Isolation Forest and Xg Boosting For Classifying Credit Card Fraudulent Transactions

Chandra Sekhar Kolli, T.Uma Devi

Abstract: ***Machine Learning based Neural Networks are extensively popular in creating machine learning models such as Text Classification, Face Recognition, Speech Recognition, Recommending new services etc. In the present hyper-competitive data-driven world these models are in huge demand since they provide high accuracy. In present days, the availability of internet and the wide variety of services like e-commerce, online shopping gained much popularity. On the other side of the coin, customers are facing adverse benefits due to fraudulent activities. Therefore, analyzing, detecting and preventing such unusual fraudulent activities in- real time is very essential and play crucial part. The main objective of this paper is to create a predictive model that capture the fraudulent transactions with high accuracy using Isolation forest & Local Outlier factor for detecting outliers that explicitly identifies anomalies and Extreme Gradient Boosting, an ensemble approach for constructing and evaluating the predictive model. A comparative study was done with existing models Logistic Regression, SVM, Random Forest with Extreme Gradient (XG) Boosting algorithm. The proposed model XG Boosting shows better performance and secured high accuracy 0.98.***

Index Terms: ***Fraudulent Activities, Isolation Forest, Local Outlier Factor, Extreme Gradient Boosting, ROC.***

I. INTRODUCTION

These days' credit cards are extensively used everywhere during money payment and online transactions. But, the physical presence of the card is not mandatory for successful completion of the transaction. These transactions are performed through third party payment gateways, such as CCAvenue, ICICI Payseal, and PayPal etc. It is not essential to use the physical card during online payment time, only few information regarding the card is enough [1]. The advancements in the technology and the same time it is open and available to all. Since it is open to all, it is giving more ways to fraudsters to explore different alternative ways and changes their plan to maintain anonymity during the fraudulent activities. Credit card fraudulent activities can be classified into (i) cloned credit cards (ii) stolen or lost (iii) no-card (card holder's information gained) (iv) identity-theft fraud [19].

Analyzing the past data forms the bedrock of the data science. With the help of statistical methods, a classification model can be developed that learns to identify when a fraudulent transaction has occurred. In the Supervised learning approach, the model is trained to analyze the previous classified data of fraudulent transaction, so it will learn to understand automatically what makes a transaction fraudulent. In supervised learning, it selects sub-sample of data randomly [22]. The model now is able to determine any new transaction falls in such category. In Un-Supervised approach, methods like anomaly detection is used to capture the fraudulent transaction as an outlier in whole dimensional space. In future, any transactions falling outside the particular region will be classified as fraudulent transactions. In the Literature, some techniques like Artificial Neural Networks, Markov Decision Model, Bayesian Networks, Decision trees were used to classify the fraudulent and legitimate transactions.

The main objective and intention of fraud detection is to classify and confirm whether a particular transaction is performed by legitimate user (card owner) or by others [2] and maximize correct predictions and maintain incorrect predictions at an acceptable level [11]. Implicit authentication [10], it is an approach that uses observations of user behavior for authentication can also be considered to differentiate legitimate users from fraudsters. Basically there are two ways to detect the fraud [3]. One is Misuse Detection and second is Anomaly Detection. For detecting misuse transaction classification algorithms can be applied. For anomaly detection, the basic profile of the user and past transactions, along with behavior of the card owner are considered during the analysis. Based on the old data, it can be decided that the new transaction is fraudulent or genuine.

During the transaction, cardholder's details sent to merchant and concern bank through internet. To secure the data, transmission is done via SSL/TLS protocol enabled services to provide end-to-end security. But there is no mechanism by to verify the authenticity of the entity [18]. The Secure Electronic Transaction protocol enables better security but the overall performance is poor [18]. One-time password (OTP) method is to provide the authentication to complete the transaction. To provide reliable, robust and more secure transaction, introducing biometric authentication method during transactions [19]. Financial Institutions and banks with an online presence must have the ability to detect [12] and respond to fraudulent activity before incurring losses. Leverage human insight and machine learning to identify advanced fraud and block more fraud while reducing fraud prevention costs.

In order to detect frauds, traditionally statistical techniques, clustering and classification techniques are used to find the fraudulent activities, but these are complex in nature. Machine learning algorithms like decision trees, Bayesian networks can be used to detect fraudulent transactions. Fraudsters are keep on refining the methods which they use and follow, hence it is very important to develop more advanced fraud detection [4] methods. In this article, a

Revised Manuscript Received on June 05, 2019.

