A MINI PROJECT REPORT ON

"Enhancing Customer Retention: Analysing Churn for Improved Business Performance"

SUBMITTED TOWARDS THE FULFILMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (B. Tech.)

Academic Year: 2023-24

By:

Bhumika Patil BTCOD282

Disha Magar BTCOD302

Maithili Sude BTCOD312

Under The Guidance of

Prof. Sushma Vispute

DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

SECTOR 26, NIGDI, PRADHIKARAN

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that, the project entitled

"Enhancing Customer Retention: Analyse Churn for Improved Business Performance"

is successfully carried out as a mini project successfully submitted by following students of "PCET's Pimpri Chinchwad College of Engineering, Nigdi, Pune-44**".**

Under the guidance of Prof. Sushma Vispute

In the fulfillment of the requirements for the B. Tech. (Computer Engineering)

Bhumika Patil BTCOD282

Disha Magar BTCOD302

Maithili Sude BTCOD312

Prof. Sushma Vispute
Project Guide

Department of Computer

Prof. Dr. K.Rajeswari
Head,
Department of Computer

Engineering

Abstract

Customer churn is a significant challenge for businesses in today's competitive landscape. It is estimated that acquiring a new customer can cost five times more than retaining an existing one. Therefore, businesses must focus on identifying and addressing the root causes of churn in order to minimize its impact.Machine learning (ML) has emerged as a powerful tool for predicting customer churn. By analyzing historical data, ML algorithms can identify patterns and trends that are associated with customer churn. This information can then be used to develop targeted retention strategies that are designed to keep customers from leaving.This project aims to develop a robust churn prediction model that can be used to identify potential churners. The model will be developed using a variety of machine learning algorithms, including logistic regression, decision trees, and random forests. The model will be evaluated on its ability to accurately predict customer churn for a defined future period.

Once the model is developed, it will be used to extract actionable insights, identifying high-risk customers and the factors contributing to their churn likelihood. This information will then be used to implement targeted retention strategies for these customers. The overall goal of this project is to reduce customer churn and increase customer loyalty, leading to improved business performance.

Index

List of Figures & Tables

Introduction

a. Problem Statement:

In today's competitive business landscape, retaining customers is crucial for sustained growth and profitability. Customer churn, the loss of customers to competitors or disengagement, poses a significant challenge. To address this issue, we embark on the 'Enhancing Customer Retention' project, focusing on predictive analytics and machine learning. The objective is to develop a robust churn prediction model to foresee potential customer attrition based on historical data and customer behaviour patterns. By proactively identifying at-risk customers, we aim to implement targeted retention strategies, such as personalized offers or improved customer service, with the ultimate goal of improving overall business performance, increasing customer satisfaction, and optimizing revenue streams. Through this initiative, we seek to harness the power of data-driven insights to fortify our customer relationships and drive sustainable growth in a competitive market.

b. Project Objectives:

- Evaluate the impact of the implemented retention strategies on reducing churn and increasing customer loyalty.
- Choose appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests) for churn prediction.
- Visualize customer behaviours and trends to gain actionable insights into potential churn drivers.
- Understand the current customer churn rate and its impact on business performance.
- Identify key features that might influence customer churn, such as customer behaviour, transaction history, customer support interactions, contracts etc.
- Utilize the trained models to predict customer churn for a defined future period.
- Extract actionable insights from the predictions, identifying high-risk customers and the factors contributing to their churn likelihood.

c. Motivation:

In today's competitive business landscape, customer retention is crucial for sustained growth and profitability. Customer churn, the loss of customers to competitors or disengagement, poses a significant challenge. It is estimated that acquiring a new customer can cost five times more than retaining an existing one. Therefore, businesses must focus on identifying and addressing the root causes of churn in order to minimize its impact.Machine learning (ML) has emerged as a powerful tool for predicting customer churn. By analyzing historical data, ML algorithms can identify patterns and trends that are associated with customer churn. This information can then be used to develop targeted retention strategies that are designed to keep customers from leaving. The motivation for this project is to develop a robust churn prediction model that can be used to identify potential churners. By proactively identifying at-risk customers, businesses can implement targeted retention strategies that are more likely to be successful. This can lead to reduced churn rates, increased customer loyalty, and improved business performance.

d. Literature Survey:

"A Prediction Model of Customer Churn Considering Customer Value: An Empirical Research of Telecom Industry in China"(Hindawi Journal) explores the phenomenon of customer churn in the telecommunications industry and proposes a prediction model to identify potential churned customers. The study focuses on the telecom industry in China and employs a logistic regression algorithm to analyze high-value customer operation data. The paper aims to understand the factors influencing customer churn, analyze the causes, and develop targeted win-back strategies. The research covers dimensions such as price, product, customer, business, and service factors, and establishes hypotheses regarding their impact on customer churn. The findings emphasize the significance of accurately identifying churn causes for formulating effective win-back strategies and underscore the importance of managing customer value. The study provides valuable insights for telecom operators to address customer churn and strengthen their competitiveness.

"Customer Churn Analysis in Telecom Industry"(IEEE Xplore) The authors propose a framework for predicting customer churn and apply it using data mining techniques, specifically Decision Trees and Logistic Regression, with the aim of identifying potential churners. The study emphasizes the importance of customer relationship management (CRM) and data mining in improving customer retention and acquisition. The paper presents the process of data acquisition, preparation, preprocessing, and extraction, culminating in the application of classification algorithms. The authors evaluate the performance of their proposed framework using various datasets, showcasing the accuracy achieved by the Decision Tree method. The research highlights the significance of accurate attribute grouping and threshold setting in achieving better prediction outcomes. Overall, the paper

underscores the significance of churn prediction in the telecom industry and the potential benefits of employing data mining techniques for customer retention.

"Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms"(Research Gate): The authors review a range of machine learning algorithms employed in various studies, including decision trees, logistic regression, neural networks, and support vector machines, highlighting their respective strengths and weaknesses in the context of churn prediction. In the Initial step, data preprocessing is performed in which filtering data and converting data into a similar form, and then we make feature selection. In the further step prediction and classification is done using the algorithms like Random Forest, XGBoost, Logistic Regression(LR). Training and testing the model with the data set, we observe the behavior of the customer and analyze them. In the final step, we do analysis based on the results obtained and predict the customer churn.

"Customer churn prediction in telecom using machine learning in big data platform"(Springer Open): Model utilizes machine learning techniques and operates on a big data platform. To assess the effectiveness of the churn prediction model, the authors employ the Area Under Curve (AUC) as the standard performance measure. The model achieves an AUC value of 93.3%, indicating its robust predictive capabilities. Social Network Analysis (SNA) features are extracted and integrated, leading to a substantial enhancement in model performance. The AUC score improves from 84% to 93.3% when SNA features are integrated. The research explores the performance of four different algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). Among these, XGBOOST demonstrates the best results and is employed for classification in the churn prediction model.

"Customer churn prediction using machine learning"(DiVA Portal): This research paper presents a machine learning-based churn prediction model tailored for a subscription-based service provider operating within the financial administration sector, with a focus on the business-to-business (B2B) context. The research compares three different machine learning approaches for churn prediction: XGBoost and Random Forest (ensemble learners), and Naïve Bayes (a single base learner). The goal is to determine which algorithm performs best in this specific application. Key evaluation metrics include accuracy, precision, recall, and F1-score. Given the imbalanced nature of the dataset, the study explores three different sampling methods: SMOTE, SMOTEENN, and RandomUnderSampler. These methods are employed to balance the dataset and improve the performance of the churn prediction model. Additionally, the findings indicate that ensemble learners (XGBoost and Random Forest) outperform a single base learner (Naïve Bayes).

Project Design

a. H/W , S/W , resources, requirements & their detail explanation

Churn prediction in Power BI involves both hardware (H/W) and software (S/W) requirements, as well as various resources. Here's a detailed explanation of each aspect:

1. Hardware (H/W) Requirements:

- Processor: A multi-core processor, preferably 64-bit architecture for handling large datasets efficiently.
- RAM: At least 8 GB of RAM, but more is recommended for handling complex calculations and data processing.
- Storage: SSD storage for faster read/write operations, especially when dealing with substantial datasets.
- Graphics Card: A dedicated graphics card is not a strict requirement for Power BI, but it can enhance visualization performance, especially for complex charts and reports.
- Network: A stable internet connection is required for accessing online data sources and sharing reports online.
- Monitor: A high-resolution monitor for better visualization and user experience.

2. Software (S/W) Requirements:

- Power BI Desktop: The primary tool for creating interactive reports and dashboards. It's available for free from Microsoft's official website.
- Power BI Service: For sharing and collaborating on reports online. It requires a Power BI Pro or Premium license for advanced features.
- Data Sources: Power BI supports a wide range of data sources, including databases, Excel files, cloud services, and more. Ensure you have access to the necessary data sources for your churn prediction analysis.

3. Resources:

- Data Sources: Access to relevant and accurate data sources is crucial. This might include customer databases, transaction history, customer interactions, etc.
- Training Data: Historical data about customer churn, including features like customer demographics, usage patterns, customer support interactions, etc.
- Machine Learning Models: Resources for developing machine learning models to predict churn. This could involve tools like Python with libraries such as scikit-learn or Azure Machine Learning services.
- Documentation: Comprehensive documentation of data sources, data cleaning processes, feature selection, model training, and evaluation methods.

4. Requirements and Their Detailed Explanation:

- Data Preparation: Prepare your data by cleaning, transforming, and aggregating it. Identify the churn variable (target) and relevant features (predictors). Cleaning involves handling missing values and outliers, ensuring data consistency.
- Feature Selection: Identify the most relevant features that might influence churn, such as customer demographics, purchase history, customer interactions, etc. Feature selection techniques like correlation analysis or feature importance from machine learning models can be employed.
- Model Selection: Choose an appropriate machine learning algorithm for churn prediction. Common algorithms include logistic regression, decision trees, random forests, or neural networks. The choice depends on the complexity of your data and the interpretability of the model.
- Model Training: Train the selected machine learning model using historical data. Split the data into training and testing sets for evaluation. Optimize the model parameters to improve its predictive accuracy.
- Evaluation: Evaluate the model's performance using metrics like accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC). Select the metrics that align with your business goals. Iterate on the model and feature selection if the performance is not satisfactory.
- Integration with Power BI: Once the model is trained and validated, integrate it with Power BI using the "Python script" or "R script" visuals. Pass new data through the model to predict churn probabilities directly within Power BI dashboards.
- Visualization: Create intuitive visualizations in Power BI to present churn predictions. Use charts, graphs, and tables to communicate insights effectively. Interactive dashboards can be shared with stakeholders for real-time monitoring and decision-making.

b. Dataset Design

customerID: Unique identifier for each customer.
gender: Customer's gender (e.g., male, female).
SeniorCitizen: Indicates if the customer is a senior citizen (e.g., 0 for no, 1 for yes).
Partner: Indicates if the customer has a partner (e.g., yes, no).
Dependents: Indicates if the customer has dependents (e.g., yes, no)
tenure: Number of months the customer has been with the company.
PhoneService: Indicates if the customer has phone service (e.g., yes, no).
MultipleLines: Indicates if the customer has multiple phone lines (e.g., yes, no, or other options).
InternetService: Type of internet service subscribed (e.g., DSL, Fiber optic).
OnlineSecurity: Indicates if the customer has online security service (e.g., yes, no,

or other options).
OnlineBackup: Indicates if the customer has online backup service (e.g., yes, no, or other options).
DeviceProtection: Indicates if the customer has device protection service (e.g., yes, no, or other options).
TechSupport: Indicates if the customer has tech support service (e.g., yes, no, or other options).
StreamingTV: Indicates if the customer has streaming TV service (e.g., yes, no, or other options).
StreamingMovies: Indicates if the customer has streaming movies service (e.g., yes, no, or other options).
Contract: Type of contract the customer has (e.g., month-to-month, one-year, two-year).
PaperlessBilling: Indicates if the customer receives paperless billing (e.g., yes, no).
PaymentMethod: Method of payment for services (e.g., electronic check, credit card, etc.).
MonthlyCharges: The monthly cost the customer incurs.
TotalCharges: The total charges incurred by the customer.
Churn: The target variable, indicating if the customer churned (e.g., yes, no).

To design the dataset effectively, you should consider the following steps:

Data Collection: Ensure that data is collected accurately and consistently. Verify that all columns are populated appropriately.

Data Cleaning: Check for missing values, outliers, and inconsistencies in the dataset. Clean the data to make it suitable for analysis.

Data Encoding: Convert categorical variables into numerical format if necessary. For example, you might use one-hot encoding for categorical features.

Feature Engineering: Create new features if needed, based on domain knowledge or data analysis.

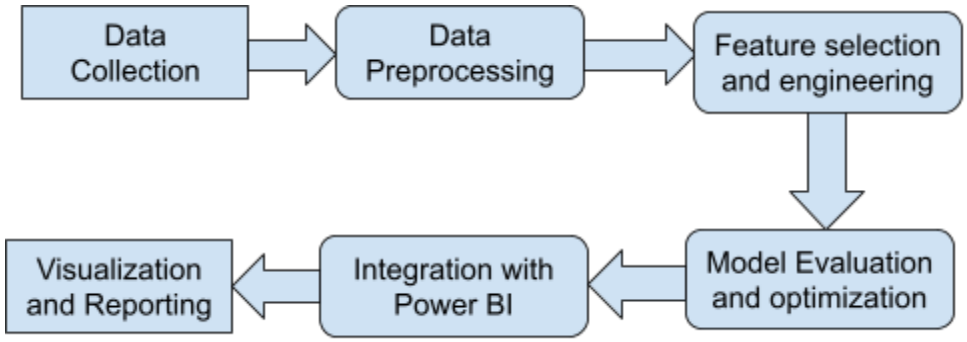
Exploratory Data Analysis (EDA): Perform EDA to understand the distribution of data, relationships between variables, and potential patterns related to churn.

Data Splitting: If you are building predictive models for churn, split the dataset into training and testing sets.

Data Documentation: Document the dataset's source, meaning of each column, and any data transformations applied.

Data Versioning: Keep track of different versions of the dataset if changes or updates occur.

c) Block diagram with explanation of each module for churn prediction in power bi



1. Data Collection and Integration:

- Explanation: In this module, data is collected from various sources such as customer databases, transaction records, customer interactions, etc. The data is then integrated and transformed into a format suitable for analysis. Power BI allows you to connect to diverse data sources and perform ETL (Extract, Transform, Load) operations to clean and integrate the data.

2. Data Preprocessing:

- Explanation: Raw data often requires preprocessing. This module involves tasks like handling missing values, outlier detection, normalization, and feature engineering. Power BI's Power Query Editor enables data cleaning and transformation. Clean and well-organized data is crucial for accurate churn prediction.

3. Feature Selection and Engineering:

- Explanation: Relevant features impacting churn are selected, and new features are created based on domain knowledge or data analysis. This module might involve statistical analysis or machine learning techniques to identify the most influential factors affecting churn. Power BI provides tools for advanced data analysis, aiding in feature selection.

4. Machine Learning Model Development:

- Explanation: Machine learning algorithms (e.g., logistic regression, decision trees, or neural networks) are trained on the preprocessed data. The model learns patterns from historical churn data. Python can be utilized within Power BI to develop and train machine learning models.

5. Model Evaluation and Optimization:

- Explanation: The trained model's performance is evaluated using metrics like

accuracy, precision, recall, etc. Optimization techniques, such as hyperparameter tuning, are applied to enhance the model's accuracy. Power BI's visualization capabilities can help in comparing predicted vs. actual churn rates and other evaluation metrics.

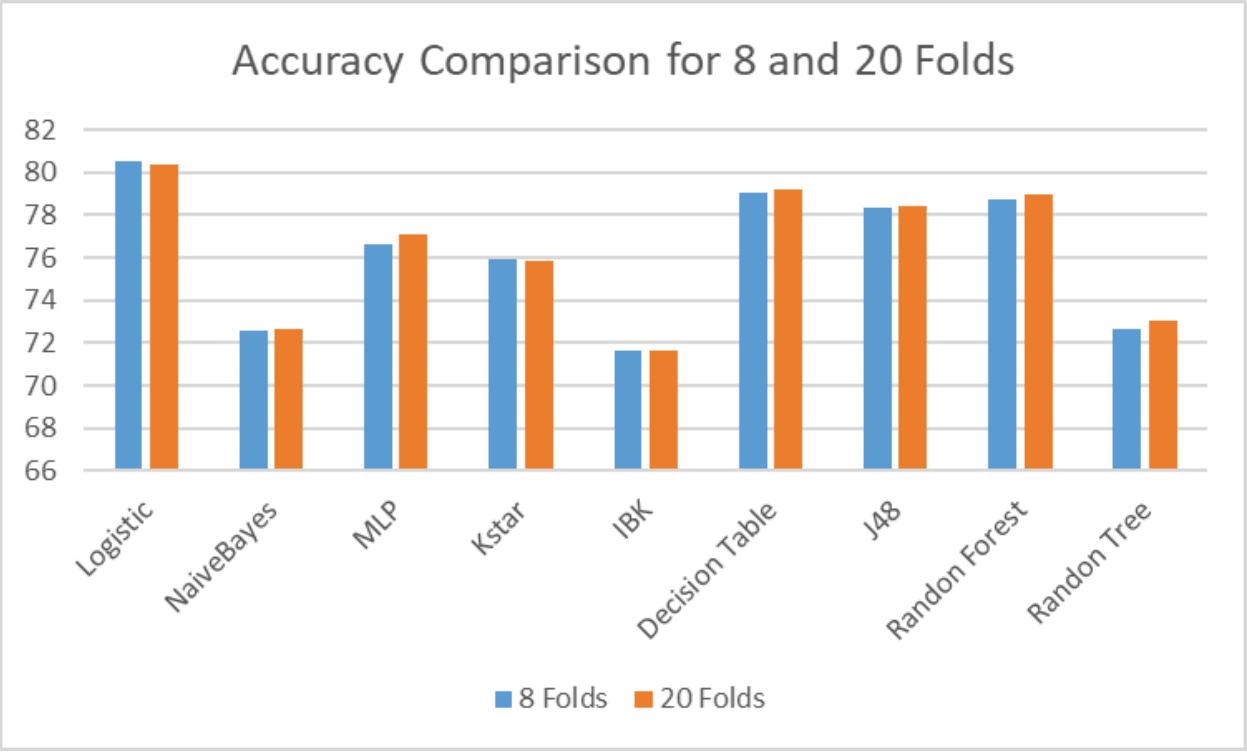
6. Integration with Power BI:

- Explanation: The trained and optimized machine learning model is integrated into Power BI using Python or R scripts. Power BI provides seamless integration with these scripting languages, allowing real-time predictions directly within the Power BI dashboard. New data points can be passed through the model to predict churn probabilities.

7. Visualization and Reporting:

- Explanation: Churn predictions and insights are visually represented using charts, graphs, and tables in Power BI. Interactive dashboards are created to allow users to explore the data and predictions.

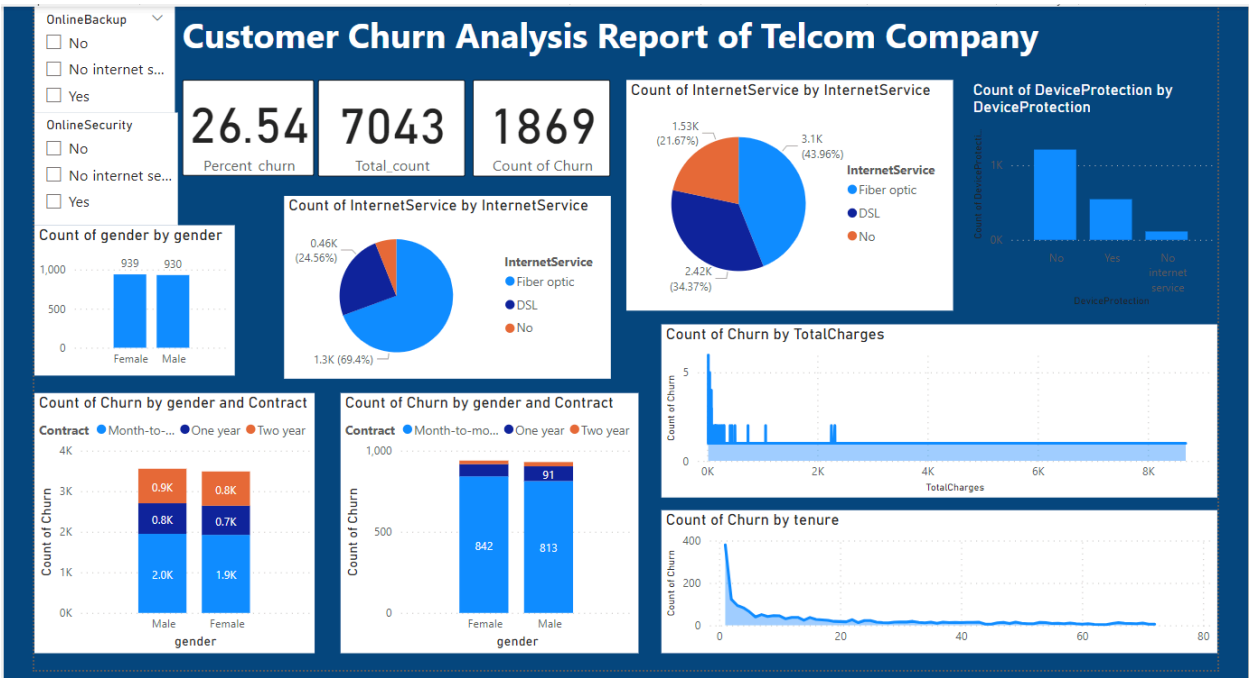
Results and discussion



Comparative analysis of all existing models:

Research Paper	Focus Area	Methods/Algorithms	Performance Metrics	Key Findings
A Prediction Model of Customer Churn Considering Customer Value: An Empirical Research of Telecom Industry in China (Hindawi Journal)	Telecom Industry (China)	Logistic Regression	Not specified	Identifies churn causes, emphasizes accurate identification for effective win-back strategies, underscores managing customer value for competitiveness.
Customer Churn Analysis in Telecom Industry (IEEE Xplore)	Telecom Industry	Decision Trees, Logistic Regression	Accuracy	Highlights importance of CRM and data mining for customer retention, emphasizes attribute grouping and threshold setting for better prediction outcomes.
Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms (Research Gate)	Telecom Industry	Decision Trees, Logistic Regression, Random Forest, XGBoost	Not specified	Reviews various ML algorithms, emphasizes data preprocessing, feature selection, and analysis based on obtained results for churn prediction.
Customer churn prediction in telecom using machine learning in big data platform (Springer Open)	Telecom Industry	Decision Tree, Random Forest, GBM, XGBOOST	Area Under Curve (AUC)	Achieves 93.3% AUC, integrates SNA features for performance enhancement, XGBOOST demonstrates best results among algorithms tested.
Customer churn prediction using machine learning (DiVA Portal)	Subscription-based B2B Services	XGBoost, Random Forest, Naïve Bayes	Accuracy, Precision, Recall, F1-score	Compares ensemble learners (XGBoost, Random Forest) with Naïve Bayes, explores various sampling methods to handle dataset imbalance, ensemble learners outperform.

Dashboard:



Conclusion:

- Predicting Churn for Improved Business Performance" aims to enhance customer retention through predictive churn analysis. By employing advanced data analytics and machine learning techniques, the project identifies potential churn patterns within the customer base.
- The precise conclusion of this business intelligence project is:Through predictive churn analysis, this project provides actionable insights for businesses to proactively identify and retain at-risk customers. By leveraging data-driven strategies, businesses can implement targeted retention efforts, optimize customer engagement, and ultimately improve overall business performance.
- Effective churn prediction facilitates resource allocation, personalized customer interactions, and tailored retention initiatives, ensuring enhanced customer loyalty, satisfaction, and long-term profitability.

References:

1. Yi Zong, Enze Pan, "A SOM-Based Customer Stratification Model", Wireless Communications and Mobile Computing, vol. 2022, Article ID 7479110, 8 pages, 2022. https://doi.org/10.1155/2022/7479110

2. K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359318.

3. Kavitha, V. & Kumar, G. & Kumar, S. & Harish, M.. (2020). Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS050022.

4. Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6