# PROFESSIONAL TRAINING REPORT

**AT**

**SATHYABAMA INSTITUTE ANDTECHNOLOGY**

**(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

**GRANDHI GOWRI SANKAR (40110404)**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

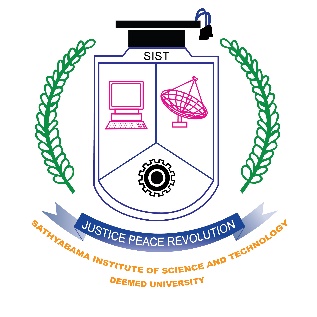
**SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI**

**CHENNAI – 600119, TAMILNADU**

**OCT 2022**

 **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

### (DEEMED TO BE UNIVERSITY)

**Accredited with Grade “A” by NAAC**

**(Established under Section 3 of UGC Act, 1956)**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119**

[**www.sathyabama.ac.in**](http://www.sathyabama.ac.in)

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonfide work of **GRANDHI GOWRI SANKAR** who carried out the project entitled “**CREDIT CARD FRAUD DETECTION USING PYTHON**” under my supervision from June 2022 to November 2022.

## Internal Guide

## DR.M.D. JEMSHIA MIRIAIAM M.E

**Head of the Department**

**Dr. L. Lakshmanan, M.E., Ph.D.**



Submitted for Viva-voce Examination held on

**Internal Examiner External examiner**

**DECLARATION**

I GRANDHI GOWRI SANKARhereby declare that the project report entitled **CREDIT CARD FRAUD DETECTION USING PYTHON** done by me under the guidance of **Dr. JEMSHIA MIRIAM** submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

## DATE:08-11-2022

**PLACE: CHENNAI SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

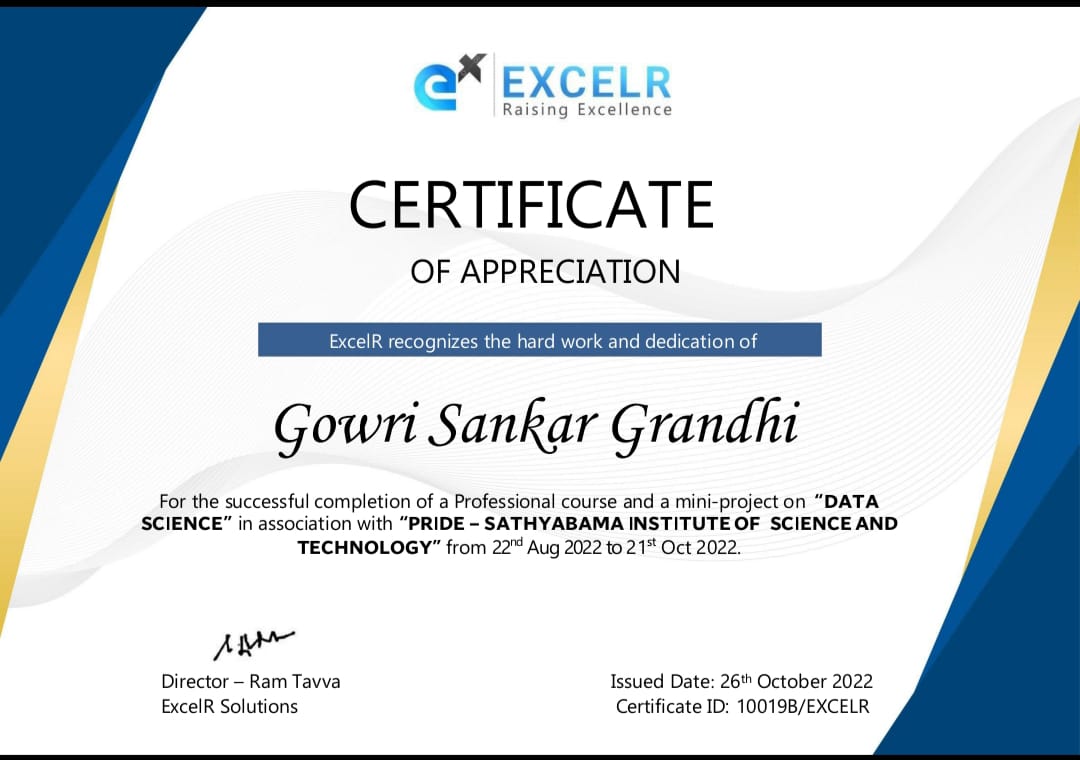
I am pleased to acknowledge my sincere thanks to the **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. Shashikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S. Vignesh Wari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. JEMISHIA MIRIAM** for her valuable guidance, suggestions, and constant encouragement that paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project

**TRAINING CERTIFICATE**



# ABSTRACT

Due to the rapid growth of the E-Commerce industry, the use of credit cards for online purchases has increased dramatically. In recent years, credit card fraud is becoming a major complication for banks as it has become very difficult for detecting fraud in the credit card system. To overcome this hardship Machine learning plays an eminent role in detecting the credit card fraud in the transactions. Modelling prior credit card transactions with data from ones that turned out to be fraudulent is part of the Card Fraud Detection Problem. In Machine learning the machine is trained at first to predict the output so, to predict the various bank transactions various machine learning algorithms are used.  The SMOTE approach was employed to oversample the dataset because it was severely unbalanced. This paper the examines and overview the performance of K-nearest neighbours, Decision Tree, Logistic regression and Random Forest, XG Boost for credit card fraud detection.The assignment is implemented in Python and uses five distinct machine learning classification techniques. The performance of the algorithm is evaluated by accuracy score, confusion matrix, f1-score, precision and recall score and au-roc curve as well.

