# ONLINE FRAUD TRANSACTION DETECTION USING MACHINE LEARNING

**Vedant Mayekar[1], Siddharth Mattha[2], Sohan Choudhary[3], Prof Amruta Sankhe[4]**

*[1-3]Student, Information Technology Department, Atharva College of Engineering, Maharashtra, India*
*[4]Asst. Professor, Atharva College of Engineering, Maharashtra, India*

---***---

***Abstract*** *- In today's world, people depend on online transactions for almost everything. Online transactions have their own merits like easy to use, feasibility, faster payments etc., but these kinds of transactions also have some demerits like fraud transactions, phishing, data loss, etc. With increase in online transactions, there is a constant threat for frauds and misleading transactions which can breach an individual's privacy. Hence, many commercial banks and insurance companies devoted millions of rupees to build a transaction detection system to prevent high risk transactions. We presented a machine learning - based transaction fraud detection model with some feature engineering. The algorithm can get experience; improve its stability and performance by processing as much as data possible. These algorithms can be used in the project that is online fraud transaction detection. In these, the dataset of certain transactions which is done online is taken. Then with the help of machine learning algorithms, we can find the unique data pattern or uncommon data patterns which will be useful to detect any fraud transactions. For the best results, the XGBoost algorithm will be used which is a cluster of decision trees. This algorithm is recently dominating this ML world. This algorithm has features like more accuracy and speed when compared to other ML algorithms.*

***Keywords*** **– Fraud detection, Machine learning, Xgboost algorithm, classification, Data pre-processing, Prediction.**

***1. Introduction*** - In today's world, we are on the verge to become a cashless world. According to various surveys and researches, people performing online transactions has increased a lot, it's expected that in future years this will go on increasing. Now, while this might be exciting news, on the other-side fraudulent transactions are on the rise as well. Even due to various security systems being implemented, we still have a very high amount of money being lost due to fraudulent transactions. Online Fraud Transaction can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card-issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users to estimate, perceive or avoid objectionable behavior, which consists of fraud, intrusion, and defaulting. Most of the time, a person who has become a victim of such fraud doesn't have any idea about it until the very end.

Necessary preventive measures can be taken to stop this abuse and the behavior of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, this is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over time.

These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time. So, in this project, what we have tried is to create a Web App for the detection of such types of frauds with the help of Machine Learning.

### *2. Scope of the project -*

Online Fraud Transaction Detection System is basically an extension of the existing system. Using this system, the algorithms will be built to go through the dataset and provide the appropriate output. In the long run, this system will be quite beneficial as it provides an efficient system to create a secure transaction system to analyse and detect fraudulent transactions. The Xgboost algorithm is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. This accuracy can be increased further by providing a huge dataset for model training. The scope of this application is very far reaching. This system can be used to detect the features of fraud transactions in a dataset which is very well applicable in various sectors like banking, insurance, e-commerce, money transfer, bill payments, etc. This will indeed help to increase security.

## *3. Literature Review -*

**"A Comparative Analysis of XGBoost"- Gonzalo, Martinez**

This work proposes a practical analysis of how this novel technique works in terms of training speed, generalization performance and parameter setup. In addition, a comprehensive comparison between XGBoost, random forests and gradient boosting has been performed using carefully tuned models as well as using the default settings. The results of this comparison may indicate that XGBoost is not necessarily the best choice under all circumstances but it has its own advantages over other algorithms.

**"Credit Card Fraud Detection in Data Mining using XGBoost Classifier" - Rahul Goyal, Amit Kumar Manjhvar, Vikas Sejwar**

In this research paper, the proposed system uses a combination of SMOTE technique followed by the Xgboost classification algorithm to classify fraud activities. SMOTE stands for Synthetic Minority Oversampling Technique. This is a technique to increase the number of cases in your dataset in a balanced way. SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority cases. They measured their performance and dignified it using only publicly available datasets for Credit card frauds by using XGBoost.

