Securing Digital Transactions: Machine Learning-Based Credit Card Fraud Detection System

Author Info
Niharikareddy Meenigea
Data Analyst
Virginia International University
USA

Abstract

The rapid growth of the e-commerce industry has resulted in an exponential increase in the use of credit cards for online purchases, resulting in an increase in associated fraud. In recent years, it has become very difficult for banks to detect fraud in the credit card system. Machine learning plays a key role in detecting credit card fraud in transactions. To predict these transactions, banks are using various machine learning techniques, collecting historical data and using new capabilities to improve their predictive power. The performance of fraud detection in credit card transactions is highly influenced by the data set sampling approach, the choice of variables, and the detection techniques used. This white paper examines the performance of logistic regression, decision trees, and random forests for credit card fraud detection. The credit card transactions data set was collected by Kaggle and contains a total of 2,84,808 credit card transactions from the European banking data set. We consider fraudulent transactions as a "positive class" and genuine transactions as a "negative class". The dataset is highly imbalanced, containing approximately 0.172% fraudulent transactions and the rest genuine transactions. The authors oversampled to balance the dataset, resulting in 60% fraudulent and 40% genuine transactions. Three techniques are applied to the dataset and the work is implemented in the R language. The technique's performance is evaluated on various variables based on sensitivity, specificity, precision, and error rate. Results show accuracies for logistic regression, decision tree, and random forest classifiers: 90.0, 94

Keywords: Fraud detection, Credit card, Logistic regression, Decision tree, Random Forest

1. INTRODUCTION

Credit card fraud is a broad term that refers to theft or fraud involved in using or making payments with the card. The intent could be to purchase goods without payment or to transfer fraudulent funds from your account. Credit card fraud is also an adjunct to identity theft. According to the U.S. Federal Trade Commission, identity theft rates remained stable in the mid-2000s, but in 2008 he increased by 21%. Despite credit card fraud, the crime most people associate with identity theft, the rate of all identity theft complaints declined in 2000, accounting for about 10 million out of 13 billion transactions annually, or 1 out of 1,300 transactions turned out to be fraudulent.

Also, 0.05% (5 out of 10,000) of all monthly active accounts were fraudulent. Today, fraud detection systems are deployed to control 1/12th of 1% of all transactions processed, yet they still cost billions of dollars. Credit card fraud is one of the biggest threats facing businesses today. However, to effectively combat fraud, it is important to first understand how fraud occurs. Credit card fraudsters carry out their fraudulent activities in a variety of ways. Simply put, credit card fraud is defined as "the use of someone else's credit card for personal reasons without the cardholder and card issuer knowing the card was used." Card fraud begins with physical card theft or key account information such as the card's account number or other information that the merchant needs to obtain during an authorized transaction. The card number, usually the Primary Account Number (PAN), is often printed on the card, and the magnetic stripe on the back contains the data in machine-readable form. Contains the following fields:

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

There are other ways to commit credit card fraud. Scammers are very talented and fast-moving people. In a conventional approach, this document identifies application fraud where an individual provides false information about themselves to obtain credit for her card. There is also lost and stolen card fraud, an important area of credit card fraud. There are smarter credit card scammers, starting with those who create counterfeit or fraudulent cards. Some people also use skimming to commit fraud. This information is stored on the magnetic stripe on the back of your credit card, or data stored on a smart chip is copied from one card to another. Website cloning and fake seller websites on the Internet are becoming popular fraud techniques for many criminals with advanced hacking skills. Such websites are designed to trick people into providing their credit card details without knowing they have been scammed.

