OPEN ACCESS

Anand P, et al. (2023) 142
J Contemp Edu Theo Artific Intel: 2023(3):100104-1
DOI: 10.71010/2023/JCETAI-e104-1

Research Article

Journal of Contemporary Education Theory & Artificial Intelligence

Enhancing Risk Assessment in Auto Insurance with Data-Driven Insights using Machine Learning

Anand Polamarasetti, MCA[1*], Rahul Vadisetty[2], Vasu Velaga[3], KishanKumar Routhu, ADP[4], Gangadhar Sadaram[5], Suneel Babu Boppana[6], Dinesh Kalla[7], Srikanth Reddy Vangala[8]

[1]Andhra University, exploretechnologi@gmail.com
[2]Wayne State University, Master of Science, rahulvy91@gmail.com
[3]Cintas Corporation, SAP Functional Analyst, vasuvelaga@gmail.com
[4]Openstack Architect, kishan1213@gmail.com
[5]Bank of America, VP DevOps/ OpenShift Admin Engineer, sadaram.gangadhar144@gmail.com
[6i]Site Technologies, Project Manager, suneelb@outlook.com
[7]Microsoft, Technical Support Engineer, kalladinesh@outlook.com
[8]University of Bridgeport, Computer Science Dept, srikanthreddy1043@microsoft.com

[*]**Corresponding author:** Anand Polamarasetti, MCA, Andhra University, Email: exploretechnologi@gmail.com

Citation: Anand P, Rahul V, Vasu V, KishanKumar R, Gangadhar S, et al. (2023) Enhancing Risk Assessment in Auto Insurance with Data-Driven Insights using Machine Learning. J Contemp Edu Theo Artific Intel: JCETAI-104-1.

Received Date: November 09, 2023; **Accepted Date:** November 20, 2023; **Published Date:** November 28, 2023

Abstract

The insurance firms, detecting auto insurance fraud is a major difficulty that may result in large financial losses. Insurance customers suffer substantial monetary losses, including higher premiums, as a result of claims that are forged. Manual inspections and rule-based techniques are the foundation of conventional fraud detection techniques, which are ineffective and unable to keep up with changing fraud trends. Machine learning (ML) is used in this investigation technique to improve fraud detection accuracy by analyzing vehicle insurance claim data. Extensive data preprocessing was applied, including handling missing values, feature selection, one-hot encoding, Min-Max normalization, and oversampling to address the severe class imbalance. A Random Forest (RF) classifier accomplished the uppermost accuracy (97.5%), outperforming Logistic Regression (LR) (87.1%) and EXtreme Gradient Boosting (XGBoost) (77.61%). Random Forest (RF) also showed superior precision (95.6%), recall (99.5%), and F1-score (97.5%), with an AUC of 0.98 from the ROC analysis, confirming its effectiveness. Despite its strong performance, limitations include dataset age and synthetic data from oversampling. The proposed approach offers an automated, scalable, and efficient fraud detection system, enhancing decision-making in the business of protection. Using machine learning (ML), this research offers an inexpensive and effective way to improve automobile insurance fraud detection.

Keywords: Auto Insurance, Risk Assessment, Insurance Fraud Prediction, Fraud Detection, Machine Learning (ML), vehicle insurance data.

1. Introduction

The insurance industry is essential for protecting individuals and businesses against financial damages resulting from unforeseen events. Vehicle insurance stands out as one of the major insurance products that protect drivers from incident-related losses and vehicle theft incidents and mechanical damage. The rising number of vehicles on roads has created an escalating market demand for auto insurance. The industry growth produces new difficulties in risk evaluation and deceptive claims which drives insurers to adopt data-backed solutions for making both accurate and efficient decisions. The category of vehicle insurance known as auto insurance formerly depended on static risk evaluation which integrated age demographics and credit history and past claim records among other variables [1]. A policyholder's basic risk profile can be assessed through these variables, but real-time driving habits alongside external environment elements, including road traffic conditions and infrastructure quality, remain outside their evaluation scope. Because of this insurer are focusing on data-based methods and real-time components to develop better risk assessment methods. Telematics combined with GPS technology gives insurers the power to review substantial datasets of live driving information thus enabling them to move toward custom risk assessment systems [2].

