DETECTION OF INSURANCE FRAUD USING MACHINE LEARNING BASED METHODS ON CLASS IMBALANCE DATASETS WITH MISSING VALUES

[1]D. Ramesh, [2]N.Mourya, [3]P.Chandu, [4]P.Chandhu, [5]Y.Mahidhar
[1]Assistant Professor, [2,3,4,5]UG Student, [1,2,3,4,5]Department of Computer Science & Engineering (AI&ML),
Geethanjali Institute 0f Science And Technology, Nellore, India

Abstract

This project, titled "Detection of Insurance Frauds Using Basic Machine Learning Based Methods on Class Imbalance Datasets with Missing Values," addresses the growing concern of fraudulent automobile insurance claims through the application of artificial intelligence and machine learning. The dataset utilized originates from an Egyptian car insurance company, comprising 1,000 claim records, of which 21.7% are fraudulent and 78.3% are legitimate—indicating a significant class imbalance that can bias conventional models toward the majority class. Furthermore, the dataset contains missing values across various features such as policy tenure, vehicle type, and claim amount, which adds complexity to the prediction task and reduces data quality.To tackle these challenges, this study implements a comprehensive machine learning pipeline incorporating robust algorithms such as Random Forest, XGBoost, Gradient Boosting, and Neural Networks etc. These models are chosen for their ability to handle non-linear data, reduce overfitting, and adapt to high-dimensional and noisy data environments. In addition, advanced preprocessing steps are integrated to optimize model performance—such as Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and imputation methods like mean, mode, and KNN-based strategies to handle missing values effectively. The ultimate goal of this project is to build a reliable, intelligent fraud detection system that not only achieves high accuracy but also provides actionable insights for insurers. By automating the identification of suspicious claims, the system reduces manual workload, improves operational efficiency, and minimizes the risk of financial losses due to fraudulent activities. This model-driven approach contributes to a more secure, data informed, and proactive insurance process, ensuring fairness for genuine claimants and improving the credibility of insurance providers

Keywords: Insurance fraud, Machine learning, imbalance dataset, missing values, SMOTE, KNN

Introduction

Insurance fraud is a significant challenge for both the insurance industry and policyholders. Fraudulent activities can result in massive financial losses for insurance companies, ultimately leading to higher premiums for honest customers. With the rise of advanced technology and machine learning, detecting and preventing insurance fraud has become more efficient. However, fraud detection is complicated by various factors, including imbalanced datasets and missing values in the data, which can affect the accuracy of machine learning models. In this introduction, we explore the concept of insurance fraud, its impact on the industry, the challenges in detecting fraud, and how machine learning methods are being employed to improve detection, especially in cases involving imbalanced data and missing values.

Insurance Fraud

Insurance fraud is an intentional act of deception or misrepresentation made by an individual or group to receive financial benefits from an insurance policy to which they are not entitled. This fraud can take many forms, including:

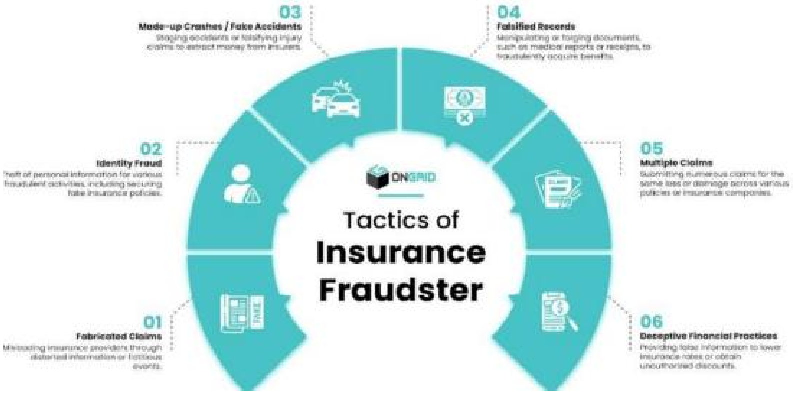

03
Made-up Crashes / Fake Accidents
Staging accidents or falsifying injury claims to extract money from insurers.
04
Falsified Records
Manipulating or forging documents, such as medical reports or receipts, to fraudulently acquire benefits.
02
Identity Fraud
Theft of personal information for various fraudulent activities, including securing fake insurance policies.
ONGRID
Tactics of
Insurance
Fraudster
05
Multiple Claims
Submitting numerous claims for the same loss or damage across various policies or insurance companies.
01
Fabricated Claims
Misleading insurance providers through distorted information or fictitious events.
FAKE
06
Deceptive Financial Practices
Providing false information to lower insurance rates or obtain unauthorised discounts.

