Lisbon School
of Economics
& Management
Universidade de Lisboa

MASTER
FINANCE

MASTER'S FINAL WORK
PROJECT

PRICE ELASTICITY IN AUTO INSURANCE: IMPACT OF PREMIUM FLUCTUATIONS ON POLICYHOLDER BEHAVIOR

FILIPA CORREIA INÊS E CORREIA VAZ

OCTOBER 2024

MASTER
FINANCE

MASTER'S FINAL WORK
PROJECT

PRICE ELASTICITY IN AUTO INSURANCE: IMPACT OF PREMIUM FLUCTUATIONS ON POLICYHOLDER BEHAVIOR

FILIPA CORREIA INÊS E CORREIA VAZ

SUPERVISION:

DR. PAULO MARTINS SILVA

OCTOBER 2024

It matters not how strait the gate,
How charged with punishments the scroll,
I am the master of my fate,
I am the captain of my soul.
— William Ernest Henley, *Invictus* (2014)

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my supervisor, Dr. Paulo Martins Silva, whose guidance and support were invaluable throughout the writing of this project. Thank you for your availability, and important insights.

Next, I would like to extend my heartfelt thanks to João Machado, who is an incredibly important person in my life and helped me a lot during this master's. Perhaps if it wasn't for him, I wouldn't have enrolled in this master's in Finance. For that, I'm profoundly grateful.

I would like to thank my dear friends and colleagues, Mariana Silva and Rafaela Correia, who shared this journey with me and were there through the struggles during this master's. Thank you both.

Finally, I would like to thank my parents, who have always supported me, provided me with everything that I needed, and allowed me to pursue this master's degree. Thank you.

ABSTRACT AND KEYWORDS

This project aims to explore the price elasticity of demand in the auto insurance market. It will use machine learning models, such as the Logistic Regression model and Gradient Boosting model, to predict how price increases can influence policyholder behavior, not disregarding other key variables such as policyholder demographics, vehicle characteristics, and payment frequency.

The models were trained using historical data from an insurance company covering the years 2020 to 2022. The Gradient Boosting model, which performed better, was also tested using price increase simulations to evaluate its performance and how it could lead to policy cancellations and revenue loss. This test revealed a nonlinear relationship. Addressing consumer behavior when there's a premium change will help insurance companies determine better strategies to retain their policyholders while staying competitive and profitable.

The findings suggest that not only do price fluctuations strongly influence policy cancellations, but other variables such as policyholder demographics, vehicle characteristics, and payment frequency also play an essential role in assessing the reasons that lead to policy cancellations. This research is important to understand how insurance companies can adapt their premiums in order to not lose customers or profitability.

KEYWORDS: Auto Insurance; Price Elasticity; Logistic Regression Model; Gradient Boosting Model; Policyholder Demographics; Vehicle Characteristics.

RESUMO E PALAVRAS-CHAVE

O objetivo deste projeto é analizar como a elasticidade de preço da procura no âmbito dos seguros de automóveis. Modelos de machine learning são utilizados, como a Regressão Logistica e o Gradient Boosting, estes são utilizados para prever como aumentos de preços conseguem influenciar o comportamento dos segurados, considerando também outras variáveis importantes como dados demográficos dos segurados, as carateristicas dos veículos e a frequência dos pagamentos.

Os modelos foram treinados com dados históricos fornecidos por uma seguradora, e cobrem os anos entre 2020 e 2022. O model Gradient Boosting, que teve uma melhor performance, foi ainda testado com várias simulações de preços para avaliar como os aumentos de preços podem levar a cancelamentos de apólices e perda de receita, este teste revelou uma relação não linear. Perceber o comportamento dos segurados perante flutuações de preços nos prémios, irá ajudar as seguradoras a compreender que estratégias utilizar para reter os seus segurados enquanto se mantêm competitivas e lucrativas.

Os resultados sugerem que não só as alterações de preços têm um forte impacto no cancelamento de apólices, como também outras importantes variáveis (dados demográficos do segurado, características do veículo e frequência de pagamento) são cruciais para avaliar potenciais razões que levem ao cancelamento de apólices. Este estudo é importante para perceber como as seguradoras podem adaptar os seus prémios de modo a que não percam clientes e lucros.

KEYWORDS: Seguro Automóvel; Elasticidade de Preço; Regressão Losgística; Gradient Boosting; Dados demográficos dos segurados; Características dos veículos.

ACRONYMS

AUC Area Under the Curve.

GBM Gradient Boosting Machine.

LR Logistic Regression.

ML Machine Learning.

ROC Receiver Operating Characteristic.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

1 INTRODUCTION

We face risks every day, whether on the road, in our jobs, at home, when traveling, or even by existing. These risks can impact our personal lives and economic situations. Insurance was created to help people cope with those risks.

Actuarial science deals with unpredictable events through statistics and risk management (Bowers et al., 1997). One important factor for insurance companies is being able to determine a price according to the nature of the risk. This price is called a premium (Antonio and Valdez, 2012). Price elasticity in the insurance industry refers to the sensitivity of the policyholder's demand to premium price changes. Understanding how these premiums influence people's behavior is equally important since it will determine whether an individual is willing to pay the premium in exchange for the security of having the risks covered (Guelman and Guillén, 2014).

According to the Insurance and Pension Funds Supervisory Authority (2023), from 2020 to 2022, there was a growth of 12.526% in the production of non-life insurance in Portugal. Regarding automobile insurance, there was also a growth of 4.976%. This trend emphasizes the need to understand consumer behavior, providing insights to help insurance companies adapt to their clients' needs, optimize pricing strategies, and enhance competitive positioning.

This project aims to explore the price elasticity of demand using machine learning models to predict how price fluctuations can influence policyholder behavior, not disregarding other key variables such as policyholder demographics, vehicle characteristics, and payment frequency. In this project, the models were also tested using price increase simulations to evaluate their performance and how they could lead to policy cancellations and revenue loss. Addressing consumer behavior when there's a price change on the premium will help insurance companies determine better strategies to retain their policyholders while staying competitive and profitable.

An insurance company database with data from 2020 to 2022 will be used. This analysis used Google Colab and Python programming language to develop and analyze Generalized Linear models (GLM) and Gradient Boosting models (GBM) (de Jong and Heller, 2008). By applying these models, it's possible to understand the relationship between premium pricing and the variables in the study and how these factors can influence demand and retention.

Regarding the project structure, in the first section, we have the literature review to introduce the topic, followed by a descriptive analysis where the raw data is analyzed, and it's possible to understand the impact of these variables in cancellations. After that, the methodology is addressed, and this section is divided into sub section, the first part refers to how the data was treated and handled, the second part is a brief theoretical

framework about the topics and models used, and the third part, refers to how the model was developed and implemented. This section is followed by the results section, where all the results obtained from the models are analyzed. The last section is about the model performance, where the GBM was tested with small price increases to see how well it could predict policy cancellations and revenue loss.

2 LITERATURE REVIEW

Studying price elasticity in insurance markets is crucial to understand how premium prices affect and influence consumer behavior. In this context, price elasticity measures the insurance demand response of the consumers/policyholders to the price variation. According to Schlesinger (2013), there's a consensus regarding the relationship between price changes and insurance demand, where price variations don't significantly change demand. By contrast, Harrington and Niehaus (2003) argues that market structures and competition between insurance companies can influence premium prices. When consumers have more choices in the market, premium changes may increase, which can influence retention and demand.

In insurance markets, it's not only the price changes that influence the consumer's behavior. One factor that plays a major role is how individuals access the value of an insurance policy based on their risk perception. Kunreuther and Pauly (2006) refers to the fact that sometimes the decision to acquire insurance is influenced by how people perceive risks, thinking there's a low probability of certain events, which can lead to underinsurance and some level of protection gap. The authors also say that this behavior could be changed through education and open communication between the consumers and the insurers since it will lead to more informed decisions regarding insurance purchases (Kunreuther and Pauly, 2006). Grace et al. (2001) also mentions that economic conditions influence consumer demand since, during economic recessions, people tend to reduce their spending on non-essential insurance products. However, when talking about automobile insurance, that fluctuation doesn't occur significantly since, in many regions, it's mandatory, so even during economic downturns, the demand tends to be stable.

The vehicle's characteristics are a critical factor that directly impacts the changes in premium prices. Garcia and Martinez (2019, as cited in Girard (2024)) indicate that newer vehicles with modern safety technologies and high safety ratings attract lower premiums due to their reduced risk profile, while those with a high theft risk or repair costs lead to higher premiums.

Regarding demographic factors, age can be an important variable in insurance pricing due to its correlation with the risk of accidents, especially among younger drivers, who are more likely to practice risky driving behaviors (Winter, 1992). This means younger drivers may face higher premiums due to their inflated accident rates (Kelly and Nielson, 2006). In contrast, another variable that can be significant in auto insurance pricing is gender. According to the Maryland Consumer Rights Coalition (2022), many insurers tend to charge women higher premiums despite the lack of a clear correlation between gender and driving ability.

Another important variable is payment frequency - monthly, quarterly, semi-annual,

and annually. Meyer and Power (1973) highlight the relationship between the payment frequency and the overall insurance cost in their study. The authors state that while more frequent payments can alleviate the financial burden on policyholders, they can also sustain additional charges, increasing the insurance cost. It is concluded that a strategy needs to be aligned according to the profile of each possible policyholder since some may benefit from more frequent payments while others from one annual payment, depending on their financial situation (Meyer and Power, 1973). The fact that the insurers offer multiple payment options increases the likelihood of renewal since more frequent payments can help policyholders with lower incomes manage their finances.

To conclude, it's possible to gain a general perception of global studies to understand trends and correlations. However, the lack of country-specific research, particularly on how the variables in the study influence auto insurance demand in Portugal, shows a significant gap in the literature. Furthermore, no research was found that particularly studies the combined influence of vehicle characteristics, demographics, and payment periodicity on price elasticity in the auto insurance market. This gap reinforces the weight of this research, which aims to offer insights into how these variables influence insurance demand in the Portuguese market from 2020 to 2022.

3 DESCRIPTIVE ANALYSIS

This section presents a descriptive analysis of key variables in the dataset. This offers insights into policyholder demographics, payment frequencies, vehicle categories, and the reasons for policy cancellations. We used visualization graphs to better visualize this trend and distribution.

