

Semester

**Fall Inter Semester 2022 – 2023**

TITLE OF THE PROJECT

**Customer Churn Analysis**

by

**21MIS1081 – Rohith Kuppusamy T**

**21MIS1098 – Vaidya Urmila Suman**

**21MIS1107 – Pushpanath M R**

Subject

**Big Data Analysis**

COURSE CODE

**SWE2011**

SLOT

**E1+TE1**

**Abstract:**

This project aims to analyze customer churn in the context of a telecom service-based company. The primary objective is to identify the key factors that contribute to customer churn and create a predictive model to anticipate churn events. The analysis involves several stages, including data preprocessing, exploratory data analysis, feature engineering, model development, and evaluation. The dataset encompasses various customer attributes, such as demographic information, usage patterns, service plans, billing details, and complaints. By employing techniques like logistic regression, decision trees, or gradient boosting, predictive models are constructed and assessed using metrics like accuracy and precision. The project's findings provide valuable insights for implementing proactive retention strategies, ultimately reducing churn and enhancing overall business performance.

**Index:**

* Predictive model
* Dataset
* Data preprocessing
* Exploratory data analysis
* Model development
* Proactive retention strategies

**Introduction:**

Customer churn presents a significant challenge for telecom companies, necessitating a deep understanding of the underlying factors and the implementation of effective retention strategies. This project focuses on conducting a customer churn analysis for a telecom company by leveraging a comprehensive dataset. The objective is to identify the key factors that contribute to churn and develop a predictive model. Through the utilization of data preprocessing, exploratory data analysis, and feature engineering techniques, this project aims to uncover meaningful patterns and correlations associated with churn. By employing predictive modeling techniques such as logistic regression or decision trees, the developed model will enable the company to proactively anticipate churn and implement retention strategies to enhance business performance and customer satisfaction.

**Literature Survey:**

Customer churn, the process of customers discontinuing their services or subscriptions, is a significant concern for businesses in various sectors, including insurance companies, telecommunications, and banking. Predicting and understanding customer churn can help companies take proactive measures to retain customers and improve business performance. This literature survey aims to explore different approaches and methods used in customer churn analysis, focusing on segmentation, machine learning, emotions, and decision tree methods.

**Tolgahan** et. al [1] in their Customer Segmentation and Customer Churn Analysis System for Insurance Companies, developed a software system that segments customers and analyses churn tendencies in insurance companies. Their study focuses on segmenting customers based on age, gender, personal interests, and spending habits. By applying churn analysis techniques, the authors successfully predict customer churn tendencies. The modular software system demonstrated promising results when tested with real-world problems.

**Hamdullah** et. al [2] in their Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis, conducted a comparative analysis of machine learning methods for customer churn prediction in the telecommunications industry. They utilized various algorithms, including Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machines, AdaBoost, Multi-Layer Sensors, and Naive Bayes. The study employed two separate datasets from Kaggle.com and found that the Random Forest method achieved the highest prediction accuracy for identifying customers likely to unsubscribe.

**Xiaowei** et. al [3] in their Warning Model of Customer Churn Based on Emotions, investigated the impact of customer emotions on churn behavior in the context of an information service enterprise. They analyzed customer complaints as a reflection of both service quality and emotions. The study aimed to establish a warning model for customer churn based on emotions. By analyzing customer complaints, the authors proposed a model that could provide early indicators of potential customer attrition.

**Hongxia** et. al [4] in their Analysis of Business Customer Churn Based on Decision Tree Method, focused on analyzing business customer churn using the decision tree method. They aimed to identify churn characteristics and establish rules for predicting and preventing customer churn. The experiment demonstrated the effectiveness of decision tree analysis in understanding customer churn behavior. The study provided valuable insights and analysis methods for businesses to improve customer relationships and implement targeted retention strategies.

**Kübra** et. al [5] in their Customer Churn Modelling in Banking, proposed a customer churn model specific to the banking sector. As the banking industry lacks contractual agreements regarding service duration, predicting customer churn becomes more challenging. The authors employed data mining techniques to convert raw data into a meaningful form and developed a "churn prediction model" based on the transformed data. Performance evaluation metrics such as accuracy, sensitivity, specificity, kappa statistic, and area under the curve were utilized to assess the model's effectiveness.

