Journal of Advanced Artificial Intelligence
Volume 2 - No.4, January 2026

Dynamic Car Insurance Pricing using Interpretable Machine Learning Models: A Comparative Study of Regression, Support Vector Machines, and Neural Networks

Yassmine Ouladhaj Kaddour
FSJES Tetouan, University Abdelmalek Essadi
Tetouan, Morocco

Asmaa Faris
FSJES Ain Sebaa, University Hassan II
Casablanca, Morocco

Mohamed Dakkoun
FSJES Tetouan, University Abdelmalek Essadi
Tetouan, Morocco

ABSTRACT

In the non-life insurance sector, dynamic pricing has emerged as a crucial component of modern auto insurance, allowing insurers to adjust premiums more precisely and enhance risk differentiation. Traditional actuarial methods, such as generalized linear models (GLMs), provide strong interpretability but often fail to capture complex nonlinear relationships or high-dimensional structures. To address these limitations, this study explores supervised learning techniques, including regression models, support vector machines (SVMs), and neural networks (NNs), to model behavioral patterns, driving conditions, and claim outcomes. The results indicate that simple linear models consistently outperform more complex approaches. Ridge, Lasso, and Linear Regression achieve comparable performance, with R^2 values around 0.943 and RMSE ranging from 97.56 to 97.71. Interpretability analyses, including permutation importance and SHAP, reveal that prior accidents are the primary determinant of pricing decisions, exerting an effect nearly ten times greater than other factors. The overall objective is to develop an accurate and interpretable model capable of estimating premiums while offering improvements over traditional actuarial methods.

General Terms

Insurance Pricing, Machine Learning

Keywords

Dynamic Pricing, Car Insurance, Machine Learning, SVM, Neural Networks.

1. INTRODUCTION

The insurance sector is a fundamental pillar of any modern economy and is currently undergoing disruptive changes, both in terms of size and technological evolution. In Morocco, car insurance is mandatory for all owners of motorized land vehicles. Its primary purpose is to cover the insured's civil liability in the event of damage caused to third parties during an accident. This includes, in particular, collision damage, comprehensive coverage, fire, and theft.
In recent years, the auto insurance sector has undergone a profound transformation driven by digitalization and the massive increase in data, both globally and nationally. These changes are largely due to the adoption of new technologies, which are pushing insurers to rethink their traditional methods. In this context, pricing, which is at the heart of the insurance business, can no longer rely solely on classic approaches. Classical actuarial models, particularly generalized linear models (GLMs), still hold a central place in practice due to their statistical rigor and interpretability, but they present several limitations when it comes to processing high-dimensional data or integrating complex interactions between variables. The general idea of dynamic pricing is based on the ability to adjust premiums more frequently and more closely to the insured's actual profile, while taking into account changes in risk and variability in driving behavior. Machine learning is emerging as a key driver of this evolution. Thanks to its ability to analyze complex data, it paves the way for more flexible and potentially more efficient pricing models.
From this perspective, this work examines the extent to which different machine learning methods can contribute to improving automotive pricing within a dynamic framework. This article focuses particularly on the ability of modern models—especially Support Vector Machines (SVMs), neural networks, and multiple linear regressions—to leverage multivariate data in order to estimate future risk more accurately. Each of these approaches offers significant advantages in the context of dynamic pricing: they allow for the processing of large datasets, the identification of sometimes complex relationships between variables, and improved predictive accuracy compared to traditional models.
Overall, the integration of these learning methods continues to be a crucial step in strengthening the performance of pricing systems

and meeting the requirements of a more dynamic and customer-oriented insurance environment.

2. LITERATURE REVIEW

A key challenge faced by insurance companies is determining the optimal insurance premium for each risk presented by clients. The risk varies considerably from one client to another due to individual characteristics, behaviors, and different environments. In this context, car insurance pricing has undergone significant changes in recent years. Actuaries generally use generalized linear models (GLM) for rate development, with the Poisson distribution used to model claim frequency and the Gamma distribution used to model claim severity [1]. These classical methods allow modeling both nonlinear behaviors and non-Gaussian residual distributions, while providing acceptable predictive performance [2]. However, even though these models are effective, they have several limitations. They become less suitable for large-scale problems, especially when dealing with high-dimensional or complex datasets, and their applications may become unreliable or inefficient [3]. These challenges have encouraged actuaries to explore more flexible and efficient algorithms, such as machine learning methods [4]. Building on these limitations, recently developed machine learning methods are transforming auto insurance pricing. Techniques such as regularized regression, Support Vector Machines (SVM), and Neural Networks (NN) currently enable insurers to forecast future claims with significantly improved precision, leading to better risk assessment and more efficient pricing models [5]. While traditional actuarial models remain important for benchmarking, it is crucial to understand their assumptions and constraints. Auto insurance pricing has historically been based on traditional actuarial models. Generalized linear models (GLMs) are a generalization of the linear regression concept. They allow for the estimation of the pure premium based on predefined rating variables and enable the study and quantification of the relationship between a response variable Y and explanatory variables [6]. In the context of auto insurance, the Poisson-Gamma and Inverse-Poisson models offer great flexibility in modeling loss distributions that exhibit overdispersion, providing insurers with tools to improve pricing accuracy. Despite their widespread adoption, these traditional models have several limitations. GLMs also assume that each data observation exists in isolation, without inherent relationships between observations [7]. These models assume a linear structure between the predictors and the link function, which limits their ability to capture complex interactions [8]. To address these limitations, machine learning has emerged as a promising alternative. “Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed” [4], and is seen as a branch of artificial intelligence, evolving over the years to become a key technology in many sectors. The underlying idea of machine learning is to make predictions based on experience gained from provided data. It is a non-parametric method that does not require any assumptions about the distribution of the data to be explained. Two key considerations influence the choice of machine learning in insurance models: first, the ability to identify inputs, which allows for the construction of multiple scenarios to evaluate the AI system; second, the ability to detect critical or problematic outputs . These systems have the potential to improve safety through accurate automation, for example by reducing the number of accidents caused by humans or by allowing systems to operate in hazardous environments without direct human control [5].

Among machine learning methods, penalized regression techniques provide a natural starting point. Multiple linear regression is a multivariate analysis method that allows explaining a quantitative variable using a linear combination of several explanatory variables, whether they are quantitative or qualitative. Regularization techniques (Ridge or Lasso) were developed to address overfitting and multicollinearity issues [9]. Modern extensions include Ridge regression, which involves adding an L2 penalty term to the sum of squared residuals to reduce the variance of the coefficients and limit overfitting [10]. In contrast, Lasso regression performs both regularization and variable selection by applying an L1 penalty. They play an essential role in reducing variance, preventing overfitting, and improving the stability of claims frequency models [11]. The Elastic Net combines L1 (Lasso) and L2 (Ridge) approaches, particularly suited to insurance data exhibiting multicollinearity. These methods are mainly used as a benchmark to assess the performance of more complex models.

