

Predicting Telecom Customer Churn: An In-depth Evaluation of Machine Learning Algorithms

Devesh Pawar¹, Yaashie Sabla², Nishil Tayal³ and Dr. Sumita Nainan⁴

¹⁻³Department of Artificial Intelligence, NMIMS, Mumbai, India

Email: deveshpawar1657@gmail.com, {sablanyaashie, ntayal106}@gmail.com

⁴Department of Electronics and Telecommunications Engineering, NMIMS, Mumbai, India

Email: sumitaa.nainan@gmail.com

Abstract— In such an advancing world, telecommunication companies confront the relentless problems of customer churn. Churning can be defined as cancelling the present company's subscription and buying a new service plan from another company. Churn Prediction is the technique that is used to detect these customers before they churn. For example, when JIO was launched, there was a significant change in the number of customers of Airtel, Idea, etc. Many of them churned and shifted to Jio as they offered economical plans. This paper focuses on utilizing machine learning algorithms like logistic regression, Extreme Gradient Boosting (XGBoost), Random Forest, etc to predict the customers who are likely to churn. We initiated the project by thoroughly exploring a Telecom Customer Churn dataset (WA_Fn-UseC_-Telco-Customer-Churn) available on Kaggle. This dataset includes several features like customer characteristics, service usage patterns and contract details. The insights from these features are important for the prediction of customer churn. Furthermore, we emphasize the development of a user-friendly front-end interface using StreamLit. This tool will enable telecom service providers to input customer information and receive real-time churn predictions, improving decision-making processes.

Index Terms— Extreme Gradient Boosting (XGBoost), Logistic Regression, Random Forest, Decision Tree, StreamLit.

I. INTRODUCTION

In today's world, telecom companies play an important role in connecting people from different parts of the world. But the competition in this field is so high that to survive in this industry, the companies must focus highly on their current customers. The effect of this competition gives rise to a significant challenge: Telecom Churn, that is the rate at which the customers change from one telecom service provider to a new telecom service provider, which became a serious concern for the telecom companies. The high churn rate not only raises a question about the services and customer satisfaction but also increases the revenue loss. Acquiring new customers is much more difficult than retaining the old ones. Customers that are price sensitive suffer from high phone rates and expensive internet plans, which leads them to search for less expensive alternatives and switch providers [3].

There are various reasons why customers leave telecom companies like price and billing issues, service quality, competitive offers, customer service experience, coverage and network quality, contract terms, etc. For example, if a customer is using Jio's network and is facing a problem of weak network and low internet speed in a new area where the customer has recently shifted. Due to this problem, they tend to switch to Airtel, which is known for its

strong internet speed in that area.

However, the average churn rate of the telecom industry is between 20% to 50%. The impact of churning could result in reputation loss, revenue loss, increase in marketing costs, etc which could eventually result in reduced profitability of the company. There are mainly 2 types of churns: a) Voluntary Churn and b) Involuntary Churn. Voluntary Churn occurs when the customer decides to switch to a different company by cancelling their current subscription. This could happen due to many reasons such as seeking better prices, and customer service. In contrast, Involuntary Churn occurs due to the actions that the company or the service provider takes without the customer's consent and results in service discontinuation due to factors like contract breaches, billing errors, etc. Customer Churn prediction becomes important because it helps to manage revenues, improve customer satisfaction, address the problems faced by the customers, allocate proper resources, and make data-driven decisions. It is valuable for long-term success in terms of profit and increasing the customer base. Churning has a negative impact not only on a company's overall reputation but also on its brand, which ultimately suffers as a result.

This study focuses on predicting customer churning by using various models like Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Decision Trees and evaluating these models' performance. It is validated with a Telco Churn Dataset imported from Kaggle. The accuracy is measured by using the following parameters: Precision, Recall, Threshold value, ROC/AUC, True Positive Rate, False Positive Rate, and K-Fold cross-validation. Our goal is to use the existing machine learning techniques to predict whether the new customer will churn or not.

II. PROBLEM STATEMENT

In this fierce market, telecom companies face intense competition in terms of customer retention. The companies must alleviate the telecom churn. This can be done using machine learning algorithms which helps in identifying the potential churners, including those who silently switch their providers. This study focuses on building an accurate predictive model using the historical data of the customers' details. This will help the telecom industries to easily identify the churns and retain at-risk customers. Successful implementation will give a valuable tool to combat churn and secure long-term success in this competitive industry.