**Keywords: Fraud Detection, Machine Learning, Logistic regression, KNN, Decision tree, random forest, XG boost .**

. When we make any transaction while purchasing any product online — a good amount of people prefers credit cards.

.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| CHAPTER No | TITLE | PAGE No |
|  | ABSTRACT | i |
|  | LIST OF ABBREVIATIONS  LIST OF DEFINATIONS | iii  iii |
| 1 | **INTRODUCTION** | 1 |
|  |  |  |
| 2 | **AIM AND SCOPE OF THE PRESENT INVESTIGATION** | 2 |
|  | 2.1 Aim | 2 |
|  | 2.2 Scope of the project | 2 |
| 3 | **MATERIALS AND METHODS, ALGORITHMS USED** | 3 |
|  | 3.1 Software Requirements | 17 |
|  | 3.2 Hardware Requirements | 17 |
|  |  |  |
| 4 | **RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS** | 17 |
| 5 | **SUMMARY AND CONCLUSIONS**  **5.1 ADVANTAGES OF CREDIT CARD FRAUD DETECTION** | 22  23 |
|  | **REFERENCES** | 25 |
|  |  |  |

# 

**CHAPTER 1**

**INTRODUCTION**

A credit card is a thin handy plastic card that contains identification information such as a signature or picture, and authorizes the person named on it to charge purchases or services to his account – charges for which he will be billed periodically. Today, the information on the card is read by automated teller machines (ATMs), store readers, bank and is also used in online internet banking system.

They have a unique card number which is of utmost importance. Its security relies on the physical security of the plastic card as well as the privacy of the credit card number.

There is a rapid growth in the number of credit card transactions which has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging

term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction.

Generally, the statistical methods and many data mining algorithms are used to solve this fraud detection problem. Most of the credit card fraud detection systems are based on artificial intelligence, Meta learning and pattern matching. The Genetic algorithms are evolutionary algorithms which aim to obtain the better solutions in eliminating the fraud.

A high importance is given to develop efficient and secure electronic payment system to detect whether a transaction is fraudulent or not. In this paper, we will focus on credit card fraud and its detection measures. A credit card fraud occurs when one individual uses other individuals’ card for their personal use without the knowledge of its owner.

When such kind of cases takes place by fraudsters, it is used until its entire available limit is depleted. Thus, we need a solution which minimizes the total available limit on the credit card which is more prominent to frauds. And, a Genetic algorithm generate better solutions as time progresses. The complete emphasis is given on developing efficient and secure electronic payment system for detecting the fraudulent.

**CHAPTER 2**

**AIM AND SCOPE OF THE PRESENT INVESTIGATION**

**2.1 AIM OF THE PROJECT**

Credit card fraud detection using machine learning and python

.

**2.2 SCOPE OF THE PROJECT**

Fraud is defined as an illegal or criminal deception intended to gain financial or personal gain. It is a purposeful act committed in violation of a law, rule, or policy with the intent of obtaining unlawful financial gain.

Several literatures on anomaly or fraud detection in this domain have previously been

published and are open to the public. Data mining applications, automated fraud detection,

and adversarial detection are among the strategies used in this domain, according to a comprehensive survey undertaken by Clifton Puha and his colleagues. Unconventional techniques, such as hybrid data mining/complex network classification algorithm, have

shown effective on medium-sized online transactions, based on network reconstruction algorithm that allows building representations of the divergence of one instance from a

reference group. Fraud detection is a difficult undertaking, and no algorithm can accurately anticipate if a transaction is fraudulent.

The following are characteristics of a good fraud detection

system:

• The frauds should be correctly identified.

• Frauds should be detected quickly.

• A genuine transaction should not be classified as

fraud.

**CHAPTER 3**

**MATERIALS AND METHODS USED**

Card transactions are always un familiar when compared to previous transaction made the customer id

unfamiliarity is a very difficult problem in real-world when are called concept drift problems [1]. Concept drift can be said as a variable which changes over time and in unforeseen ways. These variables cause a high imbalance in data. The main aim of our research is to overcome the problem of Concept drift to implement on real-world scenario.

**Materials / Requirements**:

**Windows XP, Windows 7(ultimate, enterprise)**

**SQL2008**

**Visual studio 2010**

**HARDWARE COMPONENTS:**

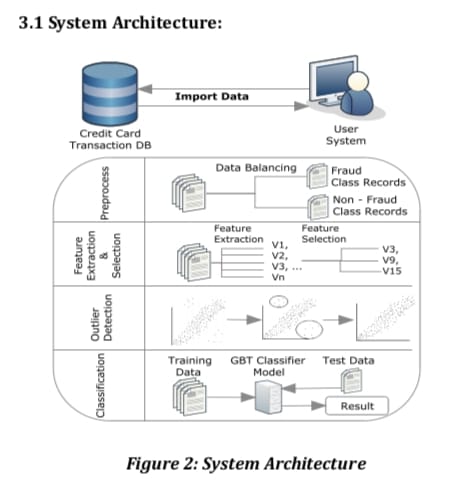
**Processor – i3**

**Min Hard Disk – 4 GB**

**Min Memory – 1GB RAM**

**ARCHITECTURE:**

1 System Architecture:



1 System Architecture: Figure 2: System Architecture Above fig shows the process of CCFDS. This system model accepts real time customer credit card transaction database.it is more important to find fraud rate of credit card.

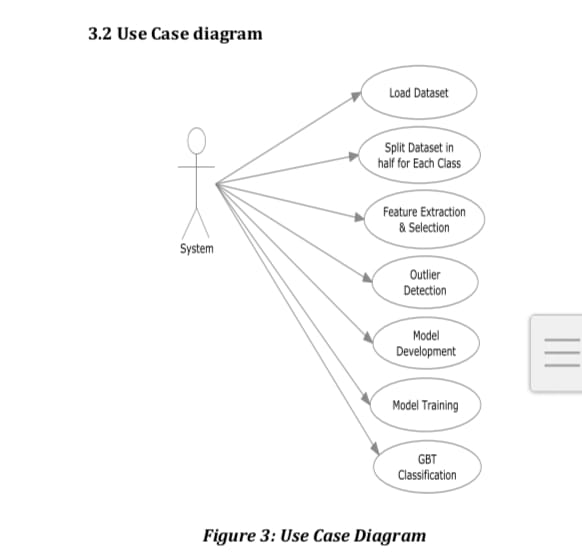
Data collection: collect input dataset based on transaction details,

Data balancing: after collecting large set of database it is necessary to understand and separate the balanced data and unbalanced data in two types of class .clas-0 indicates non-fraud and class-1 indicates fraud.

Feature extraction and selection: class-1 indicates total fraud transactions are 492 samples. In this project v1, v2 …v28 features.

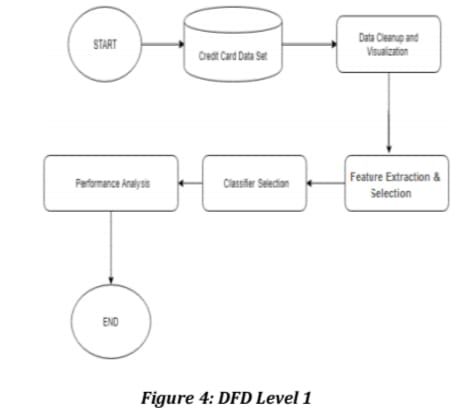
Outlier detection: It measures the distance between each similar data to the clustering technique. The values which are not follows the trained data consider as outlier.

Classification: As the dataset is imbalanced, many classifiers show bias for majority classes. Spark library is applied as a SQL-like analysis to a large amount of structured or semi-structured data. GBT Classifier does the classification of data coming through the stream. 3.2



3: **Use Case Diagram**

3.3 DFD (Data Flow Diagram) The DFD used as communication tool between system and user.it is a simple representation of the complete project process. Transaction detection activity follows three phases.1.Data exploration 2.Data prepressing 3.data classifications.



‘Figure 4: DFD Level 1 4. SYSTEM IMPLEMENTATION The proposed system is divided into several smaller units, known as module.

Figure 2: System Architecture Above fig shows the process of CCFDS. This system model accepts real time customer credit card transaction database.it is more important to find fraud rate of credit card.

Data collection: collect input dataset based on transaction details, Data balancing: after collecting large set of database it is necessary to understand and separate the balanced data and unbalanced data in two types of class. clas-0 indicates non-fraud and class-1 indicates fraud. Feature extraction and selection: class-1 indicates total fraud transactions are 492 samples. In this project v1, v2 …v28 features. Outlier detection: It measures the distance between each similar data to the clustering technique.

The values which are not follows the trained data consider as outlier. Classification: As the dataset is imbalanced, many classifiers show bias for majority classes. PY Spark library is applied as a SQL-like analysis to a large amount of structured or semi-structured data. GBT Classifier does the classification of data coming through the stream. 3.2 Use Case diagram Figure 3: Use Case Diagram 3.3 DFD (Data Flow Diagram) The DFD used as communication tool between system and user.it is a simple representation of the complete project process. Transaction detection activity follows three phases

Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

**4.1 Module Names**

**1. Data Collection**

**2. Data Balancing**

**3. Feature Extraction**

**4. Outlier Detection**

**5. Classification**

4.2 Module Description

1. **Data Collection**:

It contains 2,84,807 records of credit card transactions that happened in duration of just 2days. This dataset is very much unbalanced as it has a total of 492 fraud entries 2,84,515 genuine entries i.e.is just 0.17% of total records. The original features are masked with V1, V2, V3, ...V28. The last column here represents fraud or non-fraud class i.e. represented by0 and 1 respectively.

