University of Ibadan
Journal of Science and Logics in ICT Research

Customer Churn Prediction in Telecommunications Using Ensemble Technique

[1]✉Oladipo, I. D., [2]Awotunde, J. B., [3]AbdulRaheem, M., [4]Taofeek-Ibrahim, F. A., [5]Obaje, O. and [6]Ndunagu, J. N.

[1, 3, 5]*Department of Computer Science, University of Ilorin, Ilorin, Nigeria*
[4]*Department of Computer Science, The Federal Polytechnic Offa, Nigeria*
[6]*Department of Computer Science, National Open University of Nigeria, Abuja, Nigeria*
[1]*iodidowu@unilorin.edu.ng;* [2]*awotunde.jb@unilorin.edu.ng,* [3]*muyideen@unilorin.edu.ng,*
[4]*fatty_fatty2@yahoo.com.au,* [5]*danielgroove96@gmail.com,* [6]*jndunagu@noun.edu.ng*

Abstract

The telecommunications industry understands the hyperlink between consumer pleasure and revenue. Customer attrition is a term used in the telecommunications industry to describe when a customer-company relationship ends. The churn forecast has recently intrigued stakeholders in the telecommunications industry. They learned that retaining customers is much more cost-effective than acquiring new ones. This paper presents an ensemble-based telecom churn predictive model of machine learning (ML) algorithms such as XGBoost (extreme gradient boost), LightGBM, Random Forest (RF), and CatBoost. Using ML-based models for predictive analytics is very important in the telecommunications industry when it comes to predicting customer attrition. The ensemble method model helps the telephone company predict if a customer is likely to cancel. In addition, several algorithms such as XGBoost, LightGBM, RF, and Cat Boost have been integrated into this study using an ensemble technique called stacking, and metaheuristics have been developed with the telecoms business, to forecast customer attrition. According to the results of this study, the proposed method confirms performance in predicting customer churn with an accuracy of 92.2%.

Keywords: *Customer churn, Telecommunication, Extreme gradient boost algorithm, Random forest, LightGBM, Machine learning, Customer satisfaction, Churn prediction*

1. Introduction

Churn prediction is critical in a variety of businesses, including life insurance, finance, healthcare, and telecommunications. Over the past two decades, mobile communication has emerged as the primary channel of communication [1]. Churn prediction has grown to be a major problem for telecom firms since new technology and rivals are quickly emerging [2]. In order to offer them a retention solution, a customer churn prediction model can accurately identify probable churners [3]. Churn Prediction is becoming more precise and practical thanks to recent advances in machine learning (ML) and artificial intelligence [4-5]. It is crucial for identifying clients who are likely to leave a business or its services at an early stage [2].

Throughout the previous two decades, the telecom sector has grown to be one of the most significant industries in developed countries [6]. Due to the abundance of data available in the telecom business, data mining is essential for prediction and analysis [7]. To save money on client retention and to increase profit margins, churner prediction is the fundamental application area. In the telecom industry, data mining techniques are utilized to track customer turnover patterns.

Oladipo, I. D., Awotunde, J. B., AbdulRaheem, M., Taofeek-Ibrahim, F. A., Obaje, O. and Ndunagu, J. N. (2023). Customer Churn Prediction in Telecommunications Using Ensemble Technique, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 9 No. 1, pp. 82 – 95.

Due to the rising number of telecom users, corporations today provide a wide range of services to keep clients [8]. Churn is the term for the phenomena where a client moves service providers in order to receive better advantages and services [9]. A service provider's business will experience revenue loss if a customer switches. The probable churners can be identified by prediction, and retention strategies can be offered to them. There are numerous mining algorithms that categorize customers' behaviour into churners and non-churners.

An effective churn prediction model should effectively utilise a huge churner's recognition requires a large amount of previous data, according to existing studies [10-11]. Nevertheless, a number of shortcomings in the current models make it impossible to carry out churn forecast with great accuracy and effectiveness [12]. A lot of data is generated in the telecom business is produced that has missing values [13]. Prediction models in the literature produce poor or erroneous findings when based on this kind of data. To address this problem, data preparation is now carried out, and missing values are imputed using ML-based techniques, producing good performance and categorisation correctness.

Despite the fact that feature extraction is performed in the literature as well, some crucial aspects that are rich in information are overlooked when developing models. Additionally, model creation is done using statistical approaches, which results in poor classification results. Also, benchmark datasets are not employed in the research for model assessment, which leads to a poor depiction of the true nature of the data. Without benchmark datasets, fair comparisons between different models cannot be made. It is possible to fix the current problems and more precisely forecast churn by using an intelligent model. Researchers have also used various ML-based techniques to address the consumer problems, however these ML-based solutions are ineffective in anticipating customer attrition.

Therefore, based on the issues in the existing research, this study proposes ensemble-based classifier for churn prediction in telecommunication industry. The In this study, the Synthetic Minority Oversampling Method (SMOTE) is employed to deal the unbalanced datasets. The proposed model's contributions are as follows:

i. presents ensemble-based classifiers for the telecom industry's Customer Churn Prediction.
ii. feature selection technique, feature extraction, and SMOTE for balancing the data are all applied;
iii. duplicates, noise, and missing values are removed from the data before being prepared;
iv. apply an ensemble of ML-based techniques on the data, including CatBoost, Random Forest (RF), LightGBM, and XGBboost.
v. analyze how well the data respond to the meta algorithm.

2. Related Works

The ability to predict that a customer is at high danger of leaving while there is still time to act is a new revenue source for any firm. As a result, there has been investment in this field, as well as study and testing of a variety of approaches. The use of Support Vector Machine (SVM) to forecast churn in subscription services was one of the ways examined [14]. The purpose was to use this method to create and fine-tune a reliable churn model. The model performance was compared to that of logistic regression and RF in forecasting client attrition. The area under the curve was chosen as the major criterion for evaluating the authors' models (AUC). Their SVM models had the best AUC value of 85.14, while the RF model used as a benchmark had an AUC of 87.21.

