Prediction of automobile insurance fraud claims using machine learning

Adedotun Adedayo F.1*, Odusanya Oluwaseun A.2, Adesina Olumide S.3, Adeyiga J.A.4, Okagbue, Hilary I.1, Oyewole O.4

Abstract
Automobile insurance fraud is a significant issue for insurance firms, causing financial losses and higher premiums for policyholders. This study aims to create a predictive model for accurately identifying potential vehicle insurance fraud claims. Understanding fraud detection processes and operationalizing information communication technology is crucial for implementing corrective actions, but personally reviewing insurance claims is time-consuming and costly. This study explored machine learning algorithms to detect fraudulent vehicle insurance claims. The research evaluated AdaBoost, XGBoostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost classifiers outperformed other models with 84.5% classification accuracy, while LR classifiers performed poorly with balanced and unbalanced data. The ANN classifier performed better with unbalanced data. Performance evaluation metrics such as accuracy, precision, recall, and F1 score are utilized to assess the effectiveness of the models. The results demonstrate the effectiveness of machine learning in distinguishing between genuine and fraudulent claims, providing insurance companies with a powerful tool to proactively combat fraud and improve their overall risk management strategies. The findings of this research contribute to the insurance industry's efforts to enhance fraud detection systems, reduce financial losses, and offer more competitive insurance premiums to honest policyholders.

Keywords: Prediction, Automobile, Insurance, Fraud claims, Machine learning, Fraud detection.

Introduction
Insurance must be reliable and reasonably priced in order to provide assistance to everyone when it is needed (Platteau, De Bock & Gelade, 2017). On the other hand, insurance fraud causes havoc in the insurance industry and raises premiums for insurers and insureds. Insurance companies look for cutting-edge solutions to improve fraud detection as technology advances. Based on data obtained from kaggle.com (Roy, 2021). This study suggests that machine learning methods could be used to improve fraud detection. The primary goal of this study is to improve the detection of fraudulent vehicle insurance claims.

Insurance fraud is typically committed for financial gain. This refers to any act of deception intended to obtain a benefit that does not belong to the deceiver (Christopher & Aditi, 2020). Numerous investigations have been conducted into the cost of insurance fraud. However, because fraudsters deliberately conceal their activities, calculating the precise damages caused by insurance fraud is difficult. As a result, far fewer false claims are exposed than are actually made (King, Timms & Rubin, 2021). Fraud, whether discovered or not, may account for up to 10% of total claim expenses (Hilal, Gadsden & Yawney, 2021). According to a survey, fraud was present in 3 to 6.4% of all claim payouts (Pathmananathan & Aseh, 2021).

Given the volume of claims filed each day, it would be prohibitively expensive for insurance companies to employ staff to scrutinize each claim for signs of fraud (Gennaioli et al., 2020). Instead, many businesses employ automated algorithms to identify allegations that may necessitate additional investigation (Dexe, Franke & Rad, 2021). To detect fraud, a two-step procedure was presented (Massi, Ieva & Lettieri, 2020). First, a ‘rule-based detection algorithm’
determines which dubious assertions are most likely to be false. The highlighted claims are then sent to fraud experts, who examine them again for potential fraudulent characteristics to determine whether a more thorough investigation is required.

As a result, insurance companies must urgently figure out how to precisely define risk factors and reduce the harm caused by fraudulent claims. And determining whether a claim is fraudulent or not necessitates specialized knowledge (Huang et al., 2017), despite the fact that there are far fewer experts than claims (which are increasing). This makes it difficult for a small group of experts to properly extract, interpret, and evaluate the specifics of situations.

Furthermore, a lack of experience may exacerbate decision bias. Even when discussing the same topic, the perspectives of different specialists can differ significantly due to their unique points of view. However, a number of experts and academics have worked hard to uncover vehicle insurance fraud using machine learning algorithms and comparing the performance of these algorithms to that of fraud detection experts (Wang & Xu, 2018). Machine learning algorithms are more effective than humans at detecting fraud (Severino & Peng, 2021). Many studies on the use of machine learning for insurance claims concentrated on medical insurance rather than short-term insurance, such as auto insurance (Hanafy & Ming, 2021).

Fraud detection in car insurance is challenging due to the lack of a precise rule to classify situations. Machine learning approaches are used to identify fraud but are impacted by unequal binary class distribution. This can lead to a decrease in the effectiveness of prediction models for tiny classes. Hence, this study evaluated a range of ML algorithms, including XGBoost, logistic regression, KNearest neighbor, and random forest to discern between real and fraudulent automobile claims.

**Literature Review**

Itri et al., (2019) compared the performance of four machine learning algorithms for detecting automobile insurance fraud. Random forest outperformed the other algorithms in terms of accuracy (98.2%), precision (94.2%), recall (94.1%), F1 score (94.1%), and AUC (0.986).

