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. Utilizing 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 results 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 code 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.](image)

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).](image)
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):

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

where \( P \) is the predicted probability of the instance, it is positive, \( e \) is the mathematical constant, \( b_0 \) is the intercept term, \( b_1 \) is the coefficients, and \( X \) is the feature values vector.

The linear function of \( X \) 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):

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

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>
<thead>
<tr>
<th>Metrics</th>
<th>Computing method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>( \frac{TP + TN}{TP + TN + FP + FN} )</td>
</tr>
<tr>
<td>Precision</td>
<td>( \frac{TP}{TP + FP} )</td>
</tr>
<tr>
<td>Recall</td>
<td>( \frac{TP}{TP + FN} )</td>
</tr>
<tr>
<td>Specificity</td>
<td>( \frac{TN}{TN + FP} )</td>
</tr>
<tr>
<td>F1-score</td>
<td>( \frac{2 \times TP}{2 \times TP + FP + FN} )</td>
</tr>
</tbody>
</table>

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.
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.

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.

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.

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.
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

[20] A. A. Salarian, H. Etemadfard, A. Rahimzadegan, and M. Ghalehnovi, "Investigating the Role of


