International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Machine Learning-Based Fraud Detection In Banking Transactions

Siravuri Raghu Varma[1], Kuppli Mokshit Srinivasa[2], Nikhil[3], K Adithya[4]

[1,2,3,4] *Student, GITAM(Deemed to be University), Visakhapatnam ,Andhra Pradesh, India.*

Abstract— Financial fraud detection is crucial across sectors. This study presents a machine learning model to proactively identify fraudsters. It analyzes transactions to predict fraud, splitting data for precise training. Using machine learning, transactions are categorized as fraudulent or authentic. Results show high recall, accuracy, precision, and F1-score, indicating effective fraud prediction. This module offers a simple yet potent tool to curb financial fraud, saving costs and preserving integrity. We used algorithms like KNN algorithm, random forest algorithm, Adaboost algorithm and Decision tree classifier. Organizations can leverage such technology to detect and prevent fraudulent activities, safeguarding resources and reputation.

Keywords— *Transaction Analysis, KNN Algorithm, Random Forest, Ada Boost and Decision Tree Classifier.*

Introduction

1.1 Introduction

The employment of dishonest, unlawful, or misleading practices to obtain financial benefits is known as financial fraud. Fraud can occur in a variety of financial contexts, such as banking, insurance, taxation, corporations, and more. A rising issue is fiscal fraud and evasion, which includes money laundering, tax evasion, cc fraud, and finance fraud. Even with efforts to eradicate financial fraud, many dollars are lost annually, which has a negative impact on society and industry. Banks, retailers, and individuals have all been severely impacted by this significant financial loss. These days, there is a marked increase in fraud efforts, which emphasizes the importance of fraud detection. Regarding the certified examiners, 10% of occurrences are involved.

Financial statement fraud is less common than asset misappropriation and corruption, yet the money consequences of these types of laundering cases are not that serious. When auditing, illustrating the reasons behind a person's choice to commit fraud. The three components of the money triangle, motivation, and rationalizing work together to promote fraudulent behavior and raise the risk. The study focuses more on money fraud. Money statements include information about a company's operations and money-wise performance like income rate earnings on the company. This section presents the net income of the company after deducting expenses from revenues. An up-to-date picture of their assets, stakeholders and shareholders' stocks by the finance sheet. The cash flow statement evaluates how well a business generates enough cash to cover its debt payments, investments, and operating costs. Money notes are extra details that offer classification and more details regarding certain things.

The openings of some events, assets, and modified account policies are among the topics covered in these notes. These disclosures are essential in order to support the money presented on the money statements. Bank transaction fraud is the act of manipulating money statements to make a company appear more bigger than it used to be, and boost stock values. In auditing, the money triangle serves as a route to illustrate the reasons behind a person's choice to commit fraud. The tri-components of the money triangle motivation and rationalization where all work together to promote fraud and raise money fraud.

This hypothesis has been widely applied by auditing experts to handle and to assess financial fraud, knowledge of the fraud triangle is essential. According to Gupta and Singh, the likelihood of fraud rises when there are incentives present, such as the need to meet goals or make up for losses. The business will face pressure or temptation to engage in dishonest business activities. In addition, the absence of inspections or ineffective controls creates a suitable environment for fraud. The process of rationalization occurs when the person aims to justify the fake action which could be affected by other people and their conditions.

The interactions and activities of a single, crucial system component—the performance analyst—are depicted in the image. They are the machine learning model's primary developers. Through data-driven insights, they play a crucial role in improving their performance, boosting efficiency, and accomplishing strategic goals. Their primary responsibilities include selecting data, loading it, pre-processing, separating it, classifying it, creating machine learning models that can anticipate and produce the needed results. The first step in the performance analyst/machine learning engineer's job is choosing and loading the dataset into the model. They can now read and comprehend the data.

we leverage the Random Forest method, K-Nearest Neighbors (KNN), and AdaBoost algorithm to present a novel way to improve fraud detection in financial statements. Using these three potent techniques on datasets that have experienced substantial dimensionality reduction is part of our methodology. This method is called the "Three Algorithms on Reduced Dimensionality Datasets" approach. Our method's main goal is to solve the problem of financial data fraud detection, which frequently entails huge, intricate datasets with a variety of attributes. Our goal is to simplify the analysis process while keeping the essential information required for precise fraud detection by minimizing the dimensionality of the data.

Multiple models are combined in ensemble learning to get superior predictive performance compared to using only one model. High-dimensional data handling and identifying intricate correlations between variables are areas in which Random Forest shines. In contrast, KNN is good at finding patterns in data by comparing them using similarity metrics. By gradually training weak classifiers on various data subsets, AdaBoost improves their performance. Even when dealing with reduced-dimensional datasets, our method continuously produces fraud detection findings with excellent accuracy.

2.EXISTING SYSTEM

2.1 Literature Survey 1

Title: Evaluation of financial statements fraud detection research: A multi-disciplinary analysis

Authors: 1. Albizri, D. Appelbaum, and N. Rizzotto

Publication: Int. J. Discl. Governance, vol. 16, no. 4, pp. 206–241, Dec. 2019.

Result: point out on how damaging financial reporting fraud can be across different parts of the economy. They sat that if we bring together insights and efforts from various areas of research and detection, we could do a better job of dealing with the problem. It stresses how crucial it is to address financial fraud and its detrimental impact on the economy. It explains that one of the immediate consequences of financial reporting fraud is companies going under or facing significant financial setbacks. This, in turn, leads to a drop in stock prices and losses for investors as soon as the fraud is disclosed. The authors also talk about the significance of audit analytics. [1] They argue that audit analytics can improve the quality of audits by offering more comprehensive insights.

2.2 Literature Survey 2:

Title: An Application of Ensemble Random Forest Classifier for Detecting Financial Statement Manipulation of Indian Listed Companies.

Authors: Kalita, J., Balas, V., Borah, S., Pradhan, R.

Publication: Advances in Intelligent Systems and Computing, vol 740.

Result: this paper gives us a thorough overview of previous studies in accounting and information systems concerning the detection of financial statement manipulation. Notable companies like WorldCom, Xerox, Enron, as well as Indian firms like Satyam, Kingfisher, and Deccan Chronicle, have been involved in fraudulent financial reporting. Therefore, it's crucial to establish a robust framework for detecting financial fraud efficiently. Such a framework would benefit regulators, investors, governments, and auditors by preventing potential fraud cases. [2]

2.3 Literature Survey 3:

Title: Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China.

Authors: Yao J, Pan Y, Yang S, Chen Y, Li Y.

Publication: A Multi-Analytic Approach. Sustainability. 2019; 11(6):1579.

Result: While many researchers have focused on fraud detection in recent years, they often overlook both financial and non-financial indicators using a comprehensive approach. This study, however, examines financial statement fraud using 17 financial and 7 non-financial variables through six data mining techniques: support vector machine (SVM), classification and regression tree (CART), back propagation neural network (BP-NN), logistic regression (LR), Bayes classifier (Bayes), and K-nearest neighbour.[3]

2.4 Literature Survey 4:

Title: "An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning".

Authors: W. Xiuguo and D. Shengyong.

Publication: IEEE Access, vol. 10, pp.

Result: A promising result in leveraging textual features for enhanced financial fraud detection. [4]

2.5 Literature Survey 5:

Title: ‘‘An empirical analysis of the relation between the board of director composition and financial statement fraud’’
Authors: M. S. Beasely.

Publication: Accounting Rev., vol. 71, pp. 443–465, Oct. 1996.
Result: However, current approaches mostly focus on quantitative financial ratios, neglecting the valuable textual information, especially in Chinese managerial comments. This study aims to enhance fraud detection by integrating state-of-the-art deep learning models with numerical features from financial statements and textual data from managerial comments in annual reports of 5130 Chinese listed companies. First, a comprehensive financial index system is constructed, including both financial and non-financial indices often overlooked in previous research. Then, textual features from the section of annual reports are extracted using word vectors. Deep learning models are applied and their performances are compared using numeric data, textual data, and a combination of both. [5].

3. Proposed Method

3.1 Objective

Create a machine learning model that is capable of accurately distinguishing between which is fraud and non-fraud financial transactions. Use data preprocessing techniques to cleanse and prepare the dataset for efficient fraud analysis. Examine various machine learning algorithms which can are suitable for fraud detection, like logistic regression, decision trees, random forests, and neural networks. Train the machine learning model on historical transaction data to learn the patterns and indicators of fraud activities. Investigate for feature importance to understand the factors contributing most significantly to fraud detection. Conduct thorough testing and validation of the fraud detection system to ensure reliability and effectiveness in real-world scenarios.

