STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, September 2025, pp 1440–1457.
Published online in International Academic Press (www.IAPress.org)

Advanced Strategies for Predicting and Managing Auto Insurance Claims Using Machine Learning Models

Chadia Bekkaye [1*], Hassan Oukhouya [2,1*], Tarek Zari [1], Raby Guerbaz [1], Hicham El Bouanani[1]

[1]*Laboratory MAEGE, Department of Statistics and Applied Mathematics, FSJES Ain Sebaâ, Hassan II University of Casablanca, Morocco*
[2]*LaMSD, MSASE, Department of Economics, FSJES, University Mohammed First of Oujda, Morocco*

Abstract The high severity of automobile claims, which continues to rise, necessitates developing novel approaches for effectively handling claims. Machine Learning (ML) represents an essential solution to this issue of concern. As improving customer service remains the primary goal of auto insurers, the companies in question have naturally begun to adopt and use ML to better comprehend and evaluate their dataset more efficiently. This paper contributes scientifically to the pricing of car insurance, in particular, it focuses on the modeling of the total claims amount by ML models such as Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost) and Multi-Layer Perceptron (MLP). Further, a comparative analysis will help in this case by opting for statistical metrics (e.g. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE),...) as well as stochastic ones (e.g. Difference Score, Weight Difference Score,...) in both train and test data sets. The result shows that the SVR algorithm, originally tuned by Randomized Search CV, achieves excellent precision and surpasses other models tested, as seen in the Taylor diagram. This model, by contrast, shows less efficient visual distribution of predictions than XGBoost and MLP algorithms. The ultimate value of this study resides in the profound analysis of the data set, which can offer insurers adequate comprehension to manage these losses effectively.

Keywords Total Claim Amount, Machine Learning, Modeling, Regression Analysis, Performance Criteria, Prediction, Managing Auto Insurance Claim

AMS 2010 subject classifications 93A30, 62P05, 00A69, 62M20, 91B30, 97M30

DOI: 10.19139/soic-2310-5070-2655

1. Introduction

The increasing number and severity of car claims force us to find new methods for managing them efficiently. Estimating auto claim costs is a crucial but complex job for insurance companies. A fast, precise prediction is the key to providing insureds with accurate cost estimations. However, current methods exhibit a range of obstacles, not least manual procedures and imprecision when estimating the total cost of claims incurred. Machine Learning (ML) is one component of the solution. From that perspective, this paper aims to explore the use of ML algorithms by actuaries to evaluate risk and forecast auto insurance losses effectively. Numerous articles in the literature have addressed the problem of employing ML models for prediction in the insurance field. Smith et al. [1], for example, test several ML algorithms, such as Neural Networks (NN) and Decision Trees (DT), for identifying whether the insured made the claim or not, and examine the impact of this study on insurance results.

*Correspondence to: Chadia Bekkaye (Email: chadiabek2024@gmail.com). Department of Statistics and Applied Mathematics, FSJES Ain Sebaâ, Hassan II University of Casablanca. BP : 2634, Route des Chaux et Ciments Beausite, Casablanca 20254, Morocco.
Hassan Oukhouya (Email: oukhouya.hassan@ump.ac.ma). Department of Economics, FSJES, University Mohammed First of Oujda. BV Mohammed VI B.P. 724 Oujda 60000, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In life insurance, Devi et al. [2] adopt several ML methods to predict medical insurance costs. Their findings show that the Polynomial Regression achieves an 88% of R-squared score (R^2), closely followed by Random Forest (RF) with a score of 86% before and after scaling. Other actuaries opt for ML algorithms to assess the risk of life insurance pricing. In particular, Azzone et al. [3] propose an RF Algorithm to analyze policyholders' lapses of life insurance contracts. Further, Kaushik et al. [4] compare the Artificial Neural Network model (ANN) with the Linear Regression Model, and then decide on the ANN algorithm to predict health insurance premiums efficiently.

In the non-life insurance, ML methods are employed by various authors to tackle diverse issues within this segment, in particular. For instance, Fatima Manlaikhaf [5] uses the ML models to predict whether a customer can renew their contract or not. Additionally, Zuhermaan et al. [6] propose Support Vector Machine (SVM) to classify policyholders satisfactorily in car insurance. In another study, Huang et al. [7] mainly investigate the adoption of a wide range of driving behavior characteristics to forecast the likelihood of risks and the frequency of claims for cars currently insured with ML techniques; four classifiers were eXtreme Gradient Boosting (XGBoost), RF, SVM and ANN algorithms. The findings indicate that the XGBoost model outperforms the remaining techniques. Hanafy et al. [8] create an ML algorithm that accurately predicts car insurance claims. The analysis of the results suggests that RF performs more efficiently than alternative models. A new model was created by Liu et al. [9], namely Multi-class AdaBoost, which combines DT and adaptive boosting. The results show that the adaptive predictor is more comparable in terms of both prediction ability and interpretability. In another way, Pérez et al. [10] implemented ML techniques in an alternative claims context; it focuses on identifying fraudulent claims in insurance policies by accurately analyzing suspicious auto claims. Pesantez-Narvaez et al. [11] introduce telematic data to predict motor insurance claims. After comparing XGBoost and Logistic Regression (LR), the results show that XGBoost has a generally higher forecast accuracy.

In the regression field, Selvakumar et al. [12] perform various categories of vehicle claim amounts using ML approaches and decide that ANN is a better predictive model in their work. On the same topic, Guelman [13] describes and applies the Gradient Boosting theory to model auto insurance costs. A separate study [14] has introduced the differences of Generalized Linear Modeling (GLM) with ML approaches to predict loss expenses in vehicle insurance. In another context, Ahaggach et al. [15] propose an innovative method that blends odontological reasoning with regression models to increase the accuracy of expense estimates for auto damage repair. Results show that the RF algorithm, particularly coupled with odontological reasoning, outperforms all methods regarding R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). In a comparative conducted by Lozano-Murcia et al. [16], various ML algorithms, including RF, Bagging, ANN, DT, Boosting Trees and Linear Regression, were assessed for their performance, measuring the MAE and R^2 metrics for regression problems. A significant study was conducted by Poufinas et al. [17] to forecast the mean motor insurance costs using ML approaches, as well as SVM, DT, RF and Boosting. The results demonstrate that RF outperform the alternative models, followed by XGBoost algorithms. Although previous studies have contributed significantly to understanding auto claims, many focus narrowly on classification tasks (e.g., fraud detection, insured classification or risk evaluation) rather than direct cost estimation. ML approaches such as XGBoost, SVR, and MLP have gained importance in insurance analysis due to their capacity to handle complex, non-linear models effectively. Traditionally, actuarial models, especially GLMs, remain standard in a variety of fields. GLMs are simple to interpret and rely on well-recognized statistical principles, well suited to regulatory and pricing insurance [18]. However, GLMs presume linear relationships between both predictors and target variable, potentially restricting their predictive accuracy in the face with real, high-dimensional insurance data sets. A comparative study by Wüthrich [19] revealed that tree-based ML models perform persistently well compared to traditional preserving techniques, such as Chain-Ladder and GLMs, for prediction individual claims , notably with complex interactions and nonlinearities. Recent advancements in insurance ML include deep learning methods, Recurrent Neural Networks (RNNs) to multivariate reserving tasks [20], and Long Short-Term Memory (LSTM) for micro-level reserving [21], are used for analyzing time-series claims data. Such models can detect dynamic relationships that static models risk ignoring. Furthermore, federated learning appears as a privacy-protective option that permits collaborative model learning among insurers without shared customer-sensitive data [22]. These developments emphasize the expanding role of ML in insurance while underlining the necessity for balancing between performance, interpretability, and data privacy. In addition, several works lack rigorous interpretability

