International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

Intelligent Churn Prediction Using In E-Commerce Using Conversational Analytics

[1]S. Iliyaz, [2]P. Mounika, [3]K. Shirisha, [4]T. Saranya, [5]K. Venkatareddy, [6]V. Narendra Maruthi

[1,2,3,4,5,6]Department of Computer Science Engineering (Data Science), GATES Institute of Technology, Gooty, Andhra Pradesh, India
E-mail: [1]shaikiliyaz242003@gmail.com, [2]mounimounika70297@gmail.com, [3]kallushirisha57@gmail.com, [4]tenetisharanya@gmail.com, [5]venkatareddykonatham6@gmail.com, [6]narendranarendra42365@gmail.com

***Abstract* - Businesses must compete fiercely to win over new consumers from suppliers. Since it directly affects a company's revenue, client retention is a hot topic for analysis, and early detection of client churn enables businesses to take proactive measures to keep customers. Consequently, this study aims to advise on the optimum machine-learning strategy for early client churn prediction. The goal is to predict existing customers' responses to keep them. The study has tested algorithms like stochastic gradient booster, random forest, logistics regression, and K-Nearest Neighbours methods. The accuracy of the aforementioned algorithms are 83.9%, 82.6%, 82.9% and 78.1% respectively. We have acquired the most effective results by examining these algorithms and discussing the best among the four from different perspectives.**

Keywords: Stochastic gradient booster, Random forest, K-Nearest neighbours, Logistics regression, Machine Learning.

INTRODUCTION

The customer's concentration on the providers has prompted many new telecom associations to emerge. These new firms usually specialize in providing a specific service or product that the customer cannot find from the incumbent providers. These new firms can provide this service or product at a lower price than the incumbent providers, allowing them to capture a larger market share. This competition between the incumbent providers and the new firms has caused the rates that the associations charge to change.

Churning, in marketing terms, refers to the number of customers who stopped using a particular product. Customer churning is common with any product when there are multiple options for a single problem. Usually, customers will churn when they face difficulties or disappointments in the services rendered by the product. Any organization's primary motive should be satisfying customers and retaining existing customers. Customer churn prediction is the most important issue in adopting an industry's product.

Managing customer churn is one major challenge companies face, especially those offering subscription-based services. Customer churn, also called customer attrition, is the loss of customers caused by a change in taste, lack of proper customer relationship strategy, change of residence, and several other reasons. If businesses can effectively predict customer attrition, they can segment customers who are highly likely to churn and provide better services to them. Hence, a churn prediction model is needed in today's digitized economy to achieve high customer retention and maximize revenue.

Among the methodologies created for anticipating client churn, supervised Machine Learning procedures are widely explored.ML includes algorithms such as Decision Trees, K- Nearest Neighbours, Linear Regression, Naive Bayes, Neural Networks, Support Vector Machines (SVM), and others. Firms have begun to acquire Business Intelligence(BI) applications that anticipate churning clients. This helps in analyzing the reasons for churn and understanding the behaviour of customers who move to competitors, which aids in planning effective customer retention strategies.

In this study, we analyzed some existing algorithms together on the same dataset to determine the most effective algorithm based on accuracy. The study's main objective is to analyze machine learning algorithms for early prediction of customer churn using previously recorded customer feedback. The algorithms used in this study are SGB, Random Forest, KNN and Logistic Regression.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

The rest of the sections in the paper are organized as follows: Section 2 provides the survey on existing research work, Section 3 details the proposed system, dataset description, and modules description, Section 4 discusses the results, and Section 5 concludes the paper.

II. RELATED WORK

This section briefly surveys research works related to customer churn prediction in various industries. Omar Adwan et al. used Multi-layer Perceptron Neural Networks to predict customer churn in the telecommunications industry using real customer data from a Jordanian telecom company [1]. Analyze customer behaviour and predict potential customer loss [7].

