IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detection of online payment fraud

Nimant Gupta

Department of Artificial intelligence and data science

Central University of Andhra Pradesh, Anantapur (Andhra Pradesh)

Nimant555@gmail.com

Dr P. Sumalatha

Department of Artificial Intelligence and data science

University of Central University of Andhra Pradesh, Anantapur (Andhra Pradesh)

Sumalatha.cuap@gmail.com

Abstract **- Big data is critical to many important industries, such as healthcare, finance, manufacturing, transportation, and e-commerce. Therefore, it is very important for the financial sector, especially banking services and the development of e-commerce. Due to advancements in communication and e-commerce, online payments are now the preferred method of payment for both offline and online purchases. The banking industry is facing many problems right now because of the use of online payment or credit cards, and the number of scammers is on the rise. The goal is to detect fraudulent transactions before they are processed and approved in order to prevent financial losses and protect the business and its customers. This method involves using machine learning algorithms to analyze transactions in real time, looking for patterns or anomalies that may indicate fraud.**

Index Terms **- Big data; Machine learning algorithm; Online Payment; credit card; Real time; E-commerce**

I. Introduction

Today, a variety of financial institutions and economic organisations use big data technology in their electronic commerce systems to assist their consumers in doing online transactions from any location and at any time. Systems that are used by both honest users and criminals fall into various categories including credit card systems, telecommunications systems, insurance systems, and online auction systems [3]. Unfortunately, as these method are employed more frequently, more frauds are produced every day, particularly in the banking industry. Then, to safeguard clients and businesses from these problems with electronic crime, fraud prevention systems (FPSs) are utilised.

However, banks and their clients now face greater problems and difficulties than in the past due to fraudsters' intellect and ability to adapt to these systems. The need to identify and detect fraudulent actions necessitates the deployment of fraud detection systems (FDSs), which are more pertinent and effective.

The development and efficiency of fraud detection systems are constrained by some large data issues [3]. Then, a variety of machine learning algorithms have been put forth by numerous researchers to improve low detection accuracy, accelerate the time to detection, and decrease false warnings.

However, due to fraudsters' cunning and their existence of severely unbalanced data sets is a significant obstacle to employing ML to fraud detection. The bulk of transactions are real, with very few being fraudulent, according to several databases that are available. Researchers face a huge problem in developing an accurate and effective fraud detection system that is low on false positives but successfully detects fraudulent behaviour.

II. Relevant Research

Due to the growing use of digital transactions and the need to shield people and organisations from fraud, online payment fraud detection is a crucial topic for study and development. Here are some pertinent articles about detecting online payment fraud:

According to Liang et al. (2016)'s study, "Detecting Online Payment Fraud Using Machine Learning Techniques," the fraud detection model put out in this research is built on machine learning techniques like logistic regression and decision trees.

Phua et al. (2017)'s "Fraud Detection for Online Banking Transactions": The authors investigate various machine learning methods for identifying fraudulent online banking transactions, such as random forests and support vector machines. They provide comparisons between the effectiveness of various algorithms and insights into the choice of features and model assessment.

Zhou et al. (2018), "Deep Learning-Based Fraud Detection Using Autoencoders and Generative Adversarial Networks": The application of deep learning techniques, particularly autoencoders and generative adversarial networks (GANs), for fraud detection is examined in this article. The authors show

how their method is effective in identifying fraud trends that had not previously been noticed.
By Kumar et al. (2019), in "Fraud Detection in Online Social Networks: A Survey": An overview of fraud detection methods designed exclusively for online social networks is provided in this survey article. It discusses numerous methodologies, emphasising their advantages and disadvantages. These methodologies covered include social network analysis, machine learning, and graph-based algorithms.
Singh et al.'s "Online Payment Fraud Detection Using Network-Based Features" (2020): The authors provide a technique for detecting fraud that makes use of network-based features. taken from statistics on online payments. They use actual money transactions to illustrate how well their strategy works.
By Chen et al. (2020), "Fraud Detection in E-commerce: A Data Mining Perspective": This paper focuses on detecting fraud in e-commerce environments and provides a thorough analysis of data mining methods used for detecting fraud. The authors go over various feature selection, fraud detection classification algorithms, and data pre-treatment techniques.

III. DATASET AND ANALYSIS

For our research, we used a Kaggle [8] dataset of simulated mobile-based payment transactions. Our dataset consists of 11 columns and 6362620 rows. It has a 0.13 percent fraud transaction rate, which is severely unbalanced. We can study this data by categorising it in accordance with the different types of transactions that it contains. The dataset contains five different types of transaction labels: "Cash In," "Cash Out," "Debit," "Transfer," and "Payment." Let's look into some dataset characteristics:

A step is equivalent to one hour.

A type is an online transaction.

The amount is the amount of the transaction.

NameOrig is the customer initiating the transaction.

Oldbalance.org is the balance prior to the transaction.

NewbalanceOrig is the balance following the transaction.

NameDest: The recipient of the transaction.

OldbalanceDest: The recipient's original balance before the transaction.

NewbalanceDest: The recipient's new balance following the transaction.

IsFraud: The transaction is fraudulent.

IsFlaggedFraud: The transaction is fraudulent.

Data of a single variable are analysed using a univariate approach. Here, we'll plot a histogram for analysis.

VI. VARIOUS ALGORITHM

There are lot of various technique are there to check the fraud one. Methods are Logistic Regression, Support vector machine (SVM), K nearest neighbour, Decision tree classifier, Random Forest classifier, naïve bayes and the last one is Extreme Gradient boosting classifier. we will see all method one by one after that we will take that method which give very good accuracy.

A. Logistic Regression

The purpose of binary classification tasks using the statistical modelling technique of logistic regression is to predict the likelihood of an event occurring or not. The dependent variable in logistic regression can only have one of two potential values, commonly written as 0 or 1, because it is binary or dichotomous. Both continuous and categorical independent variables, usually referred to as predictor variables or features, are acceptable. The relationship between the independent variables and the likelihood that an event will occur is estimated using the logistic regression model. The logistic regression model converts the linear combination of independent variables into a probability value between 0 and 1 by using the logistic function, commonly referred to as the sigmoid function.

B. Support Vector Machine

SVM method is employed in pattern recognition and classification. It is a strategy for categorising or predicting patterns into two groups: fraudulent or valid. Utilising this method for binary classifications.

SVM Error = Margin Error + Classification Error.

C. K Nearest neighbour

In the KNN procedure, we categorise any incoming transaction by finding the closest point to the new transaction. If the closest neighbour is fraudulent, the transaction will be flagged as fraudulent. K is used as a tiny and odd number (usually 1, 3, or 5) to break ties. In this Euclidean distance formula is used:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

D. Decision Tree classifier

A decision tree classifier is a machine learning method that recursively splits the data based on attribute values to produce a series of decision rules, which it then uses to create predictions using a tree-like structure. It can overfit and have trouble with intricate relationships, but it is interpretable and handles different attribute types.

E. Xtreme Gradient Descent Classifier

XGBoost is a distributed gradient boosting library that has been optimised for quick and scalable machine learning model training. A number of weak models' predictions are combined using this ensemble learning technique to get a stronger prediction. Extreme Gradient Boosting, or XGBoost, is one of the most well-known and widely used machine learning algorithms because it can handle large datasets and perform at the cutting edge in many machine learning tasks like classification and regression.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F}$$

F. Navey bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. Formula for bayes theorem is:

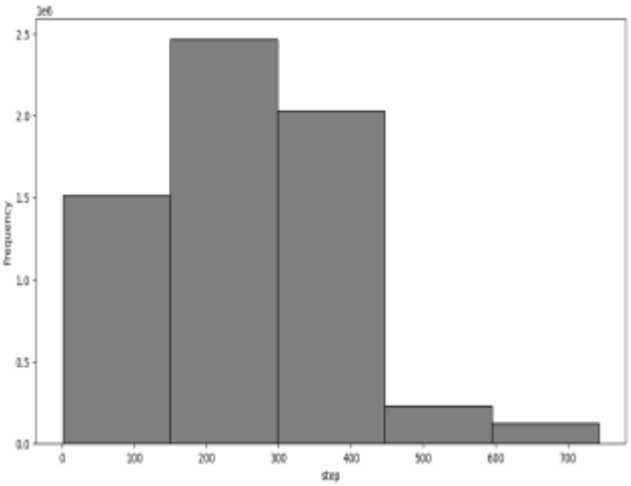
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

