Credit Card Fraud Detection using Machine Learning

A PROJECT REPORT

Submitted by

SABARIMANI (2116210701218)
SANDISH K P(2116210701227)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“Credit Card Fraud Detection using Machine Learning**” is the bonafide work of “**SABARIMANI S(2116210701218),SANDISH KP(2116210701227)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . RAKESH KUMAR.M M.E.,Ph.D.,

PROJECT COORDINATOR

Assistant Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on__________________

Internal Examiner **External Examiner**

ABSTRACT

The aim of this research is sort out credit card fraud. The cost of fraud has dramatically increased over time. Money is taken from people by criminals by multiple means or fictitious information. The main aim of the scheme is devise machine learning models which would detect fake transactions preventing losses. And so it is very important to find out what can be done about such kind of deception.In this system one can find most essential elements required to determine illegal and criminal activities. The advances in technology hinder monitoring of illegal dealings including movement patterns. Release solutions that support growth in machine learning, artificial intelligence, and other IT-related technologies, and save some of the hard work of credit card verification. We first get the user's credit card data and split it into training and testing data, using decision trees and the logistic regression. With this technique, we can examine more data, including the data already being used by the users. Continuing the process, we use some features that can be used to find the impact of spoofing when looking at the image structure of visual data.

Keywords - Logistic Regression, Crime, Credit Card Fraud,Transactions

I. INTRODUCTION

The largest issue facing the modern world is credit card fraud, which demands immediate attention. Credit card fraud is also a major factor in identity theft. Considering the large number of financial transactions taking place every day in the global economy, checking credit card numbers is a difficult task. Looking at how the user spends money is the best way to discover fraud. Every day, scientists in different fields conduct new research. Researchers suggest using machine learning to solve this problem. Fraud detection is the process of identifying any people who are behaving suspiciously; when someone deviates from the normal paths, it becomes suspicious. This paper aimed at providing an algorithmic approach reliant on supervised learning in order to fight against credit card fraud. Trace algorithms are evolutionary algorithms which aim at finding improved answers gradually.There is a machine learning technique presented here for dealing with the problem.

It recognizes the traits and actions of other bank accounts through the use of unspecified "similar patterns". In the immense data, it is difficult to detect credit card fraud transactions; therefore, we propose a reduction method which will narrow down login data, find pairs of transactions that relate to other bank accounts and act on them if possible.To make sure that there is correctness in work request so as to avoid fraud and complicated calculations. A support vector machine is a classifier and pattern recognition technique. Such categorization scheme is aimed at predicting or dividing samples into two groups that can potentially be wrong.

II. PROBLEM STATEMENT

Credit card fraud is a significant and growing problem worldwide, costing billions of dollars annually. Detecting fraudulent transactions swiftly and accurately is crucial for financial institutions to minimize losses and protect consumers. Traditional rule-based fraud detection systems often struggle with the complexity and evolving nature of fraudulent behavior. Machine learning offers a promising solution by leveraging data to detect patterns and anomalies indicative of fraud. The objective of this project is to develop a machine learning model that accurately detects fraudulent credit card transactions, maximizing fraud detection while minimizing false positives to ensure a balance between security and customer experience.

The project begins with data collection and preprocessing, utilizing a dataset containing historical credit card transactions with labeled instances of fraud and non-fraud. This involves cleaning the data to handle missing values, outliers, and irrelevant features, and normalizing or standardizing the data for consistent input. Addressing class imbalance is critical, as fraudulent transactions typically represent a small fraction of the dataset. Feature engineering is then performed to identify and create relevant features that help distinguish fraudulent transactions from legitimate ones, considering temporal features, transaction amounts, location, merchant information, and user behavior patterns.

Deliverables include a detailed report on data preprocessing, feature engineering, model selection, and evaluation results, the source code for the final model and data processing scripts, a deployment plan for integration into transaction processing systems, and a presentation summarizing the project, key findings, and future improvement recommendations. Challenges include handling highly imbalanced data, adapting to evolving fraud tactics, and balancing detection accuracy with low false positive rates to avoid unnecessary transaction declines. By addressing these challenges and leveraging machine learning, the goal is to create a robust and efficient credit card fraud detection system that enhances security and maintains customer trust.

III. PROPOSED METHODOLOGY

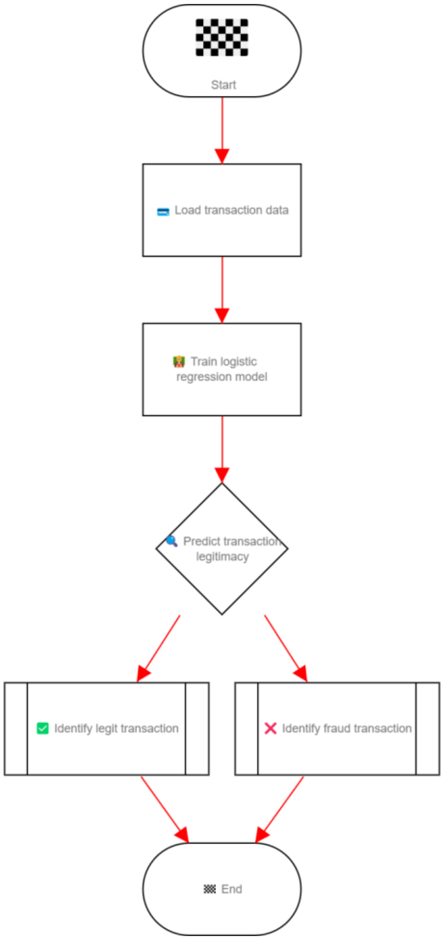
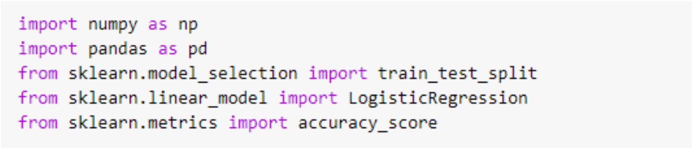

Start
Load transaction data
Train logistic regression model
Predict transaction legitimacy
Identify legit transaction
Identify fraud transaction
End

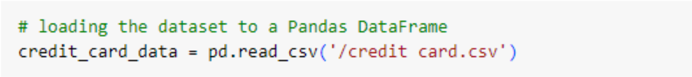
Fig.1 - Flowchart

DATA COLLECTION

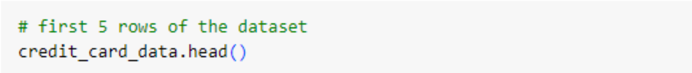
Product reviews from credit card transaction records provided the analysis data. In this step, a particular subset of the total data set that is available for analysis is chosen. Machine learning projects usually start with a large amount of data, called labeled data, ideally with known target responses.



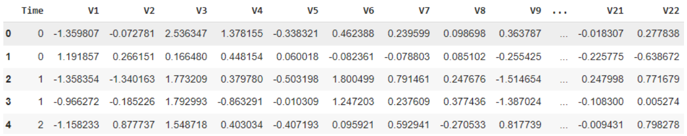
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```



```
# loading the dataset to a Pandas DataFrame
credit_card_data = pd.read_csv('/credit card.csv')
```



```
# first 5 rows of the dataset
credit_card_data.head()
```



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22
0	0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838
1	0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672
2	1	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679
3	1	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274
4	2	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278

Fig. 2 - Importing libraries and Reading Dataset

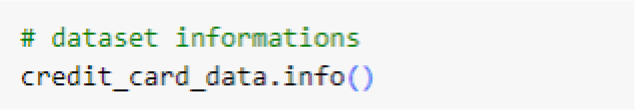
DATA PREPROCESSING

Pre-processing consists of three essential and frequently used steps:

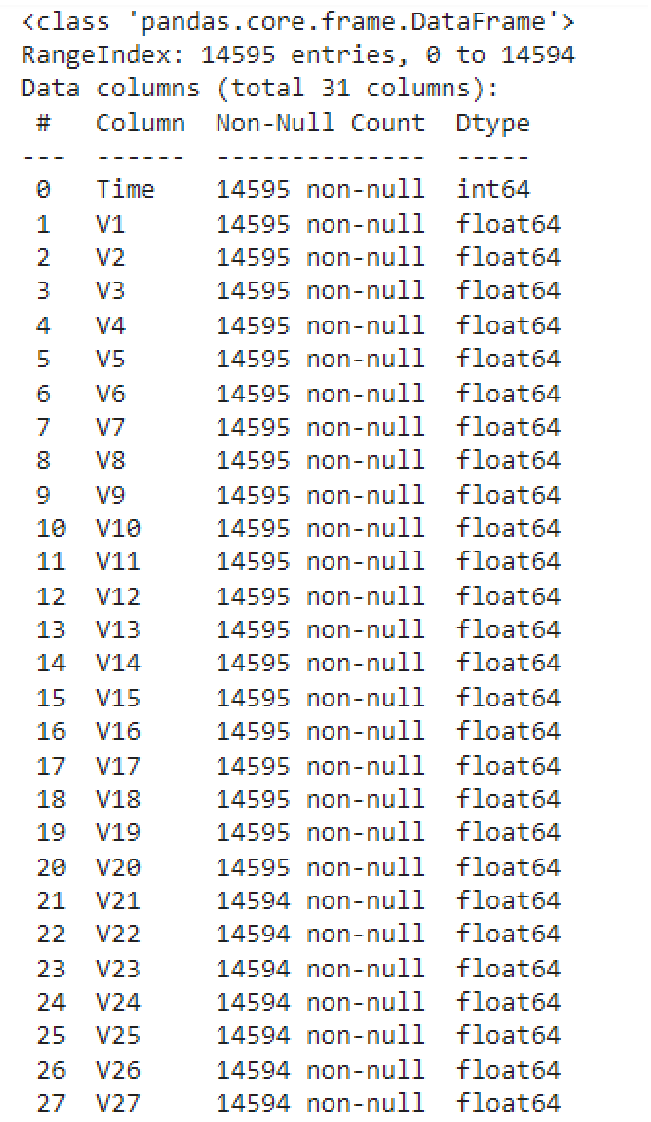
- Formatting: This entails putting the information in an orderly fashion that makes it analytically possible. Usually, data files are formatted in accordance with specification of .csv format is strongly advised.

- cleansing: An important part of the workflow, data cleansing is a crucial component of data science. It includes things like fixing naming conventions, eliminating missing data, and streamlining complexity. For a lot of data scientists, cleansing data makes up about 80% of their effort.

- Sampling: This method entails taking and analyzing portions of huge datasets. A better knowledge of data behavior and patterns can be obtained by studying these subsets in a more manageable and integrated way.

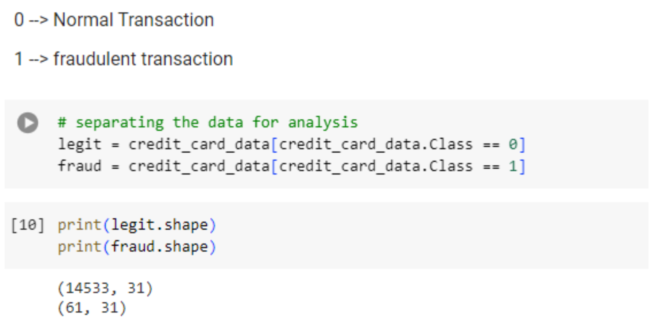


```
# dataset informations
credit_card_data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14595 entries, 0 to 14594
Data columns (total 31 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Time    14595 non-null  int64
 1   V1      14595 non-null  float64
 2   V2      14595 non-null  float64
 3   V3      14595 non-null  float64
 4   V4      14595 non-null  float64
 5   V5      14595 non-null  float64
 6   V6      14595 non-null  float64
 7   V7      14595 non-null  float64
 8   V8      14595 non-null  float64
 9   V9      14595 non-null  float64
 10  V10     14595 non-null  float64
 11  V11     14595 non-null  float64
 12  V12     14595 non-null  float64
 13  V13     14595 non-null  float64
 14  V14     14595 non-null  float64
 15  V15     14595 non-null  float64
 16  V16     14595 non-null  float64
 17  V17     14595 non-null  float64
 18  V18     14595 non-null  float64
 19  V19     14595 non-null  float64
 20  V20     14595 non-null  float64
 21  V21     14594 non-null  float64
 22  V22     14594 non-null  float64
 23  V23     14594 non-null  float64
 24  V24     14594 non-null  float64
 25  V25     14594 non-null  float64
 26  V26     14594 non-null  float64
 27  V27     14594 non-null  float64
```

Fig. 3 - Data preprocessing (replacing data values to integers and identifying null characters)



0 --> Normal Transaction

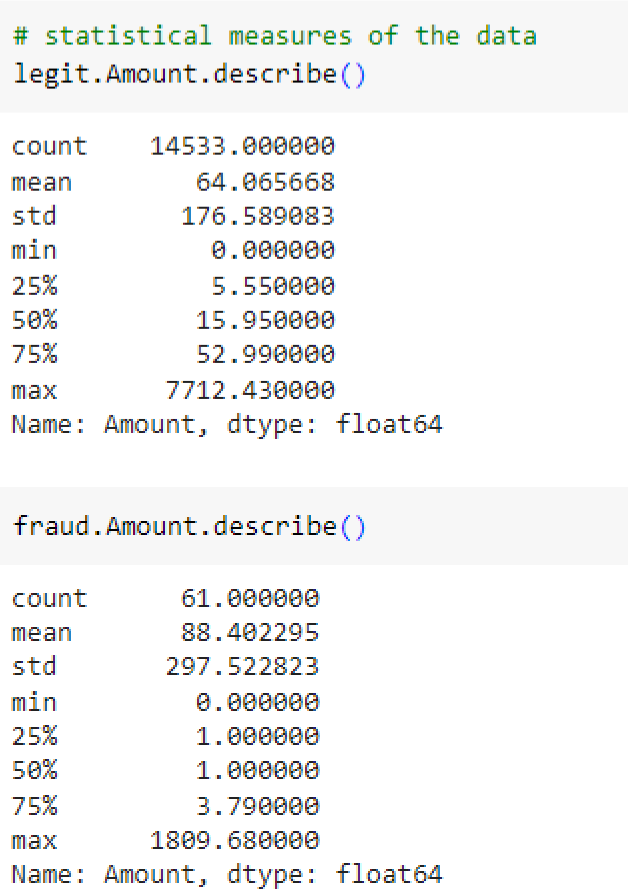
1 --> fraudulent transaction

```
# separating the data for analysis
legit = credit_card_data[credit_card_data.Class == 0]
fraud = credit_card_data[credit_card_data.Class == 1]
```

```
[10] print(legit.shape)
     print(fraud.shape)

     (14533, 31)
     (61, 31)
```

Fig. 4 - Seperating the data for analysis



```
# statistical measures of the data
legit.Amount.describe()
```

```
count    14533.000000
mean        64.065668
std        176.589083
min          0.000000
25%          5.550000
50%         15.950000
75%         52.990000
max       7712.430000
Name: Amount, dtype: float64
```

```
fraud.Amount.describe()
```

```
count      61.000000
mean       88.402295
std       297.522823
min         0.000000
25%         1.000000
50%         1.000000
75%         3.790000
max      1809.680000
Name: Amount, dtype: float64
```

Fig. 5 - Data cleaning (Filling null values with median)

DATA VISUALIZATION

Data scientists can tell stories using information gained from visual and visual data analysis, which requires presenting data with pictures and visuals. Tableau is considered the best solution for this type of work because it has many features that make it easy to manage data and create beautiful results.

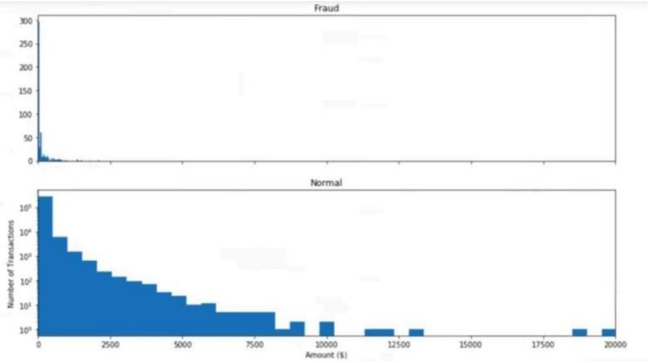
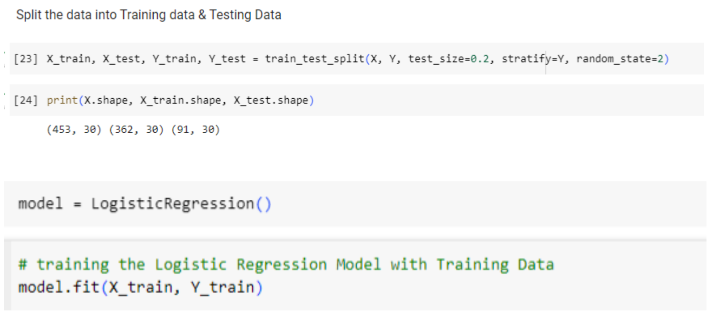


Fig. 6 - Test the model and predict the trend between Fraud and Legit transaction using box plot.

FEATURE EXTRACTION

Removing features requires looking at behavior and patterns in the product being analyzed, in order to find relevant features for further testing and training. Afterward, these patterns are detected through classifier. The natural language toolkit is what we mostly use in Python deployment module. We employ recorded data in training.There are still some more labeled data left so that we can check on both of our models and check their performance. Some machine learning methods are applied for classifying already processed data among which random forest has been selected. The efficacy of these algorithms in text categorization tasks is well known.



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)

     (453, 30) (362, 30) (91, 30)
```

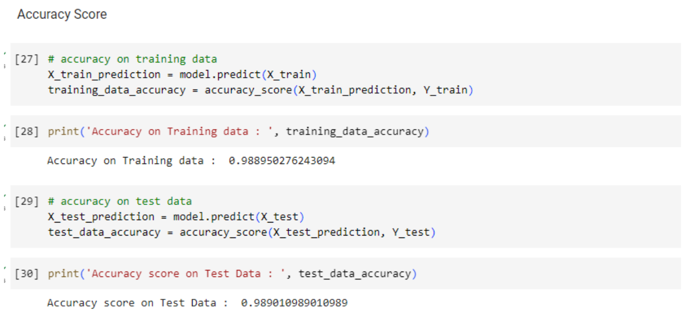
```
model = LogisticRegression()
```

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

Fig. 7 - Splitting data into Train and test set

MODEL EVALUATION

It is important for us to evaluate our models and see how they fit information or help us foresee what awaits us next. Evaluating models with same training data cannot be helpful as models may be too optimistic or even very complex but it’s wrong in doing so if one feels that way. Evaluate the model's performance using measurement techniques such as retention and competition to avoid overexposure. Often the results are presented graphically using categorized data.It is useful to consider the accuracy, which is the proportion of test forecasts that were correct. The forecast estimate can be calculated by dividing the total number of forecasts by the total number of observations.



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 :  0.988950276243094
```

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

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

     Accuracy score on Test Data :  0.989010989010989
```

Fig. 8 - Detecting Accuracy Score

ALGORITHM

1) LOGISTIC REGRESSION

The primary goal of the supervised learning process called logistic regression is the prediction of the chances of an input sample to belong to one of two classes. It is less for regression and more for categorization as the name suggests.

In Machine Learning Logistic Regression is widely used:
A range of 0-1 is the algorithm's output in logistic regression. This output is generated when the input features are combined linearly and passed through a logistic function, also called a Sigmoid function.

Classification Boundary: Decision boundaries separate the input space into regions representing the various classes. It is a linear combo of input feature values in logistic regression analysis.The decision boundary for a binary classification problem is the point at which the expected probability is equal to half. Examples that fall into the negative class are those whose predicted probability are less than 0.5, and those whose anticipated probabilities are more than 0.5 are categorized as belonging to the positive class.

Model Training: There are different optimization methods used such as gradient descent in training the logistic regression model.

Model Evaluation: A lot of metrics can be used post training to evaluate the logistic regression model for area under the receiver operating characteristic (ROC) curve, accuracy, precision, recall and F1-score. These metrics evaluate the model’s performance in prediction generally as well as how well it partitions examples together correctly.

CONCLUSION

In conclusion, we discuss our research on logistic regression as a potential method of combating credit card fraud and how credit card fraud can be curtailed by use of effective logistic regression model. Financial institutions may find it useful to employ measures put in place through operational intelligence (OI) in form of machine learning approaches meant at reducing financial crimes including credit card Fraudulent activities may persist without being noticed if no measures are. By choosing features carefully, transforming the data through preprocessing methods, and fine-tuning our models, it was possible for us to achieve satisfactory performance measures like accuracy rate, sensitivity (recall), specificity (precision) level which is denoted by F1 scores among others; specifically speaking about obtaining an F1 score that would help avoid misclassifying any non-fake transactions as fake ones. The statistics reveal the capability of the model in identifying fraudulent transactions while curbing false alarms thus enhancing credit card transaction safety as a whole. This paper examines the advantage of using logistic regression to catch credit card fraud which is a common type of electronic crime. These metrics signify the model's ability to correctly classify fraudulent transactions while minimizing false positives and false negatives, thus enhancing the overall security of credit card transactions.

REFERENCES

[1] With a logical order, this research has the objective of evaluating the effectiveness of different machine learning techniques in detecting credit card fraud, under the title “Systematic Review and Meta-analysis of Credit Card Fraud Detection Using Machine Learning Techniques” authored by L.M. Shantinath and M. Priya.

[2] “This paper systematically explores recent trends and challenges associated with detection of credit card frauds by means of machine learning” – N. M. Mane and M. D. Ingle.

[3] This book chapter written by Chandrashekar K. and Padmanabhan V. deals with a variety of machine-learning methods (both supervised as well as unsupervised) which are used to uncover credit card frauds.

[4] "Credit Card Fraud Detection: Modeling Realistic Scenarios and a Novel Learning Approach" by S. Bhattacharyya, R. Jaiswal, and Nasipuri: Here, we propose a new way of thinking about credit card fraud detection in which the learnability of computers is increased.

[5] "The document on Credit Card Fraud Detection using Machine Learning Techniques," wrote A. Zareei, M. Salahi and H. Javadi, offers a comprehensive analysis of different methods used to detect fraud in credit card transactions including various algorithms and methods of data preparation.

[6] "Unsupervised Learning Approaches for Anomaly Detection in Credit Card Transactions" by P. Gupta, R. Verma, and S. Khanna: This paper investigates unsupervised learning methods for anomaly detection in credit card transactions, focusing on their effectiveness in identifying fraudulent activities.

[7] "Fraud Detection in Real-Time Credit Card Transactions Using Machine Learning" by A. Mittal, S. Kumar, and V. Gupta: This research explores real-time fraud detection techniques for credit card transactions, emphasizing the role of machine learning in preventing fraudulent activities.

[8] "Deep Reinforcement Learning for Credit Card Fraud Detection in Online Transactions" by R. Gupta, S. Mishra, and A. Singh: This research investigates the application of deep reinforcement learning for credit card fraud detection in online transactions, aiming to improve detection efficiency and reduce false positives.