Customer Churn Analysis and Prediction in Telecommunication Sector Implementing Different Machine Learning Techniques

Samprit Gowd[1], Aarati Mohite[2], Debashish Chakravarty[2], and Sanjay Nalbalwar[1](✉)

[1] Electronics and Telecommunications, Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad, India
nalbalwar_sanjayan@yahoo.com

[2] Electronics and Telecommunications, Indian Institute of Technology, Kharagpur, India

Abstract. Nowadays, a large number of telecom industries are dependent on retaining their existing customer base, as retaining customers is found to be more profitable than acquiring new customers. Due to immensely growing competition in this industry, customers get various choices of services and privileges and hence leading them to churn. This problem encourages data scientists to search for solutions to help telecom industries. In this research, 'The orange telecom churn dataset' from Kaggle is analyzed to determine the reasons for customer churning. Different machine learning algorithms viz. Decision Tree, k-nearest neighbor, Random Forest, Naïve Bayes and XGBoost are studied and analyzed for the dataset as mentioned earlier. Results are compared to find the best algorithm to solve the problem for churn prediction. Random Forest and XGBoost algorithms performed best along with the hyperparameter optimization and hence resulted in 95.20% and 95.65% accuracies respectively. Precision-recall curve, accuracy and F-score are the different metrics utilized for the evaluation purpose.

Keywords: Churn · machine learning · XGBoost · precision-recall curve · F-score · Customer churn prediction · Customer Relationship Management

1 Introduction

We live in an era, in which there are fully-fledged businesses and rigorous competitive pressure on the companies. Thus, to survive in the market and increase their income, companies have to maintain their relationship with their customers. This approach is known as "Customer Relationship Management" (CRM), which has the motive of ensuring customers' satisfaction [1]. One of the types of CRM is "Customer churn prediction" (CCP). A company tries to build a model that predicts if a customer is planning to quit the company or minimize its purchases from a company. Companies mostly work with machine learning techniques for customer churn prediction [2]. As there is a need to predict whether customers will stop using services or not, Customer churn is considered to be a classification problem. Customers are always in search of more reassurance and splendor. Therefore, churning has turned into a common trend these days [3]. To

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 686–700, 2023.
https://doi.org/10.2991/978-94-6463-196-8_52

provide their customers with more services and offers and increase customer retention, organizations have focused on CRM. Even after focusing on CRM and providing good services, the churning of the customer cannot be stopped. Thus, we can say that churn is an endless process, but it can be predicted to reduce its rate [3].

When it comes to the telecommunications segment, there are a lot of opportunities available for the customers to get better services. The decision of a customer in the telecom industry changes as per needs or experiences. Due to this, there are a lot of chances of customers getting churned to competitors in this telecommunication industry. Collecting new customers is found to be more expensive than retaining the existing customers. Therefore, companies focus on avoiding the churning of customers. As this industry deals with high dimension data, the publication of advanced artificial intelligence and data analytics techniques further help support this rich data to address churn much more effectively.

In this research, analysis of data is carried out to identify the various reasons for churning and a predictive model is built on the telecom-based dataset using different machine learning techniques. Various results are drawn from different algorithms to obtain the best model for our telecom-based dataset. The algorithms used are Decision Tree, Random Forest, k-nearest Neighbor, Naïve Bayes, XGBoost, and Artificial neural network. Grid search is the hyperparameter optimization technique used to improve accuracy. Along with the accuracy of the algorithm time complexity is also considered to find the best algorithm. To compare the results of the build models, different evaluating parameters like Precision-recall curve, loss and accuracy graph, F-score, and accuracy are used.