The critical element of change involves risk assessment enhancement to generate fair pricing and accurate underwriting decisions and reduce financial losses [3]. The actuarial models have traditional effectiveness, yet they confront challenges while processing big volumes of intricate, unorganized data. The insurance industry experiences billions of dollars in damages from fraudulent claims because new insurance fraud cases continue to increase. Fraudulent practices involving staged collisions and untruthful injury claims and fake documents cause financial damage to insurers which forces them to raise insurance costs for legitimate policyholders. Advanced analytical tools that find hidden patterns and strange behavior patterns need implementation by insurers who want to combat their fraud challenges effectively [4]. The implementation of an automobile insurance fraud prediction system creates crucial changes in industry practices. The prediction of fraudulent insurance claims within real time needs algorithms that process historical data about claims and insurance holder conduct alongside outside indicators of fraud [5]. The current fraud detection techniques using manual reviews together with rule-

based systems prove both inefficient and time-consuming because they contain human judgment errors.

However, with the combination of AI and ML, insurers can automate fraud detection processes, reducing the chances of undetected fraud while improving operational efficiency [6]. insulars now use AI and ML systems to transform their risk evaluations as well as their fraud detection processes. The analysis of extensive amounts of structured along unstructured data by ML algorithms enables users to find hidden relationships while providing precise risk predictions. Automated systems for identifying fraud allow insurers to process new information continuously so they can find fraudulent claims before they are submitted. ML-driven models enable actual risk-based pricing for premiums which replaces generalized demographic-based charges for policyholders.

A. Motivation and Contribution of this Paper

This study originated from the rising issue of auto insurance fraud incidents that cause substantial financial damage to insurers while pushing up policy premiums for everyone. Conventional identifying fraudulent activity is inadequate and unable to change with changing fraud trends since it is manually and based on rules. By examining enormous data sets to find unconscious trends, ML enhances detection of fraud. The purpose of this research is to decrease fraudulent claims and increase accuracy, and create an automated, scalable system for a fair and secure insurance ecosystem.

- It uses a publicly available vehicle insurance dataset from Kaggle to ensure a diverse and representative sample for fraud detection.
- Implemented pre-processing such as missing value imputation, one-hot encoding, normalization, and class balancing to enhance model performance.
- Applies Implementing one-hot encoding to transform information with categories into an arrangement that is compatible with ML, preventing misinterpretation as ordinal data.
- Uses Min-Max scaling to standardize numerical attributes within [0,1] and applies oversampling techniques to address class imbalance (94% non-fraud, 6% fraud), improving model learning and reducing bias.
- Proposed a RF model and Assesses model effectiveness using key metrics comparable accuracy, precision, recall, F1-score, and ROC curvature investigation.
- Compares RF with LR and XGBoost (XGB), proving RF's superiority in fraud detection.

B. Organization of paper

This research is designed as surveys: Section II reviews associated scholarships on ML for auto insurance risk assessment. Section III outlines the methodology, including data preprocessing, model implementation, and evaluation. Section IV displays findings and conversations based on significant performance metrics. Section V accomplishes with key findings and upcoming research directions for enhancing risk assessment and policy optimization.

2. Literature Review

The research section investigates risk assessment information evaluation through ML approaches to forecast insurance claims and improve decision making in the automotive field.

Patel and Subudhi (2019) creates a new method to discover atypical claims in vehicle insurance documents using neural network-based Extreme Learning Machine (ELM). The initial stage required preprocessing of raw data before setting apart the training sets and validation sets as well as test sets. Multiple trained ELM classifiers form a pool based on different parameter settings after being used on the Pullman set. The selection of best ELM model occurs through utilizing validation set on multiple trained models. The validated model receives the testing set to identify legitimate from malicious insurance claims. The model's performance is shown through thorough tests that utilize a common auto insurance dataset [7].

Itri et al. (2019) the insurance industry fights against massive fraud issues which ML and Big Data actively work to resolve at present. This document evaluates how well famous ML algorithms function as fraud prediction tools while also checking their validity. They used the supervised approach on vehicle data claims that they received from an unnamed insurance provider. their method makes an attempt to improve how artificial intelligence generates relevant outcomes. RF proved its superiority among all available algorithms according to the study results [8].

