Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

AN EVOLUTIONARY FUSION OF OPTIMIZATION TECHNIQUES FOR CUSTOMER CHURN PREDICTION IN TELECOMMUNICATION

SABAHAT TASNEEM

PhD Degree Program, Computer Science, Government College University, Faisalabad, Pakistan.
Email: sb_tasneem@yahoo.com

MUHAMMAD YOUNAS *

Assistant Professor, Information Technology Department, Government College University, Faisalabad.
*Corresponding Author Email: myounas@gcuf.edu.pk, younas.76@gmail.com

MUHAMMAD MURAD KHAN

Assistant Professor, Comuputer Science Department, Government College University, Faisalabad.
Email: muradtariq.tk@gmail.com

UZMA JAMIL

Assistant Professor, Computer Science Department, Government College University, Faisalabad.
Email: uzma.jamil@gcuf.edu.pk

KASHIF HANIF

Associate Professor, Computer Science Department, Government College University, Faisalabad.
Email: mkashifhanif@gcuf.edu.pk

Abstract

Currently, we are living in the era of big data, which is an innovative area of research and has been developing for a couple of decades. Due to the saturated situation in the market, telecommunication companies all over the world are trying to seek a competitive edge and to explore innovative and robust solutions for dealing with vulnerabilities of Big Data. It is a challenging task for them to dig deep into bulky data of customers, which may be in Tera Bytes or Pita Bytes, for extracting scientific insights to take decision for retaining its loyal customers in future. Therefore, it is essential for telecommunication companies to turn the bulky imbalanced dataset of customers into revenue. Because it is much expensive for the telecommunication companies to acquire more new customers rather than retaining the previous once. In this study, a novice robust and efficient fusion of data optimization techniques, based upon SMOTE, Ring-Theory and Particle Swarm Optimization, has hybridized with Heterogamous Ensemble Learning to deal with the data imbalance issue and to improve the overall performance of the customer churn prediction model. The proposed model has been named Optimized Customer Churn Prediction (OCCP) scored 0.940 Kappa Statistics, 0.0298 Mean Absolute Error, 0.173 Root Mean Square Error and 0.97 Accuracy. It is required to explore more innovative fusions of optimization techniques to improve accuracy a little bit more.

Index Terms: Particle Swarm Optimization, Synthetic Minority Oversampling Technique, Ring Theory, Ensemble Leaning, Churn Prediction, Forward Feature Selection, Decision Tree.

1. INTRODUCTION

Today, Telecommunication is one of the most rapidly growing business sectors, facing fierce competition due to saturated market and commercialism. This is the sector whose progress is directly proportional to customer's satisfaction (Akter, 2022; Naz et al., 2021; Shahzad et al., 2021). Due to captivating offers from competitors, the loyal customers of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

a telecommunication company get obsessed and switch to other companies. This situation is called customer churn (Andrews, 2019; Srinivasan et al., 2023). Proactively customer churn prediction can play a vital role to retain loyal customers of a telecom company and to turn the customers' data into profit. Because, retention of loyal customer is less expensive than captivating new customers (Fareniuk et al., 2022; Saha, 2023). But unfortunately, imbalance nature of dataset is a hurdle on the way to prosperity for telecommunication companies.

Owing to disproportionate ratio of samples, majority gets authority for minority class and results into biased predictions by classifiers (Bauder & Khoshgoftaar, 2018; Buda et al., 2018; Leevy et al., 2018). So, due to imbalanced dataset, classifiers become uncertain and unreliable (Ebenuwa, 2018; Xu-Ying Liu, 2019). Hence, it is required to pay attention to control the imbalanced data issue. Thus, a lot of research has conducted to cope with it which can categorized into Classification Algorithm Level research and Data optimization level research (Aida Ali, 2015; Khoshgoftaar et al., 2007).

The Figure 1 presents the major types of data optimization techniques, which are the hot topics of research now a day. In actual, data can be optimized by using various data resampling techniques, such as data under sampling and data over sampling, or by using the Feature selection techniques, which can be discussed under the umbrella term of data level class imbalance handling techniques. On contrary side, under the umbrella term of algorithm level class imbalance handling techniques, data can be optimized by using some cost sensitive algorithms or ensemble methods (Dunja Mladenić, 1999; Jason Van Hulse 2007; Malhotra, 2015; Yin et al., 2013; Zhaohui Zheng, 2014). Figure 2 reveals the function of data under-sampling and data oversampling techniques.

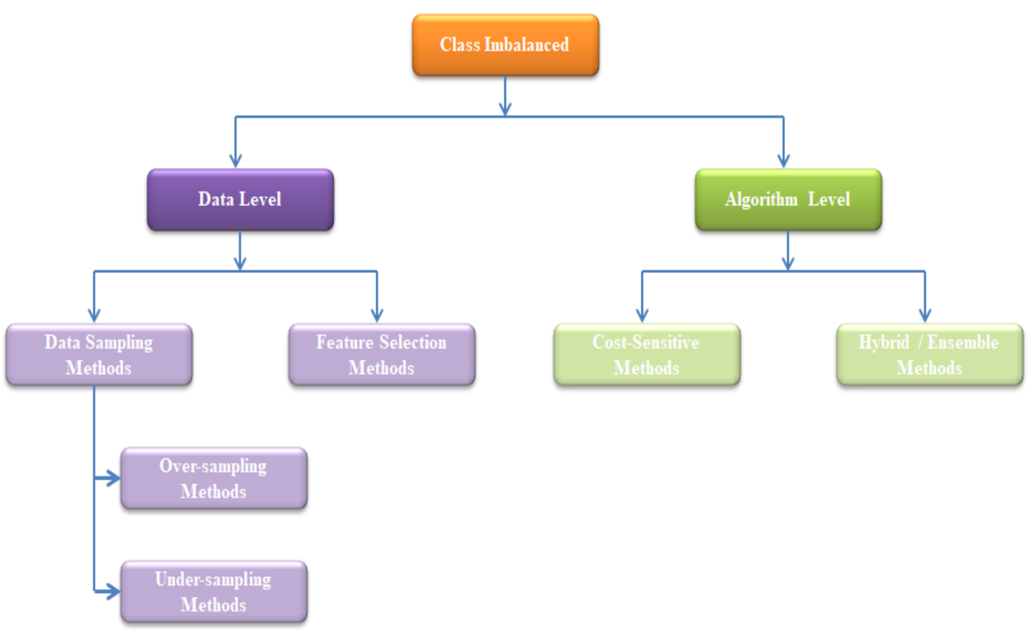


Figure 1: Types of data optimization techniques

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

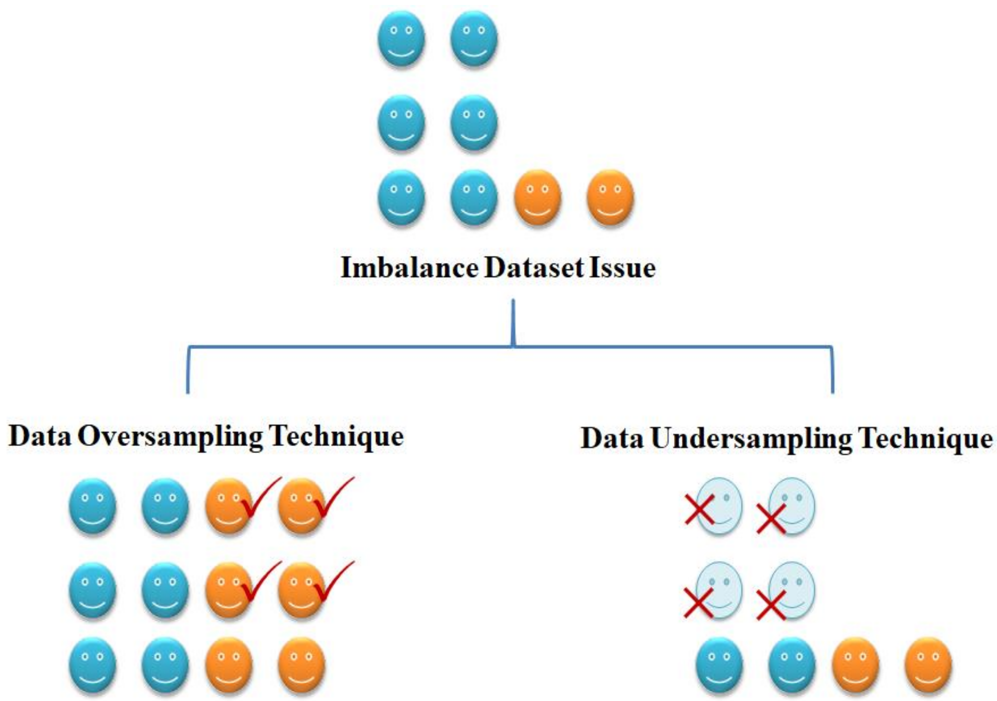


Figure 2: Function of Data Oversampling Techniques and Data Under-sampling

This research work implicates an optimized strategy to handle imbalanced data distribution, which has named as, SMOTE based Ring Particle Swarm Optimization (SRPSO). The presented SRPSO technique has embedded in a customer churn prediction model, which has named as Optimized Customer Churn Prediction (OCCP) model. Optimized Customer Churn Prediction (OCCP) is a highly intelligent model with capability of behavioral pattern recognition from customer's historical data to make churn prediction. OCCP devised a fusion of various data optimization techniques, such as Forward Feature Selection (FFS) with Logistic Regression as learning parameter; Synthetic Minority Oversampling Technique (SMOTE) based Ring Particle Swarm Optimization (SRPSO) and Ensemble Learning with stacking. The key contribution of this Optimized Customer Churn Prediction (OCCP) model are as follows:

1. The prediction accuracy has been improved by extracting valuable and correlated features in training dataset for making churn predictions.
2. The Optimized Customer Churn Prediction (OCCP) possess the capability to reduce the size of data by using the Forward Feature Selection with Logistic Regression as learning parameter
3. A fusion of SMOTE based Ring-Theory Particle Swarm Optimization (SRPSO) based upon Particle Swarm Optimization (PSO), Synthetic Minority Oversampling Technique (SMOTE) and RT- based Evolutionary algorithm (RTEA) has been used to strengthen and improve the process of feature selection by mitigating the demerits of SMOTE, PSO and RTEA

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

4. Heterogeneous Ensemble Learning (HEL) has been used with stacking, which contributed to improve the prediction performance, to improve the accuracy level of the prediction and assisted to make a robust model for churn prediction. Fusion of heterogeneous classification algorithms in Ensemble Learning (EL) performed better and improved the classification process than applying the single classifier.
5. The Optimized Customer Churn Prediction (OCCP) makes the data analyzable, coherent, salable and reliable by controlling the issue of over sampling and under-sampling.
6. The model has been evaluated by using various evaluation metrics such as Sensitivity, specificity, Area under Curve (AUC), Confusion Matrix, accuracy, precision, recall and receiver operating characteristic curve (ROC).
7. The Optimized Customer Churn Prediction (OCCP) model has the ability to identify factors behind increasing churn rate, so the Telecommunication Company become able to mitigate those issues in the bud to increase revenue.
8. The Optimized Customer Churn Prediction (OCCP) model is able to categorize the customers according to their churn rate and behavior.

