International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 13, Issue 3, March 2024)

Advancing Insurance Fraud Detection: Integrating Machine Learning Techniques and Ethical Considerations

Naina Singh[1],Prof. Pradeep Tripathi[2]
Research Scholar, Department of Computer Science & Engineering[1]
Professor , Department of Computer Science & Engineering[2]
Vindhya Institute of Technology and Science - [VITS],Satna, Madhya Pradesh

***Abstract*-- In the current era, with the widespread adoption of credit cards following demonetization, the likelihood of fraudulent activities has increased. Banks possess vast databases holding crucial business information, making them vulnerable to fraud. This issue extends across various sectors, including banking, government, corporate, and consumer domains. The growing reliance on advanced technologies like cloud and mobile computing has exacerbated the problem, rendering traditional manual detection methods, such as auditing, ineffective and costly. Consequently, financial institutions are increasingly turning to automated processes leveraging numerical and computational techniques. Data mining-based approaches have emerged as a viable solution, offering the ability to detect subtle anomalies within large datasets. By employing supervised algorithms, we aim to enhance fraud detection accuracy. Given the diverse nature of fraud and the array of data mining methods available, ongoing research seeks to identify optimal approaches for different scenarios. Financial fraud, encompassing deliberate illegal practices for financial gain, poses significant economic and societal repercussions, with credit card fraud alone resulting in substantial revenue losses annually.**

Keywords: Fraud detection, Financial fraud, Decision tree.

I. INTRODUCTION

Fraud entails the exploitation of a profit organization's system without necessarily triggering direct legal ramifications. It is a universal practice aimed at deceiving individuals or organizations for financial gain. Credit card fraud detection involves discerning between lawful and fraudulent transactions, which can be broadly categorized into traditional card-related and internet frauds. Customer fraud, perpetrated by individuals external to the organization, contrasts with management fraud, orchestrated by top-level management within the organization. Fraud detection, integral to overall fraud control, automates and streamlines the screening process, minimizing manual intervention. Credit card fraud involves unauthorized account activity, where individuals misuse another person's credit card without the cardholder's or issuer's knowledge. Rapid identification of fraud post-occurrence is crucial, with fraud detection methods continually evolving to counter offenders' strategies. Data mining, extracting knowledge from extensive datasets, encompasses supervised learning, utilizing labeled fraud and genuine cases, and unsupervised learning, employing unlabeled data. The data mining process involves defining the problem, data preparation, exploration, model building, validation, deployment, and updating. Implementation of effective fraud prevention measures and detection techniques has become imperative in today's fiercely competitive environment to mitigate financial losses and organizational setbacks caused by fraud. Commercial fraud, an organized crime, encompasses various unlawful activities, including identity theft and asset misappropriation. Data mining, extracting actionable insights from large databases, is integral to developing fraud detection systems, facilitating the identification of relevant patterns from extensive datasets. Selecting task-relevant attributes from these datasets poses a significant challenge in designing robust fraud detection systems

II. LITERATURE SURVEY

[Jarrod West, Maumita Bhattacharya]"Intelligent Financial fraud detection"

This author explains about different intelligent approaches to fraud detection which are both statistical and computational though the performance was differed each technique was shown to be reasonably capable at detecting various forms of financial fraud. The ability of the computational methods such as neural networks and support vector machines to learn and adapt to many new techniques is highly effective to the evolving of tactic fraudsters. Initial fraud detection studies focused heavily on statistical models such as logistic regression, as well as neural networks. Neural networks are used for financial applications such as forecasting. Neural network are well established history with fraud detection. But they require high computational power for training and operation, making it unsuitable for real-time function. Potential for over fitting if training set is not a good representation of the problem domain, so requires constant retraining to adapt to

new methods of fraud. In this paper the author says about the different kinds of frauds i.e., insurance fraud , mortgage fraud , health insurance fraud , telecommunication fraud , credit card fraud. Different techniques have been defined for different kinds of frauds defining the parameters like entropy, sensitivity and comparing the efficiency of the different kinds of algorithms and representing them in a graphical representation.

