QUBAHAN ACADEMIC JOURNAL
VOL. 6, NO. 1, January 2026
https://doi.org/10.48161/qaj.v6n1a2094

Effective Risk Analysis and Predictive Modeling in Motor Insurance in Saudi Arabia

Abdullah Aldaeej [1*], Hajar Aseeri [1], Atiq Siddiqui [1], Jumanah Alshehri [1] and Hafsa Alabdullateef [1]

[1] Department of Management Information Systems, Collage of Business Administration, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, 31441, Dammam, Saudi Arabia.

[*] **Corresponding author:** aaaldaeej@iau.edu.sa

ABSTRACT: Accurate prediction of motor insurance premiums that correspond with actual claims are critical to the sustainability of insurance companies. However, predicting premiums is a challenging task due to the complexity of risk factors. This study aims to identify significant risk factors and develop predictive models for motor insurance pricing within the Saudi context, using real data obtained from one of the leading insurance providers in Saudi Arabia. The dataset consists of 71,280 records and 26 features of insurance claims reported during the year of 2023. After preprocessing the data, significant risk factors are identified using Analysis of Variance (ANOVA), which are used later to build the prediction models. The findings reveal that vehicle body type and manufacturing country emerged as the most influential risk factors. The evaluation metrics (R^2, MAE, MSE) have been applied to evaluate the best-performing machine-learning pricing prediction model (Decision tree, Neural network, Generalized linear model, and Random Forest). The results of our evaluation show that the Random Forest model consistently outperformed the other models in terms of prediction accuracy. The study contributes to motor insurance industry in Saudi Arabia by supporting informed risk assessment within the Saudi Takaful insurance operations. It highlights the performance of prediction models for motor insurance pricing in Saudi Arabia.

Keywords: motor insurance; insurance premium; machine learning; risk factors; Saudi Arabia.

I. INTRODUCTION

Insurance is significant for any business as it mitigates financial losses in case of a risk realization [1]. This is done by transferring the risk to another party that will provide financial protection against potential loss in exchange for a periodic fee called the premium [2]. The insurance sector holds a fundamental role in managing risks in any economy. Globally and locally, the sector has grown rapidly. This is reflected in the fact that, in 2023 alone, the total premium reached USD 7 trillion globally, a 6.1% increase over the preceding year [3]. The insurance industry in Saudi Arabia has also witnessed high growth. As per Saudi Insurance Authority, the Gross Written Premiums (GWP) grew by 22.7% in 2023, reaching SR 53,4 billion, translating to 1.64% of GDP. Within the Saudi insurance industry, motor Insurance forms the second-largest insurance, with a 21.8% GWP share [4]. This rapid increase in insurance demand has been attributed to the Saudi government's compulsory motor and medical insurance requirements [5], and a general rise in awareness for insurance protection amongst corporations and individuals [6]. Together these global and national trends highlight the increasing importance of fair and accurate pricing mechanisms in the insurance sector.

Traditionally, insurance premiums are determined using rating tables computed by actuaries, where they consider numerous rating factors that affect the final premium [7]. In motor insurance, these rating factors depend heavily on an individual's characteristics such as his/her age, vehicle age, vehicle model [8], type of

use, coverage, territory [9], gender, and years of driver's license validity [10]. It also involves combining the expected claim frequency and severity. Actuaries typically employ these factors in a statistical model, such as Linear Regression and Generalized Linear Models (GLMs), for determining the premium price [2]. However, a key challenge with these traditional methods is that they can lead to inadequate risk differentiation, resulting in adverse selection. In this scenario, premiums for different risks are not appropriately differentiated. The premiums consequently fail to align with the risks undertaken by the company [9], which in turn negatively affects the overall business by attracting a disproportionate number of high-risk policyholders.

In the case of Saudi Arabia, the country follows the Takaful insurance system, also known as Islamic insurance, which is based on cooperation and mutual assistance principles under Islamic Sharia law [11]. Although motor insurance in Saudi Arabia is mandatory and plays a crucial role in the financial protection of vehicle owners, there remains a significant gap in research specifically addressing the unique risk factors and market dynamics, highlighting the need for accurate price prediction that ensures fair premiums [8]. Given the limitations of traditional actuarial models in capturing complex, non-linear interactions among risk factors, and the increasing need for pricing fairness within the Takaful framework, the use of advanced data-driven methods such as machine learning therefore offer a suitable approach for enhancing premium accuracy. This strengthens the connection between the identified research gap and the methodological choice used in this study.

Our objective is to address the gap in literature by presenting a comprehensive study that aim to 1) identifies the potential motor insurance risk factors faced in the Saudi context, 2) uses the Analysis of Variance to ascertain the significant risk factors affecting insurance prices, and 3) develops machine-learning-based prediction models employing the identified significant risk factors. The findings reveal that vehicle body type and manufacturing country emerged as the most influential risk factors. Among the tested models, the Random Forest model consistently outperformed the others in terms of prediction accuracy, establishing it as the best model suggested in our study.

The rest of the paper is organized as follows: Section 2 provides an overview of the literature review, discussing relevant prior research. Section 3 details the methodology employed in the study. The implementation, findings, and discussion are presented in Section 4. In Section 5, the conclusions of this study and recommendations are discussed.

II. LITERATURE REVIEW

We have organized our literature review around two streams. We first discuss the literature on risk factors influencing motor insurance premiums, and then we discuss the use of price-prediction models. Finally, we identify gaps and elaborate on how our work is positioned against the extant literature.

1. SIGNIFICANT RISK FACTORS INFLUENCING MOTOR INSURANCE PREMIUMS

As actuaries rely on risk factors for price determination, accurate and comprehensive risk consideration is crucial for ensuring the accuracy of their work. These factors typically include historical claims and demographic data; however, actual risk factors may vary based on the context and coverage of the problem. It is thus vital to identify all the associated significant risk factors. We first discuss below and identify common factors found in the literature.

Literature has identified various factors influencing premium pricing, which can broadly be categorized into insured and vehicle characteristics groups. Insured characteristics include age, gender, credit scores, and profession, while vehicle characteristics encompass factors like car age, car model, car make, and cubic capacity. A study by [12] identified car age, insured age, credit score, annual mileage, and years of no claims as major risk factors. Azaare et al. [7] suggest engine size has minimal impact on premium pricing, whereas driver age is a significant factor. Dragos and Dragos [13] identified risk preference, distance traveled by car, driver's education level, and the income-to-car price ratio as key factors influencing the purchase of voluntary motor insurance. David [2] identifies various factors, including driver age, profession, vehicle usage, bonus-malus system rating, and contract duration. The study found that premiums generally decrease

with driver and contract age but increase with a higher bonus-malus coefficient. Driving and claim histories are also significant in determining premium pricing. Factors such as annual mileage, years of no claims, and territorial clustering are identified as major risk factors [12]. The study by [14] examines claim data to identify key variables influencing claim counts, claim amounts, average loss. The findings indicate that the size of loss and coverage type are dominant factors. Similarly, another study by [9] found that territory, coverage type, accident year, reporting year, and loss size are the most significant predictors of claim outcomes.

We also found differences rooted in the problem context and the type of insurance. For instance, an analysis of car insurance pricing in Germany, Switzerland, and Austria by [15] reveals differences in data collection and criteria used. It suggests that Swiss insurers may place a greater emphasis on driver experience compared to Germany and Austria. Other predictors are related to the types of insurance. The common types of auto insurance are comprehensive and third-party. In comprehensive insurance, the insurance company covers the cost of repairing the insured car, regardless of who is at fault. In contrast, in third-party insurance, the insurance company only covers the expenses of the other party involved in the accident. For example, the study by [16] states that customers' risk behavior and claims patterns with third-party policies may differ from those with Comprehensive policies.