The usefulness and the effectiveness of decision tree and logistic regression approaches were evaluated by Dahiya and Bhatia [15]. They came to the conclusion that the decision tree technique had a higher accuracy than the logistic regression technique. Wangperawong, *et al.* [16] looked at a dataset with over 6 million customers. Only 12 criteria were used to characterize each consumer, and the data was presented as an image. The experiment was carried out with the Convolutional Neural Network (CNN) architecture, which resulted in a high AUC score of 0.743.

Burez and Van [17] compared three categorization methods: logistic regression, automated relevance determination (ARD), Neural Networks (NN), and RF. With an AUC of 83.19, the RF model provided the best match in this study. The ARD and NN models both had an AUC of 83.10, which was similar. To evaluate the churn forecasting accuracy of Multiple Regression Analysis, Logistic Regression Analysis, Multilayer Perception (MLP), and NN techniques, Ismail, *et. al.* [18] conducted an experiment. MLP had the highest overall accuracy of almost 91 percent, according to the results. Verbeke *et. al.* [7] compared different client retention techniques. The authors selected 21 approaches to simulate in both the training and testing sets and assessed their AUC and percentages of correct answers properly categorized (PCC) occurrences. Some of the algorithms available are SVM with various kernels, C4.5, RF, and k-NN. With AUC values of 0.8249 and 0.8319 in both the train and test sets, the RF obtained the greatest AUC.

Coussement *et. al.* [19] compared an optimized logistic regression, bagging, Bayesian network, and Naive Bayes were applied to eight cutting-edge data mining algorithms employing common input data, comprising genuine cross-sectional data from a large European telecoms operator, DR, NN, RF, SVM, SGB. The following conclusions are drawn as a result of the findings. First off, advances in the region underneath the reception operational characteristics curve and 34% in the highest decile lift show that analysts are becoming more conscious of the effect that their data-preparation process has on churning predictive accuracy.

Improved regression analysis technique also faces competition from more powerful individual and ensembles data gathering techniques, which is the second point. Huang *et. al.* [20] looked into the problem of customer churn in the big data platform. The researchers hoped to demonstrate that, based on the quantity, variation, and mobility of the data, big data may significantly enhance the process of churn prediction. A big data system was used to generate cracks in data from the Operation Support and Business Support divisions of China's largest telecoms corporation was required. AUC was used to evaluate the RF method.

Makhtar *et. al.* [21] proposed a technique for predicting churn in telecom using rough set theory. The Rough Set classification algorithm surpassed other algorithms such as Linear Regression, Decision Tree, and Voted Perception Neural Network, as stated in this research. The topic of unbalanced data sets, where churned customer classes are smaller than active customer classes, has been studied in several studies since it is a major issue in the churn prediction problem.

Amin *et. al.* [22] provided a novel churn prediction model based on the classifier's confidence estimation using the distance factor. They separated the dataset into two groups with high and low certainty before grouping the zones based on distance into distinct zones. They found that individuals in the region with a higher proximity factor value (i.e., high confidence client attrition and non-churn) had higher accuracy while using Nave Bayes as a classification (i.e., churn prediction and non-churn with minimal confidence) than those in the region with a smaller range factor value. On the four datasets utilized, accuracy in the last tenth iteration was (82.9%1 and 84.3%, 70.6% and 74.8%, 70.0% and 89.0%, 57.0% and 56.0%).

Andrews *et. al.* [23] employed a collection of 10,000 telecom customer information, each having 21 characteristics, 2900 of which are Belgian Telecom churners. To check the accuracy of the prediction, they employed profound the region underneath the curve rating was 0.89 using classifier model and 10-overlap crossing approval procedures. Ahmad *et. al.* [24] developed ML-based approaches on a big data infrastructure for analyzing SyriaTel's telecom statistics, which included all client information during a 9-month period. The Decision Tree, RF, GBM, and The model made extensive use of XGBoost methods. 83, 87.76, 90.9%, and 93.3% were the AUCs for the four models. The best results were obtained using XGBoost, which achieved 93.3% while using SNA features, which increased model effectiveness from 84% to 93.3%. The model was created in the Spark environment and tested.

3. Methodology

This study proposes Ensemble techniques that have revealed their advantages over single technique. However, it is not known which of the ensemble techniques can achieve the best in customer churn prediction. Therefore, the data will be subjected to stacking, a type of ensemble approach. For churn prediction, this study tends to recommend a combination of the XGBoost, LightGBM, RF, and CatBoost algorithms. The reason the models were Low variance and high bias Base models should be combined in such a manner that the ensemble model is less biased while maintaining high variance. Weak models should be merged in such a way that the robust paradigm is strengthened. The decision-making mechanism is strengthened in the face of lacking data thanks to ensemble classifier, which is another advantage over features.

3.1 Proposed Framework

In order to fill the research gaps and provide an effective approach to the study, an innovative framework has been proposed in Figure 1. The proposed framework provides a clear approach of the analysis in this study. It consists of twelve components as seen in framework.

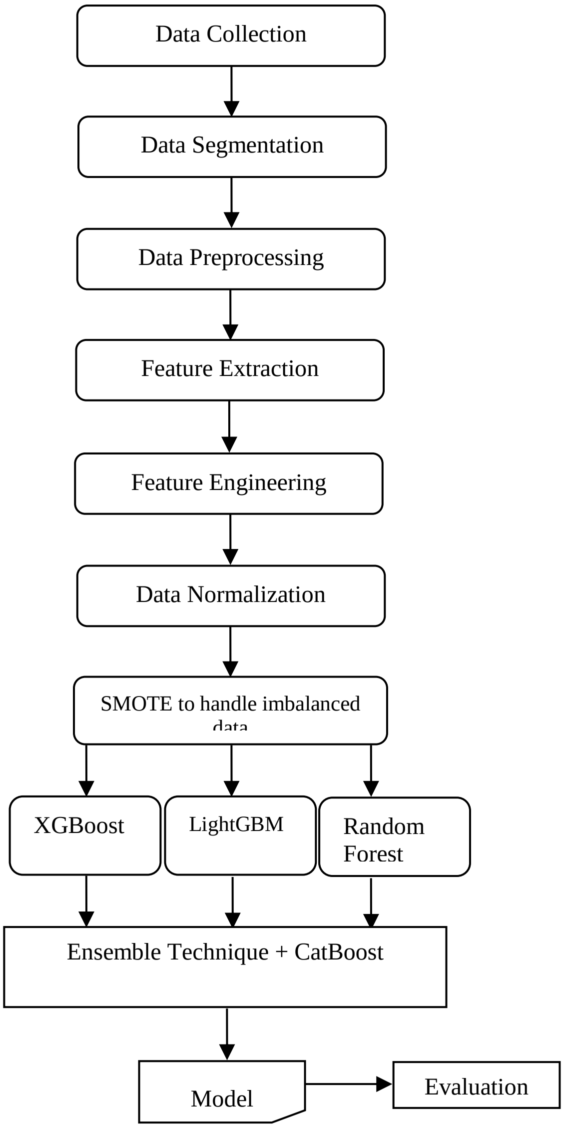


Figure 1: Theoretical Framework of the System

3.2 Dataset Collection

The customer churn dataset downloaded fromKaggle.com, is an open-source dataset with 3333 observations and 21 characteristics. The 'Churn' function displays customer churn and non-churn based on current conditions. Approximately 14.5 percent of 'Churn' is labeled as 'T,' while 84.5 percent of 'non-churn' is labeled as 'F.' the training and test datasets are made up of 80 percent (2666 occurrences) and 20 percent (667 instances) of the dataset, respectively.

3.2.1 Data Pre-processing

Data preparation is a technique for transforming the first procedure used on the churn prediction data is data pre-processing. After pre-processing the based on the data, the suggested churn forecasting model included the following activities.

3.2.2 Data Cleaning

When there are incomplete, duplicate, or noisy variables in the data, prediction becomes quite challenging. Hence, data cleaning is done to eliminate duplicate data, identify and remove noise/error values, and substitute entries with actual values calculated by the mean of each property.

3.2.3 Feature/Variable extraction

The key stage in data pre-processing is feature selection. The most significant features are chosen for the prediction model through the use of forward selection during feature selection.

3.2.4 Data Normalization

Because ML-based models are mass-based, they rely on a variety of features. Due to levels, data may be distorted into different locations, causing some values to update more quickly than others [3, 11]. It is crucial to use statistical normalization to solve this issue because it allows the calculation of the $Z - score$ function for each feature value $\left(v^{(i)}\right)$, and calculate thus:

$$Z^{(i)} = \frac{v^{(i)} - \mu}{\sigma} \qquad (1)$$

where μ is the mean of the values for a particular function $\left(v^{(i)}(i \epsilon\, 1, 2, 3, \ldots, n)\right)$ and σ is the standard deviation.

3.3 SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is a data preprocessing method algorithm that generates synthetic data points. By adding fake along the sample points linking any/all of the k examples to each underrepresented class instance outnumbered class ensemble classifiers, this class is oversampled. Using this approach, the overfitting problem is resolved and the decision region of the instances of the outnumbered class is expanded. The newly synthesized sample by SMOTE is depicted by equation (2).

$$T_{new} = T + (T_i - T) \times a(0,1) \qquad (2)$$

Where T_{new} is the newly created synthetic data, T are samples from the Minority class, T_i are chosen examples from the ensemble, and a are random numbers between 0 and 1.

3.3.1 Test Data

It is not advisable to train and test a model created on the same dataset since no one can tell if the model merely remembered the data or if it generalizes well to new, unknown data. A basic procedure of random sub sampling is used to divide a dataset into training and test sets. Eighty (80) percent of these data at random for our training set and twenty (20) percent for our test set.

3.3 Classification Algorithms

3.3.1 XGBoost (extreme Gradient Boosting)

XGBoost is a networked gradient boosting toolset that focuses on adaptability, versatility, and performance. It creates ML-based methods by employing Gradient Boosting framework. The parallel tree boosting algorithm XGBoost, commonly known as GBDT or GBM, is effective at swiftly and precisely resolving a range of data science problems. In a distributed context, the same technique may be used to solve problems with billions of cases (Hadoop, SGE, and MPI). In comparison to other gradient boosting algorithms, XGBoost is roughly ten times faster and has a high predictive power [25].

3.3.2 LightGBM

It is a gradient boosting paradigm built around decision trees that's quick, distributed, and high-performing. In contrast to previous boosting techniques, it breaks the tree leaf by leaf with the best match. When the dataset is enormous, Lite GBM performs better than every other method. On a big dataset, Light GBM runs faster than the other algorithms [28].

3.2.3 Random Forest

In RF, the tree predictors are mixed, with each tree relying on independently chosen values from a spanning all trees, a randomized vector that has an identical dispersion. RF model performs the following: Random groups are created from the original dataset (bootstrapping). To identify the appropriate split, just a random collection of characteristics is checked at each node in the decision tree. A decision tree model is fitted to each subgroup. Averaging the forecasts from all decision trees yields the final prediction [26].

3.3.4 CatBoost

CatBoost provides a gradient boosting framework that, in contrast to the standard technique, aims to solve for categorical features using a permutation-driven alternative [27]. When measures of variability contain a large number of labels (i.e. they are highly cardinal), one-hot-encoding them increases significantly the dimensionality of the dataset, proving it challenging to deal with. Unlike other ML-based techniques, CatBoost can work with explanatory data autonomously and does not require a large amount of data pre-processing [27].

3.4 Performance Evaluation Metrics

In this study, accuracy, the area under the curve (AUC) and memory rating are used to calculate assess the suggested ensemble approach for determining consumer churn. If it is expressed this equation is derived in terms of True Positive (TP), False Positive (FP), and False Negative (FN): Because the datasets utilized in this study are skewed, the fraction of accurate matches (called accuracy) would be unhelpful in evaluating model performance. The confusion matrix provides a comprehensive knowledge of the model's forecasts. The confusion matrix computes the crucial True Positive (TP), False Positive (FP), and False Negative (FN) values (FN). It would be preferable to have a single number to represent overall performance rather than many per-class F1 scores.

