Received 29-06-2025; Accepted 22-08-2025; Published 29-08-2025

Interpretable Property Insurance Premium Prediction Using Machine Learning Models

Millicent Auma Omondi momondi@aimsammi.org
African Institute for Mathematical Sciences (AIMS)
Senegal

Josue Nguinabe jnguinabe@aimsammi.org
African Institute for Mathematical Sciences (AIMS)
Senegal

John Kamwele Mutinda jkmutinda@aimsammi.org
University of Science and Technology of China (USTC)
China

Amos Kipkorir Langat moskiplangant@gmail.com
Jomo Kenyatta University of Agriculture and Technology
Kenya

Leonard Sanya lsanya@aimsammi.org
Dedan Kimathi University of Technology
Kenya

Ouraga Aime Cervert Ballou oballou@aimsammi.org
African Institute for Mathematical Sciences (AIMS)
Senegal

Jeremy Nlandu Mabiala jeremy@aimsammi.org
African Institute for Mathematical Sciences (AIMS)
Senegal

Corresponding Author: Millicent Auma Omondi

Abstract

Accurately predicting home insurance premiums is a critical challenge for insurers, as traditional methods struggle with the complexity and volume of modern data. This study leverages machine learning to address this problem, applying a range of supervised models—including linear regression, lasso regression, ridge regression, decision trees, random forests, gradient boosting, and extreme gradient boosting (XGBoost)—to a comprehensive home insurance dataset with 66 features. After preprocessing to retain 50 key variables, the models were trained and evaluated, with random forests, gradient boosting, and XGBoost emerging as top performers, achieving R-squared scores of 0.8137, 0.8014, and 0.8344, respectively. Hyperparameter tuning further improved XGBoost's performance to an R-squared of 0.8380,

Citation: Millicent Auma Omondi et al. Interpretable Property Insurance Premium Prediction Using Machine Learning Models. Journal of Artificial Intelligence and Autonomous Intelligence. Research 2025;2(2):21.

making it the preferred model. Using Shapley Additive Explanations (SHAP), we identified the top 40 influential features, such as building age, number of bedrooms, property type, and recent claim history, boosting the R-squared to 0.8799 (training) and 0.8383 (testing). These findings highlight XGBoost's potential to deliver precise premium predictions. By adopting this approach, insurers can enhance pricing accuracy, improve transparency through SHAP-driven insights, and inform fairer policy decisions.

Keywords: Home Insurance Premium, Machine Learning, Shapley Additive Explanations, XGBoost.

1. INTRODUCTION

The insurance industry is undergoing a transformative shift driven by advancements in data analytics and machine learning, enabling companies to extract actionable insights from vast datasets to make informed, data-driven business decisions [1]. By leveraging sophisticated machine learning techniques, insurers can optimize operational costs, enhance customer experiences, and improve risk management strategies. These techniques have been successfully applied across various insurance domains, including fraud detection, customer churn prevention, customer segmentation, and premium prediction [2]. In the context of home insurance, machine learning offers significant potential to refine premium pricing by analyzing diverse customer and property characteristics, ensuring that premiums reflect the underlying risks while maintaining competitiveness in the market. This study focuses on applying machine learning models to predict home insurance premiums, harnessing customer and property data to develop accurate and interpretable predictive models. In insurance, a premium represents the price paid by policyholders to secure protection against specified risks. The process of premium calculation involves assessing the probability distribution of risks to determine a price that safeguards the insurer's financial viability [3]. For home insurance, policies typically cover a range of perils, including fire, water damage, theft, property damage, and natural disasters such as earthquakes [4]. Accurate premium pricing is critical, as it ensures that insurers can cover potential losses while maintaining operational sustainability. However, the complexity of risk assessment in home insurance is compounded by the diverse factors influencing risk, such as property location, construction materials, and policyholder demographics. Traditional actuarial methods, while effective, often struggle to capture the nuanced interplay of these factors, making machine learning an attractive solution for improving predictive accuracy and efficiency. One of the most pressing challenges for insurance companies is striking a balance between acquiring and retaining customers and ensuring financial stability. Setting premiums too low can lead to insufficient funds to cover claims, risking long-term insolvency. A notable example is United Property and Casualty Insurance Co. (UPC), which faced insolvency due to its inability to cover substantial losses from widespread windstorm damage claims, ultimately forcing the company to cease operations and liquidate assets to settle outstanding obligations [5]. This case underscores the critical need for precise risk assessment and premium pricing to avoid financial distress. Conversely, setting premiums too high can deter potential customers, reducing market share and competitiveness [6]. Machine learning models address this challenge by enabling insurers to estimate premiums that align with the specific risk profiles of properties and policyholders, thereby optimizing financial stability while appealing to a broad customer base. Moreover, these models can adapt to emerging trends, such as increasing climate-related risks or shifts in property market dynamics, ensuring that insurers remain resilient in a rapidly changing environment. The growing adoption of machine learning in insurance also reflects broader industry trends toward personalization and efficiency. By analyzing historical data and real-time inputs, machine learning models can uncover patterns and correlations that traditional methods might overlook, leading to more accurate premium predictions and better risk management. For instance, predictive models can incorporate variables such as regional weather patterns, crime rates, or property age to refine premium estimates, offering a more granular understanding of risk. Additionally, the interpretability of these models is crucial for building trust with regulators and policyholders, as it ensures transparency in how premiums are determined. This study aims to apply machine learning models to predict home insurance premiums, evaluate their predictive performance to identify the most effective model, and use SHAP (SHapley Additive exPlanations) to interpret the influence of key features on premium pricing. By doing so, it seeks to contribute to the growing body of research on data-driven insurance solutions and provide practical insights for insurers aiming to balance profitability with customer satisfaction. The contributions of this work are as follows:

- Development of machine learning models to accurately predict home insurance premiums based on customer and property characteristics.
- Identification of the most effective machine learning model for premium prediction through comparative performance evaluation.
- Analysis of influential features impacting home insurance premiums using SHAP for enhanced model interpretability.

This work is structured as follows. Section 2 reviews past literature, Section 3 discusses the methodology applied, and Section 4 presents the results and findings. The conclusion and recommendations are provided in Section 5.

2. RELATED WORK

The insurance industry has increasingly adopted machine learning (ML) to tackle complex challenges in premium prediction, risk classification, and claims management, driven by the demand for data-driven decision-making in a competitive and regulated environment. As a core component of cognitive computing, ML enables insurers to analyze large datasets, identify patterns, and generate accurate predictions that traditional actuarial methods often fail to deliver. By leveraging ML algorithms, insurance companies can optimize premium pricing, improve risk assessment, and streamline operations, ultimately enhancing customer satisfaction and financial stability. This literature review synthesizes prior research on ML applications in insurance, focusing on premium cost prediction, model performance, and interpretability. While much of the existing work targets health and motor insurance, this study addresses the underexplored domain of home insurance premium prediction, drawing on insights from related fields to inform its approach. Significant research has demonstrated ML's effectiveness in predicting insurance premiums, particularly in health insurance. For example, [7] explored health insurance cost prediction using regression-based models, including linear regression, multilinear regression, and polynomial regression. The findings indicated that polynomial regression with a degree of 3 achieved the highest performance, with an R^2 score of 0.80 and an accuracy of 80.97%. The study emphasized the limitations of traditional methods, noting that the rapid growth of data makes manual calculations inefficient and time-consuming, advocating for ML adoption to handle large datasets and enhance prediction accuracy through relevant feature inclusion. Similarly, [8] investigated a variety of ML models, including artificial neural networks (ANN), gradient boosting (GB), k-nearest neighbours (KNN), support vector regression (SVR), decision trees (DT), random forests (RF), and linear regression, to predict health insurance costs. The results showed that gradient boosting achieved the highest accuracy at 92%, highlighting its potential for rapid policy-making and cost estimation. Another study, [9], applied an ANN-based regression model, achieving an accuracy of 92.72% in health insurance premium prediction, reinforcing the suitability of neural networks for complex datasets. ML models have also been applied to other insurance domains, such as motor and property insurance, with an emphasis on risk classification and claims prediction. For instance, [10] employed decision trees, random forests, and XGBoost to assess insured risk profiles in real estate insurance. The results indicated that XGBoost outperformed other classifiers, establishing it as a robust model for risk classification and premium prediction. Similarly, [11] compared SVR, random forests, and XGBoost for predicting insurance claim values. The findings revealed that SVR performed poorly, particularly for high-value claims, while XGBoost demonstrated superior performance with low error rates. To improve interpretability, feature importance plots were used to identify key predictors influencing claim values. These findings align with [12], which applied regression-based ensemble models, including XGBoost, gradient boosting, and random forests, to predict medical insurance costs. By incorporating SHAP (SHapley Additive exPlanations) and Individual Conditional Expectation (ICE) plots, the study enhanced model interpretability, identifying the most influential features affecting premium prices. The comparative analysis showed that XGBoost achieved an R^2 score of 0.864, though it required substantial computational resources, while random forests were more computationally efficient, and gradient boosting excelled in large-scale predictions. The integration of big data with ML has significantly expanded the capabilities of predictive modeling in insurance. [13] explored big data analytics in insurance, noting that the volume and complexity of modern datasets surpass the capabilities of traditional decision-making systems. The study developed four ML classifiers—Adaboost, Naive Bayes, k-nearest neighbors, and decision trees—to analyze insurance claim data. Adaboost achieved the highest performance, with a precision of 64.9%, accuracy of 66.2%, and F-score of 65.3%, while Naive Bayes performed poorly, with an accuracy of 59%, a recall of 59%, a precision of 60.2%, and F-score of 57.7%. The work underscored ML's potential to streamline claims handling and enhance risk assessments in high-volume data environments. Similarly, [14] developed a real-time health insurance cost prediction model using regression techniques, including simple linear, multiple linear, ridge, lasso, and polynomial regression. Polynomial regression produced the best results, demonstrating its effectiveness in enabling insurers to determine premiums quickly and manage healthcare expenses efficiently. Tree-based ML models have become prominent in insurance pricing due to their ability to capture complex data interactions. [15] applied gradient boosting models to insurance pricing, highlighting their superiority over traditional generalized linear models (GLMs). The study demonstrated that gradient boosting could generate tariffs that accurately reflect underlying risks, reducing adverse selection and ensuring affordable premiums for policyholders. In a related study, [16] explored tree-based methods, including regression trees, random forests, and gradient boosting, noting that decision trees serve as the foundation for these algorithms by grouping policyholders with similar risk profiles. These models depend on selecting appropriate loss functions tailored to the insurance data, enabling precise risk segmentation and pricing. The study emphasized that tree-based models provide valuable insights into data interactions, which are essential for insurers optimizing pricing strategies. Deep learning (DL) models, though less prevalent in insurance pricing, have been investigated for their predictive capabilities. [17] noted the increasing adoption of DL in risk pricing, particularly for longevity and accident risk prediction. [18] compared a DL model to logistic regression and decision trees for accident risk prediction using telemetry data. Although the DL model outperformed others in certain metrics, logistic regression was favored due to its greater interpretability. Similarly, [19] evaluated a DL model against logistic regression and random forests for driver risk categorization, again preferring logistic regression due to interpretability challenges and limited telemetry data availability. These studies highlight a critical trade-off in DL applications: while DL models can achieve high predictive accuracy, their complexity often restricts their use in regulated industries like insurance, where transparency is essential. Dynamic pricing systems, enabled by ML, have shown considerable promise in insurance. [2] developed a dynamic pricing system for online motor vehicle liability insurance using random forests, gradient boosting, and DL models. The study demonstrated that ML algorithms enabled rapid development, monitoring, and updating of pricing models, achieving efficiencies unattainable by traditional methods. The findings concluded that high-quality prediction models with fast implementation speeds provide significant benefits, such as task automation and improved customer responsiveness. However, applying ML to pricing faces challenges, particularly due to regulatory requirements. [20] emphasized the importance of cross-validation techniques to minimize bias and ensure fairness in ML-based pricing models, addressing concerns about equitable premium determination. Interpretability remains a critical consideration in ML applications for insurance. [21] investigated the interpretability of tree-based models using SHAP and feature importance techniques, noting that SHAP is particularly effective for complex models like XGBoost and neural networks, while feature importance is commonly applied to simpler models like linear regression and decision trees. Similarly, [22] applied linear regression, decision trees, and random forests to predict insurance premiums, using SHAP to explain model outputs. These studies highlight the importance of transparent models in fostering trust with regulators and policyholders, especially when determining premiums that affect customer affordability. The reviewed literature indicates that ML models, including linear regression, random forests, decision trees, gradient boosting, XGBoost, SVR, KNN, and neural networks, have been widely applied to predict insurance premiums, classify risks, and manage claims. Regression-based models enable quantification of independent variable impacts, facilitating accurate premium predictions [23]. Among these, XGBoost and gradient boosting consistently demonstrate superior performance, though their computational demands can pose challenges. While health and motor insurance have received significant attention, there is a notable gap in applying ML to home insurance premium prediction. Most studies focus on health insurance datasets, leaving home insurance—a domain with unique risk factors such as property characteristics and environmental hazards—relatively underexplored. This study aims to address this gap by applying ML models to historical home insurance data to predict premiums and identify influential features using SHAP for interpret

2.1 Research Gap and Problem formulation

People and businesses face risks daily, and these risks can lead to losses such as illness, accidents, or damage to property. Insurance companies or insurers, help protect against these risks by transferring the financial burden to themselves. They do this in exchange for a fee called a premium. FIGURE 1, illustrates this process. For insurance companies to maintain their financial stability, they must set premiums cautiously based on how much risk each customer has or their risk profile [15]. Before setting a premium, insurance companies assess the risks by evaluating different risk factors. In the case of home insurance, they look at several things, like how the building is used (e.g., as an office, warehouse, etc.), the financial health of the person buying the insurance, the building's location, what the building is made of and how it is structured, whether safety features like fire protection or surveillance are installed, and the applicant's past insurance claims [24]. This information helps insurance companies understand the risks and the amount of loss that would be incurred if a claim is made. The information also helps them suggest ways of reducing the risks from occurring by suggesting measures like surveillance installation or renovations in some parts of the house/building, etc.

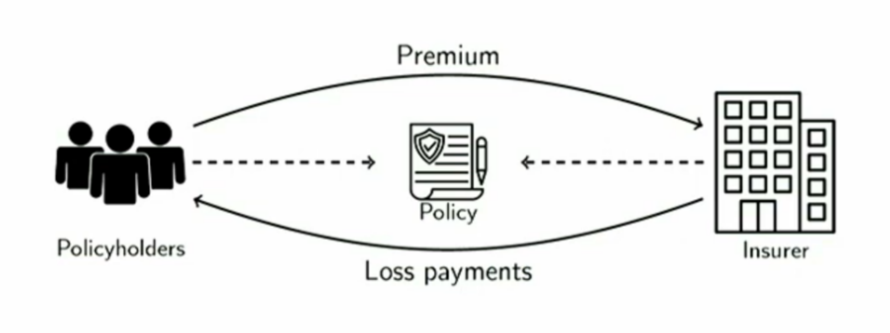


Figure 1: Insurance cycle (source [15])

While this process is detailed, it can be slow and hard to scale and sometimes depends on expert judgment, which might be biased. Machine learning offers a way to make this process fast and efficient by learning from past data [2]. Despite this, most ML models are considered ”black boxes” because it's hard to understand their decision-making process. This study aims to solve this problem by applying SHAP. It explains how each feature contributes to the prediction, therefore helping in understanding the factors that matter the most. This, in turn, makes the model more useful and transparent for insurance companies.

3. DATA AND METHODS

3.1 Data Description

This research used home insurance data from Kaggle to develop an ML model to predict home insurance premiums. The chosen dataset originated from the 'Universite de Technologié de Troyes (UTT), France' and was originally used in an R programming language course to gain insights. It includes policies spanning five years, from 2007 to 2012. UTT obtained the dataset from a home insurance company, and it contains various policy characteristics, such as building attributes, geographical zones, and coverage details, among others. In total, the dataset consists of $256,136$ observations with 66 features that describe the policy characteristics. These features provide valuable information for developing a comprehensive home insurance premium prediction ML model. We chose the 'LAST ANN PREM GROSS' variable as our target variable, as our goal is to develop a machine learning predictive model that can accurately predict home insurance premiums. The remaining variables in the dataset will serve as our features, which will be utilized to estimate the premium amount that will be charged to a client. The dataset variables are defined in Tables A1 and A2, respectively in the appendix.

3.2 Machine Learning Models

Machine learning techniques enable computers to perform tasks by learning from training data and recognising patterns without the need to program them to run a specific task explicitly. In the context of ML, data has examples, and a set of features or variables describes each example [25]. The usefulness of an ML model in a specific task is evaluated using performance metrics that improve as the model gains more experience. To evaluate the performance of ML algorithms, a range of statistical and mathematical approaches is employed [11]. Once the learning process is over, the trained models are used to predict, classify or cluster new sample (test) data based on the knowledge acquired during the learning phase. Machine learning can be divided into two main groups: supervised and unsupervised [11]. Supervised machine learning deals with labeled data and maps inputs to outputs [11]. Unsupervised machine learning algorithms look for patterns and structures in data and therefore do not need data labels. Unsupervised learning techniques are frequently used in clustering and dimension reduction [26] tasks. This work applied various supervised machine learning models to predict home insurance premiums.

3.2.1 Linear regression

Linear regression assumes a linear relationship between the explanatory (independent) variables and the dependent variable, making it a simple yet effective method for predictive modeling. It is particularly suitable for small sample sizes due to its simplicity and ease of interpretation. However, its performance may degrade when dealing with a large number of predictor variables, as it struggles to handle complex relationships or multicollinearity [27]. The goal of linear regression is to find the optimal coefficients that minimize the loss function, which is typically the sum of squared errors between the predicted and actual values. The linear regression model can be expressed as:

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_n x_{in} + \epsilon_i, \tag{1}$$

where y_i is the dependent variable for the i-th observation, θ_0 is the intercept, $\theta_1, \ldots, \theta_n$ are the coefficients for the independent variables $x_{i1}, \ldots, x_{in}$, and ϵ_i is the error term. The loss function for linear regression, which measures the model's fit, is the mean squared error (MSE), defined as:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{k=1}^{n} \theta_k x_{ik} \right)^2 . \tag{2}$$

This loss function is minimized to estimate the coefficients $\theta_0, \theta_1, \ldots, \theta_n$, ensuring the best fit to the data while maintaining simplicity and interpretability.

3.2.2 Ridge regression

Ridge regression extends linear regression by incorporating a regularization component to the loss function, addressing issues like overfitting and multicollinearity. The regularization term penalizes large coefficients, which helps improve the model's generalization performance [28]. This penalty is proportional to the sum of the squared coefficients, and as the regularization parameter increases, the coefficients are shrunk toward zero, reducing the model's variance. Ridge regression balances model complexity and predictive accuracy, making it suitable for datasets with highly correlated predictors [29]. The ridge regression model minimizes the following loss function:

$$\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{k=1}^{m}\theta_k x_{ik}\right)^2+\lambda\sum_{k=1}^{m}\theta_k^2, \tag{3}$$

where $\lambda \geq 0$ is the regularization parameter that controls the trade-off between the model fit (the first term, $\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{k=1}^{m}\theta_k x_{ik}\right)^2$) and the penalty term ($\lambda\sum_{k=1}^{m}\theta_k^2$). The parameter λ balances the importance of minimizing prediction errors with controlling coefficient size, enhancing the model's robustness to overfitting [30].

3.2.3 Lasso regression

Lasso regression, like ridge regression, adds a regularization term to the linear regression loss function but penalizes the absolute values of the coefficients rather than their squares. This L1 regularization approach encourages sparsity by shrinking less important coefficients to exactly zero, effectively performing feature selection [30]. By reducing the magnitude of coefficients, lasso regression simplifies the model while maintaining predictive accuracy, making it particularly useful when dealing with high-dimensional datasets. The lasso regression model minimizes the following loss function:

$$\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{k=1}^{m}\theta_k x_{ik}\right)^2+\lambda\sum_{k=1}^{m}|\theta_k|, \tag{4}$$

where $\lambda \geq 0$ is the regularization parameter that controls the strength of the penalty. The first term, $\sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{k=1}^{m}\theta_k x_{ik}\right)^2$, represents the sum of squared errors, while the penalty term, $\lambda\sum_{k=1}^{m}|\theta_k|$, shrinks coefficients to reduce model complexity. Lasso regression's ability to eliminate irrelevant features makes it a powerful tool for predictive modeling in insurance, where identifying key risk factors is critical. Lasso regression is efficient in reducing the number of features used in a model due to its ability to drive the coefficients (slopes) of certain features to zero. As the regularization parameter λ increases, the slope values gradually decrease.

