Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14194
e- ISSN : 2541-2019
p- ISSN : 2541-044X

Comparative Study of XGBoost, Random Forest, and Logistic Regression Models for Predicting Customer Interest in Vehicle Insurance

Gregorius Airlangga
[1)]Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia
[1)]gregorius.airlangga@atmajaya.ac.id

Submitted :Oct 4, 2024 | **Accepted** : Oct 28, 2024 | **Published** : Oct 30, 2024

Abstract: In today's competitive insurance market, accurately predicting customer interest in additional products, such as vehicle insurance, is crucial for optimizing marketing strategies and maximizing sales. This study presents a comparative analysis of three machine learning models—XGBoost, RandomForest, and Logistic Regression—to predict customer interest in vehicle insurance based on a dataset that includes demographic, vehicle, and policy-related features. The dataset was analyzed using five-fold cross-validation, and the performance of the models was evaluated using AUC-ROC, precision, recall, and F1-score. XGBoost demonstrated the highest recall (0.9525) and AUC-ROC (0.7854), making it the most effective model for identifying customers interested in vehicle insurance, though at the expense of lower precision (0.2585). RandomForest showed a more balanced trade-off between precision (0.3064) and recall (0.5341) but performed lower overall. Logistic Regression, while the most interpretable model, exhibited high variability in performance across different folds, with a lower average precision (0.2372). The findings of this research highlight that XGBoost is ideal for maximizing recall in high-volume campaigns, while RandomForest may be better suited for applications requiring fewer false positives. These results offer valuable insights into model selection based on business objectives and resource allocation.

Keywords: Vehicle Insurance Prediction; Machine Learning; XGBoost; RandomForest; Logistic Regression

INTRODUCTION

In the insurance industry, predicting customer behavior is a critical aspect of business strategy, particularly when it comes to cross-selling products such as vehicle insurance to existing health insurance policyholders (Kaswan, Dhatterwal, Sharma, & Sood, 2022; Staudt & Wagner, 2022; Tondi, 2024). Accurate predictions help companies optimize their marketing efforts and customer outreach, ensuring that they target the right customers with the right products at the right time (Chaffey & Smith, 2022; Gupta & Joshi, 2022; Wang, 2022). However, this task becomes increasingly complex when the dataset contains imbalanced classes, as is often the case when a small portion of customers express interest in an additional insurance product like vehicle insurance (Hanafy & Ming, 2021). The success of predictive modeling in such scenarios hinges not only on the choice of machine learning algorithms but also on effective preprocessing techniques that address the challenges posed by imbalanced data (Dangut, 2021). Traditional models such as Logistic Regression, RandomForest, and Gradient Boosting have been widely used in predicting customer preferences (Kiangala & Wang, 2021) . Yet, these models often struggle with skewed datasets where the majority class, customers not interested in vehicle insurance overwhelms the minority class, leading to biased predictions and suboptimal performance (Washington, 2023).

Recent advancements in machine learning have brought forward techniques to address class imbalance, including oversampling methods such as SMOTE (Synthetic Minority Over-sampling Technique) (Datta, Ghosh, & Choudhury, 2024). SMOTE generates synthetic examples for the minority class by interpolating between existing instances, which helps balance the dataset and prevents the model from being overly biased toward the majority class (Chen, Zhang, Huang, Wu, & Luo, 2022). Despite its proven efficacy in improving model performance on imbalanced datasets, many studies have yet to fully explore its impact in the context of insurance-related predictions, where customer behavior is influenced by a myriad of factors such as demographics, previous insurance history, and vehicle characteristics (Singhal, Goyal, & Singhal, 2024). The urgency for developing sophisticated preprocessing strategies is driven by the growing complexity of insurance datasets, where multiple

*name of corresponding author

factors contribute to customer decisions (Śmietanka, Koshiyama, & Treleaven, 2021). Without proper preprocessing, even the most advanced machine learning models can underperform, particularly in cases where class imbalance is prevalent (Korkmaz, 2020). The current state of the art largely focuses on improving predictive accuracy through model tuning, but there is a clear gap in the literature regarding the integration of robust preprocessing techniques like SMOTE and their comparative effectiveness when applied to the insurance domain(Bounab, Zarour, Guelib, & Khlifa, 2024).

