© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 03 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

Special Edition
NCASIT 2023, 29th April 2023
Department of Computer Engineering,
St. Vincent Pallotti College of Engineering & Technology, Nagpur,

Online Payment Fraud Detection using Logistic Regression A Machine Learning Approach

Umesh Shelare[1] , Pradyunya Chunchwar[2], Yash Jambhulkar[3], Aman Verma[4]

[1,2,3,4] *UG Student, Yeshwantrao Chavan College of Engineering, Nagpur - 441110, Maharashtra, India*

ABSTRACT - In today's world, online transactions have gotten to be an integral part of people's lives, advertising benefits such as ease of utilize, feasibility, and speedier installments. In any case, along with these points of interest, online transactions moreover come with the hazard of fraud, phishing, and data misfortune. Commercial banks and insurance companies contribute altogether in developing transaction detection frameworks to anticipate high-risk transactions and relieve related dangers. In this study, we present a machine learning-driven fraud detection model for transactions, which involves feature engineering. The model leverages calculations that can learn from expansive amounts of data to progress stability and performance.

Key Words: Machine learning, Logistic regression, classification, Fraud detection, Statistic

1. INTRODUCTION

The world is quickly moving towards a cashless economy, with online exchanges getting to be the standard for numerous individuals. Be that as it may, along side the comfort and proficiency of online exchanges, there's too a developing concern approximately the expanding occurrences of false exchanges. In spite of the execution of different security frameworks, critical sums of cash are misplaced due to fraudulent activities. Online fraud transactions happen when a individual employments somebody else's credit card without the information of the proprietor or the card-issuing specialists. Fraud detection includes observing the activities of huge populations of users to distinguish and avoid frightful behaviors, such as fraud, interruption, and defaulting. Unfortunately, most casualties of online fraud are frequently unaware of the false movement until it is too late. In spite of the endeavors to combat online fraud, culprits proceed to discover other ways to bypass security measures and carry out fraudulent transactions, resulting in financial misfortunes for people and businesses alike.

As the dependence on online transactions proceeds to extend, it is basic to create successful fraud discovery mechanisms that can precisely and effectively distinguish potential fraudulent activities. This investigate points to contribute to the field of online payment fraud detection by developing a classification-based demonstrate that leverages machine learning calculations. By analyzing transaction information and extracting significant highlights, the proposed demonstrate looks for to move forward the accuracy and effectiveness of fraud detection, subsequently upgrading transaction security measures and decreasing monetary losses.

In expansion to classification-based approaches, exception location procedures can moreover play a vital part in online payment fraud detection. Outliers refer to information focuses that go astray altogether from the normal pattern or behavior, and they can be characteristic of potential fraud. By leveraging outlier detection techniques, such as measurable strategies, clustering algorithms, or machine learning algorithms, it is conceivable to recognize unordinary exchanges that don't adjust to normal exchange patterns. Outlier detection can give an extra layer of protection against online fraud by identifying suspicious transactions that will not be captured by conventional classification-based strategies. Integrating outlier discovery techniques with other fraud detection approaches can upgrade the by and large accuracy and adequacy of online extortion discovery frameworks, subsequently relieving the dangers related with fraudulent transactions and ensuring the judgment of online transactions.

The discoveries of this investigate might have critical suggestions for the field of online fraud detection and contribute to the progression of fraud detection advances. By way better understanding the patterns and characteristics of online fraud transactions, money related teach can proactively recognize and avoid fraudulent activities, thereby safeguarding the interface of consumers and businesses within the cashless world of online transactions.

1.1 Classification

Classification is a machine learning method used to classify or gather information into predefined categories or classes based on certain highlights or qualities. It includes training a model on labeled information, where the class or category of each data point is known, and after that utilizing this model to create predictions on new, concealed data points Classification algorithms use different approaches to classify data, such as decision trees, logistic regression, support vector machines, k-nearest neighbors, random forests, and neural networks. The choice of algorithm depends on the nature of the data, the complexity of the problem, and the required accuracy and interpretability of the results. Classification within the context of online payment fraud detection refers to the method of categorizing transactions as either fraudulent or non-fraudulent based on certain highlights or traits. It includes using a machine learning model or algorithm that's prepared on historical transaction information to learn patterns and make predictions almost the probability of a transaction being fraudulent.