Denuit, Guillen and Trufin (2019) presents multivariate mixed models as an approach to modeling the simultaneous patterns between telematics measurements and claim occurrences. Using predictive distributions of claims based on historical records allows for future premium assessment. The actuary may manage insurance practice complexities through this method that handles new drivers without telematics records and contracts across various levels of seniority while considering how much drivers use their vehicles and how much telematics data they produce [9].

Wang and Xu (2018) presented work develops a fresh DL system for detecting scam involving car insurance via LDA-based text analysis algorithms. Next, LDA retrieves text-based features from claim description texts while deep neural networks process both extracted features along with numerical features to detect insurance fraud. Their proposed text analytics method achieves superior results than traditional frameworks according to experimental analyses of the real-world insurance fraud data [10].

Li et al. (2018) demonstrates proper combination of individual classifiers which leads to a multiple classifier system showing better classification accuracy. The RF and Principal Component Analysis, combined with Potential Nearest Neighbor Methods, comprise their multiple classifier system following Breiman's guidelines because, according to Breiman, these techniques depend largely on weak learner strength and diversity between learners. This paper shows RF acts as an adaptive learning system for k Potentially closest Neighbors using the idea of potential closest neighbors and monotone distance metrics. A novel voting method based on Potential Nearest Neighbors is presented in this study which replaces traditional majority vote because it addresses information losses from out-of-bag samples. An enhanced ensemble classifier becomes more effective through the proposed algorithm by increasing the separation between individual base classifiers [11].

Badriyah, Rahmaniah and Syarif (2018) develop predictive models for anomaly detection to identify fraudulent activities through the combination of distance-based Nearest Neighbor and density-based Nearest Neighbor and interquartile range statistics. The investigation employs open fraud dataset that earlier studies have used to prove their fraud detection abilities. A minority reporting open dataset of German insurance company data used as the comparative dataset. The outcome of the evaluation data is measured against results found by earlier researchers who analyzed the same dataset. Experimental data shows that the measurement output using the current study method exhibits better results compared to other instances [12].

Table I highlights key limitations in methodologies, datasets, and performance benchmarks, providing a foundation for future advancements.

Table 1: Summary of the Literature Review on ML for Enhanced Auto Insurance Risk Assessment.

References	Methodology	Performance	Advantage	Limitations & Future Work
Patel and Subudhi (2019)	Based on neural networks ELM	Demonstrated effectiveness on auto insurance dataset	Fast training time, improved fraud detection accuracy	Needs comparison with other deep learning models like CNNs or RNNs
Itri et al. (2019)	RF for fraud prediction in auto insurance	Best performance among compared ML algorithms	Demonstrated effectiveness of RF in fraud prediction	Requires exploration of ensemble methods for further improvement
Denuit, Guillen and Trufin (2019)	Pricing for insurance based on telematics using multivariate mixed models	Predictive distribution of claim characteristics	Considers behavioral data for pricing accuracy	Requires further exploration of deep learning for behavioral insights
Wang and Xu (2018)	DL with LDA-based text analytics	Outperforms traditional ML models (Random Forest, SVM)	Combines DL and text analytics to identify fraud.	Needs further validation on larger datasets and real-time applications
Li et al. (2018)	RF with monotone distance measures and Potential Nearest Neighbor-based voting	Improved classification accuracy, lower variance	Enhances RF performance with new voting mechanism	Requires testing on real-time fraud detection scenarios
Badriyah, Rahmaniah and Syarif (2018)	Detecting anomalies Interquartile distribution and Nearest Neighbor-based methods (distance and density-based) are used.	Outperformed previous studies on the German car insurance dataset	Effective for anomaly detection in fraud detection scenarios	Needs validation on more diverse datasets beyond German car insurance

3. Methodology

This study's technique uses ML to evaluate automobile insurance fraud using an organized, data-driven approach, as shown in Figure 1. 15,420 automobile insurance claims with 32 predictive variables, together with one target variable that indicates fraud or non-fraud are included in the first set of raw data on auto insurance that was acquired from Kaggle. Data preparation comprised categorical variable collection, one-hot encoding, managing missing values, data normalization, and addressing class imbalance through oversampling techniques. To assess the model, the dataset was divided into 20% testing and 80 % training groups. Because of its resilience and capacity to manage huge datasets effectively, a RF classification was used. Several decision trees were used in the harvesting process to train the RF modelling was built using bootstrapped samples, and the final prediction was derived by averaging individual tree outputs. The ROC curve, F1-score, recall, accuracy, and precision were used to assess the machine learning model's effectiveness.

