CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Project Report submitted in partial fulfillment of
The requirements for the degree of

BACHELOR OF COMPUTER APPLICATION

Of

WEST BENGAL UNIVERSITY OF TECHNOLOGY

By

Somnath Paul,	Roll No 29201217015
Shankhadip Das,	Roll No 29201217020
Arindam Saha,	Roll No 29201217060
Joykishan Sharma,	Roll No 29201217046
Imteaz Alam,	Roll No 29201217048

Under the guidance of

DEPARTMENT OF MCA

NETAJI SUBHASH ENGINEERING COLLEGE
TECHNO CITY, GARIA, KOLKATA – 700 152

2019-2020

CERTIFICATE

This is to certify that this project report titled **Credit Card fraud Detection Using Machine Learning** submitted in partial fulfillment of requirements for award of the degree Master of Computer Application of West Bengal University of Technology is a faithful record of the original work carried out by,

Somnath Paul,	Roll No 29201217015	Registration no: 172921010056
Shankhadip Das,	Roll No 29201217020	Registration no: 172921010051
Arindam Saha,	Roll No 29201217060	Registration no: 172921010011
Joykishan Sharma,	Roll No 29201217046	Registration no: 172921010025
Imteaz Alam,	Roll No 29201217048	Registration no: 172921010023

under my guidance and supervision.

It is further certified that it contains no material, which to a substantial extent has been submitted for the award of any degree/diploma in any institute or has been published in any form, except the assistances drawn from other sources, for which due acknowledgement has been made.

Date …….

_____
Guide's signature
Sofikul Mullick

Sd/_____

Head of the Department

MCA

NETAJI SUBHASH ENGINEERING COLLEGE
TECHNO CITY, GARIA, KOLKATA – 700 152

DECLARATION

We hereby declare that this project reporttitled

Credit Fraud Detection Using Machine Learning

Is our own original work carried out as an under graduate student in **Netaji Subhash Engineering College** except to the extent that assistances from other sources are duly acknowledged.

All sources used for this project report have been fully and properly cited. It contains no material which to a substantial extent has been submitted for the award of any degree/diploma in any institute or has been published in any form, except where due acknowledgement is made.

Student's names:	Signatures:	Dates:
……..	……..	……
……..	……..	……
……..	……..	……
……..	……..	……
……..	……..	……

CERTIFICATE OF APPROVAL

We hereby approve this dissertation titled

Credit Card Fraud Detection Using Machine Learning

carried out by

Somnath Paul,	Roll No 29201217015	Registration no: 172921010056
Shankhadip Das,	Roll No 29201217020	Registration no: 172921010051
Arindam Saha,	Roll No 29201217060	Registration no: 172921010011
Joykishan Sharma,	Roll No 29201217046	Registration no: 172921010025
Imteaz Alam,	Roll No 29201217048	Registration no: 172921010023

Under the guidance of

Piyali Madam

of Netaji Subhash Engineering College, in partial fulfillment of requirements for award of the Bachelor of Computer Application of West Bengal University of Technology

Date ……..

Examiners' signatures:

1. …….

2. …….

3. …….

ACKNOWLEDGEMENT AND/OR DEDICATION

We are very thankful to our college ***Netaji Subhash Engineering college*** *and we are also thankful to our project guide* ***Piyali mam*** *of our college* ***Netaji Subhash Engineering college.*** we are thankful to our training guide **Mr. *Sofikul sir*** of ENGINEERS ***STUDY CENTRE*** for giving us opportunity to do this awesome **project Credit Card Fraud detection using machine learning.**

We are also thankful to www.kaggle.com/ for providing us with the Dataset to do this project.

Date ……

ABSTRACT

Human beings always search for methods, tools, or techniques that reduce the human effort for performing a certain task efficiently. In Machine Learning, algorithms are designed in such a way that they try to learn by themselves using past experience. After learning from the past experience, the algorithms become quite capable of reacting and responding to conditions for which they are not explicitly programmed. So, Machine Learning helps a lot when it comes to fraud detection. It tries to identify hidden patterns that help in detecting a fraud which is not been previously recognized. Also, its computation is fast as compared to the traditional rule-based approaches.

