Journal of Information Systems Engineering and Management
2025, 10(19s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

TBMOR Customer Churn Predication using XGBoost Classifier

Nikita Khandelwal[1], Dr. Vikas Sakalle[2]
[1]*Research Scholar* , [2]*Associate Professor*
[1,2]*Department of Computer Science and Engineering*
[1,2]*LNCT University Bhopal M.P, India*
nikitakhandelwal0000@gmail.com , vikassakalle@gmail.com

ARTICLE INFO

Received: 24 Dec 2024

Revised: 31 Jan 2025

Accepted: 14 Feb 2025

ABSTRACT

This comprehensive study delves into the efficacy of cutting-edge machine learning models in predicting customer churn, a critical challenge faced by businesses across various sectors including Telecom, Banking, Medical, and Online Retail. By conducting a detailed comparative analysis of several predictive models, the research highlights the outstanding performance of the Proposed XGBoost Classifier. This advanced model consistently outshines its counterparts across a wide array of performance metrics: it achieves an impressive accuracy rate of up to 95%, precision as high as 99%, recall equally remarkable at 99%, F1-score peaking at 98%, and an AUC score of 96%. These figures starkly contrast with the performance metrics of traditional models such as Logistic Regression and Decision Trees, which, while useful for baseline comparisons, exhibit a broad performance spectrum with scores generally oscillating between 70% and 90% across the evaluated metrics. The study's findings reveal a clear superiority of ensemble methods, particularly the XGBoost algorithm, in modeling the complex dynamics and nuanced patterns inherent in customer churn data. This superiority is attributed to XGBoost's robustness in handling diverse data types and its proficiency in capturing intricate interactions within the data, thereby providing a more accurate and nuanced prediction of churn. Moreover, the variability in the performance of traditional models across different datasets underscores the critical importance of model selection and customization according to specific dataset characteristics. It also highlights the necessity of advanced machine learning techniques that can adapt to and efficiently process the unique challenges presented by each dataset.

Keywords: Customer churn, Telecom, Banking, Medical, Online Retail, XGBoost Classifier

1. INTRODUCTION

In the ever-evolving landscape of customer-centric industries, the ability to predict and mitigate customer churn stands as a pivotal challenge [1] that can significantly influence an organization's long-term success and sustainability. Customer churn, or the propensity of customers to cease their business relationship with a service or product, embodies a crucial metric for assessing customer satisfaction and loyalty. As businesses strive to enhance their customer retention strategies, the adoption of advanced machine learning models [2] for churn prediction has emerged as a forefront methodology, offering a nuanced understanding of customer behaviors and the factors driving churn. [3] This research paper presents an in-depth comparative analysis of various advanced machine learning models across multiple industry sectors, namely Telecom, Banking, Medical, and Online Retail, to identify the most effective strategies for predicting customer churn.

The study meticulously evaluates a suite of machine learning models, including traditional algorithms like Logistic Regression and Decision Trees, alongside more sophisticated techniques such as Gradient Boosting Machines (GBMs) and the XGBoost algorithm. By exploring these models' performance across different datasets, the research aims to unveil the strengths and limitations inherent in each approach, focusing on key metrics such as Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC). The comparative analysis provides a comprehensive overview of each model's capability to accurately predict customer churn, thereby offering invaluable insights into the most efficacious predictive models suited for various industry applications.

This introduction sets the stage for a detailed exploration of the comparative results of advanced machine learning models in predicting customer churn. By highlighting the critical importance of model selection and customization based on dataset-specific characteristics, the study endeavors to guide businesses in leveraging the predictive power of advanced algorithms. Through this research, we aim to contribute to the ongoing discourse on enhancing customer retention strategies, ultimately aiding organizations in navigating the complexities of customer churn prediction with greater precision and effectiveness.

2. LITERATURE REVIEW

Wagh et al. (2024) A huge client base generates massive amounts of data everyday in the telecom business. Getting new clients is more expensive than keeping existing ones, since churn is the act of consumers moving firms within a certain timeframe. Telecom managers and analysts are investigating why consumers cancel contracts and how holding companies use customer data. This method employs categorization to locate telecom consumers' quit subscriptions and reasons for them. This system analyses various machine learning methods to generate customer churn prediction models and identify churn causes to provide retention strategies and plans. Leave subscriptions captures customer data using Random Forest (RF), KNN, and decision tree Classifier. It provides an effective business model that analyses customer churn data and predicts churn customers so company management can respond quickly to reduce churn and financial loss. System achieves 99 % accuracy using random forest classifier for churn predictions, with classifier matrix precision, recall factor, and total accuracy of 99.09 %. Our study enhances churn prediction, expands to various business domains, and provides prediction models to retain customers and prevent churn. [1]

Wanikar et al. (2024) This research seeks a reliable communications sector customer churn prediction model. We use machine learning and analysis to increase customer churn forecast accuracy and efficiency. Bundling and bracing are used in our study to integrate model predictions and decrease variability. We also conduct awareness tests to determine how elements affect the presenting model. We calibrate the model using learning rate, layer count, and group size to determine the optimum way to analyse the disaster. We combine large quantities of data to expand the analysis of customer rivalry projections in the correspondence industry. We anticipate client-agitate forecast models to become more versatile and successful via thorough information analysis. Different AI computations were used to analyse this study's results. AI and substantial data are used to predict customer beat in the communication company. Research shows this approach accurately predicts customer agitation [2].

Usman-Hamza et al. (2024) Customer turnover is a major issue for most businesses, especially telecoms. Given the severe rivalry among telecommunications businesses and the expensive cost of acquiring new consumers, maintaining loyal subscribers is vital. Telecommunications organisations may use early dissatisfaction prediction to identify churn causes and implement creative strategies to enhance productivity, preserve market competitiveness, and avoid financial losses. To achieve this aim, effective and reliable customer churn prediction (CCP) technologies must be developed. Current CCP research suggests rule-based and machine-learning (ML) techniques to address the problem. However, rule-based CCP systems lack flexibility and robustness, and the unbalanced distribution of churn datasets hinders most classical ML approaches in CCP. ML-based CCP solutions are more effective than others. Tree-based ML classifiers provide prediction models with excellent accuracy, stability, and interpretability, unlike linear, instance, and function-based classifiers. Tree-based classifiers for CCP are usually confined to decision trees (DT) and random forests. This study examined the performance of tree-based classifiers with different computational features in CCP. The CCP performance of single, ensemble, improved, and hybrid tree-based classifiers is examined. Tree-based classifiers and their homogenous ensemble variations on CCP were also tested for prediction performance in the presence of data quality issues including the class imbalance problem (CIP). Experimental results show that tree-based classifiers outperform linear-based, instance-based, Bayesian-based, and function-based classifiers, both with and without CIP. The CIP has a major influence on the CCP performances of explored tree-based classifiers, yet a data sampling methodology and a homogeneous ensemble method can solve CIP and build efficient CCP models.[3]

Lee et al. (2023) This research presents a hybrid statistic-machine learning methodology to predict client attrition. The suggested technique dynamically determines the churn line using the statistical model's chance of customer survival, unlike existing approaches. After studying customer turnover via clustering over time, the suggested technique divided consumers into four behaviours: new, short-term, high-value, and churn, and chose machine learning models to forecast them. These factors lower the risk of misjudging attrition for consumers with longer consumption cycles. The hybrid technique was tested on two public datasets, U.K. online retail. gift vendors and Pakistan's leading E-Commerce. In the top three learning models, recall varied from 0.56 to 0.72 in the former and 0.91 to 0.95 in the latter. By forecasting customer attrition, the suggested technique helps firms retain key clients early. The hybrid technique uses less data than others. [4]

Yigit et al. (2024) Companies must employ analytical tools and revolutionary technology to succeed globally. Wilo, a major premium pump and pump system supplier, assigned Holger Jentsch, VP of Group Sales Excellence, to pilot first technology transformation initiatives in particular sales processes. Jentsch's lighthouse project was to avoid client turnover using AI-based insights. Jentsch's digital transformation effort relies on databased decision-making, therefore the case discusses key success elements. [5]