2. **Data Balancing:**

Imbalanced classes are a general issue in ML based classification where there is an abnormal count of each class. It occurs due to the fact that ML Algorithms are typically intended to improve precision by diminishing the errors. In this manner ,they don't consider the class or balancing the ratio of classes. As out of 2.84807 transactions just 492 fraud transactions exist, which makes it quite difficult to build a standard model with this much less number of fraud transactions. Thus, we use pandas in python to make it 50-50 i.e. we decrease the no. of legitimate transactions to balance it with the number of fraud transactions in equal proportion.

3. **Feature Extraction**:

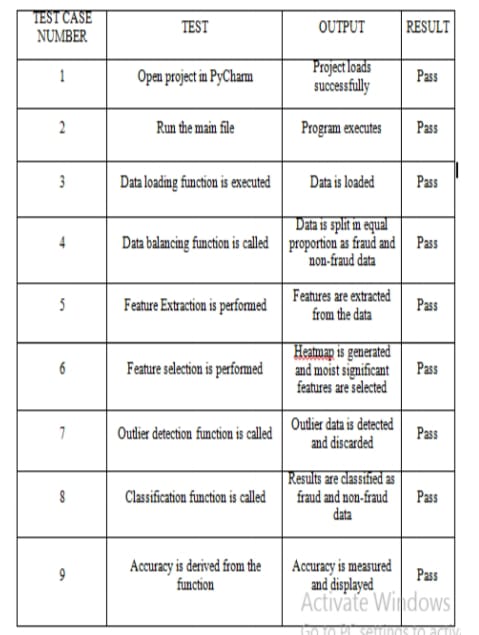
We use heatmap technique to find the significant feature that can distinguish the classes properly and ultimately that affects the accuracy of detection algorithm. Heatmap provides a good visualization of the major and minor values in the matrix as different Coloured cells that define the values. Here, rows/columns of the matrix are clustered in sets. Thus, the features which look most significant are recognized and used further for model training.

4. **Outlier Detection:**

The outlier detection technique measures the distance of each data similar to the clustering technique, but is used to find specific data and rules that are separated from the total data. The values which are not in flow of the linear graph are considered as outliers. Here our aim is to reduce the outliers to have a better trained model. We use library in python for this.

5. **Classification:**

The task of classification occurs in a wide range of applications. In a broad sense, the term could relate to any context in which some decision or Forecast is made on the basis of currently available information. It works on a set of pre-defined class session the basis of observed attributes or features. Here the aim is to establish a rule whereby one can classify a new observation into one of the existing classes. The construction of a classification procedure from a set of data for which the true classes are known has also been variously termed as pattern recognition, discrimination, or supervised learning. We use PY Spark and GBT Classifier for data streaming and classification purpose. PY Spark library is applied as a SQL-like analysis to a large amount of structured or semi-structured data. GBT Classifier does the classification of data coming throug.1. Data exploration 2. Data prepressing 3.Data classifications. Figure 4: DFD Level 14. SYSTEM IMPLEMENTATION The proposed system is divided into several smaller units, known as module.



***Project pipeline***

The project pipeline can be briefly Summary in the following four steps:

* **Data Understanding:**

Here, you need to load the data and understand the features present in it. This would help you choose the features that you will need for your final model.

* **Exploratory data analytics (EDA):**
* Normally, in this step, you need to perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary. For the current data set, because Gaussian variables are used, you do not need to perform Z-scaling. However, you can check whether there is any skewness in the data and try to mitigate it, as it might cause problems during the model building phase.
* **Train/Test split:** Now, you are familiar with the train/test split that you can perform to check the performance of your models with unseen data. Here, for validation, you can use the k-fold cross-validation method. You need to choose an appropriate k value so that the minority class is correctly represented in the test folds.
* **Model building / hyperparameter tuning:** This is the final step at which you can try different models and fine-tune their hyperparameters until you get the desired level of performance on the given data set. You should try and check if you get a better model by various sampling techniques.
* **Model evaluation:** Evaluate the models using appropriate evaluation metrics. Note that since the data is imbalanced, it is is more important to identify the fraudulent transactions accurately than the non-fraudulent ones. Choose an appropriate evaluation metric that reflects this business goal.

**Stepwise Approach:**

***Data Preparation:***

1. EDA: Data understanding, cleaning, univariate and bivariate analysis.
2. Unnecessary columns are removed.
3. Stratified train test split is performed to create train and test datasets.
4. Skewness mitigation in data is done using Power Transformer.
5. Defined functions wherever necessary to maintain conciseness in coding.

***Model Building and Evaluation Approach:***

* For this large dataset, Randomized Search CV with Stratified K Fold cross-validator is used to find the optimal hyperparameters along with the best estimator.
* learn pipelines are used:
  + Pipelines assemble several steps that can be cross-validated together while setting different parameters.
  + Advantage: Pipelines make sure that VALIDATION sets are NOT OVERSAMPLED within folds while up sampling as a part of cross-validation.
* Class imbalances are dealt (internally) within pipelines.
* Different models are built on raw imbalanced dataset as well as on balanced datasets. All models are built for all the 4 stages:
  + (1) Imbalanced Dataset (oversampling: None)
  + Balanced Datasets (2) Random Over sampler (3) SMOTE (4) ADASYN
* Logistic Regression, Decision Tree, Random Forest and XG Boost models are built for all 4 stages.
* SVM: It is Compute intensive and impractical on oversampled datasets. Hence, we train SVM on imbalanced dataset using complete train data, and for balanced datasets on stratified-subset of train data. Models are further evaluated on actual test data.
* KNN: KNN is a lazy learner, works well on small datasets, and is impractical on large datasets.

*last section concludes the paper.*

Terms

last section concludes the paper.

1. Terms

Credit is a method of selling goods or services

without the buyer having cash in hand. A credit card is only an automatic way of offering credit to a consumer. Today, every credit card carries an identifying number that speeds shopping transactions. According to Encyclopedia Britannica (no date), "the use of credit cards originated in the United States during the 1920s, when individual firms, such as oil companies and hotel chains, began issuing them to customers." However, references to credit cards have been made as far back as 1890 in Europe. Early credit cards involved sales directly between the merchant offering the credit and credit card, and that merchant's customer. Around 1938, companies started to accept each other's cards. Nowadays, credit cards allow you to make purchases with

countless third parties (Bellis, no date).In Europe, the most well-known credit card companies are arguably Barclaycard, Citibank, and American Express, offering different types of products depending on their portfolio. Depending on the product offered, the services associated with the card may be different. Interest rate, card fees, exchange rate fee, late payment fee, credit

limit, terms and conditions, are elements that can vary from one bank to another and from one product to another. In the credit card business, fraud occurs when a lender is fooled by a borrower offering him/her purchases, believing that the borrower credit card account will provide payment for this purchase. Ideally, no payment will be made. If the payment is made, the credit card issuer will reclaim the amount paid.

Today, with the expansion of e-commerce, it is on the internet that half of all credit card fraud is conducted. Fraudsters have usually connections with the affected business. In the credit card business, it can be an internal party but most likely an external party. As an external party, fraud is committed being a prospective/existing customer or a prospective/existing supplier. Three different

profiles can be identified for external fraudsters: the average offender, criminal offender, and organized crime offender (Phua et al., 2005). Average offenders display random and/or occasional dishonest behavior when there is opportunity, sudden temptation, or when suffering from financial hardship.

In contrast, the more risky external fraudsters are individual criminal offenders andorganized/group crime offenders (professional/career fraudsters) because they repeatedly disguise their true identities and/or evolve their modus operandi over time to approximate legal forms and to counter detection systems (Phua et al., 2006; Phua et al., 2004). For many companies sometimes dealing with millions of external parties, it is cost-prohibitive to manually check the majority of the external parties’ identity and activities. Indeed, to investigate each

suspicious transaction, they incur a direct overhead

cost for each of them. If the amount of a transaction is smaller than the cost of the overhead, investigating is not worthwhile even if it seems suspicious (Chan et al., 1999; Oscherwitz, 2005). In order to avoid these overheads and depending on the type of fraud committed, diverse solutions can be implemented.

2. **Types of fraud :**

2.1. Bankruptcy fraud. This section focuses on

Bankruptcy fraud and advises the use of credit report from credit bureaux as a source of information regarding the applicants’ public records as well as a possible implementation of a bankruptcy model.

Bankruptcy fraud is one of the most difficult types of fraud to predict. However, some methods or techniques may help in its prevention. Bankruptcy fraud means using a credit card while being insolvent. In other words, purchasers use credit cards knowing that they are not able to pay for their purchases.

The bank will send them an order to pay. However, the customers will be recognized as being in a state of personal bankruptcy and not able to recover their debts. The bank will have to cover the losses itself. Usually, this type of fraud loss is not included in the calculation of the fraud loss provision as it is considered a charge-off loss. The only way to prevent this bankruptcy fraud is by doing a pre-check with credit bureaux in order to be informed about the banking history of the customers. In Germany, for example, some of the most used credit bureaux are SCHUFA and CEG. SCHUFA,

**SOURCE CODE:**

|  |
| --- |
| # import the necessary packages  **import** NumPy as np  **import** pandas as pd  **import** matplotlib.pyplot as plot  **import** seaborn as suns  **from** matplotlib **import** grid spec |

**Code : Loading the Data**

|  |
| --- |
| # Load the dataset from the csv file using pandas  # best way is to mount the drive on Collab and  # copy the path for the csv file  data **=** predocs("credit.csv") |