In this project, a technique for `Credit Card Fraud Detection' is developed using Machine Learning. As fraudsters are increasing day by day. And fallacious transactions are done by the credit card and there are various types of fraud. So to solve this problem a technique is used like Logistic Regression using Machine Learning. By this transaction is tested individually and whatever suits the best is further proceeded. And the foremost goal is to detect fraud by filtering the above techniques to get better result.

Why do we use Machine Learning in Fraud Detection?

Here are some factors for why Machine Learning techniques are so popular and widely used in industries for detecting frauds:

- **Speed**: Machine Learning is widely used because of its fast computation. It analyzes and processes data and extracts new patterns from it within no time. For human beings to evaluate the data, it will take a lot of time and evaluation time will increase with the amount of data.
- **Scalability**: As more and more data is fed into the Machine Learning-based model, the model becomes more accurate and effective in prediction.
- **Efficiency**: Machine Learning algorithms perform the redundant task of data analysis and try to find hidden patterns repetitively. Their efficiency is better in giving results in comparison with manual efforts. It avoids the occurrence of false positives which counts for its efficiency.

Key words:
Credit Card Fraud Problem, Computation, Machine Learning, Data Analysis, Visualization

CONTENTS

INTRODUCTION

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

There are three core types of machine learning- supervised learning, unsupervised learning, and reinforcement learning.

1. ***Supervised Learning***: - The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term supervised refers to a set of samples where the desired output signals (labels) are already known.

2. ***Unsupervised learning***: - It is the training of an artificial intelligence algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

3. ***Semi-supervised Learning***: - It uses both labeled and unlabeled data for training- typically a small amount of labeled data and large amount of unlabeled data(because unlabeled data is less expensive and take less effort to acquire).

4. ***Reinforcement Learning***: - This learning used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.

This Credit Card Fraud Detection analysis model aims to determine that the transaction through card is fraud or not? Based on the simple tests data.

I. Gathering the Sample Data
II. Testing the Sample data
III. Predicting the future transaction (Yes or No).

Problem Definition: -

Credit card fraud detection is challenging task for the user. Online payment does not require physical card. And if anyone who know the details of card can make the transactions. Currently, card holder comes to know only after the fraud transaction is carried out. No proper mechanism is there to track the fraud transaction.

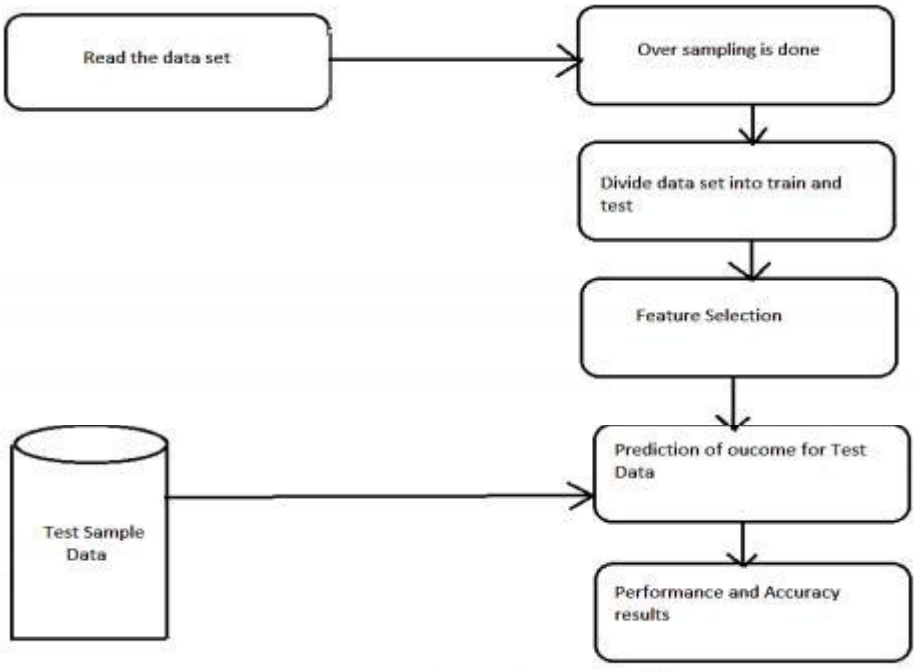
Objective: -

The overall objective of this project is listed below:

- To reduce number of fraud transaction.
- To use credit card safely for online transaction.
- To add layer of security.

