NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

MODELLING LARGE CLAIMS IN NON-LIFE INSURANCE

An application in motor insurance industry based on Extreme Value Theory and Gradient Boosting Machine Techniques

Maria Inês Trindade Vieira

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

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

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

MODELLING LARGE CLAIMS IN NON-LIFE INSURANCE

An application in motor insurance industry based on Extreme Value Theory and Gradient Boosting Machine Techniques

by

Maria Inês Trindade Vieira

Internship report presented as partial requirement for obtaining the Master's degree in Information Management/ Master's degree in Statistics and Information Management, with a specialization in Risk Analysis and Management

Advisor: Professor Manuel Vilares

ACKNOWLEDGEMENTS

I would like to thank Fidelidade which gave me the opportunity to develop my masters final work during my internship. Thank you to my team, especially Catarina Jácomo for all the support, Febio Levezinho, Rui Esteves and João Pinto for their feedback. No doubt their help pushed me and improved my work into a higher level.

In addition, I would like to thank Nova Information Management School and to all professors who provided the best study conditions during my masters journey. A special thanks goes to my adviser, professor Dr. Manuel Vilares, for his support and feedback during all my study.

To my family, for their love, encouragement and unconditional support. Especially my parents who always supported me with everything I needed during my academic life. For them, I am really thankful for all their motivation and patience.

Thank you to my colleagues and friends who lived this journey with me, always offering love and the best messages that it would be possible to finish my masters final work.

To all who directly or indirectly participated in my training, my eternal thanks. To all those who are by my side, in the most difficult moments and in the life's successes celebrations. Thanks!

ABSTRACT

In recent years, there has been an increased interest in non-life insurance sector, especially in motor insurance segment. In fact, thousands of collisions, crashes, breakdowns, and thefts are reported to Portuguese insurance companies every day. Some with bigger claims than others- the extreme events.

Extreme value theory has been used to develop models for describing the distribution of this rare events. Modelling extremes is challenging and there is usually a trade-off between being as close as possible to the "asymptotic regime" and, at the same time, having as much data as possible for estimates to be reliable. An important challenge in the application of such extreme models is the choice of a threshold, beyond which point the asymptotically justified extreme value models can provide good extrapolation.

The focus of this study will be the greatest losses. This aims to contribute to the literature of extreme values analysis - modelling extreme values based on thresholds. Moreover, this project pretends analyse the mains factors that facilitate the decision-making process by the insurance company. Several risk factors that could explain the costs' behaviour are used to model the claims' severity. The applied methodology will focus on Gradient Boosting machine learning techniques, from which a modelling process is developed on a set of data of Fidelidade Motor Insurance portfolio.

The results show that it makes sense to have two borders and to analyse factors of relevance in each segment. Regression models are built, and it is achieved, in the three main models developed, vehicle category is always very present as the main responsible variable followed by the variable's year analysis and years in portfolio.

KEYWORDS

Motor Insurance, Claims Severity, Extremal Events, Prediction Models, Gradient Boosting

RESUMO

Nos últimos anos, tem-se assistido a um grande interesse pelo setor dos seguros não vida, particularmente no segmento automóvel. De facto, milhares de colisões, despistes e avarias são comunicados diariamente nas seguradoras portuguesas. Alguns com maior gravidade e custos do que outros – os eventos extremos.

A teoria dos valores extremos é usada para desenvolver modelos extremos para descrever uma distribuição dos mesmos. A modelagem é um enorme desafio e geralmente há um *trade-off* entre escolher um valor mais próximo possível do regime assintótico do conjunto dos dados e, ao mesmo tempo, o segmento acima do valor escolhido conter o máximo de dados para existir qualidade na estimação. Um desafio importante na aplicação de tais modelos é a escolha de um valor fronteiro- a partir do qual os modelos de valores extremos podem fornecer uma boa extrapolação.

O foco deste estudo será os custos mais elevados existentes com sinistros automóveis (apenas sobre a cobertura de responsabilidade civil). O mesmo visa contribuir com literatura sobre valores extremos e modelação de valores extremos com base em fronteiras. Além disso, pretende-se neste estudo analisar as principais variáveis que explicam os modelos. Vários fatores de risco que podem explicar o comportamento dos custos são usados para modelar a gravidade dos sinistros. A metodologia é aplicada à carteira de dados da Fidelidade. Os modelos são criados utilizando a técnica de *Gradient Boosting Machine*.

Os resultados mostram que faz sentido ter duas fronteiras e analisar fatores de relevância em cada segmento. São construídos três modelos de regressão e a categoria do veículo é o fator de relevância que aparece mais presente como a principal variável responsável do modelo, seguida do fator ano do sinistro e do número de anos de carteira do segurado.

PALAVRAS-CHAVE

Seguro Automóvel, Responsabilidade Civil Automóvel, Gravidade de Sinistros, Eventos Extremos, Modelos Preditivos, Gradient Boosting

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ATSM	Automatic Threshold Selection Method
ECDF	Empirical Cumulative Distribution Function
EVA	Extreme Value Analysis
GEV	Generalized Extreme Value
GLM	Generalized Linear Models
GPD	Generalized Pareto Distribution
GBM	Gradient Boosting Machine
KS	Kolmigorov-Smirnov
Max	Maximum
Min	Minimum
MRL Plot	Mean Residual Life Plot
N	Number of elements in the data set
POT	Peaks-over-threshold
RMSE	Root Mean Square Error
Sd	Standard deviation
Th	Threshold
1st Qu.	First Quartile

3rd Qu. Third Quartile

1 INTRODUCTION

1.1 CONTEXTUAL PERSPECTIVE

The insurance activity plays an important role in the structure of modern societies, namely in motor civil liability insurance. In fact, it is an activity whose function is often carried out in adverse circumstances. Every day, Portuguese insurance companies report thousands of crashes, collisions, breakdowns and thefts.

In recent years there has been an increase in the costs of the most serious claims, mainly due to the effects of significant increases in compensation for death and compensation for moral damages.
Therefore, insurance companies need to keep a certain amount of capital buffers, or initial reserves, for prudency and regulatory reasons in case the cash flows from the premiums momentarily do not cover unexpected large claims.

However, in a short run, there is a risk that losses that occur will perhaps be much higher than expected. This risk needs to be measured and the insurance company must hold a sufficient level of capital to cover this risk with high probability. That means it is important to measure the risk premium connected with the business.

Modelling extreme events is challenging and there is usually a trade-off between being as close as possible to the "asymptotic regime" and, at the same time, having as much data as possible for estimates to be reliable. Since extreme events are rare, getting a shortage extremes dataset to analyse turns difficult to find such an appropriate large claims estimate. Unlike inferences related to the means of a random variable, when considering the extremes, it is most of the data that is discarded and the discrepant ones that are interesting.
A well-known problem in extreme value analysis (EVA) is threshold selection. There is a common question in insurance around large claims- should or should not large claims be separated from normal claims when making reserve predictions?

The aim of the study is not to look at the average behaviour of the claims that the company faces, but at the stochastic behaviour of the process of claiming its extremes.

1.2 OBJECTIVES AND RESEARCH QUESTIONS

The focus of this study will be the greatest losses. In statistical terms, our attention should be on the right tail of X. This project addresses the problem using methods developed in extreme value theory- a branch of statistics dedicated to study this type of events. This theory is usually applied in various domains, such as finance, insurance, or risk analysis of natural risks.

An important result will be the conditional distribution of excess losses at a certain threshold level. In risk analysis, and especially for applications in insurance, it is important to anticipate the losses that a particular company may face in the near future. In many applications of loss data distributions, a key concern is fitting the loss data in the tail.

Report the best ways to outline the threshold for which the cost of motor insurance claims should be defined as severe is proposed. Moreover, the idea is to model these claims classified as serious in a faster and more automatic way, allowing a lower flow of manual readjustment in the claim process in the future. The main objective of this thesis is to model extreme claims of motor insurance.
The following research questions are suggested:

- What should be considered a large claim?
- What methodology could help to identify possible thresholds?
- Should insurers handle large claims defining a threshold?
- What are the variables that explain large claims costs behaviour?

These are the major issues that have led to the development of this work. This internship report, prepared in the context of my trainee experience in Fidelidade, aims to contribute to the literature of extreme values analysis, more specifically, modelling extreme values based on thresholds.

Through the work, it is planned to clarify why it is important to look at large claims and hence the importance of choosing a frontier value.

After exposing the most recent literature (chapter 2 content), methodologies that adapt to the data in question are chosen to pick candidates for frontier values. The focus is on developing methods to select an appropriate threshold. Then, the work follows with the goal to develop predictive models and find variables that explain costs behaviour. Only after this point, error measures between frontier candidates are presented to choose the final border value. This set of precedents is applied to insurance industry in chapter 3.

To complement and to achieve better conclusions, the values of the probability of a large claim occurred and estimated risk premiums are also presented. These results will be presented in chapter 4.

2 LITERATURE REVIEW

2.1 LARGE CLAIMS

A large claim does not have a universally accepted quantitative definition. Large in the sense of magnitude is relative and must be looked from the perspective of the insurance company, the country that is inserted and, the most important part, the line of business in case (McNeil, A. J. (1997)). It could be difficult to compare in an equal way these classifications. For example, a vehicle claim of €50,000 may not be considered large in United States, but it could be significant in Portugal.

Large claims are the ones that occur less frequently, occur in irregular patterns, and have a higher magnitude than average (McNeil, A. J. (1997)). Large claims can be classified as those having a severe value and a substantial impact. This definition could be personal and relative. Moreover, it was reported in literature (Charpentier, A. (Ed.) (2014)) that the treatment of large claims depends on the quality of data available. Here comes the difficulty in predicting large claims. Extreme value theory (and associated research advances) has helped to overcome the problem previously presented. Assuming certain distributions, proven to have substantial results in these extreme segments, it is possible to define cut-off point values to separate large claims from the rest. From this point, the work of modelling and predicting extreme costs can be done.

2.2 EXTREME VALUE THEORY

Extreme value theory (EVT) is a numerical tool concerned with the modelling of extreme events in different areas. McNeil, A. J. (1997) studied EVT main uses. The extreme value theory (EVT) is generally used in finance, meteorology, and insurance areas (Uwingabire, J., 2018). The theory is relevant for large claims identification and also for modelling extreme insurance.

"Extreme value theory deals with the stochasticity of natural variability by describing extreme events with respect to a probability of occurrence", according to Marielle Pinheiro and Richard Grotjahn (2015). Since the focus of this study is on extreme claims, which can also be described in terms of a probability distribution function, we should use the information from the resultant distribution to analyse trends and the likelihood that an extreme event will occur (Pinheiro, M., & Grotjahn, R., 2015). EVT can also be defined as a theory "for assessing the asymptotic probability of extreme values, further expanding that the theory models the distribution of tail part anywhere the risk exists" (Ren, F., & Giles, D., 2007).

One possible solution to lead with large claims is undertaken by fitting a maxima series defined by

$$Mn = \max\{X1, \ldots, Xn\} \quad [1]$$

They can be defined as the r-largest values per time period (Haigh et al., 2010). This technique can be applied with different time horizons, but the most common form is to be defined in years. When block maxima approach is used, there is no simple definition of a threshold value. Instance of a fixed value, there is a set of extremes chosen (depending on the dataset used).

Choosing the appropriate block size, where the maxima values are taken, involves a trade of between the variance of the model and the systematic bias (Charpentier, A. Ed., 2014). Succinctly, taking large block sizes means that the maxima is drawn from many underlying observations. This leads to fewer data points and larger variance in the estimation. The opposite then holds for choosing smaller block sizes (Paldynski, H., 2015). Previous studies found that this type of methods have their own problems. In fact, this extreme values approach uses the block maxima for inference, and thus leave out possibly interesting data. Paldynski (2015) and Haigh (2010) researchers defend, block maxima show inefficiency in the way it uses the data.

The peaks-over-threshold (POT) approach is an alternative method to block maxima. According to Hanafy, M. (2020), peaks-over-threshold methodology is characterized by modelling exceedances above a pre-chosen threshold to distinguish the tail of extreme values (Coles, 2001; Ghil et al., 2011). A high threshold is most likely to only contain data that are truly extreme.

Using the POT-method, it is understood that all sample points that exceeds the threshold are of interest and separated from the rest of the data for further analysis. In an *Application of the Peaks-Over-Threshold Method on Insurance Data (2018*), Max Rydman defines the POT-method as one of many methods that falls under extreme value analysis, looking at extreme values of some sample that exceeds a certain threshold value.

$\mathrm{Fu}(\gamma)$ should be recognized as the probability that v lies between u and $\mathrm{u}+\gamma$ conditional on $v > u$.

$$F_u(\gamma) = \frac{F(u+\gamma)-F(u)}{1-F(u)} \qquad [2]$$

The variable Fu defines the right tail of the probability distribution. It is the cumulative probability distribution for the amount by which v exceeds u given that it does exceed u (Hanafy, M., 2020).

The generalized pareto distribution is a very popular two parameter model for extreme events formally defined as the probability of exceeding pre-determined threshold (Rydman, M., 2018) (Charpentier, A. Ed., 2014). Chiang Lee (2009) empirical results show that the general pareto (GP) method theoretically is a well-supported technique for fitting a parametric distribution to the tail of an unknown underlying extreme distribution, capturing the tail behaviour of insurance large losses very well. Hanafy, M. (2020) concluded that when data exceeds high thresholds, the generalized pareto is a useful method for estimating the tails of loss severity distributions.

Pickands (1975) shows that if it is a random quantity with distribution $F_u(\gamma)$, for a sufficient large u, assuming that the variables are independent and identically distributed (i.i.d.),

$$F(u) = P(\gamma \leq u + \gamma \,|\gamma > u) \qquad [3]$$

can be approximated by a generalized pareto distribution (GPD).

The theorem of Pickands (1975), Balkema and de Haan(1974) states that for a large class of underlying distributions, there exists a positive function $\beta(u)$ such that

$$\left|F_u(\gamma) - G_{\xi,\beta(u)}\right| = 0 \qquad [4]$$

where γ is the rightmost point of the distribution, u the threshold, $F_u(\gamma)$ the excess distribution function and $G_{\xi,\beta(u)}$ the General Pareto distribution.

The respective G distribution function is

$$G(\gamma|\xi,\beta,\mu) = \begin{cases} 1 - \left(1 + \frac{\xi(\gamma-u)}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - exp\left(-\frac{(\gamma-u)}{\beta}\right), & \xi = 0 \end{cases} \qquad [5]$$

Where $\beta>0$ is the scale parameter and ξ is the shape parameter. The function is only valid when $u\geq 0$ for $\xi\geq 0$.The data exhibit heavy tail behaviour when $\xi > 0$. The shape parameter (ξ) value increases as the tails of the distribution become heavier. Depending on the shape parameter, a Pareto, a Beta, or a Pareto distribution will be formed. The threshold is represented by u - the reference value for which GP excesses are calculated.

Recent literature, including Pešta, M., Petrová, B., Procházka, J., & Smolárová, T. study, proved generalized pareto distribution is valid only above very high thresholds. So, the estimation of the optimal threshold level seems to be a crucial issue and an important first task in EVT. No satisfactory fit for the extreme events can be obtained using a classical parametric model. Due to that fact, and after choosing a threshold, the generalized pareto distribution is a common form to use when we are modelling large claims.

More common to general pareto distribution is gamma distribution. This is a two-parameter family of continuous probability distributions. Gamma distribution has also a shape and a scale parameter- κ and θ, respectively. This distribution covers the positive skewness portion of the curve. the distribution has most of the data concentrated on the initial left side, with the extremes having a very small representation in the general conjuncture of the data (Kim, A. (2019)).

Considering $X \sim Gamma\ (\kappa, \theta)$, the respective gamma probability density function is given by

$$\frac{x^{\kappa-1}e^{-x/\theta}}{\Gamma(\kappa)\theta^{\kappa}} \qquad [6]$$

The gamma distribution can be used in a range of disciplines including the size of loan defaults and insurance claim cost (Drakenward, E., & Zhao, E., 2020).

