Dataset of an actual motor vehicle insurance portfolio

Jorge Segura-Gisbert (jorge.segura@uv.es)
Universitat de Valencia

Josep Lledó (josep.lledo@uv.es)
Universitat de Valencia

Jose M. Pavía (pavia@uv.es)
Universitat de Valencia

Data Note

Keywords: Non-life insurance, Premium, Claims, Profitability

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

Dataset of an actual motor vehicle insurance portfolio

Jorge, Segura-Gisbert [1]

Josep Lledó [1]*

Jose M. Pavía[1]

[1]Universitat de Valencia, Applied Economics Departament, Av. dels Tarongers, s/n, 46022 València, Valencia Spain.

* Correspondence:
Josep Lledó
josep.lledo@uv.es

Abstract

Advanced analytics plays a vital role in enhancing various aspects of business operations within the insurance sector, providing valuable insights that drive informed decision-making, primarily through effective database utilization. However, open access databases in the insurance industry are exceedingly rare, as they are the basis of the business, encapsulating all the risk structure of the company. This makes it challenging for researchers and practitioners to access comprehensive insurance datasets for analysis. This paper introduces an extensive database specifically tailored for non-life motor insurance, containing 105,555 rows and encompassing a wide array of 30 variables. The dataset comprises important date-related information, such as effective date, date of birth of the insured, and renewal date, essential for policy management and risk assessment. Additionally, it includes relevant economic variables, such as premiums and claim costs, for assessments of products' financial profitability. Moreover, the database features an array of risk-related variables, such as vehicle size, economic value, power, and weight, which significantly contribute to understanding risk dynamics. By leveraging the statistical analysis of this rich database, researchers could identify novel risk profiles, reveal variables that influence insured claims behaviour, and contribute to the advancement of educational and research initiatives in the dynamic fields of economics and actuarial sciences. The availability of this comprehensive database opens new opportunities for research and teaching and empowers insurance professionals to enhance their risk assessment and decision-making processes.

Keywords
Non-life insurance; Premium; Claims; Profitability

1. Introduction

Open-access databases in the insurance industry are exceedingly rare. Companies are typically reluctant to share their data as it constitutes the core of their business. During a research project conducted by the authors within a Spanish insurance company, specialized in non-life insurance, we gained access to one of their motor vehicle insurance portfolio datasets and were authorized to share the anonymized data. The data is presented in the form of a structured table, providing valuable insights into a diverse range of non-life insurance policies. Besides finding these data valuable for our own research, we believe they are also of interest to other researchers and educators. This data can be used to identify novel risk profiles, reveal variables that influence insured claims behavior, and contribute to training future actuaries.

This paper presents and describes the database designed with a collection of 105,555 rows of data and encompassing a diverse set of 30 variables. Within this dataset, one can find indispensable date-related information, including the effective date of policies, the birthdates of insured individuals, and renewal dates. Furthermore, this dataset is enriched with pivotal economic variables, notably premiums and claim costs. These economic indicators are necessary for conducting in-depth analyses of the financial viability and profitability of motor insurance products.

These data can be used in several areas of an insurance company, being relevant for marketing purposes; including customer segmentation, contract renewal processes, price renewal strategies, optimization and price sensitivity models, as well as pricing mechanisms for new business. Researches can use these data to test risk selection mechanisms and the utilization of risk underwriting standards in the insurance sector. Specifically, it could be employed to examine how insurance companies determine the acceptance or conditional acceptance, as well as the rejection, of risks by adapting their own established underwriting standards.

These data sources originate from a non-life insurance company operating in Spain. Therefore, the data have been meticulously anonymized to align with the prevailing European legislation, safeguarding individual privacy and confidentiality. For ease of access and reference, a supplementary data file in spreadsheet format is provided along with this article. Researchers and stakeholders interested in exploring these datasets further can also access them via the provided DOI link: https://doi.org/10.17632/5cxyb5fp4f.1.

2. Data description

A sample dataset has been created from a Spanish insurance company specialized in non-life insurance. The dataset is presented as a spreadsheet file covering the main operations of the company during a period of three completed years (November 2015 - December 2018), containing several variables. It encompasses comprehensive information related to the unique characteristics of different lines of motor insurance,

diverse sales and distribution channels, important client-specific details, technical factors, and historical claims data.

