Churn Prediction in Telecom sector using NLP & ML

Prof. Abhinav S. Thorat
Department of Computer Enginnering
Amrutvahini College of Engineering
Sangamner
Maharashtra, India
abhinavthorat@avcoe.org

Akash R. Raut
Department of Computer Enginnering
Amrutvahini College of Engineering
Sangamner
Maharashtra, India
akashrraut2003@gmail.com

Shivanjali A. Kadam
Department of Computer Enginnering
Amrutvahini College of Engineering
Sangamner
Maharashtra, India
shivanjalikadam89@gmail.com

Mohit M. Sakalkale
Department of Computer Enginnering
Amrutvahini College of Engineering
Sangamner
Maharashtra, India
mohitsakalkale@gmail.com

Ritesh R. Sambhus
Department of Computer Enginnering
Amrutvahini College of Engineering
Sangamner
Maharashtra, India
riteshsambhus@gmail.com

***Abstract*- Our study aims to create a strong churn prediction tool for telecom companies by combining machine learning (ML) especially the powerful XGBoost algorithm and natural language processing (NLP) together. Churn simply means when customers stop using a service or product, which is a big problem for the telecom industry. Customer churn badly affects the revenue of the company and customer loyalty. In today's fast-paced telecom world, accurately guessing when churn might happen is crucial for companies to keep customers and stay profitable. By using NLP and ML, especially XGBoost, our objective is to predict churn accurately, helping companies keep customers happy and their businesses healthy. Our tests show promising results, proving the effectiveness of our tool in reducing churn and building better customer relationships.**

***Keywords*: Machine Learning (ML), Churn prediction, Telecom industry, Natural Language Processing (NLP), XGBoost algorithm.**

SECTION I. INTRODUCTION

In any industry, customers are esteemed as the most valuable asset, as corporate profitability is often directly tied to customer numbers. The telecommunications industry encompasses businesses facilitating global communication through wired or wireless connections, spanning phone, Internet, or both mediums. These companies construct the infrastructure facilitating data transmission worldwide in text, audio, or video formats. This industry's numerous companies comprises internet service providers, cable, satellite, and landline or wireless phone operators. While wired communication is gradually advancing to wireless connections, the number of cellular users expands on a regular basis. Despite the prevalence of wireless communication, customers are divided into post-paid and pre-paid categories based on subscription payment methods, with a range of service options available. The overarching goal of telecommunications firms remains revenue growth and survival in a fiercely competitive landscape.

Customer churn, denoting a significant portion of clients dissatisfied with a telecom company's services, occurs when customers depart for another service provider. This migration of customers results in service provider changes, impacting profit margins for companies. Churn prediction systems are instrumental in mitigating this issue by pre-emptively addressing customer loss.

Various factors contribute to churn, including inadequate network infrastructure, costly services, ineffective strategies, and subpar consumer experiences. Prepaid customers, unbound to a specific company, can terminate services at any time, exacerbating churn rates. Additionally, churn detrimentally affects a company's reputation and brand value. Loyal customers, significant revenue generators, may be inclined to switch providers if dissatisfied, prompting telecom companies to reevaluate policies to prevent revenue loss.

In today's digital age, virtually every business maintains a website to promote, expand, and engage with clientele. Websites serve as platforms for customers to provide feedback and reviews, invaluable for service enhancement. However, traditional churn prediction systems predominantly rely on machine learning and historical customer data analysis. Integrating customer feedback analysis through natural language processing (NLP) can offer deeper insights into customer sentiments.

Customer relationship management (CRM) analysts prioritize forecasting customer churn and identifying underlying causes to retain current clientele. Timely marketing initiatives targeting at-risk customers are vital for churn reduction. Accurate churn prediction techniques are imperative for effective CRM strategies; inaccurate predictions hinder proactive campaign execution. The integration of advanced analytics and machine learning models enhances the precision of these predictions, allowing

for the customization of retention strategies tailored to individual customer profiles.