Insufficient methods exist in financial systems to identify fraudulent transactions. Being unable to reliably spot fraudulent activity in the middle of a high volume of transactions. It’s hard to forecast future fraud using transaction data from the past. A lack of automation in spotting unusual or suspicious trends that point to fraud.

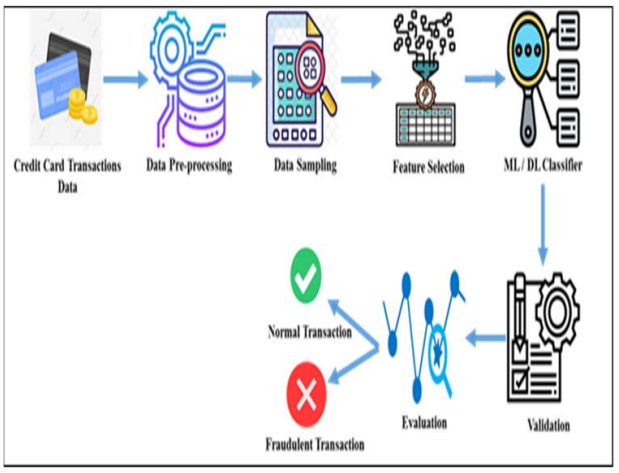

Fig-1: Methodology Flow Diagram

The interactions and activities of a single, crucial system component—the performance analyst—are depicted in the image. They are the machine learning model's primary developers. Through data-driven insights, they play a crucial role in improving their performance, boosting efficiency, and accomplishing strategic goals. Their primary responsibilities include selecting data, loading it, pre-processing, separating it, classifying it, creating machine learning models that can anticipate and produce the needed results.

The first step in the performance analyst/machine learning engineer's job is choosing and loading the dataset into the model. They can now read and comprehend the types of data found in the dataset they loaded into the model. They now sanitize the data appropriately. For training and testing purposes, the data is divided into two categories. The machine learning algorithm that the engineer uses during training aids in the appropriate classification of the data and provides us with the desired result. The engineer implements training and tests the algorithm after deciding which one to utilize for their model. when the needed output has been predicted by the algorithm. They now again evaluate the uml actions and make the required adjustments.

Computational intelligence research has seen a significant increase in interest in systems for identifying financial statement fraud. Various classification techniques have been used to automatically identify fraudulent businesses. Nevertheless, prior work has focused on creating extremely precise detection systems, ignoring the systems' interpretability. To create a highly interpretable system in terms of rule complexity and granularity.

3.2 Data Preprocessing Techniques:

-Data pre-processing is a process of removing the unnecessary data from the dataset. Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.

-This step also includes cleaning the dataset by removing unwanted or corrupted data that can affect the precision of the dataset, which makes it more easy to use.

3.3. Machine Learning Algorithm:

- Random forest: Random Forest isolates outliers by randomly choosing a feature from a given set of features and then randomly choosing a split value between the max and min values of that feature.

- Decision Trees Classifier: dividing the data recursively into subgroups according to each node's most important feature. This procedure keeps on until a predetermined endpoint is reached, like a maximum depth or purity threshold. Decision trees are useful for comprehending the underlying decision-making process since they are simple to understand and intuitive.

- AdaBoost: AdaBoost M1 by the authors Freund and Schapiro. More recently it may be similar to discrete AdaBoost because it is used for classification not for regression. AdaBoost is used to boost the performance.

3.4. Model Evaluation and Validation:

- Scikit-learn includes tools for evaluating and validating machine learning models, including cross-validation, grid search, and performance metrics such as accuracy, precision, recall, F1-score, and ROC curves.

3.5. Pipeline and Workflow:

Scikit-learn supports the creation of machine learning pipelines, allowing users to chain together multiple preprocessing steps and models into a single workflow. This simplifies the process of building and deploying a good workflow system.

3.6. Active Development and Community Support:

Scikit-learn is actively developed and maintained by a large community of contributors. It is well-documented, with extensive tutorials, examples, and API documentation available to users.

