VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgavi-590 018, Karnataka, India

A

PROJECT PHASE II REPORT

On

"DETECTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNING BASED ALGORITHM"

Submitted in Partial Fulfillment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By

BELLAM NARENDRA NATH	**1SJ18CS009**
MANJUNATH	**1SJ18CS055**
GOVINDA N	**1SJ19CS403**
H.V NAVEEN KUMAR	**1SJ19CS404**

Carried out at
B G S R&D Center,
Dept of CSE,
SJCIT

Under the guidance of

Prof. Swetha T (B E, MTech (Ph.D))
Assistant Professor
Dept of CSE, SJCIT

S J C INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHIKKABALLAPUR-562 101

2021-2022

||Jai Sri Gurudev||

Sri Adichunchanagiri Shikshana Trust®

S J C INSTITUTE OF TECHNOLOGY, CHICKBALLAPUR – 562101

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the technical Seminar work entitled "DETECTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNING BASED ALGORITHM" is a bonafied workcarried out by **BELLAM NARENDRA NATH(1SJ18CS009) MANJUNATH(1SJ18CS055) GOVINDA N(1SJ19CS403) H.V NAVEEN KUMAR(1SJ19CS404)** in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering in Eighth Semester of the Visvesvaraya Technological University, Belagavi** during the year **2021-2022.** It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements with respect to project work prescribed for the Bachelor of Engineering degree.

Signature of Guide	**Signature of HOD**	**Signature of Principal**
Prof. Swetha T	**Dr.Manjunatha Kumar B H**	**Dr. G T Raju**
Assistant Professor	Professor & HOD,	Principal SJCIT,
Dept. of CSE, SJCIT	Dept. of CSE, SJCIT	Dept. of CSE, SJCIT

Professor & HOD, Department of Computer Science & Engg., S.J.C. Institute of Technology, Chickballapur-562 101

Principal, S J C Institute of Technology, Chickballapur - 562 101

External Examiners:

Name of the Examiners	**Signature with Date**
1.....	.....
2.....	.....

DECLARATION

We **BELLAM NARENDRA NATH(1SJ18CS009) MANJUNATH(1SJ18CS055) GOVINDA N(1SJ19CS403) H.V NAVEEN KUMAR(1SJ19CS404)** Student of VIII Semester B.E in Computer Science and Engineering at S J C Institute of Technology, Chickaballapur, here by declare that this dissertation work entitled "**DETECTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNINGBASED ALGORITHM"** has been carried out at B.G.S R&D Center, Dept. of CSE, SJCIT Under the Guidance of guide **Prof Swetha T Assistant Professor, Dept. of CSE, S J C Institute of Technology, Chickaballapur** and submitted in the partial fulfillment for the award of Degree Bachelor of Engineering in the Computer Science and Engineering of Visvesvaraya Technological University, Belagavi during the academic year 2021-2022. We further declare that the report had not been submitted to another University forthe award of any other degree.

Place: Chickaballapur
Date:

BELLAM NARENDRA NATH (1SJ18CS009)
MANJUNATH (1SJ18CS055)
GOVINDA N (1SJ19CS403)
H.V NAVEEN KUMAR (1SJ19CS404)

ABSTRACT

As the world is rapidly moving towards digitization and money transactions are becoming cashless, the use of credit cards has rapidly increased. The fraud activities associated with it have also been increasing which leads to a huge loss to the financial institutions. Therefore, weneed to analyze and detect the fraudulent transaction from the non-fraudulent ones. In this wepresent a comprehensive review of various methods used to detect credit card frauds. Here weimplement different machine learning algorithms on an imbalanced dataset such as logistic regression, naïvebayes, random forest with ensemble classifiers using boosting technique. An extensive review is done on the existing and proposed models for credit card fraud detection and has done a comparative study on these techniques. So Different classification models are applied to the data and the model performance is evaluated on the basis of quantitative measurements such as accuracy, precision, recall, f1 score, support, confusion matrix.

ACKNOWLEDGEMENT

With reverential pranam, we express our sincere gratitude and salutations to the feetof his holiness **Paramapoojya Jagadguru Byravaikya Padmabhushana Sri Sri Sri Dr. Balagangadharanatha Maha Swamiji,** his holiness **Paramapoojya Jagadguru Sri Sri Sri Dr. Nirmalanandanatha Maha Swamiji** and **Pramapoojya Sri Sri Mangalnatha Swamiji ,** Sri Adichunchanagiri Mutt for their unlimited blessings.

First and foremost, we wish to express our deep sincere feelings of gratitude to our institution, **Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology,** for providing us an opportunity for completing our Project Work Phase - II successfully.

We extend deep sense of sincere gratitude to **Dr. G T Raju**, **Principal, S J C Institute of Technology, Chickballapur**, for providing an opportunity to complete the Project Work Phase - II.

We extend special in-depth, heartfelt, and sincere gratitude to HOD **Dr. Manjunatha Kumar B H, Head of the Department, Computer Science and Engineering, S J C Institute of Technology, Chickballapur,** for his constant support andvaluable guidance of the Project Work Phase - II.

We convey our sincere thanks to Project Guide **Prof. Swetha T, Assistant Professor Department of Computer Science and Engineering, S J C Institute of Technology,** for his constant support, valuable guidance and suggestions of the Project Work Phase - II.

We also feel immense pleasure to express deep and profound gratitude to Project coordinators **Prof. PradeepKumar G M and Prof Shrihari M R, Assistant Professors, Department of Computer Science and Engineering, S J C Institute of Technology,** fortheir guidance and suggestions of the Project Work phase-II.

Finally, we would like to thank all faculty members of Department of Computer Science and Engineering, S J C Institute of Technology, Chickballapur for their support.

We also thank all those who extended their support and cooperation while bringingout this project work phase- II.

BELLAM NARENDRA NATH(1SJ18CS009)

MANJUNATH (1SJ18CS055)
GOVINDA N(1SJ19CS403)
H.V NAVEEN KUMAR(1SJ19CS404)

CONTENT

LIST OF FIGURES

CHAPTER-1

INTRODUCTION

1.1 Overview

Credit card is the most popular mode of payment. As the number of credit card users is rising world-wide, the identity theft is increased, and frauds are also increasing. In the virtual card purchase, only the card information is required such as card number, expiration date, secure code, etc. Such purchases are normally done on the Internet or over telephone. To commit fraud in these types of purchases, a person simply needs to know the card details.The mode of payment for online purchase is mostly done by credit card. The details of credit card should be kept private. To secure credit card privacy, the details should not be leaked. Different ways to steal credit card details are phishing websites, steal/lost credit cards, counterfeit credit cards, theft of card details, intercepted cards etc. For security purpose, the above things should be avoided. In online fraud, the transaction is made remotely and only the card's detailsare needed. The simple way to detect this type of fraud is to analyze the spending patterns on every card and to figure out any variation to the "usual" spending patterns.Fraud detection by analyzing the existing data purchase of cardholder is the best wayto reduce the rate of successful credit card frauds. As the data sets are not available and also the results are not disclosed to the public. The fraud cases should be detectedfrom the available data sets known as the logged dataand user behavior. At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence.

1.2 Problem Statement

The card holder faced a lot of trouble before the investigation finish. And also, as all the transaction is maintained in a log, we need to maintain huge data, and also now a day's lotof online purchase are made so we don't know the person how is using the card online, we just capture the ip address for verification purpose. So there need a help from the cyber- crime to investigate the fraud.

1.3 Significance and Relevance of Work

Relevance of work includes consideration of all the possible ways to provide a solution to given problem. The proposed solution should satisfy all the user requirements and should be flexible enough so that future changes can easily done based on the future upcoming requirements like Machine learning techniques.

There are two important categories of machine learning techniques to identify the frauds in credit card transactions: supervised and unsupervised learning model. In supervised approach, early transactions of credit card are labelled as genuine or frauds. Then, the scheme identifies the fraud transaction with credit card data.

1.4 Objectives

Features Extractions from recognized facial information then data will be normalized for extracting features of good Objective of the project is to predict the fraud and fraud less transaction with respect to the time and amount of the transaction using classification machine learning algorithms such as SVM, Random Forest, Decision tree and confusion matrix in building of the complex machine learning models.

1.5 Methodology

First the Dataset is read. Exploratory Data Analysis is performed on the dataset to clearly understand the statistics of the data, Feature selection is used, A machine learning model is developed. Train and test the model and analysis the performance of the model using certainevaluation techniques such as accuracy, confusion matrix, precision etc.

1.6 Organization of the report

Chapter 1

1. **Overview:** the overview provides the basic layout and the insight about the work proposed. It briefs the entire need of the currently proposed work.
2. **Problem statement:** A problem statement is a concise description of an issue tobe addressed or a condition to be improved upon. We have identifiedthe gap between addressed or a condition to be improved upon.

3. **Significance and Relevance of Work:** We have mentioned about the contribution of our work to the society.
4. **Objectives:** A project objective describes the desired results of the work. We have mentioned about the work we are trying to accomplish in this section.
5. **Methodology:** A methodology is a collection of methods, practices, processes and techniques. We have explained in this section about the working of the project in a briefway.

Chapter 2

1. **Literature Survey:** the purpose of a literature review is to gain an understanding of theexisting resources to a particular topic or area of study. We have referred to many research papers relevant to our work in a better way.

