



LOVELY
PROFESSIONAL
UNIVERSITY

Credit Card Fraud Detection

SUBMITTED BY

NAME : SOYEL AKTER HABIB
REG NO : 12013562
SUBJECT : INT 254

[Click for Github Project Link](#)

Link : <https://github.com/Soyel-Akter-Habib/Credit-Card-Fraud-Detection.git>

Contents

Abstract

Introduction

Visualization and Preprocessing

Keywords and Definition

Proposed Architecture

Training and Testing

Prediction

Results

Conclusion

References

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 logistic regression, 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 logistic regression, 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 logistic regression and decision tree techniques

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 contain the following fields:

- 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.

The Data Set

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

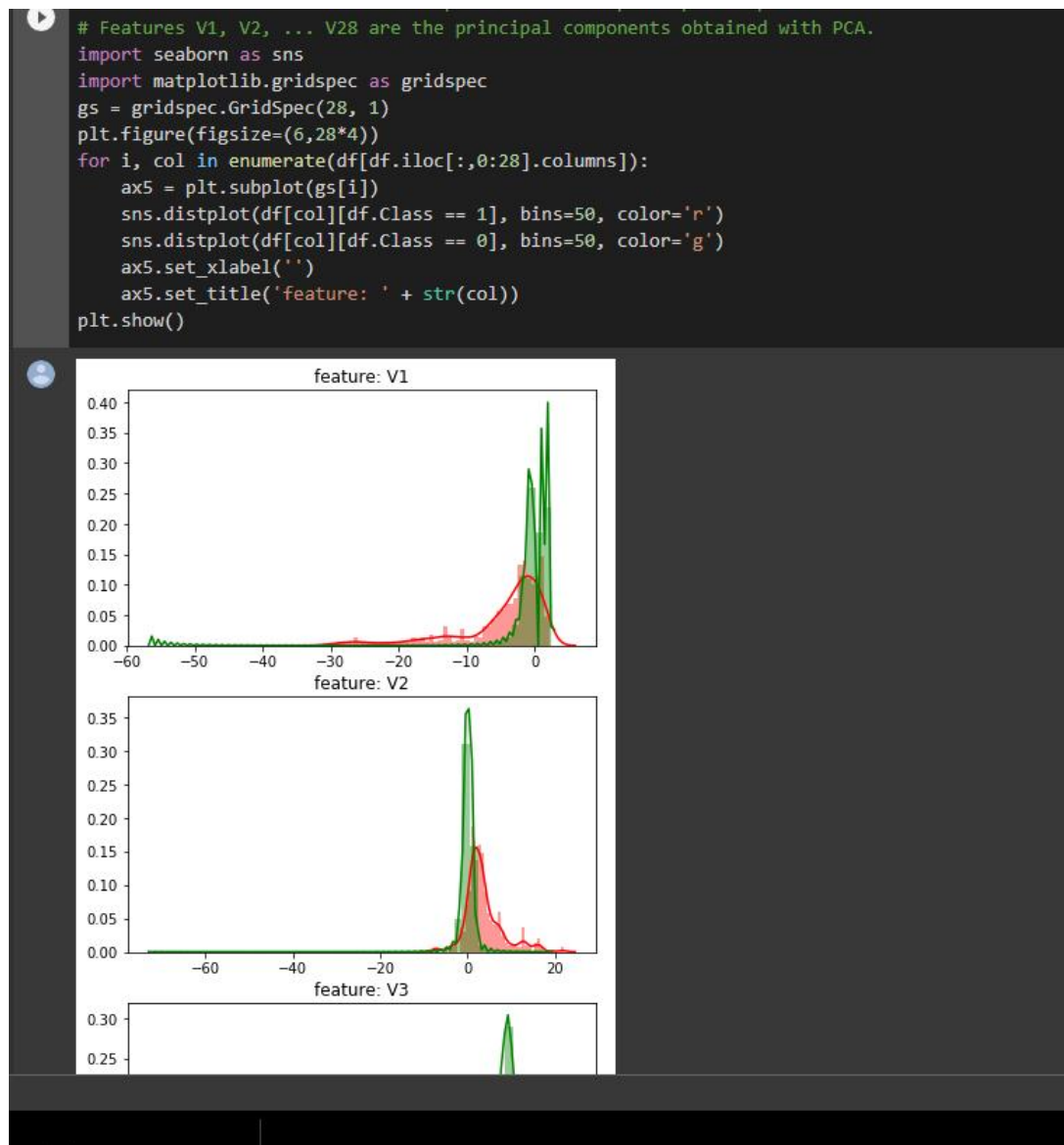
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.

