JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANOMALIES DETECTION IN CREDITCARD TRANSACTIONS

Mr. Suraj Kumar B.P[1], R. Sreehari[2], G. Bhanuteja[3], T. Tharun Kumar Reddy[4], R. Charan Kumar Reddy[5]
[1]*Assistant Professor, Department of Computer Science and Engineering*
[2,3,4,5]*Student, Department of Computer Science and Engineering*
[1,2,3,4,5]*Sir M Visvesvaraya Institute of Technology, Bangalore, India and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India*

Abstract— Credit card fraud is a problem that keeps getting worse in today's expanding financial landscape. Numerous organisations, enterprises, and governmental entities have suffered significant financial losses as a result of the sharp rise in fraud in recent years. Since the numbers are predicted to rise, several academics in this area have concentrated on employing cutting-edge machine-learning algorithms to spot fraudulent behavior early on. Both an algorithmic strategy using ensemble models like bagging and boosting and at data-level approach by using various resampling methods, including under-sampling, oversampling, and hybrid strategies, have been implemented to address this issue. So it is advantageous to have a system that can suggest the best options for credit card transactions. Machine learning algorithms can be utilized to design these systems, and frameworks like XGBoost, Logistic Regression, and Random Forest can be used to develop a prediction model. The predictive model then determines if the transaction is genuine or fraudulent using the resampled data.

I. INTRODUCTION

In the present day and age, fraudulent transactions are a significant consideration. The success of every transaction is crucial because there are numerous transactions taking place around the world. The amount that was exchanged determines the loss on the other side. Although it can appear insignificant, when all fraudulent transactions are combined, we can assess the significance of fraud detection for transactions. Therefore, in order to address this problem, we are developing a system that might advise fraudulent transactions by taking into account earlier transactions, increasing the likelihood that fraudulent transactions would be decreased. This essay focuses on an overview of various algorithms that can be applied to data analysis and prediction.

II. RELATED WORK

The study of machine learning, fraud detection, and a few machine learning techniques are covered in the following section. In order to build a system that can operate more effectively than the current system, this study is being conducted to assess the prior work that has been done in these areas.

One of the most significant problems is fraud in transactions, and preventing it is a crucial duty. The behaviour of Frauds is difficult to describe, and it is extremely tough to analyze the growing number of Fraud incidents. Each year, fraud involving credit card transactions causes enormous losses. With the development of machine learning techniques in recent years, it is advantageous to use these methods to avoid such problems and make decisions about impending events. Selecting a robust method with high accuracy is a severe challenge since customer and financial transaction behavior follows a high variation pattern. Numerous strategies are presented and used to accomplish this

For this, a model that uses Deep learning, one of the most potent techniques, and approaches decision-making similarly to humans was proposed in [1]. In this concept, the top-layer data is encoded in low-dimensional space using autoencoders, making the original data accessible upon decoding. Numerous algorithms can be used for fraud detection. Among all of those, [2] proposes an approach that involves choosing variables from the card transaction record and searching for variables with different temporal patterns between genuine and fraudulent transactions. This results in the creation of 2D and 3D subspaces. Random Forest Fraud Detection was another model suggested by [3]. This classifier uses several different decision trees. Fast training is available to address classification errors brought on by incredibly imbalanced data. Each tree is created using random information from a diverse sampling. Although a huge number of inputs are processed, this has the ability to handle high dimensional sets, which are suited for IEEE CIS data sets and can identify the most crucial properties. Some resampling strategies, such as under-sampling and over-sampling, are introduced in [4] since it is known that the count of fraudulent cases is quite small when compared to the right ones. It employs CNN (Condensed

Nearest Neighbour) for under-sampling and SMOTE (Synthetic Minority Oversampling Techniques) for over-sampling.

Simple methods can also be employed for detection, such as the classification and regression technique KNN [5]. This determines a point k-nearest neighbour without making any assumptions. It is simple to grasp but computationally intensive because the neighbours must be calculated for the full training set. One of the most popular ways for finding anomalies is this one. Here, they primarily concentrate on real-world credit card fraud detection.

[6] They primarily concentrated on online fraud detection, employing the Random Forest Algorithm (RFA) to identify fraudulent transactions and the amount of accuracy in it. As a result, they investigate Fraud Detection Solutions for Monetary Transactions Using Autoencoders [7]. Detection and prevention of fraud efforts are increasing in the current global economic context. Having an effective system for detecting financial transaction fraud could save millions of dollars from fraudulent activities. Transaction fraud is becoming increasingly common as online shopping becomes more popular. As a result, the research on fraud detection is an important one.

[8] Detection of Transaction Fraud Using Total Order Relationship and Behaviour Diversity. In this paper, they proposed a logical graph of Behaviour Profile (BP), which is a total order-based model for representing the logical relationship of attributes of transaction records, and they compute a path-based transaction probability from one attribute to another. Simultaneously, they define an information entropy-based diversity coefficient to characterize a user's transaction behavior diversity. In [9] L. Zheng and G. Liu improved Trad Boost and its applications in transaction fraud detection. AdaBoost is a boosting-based machine learning algorithm that assumes the training and testing sets have the same data distribution and input feature space. Since it updates the weight of a misclassified instance in a source domain based on the distribution distance between the instance and a target domain, and the distance computation is based on the idea of replicating Kernel Hilbert Space. In today's mobile payment age, credit card fraud detection is essential research.

[10] A different kind of loss function was suggested in this. We illustrate the detection performance of our model using Full Centre Loss (FCL), which takes angles and distances between features into account and can therefore comprehensively monitor deep representation learning. We compare FCL to other state-of-the-art loss functions. The finding suggests that FCL is a more reliable model and can outperform others. [11] Fuzzy clustering and neural networks are both used in the combined strategy that T. K. Behera and S. Panigrahi suggested. There are three phases in fuzzy clustering. The transactions are then sent on to the second step, where a fuzzy clustering technique is used to identify the pattern of credit card users based on their past transactions, after the first phase has been cleared. The transaction is categorised as suspicious, fraudulent, or lawful based on the pattern and a suspicious score that is calculated. A network-based algorithm is used to identify whether a transaction is suspect of being fraudulent or just a variance from a legitimate user.

[12] Abhinav and Amlan developed a Hidden Markov Model that can identify credit card theft without the need of counterfeit signatures. The hidden Markov Model is initially trained using the cardholder's normal behaviour, and if the fraud transaction is passed, the model detects the fraud using the trained data. This Model detects fraud transactions based on the amount spent, the location of the transaction, and the time of the transaction.Chee et al. employed twelve conventional models and hybrid approaches that make use of AdaBoost and majority voting techniques to effectively identify credit card fraud [13]. Added to the data to test the algorithm's noise's robustness. They also demonstrated that the additional noise had no effect on the majority voting process. With advancement of machine learning the recognition of transaction fraud is becoming more viable.

A transaction fraud detection method based on random forest and human detection was proposed by W. Deng and Z. Huang [14] and uses data mining to identify fraudulent transactions, which is a better model than deep network fraud detection. The information is organized into two tables: a transaction table and an identification table. Transactions are classified as either fraud (0 in the table) or not fraud (1 in the table). The transaction table and the identification table are combined to form a trained data set that is used to determine whether the transaction is legitimate or fraudulent. [15] T. Yan, Y. Li, and J published their findings on developing neural network-based fraud detection models. The LSTM and GRU models considerably outperformed basal ANN illustrating that the order of transactions in an account provides relevant information for discerning between fraudulent and genuine.

