Credit Card Fraud Detection Model

P Srinivas Reddy, Murali Krishna Chigulla,
Sai Bhavana Kancharla, Sai Rithin Reddy Kairthabad and
Nuthan Yadidya Bashpa

April 25, 2021

CREDIT CARD FRAUD DETECTION MODEL

P. Srinivas Reddy [1], CH. Murali Krishna [2], K. Sai Bhavana [3], K. Sai Rithin [4], B. Nuthan[5]

1,2,3,4,5 Department of Computer Science & Engineering, MLR Institute of Technology
Dundigal, Hyderabad, 500043

---******---

***Abstract*:**

In recent years, credit card fraud has become a major complication for banks as it became tough to detect fraud in the credit card system. To overcome this obstacle Machine learning has an important role in detecting the credit card fraud in the transactions

[1] To predict the various bank transactions various machine learning algorithms are used. This paper examines the performance Logistic regression for credit card fraud detection. Machine learning classification algorithm used here and the task is implementing in Python language. The performance of the algorithm is evaluated by accuracy score, f1-score, precision and recall score.

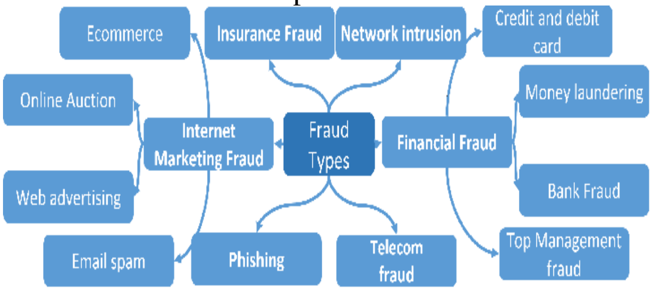
Keywords: Fraud detection, machine learning, logistic regression.

I. INTRODUCTION

Fraud can be defined as wrongdoer deception. In credit card transactions,fraud is defined as unauthorized use of an account by some unknown person . Mandatory preventions can be taken to terminate this misuse and the behavior of such fraudulent practices can be studied to reduce it and to protect against similar occurrences in the future. There are two mechanisms which are used to avoid fraud and its losses. Those two methods are known as Fraud Prevention and Fraud Detection. Fraud prevention is a method where it prevents the fraud to be happened and fraud detection is used when an unknown transaction is happened.

[1] The types of transactions are physical transaction and digital transaction. physical transaction credit card is directly involved. On the other hand, in digital transactions card is not used and the transaction happens through internet or telephone.

Credit Card Fraud is one of the massive threats to businesses today. However, to fight the fraud completely, understanding the structure of executing a fraud is important. Credit card fraudsters opt many numbers of ways to commit fraud. Card frauds is done either with the theft of the card or other information that necessarily be available to a practice a transaction.



Any card numbers are printed on the card, and a black stripe on the back of the card contains the data in a machine-readable form. It contains the name of cardholder name, card number, expiration date, CVV code, type of card and more methods to commit credit card fraud. A major challenge in applying Machine Learning to fraud detection is the presence of highly imbalanced dataset. In many available datasets, the majority of transactions are valid or genuine with an extremely small percentage of fraudulent ones. we apply classification approach i.e. Logistic regression. Our aim is to build a classifier which will be able to separate fraud transactions from genuine ones. We will compare the accuracy and effectiveness of this algorithm in detecting fraud transactions.

II. LITERATURE SURVEY

EXISTING SYSTEM:

In case of credit card fraud detection, the existing system is detecting the fraud after fraud has taken place. Existing system contains the large amount of dataset when end user comes to know about unpredictability in transaction he/she made raise complaint and then fraud detection system start it working. It first tries to find that fraud has actually happened after that it starts to track fraud location and so on. In case of existing system there is no acknowledgement of recovery of fraud and customers satisfaction.

PROPOSED SYSTEM:

The point of the proposed system is to develop a model which is capability to find the type of transaction performing by hacker from genuine user's credit card details.

A good fraud detection system should contain the following properties:

1. Must detect the frauds quickly.

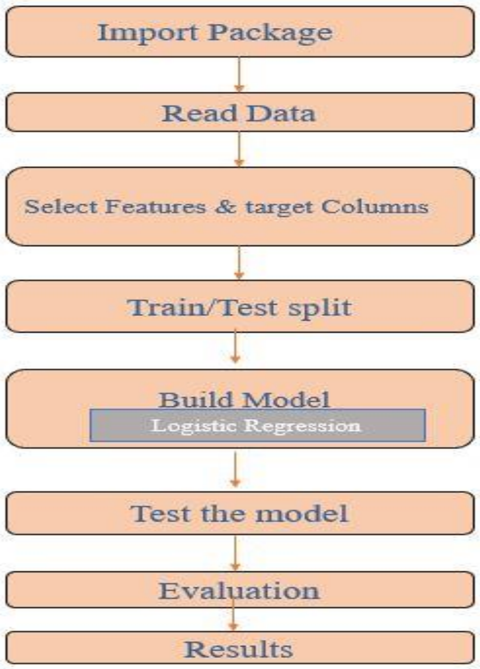
2. do not consider legitimate user as fraud user.

Advantages of Proposed system:

- Logistic Regression is one of the simple machine learning algorithms and is easy to implement.
- It provides great training efficiency in some cases. Also due to these reasons, training a model with this algorithm doesn't require high computation power.
- Logistic Regression has very high efficiency when the given dataset has features that are linearly separable.
- Logistic regression is easier to interpret and to train.

III. SYSTEM DESIGN

The figure below shows the system architecture diagram:



System architecture

A. Dataset Description:

The dataset holds transactions made by a cardholder. The given dataset consists of total 284,807 transactions, out of them 492 i.e., 0.172% transactions are fraudulent transactions. The input dataset is highly unbalanced. Since providing transaction details of an end user is considered to issue related to privacy, therefore most of the attributes in the dataset are transformed using PCA (Principal component analysis). V1, V2, V3, V4..., V28 are PCA applied attributes and remaining i.e., 'Amount', 'Time' and 'class' are non-PCA applied features, as shown in below table.

S.no	Feature	Description
1.	Amount	Transaction Amount
2.	Time	Time in seconds to specify the elapses between the current transaction and first transaction
3.	Class	1 – fraud 0- Not fraud

Table 1: Attributes of dataset

B. Working:

Logistic Regression is a classification algorithm for categorical values. In Logistic regression, we use one or more than one independent variables (X) to predict an outcome dependent variables (y). Logistic regression is similar to Linear regression but it tries to predict a categorical field or discrete target label instead of a numeric value. Such as yes/no, true/false, successful/unsuccessful, pregnant/not pregnant etc.
In Logistic regression independent variable(X) should be continuous, if categorical they must be dummied or indicator coded i.e. transform them into a numeric value. In Logistic Regression rather than fitting in a straight line or a hyperplane, the logistic regression makes use of a model which uses the logistic function i.e., sigmoid function to extract the output of a linear equation which is between 0 and 1. The function is defined as:

$$\text{logistic}(\eta)=\frac{1}{1+\exp(-\eta)}$$

In linear regression model, we have designed a model which defines the relationship between outcomes and features of a linear equation:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

Therefore, the finally expression is given as:

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

Where y is the forecast output, b0 is bias or intercept term and b1 is the coefficient for a single input value of+ (x).

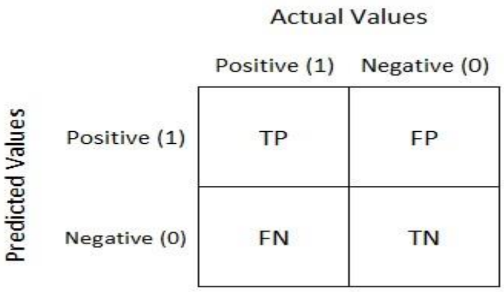
IV. EXPERIMENTAL ANALYSIS:

[3] To evaluate the results of the classification algorithms there are various parameter such as accuracy score, classification report, F1-score, confusion matrix etc.

1) Accuracy – It is the ratio of number of correct prediction to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

2) Confusion matrix - A confusion matrix is a table that is used to describe about the performance of a classifier model on a set of testing data for which the true values are known.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

3) Precision (Specificity)- The number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

4) Recall (Sensitivity) - It is the number of positive results divided by the number of all appropriate samples

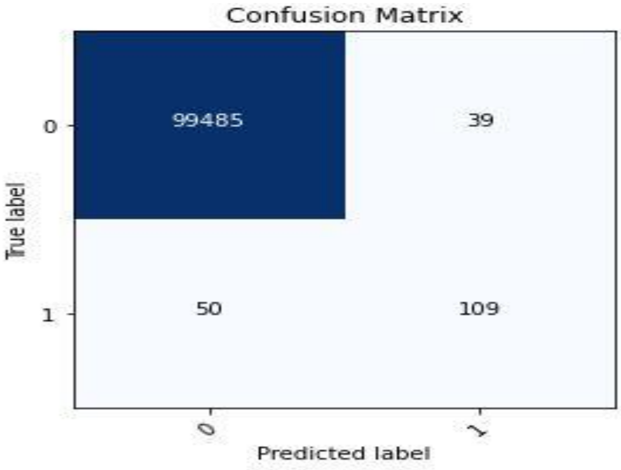
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) F1- score - F1 Score is the mean between precision and recall. The range for F1 Score is [0, 1].

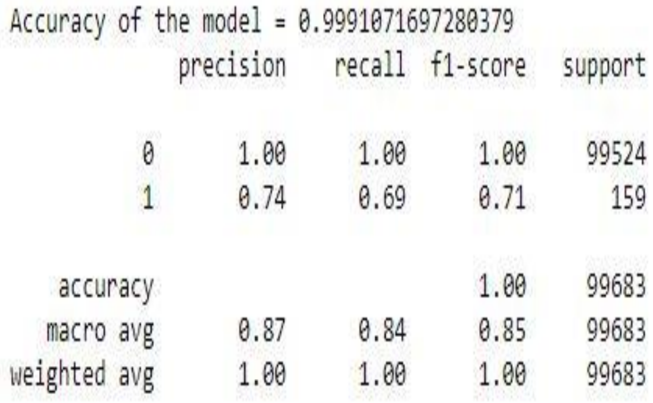
$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

V. RESULTS

In this paper, Logistic Regression algorithm was used to detect the frauds in credit card system model. To estimate the algorithm, 65% of the dataset is used for training data and 35% is used for testing data. Accuracy, F1-score, precision, and recall score are used to estimate the performance these four approaches. Evaluation of a model's performance was made in accordance to these metrics.



Classification Metrics:



```
Accuracy of the model = 0.9991071697280379
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     99524
           1       0.74      0.69      0.71       159

    accuracy                           1.00     99683
   macro avg       0.87      0.84      0.85     99683
weighted avg       1.00      1.00      1.00     99683
```

VI. CONCLUSION

Fraud detection in credit card is very serious matter in financial. The loss due to credit card fraud is increasing with the increase in various sector in which credit card is used. In this report, Logistic regression is used to detect the fraud in credit card system. Accuracy, F1-score, Confusion matrix, Sensitivity, Specificity was used to estimate the performance of the proposed system. The accuracy for the classifier was great. This technique helps to find out the credit card fraud.

This paper has surveyed the performance of Logistic Regression Algorithm.

REFERENCES

[1] Patel, Twinkle, & Ompriya, Kale. (2014). A Secured Approach to Credit Card Fraud Detection using Machin Learning. International Journal of Advanced Research in computer Engineering and technology, 3(5), 1576.

[2] Aswathy M S, Liji Samuel "Survey on Credit Card Fraud Detection".

[3] J.T. Quah and M. Sriganesh, "Real-time credit card fraud detection using Logistic Regression.

[4] N. S, Halvaiee and M. K Akbari, "A novel model for credit card fraud detection using Machine Learning.

[5] P. Srinivas Reddy, Logistic Regression in Machine Learning, International Journal of Engineering and Technology (UAE), 2019.

[6] Sowmya G, Machine Learning Strategies in engineering, Journal of Advanced Research in Machine Learning.

[7] Madhuravani B, The boosting approach to machine learning, Journal of machine learning research, 2019.