Hybrid Resampling Technique to Tackle the Imbalanced Classification Problem

Payal Gulati (gulatipayal@yahoo.co.in)
J.C. Bose University of Science and Technology, YMCA https://orcid.org/0000-0002-3294-2575

Research Article

Keywords: Imbalanced classification, data-level approach, resampling technique, undersampling, oversampling, minority class, majority class, SMOTE, NCL

Posted Date: June 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36578/v1

Hybrid Resampling Technique to Tackle the Imbalanced Classification Problem

[1] Jyoti Kashyap, [2] Dr. Payal Gulati

[1] Research Scholar, Department of Computer Engineering, J. C. Bose University of Science and Technology, YMCA Faridabad

[2] Assistant professor, Department of Computer Applications, J. C. Bose University of Science and Technology, YMCA Faridabad

ABSTRACT

In the real-world domain, many learning models faces challenge in handling the imbalanced classification problem. Imbalanced classification is a scenario where the number of data points in minority class is much lower than that of the majority. Our primary concern is the minority class, which is often neglected by learning models while predicting the values. This problem can be tackled at the data-level by using resampling techniques. In this research, hybrid of Synthetic Minority Oversampling Technique (SMOTE) and Neighborhood Cleaning Rule (NCL) is proposed to balance the data points of the classes. For experiment real-world dataset of credit card transaction has been utilized where the fraudulent (or malefactor) transaction needs to be identified. This imbalanced dataset after resampling is classified by using the logistic regression model. The experimental results depict that the learning model has correctly identified the malefactor in the balanced dataset than the original dataset. Through balancing the datasets, the proposed technique aims to enhance the performance of the learning model in order to correctly identify the cases of the minority class.

Keywords: Imbalanced classification, data-level approach, resampling technique, undersampling, oversampling, minority class, majority class, SMOTE, NCL

1. INTRODUCTION

In the domains of practical world, the primary goal of the learning model is to spot the cases of the unusual occurrence. This process of spotting becomes cumbersome for learning models in the presence of exceedingly imbalanced classes in the provided datasets. The imbalance is the issue where the cases of majority class are much more than that of minority. Our primary concern is the minority class, which is often neglected by learning models while predicting the values. The

main reason behind this is, the standard learning model aims to maximize the accuracy rate and hence the extra attention is not given to the cases of minority.

In recent years, many data scientists and researchers are digging deep into exceeding problem of imbalance. The issue of unbalancing appears in many practical world application such as the detection of oil spills through image captured by satellite [1], the detection of telephonic call that are fraud intentional [2], monitoring of defect in the gearbox of the systems [3], information retrieval [4] and diagnosis of uncommon cases of medical conditions such as cancer disease [5].

To resolve this problem of exceeding imbalance, previously many solutions were introduced in the practical world by researchers. At two levels the solutions were proposed, one is at data-level [6] and other is at the algorithmic-level [7]. Collaboration of existing standard techniques [8] was also proposed to tackle this exceeding imbalance issue.

In this research, a novel approach is proposed to compensate the problem of exceedingly imbalance in the diffusion of the classes. For experiment the data set of credit cards is utilized. Credit cards have become major part of all cashless transaction. As transaction through credit card is the most well-known technique for mode of payment in the ongoing years, the fraud exercises have expanded quickly. In this, the transaction made by malefactor is associated with the cases of the minority class where as the transaction made by the credible person is associated with the cases of majority class. Here, the data set is imbalanced by nature and available from kaggle [9].

This research aims to identify the suspicious behavior in the transactions through credit card by balancing the data set. This is done by using the hybrid resampling approach at the data-level. After that, balanced data set is classified using logistic regression. This learning model of classification is executed in order to predict the cases of minority class accurately. For the assessment of the model various performance metrics are utilized. The result shows that proposed approach has successfully spotted the cases of the minority class.

1.1 IMBALANCED PROBLEM IN CLASSIFICATION

Classification [10] is a family of supervised learning where we cut down the dataset into a given different number of categories. The primary motive of classification problem is to identify the category/class under which a new data will fall. There are generally three kinds of classification problems encountered in supervised learning which are:(a) Two-class problem (fraud/not fraud), (b) Multi - class problem (positive/negative/ neutral) and (c) Multi - label problem. (Shape and size).

In this research, we are tackling two-class (or binary) problem. Generally, two queries arise in the imbalanced classification problem [11, 12]:

(a) What imbalanced classification is?

(b) Why it is renowned as the problem?

To encounter the answers for these queries the root starts from the unevenness in the data points. Unevenness is an exceptionally acute form of imbalanced classification. Unevenness in the data point is further divided into two subparts [13, 14]:

- **Definite unevenness**: Definite unevenness appears in the binary and multi-class problems of classification. The occurrence of definite unevenness is more often found in inter-class imbalance.
- **Relative unevenness**: Relative unevenness is encountered in the multi-label problem of classification. The occurrence of the relative unevenness is spotted in intra-class imbalance.

The learning of these kinds of data is cumbersome for the standard learning models. They greed for more accuracy due to which the uneven data points remained untouched and often mistaken as noise. In the cases of spotting fraud or in diagnosing disease, the misclassifying of these uneven data points can be expensive for the learning models. This unevenness of the data points leads to the imbalanced classification.

Now that the reason behind the occurrence of imbalance and the way it can be manifested in the classification is answered, the second question remains: why is imbalanced classification considered to be a problem? It is crucial to understand that, by own, the above described phenomena do not automatically imply that they impede the task of classification. But in fact, there are certain difficulties in classification learning models associated with them [11].

The root cause behind these difficulties is the way usual learning models of classification are designed. Classification learning model efficiently run when the number of instances in each class is nearly equal. Most of the models are designed to enhance the accuracy and minimize the error. The evaluation metrics of machine learning model do not correctly examine the model performance when coping with highly imbalanced datasets.

Below are the reasons behind the reduction of accuracy of existing classification learning model on imbalanced data sets:

- Classification learning model scramble with accuracy because of distribution of classes is imbalance in data sets.
- This causes the execution of existing classification learning model to get sloped towards majority class.
- The models are accuracy driven i.e. they are bound to minimize the overall error to which the minority class contributes very little.

- Classification model assumes that the data set are balanced in nature.

1.2 PROBLEM OF IMBALANCED CLASSIFICATION IN ENCOUNTERING THE FRAUD

Encountering fraud in the transaction can be split into a problem of two-class one wants to associate each new variable: either legitimate or non-legitimate based on the learning of the transaction. In this specialized issue, a humongous amount of studies has already been considered. This specific arrangement does however welcome a slight slop towards domain of two-class. The two outcome classes are highly imbalanced to each other. This depicts that it is considered that, in a normal moment, one would expect to experience a lot more legitimate transactions as compared to the fraudulent ones. While this of course is fortunate for the industries of finance involved, it does make the learning task more cumbersome. Also these problems create hassles for learning model in learning the datasets, where the occurrence of fraud is sparse.

Throughout this research, the problem of exceeding imbalance in the diffusion of the classes will be addressed. This research can roughly be divided into a conceptual and an experimental part. The conceptual part aims to extensively learn the problem and its potential solutions, while the experimental part taken as an empirical establishment of these concepts. To resolve this unconditional issue various techniques were proposed. These are discussed in the literature review.

The rest of this paper is organized as follows. The section 2 includes the standard techniques to tackle the imbalance issue and the related work done in it. The section 3 explains the proposed method. The section 4 shows the performance of the model. The section 5 gives the conclusion of the proposed method and the suggestions for the future work.

2. LITERATURE REVIEW

2.1 METHODS TO TACKLE CLASS IMBALANCE PROBLEM

In order to compensate the problem aroused by imbalance learning, many approaches at different level have been proposed. Unbalancing of the data points impede task of standard learning model due to which it sloped towards the cases of minority class (often acknowledged as “positive” class) [15], [16].

In recent years, numerous techniques have been proposed to manage this issue, both for standard learning model and ensemble methods [17], [18]. They can be ordered into three significant families [19], [20], [21] as appeared in figure 1:

- **Data Sampling**: In this technique the data points of the training datasets are treated to produce balanced diffusion of the data points in the class**.** This allows the learning model to perform in similar fashion to standard algorithms.
- **Algorithmic Modification**: in this technique the algorithms are modified in order to address the issues of class imbalance.
- **Cost-sensitive Learning**: This approach introduces the penalty cost on the learning model for classification for incorrectly classifying the data points of minority class (in our case malefactor's data points) with the cases of majority class.

This research only concerns with data sampling approach, also known as resampling techniques.

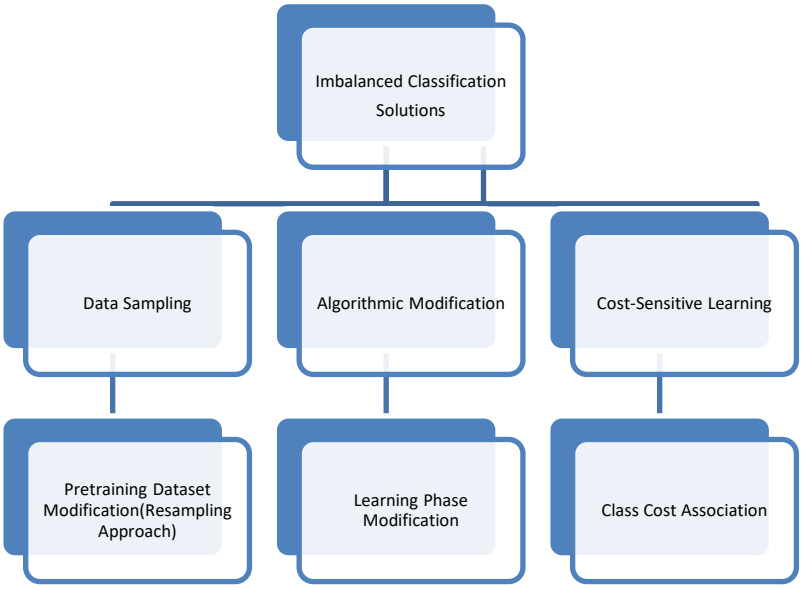

Figure 1: Categorization of solutions to Imbalanced Classification

The greater part of the learning model works most noticeably terrible within the sight of uneven diffusion of data points in the class. To overcome this issue, a data preprocessing step must be done before giving training datasets to the model. In this context, there are a few papers on resampling strategies that shows the impact of changing the diffusion of class to manage imbalanced datasets. Those works have demonstrated experimentally that, applying a preprocessing step so as to adjust the distribution of the class is typically a helpful methodology [22], [23], [24]. On account of class unbalancing issue, the data preprocessing step is performed utilizing the data level methodology, which is called resampling approach.

Approaches for the data sampling can be broken down into three aspects:

- **Down sampling or Undersampling methods**: This method involves the adjustment of the diffusion of data points in the majority class by eliminating the data points in them.
- **Up sampling or Oversampling methods**: This method involves the adjustment of the diffusion of data points in the minority class by replicating the data points in them.
- **Hybrids methods**: This method is the collaboration of both methods mentioned above. It can also term as hybrid resampling method.

2.2 RELATED WORK

2.2.1 UNDERSAMPLING METHODS

1. Marcelo Beckmann et al [25] proposed a novel downsampling method based on supervised learning algorithm. In order to balance the datasets, this method eliminates the data points on the general count of neighbor from each class. This method gave better accuracy than other Undersampling methods used for comparison. Limitation of this type of method is, there is a high chance of losing the potential data points.

2. Ivan Tomek [26] proposed a novel Undersampling method which is famous by Tomek Link method. This method adapts the property of the condensed nearest-neighbor (CNN) undersampling algorithm to eliminate the noisy data points and border line. This method differs from (CNN) because it takes both majority and minority classes into consideration. There is a requirement of modify this method to give better accuracy on the different- different learning models.