XGBoost is a machine that recognizes and categorizes fraudulent claims. Wang and Xu (2018) proposed an innovative deep-learning model for detecting vehicle insurance fraud based on latent dirichlet allocation (LDA) text analytics.

Auto-vehicle fraud can be detected using machine learning algorithms. Support vector machines were proposed by Roy and George (2017) and Rustam and Arianteri (2018) to categorize policyholders. Yan et al., (2020) combined a BP neural network with an improved adaptive genetic algorithm (NAGA) to maximize initial weight and predict accuracy. We use a sample of historical auto insurance claim data.

Machine learning can be used to detect and prevent fraudulent claims. Li et al., (2018) presents a multiple classifier system that is based on random forest, principal component analysis, and potential nearest neighbour. Huang and Meng (2019) forecast the risk likelihood and frequency of claims for an insured vehicle using Poisson regression, logistic regression, four machine learning techniques, and four risk probability models. Subudhi and Panigrahi (2018) investigated a novel fraud detection methodology based on adaptive oversampling.

Rawat et al., (2021) and Subudhi and Panigrahi (2020) used feature selection, EDA, machine learning algorithms, and Fuzzy C-means clustering to detect fraud in vehicle insurance services.

Machine learning can be used to automate fraud detection processes, allowing insurance companies to achieve faster resolutions and lower losses. Pesantez-Narvaez et al., (2019) used telemetry data to compare the effectiveness of the XGBoost and logistic regression algorithms in predicting the existence of accident claims. Wang and Xu (2018) present a novel deep learning model based on LDA text analytics for detecting automobile insurance fraud. The XGboost outperforms other existing learning algorithms in terms of performance. Dhieb et al., (2020) propose using the extreme gradient boosting (XGBoost) machine learning algorithm for insurance services.

In terms of convergence speed and prediction accuracy, Yan et al., (2020) proposed an improved adaptive genetic algorithm (NAGA) combined with a BP neural network that outperformed the traditional genetic algorithm.


Majhi et al., (2019) proposed a hybrid fuzzy clustering technique that included SSA and XGBoost Classifiers, whereas Subudhi and Panigrahi (2020) proposed a novel hybrid approach for detecting insurance fraud in automobile claims.

**Methodology**

**The Dataset**

The information was originally obtained from kaggle.com, but it was later discovered to be published by Oracle. It has 33 attributes, 33 features, and 15,420 policy claim records. Variables can be used to identify the individuals or organizations who made or received the claim.

**Modeling**

To achieve the goal, algorithms and preparation procedures should be chosen, with linear regression, logistic regression, and random forest being the best algorithms for the task.
Logistic regression
Regularization can prevent overfitting and detect fraud using logistic regression. Sci-kit learning includes regularization and class weighting.

Tree-based models
Tree-based models are less susceptible to outliers and data changes than other techniques. Random forest classifiers employ specialized trees with low bias and high variation, whereas light GBM employs weak learners with high bias and low variation to achieve low bias and low variation.

Random forest
Random forest regulates tree development and reduces variance by using Gini impurity as a selection criterion. Entropy produces different outcomes in only 2% of cases.

Metrics to Evaluate Imputation Methods
MAE, RMSE, and accuracy are used to evaluate the quality of imputation.

The following describes MAE and RMSE.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i| \\
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}
\]

where represents samples of model errors (.). We took into account the equations below in order to gauge the accuracy of imputations. Where \( Y \) stand for the imputation-predicted values and \( Y_i \) for actual values.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - Y_i| \\
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - Y_i)^2}
\]

Accuracy
Accuracy measures likelihood of correctly classifying test samples.

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{N}
\]

When N is the total number of instances, FN and FP are the number of false negatives, respectively.

Results

Data Exploration
In this phase data will be organized or managed so that it will be helpful to achieve the required goal.

Figure 1 shows the entire summary of the descriptive statistics of the data set. It is observed in the data set that the week of month shows the week of the month when the accident occurred. The average number of weeks that occurred in a month is two and the maximum number of weeks in a month is five. Similarly, the week of month claimed contains weeks in the month that the claimed in the field the mean of the week of the month is two and the max is 5. Age is the ages of individuals that make claims.

![Figure 1: Descriptive statistics summary of the dataset](image)

![Figure 2: Relation between independent and dependent variables](image)

![Figure 3: Distribution of fraudulent claims](image)

Classifier Score
Following the correction of the unbalanced dataset, the 80:20 test-train-split Python package and its machine learning tools were used to compute classifier scores for several models to compare to our proposed fraudulent detection.

Table 1 shows that KNN, XGboost, and random forest perform exponentially well on the dataset, with 93.5590,
84.0224, and 100.0000% on train data, respectively, whereas logistic regression does not. As a result, we can deduce that logistic regression is not a reliable model for this dataset. Other models, however, outperform the logistic regression model in terms of performance.

