JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detecting Auto Insurance Fraud Through MachineLearning Approaches

[1]Himanth S J, [2]Kalyan Venkatesh K J, [3]Harish, [4]Amogh C, [5]Shabana Sultana

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor
[1]Department of Computer Science and Engineering,
[1]The National Institute of Engineering, Mysuru, India.

Abstract : The global insurance industry comprises a vast number of companies, exceeding a thousand in number. Currently, the industry is actively adopting an efficient fraud management system.

We propose a solution for the insurance organizations where the insurance company agents can track the predicted fraud percentages of the claims made.The insurance agent has the provision to fill up the form with the details related to the claim. The pre-trained machine learning model utilizes the form data as input to determine the authenticity of the claim, predicting whether it is fraudulent or legitimate. A two way classification of the claim is brought about by the model as to whether the claim is a fraud or a not fraud. The insurance companies can use the output to assist them in deciding whether the claim is fraud or not. The machine learning model is structured on the Decision Tree algorithm.

***IndexTerms* - Machine Learning; Decision Tree Algorithm; Insurance fraud detection.**

I. INTRODUCTION

Having a robust fraud detection and prevention management system in place is an effective approach for analyzing data based on specific instructions within a company. With the availability of various methods for detecting fraudulent claims, the process becomes complex and requires careful consideration.

The insurance industry comprises thousands of companies worldwide, collecting over one trillion dollars in premiums annually. Insurance fraud occurs when individuals or entities make false claims to receive undeserved compensation or benefits. The estimated cost of insurance fraud exceeds forty billion dollars, making the detection of such fraud a significant challenge for the insurance industry.

The insurance industry is increasingly adopting efficient fraud management practices. While some individuals deceive insurance companies to obtain undeserved compensation, others genuinely pay their premiums.Insurance fraud can be classified into two primary categories: hard insurance fraud and soft insurance fraud. Hard insurance fraud refers to intentional acts where individuals deliberately stage accidents. On the other hand, soft insurance fraud involves falsifying certain aspects of a legitimate insurance claim. This differentiation allows for the distinction between fraudulent activities involving staged accidents and those involving falsified details within valid claims.

Having a robust fraud detection and prevention management system is an effective approach for analyzing data according to the company's specific instructions. With numerous methods available for detecting fraud claims, the process often involves complex and time-consuming investigations that span different domains of knowledge. However, the use of machine learning techniques can help overcome these challenges. By leveraging machine learning, companies can address the complexities and time constraints associated with fraud detection. Machine learning algorithms possess the capability to analyze vast amounts of data, detect patterns, and highlight anomalies that may indicate potential fraud. By employing this automated approach, the dependence on manual investigations is reduced, leading to significant time and resource savings. By training machine learning models on historical data and continually updating them, companies can enhance their fraud detection capabilities. These models learn from past fraudulent patterns and can swiftly identify potential fraud in real-time. In summary, machine learning techniques offer a solution to overcome the difficulties in fraud detection, enabling efficient analysis of data and strengthening a company's overall fraud management efforts.

II. LITERATURE SURVEY

Rama Devi Burri et al., in their study explored various machine learning techniques for efficient analysis of insurance claims. They also discussed three approaches to integrate machine learning into the insurance industry. Furthermore, they identified several obstacles to adopting machine learning for claim classification and highlighted the challenges involved in implementing machine learning solutions. The researchers conducted a performance evaluation of different algorithms for claim predictions.

Pinak Patel et al. introduced a fraud detection framework specifically designed for the healthcare industry. The framework categorizes fraudulent behavior into two main types: period-based claim anomalies and disease-based anomalies. The researchers evaluated their framework using real-world medical data and achieved efficient results in identifying fraudulent claims.

Najmeddine Dhieb et al. conducted an evaluation of the XGBoost algorithm's performance in detecting and classifying various types of auto insurance fraud claims. They compared this proposed algorithm with other cutting-edge solutions in the field. Additionally, the researchers assessed the algorithms using data analysis and exploration techniques, measuring them against multiple metrics.

Sunita Mall et al. presented a study aimed at identifying significant triggers of fraud and predicting fraudulent behavior among customers by leveraging the identified triggers. The researchers employed statistical techniques to identify and predict these triggers effectively.

Shimin LEI et al. introduced a financial fraud detection system based on XGBoost. The system was divided into two components: an automatic part and a manual part. The automatic part utilized a large database to train the model, while the manual part was responsible for monitoring transactions. To make the final decision, the researchers combined machine scoring with manual feedback, integrating both aspects into the detection process.

