Universe International Journal of Interdisciplinary Research
(Peer Reviewed Refereed Journal)
© UIJIR | ISSN (O) – 2582-6417
SEP. 2020 | Vol. 1 Issue 4
www.uijir.com

Credit Card Fraud Detection using Machine Learning Algorithms

Author Name: Muhammad Zeeshan Younas
Affiliation: *Department of Computer Science, Capital University of Science& Technology, Islamabad, Pakistan.*
E-Mail: itszeeshanyounas@gmail.com

Abstract

Credit card fraud is a severe issue in financial services area. Every year billions of dollars are lost due to credit card fraud. Credit card has been one of the most flourishing financial services by banks over the past years. However, with the rising number of credit card users, banks have been facing an escalating credit card default rate. Credit card fraud is linked with the prohibited usage of credit card material for acquisitions. In this research work, various machine learning classification techniques and methods are used to analyzed and predict the accuracy of credit card fraud detection. Dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions. Thus, Logistic Regression, Multi-Layer Perceptron, Naïve Bayes and Random Forest are used to test the variable in predicting credit fraud and by the experimental outcomes and results it's evident that Random Forest algorithm predicts the credit card fraud detection with the accuracy of 99.95% and also with good precision rate 100%.

Keywords *: Machine Learning, Classification Techniques, Credit card default, WEKA, Prediction, Analysis*

INTRODUCTION

Credit card normally refers to a card that is allocated to the consumers (cardholder), generally consumer can acquisition goods and services through credit card within limit or withdraw cash from anywhere. Banks provide many facilities to customers thorough credit card. For instance, it provide consumer to pay later in a given time by carrying it to the subsequent next bill. Fraud is a criminal or illegal cheating that targeted to fetch financial or private advantage. In circumventing damage from fraud and there are the following two approaches can be implement: fraud detection and fraud prevention. Fraud detection is desired when a fake transaction is happened by a cheaters and Fraud prevention is an active technique, where i thalts fraud from trendy in the primarily place. There are giant data model has transfigured the baking systems and by altering the financial establishments run direction. The outcome of the historical financial disaster has been gradually remedying and people are nowadays improved for opportunities and financial system(Subbas and Lahiri, 2017).Credit card fraud is linked with the prohibited usage of credit card material for acquisitions. Credit card transactions can be skillful either physically or digitally, physical transactions of the credit card are complicated while doing the transactions and in digital transactions it can be happen over the internet and phones.

Fraud is very easy if targeted people have less knowledge about transactions and online money transfer. Hacker mostly attack on innocent people those don't have complete knowledge and usage about online banking applications. Normally cardholders typically deliver expiry date and card number, in case of forgets PIN number banks provide verification numbers respective telephone or email, (Randhawa, Kuldeep, et al 2018). Credit card fraud is easy marks with no

any risks, hacker withdraw amount without owner knowledge without bank's knowledge, they transfer amount in very short time of period and then escape from specific networks. Identify to fraudsters is not easy because they use very fast tool and masterminds those have all knowledge about owner's credits cards and financial transactions moments. Fraudsters continuously attempt to variety of each fake transaction legitimate, and it makes fraud detection actual stimulating and problematic job to identify attacks. The data analytics permitted banks to method the data ambitious commerce in an improved method to tackle the actual data generated customers. Credit card default prediction is core forecast that banks are a worry with involves credit counting to healthier comprehend why customers are probable to default. Banks want to each minor detail of the customers for tracking of payment data that is added in the credit history.

Credit card fraud recognition is a problematic issue that becomes the attention of Machine Learning researchers and scientists. Nevertheless, the issue is still challenging for credit card data which suffer from class inequity as no fraud transactions over powering succeed fraud transactions making it tough for numerous machine learning algorithms to achieve good accuracy and performance, a upright illustration can be erudite from the dataset that increase the classification performance of the machine learning techniques. Machine learning is a thinkable resolution to the challenge of credit fraud prediction because of its extraordinary feature learning aptitude in large and unstable datasets.

In this work, aim of this paper to classify and categorize a well understanding between the different kinds of machine learning techniques to detection of credit card fraud/default that are continuing presently happing in this modern era. In this work the author attempts to suggestion current machine learning techniques to gain improved performance outcomes. There are the following four machine learning techniques are used in this paper to predict the credit card fraud detection accuracy namely, Logistic Regression, Multilayer Perceptron, Naïve Bayes and Random Forest by using waikato environment for knowledge analysis (WEKA) tool. Dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions. This paper discussed the results of the modern methods and it will predict the result for credit card fraud.

LITERATURE REVIEW

Y. Sayjadah, et al(2020)they have used various machine learning techniques to predict credit card fraud in bank system that based on the analysis of the results. They have proposed random forest which has prediction accuracy is more than 80%. According to them banks can use

machine learning to measure credit risk of customers before surrendering them credit card. Banks main worry in to offer treasured harvests and facilities to their consumers and in order save up with their contestants they must stay advanced and creative.

Randhawa, Kuldeep, et al (2018)they have presented credit card fraud detection by using machine learning algorithms. Some typical models that are NB, SVM, and DL have used in the empirical study. They have proposed the best MCC score is 82% that is achieved by vote. Additional assess the hybrid mockups, noise from10% - 30% has been additional into the data models.

Sarah Alexandria Ebiaredoh-Mienye, et al (2020) machine learning algorithms are ineffectual for large datasets performing classification such as large credit card data set. They have proposed stacked sparse auto encoder network to gain optimal features learning. They have introduced batch normalization methods to increase the outcomes and speed of the model and further prevent over fitting. They model was optimized by using Adamax algorithm.

Somayeh Moradi, et al (2019) they have proposed a dynamic model for credit card fraud risk to valuation that outperformance the model used. There model has a self-motivated appliance that evaluates the behavior of corrupt clients in a once-a-month basis, credit risk that include the fuzzy factors, particularly in the financial crises. Their approach can utilize changing indeterminate issues.

Vaishnavi Nath Dornadula, et al (2019) they have used novel method to identify credit card fraud detection. Various classifiers are used on three altered collections advanced assessment scores are produced for each type of classifier. These self-motivated variations in strictures lead the organization to familiarize system. They have proposed that decision tree, random forest and logistic regression provided the best results and accuracy

METHODOLOGY

Data Source:

The data set is collected from Kaggle “Credit Card Fraud Detection [9]. This dataset presents transactions that happened in two days, this data have 492 frauds out of 284,807 transactions. The dataset is unstable i.e., 0.172% of all transactions and providing transaction details of a customer is considered to issue related to confidentiality, therefore are following features in the dataset are renovated by using principal component analysis (PCA). V1, V2, V3, … , V28 are PCA applied features and rest for instance: time, class and amount are non-PCA applied features that are presented in table 1.

Table 1: Attributes and Description of the some Dataset

No	Features	Description
1	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2	Class	Transaction amount
3	Amount	0 - not fraud 1 – fraud

Architecture Diagram:

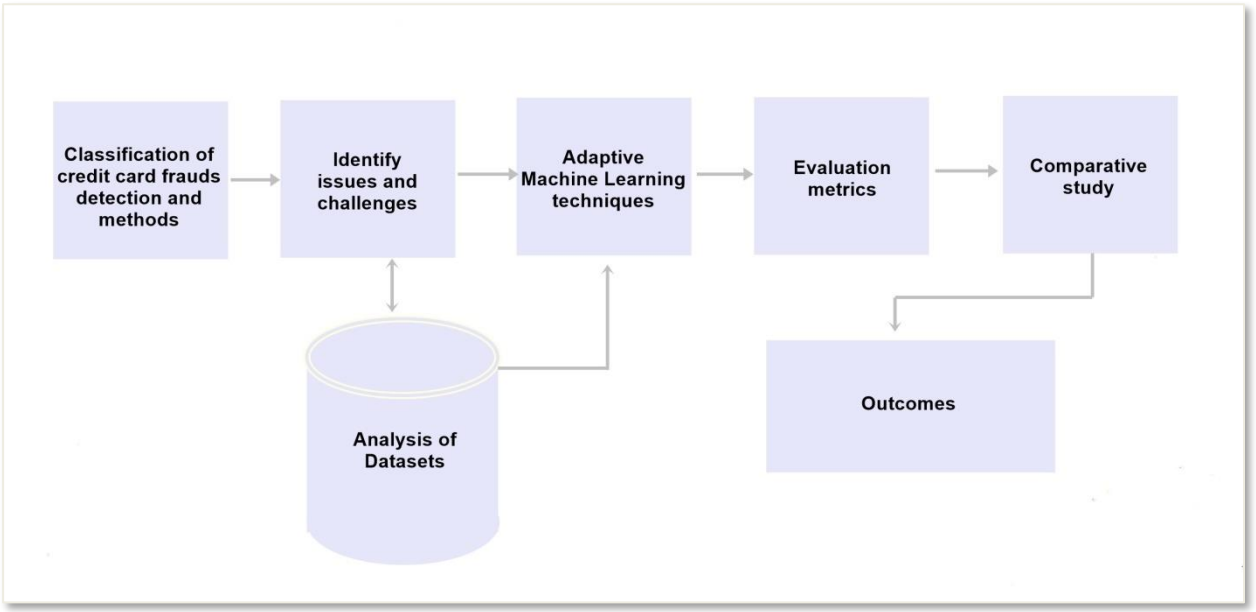


Figure 1 – Experiment workflow architecture diagram

Techniques:

Mainly there are four machine learning techniques for the prediction of credit card fraud detection:

i. Logistic Regression
ii. Multi-Layer Perceptron (MLP)
iii. Random Forest
iv. Naïve Bayes (NB)

i. Logistic Regression:

Logistic Regression is a useful method to assessment the possibility of a binary reaction based on single orextra variables (features). Logistic regression finds the best fit parameters to a nonlinear function that is known as sigmoid, it is the simple to implement and used to predict numeric value. It is an extension of the linear regression model for classification problems. (John O. Awoyemi, et al 2017). It cannot be used when the relation between independent and dependent variable are nonlinear.

ii. Multi-Layer Perceptron (MLP):

It is a feed-forward artificial neural network (ANN) and it has one extra layer that is called hidden layer. Hidden layer may be one and more than one and this layer holds the intermediary neurons. MLP uses the methods of backpropagation for training its many layers and its major benefit is individual data that is not linearly distinguishable. MLP is mostly used for handling problems that need supervised learning more likely parallel distributed processing and computational neuroscience. Application includes image recognition, speech recognition and machine translation.

iii. Random Forest:

Random forest is a supervised ML algorithm. It is based on collaborative learn model. Collaborative learn model isa category of learning where we can join various forms of techniques and algorithms or same algorithm numerous periods to form a additional influential prediction model. The random forest algorithm associations numerous algorithm of the similar kind i.e. many decision trees, subsequent in a forest of trees, hence the name "Random Forest". It can be used for both classification and regression(Devi Meenakshi, et al 2019).

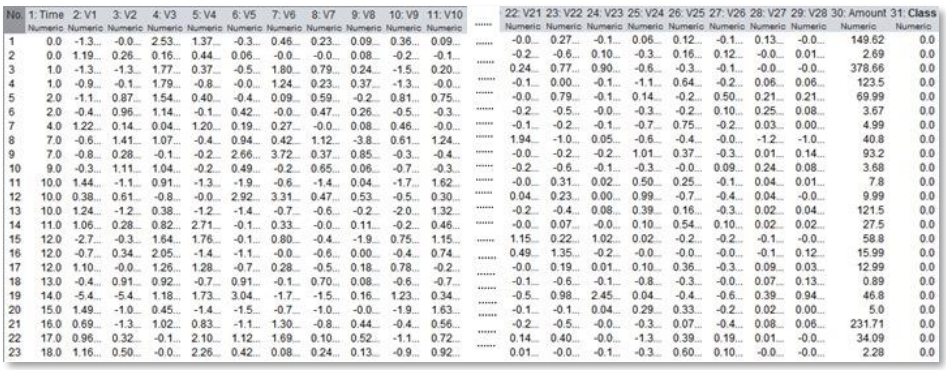
iv. Naïve bayes:

A naïve bayes classifier is a supervised ML algorithm that is used for classify the specific data into predefined classes. Naïve bayes classifier make use Bayes' theorem and it has independence between data points and attributes. Most popular uses of naïve bayes text analysis, spam filter and medical diagnosis. It uses conditional probability to classify the test dataset. Naïve bayes

requires insignificant amount of training dataset and fast to predict the classes of test data. It does not appropriate for large data and its working is not good if the features are associated.

EXPERIMENTAL RESULTS AND DISCUSSION

The proposed system has credit card fraud detection dataset which is used for classified whether the fraud is happened or not according to their features. The overall records in the dataset are distributed into two main category training and testing datasets. The proposed system applied on this data and tries to create accurate model which predict accurate results. In this proposed system, used Logistic Regression, Multi-Layer Perceptron, Random Forest and Naïve Bayes algorithms. By using WEKA tool has been calculated accuracy individually for each ML algorithm on the provided credit card fraud detection dataset. Finally analyze the outcomes by the help of Comparing Models and Confusion Matrix. In the field of machine learning, a confusion matrix, normally known as an error matrix, is a specific table design that authorizations perception of the implementation of a calculation. Each line of the matrix expresses to the instances in a predicted class while every section speaks to the cases in a real class. Firstly imported the respective dataset that comprises various variables. After the accessibility of the data, created a predictive model that is bases on Logistic Regression algorithm and this classified data based on various organized features of credit card fraud detection. After the complete data accessibility predicts the all four algorithms one by one and obtains their accuracy.



No.	1: Time	2: V1	3: V2	4: V3	5: V4	6: V5	7: V6	8: V7	9: V8	10: V9	11: V10	.....	22: V21	23: V22	24: V23	25: V24	26: V25	27: V26	28: V27	29: V28	30: Amount	31: Class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric		Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	0.0	-1.3...	-0.0...	2.53...	1.37...	-0.3...	0.46...	0.23...	0.09...	0.36...	0.09...	.....	-0.0...	0.27...	-0.1...	0.06...	0.12...	-0.1...	0.13...	-0.0...	149.62	0.0
2	0.0	1.19...	0.26...	0.16...	0.44...	0.06...	-0.0...	-0.0...	0.08...	-0.2...	-0.1...	.....	-0.2...	-0.6...	0.10...	-0.3...	0.16...	0.12...	-0.0...	0.01...	2.69	0.0
3	1.0	-1.3...	-1.3...	1.77...	0.37...	-0.5...	1.80...	0.79...	0.24...	-1.5...	0.20...	.....	0.24...	0.77...	0.90...	-0.6...	-0.3...	-0.1...	-0.0...	-0.0...	378.66	0.0
4	1.0	-0.9...	-0.1...	1.79...	-0.8...	-0.0...	1.24...	0.23...	0.37...	-1.3...	-0.0...	.....	-0.1...	0.00...	-0.1...	-1.1...	0.64...	-0.2...	0.06...	0.06...	123.5	0.0
5	2.0	-1.1...	0.87...	1.54...	0.40...	-0.4...	0.09...	0.59...	-0.2...	0.81...	0.75...	.....	-0.0...	0.79...	-0.1...	0.14...	-0.2...	0.50...	0.21...	0.21...	69.99	0.0
6	2.0	-0.4...	0.96...	1.14...	-0.1...	0.42...	-0.0...	0.47...	0.26...	-0.5...	-0.3...	.....	-0.2...	-0.5...	-0.0...	-0.3...	-0.2...	0.10...	0.25...	0.08...	3.67	0.0
7	4.0	1.22...	0.14...	0.04...	1.20...	0.19...	0.27...	-0.0...	0.08...	0.46...	-0.0...	.....	-0.1...	-0.2...	-0.1...	-0.7...	0.75...	-0.2...	0.03...	0.00...	4.99	0.0
8	7.0	-0.6...	1.41...	1.07...	-0.4...	0.94...	0.42...	1.12...	-3.8...	0.61...	1.24...	.....	1.94...	-1.0...	0.05...	-0.6...	-0.4...	-0.0...	-1.2...	-1.0...	40.8	0.0
9	7.0	-0.8...	0.28...	-0.1...	-0.2...	2.66...	3.72...	0.37...	0.85...	-0.3...	-0.4...	.....	-0.0...	-0.2...	-0.2...	1.01...	0.37...	-0.3...	0.01...	0.14...	93.2	0.0
10	9.0	-0.3...	1.11...	1.04...	-0.2...	0.49...	-0.2...	0.65...	0.06...	-0.7...	-0.3...	.....	-0.2...	-0.6...	-0.1...	-0.3...	-0.0...	0.09...	0.24...	0.08...	3.68	0.0
11	10.0	1.44...	-1.1...	0.91...	-1.3...	-1.9...	-0.6...	-1.4...	0.04...	-1.7...	1.62...	.....	-0.0...	0.31...	0.02...	0.50...	0.25...	-0.1...	0.04...	0.01...	7.8	0.0
12	10.0	0.38...	0.61...	-0.8...	-0.0...	2.92...	3.31...	0.47...	0.53...	-0.5...	0.30...	.....	0.04...	0.23...	0.00...	0.99...	-0.7...	-0.4...	0.04...	-0.0...	9.99	0.0
13	10.0	1.24...	-1.2...	0.38...	-1.2...	-1.4...	-0.7...	-0.6...	-0.2...	-2.0...	1.32...	.....	-0.2...	-0.4...	0.08...	0.39...	0.16...	-0.3...	0.02...	0.04...	121.5	0.0
14	11.0	1.06...	0.28...	0.82...	2.71...	-0.1...	0.33...	-0.0...	0.11...	-0.2...	0.46...	.....	-0.0...	0.07...	-0.0...	0.10...	0.54...	0.10...	0.02...	0.02...	27.5	0.0
15	12.0	-2.7...	-0.3...	1.64...	1.76...	-0.1...	0.80...	-0.4...	-1.9...	0.75...	1.15...	.....	1.15...	0.22...	1.02...	0.02...	-0.2...	-0.2...	-0.1...	-0.0...	58.8	0.0
16	12.0	-0.7...	0.34...	2.05...	-1.4...	-1.1...	-0.0...	-0.6...	0.00...	-0.4...	0.74...	.....	0.49...	1.35...	-0.2...	-0.0...	-0.0...	-0.0...	-0.1...	0.12...	15.99	0.0
17	12.0	1.10...	-0.0...	1.26...	1.28...	-0.7...	0.28...	-0.5...	0.18...	0.78...	-0.2...	.....	-0.0...	0.19...	0.01...	0.10...	0.36...	-0.3...	0.09...	0.03...	12.99	0.0
18	13.0	-0.4...	0.91...	0.92...	-0.7...	0.91...	-0.1...	0.70...	0.08...	-0.6...	-0.7...	.....	-0.1...	-0.6...	-0.1...	-0.8...	-0.3...	-0.0...	0.07...	0.13...	0.89	0.0
19	14.0	-5.4...	-5.4...	1.18...	1.73...	3.04...	-1.7...	-1.5...	0.16...	1.23...	0.34...	.....	-0.5...	0.98...	2.45...	0.04...	-0.4...	-0.6...	0.39...	0.94...	46.8	0.0
20	15.0	1.49...	-1.0...	0.45...	-1.4...	-1.5...	-0.7...	-1.0...	-0.0...	-1.9...	1.63...	.....	-0.1...	-0.1...	0.04...	0.29...	0.33...	-0.2...	0.02...	0.00...	5.0	0.0
21	16.0	0.69...	-1.3...	1.02...	0.83...	-1.1...	1.30...	-0.8...	0.44...	-0.4...	0.56...	.....	-0.2...	-0.5...	-0.0...	-0.3...	0.07...	-0.4...	0.08...	0.06...	231.71	0.0
22	17.0	0.96...	0.32...	-0.1...	2.10...	1.12...	1.69...	0.10...	0.52...	-1.1...	0.72...	.....	0.14...	0.40...	-0.0...	-1.3...	0.39...	0.19...	0.01...	-0.0...	34.09	0.0
23	18.0	1.16...	0.50...	-0.0...	2.26...	0.42...	0.08...	0.24...	0.13...	-0.9...	0.92...	.....	0.01...	-0.0...	-0.1...	-0.3...	0.60...	0.10...	-0.0...	-0.0...	2.28	0.0

Figure 2 – Reading the dataset

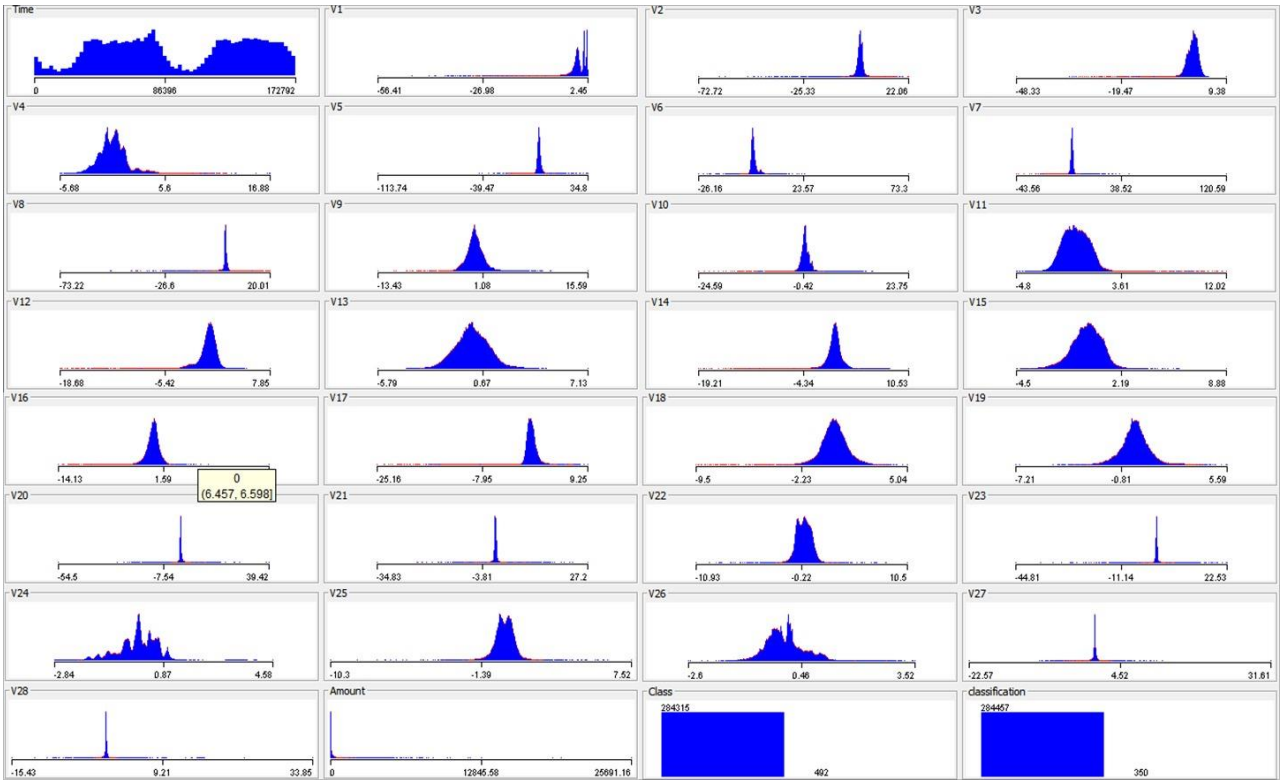


Figure 3 – Histogram of input variables in the credit card fraud dataset

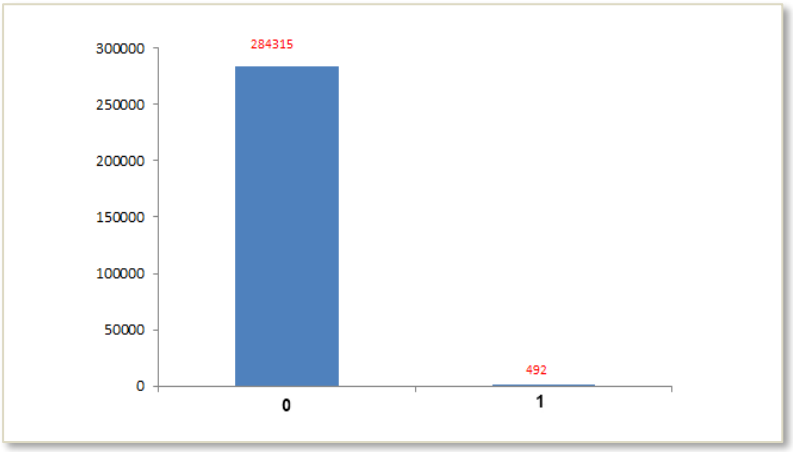


Figure 4 – Fixing targets about the fraud detection

The aim is to show the effectiveness of the following approach by performing a comparative study with four machine learning algorithms. Consequently, first show the performance of these classifiers on the raw dataset and the experiment results (including the accuracy, precision, f-measure, recall and ROC Area of each model) with the maximum accuracy, the maximum precision and the maximum f-measure accomplished by each machine learning technique and the grouping of features used in the model. The classifiers include Logistic Regression, Multi-

Layer Perceptron, Naïve Bayes and Random Forest, and the results are shown in Table 2 and Figure 5.

Table 2: Standard Metrics for 10-Fold Cross Validation Technique

Methods	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Accuracy (%)**
Logistic Regression	99	99	100	97	99.91
Multi-Layer Perceptron	99	79	99	95	99.94
Naïve Bayes	99	97	98	95	97.83
Random Forest	100	100	100	95	99.95

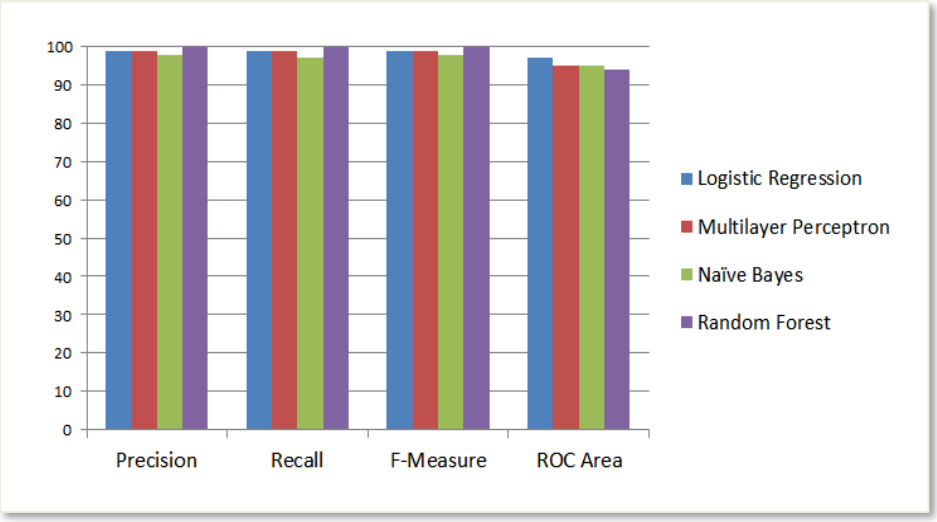


Figure 5 – Comparison of all techniques

Universe International Journal of Interdisciplinary Research
(Peer Reviewed Refereed Journal)

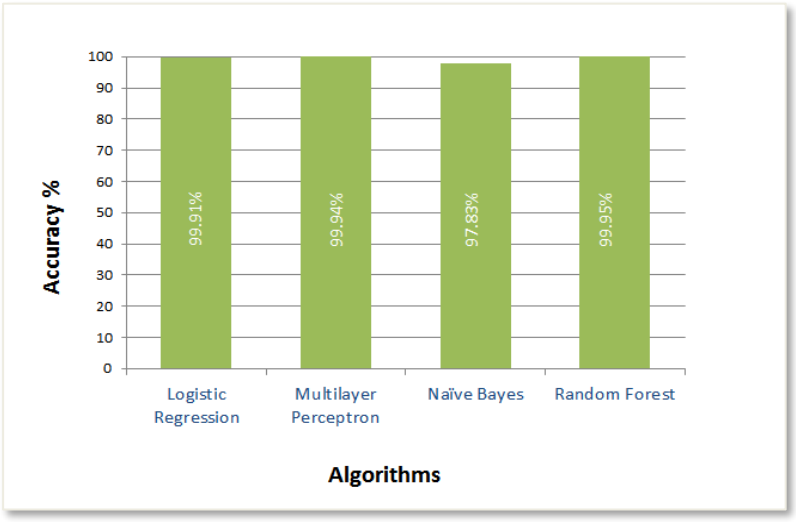


Figure 6 – Accuracy representation in graph withtheir respective algorithms

We can see that the highest accuracy (99.95%) was achieved by Random Forest with On the other hand, the highest precision(100%) was also achieved by Random Forest by using the same combination of dataset, there all four method treated with same dataset and we can see their all result. whereas the highest f-measure and recall (100%) was achieved by Random forest. Random forest provided us a good accuracy with efficient performance. The proposed method works best result 99.95% accuracy by using Random Forest algorithm, this work done by some different steps. The confusion matrix obtained by Four different algorithms is given below in the Figure 7.

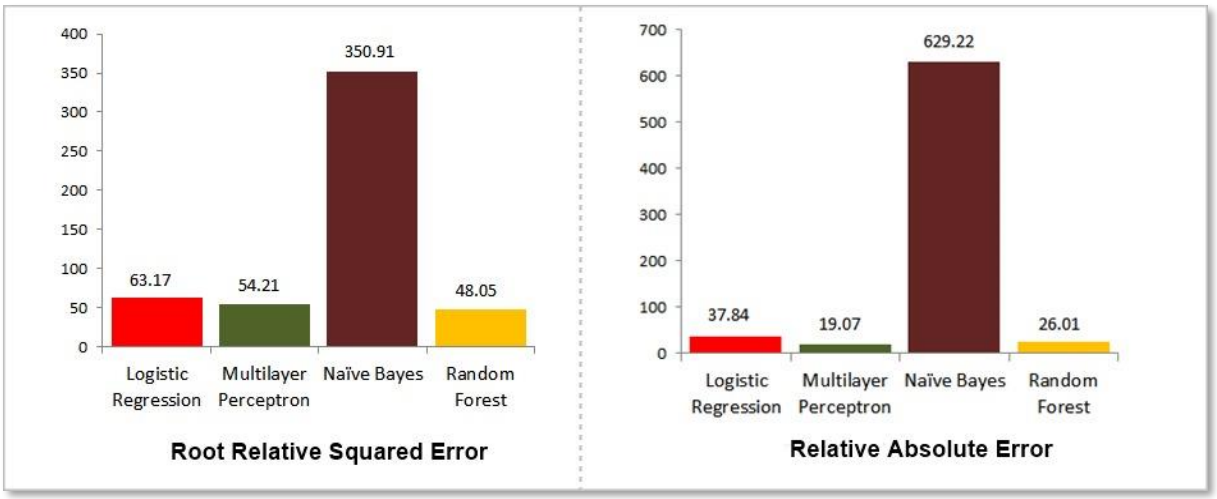


Figure 7 – Root Relative Squared Error Vs Relative Absolute Error

Figure 8 – Confusion matrix obtained using the classification algorithms

CONCLUSION

In this research work, various machine learning classification techniques and methods are used to analyzed and predict the accuracy of credit card fraud detection. Anti-fraud approaches can be adopted to prevent banks from major damages and minimize threats. The objective of the study was taken differently than the typical classification problems in that we had a variable misclassification cost. Four machine learning algorithms namely Logistic Regression, Multi-Layer Perceptron, Naïve Bayes and Random Forest are compared in terms of accuracy using the credit card fraud detection dataset. By the experimental outcomes and results it's evident that Random Forest algorithm predicts the credit card fraud detection with the accuracy of 99.95%

and also with good precision rate that is 100%.Banks can make the most of the machine learning techniques which can contribute in boosting their performance and image in the industry.

REFERENCES

1. Subba, N. and Lahiri, D. (2017) Rising credit card delinquencies tovadd to U.S. banks' worries. [online] Reuters. Available at:vhttp://Rising credit card delinquencies to add to U.S. banks' worriesv
2. Randhawa, Kuldeep, et al (2018)Credit Card Fraud Detection Using AdaBoost and Majority Voting. IEEE Access, vol. 6, pp 14277–14284 .doi:10.1109/access.2018.2806420.
3. Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran(2018). Credit Card Default Prediction using Machine Learning Techniques. Fourth International Conference on Advances in Computing, Communication Automation (ICACCA). , pp. 1–4, doi: 10.1109/ICACCAF.2018.8776802.
4. Sarah Alexandria Ebiaredoh-Mienye, Ebenezer Esenogho and (2020). Effective Feature Learning using Stacked Sparse Autoencoder for Improved prediction of Credit Card Default. Effective Feature Learning using Stacked Sparse Auto-encoder for Improved prediction of Credit Card Default.
5. Somayeh Moradi and Farimah Mokhatab Rafiei (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Moradi and Mokhatab Rafiei Financial Innovation, springer. doi.org/10.1186/s40854-019-0121-9
6. Vaishnavi Nath Dornadula and Gheeta S (2019) Credit Card Fraud Detection using Machine Learning Algorithms. International Conference On Recent Trends In Advanced Computing. Pp 631-641. doi- 10.1016/j.procs.2020.01.057
7. John O. Awoyemi, Adebayo O. Adetunmbi and Samuel A. Oluwadare (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. International Conference on Computing Networking and Informatics (ICCNI).DOI: 10.1109/ICCNI.2017.8123782
8. Devi Meenakshi, Janani , Gayathri (2019). Credit Card Fraud Detection Using Random Forest. International Research Journal of Engineering and Technology (IRJET). Vol 6(3).
9. https://www.kaggle.com/mlg-ulb/creditcardfraud