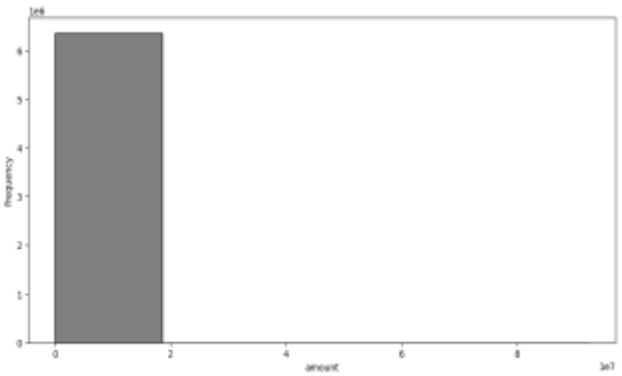
V. METHOD

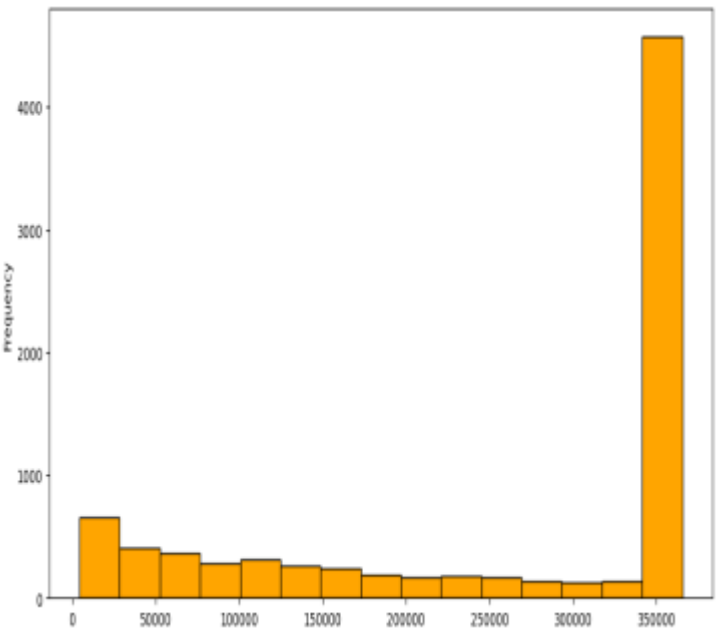
From all these Technique we will take XG Boosting technique because it gives best accuracy as compare to others technique. In this section, we find the outliers, describe our dataset split strategy and training, validation and testing processes that we have implemented in this work. All software was developed using Scikit-learn [7] ML library. We've found that fraud amount transaction ranges between 1.3-3.6 lakh. Now, we can see that among them most occurred were around 340,000-360,000 (3.4-3.6 lakh).

A. FIND THE OUTLIER

Now we will see the outlier with the help of ploting the histogram lets see:







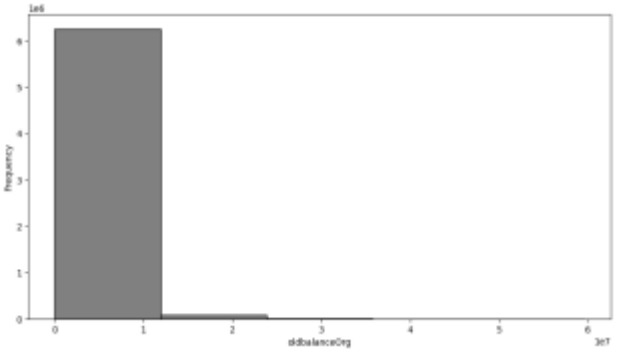
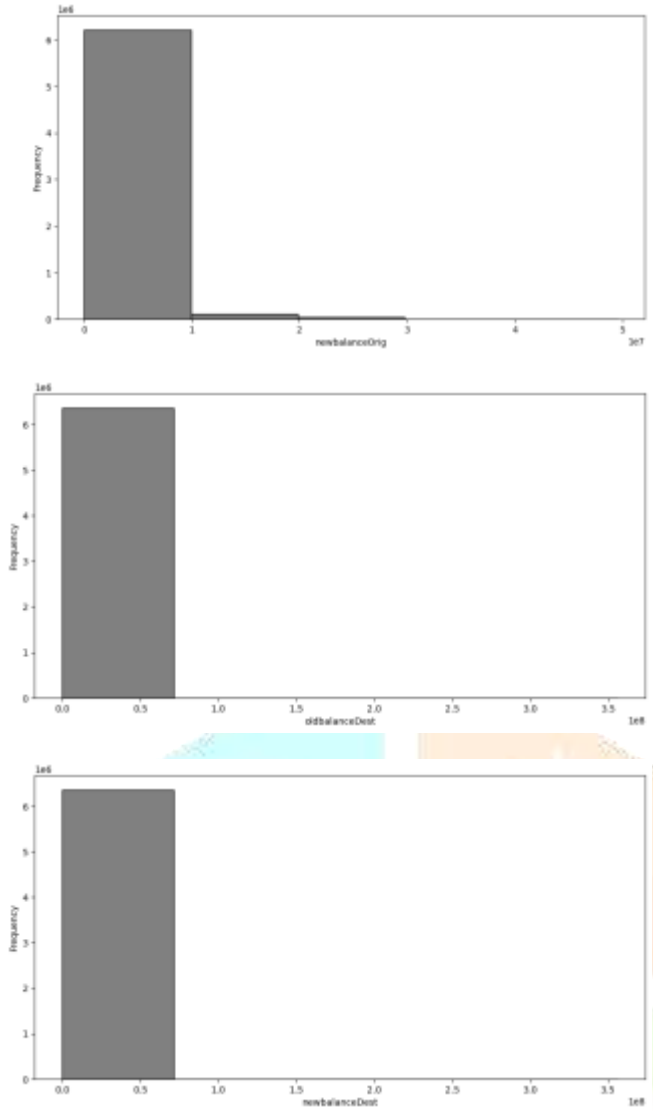


Fig 1. fraud amount transaction ranges between 1.3-3.6 lakh



When we plot the histogram with the help of univariate analysis. We could see some outliers on the plot. We'll use Quantile-based Flooring and Capping for these columns.

