Proceedings of ICFTBA 2024 Workshop: Human Capital Management in a Post-Covid World: Emerging Trends and Workplace Strategies
DOI: 10.54254/2754-1169/113/2024LD0103

Features to Involve and Collect in Future Auto Insurance Pricing

Shuang Guo[1,a,*]

[1]*Olin Business School, Washington Universality in St. Louis, St. Louis, United States*
a. serguo@deloitte.com.cn
**corresponding author*

Abstract: The property insurance industry is a highly data-driven sector. In the traditional model, insurance pricing largely depends on the self-reported risk characteristics of clients, such as age, gender, vehicle type, etc. These characteristics help insurance companies categorize clients into different risk levels and set premiums accordingly. With the development of telematics, insurance companies can now collect and analyze more dynamic data that is more directly related to risk. This study presents a comprehensive analysis of three feature selection models for vehicle insurance pricing. Utilizing regression techniques, it evaluated a multitude of factors believed to influence claim frequency and severity. Through a rigorous comparative study, key factors that significantly impact the vehicle are assessed and fit to different regression models. The article's findings indicate that factors such as policy bonus, mileage base data, and car usage are paramount in determining insurance rates. Moreover, the study revealed that incorporating these factors into the pricing model enhances its accuracy and fairness.

Keywords: Frequency-Severity Model, Machine Learning, Motor Insurance, Feature Selection, Anti-discrimination.

1. Introduction

The global commercial property and casualty insurance market is experiencing robust growth, yet it is also confronted with intensifying challenges [1]. Based on the present, insurance companies can leverage advanced data analysis technologies to more accurately identify and assess risks, thereby providing more personalized and precise insurance services to customers. This not only helps to enhance the risk management capabilities of insurance companies but also improves customer satisfaction and loyalty. At the same time, insurance companies also need to ensure that they comply with relevant laws and regulations in the process of collecting and using data, protecting the privacy rights of customers.

In car insurance, driving behavior data collected through onboard devices, such as driving speed, rapid acceleration, sharp turning, and the frequency of emergency braking, can more accurately reflect a driver's risk level. In this Methodologically, the study of Henckaerts & Antonio developed a baseline pricing model on a large portfolio with only self-reported features then propose an explainable updating mechanism to incorporate driving behavior information into the baseline tariff, which resulting in a better assessment of claim risk for both the in-sample train and out-of-sample test data [2].

DOI: 10.54254/2754-1169/113/2024LD0103

Since the outbreak of COVID-19, with the rise of the Diversity, Equity, and Inclusion (DEI) movement, the insurance industry is also facing the challenges of pricing discrimination and the need of adjusting risk classification characteristics. However, even though characteristics like sex and race are banned in some countries, David demonstrate in his research that technology may offer a solution as innovations in data analytic offer great promise for reducing the adverse effects of discrimination. As the use of telematics in auto insurance has demonstrated, innovative insurers are steadily employing new types of data to minimize this discrimination [3].

Traditionally, applied statistical modeling can be used to identify the relationship between examined variables and traffic collisions, these statistical techniques, however, have limitations with highly nonlinear data with the increasing complex data provide valuable insights into the complex interactions among traffic elements and traffic crashes, suggesting machine-learning technology and an improved deep learning model [4].

The collective evidence from Chen and Dewi's study points towards that Random Forest algorithm has become an extremely useful and efficient tool among some usually used classifiers like Support Vector Machines (SVM), K-Nearest neighbors, and linear discriminant Analysis (LDA) [5]. It is capable of handling complex datasets with a large number of variables, effectively identifying and utilizing key features to enhance the accuracy and generalization of models, possessing excellent noise tolerance. Taha, Cosgrave & Mckeever also proposed a framework for identifying and selecting features before applying predictive analytical algorithms which are the Greedy Feature Selection Algorithm (GFSA), the Laplacian Score (LS), the Spectral Algorithm (Spec), the Unsupervised Spectral Feature Selection Method (USFSM) [6]. The framework successfully identifies features like gender, location and call time are influencing the downstream accuracy gains and bring insights to assist business decision making.

As mentioned above, traditional motor insurance pricing model may lead companies to be unable to accurately classify insurers through characteristics and resulting in mistakenly grouping insurers with heterogeneous risks. Therefore, this paper aims to analyze and select a multitude of factors that are crucial to model results. Then applying analysis on the selected factors to find a trade-off between the significance of eliminating risks and the potential of discrimination. Based on the above analysis, insurance companies can gain insights from the potential alternatives to optimize pricing model and improve effectiveness of risk management with the regulatory restrictions.

