Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 8, Issue 04, April-2021

Smart Credit Card Fraud Detection Approach Using DM Techniques

Payal Boda[1], Prof. Dixita Kagathara[2]

[1]M.E. Scholar, Computer Engineering, Darshan Institute of Engineering & Technology, Rajkot
[2] Assistant Professor, Computer Engineering, Darshan Institute of Engineering & Technology, Rajkot

Abstract — *Due to the surge of intrigued in online retailing, the utilize of credit cards has been quickly extended in later a long time. Taking the card details to perform online exchanges, which is called extortion, has moreover seen more habitually. Preventive arrangements and moment extortion location methods are broadly considered due to basic monetary misfortunes in numerous industries. In this work, Naïve Bayes, D-TREE and Multiple Additive Regression Tree Classifier (MART) show for the detection of credit card fakes on the spilling transactions is explored with the utilize of diverse qualities of card transactions. I am applying Naïve Bayes, D-TREE and Multiple Additive Regression Tree Classifier (MART) algorithm to detect the CC fraud then compare the result with all algorithms for getting higher CC fraud accuracy.*

Keywords- *Data mining, Credit card fraud, Fraud detection, Naïve Bayes, D-TREE, Multiple Additive Regression Tree Classifier*

I. INTRODUCTION

Data mining aims to extract useful information from huge amount of data. In the process of data mining, large data sets are first sorted, then patterns are identified and relationships are established to perform data analysis and solve problems. The credit card payment system is one of the simplest payment methods and the most common type of financial transaction. However, it is observed that a good number of fraudulent credit card transactions are occurring. Credit card fraud means the unauthorized use of a credit card account. Thus, the fraud occurs when the third party starts unauthorized using of the credit card without the consent of the card owner.

These cards can be used for making purchases in both online and offline modes. Online credit purchases need customers to endorse payments by showing at the point of sale their personal identity numbers (PINs), while offline transactions need customers to sign purchase receipts. The method of Credit Card Fraud Detection (CCFD) mainly involves separating fraudulent financial details from genuine data. The models help to classify the pattern of fraud in the databases by applying machine learning algorithms. Various problems are associated with credit card fraud detection and hinder the direction of fraud detection, such as non-availability of real dataset, size of dataset, determining the appropriate evaluation parameters and complex actions of the fraudsters.

Generally, credit card fraud is classified as three types: Traditional card related frauds, Merchant related frauds and Internet frauds. Traditional card related frauds include application fraud, lost and stolen cards, account takeover and fake cards. Merchant related frauds include merchant collusion and triangulation. Internet frauds include site cloning, card generators and false merchant sites. In all these categories of fraud, the fraudsters obtain the information of the card without the knowledge of the cardholders and then use them for various fraudulent activities to steal the money from the account.

II. LITERATURE SURVEY

According to [1], the use of various card transaction attributes is searched with the Gradient Boosting Tree (GBT) model of the credit card fraud detection in real time on streaming Card-Not-Present transactions (CNP). To form a feature vector to be used as a training example, numerical, hand-crafted numerical, categorical, and textual attributes are merged. In this research, the main focus is two points such as character-level word embedding and sliding window-based automated training dataset generation technique. Character-level word embedding is necessary to map the name to a vector of real numbers and the name of the merchant can be used as a unique feature to detect fraudulent behavior. The sliding window-based automated training dataset generation technique is retrained by itself over time to prevent concept drift adaptively. Three features used in these experiment metrics like that encoded features, aggregated and encoded Feature, embedding, and aggregated and encoded feature. Experiments are evaluated by the following metrics; the False-Positive Rate (FPR), recall, precision, Area Under Curve (AUC). Sliding the training set increase the fraud detection performance in terms of AUC by 0.028%, and in fixed 0.3 FPR, Recall is improved by 0.029%.

