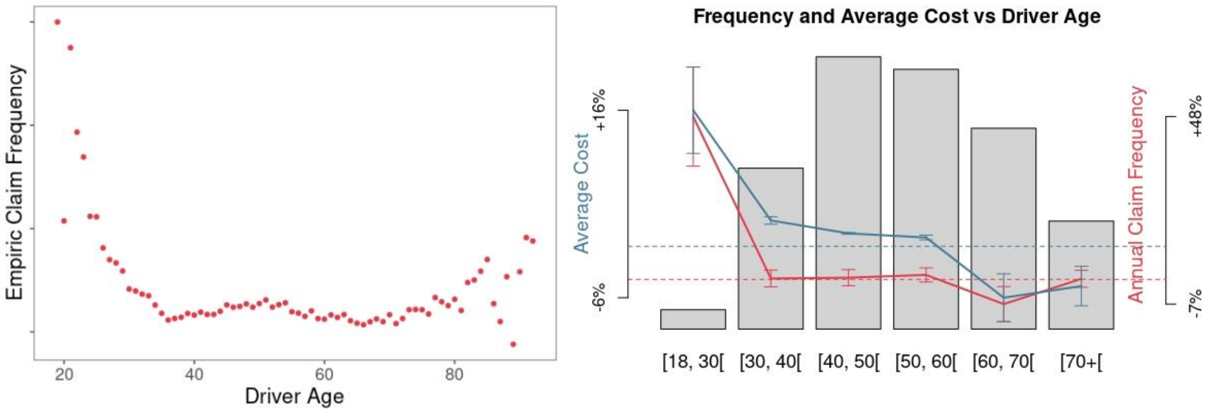


Figure 4.4 - Empiric claim frequency vs *Driver Age* (left) and Average Cost and Annual Claim Frequency vs *Driver Age* (right)

Source: Authors preparation

4.3.3.2. Payment Instalments

The *Payment Instalments* variable has four levels, depending on the payment plan chosen by the insured person. In this portfolio, most of the observations fall on annual payment, followed by semi-annual and monthly, that have similar exposure.

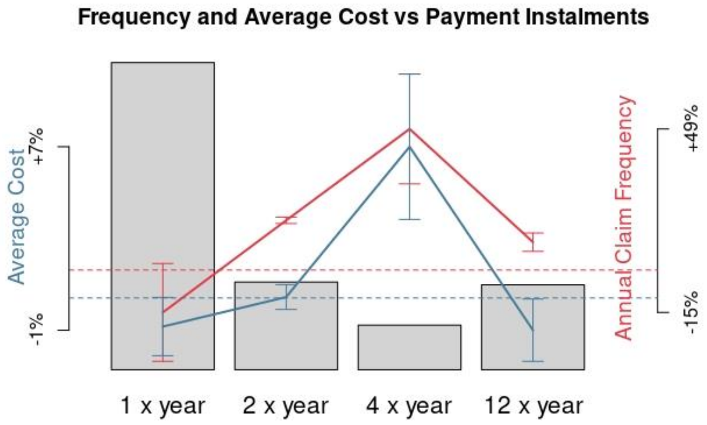


Figure 4.5 - Average Cost and Annual Claim Frequency vs *Payment Instalments*

Source: Authors preparation

Trimester payments are the least chosen option by the clients. The average cost and frequency follow a similar pattern, registering the lowest values for the annually payments and a peak for the trimester payments, with an increase above average of around 88 € in cost and 0.02 pp in frequency, representing double the average frequency, as Figure 4.5 depicts.

The fact that trimester payments can induce especially higher values of frequency could appear to be unlikely, however it has become a known fact registered by the insurance company over the years. This behaviour has been studied and a possible reason behind it could be the fact that clients that file several claims have a known behaviour of switching between insurance companies, and thereby have a preference for paying in fractions. Monthly payments have been obliged to be accompanied by direct debit type of payment since before 2016, which could lead this group of clients to select quarterly payments, the shortest period after monthly, and therefore having the opportunity to not pay the last instalment right after the claim occurrence.

4.3.3.3. Direct Debit Payment

The categorical variable *Direct Debit Payment* differentiates the policies that have direct debit as the payment option of choice. As it is possible to confirm in Figure 4.6, most of the policies in portfolio do not have direct debit, having those specific policies a lower registered frequency, as the average cost remains very similar. For policies with direct debit, frequency rises around 9% above average. This is a result of *Direct Debt Payment* being mandatory for policies with monthly *Payment Instalments*, given the bad risk that the latter represent, as depicted in Figure 4.6. This behavior is also translated by the correlation factor between the Direct Debit Payment and *Payment Instalments* variables of 0.77, as depicted in the correlation matrix on Figure C.1 in Appendix C.

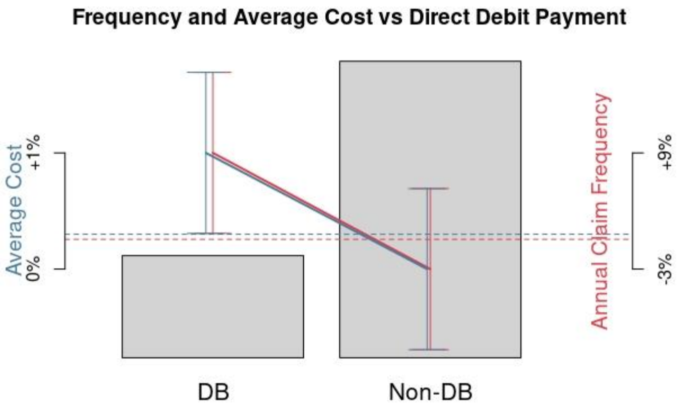


Figure 4.6 - Average Cost and Annual Claim Frequency vs *Direct Debit Payment*

Source: Authors preparation

4.3.3.4. Years of Driving

Constructed upon the variable date of driving license, *Years of Driving* is a numeric discrete variable that clearly outputs different empiric frequency values, according to the several levels, as Figure 4.7 (left) shows. In this dataset, the average time for which an insured person has had a driver's license is 27 years, being the minimum 1 and the maximum 75 years.

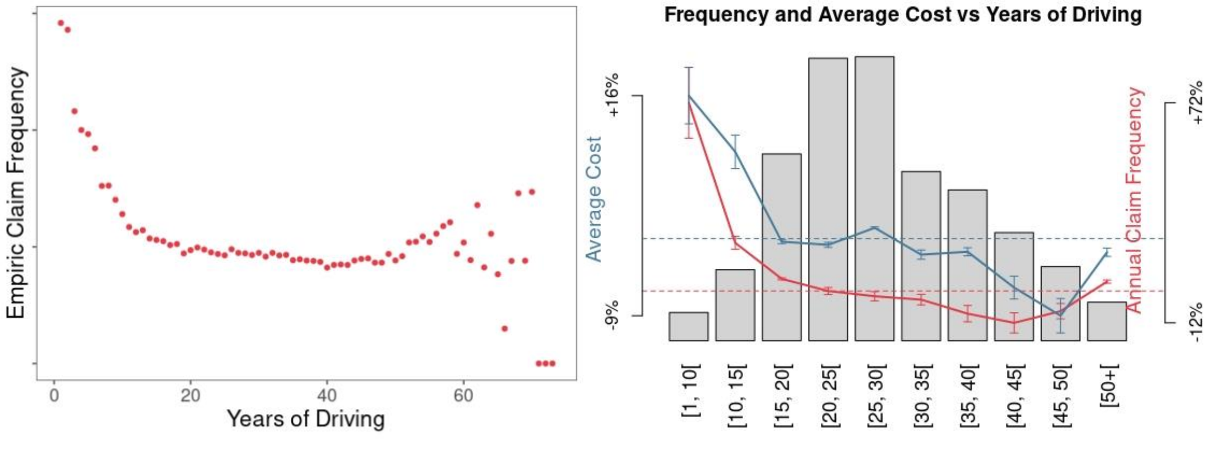


Figure 4.7 - Empiric claim frequency vs *Years of Driving* (left) and Average Cost and Annual Claim Frequency vs *Years of Driving* (right)

Source: Authors preparation

Years of Driving was grouped in 10 levels, as depicted in Figure 4.7 (right). As it would be expected given the analysis of the variable *Driver Age* presented above in this Section (where the majority of the insured people were aged between 40 and 60) the highest portion of exposure falls into levels [20, 25[and [25, 30[. Similarly, the highest average cost and frequency are registered for the first levels of *Years of Driving*.

This variable presents the same trend as *Driver Age*, with a slight increase for average cost surrounding the 25 years of driving, which could once again be explained by the usage of the vehicle by the younger descendants.

It is clear by Figure 4.7 (right) that less experienced drivers have an aggravated behaviour, especially for frequency, with observed values being much higher than the average, for this group.

4.3.3.5. Policy Time on Book

Once again another numeric discrete variable*, Policy Time on Book*, as the name indicates, represents the time, in years, for which the policy has been active. In this dataset, the policy that has been active for longest, has 62 years of time on book, being the average 4.82. As Figure 4.8 (left) illustrates, the empiric frequency decreases as time on book increases, having some outliers between levels 30 and 50, being relatively approximate to 0 for policies with more than 30 years of time on book.

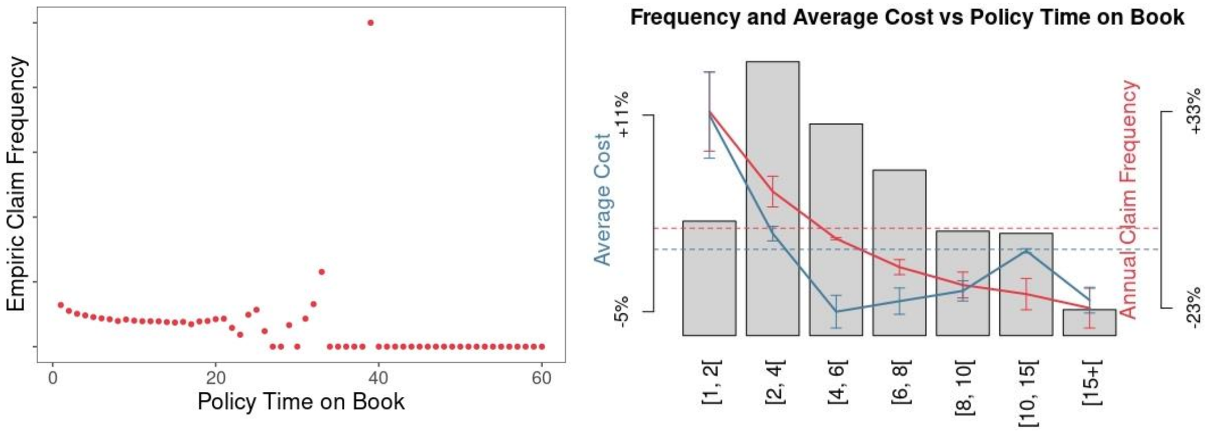


Figure 4.8 - Empiric frequency vs *Policy Time on Book* (left) and Average Cost and Annual Claim Frequency vs *Policy Time on Book* (right)

Source: Authors preparation

On Figure 4.8 (right), where the different levels of *Policy Time on Book* are grouped, there is a descending trend in frequency as time on book increases, with levels with more than 4 years of time on book being bellow average. On the other hand, the average cost also starts with a descending trend, however, from level [4,6[to [10, 15[, there is registered an increase in costs.

4.3.3.6. District

District is a categorical variable, which represents the district where the policyholder lives, ranging from code 1 to code 22. As Figure 4.9 shows, the average cost and frequency have a very different behaviour depending on each level. Districts 13, 14 and 17 are where the earned exposure is concentrated, and alongside with districts 4 and 19, represent the sample of districts with frequency above average.

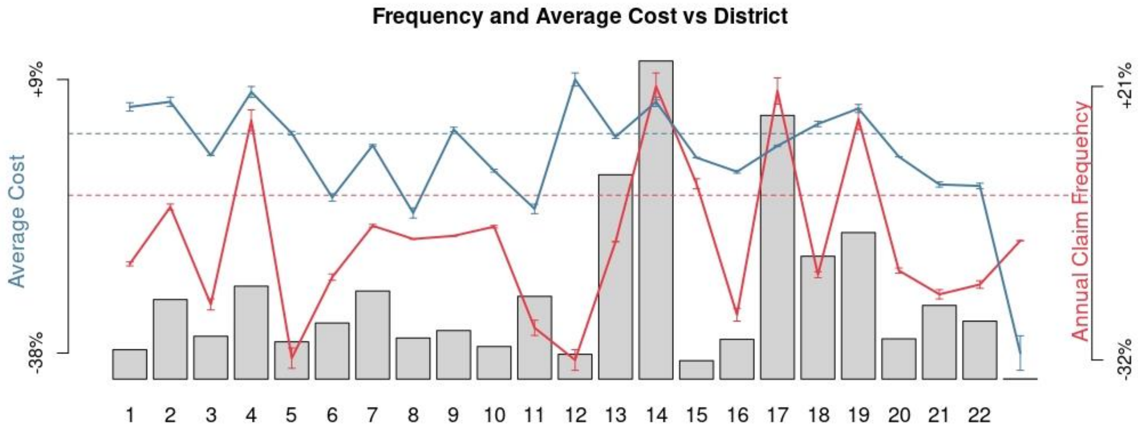


Figure 4.9 - Average Cost and Annual Claim Frequency vs *District*

Source: Authors preparation

The average cost surpasses the global average at districts 1, 2, 4, 9, 12, 14 and 19. There are several districts that present very similar behaviour, which indicates they could be grouped.

4.3.3.7. Vehicle Brand

As the name indicates, *Vehicle Brand* is the categorical variable that indicates the brand of the insured vehicle.

As Figure 4.10 depicts, for the levels with higher exposure in the portfolio, such as brands 008, 015, 023, 024, 047, 056, 057, 064, 079 and 083, the average cost and claim frequency are very similar to the global averages. For levels with lower exposure, the values in study present a very high fluctuation.

