Grenze International Journal of Engineering and Technology, July Issue

Credit Card Fraud Detection using Machine Learning Algorithms - Study of Customer Behaviour

Karanam Sravya[1], Kasthuri C M[2], Koramutla Ramesh Meghana[3] and Dr. A S Poornima[4]
[1-4]Department Of Computer Science Engineering, Siddaganga Institute of Technology
Tumkur, India

***Abstract*—Credit Card Fraud Detection Using Machine Learning Algorithm is a study of customer behaviour in bank transaction. In this we are taking an original bank transaction dataset, and dividing it to three different levels of fraud rate. We apply various machine learning algorithms to these three versions and identify the fraud rate and record the same. It reports the accuracy of each algorithm and how it varies as the fraud rate increases. It also helps us to know which attributes must be considered and necessary for identifying the fraud in transactions. This project gives the insight about three machine learning algorithms that is, Support vector machines, Local Outliers and Isolation forest for credit card fraud detection. These three algorithms are applied to three different datasets with increasing fraud rates and their accuracies have been recorded via graphs and charts.**

***Index Terms*— Credit Card Frauds, Customer Behavior patterns, Machine Learning, Isolation Forest, Local Outlier, Support Vector Machine.**

I. INTRODUCTION

In this dynamic world where there are no speed breakers in technology, innovation and modernization, people are updating to endless methods and features which make the effort of doing work easier. One of the best example for this is, Banking.

Banking is the primary chore which helps to safeguard our money and also ensure proper flow of it. The Banks secures the productivity of people by keeping their money flow intact without any errors or frauds. Banking application involves transaction of money i.e., withdraw, deposit, loans etc., Due to technological benefits, these transaction have been digitalized using credit cards, debit cards, online banking etc., which helps us to transfer money within our finger point and without actual physical movement of money.

Obviously, Banking applications adopted to these new trends to ease the transaction activity but there is a lot of disadvantages associated with it. One of the major among them is the credit card frauds. Credit card frauds have been increasing day by day due to advancing technologies and also lack of knowledge about how to use these application securely. The people are sharing their debit card and credit card details to anonymous people or phone call without understanding the consequence of it. By sharing the details it makes the offender, to easily steal the credit card or effect the flow of money. Therefore, it have become very necessary to detect these frauds to make the banking experience more hustle free and simple.

Frauds are difficult to detect as there is no trace or particular method to detect it. To distinguish a fraud and a normal transaction requires a lot of analysis and previous data history of the banking customer transactions.

Grenze ID: 01.GIJET.6.2.12

By using this data history as input to machine learning algorithms we can train our model to detect the frauds. But, the major question which arises here is that, which machine learning algorithm is the best to use. Or how different algorithms differ in accuracies when applied to different datasets. The answers to these question is provided by “Credit Card Fraud Detection Using Machine Learning Algorithms”.

Machine learning is the advancing field of artificial intelligence which helps to identify the behavior of customers more accurately and determine their patterns in various field. We can use these patterns to provide greater benefit to the customers and achieve customer satisfaction. By using this field we are identifying the customer transaction pattern, if there are anomalies in these we are recording it as a fraud, using different algorithms. And also comparing this algorithms’ accuracies which will give clarity about which algorithm is the best. The algorithms we are considering here are Isolation Forest, Local Outlier and Support Vector Machines.

In this paper we are considering frauds by taking the original credit card transaction dataset and dividing it into three versions with increasing fraud rates. And we are applying various machine learning algorithms such as, Support vector machine, Local outlier and Isolation forest to these datasets and recording its accuracies. We are going to compare the accuracy of each algorithm among three versions and also compare the accuracy of each algorithm with another. By this, it will give the clear idea about which algorithm is best suited to identify the credit card fraud rates in banking transaction.

II. LITERATURE SURVEY

Credit Card Fraud Detection is not easy as it sounds. It requires critical analysing to identify the difference between normal transaction and fraud. To identify this, a lot of solution were proposed and used. Where Machine learning has been identified as the best [1].

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a sub substitute part of artificial intelligence. Machine learning algorithm builds a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. The term machine learning was coined by Arthur Samuel in 1959, an American IBMer and pioneer in the field of computer gaming and artificial intelligence [2].

To Understand which field and method is best to detect credit card frauds. And also which all machine learning algorithms are feasible, numerous literatures and theories were proposed already and are in public usage. A Comprehensive survey “Machine Learning Methods for Analysis Fraud Credit Card Transaction” by Megasari Gusandra Saragih et al. [3] revealed that the credit card and debit card fraud activities are actively increasing day by day with advancing technologies. And also found out that the lack of knowledge about usage of credit cards is the major contributor to these frauds. They also identified that the passwords used by the customers are easily hackable. They also say about how hard is to detect the frauds made by various methods like messages, telephone calls etc., They also used local outlier algorithm to identify the frauds and showed the accuracy of it.

The similar research in the field of credit card fraud detection is “Real-Time Credit card Fraud Detection Using Machine Learning” by Anuruddha Thennakoon et al. [4], in this paper, they propose a off-centre credit-card fraud detection system by detecting four different patterns of fraudulent transactions using best suiting algorithms and also addressed the related problems. They used predictive analytics and an API module, by which the end user is notified over GUI the second fraudulent takes place using this, they can take further action based on the fraud committed.

Another literature in this domain is “Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine” by Apapan Pumsirirat and Liu Yan [5]. This tells that online transaction in growing technology is very essential and thus, detecting maximum frauds become important. It proposes two deep learning techniques that is, Auto-Encoder (AE) and Restricted Boltzmann Machine (RBM) to detect frauds in real time. They proved AE and RBM are the best suited methods in deep learning for credit card fraud detection when there is a bulk dataset using benchmark experiments with other tools. They also guaranteed in their paper that AE and RBM are best techniques to identify frauds when we have huge dataset and it gives more accurate results.

The other alternate method proposed for our approach is “Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models” by Navanshu Khare and Saad Yunus Sait [6]. In this paper they used four machine learning algorithms that is logistic regression, SVM, Decision trees and Random forests to detect the frauds and also recorded their accuracies. They concluded that Random forest algorithm is the best suitable algorithm when we take simple machine learning algorithms into account.

In the literature, “Credit Card Fraud Detection using Machine Learning and Data Science” by S P Maniraj et al. [7] tells that credit card fraud is an act of dishonesty and robbery. It lists out the common methods of fraud along with their detection methods and reviewed recent findings in this field. It also explains how machine learning can be applied to get better results in fraud detection. It illustrates Local outlier and Isolation forest algorithms and compares their accuracies. It applies the algorithms to the small dataset and records its accuracy and discusses how this accuracy varies when applied to the huge datasets.

The very different approach for credit card fraud detection is proposed in “Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm” by Sahayasakila et al. [8]. It specifies, Smote (Synthetic minority optimization technique) which is a machine learning technique used for classification of data and it also solves data imbalance problem to get better results. This technique is used to synthesize all the fraud transactions from the original fraud transaction i.e., make dataset balanced. Then this synthesized fraud transactions are optimised using whale optimization algorithm. In this paper they explained in detail how whale optimization algorithm will work for synthesized dataset.

By all these literatures we get to know that there are various methods to detect credit card frauds but, in that machine learning algorithms are first in the race.

III. Experiment Methodology

A. Data Description

We have considered three sub datasets and one main dataset and applied algorithms like Isolation factor, Local outlier, Support Vector Machine.

And recorded the output values considering their Accuracies, to compare all three algorithms and their performance. Let us consider the sub data sets in the following way with varying number of fraud transactions and valid transactions.

The dataset we have considered is highly unbalanced .

Since we have considered the real time dataset from kaggle website and these are transactions of two day time limit. Since the variable factors considered for dataset are highly confidential, we do not reveal those variables and apply algorithms for the dataset containing millions of instances.

A detailed flow of the algorithms application is described with a flow chart as shown below.

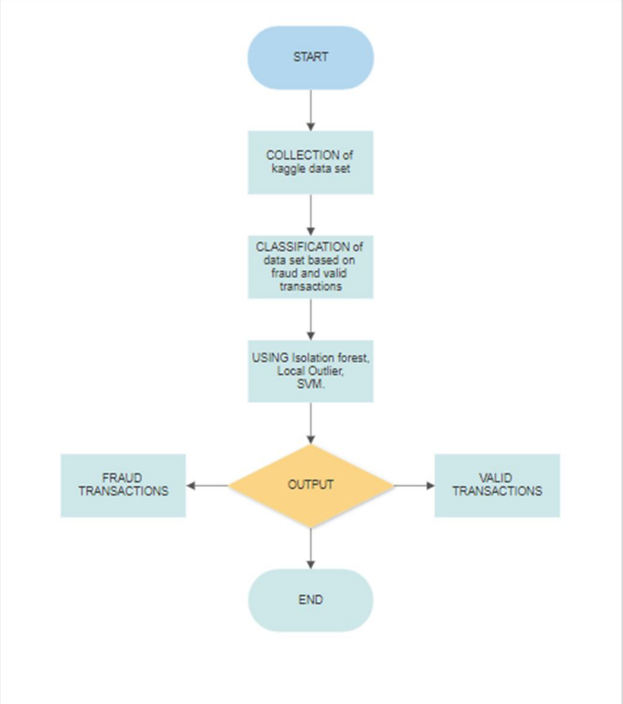
START
COLLECTION of kaggle data set
CLASSIFICATION of data set based on fraud and valid transactions
USING Isolation forest, Local Outlier, SVM.
OUTPUT
FRAUD TRANSACTIONS
VALID TRANSACTIONS
END

Figure 1. Flow Chart

B. Data Preparation

INPUT TAKEN FOR DATASET1:
Number of instances :28,481
Number of Fraud Transactions : 49
Number of Valid Transactions :28,232
INPUT TAKEN FOR DATASET 2:
Number of instances: 10,605
Number of Fraud Transactions: 9751
Number of Valid Transactions: 854
INPUT TAKEN FOR DATASET 3:
Number of Instances: 52,293
Number of Fraud Transactions: 45 184
Number of Valid Transactions: 7109

IV. RESULT ANALYSIS

Accuracy – The term Accuracy is the important measure for performance of an algorithm model. It is defined as the ratio of correctly predicted observation to total number of observations. Accuracy is the best measurer of our model but it's the case only when we have datasets of symmetrical values where values of false positives and false negatives are almost same. Therefore, we have to look at other parameters to evaluate the performance of our model.
Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.
Recall = TP/TP+FN

F1 score – It is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Although this is not as easy as accuracy to understand, This F1 is more useful than accuracy , especially if you have an uneven class distribution as in our case. Since the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Dataset1:
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :

	Precision	**recall**	**F1-score**	**support**
0	1.00	1.00	1.00	28432
1	0.26	0.27	0.26	49
Avg/total	1.00	1.00	1.00	28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :

	Precision	**recall**	**F1-score**	**support**
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49
Avg/total	1.00	1.00	1.00	28481

Support Vector Machine: 8516
Accuracy Score :
0.70099364488606442222
Classification Report :

	Precision	**recall**	**F1-score**	**support**
0	1.00	0.70	0.82	28432
1	0.00	0.37	0.00	49
Avg/total	1.00	0.70	0.82	28481

V. OBSERVATIONS FOR DATASET1

 Isolation Forest detected 73 errors versus Local Outlier Factor detecting 97 errors versus SVM detecting 8516 errors. Isolation Forest has a 99.74% more accurate than LOF of 99.65% and SVM of 70.09.
 When comparing error precision and recall for 3 models, the Isolation Forest performed much better than the LOF as we can see that the detection of fraud cases around 27% versus LOF detection rate of just 2% and SVM of 0%.
 So overall Isolation Forest Method performed much better in determining the fraud cases which is around 30%.
 We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense we can also complex anomaly detection models to get better accuracy in determining more fraudulent cases.
Dataset2:
Isolation Forest: 1823
Accuracy Score :
0.8280999528524281
Classification Report :

	Precision	**recall**	**F1-score**	**support**
0	0.03	0.03	0.03	854
1	0.91	0.90	0.91	9751
Avg/total	0.84	0.83	0.84	10605

Local Outlier Factor: 9840
Accuracy Score :
0.07213578500707214
Classification Report :

	Precision	**recall**	**F1-score**	**support**
0	0.07	0.90	0.13	854
1	0.00	0.00	0.00	9751
Avg/total	0.01	0.07	0.01	10605

Support Vector Machine: 1525
Accuracy Score :
0.8561999057048562
Classification Report :

	Precision	recall	F1-score	support
0	0.14	0.15	0.14	854
1	0.92	0.92	0.92	9751
Avg/total	0.86	0.86	0.86	10605

VI. OBSERVATION FOR DATASET2

□ Isolation Forest detected more than 1k errors versus Local Outlier Factor detecting 9k errors versus SVM detecting 1525 errors.
□ Isolation Forest has a 82.8% more accurate than LOF of 7.13% and SVM of 85.61%.
□ When comparing error precision and recall for 3 models, the SVM performed much better than the LOF.
□ So overall SVM method performed much better in determining the fraud cases which is around 77%.
□ We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense we can also complex anomaly detection models to get better accuracy in determining more fraudulent cases.

Dataset3:

Isolation Forest: 15145
Accuracy Score :
0.7103818866769931
Classification Report :

	Precision	recall	F1-score	support
0	0.01	0.02	0.02	7109
1	0.84	0.82	0.83	45184
Avg/total	0.73	0.71	0.72	52293

Local Outlier Factor: 46540
Accuracy Score :
0.11001472472415046
Classification Report :

	Precision	recall	F1-score	support
0	0.11	0.81	0.20	7109
1	0.00	0.00	0.00	45184
Avg/total	0.02	0.11	0.03	52293

Support Vector Machine: 27012
Accuracy Score :
0.4834490275945155
Classification Report :

	Precision	recall	F1-score	support
0	0.14	0.56	0.23	7109
1	0.87	0.47	0.61	45184
Avg/total	0.77	0.48	0.56	52293

VII. OBSERVATION FOR DATASET3

□ We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense we can also complex anomaly detection models to get better accuracy in determining more fraudulent cases.

C. Graphs and observations

DATASET1:

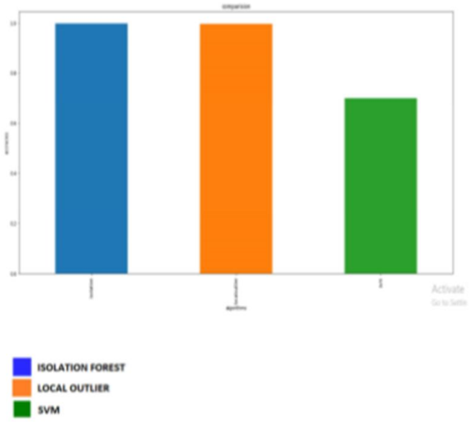

Figure 2. Comparision between three algorithms for dataset1

DATASET2:

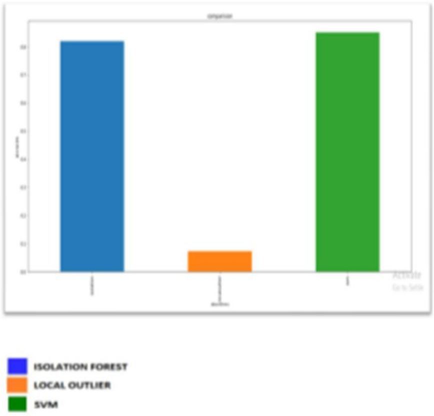

Figure 3 .Comparision between three algorithms for dataset2

DATASET3:

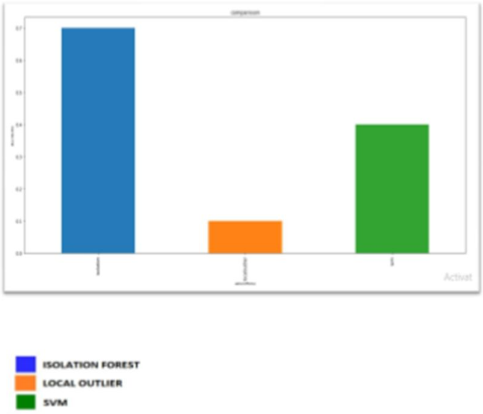

Figure 4. Comparision between three algorithms for dataset3

VIII. CONCLUSION

As we know that credit card fraud occurs when someone uses credit card or credit account to make a purchase without the authorization of owner. Thus the fraud activity can occur in various situations like when we lose our credit card or when someone steals it, or by fake calling customers and asking their card details etc., This information obtained can be used to make purchase or other transactions, either in person or by using online applications which causes error in customers money flow. Thus avoiding these frauds in bank transaction have become vital.

Many strategies have been encountered for credit card fraud detection which includes analysing and modelling past credit transactions with knowledge of the ones that turned out to be fraud. By using this model we can identify whether a new transaction is fraudulent or not.

This paper gives the insight about three machine learning algorithms that is, Support vector machines, Local Outliers and Isolation forest for credit card fraud detection. These three algorithms are applied to three different datasets with increasing fraud rates and their accuracies have been recorded via graphs and charts.

The Support vector machine algorithm is giving the best result when applied to the dataset with low fraud rates. And it is giving moderate accuracy when fraud rates are increased but the accuracy considerably decreases for large datasets. Local Outlier algorithm gives a great accuracy when applied to dataset with minimum fraud rates but fails miserably when fraud rates are increased. The Isolation algorithm gives the satisfiable accuracy for all the dataset versions. Thus it is considered as the accurate algorithm for credit card fraud detection.

By using this record we can further extend to accurately identifying the credit card frauds for differential datasets. And it can be incorporated as a feature in banking application, so as soon as the fraud occurs the user get notified and report it to the bank. Thus, this ensures error free bank transaction and achieve maximum customer satisfaction. And also eases, and helps in the further growth of technological advancements.

REFERENCES

[1] "Credit card fraud - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Credit_card_fraud.

[2] "Machine learning - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning.

[3] J. C. R. S. P. T. N. K. S. Megasari Gusandra Saragih, "Machine Learning Methods for Analysis Fraud Credit Card Transaction," International Journal of Engineering, vol. 8, no. 6S, pp. 870-874, 2019.

[4] C. B. S. P. S. M. N. K. Anuruddha Thennakoon, "Real-time Credit Card Fraud Detection Using Machine Learning," in 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Colombo, Sri Lanka, 2019.

[5] L. Y. Apapan Pumsirirat, "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 1, no. 9, pp. 18-25, 2018.

[6] S. Y. S. Navanshu Khare, "Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models," International Journal of Pure and Applied Mathematics, vol. 118, no. 20, pp. 825-838, 2018.

[7] A. S. S. D. S. S P Maniraj, "Credit Card Fraud Detection using Machine Learning and Data Science," International Journal of Engineering Research & Technology (IJERT), vol. 8, no. 09, pp. 110-115, 2019.

[8] D. K. M. A. S. V. Sahayasakila.V, "Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm," International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 5, pp. 190-192, 2019.