Credit Card Fraudulent Detection Using Machine Learning Algorithm

Dr. P. Siva Kumar [1], Preethika[2], Sivagami[3], Sridevipriya[4], Vishali[5]

[1]Professor and HOD, [2]Department of Information Technology, Manakula Vinayagar Institute of Technology, Puducherry, India.

Abstract -The fraudulent transactions that occur in credit cards end in huge financial crisis. Since the web transactions has grown rapidly, the results of digitalized process hold an enormous sharing of such transactions. So, the financial institutions including banks offers much value to the applications of fraud detection. The Fraudulent transactions can occur in different ways and in various categories. Our work mainly focuses on detecting the illegal transactions effectively. Those transactions are addressed by employing some machine learning models and therefore the efficient method is chosen through an evaluation using some performance metrics. This work also helps to select an optimal algorithm with reference to the machine learning algorithms. We illustrate the evaluation with suitable performance measures. We use those performance metrics to evaluate the algorithm chosen. Within the existing system the algorithms provide less efficiency and makes the training model slow. Hence within the proposed system we used Multilayer Perceptron and Random Forest to supply high efficiency. From these algorithms efficient one is chosen through evaluation.

Keywords — Credit card, Fraud detection, Multi-layer Perceptron, Random forest.

I. INTRODUCTION

The most accepted payment mode is MasterCard for both offline and online in today's world, it'll provide cashless shopping at every shop across the planet. It'll be the foremost suitable thanks to do online shopping, paying bills, and performing other related tasks. Hence risk of fraud transactions using MasterCard has also been increasing. Within the prevailing MasterCard fraud detection processing system, fraudulent transaction are going to be detected after transaction is completed. Master card fraud may be a huge ranging term for theft and fraud committed using or involving at the time of payment by using this card. MasterCard fraud is additionally an add on to fraud. As per the knowledge from the us Federal Trade Commission, the theft rate of identity had been holding stable during the mid- 2000s, but it had been increased by 21 percent in 2008. Albeit MasterCard fraud, that crime which most of the people accompany ID theft, decreased as a percentage of all ID theft complaints In 2000, out of 13 billion transactions made annually, approximately 10 million or one out of each 1300 transactions clothed to be fraudulent. 0.05% (5 out of each 10,000) of all monthly active accounts was fraudulent.

Today, fraud detection systems are introduced to regulate one-twelfth of 1 percent of all transactions processed which still translates into billions of dollars in losses. MasterCard Fraud is one among the most important threats to business establishments today. MasterCard fraudsters employ an outsized number of the way to commit fraud. In simple terms, MasterCard Fraud is defined as, When a private uses

II. RELATED WORK

Aman Gulati et al [2017] has addressed that are needed for building artificial neural networks for fraud detection. The Multilayer perceptron neural network model, which is taken into account together the efficient models amongst the artificial neural systems. It's a bolster forward directed quite neural system. The multilayer perceptron features a concealed layer and may convey outputs with quite two classes. A standout amongst the foremost vital parts of multilayer perceptron is planning the concealed layer i.e. the hidden layers should contain adequate neurons to grasp the knowledge included and make two distinct classes of output. Lesser the amount of neurons within the hidden layer, better the output are going to be. Zojaji et al [2016] has gave a survey on card fraud detection techniques supported technique oriented perspective. Maes et al[9] analyzed Bayesian networks and neural network for this problem.

III. TECHNIQUES OF CREDITCARD FRAUD DETECTION

Random Forest

Random forest may be a tree based algorithm which involves building several trees and mixing with the output to enhance generalization ability of the model. This method of mixing trees is understood as an ensemble method. Ensembling is nothing but a mixture of weak learners (individual trees) to supply a robust learner. Random Forest are often wont to solve regression and classification problems. In regression problems, the

 DOI : 10.35291/2454-9150.2020.0330 © 2020, IJREAM All Rights Reserved.

variable is continuous. In classification problems, the variable is categorical.

Working of Random Forest

Bagging Algorithm is employed to make random samples. Data set D1 is given for n rows and m columns and new data set D2 is made for sampling n cases randomly with replacement from the first data. From dataset D1, 1/3rd of rows are overlooked and is understood as Out of Bag samples. Then, new dataset D2 is trained to the present models and Out of Bag samples is employed to work out unbiased estimate of the error. Out of m columns, M<< m columns are selected at each node in the data set. The M columns are selected at random. Usually, the default choice of M, is m/3 for regression tree and M is sort(m) for classification tree. Unlike a tree, no pruning takes place in random forest i.e., each tree is grown fully. In decision trees, pruning is a method to avoid over fitting. Pruning means selecting a sub tree that leads to the lowest test error rate. Cross validation is used to determine the test error rate of a sub tree. Several trees are grown and the final prediction is obtained by averaging or voting.

Algorithm steps for finding the Best algorithm

Step 1: Import the dataset.

Step 2: Convert the data into data frames format. Step3: Do random oversampling using ROSE package.

Step4: Decide the amount of data for training data and testing data.

Step5: Give 70% data for training and remaining data for testing.

Step6: Assign train dataset to the models.

Step7: Choose the algorithm among 3 different algorithms and create the model.

Step8: Make predictions for test dataset for each algorithm. Step9: Calculate accuracy for each algorithm.

Step10: Apply confusion matrix for each variable.

Step11: Compare the algorithms for all the variables and find out the best algorithm.

Multi-layer perceptron

A multilayer perceptron (MLP) may be a class of feed forward artificial neural network (ANN). The term MLP is employed ambiguously, sometimes loosely to ask any feed forward ANN, sometimes strictly to ask networks composed of multiple layers of perceptron's. The hard-limiting (step) function used for producing the output prevents information on the important inputs flowing on to inner neurons. During a multilayer perceptron, the neurons are arranged into an input layer, an output layer and one or more hidden layers. A multilayer perceptron (MLP) may be a deep, artificial neural network. They're composed of an input layer to receive the signal, an output layer that creates a choice or prediction about the input, and in between those two, an arbitrary number of hidden layers that are truth computational engine of the MLP.

IV. PROPOSED TECHNIQUE

The proposed techniques are utilized in this paper, for detecting the frauds in MasterCard system. The comparison are made for various machine learning algorithms like Random Forest and multi-layer perceptron, to work out which algorithm gives suits best and may be adapted by MasterCard merchants for identifying fraud transactions. The Figure1 shows the architectural diagram for representing the general system framework.

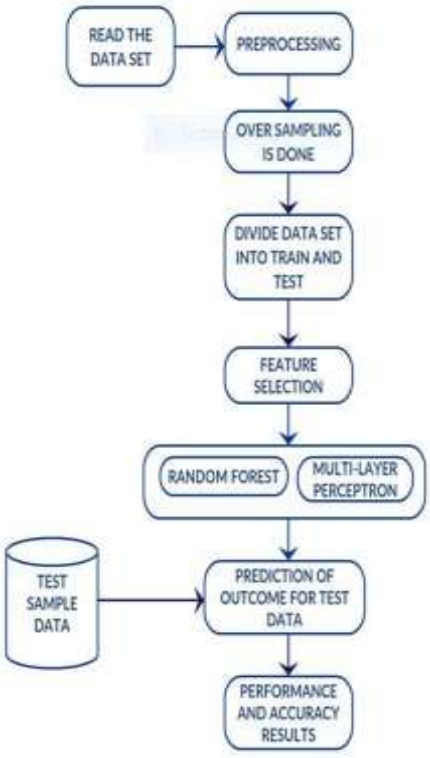

Figure1: System Architecture

Algorithm steps

Step 1: Read the dataset.

Step 2: Data set is given as input for preprocessing to make further analysis. The preprocessed data has undergone oversampling to adjust the class distribution.

Step 3: Oversampling is done on the data set to make it balanced.

Step 4: The data set is divided into train and test set to predicted the expected output.

Step 5: Feature selection are applied for the proposed models.

Step 6: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

 DOI : 10.35291/2454-9150.2020.0330 © 2020, IJREAM All Rights Reserved.

Step 7: Then retrieve the best algorithm based on efficiency for the given dataset

V. PERFORMANCE METRICS AND EXPERIMENTAL RESULTS

Performance metrics

The basic performance measures derived from the confusion matrix. The confusion matrix may be a 2 by 2 matrix table contains four outcomes produced by the binary classifier. Accuracy are derived from the confusion matrix.

Confusion Matrix

A confusion matrix may be a table that's often wont to describe the performance of a classification model (or "Classifier") on a group of test data that truth values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is usually mislabeled because the other. Most performance measures are computed from the confusion matrix. A confusion matrix may be a summary of prediction results on a classification problem. The amount of correct and incorrect predictions are summarized with count values and weakened by each class. This is often the key to the confusion matrix. The confusion matrix shows the ways during which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the kinds of errors that are being made.

		Prediction	
	-	0	1
Actual	0	TN	FP
	1	FN	TP

Figure 2: Confusion Matrix

Where,

Positive (P): Observation is positive (for example: is an apple).

Negative (N): Observation is not positive (for example: is not an apple).

True Positive (TP): Observation is positive, and is predicted to be positive.

False Negative (FN): Observation is positive, but is predicted negative.

True Negative (TN): Observation is negative, and is predicted to be negative.

False Positive (FP): Observation is negative, but is predicted positive.

Accuracy

Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

The accuracy using Random Forest Classifier is calculated as 93.17. Multilayer Perceptron has 99.8 as accuracy which is more than RFC.

Precision

Precision is given by the relation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A precision of 100% is obtained using Multilayer Perceptron which is more than RFC that has 98%.

Recall

Recall is given by the relation**:**

$$\text{Recall} = \frac{TP}{TP + FN}$$

The recall/sensitivity using Random Forest Classifier is calculated as 90.8 whereas, Multilayer Perceptron has 99.7 as recall/sensitivity which is more than RFC.

Specificity

Specificity is given by the relation:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The specificity using Random Forest Classifier is calculated as 97.3whereas, Multilayer Perceptron has 100 as recall/sensitivity which is more than RFC.

F1 Score

F1 Score is given by the relation:

$$\text{F1 Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

The F1-score using Random Forest Classifier is calculated as 99.85 whereas, Multilayer Perceptron has 99.8 as which is more than RFC.

AUC - ROC

AUC-ROC is given by the relation:

AUC –ROC = roc_auc_score(X,Y)

The AUC-ROC using Random Forest Classifier is calculated as 91 whereas, Multilayer Perceptron has 99.87 as which is more than RFC.

 DOI : 10.35291/2454-9150.2020.0330 © 2020, IJREAM All Rights Reserved.

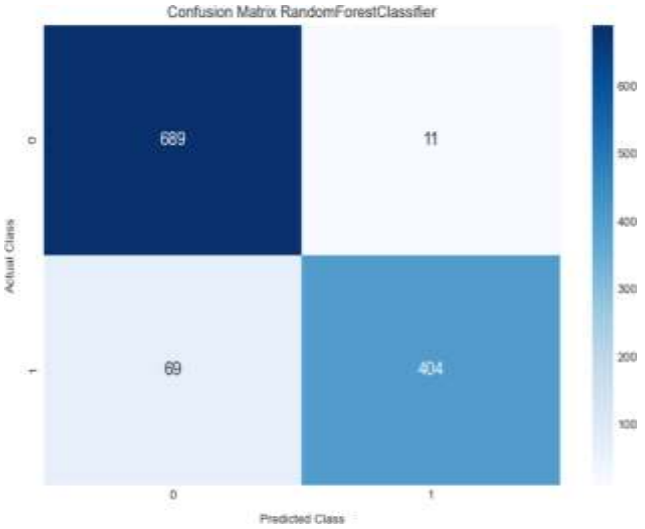

Figure 3: Confusion Matrix for RF

Here, the confusion matrix for Random Forest Classifier is shown. The True positive and True Negative values are not more accurate for the given dataset. The class zero consists of 700 data items. In that 689 is shown as True positive. The class one consists of 473 data items. In that 404 is shown as True Negative.

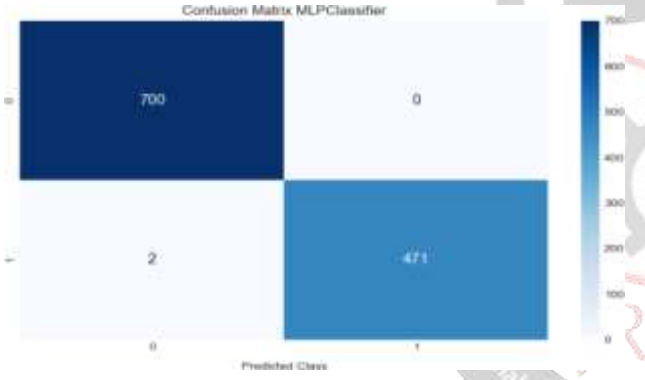

Figure 4: Confusion Matrix for MLP

Here, the confusion matrix for Multilayer Perceptron is shown. The True positive and True Negative values are accurate than Random Forest Classifier for the given dataset. The class zero consists of 700 data items. In that all 700 is shown as True positive. The class one consists of 473 data items. In that 471 is shown as True Negative.

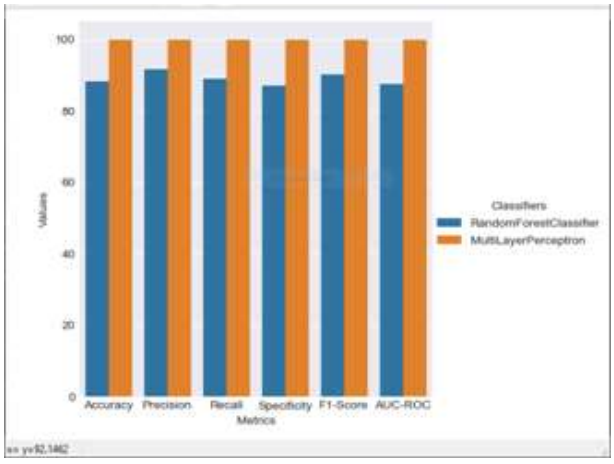

Figure 5: Graph Analysis of Algorithms

V. CONCLUSION

The Credit card fraud detection process has been a very interesting area of research for the Machine Learning researchers these years and it can also be a fascinating area of research within the coming future. This occurs due to continuous change of patterns in frauds. Optimal algorithms that address four main sorts of frauds were selected through literature, experiment the parameter tuning as shown within the methodology. In our work Multilayer Perceptron has given high accuracy and other performance metrics. However, when amount of knowledge is increased, there's some variation in performance metrics, thanks to data imbalance within the dataset. But the accuracy maintained high for MLP. Within the existing system the algorithms provides less efficiency and makes the training model slow. Hence within the proposed system we have compared Multilayer Perceptron and Random Forest. In this comparison Multilayer Perceptron has higher efficiency and it is comparatively increasing. The efficiency depends on the dataset and performance Metrix use for evaluation.

REFERENCES

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, G. N. Surname, "Random Forest for credit card fraud", *15th Int. Conf. Networking Sens. Control*, 2018.

[2] M. Zareapoor, S. K, K.R Seeja, M. Afshar Alam, "Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria", *Int. J. Comput. Appl.*, vol. 52, no. 3, pp. 35-42, 2012.

[3] D. S. Sisodia, N. K. Reddy, S. Bhandari, "Performance Evaluation of Class Balancing Techniques for Credit Card Fraud Detection", *IEEE Int. Conf. Power Control. Signals Instrum. Eng.*, pp. 2747-2752, 2017.

[4] Z. Zojaji, R. E. Atani, A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques : Data and Technique Oriented Perspective", pp. 1-26, 2016.

[5] J. O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", *2017 Int. Conf. Comput. Netw. Informatics*, pp. 1-9, 2017.

[6] R. Choudhary, H. K. Gianey, "Comprehensive Review On Supervised Machine Learning Algorithms", *2017 Int. Conf. Mach. Learn. Data Sci.*, pp. 37-43, 2017.

[7] A. C. Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten, "Cost sensitive credit card fraud detection using Bayes minimum risk", Machine Learning and Applications (ICMLA). 2013 12th International Conference, vol. 1, pp. 333-338,2013.

[8] A.Shen, R. Tong, Y. Deng, "Application of

 DOI : 10.35291/2454-9150.2020.0330 © 2020, IJREAM All Rights Reserved.

classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp.1-4,2007.

[9] S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, "Credit card fraud detection using Bayesian and neural networks", Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270, 2002.

[10] Khyati Chaudhary, Mallick Yadav, "Review of fraud detection techniques: credit card", *International Journal of Computer Applications*, vol. 45, no. 1, pp. 0975-8887,May2012.

[11] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, "Data mining for credit card fraud: A comparative study", *Decision Support Systems*, vol. 50,no.3,pp.602-613.

[12] Dighe, D., Patil, S., & Kokate, S. (2018). Detection of Credit Card Fraud Transactions Using Machine Learning Algorithms and Neural Networks: A Comparative Study. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)

[13] Aman Gulati et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 263 042039'Credit Card Fraud Detection using Neural Networks'.

 DOI : 10.35291/2454-9150.2020.0330 © 2020, IJREAM All Rights Reserved.