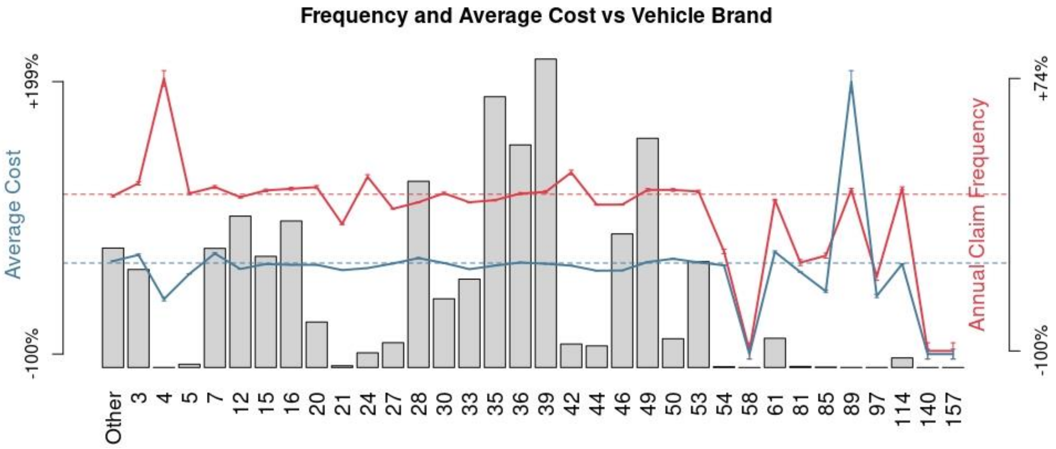


Figure 4.10 - Average Cost and Annual Claim Frequency vs *Vehicle Brand*

Source: Authors preparation

4.3.3.8. Seat Capacity

The *Seat Capacity* variable depicts the number of seats that the insured vehicle possesses. As it is clear by Figure 4.11 the vast majority of the vehicles included in this dataset have a capacity of 5 seats, followed by 2 and 4. Level 5 has an average cost and frequency spot on the global average, while the rest of the levels register some fluctuations, which is common when the exposure is concentrated in a single level.

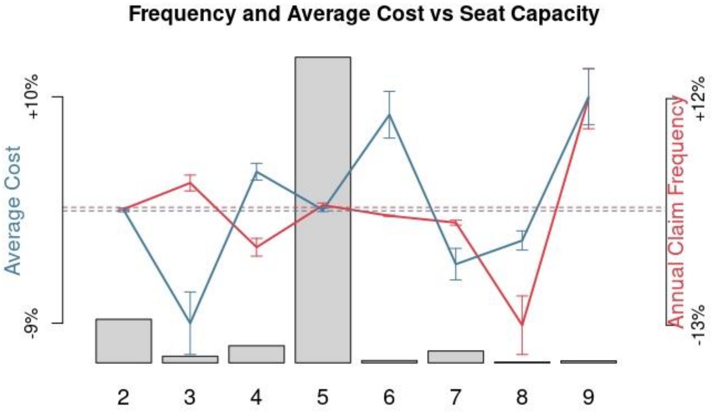


Figure 4.11 - Average Cost and Annual Claim Frequency vs *Seat Capacity*

Source: Authors preparation

4.3.3.9. Vehicle Age

Being a numeric discrete variable, created based on the year of manufacture, *Vehicle Age* portraits the age of the vehicle at the time of the policy annuity. In this dataset, the vehicles have an average of 15.61 years, with a maximum of 95 years. As is depicted on Figure 4.12 (left), the empiric frequency presents an upward trend, up to vehicles with 25 years of age, decreasing continuously until the age of 30, where the empiric frequency is 23% below average, at approximately 0. This could be due to the fact of the vehicles being quite old, and so insured people do not use them often.

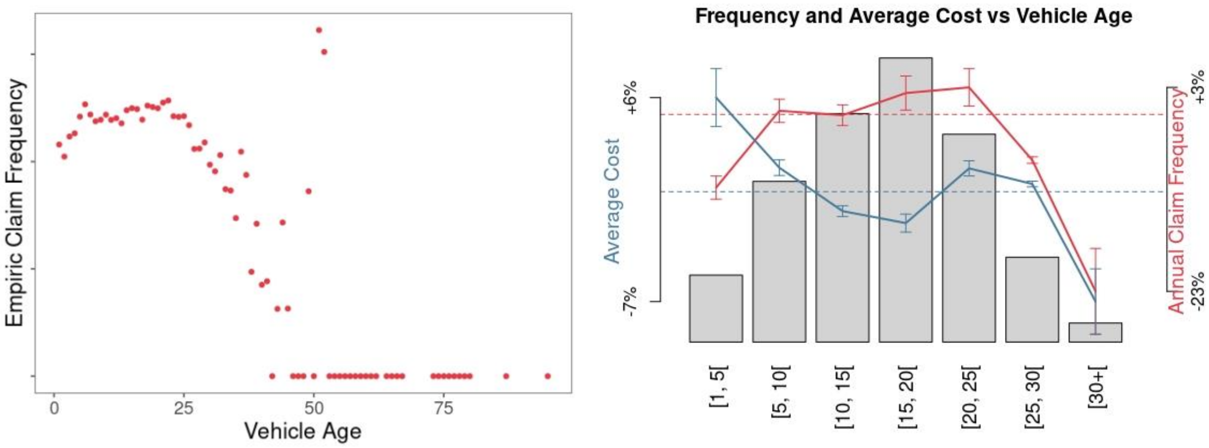


Figure 4.12 - Empiric claim frequency vs *Vehicle Age* (left) and Average Cost and Annual Claim Frequency vs *Vehicle Age* (right)

Source: Authors preparation

As could be confirmed in Figure 4.12 (right), these last mentioned vehicles represent a very small portion of the vehicles in this dataset, as the exposure is very low. The average cost decreases until level [15,20[, followed by an increase in the following level and finally reaching the minimum of 7% below average at the last level of vehicles with over 30 years of age, as mentioned above.

5. RESULTS

5.1. Fitting the Distributions

As mentioned above in Section 4.1, the response variables, *Claim Count* and *Claim Amount*, usually follow certain distributions from the exponential family. It is based on that assumption that the frequency and severity models are developed.

In this section it is presented the study of which distribution best describes the observed behavior of these variable.

5.1.1. Claim Count

Representing the number of claims reported per policy per year, this quantitative variable has been described in the section above as having the vast majority of observations allocated to 0 claims, followed by a decrease in the amount of observations for 1, 2 and 3 occurrences.

Given the behaviour of the data, it would be natural to try to fit the Poisson or the Negative Binomial distributions, known as good representations of count distributions. Table 5.1 below includes the observed probability and expected frequency considering both distributions. Records with five claims were grouped with $x = 4$, given that there were only four observations.

Table 5.1 - Modelling Claim Count: Negative Binomial and Poisson Distributions

x	O_x	p_x^{obs}	E_x^{NB}	p_x^{NB}	E_x^{Poi}	p_x^{Poi}
0	2 388 567	0.9693147	2 388 595.45	0.9693262	2 387 146.80	0.9687384
1	73 131	0.0296776	7 009.63	0.0296284	75 817.34	0.0307678
2	2 334	0.0009472	2 485.20	0.0010085	1 204.00	0.0004886
3	135	0.000055	87.47	0.0000355	12.75	0.0000052
4	14	0.000006	3.13	0.0000013	0.10	0.0000000
Total	2 464 181	1	2 464 181	1	2 464 181	1
			χ^2	68.05	χ^2	3 798.78
			p-value	≈ 0	**p-value**	≈ 0

x - Claim count; O_x - Number of observations; p_x^{obs} - Observed probability; E_x - Expected Frequency

Source: Authors preparation

Considering the Negative Binomial, the distribution parameters were estimated by maximum likelihood, achieving $\widehat{size}$=0.8146 and $\widehat{mu}$=0.0318. Given that the Chi-Square test returned a p-value of approximately 0, the hypothesis that the Negative Binomial distribution fits to the data in study is rejected at any given significance level. Figure 5.1 (left) shows that despite the rejection of the null hypothesis, there seems to be a good approximation of observed frequency.

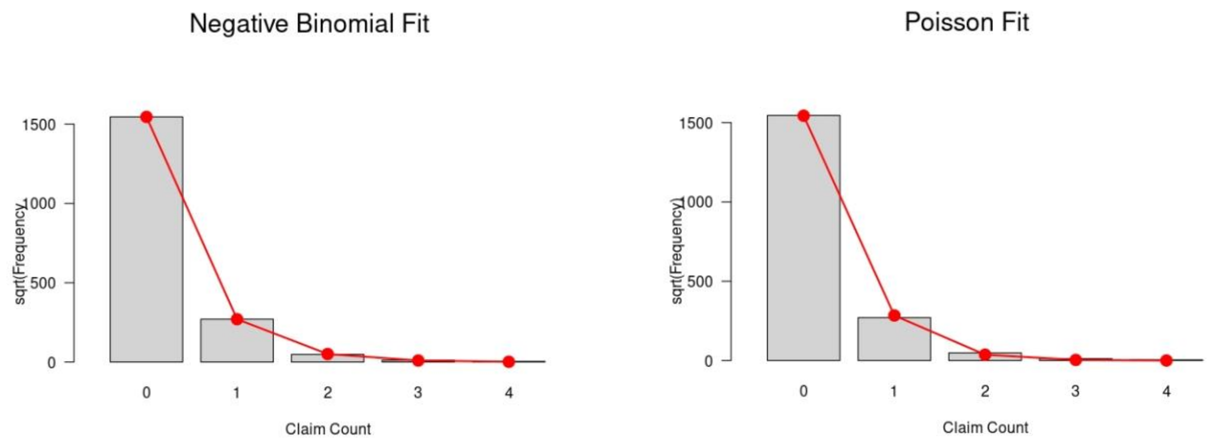


Figure 5.1 - Fitting of the Negative Binomial distribution with *size*=0.8146 and *mu*=0.0318 (left) and Poisson distribution with λ=0.03176065 (right) to *Claim Count.*

Source: Authors preparation

Similarly, the Poisson parameter lambda was estimated as being the observed mean, $\hat{\lambda}$=0.03176065. By repeating the Chi-Square test for goodness-of-fit, the p-value returned was approximately 0, being the null hypothesis that the Poisson distribution fits the data rejected. Similarly, Figure 5.1 (right) represents what seems to be a good fit to the Poisson distribution, despite the results stated.

The fact that both distributions failed the goodness-of-fit test, despite the results in Figure 5.1 indicating a good fit, could be explained by the vast majority of observations falling into zero claims filed or the size of the sample in study. This dataset is composed by real data collected by the insurance company, and it is very unlikely that real-life events can be precisely described by probability distributions, especially events whose probability is very slim, such as policies having over one claim per year. As it is clear in Table 5.1 , both distributions are able to predict the probability of filling 0 claims in one year fairly well, with that probability being more distant from the observed probability as claim count increases. Sample size can also impact the results of the goodness-of-fit test given that the Chi-Square test is sensitive to it. With very large samples, even minuscule distances between the estimate and the null hypothesis become statistically significant. When facing this problem, the rejection of the null hypothesis should not be based solely on the *p-value* (Lin et al., 2013).

Given all this information and taking into consideration that the pricing team of the company has developed previous frequency models assuming that the claim count follows a Poisson distribution, this will be the chosen assumption. Not forgetting that the observed variance was over the average, thus being in the presence of over-dispersion, the chosen error structure on EMBLEM is Poisson but the scale parameter, also known as dispersion parameter, will be estimated from the Pearson estimator, and not fixed at 1 as is usual when the mean matches the variance. This is the equivalent of assuming over-dispersion and using quasi-poisson as family parameter when developing GLM in R.

5.1.2. Claim Amount

As stated previously, *Claim Amount* is a continuous variable that represents the costs associated with the filed claims. Its distribution is right skewed, as claims with very high costs are more unlikely to happen.

Considering only the observations where there were reported claims, the following boxplots depict the distribution of claim amount in the dataset before any capping is done and below with a capping at the 99% quantile, corresponding to a maximum cost of 8 724.04 €.

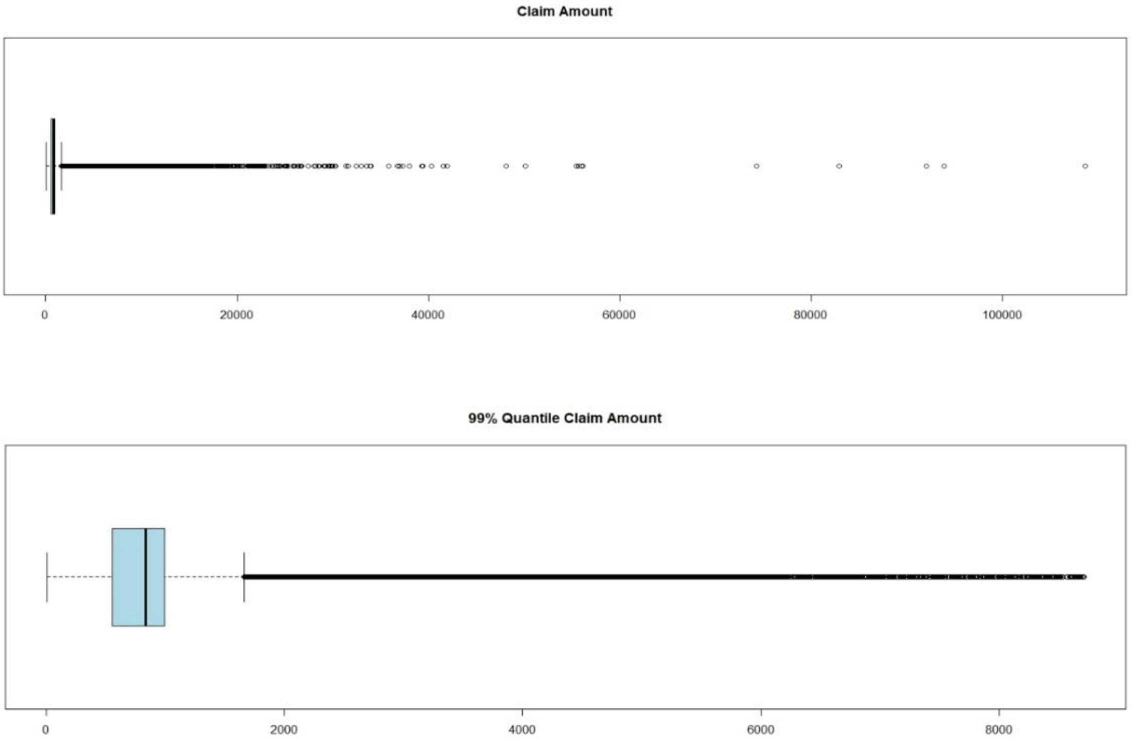


Figure 5.2 Total (top) and 99% quantile (bottom) boxplot of *Claim Amount.*

Source: Authors preparation

As was mentioned in the beginning of Section 4.3.1.2 this dataset included settled IDS claims, which do not follow any theoretical probability distribution. Because of so, those type of claims were removed from the dataset. Table 5.2 bellow presents the new descriptive statistics of claim amount.

Given this change, the quantiles of the distribution have also suffered changes, as Table 5.3 depicts. The amounts are higher, because the removed claims were associated with relatively small costs, bellow average.

Table 5.2 - Elementary descriptive statistics of claim amount, after removing IDS claims

Number of Claims with associated costs	Average Cost per Claim	Standard Deviation of Cost per Claim
22 262	1 931.94 €	3 312.96 €

Source: Authors preparation

Table 5.3 - Quantiles of claim amount, after removing IDS claims

50%	90%	92,5%	95%	99%
913.34 €	4 517.21 €	6 992.09 €	9 792.54 €	15 000.00 €

Source: Authors preparation

The following graph depicts the claim amount distribution after this alteration, limited at the new 99% percentile, 15 000 €, being visible only one peak now, at around 400 €.

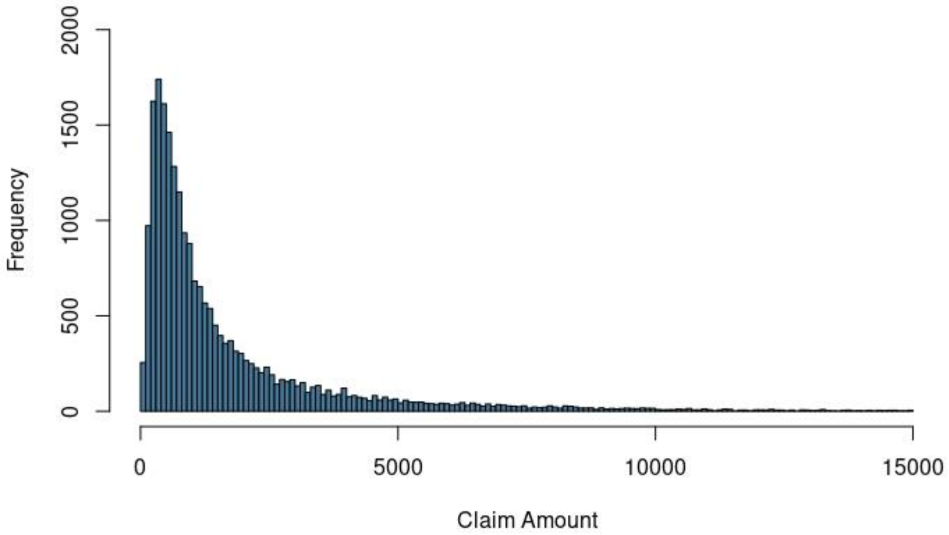


Figure 5.3 - Histogram of the Claim Amount after removing IDS claims

Source: Authors preparation

High costs have a low probability of occurrence, and normally do not follow the same probability distribution as the "usual" type of claims. Given so, there were performed several goodness-of-fit tests for the fitting to the Gamma distribution, considering different maximum limits of costs, whose results are summarized in Table 5.4 below.

Table 5.4 - Estimated parameters and p-value for the Gamma distribution, according to different limits

Claim Amount Limit	Nr. of Claims	$\widehat{\text{shape}}$	$\widehat{\text{scale}}$	V	p-value
A. 108 603.70 € (no limit)	22 262	0.7421	2 603.46	151.94	$< 2.2 \times 10^{-16}$
B. 15 000.00 € (99% quantile)	22 037	0.9001	1 901.72	37.02	$< 2.2 \times 10^{-16}$
C. 4 517.21 € (90% quantile)	20 035	1.4481	793.27	12.23	$< 2.2 \times 10^{-16}$
D. 3 359.67 € (85.36% quantile)	19 002	1.7017	587.89	2.76	0.05108

Source: Authors preparation

The dataset A comprises all claims observations and datasets B, C and D include the observations correspondent to the 99%, 90% and 83.87% quantile, respectively. The specific quantile of 85.36% was chosen as it was the highest threshold that outputted a p-value of approximately 0.05, the usual significance level. This indicates that the null hypothesis that the claim amount follows a Gamma distribution is not rejected.

Given these results, the observations to be used to model claim severity will be those obtained after applying limit D. The elementary descriptive statistics alongside with the histogram of average claim amount and density functions are depicted below.

Table 5.5 - Elementary descriptive statistics of claim amount, after removing IDS claims and considering limit D

Number of Claims with associated costs	Average Cost per Claim	Standard Deviation of Cost per Claim
19 002	1 000.45 €	772.76 €

Source: Authors preparation

Table 5.6 - Quantiles of claim amount, after removing IDS claims and considering limit D

50%	90%	92,5%	95%	99%
749.14 €	2 018.51 €	2 444.77 €	2 700.00 €	3 196.23 €

Source: Authors preparation

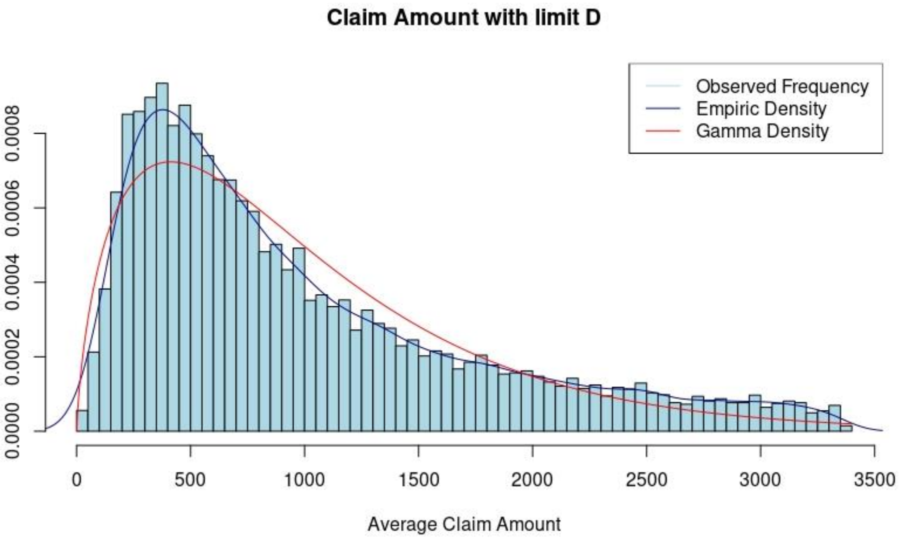


Figure 5.4 - Empiric density and Gamma density considering limit D

Source: Authors preparation

As it is clear by Figure 5.4 the aforementioned alterations to the initial claims dataset have resulted in a combination of observations that can be very well described by a Gamma distribution.

5.2. Modelling Results

In this Section, there are presented the results of the two different approaches, for both the frequency and severity modelling. Afterwards, the models are compared in terms of total deviance and residuals, in order to have a glimpse of which methodology outputs the best results.

5.2.1. Generalized Linear Models

To develop the GLM models, the dataset was imported to the EMBLEM Software, with the response variable being chosen accordingly to the problem in question: Number of reported claims alongside with the exposure to be used as an offset for the frequency model and cost of claims alongside with amount of reported claims to be used as an offset for the severity model. Both datasets were divided in training (80%) and testing (20%).

5.2.1.1. Claim Frequency

Assuming that the frequency response variable follows a Poisson distribution with overdispersion, the Emblem model was defined with Poisson error structure and the dispersion parameter was estimated using Pearson Chi-Squared statistic.

After the data selection made in Section 4.2 there was a group of 21 feature variables that could be chosen from to model claim frequency, which can be consulted in Table B.2 in Appendix B.

The established procedure to make the selection of these variables starts with the addition of those variables that were significant in previously developed models. In this case, 10 out of those 21 variables were considered as significant in the previous frequency model developed by the pricing team, and were added to the model, one at a time. With each variable that was added, it was checked if there was the need to perform variable simplification, by grouping the different levels. Firstly, it was always taken into consideration the observed frequency (levels with similar observed frequency would be grouped together), and also the observed amount of exposure (levels with little to no exposure would be grouped with the base level) or existence of unknown levels (grouped with the base level). The nature of the grouping would differ with the nature of the variable in question: for discrete variables, the grouping is done in the form of *Grouped Factors* and for the numerical variables, *Variates* (the orthogonal polynomials mentioned in Section 3.2.4.2). The selected grouping would be evaluated through the standard errors of the parameter estimates of each final level. If the standard error in percentage is smaller than 50%, then both levels are significantly different, and the grouping should remain as it is.

After the simplification of each variable, it was evaluated if that chosen variable was significant to the model or not. Comparing the nested models, the Likelihood Ratio test was performed, and the variable was kept if the *p-value* was lower than 5%, meaning that the models are significantly different.

Out of the initial variables, only eight were kept, being *Seat Capacity* and *Own Damage Cover* excluded because they did not add value to the model.

Afterwards, the process was repeated for the other 11 variables that were not included in the model that was developed years prior. Out of those, *Horse Power* was the only addition. Table 5.7 includes the listing of the final variables that were considered significant in the frequency model.

Table 5.7 - Significant Variables chosen for the GLM frequency model

Significant Variables in the Frequency Model
Fuel
Vehicle Brand
Payment Instalments
District
Driver Age
Years of Driving
Vehicle Age
Horse Power
Policy Time on Book

Source: Authors preparation

The selected variables were tested in order to detect possible interactions. It was found an interaction between *District* and *Policy Time on Book*, as is represented in Figure D.1 in Appendix D. As the policy time on book increases, the gap between the predicted values for the base level (dark blue) and other *District* levels decreases (especially considering the initial levels of *Policy Time on Book* hold the majority of the exposure). By rescaling the graph to the *District* base level, as in Figure D.2 in Appendix D, this relation becomes more evident.

This led to the final frequency model, that was composed of nine feature variables and one interaction. Table 5.8 presents the tariff structure for claim frequency, on the training set, on which all factors are statistically significant. It should be noted that the betas included in the table below have been multiplied by a constant inferior to 1, in order to proceed accordingly to the insurance company data sensitivity policy.

Table 5.8 - Tariff structure for claim frequency, using the training dataset

	Std. Error	**Std. Error %**	**Exp(β)**
Intercept	0.00886	0.3	0.05047
Fuel – G2	0.00969	9.2	0.88229
Vehicle Brand – G2	0.01037	26.7	1.01881
Payment Instalments – G2	0.00991	4.6	1.21706
Payment Instalments – G3	0.01456	3.4	1.49960
District – G2	0.01159	3	0.66522
District – G3	0.01061	3.9	0.74764
v1 Driver Age – OPoly1	0.02098	27.5	1.05781
v1 Driver Age – OPoly2	0.01117	30.8	1.01616
v2 Driver Age – OPoly1	0.00447	9.5	1.02694
v2 Driver Age – OPoly2	0.00383	17.7	0.95903
v3 Driver Age – OPoly1	0.0068	17.9	1.01802
v3 Driver Age – OPoly2	0.00552	11.5	1.02802
v1 Years of Licence – OPoly1	0.01452	5.9	0.76646
v1 Vehicle Age – OPoly1	0.00489	9.3	1.03302
v2 Vehicle Age – OPoly1	0.00436	14.6	0.95109
v2 Vehicle Age – OPoly2	0.00413	37.3	0.96922
v1 Horse Power – OPoly1	0.00787	26.4	0.95119
v1 Policy Time on Book – OPoly1	0.00615	4.1	0.84388
v1 Policy Time on Book – OPoly2	0.00367	11.3	1.01234
*District - G2*v1 Policy Time on Book*	0.01135	23.2	1.02910
*District - G3*v1 Policy Time on Book*	0.01022	23	1.02459

Source: Authors preparation

For the categorical variables and grouped factors, it is possible to directly interpretate the table above, as the last column represents the final model coefficients. For the orthogonal polynomials, there is the need of an extra step to retrieve the final values. Nonetheless, some conclusions can be drawn from this table, such as the higher risk related to clients whose *Vehicle Brand* is in included in G2 or that have chosen to pay accordingly to *the Payment Instalments* defined in G3.

5.2.1.2. Claim Severity

The approach followed to model claim severity was very similar to the one described above for claim frequency.

As stated in Section 5.1.2, the severity response variable fits best to a Gamma distribution, being this the error structure chosen to model in the EMBLEM software.

First, there were tested the five variables that were present in the prior severity model, following the same methodology as described in the Section above.

Out of the previous selection, only *District* and *Driver Age* were included in the model. Afterwards, the other variables were tested, and *Vehicle Brand* was also considered significant.

Table 5.9 - Significant Variables chosen for the GLM severity model

Significant Variables in the Severity Model
Vehicle Brand
District
Driver Age

Source: Authors preparation

This selection outputted a severity model with three feature variables, as depicted in the Table 5.9 above. It is not uncommon for severity models to include less feature variables, as it is harder to explain the cost of claims rather than the frequency in which they occur. The sample size being much smaller can also lead to this disparity, increasing the volatility.

The tariff structure of the model is represented in the table below and once again it should be noted that the betas included in the table below have been multiplied by a constant inferior to 1, in order to proceed accordingly to the insurance company data sensitivity policy.

It is now clear that clients who live in the *Districts* included in Group 3 present a higher average cost than the average client. On the other hand, clients whose *Vehicle Brand* is included in Group 2 show a lower average cost.

Table 5.10 - Tariff structure for claim severity, using the training dataset

	Std. Error	**Std. Error (%)**	**Exp(β)**
Intercept	0.00970	0.1	998.62853
Vehicle Brand – G2	0.01285	39.5	0.94864
District – G2	0.02109	19.8	0.88073
District – G3	0.02529	27.5	1.07437
v1 Driver Age – OPoly1	0.00443	17.3	0.95521

Source: Authors preparation

5.2.2. Gradient Boosting

The same dataset used to develop the GLM models was also uploaded to R Server, where the frequency and severity Gradient Boosting models were developed. The dataset was also divided in training (80%) and testing (20%).

In order to build the models, the R package gbm was used. Since the original version of this package does not support the Gamma distribution, it was used a different adaptation of the package for this model, retrieved from Harry Southworth's GitHub (Harry Southworth, n.d.) that enables that option.

5.2.2.1. Claim Frequency

The first step when developing a Gradient Boosting model is dividing the dataset into different folds in order to perform cross validation to obtain the optimal parameters. Typically, the full dataset would be used and nested cross-validation would be performed on all data, and so reassuring that the parameter selection and model validation could be performed and correctly assessed (Henckaerts et al., 2020). In this case, and because the goal is to compare the GLM to the GBM models, it was chosen to first divide the full data set in testing and training, as mentioned above, and use this last data set to develop the Gradient Boosting models. To obtain these two data sets, the data was divided into five random groups of approximate same size, mutually exclusive and stratified, to make sure the resulting subsets' response variable, in this case claim frequency, follow a similar distribution, following a very similar approach as explained on Section 3.3.3.1. One of those datasets was chosen as the testing set, and the other four as training set.

The next step would be do define a grid search to find the optimal parameters. Initially, it was experimented a model with the default parameters, number of claims as response variable, alongside with exposure as an offset and the Poisson deviance as loss function. It should be stated that this model did not take into account the overdispersion factor, as it was not an option. It was followed by some small shifts according to the previously obtained results. Out of those preliminary results, it was possible do define a grid search with the following parameter combinations:

- Number of trees ∈ {100, 250, 400, 500, 750, 1000}
- Shrinkage ∈ {0.1, 0.05, 0.01}

- Interaction depth ∈ {1, 2, 3, 4, 5}
- Minimum observations in terminal nodes =0.01$*nrows$
- Bag Fraction ∈ {0.7, 0.8, 0.9, 0.95}

This combination of parameters lead to 360 different models to be tested.

