NOVA IMS
Information Management School

MEGI

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

# IMPACT OF ADAS IN INSURANCE PRICING

Pedro Manuel Rodrigues Jordão

Dissertation presented as partial requirement for obtaining the Masters 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

# IMPACT OF ADAS IN INSURANCE PRICING

by

Pedro Manuel Rodrigues Jordão

Dissertation presented as partial requirement for obtaining the Masters degree in Statistics and Information Management, with a specialization in Risk Analysis and Management

**Advisor:** Profª Doutora Gracinda Rita Diogo Guerreiro

## NOVEMBER 2022

# DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisboa, 30 November 2022

Pedro Manuel Rodrigues Jordão

[the signed original has been archived by the NOVA IMS services]

**ACKNOWLEDGEMENTS**

Firstly, I would like to thank my advisor Profª Doutora Gracinda Guerreiro for all the support and guidance throughout this process.

I would also like to thank all my colleagues and friends who have always supported me.

Finally, a special thanks to my family who always encouraged me and pushed me to finalize this study.

# IMPACT OF ADAS IN INSURANCE PRICING


## ABSTRACT

According to EU commission statements dated of 2019, around 95% of car accidents in the EU have some type of human error involved. Autonomous cars may reduce this number significantly and improve road safety thanks to many softwares and driving aids.

It is expected that the self-driving vehicles' market grows exponentially and reaches profits of up to 620 billion euros by 2025. This is due to its environmental footprint and the reduced financial cost of electric energy. However, this increasing autonomy has raised awareness in the insurance industry. Cars' automation is not currently a factor used in insurance pricing models.

The purpose of this project is to detail the difference in claim frequency and claim amount for cars with ADAS (Advanced Driver Assistance Systems) and analyze if this should be taken into account when calculating car insurance premiums.

Generalized linear models, along with Decision Trees will be used to analyze the impact of ADAS in aggregated claim amount.

# KEYWORDS

ADAS

Car Insurance

Generalized Linear Model

Insurance Premium

Decision Tree

# ACRONYMS

ADAS – Advanced Driver Assistance Systems

AEB – Automated Emergency Brake

GLM – Generalized Linear Model

GPRS – General Packet Radio Service

HLDI – Highway Loss Data Institute

IIHS – Insurance Institute of Highway Safety

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

In this project, when using the term “autonomous cars”, it is intended to refer to cars with driving aids.

Autonomous and semi-autonomous vehicles are capable of, through instructions given by a specific hardware and software, identifying routes, traffic lights and people, giving the driver a sense of security when driving.

According to Swiss Re, in 2019 there were 30% less rear-end accidents due to Automated Emergency Brake systems (AEB).

## 1.1. HISTORY OF ADAS

Although ADAS seems like a 21st century technology, these systems were first introduced in 1920s when automatic braking systems (ABS) were developed for aircrafts.

However, it wasn´t until the 1970s that Robert Bosch patents, in partnership with Mercedes-Benz, started to be used in automobiles.

The first model to have this system was the 1971 Chrysler Imperial calling it “Sure Brake”. Nissan also had an electronic ABS system introduced in their Nissan President Sedan.

Another ADAS used since the 1960s was speed warning systems, with the 1962 Buick Wildcat having a speed indicator that could be set by the driver. When exceeding the speed defined, a sound would be heard to warn the driver to slow down.

Some of this assist systems, while not named as true ADAS technology, can be seen as early forms of driver assistance functionality.

IDTechEx, in the "Autonomous Cars, Robotaxis & Sensors 2022-2042" report, predicts that “SAE level 3 autonomous features, such as traffic jam pilot, will emerge in 2022 and become common in vehicles by 2042”.

This conclusion is supported by the evolution of ADAS adoption in the last years.

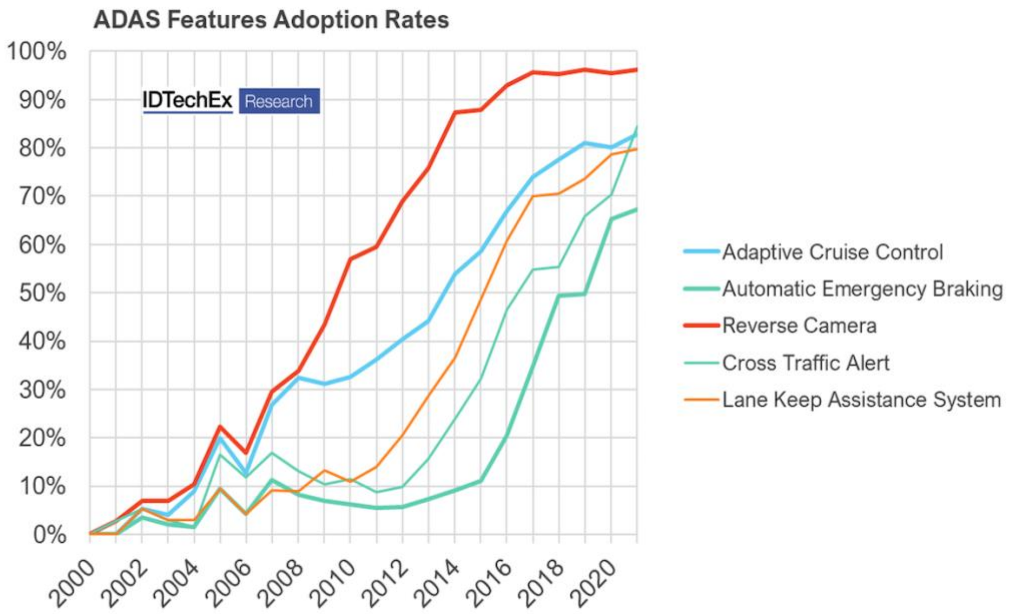



Figure 1: The progression of ADAS feature adoption over the last 20+ years. Font: IDTechEx Research

It can be observed that reverse camera is the most widely used ADAS for the past years, being present in more than 95% of 2021 models. Adaptive cruise control is also one of the most used ADAS and automatic emergency brake rose from 55% adoption rate in 2018 to 85% in 2021.

Nowadays, there are a lot of driving aids used in the auto industry such as:

- **Hands-free steering**;
- **Adaptive cruise control** allows the vehicle to adapt the speed according to the car ahead while keeping a safe distance;
- **Autobrake**: capacity of breaking when detecting a moving object ahead;
- **Lane-centering steering** prevents vehicle from crossing lanes without using the turn signal;
- **Blind spot warning** warns the driver in case of vehicle passing the blind spot area when changing lanes;
- **Cameras** are used to assist the driver when parking along with front/rear collision warnings;
- **Self-parking**.

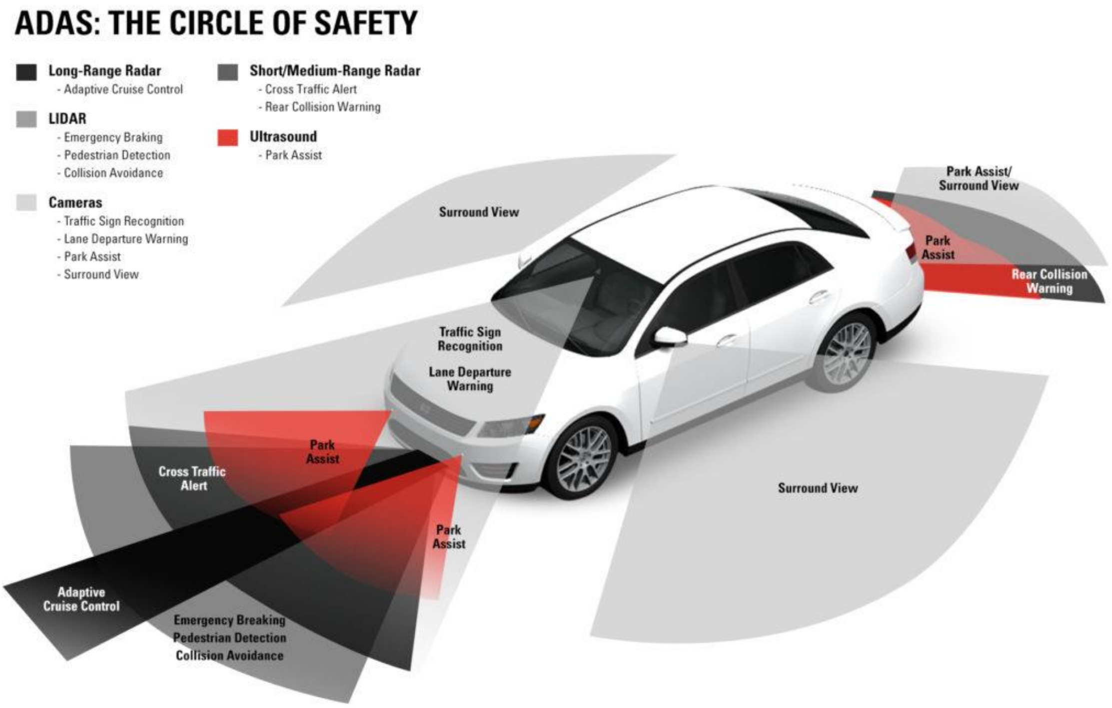



Figure 2: Common ADAS used in automotive industry. Font: Precision Autoglass

One field where this topic is still facing many challenges is the legal one by raising issues with accidents involving autonomous cars and conventional cars or both autonomous cars.

The European Union recognizes many advantages of self-driving, such as safer roads, protection of the environment, better accessibility, new jobs and economic growth.

Autonomous driving faces many challenges in the EU, such as:

**Road safety:** because roads are inevitably shared by driverless, non-automated vehicles, bicycles and pedestrians, it is essential to have appropriate safety requirements.

**Liability issues:** since part of the driving process is passed to the vehicle, liability laws must cover this transition and clarify who is liable in each case of accident.

**Data processing:** although the automated sector is also covered by data protection rules, there are not particular measures to protect driverless cars from cyberattacks.

**Ethical questions:** the EU is already drafting guidelines for artificial intelligence, but specific measures might be necessary for automated cars to respect freedom of choice and human dignity.

**Infrastructure:** to build the necessary infrastructure to support new developed technologies, it is necessary to invest in innovation and research.

Although driving aids exist and tend to be the future of driving, in most countries there are not fully autonomous cars on the road yet. However, the presence of electric and partially autonomous cars is increasing in the industry.

In Portugal, total sales of electric cars till August 2021 increased 63% from last year and in September 2021 represented around 25% in sales, although it still reflects a smaller percentage in the total number of vehicles.

Regarding electric cars vehicles' insurance, the first impression would be that insurance premium would be cheaper due to the driving aids, however sometimes it may not be the case as electric batteries are more expensive to replace and electric cars are overall costlier than conventional cars.

## 1.2. TELEMATICS

Telematics can be defined as the fusion of two technologies – informatics and telecommunications – and plays an extremely important role in ADAS implementation and growth.

Essentially, a telematics system involves a vehicle tracking device that can generate and transmit GPS and other vehicle-specific data through a wireless network via GPRS (General Packet Radio Service), satellite communication or 4G mobile data to a centralized server. This data can then be distributed to end users using optimized apps for tablets and smartphones.

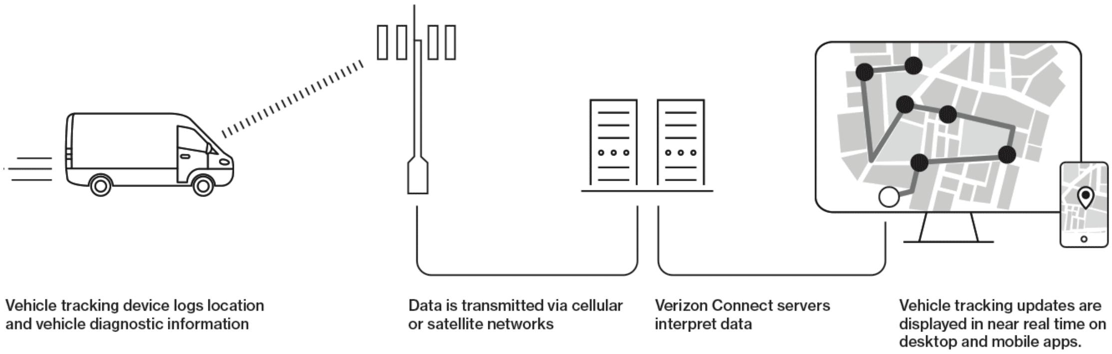



Figure 3: How does a telematic system work? Font: Verizon Connect

In the case of ADAS implementation, these end users can be other telematics systems installed in other vehicles, allowing the vehicle to receive data from its surroundings. This data can be speed, location, harsh acceleration or breaking, idling time, vehicle faults, fuel consumption, and much more.

Nowadays, with the introduction of 5G technology, data processing and transmission will increase and allow for real-time vehicle-to-vehicle communications, which will be an added value for ADAS use by receiving data about streetlights, estimated arrival time, traffic congestion, etc. and allowing each vehicle to adapt its driving to its context.

## 1.3. Background and Problem Identification

All European countries are trying to implement models to regulate autonomous cars' insurance, but most of them don't have an official legal support. However, some countries have made some developments on this matter.

The United Kingdom was one of the first countries with significant improvement regarding this matter, as legislators approved the Automated and Electric Vehicles Act in 2018.

This Act states that the insurance company is liable when the vehicle is driving itself at the time of the accident. This liability can only be excluded or reduced if there is proof that the accident resulted from prohibited software alterations or lack of software updates.

Germany also has an insurance model regarding partly automated vehicles. In this model, the car owner is liable towards the party that incurred damage whether the accident was a result of vehicle error or driver's fault, unless there's proof that the driver didn't cause the accident.

In these cases, the insurance regime would pay the victim while confirming who's the actual liable entity. This complicates the situation since there are many possible liable entities (driver, car manufacturer, fleet operator, supplier, software programmer).

In Portugal, there is not much knowledge on how this insurance regulation is approached and if autonomous cars' owners are benefited in the insurance premium. There isn't a specific law for accidents involving electric or autonomous cars. Although prices show that insuring an electric vehicle is most of the times cheaper, there isn't certainty that it relates directly to the autonomy component.

The first step into autonomous driving in Portugal was in 2019, where an autonomous vehicle was tested in a pre-defined route next to Nova School of Business & Economics in Carcavelos with an insurance by Ageas Seguros.

In the same year, a Portuguese team of students from ISEG University of Lisbon ranked 3rd in the 2019 Society of Actuaries Student Research Case Study Challenge (see Bello, F., Gudmundarson, R., Meindersma, J., Oliveira, C. and Xu, Z. (2019), *Autonomous Vehicles Insurance Policy)*. The intent was to create an insurance for autonomous vehicles regarding other types of risks rarely used in non-autonomous vehicles' insurance, such as Infrastructure Risk, Cyber Risk and Malfunction Risk.

Fully autonomous cars are only expected to appear in abundance in Portugal in 2025, however, semi-autonomous cars and other types of artificial intelligence in vehicles is and will be strongly presented in the upcoming years.

In 2019 Swiss Re introduced a risk score, in partnership with BMW, to help insurers to better price vehicles with Advanced Driver Assistance Systems (ADAS). A research conducted by the company showed that, although these new technologies can reduce loss frequencies, most insurers don´t considers in their products the impact of ADAS.

The company then partnered with Toyota Insurance Services in September 2020, who joined the ADAS risk score platform, allowing data from Lexus and Toyota into the system.

As part of building the risk score, Swiss Re studied the real impact of ADAS on claims and road safety, as well as the impact of the improvement of these technologies.

