ISSN: 0974-8571 Vol.14 No. 1 June, 2022

International Journal of Computational Intelligence in Control

INSURANCE FRAUD DETECTION MODEL: USING MACHINE LEARNING TECHNIQUES

Tuba Aziz*[1], Fareedullah[2], Mudasir Mahmood[2], Muhammad Ali Shah[2], Muhammad Ijaz Khan[2],
Muhammad Farhan[2], Muhammad Riaz[3], Muhammad Yaseen[3]
tuba.akbar513@gmail.com , fareedkamran44@yahoo.com , mudasir@gu.edu.pk , alishah@gu.edu.pk, ijazkhan@gu.edu.pk , farhan@gu.edu.pk, mriazengineer911@gmail.com , y3333tareen@gmail.com
Department of Computer Science, Muhi-ud-Din Islamic University, Azad Jammu & Kashmir*[1]
Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan[2]
Faculty of FICT, BUITEMS, Quetta[3]
Pakistan, 29050
Corresponding Author*: fareedkamran44@yahoo.com

ABSTRACT

Fraud is the activity which will cause distress to corporations. This Financial fraud has been a huge worry for many organizations across the industries, this insurance industry comprises of over and above thousands companies all across the board, which gathers greater than a trillion dollars premium every year and billions of dollars are being lost every year because of this fraud, thus detection of Insurance Fraud is a burdensome task for the insurance companies. The conventional outlook for detecting insurance fraud was completely dependent on evolving heuristics around the fraudulent indicators. The auto insurance fraud is being considered to be one of the leading categories of fraud, which will be carried out by faking accident claim. In this study, we are concentrating towards tracking down the auto insurance fraud by making use of Machine Learning Techniques (Naïve Bayes Classifier), by using this techniques time complexity will be declined and also depicts the results accurately. Gradient Boosting Classifier is latest algorithm with take less time and gives 91% Accuracy.

Keywords: ***Machine Learning, Auto Insurance, Fraud detection, Naïve Bayes Classifier.***

1. INTRODUCTION

Vehicle (auto) insurance is a legit testimony between the insurance provider and consumer to compensate the monetary expenses occurred due to any accidental damage or theft of vehicle. Insurance fraud in automobile sector happens when a person attempts to acquire financial benefits either by showing papers regarding any vehicular wreckage in a forged accident or by filing documents for previous losses or by submitting wrong or missing information about the driver. In order to boost customer happiness, companies must have an effective fraud detection and detection management system. There will be a reduction in loss adjustment costs due to an improvement in customer satisfaction. For detecting fraud claims, there are now various methods available. Analyzing data according to its own instructions is the most used. As a result, they require in-depth investigations that take a long time to complete and cover a wide range of topics. Using machine learning, we were able to completely solve this issue.

An insurer's biggest problem is fraud, which is well-known. An automobile insurance company's claim data is the topic of this thesis. Each insurer faces the possibility of bearing a hefty financial burden as a result of false claims. Because of this, it is critical to determine which statements are true and which are false. Insurance companies are unable to individually investigate every single claim because it would take too much time and money. In this research, we'll take advantage of the insurance industry's most valuable asset: data. The insurer provides us with a wide range of data concerning claims, insured people, and other variables. New insights can be gained by examining the rates of fraud in different types of claims.

Machine learning helps us identify claims that are most likely to be false. This might help cut down the list of claims that require additional investigation.. As a result, an insurer is able to uncover more false claims. Auto insurance fraud detection is one of the primary objectives of this research. In machine learning, fraud detection is a challenge because it is much less common than legitimate insurance claims. FBI's (Federal Bureau of Investigations) report produced in

2017, the United States has lost nearly $39 billion due to auto insurance fraud in 2016. This statistics high lights the extremity of the situation and therefore requires to be controlled sternly to reduce the monetary losses.

2. LITERATURE REVIEW

According to the researcher [1] focusing on detecting the auto\vehicle fraud by using, machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall. Comparing these three algorithms, decision tree, random forest and naïve Bayes algorithms the researcher concluded that decision tree and random forest have better performance than naïve Bayes. In the observation, choose sample of more than 500 data. In future work with more algorithms and finally calculate which provide more accuracy, precision, and recall.

Researchers [2] has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and discussed measures to detect them. Such measures have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms. The main tasks will be to build scoring models to predict fraudulent behavior, taking into account the fields of behavior that relate to the different types of fraud identified. As the next step focus will be upon the implementation of a 'suspicious 'scorecard on a real data-set and its evaluation.

Researcher [3] presents fraud detection method to predict and analyze fraud patterns from data. The insurance claim is a basic problem in insurance companies. Insurance insurers always have a challenge to the growing of insurance claim loss. Because there is the occurrence of claim fraud and the volume of claim data increases in the insurance companies. Machine learning model that classifies and make motor insurance claim status prediction in machine learning approach to predict and present fraud the researcher used Naïve Bayesian classifier and Decision Tree-Based algorithms. The researcher looked at model performance metrics derived from the confusion matrix. Evaluated using accuracy, Precision, Recall, and F- measure and the performance of the model was evaluated with four metrics (Accuracy, Precision, Recall, and measure). The developed motor insurance claim status prediction models have best prediction accuracy, and the two models have promising prediction accuracy. RF model prediction accuracy is slightly better than SVM model in the insurance domain specifically in motor insurance. The prediction accuracy of the model is capable of predicting the motor insurance claim status with 98.36% and 98.17% by RF and SVM classifiers respectively. As a result, RF classifier is slightly better than Multi-Class Support vector machines. Naïve Bayesian visualization provides an interactive view of the prediction results. The model commits some errors and has an accuracy of 78%.

Researcher [10] was identifying publications related to fraud detection through the use of ML techniques based on the Fraud Triangle Theory. Fraud detection is complex, as it requires the interpretation of human behavior, but this is not the only issue. The lack of data available for training or testing detection models significantly complicates the assessment of detection strategies. Even when data are available, unbalanced datasets are the norm in this domain. The large number of publications in conferences and journals representing 50% and 50% of primary studies respectively is substantial proof. In addition, results of the quality evaluation carried out for the primary of “relevance”, “limitations”, and “methodology”. The proposed reference frameworks focus on developing tools that allow auditors to perform fraud analyses more efficiently by shortening their detection time work aims to review current work related to fraud detection that uses the fraud triangle in addition to machine learning and deep learning techniques. Used the Kitchen am methodology to analyze the research works related to fraud detection from the last decade. Study provides evidence that fraud is an area of active investigation. Fraud theories associated with human behavior, this SLR reveals very little evidence from studies supporting this approach, since only one primary study was found, corresponding to 3.13% of the studies. Several works related to fraud detection using machine learning techniques the analysis and detection of fraud in which only theories related to fraud that were associated with human behavior were considered, seven primary studies (corresponding to 21.88%). As future work, it is proposed that a review focused on detecting fraud and incorporates an analysis of the availability of data and the lack of access to this resource, including other data sources as possible alternatives, should be carried out

Researcher [5] was identifying publications related to fraud detection through the use of ML techniques based on the Fraud Triangle Theory. Fraud detection is complex, as it requires the interpretation of human behavior, but this is not the only issue. The lack of data available for training or testing detection models significantly complicates the assessment of detection strategies. Even when data are available, unbalanced datasets are the norm in this domain. The large number of publications in conferences and journals representing 50% and 50% of primary studies respectively is substantial proof. In addition, results of the quality evaluation carried out for the primary of “relevance”, “limitations”, and “methodology”. The proposed reference frameworks focus on developing tools that allow auditors to perform fraud analyses more efficiently by shortening their detection time work aims to review current work related to fraud detection that uses the fraud triangle in addition to machine learning and deep learning techniques. Used the Kitchen am methodology to analyze the research works related to fraud detection from the last decade. Study provides evidence that fraud is an area of active investigation. Fraud theories associated with human behavior, this SLR reveals very little evidence from studies supporting this approach, since only one primary study was found, corresponding to 3.13% of the studies. Several works related to fraud detection using machine learning techniques the analysis and detection of fraud in which only theories related to fraud that were associated with human behavior were considered, seven primary studies (corresponding to 21.88%). As future work, it is proposed that a review focused on detecting fraud and incorporates an analysis of the availability of data and the lack of access to this resource, including other data sources as possible alternatives, should be carried out [6] present the case of using artificial intelligence to increase the client base of companies. The data used in this model is from KAGGLE. This database consists of 1000 car accidents and car insurance claims from Ohio, Illinois and Indiana from January 1, 2015 to March 1, 2015. Researcher was present three models for artificial intelligence, logistic regression, gradient boosting and random forest, Random Forest 0.821 % Logistic Regression 0.820 % Gradient Boosting 0.856. The results show that the best model can predict with 81% probability whether a potential customer would remain a customer of the company. Application of this model or similar models can be found in the detection of insurance fraud at the level of the entire industry, but also fraud at the level of certain insurance companies. The results show that Gradient Boosting is the best model can predict with 81%.

3. MATERIALS AND METHODS

Python is the programming language that we will be using for this study. It was first introduced by Guido van Rossum in 1991, when it was first released. The rise in popularity of the Python programming language over the last few years has been attributed to the increased use of Data Science. Python is an interpreted object-oriented programming and high-level language that is used in a variety of applications such as website backend code, data analysis projects, and is well-known for its use in scientific research.

In this research we use the Google Colab. ”Colab” is a product developed by Google Research. Anyone can create and run arbitrary python code in the browser with Colab, making it an excellent tool for machine learning, data analysis, and education of all kinds. Jupyter Notebooks can be used for free in Google Collaborator. Using this service is completely free, as it is a cloud-based service provided by Google. Among Colab's advantages is the fact that it doesn't require any pre-installation.

As above mentioned that it is an outstanding free version of a hosted Jupyter notebook that does not require any setup and provides access to Google processing resources such as GPUs and TPUs is provided by this service. When the browser is closed, a Colab's instance can run for up to 12 hours and 90 minutes before being declared idle and being recycled.

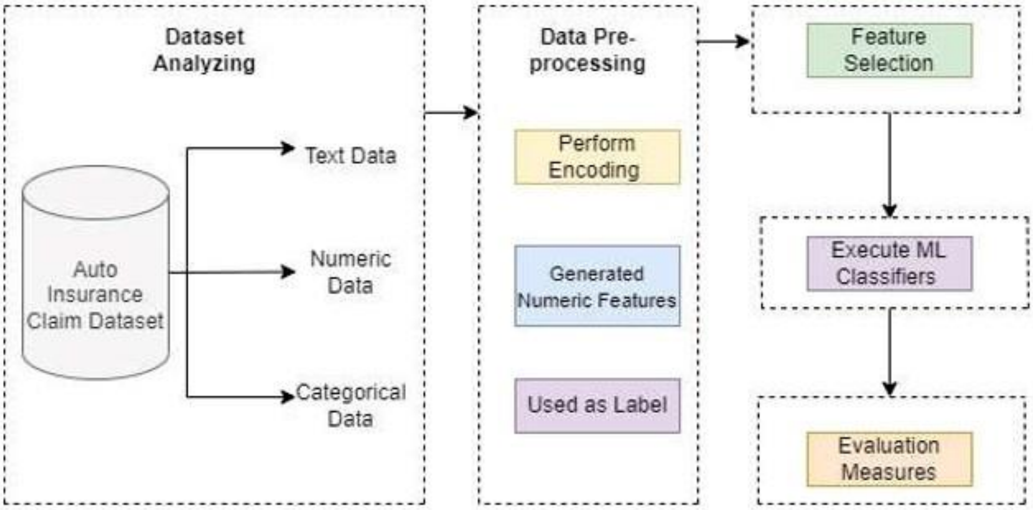
With Google Colab, you can study and construct Python machine learning models quickly. It is built on the Jupiter notebook and allows for collaboration. The notebooks can be shared and edited by the team members even when they are separated geographically. It is also possible to post the notebooks on GitHub and make them available to the broader public. Many well-known machine learning frameworks, such as PyTorch, Tensor Flow, Kara's, and Open CV, are supported by Colab. As of right now, the only limitation is that R and Scale aren't supported. Sessions and their duration are likewise restricted. Compared to the rewards, these are minor sacrifices. Different libraries are used in this research for building the model, one of them isdescribed here.

3.1. Sklearn:

scikit-learn (Sklearn) is the most convenient and powerful machine learning library in Python. It provides a powerful set of tools for machine learning and statistical modeling, including classification, regression, clustering, and dimensionality reduction, through a consistent Python interface.

David Cournapeau created it in 2007 as part of a Google Summer of Code project under the name sickie learn. Later, in 2010, FIRCA members Fabian Pedregosa, Gael Varoquaux, Alexander Gram fort, and Vincent Michel advanced the project and released the first beta version (v0.1) on February 1st.

3.2. NLTK: Python developers can utilize the NLTK framework to create applications that process human language data using statistical natural language processing and other related fields (NLP). Tokenization, parsing, categorization, stemming, tagging, and semantic reasoning are only some of the texts processing libraries included.



The Kaggle Database provided the dataset. The dataset consists of total 40 attributes and we discard those which is not suitable for our model. The insurance policy and the customer are both represented in the dataset, which also includes information on both of these parties. It also includes details regarding the incident that served as the basis for the claims. Along with policy information, this dataset includes vehicle datasets (attributes, model, accident information, etc). (Type, Tenure, etc.). The goal is to determine whether or not a claim application is fraudulent.

3.3. Feature Extraction

A dimensionality reduction technique called feature extraction divides a large amount of raw data into smaller, easier-to-process groups. These huge data sets share the trait of having many variables that demand a lot of computational power to process. The term ”feature extraction” refers to techniques that choose out specific variables and/or combine them to create features, hence minimizing the quantity of data that needs to be processed while still properly and fully characterizing the initial data set. By generating new features from the current ones, feature extraction attempts to decrease the number of features in a dataset (and then discarding the original features). The majority of the information in the original collection of features should then be summarized by this new, smaller set of features.

In this study, we have discarded unnecessary attributes from the dataset. the attributes named as: “_c39”, “auto _model”, “policy_ bind_ date”, “policy_ state”, “incident_ date”, “incident_ state”, “incident_ city”, “incident_ location”, “policy_ csl” are discarded. After removing these attributes we implement data analysis graphs and execute different machine learning classifiers.

Machine learning draws on many different areas of study, including but not limited to psychology, artificial intelligence, computer science, statistics, and optimization. Classification is a supervised learning technique used in machine learning to examine a given data set and create a model that divides the data into the appropriate number of distinct classes [8].

In the literature, there are a number of effective classification methods, such as the k-nearest-neighbor classifier [14] Bayesian networks, artificial neural networks [17] decision trees [15], [13]. Although K-nearest-neighbor approaches have the benefit of being simple to use, they are typically highly sluggish if the input data set is very large. The existence of irrelevant characteristics, however, greatly affects these [14]

To address the issue of detecting vehicle insurance fraud, we propose a number of supervised data mining classification algorithms in this section. In Chapter 2, we describe these algorithms and associated developments. In order to categories and forecast the characteristics of claims cases into ”fraud” or ”legitimate,” we observe that some of these classification algorithms have not been precisely explored in the vehicle insurance fraud business. A small amount of research has also been done on using unsupervised data mining techniques to detect fraud in auto insurance. For the sake of this thesis and our objectives, unsupervised learning data mining approaches are not relevant. As a result, our focus is mostly on supervised learning data mining approaches.

In order to achieve the goals of our thesis, this section is devoted to researching scholarly sources on the subject of detecting vehicle insurance fraud using typical data mining and machine learning approaches.

3.3. Supervised Data Mining Algorithms

The supervised learning techniques that will be discussed in this thesis make use of labelled training data so that learners are aware of the class to which each training sample belongs. Unlabeled data is used to evaluate the predictive model once it has been built using this data.

This chapter surveys a number of classifications of supervise learning-related data mining approaches; each of them produces a classifier that can be applied to prediction and classification of any classification problem. Additionally, the majority of them have been used extensively to offer fundamental solutions to issues relating to the classification of fraudulent data in vehicle insurance to manage fraud detection.

