NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Claim Severity Modelling in Automotive Insurance: Are Electric Vehicles Riskier?

A case study analysis using Portuguese Data

Ana Maria Antunes Serra

Master Thesis

presented as partial requirement for obtaining the Master 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

Claim Severity Modelling in Automotive Insurance: Are Electric Vehicles riskier?

by

Ana Maria Antunes Serra

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

Supervisor: Gracinda Rita Diogo Guerreiro

Co-Supervisor: Jorge Miguel Ventura Bravo

December, 2024

STATEMENT OF INTEGRITY

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

[Lisboa, 02/12/2024]

ACKNOWLEDGEMENTS

This journey has not been an easy one, it was filled with challenges and obstacles along the way. However, I was never alone, and without the support of those around me, I would not have made it this far. I would like to express my deepest gratitude to my advisors for their invaluable support and direction throughout this path.

To my parents and my older brother, whose unwavering encouragement and motivation have been fundamental in helping me reach this milestone. To my friends and my boyfriend, for their unconditional support, patience, and belief in me every step of the way.
Thank you all for being part of this journey.

ABSTRACT

This dissertation aims to explore the factors influencing the severity of motor insurance claims, with a specific focus on understanding the differences between electric and traditional vehicles. By identifying key variables that impact claim costs, the study seeks to assist insurance companies in making informed decisions to enhance profitability and sustainability.

The research employs a combination of statistical analysis and predictive modeling techniques, including Generalized Linear Models (GLM) and Logistic Regression. Data from insurance claims with data related to mandatory third-party liability coverage, segmented into electric and traditional vehicles, are analyzed to identify patterns and variables that significantly affect accident severity.

The study reveals distinct characteristics between electric and traditional vehicle claims. Variables such as vehicle age, geographic location, and type of accident contribute differently to the severity of claims in the two segments. For non-electric vehicles, variables such as the vehicle's gross weight, the district, the vehicle's year of construction, driver's age, years of driving experience and type of vehicle were obtained; for electric vehicles, only the vehicle's year of construction, brand, and the district were found significant.
Through the study using logistic regression, we concluded that electric vehicles have a higher probability of causing severe accidents.

The results provide actionable insights for insurance companies, enabling them to optimize premium calculations and reduce financial risks associated with claim payouts. By leveraging these findings, insurers can improve their pricing accuracy and competitive positioning in a rapidly evolving market.

Understanding the risk profiles of electric and traditional vehicles supports the development of fairer insurance policies. It offers valuable insights for insurers, policymakers, and stakeholders, providing a foundation for more effective risk management and promoting sustainable growth in the motor insurance industry.

KEYWORDS

GLM; Logistic Regression; Non-life Insurance; Pricing; Severity; Predictive Modelling; Electric Cars

Sustainable Development Goals (SDG):

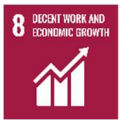


TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

GLM Generalized Linear Model

GBM Gradient Boosting Machine

FS Frequency-Severity

GAM Generalized Additive models

XGBoost Is an optimized distributed Gradient Boosting library designed to be highly efficient, flexible and portable.

AdaBoost AdaBoost algorithm is a technique for addressing binary classification challenges. This powerful algorithm enhances prediction accuracy by transforming a multitude of weak learners into robust, strong learners.

RF Random Forest is a learning method in which number of decision trees are constructed at the time of training and outputs of the modal predicted by the individual trees. RF act as a tree predictors where every tree depends on the random vector values.

ANN Artificial neural networks also shortened to neural networks (NNs) or neural nets are a branch of Machine Learning models that are built using principles of neuronal organization.

VIF Variance Inflation Factor, is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when two or more independent variables are highly correlated

1. INTRODUCTION

An insurance contract is an agreement through which the insurer assumes the coverage of specific risks, committing to provide compensation or pay the insured capital in the event of a claim, under the agreed terms.

In return, the individual or entity entering into the insurance agreement (the policyholder) is required to pay the corresponding premium, which represents the cost of the insurance.

This contract establishes the rights and obligations of both parties, including the insurer's responsibility to honor valid claims and the policyholder's duty to disclose relevant information and pay premiums promptly. It also outlines the conditions under which coverage applies, exclusions, and any limits on indemnity or insured capital.

Insurance contracts play a crucial role in risk management by enabling individuals and organizations to transfer financial uncertainty to the insurer, ensuring greater economic stability and peace of mind in the face of unforeseen events.
Sharing or pooling risks is the central concept of the insurance industry. This idea stands out for its simplicity coupled with practical utility. If risks, meaning the possibility of loss, can be distributed among various members of a group, then the burden becomes lighter for each member of that group. In this way, misfortunes that could be overwhelming for an individual become manageable for all.

Insurance companies assess risk through a process called underwriting, which involves evaluating the likelihood and potential cost of a claim being made on a policy. The goal is to determine appropriate premiums and ensure that the insurer can cover potential losses while remaining profitable. Insurers use historical data and statistical models to predict the likelihood of a claim. For instance, in automobile insurance, insurers assess variables such as the car's brand and model, the age of the policyholder, the mileage, and any previous penalties incurred by the policyholder, among others.

The insurance sector is complex, and the need to move fast, safeguarding people and places, physical and digital assets, becomes even more critical in the face of these challenges. And we are already witnessing it, insurers rely on data insights, making more informed and impactful decisions with a customer-centric focus. They are exploring the positive impacts of technology, such as generative AI, and placing technological innovations at the core of their digital transformation journey (Keeney, 2023).

In the world of automobile insurance, the development of in-vehicle telecommunication devices (telematics)-technology, wireless connectivity, machine-to-machine communication, and mobile applications powered the development of usage-based insurance (UBI) tracking mileage and driving behavior (Cunha & Bravo, 2022).

However, a significant current risk is adjusting prices to meet competition. Insurers may be compelled to continuously lower premiums to remain competitive, potentially leading to premiums so low that they fail to cover the cost of claims, which can be disastrous for the financial health of the company.

In addition to competitive pricing pressures, insurers must also adhere to minimum capital requirements and maintain adequate reserves to ensure their solvency and ability to meet claim obligations. These regulatory requirements are designed to safeguard policyholders and ensure that insurers can withstand unexpected losses or periods of high claim frequency.

Failing to balance competitive pricing with sufficient reserves and capital adequacy could undermine the insurer's ability to operate, highlighting the critical need for careful risk assessment, pricing strategies, and financial planning in the insurance industry.

Therefore, accurately predicting claim severity is crucial for insurance companies. This enables them to assess the potential financial impact of future claims and appropriately determine premium values based on the level of risk associated with prospective policyholders. By forecasting claim severity, insurers can better manage their financial reserves, set appropriate pricing strategies, and ensure their long-term solvency and competitiveness.

The Generalized Linear Model is a traditional model frequently used by actuaries in determining premiums for non-life insurance. These models are recognized for their effectiveness in model fitting and ease of interpreting results. However, they have several limitations that may reduce their effectiveness of GLMs in capturing the complexities of modern insurance risk assessment. These limitations include the assumption of a linear relationship between the predictors and the transformed response variable, with real-world insurance data often exhibiting non-linear relationships that cannot be accurately captured using this approach. Another limitation is the difficulty of handling interactions. Specifying and including all relevant interactions can be challenging and may lead to overfitting. GLMs require specifying a distribution for the response variable (eg Poisson, Gamma) which, if inappropriately chosen, introduces bias in the predictions of the model. Moreover, they are sensitive to outliers. GLMs assume that the relationships between variables are static over time. In fast-changing market conditions or emerging risks, these assumptions may fail to adapt.

To overcome these challenges and limitations, insurers often complement GLMs with novel techniques such as Gradient Boosting Machines (GBMs) and Random Forests for better handling of non-linearity and interactions. Compared to GLM, GBM does not require prior knowledge of the data structure. Alternative methods include Neural Networks (NN) for capturing complex patterns and relationships, regularization methods such as LASSO or Ridge regression to reduce overfitting, or Hybrid models combining GLMs with machine learning approaches to retain interpretability while enhancing predictive power (Clemente, Guerreiro, & Bravo, 2023). One drawback of Machine Learning techniques is that they tend to create 'black-box' models. In other words, explaining how the model works can be a challenge.

This dissertation aims to develop a claims severity model incorporating business attributes and insured environment variables to better understand portfolio behavior and identify key differences in risk profiles between electric vehicles (EVs) and non-electric vehicles (non-EVs). Generalized Linear Models (GLM) will be employed to analyze the factors influencing claim costs, while logistic regression models will be used to estimate the likelihood of severe accidents. By comparing EV and non-EV risk assessments, this research seeks to provide actionable insights for insurers, enabling more accurate pricing strategies and improved risk management tailored to the evolving automotive market.
The data used for this study is only related to mandatory third-party liability coverage.

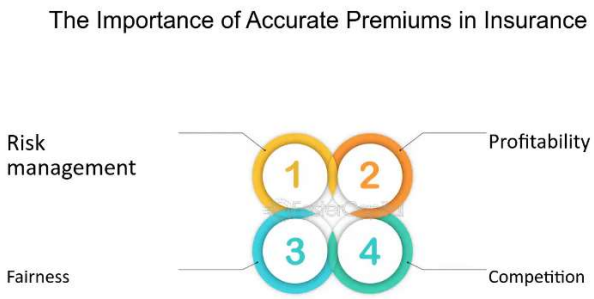
This thesis is structured into five main chapters that guide the development of the study. The first chapter presents the literature review, discussing fundamental concepts and theories related to the topic, as well as previous studies that provide context and support for the research. Next, the methodology chapter outlines the approaches and techniques used for data collection and analysis. The third chapter is dedicated to exploratory data analysis, presenting descriptive statistics and visualizations that characterize the variables and reveal initial patterns. The fourth chapter, discussion of results, and the findings of this study are presented. Finally, the conclusion and limitations chapter summarizes the study's contributions, discusses its constraints and potential biases, and suggests directions for future research.

2. LITERATURE REVIEW

2.1. RATEMAKING/PRICING

An insurance company essentially "buys" claims and "sells" risk protection. If the company acquires claims at a lower price than the insurance premium, it makes a profit, but if it acquires claims at a higher price than the premium, it incurs losses. In the value chain, a company can rely on the Law of Large Numbers, which mitigates market volatility and uncertainty - providing security in average terms.

Figure 2.1 - Spheres influenced by insurers' pricing decisions



Source: Rate making: The Power of Adjusted Premiums in Rate Making Strategies. (n.d.). FasterCapital.[1]

The determination of rates refers to setting insurance prices and calculating insurance premiums. The premium paid by the policyholder is the result of multiplying a rate, determined by actuaries, by the exposure to risk and then adjusting the premium for various rating factors (a process known as classification) (George E. Rejda, 2017).

According to Saha (2023) some of the factors that can influence the non-life insurance premiums are:

- Risk Assessment: Insurers evaluate the risk tied to insured property or liability by analyzing historical data, statistical models, and actuarial methods to estimate the likelihood of events resulting in a claim. For example, in auto insurance, factors such as driver's age, driving record, and vehicle type influence the premium;
- Loss Experience: Insurers analyze their historical loss experience to identify trends and patterns, aiding in the prediction of future claims. Crucial to this assessment is historical data covering the frequency, severity, and type of claims;

[1] Retrieved April 24, 2024, from https://fastercapital.com/content/Rate-making--The-Power-of-Adjusted-Premiums-in-Rate-Making-Strategies.html

- Underwriting Factors: "Underwriters assess specific factors related to the insured, such as credit history, location, and the value of the insured property. These factors offer insight into the policyholder's risk profile and help determine the appropriate premium;
- Market Competition: Insurers take into account their competitive position, market share, and the pricing strategies adopted by competitors when establishing premium rates.

The primary goal of setting insurance rates is to find the most economical premium that meets all the necessary objectives. A crucial aspect of this process involves recognizing all the features that can predict future losses, with the aim of charging lower premiums to low-risk groups and higher premiums to high-risk groups. By offering lower premiums to those in lower risk categories, an insurance company can attract these individuals, thus reducing its own losses and expenses. Concurrently, this strategy increases the losses and expenses of other insurance companies, as they retain a larger number of policyholders from higher-risk groups. This explains why insurance companies invest in actuarial studies, identifying all characteristics that can realistically predict future losses as accurately as possible (KAGAN, 2021).

The need to set the insurance price before knowing its cost underscores the importance and relevance of the pricing process. Insurance companies must make estimates on the quantity and magnitude of claims that may occur. Furthermore, the pressure to compete on pricing within the sector serves as an incentive for the ongoing development of the rate-setting process.

In the non-life insurance sector, there is an interest in motor insurance, as it involves managing a large number of risk events. Over the years, it has evolved significantly, but there are still many challenges that need to be addressed or improved. One of these challenges, which this dissertation aims to help address, is severity models.

2.2. FREQUENCY AND SEVERITY MODEL

Let us think of datasets where there is a large proportion of zeros, corresponding to no claims. To address this large proportion of zeros, we consider a two-part model, which is a special type of frequency-severity model. In a two-part model, one part shows whether an event (claim) occurs, and the second indicates the size of the claim (Frees, Derrig, & Meyers, 2014).

These models decompose the cost of claims into two distinct components. The frequency component analyses whether a claim occurs or not (using logistic regression) or the number of claims, for instance, through Poisson regression. The severity component examines the cost

of claims conditional on occurrence (using, for instance, Gamma or Inverse Gaussian Regression).
Both the frequency and severity of claims are random variables, introducing the risk that future experiences may deviate from past events. Therefore, it is very important to employ appropriate statistical distributions when modelling (Shi, Feng, & Ivantsova, 2015).

In general, smaller claims tend to have a higher frequency, whereas larger claims tend to have a lower frequency.

This modeling strategy assumes that the occurrence and the financial impact of insurance claims are non-correlated. The pure premium at either an individual or class level is obtained simply by multiplying the mean estimates for both claim frequency and claim severity.

However, there are two significant drawbacks in the frequency-severity model. First, the model is built using a linear predictor format, which might not adequately address the non-linear impacts of predictors in real-world situations. For example, in auto insurance, the nonlinear correlation between claim severity and the age of the insured highlights the constraints of this linear format. Second, the conventional frequency-severity model presupposes the independence between claim frequency and severity, whereas in fact, these two factors frequently show interdependence (Su & Bai, 2020).

Recent literature on insurance suggests new models that relax the assumption of independence by incorporating a shared random effect or a copula. Specifically, Czado, Kastenmeier, Brechmann, and Min (2012) introduce the dependence between claim frequency and average severity using a Gaussian copula family. Nevertheless, both selecting an appropriate copula family and estimating copula parameters prove to be challenging tasks in practical applications (Lee, Park, & Ahn, 2019).

In a study conducted by Su & Bai (2020), they forecasted both the frequency and severity of Third Party Liability (TPL) motor insurance coverage. They employed a combination of stochastic Gradient Boosting and a profile likelihood approach to estimate distribution parameters. What distinguishes their work is the incorporation of a connection between claim frequency and average claim cost. They included claim frequency as a predictor in the regression model for severity.
The results indicated that the models considering this dependency exhibited superior performance compared to other advanced models currently used.

Several actuarial models for claim severity are based on continuous distributions, while discrete probability distributions are used for frequencies. The log-normal, gamma or Inverse Gaussian Regression distributions are among the most common distributions for modelling claim severity. Other distributions for claim size are the Exponential, Weibull, and Pareto distributions (Cyprian Ondieki Omari, 2018).

2.3. GENERALIZED ADDITIVE MODELS

Similar to how GLMs are derived from linear models, additive models can be easily expanded to GAMs. A GAM represents an additive extension of the GLM family (Hastie & Tibshirani, 1986). GAMs enhance traditional GLMs by allowing the linear predictor to be influenced linearly by unknown smooth functions. Within GAMs, the linear predictor is substituted with an additive combination of parametric fits and certain smooth functions to forecast the expected response value. The response variable follows an exponential family distribution or has a known mean-variance relationship.

In GLMs, average responses are modelled as monotonic functions of linear scores. While this linearity assumption is not restrictive for categorical variables represented by binary indicators, its validity is questionable for continuous variables, which may exert a nonlinear impact on the score scale. This model maintains the additive breakdown of the score while allowing actuaries to uncover nonlinear effects of factors such as policyholder age or geographic location (geographic effect). GAMs provide a flexible, data-driven method to ascertain the optimal transformation of continuous variables for incorporation into the score scale. Specifically, continuous variables are introduced into the model through a semi-parametric additive predictor (W.N.Venables & C.M.Dichmont, 2004).

The utilization of cubic penalized regression splines in GAMs gives a more adaptable and smooth representation of the relationship between the indicators and the reaction factors than GLM's.

As mentioned before, in auto insurance, there exists a nonlinear relationship between claim severity and the insured's age (Su & Bai, 2020). Generalized additive models (GAM), address this constraint by incorporating smooth functions, derived from data, to model continuous variables. However, the additive structure of GAM models may not automatically capture intricate interactions among predictors. Although interaction terms can be physically included in the model structure, recognizing these terms can end up difficult, especially when dealing with numerous predictors. Failure to account for crucial interactions can decrease the accuracy of the prediction (Su & Bai, 2020).

2.4. TWEEDIE MODEL

Alternatively to the frequency and severity model, the total loss cost is at times directly modelled using the Tweedie distribution. This method, originally introduced by Jørgensen and de Souza (1994) and more recently reviewed by Quijano-Xacur and Garrido (2015), entails portraying aggregate claims as a combination of Poisson and Gamma distributed variables. The resulting Tweedie distribution belongs to the Exponential Family, and hence the inference procedure for GLMs directly applies. It is important to note that this approach implicitly assumes independence between the number of claims and the size of each individual claim.

In reality, there is often a correlation between the frequency and severity of claims. For example, in collision automobile insurance, claim counts and amounts might exhibit a negative correlation because drivers involved in minor accidents tend to file several claims per year. Consequently, there is a need to modify the model for aggregate claims to consider possible associations between claim frequency and severity.

Two general methods have been proposed to address this dependence between frequency and severity. Approaches by Frees and Wang (2006), Gschlößl and Czado (2007), and Frees et al. (2011) involve conditioning and use the claim count as a factor when modelling the distribution of average claim sizes (Garrido, Genest, & Schulz, 2016).

Each method has its own advantages. For example, the frequency–severity model is more flexible in modeling of the occurrence and the size of insurance claims. In contrast, with a more parsimonious specification, the Tweedie model simplifies the variable selection process (Shi, Feng, & Ivantsova, 2015).

2.5. GENERALIZED LINEAR MODELS

Generalized Linear Models are frequently used in the pricing of non-life insurance. These models encompass various factors and their relationships to estimate the cost of claims. They enable insurance companies to analyze historical data and construct statistical models linking policyholder characteristics (such as age, location, and coverage limits) to the probability and severity of claims. By fitting the model to historical data, insurers can make predictions about future claim costs and determine appropriate premium rates (Saha, 2023).

Frees et al. (2014) propose incorporating claim frequency as a covariate in the conditional mean model for average severity. An empirical investigation by Garrido, Genest, and Schulz (2016) and Shi, Feng, and Ivantsova (2015), employing an advanced version of the Frequency-Severity model, validates the significance of claim frequency as a covariate in explaining the conditional mean of average severity.
This approach, the simplest among the currently proposed models allowing for dependence between severity and frequency, appears to have potential for application in diverse non-life rate-making scenarios (Lee, Park, & Ahn, 2019).

The established models for understanding the severity of insurance claims, pioneered by Nelder and Wedderburn in 1972, involve using Generalized Linear Models. These models assume that the distribution of the response variable follows an exponential family of distributions, such as gamma, which are well-suitable for non-life insurance claims.

Moreover, a log-normal distribution is frequently applied in these models by taking the logarithm of the dependent variable and assuming a normal distribution afterward (Frees, Lee, & Yang, 2016).

These models are valuable in the Frequency-Severity approach because the means of the frequency and severity processes can be expressed, through specific transformations, as linear combinations of variables such as age, gender, and so on.

2.6. GRADIENT BOOSTING MACHINE

Introduced by Friedman in 2001, Gradient Boosting models are characterized as models comprising decision trees, where numerous weak models are combined to form a more robust predictor.

Machine Learning models, including Decision Trees, Random Forests, and Neural Networks, are increasingly utilized in non-life insurance pricing. Insurers can leverage Machine Learning algorithms to analyze large datasets, revealing patterns that traditional models may overlook. The ability of machine learning models to capture nuanced relationships between risk factors and claim outcomes contributes to enhanced pricing accuracy (Saha, 2023).

Fauzan & Hendri (2018) analyze the accuracy of XGBoost in auto-insurance claim prediction and conclude that XGBoost shows better accuracy when compared to the alternative methods: AdaBoost, Stochastic Gradient Boosting, RF, and ANN (Su & Bai, 2020). investigated the use of a Stochastic Gradient Boosting algorithm and a profile likelihood approach to estimate parameters for both the claim frequency and average claim severity distributions in a French auto insurance dataset and concluded that the approach outperforms standard models.

In contrast to other Machine Learning techniques with comparable predictive accuracy, Gradient Boosting yields interpretable results, making it especially appealing for modeling motor insurance losses. GB models simplify the representation of complex interactions, allowing their incorporation into the pricing structure.
Feature selection is integrated into the model application, offering a flexible approach when using these models for insurance pricing (Clemente, Guerreiro, & Bravo, 2023).

2.7. DIFFERENCES BETWEEN ELECTRIC AND TRADITIONAL VEHICLES

Research by McDonnell et al. (2024) on driver behavior and risk profiles of alternative energy vehicles reveal critical differences between electric vehicles (EVs), hybrid vehicles (HYBs) and traditional internal combustion engine vehicles (ICEs). EVs, with their unique design of engine, transmission, and pedal control, exhibit distinct driving behaviors compared to ICEs, including fewer harsh events, such as braking and acceleration. However, this reduction in harsh events does not translate into fewer at fault claims. In contrast, EV drivers are 4% more likely to experience an accident claim than ICE drivers, even with their lower average mileage. Logistic regression models confirm that electric vehicles have a statistically significant higher risk profile than ICEs, with travel distances identified as a key factor contributing to this increased risk.

HYBs show the fewest harsh driving events per trip but still have a 6% higher crash likelihood than ICEs, although no further risk concerns emerge in logistic models. Additionally, financial analyses highlight that more than one third of electric vehicles and HYBs incur damages exceeding €1,000 in accidents, largely attributed to the high costs of battery replacement. Combined with emerging risks like lithium-ion battery fires, these financial and safety challenges pose significant barriers to widespread EV and HYB adoption.

Statistical analyses, such as Kruskal-Wallis and Dunn's tests, confirm that driving behavior significantly differs across vehicle types, with HYBs displaying the greatest behavioral shift. These findings underscore the need for manufacturers, policymakers, and stakeholders to respond to the increased risk and financial burden associated with EVs and HYBs, which could otherwise hinder the transition to alternative fuel vehicles.

In the insurance industry, according to Valdes-Dapena (2024), there has been a long-standing correlation between horsepower and the frequency and severity of insurance claims. High-performance vehicles tend to collide more often and with greater impact, resulting in a greater number of accidents, often of greater severity. In addition, electric vehicles (EVs) lack the traditional engine noises associated with rapid acceleration and high speeds, potentially leading drivers to be less aware of their speed.

3. METHODOLOGY

3.1. GENERALIZED LINEAR MODELS

Generalized Linear Models (GLM) are a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables.

It is assumed that the outcome of the target variable is influenced by both a systematic component and a random component. The objective in modeling with Generalized Linear Models is to "explain" a significant portion of the variability in the outcome through predictors. In simpler terms, the goal is to transfer as much variability as possible from the random component to the systematic component (Goldburd, Khare, Tevet, & Guller, 2020).

3.1.1. COMPONENTS OF A GLM

A GLM consists of three main components. The first is a random component specifying the probability distribution of the response variable Y. The random component defines the conditional distribution of Y given the predictors X. The second component is the systematic component representing the linear predictor η, a linear combination of the explanatory variables:

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_p X_p, \tag{3.1}$$

where:

- $g(\mu)$ is the Link Function, which is a transformation of the mean of the response variable μ. This function ensures that the mean of the response variable is in the correct range, depending on the distribution of the response variable;
- β_0 is the intercept (constant) of the model;
- $\beta_1, \dots, \beta_p$ are the coefficients of the explanatory variables (predictors);
- $X_1, \dots, X_p$ are the explanatory variables (independent variables or predictors) used to model the response variable.

The third key component is the link function, which links the mean of the response variable $\mu = E(Y)$ to the linear predictor $g(\mu) = \eta$ or, equivalently, $\mu = g^{-1}(\eta)$. Common link functions include: Identity ($g(\mu) = \mu$) for the Normal distribution; Log ($g(\mu) = \log(\mu)$) for Poisson or Gamma distributions; or the Logit ($g(\mu) = \log(\frac{\mu}{1-\mu})$)) for the binomial distribution.

3.1.2. THE EXPONENTIAL FAMILY

The Exponential family of distributions forms the mathematical foundation for Generalized Linear Models (GLMs) and is extensively applied in insurance data analysis. According to the work developed by De Jong (2008), the structure of this family is defined by its probability density (or mass) function, which can be expressed in a unified form:

$$f(y) = \mathrm{c}(\mathrm{y}, \emptyset)\, exp\left[\frac{y\theta - a(\theta)}{\emptyset}\right]$$

(3.2)

where:

- $\mathrm{c}(\mathrm{y}, \emptyset)$**:** A normalizing constant depending on y and the dispersion parameter ϕ;
- θ**:** The canonical parameter, which determines how the mean of the distribution relates to the linear predictor in GLMs;
- $a(\theta)$**:** The cumulant function, ensuring the probability integrates to 1 and defining the mean-variance relationship;
- $\emptyset > 0$**:** The dispersion parameter, controlling variability in the distribution.

This canonical structure is central to the Exponential family and provides a unified way to represent many probability distributions. Members of this family include well-known distributions such as the Binomial, Poisson, Normal, Gamma, Inverse Gaussian, and Negative Binomial. Each of these distributions supports different types of response data, with their specific variance-mean relationships playing a critical role in model selection.

Key properties of the Exponential family are crucial in GLM theory. For example, the expected value and variance of the response variable y are derived as:

$$E[y] = \dot{a}(\theta), \qquad Var[y] = \phi\ddot{a}(\theta),$$

(3.3)

Where $\dot{a}(\theta)$ and $\ddot{a}(\theta)$, denote the first and second derivatives of with respect to $a(\theta)$ and θ. These properties facilitate the specification of the link function and the variance function $V[\mu]$ in GLMs, directly connecting the mean structure to the dispersion.

Depending on different choices of $a(\theta)$ and θ we obtain the different distributions mentioned earlier. These distributions allow for modelling diverse response types, from counts and proportions to continuous variables with non-constant variance:

- The Binomial distribution usually applies for binary or proportion data with variance depending on the mean and success probability;

- The Poisson distribution usually applies for modelling count data where the mean equals the variance;
- The Normal distribution applies for continuous data with constant variance;
- The Gamma distribution is used for modelling positive continuous variables with non-constant variance, such as waiting times or insurance claims;
- The Inverse Gaussian distribution applies to heavily skewed positive continuous data;
- The Negative Binomial distribution is an extension of the Poisson distribution for modelling count data with overdispersion.

3.1.3. PARAMETER ESTIMATION

In the process of developing a generalized linear model, the aspect that garners the most attention is the estimation of the regression parameters β. These parameters are derived by maximizing the log-likelihood function (Jong & Heller, 2008). Due to the absence of analytical solutions for these equations within the GLM framework, it's customary to employ numerical methods for resolution (Clemente C. d., 2022).

In addressing this challenge, Nelder & Wedderburn (1972) devised an algorithm, referred to as Iterative Weighted Least Squares Estimation, rooted in the Fisher scoring method, to arrive at a solution for these equations.

The maximum likelihood estimation is determined by selecting the parameter estimates that optimize the likelihood of observing the sample $y_1, \dots, y_n$. Each y_i has a probability function $f(y_i)$, which consequently relies on ϕ, if applicable. Due to the independence of y_i, their joint probability function is as follows:

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

(3.4)

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

$$l(\theta, \emptyset) = \sum_{i=1}^{n} \ln f(y_i, \theta, \emptyset)$$

(3.5)

To obtain the parameters θ and $\emptyset$ we haver to maximize the log-likelihood function represented above.

3.1.4. MODEL SELECTION

The criteria for selecting our model are not one-sided. Intuitively, the objective is to ensure that the predicted curve closely matches the observed data without over or underestimation, while also preventing overfitting.

Concerning the variable selection, there are two measures utilized to evaluate whether a variable should be incorporated into the model or not: the deviance and the Likelihood Ratio test.

3.1.4.1. AKAIKE INFORMATION CRITERION

The Akaike Information Criterion (AIC), introduced by Hirotsugu Akaike in 1973, serves as a valuable tool for comparing two models when they are not nested. It represents a balance between deviance and model complexity. Typically, we compare two models based on their AIC values, with smaller values indicating better fit.

This method relies on the log-likelihood function $loglik(\beta)$, which evaluates the fitting quality, augmented by a correction factor tied to the model's parameter count, p. This correction penalizes models with an increased number of variables. In essence, it equals the deviance plus 2 times the number of parameters, expressed as:

$$AIC = -2loglik(\beta) + 2p$$

(3.6)

3.1.4.2. DEVIANCE AND CHI-SQUARED TEST

When two models are nested (one being a sub model of the other), we typically compare them by examining their deviances. The deviance is essentially minus 2 times the log-likelihood. Assuming the two models have p_1 and p_2 parameters, respectively, where the model with p_2 parameters is the submodel of the model with p_1 parameters. The difference in deviance between the two models should asymptotically follow a chi-square distribution with $p_1 - p_2$ degrees of freedom. The test statistics are given by the equation:

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

(3.7)

The chi-squared percentage (chi-squared p-value) represents the probability of a random variable following a chi-squared distribution with $p_1 - p_2$ degrees of freedom to exceed the difference in deviance between the two models.

If this percentage is smaller than our chosen significance threshold α, then the model with more parameters is considered significantly better than the model with fewer parameters (Zhifeng, 2020).

3.1.5. QUALITY OF FITTING

This step occurs after selecting the variables with the most significant coefficients, meaning after identifying the model that best fits the data. This evaluation is conducted by analyzing deviance and residuals.

3.1.5.1. RESIDUALS

Residual analysis is crucial for diagnosing our model and is valuable for detecting deviations such as under or overestimation.

There are various expressions of residuals, one of them, the standard, is:

$$\widehat{e_l} = y_i - \widehat{\mu_l}$$

(3.8)

Deviance residuals, often used for evaluation, represent the difference between the observed value y_i and the fitted value $\widehat{y_l}$. These residuals, detailed below, are commonly assessed to evaluate the model's performance.

$$r_D = sign(y_i - \widehat{\mu_l})\sqrt{d_i}$$

$$d_i = 2y_i \log\left(\frac{y_i}{\widehat{\mu_l}}\right) - (y_i - \widehat{\mu_l})$$

(3.9)

This equation represents the difference between each fitted and observed value.

Deviance residuals are often favored for diagnosing GLMs over other types of residuals such as Pearson's residuals, response residuals, or working residuals. The choice depends on the specific characteristics of the data being modeled.

3.1.6. GENERALIZED LINEAR MODELS ADVANTAGES

Multiplicative models represent the most prevalent type of rating structure employed in pricing insurance, and they offer several advantages over alternative structures. Some of these advantages include:

- Simplicity and Practicality: Multiplicative models are straightforward and practical to implement;

- Avoidance of Negative Premiums: Additive terms in a model can lead to negative premiums, which is illogical. Multiplicative plans ensure positive premiums without the need for workarounds like minimum premium rules;
- Intuitive Appeal: Multiplicative models are more intuitively appealing. Assigning a fixed increase in auto premium, regardless of the base premium, may not align with logic. It is more sensible to express a violation surcharge as a percentage, providing a clearer understanding of the impact relative to the base premium.

3.2. LOGISTIC REGRESSION MODEL

Logistic Regression is a statistical modelling technique designed to estimate the relationship between a binary response variable Y and a set of explanatory variables. A logistic regression model is a specific case of a GLM designed to model binary or categorical outcome variables. The model is based on the principle of modelling the probability of a binary outcome (i.e. $Y \in \{0,1\}$) as a function of explanatory variables, ensuring that the predicted probabilities fall within the interval [0,1]. The model achieves this using the logit link function, which connects the linear combination of predictors to the log-odds of the binary response, i.e.,

$$g(\mu) = \ln\left(\frac{\pi}{1-\pi}\right) = \eta = x'\beta,$$

(3.10)

where π represents the probability of success (i.e., $y = 1$), x is the vector of explanatory variables, and β is the vector of coefficients to be estimated. This formulation ensures that the relationship between the predictors and the binary outcome can be modelled linearly in terms of the log-odds, a property that simplifies interpretation while maintaining mathematical rigor. By rearranging the logit transformation, the probability of success can be expressed directly as:

$$\pi = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

(3.11)

Equation *(3.11)* maps the linear combination of explanatory variables, $x'\beta$, to a value between 0 and 1, making it a suitable model for probabilities. The denominator $1 + e^{x'\beta}$ ensures that the predicted probabilities are normalized, while the numerator $e^{x'\beta}$ controls the relative likelihood of success as a function of the predictors.

The coefficients β in Logistic Regression are typically estimated using maximum likelihood estimation (MLE). This method identifies the parameter values that maximize the likelihood of observing the given data under the assumed logistic model.

The likelihood function is constructed based on the Bernoulli distribution of the binary response, where each outcome is assumed to follow the probability π predicted by the Logistic model.

Logistic Regression coefficients have intuitive interpretations. For a continuous predictor x, the coefficient β represents the change in the log-odds of success associated with a one-unit increase in x.

Specifically, if the model is expressed as $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$, then an increase in x by one unit multiplies the odds of success by e^{β}. If $\beta > 0$, the odds increase, whereas $\beta < 0$ indicates a decrease in the odds. For small values of β, the term $e^{\beta} - 1 \approx \beta$ provides an approximate percentage change in the odds.

When dealing with categorical explanatory variables, Logistic Regression incorporates indicator variables to represent different levels of the categorical factor. For a categorical variable with r levels, $r - 1$ indicator variables are used, with one level serving as the reference or baseline category. The model can be expressed as:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{r-1} x_{r-1}$$

(3.12)

where x_j is an indicator variable that takes the value 1 if the observation belongs to level j of the categorical variable and 0 otherwise. The coefficient β_0 represents the log-odds of success for the reference category, while β_j quantifies the change in the log-odds of success for level j relative to the reference category. The corresponding odds for level j are obtained by exponentiating the log-odds, giving e^{β_j}, which represents the multiplicative change in the odds of success for level j compared to the reference level. For instance, if $\beta_j > 0$, the odds of success for level j are higher than those for the reference category, while $\beta_j < 0$ indicates lower odds.

While the logit link is the most commonly used function for binary response modelling, Logistic Regression can be generalized to accommodate alternative link functions. The probit link, for example, uses the inverse of the cumulative distribution function of the standard Normal distribution, φ^{-1}, to model the relationship. Alternatively, the complementary log–log link is expressed as $\ln(-ln(1-\pi))$, which is particularly useful in survival analysis and other contexts where asymmetric behavior around $\pi = 0.5$ is expected. Despite these variations, the choice of link function retains the key property of mapping predicted values to the interval (0,1), ensuring consistency with probability theory.

Logistic regression provides a versatile and theoretically robust framework for modelling binary outcomes by linking explanatory variables to the probability of success through the logit function. It accommodates both continuous and categorical predictors, allowing for a nuanced interpretation of relationships through odds ratios and log-odds transformations. With its adaptability to include multiple predictors and its ability to generalize alternative link functions, Logistic Regression serves as a foundational tool in statistical modelling.

3.2.1. ODDS RATIO

The odds ratio is defined as the ratio between the odds of the event of interest occurring ($Y = 1$) for individuals with $x = 1$ and the odds of it occurring for individuals with $x = 0$. The odds of the event happening for individuals with $x = 1$ is given by $\frac{\pi(1)}{1-\pi(1)}$.

Similarly, the odds of the event occurring for individuals with $x = 0$ is given by $\frac{\pi(0)}{1-\pi(0)}$. Thus, the odds ratio provides a way to compare whether the probability of the event of interest is the same for individuals with $x = 1$ and $x = 0$.

3.2.2. CLASSIFICATION

A regression model can be statistically significant yet fail to accurately represent the reality under study. One way to assess the model's classification efficiency is through classification tables. To build these tables, we need to calculate the estimated probabilities of the event occurring (the endpoint), and then determine a cut-off value, c , for these probabilities. Based on this cut-off, we assume that individuals with estimated probabilities higher than c will experience the endpoint, while those with probabilities below the cut-off will not. The commonly used cut-off value is 0.5, but this is not always the most suitable. To find a more appropriate cut-off, we use graphical methods, such as the ROC curve, which helps us identify the point where the model's sensitivity and specificity are balanced.

Model sensitivity is defined as the probability of correctly predicting the occurrence of the endpoint among individuals for whom it was actually observed. Sensitivity gives us the proportion of true positives. Specificity, on the other hand, provides the proportion of false negatives, indicating the probability of correctly predicting the non-occurrence of the endpoint among individuals where it was not observed.

Table 3.1 - Classification Table

		Predicted	
		Endpoint=1	Endpoint=0
Observation	Endpoint=1	A	B
	Endpoint=0	C	D
	Accuracy	$\frac{A}{A+B}$	$\frac{D}{C+D}$

Source: Author's Preparation

In a perfect model, all cases would lie on the main diagonal. However, in practice, achieving a perfect model is very difficult, so we need to assess its predictive ability. In general, higher accuracy indicates a better model because it means the model is correctly predicting a larger proportion of outcomes, which reflects its ability to generalize well from the data. Another measure that helps evaluate the model's performance is the ROC curve. This method is widely used across various models and has already been explained in the next section.

3.2.3. PERFORMANCE MEASURERS

3.2.3.1. ROC CURVE

The ROC (Receiver Operating Characteristic) curve plots sensitivity against 1 minus specificity (false positive rate) for each threshold. Typically, the horizontal axis represents 1 minus specificity, while the vertical axis shows sensitivity. With this axis orientation, a point near zero on the x-axis (indicating high specificity) usually corresponds to a low value on the y-axis (indicating low sensitivity), and vice versa (De Jong, 2008).
Here's a breakdown of these terms:

- True Positive (TP): Instances correctly predicted as positive;
- False Positive (FP): Instances incorrectly predicted as positive when they are actually negative;
- True Negative (TN): Instances correctly predicted as negative;
- False Negative (FN): Instances incorrectly predicted as negative when they are actually positive;
- Recall/Sensitivity (True Positive Rate): The proportion of actual positive samples correctly predicted. A high recall indicates that the model captures most of the actual positive cases, reducing the risk of missing important instances;
- Specificity (True Negative Rate): The proportion of actual negative samples correctly predicted;
- Precision: The ratio of correctly predicted positive instances to all instances predicted as positive. It evaluates the accuracy of the model's positive predictions. A high precision indicates that the model is very reliable when it predicts a positive outcome;
- F1 Score: It combines precision and recall into a single metric by calculating their harmonic mean. The F1 score can be optimized by adjusting the decision threshold, balancing precision and recall for better performance;

As the classification threshold changes, the balance between Precision and Recall alters, leading to the creation of the ROC curve.
In the ROC space, a perfect classifier would occupy the top-left corner, representing high Recall and Precision simultaneously. Conversely, a random guess would produce a diagonal line from the bottom-left to the top-right, indicating an Area Under the Curve (AUC) of 0.5.

The closer the ROC curve is to the top-left corner, the better the model's performance in distinguishing between classes.

3.2.3.2. AREA UNDER THE CURVE (AUC)

The Area Under the ROC Curve (AUC) serves as a comprehensive metric for assessing the performance of a classification model. It quantifies the likelihood that the model will correctly rank a randomly selected positive instance higher than a randomly chosen negative instance. In essence, AUC reflects the model's capability to differentiate between positive and negative classes.

AUC values range from 0 to 1, where:

- 0.5 indicates random classification, suggesting the model performs no better than chance;
- 1 signifies a perfect classifier, indicating flawless performance in distinguishing between classes.

ROC curves and AUC are invaluable tools for assessing the performance of classification models. They offer a holistic perspective on a model's performance across various classification thresholds. By leveraging these metrics, one can gain deep insights into their models' capabilities and make informed decisions regarding model selection and optimization. Integrating ROC curves and AUC into the evaluation process enhances the robustness and accuracy of models, ultimately leading to more reliable and effective solutions.

3.3. VARIANCE INFLATION FACTOR

The Variance Inflation Factor (VIF) is a metric used in regression analysis to measure the degree of multicollinearity among independent variables. Multicollinearity arises when two or more independent variables are correlated, which can adversely impact the regression results.

VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity.

3.3.1. ADVANTAGES

To assess the presence of multicollinearity among independent variables in a multiple regression model, it is important to understand its implications. While multicollinearity does not reduce the overall explanatory power of the model, it can diminish the statistical significance of individual independent variables. High values of multicollinearity indicate strong collinearity among the variables, which may require adjustments to the model.

To obtain this factor we have the following formula,

$$VIF_i = \frac{1}{1 - R_i^2}$$

(3.13)

where R_i^2 is the unadjusted coefficient of determination when regressing the i-th independent variable on the remaining independent variables. As this value increases the VIF value grows, which means higher multicollinearity. If $VIF_i = 1$, then no correlation exists between the independent variable and the others (no multicollinearity). If 1 < VIF ≤ 5, variables are moderately correlated. Usually, this level of multicollinearity is not a concern. If VIF > 5, there is significant collinearity that should be addressed. Finally, if VIF > 10, severe multicollinearity is detected, requiring correction before proceeding with the analysis (Variance Inflation Factor (VIF), n.d.).

3.4. CRAMÉR'S V

Cramér's V (Kearney, 2017) is a statistical measure designed to assess the strength of association between two categorical variables, particularly when these variables have more than two levels. While the Chi-Square test is widely used to determine whether an association exists, it does not indicate the strength or practical significance of that association. Cramér's V addresses this gap by providing a measure that normalizes the Chi-Square statistic and expresses the strength of association on a standardized scale from 0 (no association) to 1 (perfect association).

The foundation of Cramér's V lies in the Chi-Square statistic (χ2), which tests the independence of two variables by comparing observed frequencies to expected frequencies under the null hypothesis. If the observed and expected frequencies align closely, the Chi-Square value is small, suggesting no significant association. Conversely, large differences yield a higher Chi-Square statistic, indicating that the variables are related.

However, the Chi-Square statistic is sensitive to sample size, often producing inflated values in larger datasets even when the relationship between variables is weak. This limitation makes the statistic difficult to interpret directly in terms of effect size.

Cramér's V builds on the Chi-Square statistic by adjusting for the size of the dataset and the smaller dimension of the contingency table. Its formula is:

$$V = \sqrt{\frac{\chi^2}{n(L-1)}}$$

(3.14)

Here, n is the sample size, and L represents the smaller number of rows or columns in the table. This normalization ensures that the measure remains unaffected by sample size, unlike the raw Chi-Square statistic.

A key advantage of Cramér's V is its consistency across studies with varying sample sizes and contingency table dimensions, making it particularly useful for comparing associations in different datasets or research contexts.

Cramér's V is particularly valuable in fields such as social science, marketing, and public health, where categorical data is frequently analyzed. For example, researchers might explore associations between educational levels and voting preferences, product categories and consumer demographics, or patient characteristics and health outcomes. In such contexts, Cramér's V can complement the Chi-Square test by quantifying the strength of detected relationships, providing a clearer picture of their practical significance. One limitation of Cramér's V is that it does not indicate the direction of the association, only its strength. Moreover, it assumes that the Chi-Square test's underlying conditions are met, including sufficient expected frequencies in the contingency table. When these assumptions are violated, the results may not be reliable.

Despite these limitations, Cramér's V is widely regarded as an effective measure for categorical data, especially when dealing with variables that have multiple levels.

Cramér's V is closely related to the phi coefficient, which is used for 2x2 tables, and the contingency coefficient. While the phi coefficient also normalizes the Chi-Square statistic, its application is limited to binary variables. The contingency coefficient adjusts the denominator differently but does not range between 0 and 1, which often makes its interpretation less intuitive. For variables with more than two levels, Cramér's V is generally preferred due to its bounded range and straightforward interpretation.

Due to the previously mentioned reasons, Cramér's V is an useful tool for researchers working with categorical data, allowing them to move beyond significance testing to evaluate the magnitude of relationships.

4. EXPLORATORY ANALYSIS OF THE DATASET

4.1. ABOUT THE DATASET

The proprietary database used in this dissertation is made up of 224,205 observations between 2019 and 2021. This database belongs to an insurance company and is real data on its claims for the years mentioned.

A division was made from this database, that is, two subsets of this database were created in which one contains data relating to electric/hybrid cars (886 observations) and the other contains the remaining observations (223,319 observations).

In addition to the response variables, this dataset includes 28 feature variables related to the client (such as the client's age, risk zone or driver's license age) and the vehicle (such as brand, age or horsepower). Table Table A.0.1 in Appendix A presents a summary of the independent variables in the original dataset.

It is also important to mention that in this study the cost variable present in the database is a standardized variable to comply with data protection constraints, i.e., this variable is the result of the cost of the claim divided by the average cost of all claims multiplied by a hundred. Thus, as these values were quite low, they were multiplied by a hundred times again to make them as realistic as possible.

Last but not least, it is also important to mention that this database represents only the mandatory civil liability coverage.

Table 4.1 - Database Information

	Total	Electric vehicle	Non-electric vehicles
Number of Claims	224 205	886	223 319
Total Amount	2 152 838 847€	7 182 358€	2 145 656 488€

Source: Author's Preparation

4.2. DATASET TREATMENT

The pure data was collected from the data warehouse of the insurer in question. As is normal in large databases, the quality of the data was not the best, and therefore this database had to be treated to achieve consistent results.

In the database, each line represented the annuity for each policy, that is, in one line we could have more than one claim, and then the cost variable was the sum of the costs divided by the number of claims there were in that annuity.

To work only with claims and to be able to have information related to each claim individually, I only considered observations with just one claim.

Of the 28 variables present in the original database, several variables were removed due to a lack of quality or a high correlation with other variables present. Two additional variables were created to assist in this study: the district and zone of the district, based on the zone variable. To carry out this study, the database was divided into two subsets, one with information relating to traditional vehicles and the other only for electric/hybrid cars.

4.2.1. MISSING VALUES

In this database there were many missing values, values with the value -1 which here corresponded to no information and some values that did not make sense for the respective variable and which were therefore removed from the database used to model the data. Regarding the claim cost variable, only claims where this variable was greater than €5 were considered.

For the variables “driver’s age” and 'years of driver's licence', a different approach was taken due to the significant lack of information available for these variables. Thus, it was assumed in these cases that the driver obtained his license at the age of 18, and these fields were calculated accordingly for observations where one of these fields was filled in, in order to obtain the other. After all this treatment, our database was left with 94 852 observations.

4.3. DESCRIPTIVE ANALYSIS

The response variable in the severity model is the claim cost, a continuous variable collected directly from the dataset. This section presents a descriptive analysis of the variables, with the aim of better describing and understanding them, which will be highly valuable and useful in developing the models.

Given the wide range of trait variables considered, the descriptive analysis in this article will focus on the variables that are likely to prove significant for the final models, considering previous models developed by the insurer and the typical behavior of the insurance market. For the graphics below, to compare non-electric vehicles with electric ones, the y-axis represents the percentage of total costs.

4.3.1. CLAIM AMOUNT

In Table 4.2 basic statistics pertaining to this variable are shown. For both vehicle types.

The lower frequency of claims for electric vehicles (EVs) compared to non-electric vehicles (non-EVs) can create a data imbalance, making it challenging to identify statistically robust patterns for EVs and reducing the model’s reliability for this subgroup. This imbalance also increases variability in severity estimates for EVs, leading to higher standard errors and complicating direct comparisons with non-EVs.

Table 4.2 - Elementary descriptive statistics of Claim Amount

	Number of claims with costs	Average Cost per claim	Standard Deviation of Cost per claim
Non-electric vehicle	80 338	3431.31€	1876.58€
Electric vehicles	502	3644.97€	1907.81€

Source: Author's Preparation

Table 4.3 illustrates the right-skewed distribution of claim costs, with 99% of claims having a cost below 8999.6€ for the non-electric vehicles and below 8629.8€ for electric vehicles.