2. Methodology

2.1. Datasets

Data is provided by the French institute of Actuaries in November 2017 used for pricing game. The datasets pg17trainpol and pg17trainclaim are sourced from the R package CASdatasets [7]. The datasets encompass 100,000 policies for private motor insurance and 14,243 claims of those policies and they are suitable for generating feature selection because they include 30 variables in total which provide a larger pool to make selection. Variables are roughly characterized into driver, secondary driver and vehicle's basic information. Detailed descriptions are included in CASdatasets manual [7]. It is important to note that the datasets have limited demographic diversity within French.

2.2. Features

2.2.1. Feature Preparation and Transformation

In these datasets, two factors that required in further models are pre-calculated. Frequency, as one of the target variables is generated by counting client_id in pg17trainclaim which is positive integer. Average claim size is also generated by summing up claim_amount with unique id_client than divided

DOI: 10.54254/2754-1169/113/2024LD0103

by Frequency. Removing negative values which come from legal recourse where driver's liability is not engaged, the target variable, claim amount, is continuous positive real number.

Besides, categorical variables with more levels impact model efficient significantly, so factor variables require regrouping to new levels. For example, vehicle carmaker contains 101 levels but there are only three major brands which are Renault, Peugeot and Citroen. Thus, the variable can be split into 3 major levels and one level represents the rest.

Moreover, pol_insee_code represents where the policyholder lives but it is in form of meaningless. The study counted the number of each 2-digit insee code to get the population density as one feature of the model, which may be highly correlated to the vehicle claims.

2.2.2. Feature Selection

Univariate Feature Selection is a relative straightforward approach to select features. As these datasets have many variables and most of them are expected to be linear correlated with targets, univariate selection is a suitable and efficient way to assess individual feature importance with high interpret ability. Various statistical measures are used to define the relationship, including correlation coefficients for numeric variables and chi-squared test for categorical variables. Variables are ranked based on their significance and strength of correlation. However, as it does not account for interactions, some potential complex association may not be captured.

Considering univariate selection focus on individuals' straightforward relationship with targets, this study suggests another feature selection method, Random Forest. Unlike feature selection methods based on statistical measures, random forest selection capture complex non-linear relationships and interactions between high-dimensional variables and targets. Also, as random forest combines multiple single trees and calculate in parallel, it is also efficiently to handle large datasets. Even random forest is a power technique, it is still needed to be aware of the probability of overfitting.

Stepwise selection is also a statistic approach use to select significant features based on the a common criterion like p-value. As the datasets include plenty of variables, it is commonly suggested to use backward selection. Starting with a full model including all potential variables, Backward Selection removes one feature at a time and assesses the impact on model performance until the model meets the minimum AIC or BIC. Backward feature selection can avoid overfitting by eliminating less important features and may improve its predictive performance by focusing on more correlated variables. Model

2.2.3. Frequency-Severity Model

This paper adopts the frequency-severity model as the approach to evaluate the selected feature performance on the target.

Frequency-Severity Model is commonly used for predicting expected number of claims and average claim amount of each with in a certain time period. It is based on historical data and generates aggregate claim amount by timing these two components:

Frequency, means the number of claims occurred. In these datasets, frequency is calculated by counting id_claim in pg17trainclaim.

Severity, means the average amount of claims. In these datasets, severity is the average claim amount which is aggregate claim amount in pg17trainclaim averaged by frequency.

The model defines N_{it} to indicate the number of claims of the i-th policyholder in the t-th policy year, and Y to indicate the claim amount of each occurrence.

$$Y_{it} = \begin{cases} (Y_{it,1}, \ldots, Y_{it,N_{it}}) & N_{it} > 0 \\ \text{No claim} & N_{it} = 0 \end{cases} \quad (1)$$

DOI: 10.54254/2754-1169/113/2024LD0103

Furthermore, the aggregated claim amount S_{it} and average claim amount A_{it} are denoted by [8]:

$$S_{it} = \begin{cases} \sum_{j=1}^{N_{it}} Y_{it,j} & N_{it} > 0 \\ 0 & N_{it} = 0 \end{cases} \quad (2)$$

$$A_{it} = \begin{cases} \frac{\sum_{j=1}^{N_{it}} Y_{it,j}}{N_{it}} & N_{it} > 0 \\ 0 & N_{it} = 0 \end{cases} \quad (3)$$

2.2.4. Regression

As the study adopts Frequency-Severity Model, two separated models are constructed with target variables, Frequency and Severity. It is a common actuarial assumption that claim frequency as a positive integer usually has a Poisson distribution or negative binomial distribution, and the distribution selected for modeling Severity are Gamma distribution and Pareto distribution. Besides, pol_sit_duration, as the exposure, is added to be an offset, for the reason that longer the policy is in force, the more probability there is for a claim that would trigger.

Predictors selected from the above feature selection methods are used as the predictors of each target. The regression models are model US (Univariate Selection), model RF (Random Forest), model SS (Stepwise Backward Selection), and also model C (Combined), which includes predictors that are selected by two or more methods in order to balance the pros and cons. Model SS covers the least features which exclude vehicle speed, vehicle type, vehicle weight and pol_insee_code. Also, the major difference between model US and model RF is the information of secondary driver.

3. Results

3.1. Model Comparison

The models are evaluated based on RMSE, the root mean square Error. It is used to measure the gain and loss on prediction performance between different models as shown in Table 1.

Table 1: Model Performance

Model	N		US		RF		SS	
Frequency/Severity	Gamma	Pareto	Gamma	Pareto	Gamma	Pareto	Gamma	Pareto
Poisson	640.89	642.91	640.77	642.90	642.93	646.82	642.91	646.92
Negative Binomial	641.86	642.82	641.77	642.79	647.86	648.24	647.86	648.26

Overall, the performance of the models has not much difference. Models with Poisson distribution to fit frequency has relative lower errors than models with negative binomial distribution, which may due to the events are relative isolated in the datasets and negative binomial is better to handle the clustering effect of concentrated events. Unsurprisingly, as there are some extreme high claim amounts in datasets, gamma distribution overall provides better prediction because it can handle skew and long-tail distribution to avoid unreasonably high claims. In addition, from the perspective of feature selection, model C includes 7 fewer variables than model US but only result in 0.12 higher error, which can be a better choice to avoid overffiting and breach of regulation. Variables that are selected by only one method has also been examined, and they are not significant in regression.

Thus, using the combined selected features as the variables to predict frequency and severity respectively by fitting Poisson and gamma distributions brings the optimal model.

DOI: 10.54254/2754-1169/113/2024LD0103

4. Discussion

4.1. Frequency Perspective

While model C is employed to explore the relationship between the selected variables and frequency, most of the variables are considered to be significant with a 99% confidence interval. As shown in Table 4, among all the variables, policy bonus, policy coverage, policy duration, pay frequency, vehicle usage, vehicle speed limit, vehicle fuel type, driver and passenger age, and vehicle age are found to be statistically significant with a p-value less than 0.001, indicating a strong association with the claim frequency, as shown in Table 2.

pol_bonus. Policy bonus which is specifically used in French bonus/malus system indicates a discount percent attached to the driver. Every time the driver gets involved in a claim, the value increases, so that the higher pol_bonus, the higher possibility that the driver would cause a claim, given the large magnitude of the effect.

pol_coverage. Policy coverage has a baseline of maxi, which covers all claims including damage like theft and assistance. The baseline covers all claims including damage, theft, and assistance. Compared to maxi, other levels naturally have a lower frequency of claims due to less coverage.

pol_sit_duration. The policy sit duration of the insurance has a significant impact on the frequency of claims, but the negative coefficient is unexpected. This may be because customers who have held insurance for a long time are more familiar with the insurance terms and claims process, leading them to drive more cautiously and to deliberate more carefully before reporting. Additionally, insurance companies often offer discounts to customers who have not made claims for a long period, which undoubtedly reduces the reports of claims.

pol_pay_freq. The semi-annual payment frequency is the baseline in this model, and it is evident that annual premium payments result in a significantly lower claim frequency compared to other levels (yearly<semi-annually<quarterly<monthly). Although it is generally to believe that the behavior of paying premiums should not significantly affect the claim frequency, insurance companies may offer different discounts or benefits for annual premium payments, which may indirectly affect the reporting of claims. To maintain long-term low costs, drivers are more motivated to reduce risky behaviors that could cause claims or to reduce the number of claims reported.

pol_payd. Mileage based policy is a recent insurance product which considering real time drive behaviors and reflect more accurate risk exposures. It is important to note that even though mileage-based indicator is not significant in this model (This may because mileage-based data only begun to emerge in recent years with less experience), this data should have important implications for pricing and risk management in short future.

pol_usage. Policy usage includes 4 levels, Alltrip (baseline), Professional, Work Private and Retired. The negative coefficients of the three levels beside baseline support cars used for trip are more likely to be involved in accidents than those used for working or retiring life.

vh_speed. Great performance and speed for a car might lead to higher risk of accidents, thus affecting claim frequency. However, it is unreasonable to say the maximum speed the car can access directly affect the driver's behavior and lead to traffic accidents.

drv_age1, drv_age2. Consequently, as age increases, there is a general decline in reaction time and energy levels, which may contribute to a higher accident rate. Additionally, it is reasonable to find that the elder age of passengers, the less claim frequency occurs, because drivers usually tend to be more cautious when there are elderly passengers in the car to avoid sudden acceleration or hard braking to reduce discomfort for passengers.

vh_age. There might be bias or discrimination within this variable and indirectly affects the claim frequency. Older vehicles may be more frequently used by drivers who are skilled or more cautious, and these drivers may inherently have a lower accident rate. Also, older vehicles may be used less

DOI: 10.54254/2754-1169/113/2024LD0103

frequently, thus covering fewer total miles and reduce the frequency. Mileage base data should be considered for better interpretation.

vh_fuel. Finally, the high degree of significance on vehicle fuel type is surprisingly unreasonable. One potential explanation is that diesel engines typically have higher torque and normally used in larger vehicles, which could affect the likelihood of accidents indirectly. However, fuel type is not suggested to consider as a good feature to predict claim frequency.

Table 2: Significant Model C Coefficients - Frequency

	Estimate	Std. Error	T value	Pr (>|t|)	Significance
(Intercept)	-1.9773	0.2632	-7.5100	0.0000	***
pol_bonus	0.7611	0.0966	7.8800	0.0000	***
pol_coverageMedian1	-0.2490	0.0453	-5.5000	0.0000	***
pol_coverageMedian2	-0.1975	0.0317	-6.2300	0.0000	***
pol_coverageMini	-1.0443	0.0710	-14.7100	0.0000	***
pol_duration	-0.0025	0.0013	-1.9800	0.0481	*
pol_sit_duration	-0.3506	0.0068	-51.8900	0.0000	***
pol_pay_freqMonthly	0.0457	0.0247	1.8500	0.0644	.
pol_pay_freqQuarterly	0.1962	0.0562	3.4900	0.0005	***
pol_pay_freqYearly	-0.0439	0.0236	-1.8600	0.0630	.
pol_usageProfessional	-0.5502	0.2110	-2.6100	0.0091	**
pol_usageRetired	-0.7695	0.2105	-3.6600	0.0003	***
pol_usageWorkPrivate	-0.7002	0.2091	-3.3500	0.0008	***
drv_sex1M	-0.0627	0.0228	-2.7500	0.0059	.
drv_age_lic1	-0.0046	0.0019	-2.3800	0.0175	*
vh_speed	0.0036	0.0010	3.7800	0.0002	***
vh_typeTourism	-0.1088	0.0463	-2.3500	0.0187	*
vh_fuelGasoline	-0.1726	0.0276	-6.2500	0.0000	***
vh_value	0.0000	0.0000	3.1100	0.0019	**
drv_age1GLM	0.0078	0.0018	4.2400	0.0000	***
drv_age2GLM	-0.0053	0.0011	-4.9700	0.0000	***
vh_ageGLM	-0.0343	0.0029	-11.9700	0.0000	***
vh_makeGLMOther	0.0907	0.0298	3.0400	0.0023	**
drv_sex2GLMM	-0.0934	0.0361	-2.5900	0.0096	**
sales_durationGLM	-0.0115	0.0053	-2.1800	0.0291	*

Note: '***' Pr (<|0.001|) '**' Pr (<|0.01|) '*' Pr (<|0.05|) '.' Pr (<|0.05|)

4.2. Claim Amount Perspective

When compared to the frequency mode, fewer variables are considered to have potential correlation with claim size as Ill as fewer variables with significance. As shown in Table 3, the impact of some variables on the claim frequency and severity is consistent with intrinsically correlation in general. For example, higher driver bonus score which represents bad drive behaviors increases accident frequency and also leads to more serious claims. Wider policy coverage with more frequent claims contains more complicated claims. In addition, the possible claim amount for elderly drivers and passengers are relatively high, but for elderly vehicles are relatively low.

However, one variable requires to pay more attention is the age of secondary driver (passenger)'s license which has a large coefficient and high significance. Variable drv_age_lic2 is a conditional

DOI: 10.54254/2754-1169/113/2024LD0103

variable given the drv_drv2 is yes in this model. Even variable drv_drv2 is less significant, it is reasonable to say cars with experienced secondary driver indirectly correlated with claim amount combined both variables.

Table 3: Significant Model C Coefficients - Claim Amount

	Estimate	Std. Error	T value	Pr (>|t|)	Significance
(Intercept)	6.1458	0.2098	29.2900	0.0000	***
pol_bonus	0.4646	0.1504	3.0900	0.0020	**
pol_coverageMedian1	-0.2213	0.0667	-3.3200	0.0009	***
pol_coverageMedian2	-0.3299	0.0485	-6.8100	0.0000	***
pol_coverageMini	0.2135	0.1110	1.9200	0.0545	.
drv_drv2Yes	0.2052	0.0799	2.5700	0.0103	*
drv_age_lic2	0.0058	0.0014	4.2700	0.0000	***
vh_fuelGasoline	0.0864	0.0372	2.3200	0.0202	*
drv_age1GLM	0.0044	0.0011	3.8400	0.0001	***
drv_age2GLM	-0.0079	0.0017	-4.6300	0.0000	***
vh_ageGLM	-0.0156	0.0038	-4.1000	0.0000	***
sales_durationGLM	0.0243	0.0079	3.0600	0.0022	**

Note: '***' Pr (<|0.001|) '**' Pr (<|0.01|) '*' Pr (<|0.05|) '.' Pr (<|0.05|)

4.3. Eliminating Discrimination

Traditionally, the risk classification systems used by insurance companies may rely on a number of easily available but potentially discriminatory variables such as gender, age, geographic location, and so on. These variables may have only an indirect relationship with actual risk behavior. It is worth noting that driver sex's significance is relatively low in this model of regression, so that abandoning the use of sex in pricing should has little impact on risk management while preventing discrimination.

With the development of data science and related technologies, insurance companies now have access to more data directly related to risk. By using more data directly related to the vehicle and driver behavior, insurance company can more accurately price the risk of individual drivers, rather than relying on group characteristics that can be discriminatory.

For example, there is no longer a need for pricing based on gender or age, but rather on policy bonus, mileage data, or the usage of individual cars.

Unlike immutable characteristics such as gender or age, drivers can improve their driving behavior and get reasonable pricing, which can also enhance driver's awareness and reduce vehicle claims. As the advances in telematics, it is expected that more innovative data sources and analysis methods will be developed, which will further improve the accuracy and fairness of insurance pricing.

5. Conclusion

This paper delves into the complex realm of motor insurance, employing multiple methodologies to enhance pricing models while circumventing discrimination. The study engaged three distinct features selection methods (multivariate selection, random forest selection and stepwise backward selection), aiming to identify the variable's significance on claim frequency and claim amounts. Subsequently, the frequency of insurance claims is modeled using Poisson and negative binomial distributions for regression analysis, capturing the discrete and probabilistic nature of claim occurrences. For the severity of claims, represented by the amount of money paid out, gamma and Pareto distributions are employed. These distributions are adept at handling positive continuous data

DOI: 10.54254/2754-1169/113/2024LD0103

that can be skewed, which is often the case with claim amounts. After regression, the aggregated claim amount is calculated by multiplying severity and frequency from the dual perspectives and evaluated using test root mean square error (RMSE). It is concluded that the model trained under the Poisson + gamma distribution results in smaller RMSE. This rigorous evaluation of different statistical models against actual claim data provided insights into which approaches best minimize prediction errors, thereby enhancing the precision of actuarial forecasts.

Based on regression analysis, the study analyzes the relationship between selected features and the target variables respectively, reveling nuanced impact on both the claim frequency and claim amount. Notably, the research concludes three potential specific variables (referred to as policy bonus, mileage base data, and car usage) that demonstrated strong associations with claims. Insurance companies can refer to these variables and develop pricing models with fairness and risk manage ability. By integrating these variables into their actuarial processes and abandoning discriminatory variables like gender and age, companies can potentially craft insurance policies that are more aligned with individual risk, thereby minimizing the risk of discriminatory practices.

References

[1] *McKinsey & Company (2023). Global Insurance Report 2023: Reimagining life.insurance[^5^].https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-report-2023-reimagining-life-insurance#/*

[2] *Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. Insurance: Mathematics and Economics, 105, 79-95.*

[3] *Cather, D. A. (2023). Addressing insurance price discrimination in an era of diversity, equity, and inclusion. Risk Management and Insurance Review, 26(3), 407-429.*

[4] *Dong, C., Shao, C., Li, J., & Xiong, Z. (2018). An improved deep learning model for traffic crash prediction. Journal of Advanced Transportation, 2018(1), 3869106.*

[5] *Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 52.*

[6] *Taha, A., Cosgrave, B., & Mckeever, S. (2022). Using feature selection with machine learning for generation of insurance insights. Applied Sciences, 12(6), 3209.*

[7] *Dutang, C., Charpentier, A., & Dutang, M. C. (2020). Package 'casdatasets'. url: https://www. openml. org/search.*

[8] *Frees, Edward W., Gee Lee, and Lu Yang. "Multivariate frequency-severity regression models in insurance." Risks 4.1 (2016): 4.*