Moving beyond regression-based methods, Support Vector Machines offer another approach for complex insurance datasets [12, 22]. SVMs are among the most effective supervised learning methods in a wide range of applications [22]. However, since their initial formulation, several significant challenges have emerged, most notably the computational difficulty of scaling SVMs to very large datasets, which limits their applicability in big data contexts.

For even more flexible modeling, neural networks have been increasingly applied in insurance pricing. Recurrent Neural Networks (RNN) are deep neural networks trained on sequential data or time series data to create a machine learning model capable of drawing conclusions based on sequential inputs. The typical architecture for pricing uses multi-layer perceptrons (MLP) with activation functions for the hidden layers and a linear activation for the output layer in regression [13]. [14] proposed a Combined Actuarial Neural Networks (CANN) approach that combines a classic actuarial model, such as a GLM, with a neural network. In addition to traditional approaches based on Poisson or negative binomial distributions for cross-sectional data, they implement a CANN training procedure with a multivariate negative binomial specification to model the dependence between policies of the same insured while capturing individual risk.

In parallel, the insurance market has increasingly adopted dynamic pricing strategies. Nowadays, the insurance market has experienced significant growth, encouraging the adoption of a dynamic pricing strategy involving frequent price adjustments to optimize sales and profitability. Dynamic pricing is defined as price changes caused by variations or differences in four key factors of market demand: (1) People (i.e., individual consumers or consumer segments), (2) Product configurations, (3) Time periods (i.e., time), and (4) Locations (i.e., places) [15]. In summary, the rise of machine learning in dynamic pricing reinforces the need for strong model explainability, especially under GDPR regulations that demand transparency in the use of personal data. Organizations must balance predictive accuracy with interpretability to ensure that model-based decisions remain reliable, transparent, and compliant.

3. METHODOLOGY

The prediction of claims is a major challenge in the insurance sector, as it enables companies to offer contracts tailored to each policyholder’s risk profile. However, risk levels vary significantly among individuals. Today, the integration of machine learning methods has profoundly transformed risk management practices in automobile insurance, enhancing the accuracy and relevance of pricing decisions. In this context, this article aims to present the approach adopted for implementing a dynamic pricing system based on advanced machine learning techniques. .

3.1 Materials and Proposed Model

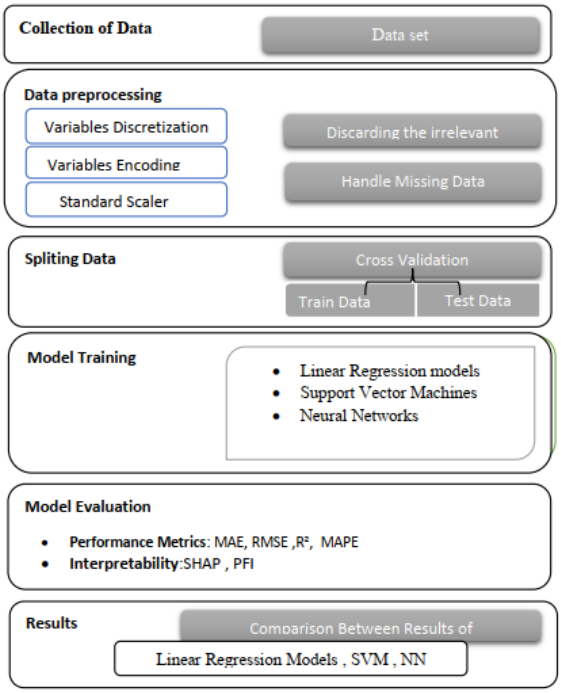
3.1.1 Data Description .

This dataset contains seven variables and 1,000 observations collected from Kaggle, consisting of synthetic data that simulate car insurance premiums calculated using a linear formula. It includes key features such as driver age, driving experience, accident history, annual mileage, and the car's manufacturing year, all of which are used to predict the insurance premium.
The dataset is well suited for exploring linear regression models, analyzing feature importance, and developing predictive models in the insurance industry. It was inspired by real-world factors influencing insurance premiums, ensuring realistic patterns and meaningful insights.

Table 1. Description of dataset

Name	**Description**
Driver Age	Driver's age at policy start.
Driver Experience	Years of driving experience.
Previous Accidents	Number of past claims.
Annual Mileage	Estimated yearly distance.
Car Manuf. Year	Year the car was made.
Car Age	Vehicle's age in years.
Premium ($)	Paid insurance premium.

3.1.2 Proposed Model



figureOverall structure of the proposed dynamic pricing model.

3.2 Data Preprocessing

3.2.1 Cleaning data and Preparation.

A preliminary exploratory analysis was conducted, and the dataset was found to be of high quality, containing neither duplicates nor missing values. All verification steps confirmed the absence of duplicated entries. Outlier detection was then performed using the interquartile range (IQR) method and visualized with boxplots. No anomalies requiring specific treatment were identified. For the construction of the target variable, the insurance premium is modeled in dollars, directly reflecting the pricing objective.

3.2.2 Descriptive Analysis.

Descriptive analysis constitutes the first phase of our methodological approach. We computed descriptive statistics for each variable in the dataset including the mean, standard deviation, median, minimum, and maximum to gain a better understanding of their dispersion and distribution. This analysis also allowed us to examine the risk distribution of the insurance premium.
A correlation matrix including all variables was used to visualize the relationships between features, complemented by histograms illustrating the distribution of each variable. Feature engineering is an essential step in data preparation. Although the initial dataset was of high quality, we created relevant derived variables to enhance the predictive performance of our model .