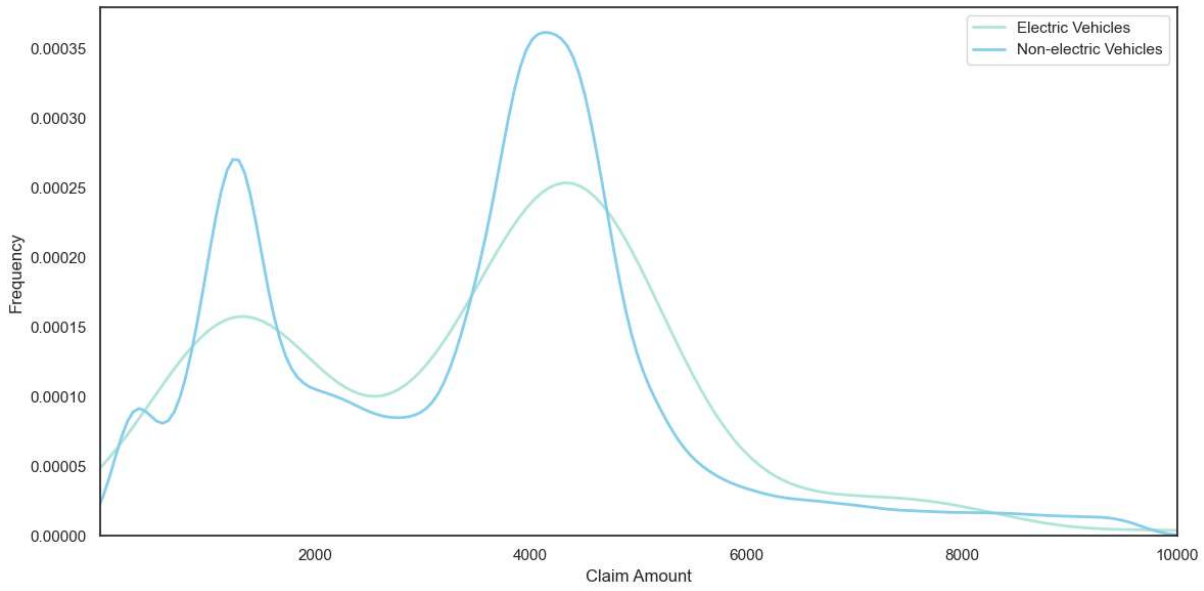
Table 4.3 - Quantiles of Claim Amount

	Min	25%	50%	75%	90%	92,5%	95%	99%	Max
Non-electric vehicle	9.1€	1661.4€	3802.7€	4440.0€	5361.3€	5891.9€	6755.5€	8999.6€	9754.1€
Electric vehicles	137.2€	2151.1€	3920.7€	4648.9€	5727.7€	6439.9€	7164.4€	8629.8€	9430.4€

Source: Author's Preparation

This threshold allows for the visualization of the claim's distribution in Figure 4.1 for both vehicle types.

Figure 4.1 - Claim Amount Density for both vehicle's type



Source: Author's Preparation

As expected, the density is much lower for claims with higher costs. The distribution for non-electric vehicles (blue line) has a higher density in the lower claim amounts (around 2000€-3000€) compared to electric vehicles. For electric vehicles (green line), the density appears

more spread out, with a peak around 4000€ and a slower decrease as the claim amounts increase.

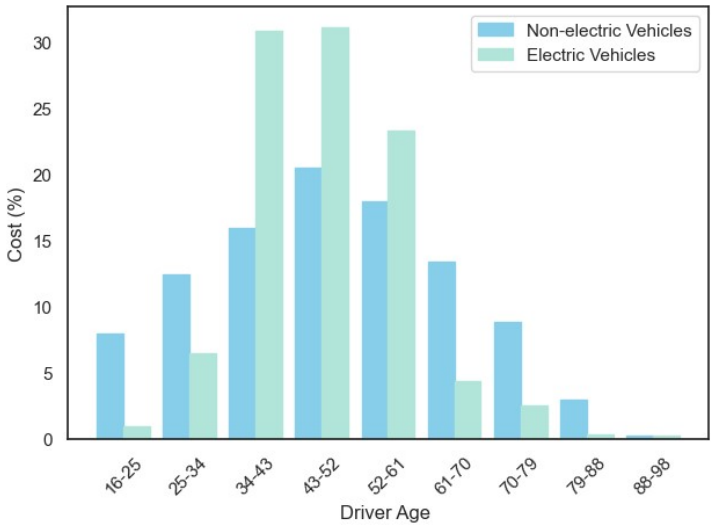
Non-electric vehicles show a sharper drop in density after the initial peak, suggesting more concentrated lower claims, whereas electric vehicles show a smoother and wider spread. This could indicate differences in claim patterns or costs between the two categories. In the sections below, we will observe the behavior of our independent variables.

4.3.2. DRIVER'S AGE

The discrete quantitative variable "Driver Age" ranges from 16 to 98 years, with an average age of 51 years for non-electric vehicles and 48 years for electric vehicles. In the figure below, we can observe how the average cost varies with the driver's age.

The driver's age, as shown in the figure below, was grouped into several categories. This segmentation was performed based on a statistical analysis of the data and its behavior.

Figure 4.2 - %Cost vs Driver's Age



Source: Author's Preparation

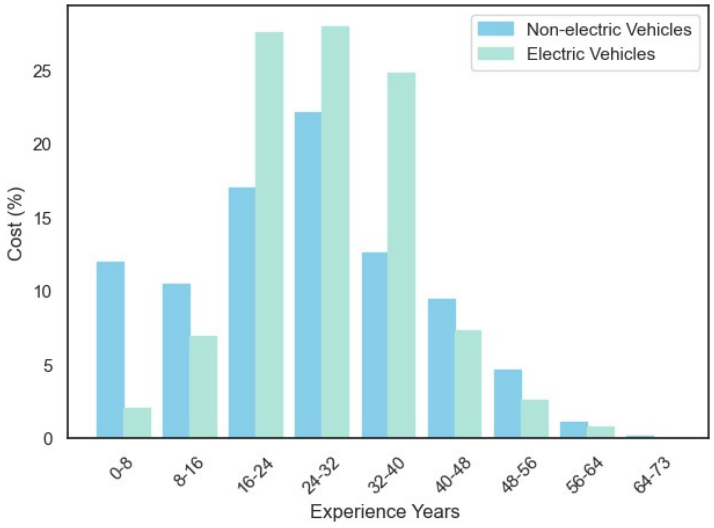
In adulthood, costs are higher, especially for electric vehicles. Reduced costs for older age groups (61-70, 70-79, etc.) in both vehicle types suggest that these drivers are involved in fewer high-cost incidents, which could be attributed to safer driving habits or lower accident rates among older drivers.

4.3.3. DRIVING EXPERIENCE

The discrete quantitative variable "Driving Experience" spans from 0 to 73 years old, with an average age of 26 years within the portfolio, for both vehicle types.

Once again, this segmentation was performed through statistical analysis of the variable and its behavior in relation to the cost.

Figure 4.3 - %Cost vs Experience Years



Source: Author's Preparation

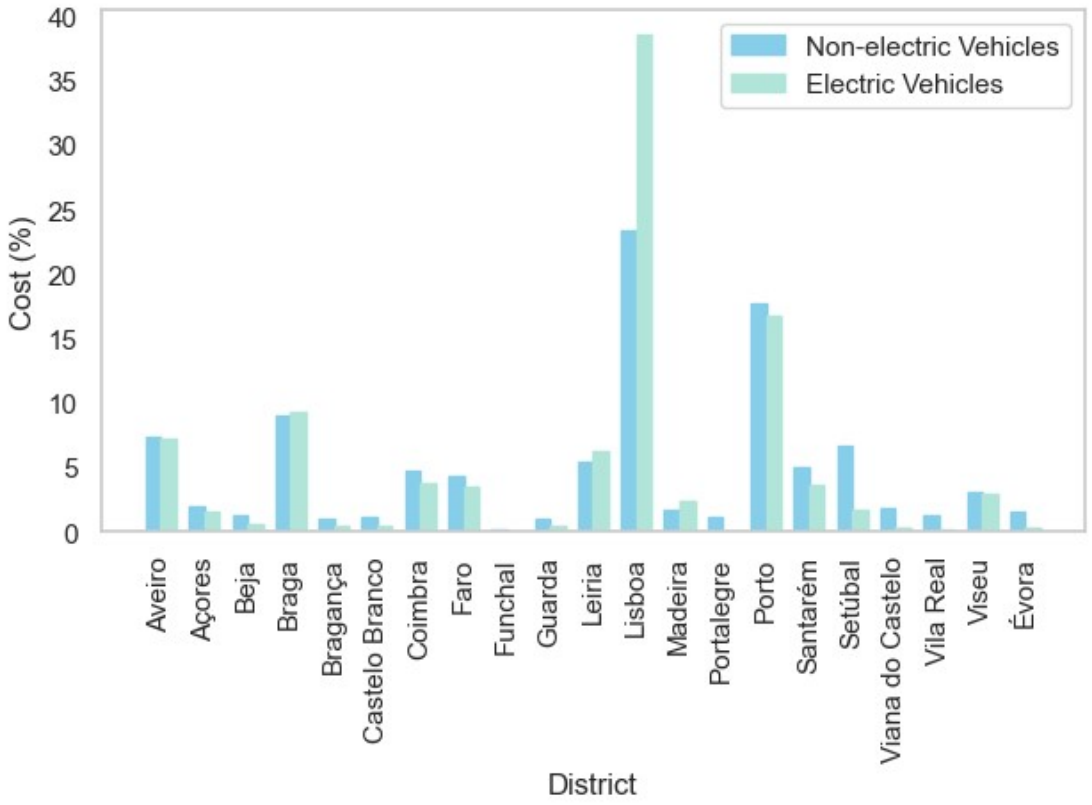
Through the graph, we can observe that for drivers with experience between 16 and 40 years, the average cost is higher. Costs are highest for both vehicle types in the 16–24 and 24–32 year intervals, suggesting that these groups may represent higher risk. Non-electric vehicles exhibit higher costs in the 0–8 and 8–16 year intervals, whereas electric vehicles surpass non-electric ones in the 16–24 and 24–32 intervals. Beyond 32 years of experience, costs decline significantly for both groups, with non-electric vehicles generally having slightly higher costs in the upper intervals. The chart highlights distinct cost patterns and potential differences in risk profiles between vehicle types, particularly at mid-range experience levels.

4.3.4. DISTRICT

From the variable 'zona,' the variable 'Distrito' was created to obtain a feature with less granularity.

In the below figure, we can observe how the average cost of accidents behaves in relation to the district where they occurred.

Figure 4.4 - %Cost vs District



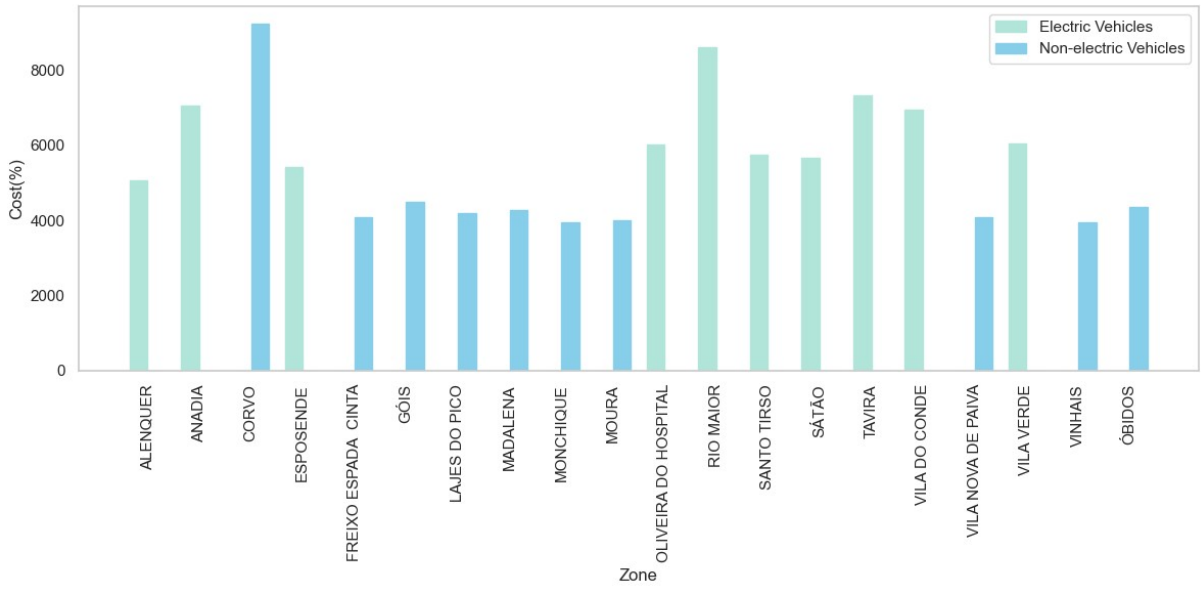
Source: Author's Preparation

We can quickly see that for both datasets, the district of Lisbon shows the highest average costs, followed by Porto.

4.3.5. ZONE

In the figure below, we can observe the 10 zones with the highest average costs for both electric and non-electric cars. We can also see that there are no shared zones between the two datasets.

Figure 4.5 - Top 10 zones with the highest average costs

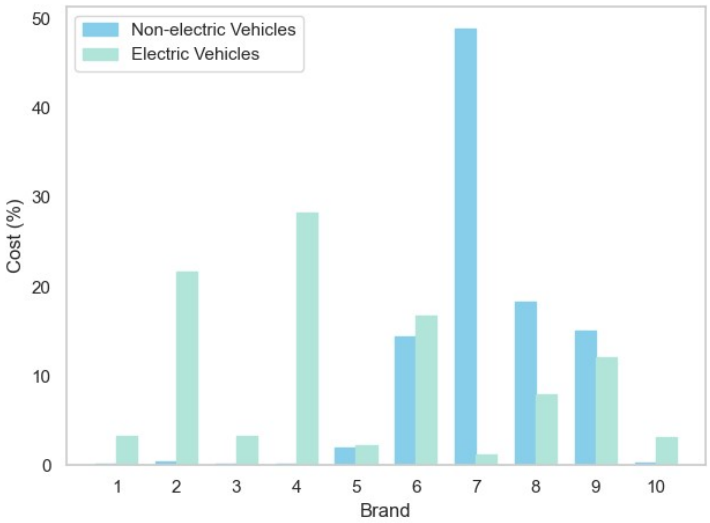


Source: Author's Preparation

4.3.6. VEHICLE BRAND

Vehicle Brand is a categorical variable that represents the brand of the insured vehicle. This variable was created to facilitate the study, since there were several distinct values for the vehicle brand variable. Brands were grouped according to the vehicle's value: 'lower-cost' brands are in group 1, while 'luxury' brands are in group 10. From the chart below, we can conclude that higher-cost brands are associated with higher claim costs. For non-electric cars, the category with the highest cost in our dataset includes brands where the vehicle's value is above €12,000 and below €19,000, which corresponds to brand category 8.

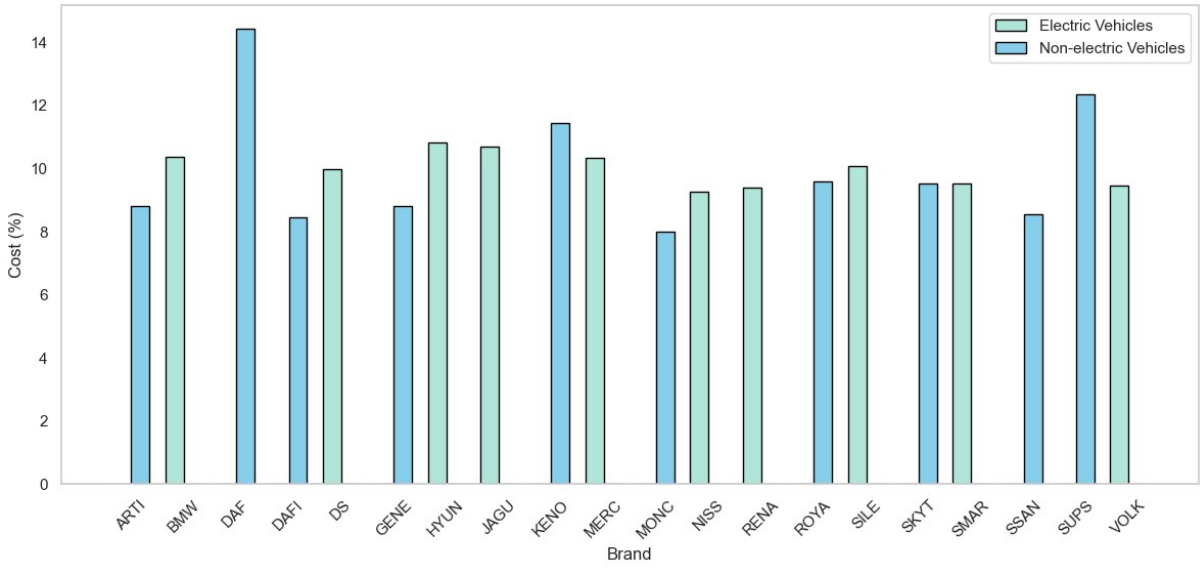
Figure 4.6 - %Cost vs Brand



Source: Author's Preparation

Similarly to the variable zona, the chart below shows the top 10 brands with the highest average cost for electric and non-electric cars.