A. Overview of Machine Learning:

Machine learning (ML) techniques are revolutionizing churn prediction in the telecommunications industry. By harnessing vast datasets comprising customer demographics, usage patterns, and interactions, ML models offer invaluable predictive capabilities. Through sophisticated preprocessing and feature engineering, raw data is transformed into actionable insights, enabling the identification of at-risk customers. ML algorithms, ranging from logistic regression to ensemble methods like random forests and gradient boosting, are meticulously trained and optimized to achieve optimal performance. Evaluation metrics such as accuracy and AUC-ROC provide valuable feedback on model effectiveness. Once deployed, these models facilitate proactive customer retention strategies, ensuring telecom companies can anticipate and address churn effectively. In summary, ML-driven churn prediction empowers telecom providers with the foresight needed to maintain customer satisfaction and business success.

B. Overview of Natural Language Processing:

Natural Language Processing (NLP) techniques are increasingly vital in enhancing churn prediction strategies within the telecommunications sector. By leveraging textual data from customer feedback, reviews, and interactions, NLP enables a deeper understanding of customer sentiment and preferences. Through text preprocessing methods such as tokenization and sentiment analysis, unstructured text data is transformed into structured information, enriching the predictive capabilities of churn prediction models. NLP techniques also facilitate topic modelling, allowing telecom companies to identify recurring themes and issues driving customer dissatisfaction. By integrating NLP insights with machine learning algorithms, telecom providers can develop more nuanced and effective customer retention strategies. In essence, NLP empowers telecom companies to extract actionable insights from textual data, ultimately improving customer satisfaction and reducing churn rates.

C. Objectives:

- To develop a Churn Prediction Model that identifies customers at risk of churning.
- To predict the factors that cause churn.
- To improve customer satisfaction by focusing on areas where improvement is required.

SECTION II. LITERATURE SURVEY

Chen Zhue et al,[1] proposed that Prediction of Telecom Company Customer Churn is addressed using the MIPCA-XGBoost Method to solve the customer churn prediction problem utilizing the MIPCA technique.

Sarkaft Saleh, Subrata [2] proposed that this study aims to understand why customers switch from one service to another and which factors affect churn prediction in the telecom industry, analyzing data and machine learning algorithms. It works on different telecom company datasets, such as IBM telco & Cell2Cell, to analyze Customer retention in telecom companies.

Sylvester Igbo Ele, Uzoma Rita Alo, Henry Friday Nweke, and Ofem Ajah Ofem [3] proposed that this study applies Support Vector Machines (SVM) and selection techniques, alongside logistic regression, for churn prediction in the telecom industry, utilizing standard machine learning algorithms for collection, preprocessing, and feature extraction.

Sharmila K. Wagh, Aishwarya A. Andhale, Kishor S. Wagh, Jayshree R. Pansare, Sarita P. Ambadekar, S.H. Gawande [4] concludes that This study aims to develop an accurate churn prediction model using machine learning algorithms such as K-Nearest Neighbors (KNN) and Decision Tree Classifier, analyzing churn data to develop retention strategies and plans to reduce customer churn.

Mr. Abhinav Sudhir Thorat, Dr. Vijay Ramnath Sonawane [5] proposed that this study highlights that Random Forest, a type of machine learning technique, is used for churn prediction, with the dataset containing features such as customer tenure, the number of voice calls, data used, and the number of customer service calls.

Samah Wael Fujo, Suresh Subramanian, Moaiad Ahmad [6] proposed that This paper focuses on developing a deep learning model for churn prediction in the telecom industry using an Artificial Neural Network (ANN) to predict customer churn and handle imbalanced data.

Denisa MELIAN, Andreea DUMITRACHE, Stelian STANCU, Alexandra [7] proposed that This study finds importance in the Telecommunications Sector in retaining customers, addressing churn from one service to another, and network-related issues, using data mining techniques and developing models.

Glory Sam, Philip Asuquo, Bliss Stephen [8] proposed that This study employs classification and clustering techniques such as XGBoost, Random Forest algorithms, SVM, Decision Trees, and KNN to find accuracy, precision, recall.

V. Kavitha, Hemanth Kumar, S. V Mohan Kumar, M. Harish [9] proposed that This study utilizes tree-based machine learning algorithms using different approaches.