Gholamiangonabadi et al. proposed a churn prediction approach using k-medoids clustering where MLPNN and SVM showed better performance [8].

Deepthi Das and Raju Ramakrishna explained that churn prediction helps businesses identify customers likely to leave in industries such as e-commerce, telecommunications, and insurance [9].

Edvaldo and Olawande applied Deep Neural Networks (DNN) for predicting customer relationships in financial sectors [10].

Edwine et al. compared KNN, Random Forest, and SVM algorithms with optimization techniques and found that optimized Random Forest performed better [11].

From the survey, it is evident that Machine Learning and Artificial Intelligence play an important role in customer churn analysis. Logistic Regression provides better interpretation, K-Nearest Neighbours offers accurate predictions, Random Forest improves performance using feature subsets, and Stochastic Gradient Booster helps achieve faster optimization.

III. PROPOSED SYSTEM

This section describes the details of the study, methodologies used and modules.

Farhad Shaikh developed a churn prediction system using classification and grouping techniques with ML and NLP methods to identify churn customers and their reasons [2].

Babu and Ananth explained that data mining techniques help discover hidden patterns in large datasets and classification methods are used to make future predictions [3].

Ismail et al. proposed an MLPNN model for predicting customer churn in a Malaysian telecommunications company and found that neural networks perform better than regression and classification methods [4–5].

Kosgey showed that hybrid models provide more accurate churn predictions compared to individual algorithms [6].

Fatih Kayaalp explained that churn analysis is widely used in subscription-based industries.

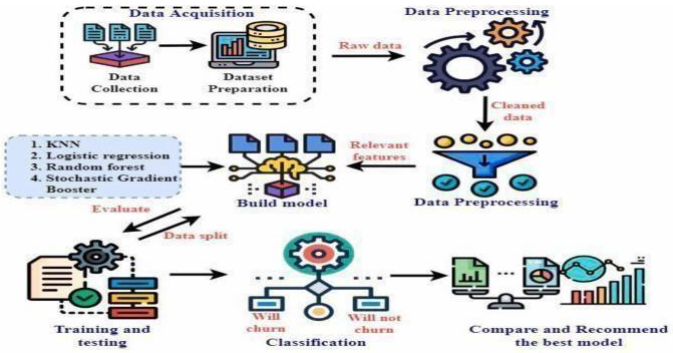


Fig.1. System Layout

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

3.1. Study flow

The goal of machine learning is to find effective solutions for future problems by learning from existing data. Customer churn prediction can be performed using various techniques such as data processing, machine learning algorithms, and hybrid methods. Decision trees are commonly used for churn prediction, but they are not always suitable for complex problems.

In this study, the dataset obtained from Kaggle contains 7044 records and 21 attributes. Initially, data preprocessing and analysis are performed to clean and prepare the dataset. The data is then divided into training and testing sets in the ratio of 70% and 30%. Machine learning algorithms such as Logistic Regression, K-Nearest Neighbours (KNN), Stochastic Gradient Booster, and Random Forest are applied to predict customer churn and evaluate the accuracy of the models.

To implement the churn prediction system, Jupyter libraries are used since they are free and open-source tools. This analysis helps organizations identify customers who are likely to churn and take necessary actions to improve customer retention and increase business profits.

3.2Methodologies Used

The system for customer churn analysis uses the following machine learning algorithms:

1. Stochastic Gradient Booster
2. Random Forest
3. K-Nearest Neighbours
4. Logistic Regression

3.2.1 Stochastic Gradient Booster (SGB)

Stochastic Gradient Boosting is a variation of the boosting algorithm in which a random subset of training data is selected at each iteration to build models instead of using the full dataset. This approach helps improve prediction accuracy and reduce bias.

Purpose

The purpose of the Stochastic Gradient Booster algorithm is to improve prediction accuracy by combining multiple weak models to create a strong predictive model. It is used for both regression and classification problems and works by sequentially correcting the errors of previous models. The algorithm randomly selects subsets of training data in each iteration, which helps reduce bias and improve overall model performance.

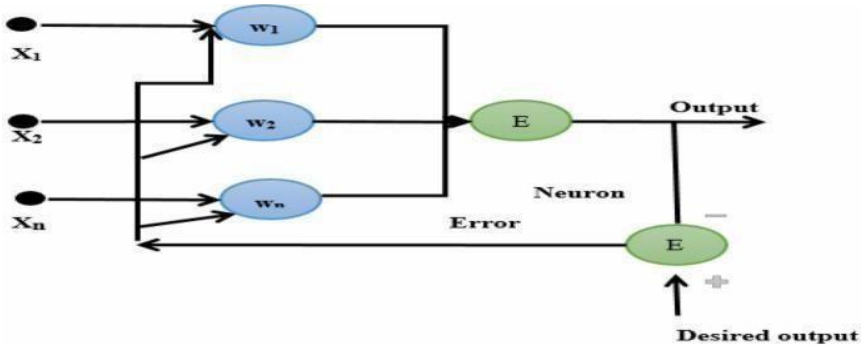


Fig. 2. Stochastic Gradient Booster Architecture[17].

3.2.2 Random Forest Model (RF)

Random Forest is suitable for large datasets and combines multiple decision trees to solve complex problems. The algorithm generates many trees and selects the final output using majority voting.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

Purpose

Random Forest algorithm is to improve prediction accuracy by combining the results of multiple decision trees. It uses majority voting to determine the final output and helps reduce overfitting. Random Forest is suitable for large datasets and can effectively handle complex relationships between variables.

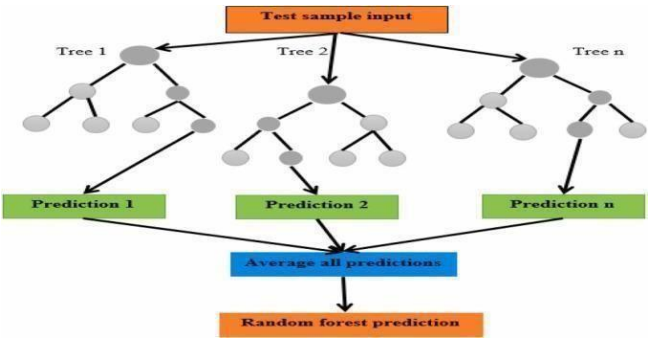


Fig. 3. Random Forest Architecture [13].

3.2.3 K-Nearest Neighbours (KNN)

KNN is a supervised machine learning algorithm used for classification and regression problems. It stores training data and classifies new data points based on the nearest neighbours.

Purpose

K-Nearest Neighbours algorithm is to classify new data points based on the similarity with existing data points. It identifies the nearest neighbours using distance measures and assigns the class based on majority voting. KNN is simple to implement and widely used for classification and regression tasks.

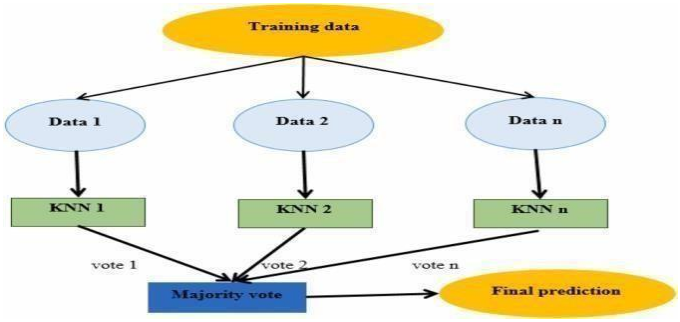


Fig.4. KNN architecture[4].

3.2.4 Logistic Regression Model

Logistic Regression is used to classify whether a customer will churn or will not churn. It estimates the probability of a customer belonging to a specific class.