A comprehensive quality control process has been conducted to ensure the reliability of the dataset. Steps have been taken to address missing, duplicated, or erroneous data, as well as to amend the presence of zero values with respect to the absolute frequency of each variable. This filtering stage holds significant importance for subsequent analyses. The process of extracting information from the dataset can be likened to a distillation process, involving a series of sequential stages. Within this process, the initial phase involves crucial tasks such as data cleaning, conditioning, and analysis of the database. These steps are vital for ensuring the dataset's integrity and suitability for analysis. Additionally, the process includes transforming or creating new variables based on existing information. It may also entail adjusting date formats and determining qualitative variables as factors, allowing for further analysis and interpretation.

3. General variables of each policy

Each entry in the datasets represents a distinct insurance contract or policy, with varying maturity dates. Consequently, each policyholder can have multiple rows in the dataset, corresponding to annual observation periods, except for policies initiated less than a year prior to the valuation date. Utilizing an internal identification code (ID, see Table 1), all policyholder-product transactions can be accurately recorded. The monetary values considered, such as premiums, are provided after deducting applicable taxes. The dataset comprises 105,555 rows and 30 columns, as indicated in Table 1. Each row in the datasets represents a policy during a period, and each column corresponds to a specific variable.

Table 1. First variables in the dataset.

Variable	Description
ID	Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
Date_start _contract	Start date of the policyholder's contract (DD/MM/YYYY).
Date_last_renewal	Date of last contract renewal (DD/MM/YYYY).
Date_next_renewal	Date of the next contract renewal (DD/MM/YYYY).
Distribution_channel	Classifies the channel through which the policy was contracted. 0 for Agent and 1 for Insurance brokers.
Date_birth	Date of birth of the insured declared in the policy (DD/MM/YYYY).
Date_driving_licence	Date of issuance of the insured person's driver's license (DD/MM/YYYY).
Seniority	Total number of years that the insured has been associated with the insurance entity, indicating their level of seniority.
Policies_in_force	Total number of policies held by the insured in the insurance entity during the reference period.
Max_policies	Maximum number of policies that the insured has ever had in force with the insurance entity.
Max_products	Maximum number of products that the insured has simultaneously held at any given point in time.

Lapse	Number of policies that the customer has cancelled or has been cancelled for nonpayment in the current year of maturity, excluding those that have been replaced by another policy.
Date_lapse	Lapse date. Date of contract termination (DD/MM/YYYY).
Payment	Last payment method of the reference policy. 1 represents a half-yearly administrative process and 0 indicates an annual payment method.
Premium	Net premium amount associated with the policy during the current year.
Cost_claims_year	Total cost of claims incurred for the insurance policy during the current year.
N_claims_year	Total number of claims incurred for the insurance policy during the current year.
N_claims_history	Total number of claims filed throughout the entire duration of the insurance policy.
R_Claims_history	Ratio of the number of claims filed for the specific policy to the total duration (whole years) of the policy in force. It provides an indication of the policy's claims frequency history.

The variable ID corresponds to a unique identification code that is assigned to the combination of policyholder and product. For instance, if there are two rows in the dataset with the same ID, it indicates that the policyholder is insured for two distinct time periods (years) while maintaining the same product. The subsequent group of variables, namely Date_start_contract, Date_last_renewal, and Date_next_renewal, are associated with important dates relevant to the insurance policy. Among these variables, Date_start_contract represents the start date of the policyholder's contract, Date_last_renewal indicates the date of the policyholder's most recent renewal, and Date_next_renewal represents the upcoming renewal date for the insurance policy. Date_birth is less than 31/12/2000 and greater than 01/01/1918 and the variable Date_driving_licence indicates the time when the legal driving license was obtained, with the minimum age requirement being 18 years in Spain. The modal year for this last variable is 2006. The format of all dates in the dataset is DD/MM/YYYY, where DD, MM, and YYYY represent the day of the month, month of the year, and year, respectively.