Sam et al. (2024) Telecommunications companies worry about customer attrition. Understanding and anticipating client turnover may improve retention and profitability. Telecommunication firms may discover unsatisfied customers early and take steps to keep them by predicting customer turnover. The telecom business creates a lot of data everyday due to its enormous customer base. Business experts and decision makers underlined that getting new consumers costs more than keeping current ones. Business analysts and CRM analysts must analyse churn data to determine customer churn causes and behaviour patterns. This research introduces a churn prediction model that employs classication and clustering to detect churn consumers and identify telecom customer churn reasons. XBoost and Random Forest outperformed K-Nearest Neighbours, Support Vector Machines, and Decision Trees in accuracy, precision, F1-Score, and recall. [6]

Singh et al. (2024) Customer attrition in the banking business happens when customers stop utilising the bank's products and services and then leave. Therefore, client retention is crucial in today's very competitive banking business. A strong customer base builds trust and referrals, attracting new customers. These considerations make lowering customer attrition essential for banks. We analyse bank data to predict which consumers will stop utilising the bank's services and become paying clients. Data is analysed using machine learning techniques and compared using multiple assessment measures. We have created a Data Visualisation RShiny tool for customer churn analysis in data science and management. This data will help the bank identify trends and retain at-risk clients. [7]

Kumar et al. (2021) Predicting client attrition is one way to keep high-value customers and improve sales chances. Service providers may avoid this turnover by identifying its reason via data analysis. However, service providers must have enough time to communicate with clients and respond positively to keep them following forecast. Thus, early churn candidate discovery is crucial for such applications. We believe that leading indicator data sources may provide this extra time to engage. In this research, we model sentiment data and socio-economic data as leading indicator sources of temporal information important to customers. We also examine the value of employing such data sources to fulfil the need for a longer time horizon to respond to customer churn prediction applications. Experimental findings from open datasets are presented to test our theory. Our analysis reveals that consumer sentiment and socio-economic factors significantly enhance churn prediction accuracy by 20% over traditional methods (P-value < 0.05). [8]

Yu et al. (2024) Decision makers need customer churn forecast to retain customers and manage company. This article reviews current concerns and challenges in studying telecommunication customer turnover to assist academics understand the key aspects driving prediction model performance. Based on 2017–2021 customer churn prediction research, two research questions were proposed: (i) what issues and challenges do researchers face in the territory of churn prediction model (CPM); and (ii) what problems in CPM are addressed by most relevant studies. This problem-based literature analysis shows that most CPM researchers struggle with data sparsity, feature selection with bias, unbalanced class distribution, and evaluation metrics. It also illustrates that inefficient models have complicated calculation, high computation time, and bad prediction outcomes. CPM is mentioned in this publication, but just for telecommunication researchers, which is a constraint. Future study will include a thorough examination of data mining and technique in many business sectors. [9]

Chinnaraj (2023) Churn prevention has long been a corporate retention goal. The telecoms business faced high customer turnover owing to crowded markets, fierce competition, changing criteria, and new alluring offers. This paper advances the area by classifying the telecom industry's churn forecast issue. A customer churn prediction (CP) model is required to monitor it. The reconstructed recurrent neural network and Elephant herding optimisation (EHO) approach are used to provide a unique framework for customer turnover forecasting. EHO is a meta-heuristic optimisation method inspired by elephant herding. EHO uses a clan operator to adjust the distance between clan elephants and a matriarch elephant. For several benchmark problems and application areas, the EHO methodology outperforms various cutting-edge meta-heuristic approaches. RRNN is updated to classify Churn Customers (CC) and regular customers. The RNN parameters are optimised by this upgraded EHO. Network utilisation is considered a retention technique if a customer leaves. This paradigm ignores the amount of users that depart depending on local network usage. The simulation and performance metrics-based comparison reveal that the novel strategy can detect churn better than relevant methods. [10]

Szeląg & Słowiński (2023) A computer experiment using bank customer attrition data shows monotonic decision rules' explanatory and predictive power. The data are partly ordinal since certain customer attribute value sets are ordered and have a monotonic connection with churn or non-churn outcomes. The Variable Consistency Dominance-based Rough Set Approach (VC-DRSA) structures the data before monotonic decision rules are inducted. The RuLeStudio and RuleVisualization programmes use an enhanced version of VC-DRSA for supervised learning. The first experiments with parameterized rule models, while the second visualises and analyses them. The monotonic decision rules reveal bank data, identifying loyal and departing consumers. Explainable AI seeks a transparent decision model that decision-makers can understand. We also compare monotonic rules' prediction power to popular machine learning methods. [11]

Liu et al. (2024) Customer retention strategies increasingly use predictive analytics to identify churners. Previous customer churn prediction relied on binary classifiers optimised for accuracy-based performance criteria. There is a growing understanding that this technique may not always maximise profit since it ignores the costs of misclassification and the advantages of precise categorization. This research predicts profit-driven customer attrition using high gradient boosting trees. This trees' class weights and other hyperparameters are optimised using Bayesian approaches based on profit maximisation. Real service provider datasets from numerous marketplaces are used for empirical analysis. The empirical findings show that the suggested model outperforms benchmark models. Bayesian optimisation and class weight adjustments increased model profitability. Compared to exhaustive grid search, model optimisation with several hyperparameters is computationally cheaper. A Bayesian optimization-based sensitivity analysis shows the model's resilience. Marketing managers may create focused marketing campaigns to retain high-churning consumer segments using the suggested approach. [12]

Bhale & Bedi (2023) Background: As pricing becomes more ubiquitous in the digital world, marketing researchers focused on consumer attrition. Customer churn research helps firms create durable competitive advantages and enhance financial results. According to Gartner (2012), a 5% churn decrease may boost a company's profitability by 25%, therefore understanding and controlling customer attrition is crucial. Objective: This research will provide the latest Scopus updates on customer churn articles. Methodology: Scopus's “Title, Abstract, and Keywords” search option was used for the bibliometric study of “customer churn”. According to the survey, customer turnover has become a prominent research topic in the recent decade, with 1,305 publications published by 2020. Customer churn is a global phenomena that has garnered the attention of specialists worldwide in the last decade, with exponential growth in publications from 2017 to 2020. The survey found that “business management” tops customer churn research. Academics may reference the research to examine consumer attrition. [13]

Baby et al. (2023) Customer churn increases revenue and customer loyalty loss in banking. The study uses Artificial Neural Networks (ANNs) to anticipate client attrition in the banking industry. Input characteristics and the learned dataset independent variable dictate the prediction. The forward propagation method and cross-validation approaches change hyperparameters during model training to improve accuracy and precision. Results show that the model predicts customer attrition 86% accurately. ANN methods predict banking client attrition better than logistic regression models. The report offers crucial insights about using machine learning to retain and improve customer retention. This technique helps banks identify churning customers and retain them. [14]

Zdziebko et al. (2024) Churn is a problem for the telecommunications business since acquiring new customers is more expensive than retaining current ones. Telecom retention departments concentrate on client loyalty and churn reduction. Identifying lucrative customers with the greatest churn risk helps guide antichurn efforts. Data about churners is typically erroneous and unclear. This research presents a fuzzy technique to modelling mobile telecoms churn intent using consumption statistics. It detects data ambiguity and illuminates churn modelling. The research tested the Mamdani and Sugeno models for developing a churn model using a small yet feature-engineered real-world dataset. Identifying churn modelling traits was another objective. Four metrics—accuracy, recall, precision, and F1-score—estimated model performance. The developed fuzzy rule-based systems show that to generalise possible churn identification factors with fuzzy rules, start with features like the change in the total amount of the invoice in the last period before the churning compared to the previous one, the total amount of the invoice in the period preceding the churning, the total amount of subscription in two months before the churning, the time of cooperation with the operator, and the number. [15]

Kumar & Logofatu (2023) Because recruiting new customers is more expensive than maintaining current ones, telecommunications, banking, insurance, and e-commerce companies must estimate client turnover. Customer churn prediction (CCP) helps firms retain customers by identifying those most likely to leave. Machine learning methods that use consumer data to uncover churn predictions are promising. A comparison of the most common supervised machine learning algorithms, including Logistic Regression, Decision tree, and Ensemble techniques including Bagging, Boosting, Stacking, and Voting, predicted telecoms customer attrition. The dataset is biassed towards non-churners, thus we examined SMOTE and SMOTEENN sampling algorithms to balance it. Our research found that machine learning can forecast client attrition. Our findings suggest that ensemble learners outperform single-base learners, and a balanced training dataset should increase classifier performance. [16]

Mishra et al. (2024) Customer churn prediction is a common challenge in most industries. ML and AI advancements have improved customer attrition prediction. This article includes most customer churn prediction methods. We examined many famous author articles. We tried to chronicle all the machine learning and deep learning methods produced and used by the world's technology giants to better understand their customers and increase their market. The DL model beat the competition in classification and prediction. Thus, DL models would ignore such irrelevant data while building data blueprints. This review article describes methods to anticipate firm metadata churn using Deep Learning, ML, and senti churn. The attrition prediction model considers context, use, customer, and support factors. Historical datasets taught deep learning and machine learning modelling. [17]

Rao et al. (2024) Banking has both possibilities and difficulties as big data and AI grow rapidly. Improving a model's capacity to categorise unbalanced datasets is a major difficulty in customer churn prediction. To handle unbalanced customer categorization, this study proposes a novel multi-strategy collaborative processing technique, IADASYN-FLCatBoost, from data and algorithm perspectives. At the data level, the traditional Adaptive Synthetic (ADASYN) sampling is improved by using the LOF (Local Outlier Factor) algorithm to eliminate outliers and processing the classification features to synthesise new minority class samples. This results in a "IADASYN" algorithm. Focal Loss is incorporated into the CatBoost ensemble learning architecture to provide a focal-aware, cost-sensitive unbalanced customer churn prediction system, FLCatBoost. The empirical study also uses the Kaggle credit card customer dataset. The staged comparison trials suggest that this paper's technique IADASYN-FLCatBoost predicts best. The proposed method outperforms 5 imbalanced classification algorithms and 20 classifiers made up of classical sampling methods and ensemble learning algorithms in classification effect, Recall, F1 score, G-mean, and Area under Precision-Recall curve (AUPRC). Further testing of the model shows that the suggested strategy is applicable to additional banks and customer churn datasets from different sectors. [18]

Karthikeya & Neerugatti (2024) Modern organisations must consider their massive client base. Their enterprises only serve the people. Companies must track how their products reach customers. And to understand customer behaviour around the product or service. This crucial information will reveal client interests related to the product or service. Companies must monitor consumer departures. Churn Rate is crucial to corporate growth or decline. It will be spectacular to predict churning clients before they do. Companies may now better target and deal with expected departing clients. This research proposes a Neural Network-based Ensemble Learning Model. Three Neural Network models will forecast, and the most frequent will be the final prediction. The proposed model outperformed individual model with 84% accuracy. [19]

Ahmad et al. (2023) Today's firms utilise targeted marketing to gain and retain customers. Google and Facebook have built their businesses on tailored ads that boost growth. Customer personality helps organisations forecast turnover. This happens in many firms when customers quit for various reasons. Due to this gap, consumer personality analysis is conducted. High imbalances were found in the dataset. CTGAN and SMOTE have been used to balance lessons. Bagging employs Random Forest (RF), boosting uses XGBoost (XGB), Light Gradient Boosting Machine (LGBM), and ADA Boost. The suggested Hybrid Model HSLR uses RF, XGB, ADA Boost, LGBM, and LR as base and meta classifiers. Three separate k-folds with 5 and 10 folds were tested. To measure classifier performance, Accuracy, Precision, Recall, F1, MCC, and ROC scores are used. Comparing SMOTE with CTGAN data returns results. The SMOTE technique had the greatest accuracy, precision, recall, F1, MCC, and Roc scores of 94.06, 94.23, 94.28, 94.05, 88.13, and 0.984. [20]

Lavanya et al. (2023) All significant firms must confront consumer loss. Companies, notably telecoms, are attempting to anticipate customer turnover since it affects revenue. Therefore, to reduce client turnover, you must understand its reasons. Customers leave a firm due to causes including new products from competitors or service problems. These factors typically lead consumers to cancel. Customer churn predictive modelling analyses previous, present, and demographic data to forecast customer defection. Customer churn prediction is a well-studied data mining and machine learning issue. Classification algorithms are often used to examine churners and non-churners. Current state-of-the-art classification algorithms do not account for financial costs and benefits during training and assessment, therefore they do not correlate with commercial aims. Over time, cost-sensitive learning (CSL) approaches for data learning have been developed since misclassification mistakes have varying prices. CSL versions of machine learning algorithms for Telecom Customer Churn Predictive Model are shown here. Also used feature selection techniques and CSL in UCI's real-time telecom dataset. The suggested CSL-ML combination surpasses state-of-the-art machine learning algorithms in prediction accuracy, precision, sensitivity, ROC curve area, and F1-score. [21]

Phumchusri & Amornvetchayakul (2024) SaaS is a subscription-based software-licensing mechanism that uses external servers. The essay provides customer attrition prediction methods for a Thai SaaS inventory management firm. This work seeks the best customer churn prediction model for a case-study SaaS inventory management firm with a high churn rate. This article presents logistic regression, support vector machine, decision tree (DT), and random forest machine learning techniques. For recall scorer, the optimised DT model outperforms previous classification algorithms with verified testing scores of 94.4% recall and 88.2% F1-score. Additionally, feature significance ratings are examined for practical insights about churn behavior-related characteristics. Thus, the data may assist the case-study organisation better identify churning consumers and improve management and marketing choices. [22]

Soleiman-garmabaki & Rezvani (2023) Companies utilise proactive measures to identify client turnover in addition to reactive ones. Getting new clients is usually more expensive than keeping them. A literature study demonstrates that machine learning predicts customer attrition most often. This research studies telecom customer turnover factors. Data mining classification techniques such neural networks, K-nearest neighbour, SVM, logistic regression, decision tree, and random forest are used. Accuracy, precision, recall, F1-score, and ROC curve are used to analyse outcomes. This study analyses how data balance, acceleration techniques, and neural networks affect each other. This study presents a hybrid classifier speed-accuracy trade-off for real-world applications, which is its main contribution. It compares classifier performance before and after data balancing. After finding effective classifiers, we use AdaBoost and XGBoost. Based on all assessment criteria, the best combinations are found. Our findings suggest that a hybrid classifier using AdaBoost and XGBoost enhances performance considerably. This study presents a more accurate combination categorization than current approaches. [23]

3. METHODOLOGY

3.1 Proposed working flowchart

Figure 1 depicts a flowchart that outlines the process of developing a predictive model using the XGBoost algorithm. The process starts with data collection, after which the data is split into training and testing datasets. The training dataset undergoes several preprocessing steps, including data cleaning to address missing values or outliers, feature engineering to derive new predictors that could improve model performance, and data transformation to scale or normalize the data and encode categorical variables. Once preprocessed, the data is fed into the XGBoost architecture to train the model.

After the model has been trained, it is then evaluated using the testing dataset. The model evaluation is based on several performance metrics: accuracy, which measures the overall correctness of the model; recall, which assesses the model's ability to detect positive instances; precision, which evaluates the model's accuracy in predicting positive instances; and the F1-score, which provides a balance between precision and recall. These metrics together give a comprehensive understanding of the model's performance and its ability to predict customer churn.

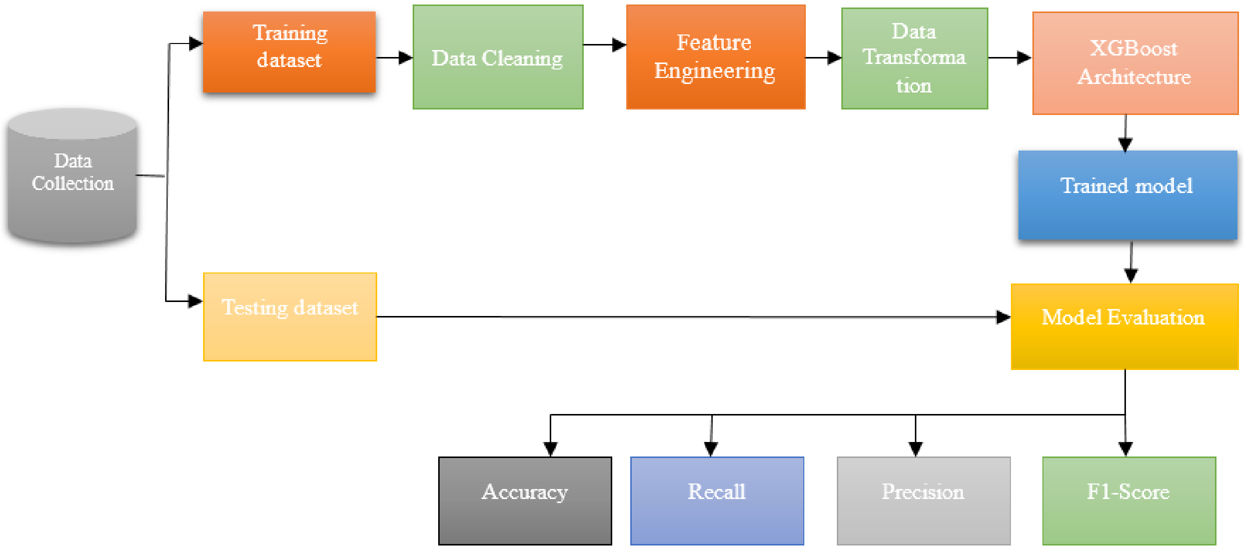

Figure 1. Proposed working flowchart

3.2 Methods for study

Logistic Regression (Baseline) : Logistic Regression serves as a foundational model for binary classification tasks. It predicts the probability that a given data point belongs to a particular category based on a logistic function. The baseline version of logistic regression typically does not incorporate any data balancing or feature selection techniques. It's straightforward, relying on the raw features (denoted as cols here) to estimate the odds of an event occurring, making it an excellent starting point for binary classification problems.

Logistic Regression (SMOTE) : Logistic Regression with SMOTE (Synthetic Minority Over-sampling Technique) applies the logistic regression algorithm after balancing the dataset. SMOTE generates synthetic samples for the minority class to balance the dataset, addressing class imbalance problems. This method uses the same features (cols) but preprocesses the data to make the classes equally represented, potentially improving model performance on imbalanced datasets.

Logistic Regression (RFE) : RFE (Recursive Feature Elimination) with Logistic Regression involves using logistic regression in conjunction with RFE for feature selection. RFE works by recursively removing the least important features (based on coefficients or feature importance scores) and rebuilding the model. This variant uses a subset of features (cols_rfe) chosen through RFE, aiming to improve model performance by focusing on the most informative variables.

Decision Tree : Decision Tree classifiers use a tree-like model of decisions where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Decision trees are intuitive and easy to interpret, making them popular for a variety of tasks. They use raw features (cols) to make classifications, potentially becoming complex as they aim to perfectly classify the training data.

KNN Classifier : The K-Nearest Neighbors (KNN) Classifier classifies data points based on the majority vote of their k nearest neighbors. It's a type of instance-based learning where the function is only approximated locally, and all computation is deferred until function evaluation. KNN is straightforward but can be computationally expensive, as it uses all available cases (cols) and relies on distance calculations between samples.

Random Forest : Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It offers improvements over single decision trees by reducing overfitting and increasing predictive accuracy, using the same set of features (cols).

Naive Bayes : Naive Bayes classifiers apply Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the class label. Despite its simplicity and the strong independence assumption, Naive Bayes can perform well in many real-world situations, especially for text classification and problems with high-dimensional data, using features (cols).

SVM (Linear and RBF) : Support Vector Machine (SVM) classifiers construct a hyperplane or set of hyperplanes in a high-dimensional space to classify data points. Linear SVM (SVM (linear)) uses a linear kernel to find the decision boundary, while SVM with Radial Basis Function (SVM (rbf)) kernel uses an RBF kernel to handle non-linearly separable data, both utilizing the same features (cols).