Existing studies typically rely on traditional methods to handle class imbalance, such as adjusting class weights or undersampling the majority class (Joloudari, Marefat, Nematollahi, Oyelere, & Hussain, 2023). While these approaches have their merits, they often result in a loss of valuable data or fail to produce significant improvements in model performance (Nguyen, Vu, Vo, & Thai, 2021). Few studies have comprehensively evaluated the benefits of combining SMOTE with ensemble models like RandomForest and XGBoost, which can better capture the underlying patterns in highly imbalanced datasets (Shin, Kim, Kim, & others, 2024; Xing et al., 2024). Our research addresses this gap by emphasizing the importance of preprocessing techniques, particularly SMOTE, in enhancing model performance in imbalanced datasets. The goal of our study is to develop a robust predictive model that accurately determines whether health insurance policyholders will be interested in vehicle insurance. We will apply SMOTE to balance the dataset, followed by the implementation of machine learning models such as RandomForest, Logistic Regression, and XGBoost, using advanced cross-validation techniques. By focusing on both the preprocessing and model development stages, we aim to demonstrate that addressing class imbalance early in the pipeline can significantly improve the performance metrics, including precision, recall, F1-score, and AUC-ROC.

The contribution of this research lies in providing an optimized preprocessing strategy that leverages SMOTE in conjunction with well-tuned machine learning models. This approach not only enhances predictive accuracy but also offers a scalable solution for insurance companies to improve their customer targeting efforts in cross-selling scenarios. Our findings will provide valuable insights into the role of preprocessing techniques in predictive modeling, particularly in industries where class imbalance is a persistent challenge. The remainder of this paper is organized as follows: the next section discusses the data collection and preprocessing methods, highlighting the application of SMOTE and its impact on model training. This is followed by a detailed description of the machine learning models and their tuning process. The results section presents the evaluation of the models using key performance metrics, and the discussion explores the implications of our findings for predictive modeling in the insurance industry. Finally, the conclusion summarizes our research and suggests future directions for improving preprocessing techniques in machine learning.

LITERATURE REVIEW

In recent years, predictive modeling in the insurance industry has gained significant attention due to its ability to enhance decision-making processes (Esfandabadi, Ranjbari, & Scagnelli, 2023). By leveraging machine learning techniques, insurance companies aim to improve the accuracy of their customer targeting, cross-selling strategies, and risk assessments (Tian, Todorovic, & Todorovic, 2023). However, one of the persistent challenges in this domain is dealing with class imbalance, where most policyholders may not express interest in a specific product, such as vehicle insurance, while a small proportion may (Hosein, 2024). Various studies have investigated both traditional and advanced machine learning techniques for predicting customer behavior, with an increasing focus on handling imbalanced data through preprocessing techniques (Zaghloul, Barakat, & Rezk, 2024). This literature review provides a critical analysis of current research, focusing on the importance of preprocessing, model selection, and the gaps that remain unaddressed (Antons, Breidbach, Joshi, & Salge, 2023). Machine learning algorithms have been widely employed in the insurance sector for customer behavior prediction. Models such as Logistic Regression, Decision Trees, RandomForest, and Gradient Boosting have shown promising results in predicting policyholder preferences (Groll, Wasserfuhr, & Zeldin, 2024). For instance, (N. K. K. Yego, Nkurunziza, & Kasozi, 2023) demonstrated the effectiveness of RandomForest in predicting the likelihood of health insurance customers purchasing additional coverage using demographic data. The study highlighted the model's ability to handle complex interactions between variables but noted its limitations when applied to imbalanced datasets. Similarly, (Sikri, Jameel, Idrees, & Kaur, 2024) used a Gradient Boosting approach to predict customer churn in the health insurance industry, achieving high accuracy by carefully tuning the model. However, the study did not account for the skewness of the dataset, which could have led to overfitting to the majority class.

A key limitation of these studies is the absence of robust methods to handle class imbalance, which is common in insurance-related datasets. (Ali, Hossain, Kona, Nowrin, & Islam, 2024) explored the use of XGBoost for predicting vehicle insurance uptake, achieving promising results but acknowledging that the imbalanced nature of the dataset (with most policyholders uninterested in vehicle insurance) posed a challenge for the model's accuracy. While (Roy et al., 2021) research focused on tuning hyperparameters for better performance, the study lacked a detailed exploration of preprocessing techniques like SMOTE, which could have significantly improved the model's performance by addressing class imbalance at an earlier stage. Handling imbalanced datasets is a crucial step in machine learning applications, especially in the insurance sector where certain outcomes (e.g., policyholder interest in vehicle insurance) are rare. Class imbalance often leads to models being biased toward the majority class, resulting in poor performance on the minority class, which is typically the class of interest. Several studies

*name of corresponding author

have proposed solutions to this issue. For instance, (Pradipta, Wardoyo, Musdholifah, & Sanjaya, 2021) introduced the Synthetic Minority Over-sampling Technique (SMOTE), which has become a widely adopted method for generating synthetic samples of the minority class to balance the dataset. SMOTE works by interpolating new data points between existing minority class examples, thereby providing the model with more balanced training data.

In the insurance domain, (Loftus, 2023) applied SMOTE to a highly imbalanced dataset of customers purchasing life insurance policies. Their study demonstrated that the use of SMOTE significantly improved the recall and F1-scores of machine learning models such as Logistic Regression and RandomForest. The authors noted that traditional resampling methods such as undersampling the majority class resulted in the loss of valuable information, which could affect model performance. However, while SMOTE improved minority class prediction, the study lacked an in-depth evaluation of how different models, when combined with SMOTE, performed across various performance metrics, such as precision and AUC-ROC. (Kotb & Ming, 2021) compared different methods of handling class imbalance in insurance datasets, including SMOTE, ADASYN (Adaptive Synthetic Sampling), and class weighting. The study found that SMOTE consistently outperformed other methods, particularly when combined with ensemble models like RandomForest and XGBoost. Despite this, the authors pointed out that SMOTE's effectiveness could vary depending on the dataset and that further research is needed to explore its impact in different insurance contexts, such as predicting cross-sell opportunities between health and vehicle insurance. Ensemble models such as RandomForest and XGBoost have gained popularity in insurance prediction due to their ability to improve accuracy by combining multiple weak learners. (Shakhovska, Melnykova, Chopiyak, & others, 2022) introduced RandomForest as a robust method for classification tasks, particularly in noisy and complex datasets like those commonly found in insurance. Research by (Hatwell, Gaber, & Azad, 2020) demonstrated that RandomForest outperformed single models like Decision Trees and Logistic Regression in predicting customer renewal rates in health insurance. However, the study revealed that while RandomForest was effective in general, it struggled with imbalanced datasets where the minority class was underrepresented. XGBoost, another powerful ensemble method, has been shown to perform well in insurance-related tasks. (N. K. Yego, Kasozi, & Nkurunziza, 2021) introduced XGBoost as an efficient implementation of gradient-boosted decision trees, optimized for speed and accuracy. (Obiora, Ali, & Hasan, 2021; Rusdah & Murfi, 2020) applied XGBoost to predict vehicle insurance claims and achieved high accuracy. However, like other models, XGBoost's performance was hindered by class imbalance. Although the study employed class weighting to mitigate this issue, the authors acknowledged that resampling techniques like SMOTE could further enhance the model's performance. Despite the effectiveness of these ensemble models, most studies focus on improving model accuracy through hyperparameter tuning and neglecting the importance of preprocessing steps, such as handling class imbalance. Research that combines advanced ensemble techniques with robust preprocessing methods is still relatively sparse, particularly in the context of predicting cross-sell opportunities in the insurance sector.

The existing literature provides substantial evidence of the benefits of machine learning models in predicting customer behavior within the insurance industry. However, there are several gaps that remain unaddressed. First, while studies have extensively explored the use of models like RandomForest and XGBoost, there is limited research on the impact of preprocessing techniques, such as SMOTE, on model performance in the context of imbalanced datasets. Second, many studies focus on optimizing model accuracy without adequately considering other performance metrics like precision, recall, and F1-score, which are crucial when dealing with imbalanced data. Third, few studies have examined how combining SMOTE with advanced ensemble models can provide a more balanced and accurate prediction for cross-sell opportunities, particularly in the context of health and vehicle insurance. Our research seeks to fill these gaps by focusing on the importance of preprocessing techniques in handling imbalanced data and their impact on predictive model performance. By combining SMOTE with well-established ensemble methods like RandomForest and XGBoost, we aim to provide a comprehensive evaluation of how these techniques can be integrated to improve model outcomes. This approach not only addresses the issue of class imbalance but also explores the effectiveness of various models across multiple performance metrics, offering a more holistic view of customer behavior prediction in the insurance industry.

METHOD

This section outlines the detailed methodology adopted in this research to address the class imbalance problem and optimize model performance in predicting customer interest in vehicle insurance. The methodology consists of several stages: data collection, data preprocessing, model development, and performance evaluation. Each stage is meticulously designed to ensure that the solution not only handles the inherent imbalance in the dataset but also maximizes predictive accuracy across key metrics. This section provides a comprehensive and rigorous explanation of each component, focusing on the mathematical aspects and ensuring that each step is aligned with state-of-the-art machine learning practices.