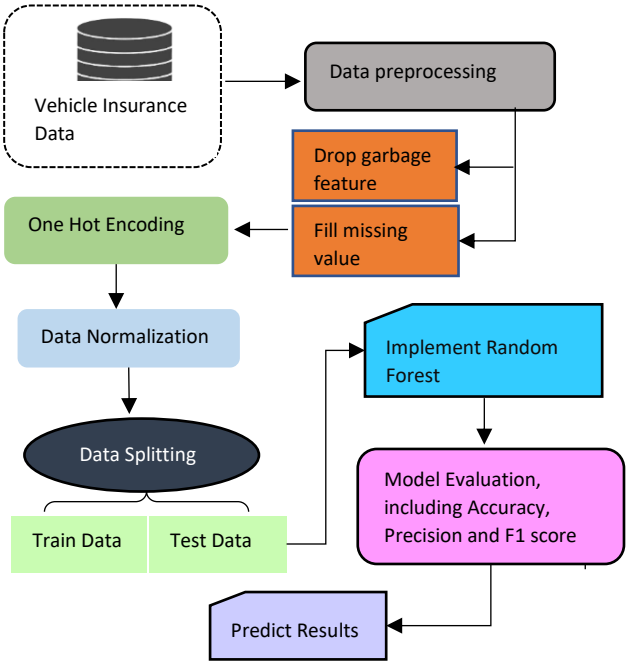

Figure 1: Flowchart for Auto Insurance.

The following steps and process of methodology are elaborate below:

A. Data Collection

The Kaggle platform was the source of the Vehicle Insurance dataset utilized in the experiment. It is made up of "carclaims.txt" files, which contain information about auto insurance claims that were taken from the Angoss Knowledge Seeker software. The data, which comprises 15420 claims from January 1994 to December 1996, is composed of 32 predictor variables and one target variable that decides whether a claim is "Fraud" or "No Fraud." The data collection comprises 430 claims on average every month, of which 14,497 are actual (non-fraud) claims (94%) and 923 are fraud instances (6%). Figure 2 shows the data shared among the two classes.

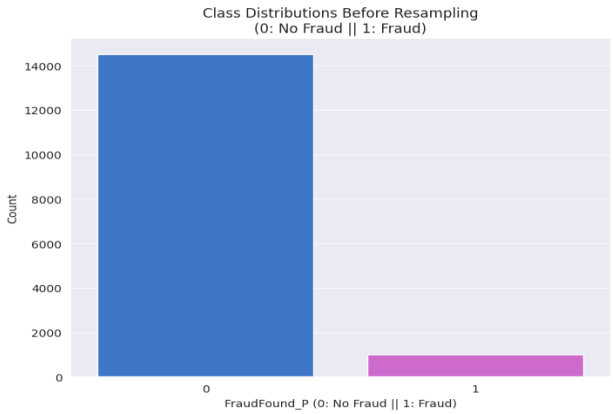

Figure 2: Bar Graph for Class Distribution.

Figure 2 illustrates the class distribution of fraud detection before resampling, where the majority class (No Fraud, labeled as 0) significantly outnumbers the minority class (Fraud, labeled as 1). The imbalance is visually evident, with non-fraudulent claims exceeding 14,000, while fraudulent claims are fewer than 1,000, indicating a severe skew in the dataset. Such an Unbalanced development of ML models might result in biased recommendations that favor the overwhelming class.

B. Data Preprocessing

Preparing the data is seen as an essential stage in both ML and data mining [13]. A nuisance, insufficient, inconsistent, or redundant data are often found in massive databases. The data needs to be processed through a number of preliminary processing processes to get it into an appropriate format in order to create a reliable model. The Key pre-processing terms are listed below:

- **Drop Garbage Feature:** Removing irrelevant, redundant, or low-variance characteristics of a dataset that don't help model performance, thereby improving efficiency and accuracy.
- **Fill Missing Values:** This involves handling incomplete data by imputing missing values using techniques like predictive, forward-fill, backward-fill, mean, median, or mode modeling to maintain data integrity and prevent bias in analysis.