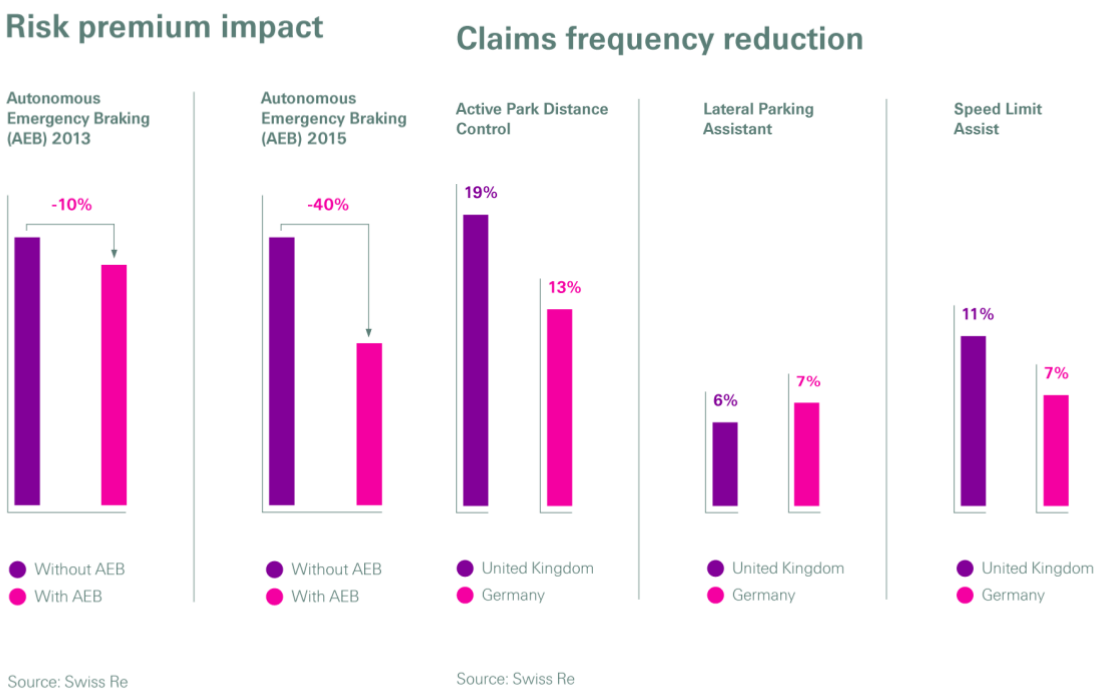



Figure 4: Risk premium impact and Claims frequency reduction. Font: Swiss Re

As seen on the graphs of Figure 4, the difference between risk premium in vehicles with and without Autonomous Emergency Braking was 10% in 2013 and 40% in 2015.

Also, claims' frequency related to active park distance control, lateral parking assistant and speed limit assistance reduced between 6% and 19% in the United Kingdom and Germany.

## 1.4. Study Objectives

The main objective of this project is to answer the question: What is the impact of ADAS in auto insurance pricing?

To answer this question, the following sub-questions are going to be approached:

- Are frequency and severity of claims different in automated cars?
- Should insurance premiums be less for cars with ADAS?
- How can the difference in aggregated claim amounts be used when pricing automated cars' insurance?

# 2. STUDY RELEVANCE AND IMPORTANCE

Usually, a car insurance's premium is based, between other factors, on driver's related factors such as age, experience, etc., but what happens in the case of vehicles with ADAS, where people's contribution to the driving process is less than usual?

This topic is important because with the exponential growth in sales of vehicles with automation features it is advisable that the impact of buying these types of cars in their insurance's premium is clear to the public so that people can take a more informed decision.

It is understandable that there is not much information surrounding this matter because it is a relatively new topic. Although cars with ADAS have been around for a long time, the diversification and market growth are still recent.

This study intends to clarify and extend the little information existent on the topic, by having both a theoretical and practical approaches.

# 3. STATE OF THE ART

One paper was published by the data analysis company LexisNexis (Khanet, J., Kohli, P. (2020), *ADAS Analysis Creates Path for Auto Insurance Pricing*, LexisNexis). According to this paper, “advanced driver assistance systems result in fewer claims and reduce insurers loss costs by up to 23%”.

The company analyzed data from 11 million vehicles, with model years ranging from 2014 to 2019, comparing industry-wide loss data to the severity and claim frequency for vehicles equipped with each of 648 possible combinations of the 11 “core” ADAS features.

Kanet said in a press release that “Insurance companies can use this new ADAS data as part of their rating segmentation to better meet the expectations of insurers’ customers, who often purchase vehicles with advanced safety features with the expectation that it will help lower their insurance rates”.

There were identified 10 ADAS features that have the highest impact on loss cost, with this being: adaptive cruise control, blind spot warning, blind spot mitigation, driver monitoring, forward collision warning, forward collision mitigation, lane departure warning, lane departure mitigation, rear collision warning and rear collision mitigation. Based on the sample analyzed, 38% of vehicles had at least one core ADAS feature, with 76% of 2019 models having an ADAS feature.

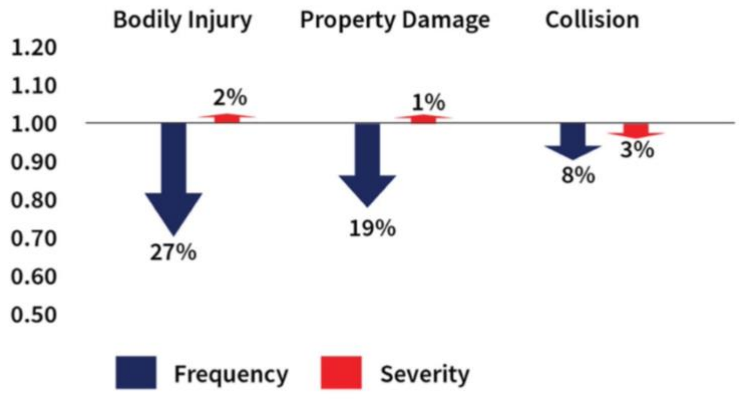



Figure 5: Claim Relativities for 1+ Core ADAS Features. Font: LexisNexis paper

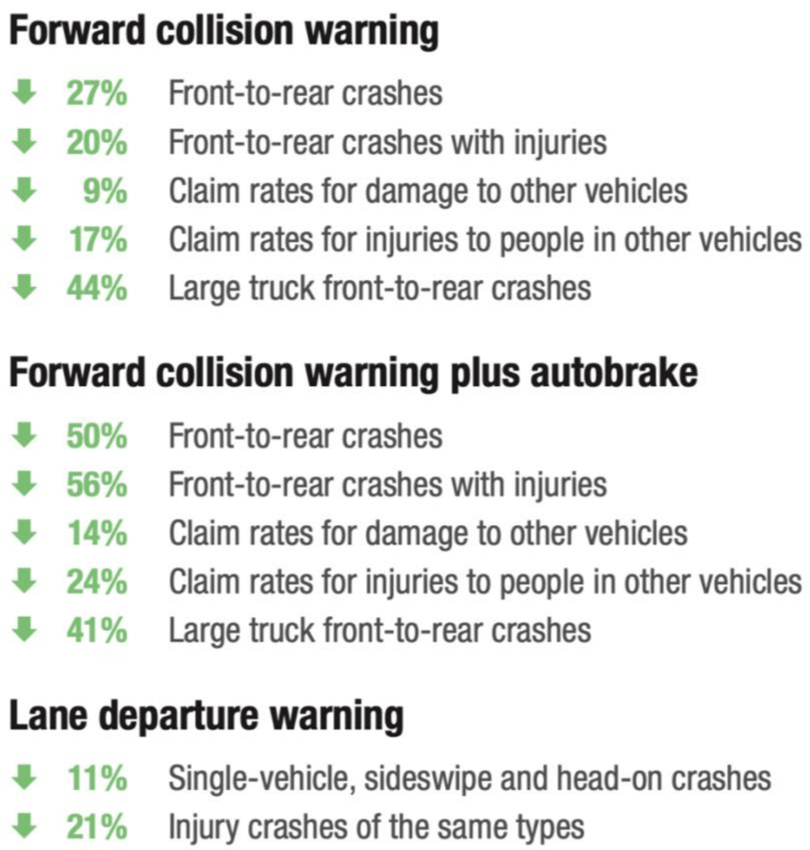
LexisNexis found that ADAS have a greater impact in claim frequency than in claim severity. Vehicles with ADAS had 8% less collision claims, 19% less property damage claims, and 27% less bodily injury claims compared to vehicles without ADAS features. In contrast, claim severity saw minimal movement as observed in the graph below.

The study concluded that “the results of this research show a clear opportunity to create significant pricing segmentation for vehicles equipped with core ADAS features”.

The Insurance Institute for Highway Safety (IIHS) and the Highway Loss Data Institute (HLDI), which data will be used in this dissertation, studied the real-world benefits of crash avoidance technologies.

The companies compared rates of “police-reported crashes and insurance claims for vehicles with and without the technologies”.

The results obtained were the following:



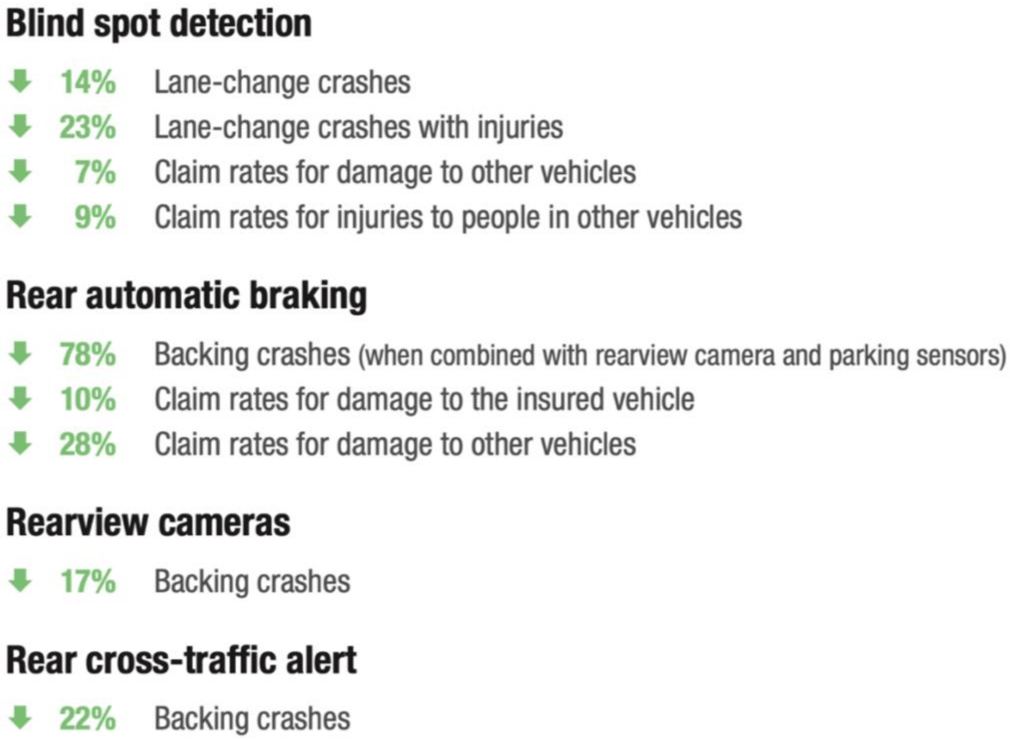



Figure 6: IIHS and HLDI's study results. Font: IIHS and HLDI

From the results it can be concluded that forward collision warning plus autobrake is one of the most effective ADAS, reducing by 50% and 56% front-to-rear crashes with and without injuries, respectively.

Another company has been very active throughout the years in studying the effect of automation on insurance premiums.

Celent is a research and consulting firm that applies information technology in the financial services industry.

The first report (Celent (2012), *A Scenario: The End of Auto Insurance: What Happens When There Are (Almost) No Accidents?*) was made in May 2012 and concluded that:

- "Celent believes that this scenario is plausible, and that the probability of a substantial reduction in auto insurance premium occurring is sufficiently high that auto insurers should devote some resources to considering the scenario and its implications for their business model and enterprise".

Four years later, in June 2016, the company published an update (Celent (2016), *The End of Auto Insurance: A Scenario or a Prediction?*), which expressed more confident conclusions and recommendations:

- "During the next 15 years, auto insurers will likely see their business shrink (and continue to shrink thereafter as driverless cars become a larger part of the fleet). Some insurers are beginning to address this existential challenge now. The rest will have to do so in a few years".

Recently, and with the major development in the auto industry, Celent published its most recent report in August 2019 (Celent (2019), *Re-rethinking The End of Auto Insurance*).

The report analyzed the impact of telematics and ADAS in the upcoming years:

| TECHNOLOGY | EFFECTIVENESS | SHARE OF THE FLEET ON THE ROAD |
|---|---|---|
| TELEMATICS | ●●● | ●●●● |
| ADAS | ●●● | ●●● |
| TELEMATICS COMBINED WITH ADAS | ●●●● | ●●● |
| AUTONOMOUS CARS | ●●●●● | ● |

Figure 7: The Impact of the Four Technologies by 2029. Font: Celent (2019), *Re-rethinking The End of Auto Insurance*

Telematics, autonomous technologies and ADAS will have a big effect in reducing accidents, auto insurance losses and premiums.

The proportion of these technologies in the fleet on the road, however, will grow differently until 2029:

- Telematics and ADAS will grow relatively quickly.
- Autonomous technologies will grow much more slowly.

# 4. METHODOLOGY

The data to be used on this work is provided by the Insurance Institute for Highway Safety (IIHS) and the Highway Loss Data Institute (HLDI) and includes the percentage related to the average of the aggregated claim amount for different car models and six insurance coverages: collision, property damage liability, comprehensive, personal injury, medical payment and bodily injury. Injury coverages only reflect claim frequency and will be analyzed separately.

In order to use the data, it will be created a dataset with the following variables:

- The coefficient for each insurance coverage for each model;
- Car's brand;
- Car's model;
- Year of manufacture;
- Level of automation.

## 4.1. SELECTION CRITERIA

In this project, cars are going to be divided in four categories based on the levels defined by U.S. National Highway Traffic Safety Administration:

**Level 0:** Traditional Vehicle with no ADAS;

**Level 1:** ADAS help the driver with braking, steering or accelerating but not simultaneously. ADAS included in this category are cameras and vibrating seat and steer warning;

**Level 2:** While the driver remains fully aware, ADAS can either brake, accelerate or steer simultaneously. The driver still has full responsibility of the driving process;

**Level 3:** In this level, an Automated Driving System (ADS) is responsible for all the driving process under certain circumstances, but the driver is still responsible to retake control of the vehicle if necessary. ADAS differentiating this level include active parking assist.

According to the U.S. National Highway Traffic Safety Administration, there are 6 levels of automation but in this study we will discard levels 4 and 5 as they are not realistic to include in vehicles built before 2020.

## 4.2. PREMIUM ESTIMATION

In non-life insurance pricing, premiums are often calculated as the expected value of the aggregated claim amounts, i.e., considering the aggregated claim amounts as the random variable

$$S = \sum_{i=0}^{N} Xi$$

with N being the number of claims in one year and $X_i$ the cost of the $i^{th}$ claim. Considering the actuarial principle of equivalence, the pure premium is given by

$$PP = E(S) = E\left[E[S|N]\right] = E(N)E(X)$$

considering variables $X_i$, i=1, …, N are i.i.d. and that the variables N and X are independent.

However, considering that the data used for this project is already a combination of the claim frequency and claim amount, we will deal with modelling the aggregate amount S.

## 4.3. GENERALIZED LINEAR MODELS (GLM)

Statistical modelling is one of the most important areas of applied statistics and can be applied in many fields of scientific research, such as economy, ecology, sociology, insurance, agronomy and medicine, to cite but a few. Amongst many statistical modelling frameworks, Generalized Linear Models (GLMs) are commonly used for estimating pricing structures in automotive insurance (Garrido, J., Genestb, C. and Schulz, J. (2016), *Generalized linear models for dependent frequency and severity of insurance claims*).

Some of the cases of GLMs are the Gamma, Gaussian, inverse Gaussian and Poisson regression models. These models are linked, since they belong to the class of Exponential Dispersion (ED) models and share the property to be described by their first two moments, mean and variance. The variance function is extremely important in ED models, since it describes the relationship between the mean and variance and characterizes the distribution.

Generalized linear models are a generalization of ordinary linear regression. The two main differences in these models are the response variable associated with a link function and the variance of each measurement following a function of its expected value.