The Distribution_channel classifies the method by which the company got the policy, with a value of 0 indicating an agent channel and a value of 1 representing insurance brokers. The agent channel refers to policies contracted directly through an agent, while the insurance broker channel refers to policies contracted through independent insurance brokers. 60% of the policies are coded with a value of 0 for this variable.

The following group of variables describes the policyholder's affiliation with the insurance company. Figure 1 illustrates the distribution of variables Seniority, Policies_in_force, Max_policies and Max_products. Specifically, the variable Seniority (Figure 1 left upper panel) indicates the length of time, in years, that the policyholder has been associated with the insurance company. It represents the policyholder's seniority or tenure within the company with an average duration exceeding 6 years. The variable Policies_in_force (Figure 1 right upper panel) quantifies the number of policies held by each policyholder within the company. On average, each policyholder maintains more than 1 policy with the company. The variable Max_policies (Figure 1

left lower panel) build upon the previous information by measuring the maximum number of products that each policyholder has held within the company. The difference between Max_policies and Policies_in_force represents the lapse in the portfolio over time. It captures the extent to which the policyholder has reduced the number of products they have with the company, indicating a potential contraction in their insurance coverage or a shift in their needs and preferences. Finally, the company offers a total of 4 different products (car, household, commerce and personal accident) for its policyholders. However, a significant majority of the policyholders, more than 99% of them, have had a maximum of 2 products (see, right lower panel in Figure 1).

