Credit Card Fraud Detection System Using Machine Learning

[1]Siddhi Bhor, [2]Purva Lokare, [3]Aishwarya Rele, [4]Harshali Rambade
[1,2,3,] Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India

***Abstract*—** Credit card fraud detection is currently occurring on a large scale everywhere. This problem stands as there has been a sharp increase in the online transactions and usage of e-commerce platforms. It is essential that credit card companies are able to identify deceitful credit card transactions so that customers do not suffer from unnecessary expenditure. Such problems can be dealt with Data Science along with Machine Learning. Credit card frauds usually occur when theft of the card is involved for any of the purposes that are not authorized or even when the fraudster is able to extract the credit card information for their own use. The credit card fraud detection system was introduced to detect such fraudulent activities. The aim of the project is to focus mainly on machine learning algorithms. Random forest algorithm, logistic regression and the SVM algorithm are being used. The results procured from the algorithms are based on accuracy, precision, recall, and F1-score. The confusion matrix is the basis for the plotting of the ROC curve. A comparison occurs of the Random Forest, Decision tree, LGBM and the Nearest neighbors algorithms and the algorithm that has the greatest accuracy, precision, recall and F1-score is decided as the best algorithm that is used to detect the fraud

***Keywords*—**: Fraud detection, Credit card, Random Forest, Decision tree, LGBM , Nearest neighbors.

I. INTRODUCTION

Credit card fraud is a massive term for fraud and larceny committed by using or involving at the time of transferring funds by using the card. There may be a dual purpose here, either to purchase stuff without paying from one's own pocket or to transfer unauthorized funds from an account. Credit card fraud also accounts for identity theft. Although credit card fraud is sort of an economic crime, it is also the crime which most people associate with ID theft. In 2000, out of 13 billion transactions made annually, approximately 10 million or one out of every 1300 transactions turned out to be fraudulent. Also, 0.05% (5 out of every 10,000) of all monthly active accounts were fraudulent. Presently, fraud detection systems are introduced to safe guard one- twelfth of one percent of all transactions processed which still does not prevent the loss of billions of dollars. Business establishments are in great jeopardy due to credit card frauds today. Therefore, to defy the fraud effectively, it is essential to first understand the mechanisms of executing a fraud. Credit card fraudsters certainly utilize a huge amount of ways to commit fraud. In simple terms, Credit Card Fraud is defined as "when a certain individual uses the other individuals' credit card for self-sufficient reasons while the owner of the card and the card issuer are in the dark about the happenings with the card". Card fraud commences either with the larceny of the physical card or with the crucial data associated with the account, that includes the card account number or other information that compulsorily has to be available to a merchant during a permissible transaction. Hence, most of the time it is difficult to identify the credit card fraud. Machine Learning is identified as one of the most accomplished techniques for fraud identification. It uses regression and classification method for guessing fraud in a credit card. Several learning algorithms have been examined for fraud detection in credit cards which comprises of neural networks, Logistic Regression Support Vector Machines and Random Forest. This project testifies the performance of above algorithms. Depending on their ability to list down whether the transaction was authorized or fraudulent and then compares them. The comparison occurs with the help of performance measure accuracy, specificity and precision.

II. LITERATURE SURVEY

The research on credit card fraud detection uses both Machine Learning and Deep Learning algorithms. In this section, the work done in two different points the methods that are readily available for fraud detection, and The

techniques that are available to handle the imbalanced data. To handle the imbalanced data some of the techniques are available. They are (a) classification methods (b) sampling methods (c) resembling techniques. Here are some of the Machine Learning algorithms that are used for credit fraud detection are support vector machine(SVM), decision trees, logistic regression, gradient boosting, K-nearest neighbor, etc

In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, Sarika jain have researched various techniques[I] for credit cards fraud detection such as support vector machines(SVM), artificial neural networks(ANN), Bayesian Networks, Hidden Markov Model, K-Nearest Neighbours (KNN) Fuzzy Logic system and Decision Trees. In their paper, they have observed that the algorithms knearest neighbor, decision trees, and the SVM give a medium level accuracy. The Fuzzy Logic and Logistic Regression give the lowest accuracy among all the other algorithms. Neural Networks, naive bayes, fuzzy systems, and KNN offer a high detention rate.

The Logistic Regression, SVM, decision trees offer a high detection rate at the medium level. There are two algorithms namely ANN and the Naïve Bayesian Networks which perform better at all parameters. These are very much expensive to train. There is a major drawback in all the algorithms. The drawback is that these algorithms don't give the same result in all types of environments.

