Online Payment Fraud Detection

Abstract: - The rise of the internet and e-commerce appears to entail the usage of online payment transactions. The increased usage of online payments is leading to a rise in fraud. However, as the number of online transactions increases, so does the number of fraud instances. Fraud detection is an important component of online payment systems since it serves to protect both customers and merchants from financial damages. In this project, we propose a fraud detection system for online payments that uses machine learning techniques to identify and prevent fraudulent transactions. Using machine learning algorithms, we can find unique data patterns or uncommon data patterns that will be useful in detecting any fraudulent transactions. The Random Forest Classifier will be utilized to get the best results. Our approach strives to improve fraud detection accuracy while reducing the amount of false positives, resulting in a more efficient and effective method for identifying and combating fraud.

Keywords: - Fraud Detection, Machine Learning, Random Forest Algorithm, SVM, Classification, Data Pre-Processing, Prediction.

I.

Online payments have become a popular means of payment. Every second, thousands of deals are completed on the internet trading platform. Because of the prevalence of network transactions, certain criminals can perpetrate crimes. Personal property is vulnerable to theft in a complex network environment, which not only affects consumer interests but also jeopardizes the network economy's healthy growth.

In recent past, security has been taken into serious account as a crucial concern. Any form of approach may be used to find frauds in an online transaction. A lot of banks have increased their security measures as a result of the large credit card scams. All systems must get proper and secure authentication.

The Random Forest Classifier may be used to tackle classification issues in noisy and complicated environments. The Random Forest Classifier was crucial in the generalization of performance in a variety of machine learning issues, including face identification, web page categorization, and handwriting digital recognition. In applications using Random Forest Classifiers, overfitting is significantly less of an issue. The minimum problem and

curse of dimensionality seldom appear in Random Forest Classifier. The Random Forest Classifier is based on the notion of mathematical learning. Since a few days ago, the Random Forest Classifier has been employed in commercial applications including credit rating analysis, bankruptcy prediction, and time series prediction and classification. Random Forest Classifier is used here to estimate the data for fraud detection.

II. LITERATURE SURVEY

The Online Transaction Fraud Detection System has been the subject of several studies. This study is conducted before beginning the project in order to comprehend the numerous approaches that have been employed in the past. The advantages and disadvantages of the current system were highlighted by this study.

These consequences have been fuelled by the increased complexity of fraud efforts and the expanding variety of attack routes [1]. We focus on online and mobile price channels as well as identity theft fraud (i.e., taking a person's personal information to commit fraud) (Amiri and Hekmat 2021). The objective is to identify external scammers who want to trigger invoices for their hobbies. Fraudsters cannot be identified based on the account access technique since they get access to the price systems as if they were the proprietors of the bills. However, the fraudster behaves differently

during a payment transaction than the account owner, and/or the price has unusual characteristics, such as an oddly high charge amount or a switch to an account in a jurisdiction that doesn't match the buyer's lifestyle context and payment behaviour. On the basis of this supposition, algorithms may detect behavioural irregularities during payment transactions.

[2] In the master card fraud and detection tactics advised with the assistance of the notion, fraud is one of the most moral challenges in the master card sector. It aims to identify the many types of unique credit card fraud and, secondly, to look at the exchange techniques utilized in fraud detection. The secondary goal is to present, examine, and analyze recently reported consequences on the identification of master card fraud. In their essay, they define commonly used terms for master card fraud and emphasize significant facts and numbers about the incident. Depending on the type of fraud that banks encounter, different actions are also provided and put into action. Implementation of a suspicious scorecard on a current information set and its evaluation are often done as a subsequent step.

[3] Even after taking a number of efforts to avoid fraud, con artists continue to look for fresh ways and methods to defraud people. As a result, we would need a

strong fraud detection system to thwart these scams, one that not only finds the fraud but also finds it accurately and before it occurs. Additionally, we would like our systems to pick up lessons from earlier frauds and be able to modify themselves to deal with brand-new fraud techniques in the future.

[4] Risk assessment models or qualitative models for assessing fraud risk are commonly mentioned in the literature on fraud risk models. To our knowledge, no quantitative solution to fraud risk management incorporates the statistical impact of the fraud detection process into risk modelling. When we address procedural, legal, and organizational risk components, we quote the relevant literature. Montague's 2010 book focuses on preventing online payment fraud, although it skips over risk management and machine learning.

[5] In recent years, people have been increasingly concerned about security. There are various methods used to identify fraud in an online transaction. Many banks have improved their security mechanisms as a result of credit card theft. All systems must be equipped with secure authentication techniques. As soon as possible, fraud must be discovered. Therefore, fast security and authentication are required for system transactions.

III. METHODOLOGY

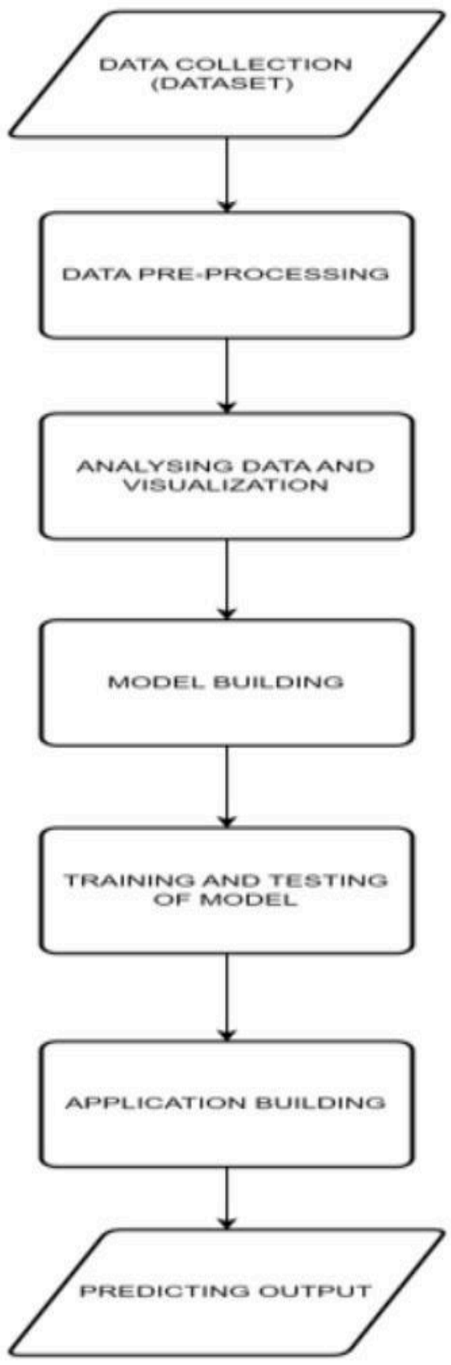

Fig 1 METHODOLOGY

A. *Data collection:*

The act of obtaining, acquiring, and combining the data that will be used to develop, test, and verify a machine learning model is known as data collection in machine learning. The fundamental phase in the machine learning pipeline is data collection for training the ML model. The accuracy of the predictions provided by machine learning (ML) systems is only as good as the training data.

B. *Data processing:*

Data processing in machine learning refers to the many processes and transformations carried out on data to make it appropriate for analysis and model training. Cleaning, preprocessing, and preparing the data for use in machine learning algorithms requires a number of procedures. The usefulness and performance of machine learning models can be significantly impacted by the quality of data processing. Before it can be used to develop models, the imported dataset has to be cleaned.

c. ***Analyzing Data and Visualization:***

Data analysis and visualization are iterative processes in machine learning, frequently requiring you to return and improve your knowledge of the data as you unearth new insights. Effective visualization and analysis aid in improved model selection, feature engineering, and the dissemination of findings to non-technical stakeholders. To ensure openness and confidence in the model's predictions, it is a crucial phase in the machine learning workflow.

D. ***Model construction:***

It entails the creation and application of prediction models to spot and stop fraudulent activity in real-time or almost real-time during online financial transactions. This is a crucial part of contemporary payment systems where fraudulent transactions are automatically detected and handled using machine learning techniques.

E. ***Model Training and Testing:***

A subset of the dataset called training data is used to train the machine learning model to identify patterns and make predictions. It is made up of input characteristics and the labels or goal values that correspond to the output features. Test data is a different subset of the dataset that is used to gauge the model's effectiveness and determine its capacity to make predictions on brand-new, unforeseen data.

F. ***Application Development:***

The process of creating software systems or applications that use machine learning models to address particular problems in the real world is known in the field of machine learning as "application building." These apps use the predictive power of machine learning models to make decisions, recommend actions, or automate operations.

G. *Predicting Output:*

Predicting output is the process of utilizing a trained model to create predictions or forecasts based on input data. These predictions, which are basically the expected outcomes or answers produced by the model, might take a variety of shapes depending on the type of problem you're attempting to address.

IV. ARCHITECTURE

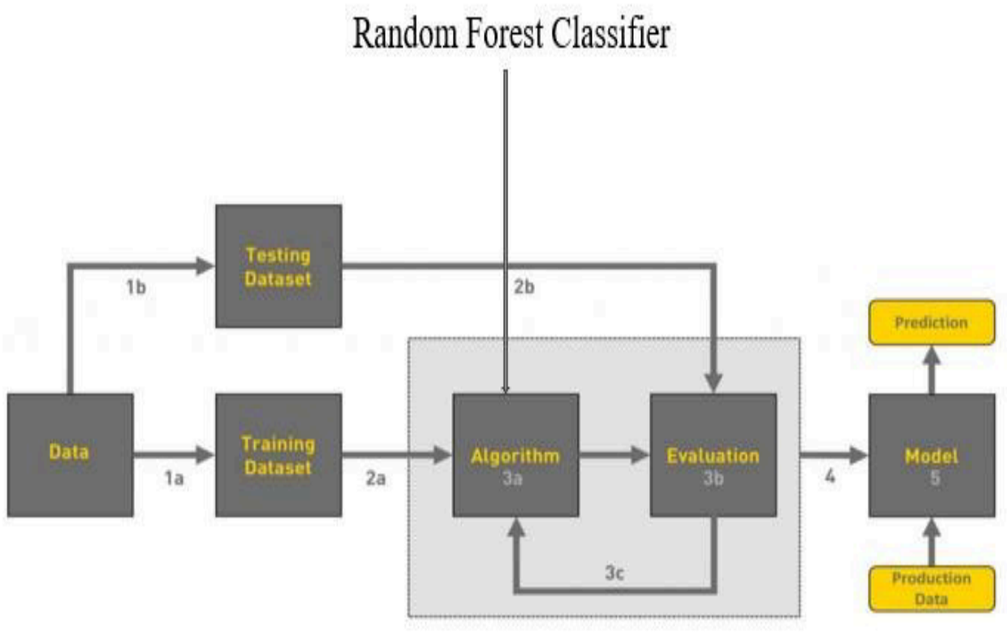

Fig 2 Random Forest Classifier

The purpose of this project report is to present the design, implementation, and evaluation of an online payment fraud detection system using machine learning. The main objectives of this project are:

- **To analyze and understand the various types of fraud**

that occur in online payment systems.

- To propose and design a machine learning-based fraud detection system that can accurately identify and prevent fraudulent transactions.
- To implement the proposed system and evaluate its performance using real-world data

To find the accuracy we can use the Random Forest Classifier Algorithm (99.6%)

V. SYSTEM IMPLEMENTATION

A. *Importing the libraries:*

Import the necessary libraries as shown in the image.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.svm import SVC
import xgboost as xgb
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report, confusion_matrix
import warnings
import pickle
```

Fig 3 Importing modules and libraries

B. *Read the dataset:*

Our dataset format might be in .csv, excel files, .txt,. Json, etc. We can read the dataset with the help of pandas.

```
# Reading the csv data
df = pd.read_csv(r'C:\Users\user\Desktop\PS_20174392719_1491204439457_logs.csv')
```

```
df
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	0	0
2	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	0	0
3	1	PAYMENT	7817.71	C90045638	53860.00	46042.29	M573487274	0.00	0.00	0	0
4	1	PAYMENT	7107.77	C154988899	183195.00	176087.23	M408069119	0.00	0.00	0	0
...	...	...	...	...	...	...	...	...	...	...	...
2425	95	CASH_OUT	56745.14	C526144262	56745.14	0.00	C79051264	51433.88	108179.02	1	0
2426	95	TRANSFER	33676.59	C732111322	33676.59	0.00	C1140210295	0.00	0.00	1	0
2427	95	CASH_OUT	33676.59	C1000086512	33676.59	0.00	C1759363094	0.00	33676.59	1	0
2428	95	TRANSFER	87999.25	C927181710	87999.25	0.00	C757947873	0.00	0.00	1	0
2429	95	CASH_OUT	87999.25	C409531429	87999.25	0.00	C1827219533	0.00	87999.25	1	0

2430 rows × 11 columns

Fig 4 Reading Data set

c. ***Data preprocessing:***

The df.isnull() method is used to verify that no values are present. We employ the sum () function to add up those null values. Two null values were discovered in our dataset, we discovered. We thus start by investigating the data.

d. ***Using A Heat Map to Check The Correlation***

I'm using a heat map to check the correlation in this instance. Using different colour combinations, it displays the data as 2-D coloured maps. Instead of numbers, it will be plotted on both axes to describe the relationship variables.

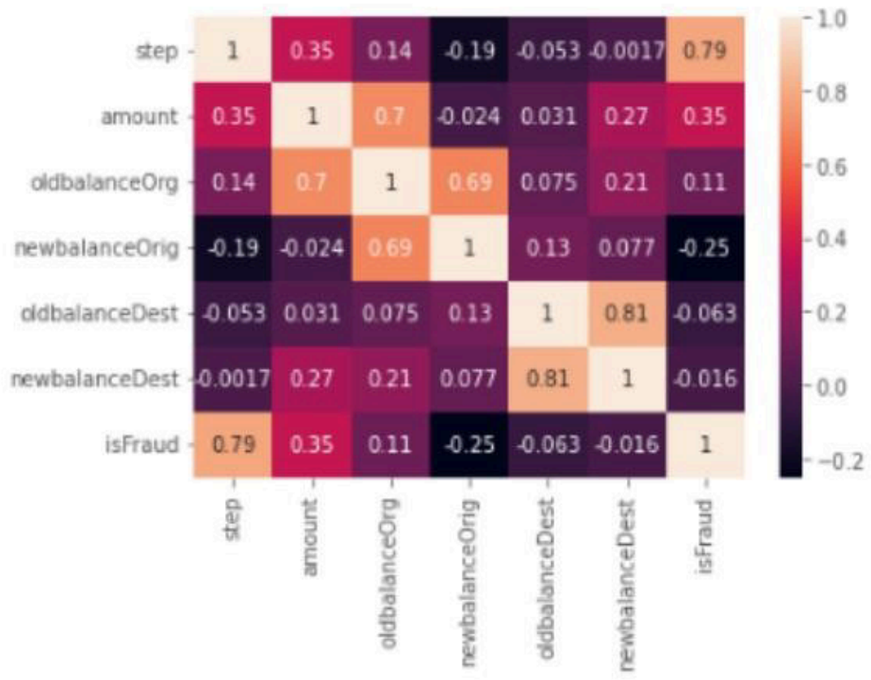

Fig 5 Heat map

e. ***Feature selection:***

Using a variety of machine learning algorithms, including Random Forest, KNN, Decision Tree, etc., I have discovered a number of metrics in this case. For the testing dataset, we are receiving the random forest model's best accuracy here.

f. ***Distribution Plots Are Now Plotted to Examine the Distribution in Numerical Data:***

Plotting the displot requires the usage of the seaborn.displot() method. The displot returns the plot 12 with the density distribution and

accepts the univariate distribution of the data variable as an input. I used distribution(displot) on the 'amount' column in this case.

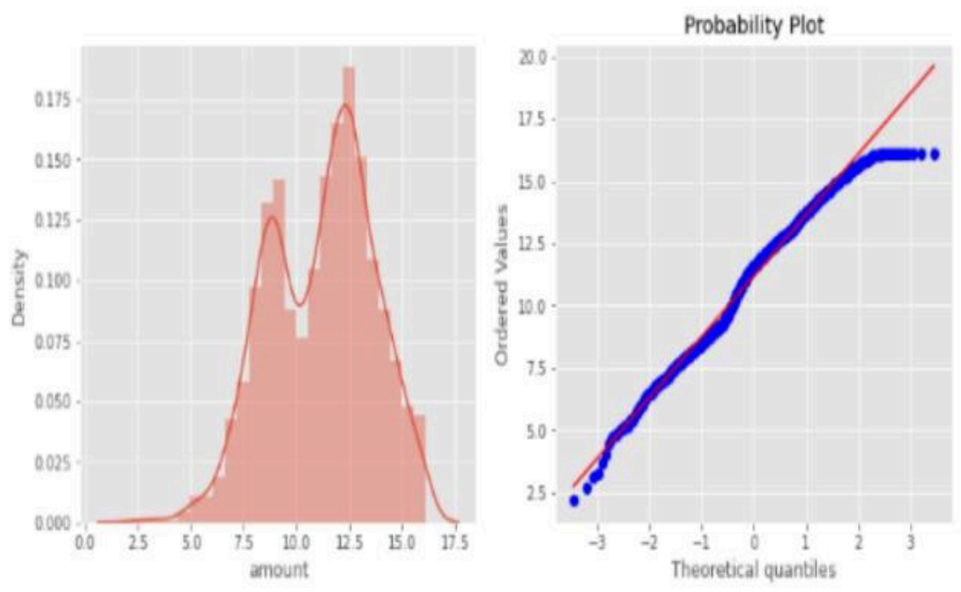

Fig 6 Distribution

G. *Converting to .pkl file:*

Now we need to convert the file to pickle file and save the model as shown below.

```
In [79]: import pickle
         pickle.dump(rfc,open('paytm.pkl','wb'))
```

Fig 7 converting to .pkl file

H. *APPLICATION BUILDING:*

- Building HTML and CSS pages
- Build python code

VI. PREREQUISITES

Several requirements must be taken into account in order to create an efficient machine learning system for detecting online payment fraud. These requirements cover data, technology, subject-matter expertise, and regulatory considerations. These are the essential requirements:

- Access to a large database of earlier internet payments, containing both honest and dishonest cas.es. The information must be accurate, well-organized, and representative of actual situations.
- Machine learning models must be trained and evaluated using data that has been clearly labelled and in which each transaction is

classified as either legal or fraudulent.

- Strict adherence to payment card industry standards (PCI DSS) and data protection laws (such as GDPR) when handling sensitive payment information. Implement tight access controls and data encryption.
- proficiency in a programming language, such as Python or R, for the creation and implementation of models for data analysis
- Considering that fraudulent transactions are often rare compared to legitimate ones, there are strategies for dealing with the class imbalance in the dataset.

VII. LIMITATIONS

There is little doubt that the following five restrictions apply to machine learning-based efforts to identify online fraud:

A. *Imbalanced Data*

Sets showing a mismatch between legitimate and fraudulent transactions may provide problems. As a result, it may be challenging to successfully detect fraud since models may become biased in favour of the dominant class.

B. *Adaptive Fraud Techniques*

Static machine learning algorithms find it challenging to keep up with fraudsters'

rapidly changing and developing techniques. New fraud patterns that the models haven't been taught to recognize may start to appear.

c. *Data Quality*

It is essential that the training data be of a high standard. A model's predictions may be inaccurate if there are gaps in the data or if it contains errors. Clear up the trash you made.

d. *Privacy Concerns*

Online fraud detection frequently entails the use of sensitive client data. Ensuring data privacy and compliance with rules such as GDPR can be difficult to achieve when adopting machine learning systems.

e. *False Positives*

Overly active fraud detection might result in a large number of false positives, which can annoy legitimate users and negatively impair the user experience. Balancing fraud detection with a low false positive rate is a difficult challenge.

It is critical to be aware of these limitations and to continually modify and enhance machine learning models and tactics for online fraud detection.

VIII. FUTURE SCOPE

The future of online payment fraud detection using machine learning seems promising, with various developing trends and advances on the horizon. Here are some important areas of

future opportunity for online payment fraud detection:

A. ***Advanced Machine Learning Models:***

Continued development of increasingly advanced machine learning models, including deep learning and reinforcement learning, to increase the accuracy and flexibility of fraud detection systems.

B. ***Explainable AI (XAI):***

Addressing the interpretability of machine learning models to make them more transparent and intelligible, enabling for better decision-making and regulatory compliance.

C. ***Defence Against Adversarial Machine Learning:***

Defending against adversarial assaults on machine learning models used in fraud detection.

D. ***Model Updating on a Constant Basis:***

Implementing systems capable of constantly updating and adapting machine learning models when new fraud tendencies arise.

IX. RESULT

Machine Learning was used to build Online Payment Fraud Detection. Because the fraud (categorical values) is the target or dependent variable, this is a classification problem. The goal of online payment fraud is to divide the available supply of usable online payments into groups that differ in quality.

To launch the application, follow these steps:

- From the start menu, launch the anaconda prompt.
- Open the folder containing your Python script.
- Now enter the command "python app.py"
- Go to the localhost to view your web page.
- Fill in the blanks, then click the submit button to view the outcome/prediction.

X.

This Project describes the creation of a machine learning model for detecting online fraud transactions using the Random Forest technique. The fundamental function of this model is to categorize the supplied dataset transactions as fraudulent or authentic. With the supplied dataset, this model produced a superior AUC score, accuracy score, and efficient output. The dataset is preprocessed together with the feature choices, and the data is then classified into several variables before being delivered to the Random Forest method model. The ultimate result is to determine if the transactions are genuine or fraudulent. This model may then be tested and trained with greater data volumes in the future to provide more precise and accurate findings.

We may argue that Random Forest is a superior tool for detecting fraud than any other

algorithm. Random forest has 99% accuracy in Python programming and 93% accuracy in R programming. As a result, it is regarded as the best approach.