3.4.1 Accuracy

It is described as the proportion of samples that the classifier successfully assigned to all of the samples in a given test dataset.

$$Accuracy = \frac{(TP\ TN)}{(TP + \mathrm{TN} + FP + FN)} \qquad (3)$$

'TN' is True Negative, 'TP' is True Positive, 'FN' is False Negative and 'FP' is False Positive.

3.4.2 Area Under Curve (AUC)

AUC is one of the most widely used measures for evaluation. It is used to solve Problems with binary categorization. AUC is the likelihood that a classifier would select a randomly selected positive example over a randomly selected minor mistake.

3.4.3 Precision

Precision is computed by dividing the number of TP by the overall number of TP and FP.

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

3.4.4 F1 Score

Both precision and recall should be examined in order to evaluate model performance fully. The F1 score is a helpful statistic that considers both of them. For a more realistic assessment of model performance, use the harmonic mean of recall and accuracy.

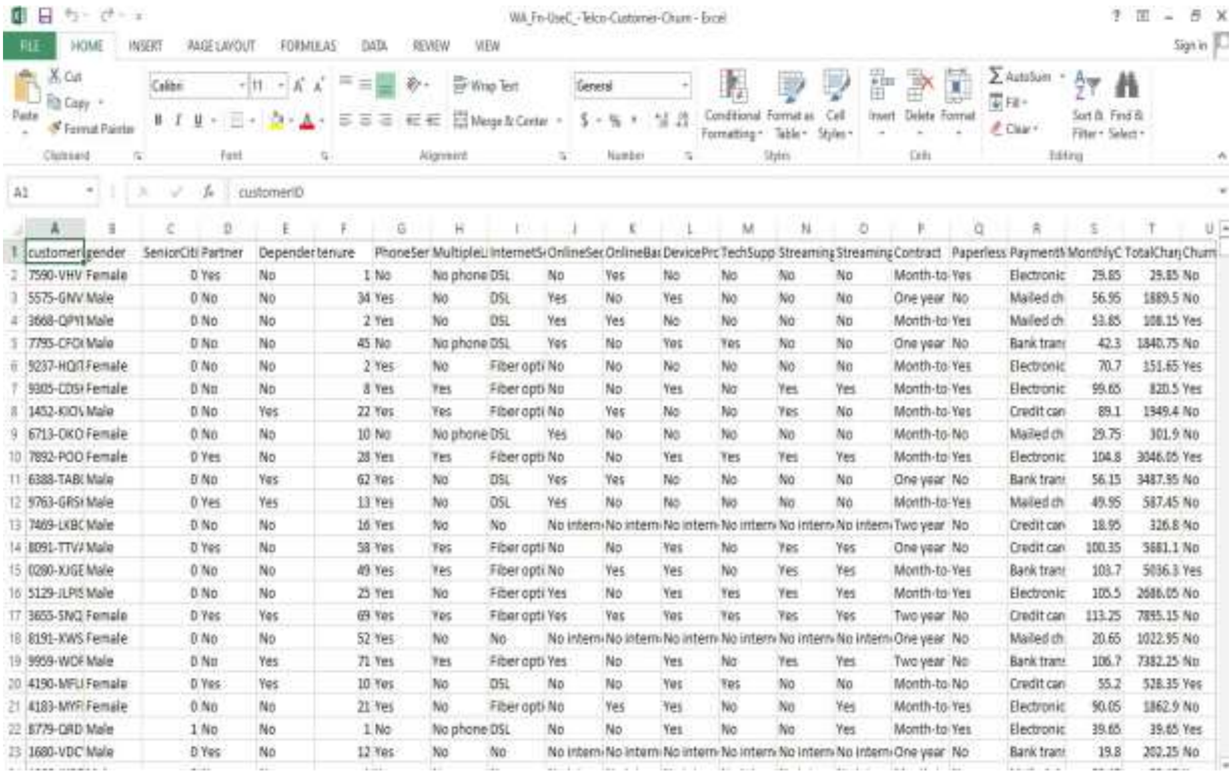
$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (6)$$

4 Implementation and Results

The VS Code was used to build the web app and Jupyter Notebook was used to build the model. In this experiment, the SMOTE library and feature extraction was used for improving the model and the train test divide function from the scikit-learn package was used to partition the data.

The customer churn information is freely available with 3333 observations and 21 characteristics [29]. The 'Churn' function displays in accordance with present situations, consumer churn and non-churn are calculated. Roughly 14.5% of 'Churn' is caused by labeled as 'T,' while 84.5% of 'non-churn' is labeled as 'F.' the training and test datasets are made up of 80% (2666 occurrences) and 20% (667 instances) of the dataset, respectively.

The smote and pandas library were used for handling the imbalanced data and also pre-processed data. Figure shown a sample of Telecommunication Customer Churn Dataset.



customer	gender	SeniorCiti	Partner	Depender	tenure	PhoneSer	MultipleL	InternetS	OnlineSec	OnlineBa	DevicePrc	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentM	MonthlyC	TotalChar	Churn
7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to	Yes	Electronic	29.85	29.85	No
5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed ch	56.95	1889.5	No
3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Mailed ch	53.85	108.15	Yes
7795-CFOI	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank tran	42.3	1840.75	No
9237-HQIT	Female	0	No	No	2	Yes	No	Fiber opti	No	No	No	No	No	No	Month-to	Yes	Electronic	70.7	151.65	Yes
9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to	Yes	Electronic	99.65	820.5	Yes
1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	No	Month-to	Yes	Credit car	89.1	1949.4	No
6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to	No	Mailed ch	29.75	301.9	No
7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber opti	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	104.8	3046.05	Yes
6388-TABC	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank tran	56.15	3487.95	No
9763-GRSI	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to	Yes	Mailed ch	49.95	587.45	No
7469-LKBC	Male	0	No	No	16	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Two year	No	Credit car	18.95	326.8	No
8091-TTVA	Male	0	Yes	No	58	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber opti	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	Bank tran	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber opti	Yes	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	105.5	2686.05	No
3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
8191-XWS	Female	0	No	No	52	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Mailed ch	20.65	1022.95	No
9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber opti	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank tran	106.7	7382.25	No
4190-MFLI	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to	No	Credit car	55.2	528.35	Yes
4183-MYFI	Female	0	No	No	21	Yes	No	Fiber opti	No	Yes	Yes	No	No	Yes	Month-to	Yes	Electronic	90.05	1862.9	No
8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to	Yes	Electronic	39.65	39.65	Yes
1680-VDC	Male	0	Yes	No	12	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Bank tran	19.8	202.25	No

Figure 2: Telecommunication Customer Churn Dataset.

