INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 3)
Available online at: www.ijariit.com

Comparative analysis on vehicle insurances fraud detection using machine learning

Sheethal H. D.
sheethalhd@gmail.com
Vidyavardhaka College of Enginnering, Mysore, Karnataka

P. Sai Pranavi
pranavip27@gmail.com
Vidyavardhaka College of Enginnering, Mysore, Karnataka

Sharanya S. Kumar
sharanya.sunilkumar34@gmail.com
Vidyavardhaka College of Enginnering, Mysore, Karnataka

Sonika Kariappa
sonika9n@gmail.com
Vidyavardhaka College of Enginnering, Mysore, Karnataka

Swathi B. H.
swathibh@vvce.ac.in
Vidyavardhaka College of Enginnering, Mysore, Karnataka

Gururaj H. L.
gururaj1711@vvce.ac.in
Vidyavardhaka College of Enginnering, Mysore, Karnataka

ABSTRACT

Now-a-days frauds have become a serious threat to the society. Fraud is an illegal way of gaining more money. Frauds are posing problems to so many people. So, fraud detection becomes very important in this current world. Fraud detection can be implemented in various fields like banking, insurance, financial sectors and information security systems. And in the field of insurance, we have different types of insurances- health, vehicle and even life insurances. Frauds occur in each of these types of insurances. There are many approaches using which fraud can be detected. Machine learning, artificial intelligence, data mining and other methods are used to detect frauds. The well-known methods used in machine learning to detect frauds are Bayesian Network, Decision trees and back propagation techniques. Many algorithms are also used to detect frauds like Naïve Bayes, KNN, Random forest. In this paper, the different techniques used for vehicle insurance fraud detection are presented along with comparative analysis.

Keywords— *Fraud, Fraud detection, Machine Learning, Insurance, Naïve Bayes, KNN, Random forest, Neural Network*

1. INTRODUCTION

Every year billions of dollars are lost due to various frauds. The Association of Certified Fraud Examiners (ACFE) defines fraud has a mistake committed by a person or entities for benefits that are not good for an individual and even the insurance companies. Internal fraud and External fraud are the different types of fraud. Internal fraud is against the insurer by the internal employees either with the help of internal or external people of the company. External fraud is against the insurer by the external employees sometimes with the help of insiders of the company.[8] There are two types of frauds soft fraud and hard fraud. Hard fraud is when individuals deliberately counterfeit a mishap. Soft fraud is when individuals has a legitimate case yet distorts some portion of the case. There are different areas where fraud usually occur such as credit cards, healthcare insurance, telecommunication, automobile insurance and online auction.[10]

Motor insurance and health insurance are more susceptible to insurance frauds. According to the survey, it is observed that for every 100 rupees earned the insurance company would be paying 213 rupees as the claim. The fraudulent activities not only affect the insurance companies but also the individuals as they have to pay higher premiums. Subsequently, to keep this from proceeding to influence one another, the inquiries should be possible by using the innovation as an answer for discover an example and afterward distinguish misrepresentation that has been happened dependent on the information that has been gathered from the past cheats as there is an expansion in the extortion cases. [12] Various techniques such as data mining, artificial intelligence, database, distributed and parallel computing, fuzzy logic, machine learning, genetic algorithms, neural networks, KNN, pattern recognition, statistics and visualization can be used for detecting and reducing the insurance frauds. Be that as it may, getting access to the information to perform fraud detection is increasingly troublesome because of the protection of the individuals to be kept up by the organization and furthermore because of the absence of very much distributed research strategies and procedures.

Whenever the insurance company cannot detect any frauds, they provide the money for the applicants. Hence there would be loss of huge amount for the company. So, to overcome this, they increase the premiums which affects the individuals. In the banking, fraud is due to the stolen credit cards, falsifying cheques, manipulation of accounts and more. Hence to detect and prevent frauds the telephone, insurance companies and the banks made use of the data analysis.[2] The Falcon extortion appraisal framework (FICO) was effectively executed in the

business of banking. Indeed, even the markets have begun to utilize the CCTV together with Point of administration (POS). Hence, the paper is structured as follows, section 1 gives introduction, section 2 gives the description about various works carried out by different researchers and the last section draws the conclusion and the future enhancement.

2. RELATED WORKS

Stijn Viaene, et al. [13] proposed strategies for detecting fraudulent claims. Insurers use automatic detection systems that help you decide whether to conduct an investigation on claims for detection of fraud. Figure1 shows a generic fraud control model is designed for P& C insurers.

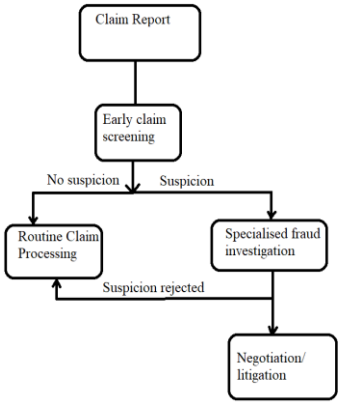

Fig. 1: Fraud control Model

The model incorporates screening, examination, arrangement /case stages. It is actualised in the safety net provider's case for taking care of procedures. Claim handling is a procedure that starts from claim occurrence and closes with the payment for the damages caused.

Sebastián M. Palacio, et al. [2] predicts fraud by utilizing semi supervised procedures and another metric that uses the cluster score which can be utilized for fraud detection which can manage the pragmatic difficulties. The principle strategy incorporates transposing unaided models into managed models utilizing the cluster score metric. It fundamentally gives the blueprint of the fringe among clusters and assesses the homogeneity of the variations from the norm in the group development. The figure 2 shows the possible clusters that result after using the unsupervised methods.

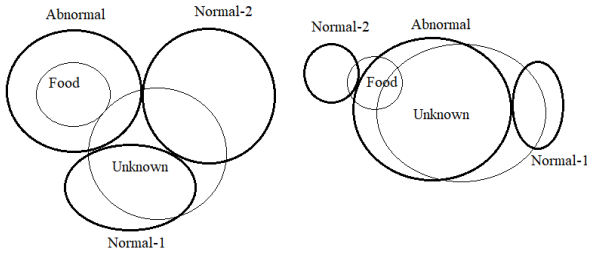

Fig. 2: Possible clusters

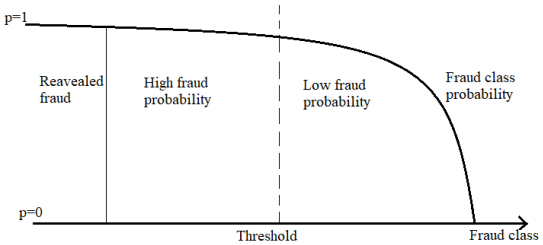

Fig. 3: A graph for the desired threshold to split the high fraud possibility cases from low fraud possibility cases

The figure 3 portrays that the limit line is drawn with the goal that we can acknowledge just fraudulent cases. Yet at the same time it very well may be unreasonable. On the off chance that we work between these limits, the instinct uncovers to us that we should be close to the lower edge which acknowledges fraudulent cases.

Ali Ghorbani, et al. [4] utilized a portion of the data mining procedures for extraction of concealed information and examples on enormous data to lead the insurance industry. Fraud is present in many industries including insurance industry. So fraud detection becomes the most important issue. One of the techniques to distinguish fraud in reported harms misfortunes is to utilize information from more seasoned identified fraud cases. We can utilize K-means clustering to distinguish fraud designs in vehicle insurance and third-party insurance. K-means clustering procedure is applied on informational collection with Euclidean separation assimilarity and divergence estimation. The aftereffect of this method gives a decent precision when contrasted and genuine measurements.

Ke Nian, et al. [9] predicts fraud using unsupervised spectral ranking for anomaly of interdependence relation. Data Mining and Machine learning are used to detect fraud cases and can reduce the economic losses Predictive strategy expand the detection rate ,limit the false positive rate and can quickly distinguish and develop the fraud plans. Clustering examination and outlier detection are the strategies utilized in Standard unsupervised anomaly identification. Density-based method is used to detect global outliers and local outliers. Local outliner factor in density based method defines the parameter neighbourhood to compare the density. To guide the learning process, the learning targets are not available. So unsupervised learning is complex than supervised learning. The challenges faced while using unsupervised learning is most of the data sets used are categorical and claim datasets forms multiple patterns which produces unsatisfactory results.

Sharmila Subudhi, et al. [3] has proposed noval approach for fraud detection in automobile insurance. Feature selection algorithm is used and its attributes are chosen from the datasets. A test set is removed from the chose characteristics and the remaining datasets are exposed to Possibilistic fuzzy C Means clustering systems for under sampling approach.

Xinxin Jiang, et al.[7] The insurance datasets are usually imbalanced and heterogeneous. And the traditional algorithms can be applied to those which are balanced and homogenous. Therefore, a parallel neural network has been proposed for such datasets which is imbalanced and heterogeneous. The cost sensitive classification tries to minimize the overall risk where the loss function of training set is the error function of cost sensitive classification. Then the optimization objective is being classified, then the cross entropy function will maximize the predictions of the required output. The objective minimization is used to find the optimal of the function. This is applied on the real-world, three insurance datasets which are extracted to test the algorithms. Due to the alternate application of the algorithms which results in imbalanced cost sensitive matrix.

Sharon Tennyson, et al. [14] theoretical literature based on the insurance contract and literature based statistical analysis. They use the PIP insurance data of the Massachusetts state. The insurance company audit uses various methods. The audit patterns reveal that 33% of the cases are evaluated. The empirical model estimates the probability of the claim audit. The

patterns of audit outcomes obtained uses the logistic model. These patterns suggest that audit have some deterrent role. Subsequently this paper helps in examining the job of claim auditing in auto mobile insurance market.

Dongxu, et al. [6] in this paper CoDetect fraud detection framework has been introduced which can detect the fraud as well as the fraud activities patterns. The major contribution in the paper are the following: Give and set up a way to deal with weighted chart in the financial network and afterward joins the properties of links and hubs. Shows the various situations in the financial fraud to define examples of the fraud regarding the graphical portrayal and the sparse matrix. Proposes the novel unsupervised framework- CoDetect, for issues of complex example revelation and the abnormality which includes identification. Assesses the outline work utilizing the manufactured and the present reality information to exhibit both efficiency and effectiveness of the proposed framework.

Rekha Bhowmik [11] this paper analyses fraud detection techniques to predict fraud patterns from the data. Data mining techniques are associated with supervised learning and unsupervised learning. Fraud detection involves Criminal offenders, Organised criminal off enders and Soft fraud. Bayesian Belief Networks algorithm is far better when compared with decision tree and back propagation. In the decision tree algorithm, it is more like if-then structure. Algorithms are utilized to isolate the information and furthermore for descriptive classification rule that can be utilized for new instance. Model performance of the model can be classified in 2 ways: Confusion matrix and ROC graph. The confusion matrix and ROC graphs are used to measure the performance levels of the classifications of a given data set.

Hojin Moon, et al. [1] uses many algorithms. Logistic regression is a well-known strategy to construct an expectation model for a double reaction variable. A different logistic regression model takes various indicator variable. In this, Evaluation and selection of variable is completed by means of a logistic regression model. Least Absolute Shrinkage and Selection Operator: LASSO strategy is a regression model that penalises the absolute size of the coefficient which makes some regression coefficient shrink to zero. The variable was chosen through LASSO variable selection by means of CV by checking AIC improvement utilizing LRT: month guaranteed, deficiency, sex, age. Irregular woodland includes more haphazardness than a solitary characterization tree. Support Vector Machine: SVM are kernel based supervised learning calculations. After the preparation stage new perception are mapped into a similar space and are ordered to a class dependent on a side of hyper plane. The best strategy in these LASSO technique is best appropriate for arranging fraudulent car claims.

Tessy Badriyah, et al. [5] predicts fraud by using Nearest neighbour-based method and statistics methods. In distance-based algorithm, first the maximum distance and the number of nearest neighbours are determined. Then, the separation of item to every single other article is determined. Later we find as many neighbours which has smallest distance value. Then the objects are labelled as fraud or not. If objects are labelled as neighbour fraud, those objects are expressed as an outlier.

3. COMPARISON

The table 1 compares various machine learning techniques like Random forest, KNN, Naïve Bayes etc with each other based on some parameters like complexity, outcome obtained, advantages, disadvantages, recommendation.

Table 1: Comparison of different methods used for fraud detection

Method used	Complexity	Results	Advantages	Disadvantages	Recomendation
Algorithmic model which is implemented in various detection Systems	Medium	Proposed system predicts fraud with low error rate.	High efficiency, Less time consumption	Complexity varies while implementing the model.	Model can be implemented educing better, latest techniques
Mini batch K-means, Isolation forest, Gaussian mixture, Bayesian mixture	High	The success rate is more with precision of 89-92%	Mini batch K-means is faster and gives the best results	Highly complex	Other ML techniques can be used to predict frauds.
K-means clustering	Low	The system extracted patterns which helped to detect fraud in insurance claims	High accuracy	There might be variations while clustering the data.	Precision of fuzzy strategy can be concentrated to get regular vulnerability of occasions. Effect of insurance market basket can be contemplated in vehicle fraud events.
Feature selective algorithm and WELM technique	Medium	Controlling the imbalanced class distribution and isolation of fraudulent cases.	Can be successfully utilized for fraud recognition in other application and conventional database.	Classifiers and proposed system performance was less.	WELM parameters can be upgraded by utilizing different optimization procedure for improving the classifier execution.
Spectral ranking method.	Low	Proposed system is applied to fraud detection and to other anomaly detection problem	Spectral optimization system can be deciphered as an estimation to an unsupervised vector machine and can determine a ranking legitimately.	Obtaining labels are time consuming costly.	On the off chance that the minority class doesn't have an adequately enormous check rate, one can decide to survey anomaly likelihood regarding a single majority class and ranking is created reasonably with this view.
Two parallel literatures	Medium	The use of audits for both deterrence and detection.	Investigative audits show that auditing patterns are consistent with the use of audits for both fraud detection and fraud deterrence.	we tend to put more weight on the results obtained when recorded statements were not counted as audits.	Recorded statements are to be considered as audits.

Parallel neural network.	High	The comparative results on the real world data set show the effectiveness of the design.	Accurate and effective results are obtained for heterogeneous dataset	CPLF and PNN network which are based on the parallel neural network with no cost-sensitive matrix are demonstrated as better classification accuracy than the DNN and RNN networks for different datasets at the same imbalance level.	To analyse multiple datasets with the help of IoT sensors.
Anomaly feature detection, CoDetect.	Medium	The proposed network can determine the frauds as well as the fraud activities.	Effective & efficient results are obtained from the proposed framework. The trace of the fraud can also be obtained.	The majority of existing methods focus on networks or features information separately, which does not utilize both information.	To study the integration of the tensor into the codetect framework for the detection of fraud.
Nearest neighbour (density based and distance based algorithm) and statistics method.	Medium	Use of selection method helped in increasing the accuracy.	The output obtained is better than that when SVM method is used.	The feature selection influences the anomalies in the dataset.	Use of selection method along with the genetic algorithm for increasing accuracy.
Naive Bayes algorithm, Decision tree algorithm, Multilayer perception algorithm.	Medium	Intrusion detection systems are the security tools used to detect anomalous and fraudulent activity from inside and outside intruders.	IDS does not require modification of production servers or host. Multilayer preceptor algorithm has highest accuracy of 99.47%.	Host based network rely upon monitoring capabilities of server.	Use of various other machine learning algorithms for the better and effective results.
Naive Bayes algorithm, Random forest, J48.	High	Percentage of premium amount and fraud claims prediction is done using three classifiers.	Big data analytics are efficient in predicting claims in both large and small volumes. Random Forest algorithm outperform the other and have an accuracy of 99.41%.	The data sets used for premium and insurance analysis are not freely available. It has different format of premium dataset attributes.	The relation between premium data and insurance claim will be discovered and classification algorithms will be customized to optimize the result.
Bayesian network, C4.5 algorithms.	Low	Using this system fraud will be reduced.	Easy to implement, robust.	Requires to check all condition.	C4.5 decision tree algorithm.
Logistic Regression, Least Absolute Shrinkage and Selection Operator, Random Forests.	High	Automobile fraud will be reduced Considerably	Good balances, robust, flexibility.	Does not consider all sub categories.	Least Absolute Shrinkage and Selection Operator.
Bayesian network, C4.5 algorithm s, Rule based algorithm.	Medium	Fraud will reduced in the sector in which implement ed.	Easy to understand , complexity is less.	All condition must be checked before implementing.	Rule based algorithm.

The comparison table compares various algorithms and methods used for the detection of insurance fraud. The implemented algorithms have various advantages, disadvantages and various improvisations which can be done to improve the efficiency of the fraud detection algorithms. For example, in one of the paper they have used nearest neighbour algorithm, statistics method along with the feature selection and the results where efficient. But if we are using genetic algorithm along with feature selection then the results would be more accurate than the before obtained results.

Similarly, in other case they have used K-means clustering has been implemented in other paper for the detection of fraud. Here the results are highly accurate, while there would be variations during the data clustering. Hence fuzzy strategy precision should be concentrated to get vulnerability of occasion. Therefore, the above comparison table contains various methods that are used to implement or for the detection of insurance fraud in order to obtain efficient and accurate results. But they even contain various advantages, disadvantages and recommendations for the better, efficient an accurate result for detecting the insurance fraud.

4. CONCLUSION

As fraud poses a serious problem in the current society, it has to be resolved. In order to resolve these problems, systems are built which predict fraud in the data given. These systems are built using various machine learning techniques like naïve Bayes, KNN, random forest, neural networks. In this paper we have discussed about various machine learning techniques and how it is implemented in the systems and how accurate it is in predicting the fraud. Later these techniques are compared using five criteria from different perspectives. In the future, various new techniques and algorithms can be implemented in systems to detect fraud with less errors and more accuracy.

5. REFERENCES

[1] A Predictive Modelling for Detecting Fraudulent Automobile Insurance Claims, Hojin Moon, Yuan Pu, Cesarina Ceglia,Journal of Theoretical Economics Letters, 2019, 9, 1886-1900

[2] Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning, Sebastián M. Palacio, Data science journal, 2019

[3] Detection of Automobile Insurance Fraud Using Feature Selection and Data Mining Techniques, Sharmila Subudhi, International Journal of Rough Sets and Data Analysis · July 2018

[4] Fraud Detection in Automobile Insurance using a Data Mining Based Approach, Ali Ghorbani and Sara Farzai, International journal of Mechatronics, Electrical and Computer technology, Vol. 8(27), Jan. 2018, PP. 3764-3771

[5] Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance, TessyBadriyah, LailulRahmaniah, IwanSyarif, 2018 IEEE, 978-1-5386-8066-7/18

[6] CoDetect: Financial Fraud Detection with Anomaly Feature Detection, Dongxu,DejunMu, LibinYang and Xiaoyan Cai, 2169-3536 2018 IEEE.

[7] Cost-sensitive Parallel Learning Framework for Insurance Intelligence Operation, Xinxin Jiang, Shirui Pan, Member, IEEE, Guodong Long, Fei Xiong, Jing Jiang, and Chengqi Zhang, IEEE transactions on industrial electronics, 2018

[8] A State-of-the-Art Review of Machine Learning Techniques for Fraud Detection Research, Sinayobye Janvier Omar, Kiwanuka Fred, Kaawaase Kyanda, 2018 ACM/IEEE Symposium on Software Engineering in Africa

[9] Auto insurance fraud detection using unsupervised spectral ranking for anomaly, KeNian, Haofan Zhang, Aditya Tayal, Thomas Coleman, Yuying Li, The Journal of Finance and Data Science 2 (2016) 58e75

[10] Fraud detection system, Aisha Abdallah, MohdAizainiMaarof, Anazida Zainal, Journal ofNetworkandComputerApplications68(2016)90–113

[11] Detecting Auto Insurance Fraud by Data Mining Techniques, Rekha Bhowmik, Journal of Emerging Trends in Computing and Information Sciences, Volume 2 No.4, April 2011

[12] Journal of computer and network Applications, Fraud detection system: A Survey, Aisha Abdallah, Mohd Aizaini Maarof, Anazida Zainal, 68 (2016) 90-113.

[13] Strategies for detecting fraudulent claims in the automobile insurance industry, Stijn Viaene, Mercedes Ayuso, Montserrat Guillen, Dirk Van Gheel, Guido Dedene, European Journal of Operational Research 176 (2007) 565–583ss

[14] Claims Auditing in automobile insurance: fraud detection and deterrence objectives, Sharon Tennyson, Pau Salsas-Forn, The Journal of Risk and Insurance, 2002, Vol. 69, No. 3, 289-308