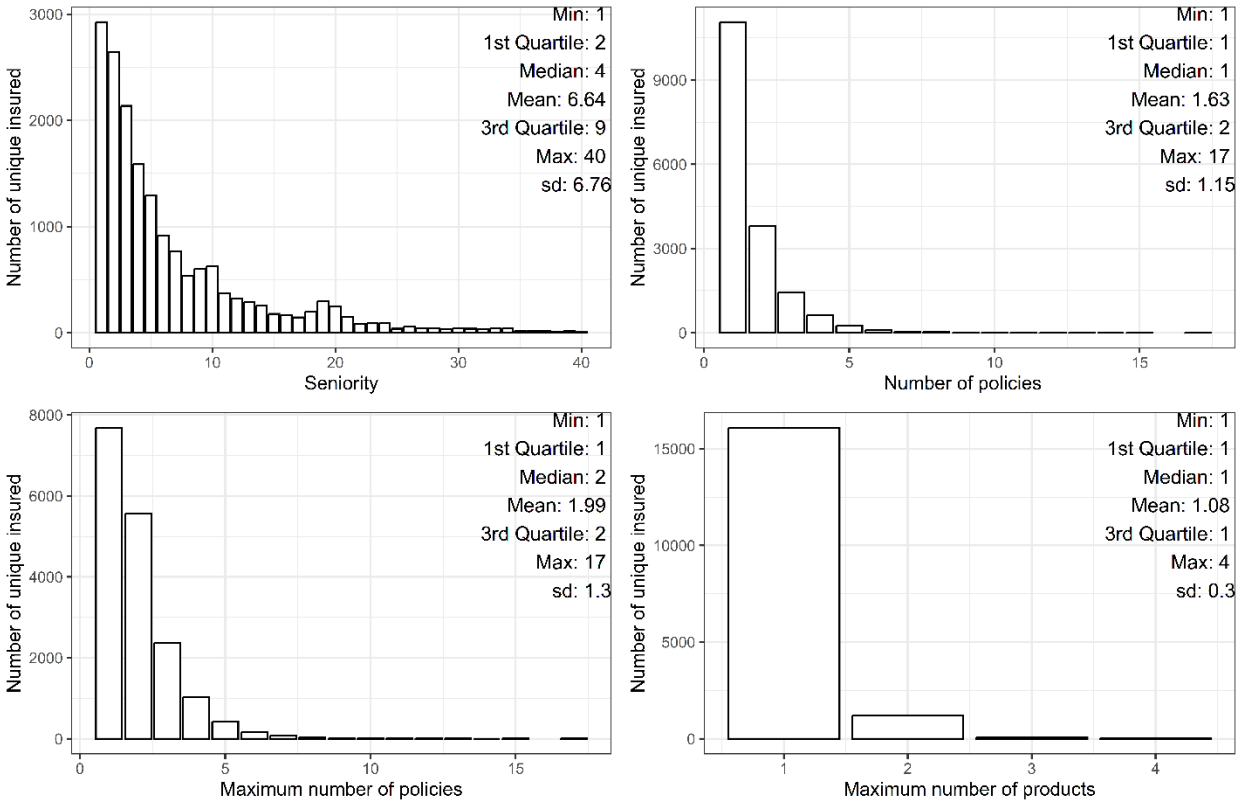

Figure 1. A graphical summary of policyholders' affiliation with the insurance company. Top left panel displays the distribution of the number of years of seniority. Top right panel shows the distribution of the number of policies that policyholders have with the insurance company. Lower left panel presents the distribution of the maximum number of policies per insured while the lower right panel exhibits the maximum number of products held by policyholders. In all panels, the year of the last renewal is 2018.

The variables Lapse and Date_lapse deserve special attention in every insurance portfolio. When a policyholder stops paying premiums and the account value of the insurance policy has already been exhausted, the policy lapses. Date_lapse only

activates under that event. When this happen, the company has the option to terminate the contractual relationship and the insured can make an independent decision of no renewing, citing various reasons for cancelling the coverage. These reasons may include, among others, no longer requiring the product, finding better prices from competitors, or the product not offering the necessary guarantees. Lapse indicates either non-permanence or permanence and can be used to measure the number and the date of policies that the company retains (does not retain). Someway related to these variables is the Payment variable, which captures the frequency of the payment method. A value of 0 indicates an annual payment method and a value of 1 a semi-annual payment method.

The following section collects quantitative variables that are related to income and expenses linked to the portfolio. The Premium variable (income) quantifies the amount received by the insurer in connection to a policy during the current year after deducting taxes but including both internal (administrative) and external (commercial) expenses. Figure 2, left panel, shows a plot density of the Premium variable, which presents an asymmetric (positive) distribution. On the other hand, Figure 2, right panel, illustrates the distribution of Cost_claims_year variable, which exhibits a comparable pattern to premium figures after excluding policies with 0 claims. The main statistics are accompanied by the right-hand panel, which also displays the average cost per claim (mean claim). This average is calculated by adding all amounts together and dividing them by the total number of claims. All the figures are in euros.

