Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

doi:10.3233/APC210174

A Relative Investigation of Various Algorithms for Online Financial Fraud Detection Techniques

C Pallavi [a, 1], Girija R [b], Vedhapriyavadhana R [b], Barnali Dey [c], Rajiv Vincent [b]
[a]*School of Electronics and Communication, VIT University, Chennai Campus, India*
[b]*School of Computer Science and Engineering, VIT University, Chennai Campus, India*
[c]*Department of Information Technology, SMIT, SMU, Sikkim, India*

Abstract. Online financial transactions play a crucial role in today's economy. It becomes an unavoidable part of the business and global activities. Transaction fraud executes thoughtful intimidations to e-commerce spending. Now-a-days, the online contract or business is fetching additional sound by knowing the types of online transaction frauds associated with, these are raising which disturbs the currency accompanying business. It has the capability to confine and encumber the contract accomplished by the intruder from an honest consumer's credit card information. In order to avoid such a problem, the proposed system is established transaction limit for the customers. Efficient data is only considered for detecting fraudulent user action and it happens only at the time of registration. Transaction which is happening for any individual is not at all known to any FDS (Fraud Detection System) consecutively at the bank which mainly issues credit cards to customers. To speak out this problem, BLA (Behaviour and Location Analysis) is executed. The FDS tracks at a credit card provided by bank. All the inbound business is directed to the FDS aimed at confirmation, authentication and verification. FDS catches the card particulars and matter to confirm that the operation is fake or genuine. The pick-up merchandises are unknown to Fraud Detection System. If the transaction is assumed to be fraud, then the corresponding bank declines it. In order to verify the individuality, uniqueness or originality, it uses spending patterns and geographical area. In case, if any suspicious pattern is identified or detected, the FDS system needs verification. The information which is already registered by the user, the system identifies infrequent outlines in the disbursement method. After three invalid attempts, the system will hinder the user. In this proposed system, most of the algorithms are checked and investigated for online financial fraud detection techniques.

Keywords. Online fraud, Secure payment, Location analysis, Fraud detection system.

1. Introduction

Online scams charge billions of dollars all over the world. Applying machine learning algorithms to predict the possibility of a transaction being a fraud transaction is one of the efficient methods. In our proposed system we will take credit card transactions, analyze the data, create the features and labels and finally apply one of the ML

[1] Rajiv Vincent, School of Computer Science and Engineering, VIT University, Chennai, India
E-mail: rajiv.vincent@vit.ac.in.

algorithms to judge the nature of the transaction as being a fraud or not. Then check the accuracy, precision, and score of the model chosen. Fraud detection is one of the major priorities for banks and commercial Organizations. This can be lectured with the use of algorithms in machine learning for calculating the accuracy. Fraud detection methods are updating promptly permitted to familiarize in latest deceitful approaches globally. But, the expansion of novel deception detection procedures makes more complicated due to the undecorated restriction of the designs and methods of conversation in fraud detection [1]. The number of fraudulent dealings is frequently found in an actual little percentage when compared to the total transactions. Hence the task of detecting fraudulent connections in a precise and effectual manner is fairly difficult and challenging. Therefore, the development of efficient methods which can discriminate occasional fraud happenings from billions of legitimate transactions seems essential. Credit card fraud happens when an unauthorized person gains access to your credit card information and uses it to make purchases. Some of the most common ways are high-level hacking of the bank account details, through Phishing attacks, when card details are overseen by some other person, lost or stolen cards and fake phone call convincing the target to share the details.

There are certain main challenges which complicate the credit card detection frauds are:

1.1. Development of Fraud designs over period

It is very toughest to report meanwhile the fraudsters are continuously find all the inventive methods to increase admittance to the design to get credit card details. Hence, it becomes very significant for all the machine learning models which is to be restructured with respect to developing arrangements in order to detect suspicious configurations. The report is showing decrease in accuracy and efficiency. Hence, all the machine learning models must be reorganized otherwise; they will fail in their objectives.

1.2. Imbalance

In the fraud detection models, there is a disparity in the classification. It makes the system harder. For genuine clients, the drawback of this challenge is worst experience. Sometimes, imbalance occurs, by weakening the truthful transactions then only fraudsters will get caught.

1.3. Model Interpretations

Model interpretations contest is accompanying with the conception of explainability because all the learning approaches calculates a score for identifying the transaction is fraud or genuine.

1.4. Feature generation are time-consuming

It has been identified and reviewed that specialists might need longer periods of time to produce a perfect feature established which reduces down the fraud recognition process.

1.5. Dissimilar misclassification significance

Various misclassification mistakes have been coming under the category of fraud detection. Misclassification of a reasonable purchase as deception is not as damaging when compared to noticing a fraud transaction as a legitimate one.

1.6. Overlying data

Various transactions might be reflected duplicitous, which may be a false positive, and in reverse, it may be a false negative. Hence, in order to obtain a truncated rate of false positive and false negative and this is the major challenge in fraud detection systems.

1.7. Absence of adaptability

It is very tough for the classification of machine learning algorithms to distinguish innovative styles of legitimate or dishonest patterns.

There are various measures to solve these challenges are:

1.8. Human-in-the-loop

This method solves the imbalance problems as well as decreases the time for detecting the features [2]. It also comprises persons supporting the replicas by providing information to detect novel patterns, features, and many dimensions of fraud.

1.9. Ensemble approach

In order to encounter the continuously emerging fake outlines, Ensemble approach chains numerous representations for a solo mission such as fraud revealing. Collaborative with respect to machine learning can detention fraud patterns to exploit outcomes and increase accurateness.

1.10. Explain ability

The concept of understandable machine learning can deliver explanations for complimentary or deteriorating transactions as fraudulent, therefore resolving the exemplary elucidation experiment. There are techniques such as surrogate modelling, maximum activation analysis, and others that provide all these benefits.

There are countless categories of credit card deceptions. Some of those deceptions are Application frauds, Electronic or Manual Credit Card Trajectories, CNP (Card Not Present), Counterfeit Card Fraud, Lost and Stolen Card Fraud, Card ID Theft, Mail Non-Receipt Card Fraud, Account Takeover and Merchant Collusion. Application Frauds is the impostor improvements contact to the presentation classification by retrieving worker particulars similar password and username and making a bogus explanation by person's particulars. This is called individuality stealing when the swindler relates for a new credit card in the board's designation. The fraudster bargains searching and essential papers in order to maintenance their deceitful presentation [3]. Electronic or Manual Credit Card Trajectories is nothing but once the swindler gains admittance to the info which is positioned on the magnetic stripe of the card which is actual personal and it can be used to acquisition in the future. Once the impostor gets contact to the account numeral and finishing date of the card, that particular card can be used for purchasing without its original presence, and then it is CNP (Card Not Present). Process of skimming and counterfeiting credit card are interrelated. A false magnetic swipe card clamps all the particulars of the unique card and it is a operational card and can be used to purchase in the future. Once the genuine cardholder mislays their card and if it becomes into the hoaxer's hand and it can be used to brand expenditures. It is tough to do this a machine is in need of PIN, but online dealings are relaxed sufficient for the swindler and comes below group of lost and stolen card fraud. Card ID Stealing is the fraudster gets the particulars of the genuine card to use the card or to produce a new version. When a user smears for a card, it proceeds selected time to appearance all the rules and regulations. If a swindler may record the card in their designation and use it to make acquisitions by interrupting in the intermediate of the conveyance. It comes under the never received issue fraud and it arises below the grouping of Mail Non-Receipt Card Fraud. Account Takeover is the furthermost communal method of deception in which an impostor strength improvement admittance to the card particulars of the inventive cardholder and several profound forms. They can even contact the credit card company and act as if they are the original cardholder and might level ask them to adjustment the address [4]. By the way of the fraudsters have all the specifics which they got done hacking or else manipulating the original cardholder, they can deliver them by way of resilient. The replacement card resolve at that moment be directed to the place assumed by the fraudster. False Merchant Sites is like a phishing violence anywhere then operator becomes stuck in a counterfeit webpage, fashioned through the swindler, which appearances actual analogous to a recognized and honest website to influence the user similar a reductions page promising the consumer to acquisition the merchandises. As soon as the compensation is finished, all the evidence is composed and the swindler usages it for upcoming procurements. When the merchant or shopkeeper licenses on the information connected to the user's card deprived of the cardholder knowing is nobody but Merchant Collusion.