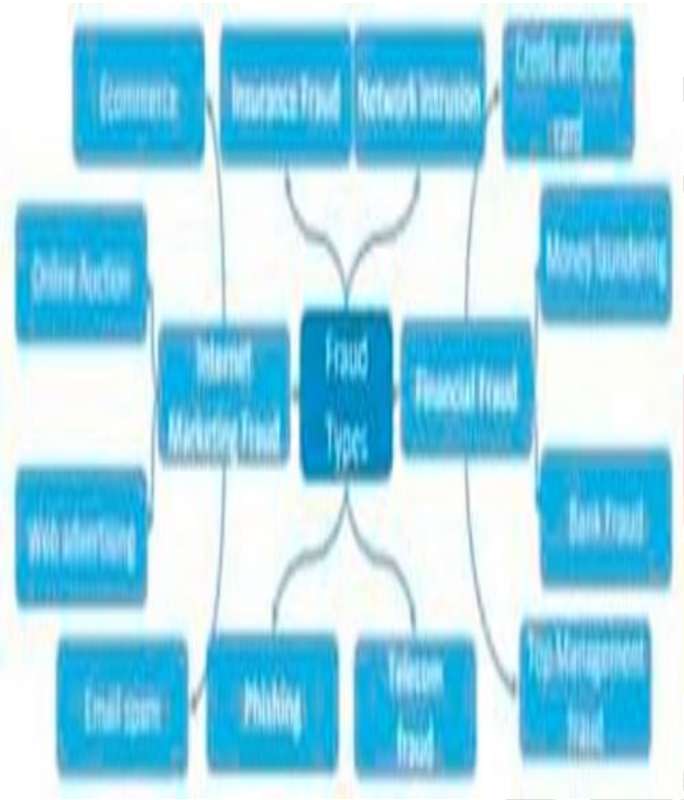
The rest of the book is described as follows. Section 2 describes relevant work on the credit card system, Section 3 describes the proposed system architecture and methodology, Section 4 presents performance analysis and results, and Section 5 presents conclusions.

A credit card, commonly referred to as a card assigned to a customer (cardholder), typically allows the customer to purchase goods and services within their credit limit or withdraw cash in advance. Credit cards give the cardholder a time advantage. H. Customers move on to the next billing

cycle, giving them time to repay later within the given time.

Credit card fraud is an easy target. Without the risk, a significant amount of money can be withdrawn in a short period of time without the owner's knowledge. Scammers are always trying to legitimize every fraudulent transaction.

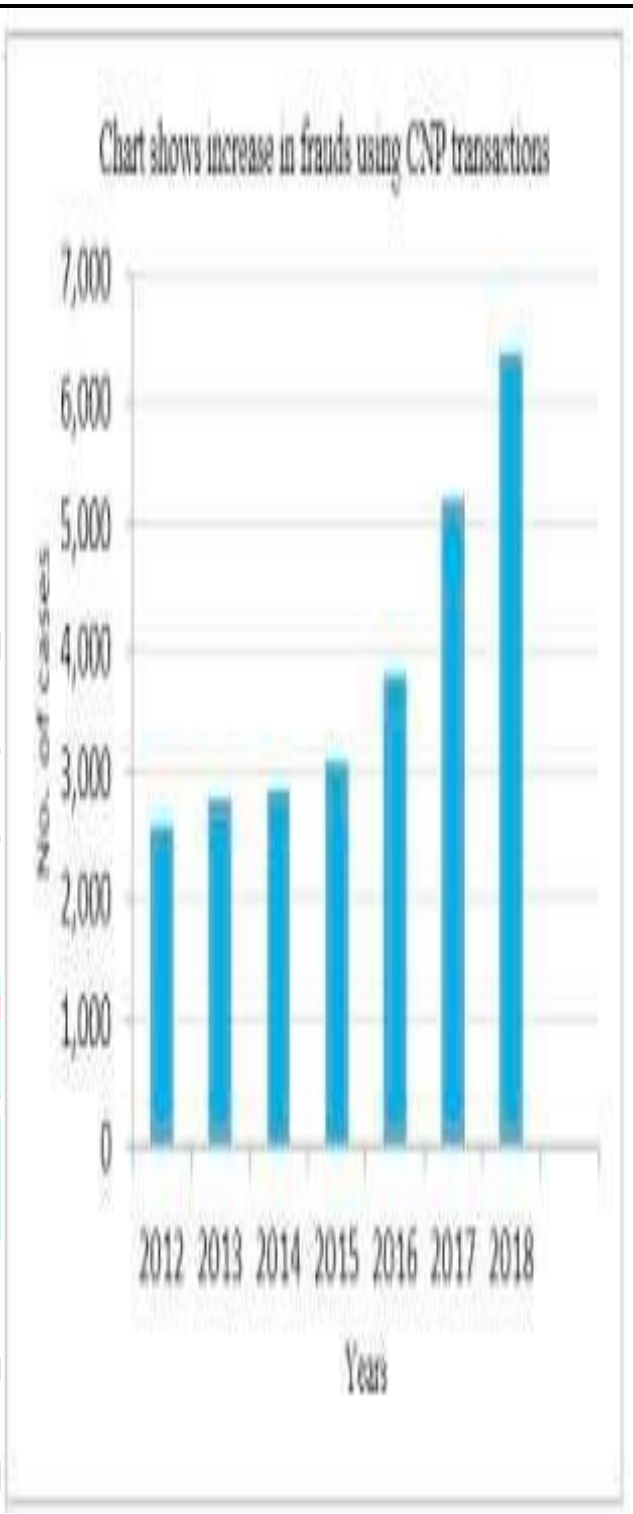
In 2017, there were 1,579 data breaches and nearly 179 million records among which Credit card frauds were the most common form with 133,015 reports, then employment or tax-related frauds with 82,051 reports, phone frauds with 55,045 reports followed by bank frauds with 50,517 reports from the statics released by FTC [10].