Table 1. The performance result of XGBoost algorithm

	Precision	Recall	F1-Score	Support
0	0.94	0.85	0.89	783
1	0.86	0.94	0.90	761
Accuracy			0.90	1544
Macro Average	0.90	0.90	0.90	1544
Weighted Average	0.90	0.90	0.90	1544

Support represents the number of samples for the target values (0 and 1).

The macro averaged precision represent the arithmetic mean of the precision of the two target values, the macro averaged recall represent the arithmetic mean of the recall of two target values and the macro averaged F1_score represent the arithmetic mean of the F1_scores of the target values (0 and 1) obtained by comparing the actual outcomes to the predicted outcomes by the Xgboost algorithm.

The weighted-averaged accuracy is obtained by taking the median of all per-class accuracy while taking into account the supports of each class, and the weighted-averaged recall is derived by taking the mean of all per-class recall while taking into account the supports of each class. The weighted-averaged F1 score is generated by averaging all per-class F1 scores while taking into account each class's support gotten as result of prediction by the Xgboost algorithm. In other words, The F1 score that has been weighted averaging is the sum of the multiplication of the f1 score with the support proportion for each of the target values.

Table 1 shows the performance of results of XGBoost algorithm with F1-Score of accuracy of 0.90, macro average score for precision and Recall is 0.90 respectively.

Table 2 shows the performance result of LightGBM algorithm with F1-Score of accuracy of 0.89, macro average score for Precision and Recall is 0.90 and 0.89 respectively.

Table 3 revealed the performance result of RF algorithm with F1-Score of accuracy of 0.89, macro average score for Precision and Recall is 0.90 and 0.90 respectively.

Table 4 shows the performance result of Cat Boost algorithm with F1-Score of accuracy of 0.84, macro average score for Precision and Recall is 0.84 and 0.84 respectively.

Table 2. The performance result of LightGBM

	Precision	Recall	F1-Score	Support
0	0.94	0.84	0.89	783
1	0.85	0.95	0.89	761
Accuracy			0.89	1544
Macro Average	0.90	0.89	0.89	1544
Weighted Average	0.90	0.89	0.89	1544

Support represents the number of samples for the target values (0 and 1).

Table 3. The performance result of Random Forest

	Precision	Recall	F1-Score	Support
0	0.95	0.84	0.89	783
1	0.85	0.96	0.90	761
Accuracy			0.90	1544
Macro Average	0.90	0.90	0.90	1544
Weighted Average	0.90	0.90	0.90	1544

Support represents the number of samples for the target values (0 and 1).

Table 4. The performance result of Cat Boost

	Precision	Recall	F1-Score	Support
0	0.89	0.78	0.83	783
1	0.80	0.90	0.85	761
Accuracy			0.84	1544
Macro Average	0.84	0.84	0.84	1544
Weighted Average	0.85	0.84	0.84	1544

Support represents the number of samples for the target values (0 and 1).

4.1 Customer Churn Predictor Application

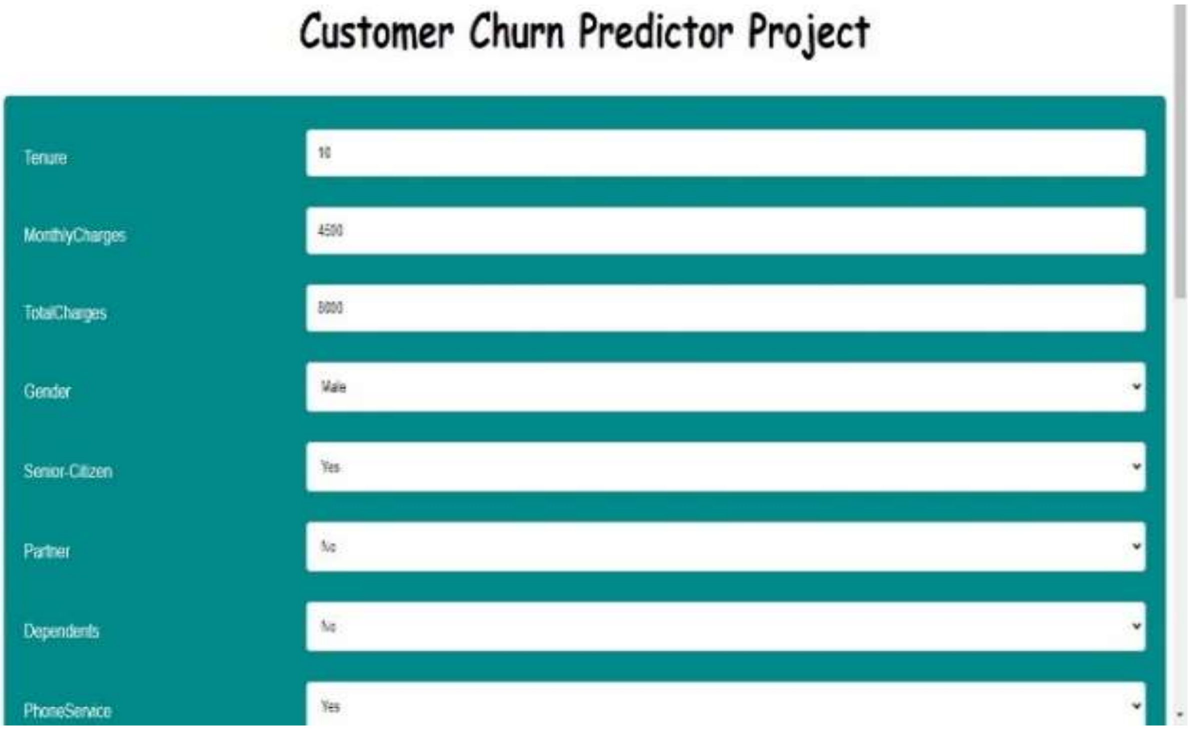


Figure 4. Customer churn prediction app.

Figure 5: Customer Churn Predictor App

4.1 Ensemble Model Outcome

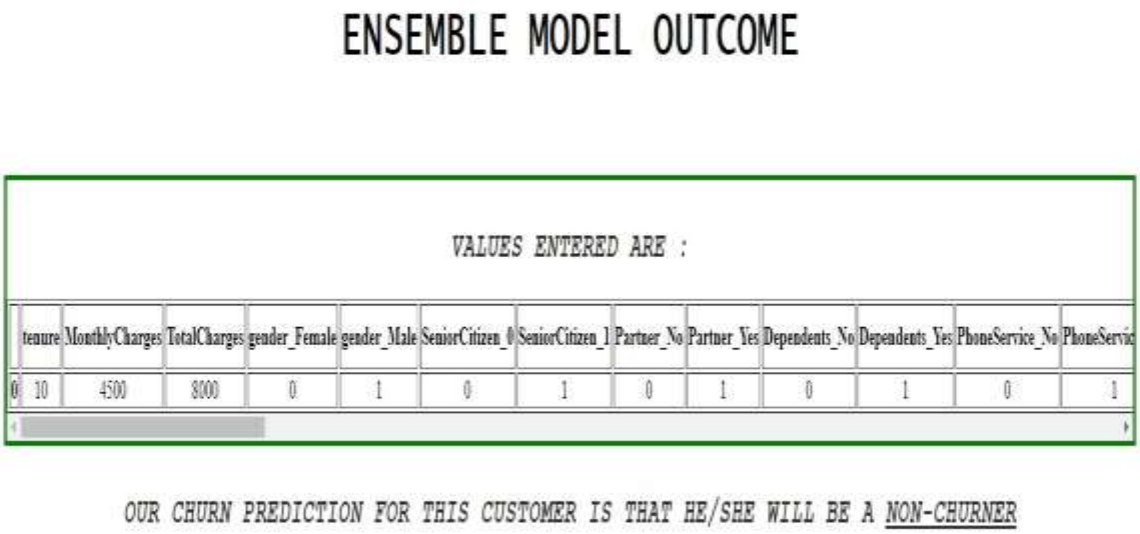


Figure 6: Ensemble Model Outcome.

4.2 Exploratory Data Analysis Carried Out

Data scientists utilize exploratory data analysis (EDA) to analyze and investigate data sets and define their key attributes, typically utilizing data visualization tools.;

Figure 7 displayed the visualization of the gender feature in the dataset to get an insight on the number of males and females that are churners and non-churners. From figure 7 it was revealed that there more female that are churn than male churners.

Figure 8 shows a visualization of the partner feature in the dataset to give a diagrammatic view of the number of single and married persons that are churners and non-churners.

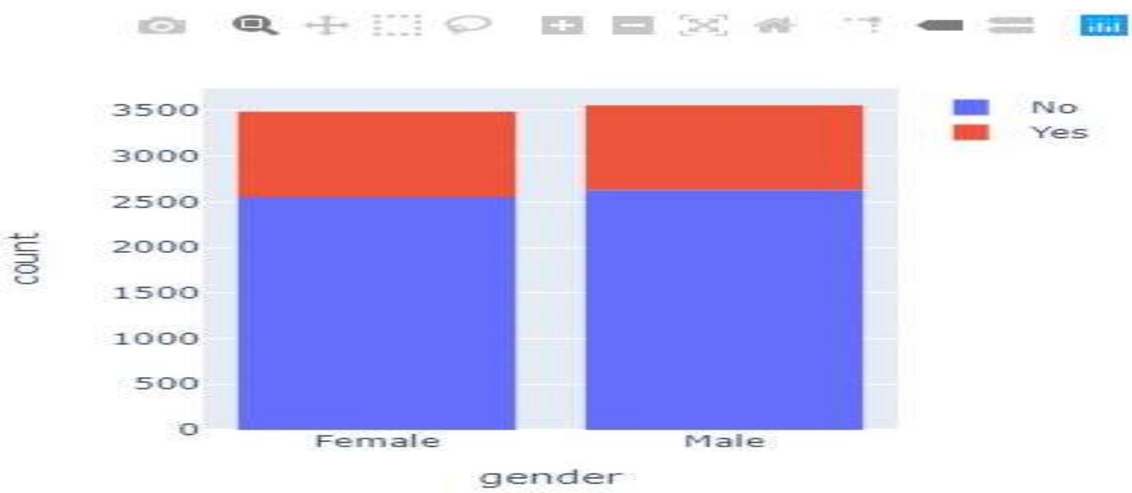


Figure 7: Gender feature

Figure 8: Partner feature

Figure 9 shows a visualization of the dependent feature in the dataset to give a diagrammatic view of the number of independent and dependent persons that are churners and non-churners.

Figure 10 revealed the visualization of the phone service feature in the dataset to get an insight on the number of those that have access to phone service, no access to phone service that are churners and non-churners.

