How to Cite:
Sahraoui, A., & Souakri, R. (2024). Enhanced insurance analytics: Precision through data and statistics. *International Journal of Economic Perspectives*, *18*(12), 2953–2967. Retrieved from https://ijeponline.org/index.php/journal/article/view/802

Enhanced insurance analytics: Precision through data and statistics

Abdelaziz Sahraoui
University of ABBES Laghrour khenchela, Algeria
Email: sahraoui.abdelaziz@univ-khenchela.dz

Roufaida Souakri
University of Shahid Mustapha Benboulaid Batna 2, Algeria
Email: Roufaida.souakri@univ-batna2.dz

***Abstract*---**This study investigates the factors influencing claims frequency and costs in the Algerian automobile insurance market, a sector currently stagnating due to reduced vehicle imports. Using comprehensive datasets from the Algerian Insurance Company (SAA), a dual-methodological approach is employed: the Poisson model is applied to analyze claims frequency, while decision tree techniques are utilized to identify key variables shaping claims behavior. The research primarily seeks to uncover the critical determinants of reported accidents and associated claims costs, addressing the central question: "What are the most significant variables influencing claims frequency and costs?" By doing so, the study aims to offer valuable insights into effective risk classification and premium adjustment strategies. The findings are anticipated to enhance understanding of claims dynamics within the Algerian insurance sector, providing actionable recommendations for insurers to improve operational efficiency and refine risk management practices. While tailored to the Algerian context, this research also holds relevance for other markets confronting similar challenges in automobile insurance.
JEL Classification: C19, C55, C88

***Keywords*---**Poisson model, decision tree, claims, automobile insurance.

1. Introduction

Insurance plays a significant role in contributing to economic growth. In the context of the Algerian insurance market, automobile insurance represents the primary sector; however, growth in this area is currently stagnating due to a

***Corresponding author**: Sahraoui, A., Email: sahraoui.abdelaziz@univ-khenchela.dz*
Submitted: 09 October 2024, Revised: 18 November 2024, Accepted: 28 December 2024

decline in vehicle imports. The experience of automobile insurance claims poses a substantial challenge for industrialized nations, as it is evaluated through metrics such as accident frequency and the associated costs.

In today's data-driven landscape, insurance companies have amassed extensive datasets. The decreasing costs of data storage and processing power have facilitated the accumulation of increasingly large volumes of information. According to industry estimates, the global volume of data collected doubles approximately every 20 months, while the extraction of value-added insights from this data shows only modest growth.

In light of this, the establishment of a comprehensive database that consolidates the insurance company's data in a coherent and uniform manner presents new opportunities for users, particularly in the realm of knowledge extraction through data mining tools. The recent advancements in actuarial methodologies raise pertinent questions that lend themselves to the application of data mining techniques.

Within the highly competitive automobile insurance market, which constitutes the largest segment of non-life insurance, companies strive to identify the factors that elucidate claims experience. By constructing risk classes based on these factors, insurers can effectively segment their portfolios and rank these classes according to claims indicators such as pure premium. The objective of this approach is to achieve an optimal alignment between claims experience and the premiums paid by policyholders. Accordingly, this study aims to identify the primary factors influencing claims experience in terms of both frequency and cost, utilizing data mining methods within a supervised learning framework.

In this work, we focus on the elements that explain the number of reported accidents. Generally, counting models (Poisson model) are used in modeling the frequency of claims. This is the model we have decided to apply in this study. To strengthen our analysis, we have chosen another classification method: decision tree techniques. We applied both models to the data collected from the Algerian Insurance Company (SAA). The primary objective of this study is to address the following research question:
"**What are the most decisive variables that can influence the increase in claims frequency and the determination of claims costs?**"

2. The theoretical aspect of insurance and the automobile branch

2.1.1. Definition of insurance

Juridical definition

According to article 2 of ordinance no. 95/07 of January 25, 1995 on insurance, as amended and supplemented by law no. 06/04 of February 20, 2006, under article 619 of the Civil Code: "insurance is a contract by which the insurer undertakes, in return for the payment of premiums or other pecuniary instalments, to provide the insured or the third-party beneficiary for whose benefit the insurance is taken out, with a sum of money, an annuity or another pecuniary benefit, in the event of the occurrence of the risk provided for in the contract".