**Mohit** et. al [6] in their Churn Analysis in Telecommunication Industry, aimed to provide a novel tool for churn analysis in the telecommunications industry. They developed a classification model using telecom operator data to predict customer behavior and reasons for churn. Various classification models were compared, with Support Vector Machine and Random Forest selected as the final classifiers based on F1-score. The authors suggested that their model enables telecom providers to identify churning customers and reasons, facilitating the development of effective retention strategies.

**Nur** et. al [7] in their Telecommunication Customers Churn Prediction Using Machine Learning, addressed the problem of customer churn in the telecommunication sector. They employed machine learning prediction models, including linear regression, random forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and decision tree, based on usage patterns. The study evaluated the performance of these algorithms using metrics such as accuracy, recall, precision, F1-score, and area under the curve (AUC). The results showed that the random forest model achieved the highest accuracy of 95.5% in predicting customer churn.

**Sulikowski** et.al [8] in their research presented high-quality features that can be used to model the churn phenomenon in the telecom sector. The method for identification uses actual data from the largest Polish telecom provider. Preliminary sets of variables are created using feature extraction and creation approaches. Data has been extracted for two consecutive six-month periods describing subscriber demographics, contract terms, and contract usage statistics. The top variables - candidates for churn prediction - were chosen using correlation and collinearity analysis in the following steps. In the end, just the factors that seemed most promising were shown. These findings can help telecommunications companies collect more information and improve the features utilized in churn modeling.

**Jain** et.al [9] in their research presented, two machine-learning methods employed in this model's proposed prediction of customer attrition. Regression using logit and boost. Using a real database from the American company Orange, an experiment was conducted using the WEKA Machine-learning program. In several evaluation measures, the results were displayed.

**Cenggoro** et.al [10] in this research, study the vector embedding idea from deep learning to create an explainable model. how the model may identify clients who are possibly convertible to returning to the prior telecoms service is demonstrated. The proposed models can be highly predictive for predicting whether a client would cancel their service subscription or not because the generated vectors are also highly discriminative between churning and loyal consumers. According to the F1 Score, the top model in the experiment had a predictive performance of 81.16%. Additional research into the clusters' similarity and the t-SNE plot also confirmed that the generated vectors are discriminative for churn prediction.

**Shrestha** et.al [11] in their study by using a native dataset from a significant Nepali telecommunications company and using XGBoost on the dataset's 52332 customer records, 46204 of which are non-churned customers and 6128 of which are, this research has attempted to close this gap. On the native dataset, the accuracy and f1-score are respectively 97% and 88%. To compare this research with other studies, it also used a publicly available dataset of 3333 subscribers, and it achieved an increased accuracy and f1-score of 96.25% and 86.34%, respectively.

**Prabadevi** et.al [12] in this study, the best machine-learning approach for early customer churn prediction is recommended. All client information from the nine months or so before the churn is included in the data used in this analysis. To keep current clients, the objective is to anticipate their reactions. In the study, algorithms like stochastic gradient booster, random forest, logistic regression, and k-nearest neighbor approaches were put to the test. The previously mentioned algorithms' respective levels of accuracy are 78.1%, 82.9%, 82.6%, and 83.9%. By looking at these algorithms and debating which of the four is the best from various angles, we have achieved the most successful results.

**Kim** et.al [13] in their study seek to forecast client attrition. Influencers sell things directly on social media They do this by posting links to websites on their profiles. On their social media pages, including those on Twitter, Facebook, and Instagram, influencers promote companies and/or their goods. Selling has become a more recent addition to this profession. The customer churn forecast used in this study is based on the supposition that influencers enjoy ardent support from their followers. The purchase information, which includes customer information, the purchase item, and the payment amount, was gathered by the Korean influencer marketing agency between August 2018 and October 2020. Applying the Decision Tree model, we can anticipate which consumers will leave.