Chandra Sekhar Kolli, Research Scholar, Computer Science Department, GITAM Deemed to be University, Visakhapatnam, India and Assistant Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India.

Dr. T. Uma Devi, Associate Professor, Computer Science Department, GITAM Deemed to be University, Visakhapatnam, India.

supervised learning algorithm Extreme Gradient Boosting algorithm is used to detect fraudulent transactions. The model is constructed using python programming language and the dataset taken from kaggle.

A. Machine Learning

Machine Learning is a very powerful and transformative tool for security aspect because of its dynamism. Traditional approaches like Rule-based methods that were designed to fight against specific attacks only. Those are not enough adaptable to address new kind of intrusions. Because rule-based methods are static and are very rigid. In addition, the strictness of those techniques block legitimate or authentic users. To achieve fraud detection, first, the major thing is preserve privacy of the user’s highly sensitive data. Fraud is a very rare event, but the main challenging issue is to find a way to bring into light the abnormal behavior in the system or application.

Particularly in some domains like banking and medical, sharing of data of individuals are not allowed, due to the policies and regulations of the organizations and firms. It is well-known fact that the neural network models work better and accurate if the more data is given during training the model. Due to the policies of the organizations and firms, complete data from that particular firms are not available to train the model fully. In that case, the model is just a homogeneous model and inaccurate results will be produced for a change in the input at real-time. In this particular context, privacy and confidentiality restrictions significantly effecting and reducing the accuracy of the models.

B. Isolation Forest

Many machine learning algorithms suffer from performance due to outliers. Outliers are less frequent than the regular data points. These data points are different in terms of their values. The working strategy is Isolation forest and Random Forest are somewhat similar. The main functionality of Isolation forest is, it analyses data and exactly identifies anomalies. In this, grouping of data is done by randomly selecting any feature and apply random split on the selected feature at maximum and minimum value. Like any other outlier detection methods, in isolation forest also anomaly score for decision making.

$$s(\mathrm{x},\mathrm{n}) = 2^{-\frac{E(\mathrm{h}(\mathrm{x}))}{c(\mathrm{n})}}$$

Where h(x) is the path length in the tree, c (n) is the average path length of unsuccessful search and n is the number of nodes in the tree. For each observation *x*, anomaly score *S*(x, n) is calculated. If the score is close to 1 then the observation is assumed to be anomaly, if the score is less than 0.5 then it is observed as normal, and if score is 0.5 it is considered as the sample doesn’t have clear distinct anomalies.

C. Local Outlier Factor

It is an un-supervised method, it computes the anomaly score of each sample and also measures the local deviation of the specified sample to its neighbors. The fundamental idea is based on the concept of Local Density. To identify the regions of similar density, compare the local density of an item with its neighbors. The set of K nearest neighbors as Nk(A). Reachability distance can be calculated as:

K (A, B) = max {k-distance (B), distance (A, B)?}

D. Random Forest

To derive any Random forest [13], Decision Tree is the rudimentary. Random Forest is basically composed of simple tree predictors [24]. Random forest is best suited for large datasets and at the same time the learning algorithm produces accurate results and handles missing data and exhibits good performance results. Random forest algorithm or classifier can be used for Classification as well as Regression tasks. It efficiently handles missing data and won't over fit the model if it there are more trees. Random Forest is well suited when the class distribution is unbalanced [15]. Single-Tree model such as decision tree are very sensitive to specific training data and more prone to over fitting of data [4]. By using ensemble methods somehow over fitting problem can be reduced by combining a group of decisions [16] thereby improving the accuracy of results. The accuracy of random forest depends upon accuracy of each individual tree as well as correlation between the groups of trees. The better accuracy of each individual tree collectively will give the best performance results for the ensemble tree. The variation and their randomness of a tree will usually come by selecting different subsets of attributes during the construction of decision tree.

The training set for each and every individual decision tree is a group of randomly chosen training data. At every internal node of the tree, it again randomly selects some subset of attributes and then computes the center. The LeftCenter and RightCenter are denoted as Class 0 and Class 1. The kth element of a center is calculated [17] with the below formulae:

$$LeftCenter[k^{th}\ \text{element}] = \frac{1}{n}\left(x_{ik} I\left(y = 0\right)\right)$$

$$RightCenter[k^{th}\ \text{element}] = \frac{1}{n}\left(x_{ik} I\left(y = 1\right)\right)$$

In the present node, each of the element in the train data set is classified by calculating Manhattan Distance between the element and the center of the node, it is calculated as:

$\text{distance(center, element)} = \sum_{i \in sub} | (center[\mathrm{i}] - i^{th} element)$ Sub is the sub-set of attributes which are randomly selected from the given set of attributes (X).

The Pseudo code for Random Forest Classification:

```
for b=1... B:
Xb, Yb = sample_with_replacement(X, Y)
Model = DecisionTree ( )
While not at terminal node and not reached max_depth:
    Select d features randomly
    Choose best split from the d features (i.e. max. information gain)
    add split to model
    models.append (model).
```

E. Ensemble - Extreme Gradient Boosting

Extreme Gradient Boosting is also known as XGBoost. It is an ensemble technique to combine diverse set of models to increase the performance of the model. Combining all the predictions from a set of models together will be termed as Ensemble Learning. Models can be dissimilar from each other because of several reasons such as difference in population, hypothesis, modeling techniques used, and difference in initial seed or combination of any of these parameters. One important thing in ensemble learning is Error. The errors of any model can be mathematically classified into Bias, Variance and irreducible error.

Bias error is used to quantify the difference between the predicted value and the actual value on an average basis. If the bias error is high, it indicates that the model is under-performing. On the other hand, variance, it quantifies how much extent the predictions are different from each other on the same data points or observations. High variance indicates that it is going to over fit on the training data. It also implies that it will give poor performance results with un-trained data.

It is difficult and hard to build a single model that works best. There are basically two types of approaches, a model that do not use randomness and the model that uses randomness. There are three different ways for ensemble, one is Bagging, Boosting and Stacking. In Bagging, it tries to develop similar learners with small set of sample data and then calculates the mean of all the calculations. In Boosting, it alters the weight of any observation based on the last classification, that is, if any of the observation is classified incorrectly, then it automatically increases the weight of the observation to balance it properly. In Stacking, the model will pool the output from different learners which lead to decrease in either bias or variance.

XG Boost used for supervised learning such as Regression and Classification problems, where the training data with multiple features are used to predict a target variable Yi. It is so popular because of its Speed and Performance. The core algorithm in XG Boost is parallelizable, it supports millions of datasets. Gradient Boosting is one of the powerful ensemble technique for constructing Predictive Models. The basic idea of boosting is to combine all the sequence of base models or trees. The model is ensemble, it takes each weaker base decision tree and builds models based on successive sequence models and generalizes the model. The gradient reflects the representation of minimizing the loss function over the entire training set.

II. LITERATURE REVIEW

Mohammad et.al. [5] reviewed different types of frauds happening with credit card usage and the existing detection techniques for detecting fraud. It is classified into two types, one is behavior fraud and application fraud. In application fraud, fraudsters will give false information or will give legitimate information of others to get a new credit card. In behavior fraud, fraudsters will get the bank account number, internet account credentials of genuine people and use them for their transactions. In recent days, a new kind of fraud detection is followed by bankers, in which banks will check the behavior of the linked cardholder [6]. In majority of the fraud detection methods, the basic principle is to analyze the behavior pattern of the card holder and detect the un-usual patterns and confirming the fraud transactions.

Amlan et.al [7] proposed a technique, in this approach misuse detection and anomaly detection [3] were combined with the help of using two-stage sequence alignment. The main use of profile analyzer is to find the similarity of any new transaction with the legitimate cardholder's past transactions. The fraudulent transactions were traced out with the help of the profile analyzer and gives input to deviation analyzer for possible orientation with the previous behavior. The final result was measured by considering the observations of the two analyzers.

Sahin and Duman [8] have made a comparative study between decision tree and SVM. By dividing the two datasets into groups in different proportion and constructed seven Decision Trees and Support Vector Models. The results show that model is comparatively better than SVM. But, the accuracy of SVM reached the same performance like decision tree when increased the size of training dataset.

Srivastava A, Kundu et al explained the credit card fraud detection using Hidden Markov Model [9]. They developed the model that simulates the sequence of operations during the transaction and shown how it can be used to detect frauds. This model is initially trained with legitimate user’s data.