2.3 THRESHOLD CHOICE

The threshold (u) choice is an important step and not that easy to find. If u is sufficiently high, the exceedances follow approximately a random sample from a GPD, as explained in the previous section. The hard question is how to find an effective way to choose the lowest threshold such that the GPD fits the sample of exceedances.

If the threshold is too high, the asymptotic arguments overstating the derivation of the model. By contrast, too low threshold will generate more excesses to estimate the parameters. Gamerman (2004) aims we should choose a low threshold, but greater than the mean value. Although the estimation becomes biased, it is preferred than a high threshold (that will produce higher variance).

The traditional methodology (quantile or percentile) is considered the fastest method to set a threshold value. In fact, the threshold definition task could be made easier if there was a fixed percentile previously defined. 90^{th}, 95^{th} and 99^{th} percentiles are the commonly used in previous studies (detailed in Rydman, M., 2018). However, due to the difference in behaviour between different data, it is not necessarily a reliable way of setting a "good threshold due to the inevitable difference between most data" (Rydman, M., 2018). Moreover, this was successfully established as described by Rydman (2018) as also a subjective methodology.

Pešta, M., Petrová, B., Procházka, J., & Smolárová, T. studied graphical methodologies for threshold choice and concluded that "the classical fixed threshold modelling approach uses graphical diagnostics, essentially assessing aspects of the model fit, to make an a priori threshold choice". An advantage of this approach is that it requires a graphical inspection of the data, comprehending their features and assessing the model fit. On the other hand, this methodology can require substantial expertise and can be rather subjective at decision time, since stability could be interpreted differently from one specialist to another. Gamerman (2004) and Coles (2001) suggests the mean residual life plot and the threshold range plots as the main diagnostics.

Mean residual life plots represent a set of possible thresholds and the respective "mean excess" (the mean value of the exceedances over the threshold, minus the threshold value) (Wang, Yingguang & Yiqing, Xia & Liu, Xiaojun , 2013). Linearity above a threshold level in this plot indicates a consistency with a property distribution (Rydman, M., 2018). When the plot starts showing a linear behaviour, a suitable threshold can be estimated (Charpentier, A. Ed., 2014). When the threshold gets too large due to the variance of the few extremes left, the plot is likely to lose its linearity.

Threshold range plots are parameter stability plots where the threshold is obtained at the point when a reasonable variation (stability) is found, and the confidence interval bars start to grow significantly. The value where the convergence of these two factors takes place should be taken as the threshold value. Looking at the results, we can find a suitable threshold at the lowest value where the plot is approximately constant (Scarrottand MacDonald, 2012).

Comparing to the mean residual life plot, this graphical method is generally preferred since it demonstrates more clearly the parameter sensitivity with respect to the threshold. In addition, parameter uncertainty is directly illustrated in the plot and the method can be generalized (Rydman, M., 2018). However, in both methodologies, there may occur cases where the plot is never linear and then graphical methodologies become an element that gives us very poor results. Depending on the data, there may be cases where the parameter stability plot will not drop any further useful information (Charpentier, A. Ed., 2014).

2.4 AUTOMATIC THRESHOLD

Given the difficulties presented above regarding graphic methods, it would be desirable to have a more automatic method where the threshold choice could be done quickly and without major uncertainties (Bader, B., Yan, J., & Zhang, X., 2018). Automatic threshold selection method (ATSM) is a simple, threshold selection method that was developed by Thompson et al. (2009). The author used simulated data in order to show the effectiveness of the automatic method.

The method consists in a sequence of goodness-of-fit tests to the exceedances over each candidate threshold in an increasing order. Under EVT, the only possible non-degenerate limiting distribution of properly rescaled exceedances of a threshold u is the GPD as $u \to \infty$.

Consider a fixed set of candidate thresholds $u_1 < \cdots < u_n$. For each threshold, there will be $n_i$excesses, i = 1, . . . , l. The sample of exceedances above u_1 should be large enough to insure reliable estimation. The sequence of general pareto assumption null hypotheses "can be stated as H_{0_i}: The distribution of the n_i exceedances above u_i follows the GPD. For a fixed u_i, many tests are available for this H_{0_i}" Bader, B., Yan, J., & Zhang, X. (2018). An automated procedure can begin with u_1 and continue until some threshold u_i provides an acceptance of H_{0_i} (Choulakian and Stephens, 2001; Thompson et al., 2009). More than the simplicity of the method, ATSM has the advantage of being computationally inexpensive. The problem, however, is that unless the test has high power, an acceptance may happen at a low threshold by chance and, thus, the data above the chosen threshold is contaminated. One could also begin at the threshold u_i and descend until a rejection occurs, but this would result in an increased type I error rate. Another disadvantage presented by the author is that the algorithm might not converge. However, this is a situation that could rarely happen.

Method	Description	How quality is measured
Mean Residual Life Plot	Graphical method. Plotting the mean excess against u, for a range of values of u, supports the selection of a threshold.	Linear trend should be identified before the threshold.
Threshold Range Plot	Graphical method. Find an appropriate threshold for General Pareto models by fitting them to a sequence of thresholds to find the lowest threshold that yields roughly the same parameter estimates as any higher threshold.	We are looking for a range of values where we can see sufficiently large bar size in the plot.
Automatic Threshold Method	A sequence of goodness-of-fit tests to the exceedances is used over each candidate threshold in an increasing order. KS Test is used in this work (by the size of the error). We are using the GP distribution and test is done using the following hypothesis: "The distribution of the exceedances above the threshold considered follows the GP model".	The final threshold is returned when the test value is smaller than 0.01.

Table 1. Summary of the different methodologies

2.5 GENERALIZED LINEAR MODELS (GLM)

This methodology is the most popular for non-life insurance pricing and modelling models (Nogueira, F. A. V. (2015)), especially due to its simplicity and the evolution of new technologies, which allowed an increase in efficiency and simplicity of handling with large amounts of data.

GLM models seek to fit a function that relates the variation of a dependent variable Y with the variation of independent variables (defined as X- variables). The independent variables, as they measure the exposure of risk factors (r risk factors), are those that can influence the occurrence of the event. The portion of variation of the dependent variable, that cannot be explained by the variation of the independent variables, it is attributed to a random variable called error (Goldburd, M., Khare, A., & Tevet, D., 2016).

A GLM model is often reasonable choice of model in non-life insurance pricing. A maximum likelihood method for estimating the regression parameters β is used in

$$g(u_i) = x_i^T = \sum_{j=1}^{r} x_{ij}\beta_j \, , i = 1, \dots n \qquad [6]$$

GLMs are used as a model that represents the relationship between a variable whose outcome is wished to predict and one or more explanatory variables. It can be used to fit a gamma regression for the average claim amount analysis, for example. The idea here in non-life actuary is to use GLM to identify which factors influence the different claims severity levels, in order to create a clearer modelling. In other words, via GLM the goal is to try to maximize the systematic component at the cost of the random component- trying to get the maximum predictive value from the explanatory variables.

2.6 GRADIENT BOOSTING MACHINE (GBM)

Although GLM method was proved to be advantageous, new methodologies (Hall, P., & Proksch, M., 2020) can provide a more accurate and transparent way to create more advanced predictive models. Gradient Boosting machine (GBM) learning technique is one of these methodologies.
In short, Gradient Boosting is defined as a machine learning technique that combines gradient-based optimization and boosting trees (Click, C., Malohlava, M., Candel, A., Roark, H., & Parmar, V., 2017). GBM technique "sequentially fit new models to provide a more accurate estimate of a response variable" (Lu, H., Karimireddy, S. P., Ponomareva, N., & Mirrokni, V., 2020). Combining multiple weak-learners, GBM creates a strong ensemble to minimize a model's loss function. (Click, C., Malohlava, M., Candel, A., Roark, H., & Parmar, V., 2017). Gradient Boosting Machine (GBM) is a procedure most used for classification or regression.

GB provides interpretable results via the relative influence of the input variables and their partial dependence plots. This is a critical aspect to consider in a business environment, where models usually must be approved by non-statistically trained decision makers (Guelman, L., 2012).

GB is known as a flexibility technique since it requires very little previous data analysis and processing- which is known as one of the most time-consuming activities in a data mining project. Using GBM, a model selection is done – this is a great help for the analysts (Guelman, L., 2012). Boosting trees use are helpful because they help to increase the technique accuracy for each variable incremental impact. Especially for non-linear models, GBMs are more accurate in their predictions (Hall, P., & Proksch, M.,

2020). However, it also decreases speed and user interpretability. This means that the final analysis of the method could be unsatisfactory (Click, C., Malohlava, M., Candel, A., Roark, H., & Parmar, V., 2017).

2.7 POINTS TO HIGHLIGHT FROM THE LITERATURE REVIEW

The literature review shows that the traditional methodology can be used as a starting point, for the analyst to have a first idea of possible threshold values.

In extreme value theory, the block maxima approach is proposed as a method to divide the observation period into equal size periods and to focus on the maximum observation in each period. As an alternative, parametric statistical methods for EVT are applied. The probability distribution of those selected observations is approximately a generalized pareto distribution (Ferreira, A., & De Haan, L., 2015). Gamma distribution is also used, especially in model extremes with regression models.

Coles (2001) comparison concluded that threshold approach is more efficient than the block maxima method in the case that the dataset is completed. In fact, as all values exceeding a certain threshold, they can serve as a basis for model fitting. In some cases, fitting distributions to block maxima data is an inefficient approach as only one value per block is used for modelling. Given this set of advantages of peak over threshold, compared to block maxima, only POT will be used in this work.

Regarding the threshold choice, we can conclude that diagnostics plots could result in subjective judgement. The ATSM approach can help to reduce this subjectivity. However, all this set of procedures does not mean that we can confirm immediately that the value chosen will be the best threshold. Here comes the importance to compare the model with other different threshold possible values. In this work different methods will be used to choose threshold candidates.

Finally, GLMs are concluded to be useful for the modelling part and to understand the focus group with more probability to have a large claim. It is of great interest for insurance companies' actuarial projects, including the ones that focus on large claims study. GBM is a good alternative method to GLM to build insurance cost models. Comparing to GLM, the utility of GBMs increases as we lead with more complex scenarios (Hall, P., & Proksch, M., 2020). GBMs are used in the practical part of this work.

3 APPLICATION TO INSURANCE INDUSTRY

3.1 DATA

In this work, the property insurance portfolio of Fidelidade is considered, which is the insurance company with the biggest market share in Portugal, including in the motor sector. In 2020, the company had about 28% of market share – an extended volume, considering that, in the same year, Portugal had more than 8 million insured vehicles. The volume of 2020 premiums in motor branch reached the value of 1,877 million euros.

The data used in practical research is a dataset that contains civil liability claims from the 1st of January 2015 to the 31st of December 2020. The claims collected are motor claims.

In Portugal, civil liability insurance in motor vehicles is mandatory by law. The fault party should be able to pay any damages caused to third parties. The minimum capital mandatory in this insurance is fixed with a value of 7,290 million euros (in Portugal, during the period here studied).

In this study, all values with a cost smaller than 100 euros are removed. These costs correspond only to costs of technical and financial appraisals and turn out not to be significant. This ensures that only the cases where the losses occurred are collected.

There are claims with more than one open process in the company's portfolio. The costs of these cases are grouped to obtain the total cost of the claim. The dataset consists of a portfolio of 512,422 motor claims. We have information on each policy, as well as the possible values that each of those variables can take.
Table 2 shows the number of loss events, percentage of the loss events including sum and percentage of loss amount. Table 3 represents the main statistics of our data.

Cost Ranges (€)	Number of events	Percentage of events (%)	Sum of costs (€)	Percentage of costs per range (%)
[0-5000[	490,320	95.69	525,234,395.54	49.13
[5000-10000[	12,386	2.42	84,914,847.06	7.94
[10000-50000[	7,778	1.52	150,169,336.47	14.05
[50000-200000[	1,577	0.31	149,141,822.79	13.95
>=200000	361	0.07	159,578,277.60	14.93
Total	512,422	100	1,069,038,679.46	100

Table 2. Frequency of loss data

N	Min	1st Qu.	Median	Mean	3rd Qu.	Sd	Skew	Kurtosis	Max
512,422	100	524	852	2,086	1,164	21,180	211	71,734	8,283,001

Table 3. Dataset main statistics (€)

Starting by analysing high percentile values, it is possible to see that there is a noticeable gap between using which percentile presented in table 4. The higher the percentile value considered, the larger the variance due to the limited amount of data treated as exceedances.

	95%	99%	99,9%
Threshold (€)	4,542	17,835	154,225
Number of exceedances	25,620	5,125	513

Table 4. Extreme percentiles

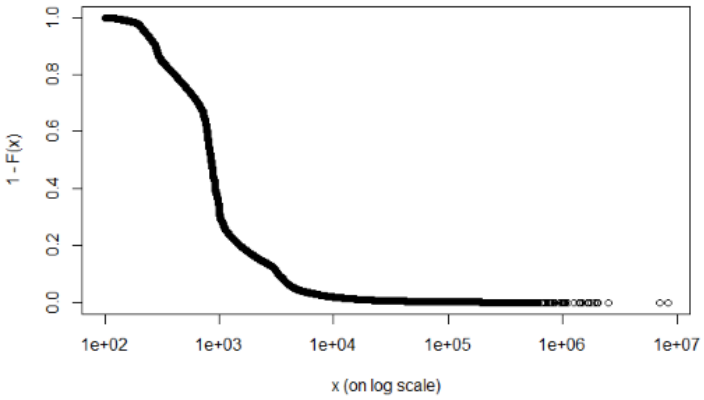


Figure 1.Cumulative function distribution representation

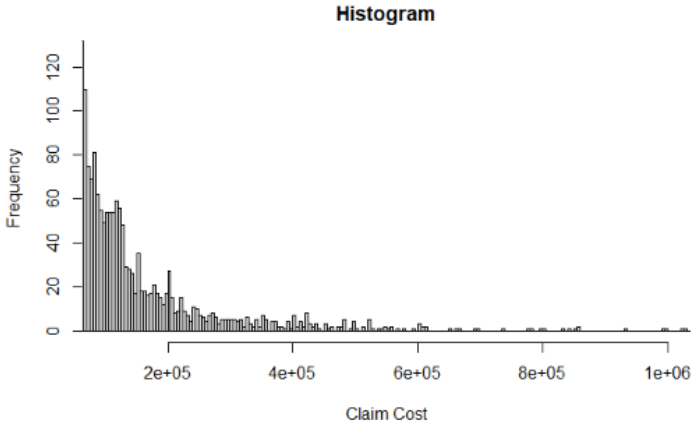


Figure 2.Representation of the cost's distribution on the histogram

Histograms, percentile plots and cumulative distribution representations are graphical tools useful to identify the behaviour of the data in general and, specifically, of its highest values – our extreme costs.

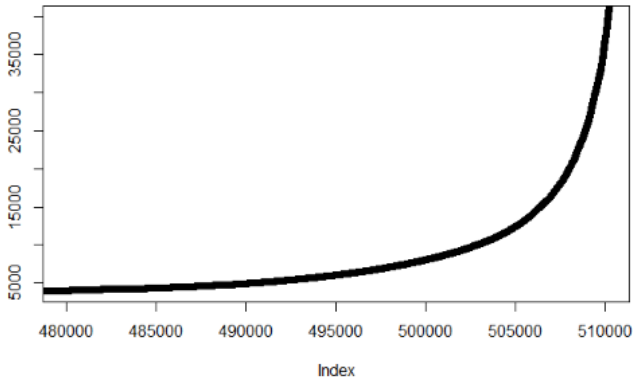


Figure 3.Percentile plot- zoom on €4,000-40,000

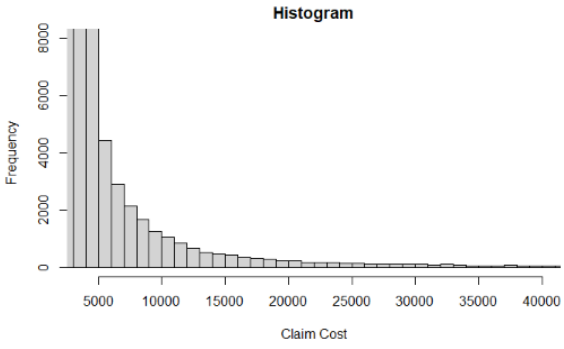


Figure 4. Histogram- zoom on €4,000-40,000 interval

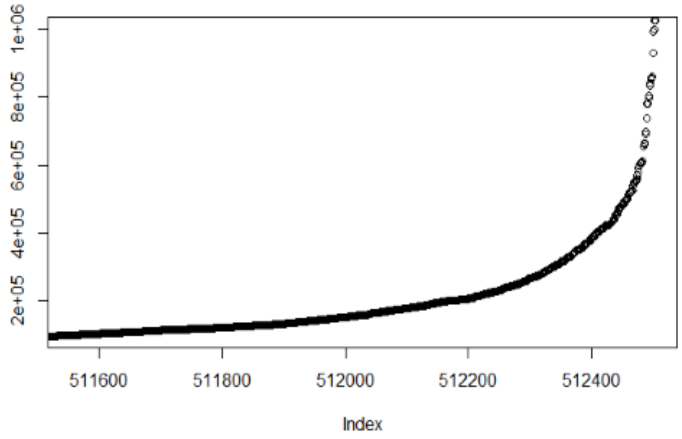


Figure 5. Percentile plot- zoom on €100,000-1000,000

Different scales and zooms are used to prove that it is not obvious to choose a frontier point. As we approach the segment with the highest costs, new cut-off hypotheses arise (where the frequency of claims turns out to be less and less and with greater variance).

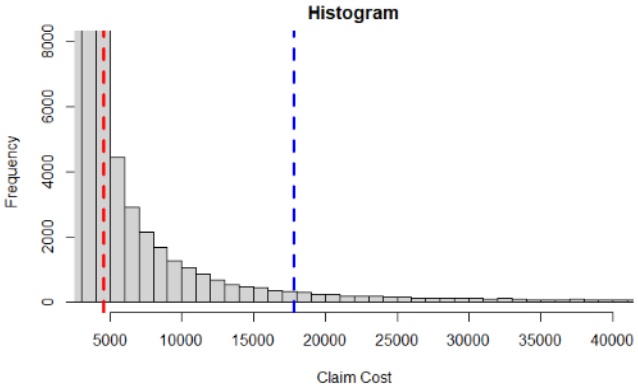


Figure 6. Histogram with 95 and 99 percentiles represented (red and blue dots lines, respectively)

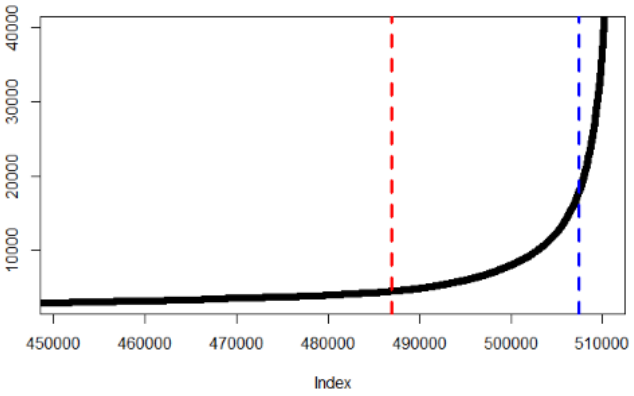


Figure 7. Percentile plot with 95 and 99 percentiles represented (red and blue dots lines, respectively)

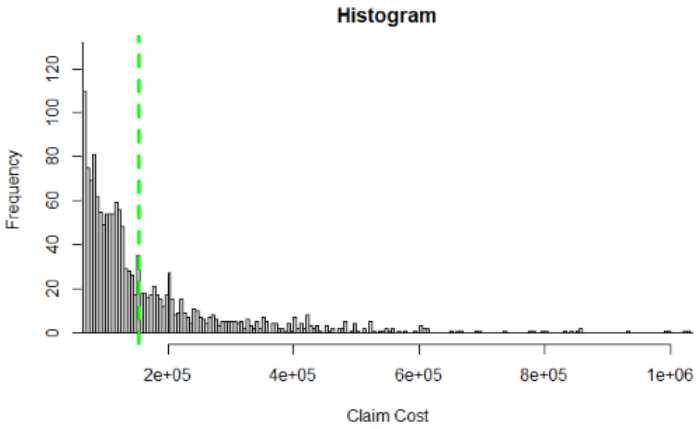


Figure 8. Histogram with 99.9 percentile represented

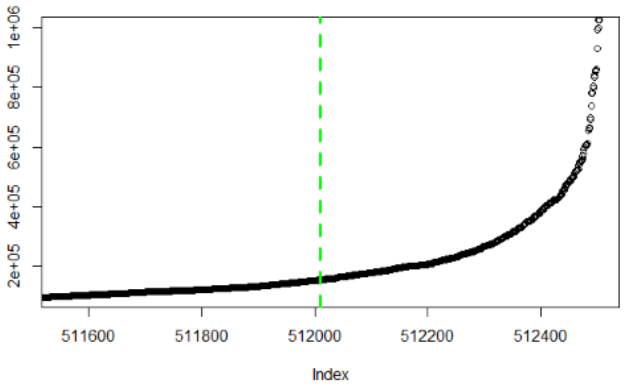


Figure 9. Percentile plot with 99.9 percentile represented

3.2 SELECTING A THRESHOLD-METHODOLOGY

Last section results gave to the reader an idea of the percentile values in terms of costs. They could be a strategic set of starting points, but how should we define if they coincide with a strong threshold?

Recalling the methods described in the literature section, comparing the literature review conclusions regarding block maxima approach and POT approach, block maxima methodology will not be used in this practical application. The main reason for this decision is that our dataset is not a time series sequence. So, the methodology to select the dataset that should be treated as exceedances will be done according to the peak-over-threshold (POT) methodology.

After discarding the block maxima hypothesis, the focus will be defining the cut location. The methodology chosen to define the threshold consists in a combination of the various methods accepted in the literature section. The central part of the POT-method is setting the threshold to understand which costs should be considered exceedances. If a fixed momentary amount is used it tends to be in the form of a simple round figure and tends to be reviewed infrequently. Some companies have been using the same fixed monetary loads for many years failing to reflect the effects of claims inflation.

After analysing the initial dataset, as a first step and as a first bounder consider, the graphical methods and the automatic method are implemented. The idea is to create a set of possible boundaries. Then, these values are compared, returning to the analysis of the percentile plot and the histogram plot (where the cut off suggestions are exposed). We still want to make sure that a balance is kept such as to limit both bias and variance from getting too large.

The main idea of this section is not to choose a final or the final values for the thresholds, but instead identify a range of values where make sense to define the final bounder.

Some experiences will be done, considering the hypothesis that we can define more than one threshold (if we identify more than one extreme behaviour).

3.2.1 MEAN RESIDUAL LIFE PLOT

The mean residual life plot (MRL) of the data is represented in Figure 10. According to the literature, the boundary should be chosen considering that the values above the threshold have a linear trend. In our data, we observe that the linear trend is concentred in the first part of the data (values less than €500,000). However, in this plot (Figure 10) is difficult to identify a possible threshold (we were only able to exclude very low values or very high values). Taking a zoom view at the values under €500,000, the plot looks somewhat linear up to a threshold of roughly €20,000, but when the threshold reaches the range between €200,000 and €500,000 the behaviour of the plot starts to shift.

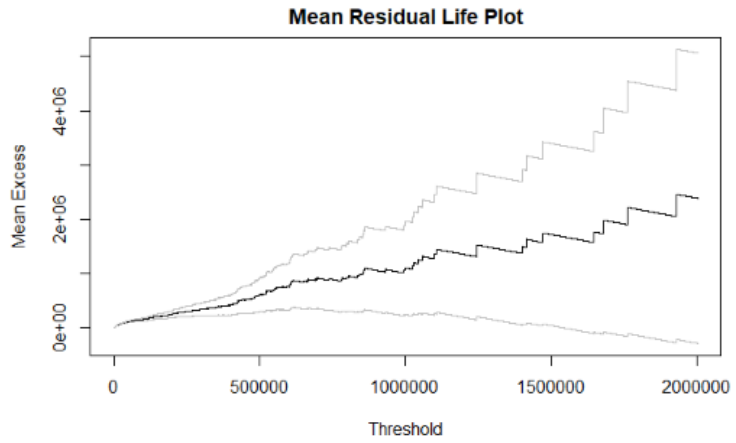


Figure 10. Empirical mean residual life plot. Plotted using mrlplot from the evd package in R. The dashed lines representing a 95% confidence interval

Since the information of the MRL as the threshold gets large is not of interest, Figure 11 x-axis has simply been restrained to a smaller range of thresholds (costs under €500,000). From this point, the theory suggests that there should exist a point at which for any larger threshold the main excess will be a linear function of the threshold and the scale parameter. Looking at Figure 12, a somewhat linear behaviour can be seen starting at a threshold of around €5,000. However, around €30,000 point it starts to show a linear behaviour with a larger gradient. Focusing on smaller values, where most of our data resides, €5,000 to 30,000 interval, it shows a more reasonable linear trend. Focus now on the plot with zoom on €5,000- 20,000 interval, Figure 13, the entire trend is linear.

Figure 11. Empirical mean residual life plot. Plotted using mrlplot from the evd package in R. The Dashed lines representing a 95% confidence interval. Zoom on the costs below €500,000

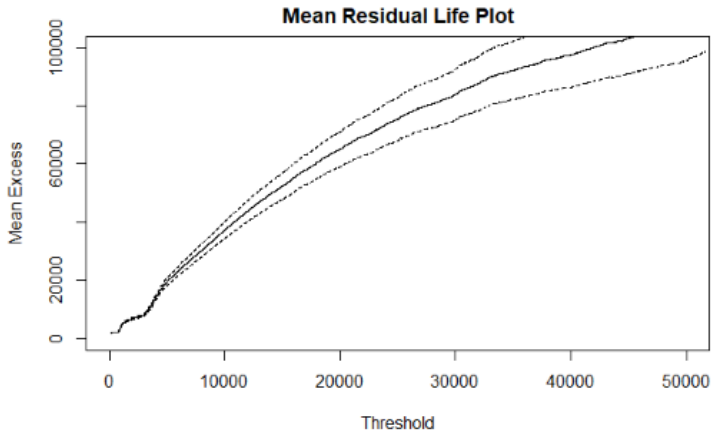


Figure 12. Empirical mean residual life plot. Plotted using mrlplot from the evd package in R. The Dashed lines representing a 95% confidence interval. Zoom on the costs below €50,000

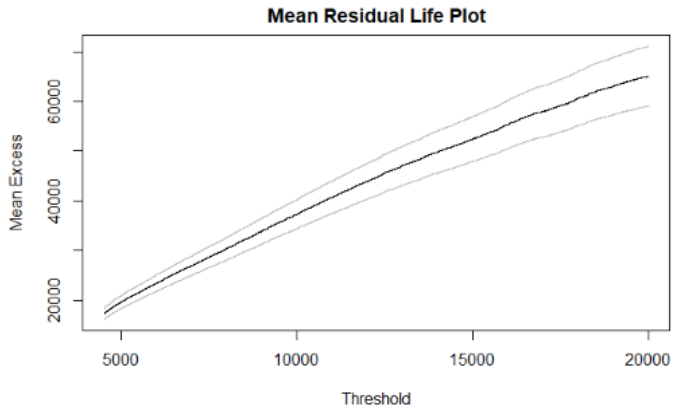


Figure 13. Empirical mean residual life plot. Plotted using mrlplot from the evd package in R. The Dashed lines representing a 95% confidence interval. Zoom in the costs below €20,000

The different results of the different limits of the mean residual life plot are shown to be poor (mainly Figure 13 results). As we approach an area where we think the identification of the boundary value will be clearer, the graph becomes completely linear, making our task more complex.

3.2.2 THRESHOLD RANGE PLOT

In this analysis, we are looking for a range of values where we can see constant and sufficiently large bars.

The first plot, Figure 14, shows some instability and a non-consistency in the way the bars grow. The results in Figure 14 (shape representation) show a higher bar size in thresholds above €15,000. This could be a good starting point and a possible value to consider.

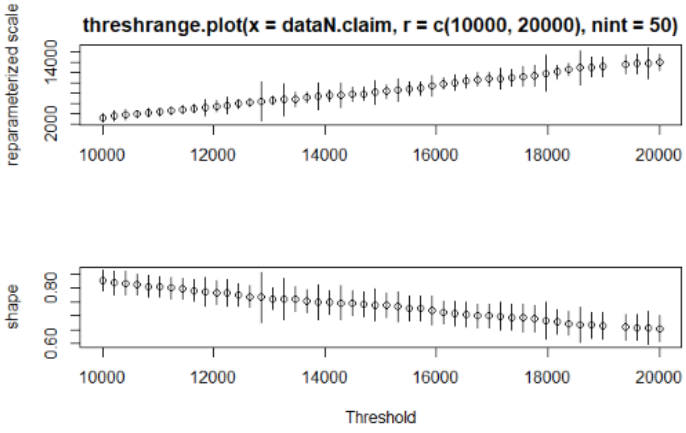


Figure 14. Threshold Selection Through Fitting Models to a Range of Thresholds. The first plot represents the reparametrized scale parameter and the second the shape parameter. Zoom on €10,000-20,000 interval

Similarly to Figure 14 results, in Figure 15 the bar size starts to grow in the right part of the plot. The biggest bars arise above the €15,000 value.

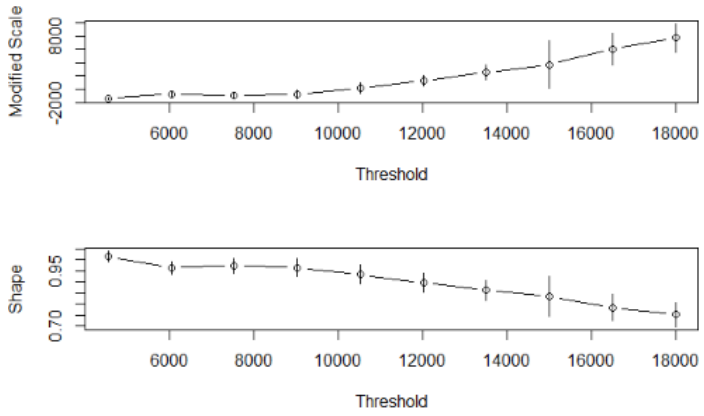


Figure 15. Threshold Selection Through Fitting Models to a Range of Thresholds. The first plot represents the reparametrized scale parameter and the second the shape parameter. Zoom on €5,000-18,000 interval

3.2.3 AUTOMATIC THRESHOLD – APPLICATION

In the automatic method, the initial value as a possible cost threshold is defined as the average value of the total costs. The highest value is set at €20,000, based on previous analyses and the company's strategic decisions. 72 values of possible thresholds will then be tested, starting with €2,085, where each one is €250 higher than the previous one (Table 5). As an alternative, 250 values will also be analysed in the same conditions, except that each one is €100 higher than the previous value.

The Kolmogorov- Smirnov (KS) test is used as a nonparametric supremum test based on the empirical cumulative distribution. Cramer-von Mises and Kolmogorov-Smirnov ones does not consider the complexity of the model (i.e., parameter number). It is not a problem when compared distributions are characterized by the same number of parameters, but it could systematically promote the selection of the more complex distributions in the other case (Hanafy, M. (2020)). 0.01 is the p value considered.

Interval (€)	N	Test	TH(€)
250	72	ks test	17,584.58
100	180	ks test	19,084.58

Table 5.Automated method- results for the first threshold

Table 5 shows the results- €17,584.58 (value closer to the 99% of our data) and €19,084.58 are the threshold candidates suggested by the automatic method. We decided to study some values below and above in order to identify some differences (based on the previous methods applied). 10,12,15,17 and 19 thousand euros turned out to be the chosen values.

The comparison with histograms and percentiles plots is done to understand if it makes sense to segment our data at the assumed points. This global view also allows us to see other possible boundary suggestions that may not have been identified in a first graphic observation. (It may be necessary to go back and explore further whether this point is plausible or not). In terms of distribution and density, no significant differences between them are detectable.

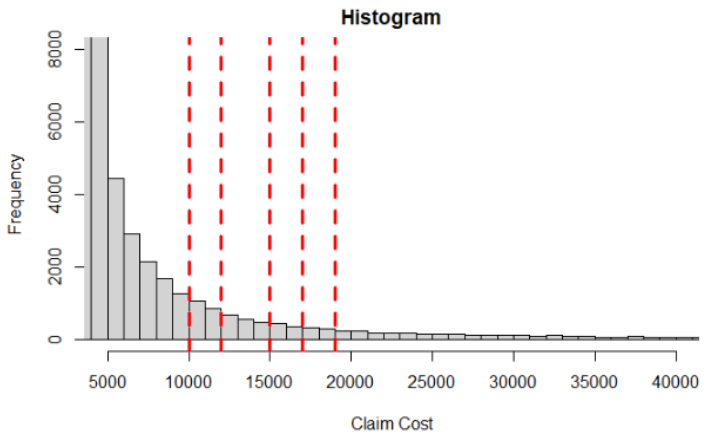


Figure 16. Histogram with threshold representations (10,12,15,17 and 19 thousand euros)

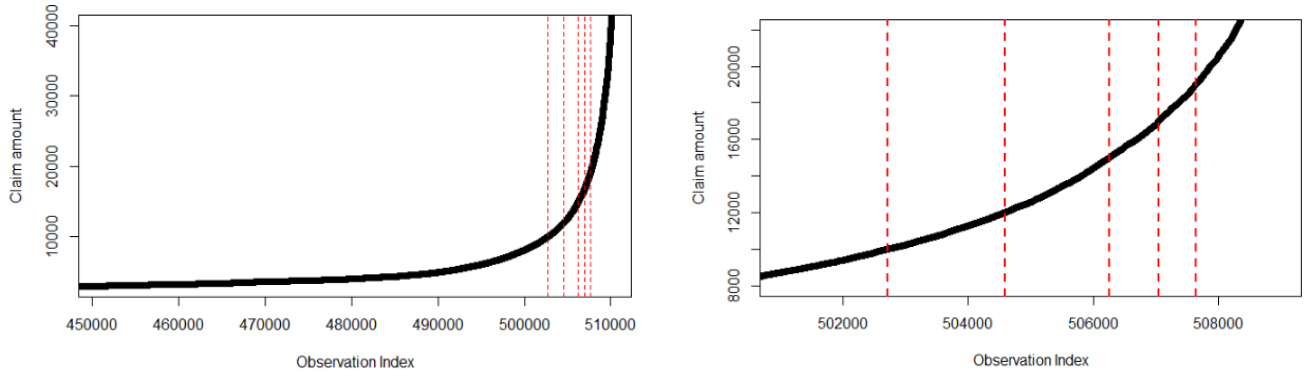


Figure 17. Percentile plot with threshold representations (10,12,15,17 and 19 thousand euros). Different zooms applied

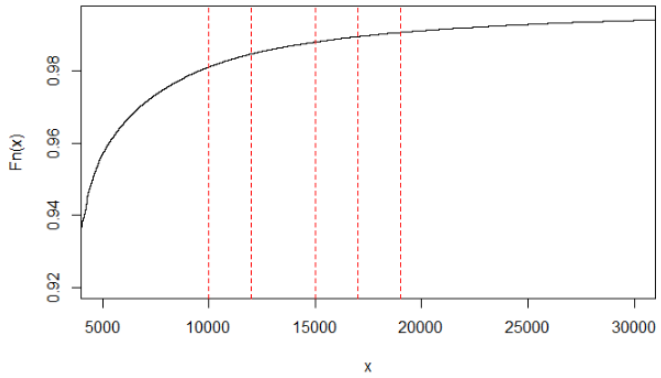


Figure 18. Empirical cumulative distribution function (ECDF) with threshold representations (10,12,15,17 and 19 thousand euros)

3.2.4 CONSIDERING A SECOND THRESHOLD

Choosing a model is not necessarily straightforward and does not have to be objectively based around which fit is theoretically better. It may occur that we have two different large claims comportments. This hypothesis should be considered not only due to low frequency of claims, but also their severity from more than a certain level. In this way, serious claims will be divided into serious claims and very serious claims.

To check this hypothesis, give the once over the final part of the dataset – the biggest costs and the smaller segment. The same methods that were used until this point will be applied in order to identify a higher threshold zone.

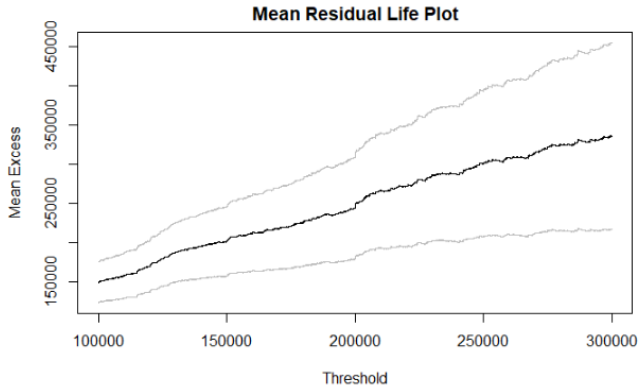


Figure 19. Mean residual life plot with zoom on €100,000-300,000 interval

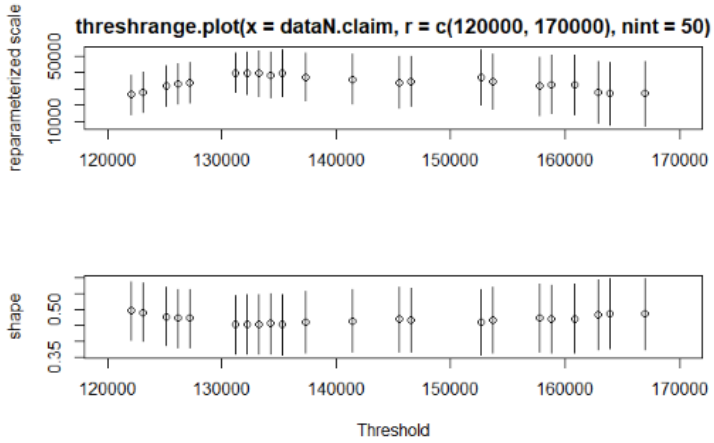


Figure 20. Threshold Selection Through Fitting Models to a Range of Thresholds. The first plot represents the reparametrized scale parameter and the second the shape parameter. Zoom on €120,000-170,000 interval

Since we do not have many candidates, in Figure 20 segment we decide to test an interval larger and where the space between the frontier candidates is chosen to be smaller (Figure 21).

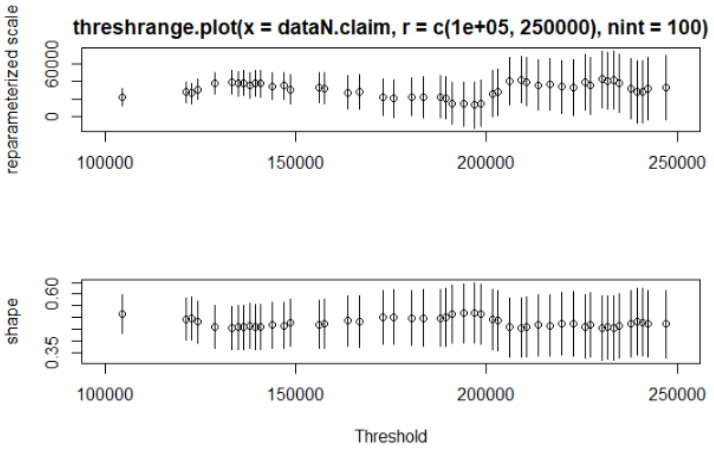


Figure 21. Threshold Selection Through Fitting Models to a Range of Thresholds. The first plot represents the reparametrized scale parameter and the second the shape parameter. Zoom on €100,000-250,000 interval

Again, it is difficult to draw strong and structured conclusions based on graphical analyses. The zoom tends to influence the analyst that bigger values are better choices.

In the automatic method, the initial value as a possible cost threshold is defined as €100,000. The highest value is set at €250,000, based on previous analyses and company's strategic decisions. 151 values of possible thresholds will then be tested, where each one is €1,000 higher than the previous one. 101 values between €100,000 and €200,000 are also tested. After the algorithm is implemented (with the ks normality tests with 0.01 p value), the result returned is €241,000. Second tentative returns a smaller value (€191,000 - Table 5). This value is closer to the 99,9% of our data.

Interval(€)	N	Test	TH(€)
1,000	151	ks test	241,000€
1,000	101	ks test	191,000€

Table 6. Automated method- results for the second threshold

Considering the results presented in Table 6, we decided to study some values below and above to identify some differences: 150, 200 and 250 thousand euros turned out to be the chosen values.

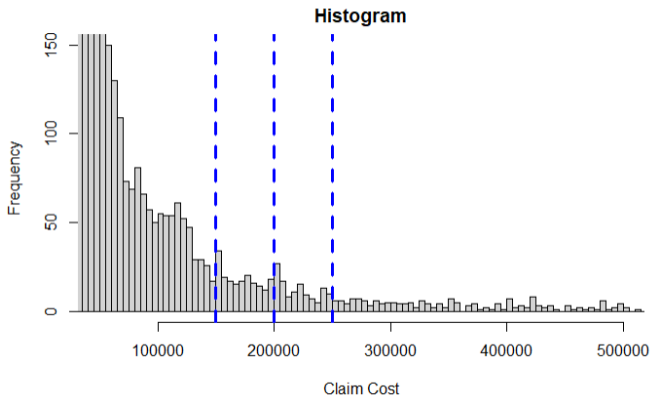


Figure 22. Representation of the possible threshold values (150,200,250 thousand euros) on the histogram

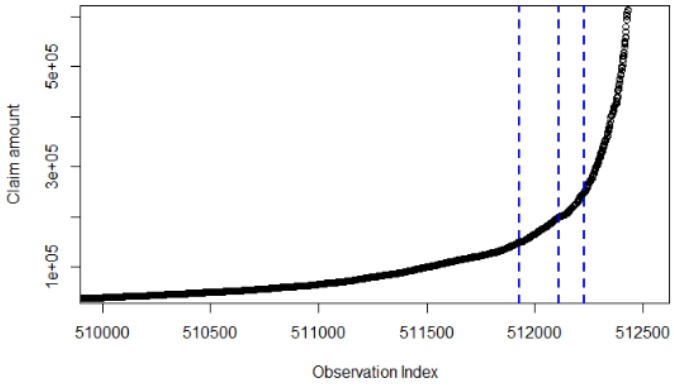


Figure 23. Representation of the possible threshold values (150, 200, 250 thousand euros) on the percentile plot

In this case (Figure 23), a change in the percentile curve is observed around the value €200,000.

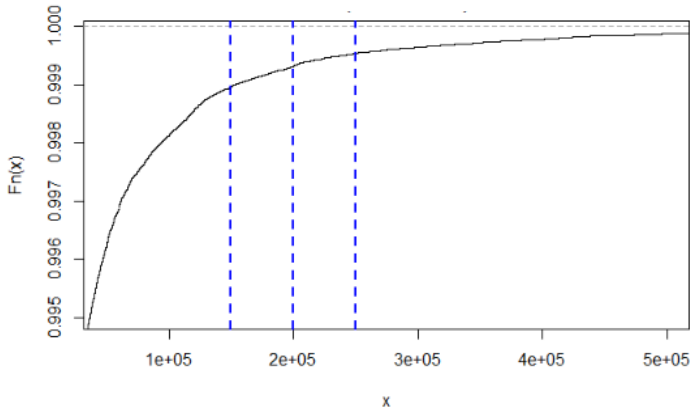


Figure 24. Empirical distribution function (ECDF) with threshold representations (150, 200 and 250 thousand euros)

In the Empirical distribution function (ECDF), a change is observed before the value €200,000, closer to the value obtained in the automatic method (Figure 24). In this set of frontier candidate values, the changes in the behaviour of the cost distribution can be identified. The value €200,000 seems to be the most adequate - justified by the combination of the different results.

3.2.5 MAIN RESULTS OF THE SECTION

This section was useful to gain structure and strong arguments to justify the final choice of the boundary's values. It allowed us to define regions where it makes sense to define the cut-off point.

Regarding the second frontier, and for greater evidence of the results, the value of €200,000 euros turns out to be chosen. As for the first frontier, from now on the objective will be to define the cut-off point: the €10,000-20,000 segment within the pre-established values (10,12,15,17 and 19 thousand euros). In the next section other methods will be used to obtain a final value.

3.3 GRADIENT BOOSTING MACHINE- METHODOLOGY

Identify characteristic factors, that may explain why some claims have such different costs, will be a great help for premiums adjustment. Based on the costs of claims, and the respective characteristics of the customer, a model is created to explain costs behaviour. The model should express which characteristics are of greater importance. In other words, it discriminates the main explanatory variables. At the end, a cost forecast is presented (based on model results).

Generalized Boosted Regression Models (GBM) will be implemented to identify certain characteristics of clients more likely to have high-cost claims. In the literature review conclusions, the usefulness of the GBM is defended and, given our work goals and robustness of the data, it is the model applied.

Within GBM, the model requires a predefined distribution of the data. The distribution to be considered should represent the weight of the data on the left side and a small amount on the right side. Gamma and pareto distribution appear to be the ones that reflect better this behaviour.

GBMs shape a sequence of trees with each tree is learning and is improving on the previous one. Starting with a week model, with only a few numbers of splits, boosting process adds new models to the ensemble successively, increasing its performance between trees.

3.3.1 Preparation for GBM creation

To create the model, it is necessary to define some hyperparameters. We implemented the following elements, presented in table 7 – called the tuning grid table.

Ntrees	number of trees to train
Max_depth	depth of each tree
Min_rows	fewest observations allowed in a terminal node
Learn_rate	rate to descend the loss function gradient

Table 7. Hyperparameters to fix – designation in the software and respective description

"Over-fitting is the tendency of the model to fit the training data too well, which means that the estimated pattern is not likely to be generalized to new data points. ", Touzani, S., Granderson, J., & Fernandes, S. (2018). In the case of a GBM model, this may happen if we use a very high number of iterations K and a too large depth of the decision trees. The question remains how to choose the right combination of hyperparameters to avoid over-fitting, and, at the same time, how to provide the best predictive accuracy.

Before starting, the model hyperparameters are optimized by a grid search process. A combination of parameters is made, the models are run and the parameters with the lowest RMSE are used. In order to classify whether the model robustly predicts claims costs, an error measure is used. In addition, to help us assess the quality of the model, the respective percentile plot is represented.

The grid search is implemented to inspect combinations of hyperparameter settings that we have already specified in a tuning grid. Since the segment below the border will not be used (the focus of the study is only the segments with costs corresponding to serious claims), the grid search methodology is not implemented in this segment. The idea is not to optimize it, but to have an overview model, reflecting general features. The following parameters are then pre-defined (based on the values used in other company studies).

GBM	**Ntrees**	**Min_rows**	**Learn_rate**
Below the threshold	5000	500	0,01

Table 8. Values used in the segment with costs bellow the first threshold the segment with costs bellow the first threshold

In the segments above the first threshold (and regardless of the final value chosen for it), 108 hyperparameter combinations are created to perform a grid search while also incorporating stopping parameters to reduce error and training time.

Learning rate	0.001; 0.01; 0.1
Minimum rows	10;50;100
Maximum depth	10;15;20
Number of trees	100;500;1,000;2,000

Table 9. Values used in combinations by parameter

H2O	Ntrees	Max_depth	Min_rows	Learn_rate
10-200k	500	20	10	0.1
12-200k	500	20	10	0.1
15-200k	500	20	10	0.1
17-200k	500	20	10	0.1
19-200k	500	20	10	0.1

Table 10. Combination of the hyperparameters for each segment that returns the smallest RMSE. The segments diverge from the value fixed in the left extreme. The right extreme is fixed in €200,000

For this segmentation, pre-definitions of borders are necessary. From the previous section results, the cost of €200,000 is already defined as a threshold. In relation to the other frontier, the lowest, the 5 values in the €10,000-20,000 segment are tested again to see which of the bounders give us a model with better behaviour. Table 10 and 11 show the best hyperparameters to be used, considering grid search results.

