A COMPARATIVE STUDY ON CLASSIFICATION ALGORITHMS FOR CREDIT CARD FRAUD DETECTION

Ankit Mohari *1, Joyeeta Dowerah*2, Kashyavee Das*3, Faiyaz Koucher*4, Dibya Jyoti Bora*5

*1,2,3,4PG, Department of Information Technology, Kaziranga University, Jorhat, Assam, India.
*5 Assistant Professor, Department of Information Technology, Kaziranga University, Jorhat, Assam, India.

ABSTRACT

The main problems within the credit card trade are ongoing a fraud. The credit card has made our life easy as we can pay easily and move without carrying any cash. Credit card gains its popularity and utilization has dramatically inflated in our day to day life, for the speedy advancement of electronic commerce technology. However, exploitation of credit card provides huge edges once used fastidiously and responsibly. Fraud activities are also increasing, and new techniques have been developed by criminals. Credit card and monetary damages are caused by fallacious activities. Such issues are tackled with Data Science, Machine Learning together with Deep Learning techniques, which cannot be exaggerated. This helps the bank and financial organizations, to detect the fraud at the early stage, and then they can reduce the ongoing fraud by not accepting the suspected transactions. The credit card company faces a huge loss if the cardholder does not detect the loss. An awfully very little quantity of data is needed by the assaulter for conducting any fallacious dealing in online transactions. During analysis work, numerous methods and outcome are reviewed, in terms of definite parameters.

Keywords: Fraud, Credit Card Fraud Detection , Machine Learning Algorithm.

I. INTRODUCTION

The most prominent form of payment is a credit card. Since the variety of credit card users is growing worldwide, fraud is improved, and fraud is increasing in the purchasing of virtual cards, where only card information such as card number, expiry date, security pin, etc. is required. These transactions are often made on the internet or via the phone. If anyone gets access to the specifics of the card, so fraud can be easily committed. Nowadays, many online transactions are made using credit cards. It is inevitable for user to keep their credit card details information confidential and the service provider must take the initiative to protect the users details; credit card fraud can be characterized, as there might be some cases where the owner and the card issuing authorities are unaware of a situation where a random person uses their card for their own personal purposes. Many frauds, which are detected worldwide, include monitoring the behaviors of user numbers in order to interpret, estimate, or prevent offensive activity that consists of fraud default and interference. Methods of fraud detection are increasingly designed to prevent the actions of offenders from adjusting to their fraudulent techniques, such frauds are known as:

- Online and Offline credit card frauds
- Card pilferage
- Account insolvency
- Interference of device
- Fraud for applications
- Forge Card
- Communication through media/telephone Fraud

Some risky countries are outlined in the figure 1, on the basis of knowledge expressed in [1] in 2012. The United Sates has introduced a minimum level of fraud rate, but credit purchases is highest. With a staggering 19%, Ukraine has the highest amount of fraud rate, along with Indonesia at an 18.3% fraud rate. The most risky nation after these two is Yugoslavia with a rate of 17.8. Malaysia having 5.9%, Turkey having 9% and lastly the
US account for the less amount of fraud rate. Figure 1 doesn’t clearly depicts the Other Country risky for credit card fraud at a rate that is 29%.

**Fig. 1:** Showing the Countries facing Credit Card Fraud

### II. LITERATURE REVIEW

A way of using machine learning for the detection of credit card fraud was suggested by K. Ratna Sree Vall et al. [2]. Supervised learning, provided the unexpected input example of the associate degree, and is designed to perform predictions. The supervised methods used in this paper are, Random Forest, Logistic Regression, Naive Bayes and a boosting technique (AdaBoost) to enhance the classification algorithm. AdaBoost or Adaptive Boosting is a boosting method, which provides us with a single "strong classifier" with the combination of multiple "weak classifiers". It was therefore concluded here that compared to logistic regression and even the Naïve Bayes techniques with a boosting technique, the random forest classifier is stronger.

Sonal Mehndiratta et.al [3] In this paper, various credit card fraud detection techniques are reviewed supported on some parameters. Here through collecting historical data, predictive analysis methods can be used to notice fraud. Different methods, such as Artificial Neural Network, Hidden Markov Model, Genetic Algorithm, Naive Bayes, KNN classifier are used here. In this paper, mainly fraud prediction is done using two phases that is feature extraction and classification and it decides to use a hybrid approach in the future for credit card fraud detection.

