IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey of Credit Card Fraud Detection using Supervised Machine Learning Algorithms

Sai Suneetha Mandava

M.tech, Department of Information Technology and Computer Applications , Andhra University,Visakhapatnam, AP, India

Email : suneethamandava184@gmail.com

.

ABSTRACT: The rapid growth in E-Commerce industry has lead to an exponential increase in the use of credit cards for online purchases and for different types of transactions . So there will be more chances for occurring fraud. Banks have many and enormous databases. Important business information can be extracted from these data stores. Fraud is an issue with far reaching consequences in the banking industry, government, corporate sectors and for ordinary consumers. Increasing dependence on new technologies such as cloud and mobile computing in recent years has encountered the problem. Physical detections are not only time consuming they are costly and they don't give accurate result. Fraud is any malicious activity that aims to cause financial loss to the other party. As the use of digital money or plastic money even in developing countries is on the rise so is the fraud associated with them. Frauds caused by Credit Cards have costs consumers and banks billions of dollars globally. Even after numerous mechanisms to stop fraud, fraudsters are continuously trying to find new ways and tricks to commit fraud. . It has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used..We have explained various techniques available for a fraud detection system such as Random Forest Classifier, K-nearest neighbors Classifier, Decision Tree Classifier, Gaussian Naive Bayes and Logistic Regression. These techniques are applied on both unbalanced data and balanced data and we provide a survey and a comparative analyses of techniques for both unbalanced data and balanced data, together with evaluation metrics. Dataset of credit card transactions is collected from kaggle and it contains a total of 2,84,808 credit card transactions of a European bank data set. It considers fraud transactions as the "class 1" and genuine ones as the "class 0" .The data set is highly imbalanced, it has about 0.172% of fraud transactions and the rest are genuine transactions. So to balance the dataset SMOTE oversampling technique has been applied to the data set, which resulted in 50% of fraud transactions and 50% genuine ones. We trained five techniques and evaluate each methodology based on certain criteria namely sensitivity, precision, accuracy and ROC AUC. Based on the criteria of different techniques ,the best technique for detecting credit card fraud is choosen. The five techniques are applied for the dataset and work is implemented in python language.

KEYWORDS: Classification, Random Forest Classifier, K-nearest neighbors Classifier, Decision Tree Classifier, Gaussian Naive Bayes and Logistic Regression.

I. INTRODUCTION

Fraud refers to the abuse of a profit organization's system without necessarily leading to direct legal concerns. Fraud is a universal act in order to deceive another person or organization for financial benefits. Credit card fraud detection is the process of identify those transactions that falls into two classes of lawful and fake transactions. These kind of frauds can be broadly classified into three categories that is traditional card related frauds, merchant related frauds and internet frauds. The fraud which is committed by individuals exterior to the organization is called as customer fraud or external fraud where when a fraud is committed by top-level management is known as management fraud or internal fraud.

Fraud detection being part of all the overall fraud control, automates and helps reduce the manual parts of a screening process. Credit card fraud is an unauthorized account activity by a person for which the account is not proposed. It is also defined as when an individual uses another individual credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card being used. And the persons using the card has not at all having the piecing together with the card holder or the issuer has no objective of making the repayments for the purchase they done. It involves identifying fraud as quickly as possible once it has been performed. Fraud detection methods are continuously developed to define offenders in familiarizing their strategies. Data mining refers to extract or mining knowledge from large amount of data.

The problem of fraud is a serious issue in e-banking services that threaten credit card transactions especially. Fraud is an intentional deception with the purpose of obtaining financial gain or causing loss by implicit or explicit trick. Fraud is a public law violation in which the fraudster gains an unlawful advantage or causes unlawful damage. The estimation of amount of damage made by fraud activities indicates that fraud costs a very considerable sum of money. Credit card fraud is increasing significantly with the development of modern technology resulting in the loss of billions of dollars worldwide each year. Fraud detection involves identifying scarce fraud activities among numerous legitimate transactions as quickly as possible. Fraud detection methods are developing rapidly in order to adapt with new incoming fraudulent strategies across the world. But, development of new fraud detection techniques becomes more difficult due to the severe limitation of the ideas exchange in fraud detection. On the other hand, fraud detection is essentially a rare event problem, which has been variously called outlier analysis, anomaly detection, exception mining, mining rare classes, mining imbalanced data etc. The number of fraudulent transactions is usually a very low fraction of the total transactions. Hence the task of detecting fraud transactions in an accurate and efficient manner is fairly difficult and challengeable. Therefore, development of efficient methods which can distinguish rare fraud activities from billions of legitimate transaction seems essential.

II. DIFFICULTIES OF CREDIT CARD FRAUD DETECTION

Fraud detection systems are prune to several difficulties and challenges enumerated bellow. An effective fraud detection technique should have abilities to address these difficulties in order to achieve best performance.

Imbalanced data: The credit card fraud detection data has imbalanced nature. It means that very small percentages of all credit card transactions are fraudulent. This causes the detection of fraud transactions very difficult and imprecise.

Different misclassification importance: In fraud detection task, different misclassification errors have different importance. Misclassification of a normal transaction as fraud is not as harmful as detecting a fraud transaction as normal. Because in the first case the mistake in classification will be identified in further investigations.

Overlapping data: Many transactions may be considered fraudulent, while actually they are normal (false positive) and reversely, a fraudulent transaction may also seem to be legitimate (false negative). Hence obtaining low rate of false positive and false negatives is a key challenge of fraud detection systems [4, 5, and 6].

Lack of adaptability: classification algorithms are usually faced with the problem of detecting new types of normal or fraudulent patterns. The supervised and unsupervised fraud detection systems are inefficient in detecting new patterns of normal and fraud behaviors, respectively.

Fraud detection cost: The system should take into account both the cost of fraudulent behavior that is detected and the cost of preventing it. For example, no revenue is obtained by stopping a fraudulent transaction of a few dollars [5, 7].

Lack of standard metrics: there is no standard evaluation criterion for assessing and comparing the results of fraud detection systems

III. CREDIT CARD FRAUD DETECTION TECHNIQUES

The credit card fraud detection techniques are classified in two general categories: fraud analysis (misuse detection) and user behavior analysis (anomaly detection).The first group of techniques deals with supervised classification task in transaction level. In these methods, transactions are labelled as fraudulent or normal based on previous historical data. This dataset is then used to create classification models which can predict the state (normal or fraud) of new records. There are numerous model creation methods for a typical two class classification task such as rule induction [1], decision trees [2] and neural networks [3].This approach is proven to reliably detect most fraud tricks which have been observed before [4], it also known as misuse detection.

The second approach deals with unsupervised methodologies which are based on account behaviour. In this method a transaction is detected fraudulent if it is in contrast with user's normal behaviour. This is because we don't expect fraudsters behave the same as the account owner or be aware of the behaviour model of the owner [5].To this aim, we need to extract the legitimate user behavioural model (e.. user profile)for each account and then detect fraudulent activities according to it. Comparing new behaviours with this model, different enough activities are distinguished as frauds. The profiles may contain the activity information of the account; such as merchant types, amount, location and time of transactions, [6].This method is also known as anomaly detection. It is important to highlight the key differences between user behaviour analysis and fraud analysis approaches. The fraud analysis method can detect known fraud tricks, with a low false positive rate. These systems extract the signature and model of fraud tricks presented in oracle dataset and can then easily determine exactly which frauds, the system is currently experiencing. If the test data does not contain any fraud signatures, no alarm is raised. Thus, the false positive rate can be reduced extremely. However, since learning of a fraud analysis system (i.e. classifier) is based on limited and specific fraud records, it cannot detect novel frauds. As a result, the false negatives rate may be extremely high depending on how ingenious are the fraudsters. User behaviour analysis, on the other hand, greatly addresses the problem of detecting novel frauds.

These methods do not search for specific fraud patterns, but rather compare incoming activities with the constructed model of legitimate user behaviour. Any activity that is enough different from the model will be considered as a possible fraud. Though, user behaviour analysis approaches are powerful in detecting innovative frauds, they really suffer from high rates of false alarm. Moreover, if a fraud occurs during the training phase, this fraudulent behaviour will be entered in baseline mode and is assumed to be normal in further analysis[7].In this paper we will introduce some supervised machine learning algorithms for credit card fraud detection and performance analysis of each algorithm.