III. METHODOLOGY

Machine Learning, a subset of Artificial Intelligence, involves training models without human intervention. It encompasses various techniques, such as Supervised, Unsupervised, and Semi-Supervised learning. In the context of data analysis, data sets are often stored in .CSV files. The data cleaning process is applied, which involves removing and filling missing values with column mean values. This step ensures data consistency and eliminates duplicate entries. Subsequently, the data set is split into training and testing data, allowing the model to be trained using the training data. Once the training process is complete, the model becomes capable of making predictions for the testing data.

Our primary goal is to train the model using pre-processed data to achieve optimal performance. To accomplish this, we utilize a Supervised classification algorithm called the Decision Tree algorithm. This algorithm is particularly well-suited for scenarios where the outcome values are discrete, such as classifying results into categories like "fraud" or "not fraud".

Fig.1. Methodology shows how the above process has been carried out.

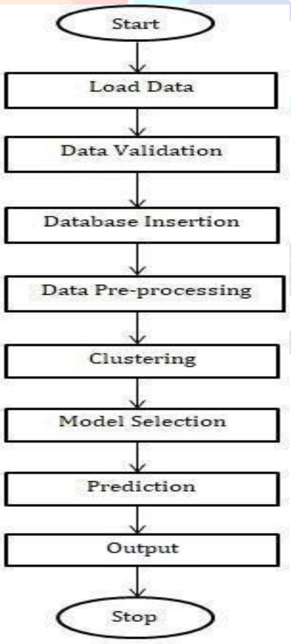


Fig. 1. Methodology

The decision tree construction process employs top-down recursive and divide-and-conquer methods. Starting with the root node, the tree represents the entire training data set.

If all samples in the training data set have the same outcome, a leaf node is created, and it is labeled with that particular class. However, if the samples have different outcomes, the decision tree selects the attribute that provides the most information and uses it to divide the data set. The node representing this attribute is then labeled with the attribute's name. These steps are repeated recursively until all samples belong to the same class or there are no more samples or attributes to split on.

IV. Implementation

We have developed a browser-based application for insurance agents to conduct surveys and input parameters into a form, resulting in a fraud or non-fraud determination.

To begin, we collected, analyzed, and pre-processed a data set comprising various parameters such as age, educational level, gender, incident type, collision type, witnesses, incident severity, premium amount, property claim, injury claim, and more. These parameters were encoded into numerical values, which were then used to train the model.

For fraud prediction, we employed a supervised learning machine learning approach called the "Decision Tree Algorithm." This algorithm is known for its ability to provide accurate results.

In this application we have user, system/application and the data set. The user refers to the insurance agent who fills the form survey.

Fig. 2. Sequence Diagram shows how the application interacts with the data set provided to train the machine learning model.

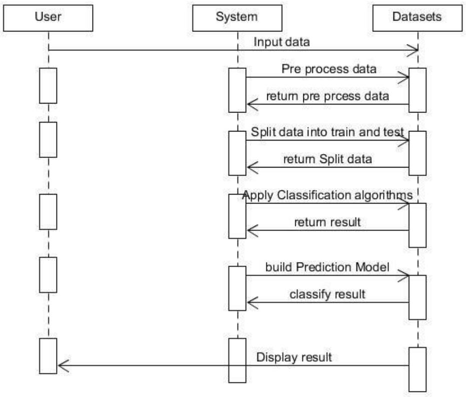


Fig. 2. Sequence Diagram

The machine learning method follows the following steps as depicted in Fig. 3. Activity Diagram :

1. Collection of data
2. Cleaning and pre-processing of the data
3. Splitting the data into training and testing sets
4. Implementation of the algorithm in the model
5. Testing the model for accuracy and precision.

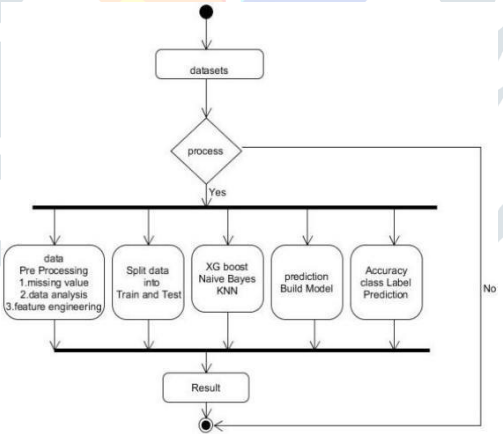


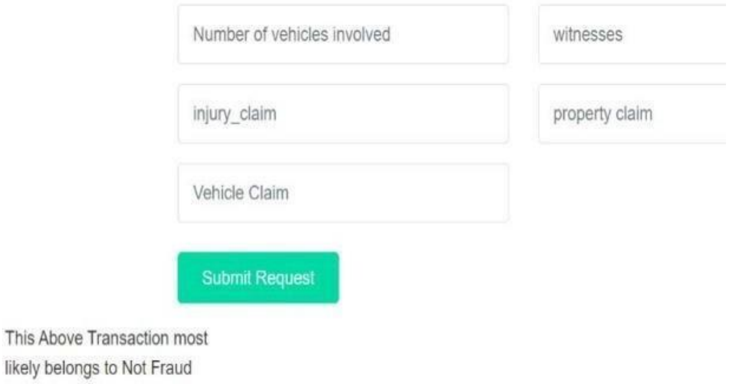
Fig. 3. Activity Diagram

The Fig. 4. Parameters List diagram lists and describes the parameters taken via the form survey.

Fig. 4. Parameters List

Sl. No.	Factor	Encoded Value
1	Age	Numerical value
2	Insurer Gender	1. Male 2. Female
3	Incident Type	1. Parked car 2. Vehicle Theft 3. Multi Vehicle Collision 4. Single Vehicle Collision
4	Collision Type	1. Front Collision 2. Rear Collision 3. Side Collision
5	Incident Severity	1. Major Damage 2. Trivial Damage 3. Minor Damage 4. Total Loss

The Fig. 5. Not fraud shows that the insurance claim is classified as “Not fraud”.



Number of vehicles involved

witnesses

injury_claim

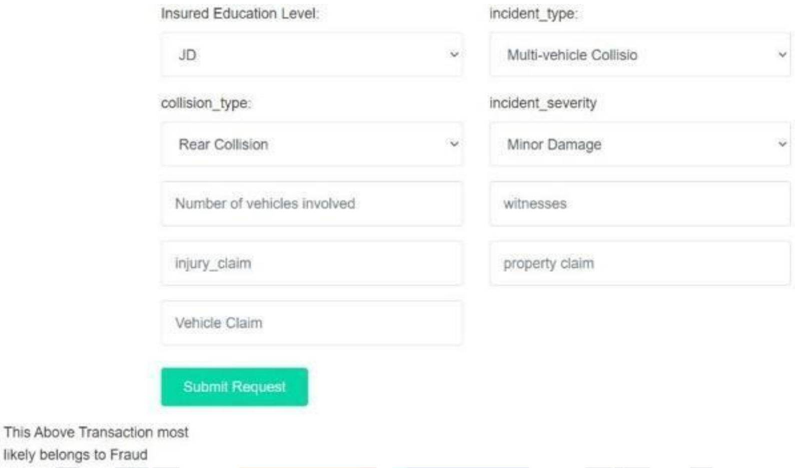
property claim

Vehicle Claim

This Above Transaction most likely belongs to Not Fraud

Fig. 5. Not fraud

The Fig. 6.Fraud shows that the insurance claim is classified as “fraud”.



Insured Education Level: JD

incident_type: Multi-vehicle Collisio

collision_type: Rear Collision

incident_severity Minor Damage

Number of vehicles involved

witnesses

injury_claim

property claim

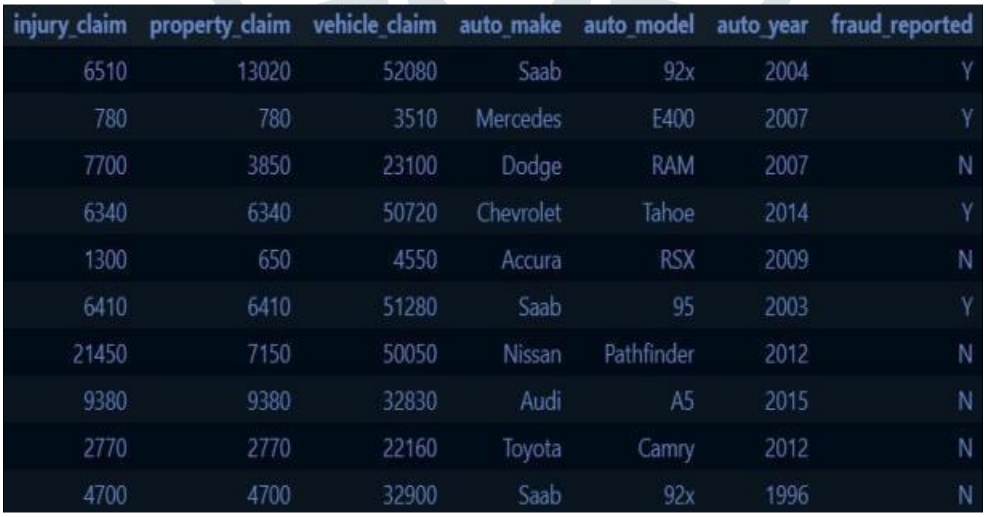
Vehicle Claim

This Above Transaction most likely belongs to Fraud

Fig. 6. Fraud

V. Result and Analysis

Fig. 7 Training data set illustrates the training data set, which includes previously collected data on claims made. Each data point in the data set has been encoded as an integer value corresponding to one of the 40 labels. The last column of the data set represents the outcome label.