H2O	Ntrees	Max_depth	Min_rows	Learn_rate
>10k	500	20	10	0.1
>12k	1,000	15	10	0.1
>15k	1,000	15	10	0.1
>17k	1,000	20	10	0.1
>19k	1,000	20	10	0.1

Table 11. Combination of the hyperparameters for each segment that returns the smallest RMSE. The segments diverge from the value fixed in the left extreme (without upper limit)

The grid search process ends here. The next paragraphs describe how the target and predictor variables are chosen. Based on data available in the company, the variables that make sense to use in cost studies (based on previous work also). The set of variables to be considered as explanatory is grouped into categories in a pre-analysis of the data. This should be the first step to be performed in this section. The models will be applied in each cost segment, since different variables can have different impacts as costs are higher or lower.

The y variable is designated as SUM_of_CTSIN_RC in the code (numeric variable). This variable is considered as the sum of costs since some claims have more than one process.

Predictor variables are: Zone, Margin, Client score, Year of analysis, Years in portfolio, Instalment payments, Fuel, Entity, Number of seats, Imported vehicle, Bonus Malus, Vehicle Age, Client Sex, Years of driving license, Classification of economic activities, Vehicle category and Vehicle weight.

In terms of interpretation, we are creating a prediction model that quantifies costs given the information described above.

Before embarking on the modelling part, it is critical to split our dataset into (at least) two groups. One of those groups is called the training set. The training set should be used for the entire model building process, and it is used to perform all the model-building steps. Another group of data, called the test set, will be used to assess the performance of the model building, to compare the resulting model against the relative performance of several candidate models.

The goal here is to test the performance of the model, using data it has not seen before. Using the training data, to compare our model to any model built on different data, would give our model an unfair advantage. This out-of-sample testing allows for a truer assessment of the model's predictive power, avoiding overfitting.

In our project we decided to fix the dataset as 70% of our total dataset and the remaining 30% corresponds to the test set. The data in each set is chosen randomly.

3.3.2 COMPARISON OF MODEL PERFORMANCE BETWEEN THRESHOLD CANDIDATES- RESULTS SECTION

The conditions and preparations made before creating the model are shown. We then proceed to expose the results given in each model ran. To measure the quality of the model, four indicators are used: variable importance plot, percentile plot, an error measure (RMSE) and partial dependence plot. The variable importance plot is used relative to the variables used and it stands out their evidence in the model. Using model predictions, percentile plot and RMSE use the test set to compare the quality of the model, also testing the overfitting hypothesis. Once it is an error measure, RMSE is especially important to define the final threshold value. Partial plots of the most important variables are included, providing a graphical representation of the marginal effect of the variable in question and to have some additional conclusions.

Starting with variable importance, after re-running our final model, we likely want to understand the variables that have the largest influence on claims cost. Variable importance helps to evaluate the

quality of the model. The software used provides a plot that shows the most influential variables. It shows the relative importance as a measure, which events the average impact each variable has across all the trees. The first variable in the plot is the most important in the model, with value fixed in 1, and the impact of all other variables are provided relative to the most important variable, returning values between 0 and 1.

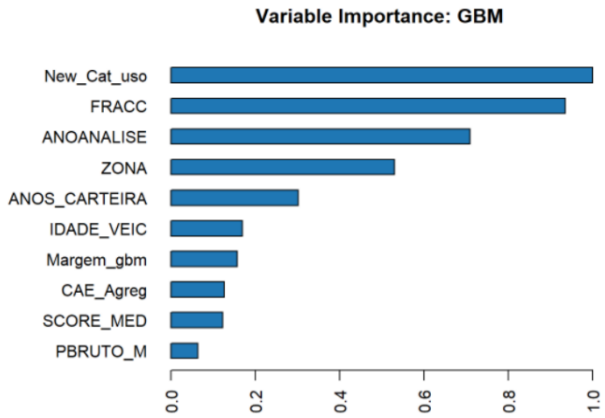


Figure 25. Variable importance plot, considering a gbm without any frontiers

If we are not considering any frontier value, vehicle category shows to be the responsible variable, followed by the instalment payments, year of analysis and accident zone.

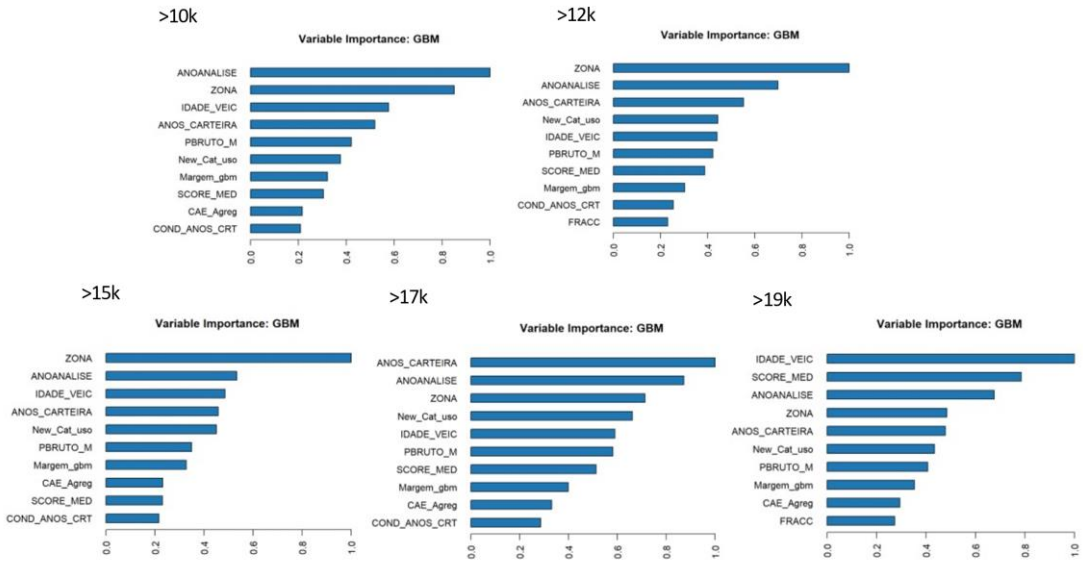


Figure 26. Variable importance plot, considering a gb model constructed considering only costs larger than the frontier candidate value, respectively

According to Figure 26, we can observe that, when using 10k,12k and 15k thresholds, ZONA and ANO ANALISE appear to be the more evident variables highlighted - they are the only ones with relative influence values greater than 60%. In the other proposed thresholds in study, these values are not so evident, being the combination of more than two variables responsible for the gb model. Above the threshold value, the results are already different from the general results. Results do not diverge that much, but the main responsible variables of the model switch position (looking only at the main ones).

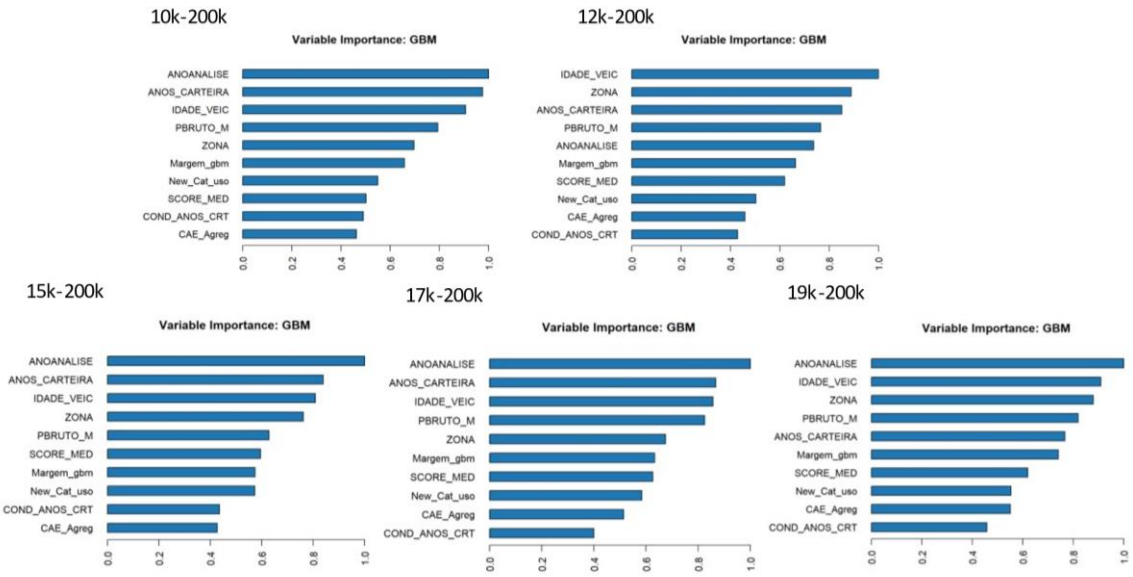


Figure 27. Variable importance plot, considering a gbm specific constructed for the segment(s) with costs larger than the frontier candidate value(s) and smaller than €200,000

Between thresholds, that is, between the first value proposed and the second frontier (fixed as €200,000) there is an increase in the volume of the importance of variables. However, the results also do not differ that much.

Overall, looking to all the variable importance plots presented in Figure 26 and Figure 27, when changing the frontier value, passing through the value 10k,12k,15k,17k and 19k, the differences between the results (in terms of variable importance) do not appear to be significant. Certainly, the error measures to consider will be a force majeure in the final frontier choice.

After studying the importance of variables, we move on to model predictions. Predictions are made based on the model. They are useful to measure the differences between what our model describes and what is the reality. In practice, we are creating new observations to evaluate if exists overfitting of the model and to test the model behaviour. This analysis is more useful if the real data considered is the data test set. Otherwise, we are testing the regression based on the observations that originated the model.

Predictions are used in this study in the percentile plot and in the RMSE. Numerically, the same number of predictions are created as the number of elements in the test set. That is, we are forecasting 30% of the initial data volume.

Percentile plot is used as a visual tool to assess the agreement between predictions and observations in different ordered percentiles of the predicted values.

The goal here is to compare the performance of gb models to the actual target variable versus the predicted target for each model. If a model fits well, then the actual and predicted target variables should follow each other closely (Goldburd, M., Khare, A., Tevet, D., & Guller, D., 2016).

In order to determine the best model, the plot should be used to check how well each model is able to predict the real claim cost in each percentile (Predictive accuracy) (Goldburd, M., Khare, A., Tevet, D., & Guller, D., 2016). Monotonicity could also be detected- monotonically would increase as the percentile increases, but the actual cost should also increase (though small reversals are okay) (Goldburd, M., Khare, A., Tevet, D., & Guller, D., 2016).

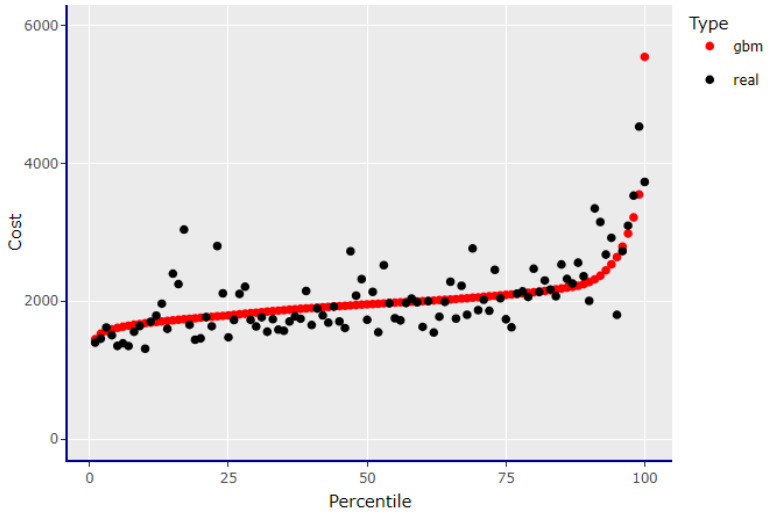


Figure 28. Percentile plot considering a gbm without any frontiers. The red dots represent the predicted values (ordinated) and the black dotes the real data (average ratio)

It is possible to observe a certain tendency in Figure 28. In fact, the real data follow the model predicted results.

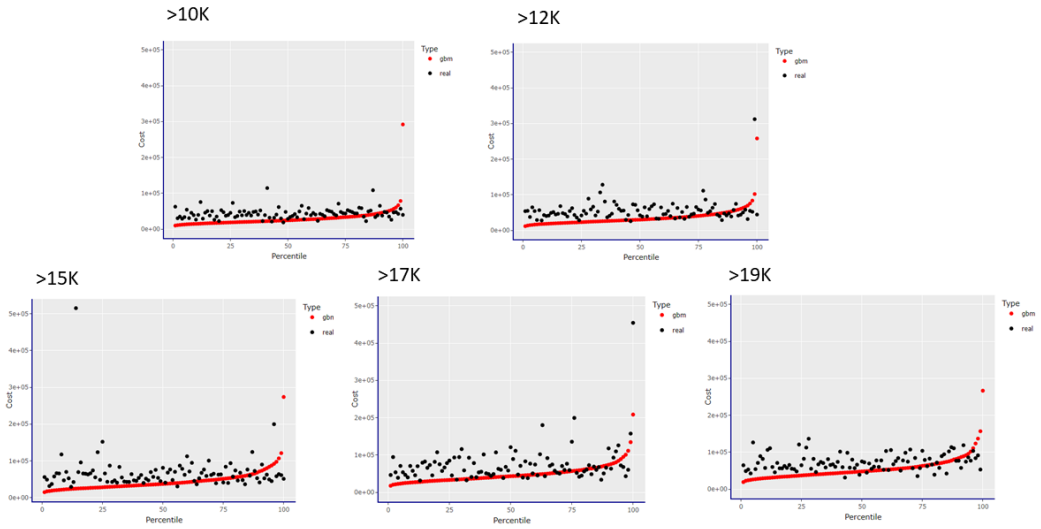


Figure 29. Percentile plot considering a gbm with a threshold. The red dots represent the predicted values (ordinated) and the black dotes the real data (average ratio)

The graphs in the figure above (Figure 29) show the model is capable to arrest the trend. Although it is a visual method with average values, it can prove that the higher frontiers better capture the trend of the data - resulting in better models.

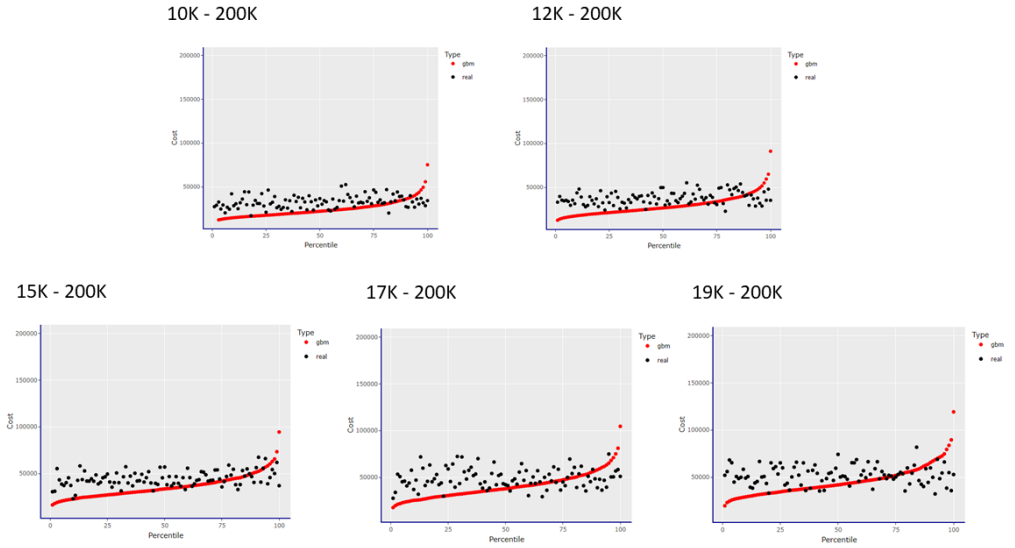


Figure 30. Percentile plot considering a two thresholds. The red dots represent the predicted values (ordinated) and the black dotes the real data (average ratio)

Looking now at the values between the two frontiers (Figure 30), analysing the five candidates for the left frontier, it becomes more difficult to identify a linear and increasing behaviour.

