International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 4, July-Aug-2022, ISSN (Online): 2395-566X

An Optimized Machine LearningAlgorithms for Solving Class Imbalance Problem in Credit Card Fraud Detection

Md Shufyan, Dr. Prashant Prashun
Dept. of Computer Science and Technology
Central University of Jharkhand,
Jharkhand, Ranchi, 835205, India
shufyanbihar96@gmail.com

Abstract- Class imbalance problem is more common with machine learning algorithm, it occurs when the ratio of data into different classes is not equal, in its data is divided into two classes, one is of majority classes and other is of minority classes. The sample present in the majority classes is too high as compared to the minority class, for very few numbers of samples is present in the minority classes.The algorithm is unable to read the data from minority classes. Thiscauses poor performance and often may cause overfitting when model get trained from skewed dataset. In this research work, to balance the dataset we applied SMOTE or Synthetic Minority Oversampling Technique in order to balance the dataset. Before balancing the dataset, it has to undergo through preprocessing phase in which we applied missing value removal and outlierdetection to reduce the dataset. When the dataset gets reduced, we applied the different algorithm like logistic regression, decision tree and extreme learning machine to detect the fraud, but it causes overfitting due to imbalance of dataset. SMOTE has been applied to balance the dataset and then ML algorithm has been applied and it has been noticed ELM is more feasible and effective as compared to remaining algorithm.

Keywords- Imbalanced data; Missing value removal; Outlier detection;SMOTE; Logistic Regression; Decision Tree; Extreme Learning Machine)

I. INTRODUCTION

Rapidly increasing data generation and data collection capability in all fields is the result of fast developments in information technology. In the business world, retailers and e-commerce have built very wide databases on commercial transactions. In various fields, a large number of scientific datasets were generated.Fraud means unethical act which is now increasing from across the world[1].It is classified both online and offline. In offline fraud information of user is stolenwhereas online fraudis performed using the Internet, phone, shopping, or without the holder of the card [2].It is assumed that the dataset used in the implementation of model is balanced. But, when it come in use, it is always not true that the dataset is balanced [3]. This issue occurs in most of the cases like network intrusion detection, medical diagnostics, fraud detection, and many more. Since the dataset contain classes, through which data is classified. In imbalance dataset there exists two classes i.e., majority classes and minority classes [4].

Class imbalance has been mostly known for complicating factor because in the distribution ofdata in classes is not uniform. Since, in most of the framework the ratio of imbalance data is upto1 in 100, 1 in 1000, 1 in 10,000 and often more [5]. As a result, the model gets biased toward majority classes.

The class imbalance problem, causes not only biasing problem and performance degradation issue butanother major and complicating issues is overfitting of model. In overfitting, the model gives 100% accuracy rate even when the model is unable to read the minority samples because the model gets biased towards the majority classes and result is predicted based on it.

In this paper, we are discussing about the skewness of data in the dataset. This is one of the suitable approaches of class imbalance problem as well as overfitting issue [6]. Here we tried to balance the dataset using SMOTE so that the number of data get equivalent in both classes i.e., minority and majority classes. Before balancing dataset, the model has to undergo through preprocessing phase in which we applied missing value removal and outlier detection to reduce the dataset. When the dataset gets reduced, we tried to trained the model with the suitable algorithms, but it causes overfitting, due to which we have tobalance the dataset using balancing technique called as SMOTE and after it we applied the logistic regression, decision tree and extreme learning machine, inwhich we observed ELMpredict better result and helpful

inresolvingthe issue of overfitting as well as fraud detection[7].

This paper is described as follows: In section II, we have discussed about the related work which has been done to resolve the issue of imbalance class problem. Section III,various method has been applied and discussed which are implemented for fraud detection.While section IV deals about the result and discussion with comparison.And the last phase reveals about conclusion and future work.

II.RELATED WORK

There are various methods which have been initiated in order to resolve the skewness of dataset to improve its accuracy [8]. Mostly,the described method depends either on algorithmic level or the data level approaches. A sampling method are used in the case of data level approaches in order to generate the sample either in the case of majority or minority classes. Sampling technique can further be partitioned into two parts i.e., undersampling and oversampling [9], [10], [11], [12]. There are number of classification algorithm which areimplementedto increase the number of samples in case of minority classes[13], [14], [15].

