International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 03 | Mar 2024 www.irjet.net p-ISSN: 2395-0072

Fraud Detection and Analysis for Insurance Claims Using Machine Learning

[1]Jaya Vani Vankara, [2]V Seshadri Naidu, [3]D Govardhan, [4]V Vivek , [5]P VNikhil

[1]Assistant Professor, Dept. of CSE, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.
[2,3,4,5]Student, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.

Abstract - *Insurance claim fraud is a serious issue that the insurance business faces. It costs insurance companies money and raises policyholders' rates. Machine learning has become a potential method for insurance claim fraud investigation and detection in recent years. Machine learning algorithms—such as data mining and deep learning techniques—have been effectively applied to identify trends and abnormalities in insurance claim data that point to fraudulent activity. These algorithms could significantly increase the precision and effectiveness of fraud detection, saving insurance firms money and giving consumers excellent safety. Machine learning algorithms can precisely analyze vast volumes of data and spot trends and abnormalities that can point to insurance claim fraud. These algorithms can look at various factors in claim data, including claim type, policyholder details, and past claim history, to identify anomalies or suspicious trends. With the help of machine learning, insurance companies will be able to build predictive models that will give each claim a Fraud Probability Score (FPS). In this project, we're focusing on identifying auto insurance fraud using machine learning. An insurance agent should be able to investigate every case and determine if it's real. But this not only takes time, but it's also expensive. Hiring and financing the skilled labor needed to review every claim filed daily is impossible. This is where machine learning comes in. In this case, we will use one of the most widely used machine learning algorithms.*

Key Words: **Fraud Insurance, XGBoost, Artificial Neural Network, Random Forest, Logistic Regression, Decision Tree, SVC.**

1. INTRODUCTION

We have discovered a significant issue with insurance fraud in this project. Claims filed to deceive an insurance company are known as false coverage claims. Since the beginning of the insurance sector, there has been a persistent problem with insurance fraud, with a significant portion of received claims being fake. Insurance firms suffer financial losses from fraudulent claims, and policyholders pay higher premiums. Machine learning algorithms, which use data mining and deep learning techniques, help identify patterns and anomalies in insurance claim data that may be signs of fraudulent conduct. These algorithms have the potential to increase the accuracy significantly. These advanced methods present the insurance sector with a viable way to improve fraud detection and lessen the effects of fraudulent claims. Insurance companies could save money, and consumers would feel safer if these algorithms significantly improved the accuracy and efficacy of fraud detection. These algorithms analyze various aspects of claim data, such as the kind of claim, policyholder information, and prior claim history, to spot abnormalities or questionable patterns. Insurance firms can create predictive models that use machine learning to assign a Fraud Probability Score (FPS) to each claim. This research primarily focuses on using machine learning to detect fraud with auto insurance.

We have chosen one of the most popular machine learning methods to do this. These algorithms produced the highest accuracy in projected results and annual expenses, amounting to billions of dollars, proving their applicability to our dataset. Insurance fraud can take many forms in different insurance realms, and it can range in severity from small-scale claim embellishment to deliberate acts of destruction or harm. Auto insurance fraud is one of insurers' most significant and well-known problems. A claims agent should look at costs due to fraudulent claims, which highlights the significance of differentiating between genuine and fraudulent claims. Although a claims agent should look into each case separately, this is frequently an expensive, time-consuming, and inefficient procedure. Examining all of the many claims that are filed every day would be very impossible. To detect and mitigate fraudulent claims, machine learning offers a practical, quick, and economical solution.

2. Literature Survey

[1] T. Badriyah, Lailul Rahmanian I. Syarif, titled 'Nearest Neighbour and Statistics Method based for Detecting Fraud m Auto Insurance" provides an overview of the nearest

neighbor method and interquartile method to detect fraud in car insurance data. The results of the conducted study, the best result using the paper, concluded that the Feature selection process improves fraud detection accuracy. The distance-based algorithm yields the best fraud detection results. - Performance measurement is superior in some cases. - The best result of fraud detection is using a distance-based algorithm.

[2] Xi Liu, Jian-Bo Yang, Dong-Ling Xu, Karim Derrick, Chris Stubbs, and Martin Stockdale titled "Automobile Insurance Fraud Detection using Evidential Reasoning Approach and Data-driven Inferential Modeling." Automobile insurance fraud detection has become crucial for lowering insurance companies' prices. Expenence-based knowledge is interpretable and reusable, but the simplistic way this knowledge is used in practice often leads to misjudgment. This paper proposes to set up a unique Evidential Reasoning (ER) rule that mixes impartial proof from each reveal primarily based on total signs and chances of fraud received from historical data Each piece of evidence is weighted and then combined conjunctively with the tights optimized using a maximum likelihood evidential reasoning (MAKER)framework for data-driven inferential modeling.

[3] K. Supraja and S.J. Saritha, titled 'Robust Fuzzy rule-based techniques to detect frauds in insurance," provide an overview of The Fuzzy Rule-based technique applied to the training data set and based on the instances of the degree of fraud or legally predicted- Concluded that this technique is used for high dimensional large datasets with accuracy. The paper discusses fuzzy rule-based techniques to improve fraud detection in vehicle insurance. The limitation of the paper mentioned is that Bayesian visualization is unsuitable for abundant data- Fuzzy Logic used to improve Fraud Detection.

[4] Richard A. Bauder and Taghi M. Khoshgoftaar, titled "Medicare Fraud Detection Using Machine Learning Methods," provides insights into the paper comparing different machine learning methods for detecting Medicare fraud and finding that supervised methods perform better than supervised or hybrid methods. This paper mentions supervised, unsupervised, and hybrid machine learning methods, as well as class imbalance reduction via oversampling and 80-20 undersampling methods.

The paper by [5] Machinya Tongesai, Godfrey Mbizo, and Kudakwashe Zvarevashe titled "Insurance Fraud Detection using Machine Learning proposes an automated fraud detection application framework using machine learning and the XGBoost method to accurately identify fraudulent insurance claims in a shorter amount of time. The paper concludes that machine learning and data analysis is an automated fraud detection framework that can help insurance companies accurately identify fraudulent claims quickly, improving the traditional claim investigation process and resulting in financial savings and a positive impact on society.

3. Project Goal

To successfully create a model that uses ML Algorithms that ultimately aid in detecting fraud insurance claims effectively and efficiently and help the insurance industries.

4. Problem Identification

Globally, there are over a thousand companies in the insurance market, and they gather premiums of over one trillion dollars annually. The most common insurance fraud is fabricating accident claims, which may be done with a vehicle. Utilizing machine learning techniques, this research aims to detect car insurance fraud.

5. Dataset Description

The dataset selected for this specific task was extracted from an online source named Kaggle, which contains nearly 1000 rows of historical data and 39 columns containing various criteria for the project.

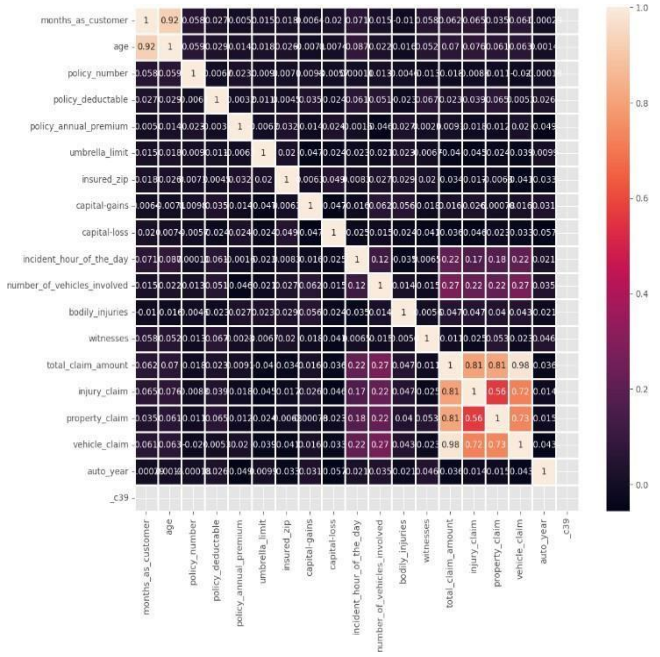

Figure -1: Heatmap of the dataset

6. Proposed Method

6.1 Data collection:

Gathering data for training a machine learning model is the initial stage in a machine learning pipeline. Data collection involves gathering information from diverse sources to address pertinent questions. The accuracy of the predictions generated by a model is inherently tied to the quality of the data used for its training. Challenges during data collection include unreliable data, missing data, and data imbalance. To mitigate these problems, data preprocessing is conducted on the collected data.

6.2 Data Preprocessing:

Because they are fragmentary, inconsistent, and devoid of specific behavioral patterns, raw data from the real world are probably not dependable. Thus, pre-processing is done after data collection to clean up the data and prepare it for machine-learning model construction.

6.3 Data cleaning:

Erroneously added or classified data can be removed manually or automatically. Imputation of data: Standard deviation, mean, and median are three common methods most machine-learning systems use to balance or fill in missing values.

6.4 Oversampling:

Biased or unbalanced datasets can be added to the underrepresented classes after being corrected using strategies like oversampling and repetition. Exploratory

6.5 Data Analysis:

Exploratory Data Analysis, abbreviated as EDA, fundamentally serves as a form of storytelling that enables the revelation of concealed insights and patterns, the identification of outliers and anomalies, and the validation of underlying structures.