In fact, there is less data, and we arrive at higher and more difficult costs to predict. Once we are working with percentiles, and percentile plot represents average costs, it is not clear whether it clearly predicts the costs closer to the upper limit (€200,000).

3.3.3 FINAL THRESHOLD DECISION CRITERIA

In order to classify whether the model robustly predicts claims costs, an error measure is used. In addition, to help us assess the quality of the model, the previous percentile plot results are also considered. The error between forecasts and actual values is run and compared. The forecasts are obtained by 2 separate models: a model where only costs below the frontier value are used and another where only costs above the frontier value are used.

The error measure used is the Root mean squared error (RMSE). In fact, it is one of the usual metrics used in regression, defined as the average distance of a sample from its observed value to its predicted value. We are interested to find the lowest RMSE, between our candidates, to get the model that predicts better samples' outcomes.

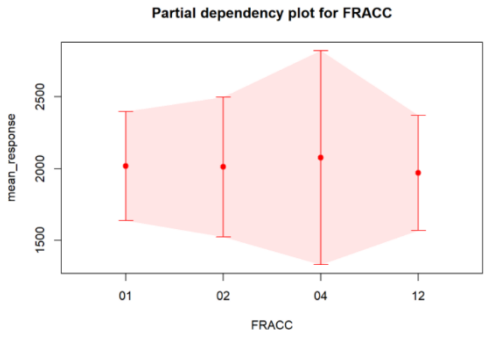
RMSE helps to prove the accuracy of the model and, comparing to other measures, it is considered less misleading when the model predicts a range of values that are far away from the general line of the observed and predicted values (Kuhn, M., & Johnson, K., 2019).

	10k	**12k**	**15k**	**17k**	**19k**
RMSE	15,632.67	15,339.37	15,525.49	15,274.05	15,443.02

Table 12. RMSE results for the different threshold candidates (10k;12k;15k;17k;19k)

As can be seen in the table presented, the smaller error appears when the threshold value is fixed in €17,000. Based on that, €17,000 will be our final choice.

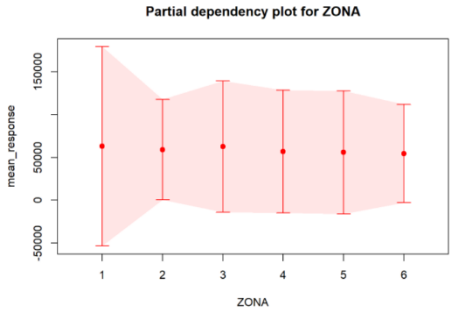
Now that the threshold value is chosen, analyse the partial dependence plot of the most important variables, in the respective model, it is done in the following lines. Partial dependence plot gives the marginal effect of a variable. It allows us to see what really influence each variable present in the model (and the respective response). The effect of a variable is also measured in change in their volatility (measured by standard deviation). The partial dependence plot helps to analyse the trend of the variable in question. Strange behaviours in this plot can judge the model. Checking the results of the most important variables is, in practice, the same as study the effects of a model ran fixing each parameter of the variable.



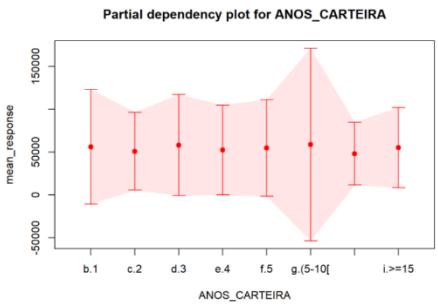
FRACC	mean_response	stddev_response
01	2017.869588	379.844875
02	2012.108490	487.722650
04	2075.526870	746.011541
12	1970.293893	401.361777

Figure 31. Partial dependence plot for instalment payments variable (FRACC) , considering a gbm without any frontiers. Mean response and standard deviation response presented in euros

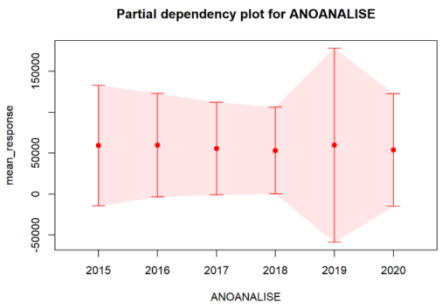
Figure 31 shows that the biggest mean response result corresponds to the quarterly payment (four payments per year) , with €2,075.53, and the smallest to the mensal option of payment (€1,970.29). In fact, the quarterly instalment payment should not be blamed for a trend towards more expensive claims. Hence the importance of dividing the dataset into segments and obtaining more correct models.



ZONA	mean_response	stddev_response
1	62944.422596	116237.793997
2	59087.108887	58522.266891
3	62742.826165	76514.436582
4	56925.424966	71646.578119
5	55843.541000	71609.339529
6	54565.730725	57427.323431



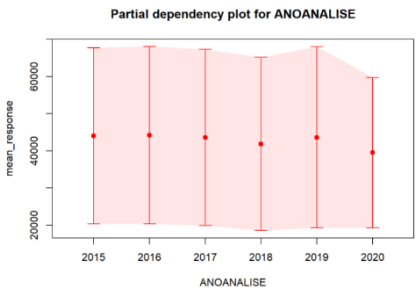
ANOS_CARTEIRA	mean_response	stddev_response
b.1	56022.935500	66952.115383
c.2	50836.628451	45319.669145
d.3	58023.296058	58915.231837
e.4	52400.122143	52230.039588
f.5	54721.668975	56496.434750
g.(5-10[	58633.263367	112236.245093
h.[10-15[	48110.674961	36606.157682
i.>=15	55088.902808	46773.299256



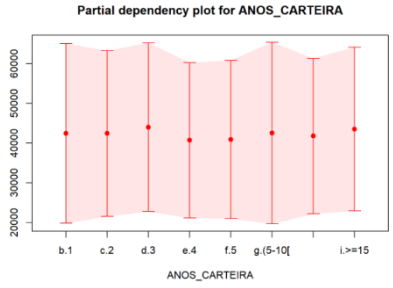
ANOANALISE	mean_response	stddev_response
2015	59252.391656	73546.543530
2016	59713.454685	63257.701618
2017	55634.845885	56556.454595
2018	53208.621484	52886.096061
2019	59730.990962	118529.770506
2020	53988.323167	68871.915455

Figure 32. Partial dependence plot for zone (ZONA), years in portfolio (ANOS_CARTEIRA) and years of analysis (ANOANALISE), respectively, considering a gbm for the costs larger than €17,000. Mean response and standard deviation response presented in euros

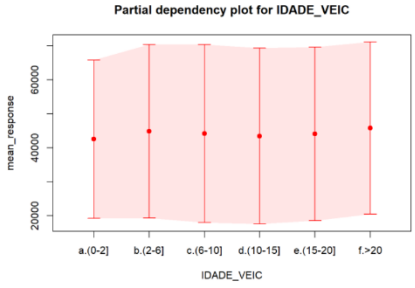
Zone 1 has the highest variance (€116,237.79) and the highest mean response value (€62,944.42) of the model. Although it is not linear, there is a decreasing trend in mean response cost per zone as the number associated with each zone increases. Regarding the number of years that the customer has been in the company, the variance is higher in the 5-10 segment (€112,236.24) and, outside of this segment, customers who have been with the company for less time appear to be responsible for higher cost claims. Finally, in relation to the year of occurrence of the accident, we observe a general decreasing trend, except for the year 2019 (€59,730.99 mean response) compared to 2020 due mainly to COVID-19 and less vehicles exposure.



ANOANALISE	mean_response	stddev_response	s
2015	44066.450055	23697.705942	
2016	44190.224248	23871.574360	
2017	43600.335504	23712.099886	
2018	41868.945419	23299.738191	
2019	43635.767997	24343.374401	
2020	39524.228161	20243.669007	



ANOS_CARTEIRA	mean_response	stddev_response	s
b.1	42476.281031	22573.817613	
c.2	42430.269932	20826.966712	
d.3	44015.950378	21200.608926	
e.4	40704.786966	19539.915166	
f.5	40901.041895	19877.754702	
g.(5-10[	42503.967312	22818.632034	
h.[10-15[	41761.393085	19557.691996	
i.>=15	43521.688151	20589.440156	



IDADE_VEIC	mean_response	stddev_response
a.(0-2]	42492.297884	23306.825459
b.(2-6]	44803.953166	25545.064042
c.(6-10]	44158.996853	26158.236435
d.(10-15]	43423.265308	25856.052092
e.(15-20]	44058.225959	25492.762418
f.>20	45735.566302	25331.561067

Figure 33. Partial dependence plot for years of analysis (ANOANALISE), years in portfolio (ANOS_CARTEIRA) and vehicle age (IDADE_VEIC), respectively, considering a gbm for the costs bigger than €17,000 and smaller than €200,000. Mean response and standard deviation response presented in euros

Focusing now on the results of the 17-200k segment (Figure 33), in general, less divergence between the average values (within the variable in question). A balanced year of analysis is obtained, although 2019 (€43,635.77 mean response) continues to have a stronger presence, especially compared to 2020 (€39,524.23 mean response). In terms of years in portfolio analysis, those with less customer time continue to have a significant importance in higher costs. In this segment, long-term clients are also more liable (due to the tendency to be more likely to had at least one claim). It is also observed a fairly balance in age of the vehicle analysis.

3.3.4 GBM FOR THE SEGMENT ABOVE SECOND THRESHOLD

In this segment, some measures are taken in advance of the model construction. This action rises from the fact that there are only 360 observations in this segment. The randomness with which these claims happen also leads us to be more careful in this segment.

Variable New_Cat_uso is transformed and renamed to New_Cat_Mud. The purpose is to make an aggregation of the factors of the variable: the use of the vehicle is removed, becoming only a category variable, and the M4 and MC factors are grouped. The variable NUM_LUG is changed since we do not have any claim in the class 10-25 seats. Then, more than nine seats class (>9) is created. The variable name is kept. To avoid overfitting, grid search is done in an identical way to what was done for the other segments.

Learning rate	0.001; 0.01; 0.1
Minimum rows	3;5;10;15
Maximum depth	3;5;10;15
Number of trees	20;50;100

Table 13.Values used in combinations by parameter, for €200,000 segment. 144 combinations are used

GBM	Ntrees	Max_depth	Min_rows	Learn_rate
>200K Segment	100	15	3	0.1

Table 14. Combination of the hyperparameters for >200k segment that returns the smallest RMSE

Due to the limited amount of data in this segment, division into data train and data test is not used. Instead of data train and test division, cross validation is used. data groups are formed, designed by folds (k folds). The procedure used is simple. Considering only one of the folds, the model is trained using the remaining folds. Then, the test of the model is done using the first fold. This step is repeated for each of the remaining. We decided to use 5 folds. The model error of the fit is estimated with the hold-out sets and this is the final model used.

Some variables in study may not make sense in this segment. Depending on the results that the model returns, we decide which variables can be useful or not. For that, with the help of partial analysis, variables are studied on a case-by-case basis to see what extent their impact on the model makes sense to be kept in the model.

Similar to the previous segments study, the partial effects of the most significant variables in the model are studied (here we assume the first three). If its behaviour does not make sense, not following any trend, the variable is removed, and a new model is run. This methodology is used successively

until we obtain the final model in which the most important variables have an adequate cost evolution behaviour.

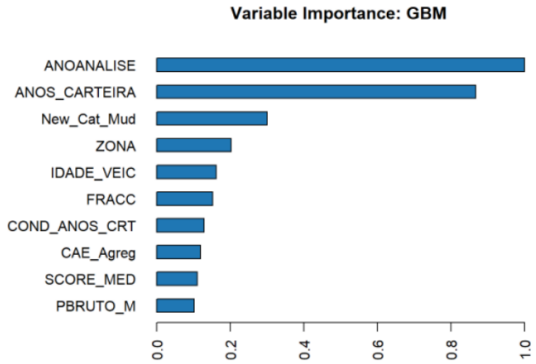
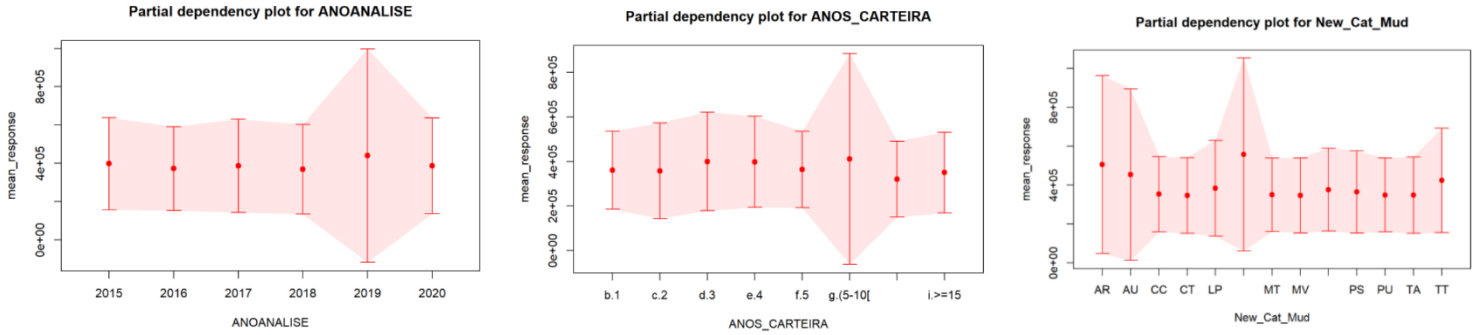


Figure 34. Variable importance plot, considering a gbm for the segment >200k



ANOANALISE	mean_response	stddev_response
2015	397863.201998	241193.370563
2016	371700.650895	219254.524639
2017	386511.654224	244146.092889
2018	368917.573810	234160.316080
2019	439490.559247	556995.840137
2020	386756.979542	250118.561711

ANOS_CARTEIRA	mean_response	stddev_response
b.1	360415.704647	175301.676658
c.2	357769.416597	215362.758399
d.3	399871.996142	221399.567786
e.4	398065.354317	205367.381850
f.5	363709.237428	171487.628627
g.(5-10[	410484.635902	473684.380063
h.[10-15[	319320.524171	170221.401548
i.>=15	349625.466322	181753.414089

New_Cat_Mud	mean_response	stddev_response
AR	505907.445058	456994.404458
AU	453965.813748	441006.574319
CC	353433.530147	193072.674557
CT	346189.474230	194222.072580
LP	383625.905722	246311.492573
M4+MC	557561.589576	495552.386830
MT	350645.432242	188326.503201
MV	347067.710006	192578.633327
OTHER	375742.620789	212776.123890
PS	364933.808402	210608.086542
PU	349070.249286	190182.451146
TA	348549.958286	195572.983115
TT	424597.541038	268378.516345

Figure 35. Partial dependence plot for years of analisis (ANOANALISE), years in portfolio (ANOS_CARTEIRA) and category (New_Cat_Mud), respectively, considering a gbm for the costs larger than €200,000. Mean response and standard deviation response presented in euros

Based on Figure 35, we can settle a huge divergence from the year 2019 and the year analysis variable is removed. The absence of an increasing or decreasing trend as the period of time the customer is with the company increases, and above all due to the divergent results of customers with 5 to 10 years of Fidelidade insurance, makes us end up discarding the variable years of Wallet. In the grouped vehicle category (New_Cat_Mud), the ARTICULATED (AR), MOTORCYCLE (M4+MC) AND BUS categories (AU) stand out. The variable is not removed.

Following will be presented the results of a new gbm model (model ran without year of analysis and years in portfolio).

