IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

**ISSN: 2454-132X**
**Impact Factor: 6.078**
**(Volume 7, Issue 3 - V7I3-1433)**
Available online at: https://www.ijariit.com

# Credit card fraud detection

*Vadlamudi Tony Titus*
*tonydtitus223@gmail.com*
*Gandhi Institute of Technology and Management,*
*Visakhapatnam, Andhra Pradesh*

*Lakshmi Choudari*
*lchoudar@gitam.edu*
*Gandhi Institute of Technology and Management,*
*Visakhapatnam, Andhra Pradesh*

*Shashank*
*121710305024@gitam.in*
*Gandhi Institute of Technology and Management,*
*Visakhapatnam, Andhra Pradesh*

*Gunavardhan*
*121710305018@gitam.in*
*Gandhi Institute of Technology and Management,*
*Visakhapatnam, Andhra Pradesh*

## ABSTRACT

***Credit card Fraud has emerged as a bigger problem with an increase in online transactions and e-commerce. A credit card fraud happens when a fraudster uses your credit card information to make unauthorized purchases in your name. Credit card companies need to address these fraudulent transactions so that customers are not charged for the goods they did not buy. In this paper, we aim to tackle such fraudulent transactions by identifying them using various machine learning algorithms like Logistic Regression, Random Forest and, eXtreme Gradient Boosting. The results of the algorithms are based on accuracy, precision, recall, and F1-score. The algorithms are compared, and the algorithm that has the greatest accuracy, precision, recall, and F1-score is considered as the best algorithm that is used to detect fraudulent transactions.***

***Keywords—*** *credit card transactions, Fraudster, SMOTE, Machine Learning.*

## 1. INTRODUCTION

Credit card fraud happens when someone, a fraudster uses your stolen credit card information to do unauthorized purchases in your name or take out cash advances using your account. Suppose you're studying your credit card statement, and you notice Rs.200/- charge at a nearby electronics store. Then you realize that you have not shopped at the store in a while. That unnatural transaction could be a sign that you've been a victim of credit card fraud.

### 1.1 How does the fraud happen?

- A fraudster digs into your trash receives discarded receipts or credit card statements that include your account number and uses that information to commit fraud.
- An identity fraudster lures you to a fraudulent website where you are tricked into providing your card number.
- Someone may have inserted a credit card scanner to steal your credit card details.

### 1.2 Need for Fraud Detection

- Credit card fraud is on the rise with the advent of modern technology.
- Credit card fraud costs consumers and financial companies billions of dollars every year, and fraudsters are constantly trying to find new laws and strategies to commit illegal acts.
- Therefore, fraudulent means have become important for banks and financial institutions to minimize their losses.
- Detection methods are being developed continuously to prevent criminals from complying with their fraudulent methods. With the rapid growth in technology and widespread internet usage, fraudsters are coming up with new ways to do fraud.

## 2. LITERATURE REVIEW

The main purpose of any credit card fraud program is to identify suspicious events and present them to the analyst while allowing normal transactions to be processed automatically. To date, credit card companies have delegated this responsibility to law-based purposes using a professional set of rules. But now, they areturning to machine learning, as it can bring major improvementsto the process.

- Higher accuracy of fraud detection: Compared to legal-based solutions, machine learning tools have higher clarity and return relevant results as they look for additional features. This is because ML technology can look at multiple data points, including very small details and patterns associated with a particular account.
- Fewer false positives: False positives happen when a system identifies a legitimate transaction as a fraud and wrongly classifies it.

- Adapt to new patterns: Unlike rule-based systems, ML algorithms are good with a constantly changing environment and financial conditions. They enable analysts to identify new patterns and introduce new rules to prevent new types of scams.

Supervised learning: Supervised learning means that a model learns from previous examples and is trained on labeled data. Here, the dataset has tags that tell the model which patterns are related to fraud and which represent normal behavior.

Unsupervised reading: Unsupervised reading is also called anomaly detection as it automatically captures unfamiliar patterns. In this case, database training comes without labels or instructions. This method is less accurate compared to supervised reading.

### 2.1 PayPal: taking an in-depth learning approach

Paying processor and payment gateway provider with more than 200 million active accounts worldwide, PayPal invests $ 300 million annually in anti-fraud technology. Ten years ago, their program relied on postponement - one of the most common learning methods used for classification. Then they added more closely monitored learning strategies, namely Extracted Trees (GBTs) and neural networks. As a result, system accuracy has been improved by 50 percent.