3.2.4 K-nearest neighbours (KNN)

KNN is used for predicting numerical targets based on a similarity measure, such as distance functions. Unlike linear or polynomial regression, KNN does not assume any underlying relationship between the features and target variables. While linear regression and multiple regression rely on the assumption of linear relationships, KNN regression leverages patterns in the data to make predictions [31]. In KNN regression, the algorithm identifies the k nearest neighbours to a given data case and predicts the target value by calculating the mean value of those neighbours. An alternative approach is using an average weighted inverse distance of the k nearest neighbours. It is worth noting that KNN uses the same distance function for both regression and classification problems [32]. Euclidean, Minkowski and Manhattan distances are commonly used for continuous variables, while the Hamming distance function is used on categorical variables.

3.2.5 Support vector regression (SVR)

Support vector machines (SVMs) were first developed for classification tasks. They were later extended to solve regression tasks [33]. In SVR, the problem is formulated as an optimisation task to find the flattest ϵ-insensitive tube that has most of the training instances. This involves minimising a convex ϵ-insensitive loss function that combines the loss function and the geometrical properties of the tube [34]. The optimisation task has a unique solution that can be solved using numerical optimisation algorithms. Like SVMs, SVR assumes that the test and training data are from the same unknown probability distribution function and are independent and identically distributed.

3.2.6 Decision trees

Decision trees have a tree-like structure with leaf nodes, internal nodes, branches and root nodes. They output their predictions by learning a dataset's decision rules. Each internal node shows the test performed on a specific feature, and each branch shows the possible outcomes of that test. The leaf or terminal nodes store the final class labels or regression predictions. The construction of a decision tree involves recursively splitting the training data into subsets based on feature values until a stopping criterion is met. This criterion can be defined by parameters, like the maximum depth of the tree or the minimum number of samples required to split a node [35].

3.2.7 Random forests

Random forests combine the prediction outputs of multiple decision trees to make predictions. It works by training numerous decision trees on a dataset with a tree-like structure and then aggregating the output from each tree to produce a final result. This algorithm handles complex datasets and captures intricate relationships between features by leveraging the ensemble nature of random forests and the ability to adjust hyperparameters efficiently. The random forest regressor's accuracy and flexibility in customization make it a powerful technique in machine learning [11].

3.2.8 Gradient boosting model

It is an ensemble method that sequentially combines the predictions of multiple weak learners, often decision trees. Its main goal is to improve a model's overall predictive performance. It does this by optimizing the model's weights based on the previous iterations' errors, thereby reducing the prediction errors gradually [11]. By leveraging the strengths of individual weak learners and iteratively learning from their mistakes, gradient boosting creates a powerful and accurate ensemble model. It is efficient in handling complex datasets and capturing complex relationships between features [36].

3.2.9 Extreme gradient boosting(XGBoost)

XGBoost is an enhanced gradient boosting algorithm. XGBoost minimises an objective function that combines a loss function with regularisation terms, such as L1 and L2 regularisation. By optimising the regularised objective function, XGBoost aims to find the best possible predictions by iteratively constructing weak models and adjusting their predictions based on the residuals. This iterative process leads to highly accurate predictions and is a key characteristic of XGBoost [37].

3.2.10 Shapley additive explanation (SHAP)

SHAP has gained popularity as a method for interpreting machine learning model predictions. It was introduced by [3] and helps identify which features contribute most to the predictions. It is based on game theory, which calculates the contribution of each player (features) to the playout (predictions). SHAP calculates a score for each feature in the model, which represents its weight to the model output. To get the scores, it considers all combinations between the features to cover all cases where all features and a subset of features are used. It is suitable for explaining complex models such as XGBoost, as it provides both local and global insights into a model's behaviour. Global interpretability involves understanding the model's overall behaviour, that is, which features are generally the most important across all predictions. It ranks features based on their influence /impact on the model's output. Local interpretability focuses on individual predictions. For example, if our model predicts the premium to be paid, SHAP breaks down the exact contributions of each feature to that prediction [38, 39].

3.2.11 Model evaluation

Performance metrics or error measurements are used to compare predicted values to actual data values [40]. The following performance evaluation metrics were used in this study:

- **R-Square Score (r2):** It measures how well the predictions fit the real values. r2 is calculated as the explained variance divided by the total variance and ranges from 0 to 1. A higher r2 value closer to 1 generally shows that the model is performing well.

$$r2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2}, \tag{5}$$

  m is the sample size, y_i is the actual value, $\bar{y}_i$ is the mean, $\hat{y}_i$ is the predicted value, $\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ represents the sum of squared regression errors, and $\sum_{i=1}^{m}(y_i - \bar{y}_i)^2$ represents the sum of squared total errors. [41].

- **Mean Square Error (MSE):** the average of the squared differences between the predicted and actual values.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2, \tag{6}$$

  m is the sample size, y_i the actual value, and $\hat{y}_i$ the prediction for the i^{th} sample. A lower MSE value shows better performance.

- **Mean Absolute Error (MAE):** Calculates the average of the absolute differences between the predictions and actual values.

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|, \tag{7}$$

where m is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value for the i^{th} sample. Like MSE, a lower MAE value shows a good performance.

4. RESULTS AND DISCUSSION

In this section the data preprocessing steps, model development, and the obtained results are discussed.

4.1 Data Pre-processing

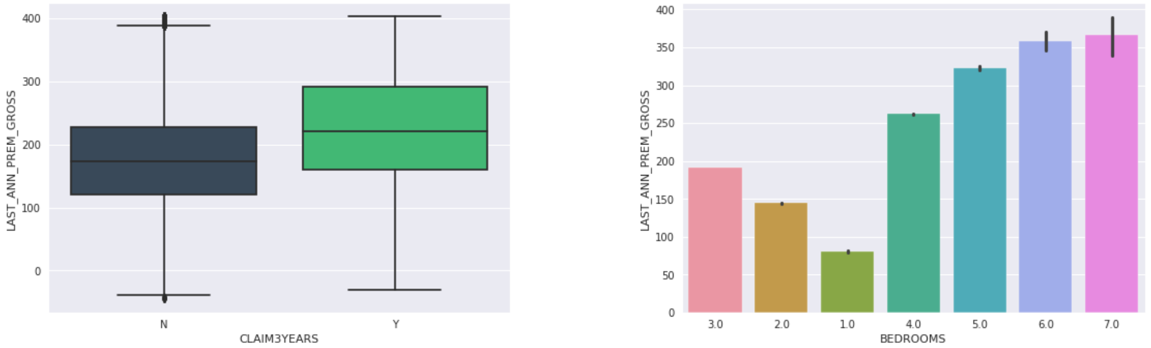
The first step was examining missing values and duplicates. The dataset had some missing values, but no duplicates. Columns with more than 70% of missing values were dropped, and for those with less than this threshold, the missing data were imputed using the mean. Two new features, ”client age” and ”building age,” were created by calculating the difference between the client’s date of birth and the construction date of the building and the date the insurance coverage was purchased, respectively.