3. When there is an overlapping issue with the imbalance of the data points in the classes, the learning task becomes more cumbersome. Pattaramon et al [27] proposed four kinds of Undersampling method which are based on neighbourhood properties. These methods known as NB-based undersampling methods. These methods differ from each other in two ways. First one is the way local search is done. And second one is in the way data points of majority class are eliminated. However, with this method there is a high chance of elimination of excessive data points from the majority class and hence the accuracy of the model can be affected.

2.2.2 OVERSAMPLING METHODS

1. The general oversampling method aims to maximize the volume of minority class by creating the replicas of the data points. However, instead of creating replicas Chawla et al [28] presented a novel method called Synthetic Minority Oversampling Technique (SMOTE) which creates the

“artificial” data points rather than creating the replicas of the minority class data points. There is a risk of excessive creation of the data points that could leads to the overfitting problem.

2. On the basis of standard oversampling method SMOTE, various researchers have proposed the variants of SMOTE. György et al [21] presented and discussed a detailed empirical comparison of 85 variants of SMOTE involving 104 imbalanced datasets for evaluation. The problem aroused in the selection of the accurate learning models and in setting the parameters to evaluate these models.

3. The smote algorithm creates the synthetic examples rather than replacing the data points of the minority class. Haibo et al [22] proposed a novel technique namely ADASYN which creates the data point among the majority and minority classes, instead of creating the artificial data points. The synthetic data generated for the minority class examples are harder to learn. So the ADASYN method uses the weighted diffusion of the data points for the minority classes. This method aims to reduce the bias issue generated due to imbalance and adjust the decision boundary towards the complex data points.

2.2.3 HYBRID METHODS

The hybrid methods are the combination of the data level methods and the algorithmic level with respective combination. The need of hybridization is to overcome the problems with the data level methods and the algorithmic level methods and also to achieve the better classification accuracy.

1. The domain of medical especially in diagnosis the major hinder faced is the problem of rare positives. To overcome this issue, Cohena et al. [31] proposed a hybrid of resampling methods. For learning process, the author utilized SVM as the learning model and presented the regularization parameter which is highly dependent on class.

2. Credit scoring falls under the binary classification problem. To improve the classification process of the credit data sets, Xu Han et al. [32] proposed hybrid of the unsupervised Undersampling method and supervised oversampling method to eliminate the issue of imbalance problem.

3. S.Ancy [31] proposed a hybrid approach called as Handling Imbalanced Data with Concept Drift using an ensemble classifier model (HIDC), which is the hybrid of the ensemble learning and the resampling techniques. This method resolves the issues of the imbalance and the concept drift. This method basically integrates the data level approach with the algorithmic approach.

Previously proposed techniques were performed either on the majority class or on the minority class. Due to this many issues generated like the problem of overfitting, potential information losses and many more. Many researchers also proposed hybrid solutions to overcome the imbalance issues. These hybrid methods are either the combination of the resampling methods or

the combination of ensemble methods with resampling methods. In this research a novel hybrid of SMOTE and NCL is proposed. Both the oversampling and the Undersampling method use the principles of supervised learning. The result of performance metrics has shown the proposed method has improved the accuracy of the learning model up to greater extent.

3. PROPOSED METHOD

The proposed method involves processes that are acquisition of the datasets, data preprocessing, resampling of the training datasets, process of classification & evaluation metrics as shown in Figure 2.

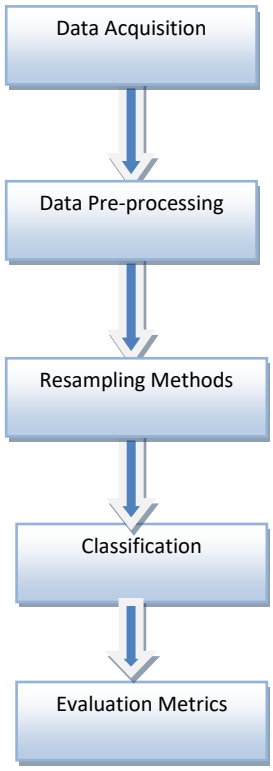

Figure 2: Pipeline of Experiments

These following steps are the explanation for each of the process on the flowchart shown in Figure 3.1.

3.1 DATA ACQUISITION

Data acquisition is the process of gathering the datasets from the sources that are publically available. It also involves the process of filtering and cleaning the datasets. In this step, the dataset is gathered from the practical-world data from kaggle [9]. The dataset is in CSV format.

This is the dataset of transactions made through credit card. The data set consists of more than 0.5 million transactions out of which less than 500 transactions falls under the cases of the minority class. There are 284315 data points associated with the majority class where as only 432 data points associated with the minority class. This is why the dataset is imbalanced and therefore used for the experiment. The exceedingly imbalance of the diffusion in the data points of the classes can be visualized in the figure 3.

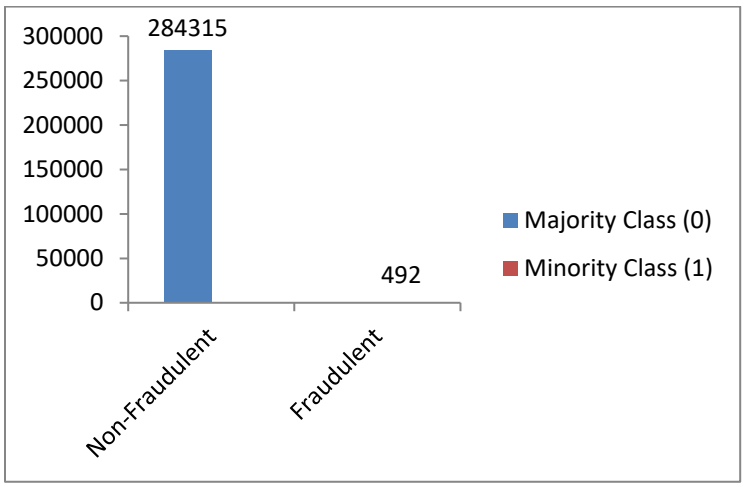

Figure 3: Imbalanced diffusion of data points in the classes

3.2 DATA PREPROCESSING

Data preprocessing is the process where the datasets are transformed, normalized and prepared for the experimentation [34]. In this step, dataset will be split into training and testing sets. The ratio of 70:30 is used for the splitting of the utilized dataset. After splitting, there are 199364 data points for training the learning model while there are 85443 data points for the testing the learning model. This is shown in figure 4.

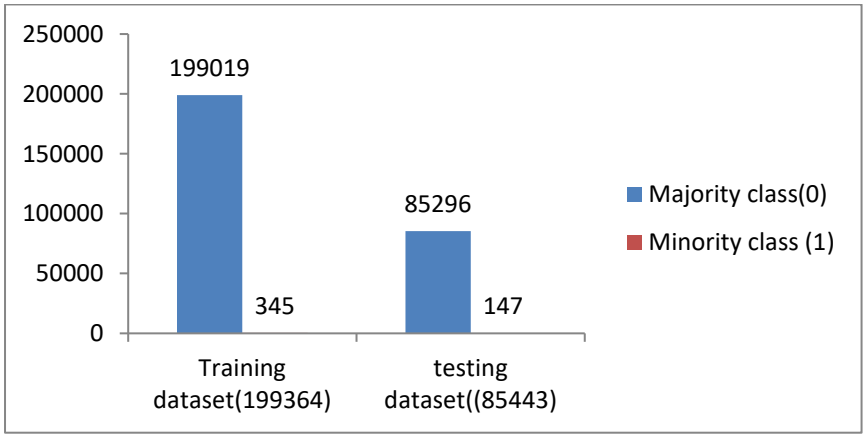

Figure 4: Splitting of the imbalanced dataset into training and testing dataset

3.3 RESAMPLING METHOD

In this step the balancing of the dataset is performed. The proposed technique is applied on the training dataset which is originally imbalanced. The proposed technique is the hybrid of oversampling and Undersampling technique. SMOTE is applied as an informed oversampling technique and NCL is applied as an informed Undersampling technique. The balancing of the number of cases of the learning dataset is done by utilizing the oversampling technique followed by the Undersampling technique.

Following are the methods which are collaborated to give novel approach in balancing the exceedingly imbalanced datasets:

3.3.1 INFORMED UNDERSAMPLING: NEIGHBORHOOD CLEANING RULE (NCL)

When the resampling process is applied on the training datasets, the Neighborhood Cleaning Rule (NCL) [35] deals with the data points of the minority class as well as the data points of the majority class separately. It utilizes the ENN (edited nearest neighbor) method to eliminate the data points of the majority class. For each data points of the training datasets, it identifies three nearest neighbors. If the data point is the part of majority class but it is misclassified by its neighbors as the part of another class, then the chosen data point is eliminated. And similar step is performed with the chosen data points of the minority class. ENN method is applied only on the majority class but the NCL method is applied on both classes.

NCL model keeps up all the data points of the class of interest F and eliminate those from the rest of the training dataset G. This procedure is cultivated in two stages. In the principal stage, ENN is utilized to locate the boisterous data points A_1 in G. Generally, 3-ENN is utilized to eliminate the data points with different class in comparison to the majority class of the three closest neighbors. Accordingly, the neighborhoods are prepared again and the set A_2 is made initially. At secondary stage, the three closest neighbor data points that have a place with G and lead to F data points misclassification is iteratively embedded in the set A_2. At last, the data is decreased by removing the data points that have a place with either sets A_1 or A_2 (i.e., $A_1 \cup A_2$).

A. Algorithm for NCL

1. Split data D into the class of interest F and the rest of data G.

2. Identify noisy data A_1 in G with edited nearest neighbor rule.

3. For each class F_i in G

if ($x \in F_i$ in 3-nearest neighbors of misclassified $y \in F$) and ($|F_i| \geq 0.5 \cdot |F|$)

Then $A_2 = \{x\} \cup A_1$

4. Reduced data $R = D - (A_1 \cup A_2)$

3.3.2 INFORMED OVERSAMPLING: SMOTE

SMOTE "Synthetic Minority Oversampling Technique" was presented by Chawla [36], this strategy produces artificial examples by utilizing the feature space as opposed to data vector. SMOTE is a generalized fabricated oversampling procedure. It means to adjust dissemination of class through arbitrarily expanding minority class models by making the artificial data points. SMOTE assists with conquering the issue of over-fitting and broaden the choice zone of the minority class examples.

The working rule of the SMOTE method is scan for the estimation of k-closest neighbors which are adjoining for each occurrence in the minority class. From that point forward, engineered occasions are made the same number of as the ideal copy rate between the minority information cases and k-closest neighbors which are picked haphazardly. The thought process is to expand the quantity of cases in minority classes. Coming up next is the recipe for the SMOTE technique condition:

$$X_{new} = X + rand(0,1) * (X' - X) \quad (1)$$

The stages of the SMOTE method include the following:

- In first stage, X represents every instance of minority class on the dataset.
- In second stage, searching for k- nearest neighbor by choosing one of the closest k represented as X'.
- The last stage includes linear interpolation between X and $X'$for building new set of instances for minority class.

3.3.3 ARCHITECTURE FOR THE HYBRID METHOD

In figure 5, the process of the hybrid resampling method is shown in detailed manner.

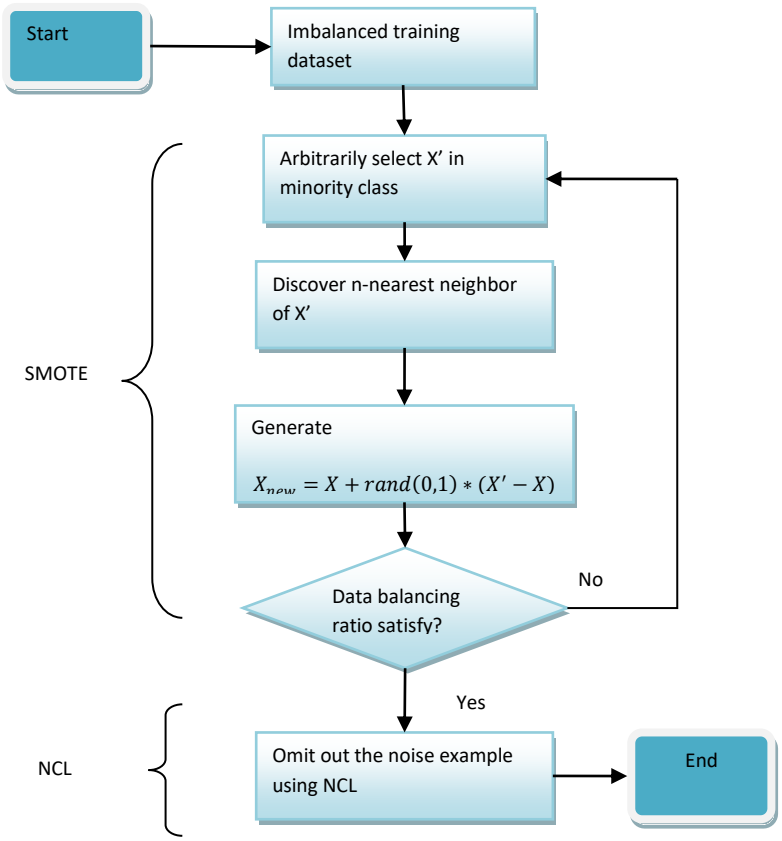

Figure 5: The Flowchart of SMOTE+NCL Algorithm

3.4 CLASSIFICATION MODEL

This is the crucial step because in this step different-different learning algorithms applied on the dataset. These datasets are normalized, reduced and balanced in the previous steps. Learning algorithms learn the patterns of the data points that are present in the dataset. The output is solely depends on the learning algorithm selected for the process. At this step the same learning algorithm can be utilized again and again or different algorithm can also be utilized [37].

In this research, Logistic Regression is utilized as the classification learning model. There are basically two phases for the classification:

- **Learning Step (Training Phase):** In this phase, learning model is trained with the balanced training datasets.

- **Classification Step:** in this phase, the learning model estimates the relations of the data points with the class. This is the testing phase and the testing datasets is utilized on the learning model.

3.4.1 LOGISTIC REGRESSION

In the family of machine learning, the elicited linear models [38] are associated with the most prominent learning models. These kinds of linear model are easy to explain. They can be executed directly does not create any hinder in doing so. Calculation capacity is more required in these kinds of learning model.

On the other hand the results drawn from these models are easily understood. These types of learning model are best suited for the learning of two-class classification problem. In the area of classification, logistic regression the most common learning model is executed. The primary motto of the logistic regression model is to recreate the approximation of the posterior diffusion i.e. $\mathrm{P}(y|X)$. In binary world, this can be further foreshortened by calculating:

$$\mathrm{P}(X): \mathrm{P}(y = 1|X) \tag{2}$$

In the logistic regression it presented as mentioned below:

$$\mathrm{P}\big(X; \beta_0, \widehat{\beta}\,\big) = \frac{e^{\beta_0 + \widehat{\beta^T} X}}{1 + e^{\beta_0 + \widehat{\beta^T} X}} \tag{3}$$

With parameters β_0 and$\widehat{\beta} = \beta_1, \ldots\ldots, \beta_\rho$. The equation can be rewritten as:

$$\log\left(\frac{\mathrm{P}(X)}{1 - \mathrm{P}(X)}\right) = \beta_0 + \widehat{\beta^T} X \quad , \tag{4}$$

Here, it is noticeable that the logistic regression model has log-odds (lhs of equation (4)) which are linear in X.

3.5 EVALUATION METRICS

This is the last step of the experimentation framework. In this step, results that are obtained in the previous steps are evaluated and interpreted by the researcher. Many techniques are available to visualize the obtained results. The researcher must interpret the result once they are visualized. If the results do not fulfill their requirements, they must apply the similar learning algorithm but with different parameters or apply the different learning algorithm to meet the desired results. In this step, the utilization of the results needs to be specified by the researcher [24].

In this evaluation step, the evaluation of the performances between the logistic regression without resampling process, logistic regression with SMOTE and NCL will be compared. Accuracy, Precision, recall, and F1-score will be applied as the evaluation indicator.

In order to grab more knowledge of the learning model other aspects of metrics are also utilized like AUROC score and Average Precision Recall score.

3.5.1 CLASSIFICATION MATRIX

This matrix is well renowned as confusion matrix [24]. In two-class data set, two by two matrix is needed. This kind of metrics tells about the expectation of happening of the testing data points. In this the cases of fraud activity falls under positive category while the non-wrong doing falls under the negative category.

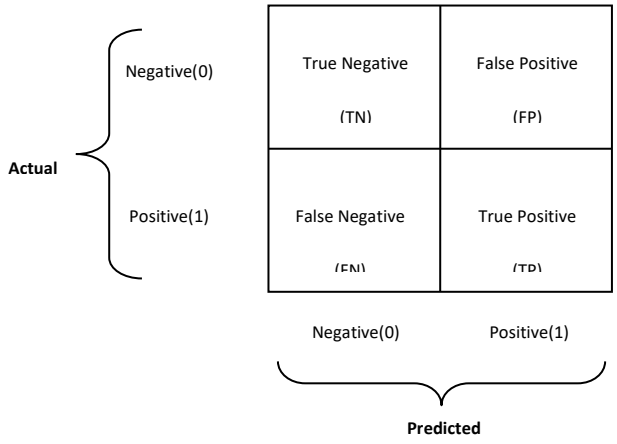

Figure 6: Confusion Matrix for Learning Model

Where,

- **True Positive**: These are the cases of malefactor's activity and accurately tell by the learning model.
- **False positive:** These are the cases of legitimates activity but learning model claims these cases with the malefactor's activity.
- **True negative:** These are the cases of non-fraudulent activity and correctly tell by the learning model.
- **False negative:** These are cases of the non-legitimate activity but learning model confuses them with legitimate.

3.5.2 ACCURACY

It gives the ratio of all cases which is correctly expected by the learning model.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \textbf{(5)}$$

3.5.3 SENSITIVITY (RECALL)

It gives the ratio of real malefactor's activity that is accurately spotted by the learning model.

$$Sensitivity = \frac{\text{True Positive}}{\text{Total Actual Positive}} = \frac{TP}{FN+TP} \quad \textbf{(6)}$$

3.5.4 PRECISION

It gives the ratio of fraud spotted accurately among the pool of expected values.

$$Precision = \frac{\text{True Positive}}{\text{Total Predicted Positves}} = \frac{TP}{TP+FP} \quad \textbf{(7)}$$

3.5.5 F_1MEASURE (F-SCORE)

It can be measured as the harmonic mean of the hit rate and the positive predicted value.

$$F_1 measure = \frac{2*(\text{Precision}*\text{Recall})}{\text{Precision + Recall}} \quad \textbf{(8)}$$

3.5.6 AREA UNDER RECEIVER OPERATING CHARACTERISTIC CURVE [40]

It is known to be one of the efficient metrics for evaluating the performance of the learning model. It shows whether the classification learning model is proficient enough to differentiate among the classes. The graph is sketched between the sensitivity and selectivity. The AUROC can be formulated in python by utilizing the roc_auc_score() function.

3.5.7 AVERAGE PRECISION-RECALL SCORE [41]

This is another important metrics especially when learning model dealing with the binary classification problem. This gives the average of the precisions at different-different thresholds that are similar to the thresholds in the area under the PR curve.

4. RESULTS

4.1 RESAMPLING PROCESS

The hybrid resampling approach is applied on the training dataset before training the learning model. The original data is exceedingly imbalanced and hence the resampling came out as a

solution for this at data-level. The figure 7 shows training datasets without resampling. The figure 8 shows that data is balanced after applying the proposed technique SMOTE+NCL on the training dataset.

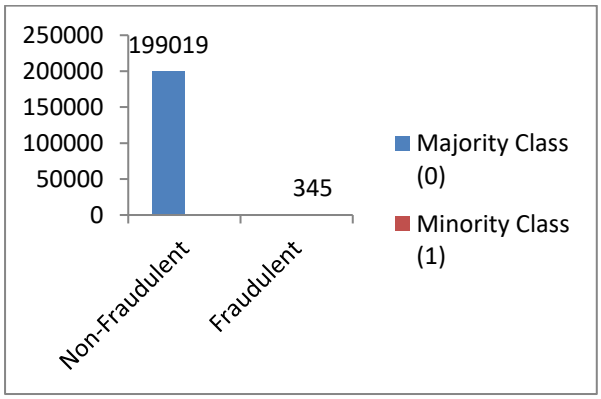

Figure 7: Imbalanced Training Dataset

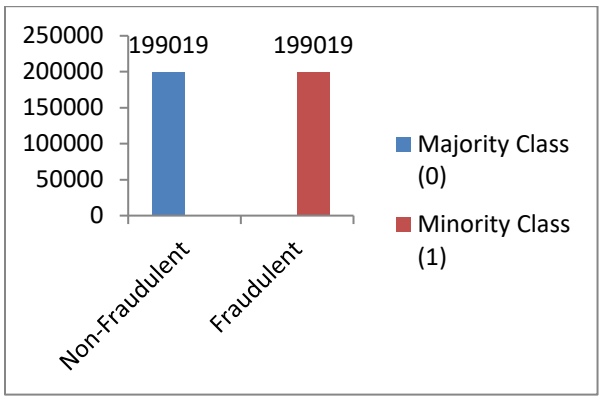

Figure 8: Balanced training dataset using SMOTE+NCL

4.2 CLASSIFICATION AND EVALUATION

For classification the logistic regression model is utilized. The results in the figure 7 shows that the model with SMOTE&NCL performed better than the model without resampling method.

4.2.1 CONFUSION MATRIX

In this the confusion matrix of two methods are compared one is without using resampling method and other one is using smote+ncl method (our proposed). Figure 9 showing confusion matrix for model without no resampling. Figure 10 is showing confusion matrix for model with smote+ncl method. From comparison it can be clearly seen the positive classes (Fraud cases) are better predicted by using our proposed method.

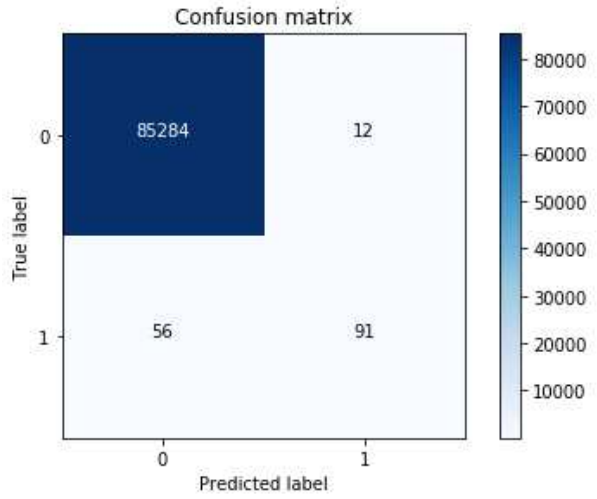

Figure 9: Confusion Matrix for no resampling method

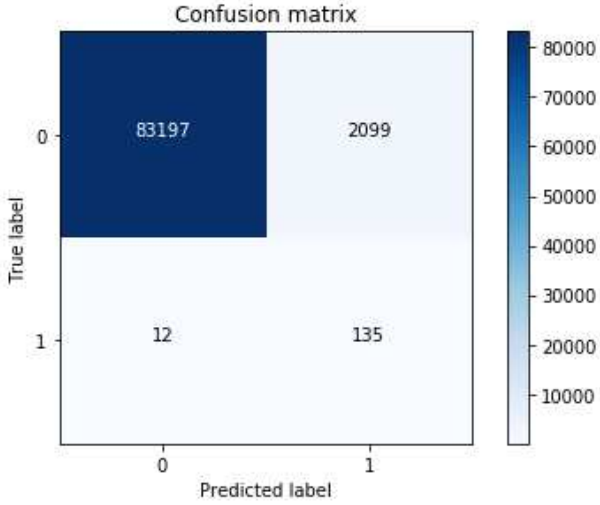

Figure 10: Confusion Matrix for SMOTE+NCL method

4.2.2 EVALUATION METRICS

In this various performance metrics are utilized to compare the methods. The result in the figure 11 shows that the model with smote+ncl performed better than the model without resampling method.

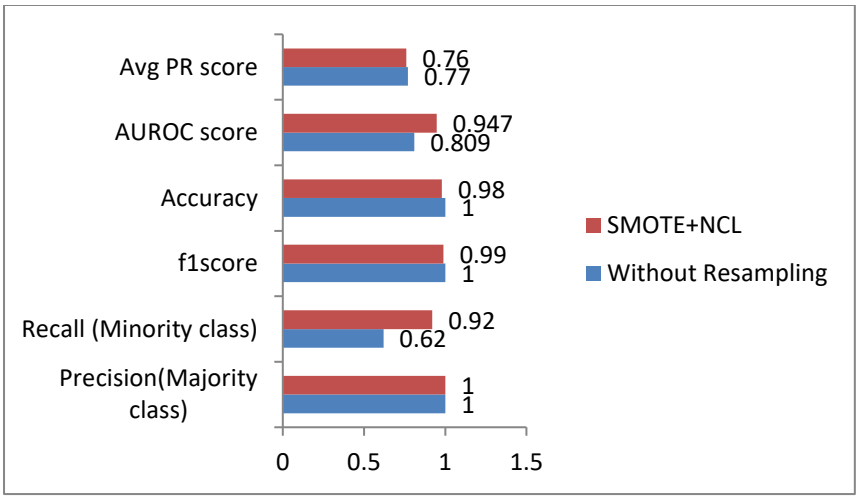

Figure 11: Graph for the comparing the results

4.2.3 AUC-ROC CURVE

In this auc-roc curve is compared between the two methods. In fig auc-roc curve for model with no resampling is shown in fig auc-roc curve for model with smote+ncl is shown. The latter one gives better result on the auc-roc curve.

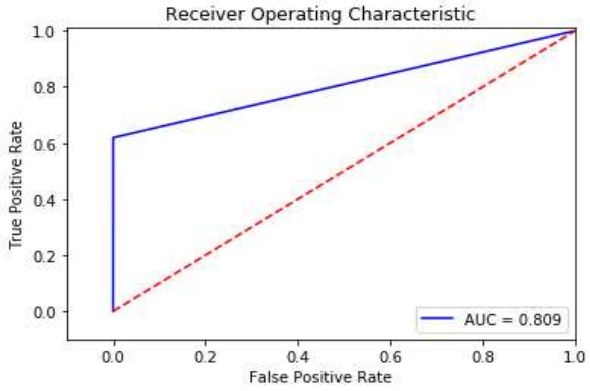

Figure 12: AUC-ROC curve for No Resampling method

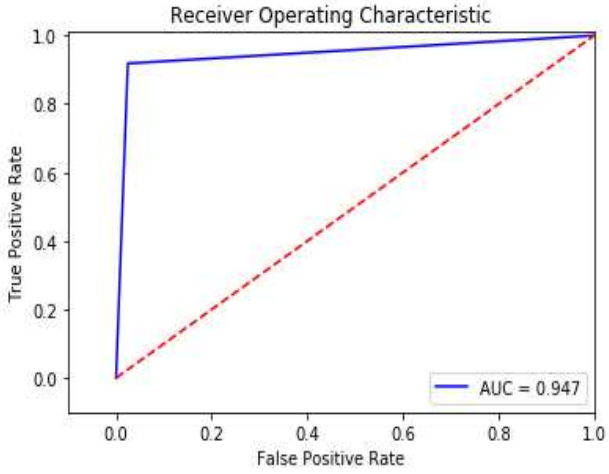

Figure 13: AUC-ROC curve for SMOTE+NCL method

5. CONCLUSION

In this research, imbalanced classification problems and their impact on the learning model has been discussed thoroughly. Previously many solutions have been proposed by researcher to tackle these issues either at the data-level or at the algorithmic-level, but here the main attention is given to the data-level approach. In this research, a novel hybrid resampling method is proposed to balance the original imbalanced classification datasets. This novel method is the hybrid of SMOTE and NCL techniques. For experiment, real-world datasets of credit card transactions have been utilized. For classification, logistic regression is utilized as the learning model. From the results, it can be concluded that the model performed better in the balanced datasets than the imbalanced ones. So, the proposed hybrid technique can be a solution in tackling the imbalanced classification problem in the machine learning process, especially to handle problems on binary class or two-class datasets. For future work, the proposed techniques can be tested on other classification learning algorithms such as SVM (support vector machine), Naïve Bayes and Decision Tree.

DECLARATIONS

Competing interests: The authors declare no competing interests.

REFERENCES

[1] M. Kubat, R. Holte and S. Matwin, Machine Learning for the Detection of Oil Spills in Satellite Radar Images, Machine Learning 30 (1998), 195–215.

[2] T.E. Fawcett and F. Provost, Adaptive Fraud Detection, Data Mining and Knowledge Discovery 3(1) (1997), 291–316.

[3] N. Japkowicz, C. Myers and M. Gluck, A Novelty Detection Approach to Classification, Proceedings of the Fourteenth Joint Conference on Artificial Intelligence, 1995, pp. 518–523.

[4] D. Lewis and J. Catlett, Heterogeneous Uncertainty Sampling for Supervised Learning, Proceedings of the Eleventh International Conference of Machine Learning, 1994, pp. 148–156.

[5] P.M. Murphy and D.W. Aha, UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Information and Computer Science, 1994.

[6] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling TEchnique. Journal of Artificial Intelligence Research, 16:321357, 2002.

[7] Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine Learning, 42, 203-231.

[8] M. V. Joshi, V. Kumar, and R. C. Agarwal. Evaluating boosting algorithms to classify rare cases: comparison and improvements. In First IEEE International Conference on Data Mining, pages 257-264, November 2001.

[9] Available from https://www.kaggle.com/mlg-ulb/creditcardfraud

[10] Gary M Weiss. Mining with rarity: a unifying framework. ACM Sigkdd Explorations Newsletter, 6(1):7–19, 2004.

[11] Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, 2013. Chapter 2.

[12] Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. J. Artif. Intell. Res. 11, 131– 167 (1999)

[13] Samir Al-Stouhi and Chandan K Reddy. Transfer learning for class imbalance problems with inadequate data. Knowledge and information systems, 48(1):201–228, 2016.

[14] Haibo He and Edwardo A Garcia. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9):1263–1284, 2009.

[15] Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, New York (2005)

[16] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Know. Data Eng. 21(9), 1263–1284 (2009)

[17] Sun,Y.,Wong,A.K.C.,Kamel,M.S.:Classificationofimbalanceddata:areview.Int.J.Pattern Recogn. Artif. Intell. 23(4), 687–719 (2009)

[18] Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recogn. 36(3), 849–851 (2003)

[19] Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor. 6(1), 1–6 (2004)

[20] López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. 250, 113–141 (2013)

[21] Chawla, N.V., Cieslak, D.A., Hall, L.O., Joshi, A.: Automatically countering imbalance and its empirical relationship to cost. Data Min. Knowl. Disc. 17(2), 225–252 (2008)

[22] Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20(1), 18–36 (2004)

[23] García, V., Sánchez, J.S., Mollineda, R.A.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowl. Based Syst.25(1), 13–21 (2012)

[24] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor. 6(1), 20–29 (2004)

[25] Available from https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

[26] Marcelo Beckmann et al. A KNN Undersampling Approach for Data Balancing. Journal of Intelligent Learning Systems and Applications, pages: 104-116 (2015)

[27] Nitesh V. Chawla. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook. Pages 853-867.

[28] Pattaramon et al. Neighborhood-based undersampling approach for handling imbalanced and overlapped data. Information Sciences 509 (2019) 47–70.

[29] Haibo He et al. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on: 1-8 June 2008.

[30] György et al. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Applied Soft Computing Journal 83 (2019).

[31] GillesCohen et al. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine. Volume 37, Issue 1, May 2006, Pages 7-18.

[32] Xu Han, Runbang Cui, Yanfei Lan , Yanzhe Kang, Jiang Deng , Ning Jia: A Gaussian mixture model based combined resampling algorithm for classification of imbalanced credit data sets. International Journal of Machine Learning and Cybernetics(2019)

[33] S.Ancy , "Handling Wireless Sensor Network by applying Dynamic Sampling in Surveillance System" Elsevier B.V.(2020)

[34] Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20(1), 18–36 (2004)

[35] J.Laurikkala, Improving identification of difficult small classes by balancing class distribution, in: Conference on Artificial Intelligence in Medicine in Europe, Springer,2001.

[36] N.V.Chawla, et al., SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res.16 (2002)321–357.

[37] García, V., Sánchez, J.S., Mollineda, R.A.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowl. Based Syst.25(1), 13–21 (2012)

[38] Rohith Gandhi. Introduction to machine learning algorithms: Logistic regression, May 2018.

[39] Available from https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

[40] Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)

[41] Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One 10(3), e0118432 (2015)

Figures

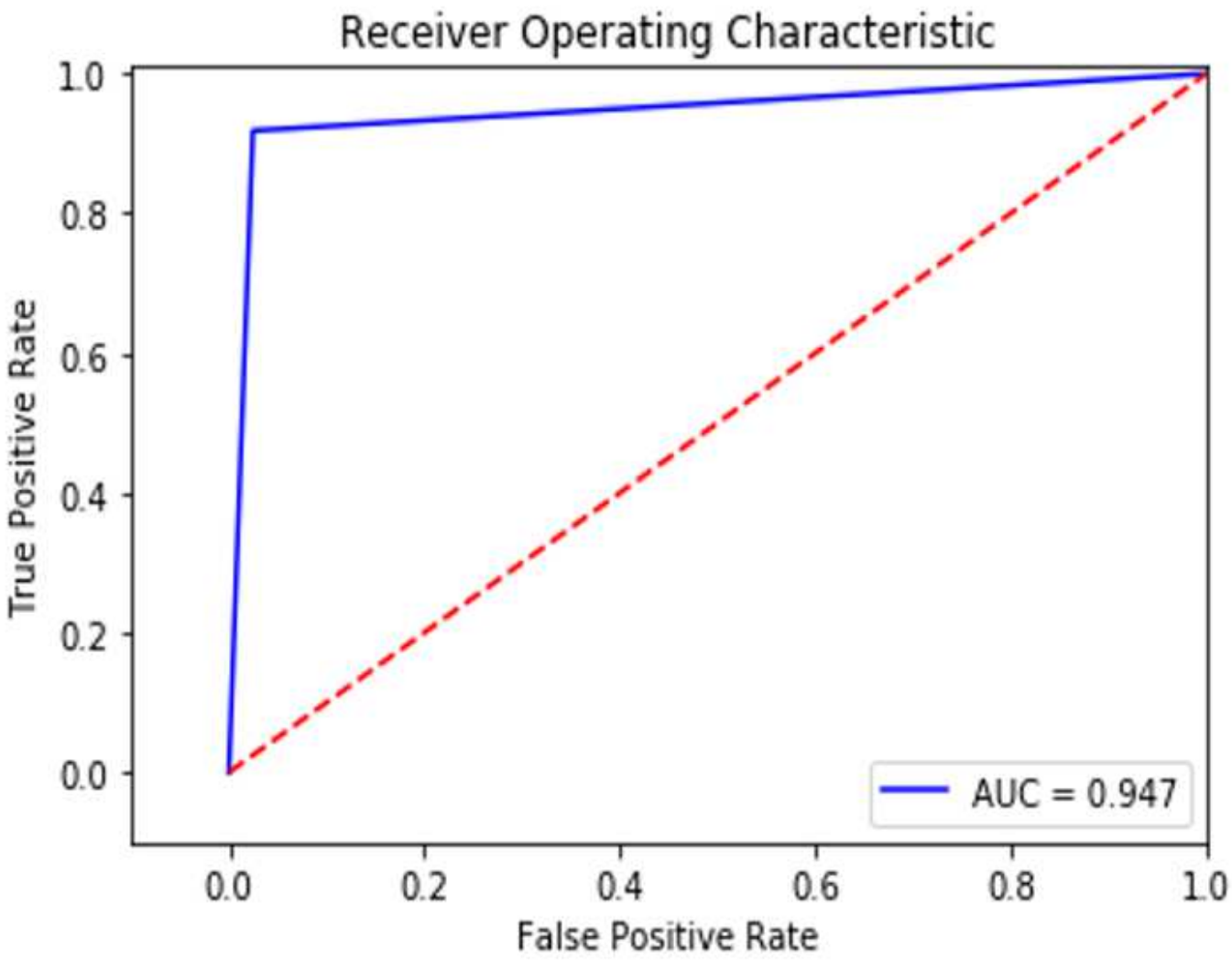

Figure 1

AUC-ROC curve for SMOTE+NCL method

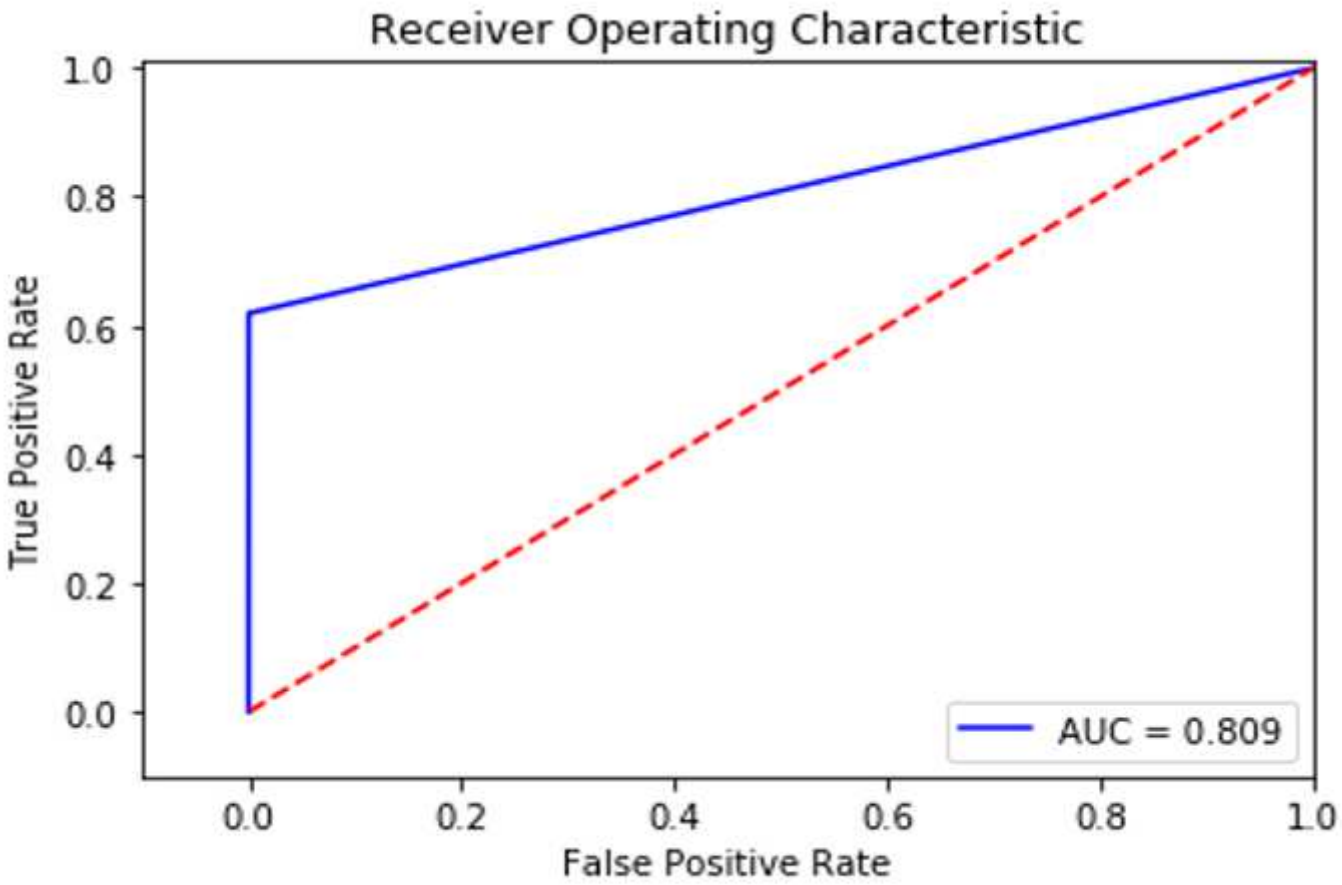

Figure 2

AUC-ROC curve for No Resampling method

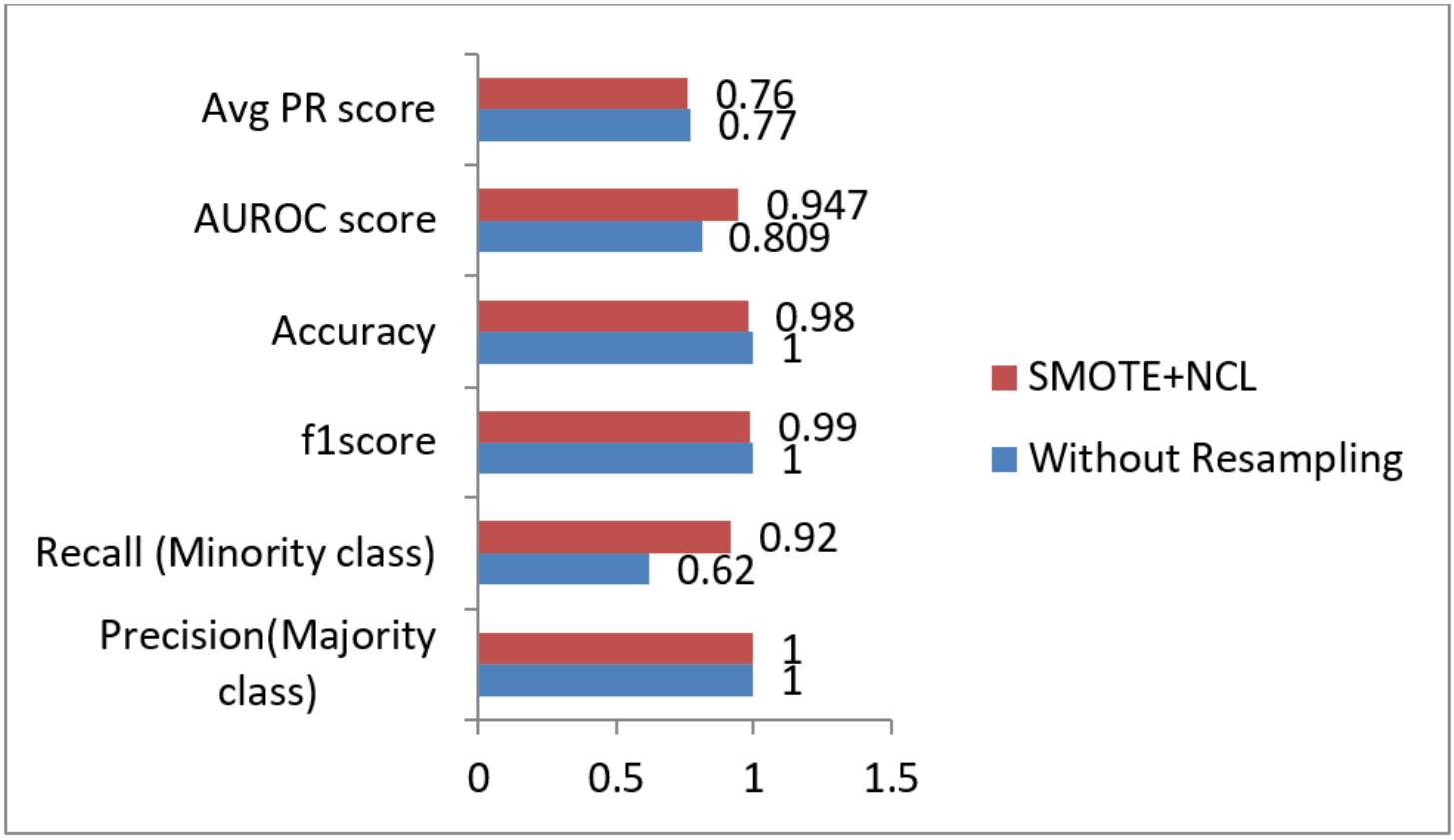

Figure 3

Graph for the comparing the results

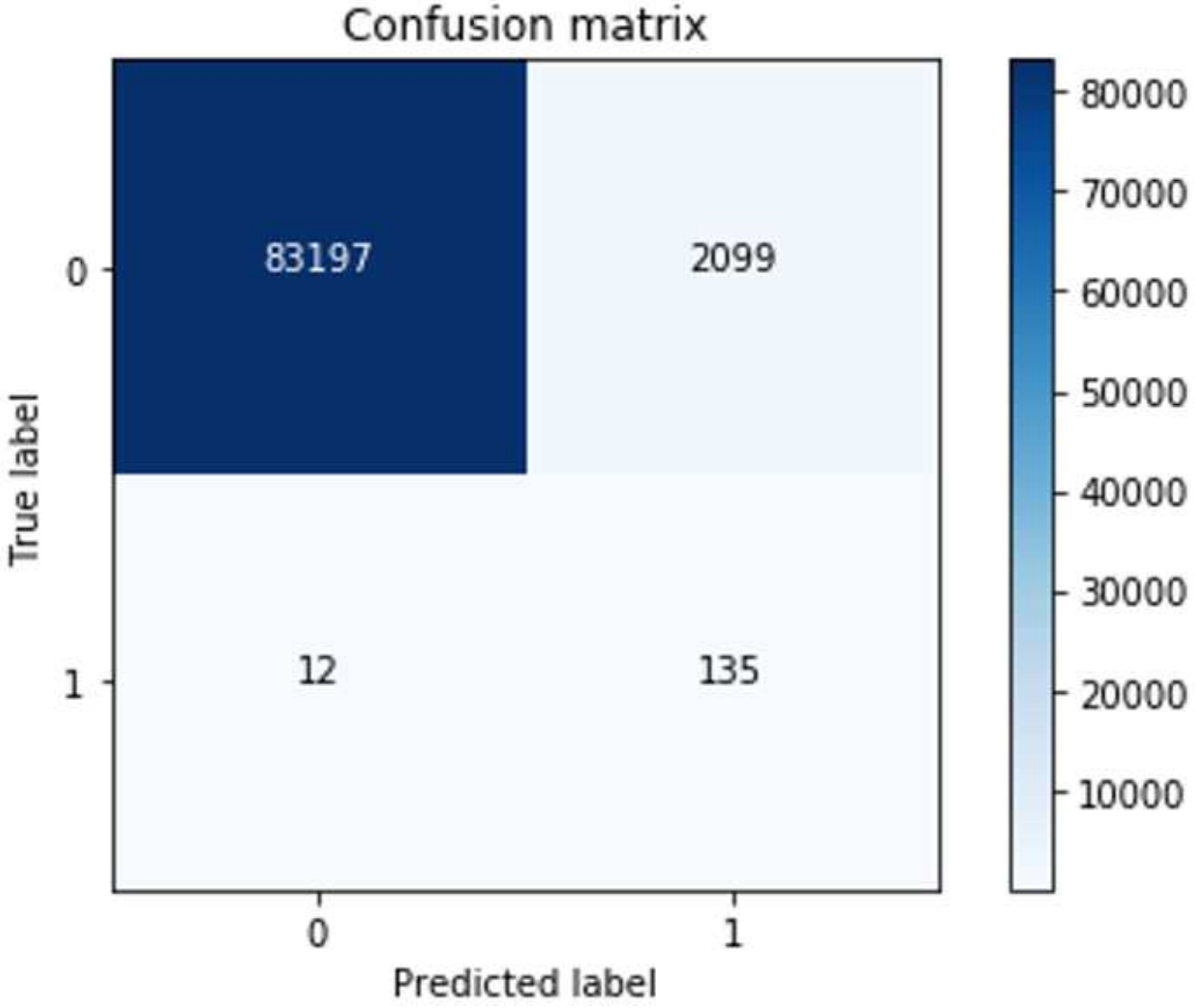

Figure 4

Confusion Matrix for SMOTE+NCL method

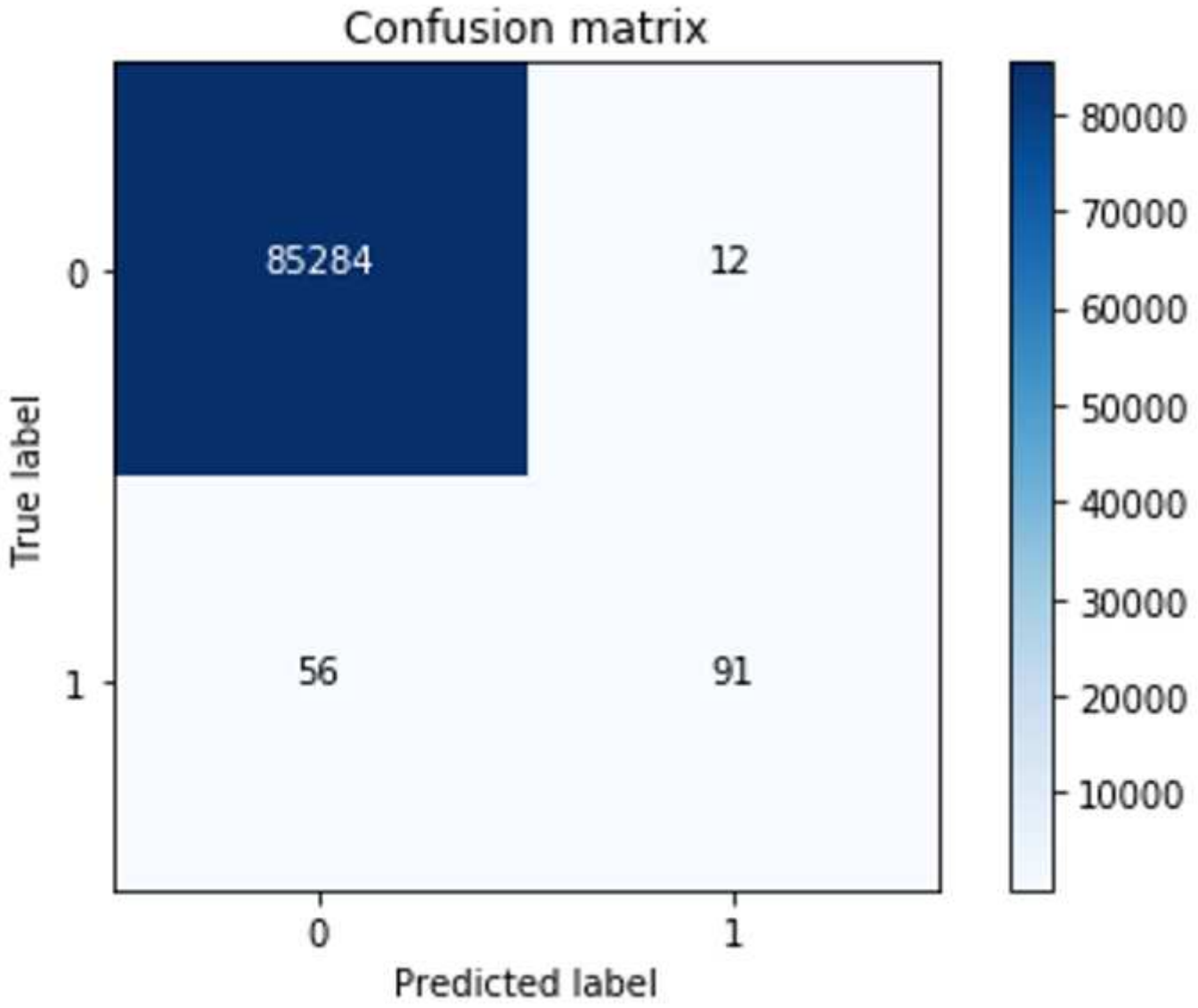

Figure 5

Confusion Matrix for no resampling method

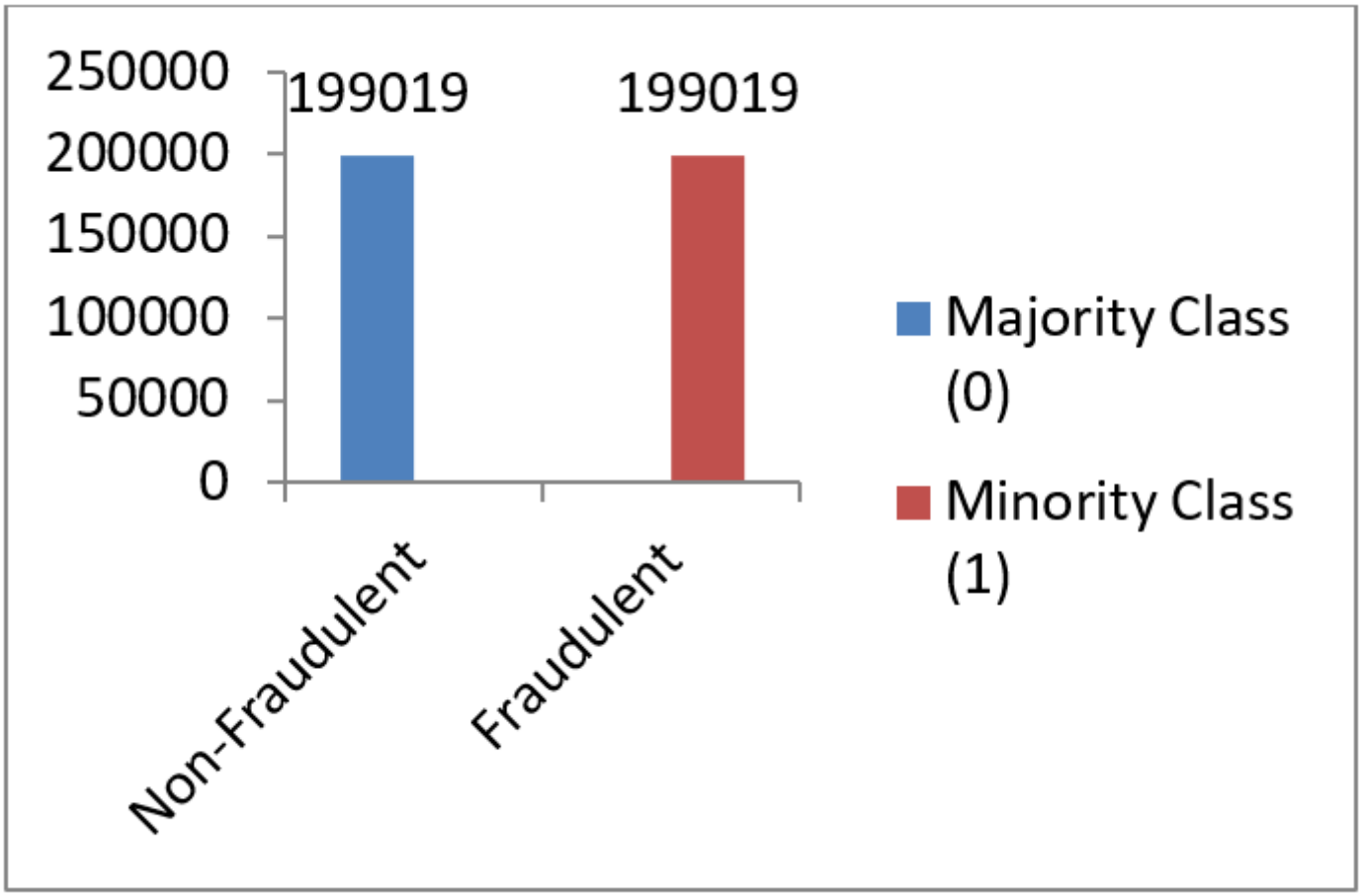

Figure 6

Balanced training dataset using SMOTE+NCL

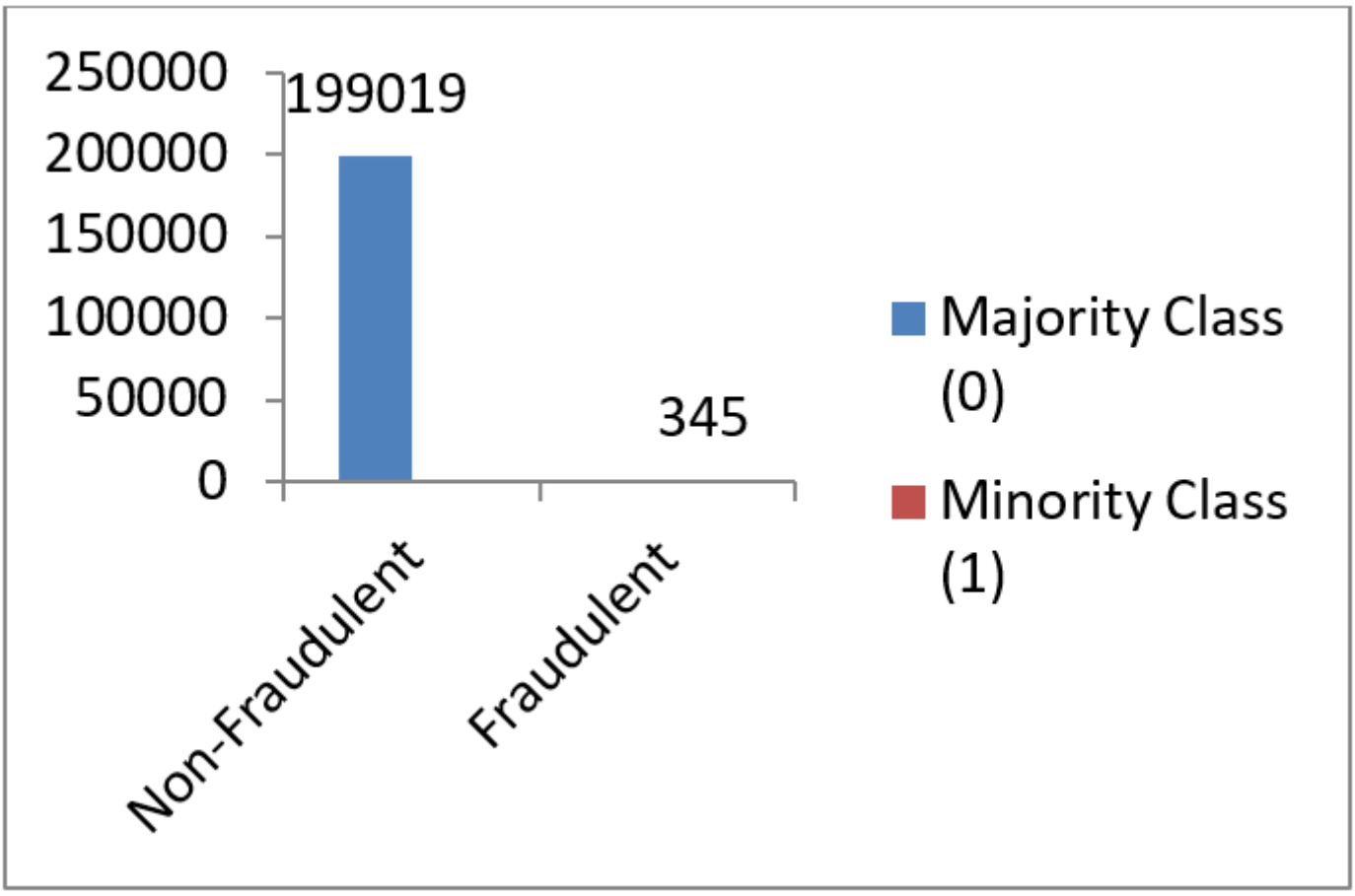

Figure 7

Imbalanced Training Dataset

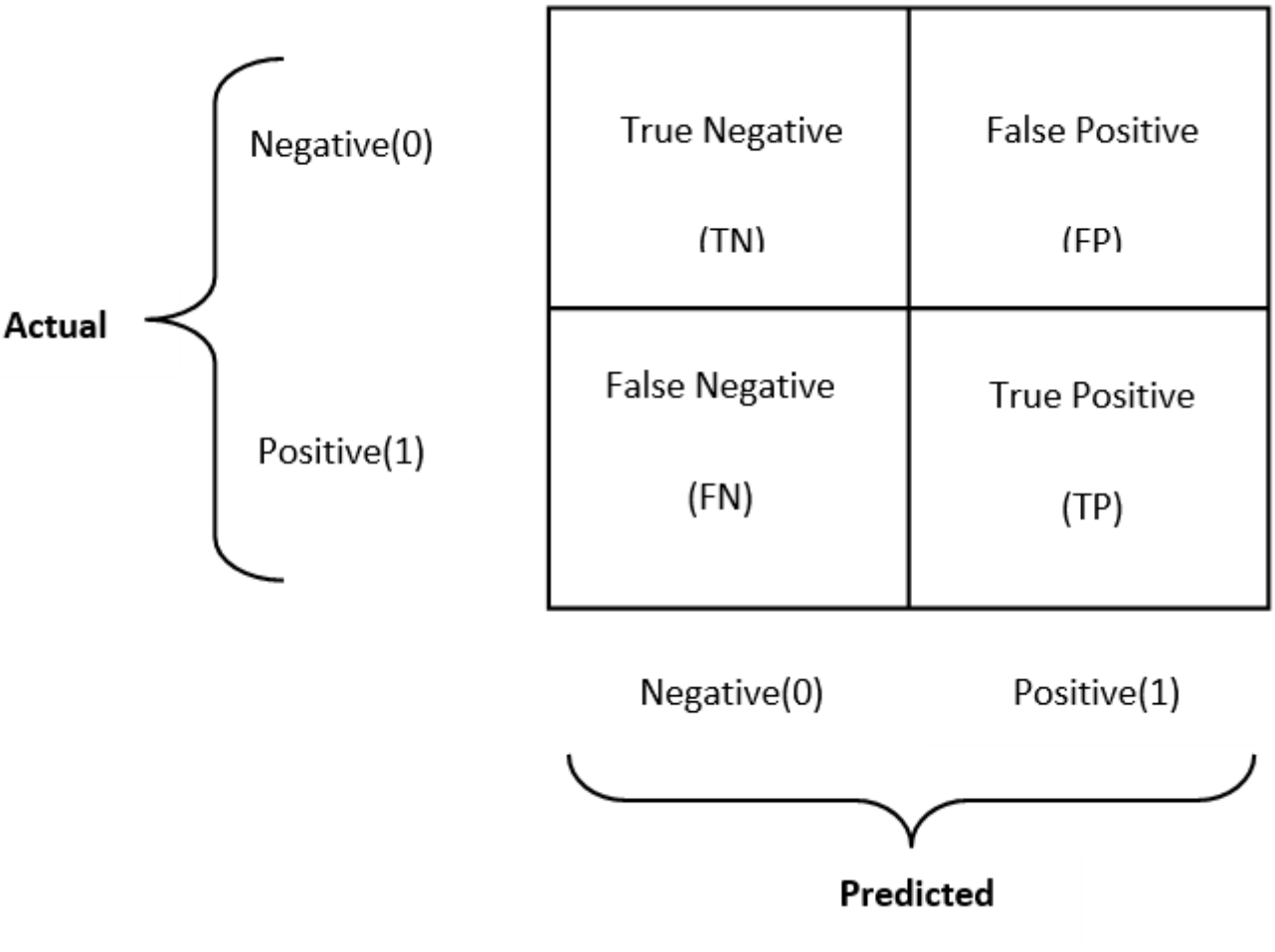

Figure 8

Confusion Matrix for Learning Model

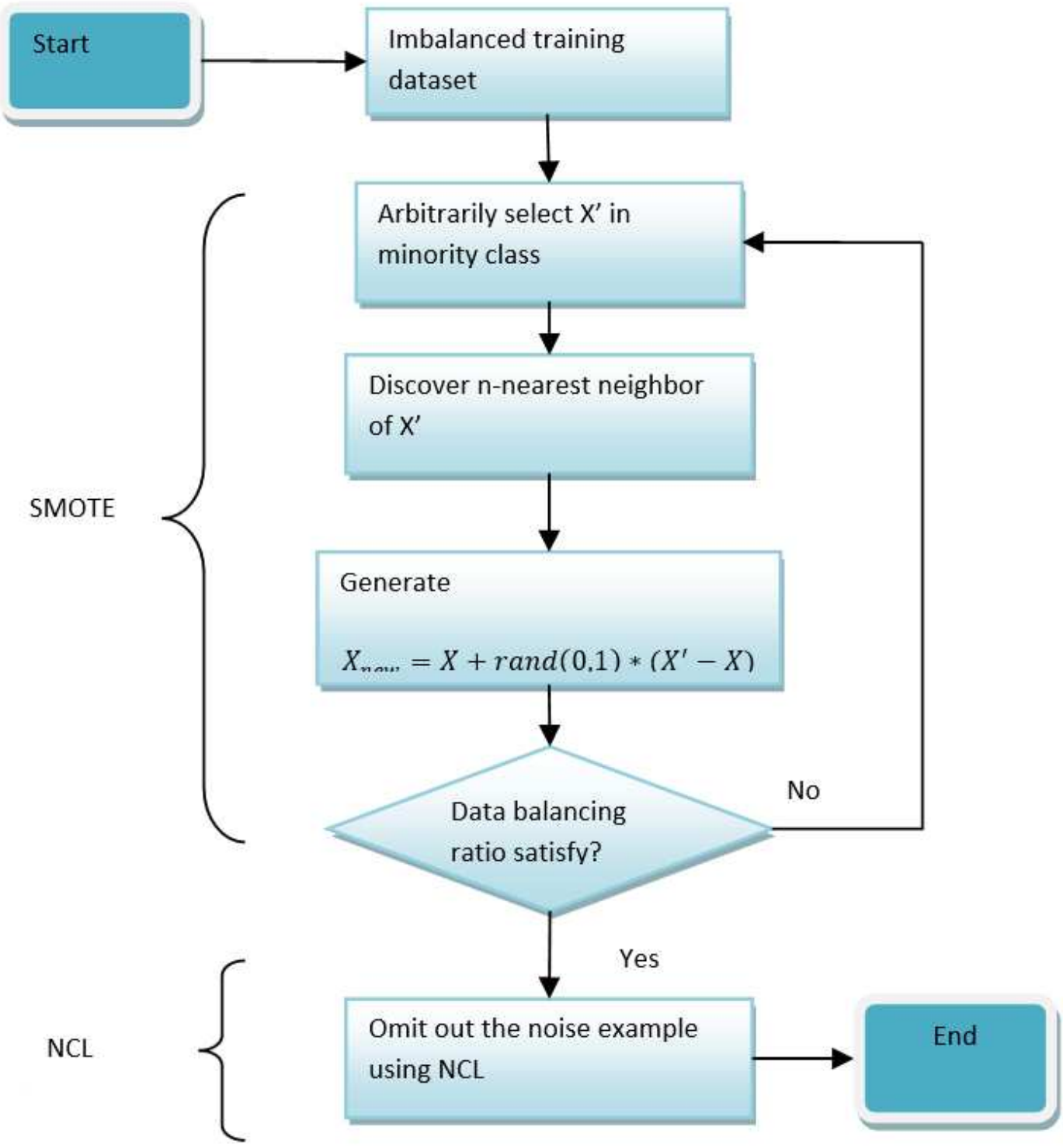

Figure 9

The Flowchart of SMOTE+NCL Algorithm

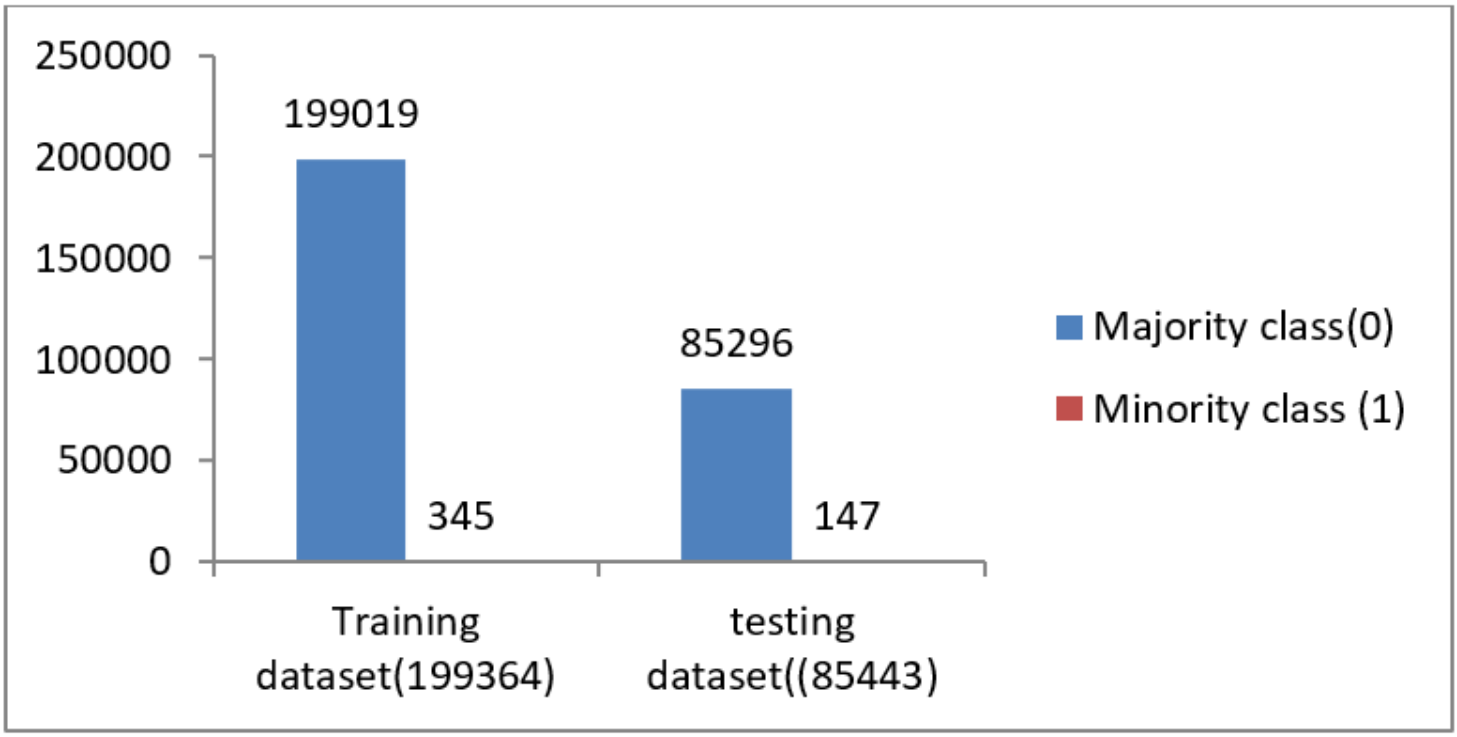

Figure 10

Splitting of the imbalanced dataset into training and testing dataset

