

CREDIT CARD FRAUD DETECTION REPORT



LOVELY
PROFESSIONAL
UNIVERSITY

Submitted to

Dr. SAGAR PANDE

Department of Computer Science and Engineering

Lovely Professional University, Phagwara, Punjab, 144411

Submitted by:

Name: Naga Sai Surya Kartheek B

Course: Machine Learning Foundation

Reg No: 11903595

Section: KM055


Course Code: INT 247

Roll No: B56

DECLARATION

To whom so ever it may concern

I, Naga Sai Surya Kartheek B, 11903595, hereby declare that the work done by me on “CREDIT CARD FRAUD DETECTION” from February 2022 to March 2022 is a record of original work for the partial fulfillment of the requirements for the award of degree B.TECH in COMPUTER SCIENCE DEPARTMENT is a record of bonfide project work carried out by me under the guidance of Dr. SAGAR PANDE.

Signature: 

Name: Naga Sai Surya Kartheek B

Reg No: 11903595

Date: 25th March 2022

ACKNOWLEDGEMENT

I would like to express my thanks to Dr. Sagar Pande for giving me a great opportunity to excel in my learning through this project.

I have achieved a good amount of knowledge through the research and the help that I got from Dr. Sagar Pande.

Apart from this, I would like to express my thanks to my parents who have supported me and helped me in every aspect despite their busy schedules.

INDEX

Serial No.	Name of Topic	Page Number
1	Introduction	6
2	Literature Review	7
2.1	Basic Terminologies	10
2.2	Credit Card Fraud Identification Methods	10
2.3	Related Work	18
3	Overview of the Project	22
3.1	Basic Working of Credit Card	22
3.2	Aim of the Project	22
3.3	Proposed System	23
3.4	Data Flow Diagram of Proposed Model	23
3.5	System Architecture of Proposed Data	23
4	Circuit Description	24
4.1	Data Pre-Processing	24
4.2	Classification Model	24
4.3	Model Evaluation	25
5	Software Details	27
6	Results	28
6.1	Results Analysis	28
6.2	Screenshots	29
7	Conclusion	31
8	Future Scope	31

9	References	32
----------	-------------------	-----------

LIST OF FIGURES

Figure No	Content	Page No.
1.1	Credit Card Fraud in US (2014-2019)	7
1.2	Credit Card Loss Global	8
1.3	Basic Framework of a Credit Card Fraud Detection System	9
2.1	Framework of Fraud Detection	10
2.2	K-distance of various neighbours	12
2.3	Reachability distance	12
2.4	Plot of LOF	13
2.5	Leaf-wise tree growth(XGBoost and Light BGM)	15
3.1	Basic Working of a Credit Card	22
3.2	Data Flow Diagram of proposed model	23
4.1	Heatmap	26
6.1	Confusion Matrix of the proposed system	28
6.2	Transaction Class Distribution graph	29
6.3	Time vs Amount graph	30
6.4	Amount per Transaction graph	30

ABSTRACT

Credit Card Fraud is one of the major ethical issues in the credit card industry. The main aims are, firstly, to identify the different types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. Due to fast growth of E-Commerce, use of credit cards for online purchases has dramatically increased and it caused an increase in the credit card fraud. As credit card has become the most popular mode of payment for online and regular purchase, frauds associated with it are rising. Research by Crowe UK and the Centre for Counter Fraud Studies at University of Portsmouth, Europe's premier fraud research centre, has found that the financial cost of fraud is \$5.38 trillion or 6.4 per cent of global GDP. Human analyst teams not only have a hard time detecting fraud as it happens — 45 percent of Financial institutions say investigations take too long to complete — but also have a penchant for mistakenly identifying legitimate account creations and transactions as fraudulent. False-positive rates are even up to 90 percent for some banks, resulting in everything from frustrating obstacles for customers as they try to clear their names to customer abandonment. To counter the frauds effectively, we must rely on the use of advanced technology. ML-based systems in particular have been shown to reduce fraud investigation times by 70 percent and improve the accuracy of fraud detection by 90 percent, with the total portion of fraud attempts detected reaching an impressive 95 percent.

CHAPTER 1

INTRODUCTION

The popularity of online shopping is growing day by day. As it stands in 2021, the number of digital buyers is at 2.14 billion. That makes 27.6 percent of the 7.74 billion people in the world. In other words, more than one out of every four people you see around you is an online shopper. With the increase of online shoppers, there has been an increase in more frauds as well. Some faux e-stores are invented from whole cloth, but many mimic trusted retailers, with familiar logos and slogans and a URL that's easily mistaken for the real thing. They offer popular items at a fraction of the usual cost and promise perks like free shipping and overnight delivery, exploiting the premium online shoppers put on price and speed.

There has seen a surge in online card fraud in recent years, and it accelerated during the pandemic as more people shopped on the internet to avoid brick-and-mortar stores, according to a report. Losses from online card fraud have reached almost \$8 billion in the year 2020, up from about \$6 billion in 2019.

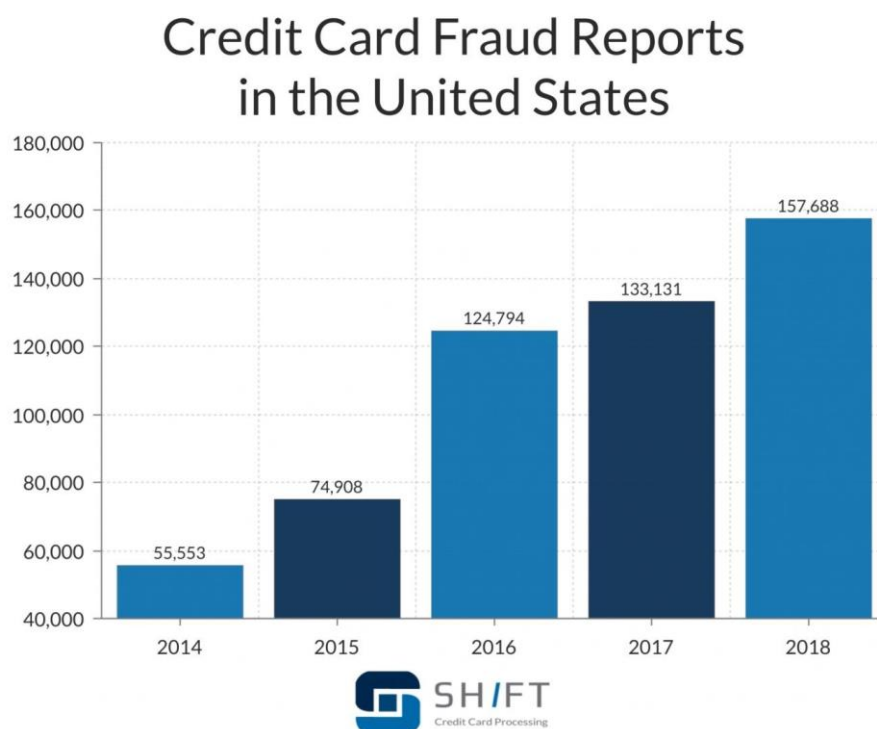


Fig 1.1 Credit Card Fraud loss in US

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Credit Card Fraud 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.

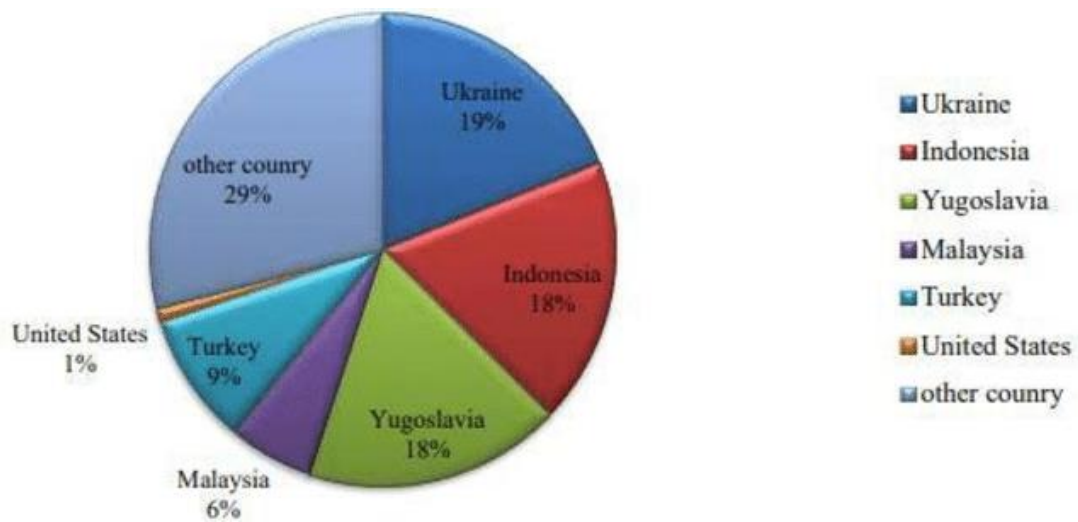


Figure 1.2 Credit Card Loss Chart Global

There are different types of credit card fraud based on the nature of fraudulent activities such as card getting stolen, obtaining cards using false information, individuals using credit cards while being unable to pay debts, bank employees stealing card details to use it remotely, individual using skimming devices to hack credit card details, etc. Credit card fraud is usually caused either by card owner's negligence with his data or by a breach in a website's security. Here are some examples:

- A consumer reveals his credit card number to unfamiliar individuals.
- A card is lost or stolen and someone else uses it.
- Mail is stolen from the intended recipient and used by criminals.
- Business employees copy cards or card numbers of its owner.
- Making a counterfeit credit card.