Here, we only cover methods for identifying false claims made under motor insurance. This issue has an impact on insurers' revenue and profit, but it also has an impact on the cost of the fraud, which might be in the millions in developing nations and billions in affluent ones. The learning strategy, benefits, and drawbacks are discussed in the following sections.

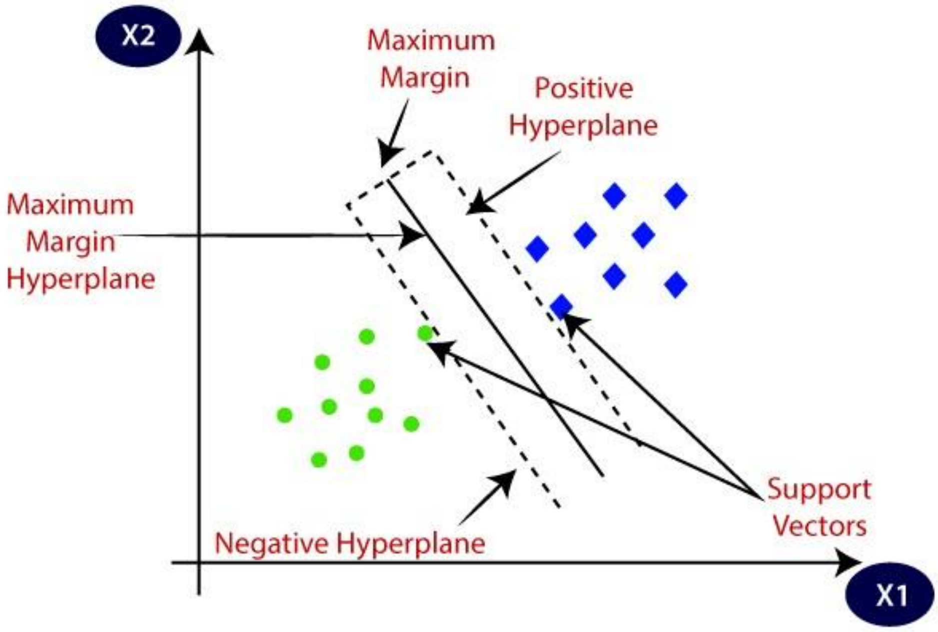
3.4. Naive Bayes Approach

A type of classifier that makes use of the Bayes Theorem is naive Bayes. It forecasts membership probabilities for each class, such as the likelihood that a specific record or piece of data falls under a specific class. The most likely class is that which has the highest likelihood.

According to (NB), an attribute value has no bearing on a particular class of attribute values. This implies that the values of other attributes have no bearing on the members of a class of a given attribute. Basically, the likelihood of this evidence with each class is determined in order to predict the class of a given attribute(s) using NB method. The class of that evidence is chosen based on which has the highest probability value [16].

3.5. Support Vector Machine SVM Approach

A novel class of machine learning techniques built on statistical learning theory is the support vector machine. The machine learning community now focuses its research on support vector machines due to their strong promotion and improved accuracy [17]. Getting a model that maximizes performance for the training data is the main goal of pattern classification. The models are created using traditional training techniques such that each input-output combination is accurately identified as belonging to the class to which it belongs. However, if the classifier is too well-suited to the training set, the model starts to memories the set rather than learning to generalize, which reduces the classifier's capacity for generalization [12]. The process of SVM is described in the following figure 3.3.



