NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Automobile Usage-Based-Insurance

Improving Risk Assessment measured through telematics

Lourenço Manuel Coelho Santiago Violante da Cunha

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

AUTOMOBILE USAGE-BASED-INSURANCE: IMPROVING RISK ASSESSMENT MEASURED THROUGH TELEMATICS

by

Lourenço Manuel Coelho Santiago Violante da Cunha

Advisor: Professor Doutor. Jorge Miguel Ventura Bravo

ABSTRACT

Accurate risk estimation with proportionate fees is the cornerstone of insurance activity, a billion-dollar service industry. Due to progressive technological development, insurance companies are now able to improve their risk assessment in the underwriting process of automobile insurance. Through the installation of Onboard-diagnostic devices or with an application in the customers' smartphones, insurance companies may measure behavioral and situational risk factors such as distance covered and driving habits. These new risk factors provide further information that helps the client's risk evaluation beyond the traditional risk factors of customer and car specific. The objective of this research is to measure the increased prediction capacity of the claim predicting model by including driver behavior variables. A Generalized Linear model was applied, that includes not only the traditional risk factors, but also additional risk factors measured by telematics, and a new model-based ensemble predictor to a dataset with more than 3 million drivers. Results show that the incorporation of driver's behavior variables increases the overall capacity of the model. By adding these behavioral risk factors, the actuarial accuracy is increased, leading to a more tailored approach of risk assessment and also awarding the clients that have safer conduct on the road and penalizing those with a more hazardous behavior prone to incurring in car accidents.

KEYWORDS

Usage-Based-Insurance; Pay-As-You-Drive; Pay-How-You-Drive; Risk Assessment; Premium Calculation

INDEX

1. INTRODUCTION

Motor insurance is one of the most important products in Non-life insurance with 77 968.5 million euros of gross written premiums in the European Union alone, during 2018. Being a strategical and continuously growing line of business in Non-life insurance, there is a need for companies to differentiate their products and gain a competitive advantage. In Sweden, pay-as-you-drive solutions have entered the market, and in Italy, the use of vehicle-installed devices has increased by 22.2% (European Insurance and Occupational Pensions Authority., 2019), to face the high incidence of fraud in the southern part of the country.

The pricing of automotive insurance has been set upon the same immutable principles and risk factors for the past decades. Currently it continues to be based upon the prediction of severity and frequency of future claims (Denuit et al., 2007), although there have been innovations. The dependent variables considered to assess the risk of frequency of claims have been mostly independent of the client's use. Risk factors such as age, driving experience, marital status, and location of residence are some of the variables that contribute to the risk profile of the driver while engine dimensions, horsepower, and the type of vehicle are the main features used for the profiling of the vehicle (Azzopardi & Cortis, 2013).

The first attempts of usage-based insurance in motor liability consisted of pay-as-you-drive schemes, where the consumers would report the distance covered during the policy term to the insurance company. This method was flawed due to the inconsistencies registered in the odometer readings (Tselentis et al., 2017). Technological development and the arrival of Insurtech companies allowed the insurance industry to improve its risk assessment models through the use of new risk factors measured by telematics (Lewis, 2017). By installing electronic devices on the client's vehicle, the insurance company can accurately monitor the annual distance covered, since it has been shown that it is a significant variable to be considered in pricing (Ayuso et al., 2019; Ferreira & Minikel, 2012; Lemaire et al., 2015). This type of insurance is called Usage-Based-Insurance (UBI), where the premium paid by the customer is based on its behavior and not on a lump sum amount (Figure 1).

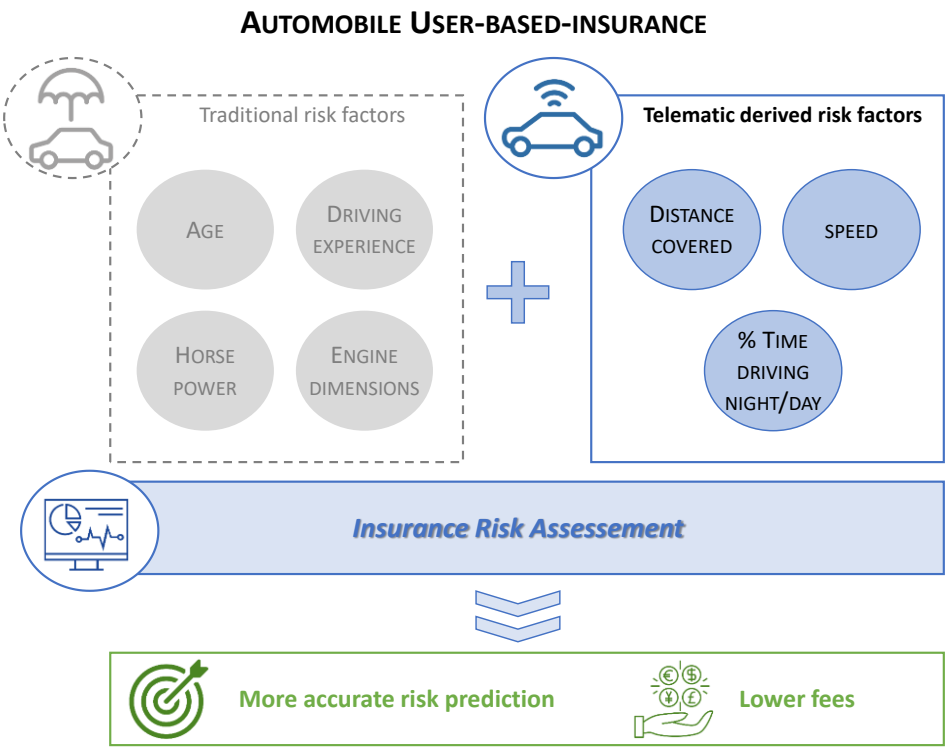

Figure 1: Schematic overview of Automobile User-Based-Insurance.

Not only the distance covered during the policy term, but other driver dependent variables (percentage of night driving and the travel speed) can be considered due to technological development that further increase the accuracy of the client's risk assessment (Ayuso et al., 2019; Baecke & Bocca, 2017).

By introducing these variables in their models, insurance companies protect themselves from risk imbalances through the correction of their pricing and gain a competitive advantage in the market. Additionally, customers may also benefit from UBI by paying a fairer premium according to their use. Furthermore, an incentive to drive safely is added, to be charged lower premiums and indirectly a societal benefit of safer roads.

New models based on machine learning techniques are being tested to predict clients' claims. Traditional methods of generalized linear regression are being compared to the new ensemble methods of random forests and neural networks to obtain the most accurate model. However, the ensemble-based models face regulatory approval complications, (Baecke & Bocca, 2017), which limit their applicability.

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The motor insurance line of business is one of the largest in Non-life insurance accounting for approximately 548 524 million dollars in 2019 of gross written premiums, being the United States of America the biggest market, with 299 536 million dollars of gross written premiums (OECD, 2019). The relevance of this line of business comes from the fact that there are now more than 1.32 billion cars on the road worldwide (Wards Intelligence, 2017) and in most countries, it is mandatory for each vehicle to be insured. The first country to make motor insurance mandatory was the United Kingdom with the Road Traffic Act 1930 (*Road Traffic Act*, 1930) followed by Germany in 1939 with the Act on the Implementation of Compulsory Insurance for Motor Vehicle Owners (*Act On Compulsory Insurance For Motorists*, 1939).

Risk classification is the cornerstone of the insurance business, where each client's risk is assessed individually and classified into similar homogeneous risk groups. Traditionally, this assessment is done only at the moment of inception of the contract, based upon the characteristics of the already existing pool of clients, and later updated regarding the occurrence or not of reported claims.

At the moment of inception, risk evaluation is done by considering variables such as age, driving experience, sex, historical law enforcement penalties and technical characteristics of the vehicle. Later on, if a claim at fault is reported, the initial risk assessment is updated based on the new information through a Bonus / Malus System. In the case of a claim at fault, the client's risk is updated and an increase in the premium may be done to compensate for the incremental risk faced by the insurance company (Malus). If there were no claims reported, the insurance company may adjust the premium, by giving the client a discount due to the lower risk incurred (Bonus).

The risk factors considered at the inception of the policy have been subject to change by regulatory obligations. Despite women drivers being more prone to incur a claim than male drivers (Aseervatham et al., 2016), the gender of the individual cannot be used for pricing, as it has been deemed discriminatory by the European Court of Justice since 2012 (*EU Rules on Gender-Neutral Pricing in Insurance Industry Enter into Force*, 2012).

This ratemaking process does not provide an equitable insurance premium for all clients, there is a mutualization (risk compensation, cross-subsidy) effect between the riskier and the safer drivers within the same risk pool. As it does not consider the use of the vehicle, the distance covered, or even the style of driving of each client (Bian et al., 2018). For example, two individuals of the same age, one driving 10 000 km/year and the other 30 000 km/year could possibly be paying the same premium.

This idea of UBI means changing how auto insurance is sold from a vehicle-year risk exposure to vehicle-kilometer-year/time (Litman, 1997) risk exposure, being a more tailored approach to each client's risk. Costumers would pay a premium based on the traditional factors mentioned combined with the distance covered during the policy term, known as Pay-as-you-drive (PAYD), and not a fixed lump sum amount. However, the initial attempt showed flaws because of the lack of technological development, insurance companies could not get reliable readings of the client's vehicle. The inability to store and analyze big datasets and the high cost of real-time data recording, data programs, and computing services hindered the development of such products (de Romph, 2013; Lee, 2014)

Owing to technological development and the appearance of Insurtech companies, there are currently new ways, until now unavailable, for the insurance industry to measure risk factors (Lewis, 2017). There is no doubt that distance covered is a significant risk variable to be considered, (Lemaire et al., 2015), and now it is possible to measure it reliably through telematics. These electronic devices can be either installed in the On-board-diagnostics (OBD) of the vehicle or by an app in the smartphone of the user (Händel et al., 2014).

Advances in European regulation help the implementation of this type of motor insurance as, in 2015, the European Parliament introduced a new regulation, where it is stated that vehicles produced in the European Union would have in-built telematics devices that would call the emergency services in case of an accident had occurred (*Regulation 2015/758*, 2015). Due to this, more vehicles will already have OBD devices in the future, making it easier for insurance companies to develop and implement this type of insurance. Furthermore, it has been shown that insurance companies can use this information without privacy leaks (Troncoso et al., 2011).

Further research (Baecke & Bocca, 2017; Lemaire et al., 2015) has shown that by adding these new risk factors, the prediction of the frequency of claims model is further improved than by the substitution of the traditional risk factors by those measured by telematics (Baecke & Bocca, 2017; Lemaire et al., 2015). An improvement of 3.58 percentage points (from an AUC of 57.7% to 60.8%) was registered with a logistic regression model by (Baecke & Bocca, 2017). Additional risk factors of driving behavior have been studied in transportation research (Ayuso et al., 2014; Ellison et al., 2015), where it was concluded that night driving, speeding, and the type of roads driven influence the risk of incurring an accident by using a dataset with nearly 16 000 drivers and applying a Weibull regression model (Ayuso et al., 2014)and with a sample of 148 drivers using Multilevel models, (Ellison et al., 2015)

It has been previously possible to introduce new variables measured by telematics in already existent models for predicting claim frequency. Notwithstanding, there are also new models of risk prediction being developed with the use of machine learning techniques (Random Forests and Artificial Neural Networks). Using these techniques, (Baecke & Bocca, 2017) have shown improvements on the traditional regression models. The model based on an artificial neural network (ANN) outperformed the logistic regression from an AUC of 60.83% to 61.74%.

In 2019, the European Insurance and Occupational Pensions Authority (EIOPA) published a thematic review on “Big Data Analytics in Motor and Health Insurance”, where 222 insurance firms from 28 Member States, 24 National Competent Authorities and 2 National Consumer Associations were involved. The insurance undertakings of this report represent 60% of the total gross written premiums of the motor and health insurance lines of businesses. Big Data Analytics (including Artifical Intelligence and Machine Learning) has had the biggest impact on the pricing and underwriting stage of the insurance business because it enables companies to better understand consumer’s needs and characteristics allowing a more personalized product development, (European Insurance and Occupational Pensions Authority, 2019).

The lack of interpretability of “black-box” models was one of the challenges pointed out since they rely on historical data which might have inherent historical biases (societal or ethical) that are reflected in the output. Additionally, if the algorithms are based upon biased datasets or rating factors, the results could create illegal price discrimination if not monitored with the adequate due diligence

and with the accepted actuarial principles. Furthermore, an algorithm that is not sufficiently explainable, transparent or auditable could risk the overall solvency position of the insurance company. The incorrect calculation of premiums, due to the “black-box” method, combined with a lack of internal controls might put in jeopardy the undertaking’s solvency, (European Insurance and Occupational Pensions Authority, 2019).

As a follow up from the thematic review, EIOPA will further assess the issue of supervision of “black-box” algorithms and how they can be monitored in practice and how it is different from other well established insurance models, such as the GLM, (European Insurance and Occupational Pensions Authority, 2019).

As shown in the thematic review, regulators prefer the use of regression models, considered to be “white-box” models where the impact of the included variables can be explained, something that the “black-box” fail to do, (Baecke & Bocca, 2017). The interpretation of the impact of each variable in these models can be made by a sensitivity analysis only, however the monitorization of the “black-box” models is under active development (European Insurance and Occupational Pensions Authority, 2019) .

1.2. STUDY OBJECTIVES AND RELEVANCE

The objective of this study is to develop a model that contributes to the transition from a PAYD pricing scheme to a Pay-how-you drive (PHYD) scheme. The difference between PAYD and PHYD is the fact that the latter, besides including the traditional risks factors and distance covered, also includes behavioral risk factors (Vaia et al., 2012). The frequency predicting model of claims will be developed through a Generalized Linear Regression Model by combining traditional and behavioral risk factors. The model will be tested in its ability to optimize the prediction capacity of pricing models and assess the improvement made by considering these new risk factors.

The second objective of this thesis is to propose a new model for predicting claims combining traditional regression models and ensemble methods, GLM Bagging. This new method aims to offer the same interpretability of the regression models and provide the benefit of ensemble methods, without having the limitations of regulatory approval.

There have been studies on PHYD insurance, (Ayuso et al., 2019), where the model developed included behavioral risk factors as the percentage of kilometers driven at night and the percentage of kilometers driven over the speed limit. This study shows a clear indication of the increased actuarial accuracy in the prediction of claims by using these risk factors for correction of the traditional models. Similar conclusions were gathered by (Baecke & Bocca, 2017) here a model of PHYD insurance was developed, with data collected from a European car insurance company with information about 6 984 customers.

There are already UBI products on the market where the customers are charged a premium based on the traditional methods, considering their age, driving experience and the characteristics of the vehicle combined with its use, as presented in Table 1 below (Husnjak et al., 2015).

Table 1: European Auto Usage-Based-Insurance (adapted from (Husnjak et al., 2015)).

Insurance Company	Country	Name of the UBI program	Insurance Concept	Technology platform	Data transmission
AXA	Italy	Autometrica	Distance-based insurance	GPS*-based	Mobile data service
Generali	Italy	Protezione Satelitare	Traditional telematics parameters	GPS*-based	Mobile data service
AXA Winterthur	Switzerland	Crash Recorder	Recording events	Event-data recorder	Event-data recorder
MAPFRE	Spain	YCAR	Traditional telematics with several risk levels within 24 hour period	GPS*-based	Mobile data service
RSA Insurance Group	United Kingdom	More than Green Wheels Insurance	Traditional telematics with several risk levels within 24 hour period	GPS*-based	Mobile data service

WGV	Germany	Young & Safe	Traditional telematics with several risk levels within 24 hour period	GPS*-based	Mobile data service

*GPS: Global Positioning System

One step further would be to develop a model that includes more risk factors that indicate driving behavior. Examples include the distance driven on highways or local roads, being highways considered safer, the distance driven on certain days of the week and the addition of breaking and acceleration data (Baecke & Bocca, 2017).

By considering these behavioral risk surrogates, insurance companies can assess the risk of each driver incurring an accident in a more tailored way. Additionally, through the surcharge of reckless driving on premiums, clients are incentivized to adapt their driving behaviour to directly benefit from discounts and indirectly decrease the overall risk of an accident, which is beneficial for both the insurance company and the client.

UBI is becoming more relevant in the insurance market where insurance companies are looking for differentiation and customers are expecting more affordable solutions (Litman, 2005). There are already well-established insurance companies with these types of insurance solutions like UnipolSai Assicurazioni S.p.A (Italy), which was the main data integrator player on the market in collaboration with Octo Telematics (Vaia et al., 2012).

The development of better UBI solutions such as PHYD insurance increases the actuarial accuracy of the risk measurement of clients. Additionally, insurance coverage becomes more affordable since customers only pay for their use of the vehicle and thus are given the possibility of reducing the premiums paid by reducing the distance driven (Litman, 2005).

For the porpuse of development of an UBI solution, regression models will be developed on a dataset provided by the Massachusetts Executive Office of Energy and Environment with information about 3 million drivers, including claims and earned exposure during 2006. The risk assessment of the drivers will be based on geographic and class risk groups as well as each driver's annual mileage.

The overall predicting capacity of the model was increased by the introduction of distance convered. These new types of behavioural variables have a biggest impact when used together with the traditional variables than as an new independent approach. These empirical findings highlight the importance of behavioural risk factors in the client's risk assessment. This is an untapped potential since both parties from an insurance contract have additional benefits. The insurance company has a more accurate risk assessment of it's clients and gain product differentiation in a already highly competitive line of business, as well as, the insured has a coverage more tailored to their needs in terms of protection and of premiums paid.

The structure of this thesis is as follows: in the first section “Background and Problem Identification” a discussion of the state of the art and previous research is presented. The second section “Study Objectives and Relevance” is where the study goals are enumerated and the intended contribution of this research is presented. The following sections are the “Methodology”, where a description of the procedures performed is presented, and the “Data” that will be used to reach said objectives, followed by the results of our empirical valuation. Finally, we highlight the conclusions and limitations encountered during the development of the investigation.

2. METHODOLOGY

The background information gathered to understand and summarize the current evolution of pricing automobile insurance practices and the challenges faced by the industry was searched in Google Scholar research engine with the keywords ("Auto Usage-based-insurance"; "Pay-as-you-drive"; "Telematics"; "Insurtech") for papers written from inception to November 2021.

The common method to price a motor insurance policy is a model considering both frequency and magnitude of claims. The focus of this paper will be to develop a frequency predicting model and sets upon the assumption that the model for predicting the monetary cost (magnitude) of the claims is obtained independently. As a standard practice in the insurance industry, the Generalized Linear Model (GLM) is used for the prediction of claim frequency, and therefore it will be the regression model used in this study.

Generalized linear models are set upon the assumptions that the explanatory variable is independently distributed, and the dependent variables follow distributions from the exponential family (binomial, poisson, multinomial or normal). Additionally, they assume a linear relationship between the transformed response in terms of link function and the explanatory variables. The estimation process is based on Maximum Likelihood Estimation therefore it relies on large sample approximations (McCullagh & Nelder, 2019).

A Generalized Linear Model is structured by 3 components: a random component which refers to the probability density function of the response variable *Y*, a systemic component that specifies the explanatory variables in the model (which can be continuous, discrete or both) and a link function that specifies the link between the random and the systematic components. The link function of *E(Y)* that the model equates to the systematic component (Agresti, 2003).

These models are an advantage in relation to the Ordinary Least Square regression models because there is no need of transformation of the response variable *Y* to have a normal distribution, the choice of the link function is distinct from the random component which gives an additional flexibility in modelling and the models are fitted with the Maximum Likelihood Estimaton, thus resulting in the optimal properties of the estimators (McCullagh & Nelder, 2019)

In this case, the response variable (*Y*) will be binary, indicating the risk of incurring in a claim and the explanatory variables will be the risk factors present in the data set. Extending the framework of normal linear models to the class of distributions derived from the exponential. It can be expressed as:

$$f(y) = \exp\left[\frac{y\theta - \psi(\theta)}{\phi} + c(Y, \phi)\right] \quad (1)$$

Where $\psi(.)$ and $c(.,.)$ are known functions, θ and ϕ are the natural and scale parameters, respectively. Regarding the link function, as our purpose is to identify if the insured incurred in a claim or not, a binary result, the link function used in the model will be the Binomial distribution.

The data will be divided into a test sample, that contains 30% of the population data, and a training sample with the remaining 70%. The training sample will be where the model is developed, and its performance will be measured on the validation sample by using the Akaike Information Criterion (AIC). The AIC is an estimator of the prediction error computed as twice the number of parameters in the model minus twice the value of the log-likelihood in the maximum given an observed sample. It estimates the relative information lost by the model, and therefore the lower the AIC, the higher the quality of the model. The resulting models will be built based upon the stepwise variable selection process, where it was subsequently added and removed predictor variables, in order to find the best performing model, lower AIC (Williams et al., 2015). The choice of this performance measurements was to allow for direct comparability with other model results. All the statistical analyses were performed using the R Studio statistical package version 1.4.1717.

Additionally, in the construction of the model, the "Annual Mileage" variable will be transformed using two methods used in Statistics: Min Max scaling method where the values of the variable are subject to a linear transformation to fit the interval between 0 and 1 and Z-score standardization method where the values are normalized based on the mean and standard deviation, (Gopal et al., 2015). This transformation was performed to reduce the high standard deviation of the original values.

Furthermore, these types of regression models are flawed in the aspect that the resulting prediction is based upon the original random sample used for its development. New methodologies have been tested to solve this issue, such as ensemble methods like random forests, decision trees and neural networks. However, these methods are of difficult interpretation, and the regulators require the use of "white box" models such as the GLM, (Baecke & Bocca, 2017).

To possibly unravel this issue, we will perform an ensemble technique that has the advantage of having the interpretability of the GLM models and the randomness of the ensemble models, called Bagging (Bootstrap Aggregation), (Breiman, 1996). Bagging is an ensemble method based on the repetition of a number of samples with replacement, where for each sample a GLM model is built. The final model is the result of the average of the models' coefficients from several samples. Due to the repetition process, the risk of building a biased model on a unique sample is mitigated, especially in a zero-inflated data set (datasets where the event in the analysis is considered rare). For this study, we will perform a Bagging GLM model based on 100 resamples. Its performance will then be compared to the GLM's based on the performance measure mentioned previously.

3. DATA

We had originally planned to measure the impact of telematics measured variables in the model predicting capacity of a claim as described in the "Study Objectives and Relevance" section. With this in mind, several insurance companies were contacted along with the Portuguese insurance regulator (Autoridade de Supervisão de Seguros e Fundos de Pensões), the Portuguese road safety association (Autoridade Nacional de Segurança Rodoviária) and the Portuguese Insurer association (Associação Portuguesa de Seguros). However, due to the lack of information available, confidentiality agreements, the General Data Protection Regulation law (*Regulation (EU) 2016/679*, 2016) and the proprietary ownership of the data, no anonymized database with the telematic variables was obtained.

Therefore, this study uses a public data set provided by the Massachusetts Executive Office of Energy and Environmental Affairs in 2010, to assess the impact that mileage has on the classical models of prediction. It contains insurance policy and claims information gathered by the Commonwealth Automobile Reinsurers and mileage readings collected by the Massachusetts Registry of Motor Vehicles. This data set was originally used for a report (Joseph & Minikel, 2010) prepared for the Conservation Law Foundation, and is available online (http://mit.edu/jf/www/payd).

To process the information into a complete dataset possible for analysis, it was necessary to link the claims data set with the policy data set. The former included the total paid losses and outstanding reserves, for all claims that matched an earned exposure period during the 2006-year policy. The latest included the earned exposure months and annual mileage estimated for each period of consistent policy endorsement conditions during the same policy year. This information was matched to 2 risk matrixes. One regarding the town where the vehicle was most likely to be garaged ("*Town Risk*"), ranging from "1" being the town with the least risk to "6" being the riskiest. A second risk matrix, classifying the use of the vehicle and the driver's experience was also employed. The R Script for the data preparation is presented in Appendix I.

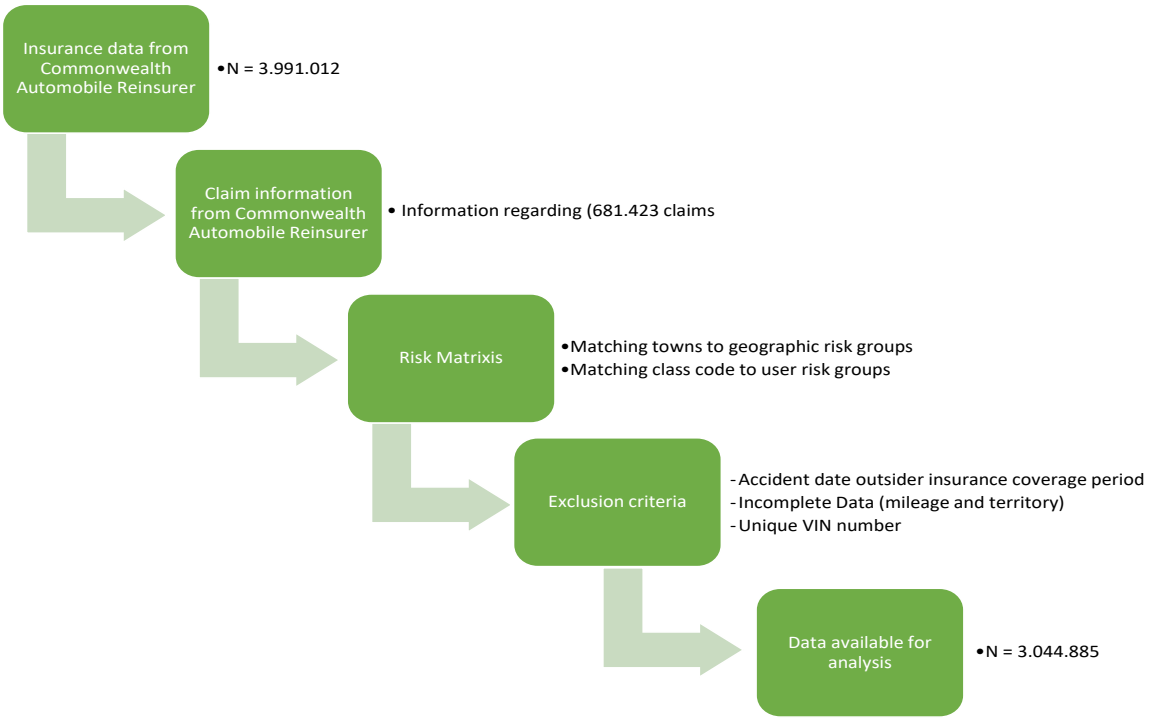

Figure 2: Data Processing Flow Diagram.

The Model will include 3 predicting variables as shown in Table 2. The explanatory variables include both traditional rating factors such as *Town Risk* (a categorical variable that represents the risk according to the accident rates and insurance rates of the town where the vehicle is garaged) and *Driver Class* (measuring the purpose of the vehicle and driver experience). Additionally, *Mileage,* a quantitative variable that indicates the estimate of annual miles traveled by the vehicle based upon the odometer readings in RMV safety inspections is used.

Table 2: Ratemaking variable description.

Ratemaking Variable	**Description**
Town Risk	Categorical variable that classifies the risk of the town where the vehicle is mainly garaged from a scale of 1 to 6.
Driver Class	Categorical variable that classifies the use and experience of the driver as: - Adult; - Business; - Occasional driver with 3 to 6 years of experience; - Occasional driver with less than 3 years of experience with driving training; - Occasional driver with less than 3 years of experience with no driving training; - Principal driver with less than 3 years of experience with driving training; - Principal driver with less than 3 years of experience with no driving training; - Senior citizen;
Annual Mileage	Quantitative variable of the estimate annual miles travelled by the vehicle based upon odometer readings in RMV safety inspection.

The descriptive statistics, presented in Tables 3 and 4, identify the differences between drivers without claims and with claims for these variables.

Table 3: Descriptive Statistics – "*Town Risk*" Variable.

Town Risk	All Sample		Drivers with no claim		Drivers with claim		Claim Ratio
	n (in k)	%	n (in k)	%	n (in k)	%	
1	578	19%	559	19%	19	14%	3.3%
2	592	19%	570	20%	22	16%	3.8%
3	343	11%	329	11%	14	10%	4.0%
4	616	20%	589	20%	27	20%	4.4%
5	568	19%	538	18%	30	22%	5.3%
6	348	11%	324	11%	24	18%	6.9%
Total	3,045	100%	2,908	100%	136	100%	4.5%

Source: Adapted from http://mit.edu/jf/www/payd

Table 4: Descriptive Statistics – "*Class*" Variable.

Driver Class	All Sample		Drivers with no claim		Drivers with claim		Claim Ratio
	n (in k)	%	n (in k)	%	n (in k)	%	
Adults	2,332	77%	2,236	77%	96	70%	4.1%
Business	41	1%	39	1%	2	2%	5.1%
Occasional w/ 3-6yr exp	113	4%	104	4%	9	6%	7.6%
Occasional w/ <3yr exp, driver training	34	1%	30	1%	4	3%	10.8%
Occasional w/ <3yr exp, no driver training	4	0%	4	0%	1	0%	13.5%
Principal w/ <3yr exp, driver training	44	1%	39	1%	5	4%	11.3%
Principal w/ <3yr exp, no driver training	13	0%	11	0%	2	1%	13.5%
Senior citizens	465	15%	445	15%	19	14%	4.2%
Total	3,045	100%	2,908	100%	136	100%	4.5%

Source: Adapted from http://mit.edu/jf/www/payd

Table 5: Descriptive Statistics – "*Annual Mileage*" Variable.

	All Sample		Drivers with no claim		Drivers with claim	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Annual Mileage	11,803	7,805	11,742	7,791	13,109	7,969

Source: Adapted from http://mit.edu/jf/www/payd

The database under analysis has an overall claim ratio of 4.5%, out of the 3,044,885 drivers under analysis, 136,401 have reported a claim. As shown in Table 3, *Town Risk 6* has the highest claim rate of 6.9% and *Town Risk 1* the lowest with 3.3%, as expected. The observations are evenly spread through the *Town Risk* matrix, having approximately 20% each of the overall sample except for Tow Risks 3 and 6, with a slightly lower percentage, 11%.

Regarding the *Driver Class* risk factor, the analyzed data is mostly made of Adults, 77%, followed by Senior Citizens, 15%, and Occasional Drivers with 3 to 6 years of experience, 4%. The Driver Class with the highest claim ratio is the Occasional driver with less than 3 years experience and no driver training and Principal drivers with less than 3 years experience and no driver training, both with 13.5%, followed by the Principal driver with less than 3 years experience with training, 11.3%.

The mean annual mileage is 11 803 miles (18 996 kilometers) while the mean annual mileage for drivers with reported claims, 13 109 miles (21 096 kilometers) is higher than those without claims, 11 742 miles (18 897 kilometers). The standard deviation of this variable is 7 805 miles (12 560 kilometers) for the complete sample, without big deviations between drivers with and without claims, meaning that is variable is highly dispersed, as shown in Figure 3. Additionally, a Kolmogorov-Smirnov test was performed and the distribution of Mileage is not statistically different from the normal distribution (p-value< 2.2E-16).

The impacts of the Z-score transformation and the Minmax transformation are shown below in Figures 4 and 5. The first shows the values normalized based on the mean and standard deviation, with the highest distance from the mean of 11.3 standard deviations and the Minmax transformation where the values are fitted in the interval between 0 and 1.

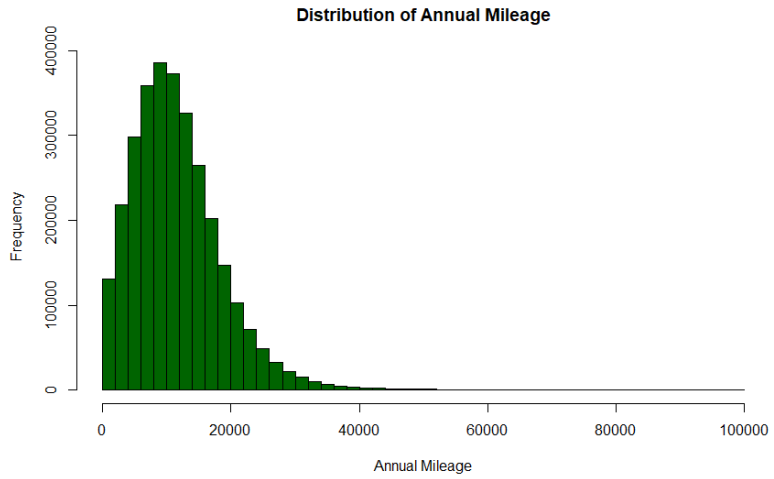

Figure 3: Distribution of Annual Mileage.

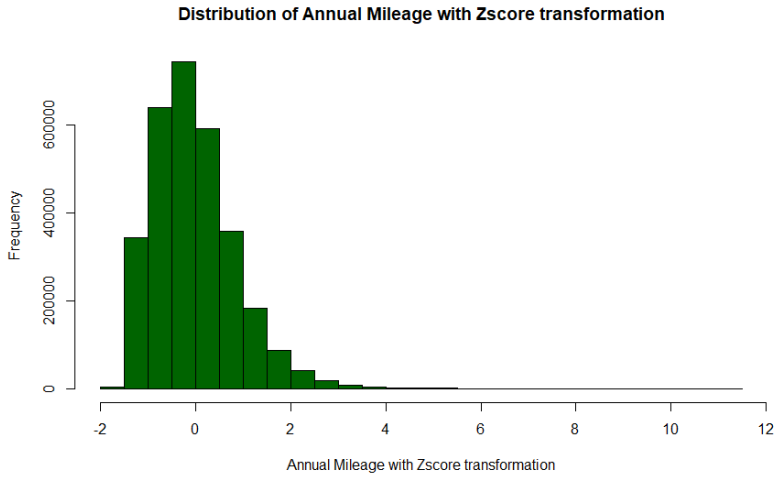

Figure 4: Distribution of Annual Mileage with Z score transformation.

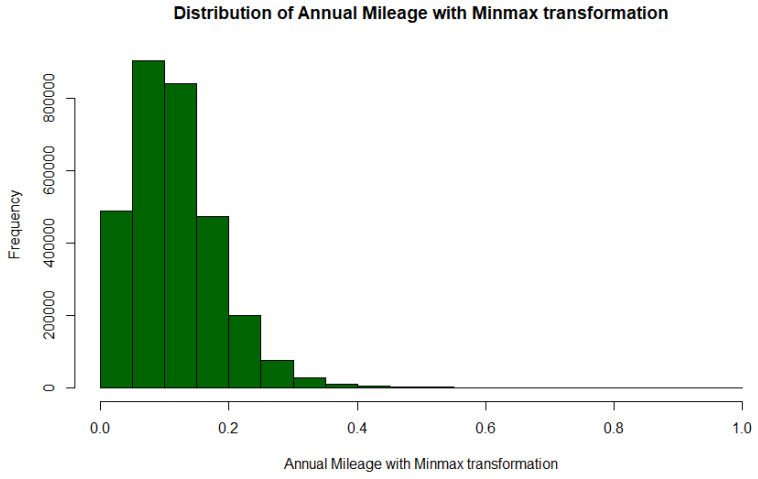

Figure 5: Distribution of the Annual Mileage with Minmax transformation.

4. RESULTS

Table 6 presents the Generalized Linear Models constructed using traditional variables *(Town Risk* and *Driver class*), the transformed annual mileage as the sole predictor variable and the combination of all variables. The resulting models were built based upon the stepwise variable selection process, as referred in the methodology.

The generalized linear model using the traditional rating variables yielded an AIC of 766 326 and the model using only *Mileage* has the worst performance of all the models, regardless of the transformation method used, showing the highest AIC of 776 897. Only the Complete models, which include both traditional rating factors and *Mileage*, are an improvement to the Classical model. The introduction of mileage improves the overall quality of the model with an AIC of 763 380. The increase in the predicting capacity of the model by the inclusion of behavioural risk factors was also found in the conclusions from the original study and from other studies (Ayuso et al., 2019; Ferreira & Minikel, 2012; Lemaire et al., 2015).

Compared to the original study of this data set (Joseph & Minikel, 2010), the results presented are similar. The authors conclude that mileage is an important predictor of insurance risk. However, its explanatory power is not strong enough to replace the traditional variables of prediction already in use in the insurance industry and therefore it should be used in conjunction with traditional rating factors.

Considering both variable transformation methods Min Max and Z-score, the models are very similar in terms of performance, having the same AIC. Additionally, a Chi-squared test was performed and both models are not statistically different.

Additionally, it has been stated that the risk of incurring a claim is not directly proportional to the mileage driven, there is a learning effect to be taken into account, (Boucher et al., 2013). This non-linear relationship can be seen more evidently in the value of the coefficients of the occasional drivers. The introduction of mileage in the model changes the value of the coefficients in occasional drivers with less than 3 years' experience with driver training (1.1046 Vs 1.0773), with no driver training (1.2688 Vs 1.2519) and with 3 to 6 years' experience (0.6438 Vs 0.6287). This effect is not so significant in the principal drivers.

Regarding the *Town Risk* variable, the introduction of mileage in the model has not made any changes in the coefficients has this variable does not depend on the driver's use of the car but the location of where it is mostly garaged.

Table 6: Generalized Linear Models.

			Minmax transformation		Zscore transformation	
Variable	Variable Discription	Classic Model	Complete Model	Mileage only	Complete Model	Mileage only
Intercept		- 3.4773	- 3.7633	- 3.2889	- 3.5204	- 3.0708
Town Risk						
	Town Risk 2	0.1376	0.1422		0.1422	
	Town Risk 3	0.1953	0.2000		0.2000	
	Town Risk 4	0.3107	0.3240		0.3240	
	Town Risk 5	0.4999	0.5368		0.5368	
	Town Risk 6	0.7614	0.8048		0.8048	
Driver Class						
	Business	0.3055	0.2710		0.2710	
	Occasional with less than 3 years experience with driver training	1.1046	1.0773		1.0773	
	Occasional with less than 3 years experience with no driver training	1.2688	1.2519		1.2519	
	Occasional with 3 to 6 years experience	0.6438	0.6287		0.6287	
	Principal with less than 3 years experience with driver training	1.1506	1.1676		1.1677	
	Principal with less than 3 years experience with no driver training	1.1049	1.0963		1.0963	
	Senior citizen	0.0430	0.1473		0.1473	
Ann_Miles_Minmax			2.0581	1.8473		
Ann_Miles_Zscore					0.1606	0.1441
AIC*		766,326	763,380	776,897	763,380	776,897

AIC*: Akaike Information Criterion

After running the 100 resamples for the GLM Bagging process, we have reached the final models with the coefficients and performance measure as shown in Table 7. The GLM Bagging model with the Min Max transformation method is the best-fit model with the lowest AIC of 763 140 when compared to the AIC of the GLM Bagging model with the Z-score transformation of 763 477.

Comparing both regression methodologies (traditional GLM and Bagging GLM), the model with the best performance is the Bagged GLM with Min Max transformation (AIC of 763 140) followed by the traditional transformed GLMs (both with an AIC of 763 380). Although the Bagged GLM with Zscore transformation used 100 resamples to build the model mitigating the risk of biased sampling, it did not showed an improvement to the traditional GLM with the same transformation method (AIC of 763 477 vs. 763 380). The R Script for the regression models is presented in Appendix II.

The coefficients between the normal GLM and the Bagging GLM are very similar. The Town Risk variable did not suffer significant differences, having Town Risk 6 the highest impact on the probability of a claim and Town Risk 2 the lowest. Considering the Driver class, drivers with less of 3 years' experience continue to be the ones with the highest impact on the probability of a claim. The impact being greater in occasional drivers with less than 3 years of experience and no driver training. The distance covered maintains as the most impactful variable when used the Min Max transformation method.

The impact of the two transformation methods could not be perceived using the traditional GLM (both models had the same AIC and the Chi-squared test performed showed that they were not statistically different), however when using the ensemble method of Bagging, a preference to use the Min Max transformation is reached since it yields the best performing model.

Table 7: GLM Bagging Model.

Variable	Variable Discription	GLM Bagging with Min Max transformation	GLM Bagging with Zscore transformation
Intercept		- 3.763349	- 3.5268
	Town Risk 2	0.141269	0.1525
	Town Risk 3	0.204404	0.1973
	Town Risk 4	0.325571	0.3294
	Town Risk 5	0.533938	0.5447
	Town Risk 6	0.816626	0.8316
Driver Class			
	Business	0.268933	0.2744
	Occasional with less than 3 years experience with driver training	1.092077	1.0744
	Occasional with less than 3 years experience with no driver training	1.244144	1.2343
	Occasional with 3 to 6 years experience	0.635739	0.6336
	Principal with less than 3 years experience with driver training	1.175751	1.1753
	Principal with less than 3 years experience with no driver training	1.068040	1.1095
	Senior citizen	0.136182	0.1334
Ann_Miles_Minmax		2.051880	
Ann_Miles_Zscore			0.1584
AIC*		763,140	763,477

AIC*: Akaike Information Criterion

5. DISCUSSION AND CONCLUSION

In this study, we highlight the importance of considering distance in the client's risk assessment for motor insurance policy acceptance. The model that yields the best result is the one that combines both traditional factors with the customer use. This compounding improvement of the prediction model has also been shown in other studies (Ayuso et al., 2019; Ferreira & Minikel, 2012; Lemaire et al., 2015), where the authors concluded that the new risk variables are meant to be used as an improvement of the currently existing models and not as a substitution.

Not only mileage but other risk variables of the client's habits can be introduced in the risk model through the use of telematics. The most recent studies already introduce the percentage of distance travelled at night during the year, the distance travelled above the speed limits and the percentage of distance travelled in urban areas (Ayuso et al., 2019).

The incorporation of this types of variables changes the way risk assessment is done in the insurance industry. The risk appraisal should not be fixed to the moment of inception of the policy but be a continuous process. The insurance companies are better aware of the risks incurred and the costumers are better protected if both sides are aware of the risks incurred. The continuous monitoring of the client's risk behavior is also a chance to invite clients to have a safer conduct on the roads, increasing society's overall road safety, incentivized by the reduction on the premium paid due to the safer conduct.

The limitations faced during the elaboration of this thesis were mainly due to the lack of information regarding the consumer's driving habits. Portuguese insurance companies and industry specific public institutions were contacted in order to obtain information for this thesis, however due to confidentiality and due to the personal data protection law, none were able to disclose an anonymized data set. This is a clear obstacle in research of new risk assessment methods. Additionally, there are incentives to develop this research field since in Portugal it is mandatory to have an 3rd party insurance policy for every vehicle driven on public roads (*Regime Do Sistema Do Seguro Obrigatório de Responsabilidade*, 2007) and is the 6th country in the European Union with the highest road fatalities per million inhabitants, according to the latest study of Eurostat with reference to 2019 (Eurostat, 2021)

Impacts of COVID-19

On the last quarter of 2019, the COVID-19 pandemic started to spread worldwide and the latest Consumer Trends Report of EIOPA with refence to the 30th of June of 2020 is focused on the impacts of the pandemic on the insurance industry in the European Economic Area (EEA).

The motor vehicle line of business experienced a decrease of 4.6% in gross written premiums, although it continues to be the most pronounced product in the non-life sector. However, this decrease is not consistent across all countries of the EEA, since it depends on the measures of each country to battle the pandemic such as lock-down and restrictions of movement.

To contain the spread of the virus, consumers changed their habits and hence their insurance needs. Possible consumer detriment may rise from the fact that products and pricing aspects may not reflect the correct risk levels, emerging a deviation between product risk level and the new consumer's needs.

Specific to the motor vehicle line of business there have been some measures relating to premium payment interruptions. These measures although important, are more addressed to the economic impact of the crisis and less to the changing risk levels of the consumers. In France, the prices for motor insurance are being adjusted based on the kilometers driven by the consumers and in Netherlands and Portugal insurance companies are returning premiums back to the consumers.

In Germany, the insurance industry was already prepared to adjust the pricing of their products since they were already taking into account, in their risk assessment, the risk of the policyholder regarding the kilometers driven (European Insurance and Occupational Pensions Authority., 2020).

Over the past years, there have been several innovations in this line of business and it is expected to continue with more product development in order to meet consumers' needs and demands. The previous adjustments influenced by the pandemic and new risk factors previously mentioned are innovations required in the motor insurance to better align the insurance company's risk assessment with the new clients' needs and demands.

Conclusion

The addition of a variable that measures the driver's behavior (distance covered) increases the overall prediction capacity of the risk assessment model, showing the potential of an add-on strategy. Behavior derived and exposure related variables measured through telematics can be used in conjunction with the existing traditional factors and not as a substitute.

Through the use of GLM Bagging prediction methodology, we attempted to mitigate the issue of having databases with excess of zeros (where a large proportion of observations did not have the predicting event, in our dataset very few drivers reported a claim, 4.5%) combining the traditional GLM with ensemble methods. Which yielded a better predicition capacity than the traditional GLM, without having the coefficients explainability or the model's transparency issues of other ensemble methods.

This technology is an untapped potential since the majority of the auto insurance products do not yet account for driver's behavior. Additionally, there is an increase in smartphone and car users that facilitates the measurement of these new metrics. Furthermore, being motor insurance mandatory in many countries, an increase in government support would accelerate this research field. The improved and continuous risk assessment of driver's behavior, constitutes an incentive for the societal benefit of having safer roads.

6. BIBLIOGRAPHY

Act On Compulsory Insurance For Motorists, (1939) (Germany).

Agresti, A. (2003). *Categorical Data Analysis* (Vol. 482). John Wiley & Sons.

Aseervatham, V., Lex, C., & Spindler, M. (2016). How do unisex rating regulations affect gender differences in insurance premiums?. *The Geneva Papers on Risk and Insurance: Issues and Practice*, *41*(1), 128–160.

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.

Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis and Prevention*, *73*, 125–131.

Azzopardi, M., & Cortis, D. (2013). Implementing automotive telematics for insurance covers of fleets. *Journal of techcnology management & innovation*, *8*(4), 59–67.

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, *98*, 69–79.

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation research part A: policy and practice*, *107*, 20–34.

Boucher, J.-P., Pérez-Marín, A. M., & Santolino, M. (2013). Pay-as-you-drive Insurance: the effect of the kilometers on the risk of accident. In *Anales Del Instituto de Actuarios Españoles* (Vol. 19, No 3, pp. 135–154).

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24(2)*, 123–140.

De Romph, E. (2013). Using BIG data in transport modelling. *Data & Modelling Magazine*, (*13)* 2013.

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and bonus-malus systems*. John Wiley & Sons.

Ellison, A. B., Bliemer, M. C., & Greaves, S. P. (2015). Evaluating changes in driver behaviour: A risk profiling approach. *Accident Analysis and Prevention*, 75, 298–309.

EU rules on gender-neutral pricing in insurance industry enter into force, (2012) (European Commission).

European Insurance and Occupational Pensions Authority. (2019). *Big Data Analytics in Motor And Health Insurance: A Thematic Review*.

European Insurance and Occupational Pensions Authority. (2019). *Consumer Trends Report 2019*.

European Insurance and Occupational Pensions Authority. (2020). *Consumer Trends Report 2020*.

Eurostat. (2021). *Road accidents: number of fatalities continues falling*.

Ferreira, J., & Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation Research Record*, 2297 (1), 97–103.

Händel, P., Ohlsson, J., Ohlsson, M., Skog, I., & Nygren, E. (2013). Smartphone-based measurement systems for road vehicle traffic monitoring and usage-based insurance. *IEEE Systems Journal*, 8(4), 1238–1248.

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, *100*, 816–825.

Joseph, F. j., & Minikel, E. (2010). *Pay-as-You-Drive Auto Insurance In Massachusetts A Risk Assessment And Report On Consumer, Industry And Environmental Benefits*.

Lee, I. J. (2014). Big Data Processing Framework of Road Traffic Collision Using Distributed CEP. *In The 16th Asia-Pacific Network Operations and Management Symposium (pp. 1-4). IEEE.*

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin: The Journal of the IAA*, *46*(1), 39–69.

Lewis, S. (2016). Insurtech: an industry ripe for disruption. *Geo. L. Tech. Rev.,* 1, 491.

Litman, T. (1997). Distance-Based Vehicle Insurance as a TDM strategy. *Transportation Quarterly*. *51*, 119-137.

Litman, T. (2005). Pay-As-You-Drive Pricing and Insurance Regulatory Objectives. *Journal of Insurance Regulation*, 23(3).

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

OECD. (2019). *Insurance business written in the reporting country*.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Troncoso, C., Danezis, G., Kosta, E., Balasch, J., & Preneel, B. (2010). PriPAYD: Privacy-friendly pay-as-you-drive insurance. *IEEE Transactions on Dependable and Secure Computing*, *8*(5), 742–755.

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis and Prevention*, *98*, 139–148.

Regime do sistema do seguro obrigatório de responsabilidade, (2007). Diário da República Portuguesa.

Regulation 2015/758, (2015) (European Parliament and of the Council).

Regulation (EU) 2016/679, (2016) (European Parliament and of the Council).

Road Traffic Act, (1930) (testimony of United Kingdom).

Vaia, G., & Trautsch, H. (2012). *Quarterly Executive*.

Wards Intelligence. (2017). *World Vehicle Population Rose 4.6% in 2016.*

Williams, B., Hansen, G., Baraben, A., & Santoni, A. (2015). A Pratical Approach to Variable Selection - a comparison of various tecnhiques. In *CAS E-Forum 2015*.

7. APPENDIX I – R SCRIPT FOR DATA PREPARATION

```
rm(list = ls())

library(tidyverse)

library(Hmisc)

library(dplyr)

expo06_amile <- read_csv("expo06_amile.csv")

str(expo06_amile)

expo06_amile$pol_id <- as.character(expo06_amile$pol_id)

expo06_amile$class4 <- as.character(expo06_amile$class4)

expo06_amile$trank <- as.integer(expo06_amile$trank)

expo06_amile$ecode <- as.integer(expo06_amile$ecode)

expo06_amile$days_overlap <- as.integer(expo06_amile$days_overlap)

all06clms <- read_csv("all06clms.csv")

all06clms$clm_id <- as.integer(all06clms$clm_id)

all06clms$pol_id <- as.character(all06clms$pol_id)

all06clms$subln_cde <- as.character(all06clms$subln_cde)

all06clms$losspaid <- as.integer(all06clms$losspaid)

all06clms$tcount <- as.integer(all06clms$tcount)

all06clms$lossreserve <- as.integer(all06clms$lossreserve)

all06clms$rcount <- as.integer(all06clms$rcount)

BaseDados1 <- left_join(expo06_amile, all06clms, by = c("vin" = "vin", "pol_id" = "pol_id"))

BaseDados1$Claim=ifelse(is.na(BaseDados1$clm_id),0,1)

BaseDados1Filtrada <- BaseDados1 %>%

  filter (BaseDados1$Claim == 0 | (BaseDados1$adate > BaseDados1$startd & BaseDados1$adate <
BaseDados1$enddate))

terrgroups <- read_csv("terrgroups.csv")

terrgroups$tgroup <- as.character(terrgroups$tgroup)

str(terrgroups)
```

```
BaseDados2 <- inner_join(BaseDados1Filtrada, terrgroups, by = "prem_twn")

BaseDados2$rateclass <- substr(BaseDados2$class4,4,4)

str(BaseDados2)

classgroups <- read_csv("classgroups.csv")

classgroups$rateclass <- as.character(classgroups$rateclass)

BaseDados3 <- inner_join(BaseDados2, classgroups, by = "rateclass")

fuel_economy_summary <- read_csv("fuel_economy_summary.csv")

fuel_economy_summary$tgroup <- as.character(fuel_economy_summary$tgroup)

BaseDados3$mileage_bin[which(BaseDados3$ann_miles<500)] <- 500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=500 & BaseDados3$ann_miles<1000)] <-
1000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=1000 & BaseDados3$ann_miles<1500)] <-
1500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=1500 & BaseDados3$ann_miles<2000)] <-
2000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=2000 & BaseDados3$ann_miles<2500)] <-
2500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=2500 & BaseDados3$ann_miles<3000)] <-
3000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=3000 & BaseDados3$ann_miles<3500)] <-
3500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=3500 & BaseDados3$ann_miles<4000)] <-
4000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=4000 & BaseDados3$ann_miles<4500)] <-
4500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=4500 & BaseDados3$ann_miles<5000)] <-
5000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=5000 & BaseDados3$ann_miles<5500)] <-
5500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=5500 & BaseDados3$ann_miles<6000)] <-
6000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=6000 & BaseDados3$ann_miles<6500)] <-
6500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=6500 & BaseDados3$ann_miles<7000)] <-
7000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=7000 & BaseDados3$ann_miles<7500)] <-
7500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=7500 & BaseDados3$ann_miles<8000)] <-
8000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=8000 & BaseDados3$ann_miles<8500)] <-
8500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=8500 & BaseDados3$ann_miles<9000)] <-
9000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=9000 & BaseDados3$ann_miles<9500)] <-
9500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=9500 & BaseDados3$ann_miles<10000)]
<- 10000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=10000 & BaseDados3$ann_miles<10500)]
<- 10500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=10500 & BaseDados3$ann_miles<11000)]
<- 11000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=11000 & BaseDados3$ann_miles<11500)]
<- 11500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=11500 & BaseDados3$ann_miles<12000)]
<- 12000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=12000 & BaseDados3$ann_miles<12500)]
<- 12500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=12500 & BaseDados3$ann_miles<13000)]
<- 13000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=13000 & BaseDados3$ann_miles<13500)]
<- 13500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=13500 & BaseDados3$ann_miles<14000)]
<- 14000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=14000 & BaseDados3$ann_miles<14500)]
<- 14500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=14500 & BaseDados3$ann_miles<15000)]
<- 15000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=15000 & BaseDados3$ann_miles<15500)]
<- 15500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=15500 & BaseDados3$ann_miles<16000)]
<- 16000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=16000 & BaseDados3$ann_miles<16500)]
<- 16500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=16500 & BaseDados3$ann_miles<17000)]
<- 17000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=17000 & BaseDados3$ann_miles<17500)]
<- 17500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=17500 & BaseDados3$ann_miles<18000)]
<- 18000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=18000 & BaseDados3$ann_miles<18500)]
<- 18500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=18500 & BaseDados3$ann_miles<19000)]
<- 19000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=19000 & BaseDados3$ann_miles<19500)]
<- 19500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=19500 & BaseDados3$ann_miles<20000)]
<- 20000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=20000 & BaseDados3$ann_miles<20500)]
<- 20500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=20500 & BaseDados3$ann_miles<21000)]
<- 21000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=21000 & BaseDados3$ann_miles<21500)]
<- 21500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=21500 & BaseDados3$ann_miles<22000)]
<- 22000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=22000 & BaseDados3$ann_miles<22500)]
<- 22500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=22500 & BaseDados3$ann_miles<23000)]
<- 23000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=23000 & BaseDados3$ann_miles<23500)]
<- 23500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=23500 & BaseDados3$ann_miles<24000)]
<- 24000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=24000 & BaseDados3$ann_miles<24500)]
<- 24500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=24500 & BaseDados3$ann_miles<25000)]
<- 25000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=25000 & BaseDados3$ann_miles<25500)]
<- 25500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=25500 & BaseDados3$ann_miles<26000)]
<- 26000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=26000 & BaseDados3$ann_miles<26500)]
<- 26500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=26500 & BaseDados3$ann_miles<27000)]
<- 27000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=27000 & BaseDados3$ann_miles<27500)]
<- 27500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=27500 & BaseDados3$ann_miles<28000)]
<- 28000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=28000 & BaseDados3$ann_miles<28500)]
<- 28500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=28500 & BaseDados3$ann_miles<29000)]
<- 29000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=29000 & BaseDados3$ann_miles<29500)]
<- 29500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=29500 & BaseDados3$ann_miles<30000)]
<- 30000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=30000 & BaseDados3$ann_miles<30500)]
<- 30500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=30500 & BaseDados3$ann_miles<31000)]
<- 31000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=31000 & BaseDados3$ann_miles<31500)]
<- 31500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=31500 & BaseDados3$ann_miles<32000)]
<- 32000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=32000 & BaseDados3$ann_miles<32500)]
<- 32500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=32500 & BaseDados3$ann_miles<33000)]
<- 33000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=33000 & BaseDados3$ann_miles<33500)]
<- 33500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=33500 & BaseDados3$ann_miles<34000)]
<- 34000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=34000 & BaseDados3$ann_miles<34500)]
<- 34500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=34500 & BaseDados3$ann_miles<35000)]
<- 35000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=35000 & BaseDados3$ann_miles<35500)]
<- 35500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=35500 & BaseDados3$ann_miles<36000)]
<- 36000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=36000 & BaseDados3$ann_miles<36500)]
<- 36500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=36500 & BaseDados3$ann_miles<37000)]
<- 37000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=37000 & BaseDados3$ann_miles<37500)]
<- 37500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=37500 & BaseDados3$ann_miles<38000)]
<- 38000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=38000 & BaseDados3$ann_miles<38500)]
<- 38500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=38500 & BaseDados3$ann_miles<39000)]
<- 39000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=39000 & BaseDados3$ann_miles<39500)]
<- 39500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=39500 & BaseDados3$ann_miles<40000)]
<- 40000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=40000 & BaseDados3$ann_miles<40500)]
<- 40500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=40500 & BaseDados3$ann_miles<41000)]
<- 41000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=41000 & BaseDados3$ann_miles<41500)]
<- 41500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=41500 & BaseDados3$ann_miles<42000)]
<- 42000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=42000 & BaseDados3$ann_miles<42500)]
<- 42500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=42500 & BaseDados3$ann_miles<43000)]
<- 43000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=43000 & BaseDados3$ann_miles<43500)]
<- 43500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=43500 & BaseDados3$ann_miles<44000)]
<- 44000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=44000 & BaseDados3$ann_miles<44500)]
<- 44500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=44500 & BaseDados3$ann_miles<45000)]
<- 45000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=45000 & BaseDados3$ann_miles<45500)]
<- 45500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=45500 & BaseDados3$ann_miles<46000)]
<- 46000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=46000 & BaseDados3$ann_miles<46500)]
<- 46500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=46500 & BaseDados3$ann_miles<47000)]
<- 47000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=47000 & BaseDados3$ann_miles<47500)]
<- 47500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=47500 & BaseDados3$ann_miles<48000)]
<- 48000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=48000 & BaseDados3$ann_miles<48500)]
<- 48500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=48500 & BaseDados3$ann_miles<49000)]
<- 49000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=49000 & BaseDados3$ann_miles<49500)]
<- 49500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=49500 & BaseDados3$ann_miles<50000)]
<- 50000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=50000 & BaseDados3$ann_miles<50500)]
<- 50500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=50500 & BaseDados3$ann_miles<51000)]
<- 51000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=51000 & BaseDados3$ann_miles<51500)]
<- 51500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=51500 & BaseDados3$ann_miles<52000)]
<- 52000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=52000 & BaseDados3$ann_miles<52500)]
<- 52500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=52500 & BaseDados3$ann_miles<53000)]
<- 53000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=53000 & BaseDados3$ann_miles<53500)]
<- 53500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=53500 & BaseDados3$ann_miles<54000)]
<- 54000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=54000 & BaseDados3$ann_miles<54500)]
<- 54500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=54500 & BaseDados3$ann_miles<55000)]
<- 55000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=55000 & BaseDados3$ann_miles<55500)]
<- 55500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=55500 & BaseDados3$ann_miles<56000)]
<- 56000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=56000 & BaseDados3$ann_miles<56500)]
<- 56500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=56500 & BaseDados3$ann_miles<57000)]
<- 57000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=57000 & BaseDados3$ann_miles<57500)]
<- 57500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=57500 & BaseDados3$ann_miles<58000)]
<- 58000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=58000 & BaseDados3$ann_miles<58500)]
<- 58500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=58500 & BaseDados3$ann_miles<59000)]
<- 59000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=59000 & BaseDados3$ann_miles<59500)]
<- 59500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=59500 & BaseDados3$ann_miles<60000)]
<- 60000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=60000 & BaseDados3$ann_miles<60500)]
<- 60500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=60500 & BaseDados3$ann_miles<61000)]
<- 61000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=61000 & BaseDados3$ann_miles<61500)]
<- 61500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=61500 & BaseDados3$ann_miles<62000)]
<- 62000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=62000 & BaseDados3$ann_miles<62500)]
<- 62500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=62500 & BaseDados3$ann_miles<63000)]
<- 63000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=63000 & BaseDados3$ann_miles<63500)]
<- 63500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=63500 & BaseDados3$ann_miles<64000)]
<- 64000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=64000 & BaseDados3$ann_miles<64500)]
<- 64500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=64500 & BaseDados3$ann_miles<65000)]
<- 65000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=65000 & BaseDados3$ann_miles<65500)]
<- 65500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=65500 & BaseDados3$ann_miles<66000)]
<- 66000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=66000 & BaseDados3$ann_miles<66500)]
<- 66500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=66500 & BaseDados3$ann_miles<67000)]
<- 67000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=67000 & BaseDados3$ann_miles<67500)]
<- 67500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=67500 & BaseDados3$ann_miles<68000)]
<- 68000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=68000 & BaseDados3$ann_miles<68500)]
<- 68500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=68500 & BaseDados3$ann_miles<69000)]
<- 69000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=69000 & BaseDados3$ann_miles<69500)]
<- 69500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=69500 & BaseDados3$ann_miles<70000)]
<- 70000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=70000 & BaseDados3$ann_miles<70500)]
<- 70500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=70500 & BaseDados3$ann_miles<71000)]
<- 71000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=71000 & BaseDados3$ann_miles<71500)]
<- 71500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=71500 & BaseDados3$ann_miles<72000)]
<- 72000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=72000 & BaseDados3$ann_miles<72500)]
<- 72500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=72500 & BaseDados3$ann_miles<73000)]
<- 73000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=73000 & BaseDados3$ann_miles<73500)]
<- 73500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=73500 & BaseDados3$ann_miles<74000)]
<- 74000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=74000 & BaseDados3$ann_miles<74500)]
<- 74500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=74500 & BaseDados3$ann_miles<75000)]
<- 75000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=75000 & BaseDados3$ann_miles<75500)]
<- 75500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=75500 & BaseDados3$ann_miles<76000)]
<- 76000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=76000 & BaseDados3$ann_miles<76500)]
<- 76500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=76500 & BaseDados3$ann_miles<77000)]
<- 77000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=77000 & BaseDados3$ann_miles<77500)]
<- 77500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=77500 & BaseDados3$ann_miles<78000)]
<- 78000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=78000 & BaseDados3$ann_miles<78500)]
<- 78500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=78500 & BaseDados3$ann_miles<79000)]
<- 79000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=79000 & BaseDados3$ann_miles<79500)]
<- 79500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=79500 & BaseDados3$ann_miles<80000)]
<- 80000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=80000 & BaseDados3$ann_miles<80500)]
<- 80500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=80500 & BaseDados3$ann_miles<81000)]
<- 81000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=81000 & BaseDados3$ann_miles<81500)]
<- 81500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=81500 & BaseDados3$ann_miles<82000)]
<- 82000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=82000 & BaseDados3$ann_miles<82500)]
<- 82500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=82500 & BaseDados3$ann_miles<83000)]
<- 83000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=83000 & BaseDados3$ann_miles<83500)]
<- 83500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=83500 & BaseDados3$ann_miles<84000)]
<- 84000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=84000 & BaseDados3$ann_miles<84500)]
<- 84500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=84500 & BaseDados3$ann_miles<85000)]
<- 85000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=85000 & BaseDados3$ann_miles<85500)]
<- 85500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=85500 & BaseDados3$ann_miles<86000)]
<- 86000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=86000 & BaseDados3$ann_miles<86500)]
<- 86500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=86500 & BaseDados3$ann_miles<87000)]
<- 87000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=87000 & BaseDados3$ann_miles<87500)]
<- 87500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=87500 & BaseDados3$ann_miles<88000)]
<- 88000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=88000 & BaseDados3$ann_miles<88500)]
<- 88500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=88500 & BaseDados3$ann_miles<89000)]
<- 89000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=89000 & BaseDados3$ann_miles<89500)]
<- 89500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=89500 & BaseDados3$ann_miles<90000)]
<- 90000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=90000 & BaseDados3$ann_miles<90500)]
<- 90500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=90500 & BaseDados3$ann_miles<91000)]
<- 91000
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=91000 & BaseDados3$ann_miles<91500)]
<- 91500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=91500 & BaseDados3$ann_miles<92000)]
<- 92000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=92000 & BaseDados3$ann_miles<92500)]
<- 92500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=92500 & BaseDados3$ann_miles<93000)]
<- 93000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=93000 & BaseDados3$ann_miles<93500)]
<- 93500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=93500 & BaseDados3$ann_miles<94000)]
<- 94000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=94000 & BaseDados3$ann_miles<94500)]
<- 94500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=94500 & BaseDados3$ann_miles<95000)]
<- 95000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=95000 & BaseDados3$ann_miles<95500)]
<- 95500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=95500 & BaseDados3$ann_miles<96000)]
<- 96000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=96000 & BaseDados3$ann_miles<96500)]
<- 96500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=96500 & BaseDados3$ann_miles<97000)]
<- 97000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=97000 & BaseDados3$ann_miles<97500)]
<- 97500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=97500 & BaseDados3$ann_miles<98000)]
<- 98000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=98000 & BaseDados3$ann_miles<98500)]
<- 98500

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=98500 & BaseDados3$ann_miles<99000)]
<- 99000

BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=99000 & BaseDados3$ann_miles<99500)]
<- 99500
```

```
BaseDados3$mileage_bin[which(BaseDados3$ann_miles>=99500 &
BaseDados3$ann_miles<=100000)] <- 100000

BaseDados4 <- inner_join(BaseDados3, fuel_economy_summary, by =
c("mileage_bin"="mileage_bin","cgroup"="cgroup","tgroup"="tgroup"))

BaseDados4_Filtrada <- BaseDados4 %>%

 filter(BaseDados4$ann_miles != -1 | BaseDados4$days_overlap != -1 |
BaseDados4$fraction_overlap != -1)

BaseDados4_Filtrada$TotaCost <- BaseDados4_Filtrada$losspaid + BaseDados4_Filtrada$lossreserve

BaseDados5_Filtrada <- BaseDados4_Filtrada %>%

 filter(!BaseDados4_Filtrada$TotaCost <= 50 | is.na(BaseDados4_Filtrada$TotaCost))

número VIN)

BaseDadosFinal <- BaseDados4_Filtrada %>%

 distinct(vin, .keep_all = TRUE)

sapply(BaseDadosFinal, function(x) sum(is.na(x)))

mean(BaseDadosFinal$ann_miles)

aggregate(x=BaseDadosFinal$ann_miles, by = list(BaseDadosFinal$cgroup), FUN = mean)

aggregate(x=BaseDadosFinal$ann_miles, by = list(BaseDadosFinal$tgroup), FUN = mean)

aggregate(x=BaseDadosFinal$earnexpo, by = list(BaseDadosFinal$cgroup), FUN = sum)

BaseDadosVersaoFinal <- subset(BaseDadosFinal, select=c(13, 24, 26, 28))

write.table(BaseDadosVersaoFinal, file="BaseDadosFinalTratada.csv", row.names=F, sep = ",")
```

8. APPENDIX II – R SCRIPT FOR REGRESSION MODELS

```
rm(list = ls())

library(tidyverse)

library(Hmisc)

library(gmodels)

library(aod)

library(pROC)

library(bigmemory)

library(dlookr)

library(dplyr)

library(caret)

library(leaps)

BaseDados <- read_csv("BaseDadosFinalTratada.csv")

View(BaseDados)

BaseDados$tgroup <- as.factor(BaseDados$tgroup)

BaseDados$rateclassdescrip <- as.factor(BaseDados$rateclassdescrip)

BaseDados$Claim <- as.factor(BaseDados$Claim)

levels(BaseDados$tgroup)

levels(BaseDados$rateclassdescrip)

prop.table(table(BaseDados$tgroup))

prop.table(table(BaseDados$rateclassdescrip))

summary(BaseDados)

sum(BaseDados$rateclassdescrip == "Occasional w/ <3yr exp, no driver training",na.rm = TRUE)

sum(BaseDados$rateclassdescrip == "Principal w/ <3yr exp, no driver training",na.rm = TRUE)

sd(BaseDados$ann_miles)

Sinistros <- BaseDados %>%

  filter(BaseDados$Claim == 1)
```

```
summary(Sinistros)

sum(Sinistros$rateclassdescrip == "Occasional w/ <3yr exp, no driver training",na.rm = TRUE)

sum(Sinistros$rateclassdescrip == "Principal w/ <3yr exp, no driver training",na.rm = TRUE)

sd(Sinistros$ann_miles)


SemSinistros <- BaseDados %>%

  filter(BaseDados$Claim == 0)

summary(SemSinistros)

sum(SemSinistros$rateclassdescrip == "Occasional w/ <3yr exp, no driver training",na.rm = TRUE)

sum(SemSinistros$rateclassdescrip == "Principal w/ <3yr exp, no driver training",na.rm = TRUE)

sd(SemSinistros$ann_miles)

CrossTable(BaseDados$Claim)

CrossTable(BaseDados$Claim, BaseDados$tgroup, digits=1, prop.r=F, prop.t=F, prop.chisq=F,
chisq=T)

CrossTable(BaseDados$Claim, BaseDados$rateclassdescrip, digits=1, prop.r=F, prop.t=F,
prop.chisq=F, chisq=T)

summary(BaseDados$ann_miles)

BaseDados$ann_miles_minmax <- transform(BaseDados$ann_miles, method = "minmax")

BaseDados$ann_miles_zscore <- transform(BaseDados$ann_miles, method = "zscore")

options(scipen = 999)

hist(BaseDados$ann_miles, main="Distribution of Annual Mileage", xlab="Annual Mileage",

    breaks = 50, col="dark green")

hist(BaseDados$ann_miles_zscore, main="Distribution of Annual Mileage with Zscore
transformation",

    xlab="Annual Mileage with Zscore transformation",

    col="dark green")

hist(BaseDados$ann_miles_minmax, main="Distribution of Annual Mileage with Minmax
transformation",

    xlab="Annual Mileage with Minmax transformation",
```

```
    col="dark green")

max(BaseDados$ann_miles_zscore)

max(BaseDados$ann_miles_minmax)

ks.test(BaseDados$ann_miles, "pnorm", mean=mean(BaseDados$ann_miles),
sd=sd(BaseDados$ann_miles))

set.seed(123456)

tds = 3/10 # proportion in training data

d = sort(sample(nrow(BaseDados), nrow(BaseDados)*tds))

train <- BaseDados[-d,] #7/10 of observations

test <- BaseDados[d,] #3/10 of observations

modelobase <- glm(Claim ~ tgroup + rateclassdescrip, family=binomial, data = train)

summary(modelobase)

modeloann_mileszscore <- glm(Claim ~ ann_miles_zscore, family = binomial, data = train)

summary(modeloann_mileszscore)

modeloann_milesminmax <- glm(Claim ~ ann_miles_minmax, family = binomial, data = train)

summary(modeloann_milesminmax)

modelototal <- glm(Claim ~tgroup + rateclassdescrip + ann_miles, family = binomial, data = train)

summary(modelototal)

modelototalminmax <- glm(Claim ~ tgroup + rateclassdescrip + ann_miles_minmax,
family=binomial,data = train)

summary(modelototalminmax)

step.model.minmax <- step(modelototalminmax, direction = "both", trace = FALSE)

summary(step.model.minmax)

modelototalzscore <- glm (Claim ~ tgroup + rateclassdescrip + ann_miles_zscore, family=binomial,
data = train)

summary(modelototalzscore)

step.model.zscore <- step(modelototalzscore, direction = "both", trace = FALSE)

summary(step.model.zscore)

anova(modelobase, modelototalzscore, test="Chisq")
```

```
anova(modelototalminmax, modelototalzscore, test="Chisq")

library(tidyverse)

library(caret)

train.control <- trainControl(method = "boot", number = 100)

model <- train(Claim ~ tgroup + rateclassdescrip + ann_miles_minmax,

        data = BaseDados[-sort(sample(nrow(BaseDados), nrow(BaseDados)*tds)),],

        method = "glm",

        trControl = train.control)

print(model)

summary(model)

modelzscore <- train(Claim ~ tgroup + rateclassdescrip + ann_miles_zscore,

        data = BaseDados[-sort(sample(nrow(BaseDados), nrow(BaseDados)*tds)),],

        method = "glm",

        trControl = train.control)

print(modelzscore)

summary(modelzscore)
```