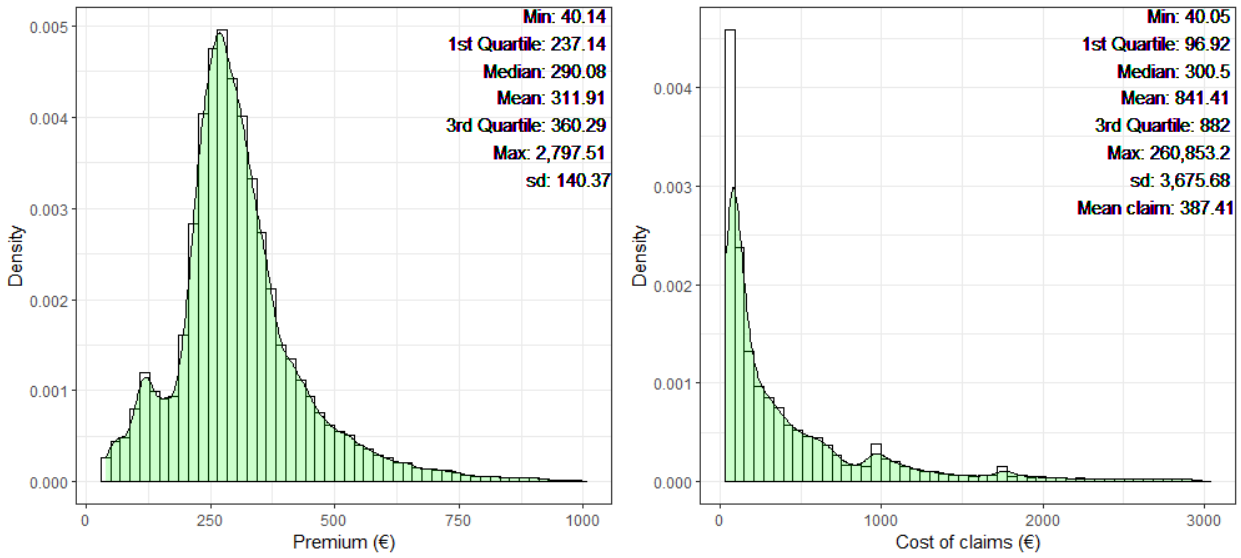

Figure 2. Graphical representation of income and expenses distributions. A density estimation of the amount of insurance premiums is displayed for the range 0-1000 in the left panel. In the right panel, a density estimation of the cost of policies that have experienced a claim is presented for the range 0-3000, but excluding claims linked to CICOS. The CICOS agreement in Spain facilitates handling claims from vehicle collisions, aiming to expedite compensation for material damage. It sets a predetermined amount agreed upon by the

involved insurance companies, regardless of the actual claim, with costs rising from around €800 in 2019 to over €1,000 in 2023. In both panels, the year of the last renewal is 2018. As can be observed, the summary statistics included in each panel have been without imposing any restrictions on the upper amounts.

Finally, the last set of variables describes characteristics related to the policy and the number of claims. These variables may encompass various factors and attributes that can influence or provide insights into the claim's behaviour of the insured. The variables N_claims_year and N_claims_history represent the number of claims that the insured has made during the current year and throughout the entire history of the policy, respectively. The R_Claims_history is defined as the ratio of the number of claims filed for a particular policy to the total duration of the policy in force, measured in whole years. This ratio serves as an indicator of the policy's historical claims frequency.

4. Factors determining risk associated of vehicles/insured

The previous variables (Table 1) establish the primary characteristics of the policy and serve as the main economic indicators. Table 2 set of variables, available in the dataset just after the ones described in Table 1, enables a comprehensive analysis of the risk undertaken by the insurance company.

Table 2. Second set of variables in the dataset.

Variable	Description
Type_risk	Type of risk associated with the policy. Each value corresponds to a specific risk type: 1 for motorbikes, 2 for vans, 3 for passenger cars and 4 for agricultural vehicles
Area	Dichotomous variable indicates the area. 0 for rural and 1 for urban (more than 30,000 inhabitants) in terms of traffic conditions.
Second_driver	1 if there are multiple regular drivers declared, or 0 if only one driver is declared.
Year_matriculation	Year of registration of the vehicle (YYYY).
Power	Vehicle power measured in horsepower.
Cylinder_capacity	Cylinder capacity of the vehicle.
Value_vehicle	Market value of the vehicle on 31/12/2019.
N_doors	Number of vehicle doors.
Type_fuel	Specific kind of energy source used to power a vehicle. Petrol (P) or Diesel (D).
Length	Length, in meters, of the vehicle.
Weight	Weight, in kilograms, of the vehicle.

The variable Type_risk allows the company to distinguish between various insured risk profiles: 1 represents motorbikes (8,502 rows, 8.1%), 2 corresponds to vans (13,212 rows, 12.5%), 3 denotes passenger cars (82,990 rows, 78.6%), and 4 pertains to agricultural vehicles (851 rows, 0.8%). The risk profiles are insured based on their predominant Area of circulation, with the rural area accounting for over 70% of the policies. Additionally, there is the possibility of including multiple regular drivers (

Second_driver) in the policy. Specifically, over 10% of the policies include at least a second regular driver.

The last set of variables of Table 2 defines the specific and technical characteristics of the insured vehicles. Firstly, Year_matriculation shows a prevalence of the years 2004-2007, reflecting the economic expansion before the onset of the financial crisis in 2008. The variables Power and Value_vehicle exhibit a wide range of numerical values and may demonstrate a degree of linear correlation, depending on the specific variable under analysis. Figure 3 illustrates the relationship between these two variables based on the Type_risk variable. Power is also typically correlated with Cylinder_capacity due to the power of a vehicle's engine often corresponds to its cylinder capacity, indicating a strong relationship between these two factors.

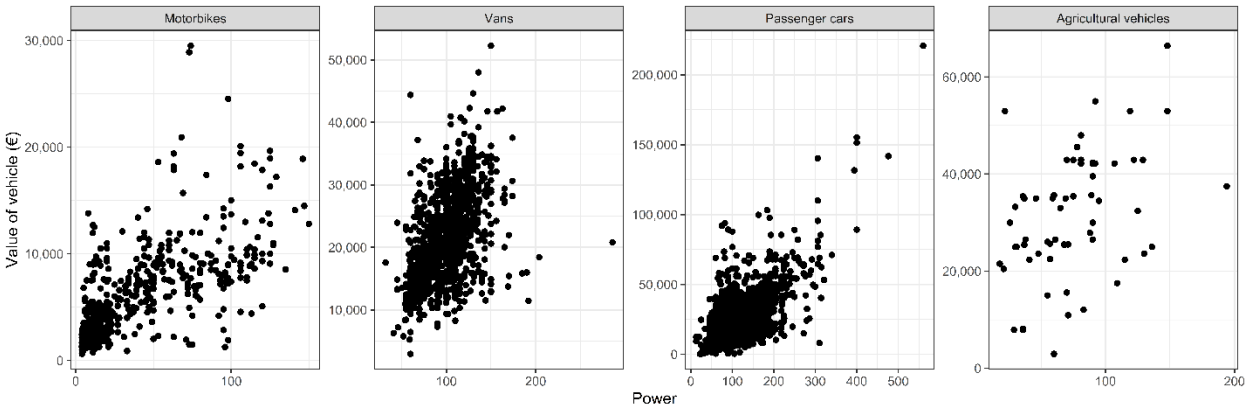

Figure 3. Relation between Power and Value_vehicle categorized by Type_risk.

Continuing the risk description, five new variables provide technical details about the vehicles. For motorbikes and agricultural vehicles, the standard value for N_doors is 0, whereas for vans and passenger cars, it is typically 5. Within the database, two different fuels are identified with the variable *Type_fuel*: petrol, accounting for 36.8% of the rows, and diesel, comprising 61.6%. Additionally, there is a total of 1.7% categorized as "NA" indicating that the fuel type is not available for these entries. The variables Length and Weight enable the determination of the risk size. Averages of Length and Weight are the 4.25 meters and 1,191 kilograms, respectively.

5. Exploring the Insurance Industry

The insurance industry plays an indispensable role in driving and supporting modern economies. At its core, the purpose of insurance is to safeguard agents and individuals against adverse events that may lead to economic losses or repercussions. This pivotal function is realized through insurance policies, which aim to provide compensation or indemnity in the event of a loss. In the digital age, the insurance sector is undergoing transformative changes, and a key driving force behind this transformation is the utilization of data. Entities within the industry are

leveraging data from various sources, both internal and external, to enhance their decision-making processes. By combining diverse information assets, insurers gain a competitive advantage, fostering innovation and sector-wide transformation.

Notably, customer behaviour and risk aversion vary among individuals in the insurance industry. Therefore, analysing the profitability of customers and policyholders becomes crucial for shaping and adapting business strategies. Insurers are wary of consistently facing losses from policyholders, making it essential to identify and engage with customers who contribute positively to the bottom line. [1] highlight, the true measure of success lies not in the sheer number of customers in a portfolio, but in the quality of the customers and the profitability of the relationships established.

The availability of open access data concerning insured populations is currently limited. In life insurance, a vast database containing valuable information can be accessed at [2] concerning ages, risk capital and quarterly values. In the realm of non-life insurance, there are existing databases that pertain to particular areas, car insurance, property insurance, and health insurance. Regarding car insurance, the database [3] contains valuable information based on upheld consumer complaints as a percentage of their total business over two years. Insurers with the lowest complaint ratios rank higher on the list, while those with the highest ratios are placed at the bottom.[4] determines the number of motorists who have completed the course and organize the data into subsets based on the year and month of course completion, motorist residency, age, and sex, as well as the course provider and delivery method. Similarly, [5] and [6] are two datasets for property and health insurance sector respectively. Nevertheless, none of the existing databases matches the comprehensive nature of the one developed in this study.

