Volume: 9 Issue: IX Month of publication: September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38225

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Improvising Balancing Methods for Classifying Imbalanced Data

Himani Tiwari[1], Dr. Sheetal Rathi[2]
[1]ME second year in computer engineering, Thakur college of engineering and technology, Mumbai, India
[2]Professor (Department of computer engineering) thakur college of engineering and technology

Abstract: Class Imbalance problem is one of the most challenging problems faced by the machine learning community. As we refer the imbalance to various instances in class of being relatively low as compare to other data. A number of over - sampling and under-sampling approaches have been applied in an attempt to balance the classes.

This study provides an overview of the issue of class imbalance and attempts to examine various balancing methods for dealing with this problem. In order to illustrate the differences, an experiment is conducted using multiple simulated data sets for comparing the performance of these oversampling methods on different classifiers based on various evaluation criteria. In addition, the effect of different parameters, such as number of features and imbalance ratio, on the classifier performance is also evaluated.

Keywords: Imbalanced learning, Over-sampling methods, Under-sampling methods, Classifier performances, Evaluation-metrices.

I. INTRODUCTION

The last decades have brought a huge development in human knowledge and computational ability and have embedded computers and automation in all areas of production and common life. The process of computerizing and automation transforms our world making it more and more digital with most of operations running automatically with computers and data flows. The increasing need for automation and more efficient algorithms for data manipulation, analysis and intelligent usage has been an extremely active research area in the intersection of computer science and statistics often referred to as machine learning. However in some areas automation is still not fully developed, especially in tasks of classification on imbalanced data. Such tasks are very common in disease diagnostics, fraud detection, quality control, failures prediction, suspicious activities detection and other cases when there is substantial imbalance between amounts of instances of different classes in the data.

This is between-class imbalance, but the problem of imbalanced learning may take two other forms: a) within-class imbalance, when predictors have skewed distributions, and b) imbalanced misclassification costs. Imbalanced distributions of predictors might be relevant both for classification and regression tasks and also for unsupervised learning problems. Imbalanced misclassification costs take place when losses of incorrect classification substantially differ between classes and this creates additional constraints in performance evaluation of learning algorithms. In this thesis we deal with between-classes imbalance leaving the other forms to future work.

Classification with imbalanced data creates hard and nontrivial problems for Statistical Learning for several reasons: 1) most standard learning algorithms suffer from substantial decrease in performance as they were designed for balanced data, 2) there is no way to assess performance of learning algorithms objectively because most well-known classification performance metrics become inefficient and misleading, 3) there is no comprehensive theory for Statistical Learning with imbalanced data, and there is no theoretically motivated way to solve it.

In past, many techniques have been proposed to address difficulty of class balance at data and algorithm level .At data level, these explanation include several dissimilar forms of resampling, such as chance oversampling with surrogate, random undersampling, and focused (or directional) oversampling, where no new samples are formed but the sample must be replaced by centralized and It is not random, centralized (or corrected).) undersampling, where the examples to be eliminated are selected intensively, oversampling is performed by produce new samples or a combination of above techniques [5, 12, 23, 4].

At the algorithm level, solution embrace adjusting the cost of each class to handle class balance (cost-sensitive learning) [58, 40, 24, 21, 50], adjusting probability estimation of leaves (when working with decision trees) or adjusting decision threshold and learning based on recognition (one type of learning) [34, 49] instead of unfairness based (two types) learning.

A. Approaches To Handle Class Imbalance Problems

1) *Data:* Level approach as resampling technique is used or a pre-processing process to rebalance the class distribution. Resampling is a common technique for this approach. The main goal is to gain more data from minority class. So, there are two approaches namely Undersampling and Oversampling used in it. If we removed the instances from majority class to balanced the datasets it's called undersampling and if we add the similar instances in minority class to balanced the ratio of the class then it's called oversampling. Now we see different approaches below in detail.[1]
2) *Undersampling:* It's a technique to balance the class distribution for classification of datasets. In imbalance classes are categories in 2 classes one is majority class and another is minority class. Under sampling basically reduced the data simply from consists datasets by eliminating some examples from majority class to balance the objectify to equalizing numbers to other given class. This method used on training datasets. Basically under sampling methods coordinate with oversampling technique of minority class and both this techniques combines and give better performances on training datasets. The random selecting is simple under sampling technique which selects the examples from majority class and deleting from training datasets.
3) *Near Miss:* It's a technique below the sample. Instead of updating the minority class by distance, this would make the majority class the same as the minority class. There are three experiments:
4) *Random Under Sampling:* Randomly remove samples from most classes, with or without replacement. This is one of the first methods used to eliminate imbalances in the data, but as category differences increase it can eliminate necessary or important samples.
5) *Oversampling:* This method used to reduction in minority class so that the there is an equal and balanced distribution in class. SMOTE is a commonly recommended in oversampling.
6) *ADASYN:* Adaptive synthetic sampling advance expands on process of SMOTE by shifting the importance of the classification boundary to that minority class which is difficult. ADASYN uses a weighted spread for not similar minority class examples as per their level of complexity in learning, where more synthetic data is created for minority class examples that are difficult to learn.
7) *SMOTE:* There are a number of processes available to oversample a dataset used in a typical classification problem (using a classification algorithm to classify a set of images, given a labelled training set of images). The most common technique is known as SMOTE: Synthetic Minority Over- sampling Technique. To illustrate how this technique works consider some training data which has s samples and f features in the feature space of the data. Note that these features, for simplicity, are continuous. As an example, consider a dataset of birds for classification. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight (all continuous). To then oversample, take a sample from the dataset and consider its k nearest neighbours (in feature space). To create a synthetic data point, take the vector between one of those k neighbours, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point. Many modifications and extensions have been made to the SMOTE method ever since its proposal. [7]