IV.RELATEDWORK

Financial fraud detection is an evolving field in which it is desirable to stay ahead of the perpetrators. Additionally, it is evident that there are still facets of intelligent fraud detection that have not been investigated. Survey of fraud detection says that there are different types of frauds and there are different computational methods for detecting the financial frauds done by the fraudsters. Different computational methods have been stated for detecting the fraud by computing various parameters for each kind of algorithm and the computing time representing with graphical view. They had taken the different datasets german credit card dataset and from different countries like china also from the available datasets they had developed computational methods for detecting the fraud and stating which algorithm is accurate. In existing system fraud detection is done using ID3 and support vector machine algorithms and a survey stating the percent of fraud happened and defining different parameters and comparing different parameters for the algorithms. Fraud detection is an important part of the modern finance industry. The system which I had proposed is fraud detection using supervised learning algorithms that is Decision tree, Random Forest, Logistic Regression, K-nearest neighbor and Naive Bayes classifier and comparing these algorithms with the accuracy acquired by these five learning algorithms. Though their performance differed, each technique was shown to be reasonably capable at detecting various

forms of financial fraud. In particular, the ability of the computational methods such as Decision trees and Bayesian classifier to learn and adapt to new techniques is highly effective to the evolving tactics of fraudsters. With the available dataset we can classify whether the user is good or bad that mean whether he will be able to repay the loan or not if he is a good user it is represented with the positive count and if the user is bad the value is represented as negative count and from these values we can calculate the sensitivity and efficiency and represent them in a graphical representation.

V. SUPERVISED LEARNING ALGORITHMS

Supervised learning algorithms are defined as the desired output is known for the input provided. In these kind of algorithms we have an input and the desired output is known and we need to map a function for these values . In these supervised learning algorithms predictions are made on the known training dataset and it will be accurate. These learning algorithms are further grouped into regression and classification problems. The supervised learning algorithm analyses the training dataset and produces a classifier. For this initially we need to collect the accurate training dataset and we need to find the accuracy of the function. It is the machine learning task of inferring a function from supervised training.

Random Forest

It is a supervised algorithm. It is a tree based algorithm. It creates several decision trees and combines their outputs to produce a good model. The process of combining the decision trees is known as ensemble process. Advantages and Disadvantages of Random Forest. It is robust to correlated predictors. It is used to solve both regression and classification problems. It can be also used to solve unsupervised ML problems. It can handle thousands of input variables without variable selection. It can be used as a feature selection tool using its variable importance plot. Intakes care of missing data internally in an effective manner. The Random Forest model is difficult to interpret. It tends to return erratic predictions for observations out of range of training data. For example, the training data contains two variable x and y. The range of x variable is 30 to 70. If the test data has x = 200, random forest would give an unreliable prediction. It can take longer than expected time to computer a large number of trees

K-Nearest Neighbor

It is one of the most used algorithms for both classification and regression predictive problems. Its performance depends on three factors: the distance metrics, the distance rule and the value of K. Distance metrics gives the measure to locate nearest neighbors of any incoming data point. Distance rule helps us to classify the new data point into a class by comparing its features with that of data points in its neighborhood. And the value of K decides the number of neighbors with whom to compare. The important question is how do we choose the factor K? In order to obtain the optimal value of K, the training and validation is segregated from the initial dataset. Now a graph based on the validation error curve is plotted to achieve the value of K. This value of K should be used for all predictions. We calculate the dominant class in the vicinity of any new transaction and classify the transaction to belong to that dominant class

Naive Bayes

It is based upon the Bayes Theorem of conditional probability; hence it is a probabilistic model that is used for automated detection of various events. It consists of nodes and edges, wherein the nodes represent the random variables and the edges between the nodes represent the relationships between these random variables and their probabilistic distribution. We calculate predefined minimum and maximum value of probabilities of a transaction being fraud or legal. Then for a new incoming transaction we see that whether it's probability of being legal is less than the minimum defined value for legal transaction and is greater than the maximum defined value for a fraud transaction. If true then the transaction is classified as a fraud

Decision Tree

It is a computational tool for classification and prediction. A tree comprises of internal nodes which denote a test on an attribute, each branch denotes an outcome of that test and each leaf node (terminal node) holds a class label. It recursively partitions a dataset using either depth first greedy approach or breadth first greedy approach and stops when all the elements have been assigned a particular class. For the partition rule to be efficient it must separate the data into groups where a single class predominates in each group. In other words, the best partition will be the one in which the subsets do

not overlap i.e. They are clearly disjoint to a maximum amount.

Logistic Regression

To combat the anomalies of linear regression where it gave values greater than 1 and less than 0, logistic regression comes into play. Despite the name being regression, LR is used for classification problems for predicting binomial and multinomial outcomes, having the goal of estimating the values of parameter's coefficients using the sigmoid function. Logistic regression is used for clustering and when a transaction is ongoing it examines the values of its attributes and tells whether the transaction should proceed or not.

VI. DATA SET USED

Credit Card fraud detection uses the records of European cardholders who made transactions using their credit cards in the month of September 2013. The dataset which has been selected and used holds the records of European cardholders who made transactions using their credit cards in the month of September 2013. This dataset holds the record of transactions that were made within two days and total transactions made within two days are 284,807 transactions from which 492 transactions were found as fraudulent which makes the dataset highly imbalanced, more oriented as the positive class i.e., fraud transactions are 0.172% out of total transactions. And the dataset is in CSV format i.e., in a format where the data values are separated by commas.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Dealing with Imbalanced Data

Resampling data is one of the most commonly preferred approaches to deal with an imbalanced dataset. There are broadly two types of methods for this i) Undersampling ii) Oversampling. In most of the cases, oversampling is preferred over undersampling techniques. The reason being, in undersampling we tend to remove instances from data that may be carrying some important information. Here we are using SMOTE oversampling technique.

SMOTE: Synthetic Minority Oversampling Technique:SMOTE is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. This approach is inspired by a technique that proved successful in handwritten character recognition (Ha & Bunke, 1997). They created extra training data by performing certain operations on real data. In their case, operations like rotation and skew were natural ways to perturb the training data. We generate synthetic examples in a less application-specific manner, by operating in "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Our implementation currently uses five nearest neighbors. For instance, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general.

VII. COMPARATIVE ANALYSIS

In order to compare various techniques we calculate the true positive, true negative, false positive and false negative generated by a system or an algorithm and use these in quantitative measurements to evaluate and compare performance of different systems. True Positive (TP) is number of transactions that were fraudulent and were also classified as fraudulent by the system. True Negative (TN) is number of transactions that were legitimate and were also classified as legitimate. False Positive (FP) is number of transactions that were legitimate but were wrongly classified as fraudulent transactions. False Negative (FN) is number of transactions that were fraudulent but were wrongly classified as legitimate transactions by the system. The various metrics for evaluation are:

1. Accuracy is the fraction of transactions that were correctly classified. It is one of the most powerful and commonly used evaluation metrics.

Accuracy (ACC)/Detection rate = (TN + TP) / (TP + FP + FN + TN)

2. Precision also known as detection rate is the number of transactions either genuine or fraudulent that were correctly classified.

Precision/Detection rate/Hit rate = TP / TP + FP

3 .Sensitivity or Recall is the fraction of abnormal records (the records that have maximum chances of being fraudulent) correctly classified by the system.

True positive rate/Sensitivity = TP / TP + FN

4. F1-score, is a measure of a model's accuracy on a dataset and is defined as the harmonic mean of the model's precision and recall.

F1-score=2*((Precision*Recall)/(Precision + Recall))

ROC Curve-Receiver Operating Characteristic curve

It is a graph displaying the performance of a classification model. It is a very popular method to measure the accuracy of a classification model. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

True positive rate = TP / TP + FN

False positive rate = FP / FP + TN

For different threshold values we will get different TPR and FPR. So, in order to visualise which threshold is best suited for the classifier we plot the ROC curve.

- When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.

- When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.

- When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

Table 1: Performance analysis of different Supervised Machine Learning Algorithms

Data Type	Classifiers	Accuracy	Recall	Precision	F1-Score	ROC - AUC
Unbalanced Data	Random Forest	100	78	98	87	0.89
	K-nearest neighbour	99.8	9	93	17	0.55
	Decision Tree	99.9	75	77	76	0.88
	Gaussian Naïve Bayes	99.3	63	16	25	0.81
	Logistic Regresssion	99.9	63	66	25	0.82
After SMOTE	Random Forest	100	100	99	99	1.00
	K-nearest neighbour	96.8	98	95	96	0.97
	Decision Tree	99.8	99	99	99	1.00
	Gaussian Naïve Bayes	86.7	74	99	84	0.87
	Logistic Regresssion	97.2	96	98	97	0.97

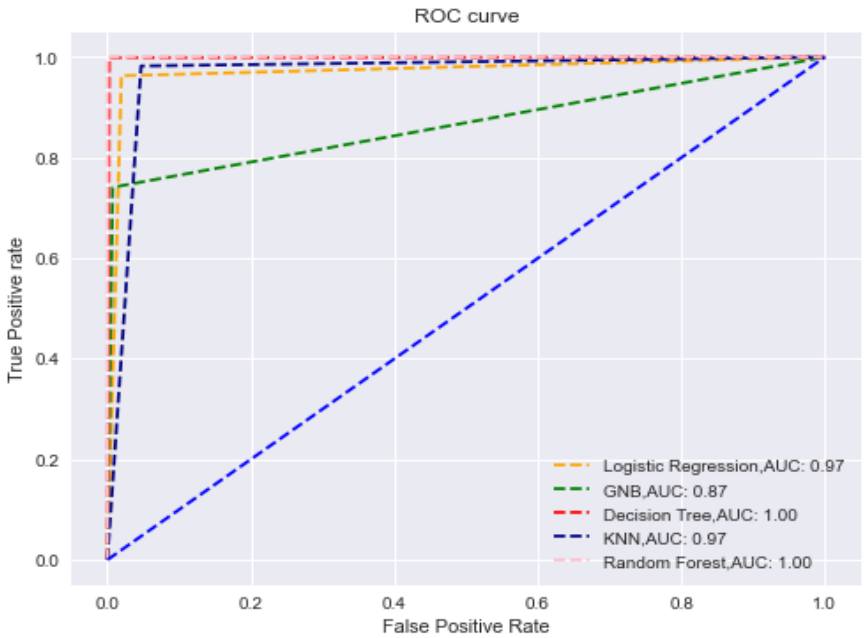

Figure-1: ROC Curve of ML Algorithms on balanced data

VIII. CONCLUSION

Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. This happens because a very minuscule number of transactions from the total transactions are actually fraudulent in nature. So to balance the dataset we used SMOTE oversampling technique. This work gives contribution towards the credit card fraud detection using the supervised learning algorithms like Logistic regression, Decision Tree, Random forest, Gaussian Naïve Bayes and K-nearest Neighbor. Accuracy, Precision, Recall, F1-score, AUC are used to evaluate the performance for the proposed system. The accuracy for logistic regression, Decision tree, Gaussian Naïve Bayes and K-nearest Neighbour and random forest classifier are 97.2, 99.8, 86.7, 96.8 and 100 respectively. By comparing all the five methods, we found that random forest, classifier is better than the remaining 4 models.

REFERENCES

[1] Https ://www.analyticsvidhya.com/blog/2016/03/practical-guide-dealimbalanced- classification-problems/.[Accessed: Oct 12, 2019].

[2] Https://www.ritchieng.com/machine-learning-evaluate-classificationmodel/.[Accessed: Oct 12, 2019].

[3] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi," Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018.

[4] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, ''Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-9.

[5] S. Dhankhad, E. Mohammed and B. Far, ' Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 122-125.

[6] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, ''Random forest for credit card fraud detection," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1-6.

[7] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, no. 3, pp. 602-613, 2011.

[8] K. Chaudhary, J. Yadav, and B. Mallick, "A review of Fraud Detection Techniques: Credit Card," Int. J. Comput. Appl., vol. 45, no. 1, pp. 975-8887, 2012.

[9] F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System," vol. 6, no. 3, pp. 311-322, 2011.

[10] O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," vol. 10, no. 1, pp. 23-27.

[11] C. Phua, D. Alahakoon and V. Lee, "Minority report in fraud detection", ACMSIGKDD Explorations Newsletter, vol. 6, no. 1, p. 50, 2004.

[12] N. Sethi and A. Gera, "A Revived Survey of Various Credit Card Fraud Detection Techniques", International Journal of Computer Science and Mobile Computing, vol. 3, no. 4, pp. 780-791, 2014.

[13] J. Awoyemi, A. Adetunmbi and S. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017.

[14] http://www.ulb.ac.be/di/map/adalpozz/imbalancedatasets.zip.[Accessed: Oct 10, 2019].

[15] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.

[16] S.Dutt, A.K.Das and S.Chandramouli, Machine Learning. Pearson 2011255_200967