2 Literature Review

No industry in the market is not affected by customer churning. It is seen that much research is done to find the reasons of customer churning and to predict churning, in the field of data science. "Mr. Saran Kumar A. and Dr. Chandrakala D studied most machine learning algorithms and stated that combining SVM with the boosting algorithms can give higher accuracy for churn prediction" [4]. "Mr. Anurag Bhatnagar Manipal and Dr. Sumit Srivastava implemented Hoeffding Tree and logistic algorithm on the data and comparing the results, concluded that the logistic algorithm works better than the Hoeffding tree algorithm" [3]. "Different machine learning algorithms viz Logistic regression, Decision tree, random forest, K-nearest neighbor are applied to the banking industry dataset in work done by Ms. Ishpreet Kaur and Ms. Jasleen Kaur. Ensembling techniques like voting and averaging are used to improve accuracy. Here Random Forest shows the best results among all the algorithms" [5]. "Xin Hu, Yanfei Yang, Lanhua Chen, and Siru Zhu worked on building the combined customer churn prediction model by using prediction results and confidence of the decision tree prediction model and neural network prediction model. This integrated model compensated for most of the shortcomings of the single prediction model" [6]. "Essam Shaaban, Yehia Helmy, Ayman Khedr, and Mona Nasr are the authors of the research paper, in which data mining techniques like Support Vector Machine, Decision Tree, and Neural Network are used with open-source software called WEKA. The best output is given by the SVM algorithm" [9].

3 Methodology

In the telecommunication industry, it is essential to obtain the causes of customer churning. Data Analysis is the process of methodically implementing statistical or logical techniques to illustrate, describe, and estimate data [7]. In this work, data exploration is carried out to reach the root causes of churning and evaluate the dataset. Several characteristics of the dataset are acquired by exploring the dataset. One of the features is that the dataset is unbalanced. Which is found to be common in the telecommunication industry. This skewness of the dataset declines the performance of algorithms to predict the customers. To solve this problem, we used different algorithms and gain the best-performing predictive model. The steps involved in the proposed work are given in the block Fig. 1.

A. Dataset

The dataset named 'The orange telecom churn dataset' is downloaded from Kaggle. It contains 2666 rows and 20 columns. A single row denotes a customer while a column gives us the customer's attributes. The dataset contains the features such as Area code, State, Account length international plan, Voice mail plan, messages, Total day calls, Total day minutes, Number of vmail, etc. The churn column is the target for prediction.

B. Data Pre-processing

Data pre-processing is the technique of data mining which converts the unprocessed data into convenient and systematic data. It includes data cleaning, data transformation, and data reduction [8]. This technique is used, as the data contain some unrelated features which can be dropped. Exploratory Data Analysis (EDA) is implemented, to carry out

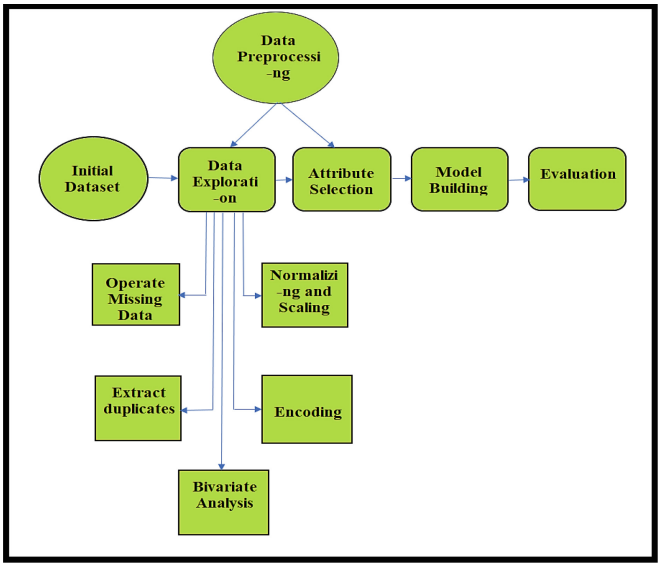

Fig. 1. Block Diagram for the methodology used.

various steps such as finding the missing values, normalizing data and scaling data. To obtain the relationship between two variables Bivariate Analysis is also applied. By finding the correlation between different attributes the unnecessary attributes are removed. Histogram, significant features and heatmap are also plotted.

C. Model Building

Different machine learning algorithms viz. Decision Tree, Random Forest, k-nearest neighbor, Naïve Bayes and XGBoost are implemented to build a churn prediction model. Artificial neural network is also implemented to achieve better churning prediction.