II. RELATED WORK

1) Chung-Chih Lin et.al (2020) Data collection on wireless sensor networks has become one of major study topics. In the literature, several studies have planned algorithms for data assortment through receiver phones. As for foreign exchange losses, these studies can be divided into two categories. This article considers busses as mobile receivers and proposes a circulated data collection algorithm to collect greatest data from sensors and extend the lifespan of wireless communication networks. Evaluation of performance of the work shows that the energy-efficient data center (EBDC) system has better accessibility and a longer lifespan than existing mechanics.
2) Thamer Khalil Esmeel et.al (2020) in the last few years, data confidentiality has been a major area of explore. Data often contains sensitive data sets, or exposure to these fields can be detrimental to the interests of the person associated with the data. To solve this problem, confidentiality technology can be used to prevent the identification of individuals by manipulating sensitive data into the database to defend sensitive information, or anonymous data may be used by third parties for uninterrupted analytical purposes. In this study, we examined a specific privacy technique, k-anonymity with different \ pmbk values for the number of different \ pmbc columns in data set. Next, calculate loss of information due to anonymity k. Anonymous files are categorized using certain machine learning algorithms (i.e. Naive Bayes, J48, and network) to check for the balance between non-data processing and data usage. If classification is correct, the best values of \ pmbk and \ pmbc are obtained. Therefore, optimum \ pmbk and \ pmbc can be used in the k-anonymity algorithm to eliminate most select number of columns in data set.

3) P. Tamilarasi et.al (2021) in the world of cloud data, cloud service providers (CSPs) provide capabilities to customers through portable virtual cloud data sources. Researchers are paying more attention to load coordination because it has a full effect on system behavior. In this article, a balanced algorithm based on prediction and virtual machine migration (VM) (PLBVM) algorithm is designed for the data cloud environment. With this algorithm the load of each server is estimated. If the future value exceeds the upper or lower limit, it indicates that the load is not balanced, which leads to VM migration. In a virtual machine migration, choose a virtual machine with a short migration time and sufficient resources. Then resume the operation on the migrated VM. The experimental results show that PLBVM achieves smaller delay and execution time in a different way.
4) Harpreet Singh et al. (2018) Data mining is a hot topic in the field of data mining. There are many programs that generate a lot of data. These data need to be sorted to gather analytical data. In order to use this analysis in the decisions made by the relevant organizations. But the data collected from the numbered sources is not balanced. This data needs to be organized to produce balanced data. Because dealing with unbalanced data will be an effective process. There are a variety of differences in the data. In the current research paper, these different types of differences will be reviewed. In this way, the processing and analysis of the data will become easy and efficient. These differences exist in the data set, such as small separation, insufficient data, overlapping categories, influence of noise data, limits etc. Using different techniques, the problems in these data are exacerbated. This allows you to enter balanced data for data mining purposes.

III. PROPOSED METHODOLOGY

In given figure 1 there is proposed system in which various approaches are used to balance the class. Sampling techniques are used to perform by adding and removing the instances of classes. Classifiers mainly perform to balance out the data. Training and Testing approaches are done by which we found how much of data is rectifying by which we finally applying our algorithms to find better performance and average of all techniques which is going to be used in it.