These models were initially formulated by John Nelder and Robert Wedderburn (Nelder, John; Wedderburn, Robert (1972). "*Generalized Linear Models*". Journal of the Royal Statistical Society. Series A (General). Blackwell Publishing. 135 (3): 370–384) as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression.

The general form of a GLM composed by the following:

**Random Component** - specifies the probability distribution of the response variable; e.g., normal distribution for Y in the classical regression model, or Binomial distribution for Y in the binary logistic regression model. This is the only random component in the model; there is not a separate error term (or any other distribution provided belongs to the Exponential Family of Distributions).

**Systematic Component** - specifies the explanatory variables ($x_1,x_2,\ldots,x_k$) in the model, more specifically, their linear combination; e.g., $\eta= \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k$.

**Link Function, g()** - specifies the link between the random and the systematic components. It indicates how the expected value of the response variable relates to the linear combination of explanatory variables.

In any GLM, the following assumptions are also considered:

- The variables $Y_1$, $Y_2$, …, $Y_n$ are independently distributed.
- Explanatory variables can be nonlinear transformations of some original variables.
- The homogeneity of variance does not need to be satisfied. In fact, it is not even possible in many cases given the model structure.
- Errors need to be independent but not normally distributed.
- Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS).

Regarding R software implementation of the generalized linear models, the first software package commonly used for fitting these models was called GLIM (6.1 – Introduction to GLMs, PennState Eberly College of Science). Today, GLMs are fit by many packages, including SAS's Genmod procedure and R's glm() function.

## 4.4. FREQUENCY

Frequency is a discrete random variable and in this project it will be used to analyze two insurance coverages (personal injury and bodily injury).

Considering N(t) the number of claims filled for a certain period of length t, the frequency can be calculated as:

$$F(t) = \frac{N(t)}{\exp(t)}$$

Exposure represents the risk present in a portfolio.

Depending on if overdispersion exist (variance is higher than expected value), claim frequency can follow a Poisson distribution or a Negative Binomial distribution.

If overdispersion doesn't occur, it can be test if the claim frequency follows a Poisson distribution, with probability mass function

$$f(x) = P\,(N = x) = e^{-\lambda}\frac{\lambda^x}{x!}\,,\ x \in N_0\,,\ \lambda > 0$$

for which it is well known that E(X) = VaR(X) = λ.

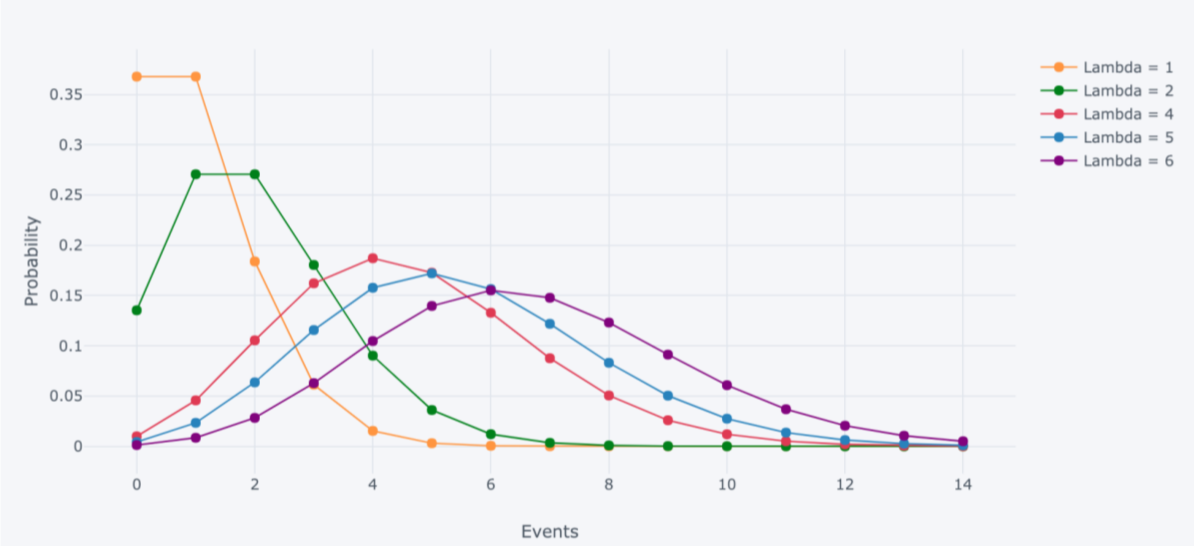



Figure 8: Poisson distribution

If the variance is higher than expected value, then the hypothesis of the Poisson distribution can be excluded and claim frequency may follow a Negative Binomial distribution with probability mass function

$$f(x) = \binom{r+x-1}{r-1} p^r (1-p)^x$$

where E(X) = μ and VaR(X) = μ(1+μ/r)

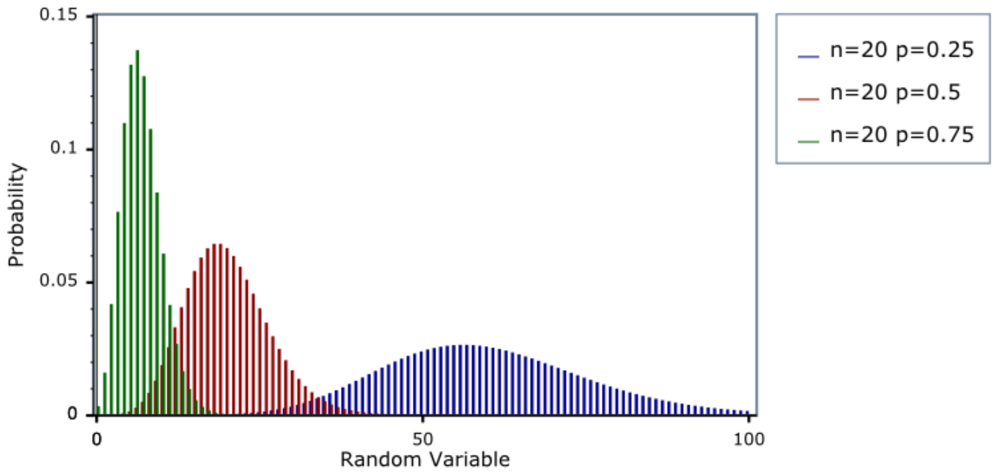



Figure 9: Negative Binomial Distribution

## 4.5. DECISION TREES

A decision tree is a representation of decision for the purpose of decision analysis. Although widely used in data mining to derive a plan to reach a particular goal, decision trees are also used in machine learning.

Decision Trees are one of the most used tools for prediction and classification. In this flowchart-like tree structure, each node represents a test on an attribute or variable, each branch is the result of the test and each terminal node denotes a class label.

<u>How can an algorithm be represented as a tree?</u>

As an example, will be considered a mock decision tree for the probability of a car accident in the Iberian peninsula. The below model considers as mock variables: country, age and brand.

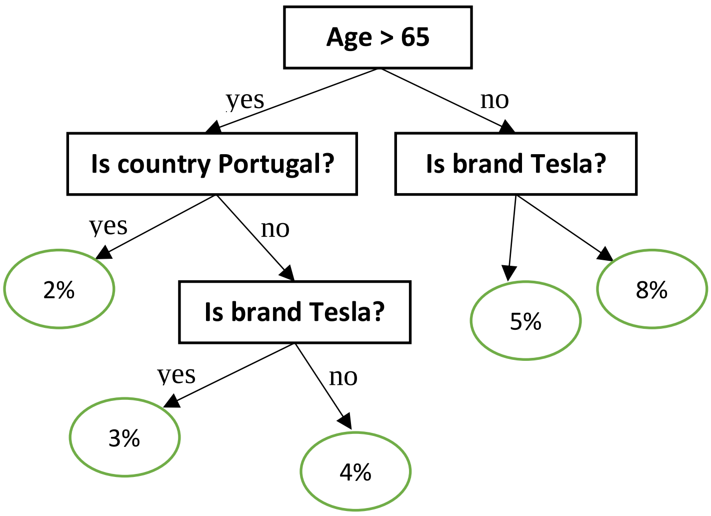



Figure 10: Example of a Decision Tree.

Decision trees start at the top, also called root. In the above image, each bold text represents an internal node and from there the tree splits into branches based on the result of the internal node. When the tree reaches the end of a branch and doesn´t have any more splits, it's called a decision/leaf which here represents the probability of a car accident.

To understand how algorithms work in decision tree's background, there are a few things to consider, such as which features to choose and which conditions to use for splitting, as well as knowing when to stop.

Recursive Binary Splitting

In this approach, all attributes/variables are considered, with different split point being tested using a cost function. The aim is to select the splits at the best cost (or lowest cost).

This method is considered recursive as the classes formed can be sub-divided using the same strategy. Based on this procedure, it is also known as the greedy algorithm, as there is an excessive desire of lowering the cost.

Cost of a split

Cost functions are used for two main areas in decision trees: classification and regression. Both cost functions have the aim of finding the most homogenous branches, i.e., branches having groups with similar responses, to make sure the test data follows a certain path.

As an example, it will be considered the prediction of houses' prices. The tree will begin splitting considering each variable in the test data. The expected value for a specific group is considered to be the mean of the responses of the data for that group.

The function above is applied to all data points and, from there, the cost is calculated for all splits. The split choses will be the one with the lowest cost.

$$\text{Classification: } G = \text{sum}\,(pk \times (1 - pk))$$

Gini scores "rate" each split according to the diversity of different response classes in the group created by the split. pk is the proportion of same class data samples present in a particular group. When pk is either 1 or 0 and G = 0, it is considered a perfect class purity because the group only has inputs from the same class. A worst class purity occurs when having a node with a 50–50 split of classes in a group, meaning for a binary classification pk = 0.5 and G = 0.5.

When to stop splitting?

The problem with a data set with many variables is when to stop splitting, as the tree will have a large number of splits and many more branches, and therefore can be harder to analyze.

One method that can be used is to define a minimum number of inputs to use in each leaf. Any leaf that considers less than that input will be ignored.

Alternatively, a maximum depth can be defined for the tree to not have longer paths from a root to a leaf than desired.

Pruning

The performance of a decision tree can increase by pruning. Branches that use variables of low importance are removed, reducing the complexity of the tree, and increasing its predictive power by reducing overfitting.

Pruning can start at the leaves or the root. The easiest method begins at leaves and removes each node with most popular class in that leaf, this change is done only if accuracy (number of correct predictions made divided by the total number of predictions made) is not affected. It's also known as reduced error pruning.

The strengths of decision tree methods are:

- Decision trees are easy to understand;
- Classification in decision trees does not require much computation;
- Both categorical and numerical variables can be used;
- Most important fields for prediction or classification are clearly indicated.

The weaknesses of decision tree methods:

- Decision trees are not the best approach for predicting the value of continuous variables.
- Errors are more common in classification of small data sets with many classes.

# 5. DATASET INFORMATION

The data used on this project is collected from the Insurance Institute for Highway Safety (IIHS) and the Highway Loss Data Institute (HLDI) website and refers to data from United States of America for vehicles from 2014 to 2020 (Insurance losses by make and model, IIHS and HLDI).

The insurance coverages are defined by IIHS and HLDI as:

- **Collision:** insures against physical damage to the vehicle in a crash if it is the user’s fault. The damage may occur from striking another vehicle or an object.
- **Property Damage:** insures against physical damage that at-fault drivers cause to other people's vehicles and property in crashes.
- **Comprehensive:** insures against theft or physical damage to insured people's own vehicles that occurs for reasons other than crashes.
- **Medical Payment:** covers injuries to insured drivers and the passengers in their vehicles but not injuries to people in other vehicles involved in the crash.
- **Personal Injury:** pays up to a specified amount for injuries, regardless of who is at fault in a collision.
- **Bodily Injury:** insures against injuries that at-fault drivers inflict on people in other vehicles.

The data is provided as a percentage related to the average of the aggregated claim amount for the first four coverages mentioned above and to the claim frequency for the last two coverages as shown in Figure 11.

| Vehicle | Collision | Property damage | Comprehensive | Personal injury | Medical payment | Bodily injury |
|---|---|---|---|---|---|---|
| Audi Q3 4dr 4WD | 14% | -3% | -11% | -29% | -31% | -10% |
| BMW X1 4dr | 3% | -7% | 4% | -5% | 3% | -8% |
| BMW X1 4dr 4WD | -5% | -12% | 6% | -29% | -22% | -21% |
| BMW X2 4dr | 13% | -16% | 56% | -4% | | |
| BMW X2 4dr 4WD | 18% | 4% | 49% | -22% | | 7% |
| Buick Encore 4dr | -15% | 14% | -43% | 13% | 30% | 38% |
| Buick Encore 4dr 4WD | -16% | -1% | -35% | 5% | 3% | 22% |
| Buick Encore GX 4dr | -37% | -46% | | | | |
| Buick Encore GX 4dr 4WD | -38% | -47% | | | | |
| Cadillac XT4 4dr | -3% | -36% | -32% | -15% | -6% | -11% |

Figure 11: Example of data from the IIHS and the HLDI

For example, regarding collision, property damage and comprehensive coverages for the model Audi Q3 4dr 4WD, if the mean of the aggregated claim amount is 2000€ then the mean of the aggregated claim amount for this specific model will be 2280€ (more 14%), 1940€ (less 3%) and 1780€ (less 11%), respectively.

For the last three coverages which only refer to claim frequency, the values of -29%, -31% and -10% for the same model will mean that this model has 29%, 31% and 10% less number of personal injury, medical payment and bodily injury claims, respectively.

The dataset was built, having a total of 498 samples and the variables mentioned before (coefficient for each insurance coverage for each model, brand, model, year of manufacture, level of automation).

Regarding the main variable in study (coefficient for each insurance coverage for each model), an adaption was made by considering the multiplicative coefficient instead of the percentage (1,14 instead of 14% and 0,97 instead of -3%, for instance) for the purpose of not having negative values.

From the main variables (brand, model and type of car), there were created the following additional variables:

<table>
<tr><td rowspan="9">Brand Group</td><td>1 - Fiat</td><td rowspan="9">Car size</td><td rowspan="2">Micro</td></tr>
<tr><td>2 – Cadillac, Chevrolet, Ford</td></tr>
<tr><td>3 – Audi, Porsche, Volkswagen</td><td rowspan="2">Mini</td></tr>
<tr><td>4 – BMW, Mini</td></tr>
<tr><td>5 – Mercedes-Benz, Smart</td><td rowspan="2">Small</td></tr>
<tr><td>6 – Acura, Hyundai, Honda, Kia, Mazda, Mitsubishi, Nissan, Toyota</td></tr>
<tr><td>7 – Land Rover</td><td>Midsize</td></tr>
<tr><td>8 - Volvo</td><td rowspan="2">Large</td></tr>
<tr><td>9 - Tesla</td></tr>
<tr><td rowspan="2">Car type</td><td>Car</td><td rowspan="2">Luxury</td><td>Luxury</td></tr>
<tr><td>SUV</td><td>Not Luxury</td></tr>
</table>

Table 1: New variables created for the dataset

# 6. RESULTS

## 6.1. Data Analysis

### 6.1.1. Categorical Variables

The first step in data analysis was to study the descriptive analysis of values per variable:

- **Car Type**

| Car | SUV |
|---|---|
| 59.4% | 40.6% |

Table 2: Descriptive analysis of values per car type

Cars represent a bigger slice of the sample than SUV, with 59.4% representativity.

- **Car Size**

| Micro | Mini | Small | Midsize | Large |
|---|---|---|---|---|
| 1.6% | 14.1% | 43.0% | 31.7% | 9.6% |

Table 3: Descriptive analysis of values per car size

Small cars represent most of the vehicles, with 43%, and Micro cars are the least common in the dataset with only 1.6%.

- **Brand Group**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 3.0% | 17.1% | 12.0% | 7.6% | 3.8% | 47.4% | 1% | 4% | 4% |

Table 4: Descriptive analysis of values per brand group

Brand group 6 (composed by Acura, Hyundai, Honda, Kia, Mazda, Mitsubishi, Nissan and Toyota) is the most common in the dataset, representing 47.4% of all vehicles.