frameworks or fail to address challenges such as feature selection, hyperparameter tuning or comprehensive model evaluation using both statistical and stochastic metrics. This paper addresses these limits by applying a complete workflow that involves robust preprocessing, modeling with XGBoost, Support Vector Regression (SVR) and Multi-Layer Perceptron (MLP), detailed evaluation metrics including the Tailor Diagram and SHAP-based interpretability for cost predictions in auto insurance. These specific models offer strong regression capabilities for modeling nonlinear relationships and improving generalization performance. Furthermore, the study is designed to improve model transparency and accuracy while promoting more efficient management of auto insurance claims from an economic perspective, which is crucial for real-world adoption by insurance companies. The rest of the paper is arranged as follows: Section 2 presents the empirical data set of car insurance along with the data preprocessing techniques executed. The theoretical framework for modeling ML approaches and the description of performance criteria are also mentioned in this study. Section 3 of this work extensively analyzes and evaluates of the various ML models, including XGBoost, SVR and MLP. The finding and their practical effects are completely clarified. Lastly, Section 4 ends the article with a summary of the outcomes alongside suggestions for additional study.

2. Research Methodology

2.1. Data set

In this study, the collected data set was sourced from Kaggle[†]. There are 1000 rows and 40 columns in total. After dropping irrelevant features and any empty columns, the new data set contains 1000 rows with 13 variables, including numerical and categorical types. The "months as customer" variable is of a numerical type, such as "age", "policy premium", "total claim amount", "injury claim", "property claim", "vehicle claim" and the last "auto year" variable. Furthermore, variables such as "insured sex", "incident type", "fraud reported", "incident severity", and "auto make" represent the categorical type. As shown in Table 1

Table 1. Description of the Data set

Variable Name	Data Type	Description
`total_claim_amount`	Numerical	Total amount claimed
`months_as_customer`	Numerical	Duration of customer membership (months)
`age`	Numerical	Age of the insured individual
`policy_annual_premium`	Numerical	Annual premium for the insurance policy
`insured_sex`	Categorical	Gender of the insured person
`incident_type`	Categorical	Type of reported incident
`incident_severity`	Categorical	Severity level of the incident
`injury_claim`	Numerical	Claim amount for injuries
`property_claim`	Numerical	Claim amount for property damages
`vehicle_claim`	Numerical	Claim amount for vehicle damages
`auto_make`	Categorical	Brand of the insured car
`auto_year`	Numerical	Year of the vehicle model
`fraud_reported`	Categorical	Whether fraud was reported (Y/N)

[†]Data set source, available link: https://www.kaggle.com/datasets

2.2. Preprocessing Data set

Before training the model, we first process the data set to verify that all inputs are correctly scaled and structured. The data set comprises 12 input variables, each corresponding to distinct features that can influence the output. These input variables are described as follows:

$$X = \{X^{(1)}, X^{(2)}, ..., X^{(12)}\}. \tag{1}$$

Each $X^{(i)}$ refers to a unique independent variable in the data set, e.g. insured's details, vehicle characteristics, claims history and other relevant variables. The target variable, which the algorithm was created to forecast, is the total claim amount (y_i), representing the cost of losses incurred for each observation i.

2.2.1. Data set Splitting and Normalization

The data set was separated into 80% training and 20% testing sets for reliable model learning and evaluation. The training data set aims to learn models from the data, while the testing set evaluates the model's ability by applying it to new cases. As the input variables vary in range and unit, we apply Min-Max normalization to standardize them within a fixed range of [0,1]. The aim is to prevent variables with larger values from dominating the model learning process. Normalization is performed using the following formula:

$$X^{(i)}_{\text{normalized}} = \frac{X^{(i)} - X^{(i)}_{\min}}{X^{(i)}_{\max} - X^{(i)}_{\min}}, \tag{2}$$

where, $X^{(i)}_{\text{normalized}}$ is the normalized value of the i-th variable, $X^{(i)}_{\max}$ is the maximum value of the i-th feature in the data set, $X^{(i)}_{\min}$ is the minimum value of the i-th feature in the data set. By using the same range for all feature scaling, Min-Max normalization ensures that no single variable dominates the learning process, thus improving model performance and stability. This preprocessing step is crucial to ensure the model can identify significant patterns while preserving numerical stability.

2.3. Machine Learning Algorithms

Based on historical auto insurance data sets, we have applied XGBoost, SVR and MLP algorithms. We then predicted the total claim amount and compared it with several parameters such as R^2, MSE, RMSE, MAE, along with Normalized Mean Squared Error (NMSE), Difference Score (DS) and Weighted Difference Score (WDS). The following section presents a description of modeling techniques.

2.3.1. XGBoost model

The XGBoost model, initiated by Chen and Guestrin [23], is a high-performance, scalable ML algorithm for gradient optimization adapted to structured data and large-scale ML tasks. It sequentially creates a set of decision trees, where every tree is designed to minimize the errors of the current iteration focusing on the residuals. The model is then used to optimize a regularized objective function that balances the trade-off between model fit and complexity:

$$\text{Objective} = \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(T_k), \tag{3}$$

where, $\mathcal{L}(y_i, \hat{y}_i)$ is the loss function measuring the error between the actual value y_i and the prediction $\hat{y}_i$, $\Omega(T_k)$ is a regularization term to control tree complexity, defined as :

$$\Omega(T_k) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2, \tag{4}$$

where, T is the number of the tree's leaves, w_j is the weight of a leaf, γ and λ are regularization hyperparameters. The principal innovations of XGBoost involve advanced tree pruning, weighted quantile sketching for optimal

split search, and support for parallel computing, providing efficient processing for both missing and large data sets (e.g., by learning the optimal direction of imputation during training). Overall, XGBoost has shown exceptional performance in various fields, notably fraud detection, recommendation systems, and predictive analysis, making it a reference algorithm for data science and ML challenges [24].

2.3.2. SVR model

The SVR model is a regression adaptation of the Support Vector Machine (SVM) approach, invented initially by Vladimir Vapnik [25]. Its principles include linear and non-linear correspondence to predict continuous output variables. The SVR algorithm aims to find a hyperplane in high-dimensional feature space that minimizes prediction errors while balancing model complexity and generalization. A significant characteristic of SVR is the loss function insensitive to ε, which tolerates errors to a margin ε, enabling the model to ignore minor deviations and concentrate on significant models. The objective function is as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*), \tag{5}$$

with constraints :

$$\begin{aligned} &y_i - \langle w, x_i\rangle - b \leq \varepsilon + \xi_i, \\ &\langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \geq 0 \quad \forall i = 1, \ldots, N. \end{aligned}$$

where, w is the vector of weights defining the hyperplane, b is the bias, $C > 0$ is a hyperparameters controlling the compromise between training error and regularization, ξ_i, ξ_i^* are the slack variables for errors outside the ε margin and ε is the sensitivity level, which ignores minor errors. By integrating kernel methods, the SVR algorithm provides predictive analysis performance. It allows nonlinear relationships to be efficiently modeled, thus ensuring its suitability for applications requiring accurate predictions, such as time-series forecasting, modeling of energy demand and financial analysis [26].