Due to rise and acceleration of E- Commerce, there has been a tremendous use of credit cards for online shopping which led to High amount of frauds related to credit cards. In the era of digitalization, the need to identify credit card frauds is necessary. Machine learning algorithms are employed to analyze 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.

Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:

- Credit Card Frauds: Online and Offline

- Card Theft
- Account Bankruptcy
- Device Intrusion
- Application Fraud
- Counterfeit Card
- Telecommunication Fraud

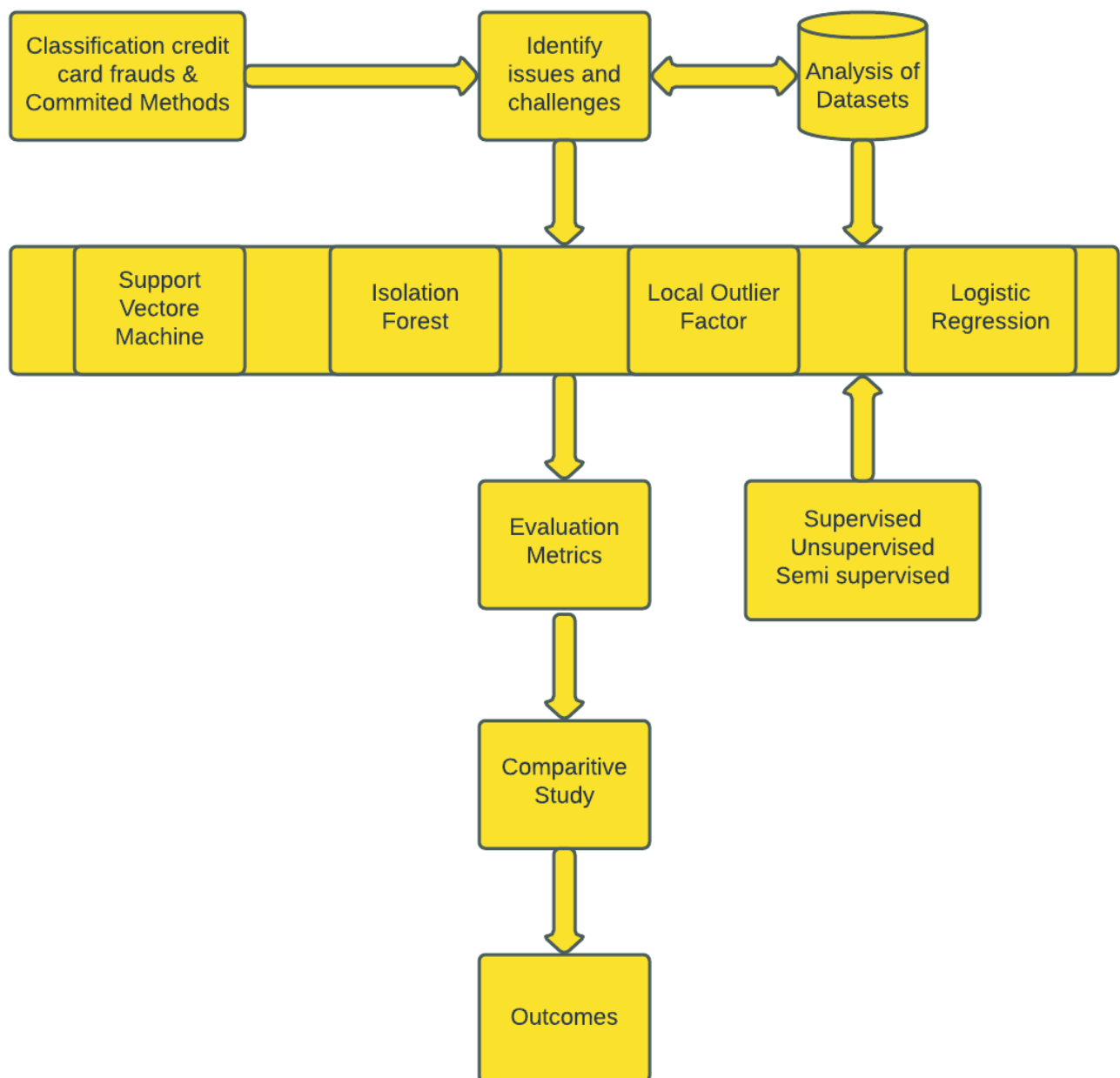


Figure 1.3 Basic Framework of a Credit Card Fraud Detection system.

CHAPTER 2

LITERATURE REVIEW

2.1 Basic Terminologies

The fraud detection system is a complex task and there is no system that has a prediction rate of 100%. No System can correctly predict any transaction as fraudulent. The properties for a good fraud detection system are:

1. Should Identify the frauds accurately.
2. Should detect the frauds quickly.
3. Should not classify a genuine transaction as fraud.

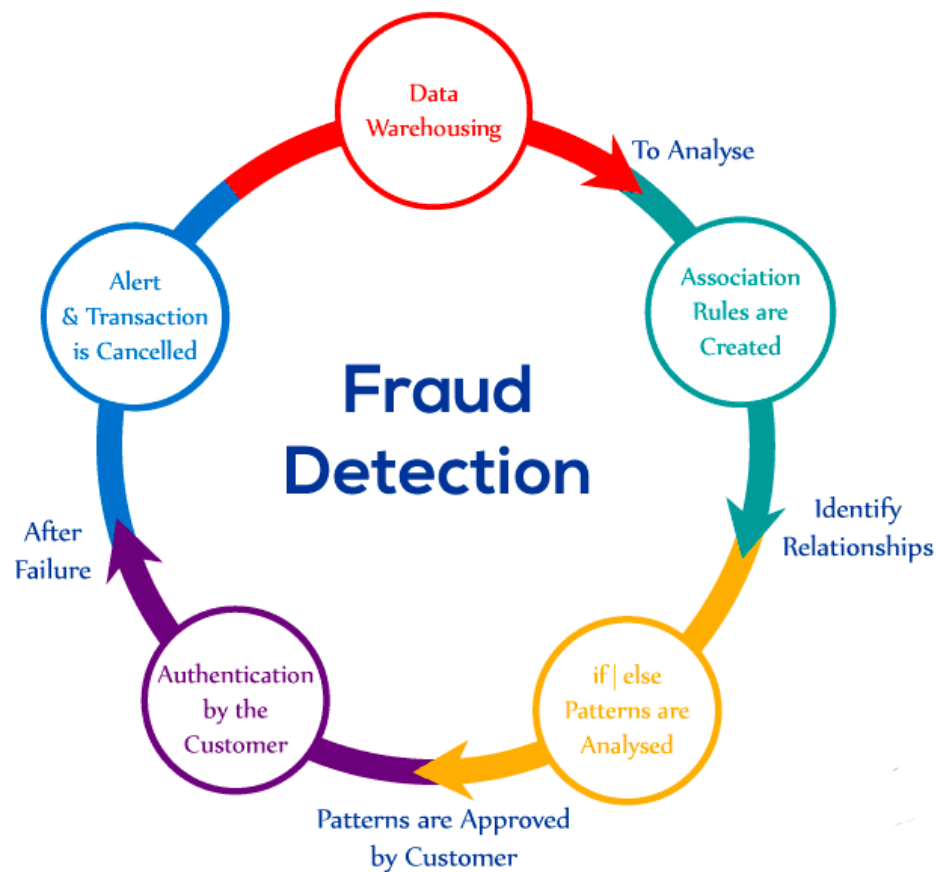


Figure 2.1 Framework of Fraud Detection.

