CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHMS

[1]Mr. Suniket Yuvraj Khairnar
suniket.22010917@viit.ac.in

[1]Mr. Om Limbhare
om.22010988@viit.ac.in

[1]Mr. Tejas Raut
tejas.22010583@viit.ac.in

[4]Dr. Jayashree Bagade
Prof. Department of Information Technology
Jayashree.bagade@viit.ac.in

Abstract – Nowadays many people use credit cards for online transaction thanks to the developed technology. Online credit card exchanges have a large portion in the number of online exchanges. Due to several flaws in this system and other issues that are emerging, many online frauds happening day by day. These fraudulent transactions can be located by examining various credit card users' past transaction history dataset's behaviors. There is a chance of a fraudulent transaction occurring if there is even the slightest departure from the established patterns of behaviour. These online frauds cause significant financial damages to the nations."CPO magazine" reports that in 2017, credit card fraud cost Americans $429 on average in lost purchases. The amount of money lost to credit card fraud in 2020 is estimated by the Federal Trade Commission to be around 149 million. Financial institutions should put cybersecurity and technological precautions in place right away to lessen the effects of credit card theft.

In this work, we study Support Vector Machine (SVM), Isolation Forest (IRF), and Local Outlier Factor(LOF) to find credit card fraud. We first describe the dataset used in the study, which includes over 284,000 credit card transactions, of which only 492 are fraudulent. We then perform data pre-processing, including feature scaling and normalization, to prepare the data for machine learning. Next, we evaluate the performance of the SVM, IRF, and LOF algorithms on the credit card fraud detection task using various evaluation metrics, including precision, recall, F1 score .

The result we got after our research is that all three algorithms are effective. SVM has high precision while IRF has high recall value. LOF has both well and hence used in real time systems. In short, our study shows the comparison between different algorithms showing their pros and cons. The motive of this research is to develop an effective credit card fraud detection system.

Keywords: Random Forest algorithm, Credit card fraud, Isolation Forest

I. INTRODUCTION

The term cybercrime refers to any illegal act that is carried out using a computer, its systems, and its online or offline applications with the intention of harming a person. When information technology is utilised to perpetrate or cover up a crime, it happens. Child pornography, spreading malware or viruses into computer systems, software piracy, spoofing, and other types of cybercrime are only a few examples. Online shopping has increased a lot due to credit card. One of the most common type of cybercrime is credit card fraud.

When a person use other's credit card for his/her own payment or withdraw money then it is credit card fraud. Taking control of an active credit card account, changing the billing address, and then reporting the credit card as stolen or lost are all strategies used in certain credit card fraud schemes in order to get a new card for use in making

fraudulent purchases. Another tactic is skimming, which is when a business employee sells the victim's credit card information to the offender. Around £844.8 million in financial fraud losses were reported in the UK in 2018. India has lost at least $100 crore daily for the previous seven years as a result of bank frauds, according to the RBI. One of the key factors making this research so crucial is the devastation that credit card fraud may wreak.

It is difficult to make a safe system to authenticate the legal transaction and prevent fake transaction. Credit card fraud is a serious problem in the economic sector that could result into significant economic losses for both the card holders and their customers. In this era of technology, we can use some machine learning algorithms to find the illegal transaction and prevent the financial loss.

The class mismatch issue, in which the number of authorised transactions is significantly higher than the number of fraudulent transactions in the dataset, is one of the trickiest issues in identifying credit card crimes. This can lead to a biased model that is more likely to predict non-fraudulent transactions and can result in missed fraudulent transactions. To address this issue, various techniques can be used, such as oversampling the minority class, under sampling the majority class, or using synthetic data generation techniques.

we used the PCA dimensionality reduction technique, and then used logistic regression as the classification algorithm to predict whether a transaction is fake or not.

The Simple-Imputer class from the scikit-learn library was used to fill in the dataset's missing values before it was scaled to a typical normal distribution using the Standard-Scaler class. Using the train_test_split function, we divided the data into a training set and a testing set, and the classification report was used to assess the model's performance.

learning techniques used in credit card fraud detection. When labelled data is not available, unsupervised learning methods like clustering, Isolation Forest, local Outlier Factor(LOF) and anomaly detection can be utilised to identify fraudulent transactions.

There is continuing research into creating more precise and effective algorithms for identifying and preventing fraud. Overall, credit card fraud detection using machine learning is a difficult and quickly changing topic.

II. LITERATURE SURVEY:

[1] In research made by Andhavarapu Bhanusri, K. Ratna Sree Valli, P. Jyothi , G. Varun Sai, R. Rohith Sai Subash , they use variour ML techniques like Logistic Regression , Random Forest (with boosting), Naive Bayes to detect online transaction scams . In their study they found that random forest with merging different type of predictions (boosting) techniques is better than any other methods to detect online scam.

[2] BORA MEHAR SRI SATYA TEJA1, BOOMIREDDY MUNENDRA2, Mr. S. GOKULKRISHNAN3 addressed the class imbalance issue of the dataset and used oversampling to finally use Random Forest classifier that got a good accuracy score. Random Forest model works better than Decision Trees.

[3] In research study of "Credit card fraud detection using machine learning ", by Hardik Manek, use Logistic Regression and Autoencoder Neural Network to detect online credit card scams. Author compared result of different ML algorithms and concluded that Autoencoder Neural Network model give best result as compared to other algorithms.

[4] In paper "fraud detection methods in credit card transactions", by Krishna Modi and Reshma Dayma applied the concept of convolutional neural network. Author found that the performance of neural network is better than convolutional neural network since the rate detect
legit transaction as fraudulent is more as compared neural network method. On the other hand, in NN, the rate of detecting fraudulent transection is far more less than CNN which helps to improve performance and accuracy of the model.

[5] Mr. Thirunavukkarasu, etal. calculate an accurate value of credit card fraud detection, i.e. 0.99948 (99.93%) by using random forest method with some new changes in Credit Card Fraud Detection technique using Machine Learning. In comparison to existing techniques, this suggested technique is suitable for bigger datasets and gives more accurate results than other algorithms. We can apply Random Forest algorithm on large data and it perform far better with large training data, but due to large training data during testing it speed may decreased will perform better with more training data, but speed during testing and application may decrease.

[6] In the study "Performance Assessment of Machine Learning Techniques for Credit Card Fraud Detection Using SMOTE and AdaBoost", Emmanuel Ileberi , Yanxia Sun, and Zenghui Wang suggest a machine learning model using Decision Trees Algorithm for detecting credit card frauds. In

this, the author creates a machine learning model that generates predictions using decision trees.
With AdaBoost, the model is improved and more accurate when categorizing transactions as fraudulent or not.

[7] Rahul Powar, Rohan Dawkhar, and Pratichi offer a machine learning model utilising the K-Nearest Neighbor (KNN) method in their paper titled "Credit Card Fraud Detection Using Machine Learning" to identify credit card frauds. This article compares current transaction details with historical transaction data to identify credit card frauds. Historical transaction data includes location, daily expenses, and transaction time of the cardholder. The author has created a model that aids in determining whether or not the new transaction is fraudulent.

[8] Implementation Of Credit Card Fraud Detection Using Support Vector Machine, by M. Amarender Reddy, Dr. Pravin R Kshirsagar, D. Akshitha, G. Alekya, and K. Divya Rosy, offers a machine learning model utilising the Support Vector Machine(SVM) algorithm for identifying credit card frauds. In this, the author preprocesses the data to prevent malware contamination and make the necessary the dataset's data balance and any inaccuracies in the data. The proposed SVM classification performs better than others, according on the author's implementation of the SVM method using a publicly accessible dataset.

[9] In research made by Heta Naik, " Credit card Fraud Detection based on Machine Learning Algorithms", author used Naive Bayes method to classify legit and nonlegit transaction. In this paper, author claims that naive bayes algorithms have very high computational speed and it is best algorithms for classification of datapoints. This help to build the model very fast and give the quick prediction

III. PROPOSED METHODOLOGY

3.1 System Architecture

Applying machine learning to determine credit card fraud is a difficult and time-consuming endeavour that needs close attention to a variety of factors such as data sources, pre-processing, feature engineering, and real-time scoring and reporting. This study looks into the effectiveness of several algorithms in identifying credit card fraud. It employs a number of methods based on machine learning, including as isolation forest (IF), local outlier factor (LOF), and support vector machines (SVM), to decide the most efficient means for credit card providers to detect fraudulent transactions**.**

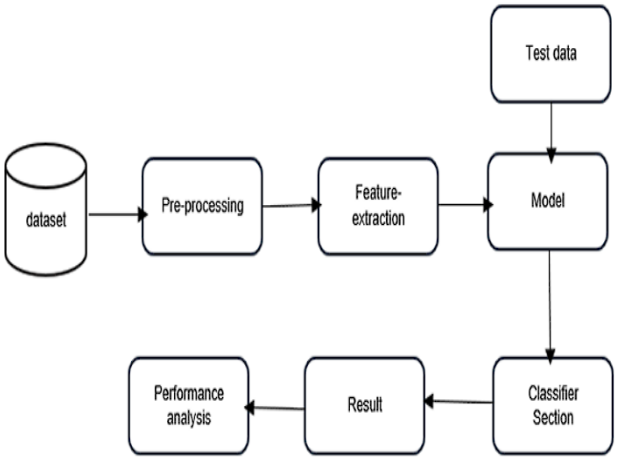

Fig.1 System Architecture

3.2 Dataset:

In this research we took transactions data from Kaggle website that contain credit card transactions made by European cardholders in September 2013 to detect frauds in the credit card online transaction. In this dataset contain 284807 transactions out of which only 492 (0.17%) transaction are non-legitimate. This show that how unbalance the dataset is it is. It is very difficult to tackle this unbalanced dataset problem to make ml model. Because of the secrecy dataset this dataset solely contains numeric data since it it gone through the PCA. The key elements created by pca are v1, v2, and v28. The only features that pca ignores are ‘amount’ and ‘time’. The label amount represents the transaction amount of transaction. The ‘time’ feature measures the seconds since the dataset's initial sale and each succeeding sale . Feature ‘class’ has indicator 1 for illegal transaction and 0 for legal transaction.

3.3 Data Pre-processing:

As an element of the machine learning process, data must be cleaned, transformed, and structured before it can be used in model training. This step is vital because it guarantees the data is in a form that the algorithm using machine learning can comprehend and that it correctly reflects the challenge that the algorithm is trying to solve.

3.4 Data Cleaning:

For applying any machine learning algorithm on dataset, it is important to identify and corrects the errors, inconsistencies, and inaccuracy of the given dataset to improve the accuracy, performance and quality of the machine learning model. This involves identifying and addressing missing data points in the dataset, either by filling in the missing data with estimated values, removing the missing data points. In Fig. 2. shows that in the given dataset does not contain any null values.

```
data.isnull().values.any()

False
```

Fig. 2 : Checking null values in datasets

Also using the drop duplicates () method we are going to remove all duplicate from the dataset. Outliers are the data points which lie at larger distance from other data points. We use boxplot technique to detect outlier in Amount column. The outliers were removed using the Inter Quantile Range (IQR) technique.

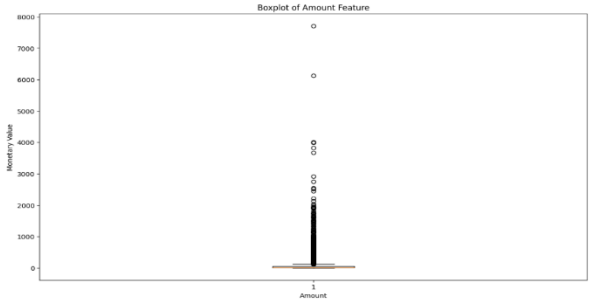

Fig. 3: Boxplot to check outlier values in dataset
From Fig.3 we say that the amount above 3000 are outlier value in our datasets.

3.5 Feature Selection:

It is the method in which we reduced the features of the dataset and consider the most relevant features and variables from a large dataset. The main aim of this technique is to reduce the complexity and dimensionality of the dataset, simplify the model so that model can give fast and more accurate output. It also increases the efficiency of the model.

3.6 Feature Scaling:

This is a pre-processing step to improve the performance of machine learning algorithms by ensuring that all features are on a similar scale. Different methods use for feature scaling like Min-max scaling in which scales the feature values to a range between 0 and 1. Robust scaling is same as min-max scaling but it is more robust to outliers.