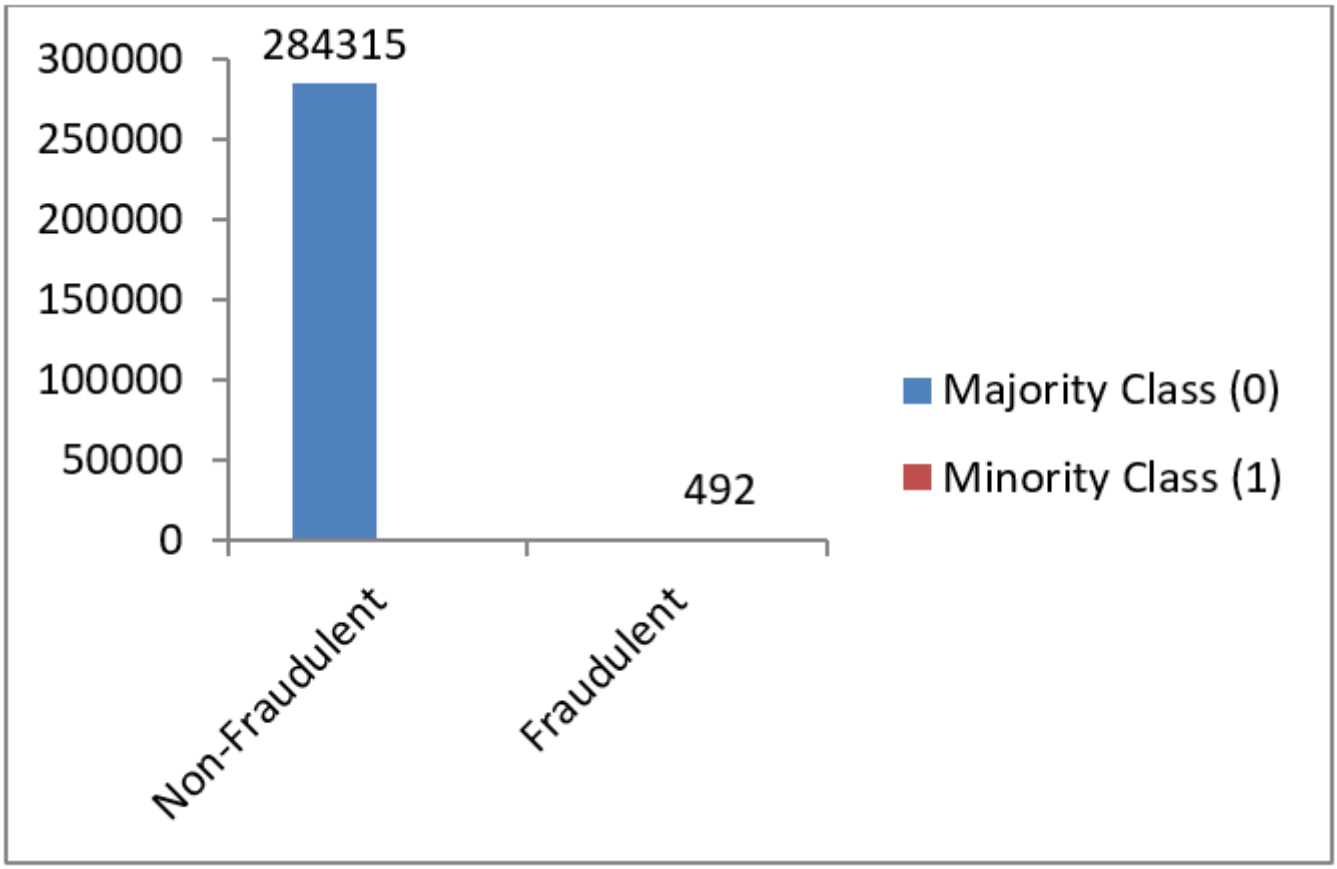

Figure 11

Imbalanced diffusion of data points in the classes

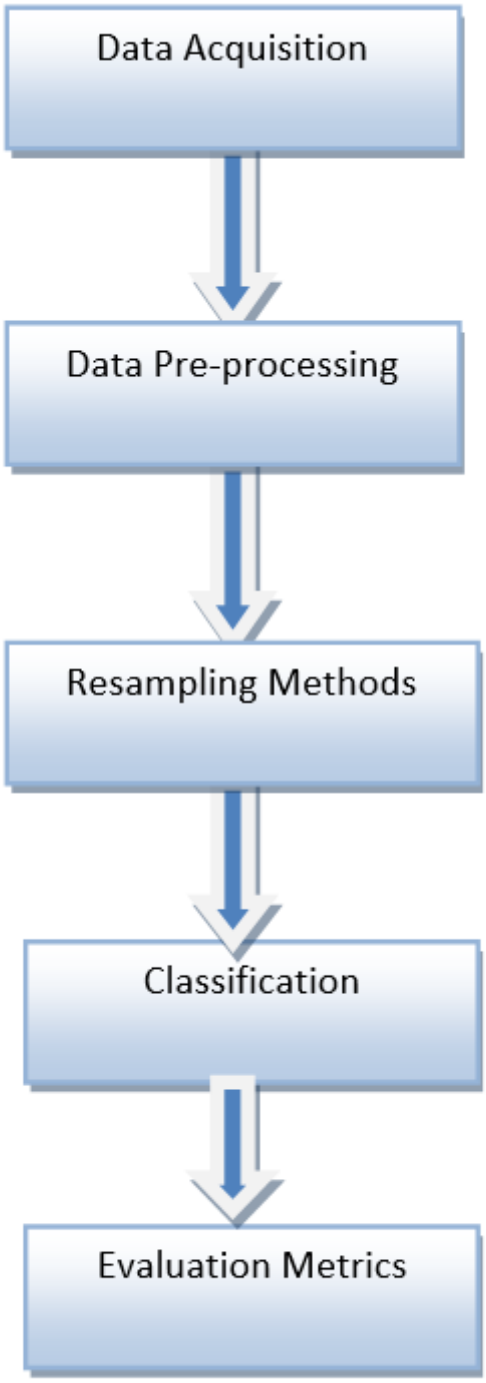

Figure 12

Pipeline of Experiments

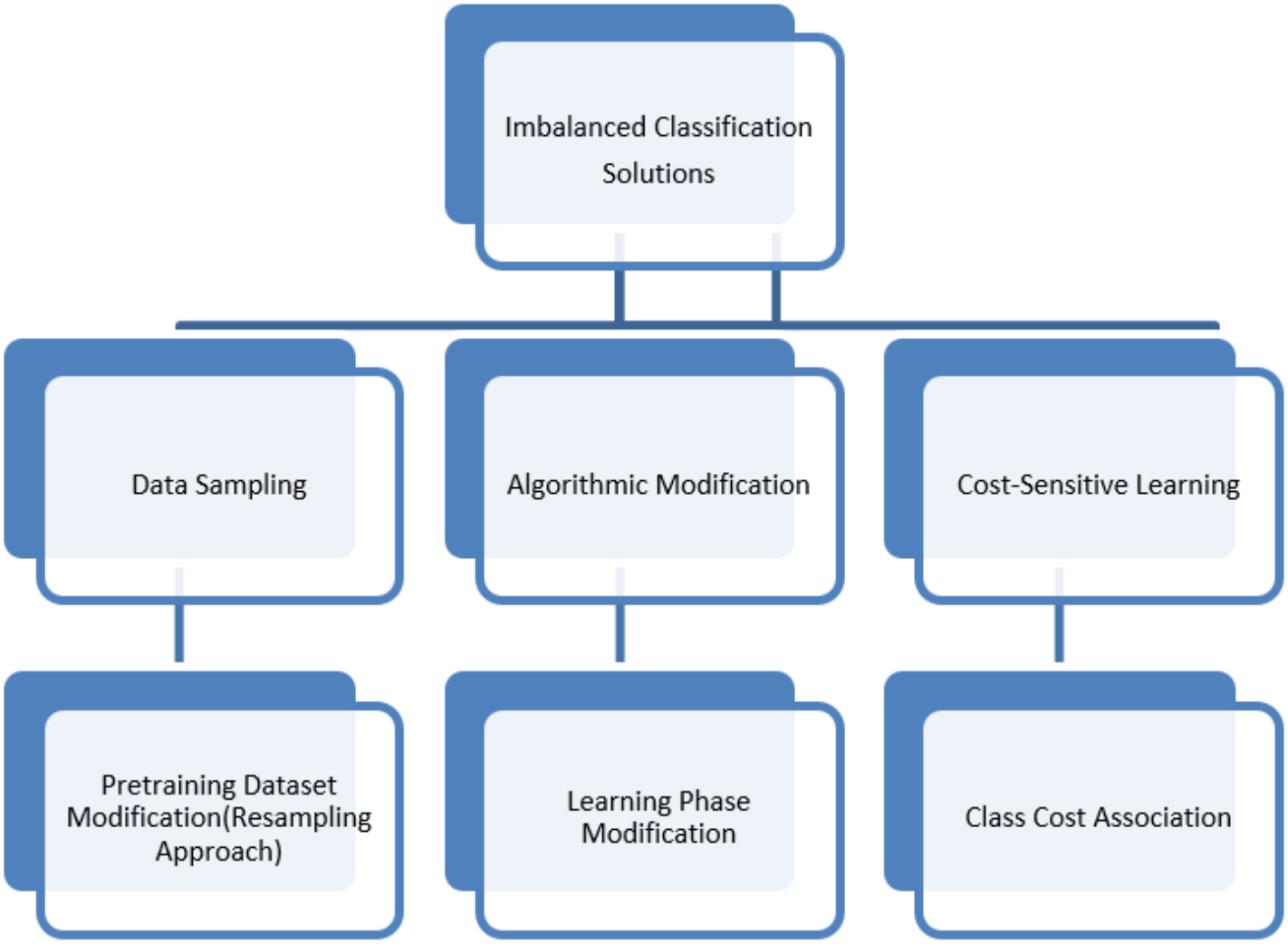

Figure 13

Categorization of solutions to Imbalanced Classification