2. REVIEW OF LITERATURE

Table 1 presents the complete summary of Literature Review for this study. The whole literature has discussed under the 4 umbrella terms, such as Classification algorithms, Data Optimization Techniques, considered Dataset and Results. Table 1 reveals that a lot of research has conducted to find out the best classification algorithms and data optimization techniques. But still the existing research is unable to achieve its goals in terms of improved accuracy. Table 1 also presents the best classification and data optimization techniques combination till now.

According to Table 1, Idris et al. (2012) achieved 0.7511 Area under Curve by hybridizing the effects of Ensemble Learning (EL), with Random Forest (RF) and K Nearest Neighbor (KNN), Particle Swarm Optimization (PSO), Principle Component Analysis (PCA), Minimum Redundancy and Maximum Relevancy (mRMR), Fisher's Ratio and F sore. This combination of classification algorithm and optimization algorithm has evaluated over Orange dataset with 260 features and 50000 instances. Over the same Orange dataset and Cell2Cell dataset with 76 features and 40000 instances, Adnan Idris et al. (2012) achieved 0.89 Area Under Curve (AUC) by hybridizing the effects of Homogeneous Ensemble Learning (Hom.EL), with AdaBoost (AB) as weak classifier, and Genetic Programming (GP), for data optimization.

In another model with aforementioned Orange and Cell2Cell dataset, Idris et al. (2013) achieved 0.699 Area under curve (AUC) by combining the effects of Ensmeble Learning, Random forest (RF), Rotation Forest (RotF), Rotation Boost (RotB), and Decorate as weak classifiers, and Minimum Redundancy and Maximum Relevancy (mRMR). Idris and Khan (2012) achieved 0.816 AUC score by combining the effects of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Ensemble Learning, with Rotation Boost (RotB) and Rotation Forest (RotF), with Principle Component Analysis (PCA), Fisher's Ratio and F Score. In another model with aforementioned Orange and Cell2Cell dataset, Adnan Idris and Khan (2017) achieved 0.85 and 0.82 Area Under Curve (AUC) on Orange and Cell2Cell dataset respectively, by combining the effects of Ensemble Learning, with RF, RotF, RotB and Support Vector Machine (SVM), with PSO, mRMR, Genetic Programming (GP). Uzair Ahmed et al. (2019) achieved 75.4% and 68.2% accuracy on aforementioned Orange and Cell2Cell dataset, respectively, by applying CNN architectures and Ensemble Learning with Genetic Programming (GP) and AdaBoost (AB).

Tianpei Xu (2021) scored 0.96% accuracy on Orange dataset with 20 features and 3333 instances by combining the effect of Ensemble Learning, with Decision tree (DT), Extreme Gradient Boosting (XGB), Linear Regression (LR) and Naïve Bayes (NB), with Equi Distance technique. While Abinash Mishra (2017) scored 91.66% accuracy on the same aforemention Orange dataset by applying Homogeneous Ensemble Learning with Random Forest (RF). Amin et al. (2017) scored highest 77.27% accuracy on the same aforementioned dataset by combining the effects of Heterogeneous Ensemble Learning (Het.EL), with Support vector Machine (SVM), Neural Network (NN), Naïve Bayes (NB) and K Nearest Neighbor (KNN), Minimal redundancy maximal relevance (mRMR). While Revati M.Wahul (2023) achieved 0.9245 Area Under Curve on the same aforementioned dataset with Ensemble Learning (EL).

Fakhar Bilal et al. (2022) achieved 92.43% and 94.7% accuracy on two datasets, Dataset1 and Dataset2, respectively, by applying the Ensemble Learning of Classification algorithms and clustering algorithms. Dataset 1 was based upon 20 features and 5000 instances. While Dataset 2 was based upon 21 features and 3333 instances. Mehpara Saghir (2019) achieved 80.8% accuracy on the same aforementioned dataset1 by applying Homogeneous Ensemble Learning (Hom.EL) with Neural Network and Bagging.

Table 1: Review of Literature

Reference	Classification	Data Optimization Techniques	Dataset	Results
Idris et al. (2012)	Ensemble Learning (RF, KNN)	PSO, PCA, mRMR, Fisher's Ratio, F sore	Orange (260/50000)	0.7511
Idris and Khan (2012)	Ensemble Learning (RotB, RotF)	PCA, Fisher's Ratio, F Score	Cell2Cell (76/40000)	AUC = 0.816
Adnan Idris et al. (2012)	Ensemble Learning (AB)	Genetic Programming	Orange (260/50000), Cell2Cell (76/40000)	AUC = 0.89
Idris et al. (2013)	Ensmeble Learning (RF, RotF, RotB, Decorate)	mRMR	Orange (260/50000), Cell2Cell (76/40000)	AUC = 0.699
Tianpei Xu (2021)	Ensemble Learning (DT, XGB, LR, NB)	Equi Distance	Orange (20/3333)	0.96%

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Adnan Idris and Khan (2017)	Ensemble Learning (RF, RotF, RotB and SVMs)	PSO, mRMR, Genetic Algorithm)	Orange (230/50000), Cell2Cell (77/40000)	Orange (AUC = 0.85), Cell2Cell (AUC = 0.82)
Uzair Ahmed et al. (2019)	CNN architectures, Ensemble Learning (GP, AB)	-	Orange (230/50000), Cell2Cell (77/40000)	Orange = 75.4%, Cell2Cell =68.2%
Fakhar Bilal et al. (2022)	Ensemble Learning (KNN, DT, GBT, RF, DL, NB) and Clustering Algorithms (K-Means, K-med, X-M, RC) with stacking	-	Datast1 (Github – 20/5000), Dataset2 (BigML – 21/3333)	Dataset1 = 94.7% Dataset2 =92.43%
Abinash Mishra (2017)	Ensemble Learning (RF),	Not mentioned	Orange (20/3333)	91.66
Amin et al. (2017)	Just-in-Time-Approach (Het.EL, Hom.EL) – SVM, NN, NB, KNN	Minimal redundancy maximal relevance (mRMR)	Churn-in-Telecom (21/3333)	Heterogeneous EL = 77.27%, Homogeneous EL = 59.24%
Mehpara Saghir (2019)	Ensemble Larning (NN + Bagging)	Not Mentioned	Churn Datase UCI (20/5000)	80.8%
Revati M.Wahul (2023)	Ensemble Learning	Exploratory Data Analytics and Feature Engineering	Orange (20 / 3333)	AUC = 0.9245

3. MATERIAL AND METHOD

In this section of the paper detailed procedure of Optimized Customer Churn Prediction (OCCP) model has presented. This section also provides details about the Orange dataset and the SMOTE based Ring-theory Particle Swarm Optimization (SRPSO) along with its algorithm and graphical representation.

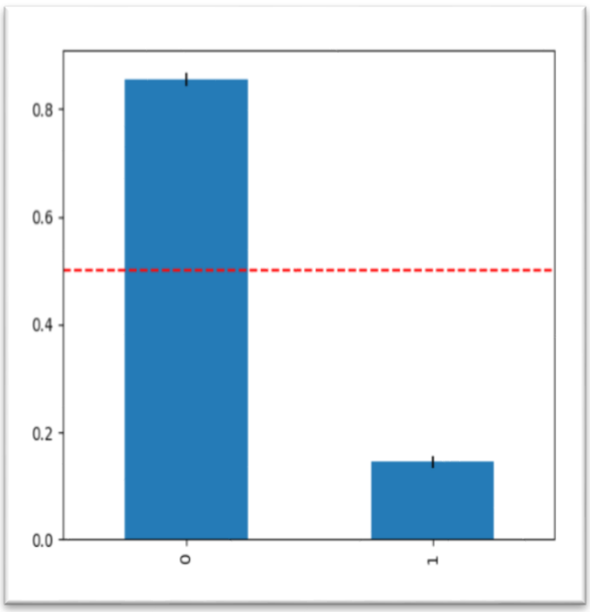
3.1 Imbalanced dataset

A Kaggle's Orange dataset of a telecommunication company was used to practice Optimized Customer Churn Prediction (OCCP) model. This dataset was consisting upon 3333 instances and 20 features Abinash Mishra (2017); Amin et al. (2019); Amin et al. (2017); (Atallah M. AL-Shatnwai 2020); Brandusoiu and Toderean (2013); Fakhar Bilal et al. (2022); (Hemlata Jain, 2020a); Irfan Ullah et al. (2019); Tjeng Wawan Cenggoro et al. (2021); (L. Y. Zhou et al., 2019).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

The division ratio of Orange dataset was (80:20) for training and testing dataset respectively. Figure 3 illustrates that the original data set was initially imbalanced and required to deal with imbalanced data issue. As minority class data samples are occurring below the red line.