Decision Tree

"The decision tree algorithm falls under the type of supervised learning algorithm which has a predefined target variable. It has two major steps, tree building and tree pruning. The tree-building includes dividing the training sets according to the values of the attributes. The dividing process continues till we get the identical values in the records of the partitions. Some branches can be removed as they may have noisy data. The pruning step includes selecting and removing the branches having a large error rate. Tree pruning is the step that enriches the predictive accuracy of the decision tree and minimizes the difficulty" [9].

Random Forest

Random Forest is the algorithm implemented for the problems of classification and regression. Numerous Decision trees are used to make a final decision. It merges the output of multiple Decision Trees, which are randomly created to generate the final output as it uses the concept of ensemble learning [5]. Each decision tree gives a prediction result during the training phase. The final decision is estimated by the Random Forest based on the majority of the results when a new data point occurs.

k-Nearest Neighbor

k-nearest neighbor algorithm saves all the obtainable data and classifies a new data point according to similarity. KNN algorithm can be applied for both Regression and Classification, but more frequently it is used for Classification problems. KNN is considered a non-parametric algorithm. As KNN does not perceive any knowledge from the training set immediately but reserves the dataset, at the time of classification it acts on the dataset. Therefore, KNN is called a lazy learner algorithm.

Naïve Bayes

A Naive Bayes classifier is considered to be a probabilistic approach. Naïve Bayes' each vector feature is considered to be independent of each other. The assumption of this classifier that the value of each feature has an independent influence on a given class is called as class conditional independence. It is used to make the computations simple, and in this sense, we call it Naïve [10]. The principle of Naïve Bayes is dependent on Bayes' theorem.

XGBoost

XGBoost uses gradient boosting to apply the decision tree algorithm. In gradient boosting new models are used to evaluate the inaccuracy of the model which are previously applied. Then both the predictions are combined to make the final prediction [10].

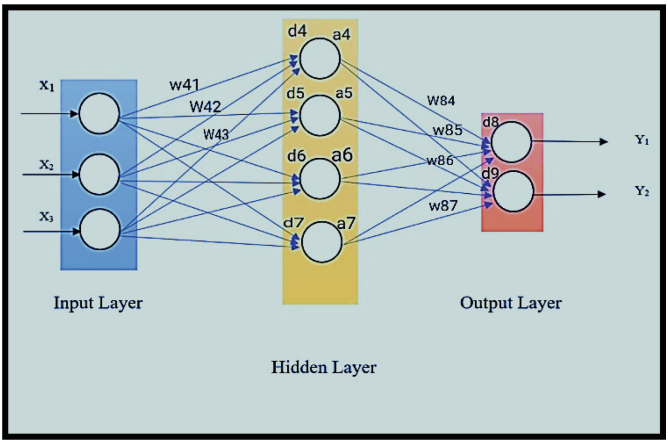

Fig. 2. Layers of ANN

Weights have an essential role in XGBoost. Weights are allocated to all the independent variables which are then fed into the decision tree which predicts results.

Artificial Neural Network

"All the decisions in the brain of a human, are taken by neural networks provided naturally in our body which are created of basic building block called "neuron". All the communications are performed in electrical signals through synapses, a connection point between dendrites and axons from preceding neurons. In the same way, in artificial neuron inputs X1, X2… Xn are taken by each neuron and given as input to the summation and activation function for decision making. The output is carried out on the basis of the joint decision taken by the whole neural network" [11]. The neuron is made up of three layers, which are shown in Fig. 2.

D. Evaluation

Precision-recall curve, F-score, and accuracy are the metrics used for the evaluation of the execution of applied algorithms. Precision and recall are found to be beneficial in cases where there is an imbalanced dataset. "Precision is computed by taking the fraction of the number of true positives divided by the addition of the true positives and false positives. The recall is the ratio of the number of true positives divided by the addition of the true positives and the false negatives" [12]. Integrating the precision and recall of the model gives the F1-score value. "Computational complexity includes the computation of time and space complexity. Time complexity gives us information about how much time is needed to execute the algorithm. It also denotes how complex the problem is to be solved. Space complexity illustrates the space required by the algorithm" [13].

4 Implementation and Results

Initially, the essential step of pre-processing the data is executed. EDA techniques are used here to analyze the data.

After analyzing the data, machine learning algorithms are applied to achieve the best fitting model. The accuracies of the different algorithms are mentioned in Table 1. One of the hyperparameter optimization techniques, Grid Search, is applied to every algorithm to improve accuracy. The improved accuracies can be seen in Table 2. Understanding the time complexity time required for the execution of algorithms is mentioned in Table 3. Best Hyperparameter values are mentioned in Table 4. For evaluation Precision, AUC, Recall, and F-score are the parameters used and precision-recall curves are plotted for the algorithms like Decision Tree, k-nearest neighbor, Random Forest, and all three types of Naïve Bayes. The values of these parameters are shown in Table 5. XGBoost and Artificial Neural Networks are also applied to the dataset.

Results obtained by analyzing the data are as follows (Figs. 3, 4, 5, and 6):

Analyzing data helps us to obtain various relations between the features which are required to gain reasons for customer churning.

Density plots for different features like 'total day charge', 'total eve charge', 'total night charge', and 'total intl charge' are plotted to get the information about distribution. The relation between the customer service calls and churning can be clearly seen in

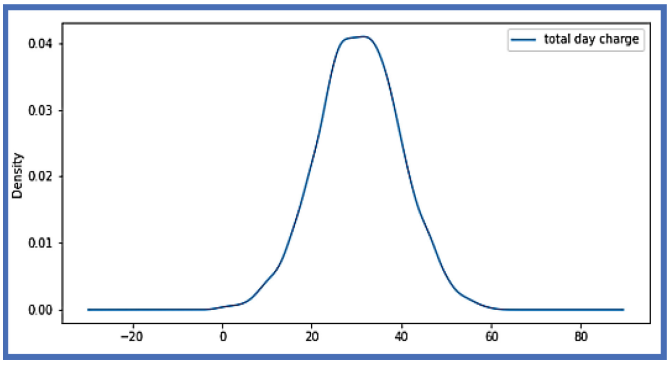

Fig. 3. Density plot for total day charge

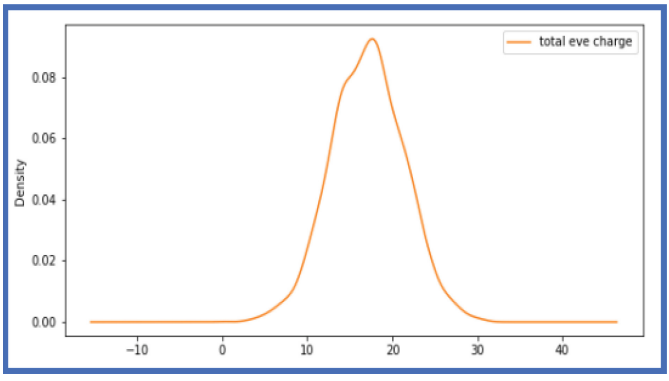

Fig. 4. Density plot for total eve charge

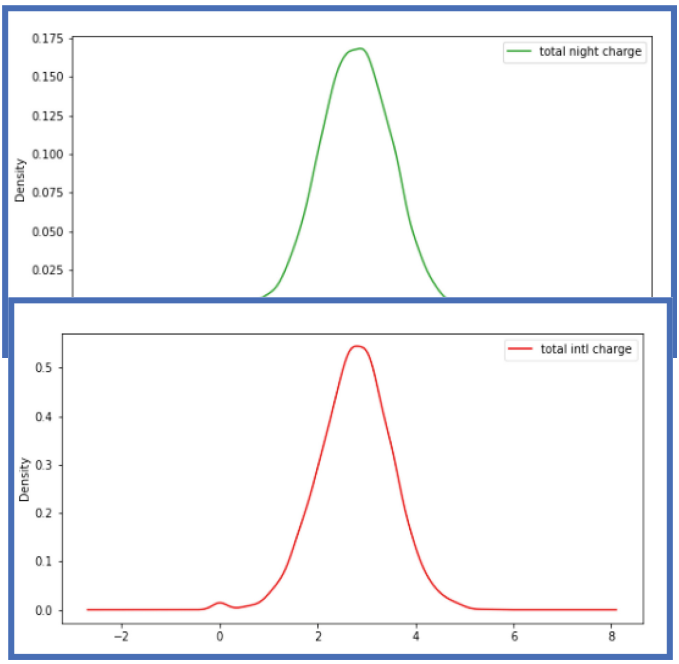

Fig. 6. Density plot for total intl charge

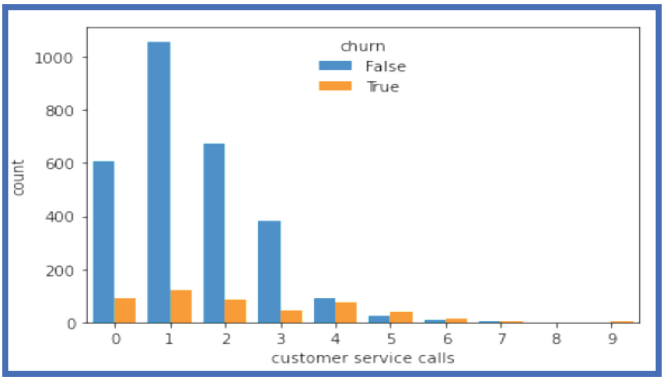

Fig. 7. Relation between Customer Service calls and Customer Churn

Fig. 7. Customer churn can be observed to be increasing after 4 and more Customer service calls. It can be observed from Figs. 8 and 9 that Customers using international plans tend to churn more than Customers with no international plans. This is not the case with the Voice mail plan. In Fig. 10 correlation plot is plotted and it gives us information that four features 'total day charge', 'total eve charge', 'total night charge', 'total intl

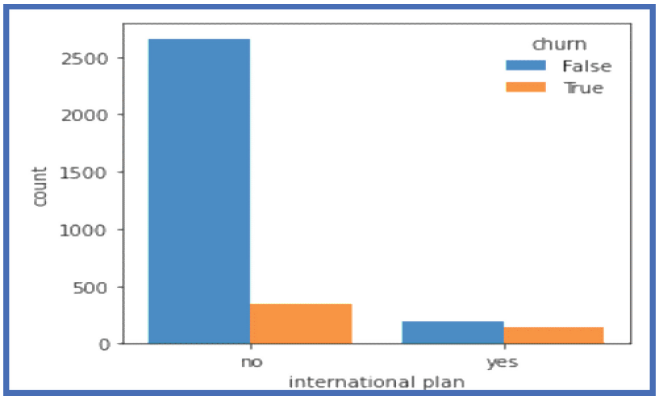

Fig. 8. Relation between International plan and Customer Churn

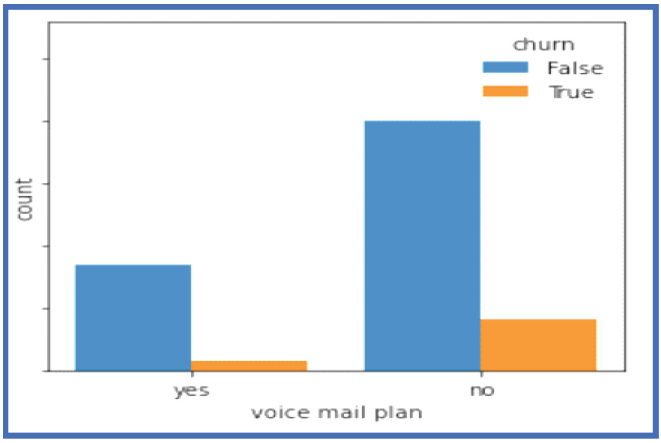

Fig. 9. Relation between Voice mail plan and Customer Churn

charge' are directly dependent on 'total day call', 'total eve calls', 'total night calls', 'total intl calls' respectively.

Precision-recall curves plotted for the applied algorithms are as follows (Figs. 11, 12, 13, 14, 15, and 16):

Artificial Neural Network is implemented on the dataset by using the sigmoid activation function and Adam optimizer. 250 epochs are given for the batch size of 60. Here ANN gives 86.80% of accuracy.

The final accuracies for all the implemented algorithms are mentioned in Table 6.

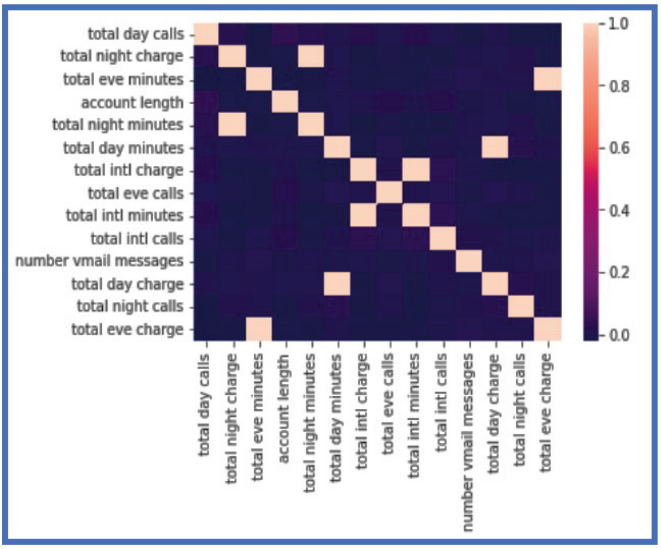

Fig. 10. Heatmap

Table 1. Accuracies of Algorithm without using Grid Search Hyperparameter Optimization

Algorithms	Accuracy (in %)
Decision Tree	93.40
Random Forest	93.85
k-nearest neighbor	88.75
Gaussian Naïve Bayes	85.15
Multinomial Naïve Bayes	63.71
Bernoulli Naïve Bayes	85.60
XGBoost	95.50

Table 2. Accuracies of Algorithm using Grid Search Hyperparameter Optimization

Algorithms	Accuracy (in %)
Decision Tree	93.40
Random Forest	95.20
k-nearest neighbor	88.30
Gaussian Naïve Bayes	88.15
Multinomial Naïve Bayes	63.71
Bernoulli Naïve Bayes	86.20
XGBoost	95.65

Table 3. Time required for execution of Algorithms

Algorithms	Time required for execution (in sec)
Decision Tree	0.010348
Random Forest	1.173684
k-nearest neighbor	0.009588
Gaussian Naïve Bayes	0.007954
Multinomial Naïve Bayes	0.010035
Bernoulli Naïve Bayes	0.010141
XGBoost	0.95737

Table 4. Best Hyperparameter values for Algorithms

Algorithms	Best Hyperparameter values
Decision Tree	random_state = 110 criterion = gini max_depth = 6 min_samples_leaf = 9
Random Forest	n_estimators = 120 random_state = 100 criterion = 'entropy' min_samples_leaf = 7 max_depth = 7
k-nearest neighbor	n_neighbors = 14 p = 2 weights = 'distance' leaf_size = 40 algorithm = 'auto' metric = 'minkowski' metric_params = None n_jobs = None
Gaussian Naïve Bayes	verbose = 1 cv = 10 n_jobs = −1
Multinomial Naïve Bayes	n_jobs = −1 cv = 5 verbose = 5
Bernoulli Naïve Bayes	alpha = 10.0 binarize = 0.0 fit_prior = True class_prior = None
XGBoost	alpha = 1.0

Table 5. Evaluating parameters values.

Algorithms	Precision	Recall	F-score	AUC
Decision Tree	0.74	0.77	0.756	0.80
Random Forest	0.90	0.79	0.80	0.84
k-nearest neighbor	0.67	0.23	0.34	0.37
Gaussian Naïve Bayes	0.62	0.26	0.50	0.47
Multinomial Naïve Bayes	0.38	0.15	0.21	0.29
Bernoulli Naïve Bayes	0.18	0.49	0.26	0.33

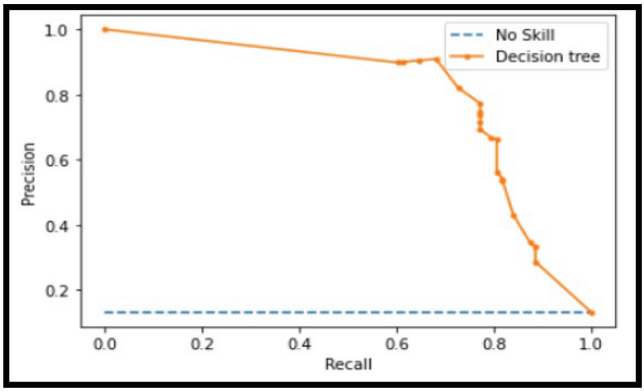

Fig. 11. Precision-recall curve for Decision Tree

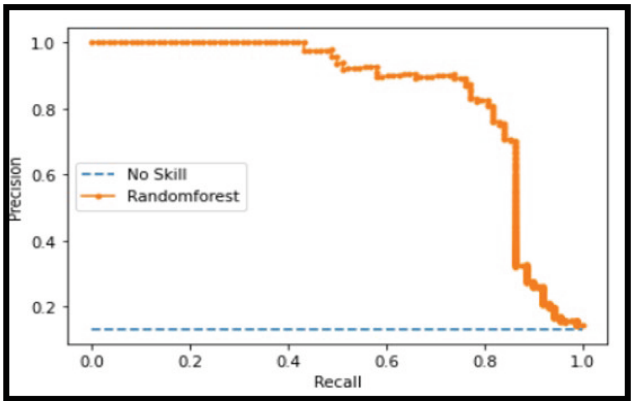

Fig. 12. Precision-recall curve for Random Forest

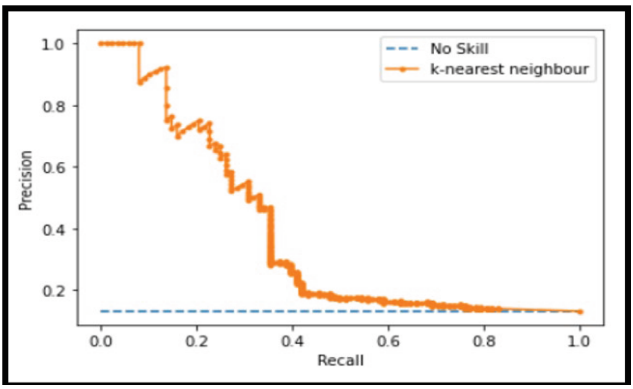

Fig. 13. Precision-recall curve for k-nearest neighbor

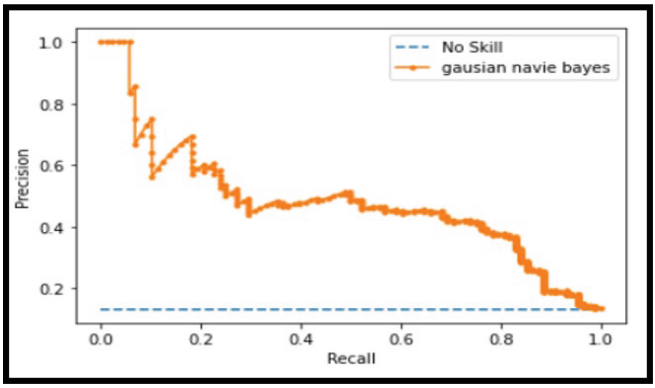

Fig. 14. Precision-recall curve for Gaussian Naïve Bayes

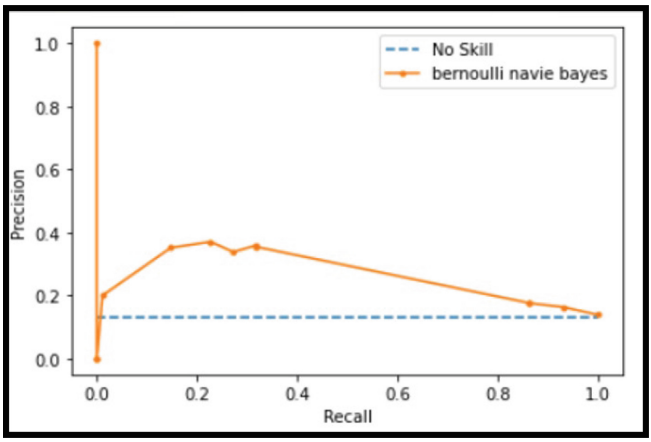

Fig. 15. Precision-recall curve for Bernoulli Naïve Bayes

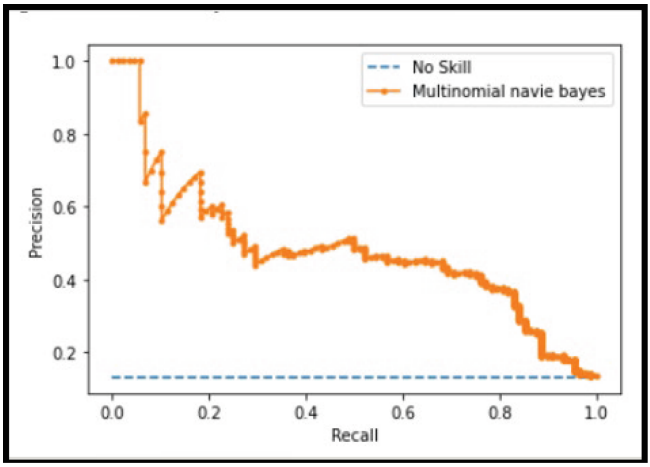

Fig. 16. Precision-recall curve for Multinomial Naïve Bayes

Table 6. Final Accuracy results for the algorithm.

Algorithms	Accuracy (in%)
Decision Tree	93.40
Random Forest	95.20
k-nearest neighbor	88.30
Gaussian Naïve Bayes	88.15
Multinomial Naïve Bayes	63.71
Bernoulli Naïve Bayes	86.20
XGBoost	95.65
Artificial Neural Network	86.80

5 Conclusion

To get the best result for the customer churn prediction in the telecom industry, some commonly known machine learning techniques are imposed on the telecom-based dataset. The use of the Grid Search hyperparameter optimization technique improves the accuracy of the algorithms. The precision-recall curve gives a visual display of the performance of the algorithms. XGBoost and Random Forest are the best performing algorithms with 95.20% and 95.65% accuracies respectively. Comparing these algorithms having the best accuracies with respect to time of execution XGBoost takes less time than Random Forest, which concludes that XGBoost is the best fitting model for this problem.

