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 Detection Using Machine Learning

[1]Ch. Vishwanath Reddy, [2]T. Rishitha, [3]P.Kavya Sri, [4]Dr.M.I.Thariq Hussan
[1,2,3]UG Scholars, [4]Professor and Head
[1,2,3,4]Department of CSE(IOT),
Guru Nanak Institutions Technical Campus (Autonomous), Hyderabad, India

Abstract : It is compelling that credit card able to pick out fraudulent credit card transactions so users can be delinquent and can be assure of their assets. Those hurdles can be address with Data Science and its eminence, along with Machine Learning, cannot be exaggerate. This project envisage to portray the modelling of a data set using machine learning and XG Boost with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem involves customizing past credit card transactions with the data matching to be fraud. This model is then used to recognize whether a new transaction is illicit or not. We precisely lean towards finding deceitful transactions and minimizing the fraud transactions using XG boost. Credit Card Fraud Detection is a complex swatch of classification. In this complex system , we have focused on pre-processing and analyzing labelled data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Random forest, XG Boost, and Decision factor on the PCA transformed Credit Card Transaction data.

***Index terms* - Credit card fraud, applications of machine learning, local outlier factor, automated fraud detection, Random forest, XG Boost, and Decision factor**

I. Introduction

Fraud in credit card transactions is illicit and unpalatable handling of an account by someone other than the owner . Some prerequisite measures can be taken into consideration to end this illicit and unlawful practices. Some unlawful practices can be analyzed and studied to minimize it and match with illicit transaction.. In other terms, Credit Card Fraud can be depicted as a instance where one uses someone else's credit card for transactions while the owner and the card issuing authorities are unaware about handling of his card. This detections including examine the population of users who regularly do transaction, avoid unlawful practices which incorporate intrusion and fraud

This is a growing absurd complication that claims the curiosity of modern machine learning and libraries and the solution can be driven out from all odds and fraudulent adversity. While the solution can be impracticable and ludicrous by various other factors such as class imbalance, only minimizing the fraudulent transaction. The patterns of transactions affix the labelled datasets to compare and contrast over the period of time and dimensional properties will changes .The statistics will dramatically changes to time and they are being ephemeral to the transactions

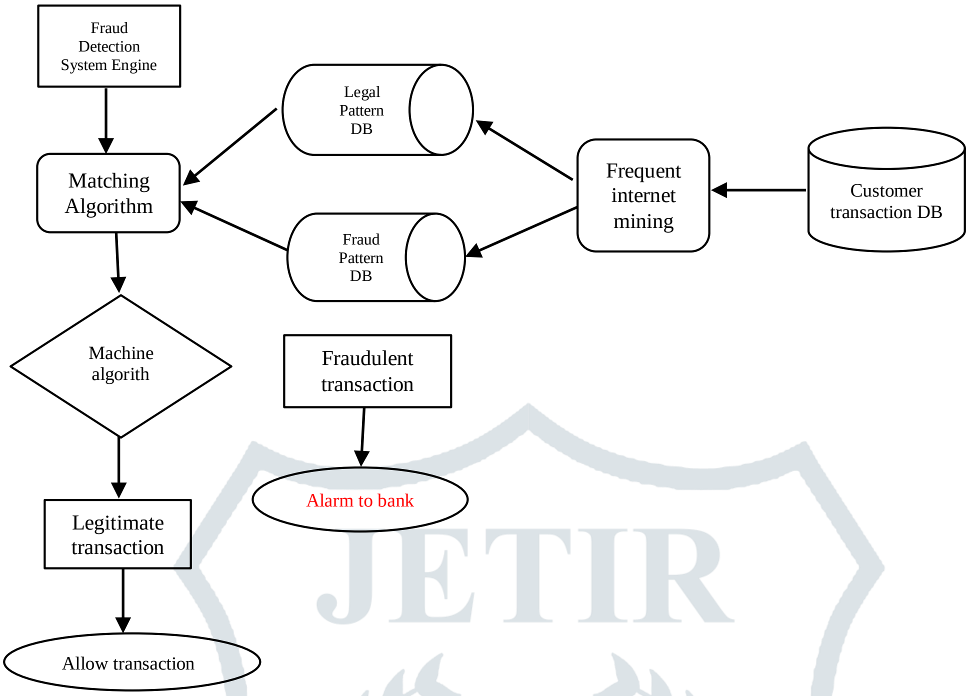
Machine learning algorithms are supervised by labelled and unlabeled datasets to analyze all the authorized transactions and sitrep the wary transactions. This report are inspected by expert by which system contacts that cardholder to examine whether the transaction was acting by him. Through this method we can find out whether the credit card transaction is fraudulent transaction or safe. This are some valid obstacles to implement the fraud detection, but high usage of payments, scanning, online transaction by automatic tools which leads to transactions to self authorize which ultimately enroute to fraudulent transactions

The inspectors come up with estimation to the machine learning system, which is used to supervise over the datasets, train the datasets and update the machine learning algorithm periodically to improve overall efficiency over the time.
These machine learning algorithms are gradually developed to confront new fraudulent tactics and strategies by criminals.

These frauds are listed as:

- Offline Credit Card Frauds
- Online Credit Card Frauds
- Telecommunication Fraud
- Bankruptcy
- Intrusion of Devices
- Application Fraud

- Counterfeit Card
- Stealing of Card



Some of the approaches used to detection of such fraud are:

- XGBoost
- K-Nearest Neighbour
- Fuzzy Logic
- Bayesian Networks
- Support Vector Machines
- Decision tree
- Logistic Regression
- Genetic Algorithm
- Artificial Neural Network

II. LITERATURE REVIEW

Fraud transactions are illicit and illegitimate intended to sequent profit and personal benefit to criminals. It's against the rules and policies in the governance and earn unauthorized unorganized financial benefits. Various literatures concern about fraud detection in the specific field large numbers of surveys are conducted and many research papers available across the internet. A comprehensive study done by many famous authors, research scholars on credit card fraud detection, Xg boost. Although these machine learning algorithms and libraries, datasets could provide a fugacious solution, later many examination were done but could match with maximum accuracy because of constant changes in trained data and there is no feasible solution to credit card fraud detection till now

A similar research was authorized by Wang and Wn fang on outlier mining and machine learning distance sum techniques to accurately predict fraudulent transaction of transacted label dataset. In this typical process, attributes are taken into consideration and users behaviors and calculates the predetermined values.

It is affected by various factors like detecting things that are detached from main system. Complex techniques such as labelled data are able to perceive illicit instances of the card during a transaction this will helps us to create representations of the deviations of one instance from another given reference datasets.

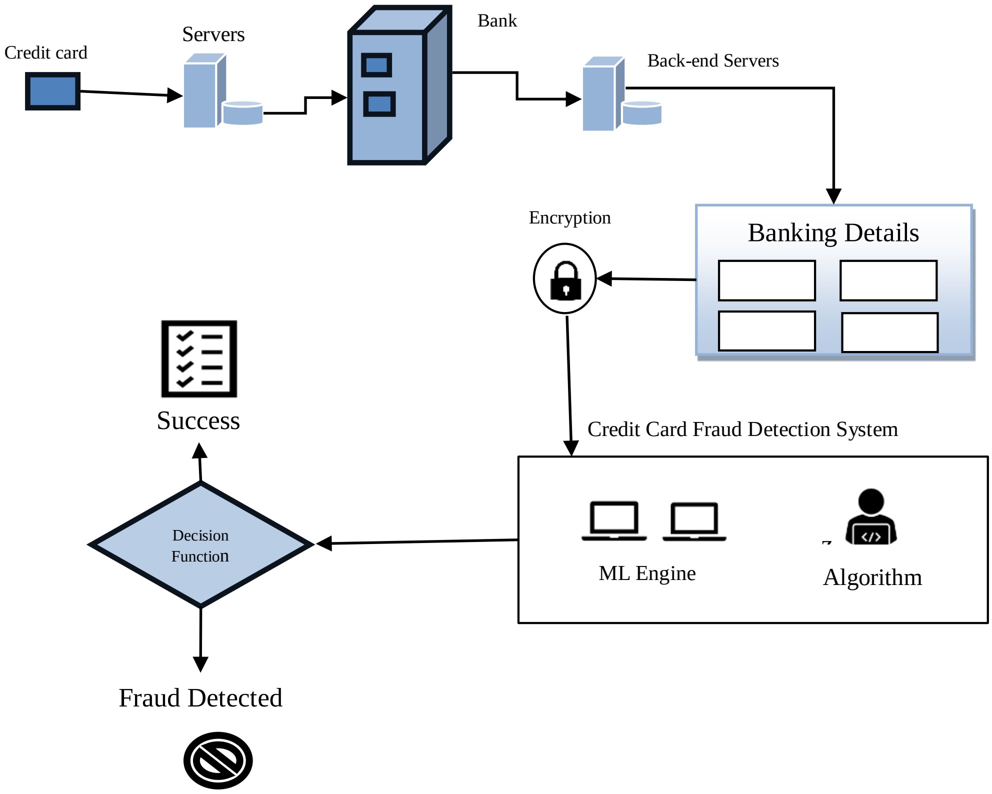
Lots of new advancements has been made to minimize the credit card fraud detection. Several attempts have been made to upgrade the wideawake estimation interaction in the case of fraudulent transaction.

In terms of unauthorized unorganized transactions, a leverage automated system would alert us and sends an estimation to the ongoing transaction denial. Leveraging of a complex process will prove their precision and accuracy in figuring out the fraudulent transaction and minimizing the number fraudulent transaction

III. RESEARCH METHODOLOGY

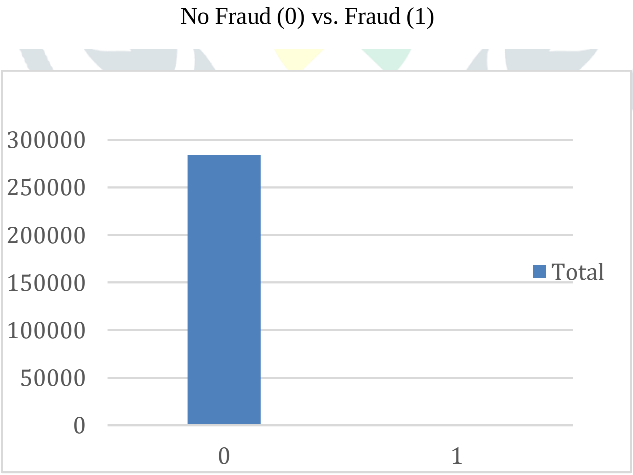
The paper proposes this approaches uses the latest machine learning algorithms to detect anomalous fraudulent transaction called outliers and XG Boost.

The full architecture diagram can be represented as

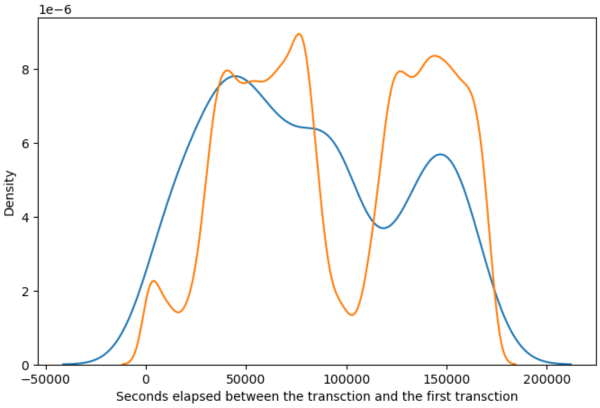


Our datasets are driven from kaggle, which is a website provides datasets.
There 31 columns out v1-v28 are used ensure date protection and protect sensitive data. Other parameters denotes amount, class time which shows the time gap between first transaction and next one. There are two classes to represent class 1 depicts valid transaction and class 0 represent fraudulent transaction

Graphs is drawn to check for fraudulent transaction and to comprehend visually



This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.



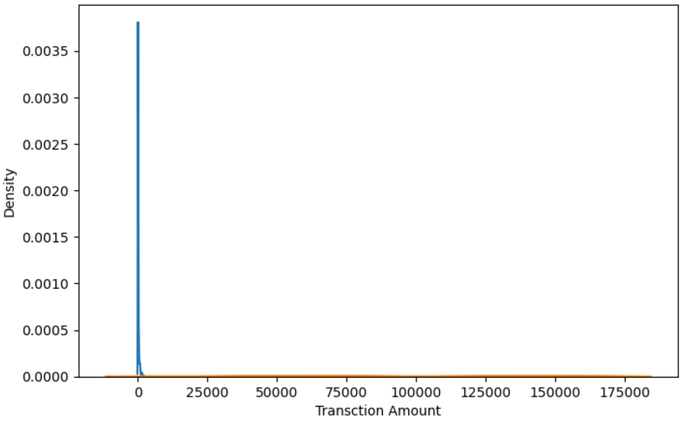
Distribution of Time Feature

The graphs depicts about transactions over three days.
The label datasets are now analyzed and processed. The amount and time are systematized and class is adjusted to make ensure fairness of evaluation .This modules explains about working of the algorithms.

The following modules fit into a model and follow outlier detection modules are applied on it

- Local Outlier Factor
- Isolation Forest Algorithm

Representation of this graph says about how much amount transacted. Most of transactions are small and come close the total amount.



Distribution of Monetary Value Feature

3.1 Libraries
The ensemble module in the sklearn comprises of classification and XG Boost.
Numpy, scipy, matplotlib libraries are efficient for data analysis and machine learning algorithms
Numpy is used for numerical python. scipy is used for scientific python
Matlplotlib is used for plotting of graphs. All this programs are executed on Jupyter notebook and Collab platform. All pseudocodes for following for their alogirthms

3.2 Local Outlier Factor

It is an Unsupervised Local Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures how isolated or "outlying" a data point is in its local neighbourhood. We precisely lean to locality by k-nearest neighbours.

The pseudocode for this algorithm is written as:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor
```

```
np.random.seed(42)
X = 0.3 * np.random.randn(100, 2)
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.concatenate((X + 2, X - 2, X_outliers))
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
plt.figure(figsize=(10, 7))
plt.title("Local Outlier Factor (LOF)")
plt.contourf(xx, yy, Z, cmap=plt.cm.Blues_r, levels=np.linspace(Z.min(), 0, 7), alpha=0.5)
normal_obs = plt.scatter(X[:200, 0], X[:200, 1], c='white', edgecolor='k', s=20)
abnormal_obs = plt.scatter(X[200:, 0], X[200:, 1], c='red', edgecolor='k', s=20)
plt.axis('tight')
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.legend([normal_obs, abnormal_obs], ['Normal Observations', 'Abnormal Observations'], loc="upper left")
plt.show()
```

By comparing how one data point differs from its nearby points, we can find data points that stand out because they are significantly different. These unusual data points are called outliers. Since the dataset is very large, we only looked at a small part of it to save time when analyzing. The conclusion is based on analyzing the entire dataset, is available in the paper

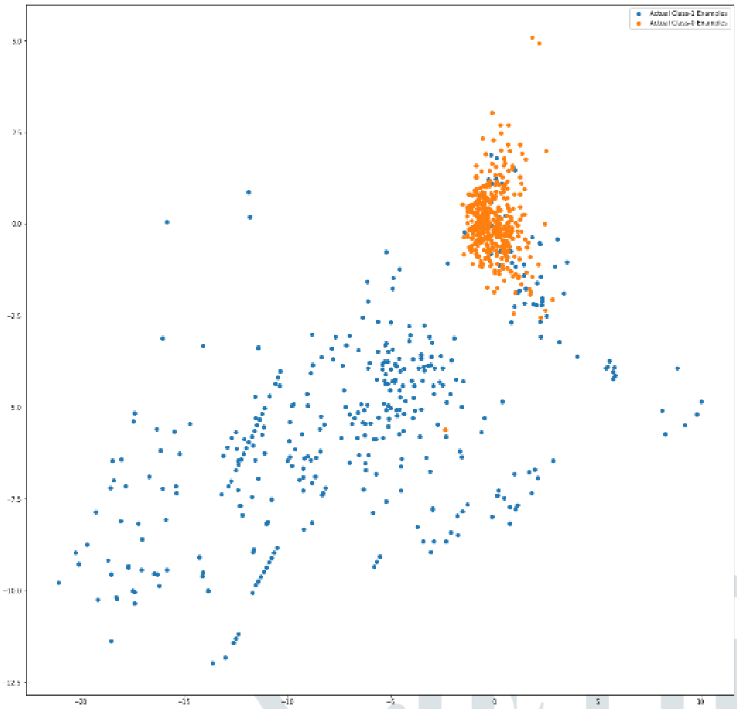
3.3 Isolation Forest Algorithm

The Isolation Forest algorithm operates by segregating data points through a process of deliberate isolation. It achieves this by arbitrarily singling out a feature and subsequently choosing a split value at random within the range defined by the maximum and minimum values of the selected feature. The essence of this method is encapsulated in recursive partitioning, which can be visualized through a tree structure. In this framework, the number of splits necessary to isolate a particular sample corresponds to the length of the path from the root node to the terminating node.

The pseudocode for this algorithm can be written as:

```
import numpy as np
from sklearn.ensemble import IsolationForest
np.random.seed(42)
X = 0.3 * np.random.randn(100, 2)
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.concatenate((X + 2, X - 2, X_outliers))
clf = IsolationForest(random_state=42)
clf.fit(X)
y_pred = clf.predict(X)
n_outliers = np.count_nonzero(y_pred == -1)
print("Number of outliers:", n_outliers)
```

From the plotting of Isolation Forest algorithm, we get the following figure:



Segregating them arbitrary gives short paths for different anomalies. when they are said to be anomalies, they are scaled as they produces shorter paths. When illicit transactions are found, the automated system promulgate them to next level authorized authority for action by testing their accuracy and testing

IV. IMPLEMENTATION

From the realms of our history, bank doesn't share the customer data due to privacy rules and security issue of the bank, so it is very hard to implement in real life . Due to security issues and bank does inculcate their customers to announce their Bank credentials details. I browsed nearly 1.57 lakhs csv file datasets records of Michigan bank, for the confidentiality of the bank only summaries of impactful approaches is depicted below. This is complex and absurd process in realistic fraud detection.

After applying this method, level 1 is still restricted to sufficient to be checked the case one by one sequentially level 1 is high probability and occurrence of being fraudsters. In order to increase maximum efficiency and effectiveness of time through adding new element in the query , this element can be sequence of their contact number, email I'd and this small scales are applicable on other levels.

V. RESULTS

The output of the program is number of false positives and detects and contrasts it with the actual value. It depicts the accuracy precision and score of the machine learning algorithms . When XG Boost is used the increase is 8 percentage of the entire datasets. XG Boost gives 91 percentage of accuracy , where others factors doesn't give like local outlier factor where isolation factor gives maximum efficiency same like XG Boost.

These results along with the classification reports for each machine learning algorithm in the output as follows
Where class 1 fraud transactions
Where class 0 means valid

Results when 8% of the dataset is used:
Results with the complete dataset is used:

```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

             precision    recall   f1-score    support
     0          1.00       1.00      1.00      284315
     1          0.33       0.33      0.33         492

  accuracy                           1.00      284807
```

```
   macro avg       0.66      0.67       0.66      284807
weighted avg       1.00      1.00       1.00      284807

Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

              precision    recall   f1-score    support
           0       1.00      1.00       1.00     284315
           1       0.05      0.05       0.05        492

   accuracy                             1.00      284807
   macro avg       0.52      0.52       0.52      284807
weighted avg       1.00      1.00       1.00      284807
```

VI. CONCLUSION

Credit card fraud detection is illicit and illegitimate practice and improbity .This paper has depicted various fraud transactions and their detections methods. This article depicts about the relationship between machine learning algorithms and pseudocodes, explaining it's implementation and experimentation. The accuracy of this project is around 91 percentage which is considered as highest one out of all algorithms Because XG Boost and isolation forest algorithm. In terms of efficiency and accuracy it is also very high in margin for consideration. As there only two transactions records in the paper only small part of the data can be available to us ,if use this project for larger datasets the accuracy levels increases as the data increases and it's efficiency increases . This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

VII. FUTURE ENHANCEMENTS

Since we could not create a automated system after a lot of insisting experiments which could not gives 100 percent accuracy. Although we began to end with better system to detect. This system could gives around 91 percentage of accuracy and effective efficiency. As any project there is chance for further reference and improvements This combined machine learning algorithms to be integrated together which improves the final result in the all factors like accuracy and efficiency. As the large complex datasets add into it ,the accuracy of the system increases and we can add more multiple algorithms to this project to bring efficiency and further improvement. This upgrades the leverage and versatility of this project. A lots of further improvement can be add up and detects the fraudulent transaction. Due to security issues bank will not subsistence for data we required. We hope someone will bring atmost accuracy that is great than 91 percentage.

VIII. REFERENCES

[1] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence

[5] “Credit Card Fraud Detection through Parenclitic Network Analysis By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008