Purpose

The purpose of Logistic Regression is to predict categorical outcomes, such as whether a customer will churn or not. It estimates the probability that a particular observation belongs to a specific class by producing probability values between 0 and 1. Logistic Regression is widely used in classification problems because it is simple to implement and easy to interpret. The algorithm helps identify the relationship between independent variables and the target variable, allowing researchers to understand which factors influence the prediction. Because of its interpretability and efficiency, it is commonly used in predictive modelling tasks such as customer churn prediction.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

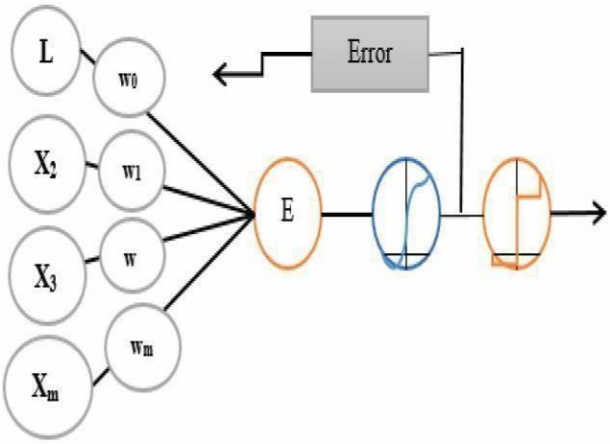


Fig. 5. Logistic Regression Architecture [22].

Dataset Description

The Customer Churn dataset was downloaded from Kaggle. It contains information about a telecommunications company providing phone and internet services to 7044 customers. The dataset shows whether customers left, stayed, or subscribed to services.

The dataset includes information about:

1. Customers who left within the last month (Churn).
2. Services subscribed by customers such as phone, internet, online security, backup, device protection, and streaming services.
3. Customer account information including tenure, contract type, payment method, monthly charges, and total charges.
4. Demographic information such as gender, age range, partners, and dependents.

IV. RESULTS AND DISCUSSION

The results were obtained using Python with Jupyter Notebook from Anaconda. Various Python libraries such as NumPy, Pandas, Matplotlib, and Seaborn were used for data processing, visualization, and analysis. Different machine learning algorithms were applied and their performances were compared.

4.1 Training and Testing Dataset Split

The customer churn dataset containing 7044 records and 21 attributes was divided into training and testing datasets in the ratio of 70:30.

The training dataset was used to train the machine learning models, while the testing dataset was used to evaluate the performance of the model.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No

