e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING & PYTHON

Samprit Ghosh*1, Koushik Nath*2

*1,2AICTE, Electronics & Communication Engineering, Techno International Newtown, Kolkata, West Bengal, India.

DOI : https://www.doi.org/10.56726/IRJMETS42136

ABSTRACT

In recent years, credit cards have taken on important roles in people's lives. The danger of fraud has increased as a result of the unexpected boom in e-commerce and the widespread usage of credit cards for online purchases. Instead, than carrying around a lot of cash, it is easier to keep credit cards on hand. But that's also dangerous now. A serious problem that is now on the increase dramatically is credit card fraud .Currently, new researchers are having a lot of trouble detecting credit card fraud. We put in place a clever algorithm that can spot any type of fraud in a credit card transaction. By identifying a trend for each consumer in between fraudulent and legal transactions, we were able to solve the issue. The pattern of each customer's transactions is predicted using the isolation forest algorithm and the local outlier factor, and decisions are then made in light of that information. All attributes are tagged similarly to avoid data mismatches.

Keywords: Automated Fraud Detection, Isolation Forest Method, Local Outlier Factor, Applications Of Machine Learning, And Data Science.

I. INTRODUCTION

As we can see, there has been a substantial rise in online payments over the past several years, and for the majority of them, credit cards are the preferred payment option. For marketing companies, credit card fraud is a serious barrier. A range of actions, such as paying taxes on another account, applying for loans using false information, and more, can be used to perpetrate fraudulent fraud .Therefore, we require an efficient fraudulent detection model in order to decrease fraudulent behavior and their losses .Numerous new tools and algorithms are available that aid in the detection of various credit card fraud activities. A fundamental knowledge of these algorithms will enable us to develop a powerful model for detecting credit card fraud. By suggesting machine learning techniques, this article assists us in identifying questionable credit card transactions. The goal of machine learning's credit card fraud detection is to lessen this kind of fraudulent conduct .This kind of scam has existed in the past, but there hasn't been any investigation into it up to this point. Bankruptcy fraud, behavioral fraud, counterfeit fraud, and application fraud are some examples of credit fraud in transactions .There have been meta-learning-based research done in the past on credit card fraud. Meta-learning has a specific upper bound. The following two aspects are described in our report: False alarm and True Positive. Because it takes little time to spot fraudulent behavior, both of these traits are crucial in capturing fraudsters. We require a stronger classifier in order for the model to perform better. Meta-learning may be utilized to integrate many classifiers .