Fig.1. Types of frauds

The Impact of Insurance

Fraud Insurance fraud has a wide range of consequences:

Financial Losses: Fraudulent claims can lead to direct financial losses for insurance companies, affecting their profitability.

Increased Premiums: To recover the costs of fraudulent claims, insurance companies often raise premiums for all policyholders, including the honest ones.

Damage to Reputation: Insurance companies may suffer from a loss of customer trust if fraud is not effectively detected and prevented.

Resource Drain: Fraud detection requires time, money, and human resources, putting a strain on an insurance company’s operations. Thus, it is essential for the insurance industry to adopt effective fraud detection methods to reduce these risks.

Literature Review

[1] Khalil, Ahmed A., Zaiming Liu, Ahmed Fathalla, Ahmed Ali, and Ahmad Salah. "Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values." In this paper, the authors address the challenges of missing data and class imbalance in automobile insurance fraud detection by utilizing an Egyptian real-life dataset and a standard dataset. They apply imputation techniques to handle missing data and employ methods such as SMOTE oversampling, under sampling, and hybrid approaches to address class imbalance. Various classifiers— including Decision Trees, Random Forests, SVMs, and Neural Networks—are trained and evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. A novel overfitting analysis is conducted to ensure model generalization. The study finds that addressing class imbalance significantly improves performance, while handling missing values has a minimal effect. The proposed methods outperform existing techniques in accuracy, highlighting the importance of dataset balancing. However, the paper does not explore the impact of feature selection techniques or the potential benefits of ensemble methods in fraud detection. Future research may investigate these areas to further enhance performance.

[2] CARRACEDO, PATRICIA, David Hervás, and Raquel Soriano-Gonzalez. "Class Imbalance in Insurance Fraud Detection Models." In this paper, the authors address the significant challenge of insurance fraud, which leads to substantial economic losses and affects pricing policies for consumers. They focus on developing a fraud detection model tailored for automobile insurance claims, employing the Random Forest Quantile Classifier. This approach optimizes classification thresholds, enhancing both specificity and sensitivity, and allows for adjustments to meet specific company requirements. Utilizing actual data from an insurance firm, the model demonstrated superior performance over other machine learning methods, achieving a balanced detection of fraudulent and non-fraudulent claims. Additionally, the model offers interpretability by elucidating variable effects on predictions, moving beyond the 'black box' nature of many algorithms. However, implementing

Random Forest models can be computationally intensive, especially with large datasets, and may require significant resources. Moreover, while the model provides interpretability, the complexity of Random Forests can still pose challenges in fully understanding the intricate relationships within the data. Overall, this research advances fraud detection techniques, offering valuable tools for professionals and researchers in financial data analysis. [3] Shamsuddin, Siti Nurasyikin, Noriszura Ismail, and R. Nur-Firyal. "Life insurance prediction and its sustainability using Machine learning approach." In this paper, the authors investigate underinsurance, where life insurance coverage is insufficient to cover expenses, impacting family financial health. They emphasize the importance of customer profiling in the insurance industry for identifying potential policyholders and highlight machine learning as a key method for effective customer segmentation. The research proposes a data mining framework utilizing various sampling methods to predict potential life insurance policyholders. Techniques like Synthetic Minority Over-sampling Technique (SMOTE), Random Under-Sampling, and ensemble methods (bagging and boosting) are employed to address imbalanced datasets. The study finds that the decision tree model performs best according to the Receiver Operating Characteristic (ROC) curve. However, Naïve Bayes outperforms in balanced accuracy, F1 score, and Geometric Mean (GM) comparisons. Ensemble models do not consistently guarantee high performance with imbalanced datasets. Despite this, ensemble and sampling methods are crucial in mitigating the imbalance issue. The study contributes to sustainable life insurance industry development by enhancing policyholder prediction accuracy and underscores the need for effective data mining approaches in addressing underinsurance and improving financial stability for families.