2.1.2 The importance of insurance in modern economies

Today, insurance is often perceived as a brake on economic development, but it has become a major branch of the economy, with insurers' sales never ceasing to grow, representing a rapidly increasing percentage of each country's Gross Domestic Product (GDP). Insurance's share of GDP can be as high as 15%, and is generally higher the higher a country's level of economic development. Insurance is a driving force behind economic and social development.

2.2 Automobile insurance in Algeria

Automobile insurance belongs to the group of insurance operations whose purpose is not the life of the insured. On the one hand, it represents a very important part of an individual's assets; on the other hand, it is exposed to accidents that cause mortality, which is why governments have made this type of insurance compulsory for its civil responsibility part. For this reason, the insurance product most familiar to the general public is automobile insurance, under which the victim of an automobile accident is compensated by the company insuring the person responsible for the accident.

3. Datamining concepts and techniques

3.1 Definition

According to Kantardzic Mehmed in his book Datamining, concepts, models, methods and algorithms, "Data mining is the set of methods and techniques designed to explore and analyze computer databases (often large ones), automatically or semi-automatically, with a view to detecting rules, associations, unknown or hidden trends, and particular structures in the data, thereby restoring the essential useful information while reducing the amount of data".

3.2 Decision trees

Decision trees are decision-support tools which, based on discriminating variables, can divide a population of individuals into homogeneous groups according to a known objective. A decision tree is a hierarchical sequence of logical rules built automatically from a database of examples. An example is made up of a list of attributes whose value determines membership of a given class. The decision tree is built by using the attributes to progressively subdivide the set of examples into finer and finer subsets.

4. **Practical part:** Extracting knowledge and interpreting results

4.1 Pretreatment

The data pretreatment phase is crucial, because the choice of descriptors and precise knowledge of the population will determine the development of prediction models. The information needed to build a good prediction model may be available in the data, but an inappropriate choice of variables or learning sample may cause the operation to fail.

4.1.1 Data source

The data for our study comes from an Automobile insurance portfolio of the Algerian insurance company (Agence of N'gaous in the wilaya of Batna) from the year 2017. This is a certain amount of information that the insurer needs when underwriting a given Guarantee.
We have an initial database with n = 538 rows and p = 55 columns.

We have carried out some preliminary work that will be crucial for the rest of the study. This involves formatting the database. However, it should be noted that, following this data verification work, our dataset now consists of 488 observations and 24 variables.

4.1.2 Variable selection and presentation

The variables are classified into four groups: those characterizing the driver, those characterizing the insured vehicle, those characterizing the insurance contract, and finally those characterizing the claims experience of the insured vehicle.

Variables that characterize the driver:

SEX: binary qualitative variable indicating the sex of the policyholder. Its modalities are: Male and Female.

AGE of the policyholder: integer-valued variable giving the policyholder's age on the date of subscription. This value is obtained from the date of birth and the year of subscription.

DRIVER'S LICENSE AGE: qualitative binary variable indicating the age of the policyholder's driver's license on the date of subscription. This value is obtained from the date of issue and the year the policy was taken out. The modalities are: plus 1 year and minus 1 year.

Commune: in our database, we have 22 communes, but after statistical analysis, we have retained the communes that represent the largest number of individuals; the others have been integrated with the nearest commune. We thus obtain four modalities for this variable:

5600 (N'gaous) ,05610(Soufiane) ,05680(OSS) ,05660(Taxlent).

