Remittances Review
May 2024,
Volume: 9, No: S 2, pp. 660-681
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)
Received: 05 March 2024, Accepted: 25 April 2024
DOI: https://doi.org/10.5281/zenodo.13283854

Ensemble Learning based Optimized Random over Sampling for Handling Class Imbalance of Customer Churn Prediction in Telecommunication

[1]Sabahat Tasneem, [2]Dr. Muhammad Younas, [3]Dr. Qasim Shafiq, [4]Dr. Muhammad Murad Khan, [5]Dr. Uzma Jamil

[1]PhD Scholar, Department of Computer Science, Government College University, Faisalabad, Pakistan.

[2]Assistant Professor, Department of Computer Science, Government College University, Faisalabad, Pakistan.

[3]Associate Professor/HoD, Department of English Language and Literature, The University of Faisalabad, Pakistan.

[4]Assistant Professor, Department of Computer Science, Government College University Faisalabad, Faisalabad, Pakistan

[5]Assistant Professor, Department of Computer Science, Government College University Faisalabad, Faisalabad, Pakistan

Abstract

This is the era of big data, which is more than a couple of decades old. In Telecommunication, it spawned a new requirement to dig deep into the imbalanced data of customers at risk of churn to gain deep insights, so a company may become able to turn the data into dollars by retaining their existing customers. But even after more than two decades of Machine Learning (ML) Model development through Big Data (BD) analytics have been passed on, the issue of class imbalance in data is still intensely grasping the focus of research. In this paper, a novice Optimized Random Oversampling Technique (OROT) has been presented to handle the data class imbalance issue and to improve the accuracy of Ensemble Learning (EL) model. The experiment has been conducted over the Cell2Cell, an open sourced, dataset with 58 features and 51047 instances. It is concluded that Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) can significantly contribute in Telecommunication for Customer Churn Prediction and Retention (CCPR) by addressing the issue of data class imbalance.

Keywords: Ensemble Learning (EL), Big Data (BD), Machine Learning (ML)

1. Introduction

Telecommunication is being generated overwhelmed and massive data volume of customer activities and its business transactions, which can be utilized to generate high valued information for proactively customer, churn prediction. So, it is indispensable to make sense of the exponentially growing Big Data, being generated by telecommunication, in millions of Tera Byte (TB) size, in real time [1, 2]. But unfortunately, a severe issue of class imbalance in Big Data can result into biased prediction towards majority class [3].

In imbalanced dataset of telecommunication, majority class becomes authority over minority class and results into low accuracy rate of a customer churn prediction by demolishing the identification of minority class [4]. This problem usually occurs due to disproportionate ratio of observations in each data class. This generates hurdles for classifiers and results into biased predictions [5-7]. That's why the reliability of classifier's prediction becomes uncertain as the ratio of imbalanced data observations increases [8, 9]. Thus it is required to intensely focus towards this issue before proceeding to Machine Learning (ML) model development. A lot of research work has been done in this regard, which can be categorized into following two major groups such as Algorithm Level models based research and Data Level Models based research [10, 11]. The Fig 2 reveals the major categories of data sampling techniques capturing the focus of researchers. In data sampling techniques, the data can be resampled randomly or can be done through an algorithmic approach. During the process of oversampling technique minority class instances are resampled by generating their replicas randomly or through an efficient algorithm [12]. On the other side, in under-sampling techniques, the observations belongs to a majority class of a particular data set are get removed randomly [13]. Fig 1 reveals the conceptual working of data under-sampling and data oversampling techniques. While feature selection techniques can select the most relevant and influential features to enhance performance of a classifier and contribute to handle class imbalance issue [14-16].

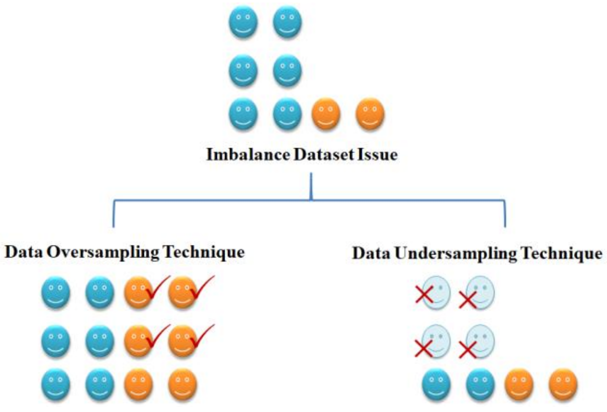

Fig 1: Working of Data Oversampling Techniques and Data Under-sampling Techniques

The Algorithm Level models-based research can be further categorized into Cost-Sensitive Methods and Hybrid / Ensemble Methods. The Cost-Sensitive Methods, a sub field of Machine Learning (ML), calculates the prediction error's cost during training a Machine Learning (ML) model. Ensemble or Hybrid techniques multiple classifiers are get combined and make predictions through Stacking, Bagging or Boosting [17-19]. Genetic programming (GP) can be used to tackle classification, optimization and feature selection [20].

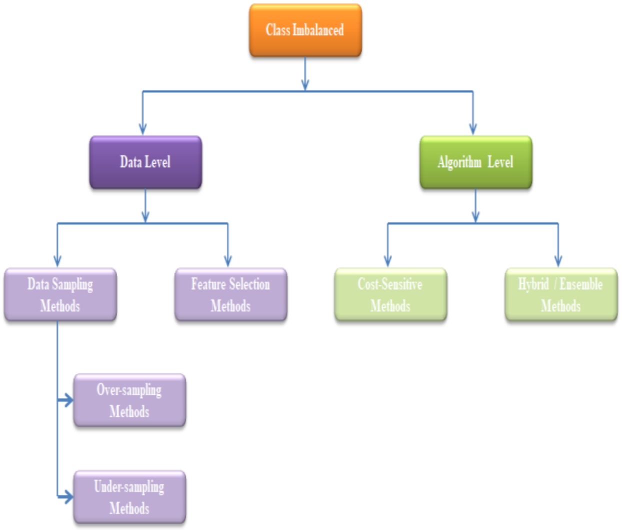

Fig 2: Categories of Class imbalance handling techniques

In this research, an Optimized Random Oversampling Technique (OROT) have been devised by combining with Ensemble Learning to tackle the issue of imbalanced dataset and to improve the performance accuracy of customer churn prediction model in telecommunication. This strategy is quite novice and can achieve the highest accuracy than its benchmark studies. Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) is full name of the devised methodology. In this research the sensational effects of Data Oversampling Technique (DOT) and Particle Swarm Optimization (PSO) have been combined with Ensemble Learning (EL) technique to extract the sensational results of this fusion. In Ensemble Learning, Random Forest (RF), AdaBoost (AB) and Decision Tree (DT) have been used as weak classifiers and stacking to make final prediction. The Model was evaluated by using state-of-the-art evaluation measurements, such as Area Under Curve (AUC) and accuracy. The devised model was simple and applied over an open-sourced dataset 'Cell2Cell' with 58 features and 51047 instances.

2. Review of literature

The review of literature in this research paper has been discussed with respect to two categories such as Data Level research and Algorithmic Level research.

Uzair et al. [21] did research on algorithm level to tackle imbalance and presented a Transfer Learning (TL) model for customer churn prediction by combining the Convolutional Neural Network (CNN) Architecture of Deep Learning (DL) with Ensemble Learning (EL) of Machine Learning. The model was applied over **Orange** with 230 features and 50000 instances and **Cell2Cell** with 77 features and 40000 instances. The model Scored 75.4% and 68.2% accuracy on Orange and Cell2Cell dataset respectively. The recorded Area Under Curve (AUC) was 0.83 and 0.74 on Orange and Cell2Cell dataset respectively. Adnan et al. [22] did research on both level of imbalance data handling research. For algorithmic solution, he presented another Ensemble Learning (EL) model based upon Random Forest (RF), Rotation Forest (RotF), RotBoost (RotB) and Support Vector Machine (SVM). For data level solution, he applied mRMR in Filter Feature Selection and Particle Swarm Optimization-based Undersampling, Genetic Algorithm in Wrapper methods. The proposed model was applied over the aforementioned datasets to evaluate the performance. The recorded Area Under curve (AUC) of **Orange** dataset was 0.85 and 0.82 of **Cell2Cell** dataset.

Devi et al. [23] did research in the field of medical. He described that the fusion of Back Propagation Neural Network (BPNN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Naïve Bayes (NB) with Tomek-linked and Redundancy based Under-sampling. The proposed model scored 0.99 Area Under Curve (AUC) over Diabetes dataset and 0.80 Area Under Curve (AUC) over Breast Cancer dataset. Ngurah et al. [24] combined the effects of Deep Neural Network (DNN) with Feature Scaling Technique and achieved 80.0% accuracy over IBM dataset with 21 features and 7043 instances. Over the aforementioned dataset Lalwani et al. [25] applied Gravitational Search Algorithm (GSA) for feature selection and AdaBoost (AB) with Extreme gradient Boosting (XGBoost). The proposed model achieved the accuracy of 84%. Pamina et al. [26] revealed that Extreme Gradient Boosting (XGB) can achieve 0.798 accuracy over the same IBM dataset when combining with Univariate Analysis and Feature Correlation techniques.

Shatnwai et al. [27] recorded highest 84% F-Score over Orange datast with 20 features and 3333 instances by applying Gradient Boost Tree (GBT) with Random Oversampler (RO), ADASYN, SMOTE and Borderline SMOTE. Caigny et al. [28] presented a Hybrid model of Decision Tree (DT), Logistic Regression (LogR) and Logit Leaf Model (LLM). The proposed model achieved 1.786 Area Under Curve (AUC) by applying Decision Rules for Segmentation of Customer over the Cell2Cell dataset with 70 features and 40000 instances. Halibas et al. [29] applied Exploratory Data Analysis (EDA), Binning (Discretization), Correlation Matrix Operator (CMO) and Pearson Correlation Coefficient (PCC) with Logistic Regression (LogR) over the IBM dataset with 21 features and 7043 instances. While IRFAN ULLAH et al. [30] scored highest 89% accuracy over two datasets by applying Random Forest (RF) classifier with Correlation Attribute Ranking Filter and Information Gain for feature selection.

Table 1 gives a bird eye view of the whole review of literature. In standalone models, it is observed that Random Forest could achieve the highest accuracy with Correlation Attribute ranking filter and Information Gain. While in hybrid methodology Extreme Gradient Boosting won the race by achieving 85% accuracy. It is also observed that hybrid methodology is being popular due to its capability to combine the effects of multiple methodologies for data level and algorithmic level-based solution for Imbalanced dataset.

Reference	**Algorithm Level Research Work**	**Data Level Research Work**	**Data set Features/Instances**	**Results**
Uzair et al. [21]	CNN architectures, Ensemble Learning (GP, AdaBoost)	-	**Orange** (230/50000), **Cell2Cell** (77/40000)	**Orange** (Acc = 75.4% AUC = 0.83), **Cell2Cell** (Acc = 68.2%, AUC = 0.74)
Adnan et al. [22]	Ensemble Learning (Random Forest, Rotation Forest, RotBoost and SVMs)	Filter and Wrapper methods (Particle Swarm Optimization-based undersampling, mRMR, Genetic Algorithm)	**Orange** (230/50000), **Cell2Cell** (77/40000)	**Orange** (AUC = 0.85), **Cell2Cell** (AUC = 0.82)
Devi et al. [23]	BPNN, KNN, SVM, NB	Tomek-linked and redundancy based under-sampling	**Diabetes** Dataset, **Breast Cancer** Dataset	**DD** (AUC = 0.99), **BCD** (AUC = 0.80)
Ngurah et al. [24]	Deep Neural Network	Feature Scaling	**IBM** (21/7043)	Acc = 80.0 %
Lalwani et al. [25]	AdaBoost, XGBoost	Gravitational Search Algorithm (GSA)	**IBM** (21 / 7000)	Acc = 84%
Pamina et al. [26]	XGBoost	Univariate Analysis, Feature Correlation	**IBM** (21 / 7000)	Acc = 0.798
Shatnwai et al. [27]	Gradient Boost Tree (GBT)	Random Oversampler (RO), ADASYN, SMOTE, Borderline SMOTE	**Orange** (20 / 3333)	F-Score = 84%

Caigny et al. [28]	Hybrid model of Decision Tree (DT), Logistic Regression (LogR) and Logit Leaf Model (LLM)	Decision Rules for Customer Segmentation	**Cell2Cell** (70 / 40000)	AUC = 1.786
Halibas et al. [29]	Logistic Regression (LogR)	Exploratory Data Analysis (EDA), Binning (Discretization), Correlation Matrix Operator (CMO), Pearson Correlation Coefficient (PCC)	**IBM** (21 / 7043)	Acc = 80.1%
IRFAN ULLAH et al. [30]	Random Forest (RF)	Correlation Attribute Ranking Filter, Information Gain	**Dataset 1** (29 / 64107), **Dataset 2** (16 / 3333)	Acc = 89 %

3. Material and Method

In this section of the study, the basic infrastructure of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) has been described in detail. The core steps of handling class imbalance technique are Data Cleaning process, Data Transformation, Data Reduction, Handling Class Imbalance, Data Optimization, Ensemble Learning, Model Evaluation and Churn Prediction. The flow of aforementioned core steps has been visualized through Fig 3.

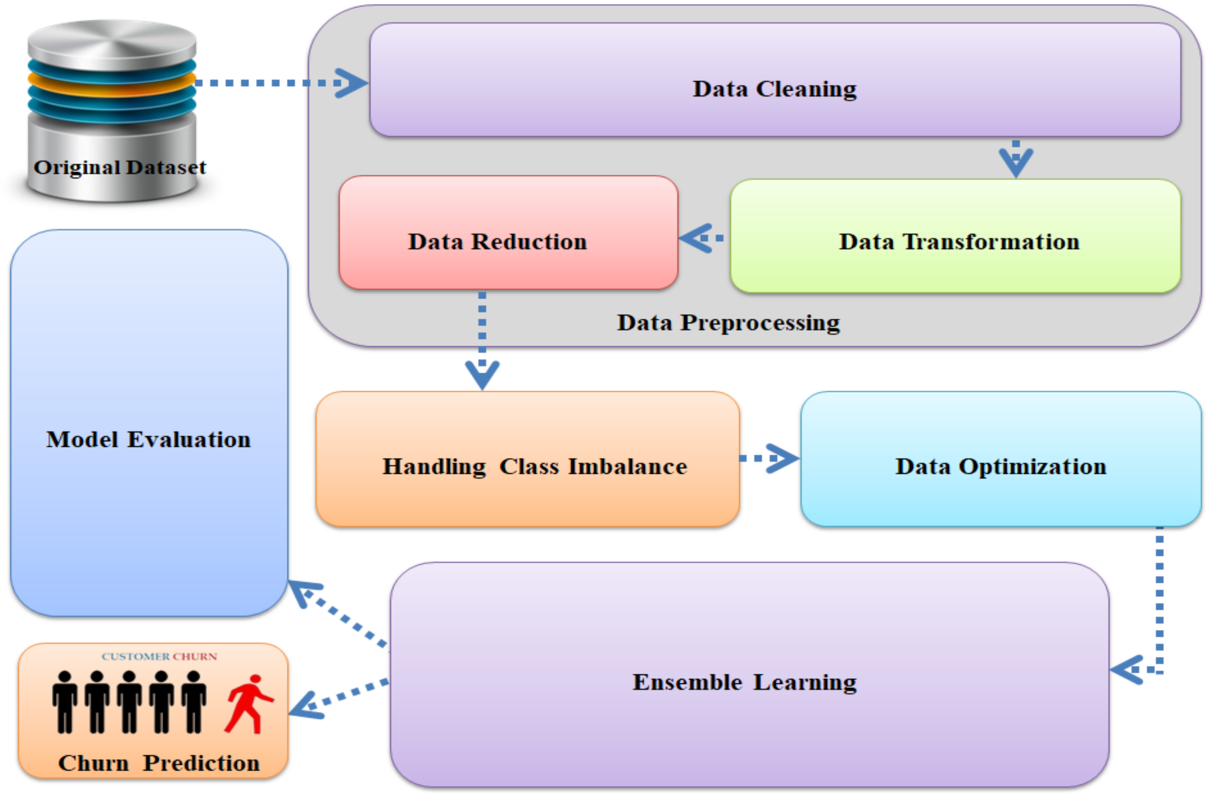

Fig 3: The basic steps of ELOROT

Data Preprocessing

Data Preprocessing is a challenging process due to outlier and missing values in data. Because this phase is required to ensure data validity and data reliability before proceeding to next phases and to improve the overall performance of the model [31, 32]. The data preprocessing phase can further comprise upon data cleaning process, data transformation and data reduction phases.

Data Cleaning

The dirty data with outliers and missing values can spoil the predictive power of a Machine Learning (ML) model and can result into low accuracy. That's the reason due to which data cleaning process is inevitable in the start. In this complex and time consuming process, the whole dataset gets analyzed to make it error free by identifying the outliers, null values and missing values [33]. According to Fig 5, in this research data preprocessing phase has been conducted initially before proceeding to data transformation process.

Data Transformation

Data transformation process is necessary to be performed because this challenging and time-consuming phase contributes to save memory. This phase can transform the categorical and other data type of features into numeric form (int) so it may become feasible for machine learning model [32, 33].

Data Reduction

Analysis of high dimensional data is a cumbersome task to be performed and can lead to inefficient performance of a Machine Learning (ML) model. Data Dimensionality reduction techniques, also known as projection, can step forward to give a solution for this problem [31, 34-36]. Because a dataset can be comprising upon many redundant or irrelevant features. In this situation it is inevitable to select the best features for improving efficiency and accuracy of Machine Learning (ML) classifiers [37] and can contribute to handle class imbalance issue [14-16]. Forward Feature Selection [38-41] has been utilized with assistance of Logistic Regression for dimensionality reduction and to select 36 best most influential features. Fig 4 Visualizes the Steps of Data Reduction process. According to Fig 4, after conducting the preprocessing phase over the dataset, the preprocessed dataset is utilized as input in this phase. In this phase, a feature is considered for next phases only when that feature fulfills the criteria of Forward Feature Selection (FFS), otherwise it gets rejected.

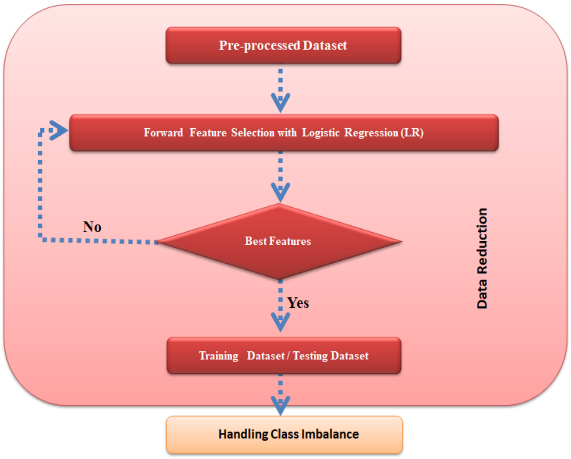
Pre-processed Dataset
Forward Feature Selection with Logistic Regression (LR)
No
Best Features
Yes
Data Reduction
Training Dataset / Testing Dataset
Handling Class Imbalance

Fig 4: Intuition of Data Reduction Process

Handling Class Imbalance

A dataset becomes imbalanced dataset if there is unequal approximation of classes. In other words this is the scenario, in which majority class of the dataset occupies the large proportion while minority class holds a smaller proportion of the dataset [42]. A classifier can perform well by achieving higher accuracy on the behalf of majority class but become poor in case of considering minority class. Hence, the traditional classifiers are unable to deal with this mystery of misbalancing datasets [43]. So, in this paper, data resampling techniques are being applied over an open sourced imbalanced dataset, called Cell2Cell, to deal with this scenario, such as, NearMiss [44, 45], SMOTE [46-50], Random Over Sampling [51, 52], Under Sampling: Tomek Links [53, 54] and ADASYN [55-57]. Data resampling techniques are performed during data preprocessing process to make a balanced dataset for training the classifiers [42]. For this purpose, data resampling techniques usually oversamples the minority class or under samples the majority class [58]. During this phase, all the aforementioned data resampling techniques have been applied over the reduced dataset and trained through Ensemble Learning. Later on, their results were evaluated by calculating their Area Under Curve (AUC), Accuracy, Precision, Recall and F1-Score. Fig 5 reveals that the best observed handling class imbalance technique was used to make a fusion with Particle Swarm Optimization (PSO) in data optimization process.

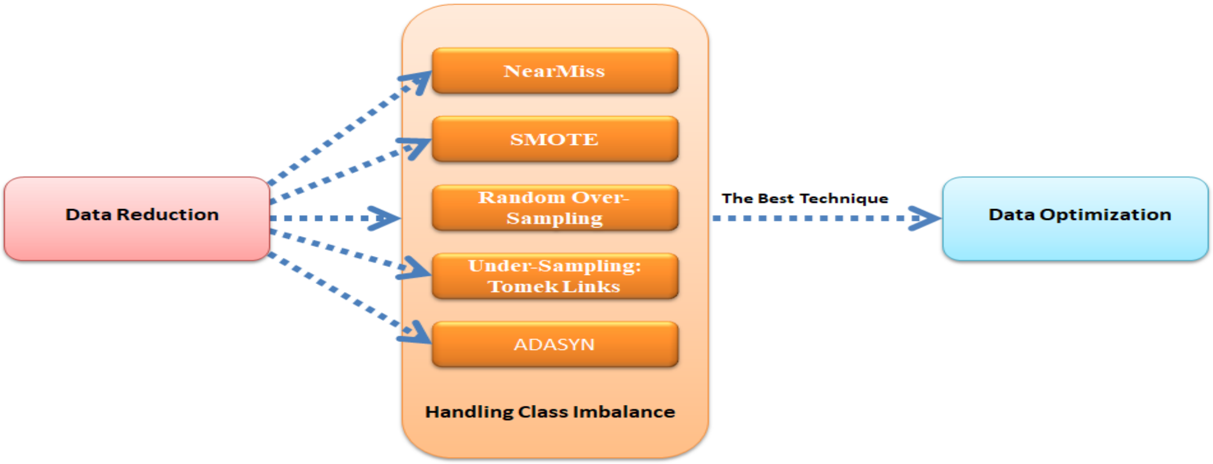

Fig 5: Intuition of Data Handling Process

Particle Swarm Optimization (PSO)

In Ensemble Learning based Optimized Random Oversampling Technique (ELOROT), an iterative and bio inspired data optimization technique, known as Particle Swarm Optimization (PSO) [59], has been unionized with Random Oversampling Technique to extract sensational results from both techniques. Particle Swarm Optimization (PSO) is basically a problem-solving technique which is inspired by bird's flock. It tries to search out the best solution of a problem in a complex solution space to minimize loses and maximizes earns. It extracts one global maximum value and one global minimum value of a flock from multiple local maximums and minimums of each swarm [60]. Table 2 presents the list of all required parameters.

Table 2: Complete Description of Parameters of Particle Swarm Optimization (PSO)

Parameter	**Description**
F	Objective Function
Vi	Velocity of the Particle
Pop	Population of Particles
W	Inertia Weight
C1	Cognitive Constant
U1, U2	Random Constant
C2	Social Constant
Xi	Position of the Particle
P_b	Personal Best
g_b	Global Best

In Particle Swarm Optimization (PSO), a flock may consist on N number of particles and each particle adjusts its velocity (Vi) and position (Xi) to keep track its personal best (pbest) and global best (gbest) by using Inertia Weight (W), Cognitive Constant (C1), Social Constant (C2) and Random Constants (U1, U2). It gives a heuristic optimum solution rather than a global optimal solution [59]. Fig 6 depicts the intuitive framework of Particle Swarm Optimization (PSO).

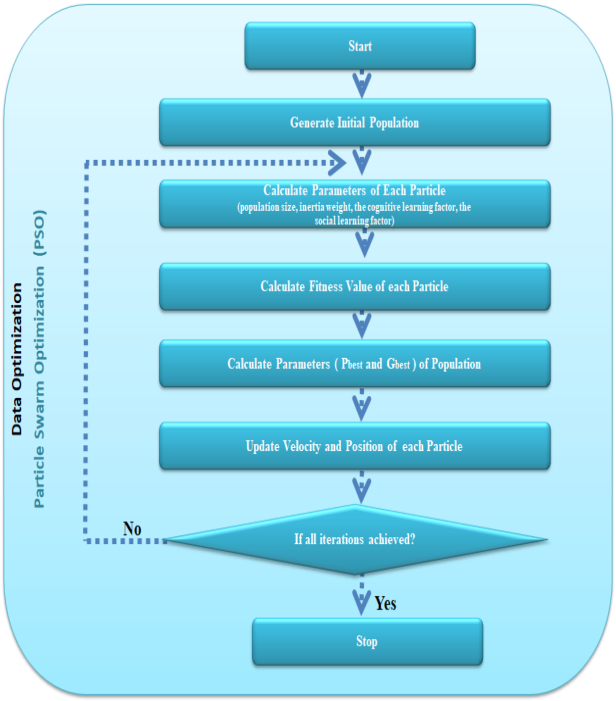
Start
Generate Initial Population
Calculate Parameters of Each Particle
(population size, inertia weight, the cognitive learning factor, the social learning factor)
Calculate Fitness Value of each Particle
Calculate Parameters (Pbest and Gbest) of Population
Update Velocity and Position of each Particle
If all iterations achieved?
No
Yes
Stop
Data Optimization
Particle Swarm Optimization (PSO)

Fig 6: Intuitive Structure of Particle Swarm Optimization (PSO)

Ensemble Learning (EL)

Ensemble Learning can improve the performance and can give a great boost to increase accuracy of customer churn prediction system [61, 62] in telecommunication rather than a standalone classifier. One of the most popular ensembles learning techniques is called stacking, which possess the capability to mitigate the issue of bias and variance. To do so, it combines a few weak classifiers with a strong classifier [47, 63]. Ensemble Learning based upon base classifiers, such as Random Forest (RF), Support Vector Classification (SVC) and K-Nearest Neighbors (KNN), and meta classifiers, such as Support Vector Classification (SVC), has been used for COVID-19 Classification through a dataset of Lung X-Ray and CT Scan images [64]. In the prediction system of break dates of river ice, stacking ensemble learning outperformed [65]. For Customer Churn Prediction in Telecommunication, Stacking Ensemble Learning based upon Random Forest (RF), Rotation Forest (RotF), RotBoost ((RotB) and Support Vector Machine (SVM) Classifiers has been applied over the same Cell2Cell aforementioned data set, which has been considered in this research [22]. Nevertheless, in Telecommunication, Ensemble Learning with stacking for Customer Churn Prediction (CCP) is yet limited and there is much more potential to dig it deep. In this study, Stacking Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) has been presented for customer churn prediction. The Stacking Ensemble Learning (SLE) utilized AdaBoost (AB) [66, 67], and Decision Tree (DT) [68, 69] and Random Forest (RF) [70, 71] as Base classifiers and Logistic Regression (LR) as Meta classifier in stacking. Fig 7 visualizes the structure of Ensemble Learning (EL) used in this study.

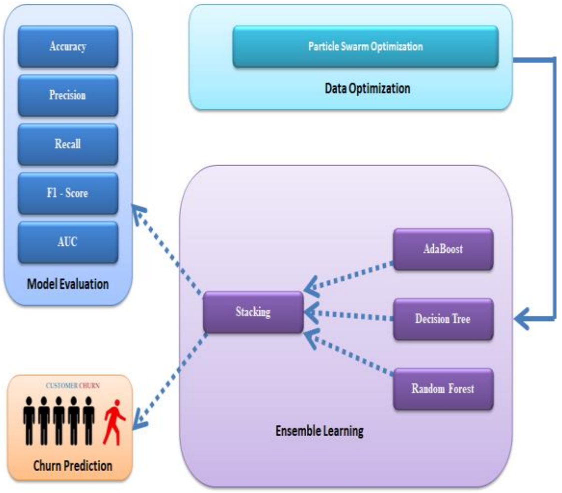

Fig 7: Intuition of Ensemble Learning and Model Evaluation

Model Evaluation

Developed model can be evaluated through Qualitative Evaluation or Quantitative valuation. In qualitative evaluation, End user gives his/her reviews about the performance of the system. While in quantitative evaluation the performance of the system is quantify in form of numbers [72]. In this study the Quantitative evaluation has been used to evaluate the performance of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT). For this purpose, Area Under Curve (AUC), Accuracy, precision, Recall and F1-Score of the Model have been calculated. Table 3 reveals aforementioned performance evaluation metrics through their equations.

Area Under Curve (AUC)

When single number evaluation is required, Area Under Curve (AUC) is best choice to evaluate the overall performance of the machine learning models. It is observed that Area Under Curve (AUC) does not depends upon decision threshold and does not get biased by probability [73-75].

Confusion Metrics

Confusion Matrix is basically a table which can visualize the summary of performance of a machine learning model [76]. It can facilitate to calculate the Accuracy, Recall, Precision, and F1-Score. It is comprises upon True Positive (TP) values, True Negative (TN) values, False Positive (FP) values and False Negative (FN) values [77]. Equation (1) in Table 3 depicts the formula of Confusion Matrix (CM).

Accuracy

Accuracy estimates how much time a machine learning model generates correct predictions in output? It can be calculated by dividing the total correct predictions by total generated predictions [76]. Equation (2) in Table 3 depicts the formula of accuracy.

Precision

Precision tries to find out how many times machine learning model generates correct positive predictions? [76]. Equation (3) in Table 3 depicts the formula to calculate Precision.

Recall

Recall as a performance evaluation metrics tries to find out whether a machine learning model possess the ability to find entire positive class instances or not? [76]. Equation (4) in Table 3 depicts the formula to calculate Recall.

F1-Score

F1-Score is also known as harmonic mean of recall and precision [78]. It possess the ability to sums up elegantly the predictive performance of multiclass or binary class classification model [76]. Equation (5) in Table 3 depicts the formula to calculate F1-Score.

Table 3: List of Performance valuation matrices

Performance Evaluation Matrices	**Equation**
Confusion Matrix [76]	$= \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$ (1)
Accuracy [76]	$= \frac{TN+TP}{TN+FP+FN+TP}$ (2)
Precision [76]	$= \frac{TP}{FP+TP}$ (3)
Recall [76]	$= \frac{TP}{TP+FN}$ (4)
F1-score [76]	$= \frac{2 * Precision * Recall}{Prcision + Recall}$ (5)

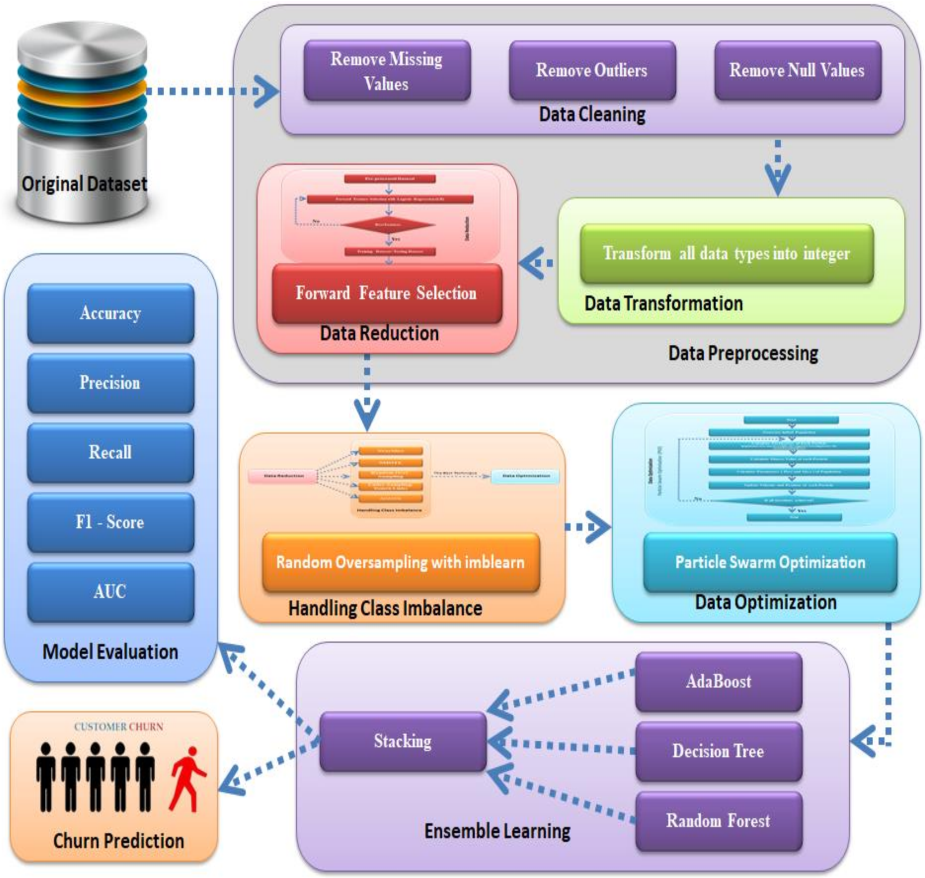

Fig 8: Ensemble Learning based Optimized Random Oversampling (ELOROT) Methodology

In this study, Fig 8 describes the complete methodology with internal structure of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT). In data preprocessing, after performing data cleaning and data transformation processes, Forward Feature Selection technique has been used to select the 35 most influential features with respect to targeted feature "Churn". Then a few data class imbalance handling techniques have been applied one by one over the extracted features to resample the dataset and compared their results to check their effects on Ensemble Learning and to choose most effective one. For this purpose their results were evaluated through performance evaluation matrices, such as, Confusion Matrix (CM), Accuracy, Precision, Recall and F1-Score. Later on a fusion of most effective data class imbalance handling technique and Particle Swarm Optimization (PSO) was designed to be used with ensemble Learning to make an Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) for customer churn prediction. The results of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) were benchmark against some studies.

4. Results and Discussion

In this section of the study, we test the effects of a few most popular Data Class Imbalance handling techniques over an open sourced dataset available on Kaggle, called Cell2Cell [11, 21, 22], with 58 features and 51047 instances. In this study's experiment, initially after performing data cleaning process and data transformation process, Forward Feature Selection (FFS) with Logistic Regression has been used for dimensionality reduction of the considered dataset. In this study, the considered data splitting ratio was 80:20 for training and testing. The rest experiment of this study has been conducted in two phases.

Phase (1):

In Phase (1), after performing Data Cleaning Process, Data Transformation Process and Data Reduction Process, various most popular imbalance data handling techniques such as, NearMiss [44, 45], SMOTE [46-50], Random Over Sampling [51, 52], Under Sampling: Tomek Links [53, 54] and ADASYN [55-57], have been applied over the dataset to make it balanced and then trained and tested in assistance of Ensemble Learning. Fig 9 visualizes the effects of aforementioned Data Resampling Techniques over dataset. Whereas Table 4 reveals the story of Dataset Shapes with or without Data Resampling Techniques in sampling numbers. While Table 5 describes the combined results of data resampling techniques with forward Feature Selection (FFS) and Ensemble Learning (EL). Fig 10 presents the visualization of recorded Area Under Curve (AUC) of all Data Resampling Techniques in assistance of Ensemble Learning (EL), without being optimized.

Goal:

The keen goal of the Phase (1) was to find out the best and most influential imbalance data handling technique in assistance of Ensemble Learning.

Comparing Results:

According to Table 5, Area Under Curve (AUC), accuracy, precision, Recall and F1-Score have been calculated of each possible combination of data resampling techniques with forward Feature Selection (FFS) and Ensemble Learning (EL).

Best Class Imbalance Handling Technique Selection Criteria:

The class imbalance handling techniques, considered in this study, were required to score highest Area Under Curve (AUC), Accuracy, Precision, Recall and F1-Score in assistance of Ensemble Learning to be selected as best technique among all.

Best Selected Technique:

From Table 5, it is observed that Random Over Sampling (ROS) fulfilled the Criteria by scoring the 0.85 Area Under Curve (AUC), 0.76 Precision, 0.82 Recall, 0.78 Accuracy and 0.792 F1-Score. In Short, in Phase (1) Random Oversampling (ROS) technique outperformed among all other data resampling techniques and selected as best technique to be considered in Phase (2).

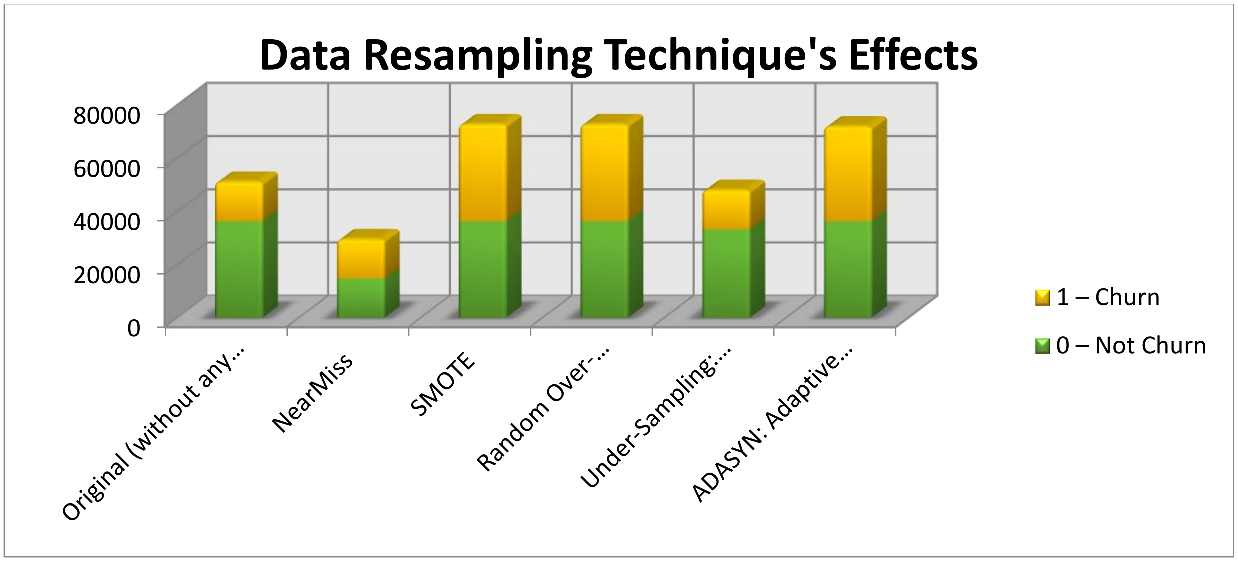

Fig 9: Effects of Data Resampling Techniques over dataset

Table 4: Dataset Shapes with / without Resampling Techniques

Dataset Shapes with / without Resampling Techniques		
Techniques	**0 – Not Churn**	**1 – Churn**
Original (without any technique)	36336	14711
NearMiss	14711	14711
SMOTE	**36336**	**36336**
Random Over-Sampling With imblearn	**36336**	**36336**
Under-Sampling: Tomek Links	33238	14711
ADASYN: Adaptive Synthetic Sampling Approach	36336	35507

Table 5: Results of Resampling Techniques with Ensemble Learning without PSO

Model	AUC	Precision	Recall	Accuracy	F1 Score
NearMiss + FFS + EL + Cell2Cell	0.74	0.67	0.66	0.67	0.665
SMOTE + FFS + EL + Cell 2 Cell	0.73	0.67	0.69	0.67	0.677
Random Over-Sampling With imblearn + FFS + EL + Cell2Cell	**0.85**	**0.76**	**0.82**	**0.78**	**0.792**
Under-Sampling: Tomek Links + FFS + EL + Cell2cell	0.59	1.000	0.00	0.69	0.001
ADASYN: Adaptive Synthetic Sampling Approach + FFS + EL + CELL2CELL	0.70	0.64	0.647	0.65	0.65

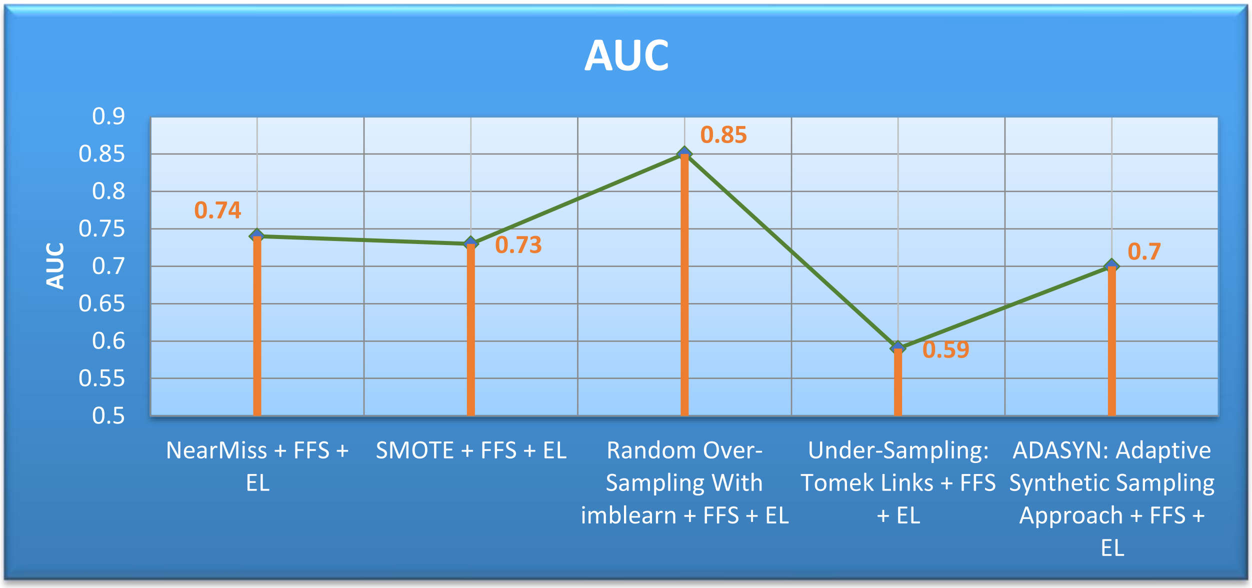

Fig 10: AUC of Resampling Techniques with Ensemble Learning without PSO

Phase (2):

In phase (2), the winning technique of Phase (1), called Random Oversampling (ROS), was considered as data resampling technique to create a fusion with Particle Swarm Optimization. So, it may become optimized as well. That fusion was named as Optimized Random Oversampling Technique (OROT). When Optimized Random Oversampling Technique (OROT) got applied over the Cell2Cell dataset to train and test in assistance of Ensemble Learning (EL) a novice technique got developed, which was named as Ensemble Learning based Optimized Random Oversampling Technique (ELOROT).

Goal:

The keen goal of the Phase (2) was to optimize Random Oversampling (ROS) technique by creating its fusion with Particle Swarm Optimization (PSO) Algorithm, so it may become able to improve the performance accuracy of a Customer Churn Prediction Model.

Comparing Results:

Area Under Curve (AUC) of the Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) was calculated to benchmark it. According to Fig 11, Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) scored 0.92 Area Under Curve (AUC).

Benchmarking Criteria

Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) was required to score highest Area Under Curve (AUC) against all other considered existing studies.

Benchmark against existing studies

In this section of the study, the validity of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) has been benchmarked against various existing studies. To do so, state-of-the-art performance evaluation technique, calls Area Under Curve (AUC), has been used. Because, in existing studies Area Under Curve (AUC) is being considered as a best choice for performance evaluation of a model in single number form [73-75]. Table 6 reveals the finale scores of a few existing studies in terms of Area Under Curve (AUC) to benchmark Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) against them. Fig 11 presents the visual representation of Table 6 and shows the superiority of Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) over the studies of Uzair et al. [21], Khoshgoftaar et al. [11] and Adnan et al. [22]. All the aforementioned studies trained and tested their models over the same dataset, considered by Ensemble Learning based Optimized Random Oversampling Technique (ELOROT). Table 6 describes that in existing studies Uzair et al. [21] recorded 0.74, Khoshgoftaar et al. [11] recorded 0.73 and Adnan et al. [22] recorded 0.82 Area Under Curve (AUC). Hence by considering the results of Table 6 and Visual representation of Fig 11, it is concluded that Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) outperformed by scoring 0.92 Area Under Curve (AUC), which is the highest recorded Area Under Curve (AUC) yet over Cell2Cell dataset in Telecommunication for customer churn prediction system.

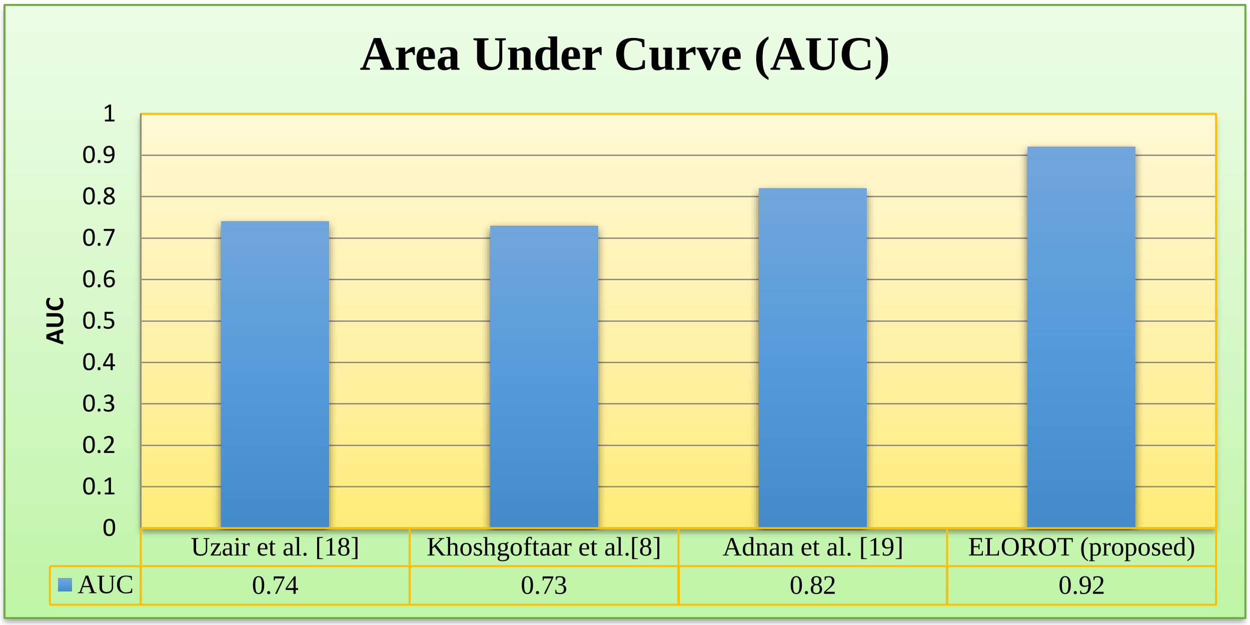

Fig 11: to compare ELROT with benchmark studies

Table 6: Ensemble Learning based Optimized Random Oversampling Technique (ELOROT) benchmark

Reference	AUC
Uzair et al. [21]	0.74
Khoshgoftaar et al.[11]	0.73
Adnan et al. [22]	0.82
ELOROT (proposed)	0.92

5. Conclusion

The major objective of this study was to find out the optimized class imbalance handling technique for developing a better customer churn prediction system with Ensemble Learning (EL) in telecommunication sector. For this purpose, a fusion of Random Oversampling (ROS) and Particle Swarm Optimization (PSO), named as Optimized Random Oversampling Technique (OROT), has been devised to make the dataset optimized before training and testing process of customer churn prediction. The proposed technique, named as Ensemble Learning based Optimized Random Oversampling Technique (ELOROT), effectively outperformed by scoring 0.92 Area Under Curve (AUC) over Cell2Cell dataset, which is the highest score yet. in future, it is essential to devise more fusions of various data resampling and data optimization techniques to achieve next levels of accuracy.

References

1. Singh, N., P. Singh, and M. Gupta, *An inclusive survey on machine learning for CRM: a paradigm shift.* Decision, 2021. 47(4): p. 447-457.
2. Mahmud, M.S., et al., *A survey of data partitioning and sampling methods to support big data analysis.* Big Data Mining and Analytics, 2020. 3(2): p. 85-101.
3. Hasanin, T., et al., *Severely imbalanced Big Data challenges: investigating data sampling approaches.* Journal of Big Data, 2019. 6(1).
4. Peng, C.-Y. and Y.-J. Park, *A New Hybrid Under-sampling Approach to Imbalanced Classification Problems.* Applied Artificial Intelligence, 2021. 36(1).
5. Bauder, R.A. and T.M. Khoshgoftaar, *The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data.* Health Inf Sci Syst, 2018. 6(1): p. 9.
6. Buda, M., A. Maki, and M.A. Mazurowski, *A systematic study of the class imbalance problem in convolutional neural networks.* Neural Netw, 2018. 106: p. 249-259.
7. Leevy, J.L., et al., *A survey on addressing high-class imbalance in big data*. Journal of Big Data, 2018. 5(1).
8. SOLOMON H. EBENUWA, M.S.S., MAMOUN ALAZAB, AND AMEER AL-NEMRAT, *Variance ranking attributes selection techniques for binary classification problem in imbalance data.* IEEE Access, 2018. 7: p. 24649 - 24666.
9. Xu-Ying Liu, J.W., and Zhi-Hua Zhou, *Exploratory undersampling for class-imbalance learning.* IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS – PART B, 2019 39(2): p. 1 - 14.
10. Aida Ali, S.M.S., Anca Ralescu, *Classification with class imbalance problem A review.* Int. J. Advance Soft Compu. Appl, Research Gate 2015. 5(3): p. 1-31.
11. Khoshgoftaar, T.M., et al., *Learning with limited minority class data.* 2007: p. 348-353.
12. Jason Van Hulse , T.M.K., Amri Napolitano *Experimental Perspectives on Learning from Imbalanced Data.* ACM Digital Library, 2007: p. 935 - 942.
13. Malhotra, R., *A systematic review of machine learning techniques for software fault prediction.* Applied Soft Computing, 2015. 27: p. 504-518.
14. Yin, L., et al., *Feature selection for high-dimensional imbalanced data.* Neurocomputing, 2013. 105: p. 3-11.
15. Zhaohui Zheng, X.W., Rohini Srihari, *Feature selection for text categorization on imbalanced data.* Explor Newsletter, 2014. 6(1): p. 81-89.
16. Dunja Mladenić, M.G., *Feature selection for unbalanced class distribution and Naïve Bayes.* Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, 1999: p. 1-10.
17. Desuky, A.S. and S. Hussain, *An Improved Hybrid Approach for Handling Class Imbalance Problem.* Arab J Sci Eng, 2021. 46(4): p. 3853-3864.

18. Napolitano, C.S.T.M.K.J.V.H.A., *RUSBoost A Hybrid Approach to Alleviating Class Imbalance.* IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2010. 40(1): p. 185 - 197.

19. Magdalena Graczyk, T.L., Bogdan Trawiński, Krzysztof Trawiński, *Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal.* Proceedings of the Second international conference on Intelligent information and database systems: Part I, Springer, 2010: p. 340–350,.

20. Khan, A., et al., *A recent survey on the applications of genetic programming in image processing.* Computational Intelligence, 2021. 37(4): p. 1745-1778.

21. Uzair Ahmed, A.K., Saddam Hussain Khan, Abdul Basit, Irfan Ul Haq, and and Y.S. Lee, *Transfer Learning and Meta Classification Based Deep Churn System for Telecom Industry.* arXiv.org, 2019: p. 1-10.

22. Adnan Idris and A. Khan, *Churn Prediction System for Telecom using Filter–Wrapper and Ensemble Classification.* The Computer Journal, IEEE, 2017. 60(3): p. 410-430.

23. Devi, D., S.k. Biswas, and B. Purkayastha, *Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance.* Pattern Recognition Letters, 2017. 93: p. 3-12.

24. Ngurah Putu Oka H and A.S. Arifin, *Telecommunication Service Subscriber Churn Likelihood Prediction Analysis Using Diverse Machine Learning Model.pdf.* 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT), 2020: p. 24-29.

25. Lalwani, P., et al., *Customer churn prediction system: a machine learning approach.* Computing, 2021. 104(2): p. 271-294.

26. J. Pamina, J.B.R., S. Sathya Bama, S. Soundarya, M.S. Sruthi, S. Kiruthika, V.J. Aiswaryadevi, G. Priyanka, *An Effective Classifier for Predicting Churn in Telecommunication.* Jour of Adv Research in Dynamical & Control Systems, 2019. 11(01): p. 221-229.

27. Atallah M. AL-Shatnwai , M.F., *Predicting Customer Retention using XGBoost and Balancing Methods.* International Journal of Advanced Computer Science and Applications (IJACSA), 2020. 11(7): p. 704-712.

28. De Caigny, A., K. Coussement, and K.W. De Bock, *A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees.* European Journal of Operational Research, 2018. 269(2): p. 760-772.

29. Alrence Santiago Halibas, A.C.M., Indu Govinda Pillai, Jay Harold Reazol, Erbeth Gerald Delvo, Leslyn Bonachita Reazol, *Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling.* IEEE ACCESS - 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), 2019: p. 1-7.

30. IRFAN ULLAH, B.R., AHMAD KAMRAN MALIK, MUHAMMAD IMRAN, SAIF UL ISLAM, SUNG WON KIM, *A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector.* IEEE ACCESS, 2019. 7: p. 60134 - 60149.

31. Fan, C., et al., *A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data.* Frontiers in Energy Research, 2021. 9.

32. Joshi, A.P. and B.V. Patel, *Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process.* Oriental journal of computer science and technology, 2021. 13(0203): p. 78-81.

33. Fakhitah Ridzuan, W.M.N.W.Z., *A Review on Data Cleansing Methods for Big Data.* Procedia Computer Science, Science Direct, 2019. 161: p. 731–738.

34. Yair Barta, N.F., Ofer Neiman, *Dimensionality reduction: theoretical perspective on practical measures.* 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019: p. 1-13.

35. Xu, H., et al., *Supervised breast cancer prediction using integrated dimensionality reduction convolutional neural network.* PLoS One, 2023. 18(5): p. e0282350.

36. Jing Zhou, D.P.F., Robert A. Stine, Lyle H. Ungar, *Streamwise Feature Selection.* Journal of Machine Learning Research, 2006. 7: p. 1861-1885.

37. Maria Luque-Rodriguez, J.M.-B., Alfonso Jimenez-Vilchez, Antonio Arauzo-Azofra, *Initialization of Feature Selection Search for Classification.* Journal of Artificial Intelligence Research 2022. 75: p. 953-983.

38. Saifudin, A., et al., *Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes.* Journal of Physics: Conference Series, 2020. 1477(3): p. 032007.

39. Etika Kartikadarma, P.A.C., Faisal Syafar, Akbar Iskandar, Arman Paramansyah, Robbi Rahim, *Application of forward selection strategy using C4.5 algorithm to improve the accuracy of classification's data set.* Journal of Population Therapeutics& Clinical Pharmacology, 2023. 30(1): p. e14–e23.

40. Erik Schaffernicht, C.M., Klaus Debes, Horst-Michael Gross, *Forward Feature Selection Using Residual Mutual Information.* Conference: ESANN 2009, 17th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 22-24, 2009, Proceedings, Research Gate, 2009: p. 583-588.

41. Mostafa, A.M., et al., *Innovative Forward Fusion Feature Selection Algorithm for Sentiment Analysis Using Supervised Classification.* Applied Sciences, 2023. 13(4): p. 2074.

42. Lian Yu, N.Z., *Survey of Imbalanced Data Methodologies.* arxiv 2021: p. 1-7.

43. Tanimoto, A., et al., *Improving imbalanced classification using near-miss instances.* Expert Systems with Applications, 2022. 201: p. 117130.

44. Joloudari, J.H., et al., *Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks.* Applied Sciences, 2023. 13(6): p. 4006.

45. Bao, L., et al., *Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets.* Neurocomputing, 2016. 172: p. 198-206.

46. Chawla, N.V., et al., *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 2002. 16: p. 321-357.

47. Lu, M., et al., *A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting.* Water, 2023. 15(7): p. 1265.

48. He, H., W. Zhang, and S. Zhang, *A novel ensemble method for credit scoring: Adaption of different imbalance ratios.* Expert Systems with Applications, 2018. 98: p. 105-117.

49. Zięba, M. and J.M. Tomczak, *Boosted SVM with active learning strategy for imbalanced data.* Soft Computing, 2014. 19(12): p. 3357-3368.

50. Alberto Fern´andez, S.G., Francisco Herrera, Nitesh V. Chawla, *View of SMOTE for Learning from Imbalanced Data_ Progress and Challenges, Marking the 15-year Anniversary.* Journal of Artificial Intelligence Research 2018. 61: p. 863-905.

51. Azadbakht, M., C.S. Fraser, and K. Khoshelham, *Synergy of sampling techniques and ensemble classifiers for classification of urban environments using full-waveform LiDAR data.* International Journal of Applied Earth Observation and Geoinformation, 2018. 73: p. 277-291.

52. Moreo, A., A. Esuli, and F. Sebastiani, *Distributional Random Oversampling for Imbalanced Text Classification.* 2016: p. 805-808.

53. Swana, E.F., W. Doorsamy, and P. Bokoro, *Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset.* Sensors (Basel), 2022. 22(9).

54. Kamaladevi M, V.V., Sekar K R, *Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification.* Annals of R.S.C.B., 2021. 25(4): p. 2182–2190.

55. Haibo, H., et al., *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* 2008: p. 1322-1328.

56. Khan, T.M., et al., *Implementing Multilabeling, ADASYN, and ReliefF Techniques for Classification of Breast Cancer Diagnostic through Machine Learning: Efficient Computer-Aided Diagnostic System.* J Healthc Eng, 2021. 2021: p. 5577636.

57. Qing, Z., et al., *ADASYN-LOF Algorithm for Imbalanced Tornado Samples.* Atmosphere, 2022. 13(4): p. 544.

58. Johnson, J.M. and T.M. Khoshgoftaar, *Survey on deep learning with class imbalance.* Journal of Big Data, 2019. 6(1).

59. Wang, K.-J., et al., *A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients.* Applied Soft Computing, 2014. 20: p. 15-24.

60. Eberhart, J.K.a.R., *particle-swarm-optimization.* Proceedings of ICNN'95 - International Conference on Neural Networks, IEEE, 1995: p. 1942-1948.

61. Nti, I.K., A.F. Adekoya, and B.A. Weyori, *A comprehensive evaluation of ensemble learning for stock-market prediction.* Journal of Big Data, 2020. 7(1).

62. Singh, R., *Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance.* International Journal of Advanced Trends in Computer Science and Engineering, 2020. 9(3): p. 3970-3976.

63. Zhu, X., et al., *An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: The example of olanzapine.* Front Pharmacol, 2022. 13: p. 975855.

64. Berliana, A.U. and A. Bustamam, *Implementation of Stacking Ensemble Learning for Classification of COVID-19 using Image Dataset CT Scan and Lung X-Ray*. 2020: p. 148-152.

65. Sun, W., et al., *Modeling River Ice Breakup Dates by k-Nearest Neighbor Ensemble.* Water, 2020. 12(1): p. 220.

66. Hornyák, O. and L.B. Iantovics, *AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics.* Mathematics, 2023. 11(8): p. 1801.

67. Ding, Y., et al., *An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification.* Applied Sciences, 2022. 12(12): p. 5872.

68. Song, Y.Y. and Y. Lu, *Decision tree methods: applications for classification and prediction.* Shanghai Arch Psychiatry, 2015. 27(2): p. 130-5.

69. Ibomoiye Domor Mienye, Y.S., Zenghui Wang, *Prediction performance of improved decision tree-based algorithms: a review.* 2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019), Science Direct, 2019. 35: p. 698–703.

70. Schonlau, M. and R.Y. Zou, *The random forest algorithm for statistical learning.* The Stata Journal: Promoting communications on statistics and Stata, 2020. 20(1): p. 3-29.

71. Solorio-Ramírez, J.-L., et al., *Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects.* Algorithms, 2023. 16(6): p. 293.

72. Dalianis, H., *Evaluation Metrics and Evaluation.* 2018: p. 45-53.

73. Bowers, A.J. and X. Zhou, *Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes.* Journal of Education for Students Placed at Risk (JESPAR), 2019. 24(1): p. 20-46.

74. Al-Hashem, M.A., A.M. Alqudah, and Q. Qananwah, *Performance Evaluation of Different Machine Learning Classification Algorithms for Disease Diagnosis.* International Journal of E-Health and Medical Communications, 2021. 12(6): p. 1-28.

75. Bradley, A.P., *The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms.* Pattern Recognition, , 1997. 30(7): p. 1145-1159.

76. Smirani, L.K., et al., *Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths.* Scientific Programming, 2022. 2022: p. 1-15.

77. Karimi, Z., *Confusion Matrix.* Research Gate, 2021: p. 1-4.

78. Michael Buckland, F.G., *The Relationship between Recall and Precision.* JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. , 1994. 45(1): p. 12-19.