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

A STUDY ON ONLINE TRANSACTION FRAUD DETECTION USING MACHINE LEARNING ALGORITHMS

SHARANYA. N
Senior Software Engineer – Development
Hyderabad, India

ABSTRACT: This paper is about Online fraud detection is crucial for protecting people and businesses from harm. As technology advances and more people use the internet, we're getting better at catching fraud. Using Math, Statistics and Machine learning algorithms helps us to find and stop fraudulent transactions. By analyzing large amounts of data and patterns, these algorithms can detect unusual or suspicious activities that may indicate fraud. In this study, my aim is to investigate and evaluate the effectiveness of machine learning algorithms, particularly in the domain of online transaction fraud detection. We seek to understand how these algorithms can be utilized to accurately identify and prevent fraudulent activities in online transactions. By analyzing different models and techniques, we aim to identify the most effective approaches for detecting fraud and minimizing financial losses. Additionally, we aim to provide insights and recommendations for improving fraud detection systems.

KEYWORDS: Online transaction, Fraud detection, Machine Learning Algorithms

RESEARCH GAP:

The current study falls short in conducting a comprehensive evaluation of model performance using machine learning algorithms like Random Forest classifier, SVC, XGB classifier, and Logistic regression, which are pivotal for achieving faster and more accurate detection of fraudulent activities.

PRIMARY OBJECTIVE:

The main objectives of this research:

- To detect the fraud transactions from the real time dataset.
- Evaluate the performance of various machine learning ensemble methods, such as Logistic Regression, XGB Classifier, Support Vector Classifier, Random Forest Classifier, in real-time fraud detection scenarios.
- Investigating the effectiveness of ensemble methods in improving the accuracy and efficiency of fraud detection systems compared to traditional static detection methods.

INTRODUCTION: In this modern world, we can see the technology is increasing day by day. Few of our daily activities are also depending on the technology. One of the major activity people are using very frequently is making a payment using Online facility for shopping to purchase goods, money transactions from source to destination banks. The online transactions are done by using different methods like Credit card, Debit card or electronic platforms like UPI, Paytm etc.,

However, in this digital era, along with the advantages, there are some disadvantages also by using online transaction like there can be risk in terms of Security or Fraud. The frequent online transactions can attract the cybercriminals and it leads to the more risk.

This research aims to explore and evaluate using Machine Learning algorithms for online transaction fraud detection, with a focus on enhancing accuracy and performance. Using the Machine Learning algorithms the study seeks to address the evolving nature of online fraud and mitigate risks associated with electronic transactions, contributing to the development of more resilient and secure digital ecosystems.

REVIEW OF LITERATURE:

According to B. B. Sagar, Pratibha Singh, S. Mallika in the year 2016 in their study "**Online transaction fraud detection techniques: A review of data mining approaches**", They have highlighted the escalating challenges of fraud in e-commerce and online banking, necessitating advanced fraud detection techniques. While data mining is commonly used, its effectiveness depends on the quality of the financial dataset. The paper reviews existing fraud detection techniques, addresses dataset issues, and proposes a hybrid approach leveraging data mining algorithms at multiple stages to enhance fraud detection accuracy.

According to John Batani in the year 2017 in their study "**An Adaptive and Real-Time Fraud Detection Algorithm in Online Transactions**", He have addressed that the persistent challenge of credit card fraud in e-commerce transactions, proposing an adaptive algorithm for real-time fraud detection. Leveraging Artificial Neural Network, Hidden Markov Model, and One-Time Password, the algorithm achieves impressive fraud detection rates and accuracy. The solution offers promise as a plugin for e-commerce sites, aiming to restore and in still confidence in online transactions amidst security concerns globally.

According to Zanin, M., Romance, M., Moral, S. and Criado, R. in the year 2018 in their study "**Credit card fraud detection through parenclitic network analysis, Complexity**", The review delves into credit card fraud detection methodologies, specifically exploring the application of parenclitic network analysis within the realm of complexity theory. By examining the intricacies of fraud networks, the study aims to enhance understanding and improve detection accuracy. Through the lens of complexity theory, it seeks to uncover underlying patterns and dynamics inherent in fraudulent activities, offering insights for more robust detection mechanisms. The research underscores the importance of embracing interdisciplinary approaches to combat evolving fraud schemes effectively. Ultimately, it advocates for leveraging complexity theory to develop innovative strategies for mitigating credit card fraud risks.