Variables that characterize the insured vehicle:

There are several variables that characterize the vehicle, including: type, brand, usage, weight, energy, power…ETC.
For the purposes of this study, we decided to keep the following variables:

Type: initially we had 9 types:
00: private vehicles without trailers
30: vehicles with a total weight exceeding 3.5 T
31: Trailers weighing over 3.5 T
32: TPM with CU exceeding 2T without undercarriage
33 : TPM trailers (quintaux)
34 : Public passenger transport
36: Road tractors only (power)
45: TP construction equipment used on public roads
50: Pneumatic tractor with trailer
In this study, we are only interested in passenger cars without trailers.

Brand: Initially, our database contained 42 brands, so we decided to select the brands representing the greatest number of individuals. In addition, we created a new modality called "other" to include all other brands marked by a small number of individuals.

Once this operation has been carried out, we obtain 4 modalities for this variable: HYUNDAI, PEUGEOT, RENAULT, Other.
Use: is a qualitative variable describing the use or category of use of the vehicle. Vehicles are classified into 6 usage categories: Business, Trade, Civil servant, Trade C. Bis, Rental, Taxi.

As with the other variables, we merged the smallest modalities. The two categories Commerce and Commerce C. Bis have been grouped together under a new modality called Commerce, and the three modalities Civil Servant, Rental and Taxi under a new modality called Civil Servant.
This gives us, for the variable Usage, three modalities, namely: Business, Commerce, Civil servant.

Weight: is a qualitative variable with three modes (Light, Heavy, Other), indicating the weight of the insured vehicle. As the two modalities Heavy and Other represent a small number of employees, we had to merge them into the modality Other.

Energy: is a qualitative variable with two modalities that indicates the engine's energy source, Gasoline (ES) or Diesel (DS).

Power: the horsepower of a vehicle refers by default to the maximum power supplied by the engine, a qualitative variable with 13 modalities, we have chosen to keep 4 of them which are: ≤4CV, 5CV, 6CV, ≥7CV.

Number of seats: is a qualitative variable initially with 7 modalities, which we have reduced to 3 modalities: 2Plac, 3Plac, 4Plac.

Variables that characterize the insurance contract:

Duration: binary quantitative variable designating the duration of the insurance contract, its modalities are: 6 months and 1 year.
Guarantees: The existing guarantees: in our database are listed in the following table:

Table 1: Automobile insurance guarantees

Guarantees	Abbreviation
Responsabilité Civile	RC
Responsabilité Civile interarabe	RC interarabe
Tout risque simple	TR simple
Dommage Collusion	DC
Dommage Collusion valeur Vénale	DC VV
Vole autoradio de véhicule	VAV
Vole incendie de véhicule	VIV

Guarantees	Abbreviation
Bris de Glaces	BDG
Bris de Glaces panoramique	BDG panoramique
Défense et recours	DR
Personnes Transportées Assurées	PTA
Assistance au véhicule	Assist véhicule
Perte d'exploitation et jouissance	PEA

Source: prepared by authors

Each guarantee is marked with a yes and a no.
Status: is a qualitative variable with three modalities: Bonus, Malus, Nul.
Premium payable: Quantitative variable. It represents the amount actually paid by the policyholder.
Variables that characterize the claims experience of the insured vehicle:
The variables describing the latest claims experience are:
Damage nature: binary qualitative variable with two modes: Material, Mixed.
Type of location: qualitative variable with two modes: During trip, In parking lot.
Rate of obsolescence (percentage): This rate is deducted from the purchase price of an asset (in our case, the vehicle) to obtain its actual value. It is a factor that insurance companies take into account when deciding on the amount of compensation. The rate of obsolescence is calculated differently by different insurers, but in general it always includes the following characteristics: Service life, age and maintenance. Modalities are: 0%, 5%, 10%, 15%, 20%, 25%.
Vehicle category: the following table lists all vehicle categories.

Table 2: Vehicle categories

Code	Catégorie de véhicule
1	Location
2	TPV-long trajet-
3	TPV - urbain -
4	TPV - transport de personnel-
5	Véhicule agricole
6	Véhicule tourisme/Autoécole/ Taxi
7	Véhicule Utilitaire léger inférieur à 3,5
8	Véhicule Utilitaire lourd supérieur à 3,5

Source: prepared by the authors

Our database contains six categories: 1, 2, 4, 6, 7 and 8.
Claim date: Corresponds to the date of the last claim.
In addition, we would like to point out that the following variables have been eliminated as unnecessary for our study: Number of impacts, type of impact, Cause, Damage observed, Amount of supplies, amount of paint, Number of days of downtime, Amount of labor, Rate of obsolescence, type of location, vehicle category, claim date.
Finally, the two variables to be explained are:
Nbr of claims-4years-: total number of claims declared by the policyholder to the company over 4 years, its modalities: zero, one, two, three and more.

Settlement amount: this is the total cost of the claim reported by the policyholder to the company during the policy period, i.e., the total cost charged by the policyholder to the company for the settlement of his claims.

4.1.3 Data cleaning and transformation

Handling missing values

Dealing with missing values is one of the most important steps in the data pre-processing phase. When it comes to the statistical replacement of missing values, care must be taken. The simplest way is to replace the missing value with the most frequent value, or the mean or median. On the other hand, we can delete observations affected by this uncertain event if its contribution to the study does not seem essential. There is no missing data in our database, as we took great care to check our data during data entry.

Variable transformation
The introduction of the "State" variable was deduced from the Bonus and Malus variables.

Discretization of quantitative variables

Discretization is the operation used to divide continuous variables into classes. It is satisfactory when it enables the creation of homogeneous classes that are distinct from one another. We performed discretization using Fisher's algorithm in **xlstat**. We chose this algorithm because it gives us the best discretization for our variables, and the results are shown in Table 3 below.

Table 3: Coding of discretized continuous variables

Variables	Classe	Codage
Age	[18 ; 31[=1 [31 ; 40[=2 [40 ; 49[=3 [49 ; 58[=4 [58 ; 69[=5 [69 ; 88] =6	Age1 Age2 Age3 Age4 Age5 Age6
Premium payable	[101 ; 9051.74 [=1 [9051.74 ; 20444.69 [=2 [20444.69 ; et plus=3	Prime1 Prime2 Prime3
Settlement amount (Montant de règlement)	[1352.39 ; 27292.65 [=1 [27292.65 ; 70799.6 [=2 [70799.6 ; et plus=3	MR1 MR2 MR3

Source: prepared by the authors using Fisher's algorithm in **xlstat**

4.2 Decision trees

Decision trees are a convenient presentation of the classification function, enabling an object to be classified. For the construction of the decision tree, we have opted for the C4.5 learning method, based on the error rate as an evaluation

measure for partitioning. We apply the model twice. First, on the dependent variable **Cost of claim**, then in relation to the variable **Claim frequency**.

4.2.1 The variable to be predicted "Cost of claim":

Application of the C4.5 algorithm resulted in the decision tree shown in Figure 1. We have manually deleted some irrelevant nodes.

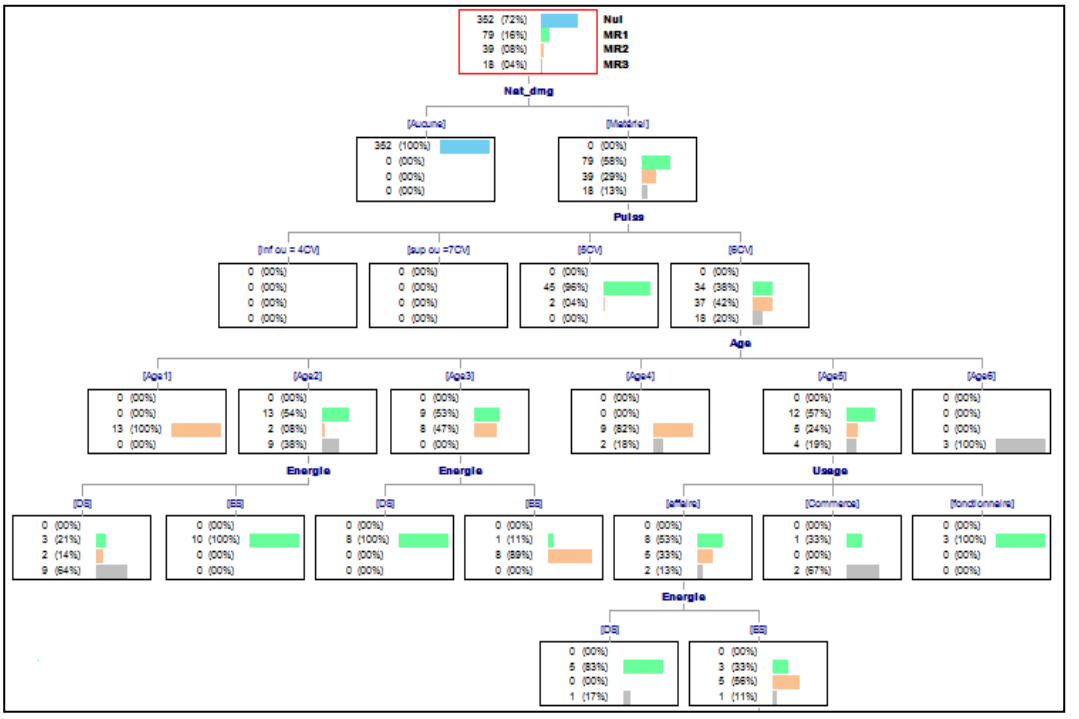

Figure 1: The C4.5 decision tree for the variable " claim cost"
Source: prepared by the authors using SIPINA software

Confusion matrix :

Table 4: Confusion matrix-1-

	Nul	MR1	MR2	MR3
Nul	352	0	0	0
MR1	0	76	2	1
MR2	0	3	34	2
MR3	0	1	2	15

Cost : 0.0225

Source: prepared by the authors using SIPINA software

The test error rate is equal to 2.3%, so we can say that when classifying an individual taken at random from the population, we have 2.3 chances out of 100 of making a wrong assignment. This low rate means that our model has good predictive power.

In order to fully understand the results derived from the tree, we explain two levels of the tree. The first vertex is called the "root" of the tree. It is located on the first level. Here we observe the frequency distribution of the variable to be predicted, " Claim cost ".

At the root of the tree, 72% of observations have been annotated "Null", 16% for the "MR1" class, 8% for the "MR2" class and 4% for the "MR3" class. The first segmentation variable chosen by the algorithm is "Nature of damage" (Nat_dmg): on the right vertex (2nd level), 45 individuals have a Material nature of damage, the proportion of Settlement Amount (MR1) = [1352.39; 27292. 65 [is 58%, the left node on the same level with a damage nature = None, concerns the Nul class, with a percentage of 100%. This vertex has no more child vertices, which is normal since it is "pure" from the point of view of the variable to be predicted.

According to the tree produced by SIPINA using method C4.5, the most significant variables in terms of the target variable Settlement Amount are:
Damage nature, Power of vehicle, Age of the policyholder, Use and Energy of vehicle.

4.2.2 The variable to be predicted "Claims frequency"

The application of C4.5 on the variable to be explained Number of claims enabled us to construct the decision tree illustrated in Figure 2.