Fraud is a priority for most people around the world, as various scams, primarily credit card fraud, have been in the news so often in recent years. With more legitimate transactions than fraudulent transactions, credit card records are noticeably imbalanced.

As progress is being made, banks are switching to his EMV card. An EMV card is a smart card that stores data on an integrated circuit rather than on a magnetic strip. While this has made some card payments more secure, it still increases the level of installment payments for cardless fraud.

According to 2017 [10], the US Payments Forum report, criminals have shifted their focus on activities related to CNP transactions as the security of chip cards were increased. Fig 2, shows the number of CNP frauds cases that were registered in respective years.



Even then there are chances for thieves to misuse the credit cards. There are many machine learning techniques to overcome this problem.

2. RELATED WORK

Ashen Teal (2007) demonstrated the efficiency of classification models for the problem of credit card fraud detection, and the authors proposed three classification models. H. Decision trees, neural networks, logistic regression. Of the three models, neural networks and logistic regression outperform decision trees. Mojica, etc. (2007) proposed a probability theory framework for making decisions under uncertainty. After confirming the Bayesian theory, a naive Bayesian classifier and a k-nearest neighbor classifier are implemented and applied to the credit card system dataset. Shahin and E. Doman (2011), citing research on credit card fraud detection, used seven classification methods that played an important role. This work included decision trees and SVMs to de-risk banks. They suggest that artificial neural networks and logistic regression classification models can help improve fraud detection performance. Y Shahin E Doman (2011) cited this work and used artificial neurals. Network and logistic regression classification, as well as the ANN classifier described, outperform the LR classifier in solving the problem under study. Here, the training set distribution was more skewed, the training set distribution was more skewed, and all models were less efficient at detecting fraudulent transactions.

Several supervised and semi-supervised machine learning methods are used for fraud detection [8]. Detect fraudulent transactions in real-time datasets using machine learning algorithms [3] such as decision trees, naive Bayesian classification, least-squares regression, logistic regression, and SVM. We use his two methods of random forests [6] to train the behavioral characteristics of normal and abnormal transactions.

3. PROPOSED TECHNIQUE:

The proposed techniques are used in this paper, for detecting the frauds in credit card system. The comparison is 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.

Performance of Logistic Regression, K-Nearest Neighbour, and Naïve Bayes are analysed on highly skewed credit card fraud data where Research is carried out on examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data.

Through supervised learning methods can be used there may fail at certain cases of detecting the fraud cases. A model of deep Auto-encoder and restricted Boltzmann machine (RBM) [2] that can construct normal transactions to find anomalies from normal patterns. Not only that a hybrid method is developed with a combination of Adaboost and Majority Voting methods [4].

The processing steps are discussed in Table 1 to detect the best algorithm for the given dataset

Table 1: Processing steps

Algorithm steps:

Step 1: Read the dataset.
Step 2: Random Sampling is done on the data set to make it balanced.
Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.
Step 4: Feature selection are applied for the proposed models.
Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.
Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

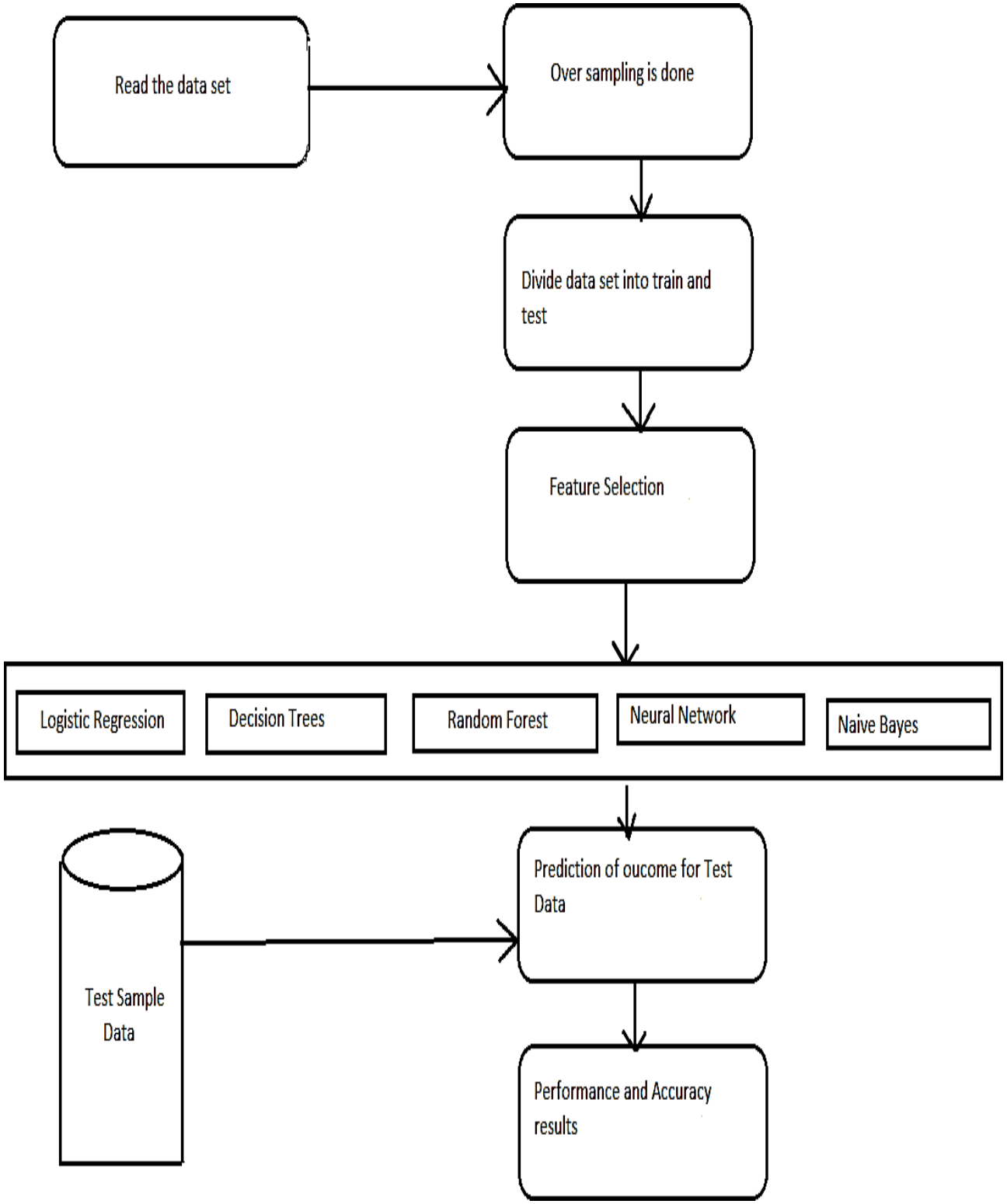


Figure1: System Architecture

3.1 Logistic Regression:

Logistic regression is one of the classification algorithms used to predict the binary value of a given set of independent variables (1/0, yes/no, true/false). Dummy variables are used to represent binary/categorical values. A special case of logistic regression is linear regression. If the outcome variable is categorical, the logarithm of the odds is used as the dependent variable and also predicts the probability of an event occurring by fitting the data to a logistic function. like that

$$O = e^{\wedge} (I_0 + I_1 * x) / (1 + e^{\wedge} (I_0 + I_1 * x)) \quad (3.1)$$

Where,

O is the predicted output

I_0 is the bias or intercept term

I_1 is the coefficient for the single input value (x).