According to I. Mettildha Mary, M. Priyadharsini, Karuppasamy. K, Margret Sharmila. F in the year 2021 in their study "**Online Transaction Fraud Detection System**", They have concluded that the growing prevalence of online credit/debit card transactions and the subsequent rise in fraud, highlighting the limitations of current detection methods. This study introduces a behavior-based approach using Support Vector Machines to enhance fraud detection accuracy, particularly focusing on predicting frauds based on transaction conduct changes. The proposed method addresses the challenges posed by the large volume of data associated with credit/debit card fraud detection, offering potential improvements in detection efficiency.

According to Rani. T.P, Suganthi. K, Magilan Saravanan, Ashish Kumar Sahu, K. Martin Sagayam, Ahmed A. Elngar in the year 2022 in their study "**Predicting Online Fraudulent Transactions Using Machine Learning**", The literature emphasizes the significant impact of fraudulent online transactions, exacerbated by technological advancements and global connectivity. Addressing the challenges of data scarcity, class imbalances, and confidentiality constraints, the study focuses on the pivotal role of machine learning in developing efficient fraud detection algorithms. Through the analysis of various machine learning models and the validation of fraud detection techniques, the research aims to provide insights and develop a prototype system capable of meeting real-world demands, thereby enhancing transaction security for customers.

According to Shayan Wangde, Raj Kheratkar, Zoheb Waghu, Prof. Suhas Lawand in the year 2022 in their study "**Online Transaction Fraud Detection System Using Machine Learning & E-Commerce**", The proposed system utilizes Behavior and Location Analysis (BLA) to prevent fraudulent transactions by restricting and blocking unauthorized credit card usage on a website. Through BLA, the system can detect suspicious activity during registration and transaction processes, leveraging user spending patterns and geographical location to verify identity and minimize false positives flagged by Fraud Detection Systems (FDS). This approach enhances security by proactively identifying and preventing potential fraud before transactions are authorized, thereby safeguarding genuine users' credit card details.

According to Mosa M. M. Megdad, Bassem S. Abu-Nasser and Samy S. Abu-Naser in the year 2022 in their study "**Fraudulent Financial Transactions Detection Using Machine Learning**", They have concluded that the comparisons between classifiers such as MLP, Random Forest, LGBM, and others reveal the Random Forest Classifier as the most effective with an accuracy of 99.97% on an unbalanced dataset, while the Bagging Classifier achieves superior performance on a balanced dataset with an accuracy of 99.96%. These findings underscore the importance of algorithm selection and dataset balancing in transaction risk detection to enhance customer experience and minimize financial loss.

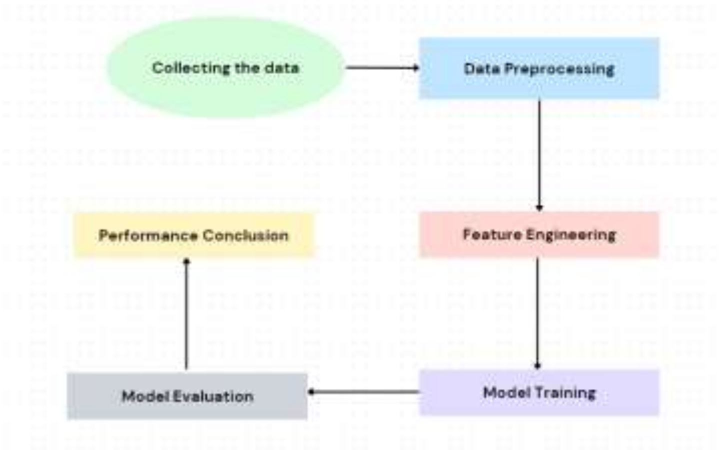
According to Taranjyot Singh Chawla in the year 2022 in their study "**Online Payment Fraud Detection using Machine Learning Techniques**", The literature emphasizes the widespread adoption of online payments, highlighting its convenience for consumers and profitability for businesses. However, this ease of use also brings forth the risk of fraud, necessitating vigilance from both consumers and businesses. The proposed model in the study aims to discern fraudulent transactions by considering factors such as payment type and recipient identity, underscoring the importance of awareness and proactive measures against internet scams.

According to Chenoori, Rathan Kumar; Kavuri, Radhika in the year 2022 in their study "**Online Transaction Fraud Detection Using Efficient Dimensionality Reduction and Machine Learning Techniques**", They have concluded that with the rapid increase in online transactions, there is a pressing need to detect fraudulent activities early to minimize losses. This study proposes machine learning models, including a statistical-based dimensionality reduction technique, and demonstrates the effectiveness of XGBoost model on the IEEE-CIS Fraud Detection dataset for accurately predicting and classifying fraudulent transactions.

According to Paolo Vanini, Sebastiano Rossi, Ermin Zvizdic & Thomas Domenig in the year 2023 in their study "**Online payment fraud: from anomaly detection to risk management**", They have concluded that the challenge of online banking fraud detection, emphasizing the need to detect fraudsters while minimizing false alarms and expected financial losses. By proposing three models - machine learning-based fraud detection, economic optimization of results, and a risk model considering countermeasures - the study demonstrates significant reductions in expected losses compared to static if-then rules. These findings highlight the viability and effectiveness of the proposed risk framework in combating online banking fraud while maintaining a low false positive rate.

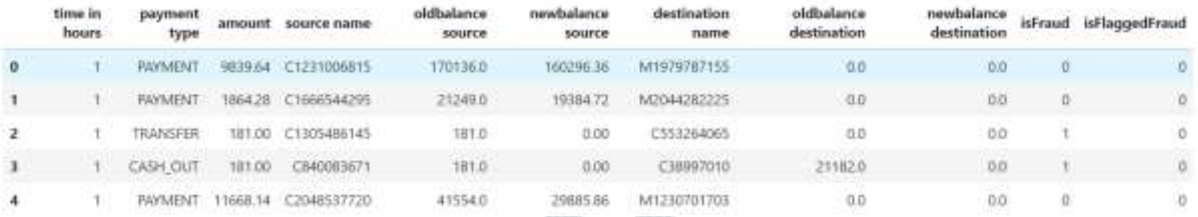
RESEARCH PROCEDURE FOLLOWED:

Source: Created using Canva software.



COLLECTING THE DATA:

In this study, Used the dataset from ***Kaggle platform online***. The data is of 210994 rows and 11 columns. The dataset is having 1442 rows of fraud data. The data is cleaned and pre-processed before applying the Machine Learning algorithms. The Data looks like below:



	time in hours	payment type	amount	source name	oldbalance source	newbalance source	destination name	oldbalance destination	newbalance destination	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

From the above table, there are 11 features.

Feature No	Feature Name	Feature Description
1	Time in hours	About the unit of time
2	Payment Type	Type of the transaction done
3	Amount	Total amount of the transaction
4	Source name	Transaction started origin
5	Oldbalance Source	Balance of the account of sender before transaction
6	Newbalance Source	Balance of the account of sender after transaction
7	Destination name	Transaction ended origin
8	Oldbalance Destination	Balance of the account of receiver before transaction
9	Newbalance Destination	Balance of the account of receiver after transaction
10	IsFraud	The value to be predicted i.e. 0 or 1
11	IsFlaggedFraud	The value to be predicted i.e. 0 or 1

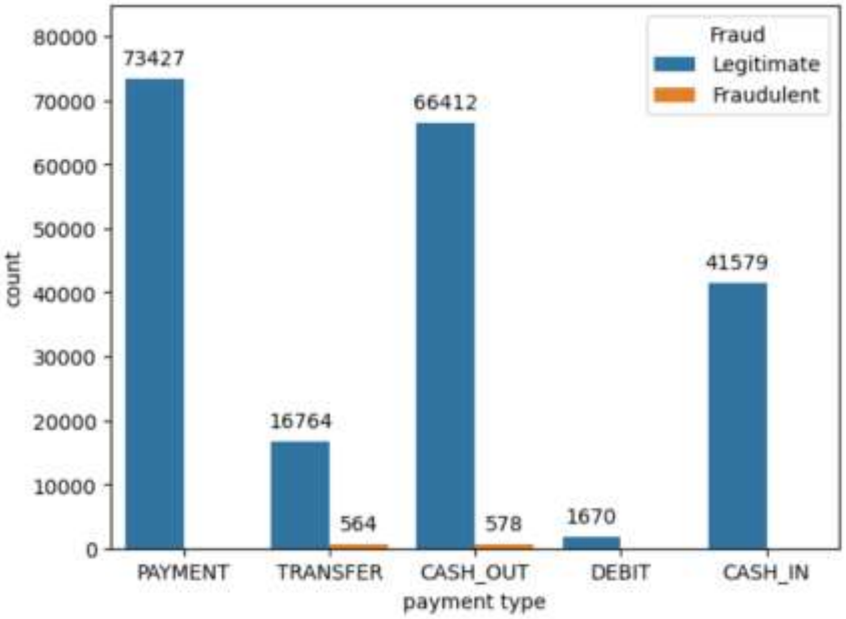
DATA PRE-PROCESSING:

The dataset has to be cleaned and normalized as it has spelling mistakes, duplicates, missing values and incorrect format of IsFraud column. The data is pre-processed by removing the unnecessary data which is not helpful for the model training and handled the missing values by using the statistical techniques such as Mean, Median and Mode. Finally corrected the spelling mistakes and converted the raw data into the valuable data to proceed further.

Now the dataset is Normalized and consists of 200994 rows and 11 columns. The dataset is having 1142 rows of fraud data.

From below graph, it is observed that the Fraud data is more by using the Transfer and Cash out Payment type options.

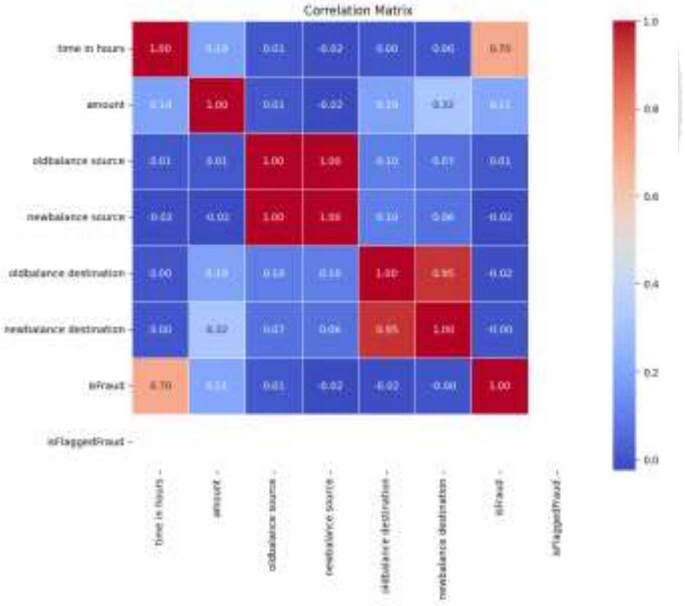
Source: Taken from the analysis using Python libraries Jupyter Notebook.



FEATURE SELECTION:

Here, Correlation Matrix analysis is used and observed the correlation matrix reveals a moderate positive correlation between the 'time in hours' and 'isFraud' features, suggesting a potential relationship between the time of the transaction and fraudulent activity. Additionally, there's a weak positive correlation between the 'amount' and 'isFraud' features, indicating that higher transaction amounts may slightly increase the likelihood of fraud.

Source: Taken from the analysis using Python libraries Jupyter Notebook.



MODEL TRAINING AND EVALUATION:

MACHINE LEARNING ALGORITHMS:

Logistic Regression: It is typically part of a predictive modelling that uses previous insights and observations to predict the probability of future events. Logistic regressions are also supervised algorithms that focus on binary classifications as outcomes, such as "1" or "0."

XGB Classifier: This classifier is an ensemble learning algorithm that utilizes a gradient boosting framework to produce highly accurate predictive models.

Support Vector Classifier (SVC): It is a supervised learning algorithm that separates data into classes by finding the hyperplane that maximizes the margin between them in a high-dimensional space.

Random Forest Classifier: Random Forest Classifier is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes output by individual trees.