3.7. Open-Source and Free:

Scikit-learn was released under a permissive open-source license (BSD), making it free to use and distribute for both commercial and non-commercial purposes. Overall, scikit-learn is like a powerful and versatile library that empowers users to explore, experiment, and deploy machine learning solutions effectively.

It is intuitive interface, comprehensive algorithms, and integration with other scientific computing tools which make it a valuable resource for researchers, practitioners, and educators in the field of machine learning and data science.

4. Results & Discussions

we leverage the Random Forest method, K-Nearest Neighbors (KNN), and AdaBoost algorithm to present a novel way to improve fraud detection in financial statements. Using these three potent techniques on datasets that have experienced substantial dimensionality reduction is part of our methodology. This method is called the "Three Algorithms on Reduced Dimensionality Datasets" approach.

Our method's main goal is to solve the problem of financial data fraud detection, which frequently entails huge, intricate datasets with a variety of attributes. Our goal is to simplify the analysis process while keeping the essential information required for precise fraud detection by minimizing the dimensionality of the data.

Multiple models are combined in ensemble learning to get superior predictive performance compared to using only one model. High-dimensional data handling and identifying intricate correlations between variables are areas in which Random Forest shines. In contrast, KNN is good at finding patterns in data by comparing them using similarity metrics. By gradually training weak classifiers on various data subsets, AdaBoost improves their performance.

4.1 Experimental Results

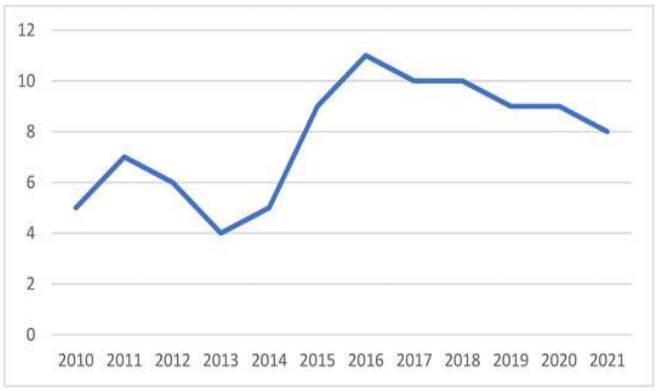

Fig-2: Screenshot displaying Article Rate

Classification Report Train

	precision	recall	f1-score	support
0	1.00	1.00	1.00	199020
1	1.00	1.00	1.00	199021
accuracy			1.00	398041
macro avg	1.00	1.00	1.00	398041
weighted avg	1.00	1.00	1.00	398041

Classification Report Test

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85295
1	1.00	1.00	1.00	85294
accuracy			1.00	170589
macro avg	1.00	1.00	1.00	170589
weighted avg	1.00	1.00	1.00	170589

Fig-3: Table displaying of Classification Scores

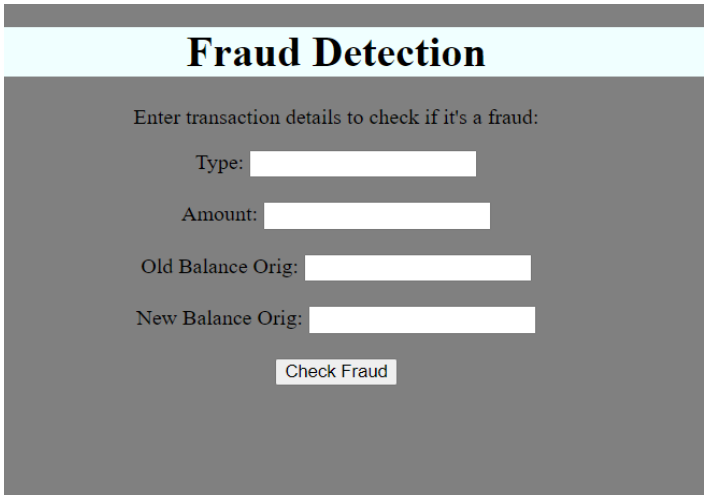

Fig-4: Figure depicting User Interface

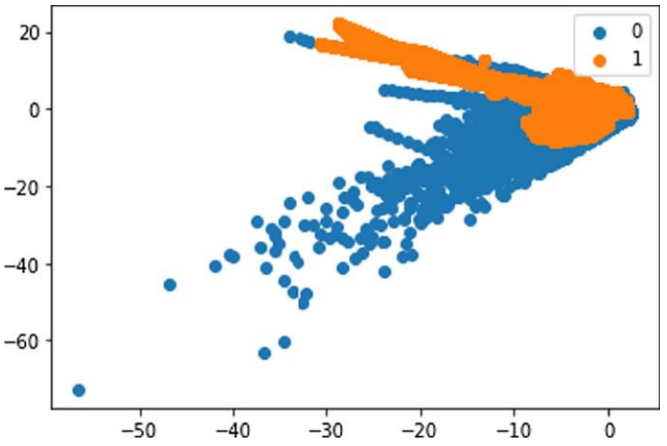

Fig-5: Figure displaying the Plot of the data set

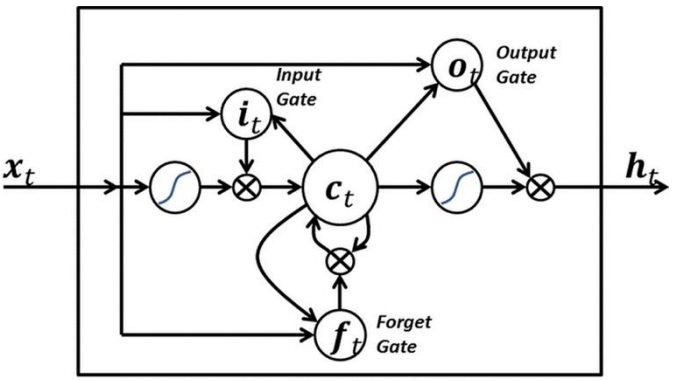

Fig-6: Displaying the unit structure

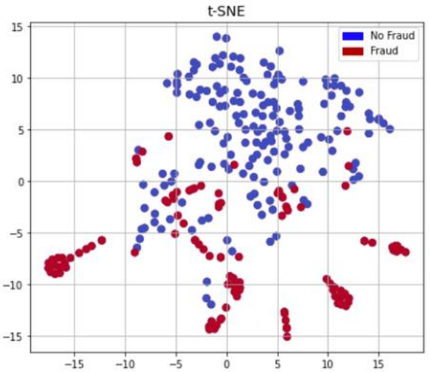

Fig-7: Finals Results being displayed in Plots

5. Conclusion and Future Enhancements

Through the application of ensemble learning and dimensionality reduction techniques, we provide regulatory authorities and financial institutions with an effective instrument to prevent fraud and protect

Even when dealing with reduced-dimensional datasets, our method continuously produces fraud detection findings with excellent accuracy. We have proven via thorough testing and analysis that our classification model either matches or surpasses the accuracy of current fraud detection methods. Additionally, we have demonstrated the superiority of our methodology in terms of accuracy and efficiency by comparing its performance with graphical methods that are frequently used in fraud detection.

Advantages: Request data, including form submissions, and query parameters, and can identify the frauds easily and the payloads, may be easily accessed with Flask. The request object, which is available within view functions, can be used to get request data.

Disadvantages: we leverage the Random Forest method, K-Nearest Neighbors (KNN), and AdaBoost algorithm to present a novel way to improve fraud detection in financial statements. Sometimes because of non-stop data collection the errors would be more as the traffic is high.

REFERENCES

[1] Albizri, D. Appelbaum, and N. Rizzotto, ‘‘Evaluation of financial statements fraud detection research: A multi-disciplinary analysis,’’ Int. J. Discl. Governance, vol. 16, no. 4, pp. 206–241, Dec. 2019.

[2] Albizri, D. Appelbaum, and N. Rizzotto, ‘‘Evaluation of financial statements fraud detection research: A multi-disciplinary analysis,’’ Int. J. Discl. Governance, vol. 16, no. 4, pp. 206–241, Dec. 2019.

[3] Yao J, Pan Y, Yang S, Chen Y, Li Y. Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach. Sustainability. 2019; 11(6):1579

[4] W. Xiuguo and D. Shengyong, "An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning," in IEEE Access, vol. 10, pp

[5] M. S. Beasley, ‘‘An empirical analysis of the relation between the board of director composition and financial statement fraud,’’ Accounting Rev., vol. 71, pp. 443–465, Oct. 1996.