IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Credit Card Fraud Detection System

Ankit Singh[1], Arman Patel[2], Mohit Katare[3], Shivam Gondhale[4]
Department of Computer Engineering[1,2,3,4]
Dhole Patil College of Engineering, Pune, Maharashtra, India

Abstract: *Since the growth in online shopping, an increasing number of people are paying online using modes of payment like credit cards to pay their bills, and like every other online interface, Credit cards are also subject to hacking. online businesses and financial companies need to detect these fraudulent activities to save their clients from getting charged for items they didn't purchase. For this purpose, banks and payment companies can implement algorithms to detect fraudulent behavior. The Credit Card Fraud Detection System includes taking into consideration past credit card transactions with the data of the ones that were fraud. This system can be used to identify a fraudulent transaction beforehand. Our objective is to identify all of the fraudulent transactions while also minimizing the incorrect fraud classifications. In this technique we first examine and pre-process the Data set to address unbalanced data, then we train our model using a Logistic regression algorithm to detect fraud.*

Keywords: Credit Card Fraud, Logistic Regression, Fraud Detection, Machine Learning, Cyber Security, etc.

I. INTRODUCTION

Credit card fraud is a term generally used for frauds committed using a payment card, such as a credit card or debit card. The criminal may want to obtain goods or services or to make payment to another account, which is owned by him. Credit card fraud can be authorized, where the genuine customer processes payment to a different account which is controlled by a hacker, or it can be unauthorized, where the owner of the card does not provide consent for the payment to proceed and the transaction is carried out by a third party. In 2018, financial losses due to unauthorized payment across payment cards and remote banking services total £844.8 million in the United Kingdom. Whereas financial companies and banks prevented a loss of £1.66 billion in transaction fraud in 2018. That is the equivalent of £2 in every £3 of attempted fraud being stopped. This is a very major problem that demands the attention of machine learning and data science engineers where the solution to this problem can be automated. This problem is challenging from the perspective of learning, as it is characterized by various factors such as data imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over time.

II. PROBLEM STATEMENT

Old ways of manually detecting fraud are time-consuming and are becoming impossible to carry out as more and more people are doing transactions online. To detect Fraud machine learning algorithms are being used.

III. OBJECTIVE

The objective of this project is to implement machine learning algorithms such as Logistics Regression to detect credit card fraud concerning time and amount of transaction

IV. RELATED WORK

In earlier studies, many methods have been used to detect fraud using unsupervised, supervised algorithms and hybrid algorithms. The variety of frauds and their patterns are changing day by day. It is necessary to have a clear understanding of the processes behind fraud detection. Here we discuss algorithms, machine learning models, and fraud detection systems used in earlier studies. In [1], after implementing an algorithm, Logistic regression gave the highest accuracy and took very low time. In [2], the authors check the performance of decision tree, Random Forest, SVM, and Logistic Regression on credit card fraud data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

In [3], the authors analyze supervised-based classification. When pre-processing the dataset using principal element analysis and normalization, all the classifiers achieved over 95.0% accuracy in comparison to the result reached before pre-processing the dataset.

In [4], Machine learning techniques like logistic regression, decision tree, and random forest were used to detect fraud in credit cards. The accuracy for logistic regression, decision tree, and random forest classifier is 90.0%, 94.3%, and 95.5% respectively.

V. TRANSACTION DATABASE

The dataset contains transactions carried out by credit cards in September 2013 by European credit card holders. This dataset presents transactions that happened 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, data cannot contain original features and more background information. 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.

VI. METHODOLOGY

The procedure that we followed to get the results are understanding problem statement and data by performing statistical analysis and visualization then checking whether the data is balanced or not. The process can be better understood with the following figure:

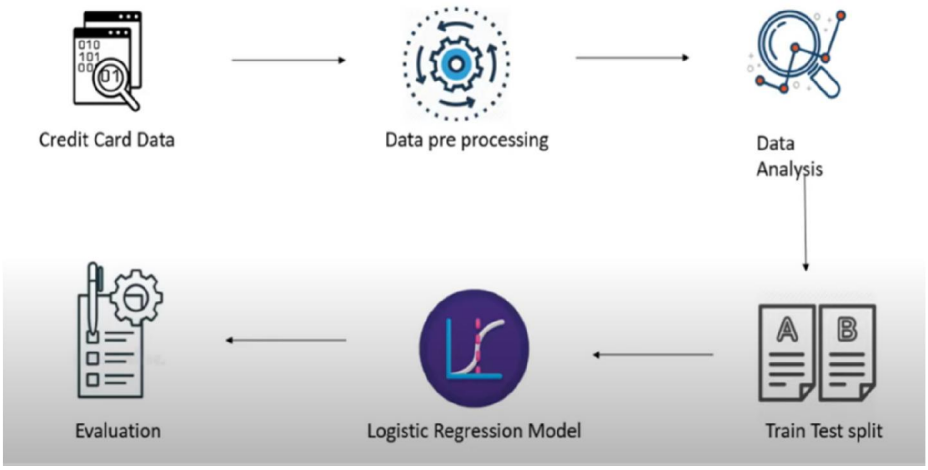

Figure.1: Process Flow

We used the dataset from Kaggle, a data analysis website that provides datasets. Inside this dataset, there are 31 columns out of which 28 are named v1-v28 to protect sensitive data. The columns other than that represent Time, Amount, and Class. Time shows the time gap between the first transaction and the following transaction. The amount is of transacted money. Class 0 represents a genuine transaction and 1 represents a fraudulent one. The first step in the process is to include the dependencies i.e., the libraries and functions needed for programming.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Figure 2: Dependencies

The Numpy is useful for making arrays
We have the dataset as a CSV file. CSV files contain Comma-separated values. It is difficult to process data from a CSV file. For this, we use Pandas.
Sklearn provides various regression, classification, and clustering algorithms.
Here we use logistic regression.

```
[ ]  #distribution of normal transaction and fraudulant transaction
     credit_card_data['Class'].value_counts()

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

This dataset is highly unbalanced

0 --> Normal or Legit transaction

1 --> Fraudulant transaction

Train_test_split lets us split our data into training and test modules.
The accuracy score will help us check the performance of our model.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Figure 3: Showing Unbalanced Data

As our data is highly unbalanced, we first balance our data without changing its nature. It means that the differences between fraudulent data and legit data are still there. It helps our machine learning model to detect fraud data correctly.
After this, we split the data into training and test data. First, we train our model using training data, and then we run it onto the test data to check the accuracy of correctly detecting the fraud. We are using the logistic regression model which is generally used for binary classification.

Spliting the data into features and targets

```
[ ]  x=new_dataset.drop(columns='Class',axis=1)
     y=new_dataset['Class']
```

Figure 4: Splitting the Data

Targets are values 0 and 1 and the features are the rest. Once separated, now we can feed it to our logistic regression model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

```
[ ]  model = LogisticRegression()

[ ]  #Training the logistic regression model with training data
     model.fit(x_train,y_train)

     LogisticRegression()
```

Figure 5: Using Logistic Regression

VI. RESULTS

```
[ ]  #accuracy on training data
     x_train_prediction = model.predict(x_train)
     training_data_accuracy = accuracy_score(x_train_prediction,y_train)

[ ]  print("Accuracy on training data :",training_data_accuracy)

     Accuracy on training data : 0.9364675984752223

[ ]  #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 : 0.9137055837563451
```

Figure 6: Accuracy

The accuracy score for our training data is 93.6. And accuracy score for our test data is 91.3. If our training accuracy is very different from our test accuracy then it means our model has been overfitted or under fitted.

VII. CONCLUSION

Credit card details of users are very sensitive so banks are unwilling to share them Since the entire dataset consists of only two days' transaction records, it's only a fraction of data that can be made available if this project were to be used on a commercial scale. Since this is based on machine learning algorithms, the program will increase its efficiency over time as more data is put into it.

REFERENCES

[1]. "Credit Card Fraud Detection using Machine Learning Methodology" 2019, Heta Naik and Prashasti Kanikar, NMIMS University, Mumbai, India.

[2]. "Credit Card Fraud Detection using Machine Learning Models" 2018, Navanshu Khare and Saad Yunus Sait, Chennai, TamilNadu.

[3]. "Credit Card Fraud Detection using Machine Learning Methodology", 2019, Hamzah Ali Shukur and Sefer Kurnaz Turkey.

[4]. "Machine Learning for Credit Card Fraud Detection System", 2018, Lakshmi and Deepthi Kavita.