Methodology:

The proposed techniques are used in this project, for detecting the frauds in credit card system. The algorithms used is Logistic Regression to determine best result and can be adapted by credit card merchants for identifying fraud transactions. Below shows the architectural diagram for representing the overall system framework.



RESOURCES USED

System Feasibility Study: The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are:

1. ECONOMICAL FEASIBILITY
2. TECHNICAL FEASIBILITY

Economical Feasibility: This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system is well within the budget and this was achieved, because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility: This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement; as only minimal or null changes are required for implementing this system.

HARDWARE & SOFTWARE REQUIREMENTS

HARDWARE USED

1. Intel Core i3 (3rd generation, 2.4 GHz, Cache 3M)
2. 4 GB DDR3 Ram
3. 500 GB Hard Disk
4. Intel HD Graphics

SOFTWARE USED

1. Anaconda (Jupyter) 2019.10
2. Python 3.7

MINIMUM REQUIREMENT FOR THIS PROJECT

1. 4GB of RAM
2. 20GB of HDD (Free Space)
3. Windows 7 or later version Operating System
4. Anaconda (Jupyter) 2019.10
5. Python 3.7

MACHINE LEARNING (BACKGROUND STUDY & METHOD)

Techniques of Supervised Machine Learning algorithms include linear and logistic regression, decision trees and support vector machines.

Supervised learning requires that the data used to train the algorithms is already labelled with correct answer supervised learning problems can further be classified into Regression and Classification problems. Both problems have as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

Regression

A regression problem is when the output variable is a real or a continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper –plane which goes through points.

Examples: - Linear Regression, Decision tree, Logistic Regression, SVR, GPR

Classification

A classification problem is when the output variable is a category, such as "red" or "blue" or" disease" and "no disease". A classification models attempts to draw some conclusion from observed values. Classification models includes Logistic Regression, Support Vector Machine,Naive Bayes

Our Project can be implemented with the help of the classification and Regression techniques:

Logistic Regression

A logistic model tree basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with k values, the node has k child nodes, and instances are sorted down one of the k branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise.

Advantages:

- Logistic regression is designed for this purpose! The dependent variableλ must be categorical, and the explanatory variables can take any form; both of which are satisfied by your problem.
- Linear combination of parameters [math]\beta[/math] and the input vectorλ will be incredibly easy to compute.
- Given that your explanatory variables are also binary, you should be ableλ to partition your input space by outcome quite well.

Disadvantages:

- we can't solve non-linear problems with logistic regression since it's λ decision surface is linear.
- If you have 3 binary predictor variables, let's say - you onlyλ have [math]2^{3} = 8[/math] possible states for the variables. Just the same, for 4 you'd have 16 states, 5, you'd have 32. Given enough data any of those cases is easily solvable, so you might not need to go through implementing logistic regression.

Derivation of Logistic Regression Equation

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression. In fact, they proposed a class of different models (linear regression, ANOVA, Poisson Regression etc) which included logistic regression as a special case.

General equation: $g(E(y)) = \alpha + \beta x1 + \gamma x2$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α,β,γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

Important Points: GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
The dependent variable need not to be normally distributed.

It does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).

Errors need to be independent but not normally distributed.

g(y) = βo + β(Age) ---- (a)
p = exp(βo + β(Age)) = e^(βo + β(Age)) ------- (b)
p=exp(βo + β(Age)) / exp(βo + β(Age)) + 1 = e^(βo + β(Age)) / e^(βo + β (Age)) + 1 ----- (c)

p = e^y / 1 + e^y --- (d)
q = 1 - p = 1 - (e^y / 1 + e^y) --- (e)

where q is the probability of failure On dividing, (d) / (e), we get,

P/1-p=e^y

After taking log on both side, we get,

Log(p/1-p) = y

log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way. After substituting value of y, we'll get:

$$\text{Log}(P/1-p)=\beta o + \beta(\text{Age})$$

This is the equation used in Logistic Regression. Here (p/1-p) is the odd ratio. Whenever the log of odd ratio is found to be positive, the probability of success is always more than 50%. A typical logistic model plot is shown below. You can see probability never goes below 0 and above 1.

DESCRIPTION OF PROJECT

LOAD THE DATASET:

The datasets contain 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.

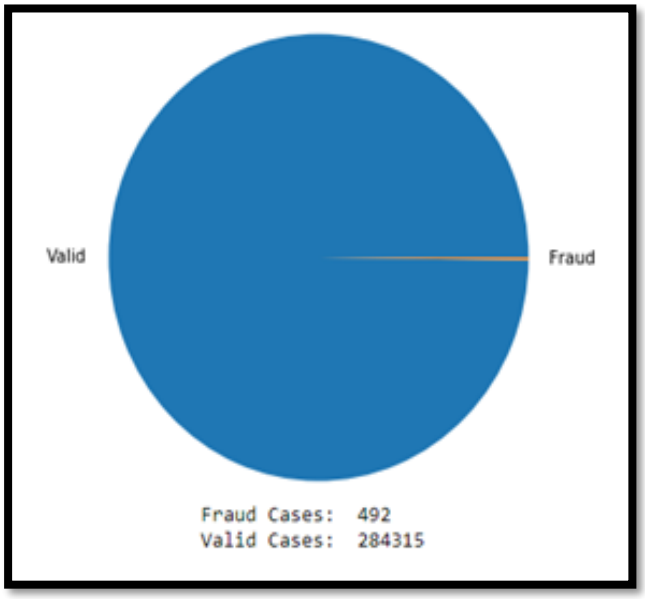
It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only 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-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

	# Time	# V1	# V2	# V3	# V26	# V27	# V28	# Amount	# Class
1	0	-1.3598071336738	-0.0727811733098497	2.53634673796914	-0.189114843888824	0.133558376740387	-0.0210530534538215	149.62	0
2	0	1.19185711131486	0.26615071205963	0.16648011335321	0.125894532368176	-0.0089830991432281 3	0.0147241691924927	2.69	0
3	1	-1.35835406159823	-1.34016307473609	1.77320934263119	-0.139096571514147	-0.0553527940384261	-0.0597518405929204	378.66	0
4	1	-0.966271711572087	-0.185226008082898	1.79299333957872	-0.221928844458407	0.0627228487293033	0.0614576285006353	123.5	0
5	2	-1.15823309349523	0.877736754848451	1.548717846511	0.502292224181569	0.219422229513348	0.215153147499206	69.99	0

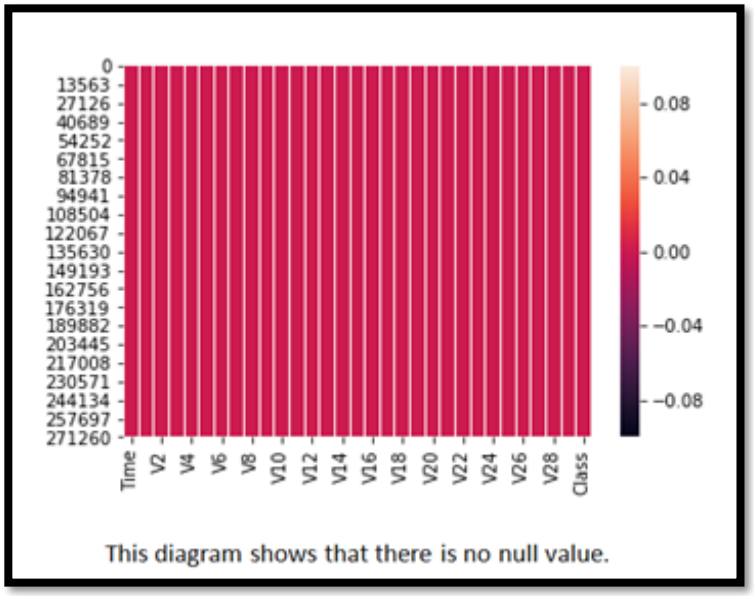
*v4, v5, v6…v24, v25 are continued with their values.

