JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud alert using Machine Learning

Sejal Makam[1], Chetna Chouhan[2], Prajwal Vairalkar[3], S. M. Ingawale[4]

E&TC Department, SKNCOE, SPPU Pune, India

smmakam25@gmail.com [1]

chetanachouhan2k@gmail.com [2]

prajwalvairalkar0702@gmail.com [3]

smingawale.skncoe@sinhgad.edu [4]

***Abstract*- The rampant growth of technology and digitization has its own consequences. The credit card transactions have also increased leading to frauds. Credit card frauds lead not only to loss of money but also the loss of identity and privacy of the user. The dataset used in our system comes from a financial institution according to a confidential disclosure agreement. The proposed methodology makes use of the Logistic Regression (LR) algorithm for determining if the transaction is a fraud or not. The dataset containing the amount, time and transaction details is fed to the system. The user can interact with the fraud detection system through GUI. The transaction is matched against the fraud detection model. If the transaction is recognized as a genuine transaction then the user can safely proceed whereas if it is recognized as a fraud transaction then the user will be alerted. The work is implemented with Python machine learning models.**

***Keywords*-** Fraud detection, Credit cards, Machine Learning, Logistic Regression, Accuracy.

I.INTRODUCTION

The revolutionary growth of technology, the payment method has been simplified by the collaboration of the financial industry and IT technology. This leads the payment method of customers from cash payment to electronic transaction and this results in a climb of the number of cases in fraud transactions. 'Fraud' in credit card transactions is unapproved and false usage of an account by someone other than the owner of that account. Financial fraud is a consistently developing threat with far arriving outcomes in the money business, corporate associations, and government. As business is running towards e-commerce credit card transactions are going sky high. With the developed e-banking system some flaws within these systems increased fraudulent transactions. For frauds, the credit card is a naive target because without any risk a significant amount of money is obtained within a short period. To commit credit card fraud, fraudsters try to steal confidential information.

The purpose of this model is to make the users aware of such frauds and alert them. Machine Learning is the modern and an efficient way to prevent such mishaps and frauds. Various Machine Learning Algorithms can be used for this purpose and Logistic Regression is one such efficient example.

II. LITERATURE REVIEW

Previously a lot of research has been done in this field. Various combinations of algorithms have been used to provide an efficient and robust system. But however, with advancement new patterns have been developed for committing frauds. The previous technology that has been used is described in this section. The GBT Classifier and Py-Spark library is applied as a SQL like analysis to a large amount of structured or semi-structured data. GBT classifier does the classification of the data coming through the stream. This has been done using the latest AI methods and hence the false alerts are reduced [1]. Many supervised Machine learning algorithms are used for detecting frauds using real time dataset. Implementation of the super classifier has been done using the ensemble learning methods. In addition, the super classifier algorithm has been compared with several supervised algorithms that are available [2]. Due to a highly imbalanced dataset, SMOTE technique was used for balancing the data. The algorithms used in the experiment were Logistic Regression, Random Forest, Naive Bayes and Multilayer Perceptron. In this paper three datasets were used for comparison between Auto-encoder and Restricted Boltzmann Machine algorithms [3]. The behavior based approach has been made use of for classification. The support vector Machine (SVM) algorithm has been used to improve its accuracy. They have incorporated new user registration with finger print images. User login is authenticated using the same. Security question, customization and verification module is added. One time password (OTP) are utilized as a supplemental thing about multi factor authentication applications [4]. The comparative study of Local Outlier Factor Algorithm and Isolation Forest algorithm is done. This proposed model is implemented in Python. Numpy and Pandas are used for simpler tasks such as data storage and transformation. For data analysis and visualization, Matplotlib is used, Seaborn is used for statistical data visualization and for algorithms we used Sklearn [5]. In this work, fraud detection using artificial intelligence is proposed. The proposed system uses logistic regression to build the classifier to prevent frauds in credit card transactions. To handle dirty data and to ensure a high degree of detection accuracy, a pre-processing step is used. The pre-processing step uses two novel main methods to clean the data: the mean-based method and the clustering-based method. The System used for this

project is Windows 10. The programming language used for the implementation of the classifier is Python. The algorithm used for this proposed system is Logistic Regression [6].

III. PROPOSED METHODOLOGY-

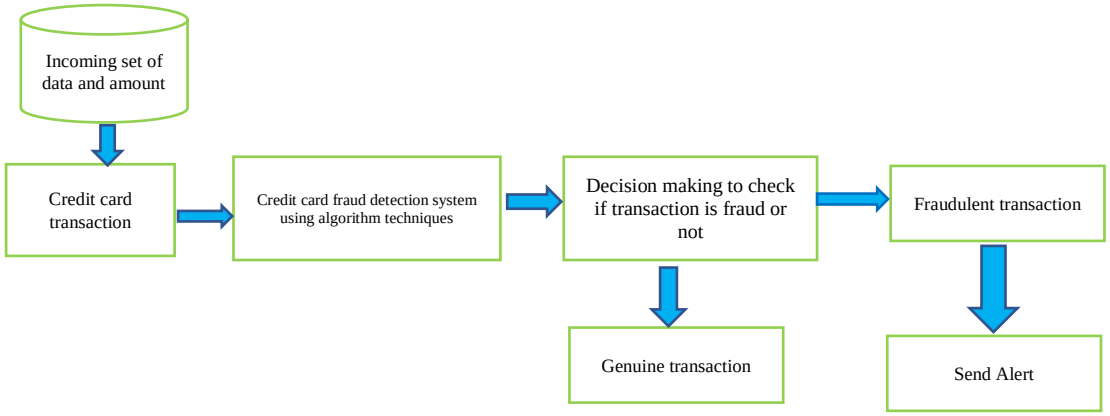

Figure 1. Block diagram of the credit card fraud detection system

The workflow begins with the previous transactions dataset whose main features are extracted through the PCA. Principal Component Analysis(PCA) is a form where unidentified variables are being readdressed. Development of classifier algorithms is the preeminent work of the training stage. After the training process it is supported with processed data.

1. The transactions that are sustained out using any credit cards are accepted with the required details.
2. This transaction is further given to Credit Card Fraud Detection System
3. The score obtained from the Credit card Fraud Detection System is further used to identify or decide the next action to be taken.
4. If the transaction is recognized as a genuine transaction, then it is sent for further processing of clearance.
5. If the transaction is recognized as fraudulent transaction, then alert or alarm is raised to highlight for the same and is stopped from further processing of that transaction.

IV. IMPLEMENTATION

A. Algorithm Description (Logistic Regression):

Logistic regression is one of the most popular machine learning algorithms, which comes under the supervised learning algorithm techniques. It is used for anticipating the categorical dependent variable using a given set of independent variables. Logistic regression anticipates the output of a categorical dependent variable. Therefore the result must be a categorical or discrete value. It can either be yes or no, 0 or 1, true or false, etc. But instead of giving the precise value as 0 and 1**,** it gives the probabilistic values which lie between 0 and 1**.** Logistic regression can mostly be employed to classify the observations and make use of different types of data and hence can easily determine the most effective variables used for the classification. The image given below shows the logistic regression function.

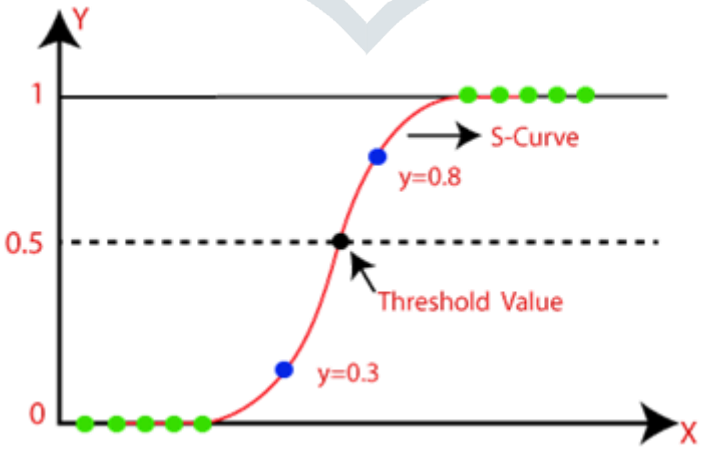

Figure 2.Logistic Regression Curve

The mathematical steps to obtain logistic regression equations are given below:
The equation of the straight line can be written as:

$$y = a0 + a1 \times x1 + a2 \times x2 + \cdots ak \times xk \quad \text{.....} \quad (1)$$

In logistic regression, y can be between 0 and 1 only, so we divide the above equation by (1 − y):

$$(y/1-y) = 0 \; for \; y = 0 \; and \; \infty \; for \; y = 1 \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Hence, the logistic regression equation is as given below:

$$\log [\, y \; /1- \; y \,] = a0 + a1 \times x1 + a2 \times x2 + \cdots ak \times xk \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

I.Testing the Classifier:

Since the cross-validation method divides the database into 10 parts, there are 10 testing data sets. Each testing data set is used to test one classifier (there are 10 classifiers). This in turn gives the model an advantage by allowing it to use the whole database for testing as well as for training. The testing process is tightly coupled with the accuracy of the model. Calculating the final accuracy involves calculating the accuracy of each classifier. Formally, let $Acck\ C$ denote the accuracy of a given trained classifier. Then, the final accuracy of the final classifier (ACC_{F^C}) is obtained based on the "average" mathematical operation

$$ACC_F^C = \frac{\Sigma_{k=1}^{10} ACC_F^C}{k}$$

II. Evaluating the Classifier:

In general, a confusion matrix is an important tool for analyzing how well a classifier can recognize records of different classes . The confusion matrix is developed on the basis on the following terms:

1) True positives (TP): positive records that are correctly labeled by the classifier.
2) True negatives (TN): negative records that are correctly labeled by the classifier.
3) False positives (FP): negative records that are incorrectly labeled positive.
4) False negatives (FN): positive records that are mislabelled negative.

Table III shows the confusion matrix in terms of the TP, FN, FP, and TN values. Relying on the confusion matrix, the accuracy, sensitivity, and error rate metrics are derived. For a given classifier, the accuracy can be calculated by considering the recognition rate, which is the percentage of records in the test set that are correctly classified (fraudulent or non-fraudulent). The accuracy is defined as

$$\textbf{\textit{Accuracy}} = \frac{(TP+TN)}{number\ of\ all\ records\ in\ the\ testing\ set}$$

Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Confusion Matrix

B. Graphical user Interface:

A graphical user interface (GUI) is an interface through which a user interconnects with electronic devices such as computers menus and other visual indicators or representations. GUIs graphically display guidance and related user commands, different text-based interfaces, where data and commands are strictly in text. GUI representations are managed by an electronic stylus such as a mouse.

For designing of user interface the Tkinter is used which is the only framework that's built into the Python standard library. The foundational element of a Tkinter GUI is the **Windows.** Windows are the containing element in which all other GUI elements . These other GUI elements, such as text boxes, labels, and buttons, are known as **widgets**. Widgets are contained inside of windows.

The application will be a desktop application/website which users can visit on their PCs and smartphones. The front end of this application will be developed using python tools. This application will have following features:

1) User Registration
2) User Login
3) Enter transaction details
4) Get alert email if fraudulent
5) Perform Payment

In the proposed system, while registration we take required information of the user such as user name, address, email id, password which is efficient to detect fraudulent activity.

To ensure a high level of security various constraints are applied to set strong password like length should be at least 6,should include one numeral, should include both uppercase and lowercase letters and also symbols .Once all the given constraints get satisfied the account will be created successfully.

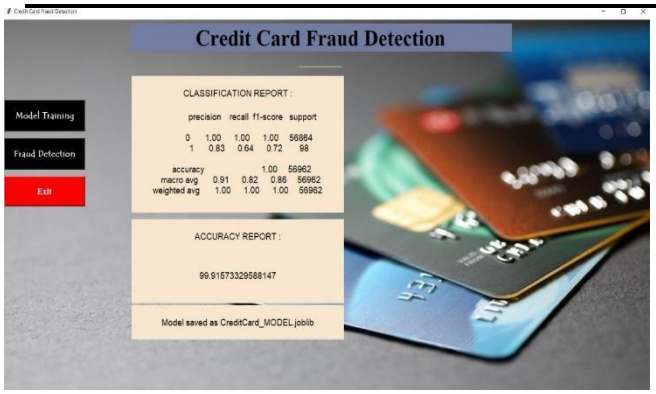

Figure 3. GUI Model

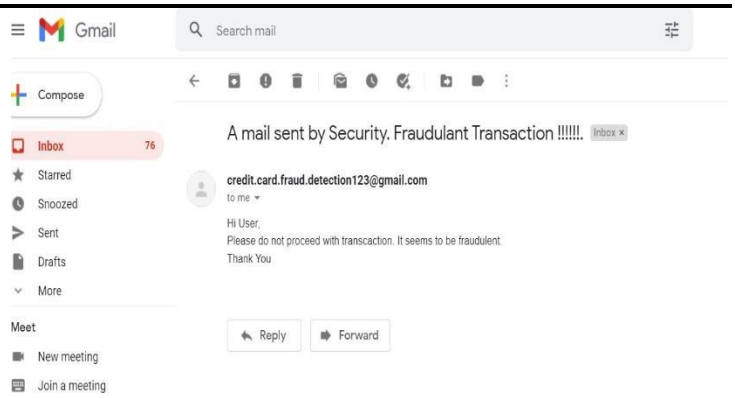

Figure 4. Email Alert

V. EXPERIMENTATION

A. Model Training:

I. Importing the necessary libraries: All the required libraries are imported in one place — so that they can be modified quickly. For this credit card data, the features that we have in the dataset are the encrypted version of PCA.

Importing Dataset

II. Data Collection: Data collection gives us the description of the dataset used. The dataset used contains 2,84,807 records of credit card transactions that happened in the duration of just 2 days. This dataset is unbalanced as it has a total of 492 fraud entries and 2,84,515 genuine entries i.e. it is just 0.17% of total records. The original features are concealed with V1, V2, V3, ...V28. The last column here stands for whether it's fraud or non-fraud transaction i.e. represented by 0 and 1 respectively.

***III. Data Balancing*:** Imbalanced classes are a general issue in ML based classification where there is an abnormal count of each class. It occurs due to the fact that ML Algorithms are typically intended to improve precision by diminishing the errors. In this manner, they don't consider the class or balancing the ratio of classes. As out of 2.84807 transactions just 492 fraud transactions exist, which makes it quite difficult to build a standard model with less number of fraud transactions. Thus, we use pandas in python to make it equal i.e. we decrease the no. of legitimate transactions to balance it with the number of fraud transactions in equal proportion.

IV. Feature Extraction: We use heatmap technique to find the significant feature that can distinguish the classes properly and ultimately that affects the accuracy of the detection algorithm. Heatmap provides a good idea about the major and minor values in the matrix as different colored cells that define the values. Here, rows/columns of the matrix are clustered in sets. Thus, the features which look most significant are recognized and used further for model training.

```
plt.figure(figsize = (40,10))
sns.heatmap(card.corr(), annot = True, cmap="tab20c")
plt.show()
```

V. Outlier Detection: The outlier detection technique measures the distance of each data similar to the clustering technique, but is used to find specific data and rules that are separated from the total data. The values which are not in sync with the linear graph are considered as outliers. Here our aim is to reduce the outliers to have a better trained model. We use the numpy library in python for this.

VI. Classification: The task of classification occurs in a wide range of applications. In a broad sense, the term could relate to any context in which some decision or forecast is made on the basis of currently available information. It works on a set of predefined classes on the basis of observed attributes or features. Here the aim is to establish a rule whereby one can classify a new observation into one of the existing classes.

VI. RESULT

70% of the data is used as training data and 30% of the data set is used as test data. The results refer to the accuracy, precision, recall, and the F1 score of the model.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56864
1	0.83	0.64	0.72	98
accuracy			1.00	56962
macro avg	0.91	0.82	0.86	56962
weighted avg	1.00	1.00	1.00	56962

```
Accuracy :  99.93153330290369
Accuracy: 99.93%
```

The accuracy of the model obtained is 99.93%. The precision for fraud and non fraud are 0.83 and 1.00 respectively. The value of recall for fraud and non-fraud cases are 0.64 and 1 respectively. Similarly, the value of F1 score is 0.72 and 1.

CONCLUSION:

Credit card fraud is without uncertainty an act of criminal dishonesty. So to control the risk of the financial loss of client and the institute the which has been approached is true stands off the requirements the data set which was obtained from the financial institute and the website and As, the study of differing research papers and the reports that were inspected has given the valuable information of the various type of data set available these research papers have opened vast windows of various information of the various type of algorithm as specifically the Random Forest, SVM and GBT classifier, which specifically work with data set which is used in training the model but the accuracy rate of the of Logistic Regression is more than the remaining algorithm as the accuracy was obtained by the equation in logistic regression which also establishes the accuracy, many other like f1 score, recall, for the ease formula is also shown in above cases.

ACKNOWLEDGEMENT

It gives us great pleasure in presenting the preliminary project paper on 'Credit Card Fraud alert using Machine Learning'. We would like to take this opportunity to thank my internal guide Prof. S.M. Ingawale and co-guides Prof. S.S Palnitkar , Prof. N.S Nikam for giving us all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful. We are also grateful to Dr. S.K. Jagtap, Head of E&TC Engineering Department, SKNCOE for her indispensable support, suggestions. We would also thank our Principal Dr. A.V. Deshpande for his great insight and motivation. Last but not least, we would like to thank my fellow colleagues for their valuable suggestions.

References

[1] Vinaya D S 1, Satish B Basapur2, Vanishree Abhay3, Neetha Natesh, "Credit Card Fraud Detection Systems (CCFDS) using Machine Learning (Apache Spark)" , International research journal of Engineering and technology (IJERT) Vol. 07 Issue 08 Aug 2020.
[2] Sahil Dhankhad , Emad A. Mohammed, Behrouz Far, " Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study", IEEE International Conference on Information Reuse and Integration for Data Science, 2018.
[3Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, " Credit Card Fraud Detection - Machine Learning methods", 18th International Symposium INFOTEH-JAHORINA, 2019.
[4] Metthilda Mary,M. Priyadarshini, Dr.Karuppasamy.K , Ms.Margret Sharmila.F, "Online transaction fraud detection System", International Conference on Advance Computing and Innovative Technologies in Engineering, (ICACITE), 2021.
[5] Pawan Kumar, Fahad Iqbal, "Credit Card Fraud Identification Using Machine Learning Approach",
[6] Hala Z Alenzi, Nojood O Aljehane, "Fraud Detection in credit cards using Logistic Regression", International Journal of Advanced Computer Science and Applications, Vol.11 ,No.12, 2022.
[7] Vinod Jain, Mayank Agrawal, Anuj Kumar, "Performance Analysis of Machine Learning Algorithms on Credit Cards Fraud Detection", International Conference on Reliability, Infocom technologies and optimization (ICRITO), June 2020.
[8] Hyder John, Sameena Naaz, "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest", International journal of Computer Science and Engineering, 2019.
[9] S P Maniraj ,Aditya Saini, Swarna Deep Sarkar Shadab Ahmed, "Credit Card Fraud Detection using Machine Learning and Data Science", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 8 Issue 09, September-2019.
[10]C.Sudha ,Dr D. Akila, "Credit card fraud detection system based on Operational and Transaction feature using SVM and Random forest", International Conference on Computation, Automation and Knowledge Management(ICCAKM),2021.
[11] Anuruddha Thennakoon, Chee Bhagyani , Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning", International Conference on Cloud Computing, Data Science and Engineering, 2019.
[12] Abrar Hayat Nadim, Ibrahim Mohammad Sayem, Aapan Mutsuddy, Mohammad Sanaullah Chowdhury, "Analysis of Machine Learning techniques for Credit Card Fraud Detection, International Conference on Machine Learning and Data Engineering (iCMLDE), 2019.
[13] Pradheepan Raghavan, Neamat El Gayar, "Fraud Detection using Machine Learning and Deep Learning", 9 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019.
[14] Shantanu Rajora ,Dong-Lin Li , Chandan Jha , Neha Bharill , Om Prakash Patel , Sudhanshu Joshi , Deepak Puthal , Mukesh Prasad, "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance", 2018.
[15] Agrawal Tina N, Patil Swati G, Ahire Swati K, Prof.N.A.Suryawanshi, "Credit card fraud detection using machine learning techniques", Resincap Journal of Science and Engineering Volume 3, Issue 2 February 2019.
Dataset link:-https://www.kaggle.com/mlg-ulb/creditcardfraud
DataCamp, website Avaliable : https://www.datacamp.com/community/tutorials/understanding-logisticregression-python