According to [2], the behavior of frauds and legitimate transactions change constantly. Also, the issue with the credit card data is that it is highly skewed which leads to an inefficient prediction of fraudulent transactions. The three various proportions of datasets were used in this study and the random under-sampling technique was used for skewed dataset. The three machine learning algorithms used in this work are Logistic Regression, Naïve Bayes, and K-Nearest

Neighbour. The performance of these algorithms is recorded and analyze that how accurately they differentiate and classify the fraud and non-fraud transactions of the credit card dataset with random under sampling method (RUS) and to check out if the performance is improved or not. The analysis is done in python and the performance of the algorithms is calculated based on precision, sensitivity, specificity, accuracy, F-measurement, and area under curve. On the basis these measurements Logistic Regression (LR) showed the optimal performance for all the data proportions as compared to Naïve Bayes (NB) and K-Nearest Neighbour (KNN).

According to [3], in a relatively small-time frame, which can range from micro to milliseconds, the mechanism of acceptance or denial of a transaction occurs and a large number of related forms of transactions occur at the same time. Therefore, to be able to distinguish between a legitimate and a fraud transaction, an appropriate Fraud Detection Mechanism must be put into work. In this paper, they used an imbalanced dataset to check the suitability of various supervised machine learning models to forecast the probability of occurrence of a fraudulent transaction. They used sensitivity, precision, and time as the deciding parameters to come to a clear conclusion. Accuracy has not been used as a parameter as it is not susceptible to imbalanced data and does not have a definitive answer. They used kNN, Naive Bayes, Decision Tree, Logistic Regression and Random Forest models for predicting the chances of occurrence of a fraudulent credit card transaction out of a given number of transactions and the analysis shows that the sensitivity of the kNN model is greater than that of Decision tree, but as time taken by kNN for testing the data is very large. To ensure that minimal time is required for prediction in the case of fraud identification, so the preferred model is the Decision Tree.

According to [4], classifier ensembles are used successfully in either data mining or data stream mining to increase the performance of single classifiers. This paper therefore proposes an Online Boosting (OLBoost) approach, which first uses the Extremely Fast Decision Tree (EFDT) as a base (weak) learner, in order to assemble them into a single strong online learner, to achieve great success in prediction with virtually no increasing memory and time costs.

According to [5], sometimes the learning models used by them are too weak to fit the large scale of data. This paper extends the fraud detection method and uses lightgbm to propose a detection algorithm. Used by many data scientists to achieve state-of-the-art results to solve many machine learning problems, it is a scalable end-to-end tree boosting method. It also implemented other classical machine learning models in this task like SVM, logistic Regression, and Xgboost. They used it to tune some key parameters like learning rate, number of estimators, a sample rate of rows, sample rate of columns, max depth of each tree, and boosting types. Experiments showed that the lightgbm model outperformed the other Logistic Regression, SVM, and Xgboost models on both Auc-Roc score and accuracy.

According to [6], this paper proposes an intelligent approach using an optimized light gradient boosting machine (OLightGBM) to detect fraud in credit card transactions. A Bayesian-based hyperparameter optimization algorithm is intelligently implemented in the proposed approach to change the parameters of a light gradient boosting machine (LightGBM). Experiments were carried out using two real-world public credit card transaction data sets consisting of fraudulent transactions and legitimate ones to demonstrate the efficacy of our proposed OLightGBM for detecting fraud in credit card transactions. The output of the proposed method was evaluated by comparison with other research findings and state-of-the-art machine learning algorithms, including random forest, logistic regression, the radial support vector machine, the linear support vector machine, kNN, decision tree, and naive bayes, based on a comparison with other approaches using the two data sets. The experimental results show that the method proposed outperformed the other machine learning algorithms and obtained the highest accuracy performance (98.40%), Area under receiver operating characteristic curve (AUC) (92.88%), Precision (97.34%), and F1-score (56.95%).

According to [7], the information of the credit card fraud detection are highly skewed or unbalanced. Due to this class imbalance problem, several generic machine learning algorithms such as Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Logistic Regression (LR) have been applied to a balanced dataset with different sampling techniques such as Oversampling, Undersampling, Both sampling, ROSE and SMOTE. In reality, these algorithms have been effective in correctly predicting non-fraudulent transactions rather than fraudulent ones. The minority class of fraudulent exchanges is ignored as noise-based exchanges. The work effectively handles this misclassification by altering the raw data provided to the classification algorithms to correctly forecast the minority class, not the majority class. According to the experiments, SMOTE sampling is better conducted on the basis of its system of synthetic sampling rather than the nearest values. Among the five methods used, logistic regression performs well with 97.04% accuracy and 99.99% precision. SMOTE sampling along with logistic regression is also recommended for the detection of credit card fraud.

According to [8], five standard machine learning classification models from heterogeneous family such as K-Nearest Neighbor (K-NN), Extreme Learning Machine (ELM), Random Forest (RF), Multilayer Perceptron (MLP), and Bagging classifier are investigated and compared with each other. An ensemble of machine learning algorithms with majority voting yields a better hybridized model that can correctly classify the fraudulent and non-fraudulent transactions. An ensemble of the machine learning algorithm is one of the novel approach for the credit card fraud detection technique. Performance parameters are measured for accuracy, sensitivity, specificity, precision, F1-Score, and Matthews Correlation Coefficient. The prediction accuracy of the proposed model is observed to be 83.83%, which is significantly improved as compared to other single classification models. The error of detection of fraud for the proposed model is decreased and the rate of prediction of fraud is increased.

III. DATASET DESRIPTION

The dataset is acquired from the Kaggle which hosts the dataset from credit card fraud detections. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, …,V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. [10]

Time	V1	V2	…	V28	Amount	Class
0	-1.35981	-0.07278	…	-0.02105	149.62	0
0	1.191857	0.266151	…	0.014724	2.69	0
1	-1.35835	-1.34016	…	-0.05975	378.66	0
1	-0.96627	-0.18523	…	0.061458	123.5	0
2	-1.15823	0.877737	…	0.215153	69.99	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .
172786	-11.8811	10.07178	…	0.823731	0.77	0
172787	-0.73279	-0.05508	…	-0.05353	24.79	0
172788	1.919565	-0.30125	…	-0.02656	67.88	0
172788	-0.24044	0.530483	…	0.104533	10	0
172792	-0.53341	-0.18973	…	0.013649	217	0

[284807 ROWS × 31 COLUMNS]

Table 1. Credit card dataset containing V1 to V28 columns, Time and Amount.

In the dataset of credit cards there are two values for classification of transactions which means that it is a binary classification problem where transactions are classified either as fraud (1) or non-fraud (0). Figure 1 demonstrate the Non Frauds 284,315 (99.83 %) of the dataset and Frauds 492 (0.17 %) of the dataset

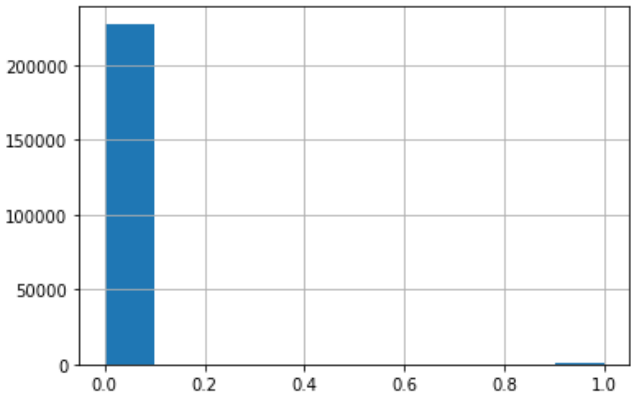

Figure 1. Class Distributions (0:Non Fraud, 1:Fraud) of original dataset

IV. VARIOUS CLASSIFICATION MODELS AND RESAMPLING METHOD

Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm. It is a form of probabilistic classifier model. It is a probabilistic classifier model, which means it can make predictions for several classes at the same time. It is based on the Bayes Theorem. Probabilistic Classifiers are those which make it possible to predict multiple classes. The decision is made based on conditional probability.[3]

Decision Tree

Decision Tree algorithm is a supervised learning algorithm. This is one of the most commonly used approaches to predictive modelling. As the name implies, the model is built in the form of a tree. In the case of a multi-dimensional analysis with multiple groups, this model may be used. The past data (also known as the past vector) is used to construct a model that can predict the output value based on the input. A tree has several nodes, each of which corresponds to one of two vectors. The tree comes to a halt at a leaf node, each of which represents a potential outcome or output. By learning basic decision rules inferred from prior data (training data), a Decision Tree can be used to construct a training model that can be used to predict the target variable's class or value.

Multiple Additive Regression Trees

MART (Multiple Additive Regression Trees) is a predictive data mining implementation of gradient tree boosting methods (regression and classification). MART is one of a group of techniques known as boosting. Boosting is a general method for improving the accuracy of any learning algorithm by fitting a collection of low-error models and then combining them to create a high-performing ensemble. MART fits a set of very simple classification trees, each requiring just a small amount of computational effort. The MART classifier is then based on a linear combination of these trees.

Synthetic Minority Oversampling Technique

SMOTE, or Synthetic Minority Oversampling Technique, is an oversampling technique that differs from standard oversampling. The minority data is duplicated from the minority data population in a traditional oversampling technique. [11] Although it increases the amount of data available, it does not provide the machine learning model with any new information or variation. SMOTE generates synthetic data using the k-nearest neighbour algorithm. SMOTE begins by selecting random data from the minority class, after which the data's k-nearest neighbours are determined. The synthetic data will be generated by combining the random data with the randomly selected k-nearest neighbour. The process is repeated until the minority class and the majority class have the same proportion. Below figure 2 is the example of SMOTE.

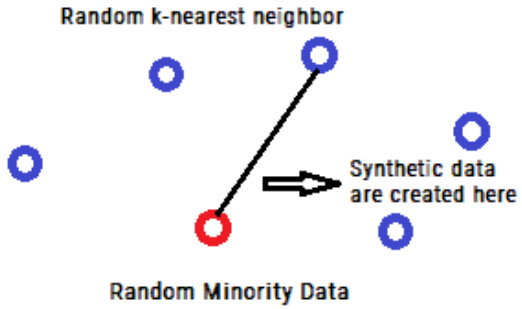

Figure 2. SMOTE [11]

Figure 3 demonstrate the applied SMOTE technique after Class =0 and Class=1 transactions are of equal ratio.