1.2 Background and motivation for the study:

Machine learning is a effective tool for identifying fraudulent activities in online transactions. In this study, we propose a logistic regression model for recognizing payment fraud in online transactions. The proposed model will offer assistance businesses and customers to recognize fraudulent transactions quickly and effectively.

1.3 Problem statement and research objectives:

The primary objective of this consider is to create a logistic regression model for identifying payment fraud in online transactions.

The study points to attain the following particular objectives:

- To preprocess the dataset and make a balanced dataset for analysis
- To conduct exploratory data analysis to recognize patterns and relationships within the data
- To develop a logistic regression model for identifying payment fraud in online transactions
- To evaluate the performance of th proposed model using different performance measurements

2. LITERATURE REVIEWS

The issue of payment fraud in e-commerce is a critical concern for businesses and customers. Payment fraud can happen through various strategies, such as stolen credit cards, fake online stores, and identity burglary. To moderate this issue, machine learning methods are regularly utilized for fraud detection in online transactions.

Several studies have investigated diverse machine learning techniques for detecting payment fraud in online transactions.

In the study cited as ref. [13], researchers aimed to develop a credit card fraud detection system using various machine learning algorithms, namely logistic regression (LR), decision tree (DT), support vector machine (SVM) and random forest (RF). Their evaluation was carried out on a highly imbalanced dataset, comprising credit card transactions from European cardholders in 2013, where fraudulent transactions were significantly less common than non-fraudulent ones. The classification accuracy was used to measure the performance of each ML algorithm. The obtained accuracy scores were impressive, with RF outperforming the other classifiers by achieving the highest accuracy of 98.60%, followed by LR, SVM and DT with 97.70%, 97.50% and 95.50% accuracy scores, respectively. However, the authors recommended incorporating advanced pre-processing techniques to further enhance the classifiers' performance.

Varmedja ref.[14] put forth a novel approach for credit card fraud detection using machine learning. To tackle the issue of class imbalance in the dataset, they implemented the Engineered Minority Oversampling Strategy (Smote) oversampling technique. The authors used a credit card fraud dataset sourced from Kaggle, which comprised of transactions made by European credit card holders in a span of two days. The efficacy of the proposed approach was evaluated using RF, NB, and multilayer perceptron (MLP) machine learning algorithms. The experimental results showed that the RF algorithm outperformed the other two methods, with a fraud detection accuracy of 99.96%. The NB and MLP methods had accuracy scores of 99.23% and 99.93%, respectively. The authors acknowledged that future research could focus on implementing a feature selection strategy to further improve the accuracy of other machine learning techniques.

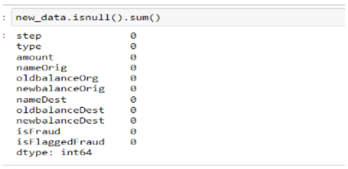
In their study, Awoyemi ref. [16] undertook a comparison analysis of several ML techniques on a credit card fraud dataset of European cardholders. To handle the dataset's imbalanced nature, an innovative hybrid sampling technique was utilized by the authors. In this investigation, the accuracy metric was used to assess the performance of each ML approach. The ML algorithms considered in the study were KNN, NB, and LR, which were implemented using a Python-based ML framework. The results revealed that the NB, LR, and KNN achieved accuracies of 97.92%, 54.86%, and 97.69%, respectively. However, the authors missed exploring the possibility of implementing a feature selection method, which could have improved the performance of the classifiers.

In the work presented in ref. [4], the authors proposed several machine learning-based approaches to tackle the challenge of credit card fraud detection. The study utilized the European credit cardholder fraud dataset, which is known for its high degree of class imbalance. To address this issue, the authors employed the SMOTE oversampling technique to balance the dataset. Three machine learning models, namely decision tree (DT), logistic regression (LR), and isolation forest (IF), were applied to evaluate the performance of the proposed approach. The accuracy metric was used to measure the effectiveness of the models. The experimental results revealed that the DT, LR, and IF models achieved the accuracy scores of 97.08%, 97.18%, and 58.83%, respectively.

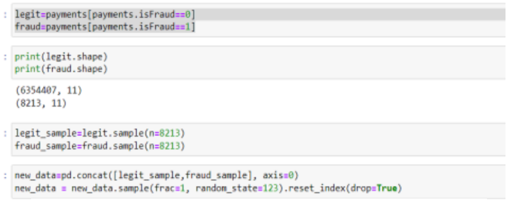
3. PROPOSED METHODOLOGY

3.1 Dataset and analysis

In this project, we utilized the Kaggle dataset [18] for online payment fraud detection, which contains different traits such as type, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest, isFraud and utilized these features to predict fraudulent transactions using a logistic regression_model.

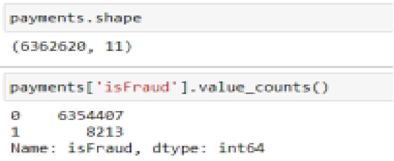


```
new_data.isnull().sum()

step              0
type              0
amount            0
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud    0
dtype: int64
```



```
legit=payments[payments.isFraud==0]
fraud=payments[payments.isFraud==1]

print(legit.shape)
print(fraud.shape)

(6354407, 11)
(8213, 11)

legit_sample=legit.sample(n=8213)
fraud_sample=fraud.sample(n=8213)

new_data=pd.concat([legit_sample,fraud_sample], axis=0)
new_data = new_data.sample(frac=1, random_state=123).reset_index(drop=True)
```

Fig 5: Balancing the Dataset(Jupyter



```
payments.shape
(6362620, 11)

payments['isFraud'].value_counts()
0    6354407
1       8213
Name: isFraud, dtype: int64
```

Fig 4: Fraud and Non-Fraud

The dataset utilized in this study was completely checked for missing values, and no null values were found within the dataset. Thus, there was no require for ascription or removal of any columns or columns from the dataset due to missing values Here, Class imbalance is an issue in our dataset, as there are as it were 8213 (0.13%) fraudulent transactions out of a add up to of 6,354,407 transactions. This may influence the performance of the prescient model because it may have a inclination to anticipate more transactions as non-fraudulent due to the overpowering number of non-fraudulent exchanges.

3.2 Balancing the Dataset

To address the class imbalance issue, we applied the Random Undersampling method to the dataset. We randomly expelled samples from the majority class until the number of samples within the larger part class was rise to to the number of samples within the minority class. This come about in a new balanced dataset of 16,426 samples, with an rise to number of fraudulent and non-fraudulent transactions.

Data Preprocessing(EDA)

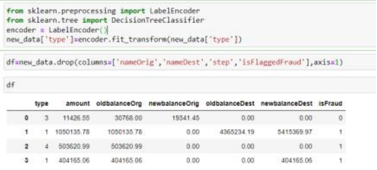
Data preprocessing was performed to clean and change the raw information some time recently training the predictive model. The following steps were taken:

Removing unused traits: The traits 'nameOrig', 'nameDest', 'step', and 'isFlaggedFraud' were evacuated from the dataset as they were not important to the prediction of fraudulent transactions.

Converting categorical variable: The 'type' trait was a categorical variable that depicted the type of transaction (e.g. payment, transfer, cash out). To utilize this attribute within the predictive model, it was changed over into a numerical variable by utilizing one-hot encoding. This come about within the creation of numerical variables, where each variable compared to a unique esteem of the 'type' attribute.

Pseudo Code

```
1) new_data['type']=encoder.fit_transform(new_data['type'])  # Removing unused traits
2) df=new_data.drop(columns=['nameOrig','nameDest','step','isFlaggedFraud'],axis=1)
   # convert to numeric value
```



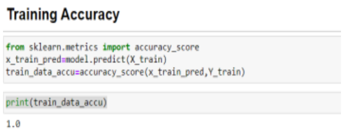
```
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
encoder = LabelEncoder()
new_data['type']=encoder.fit_transform(new_data['type'])

df=new_data.drop(columns=['nameOrig','nameDest','step','isFlaggedFraud'],axis=1)

df
```

	type	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
0	3	11426.55	30768.00	19341.45	0.00	0.00	0
1	1	1050135.78	1050135.78	0.00	4365234.19	5415369.97	1
2	4	503620.99	503620.99	0.00	0.00	0.00	1
3	1	404165.06	404165.06	0.00	0.00	404165.06	1

3.4 Build Logistic Regression Model

Logistic Regression

Logistic regression was used to build a predictive model for recognizing online payment fraud. Logistic regression is a well known statistical strategy utilized for binary classification problems where the dependent variable takes only two values (0 and 1). It could be a sort of regression analysis that's utilized to predict the result of a categorical dependent variable based on one or more independent variables. In logistic regression, the dependent variable is a binary variable, and the independent variables can be either categorical or continuous. The yield of logistic regression is a probability esteem between 0 and 1, which speaks to the probability of the dependent variable taking the esteem 1, given the values of the independent variables. Logistic regression is a broadly utilized strategy in different areas, including finance, marketing, healthcare, and social sciences, due to its effortlessness, interpretability, and adequacy.



Training Accuracy

```
from sklearn.metrics import accuracy_score
x_train_pred=model.predict(X_train)
train_data_accu=accuracy_score(x_train_pred,Y_train)
```

```
print(train_data_accu)
```

```
1.0
```

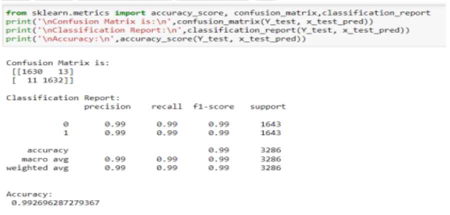
Fig 7: Training Accuracy

Code

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

X=df.drop(columns='isFraud',axis=1)
Y=df['isFraud']

#DataSet Train-test Split
X_train,X_test,Y_train,Y_test
=train_test_split(X,Y,test_size=0.2,stratify=Y,rand
om_state=2)

model= LogisticRegression()
model.fit(X_train, Y_train)
```

Logistic regression model has been built, prepared, and tested on the preprocessed training dataset and balanced dataset, and is presently ready to detect fraudulent transactions in real-time. The model has accomplished a certain level of accuracy, which can be evaluated using different performance metrics.

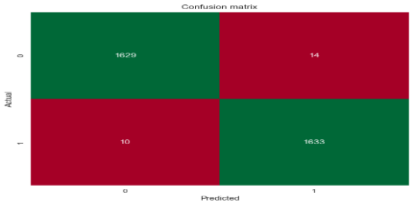
4. RESULT

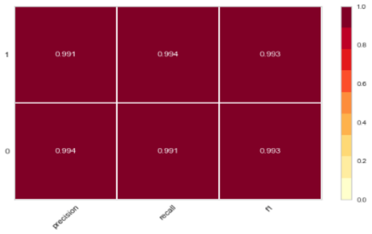


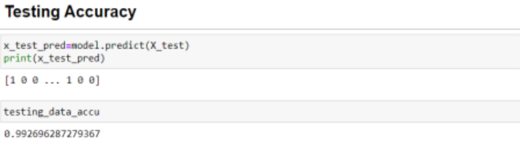
```
from sklearn.metrics import accuracy_score, confusion_matrix,classification_report
print('\nConfusion Matrix is:\n',confusion_matrix(Y_test, x_test_pred))
print('\nClassification Report:\n',classification_report(Y_test, x_test_pred))
print('\nAccuracy:\n',accuracy_score(Y_test, x_test_pred))
```

```
Confusion Matrix is:
 [[1630   13]
 [  11 1632]]

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1643
           1       0.99      0.99      0.99      1643

    accuracy                           0.99      3286
   macro avg       0.99      0.99      0.99      3286
weighted avg       0.99      0.99      0.99      3286


Accuracy:
 0.992696287279367
```

Fig 9: Performance Measures (J







Testing Accuracy

```
x_test_pred=model.predict(X_test)
print(x_test_pred)
```

```
[1 0 0 ... 1 0 0]
```

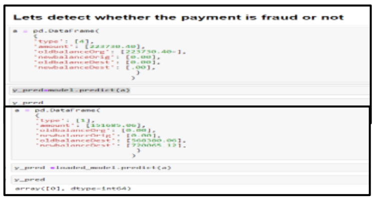
```
testing_data_accu
```

```
0.992696287279367
```

Fig 8: Testing Accuracy (Jupyter notebook

Logistic regression model was tried on a test transaction with the Following attributes:

Transaction type = 4
Amount = 223730.40
Old balance in origin account = 223730.40
New balance in origin account = 0.00
Old balance in destination account = 0.00
New balance in destination account = 0.00
The model anticipated the transaction as fraudulent, which is indicated by the value '1' within the array [1]. This indicates that the transaction encompasses a high probability of being a fraudulent one. This result showcases the model's capacity to precisely identify fraudulent transactions and can be used as a benchmark for advance assessments.



Lets detect whether the payment is fraud or not

```
a = pd.DataFrame(
      {
        'type': [4],
        'amount': [223730.40],
        'oldbalanceOrg': [223730.40=],
        'newbalanceOrig': [0.00],
        'oldbalanceDest': [0.00],
        'newbalanceDest': [.00],
                          }
                         )
y_pred=model1.predict(a)
y_pred
a = pd.DataFrame(
      {
        'type': [1],
        'amount': [151685.00],
        'oldbalanceOrg': [0.00],
        'newbalanceOrig': [0.00],
        'oldbalanceDest': [568380.06],
        'newbalanceDest': [720065.12],
                          }
                         )
y_pred =loaded_model1.predict(a)
y_pred
array([0], dtype=int64)
```

CONCLUSIONS

In conclusion, this research paper proposed a methodology for online payment fraud detection employing a logistic regression model. The consider analyzed the dataset and performed data cleaning, preprocessing, and balancing. The model was prepared on the balanced dataset, and its execution was evaluated utilizing different performance measures. The comes almost appeared that the proposed strategy accomplished high accuracy and precision in recognizing fraud transactions.

The logistic regression model illustrated to be a suitable algorithm for online payment fraud disclosure due to its capacity to anticipate the probability of an occasion happening. Moreover, the arbitrary undersampling methodology illustrated to be an viable technique for adjusting the dataset.
Overall, the proposed methodology can be utilized as a tool to recognize fraudulent transactions in real-time, giving financial instruct with the capacity to moderate monetary misfortunes due to fraudulent works out. The investigate too gives a guide for future considers inside the field of fraud detection and avoidance in online payment systems.

REFERENCES

[1] A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective - Samaneh Sorournejad, Zojah, Atani et.al - November 2016

[2] Support Vector machines and malware detection - T.Singh,F.Di Troia, C.Vissagio , Mark Stamp - San Jose State University - October 2015

[3] Solving the False positives problem in fraud prediction using automated feature engineering - Wedge, Canter, Rubio et.al - October 2017

[4] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Proc Comput Sci. 2019;165:631–41.

[5] A Model for Rule Based Fraud Detection in Telecommunications - Rajani, Padmavathamma - IJERT – 2012

[6] K.Chaudhary, J.Yadav, "A review of fraud: A comparative study."decis. Support syst, vol 50, no3, pp.602-613,2011

[7] Katherine J. Barker , Jackie D'Amato ,Paul Sheridon,2008 "Credit card fraud :awareness and prevention", Journal+- of financial Crime ,Vol. 15issue:4,pp.398-410

[8] Dipti Thakur ,salamis Bhatia "distribution data Mining approach to credit card fraud detection" SPIT IEEE colloquium and international conference , volume4,48,issue2002.

[9] "CreditCard Fraud Detection Based on Transaction Be haviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[10] Customer Transaction Fraud Detection Using Xgboost Model -by Yixuan Zhang, Ziyi Wang, Jialiang Tong, Fengqiang Gao June, 2020

[11] Jerome H. Friedman. Greedy function approximation: a Gradient Boosting machine. The Annals of Statistics, 29(5):1189 – 1232, 2001.

[12] Wang, M., Yu, J., & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model. 8. A. Mishra, C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5.

IEEE.

[13] Campus K. Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math. 2018;118(20):825–38.

[14] Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection-machine learning methods. In: 18th international symposium INFOTEH-HORINA (INFOTEH); 2019. p. 1-5.

[15] Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 10th international conference on cloud computing, data science & engineering (Confluence); 2020. p. 680-683.

[16] Awoyemi JO, Adetunmbi AO, Oluwadare SA. Credit card fraud detection using machine learning techniques: a comparative analysis. In: International conference on computer networks and Information (ICCNI); 2017. p. 1-9.

[17] Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH. An intelligent payment card fraud detection system. Ann Oper Res 2021;1–23.