2. MACHINE LEARNING TECHNIQUES FOR PRICE PREDICTION

In motor insurance literature, different techniques have been used that employ various risk factors affecting insurance pricing. Among these, the generalized linear model (GLM) is most prevalent in pricing and estimating insurance losses. For instance, Xie and Luo [9] used Generalized Linear Models (GLM) to assess the impact of various factors on the size of loss distributions in automobile insurance plans. Similarly, David [2] applied GLM to determine pure premiums based on the characteristics of policyholders. Machine learning techniques, such as artificial neural networks (ANN) and decision trees, have recently emerged as a popular insurance pricing approach, seeking higher prediction accuracy [14]. Omerasevic and Selimovic [17] indicate that data mining methods like Forward Stepwise, Decision trees, and Neural networks help select prediction variables to improve the accuracy and effectiveness of premium prediction.

3. GAPS AND CONTRIBUTIONS

The literature review indicates that while various risk factors have been considered, their relevance is highly contextual. Premium determination depends on localized risk and policy factors. Moreover, we also found that significant reliance so far is on statistical models for price prediction, while there is a rise in realization of the use of machine learning based techniques due to complexities faced in these problems. Although previous studies have examined motor insurance premium determinants globally and regionally, they often rely on limited datasets or focus on traditional statistical models without fully exploring the potential of advanced machine learning techniques.

Our study stands out by focusing on the unique risk factors associated with motor insurance in the Saudi context, providing insights tailored to the region's specific cultural and regulatory environment. Although previous research provides important insights, our work makes a notable contribution by empirically analyzing a large dataset of one of the largest motor insurance companies in Saudi Arabia. We also apply ANOVA filtering to identify the most relevant predictors before modelling, and employs various machine-learning techniques, specifically artificial neural networks, decision trees, random forest, and generalized linear models to benchmark, compare and determine the best-performing price prediction model.

In summary, previous studies explored global and regional risk factors using mainly statistical models. In addition, they often missed region-specific variables and robust model comparisons and rarely employed pre-modelling variable selection techniques. Our study addresses these gaps by including Saudi-specific risk factors, using ANOVA for variable selection, leveraging a large Saudi dataset, and systematically comparing multiple machine learning and statistical models to identify the best-performing approach for motor insurance premium prediction in Saudi Arabia. Table 1 presents a summary of previous studies on motor insurance pricing, highlighting the modeling approaches, objectives, and key results of each study. This overview provides a concise reference for understanding the variety of methods applied in the literature and the main findings reported.

Table 1. Summary of literature review on motor insurance pricing.

Study	Model Approach	Objective	Result
[12]	Generalised Additive Modelling (GAM)	Aims to fill the gap between individual-level pricing and rate regulation using the UBI database.	The study showed that the variables of car age, insured age, credit scores, annual mileage and years of no claims are major risk factors as well as the territorial clustering.
[7]	Autoregressive distributed lag (ARDL) model	Attempts to provide evidence to justify which variables are significant and needed to be considered by insurers.	The driver age characteristics are significant and should be considered in the premium calculations.
[13]	Binary logit model, Multinomial logit model	Aims to apply a causal inference approach to account for customers' price sensitivity and deduce optimal, multi-period profit maximizing premium renewal offers by estimating consumer behavior.	The study found that key factors influencing the purchase of voluntary motor insurance are risk preference, distance travelled by car, driver's education level, and income-car price ratio.
[2]	Generalized Linear Models (GLMs)	Aims to use Generalized Linear Models to calculate the pure premium for auto insurance based on observable characteristics of policyholders.	The paper observes a decrease in the premium along with an increase in the insured and contract age, and an increase along with the bonus-malus coefficient growth.
[14]	Artificial neural networks (ANN)	Aims to identify the dominant risk factors and improve the transparency of the pricing models by understanding the impact of major risk factors by measuring variable importance.	The paper examines the association between claim counts, claim amounts, average loss per claim, and major risk factors such as accident year, reporting year, territory, coverage, and size of loss. The findings indicate that the size of loss and coverage play a critical role in determining claim counts, claim amounts, and average loss per claim.
[9]	Generalized linear models (GLM)	Study of using generalized linear models (GLM) for the size of loss distributions and measure the variable importance in GLM modeling.	The study indicates that the territory, coverage type, accident year, reporting year and size of loss are the most significant factor among all factors considered in the study.
[1]	Proposed Novel data-driven model	Focuses on using a data science approach to assess the risk associated with automobile insurance policies by predicting the total claims made by new customers.	The study introduces a new model for automobile insurance risk assessment and demonstrates its effectiveness.
[17]	Forward Stepwise, Decision trees, Neural networks, Generalized linear models (GLM)	The study investigated the various types of data mining techniques to select risk factors that have an impact on insurance premium rates.	The results indicate that using data mining methods will improve the accuracy and effectiveness in predicting insurance premiums.

III. METHODOLOGY

In this section, we present the methodology used in our study. Figure 1 shows the workflow of our research methodology. The first step involves data collection and exploration. We utilize the data collected by one of the largest motor insurance companies in Saudi Arabia. This extensive and comprehensive dataset is subsequently preprocessed through necessary date conversions, variable encoding, exclusion of non-usable variables, and determination and exclusion of outliers. We then performed the analysis of variance (ANOVA) to identify the significant risk factors, which are then included in the prediction models. In step four, the prediction models are fitted and calibrated. In the last step, these models are evaluated for prediction accuracy. In the following, we outline the details of each of these steps:

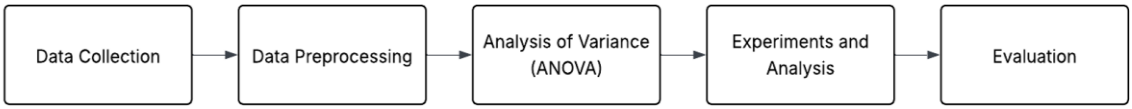

FIGURE 1. Research methodology workflow.

1. DATA COLLECTION AND EXPLORATION

The motor insurance dataset utilized in this study was collected from a leading insurance company based in Saudi Arabia, data from the year 2023 was exclusively utilized. To maintain the privacy and confidentiality of the dataset, we obtained Institutional Review Board (IRB) approval (IRB-2024-14-683). The dataset focuses on a type of insurance compulsory under Saudi law, known as Third-Party Liability insurance. The data collected consists of 26 variables that provide insights into various aspects of motor insurance policies. In total, the dataset comprises an extensive collection of 71,280 insurance policy records. The following table illustrates the variables included in the dataset:

Table 2. Risk factors.