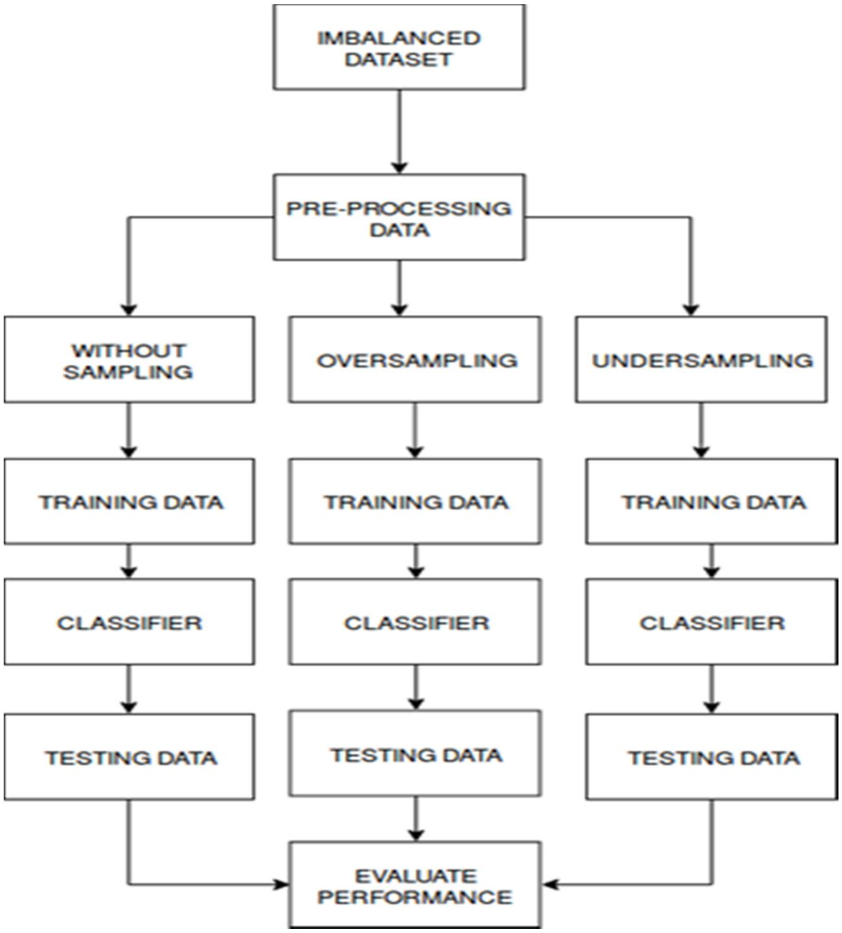


Fig 1 proposed flow

IV. SYSTEM OVERVIEW

Here, figure 1 explains the flow of arrangement in which we elucidate various components of the system as we seen there is a selection of data by which various majority and minority classes duplication are formed then new data is produce by interpolation among various majority and minority classes. Then we use oversampling and under sampling techniques. Now majority class and minority class are under sampled and oversampled and inserting synthetic examples to their nearest neighbors as taking k minority class. This practice efficiently forces decision region of minority class to become more common. [21].

As we considered the flow of system, we earlier taking a datasets by which we having 2 sampling techniques under sampling and oversampling while we taking Near Miss is an under sampling approach. In this we balance the class allotment by deducting of greater part class randomly. Now if we if the two samples of dissimilar classes are close to each other we remove samples of majority class and spaces are made among two classes if there is loss of information can occur then we using near-neighbor methods which belongs to under sampling technique.

The same for oversampling technique we were taking SMOTE that's better version which help minority class improvement which balance the class distribution by increasing the minority class value to close with other samples of majority class.

A. *Training Data*

Training data is an example of data set used in a study, i.e., used to match parameters (such as weights) of, for example, a classifier. Most of the methods of finding collaborative relationships through training data typically combine data, which means that we can identify and use explicit relationships that are not common in training data.

B. *Test Data*

The test data is data independent of the training data, but follows the same possible section as set out on the training data. If the model that matches the training data is also compatible with the set of test data, there was a slight override (see figure below). Compared to the experimental data, the combination of training data often indicates over expression. Thus, a test case is a series of examples that are used to evaluate the performance (i.e., globalization) of a fully defined category. [14]