2.3.3. MLP model

The MLP model, developed by Frank Rosenblatt [27], is a fundamental algorithm for ANN, conceived explicitly for tasks involving nonlinear correspondences between input features and output targets. The network's neuron executes a weighted sum of inputs, then implements a non-linear activation function. Mathematically, the output of the neuron is as follows:

$$z_j^{(l)} = \sum_{i=1}^{n_{l-1}} w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}, \tag{6}$$

$$a_j^{(l)} = \phi(z_j^{(l)}), \tag{7}$$

where, $z_j^{(l)}$ is the linear output (before activation) of the j-th neuron in layer l, $a_j^{(l)}$ is the active output of the j-th neuron in layer l, $w_{ji}^{(l)}$ is the weight between the i-th neuron in the layer $l-1$ and the j-th neuron in the layer l, $b_j^{(l)}$ is the bias associated with the j-th neuron of layer l, ϕ is the activation function and n_{l-1} is the number of neurons in the previous layer $(l-1)$. MLP training minimizes a loss function, typically expressed as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell\big(y_i, \hat{y}_i\big), \tag{8}$$

where, $\ell(y_i, \hat{y}_i)$ is the loss function, y_i is the target value $\hat{y}_i$ is the prediction and θ includes all model parameters (w and b). The main innovations of this model include the implementation of the backpropagation algorithm that has become an essential part of MLP training, enabling efficient calculation of gradients and weight updates in multi-layer networks [28]. Further, the model provides a practical application for solving complex tasks and reignited

interest in neural networks after a period of decline [29]. However, that decline was largely due to the limitations of single-layer perceptrons, which struggled to solve nonlinear problems.
Although the previous section describes the mathematical formulation for each model, it is also important to highlight their hypotheses and interpretative capabilities. The SVR depends strongly on the selection of kernel function (such as the radial basis function), impacting its potential to handle nonlinear models. In addition, it employs the ε-insensitive loss function that ignores minor prediction errors inside a margin. This renders the SVR robust on noise but significantly less sensitive to extreme values of demands. XGBoost, as a tree-based ensemble model, is sensitive to hyperparameters like learning rate, maximum depth, and number of estimators. These parameters monitor the complexity of the model and directly affect the risk of over-fitting or under-fitting. MLPs offer powerful function approximators but need detailed design choices such as the number of hidden layers, neurons, and activation functions. Their performance is also dependent on weight initialization and optimization methods. A clear picture of these hypotheses helps to adapt each model to the features of the claim prediction problem and to select an appropriate model choice.

2.4. Performance Criteria

R^2 measures the portion of variance in the dependent variable derived from the independent variables. High R^2 values indicate a better correlation (closer to 1). The formula is expressed as [30]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{9}$$

where y_i represents the actual values, $\hat{y}_i$ denotes the predicted values, $\bar{y}$ is the mean of the actual values, and N as the total number of observations in evaluated data set (train or test).

MSE measures the mean of the squared differences between actual and predicted values. Smaller values indicate better performance. The expression is defined as follows [31]:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{10}$$

RMSE represents the square root of the MSE, providing a measure of error in the same units as the data set. The equation is defined as [32]:

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

MAE refers to the mean of the absolute differences between actual and predicted values. The formula is expressed as [33]:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{12}$$

NMSE quantifies the mean squared error relative to the variance of the target variable. The expression is defined as follows [34]:

$$\text{NMSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{13}$$

DS measures the relative difference between predicted and actual values as a fraction of the mean of the actual values. The equation is defined as [35]:

$$\text{DS} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} y_i}. \tag{14}$$

WDS provides a weighted measure of prediction error, where weights refer to the importance or size of the actual target values. The formula is expressed as [36]:

$$\text{WDS} = \frac{\sum_{i=1}^{N} w_i |y_i - \hat{y}_i|}{\sum_{i=1}^{N} w_i y_i}. \tag{15}$$

where the weight w_i represents the actual claim amount for the i^{th} observation, i.e., $w_i = y_i$.

All experiments were performed in Python 3.10 on a Lenovo Ideapad 330 laptop with an Intel Core i5 processor and 8 GB RAM. The analysis employed the scikit-learn library (v1.6.1) for data preprocessing and model evaluation, XGBoost (v2.1.4) for gradient boosting, and TensorFlow (v2.18) for MLP model training. Hyperparameters optimization was performed with Scikit learn's RandomizedSearchCV function. These specifications aim to maintain transparency and facilitate reproducibility of findings.

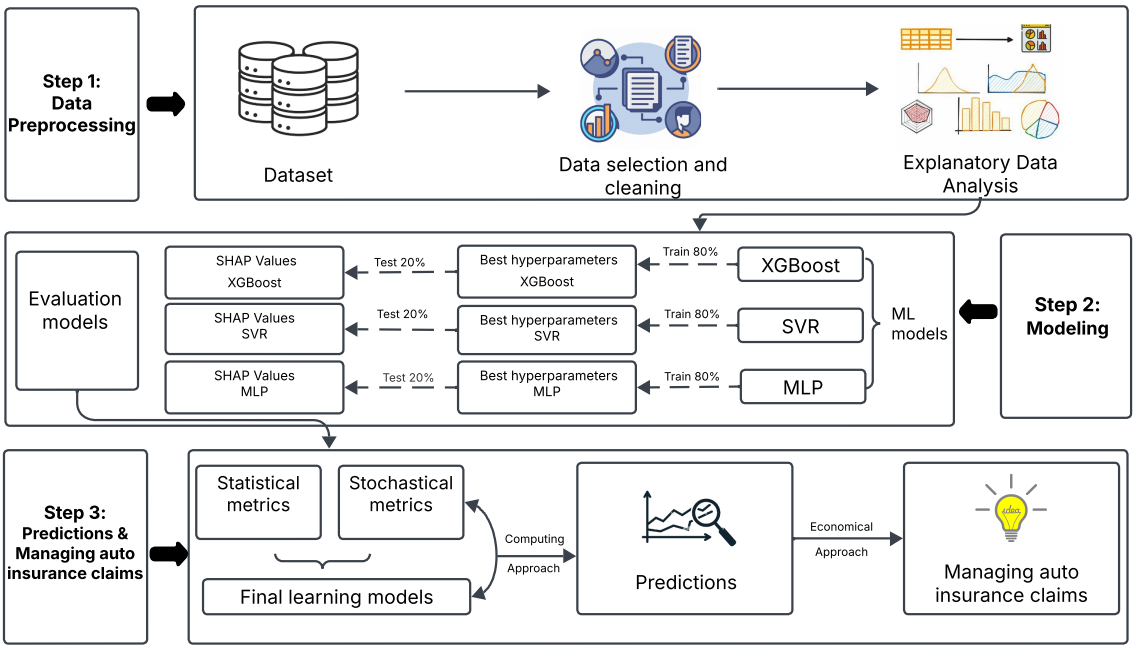


Figure 1. Work process of this research

3. Results and Discussion

The present section discusses the exploratory analysis of the data set for better understanding and interpretation of the results. Then we move on to modeling by implementing the relevant algorithms to derive the best parameters using the Randomized Search CV option. An analysis of residuals will help to reduce risks, and enable accurate prediction of the amount of claims, ensuring that the result is more closely related to actual costs, for efficient management of automobile claims.

The work process contains three key steps, as shown in Figure 1. First, the "Step 1: Data Preprocessing" represents the efficient treatment and analysis of the dataset to ensure high-quality predictions. Followed by "Step 2: Modeling", which consists of implementing ML models by setting hyperparameters and selecting features that enhance model performance. Ending with "Step 3: Predictions & Managing auto insurance claims", where the model is applied to predict new or unseen data. Statistical and stochastic metrics (computing approach) evaluate the model performance. The economic angle suggests pertinent strategies to manage auto claims effectively.

3.1. Explanatory Data Analysis (EDA)

Successful completion of the EDA ensures a complete understanding of the data set, allowing us to conduct a univariate analysis according to the "auto_make" and "fraud_reported" variables, as shown in Figures 2 and 3.