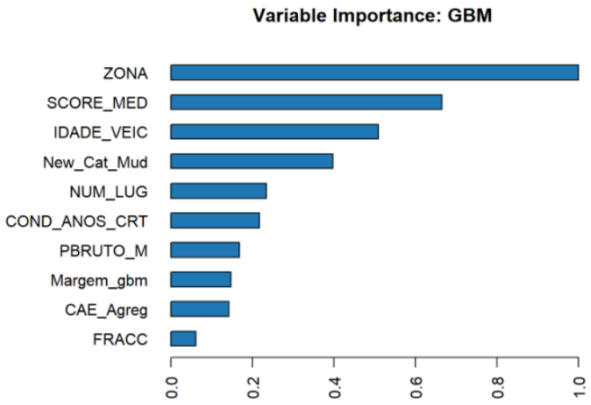
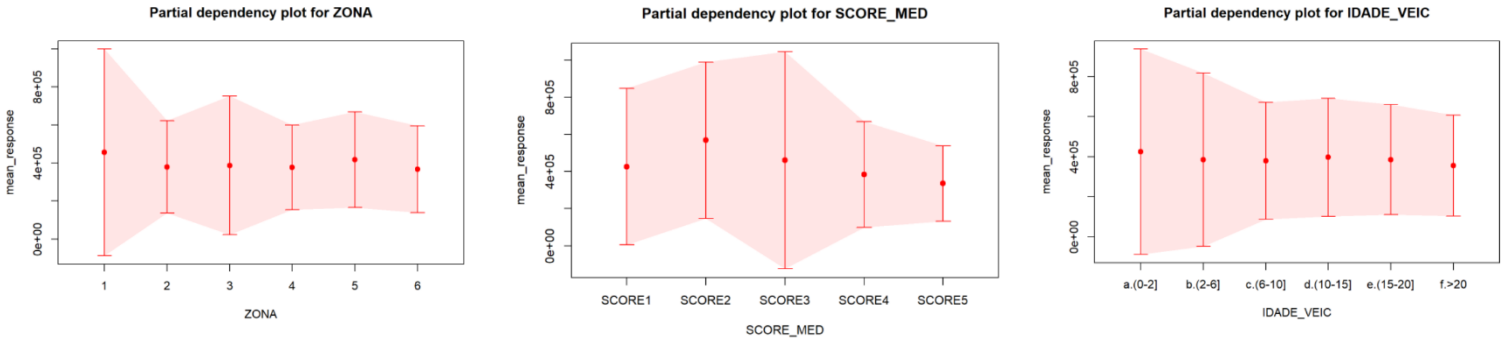


Figure 36. Variable importance plot, considering a gbm for the segment >200k without the following variables: year of analysis (ANOANALISE) and years in portfolio (ANOS_CARTEIRA)



ZONA	mean_response	stddev_response
1	455626.369563	543805.836161
2	379668.976121	242673.477700
3	387356.287150	365626.496952
4	376983.374748	222974.225895
5	417623.847772	250957.316672
6	367519.445303	228329.474334

SCORE_MED	mean_response	stddev_response
SCORE1	426192.179894	420656.634440
SCORE2	568123.111329	421470.141969
SCORE3	460256.510820	584136.916374
SCORE4	384594.856118	284569.733185
SCORE5	335383.518223	202961.754483

IDADE_VEIC	mean_response	stddev_response
a.(0-2]	424938.884404	512891.088437
b.(2-6]	385293.594551	433566.005213
c.(6-10]	379315.595972	292341.875710
d.(10-15]	397575.426389	295405.144422
e.(15-20]	386138.552766	274769.912333
f.>20	356104.126840	251287.044873

Figure 37. Partial dependence plot for zone (ZONA), client score (SCORE_MED) and vehicle age (IDADE_VEIC), respectively, considering a gbm for the costs larger than €200,000, without the following variables: year of analysis (ANOANALISE) and years in portfolio (ANOS_CARTEIRA). Mean response and standard deviation response presented in euros

For the same reasons of instability of the factors considered in each variable, ZONA and SCORE_MED are also removed. The vehicle age variable shows a decreasing trend, except for vehicles between 10 and 15 years old (€397,575.43 mean response). The differences between mean responses are not that significant in this variable, so it ends up not being removed from the model.

We move on to the next analysis of models without the 4 variables that were removed.

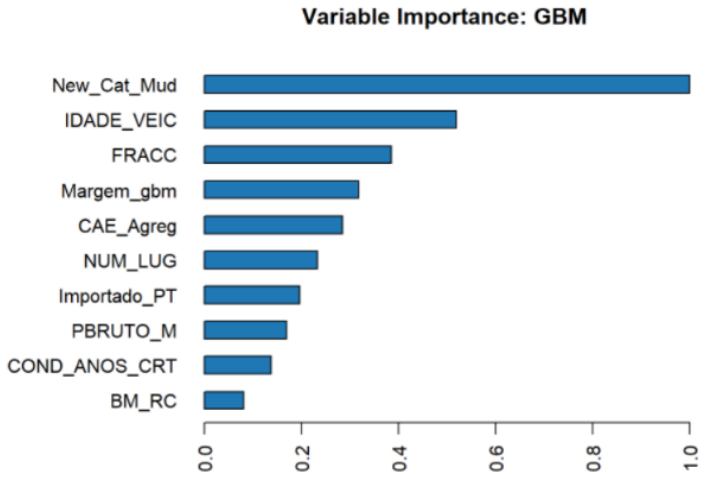
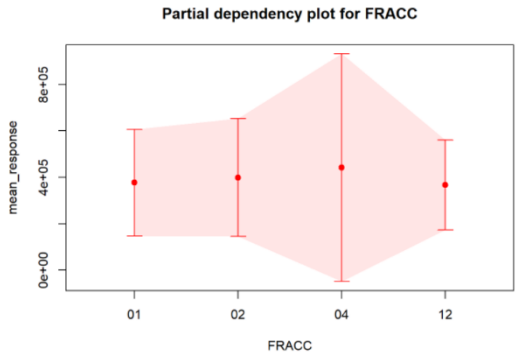
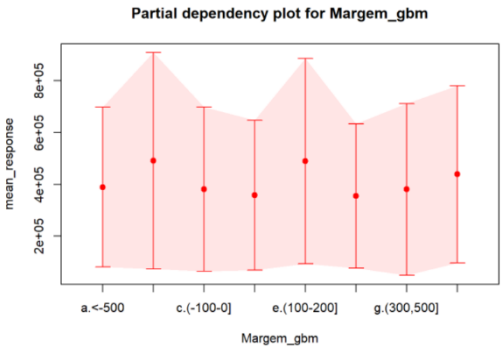


Figure 38. Variable importance plot, considering a gbm for the segment >200k without the following variables: year of analysis (ANOANALISE), years in portfolio (ANOS_CARTEIRA) , zone (ZONA) and score (SCORE_MED)



FRACC	mean_response	stddev_response
01	376756.264542	229209.306460
02	398946.418846	253965.228752
04	441329.422237	489651.056459
12	366696.427287	194040.430921



Margem_gbm	mean_response	stddev_response	st
a.<-500	389298.166443	307690.462942	
b.(-500-100]	490897.934252	417249.745795	
c.(-100-0]	380963.938670	316409.629276	
d.(0-100]	358578.098386	288730.326808	
e.(100-200]	490073.087406	396268.593133	
f.(200-300]	355561.674028	278152.770488	
g.(300,500]	380386.085230	331464.169478	
h.>500	438890.098617	341953.853400	

Figure 39. Partial dependence plot for instalment payments (FRACC) and gbm margin (Margem_gbm), respectively, considering a gbm for the costs larger than €200,000, without the following variables: year of analysis (ANOANALISE), years in portfolio (ANOS_CARTEIRA), zone (ZONA) and score (SCORE_MED). Mean response and standard deviation response presented in euros

The first two variables highlighted in terms of importance have already been studied previously. The analysis of FRACC and Margem_gbm is performed here (Figure 39). Looking to instalment payments results, the highest costs are seen in customers who pay the premium quarterly. As a rule, the higher the fractionation, the higher the average cost. This is not the case and therefore the variable ends up being removed. Regarding the margin, the fluctuations as the value increases are many. For this reason, the variable is also removed.

Figure 40 and 41 show the results of one more model regression created, this time without year of analysis (ANOANALISE) and years in portfolio (ANOS_CARTEIRA), zone (ZONA), score (SCORE_MED), instalment payments (FRACC) and margin (MARGEM_gbm)

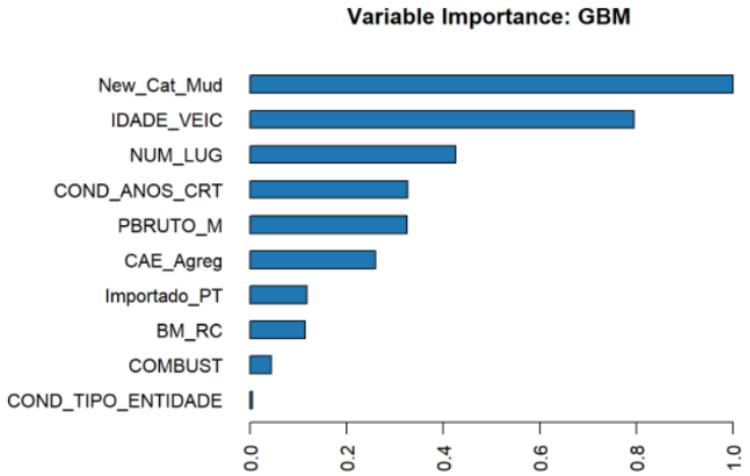
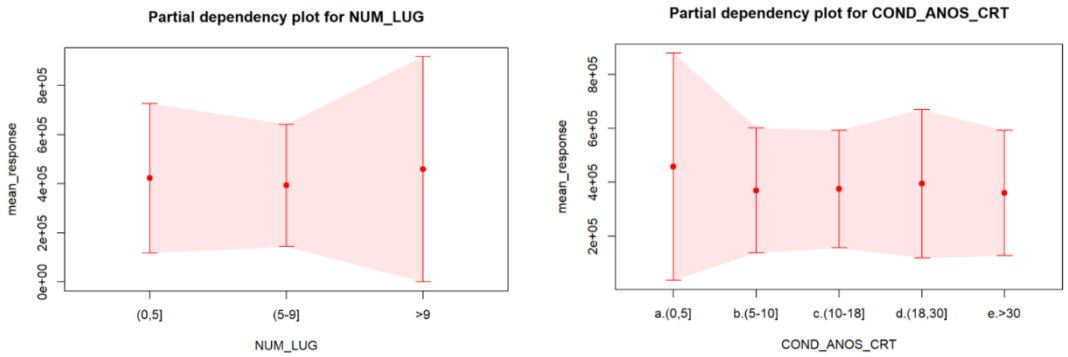


Figure 40. Variable importance plot, considering a gbm for the segment >200k without the following variables: year of analysis (ANOANALISE) and years in portfolio (ANOS_CARTEIRA), zone (ZONA), score (SCORE_MED), installment payments (FRACC) and margin




```
NUM_LUG mean_response stddev_response
(0,5] 422618.127248   304765.180563
(5-9] 393330.512230   248913.163921
   >9 458418.150487   458660.128387
```

```
COND_ANOS_CRT mean_response stddev_response
a.(0,5] 457050.738988   421316.561377
b.(5-10] 369449.288899   231623.871011
c.(10-18] 374689.258963   218321.394376
d.(18,30] 393816.596696   275141.989046
e.>30 360567.881953   232568.659371
```

Figure 41. Partial dependence plot for number of seats (NUM_LUG) and years of driving license (COND_ANOS_CRT), respectively, considering a gbm for the costs larger than €200,000, without the following variables: year of analysis (ANOANALISE), years in portfolio (ANOS_CARTEIRA), zone (ZONA), score (SCORE_MED), instalment payments (FRACC) and margin (MARGEM_gbm). Mean response and standard deviation response presented in euros

Vehicle category and vehicle age variables, and the respective partial dependence plots, have already been studied. We have the results of the number of seats in figure 41. An increasing gbm mean response cost trend is observed in vehicles with more seats. On the other side, years of driver license plot shows the newly enrolled ones have the highest average value and the highest variance of expected costs- the tendency being lower in the other classes. At the age of children or with more exposure to driving, the values increase a little, but remain stable. For these reasons the variable is not removed either. Given this framework in which the main variables make sense, we ended up considering this the final model.

After the experiments and the sequence of successively removing variables, we move on to the final analysis of the segment with costs bigger than €200,000. The predictions of the final considered model are produced and compared with the original data. In this segment, not percentile plots but decile plots are used (due to data limitation in this segment, again).

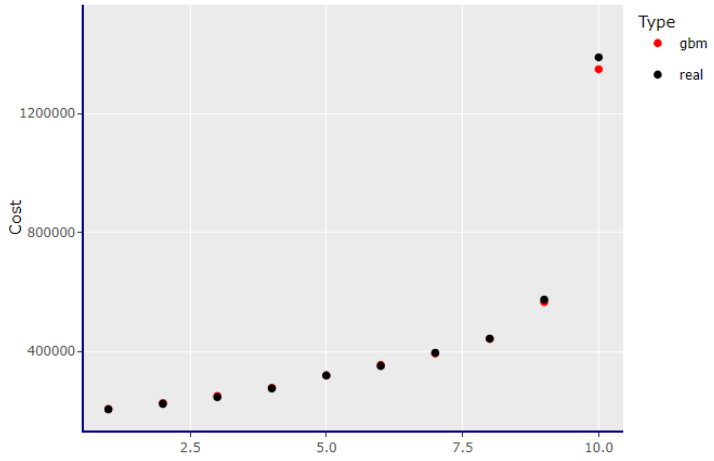


Figure 42. Decile plot considering a gbm for the final variables set, considering costs superior to €200,000. The red dots represent the predicted values (ordinated) and the black dotes the real data (average ratio)

3.3.5 MAIN RESULTS OF THIS SECTION

Without frontier	Considering only 17k frontier – segment above 17k	Considering 2 frontiers- 17k-200k segment	Considering 2 frontier->200k segment
Vehicle Category	Years in portfolio	Year of analysis	Vehicle Category
Installment payments	Year of analysis	Years in portfolio	Vehicle age
Year of analysis	Zone	Vehicle age	Number of seats
Zone	Vehicle Category	Vehicle weight	years of driving license

Table 15. Main important variables for each segment

Good preparation is essential for creating an efficient gb model. The creation and analysis of the model for each candidate for the border and the respective error measure causes the value of €17,000 to be defined as the final choice of border.

Table 15 represents the four main responsible variables in each gbm, created for each segment in analysis. From the results of the analysis of the data without any frontier, we can see that the instalment payments appear as a variable of high importance. This is confirmed to be a false indicator reenforced by the partial plot results. Hence the importance to create thresholds.

In all analyses, the vehicle category is always very present. The variables year analysis and year portfolio greatly influence 17k-200k segment and the segment with costs above 17k. Comparing the results of these two analyses, the results of the model run with the data above the frontier are hampered by the higher costs - hence the importance of the existence not only one but two fixed thresholds.

4 PROBABILITY MODELS AND RISK PREMIUM

As described in the introduction section, this chapter aims to compare the actual data to the predicted ones and still obtain probabilities of large claims. This information is complementary to the work developed in the previous chapter and it allows us to draw some more interesting conclusions.

After evaluating the results of the created models, the average cost forecasts are obtained in each costs interval considered.

The average cost values have the following interpretation: if a claim falls within a certain cost segment, its average cost is estimated (in euros).

Table 16 shows the real claim costs and predicted claim costs for each segment defined and studied in chapter 3. It is observed that the average costs of each segment predicted by the model are slightly lower than the real costs.

	<17k	**17k-200k**	**>200k**
Real mean cost (€)	1312.4	48,601.63	442,717.4
Predicted mean cost (€)	1308.9	40,682.82	438,891.75

Table 16. Comparison of real claim costs and predicted claim costs (using chapter 3 gb models). Mean costs in each segment studied

To finalize the study, the probability of a claim being in a certain range of costs (specific defined segments) is calculated. The dependent variable to be used is not the claim costs, but instead a binary variable. The variable assumes the value 0 if the cost of the claim is not in the segment in case and the value 1 if the claim is in that segment. A Bernoulli distribution is supposed, and the hyperparameters used are shown in table 17.

GBM	**Number of trees**	**Max Depth of each tree**	**Learning rate**
Probability models	3,000	3	0,01

Table 17. Hyperparameters values used in probability models

The results are analysed and presented in terms of the most important variables, in the same way that it was done in the previous chapter. Additionally, the decile plot between the observed costs and the costs predicted by the model are presented.

