NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

The impact of ADAS in the insurance world

Manuel Maria Gião Piçarra Ferreira Viegas

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

THE IMPACT OF ADAS EQUIPMENT IN THE INSURANCE WORLD

by

Manuel Maria Gião Piçarra Ferreira Viegas

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: Nuno António

February 2022

ACKNOWLEDGEMENTS

To my family, for all the encouragement.

To all my friends, for all the support and listening to me complain all the time.

To João, for the trust and for guiding me in this work.

To Professor Nuno Antonio, for all his availability and guidance.

And a special thanks to Matilde and Alex.

ABSTRACT

The automobile industry in insurance companies has been increasingly becoming more competitive and innovative, seeking not only to increase its portfolio, but especially to attract customers with fewer claims, thus increasing profitability and reducing risk. Cars have evolved their equipment, making them more “intelligent”, in order to ensure a more comfortable driving and to warn of dangers on the road. One of the reasons vehicles are becoming more intelligent are devices called ADAS (Advanced Driver-Assistance Systems). This project aims to understand the impact that these devices have on the severity and frequency, through predictive models. With this, it will be possible to delineate discounts for insurance policies of vehicles that have this equipment, while simultaneously verifying if these vehicles are less prone to accidents.

KEYWORDS

Insurance; ADAS; Predictive Models; Accidents

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADAS	Advanced Driver-Assistance Systems
APE	Automated Performance Enhancement
AUC	Area under the curve
CA	Collision Avoidance
CC	Cruise Control
CRISP DM	CRoss Industry Standard Process for Data Mining
DVA	Driver Vision Augmentation
LCA	Lateral Control Assistance
PA	Parking Assistance
ROC	Receiver operating characteristic
SEMA	Specialty Equipment Market Association
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
TPC	Tire Pressure Control

1. INTRODUCTION

Technological growth is visible in all sectors, in particular in the automotive industry, which is increasingly investing in technologies and features in its vehicles, namely new safety applications that assist and help the driver. Many of these technologies have been on the market for a long time as optional equipment for high-end vehicles, this is, as features not included in the standard set that are added per customer request. However, they have been becoming increasingly more common across all ranges, in part since consumers are more and more interested in learning about their cars' systems and investing in them (Grimm, 2003). As a result of this higher availability, they lead to a reduction in accidents by helping to find solutions to prevent accidents, injuries, and deaths on the roads. However, this potential will not be fully accomplished until consumers understand these technologies, how to use them appropriately, and avoid misusing or relying too heavily on them (Lu et al., 2005).

For automotive insurance companies, whose business depends on attracting clients with a smaller risk of having accidents, so as to avoid covering their charges when a claim occurs, it is particularly important to study and deeply understand which factors might lead to the occurrence of accidents, but even more so, which can contribute to their prevention. These factors can be related to the clients' own characteristics, for example, their age, number of years as drivers or previous accident history, but also potentially to the vehicles' characteristics, since, as mentioned, some technological features exist precisely to increase customers' safety and also to prevent accidents (Ayuso et al., 2019).

Advanced driver-assistance systems (ADAS) are one of the growing safety application areas due to the desire to reduce traffic accidents (Kala, 2016). These are safety systems designed to assist the driver while driving, using a combination of sensors to warn of potential dangers. Therefore, it is important to acknowledge the fact that ADAS might have a crucial role in reducing the risk in insurance companies' client portfolio, by enabling to identify customers whose vehicles immediately make them less prone to have accidents (Pérez-Marín & Guillen, 2019).

Nonetheless, this is still a relatively new field and, considering that road accidents are highly dependent on external factors that are not directly preventable, such as weather, the own driver's or other drivers' behaviors, it is important to study how important each of these groups of vehicles' features prove to be in accident prevention, in particular in the Portuguese market (Norris et al., 2000).

Therefore, the objective of this project is to use Machine Learning techniques, namely predictive models, to examine the impact of ADAS equipment on two of the most commonly used variables in the insurance pricing field: Frequency and Severity. These variables, together, take into account the number of accidents, total cost and clients' tenure in order to assess their profitability (Guelman, 2012). In case there is enough evidence to support the conclusion that a certain feature contributes to the prevention of accidents, there could be given a discount in order to attract customers who are less likely to have claims and translate into cost for the company. As a result, this study is not only beneficial for the company but also for the customers with these vehicles, as they can benefit from a discount.

1.1. COMPANY OVERVIEW

Ageas Group is considered one of the largest European insurance groups, headquartered in Belgium where it is a leader, and present in fourteen countries in Europe and Asia, with more than 45 thousand employees. In Portugal, there are 1,281 employees and 2,722 brokers serving around 1.7 million customers of various brands. It is composed of five commercial brands: Ageas Seguros, Ageas Pensões, Médis, Ocidental, and Seguro Directo, with Life and Non-Life insurance solutions.

The one-year internship described in this document was conducted in the non-life area, where there are different lines of business such as home insurance, car insurance, accidents at work insurance and personal accidents insurance. Additionally, it consists of the development of a project in the Pricing and Business Analytics area of the Ageas Portugal Group, more focused on Seguro Directo. This brand, despite having a small portfolio in comparison to the group's other brands, has a very strong component in innovation, investing in several projects to improve its services. For instance, it was the first car insurer to sell by telephone in Portugal and it is currently the only direct insurer of the group.

The insurance industry has invested a lot in the Data Science area, since nowadays the amount of data produced and recorded is higher and represents more value for companies. In this sector, it is essential to understand the amount of risk that is collected and to find quick and effective solutions to help in decision making.

Therefore, Data Science contains different benefits in the way prices, claims, fraud, and renewals are managed, as well as in the innovation of new products and the way they are delivered to customers. With an increasingly competitive market for companies, it is essential to embrace these new technologies and analyses in order to gain a competitive advantage, and consequently managing to increase their revenue while reducing their costs. As for Ageas in particular, the search for innovative ways to do so is present in the company's goals, for which it is constantly striving to embrace new projects, as well as to integrate and create new techniques and approaches to reach them.

1.2. PROBLEM DEFINITION

There are more and more factors that may help determine each policy's risk and allow insurance companies to calculate their cost, and while the vehicle's characteristics have long been used to personalize policy premium, it is not common to take specific vehicle equipment into account. In this sense, ADAS were selected to be studied due to the potential they might have in minimizing the likelihood of a claim and its cost.

To fulfill this purpose, an analysis of some of these vehicle features was developed so as to comprehend their behavior in the company's portfolio, through the means of predictive models. By developing models capable of predicting the claim frequency and its cost, based on, among others, the vehicle's characteristics, it is possible to learn how much ADAS contribute to these predictions, and which specific ones prove to be more relevant. Afterwards, this knowledge will be transferred to the responsible teams, who will be able to further study the potential implementation of discounts for clients with certain vehicle features. Overall, the implementation of this project is expected not to attract more clients indiscriminately, but rather to attract more clients with presumably lower risk.

2. LITERATURE REVIEW

This section describes a literature review containing the main references and concepts that serve as a theoretical basis for the project. The purpose of this section is to find the best practices to develop the project, as well as to support all the decisions that will be taken throughout the work to ensure its success.

2.1. INSURANCE AND RISK CLASSIFICATION

The fundamental concept of insurance consists in assuming risk, providing financial protection in the event of damage to customers, and transferring the risk in exchange of regular premiums. Each insured person pays a different premium according to their risk. Normally, the premium is calculated using the conditional expectation of the claim frequency with the expected value of claims, considering the observable risk characteristics (David, 2015).

To develop the ratemaking process, it is necessary for the company to quantify the risks it will assume and the premiums it will charge for assuming them. In this way, estimation of the different components is essential to create the fundamental equation that will determine if the estimated premium is likely to achieve the desired profit within the period that these same ones will be in effect (Werner et al., 2016).

The process of developing fair tariffs includes an important part called the risk classification, which consists of grouping risks into several classes with similar characteristics in order for insurers to reasonably discriminate the prices of their products fairly and equitably (Paefgen et al., 2013).

Furthermore, a fixed premium on the entire portfolio would lead to “good risks” - clients who are at a lower risk of having a claim - leaving the company and accepting a better offer from another company that considers risk classification. Consequently, the company would be left with only “bad risks” with low premiums (Henckaerts et al., 2017).

Therefore, the risk classification within the rate setting process is important so that the insurer can optimally group the risks in the portfolio, so that policyholders with the same risk profile pay the same reasonable premium rate (Antonio & Valdez, 2011).

Additionally, insurers compete for pricing based on detailed risk classification since it is a mechanism that allows them to reduce costs in their insurance policy. There are very competitive markets, and in these markets, insurers must use more risk factors to differentiate the price to other insurers.

The calculation of the premium in car insurance, which is called Insurance Pricing, is intended to be determined correctly and fairly, and it is essential for both the insurance company and the insured. With this purpose, it is typically based on a set of different features that include information about the insured person, namely age, address, engine power, license age and claim history. Moreover, pricing is based on regressions where all data collected from policies is considered and usually parted into frequency and severity models. These two models are decisive, as they allow the car insurer to make predictions about potential future claims and allow it to take the necessary steps in advance to reduce the chances of harm to the company (Verbelen et al., 2018).

2.2. ACCIDENT RATES

Road accidents are unpredictable and represent a public health problem, with high numbers of injuries and deaths every year. Nonetheless, there are some well-used security strategies and mechanisms that can prevent these events.

According to the Annual Road Safety Report, in 2019 there were 35,704 accidents with victims, which resulted in 626 fatalities and more than 45,000 people injured in Portugal, as seen in Figure 1. Throughout the past years in Portugal, the number of accidents with victims has been relatively constant, with a slight gradual increase, while the mortality rate has decreased (Relatório Anual de Segurança Rodoviária, 2019).

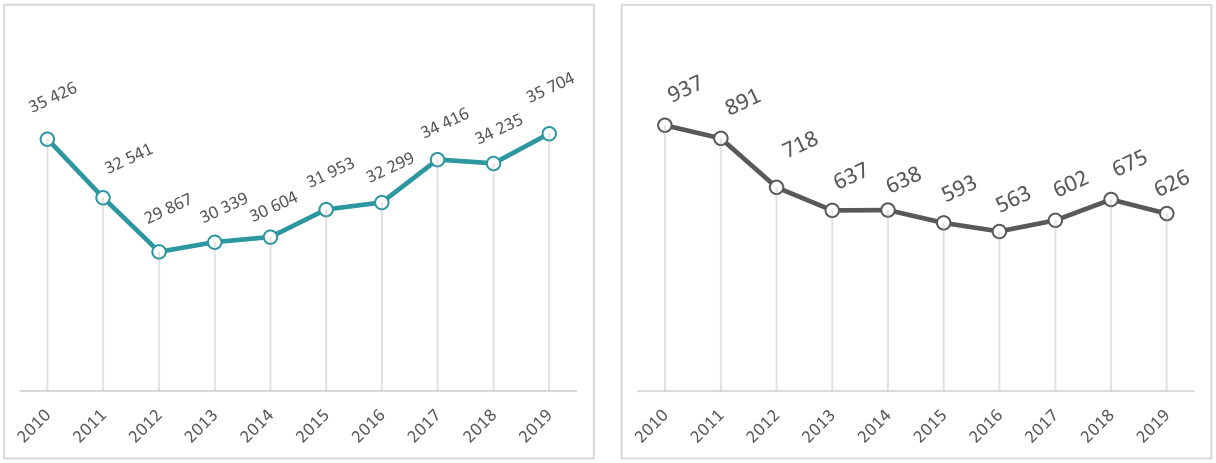

Figure 1 – Evolution of road accidents with injured people (Left) and of deaths (Right), in Portugal.
Adapted from (Relatório Anual de Segurança Rodoviária, 2019)

The most frequent type of accident in 2019 was "side collision with another moving vehicle" with 6,452 occurrences (18%), followed by "simple mistracking" (6.141 occurrences, 17%) and "pedestrian running over" (4,799 occurrences, 13%).

Moreover, car accidents are very complex and can result from several factors, the main ones being human, technical, and environmental factors. The way the streets are designed and the disposition of traffic signs, as well as the weather conditions, namely rain and storms, are some of the most important factors that influence the frequency and severity (Vorko-Jović et al., 2006).

Nonetheless, the primary cause of road accidents are human factors. According to Brookhuis et al. (2010), statistically, about 90% of road accidents are caused by human failures, such as distraction, tiredness and sleepiness at the wheel. The leading cause of fatal accidents (34%) is loss of alert, whereas alcohol accounts for 20% of all "major causal factor" accidents during the weekend, and fatigue as a "single factor" is estimated to be responsible for 7-10% of all accidents (Brookhuis et al., 2001).

2.3. ADVANCED DRIVER ASSISTANCE SYSTEMS - ADAS

Vehicles are currently essential for many people, allowing to travel in a practical and more comfortable way. However, at the same time they also put many people's lives at risk. Therefore, the demand from consumers for new security systems has been increasing.

Advances in information technologies have led to more complex road safety applications. These systems offer multiple possibilities to improve road transport, for instance by helping the driver with their detection capability, alerting in case of error, and reducing the risk of road accidents. These safety systems, better known as ADAS, have become very popular with consumers who are looking to make their car smarter and safer, since they have great advantages in terms of safety and comfort (Gerónimo et al., 2010).

These systems should be able to sense, analyze, predict, and react to the road environment, which are the key features of context-awareness. Tigadi et al. describe ADAS being technologies that provide the driver necessary information, automate difficult and repetitive tasks, and increase vehicle safety (Tigadi et al., 2016). The study of these devices has been analyzed and some studies have demonstrated the positive contribution of these ADAS to traffic accidents. In fact, implementing ADAS can lead to a 40% reduction in deaths (Golias et al., 2002).

These technologies have been on the market for a long time, as optional equipment for high-end vehicles, but these features are becoming increasingly more common across all ranges. According to a study by SEMA (2020), ADAS can be divided into different categories and applications in order to group similar equipment (John Waraniak et al., 2020).

