# Identifying the Optimal Machine Learning Model for Predicting Car Insurance Claims: A Comparative Study Utilising Advanced Techniques

**Xiaonan Li***

*Central University of Finance and Economics, Beijing, China*
**Corresponding author: lxn_cufejs@126.com*

***Abstract:*** *This study presents an investigation into the use of machine learning techniques for predicting car insurance claims, with a specific focus on identifying the optimal model for this task. By utilising advanced techniques such as SMOTEEN, ANOVA, and Chi-squared tests, the challenge of processing imbalanced data and identifying relevant features were addressed. Our evaluation of five popular and effective models, including logistic regression (LR), random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), and extreme gradient boosting (XGBoost), yields result that demonstrate the superiority of the RF model in predicting car insurance claims. Furthermore, our study illustrates the advantages of using machine learning algorithms in handling large and complex datasets, making predictions on future insurance claims, and adapting to changing circumstances, making it a valuable tool for practitioners in the insurance industry.*

***Keywords:*** *machine learning, car insurance claims, LR, RF, SVM, MLP, XGBoost*

## 1. Introduction

Predicting the likelihood of car insurance claims is a crucial task for both insurance companies and policyholders [1-3]. Accurate predictions allow insurance companies to set premiums at a level that reflects the risk of insuring a particular individual or vehicle, while also helping them to remain profitable by covering the cost of any claims that are made [4-7].

Traditional methods, such as manual analysis or rule-based systems, may not be as effective at handling large, complex datasets as machine learning algorithms [5, 8-11]. In contrast, machine learning algorithms can automatically learn from data and adapt to changing circumstances, making them more suitable for handling large and complex datasets. Another reason is that machine learning algorithms can be trained on historical data and then applied to new, unseen data, making them suitable for making predictions on future insurance claims [12-14]. However, predicting insurance claims can present several challenges. One challenge that may arise is the presence of imbalanced data, where there is a disproportionate number of instances where a claim was not filed compared to instances where a claim was filed [15-17]. This can make it difficult for a machine learning model to accurately identify and classify instances of the minority class (i.e., instances where a claim was filed) [15, 18-20]. Besides, insurance claims data can also be complex, featuring a large number of features such as policyholder characteristics, policy details, and claims history [21-23]. This can make it difficult for a machine learning model to identify the most relevant features and relationships in the data. Additionally, insurance claims data may be noisy, with errors or inconsistencies that can impact the model's ability to learn and make accurate predictions.

In this study, we aimed to identify the optimal machine learning model for predicting car insurance claims. To tackle the challenges in processing data, techniques such as SMOTEEN, ANOVA, and Chi-squared ($\chi^2$) test were utilised [24-27]. Five popular and effective models were evaluated: logistic regression (LR), random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), and extreme gradient boosting (XGBoost). These models were chosen due to their popularity and effectiveness in various machine learning tasks [13, 28-30]. By exploring multiple models, we were able to identify the one that performed best on the dataset. This information can be useful for practitioners in the insurance industry, who seek to develop accurate and reliable prediction models for car insurance claims.

## 2. Research methodology

In this study, a machine learning model was developed using the Python programming language to predict car insurance claims. The data used in the study was obtained from CarIns, a start-up insurance company, and consisted of 43 features and 58,593 rows. Each row represented information about a single policy holder and his/her car. The features included both numerical and categorical variables, some of which were binary in nature. These included numerical values such as the age of the car and policy tenure, as well as binary indicators for features such as parking sensors or cameras. Categorical variables such as transmission type, model type, and area cluster codfige were also included. The methodology outlined in this paper is depicted in Figure 1.