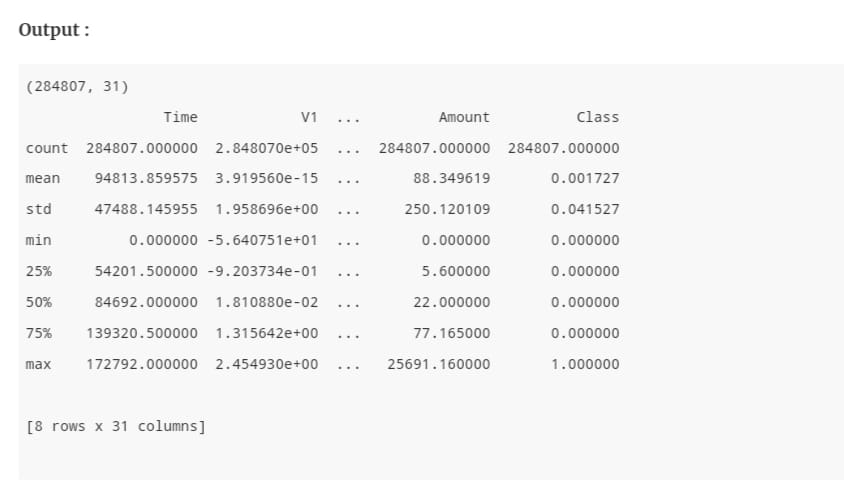
**Code : Understanding the Data**

|  |
| --- |
| # Grab a peek at the data  data. Head() |



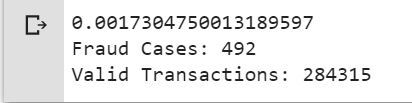
**Code : Describing the Data**

|  |
| --- |
| # Print the shape of the data  # data = data. Sample(frac = 0.1, random state = 48)  print(data. Shape)  print(data. Describe()) |



**Code : Imbalance in thedata**  
Time to explain the data we are dealing with.

|  |
| --- |
| # Determine number of fraud cases in dataset  fraud **=** data[data['Class'] **==** 1]  valid **=** data[data['Class'] **==** 0]  outlier Fraction **=** len(fraud)**/**float(len(valid))  **print**(outlier Fraction)  print('Fraud Cases: {}'.format(len(data[data['Class'] **==** 1])))  print('Valid Transactions: {}'.format(len(data[data['Class'] **==** 0]))) |



**Code: Print the amount details for Fraudulent Transaction**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | Print(“Amount details of the fraudulent transaction”)  fraud . Amount .describe() |   **Output:**  Amount details of the fraudulent transaction  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  **Code: Print the amount details for Normal Transaction**   |  | | --- | | print(“details of valid transaction”)  valid. Amount. describe() |   **Output:**  Amount details of valid transaction  count 284315.000000  mean 88.291022  std 250.105092  min 0.000000  25% 5.650000  50% 22.000000  75% 77.050000  max 25691.160000  Name: Amount, type: float64  As we can clearly notice from this, the average Money transaction for the fraudulent ones is more. This makes this problem crucial to deal with.  **Code:correlationmatrix** The correlation matrix graphically gives us an idea of how features correlate with each other and can help us predict what are the features that are most relevant for the prediction.   |  | | --- | | # Correlation matrix  format = data. Orr()  fig = Pl. Figure(fig size = (12, 9))  Sns . heatmap(format, Vmax = .8, square = True)  Pl. Show() |     In the Heat Map we can clearly see that most of the features do not correlate to other features but there are some features that either has a positive or a negative correlation with each other. For example, V2 and V5 are highly negatively correlated with the feature called Amount. We also see some correlation with V20 and Amount. This gives us a deeper understanding of the Data available to us.  **Code: s** Dividing the data into inputs parameters and outputs value format   |  | | --- | | # dividing the X and the Y from the dataset  X = data.drop(['Class'], axis = 1)  Y = data["Class"]  print(X.shape)  print(Y.shape)  # getting just the values for the sake of processing  # (its a numpy array with no columns)  xData = X.values  yData = Y.values |   **Output :**    (284807, 30)  (284807, )  We will be dividing the dataset into two main groups. One for training the model and the other for Testing our trained model’s performance.   |  | | --- | | # Using Scikit-learn to split data into training and testing sets  from Sklearn .model\_ selection import train test \_split  # Split the data into training and testing sets |   **Code : Building a Random Forest Model using scikit learn**   |  | | --- | | # Building the Random Forest Classifier (RANDOM FOREST)  from sklearn.ensemble import RandomForestClassifier  # random forest model creation  rfc = RandomForestClassifier()  rfc.fit(xTrain, yTrain)  # predictions  yPred = rfc.predict(xTest) |   **Code : Building all kinds of evaluating parameters**   |  | | --- | | # Evaluating the classifier  # printing every score of the classifier  # scoring in anything  from sklearn.metrics import classification\_report, accuracy\_score  from sklearn.metrics import precision\_score, recall\_score  from sklearn.metrics import f1\_score, matthews\_corrcoef  from sklearn.metrics import confusion\_matrix    n\_outliers = len(fraud)  n\_errors = (yPred != yTest).sum()  print("The model used is Random Forest classifier")    acc = accuracy\_score(yTest, yPred)  print("The accuracy is {}".format(acc))    prec = precision\_score(yTest, yPred)  print("The precision is {}".format(prec))    rec = recall\_score(yTest, yPred)  print("The recall is {}".format(rec))    f1 = f1\_score(yTest, yPred)  print("The F1-Score is {}".format(f1))    MCC = matthews\_corrcoef(yTest, yPred)  print("The Matthews correlation coefficient is{}".format(MCC)) |   **Output :**  The model used is Random Forest classifier  The accuracy is 0.9995611109160493  The precision is 0.9866666666666667  The recall is 0.7551020408163265  The F1-Score is 0.8554913294797689  The Matthews correlation coefficient is0.8629589216367891  **Code : Visualizing the Confusion Matrix**   |  | | --- | |  |   **Output :**  **data import**    **CODE:**  #Packages related to general operating system & warnings  import os  import warnings  warnings.filterwarnings('ignore')  #Packages related to data importing, manipulation, exploratory data #analysis, data understanding  import numpy as np  import pandas as pd  from pandas import Series, DataFrame  from termcolor import colored as cl # text customization  #Packages related to data visualizaiton  import seaborn as sns  import matplotlib.pyplot as plt  %matplotlib inline  #Setting plot sizes and type of plot  plt.rc("font", size=14)  plt.rcParams['axes.grid'] = True  plt.figure(figsize=(6,3))  plt.gray()  from matplotlib.backends.backend\_pdf import PdfPages  from sklearn.model\_selection import train\_test\_split, GridSearchCV  from sklearn import metrics  from sklearn.impute import MissingIndicator, SimpleImputer  from sklearn.preprocessing import PolynomialFeatures, KBinsDiscretizer, FunctionTransformer  from sklearn.preprocessing import StandardScaler, MinMaxScaler, MaxAbsScaler  from sklearn.preprocessing import LabelEncoder, OneHotEncoder, LabelBinarizer, OrdinalEncoder  import statsmodels.formula.api as smf  import statsmodels.tsa as tsa  from sklearn.linear\_model import LogisticRegression, LinearRegression, ElasticNet, Lasso, Ridge  from sklearn.neighbors import KNeighborsClassifier, KNeighborsRegressor  from sklearn.tree import DecisionTreeClassifier, DecisionTreeRegressor, export\_graphviz, export  from sklearn.ensemble import BaggingClassifier, BaggingRegressor,RandomForestClassifier,RandomForestRegressor  from sklearn.ensemble import GradientBoostingClassifier,GradientBoostingRegressor, AdaBoostClassifier, AdaBoostRegressor  from sklearn.svm import LinearSVC, LinearSVR, SVC, SVR  from xgboost import XGBClassifier  from sklearn.metrics import f1\_score  from sklearn.metrics import accuracy\_score  from sklearn.metrics import confusion\_matrix  **AD**  **AD**  **AD** |
|  |