III. TABLE

Title	Author	Technique	Dataset	Accuracy for Best Technique
Using deep networks for fraud detection[1]	Z. Kazemi and H. Zarrabi	Autoencoders, Deep networks	German Credit Data	82%
Using deep networks for fraud detection [2]	A. Salzar, G. Safont and L. Vergara	Alpha Integration	evaluated information from a global financial company	76%
Using Random Forest for detecting transaction fraud[3]	D. Shaohu, G. Qiu, H. Mai, and H. Yu,	Random Forest, Logistic Regression	IEEE CIS fraud dataset	89%
Using Machine Learning for detecting real time fraud detection[4]	A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi	Linear Reggression, Logistic Regression, Naïve Bayas, SVM	The dataset was produced by combining the log files for all transactions and fraud transactions.	91%
Detecting Default Payment Fraud in Credit Cards[5]	S. S. H. Padmanabhuni, A. S. Kandukuri, D. Prusti, and S. K. Rath,	Decision Tree, Logistic regression, KNN, Adaboost	UCI machine learning repository dataset	82%
Managing Credit Card Fraud Risk by Autoencoders[6]	C. -H. Chang	Auto Encoder Model.	Both training and testing were conducted using a Synthetic Dataset.	83%
Using Random Forest Algorithm for detecting the fraud[7]	M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika and E. Aswin	Classification Technique, Random Forest Algorithm.	Credit Card Dataset(public Dataset).	90%
Credit Card Fraud Detection by Deep Representation Learning With Full Center Loss for [8]	Z. Li, G. Liu, and C. Jiang	Full Centre Loss (FCL) Function.	The first dataset is made public by Kaggle, and the second is a private transaction dataset from a Chinese financial company.	85%

TrAdaBoost Enhancement and Application to Transaction Fraud Detection[9]	L. Zheng, G. Liu, C. Yan, C. Jiang, M. Zhou and M. Li	AdaBoost Method.	Dataset from Newsgroups.	75%
Detecting Transaction Fraud Using Total Order Relationship and Behavior Diversity[10]	L. Zheng, G. Liu, C. Yan, and C. Jiang	Behaviour Profile(BP), logical graph of bp(LGBP).	Kaggle, Dalpozz Datasets.	87%
Using deep networks to identify fraud in credit card transactions[11]	T. K. Behera and S. Panigrahi	Genetic algorithm, Fuzzy clustering, and neural network	The datasets developed by Panigrahi	93%
Credit Card Fraud Detection Using Hidden Markov Model[12]	A. Srivastava, A. Kundu, S. Sural and A. Majumdar	Baum-Welch algorithm and k-mean.	To create a mixture of real and fraudulent transactions, a simulator is employed.	80%
Credit Card Fraud Detection Using AdaBoost and Majority Voting[13]	K. Randeera, C. P. Lim and A. K. Nandi	Random Forest, Support vector Machine, and Logistic Regression	A data set from a Turkish bank was used.	99%
A Data Mining Based System For Transaction Fraud Detection[14]	W. Deng, Z. Huang, J. Zhang, and J. Xu	Random Forest and XG Boost.	Data on fraud transactions is derived from an online platform's transaction logs.	70%
Comparison of Machine Learning and Neural Network Models on Fraud Detection[15]	T. Yan, Y. Li and J. He	Random Forest and Manual detection	Vesta offers THE IEEE CIS data set.	92%

IV. IMPLEMENTATION

Firstly, the dataset is extracted from the sources and undergone with some pre-processing techniques. The next step deals with the analysis of data which helps us to find some relations and to understand features. The resampling techniques are introduced to deal with the class imbalance problem the dataset constitutes of. The data on further is dealt with train test split which is dividing the resampled data into train and test data accordingly so as to train the model with respective algorithm. The algorithms used to train the model are Logistic Regression (LR), Random Forest and XGBoost which are used to predict the class of the test data based on the trained data. Later, the performance of each algorithm is evaluated based on the evaluation metrics such as Recall, Precision, F1-score. The tech stack used in the project contains python libraries like NumPy, Pandas and Matplotlib for dataset manipulation and visualization. Scikit-learn and Pickle are used for model implementation and serialization. Accuracy for different Machine Learning classification algorithms used for Anomalies Detection are detected.

V. METHODOLOGY

Methodology of the system that, predicts whether the transaction is legit or fraud is done in different phases such as: obtaining the relevant dataset, pre-processing the data, Data Analysis, Train Test Split, and Evaluation.