Chapter 3

1. **System Requirements and Specifications:** System Requirements and Specifications is a document that describes the nature of a project, software or application. This sectioncontains the brief knowledge about the functional and non – functional that are needed to implement the project.

Chapter 4

1. **System Analysis:** System Analysis is a document that describes about the existing system and proposed system in the project. And also describes about advantages and disadvantages in the project.

Chapter 5

1. **System design:** System design is a document that describes about the project modules,Activity diagram, Use Case Diagram, Data Flow Diagram, and Sequence Diagram detailed in the project.

Chapter 6

1. **Implementation:** Implementation is a document that describes about the detailed concepts of the project. Also describes about the algorithm with their detailed steps. And also, about the codes for implementation of the algorithm.

Chapter 7

1. Testing: Testing is a document that describes about the
   a. **Methods of testing:** This contains the information about Unit testing, Validation testing, Functional testing, Integration testing, User Acceptancetesting.
   b. **Test Cases:** In Test Cases we contain the detailed description about program Testcases.

Chapter 8

1. **Performance Analysis:** Performance Analysis is a document that describes about the study system in detailed.

Chapter 9

1. **Conclusion and Future Enhancement:** Conclusion and Future Enhancement is a document that describes about the brief summary of the project and undetermined events that will occur in that time.

CHAPTER – 2

LITERATURE SURVEY

2.1 Credit Card Fraud Detection Techniques : Data and Technique OrientedPerspective

Authors: SamanehSorournejad, Zahra Zojaji, Amir Hassan Monadjemi.

In this paper, after investigating difficulties of credit card fraud detection, we seek to reviewthe state ofthe art in credit card fraud detection techniques, datasets and evaluation criteria.

Disadvantages

- Lack of standard metrics

2.2 Detection of credit card fraud: State of art

Authors: Imane Sadgali, Nawal Sael, Faouzia Benabbau

In this paper, we propose a state of the art on various techniques of credit card fraud detection. The purpose of this study is to give a review of implemented techniques for creditcard fraud detection, analyses their incomes and limitless, and synthesize the finding in order to identify the techniques and methods that give the best results so far.

Disadvantages

- Lack of adaptability

2.3 Credit card fraud detection using machine learning algorithm

Authors: Vaishnavi Nath Dornadulaa, Geetha S.

The main aim of the paper is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyze the past transaction details of the customers and extract the behavioral patterns.

Disadvantages

- Imbalanced Data

2.4 Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning techniques

Authors: Debachudamani Prusti, Santanu Kumar Rath

In this study, the application of various classification models is proposed by implementing machinelearning techniques to find out the accuracy and other performance parameters to identify the fraudulent transaction.

Disadvantages

- **Overlapping data.**

2.5 Detection of Credit Card Fraud Transactions using Machine Learning Algorithms andNeural Networks

Authors: Deepti Dighe, Sneha Patil, Shrikant Kokate

Credit card fraud resulting from misuse of the system is defined as theft or misuse of one's credit card information which is used for personal gains without the permission of the card holder. To detect such frauds, it is important to check the usage patterns of a user over the past transactions. Comparing the usage pattern and current transaction, we can classify it as either fraud or a legitimate transaction.

Disadvantages

- Different misclassification importance

2.6 Credit card fraud detection using machine learning algorithms andcyber security Authors: JiatongShen

As they have the same accuracy the time factor is considered to choose the best algorithm. Byconsideringthe time factor they concluded that the Adaboost algorithm works well to detect credit card fraud.

Disadvantages

- Accuracy is not getting perfectly

CHAPTER-3

SYSTEM REQUIREMENTS AND SPECIFICATION

3.1 System Requirement Specification:

System Requirement Specification (SRS) is a fundamental document, which forms the foundation of the software development process. The System Requirements Specification (SRS) document describes all data, functional and behavioral requirements of the software under production or development. An SRS is basically an organization's understanding (in writing) of a customer or potential client's system requirements and dependencies at a particular point in time (usually) prior to any actual design or development work. It's a two- way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time. The SRS also functionsas a blueprint for completing a project with as little cost growth as possible. The SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it.It is important to note that an SRS contains functional and non-functional requirements only. It doesn't offer design suggestions, possible solutions to technology or business issues, or any other information other than what the development team understands the customer's system requirements.

3.2 Hardware specification

- RAM: 4GB and Higher
- Processor: intel i3 and above
- Hard Disk: 500GB: Minimum

3.3 Software specification

- OS: Windows or Linux
- Python IDE: python 2.7.x and above
- Jupyter Notebook
- Language: Python

3.4 Functional Requirements:

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality. In this system following are the functional requirements:

- Collect the Datasets.
- Train the Model.
- Predict the results

3.5 Non-Functional Requirements

- The system should be easy to maintain.
- The system should be compatible with different platforms.
- The system should be fast as customers always need speed.
- The system should be accessible to online users.
- The system should be easy to learn by both sophisticated and novice users.
- The system should provide easy, navigable and user-friendly interfaces.
- The system should produce reports in different forms such as tables and graphs foreasy visualization by management.
- The system should have a standard graphical user interface that allows for the online

3.6 Performance Requirement:

Performance is measured in terms of the output provided by the application. Requirement specification plays an important part in the analysis of a system. Only when the requirement specifications are properly given, it is possible to design a system, which will fit into required environment. It rests largely with the users of the existing system to give the requirement specifications because they are the people who finally use the system. This is because the requirements have to be known during the initial stages so that the system can be designed according to those requirements. It is very difficult to change the system onceit has been designed and on the other hand designing a system, which does not cater to the requirements of the user, is of no use.

CHAPTER-4

SYSTEM ANALYSIS

Systems analysis is the process by which an individual studies a system such that an information system can be analyzed, modeled, and a logical alternative can be chosen. Systems analysis projects are initiated for three reasons: problems, opportunities, and directives

4.1 Existing System

- Since the credit card fraud detection system is a highly researched field, there are many different algorithms and techniques for performing the credit card fraud detection system.
- One of the earliest systems is CCFD system using Markov model. Some other various existing algorithms used in the credit cards fraud detection system includes Cost sensitivedecision tree (CSDT).
- credit card fraud detection (CCFD) is also proposed by using neural networks. The existing credit card fraud detection system using neural network follows the whale swarmoptimization algorithm to obtain an incentive value.
- It the uses BP network to rectify the values which are found error.

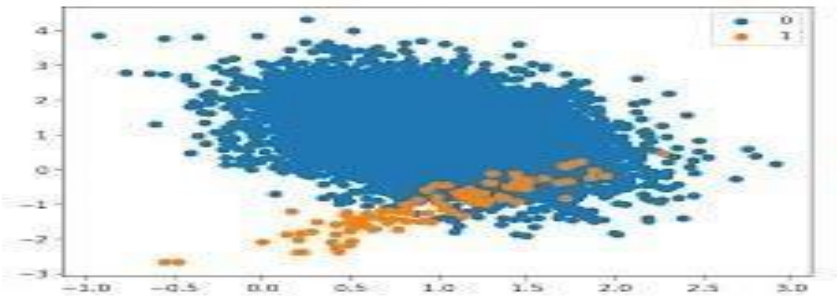


Figure 4.1.1 fraud and Non Fraud Representation

4.1.1 Limitations

- If the time interval is too short, then Markov models are inappropriate because the individual displacements are not random, but rather are deterministically related in time. This example suggests that Markov models are generally inappropriate over sufficiently short time intervals.

4.2 Proposed System

Support Vector Machine:

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane Training regression model and finding out the best one.

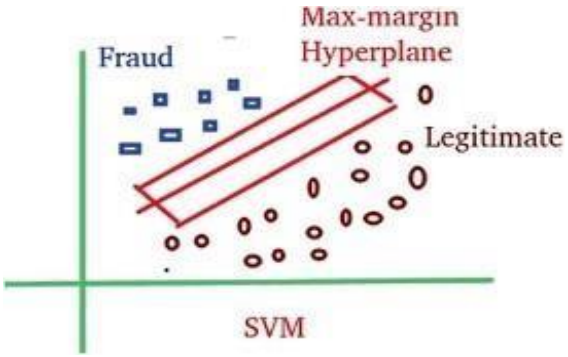


Fig 4.2.1 SVM Representation

Random Forest Classifier

Features are cheekbone to jaw width, width to upper facial height ratio, perimeter to area ratio, eye size, lower face to face height ratio, face width to lower face heightratio and mean of eyebrow height. The extracted features are normalized and finally subjected to support regression.