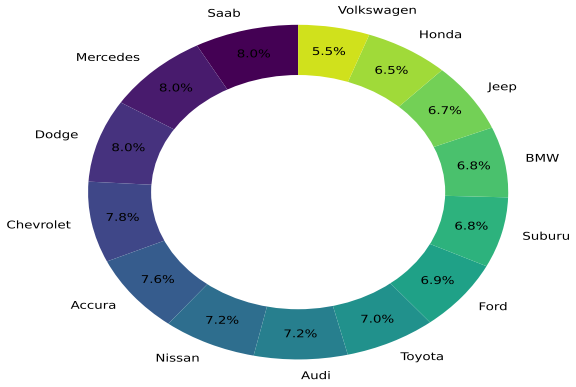


Figure 2. Different Types of Vehicle Models

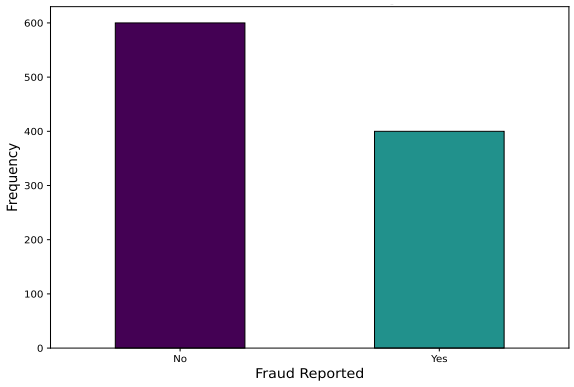


Figure 3. Fraud Distribution Histogram

The Figure 2 shows that each segment has a percentage, which indicates the contribution of that model to the total count of vehicles. For instance, the Saab segment represents 8% of the data set, meaning 8% of the vehicles listed are Saab models. However, the Figure 3 illustrates the chart that visualizes the number of claims reported as fraud or not. The majority of claims in the data set are non-fraudulent, and the fraudulent claims make up a smaller percentage of the total claims. Unlike univariate analysis, studying interactions between variables is relevant to improve data interpretation and support predictive modeling. That is illustrated in Figure 4, which discusses interaction analysis between variables to investigate how total claim amount is affected by two numeric variables (e.g., "age" and "policy annual premium"). Another comparative example will focus on the interaction between total claim amount and vehicle and property claim (see Figure 5), as well as the case of total claim amount compared with two categorical variables, e.g., "incident type" and "fraud reported" in Figure 6.

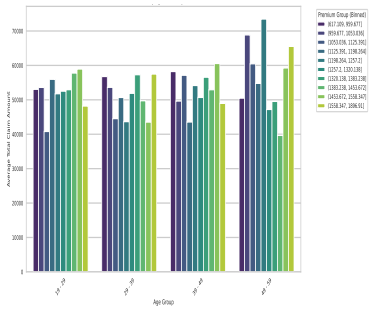


Figure 4. Total Claim Amount by Age and Policy Annual Premium

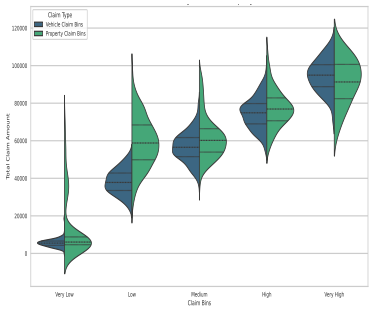


Figure 5. Total Claim Amount by Vehicle and Property Claims

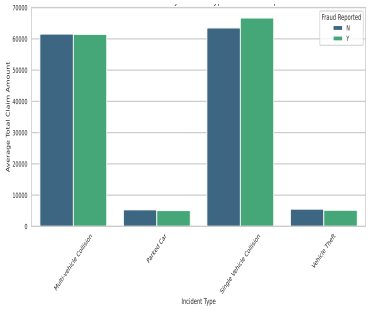


Figure 6. Total Claim Amount by Incident Type and Fraud Reported

The histogram shows that high average claim amounts in all age groups, tend to be associated with high annual premiums. The analysis reflects a correlation of insurance premiums paid with total claim amounts, illustrated in Figure 4. In addition, the split violins illustrate how the vehicle claim and property claim distributions differ. If one side of the split is larger (e.g., vehicle claim), more claims of this type dominate in that bin. A balanced width between the two sides suggests more equal contributions from both claims (see Figure 5). The last graph reveals that the "Single Vehicle Collision" type may have a higher claim amount mean compared to "Vehicle Theft" or "Parked Car", particularly in the "Fraud Reported" category. Certain types of incidents may exhibit similar patterns in the fraudulent and non-fraudulent categories, possibly suggesting that the incident's severity may not significantly impact the occurrence of fraud in that category, illustrated in Figure 6. An additional analysis is presented which compares the total claim amount with "sex" and "incident severity", respectively (as shown in Figure 7).

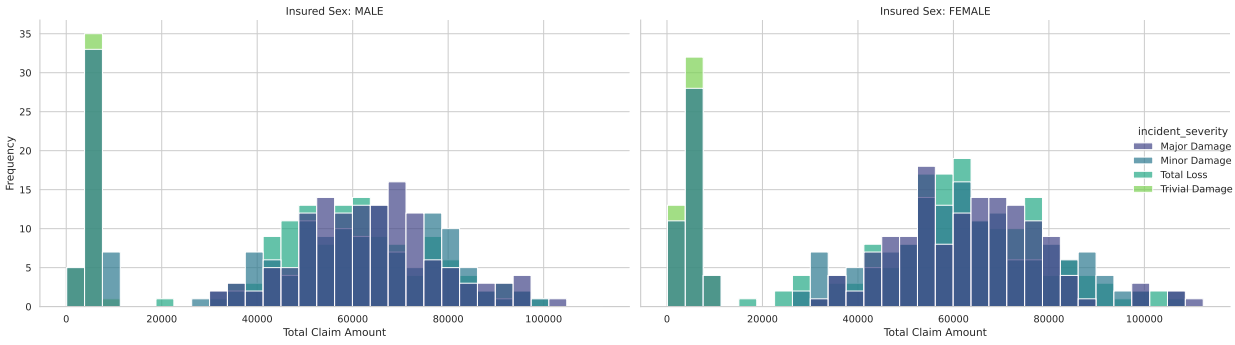


Figure 7. Total Claim Amount by Sex and Incident Severity

This plot offers an overview of the total incurred losses distribution by gender (male and female), representing different severity levels. Minor damage dominates the lower claim amounts (close to zero) for both males and females. Significant damage covers a larger range, with constant frequencies for medium to high claim amounts (see Figure 7). Correlation analysis is a crucial step in the process, used to determine the direction and degree of association between two variables, which ranges from -1 to 1.