#	Variables	Type	Description
1	Policy Number	Object	A unique identifier assigned to each policy in the dataset
2	Issue Date	Date	The date on which an insurance policy is issued
3	Policy Effective date	Date	The date when the policy becomes active, and coverage begins
4	City	Object	The specific city where the policyholder resides. There are around 900 cities in the dataset
5	Region	Object	The broader geographic area where the policyholder resides. The regions include Eastern, Northern, Central, Southern and Western
6	Policy Type	Object	Comprehensive Insurance or Third-Party Liability insurance
7	Sum Insured	Float	The vehicle market value
8	Vehicle Make	Object	The brand or manufacturer of a motor vehicle. There are around 118 different Vehicle Make in the dataset
9	Vehicle Model	Object	The specific version of a motor vehicle. There are around 700 different Vehicle Model in the dataset
10	Vehicle Manufacturing Year	Integer	The year in which a motor vehicle was produced or manufactured
11	Vehicle Body Type	Object	The general shape and configuration of a motor vehicle such as sedan and SUV
12	Vehicle Color	Object	The exterior paint on a motor vehicle
13	Vehicle Category	Object	The classification or grouping of motor vehicles such as standard, luxury and sport
14	Deductible	Integer	The portion of the claim amount that the policyholder agrees to contribute in the event of a claim
15	Repair Type	Object	Repair vehicle in workshop or agency

16	Expected Mileage	Object	The estimated number of miles a vehicle is anticipated to travel within an insurance policy period
17	Plate Type	Object	The category of license plates used on motor vehicles, indicating their purpose or use. There are 8 palate type such as private car, equipment and motorcycle
18	Insured DOB	Date	The date of birth of the policyholder
19	Insured Nationality	Object	The nationality of the policyholder
20	Insured Gender	Object	The Gender of the policyholder
21	Driver DOB	Date	The date of birth of the driver
22	Base Premium	Float	The initial amount charged by an insurance company for a policy, excluding tax and discounts
23	NCD	Integer	(No Claims Discount) is a discount given for policyholders with a claim-free record. The percentage could be 10%, 20%, 30%, 40% or 50%
24	No. of claims	Integer	The count of claims recorded in the specific policy within a specified period
25	Total amount of claims	Float	The total amount of claims recorded in the specific policy within a specified period.
26	Manufacturing Country	Object	The country where a vehicle is produced or manufactured which includes Korean, Japanese, American, Chinese, European, and Others

2. DATA PREPROCESSING

We preprocess the data to handle missing and redundant values and address inconsistencies, thereby producing a more reliable and relevant dataset for valuable insights. The detailed data processing steps are described below.

2.1 Data Conversion

Data conversion is crucial step in data preprocessing, as it standardizes variables into formats that can be effectively interpreted by machine learning algorithms, ultimately enhancing model performance. Thus, we used the driver's date of birth and vehicle manufacturing year variables to determine driver age and vehicle age.

2.2 Exclusion of Non-Usable Variables

Removing non-relevant features is crucial, as they can introduce noise, increase computational complexity, and reduce the accuracy and efficiency of machine learning algorithms by obscuring the relationships between relevant variables and the target outcome which is the process of third-party liability. These variables include policy number, issued date, policy effective date, policy type, sum insured, deductible, and repair type. Also, none of these variables were used in the literature to build prediction models. Other variables were removed due to inaccuracies in the data source. In their place, we have used alternative variables, including pairing vehicle make and vehicle model with the manufacturing country and using the region as an alternative to the city. Additionally, insured gender has been removed, as it should not be considered as a pricing factor per Saudi insurance authority. Furthermore, the insured's date of birth has been eliminated due to its redundancy with the driver's date of birth. Finally, the insured's nationality is removed because it was found incorrect in the data source. As a result, the dataset dimension was reduced to 13 variables.

2.3 Handling Missing Values

Addressing missing values is essential due to their potential impact on the analysis outcomes and the efficiency of predictive models. Missing values can be omitted or replaced by other values [18][19]. In this study, we identified the missing values in four variables: vehicle body type, NCD %, no. of claims, and total amount of claims, as shown in Table , and then applied both techniques based on the nature of the missing values.

Table 3. The missing values identified per variable.

Variable	# missing values	% of missing values
Vehicle Body Type	3,026	4.2%
NCD %	355	0.5%
No. of claims	71,132	99.8%
Total amount of claims	70,925	99.5%

2.4 Outliers Treatment

Since outliers are points that significantly deviate from the regular pattern or distribution of a dataset and thus affect model fitting, we removed these outliers. We thus evaluated the base premium variable, the only numerical variable in our dataset (Figure 2). We applied the Winsorization method, wherein the outlier values were substituted with the third quartile value [20].

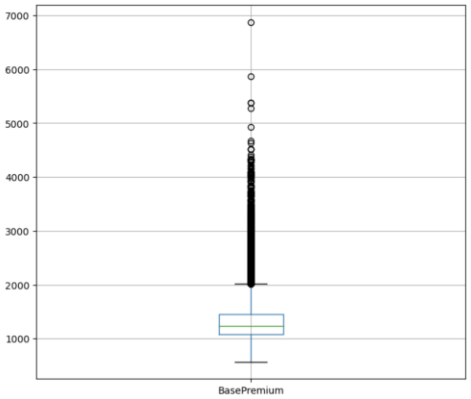

FIGURE 2. The detection of outlier in the Base Premium (dependent) variable.

2.5 Categorical Variables Encoding

In most machine learning algorithms, categorical variables cannot be directly without some form of transformation or encoding. Therefore, we used one-hot encoding techniques to transform categorical variables into numerical variables and into a format understandable by algorithms. It generates separate binary columns for each category, where a value of '1' indicates the presence of that category, while the remaining entries are filled with '0'. By doing so, machine learning algorithms can effectively handle categorical data without misinterpreting any essential ordering among the categories [21].

2.6 Data Splitting

We divided the dataset into an 80/20 split, with 80% of the data allocated for training and 20% reserved for testing. The training set was used to train the machine learning models, while the testing set provided an evaluation on unseen data. This approach is essential for assessing the model's ability to generalize to new, unobserved samples [22].

3. ANALYSIS OF VARIANCE

After cleaning and preparing the dataset, we leverage various analytics techniques, such as descriptive analytics and Analysis of Variance (ANOVA), to identify relationships between variables and develop predictive models for motor insurance pricing. In this study, ANOVA and machine learning are used as complementary steps in a unified pricing pipeline. First, we apply two-way ANOVA as a classical inferential tool to identify rating factors that have a statistically significant effect on the motor insurance outcome, providing an interpretable and transparent variable-screening step that reduces noise and the risk of overfitting. The risk factors found to be significant are then used as inputs to the machine-learning models,

which are able to capture nonlinear relationships and complex interactions and are optimized for predictive accuracy. This combined ANOVA–ML approach therefore balances actuarial interpretability and regulatory transparency with modern predictive performance in motor insurance pricing. This section will provide detailed explanations of each analytics technique used.

3.1 Descriptive Analysis

A visual exploration was conducted to gain insights into the distribution of the dataset. Various graphical methods were employed based on the variable type, whether numerical or categorical. Bar charts were utilized to visualize categorical data, while box plots were employed for numerical data. The results of the analysis will be discussed in the results section.

3.2 ANOVA Test

The ANOVA (Analysis of Variance) test determines significant differences between two or more categorical groups by testing mean differences using variance. It is commonly employed to identify significant relationships between independent and dependent variables, relying on a p-value (≤ 0.05) for significance determination and a higher F value indicating significant variables [23]. In this study, we employed two-way ANOVA, which analyzes the influence of two or more categorical predictor variables on a continuous outcome variable [24].

3.3 Predictive Analysis (Machine Learning)

We have implemented four distinct machine-learning algorithms for predicting motor insurance pricing. These models were selected because they are widely used in the literature and capable of handling the complexity and non-linearity of insurance data. Employing multiple algorithms also enhances the robustness of the analysis and allows for verification of model stability across different scenarios.

Additionally, the dataset underwent three experiments, each utilizing a different grouping of variables. We will provide an overview of each algorithm before diving into the analytical and results phase.

a) Decision Tree Regression

The Decision Tree algorithm is a supervised learning approach that can solve regression and classification tasks. Decision trees are effective tools for addressing decision-making problems, offering advantages such as interpretability, minimal data preprocessing requirements, and the ability to handle non-linear relationships effectively [25][26]. Our decision trees model is created with specific hyperparameters to optimize its performance for predicting the prices. In its hyperparameter tuning, the algorithm suggested a maximum depth of 13, a minimum of 20 samples at a leaf node, and at least 15 samples needed to split an internal node. Additionally, the model considers a maximum of 20 features when determining the best split at each node. The random state is set to 11 to ensure the reproducibility of results. The final parameters are summarized in Table 2.