Each column in the input data has an associated I coefficient (a constant real value) that must be learned from the training data.

$$y = e^{\wedge} (b0 + b1 * x) / (1 + e^{\wedge} (b0 + b1 * x)) \quad (3.2)$$

Logistic regression is started with the simple linear regression equation in which dependent variable can be enclosed in a link function i.e., to start with logistic regression, I'll first write the simple linear regression equation with dependent variable enclosed in a link function:

$$A(O) = \beta_0 + \beta(x) \quad (3.3)$$

Where

A (): link function

O: outcome variable

x: dependent variable

A function is established using two things:

1) Probability of Success(pr) and 2) Probability of Failure(1-pr).

pr should meet following criteria: a) probability must always be positive (since p >= 0) b) probability must always be less than equals to 1 (since pr <= 1). By applying exponential in the first criteria and the value is always greater than equals to 1.

$$pr = \exp (\beta_o + \beta(x)) = e^{\wedge} (\beta_o + \beta(x)) \quad (3.4)$$

For the second criteria, same exponential is divided by adding 1 to it so that the value will be less than equals to 1

$$pr = e^{\wedge} (\beta_o + \beta(x)) / e^{\wedge} (\beta_o + \beta(x)) + 1 \quad (3.5)$$

Logistic function is used in the logistic regression in which cost function quantifies the error, as it models response is compared with the true value.

$$X(\theta) = -1/m^{\star} (\sum y_I \log(he(x_I)) + (1 - y_I) \log(1 - h\theta(x_I))) \quad (3.6)$$

Where

$he(x_I)$: logistic function

y_I: outcome variable Gradient descent is a learning algorithm

3.2 Decision Tree Algorithm:

A decision tree is a type of supervised learning algorithm (with a defined target variable) used primarily in classification problems. This works for both categorical and continuous input and output variables. This technique divides a population or sample into two or more homogeneous sets (or subpopulations) based on the most important splitter/derivative of the input variables

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

TERMINOLOGY OF DECISION TREE:

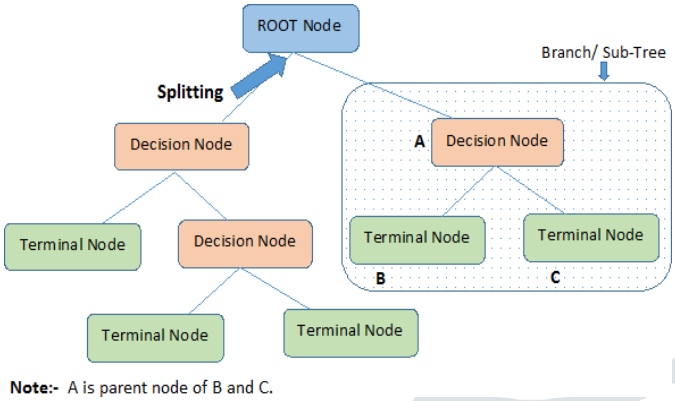
1. Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. Splitting: It is a process of dividing a node into two or more sub-nodes.
3. Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.
4. Leaf/ 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 whereas sub-nodes are the child of parent node.

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 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



Note:- A is parent node of B and C.

3.3 Random Forest:

Random Forest is a tree-based algorithm that creates multiple trees and combines them with the output to improve the generalization capabilities of the model. This method of combining trees is known as the ensemble method. Synthesis is nothing more than combining weak learners (individual trees) to produce a strong learner. Random forests can be used to solve regression and classification problems. For regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

WORKING OF RANDOM FOREST:

Create a random sample using a bagging algorithm. Given an n-by-m dataset D1, create a new dataset D2 that replaces the original data and randomly samples n cases. From dataset D1, 1/3 of the rows are omitted and called the out-of-bag sample. A new dataset D2 is then trained on this model to determine an unbiased estimate of the error using out-of-bag sampling. From my columns, M< columns are selected at each node of the dataset. The M columns are randomly selected. Usually the default choice for M is m/3 for regression trees and M is sqrt(m) for classification trees. Unlike trees, random forests do not have pruning. H. Each tree is fully grown. In decision trees, pruning is a way to avoid overfitting. Pruning means choosing the subtree with the lowest test failure rate. Cross-validation is used to determine the test failure rate of subtrees. Several trees are grown and the final prediction is obtained by averaging or voting.