Pan Tang, Wuhan, [10] proposed that This study proposes a customer churn prediction model using KNN, SVM, K-means, and XGBoost algorithm.

SECTION III. PROPOSED SYSTEM

A. System Architecture

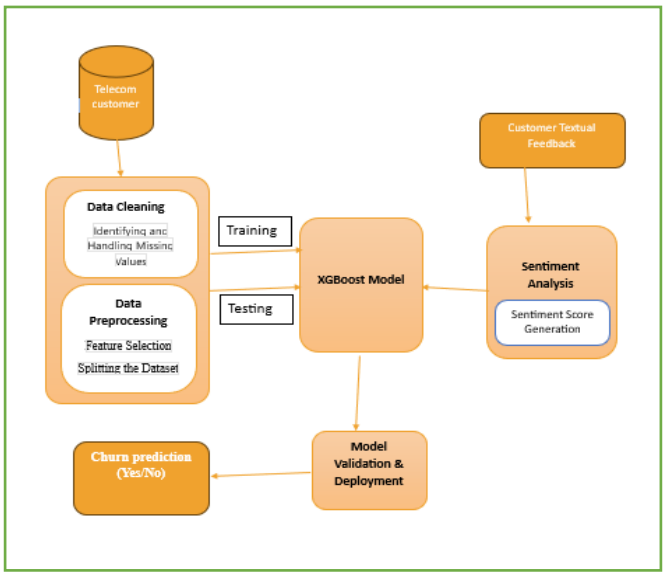

Fig. I: System Architecture

Using a dataset from the telecom sector that includes structured customer data such as customer ID, tenure, marital status, and more, the study first uses this data. The basis for forecasting churn behavior is this dataset. To complement the structured dataset, unstructured data is also gathered in the form of consumer reviews. To make sure the dataset is suitable for predictive modeling and of high quality, data cleaning is started. Only pertinent properties are kept after feature selection and data normalization is used to normalize the numerical feature scale. To effectively train the model, the dataset must be prepared using these methods.

The gathered feedback data is subjected to Natural Language Processing (NLP) using instruments such as the Sentiment Intensity Analyzer (SIA) after data cleansing. Based on consumer feedback, this natural language processing produces sentiment scores that indicate customer satisfaction levels. In order to improve feature sets for predictive modeling, these sentiment scores offer extra insights into customer sentiment that are subsequently combined with structured customer data. The XGBoost technique is used by the study for model creation after the dataset has been produced. Because of its efficiency in managing structured data and ability to support more features from sentiment analysis, XGBoost is selected. Here, the sentiment score that was calculated serves as an extra layer. To precisely assess model performance, the dataset is divided into training and testing sets. As the model is being trained, XGBoost picks up on the underlying patterns in the training data.

When the model training is finished, evaluation metrics including accuracy, precision, recall, and F1 score are used to evaluate the predicted performance. These measurements shed light on how well the model predicts churn behaviour. By employing a thorough methodology that incorporates sentiment analysis and structured data, the research seeks to offer telecom firms practical insights to improve business outcomes and maximize client retention tactics.

B. Dataset

We selected a telecom company dataset, which is available on Kaggle.com, for predicting churn customers because it includes data on both churn and non-churn customers. The dataset is publicly accessible on Kaggle as the Telco Customer Churn dataset. It contains about 21 features and 7043 rows, with the class label "churn" set to either "yes" or "no". The final defined attribute, represented as a numeric value, such as 0 or 1, is the class label. (https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data)

	Features	Feature Description
1	customerID	Customer ID
2	gender	male or female (M,F)
3	SeniorCitizen	Whether customer is Senior Citizen (0,1)
4	Partner	Does customer have partner (Yes, No)
5	Dependants	Customer dependents (Yes, No)
6	tenure	Tenure in months (0 - 72)
7	PhoneService	Customer has a phone service (Yes, No)
8	MultipleLines	Customer uses multiple lines (Yes, No, No phone service)
9	InternetService	Internet service provider (DSL, Fiber optic, No)
10	OnlineSecurity	Customer has online security or not (Yes, No, No internet service)
11	OnlineBackup	Customer has online backup or not (Yes, No, No