III. LITERATURE REVIEW

Customer Churn Prediction is an actively researched field. Despite this, an accurate implementation of the Customer Churn Prediction system remains a challenging task.

In this literature review, the researchers examined different methodologies for estimating the probability of customers discontinuing their service based on their behavioural patterns. The primary purpose of this study is to explore various machine-learning techniques to find predictions of churn.

Krishna R, CV Krishnaveni and V Krishna Prasad published "Telecom Churn using Machine Learning" [1]. The dataset contains 266 columns and 99,999 rows. Training and testing sets were divided into 70% and 30% respectively. The models they have used were Logistic Regression, Decision Tree, Random Forest and XG-Boost. Confusion Matrix, Accuracy and AUC curve are being used as the performance measure. The XG-Boost model outperforms the other models with an accuracy of 93% following logistic regression with an accuracy of 89%, Decision tree was 91% and Random Forest was 65%.

T. E Somefun, C. Azubuike, Sanjay Misra and Emmanuel Adetiba published "Predictive Analysis on Customer Churn Using Machine Learning Algorithms" [2]. Pearson's Correlation Coefficient is being used for Feature selection. The models that they used were Random Forest, XG-Boost and Hybrid Model (which was a combination of Random Forest and XG-Boost). Accuracy, True positive, True negative, False Positive, False Negative, Precision, Recall and F1-score are being used for the performance measure. The Hybrid model's accuracy was 95.32% with true positive being 81.6% and true negative being 98.92%. Random forest had an accuracy of 95.32% and XG-Boost was 95.68%. The hybrid models give better performance and better precision than Random Forest and XG-Boost.

Hemlata Jain, Ajay Khunteta, and Sumit Srivastava published "Enhancing Telecom Churn Prediction through Ensemble Techniques, Feature Engineering, and Feature Importance Analysis." [3] The dataset contains 3,333 subscribers and 57 features that have data of customers and their plain details taken from the American Telecom company Cell2Cell. The models being used here are the Random Forest and Gradient Booster Tree models. Performance is being measured by the accuracy and confusion Matrix, both before and after implementing a grid-based hyperparameter tuning approach. Random forest had an accuracy of 95% and the Gradient-boosted trees

model was 97%. The Gradient-boosted tree models give better performance and better precision than Random Forest.

V.Kavitha, S.V. Mohan Kumar, G. Hemanth Kumar, and M. Harish published “Utilizing Machine Learning Algorithms for Customer Churn Prediction in the Telecom Industry”[4]. The proposed system they followed was initially preprocessing, filtering the raw data, and feature selection to gain knowledge about the dataset. The models being used here are the Random Forest, Logistic Regression and XG-Boost models. Performance is measured by the accuracy and confusion Matrix, precision, recall, and F1-score. Random forest had an accuracy of 80%, a logistic Regression model was 79% and XG-Boost was 78%. The Random Forest models give the best performance out of the three chosen models.

Lopamudra Hota and Prasant Kumar published “Customer Churn Prediction in the Telecom Industry: A Machine Learning Approach” [5]. The dataset contains 7043 instances and 21 features. The models they used were Logistic Regression, Random Forest, AdaBoost, Gradient Boost, and Artificial Neural Networks (ANN). Confusion Matrix, precision, F1-score, and recall. Out of all the models, ANN exhibits the highest accuracy of 82.20%.

Vikant Verma, Padmavati Shrivastava, and Dhananjay Sahu published “Churn Prediction using Multiple Machine Learning Algorithms in the Telecom Sector”. The dataset contains 7043 instances and 21 columns that feature attributes such as CustomerID, Gender, and Senior citizen. The models they used were Logistic Regression, Random Forest, AdaBoost, Support Vector Machine (SVM), Naïve Bayes, XGBoost, and Artificial Neural Networks (ANN): confusion Matrix, precision, recall, and F1-measure. The Logistic Model accuracy was 82.08%. Random forest had an accuracy of 81.80%, SVM was 81.44%, AdaBoost was 81.73%, XG-Boost was 79.38%, Naïve Bayes was 75.47% and ANN with the highest accuracy of 85.59%.

IV. DATASET USED

The dataset titled “Telco Customer Churn” (WA_Fn-UseC_-Telco-CustomerChurn.csv) available on Kaggle is used for this study. This dataset includes the data of 7,043 customers. The row contains the customer IDs while the 21 columns include the customer characteristics, service usage patterns, contract information and other details. The primary objective is to predict customer churn, as indicated by the “Churn” column.

The features consist of the following:

- *Customer ID*: This feature contains the unique IDs of individual customers.
- *Gender*: This feature contains whether the customer is Male or Female.
- *Senior Citizen*: This feature contains whether the customer is a Senior Citizen or not.
- *Partner*: This feature contains whether the customer has a partner or not.
- *Dependents*: This feature contains whether the customer has dependents or not.
- *Tenure*: This feature contains the duration of the association with the company represented in months.
- *Phone Service*: This feature contains whether the customer has a phone service or not. (Yes, No)
- *Multiple Lines*: This feature contains whether the customer has multiple lines or not. (Yes, No, No Phone Service)
- *Internet Service*: This feature contains whether the customer has internet service or not. (DSL, Fiber Optic, No)
- *Online Security*: This feature contains whether the customer has online security or not. (Yes, No, No Internet Service)
- *Online Backup*: This feature contains whether the customer has an online backup or not. (Yes, No, No Internet Service)
- *Device Protection*: This feature contains whether the customer has device protection or not. (Yes, No, No Internet Service)
- *Tech Support*: This feature contains whether the customer has Tech Support or not. (Yes, No, No Internet Service)
- *Streaming TV*: This feature contains whether the customer has Streaming TV or not. (Yes, No, No Internet Service)
- *Streaming Movies*: This feature contains whether the customer has Streaming Movies or not. (Yes, No, No Internet Service)
- *Contract*: This feature contains the contract details. (Month-to-month, One Year, Two Year)
- *Paperless Billing*: This feature contains whether the customer has Paperless Billing or not. (Yes, No)
- *Payment Method*: This feature contains the customer’s payment. (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic)).

- *Monthly Charges*: This feature contains the details of amount charged to the customer monthly.
- *Total Charges*: This feature contains the details of total amount charged to the customer.
- *Churn*: This feature contains whether the customer churned or not.

V. PROPOSED SYSTEM

We have proposed a system in Fig 1 for Telecom Customer Churn that involves crucial steps to get final churn predictions.



Fig 1: Proposed System

A. Initial Data Preparation

- *Data Description*: Before starting to work on the dataset, we understand the dataset first. It includes several steps like examining its size, shape, data types, description, etc. to gain better insights.
- *Handling Missing Values*: Then we must find if there are any missing data (NaN or null values) in the dataset to handle them. This can be done either using the function `fillna` which replaces the empty spaces with zero or drop the tables that contain missing values.
- *Standardizing Column Names*: We then must make the column names consistent and user-friendly by replacing space with an underscore, uppercase to lowercase of column names as well as in column data.
- *Converting 'churn' to integer*: This step involves converting the categorical labels in the "churn" into integer values (0 or 1) for the compatibility of the machine learning models.
- *Exploratory Data Analysis*: Performing EDA helps us understand the insights, patterns, and relationships between the data. This involves visualization of the patterns between the different features as well.

B. Data Pre-Processing

Data pre-processing is essential for finding meaningful information from the raw data, which helps various companies with data-driven decision-making and finding actionable insights. This can be done by using various steps such as -

- *Splitting Dataset*: Divide the dataset into training and testing sets which helps in training and evaluating the model metrics.
- *Division of categorical and numerical features*: Dividing features between categorical and numeric is important because both these features require separate pre processing techniques when given as input to a machine learning algorithm.
- *Feature Importance*: Identify the important features required in the prediction of churning by comparing the features' mean with the global mean, finding the mutual information score to find how much a particular feature gives us the information about the target variable which is "churn". According to this study, we found that 'contract', 'online security', 'tech support', 'internet service' and 'online backup' are important features as compare to others.
- *One-Hot Encoding*: We use one-hot encoding to convert categorical data into binary format (0/1).

C. Model Training

In our study, the issue of customer churn prediction is being addressed with the help of these four machine learning algorithms: a) XGBoost b) Random Forest c) Decision Trees d) Logistic Regression. These techniques are used because they are one of the best for binary classification problems.