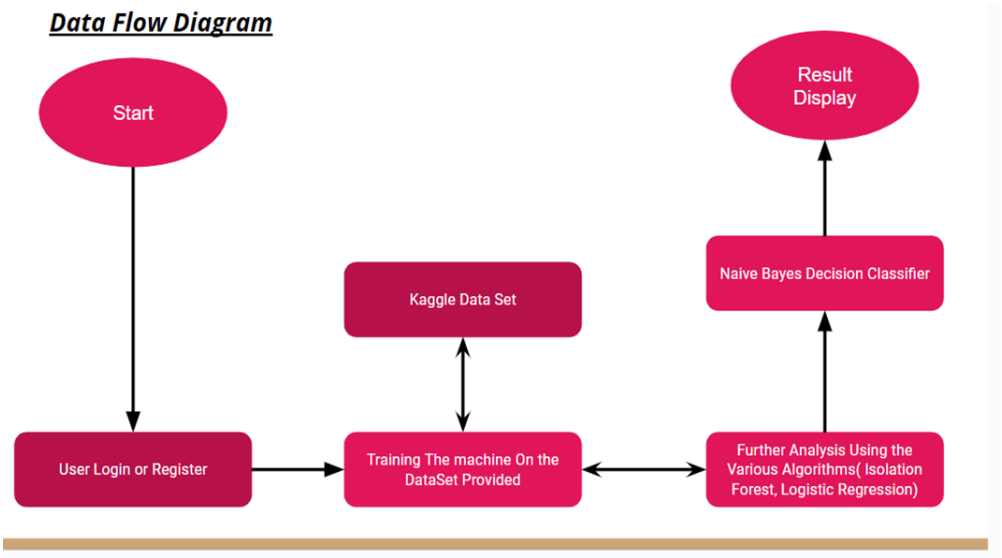
II. METHODOLOGY

Credit card use has increased along with the use of online payments. The ability to pay using a credit card is offered by many businesses. We may use our credit card to buy a variety of goods. By utilizing someone else's credit card and using their personal information to issue a credit card, people started committing fraud in this industry. In the case of online payments, electronic data can be exchanged to carry out fraudulent transactions. Credit card billing cannot prevent credit card fraud; nevertheless, fraud prevention is also necessary. If we look at the success stories of all the current systems, they are not very effective at catching fraudsters. Therefore, it is imperative to create a system that can identify fraud from the outset, assist users in reducing fraud in all of their online transactions, and alert users at the outset when their credentials are being used by someone else. This will enable him to stop this form of fraud activity before it starts and to consider how to reduce their losses. There should be restrictions on the credit card that prevent us from making purchases worth more than this sum in a single day or at a time. Losses will be lessened as a result of this. We have two analyzers that will

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

identify whether a transaction is legitimate or fraudulent: a random forest algorithm and a local factor outlier. These will aid in the projection of the score, which will indicate a more balanced outcome.

The accuracy is then decided on the basis of the outcome, which is the average between legitimate and fraudulent transactions. Given its significance, the dataset needs to be more balanced. Our dataset must now be divided into a train dataset and a test dataset. These two things are necessary for creating a problem model. Therefore, our objective is to use this dataset to train the classifier and test it .Using this approach, all methods for detecting credit card fraud are assessed and contrasted .Accuracy is defined as a portion of all exchange quantities that are clearly distinguishable. The method this study suggests makes advantage of the most recent machine learning techniques to identify outliers, or unusual activities. The following image can be used to illustrate the fundamental rough architecture diagram :



Observation

There are 492 frauds in the data set out of roughly 300,000. This yields a likelihood of fake instances of 17.2%. This revealed that there are a lot more clients that commit fraud. The data set is made up of columns that go from v1 to v28. From V1 to V28, there are several features. Furthermore, datasets do not contain any missing values. The column in the dataset is called Time & Amount. These two columns' ranges serve as the foundation for the analysis .The numerical value that may be referred to as a PCA transformation is included in the datasets. Unfortunately, we are unable to retain the original features and data information due to security concerns. The primary components are selected from columns V1 through V28. "Time" and "Amount" are the characteristics that PCA does not change .The concept of "time" is significant in this context since it is used to specify the interval between each transaction and is measured in seconds .Another element that is needed to calculate the transaction Amount is "Amount" .The most crucial component of our model, "Class," a response variable that accepts values between 1 and 0, is included here. It assigns a value of 1 for fraudulent transactions and a value of 0 for lawful ones. This model's primary objective is to forecast credit card fraud for all transactions that are accepted as online payments in order to determine whether or not the transaction is legitimate. If a transaction is legitimate, it is seen as being lawful, but a fraudulent transaction should be recognized as such .The output of this method, when it has successfully completed its task, is then shown as a graph and heat map.

Isolation Forest Algorithm

The Isolation Random Forest technique aids in the detection of anomalies. This technique reveals that anomalies are separate and sparse data points. Results from these qualities indicate that the isolation mechanism suspects abnormalities. We learned that our approach is unique from all previous ways and more accurate as a result of the factors mentioned above. Instead of the previous approach, this one incorporates an isolation algorithm, which is a strategy for anomaly identification that is more effective. Additionally, this

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

approach uses extremely little memory and has a very low time complexity. When compared to the datasets, the binary trees we create are modest.

The Isolation Random Forest technique chooses a feature at random from datasets, then finds a split value at random from the minimum and maximum value. Applying logic, there aren't many instances when anomalous observations and regular observations vary from one another. In order to isolate normal observations, we need more conditions. Score is calculated using the criteria necessary to distinguish between normal and anomalous observations. The binary decision tree with child nodes that are 0 and 1 is created using the score. Finally, if 0 is obtained, fraud is not present, and if 1 is obtained, fraud is present .By picking a feature at random, followed by a split value between the maximum and minimum values of that feature, the Isolation-Forest 'isolates' observations .The number of splitting necessary to isolate a sample is equal to the length of the route from the root node to the terminating node since recursive partitioning may be represented as a tree structure .This path length provides a gauge of our decision function and normalcy when averaged over a forest of similar random trees .For anomalies, random partitioning results in considerably shorter pathways. As a result, shorter path lengths for specific samples produced by a forest of random trees are quite likely to be outliers.

Local Outlier Factor (LOF)

The Local Outlier Factor (LOF) is an outlier algorithm that offers a way to determine how far a given data point strays from its neighbors. It comprises of outlier samples that are less dense than their neighbors. The higher and minimum values in the cluster of datasets that are distinct from its neighbors are used to choose the outlier value. If the outlier value differs from its neighbors, the system would have detected it and deemed it false. Score is calculated using the criteria necessary to distinguish between normal and anomalous observations. The binary decision tree with child nodes that are 0 and 1 is created using the score. Finally, if 0 is obtained, fraud is not present, and if 1 is obtained, fraud is present. Therefore, the local outlier factor assists us in identifying fake data that does not match well with its surroundings. It also aids in determining the departure of anomalous data from the standard deviation observed in all of the surrounding data.

Anomaly Detection

The Local Outlier Factor (LOF) is an outlier algorithm that offers a way to determine how far a given data point strays from its neighbors. It comprises of outlier samples that are less dense than their neighbors. The higher and minimum values in the cluster of datasets that are distinct from its neighbors are used to choose the outlier value. If the outlier value differs from its neighbors, the system would have detected it and deemed it false. Score is calculated using the criteria necessary to distinguish between normal and anomalous observations. The binary decision tree with child nodes that are 0 and 1 is created using the score. Finally, if 0 is obtained, fraud is not present, and if 1 is obtained, fraud is present. Therefore, the local outlier factor assists us in identifying fake data that does not match well with its surroundings. It also aids in determining the departure of anomalous data from the standard deviation observed in all of the surrounding data.

Anomaly Detection

The Local Outlier Factor (LOF) technique may be used in Python to find anomalies and detect fraud. The foundation of the unsupervised machine learning algorithm LOF is the idea that anomalies should be distinguished from the majority of the data. Each data point receives a score from LOF, which compares its local density to that of its neighbors. Normal data points are given greater values, whereas anomalies are given lower ratings, suggesting that they are in low density areas. In order to find the anomalies in our dataset, we need to select a threshold for the LOF scores. LOF is an effective method for finding abnormalities and may be applied in a variety of sectors, including banking and healthcare. When trying to detect fraud, it might be very helpful to look for outliers that match to fraudulent activity. Using LOF, we can rapidly spot dubious transactions or occurrences and take the necessary precautions to stop more fraud.

Density Based Detection of Anomaly

It uses the KNN algorithm as the basis for its working method .Assumption: Relevant data clusters at a central location, whereas irregular data are positioned farther out. depending on a density score, which may be obtained using Euclidian distance or other suitable methods depending on the data, the data points are grouped at a close proximity. Two criteria are used to classify things: nearest neighbor K: The core clustering technique

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

used in this method depends on how far apart each data point is measured, which defines how similar or clustered each piece of information is .Information's relative thickness is sometimes referred to as the Least Outlier Fraction (LOF).The separation metric is used as the foundation for the calculation.

Clustering Based Detection of Anomaly

Clustering is a remarkable technique renowned for its robustness and optimization . It is frequently utilized in unsupervised learning because of this .Assumption: Similar data points tend to cluster around particular spots. Each cluster's relative distance is determined by its shortest distance from the space's centroid .K means is a frequently used metric for classifying data. It uses the k-means method to group together sets of closely related data into clusters.

Procedure Followed During Implementation

Clustering is a remarkable technique renowned for its robustness and optimization . It is frequently utilized in unsupervised learning because of this .Assumption: Similar data points tend to cluster around particular spots. Each cluster's relative distance is determined by its shortest distance from the space's centroid .K means is a frequently used metric for classifying data. It uses the k-means method to group together sets of closely related data into clusters.

Procedure Followed During Implementation

Clustering is a remarkable technique renowned for its robustness and optimization . It is frequently utilized in unsupervised learning because of this .Assumption: Similar data points tend to cluster around particular spots. Each cluster's relative distance is determined by its shortest distance from the space's centroid .K means is a frequently used metric for classifying data. It uses the k-means method to group together sets of closely related data into clusters.

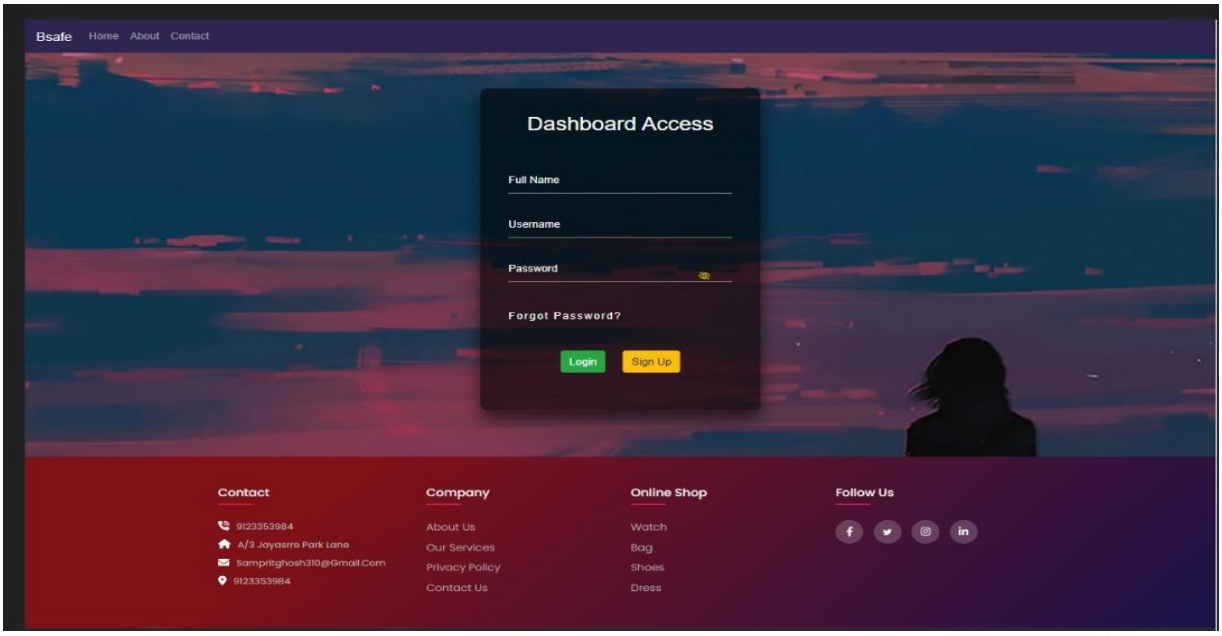
1. Download Anaconda from a trustworthy source.
2. Import the pandas, Scipy, Matplotlib, and Seaborn packages.
3. Open the dataset (.CSV file), which is a collection of data used for analytical or critical purposes.
4. Reconnoiter and navigate the dataset using the data. data and form. Describe.
5. Distinguish the training dataset from the testing dataset.
6. Use the dataset's histogram to visualize and summaries the numerical data.
7. Check if the class is 0 or 1 to ascertain the number of fraud cases.
8. Get the correlation matrix using the same approach.
9. The local outlier factor has to be identified next.
10. To get reliable findings, a random forest technique is then used.
11. The PyQt library is used to create the GUI.
12. In the file, define the constructor.
13. Write down the whole implementation within, enclosing everything in a Python file that supports a GUI.

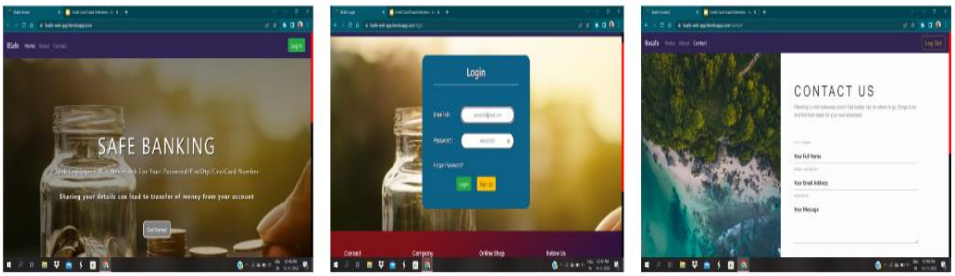
III. MODELING AND ANALYSIS

It is difficult to put this idea into effect since it requires the cooperation of banks, which are wary of sharing information owing to market competitiveness, legal reasons, and the protection of their users' data. We looked for several reference works that employed similar methods in order to gather information. One of these reference papers states: According to one of these reference publications, a sizable application data set provided by a German bank in 2006 was exposed to this method. The Amount feature's IQR analysis reveals that we lose some data from dishonest users. But it is not advised to delete it because the data for such people is substantially fewer. Instead, we would exclude any data points that are over 8000 in order to eliminate extreme outliers while also ensuring that none of our false user data was lost. During the late night and early morning hours, fewer transactions are made by legitimate users. Since most people are sleeping at this time, it also makes sense. Contrarily, the number of fraudulent transactions experiences large increases in the late hours and is much lower during the day.Let's go on to the Amount now. We can tell the feature is strongly right-skewed and has a high number of ou by viewing the distribution chart and boxplot. The credit card fraud

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

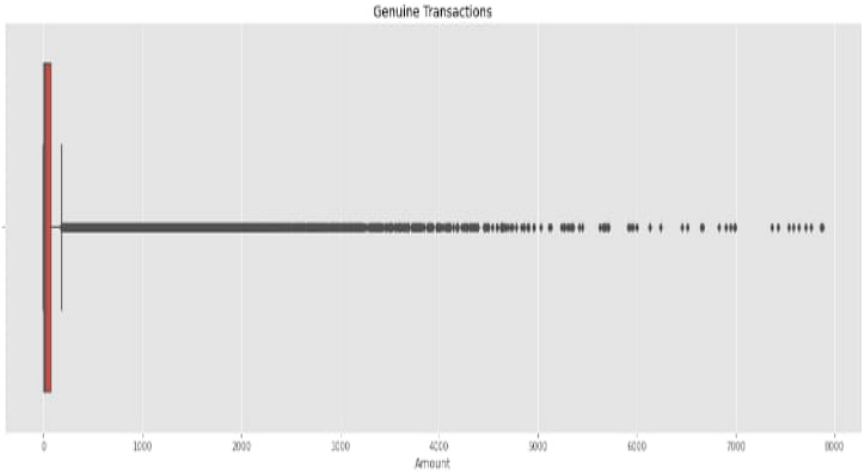
detection model may be misled by this. The Amount feature's IQR analysis reveals that we lose some data from dishonest users. However, since the data for these people is substantially fewer, it is not advised to delete it.





name	email	message
SAMPRIT GHOSH	SAMPRITGHOSH310@GMAIL.COM	Great job vai , keep it up
SAMPRIT GHOSH	SAMPRITGHOSH310@GMAIL.COM	Great job vai , keep it up
Mousumi Sengupta	mousumisengupta634@gmail.com	Hello Samprit
Mousumi Sengupta	mousumisengupta634@gmail.com	Hello Samprit
Antarik Pandit	antarikpanditison@gmail.com	Love this website
Antarik Pandit	antarikpanditison@gmail.com	Love this website
Sagnik Srimani	xabcd9537@gmail.com	Great job vai , keep it up

