International Journal of Future Generation Communication and Networking
Vol.14, No. 1, (2021), pp. 2567–2575

Credit Card Fraud Detection Using Machine Learning

[1]Dr.Vidhya.K, [2]Subhashree.M.S, [3]Swetha.S

[1]Associate Professor Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore-641022, vidhya.k@srec.ac.in
[2]Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore-641022, subhashree.1702224@srec.ac.in
[3]Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore-641022, swetha.1702235@srec.ac.in

Abstract

Frauds are the biggest threats in our day-to-day life is Fraud. Nowadays frauds are spreading all over the world and due to this financial loses are increasingly high in all sectors. Malicious behavior may be a broad term together with delinquency, fraud, intrusion, and account defaulting. Many banks find it difficult to detect the fraud in credit card system. The fraud detection performance in credit card transactions is highly affected by the sampling approach on data-set, variable selection and detection techniques. In order to resolve this problem this project comes with the fraud detection techniques that could bring a solution to this problem with a predictive analysis. A predictive model is being developed for credit card fraud detection using Spyder IDE and various machine learning algorithms are applied for the best identification of the frauds with the best accuracy for the dataset taken. The methods used in fraud detection evaluates a new hybrid approach to identify fraud detection. The credit card fraud detection is employed using machine learning algorithm namely Decision Tree, logistic regression, Random Forest and Support Vector Machine. To make the learning process efficient, we used Principal component for feature selection and a comparison is made on different machine learning algorithms with different parameters for better accuracy.

Keywords: Credit card, Fraud detection, Decision Tree, Logistic Regression, Random forest, Support Vector Machine, Spyder IDE.

I. INTRODUCTION

Nowadays most well- liked mode of payment is using credit cards in our day-to-day life. The number of credit card users are rising world-wide, that additionally results in increase in thefts and frauds also are increasing. This project proposes a credit card fraud detection system using supervised machine learning algorithms. The credit card information is confidential, the bank and the other financial enterprises doesn't want to disclose the information about the customers. Machine learning methods have contributed very well in the advancement of prediction results. These models can produce prediction results for fraud detection for higher accuracy. The Python Programming language is developed using the Spyder IDE tool for the predictive model of fraud detection. The data's collected are analyzed and pre-processed before it is used for model training and testing. Some of the machine learning algorithms were implemented such as Decision tree, Random Forest, Logistic Regression and SVM to predict the accuracy of the frauds. In the proposed model 80 percentage of data has been trained and 20 percentage of the input data has been tested. Here, the Principal Component Analysis technique is used reduce the dimensionality large data sets by transforming a large set of variables into the smaller ones that still contain the information in the large dataset. The accuracy is predicted on comparing with these machine learning algorithms. Thus, on comparing with different machine learning algorithms we tend to build a conclusion that supported the performance of different parameters such as MSE, RMSE, MAE values, the accuracy is predicted and the effectiveness of the machine learning model.

ISSN: 2233-7857 IJFGCN

II. LITERATURE SURVEY

For the detecting the fraud in the credit card ,one-of-a-kind literature survey was carried out. For the best results, researchers looked at a variety of literature surveys from various journals and conference papers. Battacharyya (2017) proposed that the mastercard fraud detection may be a serious growing drawback. The prophetical models for mastercard fraud detection square measure in active use in observe, reported studies on the utilization of information mining approaches for credit card fraud detection are comparetively low, presumable because of lack of accessible information for analysis. This paper evaluates two advanced data mining approaches to detect the better accuracy and the study is predicted on the important life information transactions. Shiyang Xuan (2016) used two forms of random forests that train the behavior options of traditional and abnormal transactions. The research worker compares these two random forests which are differentiated on the premise of their classifiers, performance on the detection of credit card fraud. The data used is of an associate e-commerce company of China that is employed to research the performance of the two types of random forests model. During this paper, the author used B2C dataset for the identification and detection of fraud from the credit cards. Therefore, the research worker completes from the result that the proposed random forests give smart results on small dataset however there are still some issues like unbalanced data that makes it less effective than the other dataset. Sahin and S. Bulkan, et al. (2017) proposed the developments within the information technology, fraud is spreading everywhere the planet, leading to vast monetary losses. Although fraud hinderance mechanisms are developed for credit card systems, these mechanisms don't stop the foremost common fraud varities like dishonorable credit card usages over virtual POS terminals or mail orders so called online credit card fraud. As a result, fraud detection becomes the essential tool and possibly the most effective way to stop such fraud sorts. During this study, a new cost-sensitive decision tree approach that minimizes the total of misclassification prices whereas choosing the cacophonic attribute at every non-terminal node is developed and therefore the performance of this approach is compared with the well-known ancient classification models on a real-world credit card data set. The results show that the cost-sensitive decision tree algorithmic rule outperforms the prevailing well-known strategies on the given drawback set with relevancy to the well-known performance metrics like accuracy and true positive rate, however conjointly a fresh out-lined cost-sensitive metric specific to credit card fraud detection domain. Consequently, the monetary losses thanks to fallacious transactions are often attenuated additional by the implementation of this approach in fraud detection systems. Ekrem DuMan (2015) published an associate that was based on genetic algorithm and scattering search. In this approach, every dealing is scored and supported based on these score transactions are divided into deceitful or legitimate transactions. They centered on an answer to reduce the incorrectly classified transactions. They merge the Meta heuristic approaches scatter search and genetic algorithmic rule. Soltani Halvayi (2014) The amount of online transactions is growing lately to an outsized variety. An enormous portion of these transactions contains credit card transactions. The expansion of the online fraud, on the opposite hand, is notable, that is mostly a result of simple access to edge technology for everybody. There has been analysis done on several models and ways for credit card fraud prevention and detection. Artificial Immune Systems is one in all them. However, organizations want accuracy beside speed with the fraud detection systems, which is not completely gained yet. Research workers address credit card fraud detection exploitation using Artificial Immune Systems (AIS), and introduce a whole new model known as AIS-based Fraud Detection Model (AFDM). We will use an associate system galvanized algorithmic rule and improve it for fraud detection. We tend to increase the accuracy up to 25%, reduce the cost up to 85%, and reduce the system latent period up to 40% compared to the base algorithm. Chen (2018) The purpose of this study is to construct a legitimate and rigorous deceitful finances detection model. The objective analysis firms that each deceitful and non-fraudulent monetary statements between the years 2002 and 2013. In the initial stage, two decision tree algorithms, together with the classification and regression trees (CART) and therefore the Chi square automatic interaction detector (CHAID) area unit applied within the choice of major variables. The second stage combines CART, CHAID, Bayesian belief network, support vector machine and artificial neural network so as to construct deceitful finances detection models.

ISSN: 2233-7857 IJFGCN

M.Puh and L.Brkic (2019) presents a comparison of three supervised machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR), 3 they used a dataset that contains credit card transactions made by European cardholders for two days in September 2013. Moreover, the dataset consists of 284,807 samples (transactions), it has 492 fraudulent transactions, 31 features; 28 numerical input variables are a result of Principal Component Analysis (PCA) transformation made by dataset provider, and two non- transformed variables, and finally the class feature. However, the challenges that appear in fraud detection system are highly dealing with an imbalanced dataset, and a non-static environment to overcome these challenges, this paper used SMOTE method, and ensembles and adaptive base learner, respectively.

III. METHODOLOGY

In this section, we'll go through the architecture of the system. On a wide scale, we use Machine Learning to detect the frauds in credit card. Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics. There are some tasks that humans perform effortlessly or with some efforts, but we are unable to explain how we perform them. For example, we can recognize the speech of our friends without much difficulty. If we are asked how we recognize the voices, the answer is very difficult for us to explain. Because of the lack of understanding of such phenomenon (speech recognition in this case), we cannot craft algorithms for such scenarios. Machine learning algorithms are helpful in bridging this gap of understanding.