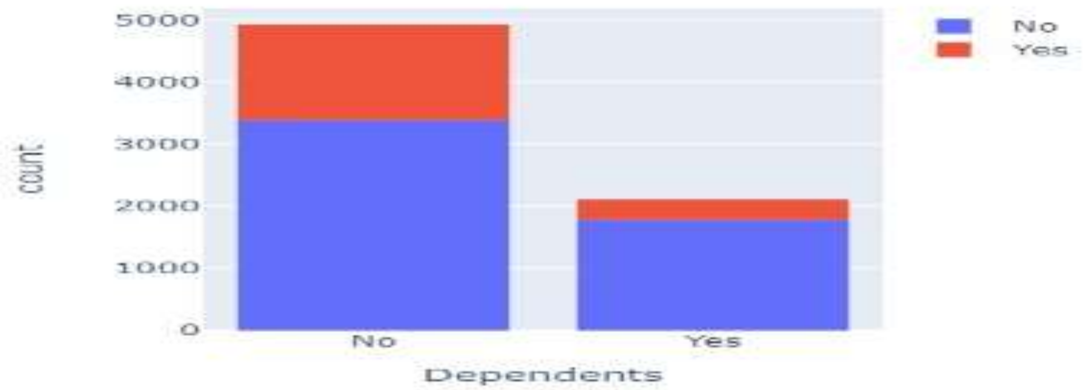


Figure 9: Dependents feature

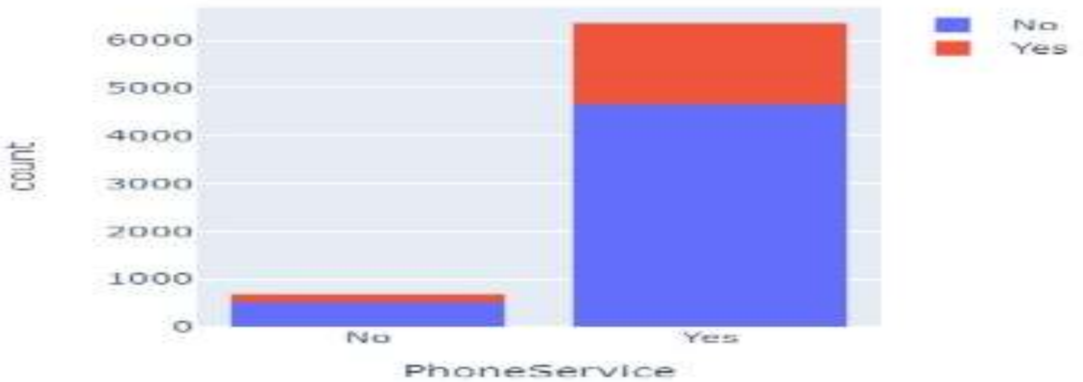


Figure 10: Phone Service feature

Figure 11 displayed the visualization of the internet service feature in the dataset to get an insight on the number of those that use DSL, Fiber Optic or no internet service that are churners and non-churners.

Figure 12 shown the visualization of the online security feature in the dataset to get an insight on the number of consumers that have no access to internet service, access to online security or no access to online security that are churners and non-churners.

Figure 13 displayed the visualization of the online backup feature in the dataset to get an insight on the number of consumers that have backup service, no access to internet service and no back up service that are churners and non-churners.

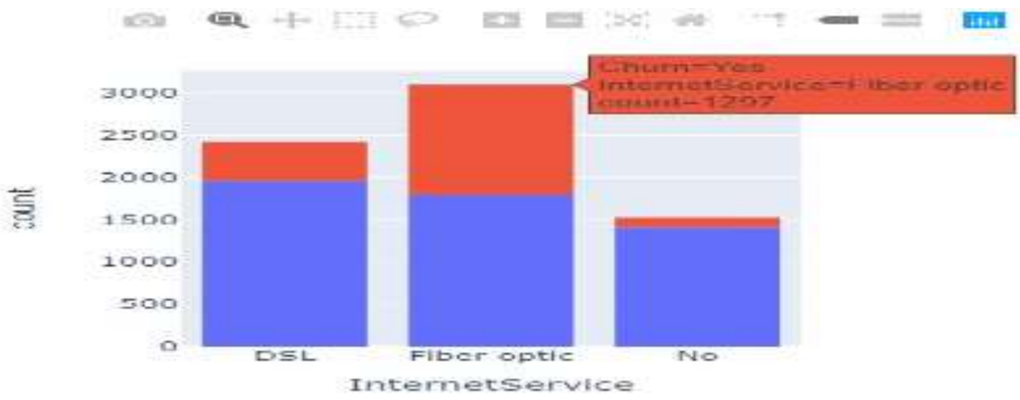


Figure 11: Internet Service feature

Figure 12: Online Security feature

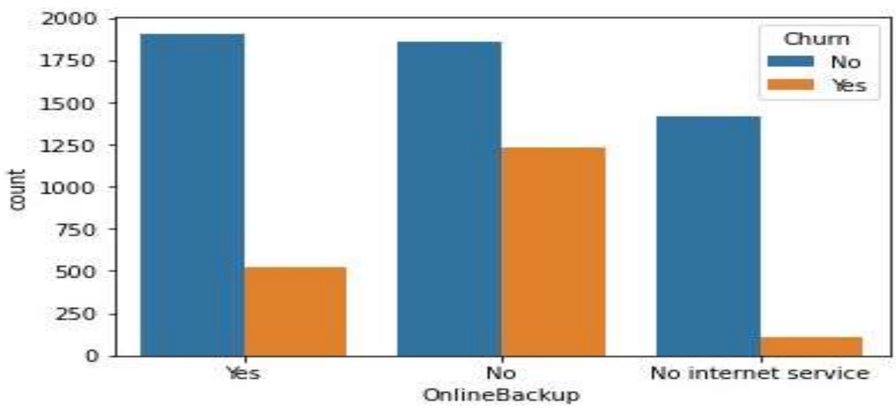


Figure 13: Online Backup feature

4.3 Comparison of the proposed models with existing approaches

Table 5: Comparison of the proposed models

Authors	Model(s)	Accuracy (%)
[22]	NB	89%
[7]	SVM + RF + KNN	83.19%
[30]	NN	91.1%
[17]	LR + ARD + NN + RF	83.19%
Proposed Model	XGB + LightGBM + RF + Cat Boost	92.2%

As revealed from table 5, the study outperformed the existing models in the literature in term of accuracy. In this study, ML-based algorithms were applied on customer churn dataset to predict customer churn in telecommunications industry. Various classification models were combined to propose an efficient system for predicting customer churn. This study makes use of four various ensemble classifiers to build a meta algorithm for the prediction of churn in the telecommunications business.