They give better results with one type of datasets and poor results with another type of dataset. Algorithms like KNN and SVM give excellent results with small datasets and algorithms like logistic regression and fuzzy logic systems give good accuracy with raw and un-sampled data used.

III. PROPOSED SOLUTION

The main aim of this project is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like that the Random Forest, Decision tree, LGBM and the Nearest neighbors algorithms. Then these three algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow for the credit fraud detection problem includes the splitting of the data, model training, model deployment, and the evaluation criteria . In this model we take the Kaggle credit card fraud dataset and pre-processing is to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the LGBM models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score is calculated for bot the models. Finally the comparison of the credit card fraud transactions more accurately.

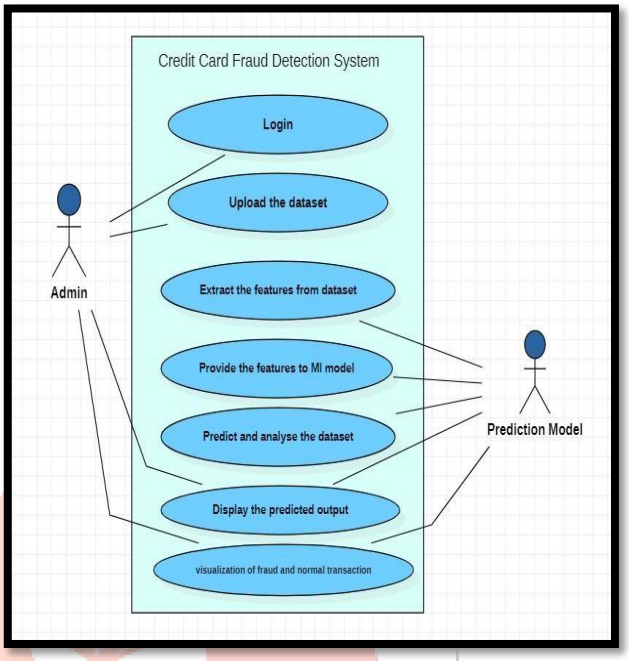


Fig. 1 Uml Diagram

Firstly Read the dataset. Random Sampling is done on the data set to make it balanced. Then Divide the dataset into two parts i.e. Train dataset and Test dataset. Step Feature selection are applied for the proposed models. Accuracy and performance metrics has been calculated to know the efficiency for different algorithms. Then retrieve the best algorithm based on efficiency for the given dataset.

- **Algorithm to be used – Random Forest, Decision tree, LGBM and the Nearest neighbors.**

IV.METHODOLOGY

The Credit card fraud detection proceeds in the following way.

Upload the dataset ,can be a single or multiple dataset the algorithm will read the dataset. Random Sampling is done on the data set to make it balanced. Then Divide the dataset into two parts i.e., Train dataset and Test dataset. Step Feature selection are applied for the proposed models. Accuracy and performance metrics has been calculated to know the efficiency for different

algorithms. Then retrieve the best algorithm based on efficiency for the given dataset.

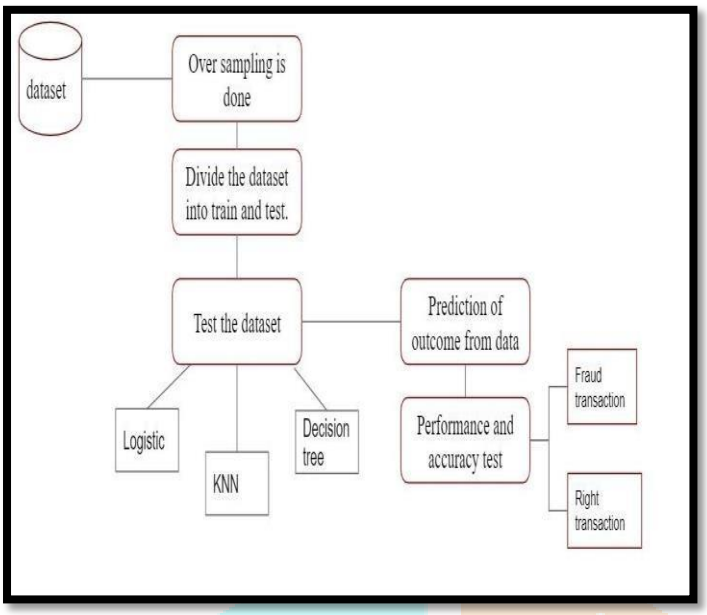
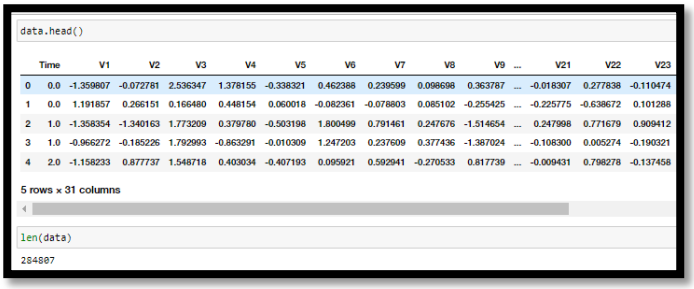