A. Acquiring the Datasets

The most important component of any ML-based application is gathering the dataset. Acquiring a dataset with sufficient data on which to create a precise anticipated model is crucial. The dataset we acquired contains 492 fraud and 284315 legit samples.

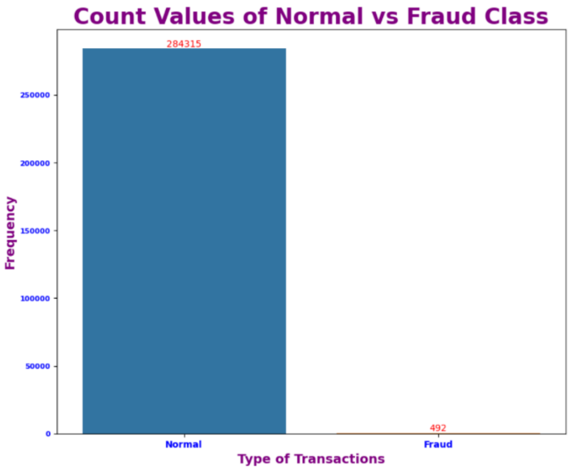
1) *Train dataset:* This covers the major part of the dataset that is used to train the learning ML model.

2) *Validate dataset*: The subset of data used to assess a model's fit to a training dataset while adjusting model hyperparameters.

3) *Test dataset*: This part of the dataset is to check the unbiased evaluation of accuracy in the model.

B. Pre-Processing the Data

Data pre-processing is the process of modifying the raw data to make it appropriate for the necessary ML application. Finding the data in the proper format is challenging since it often includes missing numbers, noise, outliers, and inconsistent data. These cannot be used to train ML models directly. Pre-processing is needed to organise the data into a structured manner and clean up the data by eliminating noise and missing information. It improves the model's precision, dependability, consistency, and effectiveness. The dataset we chosen is highly unbalanced and we use the under sampling, over sampling and SMOTE methods to convert the unbalanced dataset into balanced dataset.

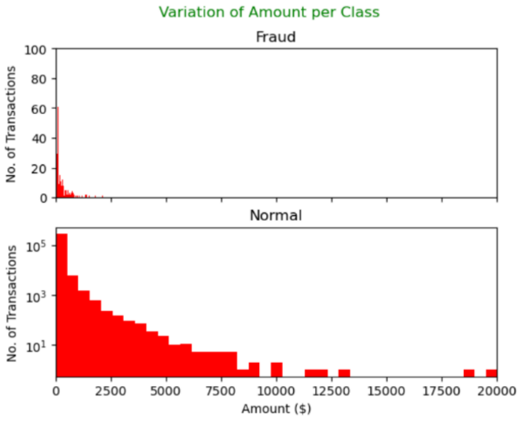


C. Data Analysis

Data analysis is the process of examining and interpreting data to uncover meaningful insights, patterns, and relationships. It involves using various statistical and computational methods to identify trends, outliers, and other key features of a dataset, and using this information to draw conclusions and make informed decisions.

Data analysis can be used in a wide variety of contexts, from scientific research and business analytics to marketing, finance, and healthcare. The goal of data analysis is to extract actionable insights from data, which can then be used to optimize processes, improve performance, and achieve better outcomes.

There are many different tools and techniques used in data analysis, including statistical methods, machine learning algorithms, data visualization tools, and more. Effective data analysis requires not only technical skills, but also a strong understanding of the underlying problem and the context in which the data is being used.



D. Train Test Split

Splitting a dataset into training and testing sets for use in a machine learning model. The dataset is represented by two arrays, X and Y which contain the input features and corresponding target labels, respectively. The oversampling, under sampling and SMOTE technique has been applied to the original dataset to balance the classes, which is often necessary when the classes are imbalanced.

The train_test_split () function from the scikit-learn library is used to randomly split the dataset into two sets - a training set and a testing set. The test size parameter is set to 0 to 1, which means that percentage of the data will be used for testing, while the remaining percentage will be used for training. The stratify parameter is set to Y, which ensures that the target class distribution is preserved in both the training and testing sets. This is important because it ensures that the model is trained and tested on a representative sample of the data, rather than one that is biased towards a particular class.

