International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

Automobile Insurance Claim Fraud Detection

Dhruvang Gondalia[1], Omkar Gurav[2], Ameya Joshi[3], Aniruddha Joshi[4], Prof. Sangeetha Selvan[5]

[1,2,3,4]UG Student, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India
[5]Assistant Professor, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India

---

Abstract - *As there has been an increase in the number of fraudulent claims in the insurance industry there is a need to curb this problem. Automobile insurance fraud is found to be the most predominant when compared to all the other types of fraudulent insurance claims. Therefore, there should be a system for the detection and prevention of such fraud and hence there comes the need to develop a system to detect insurance fraud. Many fraud detection models have been created using several algorithms and techniques. We have used Random Forest as a classifier and ADASYN to balance the dataset. This application made by us can be used by Automobile insurance providers to evaluate the claims made by their customers faster as compared to other traditional techniques which involve manual tasks, and thus the application helps ensure that while their customers claim the insurance the claim is genuine and not a fraudulent one with high accuracy as compared to traditional methods. Other techniques can also be used like the SVM, but for this particular topic, Random Forest seems to be ideal as it gives pretty good accuracy as compared to the other technique.*

Key Words: **ADASYN, Random Forest, Data Sampling, Insurance Fraud, Fraud Detection**

1.INTRODUCTION

Insurance fraud takes place when an insurance provider, advisor, adjuster, or consumer intentionally deceives for the purpose of obtaining an unlawful gain. In recent years, there is a rise in Fraudulent Insurance claims particularly in the Automobile Insurance Industry. Auto insurance fraud can range from falsifying information on insurance applications and overstating insurance claims to portraying accidents and submitting claim forms for damage or injury that never occurred to false allegations of vehicle theft. Fraud detection systems when used by Insurance Companies help them not only detect frauds but also save millions of dollars and sometimes even billions of dollars which would otherwise be given to the person who would have made a fraudulent claim. The primary aim of fraud detection algorithms is to determine whether there is any fraud or not.

2. LITERATURE REVIEW

A. Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique:

In this paper a Machine Learning approach was made for the detection of Insurance Fraud. The collected dataset is prepared on past history claims made by an insurance company. Once the dataset is prepared, they came to know that the dataset needs to be balanced so they used SMOTE to balance the dataset and used Random Forest for the prediction of the claim, So SMOTE with random forest gives accuracy up to 94%. But the model can be improved by applying other data balancing techniques such as ADASYN which is over sampling technique or oversampling technique such as bootstrapping, repetition or other classifiers that are not affected by the class imbalance.

B. Performance comparative study of machine learning algorithms for automobile insurance fraud detection:

For developing Insurance fraud detection models there are various machine learning techniques which can be used such as Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machine, Naive Bayes and many more. So, selecting one classifier with the best accuracy is a tedious task to perform. So, the Author presented a comparative study for ten most frequently used machine learning algorithms for Insurance fraud detection. The study shows that the Random Forest algorithm has the best performance for insurance fraud detection.

C. Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method:

As proposed in [2] we need a model with different data balancing techniques or a classifier which is not affected by the class imbalance. So the author tested the dataset by using an over sampling technique ADASYN in which the author tried to increase the samples of the minority class by adding the mean of columns in the particular row and he observed that ADASYN gave a better accuracy compared to SMOTE along with this the author changed base model with SVM. Which gave accuracy up to 93% again this model can be improved by using other balanced techniques that are based on more knowledge of the data patterns, rather than using statistical techniques.

D. Automobile Insurance Fraud Detection using Supervised Classifiers:

In this paper the author tested claims with different machine learning techniques such as Multi-Layer Perceptron, Decision Tree C4.5, and Random Forest with SMOTE as data balancing technique for the model. Among these techniques Random Forest proved to be the best technique for Automobile insurance fraud detection. But the Author proposed that model prediction can be improved by adding

more features like past historical claims, marital status of the policyholder, driver rating, and base policy.

E. Fraud Detection by Machine Learning:

In this paper the author told about different types of credit fraud and presented a machine learning approach which is capable of replacing human for the detection of fraud. The Author proposed that the dataset which we are going to use should have number of fraud and non-fraud in the ratio of 1:1 for best prediction he also provided different machine learning technique which are found useful for fraud detection such as logistic regression, support vector machine, boosted trees, random forest, and neural network. So if you are going to develop a fraud detection model then the model should be trained with balanced dataset and he or she need to find best fit algorithm for his problem statement, The most preferred ML algorithm is Random forest for fraud detection.

2.1 SUMMARY OF RELATED WORK

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

SN	Paper	Advantages and Disadvantages
1.	Sonakshi Harjai, Sunil Kumar Khatri, Gurinder Singh, 2019 [1]	Advantages: Used SMOTE to sample dataset and Random Forest as classifier. Disadvantages: Need other balancing technique
2.	Bouzgarne Itri, Youssfi Mohamed, Qbadou Mohammed, Bouattane Omar ,2019 [2]	Advantages: A comparative study for ten used machine-learning algorithms. The Random Forest algorithm has the best performance. Disadvantages: Biased output observed. Need a balancing technique.
3.	Charles Muranda, Ahmed Ali, Thokozani Shongwe ,2020 [3]	Advantages: ADASYN which is better than SMOTE Disadvantage: Used a poor classifier (SVM) compared to Random Forest
4.	Iffa Maula Nur Prasasti,Arian Dhini,Enrico Laoh,2020 [4]	Advantages: Balanced dataset using statistical technique and they used Random Forest as a classifier. Disadvantage: Model needs to be improved by using balanced techniques rather than statistical techniques.
5.	Yiheng Wei, Yu Qi, Qianyu Ma, Zhangchi Liu, Chengyang Shen, Chen Fang ,2020 [5]	Advantages: Give an idea about different machine learning technique which can be used to detect a fraud

3. PROPOSED WORK

We have used ADASYN sampling technique for balancing data and Random Forest classifier for classification of given cases of insurance claim as fraud or genuine.

3.1 SYSTEM ARCHITECTURE

The system architecture is given in Figure 1. Each block is described in this Section.

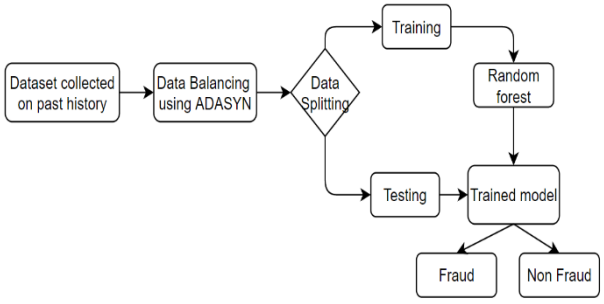

Fig. 1 Proposed system architecture

A. Dataset Collected on past history: The dataset was created using all data from previous insurance claims. The data has been checked for any missing values, redundant data, duplicates or any noise. The data is then transformed to meet the demand. The processed dataset is now passed for balancing.

B. Data Balancing using ADASYN: The processed dataset contains a large amount of Non fraud class data when compared Fraud class data. This creates a biased prediction model. To balance this dataset, we use Adaptive Synthetic Sampling Technique on the dataset. It is an oversampling technique i.e. it increases the Fraud class data to match with Non fraud class data.

C. Data Splitting: For training the model we will split the dataset in two parts for example 20% of data for testing and remaining 80% of data for training.

D. Training: We have used the Random Forest classifier which trains the data set aside for training the dataset into a model which will classify any new input case as fraud or not fraud.

E. Random Forest: The Random Forest, as the name suggests, consists of a large number of individual decision trees that act as a whole. Each individual tree in the random forest will make a class prediction, and the class with the most votes will become our model's prediction.

F. Testing: Remaining of splitted data is used for testing the model. This will give us the accuracy of our trained model.

G. Trained Model: Trained model is created after the dataset is trained using random forest and it is tested simultaneously. This model gives us a prediction whether a new insurance claim is fraud or non-fraud.

3.2 REQUIREMENT ANALYSIS

A. Software

We are going to develop an Automobile insurance fraud detection system which is deployed as a website using flask. The detection system needs python as a programming language, and a machine learning model is developed using Jupyter notebook. Database is made using SQLite. For the front end, we will be using bootstrap and HTML.

B. Hardware

In our project we have done data sampling using ADASYN and used that data for classification of genuine or fraud insurance claims using machine learning technique Random Forest. To run these modules properly we minimally require a 2 GHz processor and 8 RAM and 180 GB of HDD.

4. CONCLUSION

The fraudulent insurance claim detection model can help detect fraudulent insurance claims. Based on literature surveys we can conclude that depending entirely on imbalanced dataset will affect accuracy of our model. Therefore, to increase the accuracy of the model, in the proposed work, ADASYN should be used to balance the data. With the user giving the insurance details for the claim, the claim is fraudulent or not can be checked. Thus, the model for detecting fraudulent insurance claims successfully categorizes the fraudulent and genuine claims.

ACKNOWLEDGEMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Sangeetha Selvan for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Sonakshi Harjai, Sunil Kumar Khatri, Gurinder Singh, "Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique", Nov 2019 Available: [Online] https://ieeexplore.ieee.org/document/9036162

[2] Bouzgarne Itri, Youssfi Mohamed, Qbadou Mohammed, Bouattane Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection",30 Oct 2019 Available: [Online] https://ieeexplore.ieee.org/document/8942277

[3] Charles Muranda, Ahmed Ali, Thokozani Shongwe, "Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method",16 Oct 2020 Available: [Online] https://ieeexplore.ieee.org/document/9259322

[4] Iffa Maula Nur Prasasti,Arian Dhini,Enrico Laoh, "Automobile Insurance Fraud Detection using Supervised Classifiers",18 Oct. 2020 Available: [Online] https://ieeexplore.ieee.org/document/9255426

[5] Yiheng Wei, Yu Qi, Qianyu Ma, Zhangchi Liu, Chengyang Shen, Chen Fang, "Fraud Detection by Machine Learning",25 Oct 2020 , Available: [Online] https://ieeexplore.ieee.org/document/9361052

BIOGRAPHIES

Dhruvang Gondalia

Omkar Gurav

Ameya Joshi

Aniruddha Joshi