Figure 4 reveals that only (14.5%) samples are from minority class, denoted as "churn = 1", whereas the remaining 85.5% samples are owned by majority class, denoted as "not churn = 0". So, it is essential to convert this imbalanced dataset into balanced.



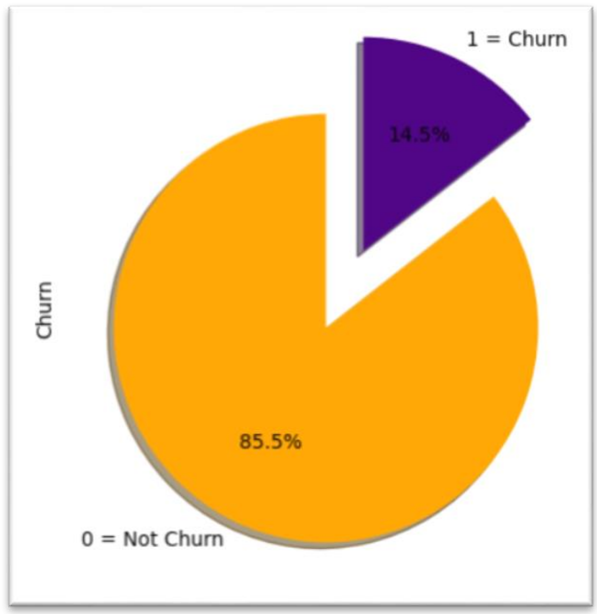


Figure 3: Imbalanced Dataset Visualization **Figure 4: Imbalanced Orange Dataset**

3.2 Optimized Customer Churn Prediction (OCCP)

Optimized Customer Churn Prediction (OCCP) model has designed to fulfill the requirements of a Telecommunication company. That's why, the proposed model has evaluated over Orange dataset, which is an opened source dataset of a telecommunication company, named as Orange, and available on Kaggle platform. Optimized Customer Churn Prediction (OCCP) possesses the capability to select the relevant features, to cope with the issue of data under sampling and oversampling. For this purpose, Forward Feature Selection, with Logistic Regression (LR) as learning algorithm, and Smote based Ring-theory Particle Swarm Optimization (SRPSO) has been introduced.

The performance of the above mentioned techniques have evaluated through Ensemble Learning (EL) with stacking. Figure 5 shows the conceptual structure of Optimized Customer Churn Prediction (OCCP) model in detail for comprehension. Optimized Customer Churn Prediction (OCCP) based upon in 8 steps. The detail of all the phases is enlisted below.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

3.3 Data preprocessing

Data Preprocessing is a cumbersome and rigorous process owing to outlier, redundancy and missing values in data. This phase is mandatory to make sure the validity, uniqueness and reliability of the dataset (Fan et al., 2021; Joshi & Patel, 2021). This phase can be further based upon data cleaning process, data transformation process and feature selection process.

Data cleaning process, also known as data auditing, is inevitable. Because, data of a telecommunication company is considered as a vital asset for its credibility to take fruitful decisions and to extract scientific insights, which gets spoiled due to data redundancy and inconsistency (Cohen et al., 2015; Fakhitah Ridzuan, 2019; Fatimah Sidi, 2012).

Figure 5 reveals that data cleaning process is the initial phase of Optimized Customer Churn Prediction (OCCP), focuses towards error identification such as outliers, missing values, redundancy and null values, error detection and error correction.

Figure 5 also reveals that data transformation process is the 2nd phase of Optimized Customer Churn Prediction (OCCP). In data transformation phase, all data types of all attributes such as float, object and Boolean, were transformed into integer data type to make sure the uniformity of dataset (Fakhitah Ridzuan, 2019; Manikandan, 2010) and to save storage and to make Machine Learning (ML) model feasible (Fakhitah Ridzuan, 2019; Joshi & Patel, 2021).

3.4 Feature Selection

Figure 5 reveals that Feature Selection process is the 3rd phase of Optimized Customer Churn Prediction (OCCP) development process, which removes anomalies, such as noisy and irrelevant data. It is essential to improve the performance, accuracy of customer churn prediction model and reduces the training time (Feng et al., 2023; S. Velliangiri, 2019). Because, too many irrelevant features of a data set can lead to bias predictions.

In Optimized Customer Churn Prediction (OCCP) model, Forward Feature Selection (FFS) (Erik Schaffernicht, 2009; Saifudin et al., 2020) with Logistic Regression (LR) has utilized to select 8 at most relevant features for tackling the problem of data over-fitting. This technique of Feature selection gave incredible boost to Optimized Customer Churn Prediction (OCCP) model.

ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

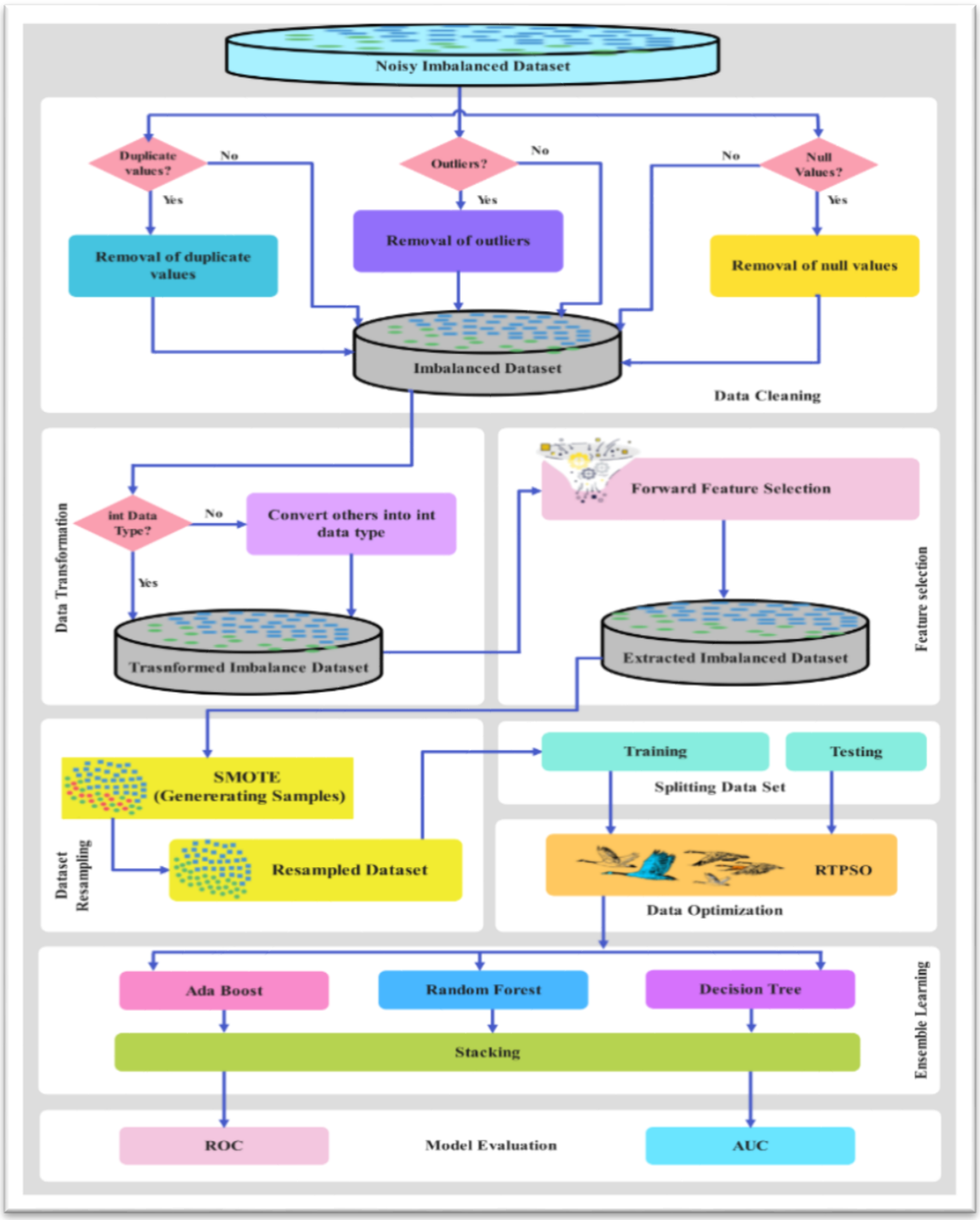


Figure 5: Optimized Customer Churn Prediction (OCCP) Model

3.5 Smote based Ring Particle Swarm Optimization (SRPSO) Fusion

As, data optimization techniques are the core keys to increase the overall quality of data and makes it relevant, efficient and scalable by reducing data size (Bian & Priyadarshi, 2024; Stergiou et al., 2023; Uddin et al., 2023; Wright, 2022). For this purpose, it tunes some hyper-parameter (Ansar & Kellis). To make Optimized Customer Churn Prediction (OCCP) model scalable, generalized, robust and efficient, a novice fusion (hybrid technique) of Synthetic Minority Oversampling Technique (SMOTE), Ring-Theory (RT)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

and Particle Swarm Optimization (PSO) has presented, which can control the class imbalance in an efficient way. This presented fusion has named as SMOTE based Ring-Theory based Particle Swarm Optimization (SRPSO). SRPSO can transform the dataset to turn it into coherent, reliable, computationally efficient and scalable for decision-making. It also improves the data analytics process for Ensemble Learning (EL).

