Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

ONLINE BANKING FRAUD DETECTION USING DISTRIBUTED CHECKPOINT APPROACH

SONIKA CHOREY

Research Scholar, G. H. Raisoni University & Assistant Professor, P.R.M.I.T. &R. Badnera-Amravati. (Correspondent Author)

NEERAJ SAHU

Assistant Professor G. H. Raisoni University, Amravati.

Abstract

Bank transaction Fraud is a deliberate act of deceit involving financial transactions to obtain a personal advantage. As the quantity of online transactions has risen, so has the number of scams. Detecting fraud is critical in the banking business to safeguard clients' funds, minimize fraud losses, and maintain profitability. Banks are using machine learning-based models because traditional fraud detection approaches are no longer adequate for identifying fraud. Skewness is a significant issue with financial transaction data, and any model's performance is data-dependent and technique- dependent. Using multiple parameters, this article compared several machine learning models Distributed Test Checkpointing approach. The research examined mobile money transactions reported on Kaggle during the previous six months. Python was used to develop the machine learning Distributed Test Checkpointing approach, while Sk learn and pandas were used to analyze the data. Analyses after the random forest, SVM with proposed novel checkpoint-based approach perform better than the other models.

Keywords: Checkpoint, Ensemble Learning, Fraud Detection, Machine Learning, Neural Network.

1. INTRODUCTION

Over the previous several years, there has been noteworthy growth in online money transactions as the younger population continues to shun physical presence in banks to avoid the bother and lengthy queues. This will expand further due to COVID19, as consumers get used to online purchases. As usual, fraud will grow as the volume of online transactions increases. According to (Grand View Research, 2019), the worldwide fraud recognition and anticipation market was cherished at about USD 17.33 billion in 2018 and was predicted to increase at a composite yearly growing quantity of 18.9 percent from 2019 to 2025 of business environment disruption caused by technology. According to a McAfee estimate study (Steve, 2018), cybercrime indemnities the world budget by $600 billion, or 0.8 percent of worldwide GDP. Digital convergence is predicted to provide new market possibilities and obviate the need for some current corporate procedures over time. With the growth of technology such as the Internet of Things, robotics, artificial intelligence, amplified reality, cloud computing, mobile banking and electronic commerce, computer-generated podia for the new-age customer have been upgraded. However, cybercrime, financial misconduct, information breaches, and individuality theft are posing a threat to the success of several enterprises. Digital fraud is one area of worry which has become a significant problem for firms in the finance and economics, health care, and e-business sectors.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

Banks and fiscal organizations previously detected fraud using rule-based procedures. Complex procedures were developed to identify and prevent scams that have previously happened. However, since fraudsters have switched their focus away from traditional tactics and toward technology, a rule- based approach will be unable to identify new frauds. Fraudsters are consistently exploiting gaps in transaction apps, and they accomplish their objective via the use of technology. That is why financial institutions are shifting their focus toward machine learning (ML) and artificial intelligence (AI) to combat scams.

In accord with (Ali, Salleh, Saedudin, Hussain, & Mushtaq, 2019), the primary obstacle to using machine learning in financial transactions is the unbalanced or skewed dataset. Machine learning models learn on data and form patterns; if there is insufficient data, the models will be unable to detect frauds effectively.

The primary goal of this research is to detect as many fraudulent transactions as possible while minimizing false positives (Choi & Lee, 2018) using checkpoint approach. False-positive refers to a legitimate transaction that the model predicts to be a fraud transaction, and false positives may result in client unhappiness and eventual customer loss.

To determine the best-fit algorithm, comparison research is conducted on several machine learning models and neural networks. Although the model accuracy may vary depending on the dataset, with this method, several models can be compared to see which ones perform better on unbalanced data sets and how to pick the best model after comparing them.

The objective of the study

- To design of novel distributed test Checkpointing approach.
- To eliminate real-time fraud by using proposed Distributed Test Checkpoint
- To provide additional security in online transaction for customers to strengthen their trust in the financial system.

2. LITERATURE REVIEW

To detect a scam, various machine learning and deep learning representations are utilized. For financial transactions datasets, both supervised and unsupervised algorithms have been applied. (Sadgali, Sael, and Benabbou, 2019) Conducted research examining the benefits and disadvantages of numerous machine learning approaches. It looked at Probabilistic Neural Network, Support Vector Machine, logistic regression, arbitrary forest, and decision tree in order to come up with a theory using a variety of representations. Another study (Bagga, Goyal, Gupta, & Goyal, 2020) employed ADASYN (Adaptive Artificial Sampling Technique for Unnecessary Information) to balance the unbalanced data and improve the correctness of the representation. We propose a classifier for pipelining and bagging. Finally, pipelining is recommended as the most operative means of noticing fraud.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

Another research (Olowookere & Adewale, 2020) developed a cost-sensitive and ensemble learning paradigms system that supports enhancing fraud recognition in unstable datasets. Cost-sensitive ensembles are constructed in this article using decision trees, MLPs, and KNN algorithms. (Raghavan & Gayar 2019) conducted a similar study in which they compared and contrasted SVMs, KNN, K-Means, Arbitrary Forest, Naive Bayes, and auto-encoders. to detect fraud. The authors determined that SVM is the superlative algorithm for big datasets and may match CNN to get the best results, whereas arbitrary forest and KNN are improved for small information sets. Though, this work focuses only on supervised learning for fraud detection.

A detailed study by Amarasinghe, Aponso, & Krishnarajah, 2018 used managed machine learning (Bayesian system, RNN, SVM, and fuzzy logic) and unverified machine learning (point outliers, K means cluster, and secreted Markov prototypical) to perceive scam and competently quantified the advantages and drawbacks of each prototypical. Additionally, it's been discovered that ANNs outperform fuzzy logic by 33 percent. Additionally, this research advocated the usage of ANNs in conjunction with genetic algorithms and compared them to other algorithms. (Shirgave, Awati, More, & Patil, 2019) analysed and evaluated numerous supervised learning algorithms' specificity, sensitivity, and accuracy (logistic regression, KNN, RF, SVM, decision tree, and Naive Bayes) for detecting scams based on a marking criterion. They recommended that models be qualified using feedback and a delayed managed sample, with each probability being aggregated to identify scams. In another study (Kurien & Chikkamannur, 2019), the writers investigated logistic regression and arbitrary forest and recommended using neuronal networks to get the best results. They also assessed the geographies and subsample proportions for unbalanced information sets. (Maniraj, Saini, Sarkar, and Ahmed, 2019) Used anomaly detection to identify fraud in this research. While the methods are accurate at 99.6%, their accuracy is only 33% when the complete dataset is used to train the prototypical. Additionally, it is mentioned that the high precision is the unbalanced dataset, and since the model has low precision, it will be unable to identify any fraudulent transactions.

(Blagus & Lusa, 2013) offered a methodology for resolving unbalanced datasets via the SMOTE function. SMOTE oversamples the marginal class information by using bootstrapping and KNN to produce more artificial explanations of the marginal class, which is scam information due to its scarcity. As is typical, logistic regression, SVM, and random forest are employed to identify financial fraud. We employed seven managed and unverified algorithms to identify fraud in this study, with those not often used in fraud recognition. The maximum critical point to grasp is that precision should not be the only criterion for prototypical assessment when applying machine learning or deep learning models to an extremely skewed information set (Wu & Radewagen, 2017). Because the information set is skewed, high accuracy is expected for the utmost representations. Additionally, the false positive rate (FPR) and false- negative rate (FNR) are significant in this situation (Choi & Lee, 2018). Thus, our main purpose is to lower this scenario's false negative and positive ratios.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

3. RESEARCH METHODOLOGY

This article used machine learning and deep learning representations to identify scams. This article will mainly discuss managed and unverified machine learning. Again, managed machine learning methods may be classified as classification or collective approaches, whereas unsupervised methods include clustering, anomaly detection, and dimensionality reduction. Each domain is represented by a single model chosen for analysis based on past research findings. As shown in Figure 1, logistic regression is used for classification in supervised learning, XGBoost and random forest are used for ensemble learning, DBscan is used for clustering, isolation forest is used for anomaly detection, and PCA is used for dimensionality reduction. Artificial neural network implementation is based on deep learning. Luque, Carrasco, Martn, & de Las Heras-Garcia de Vinuesa, 2019). The primary limitation of this study is the imbalanced information set (Luque, Carrasco, Martn, & de Las Heras-Garcia de Vinuesa, 2019).

3.1 Dataset:

This information set was derived from the Paysim artificial dataset of mobile currency transactions, which was recently uploaded on Kaggle. Nearly 6 lakh records are included in the collection, yet just 1.21 percent of records are fraudulent. The reply variable, or dependent variable, is called 'fraud,' It has a value of 1 if a transaction is fraudulent and 0 if it is not. SMOTE (Blagus & Lusa, 2013) is used to balance the skewed dataset by up-sampling the minority data, i.e., fraud data.

3.2 Methods:

A thorough examination of supervised and unverified learning, deep, and dimensionality lessening models is conducted to determine which representations perform the best in fraud recognition and have a high kappa value. The prototypical to be chosen is exceptionally dependent on the data assembly. However, we attempted to identify as many relevant models as possible during preliminary research. They could be likened to determining the best prototypical to apply to the actual information set provided by the corporation, which will have the fewest false positives and negatives.

3.2.1 Logistic regression:

When the predicted output of a model is binary data, such as zero or one, logistic regression is used (Wright, 1995). If the output value is 1, the transaction is fraudulent; if the value is 0, the transaction is legitimate. The logistic regression yield is the possibilities associated with the dependent variable group. As long as the possibility stays between 0 and 1, the range remains the same. The model's threshold is set at 0.5 during implementation. Thus, every transaction with a probability more than 0.5 is regarded authentic, whereas transactions with a probability less than 0.5 are considered fraudulent.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

3.2.2 Random Forest:

Random Forest (Breiman, 2001) is a managed machine learning technique (Fawagreh, Gaber, and Elyan, 2014) mainly used to solve classification issues. Random Forest is a classification algorithm that works in ensembles. It optimises the prediction outcome by combining many trees. Each tree verifies against a unique set of conditions. Each tree draws information from the dataset at random and calculates the likelihood of a fraudulent or legitimate transaction. The trees' majority vote will determine the outcome.

Advantages:

- By guessing missing data, the Random Forest model generates exact predictions.
- Scaling of features is not necessary.
- Overfitting may be prevented in random forest models with many data points.

Drawbacks:

- Complication while computing
- big reminiscence extent is desired
- Extended training period compared to additional machine learning representations

3.2.3 XGBoost:

XGBoost is a managed approach, more precisely, an ensemble method (Chen & Guestrin, 2016). The goal of XGBoost is to reduce the loss function to its smallest possible value. XGBoost is a gradient-boosting algorithm. The XGBoost boosting algorithms do not divide the tree leaf by leaf; instead, they split the tree depth by depth. As a result, the model may sometimes overfit the data, which may be prevented by adjusting the max depth limit appropriately. When a sole machine learning prototypical is insufficient to precisely forecast the outcome, XGBoost is used to combine several approaches.

Advantages:

- With the correct model adjustment, it provides superior accuracy.
- Capable of dealing with missing data

Drawbacks:

- There are no significant drawbacks discovered throughout the model's development, except that it takes longer to fail.

3.3 Dataset:

There is a dearth of publicly accessible statistics on financial services, particularly in the rapidly growing realm of mobile money transactions. Financial datasets are critical to a

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

wide variety of academics, especially those of us doing research in fraud detection. A factor contributing to the issue is the fundamentally personal character of financial transactions, which results in the absence of publicly accessible statistics.

As an example of how to address such a challenge, we give a synthetic dataset built using the PaySim simulator. PaySim produces an artificial dataset using aggregated information from the remote information set that matches the regular working of contacts and injects harmful activity to test scam recognition procedures' efficacy.

Content

PaySim mimics moveable currency contacts using data gathered from a month's value of monetary logs from a mobile money provider in an African nation. The initial archives were given by a multinational concern that functions as a mobile economics service in over 14 countries worldwide.

This synthetic dataset is a quarter of the original size and was developed just for Kaggle.

4. PROPOSED WORK

This article analyses seven distinct models and selects three of the highest performing representations for further investigation based on their precision, sensitivity, specificity, accuracy, MCC, BCR, and Kappa with novel check point approach. Finally, we evaluated the three models to determine which one had the most excellent match.

In order to improve the model's accuracy, the arbitrary forest method and the varImpPlot function (Archer & Kimes, 2008) are used to pick geographies based on Mean Decrease Precision and Mean Decrease Gini.

Because the information is skewed and contains roughly 1% scam information, if the prototypical correctly forecast all legitimate transactions, it will also achieve 99 percent precision. To circumvent this situation, (Blagas & Lusa, 2013) introduced the 'SMOTE' function. SMOTE is used to increase the model size of marginal data to equilibrium the amount of fraud and non-fraud transactions. The plan of novel checkpoint-based approach is for actual fraudulent activities would save banking a lot of money.

- It would also protect consumers from economic damage.
- It will increase public faith in bankers as a haven for their money.
- It would improve the quality of life in the community by reducing revenue damage.
- It would deter scammers from pursuing their deception.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

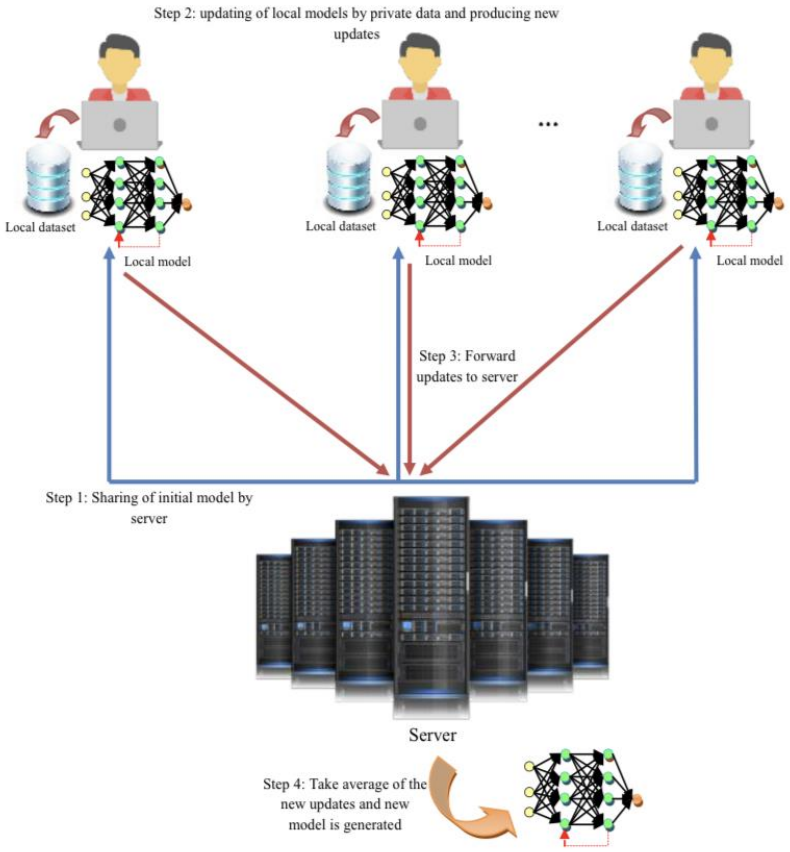

Figure 01: A framework of checking point approach for the securing the bank transaction.

4.1 Comparative analysis of the models

The dataset is split into two subsets in this study first is training and second is test. 75% of the data in the training set is missing. The model is built based on training data using checkpoint-based approach. Tenfold cross-validation (Koul, Becchio, and Cavallo, 2018) is utilized to cross-validate and checkpoint for overfitting. The dataset is divided into ten equal-sized folds using this technique. Nine of these ten folds are utilized for training the model, while the last one is used for testing. This method is done ten times, with each test using a different fold. The models are first compared without employing SMOTE, and then SMOTE is used and modified to optimize the models that provide favorable results (Probst, Wright & Boulesteix, 2018). Seven representations have been thoroughly analyzed, and the three top performers have been chosen for additional tweaking to find the most significant potential yield. While the goal should be to minimize false negatives, extra attention should be given to minimize false positives. A model with the highest prediction power is predicted for the overall balance.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