2. Methodology

2.1. Platform

For Python coding (Using Anaconda Navigator)-JupyterLab is a user interactive environment used for development purposes. JupyterLab is flexible to implement for our Project and Python Coding as well. The packages included for this comparison are Numpy, Pandas and Seaborn. NumPy adds support for matrices and 3-dimensional arrays with a huge collection of mathematical expressions to work on these arrays. In data-intensive calculation, Numpy provides a range of methods that make data manipulation in Python less complicated. Since Python is slower in execution compared to other languages during looping, Numpy helps speed up the operations by converting repetitive code to the compiled form.Pandas offers operations for operating numerical tables and time series. Seaborn is used for data visualization and gives an interface for creative and knowledgeable statistical graphics. Visualization is the major part of Seaborn which helps in understanding data. Seaborn offers the following functionalities such as Dataset-oriented API to regulate the relationship between variables, Automatic calculation and plotting of linear regression plots. it supports a high-level concept for multi-plot grids, Visualizing univariate and bivariate distribution. Matplotlib- Matplotlib library is used for plotting and getting the right plot is often achieved through trial and error. It gives an API for implanting plots into web apps using tools like Tkinter, Qt,etc and there is a methodological "pylab" interface which depends on a state machine, created closely resembling MATLAB. Pydot- Pydot is a Graphviz interface and it can analyze the DOT language. It has been developed using Python. Sklearn- The library consists of a lot of important tools for machine learning like regression, clustering,etc [5]. Ipython - It is a command shell library for various programming languages. It was originally created for Python that offers shell syntax, introspection, tab completion, and history.

2.2. Comparison Algorithms

2.2.1. Isolation Forest Classifier

Isolation Forest algorithm is a tree-based model used to detect outliers and it is separated with arbitrary breaches than a model consisting in an even class, as outliers are repeated fewer than usual results and must have prices beyond the dataset [6].

Succeeding this concept, this classifier selects a feature and at that point selects a value within the sort of this article as the split significance. By means of the previous stage, the algorithm continuously creates a tree. The quantity of essential arbitrary separations to separate a model is the tree penetration. The separation numeral, be an average of concluded a forest of such unsystematic trees, is an evaluation of ordinariness and our pronouncement purpose to distinguish outliers. Random splitting creates acknowledgeable tinier tree depths for outliers and lengthier tree depths for the break of the data models. Hence, when a forest of random trees creates smaller pathway lengths for a specific document point, this is expected to be an outlier.

2.2.2. Random Forest Classifier

Random forest is an algorithm based on ensemble learning. It is an algorithm in which the predictions are obtained by aligning different models or similar models' multiple times. The random forest algorithm functions in the same way and uses various algorithms i.e., multiple decision trees, concluding in a forest of trees, hence that's how the algorithm is named as "Random Forest" which may be used for both regression and classification purposes. Some of the benefits of this classifier are that this algorithm is not biased and it is based on the group of trees where every single tree is trained separately based on the data, therefore biasedness is lowered, it's a very strong and firm algorithm. Even if a new data point is added to the dataset it doesn't affect the overall algorithm but affects only a single tree. Thus, using this Random forest algorithm and decision trees gives the accurate percentage from the given dataset by studying its behaviour [7].

2.2.3. Decision Tree Classifier

It is the most accurate model used in data, and machine learning. This is mainly based upon decision tree which is used to look at an any item and conclude the item's target value. The Ordering tree is a model anywhere the goal variable can yield a vivid set of values from the tree structures. Regression trees are an example of decision trees where a specific variable can yield uninterrupted values. Decision trees are unique of the best machine learning procedures and are a simple representation of classifying data. It is Supervised Machine Learning where the data is split according to certain parameters [8-10].

2.2.4. Naive Bayes Classifier

Naive Bayes classifiers are founded on relating Bayes theorem through individualistic expectations amongst the structures. They are one of the easiest Bayesian network models. But they could be added with Kernel density and achieve high accuracy levels.Naive Bayes classifiers are very accessible, requiring a particular number of conditions linear in the number of features in a learning problem. Highest training can be done by calculating a closed-form form by considering linear time, rather than by repeated approximation which is used for other types of classifiers as well [11]. It has real-valued attributes estimated by assuming a Gaussian distribution and is easiest to work with, and only need mean and std from training data by calculating mean and std of input values(X) for each class to summarize the distribution.

2.2.5. Logistic Regression

Logistic regression uses a logistic expression to model a multi-variable that is dependent. In regression analysis, logistic regression is used to evaluate the criteria of a logistic model. In Spite of the name being regression, logistic regression is used for clustering for detecting binomial and multinomial results, with the aim of estimating the values of coefficients using the sigmoid function.

2.2.6. K Nearest Neighbours

The k-nearest neighbour algorithm is a non-parametric method created by Thomas Cover for regression and classification of data. The input includes all the k nearest training examples in the set. An object is grouped by the decision of its neighbours, with the object being assigned to the class which is usual among its neighbours. When k = 1, the object is allocated to the group of a nearest neighbour. In k-NN regression, the result is the data for the object.

2.3. Statistical Evaluation tools

2.3.1. Confusion matrix

The confusion matrix is comparatively simple to understand, but the terminology can be confusing. The confusion matrix represents the True Positive values, which means a class of the data matches the predicted class of the data. False Positive represents that the actual class of the data was 1 but the model predicted it to be 0. False Negative represents that the actual value of the class was 0 but it was predicted to be 1. True Negative represents that the actual value was 1 and the predicted value is also 1. Figure 1. Represents the confusion matrix.

	Actual	
Predicted	Positives (0)	Negatives (1)
Positives (0)	TP	FP
Negatives (1)	FN	TN

TP = True Positives, FP = False Positive

FN = False Negative, TN = True Negative

Figure 1. Confusion matrix

Accuracy is a measure for scoring the models. It is the part of results the model got right. It is been calculated by the below equation.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

Precision is the proportion of accurately detected positive observations to the total detected positive observations. The formula for evaluating Precision is Equation 2.

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

The F1 score is calculated using Precision and Recall. They are considered as weighted average for F1 score. Hence, this F1 score includes the false positive and False Negative. Always, accuracy is the foremost parameter, but in our proposed model, F1 score is more important than accuracy. The above case is principally for an uneven dissemination. It is calculated using the equation3.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \qquad (3)$$

Recall score is the one whose ratio is decorously anticipated positive opinions to all explanations in the genuine class. Recall has been calculated by the following equation.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

Matthews Correlation Coefficient is the coefficient considered a valid account TP, TN, FP and TN. Generally, it is regarded as a well-adjusted portion that is in use, even there are classes with different sizes. It yields a value between −1 and +1 since it is a correlation coefficient between the predicted and pragmatic binary classifications. If it is a well and perfect prediction, the coefficient value is +1. In case, if it is a random prediction, then the coefficient value is zero. The last case, if it completes mismatch between and disagreement between prediction and observation, then coefficient value is -1.

Cohen's kappa coefficient (κ) is a statistic, which is used to extend both Inter-rater dependability and also Intra-rater reliability for qualitative and categorical items. As κ considers, the opportunity of the arrangement happening by chance is generally to be a more robust degree than simple percent agreement calculation. But there is a difficulty in interpreting indices of agreement and this is the controversy for Cohen’s kappa. Many researchers have suggested very good thoughts for evaluating disagreement between items.

3. Implementation

3.1. Gathering Data

While using machine learning, the initial step is to know the problem. According to the problem definition, data should be collected. For machine learning, a dataset can be created or there are data that already exists. There are many platforms that provide the collection of datasets to solve machine learning problems.

3.2. Pre-processing the data

After the data is collected, the data needs to be processed. Without pre-processing the data or providing raw data to the model, it does not provide the expected results. Use the techniques which can provide the best form of the data which increases the accuracy of the model. If the dataset is skewed, try to balance it, perform feature selection, feature extraction, transferred learning.

3.3. Split the dataset

After cleaning the data, divide the dataset. Data can be split into train test ration, train-test-validation ratio, or use cross-validation. By splitting the dataset, the training dataset for the training of the model and the remaining for evaluating the model can be provided. This helps to avoid the over fitting of the model.

3.4. Choosing a model

After dealing with data, select the model according to the dataset, and the type of task needed to be performed like classification, clustering. Choosing an appropriate model is very important or else results will not be achieved.

3.5. Evaluate the model

After the training of the model, predict the results on the unseen dataset. If the prediction metrics provide the results which are expected, then the model is said to be ready for classifying the data. If the results are not satisfactory, retrain the model and change the parameters, fine-tune them, until the achieved results are satisfactory.

.

4. Results and Discussions

From the figure 2, it has been clearly observed, fraud representation of data. Credit card fraud occurs mostly in the time of transaction. Figure.3. represents the comparison between fraud cases and genuine cases in an hour.

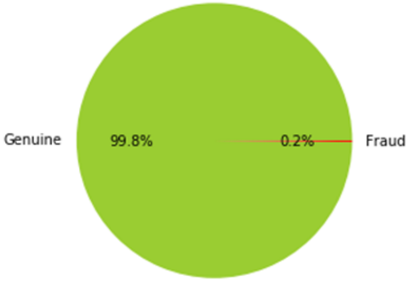


Figure 2. Pie chart representation of data

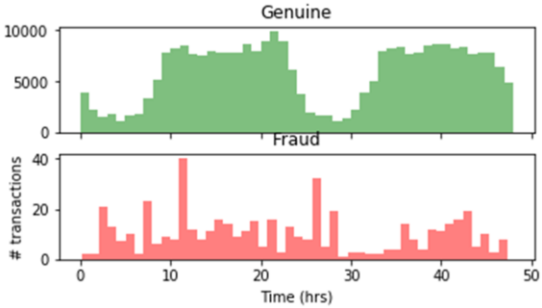
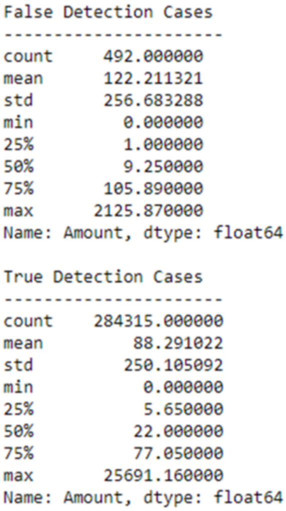


Figure 3. Transactions over time (in hrs)

Detection cases are identified, and it has been separated in such a way as false detection cases and True detection cases. It has been displayed in the table. 1.

Table 1. No. of actual and false transactions and other statistical data



```
False Detection Cases
----------------------
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64

True Detection Cases
----------------------
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

The correlation matric has been calculated and it is shown in figure. 4.

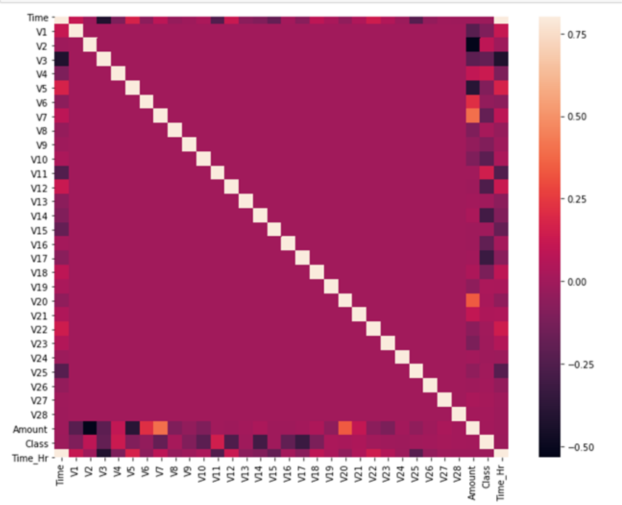


Figure 4. Correlation matrix.

5. Comparative Analysis

We compared the performance of classification models by calculating all the metrics for statistical evaluation that are generated by the algorithm in table 2. True Negative is a number of negative results that are valid and also classified as valid. True Positive is a number of positive outputs that are considered fraudulent and by the system too. False Positive is the number of positive results that are mentioned as valid but are incorrectly mentioned as fraudulent. False Negative is the number of negative results that was mentioned as fraudulent but is incorrectly mentioned as valid by the system.

Table 2. Comparison Chart

Techniques	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Matthews Correlation Coefficient**	**Cohen Kappa**
Isolation forest	99.7%	34.06%	31.6%	32.8%	32.7%	32.7%
Random forest	99.9%	96.05%	74.4%	83.9%	84.5%	83.9%
Decision tree	99.9%	96.05%	74.4%	83.9%	84.5%	83.9%
Naive Bayes	98.3%	8.4%	87.7%	15.4%	26.9%	15.2%
Logistic Regression	99.9%	80.8%	56.1%	66.2%	67.3%	66.2%
K- Nearest Neighbor	99.9%	92.1%	71.4%	80.4%	81.08%	80.4%

6. Conclusion and Future Work

After comparing all the classification algorithms, it is concluded that the random forest/decision tree classifiers are the most suited to this kind of application as the basic classifiers such as logistic regression, Naive-Bayes and K-nearest neighbours all have similar accuracy but significantly lower values in the other metrics. Thus, we can conclude that a fraud detection system that is built based on the random forest algorithm will work well as compared to other algorithms. Future work on this proposed method can enhanced to test/compare more algorithms by building a classifier from multiple models to find a superior one to increase the metric scores.

References

[1] Jain, Y. & Tiwari, N. & Dubey, S. & Jain, Sarika, "A comparative analysis o f various credit card fraud detection techniques,"*International Journal of Recent Technology and Engineering.*vol. 7, 2019.

[2] T. R. C.Sudha, "Credit card fraud detection in the internet using k nearest neighbour algorithm," *IPASJ international journal of computer science*, vol. 5, 2017.

[3] E. D. Yusuf Sahin, "Detecting credit card fraud by ann and logistic regression," 2011.

[4] F. Carcillo, Y.-A. Le Borgne, O. Caelen , Y. Kessaci, F. Oblé, and G.Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection,"*Information Sciences*, doi: 10.1016/j.ins.2019.05.042, 2019.

[5] A. S. Wheeler R, " Multiple algorithms for fraud detection.Knowledge-Based Systems," no. S0950-7051(00)00050-2, 2000.

[6] A. O. A. S. A. O. John o. Awoyemi, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *International conference on computing networking and informatics,* 2010.

[7] S. K. M. A. Masoumeh Zareapoor, "Analysis of credit card fraud detection techniques: based on design criteria," in *International journal of computer applications*, 2012.

[8] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection,"*2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India*, 2019, pp. 320-324, doi: 10.1109/CONFLUENCE.2019.8776925.

[9] Jain, Y., NamrataTiwari, S., & Jain, S. (2019). "A comparative analysis of various credit card fraud detection techniques,"*International Journal of Recent Technology and Engineering*, 7(5), 402-407.

[10] Duman, E., & Ozcelik, M. H. (2011). "Detecting credit card fraud by genetic algorithm and scatter search,"*Expert Systems with Applications*, 38(10), 13057-13063.

[11] Kumar, V. A., Kumar, V. A., Malathi, S., Vengatesan, K., & Ramakrishnan, M. (2018). Facial recognition system for suspect identification using a surveillance camera. Pattern Recognition and Image Analysis, 28(3), 410-420.