Here we explain the literature survey which has been performed in term of "Credit Card Fraud Detection" in which we look how to handle and get rid from skewness of data. The main reason behind is that the number of normal data present in the dataset is much higher than thatof fraudulent data, due to which an individual is unable to detect the fraudulent data easily. This limits the amount of data that can be analyzed and extracted from fraudulent transactions. Fraudulent activities cause huge losses, motivating researchers to find a way to recognize and prevent fraud. There are various methods and techniques through which an imbalance data and overfitting issue get resolved. According to [16], ANN have been implemented to achieve the accuracy and higher efficiency of the model.

In [17], the author uses SVM which increases the number of samples in training cases and make equivalency in both cases in order to reduce loss and to boost up the performance. In [18], outlier has been applied which generates cluster as a by products and perform efficiently with huge dataset. Sampling method has also been used in [19] through which dataset get balanced and helpful in reducing overfitting.In [20], Edited Nearest Neighbor Rule (ENN) was proposed that omits samples having different class samples in which the class of minimum two or three of its nearest neighbors. In[21], Caps Net has been used to improve the viability for fraud recognition and help to accomplish the best outcomes in precision. In [22], an under-sampling method of Tomek's link has been used which plays a major role in removing the impurities and noise from the majority data. And after that algorithm are applied to obtain the better result.

III. PROPOSED METHODOLOGY

The motive behind this research work is to balance the dataset with better accuracy and to detect the fraudulent data. We applied different algorithm in order to gain better accuracy and efficiency. In the following we briefly describe the phases which has been used in the implementation.

1.Dataset Description- the dataset is gathered from Kaggle,which contain information of European cardholders. the dataset is of September2013 which contain the transactions of2 days in which 492 data are only fraudulent and remaining dataare non-fraudulent (284,415). It shows that the dataset is extremely skewed (as shown in figure 1). The percentage of fraudulent data is only 0.172 which is very few as compared to non-fraudulent data. The dataset is in numerical form as it contains personal information of a cardholder like account number, card number, account holder name etc., since the dataset contain total 30 feature i.e., V1, V2, V3,…,V28, in which 28 are PCA converted which is called as dependent variable and remaining two are independent variable i.e., Time and Amount. Each data is defined with proper class, i.e., class 0 and class 1.Class 0 represent non-fraudulent whereas class 1 represent fraudulent.

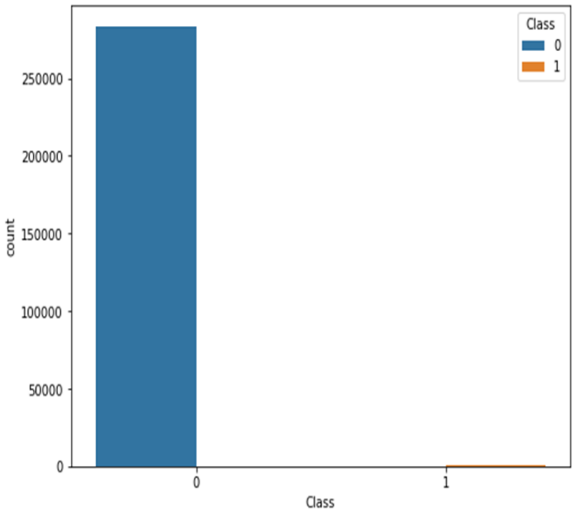

Figure 1. Imbalance Dataset.

2.Data Preprocessing- It is the one of the most important and effective phases which contain missing value removal, outlier detection and other impurities in order to improve the quality and efficacy of dataset. To get better prediction and accuracy it is essential to go through this phase so that

machine learning model can be applied smoothly. For credit card dataset we have to perform preprocessing in two steps on 28 features from 30 features.

3.Missing Values Removal-In this step we tried to remove all the zero (0) which is present in each row. Since, the model will not evaluate or predict properly if there would be presence of huge number of zero. So, there is no any presence of zero in credit card dataset.

4.Outliers- The datapoint which is dissimilar from the remaining datapoints, in such situation it needed to be removed, since the classifiers are much sensitive to the data range and distribution of attributes. Mathematically, outlier detection can be measured as:

$$P(X) = X, if\ Q1 - 1.5 \times IQR \leq X \leq Q3$$
$$reject, otherwise$$

Here, X is the instance of the features vector which lies n-dimensional space X€R.
Q1= first quartile,
Q3 = third quartile and
IQR = Interquartile range of attribute,
whereQ1, Q2,Q3, IQR€R.
Lower bound=Q1-1.5×IQR
Upper bound=Q3+1.5×IQR

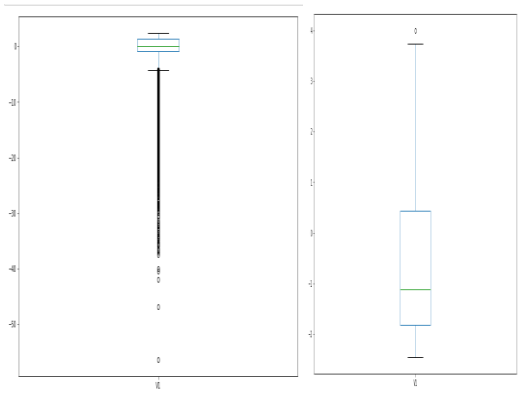

Figure 2. Boxplot of before and after outlier detection
After applying outlier, the reduced data obtained is 156350. In which only 15 data is fraudulent.

4.Data splitting-When the data is cleaned, it is split into two parts i.e., testing and training phase. After division of data, the model is being trained though training dataset and model is tested using test data to get the result.

5.Smote Algorithm- The SMOTE is one of the best balancing techniques which was developed in the year of 2002 by Nitesh Chawla in order to resolve the imbalance dataset. In it synthetic data is generated in the case of minority samples, so that data get equivalent to that of majority class samples.The steps of SMOTE is defined below:

1. For each sample in the case of minority class x_i,where i=1,2,….p, p is the samples present in the minority class, obtain the KNN called as K-Nearest Neighbor in the case of minority samples by finding the distance from all samples in the majority samples.
2. Select n from the *k* NN samples arbitrary. Where, n stand for oversampling magnification from the original data. Let's consider that the nearest neighbor data which is selected is x_j,in which j=1,2,3,….n. Thesamples in minority classes can be generated by the given formula

$$\boldsymbol{X_{new} = x_i + rand(0,1) \times (x_j - x_i)}(1)$$

Where $\boldsymbol{X_{new}}$ is the obtained synthetic data in the case of minority data.
3.In original dataset, synthesized data is added and repeat step 3 unless data get balanced. Since, SMOTE algorithm plays a vital role in balance the dataset so that overfitting can be avoided and result could be predicted better.

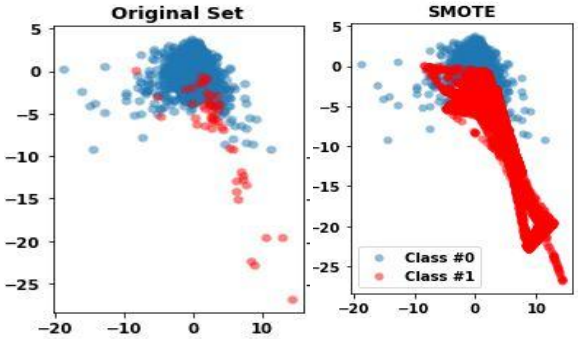

Figure 3: Scatter plot of data before and after SMOTE

6.Apply Machine Learning-Whenpreprocessing phase get completed then an individual is able to apply machine learning algorithm. We apply different algorithms to detect the fraudulent and non-fraudulent data. The motive behind to apply the different machine learning algorithm to achieve better accuracy and also to detect the fraudulent data. The algorithm which has been used is described below:

7.Logistic Regression- In classification problem Logistic regression is widely and known method of ML algorithm. The likelihood of each class's occurrence is stated as the forecast. The union of weights and input parameter are the real valued outputs of linear regression. In it the data is divide into two classes i.e.in the 0 and 1 form and helpful for detecting fraudulent and non-fraudulent data. Let us consider the circumstances which contain the combination of dependent and independent variable. The equation should be expressed in the given below:

$$Y = a_0 + a_1 + x$$

Here $a_0$could be the weight for single input variable x and $a_1$is the bias term. Though, because it will always approximate the probability that depend on every class, since the sigmoid function is used in the case of logistic regression through which outcome can be predicted as shown below in equation 4.2 it condensed the expected outcome in between 1 and 0.

$\text{Sigm}(z) = \frac{1}{1+e^{-2}}$ (3)

The equation of logistic regression is expressed in terms of sigmoid function which is shown below in equation 4.3. In the given equation there are two variable which is represented in the form x and y. Since, the logistic regression uses threshold value i.e., 0.5, if the outcome obtained is below 0.5 then it would lie in class 0 and if the outcome obtained is above the threshold value then it would lie in class 1.

$P(y = 1) = \text{sigm}(a_0 + a_1 \times X)$ (4)

In the training phase the parameter a_0 and a_1 get trained, this parameter belongs to logistic regression. As a result, the expected output can be represented with 0.5 threshold value,

$$y = 1 \text{ if } P(y = 1) \geq 0.5$$
$$y = 0 \text{ if } P(y = 1) < 0.5$$

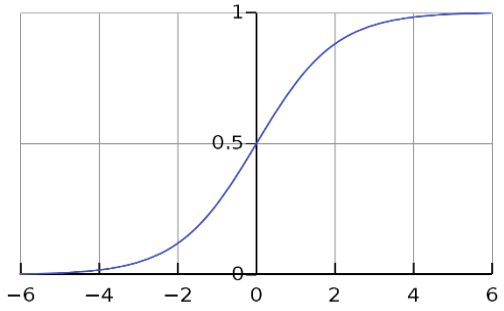

Figure 4:Graph of Sigmoid function.

8. Decision Tree-The supervised learning method which is used for regression as well as classification methods in data mining is called as decision tree. The main motive behind decision tree to resolve the issue of imbalance problem in order to predict the outcomes. In it a tree like structure gets created of the models, each data gets divided into smaller subsets so that the tree develops steadily. The main motive is to solve both regression and classification problem. The splitting of data can be decided in decision tree by entropy which is given as:

$$E = -\sum_{i=1}^{n} p_i . log_2 . p_i$$

Where, p_i is the probability of random selection.

The **algorithm** of decision tree is given below:

- Construct a tree with nodes as input feature.
- Select the attribute to predict the output from input attribute whose information gain is maximum.
- In each node of tree highest information gain is calculated.
- Repeat step 2 so that subtree can be formed with the given attribute.

9. Extreme Learning Machine- ELM is a smooth and simple algorithm that is designed to train the SLFNs known as single hidden layer feedforward neural network. In this model, the weights are arbitrarily assigned in between input layer and hidden layer, although the weights between hidden layer and output layer are analyticallycalculated.The training set can be given as:

$$D = \{(x_i, y_i) | x_i \in R^d, y_i \in R^k\}, 1 \leq i \leq n,$$

By the following equation (5), the SLFN with structure (d,m,k) can be modeled,

$$f(x_i) = \sum_{j=1}^{m} \beta_j . g(w_j . x_i + b_j) \quad (5)$$

Here, β_j means weight vector that connect thej^{th}hidden node through o/p nodes,g(.) considered as an activation function,w_j is the weight vector connecting$j^{th}$the hidden node with the i/pnodes, b_j is the bias of thej^{th} hidden node. In eq.(5) in itw_j andb_j are randomly generated, the β_j can be achieved by resolving the given linear equation (6)

$$\sum_{j=1}^{m} \beta_j . g(w_j . x_i + b_j) = y_i \quad (6)$$

The Eq. (6) may be given in matrix format as

$$H\beta = Y \quad (7)$$

In which,

$$H = \begin{matrix} g(w_1 . x_1 + b_1) & \ldots\ldots g(w_m . x_1 + b_m) \\ g(w_1 . x_n + b_n) & \ldots\ldots (w_m . x_n + b_m) \end{matrix}$$

$$\beta = [\beta_1^T, \beta_2^T, \ldots, \beta_m^T]^T$$
$$.Y = [y_1^T, y_2^T, \ldots, y_n^T]^T$$

Where, H stand for the o/p matrix of i/p layer of SLFN, which is a non-square matrix.The overall solution would beachieved by resolving and calculating the given optimization problem.

$$\min \| H\beta - Y \| \quad (8)$$

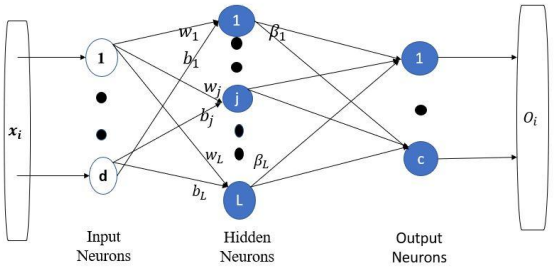

Figure 5: Architecture of Extreme Learning Machine.

10.Evaluation Metrics

When the model gets trained using trained data in machine learning, then it is necessary to evaluate its classification performance. To put it another way, we test the model with actual samples to examine the real accuracy of a models. That's why the question arises that how is the performance of the model is? We employ assessment data to measure the model's performance.

Recall tells that how much percentage are forecasted positive from total positive. Or we can also say, the ratio of TP to the sum of TP and FN.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

Precision refers to the proportion of detected cases that have been confirmed positives.

$$\text{Precision} = \frac{TP}{TP+FP} (10)$$

F1 Score is used to get test accuracy and is suitable for imbalnce data and also work effectively on such data. It may be defined using the following formula:

$$F1 = \frac{2\times(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} (11)$$

IV. RESULT AND ANALYSIS

1.For Unbalance dataset

Initially the model was trained and tested through logistic regression, on an imbalance dataset.And evaluation metrics is obtained which is shown below:

Table 1. Evaluation metrics of LR without SMOTE

	Precision	Recall	F1-Score	Support
Class-1.0	0.00	0.00	0.00	14
Class-0.0	1.00	1.00	1.00	124906
Accuracy			1.00	124920
Macro avg.	0.50	0.50	0.50	124920
Weighted avg.	1.00	1.00	1.00	124920

When model is being trained and tested through decision tree algorithm, on an imbalnce dataset. Then the evaluation metrics obtained is shown below:

Table 2. Evaluation metrics of Decision tree without SMOTE

	Precision	Recall	F1-Score	Support
Class- 1.0	0.00	0.00	0.00	14
Class- 0.0	1.00	1.00	1.00	124906
Accuracy			1.00	124920
Macro avg.	0.50	0.50	0.50	124920
Weighted avg.	1.00	1.00	1.00	124920

When model is being trained and tested through Extreme Learning Machine, on an imbalnce dataset. Then the evaluation metrics obtained from it is shown below:

	Precision	Recall	F1-Score	Support
Class- 1.0	0.00	0.00	0.00	14
Class- 0.0	1.00	1.00	1.00	124906
Accuracy			1.00	124920
Macro avg.	0.50	0.50	0.50	124920
Weighted avg.	1.00	1.00	1.00	124920

Table 3. Evaluation metric of ELM without SMOTE

Since, all three model i.e., Logistic Regression, Decision tree and Extreme Learning machine (ELM) gives accuracy of 100 percent whenapplied.From all above table, it is clearly visualized that the model is not able to read the data from both classes in order to predict the result of both classes. As it is seen, that model get biased toward the Class 0 and gives 100 percent accuracy which is termedas overfitting.

1.For Balance Dataset

The dataset is balanced using SMOTE and then logistic regression algorithm is applied on balanced dataset, the evaluation metrics achieved from it is shown in the table below:

Table4. Evaluation metrics of LR with SMOTE

	Precision	Recall	F1-Score	Support
Class- 1.0	0.89	0.96	0.92	124906
Class- 0.0	0.96	0.88	0.92	124906
Accuracy			0.92	249812
Macro avg.	0.92	0.92	0.92	249812
Weighted avg.	0.92	0.92	0.92	249812

It is observed from above table that algorithm logistic regression gives accuracy of 92%.When the decision tree is algorithm is applied on balanced dataset, then the evaluation metrics achieved is shown below:

Table5. Evaluation metrics of DT with SMOTE

	Precision	Recall	F1-Score	Support
Class- 1.0	0.98	1.00	0.97	124906
Class- 0.0	1.00	0.98	0.97	124906
Accuracy			0.97	249812
Macro avg.	0.99	0.99	0.99	249812
Weighted avg.	0.99	0.99	0.99	249812

When decision tree algorithm is applied on balanced dataset then the model gives accuracy of 97 percent on train data and 96.86 percent on test data, which means that decision tree performs well on balanced dataset and able to gives best result as compared to logistic regression.Similarly, When the Extreme Learning Machine algorithm is applied on balanced dataset, then the evaluation metrics achieved is shown below:

Table 6. Evaluation metrics of ELM with SMOTE

	Precision	Recall	F1-Score	Support
Class- 1.0	0.97	0.98	0.99	124906
Class- 0.0	0.98	0.97	0.99	124906
Accuracy			0.99	249812
Macro avg.	0.99	0.99	0.99	249812
Weighted avg.	0.99	0.99	0.99	249812

When all three model i.e., Logistic Regression, Decision tree and ELM (Extreme Learning Machine)are applied, it is observed that ELM is more effective and feasible, and gives better accuracy of 99 percent Hence, ELM gives better prediction and resultas compared to logistic regression and decision tree, which is shown in the comparison table.

Table 7. Comparison of result obtained from various algorithm on balance dataset

Algorithm	Accuracy
Logistic Regression	92%
Decision Tree	97%
Extreme Learning Machine	99%

V. CONCLUSION

The aim behind this project work is to resolve the issue of imbalance problem and to detect the fraudulent data. To get rid from imbalance issue, missing value removal and outlier detection is applied in preprocessing phase to reduce the dataset due to biasing and overfitting problem data is needed to be balanced. SMOTE is applied to balance the dataset.We can determine that extreme learning machine perform well on balance dataset and able to predictbetter accuracy rate of 0.99 percent on both train and test data,which means Extreme Learning Machineis more feasible and effective for resolving the imbalance technique and able to detect the fraudulent data withbetter accuracy and efficiencyas compared to Logistic regression and Decision tree.

REFERENCES

1. S P Maniraj, "Credit Card Fraud Detection using Machine Learning and Data Science",International Journal of Engineering Research & Technology (IJERT)Vol. 8 Issue 09, September-2019, PP.1-6.
2. Khyati Chaudhary, "A review of Fraud Detection Techniques: Credit Card'', International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012, pp.39-44.
3. Komal C Lagade et al. "Credit Card Fraud Detection System Using Data Mining'', Artificial & Computational Intelligence Published Online: June 2020
4. N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intelligent data analysis, vol. 6, no. 5, pp. 429–449, 2002.
5. N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," ACM SIGKDD explorations newsletter, pp. 1–6, 2004.
6. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," in In IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2015, pp. 159–166.
7. C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Folleco, "An empirical study of the classification performance of learners on imbalanced and noisy software quality data," Information Sciences, vol. 259, pp. 571–595, 2014.
8. Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," International journal of pattern recognition and artificial intelligence, vol. 23, no. 04, pp. 687–719, 2009.
9. W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," Information Sciences, vol. 409, pp. 17–26, 2017.
10. M. Rahman and D. N. Davis, "Cluster based under-sampling for unbalanced cardiovascular data," in Proceedings of the World Congress on Engineering, vol. 3, pp. 3–5, 2013.
11. J. Zhang, T. Wang, W. W. Ng, S. Zhang, and C. D. Nugent, "Undersampling near decision boundary for imbalance problems," in 2019 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1–8, IEEE, 2019.
12. G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote," Information Sciences, vol. 465, pp. 1–20, 2018.
13. Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognition, vol. 40, no. 12, pp. 3358–3378, 2007.
14. P. Cao, D. Zhao, and O. Zaiane, "An optimized cost-sensitive svm for imbalanced data learning," in Pacific-Asia conference on knowledge discovery and data mining, pp. 280–292, Springer, 2013.
15. C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," IEEE transactions on neural networks and learning systems, vol. 24, no. 6, pp. 888–899, 2013.
16. Devi D, Biswas SK, Purkayastha B (2019) A Cost-sensitive weighted random forest technique for credit card fraud detection. In: 2019 10th international

conference on computing, communication and networking technologies (ICCCNT). IEEE, pp 1–6
17. F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," Information Sciences, May 2019.
18. E. Kim et al., "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning," Expert Systems with Applications, vol. 128, pp. 214–224, Aug. 2019.
19. P. Hart, "The condensed nearest neighbor rule (Corresp.)," IEEE Transactions on Information Theory, vol. 14, no. 3, pp. 515–516, 1968
20. D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Transactions on Systems, Man, and Cybernetics, no. 3, pp. 408–421, 1972.
21. M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," 2019 Proc. 3rd Int. Conf. Comput. Commun. Technol. ICCCT 2019, pp. 149–153, 2019.
22. M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in Icml, vol. 97, pp. 179–186, Nashville, USA, 1997.
23. K. Randhawa, C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in IEEE Access, vol. 6, pp. 14277-14284, 2018
24. Navanshu Khare, Saad Yunus Sait: Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018.
25. O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, "Credit card fraud detection using machine learning as data mining technique," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-4, pp. 23–27, 2018.