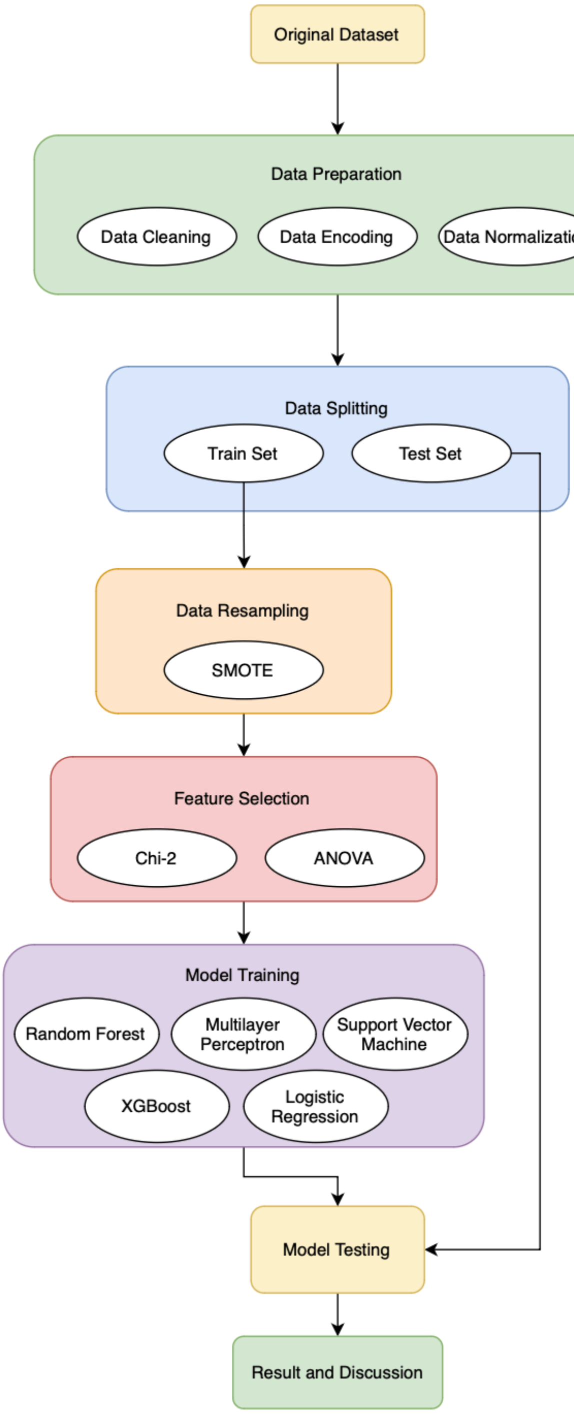

*Figure 1: Flowchart of methodology used in developing machine learning models for car insurance claims prediction.*

### *2.1. Data collection*

The numerical values were normalised, while categorical and binary variables were encoded. The dataset was found to be clean and consistent, making it suitable for pre-processing steps such as the feature selection and encoding. The target variable for the analysis was binary, indicating whether a claim had occurred (1) or not (0). As shown in Figure 2, the dataset was found to be highly imbalanced with a disproportionate number of observations belonging to the majority class (i.e., no claim) compared to the minority class (i.e., claim).

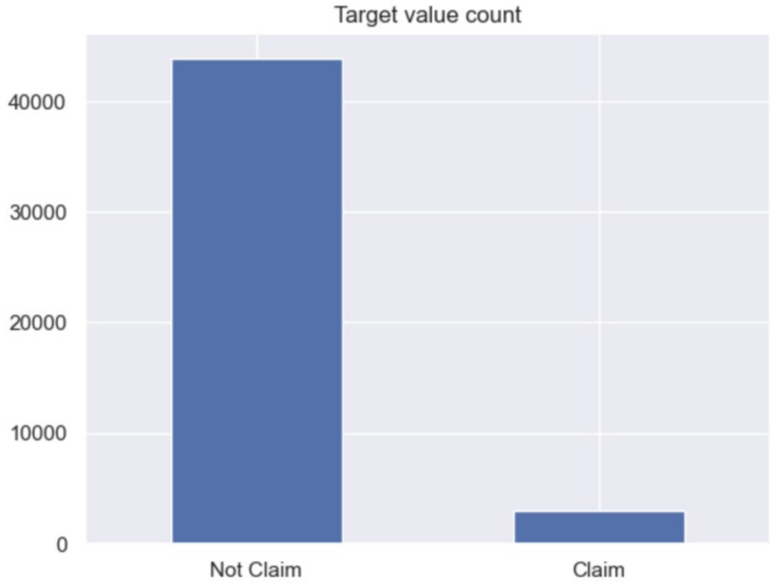

*Figure 2: Imbalance in the target variable distribution.*

### *2.2. Data pre-processing*

Evaluating the performance of a predictive model for insurance claims requires the accurate encoding of categorical features into numeric values, utilising techniques such as Label or Ordinal Encoding. The dataset is then divided into training and testing sets using the StratifiedKFold method, with a common split ratio of 80/20 for training and testing, respectively. In order to address the issue of imbalanced training data, the Synthetic Minority Oversampling Technique (SMOTE) was employed to balance the dataset and ensure a more equitable distribution of minority class instances. SMOTE selects a minority class sample and finds its k nearest neighbours, and then generates new samples along the line connecting the sample and one of its neighbours, by a random amount proportional to the distance between them. The final distribution after sampling is shown in Figure 3.

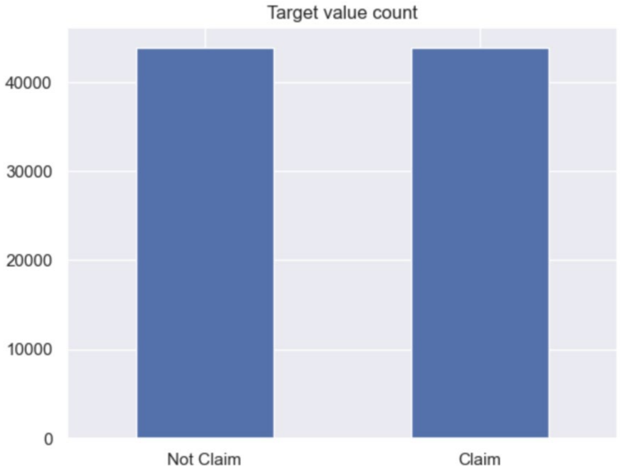

*Figure 3: Distribution of the dataset after applying the Synthetic Minority Oversampling Technique (SMOTE).*

To optimise the performance of our model and mitigate overfitting, a feature selection strategy was utilised prior to training the data. Utilising ANOVA, the most pertinent numeric features were identified by retaining only those with a score greater than 20. Similarly, we employed chi2 to select the most salient categorical features, retaining only those with a score greater than 35. The final set of selected features included: 'policy_tenure', 'age_of_car', 'age_of_policyholder', 'population_density', 'cylinder', 'width', 'max_torque', 'area_cluster', 'model', 'steering_type', and 'max_power'. After the feature selection, a standard scaler was applied to transform the data to have a zero mean and unit variance. The transformed dataset was then ready for training and evaluation.

### *2.3. Training and evaluating LR, RF, SVM, MLP, and XGBoost*

In this study, five different models were employed to predict car insurance claims. One of these models was LR, which is a widely used statistical method for analysing datasets in which one or more independent variables determine an outcome. LR is like linear regression, with the main distinction being that it makes predictions by computing the probability of the instance being predicted as positive, represented as Equation (1):

$$P = \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}} \tag{1}$$

where $\mathrm{P}$ is the predicted probability of the instance, it is positive, $\mathrm{e}$ is the mathematical constant, $\mathrm{b}_0$ is the intercept term, $\mathrm{b}_1$ is the coefficients, and $\mathrm{X}$ is the feature values vector.

The linear function of $\mathrm{X}$ activated by sigmoid function. The sigmoid function is used to scale P to [0, 1]. Accordingly, the algorithm makes predictions based on the P value and the threshold value, as illustrated in Equation (2):

$$\hat{y} = \begin{cases} 1 & if\ P \geq t \\ 0 & if\ P < t \end{cases} \tag{2}$$

where $\hat{y}$ is the predicted class, 1 represents the instance is predicted as positive (claim will occur), 0 represents the instance is predicted as negative (claim will not occur), and $t$ is the threshold.

It specifically calculates the probability that the dependent variable, a claim or no claim, is equal to one (i.e., a claim is made) given the values of the independent variables. This probability is then used to make a prediction about whether a claim will be made, typically with a threshold value of 0.5. One of the key advantages of LR is its simplicity and effectiveness in making predictions when the outcome is binary, as is the case with car insurance claims. Additionally, the model parameters can be easily interpreted in terms of odds ratio, providing valuable insight into how different features affect the outcome.

Another model that was employed in this study is RF, which is a robust algorithm for imbalanced data. It combines multiple decision trees to make predictions by taking a vote of the decision trees' predictions. A decision tree is a type of non-parametric supervised learning method that creates a tree-like structure to represent the relationship between features and labels. Each internal node, branch, and leaf represents a feature, level of feature, and class label, respectively. The main idea of a decision tree is to take all features as input and divide the data into subsets, like a tree structure, until all instances are classified perfectly, or a termination criterion is met. The selection and level of the feature at each internal node is based on a measurement such as minimising the Gini impurity or maximising information entropy. However, RF has been observed to be susceptible to over-fitting. To mitigate this issue, techniques such as limiting the maximum depth of each tree and limiting the minimum number of samples required to generate each node were employed in this study. These techniques were used to ensure that the model generalises well to unseen data and avoid overfitting.

Having examined the techniques employed to counter overfitting in RF models, another robust machine learning algorithm is SVM. As a supervised learning algorithm, an appropriate kernel function and regularization parameter were carefully selected to train the data via the SVM model. Once the training process was completed, the performance of the model was evaluated using various metrics. In order to predict the validity of an insurance claim, the features of the claim, such as policyholder information and incident details, were input into the SVM model, which then produced a prediction of the claim's validity.

The Multilayer Perceptron, also known as a feedforward neural network, is a powerful machine learning algorithm that can be used for both classification and regression tasks. By utilising optimisation algorithms such as stochastic gradient descent, the weights of the neurons can be adjusted to improve the model's performance. Its ability to learn complex non-linear relationships between the input features and the output makes it particularly suitable for tasks such as the insurance claim prediction. Furthermore, regularization techniques, such as dropout, can be implemented to counter overfitting and enhance the generalisation performance.

In addition to evaluating the capabilities of the Multilayer Perceptron algorithm for insurance claim prediction, this study also conducts a comprehensive examination of another widely adopted machine learning algorithm, XGBoost. XGBoost, an open-source implementation of gradient boosting, is a powerful ensemble technique that can be used for both regression and classification problems. One of the key advantages of XGBoost is its ability to automatically learn the optimal interactions between features and handle missing data, which makes it particularly suitable for insurance claim prediction tasks. Additionally, the use of regularization techniques in XGBoost, such as controlling overfitting, allows for the algorithm to generalise well to unseen data.

### *2.4. Measurements*

Accuracy is a widely used metric for assessing the performance of classification models, however, it is not appropriate for evaluating models trained on imbalanced datasets. In such cases, metrics such as precision, recall, Area under the Curve of Receiver Operating Characteristic (AUC-ROC), and F1-score provide a more comprehensive evaluation of the model's performance. Additionally, plots such as specificity and sensitivity plots are often generated to further assess the model's performance on imbalanced data. Furthermore, specificity in terms of the sensitivity of all models are plotted respectively. In the context of car insurance occurrence prediction, a confusion matrix can be used to evaluate the performance of a classification model. The matrix contains four entries, representing the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions made by the classifier.

● True Positives (TP) represent the number of instances where the classifier correctly predicted that a claim would occur.

● False Positives (FP) represent the number of instances where the classifier incorrectly predicted that a claim would occur, when in fact, it did not.

● True Negatives (TN) represent the number of instances where the classifier correctly predicted that a claim will not occur.

● False Negatives (FN) represent the number of instances where the classifier incorrectly predicted that a claim will not occur, when in fact, it did.

Accuracy, precision, recall, specificity, and F1-score can be computed by the entries in the matrix, with the formula listed in Table 1.

*Table 1: Metrics and computing method.*

| Metrics | Computing method |
|---|---|
| **Accuracy** | $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ |
| **Precision** | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ |
| **Recall** | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ |
| **Specificity** | $\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$ |
| **F1-score** | $\frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$ |

## 3. Results and discussion

The confusion matrix is a valuable tool for assessing the performance of binary classifiers. Table 2 presents the confusion matrix for several models (LR, RF, SVM, MLP, XGBoost). Each cell in the matrix displays the number of predictions made by the classifier for a specific class (negative or positive). The diagonal elements of the matrix represent the number of TP and TN. The off-diagonal elements represent the number of FP and false negatives FN.

*Table 2: Confusion matrix for binary classifier models.*

| Models | | Predicted Negative | Predicted Positive |
|---|---|---|---|
| LR | Truth Negative | 6547 | 4416 |
| | Truth Positive | 345 | 411 |
| RF | Truth Negative | 6207 | 4756 |
| | Truth Positive | 266 | 490 |
| SVM | Truth Negative | 5769 | 5194 |
| | Truth Positive | 255 | 501 |
| MLP | Truth Negative | 8775 | 2188 |
| | Truth Positive | 549 | 207 |
| XGBoost | Truth Negative | 8616 | 2347 |
| | Truth Positive | 478 | 278 |

In classification tasks, the models are typically evaluated based on their accuracy for each class according to the computing method in Table 1. However, when the data is imbalanced, the focus shifts to the accuracy of the model for the minority class, as this is a more challenging problem. To this end, Table 3 presents the precision, recall, F1 score and ROC-AUC of the models, specifically for the minority class predictions.

*Table 3: Evaluation of Minority Class Predictions: Precision, Recall, F1-Score and ROC-AUC.*

| Model | Precision | Recall | Specificity | F1-Score | Accuracy | AUC-ROC |
|---|---|---|---|---|---|---|
| LR | 9.14% | 53.97% | 59.72% | 14.72% | 59.37% | 56.84% |
| RF | 9.34% | 64.81% | 56.62% | 16.32% | 57.15% | 60.72% |
| SVM | 8.80% | 66.27% | 52.62% | 15.53% | 53.50% | 59.45% |
| MLP | 8.64% | 27.38% | 80.04% | 13.13% | 76.64% | 57.17% |
| XGBoost | 10.59% | 36.77% | 78.59% | 16.44% | 75.89% | 57.68% |

As can be seen in Table 3, the XGBoost model boasts the highest precision score, indicating a high proportion of true positive predictions.

However, it also has a relatively low recall, signifying a significant number of false negative instances. On the other hand, the RF and SVM models exhibit the highest recall, correctly identifying many claims that will occur. However, their precision is relatively low, indicating a high number of false positive predictions.

When it comes to Specificity, the MLP model has the highest Specificity, meaning it correctly identifies the greatest number of claims that will not occur, however, its recall is relatively low, indicating that it would predict many positive cases incorrectly.

The F1-score is a metric that combines precision and recall, and the RF and SVM models have the highest F1-scores, meaning they have a balance of high precision and recall. Multilayer Perceptron has the lowest F1-score, which indicates that it has the worst ability to balance the trade-off between precision and recall.

The XGBoost model achieves the highest accuracy, but it's worth noting that this is due in large part to its high number of TN as per the confusion matrix. While the accuracy of the SVM model is the lowest among all models, it should be noted that its recall is the highest. The lower accuracy is likely because the number of negative instances in the sample is much higher than the number of positive instances, and the model incorrectly predicted many negative samples as positive. However, it's important to note that accuracy may not be a highly informative or scientifically important metric for a model trained on an imbalanced dataset.

The AUC-ROC is a metric that measures the performance of a binary classifier. Figure 4 shows the AUC-ROC curve obtained using the five models on both the training and testing datasets. As shown is this figure, the Random Forest (RF) model has the highest AUC-ROC, indicating that it has the highest overall performance. The accuracy of the models ranges from 53.50% to 76.64%.

In summary, it is important to consider the specific business constraints and the cost associated with false positives and false negatives while selecting the best model. The RF and SVM models may be the best choice for this specific task, as they have a balance of high precision and recall, but considering the overall performance, the RF model has the highest AUC-ROC.