3.1 PREMIUM DISTRIBUTION

The figure 1 shows the distribution of gross written premiums over the three years (2020, 2021, and 2022). As we can observe, there's a concentration of values in the lower ranges, with a sharp decrease in frequency as premium value increases. The data shows a skewed distribution where most premiums are between 0 and 200 due to more basic coverage by policyholders, with a smaller number of policies having a higher premium value, either by premium adjustment due to higher claims frequency or broad and special perils coverage. Over the three years, there's a fairly similar distribution, which indicates that there has been no change in the structure of the premium.

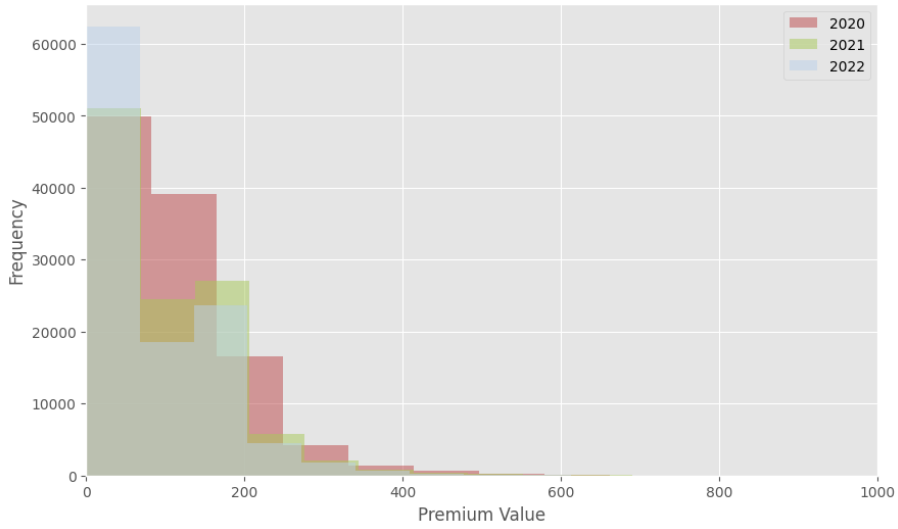

FIGURE 1: Distribution of Received Premiums by Year

3.2 AGE

This dataset includes a wide age range of policyholders, with the highest concentration between 30 and 60 years old. In figure 2, it's possible to see that the data is right-skewed, which means that there are more policyholders in the younger range, though there's a considerable portion of policyholders in the 40 to 50 age range.

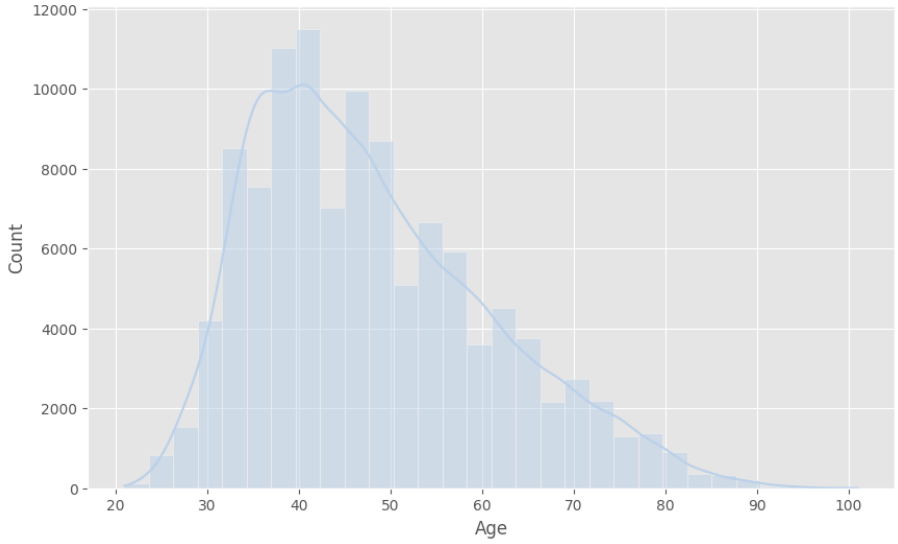

FIGURE 2: Age Distribution of Policyholders

Figure 3 further explores the distribution of policy cancellations by age group. As it can be observed the highest number of canceled policies is among the policyholders in the 31-40 and 41-50 age groups, which aligns with the previous observations since those groups also represent the largest portion of active policies. The figure also shows that ages below 30 and above 60 have fewer cancellations, this may suggest that middle age policyholders are more sensitive to price fluctuations.

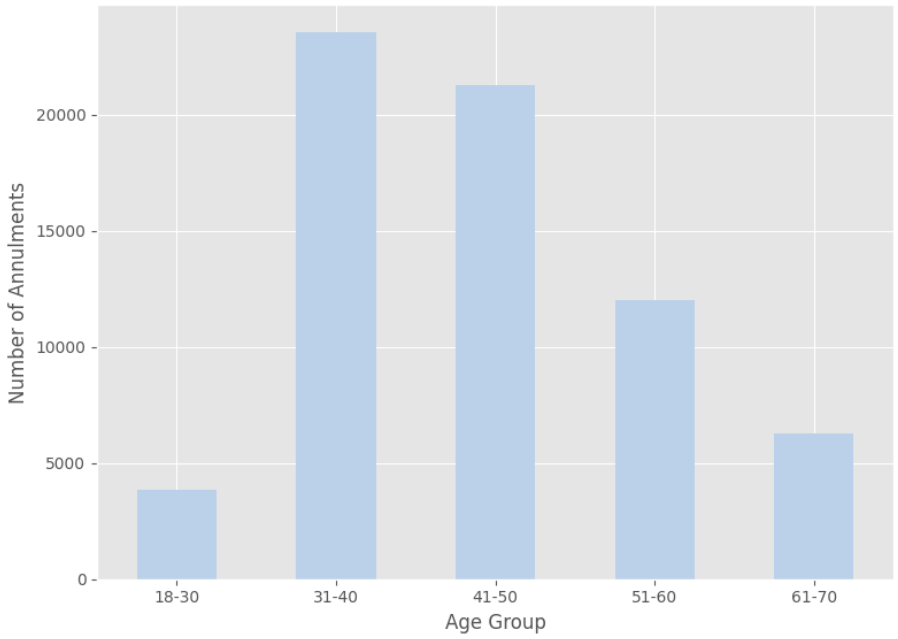

FIGURE 3: Policy Annulments by Age Group

3.3 GENDER

As observed in figure 4 the male gender is the majority class regarding policyholders, representing 70.8% of the total, while the female group represents 29.2%. Although the absolute number of cancellations is higher for males, the relative cancellation rate among males (64.9%) is only slightly higher compared to females (56.7%). This more balanced perspective accounts for the total number of policies in each group, as presented in figure 5.

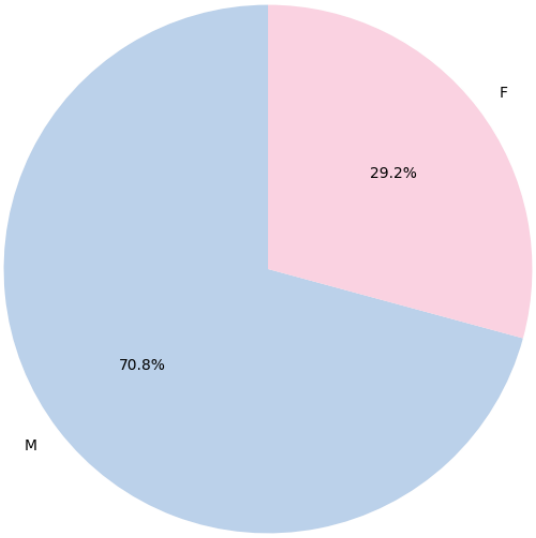

FIGURE 4: Gender Distribution

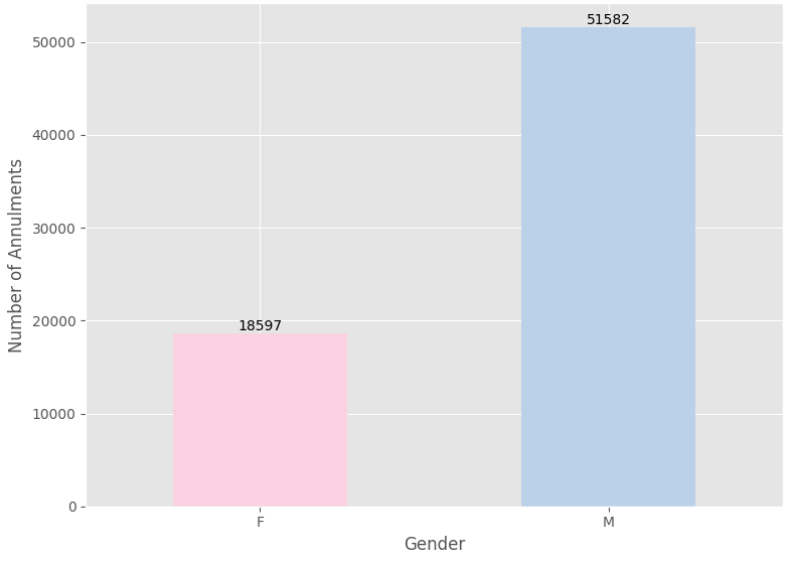

FIGURE 5: Policy Annulments by Gender

3.4 VEHICLE CATEGORY

Regarding vehicle category, most policies in this dataset are for passenger cars, followed by commercial cars and motorcycles, as can be seen in figure 6. As expected, the number of annulled policies follows the same distribution (Figure 7), with passenger cars having the highest number of annulments. This might indicate that policy cancellations are not specifically influenced by vehicle type but are more likely due to other factors such as price changes or policyholder demographics.

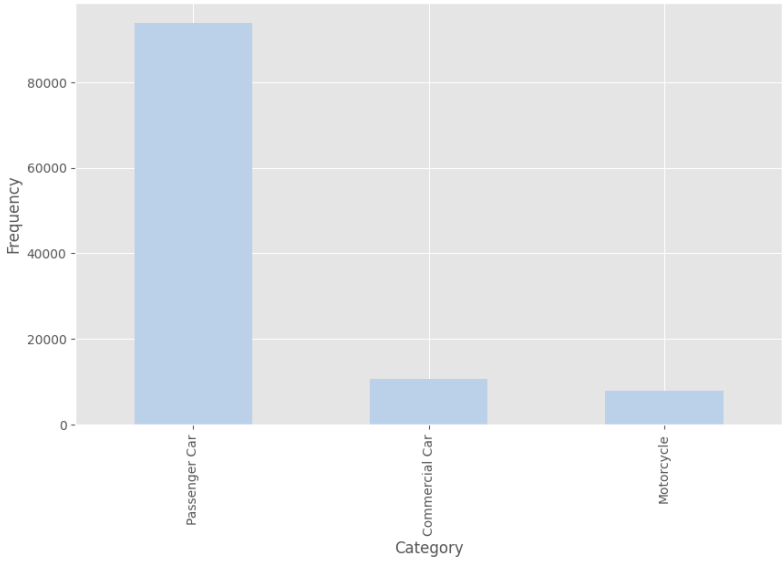

FIGURE 6: Vehicle Category Distribution

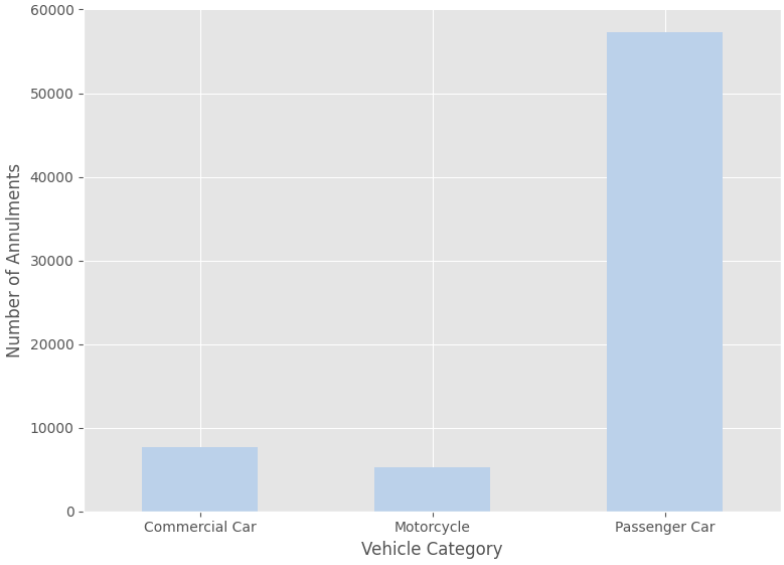

FIGURE 7: Policy Annulments by Vehicle Category

Regarding vehicle brands, figure 8 shows the top 20 brands with the most annulments. Renault, Opel, and Volkswagen are the brands with the most policy cancellations, with Renault having the highest amount of canceled policies. Understanding if these variables have an influence on policy cancellations could provide insights into customer behavior.

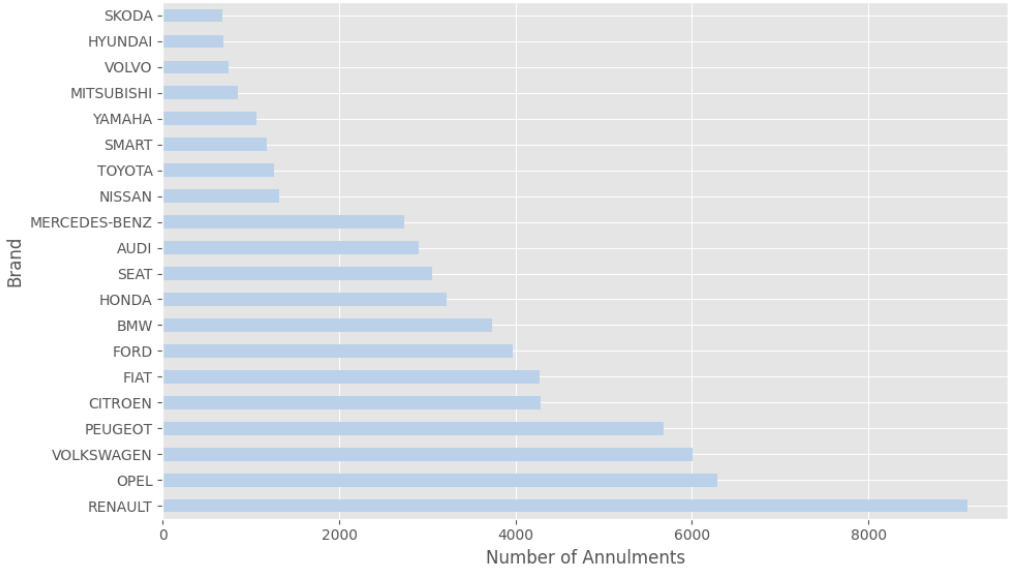

FIGURE 8: Top 20 Brands by Number of Annulments

3.5 PAYMENT FREQUENCY

Figure 9 shows that policies with an annual payment frequency have the highest number of cancellations, as well as semi-annual payments, with slightly fewer cancellations than annual payment frequency. The other two types of payment, monthly and quarterly, by contrast, show fewer canceled policies, especially quarterly payments, which show a significant decrease in policy cancellation. This may indicate that more frequent payment schedules allow policyholders to manage their finances effectively, which leads to fewer policy annulments.

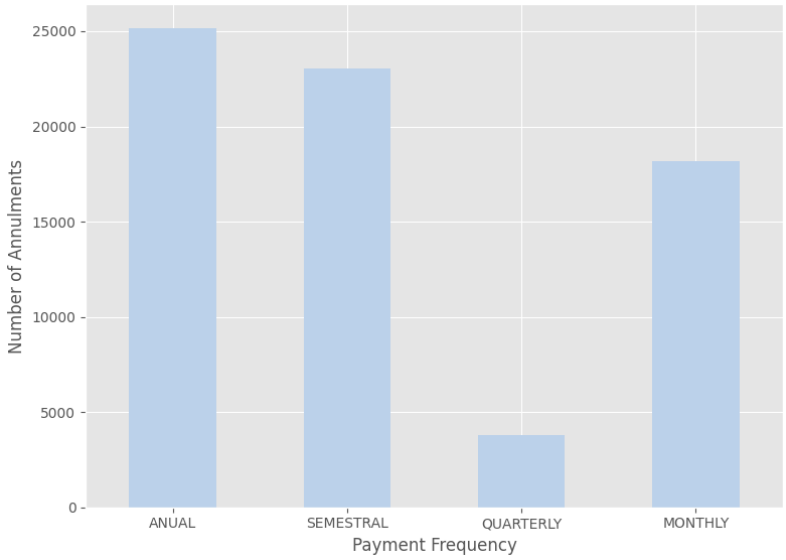

FIGURE 9: Policy Annulments by Payment Frequency

3.6 ANNULMENTS

Regarding annulment reason, this occurs for a variety of reasons, but in this dataset, the majority of cancellations are due to missing payments, as shown in figure 10. This emphasizes the policyholder's sensitivity to price changes since it may indicate that policyholders stop paying their current insurance in favor of a lower priced alternative when a better offer becomes available. Other reasons, such as policies that are transferred, false declarations, and invalid emissions, represent a smaller portion of the cancellations.

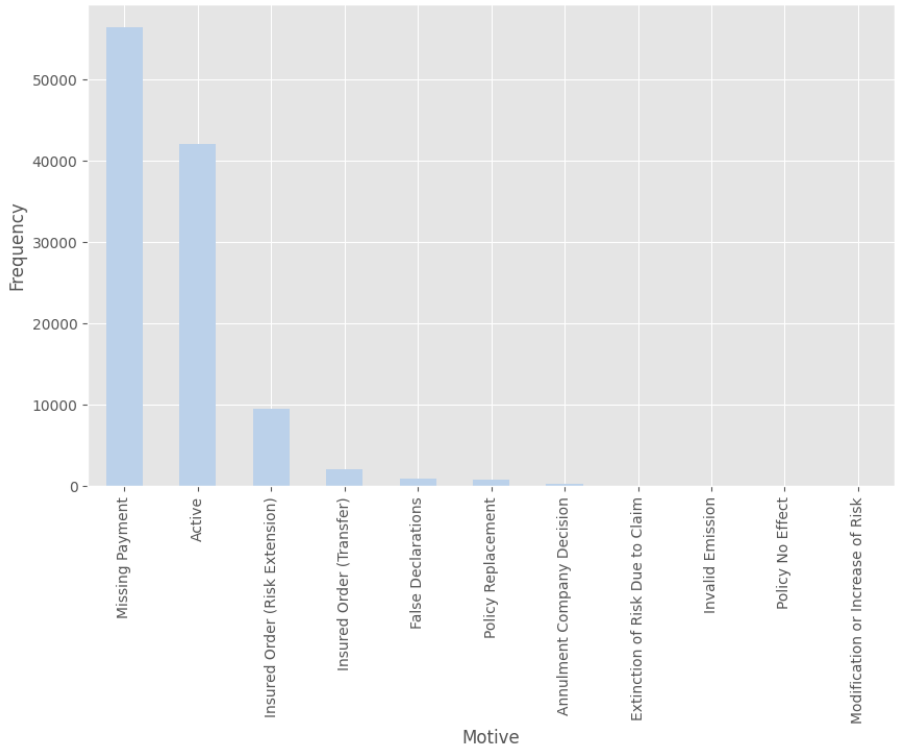

FIGURE 10: Annulment Reasons

To conclude, when referring to active and annulled policies, there are significantly more canceled policies (70179) than active ones (42111). This imbalance supports the importance of studying the drivers of policy cancellations, which are likely tied to price sensitivity.

4 METHODOLOGY

4.1 DATA

This project uses a quantitative approach to study the price elasticity of demand in the automobile insurance market from 2020 to 2022 for a Portuguese property and casualty carrier. The objective is to understand how price variations influence policy cancellation. Data-driven models were used, particularly generalized linear models, logistic regression, and Gradient Boosting models. The data was sourced from an anonymous insurance company. The original dataset has 112.290 rows and 24 variables/columns. It includes key variables such as vehicle characteristics (brand, model, vehicle category, and year), policyholder demographics (gender and age), the sum of the premiums from each year, payment frequency (monthly, quarterly, semi-annual, and annual), dates regarding the beginning and the ending of the policy, and the annulment reason. The target variable that we want to study is policy cancellation in order to define a predictive model based on price elasticity.

Two additional variables were added to the original dataset, Price Variation from 2020 to 2021 and Price Variation from 2021 to 2022. These variables will enhance the understanding of price variations' influence on consumer behavior regarding policy cancellation. They were created using the following formula:

$$\text{Price Variation}_{Year1-Year2} = \frac{\text{Premium}_{Year2} - \text{Premium}_{Year1}}{\text{Premium}_{Year1}} \tag{1}$$

In order to proceed with the study, the dataset had to be cleaned. It went through a series of preprocessing steps that are crucial to ensure accuracy and relevance. Before the treatment of missing values and outliers, some variables had errors, e.g., car brands appearing as different brands when they were the same because of spaces and misspelling errors, and they needed to be corrected. After that, it was possible to analyze where the missing values were, using Google Colab. Six variables had missing values, as can be seen in table I:

Variable	Missing Values
Annuity_ini_2020	40609
Annuity_ini_2021	18156
Annuity_ini_2022	2132
End_date_annuity_2020	40609
End_date_annuity_2021	18156
End_date_annuity_2022	2132
Date_Annulment	42111
Policyholder_Birth_date	9

TABLE I: Missing Values

Three approaches were used to deal with the missing values because, except for the variable Policyholder_Birth_date, these missing values are missing on purpose. Annuity_ini_2021 and Annuity_ini_2022 have missing data because it means that the policyholder didn't renew the policy in that year. To fix these missing values, they were filled with dates from the past (January 1st, 1970) to create an outlier so it was possible to visualize them more easily when analyzing the data. For the variable End_date_annuity_2021, End_date_annuity_2022, and Date_Annulment, a similar approach was done, but since the missing data meant that the policies were not annulled, dates in the future (December 31st, 2100) were used to fill in the missing values. To fill in the missing values in the variable Policyholder_Birth_date, which is the birth date of the policyholder, the mean value of the variable Policyholder_Birth_date was utilized for imputation.

Once the missing values were addressed, it was also necessary to handle the categorical variables to be able to proceed with the study. For the variables Frequency, Gender, and Vehicle_Category, one-hot encoding was applied. This method converts the categorical variables into new binary columns, in each of these columns a value of 0 indicates the absence of that category, while a value of 1 represents the presence. Regarding the variable Annulment, label encoding was applied, transforming the policies that had the label "Active" into a 0 and everything else into a 1.

Label encoding was applied to the variables Brand and Model. This method is similar to the one-hot encoding, however, it assigns a unique integer to each category within the variable, transforming it into a numerical value (Mukhiya and Ahmed, 2020). For this specific situation, this method was used in two steps. Firstly, the variable Brand was converted to a numerical variable using LabelEncoder, which created a new variable, Brand_Encoded. After this, a custom approach was applied to the variable Model based on the corresponding brand. For each unique Brand, a separate LabelEncoder was trained to convert the Model associated with that brand into numerical values, which created a

new variable, Model_Encoded. This method will ensure that each model has a unique value tied to its respective brand, preserving the hierarchical structure between brands and models.

Regarding the date variables which are DateTime variables, a different approach was applied. It was necessary to convert them as well since they are not considered numerical variables, and regression models wouldn't be able to interpret the temporal information. Therefore the number of days since a reference date (January 1st, 1970) was calculated for the following variables Beginning_Policy, Annuity_ini_2020, Annuity_ini_2021, Annuity_ini_2022, End_date_annuity_2020, End_date_annuity_2021, End_date_annuity_2022, Date_Annulment, and Policyholder_Birth_date. This reference date is known as the Unix epoch and is commonly used in computing systems since it represents a standard point in time for timestamp calculations (Bach, 1990). This approach allows the study to be consistent in time measurement. Furthermore, important variables such as the policyholder's age at the start of the policy and the duration of each annuity were calculated. Even though they will not be directly used in the models, this method ensures that the temporal aspects of the data are captured and that the regression model can process them effectively.

Subsequently, multicollinearity was tested so the coefficients could be reliable. Independent variables that are highly correlated lead to unstable coefficients (Murray et al., 2012). To detect multicollinearity, we used the Variance Inflation Factor (VIF), this technique measures how much the variance of the regression coefficient increases if there's a correlation between the independent variables (Shrestha, 2020). VIF is expressed as

$$VIF = \frac{1}{1 - R^2} = \frac{1}{\text{Tolerance}} \tag{2}$$

Where the tolerance is the reciprocal of the VIF. A low tolerance means that there's multicollinearity. When VIF=1, it shows that there's no correlation between the variables, if 1> VIF <5, it means that the variables, to some degree, are correlated, however, the challenge arises when VIF values are between 5 and 10 since shows that we are facing multicollinearity among the predictors in the regression model (Shrestha, 2020). When analyzing multicollinearity, some variables exhibited strong correlations, e.g., Gender_F and Gender_M showed high VIF values, 3320 to be exact, and since there are fewer policies with females as policyholders, it was decided to drop Gender_F to avoid redundant information and to improve the model stability. Since Vehicle category and Payment frequency had infinite value for VIF, one from each category had to be dropped. In the Vehicle category, the variable referring to commercial cars was dropped because this type of car policy tends to react strongly to price fluctuations due to its commercial purposes. Regarding the Payment frequency category, the variable for monthly payment frequency

was removed, despite the quarterly payment frequency having the fewest policies. This is because dropping the quarterly payment frequency would cause the VIF value for monthly frequency to be above 5, indicating some level of multicollinearity, which would require removing two variables instead of one.

After the data cleaning, the dataset has the following variables and descriptions, as can be seen in table II:

Variable	Description	Category
Variables related with Policy Information		
Last_Premium_Policy	Last premium paid for the policy	Numerical
Branch	Branch associated with the policy	Categorical
Policy DUMMY	Policy number transformed	Numerical
UNIRISCO	Risk factor related to the policy	Categorical
Beginning_Policy	Policy start date	Date Time
Maturity	Day and Month Policy Anniversary	Numerical
Date_Annulment	Date of policy annulment	Date Time
Annulment_0	No annulment (encoded as 0)	Numerical
Annulment_1	Annulment occurred (encoded as 1)	Numerical
Frequency	Payment Frequency	Categorical
Annuity_ini_2020	Start date of annuity in 2020	Date Time
Annuity_ini_2021	Start date of annuity in 2021	Date Time
Annuity_ini_2022	Start date of annuity in 2022	Date Time
End_date_annuity_2020	End date of annuity in 2020	Date Time
End_date_annuity_2021	End date of annuity in 2021	Date Time
End_date_annuity_2022	End date of annuity in 2022	Date Time
Annulment	Reason for the Annulment of the Policy	String
BONUS_ATUAL	Adjusts premiums based on claims history	Numerical
days_from_beginning_policy	Days since the policy began	Numerical
days_from_annuity_ini_2020	Days since the start of annuity 2020	Numerical
days_from_annuity_ini_2021	Days since the start of annuity 2021	Numerical
days_from_annuity_ini_2022	Days since the start of annuity 2022	Numerical
days_from_end_annuity_2020	Days since the end of annuity 2020	Numerical
days_from_end_annuity_2021	Days since the end of annuity 2021	Numerical
days_from_end_annuity_2022	Days since the end of annuity 2022	Numerical
duration_annuity_2020	Duration of annuity 2020	Numerical
duration_annuity_2021	Duration of annuity 2021	Numerical
duration_annuity_2022	Duration of annuity 2022	Numerical
days_until_annulment	Number of days until policy annulment	Numerical
Frequency_ANUAL	Frequency of payment - Annual	Numerical
Frequency_SEMESTRAL	Frequency of payment - Semi-annual	Numerical
Frequency_QUATERLY	Frequency of payment - Quarterly	Numerical
Variables related with Customer Information		
Policyholder_Birth_date	Policyholder's birth date	Date Time
Gender	Policyholder gender	Categorical
Age	Policyholder's current age	Numerical
policyholder_age_at_policy_start	Policyholder's age at policy start	Numerical
days_from_policyholder_birth	Days since policyholder's birth	Numerical
Gender_E	Gender indefinite	Numerical
Gender_M	Gender Male	Numerical

Variables related with Vehicle Information		
Brand	Vehicle brand	Categorical
Model	Vehicle model	Categorical
Brand_Encoded	Encoded vehicle brand	Numerical
Model_Encoded	Encoded vehicle model	Numerical
Vehicle_Category_Passenger Car	Vehicle category - Passenger Car	Numerical
Vehicle_Category_Motorcycle	Vehicle category - Motorcycle	Numerical
Variables related with Premium Information		
SumOfPremio_received_2020	Premium received in 2020	Numerical
SumOfPremio_received_2021	Premium received in 2021	Numerical
SumOfPremio_received_2022	Premium received in 2022	Numerical
Price Variation_2020_2021	The price variation from 2020 to 2021	Numerical
Price Variation_2021_2022	The price variation from 2021 to 2022	Numerical

TABLE II: Variables in study

4.2 THEORETICAL FRAMEWORK

4.2.1 Price Elasticity of Demand

In economics, price elasticity of demand is a fundamental concept that measures how the quantity demanded of a good or service changes in response to changes in its price (Varian, 2014). Its mathematical equation is:

$$E_p = \frac{\%\Delta Q}{\%\Delta P} \tag{3}$$

Where $\%\Delta Q$ is the percentage change in quantity demanded and $\%\Delta P$ is the percentage change in price. If the elasticity is less than zero, the good or service has negative elasticity, which means that a price increase will lead to a decrease in demand, a common pattern for most goods or services (Nicholson and Snyder, 2019). Price elasticity helps us understand consumer behavior to changes in price, which is crucial for insurance companies when setting prices (Zweifel and Eisen, 2012). This insight is important for understanding the connection between policyholder behavior and pricing strategies in the insurance industry.

4.2.2 Regression Analysis

Another model frequently used to understand how different factors influence a specific outcome is regression analysis. This method estimates the relationship between a dependent variable and one or more independent variables (Darlington and Hayes, 2017).

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4}$$

where Y is the dependent variable and what we are trying to predict, X is the independent variable and what we believe affects Y, β_0 is the intercept, indicating the value of Y when X = 0, β_1 is the coefficient that measures the impact of X on Y, and ϵ is the error term, capturing the variation in Y not explained by X (Wooldridge, 2016). When we transform Y and X in logarithms, we obtain the log-log model. In this model, a one percent change in the independent variable results in a β percent change in the dependent variable.

Categorical variables represent distinct categories or groups, e.g., gender and payment frequency. They cannot be used directly in regression models since they are nonnumerical variables. Instead, they are transformed into dummy variables, which take values of 0 or 1 to indicate the absence or presence of a particular category (Gujarati and Porter, 2009). For instance, if a variable represents three payment options (monthly, quarterly, annually), two dummy variables need to be created:

- D_1 = 1, if the payment is monthly, 0 otherwise.
- D_2 = 1, if the payment is quarterly, 0 otherwise.

These dummy variables would now be included in the regression model to account for the effect of different payment frequencies on the dependent variable.

4.2.3 Generalized Linear Models

The Generalized Linear Models (GLM) extend the classical linear model and are used to analyze the relationships between variables. The study of these models and their applicability have been growing throughout time and have been widely applied in actuarial work (de Jong and Heller, 2008). The GLM has to respect three assumptions (Agresti, 2015):

- The distribution of the variable of response Y belongs to the exponential family;
- The linear predictor, expressed as $\eta = \sum_{i=1}^{p} x_i \beta_i$, represents a linear combination of the p covariates, where x_i are the covariates and β_i are the associated parameters;
- The link function $g(\mu) = \eta$, with $g(.)$ being a monotonic and differentiable function in all its domains and $\mu = E[Y]$. While the link function connects the linear predictor to the mean of the response variable, the covariates are integrated into the model through the linear predictor. The link function should be chosen in order for the adjusted values to be fitted to the domain of μ (Nelder and Wedderburn, 1972).

Exponential Family

It is said that a random variable Y has distributions that belong to the exponential family if its Probability Density Function can be represented as:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{5}$$

where y is the observed response variable, θ is the natural parameter, ϕ is the dispersion parameter, and $b(\theta)$, $a(\phi)$ and $c(y,\phi)$ are real valued functions, where the distribution of $b(.)$ does not depend on the parameters and $a(\phi) = \frac{\phi}{\omega}$, where ω is a known constant (Turkman and Silva, 2000).

The exponential distribution family is especially useful since it includes distributions that can model a wide variety of data types while conserving a consistent framework (McCullagh and Nelder, 1989). In this family, the expected value of the response variable Y is

$$E[Y] = b'(\theta), \tag{6}$$

and its variance is given by

$$Var(Y) = b''(\theta)a(\phi). \tag{7}$$

Since this model is very flexible it is useful in GLMs, where the natural parameter θ is linked to the mean of the distribution through a canonical link function, this allows a simpler estimation and inference within these distributions. The exponential family includes several distributions, such as Normal distribution, Poisson, Gamma, and Binomial, which can be used in GLMs (de Jong and Heller, 2008).

Parameter Estimation

The maximum likelihood estimation approach is used to estimate the parameters. For a sample $(y_1, y_2, ..., y_n)$ retrieved from a distribution belonging to the exponential family, the log likelihood function is given by:

$$\ell(\theta,\phi;y_1,\ldots,y_n) = \sum_{i=1}^{n} \ln f(y_i;\theta_i,\phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \tag{8}$$

According to Dobson (2002), the maximum likelihood estimates for the coefficients in the linear predictor can be derived from a system of equations where the gradient of the log likelihood with respect to each parameter is set to zero. This is expressed as:

$$\frac{\partial\ell(\beta)}{\partial\beta_j} = 0 \iff \sum_{i=1}^{n}\frac{\partial\ell(\beta_j,y_i)}{\partial\beta_j} = 0 \iff \sum_{i=1}^{n}\left[\frac{y_i - b'(\theta_i)}{a(\phi)}\right]\frac{\partial\theta_i}{\partial\beta_j} = 0, \quad j = 1,\ldots,p, \tag{9}$$

where β is the vector of parameters.

Two methods can be applied, Newton-Raphson or Fisher scoring, since there's typically no closed-form solution for these equations. When using canonical link functions, these methods are equivalent according to McCullagh and Nelder (1989).

4.2.4 Logistic Regression

This model is used when the dependent variable is binary, which means that it only has two outcomes, e.g., success/failure (Hosmer et al., 2013). In this case, the response variable Y, which takes values of 0 or 1, follows a Bernoulli distribution with a success probability defined as $\mu = P[Y = 1]$. So its mass function probability is given by:

$$f_Y(y; \theta, \phi) = \mu^y (1-\mu)^{1-y}, \text{y= 0 ou 1} \tag{10}$$

It is easily shown that the Bernoulli distribution belongs to the exponential family by writing the equation 5 as below (Agresti, 2015):

$$f_Y(y; \theta, \phi) = \exp\left(y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right) \tag{11}$$

The natural parameter θ is the logit function of μ:

$$logit(\theta) = \log\left(\frac{\mu}{1-\mu}\right), \tag{12}$$

which results in the following equation for μ:

$$\mu = \frac{e^\theta}{1+e^\theta} \tag{13}$$

Moreover, it is known that $\phi = 1$, and the associated functions in the exponential family are: $a(\phi), b(\theta) = -log(1-\mu), and c(y, \phi) = 0$. Therefore, the expected value is $E[Y] = \frac{e^\theta}{1+e^\theta} = \mu$ and $Var[Y] = \frac{e^\theta}{(1+e^\theta)^2} = \mu(1-\mu)$. A binary response can then be modeled using GLM, taking into consideration the Bernoulli distribution with success probability μ (Hosmer et al., 2013). To relate the linear predictor to μ, the logit link function is used, as can be seen below:

$$g(\mu) = \theta = \log\left(\frac{\mu}{1-\mu}\right)$$

Logistic regression is very useful when analyzing data with binary outcomes since it allows to estimate how changes in the independent variable(s) influence the likelihood of a certain event (Hosmer et al., 2013).

4.2.5 Gradient Boosting Model

Gradient Boosting Model is an interactive algorithm that improves predictive performance by focusing on the errors of previous models, it aims to adjust the new predictor based on the residual errors left by the previous predictor (Géron, 2019). This model combines weak learners, typically decision trees, into a single strong predictive. By using gradient descent, each model in the sequence is trained to reduce a specified loss function. Therefore, the Gradient Boosting Model can handle effectively both regression and classification since it can adapt to several data types and distributions (Friedman, 2001). This model is very flexible since it allows to choose the loss function, which makes this model adaptable across several types of problems, including non-linear relationships. One of the advantages of GBM over other traditional models, like linear regression, is its capability to manage complex and non-linear interactions between variables, since it doesn't assume a predefined form for the data, it can capture complex patterns that can be missed otherwise. Moreover, this model allows regularization techniques to prevent overfitting, like subsampling. This technique introduces randomness into the training process, lowering model variance (Friedman, 2002). Regarding interpretability, in this model, we can use feature importance measures and partial dependence plots, which allows the users to have a much clearer view of the influence that individual predictors can have on the model's output. This can be useful for some industries like insurance and finance, where it's necessary to understand the factors driving predictors. Gradient Boosting Model stands out from other machine learning methods like Neural Networks since it can generate interpretable models while being highly accurate. These advantages make GBM a widely recognized and practical model for solving complex predictive tasks (Guelman and Guillén, 2014).

4.2.6 Confusion Matrix

One tool to measure the performance of classification models, especially regarding binary or multiclass problems, is the confusion matrix. It shows us the model's predictions by comparing the actual and the predicted outcome. This matrix has four important components for binary classification: True Positives, True Negatives, False Positives, and False Negatives. The respective values of these components represent the accuracy of the model in distinguishing between the positive and negative values (Sokolova and Lapalme, 2009).

Using the confusion matrix allows us to compute key performance metrics such as Precision, which indicates how many of the predicted positives are correct, Recall, which measures how many actual positives the model correctly identifies, and the F1-Score,

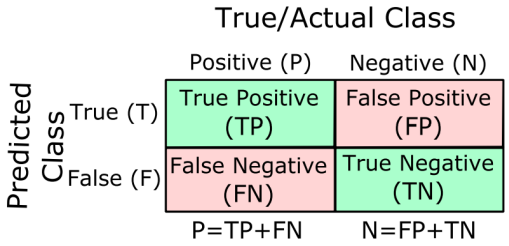

FIGURE 11: Example of a Confusion Matrix.

Source: Tharwat (2020), p. 170.

which is a combination of precision and recall metrics, and balances the compromise between the two, this metric can be useful when we are facing imbalanced datasets (Powers, 2011). This tool is essential to analyze the performance of the classification model, especially in scenarios like fraud detection in insurance, where false positives and false negatives differ significantly (Tharwat, 2020).

4.2.7 ROC Curve

The Receiver Operating Characteristic curve is a well known tool to evaluate the performance of binary classification models in various thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), and it shows the trade-off between detecting positives correctly (sensitivity) and falsely detecting negatives as positives (1-sensitivity) (Gonçalves et al., 2014). When the curve on the graph moves toward the upper left corner, it means that the model performed well since it indicates that there's a high true positive rate with a low false positive rate (Fawcett, 2006).

The figure also presents a line that represents random guessing. The desired outcome when using this tool is a curve further away from the line since it indicates a better ability to discriminate between classes. The area under the curve (AUC) summarizes the model performance, the higher the value the better. The value ranges between 0 and 1, with 1 representing a perfect classification, a value of 0.5 means that the performance of the model was not better than the random guessing. When we have a high AUC, it means that the model will have a good performance with different classification thresholds, which makes the ROC curve, when dealing with imbalanced datasets, very useful (Bradley, 1997).

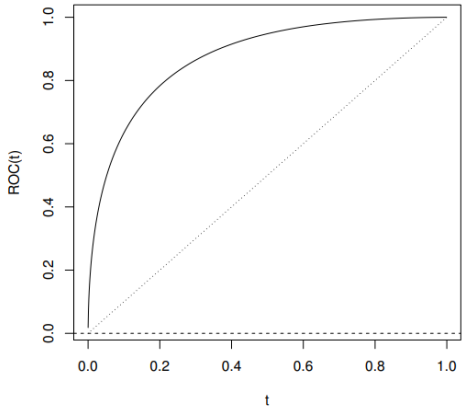

FIGURE 12: Example of a ROC curve for a bi-normal model.

Source: Gonçalves et al. (2014), p. 6.

4.3 MODEL DEVELOPMENT

The Logistic Regression model was used to predict the likelihood of policy cancellation for two periods: 2020 to 2021 and 2021 to 2022. This model was chosen due to its capability to address binary variables. Since our dataset is imbalanced, we have more canceled policies than active policies, so we had to resample it. This resampling involved reducing the number of canceled policies to match the number of active policies. With this technique, we ensure that the majority class does not influence the logistic regression model. For all the models, the data was split between training and test sets, 80% for training and 20% for testing. The `StandardScaler` (Géron, 2019) was also used to standardize features on a comparable scale. Once the dataset was balanced, we did two first logistic regression models, with price variation for each period as the only explanatory variable. This was a way to perceive the influence of price changes on policy cancellations, since it was the only variable used, it was possible to identify the direct effect of this variable on policyholder behavior. After this assessment, more explanatory variables were added to assess the model's predictive power. The variables that were chosen were price variation, age of the policyholder, gender_M if male or not (we dropped Gender_F), vehicle category for passenger cars and motorcycles, payment frequency annual, semi-annual, and quarterly, and vehicle brand and model. These variables were chosen due to their relevance to insurance companies when evaluating retention. The fact that we add more explanatory variables allows a more comprehensive model since it takes into account not only price changes but also demographics, vehicle related factors, and payment behavior, which can affect policy cancellations. To implement the logistic model for the first simpler models and the expanded ones (with explanatory variables for both periods), we used the `LogisticRegression` (Géron, 2019) class from the

scikit-learn library. Taking into consideration the imbalance of the dataset and to tune the model, besides the resample, we used the `class_weight='balanced'` (Géron, 2019) for the same reason as before.

The Logistic Regression equation for the model that evaluates the influence of the Price variation for both periods is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Price Variation} \tag{14}$$

Where p is the probability of annulment, β_0 is the intercept, and β_1 is the model coefficient that will be approached in the results section. Regarding the Logistic Regression equation for the expanded model, meaning with the explanatory variables, it is expressed as:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Price Variation} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Gender\_M} \\ + \beta_4 \cdot \text{Vehicle\_Category\_Passenger Car} \\ + \beta_5 \cdot \text{Vehicle\_Category\_Motorcycle} + \beta_6 \cdot \text{Frequency\_ANUAL} \\ + \beta_7 \cdot \text{Frequency\_SEMESTRAL} + \beta_8 \cdot \text{Frequency\_QUARTERLY} \\ + \beta_9 \cdot \text{Brand\_Encoded} + \beta_{10} \cdot \text{Model\_Encoded}\end{aligned} \tag{15}$$

Where, as previously said, the p is the probability of annulment, β_0 is the intercept, and $\beta_1, ..., \beta_{10}$ is the model coefficient that will be approached in the results section.

These models were trained on the training dataset, and the test set was used to make predictions. To analyze the likelihood of policy cancellation, we used the model coefficients of each variable to determine its influence. The coefficients for each feature were converted into odds ratios to clarify the impact of each variable on the probability of policy annulment. The odds ratio for the Price variation variable served as an indicator of price elasticity.

After the Logistic Regression model, we also used the Gradient Boosting model to further analyze the data for both periods since it can capture potential nonlinear relationships between the predictors and the target variable (Friedman, 2001). This model was not used with the simple regression model (only with the Price variation variable) but only with the Logistic regression model with the explanatory variables for both periods since the first model was only to identify if price variations would influence policyholder behavior. The Gradient Boosting model is a collective machine learning method that constructs models sequentially, where the new model tries to correct the errors of the previous one in order to improve its performance regarding the prior model. For this model, the same resampled data was used to address the class imbalance and the same explanatory variables were

used, as well as the `StandardScaler`, to have consistency and to see which model would perform better for both 2020 to 2021 and 2021 to 2022 periods. To implement the Gradient Boosting models, we used `GradientBoostingClassifier` (Géron, 2019). This model was also trained using the training dataset (80%), and the test set (20%) was used to make predictions. To evaluate the models' performance, we used classification metrics, such as accuracy, confusion matrix, precision, recall, and F1-score. To understand which variables have the most impact on the model's decision-making process, we extracted feature importance.

Following this process, it was necessary to evaluate the models' performance to access the discriminatory power. Thus, we used the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC).

5 RESULTS

5.1 LOGISTIC REGRESSION - PRICE VARIATION

From the period 2020 to 2021:

For this model, we only included the price variation_2020_2021 as the independent variable, and the model achieved an accuracy of 72.33%, which means that the model correctly predicted whether a policy was canceled (annulled) or active about 72% of the time. The confusion matrix, as well as the model results, are as follows:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	8310	129
Actual Canceled (1)	4532	3874

TABLE III: Confusion Matrix for Logistic Regression (2020-2021)

Intercept:	-0.59907178
Coefficient for price change:	-5.03161982
Odds Ratio for price variation:	0.00653

TABLE IV: Logistic Regression Results 2020-2021

As we can observe in the table, the model predicted 8310 out of 8439 active policies but misclassified 4532 canceled policies as active. This indicates that while the model strongly identifies active policies, it tends to underpredict cancellations. Regarding the classification report, it reveals a precision of 0.97 for predicting cancellations, which means that 97% of the predicted canceled policies were correct. The model has a recall of 0.46 and an F1-Score of 0.62. The overall weighted average F1-Score is 0.70, which indicates that although the model has a high precision for canceled policies, it has a low recall, meaning it's missing a large portion of actual canceled policies. When talking about coefficients, a negative coefficient of -5.03 suggests that as price increases, the likelihood of a policy remaining active significantly decreases. The odds ratio for price variation was also calculated, returning a ratio of 0.006, and this indicates that for each unit increase in price change, the probability of staying active is reduced by 99.4%. This indicates a strong connection between price increases and policy cancellations.

From the period 2021 to 2022:

For the second logistic regression model, we only included the price variation_2021_2022 as the independent variable, and the model achieved an accuracy of 71.63%, which means that the model correctly predicted whether a policy was canceled (annulled) or active about 71.63% of the time. The confusion matrix and the logistic regression results for this model are presented in tables V and VI:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	8268	171
Actual Canceled (1)	4608	3798

TABLE V: Confusion Matrix for Logistic Regression (2021-2022)

Intercept:	-0.60474774
Coefficient for price change:	-5.07911134
Odds Ratio for price variation:	0.00622

TABLE VI: Logistic Regression Results 2021-2022

In this model, 8268 active policies were correctly identified, however, the model wrongly predicted 4608 canceled policies as active. The classification report reveals a precision of 0.96 for predicting cancellations, which means that 96% of the predicted canceled policies were correct. The model has a recall of 0.45 and an F1-Score of 0.61. The overall weighted average F1-Score for this second period is 0.69, which reflects a similar performance as the previous period, with high precision but low recall for canceled policies. When we compare the other metrics to the previous period, we observe similar results. The coefficient for price change is around -5.07, meaning that for each unit that increases in price change, the probability of it remaining active decreases significantly. The odds ratio for price variation was also calculated, giving us a ratio of 0.006. The consistency across both periods indicates that price changes have a stable and significant influence on cancellations, which means that the policyholder behavior is sensitive to price changes in both periods.

5.2 LOGISTIC REGRESSION - EXPLANATORY VARIABLES

From the period 2020 to 2021:

In this model, explanatory variables were introduced, such as age, gender, vehicle characteristics, and payment frequency, as explained in section 4.3. The Logistic Regression model for this period had a higher accuracy of 78.49% than the model with only one independent variable. The corresponding confusion matrix and logistic regression results are displayed in tables VII and VIII:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	7597	842
Actual Canceled (1)	2782	5624

TABLE VII: Confusion Matrix for Logistic Regression with Explanatory Variables (2020-2021)

Intercept:	0.52120
Price Variation_2020_2021:	0.12171
Age:	0.72165
Gender_M:	1.19858
Vehicle_Category_Passenger Car:	0.83968
Vehicle_Category_Motorcycle:	0.98631
Frequency_SEMESTRAL:	0.52116
Frequency_QUARTERLY:	0.81099
Frequency_ANUAL:	0.33746
Brand_Encoded:	0.98602
Model_Encoded:	1.01763

TABLE VIII: Logistic Regression Coefficients 2020-2021

When new explanatory variables are added, it's noticeable that there's an improvement in the model performance, particularly in reducing false positive predictions of policy cancellations. The model correctly predicted 5624 out of 8604 canceled policies and 7597 out of 8439 active policies. Regarding the classification report, this shows a precision of 0.87 for predicting canceled policies, a recall of 0.67, and an F1-Score of 0.76. With these values, we can observe that this model has a more balanced performance between precision and recall when compared to the simpler model. In terms of coefficients, the price coefficient for the period 2020-2021 is 0.1217, which indicates a moderate positive relationship between price variation and the likelihood of cancellation. A significant predictor is gender Male with a coefficient of 1.1986, with male policyholders more likely to cancel policies. Another strong predictor is the vehicle model which has a positive coefficient of 1.0176, this suggests that policies for certain vehicle models are more likely to be canceled compared to other models. Specifically, for these models, the odds of policy cancellation increase by approximately 176.7%, which shows a strong influence.

From the period 2021 to 2022:

The last Logistic Regression model had a higher accuracy of 78.12% than the model with only one independent variable for the period 2021 to 2022. The corresponding confusion matrix and logistic regression results are the following:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	7527	912
Actual Canceled (1)	2774	5632

TABLE IX: Confusion Matrix for Logistic Regression with Explanatory Variables (2021-2022)

Intercept:	0.56520
Price Variation_2021_2022:	0.11176
Age:	0.73372
Gender_M:	1.18812
Vehicle_Category_Passenger Car:	0.82204
Vehicle_Category_Motorcycle:	0.96774
Frequency_SEMESTRAL:	0.50537
Frequency_QUARTERLY:	0.82403
Frequency_ANUAL:	0.30599
Brand_Encoded:	1.00437
Model_Encoded:	0.98015

TABLE X: Logistic Regression Coefficients 2021-2022

As we can observe in the table IX, the model correctly predicted 5631 canceled policies and 7527 active policies with a precision of 0.86 for predicting cancellations, which means that 86% of the predicted canceled policies were correct. The model has a recall of 0.67 and an F1-Score of 0.75. Similar to the previous period, the coefficient for the price variation is 0.1117, meaning that price fluctuations continue to influence the likelihood of policy cancellations significantly. For this period, the vehicle brand (1.0043) has a higher impact on policy cancellation than the vehicle model (0.9801). The variable Gender Male continues to have a strong influence, with a coefficient of 1.1881, meaning that male policyholders are more likely to cancel their policies compared to female policyholders, which is understandable since only 29.2% of the policies are from female policyholders.