6.6 Clustering:

To achieve the goal of having data points in the same group be more similar to each other and less different from the data points in other groups, clustering is the process of dividing the population or data points into many groups. In this case, the anticipated value is subtracted from the measured value concerning the optimal line to find the residue.

7. Methodology

1. In the initial step, we import the packages
2. Upload the CSV file (data sets)
3. We can use functions like describe and info functions tothe dataset
4. Data pre-processing
   i. Missing values (visualizing the missing values)
   ii. Handling missing values
5. Displaying correlation between features
6. Dropping the columns which are not necessary for prediction
7. Masking Multicollinearity
8. Separating the feature and target columns
9. Encoding Categorical columns
   - ➢ Extracting categorical columns
   - ➢ Drop Dummies
   - ➢ Extracting the numerical columns
   - ➢ Combining the Numerical and Categorical data frames to get the final dataset
10. Outliers Detection
11. Splitting data into training sets and test sets.
12. Scaling the numeric values in the dataset
13. Models creation and training
14. Model Performance Comparison

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$ERR = \frac{FP + PN}{P + N}$$

Fig - 2: Evaluation Metrics

Model	Recall	Precision	F1 score	Accuracy	Error rate
Random Forest	0.920000	0.900000	0.900000	0.915000	0.085000
SVC	0.860000	0.670000	0.750000	0.865000	0.135000
XGBoost	0.830000	0.840000	0.860000	0.835000	0.165000
Neural Network	0.750000	0.720000	0.720000	0.750000	0.250000
Decision Tree	0.550000	0.730000	0.730000	0.550000	0.450000
Logicstic Regression	0.470000	0.630000	0.630000	0.470000	0.530000

Fig-3: Comparison of metrics

8. CONCLUSIONS

Compared and contrasted various deep learning and machine learning models for fraud insurance detection using SVM, Random Forest, XGBoost, Decision Tree, Logistics Regression, and Neural Networks.

Out of all the models, Random Forest showed the highest accuracy, i.e., 91%, with an error rate of 0.08.

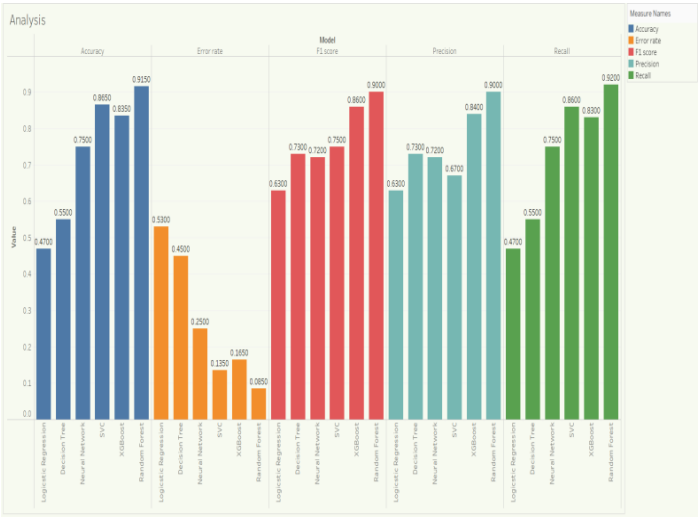

Fig -4: Graphical Analysis

9. Future scope

Future research directions are indicated, emphasizing enhancing the system's functionality, accuracy, and efficiency and adding more advanced features to the project so that consumers or policyholders can access it more conveniently and dependably.

10. REFERENCES

[1] X. Liu, J.-B. Yang, D.-L. Xu, K. Derrick, C. Stubbs, and M. Stockdale, "Automobile Insurance Fraud Detection using the Evidential Reasoning Approach and Data-Driven Inferential Modelling," 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul. 2020.

[2] Robust fuzzy rule-based technique to detect vehicle insurance fraud, K.Supraja, S.J. Saritha,01 Aug 2017.

[3] I.Sadgalian, Saelaf & Benabboua (2019). Performance of machine learning techniques in the detection of financial frauds.

[4] Nearest Neighbour and Statistics Method Based for Detecting Fraud in Auto Insurance Tessy Badriyah, Lailul Rahmaniah, Iwan Syar, 01 Oct 2018.

[5] Medicare Fraud Detection Using Machine Learning Methods Richard A. Bauder, Taghi M. Khoshgofta, 01 Dec 2017.

[6] Insurance Fraud Detection Using Machine Learning Machinya Tongesai, Godfrey Mbizo, Kudakwashe Zvarevashe, 09 Nov 2022.

[7] S. Subudhi and S. Panigrahi, "Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques," International Journal of Rough Sets and Data Analysis, vol. 5, no. 3, pp. 1–20, Jul. 2018.

[8] V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697476.

[9] Bart Baesens, S. H. (2021). Data engineering for fraud detection, Decision Support Systems.

[10] S. Ray, "A Quick Review of Machine Learning Algorithms," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com. 2019, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.