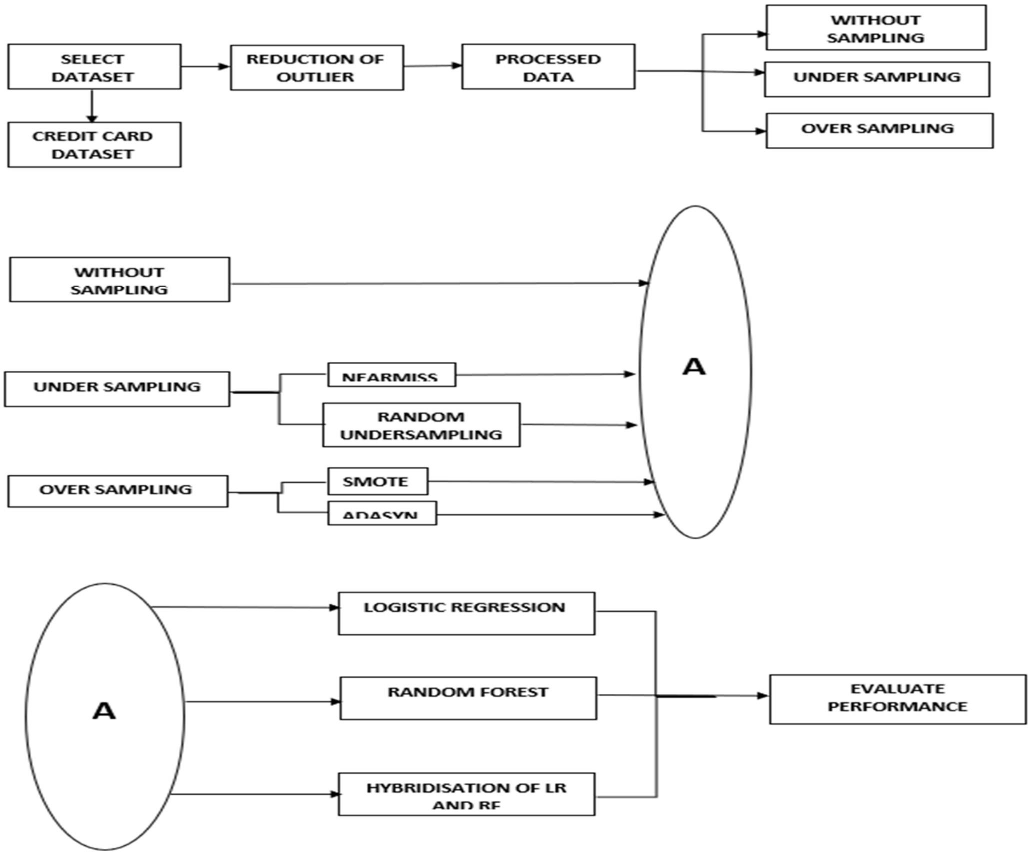


Fig .2 System Overview

V. CLASSIFIER

There are different types of classifications in machine learning. We will list the discriminators, and use the following lists to weigh the data for our system. Here we use 2 category types to implement algorithms in machine learning. [10]

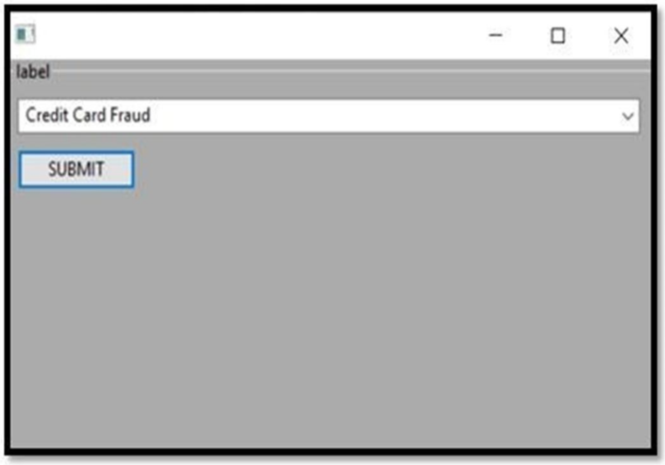
1) *Logistic Regression:* The most logical regression is the first machine-learned algorithm known to all scientists. The goal of the logic analysis model is to find the relationship between one or more properties as an independent variable and the target variable as a dependent variable.
2) *Random Forest:* A deciduous forest or a continuous decision forest is a learning method for classification, regression and other activities. It builds multiple decision trees during training and outputs a category as a category model (classification) or a pre / average prediction (review). The sudden decision forest corrects the habits of the decision tree that are not worthy of the established training. After applying the class inequality technique, we applied the method by eliminating class inequality.

VI. IMPLEMENTATION STEPS

We will first consider the degree of imbalance in the primary data. Most exchanges are not fraudulent. If we use this data structure as the basis of our predictive models and analyzes, we may get a lot of mistakes, and the algorithm may be successful because we will assume that most of the exchanges are not fraudulent.

1) Here, we use unbalanced data, namely credit card fraud.
2) We trained the trainer because I mixed it with LR and RF to get better results, because I found that the software retraction used was better than the other categories. Therefore, we use logical regression as a line constructor.
3) Two amplification techniques, namely Nearmiss, indirect undersampling and two scattering techniques, namely SMOTE and ADASYN, were used to balance the data.
4) Using the Logistic Logistics classification and balancing techniques above, we find as a measure the true average number of search rates.
5) Notice that hair resuscitation is better than resuscitation
6) The average credit card fraud rate is better than driver's insurance.

We selected a data set, as described in point 1 above. We can see in the picture that we have selected the credit card fraud cases.



label

Credit Card Fraud

Fig. 3 : Selection of the credit card fraud dataset.

As discussed in the previous section, we used two data sets and applied amplification and elongation techniques to obtain data points the following are the results of the study. The comparison of the different weighing techniques was done by regulating the software as a class. We showed the repeated number of digits for each.

After applying different weighting techniques, we obtained good results from the balanced data from the unbalanced data. See the chart in Figure 2 for the results. In the figure, we show the x-axis as a technique or method used to weigh the data, while the y-axis provides accurate mean information. Now for evaluating the model of system we use AUC by which we decide our model perform better after using balancing methods.

VII. PERFORMANCE EVOLUTION

Area under Curve (AUC) commonly used for evaluation of metrics. It is commonly used for binary classification problem. AUC of classifier is equal to probability chosen positive ratio than negative ratio chosen as an example.

1) True Positive Rate (Sensitivity): True Positive Rate is defined as -TP / (FN+TP). True Positive Rate corresponds to fraction of positive data points that are properly considered as positive, with deference to all positive data points.
2) True Positive Rate = True Positive / False Negative+ True Positive
3) False Positive Rate is defined as - FP / (FP+TN). False Positive Rate corresponds to proportion of negative data points that are incorrectly considered as positive, with reverence to all negative data points.
4) False Positive Rate = False Positive / False Positive+ True Negative

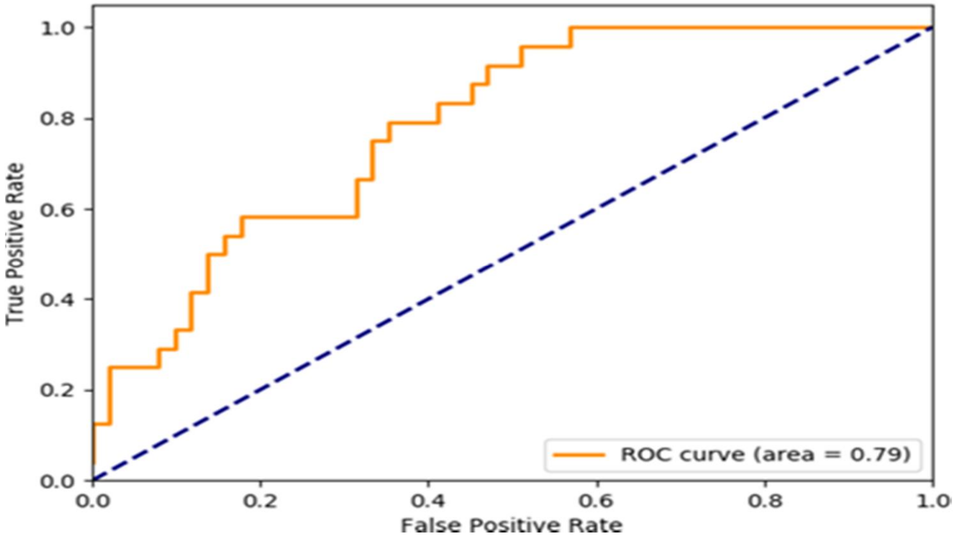


Fig 4: ROC-AUC Curve

As discussed in the earlier sections, we have used two dataset and have implemented techniques of under sampling and oversampling to get the balanced data. Below are the observations. Comparisons of different balancing techniques with the Logistic Regression as the base line classifier have been done. We have shown the average precision recall score for each. Now, we will check the dataset in terms of how imbalance our dataset is. See the graph below.

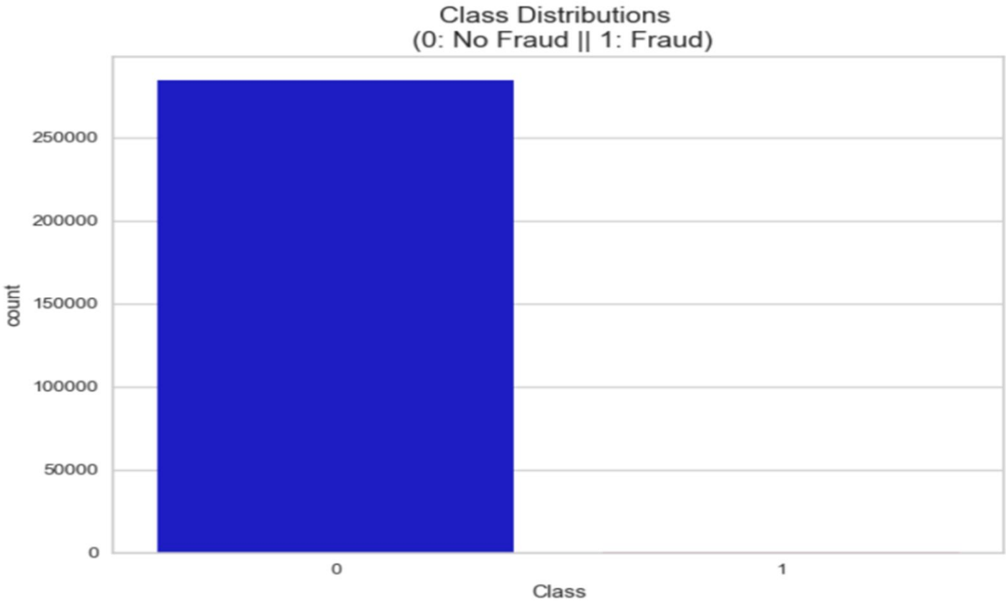


Fig 5 Graph of Imbalanced Dataset

Four types of classifiers and chose which classifier will be more efficacious in detecting fraud transactions. And I have found that the Logistic Regression classifier is more precise than the other three classifiers. Then we plot the ROC-AUC curve for all the classifiers. We then use these 2 classifiers i.e. logistic regression and random forest classifier with all the sampling techniques to get good results.

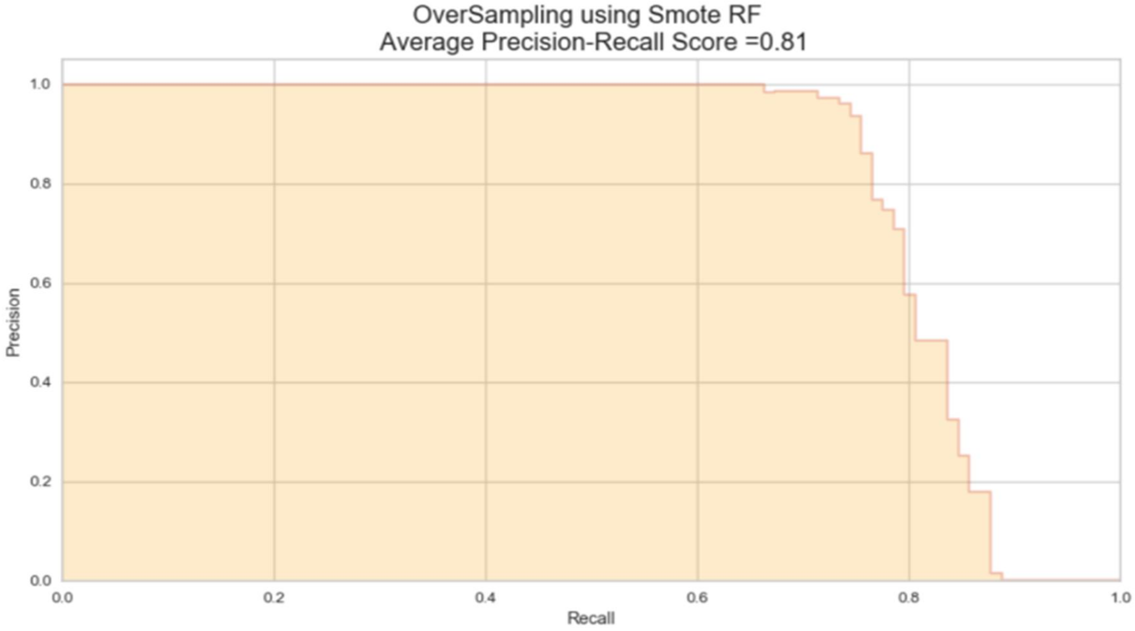


Fig 6: Average precision recall: SMOTE-LR

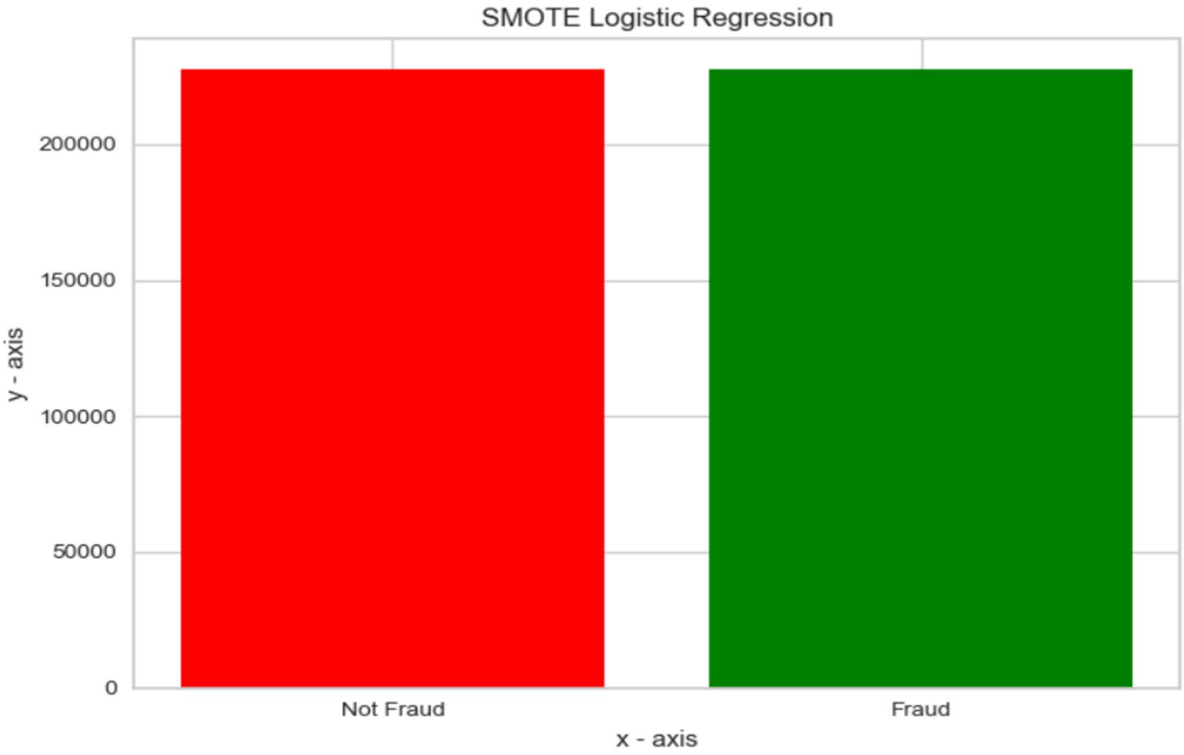


Fig 7: Average precision recall: SMOTE-RF

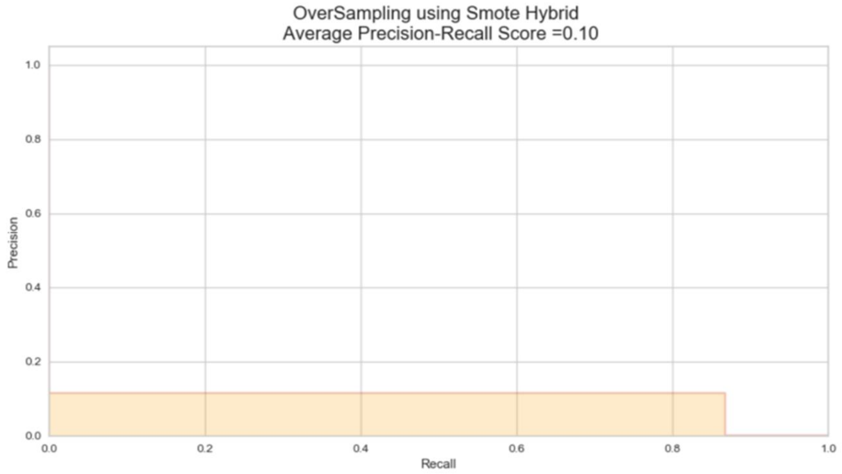


Fig 8: Average precision recall: SMOTE-Hybrid

Table 1: Observation table

Data Sets	Classifier	Balancing Techniques			
		Near Miss	Random Under sampling	SMOTE	ADASYN
Credit Card Fraud	Logistic Regression	0.02	0.72	0.78	0.79
Credit Card Fraud	Random Forest	0.64	0.91	0.81	0.82
Credit Card Fraud	Hybrid	0.00	0.05	0.10	0.10

VIII. RESULTS

We have got the desired results of balanced dataset from an imbalanced dataset after applying different balancing technique. Refer the graph in **Fig4** for the results. In the graph, we have shown the x-axis as the techniques or the methods used to balance the data-set and y-axis gives us information about the average precision recall score.