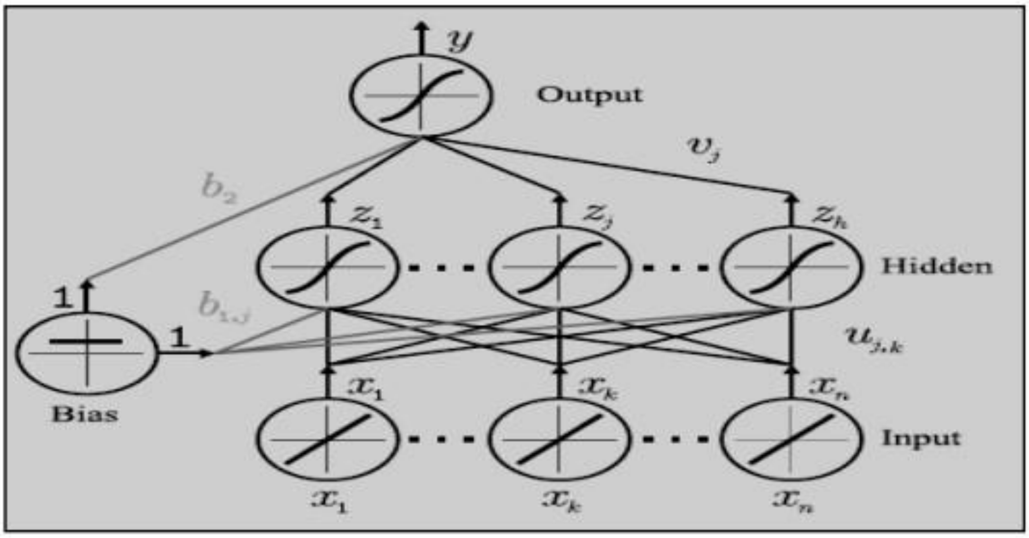
Support Vector Machine Classifier

3.6. Artificial Neural Network Approach

The primary objective of ANN is to efficiently and scalable generate nonlinear parameterized mappings between a collection of input variables and a set of output variables. Straightforward three-layer neural network (an input, hidden and an output layer). There were three layers in ANN. These layers are linked together and have adjustable weights. The main task of a processing unit in ANN is to take in signals and connections and (non-linearly) turn a weighted sum of these signals into a single outputted signal.

3.7. Logistic Regression Approach

The logistic model is a type of statistical analysis frequently used in the fields of classification and predictive analytics. Logistic regression is a statistical method used to predict whether a specific event, such as voting or not voting, will occur given a set of independent variables. The dependent variable is limited to values between 0 and



Three Layer Neural Network

1 because the outcome is a probability. Odds, defined as the probability of success divided by the probability of failure, are transformed using a log-it function in logistic regression.

Logistic regression is one type of supervised machine learning model used in the field of machine learning [16]. It's also a discriminative model, meaning it tries to establish differences between groups (or categories). It cannot, as the name suggests, generate information of the class that it is trying to predict, such as an image, in contrast to generative algorithms like naive Bayes (e.g. a picture of a cat).

3.8. K-Nearest Neighbor (K-NN) Approach

K-NN is a supervised classification technique for classifying and predicting object situations based on the characteristics shared by the objects' nearest neighbors. It is only at the server level that the K-NN algorithm is run. Until an item is categorized by a majority of its neighbors, K-NN continues to apply all of the calculations involved in the classification process. Furthermore, "Euclidean distance" is used to rank all training samples. Because induction is time-dependent, K-NN is sometimes referred to as Memory-Based Classification. As a result, this technique necessitated a lot of processing power to be applied to the training data [13]. Below, we'll take a look at one of the few works done on K-NN classifiers for the purpose of detecting vehicle insurance fraud.

K-NN was one of several data mining approaches compared using real-world examples of motor insurance fraud. There were 1,399 PIP claims from accidents in 1993 that were used to compile the sample data. Predictors that raised no red flags were also included in the analysis. The study indicated that 500-NN had an accuracy of 83.7o% and that 1-NN had an accuracy of 80.77% when employing ten-fold cross validation. When compared to other classifiers used in this study, such as NB, TAN, LS-SVM, MLP-NN, Log it model, and C4.5, K-performance NN's was the worst.

