Date of publication MAR-31, 2025, date of current version AUG-29, 2024
www.computingonline.net/ computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.1.3873

Customer Churn Prediction: A Machine Learning Approach with Data Balancing for Telecom Industry

ANURAG BHATNAGAR[1], SUMIT SRIVASTAVA[2]
[1 2]Department of Information Technology, Manipal University Jaipur, Rajasthan, India
[1] (e-mail: anurag.bhatnagar@jaipur.manipal.edu)
Corresponding author: Sumit Srivastava (e-mail: sumit.srivastava@jaipur.manipal.edu)

ABSTRACT Churn prediction is the process of identifying customers who stop using services. Churn is not only the problem in Telecom industry but also banking, insurance, gaming companies, and internet service providers are also facing this challenge. This study focuses on churn prediction in telecom industry to determine the best classification model and reduce the number of attributes in the dataset. Machine learning models like Random Forest, K-Nearest Neighbour, Decision Tree, Support Vector Machine, Logistic Regression, Bagging Classifier, Extreme Gradient Boosting, Stochastic Gradient Descent Classifier, Gaussian Naive Bayes were used. To handle imbalance data and for hyper parameter tuning, techniques like SMOTE, ENN, Under-Sampling, Over-Sampling and K-cross fold validation were used. Random Forest classifier performed exceptionally well in forecasting customer churn in the telecom sector, as evidenced by the results. Its accuracy rate was 90.30% with all attributes, and 90.90% with reduced attributes dataset. This implies that the dataset with reduced attributes may be useful for churn prediction tasks in a variety of industries, offering useful information to companies trying to reduce customer attrition. This work validates itself by comparing with four previously published research.

KEYWORDS Churn, Classification models; Imbalanced data; Hyper parameter tuning; Attributes; Validation; Accuracy

I. INTRODUCTION

Churn is not only the problem of telecommunication industry, but the term churn gets associated wherever customers and service providers are associated. Now a day better customer services or technically advanced features are attracting customers. In the telecom sector, churn analysis is the process of identifying and forecasting customer churn, or the rate at which customers discontinue utilizing a telecom company's services. Maintaining profitability and development in the telecom industry is highly dependent on client retention due to its highly competitive nature. Churn analysis is a process that looks at past data, finds trends, and creates prediction models to find out which customers are most likely to leave in the future [1], [2]. Customer churn is mostly caused by dissatisfaction with the services received, excessive fees, unsolicited offers, and harsh customer service [3].

Customers are in search of more comfort and luxury. In context of the same the churn becomes a normal trend. On the other hand, associating with new customers is costlier than retaining existing customers [4]. The establishment of client retention strategies is one of the primary goals of customer churn prediction. The danger of client attrition is rising exponentially along with the competitiveness in service sectors. As a result, developing methods to monitor devoted clients (non-churners) has become essential [5]. To accurately identify the churn analysis issue, the business must predict the customer's behaviour. Fig. 1 shows two approaches to manage customer churn: (1) Reactive & (2) Proactive. Reactive approaches involve the corporation waiting for the consumer to request a cancellation before offering the customer inducing ideas to keep them around. Proactive strategies anticipate customer attrition and provide solutions to keep those clients. The problem is a binary clas-

sification where the non-churners and churners are divided [5].

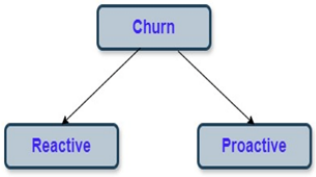


Figure 1. Classification of Churn Prediction

Machine learning models are crucial for forecasting client attrition. Accuracy, Precision, F-Measure and Recall are a few of the characteristics used to analyse the results of different machine learning models. CRM, a strategy for building and strengthening customer relationships, is utilized in industries like telecommunications, banking, and retail for customer retention, requiring BI applications [6]. The telecom sector generates significant data daily, emphasizing the cost of retaining customers over acquiring new ones, necessitating understanding of churn reasons and behaviour patterns for business analysts and CRM analysers [7]. Businesses must fiercely compete for new consumers, focusing on client retention to increase revenue. Early detection of churn allows proactive measures to retain customers [8], [9].

In this study some famous machine learning models like LogisticRegression (LR), Stochastic Gradient Descent (SGD) classifier, Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Decision Tree (DT) Classifier, Random Forest (RF) Classifier, Bagging (BGC) Classifier and Extreme Gradient Boosting (XGB) Classifier combinedly named as "Models". Hyper parameter tuning and Various data imbalance handling techniques like K-Cross Fold, With Over Sampling, Under Sampling, SMOTE, ENN known as "Techniques" are applied to predict churn of telecom customer. This study also includes description of different researches on churn prediction, briefing up of machine learning models utilized with various ensemble techniques. The accuracy of different machine learning models was retrieved from the confusion matrix and the comparison with previously published research work is also shown. As a closer remark the conclusion and the future scope of this work are illustrated.

The primary key contribution of our work are as follows:

- We have used dataset of one of the esteemed telecom company and applied preprocessing on it to handle null values and missing values. Then We have applied “Models” on the dataset and observed results of customer churn.
- After this we have used correlation matrix and dropped some of the attributes of the dataset. All “Models” were then implemented on this reduced dataset and observed the accuracy of churn prediction which was almost similar to previously calculated accuracy. This assured that attributes which were not responsible for churn were dropped correctly.
- Then hyperparameter tuning and imbalanced data handling techniques called “Techniques” combinedly were used with “Models” and accuracy to predict churn is observed.
- In this work we have compared observed results with published research also and highlighted the achieved increment in the accuracy to predict customer churn in telecom industry.
- This study is not only proposing the best suitable model to predict customer churn but also the reduced dataset which helps in predicting churners. Approximately 10% accuracy in customer churn prediction is also achieved through this work.

The remainder of the document is structured as follows: The research on customer churn prediction conducted by several researchers is presented in Section II. Machine learning models and different techniques to enhanced accuracy of classification, are the main topic of Section III. Section IV is the elaboration of the proposed methodology used to obtain reduced dataset and best machine learning classifier to predict customer churn. The graphical and tabular representation of the result observed through research work using different classification models and also retrieved the confusion matrix with comparison with previously published research work is shown in Section V. Conclusion and the future scope of this domain and is highlighted in Section VI. Telecom firms can lower customer attrition, raise customer lifetime value, and keep a competitive edge in the market by using an approach of effective churn analysis.

II. RELATED WORK

As the churn prediction is not new for service providers and researchers so a lot of work has been taken into existence. Summary of the researchers work is as follows - Praveen et al. [5] have created a thorough method that makes use of feature selection, analysis, prediction models, and data preparation to forecast customer attrition in the telecom industry. For model evaluation and hyperparameter adjustment, they employed boosting and ensemble approaches, SVM, Random Forest, Decision Trees, Logistic Regression, Naive Bayes, K-fold Cross-validation, and Random Forest. The Adaboost and XGBoost classifiers achieved high accuracy rates of 81.71% and 80.8%, respectively, through feature analysis and data pre-processing.

Asthana P. et al. [6] forecasted telecom customer attrition using Monte Carlo simulations and five classification algorithms. Naïve Bayes and Logistic Regression performed poorly, whereas two-layer Back-Propagation Network and Decision Tree were the top methods. AdaBoost M1 boosting produced better results, with an accuracy of almost 97% and an F-measure of over 84%.

Ullah B. R. et al. [7] stated that the telecom sector generates large amounts of data daily, making it crucial for decision-makers to understand customer churn. A churn pre-

Authors Name	Title	Year	Technique	Max. Accuracy
Wenjie *et al.* [10]	Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn	2016	LR, NN, DL	74.6%
Mitkees *et al.* [11]	Customer churn prediction model using data mining techniques	2017	SVM, DT, LR	80.01%
Saghir *et al.* [12]	Churn Prediction using Neural Network based Individual and Ensemble Models	2019	SVM, DT, ANN, Bagging MLP	80.86%
Kavitha *et al.* [13]	Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms	2020	DT, RF and XGBoost	80.00%
Rani *et al.* [14]	Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression	2021	LR	79.00%
Nigam *et al.* [15]	Effectual Predicting Telecom Customer Churn using Deep Neural Network	2019	Deep neural network	85.00%
Fujo *et al.* [16]	Customer Churn Prediction in Telecommunication Industry Using Deep Learning	2022	XG Boost, LR, Naive Bayes, KNN and Deep BP ANN	88.12%
Khattak *et al.* [17]	Customer churn prediction using composite deep learning technique	2023	BiLSTMCNN	81.00%
Poudel *et al.* [18]	Explaining customer churn prediction in telecom industry using tabular machine learning models	2024	Neural Network, SVC, LR, AdaBoost, XGBoost, RF, GBM	81.00%

Table 1. Previous contribution of the researchers

diction model is proposed using classification and clustering techniques, with the Random Forest algorithm performing well with 88.63% correctly classified instances. This model improves productivity, recommends promotions based on similar behaviour patterns, and enhances marketing campaigns.

Prabadevi B. et al. [8] proposed the goal of accurately forecasting customer responses by utilizing machine learning techniques and evaluating customer data prior to churn occurrences. The study predicts customer turnover using machine learning techniques such as logistic regression, random forest, k-nearest neighbors, and stochastic gradient booster, with accuracy rates of 78.1%, 82.6%, 82.9%, and 83.9%. The study highlights data pretreatment and hyperparameter optimization for enhanced model performance, as well as refining these algorithms for early customer churn prediction.

Teuku A. R. et al. [9] followed the data preparation, cleansing, and transformation, exploratory data analysis (EDA), prediction model design, and accuracy, F1 Score, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC) score analysis among the steps included in the research. The study forecasted customer attrition in the telecom industry using machine learning techniques such as XG Boost Classifier, Bernoulli Naïve Bayes, and Decision Tree. There was data purification, model building, EDA, and performance analysis. With a high accuracy of 81.59% and an F1_Score of 74.76%, the XG Boost Classifier performed admirably. In terms of efficiency, the Bernoulli Naïve Bayes and Decision Tree models outperformed.

Muthupriya V. et al. [19] proposed XGBoosted decision trees for customer churn research make use of the XGBoost algorithm, an enhanced version of gradient boosting. XGBoost is a machine learning technique that forecasts customer attrition by employing regularization techniques and parallel processing. Large datasets may be handled by it with remarkable accuracy. Customer churn analysis with XGBoost can be used to pinpoint important factors and create mitigation plans. XGBoost predicts with an accuracy rating of 82.1% when compared to other models. Some of the brief information of previous research on this domain is also shown in Table 1

III. MACHINE LEARNING MODELS & TECHNIQUES

The detailed overview of various machine learning algorithms used in churn prediction and their handling of imbalanced datasets, which is a common issue in classification tasks is studied during the review of the papers. Here's an elaboration on the machine learning techniques mentioned in the survey:

A. LOGISTIC REGRESSION (LR)

Recognized for its statistical approach to binary classification, Logistic Regression (LR) is a model that calculates the probability of a binary outcome. It is grounded in the concept of fitting data to a logistic curve, which is particularly adept at distinguishing between two possible outcomes. This model is highly valued across various fields, including spam detection, medical diagnostics, and predicting customer behavior such as the likelihood of churn. The model's popularity stems from its interpretability and straightforward implementation, with probabilities being computed within a 0 to 1 range using the logistic function. A probability threshold, typically 0.5, is used to determine class membership [19], [20].

B. STOCHASTIC GRADIENT DESCENT (SGD)

The Stochastic Gradient Descent (SGD) Classifier is an optimized linear classifier tailored for handling voluminous datasets. It's a go-to model for text classification and natural language tasks because of its capacity to process large amounts of data efficiently. SGD Classifier operates by adjusting its parameters incrementally for each training

example through a method known as stochastic gradient descent. This iterative process contributes to the model's computational efficiency and makes it well-suited for tasks requiring online learning, where data arrives sequentially. The flexibility of SGD allows it to adapt to different classification challenges, although it may require fine-tuning of hyperparameters like learning rate and regularization strength to achieve the best model performance [8].