Capping is replacing all higher side values exceeding a certain theoretical maximum or upper control limit (UCL) by the UCL value. Here we'll do 90th percentile for higher values.
Flooring is replacing all values falling below a certain theoretical minimum or lower control limit (UCL) by the LCL value. Here we'll do 90th percentile for higher values.

With the help of these two techniques, we remove the Outlier. After outlier treatment:

- We have a maximum set of distribution between 150 to 400 of step
- Amount ranges between 0-35 lakhs with more values ranging between 0-75,000
- OldbalanceOrg ranges around 0-18 lakh with more frequency around 0-375,000
- NewbalanceOrig ranges around 0-19 lakh with more frequency around 0-375,000
- OldbalanceDest ranges around 0-29 lakh with common values around 0-625,000
- NewbalanceDest ranges around 0-35 lakh with common values around 0-625,000

Let's look at the relationship with type column with isfraud column:

Out[26]:

isFraud	0	1
type		
CASH_IN	1399284	0
CASH_OUT	2233384	4116
DEBIT	41432	0
PAYMENT	2151495	0
TRANSFER	528812	4097

B. *Split the dataset*

We divided our dataset into the numerous transaction categories that are given in the dataset section. We employ transfer and cash-out transactions exclusively for our tests because they contain fraudulent transactions. We divided the related datasets for each kind into three parts: training (70%), CV (15%), and testing (15%). We use stratified sampling to get train, CV, and test divides.

C. *Model Training and Testing*

We segregated our dataset based on the different transaction categories listed in the dataset section. We specifically employ TRANSFER and CASH OUT transactions for our tests since they contain fraudulent activities. For each kind, we divided the associated datasets into three groups: training (70%), CV (15%), and testing (15%).We use stratified sampling to get splits for train/CV/test.Thanks to stratified sampling, the proportion of each class in a split remains the same as it was in the original dataset.

VI. RESULT

Python is used to construct algorithms for fraud detection. Table 3 compares the performance. The performance study of Naive Bayes, Logistic Regression, Support Vector Machine, XG Boosting Technique, K Nearest Neighbour, Decision Tree, Random Forest Classification.

Table 1: Performance comparison

S. No	Method name	Training Accuracy	Testing Accuracy	Recall Score	Precision Score
1.	**Logistic Regression**	0.89	0.90	0.90	0.90
2.	**K nearest Neighbour**	0.98	0.98	0.98	0.98
3.	**Decision Tree**	0.94	0.96	0.95	0.97
4.	**Random Forest**	0.91	0.95	0.96	0.80
5.	**Extreme gradient Boosting**	0.99	0.99	0.99	0.99
6.	**Naïve Bayes**	0.77	0.78	0.78	0.78
7.	**Support vector machine**	0.95	0.95	0.95	0.95

In This best performance gives Extreme Gradient Boosting because its test, training, Recall, precision gives best accuracy as we want.

REFRENCE

[1] Samaneh Sorournejad, Zojah, Atani et.al - *"A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspectiv"e* - November 2016

[2] L. Nahar, I. Amir, and S. Shabnam, "A Comprehensive Survey of Fraud Detection Techniques," Int. J. Appl. Inf. Syst., vol. 10, no. 2, pp. 26–32, Dec. 2015.

[3] Wedge, Canter, Rubio et.al- *"Solving the False positives problem in fraud prediction using automated feature engineering"* - October 2017

[4] Rajani, Padmavathamma – "*A Model for Rule Based Fraud Detection in Telecommunications"* - IJERT – 2012

[5] I, Benchaji and S. Douzi, *"Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection*," p. 5.

[6] L. Nahar, I. Amir, and S. Shabnam, *"A Comprehensive Survey of Fraud Detection Techniques*," Int. J. Appl. Inf. Syst., vol. 10, no. 2, pp. 26–32, Dec. 2015.

[7] A. Abdallah, M. A. Maarof, and A. Zainal, *"Fraud detection system: A survey*," J. Netw. Comput. Appl., vol. 68, pp. 90–113, Jun. 2016.

[8] A. Kundu, S. Sural, and A. K. Majumdar, *"Two-Stage Credit Card Fraud Detection Using Sequence Alignment*," in Information Systems Security, vol. 4332, A. Bagchi and V. Atluri, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 260–275.

[9] K. T. Hafiz, S. Aghili, and P. Zavarsky, *"The use of predictive analytics technology to detect credit card fraud in Canada*," in 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), Gran Canaria, Spain, 2016, pp. 1–6.