*284,807 transactions are in dataset

DATA VISUALISATION

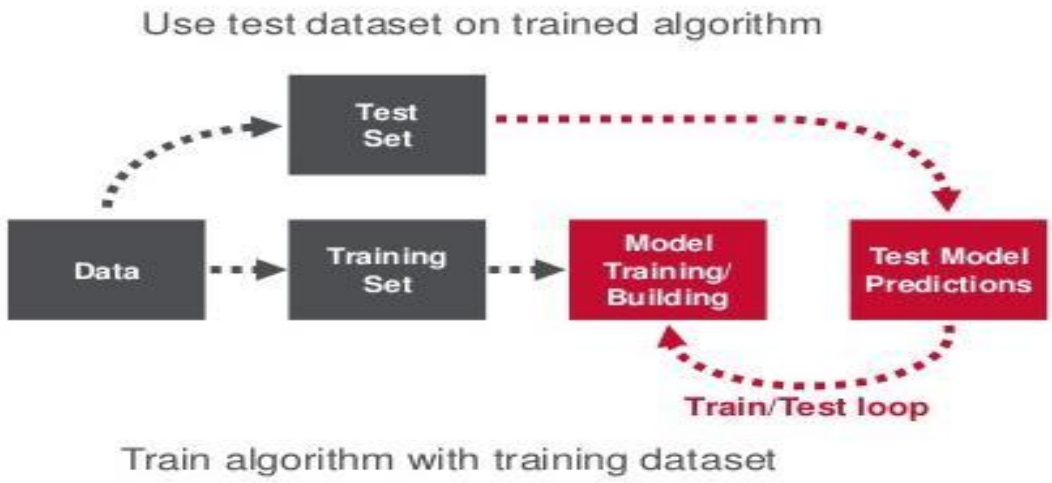


This diagram shows about the numbers of fraud and valid transactions in according to our Dataset.
In the diagram, blue colored portion represents valid transaction and orange colored portion represents fraud transactions. And after studying this diagram we can say that there is very low percentage of fraud in compare with valid transactions.



This diagram shows heatmap of our dataset, it shows non-null and null values.

FLOW CHART OF MODEL

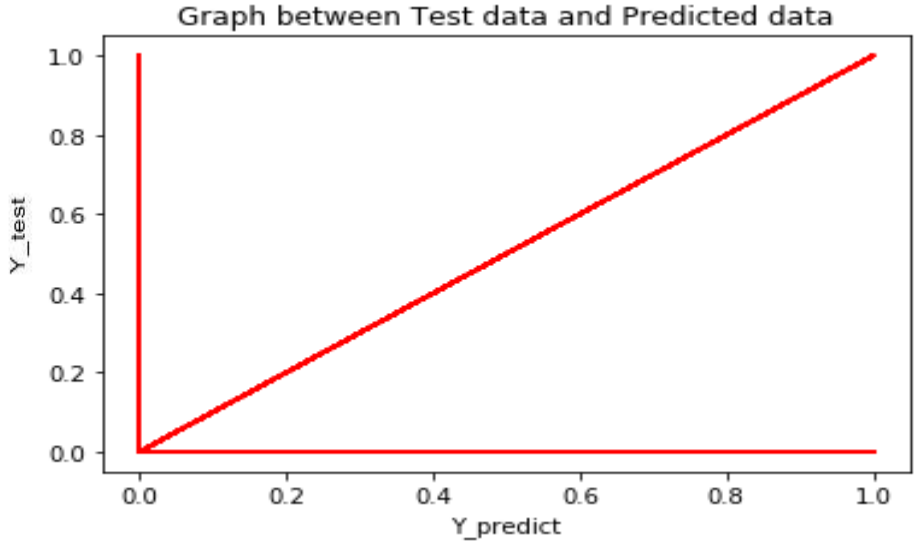


SUMMARIZING THE MODEL

In this Project, what we are doing is that we are taking tons of data on credit card transaction and feeding it into our model to train it. And when the training is complete, we test it for accuracy score. As a recommendation more the train dataset more the experience our model will get and more the experience it has better the Accuracy it will give on credit card fraud detection. In the next section we shall take a deep dive in our model's Source Code for better understanding.

MAKING SOME PREDICTION

On testing our model, we see that it makes very good accuracy result based on our actual data. This is visually expressed as below.



EFFICIENCY OF THE MODEL

Our Model gives an accuracy score of 0.99912, which means our model it 99% accurate. This accuracy score is great for prediction very precise result to determine whether the credit card transaction is fraud or valid.

```
Accuracy score: 0.99919
```

Source Code

```
# import libraries
import warnings
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 import metrics


# to ignore warning messages in output
warnings.filterwarnings('ignore')


# read the data set
data = pd.read_csv('credit_card.csv')


# no of rows and columns
print('Total rows and columns\n\n', data.shape, '\n')


# to display heatmap on null values
sns.heatmap(data.isnull())


# Dependent and independent variable
X = data.iloc[:, 1:30]
y = data['Class']


# Determine number of fraud cases and valid cases in DataSet
Fraud = data[data['Class'] == 1]
Valid = data[data['Class'] == 0]

print('Fraud Cases: ', len(Fraud))
print('Valid Cases: ', len(Valid))


# To display pie Chart on fraud/valid transaction
plt.pie([len(Valid),len(Fraud)],labels=['Valid','Fraud'],radius=1.5)
plt.show()


# splitting the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
# Build the model
clf = LogisticRegression()


# Train the classifier
clf.fit(X_train, y_train)


# test the model
y_predict = clf.predict(X_test)


# Accuracy score
a = (metrics.accuracy_score(y_test, y_predict))
print('Accuracy score:', round(a, 5))


# first 25 dataset vs first 25 predicted dataset on test dataset
print the actual and predicted labels
df1 = pd.DataFrame({'Actual': y_test, 'Predicted': y_predict})
print(df1.head(25))


# Graph between Test dataset Vs Predicted dataset
plt.plot(y_test,y_predict,color='red')
plt.title('Graph between Test data and Predicted data')
plt.xlabel('Y_predict')
plt.ylabel('X_predict')
plt.show()
```

SCOPE AND LIMITATION

The detection of credit card fraud using Machine Learning techniques have become one of the reliable approaches to counter this illegal activity. However, the process to gather real time credit card fraud data is very hard. Simply put, the situation is bleak. Fraud companies lack access to critical data about customers, and care little to help businesses (their customers) much outside of their sphere of influence.

In fact, the fraud-prevention industry, with few exceptions, is typically predatory in how they treat businesses, and their incentives could not be more misaligned. They rely on fear-tactics to over-inflate the amount of fraud that's actually happening, scare merchants about the threats of fraud as they scale, and sell safety/security instead of a focus on approving good customers. The space is overdue for change.

In the future, this study will attempt to explore more credit card fraud detections using real time data. And for improving the accuracy and precision of the model we may need to hybrid some algorithm for fast and better result.

Conclusion

Clearly, credit card fraud is an act of criminal dishonesty. This article has reviewed recent findings in the credit card field. This paper has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and discussed measures to detect them. Such measures have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms.

From an ethical perspective, it can be argued that banks and credit card companies should attempt to detect all fraudulent cases. Yet, the unprofessional fraudster is unlikely to operate on the scale of the professional fraudster and so the costs to the bank of their detection may be uneconomic.

The bank would then be faced with an ethical dilemma. Should they try to detect such fraudulent cases or should they act in shareholder interests and avoid uneconomic costs?

As the next step in this research program, the focus will be upon the implementation of a ‘suspicious’ scorecard on a real data-set and its evaluation. The main tasks will be to build scoring models to predict fraudulent behavior, taking into account the fields of behavior that relate to the different types of credit card fraud identified in this paper, and to evaluate the associated ethical implications.

References

1. For dataset https://www.kaggle.com/mlg-ulb/creditcardfraud

2. For Studying Machine Learning Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS, Author Andreas C. Müller & Sarah Guido, Publisher – O'REILLY