C. GAUSSIAN NAIVE BAYES (GNB)

The Gaussian Naive Bayes classifier is based on applying Bayes' theorem with the assumption of independence between features, a common simplification for computational ease. Despite its simplicity, the model is effective for datasets where features follow a Gaussian distribution. It's particularly useful in high-dimensional spaces, making it a strong contender for text classification and document categorization. The Gaussian Naive Bayes model is appreciated for its efficiency and capacity to perform well on small to medium-sized datasets [5], [19].

D. SUPPORT VECTOR MACHINE (SVM)

Not explicitly elaborated upon in the provided text, SVM is a powerful supervised machine learning algorithm that excels in classification and regression tasks within high-dimensional spaces. SVMs are designed to find the hyperplane that maximizes the margin between classes, thereby enhancing the model's generalization capability. This contributes to its widespread application in various domains, including image recognition, text classification, and bioinformatics.

E. K-NEAREST NEIGHBORS (KNN)

The KNN algorithm is a non-parametric, instance-based learning method that classifies a data point based on the majority vote of its nearest neighbors. It's simple yet effective for datasets with clear boundaries between classes. The algorithm's ease of use and interpretability make it a popular choice for classification tasks [2], [21].

F. DECISION TREE (DT)

Decision Trees are versatile models that segment the data into subsets based on feature values, creating a tree-like structure. They are intuitive and easy to understand but can be prone to overfitting, especially if the tree becomes too complex [22].

G. RANDOM FOREST (RF)

Random Forests improve upon the simplicity of Decision Trees by creating an ensemble of trees and aggregating their predictions to produce a more stable and accurate output. This method effectively addresses the overfitting issue common in single Decision Trees [22].

H. BAGGING CLASSIFIER (BGC)

Bagging (Bootstrap Aggregating) is an ensemble technique that increases the stability and accuracy of machine learning models by combining the predictions of multiple base learners [23]. It often uses Decision Trees as base learners and is implemented in tools like Scikit-learn [24].

I. XGBOOST (XGBC)

XGBoost stands for Extreme Gradient Boosting, which is a highly efficient and flexible gradient boosting framework. It's known for handling missing values, providing regularization to prevent overfitting, and optimizing various objective functions, making it a powerful tool for classification challenges [9].

The paper also discusses techniques used to improve model accuracy when faced with imbalanced datasets:

- K-Cross Fold Validation: This technique involves dividing the dataset into K folds and using each fold once as a validation set while training the model on the remaining folds. This helps in obtaining a more accurate estimate of the model's performance and ensuring that the model is not overfitting to a particular subset of the data [5].
- Under Sampling: This approach addresses the imbalance by reducing the size of the over-represented class. By doing so, it creates a more balanced distribution, which can help in improving the predictive performance of the model [21], [25].
- SMOTE (Synthetic Minority Over-sampling Technique): SMOTE generates synthetic samples for the minority class, thus balancing the dataset without losing valuable information. This is particularly useful in situations where the minority class is underrepresented, and traditional over-sampling would lead to overfitting [25].
- Edited Nearest Neighbor (ENN): ENN is a cleaning technique that removes any instance in the dataset that does not agree with the majority of its K nearest neighbors. This method is useful for reducing noise in the dataset, thereby improving the classifier's performance.

The literature survey suggests that various machine learning models and techniques have their strengths and weaknesses, and the choice of model or technique can significantly impact the performance of churn prediction. The authors note that the telecom industry's ability to forecast customer attrition has improved with the application of these advanced machine learning methods and techniques

IV. PROPOSED METHODOLOGY

We have taken a dataset from one of the esteemed telecom companies. The dataset was first pre-processed in which the null values and missing values were processed. Then the machine learning models which are named as "Models" in the paper were applied and churn prediction accuracy was

observed. Here Precision, Recall and F1 Score has also taken into consideration to analyse classification results between churners and non-churners. Then to obtain improved dataset

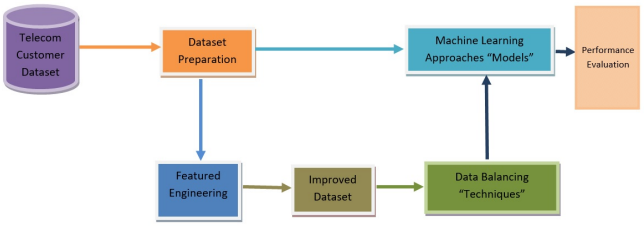


Figure 2. Proposed Customer Churn Prediction Approach

with less features the feature engineering was applied. Based on correlation matrix, attributes which has no effect on churn were identified and then dropped from dataset. Thus, the improved dataset with lesser number of attributes was identified. All "Models" were again applied on the improved dataset with reduced attribute and the accuracy of churn prediction was observed. This time the accuracy was near about same as previously obtained accuracy, this assure us that correct attributes were dropped. Now the hyper parameter tuning, and imbalanced data handling techniques were in the focus. This avoids the differences in ration of churners and non-churners so that overfitting and underfitting problem of data can be avoided to predict customer churn. Then "Models" along with the "Techniques" were applied and accuracy of predicting churners was observed again. Thus, this proposed approach for churn prediction helps us to extract reduced attribute dataset, and the best suitable churn classifier model was also proposed by comparing accuracy's. The accuracy of churn prediction then compared with earlier published results of researchers which turn into the validation of all the work done in this research paper.

V. RESULT & ANALYSIS

In the preceding section, it was outlined that this research utilizes a dataset obtained from one of the prominent telecom companies renowned for its substantial market presence and extensive customer base. The dataset underwent a meticulous cleaning process to address and rectify any missing or null values, which is a crucial step in ensuring the integrity and reliability of the data for subsequent analysis. It comprises a total of 21 attributes, among which one attribute, labelled as "Churn," serves as the target variable that the research aims to predict. The "Churn" variable is classified as dependent, meaning its outcomes are influenced by the other variables present in the dataset. The remaining attributes are categorized as independent variables, which include various factors that may contribute to a customer39;s decision to leave the service. Additionally, the dataset initially included a "CustomerID" attribute, which is simply a unique identification number assigned to each customer. This attribute was determined to be extraneous for the analysis, as it holds no significant predictive value regarding customer churn. Consequently, the decision was made to exclude this attribute from the dataset, resulting in a refined dataset that now contains 19 attributes, all poised for further analytical work. In this research, the evaluation of model performance extends beyond mere accuracy rates. It includes a comprehensive assessment of several key metrics to measure classification performance effectively. These metrics involve Precision, Recall, and the F1-Score, which collectively provide a more nuanced view of how well the models are identifying instances of churn. By focusing on these additional metrics, the research aims to ensure that the models developed are not only accurate but also reliable in minimizing false positives and false negatives, thereby enhancing the overall quality and applicability of the findings. It was observed that many researchers and using each attribute of the dataset to perform churn prediction. So, in this work we have used a correlation matrix, through which some of the attributes which didn't impact churn were identified. Thus, three attributes "gender", "Senior Citizen" and Dependents" were dropped from the dataset and the improved dataset with reduced attributes came into existence. Thus, we had two versions of the same dataset, one with all 19 attributes and the other version called as dataset "with reduced attributes" which contained 16 attributes. Dataset with all attributes was directly used with "Models" to predict customer churn and results are shown in Table 2 (Sub Section 2.1).

2.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	80.28	0.85	0.64	0.89	0.55	0.87	0.59
SGD	78.25	0.85	0.59	0.85	0.6	0.85	0.59
GNB	74.55	0.9	0.51	0.7	0.77	0.81	0.62
SVM	79.74	0.83	0.65	0.9	0.5	0.87	0.57
KNN	76.75	0.83	0.56	0.85	0.53	0.84	0.55
DT	72.7	0.82	0.48	0.81	0.48	0.81	0.48
RF	79.95	0.84	0.66	0.91	0.51	0.87	0.57
BGC	78.18	0.82	0.63	0.91	0.44	0.86	0.51
XGBC	78.82	0.84	0.62	0.89	0.51	0.86	0.56
2.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	80.02	0.85	0.64	0.89	0.55	0.87	0.59
SGD	78.25	0.85	0.59	0.85	0.6	0.85	0.59
GNB	74.55	0.9	0.51	0.74	0.77	0.81	0.62
SVM	76.74	0.83	0.65	0.9	0.5	0.87	0.57
KNN	72.75	0.83	0.56	0.85	0.53	0.84	0.53
DT	72.7	0.81	0.48	0.81	0.48	0.81	0.48
RF	79.95	0.84	0.66	0.91	0.51	0.87	0.57
BGC	78.18	0.82	0.63	0.91	0.44	0.86	0.51
XGBC	78.82	0.84	0.62	0.89	0.51	0.89	0.56

Table 2. Direct Implementation

Subsequently the dataset with reduced attributes then feed to our "Models" and churn prediction results were obtained shown in Table 2 (Sub Section 2.2). So, Table 2 shows almost no changes in the observed accuracy which proves the correct attributed were dropped. Now further work on

both versions of dataset could be proceeded. Then the dataset with all attributes and with reduced attributes was used with "Models" along with "Techniques" and the results are shown in Table 3 to Table 8.

3.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	80.88	0.86	0.63	0.89	0.57	0.87	0.6
SGD	78.84	0.87	0.57	0.85	0.62	0.86	0.6
GNB	74.48	0.9	0.5	0.74	0.76	0.81	0.6
SVM	81.33	0.85	0.67	0.92	0.5	0.88	0.57
KNN	76	0.84	0.52	0.84	0.52	0.84	0.52
DT	73.6	0.84	0.48	0.8	0.54	0.82	0.51
RF	79.2	0.85	0.6	0.88	0.52	0.86	0.56
BGC	78.4	0.83	0.59	0.9	0.45	0.86	0.51
XGBC	79.02	0.85	0.59	0.87	0.54	0.86	0.57
3.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	81.06	0.85	0.69	0.89	0.6	0.87	0.64
SGD	78.22	0.85	0.62	0.85	0.6	0.85	0.61
GNB	76.35	0.91	0.56	0.75	0.81	0.82	0.66
SVM	80.71	0.82	0.73	0.93	0.5	0.87	0.59
KNN	76.08	0.82	0.58	0.85	0.54	0.84	0.56
DT	71.2	0.8	0.49	0.79	0.5	0.8	0.5
RF	77.68	0.82	0.63	0.88	0.4	0.83	0.46
BGC	77.93	0.79	0.56	0.88	0.4	0.83	0.46
XGBC	77.86	0.82	0.63	0.88	0.52	0.85	0.57

Table 3. K Cross Fold (k=5)

4.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	75.65	0.77	0.75	0.72	0.79	0.74	0.77
SGD	74.58	0.81	0.71	0.63	0.86	0.71	0.78
GNB	74.25	0.76	0.73	0.69	0.79	0.72	0.76
SVM	77.63	0.79	0.77	0.74	0.81	0.76	0.79
KNN	76.13	0.8	0.73	0.68	0.84	0.74	0.78
DT	86.1	0.92	0.82	0.79	0.93	0.85	0.87
RF	89.54	0.94	0.86	0.84	0.93	0.88	0.89
BGC	88.43	0.92	0.86	0.84	0.93	0.88	0.89
XGBC	85.04	0.9	0.82	0.79	0.91	0.84	0.86
4.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	77.2	0.78	0.76	0.75	0.8	0.76	0.78
SGD	76.28	0.76	0.76	0.75	0.77	0.76	0.77
GNB	76.76	0.78	0.76	0.73	0.8	0.76	0.78
SVM	78.81	0.82	0.78	0.76	0.84	0.79	0.81
KNN	77.59	0.82	0.74	0.69	0.86	0.75	0.8
DT	87.36	0.93	0.83	0.81	0.94	0.86	0.88
RF	90.9	0.95	0.88	0.87	0.95	0.9	0.91
BGC	88.67	0.92	0.86	0.84	0.93	0.88	0.89
XGBC	86.2	0.9	0.83	0.81	0.91	0.85	0.87

Table 4. OverSampling

The graphical representation of observed accuracy is also shown using Fig. 4. In this figure we have represented Figure 4a to Fig. 4i. Table 8 provides a comprehensive comparison of accuracy across various models and techniques when applied to both the complete set of attributes and a dataset with reduced attributes. This table serves as a crucial reference point for evaluating the performance of each model in

5.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	76.07	0.81	0.73	0.69	0.83	0.74	0.78
SGD	76.07	0.77	0.75	0.75	0.77	0.76	0.76
GNB	74.86	0.8	0.71	0.67	0.82	0.73	0.77
SVM	76.47	0.82	0.73	0.69	0.84	0.75	0.78
KNN	73.52	0.78	0.7	0.66	0.81	0.71	0.75
DT	66.84	0.67	0.66	0.67	0.67	0.67	0.67
RF	75.53	0.78	0.73	0.71	0.8	0.75	0.76
BGC	73.12	0.74	0.72	0.72	0.74	0.73	0.73
XGBC	73.39	0.78	0.7	0.66	0.81	0.72	0.75
5.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	76.6	0.77	0.76	0.73	0.8	0.75	0.78
SGD	75.95	0.81	0.68	0.54	0.88	0.65	0.77
GNB	75.26	0.77	0.74	0.69	0.81	0.73	0.77
SVM	76.33	0.77	0.76	0.72	0.8	0.74	0.78
KNN	71.92	0.72	0.72	0.67	0.76	0.7	0.74
DT	66.97	0.65	0.69	0.66	0.68	0.66	0.68
RF	75.13	0.73	0.77	0.76	0.75	0.74	0.76
BGC	72.72	0.7	0.76	0.77	0.69	0.73	0.73
XGBC	72.05	0.71	0.73	0.71	0.73	0.71	0.73

Table 5. UnderSampling

6.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	79.18	0.8	0.78	0.76	0.82	0.78	0.8
SGD	78.12	0.83	0.75	0.7	0.86	0.76	0.8
GNB	77.59	0.79	0.76	0.74	0.81	0.77	0.79
SVM	79.47	0.81	0.78	0.76	0.83	0.79	0.8
KNN	79.52	0.86	0.75	0.69	0.89	0.77	0.82
DT	77.34	0.78	0.77	0.76	0.79	0.77	0.78
RF	84.07	0.87	0.82	0.8	0.88	0.83	0.85
BGC	81.89	0.82	0.81	0.8	0.83	0.81	0.82
XGBC	82.91	0.85	0.81	0.79	0.87	0.82	0.84
6.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	79.86	0.81	0.79	0.77	0.83	0.79	0.81
SGD	79.67	0.82	0.78	0.75	0.84	0.78	0.81
GNB	77.92	0.8	0.76	0.73	0.83	0.76	0.79
SVM	80.05	0.81	0.79	0.77	0.83	0.79	0.81
KNN	81.17	0.87	0.77	0.72	0.9	0.79	0.83
DT	79.13	0.78	0.8	0.89	0.79	0.79	0.79
RF	84.41	0.86	0.83	0.82	0.87	0.84	0.85
BGC	81.51	0.81	0.82	0.81	0.82	0.81	0.82
XGBC	83.05	0.85	0.82	0.8	0.86	0.82	0.84

Table 6. SMOTE

relation to the different techniques employed for analysis. Additionally, Fig. 3 visually represents the results observed in Table 8, facilitating a clearer understanding of the accuracy metrics. The graphical representation demonstrates that the performance outcomes show minimal to no significant variations. This consistency in results strongly indicates that the process of reducing the dataset was implemented effectively, ensuring that the essential characteristics and predictive capabilities of the original dataset were preserved.

This finding reinforces the validity of utilizing a reduced dataset without compromising accuracy, which is a critical

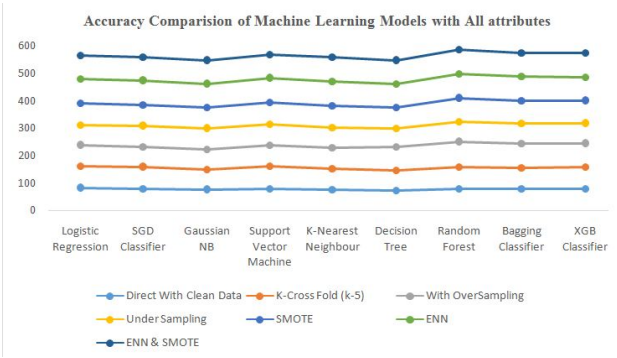
7.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	87.67	0.9	0.84	0.9	0.83	0.9	0.84
SGD	87.07	0.88	0.85	0.92	0.8	0.9	0.82
GNB	84.64	0.89	0.78	0.86	0.83	0.87	0.8
SVM	87.98	0.9	0.85	0.92	0.83	0.9	0.84
KNN	88.58	0.91	0.85	0.9	0.86	0.91	0.85
DT	84.94	0.9	0.78	0.85	0.85	0.8	0.81
RF	90.3	0.92	0.88	0.93	0.86	0.92	0.87
BGC	89.09	0.9	0.87	0.92	0.84	0.91	0.85
XGBC	88.58	0.91	0.85	0.91	0.85	0.91	0.85
7.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	86.38	0.9	0.82	0.9	0.81	0.9	0.81
SGD	86.13	0.88	0.88	0.94	0.77	0.91	0.82
GNB	83.3	0.9	0.78	0.88	0.82	0.89	0.8
SVM	85.38	0.89	0.87	0.94	0.78	0.91	0.82
KNN	87.47	0.91	0.85	0.92	0.82	0.91	0.84
DT	82.6	0.87	0.76	0.87	0.77	0.87	0.76
RF	89.16	0.9	0.9	0.95	0.81	0.93	0.85
BGC	87.17	0.88	0.85	0.92	0.78	0.9	0.81
XGBC	87.73	0.9	0.84	0.91	0.82	0.91	0.83

Table 7. ENN

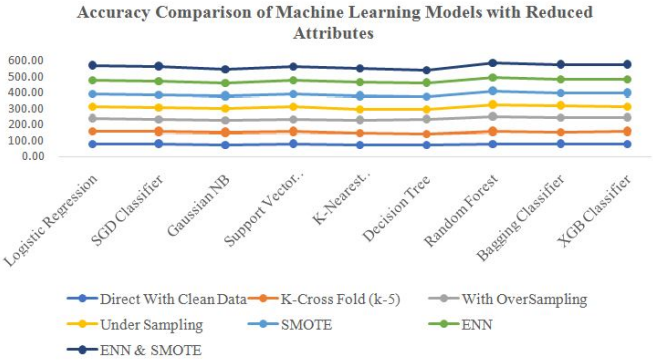
8.1-With all Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	87.67	0.85	0.64	0.89	0.55	0.87	0.59
SGD	87.07	0.85	0.59	0.85	0.6	0.85	0.59
GNB	84.64	0.9	0.51	0.74	0.77	0.81	0.62
SVM	87.98	0.83	0.85	0.9	0.5	0.87	0.57
KNN	88.58	0.83	0.56	0.85	0.53	0.84	0.55
DT	84.94	0.81	0.48	0.81	0.48	0.81	0.48
RF	90.3	0.84	0.66	0.91	0.51	0.87	0.57
BGC	89.09	0.82	0.63	0.91	0.51	0.87	0.57
XGBC	88.58	0.84	0.62	0.89	0.51	0.86	0.56
8.2-With Reduced Attributes							
ML	Accuracy	Precision		Recall		F1-Score	
		Class0	Class1	Class0	Class1	Class0	Class1
LR	87.52	0.9	0.82	0.9	0.81	0.9	0.81
SGD	88.17	0.88	0.88	0.94	0.77	0.91	0.82
GNB	85.58	0.9	0.78	0.88	0.82	0.89	0.8
SVM	88.27	0.89	0.87	0.94	0.78	0.91	0.82
KNN	88.66	0.91	0.85	0.92	0.82	0.91	0.84
DT	83.2	0.87	0.76	0.87	0.77	0.87	0.76
RF	90.06	0.9	0.9	0.95	0.81	0.93	0.85
BGC	87.27	0.88	0.85	0.92	0.78	0.9	0.81
XGBC	88.07	0.9	0.84	0.91	0.82	0.91	0.83

Table 8. ENN+SMOTE

consideration in data analysis and model development. The validation process for the findings of this research study involves a comprehensive comparison of the results obtained with those from previously published works in the field. To illustrate this comparison, Table 9 has been constructed, displaying the maximum accuracy rates achieved by various researchers, including our own contributions to the literature. This systematic presentation of data allows for an easy visual representation of how our models and techniques stand relative to prior efforts in the analysis of customer churn within the telecommunications sector. Upon reviewing



(a) Dataset with all attributes

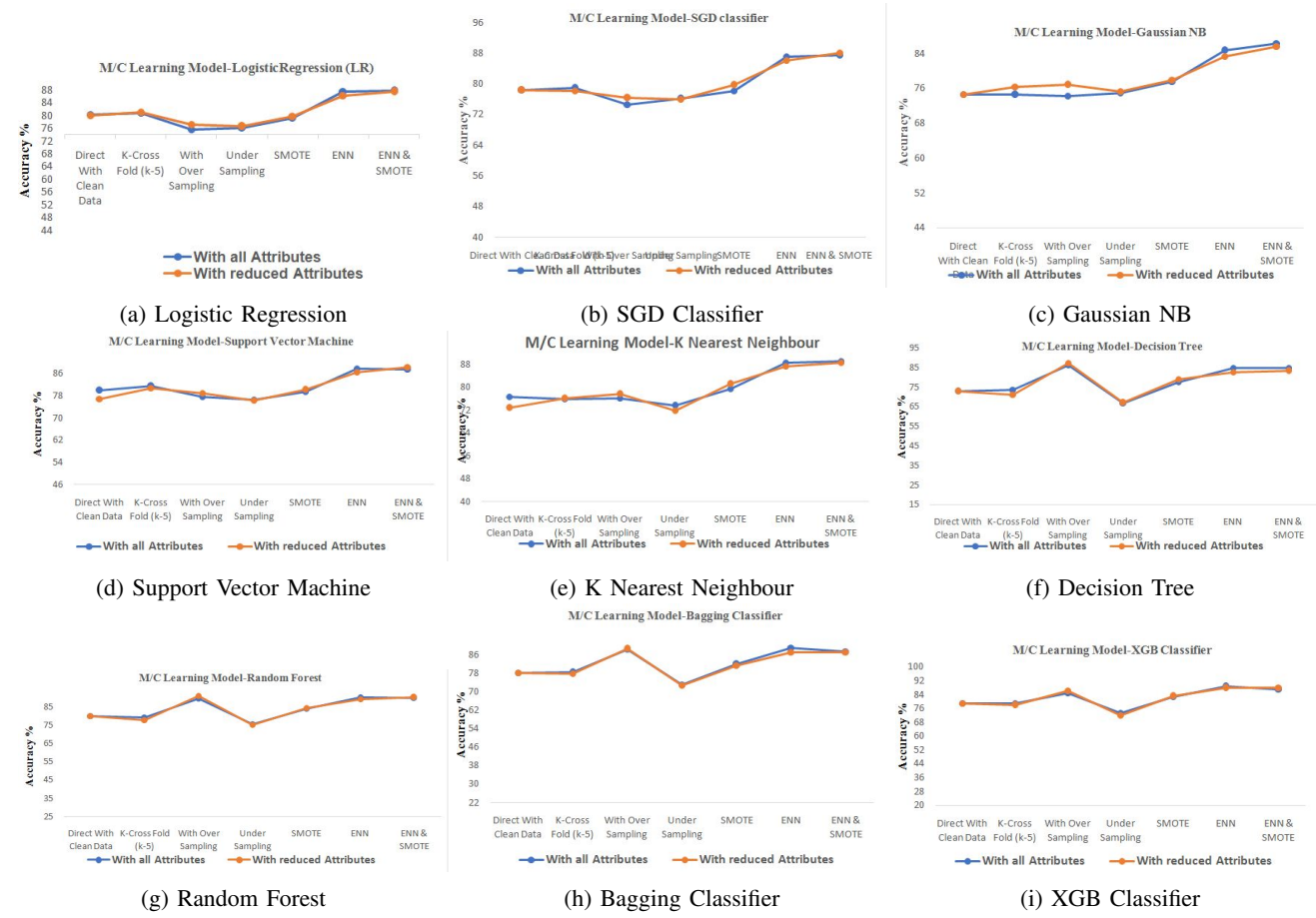


(b) Dataset with Reduced attributes

Figure 3. Accuracy Comparision of "Models" along with "Techniques"