4.2 Result Analysis

Hence, we move onto create new features using Distributed Test Checkpointing approach by changing the original features. here we fabricate three functions which creates a highly applicable feature for the domain

1. **Difference in balance**: It is a fact that the amount debited from senders account gets credited into the receivers account without any difference in cents with proper check point approach. But what if there is a difference in case of the amount debited and credited. Some could be due to the charges demand by the service providers, yet we need to flag such unusual example in the distributed environment.
2. **Flow indicator**: Also, we have to trigger flag using checkpoint when huge amount is mixed up in the transaction. From the distribution of amount, we know that we have a lot of anomalies with high amount in transactions. Hence, we consider the 75th percentile(450k) as our threshold and amount which is larger than 450k will be triggered a flag.
3. **Rate indicator**: The user, not the transaction, is flagged here as a frequent user. When a receiver receives money from a large number of people, it can act as a trigger for illegal games of chance or luck. As a result, when a recipient receives money more than 20 times, it is flagged and checked.
4. **Merchant indicator**: The customer ids in receiver begin with the letter 'M,' indicating that they are merchants with a high volume of receiving transactions. As a result, anytime there is a merchant receiver, we additionally flag with checkpoint.

4.2.1 Data Pre-processing

Before building a Distributed Test Checkpointing approach with machine learning, it is necessary to pre-process the data so that the model can train without errors and can learn more to provide better results

1. Keeping the aim in check

The goal label is plainly imbalanced in the pie chart below, as we only have 0.2 percent fraudulent data, which is insufficient for the computer to learn and identify when fraud transactions occur.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

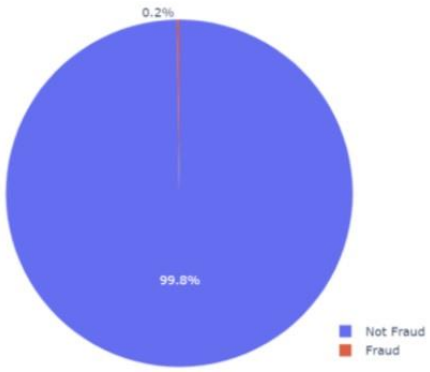

Figure 2: Fraudulent and Non fraudulent data chart

4.2.2 Split and Standardize

We build the independent and dependent features in this module, then split them into train and test data with a 70% training size. After that, we gather all of the numerical data and use the StandardScaler() method to modify the distribution so that the mean is 0 and the standard deviation is 1.

Model Building: We've successfully processed the data, and now it's time to use the Distributed Test Checkpointing technique to serve the data to the model. It takes a long time to figure out which model is optimal for our data. As a result, the checkpoint strategy to run our data through all of the classification algorithms and chose the optimal one that provides the highest level of accuracy.

Accuracy of the Logistic Regression Test: 0.96

SVC Test Accuracy: 0.97

Accuracy of Decision Accuracy: 0.97

KNN Test Accuracy: 0.98

Accuracy of the Naive Bayes Test: 0.98

Accuracy of Random Forest Test: 0.97

The proposed approach: 0.99

4.3 Evaluation of the Model

It's time to investigate the reality behind large numbers by comparing them to test data.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

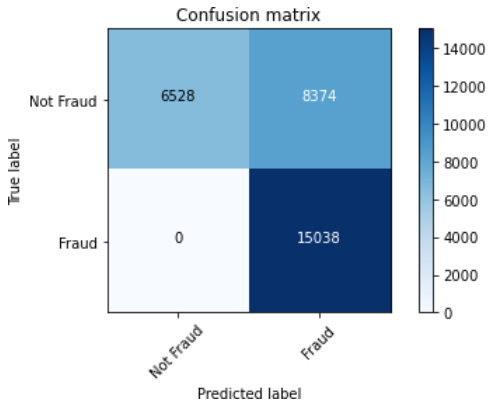

Figure 03: Confusion matrix (non-normalized) of fraud and non-fraud transaction

5. CONCLUSION

Various supervised and unsupervised models for detecting fraud are examined in this article using distributed checkpoint approach. Random forest has the lowest false-negative rate (9.33 percent), but a fairly high false-positive rate (49.33 percent). Both of these elements should be kept to a minimum with checkpoint approach & secure the data. As seen in Figures 03, the optimal model cannot be determined; hence balanced matrices: BCR is used to compare performance. The value of BCR is between -1 and +1. A checkpoint model with a value of +1 is considered ideal, whereas one with a value of -1 is considered inferior or fraud. As seen from the BCR comparison graph, all three models have BCR values close to 1, but the arbitrary forest has the maximum BCR value of the three. The issue is whether we should naively employ arbitrary Distributed Test Checkpointing approach for banking transaction fraud detection. It is entirely dependent on the order in which things are accomplished. Random forest is the most excellent performer, but it also has a high FPR. Therefore, a random forest can be used if identifying fraud is a high priority, and a single actual transaction might be compromised. If the ultimate objective is to safeguard actual transactions, even a slight amount of fraud may jeopardize them, Distributed Test Checkpointing approach should be used. As with any group, trailing even a single client may be prohibitively expensive. This article attempted to conceal all aspects of machine learning using Distributed Test Checkpointing approach, including managed and unverified learning, ensemble learning, and neuronal networks. The neural system is built, and the result is relatively decent compared to the other techniques studied here. ANNs may be used with Checkpointing algorithms for better accuracy and a low false- negative ratio in future research.

References

1. Imane S, Nawal S, Faouzia B., Performance of machine learning techniques in the detection of financial frauds 2019;
2. Bagga S., Goyal A., Gupta N., Goyal A. Credit Card Fraud Detection using Pipeling and Ensemble Learning. Procedia Computer Science 2020; 173(2): 104-112.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication Online Open Access
Vol: 41 Issue: 12-2022
DOI 10.17605/OSF.IO/65YSV

3. Toluwase O., Olumide A. A framework for detecting credit card fraud with cost- sensitive meta-learning ensemble approach. Scientific African. 2020.;
4. Pradheepan R, Neamat G, Fraud Detection using Machine Learning and Deep Learning.2019; 334-339.
5. Thushara A., Achala A, Naomi K, 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.2018; 12- 17
6. Shirgave S., Awati, C., More, R., Patil, S. A Review On Credit Card Fraud Detection Using Machine Learning. International Journal of Scientific & Technology Research. 2019; 8(1): 1217- 1220.
7. Kaithekuzhical L.K., Chikkamannur A., Detection And Prediction Of Credit Card Fraud Transactions Using Machine Learning, International Journal Of Engineering Sciences & Research Technology,2019 ;2277-9655
8. Maniraj, S, Saini A., Shadab A., Sarkar S. Credit Card Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research 2019;
9. Blagus, R., Lusa, L. SMOTE for High-Dimensional Class- Imbalanced Data. BMC bioinformatics. 2013; 1471-1479
10. Philipp P. & Wright B., Marvin W., Boulesteix S. L. Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdisciplinary Reviews: Data Mining andKnowledge Discovery.2018
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. Boughorbel S, Jarray F, El-Anbari M Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. 2017; 12(6):
13. Luque, A., Carrasco A., Martín, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019;
14. Haseeb A., Salleh, M., Saedudin, R. D., Kashif H.& Mushtaq, M. Imbalance class problems in data mining: A review. Indonesian Journal of Electrical Engineering and Computer Science. 2019; 1552-1563.
15. Grobman, S. Impact of Cybercrime Why Cyber Espionage isn't Just the Military's Problem,2019;
16. Sasikala, B & Biju, V. & Prashanth, C. Kappa and accuracy evaluations of machine learning classifiers. 2017;
17. Koul, A., Becchio, C. Cavallo A. Cross-Validation Approaches for Replicability in Psychology. Frontiers in Psychology. 2018; 9 (2) 111-124.
18. Wu, Y. & Radewagen R., 7 Techniques to Handle Imbalanced Data.2017;
19. Choi, D.Lee K., an Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation Security and Communication Networks. 2018; 1-15.