The major pitfall of using SMOTE, an oversampling technique sometimes destroys the original dataset's distribution by resampling the imbalanced dataset (Ming Gao, 2011). SMOTE produces synthetic samples of minority class through interpolation. Though owing to interpolation the samples from majority class become invaded. Ultimately, classification process of customer churn suffers (Chawla et al., 2002). To cope with imbalanced data, a fusion of SMOTE based Ring Theory Particle Swarm Optimization (SRPSO) has practiced upon Orange dataset. In SRPSO algorithm, three methodologies, which are meta-heuristic such as Ring Theory (RT), SMOTE and Particle Swarm Optimization (PSO), have hybridized. This fusion has devised to gain improved and positive results without invading a few strongest and vital majority class samples rather than utilizing single technique at a time (Shaw et al., 2021).

This fusion possess the capability to choice the best synthetic majority class samples. Figure 6 portraits the conceptual framework of Synthetic Minority Oversampling Technique (SMOTE) based Ring-Theory Particle Swarm Optimization (SRPSO) for better understanding. According to Figure 5, data optimization process in Optimized Customer Churn Prediction (OCCP) model is performed in 3 basic steps. Figure 5 reveals that during 4th phase, data resampling process takes place after feature selection process, in which SMOTE resamples the imbalanced dataset's samples to cope with the issue of data under sampling. After data resampling process on 5th phase, the dataset gets split into ratio of 80:20 to form training and testing dataset. According figure 5, on 6th step the remaining SRPSO algorithm gets performed to deal with the negative effects of Synthetic Minority Oversampling Technique (SMOTE). Table 6 describes the followed algorithm of SRPSO phase of optimization. SRPSO works on 37 basic steps, which can be visualized in Figure 6. SRPSO's core contributions are as follows:

1. SRPSO contributes to reduce execution time by making the data scalable, coherent and manageable
2. SRPSO minimizes the resource usage by reducing the size of data to be used by Ensemble Learning
3. RTPSO improves the overall performance of the Optimized Customer Churn Prediction (OCCP)
4. SRPSO improves the data quality for taking insights to make customer churn predictions and for decision-making
5. SRPSO transforms the imbalanced dataset to make it relevant and uniformed

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Table 2: Step By Step Procedure of SRPSO

Steps	Algorithm
	SMOTE
Step 1	Start
Step 2	Input Imbalanced dataset
Step 3	Count the occurrence of majority and minority classes
Step 4	Calculate Difference between majority and minority class **Difference ← Total no. of majority class instances – total no. of minority class instances**
Step 5	Initialize **i ← 1**
Step 6	Check if i <= Difference? Then go to Step 7 otherwise to **Step 11**
Step 7	Select a sample from minority class randomly
Step 8	Add the sample to training data
Step 9	Increment in I by 1 **i ← i + 1**
Step 10	Go to Step 6
Step 11	Store in balanced dataset
Steps	**RT-PSO Algorithm**
Step 12	Start Optimization
Step 13	Initialize Population (pop), Velocity (v), Prbm, maxIter and i variables
Step 14	Calculate Fitness value of each particle, find Local best (pbest) and Global best (gbest)
Step 15	If i is greater than maxIter then go to **Step 36** otherwise go to **Step 16**
Step 16	Start (Particle Swarm Optimization (PSO))
Step 17	Calculate fitness value of each population (pop_i)
Step 18	Update Local best position (Pbest) for each particle if required
Step 19	Update Global best position (Gbest) for each particle if required
Step 20	Update Inertia weight (w)
Step 21	Update Velocity (v) and Position (x,y) of each particle
Step 22	Output of Particle Swarm Optimization (PSO) consider as input to Ring Theory based Evolutionary Algorithm (RTEA)
Step 23	Start Ring Theory based Evolutionary Algorithm (RTEA)
Step 24	Initialize k as **k = 1**
Step 25	If k is greater than popsize then goto **Step 26** otherwise **Step 27**
Step 26	Increase the value of i by 1 as **(i = i +1)** and go to **Step 15**
Step 27	Perform Global search operator (R-GEO) for Global Exploration in Ring
Step 28	Store the output of **Step 27** to p
Step 29	Perform Local Search Operator (R-LDO) for Local Exploitation in Ring on p and Prbm
Step 30	Store the output of **Step 29** into p again
Step 31	Calculate Fitness value of p by using fitness function → Fitness (p)
Step 32	Calculate Fitness value of $pop_k(t)$ by using fitness function → fitness ($pop_k(t)$)
Step 33	If Fitness value of p is greater than fitness value of $pop_k(t)$ then go to **Step 32** otherwise go to **Step 35**
Step 34	Assign value of p to $pop_k(t)$ then go to **Step 35**
Step 35	Increase the value of k by 1 (k → k+1) and go to **Step 25**
Step 36	Return gbest
Step 37	End Optimization

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

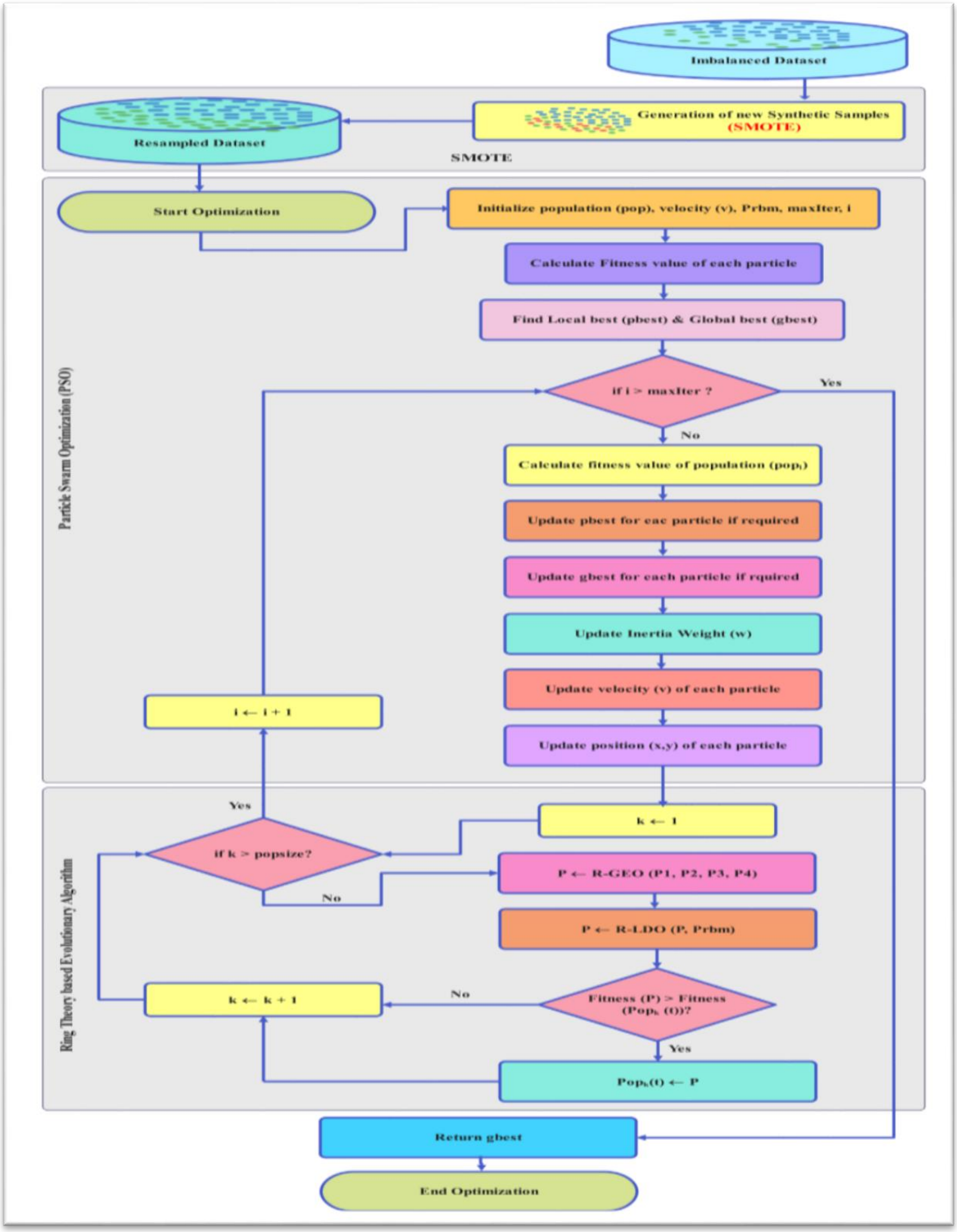


Figure 6: SMOTE based Ring-theory Particle Swarm Optimization (SRPSO) Fusion

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

3.6 Ensemble Learning

Ensemble Learning (EL), fundamental data optimization approach, contributes to control the variance and bias and gives an optimized solution for Evolutionary Optimization Algorithms (EOA). Ensemble Learning (EL) exploits the 4 main meta learner approaches, such as boosting, voting, stacking and begging, to mend the predictive power of a customer churn prediction model by improving its performance in terms of accuracy (Mahajan et al., 2023; Nazmun Nahar, 2019; Tufail et al., 2023; Zubair Hasan & Zahid Hasan, 2019). Ensemble Learning endeavors to combine the predictive power of more than one weak classifiers with stacking for generating precise and rational forecast and eradicates the prediction errors (Alshdaifat et al., 2021; Diwan et al., 2022; Lu et al., 2023; Mohammed & Kora, 2023; Nti et al., 2020; Redha Ali, 2019; Sagi & Rokach, 2018; Z.-H. Zhou, 2017).