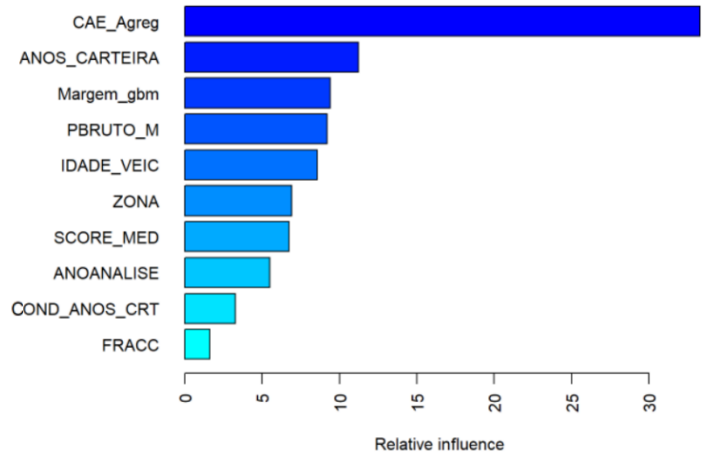


Figure 43. Variable importance plot, considering a probability gbm for the segment <17k

In figures 43 to 45 we have the results of the variables that seem to have greater importance in the models of probability of a claim having costs that fit into the segment under study. In the first segment, the variables that assume more importance in the model are classification of economic activities, years in portfolio and margin.

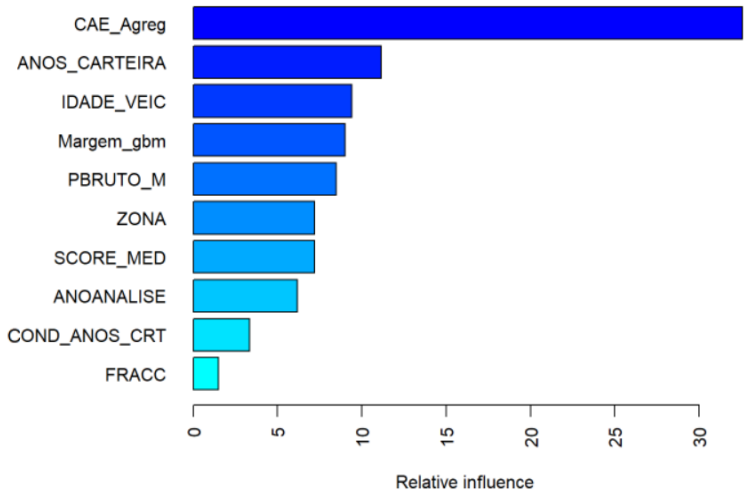


Figure 44. Variable importance plot, considering a probability gbm for the segment 17-200k

In the segment corresponding to claims costs between €17,000 and €200,000, classification of economic activities, years in portfolio and vehicle age have more representativeness in the explanation of the model. Finally, when estimating the probability of a claim costing more than €200,000 the variables classification of economic activities, years of drive license and vehicle age are the ones that appear to have greater responsibility for the biggest losses.

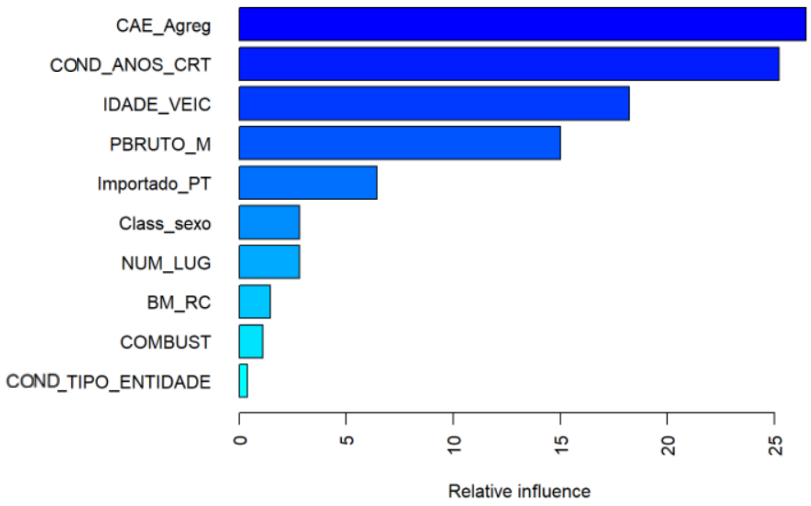


Figure 45. Variable importance plot, considering a probability gbm for the segment >200k

Based on the predicted values in each of the three models (one for each cost segment), a normalization of the probabilities obtained in each of the models is performed. These are the final predicted values to be compared with the actual values (both on the decile plot and in terms of average probability, versus average ratio).

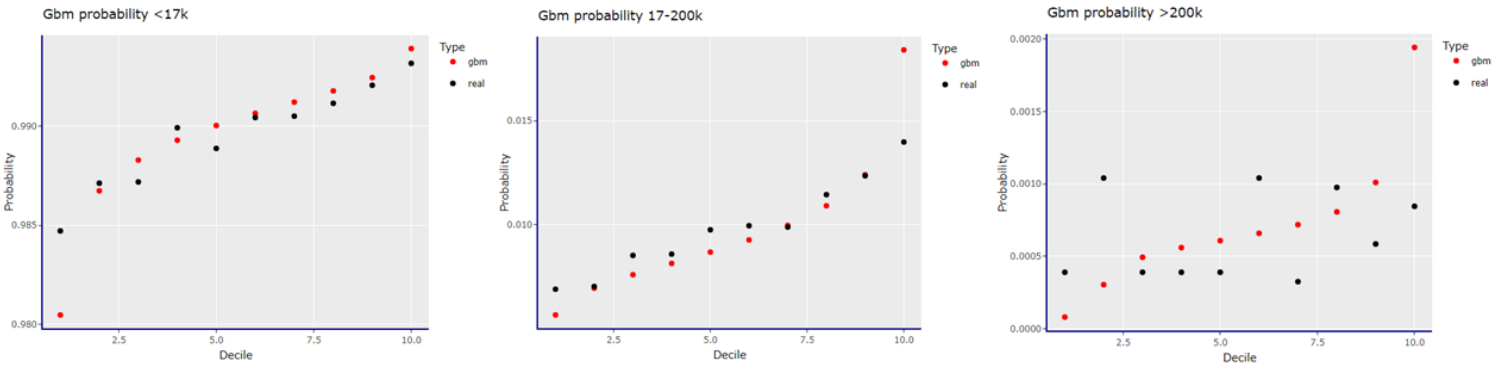


Figure 46. Decile plots considering the 3 probability gb models. The red dots represent the predicted values (ordinated) and the black dotes the real data (average ratio)

In the first and second segments, there is a trend from actual to predicted data. In the cost segment above €200,000 the trend is not so linear. The fact that there is less data in this segment (in proportion to the number of total data) influences this result. However, despite being the least assertive model, it continues to follow the model.

	<17k	**17k-200k**	**>200k**
Ratio	98.9491%	0.9806%	0.0703%
Predicted probabilities	98.9482%	0.9799%	0.0719%

Table 18. Comparison of real costs proportion and predicted probabilities. Values calculated in each segment studied

Applying the model to another future dataset, the probability values specify the probability of a motor insurance claim costing less than €17,000 is 98.95%, being between €17,000 and €200,000 is 0.98 % and being greater than €200,000 is 0.072%. The observed ratios are identical to the predicted probabilities.

In the end, the risk premium is calculated, considering the sum of the risk premium of the three segments, using the formula number 7. For the real data, the frequency considered is the real ratio (Table 18). For the predicted data, predicted probabilities values are used.

The risk premium is the expected amount of claims costs per unit of risk, in this case, per vehicle.

$$Risk\ premium\ = \sum Frequency\ x\ Mean\ Cost \qquad [7]$$

Due mainly to the influence of the predicted costs being lower than those observed, the final calculated value is lower than the observed by €77.04.

Real (€)	**Predicted (€)**
2,086.25	2,009.21

Table 19. Comparison of real risk premium and predicted risk premium

5 CONCLUSIONS

This work enables to describe what should or not should be treated as a large claim - values that are too high, with a more significant level of variance than the other claims costs. Large claims are identified in this work as costs that are above a certain value - the frontier value.

Threshold zone selection methods are used in this project. Thresholds could be identified using a methodology that combines the graphical methods (mean residual life plot and threshold range plot) and the automatic method, creating a set of possible boundaries. This is the methodology used, helping in the identification of possible thresholds.

On the first results, it is identified the range of values that makes sense to define the final bounder. Moreover, due to the variance and distribution of the data, two border zones are defined, the severe and very severe zones. It is also concluded that, in insurance, it makes sense to have borders, and to analyse factors of relevance in each segment.

Once the threshold zones are identified, regression models are built using the Gradient Boost machine learning technique. The variables that try to explain the context of certain costs are analysed. The variable importance analysis allows us to conclude with greater certainty the need of thresholds and segments in claims costs models analysis.

Based on the costs of claims, and the respective characteristics of the customer and his vehicle, gb models enable the identification of characteristic factors that could explain costs behaviour.

Model metrics were used to define thresholds values- proper error measure (RMSE) and analysis of percentile and partial plots and we concluded that €17,000 is a threshold cost value to be used. Concluding, in the three cost models developed, vehicle category is always very present as the main responsible variable followed by the variable's year analysis and year portfolio. This are the three variables that mainly explain large claims costs behaviour. Comparing the results of the 17,000-€200,000 segment with higher than €200,000 segments, the results of the model run with the data above the frontier are hampered by the higher costs – hence again the importance to define two fixed thresholds.

Cost models are always difficult to predict, but our average responses to the chosen segments and adopted measures appear to be good.

Probability models developed and the classification of economic activities, years of drive license and vehicle age are the variables that explain the biggest losses probability. Combining all the models developed, a risk premium of €2009.21 is obtained (a value very close to the real average cost).

Pre-model construction and post-model-built and adapted quality analysis was done in this work, and it is probable that they will have high predictive power in out-of-sample data. With this, it is expected that the final models can therefore be applied in future company's database.

As future work, it is suggested to test other machine learning techniques and compare other models to the presented results in this document. As a contribution to the insurance activity, it would be interesting to develop an identical work for another motor insurance portfolio or even another insurance sector branch.

6 BIBLIOGRAPHY

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. The Annals of Applied Statistics, 12(1), 310-329.

Balkema, A., and De Haan, L. (1974). "Residual life time at great age", Annals of Probability, 2, 792–804.

Behrens, C. N., Lopes, H. F., & Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. Statistical Modelling, 4(3), 227-244.

Charpentier, A. (Ed.). (2014). *Computational actuarial science with R*. CRC press.

Click, C., Malohlava, M., Candel, A., Roark, H., & Parmar, V. (2017). Gradient boosting machine with h2o. H2O. ai.

Drakenward, E., & Zhao, E. (2020). Modeling risk and price of all risk insurances with General Linear Models.

Ferreira, A., & De Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. The Annals of statistics, 43(1), 276-298.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., ... & Zaliapin, I. (2011). Extreme events: dynamics, statistics and prediction. Nonlinear Processes in Geophysics, 18(3), 295-350.

Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized linear models for insurance rating.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Systems with Applications, 39(3), 3659-3667.

Hall, P., & Proksch, M. (2020) From GLM to GBM, retrieved from https://towardsdatascience.com/from-glm-to-gbm-5ff7dbdd7e2f. 20 December 2021

Hanafy, M. (2020) Application of Generalized Pareto in Non-Life Insurance. Journal of Financial Risk Management, 9, 334-353.

Hu, Y. (2013). Extreme value mixture modelling with simulation study and applications in finance and insurance.

Hull, J. C. (2018). Risk Management and Financial Institutions. Fifth Edition, Wiley.

Kim, A. (2019). Gamma Distribution—Intuition, Derivation, and Examples. Retrieved October, 7, 2020.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

Lee, W. C. (2009). Applying Generalized Pareto Distribution to the Risk Management of Commerce Fire Insurance. Preliminary draft, Department of Banking and Finance Tamkang University, Taiwan.

Lu, H., Karimireddy, S. P., Ponomareva, N., & Mirrokni, V. (2020, June). Accelerating gradient boosting machines. In International Conference on Artificial Intelligence and Statistics (pp. 516-526). PMLR.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bulletin: The Journal of the IAA, 27(1), 117-137.

Nogueira, F. A. V. (2015). Tarifação em não-vida com um MLG, aplicação prática com a ferramenta R (Doctoral dissertation, Instituto Superior de Economia e Gestão).

Paldynski, H. (2015). Modelling Large Claims in Property and Home Insurance-Extreme Value Analysis.

Pauli, F., & Coles, S. (2001). Penalized likelihood inference in extreme value analyses. Journal of Applied Statistics, 28(5), 547-560.

Pešta, M., Petrová, B., Procházka, J., & Smolárová, T. EXERCISES FOR NON-LIFE INSURANCE.

Pickands, J. (1975). "Statistical inference using extreme order statistics", Annals of Statistics, 3, 119–131.

Pinheiro, M., & Grotjahn, R. (2015). An introduction to extreme value statistics.

Ren, F., & Giles, D. (2007). Extreme value analysis of daily Canadian crude oil. Econometrics Working Paper EWP0708, ISSN, 1485-6441.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. Update, 1(1), 2007.

Rydman, M. (2018). Application of the Peaks-Over-Threshold Method on Insurance Data.

Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. REVSTAT–Statistical Journal, 10(1), 33-60.

Uwingabire, J. (2018). Modelling extreme health insurance claims using generalized Pareto distribution.

7 ANNEXES

VARIABLES DESCRIPTION

Code	Variable	Description
SUM_of_CTSIN_RC	Sum of the costs of each claim	Cost of each claim. This variable is considered as the sum of costs since some claims have several processes open to resolve them.
ZONA	Zone	Classification of Portuguese zones between 1 and 6, where 1 is the area with the highest losses.
Margem_ gbm	Gbm margin	Margin assigned according to the gbm risk model already used by the company
SCORE_MED	Insurance intermediary score	Score assigned to the insurance intermediary.
ANOANALISE	Year of analysis	Year when the claim occurred.
ANOS_CARTEIRA	Years in portfolio	How long a client has been in the company.
FRACC	Installment payments	How the client pays his premium (annual, semi-annual, quarterly or monthly).
COMBUST	Fuel	Fuel type (1 to 5; 1 Gasoline; 2 Diesel ;3 Gasoline and LPG Gas; 4 Electricity; 5 Other).
COND_TIPO_ENTIDADE	Entity	Entity that owns the vehicle (0 1 2; 0 indefinite; 1 individual; 2 collective).
NUM_LUG	Number of seats	Number of seats per vehicle.
Importado_PT	Imported	Classify if the vehicle is imported or not (0 1).
BM_RC	Bonus Malus	adjusts the premium paid by a customer according to their individual claim history.
IDADE_VEIC	Age	Nº of years of the vehicle.

Class_SEXO	Sex	driver's sex – variable converted into individuals or companies due to data protection policies – P (individual); E (company).
COND_ANOS_CRT	Years of driving license	How long a client has a driver license.
CAE_Agreg	Classification of economic activities	Classification of Portuguese economic activities in groups.
New_Cat_Uso	Vehicle category	Vehicle main category and main use (classification used by the company) .
PBRUTO_M	Vehicle weight	Gross vehicle weight.
New_Cat_Mud	Vehicle category (modified)	Vehicle main category (adaptation of New_Cat_Uso variable, used in the >200k segment).

DESCRIPTIVE CODE- VEHICLE CATEGORY

Code	Description
AM	Animal-dragged vehicles
AR	Articulated
AU	Bus
CC	Moped
CT	Truck
EP	Forklift
LP	Passenger Car
M4	Quad bike
MC	Motorcycle
MI	Industrial Machine
MT	Mist Machine
MV	Minivan
OU	Other
PS	Heavy vehicles
QD	Quadricycle
RB	Trailer
TA	Tractor
TT	All-terrain
VC	Velocipedes (not motorized)