5.3 GRADIENT BOOSTING MODEL

From the period 2020 to 2021:

Regarding the Gradient Boosting Model, the model achieved an accuracy of 80.34% for this period, which is higher than the accuracy obtained with the Logistic Regression model for the same period. For this model, we have the following confusion matrix and results:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	7719	720
Actual Canceled (1)	2592	5814

TABLE XI: Confusion Matrix for Gradient Boosting (2020-2021)

Price Variation_2020_2021:	0.72379
Age:	0.04317
Gender_M:	0.00667
Vehicle_Category_Passenger Car:	0.00888
Vehicle_Category_Motorcycle:	0.00223
Frequency_SEMESTRAL:	0.06028
Frequency_QUARTERLY:	0.00608
Frequency_ANUAL:	0.13739
Brand_Encoded:	0.00362
Model_Encoded:	0.00784

TABLE XII: Gradient Boosting Model Feature Importance 2020-2021

As it can be observed, it's noticeable that there's an improvement in the model performance, particularly in reducing false positive predictions of policy cancellations. The model correctly predicted 5814 out of 8604 canceled policies and 7719 out of 8439 active policies. Regarding the classification report, this shows a precision of 0.89 for predicting canceled policies, a recall of 0.69, and an F1-Score of 0.78. The overall macro and weighted averages for precision, recall, and F1-Score are all around 0.80. This means that the model performs quite well across both classes. Regarding the feature importance, in the Gradient Boosting model, the price variation variable in the GBM is a far more important factor with a feature importance of 0.7237, which highlights the significant role in predicting policy cancellations. In contrast, Gender Male, and vehicle categories have much smaller scores, which suggests that they are less influential in this nonlinear model, where interactions between variables are also considered. The difference between the coefficients from Logistic Regression and feature importance from GBM is due to the distinct nature of these two modeling approaches since GBM can capture complex relationships and focus more on the most predictive variable, while Logistic Regression treats the variables equally, assuming linear effects (Friedman, 2001).

For the Gradient Boosting models, we used the ROC curve to evaluate the performance of the models, the result of the curve for the 2020-2021 period is shown in figure 13:

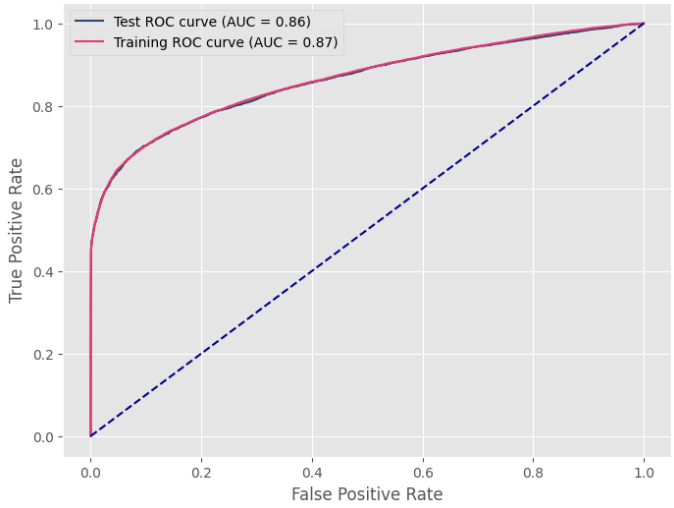

FIGURE 13: ROC Curve for Gradient Boosting Model 2020-2021

As we can observe, we have two different values for the AUC Test, 0.86, and AUC Training, 0.87. These values are very similar, which means that the model performs almost as well as the training set. This is a good indicator of the model's performance. The model is able to distinguish between annulled policies and active policies. It also means that there's no significant overfitting. The values of 0.86 and 0.87 represent a strong model that is balanced and would perform similarly well on new unseen data.

From the period 2021 to 2022:

For this period, the GBM had a high accuracy of 82.29% which means that 82% of the predicted canceled policies were correct. This value is slightly better than the value from the previous period, saying that the model performs better in terms of overall prediction accuracy. The confusion matrix and results are presented in tables XIII and XIV:

	Predicted Active (0)	Predicted Canceled (1)
Actual Active (0)	7413	1026
Actual Canceled (1)	1958	6448

TABLE XIII: Confusion Matrix for Gradient Boosting (2021-2022)

Price Variation_2021_2022:	0.77325
Age:	0.02675
Gender_M:	0.00396
Vehicle_Category_Passenger Car:	0.00524
Vehicle_Category_Motorcycle:	0.00089
Frequency_SEMESTRAL:	0.03724
Frequency_QUARTERLY:	0.00377
Frequency_ANUAL:	0.14080
Brand_Encoded:	0.00234
Model_Encoded:	0.00571

TABLE XIV: Gradient Boosting Model Feature Importance 2021-2022

Comparing this model with the model from the previous period, the model correctly predicted 6448 canceled policies, which is an increase from the prior period. However, it misclassified more active policies as canceled. Although there are more false positives in this period, the higher number of correctly predicted canceled policies improves the overall performance. Regarding the recall metric, there's an improvement in the 2021-2022 period, which indicates that the model has become better at identifying canceled policies. Despite this increase, the precision metric has slightly decreased from 0.89 to 0.86, indicating that this model has a higher false positive rate. The improvement of the F1-Score, from 0.78 to 0.81, means that the model in 2021-2022 has a better balance between precision and recall. The GBM coefficients show that the most significant predictor is price variation, with a feature importance of 0.7732, which means that price variations strongly influence policy cancellations. The features Frequency_ANUAL and Frequency_SEMESTRAL have a minor contribution compared to price variations, which indicates a smaller impact on the prediction. Variables like age, gender_M, and vehicle categories have very small values, which means that their influence in the model is minimal compared to price variation.

To evaluate the model, we used the ROC curve, as mentioned previously. For this period, the ROC curve is displayed in figure 14:

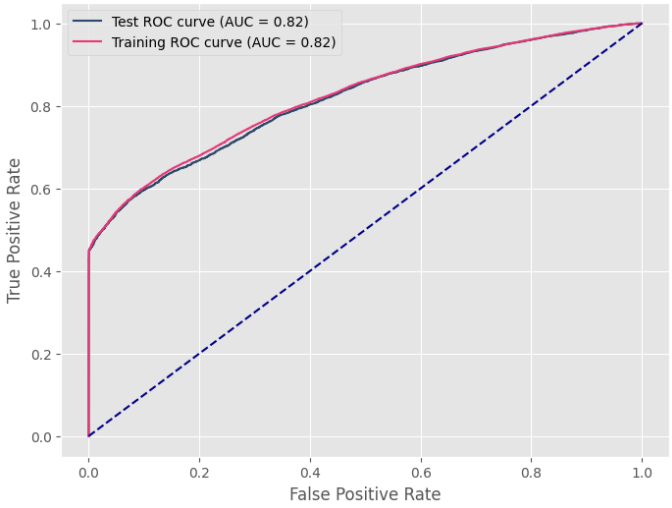

FIGURE 14: ROC Curve for Gradient Boosting Model 2021-2022

As we can observe, the AUC value is the same for the test and training, which indicates that the model is well calibrated and there's no overfitting. A value of 0.82 means that the model discriminates well between annulled policies and active policies and would perform well on new data. There was a small decrease in performance from 0.86 to 0.82 compared to the previous period. This shift could be due to data distribution changes, feature relevance, and increased noise or variability during the 2021 to 2022 period.

6 MODEL PERFORMANCE

In this section, we aim to use the Gradient Boosting Model, since it performed better, to predict, using the data from the active policies in 2022, which policies were likely to cancel if price increases by a certain percentage and how much revenue would be lost with those cancellations. The model applies different percentage increases in price, from 1% to 50%, in 5% intervals. The table XV shows the results from the model predictions:

Increase (%)	**Policies Canceled**	**Total Policies**	**Percentage Canceled (%)**
1%	20622	42111	48.97%
5%	20601	42111	48.92%
10%	21898	42111	52.00%
15%	20193	42111	47.95%
20%	20008	42111	47.51%
25%	13340	42111	31.68%
30%	32421	42111	76.99%
35%	26879	42111	63.83%
40%	28251	42111	67.09%
45%	28072	42111	66.66%
50%	28325	42111	67.26%

TABLE XV: Policies Canceled at Different Price Increases

As can be observed, with a price increase of 1%, 20622 active policies were canceled, equivalent to 48.97% of a total of 42111. At a price increase of 5%, it's seen that there's a minimal reduction in the canceled policies, with a 48.92% cancellation rate. As the price increases to 25%, the cancellation rate drastically drops to 31.68%, which may be related to the accident rate of policyholders. High accident rates often lead to higher premiums, so even if policyholders decide to leave their current insurer, they may face higher premiums elsewhere, discouraging them from switching. As the price increases to 30%, the cancellation rate has a noticeable jump and increases to 76.99%, affecting 32421 policies. This analysis shows what's expected from the insurance industry, that policyholders are very sensitive to price fluctuations. As price increases, the cancellation rate also increases, highlighting the relevance of managing premium adjustments to minimize revenue loss from canceled policies.

To understand the impact of these cancellations, when the price increases, on revenue, the model also predicted the percentage of revenue loss associated with those policies that are canceled. The table XVI shows the obtained results:

Price Increase (%)	**Revenue Loss (%)**
1%	49.28%
5%	49.61%
10%	52.79%
15%	48.93%
20%	48.52%
25%	33.65%
30%	74.59%
35%	62.60%
40%	66.00%
45%	65.58%
50%	66.14%

TABLE XVI: Revenue Loss at Different Price Increases

The table shows that at a small price increase (1% and 5%), the revenue loss remains stable at around 49%. This may suggest that the insurance company does not retain significant revenue even with minor price changes. At a 10% price increase, the revenue loss reaches almost 53%, which indicates that as the price rises also, the revenue loss increases. In contrast, when the price increases are 15% and 20%, the revenue loss drops to around 48%, suggesting that the initial shock of a price rise may have passed and cancellations have stabilized. As the price increases to 25%, there's a drop in revenue loss, which may suggest that the remaining policyholders are generating substantial revenue for the insurance company. However, after this drop, there's a spike in revenue loss, at 30% of price increase, with a 74.59% revenue loss, which may indicate a significant financial impact. From 35% to 50%, the percentages remain similar, with fluctuations between 62.60% and 66.14%, reflecting the considerable toll that higher price increases take on revenue.

In order to visualize the relationship between percentage price increases, percentage revenue loss, and percentage of cancellation, a three dimensional figure was made. The results had a few data points, we had to use interpolation to visualize the relationship better since this technique smooths the continuous surface that reflects a trend across the entire range of data. This allows to visualize intermediate values that weren't directly calculated in the original dataset. It's important to note that this technique estimates those values, and they are merely approximations. The three dimensional figure can be seen in figure 15:

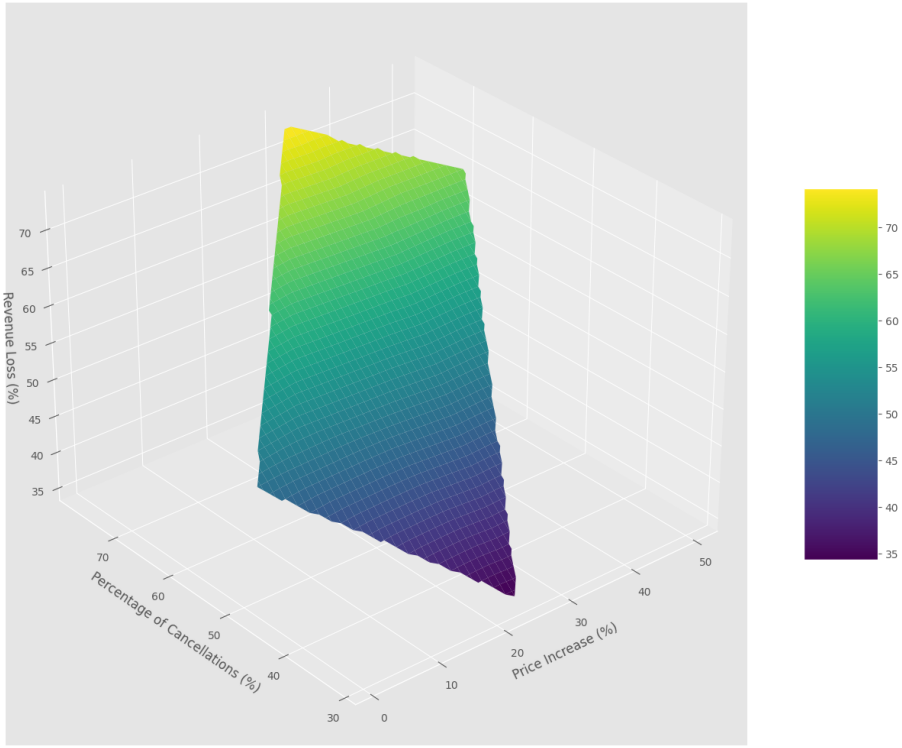

FIGURE 15: Relationship between % Revenue Loss, % Cancellation Rate, and % Price Increases

The figure shows that as the % price increase rises, the cancellation rate also increases, which leads to a higher revenue loss. This figure shows a nonlinear trend, with cancellation rates and revenue loss rising as price increases reach higher percentages. The slope in certain areas shows that beyond a certain price increase, a slight additional increase can cause significant policy cancellations and revenue loss. This visualization highlights the importance of price increases in cancellations and, consequently, in revenue loss. According to this model and data, insurers should be cautious about rising prices beyond certain thresholds, which could lead to severe revenue losses.

7 CONCLUSION

This project studied the influence of price elasticity in the auto insurance market. It tries to understand how premium changes influence policyholder behavior, especially in the context of policy cancellations. A dataset from 2020 to 2022 was used to study this behavior, and Logistic Regression models and Gradient Boosting models were implemented to predict the likelihood of cancellation. This analysis reveals that policyholders are very sensitive to premium changes, which leads to high cancellation rates.

From the results of the model, it was clear that price variations play a crucial role when it comes to predicting policy cancellations. It was also observed that policies for certain demographic groups, such as male policyholders and specific vehicle categories, were also more likely to be canceled, which highlights the importance of not only considering price variations to analyze policyholders' behavior but also accessing other variables to evaluate retention strategies.

The model was also tested through simulations of price increases, to understand the relationship between price increases and revenue loss. This showed a nonlinear relation since small price increases lead to a minimal reduction in revenue, but as price increases reach higher thresholds, there's a rise in both cancellation rates and revenue loss. This emphasizes the caution needed by insurance companies when raising their premiums since significant price increases may lead to high cancellation rates.

Although this project provides valuable insights into this topic, some limitations to this study may be found. The data used was limited to one insurance company in Portugal, and even though the dataset had a reasonable amount of information, expanding this study to more than one insurance company would be beneficial to understanding trends in price elasticity. This study only focused on auto insurance data, so future research could explore the impact of price elasticity across different types of insurance industries. Furthermore, the models that were used were trained on historical data. Although they had a good performance, external factors may have a strong impact on policyholder behavior, so they cannot capture other key variables. These external variables and market context are relevant in future research on consumer behavior.

This study concludes that not only do price fluctuations strongly influence policy cancellations, but other variables such as policyholder demographics, vehicle characteristics, and payment frequency also play an important role in assessing the reasons that lead to policy cancellations. Therefore, using machine learning models to predict policyholder behavior may be an interesting approach to understanding policyholder behavior and, with that, finding structured pricing strategies to minimize customer loss while maintaining profitability.

REFERENCES

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Antonio, K. and Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, 96(2):187–224.

Bach, M. J. (1990). *The Design of the Unix Operating System*. P T R Prentice-Hall, Englewood Cliffs, NJ.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1997). *Actuarial Mathematics*. Society of Actuaries, 2nd edition.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Coalition, M. C. R. (2022). Gender discrimination in auto insurance: A report by the maryland consumer rights coalition.

Darlington, R. B. and Hayes, A. F. (2017). *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*. The Guilford Press, New York.

de Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press, Cambridge, UK.

de Supervisão de Seguros e Fundos de Pensões, A. (2023). Relatório de evolução da atividade seguradora - 4º trimestre de 2022. `https://www.asf.com.pt/documents/42559/1543486/REAS_4T_2022.pdf/4ee240a7-1c68-afc8-42b6-a5ede71e835a`. Accessed: 2024-08-12.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38:367–378.

Girard, L. (2024). Analysis of factors influencing automobile insurance premiums in france. *Journal of Statistics and Actuarial Research*, 8(2):1–10.

Gonçalves, L., Subtil, A., Oliveira, M. R., and Bermudez, P. d. Z. (2014). Roc curve estimation: An overview. *REVSTAT – Statistical Journal*, 12(1):1–20.

Grace, M. F., Klein, R. W., and Kleindorfer, P. R. (2001). The demand for homeowners insurance with bundled catastrophe coverages. Working Paper 69, Johann Wolfgang Goethe-Universität Frankfurt am Main, Fachbereich Wirtschaftswissenschaften.

Guelman, L. and Guillén, M. (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, 41(2):387–396.

Gujarati, D. N. and Porter, D. C. (2009). *Basic Econometrics*. McGraw-Hill/Irwin, New York, 5th edition.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2nd edition.

Harrington, S. E. and Niehaus, G. (2003). *Risk Management and Insurance*. International edition. McGraw-Hill.

Henley, W. E. (2014). *Poems*. CreateSpace Independent Publishing Platform.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Hoboken, NJ, 3rd edition.

Kelly, M. and Nielson, N. (2006). Age as a variable in insurance pricing and risk classification. *Geneva Papers on Risk and Insurance - Issues and Practice*, 31(2):191–214.

Kunreuther, H. and Pauly, M. (2006). Insurance decision-making and market behavior. *Foundations and Trends in Microeconomics*, 1(2):63–127.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2 edition.

Meyer, R. L. and Power, F. B. (1973). Total insurance costs and the frequency of premium payments. *The Journal of Risk and Insurance*, 40(4):599–605.

Mukhiya, S. K. and Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing.

Murray, L., Nguyen, H., Lee, Y.-F., Remmenga, M. D., and Smith, D. W. (2012). Variance inflation factors in regression models with dummy variables. *Conference on Applied Statistics in Agriculture, 24th Annual Conference Proceedings*, 24:161–175.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nicholson, W. and Snyder, C. M. (2019). *Microeconomic Theory: Basic Principles and Extensions*. Cengage Learning, 12th edition.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Schlesinger, H. (2013). The theory of insurance demand. In Dionne, G., editor, *Handbook of Insurance*, pages 167–184. Springer New York, New York, NY.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2):39–42.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192.

Turkman, M. A. A. and Silva, G. L. (2000). *Modelos Lineares Generalizados - da Teoria à Prática*. Edições SPE.

Varian, H. R. (2014). *Intermediate Microeconomics: A Modern Approach*. W. W. Norton & Company, 9th edition.

Winter, R. A. (1992). *Moral Hazard and Insurance Contracts*, pages 61–96. Springer Netherlands, Dordrecht.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage Learning, 6th edition.

Zweifel, P. and Eisen, R. (2012). *Insurance Economics*. Springer Texts in Business and Economics. Springer, Berlin, Heidelberg.

PYTHON CODE

Since the Python code is extensive, below is the link to the Google Colab Notebook, where you can find the code used in this project.

Price Elasticity in Auto Insurance Code

DISCLAIMER

This master project was developed with strict adherence to the academic integrity policies and guidelines set forth by ISEG, Universidade de Lisboa. The work presented herein is the result of my own research, analysis, and writing unless otherwise cited. In the interest of transparency, I provide the following disclosure regarding the use of artificial intelligence (AI) tools in the creation of this project: I disclose that AI tools were employed during the development of this thesis as follows:

- AI-based research tools were used to assist in literature and data collection
- AI-powered software (Large Language Models) was used to assist in the creation of the Python code
- AI-powered software (Large Language Models) was used for formatting and sorting the references

Nonetheless, I have ensured that the use of AI tools did not compromise the originality and integrity of my work. All sources of information, whether traditional or AI-assisted, have been appropriately cited in accordance with academic standards. The ethical use of AI in research and writing has been a guiding principle throughout the preparation of this thesis.

I understand the importance of maintaining academic integrity and take full responsibility for the content and originality of this work.

Filipa Vaz, 15th of October 2024