Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

CREDIT CARD FRAUD DETECTION BASED ON XGBOOST ALGORITHM

Greeshma Arya[1], Ahmed Hesham Sedky[2], Vikas Rathi[3], Nivriti Pandey[4], Vidushi Pathak[5], Preeti Shubham[6]
1# Indira Gandhi Delhi Technical University for Women, Delhi, India
greeshmaarya@igdtuw.ac.in
2# Arab Academy for Science, Technology, and Maritime Transport, Egypt. ahsedkyy@aast.edu
3#Department of electronics and Communication Engineering Graphic Era Deemed to be university Dehradun vikasrathi@geu.ac.
4# Indira Gandhi Delhi Technical University for Women, Delhi, India
nivriti030btece19@igdtuw.ac.in
5#Indira Gandhi Delhi Technical University for Women, Delhi, India
vidushi004btece19@igdtuw.ac.in
6# Indira Gandhi Delhi Technical University for Women, Delhi, India
preeti034btece19@igdtuw.ac.in

ABSTRACT:

Creditcardfraudisagrowingprobleminthemodernworldandaffectsmillionsofpeopleeachyear.ThisPaperoutlines a credit card fraud detection system that uses machine learning to detect fraudulent activity in real-time. The system employs an ensemble of supervised learning algorithms to build a predictive model that can detect fraudulent transactions. The system detects suspicious transactions and alerts the cardholder and financial institution by utilizing data from different sources, such as credit card transactions and demographic data. The algorithm is designed to be highly accurate, efficient and scalable, and can be easily integrated into existing systems. With this system, credit card companies and financial institutions can take proactive steps to reduce the risk of fraud.

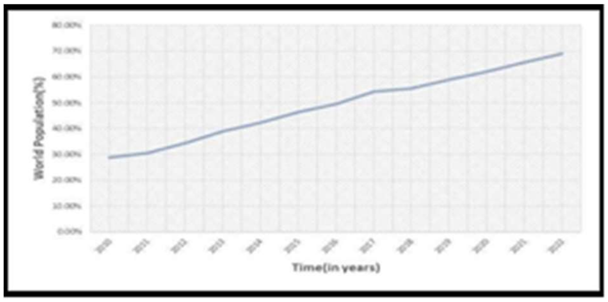
General Terms: Credit Cards, Security, Algorithms

Keywords: Fraud detection, machine learning algorithms, xg boost, k nearest neighbors, decision tree, logistic regression, support vector machine.

1. INTRODUCTION

Financial fraud is a growing problem with far-reaching repercussions for the financial industry, businesses, and government, such fraud can be characterized as Intentional deceit for the sake of financial benefit. credit card purchases have now become predominant

FIGURE 1: Growth of Internet users



means of payment for online and offline purchases. In addition, we present a comprehensive approach for detecting credit card fraud using a machine learning system.

The growth of internet users has been exponential over the past few decades as shown in the above graph. The system employs a two-stage approach, featuring a deep learning model to detect suspicious transactions and a rule-based system to further validate the accuracy of the model We will discuss the data pre-processing and feature engineering techniques used to prepare the data for the machine learning system, as well as the model selection and evaluation criteria used to determine the best model for the task. Utilizing standard methods to identify fraudulent transactions, such as the rise of big data, has rendered manual approaches tedious and impractical. Nevertheless, financial institutions have emphasized a focus on modern computational techniques to manage credit card fraud issues. The rule-based system is used to check the validity of the transactions, by comparing the transaction data with available customer information. The proposed system is evaluated using the real-world credit card transactions dataset and is shown to provide high accuracy in detecting fraudulent transactions. The proposed system is expected to help financial institutions and customersto reduce the cost of fraud and to improve their security.

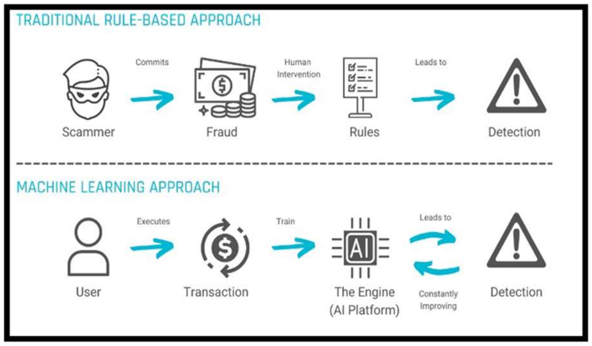


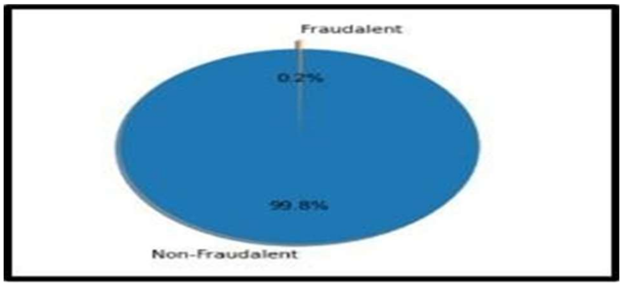
FIGURE 2: Technical Rule Based approach vs Machine Learning Approach [2]

2. LITERATURE REVIEW

Credit card purchases are often categorized using a binary system. There is difficulty in categorizing the data. The two possible outcomes of a credit card transaction in this context are as a negative class) or illegitimate (negative class) transaction (positive class). Detecting fraud can be typically done like classification issues in data mining, where the goal is to properly categorize credit card purchases as legal or illegitimate [5].

2.1. Techniques for Credit Card Fraud Detection Using Machine Learning

This work we analyzed is a comprehensive overview of various machine learning several approaches to classify credit card transactions as unlawful or lawful. Swindling with credit cards happens when sensitive information about a credit card such as card number, card verification value, and card type, all of which is exploited by a con artist for monetary gain.



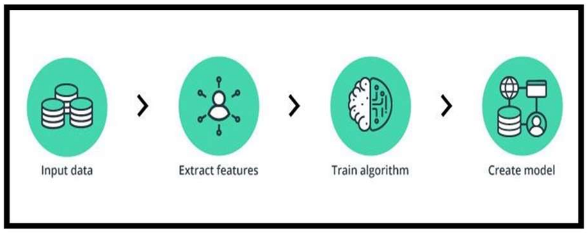


FIGURE3:Flow of Basic Machine Learning Model

This figure gives us the flow of machine learning model which involves: Data Collection->Data Preparation->Model Selection->training the model->model evaluation->model deployment Alternatively, you could misplace your credit card. Because of the unbalanced nature of the data, the SMOTE [8] technique was used to take excessive samples of it. In addition, feature selection was acted upon [11], and the data set was split in two: data used for both training and evaluation. In the experiment, we employed Logistic Regression and EM methods. The described algorithms are the primary topic of this study.

2.2. Feature Selection

Analysis is the backbone of credit card fraud detection.

Cardholders' spending habits and patterns are part of the analysis. This budget breakdown is examined by picking the right variables to measure the peculiar characteristics of credit cards. The line between what is legal and what is dishonest in business transactions is continually blurring. As a result, picking the right factors is crucial. Distinctions between the two profiles are essential for successful credit card purchase categorization. It's important to consider the factors that build the user profile and influence the strategies utilized, and detection capabilities of systems used to safeguard credit card transactions from fraud.

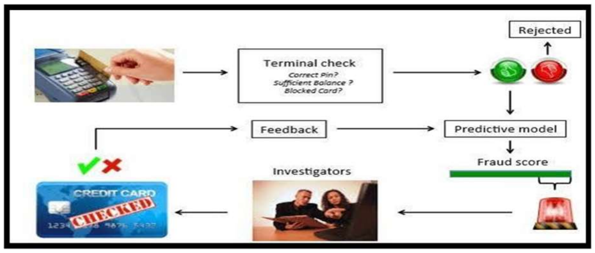


FIGURE4: Machine Learning Model for Credit Card Fraud Detection [3]

¬

3. DATASET DESCRIPTION

This data set [1] contains September 2013 credit card purchases made by consumers in Europe over the course of two days. We used a data set that indicated that 492 out of 2,84,807 total transactions were fake. This dataset is highly biased, since fraud accounts for 0.172% of all transactions. Principal component analysis was used to alter the dataset in order to keep its true nature hidden (PCA).

This figure depicts a pie chart of transactions which is made using the following steps: Determine the total number or value of transactions->Categorize the transactions->Calculate the percentage of each category->Draw and label the pie chart. Other than "time" and "volume," all characteristics (V1, V2, V3, and V28) are the primary components accessible in PCA. The Time field indicates how many seconds have elapsed between the record's first transaction and any subsequent transactions. The "amount" is the total monetary sum involved. The "class" property stands in for the class identifier and takes on the value 1 if it's forbidden or 0 otherwise.

4. PROPOSED WORKFLOW

Steps for applying various Machine Learning Algorithms
Step-1: Dataset Import.
Step-2Arrange the data into tables.
Step 3: Perform a random sampling.
Step-4: Determine the quantity of training data and testing data.
Step 5: Provide 80% of the data of dataset training and 20% for testing.
Step 6: Provide the models access to the training dataset.
Step-7: Select an algorithm from the available algorithms, and use it to build the model.
Step 8: Predict results for each algorithm based on the test dataset.
Step 9: Lastly, we'll use a confusion matrix to determine how effective each algorithm is.

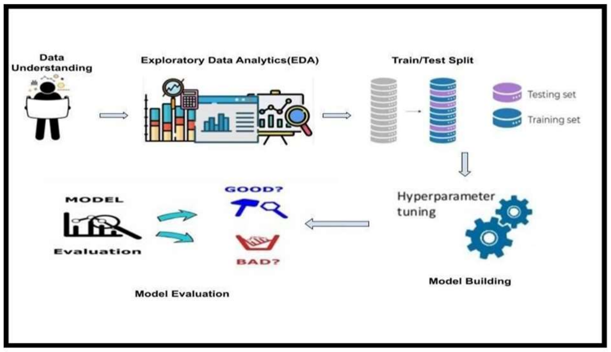
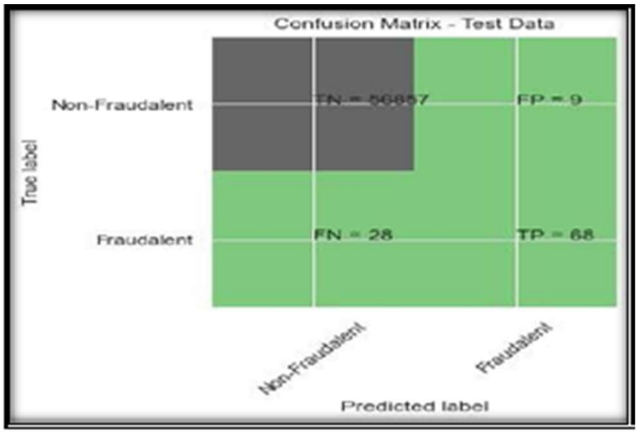


FIGURE 6: Workflow

4.1. Data Understanding

Data Understanding helps in selecting which qualities are necessary for the final model. Out of a total of 2,84,807 transactions, 492 fake transactions were found while studying the dataset used. In this dataset, the positive category (frauds) only makes up 0.172% of all transactions, indicating that something is off.



4.2. Exploratory Data Analysis

In order to avoid complications during the model building process, we checked the data for skewness and tried to reduce it. There is no need for normalization because the variables transformed by PCA are already Gaussian. If there is a skew in the data, a transform can help measure it and correct it. The conventional oversampling method will not be used because it does not improve the dataset [1], and the under-sampling method will not be used because it results in lost information.

4.3. Train/Test Data

The data is split into two sets: training and testing data in 80:20 ratio and the machine learning algorithms are applied. The training and testing set are further used in the model building phase.

4.4Model-Building/Hyper parameter Tuning

We determined which machine learning (ML) model operates well with imbalanced data in order to achieve better results on test data. The final phase permits testing and hyper parameter tuning of several models to obtain the desired degree ofp performance on the provided dataset [1].

4.5 Model Evaluation

Utilizingtheappropriateassessmenttechniqueswhileassessing the models. Because of the data's imbalance, It was observed that accurate detection of fraudulent transactions is more crucial than non-fraudulent ones. We also can't put too much stock in metrics like accuracy, recall, and F1 score right now because they all have caveats that must be met before they can be considered reliable. All potential threshold values are considered by the ROC curve. The confusion matrix's intrinsic hard cut-off of 0.5 prevents us from using it as a performance statistic

These tasks, however, become less useful when accuracy is low since more human input is required. A bank with a high transaction value will not be able to recognize transactions that are marked as non-fraudulent if the recall is low.

The strength of the model is assessed by analyzing its performance at each classification threshold using the ROC curve. Hence, we must prioritize a high recall to recognize them in order to establish the optimal model, which is crucial for protecting financial institutions from large-value fraudulent transactions.

5.EXPERIMENTDESIGN

The following is a list of the models that were utilized in this research article.

5.1 Random Forest

Methods from the realm of machine learning, such as random forest, are utilized to solve problems of classification and regression. It employs ensemble learning, a technique FIGURE7-A:MatrixofRandomForest

that pools the outputs of numerous classifiers to solve complex issues. Numerous decision trees are used in the random forest algorithm. The random forest algorithm uses a forest of several decision trees. This model is a type of ensemble learning that may be used in a variety of settings, including classification, regression, and more. During training, many decision trees are built, and the mode of the classes (classification) or the mean prediction (regression) of the individual trees is then returned [14]. It's one of the most robust and widely adopted machine learning algorithms, and it's already revolutionizing industries as diverse as banking, medicine, and advertising. Because of its ability to learn from labeled data and generate predictions on new data, Random Forest is considered a supervised learning system.

5.2 KNN

To classify new data points, a simple supervised classification approach known as KNN (K-Nearest Neighbor) can be used.KNN works by finding the closest match of an input data point to its neighbors, and then using the neighbors’ labels to make a prediction [2].

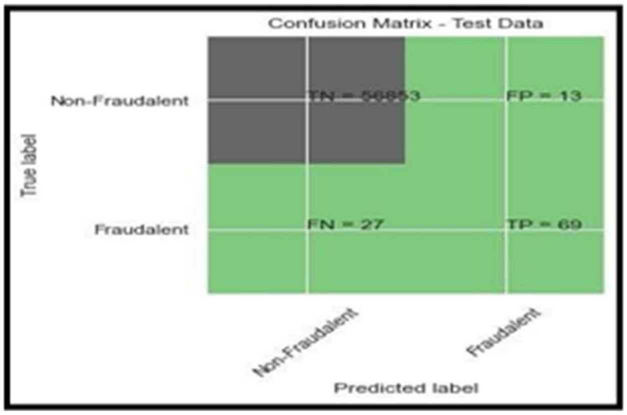


FIGURE7-B:Matrixof KNN

The algorithm is particularly useful when there is no clear boundary between classes, as it can find a pattern in the data that can be used to make a prediction.TheKNNalgorithmcanbeusedforbothclassificationandregressiontasks.Anotherapplicationisregression.KNNisnonparametricbecause it does not assume data distribution.

5.3 Logistic Regression

Classification and diagnosis are common applications of the statistical model also known as a logic model. Probability of something happening is determined by logistic regression. To what extent, for instance, a predetermined set of independent variables was utilized in making the call. In order to model the data, logistic regression employs the sigmoid function. This model is a widely used statistical technique for predictive analysis and classification problems. It is a supervised machine learning algorithm which is used to predict a binary outcome. It is used to estimate the probability of an event occurring, by fitting data to a logistic function.

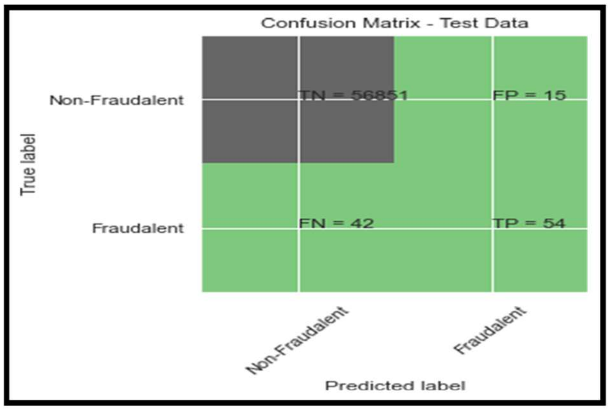


FIGURE7-C:MatrixofLogisticRegression

It is used in various fields such as medical diagnosis, credit scoring, marketing, and other areas where prediction of a binary outcome is required. Logistic regression is a powerful tool for predicting the probability of success in given situation. It is also used to identify relationships between independent variables and the probability of a dependent variable occurring.

5.4 Decision Tree:

For the purpose of classification and regression decision tree is one of the supervised learning methods. It uses a tree diagram to depict choices and their potential outcomes, such as the probabilities involved, the costs and benefits of available resources, and the value of the options being considered. It is a type of predictive analytics that can be used to make predictions about a given dataset [1]. The decision tree algorithm uses a tree-like structure, which consists of nodes, branches, and leaves [6]. Each node represents a decision (or feature) while the branches represent the possible outcomes of the decision. The leaves of the decision tree represent the predicted outcome of the model.

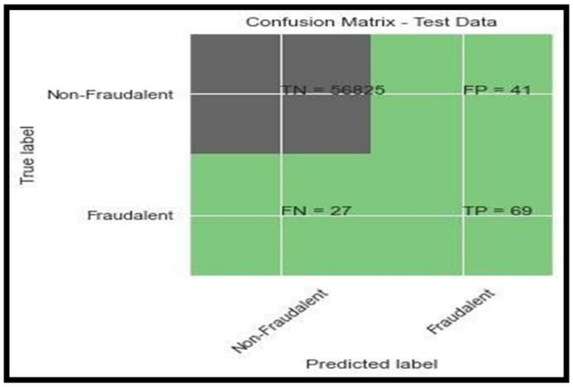


FIGURE7-D: Matrix of Decision Tree

In order to analyze the data, the algorithm divides it into smaller chunks and then builds a hierarchy out of the subsets .

5.5 SVM:

Analyses of classification and regression are especially well-suited to this approach. In a variety of contexts, you can put this method's strength and adaptability to good use. The method finds the optimal hyper plane that divides a set of points into two groups. The foundation of SVMs is the idea of decision planes, which establish the limits of possible courses of action.

A decision plane is a two-dimensional divide between classes of things in order to classify the input data, the algorithm produces a line (or hyper plane). By using SVMs, we can determine which hyper plane best separates the two groups.

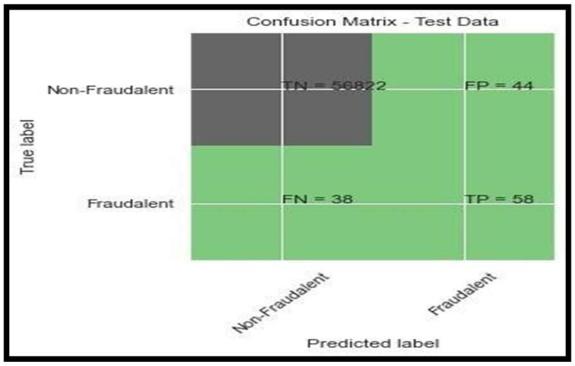


FIGURE7-E: Matrix of SVM

The term "maximum-margin hyper plane" describes this hyper surface. Non-linear classification problems can also be tackled with SVM using the kernel approach, which maps the data onto a higher dimensional space. The algorithm generates a line (or hyper plane) that best divides the given data points into classes. SVMs can be used to find the optimal hyper plane that maximizes the margin between the given two classes [9].

5.6 XG Boost:

Extreme Gradient Boosting, or XG Boost, is a technique developed by academics at the University of Washington. It is a C++ package for improving gradient-boosting training XG Boost is a gradient-boosted decision tree implementation. It is an ensemble learning technique that combines multiple weak or base learners to form a strong learning algorithm. XG Boost is an optimized version of the Gradient Boosting [10] algorithm, designed to provide higher accuracy and speed than the traditional version.

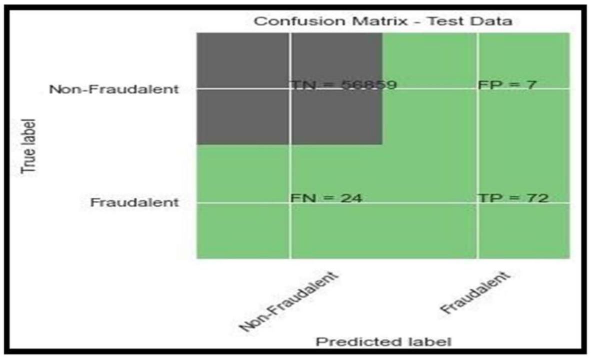


FIGURE7-F: Matrix of XG Boost

6. MODEL EVALUATION

Usings k learn. metrics. Rocaucs core for this purpose. Sk learn needs this so that the area under the curve (AUC) and receiver operating characteristic curve (ROC) metrics may be used to evaluate highly asymmetrical datasets.

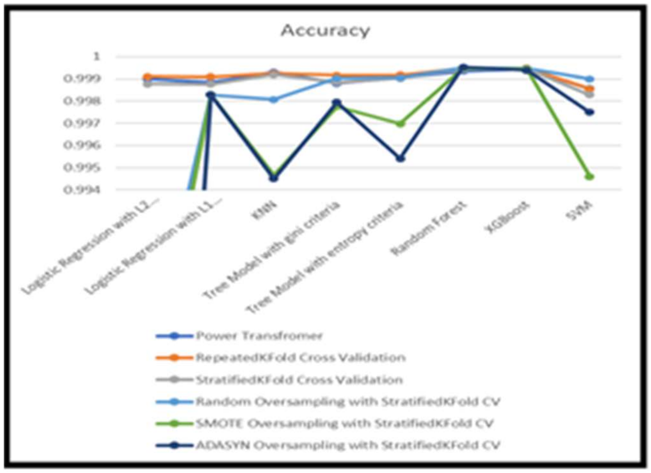


FIGURE8-A:AccuracyfordifferentModels

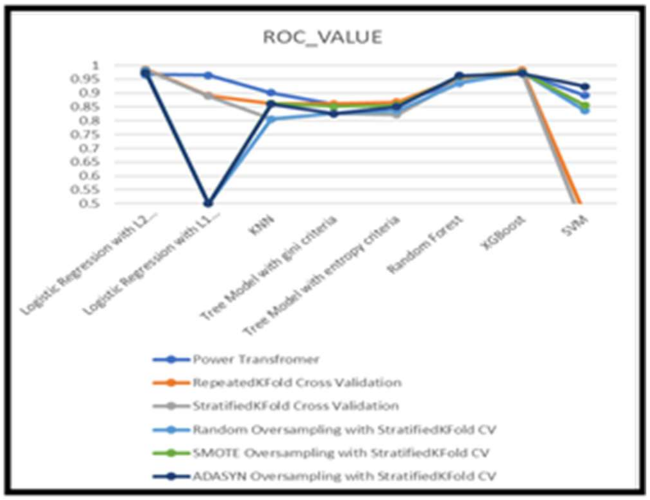


FIGURE8-B:ROC_ValuefordifferentModels

It is more detrimental to ROC to have false positives than false negatives. The ROC curve is a graph of the True Positive Rate (TPR) and False Positive Rate (FPR). The threshold with the highest TPR-FPR value in the training set is frequently the best cutoff to use. Confusion Matrix is not used as a performance metric because it has an internal hard threshold of 0.5. Accuracy metrics like Precision, Recall, and F1-Score are now vulnerable to threshold effects and should be used with caution.

7. RESULTS AND CONCLUSION

The primary objective of the paper is to determine whether a particular credit card transaction involves fraudulent activity. First, the classification is carried out utilizing Random Forest and several balancing methods. The goal was to create a tool that can reliably identify instances of credit card theft. This system should be able to identify fraudulent transactions, identify and eliminate false positives, and protect customers from fraudulent charges. The SMOTETomek balancing approach produced the greatest results out of the various techniques utilized to handle skewed data, including Under sampling, Oversampling, and SMOTE Tomek. Both artificial neural networks and XGBoost classifiers produced results that were comparable, has been discovered that using ML techniques is a fantastic way to increase the precision of credit card fraud detection. It appears that the XGBOOST model with Random Oversampling and Stratified K Fold CV generated the best results [7]. To attain the best results, we attempted to optimize the model's hyper parameters. The accuracy of a machine learning model is an evaluation metric that determines the fraction of correct predictions made by the model. It is usually measured as the proportion of correct predictions out of all the predictions made. It is used to gauge how well the model is performing and can be used to evaluate the performance of different models. Using real-time datasets and our suggested method, we can train the model while maintaining anonymity. It is possible that financial organizations, such as banks, might benefit from the suggested technology and real-time datasets to develop an effective system for identifying credit card fraud through cooperation and sharing of information. The XGBOOST model combined with Random

Oversampling and Stratified K Fold CV provides the best overall fit to identify fraudulent credit card activity.

Model Accuracy: 0.9993855444953564
XGboost roc_value:0.9852138347557161
XGBoost threshold:0.00508787808939814

8. Future work: Anomaly

Detection Algorithms can be used to identify abnormal credit card transactions that may be indicative of fraud. Additionally, they may be able to use block chain technology to track payments and transactions in real-time, making it easier to detect and prevent the fraud. Anomaly Detection Algorithms can be used to identify abnormal credit card transactions that may be indicative of fraud. These algorithms can detect the outliers and abnormal patterns in the data, which may help identify fraudulent transactions. Several supervised and unsupervised machine learning models, including decision trees, random forests, and neural networks, can be used to spot suspicious financial dealings and prevent fraud.

REFERENCES

[1] Dataset:https://www.kaggle.com/mlgulb/creditcardfraud

[2] https://www.locksmithwalnutcreekcalifornia.com/fraud- monitoring-solution-k.html

[3] https://www.researchgate.net/figure/The-Credit-Card- Fraud-Detection-Process_fig1_325658124

[4] https://images.app.goo.gl/vGADKG1WX2mkc5HY7

[5] Credit Card Fraud Detection Using Machine Learning ,ICAC3(2021)Deep Prajapati,AnkitTripathi,AnkitTripathi

[6] Credit Card Fraud Detection Using Machine Learning (ICICCS 2021) D. Tanouz,R Raja Subramanian,D. Eswar

[7] Xuan, Shiyang, etal."Random Forest for Credit Card Fraud Detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control(ICNSC)

[8] Guo S,Liu Y,Chen R,Sun X, Wang X. X, Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes. Neural Process Lett. 2019

[9] Mohammed, Emad, and Behrouz Far."Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study."

[10] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. Xgboost: extreme gradient boosting.

[11] Hemavathi D, Srimathi H. Effective feature selection technique in an integrated environment using enhanced principal component analysis. J Ambient Intell Hum Comput.2021.