XGBoost: Extreme Gradient Boosting (XGBoost) is suitable for prediction because of its high speed and the performance it gives. It also gives insights into the important features which in turn helps in understanding the factors that contribute to the customer's churn. It combines different algorithms to create a robust powerful predictive model.

Random Forest: It combines multiple decision trees to predict the customers who will churn. This ensemble learning approach helps in handling non-linear relationships and can even provide insights about the important features which further help in improving the model's accuracy. It is also robust to outliers and noisy data at the same time.

Decision Trees: Decision trees are simple to use as they require minimal effort in data preprocessing and can handle both categorical and numeric features, which makes it suitable for telecom datasets. It can automatically select the important features by prioritizing the most important attributes present at the top of the tree. Decision trees are computationally efficient and give quick predictions.

Logistic Regression: It is a first choice while working with binary classification, such as predicting customer churn (Yes/No) because it offers a clear understanding of the relationship of the features with its target variable. It can work efficiently with smaller datasets and gives good results in terms of accuracy.

D. Model Evaluation

Evaluation is an important part of any machine learning. It involves an assessment of the model's performance on unseen data with the help of various techniques like accuracy, precision, recall, ROC-AUC, etc. We can find the model's stability according to the business objective.

In this research, we evaluated our model based on the threshold value of 0.5 which was used to classify whether the customer is a churner or not, that is the accuracy of the models was calculated based on the threshold value. Using this technique, we got the accuracies as follows:

TABLE I. PREDICTION ACCURACY OF THE PROPOSED MODEL

Sr. No	Model Used	Accuracy
1.	Logistic Regression	80.16%
2.	Random Forest	77.84%
3.	XG-Boost	77.68%
4.	Decision Tree	78.65%

Logistic Regression achieved the highest accuracy of 80.16%. It outperforms all the other three models due to its simplicity and interpretability. Additionally, logistic Regression is less prone to overfitting making it suitable for

smaller datasets. Based on this evaluation, we then applied various performance metrics like precision, recall, ROC-AUC, and confusion matrix on logistic regression to further analyse it.

E. Performance Metrics

Precision

It refers to the model's ability that correctly predict true positives divided by the total number of positive predictions.

$$\text{Precision} = TP / (TP + FP)$$

Where:

- TP represents the number of correctly identified positive cases.
- FP represents the number of incorrectly identified positive cases.

Recall

It is a performance metric used to evaluate the effectiveness of the classification model. It measures the model's ability to identify all relevant instances within the data that is, out of all the actual positive cases, how many did the model correctly predict.

$$\text{Recall} = TP / (TP + FN)$$

- TP represents the number of correctly identified positive cases.
- FN represents the number of actual positive cases

F. K-Fold Cross Validation

It is a technique which is used to assess the performance of the ML model simply by dividing the dataset into K subsets or "Folds" and then training the model on K-1 of these folds, and testing it on the remaining fold in a rotation. This process is repeated K times and the performance metrics are averaged over these iterations to provide a better estimate of the model's performance.

We applied K-fold cross-validation on Logistic regression (as we got the highest accuracy of Logistic regression). After applying we enhanced the accuracy of Logistic regression and achieved 81.30% accuracy.

VI. FRONT-END DEPLOYMENT

For the practical deployment of our Customer Churn prediction model, we realised the importance of connecting the Machine Learning Algorithms with the end-user in the telecom industry.

To achieve this, we deployed the front-end using Stream Lit, using a user-friendly approach to predict whether a particular customer will get churned or not.

The user interface enables efficient decision-making and retention strategies. The user simply enters the customers' data (all the 21 features). The model will process this information and will predict the probability of that customer getting churned. By visualizing the predicted churn probabilities, telecom companies can enhance their customer satisfaction and reduce the churn rate.

By bridging the gap between machine learning and practical application, our front-end deployment helps telecom companies to better manage and prevent customers from leaving.

VII. CONCLUSION

In conclusion, Churn prediction plays an important role in the telecom industry. As the industry continues to grow, the issue of churn prediction increases. Our study into this prediction used Logistic regression, Random Forest, Decision Tree and XG-Boost to achieve high accuracy. While the algorithms predict almost similar performance, the logistic regression emerges to be the top performer with an accuracy of 81.30%.

We look forward to it, the application of this finding to various telecom datasets promises further advancements in churn prediction. The collaboration of big-data analytics and machine-learning techniques holds the potential can help telecom industries shape their future strategies.

Ultimately, it is super important for telecom companies as it helps them keep loyal customers and ensures long-lasting success in the industry.

Compared to the previous studies on telecom churn prediction, our research introduced several key aspects. Firstly, we focused on models like Logistic regression, Random Forest, XGBoost, and Decision Tree, offering a concise yet effective approach. We further applied K-fold cross-validation to enhance the accuracy. Notably, our research

extended beyond conventional accuracy metrics, incorporating the deployment of a user-friendly front-end interface, and emphasizing the practical usability of the developed models.

REFERENCES

- [1] Krishna R, CV Krishnaveni and AV Krishna Prasad, "Telecom Churn Prediction using Machine Learning". World Journal of Advanced Engineering Technology and Sciences (WJAETS), 2022.
- [2] T.E Somefun, C. Azubuike, Sanjay Misra, Emmanuel Adetiba, "Predictive Analysis on Customer Churn Using Machine Learning Algorithms", 2023.
- [3] Hemlata Jain, Ajay Khunteta, Sumit Srivastava, "Telecom Churn Prediction Using an Ensemble Approach with Feature Engineering and Importance", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), 2022
- [4] V. Kavitha, S.V Mohan Kumar, G. Hemanth Kumar, M. Harish, "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT), 2020.
- [5] Lopamudra Hota, Prasant Kumar Dash, "Prediction of Customer Churn in Telecom Industry: A Machine Learning Perspective", Computational Intelligence and Machine Learning, 2021.
- [6] Vikrant Verma, Padmavati Shrivastava, Dhananjay Sahu, "Churn Prediction using Various Machine Learning Algorithms in Telecom Sector", International Journal of Computer Applications (IJCA).
- [7] Shreyas Rajesh Labhsetwar, "Predictive Analysis of Customer Churn in Telecom Industry using Supervised Learning".
- [8] Sarkaft Saleh, Subrata Saha, "Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university", A Springer Nature Journal (SN), 2023.
- [9] Ming Zhao, Qingjun Zeng, Ming Chang, Qian Tong, Jiafu Su, "A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom industry in China", Hindawi, 2021.
- [10] Zeynep Uyar Erdem, Banu Calis Uslu, Seniye Umit Furat, "Customer Churn Prediction Analysis in a Telecommunication company with Machine Learning Algorithms", Journal of Industrial Engineering, 2021.
- [11] R. Sudharsan, E.N Ganesh, "A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy", Connection Science, 2022.
- [12] Hemlata Jain, Ajay Khunteta and Sumit Srivastava, "Telecom churn prediction and used techniques, datasets and performance measures", Springer Link, 2021.
- [13] Nasebah Almufadi, Ali Mustafa Qamar, Rehan Ullah Khan, Mohamed Tahar Ben Othman, "Deep learning based Churn Prediction of Telecom Subscribers", International Journal of Engineering Research and Technology(IJERT), 2019.
- [14] Essam Abou Kassem, Alaa Mostafa Abdelrahman, Shereen Ali Hussein, Fahad Kamal Alsheref, "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User-generated Content", International Journal of Advanced Computer Science and Applications (IJACSA), 2020.
- [15] Samah Wael Fujo, Suresh Subramanian and Moaiad Ahmad Khder, "Customer churn prediction in Telecommunication industry using Deep Learning", Informations Sciences Letters, 2022.
- [16] Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa, "Customer churn prediction in telecom using machine learning in Big Data platform", Journal of Big Data, 2019.
- [17] Lewlisa Saha, Hrudaya Kumar Tripathy, Tarek Gaber, Hatem El-Gohary and El-Sayed M. El-kenawy, "Deep churn prediction method for Telecommunication industry", 2023.
- [18] A Mahalekshmi, G Heren Chellam, "Analysis of customer churn prediction using machine learning and deep learning algorithms", International Journal of Health Sciences, 2022.
- [19] Teoh Jay Shen, Abdul Samad Bin Shibghatullah, "Customer churn prediction model for Telecommunication industry", Journal of Advances in Artificial Life Robotics, 2022.