Zarrabi et.al[4] Deep Auto encoder is proposed by the author that serves as a best extraction of the details of the features from the fraud transaction occurred in the credit card. For the class mark problems, softmax tools used here by the author. For classifying a form of fraud an AutoEncoder is used here which maps the data into a high-dimensional space. Deep learning can be said as one of the most effective methods for detecting the credit card fraud. To understand the dynamic distribution of the data in the networks types becomes difficult to understand. For extraction the best features of data with a high amount of precision and low variance the networks via Deep Auto encoder was used.

Adnan M. Al-Khatib[5] Fraud detection is often a drawback of the classification of legit transactions from deceitful transactions. Fraud detection, includes watching the defrayal behavior of users/ customers so as to work out, detect, or turning away from undesirable behavior. The utilization of credit cards is common in contemporary society. Several issues are also being faced by the developer to relate to credit card fraud detection. A number of the trendy fraud detection techniques employed by the research worker:- Neural Networks, Genetic Algorithms(GAs), Rule induction, Expert systems, Case-based reasoning(CBR), Inductive logic programming(ILP), Regression, Artificial intelligence, etc. have been used for detecting deceitful transactions. Comparative study on data mining techniques are used to detect deceitful coverage, to obtained low false alarm rate. This study shows, to obtained higher cost savings, multiple algorithms technique, can be used.
Raghavendra Patidar, et al. [6] used a dataset in combination with genetic algorithms (GA), for credit card fraud detection to train three layer of backpropagation ANN. Genetic algorithms were responsible for determining the network layout in this analysis, dealing with the network topology, the number of hidden layers and nodes in each layer.

Dilip Singh Sisodia et al. [7] the study of the output of several classifier-sampling methods when applied to the credit card fraud data set in which classes are imbalanced was presented. Principal component analysis (PCA) for real data and the variables time, quantity, and class obtained 28 main components in the data. We have found ten thousand; fifteen thousand & twenty thousand instances available from the three datasets we have implemented. The authors analyzed five over-sampling and four methods to under-sampling respectively, and few cost-sensitive and ensemble classifiers are applied.

Guanjun Liu et al. [8] Here the author used 2 kinds of random forests to coach the behavior options of traditional and deceitful transactions so compare this each and differentiated on the premise of a classifier, performance on the detection of credit card fraud. The data used is associate with an e-commerce company in China that is used to research the performance of those 2 kinds of random forest models. In this paper, to identify deceitful B2C dataset used by the authors. Random Forest classifier used which obtained better result only on the small dataset and cannot handle any imbalanced dataset, which makes it inefficient, as fraud dataset are mostly imbalanced.

A. Roy, et al. [9] proposed a deep learning method for detecting fraud transactions. 80 million, transactions has been detect as fraud. They have used cloud-based environment to obtained high performance. The researchers have concluded a dep learning method with tuning parameter for deceitful transactions detection which helps the financial institution for the prevention of illegal practices.

Dastgir Pojee et al. [10] proposed a modern process that includes payment of invoices or bills. This technique is referred to as the "No Cash" smartphone program and is primarily used by retailers through whom consumer payment services can be eased. In this situation, there is no need for the NFC-Enabled Point of Sales (PoS) Machines Approach and only cell phones are required. This system is designed which minimized the problem of the customers for bringing the card and provide them a easy payment workflow. The customer's shopping experience is enhanced as the program NoCash, which has several features, is used based on growth in the number of NFC-based mobile devices. Application clients can refer to the history of the cost and the costs will be reduce that are not necessary.

PA. Estevez et al. [11] through this research paper, we have got some idea that every year tens of billion dollars losses are estimated in global telecommunications fraud. Many researchers have implemented many techniques. The system consists of both classification and prediction module, to develop a system for the detection of subscription fraud for the fixed telephone lines. The classification module classifies subscribers according to their past historical conduct into four distinct categories: Subscription is fraudulent, otherwise fraudulent, insolvent, and regular. Some of the methods, which are implemented in the research paper, creating a data set, categories of subscriptions, system architecture, classification, and prediction module. Information over ten-thousand, real subscribers of a major telecom company in Chile was on the database and the classification module was implemented using fuzzy rules. 2.2% subscription fraud prevalence was found in this database. A multiplayer perceptron neural network was implemented for the prediction model. True fraudsters identified were 56.2 %, screening only 3.5 % of all the customers in the test dataset. This research was carried out on fixed telecommunications, but the methods suggested here may be applied to fraud in mobile communications, and also other markets. It also demonstrates the possibility of substantially avoiding telephone service fraud by examining the application details and consumer data at the time of application.

Dahee Choi et al. [12] in this paper, from 2016 to 2018 have surveyed deceitful transactions techniques using machine learning and deep learning algorithms, and analyzed the fraud detection techniques limitations and advantages. Feature selection, sampling, and applying some supervised and unsupervised algorithms to detect deceitful transactions, are the techniques included in the paper on the financial dataset. On 2015, the ultimate model was validated from the data occurring in the actual financial transaction in Korea. In conclusion, machine-learning techniques performed better and detect most deceitful transactions than the artificial neural network. In the machine learning method, the maximum detection rate and the lowest detection rate of were 1 and 0.736 respectively, and when all of the algorithms were analyzed then 0.98618 was the average detection rate.
rate. To maximum detection rate in all ratios, the lowest detection rate and the lowest detection rate of the artificial neural network were 0.914, 0.651, and 0.77228 respectively.

Ivo Correia Feedzai et al. [13] in this paper, the open-source tool, IBM Proactive Technology Online (PROTON) was presented to deal with the uncertainty. All levels of the architecture and logic of an event-processing engine strongly affected by the inclusion of uncertainty. In the complex event processing, programmatic language new capabilities as building blocks and basic primitives, were implemented to proceed the implementation of event-driven applications. At first, the application mechanism is in the domain of credit card fraud detection. To support all the varieties of uncertainty, a few Complex Event Processing (CEP) engine are there. Some limitations are also identified from the recent probabilistic engines, and there is the absence of support the uncertainty, for specifying the complex events.

Phua C et .al [14] In Online Transaction there can be numerous amount of frauds that can occur which is analyzed by viewing the behavior of the user and if there is seen any deviation in the spending behavior than it may be a fraudulent transaction so for that there are many data mining methods are used by bank and credit card companies like Decision Tree, Rule-based mining, ANN and fuzzy clustering, hidden Markov model or a hybrid approach of these methods so any one of these methods will be used to detect the behavioral activities of the customers based on the past experience, this paper mainly will compare the different techniques that would detect fraud more precisely. After comparing the algorithms it was found out that every algorithm has drawbacks as well as some advantages as ANN and Hidden Markov Model can handle large data but the imitation is that the process becomes slow and it is also expensive, it is also seen that decision tree may be easy to understand but has a limitation as it cannot handle complex data. So to increase its accuracy Precision and Recall is used so that the amount of false positive and false negative can be reduced to a certain extent and finally F1 score is found out which is the harmonic mean of precision and recall which depicts that if the F1 score is higher it indicates it as a good model. Therefore, as we can see a combination of those methods mentioned above can be used to detect deceitful transactions were adding new features and sampling methods could be performed to train the model accurately.

Aditya Saini et.al [15] in this paper, anomaly detection algorithms are used which include Isolation Forest and Local Outlier Factor. They have analyzed the dataset with various statistical methods. Only 10% sample was used for training the model with the outlier detection algorithms. Comparative analysis with two algorithms was done and the outcome was not that desirable one. As Isolation forest only can give 28% precision rate on 10% sample data and 33% precision rate on 100% dataset, and Local Outlier Factor precision rate was only 2%
3.1 Dataset: In this paper credit card, fraud detection dataset was used, which can be downloaded from data.world[16]. This dataset was first used in this research paper[17], which composed of credit card transactions from September 2013. It has 492 frauds out of 284,807 transactions that occurred in two days. The dataset is extremely unbalanced, where fraud accounts for only 0.172% [17]. The “Time” denotes the seconds elapsed between each transaction and the first transaction in the dataset. The “Amount” is the transactions amount. For confidentiality reasons, the meaning of most variables (V1-V28) is not revealed and the features have been transformed by means of principal components. Even the cardholder identifier is also not available on the dataset.

3.2 Understanding the data:

**Fig.-3:** This graph shows that the number of fraud transactions is much lower than the legit ones.

**Fig.-4:** This graph shows that the fraud transaction amount is less than 5000
Fig.-5: This graph shows the times at which transactions were done within 48 hrs. It can be seen that the least number of transactions were made during night time and highest during the days.

Fig.-6: This graph represents the amount that was transacted. We plot a heatmap to get a representation of the data and to study the correlation between our predicting variables and the class variable. This heatmap is shown below:
The dark side represent positive correlation while the lighter side represent negative correlation there is no notable correlation between features V1-V28. There are certain correlations between some of these features and Time (inverse correlation with V3) and Amount (direct correlation with V7 and V20, inverse correlation with V1 and V5).

IV. TECHNICAL TERMINOLOGIES

Some Techniques To Detect Credit Card Fraud Are Explained Below

**Naive Bayes:** In this classification method, the probability of an object associated with a particular category or class with certain feature is learned. Naive Bayes algorithm can fit model fast and provide high accuracy when applied to big data and need less training.

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

where, \(P(A)\) and \(P(B)\) are just the probability of A and B occurring, \(P(B|A)\) is the probability of event B occurring ensuring that event A is true, \(P(A|B)\) is the conditional probability of event A occurring, event B being true.

**Logistics Regression:** Logistics Regression is less inclined to the overfitting, but it can overfit in high-dimensional datasets. We can consider regularization techniques to avoid overfitting. Any big outliers will be transform into the range of 0 and 1. Its help mainly to solve classifications problem and supply us the knowledge weather the event is happening or not.

**Random Forest:** Firstly, we begin by choosing random samples from the dataset that is provided. Then, this algorithm will be used to create a decision tree for every sample that is generated. Then, the prediction result is found out for each decision tree. For each expected outcome, voting mechanism is carried out. Therefore, ultimately as the final prediction outcome the most voted prediction outcome is selected.

**AdaBoost:** Adaboost or Adaptive Boosting, improve the performance of the weak classifier, here each instance in the training dataset is weighted.

\[
\text{Initial Weight: } \text{Weight}(x_i) = \frac{1}{n}
\]
where, \( x_i \) is the \( i \)th training instance \\
n is the number of training instance.

**Artificial Neural Network:** A collection of nodes are interlinked which is designed to depict the functioning similar to the human brain which is known as Artificial Neural Network. All the other nodes that are present in the adjacent layers of the node are being assigned the weighted connection.

**Genetic Algorithm:** This algorithm is being inspired by natural evolution. So, from here the algorithm has been introduced. The binary strings used in chromosomes that describe the population of candidate solutions. It assumes that the chances of survival and the amount of reproduction is higher with the higher quality chromosomes which means it increases the health value.

**Hidden Markov Model (HMM):** A Hidden Markov model is also described as double embedded stochastic process using which highly complex stochastic processes can be generated. Within the underlying framework, a Markov process that has an unnoticed stage is presumed to be available. The definite transformation of the state's present inside the simpler Markov models is the only unknown parameters that are present.

**KNN Classifier:** In the case of classification and regression KNN is generally the non-parametric algorithm that is used. The input of this algorithm consists of K nearest training examples in the feature space for classification and regression whereas on the other hand, the output generally depends on whether the KNN belongs to classification category or regression category.

**Decision tree:** Decision tree is a tree shaped structure that expresses mainly independent attributes and dependent attributes and is a data mining technique. Classification rules that are derived from the decision trees are generally expressions of IF-THEN and each rule is needed to be produced, all the tests must succeed. Decision tree is generally a non-parametric algorithm.

**Isolation Forest:** Isolation Forest is an unsupervised anomaly detection learning algorithm that operates on the concept of isolating anomalies. The Isolation Forest algorithm aims to make it easier to separate anomalous instances in a dataset from the rest of the sample (isolate), relative to usual points. The algorithm recursively creates partitions in the sample in order to isolate a data point by randomly selecting an attribute and then randomly selecting a split value for an attribute between the minimum and maximum values permitted for that attribute.

**Local Outlier Factor:** Local Outlier Factor is an algorithm used for the identification of unattended outliers. It creates an anomaly score that reflects data points in the dataset that are outliers. It does this by calculating a given data points local density variance with respect to the data points around it.

**Steps for finding efficient Algorithm:**

1. Import the Dataset 
2. Convert the data in dataframe format 
3. Sampling the dataset 
4. Decide the amount for training and testing data 
5. Assign the train dataset to the models 
6. A comparative analysis will be carried out on these algorithms to find out the most efficient one 
7. Make predictions for test dataset for each algorithm. 
8. Measure the quality of predictions for each algorithms by using confusion matrix and precision.

**Evaluation:** There are a variety of measures for various algorithms and these measures have been developed to evaluate very different things. So it should be criteria for evaluation of various proposed method. False Positive(FP), False Negative(FN), True Positive(TP), True Negative(TN) and the relation between them are quantities which usually adopted by credit card fraud detection researchers to compare the accuracy of different approaches. The definitions of mentioned parameters are presented below:

**True Positive(TP):** The true positive rate represents the portion of the fraudulent transactions correctly being classified as fraudulent transactions.
True Negative Rate (TN): The true negative rate represents the portion of the normal transactions correctly being classified as normal transactions.

\[
\text{True Negative} = \frac{TN}{TN + FP}
\]

False Positive (FP): The false positive rate indicates the portion of the non-fraudulent transactions wrongly being classified as fraudulent transactions.

\[
\text{False Positive} = \frac{FP}{FP + TN}
\]

False Negative (FN): The false negative rate indicates the portion of the non-fraudulent transactions wrongly being classified as normal transactions.

\[
\text{False Negative} = \frac{FN}{FN + TP}
\]

Confusion matrix: The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with negative and positive classes. In this type of confusion matrix, each cell in the table has a specific and well understood name.

V. RESULT AND DISCUSSION

The following results were observed as the models - Naive Bayes, logistic regression, random forest, random forest with boosting technique, Artificial Neural Network, KNN classifier, Decision Tree, Isolation Forest, Local Outlier Factor and SVM were evaluated against the data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy Score</th>
<th>Fraud Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 GaussianNB</td>
<td>0.976905</td>
<td>0.844920</td>
</tr>
<tr>
<td>1 Logistic Regression</td>
<td>0.999166</td>
<td>0.588235</td>
</tr>
<tr>
<td>2 Random Forest Classifier</td>
<td>0.999517</td>
<td>0.754011</td>
</tr>
<tr>
<td>3 AdaBoostClassifier</td>
<td>0.999535</td>
<td>0.754011</td>
</tr>
<tr>
<td>4 Artificial Neural Network</td>
<td>0.999421</td>
<td>0.737968</td>
</tr>
<tr>
<td>5 KNeighborsClassifier</td>
<td>0.999429</td>
<td>0.727273</td>
</tr>
<tr>
<td>6 DecisionTreeClassifier</td>
<td>0.999078</td>
<td>0.721925</td>
</tr>
<tr>
<td>7 IsolationForest</td>
<td>0.000255</td>
<td>0.155080</td>
</tr>
<tr>
<td>8 LocalOutlierFactor</td>
<td>0.001475</td>
<td>0.389396</td>
</tr>
<tr>
<td>9 Support Vector Machine</td>
<td>0.997250</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Fig.-8: Comparison Results
VI. CONCLUSION

The credit card fraud detection methods have gained popularity in the past decade with the evolution of statistical models, machine learning algorithms, data mining techniques. The fraud transaction prediction has 2 phases which are feature extraction and classification. Within the first phase, the feature extraction technique is applied and within the second phase, classification is applied for fraud transaction detection. Fraud transaction detection is the major issue of prediction because of a frequent and enormous number of transactions. During this comparative research study, we tried to analyze the dataset through various graphs and also tried to detect fraud using some classification algorithms and make comparative analysis. We have got the high percentage of accuracy because of the huge imbalance between the number of valid and number of genuine transactions. By comparing all the 10 methods, we found that Local Outlier Factor fraud accuracy is greater than the rest of the algorithms.

ACKNOWLEDGEMENTS

We would like to take the opportunity to thank Dr. Dibya Jyoti Bora, Assistant Professor, Kaziranga University to provide the necessary support and suggestions for our research work.

REFERENCES


https://data.world/vlad/credit-card-fraud-detection