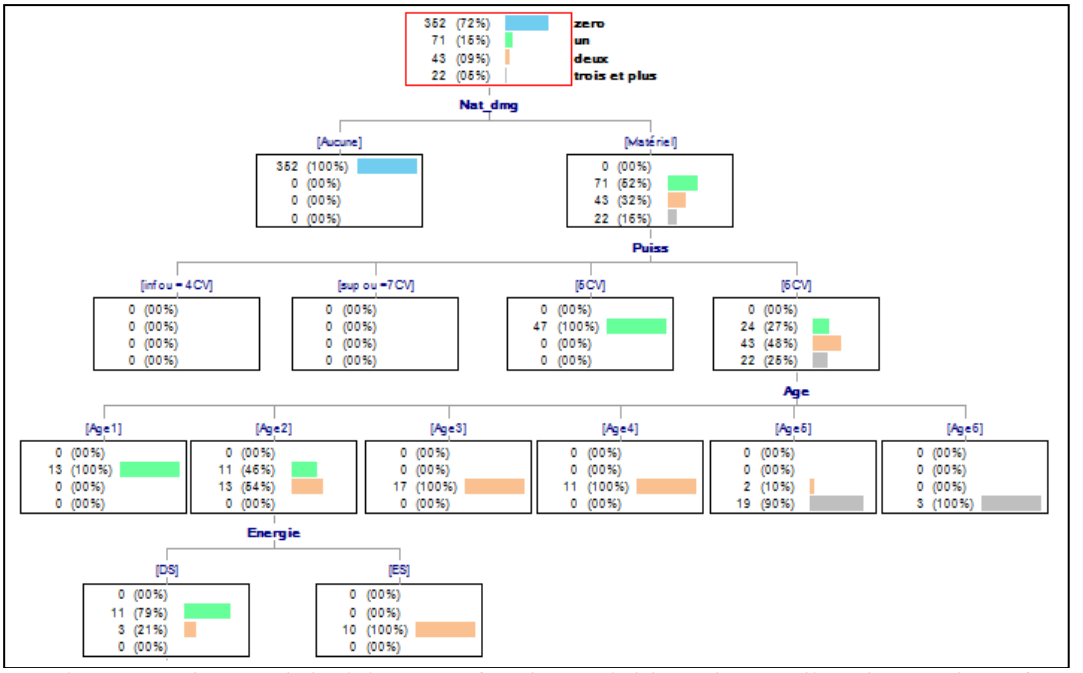

Figure 2: The C4.5 decision tree for the variable to be predicted "Number of claims"

Source: prepared by the authors using SIPINA software

Confusion matrix:

Table 5: Confusion matrix-2-

Cost : 0.0041

	zero	un	deux	trois et plus
zero	352	0	0	0
un	0	71	0	0
deux	0	0	41	2
trois et plus	0	0	0	22

Source: prepared by the authors using SIPINA software

The test error rate is estimated at 0.41%, so the model has perfectly classified almost all the observations in the sample (there are only two misclassified observations). According to the tree produced by SIPINA using method C4.5, the most significant variables in terms of the target variable Number of claims are: Damage nature, Power of vehicle, Age of policyholder and Energy.

4.3 The Generalized Linear Poisson Model

First and foremost, it is important to select the right explanatory variables for a generalized linear model (GLM). We are going to develop a methodology to optimize the explanatory variable selection phase in GLMs, i.e., in the absence of a theoretical model. First, we systematically search for all variables statistically linked to the dependent variable: the candidate variables.
Secondly, the intercorrelations between candidates are analyzed, in order to retain only those variables with a low rate of intercorrelation, i.e., the explanatory variables.

4.3.1. The choice of explanatory variables

Given a series of candidate variables, we look for the most relevant variables to explain and predict the values taken by the variable to be predicted.

4.3.1.1. Pre-selection of candidate variables

Our aim is to detect the relationship between the qualitative explanatory variables and the variable to be explained "Claims frequency". We carried out a bivariate analysis, testing the variables using SPSS software, which gave the following results: Age of driver, brand of vehicle, Power of vehicle, Number of seats, Premium and Nature of damage are dependent on the "Claims frequency" variable, while the other variables are independent.
According to the test used (Chi-square test), six variables are retained before the search for intercorrelations. However, these variables cannot yet be directly introduced into the regression. It is first necessary to study their correlations.