The ideal approach would be to divide the training dataset in six folds and then test the model with the chosen optimal parameters in the testing set. However, given the dimension of the data (the training set included 1 971 461 observations, meaning that each iteration of the inner cross validation scheme would have to iterate trough $5/6$ of the data, 1 642 884 observations), it would be very computationally exhaustive to use the full training set to obtain the optimal parameter selection. This option was tested, and it took over 20 hours to go through around 120 out of the 360 combinations, which is a very long time given that this process would have to be repeated six times (and if it were to be an option, it would take much longer than developing the GLM models, which goes against the aim of this project).

Given this restraint, it was chosen to retrieve a sample of 50 000 observations from the training set to obtain those desired parameters. This sample was apportioned in six data folds, following the methodology described in Figure 3.2.

Table 5.11 - Optimal tunning parameters per fold (and average) and out-of-sample Poisson deviance, using the 50 000 observations sample for the frequency model

Fold	Nr. of Trees	Shrinkage	Interaction Depth	Bag Fraction	OOS Poisson Deviance
1	37	0.1	2	0.95	0.2802844
2	64	0.1	2	0.95	0.2802088
3	642	0.01	2	0.8	0.2802078
4	116	0.1	1	0.95	0.2793700
5	239	0.05	2	0.95	0.2791459
6	47	0.1	4	0.95	0.2796919
Average	190	0.077	2	0.952	

Source: Authors preparation

Table 5.11 above presents the optimal set of parameters for each of the six folds, each one was selected as the combination that lead to the smallest cross validation iteration error in that fold, in this case the Poisson deviance. In Table E.1 of Appendix E, it is possible to consult the top 10 combinations that lead to the smallest cross validation error per fold.

The results are quite heterogenous, with the maximum number of optimal trees being achieved for the smallest value of shrinkage, in fold 3, a known behavior between these parameters.

Model Interpretation

In order to open the black box in which the GBM models lie, there were used interpretability tools as described in Section 3.3.4, on some of the feature variables.

Variable Importance

Variable importance introduced in Section 3.3.4.1 was used to find the most relevant variables in the frequency model. The results are visible in Figure 5.5 below.

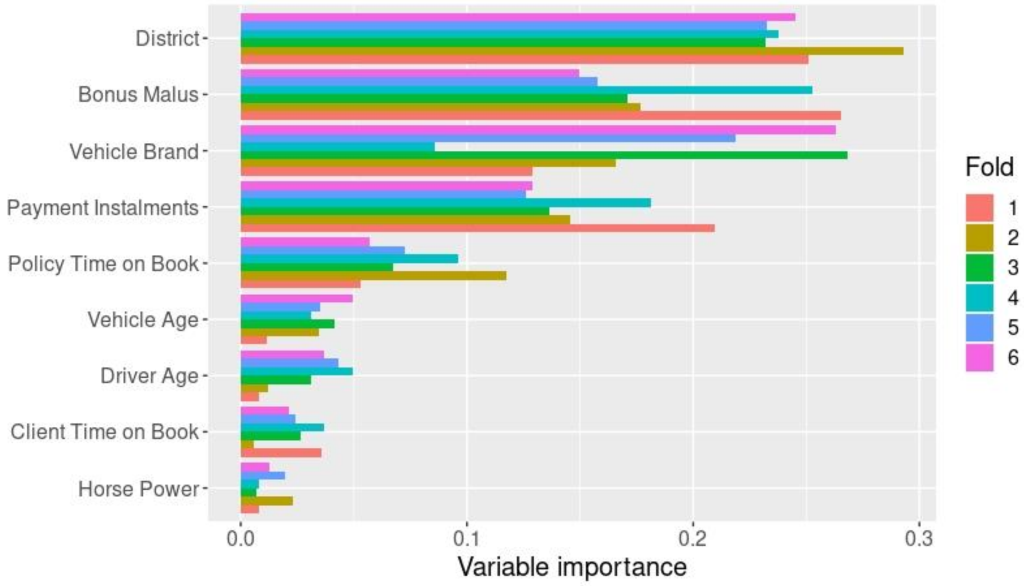


Figure 5.5 - Variable importance in the optimal GBM per data fold for frequency, considering an individual fold importance per variable above 0.1%

Source: Authors preparation

As stated in the figure above, from top to bottom are represented the variables that are the most important for the model, measured by the average variable importance over the folds. *District, Bonus Malus*, *Vehicle Brand* and *Payment Instalments* are the clear top four variables. Out of these nine variables, only *Bonus Malu*s and *Client Time on Book* were not selected in the GLM model, showing that there is a similarity between the choice of variables taken in both models.

Partial Dependence Plot

In Figure 5.6 is depicted the graphical representation of the partial dependence effect of the variable *District* in the frequency model, taken into consideration a sample of 1 000 observations.

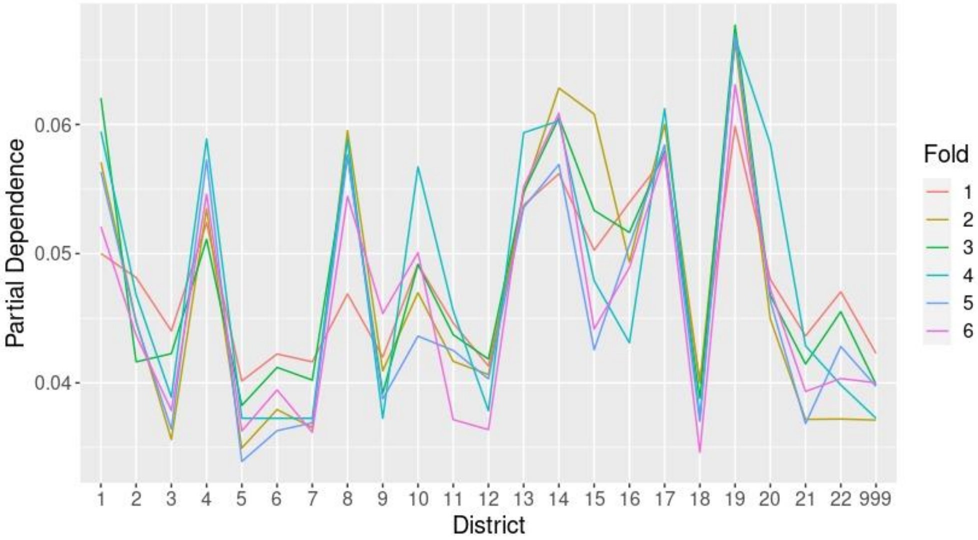


Figure 5.6 - Partial dependence plot representing the effect of the *District* on Frequency, per data fold, using a sample of 1 000 observations

Source: Authors preparation

It is clear in this figure that *Districts* 4, 8, 14, 17 and 19 present a higher risk of filling a claim, quite uniformly between all folds. Comparing to Figure 4.9 where is visible the observed frequency per *District*, the predictions depicted above follow a very similar trend to the observed values.

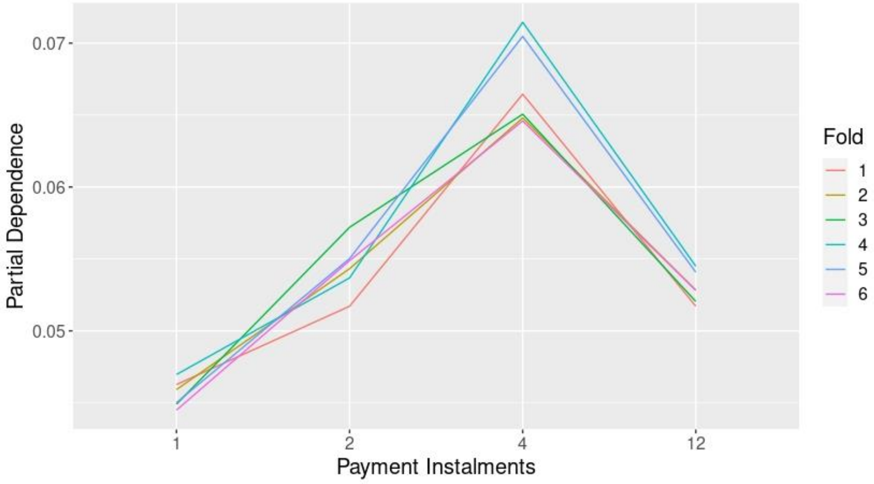


Figure 5.7 - Partial dependence plot representing the effect of the *Payment Instalments* on frequency, per data fold, using a sample of 1000 observations

Source: Authors preparation

Similarly, *Payment Instalments* GBM frequency predictions mimic the observed frequency in Figure 4.5 with policies contracted with quarterly payments posing a higher risk than semi-annually or monthly, that revolves around 0.055, and much higher than annual payments, as can be observed in Figure 5.7 above.

Individual Conditional Expectation Plot

As previously explained in Section 3.3.4.3, the ICE plot enlightens the dependence between the frequency and a certain feature of interest, in the example represented below in Figure 5.8, Vehicle *Brand*. In this case, it was selected a sample of 1 000 random policy registries from fold 5 (where the OOS Poisson deviance was the lowest), and the several individual conditional expectation lines were plotted. Each line shows how the prediction changes when the *Vehicle Brand* changes, keeping all other variables constant. The blue line represents the average of these lines, also known as partial dependence.

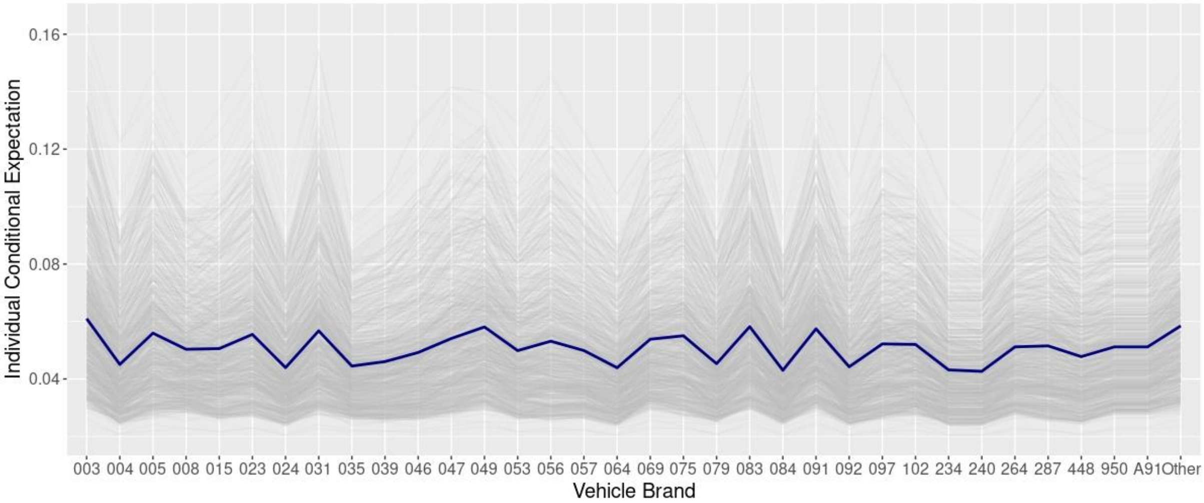


Figure 5.8 - Effect of the *Vehicle Brand* on the frequency model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering data fold 5

Source: Authors preparation

The benefit of using this method is the possibility to capture any heterogeneous relationship created by interactions. By analyzing Figure 5.8 above, the several ICE lines seem to follow the same trend as the average, however some brands such as 15, 46, 47, 75 and 102 register some predictions that do not follow the trend, being visible by the overlapping crossing lines. This could be an indicator of an interaction between *Vehicle Brand* and another variable.

In order to check for interactions between variables, Friedman's H-statistic was calculated. Table 5.12 presents an ordered list of the 10 strongest two-way interactions between the variables.

Table 5.12 - H-Statistic for the 10 strongest two-way interactions between all feature variables in the GBM frequency model, considering data fold 5

Variables	H-Statistic
(Payment Instalments, Vehicle Brand)	0.2255
(District, Policy TOB)	0.2004
(Client TOB, Vehicle Age)	0.1560
(Bonus Malus, Payment Instalments)	0.1424
(Payment Instalments, Policy TOB)	0.1355
(District, Vehicle Brand)	0.1147
(Bonus Malus, District)	0.1038
(District, Payment Instalments)	0.0868
(Bonus Malus, Vehicle Brand)	0.0867
(District, Vehicle Age)	0.0695

Source: Authors preparation

As Figure 5.8 hinted*, Vehicle Brand* does in fact seem to have an interaction with *Payment Instalments*, with a value of 0.2255 for the H-Statistic. This value can range from 0 to 1, with 0 meaning that there is no interaction present and 1 implying that the effect of both variables on frequency prediction is purely driven by the interaction. From the interactions detected above, only the interaction between *District* and *Policy Time on Book* was included in the frequency GLM model, as it was the only one that seemed significant. In theory, GBM is able to handle interactions among input variables and can fit nonlinear relationships without requiring additional input from the user (Zhang, 2015).

The graph represented in Figure 5.9 shows the effects of Vehicle Brand grouped by *Payment Instalments*, for the frequency model.

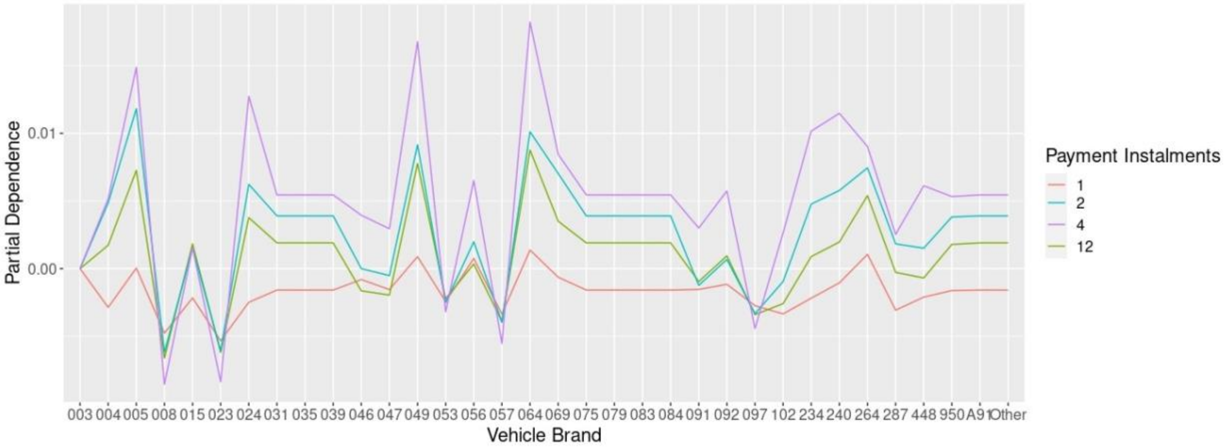


Figure 5.9 - Grouped partial dependence plot for the frequency GBM model, considering data fold 5

Source: Authors preparation

For brands 24, 49, 64, 92 and 448 it is visible that the risk associated with quarterly payments is superior than the other possible instalments. On the other hand, there are brands for which the different payment fractionations do not seem to affect the frequency predictions, such as brands 3, 53, 57 and 97.

5.2.2.2. Claim Severity

The modelling of claim severity using Gradient Boosting was performed in a very similar way to claim frequency. Following the same view, the 18 801 observations were separated into training (80%) and testing (20%). This time, the development of the model was done on the full training dataset, being there no need to collect a smaller sample given that the dimension of this dataset was not an impediment that caused computational deterrents.

Similar to the first approach, there was performed nested cross validation, by dividing the training dataset in six stratified folds, and developing a model with 5-fold cross validation for each, resulting in six combinations of optimal parameters, that would afterwards be averaged to compute the final models.

Initially there was experimented a model with the default hyperparameters, cost of claims as response variable and the Gamma deviance as loss function. Following the results that model outputted, and after some tweaking in the parameters, there was defined a grid search to find the optimal values:

- Number of trees ∈ {100, 150, 200, 250, 300, 400, 500}
- Shrinkage ∈ {0.1, 0.05, 0.01}
- Interaction depth ∈ {1, 2, 3, 4, 5}
- Minimum observations in terminal nodes =0.01$*nrows$
- Bag Fraction ∈ {0.7, 0.8, 0.9, 0.95}

This combination of parameters lead to 420 different models to be tested.

Table 5.13 - Optimal tunning parameters per fold (and average) and out-of-sample Gamma deviance, using the training dataset for the severity model

Fold	Nr. of Trees	Shrinkage	Interaction Depth	Bag Fraction	OOS Gamma Deviance
1	133	0.05	1	0.95	15.76648
2	125	0.05	1	0.95	15.76562
3	56	0.05	2	0.7	15.76658
4	33	0.1	1	0.8	15.76578
5	59	0.05	2	0.7	15.76709
6	75	0.1	1	0.95	15.76697
Average	80	0.067	1	0.84	

Source: Authors preparation

In Table 5.13 are present the optimal parameters for each one of the six folds in which the grid search iterated, chosen accordingly to the smallest cross validation iteration error in that fold, in the case of severity, the Gamma deviance. The top 10 combinations per fold for severity are available in Table E.2 in Appendix E.

The severity model requires less trees to achieve peak performance and the interaction depth is smaller, 1 in the majority of cases, which means that the models use trees with only one split as weak learners. This indicates that the models are completely additive, without interactions.

Model Interpretation

Once again, there were applied the same tools as before in order to enable some interpretation of the severity model, with the exception of interaction study, given that the severity model has an interaction depth of one, and so lacks interactions.

Variable Importance

In Figure 5.10 is available the graphical representation of variable importance per fold. The results follow a quite similar trend among folds, with *District* being the variable that adds the most value to the severity model, followed by *Vehicle Brand, Driver Age*, *License Years* and *Vehicle Weight*.

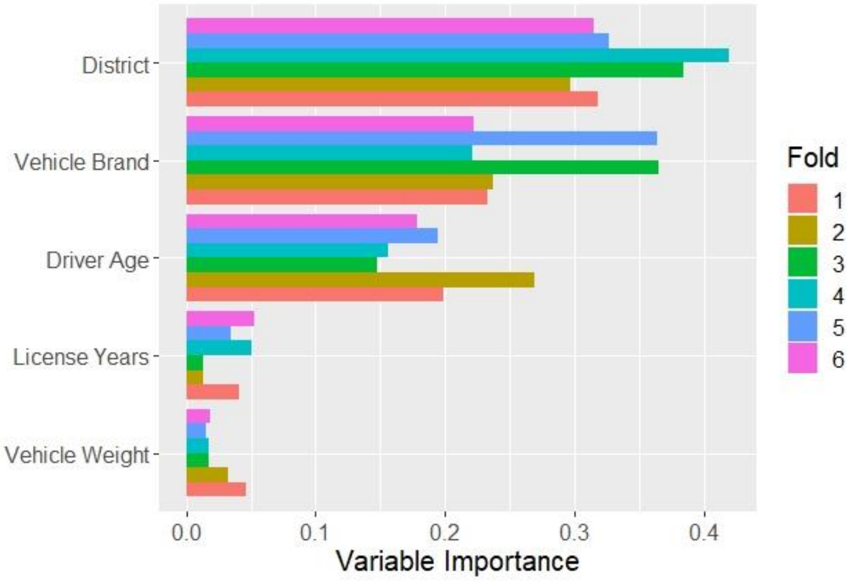


Figure 5.10 - Variable importance in the optimal GBM per data fold for severity, considering an individual fold importance per variable above 0.1%

Source: Authors preparation

There could be some concerns in this rank of variables due to the correlation between *License Years* and *Vehicle Weight* (around 0.3). However, the GBM algorithm is known for handling multicollinearity issues, by choosing between certain regressors to maximize prediction accuracy, and are therefore robust to multicollinearity problems (Sandri & Zuccolotto, 2008).

In comparison to the GLM model, the top three variables were present in both versions of the severity model, highlighting the similarity between the choice of variables in the two models.

Partial Dependence Plot

Using a sample of 1 000 observations, Figure 5.11 depicts the partial dependence effect of *Driver Age* in the severity model. When comparing the predicted average cost below with the observed average cost in Figure 4.4, there are similarities. In both representations, there is a steeper decrease down to around 35-40 years of age, followed by a mild decrease until around 60 years. The steepest decrease occurs around the mark of 70 years, with the difference between the predictions and observed values lying in the ages above 70, where the observed average cost increases but the predictions remain low.

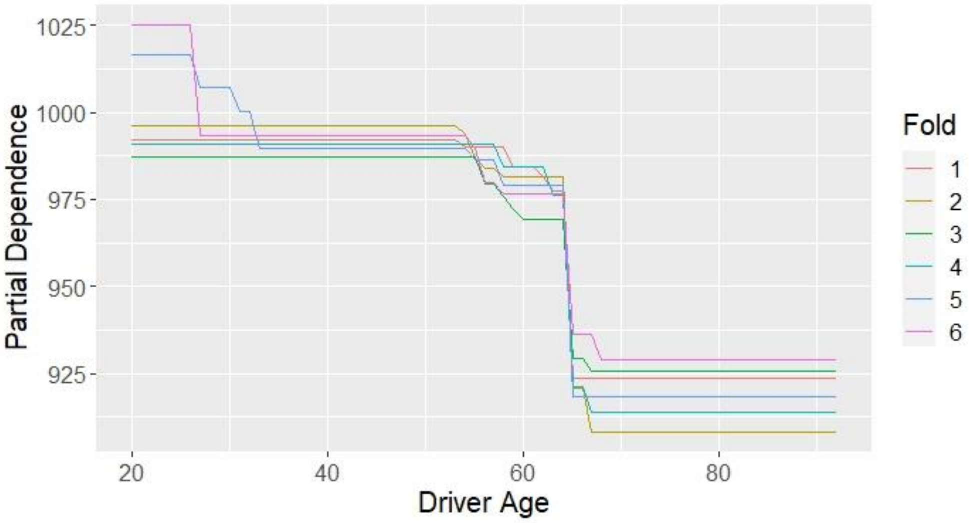


Figure 5.11 - Partial dependence plot representing the effect of the *Driver Age* on severity, per data fold, using a sample of 1000 observations on the training dataset

Source: Authors preparation

Individual Conditional Expectation Plot

Having in consideration a random sample of 1 000 observations and using data fold 2 from the training set (the one with the lowest OOS Gamma deviance), it was plotted the ICE, representing the dependence between claim severity and *District*, while keeping other variables constant. In Figure 5.12, the grey lines represent the ICE and the blue line their average, the partial dependence.

As highlighted above, the several ICE curves are concentrated around the average, especially for districts 4, 5, 7, 8, 11, 15 and 16, meaning that keeping all other risk factors constant, the severity is less sensitive to changes in these districts.

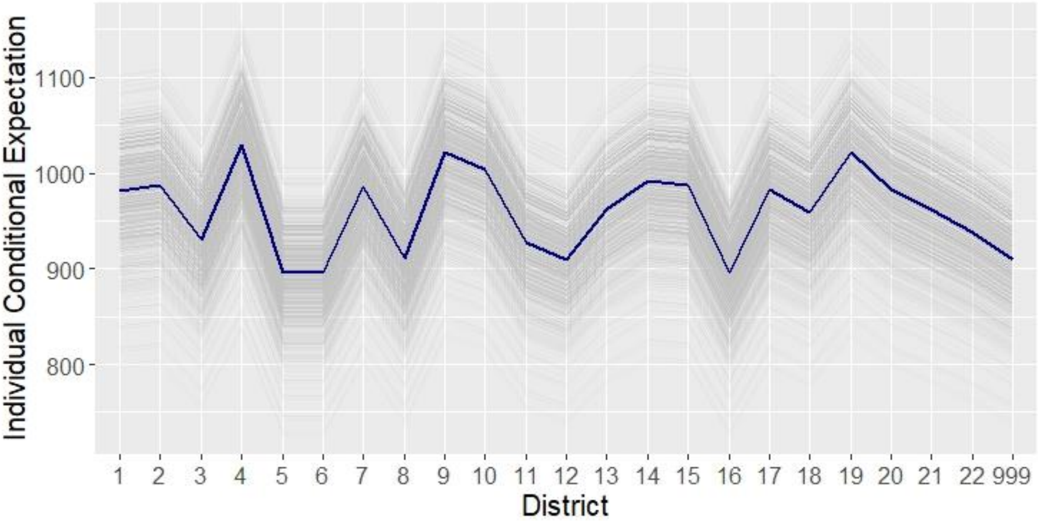


Figure 5.12 - Effect of the *District* on the severity model as Partial Dependence (in dark blue) and Individual Conditional Expectation (in grey), considering data fold 2

Source: Authors preparation

These are all districts that are not considered big cities and hold less exposure to risk. Overall, the predicted average cost associated with each *District* matches the observed values in Figure 4.9.

5.3. Comparison of the Models

In the end of the modelling process, there were obtained four different models, two for each of the response variables. In order to find an answer for the main question in hands, *can machine learning models outperform the classical GLM*, the comparison between the models is presented in this Section.

5.3.1. Choice of Variables

One important indicator that could give some insights on whether or not the Gradient Boosting model has the capability of assessing which feature variables are best to distinguish risk among policies is the choice made of which of those variables to include in the model. Given the fact that this choice in GLM is influenced not only by statistical reasons but also business knowledge, it can be very interesting to compare the choices between both model approaches.

Table 5.14 below summarizes the variables that were selected as significant in the frequency and severity GLMs and the variables that held at least 0.1% of variable importance per fold, for the frequency and severity GBMs.

Table 5.14 - Variables included in the frequency and severity models, according to both approaches: GLM and GBM (only those with over 0.1% variable importance)

Frequency		**Severity**	
GLM	**GBM**	**GLM**	**GBM**
District	*District*	*District*	*District*
	Bonus Malus		
Vehicle Brand	*Vehicle Brand*	*Vehicle Brand*	*Vehicle Brand*
Payment Instalments	*Payment Instalments*		
Policy Time on Book	*Policy Time on Book*		
Vehicle Age	*Vehicle Age*		
Driver Age	*Driver Age*	*Driver Age*	*Driver Age*
	Client Time on Book		
Horse Power	*Horse Power*		
Fuel			
Years of Driving			
			Licence Years
			Vehicle Weight

Source: Authors preparation

As the table shows, the majority of the variables were selected in both versions of each model.

This could indicate that the Gradient Boosting approach has a similar capability of selecting the feature variables that best differentiate risk as the typical approach, but it does not clearly exhibit which of the approaches is best at modelling TPL motor risk.

Regarding the frequency models, it was not surprising that the *District*, *Vehicle Brand* and *Driver Age* were included in both models, as these variables are known as being good at differentiating the risk of claim occurrence. Some other vehicle-related variables were also included in both models, such as the *Age* and *Horse Power*. The GBM model gave a great importance to the *Bonus Malus* variable, but it was not chosen to be a part of the GLM given that the vast majority of the exposure was concentrated in the maximum level (attributed to clients without claims), and the model outputted a fairly good prediction without its addition.

On the other hand, the *Payment Installments* was considered significant for both approaches, which was surprising, as typically it translates the financial possibilities of the client and is not known to lead to differences in the risk of claim occurrence. However, there were studied several interactions whilst the development of the GLM model, and there was not significant proof that these results could be driven from other variables. This has been a known behavior for the company throughout the years, not only for this specific cover and line of business, but generally. It could be an interesting behavior to analyze in further studies.

For the severity models, both approaches included the *District, Vehicle Brand* and *Driver Age*, all of which known as good risk differentiating variables for claim severity. The GBM has also given importance to *License Years* and *Vehicle Weight*, in a much smaller portion comparing to the previous two.

5.3.2. Total Deviance

One way to evaluate the accuracy of the models, is to calculate the total deviance. Table 5.15 presents the total Poisson and Gamma deviance, for the frequency and severity models, respectively, following each of the approaches in study.

Table 5.15 - Total Poisson and Gamma deviance for the frequency and severity models, respectively, considering the two sub-samples and all data

Sample	**Total Poisson Deviance**		**Total Gamma Deviance**	
	GLM	**GBM**	**GLM**	**GBM**
Training (80%)	432 449	428 621	9 406	10 545
Testing (20%)	107 896	106 773	2 335	2 624
All (100%)	540 362	535 685	11 742	13 209

Source: Authors preparation

For the frequency modelling, the Gradient Boosting has the lowest total deviance for all samples, indicating that this approach leads to a better fit. On the other hand, for the severity model it is the Generalized Linear Model approach that outputs the smallest values of total deviance, considering all samples.

The fact that the frequency GBM has the lowest total deviance goes in accordance with the assumption taken in this project that machine learning models have the ability to improve prediction accuracy.

The contrasting (but approximate) severity results could be due to the severity sample size, that had around 18 000 observations. It is known that machine learning models are a good choice for big data, but in small sets of samples, their performance can be reduced.

Other reason could be the volatility associated with severity modelling, that is known for being harder to model than frequency.

5.3.3. Residuals

Defined above in Section 3.2.4.4 as the distance between the observed and the fitted values, the analysis of the residuals can be a great tool to better understand how the data points deviate from the model and could also give important intel on whether or not the chosen distribution is in fact a good match for the data in question.

5.3.3.1. Frequency Residuals

The deviance residuals were calculated using the square root of each observation contribution to the deviance, multiplied by the signal of the difference between the observations and the predictions.

Figure 5.13 top row depicts the residuals obtained for the frequency modelling using the GLM approach. The layers of residuals describe policies with the same number of claims, being the majority concentrated bellow the horizontal axis, near 0 (as most policies did not register claims). Both training and testing datasets show a similar pattern of residuals, ranging from -0.5 to 6.

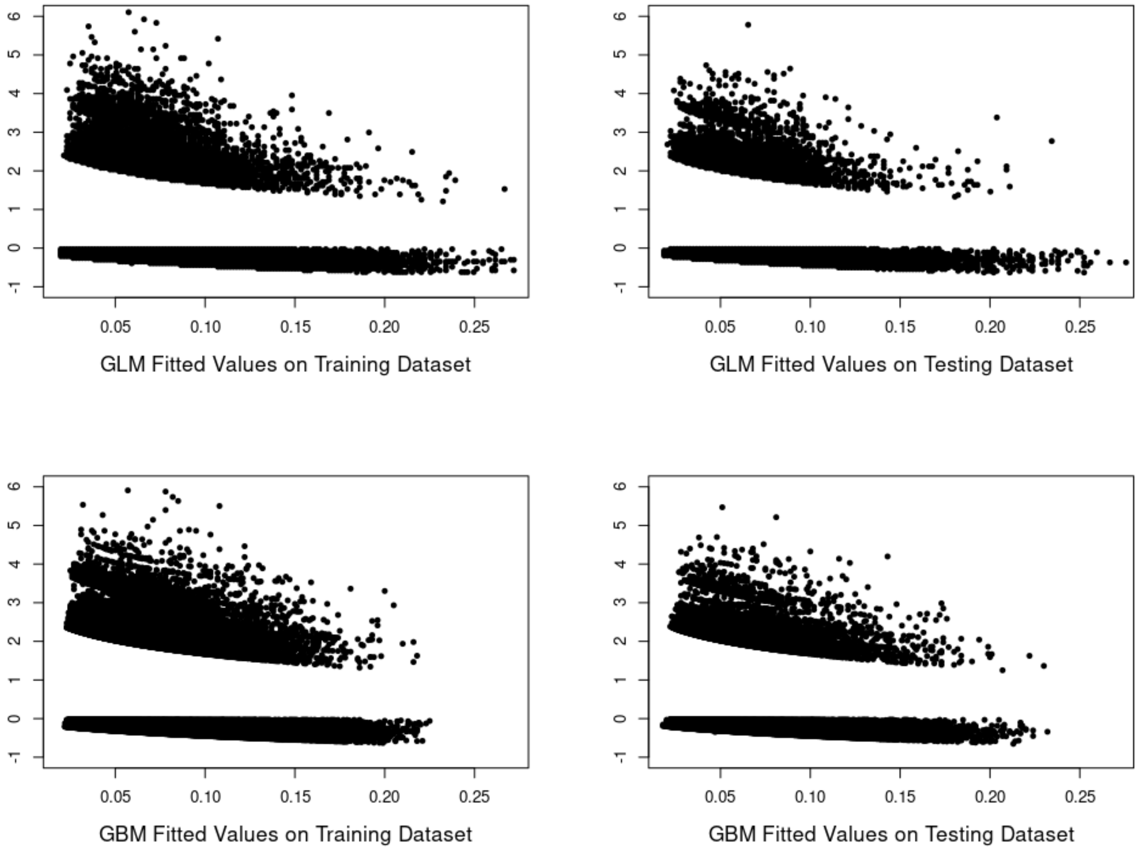


Figure 5.13 - Deviance residuals for the fitted values of the **GLM** (top) and **GBM** (bottom) frequency models, considering the training dataset (left) and testing dataset (right)

Source: Authors preparation

On the other hand, the bottom row represents the deviance residuals for the Gradient Boosting frequency model.

In this case, the residuals follow a very similar pattern as the ones represented above for the GLM. The exception in the GBM model is that its maximum fitted value is around 0.23, whereas the maximum fitted value according to GLM is set higher, around 0.26.

5.3.3.2. Severity Residuals

The severity deviance residuals were calculated and plotted against the fitted values, for each of the modelling approaches.

Collected from the EMBLEM software, Figure 5.14 top row translates the GLM deviance residuals plotted against the fitted values, considering both the training set and the testing set.

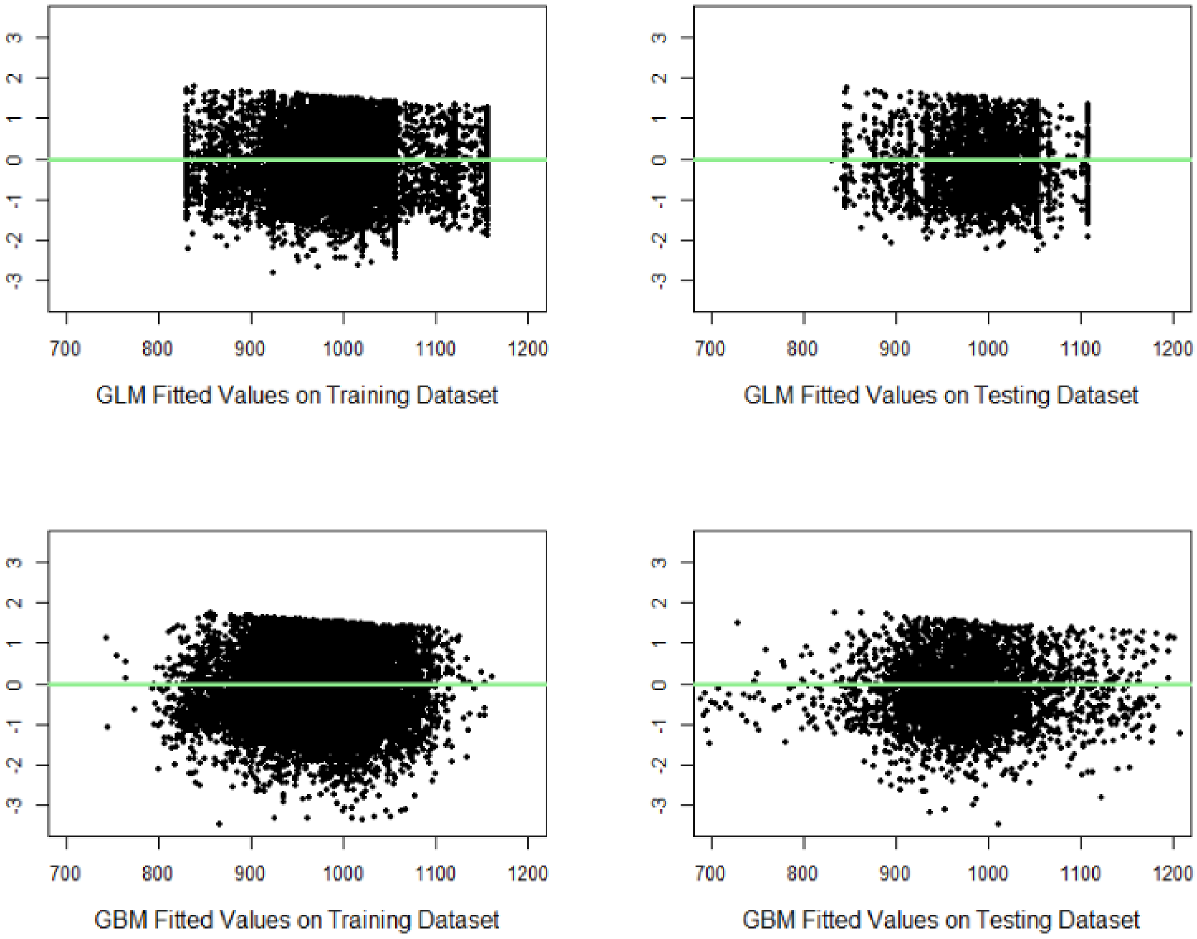


Figure 5.14 - Deviance Residuals for the fitted values of the **GLM** (top) and **GBM** (bottom) severity model, considering the training dataset (left) and testing dataset (right)

Source: Authors preparation

It is clear that the residuals are centered around 0 and concentrated between -2 and 2, for both samples. The variance is constant and does not depict any trend.

In the bottom row it is possible to observe that for both samples the residuals are randomly centered around 0 and concentrated mainly around -2 and 2, with some outliers around 3. In the training set the variance appears quite constant, however in the testing set there is a small increasing trend in the beginning and some dispersion. This could be an indication that the severity GBM is outperformed by the GLM, as the top right plot in Figure 5.14 did not present the same behavior.

According to all this, there seems to be a general good fitting of the models, regardless of the approach in consideration.

6. CONCLUSIONS

The insurance field has registered a significant growth over the last few years in Portugal, and with that comes the rise of competitiveness. It is in every insurance company's best interest to make sure that their prices correctly reflect the risks they are underwriting, which is partially done through the development and implementation of fair pricing models. It only makes sense that as new forms of modelling claim frequency and severity emerge, and when the scientific community has delivered several proofs that these new approaches could in fact provide equally or better results, that these companies begin to test the added value that could come with new forms of modelling, in this case, supervised machine learning.

In this work project it was developed the claim frequency and severity models, through the application of two different approaches: the classical approach followed by the insurance company, the Generalized Linear models, using the EMBLEM software and a machine learning approach, the Gradient Boosting models, using the R Studio software, namely the *gbm* package.

Firstly, the data was recovered, grouped together and outliers were removed. There was performed a descriptive and exploratory analysis of the data that allowed for a better understanding on how the behavior of the response variables changed according to each feature variable level. The distributions of *Claim Count* and *Claim Amount* were studied in order to find which theoretical distribution best fitted the data, and after some fittings it was concluded that the Poisson (allowing for over-dispersion) and Gamma were the best fit to model claim frequency and claim severity, respectively.

Both modelling approaches were followed, and it was conducted a preliminary analysis that allowed for some interpretation of the model's estimations. Regarding GLM, the coefficients values allowed for a perception of which factor's levels were related to a higher risk profile. For the GBM, the several interpretability tools presented granted the possibility of having a glimpse behind the machine learning decisions and thus unveiled similarities between both approaches.

To assess if the new approach has the capability of improving the accuracy of the models, the total deviance was compared, having in consideration the training and testing sets, and the whole samples. Taking the results into consideration, the Gradient Boosting only outperforms the classical approach for the modelling of claim frequency, as for severity the GLM remains with the lowest deviance, regardless of the sample taken into account. The residuals were also plotted, and the results indicated that there was an overall good fitting of the models, regardless of the approach. Although, it should be noted that allowing for overdispersion in the frequency modelling leads to a better fit. In terms of processing, GLM can be more time consuming to develop, given that even though the GBM takes a large time to run, it can be done independently dispensing any human input besides the initial treatment of the data, that is common for both approaches. Other than that, both models include a very similar selection of significant variables, meaning that Gradient Boosting results could be used to aid the selection of variables to consider in the GLM, by setting a starting point that most likely includes the most significant variables.

It can now be concluded that this project fulfilled its goal, as it was possible to develop the claim frequency and severity according to the Generalized Linear model and the Gradient Boosting model approaches and draw conclusions from the findings. From now on, the insurance company has a steadier ground to start implementing machine learning models in their pricing practices.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

There were encountered some constraints as this project was being developed, most of which were manageable.

First, the treatment of the data that took a very long time due to its magnitude and complexity. The code had to be translated from another programming language, and several adaptations were made.

Other issue stood with the size of the sample used to model claim frequency, that had over 2 million records, and lead to the usage of a sample to achieve the optimal hyperparameters for the Gradient Boosting model. A suggestion for future works could be to retrieve this sample from the beginning and develop the GLM based upon that same sample, so that the results could be better compared. Another suggestion could be the development of the frequency GBM using the full dataset, even if it took days to run, to confirm if the gain in performance compared to the model based on the sample justified the time it takes to complete.

Lastly, there were some restraints related to the *gbm* R Server package*,* which did not support the possibility to use the Gamma loss function. It was overcame by using a different adaptation of the package. Another flaw in this package was the inability to adapt to a quasi-poisson loss function, as it was not supported.

A major limitation associated with the GBM lies in the implementation of the findings in order to calculate the premium to charge the client. As the GLM outputs several coefficients that can be directly used to perform the calculations in any typically used insurance software, the GBM does not give a direct output. In order to obtain a premium, there would be the need to execute both the severity and frequency models on R, giving each client's specific information as an input. This is a big inconvenience, as there would be many implementation constraints in comparison with the current methodology.

For future works, there is the recommendation of complementing the analysis performed using different packages that develop Gradient Boosting, such as *xgboost*, *h2o* or *lightGBM* (available in *R Studio*) or even other machine learning methods, such as Neural Networks, for example.

BIBLIOGRAPHICAL REFERENCES

Circular n.º 2/2022, de 15 de março, (2022) (testimony of Autoridade de Supervisão de Seguros e Pensões).

Boehmke, B., & Greenwel, B. (2020). *Hands-On Machine Learning with R*. CRC Press.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Taylor & Francis.

Chen, J., & Shao, J. (1993). Iterative Weighted Least Squares Estimators. *The Annals of Statistics*, *21*(2), 1071–1092. https://doi.org/10.1214/aos/1176349165

Choi, H. I. (2017). *Lectures on Machine Learning. Lecture 4: Exponential family of distributions and generalized linear model (GLM)*. *Draft: version 0.9.2*, 1–20.

de Jong, P., & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

Eling, M., Nuessle, D., & Staubli, J. (2021). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *Geneva Papers on Risk and Insurance: Issues and Practice*. https://doi.org/10.1057/s41288-020-00201-7

Fauzan, M. A., & Murfi, H. (2018). The Accuracy of XGBoost for Insurance Claim Prediction. *International Journal of Advances in Soft Computing and Its Applications*, *10*(2), 159–171.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing StatisticaPeeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectationl Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Grize, Y. L., Fischer, W., & Lützelschwab, C. (2020). Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry*. https://doi.org/10.1002/asmb.2543

Gudmundarson, R. L., Guerra, M., & de Moura, A. B. (2021). *Minimizing Ruin Probability Under Dependencies for Insurance Pricing*. http://arxiv.org/abs/2108.10075

Guerreiro, G. (2016). *Manual de Construção de Tarifas com R*. Textos de Apoio, FCT NOVA.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, *9*(2), 1–23. https://doi.org/10.3390/risks9020042

Harry Southworth. (n.d.). *Harry Southworth gbm*. GitHub. Retrieved July 15, 2022, from https://github.com/harrysouthworth/gbm

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning* (2nd ed.). https://hastie.su.domains/Papers/ESLII.pdf

Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2020). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal*, 1–31. https://doi.org/10.1080/10920277.2020.1745656

Hetherington, A. (2020, May 30). *New Actuaries Must Know About Machine Learning | by Andrew Hetherington | Towards Data Science*. https://towardsdatascience.com/new-actuaries-must-know-about-machine-learning-846dd65647d9

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Lin, M., Lucas, H. Jr. C., & Shmueli, C. (2013). Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, *24*(4), 906–917.

Malhotra, R., & Sharma, S. (2018). *Machine Learning in Insurance*.

Molnar, C. (2021). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. In *Source: Journal of the Royal Statistical Society. Series A (General)* (Vol. 135, Issue 3).