There are different types of sensors that can be combined to create a variety of different ADAS solutions. These sensors – Lidar, Cameras, Radar (long and short/medium range), and Ultrasonics – provide information about the environment surrounding the vehicle and may even act if necessary, thus benefiting the driver. As presented in Figure 2, the sensors that are part of the vast series of ADAS solutions are already available as equipment in different vehicles (Kukkala et al., 2018).

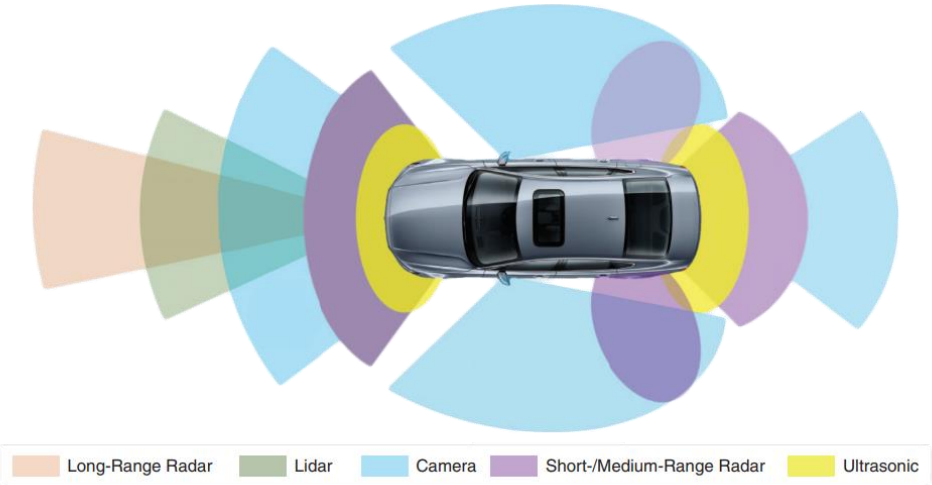

Figure 2 – Sensors integrating different ADAS solutions in a vehicle. *Adapted from* (Kukkala et al., 2018)

2.4. IMBALANCED DATASETS

In classification problems, Machine Learning algorithms assume that the number of observations in each target class is identical. However, in many real situations, this does not occur, and one or more class has more frequency than the others (Krawczyk, 2016).

These cases are called imbalanced data problems, which describe a dataset where one or some of the classes have a much greater number of instances than the others. The most prevalent class is known as the majority class, while the smallest is called the minority class (Weiss, 2004). The minority class is often the most important concept to learn, and it is difficult to identify because it can be associated with rare and significant cases or because it is expensive to obtain data from these examples.

Regarding the multi-class problem, instead of the binary classification one, this issue is common and since there can be multiple minority classes, it is more difficult to solve. Since most standard learning algorithms consider a balanced training set, this can produce sub-optimal classification models (Sun et al., 2007).

Furthermore, it is important to note that for imbalanced data problems sample size plays an important role in determining the "goodness" of a classification model. In fact, it has been proven that as the size of the training set increases, the large error rate caused by the disparity of class distribution decreases. The reason for this is that, when there is a large amount of data, there will inevitably be more information regarding the minority class as well, which benefits the classification model since it becomes better fit to distinguish rare samples (minority) from the majority (Sun et al., 2007).

2.5. SMOTE

Synthetic Minority Oversampling Technique **(**SMOTE) is one method to solve the problem of imbalanced datasets explained above. It is an over-sampling approach in which the minority class is over-sampled by creating synthetic instances and adding them to the training dataset. This is done by performing certain operations on real data, functioning in the feature space instead of data space. This way, the algorithm is based on the values of the features and their relationship, rather than considering the data points as a whole (Chawla et al., 2002).

More specifically, this process works as follows: in each minority data point its k-nearest neighbors of the same class are calculated and randomly selected, according to the over-sampling rate. Lastly, new instances are created along the line between the minority instance and the selected nearest neighbor (Han et al., 2005).

Chawla explored the approach of using SMOTE within a boosting classifier, which improved the recall achieved and did not cause a significant degradation in precision (Chawla et al., 2002). While boosting classifiers improve the predictive accuracy of models by focusing on difficult examples that belong to all the classes, the SMOTE algorithm improves the performance of a classifier only on the minority class examples. Therefore, the embedded SMOTE algorithm forces the boosting algorithm to focus more on difficult examples that belong to the minority class than to the majority class (Chawla et al., 2003).

2.6. EVALUATION METRICS

The evaluation of the classifier's performance must be carried out using specific metrics to take into account the class distribution. Usually, accuracy is the most used measure for these purposes. However, for classification with the class imbalance problem, accuracy is no longer a proper measure, since it does not distinguish between the numbers of correctly classified examples of different classes. Accuracy represents the ratio of correctly predicted instances and is calculated by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

Nevertheless, it is possible to obtain other metrics from the confusion matrix (see Table 1) to measure the classification performance of both positive and negative classes, independently. Since for this classification scenario it is important to achieve good results for both classes, it is necessary to combine the individual measures of the positive and negative classes, as none of them alone is sufficient.

	Predicted Class Positive	**Predicted Class Negative**
Actual Class Positive	True Positive TP	False Negative FN
Actual Class Negative	False Positive FP	True Negative TN

Table 1 – Confusion Matrix

Precision is the ratio of True Positives to all Positives that allows to understand how many instances that were classified as positive are actually positive.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Sensitivity, also known as Recall, is the ratio of correct Positive predictions by all True Positives and it shows how many instances of the positive class were predicted correctly.

$$Sensitivity(or\ Recall) = \frac{TP}{TP + FN} \quad (3)$$

Specificity is the ratio of correct Negative predictions by all True Negatives that shows how many instances of the positive class were predicted correctly.

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

Moreover, precision and recall are not sensitive to data distributions, and they are inversely related to each other. However, from an analytical point of view it is crucial to increase recall without sacrificing accuracy, which can easily occur since, by putting more weight on common classes than on rare classes, accuracy fails to perform well on rare classes and can be a misleading indicator. In this way, it is possible and beneficial to combine the Recall and Specificity, allowing the evaluation of the algorithm's effectiveness in a single class, positive and negative, respectively (Bekkar et al., 2013).

The Receiver Operating Characteristic **(**ROC) Curve is a two-dimensional graph that plots the TP rates and FP rates, summarizing classifier performance over a range of tradeoffs between these two rates, which are also known by specificity and sensitivity. These graphs are considered useful for evaluating classifiers and offering a stronger measure of classification performance than measures such as accuracy and error rate (Fawcett, 2006).

In ROC space, seen in Figure 3, the closer the classifier gets to the top left corner, the better it is, as the corner represents the optimal performance of the model. The upward diagonal represents a random performance, so if the ROC curve is below this diagonal, it means that the performance is worse than a random model. In order to achieve a good model, the ideal is to have the points above and as far away from this diagonal as possible (Fernández et al., 2018).

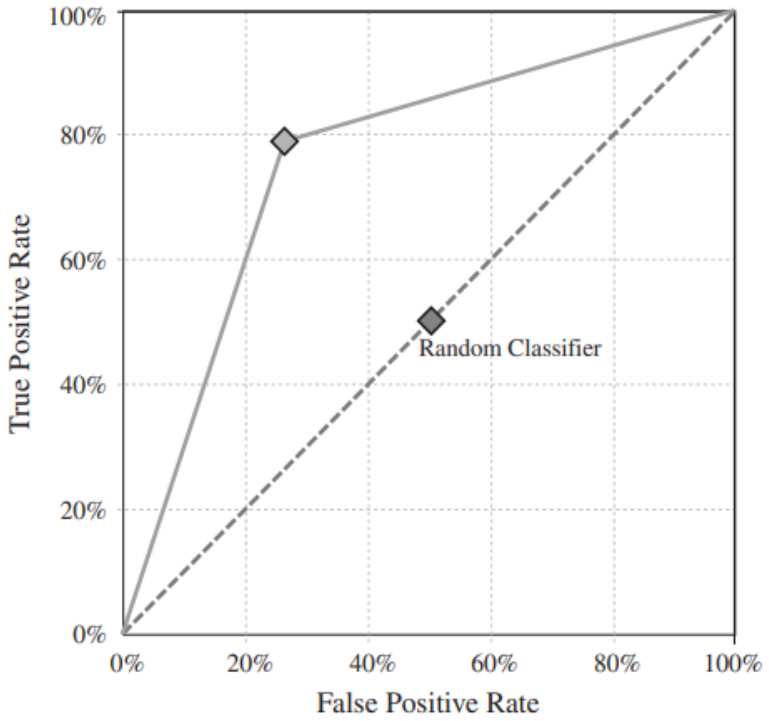

Figure 3 – ROC Curve. Source: (Galar et al., 2012)

Area Under the Curve (AUC) estimates the entire two-dimensional area underneath the curve. It is described as an alternative single-number measure for evaluating the predictive ability of learning algorithms, providing a summarized performance of tradeoff between sensitivity and specificity (Huang & Ling, 2005). With this measure it is possible to compare models and see which has better performance on average (Galar et al., 2012).

2.7. SHAP

Feature interpretation is the essential key to this project, as its success comes from understanding the impact of ADAS equipment on both target variables: frequency and severity. When Machine Learning models are developed, it is often difficult to interpret them, especially with the use of complex algorithms. The Shapley Additive exPlanations (SHAP) proposed by Lundberg et al. (2017) is an excellent approach to the interpretation of the output as, being based on game theory, it provides the contribution of each variable. Through the Python package developed by the same authors, it is easily possible to obtain estimates of these values and a series of visualizations are also available to analyze the results (Lundberg et al., 2017).

This technique was used in an article to study the impact of features on the accident duration in an XGBoost algorithm. Here, the authors mention that SHAP, in addition to being able to evaluate the contribution and its direction of the impacts of each feature, can also extract complex and nonlinear joint impacts of features (Parsa et al., 2020).

3. METHODOLOGY

3.1. RESEARCH FRAMEWORK

Data Mining has had a tremendous growth, and efforts have been made to establish standards in the area. The use of a standard framework is essential, otherwise the success of the project is highly dependent on the skills of those in charge of it. Therefore, to develop a successful data mining project it is necessary to establish standards that help in understanding the business problems through data mining tasks, proposing appropriate data transformations and data mining techniques (Wirth & Hipp, 2000).

The research data mining methodology adopted in this work is CRoss Industry Standard Process for Data Mining (CRISP-DM). This methodology provides a framework to guide a data mining project, aiming to make projects more manageable, reliable, while also reducing time and costs (Shafique & Qaiser, 2014). The choice of this CRISP-DM is based on being considered ‘‘facto standard’’ for developing data mining projects, as well as being the most commonly used (Marbán et al., 2009).

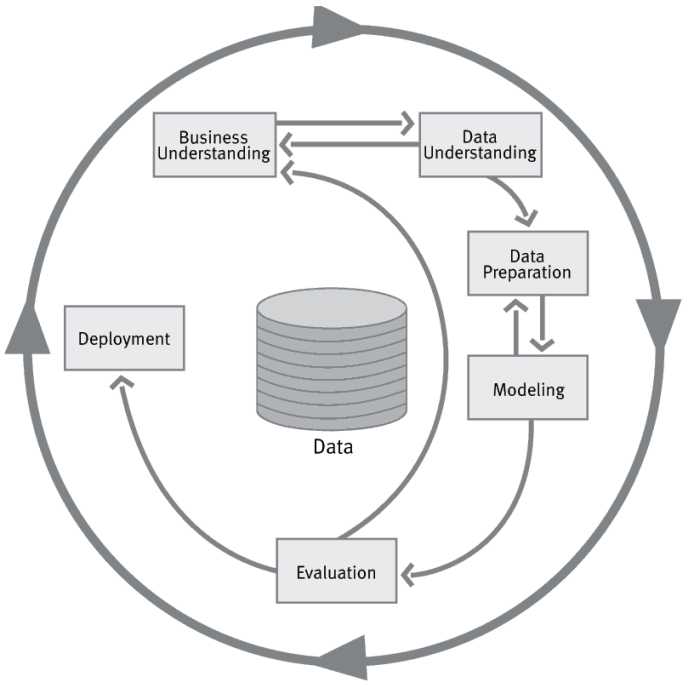

Figure 4 – CRISP-DM research methodology. Source: (Shafique & Qaiser, 2014)

One of the success factors of CRISP-DM is that it is neutral in relation to the industry and application, being also independent of the chosen data mining tool. CRISP-DM includes a structure with six phases that interact with each other iteratively during project development throughout the project lifecycle (Shafique & Qaiser, 2014). The phases presented in Figure 4 are described throughout the thesis.

The first stage of this methodology corresponds to Business Understanding, where the goals of the project are defined according to the company’s business perspective, as well as the plan to reach them. All this information is explained in Chapter 1.

3.2. TOOLS AND TECHNOLOGIES

In this thesis, the main tool used was Python. This is one of the most used programming languages for data science and, due to its large community of users and being open source, it contains a countless number of useful libraries for scientific computing and machine learning (Raschka, 2015, 4).

For the development of machine learning programming tasks in this project, the scikit-learn library was used, which is one of the most popular and accessible open-source libraries to date (Hao & Ho, 2019). The Jupyter Notebook is a browser-based graphical interface for the IPython shell and was used throughout the development of the project since it allows users to include formatted text, static and dynamic views, as well as mathematical equations (VanderPlas, 2016, 2).

To extract data relating to customer policies, the SAS® Enterprise Guide was used. This is the main tool used by analysts within Ageas group, and it consists of a point-and-click tool, which allows users to access, transform, analyze, and export data.

3.3. DATA COLLECTION

In this section, the entire process of collecting, preparing, and processing data will be presented. Data collection in this project was complex, as it involved not only managing the data request from a supplying company, but also creating a sample that was representative of the company's portfolio. In addition, this section also describes the entire data processing flow as well as the practices used.

3.3.1. Portfolio

Seguro Directo's portfolio has more than 325,000 policies, but it was necessary to create a scope for the sample in order to observe only the policies that meet the pre-selected requirements based on the research question.

To carry out this analysis, it was necessary to acquire data from an external company that provides the Vehicle Identification Number (VIN). This unique 17-digit serial number, consisting of numbers and characters, is assigned to a vehicle when it is manufactured. This serial number provides information about the vehicle, each part of its composition being specific to certain characteristics. For example, the first three digits identify the country of manufacture, vehicle manufacturer and the type of vehicle, from the fifth to the eighth digit it identifies the vehicle's brand, body style, engine size, model and series, etc.

Although the whole VIN is relevant for this project, the existing agreement with the data provider is only up to the eighth digit. Therefore, to develop this project it was necessary to obtain the last nine digits of the VIN to know which options each car has, so the supplier company provided a buffer of ten thousand full VIN every year, but more than that represents an extra fee for each car. For this reason, the sample for this study was limited to these 10,000 plates, without incurring in extra charges.

Before selecting the vehicles for the sample, four filters were applied in order to restrict the universe of vehicles within the company, which are presented below.

- Only vehicles of Seguro Directo insurance brand were chosen, due to the small sample and since it was the brand that requested the project.

- Vehicles with less than five years of age. As mentioned in the Literature Review, the proportion of ADAS is more significant in more recent vehicles, thus limiting the age allows for a greater quantity of these systems in the sample as well as their diversity.
- Vehicle category only includes Light, Mixed and Trucks. The selected categories retain a higher percentage of ADAS and policies. The others were not considered, not only due to their reduced application of ADAS, but also to the low relevance of these vehicles in the portfolio.
- Select only own damages option, which is a set of covers that protect the vehicle against unpredicted self-damage. So as to be able to calculate the costs associated with vehicles with these systems, it is necessary to guarantee an insurance with own damages, since this information is only available in the database where there is this condition.

After applying the filters to the company's portfolio, we had 40,555 policies, from which 10,000 were selected for the sample, which was the number stipulated by the company providing the data.

3.3.2. Sample Selection

Before ordering the data provider for all the data, it was necessary to sample two hundred plates to understand if the data received was suitable for analysis. In order to obtain a meaningful sample that reflects the weight of all vehicle types in the portfolio, the selection was based on the weight of each vehicle version, and the choice of entries was random.

At this stage it was noticed that Asian brands do not have optional equipment, since they offer multiple versions instead of just one with the chance to add additional equipment. Thus, it was necessary to make two separate samples, one for Asian vehicle brands completely acquired from Seguro Directo's portfolio, according to the same filters as explained above, and another consisting of 10,000 vehicles for non-Asian brands.

After having validated that the data contained the desired information, of the two hundred plates, only 107 were returned with information since at this stage in the project we still did not know that Asian brands did not have optional equipment as well as certain specific brands. The selection of ten thousand plates was carried out using the same method as in the initial sample. For Asian brands the weight of their versions in the portfolio was also used to select this part of the sample, which led to the selection of 1,715 plates. Considering the whole sample was small, it was agreed that the 107 plates obtained initially would also be included to increase the sample size. Overall, the total number of plates selected was 11,822.

Once the sample was selected, a few analyses were carried out to verify if the plates reflected the portfolio's information regarding the number of claims, premium and cost. It was discovered that, on average, the sample had the same behavior as the rest of the portfolio, so the request for information of equipment of non-Asian plates was made and the process began to be set up to obtain information through our databases for Asians plates.

3.3.3. Extracting data

Regarding the process of actually extracting the data from the company's database, both for Asian and non-Asian vehicle brands, this was done using SAS Enterprise Guide.

For the Asian brands, the process created for this purpose was done in order to obtain the data in the same format and with the equivalent level of detail that was provided by the supplier company for the non-Asian plates. This way, the whole sample was as consistent as possible in terms of the information contained.

As for the information regarding clients and their policies' details, this had to be extracted from another table in SAS. This table is structured by accident period, so it contains several records for each policy, since for every month there is a new line for each registration within the company. It has more than 370 thousand records and 35 variables.

Prior to joining these two tables, it was necessary to change this last one related to policies' information to contain granularity per plates as in the equipment table. With this done, there are, overall, 11,822 observations and 38 variables.

3.4. DATA UNDERSTANDING

Following the steps of the chosen methodology, this section includes accessing the data that was previously collected and exploring it in order to identify data quality issues and to discover early insights and what their impact on the rest of the project could be. This phase is essential to avoid problems in other phases of the project, which can lead to better results and less time wasted.

Before analyzing the data, it was necessary to create the two target variables: Frequency and Severity, which allow to identify the number of claims generated by the policy and its cost. Frequency is the number of claims on the policy's exposure and severity takes into account the average cost, being the total cost over the number of claims.

$$Frequency = \frac{Number\ of\ claims}{Exposure} \tag{5}$$

$$Severity = \frac{Total\ cost}{Number\ of\ claims} \tag{6}$$

Considering the research question, these variables, although initially continuous, were changed into classes. The reason for this it that the goal of this project was not to create a model that has good performance predicting frequency or severity values, but rather that proves the importance of each ADAS equipment on each of them. Therefore, these transformations were applied, and frequency was turned into a binary variable, since we only want to analyze which equipment contributes to having frequency, with the distribution of each class shown in Table 2.

Frequency Binary	Frequency values
Frequency	[0]
No Frequency	>0

Table 2 – Frequency classes

In the case of severity, it was decided to only consider the observations that had a frequency greater than zero, considering the objective is to understand how the equipment impacts specific classes of cost and not whether there is a cost. Therefore, these variable's values were grouped into three classes, as shown in Table 3, in order to maintain information on the distinct classes of different cost levels:

- The first class represents lower costs, with severity between 1€ to 350€;
- The second class is composed of medium costs, with values between 350€ to 1,000€;
- The third class, with costs between 1,000€ and 30,000€, which represents the maximum cost value average in the sample, constitutes the class with higher costs.

Severity Class	Severity Interval
Class 1 (Low Cost)	[1€;350€]
Class 2 (Medium Cost)	]350€; 1,000€]
Class 3 (High Cost)	]1000€; 30,000€]

Table 3 – Severity classes

Once the data collection process was finished, the raw data contained 11,823 records and 47 variables, with 35 variables regarding customers' personal information, vehicles' general details, policies, and number of claims and 12 variables regarding vehicles and their equipment, grouped into several categories. Descriptive statistics were developed to understand each relevant variable, as well as the behavior of the target variables frequency and severity.

Analyzing the two target variables, it was noticed that there was a problem of imbalanced data (Figure 5), especially on frequency, since there were few claims. In frequency target, the weight of the minority class is 2,856 (24.16%) which represents cars that do not have a frequency, against 8,966 (75.84%) of cars with frequency. Severity target has a more balanced distribution: in the Low Cost class it has 995 (35.12%), 997 (35.19%) in the Medium Cost class and finally 841 (29.69%) in the High Cost class.

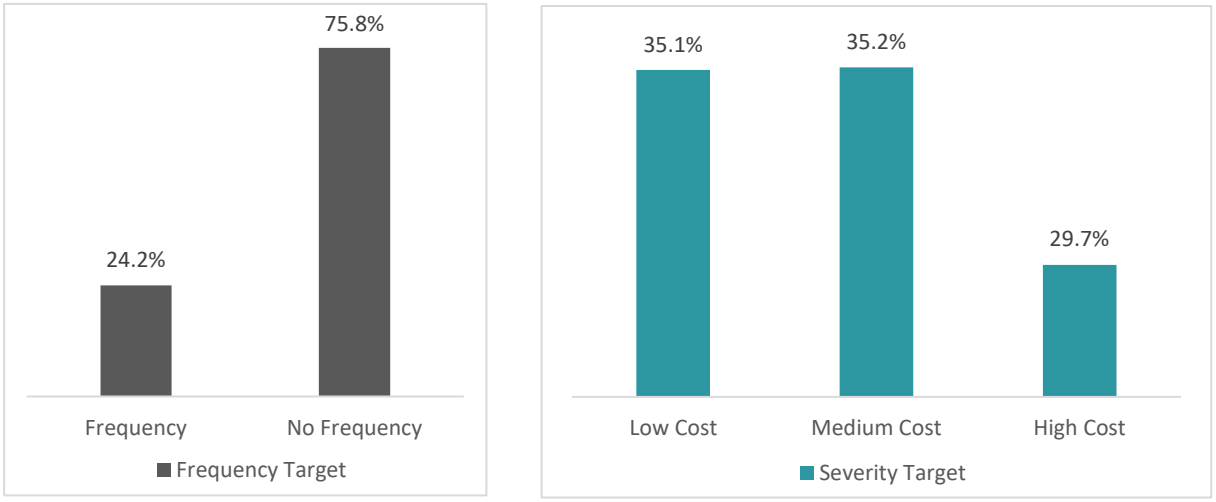

Figure 5 – Distribution of values in the target variables (Frequency and Severity)

Regarding the equipment received, it was immediately possible to identify the percentage of equipment that are ADAS, considering the company providing this data identified them as being all gear that belong to two groups: Active Safety and Assisted Driving. As shown in Figure 6, of the 361 devices, 50 are ADAS devices and the remaining are other devices, such as paint, tires, and chassis information. This step was important since it enabled to immediately identify certain characteristics to be removed. In fact, since the sole purpose of this project is to identify the impact of ADAS in the target variables, not of any equipment, only ADAS equipment were selected for the rest of the analysis.

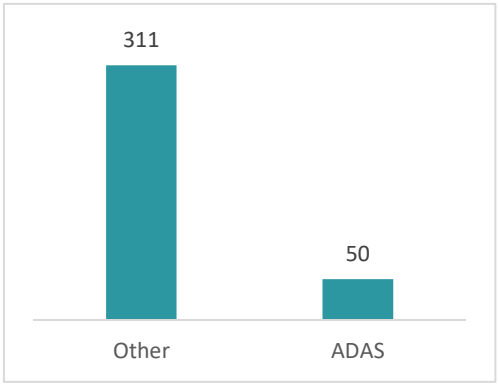

Figure 6 – Type of equipment in the portfolio's vehicles

Furthermore, looking in more detail at ADAS, there were some pieces of equipment that did not fit into this group, due to the fact that they do not possess the characteristics of ADAS, as defined in the literature review. A total of 29 devices were excluded from the feature list, including some such as lights and headlights.

Additionally, the median of ADAS equipment per vehicle is 5, the maximum equipment is 18 and the minimum is 1. Also, as seen in Figure 7, the equipment that has the greatest representation in the sample, with 10,031 vehicles containing it, is Stability Control and the equipment with less is Traffic Signal Recognition, which is present only in one vehicle.

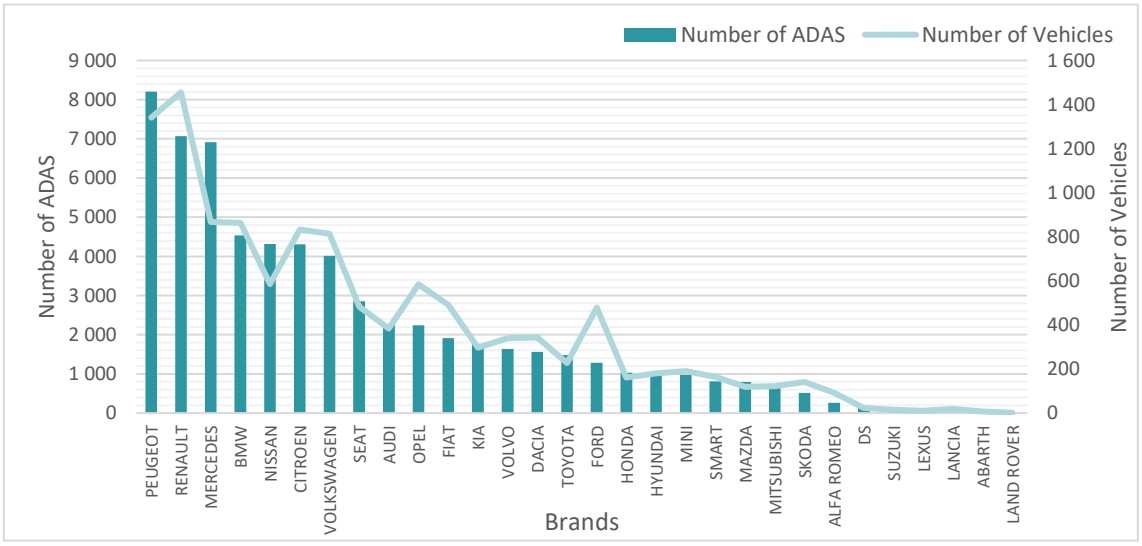

Figure 7– Number of ADAS and vehicles by brand

Finally, a more general analysis of the sample enabled to find that there are 30 different brands in analysis, and the most frequent one is Renault. Moreover, the brand with a higher number of ADAS in total is Peugeot, while Land Rover is the one with the smallest number of these devices found, in part because it is the brand with a smaller number of cars. When considering the ratio between the number of vehicles and the number of ADAS in each brand, it is possible to see that Mercedes is the brand with the highest value, making it the brand with more equipment per vehicle.

3.5. DATA PREPARATION

Per accordance with CRISP-DM phases, following data collection and understanding tasks, the dataset should be prepared for modelling. As such, data cleaning and feature selection tasks were performed and will be detailed in this section.

3.5.1. Data Cleaning

Data cleaning is a very important step since raw data can have many problems that can compromise the quality of a model and its results and analysis. In this dataset, these problems mostly result from errors in the database creation process, more specifically in the lack of necessary conditions/rules to guarantee the quality of the data at the moment of its creation.

Missing Values

The treatment of missing values should be one of the first stages. There are eleven variables with at least one missing value, of which five have less than 10% of absent observations. The remaining six variables have a high percentage of missing values, as they are related to the number and cost of claims, in which, for the cases where there are no claims, all related variables take null values. Table 6 contains these variables and their representation, and it can be found on the appendix.

All values were handled by filling in the missing values (imputation), but different rules and approaches were applied for each case. The variables related to the clients' geographical area were filled in based on the vehicle's brand and the rationale used was to choose the most frequent value within the brand where each observation with missing values belongs to. As for the drivers' age and vehicle capital, these were filled in using the median, since this measure is more recommended than the mean, as it prevents the bias of extreme values. Total premium was handled using KNN imputation technique, which takes into account the total premium value of its neighbors, thus enabling for a more fitted imputation based on the observations more similar in terms of the other non-null variables. Finally, all variables related to number and cost of claim were filled in with zero because of the way the database is done: if there is no claim record, it means that it does not exist.

Outlier Detection

Secondly, it was necessary to understand whether there were outliers since these values can be very disruptive to models. For this purpose, a simple univariate technique was used through a visualization of the distribution of variables using boxplots. Additionally, to complement the detection of these values, a multivariate technique was also used through scatter plots, which allow the visualization between two dimensions.

In this stage, 0.085% of the observations were removed, from the variables Frequency, Total Cost, Medium Cost and Driver Age.

Data Transformation

Additionally, as was previously noted, the number of ADAS equipment was high and some of them not very significant, leading to a high dimensionality of the dataset. For this reason, it was necessary to create groups, keeping the information and managing to identify them in the future to apply the potential discount. Based on a study by SEMA (2020), the 21 ADAS equipment in this study were divided into seven different groups, each representing a new feature, displayed in Figure 7. These variables consist of the sum of equipment that each vehicle has in each group (John Waraniak et al., 2020).

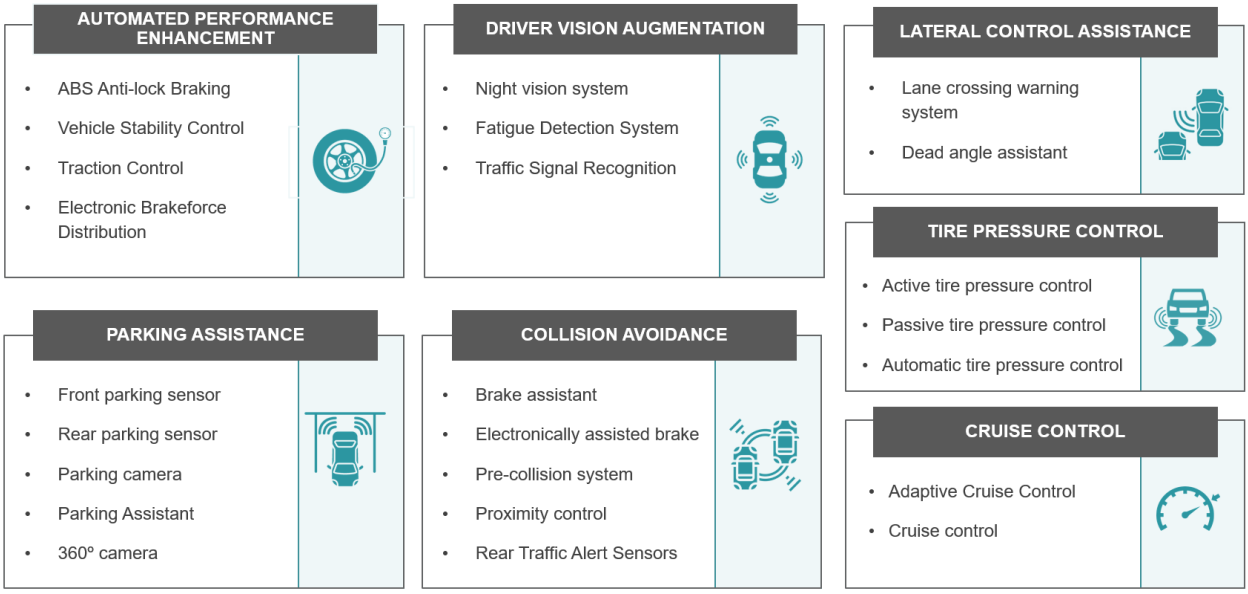

Figure 8 – ADAS groups created and respective equipment

Another processing step implemented was the one-hot encoding of categorical variables. In another words, to represent a categorical variable that can take k different values, k - 1 dummy variables were defined for each categorical variable. This task was performed because the machine learning models that will be trained during the modeling phase require all input resources to be numerical. The variables that were transformed with this approach are: Brand, Model, Version, District, NB_RW and Equipment. After this task, the dataset had a total of 86 input variables and 2 target variables.

Some other variables were created as well. The variable that identifies vehicles with more than 5 ADAS was created (Figure 8) based on the median of ADAS equipment per vehicle, as seen in the data exploration.

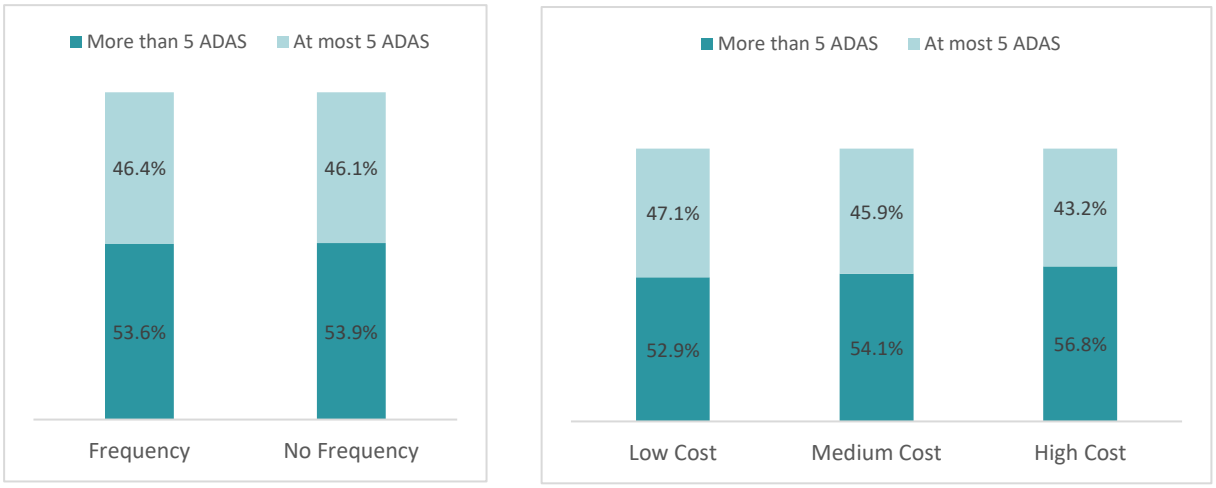

Figure 9 – Distribution of ADAS per each target class (Frequency and Severity)

Correlation

Moreover, it is important to check for variable correlation, since correlated features increase the complexity of the algorithm, thus increasing the risk of errors, without providing additional information. Therefore, an analysis was performed to identify and eliminate some variables with high correlations, to avoid problems of multicollinearity and redundancy. Pearson correlation was performed for the numerical variables and out of the 86 considered, 15 were highly correlated. For all pairs with a correlation coefficient greater than 0.5, they were either eliminated or a combination of variables was made in order to not lose the information of any of the variables, thus helping to reduce the dimensionality of the dataset.

After having handled all these steps, the plates identifier was removed and in total, 11,627 observations remained in the dataset.

3.5.2. Feature Selection

Feature selection is used to reduce dimensionality choosing a shorter feature list with the most relevant variables, resulting in better performance, lower computational cost, and better interpretability of the model (Tang et al., 2014).

Four methods were established for feature selection, to choose only relevant variables from a model and decrease dimensionality. The first two are Shrinkage methods: Lasso and Ridge, also known as regularization. This approach penalizes the magnitude of coefficients of features while also minimizing the error between predicted and actual observations. The other two are sequential selection methods: Forward and Backward regression models.

Ridge and Lasso regression

Ridge regression performs L2 regularization, thus penalizing the least squares loss by applying a ridge penalty on the regression coefficients.

$$\widehat{\beta}^{\ ridge} = arg\ _{\beta}\min \left\{ \sum_{i=1}^{N} \left((y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j) \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \quad (7)$$

Here $\lambda \geq 0$ is a tunning parameter that controls the amount of shrinkage. The penalty shrinks the regression coefficient estimate towards zero, but not exactly zero. Lasso regression is a particular case of the penalized least squares regression with L1 penalty function (Muthukrishnan & Rohini, 2017).

$$\widehat{\beta}^{\ lasso} = arg\ _{\beta}\min \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (8)$$

Lasso increases the interpretability of the model, eliminating irrelevant variables, consequently reducing the variability of the estimators by shirking some to exactly zero. The main difference between the two regressions is that in ridge regression, the model complexity is reduced by decreasing the magnitude of the coefficients, but it never shrinks coefficients to exactly zero. However, Lasso regression tends to make coefficients to absolute zero (Muthukrishnan & Rohini, 2017).

Forward and Backward regression

Forward begins with the baseline model and features are added, creating a sequence of models that increase sizing. In each step the feature which gives the lowest prediction error is chosen. This is an iterative process, which means that once the feature enters the model, it is never removed, only other variables are gradually included as well. This is repeated until all features are included in the model and the best combination is chosen by identifying at which step there is no longer a great improvement of model performance by adding new features.

Backward elimination works similarly to Forward selection, but it starts with the whole feature set and all features are removed one by one, until there is only one variable remaining in the model, which is the one that best explains the target variable. As before, once a feature is excluded, it is not added back to the model again (Andersen & Bro, 2010; Bursac et al., 2008).

After the four approaches were executed, the best predictor was outlined for each method. Then, by assessing, combining, and comparing the four feature lists obtained, the set of relevant variables was chosen to be used in the modeling stage.

It is necessary to note that the variables related to the equipment were taken into special consideration, to be easily selected since they could not all be excluded, considering the goal of this analysis. Therefore, they were compared among each other, and only the most relevant ones were selected, thus allowing to guarantee the research question.

This process was carried out for the two target variables, frequency and severity, and as expected due to the similar nature of the problems, the results were very identical. It is important to mention that the models must be comparable, being able to understand the relationship between the two targets, so in this sense, the variables of both models must be the same. Displayed in Table 4 is the final feature list used to train both models in the analysis.

Final Variables Selected
More than 5 ADAS
ADAS optional
Exposure
Vehicle's Capital
Driver's Age
Number of Claims Last 5 Years
Average Premium
Automated Performance Enhancement
Forward Collision Avoidance
Parking Assistance
Tire Pressure Control
Lateral Collision Avoidance
Driver Vision Augmentation
Cruise Control

Table 4 – List of selected variables

3.6. MODELLING

In this section, the results of the models applied to the two target variables, frequency and severity, are presented and compared through the metrics described in the literature review. The dataset was partitioned into training and testing sets, with, respectively, 70% of the data used to train the classifiers, and the remaining 30% used to evaluate their performance and ability to generalize.

Cross validation was used to evaluate and compare the algorithms. With this method, the dataset is divided into k equal subsets and in each iteration a different fold is chosen for validation and the remaining k-1 folds are trained. This way, more diversity is ensured and there is more confidence in the consistency of the results and in the performance of the model since the validation subsets are different every iteration. For the following models, a 10-fold cross-validation was performed.

At this stage, multiple algorithms were considered in order to find the fittest one for the data. Each of them was tuned with several hyperparameters through the use of a Grid Search Method, which considers several combinations of hyperparameters, providing the best possible combination among all the possible ones. Each of the considered algorithms was exhaustively evaluated for its performance and select the optimum values from a list of parameters provided.

3.7. EVALUATION

The best performing models for frequency were four Ensemble classifiers, three boosting models: AdaBoost (Adaptive Boosting), Gradient Boosting, XGBoost and one bagging model: Random Forest. Out of all, the winning one was XGBoost Classifier, according to the evaluation metrics used.

Analyzing the performance of the chosen model, as seen in Figure 9 and Figure 10, in the test set it achieved an AUC of 84%, a Precision score of 79% and a Recall of 76%. The AUC measure was decisive in choosing the algorithm because, as was explained in the literature review, it reflects in one single number the ability of a classifier to distinguish between classes.

Moreover, it is relevant to emphasize that, due to the use of SMOTE to smooth the imbalance in the data, it was possible to obtain a less significant difference between the two classes of Frequency, both in terms of precision and in recall. In fact, XGBoost is the model that obtained higher precision for the frequency class, while still managing to achieve high precision for the no frequency class, thus resulting in a less accentuated difference between the two classes.

Additionally, as seen in Figure 10, it is also important to state that there is no significant evidence of overfitting in any model, regarding accuracy and AUC.

Figure 10 – Precision and recall for the Frequency model, in the test set

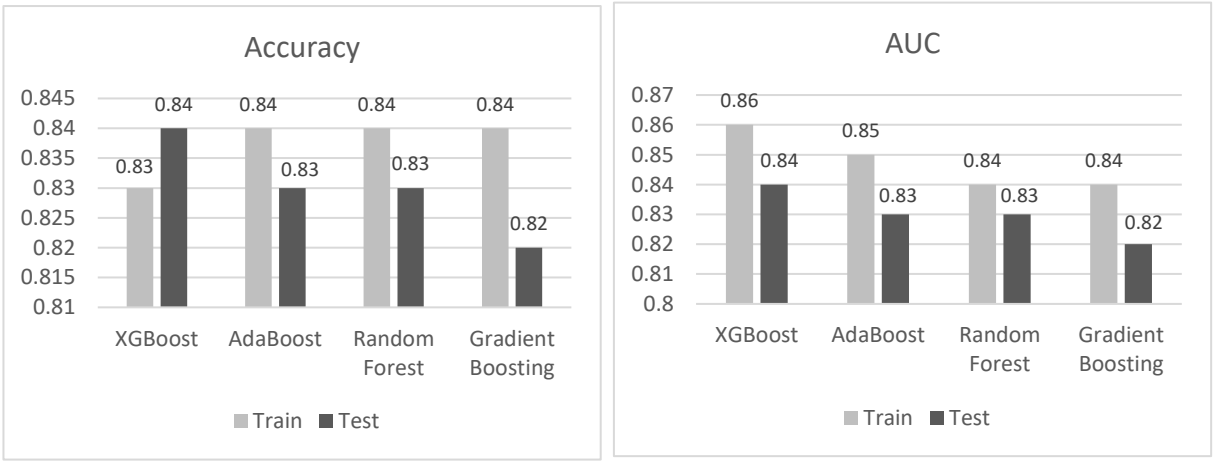

Figure 11 – Accuracy and AUC for the Frequency model

Regarding the severity target, AdaBoost and Gradient Boosting provided a good performance as well as two other algorithms: Linear Discriminant Analysis (LDA) and Gaussian Naïve Bayes (Gaussian NB). Taking into account the four main metrics used to evaluate their performance, the best one was Gradient Boosting Classifier, which reached an AUC score of 59%, with both precision and recall of 42%, all in the test set.