[Rasa kanapickiene, Zivile Grundiene] "The model of fraud detection by means of financial ratios"

This author explains about how financial ratios are analysed in order to determine the most fraud-sensitive ratios of financial statements with regard to company managers' and employees' motivation to commit fraud. It was found out that in most cases fraud is committed to show that the company keeps growing and to fulfill obligation conditions. Literary sources offer a wide range of such ratios. Theoretical analysis showed that profitability, liquidity, activity and structure ratios are analyzed most often. Theoretical survey revealed that, in scientific literature, financial ratios are analyzed in order to designate which ratios of the financial statements are the most sensitive in relation with the motifs of executive managers and employees of companies to commit frauds. The logistic regression model of fraud detection in financial statements has been developed.

[Fletcher H. Glancy, Surya B. Yadav] "A computational model for financial reporting fraud detection"

This author explains that the computational fraud detection model is possible to detect financial exposure fraud from the text of annual filings with the Security and Exchange Commission. The model is generalizable because it specifies automatable steps that can be adapted to other domains and genres. A potential application for CFDM is to screen companies for investigation of potential fraud by the SEC (Security and exchange commission). Additional potential applications include financier analysis, e-mail spam detection, and business intelligence validation. A computational fraud detection model (CFDM) was proposed for detecting fraud in financial reporting. CFDM uses a quantitative approach on textual data. It incorporates techniques that use essentially all of information contained in the textual data for fraud detection. Extant work provides a foundation for detecting deception in high and low synchronicity computer- mediated communication (CMC). CFDM provides an analytical method that has the potential for automation. It was tested on the Management's Discussion and Analysis from 10-K filings and was able to distinguish fraudulent filings from non-fraudulent ones. CFDM can serve as a screening tool where deception is suspected.

Siddhartha Battacharya, Sanjeev jha , Kurian Thanakunnel, J Christopher Westland: Data Mining for credit card fraud:

This author says that with the growth in credit card transactions, as a share of the payment system ,there has also been increase in the credit card fraud and most of the U.S consumers are noted to be significantly concerned about identity fraud. While predictive models for credit card fraud detection are in active in use practice, reported studies on the use of web data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. In this paper the author evaluates two advanced data mining approaches , support vector machines and random forests. Together with well known logistic regression as part of an attempt to better detect credit card fraud. In this paper the Statistical fraud detection methods have been divided in to two broad categories: supervised and unsupervised. In supervised fraud detection methods , models are estimated based on the samples of fraudulent and legitimate transactions to classify new transactions as fraudulent or legitimate. In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both thee fraud detection methods predict the probability of fraud in any given transaction. Predictive models for credit card fraud detection are in active use. Other techniques reported for credit card fraud detection include case based reasoning and hidden Markov models. Support vector machines and random forests are sophisticated data mining techniques which have been noted in recent years to show superior performance across different applications. The choice of these two techniques together with the logistic regression is based on their accessibility for practitioners and noted performance advantages.

III. RELATED WORK

The realm of financial fraud detection is continuously evolving, with a focus on outpacing perpetrators. However, there are still aspects of intelligent fraud detection that remain unexplored. Surveys on fraud detection reveal the existence of various fraud types and computational techniques aimed at detecting fraudulent activities perpetrated by fraudsters. These techniques involve computing different parameters for each algorithm and representing computing time graphically. Researchers have utilized diverse datasets, including the German credit card dataset and data from other countries

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 13, Issue 3, March 2024)

like China, to develop computational methods for fraud detection and evaluate the accuracy of different algorithms. Current systems employ ID3 and support vector machine algorithms for fraud detection, analyzing the percentage of fraud occurrences and comparing various parameters across algorithms. Fraud detection plays a crucial role in the modern finance industry. The proposed system involves fraud detection using supervised learning algorithms such as decision tree learning and Naive Bayes classifier, with a focus on comparing their building times. Despite variations in performance, each technique demonstrates reasonable capability in detecting different forms of financial fraud. Particularly, the adaptability of computational methods like decision trees and Bayesian classifiers to learn and respond to evolving fraud tactics is highly effective. By leveraging available datasets, users can be classified as good or bad based on their ability to repay loans, with positive and negative counts representing each category, and sensitivity and efficiency can be calculated and graphically represented.

IV. Types Of Frauds

There are different types of frauds they are: credit card fraud, financial fraud, mortgage fraud, insurance fraud , telecommunication fraud.

Credit card fraud:

This fraud is defined as the method of purchasing and marketing goods without having money. It is a small plastic card to provide the credit service to the customer. Now a days credit card plays a important role in automated business and online money transaction area which is increasing every year. With the growth of usage of the credit card, fraudsters are finding more opportunities to commit the fraud which causes huge loss to cardholders and banks. Credit card fraud is classified in to two two types:

- Offline credit card fraud:

This kind of fraud is done physically which means the plastic card is stolen by fraudsters and using the card in stocks or supplies or stores or for different purposes as an actual owner. It is an unusual type of fraud because financial organizations will immediately block the card immediately when the card holders report about the theft.

- Online credit card fraud:

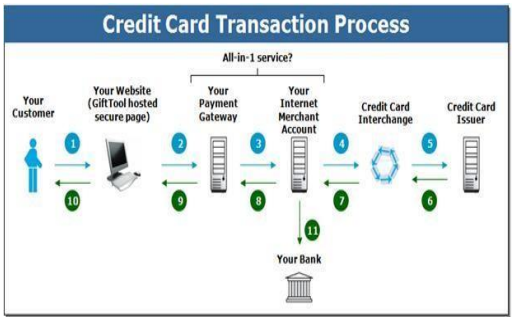
This kind of fraud is popular and it is very dangerous, the credit card's information is stolen by the fraudsters to be used in future online transactions by internet or by phone. This kind of fraud is also called as "cardholder not existing" fraud. The card holders can be obtained by the fraudsters through the skimming, phishing or credit card generators.

There is another classification for credit card fraud they are application fraud and behavioral fraud. This classification is based on fraudster's strategy on compelling the fraud. Application fraud occurs when the user enters any wrong evidence and wrong details in to the presentation for opening a new credit card. Fraudsters may use other persons information to obtain credit cards or get their new credit cards by using false information with the intention of the never repaying the purchases. Behavioral fraud occurs when fraudsters obtain credit card holder details to use them later for sales which are made on a cardholder present basis.

V. Supervised Learning Algorithms

Supervised learning algorithms are defined as the desired output is known for the input provided in these kind of algorithms we have an input and the desired output is known and we need to map a function for these values . In these supervised learning algorithms predictions are made on the known training dataset and it will be accurate. These learning algorithms are further grouped into regression and classification problems. The Supervised learning algorithms uses a supervised training data where it contains supervised examples. The supervised learning algorithm analyzes the training dataset and produces an classifier. For this initially we need to collect the accurate training dataset and we need to find the accuracy of the function. It is the machine learning task of inferring a function from supervised training data. The training data consists of training examples. In supervised learning , each example is a pair consisting of an input object and a desired output value. a supervised learning algorithm.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 13, Issue 3, March 2024)



Introduction to decision tree algorithm:

To find an optimal way to classify the learning set initially we need to minimize the depth of the tree. To minimize the tree we need some function information gain. In order to define information gain precisely we need to calculate entropy first.

Entropy:

Without loss of generality, that the resulting decision tree classifies instances into two categories, we'll call them P(positive)and N(negative).

Given a set S, containing these positive and negative targets, the entropy of S related to this boolean classification is: Entropy(S)=

- P(positive)log2P(positive) - P(negative)log2P(negative) P(positive): proportion of positive examples in S P(negative): proportion of negative examples in S Information gain:

To minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice. We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node. The information gain, Gain(S,A) of an attribute A,

Gain(S,A)= Entropy(S) - Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv) The intention of this ordering is:

1. To create small decision trees so that records can be identified after only a few decision tree splitting.

2. To match a hoped for minimalism of the process of decision making

Information gain:

To minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice. We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.

The information gain, Gain(S,A) of an attribute A,

Gain(S,A)= Entropy(S) - Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv) The intention of this ordering is:

1. To create small decision trees so that records can be identified after only a few decision tree splitting.

2. To match a hoped for minimalism of the process of decision making Maximum(Gain(S,A)). Create child nodes of this rootNode and add to rootNode in the decision tree.

For each child of the rootNode, apply ID3(S,A,V) recursively until reach node that has entropy=0 or reach leaf node. End ID3.

Naïve bayes classifier:

Introduction to Bayesian Classification The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Uses of Naive Bayes classification:

1. Naive Bayes text classification The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 13, Issue 3, March 2024)

2. Spam filtering: Spam filtering is the best known use of Naive Bayesian text classification. It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email (sometimes called "ham" or "bacn").[4] Many modern mail clients implement Bayesian spam filtering. Users can also install separate email filtering programs. Server-side email filters, such as DSPAM, Spam Assassin, Spam Bayes, Bogofilter and ASSP, make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within mail server software itself. 3. Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering (http://eprints.ecs.soton.ac.uk/18483/) Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. It is proposed a unique switching hybrid recommendation approach by combining a Naive Bayes classification approach with the collaborative filtering. Experimental results on two different data sets, show that the proposed algorithm is scalable and provide better performance–in terms of accuracy and coverage–than other algorithms while at the same time eliminates some recorded problems with the recommender systems.

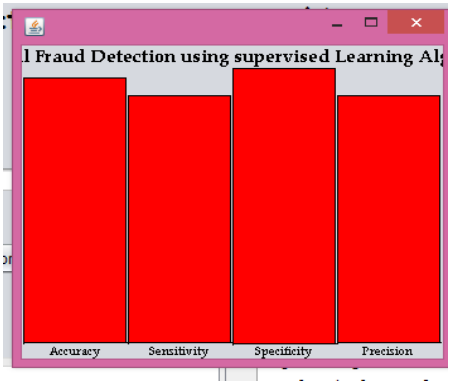


Fig 1.Bar Chart for accuracy measures

VI. CONCLUSION

Credit card fraud has become more and more rampant in recent years. To improve merchants' risk management level in an automatic and effective way, building an accurate and easy handling credit card risk monitoring system is one of the key tasks for the merchant banks. One aim of this study is to identify the user model that best identifies fraud cases. There are many ways of detection of credit card fraud. If one of these or combination of algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks. And a series of anti-fraud strategies can be adopted to prevent banks from great losses before and reduce risks. This paper gives contribution towards the credit card fraud detection using the supervised learning algorithms.

REFERENCES

[1]Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4, Issue 2, 2009 .
[2]Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .
[3]Vladimir Zaslavsky and Anna Strizhak," credit card fraud detection using selforganizing maps", information & security. An International Journal, Vol.18,2006.
[4]L. Mukhanov, "Using bayesian belief networks for credit card fraud detection," in Proc. of the IASTED International conference on Artificial Intelligence and Applications, Insbruck, Austria, Feb. 2008, pp. 221– 225.
[5]John T.S Quah,M Sriganesh "Real time Credit Card Fraud Detection using Computational Intelligence" ELSEVIER Science Direct,35 (2008) 1721-1732.
[6]Joseph King –Fung Pun, "Improving Credit Card Fraud Detection using a Meta Heuristic Learning Strategy" Chemical Engineering and Applied Chemistry University of Tornto 2011.
[7]Kenneth Revett,Magalhaes and Hanrique Santos "Data Mining a Keystroke dynamic Based Biometric Dtatabase Using Rough Set" IEEE
[8]Linda Delamaire ,Hussein Abdou and John Pointon, "Credit Card Fraud and Detection technique", Bank and Bank System,Volume 4, 2009.