Noll, A., Salzmann, R., & Wüthrich, M. v. (2018). Case Study: French Motor Third-Party Liability Claims. *SSRN Eletronic Journal*, 1–39. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3164764

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGboost versus logistic regression. *Risks*, *7*(2). https://doi.org/10.3390/risks7020070

Quan, Z., & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), 377–407. https://doi.org/10.1515/demo-2018-0022

Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611–628.

Smith, K. A., Wilis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: a case study. *Journal of the Operational Research Society*, *51*(5), 532–541.

Staudt, Y., & Wagner, J. (2021). Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks*, *9*(3). https://doi.org/10.3390/risks9030053

Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PLoS ONE*, *15*(8 August 2020). https://doi.org/10.1371/journal.pone.0238000

Tober, S. (2020). *Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling*. 1–29.

Willis Tower Watson. (2020). *Advanced analytics: Are insurers living the dream?* https://www.willistowerswatson.com/en-US/Insights/2020/01/advanced-analytics-are-insurers-living-the-dream-2019-2020-P-C-insurance-advanced-analytics

Zhang, Y. (2015). A Gradient Boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*(Part B), 308–324.

Zhifeng, X. (2020). *Best Practice of Risk Modelling in Motor Insurance-Using GLM and Machine Learning Approach*. 1–37. http://hdl.handle.net/10400.5/20405

APPENDIX A

RELEVANT LITERATURE

Table A.1 - Summary of relevant literature related to the topic in question

Title	Authors	Variable in study	Models Applied	Main results/findings
Predictive analytics of insurance claims using multivariate decision trees	(Quan & Valdez, 2018)	Claim occurrence, using a dataset containing six coverages for buildings, vehicles and equipment of the US government.	Tree-based models with multivariate response variables, namely CART, Random Forest and XGBoost.	There was an improvement in prediction accuracy when applying the models based on multivariate trees when comparing to univariate trees.
The Accuracy of XGBoost for Insurance Claim Prediction	(Fauzan & Murfi, 2018)	Claim Frequency of a large dataset with many missing values	XGBoost, AdaBoost, Stochastic GB, Random Forest and Neural Networks.	XGBoost has a better accuracy in terms of normalized Gini than the other models.
Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression	(Pesantez-Narvaez et al., 2019)	Claim occurrence, using a dataset containing telematics data.	Logistic regression and XGBoost.	XGBoost requires more attention to match the predictive performance of the logistic regression and is harder to interpretate, only increasing the predictive performance slightly.
Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods	(Henckaerts et al., 2020)	Frequency and severity, based on MTPL portfolio from a Belgian insurer from 1997.	Machine learning with decision threes (simple regression trees, random forests and boosted trees).	Boosted trees outperform the classical GLM model.
Machine learning applications in nonlife insurance	(Grize et al., 2020)	Online motor insurance (random selection of 10	Commercial product containing over 50 standard ML	The best performing model was a Light Gradient Boosted Trees Regressor with Early

		000 customer profiles collected from an insurance company platform). The response variable is the premium offered for a given customer profile.	algorithms, such as Gradient Boosting models, deep-learning models, random forests, etc.	Stopping (Gamma Loss). Without ML, dynamic pricing would not have been possible. The fast development, monitoring and updating of the models could only be achieved by using ML, which provides high model prediction quality and speed of implementation.
Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling	(Tober, 2020)	Claim frequency for an all-risk insurance tariff, based on data collected from a Swedish insurance company.	Three tree-based ML models, namely simple decision trees, random forests and Gradient Boosting machines.	The gradient boost and random forest trees outperform the individual decision trees.
Best Practice of Risk Modelling in Motor Insurance - Using GLM and Machine Learning Approach	(Zhifeng, 2020)	Risk premium, using five main perils provided by Liberty Insurance.	Classical GLM improved with ML, namely root in penalized GLM and XGBoost.	The top ranked variables from penalized GLM play an important role. The XGB detected interactions and added valuable information to the model.
Stochastic Gradient Boosting frequency-severity model of insurance claims	(Su & Bai, 2020)	Claim frequency and severity, using a TPL dataset.	Combination of a stochastic gradient boost algorithm and a profile likelihood approach: GLM, GAM and D-FSBoost.	The dependent models have a better performance, being superior to other state-of-the-art models when it comes to predicting claim frequency and severity.
Case Study: French Motor Third-Party Liability Claims	(Noll et al., 2018)	Claim frequency on a French TPL dataset.	GLM as a benchmark, regression trees, boosting machine and	Boosting machine and neural networks produce very similar results, better than the GLM however the

			neural networks.	authors state that the GLM could have been better studied).
Assessing the Performance of Random Forests for Modeling Claim Severity in Collision Car Insurance	(Staudt & Wagner, 2021)	Claim severity of a collision dataset, from a Swiss insurance company.	GLM and GAM as benchmark and two random forest models (one for claim severity and other for the log-transformed claim severity).	The use of the log-transformation doesn't lead to any improvements in the random forest model. Nevertheless, this model is the best to explain right-skewed claims. Globally, GAM has a better performance.

APPENDIX B

VARIABLES IN STUDY

Table B.1 - Summary of the initial proposed feature variables to be used to develop the frequency and severity models

Variable Original Name	Levels	Description
Cust_TipoProf	*Self-employed, employed, self-employed and employed, unknown*	Type of profession of the client.
Cust_UEN	*RIF, ZRT*	Type of client (RIF – particular, ZRT – direct channel).
Cust_HabAcademicas	*doesn't know how to read or write, 4th grade, 9th grade, 12th grade, technical course, bachelor, degree, master, doctorate, unknown*	Academic qualifications of the client.
Cust_IndFilhos	*Yes, No, unknown*	Flag variable, *Yes* in case the client has children, *No* otherwise.
Cust_ConcelhoT	308 different levels, from *C_CT1* to *C_CT308*, *unknown*	County of living of the policy undertaker.
Cust_DistritoT	22 different levels, from *C_DT1* to *C_DT22*, *unknown*	District of living of the policy undertaker.
Cust_ZonaCirc	*ZonaA, ZonaB, ZonaC, ZonaD, ZonaE, unknown*	Usual circulation area of the undertaker.
Cust_EstadoCivil	*Married, Single, Widower, Divorced, unknown*	Marital status of the policy undertaker.
Cust_IdadeT	*0-17, 18* to *85* (individually), *85+, unknown*	Age of the policy undertaker.
Cust_AntigClient	*1* to *21* (individually), *21+*, *999*	Seniority of the policy undertaker as a client in the company.
Pol_Frac	*1 x year, 2 x year, 4 x year, 12 x year*	Payment instalments of the policy.
Pol_Delegacao	22 different levels, from *P_D1* to *P_D22*	Distribution channel associated with the agent responsible for the policy.

Pol_TipoAgente	17 different levels, from *P_TA1* to *P_TA17*	Distinguishes the agents between different levels of achievements and volume.
Pol_DebitoDireto	*Non-DB*, *DB*	If the policy payments come from direct-charge or not.
Pol_AntigApolice	*1* to *21* (individually), *21+*	Time on book of the policy.
Obj_Lob	148 different levels, from *O_L1* to *O_L148*	Line of business.
Obj_ConcelhoC	309 different levels, from *O_CC1* to 0_*CC309*.	County of living of the usual driver.
Obj_ProfissaoC	364 distinct levels from *O_PF1* to *O_PF364*, *unknown*	Profession of the usual driver.
Obj_Marca	708 different levels from *O_M1* to *O_M708*, *unknown*	Brand of the vehicle.
Obj_Lotacao	*2, 3, 4, 5, 6, 7, 8, 9, 11+, 999*	Capacity of the vehicle.
Obj_Cilindrada	*1-50, 50-125, 125-250, 250-400, 400-500, 500-600, 600-700, 700-800, 800-900, 900-1000, 1000-1100, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 2000-2100, 2100-2200, 2200-2300, 2300-2400, 2400-2500, 2500-3000, 3000-3500, 3500-4000, 4000-4500, 4500-5000, 5000+, 999*	Engine capacity of the vehicle.
Obj_HP	*0-50, 50-100, 100-150, 150-200, 200+*	Vehicle power measured in horsepower.
Obj_KW	*0-50, 50-100, 100-150, 150-200, 200+*	Vehicle power measured in killowatts.
Obj_SistemaSegTravagem	*ABS, Disco, BWW, unknown*	Brake safety system of the vehicle.
Obj_PesoBruto	*<50, 50-500, 500-600, 600-700, 700-800, 800-900, 900-1000, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 200-2100, 2100-2200, 2200-2300, 2300-*	Gross weight of the vehicle.

	2400, 2400-2500, 2500-2600, 2600-2700, 2700-2800, 2800-2900, 2900-3000, 300-3100, 3100-3200, 3200-3300, 3300-3400, 3400-3500, 3500+, 999	
Obj_DissFurto	18 different levels, from *O_DF1* to *O_DF18*, *unknown*, *withoutDF*	Theft deterrent present in the vehicle.
Obj_ValorNovo	*<7000, 7000-10000, 10000-15000, 15000-20000, 20000-25000, 25000-30000, 30000-35000, 35000-50000, 50000-100000, 100000-500000, 500000+*	Initial value of the vehicle, as if it was new.
Obj_Combustivel	8 distinct levels, from *O_F1* to *O_F2*, *without fuel, other, unknown*	Fuel of the vehicle
Obj_DistritoC	22 different levels, from *O_DC1* to *O_DC22*, *unknown*	District of living of the usual driver.
Obj_EscalaoBM	*-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14*	Bonus Malus value.
Obj_AnosCarta	*1* to *21* (individually), *21+, 999*	Seniority of driver's license.
Obj_IdadeVeic	*1* to *30* (individually), *30+, 999*	Age of the vehicle.
Obj_IdadeC	*0-17, 18* to *85* (individually), *85+, unknown*	Age of the usual driver.
Cov_CapCob	*CapMin, CapMax*	Flag variable, *CapMax* if the policy has the optional 59M TPL capital, or *CapMin* ontherwise.
NBexe	*RN, NB, FNB*	Flag variable, *RN* in case of renovation, *NB* in case of New Business and *FNB* in case of fake new business.
OwnDamage	*Yes, No*	Flag variable, *Yes* in case the policy has the own damage coverage, *No* otherwise.

Table B.2 - Final feature variables used to develop the frequency and severity models

Variable Original Name	Variable Simplified Name
Cust_UEN	UEN
Cust_AntigClient	Client Time on Book
Pol_Frac	Payment Instalments
Pol_Delegacao	Agent Delegation
Pol_DebitoDireto	Direct Debit Payment
Pol_AntigApolice	Policy Time on Book
Obj_Marca	Vehicle Brand
Obj_Lotacao	Vehicle Seats
Obj_Cilindrada	Engine Capacity
Obj_HP	Horse Power
Obj_PesoBruto	Vehicle Weight
Obj_ValorNovo	Vehicle Value as New
Obj_Combustivel	Fuel
Obj_DistritoC	District
Obj_EscalaoBM	Bonus Malus
Obj_AnosCarta	Years of Driving
Obj_IdadeVeic	Vehicle Age
Obj_IdadeC	Driver Age
Cov_CapCob	Cover Capital
NBexe	New Business
OwnDamage	Own Damage Cover

APPENDIX C

CORRELATION MATRIX

Figure C.1 – Correlation matrix between the 36 initial variables

Factor	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.	33.	34.	35.	36.
1. Cust_TipoProf	1																																			
2. Cust_UEN	0,02	1																																		
3. Cust_HabAcademica	0,04	0,13	1																																	
4. Cust_IndFilhos	0,06	0,31	0,22	1																																
5. Cust_ConcelhoT	0,07	0,22	0,19	0,2	1																															
6. Cust_ZonaCirc	0,02	0,09	0,09	0,07	1	1																														
7. Cust_EstadoCivil	0,01	0,26	0,09	0,23	0,07	0,02	1																													
8. Cust_IdadeT	0,01	0,04	0,11	0,21	0,03	0,02	0,32	1																												
9. Cust_AntigCliente	0,03	0,06	0,1	0,2	0,08	0,05	0,06	0,09	1																											
10. Pol_Frac	0,02	0,02	0,07	0,07	0,18	0,07	0,07	0,14	0,14	1																										
11. Pol_Delegacao	0,03	0,97	0,13	0,25	0,78	0,72	0,13	0,02	0,05	0,15	1																									
12. Cust_Distrito	0,03	0,16	0,13	0,11	1	0,94	0,05	0,02	0,05	0,14	0,72	1																								
13. Pol_TipoAgente	0,02	0,88	0,06	0,24	0,32	0,12	0,11	0,02	0,05	0,07	0,31	0,15	1																							
14. Pol_DebitoDireto	0,01	-0,02	0,11	0,08	0,23	0,08	0,08	0,18	0,17	0,77	0,19	0,17	0,11	1																						
15. Pol_AntigApolice	0,02	0,11	0,05	0,16	0,06	0,04	0,05	0,08	0,38	0,16	0,05	0,04	0,05	0,18	1																					
16. Obj_Lob	0,03	0,99	0,11	0,29	0,06	0,11	0,26	0,07	0,18	0,17	0,28	0,1	0,28	0,21	0,39	1																				
17. Obj_ConcelhoC	0,07	0,24	0,19	0,2	0,92	0,87	0,07	0,03	0,08	0,19	0,79	0,95	0,32	0,23	0,06	0,06	1																			
18. Obj_Profissao	0,03	0,3	0,26	0,16	0,07	0,15	0,27	0,1	0,09	0,13	0,19	0,14	0,12	0,15	0,1	0,15	0,07	1																		
19. Obj_Marca	0,01	0,04	0,04	0,04	0,06	0,07	0,04	0,03	0,02	0,07	0,05	0,05	0,02	0,08	0,02	0,04	0,06	0,04	1																	
20. Obj_Lotacao	0	0,04	0,03	0,02	0,07	0,05	0,03	0,03	0,02	0,04	0,05	0,05	0,02	0,06	0,02	0,06	0,07	0,05	0,18	1																
21. Obj_Cilindrada	0,01	0,04	0,04	0,04	0,05	0,06	0,04	0,03	0,02	0,07	0,04	0,04	0,02	0,1	0,02	0,04	0,05	0,05	0,28	0,21	1															
22. Obj_HP	0,01	0,02	0,07	0,06	0,08	0,04	0,02	0,06	0,07	0,05	0,05	0,05	0,03	0,11	0,08	0,11	0,08	0,1	0,23	0,1	0,32	1														
23. Obj_KW	0,02	0,07	0,09	0,12	0,11	0,06	0,04	0,08	0,16	0,08	0,08	0,07	0,05	0,14	0,19	0,23	0,11	0,12	0,22	0,08	0,31	0,61	1													
24. Obj_SistemaSegTravagem	0,01	0,01	0,03	0,01	0,06	0,01	0	0,01	0,01	0,01	0,03	0,03	0,02	0,02	0,02	0,03	0,06	0,03	0,02	0,01	0,02	0,05	0,05	1												
25. Obj_PesoBruto	0,01	0,05	0,05	0,1	0,06	0,06	0,05	0,03	0,07	0,11	0,05	0,05	0,03	0,14	0,09	0,09	0,06	0,05	0,12	0,18	0,17	0,29	0,34	0,03	1											
26. Obj_DissFurto	0,01	0,05	0,04	0,03	0,06	0,04	0,02	0,01	0,01	0,03	0,05	0,04	0,02	0,06	0,02	0,07	0,07	0,05	0,03	0,03	0,04	0,17	0,2	0,11	0,04	1										
27. Obj_ValorNovo	0,01	0,07	0,05	0,08	0,07	0,04	0,02	0,02	0,04	0,04	0,05	0,04	0,03	0,06	0,05	0,08	0,07	0,06	0,11	0,05	0,17	0,36	0,32	0,05	0,18	0,11	1									
28. Obj_Combustivel	0,01	0,06	0,02	0,04	0,05	0,02	0,03	0,05	0,07	0,07	0,04	0,03	0,03	0,08	0,18	0,19	0,05	0,07	0,1	0,21	0,28	0,13	0,09	0,02	0,13	0,03	0,07	1								
29. Obj_DistritoC	0,03	0,17	0,13	0,11	0,95	0,81	0,05	0,02	0,05	0,14	0,72	0,95	0,15	0,17	0,04	0,1	1	0,14	0,05	0,05	0,04	0,05	0,07	0,03	0,05	0,04	0,04	0,03	1							
30. Obj_EscalaoBM	0,01	0,02	0,02	0,06	0,03	0,02	0,08	0,11	0,05	0,08	0,02	0,02	0,01	0,1	0,04	0,11	0,03	0,04	0,01	0,01	0,01	0,03	0,07	0,01	0,03	0,01	0,02	0,01	0,02	1						
31. Obj_AnosCarta	0,01	0,01	0,07	0,17	0,03	0,03	0,21	0,29	0,1	0,14	0,03	0,03	0,02	0,17	0,08	0,07	0,03	0,08	0,03	0,02	0,02	0,04	0,09	0,01	0,03	0,01	0,02	0,03	0,03	0,24	1					
32. Obj_IdadeVeic	0,01	0,04	0,07	0,04	0,04	0,04	0,03	0,03	0,05	0,07	0,03	0,03	0,02	0,13	0,08	0,05	0,04	0,05	0,07	0,06	0,1	0,18	0,17	0,04	0,07	0,07	0,1	0,08	0,03	0,02	0,03	1				
33. Obj_IdadeC	0,01	0,02	0,1	0,19	0,02	0,02	0,24	0,88	0,09	0,14	0,02	0,02	0,02	0,18	0,09	0,07	0,02	0,09	0,03	0,03	0,03	0,06	0,09	0,01	0,03	0,01	0,02	0,06	0,02	0,14	0,31	0,03	1			
34. Cov_CapCob	0,01	-0,03	0,1	0,03	0,14	0,06	0,03	0,05	0,07	0,03	0,14	0,09	0,09	0,03	0,1	0,21	0,14	0,17	0,09	0,05	0,09	0,17	0,15	0,05	0,08	0,18	0,17	0,07	0,09	0,03	0,06	0,19	0,04	1		
35. NBexe	0,03	0,04	0,07	0,12	0,09	0,04	0,04	0,14	0,54	0,12	0,07	0,07	0,08	0,1	0,71	0,38	0,09	0,15	0,04	0,03	0,04	0,08	0,21	0,02	0,18	0,03	0,09	0,05	0,07	0,11	0,17	0,13	0,15	0,03	1	
36. OwnDamage	0,01	0,03	0,16	0,03	0,17	0,08	0,04	0,08	0,03	0,13	0,13	0,12	0,07	-0,15	0,05	0,14	0,17	0,2	0,18	0,13	0,24	0,45	0,44	0,12	0,26	0,47	0,44	0,12	0,12	0,08	0,04	0,57	0,07	-0,3	0,04	1

APPENDIX D

GENERALIZED LINEAR MODEL INTERACTIONS

Figure D.1 - Predicted values considering *Policy Time on Book* (Pol_AntigApolice) and *District* (Obj_DistritoC_Cod), for the frequency model. Scale is omitted to proceed accordingly to the insurance company data sensitivity policy

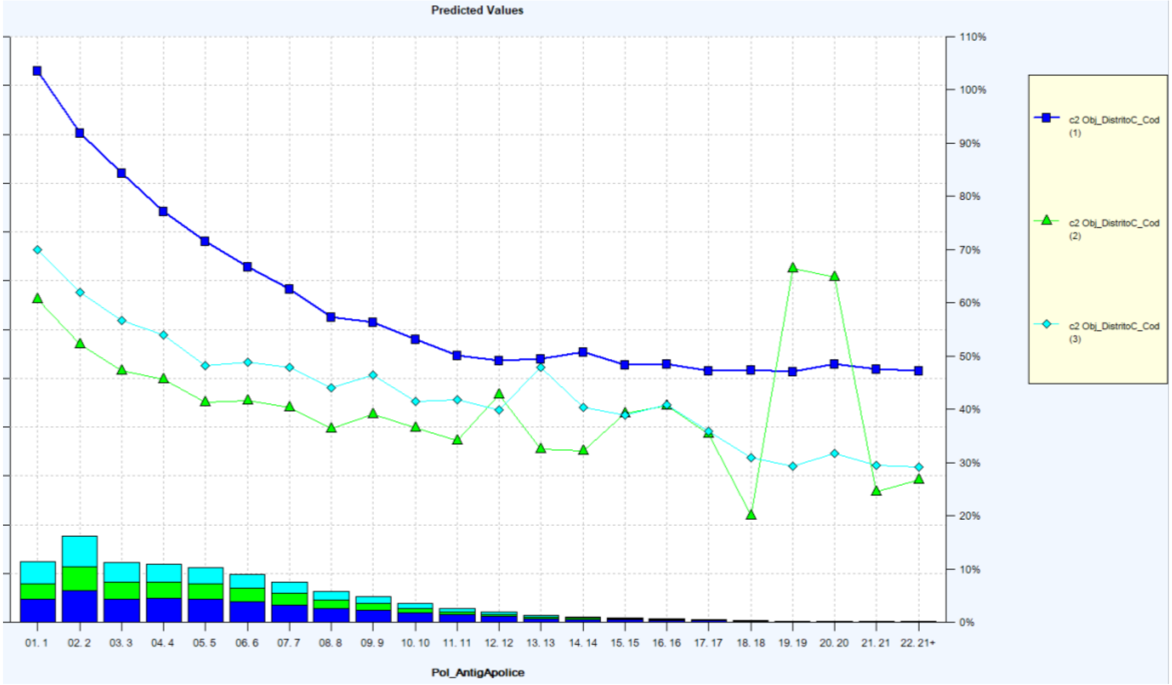
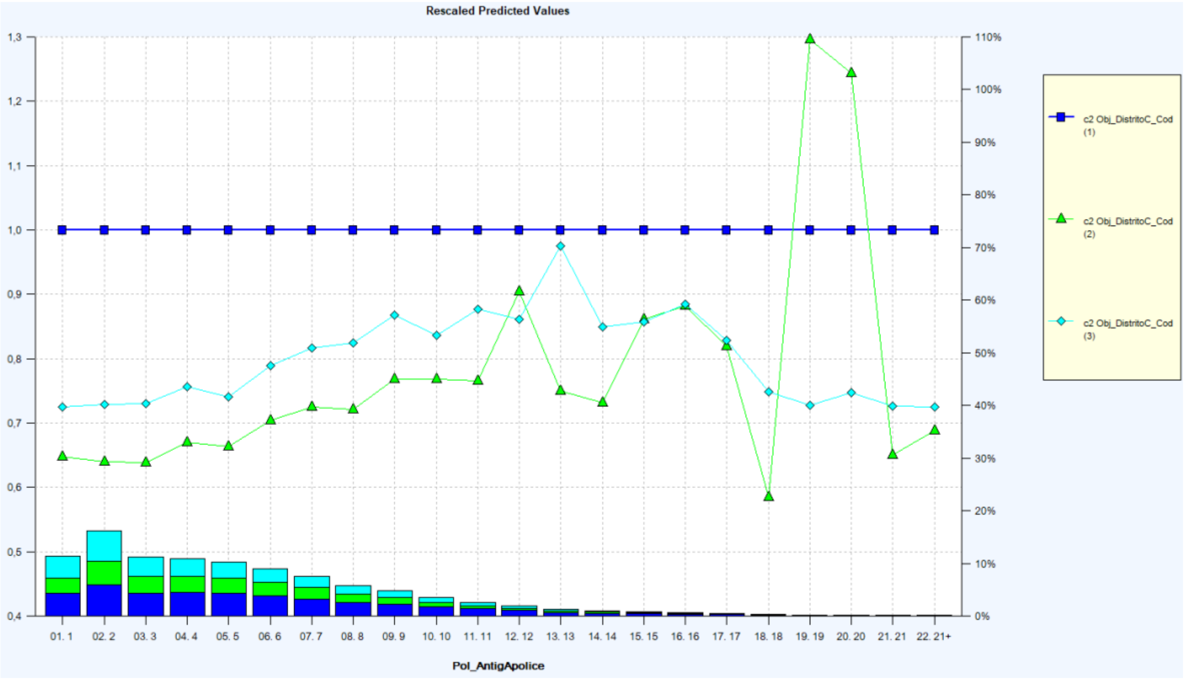


Figure D.2 - Rescaled predicted values considering *Policy Time on Book* (Pol_AntigApolice) and *District* (Obj_DistritoC_Cos), for the frequency model



APPENDIX E

OPTIMAL PARAMETERS FOR GRADIENT BOOSTING

Table E.1 - Top 10 parameters per fold ordered by smallest out-of-sample Poisson deviance, using the 50 000 observations sample for the frequency model

Fold	Nr. of Trees	Shrinkage	Interaction Depth	Bag Fraction	OOS Poisson Deviance
1	37	0.1	2	0.95	0.2802844
1	37	0.1	2	0.95	0.2802844
1	37	0.1	2	0.95	0.2802844
1	37	0.1	2	0.95	0.2802844
1	37	0.1	2	0.95	0.2802844
1	37	0.1	2	0.95	0.2802844
1	71	0.05	2	0.9	0.2803413
1	71	0.05	2	0.9	0.2803413
1	71	0.05	2	0.9	0.2803413
1	71	0.05	2	0.9	0.2803413
2	64	0.1	2	0.95	0.2802087
2	64	0.1	2	0.95	0.2802087
2	64	0.1	2	0.95	0.2802087
2	64	0.1	2	0.95	0.2802087
2	64	0.1	2	0.95	0.2802087
2	64	0.1	2	0.95	0.2802087
2	96	0.1	1	0.95	0.2802124
2	96	0.1	1	0.95	0.2802124
2	96	0.1	1	0.95	0.2802124
2	96	0.1	1	0.95	0.2802124
3	642	0.01	2	0.8	0.2802077
3	642	0.01	2	0.8	0.2802077
3	78	0.1	1	0.8	0.2802383
3	78	0.1	1	0.8	0.2802383
3	78	0.1	1	0.8	0.2802383
3	78	0.1	1	0.8	0.2802383
3	78	0.1	1	0.8	0.2802383
3	78	0.1	1	0.8	0.2802383
3	119	0.05	2	0.9	0.2802488
3	119	0.05	2	0.9	0.2802488
4	116	0.1	1	0.95	0.2793699
4	116	0.1	1	0.95	0.2793699
4	116	0.1	1	0.95	0.2793699
4	116	0.1	1	0.95	0.2793699
4	116	0.1	1	0.95	0.2793699
4	90	0.1	1	0.95	0.2794119
4	239	0.05	1	0.95	0.2794256

4	239	0.05	1	0.95	0.2794256
4	239	0.05	1	0.95	0.2794256
4	239	0.05	1	0.95	0.2794256
5	239	0.05	2	0.95	0.2791458
5	239	0.05	2	0.95	0.2791458
5	239	0.05	2	0.95	0.2791458
5	239	0.05	2	0.95	0.2791458
5	239	0.05	2	0.95	0.2791458
5	892	0.01	2	0.95	0.2791905
5	857	0.01	2	0.9	0.2792089
5	757	0.01	3	0.95	0.2792277
5	160	0.05	2	0.9	0.2792307
5	160	0.05	2	0.9	0.2792307
6	47	0.1	4	0.95	0.2796919
6	47	0.1	4	0.95	0.2796919
6	47	0.1	4	0.95	0.2796919
6	47	0.1	4	0.95	0.2796919
6	47	0.1	4	0.95	0.2796919
6	47	0.1	4	0.95	0.2796919
6	36	0.1	3	0.8	0.2797060
6	36	0.1	3	0.8	0.2797060
6	36	0.1	3	0.8	0.2797060
6	36	0.1	3	0.8	0.2797060

Table E.2 - Top 10 parameters per fold ordered by smallest out-of-sample Gamma deviance, using the training dataset for the severity model

Fold	Nr. of Trees	Shrinkage	Interaction Depth	Bag Fraction	OOS Gamma Deviance
1	133	0.05	1	0.95	15.7664842
1	133	0.05	1	0.95	15.7664842
1	133	0.05	1	0.95	15.7664842
1	133	0.05	1	0.95	15.7664842
1	133	0.05	1	0.95	15.7664842
1	133	0.05	1	0.95	15.7664842
1	493	0.01	1	0.9	15.7665069
1	64	0.1	1	0.9	15.7665218
1	64	0.1	1	0.9	15.7665218
1	64	0.1	1	0.9	15.7665218
2	125	0.05	1	0.95	15.7656292
2	125	0.05	1	0.95	15.7656292
2	125	0.05	1	0.95	15.7656292
2	125	0.05	1	0.95	15.7656292
2	125	0.05	1	0.95	15.7656292
2	125	0.05	1	0.95	15.7656292

2	58	0.1	1	0.9	15.7656960
2	58	0.1	1	0.9	15.7656960
2	58	0.1	1	0.9	15.7656960
2	58	0.1	1	0.9	15.7656960
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	56	0.05	2	0.7	15.7665879
3	59	0.05	2	0.8	15.7667829
3	59	0.05	2	0.8	15.7667829
3	59	0.05	2	0.8	15.7667829
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	33	0.1	1	0.8	15.7657817
4	180	0.05	1	0.95	15.7658257
4	180	0.05	1	0.95	15.7658257
4	180	0.05	1	0.95	15.7658257
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	59	0.05	2	0.7	15.7670927
5	33	0.05	5	0.8	15.7671482
5	33	0.05	5	0.8	15.7671482
5	33	0.05	5	0.8	15.7671482
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	75	0.1	1	0.95	15.7669788
6	149	0.05	1	0.95	15.7670874
6	149	0.05	1	0.95	15.7670874
6	149	0.05	1	0.95	15.7670874