Data visualisation was carried out to help in identifying patterns and relationships between variables (correlation), outliers and the data’s distribution. The presence of outliers in the dataset was addressed using the interquartile range (IQR) and the Gaussian-based Winsorizer method. The two methods manage outliers by capping extreme values. The categorical data were encoded into a numerical format using label encoding. Data standardisation ensured the numerical variables were on a comparable scale. This was done by calculating z-scores using the `scipy.stats.zscore()` function.

FIGURE 2a, shows that clients who had reported a claim in the past three years paid higher premiums compared to those who had not reported any claims in the same period. This suggests that a client’s claim history influences the insurance premium charged. Additionally, FIGURE 2b, demonstrates that properties with more bedrooms are associated with higher insurance premiums. This suggests that the size of the property impacts the cost of insurance.

A correlation plot provides a graphical representation of the relationships between variables with values between -1 and 1. The relationship could be either positive or negative. A strong positive relationship is shown by a value near 1, while a strong negative relationship is shown by a value near -1.

The highly correlated variables were removed to avoid redundancy. The heatmap in FIGURE 3, illustrates the dropped highly correlated variables with a correlation threshold of 0.90. In the end, 50 variables were left after dropping 6 highly correlated variables.



(a) Premium vs Claim3years.

(b) Premium vs Number of bedrooms.

Figure 2: Premium values under different feature: (a) Claim3years, (b) Number of bedrooms.

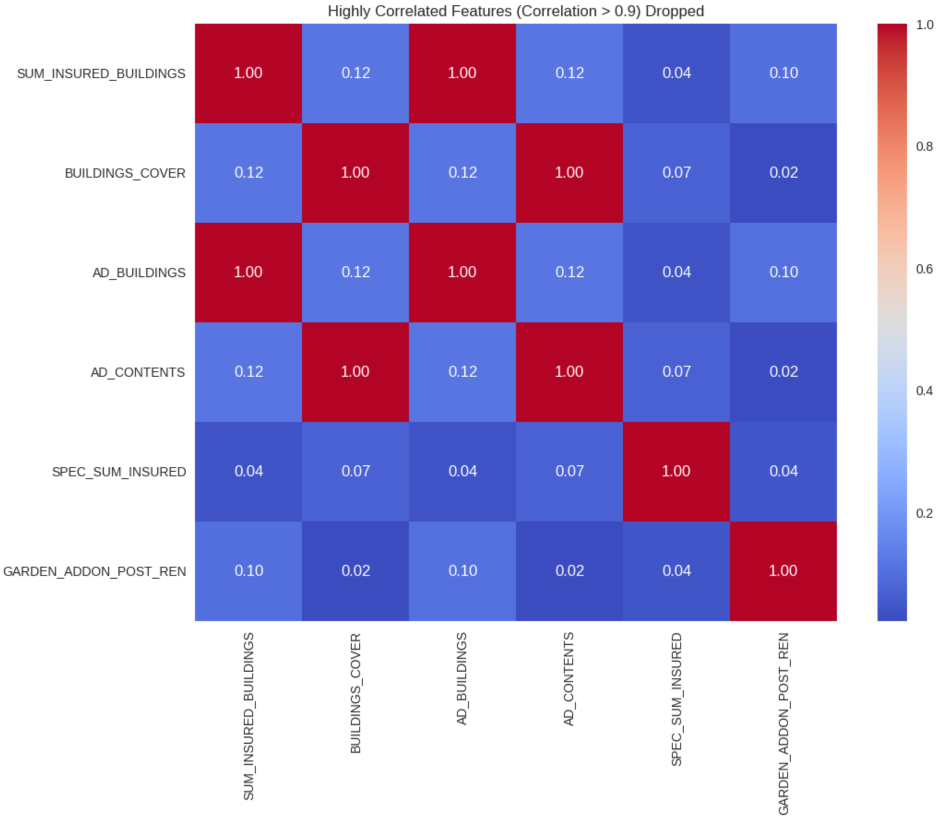


Figure 3: Highly correlated values.

4.2 Fitting and Evaluating the Models

Before fitting the supervised ML models, the dataset was divided into two: 80% for training and 20% for testing. The random seed value was set 42 for reproducibility. To mitigate overfitting, k-fold cross-validation (CV) was applied, dividing the training set into 5 folds. The training set is partitioned into five folds through five iterations. Each iteration used four folds for training and the remaining one fold for validation. The results across all the folds were then averaged [42].

4.2.1 Selecting the best model

XGBoost had the highest r2 score and the lowest values for both MSE and MAE on the test set. These results showed that XGBoost outperformed the other models in accurately predicting home insurance premiums. In this context, XGBoost is therefore regarded as the most effective model for predicting home insurance premiums. On the other hand, support vector regression (SVR) performed poorly in providing accurate premium predictions. It had relatively higher values for both MAE and MSE compared to the other models. Therefore, SVR may not be the most suitable choice for predicting home insurance premiums in our case. This can be seen in Table 1.

Table 1: Model Performance on test set: Best Scores are in Bold

Model	R Squared	R Squared (CV)	MSE	MAE
Linear Regression	0.7161	0.7125	2224.03	35.97
Ridge Regression	0.7161	0.7126	2224.03	35.97
Lasso Regression	0.7064	0.7026	2299.88	36.53
Decision Tree	0.6374	0.6336	2840.42	36.32
Random Forest	0.8137	0.8163	1458.58	26.34
KNN	0.6556	0.6516	2697.71	36.50
Gradient Boosting	0.8014	0.7978	1515.83	25.43
XGBoost	**0.8344**	**0.8305**	**1297.52**	**25.00**
SVR	0.0041	0.0050	7801.03	69.14

[1] Evaluation of various regression models. R Squared (CV) refers to the average R Squared from 5-fold cross-validation.

[2] Bolded values indicate the best performance for each metric on the test set.

4.2.2 Hyperparameter tuning

The study applied 9 machine learning models to the prediction problem. These models were later narrowed down to the top 3 best-performing models, that is, random forest, gradient boosting and XGBoost. This was to save time on tuning the models' hyperparameters by focusing on the best-performing ones. For hyperparameter tuning, the GridSearchCV was applied to the 3 models to test different combinations of hyperparameters to improve their performance. These 3 ML models are all supervised ensemble machine-learning techniques that are known for their effectiveness in regression tasks.

Fine-tuning the hyperparameters led to improved predictive performance for these three models. XGBoost still outperformed both gradient-boosting and random forest models in predicting home insurance premiums, as seen in TABLE 2.

Table 2: Model Performance on test set after Hyperparameter Tuning: Best Scores are in Bold

Model	R Squared	R Squared (CV)	MSE	MAE
Random Forest	0.8173	0.8134	1430.48	26.53
Gradient Boosting	0.8307	0.8279	1325.36	25.43
XGBoost	**0.8380**	**0.8352**	**1269.08**	**24.62**

Model evaluation metrics after hyperparameter tuning.