The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

VISUALIZATION AND PREPROCESSING:



For some of the features, both the classes have similar distribution. So, I don't expect them to contribute towards classifying power of the model. So, it's best to drop them and reduce the model complexity, and hence the chances of overfitting. Ofcourse as with my other assumptions, I will later check the validity of above argument.

Now, it's time to split the data in test set (20%) and training set (80%). I'll define a function for it.

Numpy :

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas :

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.

PROPOSED ARCHITECTURE :

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.

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.

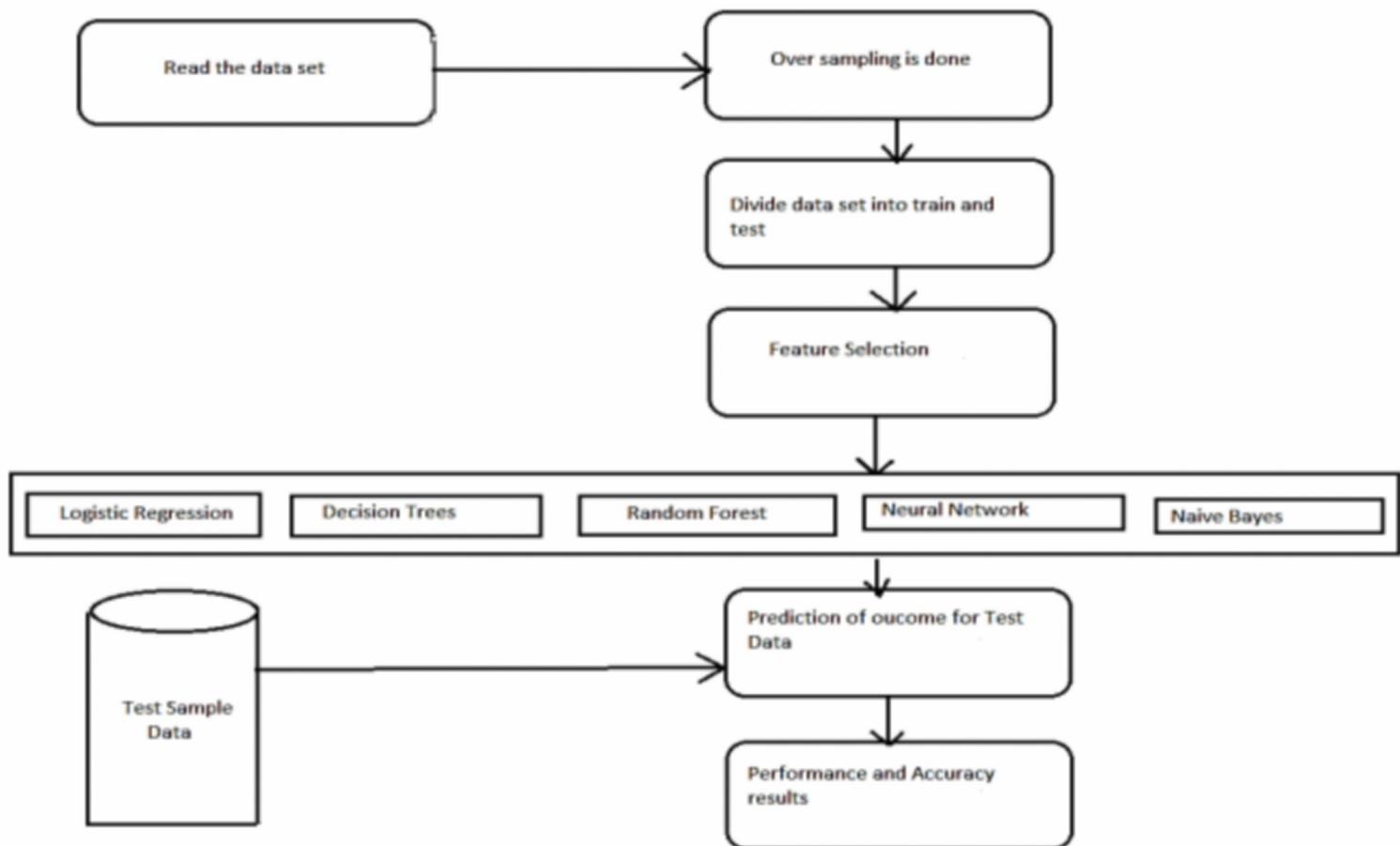


Figure1: System Architecture

YouTube

The Complete Machine Learning

Fraud-Detection-Solution.ipynb

WhatsApp

11809946

https://colab.research.google.com/drive/1xvZISzgGMDR6CDJucRi_G7-TrH7bwFLZ#scrollTo=qw