Currently, PayPal is taking a new, in-depth approach to learning, benefiting from the huge amounts of fraudulent data collected over the years. In-depth learning models have already been shown to be 10 to 20 percent more accurate than machine learning algorithms in detecting real-time fraud. PayPal's latest fraudulent loss rate is 0.28 percent.

### 2.2 Amazon: Amazon Web Services (AWS) customer database training models

This year the largest online retailer publicly launched the fully owned Amazon Fraud Detector. The technology is powered by machine learning and the company's 20 years of experience in combating online scams.

To take advantage of the Fraud Detector, AWS customers must be integrated into the service via an Application Programming Interface (API) and feed historical program information, as well as symbols that indicate fraudulent and legal transactions. Amazon combines this information with its data to create models that identify signs of identity theft or laundry.

## 3. METHODOLOGY

The purpose of this paper is to identify fraudulent credit card transactions so that customers of credit card companies can be charged for items they did not purchase.

### 3.1 Challenges involved in credit card frauddetection

- Large Data is processed every day, and the model build must be fast enough to respond to the scam in time.
- It is highly imbalanced data which makes it hard to detect fraudulent transactions.
- To maintain the privacy of the customer, we onlyget the transformed data to work on.
- Data can be misclassified as not all fraud transactions are identified and reported.
- Adaptive techniques used against the model by the scammers.

### 3.2 Ways to deal with these challenges

- The model used should be simple and fast enough to find the unusual and set it as a deceptive transaction very quickly.
- Data inequalities can be eliminated using strategies such as SMOTE, where we create fewer input data to match the bulk category.
- By protecting user privacy, data size can be reduced.
- A more reliable source, double-check, at least for model training, should be considered.
- We can simplify the model and translate it so that when the fraudster adapts to it in some way, we have a new and used model.

The Database is taken from Kaggle. The databases contain a credit card transaction in September 2013 by European cardholders. This database presents a two-day transaction, in which we have 492 frauds of 284,807 payments. However, the database does not accurately measure, where the positive (fraud) category accounts for 0.172% of all transactions. Due to privacy issues, input variables are converted to values using principal component analysis (PCA) modification.

Libraries used in the project are pandas. It is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool. Seaborn is a Python data visualization library based on Matplotlib. It provides an interface for distribution plots of data and informative statistical graphics, and lastly, Scikit- learn, which provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

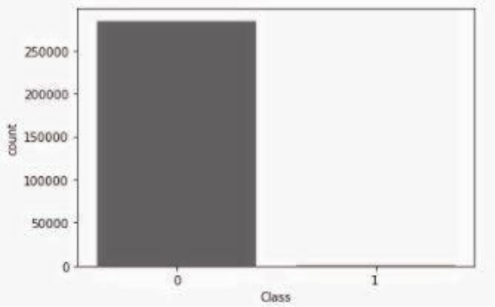


**Fig. 1. Transactions Distribution**

The above graph shows that the number of fraudulent transactions is much lower than the legitimate ones. It is clear from the above distribution that the number of fraudulent transactions is very low. Directly running any classifier will yield a higher accuracy score as one of the classes is highly represented in the data.

Using Kdeplot using a Seaborn library, we visualize the data against the time and amount to have a deep understating of the data we are dealing with.

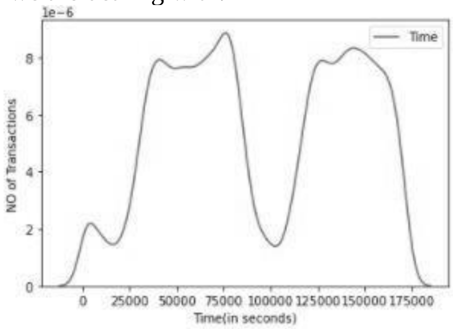


**Fig. 2. Time vs. Transactions**

The above graph shows the times at which transactions were done. It is observed that the least number of transactions were made during night time and highest were done during the day.

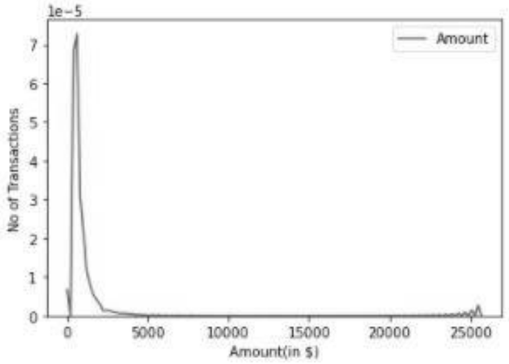


**Fig. 3. Amount vs. Transactions**

The above graph represents the amount that was transacted. A majority of transactions are relatively small, and only a handful of transactions come close to the maximum transacted amount.

### 3.3 Machine Learning (ML) Algorithms to be used are

- Logistic Regression
- Random Forest Algorithm
- eXtreme Gradient Boosting Algorithm

### 3.3.1 Logistic Regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- The dependent variable is binary in nature, having data coded as either 1 (represents success) or 0 (represents failure).
- It is used for various classification problems such as spam detection, Diabetes prediction, cancer detection, etc.

### 3.3.2 Random Forest Algorithm

- Random forest is also a supervised learning algorithm that is used for both classifications as well asregression.
- Random forest algorithm generates decision trees using data samples and then gets the prediction from each of them and finally selects the best solution through voting.
- It is an ensemble method that is better than a single decision tree because it reduces the over-fitting by averaging the result.

### 3.3.3 eXtreme Gradient Boosting Algorithm

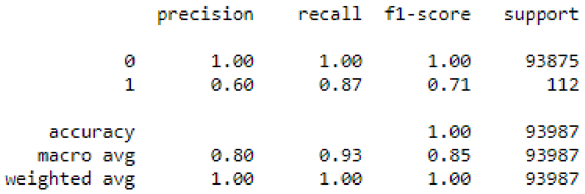
- In a regular machine learning model, such as decision tree, we use a single model.
- Whereas, boosting uses a more iterative approach. It is still technically an ensemble technique in that many models are combined together to perform the final one.
- With each new model being trained to correct the errors made by the previous models.
- Gradient Boosting is an approach where new models are trained to predict the errors of prior models.

## 4. IMPLEMENTATION

This idea is difficult to implement in real life because it requires cooperation from banks, which are not willing to share information due to their market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers that follow similar approaches and gathered results. 28 out of 31 columns of the dataset are PCA generated. We do not have original features and more background information about the data (V1, V2, V28).

### 4.1 Logistic Regression

Implementing logistic regression Algorithm directly to classify fraudulent and non-fraudulent transactions. This approach is not correct. It's just to showcase what is wrong. Since our dataset is highly imbalanced, the accuracy is only reflecting the underlying class distribution, which in this case is non-fraudulent cases. Therefore, we need to compute other metrics like precision, recall, F1 score using a confusion matrix to detect overfitting.



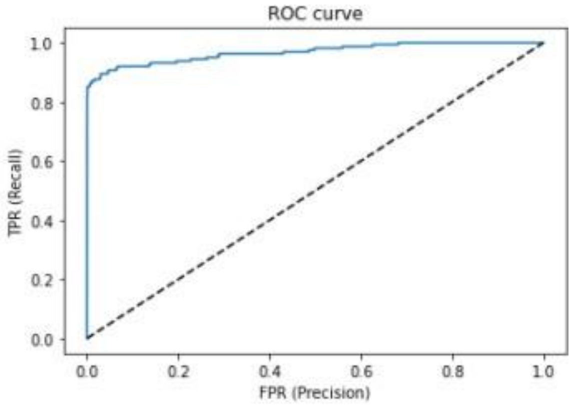
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93875
           1       0.60      0.87      0.71       112

