Fraud Detection and Prevention in Banking Financial Transaction with machine learning using R

Ankita Layek [1]

[1] Symbiosis Institute of Operations Management, Nashik

ABSTRACT

Fraud is an intentional act of deception involving financial transaction for the purpose of personal gain. With the increased number of online transaction, frauds have also increased. In banking sector detecting fraud is important to keep customers' money safe and also to reduce the losses from fraud and keep company profitable. Traditional fraud detection methods are no more sufficient in detecting frauds so banks are adopting machine learning based models. One major problem with the financial transaction data is its skewness. Performance of any model depends on dataset and the technique applied. This paper has compared seven machine learning models (logistics regression, random forest, XGBoost, DBscan, Artificial neural network, isolation forest, Principle component analysis with Support vector machine) with the help of several parameters as accuracy, sensitivity, specificity, precision, balanced accuracy (BCR), Matthews correlation coefficient (MCC), kappa value. The study was done for a period of four months on Paysim synthetic dataset of mobile money transactions published on kaggle. The machine learning models were created using R and data analysis was done with the help of tableau. Post analysis It is found that random forest and XGBoost is providing better result than other models.

Keywords:

Machine learning, neural network, confusion matrix, supervised learning, unsupervised learning, ensemble learning, fraud detection, gradient boosting.

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

1. Introduction

Recently for few past years there is a sharp increase in online money transaction as new generation is continuously avoiding physical presence in banks to avoid hassle and long queue. Presently due to COVID19 as people are becoming habituated with online transactions so this will further increase. As always with the increasing number of online transactions, frauds will also increase. In 2018 global fraud detection and prevention market was valued by nearly USD 17.33 billion and it was expected to grow from 2019 to 2025 at a compound growth rate of 18.9% annually according to (Grand View Research, 2019), as a result of technology disruption in business environment. McAfee estimation report by (Steve, 2018) shows that among all the financial fraud only cybercrime itself costs $600 billion globally, which is 0.8% of global gross domestic product. Over time, digital convergence is expected to create new market opportunities and make certain existing business processes obsolete. With the advancement of technologies as Internet of Things (IOT), Robotics, Artificial Intelligence (AI), Augmented Reality (AR), cloud computing, and mobile banking & e-transactions has enhanced virtual platforms for the new-age consumer. However, along with this increasing number of cyber-crimes, financial crimes, data breaches, and identity thefts are challenging the growth of many organizations. One area of concern is digital fraud, which has become an important issue for businesses operating in the banking and finance, healthcare, and e- commerce areas.

Previously banks and financial institutions used Rule based methods to detect fraud. Complex rules were made to detect those frauds which had already occurred in past or to avoid some common fraud. But now a day's fraudsters have shifted from conventional methods to use of technology so rule based approach is not able to detect new frauds. Fraudsters are continuously finding loopholes in the transaction applications and by using technology they are fulfilling their motive. That's why financial institutions are moving towards machine learning and artificial intelligence to reduce fraud.

According to (Ali, Salleh, Saedudin, Hussain & Mushtaq, 2019) The main difficulty in implementing machine learning is the imbalanced or skewed dataset in case of financial transactions. Machine learning models learns from data and

creates a pattern and if the data is less it will not be able to identify the frauds correctly.

Main objective of this study is to identify maximum fraud transactions while keeping the false positives low (Choi & Lee, 2018). False positive denotes the genuine transaction that is predicted as fraud transaction by the model. False positives can lead to customer dissatisfaction and may result in loss of customer.

In this research paper a detailed study on machine learning has been conducted to detect fraudulent transactions. A comparative study is done on different machine learning model and also neural networks to find the best fitted algorithm. Although the model accuracy will differ for different dataset but by this way different models can be compared and it can be checked that which models are working better on imbalanced data set and how can we select the best model after comparing them. In this paper applied models are logistics regression, random forest, Dbscan, isolation forest, ANN, XGboost, PCA with LR.

2. Literature Review

Several machine learning and deep learning models are applied to detect fraud. In machine learning there are mainly supervised and unsupervised algorithm has been used for financial transactions dataset. A study was done by (Sadgali, Sael & Benabbou, 2019), has compared several machine learning models in terms of their advantages and drawbacks. It has analyzed PNN, SVM, logistic regression, random forest, decision tree and tried to provide solution by combining several models. Another study by (Bagga, Goyal, Gupta & Goyal, 2020) has used ADASYN (Adaptive Synthetic Sampling Method for Imbalanced Data) to make the unbalanced data balanced and to improve the accuracy of the model Pipelining and bagging classifier is proposed. Finally, pipelining is proposed as best method for fraud detection.

Another study was conducted by (Olowookere & Adewale, 2020) which proposed a framework consist of cost sensitive learning paradigm and ensemble learning paradigm which helps in improving fraud detection in imbalanced dataset. In this paper cost sensitive ensemble is built from decision tree, MLP and KNN algorithm. Similar studies were done by (Raghavan & Gayar,2019) has discussed and compared SVMs, KNN, K-Means, Random Forest, Naïve Bayes, auto-encoder to detect fraud. From the study authors concluded that for large dataset SVM is best and it can be combined with CNN to get best result and for small dataset random forest and KNN can provide better result. But this study is limited to supervised learning to detect fraud.

A detail research by (Amarasinghe, Aponso & Krishnarajah, 2018) included supervised machine learning (Bayesian network, RNN, SVM, Fuzzy logic) and unsupervised machine learning (point outliers, K means clustering, hidden markov model) to detect fraud and skillfully specified the advantages and disadvantages of these model. It is also found that ANN provides 33% better accuracy than fuzzy logic. This paper also proposed to use ANN with genetic algorithm and to compare it with other algorithm.

(Shirgave, Awati, More & Patil, 2019) in their research has analyzes several supervised learning algorithms (logistic regression, KNN, RF, SVM, decision tree, naive bayes) and compared specificity, sensitivity and accuracy of these models to identify fraud based on scoring rule. They proposed the models to be trained using feedback and delayed supervised sample and it will aggregate each probability to detect fraud.

In another research by (Kurien & Chikkamannur, 2019), the authors have analyzed logistics regression and random forest and it suggested of using neural network for better result and also analyzed the features and subsample ratios for imbalanced datasets. (Maniraj, Saini, Sarkar & Ahmed, 2019) in this paper has implemented anomaly detection to detect fraud. Local outlier factor and isolation forest is applied in data although the accuracy of the algorithms reaches up to 99.6% but its precision is only 33% when the entire dataset is taken for training the model. It is also discussed that the reason behind getting high accuracy is the imbalanced dataset and as precision is less model will not be able to detect much fraud transactions.

(Blagus & Lusa, 2013) has proposed a framework to deal with imbalanced dataset using SMOTE function. SMOTE oversamples the minority class data by using bootstrapping and KNN to create additional synthetic observations of that minority class, which is the fraud data in this case as it is low in number.

As in most of the cases we see that for detecting banking fraud logistics regression, SVM, random forest is used. In this research we have

implemented seven supervised and unsupervised algorithms to find fraud including which are not generally used in fraud detection.

The most important concept we should be clear about is that when we apply machine learning or deep learning model on a highly imbalanced data set, accuracy should not be the only factor for model evaluation (Wu & Radewagen, 2017) as the dataset is skewed getting high accuracy is normal for most of the models. Here false positive ratio(FPR) and false negative ratio(FNR) are also important (Choi & Lee, 2018). So our main objective in this case is to reduce false negative ratio and false positive ratio.

3. Research Methodology

For detecting fraud in this paper machine learning and deep learning models were adopted. This paper has mainly concentrated on supervised and unsupervised machine learning. Again supervised machine learning can be divided to classification and ensemble techniques and unsupervised machine learning can be divided into clustering, anomaly detection and dimensionality reduction technique. One model from every domain is selected for analysis and the models are selected based on the previous research results. From figure 1 we see that in supervised learning logistic regression is selected from classification domain, XGBoost and random forest from ensemble domain, DBscan from clustering domain, isolation forest from anomaly detection and PCA from dimensionality reduction domain. From deep learning artificial neural network is implemented. The main problem for this research is the imbalance dataset (Luque, Carrasco, Martín & de Las Heras-Garcia de Vinuesa, 2019).

3.1 Dataset:

This dataset istaken from Paysim synthetic dataset of mobile money transactions published on kaggle. The dataset contains nearly 6 lakhs data in which only 1.21% data is fraud. Feature 'fraud' is the response variable or the dependent variable and it takes value 1 if a transaction is fraud and 0 otherwise. To balance the skewed dataset SMOTE (Blagus & Lusa, 2013) is used to up-sample the minority data i.e. the fraud data in the dataset.

3.2 Methods:

A detailed research is done on supervise and unsupervised learning models, deep learning model and dimensionality reduction model to find out which models fits best in case of finding fraud detection

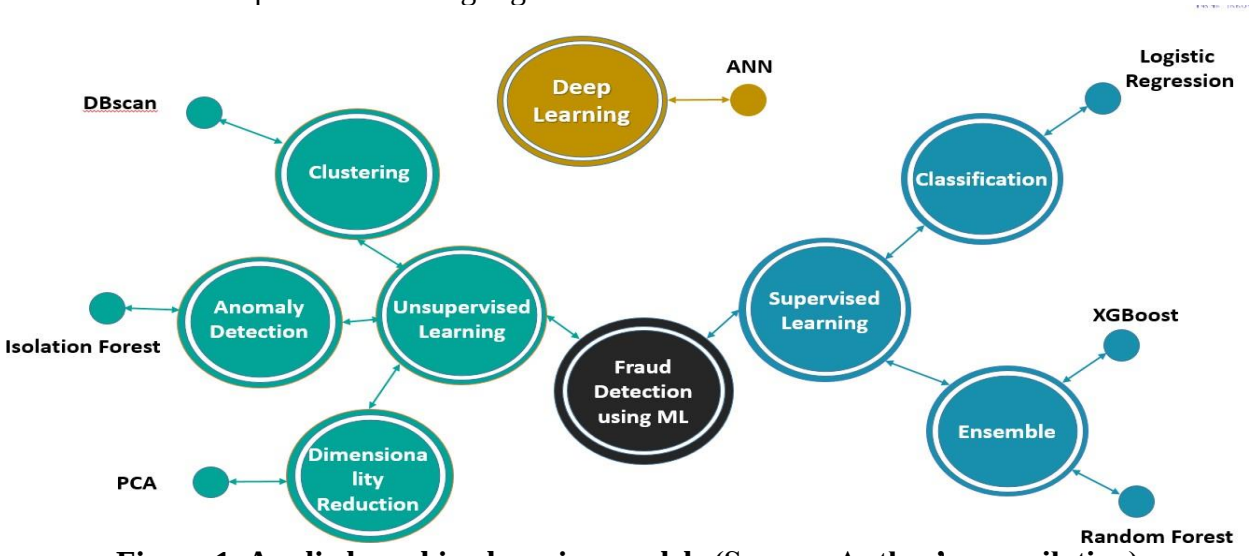


Figure 1. Applied machine learning models (Source: Author's compilation)

The model to be chosen highly depends on the structure of the data but in preliminary research we tried to find out as many relevant models as possible, so that these models can be compared to find the best model, to be applied on the actual dataset provided by the company which will have less amount of false positives and false negatives and having high accuracy with a good kappa value (Sasikala, Biju & Prashanth).

3.2.1 Logistic regression:

Logistic regression (Wright, 1995) is applied when the output of a model expected to be a binary data, where output is zero or one. If the output is 1, the transaction is fraud and if output is

0, the transaction is genuine. In logistic regression, the output returned from the model is the probabilities of the dependent variable class. As the probability remains within zero and one, the same is the range here too. While Implementing the model threshold limit is set to 0.5. So any probability value less than 0.5 will be considered as genuine transaction, probability value less than 0.5 are fraud transaction.

3.2.2 Random Forest:

Random Forest (Breiman, 2001) is a supervised machine learning method (Fawagreh, Gaber & Elyan, 2014), majorly use for classification problems. Random Forest is an ensemble classifier. It uses a combination of trees to improve the prediction result. Each tree checks on different condition. Each tree randomly selects features from the dataset and provides probability of a transaction being fraud or genuine. Final result will be based on the majority votes given by the trees.

Advantages:

1. Random Forest model produces highly accurate predictions by estimating missing data.
2. Feature scaling is not required.
3. With large number of data overfitting can be avoided in random forest.

Drawbacks:

1. Complexity: While computing large memory size is needed
2. Longer training period compare to other machine learning models

3.2.3 XGBoost:

XGBoost (Chen & Guestrin, 2016) is a supervised technique specifically an ensemble technique. The objective of and XGBoost is to minimize the loss function. XGBoost follows the principle of gradient boosting. XGBoost boosting algorithms does not split the tree leaf-wise, it splits depth wise. Due to this sometimes model can overfit on the dataset which can be avoided by correctly setting max depth parameter. When one single ML model is not enough to predict the output accurately XGBoost is used for using combination of methods together.

Advantages:

1. Provides high accuracy with proper model tuning
2. Can handle missing data

Drawbacks: There are no major disadvantages detected while preparing the model only that the model takes quite a while to fail.

3.2.4 DBSCAN (Density-ased spatial clustering of applications with noise):

DBSCAN (Ester, Kriegel, Sander & Xu, 1996) is a density based clustering method which identify cluster in a data set contains noise or outliers. Two important parameters are MinPts & eps. MinPts is the minimum points need to form a cluster. eps is the radius of neighborhood around a point. So basically DBSCAN creates cluster with those points nearer to each other and the points which cannot be included in any clusters are pointed as outliers. Optimal eps value can be found by KNN distances using kNNdistplot() function. The aim is to find the elbow point or knee and that point is the optimal eps value.

3.2.5 PCA (Principle Component Analysis):

Principle Component Analysis (Jolliffe, 2002) is generally use for analyzing dataset that contains multiple variables. The output of PCA creates a more compact feature space by combining the feature values, called the principal components. Principle components are orthogonal to each other's and ordered such that the retention of variation present in the original variables decreases as we move down in the order.

Advantages:

1. Removes co-related features
2. Computation speed is faster in other words, training time reduces with less number of features

Drawbacks:

1. Principal Components are not as interpretable as original features.
2. Data needs to be standardized before implementing PCA.
3. For classification SVM model is used which takes long time to train the model.

3.2.6 Isolation Forest:

Isolation forest (Liu, Ting & Zhou, 2009) is an anomaly detection technique It is an unsupervised machine learning algorithm. Builds on the basis of random forest. This model first selects a point to

isolate. A random feature is selected and then a random split value (between max and min value of that selected feature) is selected. Split value is nothing but the number conditions required to isolate a point from the rest of the dataset.

Outliers are easily separable and does not need more number of cuts where for inliers it's very difficult to separate and needs more number of cuts to isolate it.

Advantages:

1. This method works very well even with a data set containing small number of data points to detect outliers
2. Isolation Forest, we can not only detect anomalies faster but we also require less memory compared to other algorithms.

Drawbacks:

1. The main disadvantage with this method is the procedure this method uses to while branching of the trees is done. This method introduces bias that may lead to reduction in the efficiency of this method in ranking the data points.

3.2.7 Artificial Neural Network:

ANN (von der Malsburg & Christoph, 1986) is mostly used in prediction for nonlinear statistical dataset. That means these is a non- linear relationship between input and output. Also, ANN is generally called as a neural network.

This algorithm assigns random weights to all the linkage at start. Then using input and the nodes it finds activation rate of the hidden node. Then using activation rate of hidden node and linkages to output, it finds activation rate of output node. This way it finds the error rate and recalibrate the linkages between hidden node and output node. Cascades down the error to hidden node and recalibrate the weightage within hidden node and input nodes. This process will continue until a specific criterion is met.

Advantages:

1. Model is fast, takes few second to provide output.
2. Provides high accuracy in the output

Drawbacks:

1. Dependent on JDK 8/9/10/11/12/13. So it is compulsory to have JDK installed in the system. JDK 14 is not supported by ANN, which is latest version of JDK.
2. ANNs can only work with numerical information. Dataset must be changed to numerical values before applying ANN.

3.3 Proposed Work:

Seven different models are analysed in this paper and selected 3 top performing models for further analysis depending on Accuracy, Sensitivity, Specificity, Precision, MCC, BCR, Kappa. Finally compared those three model and come up with the best fitted model.

For the purpose of increasing accuracy of the model in these research random forest algorithm and varImpPlot function (Archer & Kimes, 2008) is used to select the features based on Mean Decrease Accuracy and Mean Decrease Gini.

As the data is imbalanced and with nearly 1% of fraud data so if the model predicts all the transaction as genuine then also it will give 99% accuracy. To avoid this scenario (Blagas & Lusa, 2013) proposed 'SMOTE' function. SMOTE is used to up-sample the minority data to balance the fraud and non-fraud transaction data.

3.4 Confusion Matrix:

Confusion matrix (Beauxis-Aussalet & Hardman, 2014) describe the performance of a classification model for which the actual output is already known. Confusion matrix gives us the count of true positive (fraud transaction is predicted as fraud by the model), true negative (genuine transaction is predicted as genuine by the model), false positive (a genuine transaction is predicted as a fraud by the model) and false negative (fraud transaction is predicted as genuine by the model).

	Genuine	Fraud
Genuine	True Negative	False Positive
Fraud	False Negative	True Positive

From the confusion matrix model evaluation parameters can be calculated.

Accuracy is the most important parameter for model evaluation but this is not true in case of imbalanced dataset. (Wu & Radewagen, 2017) in their study has specified several evaluation metrics used for imbalanced dataset. Here false negative ratio and false positive ratio are two important factor and these should be kept

minimum. Other parameters include Sensitivity, Specificity, Balanced accuracy(BCR) (Layaq & Manjula. 2020), Precision, MCC (Boughorbel, Jarray & El-Anbari, 2017), Kappa.
In this study below parameters were used to evaluate the models.

- Sensitivity = Sensitivity denotes actual fraud transactions predicted as fraud by the model.

Sensitivity =True Positive / (True Positive +False Negative)

- Specificity = Specificity denotes actual negative cases that are correctly predicted as negative

Specificity = True Negative / (False Positive + True Negative)

- Overall accuracy = It is the defined as the fraction of the correct prediction done by the model.

Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)

- Precision = Precision basically measures correctly predicted positive observation as a percentage of total positive prediction. (Shirgave, Awati, More & Patil, 2019) has used precision for different model comparison and proposed model which had high precision and specificity.

Precision = True Positive / (True Positive + False Positive)

- MCC = MCC measures the quality of the binary classification. It ranges from -1 to +1, where +1 identifies a perfect model and -1 a poor model. MCC provides a balanced measure by including True Positive, True Negative, False Positive and False Negative. (Bagga, Goyal, Gupta & Goyal, 2020) in their research has used MCC as a model evaluation matric for imbalanced data.

MCC = True Positive * True Negative - False Positive * False Negative) / sqrt((True Positive + False Positive) * (True Positive +False Negative) * (False Positive + True Negative) * (True Negative + False Negative)

- Balanced Accuracy(BCR) = BCR combines sensitivity and specificity matric and generate a balanced result. It also ranges from -1 to +1. Value near to +1 denotes a good model. (Bagga, Goyal, Gupta & Goyal, 2020) has also proposed BR as a model evaluation matric for imbalanced data.

BCR = ½*(True Positive / Total Positive + True Negative / Total Negative)

- Kappa= Kappa evaluates the prediction performance of the classifier model.

- FPR (False Positive Rate): FPR shows the frequency of classifying a genuine transaction as fraud transaction by the model

FPR = False Positive / (False Positive + True Negative)

- FNR (False Negative Rate): FPR shows the frequency of classifying a fraud transaction as genuine transaction by the model

FNR = False Negative / (False Negative + True Positive)

4. Analysis

4.1. Exploratory Data Analysis

From the bar chart it is seen that the data set is highly skewed dataset. Number of fraud transaction is only 7200 (1.21%). Where the number of genuine transaction is 587443

Figure 2: Count of fraud and non-fraud payment (Source: Author's compilation)

The fraud transaction amount percentage is 16.96. So it is clear that even if the data having less fraud records but the amount transferred in fraud transactions are huge.

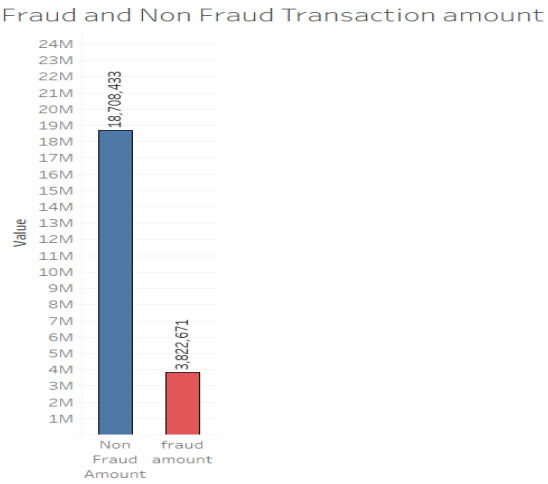


Figure 3: Fraud and non-fraud transaction amount (Source: Author's compilation)

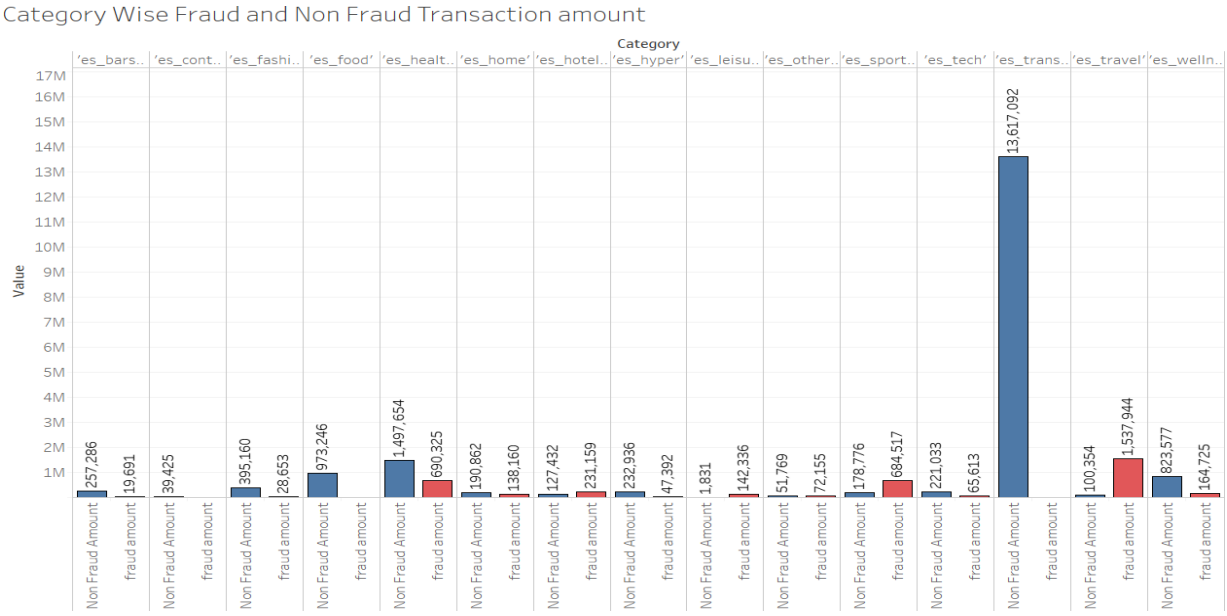


Figure 4: Category wise fraud and non-fraud transaction amount (Source: Author's compilation)

We can analyse that in some category transactions are high but the fraud is less and vice versa. So we can interpret that category is providing valuable insights to the output

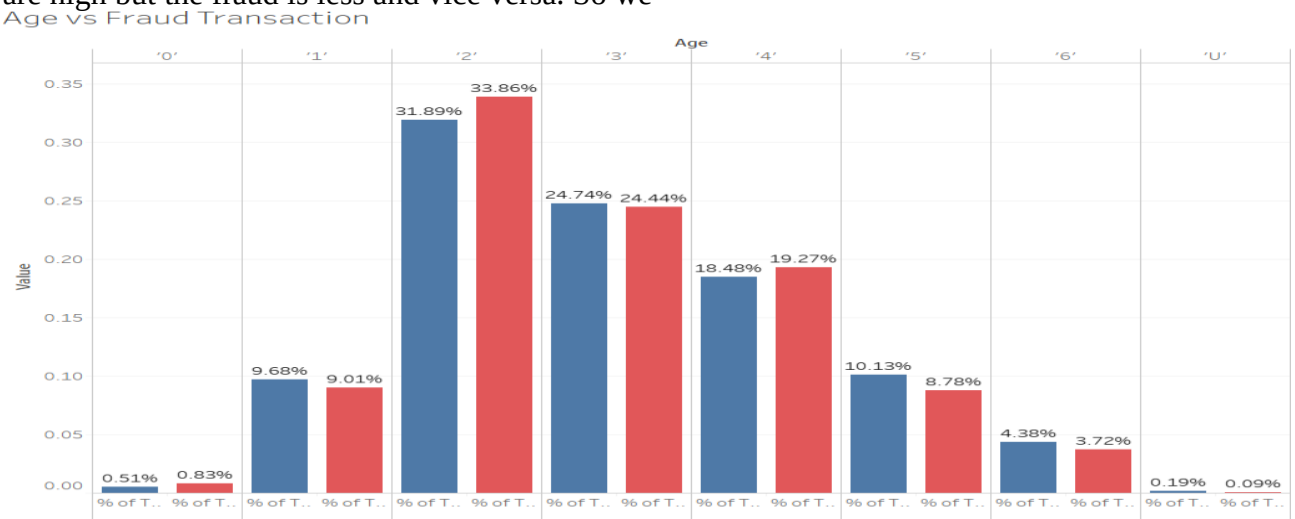


Figure 5: Age vs Fraud transaction (Source: Author's compilation)

Here in the dataset age 0-6 and U refers to-

0 <= 18years	4 -> 46-55 year
1 -> 19-25 year	5 -> 56-65 year
2 -> 26-35 years	6 > 65 years
3 -> 36-45 years	U -> unknown

It has been seen from the bar chart that the group is having large number of transaction is also having high amount of fraud transaction. So age may not be contributing much to the output

4.2. Comparative analysis of the models

In this study the dataset is divided in to training set and test set. Training set contains 75% of the data. The model is built on training set. To cross validate and to check overfitting 10-fold cross validation (Koul, Becchio & Cavallo, 2018) is used for this study. This method divides the dataset into 10 folds of equal size. Out of these 10 folds, 9 folds are used to train the model and one fold is chosen for testing. This procedure is repeated 10 times; each time a different fold is selected for testing purpose. At first the models are compared without using SMOTE then on those models which is giving good result SMOTE is applied and tuned to refine (Probst, Wright & Boulesteix, 2018).

Below seven models have been analysed in detail and three best performing models have been selected for further tuning to get the best possible output from those models. Aim is to reduce false negatives but special care should be taken on keeping false positive as minimum as possible. An overall balance model is expected with maximum prediction power.

The criteria for selecting best models are having high accuracy, sensitivity, specificity, precision, MCC, BCR, Kappa and low FPR and FNR. Below table represent the output of each machine learning model obtained from R programming.

Table 1: Model comparison (Source: Author's compilation)

Model	Accuracy	Sensitivity	Specificity	Precision	MCC	BCR	Kappa
Logistic Regression	0.9932	0.5685	0.9988	0.8581	0.6954	0.7837	0.6806
Random Forest	0.9955	0.7364	0.9988	0.8934	0.8089	0.8676	0.8050
XGBoost	0.9951	0.7110	0.9988	0.8863	0.7915	0.8549	0.7870
DBSCAN	0.9928	0.5688	0.9983	0.8164	0.6781	0.7836	0.6669
Isolation Forest	0.8253	0.9914	0.8231	0.0683	0.2357	0.9073	0.1062
PCA	0.9907	0.2956	0.9998	0.9573	0.5293	0.6477	0.4484
ANN	0.9942	0.7062	0.9978	0.8013	0.7494	0.8520	0.7600

For visualization and easy interpretation above table is plotted with line diagram by the use of tableau.

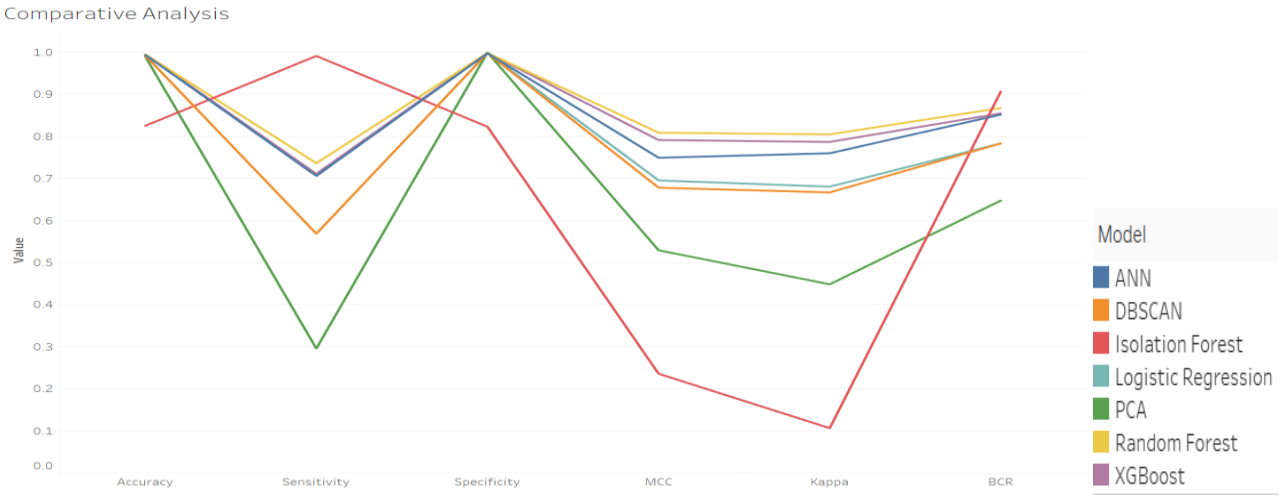


Figure 6: Comparative analysis of models (Source: Author's compilation)

