Fraud Detection in Credit Card using Machine Learning.

[1]Maneeksha C Ashok
Department of Computer Application
Amal Jyothi College of Engineering
Kottayam, India
maneekshacashok1997@gmail.com

[2]Shelly Shiju George
Asst.prof Department of Computer Application
Amal Jyothi College of Engineering
Kottayam, India
shellyshijugeorge@amaljyothi.ac.in

Abstract—The fast expansion of the e-commerce Sector has resulted in an exponential rise in the usage of credit cards for the online purchases, resulting in an increase in fraud. The bank does seem to have trouble discovering suspicious transactions. Credit card fraud is recognized via machine learning. Different types of machine learning methods are used by the bank to predict these purchases, collecting data history to predict fraud opportunities. The data-set sampling technique, variable selection, and detection algorithms are all used, and they all have a significant impact on the performance of credit card fraud detection in transactions. The efficacy of logistic regression a machine learning algorithm is used for finding the forgery in credit card in this work. A credit card transaction data set was gathered via Kaggle, and it comprises a total of 2,84,808 credit card transactions from a European bank. It classifies fraudulent transactions as "positive class" and authentic transactions as "negative class." The data set is substantially skewed, with around 0.172% of transactions are under forgery and the remainder being legitimate. Oversampling is used to measure an unbalanced set of data, resulting in 70% fraudulent activities and 30% legal activities. The database is under three modes, and the process is completed with Colab. Strategic sensitivity, clarity, accuracy, and error rate are all considered when evaluating their effectiveness. The accuracy of the reversal of objects is 92.0 percent, which is very high. According to the findings of this study, logistic regression can be used with a high degree of accuracy. Credit Card Fraud, Machine Learning, Attributes, Algorithms, Accuracy — Colab, Credit Card Fraud, Machine Learning, Attributes, Algorithms, Accuracy

I. INTRODUCTION:

Credit card fraud is a broad word for theft and fraud perpetrated using or utilizing a credit card at the moment of payment. The policy may be to purchase something without having to pay or withdraw money from the account without permission. User IDs are also stolen as a result of credit card fraud. The US Federal Trade Commission reports that identity theft rates stabilized by the mid-2000s, but increased by 21% in 2008. Despite the fact that the credit card.

Fraud, the crime of many people linked to identity theft, has decreased as part of the identity theft reports, Annually

2000, 10 million transactions, or 1300 transactions, were found to be counterfeit out of the total of 13 billion transactions. Furthermore, fraud was defined as 0.05 percent of all monthly active accounts (5 out of 10,000). One-twelfth of all jobs are monitored by fraud detection systems, resulting in billions of dollars in losses. Credit card fraud is one of the most serious problems businesses face today. However, in order to successfully prevent fraud, it is necessary to first comprehend the processes of fraud execution. Credit card thieves use a variety of fraudulent methods. Credit card fraud is defined as "when someone uses another person's credit card for personal reasons while the cardholder and card issuer are unaware of its use.[1]

II. DESCRIPTION

Destruction of the physical card or crucial account data, including the card account number and other records that must be disclosed to a merchant during a permitted transaction., is the starting point for card fraud. Card numbers, such as the Primary Account Number (PAN), are frequently reprinted on the card, and data is stored in machine-readable format on a magnetic stripe on the back.
It contains the following Fields:

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

Credit card fraud may be done in a variety of ways. Fraudsters are highly skilled and quick-thinking individuals. This study uses the traditional technique to identify Application Fraud, which occurs when a person provides false information about himself in order to obtain a credit card. There's also unlawful use of Lost and Stolen Cards, which accounts for a sizable portion of credit card fraud. There are more enlightened credit card criminals, beginning with those who create Fake and Doctored Cards, as well as those who conduct fraud by Skimming. They'll get this information via the magnetic strip on the back of the credit card or the data on the smart chip, which will be replicated from card to card. For many criminals with a sophisticated talent for hacking, site cloning and false merchant sites on the Internet are becoming a popular means of fraud. These kinds of websites are designed to trick consumers into giving their credit card.

DOI : 10.5281/zenodo.6372596
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Key facts

- Any type of fraud involving a credit card, such as a credit card or a bank card, is called a credit card fraud. It can be the acquisition of goods or services, or the transfer of payment to a criminal account.
- In 2018, the loss of illicit financial fraud using payment and banking cards amounting to £ 844.8 million in the United Kingdom low- and middle-income countries comprising approximately three quarters of CVD deaths.
- In 2015, 82 percent of the 17 million premature deaths (before 70 years) due to the absence of infectious diseases occurred in low- and middle-income countries, and CVDs accounted for 37 percent [2]

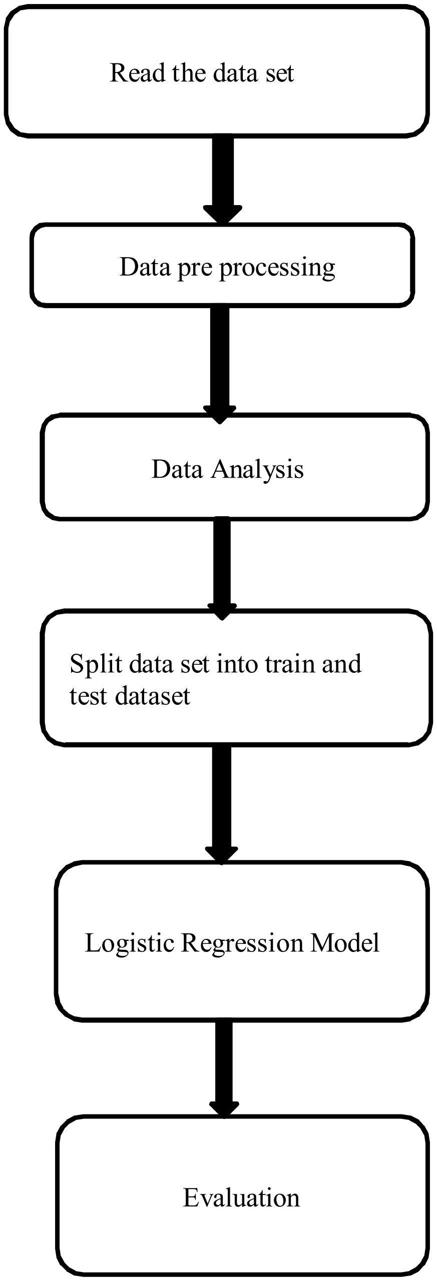
. Machine Learning

Machine learning is a study of computer algorithms that can learn and develop based on their experience and data (ML). Modification is like a cornerstone. Machine learning algorithms create a data-based training model to make predictions or judgments without the need for explicit planning. Machine learning algorithms make predictions or judgments by creating a data-based training model without being programmed in person. When traditional algorithms are difficult or impossible to do, machine learning is used in many applications, including medicine, spam filtering, voice recognition, and computer vision. [3]

Machine learning is essential because it helps firms to spot trends in customer behavior and organizational operational patterns while also assisting in the development of new products. Many of today's most successful organizations, such as Facebook, Google, and Uber, use machine learning. For many businesses, machine learning has become a significant competitive differentiation. [3]

III. PURPOSE OF THE STUDY

This study uses the traditional technique to identify Application Fraud, which occurs when a person provides false information about himself in order to obtain a credit card. There's also unlawful use of Lost and Stolen Cards, which accounts for a sizable portion of credit card fraud. There are more enlightened credit card criminals, beginning with those who create Fake and Doctored Cards, as well as those who conduct fraud by Skimming. They will receive this information through a magnetic stripe on the back of a credit card or data on a smart chip, which will be copied from one card to another.[1] In this system, the methods presented are used to for personal use while Credit card fraud is defined as "when a person uses a credit card to make a fraudulent purchase using another person's credit card for personal use while the cardholder and card issuer are unaware of it." Credit card fraud begins with credit card theft. real card or other important data associated with the account, including the card account number or other information that must be obtained from the merchant at the time of authorized purchase. The architectural design to reflect the overall structure of the system is shown in Figure 1.


Read the data set
Data pre processing
Data Analysis
Split data set into train and test dataset
Logistic Regression Model
Evaluation

IV. METHODOLOGY

This paper is all about the process and analysis of the credit card records of customers who made transactions and generating pattern or identifying the pattern or behavior of the transactions with the help of some algorithms and machine learning tools. The developing tool used is Collaboratory (Colab) [1]

COLAB

Google Research's Colaboratory, or "Colab" for short, is a product. Colab is a web-based Python editor that allows anyone to write and run arbitrary Python code. It's notably useful for machine learning, data analysis, and education. Colab is a dedicated Jupyter notebook platform that demands no deployment and provides free usage of computer resources, including GPUs.[4]

Properties of Colab

- Python may be used to implement and run programs.
- It is necessary to document the code that supports the mathematical computations.
- Make a new notebook.
- Existing notebooks should be uploaded.
- Use the google link to share the notebooks.
- Data from Google Drive can be imported.
- Google Drive notebooks can be saved to/from.
- GitHub notebooks can be imported and published.[4]

V. IMPLEMENTATION

Here logistic regression is performed on given data set in order to create or to make out a few useful models for predicting the fraud in the credit card transactions.

Data source

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 divides transactions into two categories: "positive class" and "negative class." The data set is substantially skewed, with around 0.172 percent of transactions being fraudulent and the remainder being legitimate. To balance the data set, oversampling was performed, resulting in 60% fraudulent transactions and 40% real transactions. [1]

Logistic Regression:

One of the dividing algorithms, logistic regression is used to predict binary values in a set of independent variables (1/0, Yes / No, True / False). Dummy variables are used to represent binary values or by categories. If the variance of the result is classified, the odds log is used as the variance depending on the regression. By inserting data into the configuration function, it can also anticipate the event occurring. [5]

$O = e^\wedge (I0 + I1*x) / (1 + e^\wedge (I0 + I1*x))$

Here,

- O predicted result
- I0 is a bias or termination term

For a single input value, I1 is a coefficient (x). s
From the training data, the coefficient of I (the actual fixed value) for each column in the input data should be calculated.

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

To get started with a regression, write a simple number of regression fluctuations with the prediction surrounding the link function: To start with a regression, write a simplified regression model with a closed retraction coefficient on the scale:

$A(O) = \beta 0 + \beta(x)$

Were

- A (): link function
- O: variable effect
- x: dependent change

The work is started using two things:
1) Chances of Success (pr) and
2) Chances of Failure (1-pr).

The Pr must meet the following conditions:
a) chances should always be positive (from p> = 0)
b) Chances should always be less than equal to 1 (since pr <= 1).
By applying exponential to the initial conditions and the value remains greater than its equivalent

1. $pr = exp (\beta o + \beta(x)) = e^\wedge (\beta o + \beta(x))$

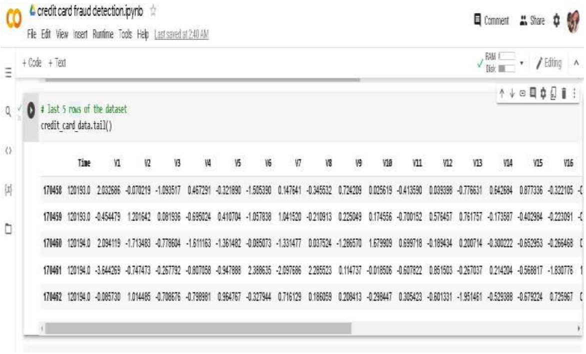
In the second determinant, the same exponential is divided by adding 1 to it so that the value is less than equal

$1\ pr = e^\wedge (\beta o + \beta(x)) / e^\wedge (\beta o + \beta(x)) + 1$

Moving function is used in depreciation when the cost function measures the error, as it shows the answer compared to the actual value.

Firstly, imported all the attributes of the python libraries which is needed for running the python code. after importing the libraries, the dataset is uploaded to the Colab note book.

Figure2: Imported Dataset.

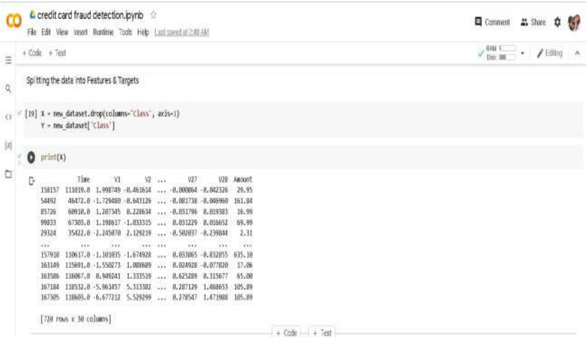

credit card fraud detection.ipynb
File Edit View Insert Runtime Tools Help Last saved at 2:40 AM
Comment Share
+ Code + Text
RAM
Disk
Editing

```
# last 5 rows of the dataset
credit_card_data.tail()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
170458	120193.0	2.032686	-0.070219	-1.093517	0.467291	-0.321890	-1.505390	0.147641	-0.345532	0.724209	0.025619	-0.413590	0.039398	-0.776631	0.642684	0.877336	-0.322105
170459	120193.0	-0.454479	1.201642	0.081936	-0.695024	0.410704	-1.057838	1.041520	-0.210913	0.225049	0.174556	-0.700152	0.576457	0.761757	-0.173587	-0.402984	-0.223091
170460	120194.0	2.094119	-1.713483	-0.778604	-1.611163	-1.361482	-0.085073	-1.331477	0.037524	-1.286570	1.679909	0.699718	-0.189434	0.200714	-0.300222	-0.652953	-0.266468
170461	120194.0	-3.644269	-0.747473	-0.267792	-0.807058	-0.947888	2.388635	-2.097686	2.285523	0.114737	-0.018506	-0.607822	0.851503	-0.267037	0.214204	-0.568817	-1.830776
170462	120194.0	-0.085730	1.014485	-0.708676	-0.798981	0.964767	-0.327944	0.716129	0.186059	0.208413	-0.298447	0.305423	-0.601331	-1.951461	-0.529388	-0.679224	0.725967

Once loading the dataset, the data is checked for null values, although there are none in this dataset. When the values are compared, the dataset has more legitimate transaction values than fraud transactions. Because the data is in an unbalanced state, it must be under-sampled in order to modify it. The dataset is separated into two sets after under-sampling: training data and testing data. [3]

Figure3: Dataset are distributed.


credit card fraud detection.ipynb
File Edit View Insert Runtime Tools Help Last saved at 2:40 AM
Comment Share
+ Code + Text
RAM
Disk
Editing
+ Code + Text

Splitting the data into Features & Targets

```
[19] X = new_dataset.drop(columns='Class', axis=1)
     Y = new_dataset['Class']
```

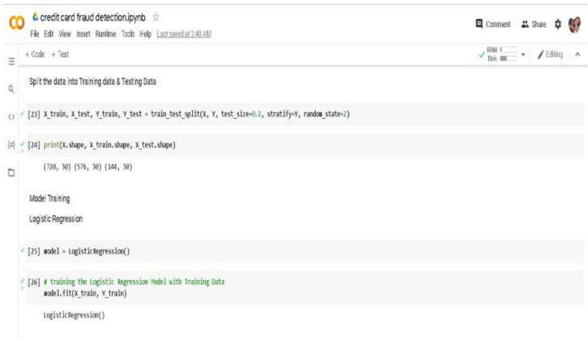
```
print(X)
```

```
            Time        V1        V2  ...       V27       V28  Amount
158157  111019.0  1.998749 -0.461614  ... -0.000864 -0.042326   29.95
54492    46472.0 -1.729480 -0.643126  ... -0.081738 -0.046960  161.84
85726    60910.0  1.207345  0.228634  ... -0.031796  0.019383   16.99
99833    67303.0  1.198617 -1.033315  ...  0.031229  0.016652   69.99
29324    35422.0 -2.245070  2.129219  ... -0.502037 -0.239844    2.31
...          ...       ...       ...  ...       ...       ...     ...
157918  110617.0 -1.101035 -1.674928  ...  0.033865 -0.832855  635.10
163149  115691.0 -1.550273  1.088689  ...  0.024928 -0.077820   17.06
163586  116067.0  0.949241  1.333519  ...  0.625289  0.315677   65.00
167184  118532.0 -5.961457  5.313382  ...  0.287129  1.468653  105.89
167305  118603.0 -6.677212  5.529299  ...  0.278547  1.471988  105.89

[720 rows x 30 columns]
```

Figure4: The logistic regression model is trained using the training data

After training the data with logistic regression finding the accuracy of both the datasets.


credit card fraud detection.ipynb
File Edit View Insert Runtime Tools Help Last saved at 2:40 AM
Comment Share
+ Code + Text
RAM
Disk
Editing

Split the data into Training data & Testing Data

```
[23] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
```

```
[24] print(X.shape, X_train.shape, X_test.shape)

     (720, 30) (576, 30) (144, 30)
```

Model Training

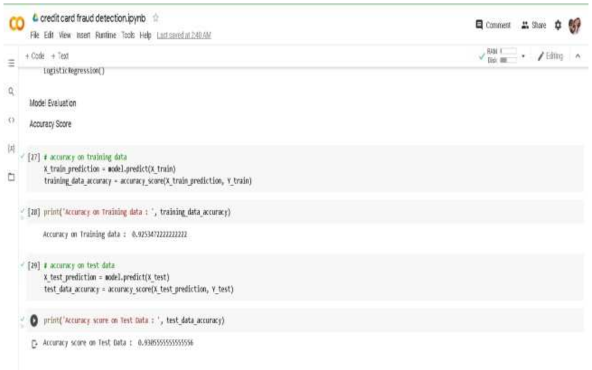
Logistic Regression

```
[25] model = LogisticRegression()
```

```
[26] # training the Logistic Regression Model with Training Data
     model.fit(X_train, Y_train)

     LogisticRegression()
```

Figure5: Finding the accuracy


credit card fraud detection.ipynb
File Edit View Insert Runtime Tools Help Last saved at 2:40 AM
Comment Share
+ Code + Text
RAM
Disk
Editing

```
LogisticRegression()
```

Model Evaluation

Accuracy Score

```
[27] # accuracy on training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
[28] print('Accuracy on Training data : ', training_data_accuracy)

     Accuracy on Training data :  0.9253472222222222
```

```
[29] # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy score on Test Data : ', test_data_accuracy)

Accuracy score on Test Data :  0.9305555555555556
```

VI. CONCLUSION

This article uses machine learning strategies such as decentralization to detect fraud on credit card networks. The performance of the proposed system is tested using sensitivity, specificity, accuracy, and error level. Test and training data are almost identical in accuracy.

As a result, depreciation may be used to detect credit card fraud.

REFERENCES

[1] https://ieeexplore.ieee.org/abstract/document/8123782

[2] https://ieeexplore.ieee.org/abstract/document/5762457

[3] https://en.wikipedia.org/wiki/Credit_card_fraud

[4] https://towardsdatascience.com/machine-learning-for-credit-card-fraud-detection-a-Jupyter-book-for-reproducible-research-8ca5edad7b5d

[5] https://www.javatpoint.com/logistic-regression-in-machine-learning

DOI : 10.5281/zenodo.6372596
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam