2020 IEEE Pune Section International Conference (PuneCon)
Vishwakarma Institute of Technology, Pune India. Dec 16-18, 2020

Handling Imbalanced Dataset Classification in Machine Learning

Seema Yadav
Department of Computer Engineering
and Information Technology
Veermata Jijabai Technological Institute, Matunga
Mumbai, India
ssyadav_p17@ce.vjti.ac.in

Girish P. Bhole
Department of Computer Engineering
and Information Technology
Veermata Jijabai Technological Institute, Matunga
Mumbai, India
gpbhole@ce.vjti.ac.in

2020 IEEE Pune Section International Conference (PuneCon) | 978-1-7281-9600-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/PuneCon50868.2020.9362471

***Abstract*— Real world dataset consists of normal instances with lesser percentage of interesting or abnormal instances. The cost of misclassifying an abnormal instance as normal instance is very high. The majority class is normal class whereas minority class is the abnormal one. Researchers in data mining and machine learning are looking out numerous strategies to resolve issues associated with dataset that is unbalanced and also the challenges featured in way of life. Irregular distribution in the dataset is the motive behind declining performance of classifier. There are mainly two methods, algorithm based and data level based, the utmost widespread methodology associated to the current is hybrid method. The task of decision making and overall classification accuracy is affected due to bias for majority class. Ensemble technique is an effective technique. The objective of study is providing background related to imbalance class issues, way out to confront the disputes and challenges in studying unbalanced data. In support to experimental result accompanied on one of the dataset, ensemble technique in adjacent to different strategies of data-level offers improved outcomes. The fusion of techniques is going to be advantageous for several applications in real-life like intrusion detection, medical diagnosis, software defect prediction, etc.**

Keywords*— Classifier, Imbalance class, Minority and majority class, Bias, Sampling, Machine learning.

I. INTRODUCTION

Classification is based on predictive modelling that involves allocating class labels to individual observations. A predicted class is generated from the classification model. Each example consists of class label and observations. The observation is the input and accompanying class label is the output. The number of classes is fixed while describing the problem. The predicted class generated from the classification model is in the form of discrete category, prediction is required for making decision. Alternately we can choose to estimate the likelihood of class membership in place of class label. There are problems due to skewed distribution of classes in machine learning field and data mining. Additional examples belong to one class and less number of examples in other is the class imbalance problem. Extra examples are there in majority class [1] in contrary to fewer number of examples related to minor class. In number of applications the most interesting and essential class is the minority category. The cause behind the growth of imbalance problem is due to uneven class distribution. Lesser amount of patterns existing in a class are known as exceptional events. The society is affected due to this. In daily life rare events are not frequently found and prediction tasks is affected. The performance of model is deteriorated in mining from the skewed domains, ordinarily in predicting minority class instance. The imbalance problem in class is significant with in the field of data mining. Classifiers fabricated using data mining techniques are very useful. Thus the activity of taking decision is easy for the managers, still classification of unbalanced data is a challenge for normal classification models. The misclassification of examples associated with minor class costs more compared to major class classification [2]. Several real-life applications face class imbalance problem for instance in medical analysis of the unusual disease patients suffering from unusual disease is fewer, bioinformatics, frauds in transaction related to credit card, fraud phone calls, fraudulent claim of insurance, detection of network intrusion, grading cancer malignancy, biomedical [3] etc.

A. Problem Statement

Applications in real life are suffering due to imbalance problem. The imbalanced data is still a challenge for normal classification models. The traditional classification models are designed assuming that there are same number of samples for each class. The result is poor prediction performance and decision making task precisely for the minority category. The main problem is with minority category as it is more significant and there is possibility of error in classifying minority category instances. By means of machine learning and data mining algorithm classifier model can be built that guides the decision making activity. The purpose of designing the classifier model is to avoid misclassification of minority class instances and improve the performance of classifier using the different approaches used to handle imbalanced data. The ensemble or hybrid approach is required for performance improvement of classifier.

II. LITERATURE REVIEW

The number of example that fits in each class is called as class distribution. Disproportion occurs at that time if one class has extra examples and other has fewer examples in the training dataset. For instance, if we are collecting measurement of flowers, 80 samples of flowers belong to one species and 20 samples of flowers belong to second type of species, only this samples are included in our training dataset. This signifies the imbalanced classification problem. The imbalance classes in a dataset can be represented in terms of a ratio. For instance, in imbalance binary classification problem with an imbalance ratio of 1 to 100 (1:100), means for each one example in one category there are 100 examples with in the alternative class.

A. Related Work

1) Classification of Binary Imbalanced Data

The enlightened branch well-thought-out for learning imbalance category downside is binary classification. In numerous applications binary unbalanced classification is being presented in real-life, like sick additionally as healthy patients in drugs, legitimate as well as illegitimate actions in safety of computer. Structure defines the connection between majority class and minority class. There are several direct methods such as: -Analysis of classes, extremely imbalanced classes, Classifier's output adjustment, learning imbalance for ensemble to balance the category distribution.

2) Learning Class Imbalance

More attention is paid by the research community of machine learning in learning class imbalance problem. From the observation it is found that the class proportion differs. Classification of customers, imbalance in credit scoring, grading cancer malignancy are few examples of class imbalance problem. The bias for major class is the problem associated with class imbalance learning. In case of cancer malignancy grading more concern is for detection of minority class in comparison with majority class.

3) ELM (Extreme Learning Machine)

There are conventional ELM [6] in which all the samples are considered equally important due to which the accuracy predicted is biased for major class. In order to beat this limitation several variations of ELM are suggested like weighted and class specific cost-regulation ELM etc. for effective management of imbalance class problem. CS-ELM (class specific ELM) is a variation of weighted ELM and it handles the imbalance problem more efficiently and is less complex as compared to weighted ELM. In weighted ELM, weights are assigned to the training instances and is not required in (CS-ELM) class specific ELM. In CCR-ELM, regularization parameters specific to class are used. While computing the specific class regularized parameters, the class distribution is considered. In CCR-ELM during computation of regularized parameter class overlap and class distribution is not considered.

4) Deep Learning with Class Imbalance

According to the survey it is found that study of class imbalance problem is completed systematically using traditional machine learning model and most of the research is focused on computer vision with CNN in addition to this big data property is not considered [7]. Inspite of recent developments in deep learning and its acceptance very little work is done in the zone of deep learning with class balance. The traditional methods for class imbalance like cost-sensitive learning and data sampling can be appropriate in deep learning but advanced method where neural network feature learning are used shows good results. While doing survey most of the areas such as data complexity, performance interpretation, architecture testing, big data applications, ease of use and generalization to various domains were engrossed.

5) Categories of Various Domains of Class Imbalance

There are four main categories of domains based on the study carried out as shown in Fig.1, which is again categories into nine. To handle imbalance class problem 18 different approaches are recommended in past. More than one approach is used in some techniques to tackle the problem. [1]

Class imbalance domain categories are as follows:

Following 18 approaches are used by proposed technique:

1. Random principle
2. Geometric mean
3. Bagging
4. Clustering
5. Noise Filter
6. Boosting
7. Genetics
8. Genetics
9. Fuzzy Logic
10. Fuzzy rule base
11. Nearest Neighbor
12. Rough sets
13. Rotation network
14. Neural Network
15. Kernel function
16. Immune network
17. Principal component analysis
18. Support vector machine
19. Greedy divide and conquer

B. Reasons for The Imbalance Class Problem

The imbalance in distribution of samples associated to class in classification is based on predictive modelling. There are two foremost groups of causes generally considered for the imbalance, sampling data and domain properties. While collecting the examples from the problem domain the imbalance occurs. This may embody biases led during data collection, and error created throughout data collection.

1) Biased sampling 2) Measurement error

Mistakes occurs during collection of observations. One sort of error is applying the incorrect category label to several examples. Alternately it happens that the processes or systems from that examples were collected might be broken or impaired to cause the imbalance. Improved sampling technique is used in cases were the imbalance is caused by a sampling bias or measure error and therefore the imbalance is corrected. This is due to unfair representation of training dataset. The imbalance may well be the property of problem domain.

C. Methods to Tackle Imbalanced Data

The methods generally employed for learning class imbalanced data are: [1][4].

a) Algorithm level approach. b) Data level methods.
c) Hybrid method and d) Feature based approach. Fig 1. represents the categories of class imbalance domain.

- *Data level method:* To balance the distribution modification is done in group of examples. By adding or deleting difficult samples balancing can be done. Modification in training set is done in standard learning algorithm. Generation of novel instances and adding them to minority class is referred as oversampling whereas removing the instances from the majority class is under-sampling.

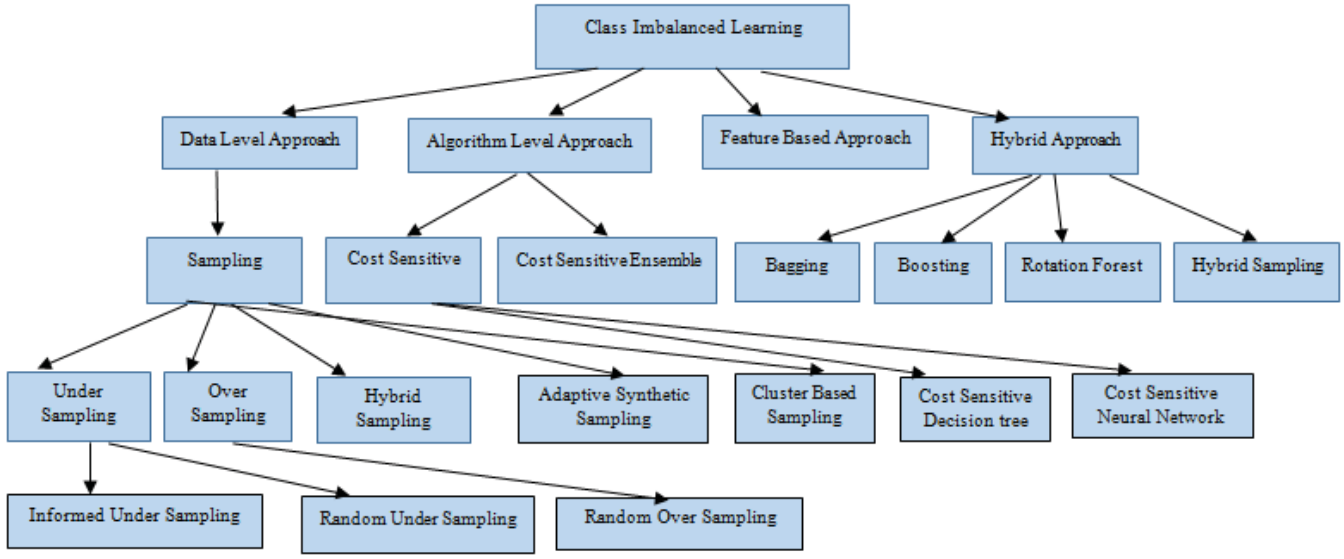

Fig. 1. Categories of class imbalance domain [5]

- *Algorithm-level approach*: Algorithm that already exists is modified, more inclination is towards instances appropriate for majority class. Clear thought is essential for modifying algorithm for learning. Most prevalent method is cost-sensitive approach. Cost of misclassification is considered in cost learning techniques; high cost is allocated for misclassification of minority class. Given learner is charged more that comprises of unpredictable fine for individual collection of samples that are concerned. As a result, high cost is allocated to objects less in number. The objective of this method is to minimize the cost related to misclassification of minority class samples.
- *Hybrid methods*: Data and algorithm level advantages are integrated in hybrid method. Focus is on integration of those methods to extract their robust points and cut back their flaws. We tend to get a proficient and strong learner as a consequence of integration of data level solutions and classifier ensembles Hybridization of sampling and cost-sensitive learning is proposed in some applications.
- *Feature selection method:* The aim of feature selection method is to select most optimal subset of features from all available features.

D. Class Imbalance Problem in Real-world Applications

Various real world applications, faces the matter of irregular data distribution. There's plenty of development despite the fact learning an imbalanced dataset. Vital problems are preventing mischievous attacks, detecting severe diseases, managing Facebook incomparable activities and Twitter, and the social networks to regulate uncommon conditions while monitoring the systems.

Real-world applications with unbalanced data are:

1)Video mining 2) Text mining 3) Sentiment analysis 4) System monitoring in industry 5) Behavior analysis 6) Cancer malignancy grading etc. 7) Prediction of software defects.

III. METRICS USED FOR ASSESSMENT OF IMBALANCED CLASS DATASET

Metrics used for assessment is meant to be a severe problem. Valuation Metrics may be a pointer for the performance measure of algorithms employed in machine learning [6]. Accuracy, besides error rate, are the standard metrics used for assessment. Since the accuracy is biased towards the category having additional samples no matter the category having fewer samples, it is not correct to use it to handle class imbalance problem and the performance worsens. From the confusion matrix, we can develop metrics public for the two-class problem as presented in Table I. Assessment metrics commonly related with class imbalance [4] are recall, sensitivity, precision, geometric mean(g-mean), specificity, F-measure as shown in Table II. The classification performance of every class are often monitored by means of sensitivity and specificity. Table I. shows the confusion matrix used for classification.

TABLE I. CONFUSION MATRIX USED FOR CLASSIFICATION

		EXPECTED CLASS	
		+ve	**-ve**
ACTUAL CLASS	**+ve**	*tp*	*fn*
	-ve	*fp*	*tn*

tp – true positive; fp - false positive; tn – true negative;
fn- false negative; sp -specificity; se -sensitivity;
A – accuracy; P – precision; R – recall;

The assessment metrics associated with class imbalance problem are as shown below in Table II.

TABLE II. ASSESMENT METRICS RELATED WITH CLASS IMBALANCE

Sr. No	Metrics	Formula
1.	Sensitivity	$se = \frac{tp}{fp + tp}$
2.	Specificity	$sp = \frac{tn}{tn + fp}$
3.	Accuracy	$A = \frac{tn + tp}{tn + tp + fp + fn}$
4.	Precision	$P = \frac{tp}{tp + fp}$
5.	Recall	$R = \frac{tn}{fp + tn}$
6.	F-measure	$F\text{-}measure = \frac{(1+\beta^2)*(P*R)}{\beta^2 * R + P}$
		$F\text{-}measure = \frac{2*P*S}{S+P}$
7.	G-mean	$G\text{-}Mean = \sqrt{tprate + tnrate}$
8.	AUC(Area under curve)	$AUC = \frac{tprate + tnrate}{2}$
9.	Total cost	$Total\ cost = (fn.\ C_{fn}) + (fp.\ C_{fp})$

AUC (Area under curve) and ROC (Receiver Operating Characteristic) curve are the most standard measures for class imbalanced data. Illustration of ROC Curve may be completed by graph plot tp Vs fp on the coordinate axes. By means of a receiver operational curve, the performance of classifier is summarized and visualized. Cost curve and cost matrix is used in cost-sensitive matrix. Cost matrix is employed to seek out the value related to the examples that are classified C (i, j) describes price related to examples classified as class j in place of class i. Total cost may be calculated as shown in Table II.

IV. PROSPECTS AND CHALLENGES

Number of opportunities and challenges to deal with imbalance data are as follows [2]

- Classifying instances for multiple label and multiple instance imbalanced dataset,
- Learning in Imbalanced Big data,
- Imbalanced data stream learning,
- Imbalanced Regression,
- Unsupervised and semi-supervised learning from data that is imbalanced,
- Multi class imbalanced classification.

V. EXPERIMENTAL SETUP AND RESULTS

i. Dataset- Selected dataset for experimentation is telecom customer churn prediction from kaggle [13].

ii. Dataset Objective - The dataset objective is labeling voluntary and involuntary churn customer.

iii. Data set description– Customer attrition, is additionally, referred to as customer churn, turnover of customer, loss of customer or client. Telecomm firms, net facility provider and insurance firms typically use client attrition examination and rate of attrition as their significant business metrics since charge of holding current customer is much but exploit a replacement one. Companies typically makes a distinction between involuntary churn and voluntary churn. Customer decision to switch to another company or service provider is voluntary churn and involuntary churn happens because of situations like transfer of customer to an extended period safe keeping facility, death, transfer to a remote place. An analyst target voluntary churn because of customer relationship mechanism of company. The dataset contains 7043 samples or instances as well as 21 attributes.

Predictor variables: The number of predictor variables are 20. The variables are as follows:

Customer ID, Gender, Senior Citizen, Partner, Dependents, tenure, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges

Response variable: Class – churn and not churn. There are in total 7043 observations out of which 5174 customers are normal customer and 1869 customers are churn. The proportion is 2.77:1.

iv) Result- Output: ['churn', 'not churn']

v) Method –

Steps:

1) Application of RUS (Random Under Sampling) for dataset balancing –RUS is a random sampling method, that is applied to balance the dataset. In this method the examples from the majority class are randomly selected and deleted from the training dataset. The instances from the majority class are discarded randomly until a more balanced distribution is reached.
2) Multilayer Perceptron, KNeighbour, RFC, Logistic Regression, and SVM classifier models are applied on the dataset – Different classifier model as mentioned above are applied on the sampled dataset to evaluate and analyze their performance based on the metrics accuracy, precision, recall and ROC curve.

vi) Results

Fig. 2 shows the class distribution.

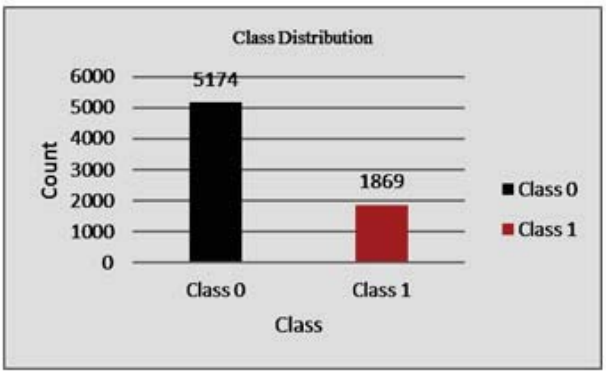

Fig. 2. Class distribution

From Fig.2 we can see that Majority category count and Minority category count are 5174 and 1869.

Table III. shows the metrics used for evaluating classifiers i.e. accuracy, precision, recall and ROC score used to determine the performance of 6 classifiers.

TABLE III. METRICS USED FOR EVALUATING CLASSIFIERS

Classifiers	Accuracy	Precision	Recall	ROC Score
Logistic Regression	0.769	0.747	0.812	0.727
Random Forest	0.853	0.826	0.882	0.886
KNeighbour	0.739	0.717	0.785	0.764
MLP	0.718	0.723	0.715	0.731
Linear SVC	0.571	0.553	0.585	0.584
Naïve Bayes	0.749	0.815	0.831	0.866

Fig.3. shows the accuracy measurement for classifiers.

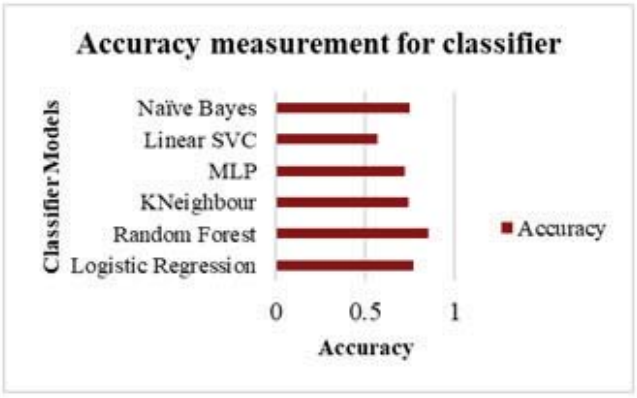

Fig. 3. Accuracy measurement for classifier

Fig.4. shows the precision measurement for classifier.

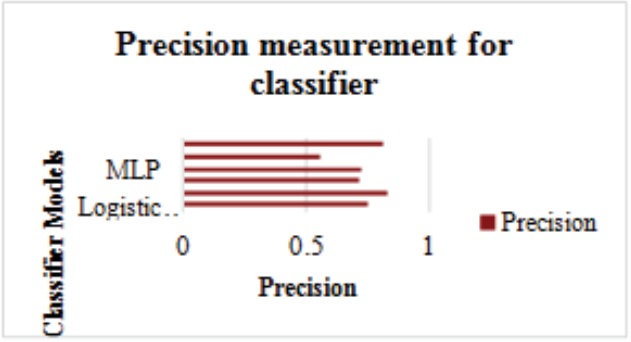

Fig. 4. Precision measurement for classifier

From Table III., Fig.3., Fig.4., Fig.5 and Fig.6. it is determined that the performance of Random Forest is good in terms of Recall Accuracy, Precision and ROC-AUC score in comparison to other models for classification.

Fig. 5 shows the recall measurement for classifiers.

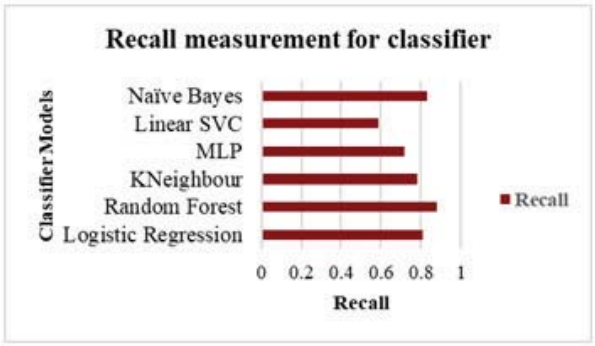

Fig. 4. Recall measurement for classifier

Fig.6 shows the receiver operating characteristics curve.

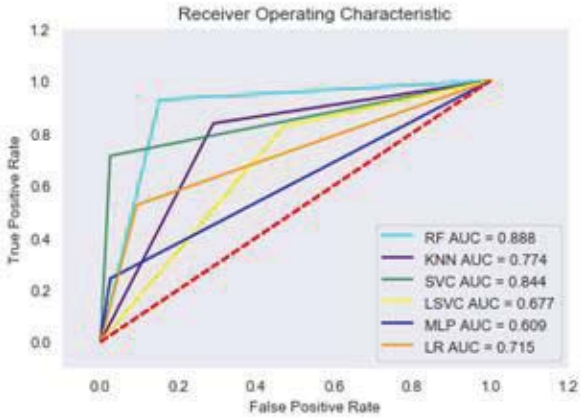

Fig. 5. Receiver operating characteristic for classifier

VI. CONCLUSION

The class imbalance problem is summarized in this paper, solutions for handling it, results based on experiment and challenges in research related to applications in real-life besides different traits of imbalanced learning is discussed. The experiment is performed on telecom customer churn prediction dataset. The performance of Random Forest is good in terms of Recall, Accuracy, Precision and ROC-AUC score in comparison to other models for classification. Despite most of the research work on learning imbalanced data still there are glitches which needs to be considered and novel techniques should be developed or the method that exists, modification should be done for overcoming the limitations.

Some of the directions that can be well-thought-out for resolving problems related to imbalance data:

1) Importance should be given to nature and structure of instances associated with minority class for clear understanding of the problems.
2) Advanced methods for learning multiple-class imbalanced data.

3) New responses are envisioned for learning multiple- instance and multiple-label.
4) Introduction of new efficient clustering algorithms for uneven distribution of object groups.
5) Development of new methods for deep analysis of properties of rare samples in consideration with imbalanced regression.

REFERENCES

[1] Guo Haixiang, Li Yijing Jennifer Shang, Gu Mingyun Huang Yuanyue, Gong Bing. "Learning from class imbalanced data: Review of methods and applications". Elsevier Journal Expert Systems with Applications.Vol 73. pp.220-239. January 2017.

[2] Bartosz krawcyk. "Learning from imbalanced data: Open challenges and future directions". Springer Review. 2016.

[3] Krawczyk, B., Galar, M., Jelen, Herrera,F. "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy". Application Soft Computing. Vol.38, pp.714–726. 2016.

[4] Shaza M. Abd Elrahman, Ajith Abraham. "A Review of Class Imbalance Problem", Dynamic Publisher, Journal of Network and Innovating Computing, Vol.1. pp.332-340. 2013.

[5] Seema Yadav, Girish P. Bhole. "Learning from imbalanced data in classification". International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-VIII, Issue-V. 2020.

[6] Bhagat Singh Raghuwanshi , Sanyam Shukla ."Class-specific extreme learning machine for handling binary class imbalance problem". Elsevier Journal of Neural Network.Vol 105 pp.206-217, 2018

[7] Justin M. Johnson, Taghi M. Khoshgoftaar . "Survey on deep learning with class imbalance". Survey Peper Journal of Big Data. 2019.

[8] Buda, Mateusz and Maki, Atsuto and Mazurowski, Maciej."A systematic study of the class imbalance, problem in convolutional neural networks", Elsevier Journal Neural Networks, Vol.106, pp.249- 259. 2018.

[9] Ronaldo C. Prati, Gustava E.A.P.A. Batista, Maria Carolina Monard "Data Mining with Imbalanced Class Distribution: Concepts and Methods",4th International Conference on Artificial Intelligence .2019.

[10] Thammasiri, Dech and Hengpraprohm, Supoj and Hengpraprohm, Kairung and Mukviboonchai, Suvimol."Imbalance Classification Model for Churn Prediction", Elsevier Journal Advanced Science Letters, Vol.24. pp.1348-1351. 2018.

[11] Lin, Wei-Chao and Tsai, Chih-Fong and Hu, Ya-Han and Jhang Jing-Shang. "Clustering-based undersampling in class, imbalanced data", Elsevier Journal Information Sciences, Vol.409, pp.17-26 .2017

[12] Jian, Chuanxia and Gao, Jian and Ao, Yinhui ."A new sampling method for classifying imbalanced data based on support vector machine ensemble", Elsevier Journal Neurocomputing, Vol.193, pp.115–122 .2016.

[13] Kaggle.com, 'Telecom Churn Prediction', 2018. [Online]. Available: https://www.kaggle.com/bandiatindra/telecom-churn-prediction. [Accessed: 30- Sep - 2020].