**3.1 SOFTWARE REQUIREMENTS:**

* Operating system : Windows 8/10.
* IDE Tool : PyCharm
* Coding Language : Python 3.6
* APIs : Numpy, Pandas,PySpark, Matplotlib
  1. **HARDWARE REQUIREMENTS:**
* Processor : Pentium i3 or higher.
* RAM : 4 GB or higher.
* Hard Disk Drive : 20 GB (free).
* Peripheral Devices : Monitor, Mouse and Keyboard

**CHAPTER 4**

**RESULTS AND DISCUSSION**

1. **RESULT AND ANALYSIS**

**Step 1**: Install the required software and files like PTHON

3.6 and pycham IDE

**Step 2**: Create a folder to load credit card database

**Step 3**: Open pycharm ide

**Step 4**: Create a new project

**Step 5**: Import the credit card database to new project

**Step 6**: Write executable credit card fraud detection

program code

**Step** 7: Verify and debug the code to get better and good

result

**Step** 8: Run the project code

Step 9: Check he final result

------------------\*\*\*\*\*\*\*-----------------------------

**Total records : 284807**

**Fraud record : 492**

**Non-fraud record: 284315**

**True positive : 79**

**True negative : 77**

**False positive : 4**

**False negative : 6**

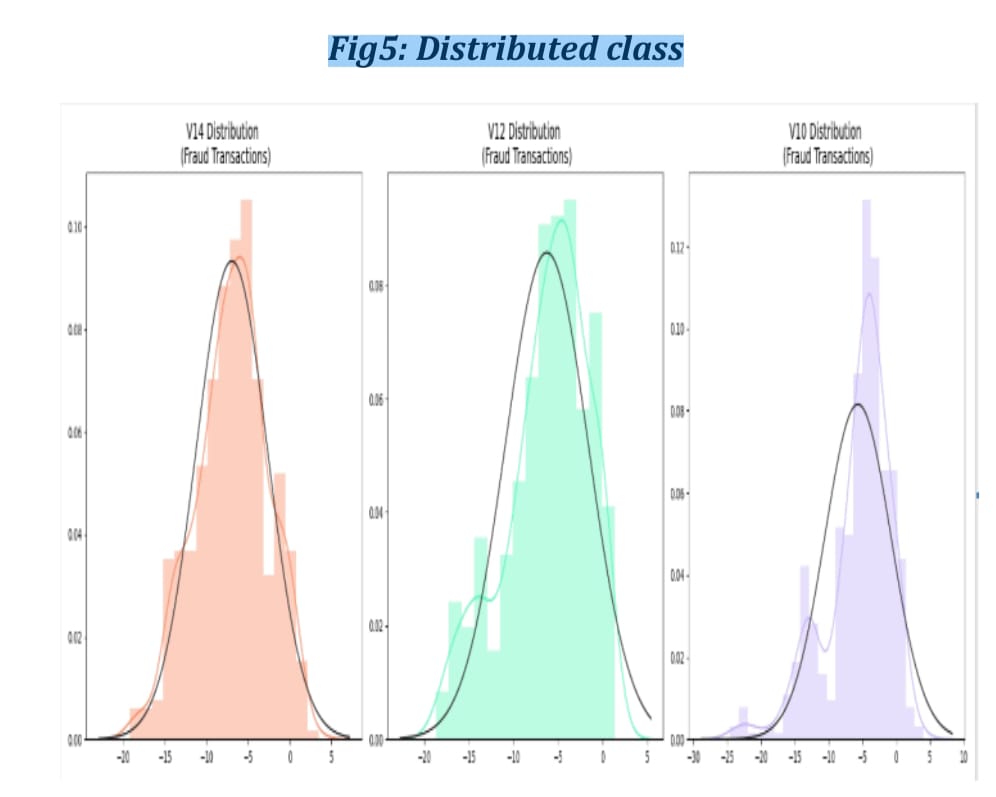
**Recall : 0.9294117647058824**

**Precision** : 0.**9518072289158826**

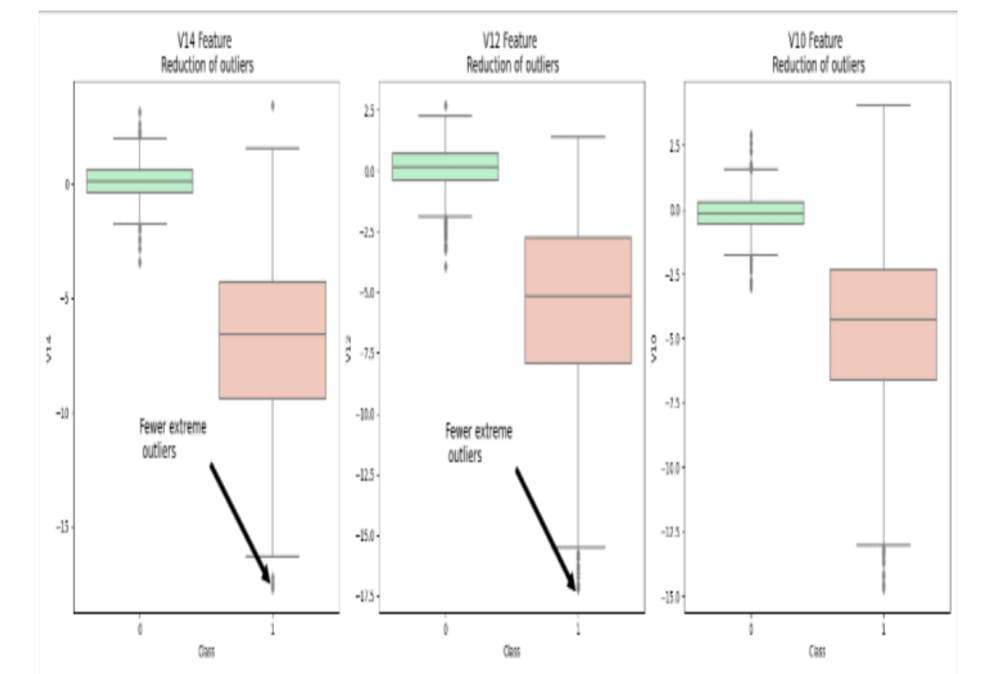
**Accuracy** : 92.9411764



**Fig5: Distributed class**



**Fig6: Fraud transaction detection of features**



**fig 7:Feature reduction of outliers**

**CHAPTER 5**

**SUMMARY AND CONCLUSION**

With the development of electronic financial transaction technology and the emergence of simple payment, the risk of fraudulent payment and fraudulent payment increases as the authentication process is simplified.

The types of fraudulent use of credit cards include theft and loss ,identity theft, new card not received, card forgery, and card information theft. In particular, as phishing ,pharming as well as card information leakage due to card information leakage, card information theft accidents a reoccurring. In response, the government tried to deal with electronic financial fraud by implementing the 'e-financial fraud prevention service'. It is difficult to cope with financial fraud by simply setting the existing keyboard security, public certificate, and additional password.

**REFERENCES:**

[1] Donald V. MacDougall, Richard G. Mosley, Garioch J. l.

Saunders; Credit card crime in Canada: Investigation -

Prosecution; The Canadian Association of Crown Counsel;

page 1-56; January 1985.

[2] Isabelle Sender; Detecting and combating fraud; Chain

Store Age; New York; Vol. 74; Issue 7; Page 162; July 1998.

3] Elkford Dean, Raj Thomas, Lorry; Visa security center;

Personal meetings; January 7 and February 11,1999.

[4] Gyatso Kim and Selig Lee, “2014 Payment Research”,

Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[5] EWT Ngai, Yong Hu, HY Wong, Yujun Chen, Xin Sun,

“The Application of Data Mining Techniques in Financial