C. One Hot Encoding

A method for converting categorical data into a numerical format suitable for machine learning models is called "one-hot decoding." For every, it generates binary columns unique category, assigning 1 to the present category and 0 to the others. This method prevents the model from misinterpreting categorical data as ordinal. While effective, it increases dimensionality, especially with high-cardinality features, requiring careful handling to optimize computational efficiency.

D. Data Normalization

The development of clambering each attribute value in a record inside the interval [0, 1] is known as data normalization. The determination of normalization is to lessen the impression of highly valued data elements on the performance of classifiers [14]. The raw dataset D was subjected to the Min-Max normalization approach, which can be quantitatively stated as shown in Equation (1):

$$Norm_D = \frac{D - D_{min}}{D_{max} - D_{min}} \qquad (1)$$

where $Norm_D$ is the normalized form of the inventive dataset, D_{min}, and D_{max} denote the min and max values of the elements in D.

E. Class Oversampling

The model can better understand the patterns and distinctions across classes if it is exposed to more instances of the minority class. When the original dataset's breakdown of classes is unbalanced, leading to a model that is skewed towards the dominant class, this approach may be helpful. The accuracy of the predictive algorithm may be increased by oversampling, particularly when the cost of false negatives and the failure to discover the minority class are significant. The distribution of the utilized dataset upon resampling is shown in Figure 3.

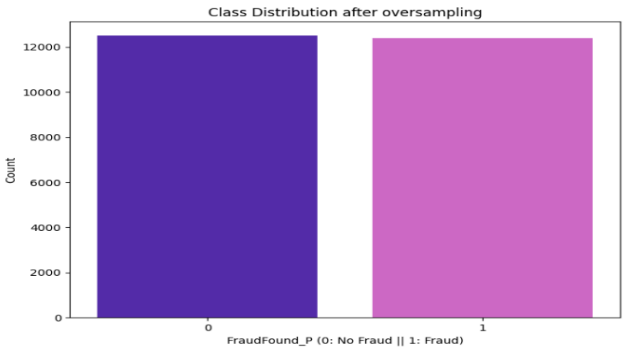

Figure 3: Class Distribution of the Vehicle Insurance Data.

Figure 3 illustrates the distribution of predicted fraud labels ("Fraud Found P") in a vehicle insurance dataset. The x-axis represents the predicted labels: 0 for "No Fraud" and 1 for "Fraud." The y-axis demonstrations the count of each label. Both categories have approximately equal counts, around 12,000, indicating a balanced dataset in terms of predicted fraud instances.

A. Data Splitting

The basic premise is to partition the dataset into separate parts for testing as well as training. Partitioning the dataset into 80% training and 20% testing sets ensures effective model training and evaluation, enabling accurate prediction of risk factors and insurance claims using data-driven insights.

B. Proposed Random Forest Model

RF is an ensemble approach that creates a single forecast by combining the output of several regression trees. The main idea is bagging, which involves randomly selecting some amount to provide training data and fitting it into a regression tree [15]. This sample was chosen at random. RF is a collective method

that combines the outcomes of several regression trees to get a single forecast. Bagging is the main idea, which is the process of choosing a bootstrap sample, a random selection of training data, and fitting it into a regression tree. Any data point that has already been chosen may be utilized one more. N data points are randomly chosen from the dataset, and the data points are then substituted for them that are already there, a bootstrap sample may be created. Any data point has a 1/N probability of getting selected. Decision tree estimation techniques are combined to create RF $\{h(X', \Theta k, k = 1,2,\ldots.,)\}$ Each DT is computed using a random vector's outputs $\{\Theta k\}$, It is equally distributed across all of the decision trees in the forest and separately sampled. Upon completion of training, the average output of all DT on sample X′ is used to produce forecasting, as indicated by Equation (2).

$$\hat{f} = \frac{1}{k}\sum_{i=1}^{k} h(X', \Theta k) \qquad (2)$$

here $\hat{f}$ is the final prediction and k is the number of DT.

A. Model Evaluation

The performance metrics used to evaluate model effectiveness included comparing actual observations with predictions. The evaluation matrix included accuracy, precision, recall, and F1-score for assessment risk assessment results in vehicle insurance. The main tool for comparing the model's estimated and actual results is a confusion matrix, which gives a clear evaluation of the model's performances. It shows the count of TP, TN, FP, and FN. The class-specific metrics, comprise accuracy, precision, recall, f1-score, and ROC, were computed independently for precise risk classification and assessment.

- **TN:** List the number of entries whose real classification was minus and whose grouping was correctly identified as negative by the algorithm [16].
- **FP:** List the number of entries whose true category was unfavorable and whose affirmative category was mistakenly identified by the algorithm.
- **FN:** List the number of entries whose real categorization was positive but whose unfavorable classification the algorithm misidentified.
- **TP:** List the total amount of entries for which the algorithm correctly identified their category as favorable but whose actual categorization was positive.

Accuracy: Accuracy is an overall metric that approximations the classifier's exactness. Equation (3) calculates Accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (3)$$

Precision: This is an indicator of how many genuine positives the model reports in relation to how many positives it promises. Equation (4) below provides the accuracy value for a single class:

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

Recall: Recall, or sensitivity, which was determined using Equation (5) is the percentage of Real Positive instances that are accurately Predicted Positive.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1 score: The sung median of precision and awareness is acknowledged as the F1-score. It is sometimes referred to as the Dice Similarity Agreement or the Sorensen–Dice Coefficient. The ideal value is 1. The following Equation (6) illustrates how the F1-score is calculated:

$$F1\ score = \frac{2.TP}{2.TP+FP+FN} \qquad (6)$$

ROC: A classification model generates its performance data for various threshold values through the graphical Headset Operating Characteristic (ROC) curve. The True Positive Rate (TPR) and False Positive Rate FPR) appear together on a plot which depicts the model sensitivity against specificity.
The performance metrics analyze test set outputs from the model for effectiveness evaluation.

4. Result And Discussion

The section outlines how the proposed approach was executed together with its performance assessment. The research execution took place on a system equipped with Ubuntu 22.04 LTS and an AMD Ryzen 9 5900X mainframe and 64 GB of RAM. A set of ML algorithms underwent testing to determine their capacity in identifying auto insurance risks by analyzing data-driven insights according to Table II. The RF model led the performance metrics evaluations with 97.5% accuracy as it proved effective in insurance risk assessment. Few false positives occur in the model with its 95.6% precision but the 99.5% recall level protects against missing risky policyholders. The F1-score evaluation of 97.5% demonstrates that the model effectively maintains a balanced relationship between precision and recall for reliable auto insurance risk assessment.

Table 2: Experimental Results of Random Forest Performance for auto insurance fraud.

Measures	Random Forest
Accuracy	97.5
Precision	95.6
Recall	99.5
F1-score	97.5

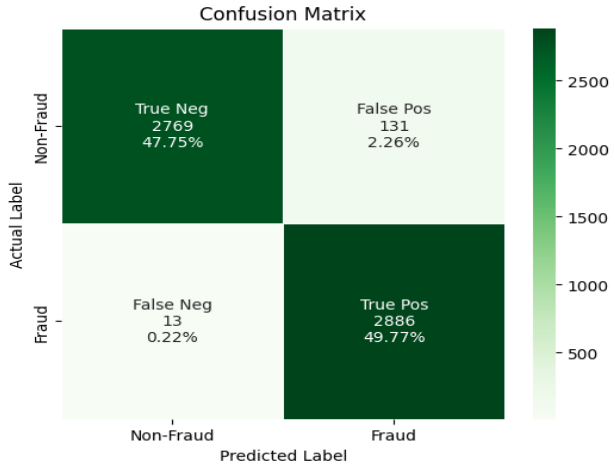

Figure 4: Confusion Matrix for Random Forest.

The confusion matrix in Figure 4 evaluates a vehicle insurance fraud detection model. It shows 2769 TN (correctly identified non-fraud), 2886 TP (correctly identified fraud), 131 FP (non-fraud misclassified as fraud), and 13 FN (fraud misclassified as non-fraud).

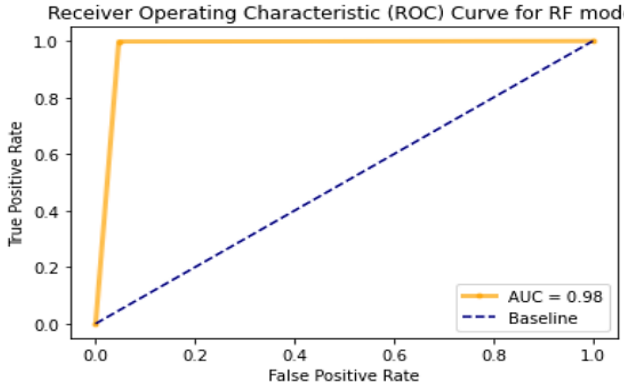

Figure 5: ROC Curve for Random Forest.

Figure 5 evaluates a RF model for vehicle insurance fraud, showing its TPR vs. FPR. The model's orange line is near perfect, with an AUC of 0.98, indicating strong performance. Compared to the baseline (dashed line), the model excels in distinguishing fraudulent claims.

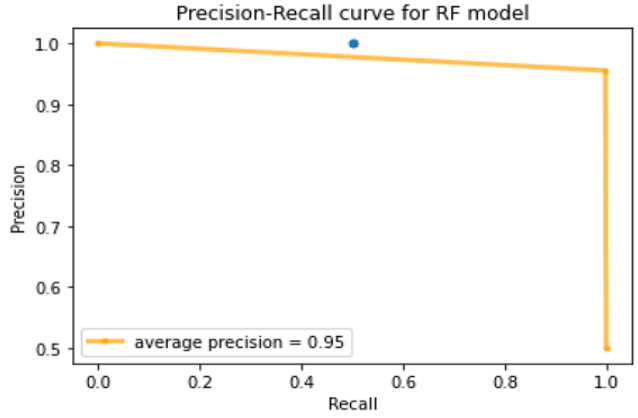

Figure 6: Precision-Recall Curve for Random Forest.

Figure 6 shows the Precision-Recall curve evaluates a RF model for vehicle assurance fraud. It shows a high average precision of 0.95, indicating the model's accuracy. The curve highlights the precision-recall trade-off, with the model maintaining high precision as recall increases. A blue dot marks a specific operating point.

A. Comparative Analysis

A comparative analysis of different ML models for risk assessment in auto insurance is presented in Table III. The RF model demonstrates the highest classification presentation, achieving 97.5% accuracy and an F1-score of 97.5%, demonstrating its toughness in risk assessment. The LR model follows with an accuracy of 87.1%, showing a strong precision (93.1%) but a lower recall (62.4%), suggesting that it may struggle with identifying high-risk cases effectively. Meanwhile, XGBoost achieves an accuracy of 77.61%, with a recall of 85.66%, making it a viable option for detecting high-risk cases, but its F1-score (68.56%) highlights room for improvement. For auto insurance risk assessment RF proves to be the most successful model when compared to both LR and XGBoost because of its superior accuracy level and balanced performance results.

Table 3: Comparison of Model Performance for auto insurance fraud.

Model	Accuracy	Precision	Recall	F1-score
LR[17]	87.1	93.1	62.4	93.1
XGB[18]	77.61	76.61	85.66	68.56
RF	97.5	95.6	99.5	97.5

The use of data-driven algorithms enables the risk assessment model to exceed traditional accuracy rates by reaching 97.5% while automatically finding elaborate risks that do not need human intervention. Modern actuarial approaches, which update to changing risk elements, allow the model to offer strong adaptive classification. The data-driven tool operates with automated and scalable capability to improve risk assessments in auto insurance through its efficient intelligent insights.

5. Conclusion and Future Work

The insurance industry faces substantial difficulties from vehicle insurance fraud because both costs rise and customer faith deteriorates. The increasing technological adoption has brought about complex fraudulent activities that create obstacles for organizations to stop and recognize them. A method based on ML data enables detecting fraudulent auto insurance claims in this study. The investigation relies on insurance data with advanced data preprocessing strategies and balanced class techniques to create a trustworthy model. Although the proposed model achieves high accuracy, it has certain limitations. The dataset may not reflect modern fraud trends, and the use of oversampling to address class imbalance could introduce synthetic data that differs from real-world fraud patterns DL models can be investigated in future studies (e.g., neural networks, transformers) for enhanced fraud detection, integrate real-time streaming data for proactive detection, and expand datasets with recent and diverse claims for improved generalizability.