Figure 4.7 - Top 10 brands with the highest average costs

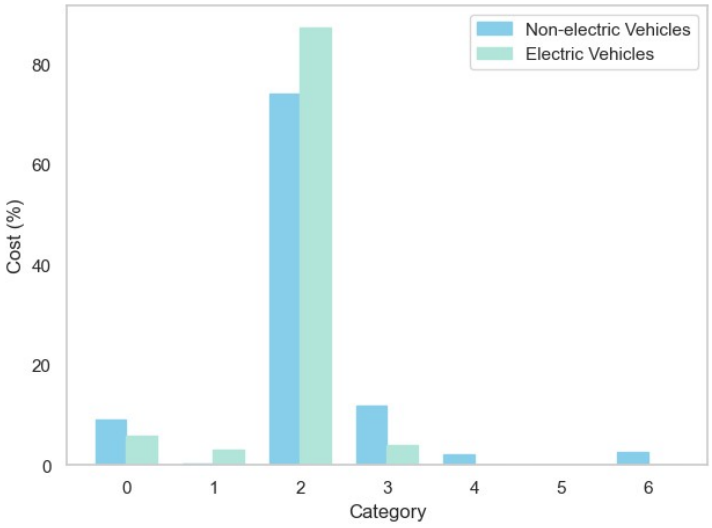


Source: Author's Preparation

4.3.7. VEHICLE TYPE

As expected, the category representing the highest costs in our dataset is category 2, which corresponds to passenger cars. This occurs because it also represents the majority of our insured vehicles.

Figure 4.8 - %Cost vs Category



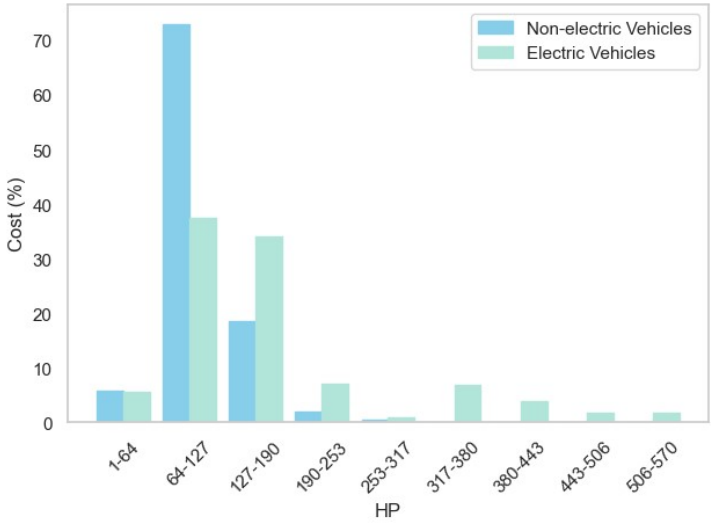
Source: Author's Preparation

Where each value corresponds respectively to: 0: 'Camionetas', 1: 'Ciclomotores', 2: 'Ligeiros', 3: 'Mistos', 4: 'Motociclos', 5: 'Pesados', 6: 'Pickup'.

4.3.8. HORSEPOWER

The chart below shows that, on average, cars with low to medium power—those in a more economical range—have higher costs. We can also observe that electric vehicles tend to have higher power.

Figure 4.9 - %Cost vs Horsepower

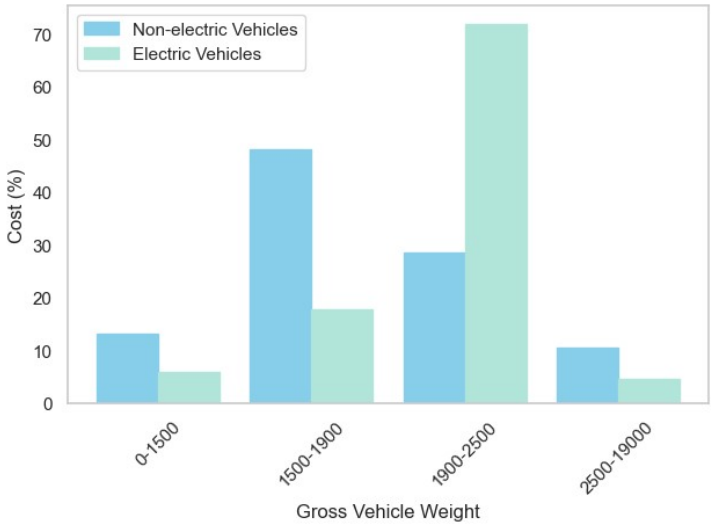


Source: Author's Preparation

4.3.9. GROSS VEHICLE WEIGHT

The chart below shows the distribution of the gross vehicle weight, for both electric and non-electric cars, by average cost. We observe that most accidents involved vehicles with a gross weight between 1500 and 2500 kg, which are small to medium-sized vehicles.

Figure 4.10 - %Cost vs Gross Vehicle Weight

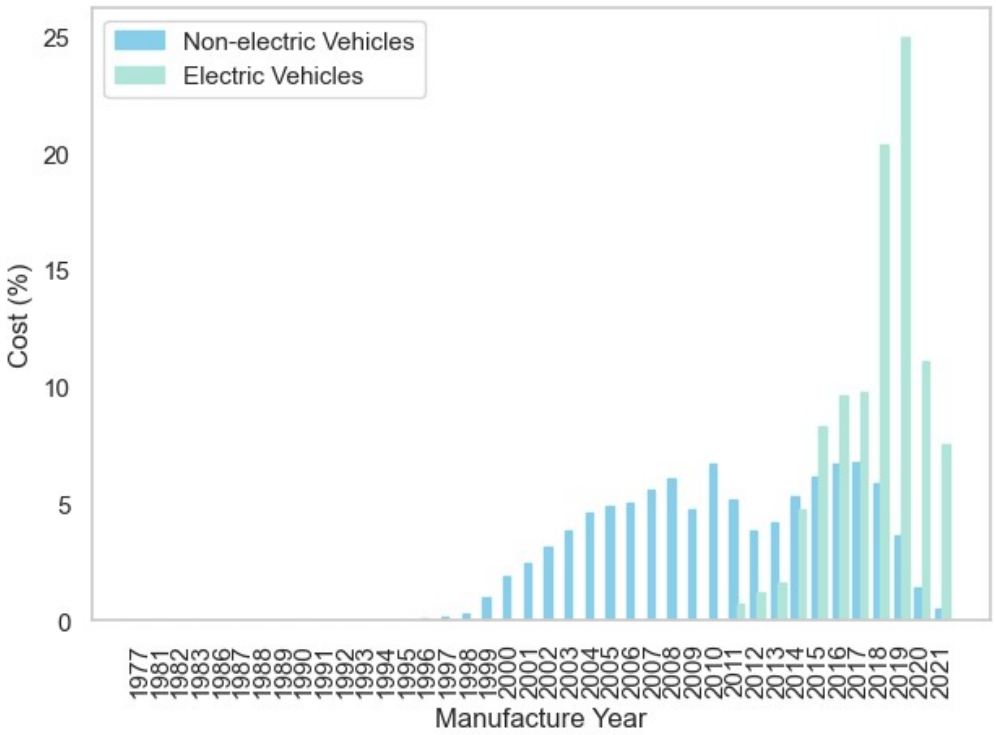


Source: Author's Preparation

4.3.10. MANUFACTURE YEAR

Regarding the construction year of the vehicles in our dataset, we observe that, in general, they are vehicles from the year 2000 onwards. As expected, electric vehicles tend to be newer than non-electric vehicles.

Figure 4.11 - %Cost vs Manufacture Year



Source: Author's Preparation

5. RESULTS AND DISCUSSION

This chapter focuses on presenting the results from the models, including an analysis of data distribution, model performance, and the interpretation of key findings. The results of the claim severity models will be displayed, examining how different variables influence the severity of accidents for both electric and non-electric vehicles. The analysis will provide a detailed overview of the model outcomes, highlighting their implications for risk assessment and future model improvements.

For the models created in this section, categorical variables were used, and all variables were transformed into categorical ones in order to optimize the results of the models. Here is a summary of these variables:

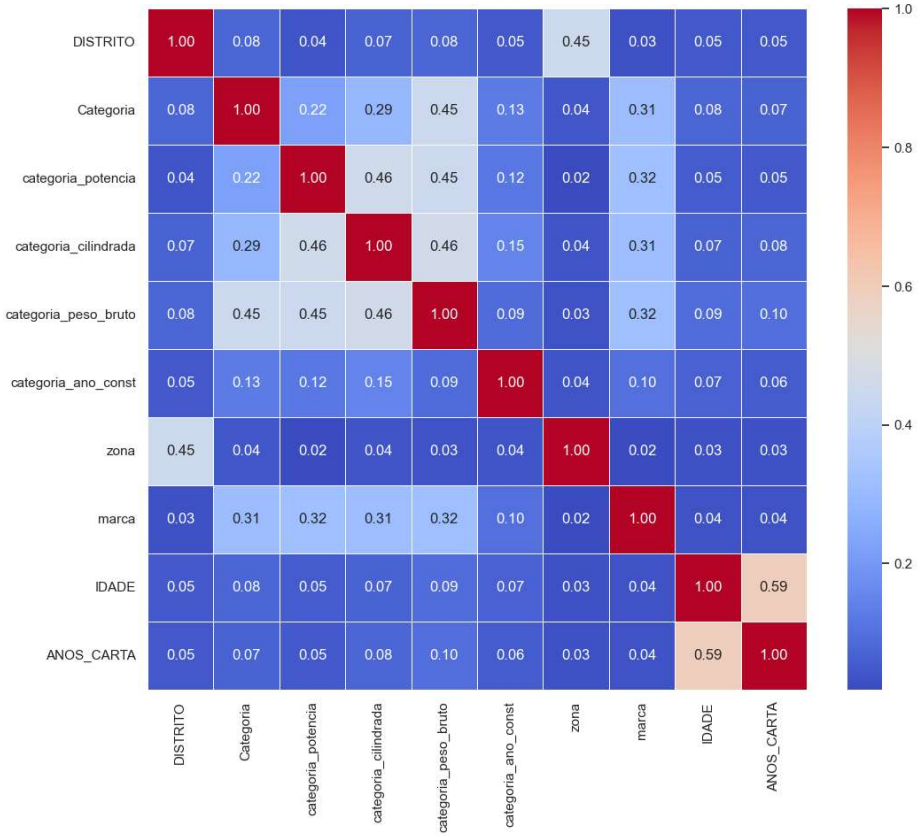
Table 5.1 - Categorical variables

CATEGORICAL VARIABLE	VALUES
CATEGORIA_POTENCIA	'Até 75 cv', '76-100 cv','101-150 cv','151-200 cv', '> 200 cv'
CATEGORIA_CILINDRADA	'Até 1200', '1201-1500','1501-2000', '> 2000'
CATEGORIA_PESO_BRUTO	'Até 1500 kg', '1501-2000 kg','2001-2700 kg', '> 2700 kg'
CATEGORIA_ANO_CONST	'Até 2006', '2007-2013','2014-2020', '> 2020'
MARCA	1-10
IDADE	1 (ages between 16 and 30), 2(ages between 31 and 50), 3 (ages between 51 and 60), 4 (ages between 61 and 70) and 5 above 70 years
ANOS_CARTA	1 (experience under 5 years), 2 (experience between 6 and 15 years), 3 (experience between 16 and 25 years), 4 (experience between 26 and 35 years), 5 (experience between 36 and 45 years), 6 (experience between 46 and 55 years) and 7 above 56

5.1. CORRELATION MATRIX

Before modeling the data, a correlation matrix was created to check for highly correlated variables, for both electric and non-electric vehicles, as this can impact the models later on. In the figure below, we can see the result of the correlation matrix for non-electric vehicles based on Cramér's V values, which measures the strength of association between pairs of categorical variables for non-electric vehicles.

Figure 5.1 - Correlation Matrix for Non-Electric Vehicles



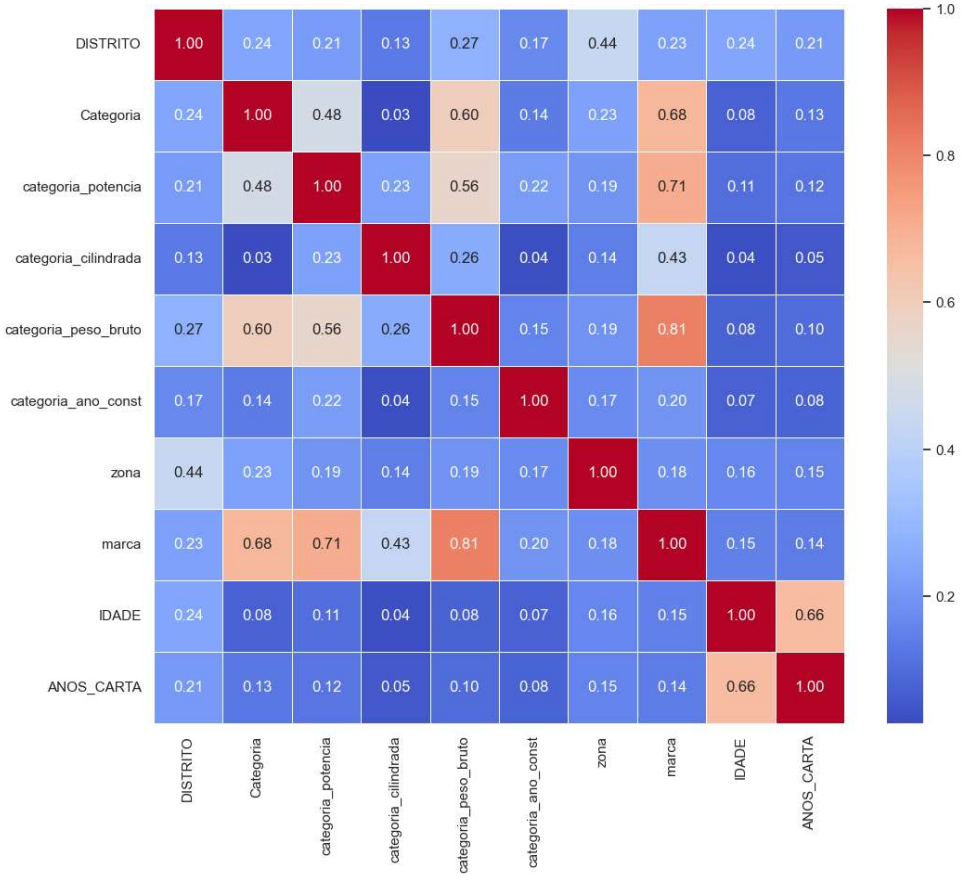
Source: Author's Preparation

There are several notable high correlations, such as:

- *categoria_potencia* and *categoria_cilindrada* (0.46), indicating a moderate correlation between power and cylinder category.
- *categoria_peso_bruto* and *categoria_potencia* (0.45), suggesting a moderate correlation between weight and power categories.
- *IDADE* and *ANOS_CARTA* (0.59), indicating a relatively high correlation between age and experience years.

For electric vehicles, the same graphic was built.

Figure 5.2 - Correlation Matrix for Electric Vehicles



Source: Author's Preparation

In the graph above, we can observe that the variable *categoria* shows a strong correlation with the variables *marca* and *categoria_peso-bruto*, as well as with the variable *categoria_potencia*.

Variable *categoria_peso_bruto* also shows a strong correlation with the variable *marca*. Additionally, the variables *ANOS_CARTA* (years of driving license) and *IDADE* (age) exhibit a strong correlation with each other as well.