3.7 Feature Correlation and Selection:

Feature correlation refers to the relationship between different features in a dataset. If two or more features are highly correlated, then they may contain redundant information and can be removed from the dataset without affecting the performance of the model.

Fig. 4 shows the heatmap for the correlation of the different feature in dataset. Given the

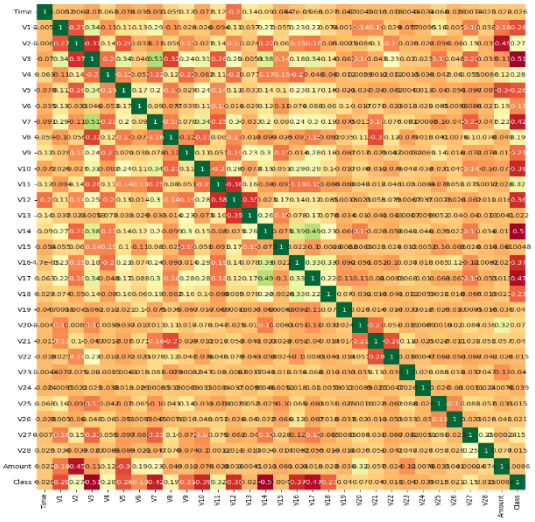

Fig. 4: Boxplot to check outlier values in dataset

size of the dataset, it can be seen that the heatmap is not providing a lot of information.

So, we write a correlation function which study the heatmap and show the feature which are related to each other. We have to pass the dataset and threshold to the correlation function

In Fig.4 shows that for there is no any feature which are 70 percent correlated. So, we can say that there is not any correlation in features in our dataset. This indicates that slight changes in one property have a very small impact on changes in another feature.

```
def correlation(dataset, threshold):
    col_corr = set()  # Set of all the names of correlated columns
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if abs(corr_matrix.iloc[i, j]) > threshold: # we are inter
                colname = corr_matrix.columns[i]  # getting the name o
                col_corr.add(colname)
    return col_corr
```

```
corr_features = correlation(data1, 0.7)
len(set(corr_features))
```

```
1
```

```
corr_features
```

```
{'V7'}
```

Fig. 4: function to check feature correlation

3.8 Selected Best algorithm for Fraud detection:

3.8.1 The Isolation Forest algorithm:

The Isolation Forest algorithm(IF) is foundation of the concept that anomalies in the dataset are easy to isolate from the dataset. It builds number of isolation trees(binary tree) recursively partitioning the data. In this algorithm, different isolation trees are built by selecting the random feature and split value within the given range of the feature. The process continues until every data point is isolated in a separate leaf node.

Here are the steps of the Isolation Forest algorithm:

1. Pick a random subset of data points from the dataset.
2. Pick a random feature from the dataset.
3. Then from max(upper limit) and min (lower limit) of the chosen feature , select the split value.
4. Divide the data points into two groups: feature value less and grater than the split value.
5. Repeat steps 2-4 until each data point is isolated in its own tree.
6. Find out the anomaly score for every data point as the average path length of the data point across all trees in the forest. Anomalies are the data points with scores significantly higher than the average score.

In Fig.5, seeing the confusion matrix for isolation forest algorithm, we can evaluate the performance of algorithm by using confusion matrix.

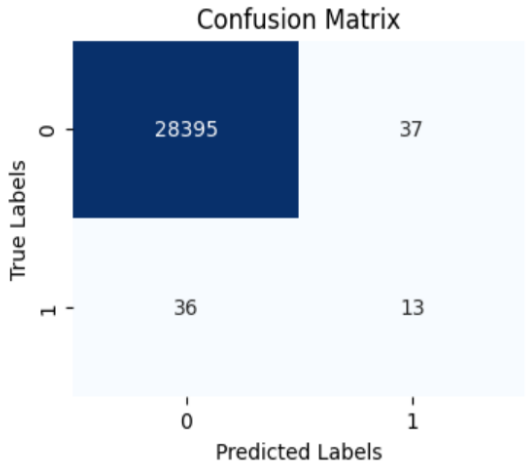

Fig. 5 Confusion Matrix of IFA

3.8.2 The Local Outlier Factor (LOF):

The Local Outlier Factor (LOF) algorithm is a unsupervised anomaly detection method used in machine learning and data mining. It calculates a score for each data point in a dataset based on its proximity to its k-nearest neighbours, and comparing this score to the scores of its neighbours. LOF algorithm follows below steps:

1. Calculate the distance of each data point to all other points in the dataset.
2. Pick a value for k, the number of closest neighbours to take into account. This is a hyperparameter that is responsible to improve performance.
3. For each data point, find its k-nearest neighbours based on the distances calculated in step 1.
4. Calculate LRD for each point. It measures how close the point's neighbours are.

Calculate the local outlier factor (LOF) for each point. This is defined as the ratio of a point's LRD to the average LRD of its k-nearest neighbours. If a point's LOF is significantly greater than 1, it is considered an outlier.

In Fig.6, seeing the confusion matrix for Local outlier factor algorithm(LOF), we can judge the outcomes with below confusion matrix.

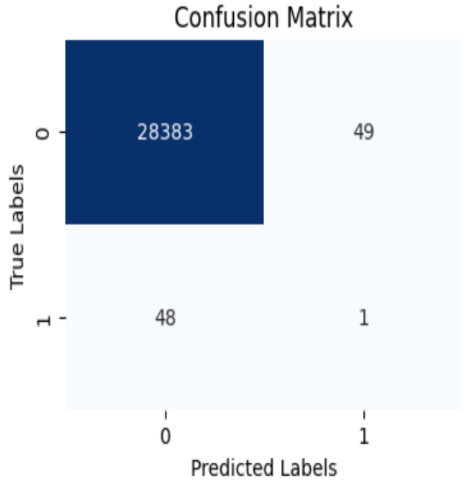

Fig. 6 Confusion Matrix for LOF

3.8.3 Support Vector Machine (SVM):

Support Vector Machine(SVM) is a supervised machine learning algorithms for regression, classifier and outlier detection. In this algorithm we find hyperplane which separates two classes.

Here are the steps of the SVM algorithm:

1. Given a dataset with labelled data, SVM first maps the data to a high-dimensional space.

2. SVM is then used to determine the optimum hyperplane that distinguishes the two classes with maximum amount of gap. Support vectors are the data points that are closest to the hyperplane.
3. If the data is not linearly separable, SVM can use a soft-margin approach.

In Fig.6, seeing the confusion matrix for Support vector machine technique , we can evaluate the outcomes of the SVM by using below matrix table :

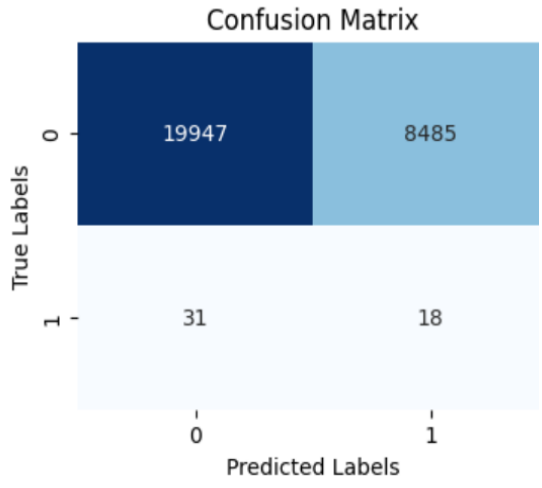

Fig.7 Confusion matrix of SVM

IV. RESULT AND DISCUSSION

A confusion matrix is a table by which we can conclude how well a machine learning model is performing. The matrix compares the predicted values with the actual values.

Actual/Predicted	Positive(1)	Negative(0)
Positive(1)	TP(True positive)	FN (False Negative)
Negative(0)	FP (False Positive)	TN (True Negative)

Table 1: Analysing Confusion Matrix

Table 1. indicate that the confusion matrix From the confusion matrix, different metrics including sensitivity, specificity, accuracy, and error rate are obtained by using formulae for the same.

Now we will compare confusion matrix for different algorithms to check the accuracy of different models.

Result of Isolation Forest algorithm is tabulated below:

Type	Score
Accuracy of model	99.74%
Precision of model	99.87%
Recall of model	99.76%
F1-score of model	99.96%

Table 2: Value calculated by IFA

Result of Local outlier factor algorithm is tabulated below:

Type	Score
Accuracy of model	99.65%
Precision of model	99.87%
Recall of model	99.76%
F1-score of model	99.92%

Table 3 : Value calculated by LOF

Result of Support vector machine is tabulated below :

Type	Score
Accuracy	70.09%
Precision	99.05%
Recall	70%
F1-score	82%

Table 4 : Value calculated by SVM

Compares the accuracy of the respective models.

Model	**Accuracy (Precision)(In percentage)**
Isolation forest (IF)	**99.74%**
Local outlier factor (LOF)	**99.65%**
Support vector machine (SVM)	**70.09%**

Table 5: Accuracy of respective models

From Table 5, we can say that Isolation Forest algorithm and Local outlier factor both give high accuracy as compared to support vector machine having accuracy about 70%.

IF algorithm outperformed both SVM and LOF in terms of all the evaluated metrics. The IF algorithm achieved the highest precision, recall, F1-score, as well as the highest accuracy among the three algorithms.

On the other hand, SVM and LOF also performed reasonably well in the credit card fraud detection

task, with LOF give high accuracy as compared to SVM.

V. CONCLUSION:

In this study, we examined a number of machine learning(ML) approaches, together with SVM, Isolation Forest, and LOF, that identify fraudulent activity in transactions made with credit cards. We can forecast the chances of a fraud taking place resulting from a credit card transaction if these methods are used. To stop frauds, stop major losses for banks, and reduce risks, substantial steps and procedures can be taken. Performance is measured using accuracy, support, recall, precision, and f1-score. By comparing all the three algorithms, we found that Isolation Forest performs better than the Local Outlier Factor and SVM.

VI. FUTURE SCOPE:

By doing this research on various classification algorithms, we conclude that performance of isolation forest algorithms is better than LOF and SVM.

The algorithm may identify a transaction as anomalous but may not have enough information to determine whether it is actually fake or legal.

The isolation forest can be helpful in detecting credit card fraud, they should not be relied on solely for determining the nature of individual transactions. In the future, we can work to solve this problem by using certain methods.

VII. REFERENCES:

[1] Gupta, Shalini, and R. Johari. "A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant." Communication Systems and Network Technologies (CSNT), 2011 International Conference

[2] 'A history of algorithms' understanding data science main algorithms by devoteam.

[3] "Anomaly detection with Isolation Forests" by Liu F T, Ting K M, and Zhou Z in ICDM 2008, Eight's IEEE International Conference on Data Mining

[4] "How to perform anomaly detection with the Isolation Forest algorithm" from towardsdatascience.com

[5] Andhavarapu Bhanusri, etal. Paper on "Credit card fraud detection using ML techniques."

[6] BORA MEHAR SRI SATYA TEJA1, BOOMIREDDY MUNENDRA2, Mr. S. GOKULKRISHNAN3 "Credit Card Fraud Detection", IRJET, March 2022

[7] Emmanuel Ileberi, Yanxia Sun & Zenghui Wang , "A MACHINE LEARNING BASED CREDIT CARD FRAUD DETECTION USING THE GA ALGORITHM FOR FEATURE SELECTION " , Journal of Big Data volume 9, Article number: 24 (2022)

[8] M. Amarender reddy 1, Dr. Pravin R Kshirsagar2 , D. Akshitha3, G.Alekya4, K. Divya rosy5 , "IMPLEMENTATION OF CREDIT CARD FRAUD DETECTION USING SUPPORT VECTOR MACHINE" , Journal of Engineering Science , Vol 12

[9] Vaishnavi Nath Dornadula a, S Geetha a , "Credit Card Fraud Detection using Machine Learning Algorithms" , 2nd International Conference on Recent Trends in Advanced Computing ICRTAC - DISRUP - TIV INNOVATION , 2019 November 11-12, 2019

[10] AlEmad, Meera, "Credit Card Fraud Detection Using Machine Learning" (2022).

[11] SamanehSorournejad1, Zahra Zojaji2 , Reza Ebrahimi Atani3 , Amir Hassan Monadjemi4 , " A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective "