Neural approach for credit card Fraud Detection

Name - Vikash Rajput
Roll No. - 1901216

Introductio n & Problem Statement

Now-a-days, most of us are using credit card for buying the necessary goods which are so much in need but can’t afford at the moment. In order to meet the needs credit card are used and the fraud associated with it.. The fraud in credit card transaction occurs when the stealer uses the other person card without authorization of the respective person by stealing the necessary information like PIN, password and other credentials with or without the physical card.

Credit card fraud is happening in all organization such as appliances industry,automobile industry,bank and so on. Many of the process like machine learning algorithmic approaches are applied to identify the fraud in the credit card.

Problem Definition

Credit Card fraud involves stealing the essential credentials from the cardholder and using it unauthorized manner by the fraudsters either by using phone calls or SMS and withdraw their cash and use for online payments.
Credit card fraud is an offensive activity, carried out by an unauthorized person by cheating innocent. This fraud in credit card may also happen using some software Applications that are under the control of fraudsters.

The credit card fraud detection takes place as :the user or the customer enters the necessary credentials in order to make any transaction using credit card and the transaction should get approved only upon being checked for any fraud activity. For this to happen, we first pass the transaction details to the verification module where, it is classified under fraud and non-fraud categories. Any transaction that is put under fraud category is rejected. Otherwise , the transaction gets approved.

Classification of Credit card frauds

Application fraud :- When a fraudster acquires the controls over the application , steals the credentials of customer, and makes fake account and then the transactions takes place.

Electronic and manual card imprints:-In this form of fraud, the fraudster skim the information from the magnetic stripe which is present on card then uses the credentials and fraud transactions are carried out

Card not present:- The type of fraud in which credit card is not physically present during transactions . Mainly during online transaction fraudster uses credential of customers.

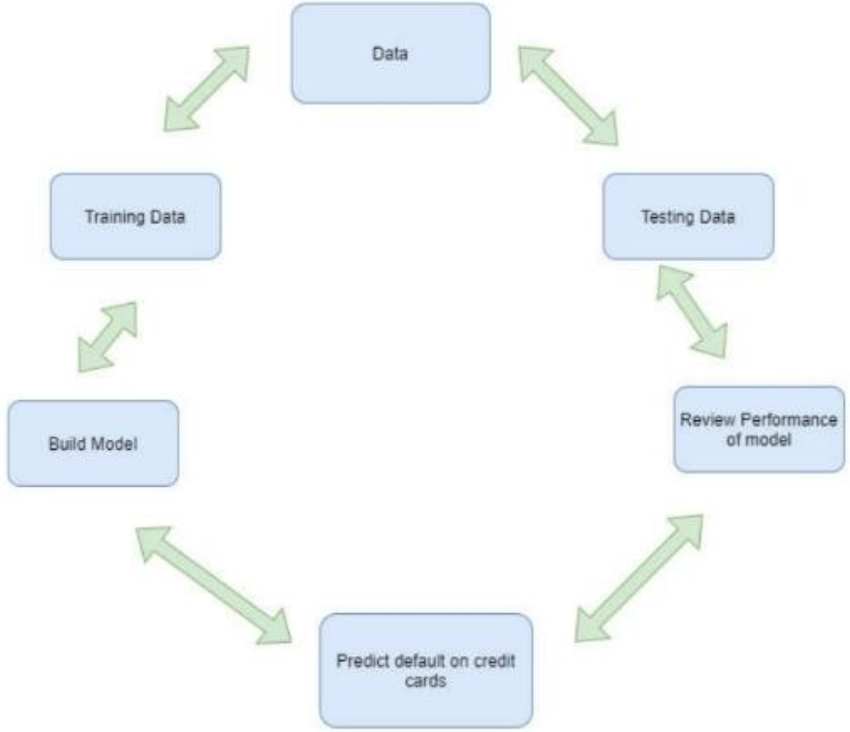
Lost/Stolen card:- This type of fraud is due to lossing of the card by the cardholder or stealing card by the fraudster from the cardholder and used it for the transaction.

Account takeover:- In this type of fraud fraudster takes the complete control over the account holder and make a fraud .

Objective

>The main objective of this project is to find fraudulent transactions in credit card transactions using payment history of the customer using neural networks on random data samples

Methodology



Data

Testing Data

Review Performance of model

Predict default on credit cards

Build Model

Training Data

Data Description

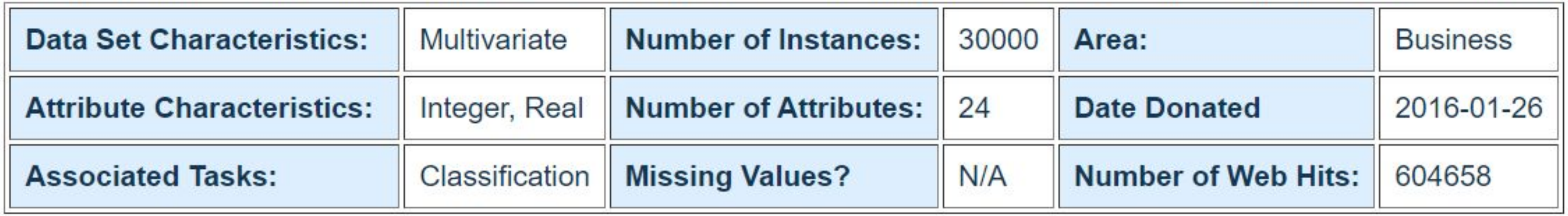
The dataset contains 25 variables out of which one is a dependent variable i.e. default payment next month(categorial variable) and 23 are independent variables and 1 is a non-predictive variable.

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Dataset description :-

- Number of features = 23
- Number of attributes = 24
- Number of patterns = 30000
- Category of model = Classification
- Classes = 0 and 1



Data Set Characteristics:	Multivariate	**Number of Instances:**	30000	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	24	**Date Donated**	2016-01-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	604658

Data Preparation

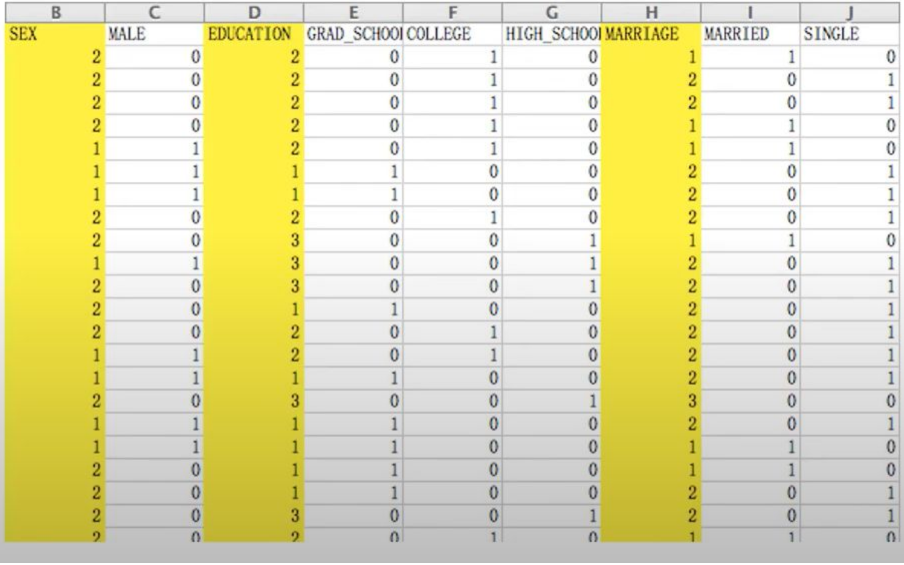
21 attributes including gender, age, education and our label

30,000 records

Male	GradSchool	Age	…	If default
0	0	24	…	0
0	0	26	…	1
0	1	34	…	1
1	0	37	…	0
….	….	….	….	….

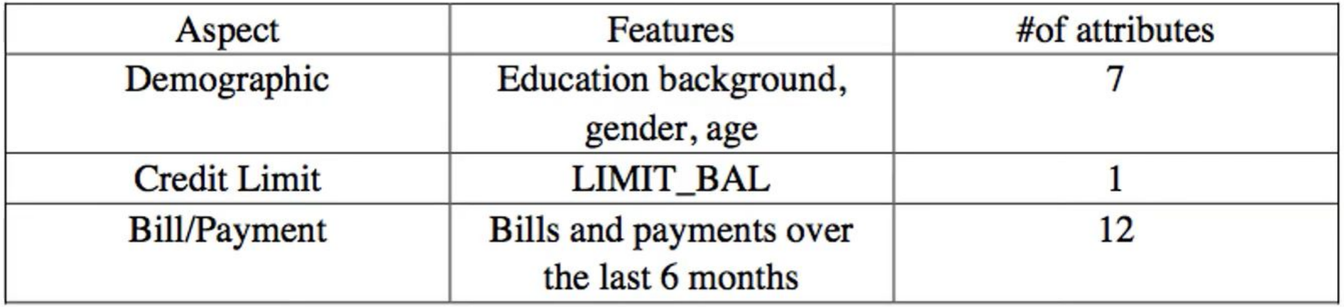
Dealing with Dummy Variables

Original dataset contains categorical variables "education", "sex", "marriage" that have numeric values ranging from 1-4. We converted them into dummy variables to fit models on them.



B	C	D	E	F	G	H	I	J
SEX	MALE	EDUCATION	GRAD_SCHOOI	COLLEGE	HIGH_SCHOOI	MARRIAGE	MARRIED	SINGLE
2	0	2	0	1	0	1	1	0
2	0	2	0	1	0	2	0	1
2	0	2	0	1	0	2	0	1
2	0	2	0	1	0	1	1	0
1	1	2	0	1	0	1	1	0
1	1	1	1	0	0	2	0	1
1	1	1	1	0	0	2	0	1
2	0	2	0	1	0	2	0	1
2	0	3	0	0	1	1	1	0
1	1	3	0	0	1	2	0	1
2	0	3	0	0	1	2	0	1
2	0	1	1	0	0	2	0	1
2	0	2	0	1	0	2	0	1
1	1	2	0	1	0	2	0	1
1	1	1	1	0	0	2	0	1
2	0	3	0	0	1	3	0	0
1	1	1	1	0	0	2	0	1
1	1	1	1	0	0	1	1	0
2	0	1	1	0	0	1	1	0
2	0	1	1	0	0	2	0	1
2	0	3	0	0	1	2	0	1
2	0	2	0	1	0	1	1	0

Data Preparation: Features



Aspect	Features	#of attributes
Demographic	Education background, gender, age	7
Credit Limit	LIMIT_BAL	1
Bill/Payment	Bills and payments over the last 6 months	12

Result Analysis

Classification Techniques:

List of Algorithms used for credit card fraud detection implementation

1. Logistic Regression
2. SLP(Single layer Perceptron)
3. MLP(Multilayer Perceptron)

1. Logistic Regression :-

Library Used

Sklearn
Scikit-learn is a free software machine learning library for the Python programming language. It features various classification ,regression and clustering algorithms including support vector machines,random forests ,gradient boosting, k-means

Class Used

Logistic Regression
Logistic Regression is a machine learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression the dependent variable is a binary variable that contains data coded as 1 or 0

Algorithms

```
Step-1 : Loading of dataset
            data= pd.read_excel(“filename.xls”)
Step-2 : divide data into features and target then
      Check basic  information
      If Y = 0    normal transaction
      if Y = 1  fraud transaction

Step-3: Scaling and Normalization of data
      Using StandardScaler()
      Using one-hot-encoding

Step 4: Splitting data into train , test dataset
      Using train_test_split()

Step-5  Train the model then fit the trained model
      Train the data using LogisticRegression
      Model  = LogisticRegression()

Step-6 Calculating the number of fraud, normal transaction
      And calculate accuracy,precision,recall
```

2. SLP(Single layer Perceptron) :-

1. Single layer perceptron(SLP) is a feed-forward neural network
2. SLP follows

   i) one perceptron neuron(threshold activation function)

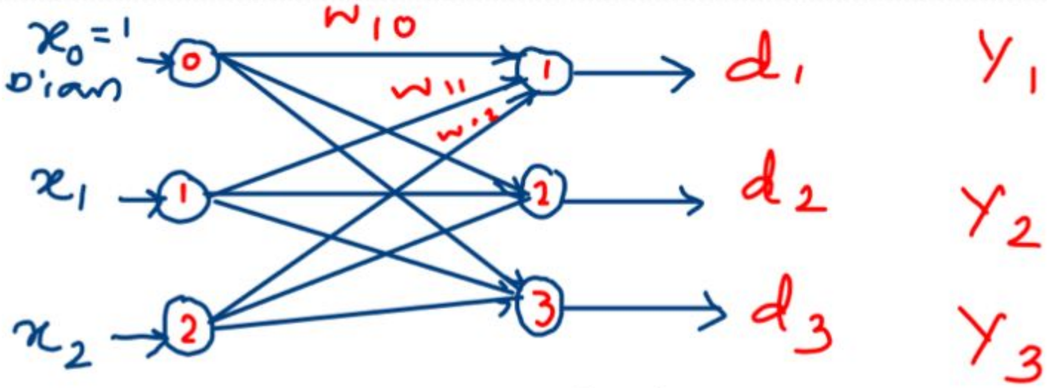
   ii)classifier model for binary classification of linearly separable patterns

It has two layers i)input layer ii)output layer

Algorithms

```
Step-1 : Loading of dataset
            data= pd.read_excel(“filename.xls”)
Step-2 : divide data into features and target then
      Check basic  information
      If Y = 0    normal transaction
      if Y = 1  fraud transaction

Step-3: Scaling and Normalization of data
      Using StandardScaler()
      Using one-hot-encoding

Step 4: Splitting data into train , test dataset
      Using train_test_split()

Step-5  Train the model then fit the trained model
      Train the data using Perceptron
      Slp = Perceptron()

Step-6 Calculating the number of fraud, normal transaction
      And calculate accuracy,precision,recall
```

Architecture for SLP



Inputs / Input neuron

Input layer

Output neurons

Output Layer

3. MLP(Multilayer Perceptron) :-

MLP(Multilayer Perceptron) is a class of feed-forward artificial neural network .

It has three layers i)input layer, ii) hidden layer(one hidden layer) , iii) Output layer.

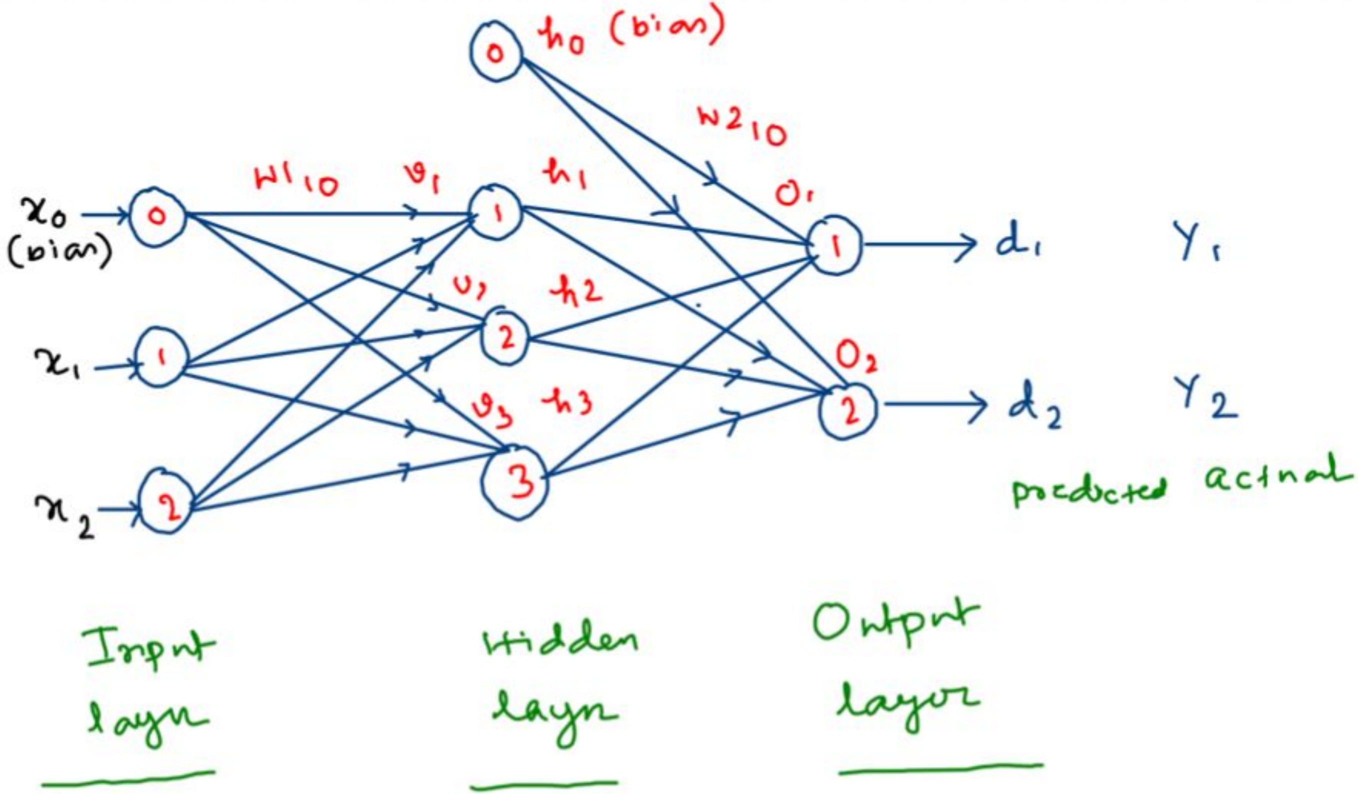
In the hidden layers I used the rectifier function as the activation function and sigmoid function for the output layer

The weight of the neurons are updated by the back propagating the errors.

Algorithm

```
Step-1 : Loading of dataset
          data= pd.read_excel("filename.xls")
Step-2 : divide data into features and target then
     Check basic  information
     If Y = 0    normal transaction
     if Y = 1  fraud transaction

Step-3: Scaling and Normalization of data
     Using StandardScaler()
     Using one-hot-encoding

Step 4: Splitting data into train , test dataset
     Using train_test_split()

Step-5  Train the model then fit the trained model
     Train the data using MLPClassifier
     Clf = MLPClassifier()

Step-6 Calculating the number of fraud, normal transaction
     And calculate accuracy,precision,recall
```

Architecture of MLP



Result and Discussion :-

Dataset :- Dataset is used the transactions made by customer in a bank . It consist of 24 columns and 30000 rows(patterns) ,in which 23 are the features and the one class is the Y (target) class which decides about whether the transaction is fraud or normal. If Y = 1 means transaction is fraud ,if Y = 0 means transaction is normal

Evaluation measure:

The end result is evaluated based on the confusion matrix and precision, recall, and accuracy is calculated. It contains two classes i).actual class and ii).predicted class

Accuracy :-

Accuracy = (true positive + true negative) / total(TP+TN+FP+FN)

Precision :-

Precision = (true positive) / (true positive + false positive)

Recall :-

Recall = (true positive) /(true positive + false negative)

Result Comparison -

Algorithms	Accuracy	Precision	Recall
1. Logistic Regression	0.78353	0.5405	0.1519
2. SLP(Single layer Perceptron)	0.6588	0.1927	0.1693
3. MLP(Multilayer Perceptron)	0.7667	0.482	0.2910

Conclusion :

In this project , we successfully implemented machine learning algorithms based on Artificial Neural Network 1. Logistic regression , 2. Single layer perceptron(SLP), 3. Multilayer Perceptron(MLP) for detecting the credit cards fraud from the given dataset and sample.

Also, Implemented algorithms Logistic Regression , SLP(single layer perceptron) and MLP(Multilayer Perceptron) using Scikit-learn libraries and also used pandas , numpy and matplot

Compared the accuracy , precision, recall of all three algorithms Logistic Regression, SLP, MLP