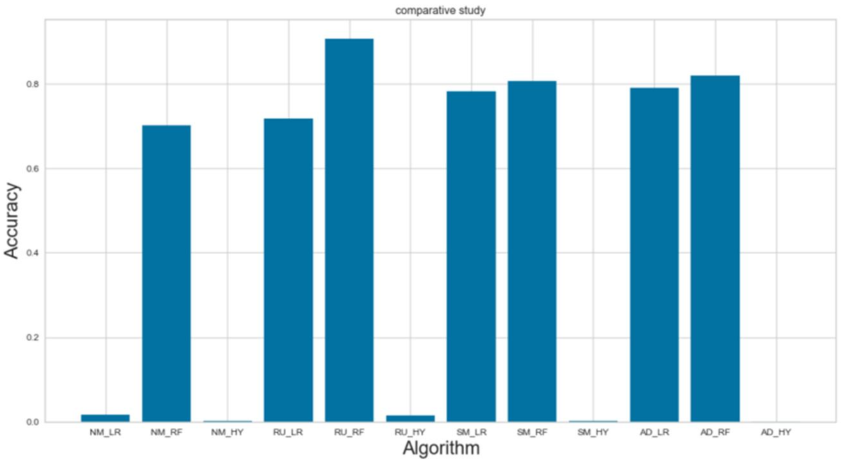


Fig 9: Accuracy performance of algorithm

IX. CONCLUSION

To conclude, we have discussed the class imbalance problem and look into different approaches used to solve it. We have also explored many different methods and algorithms to improve the class imbalance in the data sets, this includes learning about the data level approaches and the algorithm approaches. In the proposed system, we have proposed the method of oversampling and under sampling methods can be used for tackling the imbalance class problems. We have learned about the oversampling and undersampling techniques, we came to know that only after using any of the above-mentioned methods, we can overcome the problem of data imbalance. We find the novel approach to address the issue of class imbalance is sampling, oversampling and undersampling can be used to take care of class imbalance problem.

X.FUTURE SCOPE

We can take several possible directions in the future. We would like to extend our study to multi-class problems. All the methods proposed in this paper addresses two-class cases so far. Even, we can use Ensemble learning on a large collection of real datasets. In addition, we can extend our work in Big Data domain. Here, we have worked on data level approaches & would plan to conduct experimental evaluation on algorithm level approaches.

REFERENCES

[1] Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl, 7(3), 176-204.

[2] Bennin, K. E., Keung, J., Phannachitta, P., Monden, A., & Mensah, S. (2017). Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. IEEE Transactions on Software Engineering, 44(6), 534-550.

[3] Gong, L., Jiang, S., Bo, L., Jiang, L., & Qian, J. (2019). A Novel Class-Imbalance Learning Approach for Both Within- Project and Cross-Project Defect Prediction. IEEE Transactions on Reliability

[4] Cosma;Aboozar Taherkhani Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning 2020 International Joint Conference on Neural Networks (IJCNN) Year: 2020 DOI: 10.1109/IEEE Glasgow, UK

[5] Behzad Mirzaei;Bahareh Nikpour;Hossein Nezamabadi-Pour An under-sampling technique for imbalanced data classification based on DBSCAN algorithm 2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS) Year: 2020 DOI: 10.1109/IEEE Mashhad, Iran

[6] Wei Wang;Mengjun Zhang;Li Zhang;Qiong BaiImbalanced Data Classification for Multi-Source Heterogenous Sensor Network IEEE Access Year: 2020

[7] Jaewoong Kang;Mye Sohn Recursive Undersampling-Based Decision Boundary Alignment for Imbalanced Radiology Image 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) Year: 2020

[8] JUN-HAI ZHAI;SU-FANG ZHANG;MO-HAN WANG;YAN LI A Three-stage Method for Classification of Binary Imbalanced Big Data 2020 International Conference on Machine Learning and Cybernetics (ICMLC) Year: 2020

[9] Baofeng Yao;Lei Wang An Improved Under-sampling Imbalanced Classification Algorithm 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) Year: 2021

[10] Salma El Hajjami;Jamal Malki;Alain Bouju;Mohammed Berrada A Machine Learning based Approach to Reduce Behavioral Noise Problem in an Imbalanced Data: Application to a fraud detection 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) Year: 2020

[11] Xun Dong;Hongli Gao;Liang Guo;Kesi Li;Andongzhe Duan Deep Cost Adaptive Convolutional Network: A Classification Method for Imbalanced Mechanical Data IEEE Access Year: 2020

[12] Zonghai Zhu;Zhe Wang;Dongdong Li;Wenli Du Globalized Multiple Balanced Subsets With Collaborative Learning for Imbalanced Data IEEE Transactions on Cybernetics Year: 2020

[13] Qibin Wang;Lingqiao Li;Xipeng Pan;Huihua YangClassification of Imbalanced Near-infrared Spectroscopy Data 2020 12th International Conference on Advanced Computational Intelligence (ICACI) Year: 2020

[14] Chung-Chih Lin;Chih-Yung Chang An Energy Balanced Data Collection Mechanism for Maximizing Throughput using Uncontrolled Mobile Sink in WSNs 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan) Year: 2020

[15] Harpreet Singh Bror Balancing of the Imbalance data classification using Data Intrinsic characteristics 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) Year: 2018

[16] Thamer Khalil Esmeel;Md Munirul Hasan;Muhammad Nomani Kabir;Ahmad Firdaus Balancing Data Utility versus Information Loss in Data-Privacy Protection using k-Anonymity 2020 IEEE 8th Conference on Systems, Process and Control (ICSPC) Year: 2020

[17] P. Tamilarasi;D. Akila Prediction based Load Balancing and VM Migration in Big Data Cloud Environment 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM) Year: 2021