Dataset Description

The dataset used in this study was provided by an insurance company and contains demographic, vehicle-related, and policy-related information about policyholders and can be download from (Kumar, 2021). Each sample in the dataset represents a customer, and the target variable (response) indicates whether the customer is interested in purchasing vehicle insurance. Formally, let $(D = \{(x_i, y_i)\}_{i=1}^{N})$ represent the dataset, where $(x_i \in$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14194
e- ISSN : 2541-2019
p- ISSN : 2541-044X

R^d) is a (d)-dimensional feature vector for customer (i), including attributes such as age, gender, driving license status, vehicle age, annual premium, and policy sales channel. ($y_i \in \{0,1\}$) is the binary response variable, where ($y_i = 1$) indicates that the customer is interested in vehicle insurance, and ($y_i = 0$) indicates no interest. The dataset contains (N) samples, with (N_0) samples from the majority class (($y = 0$)) and (N_1) samples from the minority class (($y = 1$)), where ($N_0 \gg N_1$), making the dataset highly imbalanced. The features can be grouped as follows: demographics (age, gender, region code), vehicle information (vehicle age, vehicle damage), policy information (annual premium, policy sales channel, driving license, vintage), and response (target), which indicates whether the customer is interested in vehicle insurance.

Data Preprocessing and Handling Class Imbalance

One of the primary challenges in this research is the significant class imbalance in the dataset. Most customers do not show interest in vehicle insurance (($y = 0$)), while only a small proportion express interest (($y = 1$)). This imbalance can lead to biased predictions, where the model favors the majority class, resulting in poor predictive performance for the minority class. Let the imbalance ratio be defined as Imbalance Ratio $= \frac{N_0}{N_1}$ where (N_0) and (N_1) are the numbers of samples in the majority and minority classes, respectively. In typical insurance datasets, this ratio can be extremely high, making it difficult for standard classifiers to capture patterns related to the minority class. To address the class imbalance problem, we apply the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is an advanced oversampling method that generates synthetic samples of the minority class by interpolating between existing minority class samples. This technique avoids simply duplicating samples, which can lead to overfitting, and instead generates new samples that enrich the decision space of the minority class.

Mathematically, the process works as follows. For each sample (x_m) in the minority class, we identify its k-nearest neighbors ($\mathcal{N}_k(x_m)$) from the same class. A new synthetic sample ($x_{\text{synthetic}}$) is generated using the formula $x_{\text{synthetic}} = x_m + \lambda(x_{\text{knn}} - x_m)$ where ($x_{\text{knn}} \in \mathcal{N}_k(x_m)$) is a randomly selected neighbor, and (λ) is a random number between 0 and 1. This ensures that the new sample is a linear interpolation between two minority class samples. The SMOTE algorithm repeats this process for a specified number of iterations, generating new synthetic samples until the dataset reaches a more balanced distribution of classes. After applying SMOTE, the modified dataset, (D'_{train}), contains approximately equal proportions of both classes $N_0 \approx N_1$.

This preprocessing step ensures that the machine learning models trained on this data can learn from both classes more effectively. After applying SMOTE, we perform feature scaling to normalize the continuous features in the dataset, ensuring that they are all on the same scale. Specifically, we use StandardScaler, which transforms each feature (x_j) to have zero mean and unit variance $x'_j = \frac{x_j - \mu_j}{\sigma_j}$ where (μ_j) is the mean of feature (x_j), and (σ_j) is its standard deviation. The dataset is then split into training and test sets, ensuring that the training set is used for model development and the test set is reserved for final evaluation. Stratified splitting is used to maintain the class distribution across both sets.

Machine Learning Models

To predict whether customers are interested in vehicle insurance, we employ three machine learning models: RandomForest, XGBoost, and Logistic Regression. Each model is tailored to handle the class imbalance and is evaluated using cross-validation. RandomForest is an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. The RandomForest model constructs (T) decision trees, each trained on a bootstrap sample of the training data. The final prediction is obtained by averaging the predictions of all trees. Let (T) be the total number of trees in the forest. For each tree (t), the training data is sampled with replacement to create a bootstrap sample ($D^{(t)}$), and a decision tree ($f_t(x)$) is trained on this sample. The overall prediction for a sample (x) is given by $\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$. To handle class imbalance, the RandomForest model applies class weighting, where the weight for each class is inversely proportional to its frequency in the dataset $w_c = \frac{N}{N_c}$ where (N_c) is the number of samples in class (c), and (N) is the total number of samples.

XGBoost is a powerful gradient boosting algorithm that builds an ensemble of decision trees sequentially. Each new tree corrects the errors of the previous trees by minimizing a regularized objective function. The objective function for XGBoost is defined as $\mathcal{L}(\Theta) = \sum_{i=1}^{N} \text{l}(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$ where ($\text{l}(y_i, \hat{y}_i)$) is the loss function (logistic loss for binary classification), and ($\Omega(f_t)$) is a regularization term that penalizes the complexity of the trees. The predictions are updated in each iteration based on the gradient of the loss function. To handle class imbalance, XGBoost uses the parameter $scale\_pos\_weight$, which adjusts the gradient updates for the minority class scale_pos_weight $= \frac{N_0}{N_1}$. This ensures that the minority class samples have a higher impact on the overall model training process.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14194
e- ISSN : 2541-2019
p- ISSN : 2541-044X

Logistic Regression is a simple yet effective model for binary classification, where the probability of a positive outcome is modeled using the logistic function $P(y = 1|x) = \frac{1}{1+e^{-\beta^T x}}$, where (β) is the vector of coefficients. The model is trained by maximizing the weighted log-likelihood function, which is adjusted for class imbalance $\mathcal{L}_{\text{wihe}} = \sum_{i=1}^{N} w_i \log P(y_i|x_i)$, where (w_i) represents the class weight for sample (i).

Cross-Validation Strategy

To ensure robust evaluation and mitigate overfitting, we employ Stratified K-Fold Cross-Validation with ($k = 5$) folds. In each fold, the dataset is split into ($k - 1$) training folds and 1 validation fold, ensuring that the class distribution is preserved in each fold. The model is trained on the training folds and evaluated on the validation fold, with the process repeated for all folds. Formally, let $(F = \{(D_{\text{train}}^{(1)}, D_{\text{val}}^{(1)}), \dots, (D_{\text{train}}^{(k)}, D_{\text{val}}^{(k)})\})$ represent the stratified folds. For each fold (i), we compute the validation performance metrics, such as AUC-ROC, precision, recall, and F1-score. The final performance is obtained by averaging the metrics across all folds Average Score $= \frac{1}{k}\sum_{i=1}^{k} \text{score}^{(i)}$. This cross-validation strategy ensures that the models are evaluated across multiple subsets of the data, reducing the risk of biased estimates and overfitting. This comprehensive methodology integrates advanced preprocessing techniques like SMOTE with well-tuned machine learning models and employs rigorous cross-validation. Each stage of the process is designed to optimize model performance, particularly in the context of class imbalance, ensuring that the solution is robust, scalable, and capable of delivering accurate predictions.

RESULT

This section presents the results obtained from the application of the XGBoost, RandomForest, and Logistic Regression models as presented in the table 1. These models were evaluated using five-fold cross-validation, and the performance metrics assessed include AUC-ROC, precision, recall, and F1-score. The results are summarized for each fold, followed by the overall performance averaged across all folds. The XGBoost model's performance metrics are summarized across five folds. The average AUC-ROC score was 0.7854 with a standard deviation of 0.0011, indicating that the model maintained consistent performance across all validation folds. The model demonstrated excellent recall, with an average value of 0.9525, showing that XGBoost was able to correctly identify most of the positive cases, specifically customers interested in vehicle insurance. However, the precision was relatively low at 0.2585, indicating that many of the predicted positives were false positives. This trade-off between precision and recall resulted in an overall F1-score of 0.4066, which strikes a balance between precision and recall.

Looking at the individual folds, the AUC-ROC values ranged from 0.7840 to 0.7874, demonstrating consistent model performance. The precision values across folds ranged from 0.2576 to 0.2594, while recall values ranged from 0.9507 to 0.9560. The F1-scores were similarly stable, ranging from 0.4054 to 0.4081. These results suggest that XGBoost is highly effective at recalling positive cases but may predict more false positives than desired. The RandomForest model showed slightly lower performance compared to XGBoost, with an overall AUC-ROC of 0.6826 and a standard deviation of 0.0020. The model's precision was higher than XGBoost, averaging at 0.3064, which indicates a lower rate of false positives. However, this improvement in precision came at the cost of lower recall, which averaged at 0.5341, showing that RandomForest missed many positive cases. The overall F1-score of 0.3894 reflects this trade-off, indicating a more balanced but less accurate model compared to XGBoost.

Analyzing the individual fold results for RandomForest, the AUC-ROC values ranged from 0.6804 to 0.6864, demonstrating consistency. Precision ranged from 0.3042 to 0.3092, while recall values varied more widely, ranging from 0.5303 to 0.5420. This variability in recall suggests that the RandomForest model struggled to consistently identify positive cases, but its precision remained stable. The F1-scores ranged from 0.3866 to 0.3937, slightly lower than XGBoost, but more balanced between precision and recall. The performance of the Logistic Regression model was the most variable among the three models. The overall AUC-ROC score was 0.6965, but with a relatively high standard deviation of 0.0780, indicating inconsistency across folds. The precision for Logistic Regression was the lowest among the models, averaging 0.2372, which suggests a higher number of false positives. The recall, however, was moderate, averaging 0.7124, meaning that the model was somewhat effective in identifying positive cases, though not as consistently as XGBoost. The overall F1-score was 0.3557, reflecting the challenges in balancing precision and recall.

In the individual folds, the AUC-ROC scores for Logistic Regression ranged from 0.5921 to 0.7628, highlighting the model's inconsistency. Precision values ranged from 0.1728 to 0.2747, while recall values fluctuated significantly, from 0.5163 to 0.8354. The wide range in F1-scores, from 0.2637 to 0.4131, further illustrates the variability of Logistic Regression's performance. These results indicate that while Logistic Regression performed reasonably well in some cases, it struggled to maintain consistent performance, especially when it came to precision. Across all models, XGBoost demonstrated the highest recall, making it the most effective model for identifying customers interested in vehicle insurance. However, its low precision indicates a higher number of false positives. RandomForest, on the other hand, offered a more balanced trade-off between precision and recall, although its overall performance was lower than XGBoost in terms of AUC-ROC. Logistic

*name of corresponding author

e- ISSN : 2541-2019
p- ISSN : 2541-044X

Regression, while occasionally performing well in terms of recall, exhibited the highest variability and the lowest precision, leading to overall lower performance compared to the other models.

DISCUSSIONS

The results of this study provide insights into the performance of three machine learning models: XGBoost, RandomForest, and Logistic Regression applied to the task of predicting customer interest in vehicle insurance as presented in the table 1. The evaluation metrics, particularly AUC-ROC, precision, recall, and F1-score, reveal key trade-offs in the performance of each model and highlight important considerations for their practical use in real-world applications. The XGBoost model consistently outperformed the other models in terms of AUC-ROC and recall. An AUC-ROC score of 0.7854 indicates that XGBoost is highly capable of distinguishing between customers who are interested in vehicle insurance and those who are not. The model's high recall of 0.9525 further demonstrates its effectiveness at capturing most of the positive cases, i.e., customers who are interested in vehicle insurance. This makes XGBoost a valuable tool in situations where the primary concern is ensuring that the majority of interested customers are identified. However, XGBoost's low precision of 0.2585 points to a high number of false positives, where the model incorrectly classifies customers as being interested in vehicle insurance. This trade-off between high recall and low precision must be considered when deploying XGBoost in practice. In contexts where false positives are costly, such as personalized marketing campaigns, the model's low precision may lead to inefficient use of resources by targeting customers who are not genuinely interested.

The RandomForest model, on the other hand, offered a more balanced performance between precision and recall. Its AUC-ROC of 0.6826, though lower than that of XGBoost, still indicates moderate classification ability. RandomForest's precision was higher than XGBoost's at 0.3064, suggesting that it produced fewer false positives. However, this improvement in precision came at the cost of recall, which was significantly lower at 0.5341. This means that while RandomForest is more conservative in predicting positive cases, it misses many customers who are actually interested in vehicle insurance. Therefore, RandomForest might be more suitable in scenarios where the goal is to minimize false positives, such as when offering personalized discounts or incentives that carry a significant cost. However, its lower recall might make it less effective in ensuring broad customer coverage.

The Logistic Regression model exhibited the most variable performance among the three models, with an average AUC-ROC of 0.6965 and wide fluctuations in both precision and recall across the five folds. Although Logistic Regression achieved a moderate recall of 0.7124, its precision was the lowest at 0.2372, indicating a high rate of false positives. This inconsistency in performance suggests that Logistic Regression may not be well-suited for this classification task, particularly in situations where stable and reliable predictions are critical. The variability in the model's metrics, particularly the large standard deviation in AUC-ROC and the wide range of precision and recall values, indicates that Logistic Regression struggled to generalize effectively across different folds of the dataset. This inconsistency can be problematic in practice, as it implies that the model's predictions may be unreliable when applied to new data.

Table 1. Comparison of Model Performance

Model	Average AUC-ROC	Precision	Recall	F1-Score
XGBoost	0.7854 ± 0.0011	0.2585 ± 0.0006	0.9525 ± 0.0019	0.4066 ± 0.0009
RandomForest	0.6826 ± 0.0020	0.3064 ± 0.0018	0.5341 ± 0.0044	0.3894 ± 0.0024
Logistic Regression	0.6965 ± 0.0780	0.2372 ± 0.0437	0.7124 ± 0.1445	0.3557 ± 0.0670

