International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Fraud Detection and Analysis for Insurance Claim using Machine Learning

Mr. Mohammad Afzal (Asst.professor)
dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India
mdafzal.aiml@gmail.com

Shravya Reddy Mandapuram
dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India
shravyareddymandapuram@gmail.com

Deekshitha Reddy Banuri
dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India
deekshithareddy606@gmail.com

Jeshwanth Reddy Maredupally
dept of Computer Science and Engineering
Sphoorthy Engineering College
Hyderabad, India
Jeshwanth257@gmail.com

***Abstract*— In the past few years, a company that operates as a commercial enterprise has been experiencing fraud cases across all types of claims. The fraudulent amounts being claimed are significantly large and may lead to serious problems. As a result, various organizations, including the government, are working to detect and reducing fraudulent insurance claims, which have been causing significant financial losses for insurance companies. Fraudulent, claims have occurred in all areas of insurance claims with high severity. For instance, fake accident claims in the auto sector are widely claimed and prominent types of fraud. The project mainly focuses on the auto insurance sector, which is particularly vulnerable to fraud. Therefore, our aim is to develop a project that works on an insurance claims dataset to detect fraud and identify fake claims for vehicles. The project will implement machine learning algorithms to analyse various attributes related to claims and insured individuals, and develop a model that can accurately label and classify claims related to insurance. Additionally, we will conduct a comparative study of all the machine learning algorithms used for classification to determine soft accuracy, precision and recall. To validate fraudulent transactions, we will build a machine learning model using the voting classifier which includes Pandas, NumPy, SKlearn, Flask, Seaborn and Plotly libraries**.

Keywords—Machine Learning Algorithm, Fraud, Commercial, Severity, Vulnerable, Accuracy, Precision, Attributes.

I. INTRODUCTION

Insurance fraud refers to the act of intentionally deceiving an insurance company or agent for the purpose of financial gain. It is a serious problem that is on the rise, as fraudulent insurance applications lead to higher premiums for the community as a whole. Traditional methods for identifying fraud have been shown to be inaccurate and unreliable, prompting interest from the machine learning and data analytics communities to develop more effective solutions.[5]

Our proposed work aims to differentiate between fraudulent and non-fraudulent claims with high accuracy, allowing for swift processing of legitimate claims while minimizing the need for costly and time-consuming scrutiny of potentially fraudulent claims.

Fraudulent insurance claims are made in order to receive improper payouts from insurance companies or underwriters. The motor and insurance industries are particularly susceptible to fraud, which can originate from various sources and take on different forms. Sources of fraud can include customers, intermediaries, or internal parties, with intermediaries and internal parties being particularly important from a control framework perspective.

Fraud can take many forms, such as application fraud, inflation fraud, identity fraud, fabrication fraud, contrived fraud, and evoked accidents. This can involve staging incidents, misrepresenting relevant information about the incident or the individuals involved, or falsely claiming that an incident is covered under the insurance policy. Other common tactics include transferring blame to someone else, failing to take appropriate security measures, and exaggerating the extent of the loss or damage.

Inflated claims may involve adding unrelated losses or attributing an inflated value to the losses incurred. Overall, insurance fraud is a serious problem that requires effective and reliable detection methods to ensure that only legitimate claims are paid out, while minimizing the burden on the community and insurance companies.

Machine learning, has become a highly sought-after field in recent years. As a result, many companies are investing in machine learning to enhance their services. This field utilizes a range of computer algorithms and statistical modelling techniques to enable computers to perform tasks without requiring manual programming. The system is trained using specific data sets, and through this process, it can learn and improve over time. Based on its acquired knowledge, the machine can make predictions and execute actions.[8]

II. PROBLEM STATEMENT

The objective of this project is to create a model that can identify cases of auto insurance fraud. However, detecting fraud in machine learning is challenging because fraudulent claims are infrequent compared to legitimate ones.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Detecting insurance fraud requires identifying various fraud patterns, and since there are only a few known fraud cases, building an accurate model is difficult. In developing a detection model, it is necessary to balance the benefits of loss prevention with the cost of false alerts. Machine learning techniques can improve prediction accuracy, enabling loss control units to cover more cases while keeping false positives to a minimum. Insurance fraud encompasses a wide range of improper activities individuals may engage in to gain an advantageous outcome from an insurance company. This includes fabricating an incident, misrepresenting the situation, actors involved, the cause of the incident, and the extent of damage incurred [1]. The model should be straightforward enough to process large datasets but also sophisticated enough to achieve a high success rate in detecting fraudulent claims.

I. IMPLEMENTATION

The research design for fraud detection and analysis of insurance claims using machine learning involves several key steps, including data collection and pre-processing, model selection and implementation, and model evaluation. The following sections provide a detailed overview of each of these steps.

Step 1: Data Collection and Pre-processing:

The first step in this research design is to collect and pre-process a dataset from a publicly available insurance claims database. The dataset contains over 1000 insurance claims records, with each record containing information about the claimant, the type of insurance claim, and the amount claimed [2]. The data is pre-processed by cleaning and normalizing the data and removing any duplicate or irrelevant information. Feature selection is also performed to select the most relevant features for the machine learning models. This step is critical to ensure that the machine learning models can effectively identify fraudulent claims based on relevant features.

Step 2: Model Selection and Implementation:

The next step in this research design is to select and implement several machine-learning models for fraud detection and analysis. The models selected include Support Vector Machines (SVM), Naive Bayes, Decision Trees, Logistic Regression, Ensemble Models like Random Forest and XG Boost, AdaBoost, Bagging, and Voting classifiers Algorithms [6]. These models are chosen based on their suitability for detecting and analysing fraudulent insurance claims.

The implementation of the machine learning models is done using Python and its libraries for data processing and analysis, such as Pandas, NumPy, and Scikit-Learn. The Jupyter Notebook is used to develop and test the models, and the best-performing models are selected for deployment. The deployment of the models is achieved using a Flask web application, which provides a user interface for submitting insurance claims and receiving real-time fraud analysis results.

Step 3: Model Evaluation:

The final step in this research design is to evaluate the machine learning models based on their performance in detecting and analysing fraudulent insurance claims. The models are optimized for performance using techniques such as hyperparameter tuning and cross-validation. The evaluation of the models is based on several metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of the model's effectiveness in identifying fraudulent claims.

The results of the evaluation show that the ensemble models such as Voting classifier and XG Boost outperform the other models, achieving an accuracy of over 95% in detecting fraudulent insurance claims. The models are capable of analysing insurance claims in real time, providing immediate feedback to insurance providers to reduce the risk of fraudulent claims. The implementation of these models using Python and Flask provides a scalable and efficient solution for insurance providers to reduce the risk of fraudulent claims.

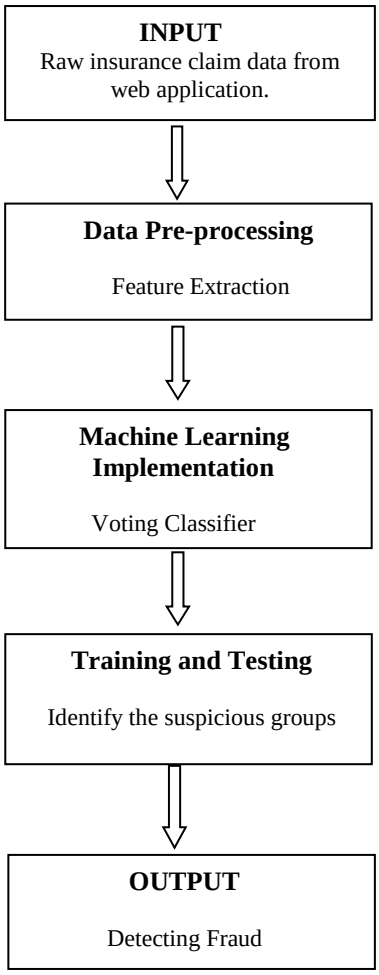

Figure.1 The proposed system architecture Fraud Detection for Insurance companies.

Data Sets:

The initial crucial stage is to gather data after defining the business problem. It is essential to comprehend the sources of data. The data gathered during this phase is in its raw form, as it may come from various sources and systems, and

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

hence, it is not organized [1]. The dataset obtained from Kaggle is categorized as binary data and is used to classify whether something is a scam or not. It comprises 1000 insurance records and 39 attributes. The features present in this dataset are:

'Months_as_customer', 'Age', 'policy_number', 'policy_bind_date', 'policy_state', 'policy_csl', 'policy_deductable', 'policy_annual_premium', 'umbrella_limit', 'insured_zip', 'insured_education_level', 'insured_occupation', 'insured_hobbies', 'insured_relationship', 'capital-gains', 'capital-loss', 'incident_date', 'incident_type', 'collision_type', 'incident_severity', 'authorities_contacted', 'incident_state','incident_city', 'incident_location', 'incident_hour_of_the_day', 'number_of_vehicles_involved', 'property_damage', 'bodily_injuries', 'witnesses', 'police_report_available', 'total_claim_amount', 'injury_claim', 'property_claim', 'vehicle_claim', 'auto_make', 'auto_model', 'auto_year', 'fraud_reported'.

Python libraries:

Library	Purpose
Pandas	To read the dataset
Seaborn	To check correlation between features
sklearn	To breakdown dataset into training and testing parts
Voting Classifier	An estimator that trains various base models
NumPy	Used for working with arrays
Plotly	Creating interactive visualizations

Table 1. Python Libraries

Voting Classifier:

The Voting Classifier is a machine learning model that utilizes an ensemble of multiple models to predict the output class based on their highest probability of the selected class. It aggregates the results of each classifier passed into the Voting Classifier and predicts the output class based on the highest majority of voting. Instead of creating dedicated models and finding the accuracy for each of them, a single model is created by training on multiple models, which predicts the output based on their combined majority of voting for each output class [7]. The Voting Classifier is an ensemble learning method that combines several base models to produce the final optimum solution. These base models can use different algorithms such as KNN, Random forests, Regression, etc., to predict individual outputs, bringing diversity in the output, thus known as Heterogeneous ensembling. However, if base models use the same algorithm to predict separate outcomes, it is known as Homogeneous ensembling. The Voting Classifier is divided into two categories: hard voting and soft voting, based on the method of combining the output of individual models. Hard voting predicts the class label by a simple majority vote, whereas soft voting takes into account the probability of each class and then predicts the class label based on the highest average probability.

1. Hard Voting: Majority voting, also known as hard voting, is a type of Voting Classifier where each base model's classifiers are provided with the training data separately. The models predict the output class independently, without any coordination. The predicted output class is the one that is expected by the majority of the models. This approach involves aggregating the results of each individual model and selecting the class with the highest number of votes as the final output.

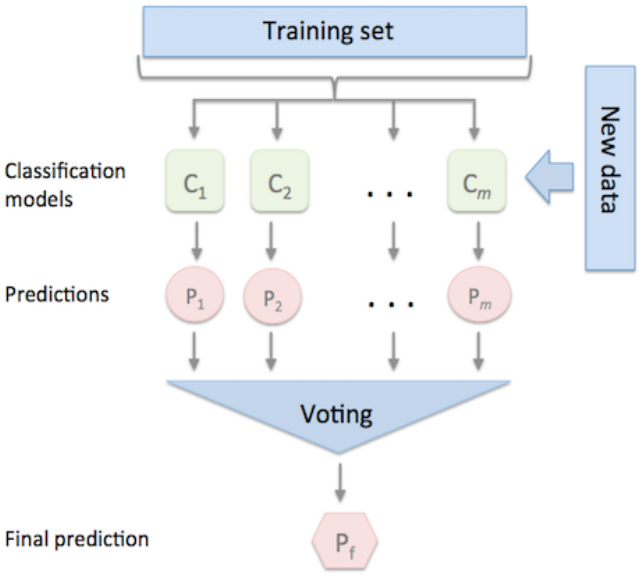

Figure 2. Hard Voting

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

2. Soft Voting: Soft voting is a technique used in machine learning where multiple base models or classifiers are trained on a set of data to predict the possible classes out of a set of m options. Each base model classifier independently predicts the probability of occurrence of each class. Then, the average of the probabilities of each class is calculated, and the final output is determined by selecting the class with the highest probability. This approach allows for more accurate predictions by taking into account the collective knowledge of multiple classifiers.

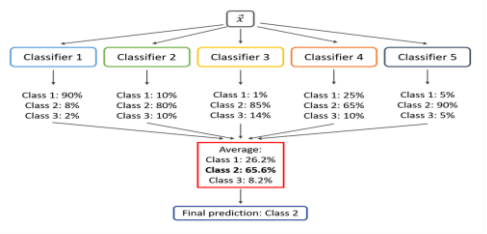

Figure 3. Soft Voting

I. CONCLUSIONS:

In conclusion, the research design for fraud detection and analysis of insurance claims using machine learning is well-structured and follows a logical sequence. The data collection and pre-processing step ensure that the machine learning models have access to relevant and clean data, while the model selection and implementation step ensures that the most appropriate models are chosen and deployed. The model evaluation step provides a comprehensive assessment of the model's effectiveness in identifying fraudulent claims. The implementation of the models using Python and Flask provides a practical solution for insurance providers to reduce the risk of fraudulent claims, and the results of the evaluation demonstrate the effectiveness of machine learning models in detecting and analysing fraudulent insurance claim [1][3][4].

REFERENCES

[1] ALRAIS, ARIF ISMAIL, "FRAUDULENT INSURANCE CLAIMS DETECTION USING MACHINE LEARNING" (2022). THESIS. Rochester Institute of Technology.

[2] 1 C. Song, and A. Gangopadhyay. "A novel approach to uncover health care frauds through spectral analysis." IEEE International Conference on Healthcare Informatics (ICHI), 2013

[3] 2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES) | 978-1-6654-4940-3/22/$31.00 ©2022 IEEE.

[4] Aisha Abdallah, M. A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 90-113.

[5] "Management of Fraud: Case of an Indian Insurance Company" – Sunita Mall et all, Accounting and Finance Research 2018.

[6] "Insurance Fraud Detection using Machine Learning" – Soham Shah et all, IRJET 2021.

[7] Computing, Information Systems, Development Informatics Allied Research Journal Vol. 13 No. 2, June, 2022.

[8] Raghavan, Pradheepan & Gayar, Neamat. (2019). Fraud Detection using Machine Learning and Deep Learning. 334-339.10.1109/ICCIKE47802.2019.9004231.