Fig 2 Methodology

Step 1:	Import the dataset
Step 2:	Convert the data into data frames format
Step3:	Do random oversampling using ROSE package
Step4:	Decide the amount of data for training data and testing data
Step5:	Give 70% data for training and remaining data for testing.
Step6:	Assign train dataset to the models
Step7:	Choose the algorithm among 3 different algorithms and create the model
Step8:	Make predictions for test dataset for each algorithm
Step9:	Calculate accuracy for each algorithm
Step10:	Apply confusion matrix for each variable
Step11:	Compare the algorithms for all the variables and find out the best algorithm.

Table 1. Algorithm steps for finding the Best algorithm

IV. Results



```
data.head()
```

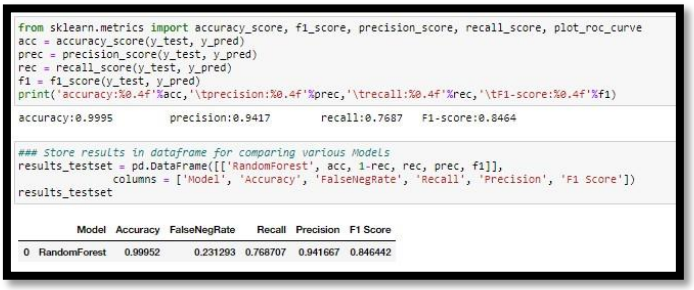
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458

5 rows × 31 columns

```
len(data)
```

284807

Fig 3.Data set



```
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, plot_roc_curve
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred)
rec = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print('accuracy:%0.4f'%acc,'\tprecision:%0.4f'%prec,'\trecall:%0.4f'%rec,'\tF1-score:%0.4f'%f1)
```

```
accuracy:0.9995          precision:0.9417          recall:0.7687    F1-score:0.8464
```

```
### Store results in dataframe for comparing various Models
results_testset = pd.DataFrame([['RandomForest', acc, 1-rec, rec, prec, f1]],
             columns = ['Model', 'Accuracy', 'FalseNegRate', 'Recall', 'Precision', 'F1 Score'])
results_testset
```

	Model	Accuracy	FalseNegRate	Recall	Precision	F1 Score
0	RandomForest	0.99952	0.231293	0.768707	0.941667	0.846442

Fig4.Algorithm

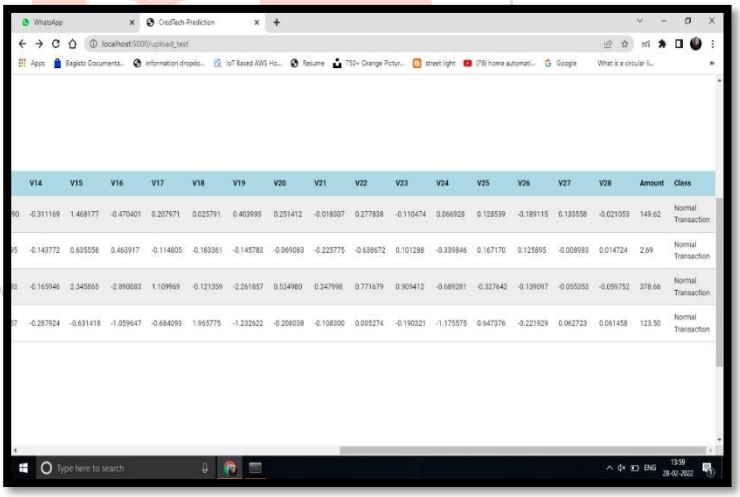


Fig 5.Output

V. Conclusion

In this paper, Machine learning technique like Logistic regression, SVM and Random forest were used to detect the fraud in credit card system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system By comparing all the three method, the best

Algorithm is found

VI. References

i. Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

ii. CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 A Comprehensive Survey of Data Mining-based Fraud Detection Research published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

iii. Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

iv. Research on Credit Card Fraud Detection Model Based on Distance Sum by Wen-Fang YU and Na Wang published by 2009 International Joint Conference on Artificial Intelligence

v. Credit Card Fraud Detection through Parenclitic Network Analysis- By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages