A Novel Technique to Defraud Credit Card by Handling Class Imbalance Problem Using Machine Learning

Kaneez Zainab
Amity University, Lucknow Campus, Uttar Pradesh, India
kaneez_srm@yahoo.com

Namrata Dhanda
Amity University, Lucknow Campus, Uttar Pradesh, India

Syed Qamar Abbas
Ambalika Institute of Technology & Management, Uttar Pradesh, India

Article Info
Page Number: 82 – 91
Publication Issue:
Vol 71 No. 2 (2022)

Abstract
Typically, classification algorithms work poorly when confronted with unbalanced datasets, and the resulting effects are skewed against the majority class. As a result, an effective model is required to identify unbalanced data, particularly in the context of fraud detection. For these types of issues, the classifier's accuracy is not trusted because the cost of predicting a fraud sample as a non-fraud sample is extremely high. In general, imbalanced learning happens when some types of data distributions significantly outnumber other data distributions in the instance space. There is a need of technique such as under sampling or oversampling in order to learn from unbalanced datasets. A novel over sampling method has been suggested for learning from unbalanced datasets in this paper. The basic impression here is that a weighted distribution for diverse outnumbered class instances has been utilized depending on their degree of complexity to learn, with more pretended evidence leading to the outnumbered ones, being more troublesome to learn. As a result, with regard to data distributions, the suggested approach improves learning first by bringing down the bias familiarized using class difference, and then by pliantly conveying the classification judgement boundary toward challenging instances.

Article History
Article Received*: 10 December 2021*
Revised*: 19 January 2022*
Accepted*: 08 February 2022*
Publication*: 21 March 2022*

Keywords: Fraud Detection, Imbalanced dataset, SMOTE, Machine Learning, Boosting Algorithms, Over Sampling, XG Boost, Classification Algorithm.

INTRODUCTION

Now a day's customers have serious concerns about safety in online transactions. There are many different types of credit card frauds performed by hacker. For detecting and preventing the credit card fraud transaction various methods available in market. Some popular detecting techniques for credit card scams are HMM, Data mining, Biometrics, SVM, Bayesian Network, Neural Network, etc. But the problem with all these methods are, some methods are detecting but not preventing transaction at the same time (Taha, et al., 2020 ; Pandey, et al., 2017).

As rising cases of cyber frauds, the Reserve Bank is working to enhancing security. Especially, today a digital transaction is observing an important growth. The main objective behind electronic payment growth is it removes the limitations of traditional commerce. A user does not have to stand in long queue and personally visits the Bank to settle transaction. There are two ways to perform electronic payments that are online or offline. Online payment can identify as virtual payment. In online payment, it requires account holder name, PIN, card number, expiry date, etc.

sensitive information. Offline payment can identify as physical payment. For offline payment, presence of cardholder and PIN are required (Moreira, F.R., Nunes, et al., 2020).

Resampling techniques

Generally, all the available datasets of credit card transactions are imbalanced. One of the most challenging aspects of statistical modelling is resolving an imbalanced dataset with an Inherently Imbalanced class distribution. A machine learning algorithm's primary objective is to effectively identify these uncommon occurrences into their respective groups. When working with an imbalanced dataset, the primary impediment is when one class exceeds another, resulting in a paradigm that is severely under fitted due to its inability to successfully distinguish the minority class. Thus, the principle of oversampling is introduced, which resolves this issue (Ghorbani, R. and Ghousi, R., 2020.).

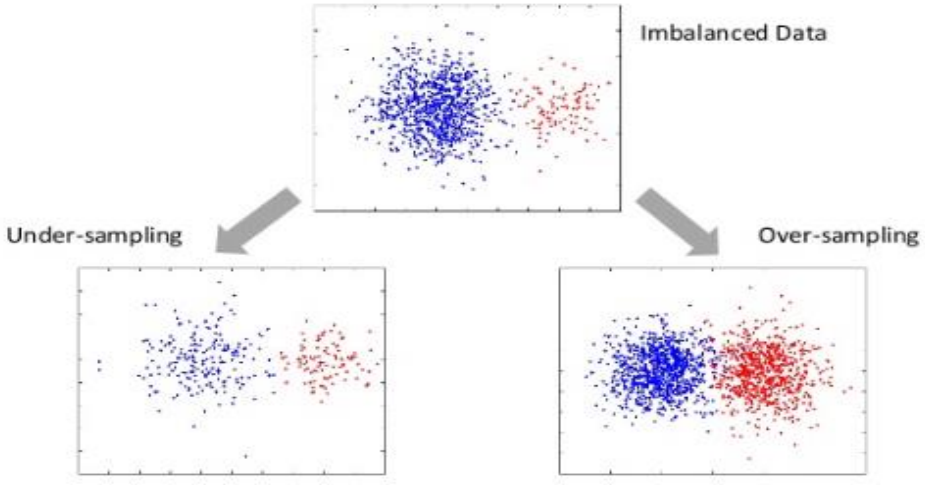
There are two types of resampling Techniques:

Under sampling aims to reduce the proportion of members of the dominant class in the training package. As a result, the training set's total number of records is significantly decreased. This means the preparation time is significantly shortened during classification. Due to the fact that we are working with very large databases, there is also a huge memory savings. However, since we are excluding members of the dominant class, it is likely that we would miss a significant amount of important knowledge if we exclude records that may aid our classifier in developing a reliable model (Mrozek, P., et al,. 2020 ; Guzmán-Ponce, et al,. 2021).

Oversampling aims to improve the representation of representatives of minority classes in the training package. The benefit of oversampling is that no material from the initial training collection is missing and all members of the minority and majority groups are retained. The downside, though, is that we significantly expand the scale of the training package. As a result, we lengthen the training period and increase the amount of memory available to store the training package. Due to the fact that we are working with very high-dimensional datasets, we must exercise caution in order to keep time and memory complexity within acceptable bounds. If the time needed to resample is ignored, under sampling outperforms oversampling in terms of time and memory complexity. Given this, oversampling must outperform under sampling in terms of classification efficiency in order to be viable. Previous research has not established conclusively whether under sampling or oversampling is better for classification results. Contradictory findings are most certainly the product of combining various datasets and classification algorithms. Further, the process of resampling is probably domain- and problem-specific (Mohammed, R., Rawashdeh, J. et.al 2020)

The difference of oversampling and under sampling is depicted in fig. 1.

Fig. 1: Under-Sampling Vs Over-Sampling



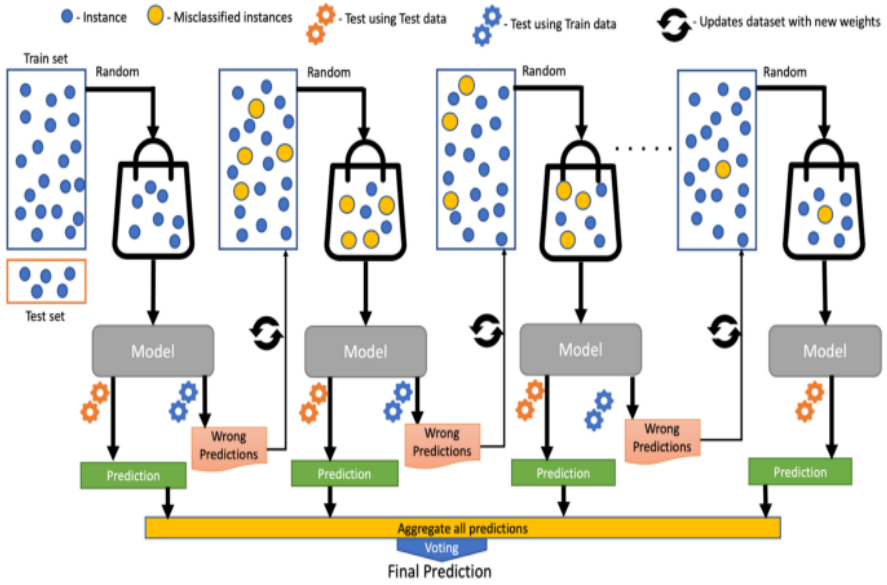
As in this paper, we are focused on oversampling technique, so the most popular oversampling technique is depicted below:

Synthetic Minority Oversampling Technique (SMOTE): SMOTE is a simple but efficient algorithm that outperforms random oversampling in a variety of low-dimensional problems. SMOTE is one of the few resampling techniques that consistently has superior performance from both random under and oversampling (Ishaq, A.,et,al. 2021; Raghuwanshi, B.S. and Shukla, S., 2020.)

Boosting Algorithms

The boosting algorithm generates new weak learners (models) and integrates their predictions sequentially to increase the model's overall efficiency. For each erroneous prediction, a higher weight is given to wrongly classified instances and a lesser weight to appropriately classified instances. Weak learner models with superior performance are weighted more heavily in the final ensemble model. Boosting never modifies the prior predictor and only corrects the subsequent predictor by error learning. (Atir, M. and Haydoutov, M., 2020 ; Bentéjac, C.; et al, .2021)

Fig. 2: Internal working of boosting algorithm



The remaining work is compiled in the subsequent manner. In literature review, we will discuss prior approaches to the issue of imbalanced datasets and further explain our solution. Proposed Methodology explain the resampling techniques that have been proposed. Experimental Evaluation describes our experimental design, as well as the datasets, classifiers, and measurement metrics that are in our analytical evaluation. Additionally, we explain the outcomes of our experiments in this section. Finally, we bring our conclusions to a close.

LITERATURE REVIEW

Minjae Son et al.(2021) suggested a scheme for oversampling based on “Conditional generative adversarial network” (CGAN) and borderline class. Further, they accurately identified a marginal class on account of minority-majority dataset. Then, they created data for the borderline class, contrasting it with the CGAN. They performed a series of tests to show the proposed scheme's success on a variety of imbalanced datasets.

TM Alam et al.(2020) hypothesised that models built using various ML practices are substantially similar or dissimilar, and that resampling methods significantly increase the result of the suggested representations. One-way ANOVA is considered as a hypothesis checking technique that results to determine the validity of the findings. To check the analysis, the split procedure was utilized, where the data was divided in segments named as training and test . On imbalanced datasets, the findings from credit card data indicate a precision of 66.9 percent for the clients of Taiwan, percentage of 70.7 for the South German customer’s credit dataset, and 65 percent for the Belgian client’s

credit card statistics. In comparison, their suggested methods greatly increase precision to 89 percent on the Taiwan customer's credit dataset, 84.6 percent on the South German customer's credit dataset, and 87.1 percent on the Belgian customer's credit dataset. The findings indicated that classifiers perform better on a balanced dataset than on an imbalanced dataset.

Based on the divide-and-conquer principle, Zhenchuan Li et al. (2021) suggested a novel hybrid approach for resolving the issue of class imbalance with overlap. To begin, an abnormality discovery model was trained on minority samples in order to exclude both a few minority class outliers and a large number of bulk specimen samples making use of the initial data. The residual specimen converged to form an alternating subset that results in lower variation of ratio. Afterwards, a non-linear classifier was used to effectively separate this difficult overlapping subset. They suggested a novel evaluation metric, named as Dynamic Weighted Entropy (DWE), to identify the intersected subset's consistency in order to obtain desirable properties. This being a trade-off between the several minority class outliers omitted and the percentage of imbalance category in the overlaying subgroup. Making use of DWE, time spent for looking appropriate hyper-parameters was significantly reduced. Substantial studies that make use of Kaggle fraud detection dataset along with a massive digital transactions showed, their approach exceeds state-of-the-art methods by a substantial gap.

Some other related work is illustrated as below:

Year	Procedure(s)	Outcomes	Shortcomings
2013	Wrapped in BMR and including Linear Regression, Random forest and Decision Trees	BMR Wrapping produces superior performance.	To rebalance the data, under sampling is used.
2017	Compare 11 classifiers that were put to the test on 71 different datasets.	The effects of Gradient Boosting Decision Trees are easier and quicker.	As a metric, use Accuracy and AUC.
2018	Gaussian Naïve Bayes, Random Forest and Best Brains Exchange	BBE has the highest accuracy, but RF outperforms BBE when dealing with massive data sets.	Accuracy is used as a metric, with no cost constraints or balance.
2018	RF and XGBoost dual classifier	Results that are similar	no balancing, Orthodox methods were used
2018	SVM for just one class and T2 control charts	High precision and a low false positive rate	no balancing, Accuracy as a metric, real-world data (which can't be compared),
2018	Autoencoders with RBM	Improved Area Under Curve	no balancing, As a metric, AUC is used
2013	Decisions Trees that are cost-sensitive	The results were superior to those of the DT, ANN, and SVM algorithms.	No balancing
2019	Cost-sensitive SVM, 21 datasets	Various datasets were used, better results were obtained.	no balancing, As a metric, AUC is used
2018	Linear Regression that are cost-sensitive	Improved Area Under Curve	AUC as a metric with no balancing, credit score issue rather than a fraud detecting issue.
2019	Cost-sensitive SVM, 21 datasets	Improved Area Under Curve	No use of Swindle Revealing dataset, Area Under Curve as measure

Year	Procedure(s)	Outcomes	Shortcomings
2019	Cost-sensitive weighted random forest	G-mean, F-measure and Area Under Curve values	Area Under Curve as a metric with no balancing,
2020	ICSRealBoost ,ICSAdaBoost, and ICSGentleBoost	Improved F-Score	Datasets for Fraud Detection are not being used.

PROPOSED METHODOLOGY

As the credit card dataset is highly imbalanced, we propose an approach for balancing it in more effective manner so that it can be used for smooth and effective learning by machine learning algorithms.

Proposed Dataset Balancing Algorithm

Input:

D : Training Dataset
$\{x_i, y_i\}$: Sample from D
n : Number of samples in D
n_r : No of instances of minority class
n_x : No. of instances of majority class
d : No. of features
T : Threshold for the overall amount of imbalance that can be allowed
Given Conditions : $n_r + n_x = n$ and $n_r \preceq n_x$

Procedure:

1) Determine the Imbalance Degree(ID)

$$ID = n_r / n_x$$

2) If (ID < T)

a) Calculate SS which is number synthetic samples from the minority class would be needed:

$$SS = (\mathbf{n_x} — \mathbf{n_r}) \times \mathbf{B_L}$$

The balance level of the produced synthetic samples is B_L. There is a complete equilibrium of two groups if $B_L = 1$.

b) Calculate the ratio RM_i for each minority sample(x_i) by finding the k-nearest neighbours based on Euclidean distance.

$$RM_i = \mathbf{NS_i / K}$$

where NS_i is the no. of instances in x_i's K closest neighbours that are members of the majority class, thus value of ri lies between 0 and 1.

c) Normalize $\mathbf{DD} \leftarrow RM_i / \sum RM_i$**,** such that DD is density distribution

d) For each minority data point , no. of synthetic sample is intended such as:

$$SS_i = DD \times \mathbf{SS}$$

e) Create ss_i data specimens for the outnumbered category , example x_i using the steps below:

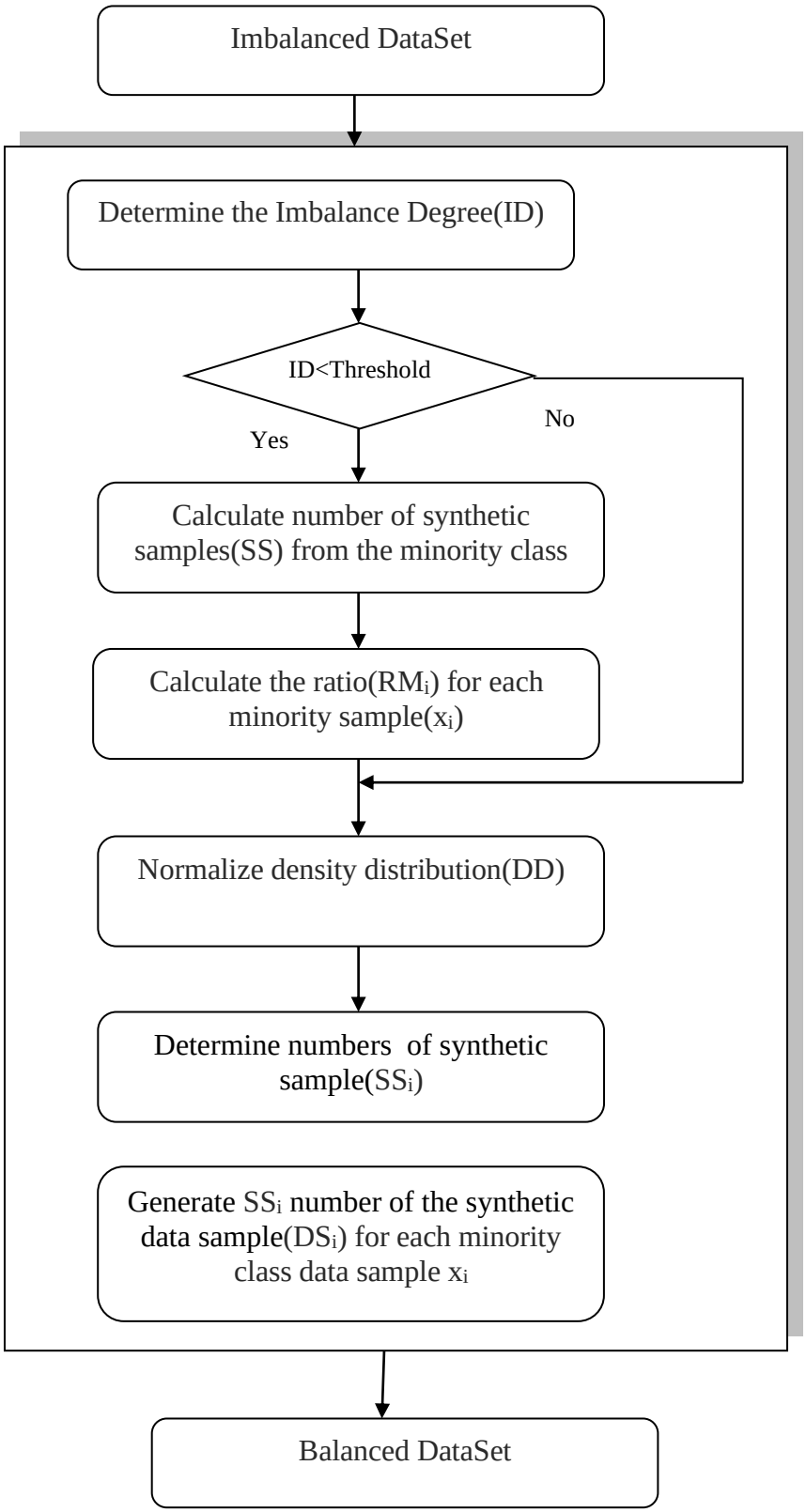
Repeat from **1 to SS_i**:

i. For data x_i, pick one minority data sample(x_p), at random from the K closest neighbours.
ii. Generate the synthetic data sample(DS_i):

$$\mathbf{DS_i = x_i + (x_p - x_i) \times W}$$

where $\mathbf{(x_p - x_i)}$ results in the variance vector in d-dimensional spaces, W being the arbitrary number which lies between 0 and 1

The flow of proposed algorithm is illustrated as below:



For getting better results we exploited Stratified K Fold Cross Validation approach. Executing the idea of stratified sampling making use of cross-validation that make certain, the training and test sets results into alike ratio of the features interested of the actual repository. On the execution with the target variable make sure that the outcome of CV results in a near approximation of error generalized.
The suggested approach used Extreme Gradient Boosting (XGBoost) procedure in order to evaluate the effectiveness of the presented resampling approach.

EXPERIMENTAL EVALUATION

Dataset

We utilize the Credit Card Dataset in this paper, which contains 284,807 transactions with 492 fraud cases (0.172 percent fraudulent) and two class values ("1" when there is fraud and "0" when there isn't).

Evaluation Metrics

When XGBoost algorithm was applied on over sample dataset (through SMOTE and Proposed approach), obtained confusion matrices are presented in Figure 3 and Figure 4 respectively.

Fig. 3. XGBoost (SMOTE Oversampling with StratifiedKFold CV)

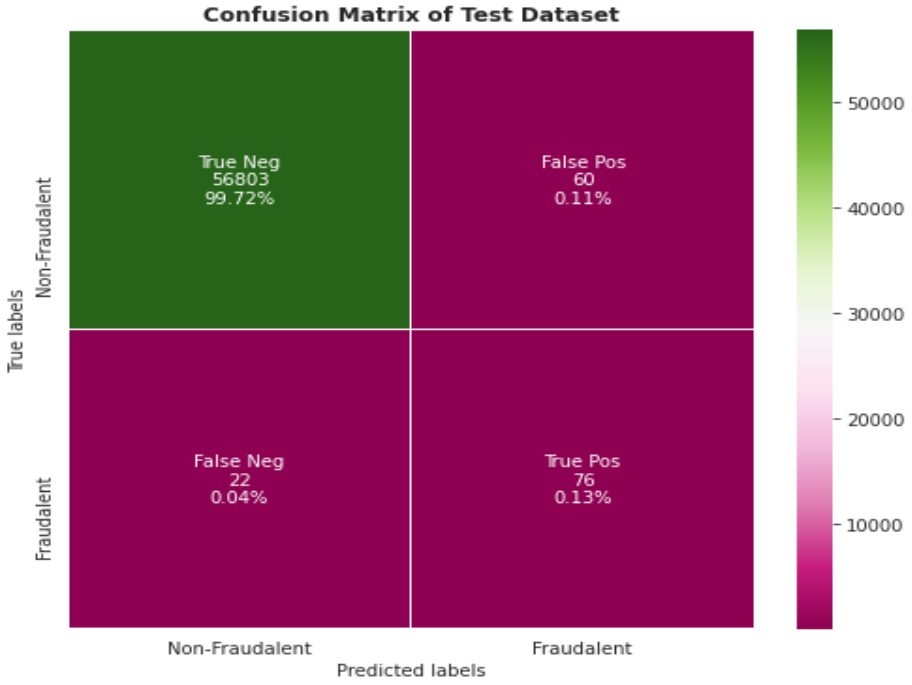


TABLE I. PERFORMANCE METRICS OF XGBOOST (SMOTE OVERSAMPLING WITH STRATIFIED K FOLD CV)

	Precision	Recall	F1 Score	Support
Non Fraudulent	0.99961	0.99894	0.99927	56863
Fraudulent	0.55552	0.77551	0.64957	98
Accuracy	0.99856	0.99856	0.99856	0.99856
Macro Avg	0.77921	0.88722	0.82442	56961
Weight Avg	0.99885	0.99856	0.99867	56961

Fig. 4. XGBoost (Proposed Oversampling with Stratified KFold CV)

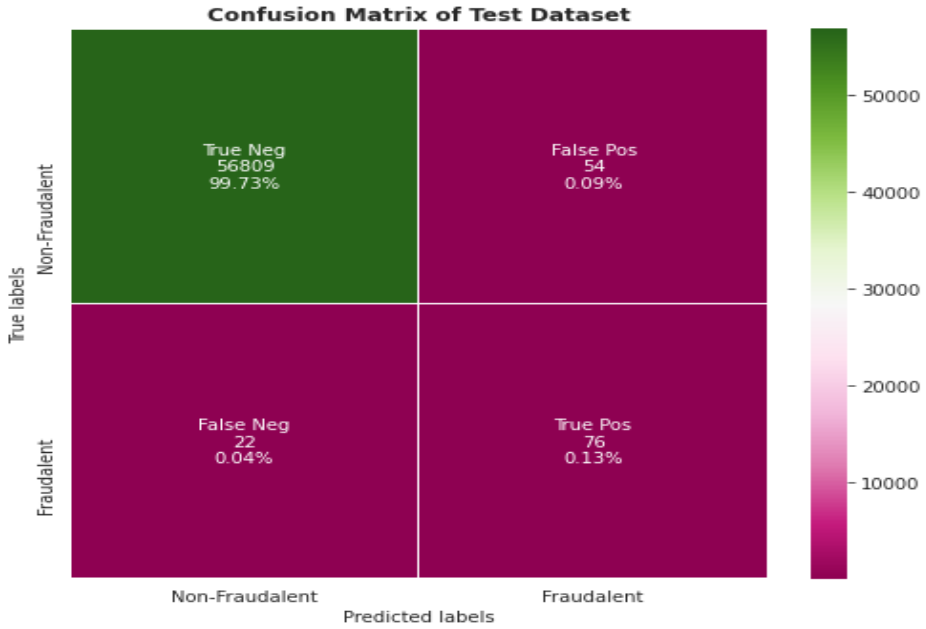


TABLE II. PERFORMANCE METRICS OF XGBOOST (PROPOSED OVERSAMPLING WITH STRATIFIED KFOLD CV)

	Precision	Recall	F1 Score	Support
Non Fraudulent	0.99961	0.99905	0.99933	56863
Fraudulent	0.58461	0.77551	0.66666	98
Accuracy	0.99866	0.99866	0.99866	56961
Macro Avg	.79211	0.88728	0.83299	56961
Weight Avg	0.99889	0.99866	0.99875	56961

CONCLUSION

The suggested work, overviewed the issue of credit card scam uncovering and presented strategies for reducing data unbalance using resampling SMOTE as a pre-process. We also proposed a novel resampling approach for balancing it in more effective manner so that it can be used for smooth and effective learning by machine learning algorithms. We used the used Extreme Gradient Boosting (XGBoost) algorithm for measuring the performance of proposed approach. We compare XGBoost on resampled data set (through SMOTE and proposed approach) by accuracy, precision, recall and F1-score measures. Finally, we found that our proposed approach outperforms the existing approach.

REFERENCES

[1]. Alam, T.M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U., Shabbir, S., Li, J. and Khushi, M., 2020. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. IEEE Access, 8, pp.201173-201198.

[2]. Atir, M. and Haydoutov, M., 2020, February. Tree-Based Bagging and Boosting Algorithms for Proactive Invoice Management. In 2019 International Conference on Advances in the Emerging Computing Technologies (AECT) (pp. 1-6). IEEE.

[3]. Bahnsen AC, Stojanovic A, Aouada D, Ottersten B (2013) Cost sensitive credit card fraud detection using Bayes minimum risk. In: 2013 12th international conference on machine learning and applications, vol 1. IEEE, pp 333–338

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), pp.1937-1967.

[4]. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp 785–794

[5]. Credit Card / Fraud Detection - dataset by vlad | data.world.

[6]. 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

[7]. Dhankhad S, Mohammed E, Far B (2018) Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In: 2018 IEEE international conference on information reuse and integration (IRI), IEE, pp 122–125

[8]. Guzmán-Ponce, A., Sánchez, J.S., Valdovinos, R.M. and Marcial-Romero, J.R., 2021. DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem. Expert Systems with Applications, 168, p.114301.

[9]. Ghorbani, R. and Ghousi, R., 2020. Comparing different resampling methods in predicting Students' performance using machine learning techniques. IEEE Access, 8, pp.67899-67911.

[10]. Iranmehr A, Masnadi-Shirazi H, Vasconcelos N (2019) Cost-sensitive support vector machines. Neurocomputing 343:50–64

[11]. Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V. and Nappi, M., 2021. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access, 9, pp.39707-39716.

[12]. Li, Z., Huang, M., Liu, G. and Jiang, C., 2021. A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. Expert Systems with Applications, 175, p.114750.

[13]. Mrozek, P., Panneerselvam, J. and Bagdasar, O., 2020, December. Efficient Resampling for Fraud Detection During Anonymised Credit Card Transactions with Unbalanced Datasets. In 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC) (pp. 426-433). IEEE.

[14]. Mohammed, R., Rawashdeh, J. and Abdullah, M., 2020, April. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 243-248). IEEE.

[15]. Mohammed RA, Wong KW, Shiratuddin MF, Wang X (2018) Scalable machine learning techniques for highly imbalanced credit card fraud detection: a comparative study. In: Pacific rim international conference on artificial intelligence. Springer, Cham, pp 237–246

[16]. Moreira, F.R., Nunes, R.R., Giozza, W.F. and Nze, G.A., 2020, June. Optimization of the performance of an online payment application by the improviment of its infrastructure. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-2). IEEE.

[17]. Pandey, D., Tiwari, R.G. and Kumar, P., Machine Learning: Adaptive Negotiation Agents in E-Commerce, International Journal of Computer Applications, Vol 166(10), pp 21-30, 2017.

[18]. Park Y, Luo L, Parhi KK, Netoff T (2011) Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia 52(10):1761–1770

[19]. Pumsirirat, A. and Yan, L., 2018. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of advanced computer science and applications, 9(1), pp.18-25.

[20]. Raghuwanshi, B.S. and Shukla, S., 2020. SMOTE based class-specific extreme learning machine for imbalanced learning. Knowledge-Based Systems, 187, p.104814.

[21]. Sahin Y, Bulkan S, Duman E (2013) A cost-sensitive decision tree approach for fraud detection. Expert Syst Appl 40(15):5916–5923

[22]. Sharifnia, E. and Boostani, R., 2020. Instance-based cost-sensitive boosting. International Journal of Pattern Recognition and Artificial Intelligence, 34(03), p.2050002.

[23]. Son, M., Jung, S., Jung, S. and Hwang, E., 2021. BCGAN: A CGAN-based over-sampling model using the boundary class for data balancing. The Journal of Supercomputing, pp.1-25.

[24]. Taha, A.A. and Malebary, S.J., 2020. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, pp.25579-25587.

[25]. Tran PH, Tran KP, Huong TT, Heuchenne C, HienTran P, Le TMH (2018) Real time data-driven approaches for credit card fraud detection. In: Proceedings of the 2018 international conference on e-business and applications. pp 6–9

[26]. Vikas Khullar, Raj Gaurang Tiwari, Ambuj Kumar Agarwal and Soumi Dutta, "Physiological Signals based Anxiety Detection using Ensemble Machine Learning", International Conference on Cyber Intelligence and Information Retrieval(CIIR 2021), 2021

[27]. Xie, Y., Qiu, M., Zhang, H., Peng, L. and Chen, Z., 2020. Gaussian Distribution based Oversampling for Imbalanced Data Classification. IEEE Transactions on Knowledge and Data Engineering.

[28]. Zhang, C., Liu, C., Zhang, X. and Almpanidis, G., 2017. An up-to-date comparison of state-of-the-art classification algorithms. Expert Systems with Applications, 82, pp.128-150.

[29]. Zhang, L., Ray, H., Priestley, J. and Tan, S., 2020. A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. Journal of Applied Statistics, 47(3), pp.568-581.