4.3 Feature Importance and Final Xgboost Model

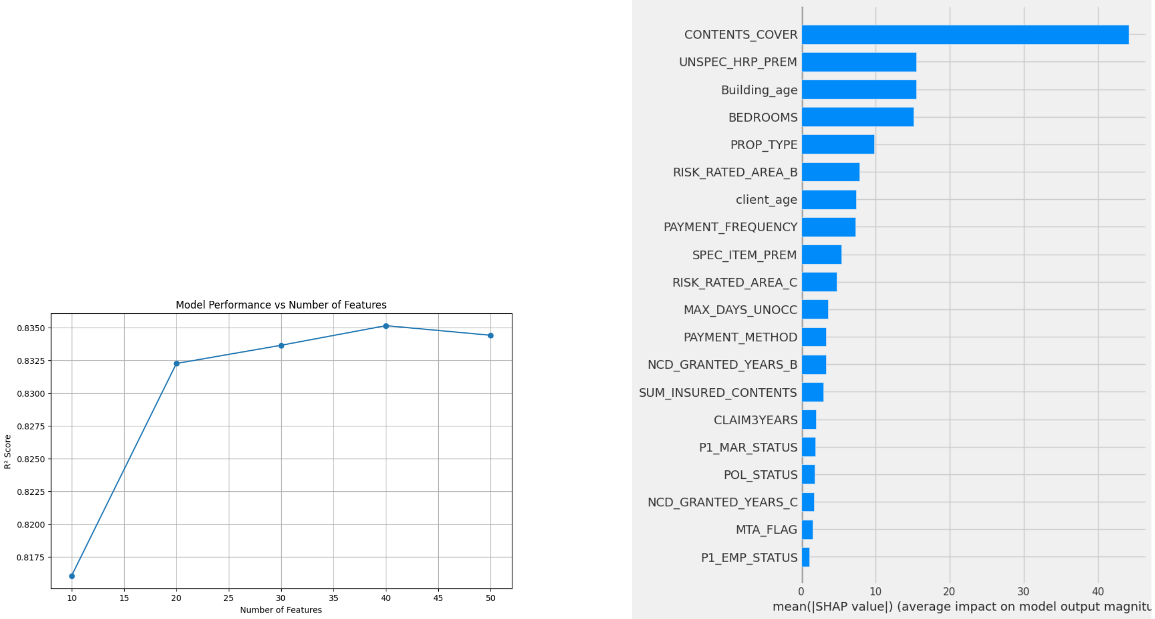
The study focused on XGBoost as it had the best performance compared to random forest and gradient boosting, and proceeded to apply the Shapley additive explanations (SHAP) method on the XGBoost model to determine the most influential features, improving its interpretability capabilities. SHAP assigns a numerical value to each feature, indicating its contribution to the model's predictions. The most influential feature has the highest value. SHAP is applied to determine the top k features that highly contribute to predictions. A SHAP summary plot is a visual representation that illustrates the contribution of each feature to the model's prediction. Features are listed on the vertical axis according to their level of importance, and horizontal bars represent the contribution of each feature to the prediction [43]. In this study, the k values were 10, 20, 30, 40 and 50. The top 40 feature, as seen in FIGURE 4a, produced the highest results with an R-squared score of 0.8551 and 0.8351 for the train and test sets, respectively, as seen in TABLE 3, before hyperparameter tuning. Therefore, these 40 features were selected to fit the final model, ensuring its focus on the most influential features. This suggests that instead of insurance companies collecting a lot of information from prospective clients, they could narrow it down to only a few key features that significantly contribute to the machine learning models. We only highlighted the top 20 important features on the SHAP summary plot as seen in FIGURE 4b.

Table 3: XGBoost Performance on Train and Test after applying SHAP

	R Squared	MSE	MAE
Train	0.8551	1120.97	23.50
Test	0.8351	1291.48	24.95

Performance metrics for XGBoost model after applying SHAP.

According to FIGURE 4b, the feature 'CONTENTS COVER' was the most influential factor. It indicates whether personal objects are under the cover or not. Additionally, the number of bedrooms in the house, the age of the building, and the property type were also identified as significant features. Adding such important features improves the model's ability to consider important factors that impact insurance premiums accurately.



(a) Top-k Features.

(b) Important Features.

Figure 4: Visualization of SHAP feature importance.

4.3.1 Fitting the final model: Xgboost

Hyperparameter tuning using GridSearchCV was applied on the XGBoost model fit on the 40 features. A combination of hyperparameters like `n_estimators`$(50, 100, 200)$, `max_depth`$(3, 5, 7)$, `learning_rate`$(0.01, 0.1, 0.2)$, and `gamma` $(0, 0.1, 0.2, 0.3, 0.4)$. This improved the model's performance slightly with an r2 score of 0.8799 and 0.8383 on the train and test sets, respectively, as seen in TABLE 4. This implies that approximately 84% of the premiums are correctly predicted by the model. The MAE and MSE are consistent across the train and test sets, implying that the model generalises well without overfitting.

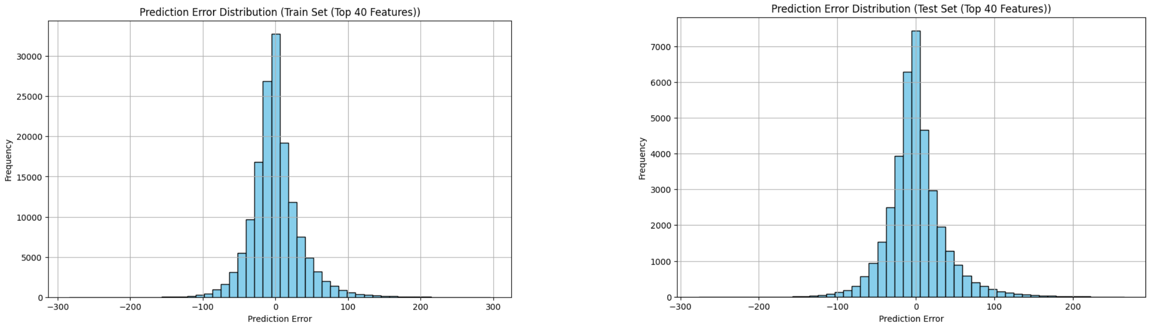
Table 4: Hyperparameter-Tuned: Final XGBoost Performance on Train, Test, and Cross validation (CV)

	R Squared	**MSE**	**MAE**
Train	0.8799	929.27	21.37
Test	0.8383	1266.87	24.56
CV	0.8351	1278.71	24.64

Performance metrics for the XGBoost model after hyperparameter tuning.