the data presented in Table 9, it becomes evident that our models, when paired with the specified techniques, exhibit robust performance across all listed formats. Notably, our research has yielded a significant improvement in accuracy, with an increase of approximately 8 to 10 percent in predicting which customers are likely to churn. This enhancement underscores the effectiveness of our approach in addressing this crucial issue faced by telecom companies. Furthermore, among the various machine learning algorithms evaluated in this study, the Random Forest algorithm stands out as particularly effective. It has achieved an accuracy rate exceeding 90 percent, establishing its position as the most reliable model for predicting customer churn in the telecommunications industry. This high level of accuracy not only highlights the strength of the Random Forest method but also reinforces the overall success of our research efforts in contributing valuable insights into customer retention strategies.

Notably, the Random Forest classifier emerged as the most effective, achieving 90.30% accuracy with all attributes and 90.90% with a reduced attribute set as illustrated in Table 10. Comparative analysis with prior research indicates the suitability of the reduced attribute dataset for churn prediction across industries can be given preference to be used with classification models. All other authors in Table 10 have used all the parameters available in the dataset



(a) Logistic Regression (b) SGD Classifier (c) Gaussian NB

(d) Support Vector Machine (e) K Nearest Neighbour (f) Decision Tree

(g) Random Forest (h) Bagging Classifier (i) XGB Classifier

Figure 4. Accuracy Comparison of "Models" with all and with reduced attributes

9.1- Accuracy Comparison by Using All Attributes									
Techniques/Models	LR	SGD	GNB	SVM	KNN	DT	RF	BGC	XGBC
Direct With Clean Data	80.28	78.25	74.55	79.74	76.75	72.7	79.95	78.18	78.82
K-Cross Fold (K-5)	80.88	78.84	74.48	81.33	76	73.6	79.2	78.4	79.02
With Over Sampling	75.65	74.58	74.25	77.63	76.13	86.1	89.54	88.43	85.04
Under Sampling	76.07	76.07	74.86	76.47	73.52	66.84	75.53	73.12	73.39
SMOTE	79.18	78.12	77.59	79.47	79.52	77.34	84.07	81.89	82.91
ENN	87.67	87.07	84.64	87.65	88.58	84.66	90.3	89.09	88.58
ENN and SMOTE	87.73	87.47	86.28	87.37	89.16	84.89	89.86	87.57	87.17
9.2- Accuracy Comparison by using Reduced Set of Attributes									
Techniques/Models	LR	SGD	GNB	SVM	KNN	DT	RF	BGC	XGBC
Techniques/Models	LR	SGD	GNB	SVM	KNN	DT	RF	BGC	XGBC
Direct With Clean Data	80.02	78.25	74.55	76.74	72.75	72.7	79.95	78.18	78.82
K-Cross Fold (K-5)	81.06	78.22	76.35	80.71	76.08	71.2	77.68	77.93	77.86
With Over Sampling	77.2	76.28	76.76	78.81	77.59	87.36	90.9	88.67	86.2
Under Sampling	76.6	75.95	75.26	76.33	71.92	66.97	75.13	72.72	72.05
SMOTE	79.86	79.67	77.92	80.05	81.17	79.13	84.41	81.51	83.05
ENN	86.38	86.13	83.3	86.38	87.47	82.6	89.16	87.17	87.73
ENN and SMOTE	87.52	88.17	85.58	88.27	88.66	83.2	90.06	87.27	88.07

Table 9. Accuracy Comparison

whereas we have used same dataset to apply "Models" and "Techniques". After meticulous examination of churn prediction using machine learning, our work highlighting the Random Forest classifier's efficacy and advocating for reduced attribute datasets. Decisively, the authors assert the superiority of the Random Forest model in telecom sector churn prediction. The study is bolstered by references to prior works, and it sets forth avenues for future research

Authors/Models	LR	GNB	KNN	DT	RF	SVM	XGBC
Muthupriya V. et al. [19]	76.57	77.07	-	-	81.21	80.21	82.1
Prabadevi B. et al. [8]	82.6	-	78.1	-	82.9	-	-
Teuku Alif et al. [9]	80.89	-	78.96	79.89	81.24	81.24	81.59
Lalwani, P. et al. [5]	80.45	75.86	79.64	80.14	78.04	80.21	80.8
With All Attributes	87.67	84.64	88.58	86.1	90.3	87.98	88.58
With Reduced Attributes	87.52	85.58	88.66	87.36	90.9	88.27	88.07

Table 10. Accuracy Comparison with previously published researchers

aimed at enhancing churn prediction precision.

VI. CONCLUSIONS

The paper aims to ascertain the most optimal machine learning model for churn prediction while exploring the impact of reducing the dataset's attributes on predictive accuracy. The authors utilized an array of machine learning models, encompassing Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest, Bagging Classifier, Stochastic Gradient Descent Classifier, Gaussian Naive Bayes, and Extreme Gradient Boosting (XGB) Classifier combinedly named as "Models" in the work. These "Models" underwent testing on a clean dataset, followed by application of techniques such as K-Cross Fold validation, Over Sampling, Under Sampling, SMOTE (Synthetic Minority Over-sampling Technique), and Edited Nearest Neighbour (ENN) combinedly known as "Techniques" to address data imbalance and enhance model efficacy. Conclusively in this work we can say that for LR, GNB, KNN, DT, RF, SVM and XGBC have used same dataset as used by all four authors. With all attributes and with reduced attributes to predict customer churn in telecom industry our stated all classifiers have achieved better churn prediction accuracy shown in Table 10. Results are presented through tables and graphs, showcasing accuracy, precision, recall, and F1 score metrics for each model. The work advocate for future investigations into reinforcement learning, deep learning, and temporally ordered datasets to further refine churn prediction accuracy.