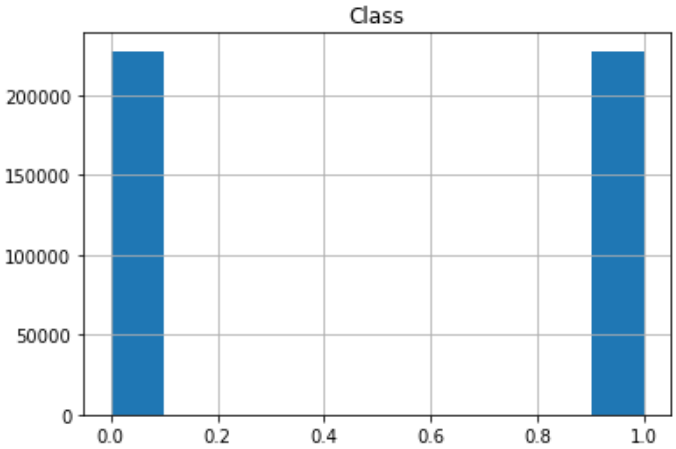

Figure 3. Equally Distributed Classes

V. PROPOSED SYSTEM

Proposed Flowchart

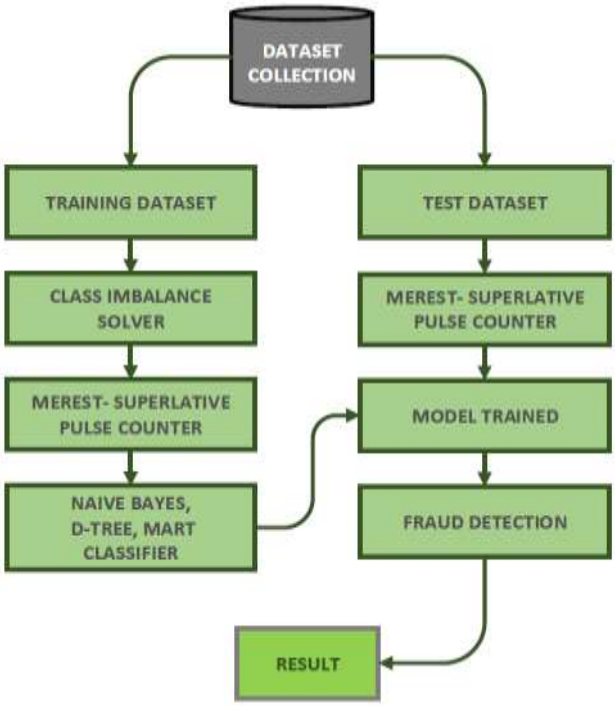

Figure 4. Proposed Flowchart Method

Proposed Algorithm

BEGIN
Step 1: Take input from Dataset
Step 2: Data-preprocessing from Dataset
Step 3: Divide Training and Testing data from Dataset
Step 4: Utilize class imbalance solver technique on Dataset
Step 5: Apply Merest – Superlative pulse counter on Dataset
Step 6: Train Model using Naïve Bayes, D-TREE and Multiple Additive Regression Tree (MART) Classifier algorithm
Step 7: Model Trained
Step 8: Fraud Detection
Step 9: Result
End

Description of Proposed Approach

- It starts with the data collection; here in this step, the collected input data is in the form of csv files.
- A process to gather context to the input data. Understanding the data for preprocessing and cleaning of datasets. The two columns 'amount' and 'time' were not normalized. The remaining columns were normalized using Principal Component analysis.
- Dataset then divided into training dataset and test dataset among them 80% of the data will be used for training the model while rest 20% will be used for testing the model, which will be highly skewed or imbalanced.
- Now utilize class imbalance solver technique SMOTE on dataset, which is used to balance class distribution by randomly increasing minority class examples by replicating them.
- Then apply Merest – Superlative pulse counter on dataset. It is normalize the data for every feature the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1 and every other value gets transformed into decimal between 0 and 1.

- After the data segregation, the data are fed into machine learning algorithm like Naïve Bayes, D-TREE and Multiple Additive Regression Tree (MART) Classifier. This step is mainly done using training data to teach the machine to increase its predictive accuracy.
- Once the data have learnt enough, our learned model will be ready for testing.
- The learned model is tested using test data to check its predictive accuracy.
- If the predictive accuracy is up to the desired level, then the model is deployed.

VI. PERFORMANCE EVALUTION

Performance evaluations were done for the three different classification techniques namely NB, D-TREE and MART for the SMOTE technique used. The four elementary matrices through which performance evaluations are predicted are as: True Positive (TP), True negative (TN), False positive (FP) and false Negative (FN).

- The transaction cases which are not fraud and the system model has predicted as not fraud as True Positive(TP).[9]
- The transaction cases which are fraud and the system model has predicted as fraud as True Negative(TN).[9]
- The transaction cases which are fraud and the system model has predicted as not fraud as False Positive(FP).[9]
- The transaction cases which are not fraud and the system model has predicted as fraud as True Neagtive(TN).[9]

Accuracy:

It is defined as the ratio of total number of predicted transactions that are correct.[2]

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

Precision:

The proportion of positive observed values correctly predicted as positive. It is also called as True Positive Rate (TPR). [2]

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

Recall:

The proportion of positive (fraud) predictions that are actually correct. [2]

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F-measure:

F-measure gives the accuracy of the test which means that it gives the accuracy of experiments performed. It uses the both precision and recall to compute its value. The best value for f1 score is considered at value 1. [2]

$$\text{F-measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

Confusion Matrix of Credit Card Dataset [2]

	Predicated Fraud	**Predicated Non Fraud**
Actual Fraud	TP	FP
Actual Non Fraud	FN	TN

Table 2. Confusion Matrix

VII. COMPARISON & RESULT

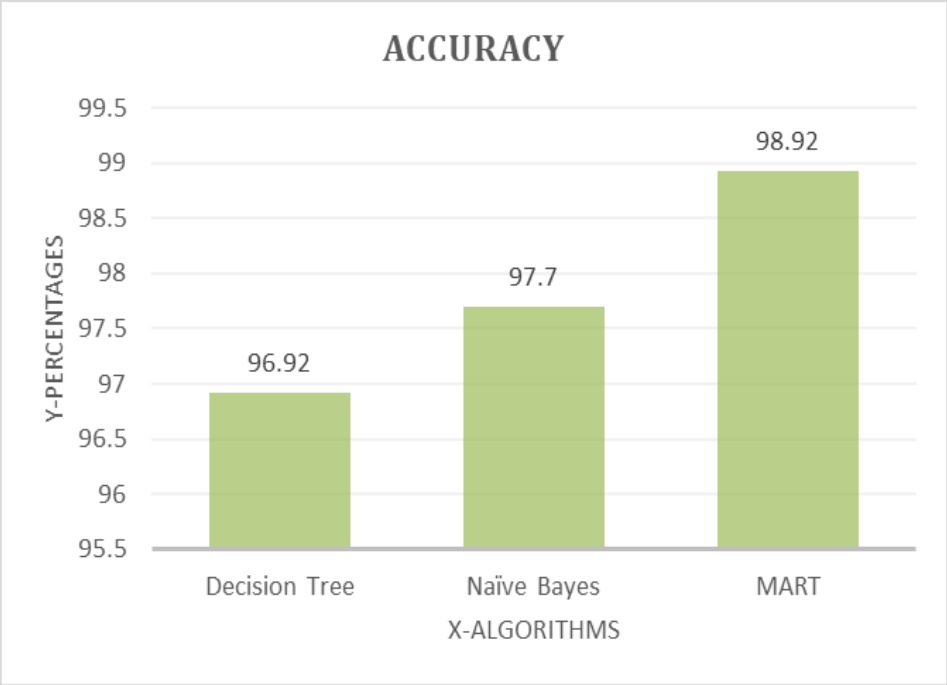

Figure 5. Comparison of Algorithms Base on Accuracy

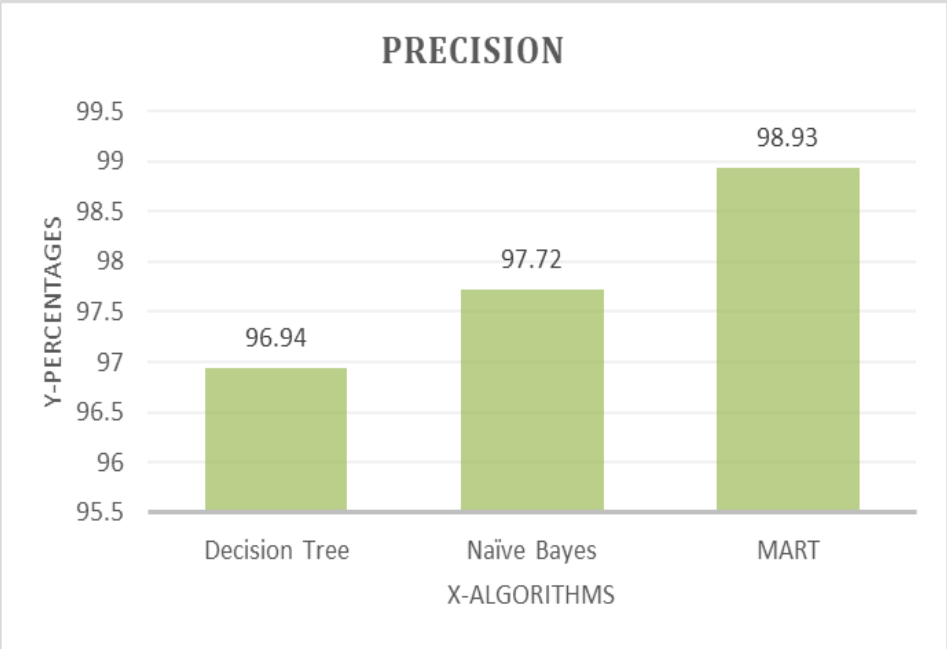

Figure 6. Comparison of Algorithms Base on Precision

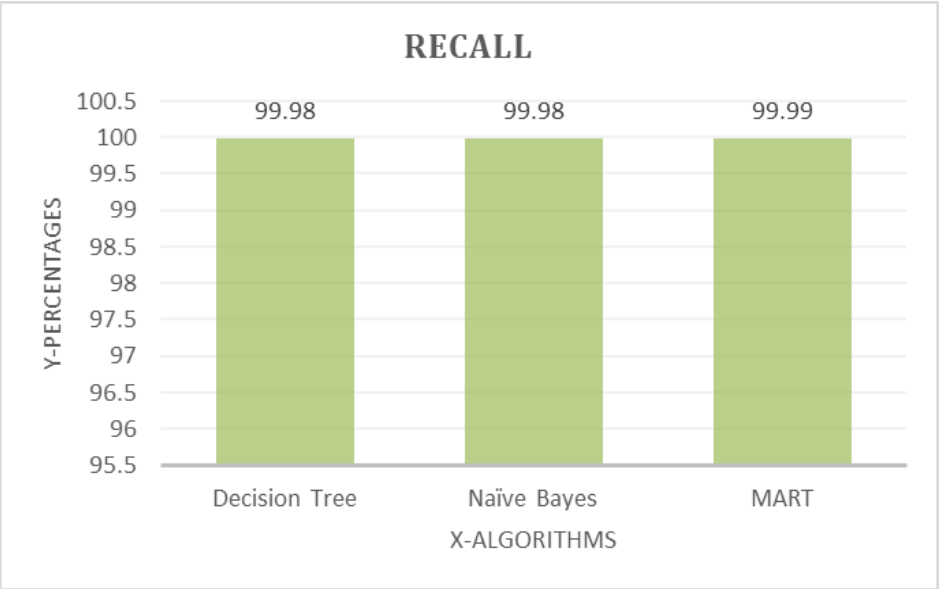

Figure 7. Comparison of Algorithms Base on Recall

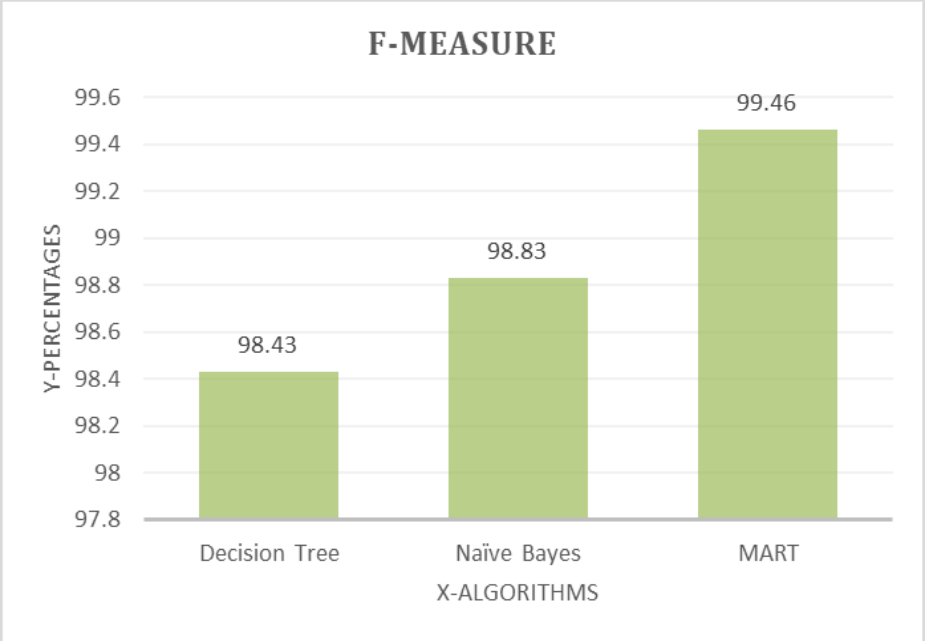

Figure 8. Comparison of Algorithms Base on F-Measure

Algorithms	Accuracy	Precision	Recall	F-Measure
MART	**98.92**	**98.93**	**99.99**	**99.46**
Naïve Bayes	97.70	97.72	99.98	98.83
Decision Tree	96.92	96.94	99.98	98.43

Table 3. Performance Results

The results demonstrated here are obtain by using SMOTE technique, so the findings provided here are based on SMOTE strategy. Performance of these three algorithms such as MART, Naïve Bayes and Decision Tree are compare based on their accuracy, precision, recall and f-measure. The above figure 5 indicates that among of these three algorithms, MART have achieved highest accuracy of 98.92% as compare to Naïve Bayes and Decision Tree algorithms and figure 6,

figure 7 and figure 8 also indicates that compare to other two algorithms, MART achieves the better result in other parameters such as precision, recall and f-measure. Table 3 illustrate the performance results of these three algorithms.

VIII. CONCLUSION AND FUTURE WORK

The research work was carried out with the purpose of comparing the ability of machine learning algorithms as to how accurately they differentiate and classify the fraud and non-fraud transactions of the credit card dataset with SMOTE technique and to check out if the performance is improved or not. Multiple Additive Regression Trees (MART) showed the optimal performance for all the data proportions as compared to Naı¨ve Bayes (NB) and Decision Tree (D Tree). MART was successful in getting higher accuracy as compared to Naı¨ve Bayes and Decision Tree. The MART showed the maximum accuracy of 98.92%, NB showed 97.70% and D Tree showed 96.92%. Also MART shows the better Precision, Recall and F-Measure as compare to NB and D-Tree technique.

In the future, there can be other resampling methods, which could be put to application for the skewed dataset for credit card fraud detection. These methods could be improved to achieve better results. Also using our statistics could be compared with the other techniques like Logistic Regression, K-Nearest Neighbour, Random-Forest, Support Vector Machine, and Neural Network.

IX. REFERENCES

[1] Ali Ye, silkanat(B), Barı, s Bayram, Bilge K¨oro˘glu, and Se, cil Arslan, “An Adaptive Approach on Credit Card Fraud Detection Using Transaction Aggregation and Word Embeddings” © IFIP International Federation for Information Processing 2020.

[2] Fayaz Itoo, Meenakshi, Satwinder Singh, “Comparison and analysis of logistic regression, Naı¨ve Bayes and KNN machine learning algorithms for credit card fraud detection” © Bharati Vidyapeeth’s Institute of Computer Applications and Management 2020.

[3] Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal, “Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison” © 2020 IEEE.

[4] Aye Aye Khine, Hint Wint Khin, “Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree” © 2020 IEEE.

[5] Dingling Ge, Shunyu Chang, “Credit Card Fraud Detection Using Lightgbm Model” © 2020 IEEE.

[6] Altyeb Altaher Taha, Sharaf Jameel Malebary, “An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine ” © 2020 IEEE.

[7] J. V. V. Sriram Sasank, G. Ram sahith, K.Abhinav, Meena Belwal, “Credit Card Fraud Detection Using Various Classification and Sampling Techniques: A Comparative Study” Proceedings of the Fourth International Conference on Communication and Electronics Systems (ICCES 2019).

[8] Debachudamani Prusti, Santanu Kumar Rath, “Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning techniques”, 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT 2019).

[9] Kaithekuzhical Leena Kurien, Dr. Ajeet Chikkamannur, “Detection And Prediction Of Credit Card Fraud Transactions” © International Journal of Engineering Sciences & Research Technology (IJESRT 2019)

[10] https://www.kaggle.com/janicechen1280/creditcard

[11] https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5