4.4 Prediction Error Distribution Analysis

FIGURE 5a and FIGURE 5b, show the prediction error distribution for training and testing sets, respectively. The distribution of prediction errors helps in evaluating the consistency and generalizability of the final XGBoost model in predicting home insurance premiums. The error distributions in both cases are centred around zero, with a symmetrical bell-shaped curve. This shows that the model's predictions are not biased or skewed, implying that the errors are relatively balanced in both directions. The training set, FIGURE 5a, shows a slightly narrower contribution due to the models' exposure to the data during training. The testing set, FIGURE 5b, displays a similar shape, demonstrating that the model generalises well to unseen data. These patterns enhance the reliability of the model as reported by the evaluation metrics, that is, the R-squared, MSE and MAE.



(a) Train: prediction error distribution using the top 40 SHAP-ranked features.

(b) Test: prediction error distribution using the top 40 SHAP-ranked features.

Figure 5: Prediction error distributions with top-40 SHAP-ranked features on train and test sets.

5. CONCLUSION

This study applied various supervised regression-based models to predict home insurance premiums based on customer and property traits. XGBoost outperformed the other applied regression models. The SHAP-based feature selection method was applied on the clean dataset that had 50 features. SHAP showed that the top 40 most influential features yielded the best performance, demonstrating its effectiveness in improving model interpretability while maintaining high performance. This proves that reduced features can still obtain good predictive results. Fewer features also tend to lower the complexity of the model. The final XGBoost model fit on the top 40 important features had an R-squared score of 0.8799 and 0.8383 on the train and test set, respectively. The MAE and MSE on both the training and test sets were consistent, illustrating the model's generalizability. On the other hand, the predictive error distribution plots showed the robustness of the model in capturing the main structure of the data on different premium price changes.

In conclusion, the results show that applying SHAP feature selection provides a reliable, interpretable and high-performing model for insurance premium prediction. XGBoost model can be applied by insurance companies in building premium prediction models due to its high performance.

5.1 Future Work

One limitation of the study is the inaccessibility of a local company's home insurance data and the data is possibly outdated covering the years 2007 to 2012. For future research, we recommend applying similar machine learning regression models to a more recent and locally sourced insurance data set. This would allow the model to be evaluated in real time setting and compare its performance on different companies data.

We acknowledge that our work only focused on machine learning models to predict home insurance premiums. In future work, we propose to apply deep learning methods on premium predictions. We applied stochastic AI implying that the predicted outcomes maybe uncertain and vary even with the same initial set conditions and actions.

Abbreviations

ML	Machine Learning
SVM	Support Vector Machine
SVR	Support Vector Regression
DL	Deep Learning
CV	Cross-validation
NN	Neural Network
MSE	Mean Squared Error
MAE	Mean Absolute Error
KNN	K-Nearest Neighbors
UPC	United Property and Casualty Insurance Co.
XGBoost	Extreme Gradient Boosting
RF	Random Forest
DT	Decision Tree
ANN	Artificial Neural Network
r2	R-Squared
SHAP	Shapley additive explanations

References

[1] Angra S, Ahuja S. Machine learning and its applications: a review. In: International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE; 2017;57-60.

[2] Grize YL, Fischer W, Lützelschwab C. Machine learning applications in nonlife insurance. Appl Stoch Models Bus Ind. 2020;36:523-537.

[3] Tsanakas A, Desli E. Measurement and pricing of risk in insurance markets. Risk Anal. 2005;25:1653-1668.

[4] Lyubchich V, Newlands NK, Ghahari A, Mahdi T, Gel YR. Insurance risk assessment in the face of climate change: integrating data science and statistics. WIREs Computational Stats.

2019;11:e1462.

[5] Saunders J. Insolvent insurer united property & casualty headed to receivership. Daily business review, February 2023.

[6] Societies in Kenya. Ministry of agriculture, livestock, fisheries and co-operatives. 2021.

[7] Shreekar C, Kiran M, Sumanth D, Jeevan P. Costprediction of health insurance. Int Res J Eng Technol. 2023;10.

[8] Aminul Islam M, Nag A, Chandra P, Ahmed Fahim SMF, Hoque MM. Healthcare cost patterns and prediction: investigating personal datasets using data analytics. 2023. Techexiv preprint: https://www.techrxiv.org/doi/full/10.36227/techrxiv.24457090.v1

[9] Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine learning-based regression framework to predict health insurance premiums. Int J Environ Res Public Health. 2022;19:7898.

[10] Sahai R, Al-Ataby A, Assi S, Jayabalan M, Liatsis P, Loy CK et al. Insurance risk prediction using machine learning. In: The International Conference on Data Science and Emerging Technologies. Springer. 2022:419-33.

[11] https://thesis.eur.nl/pub/70185/Predicting-insurance-premiums-with-Machine-Learning-Thesis-Sven-Groen.pdf

[12] Orji U, Ukwandu E. Machine learning for an explainable cost prediction of medical insurance. Mach Learn Appl. 2024;15:100516.

[13] Alamir E, Urgessa T, GopiKrishna T, Ellappan V. Application of machine learning with big data analytics in the insurance industry; 2020.

[14] Panda S, Purkayastha B, Das D, Chakraborty M, Biswas SK. Health insurance cost prediction using regression models. In: International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON).IEEE. 2022;1:168-173.

[15] Henckaerts R. Insurance pricing in the era of machine learning and telematics technology; 2021.

[16] Henckaerts R, Côté MP, Antonio K, Verbelen R. Boosting insights in insurance tariff plans with tree-based machine learning methods. N Am Actuarial J. 2021;25(2):255-85.

[17] Levantesi S, Nigri A, Gabriella P, et al. Longevity risk management through machine learning: state of the art. Ins Markets Companies. 2020;1:11-20.

[18] Paefgen J, Staake T, Thiesse F. Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. Decis Support Syst. 2013;56:192-201.

[19] Baecke P, Bocca L. The value of vehicle telematics data in insurance risk selection processes. Decis Support Syst. 2017;98:69-79.

[20] Kristiansson J. Double machine learning for insurance price optimization; 2023.

[21] Jones KI, Sah S. The implementation of machine learning in the insurance industry with big data analytics. Int J Data Inform Intell Comput. 2023;2:21-38.

[22] Rao AR, Jain R, Singh M, Garg R. Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data. Healthc Anal. 2024;6:100351.

[23] Hossen S. Medical insurance cost prediction using machine learning [PhD thesis], 10; 2023.

[24] `https://baleario.com/what-is-a-home-insurance-risk-assessment/`

[25] Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. Sensors (Basel). August 2018;18:2674.

[26] Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Berry MW, Mohamed A, Yap BW, editors. Supervised and unsupervised learning for data science. Cham: Springer International Publishing; 2020:3-21.

[27] Yang X, Wen W. Ridge and lasso regression models for cross-version defect prediction. IEEE Trans Reliab. 2018;67:885-96.

[28] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55-67.

[29] Jackson SS. Chapter 3. ridge regression; 2023.

[30] `https://www.stat.uchicago.edu/~yibi/teaching/stat224/L18.pdf`

[31] Barrash S, Shen Y, Giannakis GB. Scalable and adaptive knn for regression over graphs. In: 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). IEEE. 2019:241-245.

[32] Sayad S. K nearest neighbors. technical report. University of Toronto; 2010.

[33] Bathla G. Stock price prediction using LSTM and svr. In: Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE; 2020:211-214.

[34] Awad M, Khanna R, Awad M, Khanna R. Support vector regression. Efficient learning machines: theories, concepts, and applications for engineers and system designers; 2015:67-80.

[35] Tike A. A medical price prediction system; 2018.

[36] Poufinas T, Gogas P, Papadimitriou T, Zaganidis E. Machine learning in forecasting motor insurance claims. Risks. 2023;11:164.

[37] Nalluri M, Pentela M, Eluri NR. A scalable tree boosting system: Xg boost. Int J Res Stud Sci Eng Technol. 2020;7:36-51.

[38] Ekici B. Interpretable machine learning for auto insurance fraud detection using Shapley additive explanations [PhD thesis]. Tilburg University; 2025.

[39] Wang H, Liang Q, Hancock JT, Khoshgoftaar TM. Feature selection strategies: a comparative analysis of shap-value and importance-based methods. J Big Data. 2024;11(1):44.

[40] Botchkarev A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. 2018. ArXiv preprint: https://arxiv.org/pdf/1809.03006

[41] Cai J, Xu K, Zhu Y, Hu F, Li L. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Appl Energy. 2020;262:114566.

[42] Berrar D. Cross-validation. 2019.

[43] Antonini AS, Tanzola J, Asiain L, Ferracutti GR, Castro SM, et al. Machine learning model interpretability using shap values: application to igneous rock classification task. Appl Comput Geosci. 2024;23:100178.

6. Appendix

Table A1: Description of the first half of variables in the home insurance dataset, including their type (discrete or continuous) and role as input or output in the predictive model. These variables capture customer, property, and policy characteristics for premium prediction.

Variable	Description	Type	Input/ Output
QUOTE_DATE	Day the quotation was made	Discrete	Input
COVER_START	Beginning of the cover payment	Discrete	Input
CLAIM3YEARS	3-year loss history	Discrete	Input
P1_EMP_STATUS	Client's professional status	Discrete	Input
P1_PT_EMP_STATUS	Client's part-time professional status	Discrete	Input
BUS_USE	Commercial use indicator	Discrete	Input
CLERICAL	Administration office usage indicator	Discrete	Input
AD_BUILDINGS	Building coverage - self damage	Discrete	Input
RISK_RATED_AREA_B	Geographical classification of risk - building	Continuous	Input
SUM_INSURED_BUILDINGS	Assured sum - building	Continuous	Input
NCD_GRANTED_YEARS_B	Bonus malus - building	Continuous	Input
AD_CONTENTS	Coverage of personal items - self damage	Discrete	Input
RISK_RATED_AREA_C	Geographical classification of risk - personal objects	Continuous	Input
SUM_INSURED_CONTENTS	Assured sum - personal items	Continuous	Input
NCD_GRANTED_YEARS_C	Malus bonus - personal items	Continuous	Input
CONTENTS_COVER	Coverage - personal objects indicator	Discrete	Input
BUILDINGS_COVER	Cover - building indicator	Discrete	Input
SPEC_SUM_INSURED	Assured sum - valuable personal property	Continuous	Input
SPEC_ITEM_PREM	Premium - personal valuable items	Continuous	Input
UNSPEC_HRP_PREM	Unknown premium component	Continuous	Input
P1_DOB	Date of birth of the client	Discrete	Input
P1_MAR_STATUS	Marital status of the client	Discrete	Input
P1_POLICY_REFUSED	Policy emission denial indicator	Discrete	Input
P1_SEX	Customer sex	Discrete	Input
APPR_ALARM	Appropriate alarm indicator	Discrete	Input
APPR_LOCKS	Appropriate lock indicator	Discrete	Input
BEDROOMS	Number of bedrooms	Continuous	Input
ROOF_CONSTRUCTION	Code of roof construction type	Continuous	Input
WALL_CONSTRUCTION	Code of wall construction type	Continuous	Input
FLOODING	House susceptible to floods	Discrete	Input
LISTED	National heritage building status	Continuous	Input
MAX_DAYS_UNOCC	Number of days unoccupied	Continuous	Input
NEIGH_WATCH	Vigils of proximity present	Discrete	Input

Table A2: Description of the second half of variables in the home insurance dataset, including their type (discrete or continuous) and role as input or output in the predictive model. These variables further detail property and policy characteristics for premium prediction.

Variable	Description	Type	Input/ Output
OCC_STATUS	Occupancy status	Discrete	Input
OWNERSHIP_TYPE	Type of membership	Continuous	Input
PAYING_GUESTS	Presence of paying guests	Continuous	Input
PROP_TYPE	Type of property	Continuous	Input
SAFE_INSTALLED	Safe installation indicator	Discrete	Input
SEC_DISC_REQ	Premium reduction for security	Discrete	Input
SUBSIDENCE	Subsidence indicator	Discrete	Input
YEARBUILT	Year of construction	Continuous	Input
CAMPAIGN_DESC	Description of the marketing campaign	Continuous	Input
PAYMENT_METHOD	Method of payment	Discrete	Input
PAYMENT_FREQUENCY	Frequency of payment	Continuous	Input
LEGAL_ADDON_PRE_REN	Legal fees option before 1st renewal	Discrete	Input
LEGAL_ADDON_POST_REN	Legal fees option after 1st renewal	Discrete	Input
HOME_EM_ADDON_PRE_REN	Emergencies option before 1st renewal	Discrete	Input
HOME_EM_ADDON_POST_REN	Emergencies option after 1st renewal	Discrete	Input
GARDEN_ADDON_PRE_REN	Gardens option before 1st renewal	Discrete	Input
GARDEN_ADDON_POST_REN	Gardens option after 1st renewal	Discrete	Input
KEYCARE_ADDON_PRE_REN	Key replacement option before 1st renewal	Discrete	Input
KEYCARE_ADDON_POST_REN	Key replacement option after 1st renewal	Discrete	Input
HP1_ADDON_PRE_REN	HP1 option before 1st renewal	Discrete	Input
HP1_ADDON_POST_REN	HP1 option after 1st renewal	Discrete	Input
HP2_ADDON_PRE_REN	HP2 option before 1st renewal	Discrete	Input
HP2_ADDON_POST_REN	HP2 option after 1st renewal	Discrete	Input
HP3_ADDON_PRE_REN	HP3 option before 1st renewal	Discrete	Input
HP3_ADDON_POST_REN	HP3 option after 1st renewal	Discrete	Input
MTA_FLAG	Mid-term adjustment indicator	Discrete	Input
MTA_FAP	Bonus up to date of adjustment	Continuous	Input
MTA_APRP	Premium adjustment for mid-term	Continuous	Input
MTA_DATE	Date of mid-term adjustment	Discrete	Input
LAST_ANN_PREM_GROSS	Total premium for the previous year	Continuous	Output
POL_STATUS	Policy status	Discrete	Input
Police	Policy number	Discrete	Input