Fraud-Detection-Solution.ipynb

File Edit View Insert Runtime Tools Help Last saved at 2:30 AM

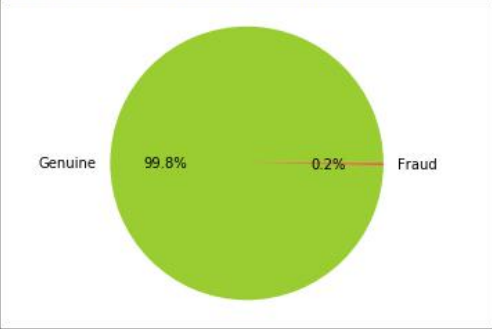
+ Code + Text

[]

V24 284807 non-null float64
V25 284807 non-null float64
V26 284807 non-null float64
V27 284807 non-null float64
V28 284807 non-null float64
Amount 284807 non-null float64
Class 284807 non-null int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
None
'Since all variables are of float and int type, so this data is easy to handle for modeling'

Check Class variables that has 0 value for Genuine transactions and 1 for Fraud
print("Class as pie chart:")
fig, ax = plt.subplots(1, 1)
ax.pie(df.Class.value_counts(), autopct='%1.1f%%', labels=['Genuine', 'Fraud'], colors=['yellowgreen', 'r'])
plt.axis('equal')
plt.ylabel('')

Class as pie chart:
<matplotlib.text.Text at 0x7f6e969012e8>



[]

#plot Time to see if there is any trend
print("Time variable")
df["Time_Hr"] = df["Time"]/3600 # convert to hours
print(df["Time_Hr"].tail(5))
fig, (ax1, ax2) = plt.subplots(2, 1, sharex = True, figsize=(6,3))
ax1.hist(df.Time_Hr[df.Class==0], bins=48, color='g', alpha=0.5)

Logistic Regression:

Logistic Regression is one of the classification algorithm, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function. Such as $O = e^{(I_0 + I_1 * x)} / (1 + e^{(I_0 + I_1 * x)})$ 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. $y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$ (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: $A(O) = \beta_0 + \beta(x)$ (Where $A()$: link function O : outcome variable x : dependent variable A function is established using two things: 1) Probability of Success(p_r) and 2) Probability of Failure($1-p_r$). p_r should meet following criteria: a) probability must always be positive (since $p \geq 0$) b) probability must always be less than equals to 1 (since $p_r \leq 1$). By applying exponential in the first criteria and the value is always greater than equals to 1. $p_r = \exp(\beta_0 + \beta(x)) = e^{(\beta_0 + \beta(x))}$ For the second criteria, same exponential is divided by adding 1 to it so that the value will be less than equals to 1 $p_r = e^{(\beta_0 + \beta(x))} / e^{(\beta_0 + \beta(x))} + 1$ (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. $X(0) = -1/m * (\sum y_i \log(h_0(x_i)) + (1-y_i) \log(1-h_0(x_i)))$ Where $h_0(x_i)$: logistic function y_i : outcome variable Gradient descent is a learning algorithm.

Algorithm steps for finding the Best algorit

- 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.

How does Credit Card Fraud work?

A credit card is one of the most used financial products to make online purchases and payments such as gas, groceries, TVs, traveling, shopping bills, and so on because of the non-availability of funds at that instance. Credit cards are of most value that provide various benefits in the form of points while using them for different transactions. There are several categories of credit card fraud that are prevalent in today's time:

- **Lost/Stolen cards:** People steal credit cards from the mail and use them illegally on behalf of the owner. The process of blocking credit cards that have been stolen and re-issuing them is a hassle for both customers and credit card companies. Some financial institutions keep the credit cards blocked until it is verified that the rightful owner has received the card.
- **Card Abuse:** The customer buys goods and items on the credit card but has no intention to pay back the amount charged by the bank for the same. These customers stop answering the calls as the deadline to settle the dues approaches. Sometimes they even declare bankruptcy—this type of fraud results in losses of millions every year.
- **Identity Theft:** The customers apply illegitimate information, and they might even steal the details of a genuine customer to apply for a credit card and then misuse it. In such cases, even card blocking can not stop the credit card from falling into the wrong hands.
- **Merchant Abuse:** Some merchants show illegal transactions (that never occurred) for money laundering. For performing these illicit transactions, legal information of genuine credit card users is stolen to generate replicas of the cards and use it for illegal work.

Challenges Faced :

There were many challenges faced by us during the project. The very first issue we faced was of dataset. We wanted to deal with raw images and that too square images as CNN in Keras as it was a lot more convenient working with only square images. We found existing dataset for that hence we decided to make our project. Second issue was to select a filter which we could apply on our images so that proper features of the images could be obtained and hence then we could provided that image as input for CNN model. We tried various filter including binary threshold.. More issues were faced relating to the accuracy of the model we trained in earlier phases which we eventually improved by increasing the input image size and also by improving the dataset.