Training Metrics

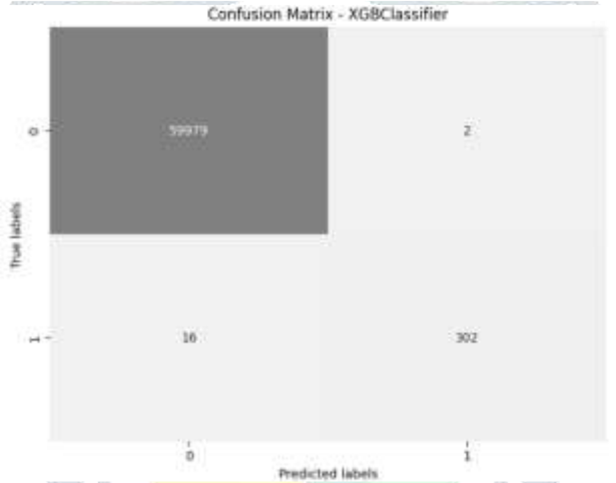
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	1.00	0.99	0.83	0.90
XGBClassifier	1.00	1.00	1.00	1.00
SVC	1.00	0.86	0.35	0.50
RandomForestClassifier	1.00	1.00	0.99	1.00

Validation Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	1.00	0.98	0.83	0.90
XGBClassifier	1.00	0.99	0.95	0.97
SVC	1.00	0.84	0.37	0.52
RandomForestClassifier	1.00	1.00	0.93	0.97

The best-performed model is **XGBClassifier.**

Source: Taken from the analysis using Python libraries Jupyter Notebook.



CONCLUSION:

After evaluating the models, it is evident that the XGBClassifier performed exceptionally well in detecting fraudulent online transactions based on the validation set. It demonstrated superior accuracy, precision, recall, and F1-score compared to alternative models. This indicates its effectiveness in accurately identifying fraudulent activities while minimizing errors. Overall, the XGBClassifier achieved the best balance between precision and recall, making it the top choice for predicting whether a transaction is legitimate or fraudulent. Therefore, for organizations seeking a reliable model for online fraud detection, the XGBClassifier emerges as the optimal choice, offering superior predictive capabilities and robust performance.

FURTHER RESEARCH:

In further research, one advanced technique that holds significant promise for online transaction fraud detection is Deep Learning, particularly in the form of Convolutional Neural Networks (CNNs). CNNs have shown remarkable success in various image recognition and natural language processing tasks, but their potential in fraud detection remains largely untapped. By leveraging CNNs, researchers can extract intricate patterns and features from transaction data, including transaction timestamps, amounts, and user behaviors. This advanced technique has the potential to enhance the accuracy and efficiency of fraud detection systems, leading to fewer false positives and false negatives. Implementing CNNs in online transaction fraud detection could result in improved security for online transactions, reduced financial losses for businesses and individuals, and increased trust and confidence in digital payment platforms.

REFERENCES:

1. B. B. Sagar, Pratibha Singh, S. Mallika on "**Online transaction fraud detection techniques: A review of data mining approaches**" 2016.
2. John Batani on "**An Adaptive and Real-Time Fraud Detection Algorithm in Online Transactions**" 2017.
3. Zanin, M., Romance, M., Moral, S. and Criado, R. **"*Credit card fraud detection through parenclitic network analysis, Complexity*"** 2018.
4. I. Mettildha Mary, M. Priyadharsini, Karuppasamy. K, Margret Sharmila. F on "**Online Transaction Fraud Detection System**" 2021.
5. Rani. T.P, Suganthi. K, Magilan Saravanan, Ashish Kumar Sahu, K. Martin Sagayam, Ahmed A. Elngar on ***"Predicting Online Fraudulent Transactions Using Machine Learning"*** 2022.
6. Shayan Wangde, Raj Kheratkar, Zoheb Waghu, Prof. Suhas Lawand **"*Online Transaction Fraud Detection System Using Machine Learning & E-Commerce*"** 2022.

7. Mosa M. M. Megdad, Bassem S. Abu-Nasser and Samy S. Abu-Naser "***Fraudulent Financial Transactions Detection Using Machine Learning***" 2022.
8. Chenoori, Rathan Kumar; Kavuri, Radhika on "**Online Transaction Fraud Detection Using Efficient Dimensionality Reduction and Machine Learning Techniques**" 2022.
9. Taranjyot Singh Chawla "***Online Payment Fraud Detection using Machine Learning Techniques***" 2022.
10. Paolo Vanini, Sebastiano Rossi, Ermin Zvizdic & Thomas Domenig on "**Online payment fraud: from anomaly detection to risk management**" 2023.

SOFTWARE:
Python, Jupyter Notebook

DECLARATION:
I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution are listed below:

Dataset used:
The dataset is used from the Kaggle platform online.

References:
Past study research papers are taken from Google scholar to study and understand the research gap.