There are two basic types of Ensemble Learning (EL) such as Heterogeneous Ensemble Learning (Het.EL) and Homogeneous Ensemble Learning (). In Hom.EL, Base Classifiers are, always, of same typed, while in Het.EL Base Classifiers are, always, of different types. In Heterogeneous Ensemble Learning (Het.EL), different base classifiers generate bias predictions, by using which Meta Learner become proficient to diminish overall bias of final generated prediction as well (Saleh et al., 2022). Figure 5 reveals that Optimized Customer Churn Prediction (OCCP) model is based upon Heterogeneous Ensemble Learning (HEL) with stacking (meta learner), during 7th step. For this purpose, Random Forest (RF) (Schonlau & Zou, 2020; Solorio-Ramírez et al., 2023), Decision Tree (DT) (Ibomoiye Domor Mienye, 2019; Song & Lu, 2015) and AdaBoost (AB) (Ding et al., 2022; Hornyák & Iantovics, 2023) have used as weak learners.

3.7 Performance Evaluation

After developing the customer churn prediction model 8th step, the model gets evaluated through quantitative evaluation or qualitative evaluation. End user, n qualitative evaluation, provides reviews about the performance of the customer churn prediction model. Whereas in quantitative evaluation, the overall performance of the customer churn prediction model gets quantify in numbers (Dalianis, 2018). In this research work, the quantitative evaluation has adopted to evaluate and weigh the performance of Optimized Customer Churn Prediction (OCCP). For this purpose, Accuracy, Kappa Statistics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of the Model have been calculated.

Table 3 reveals aforementioned performance evaluation metrics along with their equations. Eq.1 and Eq.2 in Table 3 represents the formula to calculate confusion matrix and accuracy, respectively. While, Eq. 3 and Eq.4 represents the formula of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) respectively. Hence, Figure 7 presents the complete performance evaluation stick of Cohen' Kappa Statistics, which can be denoted by Greek letter "k". it can measure the interrater reliability. Hence, the degree of agreement between more than one evaluators is called

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

interrater reliability (Nichols et al., 2010; Vieira et al., 2010). It determines the rate of consistency among all measurements spawned by all evaluators. Its range can be from -1 to +1. If all the observers get perfectly agreed upon all objects, it means the interrater reliability is perfect (Cross, 1996; McHugh, 2012; Xu & Lorber, 2014).

Table 3: List of Performance Valuation Matrices

Performance Evaluation Matrices	Equation	Eq. No:
Confusion Matrix *(Smirani et al., 2022)*	$= \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$	Eq. 1
Accuracy *(Smirani et al., 2022)*	$= \frac{TN + TP}{TN + FP + FN + TP}$	Eq. 2
Root Mean Square Error (Schubert et al., 2017)	$= \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$	Eq. 3
Mean Absolute Error (Chai & Draxler, 2014; Karunasingha, 2022)	$= \frac{\sum_{i=1}^{n} \lvert Actual_i - Predicted_i \rvert}{n}$	Eq. 4

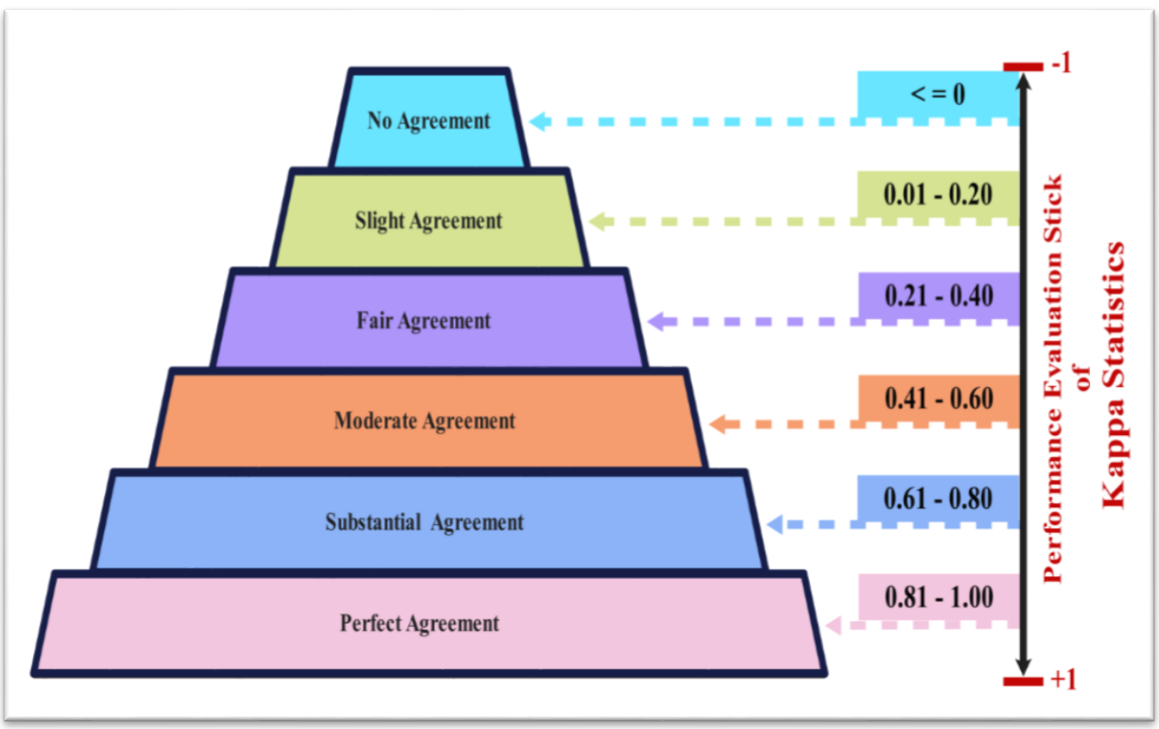


Figure 7: Complete range of Cohen's Kappa Statistics

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

4. RESULT AND DISCUSSION

4.1 Novelty of OCCP in context of Ensemble Learning

In this section of the paper, the total novelty of this research work, in context of Ensemble Learning, has discussed. The literature review reveals that a lot of work has done through Ensemble Learning (EL) in the field of customer churn prediction. Unfortunately, there is still need to reform this technique to improve its accuracy in such customer churn prediction models. The literature review also reveals that the researchers have contributed a lot in this regard. They tried their level best to improve the accuracy and performance of their models by creating fusions of various techniques. After observing their pits and falls, a robust and efficient model, named as Optimized Customer Churn Prediction (OCCP), has been devised to anticipate the risky customers, who can nearly in future turn their loyalty to another telecommunication company. Here in the Table 1 and 2, enlisted below, the total novelty of the OCCP has been revealed in contrast to other existed models, discussed in literature review. Under the umbrella term of Ensemble Learning, this contrast can been be discussed in more broader terms, such as Supervised Machine Learning Algorithms (SMLA) for Ensemble Learning (EL), Unsupervised Machine Learning Algorithms (UMLA) for Ensemble Learning (EL), Optimization Algorithms, Feature Selection Algorithms and Feature Extraction Algorithms.

4.2 Novelty of Ensemble Learning (EL) in terms of Classification Algorithms

In this context, the Table 4 presents total novelty of OCCP in terms of Supervised Machine Learning Algorithms (SMLA) and Unsupervised Machine Learning Algorithms (UMLA). The most prominent supervised and unsupervised classification algorithms of Machine Learning (ML) have been enlisted below in Table 4, such as Stochastic Gradient Booting (SGD), Deep Convolutional Neural Network (DCNN), Gaussian Process Regression (GPR), Support Vector regression (SVR), Regression Tree (RT), Ensemble Regression (ER), Naïve Bayes (NB), Decorate (D), AdaBoost (AB), Support Vector Machine (SVM), Decision Tree (DT), Extreme Gradient Boosting (XGB), Gradient Boosting Tree (GBT), K Nearest Neighbor (KNN), Decision Learning (DL), K Means (KM), K Medoids (KMed), Regression Clustering (RC), X Means (XM), Logistic Regression (LR), Random Forest (RF), Rotation Forest (RotF), Rotation Boosting (RotB), Genetic Programming (GP) and Artificial Neural Network (ANN). It can be observed from Table 4 that the OCCP is quite novice model, because AdaBoost (AB), Decision Tree (DT) and Random Forest (RF) have been ensemble in OCCP for the first time. This combination of classification is quite novice and has not been yet.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Table 4: Novelty of Ensemble Learning (EL) In Terms of Classification Algorithms

Ref.	SMLA																								UMLA	
	SGD	DCNN	GPR	SVR	RT	ER	NB	D	AB	RT	SVM	DT	XGB	GBT	KNN	DL	KM	KMed	RC	XM	LR	RF	RotF	RotB	GP	ANN
Adnan Idris and Khan (2017)									✓													✓	✓		✓	
Idris and Khan (2012)																							✓	✓		
Adnan Idris et al. (2012)									✓																✓	
Amin et al. (2017)											✓															
Idris et al. (2013)								✓														✓	✓	✓		
Hussain et al. (2019)			✓	✓	✓	✓																				
Tianpei Xu (2021)							✓					✓	✓								✓					
Revati M.Wahul (2023)	✓								✓					✓								✓				
Pratiwi et al. (2021)		✓																								
Mehpara Saghir (2019)																										✓
Amin et al. (2017)							✓				✓				✓											✓
Abinash Mishra (2017)																						✓				
Fakhar Bilal et al. (2022)							✓					✓		✓	✓	✓	✓	✓	✓	✓	✓					
Uzair Ahmed et al. (2019)									✓																✓	
OCCP									✓			✓										✓				

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Novelty of Ensemble Learning (EL) in terms of Optimization and Feature Reduction Techniques

The Table 5 illustrates the total novelty of OCCP in terms of techniques and feature reduction techniques. Table 5 is consisted upon a few references of the papers, which got published under the umbrella term of Ensemble Learning (EL) Technique. By observing this Table 5 it is concluded that a very less work has been conducted in the area of Machine Learning (ML) model's optimization, feature selection and feature extraction. The literature review reveals that the most prominent techniques of model optimization are Particle Swarm Optimization (PSO), Ring Theory (RT) and SMOTE. It also reveals that most commonly techniques for feature selection are minimum Redundancy and Maximum Relevancy (mRMR) and Genetic Programming (GP). While, Forward Feature Selection (FFS) has not been used in existing studies of Ensemble Learning (EL). It also reveals that the SMOTE, Ring theory (RT) and Particle Swarm Optimization (PSO) have not been hybridized yet.

