Credit Card fraud detection

[1]Ankit Patel, [2]Yash Nanda, [3]Stevina Correia
[1]Student, [2]Student, [3]Faculty
[1]Information Technology,
[1]Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.

Abstract: The constant growth of different e-commerce platforms has contributed to an increase in the use of credit cards for payment purposes. An increase in credit card usage makes it vulnerable to hackers intending to exploit the system. As it becomes a popular method for payment purposes, a surge in fraud has been observed. Using the information captured from the transaction, numerous things can be deduced which might have detrimental effects. The current scenario requires efficient techniques for the detection of frauds which might reduce the cases of it which will directly impact the loss caused by it. Different techniques have been implemented for fraud detection and some of them show promising results in the detection of fraud. This paper shows the comparison between the accuracy of some supervised learning classification algorithms which are used for identifying fraudulent transactions.

I. INTRODUCTION

In the future, we intend to build a society where there is minimal cash usage. If facts are taken into consideration, World Payment reports state that today, non-cash transactions have increased by 27% globally and are expected to almost double in 2022 as compared to its usage in 2017. The reason might be attributed to the shifting of user preferences, increased e-commerce usage, and governmental support [1]. While it aligns with our idea of a cashless society, the statistics say that fraudulent transactions are also increasing. EMV (Europay, MasterCard, Visa) smart chips is a globally-adopted standard for chip-based debit-card and credit-card transactions to ensure security against fraudulent transactions. But even it has not been sufficient to counter this problem. An approximate of 25 billion dollars was lost due to credit-card fraud worldwide in 2018 [3].

Figure 1: Worldwide non-cash transactions [1]

As seen in the figure 1, x-axis represents Year and y-axis represents the amount of credit cards in millions. It illustrates the continuous rise in the usage of credit cards over the years.

So, let's start with the basics. Credit card fraud can be defined as an identity theft that occurs when someone intentionally uses your card to conduct a transaction without your knowledge about it [3]. It can be performed in numerous ways which include stealing a card, using a misplaced card, using duplicate cards which have been mitigated by EMV, using someone's mailbox to intercept the card, using your information to issue a new card, card-not-present fraud which is having the card number but not the physical card [3]. Usage of analytics to detect, collect data and act on the fraud based on the patterns in data is known as monitoring for credit card fraud. The Association of Certified Fraud Examiners (ACFE) recommends consistent analysis and data-monitoring as effective ways to control fraud [2].

This paper focuses on using anomaly detection technique to find patterns which do not match with the expected behaviour, also known as outliers. It has many applications which include identifying patterns in a network that signals a hack, finding a tumour in an MRI scan, fraud detection in credit card transactions, etc. In the context of data mining, anomaly detection can be defined as identifying rare observations that differ from most of the data. For example, if all the values of a parameter 'X' range from a numerical value 1 to 10 and there exists a value 2000, it is an outlier as the value expected for that particular parameter is between 1 and 10. They can be classified as:

1. Global Outliers/ Point anomalies
   If the value is outside the range of the entire dataset, it is considered a global outlier [4]. To give a real-world example, if an individual who does not deposit more than Rs.10000 in his/her account per month, deposits an amount of 1 Lakh in a month, twice is considered a global outlier as it has never occurred in the customer's deposit history. If the time-series data is analysed, it will show a sudden rise which will raise questions against it, the reasons which could be attributed to laundering or fraud etc.
2. Contextual Outliers

If the value is different from the rest of the data in the same context, it is known as a Contextual Outlier [4]. However, it should be noted that the same value might not be an outlier in a different context. For instance, it is normal to assume that during Black Friday, the sales boost up. But, if the sales remain the same or go down, it will be considered as an outlier. But the sales remaining constant might not be an outlier in a different context, say a regular weekday.

3. Collective Outliers

   A set of values that differ from the entire dataset are known as Collective outliers [4]. However, it should be noted that individual values might not be anomalous but as a group, they differ. For instance, a decrease in sales of a particular product might not indicate an anomaly, but if it is found that there is a relation between the drop in sales for 5 products, it highlights a bigger issue which when considered together, might be an anomaly.

II. DATASET

In this paper, the dataset used was published on Kaggle. The dataset contains transactions made by credit cards by European users in 2013. The data presents transactions that occurred on two days of September 2013. The dataset contains 492 frauds out of 284,809 transactions. The dataset suffers from class imbalance as the positive class(frauds) account for 0.172% of the entire dataset.
Due to confidentiality issues, the dataset has 28 of 31 feature points converted into numeric value which are the result of PCA transformation [5]. The remaining three feature points are 'Time', 'Amount' and class of transaction (fraud or not). The feature time contains the seconds passed between each transaction and the first transaction. The Amount feature describes the amount of transaction. The Class feature describes the category of transaction (1 for fraud and 0 otherwise).

III. METHODOLOGY

In this section, we are going to discuss the various classification techniques that can be used for fraud detection. The techniques mainly use either of the two approaches i.e supervised learning and unsupervised learning. In supervised learning, the function or the technique creates a model based on the train and test data points that we provide. The train data points are the ones that have already been categorized, thus helps in creating the model. Using this model, it maps an input data point to one of the categories. Whereas in Unsupervised learning, the model works on its own to discover and collect information. It mainly works with unlabelled data. Here we are going to use the supervised learning approach as we can divide the dataset into training and testing data which helps to create a model. Also, the dataset is categorized into the category whether it is a fraudulent transaction or not. We are going to use the Isolation Forest, Local Outlier Factor and Support vector machine algorithms of the Supervised learning approach to detect fraud.

Isolation Forest Algorithm

Isolation Forest algorithm is different from the other outlier detection methods as it profiles the outliers instead of the valid data points. The splitting of the data point is based on the time taken to split the point. For example, if a point is not an outlier, it will have many data points around which will make it difficult to isolate. On the other hand, if it is an outlier, it will be far away from the normal data point, thus the time taken to discover it will be higher using which it can be declared as an outlier. This algorithm can work with huge datasets and multiple dimensions which is an advantage. It takes advantage of the fact that outliers are less and different from the normal data points [4].

Local Outlier Factor Algorithm

It is an approach based on the density of a point for which it relies on its k-nearest neighbours. By calculating the ratio of the average density of neighbours to the density of point, a numerical value is assigned to the data point done by the LOF method [6]. To compute LOF, the following steps are followed:

1. Distance between the two observation pairs is calculated
2. kth nearest neighbour is found which is used for calculating the distance between it and the observation
3. Calculate Local Reachability Distance (LRD)
4. Calculate Local Outlier Factor (LOF)

LOF value less than 1 indicates a valid data-point while LOF value greater than 1 indicates an outlier

Support Vector Machine Algorithm

It is a supervised learning model that is used for both classification and regression techniques. It is commonly abbreviated as SVM. SVM uses Hyperplanes and Support Vectors for analyzing the data. A hyperplane is a decision boundary that classifies data points. Hyperplane depends upon the number of features used to discriminate the classes. For example, if there are 2 features then the hyperplane is a line and if there are 3 feature points then the hyperplane is a 2-dimensional plane. The support vector is the data points that influence the orientation and position of the data points. Using support vectors, we can maximize the margin of the classifier. For a given dataset, first, we divide it into training and testing data. Using the training data points, where each belongs to one or the other of the two categories, it tries to define a hyperplane between those points. Using this hyperplane, the SVM model assigns new examples to one of the categories. The figure 2 shows the working of SVM model. As seen in the figure the Hyperplane A and B divide the class red and blue.

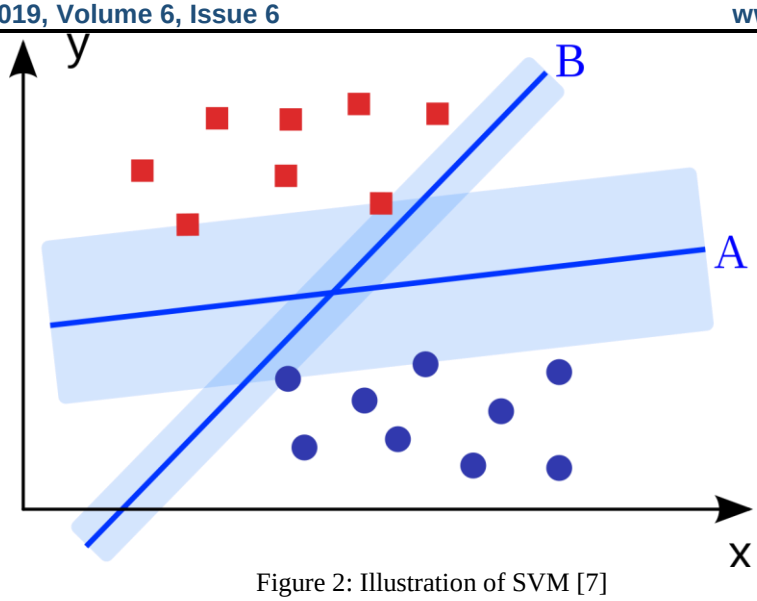

Figure 2: Illustration of SVM [7]

IV. RESULTS

Technique	Errors Detected	Accuracy	Precision	Recalls
Isolation forest	73	99.62%	27%	26%
Local Outlier Factor	98	99.48%	3%	3%
Support Vector Machine	8512	70.04%	0.02%	37%

Table 1: Comparison of Results Between Various Classifiers

Analysis of Result:

- Since the Isolation forest algorithm is sensitive to global outliers, it gives us an accuracy of 99.62% which is greater than the Local Outlier Factor.
- The Local Outlier Factor model gives the correct result to only data points that are closer to the outlier cluster. The data points away from the cluster are not classified correctly. Hence the no. of errors increased.
- The Support Vector Machine model gave a large number of errors, as the hyperplane positioned in the dataset did not correctly classify the data points due to the lack of a number of feature points.

V. CONCLUSION

The Isolation Forest algorithm provides us with the most optimal results, as it overcomes the errors of the Local outlier Factor algorithm which is not sensitive to global outliers and the support vector machine is not able to identify the transactions correctly. Hence using the Isolation Forest algorithm, the detection of fraudulent credit card transactions can be easily detected.

REFERENCES

1. https://worldpaymentsreport.com/
2. Report to the Nations on Occupational Fraud and Abuse: 2014 Global Fraud Study, 2014
3. https://shiftprocessing.com/
4. https://towardsdatascience.com/
5. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
6. V. Ceronmani Sharmila, K. K. R., S. R., S. D. and H. R., "Credit Card Fraud Detection Using Anomaly Techniques," *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, CHENNAI, India, 2019, pp. 1-6.
7. https://binged.it/37LhJCO