    accuracy                           1.00     93987
   macro avg       0.80      0.93      0.85     93987
weighted avg       1.00      1.00      1.00     93987
```

**Fig. 4. Performance of Logistic Regression**

We observed that our model is predicting non-fraudulent transactions very well. However, they are not predicting fraudulent transactions that well. For fraudulent transactions, precision is 60%, recall is 87%, and F1-score of 71% only.

### 4.2 So, which metric is most important?

- Precision is an important metric - We want our model to be precise. If our model is not predicting correctly, then banks would block the genuine transactions, which might lead to customer dissatisfaction.
- Recall is also an important metric - If recall is low, then our model is not predicting fraudulent transactions well. Bank will have to incur losses as these transactions are not getting detected correctly.
- Therefore, we consider the F1 score, which is a harmonic mean of precision and recall scores.
- To compare the models an ROC curve (Receiver Operating characteristic Curve) which is a graph showing the performance of a classification model at all classification thresholds is used.
- Using ROC curve we determine the AUC that stands for "Area under the ROC Curve." Thatis, AUC measures the entire two-dimensional areaunderneath the entire ROC curve



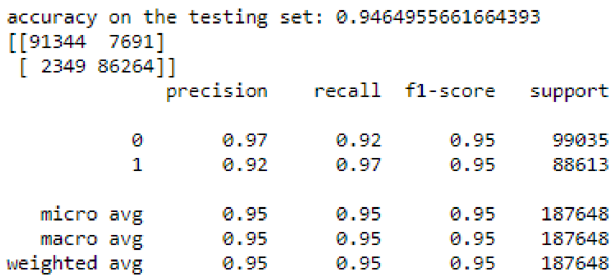
```
Area under curve (AUC):  0.967970973015826
```

**Fig. 5. ROC and AUC (Logistic Regression)**

Though AUC is high, we are not satisfied with our results as it is not capturing fraudulent transactions that well. Our model is predicting non-fraudulent transactions better than fraudulent transactions. We also know that this is happening because of a highly imbalanced dataset. So, what should we do? Following is a technique I have used to tackle the issue of an imbalanced dataset.

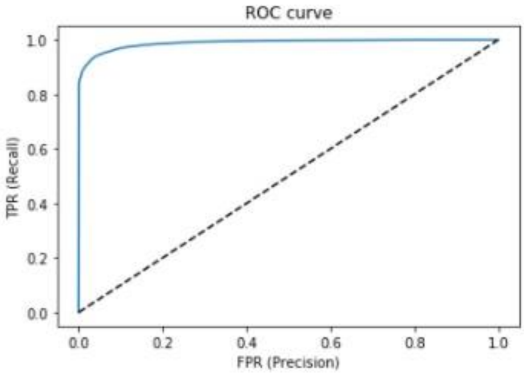
### 4.3 Synthetic Minority OversamplingTechnique (SMOTE)

SMOTE is a sampling method in which samples are made for a small group. This algorithm helps to overcome the problem of overcrowding caused by random sampling. It focuses on the space gap to produce new events with the help of translation between beautiful scenes that lie together. The problem with unequal segregation is that there are very few examples of the minority category of the effective learning limit. One way to solve this problem is to sample examples from a small group. This can be achieved by simply repeating examples from the minority category in the training database before installing the model. This may limit class distribution but does not provide additional details for the model. Development of repetition of examples from the minority category to include new examples from the minority category. This is a type of table data addition and can be very effective.



```
accuracy on the testing set: 0.9464955661664393
[[91344  7691]
 [ 2349 86264]]
              precision    recall  f1-score   support

           0       0.97      0.92      0.95     99035
           1       0.92      0.97      0.95     88613

   micro avg       0.95      0.95      0.95    187648
   macro avg       0.95      0.95      0.95    187648
weighted avg       0.95      0.95      0.95    187648
```

**Fig. 6. SMOTE**



```
Area under curve (AUC):  0.9890538533977183
```

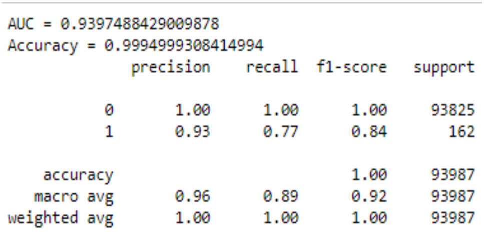
**Fig. 7. ROC and AUC (SMOTE Technique)**

AUC is high (98.9%). Along with it, our F1-score, precision, and recall scores are also high (95%, 92%, and 97%, respectively).

With this approach, we are able to correctly predict 92% of fraudulent cases correctly. After oversampling the minority class, our detection rate improved significantly. The machine has variable and sufficient data to work with and classify them correctly compared to the previous algorithm without SMOTE technique.

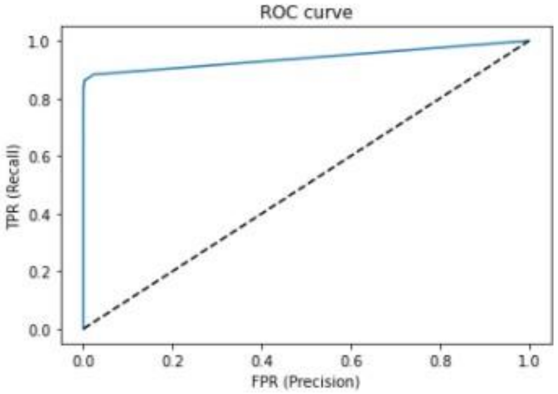
### 4.4 Random Forest Algorithm

Random forest algorithm generates decision trees using the given data samples and then gets the prediction from each of them and finally selects the best solution through voting.