- **Luxury vs Non-Luxury**

| | **Luxury** | **Non-Luxury** | |
|---|---|---|---|
| | 19.9% | 80.1% | |

Table 5: Descriptive analysis of values - Luxury vs Non-Luxury

Non-Luxury cars are the most representative in the data set, with 80.1% of vehicles fitting in this category.

- **Level of automation**

| | **Level 0** | **Level 1** | **Level 2** | **Level 3** | |
|---|---|---|---|---|---|
| | 4% | 59% | 31% | 6% | |

Table 6: Descriptive analysis of values per level of automation

Levels 1 and 2 are the most common levels of automation in the dataset, with level 1 being the predominant category (59% of vehicles).

### 6.1.2. Quantitative variables

Some initial analysis of mean, standard deviation, maximum and minimum were performed as shown in the tables below:

| ***Mean [Standard Deviation]*** | *Collision* | *Property Damage* | *Comprehensive* | *Personal Injury* | *Medical Payment* | *Bodily Injury* |
|---|---|---|---|---|---|---|
| *Level 0* | 0.977 [0.004] | 0.825 [0.004] | 0.861 [0.004] | 0.961 [0.005] | 0.882 [0.005] | 0.854 [0.004] |
| *Level 1* | 0.999 [0.202] | 0.929 [0.238] | 0.879 [0.199] | 1.046 [0.398] | 1.045 [0.471] | 0.923 [0.372] |
| *Level 2* | 1.045 [0.321] | 0.769 [0.202] | 1.089 [0.351] | 0.697 [0.265] | 0.635 [0.297] | 0.633 [0.234] |
| *Level 3* | 1.888 [0.831] | 0.858 [0.245] | 1.422 [0.414] | 0.567 [0.156] | 0.497 [0.175] | 0.612 [0.202] |

Table 7: Analysis of mean and standard deviation

Regarding mean, lower levels of automation have better results in Collision and Comprehensive coverages, and higher levels of automation have less aggregated claim amounts on the remaining four coverages.

On the coverages Personal Injury, Medical Payment and Bodily Injury, vehicles with level 1 automation have a much higher standard deviation. This can be explained by the wider range of values as shown by the difference between the minimum and maximum n table 7. On the other three coverages, standard deviation tends to be higher for vehicles with level 3 automation, probably due to this level of automation having less data.

| *Min/Max* | *Collision* | *Property Damage* | *Comprehensive* | *Personal Injury* | *Medical Payment* | *Bodily Injury* |
|---|---|---|---|---|---|---|
| *Level 0* | 0.72<br>1.22 | 0.49<br>1.33 | 0.6<br>1.29 | 0.72<br>1.85 | 0.73<br>2.14 | 0.75<br>1.74 |
| *Level 1* | 0.42<br>1.52 | 0.31<br>1.71 | 0.30<br>1.69 | 0.30<br>2.44 | 0.57<br>2.81 | 0.49<br>2.37 |
| *Level 2* | 0.49<br>2.32 | 0.33<br>1.47 | 0.39<br>2.05 | 0.38<br>1.75 | 0.39<br>1.91 | 0.36<br>1.76 |
| *Level 3* | 0.73<br>3.36 | 0.33<br>1.26 | 0.20<br>2.04 | 0.43<br>1.13 | 0.48<br>1.32 | 0.44<br>1.40 |

Table 8: Analysis of minimum and maximum

Considering that negative results correspond to a lower aggregated claim amount, we can observe initially that lower levels of automation (levels 0 and 1) tend to have better results on Collision and Comprehensive coverages, and higher levels of automation (levels 2 and 3) have better results on Property Damage, Personal Injury, Medical Payment and Bodily Injury.

It was also analyzed the distribution of values for the insurance coverages collision, property damage and comprehensive.

In the following figures, the vertical axis represents the number of samples registered and the horizontal axis represent the coefficient intervals in which the samples were counted. The dotted line is the regression line used to visually interpret the distribution of values in the dataset.

The purpose of this analysis is to have an idea of the distributions to be tested in the next section.

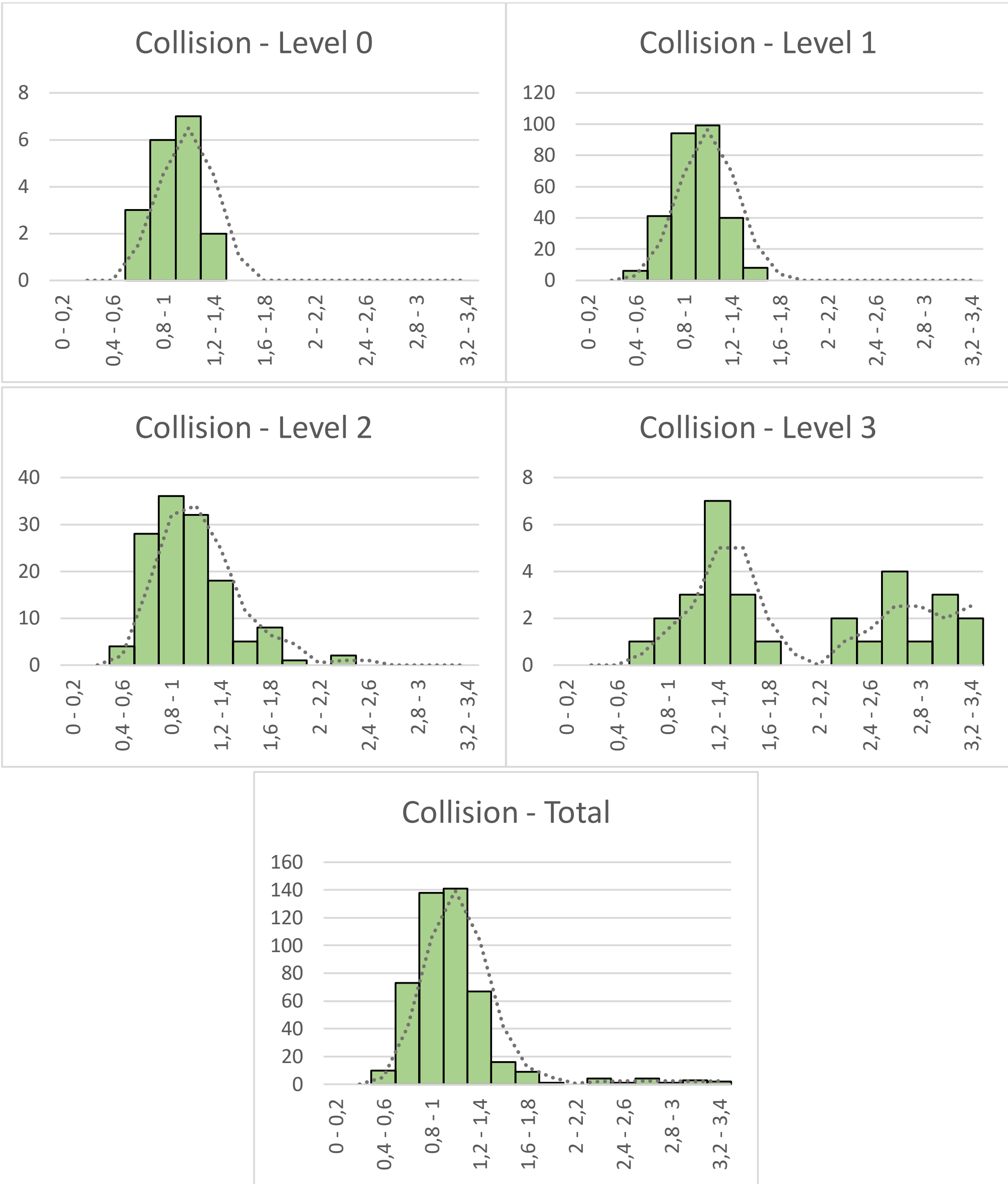



Figure 12: Distribution of values - Collision

The distribution of level 3 vehicles is irregular meaning a higher dispersion of results and matching the results obtained on the standard deviation analysis above.

For levels 0 and 1, the highest number of samples is registered between 1 and 1.2, meaning that vehicles have between 0% and 20% more aggregated claim amounts than average. For level of automation 2, the predominance is between 0.8 and 1 and for level 3, the interval 1.2 to 1.4 is the most common.

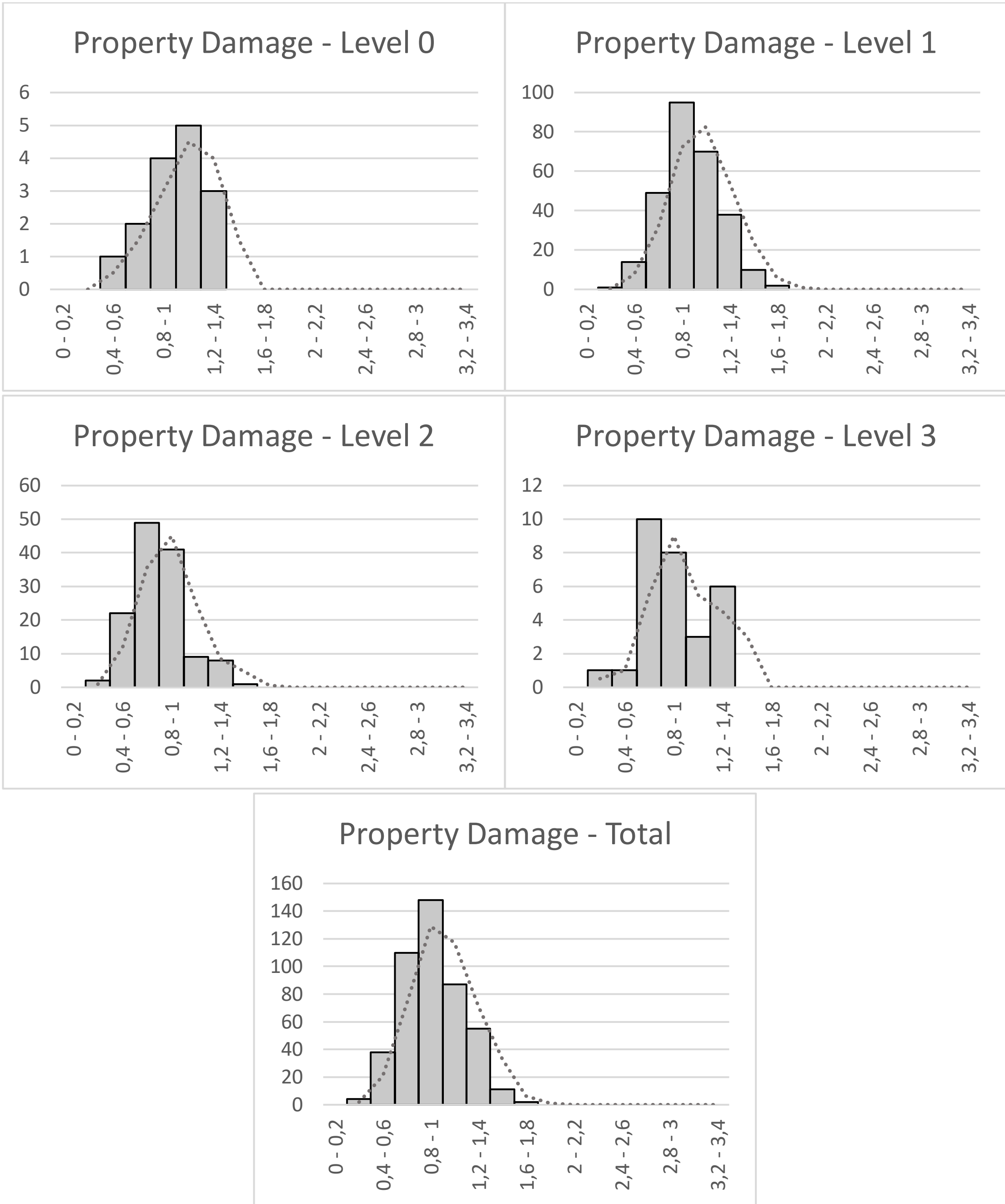



Figure 13: Distribution of values - Property Damage

The dispersion of results for levels 2 and 3 is much lower comparing to the variable collision which also corroborates the values of standard deviation calculated.

For levels of automation 2 and 3, the highest number of samples is registered between 0.6 and 0.8, meaning that vehicles have between 20% and 40% less aggregated claim amounts than average. For level of automation 1, the most common values are between 0.8 and 1 and for level 0, the interval 1 to 1.2 is predominant one.

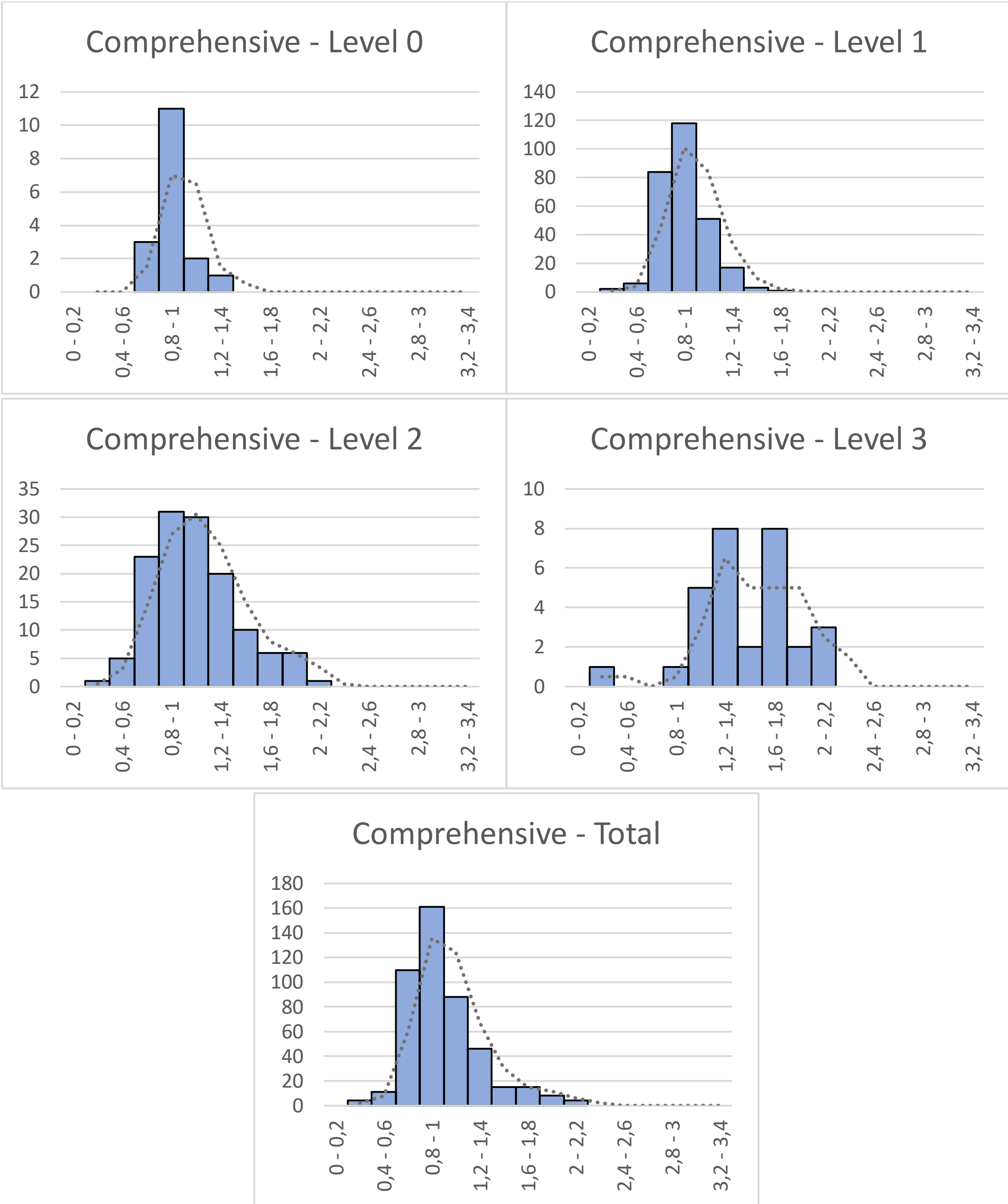



Figure 14: Distribution of values - Comprehensive

As observed on the collision coverage, the distribution of the level 3 vehicles is also more disperse, matching the results obtained on the standard deviation analysis above.

For levels of automation 0, 1 and 2, the highest number of samples is registered between 0.8 and 1, meaning that vehicles have between 20% and 0% less aggregated claim amounts than average. For level of automation 3, there are two predominant intervals: 1.2 to 1.4 and 1.6 to 1.8.

## 6.2. Distribution Fitting

### 6.2.1. Correlation analysis

The correlation between brand group and level of automation was analyzed to decide if both variables should be used in the study. The following tables represents the descriptive analysis of samples per level of automation and brand group.

| Brand Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Level 0 | 2 | 4 | 2 | 0 | 1 | 9 | 0 | 0 | 0 |
| Level 1 | 11 | 61 | 41 | 16 | 2 | 160 | 0 | 0 | 0 |
| Level 2 | 2 | 13 | 17 | 20 | 16 | 63 | 5 | 18 | 0 |
| Level 3 | 0 | 4 | 0 | 2 | 0 | 4 | 0 | 2 | 20 |

Table 9: Descriptive analysis per level of automation and brand group

It can be observed that the data is not evenly distributed. The correlation coefficient was calculated and the value 0.37 was obtained. As this is not a high correlation, both variables will be used in the study.

### 6.2.2. Collision

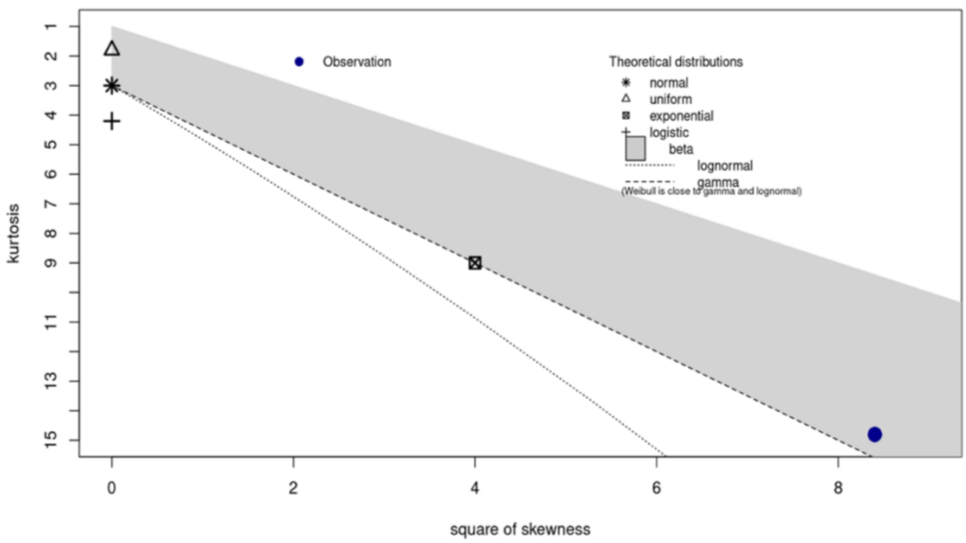



Figure 15: Collision distribution fitting - Cullen and Frey graph

For a Gamma (9.82 , 0.11) was obtained a p-value of $2.2\times10^{-16}$ and therefore the null hypothesis "the variable Collision ~ Gamma (9.82 , 0.11)" should be rejected.

The same approach was applied, but splitting the Collision per level of automation:

| | Distribution | p-value | Expected value | Variance |
|---|---|---|---|---|
| **Level 0** | Gamma (38.474 , 0.025) | 0.09826 | 0.962 | 0.024 |
| **Level 1** | IG (1.004 , 21.986) | 0.09826 | 1.004 | 0.046 |
| **Level 2** | IG (1.045 , 11.430) | 0.8683 | 1.045 | 0.100 |
| **Level 3** | Gamma (5.050 , 0.383) | 0.8302 | 1.934 | 0.741 |

Table 10: Collision distribution fitting per level of automation

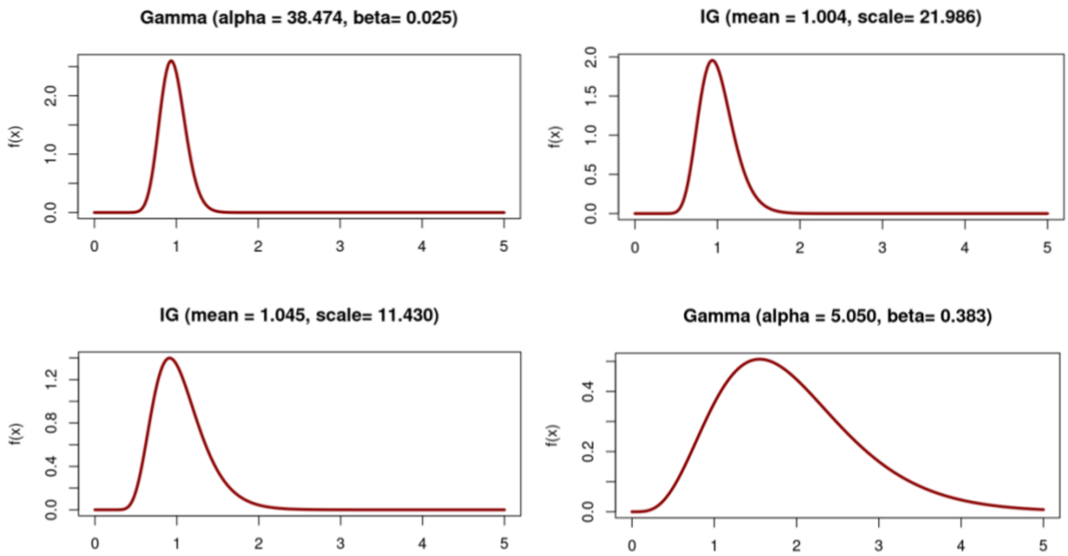



Figure 16: Adjusted distributions for Collision per level of automation

It can be observed that cars with higher level of automation tend to have, in average, higher aggregated claim amounts.

To further analyze the impact of other variables in Collision data, a decision tree was made to identify which variables are the most significant according to each partition. The variables considered were brand group, car type, car size, luxury and level of automation.

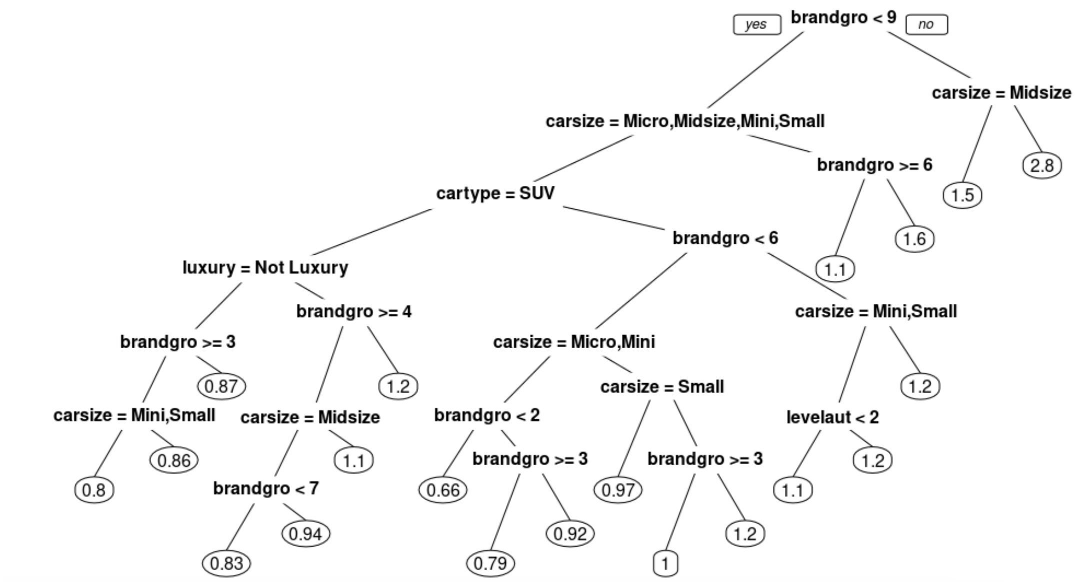



Figure 17: Decision tree for variable Collision

The most significant variable is the brand group, in which the partition is made between cars from brand group 9 (Tesla) and others.

The worst result is 2.8, which means that Midsize Teslas have 180% more aggregated claim amounts than the average.

The best result is 0.66 for a Micro or Mini SUV from the brand groups 1 or 2 (Fiat, Chevrolet, Cadillac or Ford), meaning these vehicles have 34% less aggregated claim amount than average.

For these two specific branches, the variable level of automation does not have an impact. However, it can be observed that this variable has a ‘negative’ impact in 2 out of 20 branches, meaning that cars with higher level of automation tend to have more aggregated claim amounts than the average. Although in this case level of automation has a negative impact on the results, it is n line with the mean analysis previously made.

Repeating the same approach but ignoring the variable brand group, the following result is obtained:

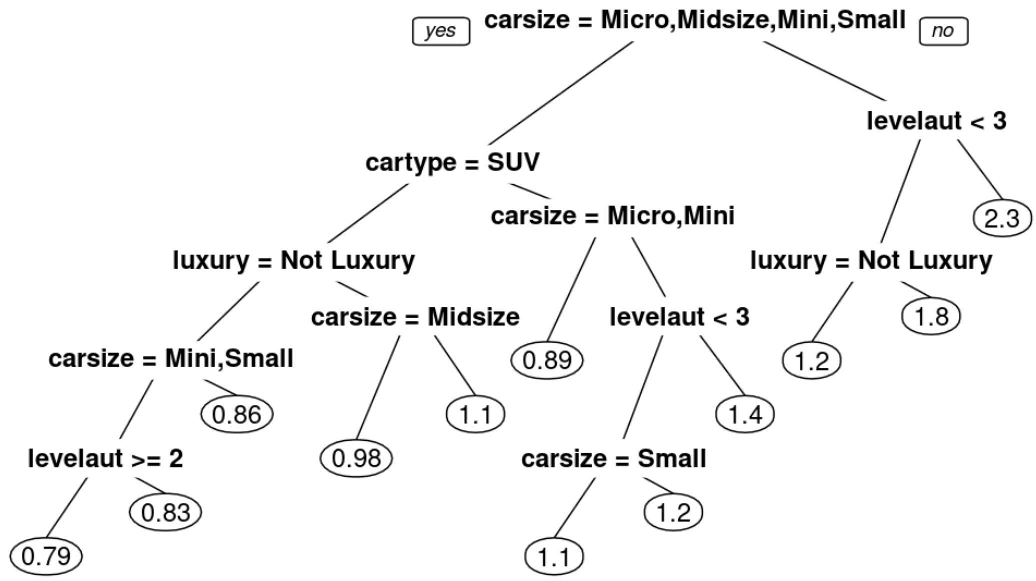



Figure 18: Decision tree for variable Collision (excluding variable brand group)

In this case, the level of automation has an impact in 8 of 12 branches. In 6 of those branches, the variable has a negative impact but on the two left branches (Mini or Small not luxury SUVs) the variable has a ‘positive’ impact, meaning that cars with level of automation 2 or 3 have 21% less aggregated claim amounts than average.

### 6.2.3. Property Damage

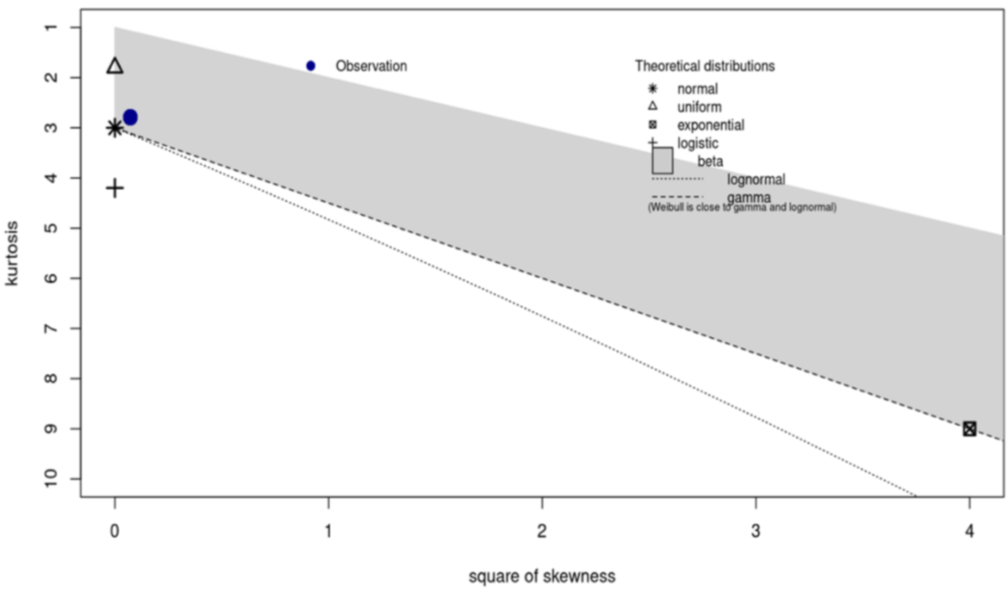



Figure 19: Property Damage distribution fitting - Cullen and Frey graph

To further analyze the variable, the test functions for the Gamma and Inverse Gaussian distributions were used.

It was obtained a p-value of 0.3476 using a transformation to gamma variables for a IG (0.92 , 11.11). For a significance level of 0.05, we shouldn´t reject the null hypothesis “the variable Property Damage ~ IG (0.92 , 11.11)”.

Under the assumption that Property Damage ~ IG (0.92 , 11.11), a GLM was used to study the impact of the other variables in the data and the results were:

| | **Estimate** | **Standard Error** | **Pr(>\|t\|)** |
|---|---|---|---|
| **(Intercept)** | -0.03272 | 0.11056 | 0.76401 |
| **brandgroup2** | 0.38497 | 0.06964 | $5.58 \times e^{-8}$ |
| **brandgroup3** | 0.16752 | 0.07311 | 0.022409 |
| **brandgroup4** | -0.06533 | 0.07861 | 0.406390 |
| **brandgroup5** | 0.09592 | 0.08910 | 0.282304 |
| **brandgroup6** | 0.26911 | 0.06535 | $4.57 \times e^{-5}$ |
| **brandgroup7** | 0.67858 | 0.14167 | $2.29 \times e^{-6}$ |
| **brandgroup8** | 0.02427 | 0.08856 | 0.784227 |
| **brandgroup9** | 0.18783 | 0.10924 | 0.086258 |
| **carsizeMicro** | -0.08970 | 0.10775 | 0.405606 |
| **carsizeMidsize** | -0.19229 | 0.04376 | $1.40 \times e^{-5}$ |
| **carsizeMini** | -0.08792 | 0.02506 | 0.124290 |
| **carsizeSmall** | -0.16012 | 0.04356 | 0.000267 |
| **cartypeSUV** | -0.11955 | 0.02506 | $2.51 \times e^{-6}$ |
| **luxuryNotLuxury** | -0.10469 | 0.04792 | 0.029453 |
| **levelauto1** | 0.02824 | 0.06342 | 0.656335 |
| **levelauto2** | -0.14786 | 0.06730 | 0.028540 |
| **levelauto3** | -0.14377 | 0.10105 | 0.155507 |

Table 11: Property Damage GLM results

From the GLM results, it can be concluded that for a significance level of 0.05, the null hypothesis “Variable has a different distribution than the base variable” shoudn´t be rejected for the following (considering base variables as brandgroup1, carsizeLarge, cartypecar, luxury and levelauto0):

- Brand group 2
- Brand group 3
- Brand group 6
- Brand group 7
- Car size equal to midsize
- Car size equal to small
- Car type equal to SUV
- Not Luxury
- Level auto 2

The same approach was applied, but splitting the Property Damage per level of automation:

| | Distribution | p-value | Expected value | Variance |
|---|---|---|---|---|
| **Level 0** | IG (0.990 , 14.361) | 0.5897 | 0.990 | 0.068 |
| **Level 1** | IG (0.972 , 14.040) | 0.2363 | 0.972 | 0.065 |
| **Level 2** | IG (0.792 , 10.726) | 0.5992 | 0.792 | 0.046 |
| **Level 3** | Gamma (10.989 , 0.082) | 0.1538 | 0.901 | 0.074 |

Table 12: Property Damage distribution fitting per level of automation

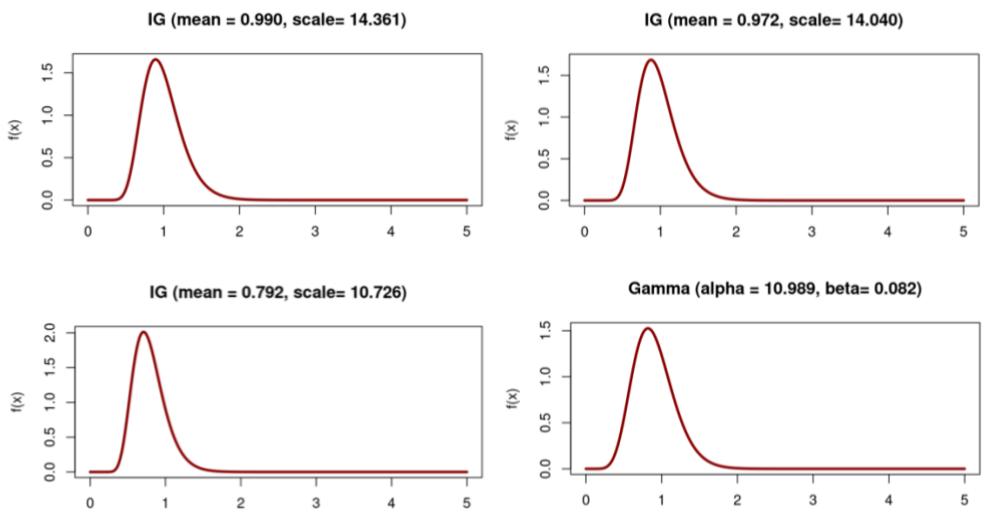



Figure 20: Adjusted distributions for Property Damage per level of automation

It can be observed that cars with higher level of automation tend to have less aggregated claim amounts, with cars with level of automation 2 and 3 having 20.8% and 9.9% less aggregated claim amounts than average, respectively.

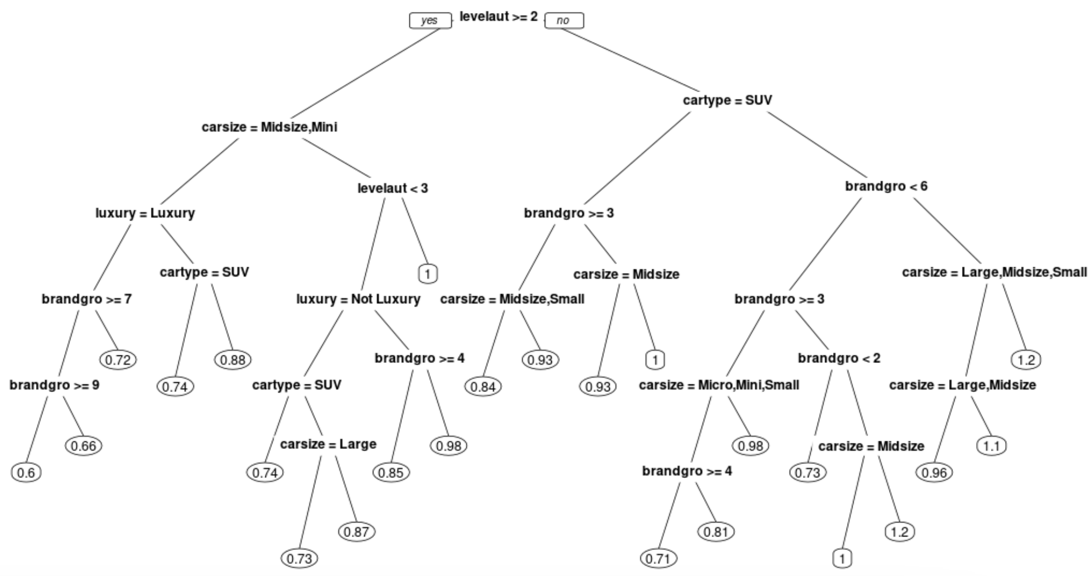



Figure 21: Decision Tree for variable Property Damage

Although the GLM results indicated that level of automation didn´t have an impact in Property Damage data, the Decision Tree results show that this is the variable with most impact in the test data.

The best result was obtained for Mini/Midsize Luxury Teslas with level of automation 2 or 3. These cars have 40% less Property Damage aggregated claims amounts than average.

### 6.2.4. Comprehensive

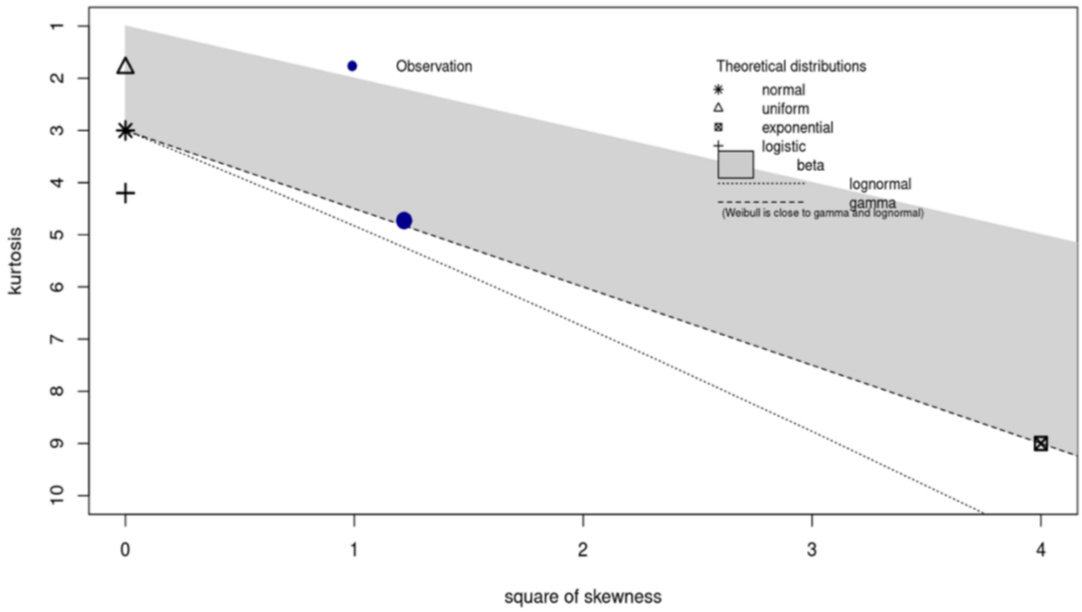



Figure 22: Comprehensive distribution fitting - Cullen and Frey graph

For a Gamma (11.40, 0.09) was obtained a p-value of 0.003413 and therefore the null hypothesis “the variable Comprehensive ~ Gamma (11.40, 0.09)” should be rejected.

The same approach was applied, but splitting the Comprehensive per level of automation:

| | Distribution | p-value | Expected value | Variance |
|---|---|---|---|---|
| **Level 0** | Gamma (31.744 , 0.029) | 0.8828 | 0.921 | 0.027 |
| **Level 1** | Gamma (20.538 , 0.044) | 0.5725 | 0.904 | 0.040 |
| **Level 2** | Gamma (9.641 , 0.113) | 0.9618 | 1.089 | 0.123 |
| **Level 2 + 3** | Gamma (8.569 , 0.134) | 0.2275 | 1.148 | 0.154 |

Table 13: Comprehensive distribution fitting per level of automation

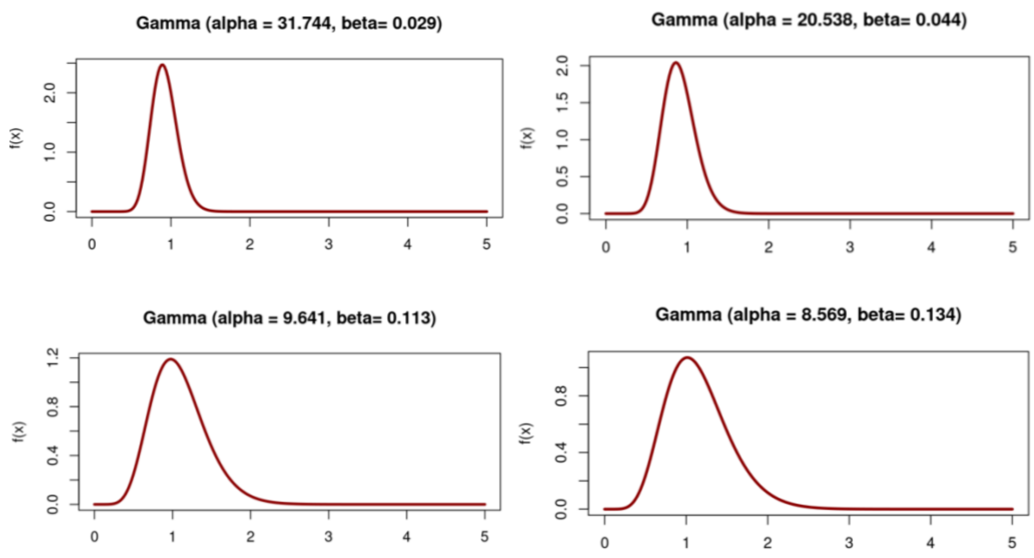



Figure 23: Adjusted distributions for Comprehensive per level of automation

It was not possible to fit a distribution to the variable Comprehensive for cars with level of automation 3. The reason for this may be because of the few test samples available for this level of automation. For this reason, an additional analysis was made to a new variable combining comprehensive results for cars with level of automation 2 and 3.

Using the same approach as in the variable Collision, a decision tree was made to identify which variables have the most impact in the data according to each partition.

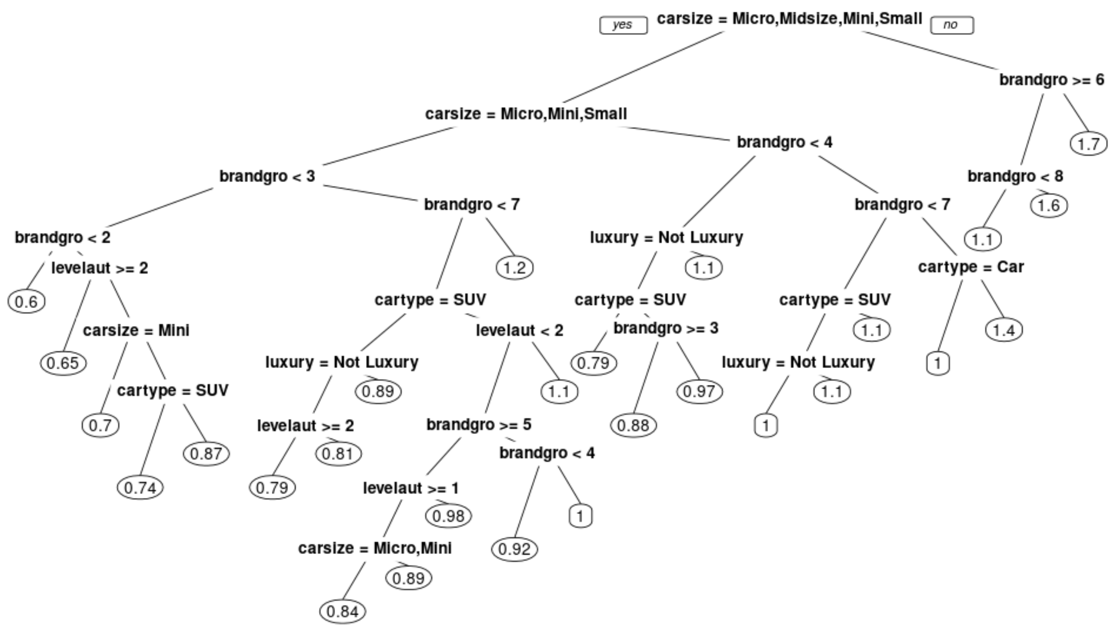



Figure 24: Decision tree for variable Comprehensive

The most significant variable is the carsize, in which the partition is made between Micro/Midsize/Mini/Small and others (Large).

The worst result is 1.7, which means that Micro/Midsize/Mini/Small cars from the brand groups 6, 7, 8 or (Toyota, Hyundai, Kia, Acura, Honda, Mazda, Mitsubishi, Nissan, Land Rover, Volvo or Tesla) have 70% more aggregated claim amounts than the average.

The best result is 0.6 for a Micro/Mini/Small car from the brand groups 1 or 2 (Fiat, Chevrolet, Cadillac or Ford), meaning these cars have 40% less aggregated claim amounts than average.

For these two specific branches, the variable level of automation does not have an impact. However, it can be observed that this variable has an impact in 9 out of 20 branches.

### 6.2.5. Personal Injury

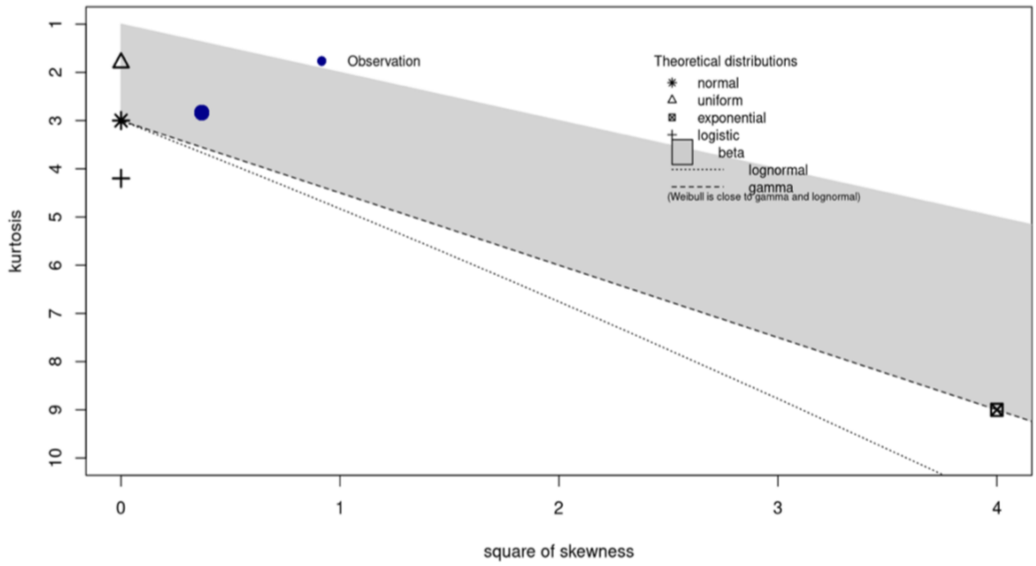



Figure 25: Personal Injury distribution fitting - Cullen and Frey graph

To further analyze the variable, the test functions for the Gamma and Inverse Gaussian distributions were used.

