Credit Card Fraud Detection Using Isolation Forest

Prof. Vikas Palekar[1], Samruddhi Kharade[2], Hrushikesh Zade[3], Sajjad Ali[4], Kalyani Kamble[5], Shriviraj Ambatkar[6]

[1]Assistant Professor, Dept. of Computer Science & Engineering, Datta Meghe Institute of Engineering Technology & Research, Wardha, Maharashtra, India
[2,3,4,5,6]Student, Dept. of Computer Science & Engineering, Datta Meghe Institute of Engineering Technology & Research, Wardha, Maharashtra, India

---***---

Abstract -*With the evolution of recent technology, employment of credit cards has increased. As credit card becomes the trendiest kind of payment for online payment moreover as manual payment, the numbers of Credit card frauds are increasing day by day. It's necessary to curb the credit card frauds because it causes brobdingnagian quantity of economic loss. So, we tend to find the dishonorable and real transactions by victimization the conception of sophistication. And additionally the accuracy of Isolation Forest and Local Outlier issue is calculated to point out the most effective rule for fraud detection. Machine Learning consists of many algorithms which will be employed in fraud detection like Random Forest, Local Outlier Factor, Isolation Forest, Naive mathematician, K-nearest Neighbors, Neural Networks, etc which will be employed in fraud detection. And that we are about to use Isolation Forest and Local Outlier in our project.*

Keywords: Credit card, Fraud Detection, Random forest, Local Outlier Factor, Financial.

1. INTRODUCTION

One of the method by that credit card fraud is feasible is by obtaining access to the purloined credit cards and other is by exploiting or to steal the main points of the card via online transaction. Whereas detecting this type of fraud we tend to face several challenges. Millions & billions of transactions occur per minute everywhere the planet.

Detecting that among all the transactions is occurred and which one is real could be a task. The number of analysis disbursed during this field to find the fraudulent transactions is low because of the shortage of availableness of datasets. Because the details of the users are confidential, acting on the important data sets becomes not possible.

Machine learning algorithms are loosely divided into supervised and unsupervised algorithms. In supervised algorithms, a preset set of data is provided for training the system. The system tries to predict the results supported the already offered results i.e., training dataset. Whereas just in case of unsupervised algorithms the system tries to search out the patterns directly from the instance provided.

The random forest algorithm is a supervised classification formula. It's used for classification because it provides correct results than the other classification algorithmic rule.

Local Outlier factor is an anomaly detection algorithmic rule. The outlier is merely a word for anomalies. Anomaly represents the abnormal behavior or a deviation from the traditional behavior of an information points concerning certain attributes. Outliers have a unique statistical property.

The Isolation Forest 'isolates' observations by indiscriminately choosing a feature and then every which way selecting a split value between the maximum and minimum values of the chosen feature. Since recursive partitioning is portrayed by a tree structure, the amount of splitting needed to isolate a sample is equivalent to the path length from the root node to the terminating node.

Applications of those algorithms are speech recognition, banking sector, health care, pattern recognition, etc. At the tip of the project, two algorithms Local Outlier Factor and Isolation Forest is compared to ascertain which one provides the simplest result for fraud detection.

1.1 RELATED WORK

According to Koufakou E.G. Ortiz, M. Georgiopoulos, K.M. Reynolds, the earliest approaches to observe outliers, applied mathematics model primarily based strategies, assume that a constant quantity model describes the distribution of the information (e.g., traditional distribution), and are principally single-dimensional or univariate. The constraints of those approaches embrace the problem to search out the proper model for every dataset and application, still because of the undeniable fact that their potency declines because the information spatiality will increase. Another issue with high dimensional datasets is that the dataset abate dense, which makes the lentiform hull more durable to work out ("Curse of Dimensionality"). There are unit strategies to assist alleviate this drawback, e.g. Principal part Analysis. Another plan to handle higher spatiality information is to arrange the information points in layers, supported the thought that shallow layers tend to contain outliers a lot of typically than the deep layers, these ideas but area unit impractical for quite a pair of dimensions. Distance-based approaches don't create assumptions for the distribution of the information since they cypher distances among points. These approaches become impractical for giant datasets (e.g. nearest neighbor technique has quadratic complexness concerning the dataset size). There are enhancements on the first distance-based algorithms, e.g. Knorr et al., wherever associate outlier is outlined as associate object O in a very dataset T that has a minimum of filled with the objects in T any than distance D from it. The complexness of their approach continues to be exponential within the variety of nearest neighbours. Bay and Schwabacher propose a distance-based technique and claim its complexness is on the point of linear in observe. [1]

According to V. Ceronmani Sharmila, Kiran Kumar, the business is stuffed with practices that area unit dishonest in nature. The core objective is to primarily observe the assorted credit card frauds and so, check on the various algorithms associated create a knowing call. The target is to indicate and analyze recently revealed outcomes in credit card fraud detection. This shows the common terms in credit card fraud and shows key statistics and figures during this field. Supported the versions of fraud incurred by the establishments, several precautions area unit taken and place into work. The ideas given during this paper motivate tons of potency. The flexibility of the assorted methodologies taken a glance here within the limitation of the credit card. Though there are several problems once a true credit card customers area unit wrong framed as anomalies. The problem with making enterprise from the online have it in such how that each the card and also the holder needn't be a gift within the premises. It's therefore terribly laborious for the business person to search out whether or not the client is that the true owner or not. Payment card fraud is currently a significant issue across the world. Enterprises and firms shed large prices yearly owing to pretend and seamsters endlessly finding alternate ways to try to unethical activities. The half to appear forward is that the undeniable fact that these varieties of malicious activities have sure forms of formats that they follow and area unit somewhat easier to search out its root path and any details concerning it. During this article we tend to commit to checking malicious group action through the rules still like the genetic algorithm. As we are going to see that artificial neural network. computing will be programmed and trained in such how that they'll imitate a true brain, but it's nearly not possible for AI to succeed in the amount of subtleness and detail because the human brain, it's almost like the unit in our brain that creates core choices. [2]

According to Pawan Kumar Fahad Iqbal, this paper represents a groundwork on European credit card holders, wherever information standardization is completed before doing Cluster Analysis of the dataset. The information is MLP trained and Machine learning algorithms area unit used for generating correct results. Promising results area unit obtained by mistreatment normalized information. Each supervised and unsupervised learning are utilized in this paper. This paper aimed to search out new strategies for fraud detection which will increase the accuracy of results. [3]

According to Ms. Amruta D. Pawar, Prof. Prakash N. Kalavadekar, Y Kou, C. Lu, S. Sinvongwattana and Y. Huang has bestowed a survey on numerous fraud detection techniques like credit card fraud detection, telecommunication fraud detection, and intrusion detection system. for every of this class completely different detection techniques area unit mentioned by them.[4]

M. Breunig, H-P Kriegel, R. T. Ng, and J. electric sander planned density-based Local Outlier Factor (LOF) technique to search out outliers. They allotted every object a degree of being the associate outlier. This degree is named as Local Outlier Factor (LOF) of the associate object. The degree depends on however the object is isolated from close objects. All previous strategies contemplate outliers as binary property however this technique assigns the degree of being the associate outlier. But this technique doesn't work on a high dimensional dataset and additionally needs high computation. [5]

1.2.PROBLEM STATEMENT

- Credit card fraud detection becomes difficult because of 2 major reasons –first the profile of traditional and dishonorable behaviors modification perpetually and second, credit card information sets are extremely skewed.
- Due to digitalization of all the services like banking, selling etc. The utilization of any quite payment cards for transaction could be a terribly traditional method for each individual. Are it becomes terribly difficult to work out the real and fallacious transactions.

1.3. OBJECTIVES

- To preprocess dataset and split it into training and testing phase.
- Perform implementation of Isolation forest algorithm.
- Perform comparison with local outlier algorithm for accuracy analysis.
- To design a classifier to classify fraud and real transactions to maximize accuracy level.

2. METHODOLOGY

The most vital role has been contended by the dataset. Because it is not possible to figure on a real-time database, we tend to are using the creditcard.csv dataset from Kaggle which incorporates 2,83,807 entries. The various parameters employed in the database are time, class, amount, location &, etc. Such kind of total of thirty-one parameters is used. V1 - V28 are the fields of a PCA dimensionality reduction to safeguard the identity and sensitive options of a user.

In this project, we are aiming towards determination of transactions with a high likelihood to bear credit card fraud. We'll build and use the subsequent two machine learning algorithms:

- Local Outlier Factor (LOF)

  The anomaly score of every sample is named the local Outlier factor. It calculates the local deviation of the density of a given sample in relation to its neighbors. It's known as local because the anomaly score depends on the object isolated the thing is with relation to the encircling neighborhood.

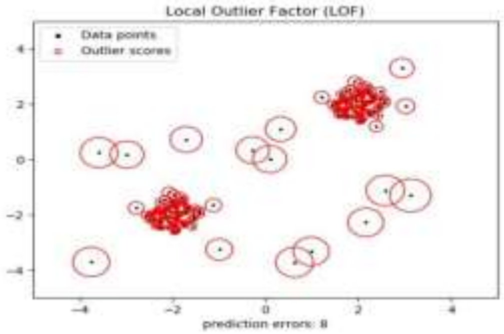


Fig 2.1: Local Outlier Factor

- Isolation Forest algorithm

  The Isolation Forest isolates the observations by indiscriminately choosing a feature by randomly selecting a value and splitting it in between the atmost maximum and minimum values of the chosen feature. Also the recursive partitioning method is denoted by a tree-like structure, the quantity of splitting required to isolate a sample is comparable to the trail length from the root node to the terminating node.

3. MODELING & ANALYSIS

3.1. REQUIREMENTS

Software requirements**:**

- Application Used: Python (jupyter notebook) version-3.6.5
- Operating System: Windows 10(x86)

Data set Requirements:

- Credit card dataset (.csv file containing 2,84,807 records)

3.2. PHASES OF PROJECT

We are completing this fraudulent transactions detection activity in following three phases,

1) Data Exploration

Steps: a) Load dataset
b) Preprocess dataset
c) Perform graphing
d) Display dataset

2) Data Preprocessing

Steps: a) Load dataset
b) Remove Null values
c) Split dataset
d) Move to training phase

3) Data Classification

Steps: a) Train the dataset
b) Develop classifier
c) Isolation Forest
d) Perform Classification

The first phase involves loading the dataset also known as the Data exploration phase. Data exploration is the process just like data analysis, we have used visual exploration to understand what is there in the dataset along with its characteristic. We have used the data set from the Kaggle website, it contains various parameters such as amount, class, time and others are reduced using the PCA dimensionality reduction process. The dataset is explored and represented to generates descriptive statistics that summarize the central tendency, dispersion, and form of a dataset's distribution for the given series object. All the calculations are performed by excluding Null values. By victimization this represented knowledge, the histogram is generated to show it.

The second phase involves Data Preprocessing. It again loads the dataset and removes all the null values and garbage values from the dataset to improve its efficiency. In this phase itself, we have to split the dataset into training and testing phases. Here we mostly work on the training phase by describing class 0 as genuine transaction and class 1 as a fraudulent one. In training the dataset, fraudulent and genuine entries are provided randomly to amplify the quality and thus more realistic data is generated. A correlation matrix is provided to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analysis.

Third and the last phase is Data classification. It is simply a task of inputting training data set for which the classes are pre-labeled for the algorithm to learn from. The model is then used by inputting a different dataset for which the classes are not defined and then the model predicts the class to which it resembles using the learning from the training set. Both the algorithm has to be applied to get a productive result determining the outcome using terms as precision, recall, f1-score, and support.

4. RESULTS & DISSCUSSIUONS

4.1. CORRELATION MATRIX

The correlation matrix is a heat map which is used to check whether there is any correlation between different parameters and different variables in our dataset.

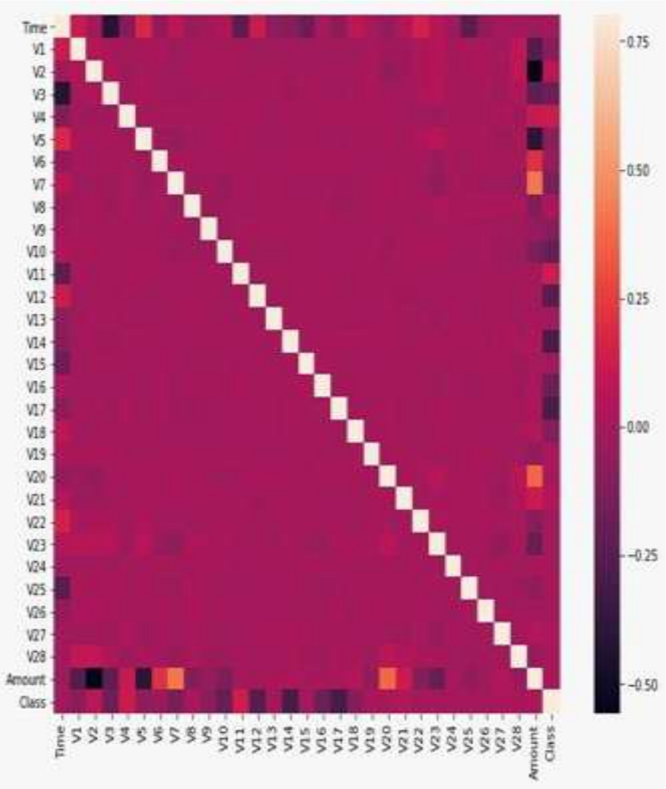


Fig -4.1.: Heat Map

The above figure is a result of pyplot, it uses seaborn and SNS heatmap. It gives our simple correlation matrix a visual look and also makes it easier for analysis purpose. All the 31 parameters V1-V29, class and amount is present at both X-Y axes with a range from -0.75 to +0.50.

4.2. HISTOGRAMS

The histograms are used in the project to easily analyze the fraudulent and real transactions. We can use the matplotlib for this purpose. We can also change the size of the plot accordingly.

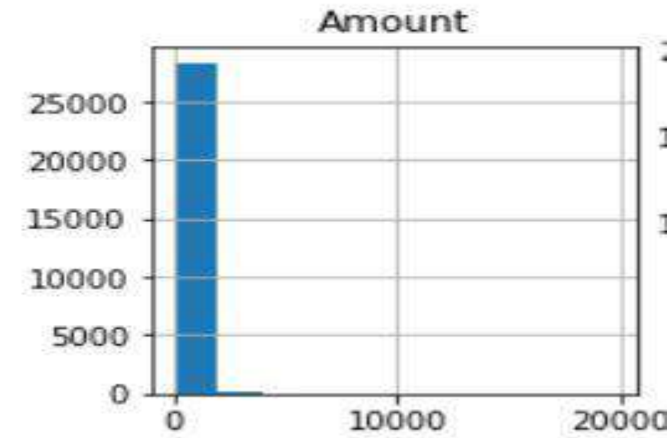


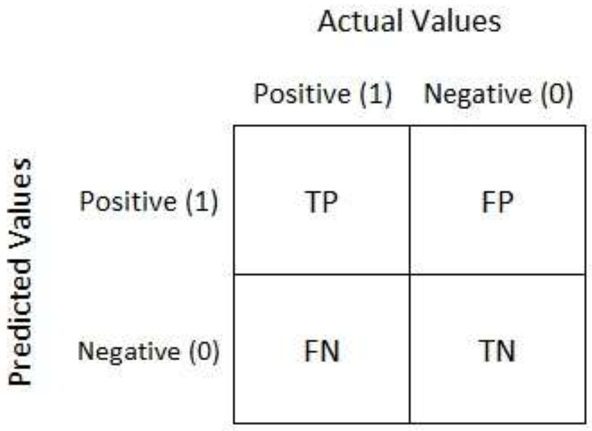
Fig -4.2.: Bar graph for amount

The above histogram shows the average amount transacted by all the customers and the number of transactions that are fraudulent and the number of transactions that are genuine. If the transaction is at 0 means it is a genuine one and if it is at 1 it is a fraudulent transaction. The graph shows 98% of the transactions as genuine and 2% is considered as fraudulent.

In this way we can analyze all the 31 parameters using simple histograms.

4.3. PERFORMANCE METRICS

There are numerous metrics to evaluate the performance of varied machine learning algorithms. However, we've got used solely major four metrics.



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

Fig -4.3.: Confusion metrics

The confusion matrix is employed for calculating these performance metrics. The terms utilized in the confusion matrix square measure as follows:

Tp: true positive values

Tn: true negative values

Fp: false-positive values

Fn: false negative values

- **Precision**:

  It is the quantitative relation of Tp/(Tp+Fp). The best value is one and the worst value is zero.

- **Recall**:

  It is the quantitative relation of Tp/(Tp+Fn). It finds all the possible positive samples.

- **f1_score**:

  It is a weighted average of the precision and recall.

- **Support**:

  It is the number of true values in every number of target values.

5. CONCLUSIONS

At present credit card fraud detection is one of the major issue in current transaction that are made online. To avoid such situation it is mandatory to design a classifier that can classify which transaction are fraud and which are real ones.

We are going to divide the dataset into training and testing phase for implementing further processes. We have used K-means algorithm which is an unsupervised in nature for clustering. We need a dataset which have both the entries fraudulent and genuine as we require this for training phase.

Thus by K-Means clustering and machine learning algorithms (Isolation forest and local outlier factor) which can be best adapt to the change in scenario taking place can be used and developed on a very large scale to detect the fraudulent transactions and used to ensure the credibility of payment system.

REFERENCES

[1] Koufakou1 E.G. Ortiz1 M. Georgiopoulos1 K.M. Reynolds, "A Scalable and Efficient Outlier Detection Strategy for Categorical Data" 1st International Conference on Innovations in Information and Communication Technology(ICIICT), 2019. pp.420-690

[2] V. Ceronmani Sharmila Kiran Kumar R Sundaram R, Samyuktha, "Credit Card Fraud Detection Using Anomaly Techniques" pp.398-410.

[3] Pawan Kumar Fahad Iqbal, "Credit Card Fraud Identification Using Machine Learning Approaches" 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT).

[4] Ms. Amruta D. Pawar1, Prof. Prakash N. Kalavadekar, Ms. Swapnali N. Tambe, "A Survey on Outlier Detection Techniques for Credit Card Fraud Detection. p- ISSN: 2278 8727 Volume 16, Issue 2, Ver. VI (Mar-Apr. 2014),IOSR Journal of Computer Engineering (IOSR-JCE).

[5] Y. Sahin and E. Duman, "Distinguishing charge card misrepresentation by choice trees and bolster vector machines", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2011 Vol I, IMECS 2011, March 2011.

[6] Hawkins, S., He, H., Williams, G., and Baxter, R., "Outlier Detection Using Replicator Neural Networks", Proc. of the Fifth Int'l Conference Data Warehousing and Knowledge Discovery, pp.170-180, 2002.

[7] Knorr, E., Ng, R., and Tucakov, V., "Distance-basedoutliers: Algorithms and applications", VLDB Journal, 2000.

[8] Bolton, R.J., Hand, D.J., "Statistical fraud detection: A review", Statistical Science, 17, pp. 235–255, 2002.

[9] Penny, K.I., Jolliffe, I.T., "A comparison of multivariate outlier detection methods for clinical laboratory safety data",The Statistician, Journal of the Royal Statistical Society, 50, pp. 295–308, 2001.