Table 2. Hyperparameters of the decision tree regression model.

Hyperparameter	Value
criterion	squared_error
max_depth	13
min_samples_leaf	20
min_samples_split	15
max_features	20
random_state	11
splitter	best
min_weight_fraction_leaf	0.0
max_leaf_nodes	None
ccp_alpha	0.0

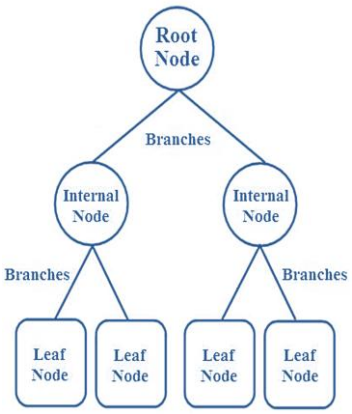

FIGURE 3. Decision tree model structure [25].

b) Neural Networks

Neural networks, also known as artificial neural networks, are models used for classification and regression. They are inspired by the biological activity in the brain, where interconnected neurons learn from experience. The key strength of neural networks lies in their exceptional predictive performance. Their structure captures complex relationships between predictors and outcome variables, which is often challenging for other predictive models [27, 28]. Our model consists of two hidden layers, each with 100 neurons. The activation function used is 'logistic', and the model is configured to run for a maximum of 5000 iterations during training. The final parameters are summarized in Table 3.

Table 3: Hyperparameters of the neural networks model.

Hyperparameter	Value
hidden_layer_sizes	(100,100)
activation	'logistic'
max_iter	5000
solver	adam
alpha	0.0001
batch_size	auto
learning_rate	constant
learning_rate_init	0.001
power_t	0.5
shuffle	True
random_state	None
tol	1e-4
verbose	False
warm_start	False
momentum	0.9
nesterovs_momentum	True
early_stopping	False
validation_fraction	0.1
beta_1	0.9
beta_2	0.999
epsilon	1e-8

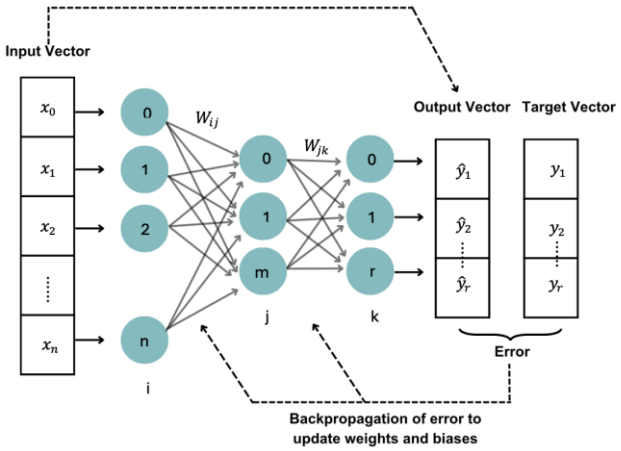

FIGURE 4. Neural networks model structure [14].

c) Generalized Linear Models (GLMs)

GLMs have emerged as a widely adopted approach in non-life insurance pricing. These models are an extension of the Gaussian linear model framework derived from the exponential family, as demonstrated by [2]. In this model, we only specified the Gaussian family to predict the dependent variable. The final parameters are summarized in Table 4.

Table 4. Hyperparameters of the generalized linear models.

Hyperparameter	Value
family	Gaussian
link	identity
offset	None
exposure	None
freq_weights	None
var_weights	None
missing	none

d) Random Forest

Random forest is a supervised learning algorithm that constructs an ensemble of multiple decision trees to achieve a more accurate prediction. It builds multiple decision trees on random subsets of the training data and then combines their outputs [29, 30]. Our model was created with 1000 trees. It requires a minimum of 4 samples to split a node and 4 samples in a leaf node. It uses the square root of the number of features to determine the best split and has a maximum depth of 20 for each tree. The final parameters are summarized in Table 5.

Table 5. Hyperparameters of the random forest model.

Hyperparameter	Value
n_estimators	1000
min_samples_split	4
min_samples_leaf	4
max_features	sqrt
max_depth	20
criterion	squared_error

min_weight_fraction_leaf	0.0
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	True
oob_score	False
n_jobs	None
random_state	None
verbose	0
warm_start	False
ccp_alpha	0.0
max_samples	None

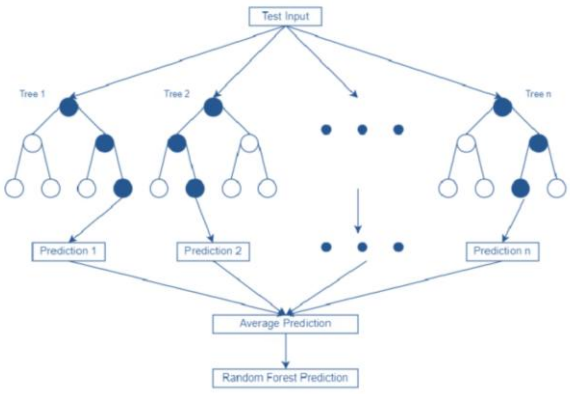

FIGURE 5. Random forest model structure [14].

e) Evaluation Metrics

We used multiple evaluation criteria to measure the effectiveness of the fitted models. The measures include:

- R-squared (R2): A statistical metric used to assess the goodness of fit of a regression model. It is a value that ranges between 0 and 1, and higher values indicate a perfect fit of the model to the data [31].
- Mean Absolute Error (MAE): An effective metric used to evaluate the accuracy of regression models. It calculates the average absolute difference between predicted and target values [32].
- Mean Square Error (MSE): A metric used to evaluate the performance of predictive models. It calculates the average of the squared differences between predicted and actual target values in a dataset. The main purpose of MSE is to evaluate the accuracy of a model's predictions by quantifying how closely they match the true values [33].

IV. RESULTS AND DISCUSSION

Following data preprocessing, the dimensionality was effectively reduced to include 10 categorical variables and one numerical variable (the dependent variable), while retaining the complete policy records. Below is a detailed description of the analysis results.

1. MISSING VALUES TREATMENT

We identified missing values in four variables: 3,026 in Vehicle Body Type, 355 in NCD%, 71,132 in No. of claims, and 70,925 in the total amount of claims, as explained in section Handling Missing Values. Both missing value-handling techniques were implemented. Firstly, we replaced 3,026 missing values of the 'Vehicle Body Type' variable with 'Others' and corrected 355 missing NCD% values using the original values

identified in the data source. Furthermore, we omitted the variables 'No. of claims' and 'Total amount of claims' because a significant portion of records was missing.

2. OUTLIER DETECTION AND RESOLUTION

We detected the outliers in the base premium variable, the numerical variable. FIGURE .a displays the distribution of the base premiums before addressing the outliers, showing a significant skewness in the data that could potentially influence the analysis. In contrast, FIGURE .b presents the base premium distribution after the outliers have been handled, showcasing the skewness changes that occurred due to outlier handling.