References

1. G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, 2018. doi: 10.1109/ICICCT.2018.8473034.
2. P. Li, B. Shen, and W. Dong, "An anti-fraud system for car insurance claim based on visual evidence," *arXiv Prepr. arXiv1804.11207*, 2018.
3. S. Subudhi and S. Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud," in *Proceedings - 2nd International Conference on Data Science and Business Analytics, ICDSBA 2018*, 2018. doi: 10.1109/ICDSBA.2018.00104.
4. R. Roriz and J. L. Pereira, "Avoiding Insurance Fraud: A Blockchain-based Solution for the Vehicle Sector," *Procedia Comput. Sci.*, vol. 164, pp. 211–218, 2019, doi: https://doi.org/10.1016/j.procs.2019.12.174.
5. C. Oham, R. Jurdak, S. S. Kanhere, A. Dorri, and S. Jha, "B-FICA: BlockChain based Framework for Auto-Insurance Claim and Adjudication," *Proc. - IEEE 2018 Int. Congr. Cybermatics 2018 IEEE Conf. Internet Things, Green Comput. Commun. Cyber, Phys. Soc. Comput. Smart Data, Blockchain, Comput. Inf. Technol. iThings/Gree*, pp. 1171–1180, 2018, doi: 10.1109/Cybermatics_2018.2018.00210.

6. V. Kolluri, "An In-Depth Exploration of Unveiling Vulnerabilities: Exploring Risks in AI Models and Algorithms," *Int. J. Res. Anal. Rev.*, vol. 1, no. 3, 2014.
7. D. K. Patel and S. Subudhi, "Application of extreme learning machine in detecting auto insurance fraud," in *2019 international conference on applied machine learning (ICAML)*, 2019, pp. 78–81.
8. B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019, pp. 1–4. doi: 10.1109/ICDS47004.2019.8942277.
9. M. Denuit, M. Guillen, and J. Trufin, "Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data," *Ann. Actuar. Sci.*, vol. 13, no. 2, pp. 378–399, 2019.
10. Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decis. Support Syst.*, 2018, doi: 10.1016/j.dss.2017.11.001.
11. Y. Li, C. Yan, W. Liu, and M. Li, "A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification," *Appl. Soft Comput.*, vol. 70, pp. 1000–1009, 2018, doi: https://doi.org/10.1016/j.asoc.2017.07.027.
12. T. Badriyah, L. Rahmaniah, and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," in *Proceedings of the 2018 International Conference on Applied Engineering, ICAE 2018*, 2018. doi: 10.1109/INCAE.2018.8579155.
13. A. Immadisetty, "Edge Analytics vs. Cloud Analytics: Tradeoffs in Real-Time Data Processing," *J. Recent Trends Comput. Sci. Eng.*, vol. 13, no. 1, pp. 42–52, 2016.
14. N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations," in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2019, pp. 1–5. doi: 10.1109/ICVES.2019.8906396.
15. Y. H. Rajarshi Tarafdar, "Finding Majority for Integer Elements," *J. Comput. Sci. Coll.*, vol. 33, no. 5, pp. 187–191, 2018.
16. K. Faramarz, Z. S. Ahad, and D. Gholamhosseyn, "Identifying the effective factors in the profit and loss of vehicle third party insurance for insurance companies via data mining classification algorithms," *Indian J. Sci. Technol.*, vol. 9, no. 18, 2016, doi: 10.17485/ijst/2016/v9i18/93767.
17. H. Moon, Y. Pu, C. Ceglia, and others, "A predictive modeling for detecting fraudulent automobile insurance claims," *Theor. Econ. Lett.*, vol. 9, no. 06, p. 1886, 2019.
18. S. K. Majhi, S. Bhatachharya, R. Pradhan, and S. Biswal, "Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection," *J. Intell. \& Fuzzy Syst.*, vol. 36, no. 3, pp. 2333–2344, 2019.
19. Kuraku, D. S., Kalla, D., & Samaah, F. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*, *9*(12).
20. Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1)*, *191*, 2-7.
21. Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.
22. Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.
23. Routhu, K., & Jha, K. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *Available at SSRN 5106490*.
24. Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. *Available at SSRN 5102662*.