**"Customer Transaction Fraud Detection Using Xgboost Model" - Yixuan Zhang, Jialiang Tong,Ziyi Wang,Fengqiang Gao**

This paper puts forward a Xgboost-based fraud detection algorithm. They firstly performed data cleaning for the purpose of putting out some anomalies. In addition, to solve the unbalanced distribution problem of labels, the SMOTE (Synthetic Minority Oversampling Technique) was used to oversample the minority class. While for categorical features, label encoding algorithm is used to encode categorical data. Finally, a highly effective and widely used algorithm that is Xgboost was implemented for the classification

**"Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection" - Dr. C. Victoria Priscilla, D. Padma Prabha**

This paper presented the influence of optimization in the XGBoost model for handling class imbalance in the dataset effectively by itself. The best parameters are identified using the RandomizedSearchCV method available in the scikit-learn package in python. The mathematical derivative of XGBoost is discussed and the experiment was conducted with two real-world imbalanced datasets by integrating different sampling methods. Our findings proved that the proposed XGBoost can achieve higher accuracy for extremely imbalanced data without sampling.

## *4. Proposed System*

In this system, we have used the Xgboost algorithm which also works based on the decision-making trees. This algorithm has recently become popular dues to its advantages like fast, efficient, more accurate etc. the training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. It basically classifies the transaction in only two states that are either fraud or legitimate transactions.

### 4.1 Algorithm:

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

xgboost algorithm is based on gradient boosted decision trees. With the help of these decision trees, it classified the data as fraud or not. As it is based on a decision tree it can give us pretty good accuracy and also efficiency. This algorithm has some key features which are optimal results and high speed.

Some of the advantages of the XGBoost algorithm:

**a. Regularization:** XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why XGBoost is also called a regularized form of GBM (Gradient Boosting Machine).

While using the Scikit Learn library, we pass two hyper-parameters (alpha and lambda) to XGBoost related to regularization. alpha is used for L1 regularization and lambda is used for L2 regularization.

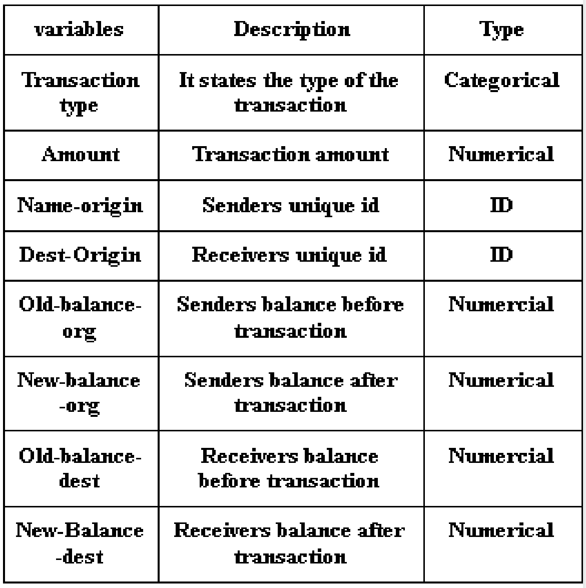
**b. Parallel Processing:** XGBoost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model.
While using Scikit Learn library, nthread hyper-parameter is used for parallel processing. nthread represents the number of CPU cores to be used. If you want to use all the available cores, don't mention any value for nthread and the algorithm will detect automatically.

**c. Handling Missing Values:** XGBoost has an in-built capability to handle missing values. When XGBoost encounters a missing value at a node, it tries both the left and right-hand split and learns the way leading to a higher loss for each node. It then does the same when working on the testing data.

**d. Cross-Validation:** XGBoost allows users to run cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we have to run a grid-search and only limited values can be tested.

**e. Effective Tree Pruning**: A GBM would stop splitting a node when it encounters a negative loss in the split. Thus, it is more of a greedy algorithm. XGBoost on the other hand makes splits up to the max_depth specified and then starts pruning the tree backwards and removing splits beyond which there is no positive gain.

***4.2 Dataset***: The dataset plays an important role in classifying the model. The dataset has been taken from the official Kaggle website, which is a well-informed data science organization. This dataset has details of millions of transactions out of which some of them are fraud transactions. This makes the development of the system more fluent and reliable. This dataset contains information on the rising risk of digital financial fraud, emphasizing the difficulty in obtaining such data. The main technical challenge it poses to predicting fraud is the highly imbalanced distribution between positive and negative classes in 6 million rows of data. The parameters of this dataset are Transaction type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest.



| variables | Description | Type |
|---|---|---|
| Transaction type | It states the type of the transaction | Categorical |
| Amount | Transaction amount | Numerical |
| Name-origin | Senders unique id | ID |
| Dest-Origin | Receivers unique id | ID |
| Old-balance-org | Senders balance before transaction | Numercial |
| New-balance -org | Senders balance after transaction | Numerical |
| Old-balance-dest | Receivers balance before transaction | Numercial |
| New-Balance -dest | Receivers balance after transaction | Numerical |

**Fig4.2-Dataset**

### *4.3 Data Flow Diagram* -

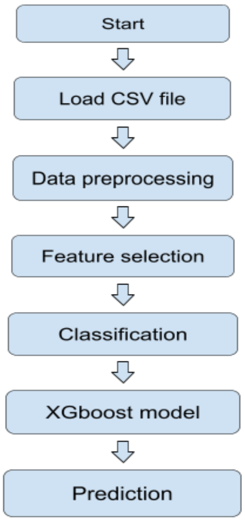


**Fig4.3 - Data Flow Diagram**

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various sub-processes the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships. It shows the complete procedure of how the data is first pre-processed and then prepared to be fed to the algorithm. Then the data is classified into fraud and legitimate transactions.

First, the dataset is fed to the model along with the libraries which are needed. Then the dataset which is being used is cleaned by removing the null values and other anomalies. Later the feature selection of the dataset is done where the input parameters are decided which will be given to the model and on its basis, the model has to do the prediction part. After that, the dataset is divided into 70% and 30%. Then the Xgboost model is trained on 70% data and tested on the remaining data.

## *5. HARDWARE & SOFTWARE USED*

### 5.1 Hardware: -

OS version: Windows 10 64-bit
Storage: 5 GB SSD
CPU: Intel Core i5-8400
Memory: 4 GB RAM

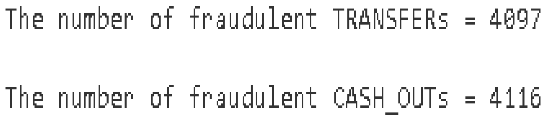
### 5.2 Software: -

IDE Used: Visual Studio Code, Jupyter, Google Chrome
Language: Python

## *6. Result –*

In this paper, we have proposed a model which can predict whether the transactions are fraud or not. We have used the XGBoost algorithm which has a lot of advantages over other algorithms. Xgboost (Extreme gradient boosting) is one of the well-known gradient boosting techniques having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. Which provides better results. Overall, in this data set, there were a total of 8213 transactions which were fraudulent transactions.

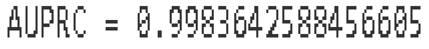
On the training dataset we have received an accuracy of 0.99



```
The number of fraudulent TRANSFERs = 4097

The number of fraudulent CASH_OUTs = 4116
```

**Fig 6.1-Classification result**



```
AUPRC = 0.9983642588456605
```

**Fig 6.2-Accuracy result**

## *Conclusion –*

This paper represents the development of a machine learning model to detect online fraud transactions using gradient boosting xgboost algorithm. The basic feature of this model is to classify the given dataset transactions as a fraudulent or genuine transaction. With the given dataset, this model has proved to result in better AUC score, accuracy score and efficient output. The dataset is preprocessed along with the feature selections, the data is then sent to classification into various factors before letting it to xgboost algorithm model. The final output is to obtain the transactions as true or fraudulent. This model can be then tested and trained with the larger data volume in future, so as to get more precise and accurate results. The model can also be upgraded to test dynamic datas in future for more advanced research.

## References:

1. K.Chaudhary, J.Yadav, "A review of fraud: A comparative study."decis. Support syst, vol 50, no3, pp.602-613,2011
2. Katherine J. Barker , Jackie D'Amato ,Paul Sheridon,2008 "Credit card fraud :awareness and prevention", Journal+- of financial Crime ,Vol. 15issue:4,pp.398-410
3. Dipti Thakur ,salamis Bhatia "distribution data Mining approach to credit card fraud detection" SPIT IEEE colloquium and international conference , volume4,48,issue2002.
4."CreditCard Fraud Detection Based on Transaction Be haviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.
5. Customer Transaction Fraud Detection Using Xgboost Model -by Yixuan Zhang, Ziyi Wang, Jialiang Tong, Fengqiang Gao June, 2020
6. Jerome H. Friedman. Greedy function approximation: a Gradient Boosting machine. The Annals of Statistics, 29(5):1189 – 1232, 2001.
7. Wang, M., Yu, J., & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model.
8. A. Mishra, C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.