International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 5, Sept-Oct-2024, ISSN (Online): 2395-566X

Fraud Detection in Financial Transactions using Machine Learning

Professor Syeeda, Abhisek Mohanty
School of Science and Computer Application
CMR University (Lakeside campus)Bengaluru, India

Abstract- Banking system vulnerabilities have made us vulnerable to fraudulent activities that seriously harm the bank's reputation and financial standing in addition to harming clients. An estimated large sum of money is lost financially each year as a result of financial fraud in banks. Early discovery aids in the mitigation of the fraud by allowing for the development of a countermeasure and the recovery of such losses. This research proposes a machine learning-based method to effectively aid in fraud detection. In order to combat counterfeits and minimise damage, the artificial intelligence (AI) based model will expedite the check verification process. In order to determine the association between specific parameters and fraudulence, we examined a number of clever algorithms that were trained on a public dataset in this article.

Index Terms- cybercrime,machine learning techniques.

I. INTRODUCTION

Fraudulent sales and sensors have long been intertwined. In the modern era of the Internet, fraudulent transactions are occurring more frequently than ever before and are the primary source of financial losses. The amount of thrift brought about by sale fraud was over $28 billion in 2019, $30 billion in 2020, and above $32 billion by 2021. The worldwide rate

It is projected that sales fraud would continue to climb, reaching $34 billion in 2022. Therefore, in order to celebrate and screen fiscal deals, banks and fiscal service providers might require an automated fraud finding tool. Fraud detection tools are designed to identify anomalous behaviour patterns in a vast amount of transactional data and then utilise those patterns to analyse and monitor new business opportunities (2, 3).

Machine literacy is an artificial intelligence (AI) function systems improve and on their literacy (1- 10).

In order to find patterns in the data and form better judgements later on based on the exemplifications we provide, the literacy process starts with compliances or data, which are comparable to exemplifications, firsthand experience, or instruction (11–20). The main goal is to enable computers to learn on their own without human assistance or support for acclimatise behaviour in the aftermath (21– 30).

Networks with the capacity for unsupervised learning are referred to as "deep literacy," a subclass of "machine literacy" in artificial intelligence.It lacks a name or shape (31–40). By utilising profile photos and identifying the variations between them, deep literacy is a style that encourages face discovery and determines if it is genuine or fraudulent (41– 50).

It has been shown that machine literacy is incredibly effective at identifying and classifying fraudulent transactions (51–60). In another way, agreat number of sale reports may be used to train and validate fraud classifier) 61- 70). Given the fact that it was overseen, Since literacy has proven very effective in identifying fraudulent transactions, transactional fraud analysis technology will continue to advance (71–76). Even a slight improvement in the classifier can save a business a significant amount of plutocrat.

II. LITERATURE REVIEW

The fraudulent deals have been the subject of numerous research. The study in (77) found that people all over the world were more likely to become accustomed to using cards than cash in their daily lives due to the rapid advancement of technology. The MasterCard emerged as a fairly functional tool for online commerce. This excessive increase in consumption results in significant harm as well as fraud instances. The fraud deals had a real influence on the financial situation, thus it was imperative to put an end to them. The anomaly discovery was playing a significant role in the fraud discovery process. This essay focuses mostly on determining if the sale was legitimate or fraudulent.

They demonstrated bidirectional gated intermittent units and bidirectional long short-term memory models. They also use machine learning techniques and deep literacy. However, their

© 2024 IJSRET

model outperforms the machine learning classifiers, with a score of 91.37.

The authors of(78) focused a great deal on the solution to the bracket imbalance issue, exploring the solution for fraud detection through the use of machine literacy algorithms. Additionally, they identify the best outcomes and weaknesses they obtain by using a dataset labelled with credit card fraud. They conclude that when the data are mostly imbalanced, the imbalanced bracket is ineffective. The writers of this paper establish that the being styles were high-end and provide multiple misleading warnings.

The DT, LR, and RF algorithms were used in the study (79) to gauge the effectiveness of the credit card fraud identification process. Oversampling was required in order to account for a somewhat unstable dataset. Following oversampling, 60 legitimate and 40 illegal transactions are made. The R programming language was utilised to carry out these methods. Measurements are also made of particularity, error rates, perceptivity, and LR's delicacy, which are 90.0, RF95.5, and DT94.3. The RF method works well with each of them.

An automatic bracket and a guiding bracket were employed in the study (80) to identify frauds, and a unique machine learning method was compared to identify the frauds. SVM, LR, and RF were applied. In order to advance a trouble scoring model, they learn about objectives. Every method was put to the test, and RF was completed correctly and as elegantly as possible. Additionally, this method operates well on a huge dataset and was a pleasure to work with. The outcome confirmed that type of algorithm works quite well in practice.

Nowdays financial fraud detection is crucial to lowering cybercrimes and is also important from a commercial perspective, a lot of literature on the subject is available. A few scholars have also reviewed the literature on papers that were written between the years 2000 and 2010.

Researchers usually use outlier detection approaches with highly imbalanced datasets to identify financial fraud (Jayakumar et al., 2013). Financial frauds of many kinds are also conceivable. According to one article (Jans et al., 2011), there are four types of financial fraud. The study primarily focuses on transaction fraud as it relates to mobile payments. Many methods have been tried and tested to identify financial fraud.

Phua et al. (2004) employed Decision Trees, and Neural Networks to identify motor insurance fraud.l the headings in the main body of your paper are numbered (automatically).In order to identify financial statement fraud in Chinese enterprises, Ravishanker et al. (2011) employed, and neural networks in another study.

Cost-sensitive Decision Trees (Sahin et al., 2013) and density-based clustering (Dharwa et al., 2011) have both been applied to credit card fraud.

Sorourenejad et al. (2016) address machine learning-based techniques that include clustering, SVM, HMM (Hidden Markov Models), ANNs (Artificial Neural Networks), and other techniques in both supervised and unsupervised learning scenarios.

III. RESEARCH METHODOLOGY

- We experimented with two supervised machine learning methods, Random Forest and Logistic Regression, to address the fraud detection issue.
- In an effort to rectify the dataset's class imbalance, we also experimented with undersampling.
- Cross-validation was used in the model development process to prevent overfitting and provide consistent performance.
- The models' performances were compared using performance metrics such as the Area Under the Curve (AUC) and Confusion Matrix.

Fig. 1.Steps followed to solve the fraud detection problem

Limitations of Study

- Under such circumstances, we ought to assess unsupervised methods that surpassed the purview of this investigation.
- we assessed the efficiency of employing particular supervised address transaction .The following are the methods' limitations that were used in this study:
- To train the algorithms, we used a pre-labeled dataset. Nevertheless, labelled data is typically hard to come by, thus using might practical. these situations, ought to assess unsupervised methods, which were outside the purview of this investigation.
- The study takes into account data from digital transactions, such as the the balance, and the transaction time. Other financial transactions, like credit card fraud,

© 2024 IJSRET

might not be affected by the same characteristics that aided in the detection of fraud.

- We assessed Random Forest and Logistic Regression as two machine learning algorithms. Even though the study's outcome using these algorithms is good, other approaches must be considered in order to decide which algorithm is optimal for this particular use case.
- Owing to the substantial amount of data, \computational capability. These methods could aid in enhancing the study's findings even more.

IV. METHODOLOGY

1. Dataset Sources

Kaggle

For the purpose of identifying fraudulent sales, we gathered the "fraudulent sale" dataset from Kaggle Depository. Ten characteristics and 6362620 records make up the dataset. The greatest sale reported in this data set amounts to 1991430 USD, while the average value of all trades is 144972 USD. Nevertheless, the distribution of the financial value of all trades is strongly skewed to the right, as you may be speculating based on the mean and outside at this point. Only a tiny percentage of transactions actually approach the outside, and the majority of deals are quite minor in size.

- The kind of transaction
- The amount transacted
- The recipient's and customer's IDs
- The recipient's and customer's old and new balances
- The transaction's time step
- Whether or not the transaction was fraudulent

.

1. Phases of the MethodologyProject Outcomes

Having a thorough understanding of the data set; reporting on its summary, each variable it contains, and any appropriate visualisations

Investigative Data Analysis Report on analysis carried out, important conclusions, and a thorough explanation of the data slices taken into consideration

Theory on the distinction between transactions that involve fraud and those that do notCharts and visualisations that highlight non-fraudulent transactions; the analysis's Python code .

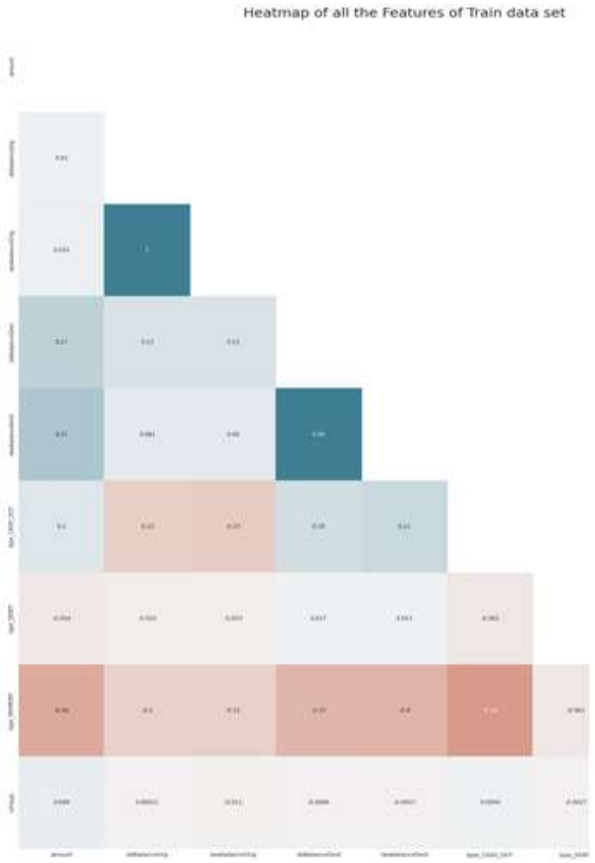
Modeling: • Summarise the outcomes of the many approaches that were tested; including data transformations, iterations, and the thorough modelling approach

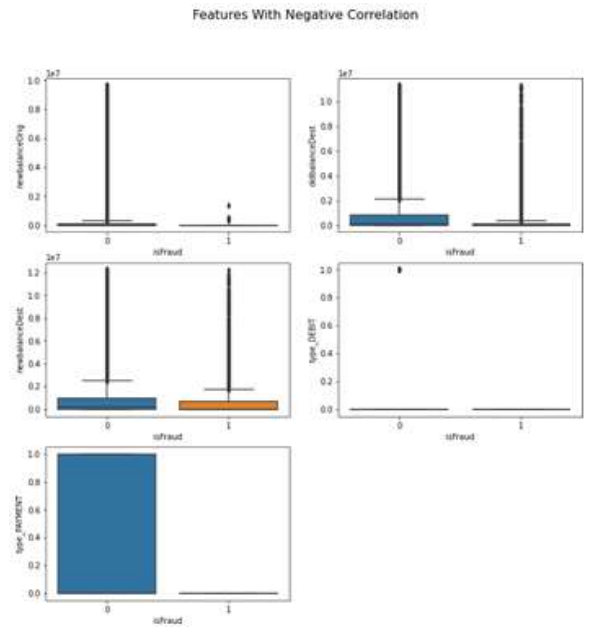
Final Project paper: This paper provides an overview of the work completed throughout the project, highlighting significant discoveries, contrasting various technique.

Instruments Employed

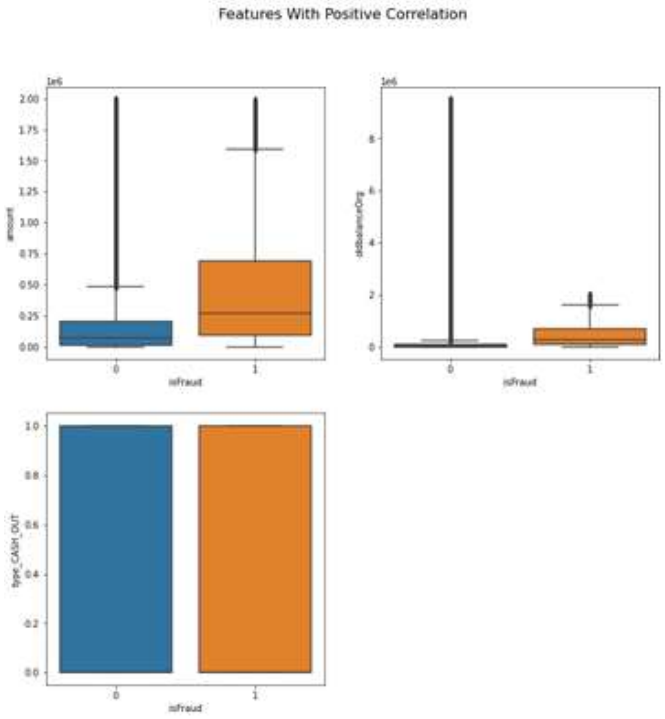
The following describes these libraries:

- Sklearn: utilised for tasks involving machine learning
- Seaborn: utilised to produce graphs and images
- Pandas is a tool for reading and modifying data.