**Classification Report**

In Logistic regression, the accuracy of this model did not improve significantly. As a result, we can conclude that logistic regression is not a trustworthy model for this dataset. Other algorithms, such as KNN, XGboost, and random forest, perform well on the dataset, yielding 96, 89, and 100%, respectively. We can say that these algorithms can be used to produce accurate results with new and massive amounts of data. After fine-tuning the model, KNN’s accuracy has increased from 89 to 96%. The accuracy of XGBoost has increased from 84 to 89%. The logistic regression model did not improve after fine tuning, and the random forest model performed exceptionally well, most likely due to overfitting on the default value of the model. After fine-tuning the model to produce more realistic results, the best result was 98.5%.

The other matrices of the models are shown in the following Table 2.

**Discussion**

The prediction of automobile insurance fraud claims using machine learning has emerged as a critical area of research in the insurance industry due to the substantial financial losses and higher premiums resulting from fraudulent activities. This study aimed to develop a predictive model that accurately identifies potential fraudulent claims, providing insurance companies with a powerful tool to proactively combat fraud and improve overall risk management strategies. The findings from this research shed light on the effectiveness of machine learning algorithms in distinguishing between genuine and fraudulent claims, making valuable contributions to the insurance industry’s efforts to enhance fraud detection systems. The study explored several machine learning algorithms to detect fraudulent automobile insurance claims, including AdaBoost, XGboostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost demonstrated superior performance among these classifiers, achieving an impressive 84.5% classification accuracy. These results highlight the potential of ensemble learning methods to effectively handle complex relationships within the data and make accurate predictions. However, the study also revealed that the performance of LR classifiers was subpar, particularly with both balanced and unbalanced data. The ANN classifier showed promise in dealing with unbalanced data, indicating its utility in scenarios where fraudulent claims are less prevalent.

The study explored several machine learning algorithms to detect fraudulent automobile insurance claims, including AdaBoost, XGboostNB, SVM, LR, DT, ANN, and RF. AdaBoost and XGBoost demonstrated superior performance among these classifiers, achieving an impressive 84.5% classification accuracy. These results highlight the potential of ensemble learning methods to effectively handle complex relationships within the data and make accurate predictions. However, the study also revealed that the performance of LR classifiers was subpar, particularly with both balanced and unbalanced data. The ANN classifier showed promise in dealing with unbalanced data, indicating its utility in scenarios where fraudulent claims are less prevalent.

**Conclusion**

In this study, we successfully developed a predictive model for identifying potential automobile insurance fraud claims using machine learning techniques. The dataset utilized in building this insurance predictive model was obtained from Kaggle and covered the years 1994–1996, providing valuable historical information on insurance claims and their outcomes. However, to further validate and enhance the robustness of our proposed solution, it is essential to consider collecting new datasets from the last 2 to 5 years. These updated datasets would allow us to evaluate the model’s performance against more recent and diverse insurance scenarios, ensuring its applicability in real-world situations. Testing the predictive model’s generalization capabilities is highly recommended to assess it on different datasets from various sources. This cross-validation process would enable us to measure the model’s effectiveness across different insurance environments and ascertain its reliability in detecting fraudulent claims beyond the dataset on which it was trained. By exposing the model to various scenarios, we can gain deeper insights into its strengths and limitations and make necessary adjustments to improve its overall performance.

To achieve optimal results, we suggest exploring various hyperparameter settings during model development.
Randomly testing combinations of parameters or using a predetermined set can help identify the most suitable configurations for the machine learning algorithms. Additionally, it is crucial to test the model on datasets with different characteristics and distributions to ensure its adaptability to varying data environments. The performance of the model in diverse surroundings will validate its reliability and efficiency in real-world applications.

Reducing the number of characteristics used in the prediction process is recommended to address computational costs and optimize the model's efficiency. Feature selection and dimensionality reduction techniques can be applied to retain only the most relevant and informative features while discarding redundant or irrelevant ones. This step would streamline the model's computation time and memory requirements, making it more feasible for practical implementation in insurance fraud detection systems.

Throughout the study, we employed supervised machine learning techniques, including random forest, KNN, LR, and XGBoost, to build the insurance claims predictive models. Out of these four algorithms, random forest and KNN demonstrated exceptional performance on the dataset. Their ability to handle complex relationships in the data and effectively distinguish between genuine and fraudulent claims showcases their potential as reliable tools for insurance fraud detection.

In conclusion, this research contributes significantly to the domain of insurance fraud detection by demonstrating the efficacy of machine learning techniques in predicting automobile insurance fraud claims. However, it is important to acknowledge that the predictive model's reliability and generalizability can be further enhanced by incorporating more recent datasets, cross-validating on diverse environments, exploring hyperparameter settings, and optimizing feature selection. By continuously refining and updating the model with new data, insurance companies can proactively combat fraudulent activities, minimize financial losses, and provide more competitive premiums to their honest policyholders.

References
Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019, October). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In 2019...
Prediction of automobile insurance fraud claims using machine learning


Ghorbani, A. and Farzai, S., 2018. Fraud detection in automobile