E. Evaluation

Evaluation is a crucial step in developing and fine-tuning machine learning algorithms, and accuracy score is one of the most commonly used evaluation metrics in supervised learning. Accuracy score measures the proportion of correctly classified instances in a dataset, and is calculated as the number of correct predictions divided by the total number of predictions. For example, if a model correctly predicts 90 out of 100 instances, its accuracy score would be 90%. In addition to accuracy score, there are many other evaluation metrics that can be used depending on the type of problem and the nature of the data. For example, precision, recall, F1 score are commonly used in classification problems, while mean squared error, R-squared, and coefficient of determination are commonly used in regression problems. In the proposed project the Evaluation metrics are found with all possible algorithms used.

IV. CONCLUSION

Data patterns can be efficiently found using machine learning algorithms. Machine learning has become a crucial component in resolving issues in a number of study fields as a result of the growth of big data. Big data technologies and machine learning algorithms collude to address a plethora of issues. Here, we have used algorithms like Logistic Regression, Random Forest, and XGBoost alongside data-level tactics including under sampling, oversampling, and hybrid techniques. In the performance tests harnessing the aforementioned methodologies, the Random Forest using the Synthetic Minority Over-sampling Technique (SMOTE) shows better performance than other models.

REFERENCES

1. Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017, pp. 0630-0633, doi: 10.1109/KBEI.2017.8324876.
2. A. Salazar, G. Safont and L. Vergara, "A New Method for Fraud Detection in Credit Cards Based on Transaction Dynamics in Subspaces," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 722-725, doi: 10.1109/CSCI49370.2019.00137
3. D. Shaohui, G. Qiu, H. Mai, and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 144-147, doi: 10.1109/ICCECE51280.2021.9342259.
4. A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 488-493, doi: 10.1109/CONFLUENCE.2019.8776942
5. S. S. H. Padmanabhuni, A. S. Kandukuri, D. Prusti, and S. K. Rath, "Detecting Default Payment Fraud in Credit Cards," 2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), 2019, pp. 15-153, doi: 10.1109/ICISGT44072.2019.00018.
6. C. -H. Chang, "Managing Credit Card Fraud Risk by Autoencoders: (ICPAI2020)," 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), 2020, pp. 118-122, doi: 10.1109/ICPAI51961.2020.00029.
7. M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), 2019, pp. 149-153, doi: 10.1109/ICCCT2.2019.8824930.
8. Z. Li, G. Liu, and C. Jiang, "Deep Representation Learning With Full Center Loss for Credit Card Fraud Detection," in IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 569-579, April 2020, doi: 10.1109/TCSS.2020.2970805.
9. L. Zheng, G. Liu, C. Yan, C. Jiang, M. Zhou and M. Li, "Improved TrAdaBoost and its Application to Transaction Fraud Detection," in IEEE Transactions on Computational Social Systems, vol. 7, no. 5, pp. 1304-1316, Oct. 2020, doi: 10.1109/TCSS.2020.3017013.
10. L. Zheng, G. Liu, C. Yan, and C. Jiang, "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity," in IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 796-806, Sept. 2018, doi: 10.1109/TCSS.2018.2856910.
11. T. K. Behera and S. Panigrahi, "Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network," 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015, pp. 494-499, doi: 10.1109/ICACCE.2015.33.
12. A. Srivastava, A. Kundu, S. Sural and A. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," in IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37-48, Jan.-March 2008, doi: 10.1109/TDSC.2007.70228.
13. K. Randeera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in IEEE Access, vol. 6, pp. 14277-14284, 2018, DOI: 10.1109/ACCESS.2018.2806420.
14. W. Deng, Z. Huang, J. Zhang, and J. Xu, "A Data Mining Based System For Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 542-545, doi: 10.1109/ICCECE51280.2021.9342376.
15. T. Yan, Y. Li and J. He, "Comparison of Machine Learning and Neural Network Models on Fraud Detection," 2021 IEEE International Conference on Artificial Intelligence and Comput