3.2.3 Correlation Matrix.

Correlation analysis allows for the identification of linear relationships between variables and the detection of potential multicollinearities. figure 1 presents the Pearson correlation matrix between all the variables.
As part of the data exploration and preprocessing phase, a correlation matrix was calculated to analyze the linear relationships between several explanatory variables and the target variable (the insurance premium). This analysis provided an initial overview of the dependency structure within the dataset, allowing for the identification of potential redundancies among the variables.
By examining the pairwise correlations, it is possible to detect multicollinearity issues that can affect the stability and performance of certain models, particularly regression-based models. The results reveal that the most influential predictive factor for the insurance premium is the history of previous accidents, with a strong correlation value of 0.85, indicating that drivers with a higher number of past accidents generally pay higher premiums. This confirms the central role of claims history in risk assessment.
On the other hand, driver experience is negatively correlated with the insurance premium (correlation = -0.28), suggesting that more experienced drivers benefit from a more favorable risk profile. Factors related to vehicle usage and condition, such as annual mileage and car age, have only a low to moderate positive impact, while driver age and car manufacturing year are practically uncorrelated with the premium. Overall, the low correlations observed among the explanatory variables suggest minimal multicollinearity, which is favorable for the application of regression and machine learning models.

3.3 Variable Distributions

The analysis of the distributions shows that the explanatory variables in the portfolio are generally well balanced and do not reveal any notable irregularities. Driver age and driving experience display an almost uniform distribution, representing young, adult, and experienced profiles. Accident history, although limited, shows a higher proportion of drivers with no past claims, consistent with the typical structure of automobile insurance portfolios. Annual mileage and the vehicle's manufacturing year capture both low- and high-mileage drivers as well as recent and older vehicles. Overall, the distribution of insurance premiums exhibits an asymmetric

Table 2. Descriptive statistics

	Age	Exp.	Accid.	Mileage	Manuf.	Car Age	Premium
count	1000	1000	1000	1000	1000	1000	1000
mean	43.82	20.02	2.54	27.73	2012.08	11.73	1427.99
std	14.99	11.68	1.72	13.12	6.81	6.71	404.45
min	18	0	0	5.01	2000	0	273.61
50%	44	20	3	28.46	2012	12	1446.15
max	69	39	5	49.99	2023	23	2562.78

shape centered around average values, reflecting realistic market behavior. These characteristics ensure a reliable and representative dataset suitable for pricing modeling.

3.3.1 Feature Engineering.
Driver risk is categorized into three levels: Low risk with values between 0 or 1 for those having accidents, Medium risk for 2 to 3 accidents, and High risk for higher values. Normalization was carried out due to the necessity of handling Support Vector Machines and Recurrent Neural Networks, which are sensitive to the scales of variables. The dataset is split into two parts, comprising 80% of the observations for training, and a test set comprising 20% of the observations.

Table 3. Description of Risk Categories for the Variable 'Previous Accidents'

Risk Category	Description	Data Size
Low Risk	≤ 1 accident	340
Medium Risk	$2 \leq accidents \leq 3$	316
High Risk	> 3 accidents	344

3.3.2 Normalization and split of the data.

The data were normalized differently depending on the models. For linear regression models and SVMs, the StandardScaler normalization was applied, uniformly scaling the variables with a mean of 0 and a variance of 1.

$$Z = \frac{X - \mu}{\sigma}$$

On the other hand, for the neural network, a MinMaxScaler normalization was used to standardize all features in the range $[0, 1]$, which facilitates learning and optimizes the convergence of neural algorithms.

$$Z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

3.4 Machine Learning Models

3.4.1 Linear Regression Models and Regularization . Multiple linear regression serves as a baseline model to establish an interpretable reference for pricing,taking into account all relevant driver and vehicle characteristics . The basic linear regression formula is given by:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Where y_i is the dependent variable (Insurance Premium), x_{ij} are the values of the explanatory variables (driver/vehicle features), β_j are the model coefficients (weights),and ϵ_i are the residuals (prediction errors).

In **Ridge regression**, a penalty equivalent to the square of the magnitude of the coefficients is applied to the cost function (minimizing the sum of squared errors), ensuring that they are small but not zero. This process is known as L2 regularization .

$$Cost(\boldsymbol{\beta})Ridge = \sum i = 1^n (y_i - \hat{y}i)^2 + \lambda \sum j = 1^p \beta_j^2$$

Lasso regression (Least Absolute Shrinkage and Selection Operator), on the other hand, applies an absolute value penalty term, which can reduce some coefficients to zero, thereby removing the corresponding feature from the model. This method is known as L1 regularization.

$$Cost(\boldsymbol{\beta})Lasso = \sum i = 1^n (y_i - \hat{y}i)^2 + \lambda \sum j = 1^p |\beta_j|$$

Finally, the ElasticNet model combines the two approaches (L1 and L2), offering a better balance, which is particularly useful for managing correlated data. However, this family of models stands out for its simplicity and its ability to provide a solid foundation for interpretability

3.4.2 Support Vector Machines.
Support vector machines (SVM), our second family of models, are capable of capturing nonlinear relationships through the use of different kernels. We tested three kernels: linear, RBF (radial basis function) capable of capturing nonlinear relationships, and polynomial of degree 3. Each model was trained after normalizing the explanatory variables using a StandardScaler, which is crucial for the proper functioning of SVMs.

(1) **Linear Kernel:**

$$K(x_i, x_j) = x_i^T \cdot x_j \tag{1}$$

(2) **RBF Kernel (Radial Basis Function):**

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{2}$$

where γ defines the influence of a single training example.

(3) **Polynomial Kernel:**

$$K(x_i, x_j) = (x_i^T \cdot x_j + c)^d \tag{3}$$

where $d = 3$ represents the polynomial degree.

The objective function of the SVR is defined as:

$$\min_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{4}$$

where C is the regularization parameter and ξ_i represents errors beyond the margin ϵ.

3.4.3 Neural Networks.
The third model considered is the Neural Network (NN), a machine learning approach that enables the identification of complex nonlinear patterns and supports data-driven decision-making. Two network architectures were evaluated. The first corresponds to a shallow neural network composed of two hidden layers with 50 and 30 neurons, respectively, which is suitable for modeling moderately complex relationships. The second corresponds to a deep neural network featuring three hidden layers with 100, 50, and 25 neurons, allowing the model to capture hierarchical and abstract representations of the underlying data.
The Mean Squared Error (MSE) was employed as the loss function, as it is appropriate for regression tasks and aims to minimize the average squared difference between predicted and observed values. Model training was performed using the Adam optimizer, specified through the parameter `solver='adam'` in the `MLPRegressor`. This optimizer adaptively adjusts the learning rate during training, ensuring stable and efficient convergence. In addition, input features were scaled using the `MinMaxScaler`, a crucial preprocessing step for neural networks to improve training stability and overall performance.
The inputs are normalized using *MinMaxScaler*, ensuring consistent training. We employed the Adam optimizer (*Adaptive Moment Estimation*), which automatically adjusts the learning rate for stable convergence.
The ReLU (Rectified Linear Unit) activation function introduces non-linearity into the network, as shown in Eq. (5):

$$ReLU(x) = \max(0, x) \tag{5}$$

Forward propagation through each hidden layer computes activations by applying weights, biases, and the ReLU function:

$$h^{(l)} = ReLU(W^{(l)}h^{(l-1)} + b^{(l)}) \tag{6}$$

where $h^{(l)}$ is the output of layer l, $W^{(l)}$ is the weight matrix, and $b^{(l)}$ is the bias vector.
To minimize prediction errors, the Mean Squared Error (MSE) loss function is utilized:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

Finally, the Adam optimizer updates the network weights adaptively:

$$W_t = W_{t-1} - \alpha\frac{m_t}{\sqrt{v_t} + \epsilon} \tag{8}$$

where α is the learning rate, and m_t, v_t represent the adaptive first and second moments, respectively. The final output layer produces the price prediction through a linear transformation:

$$\hat{y} = W^{(out)}h^{(L)} + b^{(out)} \tag{9}$$

where $\hat{y}$ represents the predicted price, $W^{(out)}$ is the weight matrix of the output layer, $h^{(L)}$ is the activation from the last hidden layer, and $b^{(out)}$ is the bias term.

3.5 Evaluation and Validation Methods

3.5.1 Evaluation metrics.

The metrics used to evaluate pricing models, which involve a regression problem, include several complementary values. Among the main evaluation metrics, we calculate the MAE (Mean Absolute Error), which can be used to assess the model's accuracy and measure the average of the absolute differences between the predictions and the actual observed values

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

The RMSE (Root Mean Squared Error), which is a standard measure for evaluating the quality of predictions from a model, represents the average difference between predicted and actual values, while penalizing larger errors more heavily and is sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

The MAPE (Mean Absolute Percentage Error) is used to evaluate the accuracy of a forecasting or predictive model by expressing the average prediction error as a percentage of the actual values.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{12}$$

The coefficient of determination R^2 expresses the proportion of the variance of the dependent variable Y that is explained by the independent variable X of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

3.5.2 Cross Validation. To ensure the robustness and generalization of our models, a k-fold cross-validation strategy with $k = 5$ was adopted. This technique splits the dataset into 5 equal-sized subsets, where each subset successively serves as the test set, while the other 4 serve for training.
The final performance is calculated as the average of the performances across the 5 folds:

$$Score_{CV} = \frac{1}{k}\sum_{i=1}^{k}Score_i \tag{14}$$

Where $Score_i$ represents the evaluation metric (MAE, RMSE, R²) on the i-th fold. This procedure helps reduce the variance associated with a single train-test split and enables the detection of potential overfitting, thereby providing a more reliable estimate of the predictive performance of models.

3.5.3 Importance of variables.

The analysis of feature importance helps identify the factors that most significantly influence dynamic pricing. Two complementary approaches are used:

3.5.4 Permutation Feature Importance.

Permutation Feature Importance evaluates the contribution of each variable by randomly shuffling its values and measuring the resulting degradation in the model's performance. The importance of a variable j is calculated as:

$$Importance_j = Score_{original} - Score_{permuted_j} \tag{15}$$

Where $Score_{original}$ is the performance of the model with the original data, and $Score_{permuted_j}$ is the performance after permuting the variable j. The table **??** presents the results of the Permutation Feature Importance for the best model (Ridge).

3.5.5 SHAP Values (Shapley Additive Explanations).

SHAP values provide a unified framework for explaining the contribution of each variable to individual predictions, grounded in cooperative game theory. The SHAP value for a variable j and an observation i is defined as:

$$\phi_j(i) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x_i) - f_S(x_i)] \quad (16)$$

where F denotes the set of all variables, S is a subset of variables not including j, and $f_S(x_i)$ represents the model's prediction using only the variables in S.

SHAP works by decomposing a model's output into the sum of the contributions of each feature. It computes a value that represents how much each feature influences the model's prediction. These values can be used to understand the relative importance of each feature and to provide human-interpretable explanations of the model's decisions. By assigning a contribution to every input feature, SHAP shows how—and to what extent—each feature affects the final prediction.

All figures presented in this paper were regenerated as original high-resolution images to ensure clarity, readability of text labels, and proper rendering without distortion when zoomed in the final published version.

4. RESULTS AND DISCUSSION

To provide a comprehensive evaluation of the proposed dynamic pricing framework, multiple machine learning paradigms were assessed, including linear regression models, support vector machines with various kernels, and neural networks with different architectures. The models were evaluated under several experimental scenarios using complementary performance metrics (MAE, RMSE, MAPE, and R²). Furthermore, a 5-fold cross-validation strategy was employed to assess robustness and generalization across different data splits. This extensive evaluation framework enables a reliable comparison of model performance and confirms the consistency and stability of the obtained results. The **Table4** presents the main supervised learning performance models applied to dynamic pricing of car insurance. All models were trained on 800 observations representing 80% of the data and evaluated on 200 observations, 20% as test data. However, StandardScaler normalization was used for linear regression models and SVMs, while MinMaxScaler was applied to Neural Networks (NN) .

4.1 Discussion of Model Results

All four linear regression models were trained using *StandardScaler* to normalize the explanatory variables, ensuring that all features contribute proportionally to the learning process, regardless of their original units.
The standard multiple linear regression model shows strong performance, with a MAE of 76.94$, an RMSE of 98.08$, and a coefficient of determination R^2 of 0.943. These results indicate that the model explains 94.3% of the observed variance in the insurance premium and maintains an average prediction error of less than 77$. The absence of regularization gives the model full flexibility in estimating the coefficients, which is appropriate for our problem, where multicollinearity although present remains at a moderate level.
Ridge regression, which introduces an L2 penalty to limit the magnitude of the coefficients, achieves the best overall performance, with a MAE of 76.90$, an RMSE of 98.10$, and an R^2 of 0.943. The minimal difference with linear regression (MAE of 0.04$) indicates that multicollinearity between the variables does not significantly affect the model's performance. However, the L2 penalty provides slightly increased stability to the coefficient estimates.
Lasso regression, based on L1 regularization, yields very similar performance, with an MAE of 77.05$, an RMSE of 98.25$, and an R^2 of 0.942. The advantage of lasso lies in its ability to perform automatic variable selection by zeroing out certain coefficients; however, in our context, this selection does not lead to significant gains in predictive performance.
The ElasticNet model, which combines both L1 and L2 regularization, shows noticeably lower performance, with a MAE of 122.24$, an RMSE of 149.50$, and an R^2 of 0.867. This performance degradation, corresponding to a 59% increase in MAE compared to the ridge model, suggests that the default hyperparameters ($\alpha = 1.0$, $l1\_ratio = 0.5$) are less optimal for this problem. Fine-tuning these parameters could considerably improve its performance.

Table 4. Comparison of Model Performances

Model	MAE	RMSE	R^2	MAPE (%)
Regression Models (StandardScaler)				
Linear Regression	76.94	98.08	**0.943**	6.05
Ridge	76.90	98.10	**0.943**	6.05
Lasso	77.05	98.25	0.942	6.07
Neural Networks (MinMaxScaler)				
NN (3 layers)	78.61	102.15	0.938	6.14
NN (2 layers)	93.95	120.56	0.913	7.18
Support Vector Machines (StandardScaler)				
Linear SVR	92.98	115.37	0.921	7.39
ElasticNet	122.24	149.50	0.867	10.19
Poly SVR	284.19	341.23	0.306	24.38
RBF SVR	302.30	365.03	0.206	26.00

4.1.1 Neural Networks Performance.

Both neural network architectures were trained using *MinMaxScaler* to normalize the data to the $[0, 1]$ range. The shallow neural network (NN, 2 layers), with an architecture comprising two hidden layers of 50 and 30 neurons respectively, shows satisfactory performance with a MAE of 93.95$, an RMSE of 120.56$, and an R^2 of 0.913. This simple architecture allows the network to learn moderate nonlinear transformations of the input features. However, the shallow depth limits its ability to capture complex hierarchical representations. The R^2 value of 0.913 indicates good explanatory power, capturing 91.3% of the variance in the insurance premiums.
The deep neural network (NN, 3 layers – Deep), with a three hidden layer architecture of 100, 50, and 25 neurons, demonstrates remarkable performance, achieving a MAE of 78.61$, an RMSE of 102.15$, and an R^2 of 0.938. The additional layers enable the network to form more abstract and hierarchical representations of the features, capturing subtle patterns that a shallow network cannot model.
Despite their flexibility and ability to model nonlinear relationships, neural networks do not outperform simple linear models. This observation suggests that the relationship between explana-

tory variables and insurance premiums is largely linear, making the added complexity of neural networks unnecessary. Nevertheless, the deep network remains a viable alternative and could become advantageous in future scenarios involving more complex data integration.

4.1.2 Support Vector Machines Analysis.

The training was carried out on the three SVM models using *StandardScaler* and differed by their kernel function, which determines how the data is transformed in the feature space. The SVR model with a linear kernel achieves satisfactory performance with an MAE of 92.98$, an RMSE of 115.37$, and an R^2 of 0.921. Although these results are about 21% lower than the best linear regression models in terms of MAE, they remain reasonable and confirm that the SVM approach can be viable for insurance pricing.
The SVR model with an RBF (Radial Basis Function) kernel shows poor performance, with an MAE of 302.30$, an RMSE of 365.03$, and a limited R^2 of 0.206. These results indicate that the model only explains 20.6% of the variance in premiums and that it makes an average error of 300$. This suggests that the nonlinear transformation induced by the Gaussian kernel does not match the underlying structure of the data.
The SVR model with a polynomial kernel (degree 3) also shows clearly insufficient performance, with a MAE of 284.19$, an RMSE of 341.23$, and an R^2 of 0.306. Although slightly better than RBF SVR, this model remains largely unsuitable for the problem. The degree 3 polynomial kernel creates feature combinations up to the third order, theoretically capturing complex interactions between variables. Hyperparameter optimization could improve performance, but the significant gap with the linear models suggests that the additional effort would not be worthwhile.

Figure 4 illustrates the distribution of key variables in the automobile insurance dataset—driver age, driving experience, previous accidents, vehicle characteristics, and premium values—providing an overview to identify variability, skewness, and potential outliers before model training.

Table 5. Cross-Validation Results (5-Fold)

Model	**RMSE**	**Std Dev**
Linear Regression	97.71	± 2.19
Ridge	97.71	± 2.28
Lasso	97.56	± 2.31
ElasticNet	145.72	± 5.04
Linear SVR	112.36	± 3.72
RBF SVR	360.64	± 10.98
Poly SVR	336.74	± 10.07
NN (2 layers)	113.23	± 6.69
NN (3 layers)	105.63	± 7.25

The results of the 5-fold cross-validation confirm the superiority of linear models for this regression problem. Simple linear regression, Ridge, and Lasso show almost identical performance with exceptionally low RMSEs (97.56 –97.71) and minimal standard deviations (±2.19 –2.31), indicating excellent stability across different data splits. Neural networks and Linear SVR produce results, but their higher standard deviations reveal increased sensitivity to the composition of the training data. In contrast, RBF SVR and Poly SVR kernel models fail dramatically with RMSEs exceeding 330. These results demonstrate that the underlying relationship between the variables is essentially linear, Shows that complex non-linear methods are unnecessary, confirming that regularized linear models are the best choice for performance, stability, and interpretability.

The Figure 5 shows the scatter plots of predicted versus actual values for the nine models on the test set. Linear regression, Ridge, and Lasso show an almost perfect alignment of the points along the diagonal, confirming their excellent R² scores. L1 regularization for Lasso and L2 for Ridge do not bring any significant improvement, which suggests that the model does not exhibit strong multicollinearity or overparameterization. However, ElasticNet, Linear SVR, and the 2- and 3-layer neural networks show a clear linear relationship, but with a slightly greater dispersion of the points. The three-layer neural network demonstrates superiority over the two-layer network, suggesting that a deeper architecture better captures potential nonlinearities, even though these remain limited in this dataset.The RBF SVR and Poly SVR models reveal predictions that form a horizontal plateau around 1400-1500, regardless of the actual values. This underfitting is probably due to a complexity that is unsuitable for the dominant linear structure.

Different regression models, which allows for the evaluation of both their bias and their stability. The linear regression, Ridge, and Lasso models show errors that are generally centered around zero, with a slightly negative mean, indicating a moderate tendency to underestimate. ElasticNet stands out with a mean error very close to zero, reflecting an almost zero bias. However, its distribution is a bit wider, revealing greater variability in the predictions. The linear SVR models, on the other hand, show a positive mean error, indicating a slight overestimation, with a dispersion comparable to that of the linear models. Non-linear SVR models, particularly the RBF and polynomial ones, show the largest deviations. Their distributions are widely dispersed, with very high extreme errors, indicating more systematic instability, which limits their generalization capacity in this context.However, neural networks show relatively robust intermediate behaviors, with a slight bias depending on the number of layers.
The analysis of variable importance by permutation, summarized in Table 6 and Figure 7, confirms an explicit ranking of risk variables in the insurance pricing process.
The most dominant factor is the history of previous accidents (*Previous Accidents*), with an importance of $\mathbf{1.494}$. This means that if this variable were randomly shuffled (permuted), the model's error would increase on average by 149.4%. This result is consistent with the high correlation coefficient of $\mathbf{0.85}$ observed during the exploratory analysis, which confirms the central role of past risk in pricing.
The secondary factors are Driver Experience and Annual Mileage, whose importance values are significant (0.144 and 0.129 respectively). The importance of these variables is consistent with actuarial practices, as they are excellent indicators of risk exposure and driving profile.
Other variables such as Car Age, Driver Age, and the categorized risk level (*Risk_Level*) have low or even negligible importance (negative for *Car Manufacturing Year*). The low importance of the categorized variable *Risk_Level* (grouping *Previous Accidents*) suggests that the raw information (exact number of accidents) is more relevant than its categorized version.

Table 6. Permutation feature importance for the Ridge model

Variable	**Importance**	**Std. Dev.**
Previous Accidents	1.494	± 0.037
Driver Experience	0.144	± 0.004
Annual Mileage (×1000 km)	0.129	± 0.005
Car Age	0.118	± 0.007
Driver Age	0.053	± 0.003
Risk Level	0.0002	± 0.0001
Car Manufacturing Year	-0.00015	± 0.00012

This visualization summarizes the distribution of SHAP values for each variable across all observations. Each point represents an individual prediction, with a color representing the variable's value (red for high values, blue for low values). For Previous Accidents, there is significant dispersion across the entire range of the axis, confirming its predominant role as well as the high variability of its impact. Red points consistently cluster on the positive side of the axis, indicating a significant increase in the predicted premium, while blue points are mostly on the negative side, reducing the premium. For Driver Experience, Annual Mileage, and Car Age , the distributions are more concentrated around zero, with SHAP values varying roughly between -200 and +200, but show marked trends for experience, with red values (high experience) tending towards negative impacts (premium reduction), while blue values (low experience) have positive impacts. The variables Risk Level and Car Manufacturing Year show negligible effects, strongly concentrated around zero, confirming their low contribution to the model.

This waterfall chart (Figure 9) illustrates how the model constructs its prediction for a specific observation. Starting from a base value of $E[f(x)] = 1439.762$ (the average prediction), each bar represents the additive contribution of a feature to reach the final prediction $f(x) = 1498.196$. For the analyzed observation, the normalized value of previous accidents is 0.854, contributing $+302.9$ to the prediction and representing the most significant impact. Vehicle age, with a normalized value of -1.301, reduces the premium by -146.96, while driver experience 1.198 decreases it by -127.23. Driver age 0.813 contributes positively with $+63.93$. The remaining variables have minor contributions. This additive decomposition provides a clear explanation of why this client receives a premium of 1498.196, ensuring transparency in the model's decision-making process.
This chart Figure 10 summarizes the overall importance of each variable by calculating the average of the absolute SHAP values across all predictions. Previous Accidents dominates overwhelmingly with an average importance of around 300, far surpassing all other variables. This predominance confirms that the history of claims is the most determining factor in pricing. The secondary variables are Driver Experience, Annual Mileage , and Car Age , with similar importance around 80-90, forming an intermediate group of moderately influential predictors. "Driver Age" has a lower importance (around 50), suggesting a limited but still notable role. Finally, Risk Level and Car Manufacturing Year show near-zero importance, indicating that they provide virtually no additional information to the model once the other variables are accounted for This hierarchy perfectly corresponds to the results obtained by permutation analysis, validating the robustness of the interpretation.

5. CONCLUSION

This study compared nine machine learning models for car insurance pricing and found that simple linear models significantly outperformed complex approaches. Ridge, Lasso, and Linear Regression achieved nearly identical exceptional performance with R-squared scores of 0.943 and RMSE values between 97.56 and 97.71, with differences so minimal they are practically equivalent. This demonstrates that the relationship between risk factors and insurance premiums is fundamentally linear. Analysis revealed that previous accidents dominate pricing decisions with importance ten times greater than other factors, while driver experience, annual mileage, and vehicle age play secondary roles. The linear models offer key advantages including exceptional accuracy for fair pricing, complete transparency through SHAP analysis for regulatory compliance, minimal computational requirements for real-time systems, and remarkable stability with standard deviations below 2.5. The near-identical performance of all three linear variants indicates that regularization provides minimal benefit for this dataset, though Ridge or Lasso remain preferable for production deployment due to enhanced stability. Overall, the results suggest that making the model more complex brings limited gains, confirming that well-fitted linear models represent an optimal balance between performance, readability, and operational efficiency for car insurance pricing. However, it is important to note that these conclusions are partly influenced by the synthetic and predominantly linear structure of the dataset. Future research should therefore focus on validating these findings using real-world insurance data, incorporating telematics variables and temporal dynamics to further assess the robustness of the proposed pricing framework .

6. REFERENCES

[1] Modugno, L., Perin, G., & Bernardi, A. (2023). Modelling motor insurance claim frequency and severity using gradient boosting. *Risks*, 11(3).

[2] Henckaerts, R., Frees, E. W., & Antonio, K. (2023). Neural networks for insurance pricing with frequency and severity data: A benchmark study from data preprocessing to technical tariff. *North American Actuarial Journal*, 29(3).

[3] Mahendran, A., Thompson, H., & McGree, J. M. (2023). A model-robust subsampling approach for generalized linear models in big data settings. *Journal of Computational and Graphical Statistics*, 32(2).

[4] Rayadurgam, S., & Byun, T. (2020). Manifold for machine learning assurance. In *Proceedings of ICSE-NIER 2020*.

[5] Burton, S., & Herd, B. (2023). Addressing uncertainty in the safety assurance of machine learning. *IEEE Software*.

[6] Haris, M., & Arum, P. (2022). Poisson-gamma and inverse-Poisson models for insurance claim modeling. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 16(2).

[7] Novkaniza, A. (2025). A posteriori premium rate calculation using Poisson-gamma hierarchical generalized linear model.

Journal of Theoretical and Applied Mathematics (JTAM), 12(4).

[8] Wüthrich, M. V., & Merz, M. (2023). *Statistical foundations of actuarial learning and its applications*. Springer Actuarial.

[9] McGuire, G., Taylor, G., & Miller, H. (2021). Self-assembling insurance claim models using regularized regression and machine learning. *Variance*, 14(1).

[10] Kennard, R., & Hoerl, A. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1).

[11] Sun, B., Lin, X., Furman, E., & Wüthrich, M. (2024). Nonparametric intercept regularization for insurance claim frequency regression models. *ASTIN Bulletin*, 54(1).

[12] Vapnik, V. (2000). *The nature of statistical learning theory* (2nd ed.). Springer.

[13] Wüthrich, M. V. (2019). Neural network applications for actuarial risk modelling. *ASTIN Bulletin*, 49(3).

[14] Duval, F., Boucher, J.-P., & Pigeon, M. (2024). Telematics combined actuarial neural networks for cross-sectional and longitudinal claim count data. *ASTIN Bulletin*.

[15] Kopalle, P. K., Pauwels, K., Akella, L. Y., & Gangwar, M. (2023). Dynamic pricing: Definition, implications for managers, and future research directions. *Journal of the Academy of Marketing Science*, 51(1).

[16] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

[17] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate data analysis* (8th ed.). Pearson.

[18] Haris, M., & Arum, P. (2022). Negative binomial regression and generalized Poisson regression models on the number of traffic accidents in Central Java. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 16(2).

[19] Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.

[20] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2022). Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7).

[21] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2022). Fooling LIME and SHAP: Adversarial attacks and robustness improvements. *Machine Learning*, 111(5).

[22] Cervantes, F., & Gomez, R. (2020). Support vector machines applications in insurance risk prediction. *Journal of Insurance Analytics*, 8(1).

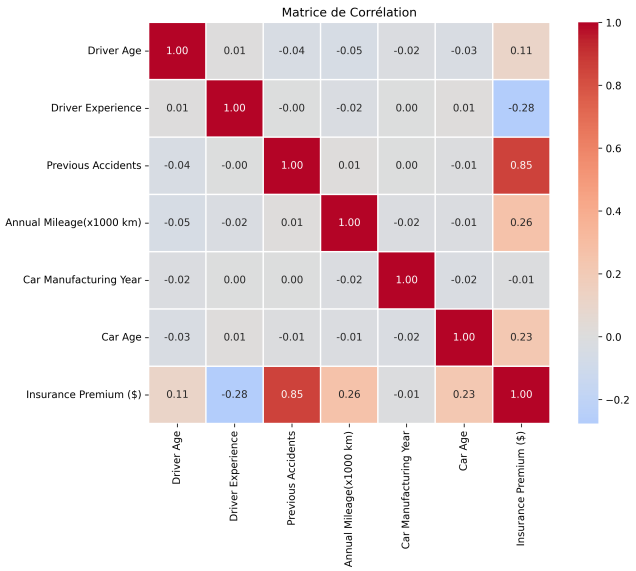


Fig. 1. Pearson correlation matrix between the variables of the dataset.

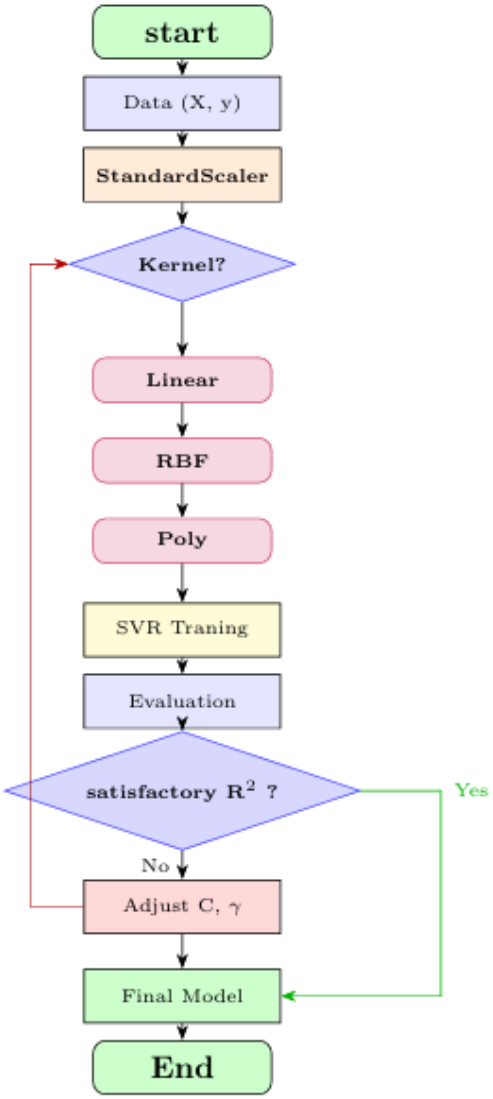


Fig. 2. SVR algorithm for dynamic pricing.

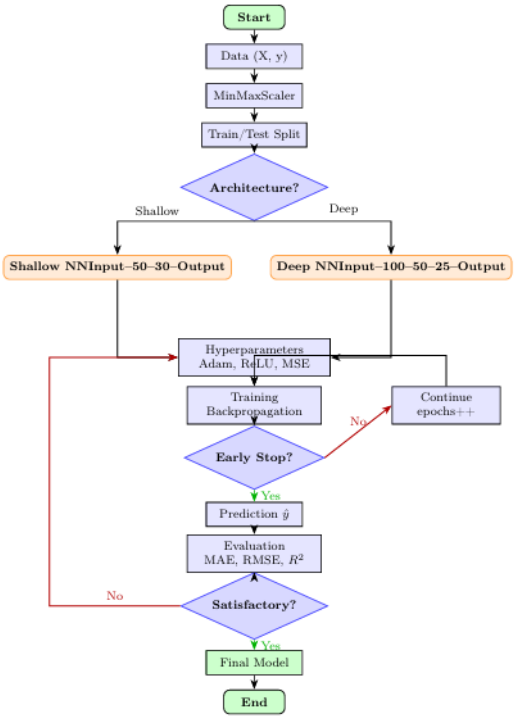


Fig. 3. Neural Network algorithm for dynamic pricing.

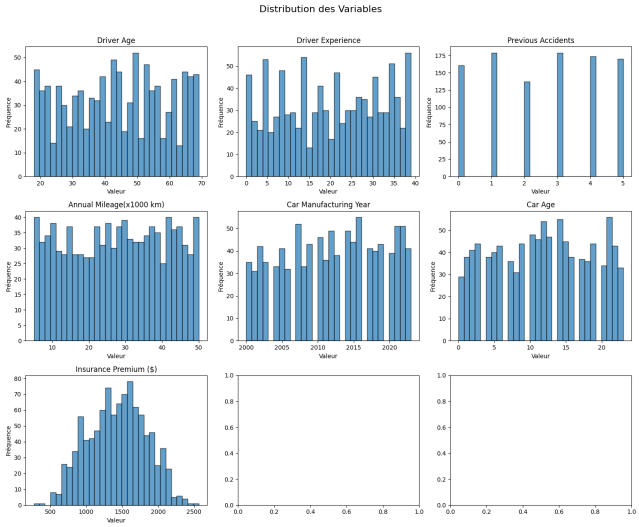


Fig. 4. Distribution of variables in the automobile insurance dataset.

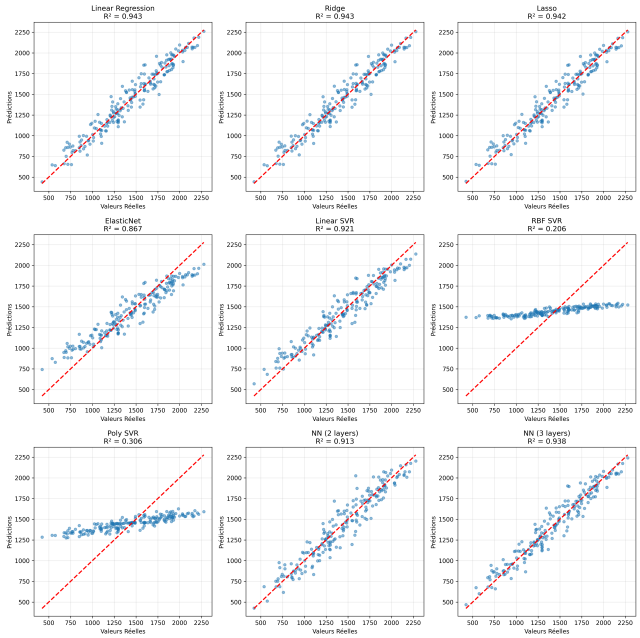


Fig. 5. Predicted versus actual insurance premiums for the nine models on the test set.

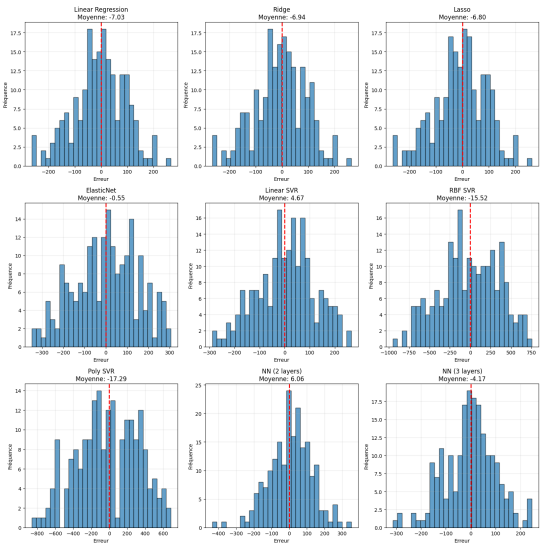


Fig. 6. Distribution of prediction errors for the different regression models.

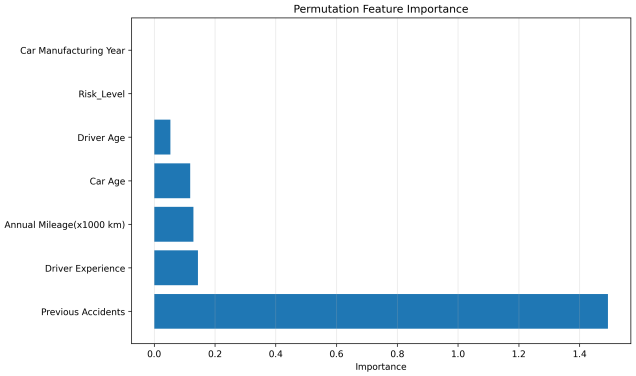


Fig. 7. Permutation feature importance for the Ridge regression model.

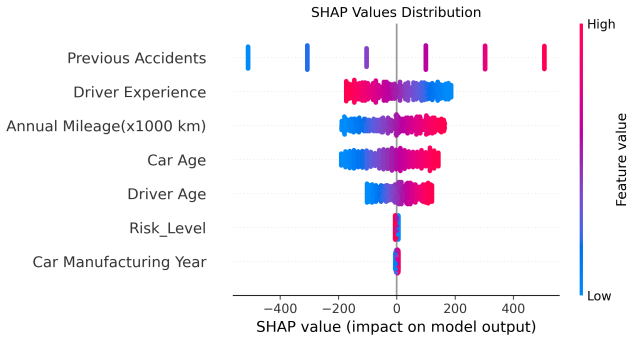


Fig. 8. SHAP summary plot showing the distribution and impact of features on premium predictions.

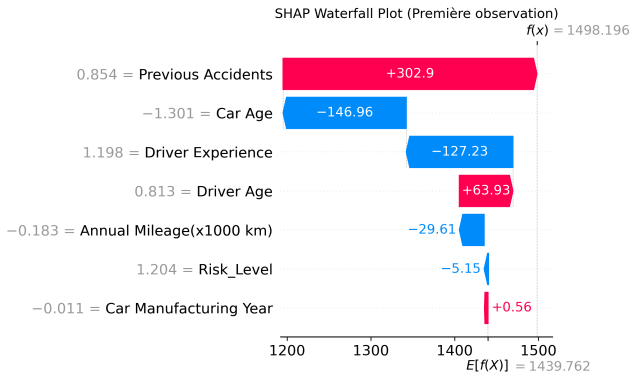


Fig. 9. SHAP waterfall plot explaining the prediction for a specific observation.

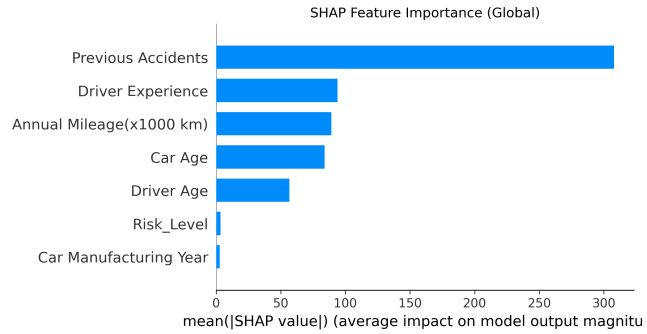


Fig. 10. Global feature importance based on mean absolute SHAP values.