Working of Our Application



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

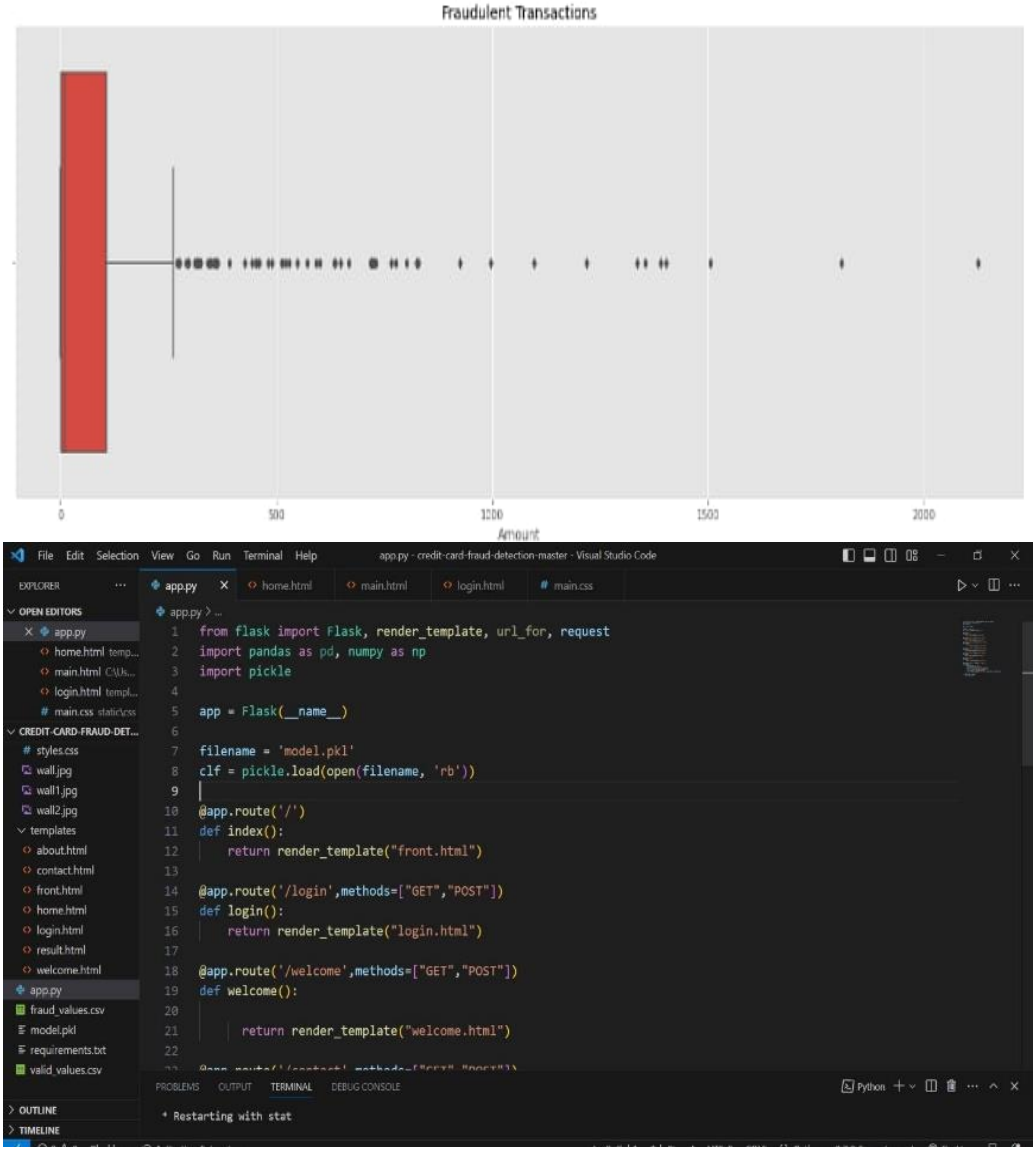


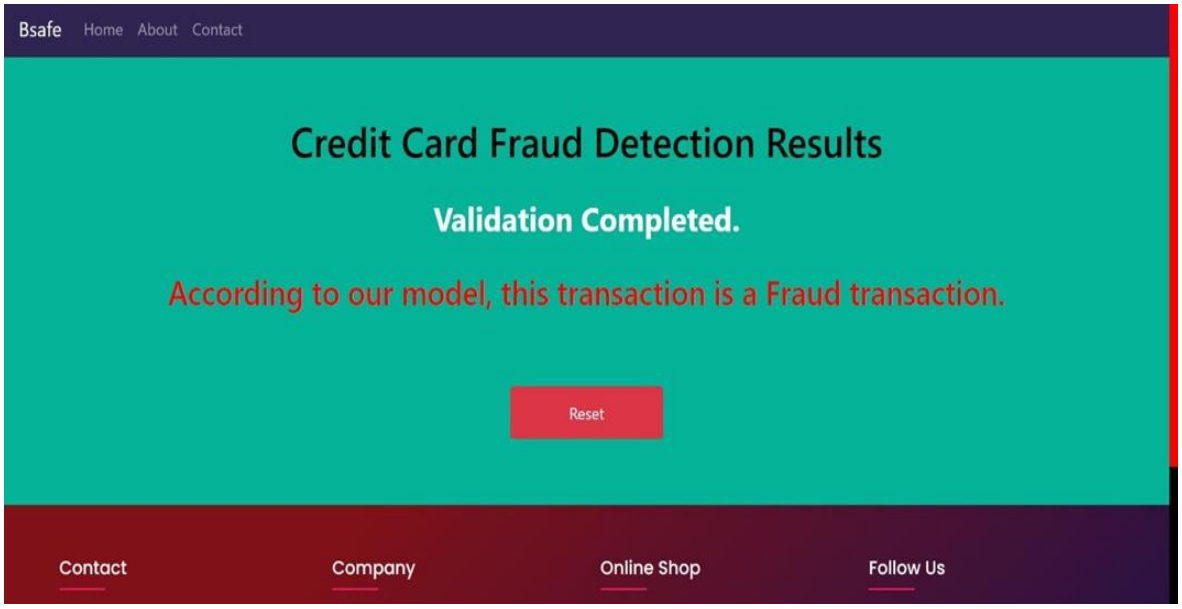
Figure: view of building.

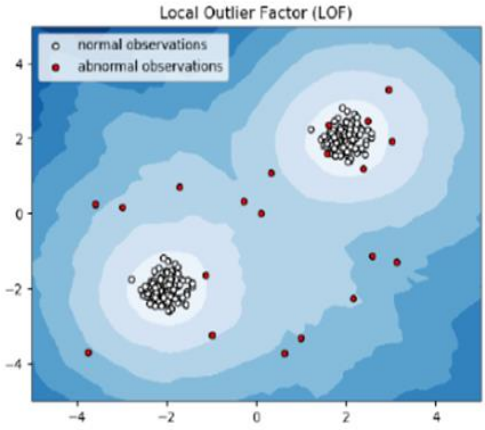
IV. RESULTS AND DISCUSSION

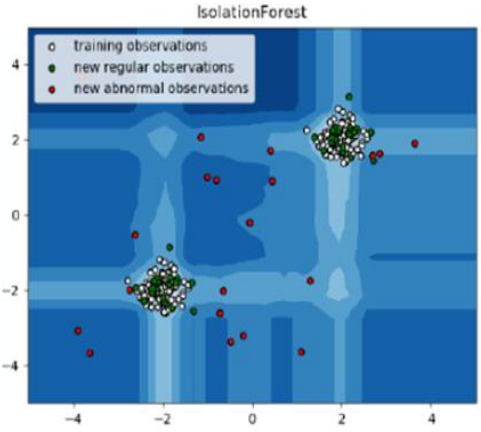
The code displays the results after comparing the true numbers to the amount of false positives it discovered. This is used to evaluate the algorithms' precision and accuracy. 10% of the entire dataset was made up of the dataset component we selected for quicker testing .The subset of the dataset we utilized for expedited testing was 10% of the overall dataset. At the end, which also takes use of the complete dataset, both outcomes are presented. Class 0 indicates that the transaction was determined to be genuine, while class 1 indicates that the transaction was determined to be fraudulent in the output shown below. These results are also provided together with the categorization report for each technique. This result was compared to the class values in order to exclude any potential false positives. It's crucial to remember that credit card fraud detection systems frequently combine these methods and regularly upgrade their algorithms to keep up with new fraud schemes. In order to efficiently identify and look into fraud instances, financial institutions also rely on consumer reports, manual checks, and cooperation with law enforcement organizations. Another algorithm frequently used with Isolation Forest for fraud detection is Local Outlier Factor (LOF). A data point's local density variation from its neighbors is determined by LOF. A data point is regarded as an outlier if its density is considerably lower than that of its neighbors .These algorithms allow for the prediction of the likelihood that a credit card transaction will be fraudulent. The algorithms may assign an anomaly score or a fraud chance to each transaction by