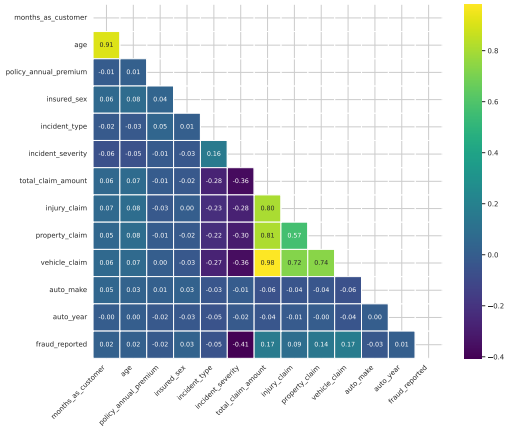


Figure 8. Correlation Plot of Numerical & Categorical Variables

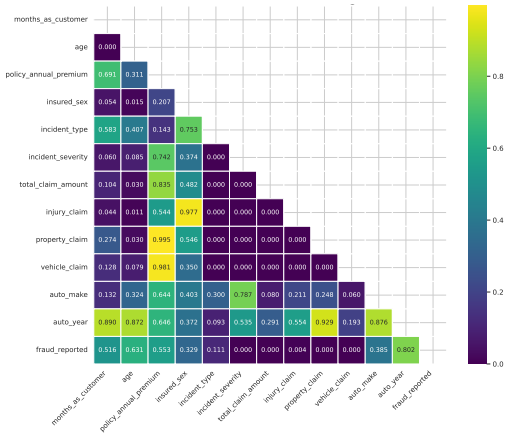


Figure 9. P-value Correlation Plot of Numerical & Categorical Variables

The correlation graph illustrates a strong correlation (close to 1) between the following variables: total claim amount with claims-related variables, age with months as a customer. However, there is a weak negative correlation between incident severity and fraud reported, incident severity with total claim amount and claims-related variables shown in Figure 8. In addition, p-value < 0.05 suggests a substantial connection between the factors in question. Thus, the correlation between total claim amount and claims-related variables (injury, property, vehicle), while higher p-values indicate no significant correlation, e.g. property and injury claims with policy annual premium, insured sex with injury claims, auto year with property claims, ...etc which is illustrated in Figure 9.

3.2. Hyperparameters with modeling

3.2.1. Models Randomized Search Cross Validation results

Randomized Search CV represents a machine learning technique explicitly designed to identify the ideal hyperparameters for a model by randomly selecting combinations of parameters and evaluating them with cross-validation. We evaluated model performance using R^2, MSE, RMSE, and MAE, along with NMSE, DS and WDS to assess multi-step prediction accuracy. The Randomized Search enables the obtaining of the best parameters for the XGBoost, SVR and MLP models, summarized in Table 2.

3.2.2. Hyperparameters Summary

Random searches were conducted to find hyperparameters using a 5-fold blocked cross-validation technique to achieve optimal performance for each model. A 5-fold cross-validation procedure was followed to provide robust model evaluation and hyperparameters tuning. The data set was split into five folds with random mixing to minimize bias and retain reproducibility. Table 2 shows the hyperparameters adjusted for the algorithms in this paper. The hyperparameters adjusting on the models in the current research, where subsample is the data fraction per boosting round, reg-alpha is L1 regularization term for feature selection, reg-lambda is L2 regularization term to reduce overfitting, n-estimators is the number of trees, max-depth is the tree depth limit, learning-rate is the model update step size, colsample-bytree is the features per tree, kernel is the function type, gamma is influence range of points, epsilon is the margin of tolerance, c is the regularization strength, solver is the optimization algorithm, max-iter is the training iterations, hidden-layer-sizes is the hidden layers setup, alpha is the L2 regularization, and activation is the neuron activation function.

3.2.3. SHAP Values Summary

SHAP values analyzes the effect of features on model prediction. This shows the influence of each input on the output, thus guaranteeing a correct and coherent understanding of the model's individual and global behavior. In

Table 2. The hyperparameters adjusting on the models in this study.

Model	Parameters	Range	Best Parameters
XGBoost	Subsample	[0.6, 0.8, 1.0]	1.0
	reg-lambda	[1.0, 1.5, 2.0]	1.5
	reg-alpha	[0, 0.1, 0.5, 1.0]	1.0
	n-estimators	[100, 200, 300, 400]	400
	max-depth	[3, 5, 7, 10]	5
	learning-rate	[0.01, 0.05, 0.1, 0.2]	0.05
	colsample-bytree	[0.6, 0.8, 1.0]	0.6
SVR	Kernel	['rbf', 'poly', 'sigmoid']	rbf
	gamma	['scale', 'auto', 0.01, 0.1, 1]	auto
	epsilon	[0.01, 0.1, 0.2, 0.5]	0.01
	C	[0.1, 1, 10, 100]	10
MLP	Solver	['adam', 'sgd']	adam
	max-iter	['adaptive', 'constant']	adaptive
	hidden-layer-sizes	[(50,), (100,), (50,50), (100,50)]	(100,50)
	alpha	[0.0001, 0.001, 0.01, 0.1]	0.1
	activation	['relu', 'tanh', 'logistic']	tanh

Figure 10, we illustrate the influence of each factor on the output by XGBoost model, as well as the effect on the SVR model output (as shown in Figure 11) and the influence on the MLP model outcome (see Figure 12).

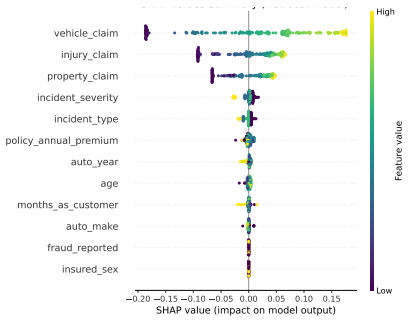


Figure 10. SHAP Values Summary for XGBoost model

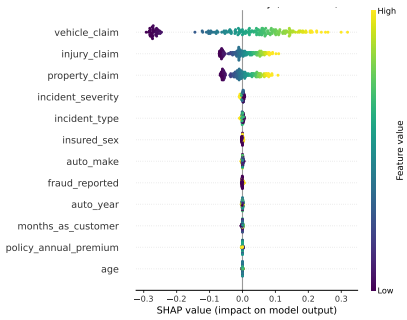


Figure 11. SHAP Values Summary for SVR model

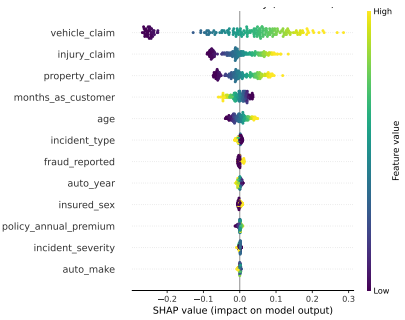
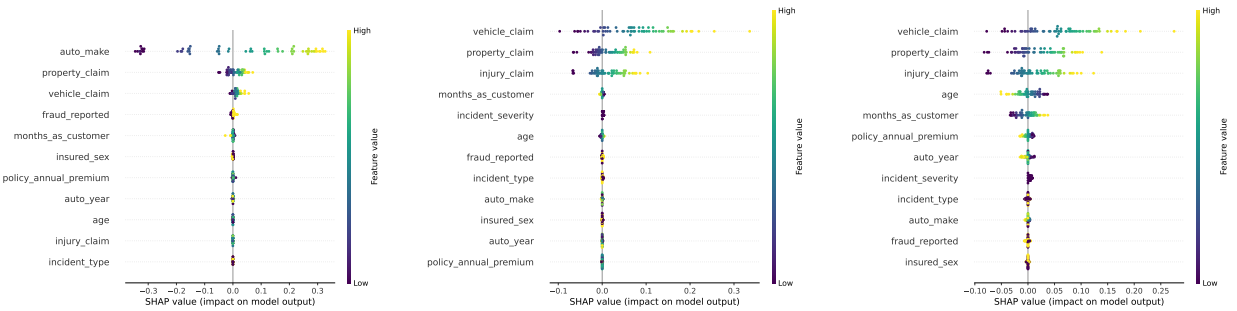


Figure 12. SHAP Values Summary for MLP model

Since the target variable is the total claim amount, the SHAP summary graph offers a clear view of the impact of each characteristic on the predicted claim amount. After analyzing the three models, we conclude that the variables "vehicle claim", "injury claim" and "property claim" have the most decisive influence on the prediction of the total claim amount; e.g. as expected, higher claims for each category increase the total claim amount (yellow dots on the right). This is followed by severity/incident type, which also plays a role in XGBoost (see Figure 10) and SVR as illustrated in Figure 11, but less dominantly. By contrast, the insured's demographics and policy details have minimal impact. XGBoost and SVR exhibit a slightly larger importance of variables. At the same time, MLP offers a slightly broader range of SHAP values and an important range of variables, i.e. months as a customer and age (as shown in Figure 12). Overall, claim-related characteristics consistently dominate model

predictions, reinforcing their influence on identifying total claim amounts. To improve the interpretability of the model's prediction of total_claim_amount, we carried out a SHAP analysis stratified by incident_severity variable. This subgroup approach allows the feature contributions across each severity level of incident, uncovering how input variables impact predictions variously across severity contexts. This analysis quantifies the significance of individual variables by severity level, in line with the objective to interpret total_claim_amount with finer granularity.