3.9. Ada-Boost

Ada-Boost is one of the most popular algorithms for building robust classifications using linear classifications of members. The classification of participants is chosen to minimize errors at each step during training. Ada Boost provides a simple and effective method for creating custom classifications. The performance of the ensemble depends on the differences between the member groups as well as the performance of each member group. However, the current Ada Boost algorithm focuses on the problem of detection of vehicle insurance fraud [15]

3.10. Confusion Matrix

Confusion matrix is used for performance measurement in machine learning classification on the test data. It consists of the 2 dimensions of classes, in which one is actual class and the other is predicated class. Using actual and predicated class, confusion matrix is built as True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN)

3.11. Precision

Precision is the percentage of positive samples that are accurately identified as True Positives out of the total number of positive samples (either correctly or incorrectly).

$$Precision = \frac{TP}{(TP + FP)}$$

3.12. Recall

The recall is the fraction of positive samples that were appropriately labelled as such. In order to evaluate how well a model works, we look at its recall. When there are more positive samples found, recall is said to be higher.

$$Recall = \frac{TP}{(TP + FN)}$$

3.13. F-Measure

The F1 score is the harmonic mean of the recall and accuracy scores. The F1 Score scale is from 0 to 1. A classifier's accuracy (as measured by a percentage of correctly classified examples) and its stability are both revealed (it does not miss a significant number of instances.

With high precision but low recall, you get a very precise result, but you also overlook many cases that are hard to categories. Generally speaking, our model Performs better as the F1 Score rises. It has a mathematical expression as:

$$F - Measure = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

3.14. Accuracy

Classification when we say something is accurate, we usually mean it to be accurate. It measures the proportion of accurate predictions to all input samples. We calculate the classification accuracy of 13 different classifiers, and evaluate our model.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

4. RESULTS AND DISCUSSION

The purpose of this research is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that there are far fewer frauds than legitimate insurance claims. This type of problem is called imbalanced classification.

Cheating is unethical and harmful to the company. By building models that can classify auto insurance fraud, insurers can reduce losses. The less you lose, the more you earn.

Insurance fraud is intentionally fraudulent claims made against insurance companies or agents for proceeds. Fraud can be committed by contractors, ensured persons, third party agents or professionals who assist the claimant at different stages of the transaction. Insurance agents and company employees can also commit insurance fraud. Common fraud includes "stuffing" or claim inflation, misrepresentation of facts in insurance claims, filing claims for injuries or damages that never occurred, and fabricating accidents.

Classifiers Name	**Precision** (%)	**Accuracy** (%)	**Recall** (%)
Logistic Regression	0.6125%	0.55%	0.324%
Support Vector Classifier	0.125%	0.69%	0.3030%
KNN Classifier	0.5%	0.48%	0.25%
Decision Tree	0.89%	0.85%	0.84%
Random Forest Classifier	0.88%	0.87%	0.88%
Linear Discrimination Analysis	0.87%	0.88%	0.90%
Ada Boost Classifier	0.87%	0.87%	0.89%
Gradient Boosting Classifier	0.92%	0.90%	0.89%
XGBoost Classifier	0.91%	0.89%	0.89%
Cat Boost Classifier	0.89%	0.89%	0.90%
Extra Tree Classifier	0.87%	0.88%	0.90%
Light GBM Classifier	0.87%	0.88%	0.90%

In the above table, you can see that, Logistic regression has 0.61 precision values, 0.55 accuracy and 0.324 recalls. The least performance shows by the KNN classifiers, with 0.125 precision value, 0.48 accuracy and 0.30 recall value. The best classifier among all of these twelve algorithms is Gradient Boosting Classifier with 0.90 accuracy, 0.92 precision and 0.89 recall values. Random Boost Classifier and Ada Boost Classifier gives the 0.87 accuracy. Linear Discrimination Analysis, Extra Tree Classifier and Light GBM Classifier gives the 0.88 accuracy. Following XG Boost and Cat Boost classifier gives the 0.89 accuracy which are the second highest score among all of these classifiers.

Conclusions

The research of motor insurance fraud detection is certainly not exception. Despite the recognized significance of data extraction methods in the detection of motor insurance fraud, the aforementioned framework or an organized review of their usage in motor insurance fraud detection studies is missing. Because of this, first of all, in this study, we carried out a review of academic articles and delivered an extensive bibliography and categorization framework for the usage of data mining and machine learning in motor insurance fraud detection. In this sense, our goal was to provide valued information for both scholars and practitioners of the respective fields in which certain techniques can be used in motor insurance fraud detection, and to report and assemble an organized review of the flourishing publications regarding motor insurance fraud detection. After the comprehensive literature review, along with the different motor insurance fraud schemes, in this study we presented an aggregated ranking of the most important motor insurance fraud indicators. While creating this aggregated ranking, we took the indicators presented in the literature into consideration, as well as the indicators recommended by the industry experts during our interviews and the ones present in our survey. Thus, the aggregated ranking can serve as a really good starting point for any future scholar or insurance company who wants to look into any part of motor insurance fraud detection. It can be helpful when insurance companies are working on their own accident reporting statements available on their website, or on their online claim reporting. Motor insurance fraud detection is a developing field where it is best to outpace the perpetrators. Furthermore, it is obvious that many aspects of intelligent fraud detection are still to be looked into. In the following section we will showcase some of the most important open issues related to motor insurance fraud detection and also recommend areas for possible upcoming research. Regarding the cost saving ability of the motor insurance fraud detection methods, it would be useful to examine the combination of our suggested method and the approach suggested handling the total damage (where the amount of compensation is equal to the value of the insured vehicle) and the partial damage claims separately.

REFERENCES

1. Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., ... & Dandona, L. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *The Lancet Planetary Health*, *3*(1), e26-e39.
2. George, A. (2017). Abstracts of 2016 Conference. *NIGERIAN JOURNAL OF DERMATOLOGY*, *7*(1).
3. Al-Amir Khadim, M. A., Al-Wahid, W. A. A., Hachim, D. M., & Sopian, K. (2021). Experimental study of the performance of cylindrical solar still with a hemispherical dome. *Smart Science*, *9*(1), 30-39.
4. Sánchez-Aguayo, M., Urquiza-Aguiar, L., & Estrada-Jiménez, J. (2021). Fraud detection using the fraud triangle theory and data mining techniques: A literature review. *Computers*, *10*(10), 121.

5. Huang, S. Y., Lin, C. C., Chiu, A. A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, *19*, 1343-1356..
6. Todevski, D. (2021). Fraud detection in insurance with machine learning model. *Knowledge-International Journal*, *47*(5), 881-886.
7. El Emary, I. M. (Ed.). (2020). Handbook of research on artificial intelligence and soft computing techniques in personalized healthcare services/edited. In *International Conference on E-Commerce and Internet Technology (ECIT)(pp* (Vol. 245, p. 247).
8. Erdal, H., Erdal, M., Simsek, O., & Erdal, H. I. (2018). Prediction of concrete compressive strength using non-destructive test results. *Computers and Concrete*, *21*(4), 407-417.
9. Akhter, I., Jalal, A., & Kim, K. (2021, January). Pose estimation and detection for event recognition using Sense-Aware features and Adaboost classifier. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)* (pp. 500-505). IEEE.
10. Singh, K., Rajora, S., Vishwakarma, D. K., Tripathi, G., Kumar, S., & Walia, G. S. (2020). Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. *Neurocomputing*, *371*, 188-198..
11. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.
12. Yan, X., Zhu, J., Kuang, M., & Wang, X. (2019). Aerodynamic shape optimization using a novel optimizer based on machine learning techniques. *Aerospace Science and Technology*, *86*, 826-835..
13. Giglioni, V., García-Macías, E., Venanzi, I., Ierimonti, L., & Ubertini, F. (2021). The use of receiver operating characteristic curves and precision-versus-recall curves as performance metrics in unsupervised structural damage classification under changing environment. *Engineering Structures*, *246*, 113029.
14. Yildirim, M., & Cinar, A. (2020). Classification of Alzheimer's Disease MRI Images with CNN Based Hybrid Method. *Ingénierie des Systèmes d Inf.*, *25*(4), 413-418.
15. Serra, A., Fratello, M., Cattelani, L., Liampa, I., Melagraki, G., Kohonen, P., ... & Greco, D. (2020). Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. *Nanomaterials*, *10*(4), 708.
16. Velu, A. (2021). Application of logistic regression models in risk management. *Management*, *8*(4).
17. Wu, D., Zheng, S. J., Bao, W. Z., Zhang, X. P., Yuan, C. A., & Huang, D. S. (2019). A novel deep model with multi-loss and efficient training for person re-identification. *Neurocomputing*, *324*, 69-75.