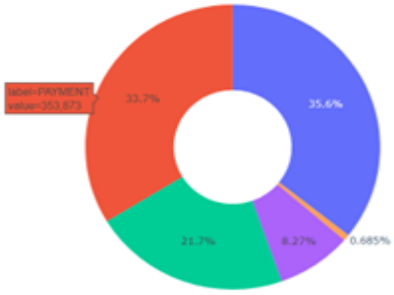
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

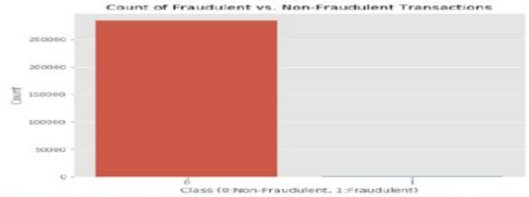
looking at many aspects of it, including the transaction amount, location, time, and other pertinent information.



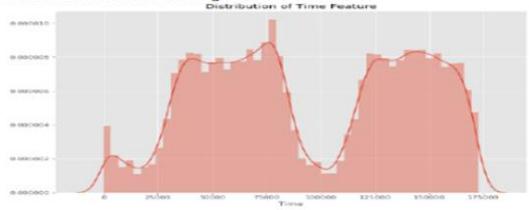








This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.



Global Link : https://bsafe.onrender.com

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

Test Case (sample split)	Assumption	Description	Expected Output	Actual Output		Log Message
				Isolation Forest Algorithm- Algorithm I Accuracy(%)	Local Outlier Factor - Algorithm II Accuracy(%)	
10:90	Algorithm -I will perform better	Check for accuracy at 10% training of data	99.70505	99.75071	99.65942	Success
15:85	Algorithm -II will perform better	Check for accuracy at 15% training of data	99.71675	99.75421	99.67931	Fail
20:80	Algorithm -II will perform better	Check for accuracy at 20% training of data	99.73485	99.69628	99.77352	Success
25:75	Algorithm -I will perform better	Check for accuracy at 25% training	99.73311	99.77107	99.69523	Success

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

		of data				
30:70	Algorithm -I will perform better	Check for accuracy at 30% training of data	99.73425	99.77645	99.69218	Success

The programmer produced adequate results when we tried it in various test environments. The accuracy based on Algorithm, is shown in the following chart .the two methods utilized, namely the Local Outlier Factor Algorithm and the Isolation Forest .The Outlier Fraction values frequently fluctuate depending on the chosen situations .The aforementioned graphic is self-explanatory and characterizes the testing outcomes .The graph for the Isolation Forest Algorithm indicates an increase up to a 20% margin, then it shows a steady development witsh the potential to boost accuracy starting at a 25% margin.

V. CONCLUSION

It goes without saying that using a credit card fraudulently is a crime. This page lists the most common fraud schemes and describes how to recognize them. It also highlights current academic work in the field. Along with the strategy, this study has provided a thorough description of how machine learning may be applied to improve fraud detection. Even though the method is over 99.6% accurate, its precision is still just 28% when only a tenth of the data set is taken into account. When the system is fed the entire dataset, the accuracy rises to 33%.Given the enormous difference in the quantity of legitimate data, it is reasonable to expect such a high accuracy rate.

VI. REFERENCES

[1] KATE SMITH1, VINCENT LEE1, CLIFTON PHUA1, and ROSS GAYLER2 School of Business Systems, Faculty's "A Comprehensive Survey of Data Mining-based Fraud Detection Research" Monash University, Wellington Road, Clayton, Victoria 3800, Australia, Department of Information Technology

[2] Researcher, GJUS&T Hisar HCE, Sonepat published "Survey Paper on Credit Card Fraud Detection by Suman" in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 3, March 2014.

[3] International Journal of Computer Trends and Technology (IJCTT), issue 8, number 1, February 2014, Renu, Suman, "Analysis on Credit Card Fraud Detection Methods."[7] "Credit Card Fraud Detection with a Neural-Network," Proc. IEEE First Int. Conf. on Neural Networks, 2014. Sushmito Ghosh and Douglas L. Reilly.

[4] Deepak pawar, Swapnil Rabse, Sameer Paradkar, and Naina Kaashi, "Detection of Fraud in Online Credit Card Transactions," International Journal of Technical Research and Applications e-ISSN: 2320-8163.