It was obtained a p-value of 0.9448 for a Gamma (6.93 , 0.15). For a significance level of 0.05, we shouldn´t reject the null hypothesis “the variable Personal Injury ~ Gamma (6.93 , 0.15)”.

| | Estimate | Standard Error | Pr(>\|t\|) |
|---|---|---|---|
| **(Intercept)** | 1.41777 | 0.14306 | $< 2\times e^{-16}$ |
| **brandgroup2** | -0.24059 | 0.09939 | 0.01594 |
| **brandgroup3** | 0.15024 | 0.10622 | 0.15801 |
| **brandgroup4** | 0.57092 | 0.12811 | $1.09\times e^{-5}$ |
| **brandgroup5** | -0.21432 | 0.13674 | 0.11785 |
| **brandgroup6** | -0.24699 | 0.09812 | 0.01222 |
| **brandgroup7** | -0.36202 | 0.16901 | 0.03280 |
| **brandgroup8** | 0.23910 | 0.15966 | 0.13505 |
| **brandgroup9** | 0.29894 | 0.16346 | 0.06818 |
| **carsizeMicro** | -0.48067 | 0.07531 | $4.91\times e^{-10}$ |
| **carsizeMidsize** | -0.16848 | 0.05743 | 0.00354 |
| **carsizeMini** | -0.31396 | 0.06212 | $6.66\times e^{-7}$ |
| **carsizeSmall** | -0.22151 | 0.05583 | $8.62\times e^{-5}$ |
| **cartypeSUV** | -0.23768 | 0.02494 | $< 2\times e^{-16}$ |
| **luxuryNotLuxury** | -0.21639 | 0.07147 | 0.00263 |
| **levelauto1** | -0.02428 | 0.04857 | 0.61734 |
| **levelauto2** | 0.12712 | 0.05564 | 0.02287 |
| **levelauto3** | 0.14483 | 0.11668 | 0.21525 |

Table 14: Personal Injury GLM results

From the GLM results, it can be concluded that for a significance level of 0.05, the null hypothesis “Variable has a different distribution than the base variable” shoudn´t be rejected for the following (considering base variables as brandgroup1, carsizeLarge, cartypecar, luxury and levelauto0):

- Brand group 2
- Brand group 4
- Brand group 6

- Brand group 7
- Car size equal to micro
- Car size equal to midsize
- Car size equal to mini
- Car size equal to small
- Car type equal to SUV
- Not Luxury
- Level auto 2

The same approach was applied, but splitting the Personal Injury per level of automation:

| | Distribution | p-value | Expected value | Variance |
|---|---|---|---|---|
| **Level 0** | IG (1.330 , 17.420) | 0.8773 | 1.330 | 0.135 |
| **Level 1** | IG (1.195 , 9.350) | 0.6597 | 1.195 | 0.183 |
| **Level 2** | IG (0.845 , 7.618) | 0.6957 | 0.845 | 0.079 |
| **Level 3** | IG (0.630 , 12.095) | 0.8252 | 0.630 | 0.021 |

Table 15: Personal Injury distribution fitting per level of automation

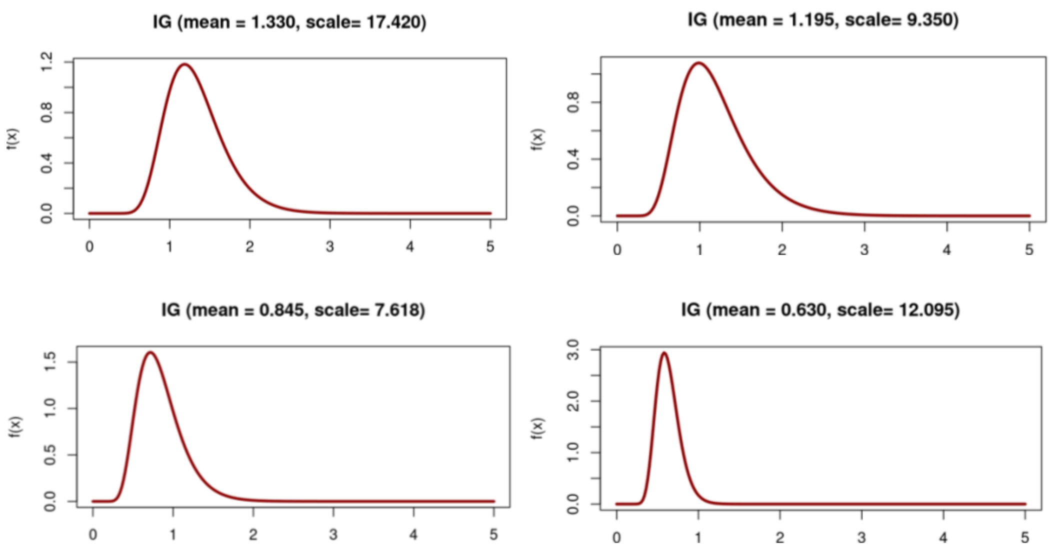



Figure 26: Adjusted distributions for Personal Injury per level of automation

It can be observed that cars with higher level of automation have less Personal injury claims. Cars with level of automation 3 have, on average, 37% less claims and cars with level of automation 0 have 33% more claims than average.

It was also used a Decision Tree to analyze the Personal Injury data and the impact of the other variables.

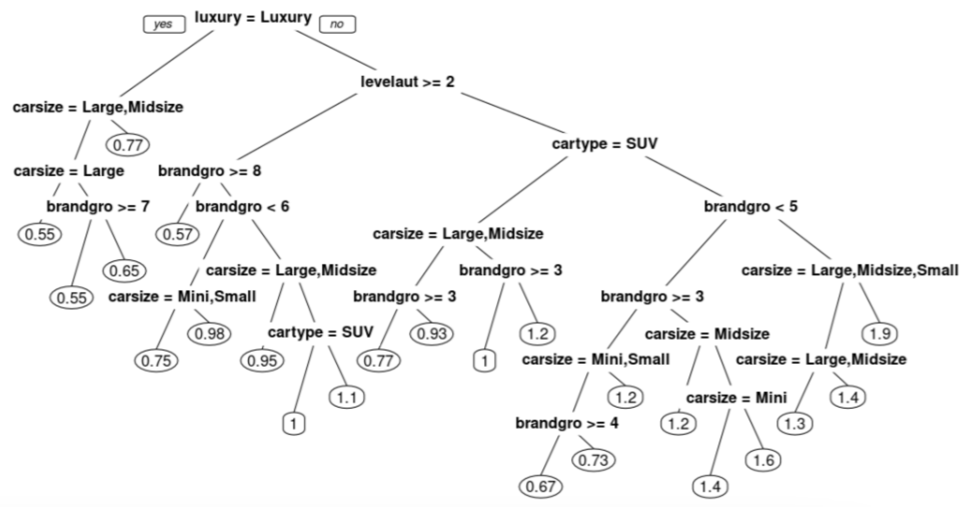



Figure 27: Decision Tree for variable Personal Injury

Luxury is the variable with most impact in the Personal Injury data.

The best results was obtained for Land Rover/Volvo/Tesla Large Luxury cars, with these cars having 45% less Personal Injury claims than average.

Although level of automation is not the most significant variable in the data, it has an impact in 19 out of 23 branches, with the best result obtained for cars with higher level of automation.

### 6.2.6. Medical Payment

Using the function *descdist()* in R, the following graph was obtained for the Medical Payment coverage:

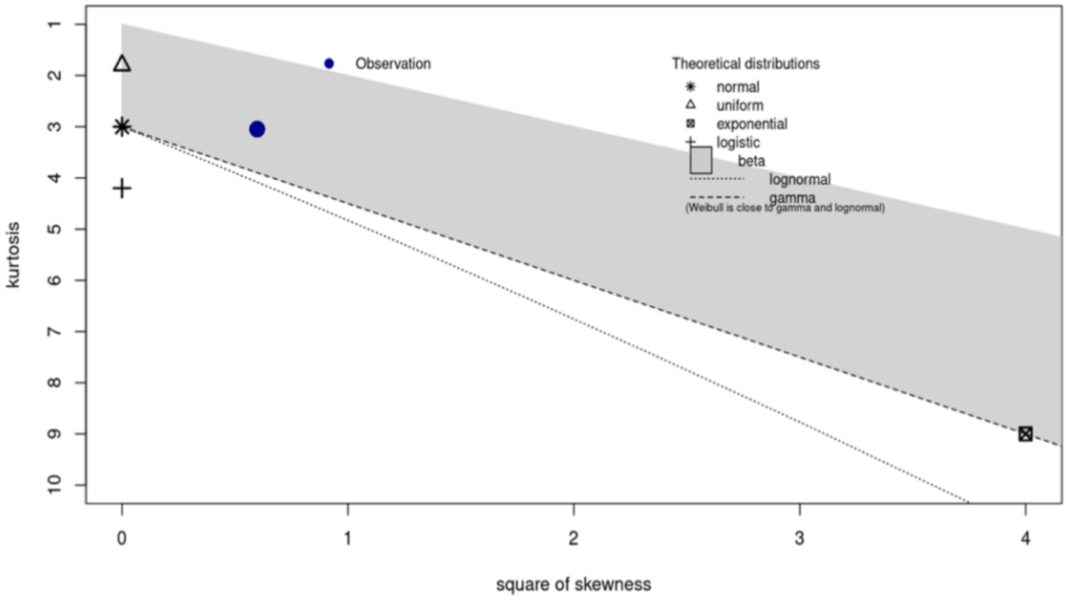



Figure 28: Medical Payment distribution fitting - Cullen and Frey graph

To further analyze the variable, the test functions for the Gamma and Inverse Gaussian distributions were used.

When using the functions *gamma_fit* and *gamma_test*, it was obtained a p-value of 0.3126 for a Gamma (6.03 , 0.19). For a significance level of 0.05, we shouldn´t reject the null hypothesis “the variable Medical Payment ~ Gamma (6.03 , 0.19)”.

| | Estimate | Standard Error | Pr(>\|t\|) |
|---|---|---|---|
| **(Intercept)** | 1.63320 | 0.17970 | $< 2\times e^{-16}$ |
| **brandgroup2** | -0.58812 | 0.13142 | $1.03\times e^{-5}$ |
| **brandgroup3** | -0.20017 | 0.13862 | 0.14961 |
| **brandgroup4** | 0.27956 | 0.17755 | 0.11625 |
| **brandgroup5** | -0.75427 | 0.16157 | $4.30\times e^{-6}$ |
| **brandgroup6** | -0.56520 | 0.13070 | $1.99\times e^{-5}$ |
| **brandgroup7** | -0.58967 | 0.20010 | 0.00342 |
| **brandgroup8** | 0.13860 | 0.22176 | 0.53236 |
| **brandgroup9** | -0.07756 | 0.20216 | 0.70148 |
| **carsizeMicro** | -0.51939 | 0.08419 | $1.87\times e^{-9}$ |
| **carsizeMidsize** | -0.17426 | 0.06889 | 0.01185 |
| **carsizeMini** | -0.37123 | 0.07314 | $6.24\times e^{-7}$ |
| **carsizeSmall** | -0.26830 | 0.06699 | $7.56\times e^{-5}$ |
| **cartypeSUV** | 0.25927 | 0.02746 | $< 2\times e^{-16}$ |
| **luxuryNotLuxury** | -0.12397 | 0.07920 | 0.11842 |
| **levelauto1** | -0.01730 | 0.05302 | 0.74436 |
| **levelauto2** | 0.16743 | 0.06141 | 0.00672 |
| **levelauto3** | 0.21007 | 0.13645 | 0.12457 |

Table 16: Medical Payment GLM results

From the GLM results, it can be concluded that for a significance level of 0.05, the null hypothesis "Variable has a different distribution than the base variable" shoudn´t be rejected for the following (considering base variables as brandgroup1, carsizeLarge, cartypecar, luxury and levelauto0):

- Brand group 2
- Brand group 5
- Brand group 6

- Brand group 7
- Car size equal to micro
- Car size equal to midsize
- Car size equal to mini
- Car size equal to small
- Car type equal to SUV
- Level auto 2

The same approach was applied, but splitting the Medical Payment per level of automation:

| | Distribution | p-value | Expected value | Variance |
|---|---|---|---|---|
| **Level 0** | IG (1.444 , 11.673) | 0.6728 | 1.444 | 0.258 |
| **Level 1** | Gamma (7.573 , 0.170) | 0.6484 | 1.287 | 0.219 |
| **Level 2** | Gamma (8.587 , 0.104) | 0.805 | 0.893 | 0.093 |
| **Level 3** | IG (0.641 , 12.235) | 0.1591 | 0.641 | 0.022 |

Table 17: Medical Payment distribution fitting per level of automation

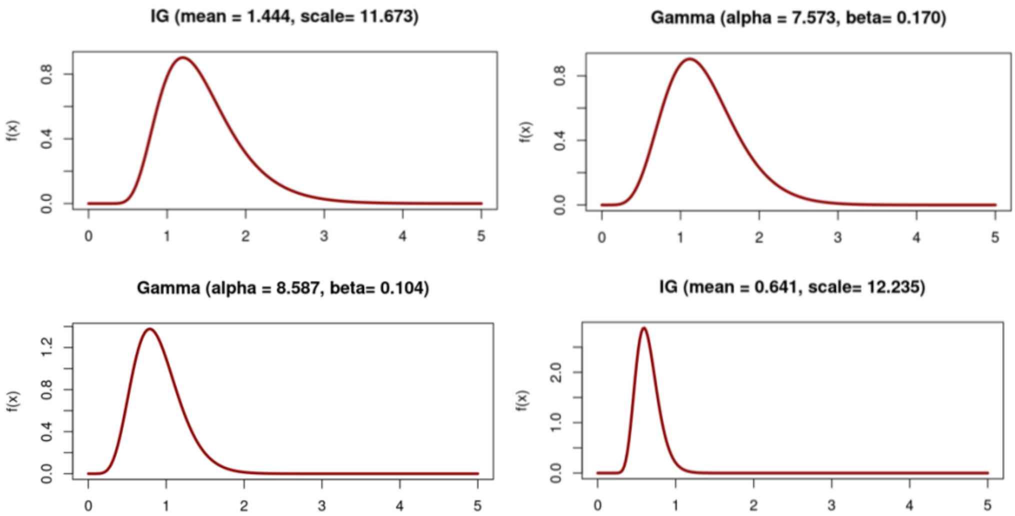



Figure 29: Adjusted distributions for Medical Payment per level of automation

It can be observed that cars with higher level of automation have less Personal injury claims, with cars with level of automation 3 having 35.9% less claims than average, whereas cars with level of automation 0 have 44.4% more claims than average.

Using decision tree methodology, the following results were obtained:

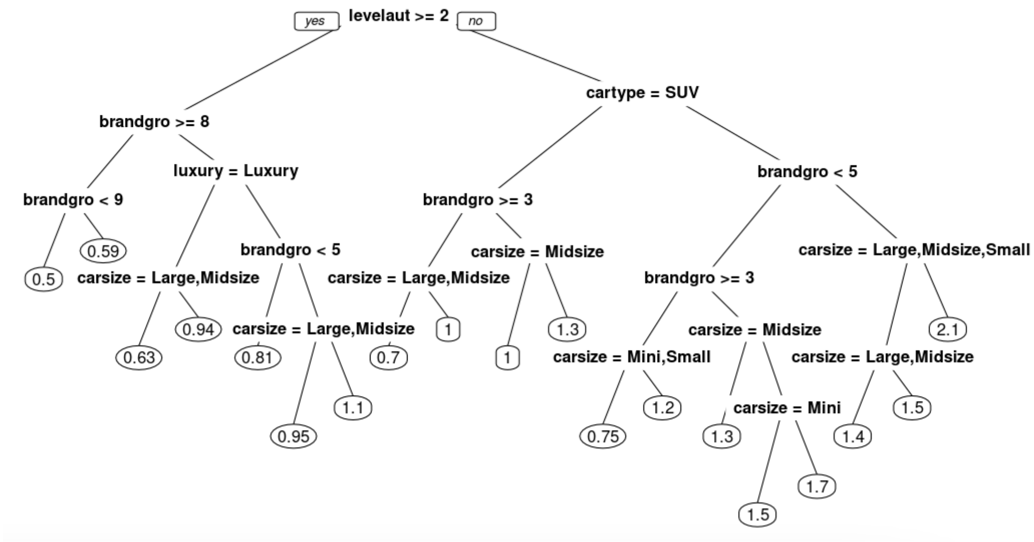



Figure 30: Decision Tree for variable Medical Payment

Based on the Decision tree produced for this variable, level of automation is the most significant variable in the Medical Payment data, dividing car with levels of automation 2 or 3 and others.

The best result was obtained for Volvo cars with level of automation 2 or 3, i. e., these cars have 50% less Medical Payment claims than average.

In this insurance coverage, the variable level of automation has a positive impact in the test data, with cars with higher level of automation having less claims than average.

### 6.2.7. Bodily Injury

Using the function *descdist()* in R, the following graph was obtained for the Bodily Injury coverage:

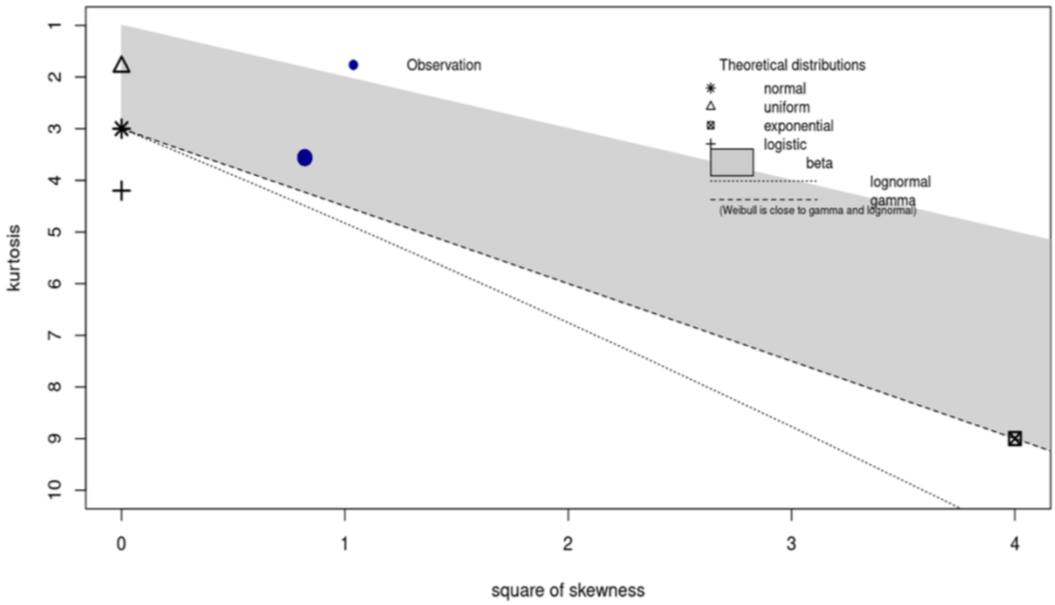



Figure 31: Bodily Injury distribution fitting - Cullen and Frey graph

To further analyze the variable, the test functions for the Gamma and Inverse Gaussian distributions were used.

When using the functions *ig_fit* and *ig_test*, it was obtained a p-value of 0.4255 using a transformation to normality for a IG (1.04 , 8.35). For a significance level of 0.05, we shouldn´t reject the null hypothesis “the variable Bodily Injury ~ IG (1.04 , 8.35)”.

| | Estimate | Standard Error | Pr(>\|t\|) |
|---|---|---|---|
| **(Intercept)** | -0.171808 | 0.144957 | 0.23671 |
| **brandgroup2** | 0.567685 | 0.100720 | $3.51 \times e^{-8}$ |
| **brandgroup3** | 0.232916 | 0.103862 | 0.02553 |
| **brandgroup4** | -0.125685 | 0.112710 | 0.26554 |
| **brandgroup5** | 0.344039 | 0.120882 | 0.00468 |
| **brandgroup6** | 0.415403 | 0.097154 | $2.44 \times e^{-5}$ |
| **brandgroup7** | 0.662889 | 0.154369 | $2.26 \times e^{-5}$ |
| **brandgroup8** | -0.043924 | 0.120155 | 0.71490 |
| **brandgroup9** | 0.231582 | 0.141072 | 0.10155 |
| **carsizeMicro** | 0.447522 | 0.173582 | 0.01033 |
| **carsizeMidsize** | -0.130987 | 0.049805 | 0.00890 |
| **carsizeMini** | 0.131795 | 0.069834 | 0.05993 |
| **carsizeSmall** | -0.014886 | 0.050104 | 0.76656 |
| **cartypeSUV** | -0.239587 | 0.028788 | $1.81 \times e^{-15}$ |
| **luxuryNotLuxury** | -0.038423 | 0.053036 | 0.46925 |
| **levelauto1** | -0.003529 | 0.080671 | 0.96513 |
| **levelauto2** | -0.184016 | 0.084562 | 0.03020 |
| **levelauto3** | 0.193789 | 0.124448 | 0.12030 |

Table 18: Bodily Injury GLM results

From the GLM results, it can be concluded that for a significance level of 0.05, the null hypothesis “Variable has a different distribution than the base variable” shoudn´t be rejected for the following (considering base variables as brandgroup1, carsizeLarge, cartypecar, luxury and levelauto0):

- Brand group 2
- Brand group 3
- Brand group 5

- Brand group 6
- Brand group 7
- Car size equal to micro
- Car size equal to midsize
- Car type equal to SUV
- Level auto 2

The same approach was applied, but splitting the Bodily Injury per level of automation:

| | **Distribution** | **p-value** | **Expected value** | **Variance** |
|---|---|---|---|---|
| **Level 0** | IG (1.189 , 17.964) | 0.9443 | 1.189 | 0.094 |
| **Level 1** | Gamma (9.804 , 0.116) | 0.1393 | 1.137 | 0.132 |
| **Level 2** | Gamma (11.819 , 0.070) | 0.262 | 0.827 | 0.058 |
| **Level 3** | IG (0.793 , 12.201) | 0.9981 | 0.793 | 0.041 |

Table 19: Bodily Injury distribution fitting per level of automation

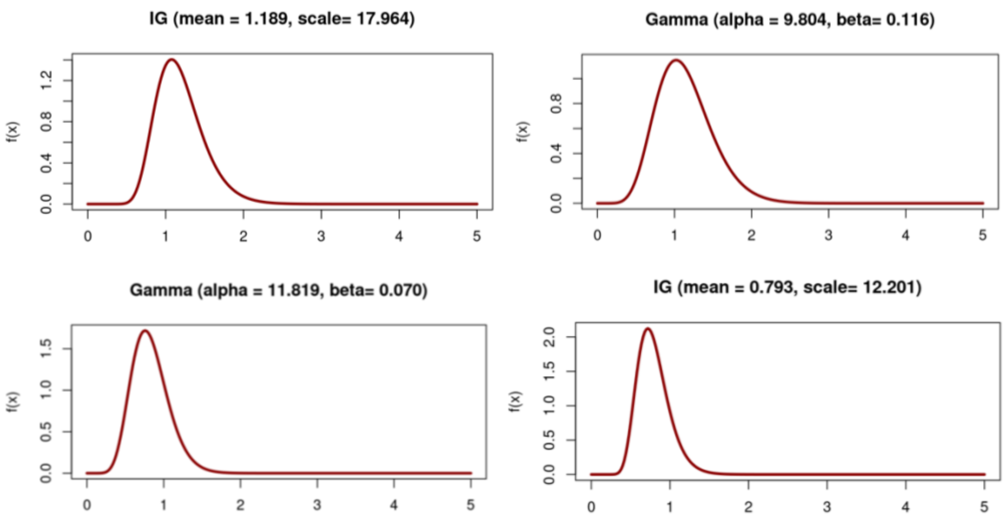



Figure 32: Adjusted distributions for Bodily Injury per level of automation

It can be observed that cars with higher level of automation have less Personal injury claims, with cars with level of automation 3 having 20.7% less claims than average and car with level of automation 0 having 18.9% more claims than average.

To further analyze the impact of other variables in Bodily Injury data, a decision tree was made to identify which variables are the most significant according to each partition.

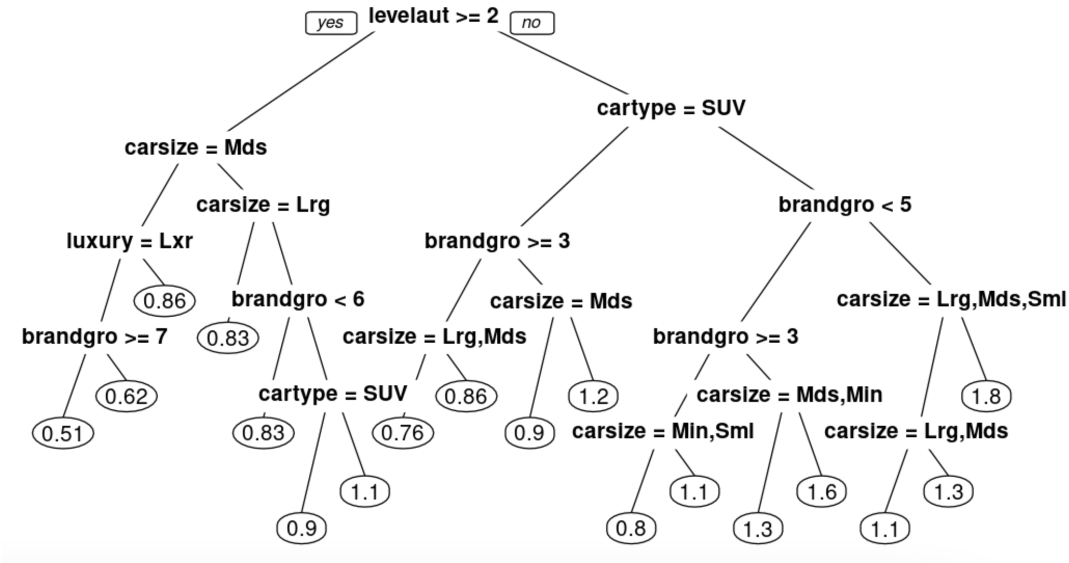



Figure 33: Decision Tree for variable Bodily Injury

It can be observed that the variable level of automation is the most significant variable in the Bodily Injury data. The partition is made between cars with level of automation equal or greater than 2 and others.

Overall, cars with higher level of automation tend to have less Bodily Injury claims than average, with the best result obtained for Land Rover/Volvo/Tesla Midsize Luxury cars levels of automation 2 or 3.

# 7. CONCLUSION

In this study, were used to two main statistical models to analyze the impact of ADAS in insurance premiums: Generalized Linear Models and Decision Trees.

Using the distribution fitting functions in R, it was possible to fit 4 of the insurance coverage to a distribution:

- Property Damage ~ IG (0.92 , 11.11)
- Personal Injury ~ Gamma (6.93 , 0.15)
- Medical Payment ~ Gamma (6.03 , 0.19)
- Bodily Injury ~ IG (1.04 , 8.35)

Decision Trees were used to analyze all coverages and it was observed that the variable level of automation has an impact in all coverages.

It was also used the same distribution fitting method but for each level of automation.

Based on the results obtained using this methodology, the coefficients to be applied to an average insurance premium would be:

| ***Mean*** | *Collision* | *Property Damage* | *Comprehensive* | *Personal Injury* | *Medical Payment* | *Bodily Injury* |
|---|---|---|---|---|---|---|
| *Level 0* | 0.962 | 0.990 | 0.921 | 1.330 | 1.444 | 1.189 |
| *Level 1* | 1.004 | 0.972 | 0.904 | 1.195 | 1.287 | 1.137 |
| *Level 2* | 1.045 | 0.792 | 1.089 | 0.845 | 0.893 | 0.827 |
| *Level 3* | 1.934 | 0.901 | 1.148* | 0.630 | 0.641 | 0.793 |

Table 20: Expected value for each insurance coverage per level of automation

To note that for Comprehensive coverage and level of automation 3, as it was not possible to fit any distribution, the expected value in table 19 refers to test data for level of automation 2 and 3.

It can observed that for coverages Collision and Comprehensive, the aggregated claim amount is higher for vehicles with higher level of automation, which can be easily related to the fact that these vehicles tend to be more expensive and therefore the cost to repair will be higher.

Considering that all coverages would weigh the same in the premium calculation, the proposed coefficients to be applied for each level of automation are:

| | *Overall* |
|---|---|
| *Level 0* | 1.139 |
| *Level 1* | 1.083 |
| *Level 2* | 0.915 |
| *Level 3* | 1.007 |

Table 21: Overall expected value per level of automation

Based on these results, we can conclude that the vehicles that would benefit from the implementation of this variable on the insurance industry would be the vehicles with level of automation 2 with an average reduction of the insurance premium of 8.5%.

It would be expected that level 3 would have an overall expected value less than average, but this is not the case and can be explained by the few data samples in the dataset.

Overall, we can conclude that the level automation should be taken in consideration when calculating insurance premiums, since it has been observed in this study that the level of automation has a significant impact in the claim frequency and aggregated claim amount.

# 8. REFERENCES

Accenture (2017), *Insuring Autonomous Vehicles*;

Bello, F., Gudmundarson, R., Meindersma, J., Oliveira, C. and Xu, Z. (2019), *Autonomous Vehicles Insurance Policy;*

Bonat, W. H. and Jorgensen, B. (2016), *Multivariate covariance generalized linear models, Journal of the Royal Statistical Society: Series C (Applied Statistics)*;

Celent (2012), *A Scenario: The End of Auto Insurance: What Happens When There Are (Almost) No Accidents?*;

Celent (2016), *The End of Auto Insurance: A Scenario or a Prediction?*;

Celent (2019), *Re-rethinking The End of Auto Insurance*;

CNseg (2019), *Revista Jurídica de Seguros*;

Charpentier, A. (2016), *Computational Actuarial Science with R*, CRC Press, Taylor & Francis;

European Parliament (2019), *Car insurance: new EU rules to better protect victims of road accidents*;

European Commission (2018), *On the road to automated mobility: An EU strategy for mobility of the future;*

European Parliament (2019), *Self-driving cars in the EU: from science fiction to reality*;

Hand, C. (2020), *Car insurance must evolve for autonomous vehicles;*

Garrido, J., Genestb, C. and Schulz, J. (2016), *Generalized linear models for dependent frequency and severity of insurance claims;*

Guerreiro, G. (2017), *Manual de Construção de Tarifas com R*, Texto de apoio da unidade curricular Gestão do Risco em Atuariado Não Vida, FCT NOVA;

Gupta, P. (2017), *Decision Trees in Machine Learning*;

Highway Loss Data Institute, Insurance Institute for Highway Safety (2020), *Real-world benefits of crash avoidance technologies*;

Insurance Institute for Highway Safety (IIHS) and Highway Loss Data Institute (2021), Retrieved from https://www.iihs.org/ratings/insurance-losses-by-make-and-model;

Jeffs, J. (2022), *A History of ADAS: Emergence to Essential*;

Khanet, J., Kohli, P. (2020), *ADAS Analysis Creates Path for Auto Insurance Pricing*, LexisNexis;

Kennedy, E., *Quais os desafios das seguradoras na era dos veículos autónomos?*;

Kida, Y. (2019), *Generalized linear models;*

*KPMG Canada (2015), "Marketplace of Change: Automobile Insurance in the Era of Autonomous Vehicles";*

*KPMG UK (2015), "Connected and Autonomous Vehicles – The UK Economic Opportunity";*

Lutkevich, B. (2019), *self-driving car (autonomous car or driverless car)*;

Mageira, M. and Marcacci, S. (2021), *New developments in Swiss Re's ADAS Risk Score and how it impacts insurance claims*;

McBride, N. (2015), *The ethics of driverless cars*;

Nelder, John; Wedderburn, Robert (1972). "*Generalized Linear Models*". Journal of the Royal Statistical Society. Series A (General). Blackwell Publishing. 135 (3): 370–384.

Punev, A. (2020), *Autonomous Vehicles: The Need for a Separate European Legal Framework*;

Swiss Re (2018), *How smart tech could lower car insurance premiums*;

Swiss Re (2018), *Introducing the ADAS risk score*;

Verizon Connect (2022), *What is Telematics*;

2022

Title: Impact of ADAS in Insurance Pricing

Subtitle:

Pedro Manuel Rodrigues Jordão