Table 2: Algorithm steps for finding the best algorithm

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

4. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS:

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 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).

$$\text{Accuracy}=A+B/C+D \quad (4.1)$$

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).

$$\text{Error rate}=F+E/C+D \quad (4.2)$$

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).

$$\text{Sensitivity}=A/C \qquad (4.3)$$

Specificity:

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

Specificity = B/D. (4.4) Accuracy, error rate, sensitivity, and specificity of

are used to report the performance of credit card fraud detection systems.

In this paper, three machine learning algorithms are developed to detect fraud in credit card schemes. To evaluate the algorithm, 70% of the dataset is used for training and 30% for testing and validation. Accuracy, error rate, sensitivity, and specificity are used to evaluate various variables of the three algorithms, as shown in Table 3. Logistic Regression Accuracy results are displayed. Decision tree and random forest classifiers are 92.7, 95.8 and 97.6 respectively. Comparative results show that random forest outperforms logistic regression and decision tree techniques..

Table 3: Performance analysis for three different algorithms

Feature Selection	Logistic regression	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

5. CODE

1.) Importing Libraries:

The first step of all the projects will be always importing the required libraries.

```
# import libraries for ProjectGurukul Credit Card Fraud Detection Project us
Machine Learning:

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix , accuracy_score,
classification_report
```

2.) Load the dataset:

Load the dataset we have downloaded above which is creditcard.csv file.

```
#Loading the dataset to a Pandas Dataframe

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

```
# let's see first 5 rows of the dataset:

credit_card_data.head(5)
```

```
   Time        V1        V2        V3        V4        V5        V6        V7  \
0   0.0 -1.359807 -0.072781  2.536347  1.378155 -0.338321  0.462388  0.239599
1   0.0  1.191857  0.266151  0.166480  0.448154  0.060018 -0.082361 -0.078803
2   1.0 -1.358354 -1.340163  1.773209  0.379780 -0.503198  1.800499  0.791461
3   1.0 -0.966272 -0.185226  1.792993 -0.863291 -0.010309  1.247203  0.237609
4   2.0 -1.158233  0.877737  1.548718  0.403034 -0.407193  0.095921  0.592941

         V8        V9  ...       V21       V22       V23       V24       V25  \
0  0.098698  0.363787  ... -0.018307  0.277838 -0.110474  0.066928  0.128539
1  0.085102 -0.255425  ... -0.225775 -0.638672  0.101288 -0.339846  0.167170
2  0.247676 -1.514654  ...  0.247998  0.771679  0.909412 -0.689281 -0.327642
3  0.377436 -1.387024  ... -0.108300  0.005274 -0.190321 -1.175575  0.647376
4 -0.270533  0.817739  ... -0.009431  0.798278 -0.137458  0.141267 -0.206010

        V26       V27       V28  Amount  Class
0 -0.189115  0.133558 -0.021053  149.62      0
1  0.125895 -0.008983  0.014724    2.69      0
2 -0.139097 -0.055353 -0.059752  378.66      0
3 -0.221929  0.062723  0.061458  123.50      0
4  0.502292  0.219422  0.215153   69.99      0

[5 rows x 31 columns]
```

```
# let's see last 5 rows of our dataset:

credit_card_data.tail()
```

```
           Time         V1         V2        V3        V4        V5  \
284802  172786.0 -11.881118  10.071785 -9.834783 -2.066656 -5.364473
284803  172787.0  -0.732789  -0.055080  2.035030 -0.738589  0.868229
284804  172788.0   1.919565  -0.301254 -3.249640 -0.557828  2.630515
284805  172788.0  -0.240440   0.530483  0.702510  0.689799 -0.377961
284806  172792.0  -0.533413  -0.189733  0.703337 -0.506271 -0.012546