(a) XGBoost (b) SVR (c) MLP

Figure 13. SHAP Value Summary Plots for the `Major Damage` Subgroup Across Different Models.

The corresponding SHAP graphs for the `Major Damage` subgroup across three regression models (XGBoost, SVR, and MLP) are illustrated in Figure 13. In all models, the more influential variables remain relatively consistent, with vehicle_claim, property_claim, and injury_claim dominating the predictive variance explanation for total_claim_amount. The XGBoost model also emphasizes the significance of the auto_make variable, whereas the MLP model shows a more varied range of influential features, notably age and months_as_customer.
For brevity, detailed analysis of the SHAP values representing the other severity levels such as `Minor Damage`, `Total Loss`, and `Trivial Damage`, are provided in the Appendix (see Figure 19, Figure 20, and Figure 21, respectively). These visualizations complete the analysis of the subgroup of `Major Damage` and enable a comprehensive comparison of the feature contributions across various levels of incident_severity.

3.3. Residual Analysis

3.3.1. Performance comparison of XGBoost, SVR and MLP

This section presents the performance evaluation of the XGBoost, SVR and MLP models using various performance metrics, including R^2 score, MSE, RMSE, MAE, NMSE, DS and WDS. The findings are available in Table 3.

Table 3. Comparison of performance metrics for XGBoost, SVR, and MLP models on both training and test sets.

Models	XGBoost		SVR		MLP	
	Train	Test	Train	Test	Train	Test
R^2 Score	0.9999	0.9918	0.9960	0.9950	0.9878	0.9864
MSE	0.0000	0.0004	0.00021	0.00026	0.0006	0.0007
RMSE	0.0011	0.0207	0.0147	0.0162	0.0259	0.0268
MAE	0.0008	0.0100	0.0055	0.0060	0.0183	0.0187
NMSE	0.0000	0.0081	0.0039	0.0049	0.0121	0.0135
DS	0.0017	0.0222	0.0117	0.0135	0.0388	0.0416
WDS	0.0008	0.0126	0.0060	0.0067	0.0196	0.0203

Referring to Table 3, the performance criteria suggest that SVR outperforms the other models on test sets, with lower errors (MSE, RMSE, MAE) and the highest R^2 score. SVR shows consistent performance in the train and test sets, indicating good generalization. Followed by XGBoost, which reflects a minor overfitting, resulting in slight increases in test errors. MLP has the highest errors and lowest R-scores, making it the weakest model. Regarding NMSE, DS and WDS, SVR remains stable in both phases. In contrast, XGBoost achieves the smallest values for training but increases values for testing, thus suggesting some overfitting. MLP exhibits the highest dispersion, indicating instability. For better graphical visualization of the results, we illustrate the evaluation metrics in Figure 14.

3.3.2. Taylor Diagram

The Taylor diagram visually illustrates the performance of three models (XGBoost, SVR and MLP) based on their correlation and standard deviation. The Figure 15 shows the comparison relative to the reference point (red dot).

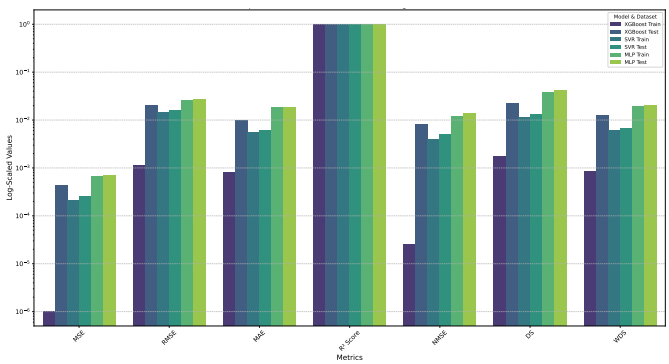


Figure 14. Histogram for comparison between performance metrics on train and test sets

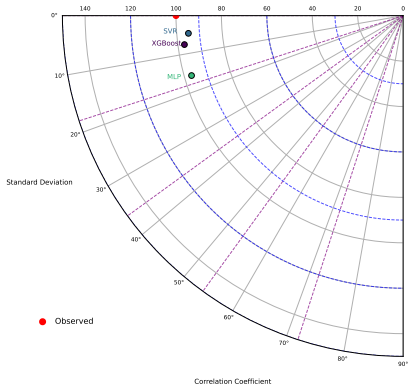


Figure 15. Taylor Diagram

Since the red dot represents the reference point (perfect model), the model closest to this point becomes the best choice. Thus, we notice that SVR has a good combination of high correlation and low standard deviation, reflecting that it performs well against the evaluation criteria (as illustrated in Figure 15).
While the Taylor Diagram validates SVR's strong predictive ability_surpassing both XGBoost and MLP in the

majority of numerical metrics_ its black-box nature introduces a limit in terms of interpretability. In light of the transparency in insurance, especially in the General Data Protection Regulation (GDPR), further work is likely to focus on integrating explicitness techniques for SVR, like rule extraction or interpretive substitution models, to guarantee both regulatory conformance and confidence in the model's decisions.

3.4. Prediction and managing auto insurance claims

3.4.1. Model's predictions

This section represents the model prediction, described as the output after input data processing. Each model must estimate the predicted values and compare them with the actual ones, as shown in Figure 16, which represents the predictions retained by the XGBoost model, Figure 17 shows the predictions of the SVR model and Figure 18 those of the MLP algorithm, compared to current values.

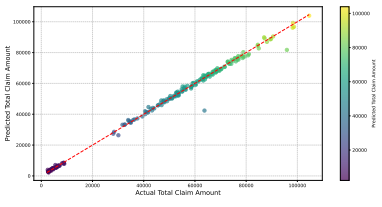


Figure 16. XGBoost: Actual vs Predicted Claims

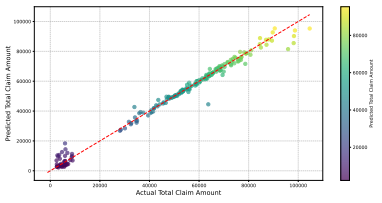


Figure 17. SVR: Actual vs Predicted Claims

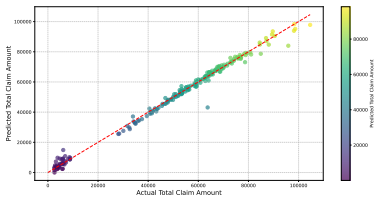


Figure 18. MLP : Actual vs Predicted Claims

The scatter plot shows the variation between the total amounts of actual and predicted claims. The points closely follow the red line (y=x), reflecting a strong correspondence between the actual and predicted values. The colored bar denotes the estimated claim cost, with colors varying from purple (minimum values) to yellow (maximum values). We note a few outliers, but the XGBoost model generally performs well as shown in Figure 16. On the last scatter plot (see Figure 18), the MLP model gives predictions aligned with the red diagonal (following the actual values) with minor outliers as well. The SVR model demonstrated strong overall performance but struggled to predict extreme claim amounts, especially in the tails of the distribution, as illustrated in Figure 17 . This is attributable to the model's sensitivity towards the kernel function and ε-insensitive loss, which can restrict its ability to treat efficiently large deviations.

3.4.2. Managing auto insurance claims

From an economic point of view, auto claims management usually aims to ensure a balance between cost-efficiency and equitable compensation. Insurance companies opt for data-driven modeling to assess claims, predict future risks, and ensure they are handled quickly and accurately. Key tasks include analyzing damage, specifying responsibility and estimating repair costs. The algorithms will help insurance companies to detect fraud and optimize the amounts paid out, thus minimizing unnecessary expenses. In addition, they study risk profiles to adjust premiums, ensuring adequate coverage for the insured and healthy finances for the insurer. Efficiency in claims management reflects reduced operational costs and the adoption of durable pricing strategies. Subrogation procedures, operated by insurers to recover costs from third parties responsible for the accident, also reduce overall expenditure. Insurance companies that balance between operational efficiency, accurate claim payments and cost control strategies can maintain financial stability while delivering value to their clients. This approach is consistent with the sector's evolution towards data-driven management to ensure competitiveness in dynamic markets [37, 38].

4. Conclusion

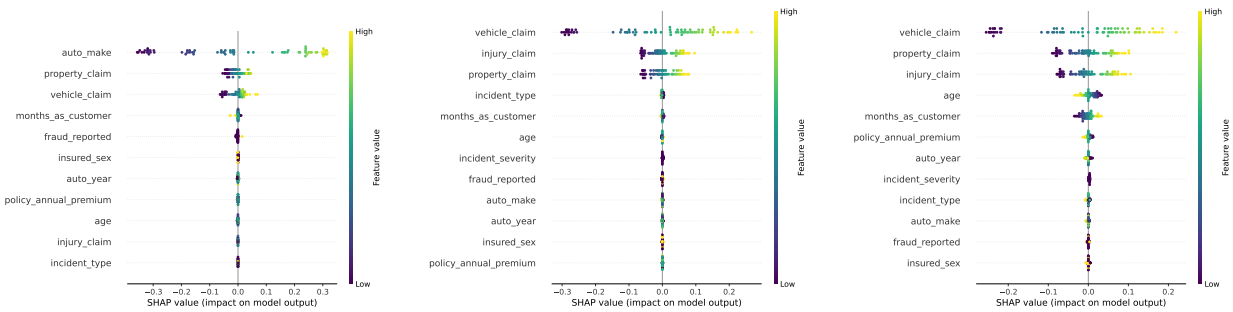
In this research, we evaluate the efficiency of various ML models for predicting total claim amounts in auto insurance. Three models, namely XGBoost, SVR, and MLP, have been selected for their ability to handle complex data and provide accurate predictions of claim payouts. Based on multiple evaluation metrics, the SVR model demonstrated excellent predictive accuracy among the models evaluated. SVR is the best-performing model on the test set, with the highest R^2 (0.9950) and lowest error values (MSE, RMSE, MAE, NMSE, DS, WDS), suggesting better generalization. The XGBoost performs well in the training set, but shows signs of overfitting, with a slight increase in test errors. In contrast, the MLP model has the lowest performance, with the highest errors and the lowest R^2, reflecting instability and weak generalization. Regarding scatter plot visualization, we note the existence of a few outliers in SVR forecasts, which reflects that the model appears inadequate at extremes. XGBoost, on the other hand, produces a better visual distribution of predictions, despite achieving slightly worse metrics than SVR, reflecting its overall accuracy. These results demonstrate the trade-offs between evaluation based on performance metrics and visual coherence. They also emphasize the importance of carefully choosing models depending on the specific objectives of loss prediction. For efficient auto claim management, insurers can also opt for hybrid approaches that maintain predictive accuracy and stability through techniques for detecting and correcting outliers. In economic terms, they can also employ models to optimize reserve allocations and reduce unnecessary payments. In addition, introducing telematic policies for safe driving and using automation to simplify claims reduces operational costs and improves efficiency.

Since the models selected in this study are relatively complex, achieving predictions is challenging to grasp. To improve interpretability and better understand the decision-making process of these complex models, techniques such as SHAP are employed to explain feature contributions and individual predictions. Additionally, missing or excluded features may influence the model's efficiency. Although the SVR model presented a high performance in most evaluation metrics, a detailed analysis of model predictions reveals limits in terms of precisely estimating small and large loss components. This issue appears evident when visualizing predicted versus actual values, exactly as the model seems to underfit the extremes. As a future work, hybrid modeling approaches (e.g., SVR coupled with quantile regression), claim severity segmentation, or the use of other alternative loss functions, such as quantile loss, could investigated to boost predictive accuracy for extreme loss values and improve model generalization in asymmetrical data conditions. To enhance the hybrid solution, we also plan to introduce methodological principles that address potential biases in the data set and reinforce model robustness. In particular, strategies include bias-variance analysis, resampling methods (e.g., SMOTE) to mitigate imbalance, and synthetic noise injection to assess robustness. Additionally, the inclusion of domain-specific characteristics transformation (e.g., log-transformed premium or loss severity ratios) may further boost generalization and sensitivity to infrequent high-loss claims, as well as addressing interpretability challenges discussed in Section 3.3.2.

Appendix A: SHAP Analysis by incident_severity Type

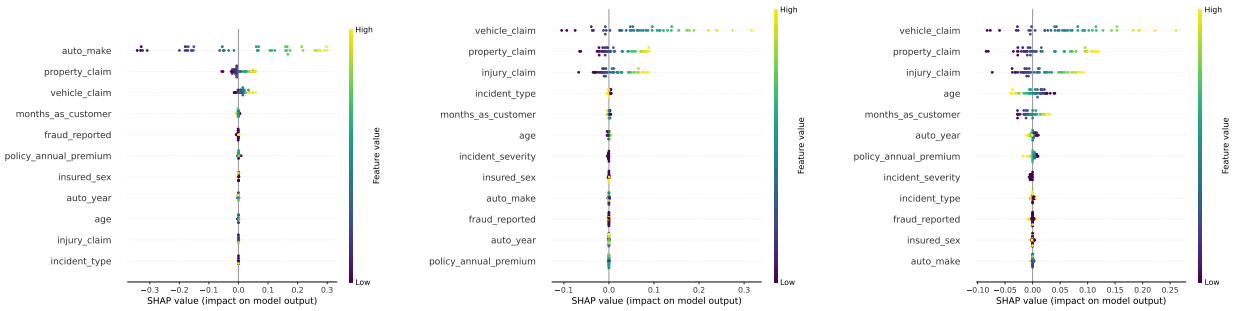
A.1. Case: `Minor Damage`

A.2. Case: `Total Loss`



(a) XGBoost (b) SVR (c) MLP

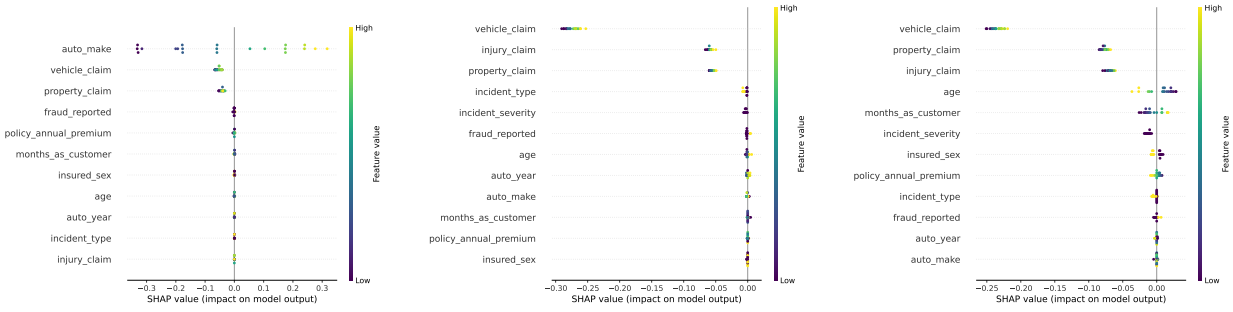
Figure 19. SHAP Value Summary Plots for the `Minor Damage` Subgroup Across Different Models.