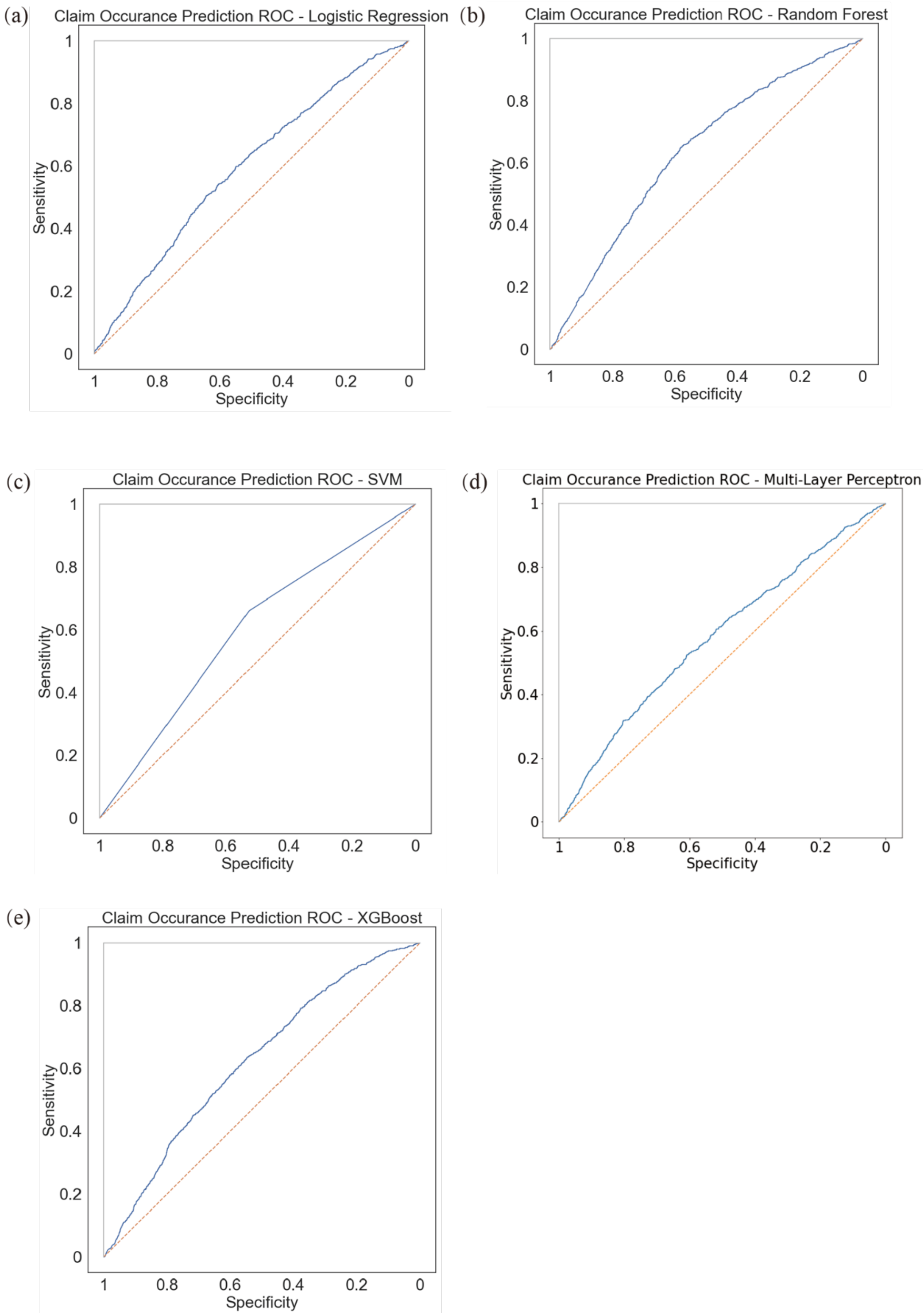

*Figure 4: Comparison of AUC-ROC Performance of Five Models on Training and Testing Datasets: (a) LG, (b) RF, (c) SVM, (d) LP, and (e) XGBoost.*

## 4. Conclusion

This study aimed to identify the optimal machine learning model for predicting car insurance claims. Here are the conclusions:

1) The challenges of processing data such as handling imbalanced data and identifying the most relevant features can be addressed using techniques such as SMOTEEN, ANOVA, and Chi-squared tests.

2) By evaluating five popular and effective models, we found that the Random Forest (RF) model performed the best on the dataset. This information can be useful for practitioners in the insurance industry, who seek to develop accurate and reliable prediction models for car insurance claims.

3) Machine learning algorithms can automatically learn from data, adapt to changing circumstances and can be trained on historical data making them suitable for making predictions on future insurance claims.

4) The use of machine learning models can be an effective solution for predicting car insurance claims and these models can provide valuable insights to insurance companies and policyholders.

## References

*[1] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," Int. J. Adv. Soft Comput. Appl, vol. 10, no. 2, pp. 159-171, 2018.*
*[2] D. Huangfu, "Data Mining for Car Insurance Claims Prediction," WORCESTER POLYTECHNIC INSTITUTE, 2015.*
*[3] S. Matthews and B. Hartman, "Machine Learning in Ratemaking, an Application in Commercial Auto Insurance," Risks, vol. 10, no. 4, p. 80, 2022.*
*[4] K. Weerasinghe and M. Wijegunasekara, "A comparative study of data mining algorithms in the prediction of auto insurance claims," European International Journal of Science and Technology, vol. 5, no. 1, pp. 47-54, 2016.*
*[5] P. Hanafizadeh and N. R. Paydar, "A data mining model for risk assessment and customer segmentation in the insurance industry," International Journal of Strategic Decision Sciences (IJSDS), vol. 4, no. 1, pp. 52-78, 2013.*
*[6] T. Seo, K. H. Park, and H. Chung, "SOCAR: Socially-Obtained CAR Dataset for Image Recognition in the Wild," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 430-438.*
*[7] W. Breuer, A. Haake, M. Hass, and E. Sachsenhausen, "Silence is Silver, Speech is Gold: The Benefits of Machine Learning and Text Analysis in the Financial Sector," in The Monetization of Technical Data: Springer, 2023, pp. 69-86.*
*[8] K. A. Smith, R. J. Willis, and M. Brooks, "An analysis of customer retention and insurance claim patterns using data mining: A case study," Journal of the operational research society, vol. 51, no. 5, pp. 532-541, 2000.*
*[9] J. S. Kong et al., "Machine learning-based injury severity prediction of level 1 trauma center enrolled patients associated with car-to-car crashes in Korea," Computers in biology and medicine, vol. 153, p. 106393, 2023.*
*[10] D. Agarwal and K. Tripathi, "A Framework for Structural Damage detection system in automobiles for flexible Insurance claim using IOT and Machine Learning," in 2022 International Mobile and Embedded Technology Conference (MECON), 2022: IEEE, pp. 5-8.*
*[11] M. Hanafy and R. Ming, "Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study," Applied Artificial Intelligence, pp. 1-32, 2022.*
*[12] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," Ieee Access, vol. 5, pp. 7776-7797, 2017.*
*[13] O. Stucki, "Predicting the customer churn with machine learning methods: case: private insurance customer data," 2019.*
*[14] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in 2013 Sixth international conference on contemporary computing (IC3), 2013: IEEE, pp. 404-409.*
*[15] Y. Liu, Y. Liu, X. Bruce, S. Zhong, and Z. Hu, "Noise-robust oversampling for imbalanced data classification," Pattern Recognition, vol. 133, p. 109008, 2023.*
*[16] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD explorations newsletter, vol. 6, no. 1, pp. 20-29, 2004.*
*[17] C. Diamantini and D. Potena, "Bayes vector quantizer for class-imbalance problem," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 638-651, 2008.*
*[18] L. Huang, T. Song, and T. Jiang, "Linear regression combined KNN algorithm to identify latent defects for imbalance data of ICs," Microelectronics Journal, vol. 131, p. 105641, 2023.*
*[19] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," Engineering Applications of Artificial Intelligence, vol. 37, pp. 368-377, 2015.*
*[20] A. A. Salarian, H. Etemadfard, A. Rahimzadegan, and M. Ghalehnovi, "Investigating the Role of*

*Clustering in Construction-Accident Severity Prediction Using a Heterogeneous and Imbalanced Data Set," Journal of Construction Engineering and Management, vol. 149, no. 2, p. 04022161, 2023.*

*[21] W. Xu, S. Wang, D. Zhang, and B. Yang, "Random rough subspace based neural network ensemble for insurance fraud detection," in 2011 Fourth International Joint Conference on Computational Sciences and Optimization, 2011: IEEE, pp. 1276-1280.*

*[22] S. Baran and P. Rola, "Prediction of motor insurance claims occurrence as an imbalanced machine learning problem," arXiv preprint arXiv: 2204. 06109, 2022.*

*[23] S. Meng, Y. Gao, and Y. Huang, "Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees," Insurance: Mathematics and Economics, 2022.*

*[24] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," International journal of pattern recognition and artificial intelligence, vol. 23, no. 04, pp. 687-719, 2009.*

*[25] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in European conference on machine learning, 2004: Springer, pp. 39-50.*

*[26] C. Cardie and N. Howe, "Improving minority class prediction using case-specific feature weights," 1997.*

*[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.*

*[28] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.*

*[29] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression," Risks, vol. 7, no. 2, p. 70, 2019.*

*[30] J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley, "Pruning decision trees with misclassification costs," in European Conference on Machine Learning, 1998: Springer, pp. 131-136.*