                V6        V7        V8        V9  ...       V21       V22  \
284802 -2.606837 -4.918215  7.305334  1.914428  ...  0.213454  0.111864
284803  1.058415  0.024330  0.294869  0.584800  ...  0.214205  0.924384
284804  3.031260 -0.296827  0.708417  0.432454  ...  0.232045  0.578229
284805  0.623708 -0.686180  0.679145  0.392087  ...  0.265245  0.800049
284806 -0.649617  1.577006 -0.414650  0.486180  ...  0.261057  0.643078

                V23       V24       V25       V26       V27       V28  Amount  \
284802  1.014480 -0.509348  1.436807  0.250034  0.943651  0.823731    0.77
284803  0.012463 -1.016226 -0.606624 -0.395255  0.068472 -0.053527   24.79
284804 -0.037501  0.640134  0.265745 -0.087371  0.004455 -0.026561   67.88
284805 -0.163298  0.123205 -0.569159  0.546668  0.108821  0.104533   10.00
284806  0.376777  0.008797 -0.473649 -0.818267 -0.002415  0.013649  217.00

        Class
284802      0
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
...
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
```

```
# checking number of missing values:

credit_card_data.isnull().sum()
```

```
# Find distribution of Normal transaction or Fraud transaction:

credit_card_data['Class'].value_counts()
```

```
0    284315
1       492
Name: Class, dtype: int64
```

Now we will separate Normal and Fraud transactions, and analyze and visualize that fraud and normal data :

```
# Separating the data:
normal = credit_card_data[credit_card_data.Class == 0]
fraud = credit_card_data[credit_card_data.Class == 1]
# check shape
print(normal.shape)
print(fraud.shape)

#visualize the data:
labels = ["Normal", "Fraud"]
count_classes = credit_card_data.value_counts(credit_card_data['Class'], sort= True)
count_classes.plot(kind = "bar", rot = 0)
plt.title("ProjectGurukul")
plt.ylabel("Count")
plt.xticks(range(2), labels)
plt.show()
```

```
#visualize the data:

labels = ["Normal", "Fraud"]
count_classes = credit_card_data.value_counts(credit_card_data['Class'], sort= True)
count_classes.plot(kind = "bar", rot = 0)
plt.title("ProjectGurukul")
plt.ylabel("Count")
plt.xticks(range(2), labels)
plt.show()
```

```
# statistical measures of the data:

normal.Amount.describe()
```

```
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, dtype: float64
```

```
fraud.Amount.describe()
```

```
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
```

We are just performing Exploratory data analysis, just follow along to understand the dataset better. And make it better so that our model can detect fraud and normal transactions accurately and efficiently.

```
# Compare values of both transactions:
credit_card_data.groupby('Class').mean()

# Now we will build a sample dataset containing similar distribution of normal transaction and fraud transaction:
normal_sample = normal.sample(n=492)
# Concat two data ( normal_sample and fraud) to create new dataframe which consist equal number of fraud transactions and normal transactions, In this way we balance our dataset (As our dataset is highly unbalanced initially) :
credit_card_new_data = pd.concat([normal_sample, fraud], axis=0)
Let's see our new dataset:
credit_card_new_data

# Analyse our new dataset:
credit_card_new_data['Class'].value_counts()
```

4) Splitting the data:

After analyzing and visualizing our data, now we will split our dataset in X and Y or say in features and labels:

```
# Splitting data into features and targets
X = credit_card_new_data.drop('Class', axis=1)
Y = credit_card_new_data['Class']

# splitting the data into training and testing data:
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.2, stratify = Y, random_state= 2)
print(X.shape, X_train.shape, X_test.shape)
```

5.) Creating Logistic Regression Model:

Now we will create the machine learning model.

```
# Creating Model:
model = LogisticRegression()
# training the Logistic Regression model with training data:
model.fit(X_train,Y_train)
```

6.) Model Evaluation:

After fitting the data into the model we have to perform model evaluation to check the accuracy of the model.