```
AUC = 0.9397488429009878
Accuracy = 0.9994999308414994
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93825
           1       0.93      0.77      0.84       162

    accuracy                           1.00     93987
   macro avg       0.96      0.89      0.92     93987
weighted avg       1.00      1.00      1.00     93987
```

**Fig. 8. Random Forest Algorithm**



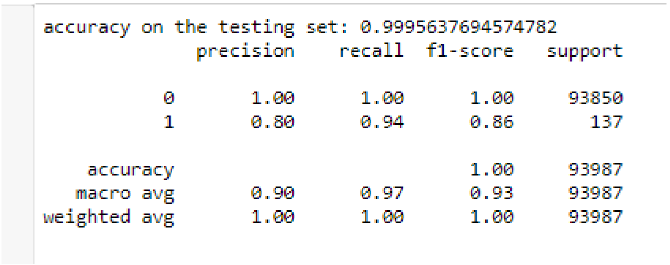
```
Area under curve (AUC):  0.9397488429009878
```

**Fig. 9. ROC and AUC (Random Forest)**

With the Random forest algorithm, we observed that the AUC very low compared to the SMOTE technique we used earlier, even precision, recall and f1-score are not that great.

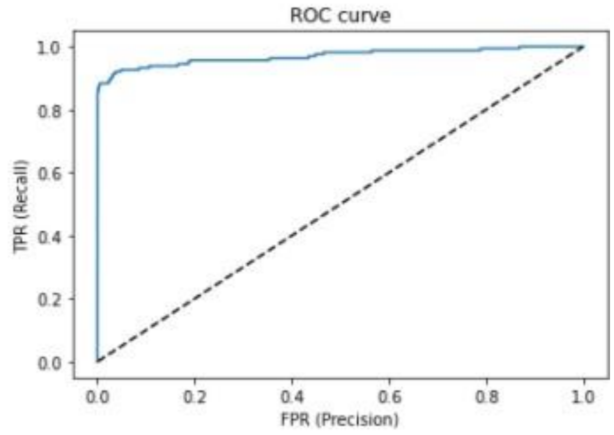
### 4.5 eXtreme Gradient Boosting Algorithm

eXtreme Gradient Boosting is a boosting technique where we implement gradient boosted decision trees designed for speed and performance.



```
accuracy on the testing set: 0.9995637694574782
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93850
           1       0.80      0.94      0.86       137

    accuracy                           1.00     93987
   macro avg       0.90      0.97      0.93     93987
weighted avg       1.00      1.00      1.00     93987
```

**Fig.10. xgboost Algorithm**



```
Area under curve (AUC):  0.9697731855667729
```

**Fig.11. ROC & AUC(XGBOOST)**

XGBoost algorithm definitely produced better results than the random forest and has a higher AUC score. However still, the logistic regression with SMOTE technique outperforms both xgboost and the Random Forest algorithms.

## 5. CONCLUSION

This paper put forwards various machine learning algorithms that can effectively tackle the imbalanced data set of transactions and provide good results in detecting fraudulent transactions. Techniques such as minority oversampling were used to deal with the unevenness of the dataset to get better results. Various performance metrics such as precision, recall, and F1 score are used to evaluate the models. It is observed that the transactional behavior of both the customers and fraudsters is changing with the period. These changes need to be incorporated into the model concurrently so that it can be adapted and classify transactions effectively.

## 6. FUTURE ENHANCEMENTS

While the main objective of this paper has been achieved, I would like to work on various other machine learning algorithms to see ifthe AUC score can be further improved and effectively classify fraudulent and normal transactions. I would also like to use techniques such as cross-validation and bagging to address the issue of an imbalanced dataset. There is much room for improvement in this project. With a better dataset, the performance also can be greatly improved.

## 7. REFERENCES

[1] "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.

[2] 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.

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

[4] "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by the 2009 International Joint Conference on Artificial Intelligence.

[5] "Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, Regino Criado, and Article ID 5764370, 9 pages.