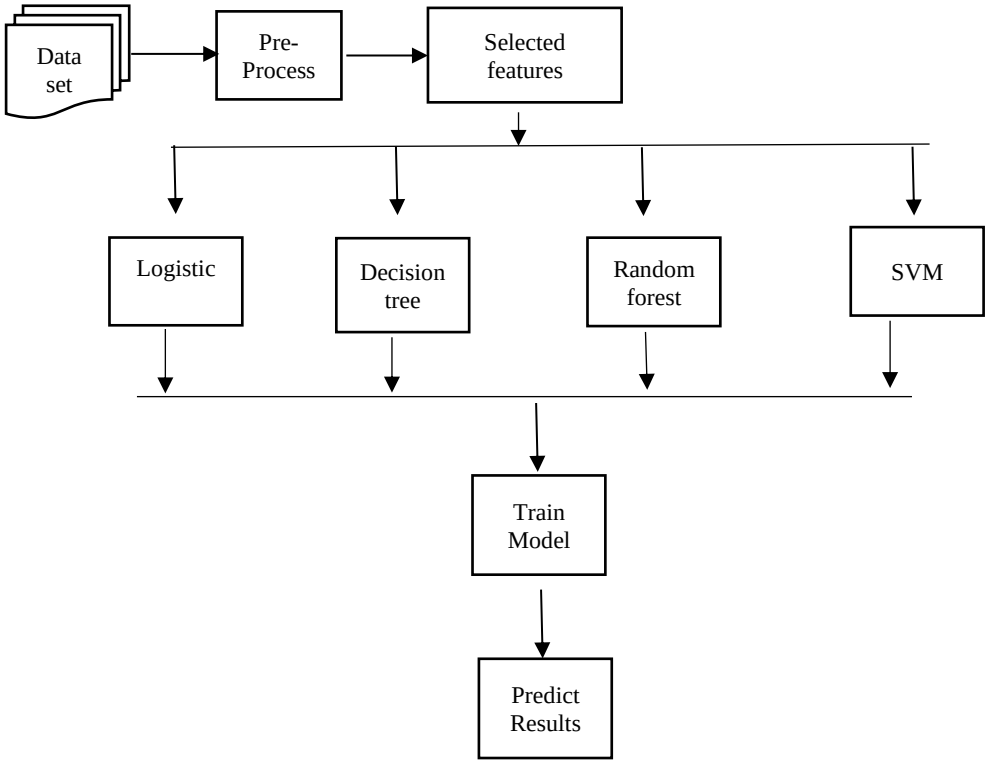

Figure 1. Block Diagram

A. Data Visualization:

Our paper suggests the latest machine learning algorithms to detect fraudulent in credit card. A large amount of information represented in graphic form are very easy to analyze and understand. In our approach, the confusion matrix and the data histogram are shown as data visualization part. But data visualization is not only important for data analysts and data scientists, it is important to understand data visualization in any career. Whatever you work in tech design, marketing and finance you need to visualize data. This shows the importance of data visualization.

ISSN: 2233-7857 IJFGCN

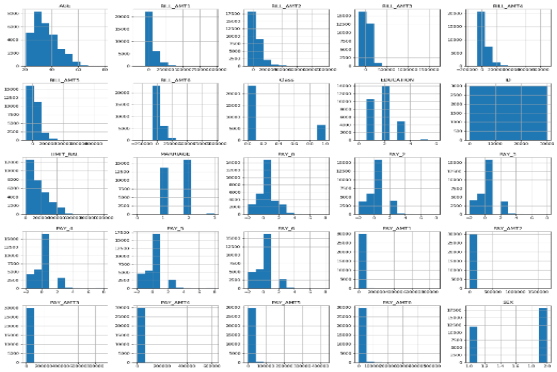

Figure 2. Data Attributes

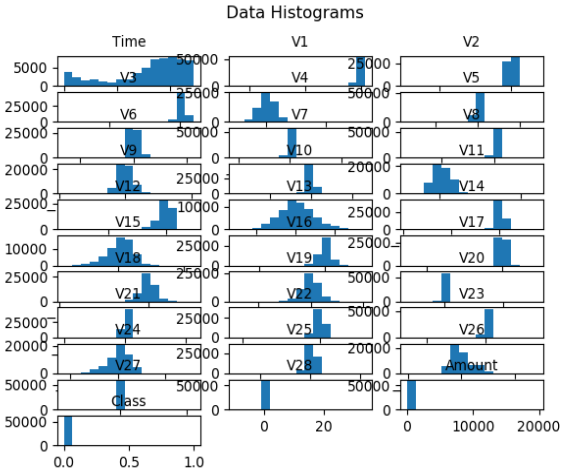

Figure 3. Data Visualization

B. Data pre-processing:

The pre-processing purpose is to convert raw data into clean data that fits machine learning. Clean and structured data allows a data scientist to get more precise results from an applied machine learning. This technique includes cleaning, sampling and data formatting.

Data set has been pre-processed for converting the string attributes to numerals and missing data records are dropped. The pre-processed data set is stored in file called “dataset.csv”, that is given as input for machine learning models. Data pre-processing is a process of converting raw data and make it suitable for machine learning model. It is the crucial step while creating a machine learning model. If we doing any operation with data, it is very important to clean it and put in a formatted way.

C. Dataset splitting

In training data set, the data is trained to a model and define its optimal parameters it has to learn from data. A dataset used for machine learning should be divided into three subsets are

- Testing
- Training
- Validation Sets

ISSN: 2233-7857 IJFGCN

A test set is necessary for an evaluation of the trained model and its capacity for generalization. The latter means a model ability to identify patterns in new unseen data after been trained over a training data. It's crucial to use different subsets for testing and training to avoid model overfitting, which is the incapacity for generalization.

D. Training and Testing:

Testing and training process for the classification of dataset in machine learning is extremely necessary. Every step should be chosen carefully by the researcher. The studies in the literature are typically theoretical. There is no useful model for selecting samples in the training and testing process. Therefore, there is need for resources in machine learning which discuss the training and testing process in detail and offer new recommendations different sampling theorems.

The test set is a set of observations used to detect the performance of model using some performance metric. It is important that observation from the training set are not included in the test set. If the test set contain examples from the training set, it will be difficult to know whether the algorithm has learning to generalize from the training set or simply memorized it.

The following algorithms are used to build the model and train the model for detection of frauds in credit card transactions:

- Logistic Regression
- Decision Tree
- Random forest
- SVM model

IV. EXPERIMENTAL RESULTS

In this chapter the experimental results are shown. The model is created in Python version 3.6 software platform. Implemented the machine learning algorithm on the given dataset for credit card fraud detection shows that Random forest model outperforms other models. The accuracy is high compared to other three machine learning algorithm.

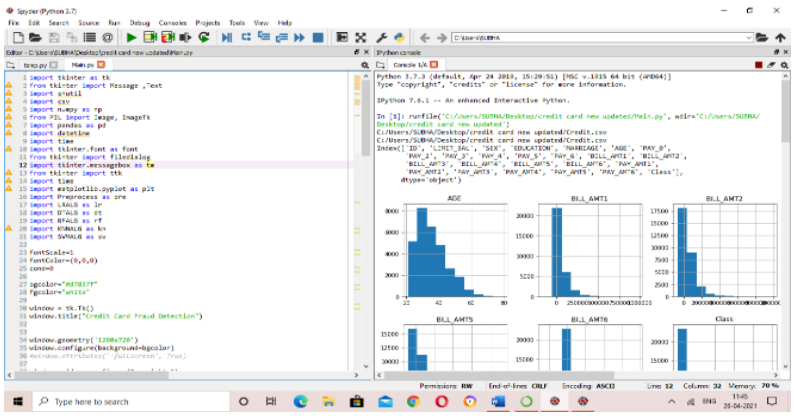

Figure 4. Pre-Processed Output(1)

ISSN: 2233-7857 IJFGCN

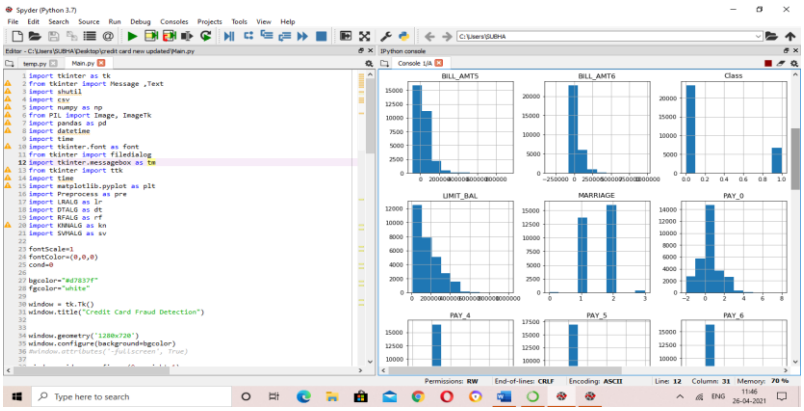

Figure 5. Pre-Processed Output(2)

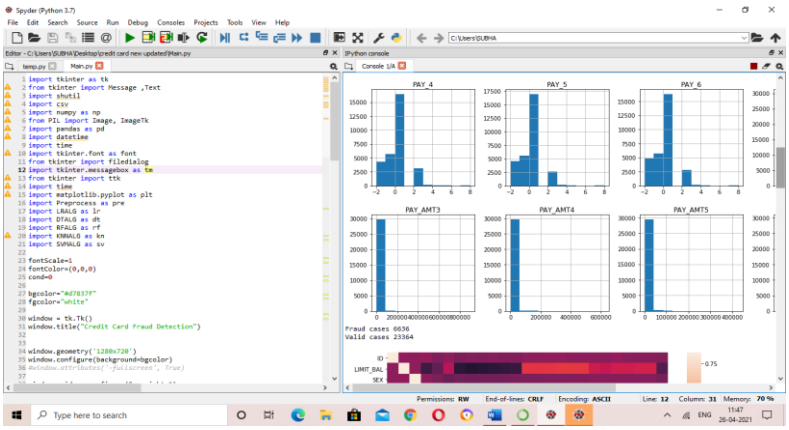

Figure 6. Pre-Processed Output(3)

The fig.(4) (5) (6) represent the pre-process of data set. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

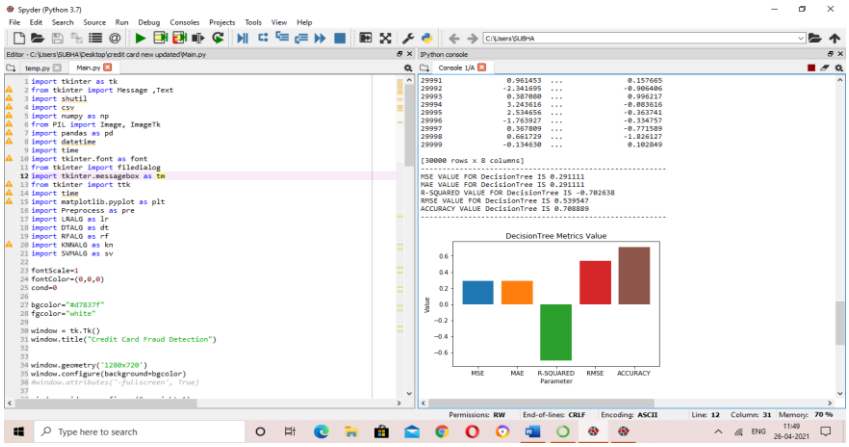

Figure 7. Accuracy rate for Decision Tree

ISSN: 2233-7857 IJFGCN

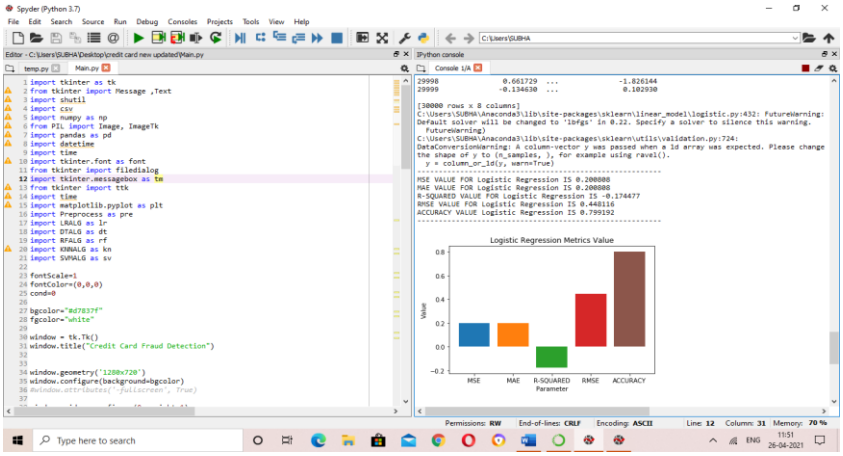

Figure 8. Accuracy rate for Logistic Regression

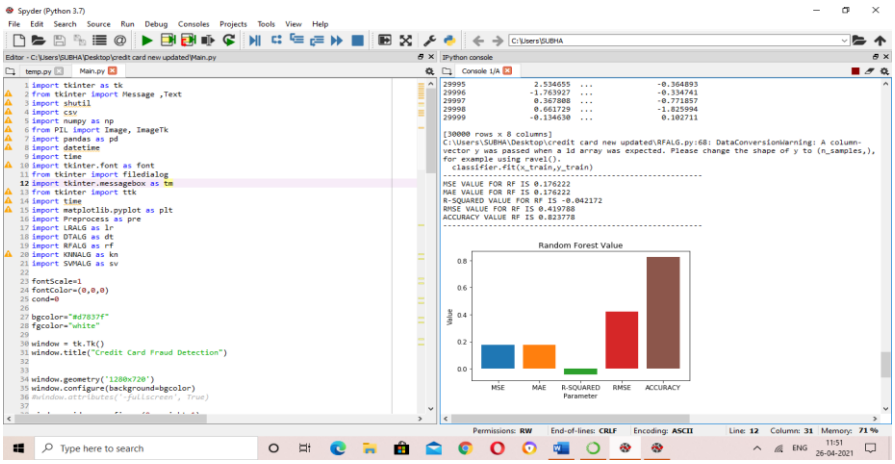

Figure 9. Accuracy rate for Random Forest

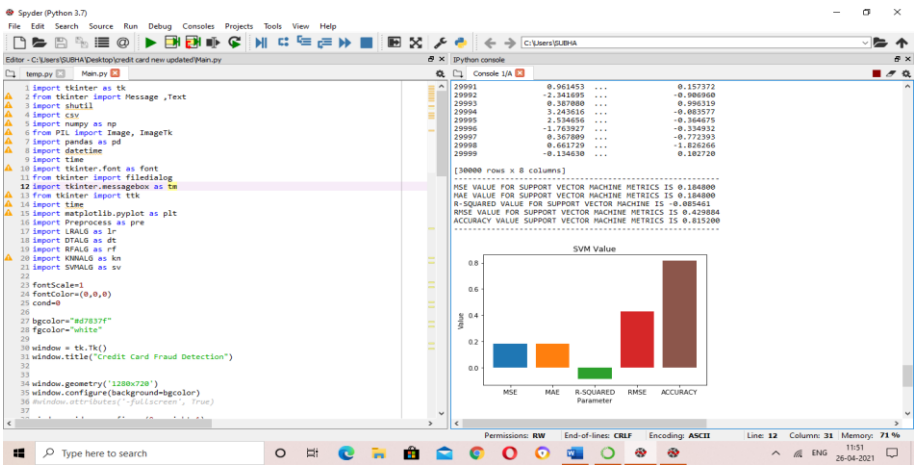

Figure 10. Accuracy rate for SVM Model

The figure 7, 8, 9, 10 depicts the accuracy rate of fraud detection.

- Accuracy rate for Decision Tree is 71%
- Accuracy rate for Logistic Regression is 79%
- Accuracy rate for Random Forest is 82%

ISSN: 2233-7857 IJFGCN

- Accuracy rate for SVM Model is 81%

Final results of each algorithm are shown with accuracy and the best algorithm is identified.

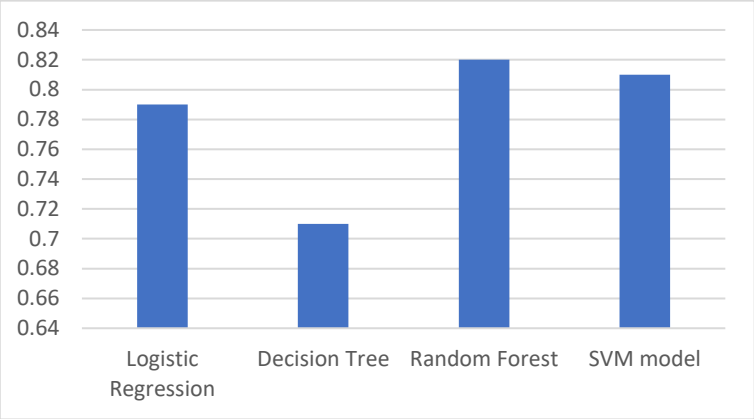

Figure 11. Accuracy Comparison for fraud detection

Table 1. Analysis of error detection and accuracy table

ALGORITHM	MSE	MAE	R-SQUARED VALUE	RMSE VALUE	ACCURACY
DECISION TREE	0.281212	0.2712	-0.644741	0.53029	0.718=71%
LOGISTIC REGRESSION	0.200000	0.2008	-0.174477	0.44811	0.79=79%
RANDOM FOREST	0.176222	0.1792	-0.042172	0.41978	0.82=82%
SVM	0.184800	0.1843	-0.085461	0.429884	0.81=81%

V. CONCLUSION

This method proves accuracy in finding out the fraud transaction and minimizing the number of false alerts. The use of this algorithms in credit card fraud detection results in predicting or detecting the fraud probably in very short period of time after the transaction has been made. This will eventually prevent the customer from huge losses and also reduce risks. A new method for data generation of imbalanced data set minority class was proposed to enhance fraud detection in credit card by PCA and machine learning algorithms as an oversampling strategy. PCA algorithms have been applied in huge areas, our domain aims to handle imbalanced data set issue by generating new minority classes instances to gain new training sets. Applying this algorithm into credit card fraud detection system aims to reduce fraudulence transaction and decrease the number of false alerts.

References

[1] Y. Sahin, S. Bulkan, and E. Duman, ``A cost-sensitive decision tree approach for fraud detection," Expert Syst. Appl., vol. 40, no. 15, pp. 59165923, 2017.

ISSN: 2233-7857 IJFGCN

[2] A. O. Adewumi and A. A. Akinyelu, ``A survey of machine-learning and nature-inspired based credit card fraud detection techniques," Int. J. Syst. Assurance Eng. Manage., vol. 8, no. 2, pp. 937953, 2017.

[3] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar, ``Credit card fraud detection using hidden Markov model," IEEE Trans. Depend. Sec. Com- put., vol. 5, no. 1, pp. 3748, Jan. 2008.

[4] J. T. Quah and M. Sriganesh, ``Real-time credit card fraud detection using computational intelligence," Expert Syst. Appl., vol. 35, no. 4, pp. 17211732, 2008.

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, ``Data mining for credit card fraud: Acomparative study," Decision Support Syst., vol. 50, no. 3, pp. 602613, 2011.

[6] N. S. Halvaiee and M. K. Akbari, ``A novel model for credit card fraud detection using articial immune systems," Appl. Soft Comput., vol. 24, pp. 4049, Nov. 2014.

[7] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, ``Credit card fraud detection: A fusion approach using DempsterShafer theory and Bayesian learning," Inf. Fusion, vol. 10, no. 4, pp. 354363, 2009.

[8] N. Mahmoudi and E. Duman, ``Detecting credit card fraud by modified fisher discriminant analysis," Expert Syst. Appl., vol. 42, no. 5, pp. 25102516, 2015.

[9] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, ``Association rules applied to credit card fraud detection," Expert Syst. Appl., vol. 36, no. 2, pp. 36303640, 2009.

[10] John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadaren Awoyemi, "Credit card fraud detection using machine learning techniques: A comparative analysis."International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9, 2017.

ISSN: 2233-7857 IJFGCN