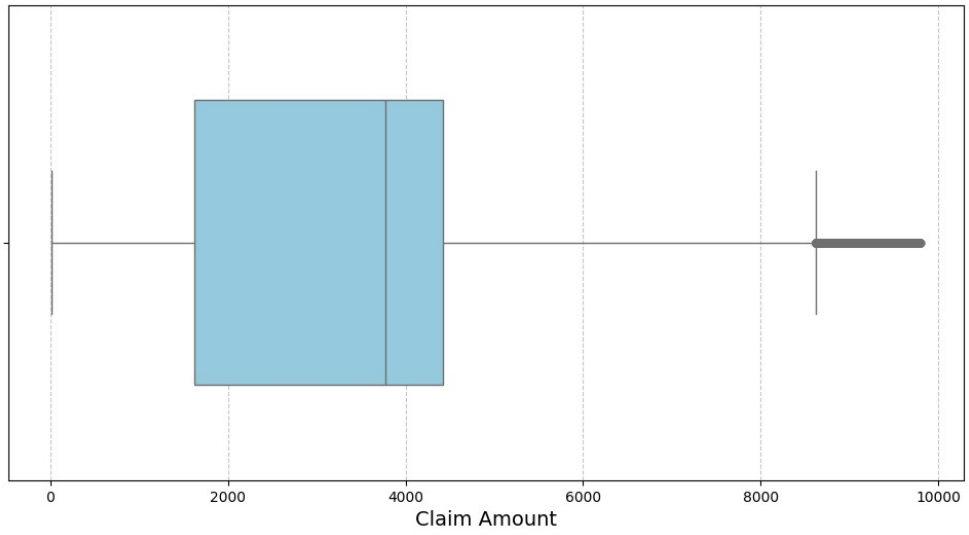
5.2. FITTING THE DISTRIBUTIONS

Typically, the response variable claim amount follows a distribution from the Exponential family. To understand how well my data fit this distribution, I used the AIC method and also attempted various transformations on the data to achieve better results.

5.2.1. NON-ELECTRIC VEHICLES

For this dataset, the claim amount behaves as shown below:

Figure 5.3 - Boxplot of Claim Amount for Non-Electric Vehicles



Source: Author's Preparation

The boxplot shows a slight right tail, suggesting that the distribution of the Claim Amount is slightly skewed with a tendency towards higher values due to the presence of outliers.

After calculating the AIC values for various transformations (square root, Box-Cox, and logarithmic), the best results were obtained with the logarithmic transformation, yielding the following results:

Table 5.2 - AIC value for different distributions with logarithmic transformation, for Non-Electric Vehicles

AIC (Gaussian)	169 295,69
AIC (Inv_Gaussian)	1 916 298,38
AIC (Gamma)	173 599,12

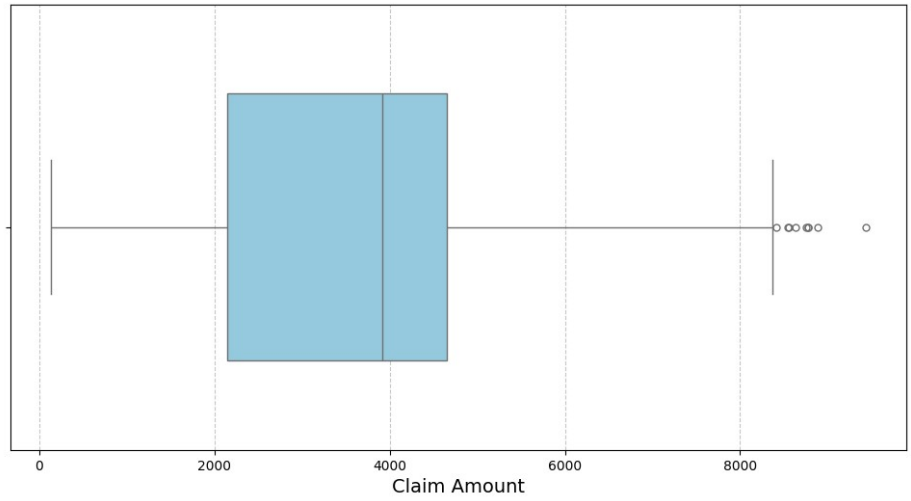
Source: Author's Preparation

Upon analyzing the table above, we conclude that the distribution that presents the lowest value is the Normal distribution with a logarithmic transformation to the response variable, making it the distribution that best fits the data.

5.2.2. ELECTRIC – VEHICLES

For electric vehicles, the claim amount variable exhibits the following behavior:

Figure 5.4 - Boxplot of Claim Amount for Electric Vehicles



Source: Author's Preparation

The distribution of our dependent variable Claim Amount for non-electric cars is relatively symmetrical, with most values concentrated between 2000 and around 4000, and there are no visible outliers that deviate significantly from this range.

The same reasoning was applied to the electric vehicle dataset, where the best results were obtained without any transformation, using the original data. The distribution that best fits, according to the AIC value, is also a Normal distribution with a logarithmic transformation to the response variable, as we can confirm from the table below.

Table 5.3 - AIC value for different distributions with logarithmic transformation, for Electric Vehicles

AIC (Gaussian)	9 005,22
AIC(Inv_Gaussian)	12 834,68
AIC (Gamma)	9 006,22

Source: Author's Preparation

5.3. GENERALIZED LINEAR MODELS

We model the data using generalized linear models for both datasets. Based on the variable *DISTRITO*, we created the variable *zona_distrito*, which is broader and allows for fewer possible values, that is, we were able to group the values into larger sets with lower granularity. This variable can take the following values: Alentejo, Algarve, Centro, Insular, Interior, Metropolitana, and Norte. To use this new variable along with the other categorical variables in the models, dummy variables were created with values of 0 or 1 for each of the categories. For each variable, the reference category is the most frequent.

The Python module statsmodels was used for this models.

Table 5.4 summarizes the reference categories. Lastly, the datasets were divided into 80% for training and 20% for testing.

Table 5.4 - Reference category per feature

VARIABLE	NON-ELECTRIC VEHICLES	ELECTRIC-VEHICLES
MARCA	Group *7*	Group 4
DISTRITO	Group *Lisboa*	Group *Lisboa*
ZONA_DISTRITO	Group *Zona Metropolitana*	Group *Zona Metropolitana*
ANOS_CARTA	Group 26-35	Group 26-35
CATEGORIA	Group Ligeiros	Group Ligeiros
POT	Group *101-150 cv*	Group *151-200 cv*
CC	Group *1501-2000*	Group *Até 1200*
PB_VEIC	Group *1501-2000 kg*	Group *2001-2700 kg*
ANO_CONST	Group *2007-2013*	Group *2014-2020*
IDADE	Group 31-50	Group 31-50

Source: Author's Preparation

5.3.1. NON-ELECTRIC

To model identification strategy required testing first with all the candidate features to get an initial general idea of how they behave and check if they are significant. For this, the stepwise regression method was used, and the chosen type was backward elimination, this method was chosen considering the number of variables that exist. We removed the variables that are not significant for the model, meaning those with a p-value greater than 5%, and we are left with:

Table 5.5 - GLM Regression Results for Non-Electric Vehicles

	Coef	*Std err*	*Z*	*P>|z|*	*[0.025*	*0.975]*
Const	2.0710	0.001	2782.453	0.000	2.070	2.072
DISTRITO_Aveiro	0.0071	0.001	4.752	0.000	0.004	0.010
DISTRITO_Açores	-0.0106	0.003	-3.913	0.000	-0.016	-0.005
DISTRITO_Braga	0.0079	0.001	5.734	0.000	0.005	0.011
DISTRITO_Funchal	-0.0236	0.010	-2.371	0.018	-0.043	-0.004
DISTRITO_Portalegre	-0.0099	0.004	-2.775	0.006	-0.017	-0.003
DISTRITO_Porto	0.0020	0.001	1.947	0.052	-1.34e-05	0.004
DISTRITO_Santarém	-0.0079	0.002	-3.432	0.001	-0.012	-0.003
DISTRITO_Viseu	-0.0079	0.003	-3.015	0.003	-0.013	-0.003
Categoria_Camionetas	0.008	0.002	4.722	0.000	0.005	0.011
Categoria_Ciclomotores	-0.0209	0.007	-2.893	0.004	-0.035	-0.007

Categoria_Mistos	0.0082	0.001	6.286	0.000	0.006	0.011
Categoria_Motociclos	-0.0141	0.003	-5.404	0.000	0.019	-0.009
Categoria_Pickup	0.0112	0.003	3.682	0.000	0.005	0.017
Categoria_peso_bruto_> 2700 kg	0.0033	0.002	1.787	0.074	-0.000	0.007
Categoria_ano_const_2014-2020	-0.0040	0.001	-5.201	0.000	-0.006	-0.003
IDADE_61-70	-0.0023	0.001	-2.159	0.031	-0.004	-0.000
ANOS_CARTA_0-5	0.0026	0.001	1.911	0.056	-6.81e-0.5	0.005
ANOS_CARTA_16-25	-0.0033	0.001	-3.643	0.000	-0.005	-0.002
zona_distrito_Zona Centro	0.0041	0.002	2.372	0.018	0.001	0.007

Source: Author's Preparation

The table below divides the variables according to the sign of the coefficient obtained for each of them:

Table 5.6 - Significant Variables chosen in the GLM severity model for Non-Electric Vehicles

Variables with Positive Effect	*Variables with Negative Effect*
DISTRITO_Aveiro	*DISTRITO_Açores*
DISTRITO_Braga	*DISTRITO_Funchal*
DISTRITO_Porto	*DISTRITO_Portalegre*
Categoria_Camionetas	*DISTRITO_Santarém*
Categoria_Mistos	*DISTRITO_Viseu*
Categoria_Pickup	*Categoria_Ciclomotores*
Categoria_peso_bruto_> 2700 kg	*Categoria_Motociclos*
ANOS_CARTA_0-5	*Categoria_ano_const_2014-2020*
zona_distrito_Zona Centro	*IDADE_61-70*
	ANOS_CARTA_16-25
	zona_distrito_Zona Algarve
	zona_distrito_Zona Interior

Source: Author's Preparation

Through this summary of the model results, we can draw several conclusions.

It is important to note that since a logarithmic transformation was applied to the data, the coefficient presented above in Table 5.5 was also affected by this adjustment.

Variables with a positive effect, with a coefficient above zero, mean that, for example, for variable *Distrito*, it indicates that, in relation to their reference variable (which is DISTRITO_Lisboa), this category has higher costs.

For example, in the case of the categorical variable DISTRITO_Aveiro, it means that being from this district increases the expected value by 0.0071 compared to the reference variable. The same reasoning applies to variables with a negative coefficient, but in that case, the expected value decreases compared to the reference variable.

For example, DISTRITO_Funchal has a negative coefficient wich means that being from this district decreases the expected value by 0.0236 compared to the reference variable. This can be related to reduced traffic intensity or different driving patterns.

The results of the GLM model demonstrate that various geographic, demographic, and vehicle characteristics significantly influence claim costs. Districts with higher urban density, such as Aveiro and Braga, showed slightly higher costs, while more isolated regions, such as the Azores and Funchal, had lower costs, reflecting differences in traffic intensity and risk exposure. Heavier vehicles (>2700 kg) were associated with marginally higher costs, consistent with their potential to cause greater damage in accidents. We can also observe that newer vehicles (2014 to 2020) and more experienced drivers (between 16 and 25 years) contributed to lower costs, aligning with the expectation of reduced risks for these groups.

Specific vehicle categories also exhibited significant impacts, with some categories indicating higher costs due to the severity of incidents. These findings are consistent with previous studies highlighting the influence of regional, demographic, and vehicle characteristics on claim costs, underscoring the importance of modeling these variables to improve risk prediction and management in automotive insurance.

Therefore, we obtained a severity model with these 21 variables. The results of the model indicate that it does not explain the intended variable well, which is not uncommon for this type of model. This is because, for example, it is generally more challenging to explain the cost of claims compared to frequency.

According to previous studies modeling claim severity is inherently more challenging than frequency due to the nature and variability of the data. Severity distributions are typically highly asymmetric highly skewed, with a long tail of high-cost claims, making them difficult to model and interpret.

5.3.2. ELECTRIC VEHICLES

Similarly to what was done for non-electric vehicles, testing the model with all the variables was the first thing to do and only then removed the variables that were not significant but, in this case, a significance level of 10% was used due to the reduced sample size.

For electric vehicles, unlike what happened with traditional vehicles, a better result was obtained in the model without using the created categorical variables. Therefore, the only variables used as categorical were those that were originally categorical.

Table 5.7 - GLM Regression Results for Electric Vehicles

	Coef	*Std err*	*Z*	*P>|z|*	*[0.025*	*0.975]*
Const	8.4947	4.639	1.831	0.067	-0.598	17.587
ANO_CONST	-0.0031	0.002	-1.385	0.166	-0.008	0.001
DISTRITO_Leiria	0.0356	0.019	1.848	0.065	-0.002	0.073
DISTRITO_Setúbal	-0.0571	0.045	-1.274	0.203	-0.145	0.031
Marca_6	0.0292	0.013	2.307	0.021	0.004	0.054
Marca_8	0.0334	0.018	1.857	0.063	-0.002	0.069

Source: Author's Preparation

The table below divides the variables according to the sign of the coefficient obtained for each of them:

Table 5.8 - Significant Variables chosen in the GLM severity model for Electric Vehicles

Variables with Positive Effect	*Variables with Negative Effect*
DISTRITO_Leiria	*ANO_CONST*
marca_6	*DISTRITO_Setúbal*
Marca_8	

Source: Author's Preparation

The interpretation of the variables follows the same logic as previously explained.

Categorical variables related to vehicle brands stand out in the model. The brands in group 6 have a positive coefficient (0.0292) and are statistically significant (P>|z| = 0.021), suggesting that these brands may be associated with a higher severity in electric vehicles.

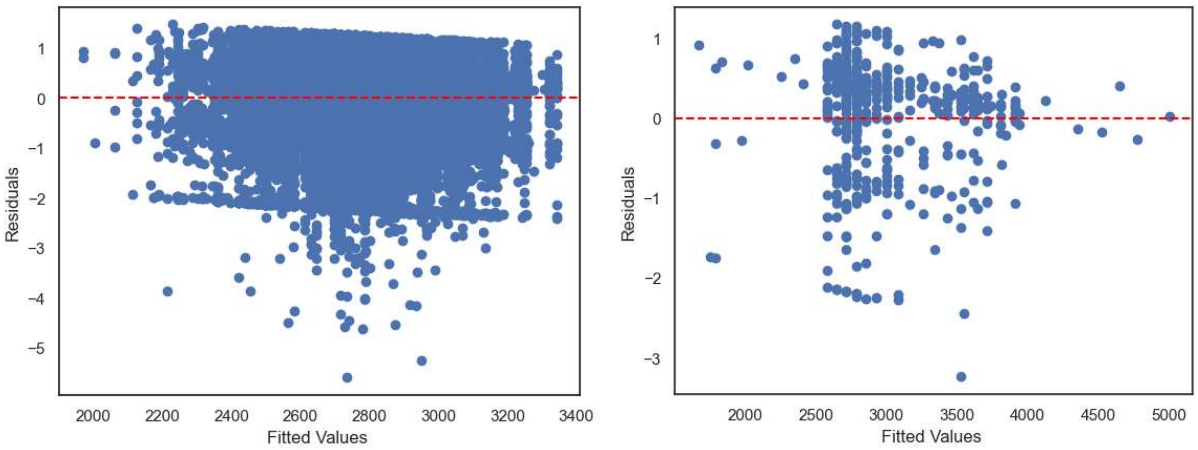
Brands in group 8 also have a positive coefficient (0.0334) and are marginally significant (P>|z| = 0.063), indicating a potential influence, though with a lower confidence.

It is also important to remember that our data on electric vehicles is a very small sample, which negatively impacts on the reliability of the results.

5.3.3. RESIDUALS

The residuals were calculated for each dataset, as represented below. This value was obtained by taking the difference between the observed value and the value predicted by the model.

Figure 5.5 - GLM Fitted Values Residuals



Source: Author's Preparation

The residuals for non-electric vehicles (left graph) range between -5 and 1, but they are more concentrated between -2 and 1.

In the right graph, pertaining to the electric vehicle dataset range between -3 and 1, we see a large dispersion of values due to the lack of data. However, we can observe that most observations are closer to zero.