2.2 Credit Card Fraud Identification Methods

Credit Card Fraud Identification Method are split into:

- Unsupervised

- Supervised

2.2.1 Unsupervised

Unsupervised Machine Learning methods use unlabeled data to find patterns and dependencies in the credit card fraud detection dataset, making it possible to group data samples by similarities without manual labeling. Example – Principal Component Analysis, Local Outlier Factor, Support Vector Machine, Isolation Forest.

Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed. The underlying data can be measurements describing properties of production samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals or trials of a DOE-protocol, for example. PCA is one of the most popular techniques for Anomaly Detection. PCA searches for correlations among features which in the case of credit card transactions, could be time, location, and amount of money spent and determines which combination of values contributes to the variability in the outcomes. Such combined feature values allow the creation of a tighter feature space named principal components.

Local Outlier Factor (LOF)

LOF is an algorithm used for Unsupervised outlier detection. It produces an anomaly score that represents data points which are outliers in the data set. It does this by measuring the local density deviation of a given data point with respect to the data points near it. is another of the most popular Anomaly Detection methods. To calculate LOF, the number of neighboring data points is considered to figure out its density and compare it to the density of other data points. If a certain data point has a substantially low density compared to its close neighbors, it is an outlier.

Working of LOF: Local density is determined by estimating distances between data points that are neighbors (k-nearest neighbors). So for each data point, local density can be calculated. By comparing these we can check which data points have similar densities and which have a lesser density than its neighbors. The ones with the lesser densities are considered as the outliers. Firstly, k-distances are distances between points that are calculated for each point to determine their k-nearest neighbors. The 2nd closest point is said to be the 2nd nearest neighbor to the point. Here is an image which represents k-distances of various neighbors in the cluster of a point:

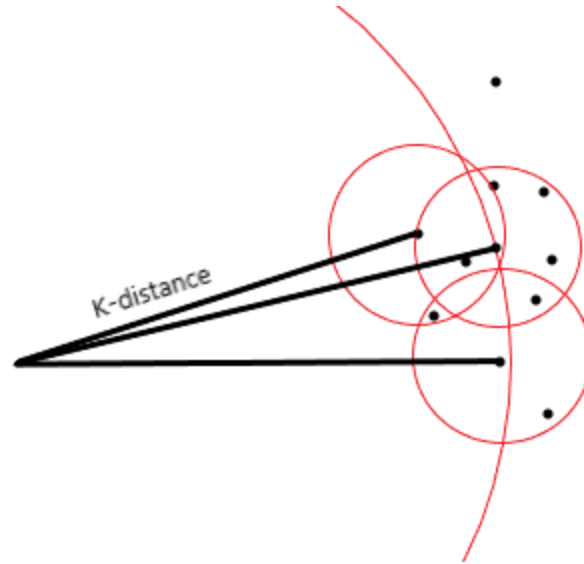


Figure 2.2 K-distance of various neighbours.

This distance is used to calculate the reachability distance. It is defined as the maximum of the distance between two points and the k-distance of that point. Refer to the following equation, where B is the point in the center and A is a point near to it.

$$\text{reachability-distance}_k(A,B) = \max\{k\text{-distance}(B), d(A,B)\}$$

Here is an image which represents reachability distance of a point to various neighbors:

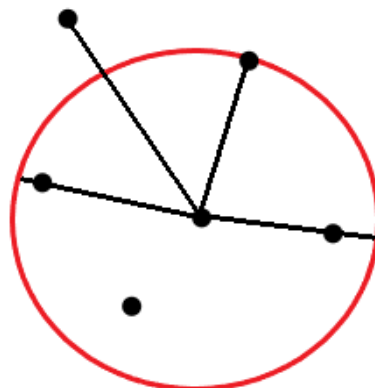


Figure 2.3 Reachability distance

As you can see in the image given above, for points inside the circle the k-distance is considered and for points outside the cluster, the distance between points is considered.

Now, reachability distances to all of the k-nearest neighbors of a point are calculated to determine the Local Reachability Density (LRD) of that point. The local reachability density is a measure of the density of k-nearest points around a point which is calculated by taking the inverse of the sum of all of the reachability distances of all the k-nearest neighboring points. The closer the points are, the distance is lesser, and the density is more, hence the inverse is taken in the equation.

$$\text{lrd}_k(A) := 1 / \left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right)$$

The calculation of Local outlier factor (LOF) is done by taking the ratio of the average of the lrd's of k number of neighbors of a point and the lrd of that point. Here is the equation for LOF:

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}_k(B)}{\text{lrd}_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}_k(B)}{|N_k(A)| \cdot \text{lrd}_k(A)}$$

So, in the equation, if the density of the neighbors and the point are almost equal we can say they are quite similar; if the density of the neighbors is lesser than the density of the point we can say the point is an inlier i.e. inside the cluster, and if the density of the neighbors is more than the density of the point we can say that the point is an outlier. Refer to the following illustration:

LOF ~ 1 => Similar data point

LOF < 1 => Inlier (similar data point which is inside the density cluster)

LOF > 1 => Outlier

Here is an image of the plot of LOF on a data set:

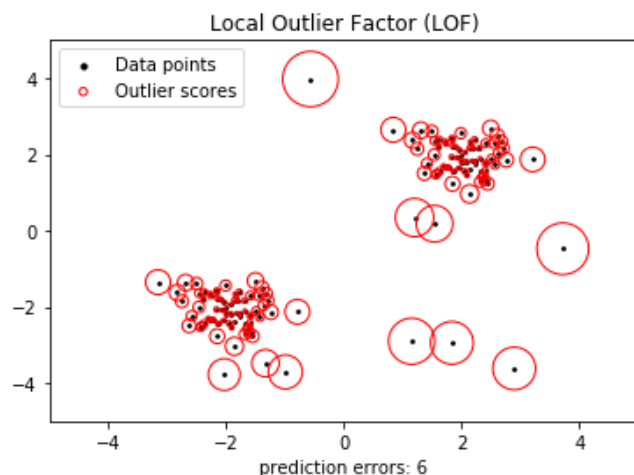


Figure 2.4 plot of LOF

Support Vector Machine (SVM)

Support Vector Machine is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

The idea behind One-class SVM is to train only on a solid number of legitimate transactions and then identify anomalies or novelties by comparing each new data point to them.

Keywords of SVM

1. **Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.
2. **Support Vectors:** The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector.
3. **Margin:** SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin.
4. **Optimal Hyperplane:** The hyperplane with maximum margin is called the optimal hyperplane.

Isolation Forest

Isolation Forest (IF) is an Anomaly Detection method from the Decision Trees family. The main idea of IF, which differentiates it from other popular outlier detection algorithms, is that it precisely detects anomalies instead of profiling the positive data points. Isolation Forest is built of Decision Trees where the separation of data points happens first because of randomly selecting a split value amidst the minimum and maximum value of the chosen feature.

If we have a set of legitimate transactions, the Isolation Forest algorithm will define fraudulent credit card transactions because of their values — which are often very different from the values positive transactions have (i.e. they take place further away from the normal data points in the feature space).

2.2.2 Supervised

Supervised Machine Learning methods use labelled data samples, so the system will then predict these labels in future unseen before data. Example – XGBoost(Extreme Gradient Boosting) and Light GBM(Gradient Boosting Machine), KNN, Random Forest, Logistic Regression.

XGBoost(Extreme Gradient Boosting) and Light GBM(Gradient Boosting Machine)

XGBoost (Extreme Gradient Boosting) and Light GBM (Gradient Boosting Machine) are a single type of gradient-boosted Decision Trees algorithm, which was created for speed as well as maximizing the efficiency of computing time and memory resources. This algorithm is a blending technique where new models are added to fix the errors caused by existing models.

Light GBM differs from other tree-based techniques only in that it follows a leaf-wise direction to build conditions instead of a level-wise direction. In general, the idea behind all tree-based gradient boosting based algorithms is the same.

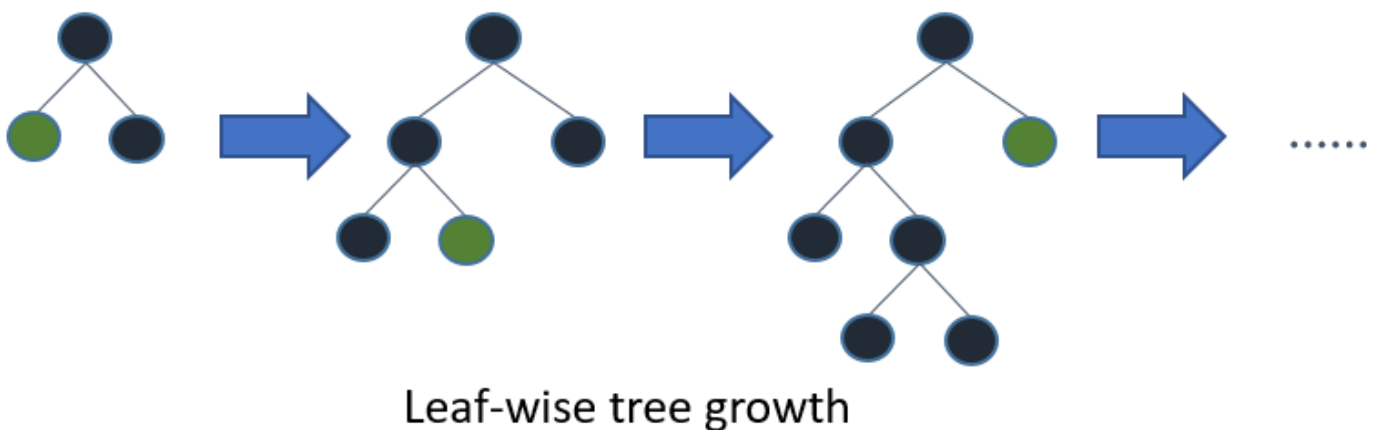


Figure 2.5 Leaf wise tree growth

To classify a transaction as a fraudulent charge, the result (probability) of many Decision Trees is summarized — whereas every future tree improves its results based on of the errors made by its predecessors.

KNN

K-Nearest Neighbors is a Classification algorithm that counts similarities based on the distance in multi-dimensional space. The data point, therefore, will be assigned the class that the nearest neighbors have. This method is not vulnerable to noise and missing data points, which means composing larger datasets in less time. Moreover, it is quite accurate and requires less work from a developer in order to tune the model.

Random Forest

Random Forest is a classification algorithm that is comprised of many Decision Trees. Each tree has nodes with conditions, which define the final decision based on the highest value.

The Random Forest algorithm for fraud detection and prevention has two cardinal factors that make it good at predicting things. The first one is randomness, meaning that the rows and columns of data are chosen randomly from the dataset and fit into different Decision Trees. Say Tree Number 1 receives the first 1,000 rows, Tree Number 2 receives Rows 4,000 to 5,000, and the Tree Number 3 has Rows 8,000 to 9,000.

The second factor is diversity, meaning that there's a forest of trees that contribute to the final decision instead of just one decision tree. The biggest advantage here is that this diversity decreases the chance of model overfitting, while the bias remains the same.

Implementation in Scikit-learn

For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Where,

- $ni_{sub(j)}$ = the importance of node j
- $w_{sub(j)}$ = weighted number of samples reaching node j
- $C_{sub(j)}$ = the impurity value of node j
- $left(j)$ = child node from left split on node j
- $right(j)$ = child node from right split on node j

The importance for each feature on a decision tree is then calculated as:

$$fi_i = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k}$$

Where,

- $fi_{\text{sub}(i)}$ = the importance of feature i
- $ni_{\text{sub}(j)}$ = the importance of node j

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j}$$

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in \text{all trees}} normfi_{ij}}{T}$$

Where,

- $RFfi_{\text{sub}(i)}$ = the importance of feature i calculated from all trees in the Random Forest model
- $normfi_{\text{sub}(ij)}$ = the normalized feature importance for i in tree j
- T = total number of trees

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

2.3 Related Work

Kuldeep Randhawa et al. proposed a technique using machine learning to detect credit card fraud detection. Initially, standard models were used after that hybrid models came into picture which made use of AdaBoost and majority voting methods. Publically available data set had been used to evaluate the model efficiency and another data set used from the financial institution and analysed the fraud. Then the noise was added to the data sample through which the robustness of the algorithms could be measured. The experiments were conducted based on the theoretical results which show that the majority of voting methods achieve good accuracy rates in order to detect the fraud in the credit cards. For further evaluation of the hybrid models noise of about 10% and 30% has been added to the sample data. Several voting methods have achieved a good score of 0.942 for 30% added noise. Thus, it was concluded that the voting method showed much stable performance in the presence of noise. [1]

Abhimanyu Roy et al. proposed deep learning topologies for the detection of fraud in online money transaction. This approach is derived from the artificial neural network with in-built time and memory components like long-term short-term memory and several other parameters. According to the efficiency of these components in fraud detection, almost 80 million online transactions through credit card have been pre-labeled as fraudulent and legal. They have used high performance distributed cloud computing environment. The study proposed by the researchers provides an effective guide to the sensitivity analysis of the proposed parameters as per the performance of the fraud detection. The researchers also proposed a framework for the parameter tuning of Deep Learning topologies for the detection of fraud. This enables the financial institution to decrease the losses by avoiding fraudulent activities. [2]

Prajat Save et al. have proposed a model based on a decision tree and a combination of Luhn's and Hunt's algorithms. Luhn's algorithm is used to determine whether an incoming transaction is fraudulent or not. It validates credit card numbers via the input, which is the credit card number. Address Mismatch and Degree of Outlierness are used to assess the deviation of each incoming transaction from the cardholder's normal profile. In the final step, the general belief is strengthened or weakened using Bayes Theorem, followed by recombination of the calculated probability with the initial belief of fraud using an advanced combination heuristic. [3]

Zahra Kazemi et al. proposed Deep autoencoder which is used to extract the best characteristics of the information from the credit card transaction. This will further add softmax software to resolve the class labels issues. An overcomplete autoencoder is used to map the data into a high dimensional space and a sparse model was used in a descriptive manner which provides benefits for the classification of a type of fraud. Deep learning is one of the most motivated and powerful techniques being employed for the detection of fraud in the credit card. These types of networks have a complex distribution of data which is very difficult to recognize. Deep autoencoder has been used in some stages to extract the best features of the data

and for the classification purposes. Also, higher accuracy and low variance are achieved within these networks. [4]

Sharmistha Dutta et al. presented a study on the commonly found crime within the credit card applications. There are certain issues faced when the existing non-data mining approaches are applied to avoid identity theft. A novel data mining layer of defence is proposed for solving these issues. For detecting the frauds within various applications, two algorithms named Communal Detection and Spike Detection which generate novel layer. There is a large moving window, higher numbers of attributes and numbers of link types available which can be searched by CD and SD algorithms. Thus, results can be generated by the system by consuming a huge amount of time. Since the attackers do not get time to modify their behaviors with respect to the algorithms being deployed in real time, there is no true evaluation achieved even after a regular update of the algorithms. Therefore, it is not possible to properly demonstrate the concept of adaptability. These issues can be resolved by making certain enhancements in the proposed algorithm in future work. [5]

Krishna Modi et al. investigated several techniques that were used for detecting the fraudulent transactions and provided a comparative study amongst them. The fraudulent transactions can be detected by utilizing either one of these or integrating any of these methods. The model can possibly be trained in a more accurate manner by adding new features. Several data mining techniques are being used by bank and credit card companies for detecting fraud behaviors. The normal usage pattern of clients depending upon their past activities can be identified by applying any of these methods. Therefore, a comparative analysis is made here by studying different fraud detection techniques proposed over the years. [6]

Dastgir Pojee et al. proposed a novel mechanism using which the payment of invoice or bill is initiated. This approach is named as „NoCash“ mobile application which is mainly used by the merchants through which the payment facility of clients can be eased. There is no need for NFC-Enabled Point of Sales (PoS) Machines in this approach and only the mobile phones are required. Minimizing the burden of clients for bringing cards when outside, by providing easy payment transferring mechanisms is the only aim for which this system is designed. The client's experience of shopping is improved when NoCash application that includes many features is applied on the basis of the increase in a number of NFC-based mobiles. To provide benefits to merchants, the fraud activities are minimized using this proposed application. The application clients can be related to the expense history and minimize any unwanted costs using this proposed method. [7]

Dilip Singh Sisodia et al. presented the evaluation of the performance of several sampling techniques on the classifier when they are applied on credit card fraud data set with the class imbalance. The principal

component analysis (PCA) is applied to real data as well as the variables time, amount and class to achieve 28 principal components that are included within the data. There are ten thousand, fifteen thousand and twenty thousand instances available respectively within the three datasets. This approach applied five over-sampling and four under-sampling approaches. Further, on the data, few cost sensitive and ensemble classifiers are applied. [8]

Luis Vergara et al. proposed an improvement in the performance of credit card fraud detection by developing various methods that are based on signal processing. A variant of the traditional iterative amplitude adjusted Fourier transform (IAAFT) and the iterative surrogate signals on graph algorithms (ISSG) are present within the proposed methods. Improving the training of detectors is the major aim of this approach. The surrogate samples are generated from original fraud samples in this mechanism. The variance of the estimate is reduced here such that the training of detectors can be improved. Due to the presence of various issues and the constant change of patterns present in the data stream it is important to provide a reliable augmentation of the target scarce population of frauds. The real data was used in this experiment to demonstrate the capabilities of proposed methods such that the performance of detection can be improved. The ROC curves and KPIs which are commonly used in financial business were used in this research to measure the capabilities. [9]

Vimala Devi. J et al. To detect counterfeit transactions, three machine-learning algorithms were presented and implemented. There are many measures used to evaluate the performance of classifiers or predictors, such as the Vector Machine, Random Forest, and Decision Tree. These metrics are either prevalence-dependent or prevalence-independent. Furthermore, these techniques are used in credit card fraud detection mechanisms, and the results of these algorithms have been compared [10]

Popat and Chaudhary supervised algorithms were presented Deep learning, Logistic Regression, Naive Bayesian, Support Vector Machine (SVM), Neural Network, Artificial Immune System, K Nearest Neighbour, Data Mining, Decision Tree, Fuzzy logic-based System, and Genetic Algorithm are some of the techniques used. Credit card fraud detection algorithms identify transactions that have a high probability of being fraudulent. We compared machine-learning algorithms to prediction, clustering, and outlier detection. [11]

Shiyang Xuan et al. [21] For training the behavioural characteristics of credit card transactions, the Random Forest classifier was used. The following types are used to train the normal and fraudulent behaviour features Random forest-based on random trees and random forest based on CART. To assess the model's effectiveness, performance measures are computed. [12]

Dornadula and Geetha S. Using the Sliding-Window method, the transactions were aggregated into respective groups, i.e., some features from the window were extracted to find cardholder's behavioral

patterns. Features such as the maximum amount, the minimum amount of a transaction, the average amount in the window, and even the time elapsed are available. [13]

Sangeeta Mittal et al. To evaluate the underlying problems, some popular machine learning algorithms in the supervised and unsupervised categories were selected. A range of supervised learning algorithms, from classical to modern, have been considered. These include tree-based algorithms, classical and deep neural networks, hybrid algorithms and Bayesian approaches. The effectiveness of machine-learning algorithms in detecting credit card fraud has been assessed. On various metrics, a number of popular algorithms in the supervised, ensemble, and unsupervised categories were evaluated. It is concluded that unsupervised algorithms handle dataset skewness better and thus perform well across all metrics absolutely and in comparison, to other techniques. [14]

Deepa and Akila For fraud detection, different algorithms like Anomaly Detection Algorithm, K-Nearest Neighbour, Random Forest, K-Means and Decision Tree were used. Based on a given scenario, presented several techniques and predicted the best algorithm to detect deceitful transactions. To predict the fraud result, the system used various rules and algorithms to generate the Fraud score for that certain transaction. [15]

CHAPTER 3

OVERVIEW OF PROJECT

3.1 Basic Working of Credit Card

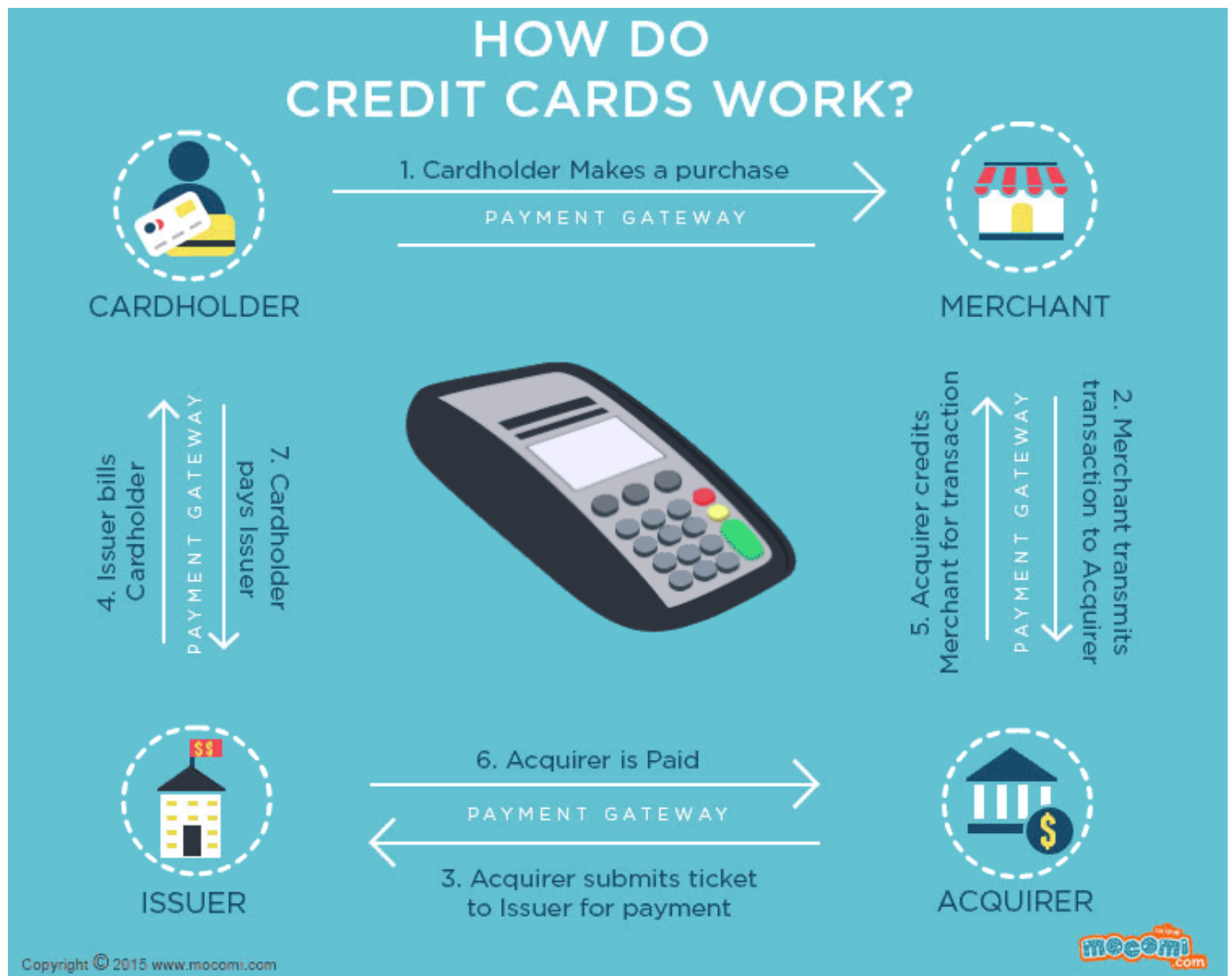


Figure 3.1 Working of a Credit Card

3.2 Aim of the Project

The aim of this project is to predict whether a credit card transaction is fraudulent or not, based on the transaction amount, location and other transaction related data. It aims to track down credit card transaction data, which is done by detecting anomalies in the transaction data.

3.3 Proposed System

In proposed system we use Random Forest Algorithm (RFA) for finding the fraudulent transactions and the accuracy of those transactions. This algorithm is based on supervised learning algorithm where it uses decision trees for classification of the dataset. After classification of dataset a confusion matrix is obtained. The performance of Random Forest Algorithm is evaluated based on the confusion matrix.

3.4 Data Flow Diagram of Proposed Model:

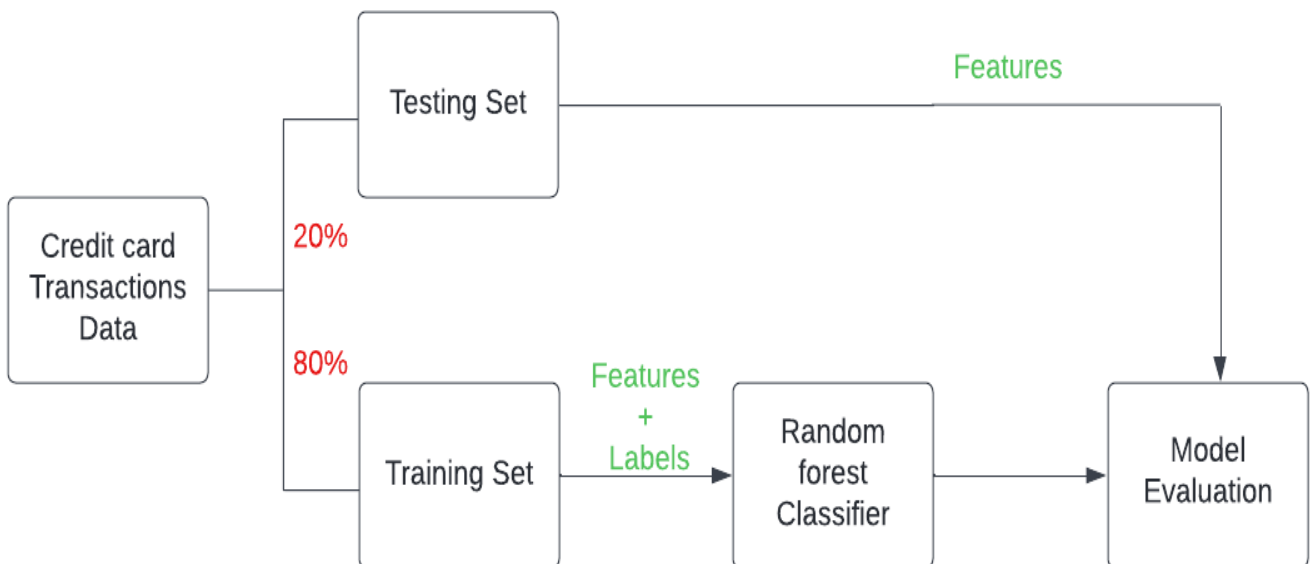


Figure 3.2 Data Flow Diagram of Proposed Model

3.5 System Architecture of Proposed Data

1. We give the system a dataset of various credit card transactions
2. The Machine Learning algorithm uses these as inputs.
3. The system will analyze the data and will show us all the legit and fraud transactions. The algorithm uses this data to observe a pattern to detect anomalies.
4. The legit transactions will be allowed to pass unless there is some suspicious transaction.
5. In case a fraudulent transaction occurs, the system alerts the user who in turn can take necessary steps such as blocking the card to prevent more frauds.

CHAPTER 4

CIRCUIT DESCRIPTION

4.1 Data Pre-Processing

Data pre-processing involves preparing the dataset to train the machine learning model. The data pre-processing step is crucial and should transform the data in a way that can be processed by the selected machine learning algorithm. For example, most classification algorithms will not be able to understand the text in the data, and hence not performing data pre-processing will lead to errors.

Common data pre-processing steps involve – imputing or dropping records containing missing values, label encoding categorical data, one hot encoding labelled data, scaling the data, and performing train-test splits on the dataset.

As this dataset does not contain any missing values or categorical data, most data pre-processing steps are not needed. The data is taken and first split into the predictors i.e. X and the outcome i.e. Y. X contains 284807 data records with 30 features each while Y contains 284807 data records with one column – class.

```
In [31]: X =data.iloc[:, :-1]
         Y =data.iloc[:, -1]
         X_train, X_test, Y_train, Y_test=train_test_split(X, Y, test_size=0.2, random_state=42)
```

The train-test split divides the dataset into a training set and testing set. The training set is used to train the machine learning model while the testing set is used to evaluate the model. The test size of 0.2 indicates that 20% of the dataset is chosen to be the testing set. Hence, the training set contains 227845 records while the testing set contains 56962 records.

4.2 Classification Model

The credit card fraud detection problem is a classification problem, as it involves classifying a credit card transaction to be in either of the two classes – valid or fraudulent. As mentioned, there are several classification algorithms available such as Linear Classifiers, Naïve Bayes Classifier, Support Vector Machines, Nearest Neighbour Classifier, Decision Trees, etc. For this problem, a Random Forest Classifier is implemented which is an extension of the Decision Tree classifier.

```
In [32]: ► classifier=RandomForestClassifier()
classifier.fit(X_train, Y_train)
Y_pred=classifier.predict(X_test)
```

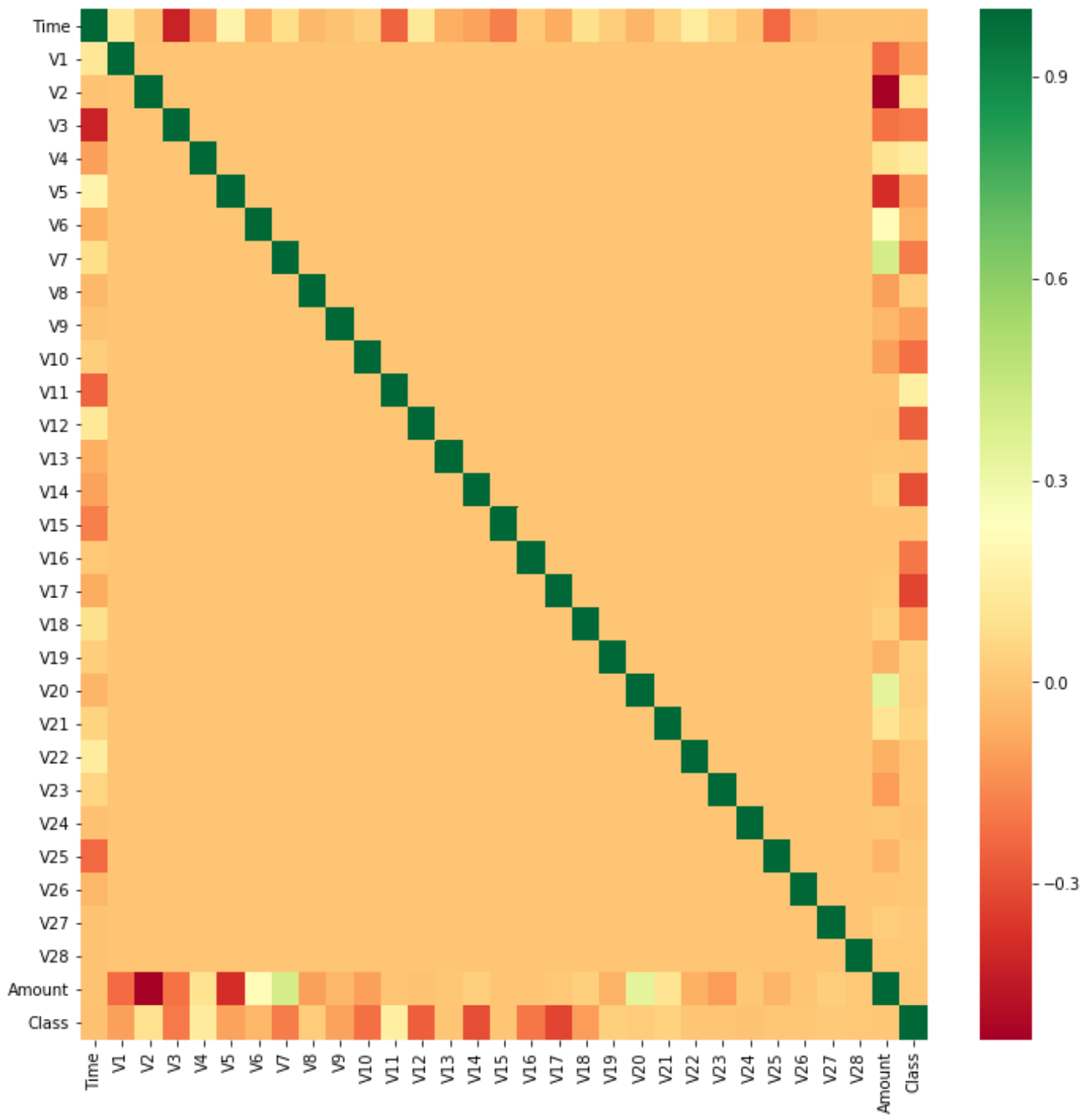
In the above code, an object of the RandomForestClassifier class belonging to the sklearnlibrary is created. Using the fit function of this class, the model is trained using the training set. Finally, the predict function gives a prediction for the values of features in the testing set.

4.3 Model Evaluation

Every machine learning model must be evaluated on the task that it performs. Model evaluation involves asking the model to predict the values for unseen data records based on what it has learnt. This has been done above and is stored in Y_pred. Y_pred are the values as predicted by the model which must be compared against the true values i.e. Y_test. As this is a classification problem, we can evaluate the model using metrics such as accuracy, precision, and recall.

```
In [34]: ► print("Model Accuracy:", round(accuracy_score(Y_test, Y_pred),4))
print("Model Precision:", round(precision_score(Y_test, Y_pred),4))
print("Model Recall:", round(recall_score(Y_test, Y_pred),4))
```

- An accuracy of the model determines how many data records the model predicted correct values.
- Precision indicates the correctness of all those records that were predicted to be positive
- Recall of a model indicates how many truly positive values were identified correctly.



Chapter 4.1 Heatmap

CHAPTER 5

SOFTWARE DETAILS

5.1 Tools

The proposed model uses Python, NumPy, Pandas, Sci-kit learn, Seaborn, Matplotlib.

- Numpy and Pandas – For Analyzing the data.
- Sci-kit learn – Used for Numerical Computation.
- Matplotlib – Used for data visualization and graphical plotting.
- Seaborn – Used for statistical visualization.

5.2 System Requirements

- Operating Systems – Windows 7 and above
- Ram: 4GB and above
- Hard Disk – 128GB SSD or 1TB HDD
- Processor: I3 and above

CHAPTER 6

RESULTS

6.1 Result Analysis

We will analyze the result using Confusion Matrix.

Terminology

- True positive (TP) - A test result that correctly indicates the presence of a condition or characteristic
- True negative (TN) - A test result that correctly indicates the absence of a condition or characteristic
- False positive (FP) - A test result which wrongly indicates that a particular condition or attribute is present
- False negative (FN) - A test result which wrongly indicates that a particular condition or attribute is absent

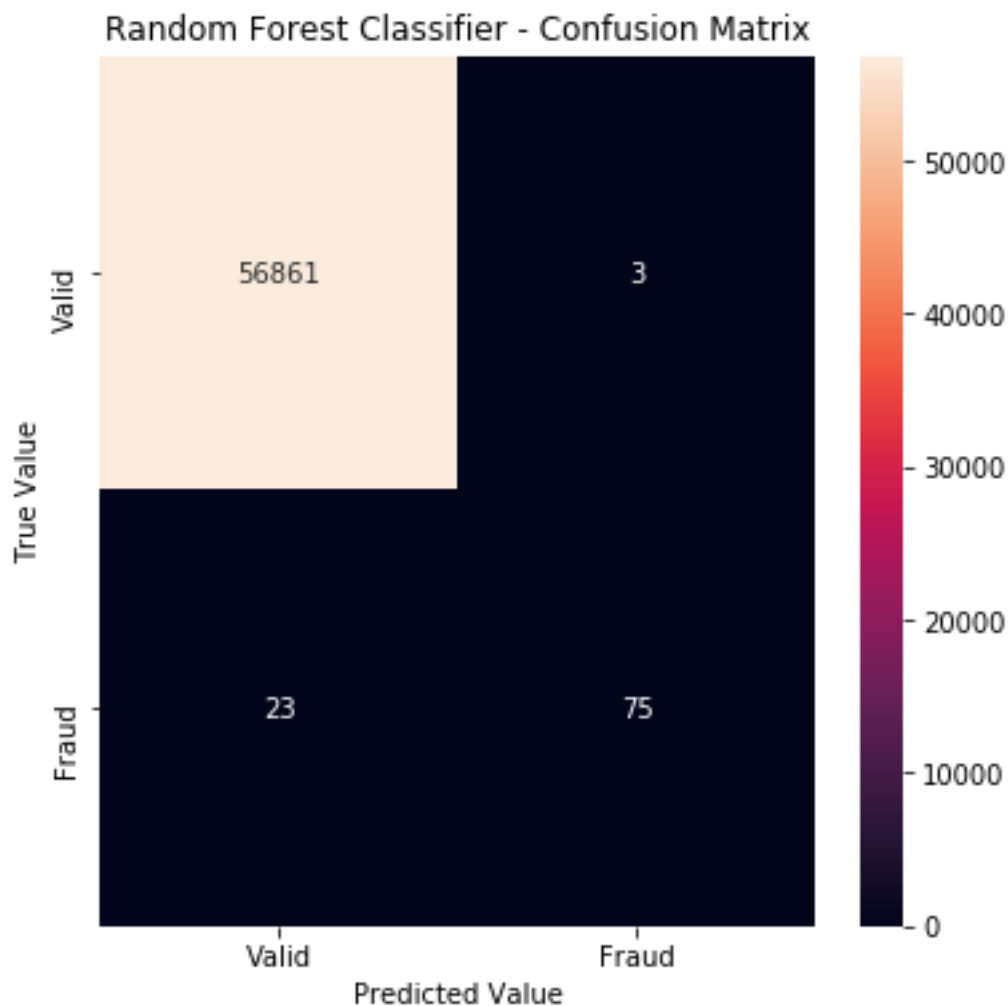


Figure 6.1 Confusion matrix of the proposed system

1. As seen from the Confusion matrix, the model was correctly able to classify 56861 as Legit and 75 records as fraudulent.
2. However, it incorrectly identified a legit transaction as fraudulent 3 times which is 3 False Negatives.
3. It also incorrectly identified 23 fraudulent transactions as legit which is 23 False Positives.

6.2 Screenshot

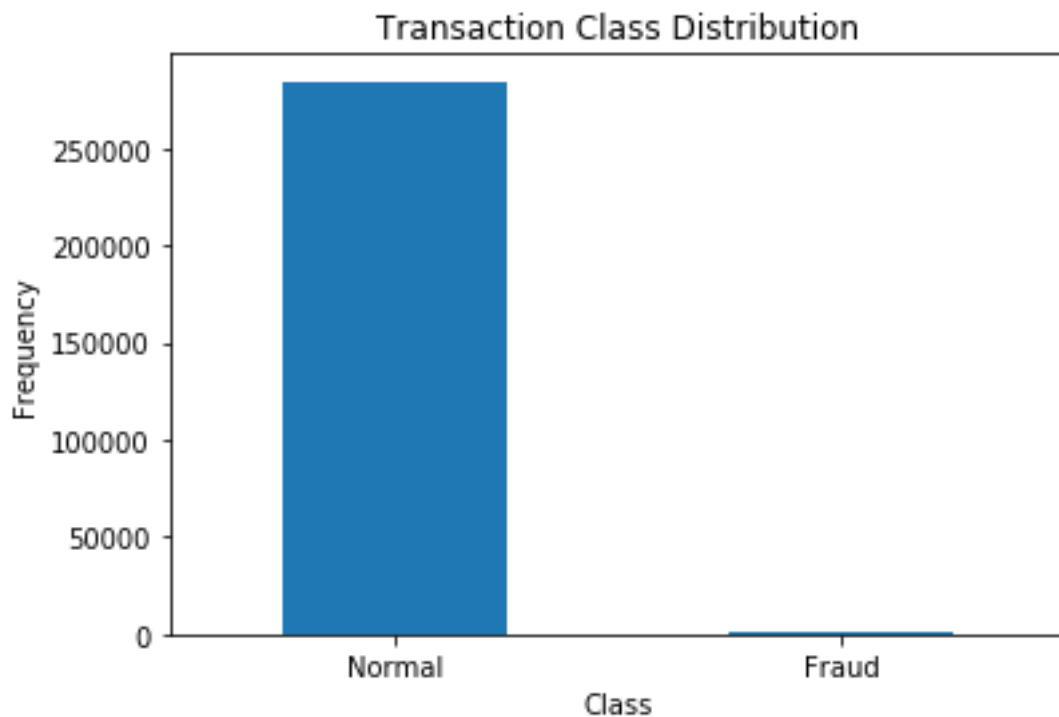


Figure 6.2 Transaction Class Distribution

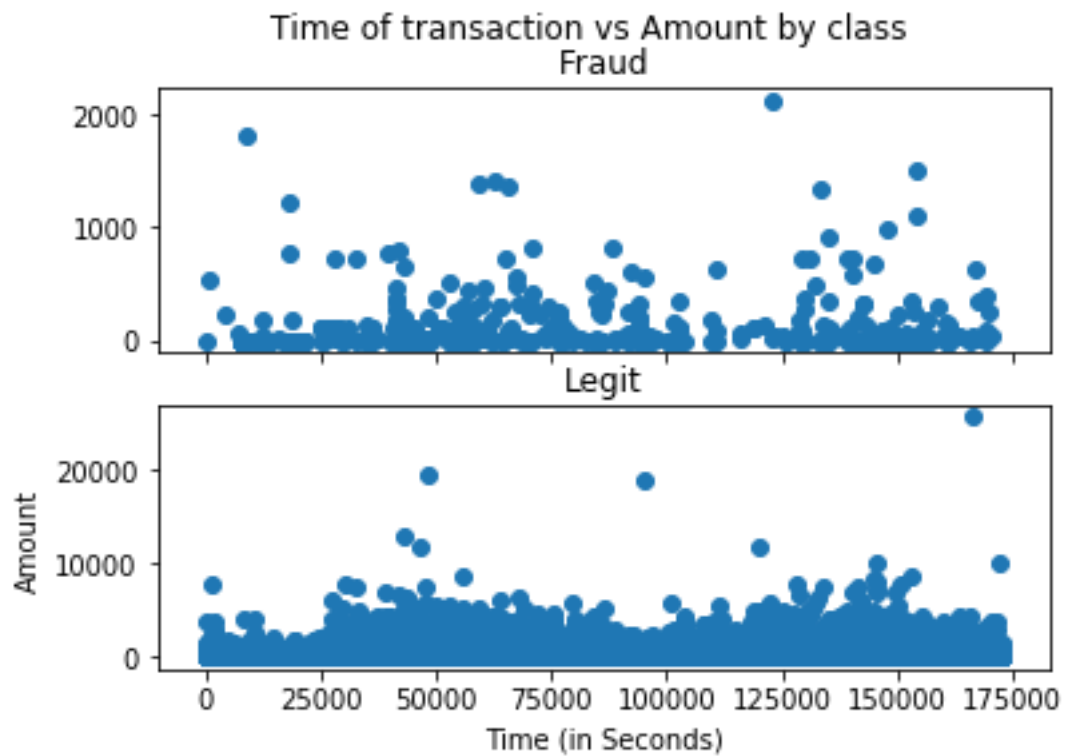


Figure 6.3 Time vs Amount

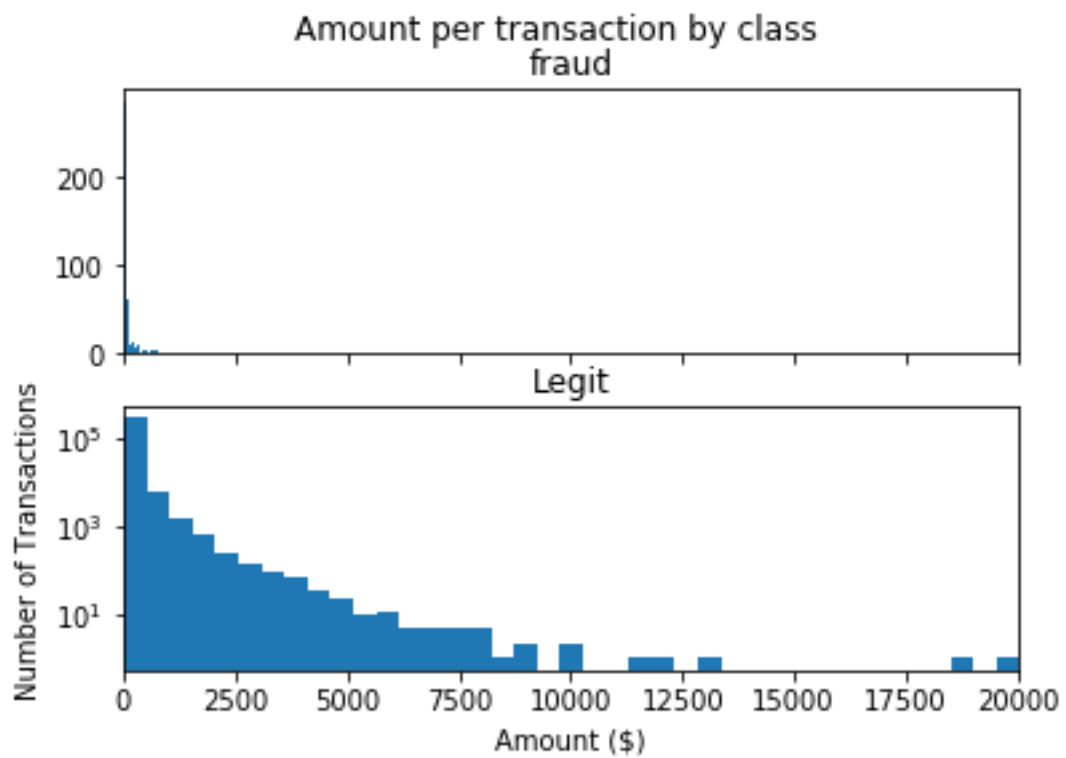


Figure 6.4 Amount per Transaction

CHAPTER 7

CONCLUSION

Clearly, credit card fraud is an act of criminal dishonesty. The detection of fraud is a vital research field. This issue opens door for many AI-based systems to counter fraud. This report has reviewed recent findings in the credit card fraud. This report has also identified the different types of frauds such as Credit Card Frauds: Online and Offline, Card Theft, Account Bankruptcy, Device Intrusion, Application Fraud, Counterfeit Card, Telecommunication Fraud. In this research, we have proposed a model with an accuracy of 99.95%, precision of 96.15% and recall of 0.76.

Advantages of Credit Card Fraud Detection System

- Fraud which is detected using existing purchase data of card holder is a way to reduce the rate of frauds.

Limitations

- The Credit Card Fraud Detection system uses Machine Learning to process. Machine learning needs data to train itself so it can predict well. But there isn't sufficient data for the system to train itself as original data of transactions cannot be available.
- Every day a new method of fraud is being used so the system cannot detect it.

CHAPTER 8

FUTURE SCOPE

As the next step in this research, the focus should be on getting a 100% accuracy and precision. It is possible when we implement more algorithms. We intend to enhance the performance and take the security and privacy of the data in real time into consideration.

REFERENCES

- [1] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277-14284, 2018.
- [2] Roy and J. Sun and R. Mahoney and L. Alonzi and S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," in *Systems and Information Engineering Design Symposium (SIEDS)*, pp. 129-134, 2018.
- [3] P. Save, P. Tiwarekar, K. N., and N. Mahyavanshi, —A Novel Idea for Credit Card Fraud Detection using Decision Tree, *Int. J. Comput. Appl.*, vol. 161, no. 13, pp. 6–9, 2017, doi: 10.5120/ijca2017913413.
- [4] Zarrabi, H. Kazemi, "Using deep networks for fraud detection in the credit card transaction," *IEEE 4th International Conference in Knowledge-Based Engineering and Innovation (KBEI)*, pp. 0630-0633, 2017.
- [5] S. Dutta, A. K. Gupta and N. Narayan, "Identity Crime Detection Using Data Mining," *3rd International Conference on Computational Intelligence and Networks (CINE)*, Odisha, pp. 1-5, 2017.
- [6] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," *International Conference on Intelligent Computing and Control (I2C2)*, Coimbatore, pp. 1-5, 2017.
- [7] D. Pojee, S. Zulphekari, F. Rarh, and V. Shah, "Secure and quick NFC payment with data mining and intelligent fraud detection," *2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, pp. 148-152, 2017.
- [8] D. S. Sisodia, N. K. Reddy and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, pp. 2747-2752, 2017.
- [9] L. Vergara, A. Salazar, J. Belda, G. Safont, S. Moral and S. Iglesias, "Signal processing on graphs for improving automatic credit card fraud detection," *International Carnahan Conference on Security Technology (ICCST)*, Madrid, pp. 1- 6, 2017.
- [10] Vimala Devi and K. S. Kavitha, —Fraud Detection in Credit Card Transactions by using Classification Algorithms, *Int. Conf. Curr. Trends Computer. Electr. Electron. Commun. CTCEEC* 2017, pp. 125–131, 2018, doi: 10.1109/CTCEEC.2017.8455091.
- [11] R. Popat and J. Chaudhary, —A Survey on Credit Card Fraud Detection Using Machine Learning, *Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI* 2018, no. Icoei, pp. 1120–1125, 2018, doi: 10.1109/ICOEI.2018.8553963.

- [12] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, —Random Forest for credit card fraud detection, || ICNSC 2018 - 15th IEEE Int. Conf. Networking, Sens. Control, pp. 1–6, 2018, doi: 10.1109/ICNSC.2018.8361343.
- [13] V. N. Dornadula and S. Geetha, —Credit Card Fraud Detection using Machine Learning Algorithms, || Procedia Comput. Sci., vol. 165, pp. 631–641, 2019, doi: 10.1016/j.procs.2020.01.057.
- [14] S. Mittal and S. Tyagi, —Performance evaluation of machine learning algorithms for credit card fraud detection, || Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2019, pp. 320–324, 2019, doi: 10.1109/CONFLUENCE.2019.8776925.
- [15] M. Deepa and D. Akila, —Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques, || Int. J. Innov. Res. Appl. Sci. Eng., vol. 3, no. 6, p. 483, 2019, doi: 10.29027/ijirase.v3.i6.2019.483-489.