(a) XGBoost (b) SVR (c) MLP

Figure 20. SHAP Value Summary Plots for the `Total Loss` Subgroup Across Different Models.

A.3. Case: ***`Trivial Damage`***



(a) XGBoost (b) SVR (c) MLP

Figure 21. SHAP Value Summary Plots for the `Trivial Damage` Subgroup Across Different Models.

Declarations

Funding

This research received no external funding.

Availability of data and materials

The data set used and/or analyzed during the current research is available on the CSE website: https://www.kaggle.com/datasets

Competing interests

The authors have no conflicts of interest or personal relationships that could have biased the research.

Acknowledgments

We would like to express our sincere gratitude to Professor David G. Yu, Coordinating Editor of Statistics, Optimization & Information Computing, and the anonymous reviewers for their valuable comments and constructive feedback, which have greatly improved the quality of this article.

REFERENCES

1. K. A. Smith, R. J. Willis, and M. Brooks, "An analysis of customer retention and insurance claim patterns using data mining: A case study," *Journal of the operational research society*, vol. 51, no. 5, pp. 532–541, 2000.
2. M. Shyamala Devi, P. Swathi, M. Purushotham Reddy, V. Deepak Varma, A. Praveen Kumar Reddy, S. Vivekanandan, and P. Moorthy, "Linear and ensembling regression based health cost insurance prediction using machine learning," in *Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2*, pp. 495–503, Springer, 2021.
3. M. Azzone, E. Barucci, G. G. Moncayo, and D. Marazzina, "A machine learning model for lapse prediction in life insurance contracts," *Expert Systems with Applications*, vol. 191, p. 116261, 2022.
4. K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine learning-based regression framework to predict health insurance premiums," *International journal of environmental research and public health*, vol. 19, no. 13, p. 7898, 2022.
5. F. Manlaikhaf, "Prediction of the insurance contract renewal for vehicle," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 17, 2023.
6. Z. Rustam and N. P. A. A. Ariantari, "Support vector machines for classifying policyholders satisfactorily in automobile insurance," in *Journal of Physics: Conference Series*, vol. 1028, p. 012005, IOP Publishing, 2018.
7. Y. Huang and S. Meng, "Automobile insurance classification ratemaking based on telematics driving data," *Decision Support Systems*, vol. 127, p. 113156, 2019.
8. M. Hanafy and R. Ming, "Machine learning approaches for auto insurance big data," *Risks*, vol. 9, no. 2, p. 42, 2021.
9. Y. Liu, B.-J. Wang, and S.-G. Lv, "Using multi-class adaboost tree for prediction frequency of auto insurance," *Journal of Applied Finance and Banking*, vol. 4, no. 5, p. 45, 2014.
10. J. M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J. I. Martín, "Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance," in *Pattern Recognition and Data Mining: Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, 2005, Proceedings, Part I 3*, pp. 381–389, Springer, 2005.
11. J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—xgboost versus logistic regression," *Risks*, vol. 7, no. 2, p. 70, 2019.
12. V. Selvakumar, D. K. Satpathi, P. P. Kumar, and V. V. Haragopal, "Predictive modeling of insurance claims using machine learning approach for different types of motor vehicles," *Accounting and finance*, vol. 9, no. 1, pp. 1–14, 2021.
13. L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3659–3667, 2012.
14. A. A. Wilson, A. Nehme, A. Dhyani, and K. Mahbub, "A comparison of generalised linear modelling with machine learning approaches for predicting loss cost in motor insurance," *Risks*, vol. 12, no. 4, p. 62, 2024.
15. H. Ahaggach, L. Abrouk, and E. Lebon, "Enhancing car damage repair cost prediction: Integrating ontology reasoning with regression models," *Intelligent Systems with Applications*, p. 200411, 2024.
16. C. Lozano-Murcia, F. P. Romero, J. Serrano-Guerrero, and J. A. Olivas, "A comparison between explainable machine learning methods for classification and regression problems in the actuarial context," *Mathematics*, vol. 11, no. 14, p. 3088, 2023.
17. T. Poufinas, P. Gogas, T. Papadimitriou, and E. Zaganidis, "Machine learning in forecasting motor insurance claims," *Risks*, vol. 11, no. 9, p. 164, 2023.
18. E. Ohlsson and B. Johansson, *Non-life insurance pricing with generalized linear models*, vol. 174. Springer, 2010.
19. M. V. Wüthrich, "Machine learning in individual claims reserving," *Scandinavian Actuarial Journal*, vol. 2018, no. 6, pp. 465–480, 2018.

20. P. Cai, A. Abdallah, and P. Jeganathan, "Recurrent Neural Networks for Multivariate Loss Reserving and Risk Capital Analysis," *arXiv preprint arXiv:2402.10421*, 2024.
21. I. Chaoubi, C. Besse, H. Cossette, and M.-P. Côté, "Micro-level reserving for general insurance claims using a long short-term memory network," *Applied Stochastic Models in Business and Industry*, vol. 39, no. 3, pp. 382–407, 2023.
22. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
23. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
24. H. Oukhouya and K. El Himdi, "Machine Learning Models-Based Forecasting Moroccan Stock Market," in *International Conference on Logistics Operations Management*, pp. 56–66, Springer, 2024.
25. V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
26. H. Oukhouya and K. E. Himdi, "A comparative study of ARIMA, SVMs, and LSTM models in forecasting the moroccan stock market," *International Journal of Simulation and Process Modelling*, vol. 20, no. 2, pp. 125–143, 2023.
27. F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.
28. P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavior science," *PhD thesis, Harvard University*, 1974.
29. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
30. N. Draper, *Applied regression analysis*. McGraw-Hill. Inc, 1998.
31. H. Theil, G. Beerens, C. Tilanus, and C. B. De Leeuw, *Applied economic forecasting*, vol. 4. North-Holland Publishing Company Amsterdam, 1966.
32. T. Chai and R. R. Draxler, "Root mean square error (RMSE) or Mean Absolute Error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.
33. M. A. Rawlins, C. Willmott, A. Shiklomanov, E. Linder, S. Frolking, R. B. Lammers, and C. Vörösmarty, "Evaluation of trends in derived snowfall and rainfall across eurasia and linkages with discharge to the arctic ocean," *Geophysical Research Letters*, vol. 33, no. 7, 2006.
34. D. R. Legates and G. J. McCabe Jr, "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation," *Water resources research*, vol. 35, no. 1, pp. 233–241, 1999.
35. C. J. Willmott, "On the validation of models," *Physical geography*, vol. 2, no. 2, pp. 184–194, 1981.
36. C. J. Willmott, S. M. Robeson, and K. Matsuura, "A refined index of model performance," *International Journal of climatology*, vol. 32, no. 13, pp. 2088–2094, 2012.
37. J. D. Cummins and M. A. Weiss, "Systemic risk and the US insurance sector," 2014.
38. H. Oukhouya, A. El Rhiouane, R. Guerbaz, T. Zari, and K. El Himdi, "Balancing Returns and Responsibility: Markowitz Optimization for ESG-Integrated Portfolios in Morocco," *Computational Economics*, pp. 1–26, 2025.