# CHAPTER 1

# INTRODUCTION

## PROBLEM STATEMENT

In today's mobile-dependent society, individuals often find themselves in situations, where their mobile phones run out of battery while they are away from home or office. This poses a significant inconvenience as people rely heavily on their smartphones for communication, information, and various daily tasks. While carrying chargers is an option, individuals may not always find accessible power sockets in public places even if they have their chargers with them. Therefore, there is a need for a project that addresses this issue by developing a smart mobile charging system for public places.

Online payment fraud detection using machine learning relies on various features and techniques to identify potentially fraudulent transactions. The choice of features plays a crucial role in the effectiveness of the detection system. Here are key features commonly used in such systems:

## PROBLEM OBJECTIVES

The primary objective of the online payment fraud detection project using machine learning is to develop a robust and efficient system that enhances the security of online payment transactions. The overarching goal is to create a sophisticated fraud detection mechanism capable of identifying and preventing unauthorized activities in real-time. The project aims to strike a crucial balance by minimizing false positives, ensuring that legitimate transactions are not erroneously flagged as fraudulent, thereby maintaining a seamless and positive user experience. The system will be designed to integrate seamlessly with existing online payment platforms, providing a continuous, real-time analysis of transactions. Additionally, the project aims to optimize fraud detection accuracy through careful selection and training of machine learning models, and to ensure adaptability to emerging fraud patterns. Continuous model improvement processes will be established to refine the system's capabilities based on feedback

from detected fraud cases. Scalability and compliance with regulatory standards are also key objectives, ensuring that the system can handle increasing transaction volumes while adhering to legal and ethical guidelines. Documentation, reporting, and a user-friendly experience are

prioritized aspects to provide transparency, accountability, and an effective defense against the evolving landscape of online payment fraud.

* 1. **PROJECT OVERVIEW**
* **Transaction Amount:** Unusual spikes or irregular transaction amounts may indicate fraudulent activity.
* **Transaction Frequency:** Rapid, unexpected changes in transaction frequency might signal fraudulent behavior.
* **Geolocation:** Analyzing the geographic location of transactions helps detect anomalies, especially if transactions occur in multiple distant locations simultaneously.
* **Device Information:** Examining the device used for the transaction, including device type, operating system, and IP address, can reveal irregularities.
* **User Behavior Analytics:** Establishing a baseline of typical user behavior helps identify deviations from normal patterns.
* **Time of Transaction:** Unusual transaction times or patterns, such as transactions occurring at odd hours, can be indicative of fraud.
* **Merchant and Merchant Category Code (MCC):** Monitoring the types of merchants involved in transactions and their corresponding MCC helps identify suspicious activities.
* **Card Verification Value (CVV) and Expiry Date:** Validating the CVV and expiry date adds an extra layer of security and helps detect potential fraud,
* **User Authentication Information:** Analyzing the success or failure of user authentication attempts can be crucial in detecting unauthorized access.
* **Social Network Analysis:** Examining relationships between users and their network can reveal abnormal transaction patterns.
* **Email Address and User Information:** Analyzing the validity and consistency of email addresses and user information helps in identifying suspicious accounts.
* **Velocity Checks:** Monitoring the rate of transactions within a specific time frame helps detect rapid, abnormal activity.
* **Machine Learning Predictions**: Predictions from machine learning models, based on historical data, are fundamental features in fraud detection.
* **Transaction Correlation**: Examining correlations between different transactions, especially across multiple accounts, can reveal coordinated fraud attempts.

By combining and analyzing these features, a machine learning model can learn to distinguish between legitimate and fraudulent transactions, contributing to the effectiveness of an online payment fraud detection system. The specific set of features used may vary depending on the characteristics of the data and the chosen machine learning algorithms.

# 

# CHAPTER – 2

# LITERATURE SURVEY

## REVIEW OF LITERATURE

## As there is a huge increase in online transactions, the fraud related to online transactions has also increased, due to this many companies and researchers are working towards building the best model for the detection of fraud in online transactions. This encourages researchers like us to work towards finding the best-suited model for online payment fraud detection.

## According to research done by Noor Saleh Alfaiz et al, [2] [3] they have proposed K-Nearest Neighbours (KNN) under- sampling technique along with the Categorical boosting [Cat Boost] algorithm as the best model for credit card fraud detection. Olawale Adepoju et al [4] [5] [6] research paper has considered exceptionally distorted data on credit card fraud and has concluded that the Logistic Regression algorithm is the best-suited algorithm for credit card fraud detection.

## Mosa M. M. Megdad et al [7] [8] research paper has concluded that the Random Forest classifier is the best classifier for an unbalanced dataset and the Bagging classifier is the best classifier for a balanced dataset [9] [10].

## In the paper written by Emmanuel Ileberi, we get to know the implementation of ML frameworks and the use of the Synthetic Minority over-sampling Technique (SMOTE) [11] [12] to solve the problem of class imbalance. The outcome of this research was that AdaBoost had a greater impact on ML methods and yielded better results [12].

## “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.

# CHAPTER - 3

# METHODOLOGY

## BLOCK DIAGRAM

## 

Fig 3.1.1: Block Diagram

## When it comes to finding fraudulent online payment transactions, data analysis is crucial. Banks and other financial institutions can adapt the required defences against these frauds with the aid of machine learning techniques. Many businesses and organizations are investing a lot of money in the development of these machine learning systems to determine whether a specific transaction is fraudulent. Machine learning techniques assist these organizations in identifying frauds and preventing their clients who may be at risk for such frauds and occasionally sustain losses as a result. The research’s data set came from the open platform ”kaggle.” Due to privacy concerns, it is challenging to obtain real-time data sets; therefore, a data collection big enough to conduct the research was taken. The data set has 1048576 records and 11 columns. This data set includes attributes like type (type of payment), amount, ”nameOrig” (customer initiating the transaction), ”pldbalance- Org” (balance before the transaction), ”newbalanceOrig” (balance after the transaction), ”nameDest” (recipient of the transaction), ”pldbalanceDest” (initial

## recipient balance prior to the transaction), ”newbalanceDest” (the new balance recipient after the transaction), and isFraud which (0 if the transaction is legitimate and 1 if the transaction is fraudu- lent). The figure 3.1.2. shows all the features

## 

## 

## 

## 

## 

## Fig 3.1.2: Dataset

## in the dataset. Whether a particular transaction is fraudulent or not depends highly on the type of the trasaction, figure 3.1.3 below shows the types of transactions and the percentage of the same in our dataset.

## 

## Fig 3.1.3: Distribution of Transaction Type

## Analysing Missing values:

## Before using the data in the model, it is important to pre-process the data downloaded from the dataset. The next step is to check for any missing values in our dataset. It can be seen in the figure 3.1.4. that there are no missing values in our dataset.

## 

## 

## Fig 3.1.4: Missing values

## Investigating the Correlation between the Features:

## Even though our data has a large number of features, not all of them are contributing to our target feature. Figure 3.1.6 shows how the features are related to each other. With the help of the correlation matrix, we have narrowed down the list of important features that can help us predict our target feature.However, we will again train the models including all the features which were rejected earlier to compare the results of the models.

## 

## Fig 3.1.6: Correlation Matrix

## Data Preparation:

## For the machine learning model to give accurate, high-quality results, the data used to train and test it should be well-prepared. One of the most important steps in data mining is getting the data ready. There are many things to consider, such as how to deal with missing data, duplicate values, removing redundant features from data using correlation matrix and feature selection methods, how to deal with the fact that data isn’t balanced, etc. The quality of the techniques used to prepare the data has a lot to do with how well machine learning works. If the data is not prepared well, it could take a long time to run the models and cost a lot of money. Because of all of these things, the most difficult and time-consuming part of the data mining process is getting the data ready.

## Feature Selection

## Feature selection is one of the approaches that helps models perform even better after data cleansing and feature correlation analysis. This method is used to eliminate unnecessary variables, which leads to a smaller feature space and could improve the performance of the model. In our dataset two features ”namedest” and ”nameOrig” were of less significance as compared to other features, however to compare the same we will be running the models without these features and then including these two features.

## Modelling Approach

## Modelling is a very important aspect in machine learning. After the final data preparation, which includes steps like handling the class imbalance and feature selection, the proposed models are implemented on the processed or prepared data. The detailed explanation and working of the proposed models are discussed in this section:

## Logistic Regression

## Logistic Regression is the classification of algorithm into multiple categorical values. It includes the use of multiple independent variables which are used to predict a particular outcome of a variable which is dependent on all the independent variables use to train the model. Logistic Regression is similar to linear regression, it predicts a target field rather than a numeric

## one Zanin et al. (2018). Like predicting True or False, successful or unsuccessful in our case it is fraudulent or non fraudulent. The figure below explains the logistic regression:

## 

## Fig 3.1.7: Logistic Regression

## Random Forest Classifier

## The random forest model is made up of many decision trees that are all put together to solve classification problems. It uses methods like feature randomization and bagging to build each tree. This makes a forest of trees that don’t have anything in common with each other. Every tree in the forest is based on a basic training sample, and the number of trees in the forest has a

## direct impact on the results.Bahnsen et al. (2016) Tsest

## 

## 

## Fig 3.1.8: Random Forest Algorithm

## Decision Tree

## Decision tree is a supervised machine learning algorithm which uses a combination of rules to make a particular decision, just like a human being. The motive behind decision tree is that one uses the dataset features to create yes or no questions and split the dataset until and unless we isolate all the datapoints those belong to each class.Choi and Lee (2017)

## 

## Fig 3.1.9: Decision Tree Algorithm

## FLOW CHART

## 

## 

## Fig 3.2.1: Flow Chart

## 3.3 ALGORIHTM

## Step 1: Acquired a huge number of data from Kaggle and loaded it to the system.

## Step 2: Check if there are any duplicates, missing values. If so, remove them.

## Step 3: Split the data into training set for teaching the machine and testing set for checking how well the machine has learnt

## Step 3: The raw data should be processed before applying to ML algorithms for which we have

## used categorical encoding and scaling techniques

## Step 4: train the data using respective algorithms

## Step 5: predict the results for the test data

## 

## Step 6: Check the correctness of the result using different performance metrics.

# 

# 

# 

# CHAPTER – 4

# IMPLEMENTATION

## SOFTWARE REQUIREMENTS

## Programming Language :  Python

## Machine Learning Libraries :  Scikit-Learn

## Data Processing and Analysis: Numpy

## Data Manipulation and Visualization Tools : Matplotlib, Seaborn

## Code Development Environment :  Jupyter Notebook

## 4.2 HARDWARE REQUIREMENT

## Central Processing Unit (CPU):

## A multi-core processor with a high clock speed is beneficial for training machine learning models. More cores can handle parallel processing, which is useful for tasks like feature extraction and model training.

## Random Access Memory (RAM):

## Sufficient RAM is crucial, especially when working with large datasets or complex models. A minimum of 16 GB is recommended for moderate-sized datasets and models. For larger datasets or more complex models, consider 32 GB or more.

## Graphics Processing Unit (GPU):

## GPUs are essential for accelerating the training of deep learning models. If using deep neural networks (e.g., convolutional neural networks or recurrent neural networks), consider a high-end GPU. NVIDIA GPUs (such as those in the GeForce, Quadro, or Tesla series) are commonly used for machine learning tasks.

## Storage:

## Fast storage is important for reading and writing large datasets efficiently. Consider using Solid State Drives (SSDs) for better performance compared to Hard Disk Drives (HDDs). The storage capacity depends on the size of your datasets, but a minimum of 512 GB SSD is recommended.

## Networking:

## A fast and reliable internet connection is essential, especially if the system involves real-time processing of transactions. Low-latency networking is crucial for timely fraud detection.

## Dedicated Server or Cloud Services:

## Depending on the scale of your application, you may choose to use a dedicated server or cloud services. Cloud platforms such as AWS, Google Cloud, or Azure provide scalable resources, allowing you to adjust the infrastructure based on the workload.

## Additional Considerations:

## Parallel Processing:

## If your machine learning framework supports it, consider hardware that supports parallel processing, such as multi-GPU setups or distributed computing.

## Cooling:

## Intensive machine learning tasks generate heat. Ensure that the hardware setup has proper cooling solutions to prevent overheating.

## Scalability

## Design the system with scalability in mind. If the volume of transactions is expected to increase, plan for a scalable infrastructure that can handle the growth.

## Redundancy:

## Implement redundancy and backup systems to ensure continuity in case of hardware failures.

# 

# CHAPTER – 5

# OBSERVATIONS

# 5.1 EXPERIMENTAL SETUP

In online payment fraud detection using machine learning, the experimental setup involves the orchestration of various components to create a robust and accurate system. The process typically begins with the collection of a diverse and representative dataset comprising legitimate and fraudulent transactions. This dataset serves as the foundation for training and evaluating machine learning models. Features extracted from transactional data, such as transaction amount, frequency, location, and user behavior patterns, are crucial for model learning. The dataset is then divided into training, validation, and test sets to ensure the model's generalizability. Various machine learning algorithms, such as supervised learning classifiers (e.g., decision trees, support vector machines, or neural networks

), are applied to the training set to learn patterns indicative of fraudulent behavior. Hyperparameter tuning and cross-validation techniques are often employed to optimize model performance. Additionally, anomaly detection methods may be integrated to identify outliers in transaction patterns. The experimental setup also includes real-time monitoring components to assess the model's performance in a dynamic environment. Continuous feedback loops and model retraining are implemented to adapt to evolving fraud patterns and maintain a high level of accuracy in detecting fraudulent activities in online payment systems. Overall, a well-designed experimental setup is crucial to the success of machine learning-based online payment fraud detection systems, ensuring they remain effective and adaptive in the face of emerging threats.

# 5.2 PARAMETERS WITH FORMULAS

# The main goal of this research is to use supervised machine learning techniques and Artificial Neural Networks together and see if our proposed method improves the model’s performance more than other state-of-the-art methods.For the comparative study, we did the following two experiments: [1] Figuring out how well a model works by training it on all of the features in the dataset. [2] Figuring out how well the model works by training it on certain features (Eliminating namedest and name orig features).We chose metrics like recall, specificity, F1-score, AUC score, AUC-ROC curve, and the geometric mean of recall and specificity so that we could compare how well our models worked.Because our data isn’t balanced, we can’t judge how well models work based on how accurate they are.Rambola et al. (2018) also compares the results based on the confusion matrix as in this case it is better to compare the True Positive and True negatives and decide based on that that how much accuracy is achieved in our models.The confusion matrix is explained below: • True Positive(TP):It shows that the given model has done a good job of figuring out non-fraudulent cases as non-fraud (positive). • False Positive(FP): It shows that the model didn’t get the prediction right, fraudulent cases were identified as non-fraud (positive). • False Negative(FN):It shows that the model didn’t get the prediction right, nonfraudulent cases were identified as fraudulent (negative). • True Negative(TN):It shows that the model has been able to accurately predict fraudulent cases as fraudulent (negative). Precision and specificity show the number of transactions that are considered to be fraud and are frauds. On the other hand, recall/sensitivity values show what percentage of real fraud transactions are correctly classified.F1-score is the average of the notes between Precision and Recall and for better classification, should be close to 1 Kalbande et al. (2021).The geometric mean is the sum of both specificity and sensitivity. We chose this metric to judge our models because it works well with unbalanced data Bahnsen et al (2016). Due to how unbalanced our data is, the most important evaluation metric in our research is the recall and AUC score. We’ve used the confusion matrix to figure out the above evaluation metrics. the equations that follow:- After using the different data

# 

# preparation methods talked about in the above sections, we now have our final dataset to use with our chosen models.To see how well our proposed method of using undersampling for handling Imbalance in data works and eliminating two fetaures(namedest and nameorig),We have done two different case studies for each of the algorithms and used the above-discussed evaluation metrics to rate them.

# 

# 

# CHAPTER – 6

# DISCUSSION OF RESULT

# 6.1 RESULTS

# Considering the results of all the above mentioned supervised machine learning algorithms we came to know Random forest algorithm is the best suited algorithm for the detection of online transaction fraud with accuracy of 99.994%, precision score of 0.9548, recall score of 0.5075, Log-loss value of 0.018884, F1 score of 0.66274 as shown in the below TABLE 6.1. As we know, lower the Log-loss value higher is the performance of the algorithm. The lowest value of Log- loss can be observed in Random Forest as 0.01888and with respect to accuracy, Decision tree and Random Forest algorithms we observe that both yields similar percentage. However, looking into other performance metrices we conclude that Random Forest exhibits superior results.

# 

# 

# Table6.1Accuracy,precision,recall,f1 score and log loss when different algorithms were used.

# 

# Fig 6.1.1: Plot Accuracy Score

# 

# 

# Fig 6.1.2 Plot of F1 Score

# 

# Fig 6.1.3 Plot of Precision Score

# 

Fig 6.1.4: Plot of Recall Score

# 

# 

# Fig 6.1.5 Plot of Logg Loss

# 

# Table 6.1.2: Confusion Matrix of Random Forest Algorithm

# 

# 

# Fig 6.1.6: Accuracy of Decision Tree Algorithm

# 

# 

# 

# 

# Fig 6.1.7: Accuracy of Random Forest Algorithm

# 

# 

# 

# 

# Fig 6.1.8: Result

# CHAPTER – 7

# CONCLUSIONS

# 7.1 JUSTIFICATIOS

**1. Advanced Threat Detection**: The online landscape is constantly evolving, and traditional rule-based systems may struggle to keep pace with sophisticated and rapidly changing fraud tactics. Machine learning algorithms can adapt and learn from patterns in data, enabling the detection of subtle, complex fraud patterns that may go unnoticed by rule-based systems.

**2. Real-Time Analysis and Response:** Machine learning facilitates real-time analysis of transactions as they occur. This capability is crucial for quickly identifying and responding to potential fraud, minimizing the impact on both businesses and users. Rapid response times can significantly reduce the financial losses associated with fraudulent transactions.

**3.** **Minimization of False Positives:** Traditional fraud detection systems often suffer from high false positive rates, leading to legitimate transactions being mistakenly flagged as fraudulent. Machine learning models can be fine-tuned to optimize accuracy, reducing false positives and ensuring a smoother and more trustworthy user experience. This is particularly important in the context of online payments, where user convenience is paramount.

**4. Adaptability to Dynamic Fraud Patterns:** Fraudsters continually evolve their tactics to bypass security measures. Machine learning models excel at adapting to changing patterns by continuously learning from new data. This adaptability allows the fraud detection system to remain effective over time, even as fraud techniques become more sophisticated.

**5**. **Scalability and Efficiency:** As online transaction volumes increase, scalability becomes crucial. Machine learning algorithms, when properly implemented, can efficiently process large amounts of data and scale to meet the demands of growing transaction volumes. This ensures that the fraud detection system remains effective and responsive even in high-traffic scenarios.

**7.2 PARAMETERS IMPROVED**

**1.Feature Selection and Engineering:** Careful selection and engineering of features such as transaction amount, frequency, geolocation, and user behaviour contribute significantly to the performance of machine learning models. Continuous refinement of feature sets based on their relevance to fraud patterns can enhance detection capabilities.

**2.Algorithm Selection:** Choosing the right machine learning algorithms is crucial. Experimenting with different algorithms, including decision trees, random forests, support vector machines, and neural networks, allows for the identification of the most suitable model for the specific characteristics of the fraud detection problem

**3.Continuous Model Training and Updating:** Implementing mechanisms for continuous model training and updating allows the system to adapt to evolving fraud patterns. Regularly retraining the model with new data ensures that it remains effective over time.

**4.Anomaly Detection Techniques**: Incorporating advanced anomaly detection techniques, such as isolation forests or one-class SVMs, can enhance the ability of the system to identify rare and previously unseen fraudulent patterns.

**5.Data Quality and Preprocessing:** Ensuring the quality of input data and implementing robust preprocessing techniques, including handling missing values, dealing with outliers, and scaling features appropriately, improves the overall reliability of the machine learning model.

Top of Form

# REFERENCES

Bahnsen, A. C., Aouada, D., Stojanovic, A. and Ottersten, B. (2016). Detecting credit card fraud using periodic features.

Behera, T. K. and Panigrahi, S. (2015). Credit card fraud detection: A hybrid approach using fuzzy clustering neural network.

Choi, D. and Lee, K. (2017). Machine learning based approach to financial fraud detection process in mobile payment system, IT CoNvergence PRActice (INPRA) 5.

Ileberi, E., Sun, Y. and Wang, Z. (2021). Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost, IEEE Access 9.

Jain, V., Agrawal, M. and Kumar, A. (2020). Performance analysis of machine learning algorithms in credit cards fraud detection.

Kalbande, D., Prabhu, P., Gharat, A. and Rajabally, T. (2021). A fraud detection system using machine learning.

Kaur, P., Sharma, A., Chahal, J. K., Sharma, T. and Sharma, V. K. (2021). Analysis on credit card fraud detection and prevention using data mining and machine learning techniques.

Kolodiziev, O., Mints, A., Sidelov, P., Pleskun, I. and Lozynska, O. (2020). Automatic machine learning algorithms for fraud detection in digital payment systems, EasternEuropean Journal of Enterprise Technologies 5.

Rai, A. K. and Dwivedi, R. K. (2020). Fraud detection in credit card data using unsupervised machine learning based scheme. 18

Rambola, R., Varshney, P. and Vishwakarma, P. (2018). Data mining techniques for fraud detection in banking sector.

Ranjan, P., Santhosh, K., Kumar, A. and Kumar, S. (2022). Fraud detection on bank payments

using machine learning.

Saputra, A. and Suharjito (2019). Fraud detection using machine learning in e-commerce, International Journal of Advanced Computer Science and Applications 10.

Singh, P., Chauhan, V., Singh, S., Agarwal, P. and Agrawal, S. (2021). Model for credit card fraud detection using machine learning algorithm.

Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S. and Kuruwitaarachchi, N. (2019). Real-time credit card fraud detection using machine learning.

Wang, S., Minku, L. L. and Yao, X. (2015). Resampling-based ensemble methods for online class imbalance learning, IEEE Transactions on Knowledge and Data Engineering 27.

Yan, T., Li, Y. and He, J. (2021). Comparison of machine learning and neural network models on fraud detection. Yee, O. S., Sagadevan, S. and Malim, N. H. A. H. (2018). Credit card fraud

detection using machine learning as data mining technique, Journal of Telecommunication, Electronic and Computer Engineering 10.

Zanin, M., Romance, M., Moral, S. and Criado, R. (2018). Credit card fraud detection through parenclitic network analysis, Complexity 2018.