The classifiers utilized in this investigation are as follows; XGBoost, LightGBM, RF and CatBoost. After ensemble technique was applied for the combination of various algorithms to form a meta algorithm, an accuracy of 92% was achieved. The application of ML-based algorithms for predicting analysis is very important in the aspect in the telecommunications business of estimating customer churn. The system created using ensemble technique will help telecommunication companies to predict if their customers are potential churners. Results in this study show that the proposed method proves its performance in predicting customer churn with an accuracy of 92.2%. From these results, it can be concluded that the proposed method performed better than existing models stated in this study.

5. Conclusion

Predicting customer attrition is a serious issue for telecom firms. Identification of consumers who are dissatisfied with the services offered enables businesses to improve their weak points, pricing strategies, marketing initiatives, Leveraging consumer preferences to reduce the causes of turnover. In literature, a number of techniques are employed to predict client churn. The suggested approach concentrated on providing different ensemble classifiers for highly accurate customer churn prediction in the telecom industry.

The churn prediction dataset, which was gathered from an open data source, was then used to evaluate the proposed models. The proposed models revealed an accuracy of 92.2%, and outperformed other state-of-the-art models in terms of classification and prediction accuracy. Big data analytics can be used to further this research in the future. The amount of consumer satisfaction with telecom services can be determined through social network analysis, and these services can subsequently be provided to lower the churn rate.

References

[1] Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Computer Science*, 8, e854.

[2] Shabankareh, M. J., Shabankareh, M. A., Nazarian, A., Ranjbaran, A., & Seyyedamiri, N. (2022). A stacking-based data mining solution to customer churn prediction. *Journal of Relationship Marketing*, 21(2), 124-147.

[3] Abdulsalam, S. O., Ajao, J. F., Balogun, B. F., & Arowolo, M. O. (2022). A Churn Prediction System for Telecommunication Company Using Random Forest and Convolution Neural Network Algorithms. *EAI Endorsed Transactions on Mobile Communications and Applications*, 7(21).

[4] Amin, A., Khan, C., Ali, I., & Anwar, S. (2014). Customer churn prediction in telecommunication industry: With and without counter-example. *In Nature-Inspired Computation and Machine Learning*: *13th Mexican International Conference on Artificial Intelligence, MICAI 2014*, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part II 13 (pp. 206-218). Springer International Publishing.

[5] De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & Von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51(1), 91-105.

[6] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, 7, 60134-60149.

[7] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, 218(1), 211-229.

[8] Bhattacharyya, J., & Dash, M. K. (2022). What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019. *FIIB Business Review*, 11(3), 280-302.

[9] Lin, L., Guo, Z., & Zhou, C. (2023). Failure to maintain customers: antecedents and consequences of service downgrades. *Journal of Service Theory and Practice*.

[10] Wu, F., Lyu, F., Ren, J., Yang, P., Qian, K., Gao, S., & Zhang, Y. (2023). Characterizing Internet Card User Portraits for Efficient Churn Prediction Model Design. *IEEE Transactions on Mobile Computing*.

[11] de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, 34(14), 11751-11768.

[12] Toor, A. A., & Usman, M. (2022). Adaptive telecom churn prediction for concept-sensitive imbalance data streams. *The Journal of Supercomputing*, 78(3), 3746-3774.

[13] Chinnaraj, R. (2022). Bio-Inspired Approach to Extend Customer Churn Prediction for the Telecom Industry in Efficient Way.

[14] Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1), 313-327.

[15] Dahiya, K., & Bhatia, S. (2015, September). Customer churn analysis in telecom industry. *In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(*Trends and Future Directions) (pp. 1-6). IEEE.

[16] Wangperawong, A., Brun, C., Laudy, O., & Pavasuthipaisit, R. (2016). Churn analysis using deep convolutional neural networks and autoencoders. *arXiv preprint arXiv*:1604.05377.

[17] Burez, J., & Van den Poel, D. (2008). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(4626-4636), 2-11.

[18] Ismail, M. R., Awang, M. K., Rahman, M. N. A., &Makhtar, M. (2015). A multi-layer perceptron approach for customer churn prediction. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213-222.

[19] Coussement, K., Lessmann, S., &Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn

prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, 27-36.

[20] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Application*s, 39(1), 1414-1425.

[21] Makhtar, M., Nafis, S., Mohamed, M. A., Awang, M. K., Rahman, M. N. A., & Deris, M. M. (2017). Churn classification model for local telecommunication company based on rough set theory. *Journal of Fundamental and Applied Sciences*, 9(6S), 854-868.

[22] 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.

[23] Andrews, R., Zacharias, R., Antony, S., & James, M. M. (2019). Churn prediction in telecom sector using machine learning. *International Journal of Information*, 8(2).

[24] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1), 1-24.

[25] Awotunde, J. B., Folorunso, S. O., Imoize, A. L., Odunuga, J. O., Lee, C. C., Li, C. T., & Do, D. T. (2023). An Ensemble Tree-Based Model for Intrusion Detection in Industrial Internet of Things Networks. *Applied Sciences*, 13(4), 2479.

[26] Wu, W., Xia, Y., & Jin, W. (2020). Predicting bus passenger flow and prioritizing influential factors using multi-source data: Scaled stacking gradient boosting decision trees. *IEEE Transactions on Intelligent Transportation Systems*, 22(4), 2510-2523.

[27] Zhou, L., Fujita, H., Ding, H., & Ma, R. (2021). Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting. Applied *Soft Computing*, 110, 107672.

[28] Gu, Q., Sun, W., Li, X., Jiang, S., & Tian, J. (2023). A new ensemble classification approach based on Rotation Forest and LightGBM. *Neural Computing and Applications*, 1-22.

[29] Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013, September). Telecommunication subscribers' churn prediction model using machine learning. *In Eighth international conference on digital information management (ICDIM 2013)* (pp. 131-136). IEEE.

[30] He, Y., He, Z., & Zhang, D. (2009, August). A study on prediction of customer churn in fixed communication network based on data mining. *In 2009 sixth international conference on fuzzy systems and knowledge discovery* (Vol. 1, pp. 92-94). IEEE.