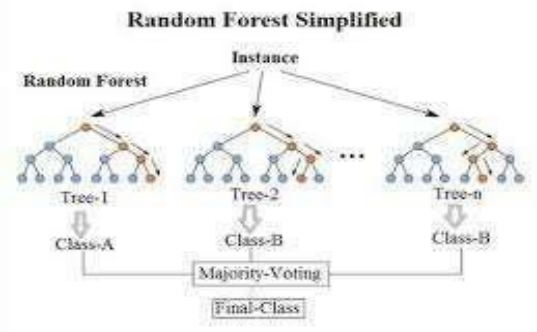
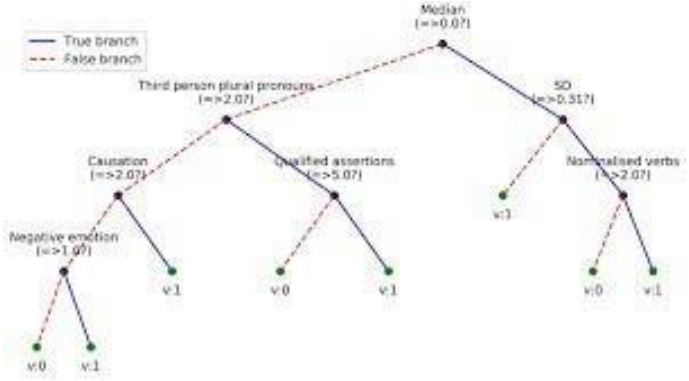


Fig 4.2.2 Simplified Random Forest algorithm

Decision Tree

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a formof supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.



4.2.3 Decision tree Algorithm

4.2.1 Advantages

- Support vector machine works comparably well when there is an understandablemargin of dissociation between classes.
- SVM is effective in instances where the number of dimensions is larger than thenumber of specimens.
- Simple to understand and to interpret.
- Requires little data preparation.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of datapoints used to train the tree.
- Able to handle both numerical and categorical data.
- Random forest classifier can be used to solve for regression or classification problems.
- The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample.

CHAPTER-5

SYSTEM DESIGN

5.1 Project Modules

Entire project is divided into 3 modules as follows:

Data Gathering and pre processing

Training the model using following Machine Learning algorithms

i. SVM
ii. Random Forest Classifier
iii. Decision Tree

Module 1: Data Gathering and Data Pre processing

a. A proper dataset is searched among various available ones and finalized with the dataset.
b. The dataset must be preprocessed to train the model.
c. In the preprocessing phase, the dataset is cleaned and any redundant values, noisy dataandnull values are removed.
d. The Preprocessed data is provided as input to the module.

Module 2: Training the model

a. The Preprocessed data is split into training and testing datasets in the 80:20 ratio to avoidthe problems of over-fitting and under-fitting.
b. A model is trained using the training dataset with the following algorithms SVM, Random Forest Classifier and Decision Tree
c. The trained models are trained with the testing data and results are visualized using bar graphs, scatter plots.
d. The accuracy rates of each algorithm are calculated using different params like F1 score, Precision, Recall. The results are then displayed using various data visualization tools for analysis purpose.
e. The algorithm which has provided the better accuracy rate compared to remaining algorithms is taken as final prediction model.

Module 3: Final Prediction model integrated with front end

a. The algorithm which has provided better accuracy rate has considered as the finalprediction model.

b. The model thus made is integrated with front end.

c. Database is connected to the front end to store the user information who are using it.

SYSTEM ARCHITECTURE

Our Project main purpose is to making Credit Card Fraud Detection awaring to people from credit card online frauds. the main point of credit card fraud detection system is necessaryto safe our transactions & security. With this system, fraudsters don't have the chance to make multiple transactions on a stolen or counterfeit card before the cardholder is aware of the fraudulent activity. This model is then used to identify whether a new transaction is fraudulent or not. Our aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

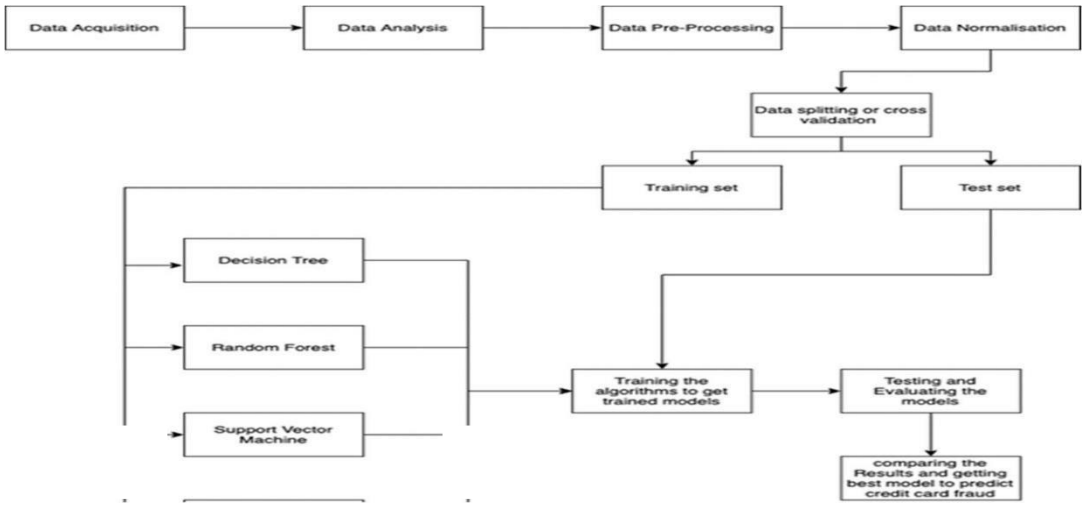


Fig 5.1 System Architecture

5.2 Activity diagram

Activity diagram is an important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity toanother activity. The activity can be described as an operation of the system. The control flow is drawnfrom one operation to another. This flow can be sequential, branched, or concurrent. Activity diagramsdeal with all type of flow control by using different elements such as fork, join, etc. The basic purposesof activity diagram are it captures the dynamic behavior of the system. Activity diagramis used to showmessage flow from one activity to another Activity is a particular operation of the system. Activitydiagrams are not only used for visualizing the dynamic nature of a system, but they are also used to construct the executable systemby using forward and reverse engineering techniques.The only missing thing in the activity diagram is the message part.

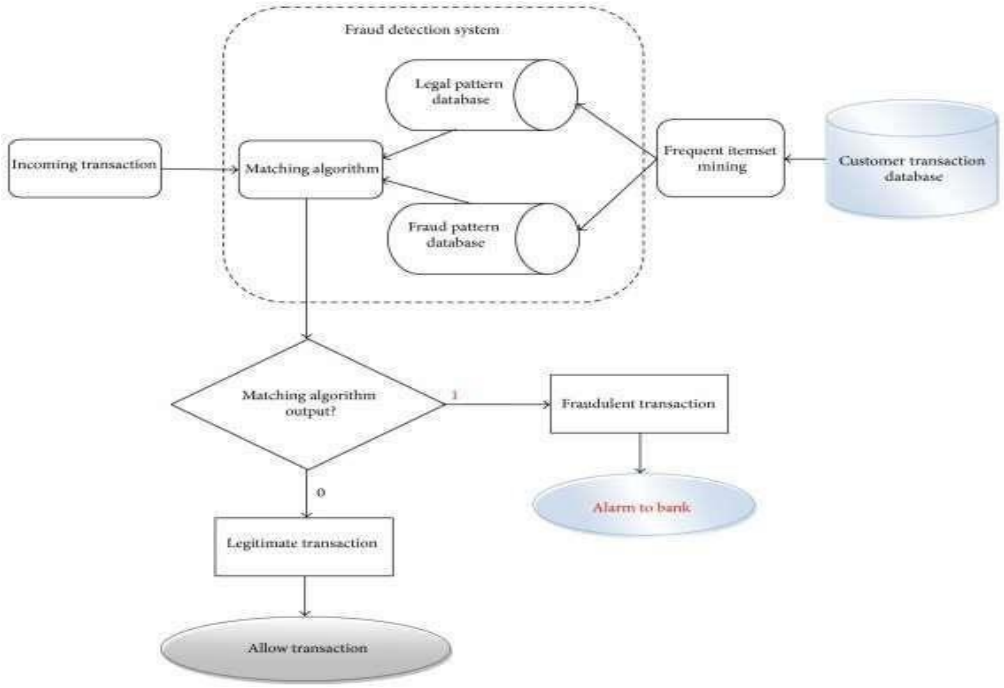


Fig 5.2 Activity Diagram

5.3 Use case diagram

In UML, use-case diagrams model the behavior of a system and help to capture the requirements of the system. Use-case diagrams describe the high-level functions and scope of a system. These diagrams also identify the interactions between the system and its actors. The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not how the system operates internally. Use-case diagrams illustrate and define the context and requirements of either an entire system or the important parts of the system. You can model a complex system with a single use-case diagram, or create many use-case diagrams to model the components of the system. You would typically develop use- case diagrams in the early phases of a project and refer to them throughout the development process.

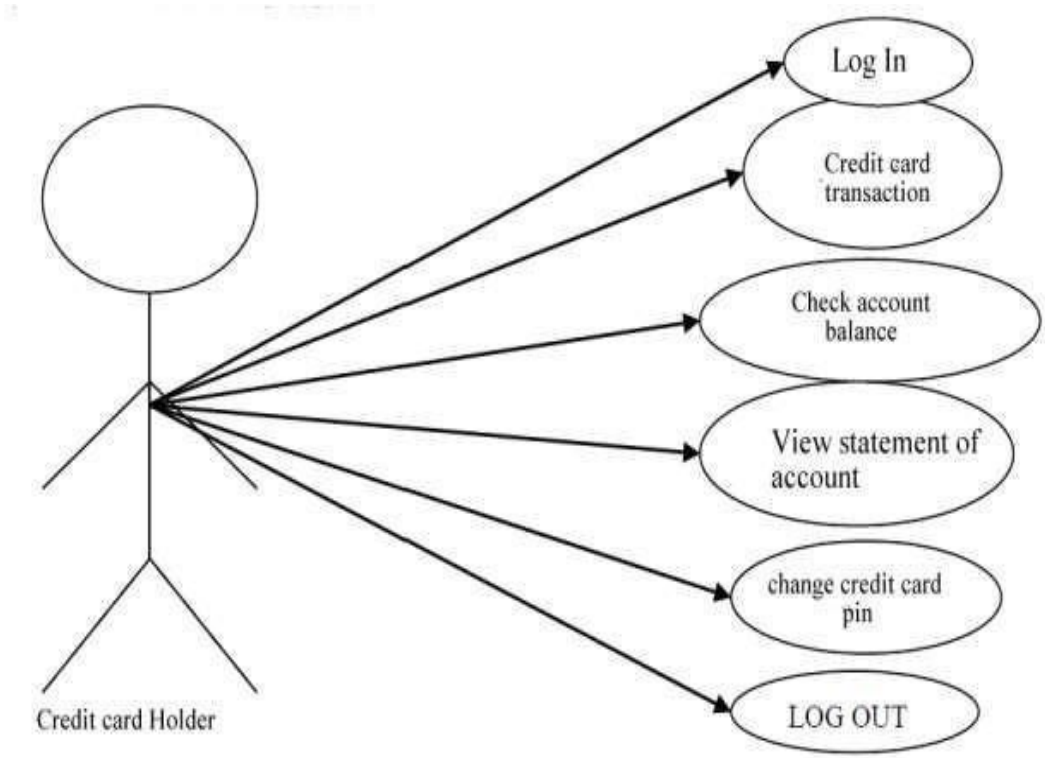


Fig 5.3 Use case Diagram

5.4 Sequence Diagram

The sequence diagram represents the flow of messages in the system and is also termed as an event diagram. It helps in envisioning several dynamic scenarios. It portrays the communication between any two lifelines as a time-ordered sequence of events, such that these lifelines took part at the run time. In UML, the lifeline is represented by a vertical bar, whereas the message flow is represented by a vertical dotted line that extends across the bottom of the page. It incorporates the iterations as well as branching.

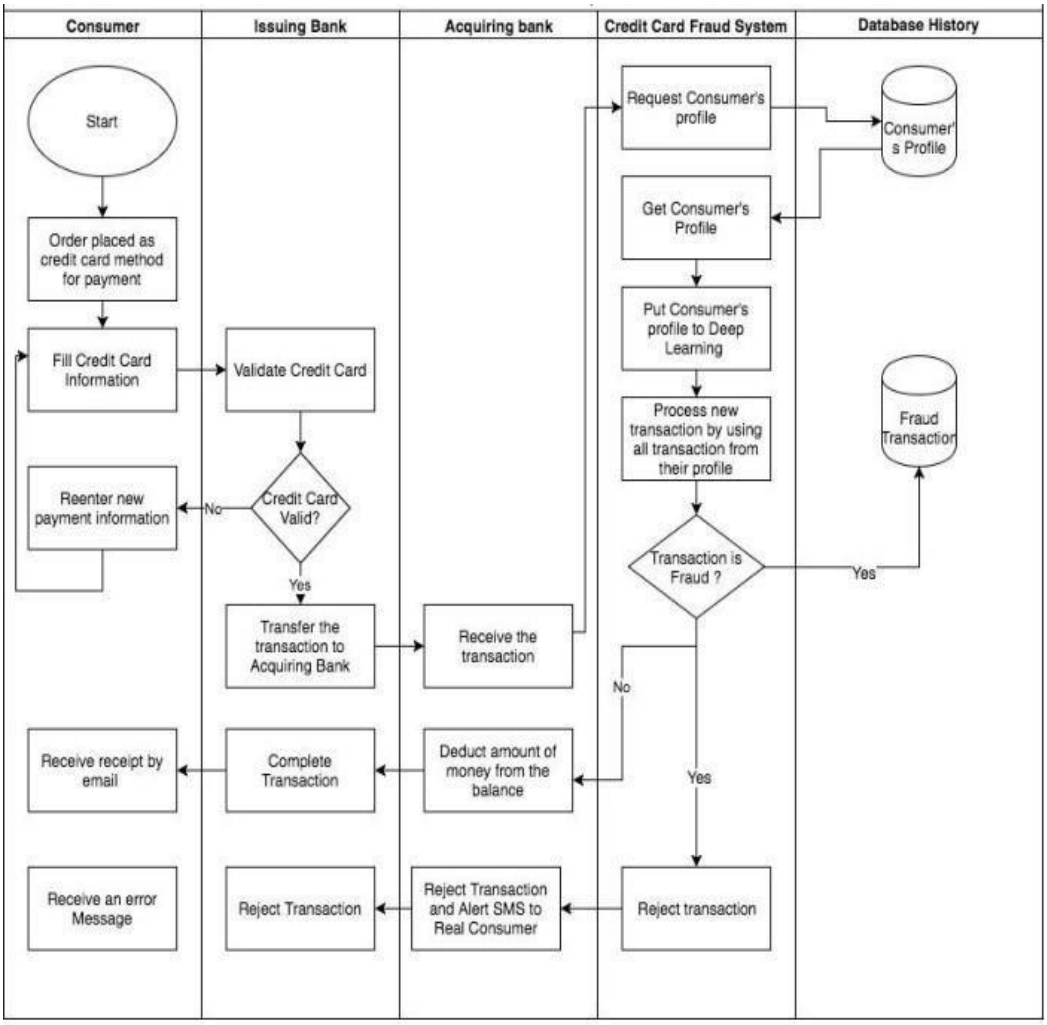


Fig 5.4 Sequence diagram

5.5 Data Flow Diagram

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirementgraphically. It can be manual, automated, or a combination of both. It shows how data entersand leaves the system, what changes the information, and where data is stored. The objectiveof a DFD is to show the scope and boundaries of a system as a whole. It may be used as a communication tool between a system analyst and any person who plays a part in the order that acts as a starting point for redesigning a system. The DFD is also called as a data flow graph or bubble chart.

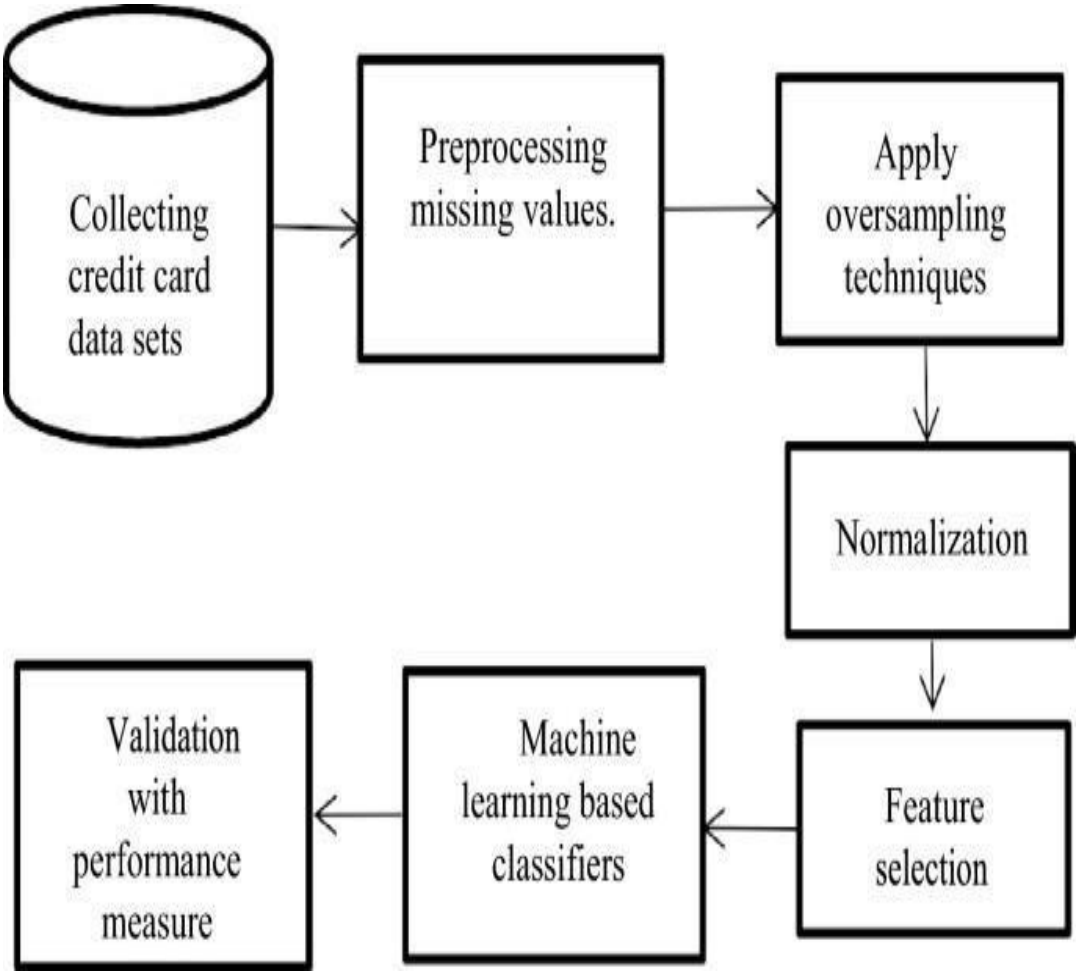


Fig 5.5 Data Flow diagram

CHAPTER-6

IMPLEMENTATION

6.1 Algorithm

Step 1: Import dataset

Step 2: Convert the data into data frames format Step3:

Do random oversampling using ROSE package

Step4: Decide the amount of data for training data and testing data

Step5: Give 80% data for training and remaining data for testing.