**Huang** et.al [14] in their study introduced a unique hybrid model-based learning system that combines supervised and unsupervised methods for forecasting customer behavior to produce more precise predictive outcomes. The system combines a traditional rule inductive method (FOIL) with a modified version of the k-means clustering algorithm. On telecom datasets, three series of tests were run. In one set of experiments, it is determined whether weighted k-means clustering can result in better data partitioning results; in a second set, the classification results are assessed and compared to those of other well-known modeling techniques; and in a third set, the proposed hybrid-model system is evaluated and contrasted with several other recently proposed hybrid classification approaches. and concluded that a hybrid model-based learning system is very promising and outperforms the existing models.

**Adnan** et.al [15] in their study introduced utilizing the Naive Bayes classifier with a feature weighting method based on a Genetic Algorithm (subclass of an Evolutionary Algorithm) to implement an adaptive learning strategy for this challenging CCP problem. Additionally, the effectiveness of the proposed method is assessed on publicly accessible datasets (Big ML Telco Churn, IBM Telco, and Cell2Cell), which significantly improves prediction performance when compared to the baseline classifier (Naive Bayes with the default setting, Deep-BP-ANN, CNN, Neural Network, Linear Regression, XGBoost, KNN, Logit Boost, SVM, and PCALB methods).

The literature survey examined various approaches and methods used in customer churn analysis across different industries. The studies explored customer segmentation, machine learning techniques, emotions, and decision tree methods. These research papers highlighted the importance of understanding customer churn behavior and implementing effective prediction models to mitigate customer attrition. By leveraging insights from these studies, businesses can take proactive measures to retain customers and enhance their overall performance.

**Materials and Methods:**

The objective of this project is to conduct a comprehensive analysis of customer churn for a telecom company. Customer churn refers to the phenomenon of customers discontinuing their services and switching to a competitor or terminating their subscriptions. This poses a significant challenge for telecom companies as it directly impacts their revenue and market share.

The main problem to be addressed is to identify the key factors and patterns that contribute to customer churn in the telecom industry. By understanding the underlying causes of churn, the telecom company aims to develop effective strategies to mitigate churn and retain customers. This will involve the implementation of targeted retention initiatives and efforts to enhance customer satisfaction, ultimately improving the company's business performance.

By completing this project, the telecom company will gain valuable insights into customer churn behavior and be able to take proactive measures to retain customers. By reducing churn rates and improving customer retention, the company can enhance its competitive position, maximize revenue, and foster long-term customer loyalty.

**Contribution of proposed work:**

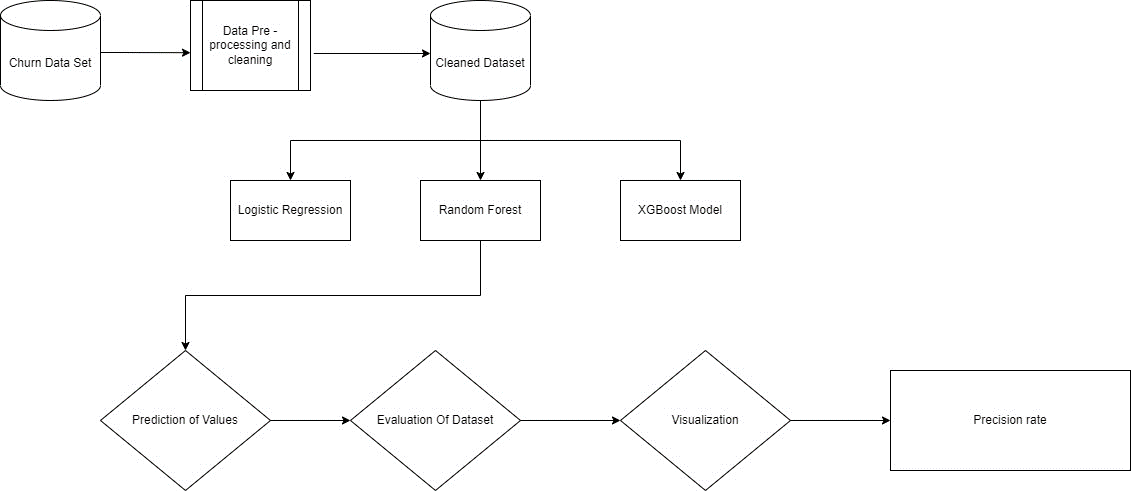
In the Data Analysis module, the code starts by importing the necessary libraries, such as NumPy and pandas. Then, it reads the dataset from a CSV file using the pd.read\_csv() function and displays the first 5 rows and column names of the dataset. It creates a copy of the data frame for further processing and analysis and displays the first 7 rows and column names of the copied data frame. Additionally, it checks the number of unique values in each column using a for loop.