injury_claim	property_claim	vehicle_claim	auto_make	auto_model	auto_year	fraud_reported
6510	13020	52080	Saab	92x	2004	Y
780	780	3510	Mercedes	E400	2007	Y
7700	3850	23100	Dodge	RAM	2007	N
6340	6340	50720	Chevrolet	Tahoe	2014	Y
1300	650	4550	Accura	RSX	2009	N
6410	6410	51280	Saab	95	2003	Y
21450	7150	50050	Nissan	Pathfinder	2012	N
9380	9380	32830	Audi	A5	2015	N
2770	2770	22160	Toyota	Camry	2012	N
4700	4700	32900	Saab	92x	1996	N

Fig. 7.Training data set

MODEL ACCURACY

Fig. 8. "Decision Tree Efficiency," presents the performance evaluation of the Decision Tree algorithm. The original data set is divided into training data (80%) and testing data (20%). The algorithm assesses its accuracy by comparing the predicted outputs of the testing data set with the original data set. The accuracy of the model improves as the number of correctly predicted records increases.

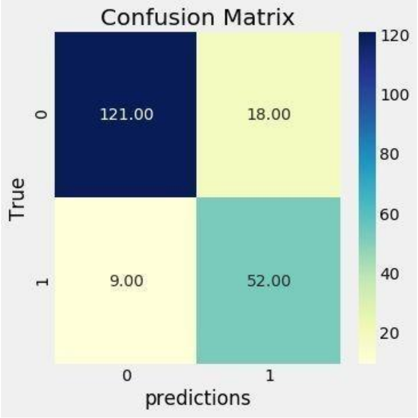


Fig. 8. Decision Tree efficiency

Fig. 9. "Graphical Analysis of Tested Data," provides a visual representation of the fraud claims observed in the testing data set. This graphical analysis presents a tabulated summary of the fraudulent claims within the testing data set.

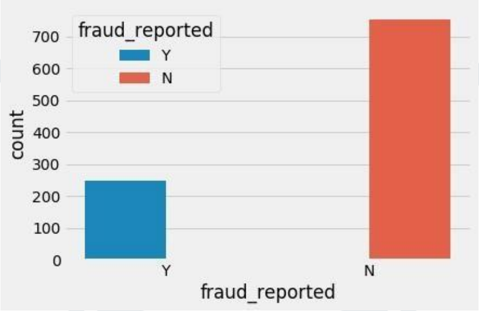


Fig. 9. Graphical Analysis of Tested Data

The tabulated data indicates that out of 1000 claims, 753 were reported as "Not Fraud," while 247 claims were reported as "Fraud”.

VI. Conclusion and Future Enhancements

Insurance fraud detection is a complex and challenging task that the insurance industry has faced since its inception. The proposed system aims to develop a solution that can effectively minimize potential frauds with a high level of accuracy.

In this project, the Decision Tree algorithm has demonstrated promising results in detecting vehicle fraud. The proposed machine learning model utilizing the Decision Tree algorithm achieves an accuracy rate of 86.5%.

The system has the potential for scalability to predict insurance fraud in bulk. Enhancements can be made by incorporating a feedback module to gather additional information and improve the accuracy of fraud prediction. Furthermore, an upgrade can be implemented by adding a report module, which enables the branch in-charge to generate a complete report by clicking on the report button in cases where the output indicates fraud. This report can then be promptly sent to the nearby police station for further investigation.

References

[1] Rama Devi Burri et al, 2021 presented several machine learning techniques to analysis insurance claims efficiently. They also mentioned three ways to transform machine learning techniques into insurance industry. Additionally they specified different resistances for adapting machine learning to classify claims and challenges in implementing machine learning. Also they evaluated the performance of different algorithms for claim predictions.

[2] Jing Li, JianjunShi Jionghua Jin and Kuei-Ying Huang, May 2009, “A survey on statistical methods for health care fraud detection”, Springer Science.

[3] Pinak Patel et al proposed a fraud detection framework for health care industry. They classified the fraudulent behaviour in two categories period based claim anomalies and disease based anomalies. Their framework was evaluated on real world medical data which showed efficient results to determine fraud claims.

[4] K. Branting, T. Champney and F. Reeder, J. Gold, 2016 ,“ Graph Analytics for Health care Fraud Risk Estimation”, pp. 845-851, ICTAI.

[5] Travaille, Dallas Thornton, Roland M Muller and Jos Van Hillegersberg, 2011, “Electronic fraud detection in U.S”., Proc.7th Americans conference on information systems, pp.1-10.

[6] “Management of Fraud: Case of an Indian Insurance Company” – Sunita Mall et all, Accounting and Finance Research 2018 http://www.sciedu.ca/journal/index.php/%20afr/a rticle/download/13474/8333.