Fig. 6. The sample rows of the dataset

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

4.2 Dataset Preprocessing

The raw dataset contained attributes with different data types such as objects and categorical values. These values were converted into suitable numerical formats for machine learning algorithms. Label Encoding and One Hot Encoding techniques were applied to convert categorical data into numerical form.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
None
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   int64
 1   gender            7043 non-null   int64
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   int64
 4   Dependents        7043 non-null   int64
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   int64
 7   MultipleLines     7043 non-null   int64
 8   InternetService   7043 non-null   int64
 9   OnlineSecurity    7043 non-null   int64
 10  OnlineBackup      7043 non-null   int64
 11  DeviceProtection  7043 non-null   int64
 12  TechSupport       7043 non-null   int64
 13  StreamingTV       7043 non-null   int64
 14  StreamingMovies   7043 non-null   int64
 15  Contract          7043 non-null   int64
 16  PaperlessBilling  7043 non-null   int64
 17  PaymentMethod     7043 non-null   int64
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   float64
 20  Churn             7043 non-null   int64
dtypes: float64(2), int64(19)
memory usage: 1.1 MB
None
```

Fig. 7a. Dataset description.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
0	7590	1	0	1	0	1	0	3	1	0	...
1	5575	2	0	0	0	34	1	0	1	1	...
2	3668	2	0	0	0	2	1	0	1	1	...
3	7795	2	0	0	0	45	0	3	1	1	...
4	9237	1	0	0	0	2	1	0	2	0	...

5 rows × 21 columns

Fig. 7b. Dataset samples after conversion.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	In
customerID	1.000000	0.006073	-0.002197	-0.026779	-0.012816	0.007805	-0.006252	0.007550	
gender	0.006073	1.000000	-0.001874	-0.001808	0.010517	0.005106	-0.006488	0.001806	
SeniorCitizen	-0.002197	-0.001874	1.000000	0.016479	-0.211185	0.016567	0.008576	0.071049	
Partner	-0.026779	-0.001808	0.016479	1.000000	0.452676	0.379697	0.017706	0.061417	
Dependents	-0.012816	0.010517	-0.211185	0.452676	1.000000	0.159712	-0.001762	-0.011900	
tenure	0.007805	0.005106	0.016567	0.379697	0.159712	1.000000	0.008448	0.176459	
PhoneService	-0.006252	-0.006488	0.008576	0.017706	-0.001762	0.008448	1.000000	-0.844955	
MultipleLines	0.007550	0.001806	0.071049	0.061417	-0.011900	0.176459	-0.844955	1.000000	
InternetService	-0.012230	-0.000863	-0.032310	0.000891	0.044590	-0.030359	0.387436	-0.381534	
OnlineSecurity	-0.000409	-0.000214	-0.208709	0.056157	0.179614	0.085500	0.146522	-0.249780	
OnlineBackup	-0.006665	0.000788	-0.170002	0.059540	0.161106	0.107643	0.164540	-0.244690	
DeviceProtection	-0.008100	0.005642	-0.172926	0.064584	0.157003	0.107656	0.156631	-0.237032	
TechSupport	-0.004863	0.002805	-0.217566	0.047420	0.173036	0.084902	0.145215	-0.247977	
StreamingTV	-0.008606	0.002992	-0.155266	0.054605	0.146505	0.078087	0.179510	-0.246600	
StreamingMovies	-0.012090	0.002082	-0.149000	0.051632	0.136652	0.081169	0.175257	-0.241952	

Fig. 7c. Sample Dataset after Label encoding and One hot encoding.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

4.3 Prediction using KNN Algorithm

The KNN algorithm was applied using Grid Search CV for hyperparameter tuning. The model achieved a training score of 0.7849, testing accuracy of 0.7773, and AUROC value of 0.7817.

KNN AUROC: 0.7814042078747963

	precision	recall	f1-score	support
0	0.81	0.92	0.86	518
1	0.63	0.39	0.48	187
accuracy			0.78	705
macro avg	0.72	0.65	0.67	705
weighted avg	0.76	0.78	0.76	705

Fig.8a.Final Accuracy of KNN

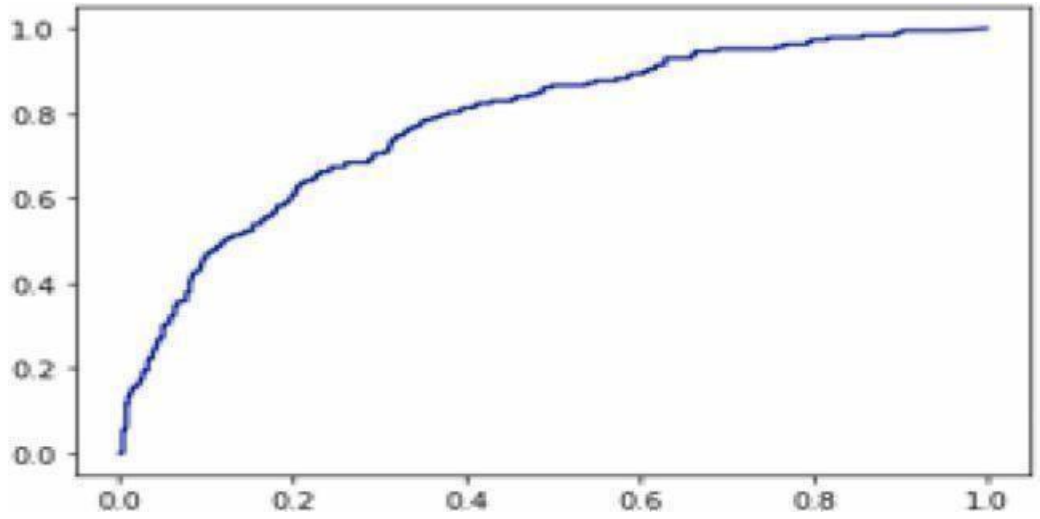


Fig.8b.Predicted graph of KNN

4.4. Prediction of logistic regression algorithm

The prediction results of the LR algorithm by applying Grid Search CV for hyperparameter tuning are displayed in Fig. 9a. and 9b. The best LR Training Score was 0.7975697081209744, test Performance was 0.7829787234042553 and then AUROC was 0.8269877975760328.

```
Tuned LR Parameters: {'C': 0.05}
Best LR Training Score:0.8002514696033005
 LR Test Performance: 0.7843971631205674
 LR AUROC: 0.8176036999566413
```

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

Fig. 9a. Final accuracy of LR

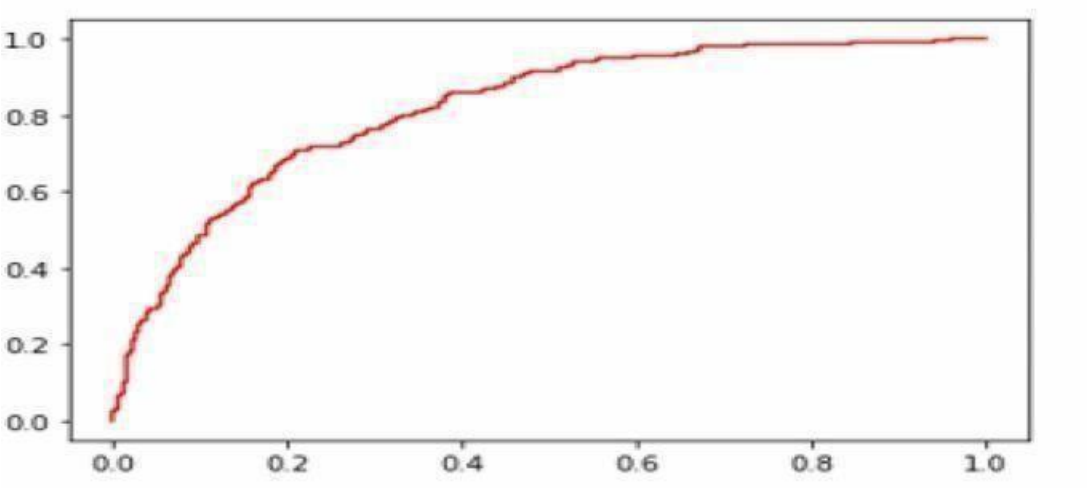


Fig .9b. Predicted graph for LR

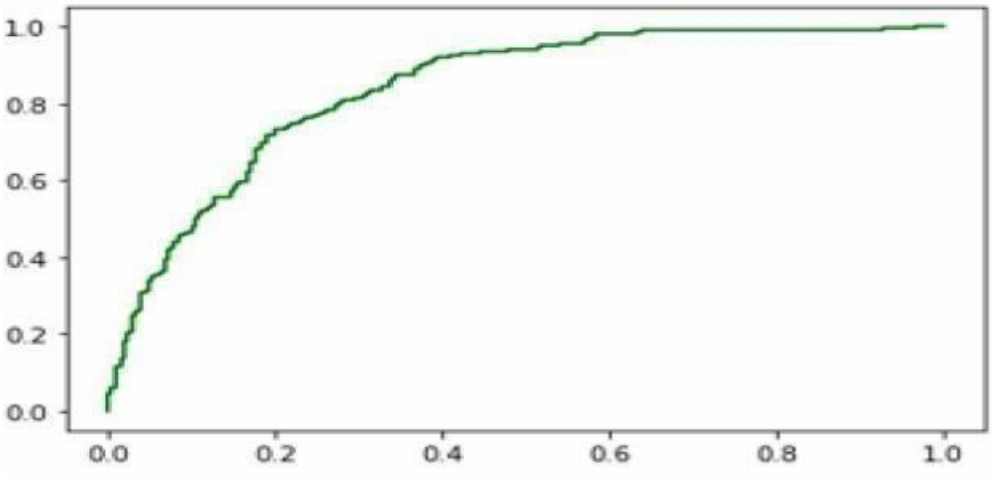
4.5. Prediction of random forest algorithm

The prediction results of the RF algorithm by Randomized Search CV for hyperparameter tuning are displayed in Fig. 10a. and 10b. The best RF Training Score was 0.8038805992445953, test Performance was 0.7872340425531915 and AUROC was 0.829097309685545

RF AUROC: 0.839097309685545

	precision	recall	f1-score	support
0	0.82	0.91	0.86	518
1	0.64	0.46	0.53	187
accuracy			0.79	705
macro avg	0.73	0.68	0.70	705
weighted avg	0.77	0.79	0.78	705

Fig.10a.Final Accuracy of random forest



International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

Fig. 10b. Predicted graph for Random Forest.

4.6 Prediction using Stochastic Gradient Booster Algorithm

The Stochastic Gradient Booster algorithm was applied with Randomized Search CV for hyperparameter tuning. The model achieved a training score of 0.8067, testing accuracy of 0.7914, and AUROC value of 0.8396, which is higher compared to the other models.

SGB AUROC: 0.8396754279107219

	precision	recall	f1-score	support
0	0.82	0.91	0.87	518
1	0.65	0.46	0.54	187
accuracy			0.79	705
macro avg	0.74	0.69	0.70	705
weighted avg	0.78	0.79	0.78	705

Fig. 11a. Final accuracy of stochastic gradient booster.

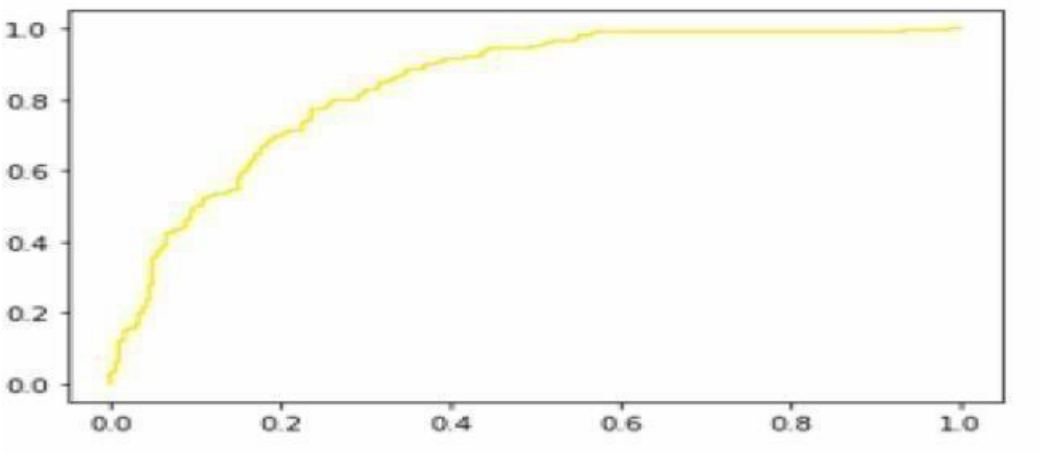


Fig. 11b. Predicted Graph for stochastic gradient booster.

4.7 Overall Comparison of Algorithms

The four algorithms LR, KNN, RF, and SGB were compared based on their accuracy and performance. The results show that SGB achieved the highest accuracy (0.839) compared to RF(0.829), LR(0.826), and KNN (0.781). ROC and

AUC metrics were also used to evaluate the models. Among all the algorithms, SGB provided the best performance for customer churn prediction.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

Fig.13. Comparative analysis of KNN, LR, RF, and SGB.

V. CONCLUSION

The results were analyzed to evaluate the performance of different machine learning algorithms for customer churn prediction. Predicting customer churn is important for organizations because it helps improve customer retention and increase revenue. In this study, four algorithms— Logistic Regression, KNN, Random Forest, and Stochastic Gradient Booster—were analyzed using ROC and AUC evaluation metrics. Among these models, the Stochastic Gradient Booster performed the best with an AUC of 0.84, while KNN showed the lowest performance with an AUC of 0.781.

REFERENCES

[1] Omar Adwan, Hossam Faris, Khalid Jaradat, Osama Harfoushi, Nazeeh Ghatasheh, Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis, Life Sci. J. 11 (3) (2014) 75–81.

[2] Mohammad Ridwan Ismail, Mohd Khalid Awang, M. Nordin A. Rahman, Mokhairi Makhtar, A multilayer perceptron approach for customer churn prediction, International Journal of Multimedia and Ubiquitous Engineering 10 (7) (2015) 213–222.

[3] Anuj Sharma, DrPrabin Kumar Panigrahi, A neural network based approach for predicting customer churn in cellular network services, Int. J. Comput. Appl. 2 (11) (2011) 26–31.

[4] Fatih Kayaalp, Review of customer churn analysis studies in telecommunications industry, Karaelmas Science & Engineering Journal 7 (2) (2017).

[5] Kamorudeen A. Amuda, Adesesan B. Adeyemo, Customers Churn Prediction in Financial Institution Using Artificial Neural Network, 2019, 11346 arXiv preprint arXiv:1912.

[6] Anam Bansal, Churn prediction techniques in telecom industry for customer retention: a survey, J. Eng. Sci. 11 (4) (2020) 871–881.

[7] Vrushabh Jinde, Savyanavar Amit, " customer churn prediction system using machine learning.", International Journal of Advanced Science and Technology 29 (5 (2020) 7957–7964.

[8] Ahmed Iqbal, Shabib Aftab, A classification framework for software defect prediction using multi-filter feature selection technique and MLP, Int. J. Mod.Educ. Comput. Sci. 12 (1) (2020).

[9] Sunday A. Amatare, A.K. Ojo, Predicting customer churn in telecommunication industry using convolutional neural network model, IOSR J. Comput. Eng. 22 (3) (2020)54–59.

[10] M. Feindt, U. Kerzel, The NeuroBayes neural network package, Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrom. Detect. Assoc. Equip. 559 (1)(2006) 190–194.

International Current Journal of Engineering and Science (ICJES)
Volume 5, Issue 4, pp 41-51, April-2026
ISSN (Online): 2583-7257
Article DOI: https://doi.org/10.47001/ICJES/2026.504006

[11] Sun-Chong Wang, Artificial neural network, in: Interdisciplinary Computing in Java Programming, Springer, Boston, MA, 2003,pp. 81–100.

[12] Nabahirwa Edwine, Wenjuan Wang, Wei Song, Denis Ssebuggwawo, Detecting the risk of customer churn in telecom sector: a comparative study, Math. Probl Eng. 2022 (2022). Article ID 8534739, 16 pages.

[13] Navid Khaledian, Farhad Mardukhi, CFMT: a collaborative filtering approach based on the nonnegative matrix factorization technique and trust relationships, J. Ambient Intell. Hum. Comput. (2022) 1–17.

[14] Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi, Customer churn prediction system: a machine learning approach, Computing (2022) 1–24.

****