When comparing the three models, it becomes clear that each has its strengths and weaknesses. XGBoost's strength lies in its ability to capture the majority of positive cases, making it ideal for applications where recall is prioritized, such as in early detection or exploratory marketing strategies where missing potential leads is more detrimental than targeting uninterested customers. However, the model's low precision must be addressed, possibly by adjusting the decision threshold or combining XGBoost with another model to improve precision. RandomForest, with its more balanced precision and recall, may be better suited to applications where the cost of false positives is higher, such as in targeted campaigns or resource-intensive outreach programs. The main drawback of RandomForest is its lower recall, which means it might miss valuable opportunities to engage interested customers. Logistic Regression, while straightforward and interpretable, does not perform consistently enough for this task and would likely require further tuning or modification, such as incorporating regularization or feature engineering, to improve its stability and overall effectiveness. In practice, the choice of model should be guided by the specific business objectives and the trade-offs that stakeholders are willing to make between recall and precision. For instance, if the goal is to cast a wide network and identify as many interested customers as possible, XGBoost would be the preferred model. However, if the objective is to reduce the number of false positives and avoid wasted resources, RandomForest might be more appropriate. In either case, improving precision through post-processing techniques, such as adjusting the decision threshold, recalibrating probabilities, or implementing an ensemble of models, could further enhance the performance of the chosen approach. Overall,

*name of corresponding author

the results of this study demonstrate that machine learning models can effectively support decision-making in the context of predicting customer interest in vehicle insurance. The specific model choice should align with the business priorities, whether that involves maximizing recall to identify as many interested customers as possible or balancing precision and recall optimizing resource allocation.

CONCLUSION

This study evaluated the performance of three machine learning models: XGBoost, RandomForest, and Logistic Regression in predicting customer interest in vehicle insurance using a dataset that included demographic, vehicle, and policy-related information. The models were assessed using five-fold cross-validation, with performance metrics including AUC-ROC, precision, recall, and F1-score. Each model exhibited strengths and weaknesses, and the results revealed important trade-offs between precision and recall. XGBoost emerged as the top-performing model in terms of recall and AUC-ROC, achieving a high recall of 0.9525 and an AUC-ROC of 0.7854. This demonstrates that XGBoost is highly effective at identifying customers who are interested in vehicle insurance. However, its low precision of 0.2585 indicates a high rate of false positives, which could lead to inefficiencies in practical applications where minimizing false positives is important. RandomForest, while slightly less effective in terms of overall classification ability with an AUC-ROC of 0.6826, provided a more balanced trade-off between precision (0.3064) and recall (0.5341). This suggests that RandomForest is more conservative in predicting positive cases and may be better suited for applications where the cost of false positives is a key concern. Logistic Regression, despite being the simplest model, showed the most variability in performance and struggled to maintain consistency across different folds of the data. Its low precision and moderate recall made it the least effective of the three models, suggesting that it may not be the most appropriate choice for this classification task without further tuning or modification. In conclusion, the choice of the best model for predicting customer interest in vehicle insurance depends on the specific business goals and the desired balance between precision and recall. XGBoost is ideal for situations where identifying as many interested customers as possible is the priority, even at the expense of some false positives. RandomForest offers a more balanced approach, with fewer false positives but at the cost of missing some interested customers. Logistic Regression, while interpretable, requires improvement to be competitive with the other models. Further refinements, such as hyperparameter tuning, threshold adjustments, or ensemble methods, could enhance model performance and address the trade-offs observed in this study.

REFERENCES

Ali, M. S., Hossain, M. M., Kona, M. A., Nowrin, K. R., & Islam, M. K. (2024). An ensemble classification approach for cervical cancer prediction using behavioral risk factors. *Healthcare Analytics*, *5*, 100324.

Antons, D., Breidbach, C. F., Joshi, A. M., & Salge, T. O. (2023). Computational literature reviews: Method, algorithms, and roadmap. *Organizational Research Methods*, *26*(1), 107–138.

Bounab, R., Zarour, K., Guelib, B., & Khlifa, N. (2024). Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN. *IEEE Access*.

Chaffey, D., & Smith, P. R. (2022). *Digital marketing excellence: planning, optimizing and integrating online marketing*. Routledge.

Chen, Q., Zhang, Z.-L., Huang, W.-P., Wu, J., & Luo, X.-G. (2022). PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets. *Neurocomputing*, *498*, 75–88.

Dangut, M. D. (2021). *Application of data analytics for predictive maintenance in aerospace: an approach to imbalanced learning.*

Datta, S., Ghosh, C., & Choudhury, J. P. (2024). Classification of imbalanced datasets utilizing the synthetic minority oversampling method in conjunction with several machine learning techniques. *Iran Journal of Computer Science*, 1–18.

Esfandabadi, Z. S., Ranjbari, M., & Scagnelli, S. D. (2023). Prioritizing risk-level factors in comprehensive automobile insurance management: A hybrid multi-criteria decision-making Model. *Global Business Review*, *24*(5), 972–989.

Groll, A., Wasserfuhr, C., & Zeldin, L. (2024). Churn Modeling of Life Insurance Policies Via Statistical and Machine Learning Methods. *Journal of Insurance Issues*, *47*(1), 78–117.

Gupta, S., & Joshi, S. (2022). Predictive analytic techniques for enhancing marketing performance and personalized customer experience. *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, 16–22.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, *9*(2), 42.

Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. *Artificial Intelligence Review*, *53*, 5747–5788.

Hosein, P. (2024). A data science approach to risk assessment for automobile insurance policies. *International Journal of Data Science and Analytics*, *17*(1), 127–138.

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Applied Sciences*, *13*(6), 4006.

Kaswan, K. S., Dhatterwal, J. S., Sharma, H., & Sood, K. (2022). Big data in insurance innovation. *Big Data: A*

*name of corresponding author

Game Changer for Insurance Industry, 117–136.

Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications*, *4*, 100024.

Korkmaz, S. (2020). Deep learning-based imbalanced data classification for drug discovery. *Journal of Chemical Information and Modeling*, *60*(9), 4180–4190.

Kotb, M. H., & Ming, R. (2021). Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. *International Journal of Advanced Computer Science and Applications*, *12*(9).

Kumar, A. (2021). *Health Insurance Cross Sell Prediction Dataset*. Retrieved from https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction/data

Loftus, J. (2023). *An assessment of the effectiveness of using data analytics to predict death claim seasonality and protection policy review lapses in a life insurance company*.

Nguyen, H., Vu, T., Vo, T. P., & Thai, H.-T. (2021). Efficient machine learning models for prediction of concrete strengths. *Construction and Building Materials*, *266*, 120950.

Obiora, C. N., Ali, A., & Hasan, A. N. (2021). Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance. *2021 IEEE PES/IAS PowerAfrica*, 1–5.

Pradipta, G. A., Wardoyo, R., Musdholifah, A., & Sanjaya, I. N. H. (2021). Radius-SMOTE: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. *IEEE Access*, *9*, 74763–74777.

Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., … Ali, I. (2021). An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. *Complexity*, *2021*(1), 9953314.

Rusdah, D. A., & Murfi, H. (2020). XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*(8), 1336.

Shakhovska, N., Melnykova, N., Chopiyak, V., & others. (2022). An Ensemble Methods for Medical Insurance Costs Prediction Task. *Computers, Materials \& Continua*, *70*(2).

Shin, Y., Kim, M., Kim, H., & others. (2024). Towards unbalanced multiclass intrusion detection with hybrid sampling methods and ensemble classification. *Applied Soft Computing*, *157*, 111517.

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, *14*(1), 13097.

Singhal, N., Goyal, S., & Singhal, T. (2024). *Potential, Risks, and Ethical Implications of Decentralized Insurance*. Springer.

Śmietanka, M., Koshiyama, A., & Treleaven, P. (2021). Algorithms in future insurance markets. *International Journal of Data Science and Big Data Analytics*, *1*(1), 1–19.

Staudt, Y., & Wagner, J. (2022). Factors Driving Duration to Cross-Selling in Non-Life Insurance: New Empirical Evidence from Switzerland. *Risks*, *10*(10), 187.

Tian, X., Todorovic, J., & Todorovic, Z. (2023). A Machine-Learning-Based Business Analytical System for Insurance Customer Relationship Management and Cross-Selling. *Journal of Applied Business \& Economics*, *25*(6).

Tondi, M. (2024). *THE RECONFIGURATION OF CUSTOMER VALUE PROPOSITION IN THE INSURANCE INDUSTRY*.

Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing \& Management*, *59*(6), 103085.

Washington, A. L. (2023). *Ethical Data Science: Prediction in the Public Interest*. Oxford University Press.

Xing, Q., Yu, C., Huang, S., Zheng, Q., Mu, X., & Sun, M. (2024). Enhanced Credit Score Prediction Using Ensemble Deep Learning Model. *ArXiv Preprint ArXiv:2410.00256*.

Yego, N. K. K., Nkurunziza, J., & Kasozi, J. (2023). Predicting health insurance uptake in Kenya using Random Forest: An analysis of socio-economic and demographic factors. *Plos One*, *18*(11), e0294166.

Yego, N. K., Kasozi, J., & Nkurunziza, J. (2021). A comparative analysis of machine learning models for the prediction of insurance uptake in kenya. *Data*, *6*(11), 116.

Zaghloul, M., Barakat, S., & Rezk, A. (2024). Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. *Journal of Retailing and Consumer Services*, *79*, 103865.

*name of corresponding author