The residual analysis suggests that the model aligns with its assumptions, particularly for non-electric vehicles with no discernible trend or pattern. This consistency indicates that the assumptions of homoscedasticity and independence are reasonably met.

High residual values reflect instances where observed claim costs exceed predictions, potentially due to unaccounted factors like extreme claims. Low residual values indicate overpredictions, suggesting that the model may not fully capture certain patterns. For electric vehicles a larger dispersion is observed due to the limited dataset, indicating potential challenges in accurately modeling claim costs in this group.

5.4. LOGISTIC REGRESSION

A new variable was created in the dataset based on the cost of the claim, where costs above 4 500€ are considered severe. Thus, this variable *acidente_grave* takes on only 'yes' and 'no' values, which are represented as 1 and 0, respectively.

The goal is to understand the probability of a claim resulting in a severe accident and to analyze the differences between electric and non-electric cars.

The Python library scikit-learn was used for this models.

5.4.1. NON-ELECTRIC VEHICLES

5.4.1.1. VARIABLE SELECTION

The variable selection was made based on the importance of each variable for the model, meaning that variables with a p-value below 5% were selected. Alongside this analysis, the VIF value for each variable was also considered, and those with excessively high values were removed.

Thus, the variables included in the final model are as follows

Table 5.9 - Python Output for Significant Variables for Non- Electric Vehicles

	Coef	*Std err*	*Z*	*P>|z|*	*[0.025*	*0.975]*
Const	-1.1719	0.020	-58.920	0.000	-1.211	-1.133
DISTRITO_Aveiro	0.1244	0.038	3.247	0.001	0.049	0.200
DISTRITO_Açores	-0.2093	0.076	-2.740	0.006	-0.359	-0.060
DISTRITO_Braga	0.1632	0.035	4.720	0.000	0.095	0.231
DISTRITO_Guarda	-0.1986	0.101	-1.973	0.048	-0.396	-0.001
DISTRITO_Viseu	-0.1340	0.058	-2.291	0.022	-0.249	-0.019
Categoria_Camionetas	0.0752	0.040	1.869	0.062	-0.004	0.154
Categoria_Mistos	0.0932	0.035	2.648	0.008	0.024	0.162
Categoria_Motociclos	0.1349	0.070	1.930	0.054	-0.002	0.272
Categoria_Pickup	0.1017	0.070	1.453	0.146	-0.035	0.239
Categoria_potencia_151-200 cv	0.0517	0.035	1.472	0.141	-0.017	0.120
Categoria_cilindrada_< 1200	-0.0426	0.027	-1.559	0.119	-0.096	0.011
Categoria_ano_const_>2020	0.2221	0.076	2.916	0.004	0.073	0.371
IDADE_16-30	0.0510	0.030	1.690	0.091	-0.008	0.110
IDADE_61-70	-0.0639	0.032	-1.992	0.046	-0.127	-0.001
IDADE_>70	-0.0908	0.043	-2.679	0.007	-0.157	-0.024
ANOS_CARTA_16-25	-0.0682	0.026	-2.645	0.008	-0.119	-0.018
ANOS_CARTA_36-45	-0.0866	0.032	-2.695	0.007	-0.150	-0.024

Source: Author's Preparation

The table below divides the variables according to the sign of the coefficient obtained for each of them:

Table 5.10 - Selected Variables for the Logistical Regression Model

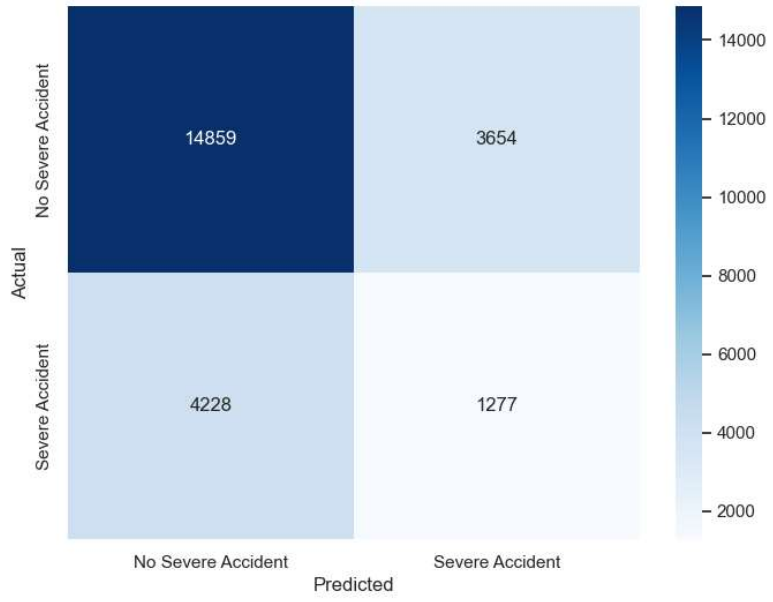
Variables with Positive Effect	*Variables with Negative Effect*
DISTRITO_Aveiro	*ANO_CONST*
DISTRITO_Braga	*DISTRITO_Açores*
Categoria_Camionetas	*DISTRITO_Guarda*
Categoria_Mistos	*Categoria_cilindrada_< 1200*
Categoria_Motociclos	*IDADE_26-35*
Categoria_Pickup	*IDADE_>70*
Categoria_potencia_151-200 cv	*ANOS_CARTA_16-25*
Categoria_ano_const_> de 2020	*ANOS_CARTA_36-45*
IDADE_16-30	*DISTRITO_Viseu*

Source: Author's Preparation

From this output, we can observe for example that higher ages and more driving experience have a negative effect, meaning there is a lower probability of causing a serious accident with these characteristics.

To evaluate the model, the chart below was created:

Figure 5.6 - Confusion Matrix for Non-Electric Vehicles



Source: Author's Preparation

This chart is represented in the table below, where we can see the information in more detail.

Table 5.11 - Classification Table for Non-Electric Vehicles

	Precision	Recall	F1-Score	Support
0	0.78	0.80	0.79	18513
1	0.26	0.23	0.24	5505
Accuracy			0.67	24018
Macro avg	0.52	0.52	0.52	24018
Weighted avg	0.66	0.67	0.67	24018

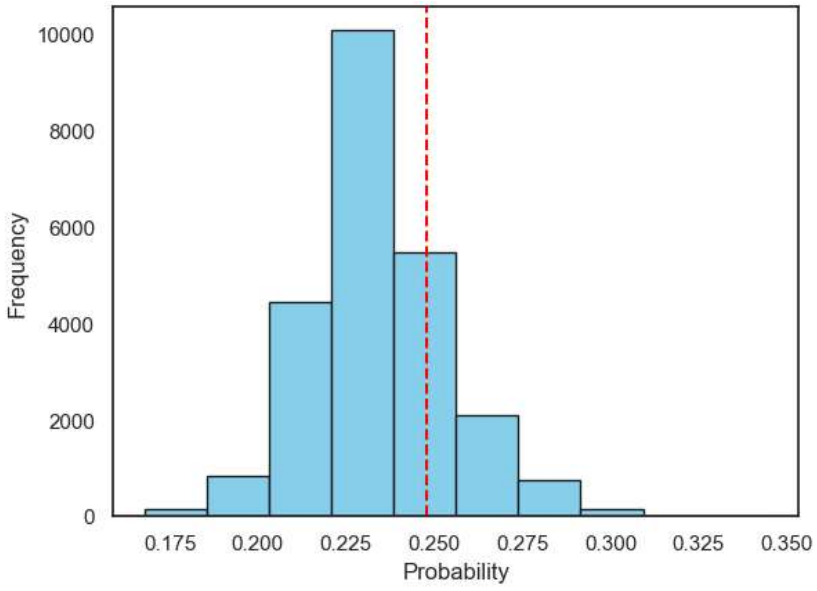
Source: Author's Preparation

This means that the model achieved an accuracy of 67%, indicating that approximately 67% of the model's predictions are correct.

Regarding the model's precision, which is the ratio of true positives to the total predicted positives, the value was 78% for class 0 and 26% for class 1 (with an severe accident). The recall for class 0 is 80%, meaning that the model correctly identifies 80% of the positive cases in this class, while it has more difficulty with class 1. The F1-score reflects the values of precision and recall, further demonstrating that the model performs well for class 0 but struggles significantly with class 1.

Due to the imbalance in our dataset, where there are more observations with the variable *'acidente grave'* equal to 0 than with a value of 1 (indicating a severe accident), the threshold used by the model was set at 0.248. This value was obtained by calculating the proportion of severe accidents in our dataset. Therefore, probabilities above this value were considered as severe accidents, as we can see in the graph below.

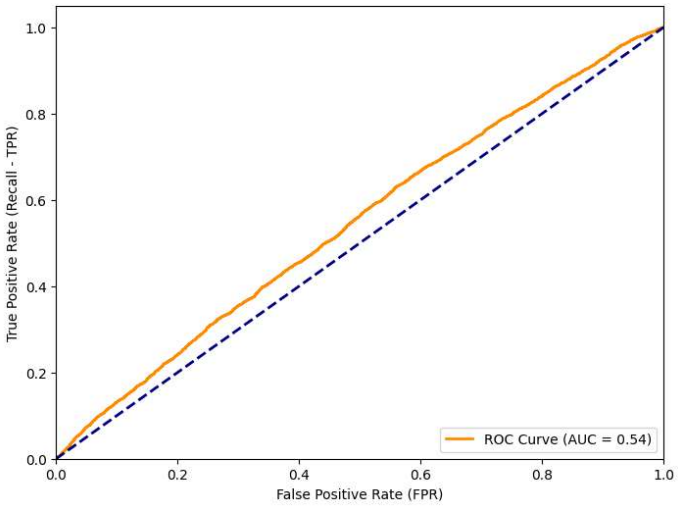
Figure 5.7 - Distribution of Predicted Probabilities for Severe Accidents



Source: Author's Preparation

To observe how the False Positive Rate (FPR) behaves in relation to the True Positive Rate (TPR), a ROC curve was created with the FPR on the x-axis. The area under the curve was calculated as a metric, where the closer this value is to 1, the better our model can predict the results.
The value obtained for our model was 0.54 which indicates that the model has a very limited ability to distinguish between the predicted classes, since an AUC of 0.5 represents a model with no discriminatory power.

Figure 5.8 - ROC Curve for Non-Electric Vehicles



Source: Author's Preparation

5.4.2. ELECTRIC VEHICLES

The variable selection for these vehicles was made based on the importance of each variable for the model, meaning that variables with a p-value below 5% were selected. Alongside this analysis, the VIF value for each variable was also considered, and those with excessively high values were removed.

Table 5.12 - Python Output for Significant Variables for Electric Vehicles

	Coef	*Std err*	*Z*	*P>|z|*	*[0.025*	*0.975]*
Const	-45.1623	117.825	-0.383	0.701	-276.094	185.770
ANO_CONST	0.0219	0.058	0.376	0.707	-0.093	0.136
Marca_9	-1.0670	0.435	-2.454	0.014	-1.919	-0.215
ANOS_CARTA_6-15	0.5925	0.423	1.401	0.161	-0.236	1.421
ANOS_CARTA_26-35	0.3372	0.253	1.335	0.182	-0.158	0.832

Source: Author's Preparation

Thus, the variables included in the final model are as follows:

Table 5.13 - Selected Variables for the Logistical Regression Model

Variables with Positive Effect	*Variables with Negative Effect*
ANO_CONST	*Marca_9*
ANOS_CARTA_6-15	
ANOS_CARTA_26-35	

Source: Author's Preparation

The variable Marca_9, representing a group of more expensive vehicle brands, has a statistically significant negative coefficient (-1.0670, P>|z| = 0.014). This indicates that vehicles from high-end brands are associated with lower odds of severe claims compared to the reference category. A possible explanation is that owners of these more expensive vehicles may adopt safer driving behaviors or have access to advanced safety features, reducing the likelihood of serious consequences.

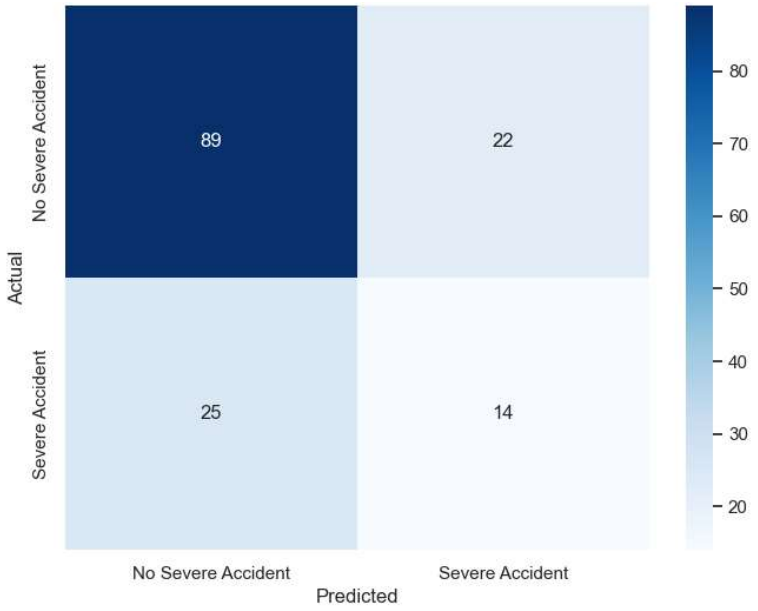
This is consistent with other studies, including international research such as the study by Wu et al. (2020), which also states that vehicles from higher-end brands, such as SUVs, are associated with less severity in accidents.

The other variables have a positive effect, meaning that observations with these characteristics have higher odds of severe claims compared to the reference category.

For example, the variable ANOS_CARTA_4.0 (between 26 and 35 years) with a coefficient of 0.3372 indicates that experienced drivers are associated with higher odds of severe claims compared to the reference category.

Just like for non-electric vehicles, the following chart was also constructed:

Figure 5.9 - Confusion Matrix for Electric Vehicles



Source: Author's Preparation

This chart is represented in the table below, where we can see the information in more detail.

Table 5.14 - Classification Table for Electric Vehicles

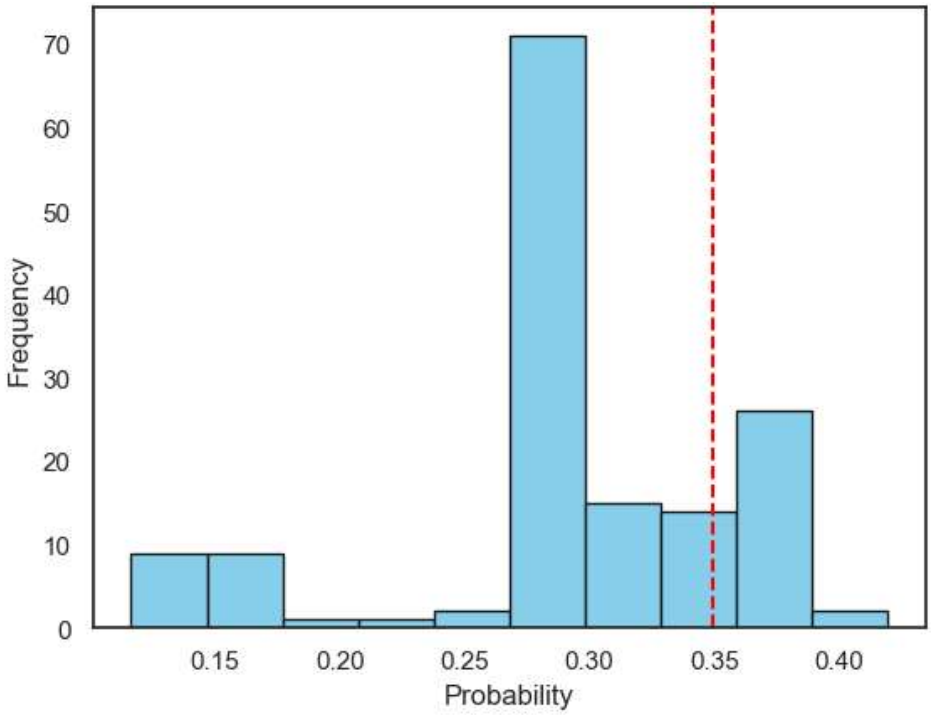
	Precision	Recall	F1-Score	Support
0	0.78	0.80	0.79	111
1	0.39	0.36	0.37	39
Accuracy			0.69	150
Macro avg	0.58	0.58	0.58	150
Weighted avg	0.68	0.69	0.68	150

Source: Author's Preparation

This means that the model achieved an accuracy of 69%, indicating that approximately 69% of the model's predictions are correct, which is reasonable compared to similar studies.

