International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 30 No. 1 December 2021, pp.504-507

Detection of Vehicle Insurance Claim Fraud

A Fraud Detection Use-Case for the Vehicle Insurance Industry

Dilkhaz Y. Mohammed

Technical College of Engineering, Duhok Polytechnic University

Duhok, Iraq

Abstract— Insurance fraud has accompanied insurance since its inception, but the manner in which these practices and their methods of operation have evolved over time, and the volume and frequency of insurance fraud incidents have recently increased. Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident. Some common examples include staged accidents where fraudsters deliberately "arrange" for accidents to occur; the use of phantom passengers, where people who were not even at the scene of the accident claim to have suffered grievous injury, and making false personal injury claims where personal injuries are grossly exaggerated. The typical analysis of these datasets includes Algorithms is implemented on the Weka tool depends upon real info represented through from Oracle Databases. In this paper, focusing on detecting vehicle fraud by using, machine learning algorithms, and also the final analysis and conclusion based on performance steps, revealed that J48 is more accurate than Random Forest, Random Tree, Bayes Net and Naïve Bayes but Random Tree has the lowest classification accuracy.

Keywords— J48, Random Forest, Random Tree, Bayes Net, Naïve Bayes Weka.

I. INTRODUCTION

Insurance fraud is very old, and some preventive measures date back to the middle Ages. At that time, failure to comply with ethical rules may result in very severe penalties. The punishment of insurance fraudsters was severely reduced, as when a merchant ship was rescued in Jascoyne Bay in the 15th century, an inspection of its cargo revealed that the ship was full of stones, while the shipping policy mentioned fabrics, and in 1570, the captain and the insurance broker were both found guilty and sentenced to death. In 1598, the Amsterdam Rules of the Prague and Antwerp Acts provided for physical and financial penalties for captains of navigators and holders of documents in case of fraud. The fraud may be internal or external, occurring at any stage of the contract. Excessive insurance, forgery, corruption, intentional destruction or deterioration of property. The fraudulent act draws the attention of insurance companies to two matters: the assessment of fraud and the effect of fraudulent acts on the amount of premiums.

This project's purpose is to develop a model that can detect motor insurance fraud. The difficulty with machine learning fraud detection is that fraud is significantly less common than legitimate insurance claims. Given the variety of fraud methods and the relatively low ratio of recognized fraud in typical samples, insurance fraud detection is a difficult challenge. While developing detection models, the cost of false alerts must be balanced against the cost of loss avoidance. Machine learning approaches improve forecast accuracy, allowing loss control units to cover more territory with fewer false positives. Insurance fraud refers to a variety of unethical behaviors that a person may engage in in order to obtain a favorable outcome from an insurance company. This could include arranging the incident, misrepresenting the circumstances, adding significant individuals and the incident's cause, and lastly the scope of the incident.

Corresponding Author: Dilkhaz Y. Mohammed

II. Methods and Techniques used for Fraud Detection

A. Dataset Collector

The dataset came from the Oracle Database. The dataset contains information about the insurance policy as well as information about the consumer. It also contains information about the accident that was used to make the claims. This dataset contains vehicle datasets (attributes, model, accident details, etc.) along with policy details (type, tenure, etc.). The target is to detect if a claim application is fraudulent or not

B. Data Pre-Processing

The dataset used for our research paper consists of a total of 15420 instances and more than 30 attributes. Before applying techniques (algorithms), usually some preprocessing is performed on the dataset. It is necessary to improve the quality of the data to accomplish data processing. There are a small number of techniques used for data processing, such as data aggregation, data sampling, data discretization, variable transformation, and dealing with missing values.

C. Using Techniques with Weka

There are some model evaluation techniques that you can choose from, and the Weka machine learning workbench uses four of them to prepare your model within the entire training dataset. After that, evaluate the model on the same dataset. Manually divide your dataset using another program. Percentage Split: randomly divide your dataset into training and test segments. When the researcher evaluates a model, each testing partition is evaluated. Cross-Validation: split the data into k-partitions or folds. Train a model on all of the partitions except one that is usually held away as a test set. After that, calculate the average performance of all k models.

The researcher can see these techniques in the Weka tool explorer. On the classify tab after you have loaded a dataset, each test option has a time. Evaluation options are concerned with determining the performance of a model on unseen data. Predictive modeling aims to build a model that performs best in a situation that we do not entirely understand, in the future with new unknown data. To achieve the best results, we must employ these types of robust statistical techniques. Estimate the performance of the model in this situation, and the performance summary is provided in Weka when you evaluate a model. Whenever evaluating a machine learning algorithm on Due to a classification issue, you are given a massive total of performance information to digest because of classification.

This could be the most analyzed type of predictive modeling issue. And there are numerous approaches to considering the performance of classification algorithms. Therefore, the first thing to note in the performance is the Classification algorithms' classification accuracy, which is defined as the ratio of In comparison to all forecasts, the number of correct predictions is small. It is frequently presented as a percentage, where 100% is the first and foremost thing an algorithm is capable of performing. The second one is accuracy. By class, take note of the real positive and false positive rates. The researcher get the predictions for each class, which may be instructive. From the class break down, you get the problem that is uneven, or you will find more than two classes. As well as the last one, the confusion matrix, a table showing the number of predictions for each class in comparison to the number of instances that actually participate in each class. The fraud data set worldwide is split into two. Training dataset with a 66% percentage and a testing dataset with a 34% percentage of the whole dataset, and that is applied by using the default setting of the Weka tool

III. Results and Discussion

In our experiment, the researchers applied different algorithms to the fraud detection data set. The pre-processed dataset, which consists of 15420 data instances, is converted to an *.ARFF file to be used by the Weka tool.

Each test option has a time. Therefore, the results are obtained according to two test options, which are:

1. Test Split evaluation, which divides the input data set into 66% for training data and 34% for testing data.
2. 10 Fold Cross-Validation.

The results from the applied classification algorithms in the two approaches will be evaluated according to four performance measures, which are the classification accuracy, recall, the F-measure, also called the F-Score and MCC.

In the case of dividing the input data set into 66% for the training data and the remaining 34% for testing, the results are

shown in Table 1 and 2, which provides a clear comparison among the selected classifiers according to accuracy, precision, recall, F-measure and MCC which shows that:

Table 1 Test split and accuracy results.

Algorithm Used	Correctly Classified Instances	Incorrectly Classified Instances	Precision	Recall	F- Measure	MCC
J48	99.9935	0.0065	1.000	1.000	1.000	1.000
Random Forest	99.9805	0.0195	1.000	1.000	1.000	1.000
Random Tree	89.0921	10.9079	0.946	0.951	0.948	0.923
Bayes Net	99.7665	0.2335	0.996	0.998	0.997	0.996
Naïve Bayes	99.7147	0.2853	0.996	0.998	0.997	0.996

From the accuracy point of view, J48 was correctly classified about 99.9935% of the data. That means 15419 items out of 15420 instance. Random Forest outperformed Bayes Net and Nave Bayes, which correctly identified approximately 99.98.5% of the data. It is evident that the accuracy of Random Tree achieved the lowest accuracy of 89.0921% among the other classifiers, although it has the lowest precision, recall, f-measure, and mcc of the other classifiers. As shown in the below table, which summarizes the number of correct and incorrect predictions.

Table 2 Test split and confusion matrix results.

Algorithm Used	Classified as
J48	5008 1 0 | a = Liability
	0 5962 0 | b = Collision
	0 0 4449 | c = All Perils
Random Forest	5007 1 1 | a = Liability
	0 5962 0 | b = Collision
	1 0 4448 | c = All Perils
Random Tree	4762 153 94 | a = Liability
	168 5274 520 | b = Collision
	105 642 3702 | c = All Perils
Bayes Net	5001 0 8 | a = Liability
	0 5953 9 | b = Collision
	19 0 4430 | c = All Perils
Naïve Bayes	5000 0 9 | a = Liability
	0 5946 16 | b = Collision
	19 0 4430 | c = All Perils

It is obvious from Figure 1. That a comparison is applied to our five classifiers due to precision, recall, F-measure, and mcc, which shows us that J48 has the highest accuracy of the other classifiers. It does not mean that it performs well in other results. The Random Forest classifier performs well in all the results. The overall performance of Bayes Net is very close to that of Nave Bayes results. A random tree has lower accuracy than all other classifiers

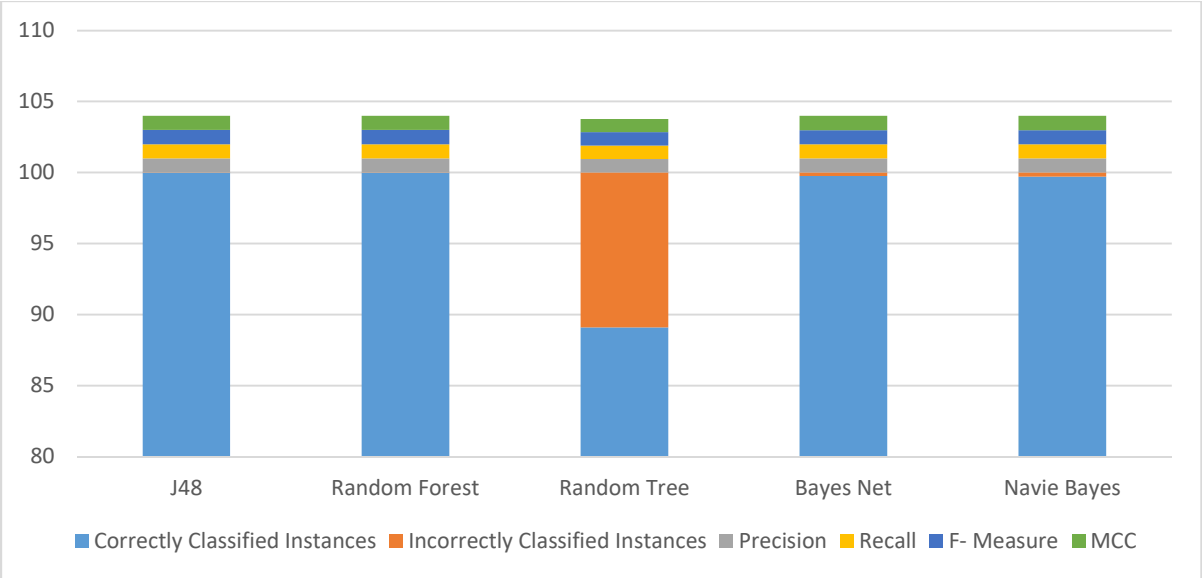

Figure 1 Performance measures results of used classifiers.

IV. CONCLUSION

The field of fraud analytics research is a growing one. Innovating, developing, deploying, and evaluating new models, software tools, and procedures to aid in the fight against terrorism is the focus of research. The final results after each Weka algorithm has been applied us used the following methods in our research: Nave Bayes, Bayes Net, J48, Random Forest, and Random Tree are the five primary classification algorithms. Finally, some researchers could modify this research by using different methods. Others could use different test options to test the performance of the classification algorithms.

REFERENCES

[1] Baesens B, Höppner S, Verdonck T. Data engineering for fraud detection. Decision Support Systems. 2021 Jan 12:113492.

[2] Pourhabibi T, Ong KL, Kam BH, Boo YL. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems. 2020 Jun 1; 133:113303.

[3] Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences. 2019 Apr 1; 479:448-55.

[4] Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection-machine learning methods. In2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) 2019 Mar 20 (pp. 1-5). IEEE.

[5] Johnson JM, Khoshgoftaar TM. Medicare fraud detection using neural networks. Journal of Big Data. 2019 Dec; 6(1):1-35.

[6] Thennakoon A, Bhagyani C, Premadasa S, Mihiranga S, Kuruwitaarachchi N. Real-time credit card fraud detection using machine learning. In2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2019 Jan 10 (pp. 488-493). IEEE.

[7] Makki S, Assaghir Z, Taher Y, Haque R, Hacid MS, Zeineddine H. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access. 2019 Jul 8; 7:93010-22.