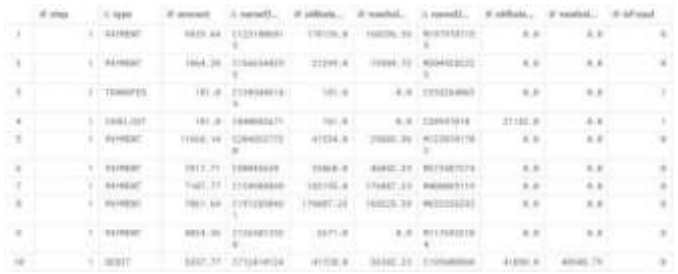




International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 5, Sept-Oct-2024, ISSN (Online): 2395-566X



V. SYSTEM ANALYSIS



In order to identify fraud, we examined the data on financial transactions. Predictive modelling, exploratory analysis, and data cleansing.

The data types were transformed, looked for missing values, and compiled the variables in the data as part of the data cleaning was examined in an exploratory analysis, delving deeply into each variable—, in particular. It was found some derived variables that may be useful in the detection of fraud. In order to better visualise the data and draw conclusions, we also generated a variety of graphs.

This work was explored with the Random Forest and Logistic Regression algorithms in predictive modelling. We found that Random Forest yields near-perfect precision and recall scores in this application. It was attempted to enhance the logistic regression outcomes by undersampling, but the outcomes remained unchanged due to a significant amount of discarded data. Using cross-validation, we made sure the models were not overfit.

It can be concluded that this labelled dataset successfully detects financial transaction fraud, and Random Forest is the most effective algorithm for this task.

V. CONCLUSION

In summary, we have effectively created our system for detecting fraudulent transactions in financial data. This framework will help in understanding the nuances of fraud detection, including choosing the best machine learning technique, correcting class imbalance, and creating derived variables that could facilitate class separation.We investigated Random Forest and Logistic Regression as two machine learning approaches. The Random Forest algorithm's significantly superior results than Logistic Regression's show that tree-based algorithms perform effectively on transaction data with highly-differentiated classes. In order to differentiate the classes more effectively than the raw data, this emphasises even more how crucial it is to conduct comprehensive exploratory analysis.

1. Suggestions

We demonstrated through this experiment that fraud can be detected in financial transaction data with a very high degree of accuracy, even in the presence of a high-class imbalance. We offer the following recommendations that come from this effort: • The most effective way to detect fraud in transaction data when the receiver and originator's balances and transaction amount are available is to use tree-based algorithms such as Random Forest.

- It's crucial to select the appropriate features when to visualise the distinction between fraudulent and non-fraudulent transactions.
- A number of sampling techniques, such as SMOTE, undersampling, and oversampling, can be used to lessen the high-class imbalance that is frequently seen in fraud detection scenarios. However, these techniques do have some computational limitations, especially when dealing with big data sets..
- Selecting the appropriate metric requires careful consideration in order to assess the effectiveness of fraud detection systems. Recall is a useful metric that indicates whether or not a significant proportion of fraudulent transactions are accurately classified. Accuracy alone shouldn't be our main guide since it might be deceptive.

REFERENCES

1. E. Ngai et al., Decision Support Systems 50, 2011, 559–569, The Application of Data Mining Techniques in

© 2024 IJSRET

Financial Fraud Detection: A Classification Framework and an Academic Review of Literature
2. Albashrawi et al., in Journal of Data Science 14(2016), 553-570, Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015.
3. Synthetic Financial Datasets for Fraud Detection, TESTIMON at NTNU, was obtained from https://www.kaggle.com/ntnu-testimon/paysim1.
4. A Novel Approach to Clustering Utilising Multivariate Outlier Identification by Jayakumar et al. Data Science Journal 11 (2013): 69–84
5. Jans et al., Expert Systems with Applications 2011; 38: 13351–13359, A Business Process Mining Application for Internal Transaction Fraud Mitigation
6. Phua et al., Classifying Skewed Data in Fraud Detection: A Minority Report. 2004; 6: 50–59 in ACM SIGKDD Explorations Newsletter.
7. Dharwa et al., A Based on Data Mining with Hybrid Approach, International Journal of Computer Applications, 2011; 16: 18–25.
8. A cost-sensitive decision tree method for fraud detection was developed by Shin et al. and published in Expert Systems with Applications in 2013; 40: 5916–5923.
9. Sorourenejad et al., A Survey of Methods for Detecting Credit Card Fraud: An Approach Focused on Data and Techniques, 2016.
10. Wedge et al., "Automated Feature Engineering Solves the False Positives Problem in Fraud Prediction," Machine Learning and Knowledge Discovery in Databases, pp. 372–388, 2018

© 2024 IJSRET