IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud Detection using Different Machine Learning Models

Priyanka Sharma[1*] and *Santoshi Pote*[2]

[1]Department of Electronic and Communication, Usha Mittal Institute of Technology, SNDT University, Mumbai, India.

[2]Department of Electronic and Communication, Usha Mittal Institute of Technology, SNDT University, Mumbai, India

Abstract: Credit card fraud is an event problem and fraud detecting techniques getting more sophisticated each day. It has cost banks and their customers a loss of billions of rupees. The techniques used now a day detects the anomaly only after the fraud transaction takes place. The intruders have found ways to crack the system loopholes and defeat the security. Thus, Artificial Intelligent (AI) algorithms are used to detect the behaviour of such activity by learning the past behaviour of the transaction of the user not only this, but data is also unbalanced so SMOTE is used. Various concepts and types have been introduced in this paper related to credit card anomaly detection systems like decision tree, random forest, Neural Networks, and Auto-encoder. A comparative study is done in this paper on the bases of accuracy, precision, recall and AUC curve.

Keywords: Credit card fraud, Artificial Intelligent, SMOTE, decision tree, random forest, Neural Network and Auto-encoder.

1. INTRODUCTION:

These days, a credit card is a well-known method of transactions used in many scenarios like online shopping, bill payments, bank-to-bank money transfer, etc. Due to a rise in demands for online transactions, credit card users have increased leading to credit card frauds. A credit card is a payment card provided to clients as an arrangement of payment. There are heaps of advantages in utilizing Visas, for example:

- **Purchasing convenience:** The credit card allows customers to buy anything at any time, place or amount without conveying the money.
- **Hold the records of custom credit:** Great financial record is regularly significant in identifying steadfast clients. This history is important for a credit card, yet additionally for other budgetary administrations like credits, rental applications, or even a few employments.
- **Purchase security:** A credit card may likewise offer clients, extra assurance if the bought stock gets lost, harmed, or taken. Furthermore, some charge card organizations give protection to enormous buys.

Despite all referenced advantages, the issue of misrepresentation is difficult in e-banking administrations that undermine charge card exchanges particularly. In the era of digitalization, the need to identify credit card frauds is mandatory. Further, the individual utilizing the card has no association with the cardholder or backer and has no aim of either reaching the proprietor of the card or making reimbursements for the buys made[1].

For fraud, a credit card is a basic objective in light of the fact that with no hazard, an immense measure of cash can be picked up inside a brief time frame. To commit fraud, fraudsters endeavour to take delicate information like card number, financial balance information, and CVV number. These fraudulent transactions are so legitimate that it makes fraud

detection a difficult issue. In this manner, a great misrepresentation recognition framework ought to have the option to distinguish the extortion exchange precisely and should make the discovery conceivable progressively exchanges.

In the detection of credit card fraud, both unsupervised and supervised learning is investigated. Supervised learning uses defined data sets to train and make correct learning by adjusting the learning rate parameters. The disadvantage of supervised learning is that if new fraud transactions happen that don't coordinate with the records of the database, at that point this transaction will be viewed as real. Although, unsupervised learning acquires new transaction knowledge and discovers anomalous trends from new transactions. This uncontrolled learning is tougher than supervised learning, as we need effective methods for identifying irregular behaviours.

Deep learning is another innovation that as of late pulled in a lot of consideration in the field of AI. By reconstructing deep structures such as the neural networks on the human brain, it greatly increases the accuracy of abstract representations. In this paper, we had compared deep learning algorithms Auto-encoder and Neural Network with machine learning algorithms such as decision tree, random forest using SMOTE (Synthetic Minority Oversampling Technique) to overcome the problem of the imbalanced dataset. The finding of these algorithms will also help us to find the best algorithm for detecting credit card fraud. The remaining paper as per the following section 2 explains all the current system use in fraud detection. Followed by section 3 portrayed the proposed technique, section 4 shows the performance analysis and results and last section 5 shows the conclusion.

2. RELATED WORK:

According to [2], they have proposed the utilization of HMM in credit card extortion discovery. The proposed technique for finding the spending profile of cardholders, just as the utilization of this information in choosing the estimation of perception images and an underlying appraisal of the model parameters. It has likewise been clarified how the HMM can identify whether an approaching exchange is fake or not. Trial results show the presentation and adequacy of our framework and show the convenience of learning the spending profile of the cardholders. Relative examinations uncover that the Accuracy of the framework is near 80% over a wide variety in the information. The framework is moreover adaptable for taking care of huge volumes of exchanges.

Ayahiko Niimi [3], led tests that affirm that deep learning has a similar precision as the Gaussian kernel SVM. Likewise, the 10-fold cross-validation analysis demonstrates that it is deep learning offers higher exactness. In this experiment, they had utilized the H2O library for deep learning, with the deep learning modules are written in Java were actuated each time. Thusly, they can't evaluate the execution time. Deep learning parameter alteration is troublesome. By upgrading the parameters, it is conceivable to build the learning exactness.

According to [4], examination uncovers a relative execution of CFLANN, MLP, and Decision Tree more than two unique informational collections for credit card fraud detection. The outcome shows that in both the informational collection MLP outflanked CFLANN and Decision Tree in misrepresentation recognition. Even though FLANN with other info development has been effectively utilized in different regions like an expectation in which FLANN performed better than MLP however in MasterCard extortion location MLP has marginally an edge over CFLANN.

Pooja Chougule, A.D. Thakare and others [5], work mirrors an endeavor to distinguish false card transactions by utilizing k-means alongside a genetic algorithm. Genetic Algorithm is an incredible optimization method. The k-means algorithm bunches the MasterCard transaction dependent on autonomous quality qualities. Be that as it may, with the expansion in the information size, it brings about anomalies. Consequently, to give enhanced recognition of cheats, they had utilized a hereditary calculation. The huge outcomes by the proposed model are seen over straightforward K-means and Simple Genetic Algorithm.

M.Suresh Kumar, V.Soundarya and others [6] proposed the Random Forest Algorithm (RFA) for finding the false transactions and the precision of those transactions. This algorithm depends on a supervised learning algorithm where it utilizes choice trees for classification of the dataset. After the classification of the dataset, a confusion matrix is acquired. The presentation of the Random Forest Algorithm is assessed depending on the confusion matrix. The outcome got from handling the dataset gives a precision of around 90-95%.

3. PROPOSED TECHNIQUE:

This paper uses the methods suggested identifying credit card fraud. Comparisons are made with various machine learning algorithms, such as decision tree, random forest and deep learning, including auto-encoders or neural networks, which algorithm is better suited to classify fraud transactions by credit card dealers.

3.1. Dataset

In 2013, 284,807 European data sets were used for two days. It includes 492 fraud transactions listed as 1 and 0 for other transactions. The fraud-to-non-fraud transaction ratio is 0.17%, which reveals a very imbalanced data collection. The original features are not seen on this data set due to customer confidentiality and include 28 PCA mapping features plus two unmapped features known as the time and transaction number. As the data is very imbalanced, SMOTE is used to align the data collection for machine learning algorithms.

3.2. Machine Learning Algorithms

For Machine learning algorithms SMOTE is used to overcome highly imbalance data. The over-sampling technique by the synthesized minority reduces the non-fraud transaction. The parameters of SMOTE() function synthesizes the confluence. The result of this technique will be compared with other algorithm is used for comparison

Decision Tree

A decision tree is a kind of supervised learning algorithm that is primarily used in problem classification. It operates both for the category and continuous variables of input and output. In this technique, the population test is separated into at least two homogenous sets depending on the most appropriate input divider/differentiator. Decision trees appear to have a significant variation when separate training and testing sets with the same data are used because they are over fitting with the training data. This contributes to poor data output. Unfortunately, this restricts the use of predictive modelling decision trees[7].

Random Forest

Random forests are tree-based algorithms that require the formation and combination of many trees with outputs to improve the model's generalization. It is known as an ensemble approach for the combination of trees. Combining is simply a mixture of weak students (individual trees) generating a strong student [8]. Random forests can be used to solve problems with regression and classification. The dependent variable is constant in cases of regression problems.

3.3. Deep Learning Algorithm

Auto-encoder

Auto-encoders are unsupervised learning techniques that affect the neural network of portrait learning assignments. It is designed with the ultimate purpose of forcing a device bottleneck that provides a packed information portrait of the first information. This strain and subsequent rework would be a very difficult job if the knowledge highlights were each separate from one another. In any case, if there is an information structure, this structure can be used by scientists and the machine bottleneck to drive contribution.

Neural Network

Neural networks are a set of algorithms, demonstrated freely after the human mind, that are intended to perceive designs. The systems are worked from singular parts approximating neurons, normally called units or just "neurons." Each unit has some number of weighted sources of info. These weighted information sources are added together at that point went through an activation function to get the unit's yield.

There are fundamentally three sorts of nodes in neural network:

- Input unit: Provides network information from outside world. These nodes do not compute they simply pass the information on to the hidden nodes.

- Hidden unit: It calculations and transfers the information from input nodes to output nodes. A hidden nodes forms a set of "Hidden Layer". Although there may be one input layer and only one output layer in a feed-forward network, it may have no or several Hidden Layers.
- Output unit: The output nodes are called the "Output layer" collectively and are responsible for computations and transmission of information from the network to the outside world.

In this study to implement Neural Network, we had used the PyTorch. It generally utilizes the style and intensity of python which is understand easy and used. It core give two primary component, for example, a n-dimensional Tensor, like numpy yet can run on GPUs and programmed separation for building and preparing neural systems.

The experimental NN consisted of 4 hidden layers and each layer is backed with a non-linear activation function – Rectified Linear Unit (ReLU). The input features of each hidden layer are set to 30, 50, 32 and 16 respectively. Deeper networks have been shown to produce better performance than those with fewer layers. After this experiment, we started slowly by increasing a smaller number of layers to obtain appropriate results. Therefore, based on extensive analysis, the best hyper-parameters were chosen. More network improvement resulted in more machine time and the results were not so different from the design chosen. Adaptive Moment Estimation (Adam) is a stochastic gradient descent (SGD) and RMSprop-based optimizer, accomplished weight optimization.

4. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS:

These are the result of machine learning algorithms decision tree and random forest appeared fig 1, 2, 3 and 4 respectively as we referenced above that the dataset was isolated for training and testing in a proportion of 80:20. The basic performance measures derived from the AUC and confusion matrix. The confusion matrix is a 2 by 2 matrix table contains four results delivered by the paired classifier. Different estimates, for example,accuracy, precision, recall and F1 score are gotten from the confusion matrix.

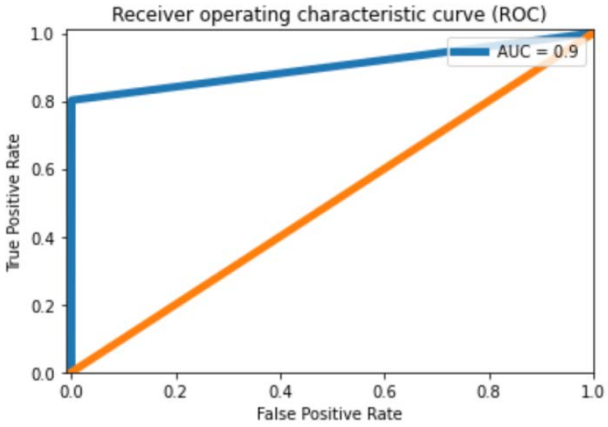


Figure 1: AUC of Decision Tree using SMOTE

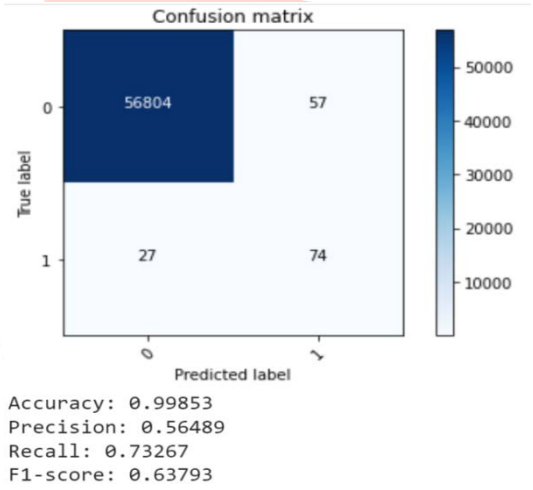


Figure 2: Confusion Matrix of Decision Tree using SMOTE

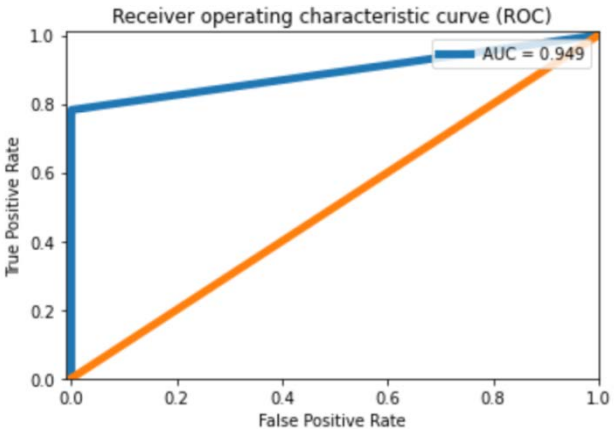


Figure 3: AUC of Random Forest using SMOTE

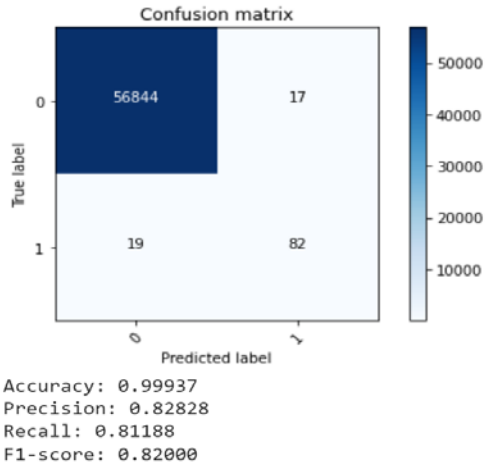


Figure 4: Confusion Matrix of Random Forest using SMOTE

Here, we move on to the deep learning algorithm which are Auto-encoder and Neural network. The results are shown in Fig. 5,6,7,8.

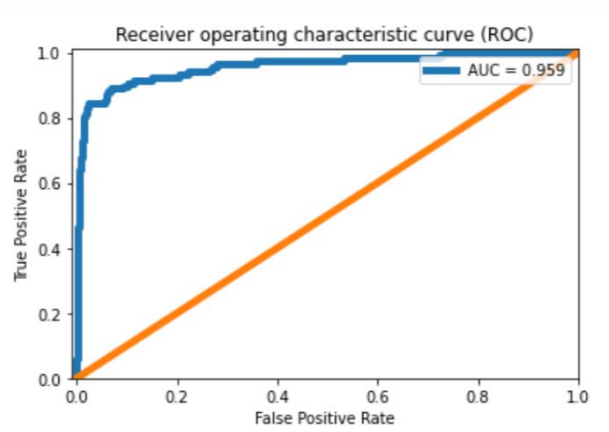
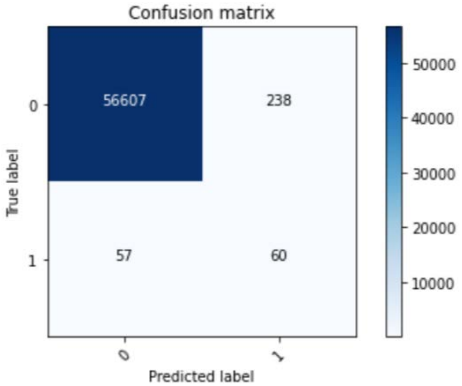


Figure 5: AUC of Auto-encoder



Accuracy: 0.99482
Precision: 0.20134
Recall: 0.51282
F1-score: 0.28916

Figure 6: Confusion Matrix of Auto-encoder

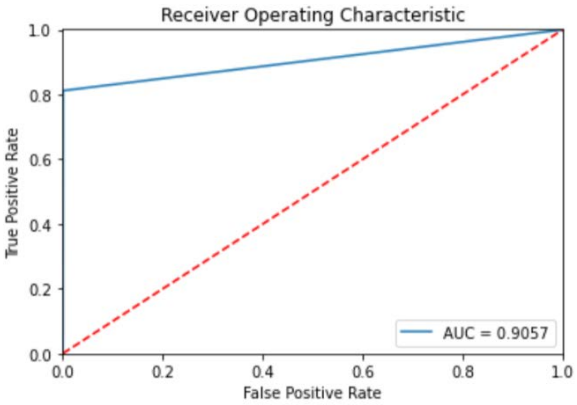
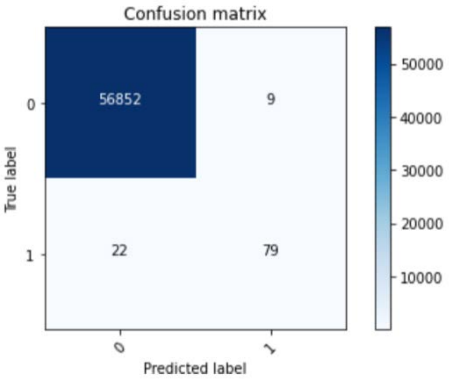


Figure 7: AUC of Neural Network



Accuracy: 0.99946
Precision: 0.89773
Recall: 0.78218
F1-score: 0.83598

Figure 8: Confusion Matrix of Neural Network

Accuracy, precision, recall and F1-score are utilized to report the presentation of the framework to identify the fraud in the credit card. In this paper, two machine learning algorithms and two deep learning algorithms are developed to detect the fraud in credit card system. To evaluate the algorithms, 80% of the dataset is used for training and 20% is used for testing and validation. Accuracy, precision, recall, F1-score are used to evaluate for different variables for three algorithms as shown in Table 1.

Table 1: Performance analysis for different algorithms

Algorithms	Decision Tree	Random Forest	Auto-encoder	Neural Network
Accuracy	99.85%	99.93%	99.48%	99.94%
Precision	56.48%	82.82%	20.13%	89.77%
Recall	73.26%	81.18%	51.28%	78.21%
F1-score	63.79%	82.00%	28.91%	83.59%

The accuracy result is shown for decision tree; random forest, auto-encoder and neural network are 99.85%, 99.93%, 99.48%, and 99.94% respectively. In such a case where data is critical, system cannot rely only on accuracy. System has to be more precise than being accurate. It should detect less number of false positive and false negative cases.

5. CONCLUSION

In order to make final comparison of all the above algorithms with respect to their classification accuracy, best result have been take from Table 1.

As shown in the confusion matrix above, a fine-tuned Neural Network based system has detected less number of false positives compared to other counterparts hence giving highest precision. Random forest gives almost the same results but has ~7% difference in its precision which is a lot while handling such a sensitive data. While NN can be fine-tuned further for better results whereas more number of trees in RF will create lot of confusion. Adding more layers will make Auto-encoders more complex to train resulting in delayed output. The comparative results show that the neural network performs better than other three algorithms.

In future one can further fine-tuning hyper-parameters the neural network, perform boosting techniques on different Machine Learning algorithms. One can also compare the results of different deep learning libraries like fast.ai.

REFERENCES

[1] T. P. Bhatla, V. Prabhu, and A. Dua, "Understanding Credit Card Frauds," *Cards Bus. Rev.*, vol. 1, no. 6, pp. 1–15, 2003, doi: 10.1.1.431.7770.

[2] A. Srivastava, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection using Hidden Markov Model," *IEEE Trans. Dependable Secur. Comput.*, vol. 5, no. 1, pp. 37–48, 2008, doi: 10.1109/TDSC.2007.70228.

[3] A. Niimi, "Deep learning for credit card data analysis," *2015 World Congr. Internet Secur. WorldCIS 2015*, pp. 73–77, 2015, doi: 10.1109/WorldCIS.2015.7359417.

[4] M. K. Mishra and R. Dash, "A comparative study of chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection," *Proc. - 2014 13th Int. Conf. Inf. Technol. ICIT 2014*, vol. 228, no. August 2013, pp. 228–233, 2014, doi: 10.1109/ICIT.2014.25.

[5] P. Chougule, A. D. Thakare, P. Kale, M. Gole, P. Nanekar, and A. K. Algorithm, "Genetic K-means Algorithm for Credit Card Fraud Detection," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1724–1727, 2015.

[6] M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," *2019 Proc. 3rd Int. Conf. Comput. Commun. Technol. ICCCT 2019*, pp. 149–153, 2019, doi: 10.1109/ICCCT2.2019.8824930.

[7] X. Niu, L. Wang, and X. Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised," 2019, [Online]. Available: http://arxiv.org/abs/1904.10604.

[8] S. Venkata Suryanarayana, G. N. Balaji, and G. Venkateswara Rao, "Machine learning approaches for credit card fraud detection," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 917–920, 2018, doi: 10.14419/ijet.v7i2.9356.

[9] SamanehSorournejad, Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective," no. November, 2016, [Online]. Available: http://arxiv.org/abs/1611.06439.

[10] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," *Proc. Hawaii Int. Conf. Syst. Sci.*, vol. 3, pp. 621–630, 1994, doi: 10.1109/hicss.1994.323314.