Table 5: Novelty of Ensemble Learning (EL) In Terms of Optimization and Feature Reduction Techniques

Reference	Optimization			Feature Selection Algorithm							Feature Extraction	
	PSO	RT	SMOTE	mRMR	FFS (LR)	EDA	WE	GP	FR	FS	PCA	ED
Adnan Idris and Khan (2017)	✓	-	-	-	-	-	-	✓	-	-	-	-
Idris and Khan (2012)	-	-	-	✓	-	-	-	-	✓	✓	✓	-
Adnan Idris et al. (2012)	-	-	-	-	-	-	-	✓	-	-	-	-
Amin et al. (2017)	-	-	-	✓	-	-	-	-	-	-	✓	-
Idris et al. (2013)	-	-	-	✓	-	-	-	-	-	-	-	-
Hussain et al. (2019)	-	-	-	-	-	-	✓	-	-	-	-	-
Tianpei Xu (2021)	-	-	-	-	-	-	-		-	-	-	✓
Revati M.Wahul (2023)	-	-	-	-	-	✓	-	-	-	-	-	-
Pratiwi et al. (2021)	-	-	-	-	-	-	-	-	-	-	-	-
Mehpara Saghir (2019)	-	-	-	-	-	-	-		-	-	-	-
Amin et al. (2017)	-	-	-	✓	-	-	-		-	-	-	-
Abinash Mishra (2017)	-	-	-	-	-	-	-	-	-	-	-	-
Fakhar Bilal et al. (2022)	-	-	-	-	-	-	-	-	-	-	-	-
Uzair Ahmed et al. (2019)	-	-	-	-	-	-	-	-	-	-	-	-
(OCCP)	✓	✓	✓		✓	-	-	-	-	-	-	-

Performance Evaluation of OCCP in terms of Accuracy, Kappa Statistics and Error evaluation

In this section of the paper, the performance Optimized Customer Churn Prediction (OCCP) has benchmarked against standalone models by Hemlata Jain (2020b) in terms of accuracy, root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Kappa Statistics. For this purpose Table 6 has presented for comprehension. Table 6 reveals

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

that Optimized Customer Churn Prediction (OCCP) model scored highest score than both stand-alone models of Hemlata Jain (2020b), Logit Boost (LogB) and Logistic Regression (LogR). Optimized Customer Churn Prediction (OCCP) scored 0.940 Kappa Statistics, 0.0298 Mean Absolute Error, 0.173 Root Mean Square Error and 0.97 Accuracy.

According to Kappa Statistics evaluation stick, Optimized Customer Churn Prediction (OCCP) stands 1st by achieving Prefect Agreement tag. It also stands 1st in terms of accuracy. Graphical representation of Table1 has presented below from Figure 8-11. Figure 8 visualizes the Kappa Statistics score, Figure 9 depicts Mean Absolute Error, Figure 10 symbolizes the Root Mean Square Error (RMSE) and Figure 11 visualizes the Accuracy of Optimized Customer Churn Prediction (OCCP).

Table 6: Performance Evaluation against Stand-Alone Models.

Evaluation Matrices	Logit Boost (Hemlata Jain, 2020b)	Logistic Regression (Hemlata Jain, 2020b)	OCCP (EL) (Proposed)
Kappa Statistics	0.0956	0.1176	0.940
Mean Absolute Error	0.2261	0.2237	0.0298
Root Mean Square Error	0.3357	0.3352	0.173
Accuracy	85.1785%	85.2385%	0.97

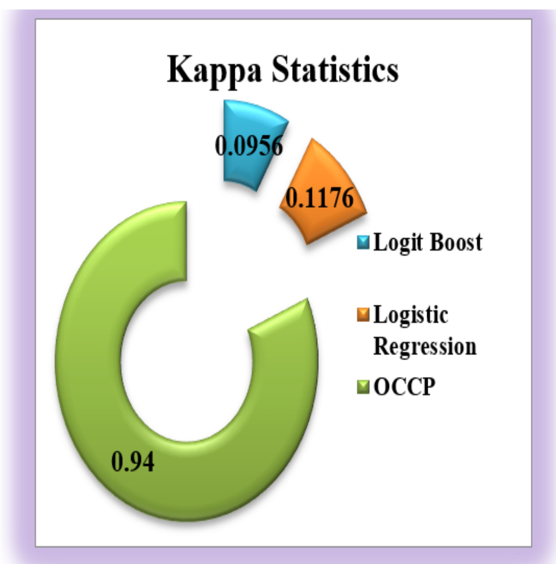


Figure 8: Performance evaluation of OCCP in terms of Kappa statistics

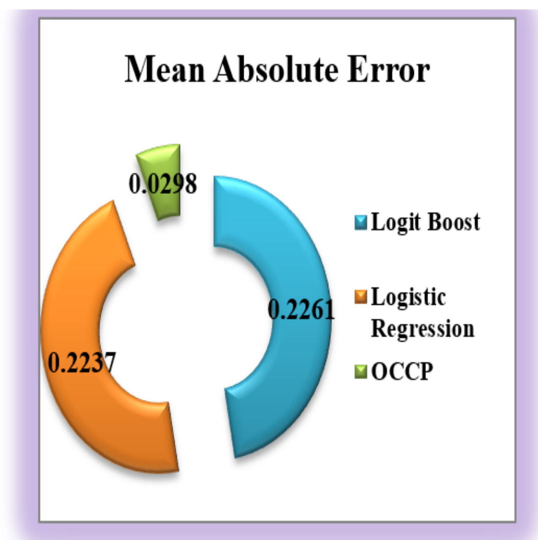


Figure 9: Performance evaluation of OCCP in terms of Mean Absolute Error

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

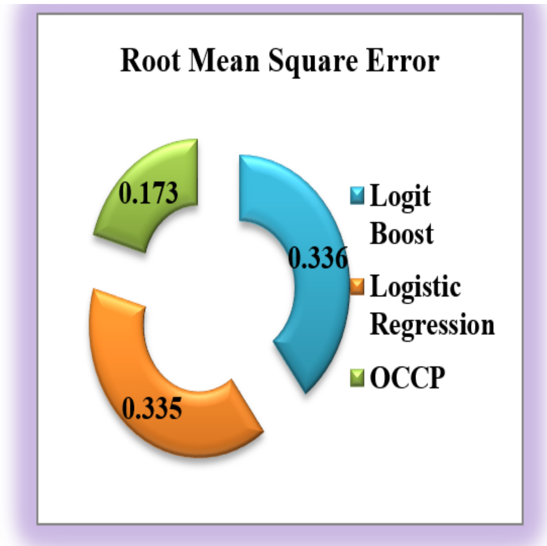


Figure 10: Performance evaluation of OCCP in terms of Root Mean Squared Error

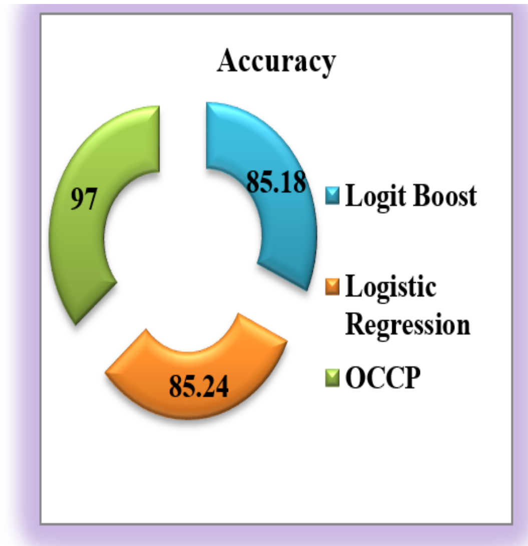


Figure 11: Performance evaluation of OCCP in terms of Accuracy

4.3 Benchmark against more studies in terms of accuracy

In this section of the paper, Optimized Customer Churn Prediction (OCCP) has benchmarked against more studies. Table 7 presents the summary of this section in detail. According to table 7, Amin et al. (2019) scored 84.30% accuracy with technique of Estimation of certainty of classifiers through Distance factor. While Tjeng Wawan Cenggoro et al. (2021) scored 89.82% accuracy by applying Deep Learning and Cluster Similarity technique. Hence, Brandusoiu and Toderean (2013) 88.56% accuracy by applying Support Vector Machine(SVM) with Polynomial Kernel). Hemlata Jain (2020b) scored 85.24% accuracy by applying Logistic Regression and Logit Boost as standalone models. Thus, Irfan Ullah et al. (2019) scored 88.63% accuracy with classification of Random Forest (RF). L. Y. Zhou et al. (2019) achieved 91.67% accuracy by using Support Vector Machine with Polynomial Kernel Function. Amin et al. (2017) recorded 77.27% accuracy by using Heterogeneous Ensemble Learning (Het.EL). Abinash Mishra (2017) scored 91.66% accuracy by using Ensemble Learning with Random Forest (RF) classifier and Bagging. Fakhar Bilal et al. (2022) scored 92.43% accuracy by hybridizing Classification and Clustering Algorithms. While Optimized Customer Churn Prediction (OCCP) achieved 97% accuracy, which is the highest accuracy achieved yet. Figure 12 also visualizes the Table 7 for better understanding.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

Table 7: Benchmark Studies

Refference	Classification	Accuracy
Amin et al. (2019)	Estimation of certainty of classifiers through Distance factor	84.30%
Tjeng Wawan Cenggoro et al. (2021)	Deep Learning and Cluster Similarity	89.82%
Brandusoiu and Toderean (2013)	Support Vector Machine (Polynomial Kernel)	88.56%
Hemlata Jain (2020b)	Logistic Regression and Logit Boost	85.24%
Irfan Ullah et al. (2019)	Random Forest	88.63%
L. Y. Zhou et al. (2019)	Support Vector Machine (Polynomial Kernel Function)	91.67%
Amin et al. (2017)	Just-in-Time-Approach (Heterogeneous EL)	77.27%
Abinash Mishra (2017)	Ensemble Learning (Random Forest), Bagging, Boosting	91.66%
Fakhar Bilal et al. (2022)	Hybrid Model of Classification and Clustering Algorithms with stacking	92.43%
Proposed (OCCP)	Ensemble Learning	97.00%

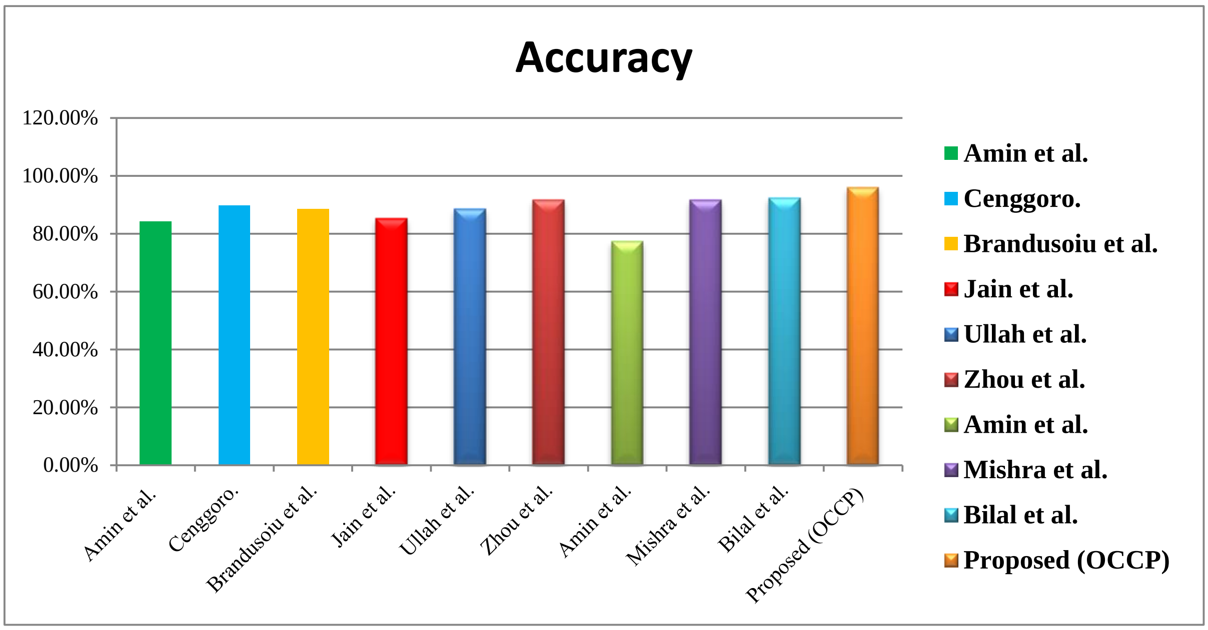


Figure 12: Optimized Customer Churn Prediction (OCCP) benchmark against existing studies

5. CONCLUSION

In this era of fierce competition, the telecommunication companies are required to anticipate their risky customers, who can switch to competitors nearly in future, in advance to increase their revenue. This study is basically a contribution to cope with aforementioned problem of telecommunication companies. The results of this study reveals that Optimized Customer Churn Prediction (OCCP) model performed well by embedding the fusion of optimization techniques, named as Smote based Ring-theory Particle Swarm Optimization (SRPSO). This fusion contributed to make dataset more concise, scalable, relevant, coherent, manageable, uniform and consistent. It is also concluded that Ensemble Learning can perform well by hybridizing with optimization techniques and can achieve the highest level of accuracy up to 97%. In future, a few more optimization techniques can be combined with Ensemble Learning (EL) to improve its results and can be experimented over big data platform.

Reference

1) Abinash Mishra, U. S. R. (2017). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. *Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)*, 721-725. doi: 10.1109/ICICI.2017.8365230

2) Adnan Idris, Asifullah Khan, & Lee, Y. S. (2012). Genetic Programming and Adaboosting based churn Prediction for Telecom. *IEEE International Conference*, 1328-1332.

3) Adnan Idris, & Khan, A. (2017). Churn Prediction System for Telecom using Filter–Wrapper and Ensemble Classification. *The Computer Journal, IEEE, 60*(3), 410-430. doi: 10.1093/comjnl/bxv123

4) Aida Ali, S. M. S., Anca Ralescu. (2015). Classification with class imbalance problem A review. *Int. J. Advance Soft Compu. Appl, Research Gate, 5*(3), 1-31.

5) Akter, N. (2022). Service Quality and Customer Retention in COVID-19 Period at Telecom Industry. *Social Science Research Network (SSRN)*, 9.

6) Alshdaifat, E. a., Al-hassan, M., & Aloqaily, A. (2021). Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers. *ICT Express, 7*(3), 342-349. doi: 10.1016/j.icte.2020.11.005

7) Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research, 94*, 290-301. doi: 10.1016/j.jbusres.2018.03.003

8) Amin, A., Al-Obeidat, F., Shah, B., Tae, M. A., Khan, C., Durrani, H. U. R., & Anwar, S. (2017). Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing, 76*(6), 3924-3948. doi: 10.1007/s11227-017-2149-9

9) Andrews, R. (2019). Churn Prediction in Telecom Sector Using Machine Learning. *International Journal of Information Systems and Computer Sciences, 8*(2), 132-134. doi: 10.30534/ijiscs/2019/31822019

10) Ansar, M., & Kellis, M. Machine Learning and Optimization Methods for Enhancing Information Extraction in Big Data Analysis. *integration, 1*, 2.

11) Atallah M. AL-Shatnwai , M. F. (2020). Predicting Customer Retention using XGBoost and Balancing Methods. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(7), 704-712. doi: 10.14569/IJACSA.2020.0110785

12) Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Inf Sci Syst, 6*(1), 9. doi: 10.1007/s13755-018-0051-3

13) Bian, K., & Priyadarshi, R. (2024). Machine learning optimization techniques: a Survey, classification, challenges, and Future Research Issues. *Archives of Computational Methods in Engineering*, 1-25.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

14) Brandusoiu, I., & Toderean, G. (2013). Churn Prediction in the Telecommunications Sector Using Support Vector Machines. *ANNALS OF THE ORADEA UNIVERSITY. Fascicle of Management and Technological Engineering., XXII (XII), 2013/1*(1). doi: 10.15660/auofmte.2013-1.2772

15) Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw, 106*, 249-259. doi: 10.1016/j.neunet.2018.07.011

16) Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions, 7*(1), 1525-1534.

17) Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. doi: 10.1613/jair.953

18) Cohen, B., Vawdrey, D. K., Liu, J., Caplan, D., Furuya, E. Y., Mis, F. W., & Larson, E. (2015). Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings. *Policy Polit Nurs Pract, 16*(3-4), 117-124. doi: 10.1177/1527154415603358

19) Cross, S. (1996). Kappa statistics as indicators of quality assurance in histopathology and cytopathology. *Journal of clinical pathology, 49*(7), 597-599.

20) Dalianis, H. (2018). Evaluation Metrics and Evaluation. 45-53. doi: 10.1007/978-3-319-78503-5_6

21) Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification. *Applied Sciences, 12*(12), 5872. doi: 10.3390/app12125872

22) Diwan, S., Thakur, G. S., Sahu, S. K., Sahu, M., & Swamy, N. K. (2022). Predicting Heart Diseases through Feature Selection and Ensemble Classifiers. *Journal of Physics: Conference Series, 2273*(1), 012027. doi: 10.1088/1742-6596/2273/1/012027

23) Dunja Mladenić, M. G. (1999). Feature selection for unbalanced class distribution and Naïve Bayes. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia*, 1-10.

24) Ebenuwa, S. H., Sharif, M. S., Alazab, M., & Al-Nemrat, A. (2018). Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE ACCESS, 7*, 24649 - 24666.

25) Erik Schaffernicht, C. M., Klaus Debes, Horst-Michael Gross. (2009). 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*, 583-588.

26) Fakhar Bilal, S., Ali Almazroi, A., Bashir, S., Hassan Khan, F., & Ali Almazroi, A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Comput Sci, 8*, e854. doi: 10.7717/peerj-cs.854

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

28) Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research, 9*. doi: 10.3389/fenrg.2021.652801

29) Fareniuk, Y., Zatonatska, T., Dluhopolskyi, O., & Kovalenko, O. (2022). Customer churn prediction model: a case of the telecommunication market. *Economics, 10*(2), 109-130. doi: 10.2478/eoik-2022-0021

30) Fatimah Sidi, P. H. S. P., Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, Aida Mustapha. (2012). Data quality: A survey of data quality dimensions. *2012 International Conference on Information Retrieval & Knowledge Management, IEEE*, 300-304.

31) Feng, W., Gou, J., Fan, Z., & Chen, X. (2023). An ensemble machine learning approach for classification tasks using feature generation. *Connection Science, 35*(1). doi: 10.1080/09540091.2023.2231168

32) Hemlata Jain, A. K., Sumit Srivastava. (2020a). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Procedia Computer Science ScienceDirect, 167*, 101-112. doi: https://doi.org/10.1016/j.procs.2020.03.187

33) Hemlata Jain, A. K., Sumit Srivastava. (2020b). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science, 167*, 101–112.

34) Hornyák, O., & Iantovics, L. B. (2023). AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics. *Mathematics, 11*(8), 1801. doi: 10.3390/math11081801

35) Hussain, L., Saeed, S., Idris, A., Awan, I. A., Shah, S. A., Majid, A., . . . Chaudhary, Q. A. (2019). Regression analysis for detecting epileptic seizure with different feature extracting strategies. *Biomed Tech (Berl), 64*(6), 619-642. doi: 10.1515/bmt-2018-0012

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

37) Idris, A., & Khan, A. (2012). Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers. 23-27. doi: 10.1109/inmic.2012.6511498

38) Idris, A., Khan, A., & Lee, Y. S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence, 39*(3), 659-672. doi: 10.1007/s10489-013-0440-x

39) Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Computers & Electrical Engineering, 38*(6), 1808-1819. doi: 10.1016/j.compeleceng.2012.09.001

40) Irfan Ullah, Basit Raza, Saif Ul Islam, Ahmad Kamran Malik , Muhammad Imran, & Wonkim, S. (2019). A Churn Prediction Model Using Random Forest Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification. *IEEE ACCESS, 7*, 60134 - 60149.

41) Jason Van Hulse , T. M. K., Amri Napolitano (2007). Experimental Perspectives on Learning from Imbalanced Data. *ACM Digital Library*, 935 - 942. doi: 10.1145/1273496.1273614

42) Joshi, A. P., & Patel, B. V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental journal of computer science and technology, 13*(0203), 78-81. doi: 10.13005/ojcst13.0203.03

43) Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences, 585*, 609-629.

44) Khoshgoftaar, T. M., Seiffert, C., Hulse, J. V., Napolitano, A., & Folleco, A. (2007). Learning with limited minority class data. 348-353. doi: 10.1109/icmla.2007.76

45) Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data, 5*(1). doi: 10.1186/s40537-018-0151-6

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

46) Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting. *Water, 15*(7), 1265. doi: 10.3390/w15071265

47) Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A Review. *Healthcare (Basel), 11*(12). doi: 10.3390/healthcare11121808

48) Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing, 27*, 504-518. doi: 10.1016/j.asoc.2014.11.023

49) Manikandan, S. (2010). Data transformation. *J Pharmacol Pharmacother, 1*(2), 126-127. doi: 10.4103/0976-500X.72373

50) McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia media, 22*(3), 276–282.

51) Mehpara Saghir, Z. B., Saba Bashir, Farhan Hassan Khan. (2019). Churn Prediction using Neural Network based Individual and Ensemble Models. *Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan*, 634 -639.

52) Ming Gao, X. H., Sheng Chen , Chris J. Harris. (2011). On combination of SMOTE and particle swarm optimization based radial basis function classifier for imbalanced problems. *Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA*, 1146-1153

53) Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences, 35*(2), 757-774. doi: 10.1016/j.jksuci.2023.01.014

54) Naz, F., Alshaabani, A., Rudnák, I., & Magda, R. (2021). Role of Service Quality in Improving Customer Loyalty towards Telecom Companies in Hungary during the COVID-19 Pandemic. *Economies, 9*(4), 200. doi: 10.3390/economies9040200

55) Nazmun Nahar, F. A., Md. Arif Istiek Neloy, Vicky Barua, Mohammad Shahadat Hossain, Karl Andersson. (2019). A Comparative Analysis of the Ensemble Method For Liver Disease Prediction. *International Conference on Innovation in Engineering and Technology (ICIET), IEEE*, 1-6.

56) Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. *The Quality Assurance Journal, 13*(3-4), 57-61.

57) Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data, 7*(1). doi: 10.1186/s40537-020-00299-5

58) Pratiwi, R. A., Nurmaini, S., Rini, D. P., Rachmatullah, M. N., & Darmawahyuni, A. (2021). Deep ensemble learning for skin lesions classification with convolutional neural network. *IAES International Journal of Artificial Intelligence (IJ-AI), 10*(3), 563. doi: 10.11591/ijai.v10.i3.pp563-570

59) Redha Ali, R. C. H., Barath Narayanan Narayanan, Supun De Silva. (2019). Deep learning ensemble methods for skin lesion analysis towards melanoma detection. *2019 IEEE National Aerospace and Electronics Conference (NAECON), IEEE*, 311-316.

60) Revati M.Wahul, A. P. K., Prabhakar N.Kota. (2023). An Ensemble Learning Approach to Enhance Customer Churn Prediction in Telecom Industry. *International Journal of Intelligent System And Application In Engineering, 11*(9), 258–266.

61) S. Velliangiri, S. A., S Iwin Thankumar joseph (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science, Science Direct, 165*, 104-111.

62) Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery, 8*(4). doi: 10.1002/widm.1249

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

63) Saha, S. S. S. (2023). Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. *SPRINGER Nature Journal, SN Applied Sciences, 5*(173), 1-14.

64) Saifudin, A., Ekawati, Yulianti, & Desyani, T. (2020). Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes. *Journal of Physics: Conference Series, 1477*(3), 032007. doi: 10.1088/1742-6596/1477/3/032007

65) Saleh, H., Mostafa, S., Alharbi, A., El-Sappagh, S., & Alkhalifah, T. (2022). Heterogeneous Ensemble Deep Learning Model for Enhanced Arabic Sentiment Analysis. *Sensors (Basel), 22*(10). doi: 10.3390/s22103707

66) Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata, 20*(1), 3-29. doi: 10.1177/1536867x20909688

67) Schubert, A.-L., Hagemann, D., Voss, A., & Bergmann, K. (2017). Evaluating the model fit of diffusion models with the root mean square error of approximation. *Journal of Mathematical Psychology, 77*, 29-45.

68) Shahzad, A., Yaqub, R. M. S., Di Vaio, A., & Hassan, R. (2021). Antecedents of customer loyalty and performance improvement: Evidence from Pakistan's telecommunications sector. *Utilities Policy, 70*, 101208. doi: 10.1016/j.jup.2021.101208

69) Shaw, S. S., Ahmed, S., Malakar, S., Garcia-Hernandez, L., Abraham, A., & Sarkar, R. (2021). Hybridization of ring theory-based evolutionary algorithm and particle swarm optimization to solve class imbalance problem. *Complex & Intelligent Systems, 7*(4), 2069-2091. doi: 10.1007/s40747-021-00314-z

70) Smirani, L. K., Yamani, H. A., Menzli, L. J., Boulahia, J. A., & Huang, C. (2022). Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths. *Scientific Programming, 2022*, 1-15. doi: 10.1155/2022/3805235

71) Solorio-Ramírez, J.-L., Jiménez-Cruz, R., Villuendas-Rey, Y., & Yáñez-Márquez, C. (2023). Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects. *Algorithms, 16*(6), 293. doi: 10.3390/a16060293

72) Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry, 27*(2), 130-135. doi: 10.11919/j.issn.1002-0829.215044

73) Srinivasan, R., Rajeswari, D., & Elangovan, G. (2023). Customer Churn Prediction Using Machine Learning Approaches. 1-6. doi: 10.1109/iceconf57129.2023.10083813

74) Stergiou, K., Ntakolia, C., Varytis, P., Koumoulos, E., Karlsson, P., & Moustakidis, S. (2023). Enhancing property prediction and process optimization in building materials through machine learning: A review. *Computational Materials Science, 220*, 112031.

75) Tianpei Xu, Y. M. a. K. K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. *Applied Siences, MDPI, 11*(4742), 1-12. doi: 10.3390/ 10.3390/app11114742

76) Tjeng Wawan Cenggoro, Raditya Ayu Wirastari, Edy Rudianto, Mochamad Ilham Mohadi, Dinne Ratj, & Pardamean, B. (2021). Deep Learning as a Vector Embedding Model for Customer Churn. *Procedia Computer Science 179 (2021) 624–63, 179* 624–663.

77) Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics, 12*(8), 1789. doi: 10.3390/electronics12081789

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 09:2024
DOI: 10.5281/zenodo.13859395

78) Uddin, Mohi, K. M., Biswas, N., Rikta, S. T., & Dey, S. K. (2023). Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Computer Methods and Programs in Biomedicine Update, 3*, 100098.

79) Uzair Ahmed, Asifullah Khan, Saddam Hussain Khan, Abdul Basit, I. U. H., & Lee, Y. S. (2019). Transfer Learning and Meta Classification Based Deep Churn System for Telecom Industry. *arXiv.org*, 1-10.

80) Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010). *Cohen's kappa coefficient as a performance measure for feature selection.* Paper presented at the International conference on fuzzy systems.

81) Wright, S. J. R., Benjamin. (2022). *Optimization for data analysis*: Cambridge University Press.

82) Xu-Ying Liu, J. W., and Zhi-Hua Zhou. (2019). Exploratory undersampling for class-imbalance learning. *IEEE Transactions On Systems, Man And Cybernetics – PART B, 39*(2), 1 - 14.

83) Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of consulting and clinical psychology, 82*(6), 1219.

84) Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing, 105*, 3-11. doi: 10.1016/j.neucom.2012.04.039

85) Zhaohui Zheng, X. W., Rohini Srihari. (2014). Feature selection for text categorization on imbalanced data. *Explor Newsletter, 6*(1), 81-89.

86) Zhou, L. Y., M. Amoh, D., K. Boateng, L., & A. Okine, A. (2019). Combined Appetency and Upselling Prediction Scheme in Telecommunication Sector Using Support Vector Machines. *International Journal of Modern Education and Computer Science, 11*(6), 1-7. doi: 10.5815/ijmecs.2019.06.01

87) Zhou, Z.-H. (2017). Ensemble Learning. *Encyclopedia of Biometrics, SPRINGER*, 270–273.

88) Zubair Hasan, K. M., & Zahid Hasan, M. (2019). Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease. *882*, 415-426. doi: 10.1007/978-981-13-5953-8_34