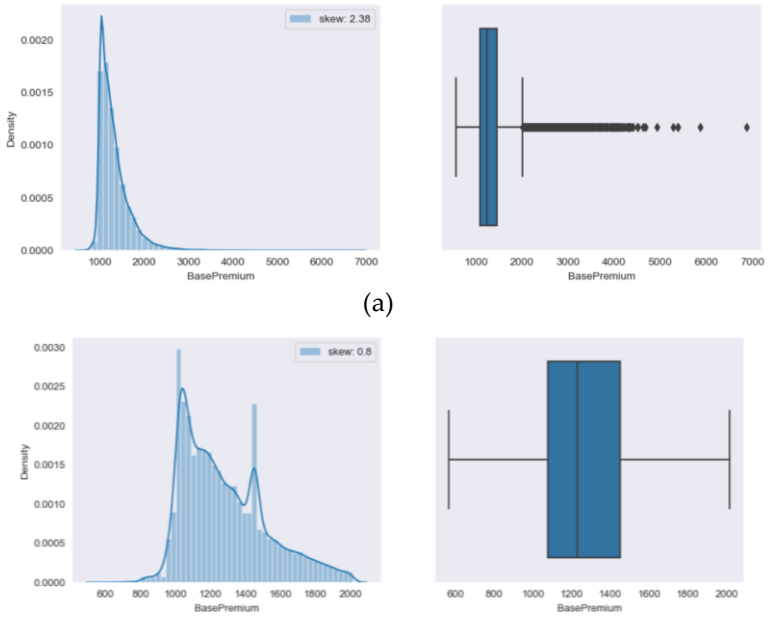

FIGURE 6. Outlier detection and treatment - base Premium variable.

3. DESCRIPTIVE ANALYSIS

We performed a descriptive analysis of the variable to understand the distribution needed in evaluating their role. FIGURE shows the distribution of each input variable. As shown in FIGURE .a, the region variable is distributed into four main categories: Eastern, Northern, Central, and Others, which consist of the Western and Southern regions, with respective record counts of 27,045, 20,538, 11,504, and 9,167. In FIGURE .b, the white color is dominant among all other color categories in the vehicle color variable with 44,060 values, followed by others with 13,094 values. The other category consists of 35 colors, such as blue, red, and green.

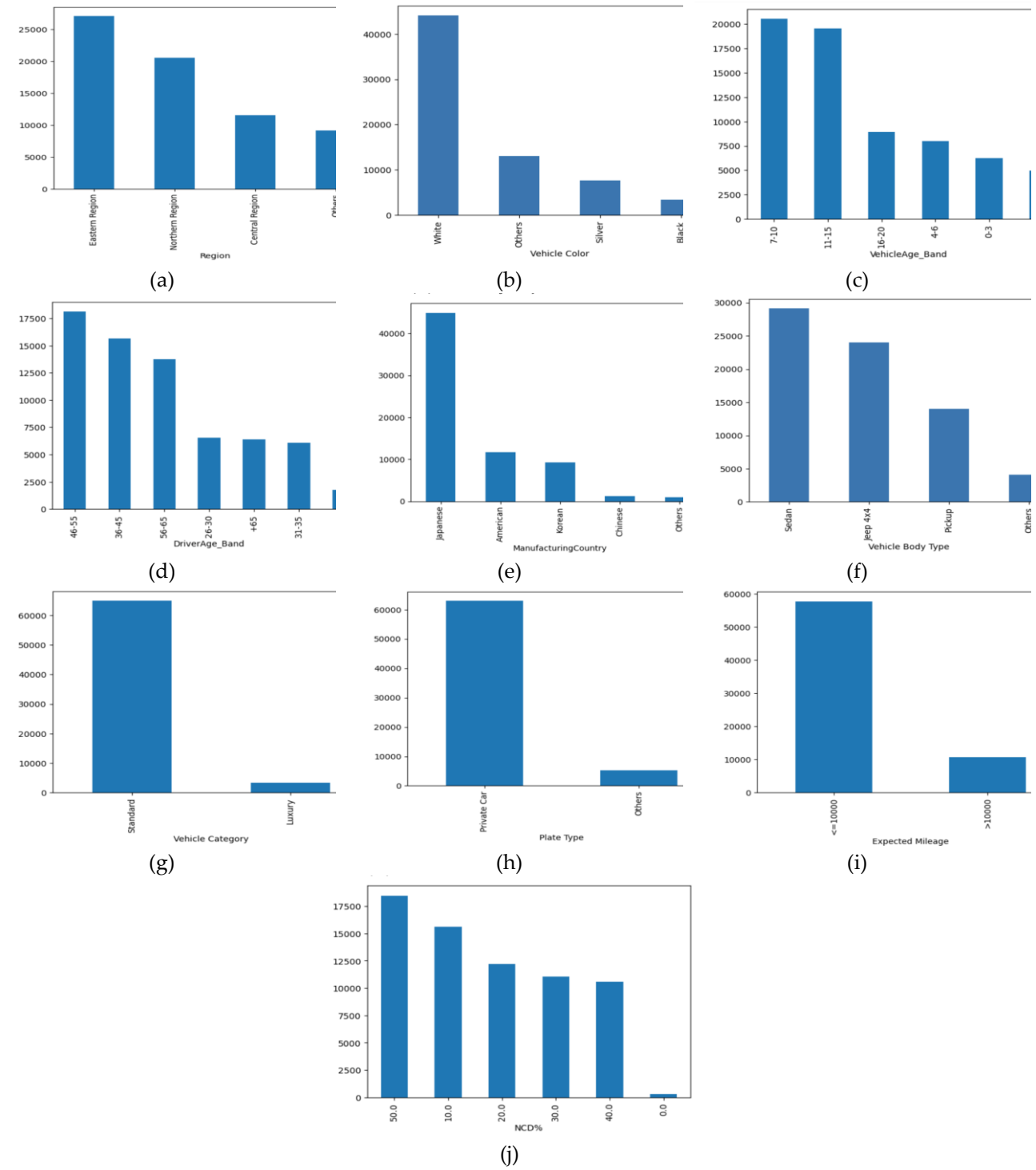

FIGURE 7. Distribution of categorical variables.

FIGURE .c shows the distribution of insured vehicles based on the vehicle age band, which illustrates that the most insured vehicles fall into the 7-10 years category with 20,526 vehicles, followed by the 11-15 years category with 19,569 vehicles and the 16-20 years category with 8,938 vehicles. The remaining 19,221 vehicles

are divided into the 4-6 years, 0-3 years, and more than 20 years categories. Similarly, FIGURE .d showcases the distribution of insured vehicles based on the driver's age, showing the following breakdown: the 46-55 category with 18,131 vehicles, the 36-45 category with 15,675 vehicles, and the 56-65 category with 13,823 vehicles. The remaining 20,625 vehicles are distributed across the 26-30, 31-35, more than 65, and less than 25-year age bands.

The majority of the insured vehicles were manufactured in Japan, with 44,957 vehicles, as illustrated in FIGURE .e, while the rest of the insured vehicles were manufactured in America, South Korea, China, and other countries. As in FIGURE .f, the distribution of the vehicle body type is divided into sedan, Jeep 4×4, and others, including vans, trucks, buses, motorcycles, and equipment. Most insured vehicles fall into the sedan and Jeep 4x4 categories, with respective values of 29,119 and 24,054.

FIGURE .g, 7.h, and 7.i show the binary variables along with their dominant categories. The vehicle category variable's dominant category is "Standard," with a count of 64,932. The plate type variable's dominant category is "Private Car", with a count of 62,936. Lastly, the expected mileage variable's dominant category is "<=10000", with a count of 57,663. Finally, FIGURE .j shows the distribution of the no-claim discount (NCD) variable. The counts for each category are as follows: 50% with a count of 18,460, 10% with a count of 15,624, 20% with a count of 12,222, 30% with a count of 11,061, 40% with a count of 10,561, and 0% with a count of 326.

4. ANALYSIS OF VARIANCE

Given that we have ten potential drivers of insurance price premium, we then performed an analysis of variance to see which of these has any impact of the dependent or response variable. The ANOVA test results are shown in FIGURE . Clearly, all variables are statistically significant based on the p-values, which are less than 0.05, except for the 'PlateType' variable, where the p-value is 0.081, which is greater than 0.05. Additionally, the results indicate that 'Vehicle Body Type' and 'Manufacturing Country' are the most influential variables, as they have the highest F-values, with values of 2,802.89 and 2,792.97, respectively.

	df	sum_sq	mean_sq	F	PR(>F)
VehicleAge_Band	5.0	1.300807e+08	2.601614e+07	686.518475	0.000000e+00
DriverAge_Band	6.0	3.302463e+08	5.504105e+07	1452.432933	0.000000e+00
Region	3.0	1.598894e+08	5.329646e+07	1406.396512	0.000000e+00
VehicleColor	3.0	9.127337e+07	3.042446e+07	802.845961	0.000000e+00
ManufacturingCountry	4.0	4.233661e+08	1.058415e+08	2792.965247	0.000000e+00
VehicleBodyType	3.0	3.186449e+08	1.062150e+08	2802.819846	0.000000e+00
VehicleCategory	1.0	1.934627e+07	1.934627e+07	510.512789	1.223132e-112
PlateType	1.0	1.156944e+05	1.156944e+05	3.052965	8.059420e-02
ExpectedMileage	1.0	1.221174e+06	1.221174e+06	32.224546	1.378766e-08
C(NCD)	5.0	1.172720e+07	2.345440e+06	61.891876	1.286449e-64
Residual	71247.0	2.699959e+09	3.789576e+04	NaN	NaN

FIGURE 8. ANOVA test result.

When comparing our results with related work, we validated the significance of factors commonly identified in previous studies. For instance, car age and annual mileage were found to be influential by [12], while driver age was highlighted by [7, 2], which aligns closely with our analysis. Moreover, we introduced new significant risk factors, namely Vehicle body type and Manufacturing Country, that have not been noticeably covered in the related work. We also acknowledge that additional risk factors, such as size of the loss, years of no claims, policyholder's claim history, and credit scores, are commonly considered in motor insurance pricing [9, 12, 14]. However, due to our dataset's limitations, these factors were not included in our

study. Future studies with more comprehensive datasets could explore the impact of these factors on motor insurance pricing.

5. PREDICTIVE MODELING

Three different experiments were conducted on the dataset, each employing a unique grouping of variables. These experiments aim to evaluate and compare the performance of various machine learning algorithms in predicting motor insurance prices, utilizing distinct groupings of variables for each experiment. To facilitate the evaluation of the model in a relative sense and benchmark the performance considering industry practices, we used GLM as a benchmark, where any improvement over it will demonstrate a better-performing approach. Detailed explanations of the methodologies and outcomes of each experiment is provided in subsequent subsections.

- Experiment 1: The objective of this experiment is to determine which machine learning algorithm performs best in predicting the prices, when all nine significant variables are including: VehicleAge_Band, DriverAge_Band, Region, Vehicle Color, Manufacturing Country, Vehicle Body Type, Vehicle Category, Expected Mileage, and NCD, ensuring that no significant variables are omitted or excluded from the analysis. Table presents the results of experiment 1, where the Random Forest model achieves the highest R-squared (R2) value of 0.391.

This indicates that, relative to the other models considered, Random Forest provides the strongest explanatory power for variation in the base premium, even though the overall R^2 remains moderate. Additionally, the Random Forest model exhibits the lowest mean absolute error (MAE) and mean squared error (MSE), suggesting comparatively smaller average deviations between its predictions and the observed values. In terms of comparison with the benchmarking GLM model, the Random Forest and the Decision Trees outperformed GLM, while the Neural Network performed slightly below GLM.

Table 8. Results of the model performance in experiment 1**.**

	Decision Trees	**Neural Network**	**GLM**	**Random Forest**
R^2	0.367767	0.345056	0.352242	0.393055
MAE	147.007637	147.484208	149.415192	144.240280
MSE	37262.566123	38601.089002	38177.594052	35772.156352

- Experiment 2: This experiment aims to determine the most effective model in predicting the prices, focusing on the most significant variables identified through the ANOVA test, which yielded an F-value greater than 1000. These variables include Region, Vehicle Body Type, Manufacturing Country, and DriverAge_Band. Table illustrates the results of the experiment. Decision Trees, Neural Networks, and Random Forest models perform similarly, with R-squared (R2) values of 0.35, 0.352, and 0.351, respectively. The MAE and MSE for these models are approximately 150.00, 150.13, 150.06 and 38277.30, 38215.11, 38250.91, respectively. These results indicate that, when restricted to only the most influential variables, the models perform at a comparable and moderately predictive level. All three machine learning models here outperformed the benchmarking GLM model.

Table 9. Results of the model performance in experiment 2.

	Decision Trees	**Neural Network**	**GLM**	**Random Forest**
R^2	0.350550	0.351605	0.325137	0.350998
MAE	150.00425	150.125593	153.9730734	150.064019
MSE	38277.300308	38215.109708	39775.050164	38250.911429

- Experiment 3: This experiment aims to identify the best model in predicting the prices by excluding the least significant variables, which were identified through an ANOVA test with an F-value of less than 100. The excluded variables in this experiment were PlateType, NCD, and Expected mileage. Table presents the results, where the Random Forest model achieved the highest R-squared (R2) value of 0.38, as well as the lowest MAE and MSE values of 145.38 and 36346.57, respectively. These results indicate that the Random Forest model performed slightly better than the other models in this reduced-variable setting. As in Experiment 2, all three machine learning models outperformed the benchmarking GLM model.

Table 10. Results of the model performance in experiment 3.

	Decision Trees	Neural Network	GLM	Random Forest
R^2	0.368886	0.355731	0.349893	0.383308
MAE	146.577918	147.034412	149.071710	145.38311
MSE	37196.574098	37971.891040	38316.01487	36346.567031

Based on the comprehensive analysis of the results from the three distinct experiments, it is clear that the models show relatively similar levels of performance. However, upon closer examination, the Random Forest model consistently comparatively demonstrates a superior performance across all experiments. The Random Forest model achieved the highest R-squared (R2) values of 0.39, 0.35, and 0.38 in the three experiments, respectively, and maintained the lowest MAE and MSE values, indicating smaller average prediction errors relative to the other models and GLM.

It is important to note that the R^2 values in the range of approximately 0.35–0.39 indicate moderate, rather than high, predictive strength. This implies that a substantial portion of the variability in the base premium remains unexplained by the observable variables used in this study. In practice, this is not unexpected: pricing decisions in motor insurance are often influenced by additional risk characteristics and underwriting rules that were not available in our dataset (for example, detailed claim history, severity of past losses, credit-based measures, or company-specific pricing policies), as well as inherent randomness in claim occurrence and insurer behavior. As a result, the models should be interpreted as providing useful but partial explanatory and predictive power. They are particularly valuable for relative pricing, risk segmentation, and scenario analysis, but they are not intended to yield perfectly precise predictions at the individual policy level.

Furthermore, the analysis indicates that including all significant variables is beneficial for improving model performance, as shown in Experiment 1. When all nine significant variables were included, the Random Forest model achieved the best overall performance with the highest R2 value of 0.39 and the lowest MAE and MSE values. In contrast, when only the most significant variables were included in Experiment 2, the models performed similarly, with R2 values around 0.35. Similarly, when the least significant variables were excluded in Experiment 3, the Random Forest model still outperformed the other models, but the performance was less pronounced than in Experiment 1. This suggests that even variables with relatively smaller individual effects can contribute incrementally to predictive performance when combined within flexible models such as Random Forests. Figure 9 visualizes the performance comparison across models over the three experiments.