[4] Alcaide, Diogo Cunha, and Rui Alexandre Henriques Gonçalves. "Predicting lapse rate in life insurance: An exploration of machine learning techniques." In this paper, the authors address fraud detection in the automobile insurance sector, which experiences the highest rate of fraudulent claims in the industry. They propose a methodology that combines resampling techniques and backward elimination for feature selection to tackle data imbalance, significantly improving the accuracy of machine learning models in identifying fraudulent activities. The study conducts a comprehensive comparison of seven major machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, k-nearest Neighbors, Naive Bayes, XGBoost, and Support Vector Machine. These models are evaluated based on precision, recall, and F1 score across three different datasets to determine their effectiveness in fraud detection. The findings reveal that the Random Forest algorithm is the most efficient, consistently achieving F1 scores above 0.92. This highlights the vital role of machine learning in detecting and preventing fraud in the automotive insurance industry

This study explores the application of artificial intelligence (AI) and Machine Learning (ML) techniques, particularly Random Forest classification, to predict car insurance risks using publicly available datasets from Kaggle [8] [9]. By implementing feature extraction and classification methodologies, this research demonstrates the effectiveness of AI-driven predictive models in enhancing risk assessment accuracy and operational efficiency in the insurance sector.

Proposed Methodology

The methodology for detecting insurance fraud using machine learning focuses on several stages, from data collection and preprocessing to model selection, training, and evaluation. The goal is to develop a system that accurately identifies fraudulent claims while handling challenges like class imbalance and missing values.

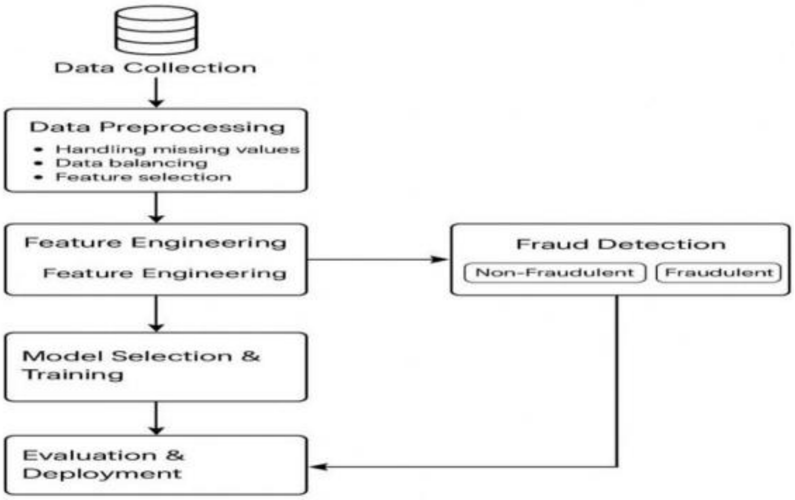


Figure.2. Fraud Detection System Architecture

Data Collection

The first step is to gather a dataset that contains both fraudulent and legitimate claims. In most real-world scenarios, datasets are often imbalanced, where fraudulent claims constitute a small proportion of total claims. The data is collected from an insurance company or a publicly available dataset, ensuring it includes relevant features like claim amount, claim history, policyholder details, and claim status

Pre-processing

Before model training, the data undergoes preprocessing, which includes:

Handling Missing Values: Missing data can lead to inaccurate predictions. Common strategies to address this include: Imputation (e.g., using mean, median, or KNN imputation) to replace missing values in numeric fields. For categorical data, missing values are either imputed using the mode or handled by marking them as a separate category.

Handling Class Imbalance

To address the class imbalance, the following techniques are applied:

SMOTE (Synthetic Minority Over-sampling Technique): This technique generates synthetic instances of the minority class (fraudulent claims) by interpolating between existing instances.

Random Oversampling: The minority class is oversampled by duplicating instances of fraudulent claims.

Cost-sensitive Learning: Machine learning algorithms are modified to assign higher penalties for misclassifying the minority class to ensure that the model learns to identify fraud more effectively.

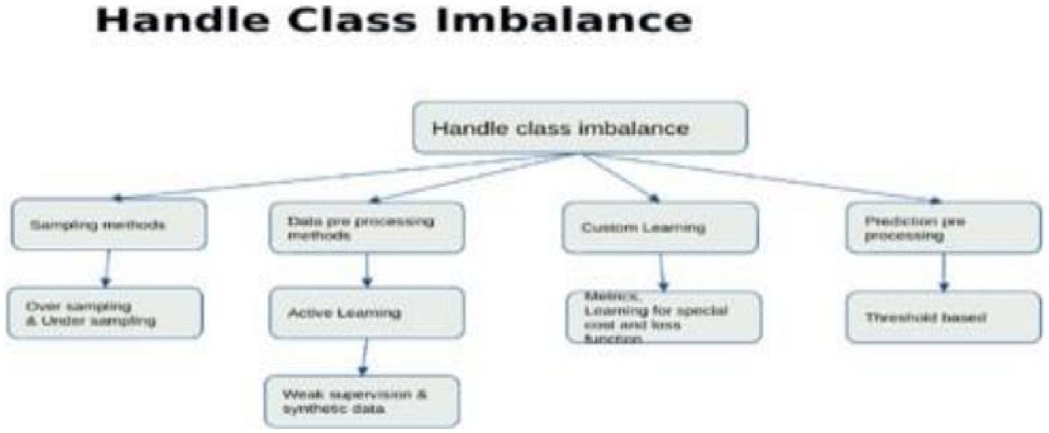

Handle Class Imbalance
Handle class imbalance
Sampling methods
Data pre processing methods
Custom Learning
Prediction pre processing
Over sampling & Under sampling
Active Learning
Metrics, Learning for special cost and loss function
Threshold based
Weak supervision & synthetic data

Fig.3. Handling class imbalance

Model Selection & Training

Several machine learning models are trained to detect fraudulent claims. The selection of models is based on their ability to handle imbalanced data and their suitability for the task.

Common models include: Logistic Regression: A simple, interpretable algorithm often used for binary classification tasks.

Random Forests: An ensemble method that combines multiple decision trees and is robust to overfitting.

XGBoost: A gradient boosting algorithm that excels in handling large datasets and imbalanced classes.

Support Vector Machines (SVM): Used for classification tasks with high-dimensional feature spaces. The models are trained on the preprocessed data using cross-validation to ensure generalization and prevent overfitting.

Evaluation Metrics

Class balance along with anticipated outcomes are just two of the many factors that go into choosing the optimal metrics for assessing a classifier's performance in a specific set of data in classification challenges. A classifier may be evaluated on one performance parameter while being unmeasured by the others, and vice versa. As a result, the generic assessment of performance of the classifier lacks a defined, unified metric. This study uses a number of metrics, including F1 score, accuracy, precision, recall, and recall, to assess how well models perform. The subsequent four categories are where these metrics are derived from: True Positives (TP): instances in which both the model prediction and the actual class of the occurrence were 1 (True). False Positives (FP) are situations in which the model predicts a value of 1 (True), but the actual class of the occurrence was 0 (False). True Negatives (TN): an instance in which both the model prediction and the true class of the occurrence were 0 (False). False Negatives (FN) are situations in which the model predicts 0 (False) but the true class of the occurrence was 1 (True).

Accuracy– The mean amount of accurate predictions is used to characterize the accuracy measure. This isn't quite as strong, though, given the imbalanced sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Precision, also known as positive predictive value, gauges the capacity of a model to pinpoint the right examples for every class. For multi-class classification with unbalanced datasets, this is a powerful matrix.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

Recall – This metric assesses how well a model detects the true positive among all instances of true positives.

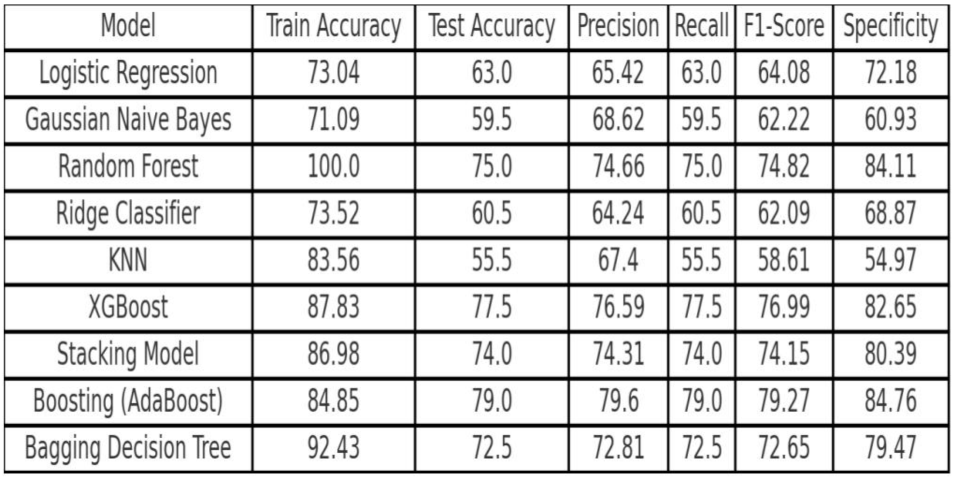
$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

F1-score – referred to as an F-measure or balanced F-score It might be characterized as a recall as well as precision weighted average.

$$F1_{Score} = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (4)$$

Results & Analysis

Evaluation criteria



Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	Specificity
Logistic Regression	73.04	63.0	65.42	63.0	64.08	72.18
Gaussian Naive Bayes	71.09	59.5	68.62	59.5	62.22	60.93
Random Forest	100.0	75.0	74.66	75.0	74.82	84.11
Ridge Classifier	73.52	60.5	64.24	60.5	62.09	68.87
KNN	83.56	55.5	67.4	55.5	58.61	54.97
XGBoost	87.83	77.5	76.59	77.5	76.99	82.65
Stacking Model	86.98	74.0	74.31	74.0	74.15	80.39
Boosting (AdaBoost)	84.85	79.0	79.6	79.0	79.27	84.76
Bagging Decision Tree	92.43	72.5	72.81	72.5	72.65	79.47

Conclusion

The study on insurance fraud detection using machine learning models demonstrated that ensemble techniques, particularly Boosting (AdaBoost), yielded the best performance. AdaBoost achieved the highest test accuracy of 79%, with a balanced precision, recall, and F1-score, making it the most reliable model for fraud detection. Random Forest exhibited overfitting with perfect training accuracy but lower generalization on test data. XGBoost and Stacking models also performed well, reinforcing the effectiveness of ensemble learning. Simpler models like Logistic Regression and Naïve Bayes struggled with accuracy, while K-Nearest Neighbors performed the worst due to its sensitivity to high-dimensional data. The study highlights the importance of hyperparameter tuning, feature engineering, and cross-validation in improving fraud detection accuracy. Future work can focus on integrating deep learning models, refining feature selection techniques, and deploying the model in realworld fraud detection systems to enhance predictive capabilities.

Future Scope

Deep Learning Integration : Advanced deep learning models, such as neural networks and transformers, can be utilized to capture complex fraud patterns, improving detection accuracy and adaptability.

Automated Feature Engineering : Techniques like deep feature extraction, reinforcement learning, and automated feature selection can enhance model performance by identifying the most relevant fraud indicators.

Handling Class Imbalance : Implementing sophisticated resampling strategies, anomaly detection methods, and cost-sensitive learning can help address the challenges of highly imbalanced fraud datasets.

Real-Time Fraud Detection : Deploying the model in a real-world fraud detection system with continuous learning mechanisms will allow it to adapt dynamically to evolving fraud trends.

Industry Collaboration & Explainability : Working closely with insurance companies to validate models on live datasets and incorporating explainable AI solutions will improve transparency, trust, and regulatory compliance in fraud prevention

References

1. Breiman, L. (2001). "Random forests." Machine Learning, 45(1), 5-32.
2. Freund, Y., & Schapire, R. E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of Computer and System Sciences, 55(1), 119-139.
3. Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785- 794.
4. He, H., & Garcia, E. A. (2009). "Learning from imbalanced data." IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

5. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). "Calibrating probability with under sampling for unbalanced classification." 2015 IEEE Symposium Series on Computational Intelligence (SSCI), 159-166.
6. Lemieux, T. (2019). "Financial fraud detection using machine learning techniques." Journal of Finance and Data Science, 5(1), 1-18.
7. Zhang, Y., & Zhou, X. (2021). "Handling missing data in machine learning: A survey." Applied Intelligence, 51(7), 4772-4791.
8. Ramesh Chandra AdityaKomperla, "CAR INSURANCE RISK PREDICTION USING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING" 2022 | Global Journal of Advanced Engineering Technologies and Sciences, 9(4), 1-8. https://doi.org/10.29121/gjaets.2022.04.01
9. Ramesh Chandra AdityaKomperla , "OPTIMIZING HEALTHCARE INSURANCE PREDICTIONS THROUGH DEEP LEARNING AND SCO-ENHANCED HYPERPARAMETER TUNING". (2022). Global Journal of Advanced Engineering Technologies and Sciences, 9(11), 1-9. https://doi.org/10.29121/gjaets.2022.11.01