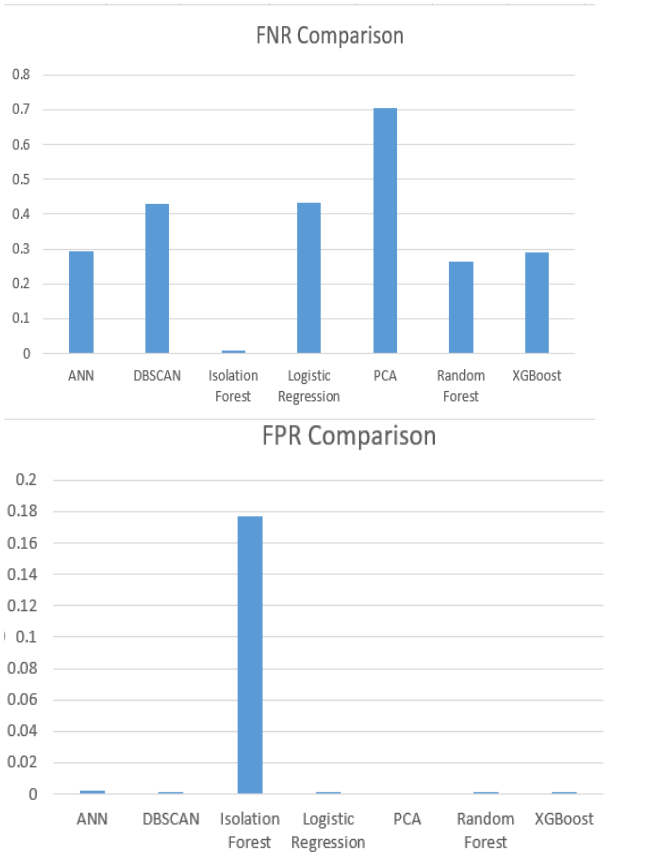


Figure 7: FNR Comparison

(Source: Author's compilation)

Figure 8: FPR Comparison

It can be interpreted from the above diagram that even if isolation forest is having high sensitivity and BCR but this model is performing poorly as it is having lowest specificity, MCC and Kappa above all though false negative ratio is less but false positive ratio is too high which indicates that isolation forest is not fit for detecting frauds.
For all other model false positive is low. In this paper we have not considered PCA also for future analysis for its high FNR and low sensitivity and low kappa value. As false negative is high the model is not fulfilling the objective of this study.
Though all other models are performing well but for further analysis random forest, XGBoost and ANN is selected for having less False Positive Ratio & False Negative Ratio and having top three highest value in all other performance evaluation indicator (accuracy, sensitivity, specificity, precision, BCR, MCC, kappa value).

4.3 Model Tuning

To up-sample the fraud data SMOTE is used on training set for better prediction. Before using smote dataset had 3412 fraud transaction in training set, after upsampling the data with smote now we have 174012 fraud transaction. Machine learning model's prediction power increases with large dataset. In this paper shortlisted models are built on this upsample data.

4.3.1 Random Forest:

For random forest model tuning (Probst, Wright & Boulesteix, 2018) two hyper parameters are finally selected after several computations with different hyper parameters. Those are:

1. Mtry is the number of variables randomly sampled at each split. For better result optimum mtry should be searched for which the out of bag error is minimum.
2. Ntree is the number of trees to grow. (Hastie & Friedman, 2009, page 596) also state that "the average of fully grown trees can result in too rich a model, and incur unnecessary variance". So there can be some overfitting due to use of fully grown trees which may be showing up if more number of trees are added. To avoid this optimum number of trees should be selected.

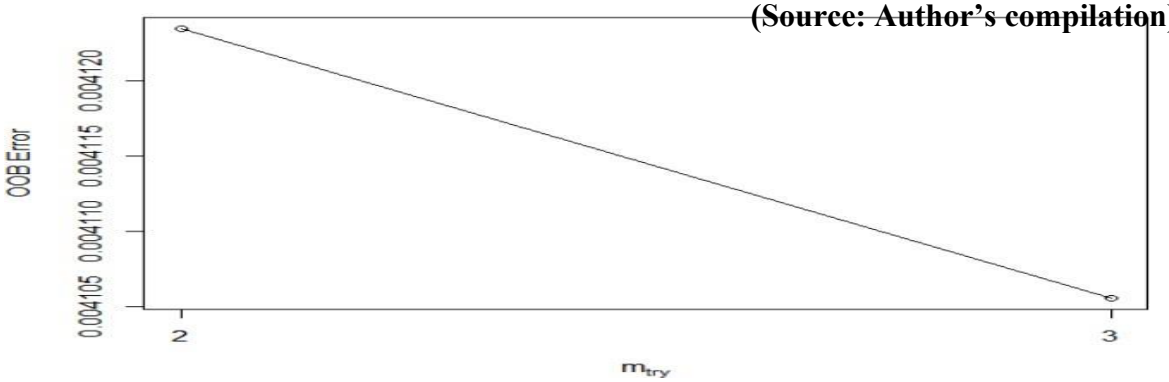


Figure 9: mtry vs out of bag error plot
(Source: Author's compilation)

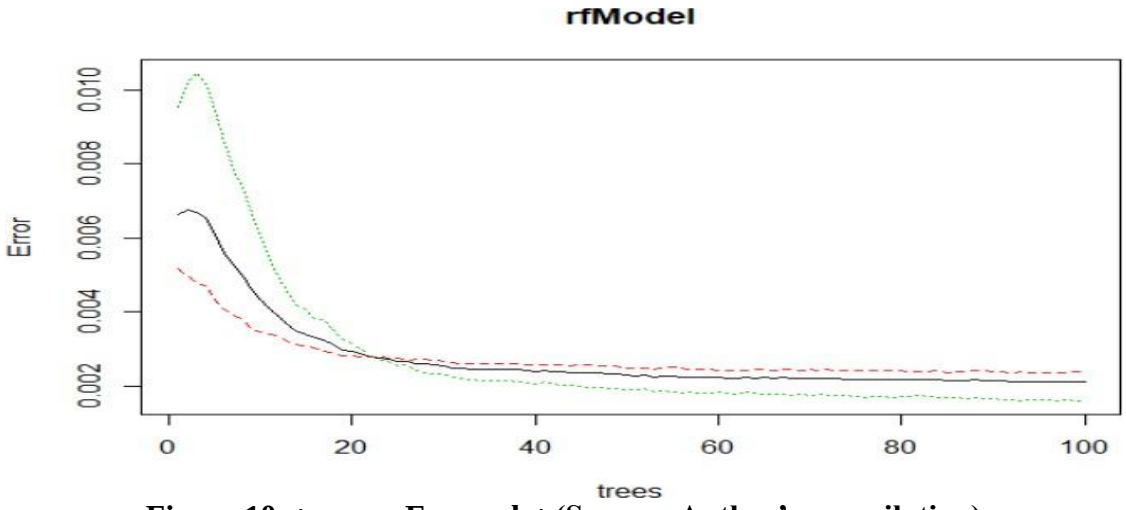


Figure 10: trees vs Error plot (Source: Author's compilation)

From fig 9 it is seen that at mtry=3 misclassification (OOB Error) is minimum. So the optimal mtry is 3 which will be used for model creation.

Ntree is set as 100 as from the graph it is seen that at ntree =100 the error is minimum.

Post application of smote and tuning a balanced output is obtained with False Positive rate and False Negative rate of 0.77% and 9.3%. Final accuracy obtained is 99.12% with a balanced accuracy of 94.94%

4.3.2 XGBoost:

As XGBoost only accept numerical value so the dependent variable "fraud" should be converted to numeric value along with all the dependent variable.

Post application of Smote below hyper parameters are selected for model tuning (Li, 2019) to get the best result.

```
# Fitting XGBoost to the Training set
#install.packages('xgboost')
library(xgboost)
classifier = xgboost(data = as.matrix(smoted_data[-8]),
                     label = smoted_data$fraud,
                     nrounds = 600,
                     objective="binary:logistic",
                     booster='gbtree',
                     eval_metric="auc",
                     verbose= FALSE,
                     eta=0.3,
                     gamma=1,
                     lambda=5,
                     subsample=.8,
                     min_child_weight=1,
                     max_depth=4)
```

Figure 11: XGBoost model tuning (Source: Author's compilation)

After training the model it is tested with test set and from the output false positive rate and false negative rate is calculated that is 0.45% and 12.4%. It can be observed that after rectifying the model false negative rate has decreased but false positive rate has increased a bit. An overall accuracy of 99.4% is obtained but balanced accuracy we got 93.55%.

4.3.3 Artificial Neural Network:

For applying ANN, data set must be scaled except the dependent variable which should be in factor datatype. For model tuning number of hidden layer selected is 7 and number of nodes in hidden layer is 3 to get best output. One epoch is when an entire dataset is passed through the neural network once. With increasing number of epocs better output can be obtained but after a certain epocs model starts overfitting on the dataset. So model needs to be checked with test set to stop overfitting. If the model is providing good result on training set but in case of test set its not performing well then it can be assumed that the model is overfitted on training set. Below is the snippet of the model building and tuning.

```
library(h2o)
h2o.init(nthreads = -1)
model = h2o.deeplearning(y = 'fraud',
                         training_frame = as.h2o(training_set),
                         activation = 'tanh',
                         hidden = c(7,3),
                         distribution= 'Auto',
                         epochs = 300,
                         seed=5,
                         nfolds=4,
                         model_id = "deep_model",
                         train_samples_per_iteration = -1)
```

Figure 12: ANN model tuning (Source: Author's compilation)

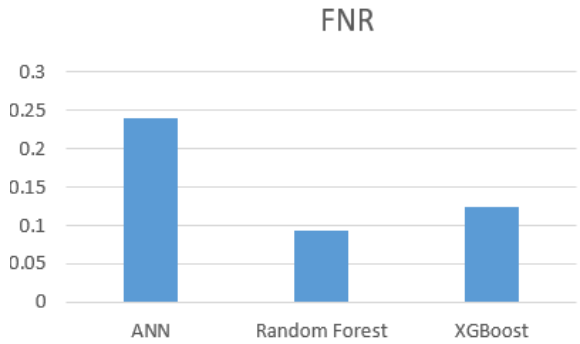
Post testing the model false positive rate and false negative rate is calculated that is 0.45% and 12.4%. Overall accuracy obtained is 99.4% with a balanced accuracy of 87.85%. In this model SMOTE is not used as application of smote is decreasing the performance.

4.4 Comparison of analyzed model:

False positive rate(FPR), False Negative Rate(FNR) and BCR is calculated for all the three model and represented in below table for comparison.

Table2: Model comparison based on FPR FNR & BCR (Source: Author's compilation)

Model	FPR	FNR	BCR
Random Forest	0.77%	9.33%	0.949469
XGBoost	0.45%	12.44%	0.935504
ANN	0.24%	24.06%	0.878531



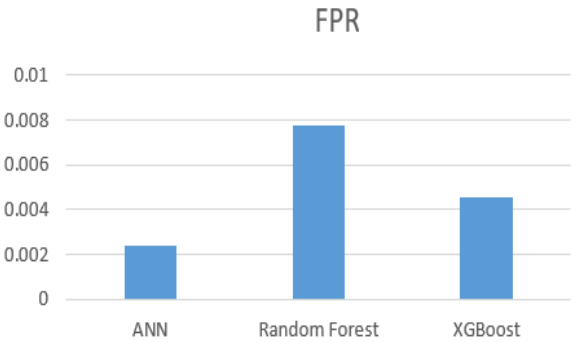


Figure 13: FNR comparison
Figure 14: FPR comparison
(Source: Author's compilation)

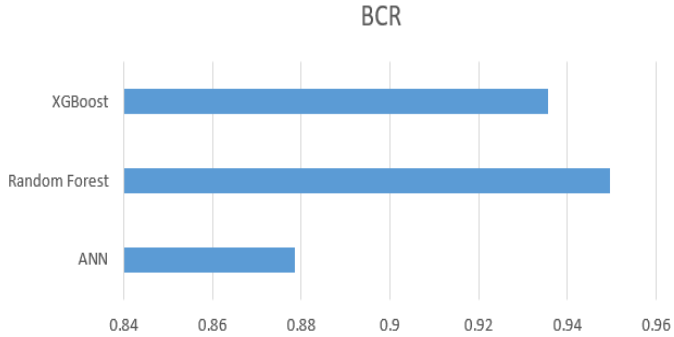


Figure 15: BCR comparison (Source: Author's compilation)

5. Conclusion and Recommendations

In this paper different supervised and unsupervised models are studied for detecting fraud. From figure 13 and figure 14 it can be seen that random forest is having least false negative rate (9.33%) but having comparatively high false positive rate. Our aim is to keep both these as minimum. As from figure 13 and figure 14 best model cannot be concluded so to compare the performance a balanced matric is used: BCR. BCR ranges from -1 to +1. +1 signifies a perfect model and -1 signifies poor model. We can see in BCR comparison graph that all the three model is having BCR value close to 1 but random forest is having the highest BCR among the three models.

Now the question is should we blindly use random forest model for fraud detection in banking transaction?

It totally depends on priority. Obviously random forest is performing best but it is also having high FPR. So If detecting fraud is the priority where a little genuine transaction can be compromised, random forest can be applied. If protecting genuine transaction is the ultimate priority where a little fraud transaction can be compromised, XGboost should be applied. As for any bank losing even a single customer can be costly.

This paper has tried to cover every domain of machine learning supervised, unsupervised, ensemble and neural network. In neural network ANN is implemented and a comparatively good output is obtained with respect to other methods analyzed here. In future studies ANN can be combined with genetic algorithm to obtain higher accuracy and low false negative ratio.

Acknowledgement

The authors wish to acknowledge Symbiosis Institute of Operation Management for providing the different facilities.

Conflict of Interest: Not Applicable

Funding: Not applicable

Ethical approval: Not applicable

References

[1] Sadgali, Imane & Sael, Nawal & Benabbou, Faouzia. (2019). Performance of machine learning techniques in the detection of financial frauds.

[2] Bagga, Siddhant & Goyal, Anish & Gupta, Namita & Goyal, Arvind. (2020). Credit Card Fraud Detection using Pipeling and Ensemble Learning. Procedia Computer Science. 173. 104-112. 10.1016/j.procs.2020.06.014.

[3] Olowookere, Toluwase & Adewale, Olumide. (2020). A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. Scientific African. e00464. 10.1016/j.sciaf.2020.e00464.

[4] Raghavan, Pradheepan & Gayar, Neamat. (2019). Fraud Detection using Machine Learning and Deep Learning. 334-339. 10.1109/ICCIKE47802.2019.9004231.

[5] Amarasinghe, Thushara & Aponso, Achala & Krishnarajah, Naomi. (2018). Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions. ICMLT '18: Proceedings of the 2018 International Conference on Machine Learning Technologies. 12-17. 10.1145/3231884.3231894.

[6] Shirgave, Suresh & Awati, Chetan & More, Rashmi & Patil, Sonam. (2019). A Review On Credit Card Fraud Detection Using Machine Learning. International Journal of Scientific & Technology Research. 8. 1217-1220.

[7] Kaithekuzhical Leena Kurien & Dr. Ajeet Chikkamannur, Detection And Prediction Of Credit Card Fraud Transactions Using Machine Learning, International Journal Of Engineering Sciences & Research Technology,2019, ISSN: 2277-9655

[8] Maniraj, S & Saini, Aditya & Ahmed, Shadab & Sarkar, Swarna. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research and. 08. 10.17577/IJERTV8IS090031.

[9] Blagus, Rok & Lusa, Lara. (2013). SMOTE for High-Dimensional Class-Imbalanced Data. BMC bioinformatics. 14. 106. 10.1186/1471-2105-14-106.

[10] Probst, Philipp & Wright, Marvin & Boulesteix, Anne-Laure. (2018). Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 10.1002/widm.1301.

[11] Grand View Research, Fraud Detection & Prevention Market Size, Share & Trends Analysis Report By Component, By Solutions, By Services (Professional Services, Managed Services), By Application, By Organization, By Vertical, And Segment Forecasts, 2019 – 2025, 2019

[12] Fawagreh, Khaled & Gaber, Mohamed & Elyan, Eyad. (2014). Random Forests:

From Early Developments to Recent Advancements. Systems Science & Control Engineering. 2. 10.1080/21642583.2014.956265.

[13] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[14] Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.

[15] Ester, Martin & Kriegel, Hans-Peter & Sander, Joerg & Xu, Xiaowei. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD. 96. 226-231.

[16] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.

[17] Jolliffe, Ian. (2002). Principle Component Analysis. 24. 417-441.

[18] von der Malsburg, Christoph. (1986). Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Brain Theory. 245-248. 10.1007/978-3-642-70911-1_20.

[19] Li, Yingchang. (2019). [R Code] Tuning Randomforest and XGBoost.

[20] Beauxis-Aussalet, Emma & Hardman, Lynda. (2014). Simplifying the Visualization of Confusion Matrix.

[21] Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE 12(6): e0177678. https://doi.org/10.1371/journal.pone.0177678

[22] Luque, A. & Carrasco, Alejandro & Martín, Alejandro & de Las Heras-Garcia de Vinuesa, Ana. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 91. 6829. 10.1016/j.patcog.2019.02.023.

[23] Ali, Haseeb & Salleh, Mohd & Saedudin, Rd & Hussain, Kashif & Mushtaq, Muhammad. (2019). Imbalance class problems in data mining: A review. Indonesian Journal of Electrical Engineering and Computer Science. 14. 10.11591/ijeecs.v14.i3.pp1552-1563.

[24] Layaq & Manjula. 2020. A Recapitulation of Imbalanced Data Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (p. 217–244). American Psychological Association.

[25] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

[26] Grobman, Steve. (2018). Impact of Cybercrime Why Cyber Espionage isn't Just the Military's Problem

[27] Sasikala, B & Biju, Vinai & Prashanth, C.. (2017). Kappa and accuracy evaluations of machine learning classifiers. 20-23. 10.1109/RTEICT.2017.8256551.

[28] Archer, Kellie & Kimes, Ryan. (2008). Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis. 52. 2249-2260. 10.1016/j.csda.2007.08.015.

[29] Koul, Atesh & Becchio, Cristina & Cavallo, Andrea. (2018). Cross-Validation Approaches for Replicability in Psychology. Frontiers in Psychology. 9. 1117. 10.3389/fpsyg.2018.01117.

[30] Wu, Ye & Radewagen, Rick. (2017). 7 Techniques to Handle Imbalanced Data. https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

[31] Choi, Dahee & Lee, Kyungho. (2018). An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation. Security and Communication Networks. 2018. 1-15. 10.1155/2018/5483472.