4.3.1.2. Eliminating highly correlated variables

Too high correlation between two explanatory variables is detrimental to the quality of the regression. It is therefore necessary to detect strongly inter-correlated variables. At the end of the previous phase, we examine the interdependencies between the six candidate variables, using Cramer's V statistical test. Traditionally, to establish whether there is an effect between the

two categorical variables crossed in a contingency table, we use the Chi-square test (χ^2). Cramer's V test is used to compare the strength of the relationship between the two variables under study.

$$V = \sqrt{\frac{\chi^2}{\chi^2 max}} = \sqrt{\frac{\chi^2}{n\times[\min(l,c)-1}}$$

The closer V is to zero, the less dependent the variables studied are. It will be 1 when the two variables are completely dependent. So, the closer V is to 1, the stronger the link between the two variables studied. We performed the Cramer's V test on the variables using SPSS software, giving the following results in Table 6:

Table 6: Results of Cramer's V test on candidate variables

Les variables	Age du Conducteur	Marque du Véhicule	Prime	Puissance du véhicule	Nature de dommage	Nombre de places
Age du Conducteur	**1**	**0.128**	**0.315**	**0.116**	**0.143**	**0.150**
Marque du Véhicule	**0.128**	**1**	**0.145**	**0.177**	**0.171**	**0.165**
Prime	**0.315**	**0.145**	**1**	**0.516**	**0.191**	**0.165**
Puissance du véhicule	**0.116**	**0.177**	**0.516**	**1**	**0.296**	**0.325**
Nature de dommage	**0.143**	**0.171**	**0.191**	**0.296**	**1**	**0.236**
Nombre de places	**0.150**	**0.165**	**0.165**	**0.325**	**0.236**	**1**

Source: prepared by the authors using SPSS software

Table 6 shows:
The absence of strong correlations between the candidate variables except between the two variables Vehicle Power and Premium. The following variables will therefore be retained directly: Age of driver, brand of vehicle, Nature of damage and Number of seats in the regression, and to decide which of the two correlated variables to retain for the Poisson regression, we decided to run the GLM Poisson procedure on each of the two variables and see which produces the best result.

4.3.2. Implementing Poisson Regression

We applied Poisson regression under **Xlstat** on the explanatory variables retained in the previous phase, once with the variable Vehicle Power and then with the variable Premium. On the basis of several model quality criteria, we decided to eliminate the Premium variable and keep the Vehicle Power variable. We report below the results, with comments, of the best Poisson regression:

4.3.2.1. Model evaluation:

Fit coefficients:
Table 7 shows a series of statistics for the independent and adjusted models.

Table 7: Evaluation of results

Statistique	Indépendant	Complet
Observations	488	488
Somme des poids	488,000	488,000
DDL	487	473
-2 Log (Vraisemblance)	1323,851	576,754
R^2(McFadden)	0,000	0,564
R^2 (Cox and Snell)	0,000	0,784
R^2(Nagelkerke)	0,000	0,839
AIC	1325,851	606,754
SBC	1330,041	669,608
Déviance	891,369	144,272
Khi² de Pearson	1200,153	177,859
Itérations	0	20

Source: prepared by the authors using Xlstat software

A- Les critères d'Akaike et Schwartz

Table 8 allows us to compare the model under study with the trivial "independent" model, reduced to the observation. The Akaike and Schwartz criteria are used to compare the two models. The best model is the one for which AIC and SBC are the lowest.

The result table shows that: AIC(Model) < AIC(Independent) for the AIC criterion and SBC(Model) < SBC (Independent) for the SCB criterion. This leads us to conclude that the model studied is better than the independent model. We deduce that the explanatory variables contribute to the explanation of the "Claims frequency" variable.

Le tableau 8, nous permet de comparer le modèle étudier avec le modèle trivial « indépendant », réduit à la constate. Les critères d'Akaike et Schwartz permettant de comparer les deux modèles. Le meilleur étant celui pour lequel AIC[1] et SBC[2] sont les plus bas.

B- Déviance

To assess the goodness of fit of a GLM, it is common in practice to calculate the Deviance statistic.
According to Table 7: Deviance (Model) < Deviance (Independent), then the model is good.

[1] Le critère d'information d'Akaike (Akaike's Information Criterion).
[2] Le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

C- R^2-like

R^2 (McFadden): coefficient between 0 and 1 that measures the goodness-of-fit of the model.
R^2 (Cox and Snell): coefficient between 0 and 1, measuring good model fit.
R^2(Nagelkerke): coefficient between 0 and 1, measuring the goodness of fit of the model.
When these values are close to 0, we say that the regression is not good. In our case, the regression model is good.

4.3.2.2. Parameter significance

Table 8: Parameter significance

Source	**DDL**	**Khi² (LR)**	**Pr > LR**
Puiss	3	9,750	0,021
Nbr _places	2	0,836	0,658
Marque	3	1,208	0,751
Natur_domg	1	706,426	< 0,0001
Age_conducteur	5	3,202	0,669

Source: prepared by the authors using **Xlstat** software

This table is only of interest if there is more than one explanatory variable. Here, we are testing the adjusted model against a test from which we would have removed the variable from the row of the table in question. If the probability Pr > LR is less than a set significance level (typically 0.05), then the variable's contribution to model fitting is significant. Otherwise, it can be removed from the model. The variables Damage nature and Vehicle power on the red background are the significant variables, the most significant being Damage nature. However, Number of seats, brand of vehicle and Age of policyholder are not significant. They can be removed from the model.

4.3.2.3. Parameter estimation

Model parameters
For each model constant and variable, the parameter estimates, corresponding standard deviation, Wald Chi² and corresponding p-value are displayed.

Table 9: Paramètres du modèle

Source	Valeur	Erreur standard	Khi² de Wald	Pr > Khi²
Constante	-20,621	1438,188	0,000	0,989
Puiss-5CV	0,000	0,000		
Puiss-6CV	0,363	0,183	3,925	0,048
Puiss-Inf ou = 4CV	0,209	0,230	0,824	0,364
Puiss-Sup ou = 7CV	-0,103	0,165	0,389	0,533
Nbr de place-2Plac	0,000	0,000		
Nbr de place-3Plac	0,181	0,318	0,326	0,568

Source	Valeur	Erreur standard	Khi² de Wald	Pr > Khi²
Nbr de place-4Plac	0,057	0,307	0,035	0,851
Marque-Autre	0,000	0,000		
Marque-Hyundai	0,122	0,155	0,624	0,430
Marque-PEUGEOT	0,005	0,140	0,001	0,971
Marque-RENAULT	0,144	0,165	0,763	0,383
Natur_domg-Aucune	0,000	0,000		
Natur_domg-Matériel	21,011	1438,188	0,000	0,988
Age_conduct-Age1	0,000	0,000		
Age_conduct-Age2	0,027	0,194	0,020	0,888
Age_conduct-Age3	0,213	0,174	1,494	0,222
Age_conduct-Age4	0,215	0,193	1,248	0,264
Age_conduct-Age5	0,234	0,186	1,583	0,208
Age_conduct-Age6	0,277	0,257	1,165	0,280

Source: prepared by the authors using Xlstat software

The blue boxes represent the coefficients of the modalities that do not contribute to the explanation of the target variable Claims frequency.

5. Comparison between decision tree and Poisson GLM according to target variable Claims frequency

Table 10: Comparison between the decision tree and the Poisson GLM

Decision tree	**Poisson GLM**
Relevant variables in descending order	
Damage nature Vehicle power Age of policyholder Vehicle energy	Damage nature Vehicle power

Source: prepared by the authors

It is clear that the explanatory variables extracted by the Poisson GLM are those that appear at the top of the decision tree.

6. Conclusion

In view of the serious accident situation in Algeria, we recommend that the variables we have obtained be taken into account to increase the value of the premium, in order to encourage policyholders to respect the highway code and reduce the devastation caused by these accidents.

8. References

Toseti, A., & al. (2000). *Comptabilité, réglementation, actuariat*. Ed Economica, Paris.

Besson, J. L., & Patrat, C. (2005). *Assurance non vie, modélisation et simulation*. Ed Economica, Paris.

Tuffery, S. (2012). *Datamining et statistique décisionnelle : l'intelligence des données* (4th ed.). Edition TECHNIP, Paris.

Rakotomalala, R. (2005). *Arbres de décision*. MODULAD, France. 25 pages, Num 3

WWW.SAA.dz

WWW.cnc.dz