References

[1] A. Bhatnagar and S. Srivastava, "Performance analysis of hoeffding and logisitic algorithm for churn prediction in telecom sector," in 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, 2020, pp. 377–380.

[2] A. Bhatnagar and S. Srivastava, "A robust model for churn prediction using supervised machine learning," in 2019 IEEE 9th international conference on advanced computing (IACC). IEEE, 2019, pp. 45–49.

[3] A. Saran Kumar and D. Chandrakala, "A survey on customer churn prediction using machine learning techniques," International Journal of Computer Applications, vol. 975, p. 8887, 2016.

[4] A. J. Petkovski, B. L. Risteska Stojkoska, K. V. Trivodaliev, and S. A. Kalajdziski, "Analysis of churn prediction: A case study on telecommunication services in macedonia," in 2016 24th Telecommunications Forum (TELFOR), 2016, pp. 1–4.

[5] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, vol. 104, no. 2, pp. 271–294, 2022.

[6] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.

[7] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," IEEE access, vol. 7, pp. 60 134–60 149, 2019.

[8] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," International Journal of Intelligent Networks, vol. 4, pp. 145–154, 2023.

[9] T. A. R. Akbar and C. Apriono, "Machine learning predictive models analysis on telecommunications service churn rate," Green Intelligent Systems and Applications, vol. 3, no. 1, pp. 22–34, 2023.

[10] W. Bi, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1270–1281, 2016.

[11] I. Mitkees, S. Badr, and A. Elseddawy, "Customer churn prediction model using data mining techniques," 12 2017, pp. 262–268.

[12] M. Saghir, Z. Bibi, S. Bashir, and F. H. Khan, "Churn prediction using neural network based individual and ensemble models," in 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). IEEE, 2019, pp. 634–639.

[13] V. Kavitha, G. H. Kumar, S. M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms," International Journal of Engineering Research & Technology (2278-0181), vol. 9, no. 05, pp. 181–184, 2020.

[14] K. S. Rani, S. Thaslima, N. Prasanna, R. Vindhya, and P. Srilakshmi, "Analysis of customer churn prediction in telecom industry using logistic regression," International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, pp. 2347–5552, 2021.

[15] B. Nigam, H. Dugar, and M. Niranjanamurthy, "Effectual predicting telecom customer churn using deep neural network," Int J Eng Adv Technol (IJEAT), vol. 8, no. 5, 2019.

[16] S. W. Fujo, S. Subramanian, M. A. Khder et al., "Customer churn prediction in telecommunication industry using deep learning," Information Sciences Letters, vol. 11, no. 1, p. 24, 2022.

[17] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," Scientific Reports, vol. 13, no. 1, p. 17294, 2023.

[18] S. S. Poudel, S. Pokharel, and M. Timilsina, "Explaining customer churn prediction in telecom industry using tabular machine learning models," Machine Learning with Applications, vol. 17, p. 100567, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827024000434

[19] V. Muthupriya, R. Narayanan, S. Nakeeb, and A. Abhishek, "Customer churn analysis using xgboosted decision trees," Indonesian Journal of Electrical Engineering and Computer Science, vol. 25, no. 1, pp. 488–495, 2022.

[20] H.-A. Park, "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain," Journal of Korean academy of nursing, vol. 43, no. 2, pp. 154–164, 2013.

[21] D. Ramadhanti, A. Larasati, A. Muid, and E. Mohamad, "Building customer churn prediction models in indonesian telecommunication company using decision tree algorithm," in AIP Conference Proceedings, vol. 2654, no. 1. AIP Publishing, 2023.

[22] J. Balogun, F. Kasali, and I. Akinyemi, "Development of classification model for the prediction of churn among customers using decision tree algorithm," Journal of Computer Science and Its Application, vol. 28, pp. 45–53, 08 2022.

[23] M. K. Awang, M. Makhtar, N. Udin, and N. F. Mansor, "Improving customer churn classification with ensemble stacking method," International Journal of Advanced Computer Science and Applications, vol. 12, no. 11, 2021.

[24] L. Makurumidze, W. S. Manjoro, and W. Makondo, "Implementing random forest to predict churn," International Journal of Computer Science and Mobile Computing, vol. 11, no. 2, pp. 75–84, 2022.

[25] S. Gore, Y. Chibber, M. Bhasin, S. Mehta, and S. Suchitra, "Customer churn prediction using neural networks and smote-enn for data sampling," in 2023 3rd international conference on artificial intelligence and signal processing (AISP). IEEE, 2023, pp. 1–5.

MR. ANURAG BHATNAGAR He has done his B.E. in Information Technology from Govt. Engineering College Ajmer, his M.Tech. (CSE) is from RITE Bhankrota which is affiliated with Rajasthan Technical University Kota, Rajasthan. He is pursuing his PhD from Manipal University Jaipur under the supervision of Dr. Sumit Srivastava. He has 8 patents and 4 copyrights in his research career till now. He has Q-1 and Q-3 publication in his research. Currently he is working as Asstt. Prof in IT at Manipal University Jaipur.

DR. SUMIT SRIVASTAVA He has done his MCA from BITS RANCHI, MTECH from KARNATAKA OPEN UNIVERSITY and his PHD is from UNIVERSITY OF RAJASTHAN, Jaipur. He is currently working in Manipal University Jaipur as Prof. in IT. He is also serving as Director of the School of Information Security and Data Science. He was the Faculty Coordinator for the Summer University Program in collaboration with the University of Applied Sciences, Western Switzerland. He is also an editor in all versions of the International Conference on Smart IoT Systems: Innovations in Computing (SSIC 2019) held at MUJ and a Member EXECOM-IEEE DELHI SECTION 2015-16. Dr Sumit is IEEE-Senior Member and has a ACM-Professional Membership.