The dataset utilized in this paper comprises information from an insurance company in Spain, spanning the period from 2015 to 2018. The insurance policy is renewable annually and is automatically renewed until either the insured opts to cancel or the company cancels due to non-payment. The database inherently includes records of new business entries and cancellations within the insured's portfolio. The reference policy's frequency and claim costs consist of various claim types, such as material damage, body injury, vehicle all-risk coverage, travel assistance-related claims, glass amounts, and vehicle theft. Regarding the insured person's premiums, these are comprised of the total sum of payments for different guarantees and products included in the policy.

Apart from the variables available in the dataset, additional variables necessary for various analyses, such as client profitability, can be derived. Client profitability can be determined by considering the premiums paid during a specific period and the claims incurred, considering both closed file (claims already settled) and provisioned file (claims yet to be settled). Non-life insurance often involves calculating the claims

frequency, which is the ratio of the number of claims to the exposure period of the insured contract in days. Another significant calculation is the claims ratio, expressed as a percentage, which indicates the portion of premium consumed by the claims experience. A ratio exceeding 100% signifies losses in the technical aspect of the business.

Each of the mentioned economic indicators can be applied to analyse and measure the variables that influence claim behaviour. Moreover, these indicators can be utilized to classify and group variables based on their detection patterns using machine learning techniques. By employing classification methods, we can categorize the variables effectively, allowing for a deeper understanding of their impact on the claims process. These techniques offer valuable insights into the complex relationships between the variables and the overall insurance claims behaviour. Finally, it is worth noting that while these databases provide valuable insights, their coverage might not be exhaustive and may have certain limitations. As the insurance industry increasingly embraces data-driven approaches and digital transformations, the expansion and enrichment of these databases, along with the inclusion of other lines of business, become essential for insurers to make more informed decisions, create personalized policies, and enhance risk management strategies.

Declarations

Ethics approval and consent to participate

The relevant informed consent was obtained by the company from the insured in the moment of contracting the product. Data is offered anonymised.

Consent for publication

Not applicable.

Competing interests

The authors declare to have no known competing financial interests or personal relationships that may have influenced the work reported in this paper.

Availability of data and materials

All the data used are openly accessible in the supplementary material and have been appropriately described within the paper.

Funding

The authors acknowledge the support of Generalitat Valenciana through project AICO/2021/257 (Consellería de Educación, Universidades y Empleo) and Ministerio de Economía e Innovación through project PID2021-128228NB-I00.

Authors' information

Jorge Segura-Gisbert: Methodology, Software, Resources and Data curation. **Josep Lledó**: Methodology, Software, Visualization, Data curation, Investigation, Writing–original draft and editing; **Jose M. Pavía**: Conceptualization, Methodology, Supervision, Funding acquisition, Investigation, Writing–review & editing.

References

1. Venkatesan, R., Kumar, V. A customer lifetime value framework for customer selection and resource allocation strategy. In: Journal of Marketing. 2004. Vol. 68 (4). p. 106–125.
2. Lledó, J. and Pavia, J.M. Dataset of an actual risk-life insurance portfolio. In: Data in Brief. 2022. 45. 108655. doi: 10.1016/j.dib.2022.108655.
3. New York State Department of Financial Services. DFS automobile insurance company complaint rankings: beginning 2009. In: NY Open Data. 2022. https://links.uv.es/lQtNU63
4. New York State Department of Financial Services, Motor vehicle point & insurance reduction program (PIRP) participation: Five-year window. In: NY Open Data. 2022. https://links.uv.es/qYWU9Tg
5. New York State Department of Financial Services, Total property insurance premiums written annually in New York: Beginning 1998. In: NY Open Data. 2022. https://links.uv.es/bhc9akR
6. Department of Health and Human Services. Health insurance marketplace: summary enrollment data for the initial annual open enrollment period. In: ASPE, issue brief. 2021. https://links.uv.es/TLbW7X5

Figures

Figure 1

Figure 2

Figure 3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tabla1y2.xlsx
- Motorvehicleinsurancedata.csv
- Motorvehicleinsurancedata.csv