Regarding the precision of the model, which is the ratio of true positives to the total predicted positives, the value was 78% for class 0 and 39% for class 1 (with an severe accident). The recall for class 0 is 80%, which means that the model correctly identifies 80% of the positive cases in this class, while it has more difficulty with class 1.
The F1 score reflects the precision and recall, further demonstrating that the model performs well for class 0 but struggles significantly with class 1.

Figure 5.10 - Distribution of Predicted Probabilities for Severe Accidents

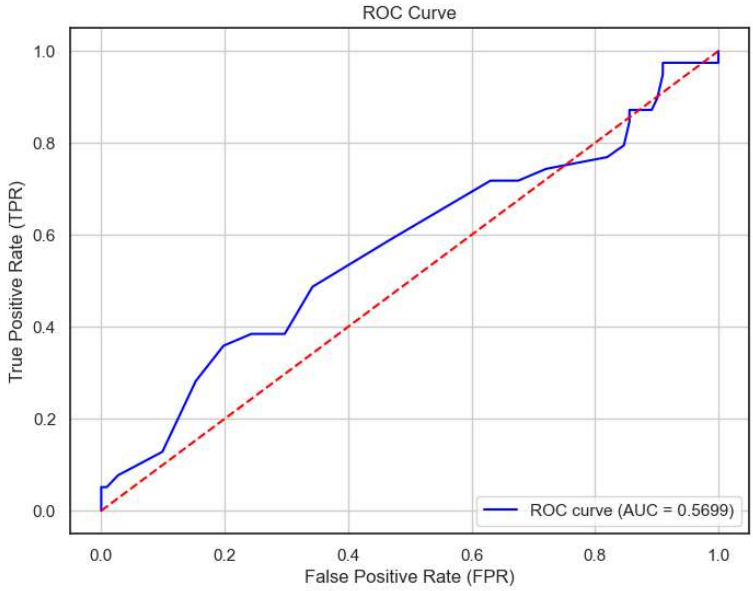


Source: Author's Preparation

Due to the imbalance in our dataset, where there are more observations with the variable *'acidente grave'* equal to 0 than with a value of 1 (indicating a severe accident), the threshold used by the model was set at 0.3495. This value was obtained by calculating the proportion of severe accidents in our dataset, so probabilities above this value were considered severe accidents.

To observe how the False Positive Rate (FPR) behaves in relation to the True Positive Rate (TPR), a ROC curve was created with the FPR on the x-axis. The area under the curve was calculated as a metric, where the closer this value is to 1, the better our model can predict the results. The value obtained for our model was 0.57 witch once again indicates that the model has a very limited ability to distinguish between the predicted classes, since an AUC of 0.5 represents a model with no discriminatory power.

Figure 5.11 - ROC Curve for Electric Vehicles



Source: Author's Preparation

This curve is somewhat atypical, this can be justified by the characteristics of the dataset like class imbalance that typically poses challenges for predictive models, as they tend to favor the majority class, leading to reduced discriminatory power for the minority class. Furthermore, the small sample size amplifies the variability and uncertainty in the predictions, contributing to the shallow curve observed. This limited data volume may also be difficult for the model's ability to capture meaningful patterns or relationships, resulting in predictions that are only slightly better than random guessing. As a result, the observed ROC curve reflects the inherent difficulties associated with modeling severe accidents under these constraints.

5.4.3. COMPARISON TEST

The Z-test for proportions is a statistical test used to determine whether there is a significant difference between two population proportions. It is particularly useful when comparing categorical data.

In this case, the Z-Test compares the probabilities of severe accidents between electric and non-electric cars. The results obtained were as follows:

Table 5.15 - Output of Z-test

	Electric Cars	Non-Electric Cars
Severe accident proportion	0.2337	0.2886
Z-test Statistic	2.8893	
P-value	0.0039	

Source: Author's Preparation

Based on the results of the test, electric cars have a slightly higher probability of causing severe accident compared to non-electric cars.

However, these results should be analyzed with caution due to the sample size of electric vehicles.

6. CONCLUSIONS AND FUTURE WORK

Insurance companies to ensure that their pricing accurately represents the risks they are covering. This is achieved in part by creating and using adequate pricing models. As innovative methods for modelling claim severity are introduced, and as the scientific community presents evidence that these new methods can yield comparable or superior outcomes, it is logical for these companies to explore the potential benefits of adopting these new modelling techniques, particularly supervised machine learning.

In this project, a study was conducted on the variables that most impact the cost of claims and whether these vary between non-electric and electric vehicles, also addressing the question of whether vehicles have more or less severe accidents in terms of cost.

First, considering that this work was based on real data from an insurance company, significant initial data cleaning was necessary. This included removing rows with multiple missing variables and removing outliers. Through exploratory analysis, we obtained a more detailed view of the data and the variables. Regarding our response variable, we concluded that it follows a Normal distribution in both datasets with a logarithmic transformation.

For the variables that best explain the claim amount: for non-electric vehicles, variables such as the vehicle's gross weight, the district, the vehicle's year of construction, driver’s age, years of driving experience and type of vehicle were obtained; for electric vehicles, only the vehicle's year of construction, brand, and the district were found.

Through the study using logistic regression, we concluded that electric vehicles have a higher probability of causing severe accidents.

Modelling the severity of the claim is challenging due to several factors, including the complex influences of multiple variables, such as the nature of the incidents and characteristics of insured items. Furthermore, data quality issues, like incomplete, inconsistent, or outlier-prone data, can lead to unreliable outcomes. The skewed distribution of claims, where a small number of high-cost claims dominate total costs, complicates predictions. Nonlinear relationships among variables further complicate standard modelling techniques. These challenges require sophisticated techniques to accurately model the severity of the claim.

In the future, it could be interesting to try other methods to model the severity of claims, such as GBM, XGBoost, or even neural networks. Regarding the data set itself, it could be beneficial to aggregate more variables or obtain additional variables that contribute more to the model. For this specific study, having more observations of electric vehicles would have been very important, but until 2021, electric vehicles were still relatively few.

BIBLIOGRAPHICAL REFERENCES

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. Hippokratia, 14(Suppl 1), 23.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/

Andersson, H. (2022). Study for identification and contribution of extratropical cyclone claims in the home insurance portfolio [masterThesis, Instituto Superior de Economia e Gestão]. https://repositorio.ulisboa.pt/handle/10400.5/26571

Boehmke, B., & Greenwell, B.M. (2019). Hands-On Machine Learning with R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377

Clemente, C. de M. (2023a). A refreshed vision of Non-Life Insurance Pricing—A Generalized Linear Model and Machine Learning Approach [masterThesis]. https://run.unl.pt/handle/10362/149112

Clemente, C., Guerreiro, G. R., & Bravo, J. M. (2023b). Gradient Boosting in Motor Insurance Claim Frequency Modeling. CAPSI 2023 - 23rd Conference of the Portuguese Association for Information Systems (23ª Conferência da Associação Portuguesa de Sistemas de Informação), pp. 53–69. DOI: 10.18803/capsi.v23.53-69.

Clemente, C., Guerreiro, G. R., & Bravo, J. M. (2023c). Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting. Risks, 11(9), 1–20. https://doi.org/10.3390/risks11090163

Cunha, L. & Bravo, J. M. (2022). Automobile Usage-Based-Insurance: Improving Risk Management using Telematics Data. CISTI'2022 - 17th Iberian Conference on Information Systems and Technologies. (CISTI), Madrid, Spain, 2022, pp. 1-6, IEEE Computer Society. https://doi.org/10.23919/CISTI54924.2022.9820146

de Jong, P., & Heller, G. Z. (2008). Generalized linear models for insurance data. (1 ed.) (International Series on Actuarial Science). Cambridge University Press (CUP). https://doi.org/10.1017/CBO9780511755408

Dilmegani, C. (2024). Insurance Pricing: Determination & New Methods in 2025. AIMultiple. Retrieved February 3, 2025, from https://research.aimultiple.com/insurance-pricing/

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

Félix, C. C. (2019). The use of business attributes in motor insurance pricing: Case study of a portuguese insurance company [masterThesis]. https://run.unl.pt/handle/10362/84972

Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.). (2014). Predictive Modeling Applications in Actuarial Science: Volume 1: Predictive Modeling Techniques (Vol. 1). Cambridge University Press. https://doi.org/10.1017/CBO9781139342674

Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate Frequency-Severity Regression Models in Insurance. Risks, 4(1), Artigo 1. https://doi.org/10.3390/risks4010004

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics, 70, 205–215. https://doi.org/10.1016/j.insmatheco.2016.06.006

Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized Linear Models for Insurance Rating. Casualty Actuarial Society. https://books.google.pt/books?id=HHAayAEACAAJ

Gonçalves, A. M. L. (2013, September 10). Regressão logística aplicada à pesquisa de preditores de morte. Regressão logística aplicada à pesquisa de preditores de morte. Regressão logística aplicada à pesquisa de preditores de morte. https://estudogeral.uc.pt/handle/10316/33697

Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. Risks, 9(2), Artigo 2. https://doi.org/10.3390/risks9020042

Hastie, T. J. (2017). Generalized additive models. In Statistical models in S (pp. 249-307). Routledge.

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. North American Actuarial Journal, 25(2), 255–285. https://doi.org/10.1080/10920277.2020.1745656

Hosmer, D.W. and Lemeshow, S. (2000). Application of Logistic Regression with Different Sampling Models. In Applied Logistic Regression (eds W.A. Shewhart, S.S. Wilks, D.W. Hosmer and S. Lemeshow). https://doi.org/10.1002/0471722146.ch6

Insurtech.com.br, R. (2023, dezembro 8). Perspectivas para 2024: O futuro do seguro é orientado por dados, eficiente e personalizado. INSURTECH.COM.BR. https://www.insurtech.com.br/seguros/perspectivas-para-2024-o-futuro-do-seguro-e-orientado-por-dados-eficiente-e-personalizado/

Islam, M., & Chowdhury, R. (2017). Exponential Family of Distributions (pp. 23–30). https://doi.org/10.1007/978-981-10-3794-8_3

Kagan, J. (2021). Actuarial Rate: What it Means, How it Works. Investopedia. https://www.investopedia.com/terms/a/acutarial-rate.asp

Kearney, M. (2017). Cramér's v. In The sage encyclopedia of communication research methods. SAGE Publications, Inc. Vol. 4, pp. 290-290. https://doi.org/10.4135/9781483381411

Lee, W., Park, S. C., & Ahn, J. Y. (2019). Investigating dependence between frequency and severity via simple generalized linear models. Journal of the Korean Statistical Society, 48(1), 13–28. https://doi.org/10.1016/j.jkss.2018.07.003

McDonnell, K., Sheehan, B., Murphy, F., & Guillen, M. (2024). Are electric vehicles riskier? A comparative study of driving behaviour and insurance claims for internal combustion engine, hybrid and electric vehicles. Accident Analysis & Prevention, 207, 107761. https://doi.org/10.1016/j.aap.2024.107761

Molnar, Christoph & Casalicchio, Giuseppe & Bischl, Bernd. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. 10.1007/978-3-030-65965-3_28.

Natekin, A., & Knoll, A. (2013). Gradient Boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021

Ohlsson, E., & Johansson, B. (2010). Non-Life Insurance Pricing with Generalized Linear Models. Springer. https://doi.org/10.1007/978-3-642-10791-7

Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. Journal of Mathematical Finance, 8(1), Artigo 1. https://doi.org/10.4236/jmf.2018.81012

Pereira, R. J. V. G. (2017). Tarifação a priori—Estudo de uma carteira automóvel [masterThesis]. https://run.unl.pt/handle/10362/110485

Pereira, R. J. V. G. (2017). Tarifação a priori—Estudo de uma carteira automóvel [masterThesis]. https://run.unl.pt/handle/10362/110485

Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine Learning in Forecasting Motor Insurance Claims. Risks, 11(9), Artigo 9. https://doi.org/10.3390/risks11090164

Rejda, G. E. (2005). Risk management and insurance. Person Education Inc, 13, 44-55.

Rodrigues, C. G. (2020). Avaliação de desempenho de modelos de regressão logística multivariada através de curvas ROC num estudo de RN de muito baixo peso [masterThesis]. https://repositorium.sdum.uminho.pt/handle/1822/70801

Saha, S.N. (2023). Utilizing Data Analytics in the Pricing of Non-Life Insurance. Linkedin. https://www.linkedin.com/pulse/utilizing-data-analytics-pricing-non-life-insurance-saha

Shannon, D., Murphy, F., Mullins, M., & Eggert, J. (2018). Applying crash data to injury claims—An investigation of determinant factors in severe motor vehicle accidents. Accident Analysis & Prevention, 113, 244–256. https://doi.org/10.1016/j.aap.2018.01.037

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. Insurance: Mathematics and Economics, 64, 417–428. https://doi.org/10.1016/j.insmatheco.2015.07.006

Silva, J. M. C. da. (2021). Risk modeling journey—GLM and impact analysis [masterThesis, Instituto Superior de Economia e Gestão]. https://www.repository.utl.pt/handle/10400.5/23313?locale=en

Su, X., & Bai, M. (2020). Stochastic Gradient Boosting frequency-severity model of insurance claims. PLOS ONE, 15(8), e0238000. https://doi.org/10.1371/journal.pone.0238000

The Investopedia Team. (2024). Variance Inflation Factor (VIF). Investopedia. https://www.investopedia.com/terms/v/variance-inflation-factor.asp

Valdes-Dapena, P. (2024, janeiro 18). Teslas crash more than gas-powered cars. Here's why | CNN Business. CNN. https://www.cnn.com/2024/01/18/business/why-do-people-keep-crashing-teslas/index.html

Venables, B., & Dichmont, C. (2004). GLMs, GAMs and GLMMs: An overview of theory for applications in fisheries research. Fisheries Research, 70, 315–333. https://doi.org/10.1016/j.fishres.2004.08.011

Vieira, M. I. T. (2022). Modelling large claims in non-life insurance: An application in motor insurance industry based on extreme value theory and Gradient Boosting machine techniques [masterThesis]. https://run.unl.pt/handle/10362/144999

Wu, W.-J., Li, C.-S., & Peng, S.-C. (2020). The relationships between vehicle characteristics and automobile accidents. Risk Management and Insurance Review, 23(4), 331–377. https://doi.org/10.1111/rmir.12163

Xu, Z. (2020). Best practice of risk modelling in motor insurance: Using GLM and Machine Learning approach [masterThesis, Instituto Superior de Economia e Gestão]. https://www.repository.utl.pt/handle/10400.5/20405

APPENDIX A

Variables in Study

Variable Name	Description	Levels
N_REGISTO	It is a unique key for each Record	224 205 levels
VALOR_VEIC	The vehicle's current value	24 583 levels
VALOR_VEIC_NOVO	Vehicle price if it were new	20 504 levels
VEICULO_VALOR_EXTRAS	The value of the vehicle's extras (rims, etc.)	3 580 levels
DESC_CONC_RISCO	The vehicle's circulation area	312 levels
Categoria	Vehicle category (passenger, etc.)	11 levels
ANO_CONST	Year of vehicle construction	71 levels
TYLACODE	It is also a code that will not be relevant to the study	28 462 levels
MARCA_VEIC	Vehicle brand	350 levels
MODELO	Vehicle model	12 510 levels
VERSAO	Vehicle version	34 464 levels
NUM_LUG	Number of seats	98 levels
POT	Power	479 levels
CC	Engine displacement	1 688 levels
TARA_VEIC	Vehicle's curb weight	2 692 levels
PB_VEIC	Vehicle's gross weight	1 344 levels
COMBUST	Fuel type	5 levels
COR	Vehicle color	27 levels
FORMA	Type of body (pick-ups, for example)	3 levels
SEGURADO_IDADE	Policyholder's age	87 levels
COND_ANOS_CRT	Driver's years of license	92 levels
CAP_CCC	Collision, crash, and rollover coverage capital (own damage)	25 618 levels
CAP_RC	Civil Liability coverage capital	12 levels
SIN_RC	Number of claims	4 levels
SIN_CCC	Number of claims with CCC (Collision, Crash, and Rollover)	6 levels
CT_RC	Total Civil Liability cost	98 536 levels
CT_CCC	Total CCC cost	25 571 levels

Table A.0.1 - Summary of the initial proposed feature variables

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa