Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

Customer Churn Prediction

Lavina Anand Parulekar and Sanika Abasaheb Sardesai
SSPM's College of Engineering
Computer Engineering
Mumbai University

Prajkta Shriram Jamsandekar and Sampada Sanjay Parkar
SSPM's College of Engineering
Computer Engineering
Mumbai University

Sawant S.P (Prof.)
SSPM's College of Engineering
Computer Engineering
Mumbai University

Abstract:- In today's highly competitive business landscape, customer retention revenue preservation, customer experience improvement, and marketing optimisation are critical factors for sustained growth and profitability. Customer churn prediction is discontinuing their services or purchases, which presents a significant challenge for businesses across various industries. This project focuses on developing a predictive model to expect customer churn in the banking sector using machine learning techniques. The project involves the collection and analysis of historical customer data, confined account activity, transaction history, demographic information, and customer service interactions. By extracting the right features from this data, a machine learning model is trained to forecast which bank customers are at the highest risk of churning. A critical step in this study was the selection of relevant features that influence customer churn. Feature selection was guided by domain knowledge and feature importance analysis. The different classifiers were used and then trained on the training dataset further ensuring the model's optimal performance. The model's performance is assessed through various evaluation metrics, including accuracy, precision, and recall. Additionally, the project explores a model illustration to uncover the influential factors contributing to customer churn within the banking context. This project's outcomes can empower banks to take proactive measures in retaining customers, enhancing their overall experience, and thereby preserving revenue streams. By addressing customer churn, banks can foster long-term relationships, reduce customer acquisition costs, and boost their competitiveness in the financial industry. The results of this project are expected to assist businesses in proactively retaining customers by targeting those at the highest risk of churning. Ultimately, reducing customer churn can lead to increased customer satisfaction, revenue, and long-term business sustainability.

I. INTRODUCTION

Customer churn refers to the problem where customers stop doing business with a company or organization. It is a critical thing for businesses to understand and predict. It is the rate at which customers stop doing business with a particular company. It's a critical thing because retaining existing customers is sometimes more costly than acquiring new customers. This project aims to tackle the high customer churn rate in the Banking Sector. In the banking industry, retaining existing customers is sometimes more costly than acquiring new ones. Loyal, long-term customers use multiple services offered by the bank, such as savings accounts, loans, credit cards, and investment products. Losing customers can result in a significant loss of revenue. Due to the high customer churn rate, industries have to face problems like loss of revenue, reduced profitability, and damage to the brand's reputation. We are doing this project to develop a predictive model that can identify customers at risk of churning. It is a classification model. The limitations of this predictive model are data availability, time, and resources. The accuracy of the churn prediction model depends on the Quality of data, representativeness of data, fewer variations in data, feature selection, dimensionality reduction, time, etc. We used Random forest, KKN, XG-Boost, Naive Bayes, etc. to check the accuracy of the model on the dataset. From that, we observed that random forest gives higher accuracy. To implement this project we were required to implement certain steps like Data collection, selection of specific features to train the model, model building, training of the model, validation of the model's result, prediction of newly incoming data. The importance of the customer churn prediction is that it helps to Predicting customer churn that allows companies to proactively take actions to retain their valuable customers, minimize revenue loss, and improve customer satisfaction and loyalty.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

II. RELATED WORK

Several studies shows that different techniques are used tofind out Customer Churn Rate in different fields:

- Amol, Chole,in 2023 evaluated Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Convolutional Neural Network (CNN) algorithmsfor predicting NonChurn customers.
- Ziyu Zhu,in 2023 study explored Logistic Regression and Random Forest algorithms, reporting AUC values of 78%and 69%˙
- Vani Haridasan ,in 2023 study introduced the Arithmetic Optimization Algorithm (AOA) and Stacked Bidirectional Long Short-Term Memory (SBLSTM), demonstrating the effectiveness of the AOASBLSTM model in classifying CR and NCR classes
- Micheal Olaolu Arowolo , in 2022 study compared CNN and Random Forest classifiers, reporting prediction accuraciesof 94% and 91 %, for forecasting the churn rate in telecom companies.
- Dr. K. Geetha ,in 2022 study investigated Decision Trees, Random Forests, and XGBoost algorithms, capable of efficiently excluding non-essential data and forecasting the training model ahead of time.
- Sena Kasim ,in 2022 research identified Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and K-NN algorithms, indicating the suitability of ten features for churn analysis.
- Edemealem Desalegn Kingawa,in 2022 study explored Logistic and Multiple Regression, SVM, DNN, and Random Forest algorithms, ultimately selecting a Deep Neural Network model for prediction, with a recommendation to train and test it on a large dataset
- Sulim Kima,in 2022 study focused on Decision Trees for churn prediction in e-commerce, comparing original data with predicted results to accurately identify churners, highlighting how advancements in e-commerce have provided customers with numerous purchasing options.
- Dr. S.B. Kishor,in 2022 study utilized Deep Learning (DL), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN), preprocessing raw data to convert categorical values and balancing the dataset to 7032 instances, emphasizing the telecom industry's dependence on customer feedback.
- Madhavi Kasa,in 2021 study compared Random Forest, Decision Tree, and Na¨ıve Bayes algorithms, highlighting the superior efficiency of the rough set-based model in classifyingchurns from non-churns.

III. PROBLEM STATEMENT

In today's highly competitive business world, customer retention is a critical factor for sustained growth and profitability. Organizations face the challenge of high customer churn rates, which are affecting revenue and decreasing the ability to achieve long-term success. To address this issue, we need an advanced customer churn prediction system based on data analytics and machine learning techniques. This system should accurately identify customers at risk of churning and provide actionable measures to implement retention strategies. Developing a predictive model capable of accuratelyidentifying customers who are at risk of churn in the future. This involves analyzing historical customer data, transaction records, interaction history, and other relevant information to recognize churn patterns.

IV. KEY HIGHLIGHTS

➢ *Retention*

The primary goal is to retain valuable customers. By identifying those at risk of churning, a business can implement strategies to get them back.

➢ *Revenue Preservation*

Churn prediction helps businesses protect their revenue by preventing customer loss. Retaining existing customers is often more cost-effective than acquiringnew ones.

➢ *Customer Experience Improvement*

Understanding why customers churn provides the areas where improvements are needed. This information can guide product and serviceenhancements.

➢ *Marketing Optimization*

Churn prediction enables businesses to allocate marketing resources more efficiently. So that companies can more focus on retaining high-risk customers and serving the loyal ones.

V. PROPOSED SYSTEM

A proposed system for customer churn prediction involves a detailed approach to identify and predict customers who are likely to leave a bank or stop using its services.

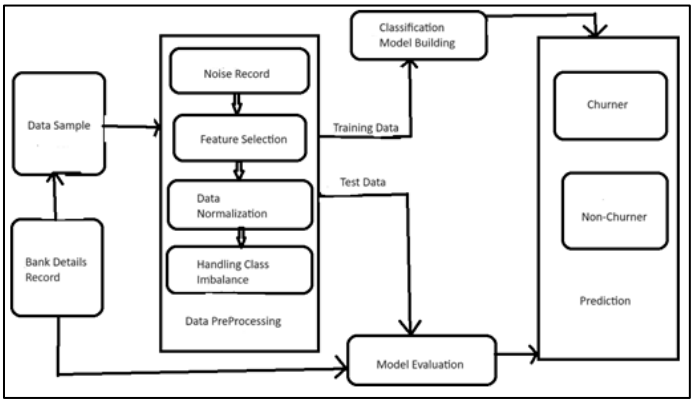

Fig. 1. Flowchart of proposed system

Here's a detailed outline of Customer Churn Prediction asystem that involves the following steps:

- Collection of Bank Details Record.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

- Data Preprocessing.
- Classification model building.
- Splitting data into training data and test data.
- Providing training data to the Classification model.
- Providing test data to the Evaluation model.
- Prediction of Churner And non-Churner.

A. Algorithms

In this project, we are using methods for model building such as classification algorithms which can detect the accuracy of given input data, and from that, we can compare which Machine learning algorithms are best for implementing. In this project, our methodology involves the use of classification algorithms to build a predictive model that can accurately detect potential customer churn. We will collect and preprocess relevant customer data, perform exploratory data analysis to gain insights into customer behavior, select important features, and train a machine-learning model to predict churn. We will evaluate the model's performance using various metrics and use clustering algorithms to segment customers based on churn probability or behaviour. Additionally, we will implement an early warning system to alert stakeholders of potential churnrisks. To train the model we are using logistic regression, naive Bayes, Decision tree, Random forest, k-neighbour, SVM, and XG Boost classification algorithms.

➢ *Logistic Regression*

Logistic Regression is a statistical method used for binary classification, which means it's primarily used to predict oneof two possible outcomes based on a set of input variables. In logistic regression, the dependent variable is categorical, representing a binary outcome (e.g., yes/no, 1/0, true/false).

➢ *Naive Bayes*

Naive Bayes classifier is a simple probabilistic machine learning algorithm that is based on Baye's theorem. it is particularly useful for classification tasks where we can predict the categorical data.

➢ *Decision Tree*

A decision tree algorithm is a popular machine learning technique used for both classifications. It isa tree-like model that makes decisions or predictions by recursively splitting the data into subsets based on the most significant features.

➢ *Random Forest*

Random Forest is an ensemble machine learning algorithm that works by constructing a multitude of decision trees during training and combining their predictions to make more accurate and robust predictions. Random Forest is known for its high predictive accuracy. It excels at capturing complex relationships within the data, which is often necessary when predicting churn.

➢ *k-Nearest Neighbour*

K-NN is a lazy learner while K-Means is an eager learner. An eager learner has a model fitting that means a training step but a lazy learner does not have a training phase.

➢ *SVM*

SVMs are particularly effective in solving classification problems, and they work by finding the optimal hyperplane that best separates data into different classes.

➢ *XG Boost*

It is known for its exceptional predictive performance, speed, and versatility in handling various types of machine learning tasks, particularly regression and classification.

B. Methodology

The methodology for customer churn prediction involves a systematic approach to build and deploy predictive models that can identify customers likely to churn.

Here's a step-by-step methodology for customer churn prediction:

➢ *Data Collection and Preprocessing:*

- *Data Gathering:* Collecting relevant customer data from various sources
- *Data Cleaning:* Removing duplicates, handling missing values, and addressing data inconsistencies to ensure data quality.
- *Feature Engineering:* Creating new features or transforming existing ones to extract relevant information that can improve the predictive model's performance.

➢ *Exploratory Data Analysis (EDA):*

Visualizing and exploring the data to get knowledge into customer behaviour,patterns, and potential churn drivers.

➢ *Feature Selection*

Removing irrelevant or highlycorrelated features to avoid overfitting.

➢ *Model Selection:*

Choosing appropriate machine learningalgorithms for churn prediction.

➢ *Model Training*

Splitting the dataset into training and testing sets for model training and evaluation.

➢ *Model Evaluation*

Assessing the model's performance using metrics like accuracy, precision, recall, F1 score, and AUC-ROC on the test set.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

➢ *Customer Segmentation*

Utilizing clustering algorithms (e.g., k-means, hierarchical clustering) to segment customersbased on churn probability or behaviour.

➢ *Early Warning System*

With the help of the threshold churn rate, we can alert businesses so they can avoid loss.

➢ *Retention Strategies and Recommendations*

By applying different strategies And Providing actionable recommendations on how to improve customer engagementand satisfaction

VI. IMPLEMENTATION SETUP

Implementation Setup for Customer Churn Prediction Modelinvolves the following :

A. Details about Input to Systems

In customer churn prediction projects Input data plays an important role in training and testing the predictive model. The quality and accuracy of Input data have a high impact on the output of the predictive model. In our Project, we will use a banking dataset for training our prediction model by using various machine learning algorithms This input data includes the following points:

➢ *Customer Demographics*

Personal information about thecustomer, such as Name, age, gender, and geography.

➢ *Customer Account Information*

Details about the customer's accounts, CustomerId, Tenure.

➢ *Financial History*

Information related to the customer's credit score, HasCrCard, and Estimated salary.

➢ *Customer Interaction*

Data about the customer's interactions with the bank such as Is, NumOfProducts,ActiveMember.

➢ *Transaction History*

Records of the customer's past transactions such as Balance, when the customer is Exited.

B. Performance Evaluation Parameters

Performance matrices are the set of matrices that help to assess the effectiveness of the predictive model. In our project, we will use the following matrices.

➢ *Confusion Matrix*

- True Positives (TP): The number of correctly predicted churn cases.
- True Negatives (TN):The number of correctly predicted non-churn cases.
- False Positives (FP):The number of non-churn cases predicted as churn.
- False Negatives(FN):The number of churn cases predictedas nonchurn.

➢ *Accuracy (ACC)*

$$(TP + TN)/(TP + TN + FP + FN) \quad (1)$$

Measures the overall correct predictions.

➢ *Recall*

- *Google Sheets(For CSV)*

Google Sheets allows easy import and management of CSV data, enabling collaborative data preprocessing and analysis for customer churn prediction project, accessible from any device with internet connection.

- *Hardwares*

Windows 10 / Windows 11 OS

VII. RESULT

The classification models were implemented with python 3.12.1, Churn modelling in bank dataset on laptop with intel core i5 processor.

Table 1 Performance Metrics for Classifiers

Parameters	Logistic Regression	Naive Bayes	Random Forest
Accuracy	0.8	0.791	0.876
Recall	0.065	0.085	0.502
ROC	0.519	0.522	0.733

$$TP/(TP + FN) \quad (2)$$

Measure the proportion of actual churn cases that were correctly predicted.

➢ *ROC Curve (Receiver Operating Characteristic)*

Graph of the true positive rate against the false positive rate.

➢ *AUC-ROC (Area under the ROC Curve):*

Compare the overall performance of the matrices based onChurn and non-churn cases.

➢ *Gini Coefficient*

Measures the inequality in the model's performance based oncorrectly predicting churn and non-churn.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

➢ *Entropy*

Entropy can be used to evaluate the impurity of a set of customer data with respect to churn rate.

C. Software and Hardware Setup

➢ *Softwares*

Following Software required for the implementation of Customer Churn Prediction Model.

- *Jupyter*

Jupyter Notebook facilitates customer churn prediction projects by providing an interactive environment for data exploration, visualization, and iterative model development, enhancing collaboration and documentation through code, visualizations, and explanatory text in a single, shareable document.

- *Chrome*

Chrome can leverage Jupyter Notebook to access web-based data sources and seamlessly integrate with Python libraries for data analysis and model building, streamlining the process of customer churn prediction within a familiar browser environment.

Table 2 Performance Metrics For Classifiers

Parameters	K-NN	SVM	XGBoost
Accuracy	0.770	0.788	0.867
Recall	0.083	0.049	0.552
ROC	0.508	0.506	0.746

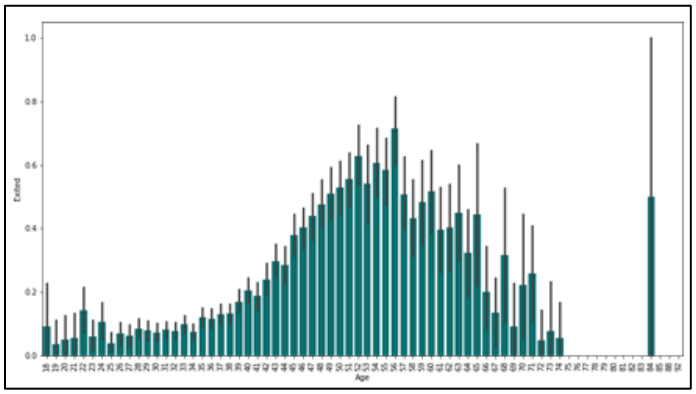

Fig. 2. Graph of Exited Vs Age

The above graphs show us the churn rate with age. Visualizing where age is taken on the x-axis and exit istaken on the y-axis.

VIII. CONCLUSION

Through the implementation of various classifiers including Random Forest, KNN, XGBoost, Decision Tree etc. for customer churn prediction, it was observed that Random Forest gives higher accuracy, particularly after hyper parameter tuning. This indicates the efficacy of Random Forest in handling complex data and optimizing predictive performance, highlighting its suitability for customer churn prediction tasks.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our supervisor, Prof. Sawant S.P., for his invaluable guidance, patience, and support throughout this research. His insightful feedback and constructive criticism have been instrumental in shaping this work. we would also like to thank the participants who generously gave their time and shared their experiences for this study. Finally, we would like to thank my family and friends for their unwavering support and encouragement.

REFERENCES

[1]. Assef Jafar Abdelrahim Kasem Ahmad* and Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. Journal Of Big Data, 2019.
[2]. Siva Bala Krishnan Akshara Santharam. Survey on customer churnprediction techniques IRJET, 2018.
[3]. Rushikesh Chavan Shradha Mandan Amol Chole, Trupti Mate. Customers churn prediction using machine learning and deep learning.JETIR, 2023.
[4]. Dr.K.Geetha. Customer churn prediction. Turkish Journal ofComputer and Mathematics Education, 2022.
[5]. Alaa Mostafa Abdelrahman Fahad Kamal Alsheref Essam Abou elKassem, Shereen Ali Hussein. Customer churn prediction model and identifying features to increase customer retention based on user generated content. IJACSA, 2020.
[6]. Edemealem Desalegn Kingawa Tulu Tilahun Hailu. Customer churn prediction using machine learning techniques: the case of lion insurance. Asian Journal of Basic Science Research, 2022.
[7]. Prateek Anand Kiran Dhangar. A review on customer churn prediction using machine learning approach. IJIERT, 2021.
[8]. N.G.L. Prasanna R.Vindhya K.Sandhya Rani, Shaik Thaslima and P. Srilakshmi. Analysis of customer churn prediction in telecom industry using logistic regression. IJIRCST, 2021.
[9]. Shreyas Rajesh Labhsetwar. Predictive analysis of customer churn in telecom industry using supervised learning. IJSC, 2020.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1246

[10]. Sena KASIM Levent Ç ALLI. Using machine learning algorithms to analyze customer churn in the software as a service (saas) industry. JESS, 2022.
[11]. Dr. S.B. Kishor Madhuri D. Gabhane, Dr. Aslam Suriya. Churn prediction in telecommunication business using cnn and ann. Journal of Positive School Psychology, 2022. [12] Ming-Chang Qian Tong Ming Zhao, Qingjun Zeng and Jiafu Su. A prediction model of customer churn considering customer value: An empirical research of telecomindustry in china. Hindawi, 2021.
[12]. Mukta Sawant Shradhha Rajput Prof. Abhay Gaidhani Mittal Patil, Mahima Rawal. Customer churn prediction. IJCRT, 2021.