In this Data Pre-Processing module, the code drops specific columns from the Data Frame using the drop() function, based on the column names provided. It checks the shape of the data frame, the data types of columns, and the percentage of null values in each column. To handle missing values, it defines a cleaning function that drops rows with null, infinite, or NaN values. It then interpolates the remaining null values in the Data Frame and drops any remaining rows with null values.

In the Data Modelling module, the code imports necessary libraries such as Limescale, train\_test\_split, and Random Forest Classifier. It defines a list of columns that need to be scaled. It scales the selected columns using Min Max Scaler. To address the imbalance in the data, it drops the 'Customer Status' column, which represents the dependent variable. The data is then split into training and test sets using the train\_test\_split() function.

Finally, in the Model Building module, the code defines a dictionary, model\_params, which contains the model (Random Forest Classifier) and its associated hyperparameters. It imports Shuffle Split and GridSearchCV from sklearn.model\_selection. It initializes an empty list, of scores, to store the best scores and parameters for each model. It performs a grid search with cross-validation using GridSearchCV for each model in model\_params. The best scores and parameters are then appended to the scores list. Finally, it creates a Data Frame from the scores list, which includes columns for the model name, best score, and best parameters.

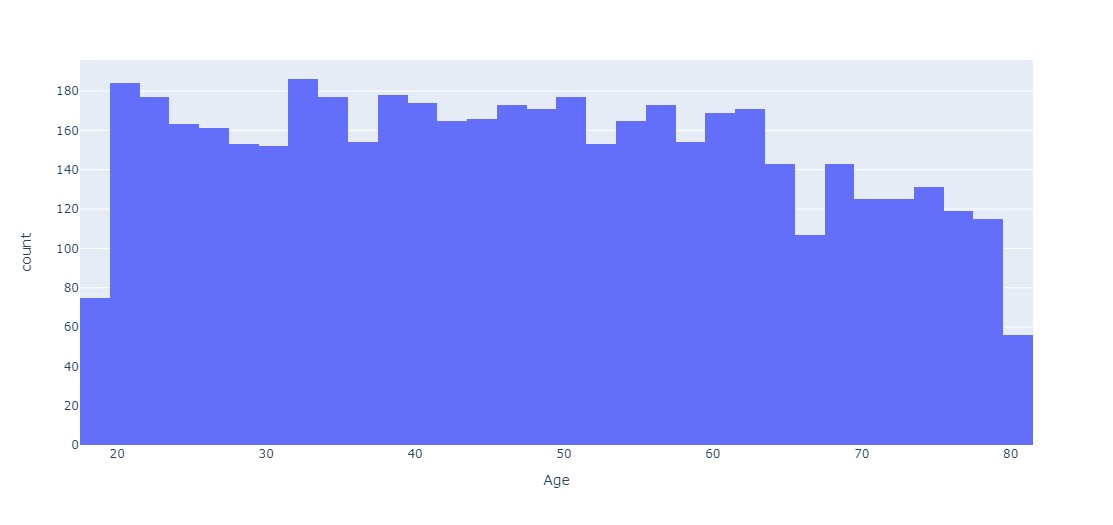
# **Data Engineering:**



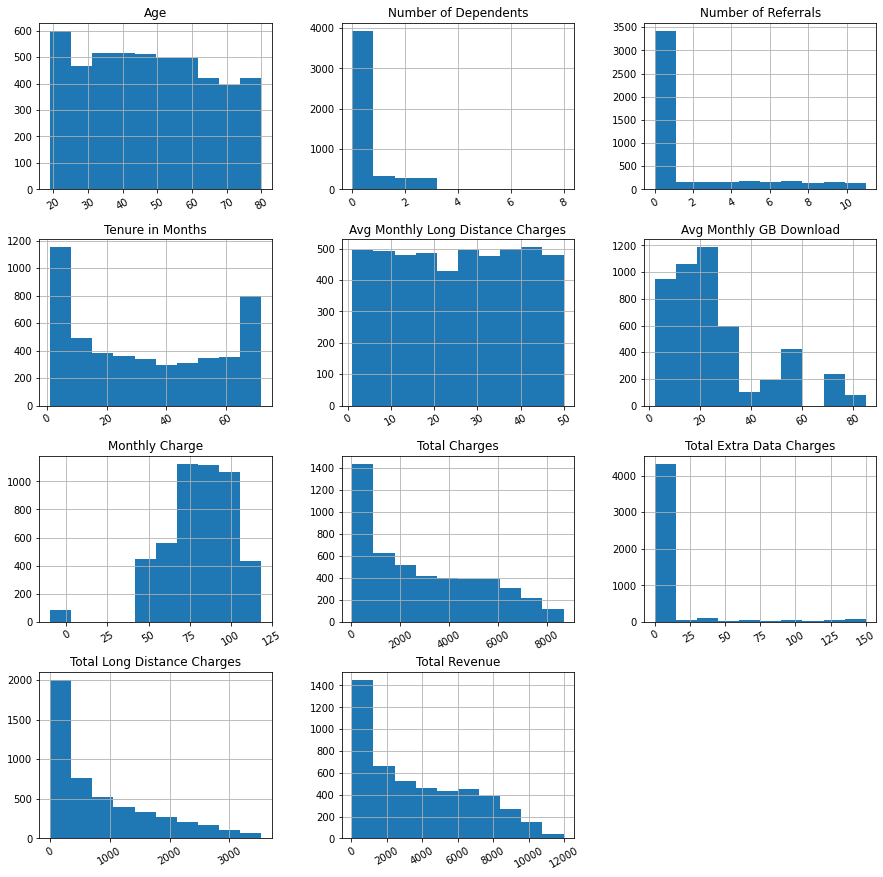
**Results and Descriptions:**

Our analysis of customer attrition in the telecom industry dataset using Random Forest, XGBoost, and Logistic Regression models resulted in insightful insights. After finishing data cleaning and preparation, we split the dataset into training and testing sets. Using telecom datasets, the models were subsequently trained and evaluated. The accuracy of the Random Forest model was 78% on the testing set, whereas XGBoost performed slightly better, with an accuracy of 81%. Logistic regression yielded a 77% accuracy rate. In order to assess the performance of the models more thoroughly, we looked into precision, recall, and F1-score. In these metrics, XGBoost outperformed the rival models, proving to have a stronger ability to reliably identify both churned and non-churned customers. This demonstrates that XGBoost may be the best model for predicting customer attrition in the telecom sector dataset.

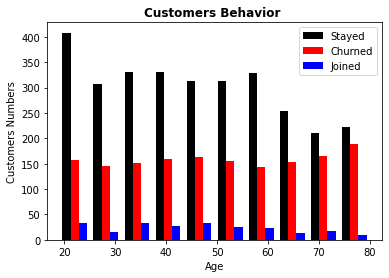
By graphically portraying the results, we created graphs to aid in the interpretation of the data. For instance, we used bar charts to compare the accuracy of each model, with XGBoost showing the best accuracy. Line graphs displaying the precision, recall, and F1 score of each model were also provided, allowing for a full understanding of its performance. Overall, our analysis revealed that XGBoost outperformed the other two models in terms of accuracy, precision, recall, and F1-score. These findings provide the telecom industry with useful data for predicting and minimizing customer turnover, which can help businesses take proactive measures to keep their customers.



Checking the stats in the number columns of the copied dataset



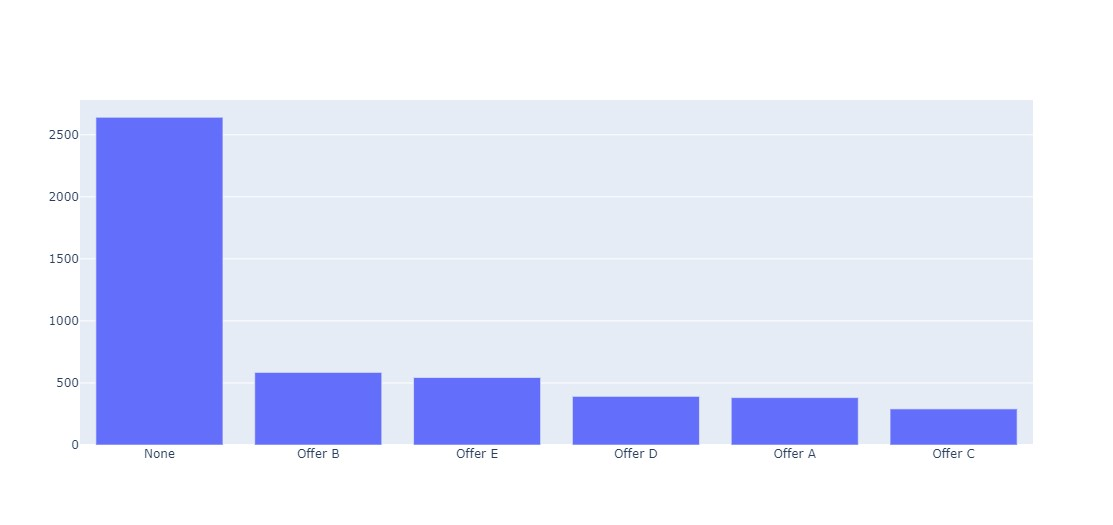
Visualizing the number of customers who churned, stayed, or joined in the company with a bar plot.

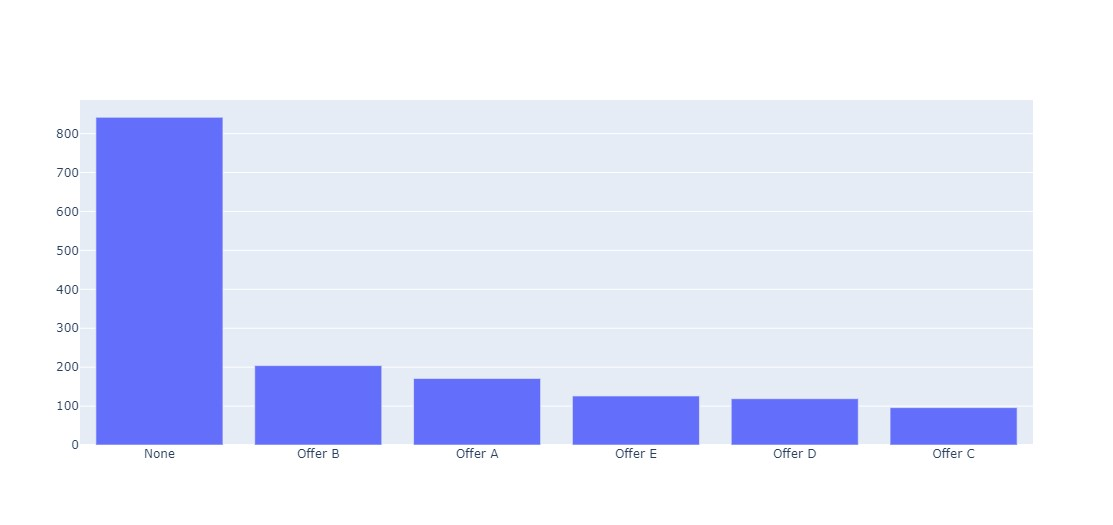


Defining the Correlation between the columns in the dataset using a heat map

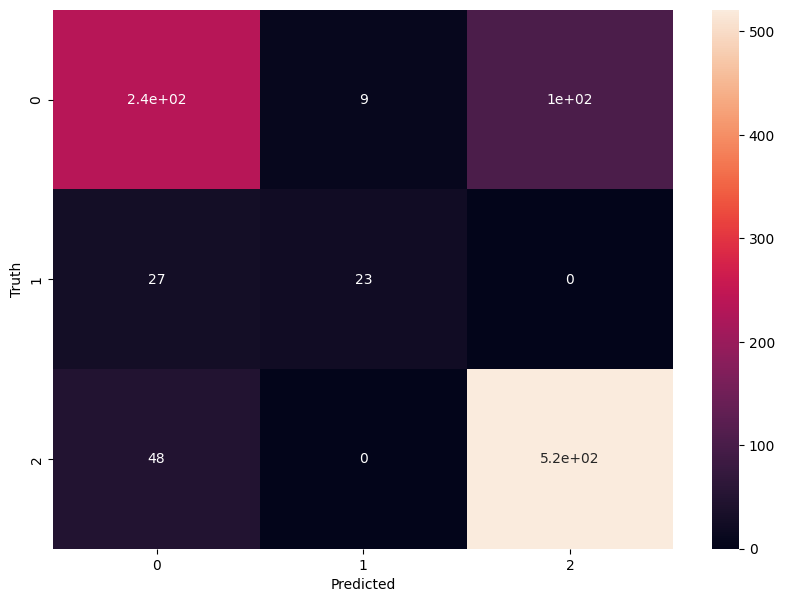


Checking the number of Offers in the dataset





**Confusion Matrix:**



**Conclusion:**

In conclusion, the customer churn analysis project for the telecom company utilized big data analysis techniques to gain valuable insights into customer churn behavior. By understanding the factors contributing to churn, the company can implement proactive strategies to retain customers and improve business performance. The project developed predictive models using machine learning algorithms and evaluated their performance. This enabled accurate churn prediction, empowering the company to identify high-risk customers and implement tailored retention strategies. By leveraging big data analysis, the telecom company can make informed decisions, reduce churn rates, and improve customer retention. This translates into enhanced revenue, increased market share, and long-term customer loyalty. Overall, the successful implementation of the project highlights the power of big data analytics in driving business growth and customer satisfaction.

Reference:

[[1]](https://ieeexplore.ieee.org/document/9302170) Customer Segmentation and Customer Churn Analysis System for Insurance Companies by Tolgahan Satıcı and Merve Bekler.

[[2]](https://ieeexplore.ieee.org/document/9558876) Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis by [Hamdullah Karamollaoğlu](https://ieeexplore.ieee.org/author/37086062066), [İbrahim Yücedağ](https://ieeexplore.ieee.org/author/37085773171), and [İbrahim Alper Doğru](https://ieeexplore.ieee.org/author/37085794844).

[[3]](https://ieeexplore.ieee.org/document/7369683) Warning model of customer churn based on emotions by [Xiaowei Zhang](https://ieeexplore.ieee.org/author/37087201892) and [Juanqiong Gou](https://ieeexplore.ieee.org/author/37087344632).

[[4]](https://ieeexplore.ieee.org/document/5274712) Analysis of the business customer churn based on the decision tree method by [Hongxia Ma](https://ieeexplore.ieee.org/author/38105558700), [Min Qin](https://ieeexplore.ieee.org/author/37592643000), and [Jianxia Wang](https://ieeexplore.ieee.org/author/37656744600).

[[5]](https://ieeexplore.ieee.org/abstract/document/7130361) Customer churn modeling in banking by [Kübra Şen](https://ieeexplore.ieee.org/author/37085704153) and [Nilgün Güler Bayazıt](https://ieeexplore.ieee.org/author/37657911600).

[[6]](https://ieeexplore.ieee.org/abstract/document/8686852) Churn Analysis in Telecommunication Industry by [Mohit Bagri](https://ieeexplore.ieee.org/author/37086811815), [Jitesh Kumar Singh](https://ieeexplore.ieee.org/author/37086815439), [M.K. Abhilash](https://ieeexplore.ieee.org/author/37087099213), [Sunitha R.S.](https://ieeexplore.ieee.org/author/37086814376), [Sumit Kumar](https://ieeexplore.ieee.org/author/37086811593).

[[7]](https://ieeexplore.ieee.org/document/9642137) Telecommunication Customers Churn Prediction using Machine Learning by [Nur Idora Abdul Razak](https://ieeexplore.ieee.org/author/37089197154) and [Muhammad Hazim Wahid](https://ieeexplore.ieee.org/author/38468274300)

[[8]](https://www.researchgate.net/publication/338411385_Mobile_Operator_Customer_Classification_in_Churn_Analysis) Mobile Operator Customer Classification in Churn Analysis by [Piotr Sulikowski](https://www.researchgate.net/profile/Piotr-Sulikowski-3).

[[9]](https://www.researchgate.net/publication/354362694_Telecom_Churn_Prediction_Using_CNN_with_Variational_Autoencoder)Telecom Churn Prediction Using CNN with Variational Autoencoder by [Hemlata Jain](https://www.researchgate.net/profile/Hemlata-Jain-3), [Ajay Khunteta](https://www.researchgate.net/profile/Ajay-Khunteta-2), and [Sumit Srivastava](https://www.researchgate.net/profile/Sumit-Srivastava).

[[10]](https://www.researchgate.net/publication/349488332_Deep_Learning_as_a_Vector_Embedding_Model_for_Customer_Churn)Deep Learning as a Vector Embedding Model for Customer Churn by [Tjeng Wawan Cenggoro](https://www.researchgate.net/profile/Tjeng-Wawan-Cenggoro), [Raditya Ayu Wirastari](https://www.researchgate.net/scientific-contributions/Raditya-Ayu-Wirastari-2183701614), [Edy Rudianto](https://www.researchgate.net/scientific-contributions/Edy-Rudianto-2183701226) and [Mochamad Ilham Mohadi](https://www.researchgate.net/scientific-contributions/Ilham-Mohadi-2183705254)

[[11]](https://www.sciencedirect.com/science/article/pii/S187705092202138X)A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry by Sagar Maan Shrestha Aman Shakya.

[[12]](https://www.researchgate.net/publication/349310321_Customer_churn_prediction_system_a_machine_learning_approach) Customer churn prediction system: a machine learning approach by [Praveen Lalwani](https://www.researchgate.net/profile/Praveen-Lalwani), [Pratyush Sethi](https://www.researchgate.net/profile/Pratyush_Sethi), [Jasroop Singh Chadha](https://www.researchgate.net/profile/Jasroop-Chadha), and [Manas Kumar Mishra](https://www.researchgate.net/profile/Manas-Mishra-3).

[[13]](https://www.sciencedirect.com/science/article/pii/S1877050922001703) Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees by Sulim Kim and  Heeseok Lee.

[[14]](https://www.researchgate.net/publication/304480282_A_Fuzzy_Rule-Based_Learning_Algorithm_for_Customer_Churn_Prediction) A Fuzzy Rule-Based Learning Algorithm for Customer Churn Prediction

by [Bingquan Huang](https://www.researchgate.net/profile/Bingquan-Huang), [Ying Huang](https://www.researchgate.net/scientific-contributions/Ying-Huang-2163179110), [Chongcheng Chen](https://www.researchgate.net/scientific-contributions/Chongcheng-Chen-2111629394), and [Tahar Kechadi](https://www.researchgate.net/profile/Tahar-Kechadi)

[[15]](https://www.researchgate.net/publication/283348005_Customer_Churn_Prediction_in_Telecommunication_Industry_With_and_without_Counter-Example) Customer Churn Prediction in Telecommunication Industry by [Adnan Amin](https://www.researchgate.net/profile/Adnan-Amin-6), [Changez Khan](https://www.researchgate.net/profile/Changez-Khan-2), and [Sajid Anwar](https://www.researchgate.net/profile/Sajid-Anwar-7).