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.

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 subnodes, 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.*

PERFORMANCE METRICS AND EXPERIMENTAL RESULTS:

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). $Accuracy = \frac{A+B}{C+D}$

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 the total number of the dataset(C+D).

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

Whereas,

E=False Positive

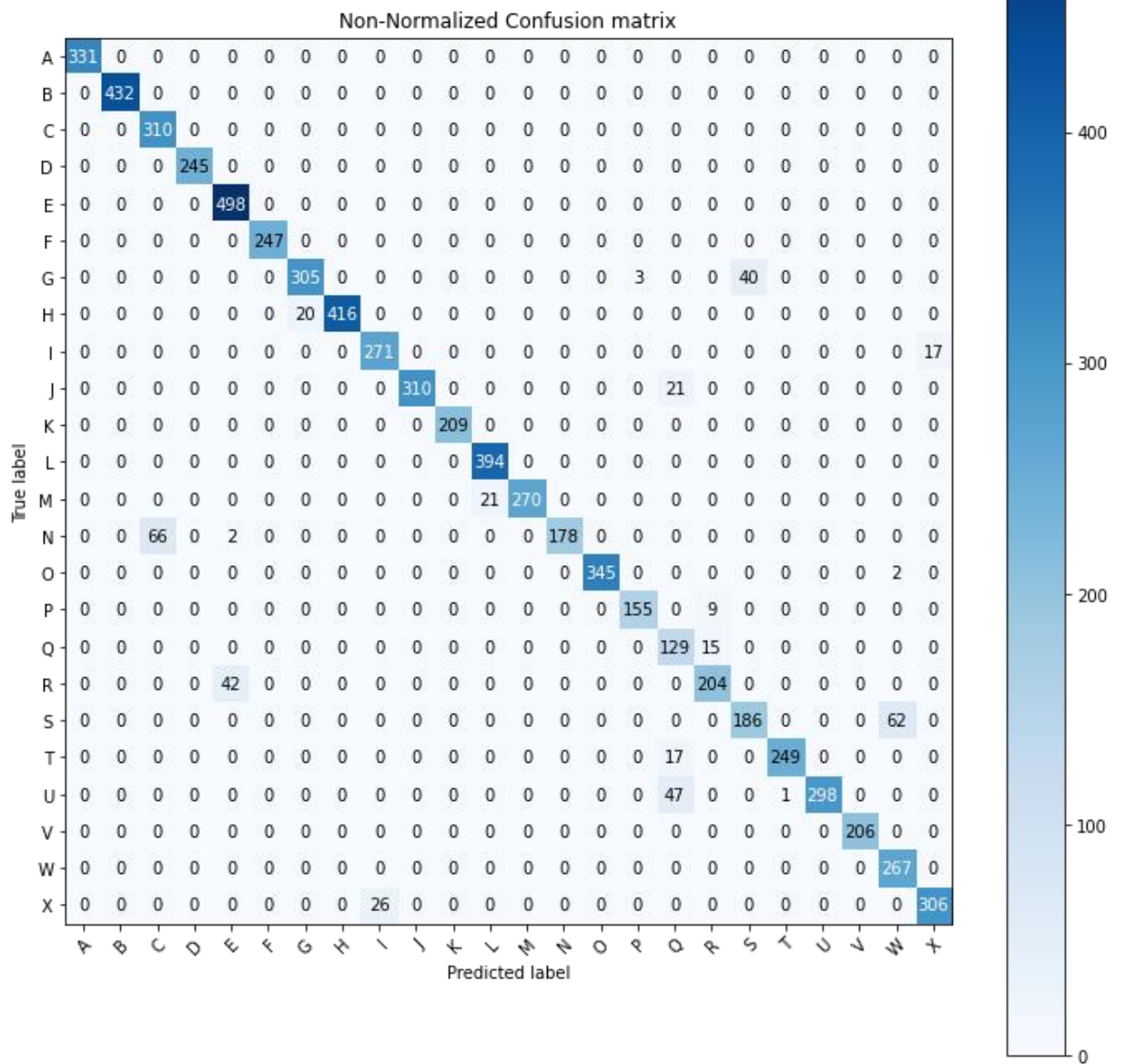
F=False Negative

C=Positive

D=Negative

Results :

We have achieved an accuracy of 94.2% in our model using only cnn of our algorithm , which is a better accuracy then most of the current research papers on american sign language.



Conclusion :

In this paper, Machine learning technique like Logistic regression, 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 logistic regression, Decision tree and random forest classifier are 90.0, 94.3, and 95.5 respectively. By comparing all the three method, found that random forest classifier is better than the logistic regression and decision tree.

Reference:

- [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.