		internet service)
12	DeviceProtection	Customer has device protection or not (Yes, No, No internet service)
13	TechSupport	Tech support for customer (Yes, No, No internet service)
14	StreamingTV	streaming TV (Yes, No, No internet service)
15	StreamingMovies	Customer has streaming movies or not (Yes, No, No internet service)
16	Contract	The contract term of the customer (Month-to-month, One year, Two year)
17	PaperlessBilling	Customer has paperless billing or not (Yes, No)
18	PaymentMethod	Payment method (Mailed check, Electronic check, Credit card ,Bank transfer (automatic))
19	MonthlyCharges	Monthly amount charged (18.3 – 119)
20	TotalCharges	Total amount charged (18.8 - 8680)
21	Churn	(Yes, No)

Table I. Dataset Description Table

C. *Data Preprocessing*

Data preparation is crucial to ensure that the data is clean, consistent, and suitable for machine learning models. It involves a series of systematic steps, including addressing missing values, standardizing data, encoding categorical variables, and eliminating noise or outliers. The dataset to be used consists of 7043 rows and 21 columns. The first step of preprocessing involves searching for null values within the dataset and filling them. Features like 'MonthlyCharges' and 'TotalCharges' are converted to numeric values to handle errors and fill the missing values.

The dataset consists of various categorical variables such as 'Gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', etc., which are to be converted into numerical format using one-hot encoding for further processing and model training. To obtain dummy values for the categorical variables, a new column for each unique value in the categorical columns is created. The first dummy column for each category is dropped to avoid redundancy and multicollinearity. The target column 'Churn' contains categorical data which is to be converted into binary numerical labels 0 and 1.

Feature Selection:

Building predictive models requires carefully selecting a subset of important features. Features such as 'CustomerID' that have no impact on the model's prediction are filtered out. The target column 'Churn' is dropped for the purpose of model training.

D. *Data Analysis*

The key to predicting customer churn lies in uncovering hidden patterns within their data. This study adopts a unique approach by integrating two methodologies: ML and NLP.

In our data analysis, we delve into various aspects of customer behavior to pinpoint potential indicators of churn. Initially, we investigate the relationship between a customer's tenure and their propensity to churn. Subsequently, we explore how different usage patterns, such as phone line or internet service subscriptions, impact churn rates. Additionally, we examine the influence of billing preferences, such as paperless billing and payment methods, on churn behavior. Moreover, we delve into demographic factors like age, gender, and geographical location to understand their contribution to churn. Lastly, we analyze the correlation between monthly spending habits and customer churn. Through the use of charts, graphs, and statistical methods, our objective is to uncover any discernible connections between these factors and customer attrition, thereby identifying the most significant predictors of churn.

To gain insights into these relationships, we employ a correlation matrix to visualize the strength of connections between various customer features. This visualization aids in identifying initial patterns, such as whether senior citizens with high spending habits (MonthlyCharges) tend to exhibit greater retention. Additionally, we delve deeper into the analysis by exploring how factors like contract duration and customer service interactions impact churn rates. This comprehensive examination allows us to uncover nuanced insights and refine our churn prediction model further.

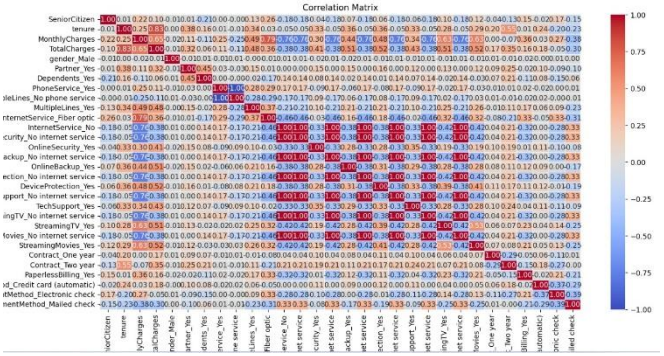

Fig 2. Correlation Matrics

SECTION IV. MODEL BUILDING

A. ML Model:

1. Random Forest:

The Random Forest machine learning technique generates many decision trees and aggregates their forecasts to arrive at a final prediction. Decisions are made by each decision tree in the forest using a random subset of characteristics once it has been trained on a portion of the data. Random Forest is able to manage complicated interactions in the data and produce reliable forecasts by combining the predictions of several trees. It has a reputation for being extremely accurate and capable of handling big datasets with plenty of attributes.

2. Logistic Regression

Logistic Regression is a statistical tool mainly used for tasks where we need to classify things into two groups For example, the model may examine past customer data, including payment details, use trends, demographics, and customer service exchanges. It works by looking at the information we have and calculating the chances of something belonging to one group or the other. These chances are turned into probabilities, which tell us how likely it is for each data point to belong to each group. The tool then figures out which factors influence these probabilities the most and adjusts them to make the predictions as accurate as possible. Once it's trained, it can take new information and predict which group each new data point belongs to, based on the probabilities it calculated

3. XGBoost

The gradient boosting algorithm's advanced version, known as XGBoost, has become popular for its excellent accuracy and performance in predictive modeling applications. It builds a sequence of decision trees one after the other, each one learning from the errors of the one preceding it. XGBoost decreases the discrepancy between the expected and actual results by employing a gradient boosting strategy to enhance the learning process.

Objective Function: The objective function in XGBoost combines the loss function with regularization terms. It's formulated as:

$$Objective = \sum(Loss(y\_i, \hat{y}\_i)) + \sum(\Omega(f\_k))$$

Where:

- **Loss(y_i, ŷ_i):** Loss function measuring the difference between predicted (ŷ_i) and actual (y_i) values.
- **Ω(f_k):**regularization word for every tree in the group.

Because XGBoost can effectively handle huge datasets with high dimensionality, it is thought to be the best model for churn prediction. Because it can automatically manage feature engineering, feature selection, and missing values, it is appropriate for complicated real-world datasets that are frequently used in churn prediction tasks. Regularization strategies are also used by XGBoost to avoid overfitting and ensure the model's applicability to fresh data. Because of its speed and scalability, it can handle large volumes of data rapidly, which makes it the perfect option for real-time applications like telecom churn prediction. Furthermore, XGBoost offers comprehensible outcomes, enabling analysts to understand the elements causing churn and extract useful information for maintaining clients tactics.

B. NLP Model:

Natural Language Processing (NLP) can be used as a additional layer as it does analysis on customer feedback for churn prediction. NLP can be used to determine the sentiment of customer feedback. Positive sentiments might indicate satisfaction, while negative sentiments could signal dissatisfaction or intention to churn. The sentiment analysis is done with the use of Sentiment Intensity Analyzer (SIA). The Sentiment Intensity Analyzer (SIA) is a tool within the TextBlob library in python. It's specifically designed to analyze the sentiment of text data. Unlike other sentiment analysis tools that simply categorize text as positive, negative, or neutral, SIA provides a more nuanced analysis by calculating a sentiment score. These scores reflect the positive or negative connotation of each word.

SIA doesn't use pre-trained data in the same way as some machine learning models. SIA relies on a pre-built lexicon containing words with associated sentiment scores. The sentiment lexicon is itself pre-trained. This lexicon is a large list of words with pre-assigned sentiment scores.

SIA generates a compound score that reflects the general emotions of the piece of text. The "compound" score in the lexicon, represents the overall sentiment polarity (positive, negative, or neutral). This score normally varies from -1 (very negative) to +1 (highly positive), with zero representing neutrality. SIA is relatively easy to use. Hence, Sentiment analysis with SIA is fast and efficient.

SECTION V. RESULT & ANALYSIS

We conducted an analysis and comparative study of the churn prediction problem, where customers are churning from one service to another. Through this, we aimed to identify which parameters affect customer churn and to find solutions to mitigate this problem using machine learning techniques, algorithms, and Natural Language Processing (NLP). We utilized various algorithms for this comparative study, such as XGBoost (Extreme Gradient Boosting), Random Forest, Decision Tree, and Logistic Regression. With the help of these algorithms, we determined the accuracy of each: XGBoost Accuracy (0.8292122072391767), Random Forest Accuracy (0.8077288857345636), Decision Tree Accuracy (0.7204329311568488), and Logistic Regression Accuracy (0.7926969481902059). According to our study and results, the XGBoost algorithm exhibited the highest accuracy; therefore, we used the XGBoost algorithm in our model. After obtaining these prediction results, we applied NLP for sentimental analysis to generate a sentiment score or customer feedback. Thus, our model is beneficial for solving the customer churn problem in telecommunication companies.

A. Confusion matrix

To evaluate a classification model's performance on a set of test data where the real values are known, a confusion matrix is a useful tool that we employ in our research. It presents the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions, giving a clear visual representation of the model's performance.

1.True Positive (TP): These are examples in which the model properly classified them as positive, meaning that positive cases were correctly detected by the model. (TP is 925).

2. True Negative (TN): This category includes situations that the model accurately identified as negative, proving its capacity to do so. (TN =195)

3. False Positive (FP): These cases indicate Type I errors since the model misclassified them as positive. Stated differently, the model projected a favorable outcome, while the actual event was unfavorable. (FP = 111)

4. False Negative (FN): These cases show Type II errors because the model misclassified them as negative. When the real result was positive, the model projected a negative outcome. (FN = 178)

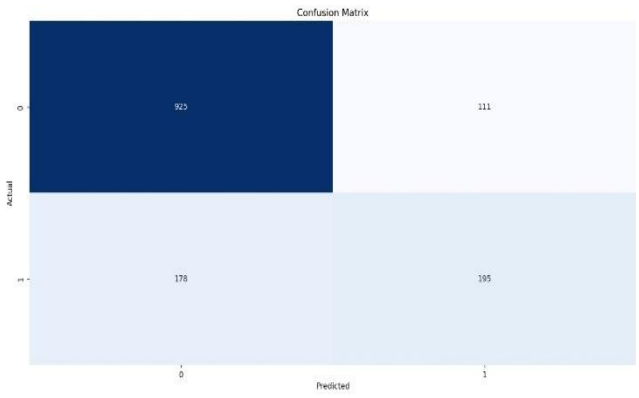

Fig. 3. Confusion Matrix

B. Performance Metrics:

To assess the effectiveness of our machine learning model, we utilized several standard evaluation metrics tailored to the specific classification problem at hand.

1.Precision:

Precision quantifies the accuracy of positive predictions, representing the ratio of correctly predicted positive observations to the total predicted positives. It is especially useful when there is a large cost associated with false positives.

Formula: Precision = TP / (TP + FP)

2.Accuracy:

By showing the ratio of accurately predicted data to total observations, accuracy indicates how accurate the model is overall. It is suitable for balanced class distributions but may be misleading in the presence of class imbalance.

Formula: Accuracy = (TP + TN) / Total Observations

3. F1 Score:

As the harmonic mean of recall and accuracy, the F1 score finds a balance between the two. It offers a comprehensive assessment of model performance.

Formula: F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

4.Recall:

Recall focuses on identifying type-II errors (FN) and does not directly measure type-I errors (false positives). It signifies the ratio of correctly predicted positive observations to the actual positives in the dataset.

Formula: Recall = TP / (TP + FN)

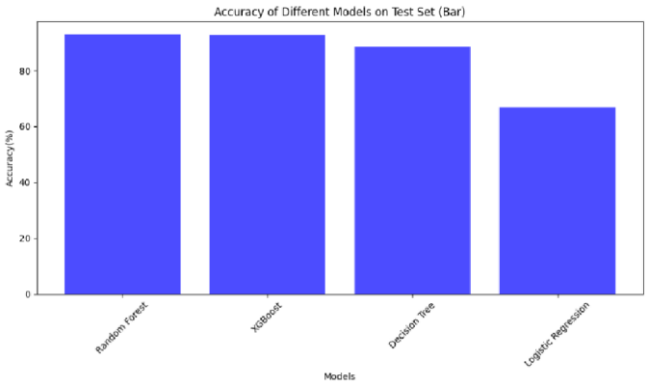

Fig. 4. Accuracy Bar-Graph

SECTION VI. CONCLUSION

In the present competitive telecom market, customers frequently switch between service providers, posing a retention challenge. To address this, we employ machine learning algorithms and Natural Language Processing (NLP). Our project utilizes the XGBoost algorithm for model implementation and conducts a comparative study with Random Forest. Our proposed prediction model combines XGBoost and Random Forest to forecast customer churn in telecommunications companies. By employing Random Forest, XGBoost, and Logistic Regression, we achieve higher accuracy compared to other algorithms. Leveraging customer service plan data, our model accurately predicts potential churners, enabling telecom companies to offer targeted incentives to retain customers. Results demonstrate the effectiveness of our churn model, with XGBoost achieving superior accuracy. Additionally, NLP aids in

generating sentiment scores from customer feedback, enhancing model performance.

AKNOWLEDGEMENT

We would like to extend our heartfelt thanks to our guide, Prof. Abhinav Thorat sir, for his invaluable guidance, unwavering support, and mentorship throughout this project. His expertise, encouragement, and constructive feedback have been instrumental in shaping our research journey and contributing to the success of this endeavour.

REFERENCES

[1] Chen Zhue(2023);" Prediction of Telecom Customer Churn Based on MIPCA-XGBoost Method". ISSN: 2832-6024 | Vol. 3, No. 1, 2023.

[2] Sarkaft Saleh ,Subrata Saha(2023);" Customer retention and churn prediction in the telecommunication Industry" 27 February 2023 / Accepted: 16 May 2023.

[3] Sylvester Igbo Ele , Uzoma Rita Alo 2,*, Henry Friday Nweke 3,and Ofem Ajah Ofem 1(2023);" Regression Based Machine Learning Framework for Customer Churn Prediction in Telecommunication Industry". Journal of Advances in Information Technology, Vol. 14, No. 5, 2023.

[4] Sharmila K. Wagh , Aishwarya A. Andhale , Kishor S. Wagh ,Jayshree R. Pansare , Sarita P. Ambadekar ,S.H. Gawande(2023);" Customer Churn Prediction in Telecom Sector using Machine Learning Techniques".S2666-7207(23)00144-3 RIOC 100342.11 Novembber 2023.

[5] Mr.Abhinav Sudhir Thorat,Dr. Vijay Ramnath Sonawane(2022);" A Random Forest Churn Prediction Model: An Investigation of Machine Learning Techniques for Churn Prediction and Factor Identification in the Telecommunications Industry" Digital Object Identifier 10.1109/ACCESS.2019.2914999.

[6] Samah Wael Fujo,Suresh Subramanian, Moaiad Ahmad Khder(2022);" Customer Churn Prediction in Telecommunication Industry Using Deep Learning"Inf.Sci.Lett.11.No.1185 -198(2022).

[7] Denisa MELIAN 1,Andreea DUMITRACHE 2,Stelian STANCU 3,Alexandra NASTU 4(2022);" Customer Churn Prediction in Telecommunication Industry. A Data Analysis Techniques Approach" ISSN: 2068-0236 | e-ISSN: 2069-9387Covered in: Web of Science (WOS); EBSCO; ERIH+; Google Scholar; Index Copernicus; Ideas RePeC; Econpapers; Socionet; CEEOL; Ulrich ProQuest; Cabell, Journalseek; Scipio; Philpapers; SHERPA/RoMEO repositories; KVK; WorldCat; CrossRef; CrossCheck.

[8] Glory Sam a*, Philip Asuquo a∗,Bliss Stephen (2021);" Customer Churn Prediction using Machine Learning Models" Journal of Engineering Research and Reports · February 2024
DOI: 10.9734/jerr/2024/v26i21081.

[9] V. Kavitha, Hemanth Kumar, S. V Mohan Kumar, M. Harish,(2020);" Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT)http://www.ijert.org ISSN: 2278-0181 IJERTV9IS050022(This work is licensed under a Creative Commons Attribution 4.0 International License.).

[10] pan Tang Wuhan,Hubei(2020);"Telecom Customer Churn Prediction Model Combing K-means and XGBoost Algorithm"Wuhan University of Technology School of Management Wuhan,Hubei,China.