The future work of the research can involve the implementation of more advanced algorithms and the merging of algorithms to achieve the best outcomes.

References

1. Prof.Andrea Pietracaprina, Prof. Geppino Pucci, "Machine learning techniques for customer churn prediction in banking environments", 2015–16
2. Oskar Sucki, "Predicting the customer churn with machine learning methods - CASE: private insurance customer data", 2019
3. Mr. Anurag Bhatnagar, Dr. Sumit Srivastava "International Conference on Computation, Automation and Knowledge Management (ICCAKM) Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector", 2020
4. Saran Kumar A., Chandrakala D., "International Journal of Computer Applications, A Survey on Customer Churn Prediction using Machine Learning Techniques", 2016
5. Ishpreet Kaur, Jasleen Kaur "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning", 2020
6. Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu "IEEE 5th International Conference on Cloud Computing and Big Data Analytics Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network", 2020
7. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html
8. Deepak Jain, "Data Preprocessing in Data mining, Geeks for geeks", 2021

9. Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr "International Journal of Engineering Research and Applications, A Proposed Churn Prediction Model", 2012
10. Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi, "Customer Churn prediction system: a machine learning approach", 2022
11. Yasser Khan, Shahryar Shafiq, Abid Naeem, Sabir Hussain, Sheeraz Ahmed, Nadeem Safwan, "Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry", 2019
12. Jason Brownlee, "How to use ROC curves and precision Recall curves for classification in python", 2018
13. Qiang Gao, Xinhe Xu "26th Chinese Control and Decision Conference (CCDC), The Analysis and Research on Computational Complexity", 2014