Step6: Assign train dataset to the models

Step7: Choose the algorithm among 3 different algorithms and create the model

Step8: Make predictions for test dataset for each algorithm

Step9: Calculate accuracy for each algorithm

Step10: Apply confusion matrix for each variable

Step11: Compare the algorithms for all the variables and find out the best algorithm.

CODE: -

Importing Libraries

```
!pip install tensorflow
# for numerical operations
import numpy as np
# to store and analysis data in dataframes
import pandas as pd
# data visualization
import matplotlib.pyplot as plt
import seaborn as sns
# python modules for data normalization and splitting
from sklearn.preprocessing import RobustScaler
from sklearn.model_selection import train_test_split
# python modules for creating training and testing ml algorithms
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
# python modules for creating training and testing Neural Networks
import tensorflow as tf
from tensorflow.keras.models import load_model
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dropout,Dense
# evaluation
From sklearn.metrics import accuracy_score,confusion_matrix,classification_report,precision_score,recall_score, f1_score,roc_auc_score
import systemcheck
```

Data Acquisition

```
data = pd.read_csv('creditcard.csv')
data
```

Data Analysis

```
data.shape
data.info()
data.describe()
sns.countplot(x='Class', data=data)
print("Fraud: ",data.Class.sum()/data.Class.count())
Fraud_class = pd.DataFrame({'Fraud': data['Class']})
Fraud_class. apply(pd.value_counts). plot(kind='pie',subplots=True)
fraud = data[data['Class'] == 1]
valid = data[data['Class'] == 0]
fraud.Amount.describe()
plt.figure(figsize=(20,20))
plt.title('Correlation Matrix', y=1.05, size=15)
sns.heatmap(data.astype(float).corr(),linewidths=0.1,vmax=1.0,
        square=True, linecolor='white', annot=True)
```

Data Normalization

```
rs = RobustScaler()
data['Amount'] = rs.fit_transform(data['Amount'].values.reshape(-1, 1))
data['Time'] = rs.fit_transform(data['Time'].values.reshape(-1, 1))
data
```

Considering inputs columns and output column

```
X = data.drop(['Class'], axis = 1)
Y = data["Class"]
```

Data splitting

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 1)
X_train
X_test
Y_test
```

```
def evaluate(Y_test, Y_pred):
print("Accuracy: ",accuracy_score(Y_test, Y_pred))
print("Precision: ",precision_score(Y_test, Y_pred))
print("Recall: ",recall_score(Y_test, Y_pred))
print("F1-Score: ",f1_score(Y_test, Y_pred))
print("AUC score: ",roc_auc_score(Y_test, Y_pred))
print(classification_report(Y_test, Y_pred, target_names = ['Normal', 'Fraud']))
conf_matrix = confusion_matrix(Y_test, Y_pred)
plt.figure(figsize =(6, 6))
sns.heatmap(conf_matrix, xticklabels = ['Normal', 'Fraud'],
yticklabels = ['Normal', 'Fraud'], annot = True, fmt ="d");
plt.title("Confusion matrix")
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.show()
```

Creating algorithms, Training, Testing and Evaluating

Creating Support Vector Classifier

```
svm = SVC()
# Training SVC
svm.fit(X_train, Y_train)
# Testing SVC
Y_pred_svm = svm.predict(X_test)
# Evaluating SVC
evaluate(Y_pred_svm, Y_test)
```

Random forest model creation

```
rfc = RandomForestClassifier()
# training
rfc.fit(X_train, Y_train)
# Testing
Y_pred_rf = rfc.predict(X_test)
```

```
# Evaluation
evaluate(Y_pred_rf, Y_test)
```

Decision tree model creation

```
dtc = DecisionTreeClassifier()
dtc.fit(X_train, Y_train)
# predictions
Y_pred_dt_i = dtc.predict(X_test)
evaluate(Y_pred_dt_i, Y_test)
```

#Random forest balanced weights

```
from sklearn.ensemble import  RandomForestClassifier
# random forest model creation
rfb = RandomForestClassifier(class_weight='balanced')
rfb.fit(X_train, Y_train)
# predictions
Y_pred_rf_b = rfb.predict(X_test)
evaluate(Y_pred_rf_b, Y_test
```

CHAPTER-7

TESTING

Testing is a process of executing a program with intent of finding an error. Testing presents an interesting anomaly for the software engineering. The goal of the software testing is to convince system developer and customers that the software is good enough for operational use. Testing is a process intended to build confidence in the software. Testing is a set of activities that can be planned in advance and conducted systematically. Software testing is often referred to as verification & validation.

7.1 Unit Testing

In this testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In this testing step each module is found to working satisfactorily as regard to the expected output from the module. There are some validation checks for fields also. It is very easy to find error debut in the system.

7.2 Validation Testing

At the culmination of the black box testing, software is completely assembled as a package, interfacing errors have been uncovered and corrected and a final series of software tests. Askingthe user about the format required by system tests the output displayed or generated by the systemunder consideration. Here the output format is considered the of screen display. The output formaton the screen is found to be correct as the format was designed in the system phase according to the user need. For the hard copy also, the output comes out as specified by the user. Hence the outputtesting does not result in any correction in the system.

7.3 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input: identified classes of valid input must be accepted.

Invalid Input: identified classes of invalid input must be rejected.

Functions: identified functions must be exercised.

Output: identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.4 Integration Testing

Data can be lost across an interface; one module can have an adverse effort on the other sub functions when combined may not produces the desired major functions. Integrated testingis the systematic testing for constructing the uncover errors within the interface. The testing was done with sample data. The Developed system has run successfully for this sample data. The need for integrated test is to find the overall system performance.

7.5 User acceptance testing

User Acceptance Testing is a critical phase of any project and requires significant participation bythe end user. It also ensures that the system meets the functional requirements. Some of my friendswere who tested this module suggested that this was really a user-friendly application and givinggood processing speed.

CHAPTER-8

PERFORMANCE ANALYSIS

4.1 Performance metrics:

The basic performance measures derived from the confusion matrix. The confusion matrix is a 2 by 2 matrix table contains four outcomes produced by the binary classifier. Various measures such as sensitivity, specificity, accuracy and error rate are derived from theconfusion matrix.

Accuracy: Accuracy is calculated as the total number of two correct predictions(A+B) divided by thetotal number of the dataset(C+D). It is calculated as (1-error rate).

Accuracy=A+B/C+D

Whereas,

A=True Positive B=True Negative

C=Positive D=Negative

Error rate:

Error rate is calculated as the total number of two incorrect predictions(F+E) divided by thetotal number of the dataset(C+D).

Error rate=F+E/C+D

Whereas,

E=False Positive

F=False Negative

C=Positive

D=Negative

Sensitivity:

Sensitivity is calculated as the number of correct positive predictions(A) divided by thetotal number of positives(C).

Sensitivity=A/C

Specificity: Specificity is calculated as the number of correct negative predictions(B) divided by the total number of negatives(D).

Specificity=B/D.

DATA ANALYSIS

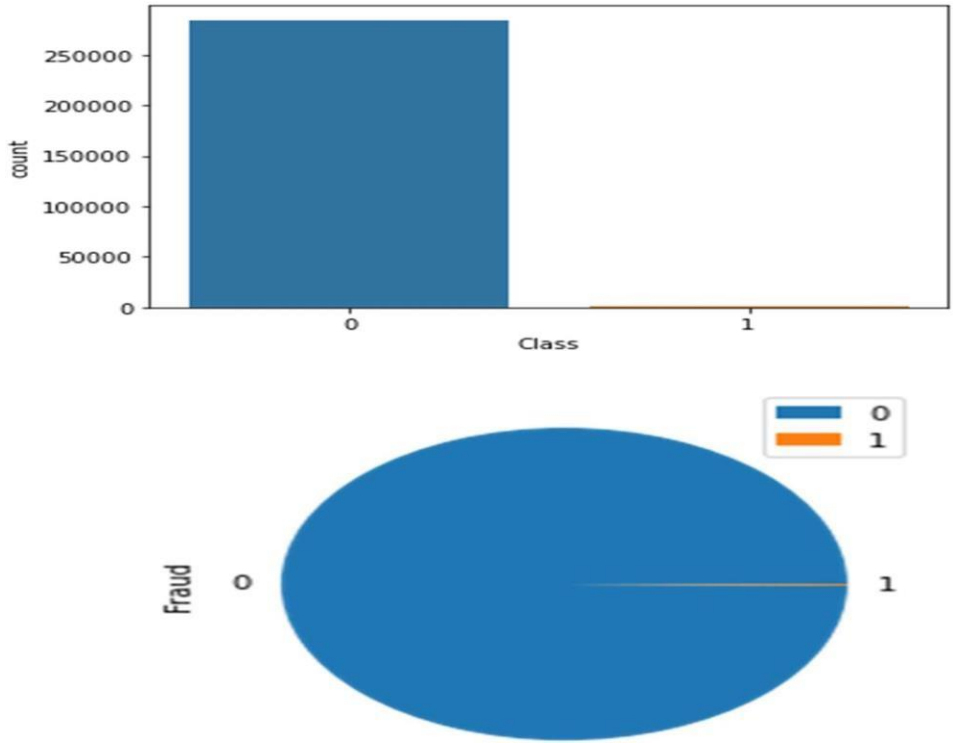


Fig 8.1 Dataset analysis

SUPPORT VECTOR MACHINE

Accuracy: 0.9994557775359011

Precision: 0.6781609195402298

Recall: 0.9516129032258065

F1-Score: 0.7919463087248322

AUC score: 0.975560405918703

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	56900
Fraud	0.68	0.95	0.79	62
accuracy			1.00	56962
macro avg	0.84	0.98	0.90	56962
weighted avg	1.00	1.00	1.00	56962

RANDOM FOREST

Accuracy: 0.9995611109160493

Precision: 0.7701149425287356

Recall: 0.9305555555555556

F1-Score: 0.8427672955974842

AUC score: 0.9651019999609383

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	56890
Fraud	0.77	0.93	0.84	72
accuracy			1.00	56962
macro avg	0.89	0.97	0.92	56962
weighted avg	1.00	1.00	1.00	56962

DECISION TREE

Accuracy: 0.9992802219023208

Precision: 0.7241379310344828

Recall: 0.7875

F1-Score: 0.7544910179640718

AUC score: 0.8935390369536936

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	56882
Fraud	0.72	0.79	0.75	80
accuracy			1.00	56962
macro avg	0.86	0.89	0.88	56962
weighted avg	1.00	1.00	1.00	56962

CHAPTER-9

CONCLUSION & FUTURE ENHANCEMENT

Nowadays, in the global computing environment, online payments are important, because online payments use only the credential information from the credit card to fulfill an application and then deduct money. Due to this reason, it is important to find the best solution to detect the maximum number of frauds in online systems.

Accuracy, Error-rate, Sensitivity and Specificity are used to report the performance of the system to detect the fraud in the credit card. In this paper, three machine learning algorithmsare developed to detect the fraud in credit card system. To evaluate the algorithms, 80% of the dataset is used for training and 20% is used for testing and validation. Accuracy, error rate, sensitivity and specificity are used to evaluate for different variables for three algorithms. The accuracy result is shown for SVM; Decision tree and random forest classifier are 99.94, 99.92, and 99.95 respectively. The comparative results show that the Random Forest performs better than the SVM and decision tree techniques.

Future Enhancement

Detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together asmodules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

BIBLIOGRAPHY

[1] B.Meena, I.S.L.Sarwani, S.V.S.S.Lakshmi," Web Service mining and its techniques in Web Mining" IJAEGT,Volume 2,Issue 1 , Page No.385-389.

[2] F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2019.

[3] G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2019, ISSN ISSN: 2277-1581.

[4] K. Chaudhary, B. Mallick, "Credit Card Fraud: The study of its impact and detection techniques", International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, pp. 31-35, 2019, ISSN ISSN: 2277-5420.

[5] M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. Sid- Ahmed, "Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2017.

[6] R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2018.

[7] S. Patil, H. Somavanshi, J. Gaikwad, A. Deshmane, R. Badgujar, "Credit Card Fraud Detection Using Decision Tree Induction Algorithm", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4, no. 4, pp. 92-95, 2020, ISSN ISSN: 2320-088X.

[8] S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick,"Credit card fraud detection using Bayesian and neural networks", Proceedings of the 1st international naiso congresson neuro fuzzy technologies, pp. 261-270, 2017.

[9] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C.Westland, "Data mining for credit card fraud: A comparative study", Decision Support Systems, vol. 50, no. 3, pp. 602-613, 2019.

[10] Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2018 International Symposium, pp. 315-319, 2018.

APPENDIX

Appendix A: Screen Shots

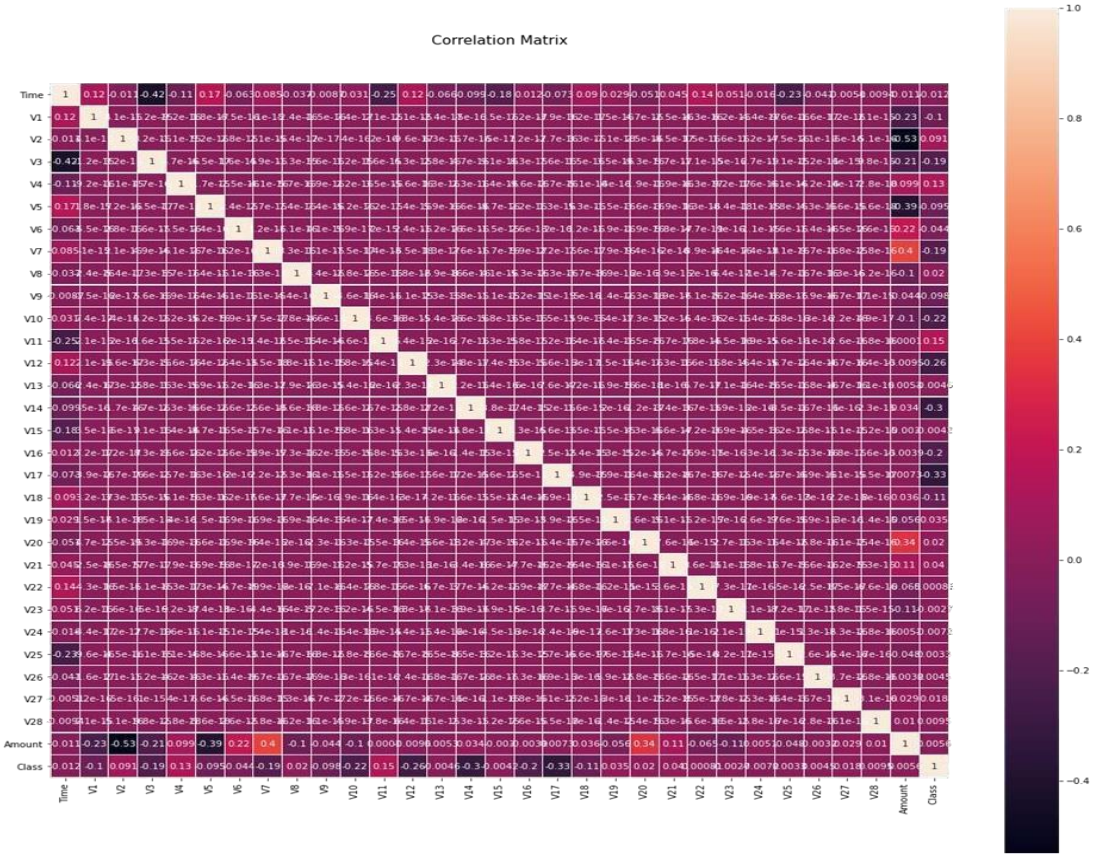
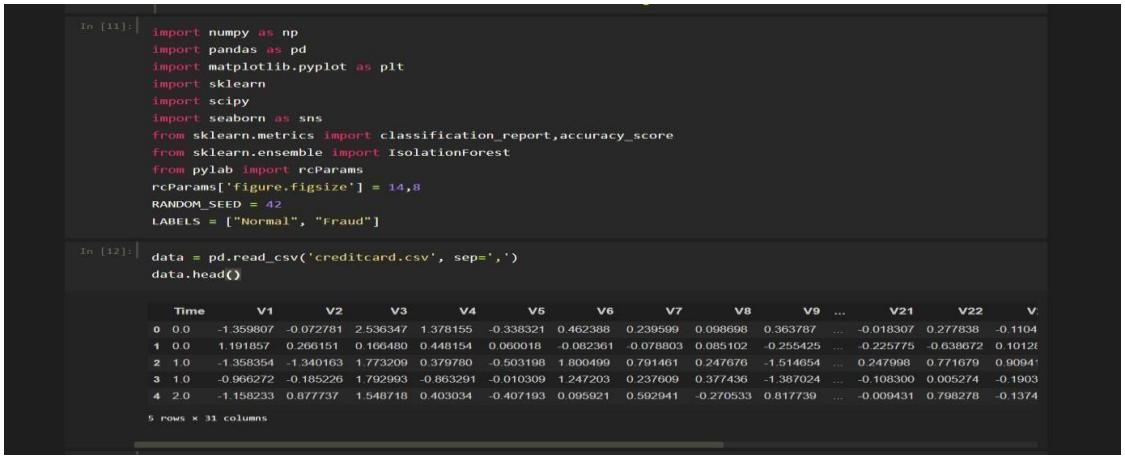


Fig 1 Correlation Matrix



```
In [11]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import sklearn
import scipy
import seaborn as sns
from sklearn.metrics import classification_report,accuracy_score
from sklearn.ensemble import IsolationForest
from pylab import rcParams
rcParams['figure.figsize'] = 14,8
RANDOM_SEED = 42
LABELS = ["Normal", "Fraud"]

In [12]:
data = pd.read_csv('creditcard.csv', sep=',')
data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.1104
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.1012
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.9094
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.1903
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.1374

5 rows × 31 columns

Fig 2 Dataset

```
df = _
data = df.df

fraud = data[data['Class'] == 1]
valid = data[data['Class'] == 0]
outlierFraction = len(fraud)/float(len(valid))
print(outlierFraction)
fraud.Amount.describe()
```

```
Last executed at 2021-01-21 14:36:13 in 166ms
```

```
0.0017304750013189597
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

Fig 3 Data set reading code

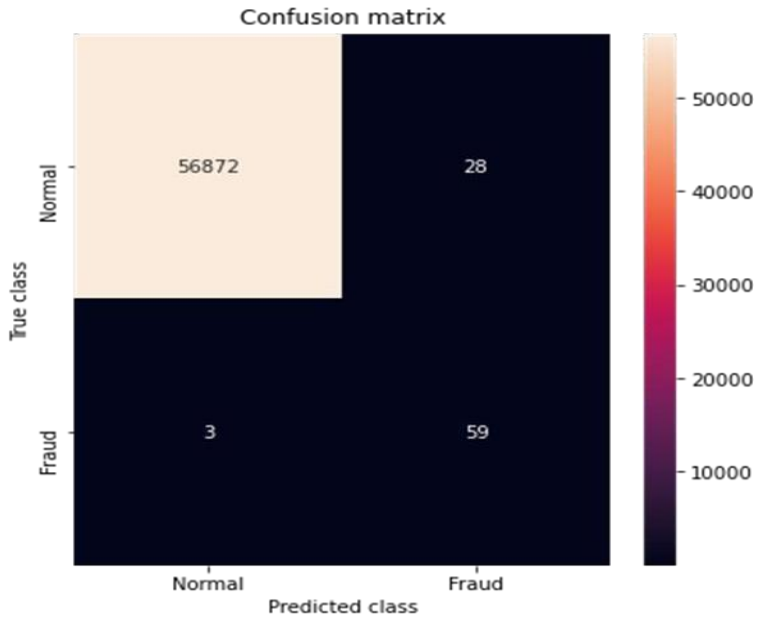


Fig 4 Confusion Matrix

Appendix B: Abbreviations

CCFD – Credit Card Fraud Detection

CSDT – Cost Sensitive Decision Tree

ML – Machine Learning

SVM – Support Vector Machine

URL – Uniform Resource

PAPER PUBLICATION DETAILS

The paper has been published in International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) journal Volume 02, Issue 01 July 2022 and entitled as “Detection of Credit Card Fraud Transaction Using Machine Learning Based Algorithm”. The authors of this paper are Prof Swetha T, Bellam Narendra Nath, Manjunath,Govinda N, H.V Naveen Kumar.

This paper shows the insights about the project which consists of Abstract, Introduction, Literature Review, Methodology, Implementation, Conclusion and so on.


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Bellam Narendra Nath
Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm
IN IJARSCT, VOLUME 2, ISSUE 1, JULY 2022
Certificate No: 072022-A285
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-5742
www.doi.org
www.crossref.org
6.252
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Manjunath
Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm
IN IJARSCT, VOLUME 2, ISSUE 1, JULY 2022
Certificate No: 072022-A286
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-5742
www.doi.org
www.crossref.org
6.252
Scientific Journal Impact Factor
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
Govinda N
Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm
IN IJARSCT, VOLUME 2, ISSUE 1, JULY 2022
Certificate No: 072022-A287
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-5742
www.doi.org
www.crossref.org
6.252
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief


INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
IJARSCT
CERTIFICATE OF PUBLICATION
INTERNATIONAL STANDARD SERIAL NUMBER ISSN NO: 2581-9429
THIS IS TO CERTIFY THAT
H V Naveen Kumar
Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India
HAS PUBLISHED A RESEARCH PAPER ENTITLED
Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm
IN IJARSCT, VOLUME 2, ISSUE 1, JULY 2022
Certificate No: 072022-A288
www.ijarsct.co.in
Crossref
DOI: 10.48175/IJARSCT-5742
www.doi.org
www.crossref.org
6.252
Journal Impact Factor
www.sjifactor.com
Editor-in-Chief

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, July 2022

Impact Factor 6.252

Detection of Credit Card Fraud Transactions using Machine Learning based Algorithm

Swetha T[1], Bellam Narendra Nath[2], Manjunath[3], Govinda N[4], H V Naveen Kumar[5]

Project Guide, Department of Computer Science and Engineering[1] Projecties, Department of Computer Science and Engineering[2,3,4,5] Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology, Chikkaballapura, Karnataka, India

Abstract: The rapid growth in E-Commerce industry has lead to an exponential increase in the use of credit cards for online purchases and consequently they has been surge in the fraud related to it .In recent years, For banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. This paper investigates the performance of SVM, decision tree and random forest for credit card fraud detection. Dataset of credit card transactions is collected from kaggle and it contains a total of 2,84,808 credit card transactions of a European bank data set. It considers fraud transactions as the "positive class" and genuine ones as the "negative class" .The data set is highly imbalanced, it has about 0.172% of fraud transactions and the rest are genuine transactions. The author has been done oversampling to balance the data set, which resulted in 60% of fraud transactions and 40% genuine ones. The three techniques are applied for the dataset and work is implemented in R language. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy and error rate. The result shows of accuracy for SVM, Decision tree and random forest classifier are 90.0, 94.3, 95.5 respectively. The comparative results show that the Random forest performs better than the SVM and decision tree techniques.

Keywords: Fraud detection, Credit card, SVM, Decision tree, Random forest.

I. INTRODUCTION

Credit card fraud is a huge ranging term for theft and fraud committed using or involving at the time of payment by using this card. The purpose may be to purchase goods without paying, or to transfer unauthorized funds from an account. Credit card fraud is also an add on to identity theft. As per the information from the United States Federal Trade Commission, the theft rate of identity had been holding stable during the mid 2000s, but it was increased by 21 percent in 2008. Even though credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints In 2000, out of 13 billion transactions made annually, approximately 10 million or one out of every 1300 transactions turned out to be fraudulent.

Also, 0.05% (5 out of every 10,000) of all monthly active accounts was fraudulent. Today, fraud detection systems are introduced to control one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses. Credit Card Fraud is one of the biggest threats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of ways to commit fraud. In simple terms, Credit Card Fraud is defined as "when an individual uses another individuals' credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used". Card fraud begins either with the theft of the physical card or with the important data associated with the account, including the card account number or other information that necessarily be available to a merchant during a permissible transaction. Card numbers generally the Primary Account Number (PAN) are often reprinted on the card, and a magnetic stripe on the back contains the data in machine-readable format. It contains the following Fields:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

- Name of card holder
- Card number
- Expiration date
- Verification/CVV code
- Type of card

There are more methods to commit credit card fraud. Fraudsters are very talented and fast moving people. In the Traditional approach, to be identified by this paper is Application Fraud, where a person will give the wrong information about himself to get a credit card. There is also the unauthorized use of Lost and Stolen Cards, which makes up a significant area of credit card fraud. There are more enlightened credit card fraudsters, starting with those who produce Fake and Doctored Cards; there are also those who use Skimming to commit fraud. They will get this information held on either the magnetic strip on the back of the credit card, or the data stored on the smart chip is copied from one card to another. Site Cloning and False Merchant Sites on the Internet are getting a popular method of fraud for many criminals with a skilled ability for hacking. Such sites are developed to get people to hand over their credit card details without knowing they have been swindled.

Rest of the paper is described as follows: section 2 describes the related work about the credit card system, section 3 described the proposed system architecture and methodology, section 4 shows the performance analysis and results, section 5 shows the conclusion.

II. RELATED WORK

A.Shen etal (2017) demonstrate the efficiency of classification models to credit card fraud detection problem and the authors proposed the three classification models ie., decision tree, neural network and logistic regression. Among the three models neural network and logistic regression outperforms than the decision tree. M.J.Islam et al (2017) proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbor classifier is implemented and applied to the dataset for credit card system.Y. Sahin and E. Duman(2019) has cited the research for credit card fraud detection and used seven classification methods took a major role .In this work they have included decision trees and SVMs to decrease the risk of the banks. They have suggested Artificial Neural networks and Logistic Regression classification models are more helpful to improve the performance in detecting the frauds. Y. Sahin, E. Duman(2020) has cited the research , used Artificial Neural Network and Logistic Regression Classification and explained ANN classifiers outperform LR classifiers in solving the problem under investigation. Here the training data sets distribution became more biased and the distribution of the training data sets became more biased and the efficiency of all models decreased in catching the fraudulent transactions.

IJARSCT

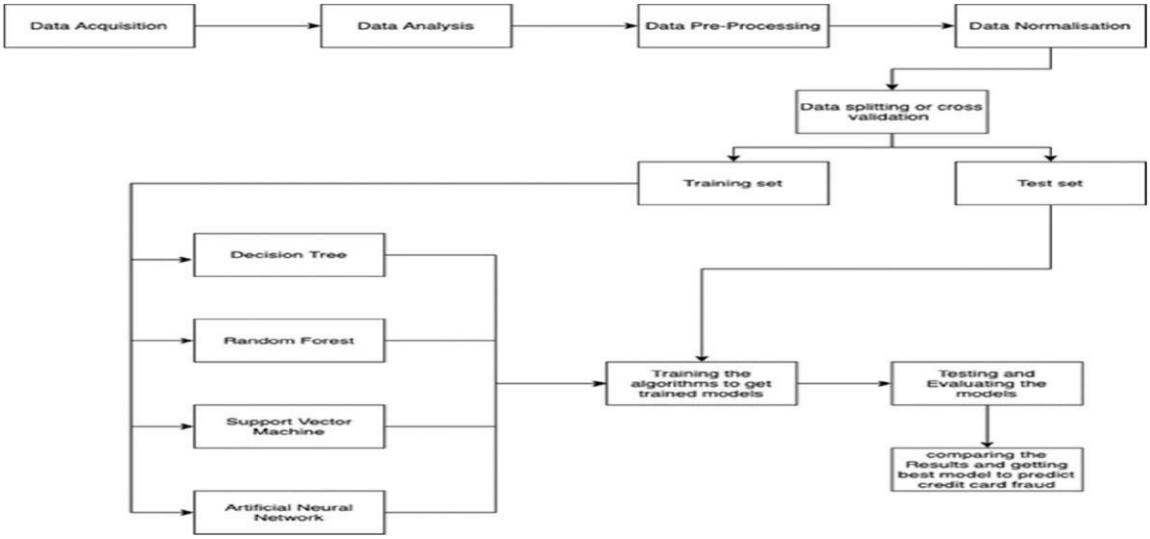
ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, July 2022

Impact Factor 6.252

III. PROPOSED TECHNIQUE



The proposed techniques are used in this paper, for detecting the frauds in credit card system. The comparison are made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions. The Figure1 shows the architectural diagram for representing the overall system framework

IV. DECISION TREE ALGORITHM

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

4.1 Types of Decision Tree

1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree.
2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree

4.2 Terminology of Decision Tree

1. Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. Splitting: It is a process of dividing a node into two or more sub-nodes.
3. Decision Node: When a sub-node splits into further sub- nodes, then it is called decision node.
4. Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.
5. Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree.
7. Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

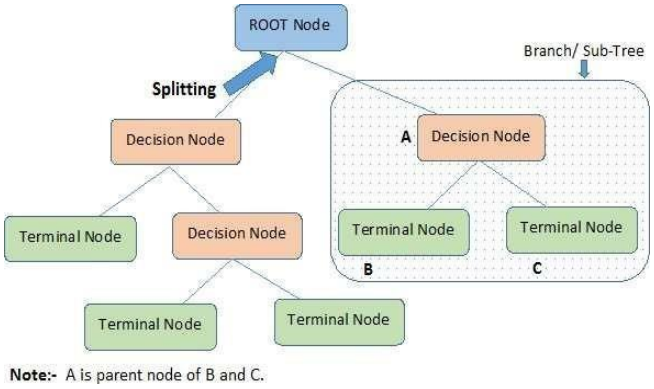
4.3 Working of Decision Tree

Decision trees use multiple algorithms to decide to split a node in two or more sub- nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

1. Gini Index
2. Information Gain
3. Chi Square
4. Reduction of Variance



V. RANDOM FOREST

Random forest is a tree-based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

5.1 Working of Random Forest

Bagging Algorithm is used to create random samples. Data set D1 is given for n rows and m columns and new data set D2 is created for sampling n cases at random with replacement from the original data. From dataset D1,1/3rd of rows are left out and is known as Out of Bag samples. Then, new dataset D2 is trained to this models and Out of Bag samples is used to determine unbiased estimate of the error. Out of m columns, M << m columns are selected at each node in the data set. The M columns are selected at random. Usually, the default choice of M, is m/3 for regression tree and M is sqrt(m) for classification tree. Unlike a tree, no pruning takes place in random forest i.e, each tree is grown fully. In decision trees, pruning is a method to avoid over fitting. Pruning means selecting a sub tree that leads to the lowest test error rate. Cross validation is used to determine the test error rate of a sub tree. Several trees are grown and the final prediction is obtained by averaging or voting.

Step 1: Import the dataset

Step 2: Convert the data into data frames format Step3: Do random oversampling using ROSE package

Step4: Decide the amount of data for training data and testing data

Step5: Give 70% data for training and remaining data for testing.

Step6: Assign train dataset to the models

Step7: Choose the algorithm among 3 different algorithms and create the model

Step8: Make predictions for test dataset for each algorithm Step9: Calculate accuracy for each algorithm

Step10: Apply confusion matrix for each variable

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, July 2022

Impact Factor 6.252

Step11: Compare the algorithms for all the variables and find out the best algorithm.

Table 2: Algorithm steps for finding the best algorithm

VI. SUPPORT VECTOR MACHINES

(SVMs) are a popular machine learning method for classification, regression, & other learning tasks. LIBSVM is a library for Support Vector Machines (SVMs). A typical use of LIBSVM involves two steps: first, training a data set to obtain a model & second, using the model to predict information of a testing data set. For SVC & SVR, LIBSVM can also output probability estimates. Many extensions of LIBSVM are available at libsvm tools. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

1. Set up the training data for model creation
2. Set up SVM's parameters
3. SVM Trainer
4. SVM Predictor

VII. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS

7.1 Performance Metrics

The basic performance measures derived from the confusion matrix. The confusion matrix is a 2 by 2 matrix table contains four outcomes produced by the binary classifier. Various measures such as sensitivity, specificity, accuracy and error rate are derived from the confusion matrix. Accuracy:

Accuracy is calculated as the total number of two correct predictions(A+B) divided by the total number of the dataset(C+D).It is calculated as (1-error rate).

Accuracy=A+B/C+D (4.1)

Whereas, A=True Positive B=True Negative C=Positive D=Negative

Error rate:

Error rate is calculated as the total number of two incorrect predictions(F+E) divided by the total number of the dataset(C+D).

Error rate=F+E/C+D (4.2)

Whereas, E=False Positive F=False Negative, C=Positive, D=Negative

Sensitivity:

Sensitivity is calculated as the number of correct positive predictions(A) divided by the total number of positives(C).

Sensitivity=A/C (4.3)

Specificity:

Specificity is calculated as the number of correct negative predictions(B) divided by the total number of negatives(D).

Specificity=B/D. (4.4)

Accuracy, Error-rate, Sensitivity and Specificity are used to report the performance of the system to detect the fraud in the credit card.

In this paper, three machine learning algorithms are developed to detect the fraud in credit card system. To evaluate the algorithms, 80% of the dataset is used for training and 20% is used for testing and validation. Accuracy, error rate, sensitivity and specificity are used to evaluate for different variables for three algorithms as shown in Table 3. The accuracy result is shown for SVM; Decision tree and random forest classifier are 92.7, 95.8, and 97.6 respectively. The comparative results show that the Random forest performs better than the SVM and decision tree techniques. Table 3: Performance analysis for three different algorithms

Feature Selection	SVM	Decision tree	Random Forest

For 5 variables	87.2	89	90.1
For 10 variables	88.6	92.1	93.6
For all Variables	90.0	94.3	95.5

VIII. CONCLUSION

In this paper, Machine learning technique like SVM, Decision Tree and Random Forest were used to detect the fraud in credit card system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system. The accuracy for SVM, Decision tree and random forest classifier are 90.0, 94.3, and 95.5 respectively. By comparing all the three methods, found that random forest classifier is better than the and decision tree.

REFERENCES

[1]. Andrew. Y. Ng, Michael. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes", Advances in neural information processing systems, vol. 2, pp. 841-848, 2002.
[2]. A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.
[3]. A. C. Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten, "Cost sensitive credit card fraud detection using Bayes minimum risk", Machine Learning and Applications (ICMLA). 2013 12th International Conference, vol. 1, pp. 333-338, 2013.
[4]. B.Meena, I.S.L.Sarwani, S.V.S.S.Lakshmi," Web Service mining and its techniques in Web Mining" IJAEGT,Volume 2,Issue 1 , Page No.385-389.
[5]. F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2011.
[6]. G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194198, 2012, ISSN ISSN: 2277-1581.
[7]. K. Chaudhary, B. Mallick, "Credit Card Fraud: The study of its impact and detection techniques", International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, pp. 31-35, 2012, ISSN ISSN: 2277-5420.
[8]. M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. Sid- Ahmed, "Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2007.
[9]. R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2000.
[10]. S. Patil, H. Somavanshi, J. Gaikwad, A. Deshmane, R. Badgujar, "Credit Card Fraud Detection Using Decision Tree Induction Algorithm", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4, no. 4, pp. 92-95, 2015, ISSN ISSN: 2320-088X.
[11]. S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, "Credit card fraud detection using Bayesian and neural networks", Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261- 270, 2002.
[12]. S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, "Data mining for credit card fraud: A comparative study", Decision Support Systems, vol. 50, no. 3, pp. 602-613, 2011.
[13]. Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.
[14]. Selvani Deepthi Kavila,LAKSHMI S.V.S.S.,RAJESH B " Automated Essay Scoring using Feature Extraction Method " IJCER ,volume 7,issue 4(L), Page No. 12161-12165.
[15]. S.V.S.S.Lakshmi,K.S.Deepthi,Ch.Suresh "Text Summarization basing on Font and Cue-phrase