```
# Model Evaluation
X_train_pred = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_pred, Y_train)
print('Accuracy of Training data:', training_data_accuracy)
```

```
In [30]: # Model Evaluation
         X_train_pred = model.predict(X_train)
         training_data_accuracy = accuracy_score(X_train_pred, Y_train)
         print('Accuracy of Training data:', training_data_accuracy)
Accuracy of Training data: 0.9504447268106735
```

As you can see the model we have created gives 95% accuracy on training data. The accuracy is very good as we are training our model on very less data. So on considering that our model accuracy is good.

```
# classification report of the model on training data:
print(classification_report(X_train_pred, Y_train))
```

Now evaluating our model on test data:

```
# accuracy on test data:
X_test_pred = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_pred, Y_test)
print('Accuracy of Testing data:', test_data_accuracy)

# confusion matrix and classification report of test data:
print(confusion_matrix(X_test_pred, Y_test))
print(classification_report(X_test_pred, Y_test))
```

```
# accuracy on test data:

X_test_pred = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_pred, Y_test)

print('Accuracy of Testing data:', test_data_accuracy)
```

```
Accuracy of Testing data: 0.8984771573604061
```

```
print(confusion_matrix(X_test_pred, Y_test))
```

```
[[94 15]
 [ 5 83]]
```

```
print(classification_report(X_test_pred, Y_test))
```

```
              precision    recall  f1-score   support

           0       0.96      0.90      0.93       105
           1       0.90      0.96      0.93        92

    accuracy                           0.93       197
   macro avg       0.93      0.93      0.93       197
weighted avg       0.93      0.93      0.93       197
```

6. CONCLUSION

This article used machine learning techniques such as logistic regression, decision trees, and random forests to detect fraud in credit card schemes. Evaluate the performance of the proposed system using sensitivity, specificity, precision, and error rate. The accuracies of logistic regression, decision tree, and random forest classifiers are 90.0, 94.3, and
95.5 respectively. Comparing all three methods, the random forest classifier was found to outperform logistic regression and decision trees.

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. Bunsen, A. Stojakovic, D. Aoudad, B. Otters ten, "*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] 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. 194-198, 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 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.

[16] 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.

[17] 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.

[18] 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, 2012, ISSN: 2277-1581.

[19] 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.

[20] 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.

[21] R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2000.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] S.V.S.S.Lakshmi,K.S.Deepthi,Ch.Suresh "Text Summarization basing on Font and Cue-phrase

There are other ways to commit credit card fraud. Scammers are veritably talented and fast- moving people. In a conventional approach, this document identifies operation fraud where an individual provides false information about themselves to gain credit for her card. There's also lost and stolen card fraud, an important area of credit card fraud. There are smarter credit card scammers, starting with those who produce fake or fraudulent cards. Some people also use skimming to commit fraud. This information is stored on the glamorous stripe on the reverse of your credit card, or data stored on a smart chip is copied from one card to another. Website cloning and fake dealer websites on the Internet are getting popular fraud ways for numerous culprits with advanced hacking chops. similar websites are designed to trick people into furnishing their credit card details without knowing they've been scammed.
The rest of the book is described as follows. Section 2 describes applicable work on the credit card system, Section 3 describes the proposed system armature and methodology, Section 4 presents performance analysis and results, and Section 5 presents conclusions.

A credit card, generally appertained to as a card assigned to a client(cardholder), generally allows the client to buy goods and services within their credit limit or withdraw cash in advance. Credit cards give the cardholder a time advantage.H. guests move on to the coming billing cycle, giving them time to repay latterly within the given time.
Credit card fraud is an easy target. Without the threat, a significant quantum of plutocrat can be withdrawn in a short period of time without the proprietor's knowledge. Scammers are always trying to legitimize every fraudulent sale.
In 2017, there were,579 data breaches and nearly 179 million records among which Credit card frauds were the most common form with,015 reports, also employment or duty- related frauds with,051 reports, phone frauds with,045 reports followed by bank frauds with,517 reports from the statics released by FTC(10).

Fraud is a precedence for utmost people around the world, as colorful swindles, primarily credit card fraud, have been in the news so frequently in recent times. With further licit deals than fraudulent deals, credit card records are noticeably imbalanced.
As progress is being made, banks are switching to his EMV card. An EMV card is a smart card that stores data on an intertwined circuit rather than on a glamorous strip. While this has made some card payments more secure, it still increases the position of investiture payments for cardless fraud.