Credit Card Fraud Detection Using Machine Learning

Shivani Varma[1] Dynaneshwari Badarkhe[2] Pratiksha Dongre[3] Prof Mrs. Priti Dhanke[4]
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Government College of Engineering Nagpur, India

Abstract— Credit cards have surpassed cash as the most widely used payment method in recent years. As technology progresses, the number of fraud cases rises, necessitating the creation of a fraud detection algorithm that can reliably detect and remove fraudulent behavior. The Isolation Forest Technique, a machine learning-based classification system, is used in this study to deal with the very imbalanced dataset. Finally, this study will calculate the accuracy, precision. Because of the rapid growth of the internet, online shopping has become a vital aspect of everyone's lifestyle. Most people use MasterCard to pay for things online. It's a simple matter of looking, and consumers will find what they're looking for on your visual display device or on their smart phone. Purchases made online.

Keywords: Accuracy, Fraud Detection, Credit Card, Machine Learning, Isolation Forest Algorithm

I. INTRODUCTION

The main goal of this study is to find out how to spot fraudulent credit card transactions. To do so, fraudulent and non-fraudulent transactions must be classified. The main goal is to develop a fraud detection algorithm that uses machine learning-based classification methods to uncover fraudulent transactions in a short amount of time with high accuracy. As technology advances, cash payments are becoming less common, while online payments are becoming more common, allowing fraudsters to carry out anonymous transactions.

Only the card number, expiration date, and CVV are necessary in various forms of online payment, and this information can be lost without our knowledge; in some circumstances, we are unaware that our information is being taken.

Machine learning is a common practise. Machine learning is not without its drawbacks. As a result, a highly unbalanced data set (Data set from Kaggle) is considered in order to determine the optimal algorithm to utilise in conjunction with its problems.

Even if the proposed method is good or not, it provides us an accuracy of around 99.9% because it is heavily imbalanced.

So, like in, the Isolation Forest Algorithm is utilised to accurately forecast fraudulent transactions of a credit card transaction dataset.

We do not know that our data has been leaked from online purchases where fraudsters use phishing techniques to grab the details. He only requires card details for some purchases to commit fraud, and the user may not be aware that his/her credit card information has been leaked. The details of the card should be kept confidential. But it is not always in our hands. Information may be leaked due to phishing websites, and the card itself may be lost or stolen. the best way to determine whether a transaction is fraudulent or not is to use existing data to determine the customer's spending pattern and machine learning to determine whether a is genuine or not.

Types of frauds:

- Online and offline
- Card theft
- Data phishing
- Application fraud
- Telecommunication fraud

The practice of discovering data points that deviate from the norm in some way is known as anomaly detection, also known as outlier detection. The topic has been carefully explored, and the solutions developed are employed in a variety of applications, including MasterCard fraud detection and e-commerce. Depending on the subject of investigation, the source of abnormalities varies. They'll emerge as a result of fraud attempts, detector weaknesses, or user blunders in a sophisticated input interface.

As a result, a thorough understanding of the 'black box' that exists between the mechanism that generates the data and the data itself is required. It could be extremely beneficial, if not vital, in spotting attention-getting outliers. This is one of the reasons why domain-specific research has become so popular in the field of anomaly detection. This thesis is an example of a domain-specific research project that focuses on acquiring information.

II. BACKGROUND AND RELATED WORK:

Fraud is a criminal offence in any form, and credit card fraud means stealing money. Many research have been conducted to determine if a transaction is fraudulent or not. Many issues remain, and efforts are still being made to address them.

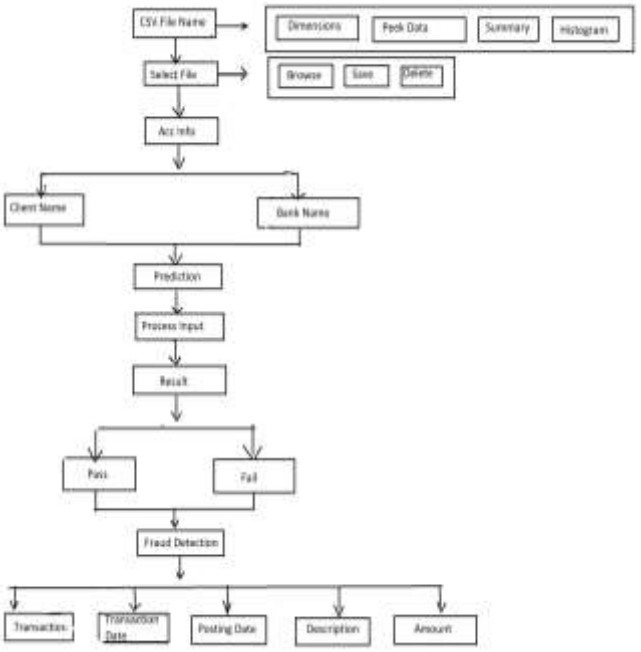
To begin, many people employed Data Mining Techniques to detect fraudulent transactions using a traditional methodology, which is not customary, and fraudsters these days are so clever that they can do it.

As demonstrated in [11], outlier detection and removal techniques are utilised to accurately forecast fraudulent transactions in a credit card transaction dataset. To cope with detecting unusual activity, outlier data is used. Even after a large number of proposed algorithms and mechanisms to stop fraudulent transactions, fraudsters are so cunning that they always try to come up with new ways to make anonymous transactions, and sometimes even the proposed algorithm is unable to determine whether the transaction is fraudulent or not. To prevent these scams, the proposed algorithm should be designed to learn from previous frauds and apply it to future frauds, potentially detecting fraud before it occurs.

III. PROPOSED METHODOLOGY:

In this suggested system, we will investigate various fraud detection strategies that are both user-friendly and secure. This system investigates the viability of using outlier mining to detect credit card fraud, applies outlier detection mining

based on distance sum to credit card fraud detection, and offers these detection processes and their empirical approach.



An overview of the workflow. (The workflow and steps of our implementation to achieve the goal are depicted in Figure 1.)

The dataset covers transactions of Canada Scotia Bank credit card holders for two days, and it includes the v1-v28 PCA [4] feature due to confidentiality concerns, as well as Time, Amount, and Class with 0 and 1, where 1 indicates fraud and 0 indicates non-fraud.

	V1	V2	V3	V4	V5
0	-1.359807	-0.072781	2.536347	1.378155	-0.338321

Table 1: Dataset

1	1.191857	0.266151	0.166480	0.448154	0.060018
2	-1.358354	-1.340163	1.773209	0.379780	-0.503198
3	-0.966272	-0.185226	1.792993	-0.863291	-0.010309
4	-1.158233	0.877737	1.548718	0.403034	-0.407193
5	-0.425966	0.960523	1.141109	-0.168252	0.420987

Table 1: shows a sample of the dataset as well as a few features v1-v5.

Data Distribution Visualization (As shown in fig. 2, it is extremely unbalanced.)

Better accuracy can be gained utilising the Outlier Data mining technique, as previously stated. Better results can be obtained now that the bias and extreme values have been detected and removed, therefore our objective now is to find and remove the extreme outliers. To visualise the correlation between variables, a heat map will be created first.

IV. Related Work:

This work is based on a study of Canada (Scotia Bank) credit card holders, in which data is normalised before the dataset is clustered. Machine learning techniques are used to generate reliable results after the data has been MLP educated. The use of normalised data yields promising outcomes. In this research, both supervised and unsupervised learning methods were applied. The goal of this study was to discover new fraud detection technologies that could improve the accuracy of results.

V. Dataset:

The files contain credit card transactions made by Canada(Scotia Bank) cardholders. This dataset shows transactions that took place in the last two days, with 492 frauds out of 284,807 total transactions. The dataset is severely skewed, with the positive category (frauds) accounting for only 0.17% of all transactions. It only has numerical input variables that have undergone a PCA transformation. Unfortunately, due to confidentiality concerns, we are unable to provide the first options and extensive background information on the information. The primary portions obtained with PCA are options V1, V2,... V28;

VI. Techniques:

A. Unsupervised Learningg:

Learning without supervision is a category of Machine Learning techniques for detecting patterns in data. The data fed into an unsupervised algorithmic programme isn't labelled, implying that only the input variables are provided with no output variables. In unsupervised learning, the algorithms are left to their own devices to discover interesting

patterns in the data. The clearly depicted in the diagram and referenced too.

B. Supervised Learning:

Learning under supervision in the algorithm tries to deduce information from prior samples. In supervised learning, on the other hand, the system tries to find patterns straight from the supplied instance. As a result, if the dataset is labelled, it falls under the supervised disadvantage.

C. Local Outlier Factor (LOF):

The Local Outlier Factor (LOF) rule is an unattended anomaly detection technique that computes a given information's native density deviation in relation to its neighbours. It considers samples that have a significantly lower density than their neighbours to be outliers. This example demonstrates how to utilise LOF for outlier detection, which is Scikit-default learns use case for this figure. There are no predictions, decision function, or score samples techniques once LOF is used for outlier identification.

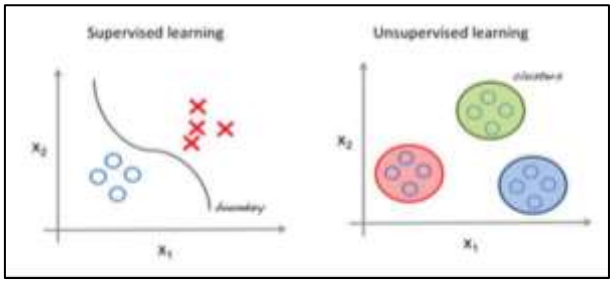
D. Isolation Forest:

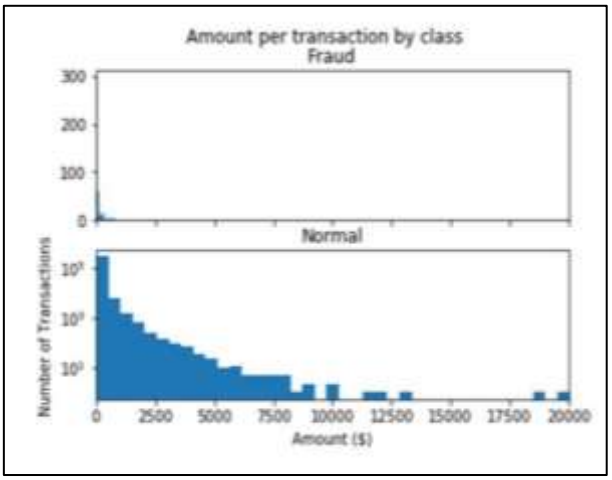
This technique's name is derived from its central concept. Each point in the data is isolated and classified as an outlier or an inlier by the algorithm. This technique has the advantage of working with a large data collection and multiple dimensions. The Advantages of Using Isolation Forests for Anomaly Detection Isolation Forests are one of the most recent ways for detecting anomalies. The approach is based on the concept that anomalies are rare and distinct data points.

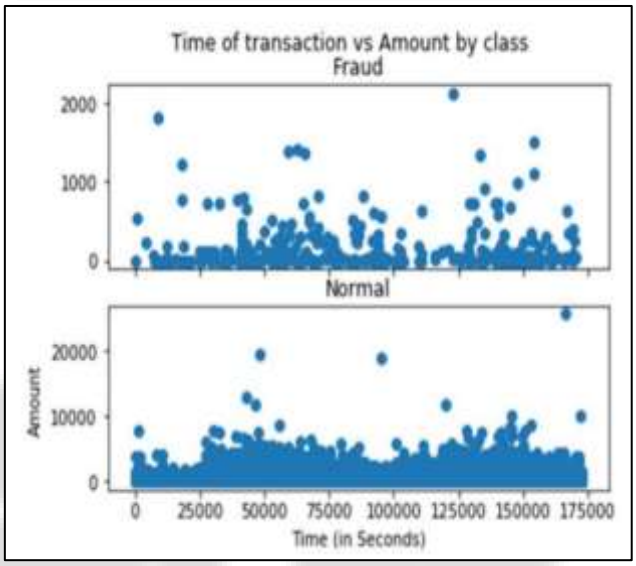
VII. OBSERVATIONS:

Isolation Forest discovered 73 errors, Local Outlier Factor discovered 97 errors,. Isolation Forest is 99.74 percent more accurate than LOF, which is 99.65 percent. When comparing error precision and recall for two models, we can observe that the Isolation Forest outperformed the LOF, with a fraud detection rate of roughly 27% vs a LOF detection rate of only 2%. As a result, the Isolation Forest technique fared significantly better in deciding the fraud instances, which were roughly 30 in total. We can improve this accuracy by increasing the sample size or using deep learning techniques, but at the cost of increased machine costs.

We can also utilise advanced anomaly detection methods to improve the accuracy of determining additional fraudulent cases.







VIII. CONCLUSIONS:

The usage of MasterCard for different purposes is become unavoidable. The increased use of credit cards will also lead to an increase in deceptive actions and a significant loss of money. In order to build cost-effective and accurate fraud detection tools, a great deal of analysis work has been done in this subject. As a result, a thorough assessment of previously planned algorithms is essential. The survey of current studies on credit card fraud detection is presented in this paper. The uniqueness of previous investigations is mentioned. The longer-term goal is to create an effective data processing algorithm for detecting MasterCard fraud.

ACKNOWLEDGMENT:

I'd like to use this occasion to express my gratitude to Prof. Priti Dhanke for providing the essential support and supervision in order to complete the research project. Her willingness and passion in encouraging us to develop our technical and creative ideas, which ultimately led to the project's success, is also something I admire. We would like to express our gratitude to the academic members of the Computer Science and Engineering Department for their unwavering support and willingness to help us with whatever we needed for our project.

REFERENCES:

[1] https://www.researchgate.net/publication/336800562_Credit_Card_Fraud_Detection_using_Machine_Learning_and_Data_Science
[2] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8
[3] https://ieeexplore.ieee.org/document/8717766
[4] https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc
[5] https://asianssr.org/index.php/ajct/article/view/1151
[6] https://ieeexplore.ieee.org/document/8123782
[7] https://www.researchgate.net/profile/Karl-Tuyls/publication/254198382_Machine_Learning_Techniques_for_Fraud_Detection/links/555f695508ae6f4dcc926e88/Machine-Learning-Techniques-for-Fraud-Detection.pdf
[8] https://ieeexplore.ieee.org/document/9121114