Observing Figure 11 below in more detail, it is possible to see that in the test set the class which obtained lower precision in all models is Medium Cost, though with very little difference, and regarding recall, the Low Cost class achieved slightly higher results for all models. For both metrics, Gradient Boosting yielded, overall, the best results.

As for Figure 12, despite the slightly higher evidence of overfitting, it is not too significant and the Gradient Boosting Classifier achieves higher scores than the other models in both metrics, for both the train and test set.

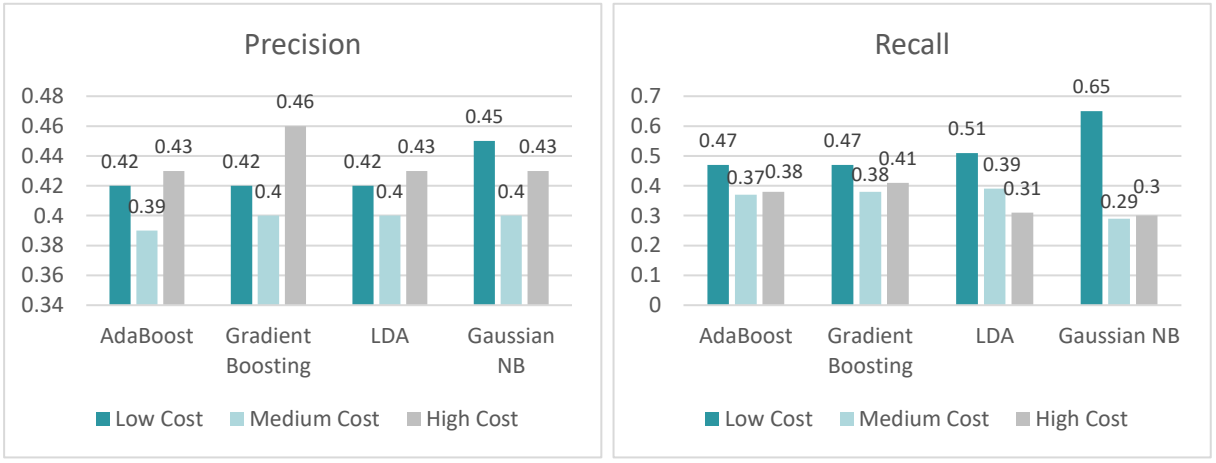

Figure 12 – Precision and recall for the Severity model, in the test set

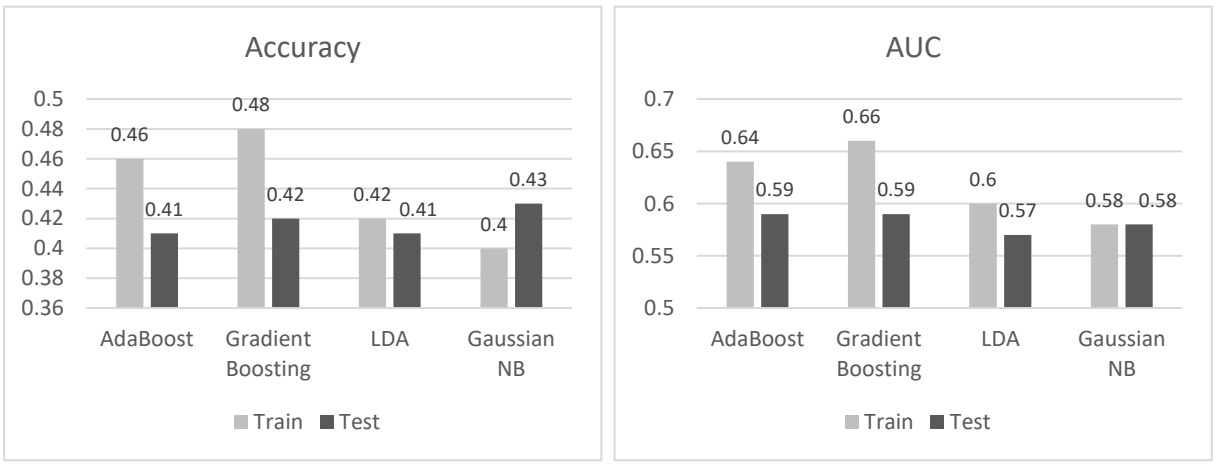

Figure 13 – Accuracy and AUC for the Severity model

The continuation of the evaluation, such as the results related to the feature selection where the results of the ADAS equipment are analyzed through the SHAP tool and the evaluation of the impacts of these features for the two target variables is dealt with in section 4 further on, where the results and the discussion are found.

3.8. DEPLOYMENT

According to CRISP-DM, after the evaluation of the results the strategy for the deployment of the project is defined, which is crucial for its success. As this is a pilot project, the objective was to prove that there is an impact of these equipment on the frequency and severity. As such, implementation, monitoring, and maintenance are out of scope. However, this section presents a proposal for the deployment plan.

The deployment of this project will depend on its feasibility on the company side. If the Group decides to invest on the development of this project, the new model must be able to quantify the impact of the equipment and understand what the implementation's risk and cost would be, and a few criteria would have to be met in order to develop it successfully.

Firstly, it is necessary to understand how the information about vehicle options would be collected, since at this moment not all vehicles have the same level of detail regarding this, which compromises the results of the analysis. Furthermore, defining when to acquire this information is crucial to decide what benefits to give to customers with specific equipment. For this, two possible solutions are proposed: the first would be to obtain this information at the time of the customer's insurance simulation on the company's website, by requiring them to fill out a form concerning which ADAS equipment their vehicle has. This would then have to be validated to guarantee the veracity of the information and an adjustment of the premium would be made afterwards, upon confirmation. The second option would be similar to the first one, only it would not be the customer stating the vehicle's options, but rather this information would be automatically available through a web service solution provided by the same company that provides information on the Seguro Directo's equipment. Even though this service has associated costs for the company, it brings added value in terms of information, which can be extremely useful and innovative for future projects and strategies.

For this reason, I recommend the second option, as it allows to immediately apply the corresponding discount in the policy, thus reducing efforts in the underwriting area with the application of discounts, while also potentially attracting more customers who can immediately be aware of the direct discount they benefit from.

Concerning the model, all the processes of data extraction, data preparation, feature selection and modeling must be automated and integrated into a pipeline in order to produce the results passing through the entire workflow of the model. This way, the same process is fed with new data and produces results that should then be evaluated regularly, since the increase in ADAS has a high expected growth. The marketing team together with the pricing team should evaluate the results and reconsider the discount campaigns if necessary.

Nonetheless, due to this constant innovation and growth of new technologies, it is vital to be aware of the potential need to update this model. This is because the emergence of new relevant equipment is inevitable, so the model must be changed to accommodate the new features.

4. RESULTS AND DISCUSSION

The following section discusses the results obtained from the frequency and severity models. To answer the research question, it is necessary to be able to interpret the contribution of each variable in the models. The algorithms used, despite generally presenting better results compared to simpler ones, are not very transparent, which makes them poorly interpretable, and known as black box algorithms. So, a solution to this problem is to use SHAP. This tool allows to identify and visualize how and how much the variables are contributing to the predictions.

Frequency Model

As seen in Figure 13 found below, it is possible to analyze the contribution of each variable and its impact. All equipment groups have a negative impact on the target variable, as well as the variable "More than 5 ADAS" being the variable that contributes most after the Exposure and Number of Claims Last 5 years, which clearly has a high weight for the model. Tire Pressure Control (TPC) is the group that stands out the most in the results, unlike Driver Vision Augmentation (DVA), which according to the SHAP value represents the least importance among the other groups.

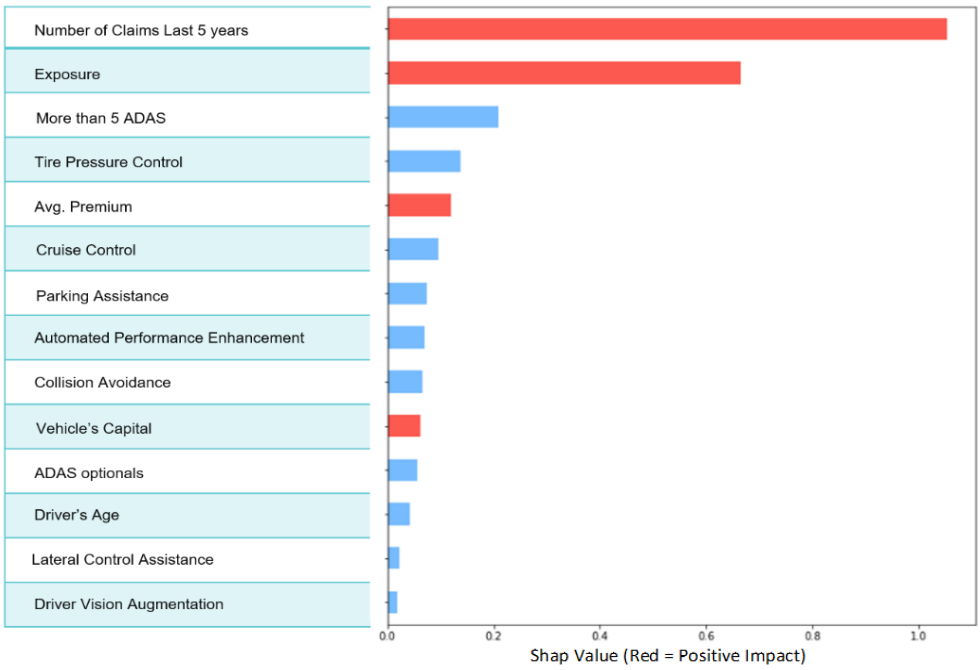

Figure 14 – Simplified SHAP values for the Frequency model

To better understand the positive and negative relationships of the predictors in the target variable, a summary plot for the SHAP values (Figure 14) was displayed, which shows a set of points for each variable, where each point represents an observation of the dataset. Blue dots mean that the observation in this variable has a low value, whereas red dots indicate high values. The position of the point along the horizontal axis indicates the contribution of the SHAP value to the target variable, in the sense that the further it is from the center, the bigger its contribution.

Analyzing Figure 14, it is possible to understand that most groups of equipment have a clear division of colors, with blue and red in opposite directions, which indicates that they are good predictors to interpret frequency because by changing the value of the variable, we can clearly perceive its effect on the target variable. None of the equipment groups varies much on the horizontal axis, instead they are all located near the center, which shows little relevance. However, they are all on the left side of the graph, which indicates that they have a negative impact on frequency. This result supports the hypothesis that, in general, ADAS equipment contributes to less accidents, and consequently, less claims and frequency.

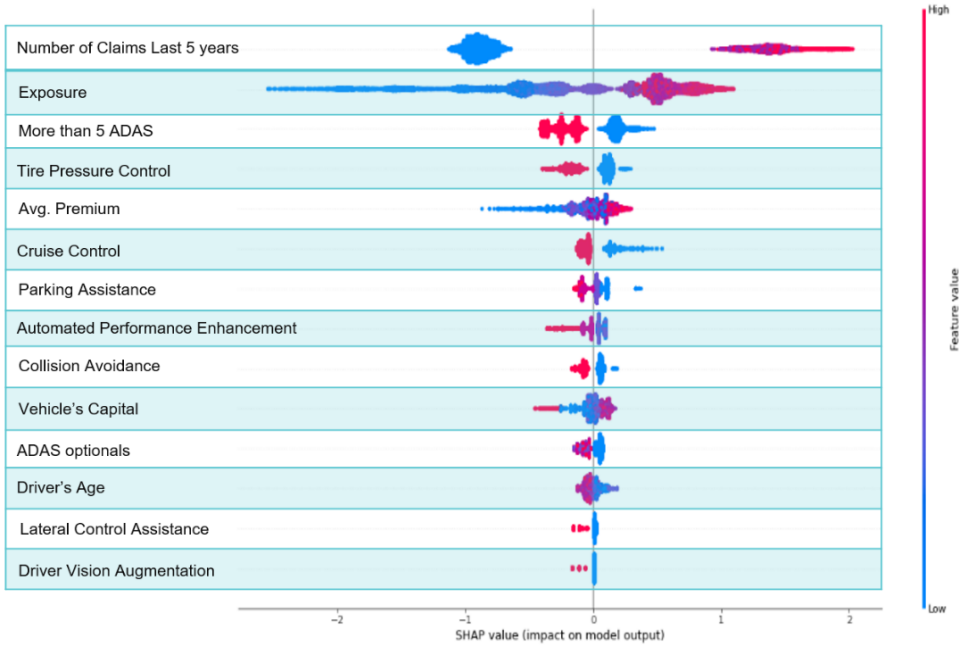

Figure 15 – SHAP summary plot for the Frequency model

In order to summarize the results of the ADAS impacts on the target frequency variable and acquire a better understanding of which are more relevant, three distinct groups/labels were created.

The first group represents the equipment that appears to have a bigger effect on this model and is composed of TPC and Cruise Control (CC). For simplicity purposes, this group is labeled "High Importance" since, when compared to the other ADAS, these are the ones that impact the frequency in a more negative way, thus helping to reduce the loss ratio. The second group, called "Medium Importance", contains Parking Assistance (PA), Automated Performance Enhancement (APE) and Collision Avoidance (CA).

Finally, the variables Lateral Control Assistance (LCA) and DVA constitute the group called "Low Importance", since this group, despite having a negative impact, is not very relevant in explaining the frequency.

Severity Model

The analysis of these results should be carried out in two perspectives: looking only at the ADAS equipment and their global importance according to SHAP, but also looking at the behavior of these

equipment within each severity class. With the latter approach it is possible to attribute the same labels as for the frequency model – High, Medium and Low Importance -, making the comparison between the results of the two models easier.

Observing Figure 15 below, it is possible to see the APE and CA variables have a higher impact compared to the rest of the equipment, although both are impacting only two of the classes: high and low cost. After these, the variable LCA has an impact only on the same two classes: low-cost and high-cost. TPC and PA display lower values, but with representation in all three classes of severity. Finally, CC, DVA and More than 5 ADAS variables have no impact on the target severity variable. The summary plot referring to the target severity variable can be consulted in the appendix section (Figures 23, 24 and 25).

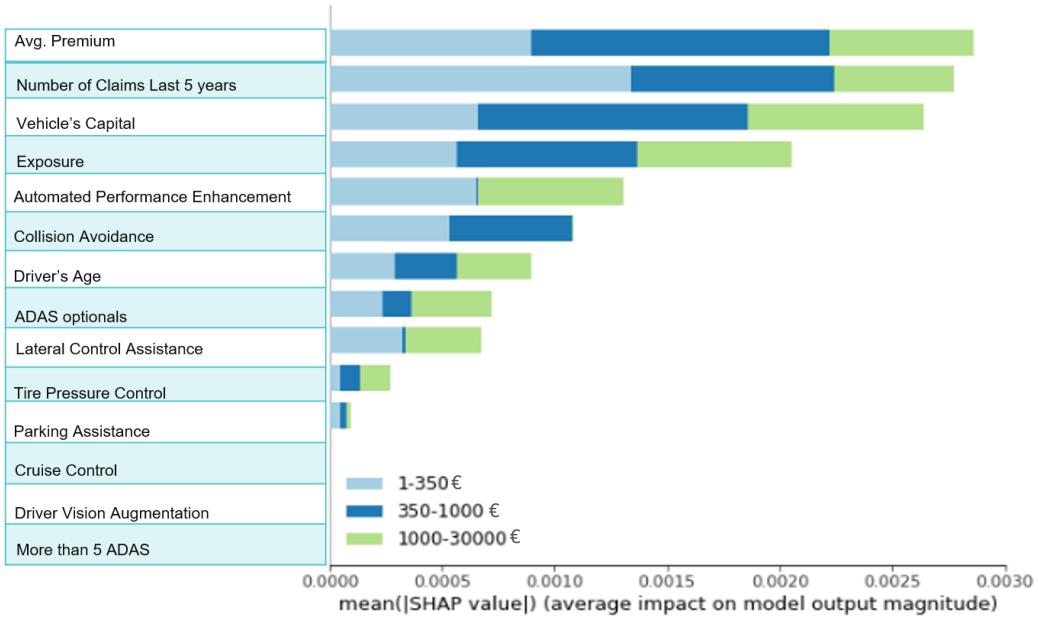

Figure 16 – Simplified SHAP values for the Severity model

Analyzing in more detail the results by class, as can be seen in Table 5 below, it was observed that in the low cost (containing observations with costs between 1€-350€), the variables that are more important are the APE and CA, the LCA equipment presents a “Medium Importance”, and equipment with “Low Importance” are TPC and PA.

Secondly, in the medium cost class, only three equipment are representative: CA with “High Importance”, the TPC and PA with “Low Importance” for the target variable.

Lastly, in the high cost class, the APE has “High Importance”, the LCA with “Medium Importance” and with “Low Importance” there are the equipment TPC and PA.

Overall, when combining the results from both models (Table 5), some ADAS groups displayed different importance and should be briefly analyzed separately. This way, it is possible to directly address the business goals of this work and further on deliver the most relevant insights obtained to the responsible teams as mentioned previously.

	TPC	CC	PA	APE	CA	LCA	DVA
Frequency	High Imp	High Imp	Med Imp	Med Imp	Med Imp	Low Imp	Low Imp
Severity							
Low Cost	Low Imp	-	Low Imp	High Imp	High Imp	Med Imp	-
Medium Cost	Low Imp	-	Low Imp	-	High Imp	-	-
High Cost	Low Imp	.	Low Imp	High Imp	-	Med Imp	-

Table 5 – ADAS importance for each target variables
(red – positive impact, blue – negative impact)

It can be seen that the variables with higher impact in the Frequency model are some of those who are less relevant for all classes of severity. In fact, TPC is considered as "Low Importance" for all cost classes and CC has no impact at all for them, while on the Frequency model these are the "High Importance" ones. This could be explained by the fact that the Frequency model is taking into account the records with no claims, which, on the severity model would represent a cost of 0 and were thus excluded. Therefore, perhaps these variables are more relevant at explaining cases where there are no accidents at all, and not exactly how much their costs are, partly since it could be that vehicles with these ADAS have few claims, thus very often a Frequency of 0, but no costs at all. Consequently, these variables possibly contribute to preventing the frequency of claims and, in case of a claim, they will not greatly impact its cost.

As for APE and CA, these groups reveal "Medium Importance" for the Frequency model and "High Importance" in some of the severity classes. Looking more in detail into them, they are composed of features such as Electronically assisted brake (EAB) and Anti-lock braking system (ABS), which guarantee stability and braking in case of danger. This way, even though these ADAS groups do not prevent entirely the occurrence of an accident, as seen by the negative "Medium Importance" in frequency, they possibly contribute for it to be a less severe accident, hence with a lower cost, thus explaining the positive importance in Low Cost class and a negative importance in the other classes.

Therefore, all the groups mentioned above are the main ones that should be considered by the responsible teams, since they demonstrate a promising potential to be applied a discount.

Regarding PA, LCA and DVA, these groups do not exhibit consistent importance for both models combined. For this reason, considering the attribution of discounts has associated costs for the company, these three groups do not reveal a sufficient importance to be explored in an initial phase.

Finally, another result worth mentioning is that, in both models, the variables that contribute the most for each target are not related to the vehicle's equipment, but rather related to the client's claims history and exposure, and in the case of severity, to the policy's premium and vehicle's capital. This reflects the previously stated hypothesis that there are certain factors that have a much greater and more direct impact on these target variables, which must be considered.

5. CONCLUSIONS

This work focuses on understanding the impact of ADAS equipment on frequency and severity, to serve as evidence for a future project where their impacts will be quantified and discounts implemented or, if possible, incorporate them as a risk factor in the calculation of the premium. Therefore, the aim is to evaluate whether ADAS has a positive or negative impact and its weight for the models.

To achieve the intended objective, according to the Pricing area's requirements, two models were developed: one for Frequency and one for Severity. For the Frequency model, the goal was to assess whether or not the ADAS equipment had an impact on this variable, and hence the variable was transformed into binary. In the case of Severity, it was essential to understand the importance of equipment for different values, therefore, three different classes were created: low cost, medium cost and high cost. This way, it was necessary to develop two classification models, a binary frequency model and multi-class severity model.

In the development of the project, the methodology chosen was CRISP-DM, which helped to manage and plan each phase and encouraged best practices for the project. The data collection process was complex, due to the need to create two representative samples of the portfolio for Asian and non-Asian brands, and the high cost of acquiring information about vehicle options led to these samples being small.

Regarding modeling, for both models, different algorithms were implemented with 10-fold cross validation using a grid search in order to find the optimal parameters. Afterwards, they were evaluated through the metrics presented in the literature review, and the best algorithms for each model were compared and chosen. For the frequency model, XGBoost had the best performance with an average cross validation AUC of 84%, 79% of precision and 6% of Recall. In the severity model, the algorithm with the best results was Gradient Boosting with an AUC of 59%, precision of 42% and a recall of 42%.

Furthermore, in order to evaluate the impact of the equipment, it was necessary to analyze the importance of the features in the models. However, since the use of black box algorithms makes this process difficult, the use of the SHAP tool was necessary to allow for a clearer evaluation of the results through the SHAP value, also providing a number of graphs which helped in the interpretation of the ADAS impact.

With this project it was possible to see that, in general, ADAS equipment do in fact contribute for these models. Regarding the frequency model, it was observed that all variables have a negative impact on this target, in particular TPC and CC which have a high impact. As for the severity model, APE and CA have the highest impact on severity, especially on the low-cost class, where this impact was positive, contrarily to the other classes. These four groups of ADAS equipment reveal the highest potential for success in the application of discounts to customers and should be further explored in an initial phase.

Nonetheless, there are many factors that lead to road accidents, and vehicles' equipment are only a part of it, so the implementation of discounts based on them must be assessed very carefully. This project will be presented to the marketing and pricing teams, in order to study the feasibility of a future study with more information and so as to quantify the impact of ADAS equipment in terms of frequency and severity.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Firstly, there were several limitations regarding data. As a result of budget restrictions, only 10,000 records were used for this project, when the sample was initially thought to consist of 50,000 records. Due to the scarcity of data on vehicles' equipment, a larger sample is expected to yield better results, so, in future studies, an effort to obtain more data is necessary. Moreover, there are certain inconsistencies in the databases' historical data which make the entire data preparation process more complex and time-consuming, beside compromising the quality of the data. Additionally, the provided data on vehicle equipment was not leveled for all vehicles or brands, that is, certain brands contained much more detail about equipment on their vehicles, not because they truly contain more equipment, but because they were registered, while some other brands did not register everything. For this reason, the data can lead to wrong conclusions because it does not reflect the truth. Consequently, it could be relevant to study the possibility of finding a company that provides this data with more consistency and coherence.

Secondly, in the deployment phase a barrier is put up because customers simulate the commercial premium on the site before joining the insurance. It was suggested that the information of vehicle equipment must be verified at the time of simulation to generate a discount immediately, but to obtain this information there is an associated cost that would not be viable to spend due to the huge amount of simulations that are done every day.

Thirdly, despite having become very popular, there is still little information regarding the growth of ADAS equipment and the application of models or studies that evaluate their performance. Additionally, the identification of these equipment is still unclear, as each brand assigns its own name to each equipment, which greatly complicates data preparation and time spent on processing it to harmonize them. Thus, it would be beneficial to encounter new ways to group equipment or to be able to distinguish the differences in equipment names between brands.

Another limitation and one of the greatest ones, is that there is no possibility of identifying the cause of the accident, which often leads to penalizing equipment for the sole fact that a vehicle that had a claim contains it. That is, for example, if a vehicle has an accident when parking we will not only blame Parking Assistance but all the equipment in that vehicle.

One more important note is that the cost considered includes the deductibles, which are a portion of an insured loss that an insured person must pay before the insurer covers the remaining part, dependable on the contract, so it is not possible to know what the real cost of the claim was. This leads to a misinterpretation of costs since, for instance, an individual may have a lower cost than another for having a higher deductible, although having a higher cost claim. Therefore, it would be crucial to acquire information from the company's databases regarding the deductibles or the real cost of claims, not only the cost that the company incurred with the claim, to have more reliable and consistent results in the target severity variable.

Finally, due to inconveniences related to the pandemic situation there was a significant delay in the assignment of an internship project, caused by the cancellation of the initial project. Therefore, the time to develop the work project was greatly reduced, limiting the testing of other techniques and solutions.

BIBLIOGRAPHY

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, *24*(11–12), 728–737. https://doi.org/10.1002/CEM.1360

Antonio, K., & Valdez, E. A. (2011). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis 2011 96:2*, *96*(2), 187–224. https://doi.org/10.1007/S10182-011-0152-7

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, *46*(3), 735–752. https://doi.org/10.1007/s11116-018-9890-7

Bekkar Mohamed, Hassiba Kheliouane Djemaa, & Taklit Akrouf Alitouche. (2013). *Evaluation Measures for Models Assessment over Imbalanced Data Sets*. Journal of Information Engineering and Applications.

Brookhuis, K. A., Waard, D. de, & Janssen, W. H. (2001). Behavioural impacts of Advanced Driver Assistance Systems–an overview. *European Journal of Transport and Infrastructure Research*, *1*(3), 245–253. https://doi.org/10.18757/EJTIR.2001.1.3.3667

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*(1), 1–8. https://doi.org/10.1186/1751-0473-3-17/TABLES/6

Chawla, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/JAIR.953

Chawla, N. v., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, *2838*, 107–119. https://doi.org/10.1007/978-3-540-39804-2_12

David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, *20*, 147–156. https://doi.org/10.1016/S2212-5671(15)00059-3

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/J.PATREC.2005.10.010

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. *Learning from Imbalanced Data Sets*. https://doi.org/10.1007/978-3-319-98074-4

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Gerónimo, D., López, A. M., Sappa, A. D., & Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1239–1258. https://doi.org/10.1109/TPAMI.2009.122

Golias, J., Yannis, G., & Antoniou, C. (2002). Classification of driver-assistance systems according to their impact on road safety and traffic efficiency. *Transport Reviews*, *22*(2), 179–196. https://doi.org/10.1080/01441640110091215

Grimm, K. (2003). Software Technology in an Automotive Company-Major Challenges. *25th International Conference on Software Engineering, Proceedings. (pp. 498-503). IEEE. https://ieeexplore.ieee.org/document/1201228*

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659–3667. https://doi.org/10.1016/J.ESWA.2011.09.058

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A New Over-Sampling Method in *Imbalanced Data Sets Learning. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3644 LNCS*, 878–887. https://doi.org/10.1007/11538059_91

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language: Journal of Educational and Behavioral Statistics, 44(3), 348-361. https://doi.org/10.3102/1076998619832248

Henckaerts, R., Antonio, K., Clijsters, M., & Roel, V. (2017). A Data Driven Binning Strategy for the Construction of Insurance Tariff Classes. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3052174

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

John Waraniak, Michael McSweeney, Michael Kress, Brian Ellis, Alex Lybarger, Joshua Every, Blaine Ricketts, Darek Zook, David Karls, Jake Rodenroth, & Greg Potter. (2020). *Resource-Doc-Aug-15-ADAS-White-Paper*.

Kala, R. (2016). Advanced Driver Assistance Systems. *On-Road Intelligent Vehicles*, 59–82. https://doi.org/10.1016/B978-0-12-803729-4.00004-0

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/S13748-016-0094-0/TABLES/1

Kukkala, V. K., Tunnell, J., Pasricha, S., & Bradley, T. (2018). Advanced Driver-Assistance Systems: A Path Toward Autonomous Vehicles. *IEEE Consumer Electronics Magazine*, *7*(5), 18–25. https://doi.org/10.1109/MCE.2018.2828440

Lu, M., Wevers, K., & van der Heijden, R. (2005). Technical feasibility of advanced driver assistance systems (ADAS) for road traffic safety. *Transportation Planning and Technology*, *28*(3), 167–187. https://doi.org/10.1080/03081060500120282

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems 30.

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*, *34*(1), 87–107. https://doi.org/10.1016/J.IS.2008.04.003

Muthukrishnan, R., & Rohini, R. (2017). LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*, 18–20. https://doi.org/10.1109/ICACA.2016.7887916

Norris, F. H., Matthews, B. A., & Riad, J. K. (2000). Characterological, situational, and behavioral risk factors for motor vehicle accidents: a prospective examination. In *Accident Analysis and Prevention* (Vol. 32). https://doi.org/10.1016/S0001-4575(99)00068-8

Paefgen, J., Staake, T., & Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems*, *56*(1), 192–201. https://doi.org/10.1016/J.DSS.2013.06.001

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (Kouros). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, *136*, 105405. https://doi.org/10.1016/J.AAP.2019.105405

Pérez-Marín, A. M., & Guillen, M. (2019). Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. *Accident Analysis & Prevention*, *123*, 99–106. https://doi.org/10.1016/J.AAP.2018.11.005

Raschka, S. (2015). *Python machine learning : unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics :Packt Publishing Ltd*

Relatório Anual de Segurança Rodoviária. (2019). Autoridade Nacional de Segurança Rodoviária (ANSR)

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *Nternational Journal of Innovation and Scientific Research, 12(1), 217-222*.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378. https://doi.org/10.1016/J.PATCOG.2007.04.009

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. *Algorithms and Applications, 37.*

Tigadi, A., Gujanatti, R., & Gonchi, A. (2016). ADVANCED DRIVER ASSISTANCE SYSTEMS. *International Journal of Engineering Research and General Science*, *4*(3). www.ijergs.org

VanderPlas Jake. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc.

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(5), 1275–1304. https://doi.org/10.1111/RSSC.12283

Vorko-Jović, A., Kern, J., & Biloglav, Z. (2006). Risk factors in urban road traffic accidents. *Journal of Safety Research*, *37*(1), 93–98. https://doi.org/10.1016/J.JSR.2005.08.009

Weiss, G. M. (2004). Mining with rarity. *ACM SIGKDD Explorations Newsletter*, *6*(1), 7–19. https://doi.org/10.1145/1007730.1007734

Werner Geoff, Modlin Claudine, & Watson Willis. (2016). *Basic Ratemaking*. *Casualty Actuarial Society* (Vol. 4, pp. 1-320)

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (Vol. 1, Pp. 29-40)*.

7. APPENDIX

Variable	% Missing
District	0.49%
Urban_Rural_Areas	0.30%
Driver Age	0.30%
Vehicle Capital	5.2%
Total Premium	0.01%
Current Total Cost	98.60%
Current Total IBNR Cost	98.59%
Current Total Number of Claims	98.59%
Number of Claims Last 5 Years	22.09%
Number of Claims Last 3 Years	14.63%
Number of Claims Last 1 Year	5.62%

Table 6 – Variables with missing values

ADAS Feature	Equipments	Definition
COLLISION AVOIDANCE	Brake assistant	Detects circumstances in which emergency braking is required by measuring the speed with which the brake pedal is depressed and applies maximum braking pressure automatically.
	Electronically assisted brake	Ensures that the vehicle remains directionally stable and steerable even during emergency braking on slippery roads.
	Pre-collision system	Detects impending collision while traveling forward and alerts driver
	Proximity control	Detects pedestrians in front of vehicle and alerts driver to their presence
	Rear Traffic Alert Sensors	It is active once the vehicle is shifted into reverse. When backing, you will receive a visual or auditory warning if an approaching vehicle enters the rear cross traffic alert detection areas.

AUTOMATED PERFORMANCE ENHANCEMENT	ABS anti-lock braking system	Detects potential collisions while traveling forward and automatically applies brakes to avoid or lessen the severity of impact.
ADVANCED CRUISE CONTROL	Adaptive Cruise Control	Controls acceleration and/or braking to maintain a prescribed distance between it and a vehicle in front. May be able to come to a stop and continue
	Cruise control	System which can maintain the speed of a car at a desired level. The standard cruise control systems can take over the throttle once the driver activates Cruise Control and sets the desired speed.
LATERAL COLLISION AVOIDANCE	Lane crossing warning system	Monitors vehicle's position within driving lane and alerts driver as the vehicle approaches or crosses lane markers.
	Dead angle assistant	Detects vehicles to rear in adjacent lanes while driving and alerts driver to their presence
PARKING ASSISTANCE	Front parking sensor	Proximity sensors on front of vehicles designed to alert the driver of obstacles while parking. These systems use either electromagnetic or ultrasonic sensors.
	Rear parking sensor	Proximity sensors which alarm the driver when the vehicle gets too close to an object.
	Parking camera	Front or/and rear parking camera
	Parking Assistant	The system is designed to help you park alongside a detected vehicle or vehicles. The system will then automatically move your vehicle into the spot.
	360º camera	Uses cameras located around vehicle to present view of surroundings
DRIVER VISION AUGMENTATION	Night vision system	A system that aids driver vision at night by projecting enhanced images on instrument cluster or heads-up display.
	Fatigue Detection System	Detects pedestrians in front of vehicle and alerts driver to their presence.
	Traffic Signal Recognition	Ensures that the current speed limit and other road signs are displayed to the driver on an ongoing basis
TIRE PRESSURE CONTROL	Active tire pressure control	Uses a sensor mounted in the wheel to measure air pressure in each tire. When air pressure drops 25% below the manufacturer’s recommended level, the sensor

		transmits that information to your car’s computer system and triggers your dashboard indicator light.
	Passive tire pressure control	It works with car's Antilock Braking System's (ABS) wheel speed sensors. If a tire's pressure is low, it will roll at a different wheel speed than the other tires. This information is detected by your car's computer system, which triggers the dashboard indicator light.
	Automatic tire pressure control	Compensates for fluctuations in the air pressure and temperature of vehicle tires and automatically adjusts the tire pressure.

Table 7 – Description of ADAS equipment per created groups

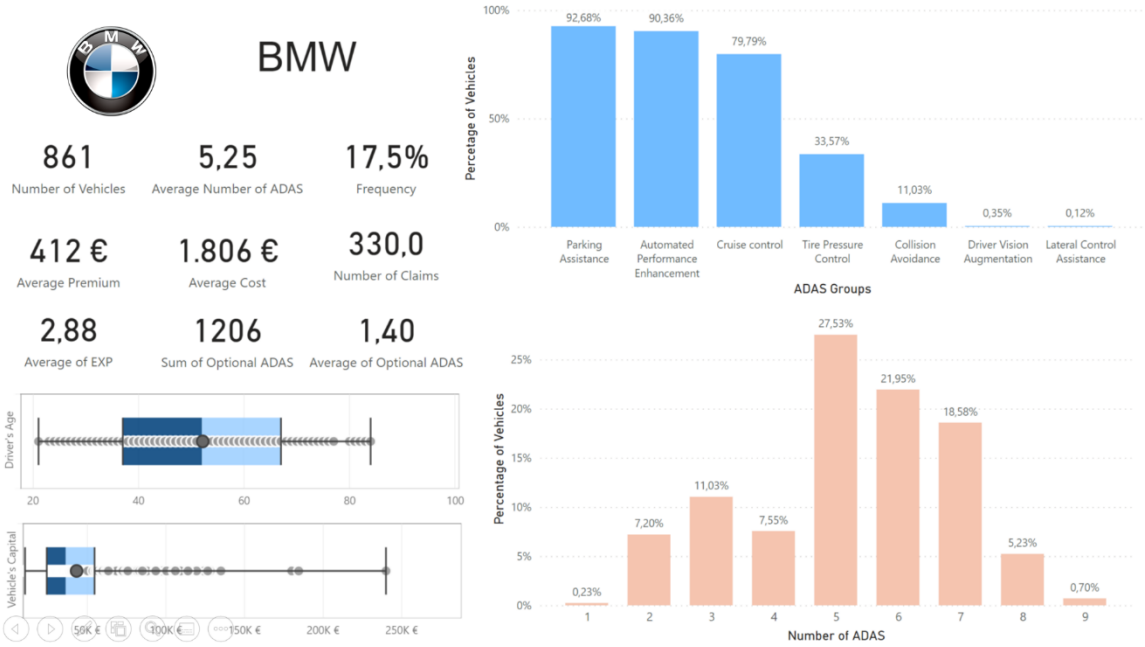

Figure 17 – Example of descriptive analysis by brand

A descriptive analysis was created with the vision for each of the brands in the sample chosen for analysis. As shown in Figure 17 for the BMW brand, a set of the main metrics and some graphs related to the policy are presented, namely the age of the driver and vehicle capital, as well as the percentage of vehicles by ADAS groups and number of ADAS. The creation of these analyzes aims to help the Marketing and Pricing teams so that it is possible to have a more detailed notion of each brand.

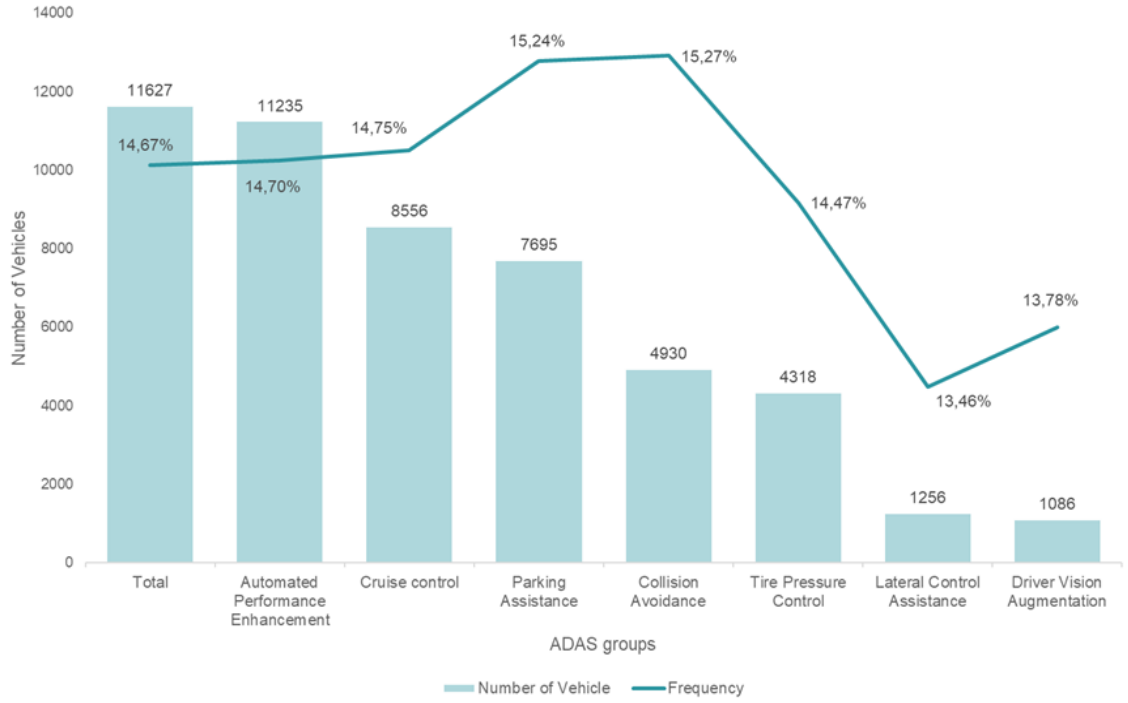

Figure 18 – Number of vehicles and frequency per ADAS groups

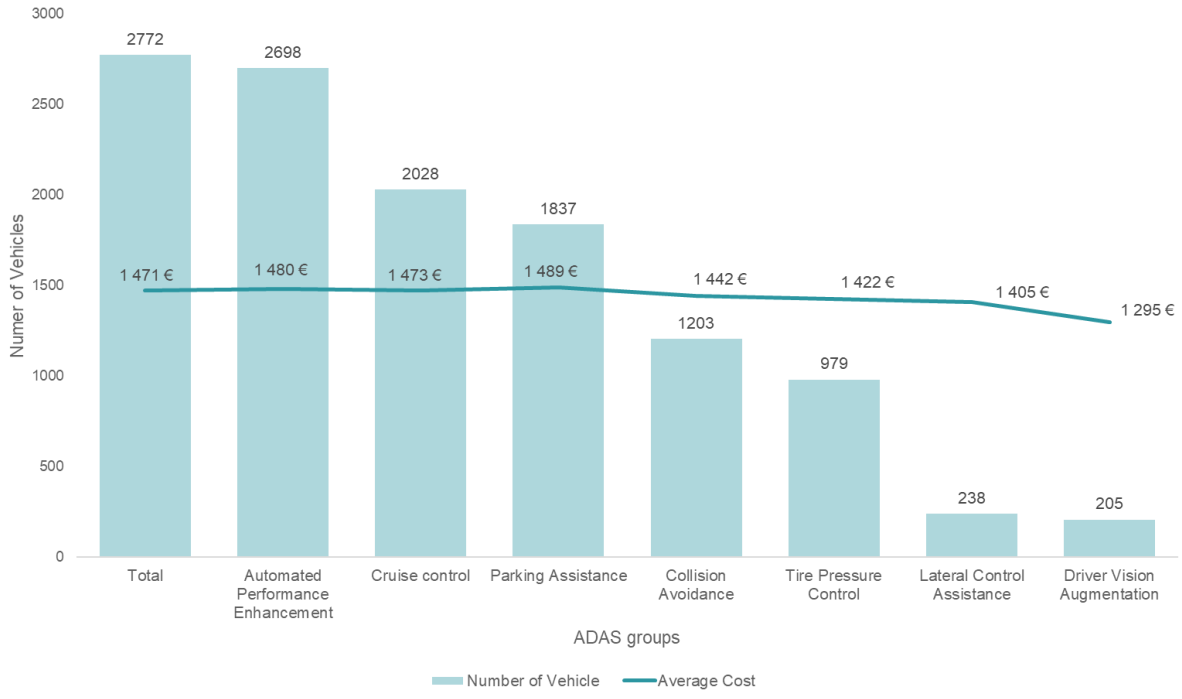

Figure 19 - Number of vehicles and severity per ADAS groups

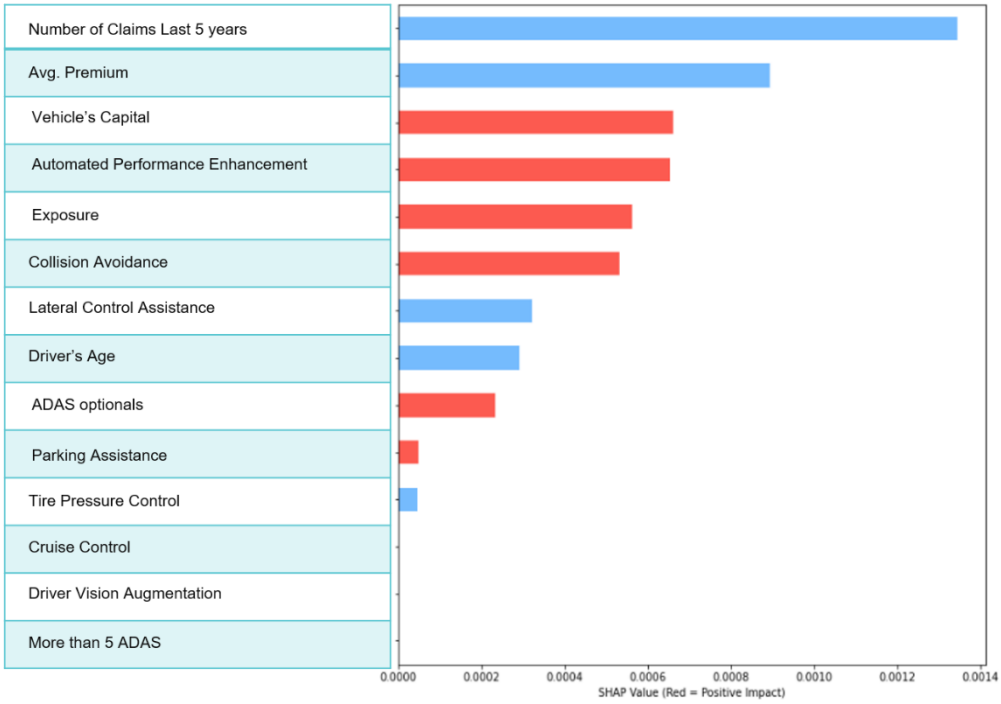

Figure 20 – Simplified SHAP values for Severity Low Cost class

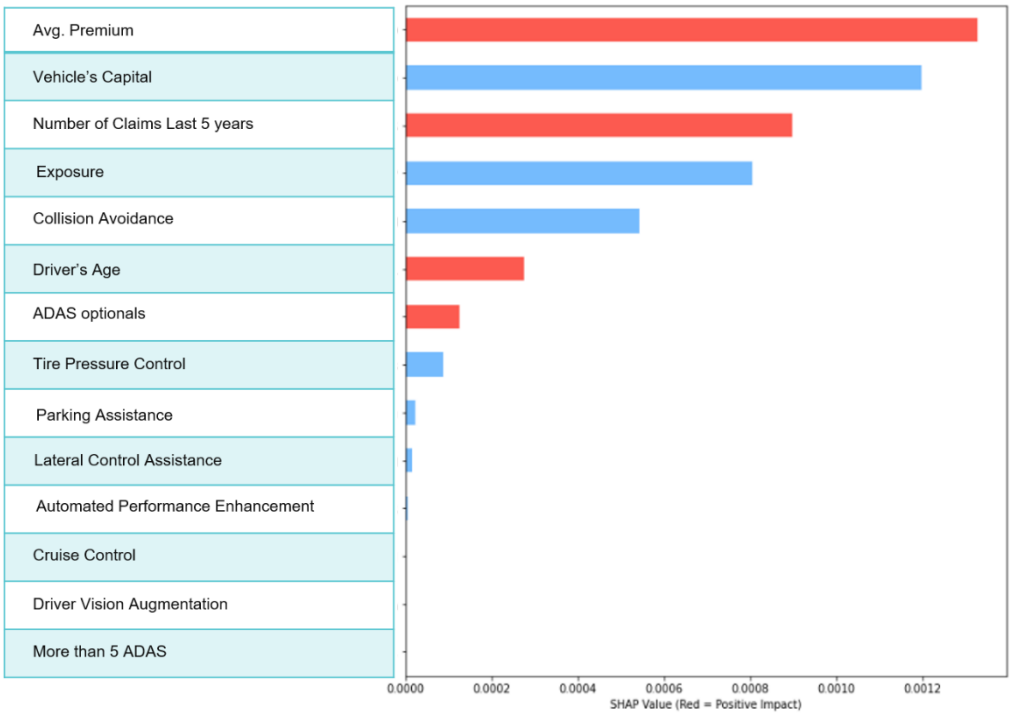

Figure 21 - Simplified SHAP values for Severity Medium Cost class

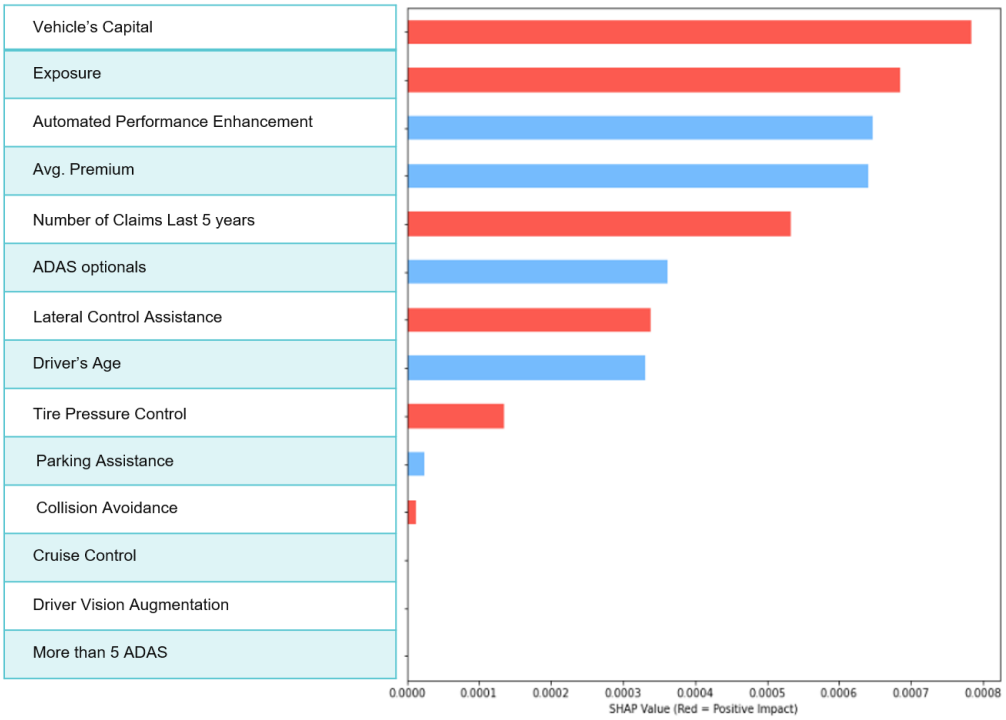

Figure 22 - Simplified SHAP values for Severity High Cost class

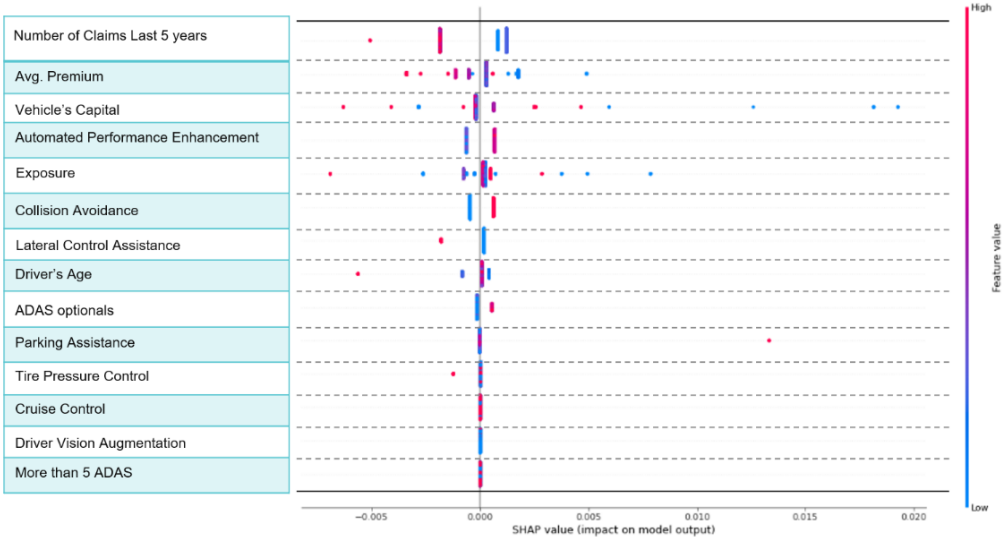

Figure 23 - SHAP summary plot for Severity Low Cost class

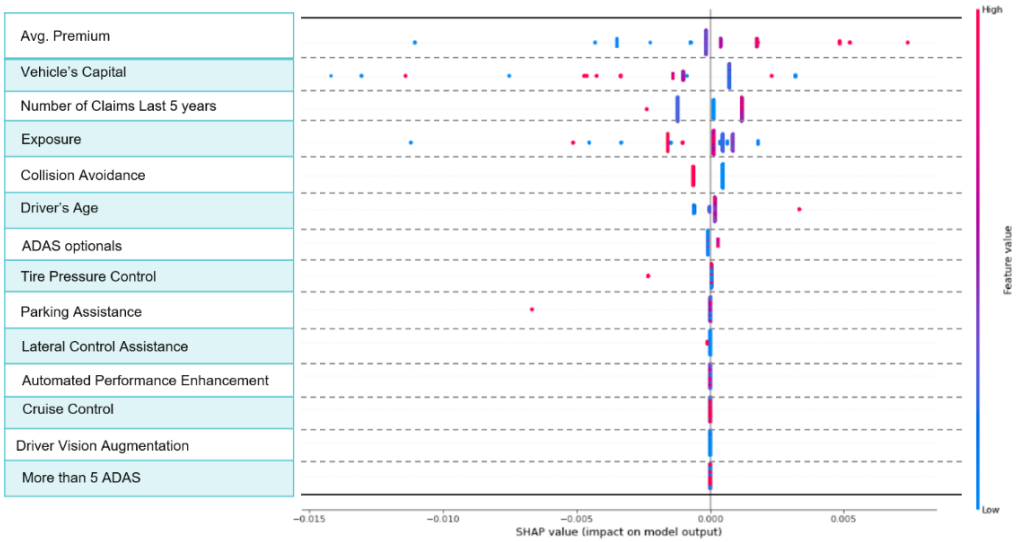

Figure 24 - SHAP summary plot for Severity Medium Cost class

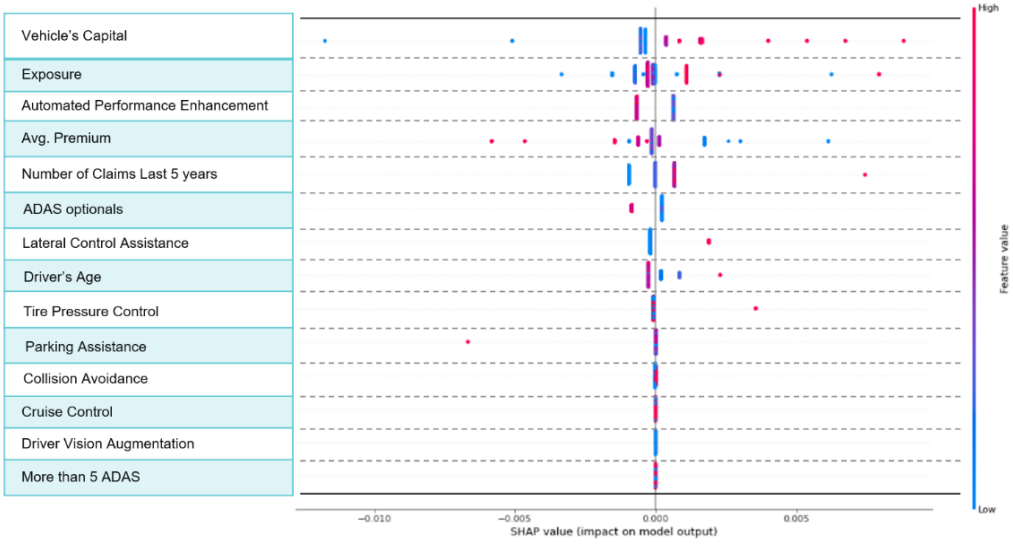

Figure 25 - SHAP summary plot for Severity High Cost class