The performance metrics reported in Tables 8–10 is all computed on the same evaluation sample, which allows a fair comparison of the generalization performance of the different models. Across the three experimental settings, the Random Forest model consistently attains the highest R^2 and the lowest MAE and MSE among the considered approaches, suggesting that it achieves a more favorable bias–variance trade-off in this application. This behavior is consistent with the theoretical properties of Random Forests, where aggregating many decision trees reduces variance and mitigates overfitting compared with a single tree. By contrast, GLM is more restrictive in capturing nonlinearities and interactions, and the neural network may require richer input features or larger sample sizes to fully realize its potential. Nevertheless, the gains of Random Forest over the other models are moderate, and the R^2 values remain in a moderate range, indicating

that a substantial portion of the variability in premiums is still unexplained. We therefore interpret the models as useful tools for relative risk segmentation and pricing support rather than as perfectly accurate predictors at the individual policy level, and we acknowledge the absence of a more detailed residual and train–test gap analysis as a limitation that future work could address.

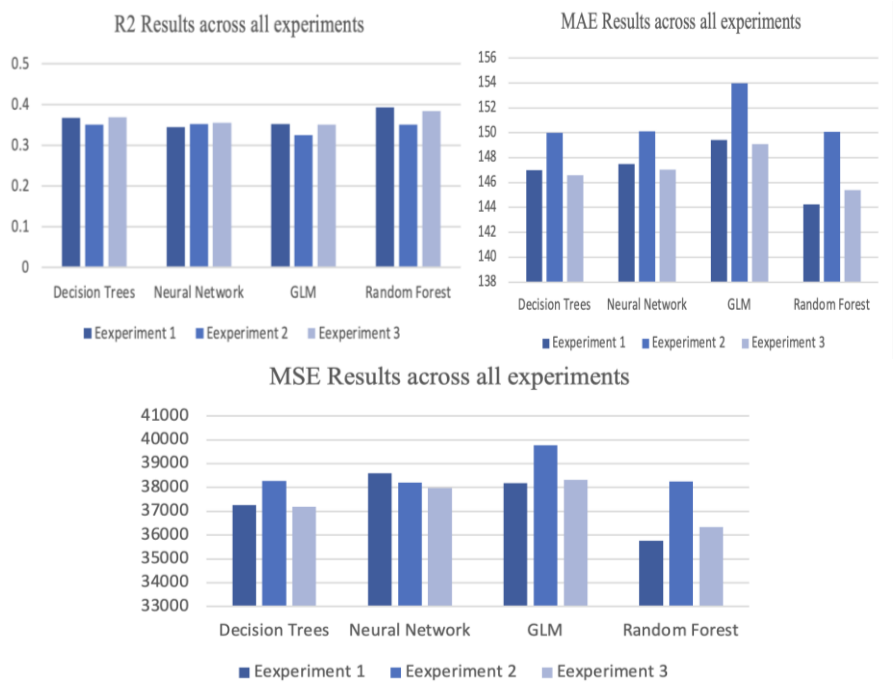

FIGURE 9. R^2, MAE, and MSE of models across three experiments.

To relate these findings to previous studies, GLM has been the most prevalent method for estimating premiums, as reported by [2, 9]. Our study demonstrates that machine learning models, particularly Random Forest, outperform GLM in prediction accuracy. This aligns with observations by [17], who highlighted the benefits of data mining techniques including decision trees and neural networks for selecting predictive variables and improving model performance. Unlike previous studies, which mainly focused on traditional statistical methods or individual machine learning models, our experiments comparatively evaluate multiple machine learning approaches on a Saudi Takaful dataset, thereby extending previous research and providing evidence for the superiority of ensemble methods in this context.

Table 6. Comparative performance of machine learning models across experiments.

Model	R^2 Exp 1	R^2 Exp 2	R^2 Exp 3	MAE Exp 1	MAE Exp 2	MAE Exp 3	MSE Exp 1	MSE Exp 2	MSE Exp 3
Decision Trees	0.37	0.35	0.37	147.01	150.00	146.58	37262.57	38277.30	37196.57
Neural Network	0.35	0.35	0.36	147.48	150.13	147.03	38601.09	38215.11	37971.89
GLM	0.35	0.33	0.35	149.42	153.97	149.07	38177.59	39775.05	38316.01
Random Forest	0.39	0.35	0.38	144.24	150.06	145.38	35772.16	38250.91	36346.57

The findings of this study contribute to understanding risk factors and model selection in motor insurance pricing within Saudi Arabia. The insurance companies can gain insights by replicating our methodological procedure on their dataset, promoting data-driven approach for insurance pricing strategies. They provide practical evidence that a combined ANOVA–ML pipeline can support moderate but meaningful predictive performance and interpretable risk factor assessment.

The study has some limitations that might undermine its findings. First, the dataset was limited motor insurance claims reported in only one leading insurance company in Saudi Arabia at 2023. This might affect the generalizability of our findings to other companies and context. However, the size of our dataset (71,280 records) was adequate to build prediction models that are tailored to one company. Second, the identification of feature significant was based on ANOVA measure only. Future studies are encouraged to adopt multiple approaches for feature selection, incorporating additional risk factors, and leveraging richer policyholder and claims histories. Third, we adopted only four ML models. Despite their commonality in the literature, future studies should explore diverse modeling techniques (e.g., gradient boosting, generalized additive models, or hybrid actuarial–ML frameworks) that were not considered in our study, to further advancing the field of motor insurance pricing.

V. CONCLUSION

Motor insurance is a critical sector that plays a significant role in fostering economic growth. Competitive pricing is a foundation and key success factor for insurance companies, as they should identify a fair premium that covers the expected losses to generate profit and remain competitive in the insurance market. Our study aims to determine the most significant risk factors that impact motor insurance pricing in Saudi Arabia and identify the most effective pricing predictive techniques. Firstly, we collected the motor insurance dataset from a Saudi insurance company. Subsequently, we employed various data preprocessing techniques and conducted an ANOVA test to identify the major significant risk factors based on the p and F values. Our findings showed that all factors included in the test are significant except the plate type variable. Among the significant factors, Vehicle body type and manufacturing country emerged as the most influential, demonstrating high F values. We also implemented four different machine learning models: Decision Trees, Neural Networks, a Generalized Linear Model, and Random Forest. The results indicated that these models performed relatively similarly, with the Random Forest model being identified as the optimal choice.

Our study examined the motor insurance dataset in Saudi Arabia, which had not been explored by researchers. We looked at various risk factors and carefully evaluated different ways to choose models. This helped us learn how motor insurance prices can be set in the kingdom. The insights gained from this study can support Saudi insurance companies in enhancing their pricing strategies and improving the accuracy of their risk assessments, thereby enabling more efficient operations and better customer service. In particular, the study demonstrates how data-driven models can strengthen pricing accuracy and inform decision-making within Saudi Takaful insurance operations. The identification of key risk factors such as vehicle body type and country of manufacture provides clear direction for refining pricing policies, while the superior performance of the Random Forest model highlights the value of adopting advanced analytical tools to improve operational efficiency and reinforce risk assessment practices.

We offer the following recommendations for future research based on the study results. Firstly, we suggest incorporating a more comprehensive dataset (across multiple insurance companies) that includes a wide array of critical risk factors, especially those not examined in our study, such as loss history information. This addition is expected to greatly improve our understanding of the model and its predictive capability. Secondly, we recommend collecting data from multiple insurance companies across Saudi Arabia to gain a broader perspective and potentially discover unique insights. Lastly, exploring alternative modeling approaches can provide new insights and lead to more effective predictive strategies.

Funding Statement

The authors have not received any funding for this work.

Author Contributions

Abdullah Aldaeej: Conceptualization, Investigation, Data curation, Writing – review and editing. Hajar Aseeri: Data collection, Data curation, Investigation, Writing - original draft. Atiq Siddiqui: Methodology, Writing – review and editing. Jumanah Alshehri: Methodology, Writing – review and editing. Hafsa Alabdullateef: Literature review, Writing – review and editing

Conflicts of interest

The authors declare that they have no conflict of interest or financial interest that could have appeared to influence the work reported in this paper.

Data availability

The authors used data from an insurance company in Saudi Arabia, representing motor insurance records in 2023.

Ethics approval

To maintain and protect the privacy and confidentiality of the dataset, the authors obtained Institutional Review Board (IRB-2024-14-683).

REFERENCES

1. Hanafy, M., & Ming, R. (**2022**). Classification of the insureds using integrated machine learning algorithms: A comparative study. *Applied Artificial Intelligence, 36*(1).
2. David, M. (**2015**). Auto insurance premium calculation using generalized linear models. In D. Airinei, C. Pintilescu, D. Viorica, & M. Asandului (Eds.), *Globalization and higher education in economics and business administration—GEBA 2013* (Vol. 20, pp. 147–156). Elsevier.
3. Swiss Re Institute. (**2024**). *World insurance report 2024*. Swiss Re Institute.
4. Insurance Authority. (**2023**). *Insurance sector reports*. Insurance Authority of Saudi Arabia.
5. Argaam. (**2024**). *S&P expects GCC insurance sector to grow driven by Saudi market*. ArgaamPlus.
6. Statista. (**2024**). *Insurances—Saudi Arabia: Market forecast*. Statista.
7. Azaare, J., Wu, Z., & Ahia, B. N. K. (**2022**). Exploring the effects of classical auto insurance rating variables on premium in ARDL: Is the high policyholders' premium in Ghana justified? *SAGE Open, 12*(4), 21582440221134219.
8. Yang, Y., Qian, W., & Zou, H. (**2018**). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics, 36*(3), 456–470.
9. Xie, S., & Luo, R. (**2022**). Measuring variable importance in generalized linear models for modeling size of loss distributions. *Mathematics, 10*(10), 1630.
10. Gómez-Déniz, E., & Calderín-Ojeda, E. (**2021**). A priori ratemaking selection using multivariate regression models allowing different coverages in auto insurance. *Risks, 9*(7), 137.
11. Alhumoudi, Y. (**2013**). *Islamic insurance takaful and its applications in Saudi Arabia* (Doctoral thesis). Brunel University.
12. Xie, S., & Shi, K. (**2023**). Generalized additive modelling of auto insurance data with territory design: A rate regulation perspective. *Mathematics, 11*(2), 334.
13. Dragos, C. M., & Dragos, S. L. (**2017**). Estimating consumers' behavior in motor insurance using discrete choice models. *E&M Ekonomie a Management, 20*(4), 88–102.
14. Xie, S. (**2021**). Improving explainability of major risk factors in artificial neural networks for auto insurance rate regulation. *Risks, 9*(7), 126.
15. Laas, D., Schmeiser, H., & Wagner, J. (**2016**). Empirical findings on motor insurance pricing in Germany, Austria, and Switzerland. *The Geneva Papers on Risk and Insurance – Issues and Practice, 41*(3), 398–431.
16. Hosein, P. (**2023**). A data science approach to risk assessment for automobile insurance policies. *International Journal of Data Science and Analytics*.
17. Omerasevic, A., & Selimovic, J. (**2020**). Risk factor selection with data mining methods for insurance premium ratemaking. *Zbornik Radova Ekonomskog Fakulteta u Rijeci, 38*(2), 667–696.
18. Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (**2017**). *Data mining for business analytics: Concepts, techniques, and applications in R* (1st ed.). Wiley.

19. Salih, M. S., Ibrahim, R. K., Zeebaree, S. R., Asaad, D., Zebari, L. M., & Abdulkareem, N. M. (2024). Diabetic prediction based on machine learning using PIMA Indian dataset. *Communications on Applied Nonlinear Analysis*, *31*(5s), 138-156.
20. Abuzaid, A., & Alkronz, E. (**2024**). A comparative study on univariate outlier winsorization methods in data science context. *Statistica Applicata – Italian Journal of Applied Statistics*.
21. Nitika, S. (**2025**). *One-hot encoding using categorical data*. Analytics Vidhya.
22. Liu, H., & Cocea, M. (**2017**). Semi-random partitioning of data into training and test sets in granular computing context. *Granular Computing, 2*(4), 357–386.
23. Simkus, J. (**2023**). *ANOVA test statistics: Analysis of variance*. Simply Psychology.
24. Daines, R. (**2024**). *Two-way ANOVA*. Statistics resources, LibGuides.
25. Gurucharan, K. (**2020**). *Machine learning basics: Decision tree regression*. Medium.
26. Zebari, D. A., Sulaiman, D. M., Sadiq, S. S., Zebari, N. A., & Salih, M. S. (2022, September). Automated Detection of Covid-19 from X-ray Using SVM. In *2022 4th International Conference on Advanced Science and Engineering (ICOASE)* (pp. 130-135). IEEE.
27. Ahmed, F. Y., Masli, A. A., Khassawneh, B., Yousif, J. H., & Zebari, D. A. (2023). Optimized Downlink Scheduling over LTE Network Based on Artificial Neural Network. *Computers*, *12*(9), 179.
28. Salih, M. S., Zebari, N. A., Masoud, R., & Zebari, D. A. (2025). Deep Transfer Learning and Feature Fusion for Improving Facial Expression Recognition on JAFFE Dataset. *Applied Computing Journal*.
29. Mohammed, M. A., Lakhan, A., Zebari, D. A., Abdulkareem, K. H., Nedoma, J., Martinek, R., ... & Tiwari, P. (2023). Adaptive secure malware efficient machine learning algorithm for healthcare data. *CAAI Transactions on Intelligence Technology*.
30. Donges, N. (**2024**). *Random forest: A complete guide for machine learning*. Built In.
31. Quantified Trading. (**2024**). *R-squared: Definition, formula, uses, and pros and cons*. Quantified Strategies.
32. Ahmed, M. W. (**2023**, August 24). *Understanding mean absolute error (MAE) in regression: A practical guide*. Medium.
33. Encord. (**2023**). *Mean square error (MSE)*. Encord machine learning glossary.

Appendix

This section provides the technical details of the data preprocessing, descriptive analysis and analysis of variance steps carried out in this study. The implementations were conducted using Python and common libraries for data manipulation and visualization.

Data Preprocessing

The dataset was prepared for analysis by handling missing and redundant values and addressing inconsistencies to create a clean and reliable dataset.

Libraries: Python with pandas and numpy were primarily used for these operations.

Steps:

- Handling missing values and redundant columns.
- Converting dates into ages for drivers and vehicles.
- Outlier detection and treatment for numerical variables (base premium) using the Winsorization method, replacing extreme values with the third quartile (Q3) as suggested by (Abuzaid & Alkronz, 2024).
- Encoding categorical variables using One-Hot Encoding to convert them into numerical format suitable for machine learning algorithms (Nitika, 2025).

Descriptive Analysis

A visual exploration of the dataset was conducted to understand the distribution and patterns in the data.

Techniques:

- Bar charts for categorical variables.
- Box plots for numerical variables.

ANOVA Test

A two-way Analysis of Variance (ANOVA) was performed to identify significant relationships between categorical predictors and the continuous outcome variable (base premium).