Ekrem Duman [20] implemented fraud detection system in Turkish Bank using Migrating Birds Optimization algorithm. In his implementation, he considered Saved Limit Ratio as the main performance attributed. This algorithm obtained 93.9% accuracy. Wen-Fang Yu and Na Wang [21] used genetic algorithm – Outlier Data-Mining approach to detect fraudulent data points in the dataset. They achieved accuracy around 89.4% when the outlier threshold is 12.

III. MODEL CREATION AND EXPERIMENT

A. Data Preparation

The Dataset CreditCards.csv, is taken from Kaggle dataset owned by Google. This dataset consists 2, 84,807 record of credit card transactions, 31 different parameters or attributes in the dataset. In the proposed model, Local Outlier Factor for calculating Anomaly Score and Isolation forest algorithm were used. In the dataset, the attribute "Class” value is 1.0000, this implies that such transactions are fraudulent transactions, if “Class” value is 0.00000 are considered to genuine transactions. For constructing model, Python Programming Language is used.

B. Data Analysis & Outlier Detection using Isolation Forest and Local Outlier Factor

When analyzing and evaluating the attributes in the dataset, it is observed that there is much skew between genuine transactions fraudulent transactions. Philip K chan [21] suggested if the distribution is 50:50, then it will produce an ideal model.

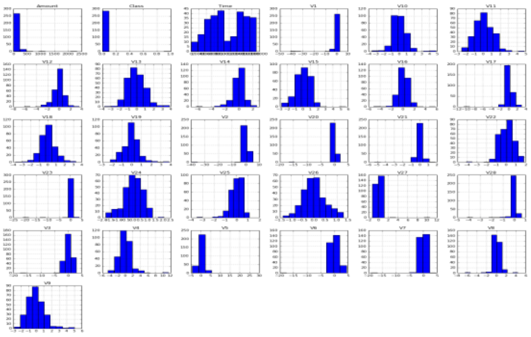

Fig 1. Histogram that illustrates Legitimate Transactions versus Fraud transactions.

When the dataset was analyzed, total number of transactions are 28481, number of fraud cases are 49, number of legitimate transactions are 28432. When applied for the outlier detections, outlier fraction is 0.001723. There is a huge disparity between numbers of legitimate transactions with fraudulent transactions. Correlation matrix was constructed to see the strong correlation between different variables in the dataset, Fig 2 shows heat map of correlation matrix, it shows which features are important for the overall classification.

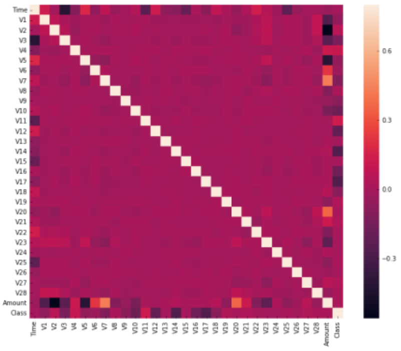

Fig 2. Heat Map of Correlation matrix

For finding the anomalies, Isolation forest and Local Outlier Factor algorithms were used. Local Outlier Factor computes the anomaly score of each sample, and measures local deviation or density of the given sample with respect to its neighbors. On the other hand, Isolation Forest depends on how much extent the data objects are isolated with respect to the neighbors by randomly selecting the split value from the sample. Table I shows the comparison of the test results of both the approaches.

Local Outlier Factor	Accuracy= 99.6708			
		Precision	Recall	Support
	Class 0	1.00	1.00	22748
	Class 1	0.00	0.00	37
Isolation Forest	Accuracy =99.714			
	Class 0	1.00	1.00	22748
	Class 1	0.13	0.14	37

Table I. Comparison of test results of Local outlier and Isolation Forest

Exploratory Data Analysis

The fraud detection model is used to analyze and verify whether the incoming transaction is legitimate or not. Machine Learning algorithms such as Logistic Regression, SVM, Random Forest are used to create classification models and compare the results with Extreme Gradient Boosted Tree. The credit card transactions dataset contains 2, 84,807 real time transactions it includes 0.17% fraudulent transactions i.e. 492 transactions are identified as fraudulent. The dataset contains 30 attributes or features, all are of type numerical data only. After applying the Principal Component Analysis and transformation it results 28 attributes. One particular feature "Class", if its values is 1 that implies it is Fraud transaction, and 0 means legitimate transaction. Fig 3. Show the sequence of steps followed in creating and evaluating models.

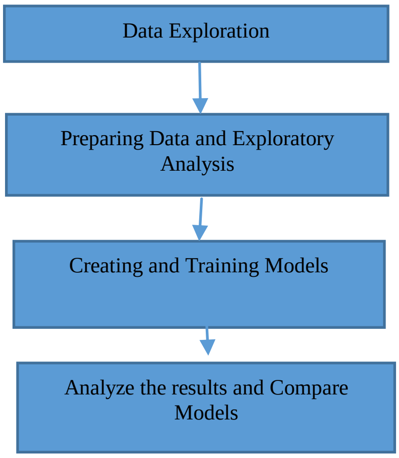
Data Exploration
Preparing Data and Exploratory Analysis
Creating and Training Models
Analyze the results and Compare Models

Fig 3. Sequence of steps for Evaluating Models

Before preprocessing the dataset, first the dataset is divided into features and response variables. It is considered that the test size is of 20% and divide the samples based on the response variable, because 0.17% (492) fraudulent transactions exists in dataset. The exploratory data analysis was performed and compared based on different quantities, and observed that the values are right-skewed, the same thing is represented in Fig. 4. Histogram of the transactions. It is very difficult to understand from the histogram because there are many outliers which cannot be seen.

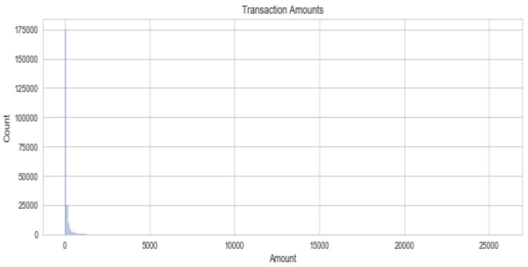

Fig 4. Histogram of the transactions amount.

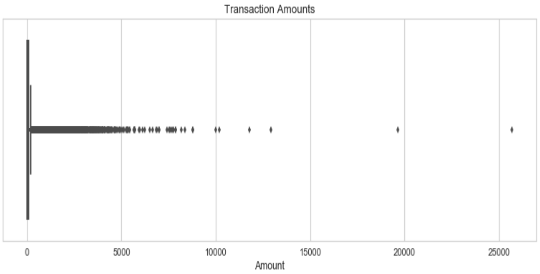

Fig 5. Boxplot to visualize outliers.

Fig 5. Boxplot to visualize the outliers in the data. It can be observed that, there are no outliers on the left and there are huge number of outliers on the right. This indicates that the amount are certainly seem right-skewed. Box-cox transformation applied to bring the transaction amount close to the normal distribution. After applying the transformation plotted Box-Cox transformation. Fig 6. Showing the resultant transaction amount. Again calculated the skew, now it is almost close to zero, i.e. 0.1142. That indicates and confirms that the

transformation removed the skewness in the data.

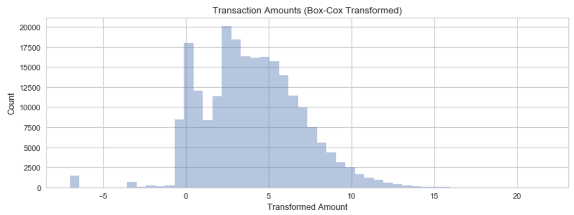

Fig 6. Transaction amounts after Box-Cox Transformation

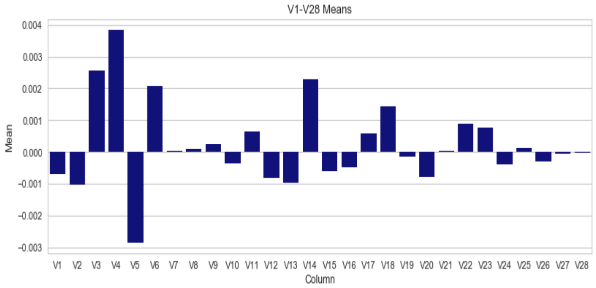

Fig 7. Histogram for Principal Component Variables in the dataset.

To compare the descriptive statistics, histogram is plotted for PCA variables and is shown in. To estimate the dependency between any two attributes, the non-parametric method called Mutual Information can be applied, which can capture any kind of statistical dependency among the variables. If the mutual information is 0, that tells no dependency and higher value indicates higher dependency between the variables. In data set there are more training samples, so mutual information surely works the best. Table II shows the calculated mutual information of all the variables.

V17	V14	V10	V12	V11
0.00803	0.00797	0.00735	0.00735	0.00660

V16	V4	V3	V18	V9
0.00579	0.00484	0.00475	0.00402	0.00399

V7	V2	V21	V27	V6
0.00394	0.00308	0.00230	0.00227	0.00226

V5	V1	V8	V28	Time
0.00225	0.0019	0.00184	0.00175	0.00172

Amount	V19	V20	V23	V24
0.00141	0.00132	0.00113	0.00082	0.00059

V26	V22	V25	V15	V13
0.00045	0.00038	0.00037	0.0002	0.00020

Table II. Mutual Information (Statistical Dependency) of each PCA Variable.

Modeling

After the preparation of the data and thorough exploratory data analysis, different algorithms like linear regression, support vector machine and random forest are experimented for comparative study.

Logistic Regression

Fig 8 and Fig 9 shows the confusion matrix for the logistic regression without normalization of data and with normalization of data respectively and classification accuracy.

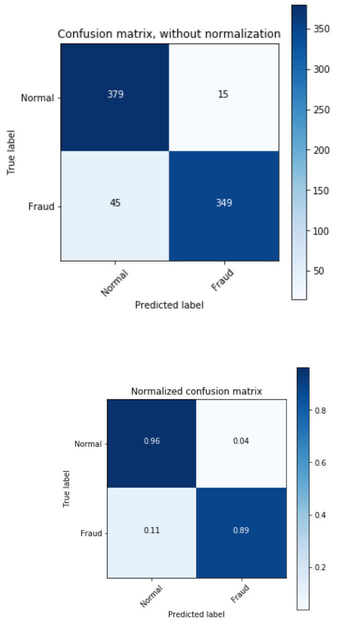

Fig. 8 Logistic Regression confusion matrix without normalization and with normalization- Classification on the training data.

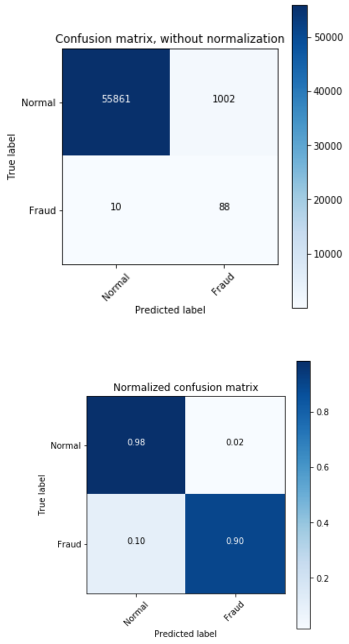

Fig. 9 Linear Regression confusion matrix without normalization and with normalization- Classification on the test data.

The classification is examined to find the optimal point where the model is going to fit with high score in detecting fraud i.e. Recall Score and Precision score. Recall score will be used to identify where exactly loss happening and Precision score will be useful for improving the overall system accuracy. Fig 10 shows the precision and recall scores for the classification threshold.

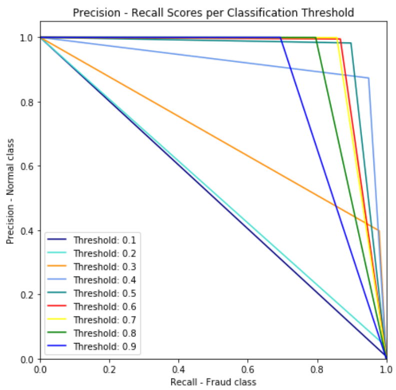

Fig 10. Precision and Recall score of the classification threshold.

SVM Model

SVM is a linear classifier [14] which attempts to learn the maximum margin by separating the two classes in the data. The separation is determined by a small subset of instances called support vectors. To further examine the accuracy and performance, SVM algorithm is explored with the same dataset which already analyzed at the beginning. During SVM implementation, the best performance was obtained RBF kernel function. To avoid over fitting problem, 5-fold cross validation is used since there are very less fraudulent transactions in dataset. Analysis and observations illustrated in fig. 11 confusion matrix.

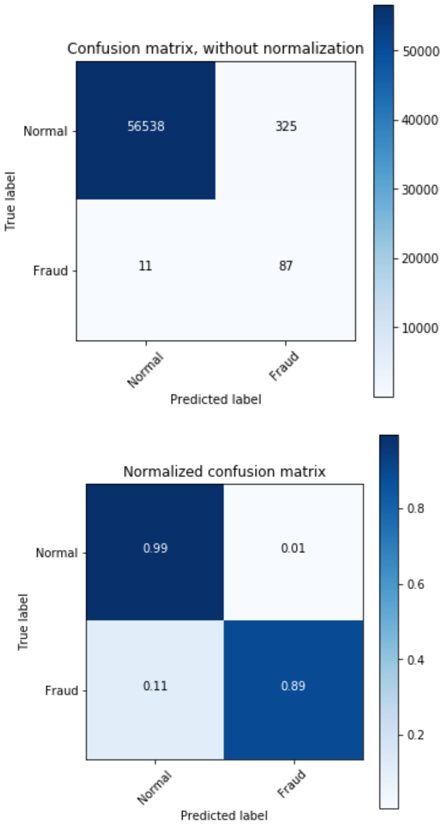

Fig. 11 SVM confusion matrix without normalization and with normalization- Classification on the test data.

RANDOM FOREST

In this method, there is no need to rescale the data for the tree-based models. The random forest taken much longer time to train the large datasets. According to the cross-validation, the Random Forest performed well. Table III illustrates the classification report of the Random Forest model.

Classification Results of Random Forest				
	Precision	Recall	F1 Score	Support
0	0.99974	0.99982	0.99978	56864
1	0.89247	0.84694	0.86911	98
Average	0.99955	0.99956	0.99956	56962

Table III. Classification Report of the Random Forest

EXTREME GRAIDENT BOOSTING

Extreme Gradient Boosting is also referred as XG Boost, it is chosen in implementation since its execution speed and learning is comparatively good. Stratified k-fold is used where k value considered as 10. Area under the precision-recall curve was used to illustrate classification accuracy. It is illustrated in Fig.12 area under the precision-recall curve and its accuracy is 0.979.

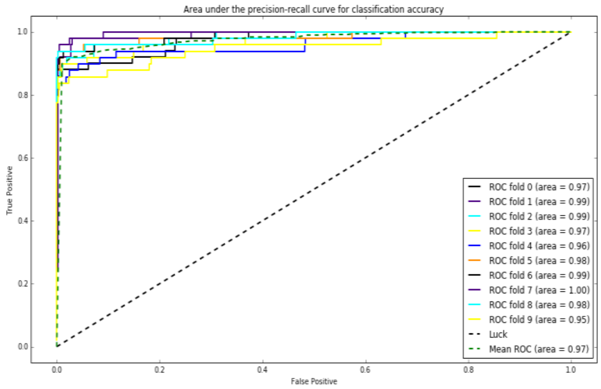

Fig.12 Area under the precision-recall curve for classification accuracy.

From the experimental results, it is observed that the XG Boost algorithm achieved area under the precision-recall score 0.980. Table IV illustrates the comparison of ROC Curve precision values with other algorithms. It is very insensitive to the class imbalance and hence it can be considered and preferred as an evaluation metric for estimating the performance of the learning model.

	Area Under Curve Precision Values
SVM	0.8075
Decision Tree	0.8773
Random Forest	0.8843
XG Boost	0.980

Table IV. Comparison of area under the curve – precision values of different algorithms

IV. CONCLUSION

Fraud Detection Model need to be more sophisticated to maintain the pace with the fraudsters and act in response within a short span to time when it is discovered. Financial Institutions and Banking sectors who want to defend themselves against fraud need to update their technology time-to-time and must be superior, faster re-learning & re-skilling in finding the better solutions that can continuously evolve and easy to use and maintain. Machine learning models has to redefine the traditional strategies & previous tools in fraud management. To explore the better analytical accuracy and to improve the performance of fraud detection models, logistic regression, support vector machine, random forest are compared with the Extreme Gradient Boosting algorithm. These models are tested on same credit card data set and the accuracy is evaluated. Among all these models, XG Boost outperforms in terms of performance metrics accuracy, precision and recall. The problem is if the dataset increases more it may lead to over fitting problem. This can be considered as a future work to avoid the over fitting problem in real-time fraud detection. Application of Deep Learning concepts perform better analysis and can deliver risk scores in real-time with better accuracy. Such models can also be applied to derive outcome measurements such as a statistical risk. The effectiveness of the risk score will depends upon the model, which detects anomalies from patterns, identify matches to known patterns, and uncover new patterns.

REFERENCES

1. Vlasselaer, V. V., Bravo, C., O., Eliassi-Rad, T., Akoglu, L.,and Snoeck, "A novel approach for automated credit card transaction fraud detection using network-based extensions", Decision Support Systems, 38-48,2015.
2. Duman E, "Detecting credit card fraud by genetic algorithm and scatter search", 13057-13063, 2011.
3. W. H., and Vardi, "A hybrid high-order markov chain model for computer intrusion detection", "Journal of Computational and Graphical Statistics", Volume 10, Issue 2, PP 277-295,2001.
4. Bhattacharyya S, Jha S, Tharakunnel K, and Westland JC, "Data mining for credit card fraud ,a comparative study", Decision Support Systems 50, 602-13, 2011.
5. Behdad M, Barone L, Bennamoun M, and French, "Natureinspired techniques in the context of fraud detection", IEEE Transactions on Systems Man and Cybernetics Part C, 1273-1290, 2011.
6. Quah J. T. S, and Sriganesh M, "Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications", 1721-1732, 2008.
7. Kundu A, Panigrahi S, Sural, S, and Majumdar A. K,"Blastssaha hybridization for credit card fraud detection", IEEE Transactions on Dependable and Secure Computing, 309-315, 2009.
8. Sahin Y, and Duman E, "Detecting credit card fraud by decision trees and support vector machines", "Lecture Notes in Engineering and Computer Science", 2011.
9. Srivastava A, Kundu , Sural S, and Majumdar A, "Credit card fraud detection using hidden markov model", IEEE Transactions on Dependable and Secure Computing, Volume 5 Issue 1, 37-48, 2008.
10. Shi E, Niu Y, Jakobsson M, and ChowR, "Implicit Authentication through Learning User Behavior" International Conference on Information Security (Volume 6531, PP 99-113). Springer-Verlag-2010.
11. Stream Base, 2008 Entrust www.entrust.com.
12. https://www.insurancehound.co.uk/claims/fraud/definitive-guide-next-generation-fraud-prevention-31031#.
13. Breiman L, Random forests. Machine Learning, 5-32, and 2001.
14. B. Schlkopf, C. J. C. Burges, and A. J. Smola, editors. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, Massachusetts, 1999.
15. Khoshgoftaar T. M, Golawala M and Van Hulse J, "An Empirical Study of Learning from Imbalanced Data Using Random Forest", in 19th IEEE International Conference on Tools with AI PP 310–317, 2007.
16. Dietterich T.G, "Ensemble methods in machine learning. 1-15,2000.
17. Abeel T,and Saeys Y, "A Machine Learning Library, Journal of Machine Learning Research", 931-934, 2009.
18. Rexha B, "Increasing user privacy in online transactions with X.509 v3 certificate private extensions and smartcards", Seventh IEEE International Conference on E-Commerce Technology, 293 - 300, 2005.
19. Statistic Brain Research Institute, Credit Card Fraud Statistics, Available: http://www.statisticbrain.com/credit-cardfraud-statistics, 2014.
20. Ekrem Duman, Ayse Buyukkaya, and Ilker Elikucuk, "A Novel and Successful Credit Card Fraud Detection System Implemented in a Turkish Bank" in IEEE 13th International Conference on Data Mining Workshops, pp 162 –171, 2013.
21. Wen Fang Yu, Na Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum" in International Joint Conference on Artificial Intelligence 2009, PP 353 – 356, 2009.
22. Eesha Goel, Abhilasha and Abhilasha, "Fraud Detection Using Random Forest Algorithm", International Journal of Computer Science Engineering (IJCSE), Volume 5, Sep 2016.

AUTHORS PROFILE

Chandra Sekhar Kolli is pursuing PhD in Computer Science from GITAM (Deemed to be University), Visakhapatnam, India and working as an Assistant professor in Koneru Lakshmaiah Education Foundation (Deemed to be University), Vaddeswaram, Guntur, India. His research interests includes Data Analytics, Internet of Things, and Security & Privacy.

Dr. T. Uma Devi is working as Associate Professor in the Department of Computer Science, GITAM (Deemed to be University), Visakhapatnam, and India. Her research interests include Mining Biological Data, Data Analytics, Soft Computing and Cloud Computing.