LGBM Classifier : Light Gradient Boosting Machine (LGBM) Classifier is a gradient boosting framework that uses tree-based learning algorithms and is designed for speed and efficiency. It grows trees leaf-wise (best-first) rather than level-wise, achieving better accuracy with lower memory usage, using features (cols).

XGBoost Classifier : XGBoost (Extreme Gradient Boosting) Classifier is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework, offering advanced regularization to prevent overfitting, using features (cols).

Gaussian Process : Gaussian Process Classifier (GPC) applies the Bayesian classification approach with Gaussian processes, providing a probabilistic prediction. It's particularly useful for estimating uncertainty and performing well on smaller datasets, using features (cols).

AdaBoost : AdaBoost (Adaptive Boosting) focuses on converting weak learners into strong ones by adaptively changing the distribution of training data based on the performance of previous models, emphasizing more on the incorrectly classified instances. It's used with decision trees as base learners, using features (cols).

Gradient Boost : Gradient Boost classifiers build an additive model in a forward stage-wise fashion, allowing for the optimization of arbitrary differentiable loss functions. Each new tree corrects errors made by previously trained trees, using features (cols).

LDA and QDA : Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are classifiers with a linear and quadratic decision surface, respectively. They assume that different classes generate data based on different Gaussian distributions, using features (cols).

MLP Classifier : Multi-Layer Perceptron (MLP) Classifier is a deep, artificial neural network that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. MLP utilizes backpropagation for training the network, handling complex patterns in data, using features (cols).

Bagging Classifier : Bagging (Bootstrap Aggregating) Classifier improves the stability and accuracy of machine learning algorithms, particularly decision trees, by constructing multiple versions of a predictor and using their average to make predictions. It reduces variance and helps to avoid overfitting, using features (cols).

3.3 Preprocessing Steps

1. Data Cleaning:
    - Missing Values: Identify and impute missing values. The strategy can vary from mean/mode imputation, forward/backward filling, to predictive modeling methods, depending on the nature of the data.
    - Outliers: Detect and handle outliers either by removing, capping, or transforming them, as outliers can skew the results of the predictive model.
2. Feature Engineering:
    - Feature Creation: Derive new features that can capture customer behavior more effectively. For example, aggregating transaction data to create features like average transaction value, frequency of transactions, etc.
    - Dimensionality Reduction: Apply techniques like PCA (Principal Component Analysis) to reduce the number of variables and focus on the most informative features, especially important in high-dimensional data like medical records or online retail browsing data.
3. Data Transformation:

  - Normalization/Standardization: Scale numerical data to reduce the influence of vastly different ranges on the prediction model. Standardization (z-score normalization) or Min-Max scaling are common approaches.
  - Encoding Categorical Variables: Convert categorical variables into a form that can be provided to ML algorithms via methods like one-hot encoding, label encoding, or using embedding layers for deep learning models.

Industry-Specific Considerations

- Telecommunications:
  - Focus on usage patterns, contract details (e.g., contract duration, type of plan), and service-related features (e.g., number of service calls).
  - Analyze call detail records (CDRs) to extract behavior patterns such as peak usage times, call failures, etc.
- Banking:
  - Key features might include account balance, product holdings, transaction history, and interaction with digital channels.
  - Customer segmentation based on financial behavior could be useful in tailoring the churn prediction models.
- Online Retail:
  - Session data analysis is crucial, including page views, time spent on site, cart abandonment rate, and purchase history.
  - Customer feedback and product return data can also be significant predictors of churn.
- Medical Industry:
  - Patient records need careful handling, ensuring compliance with data protection regulations (e.g., HIPAA in the USA).
  - Features could include treatment history, appointment no-shows, medication adherence, and outcomes of previous treatments.

3.3 Proposed algorithm

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm widely used in classification tasks, including customer churn prediction across various industries such as telecommunications, banking, online retail, and the medical industry. Its architecture and methodology can be tailored to these specific domains for optimal performance. Here's a step-wise, detailed description of the XGBoost architecture, particularly in the context of customer churn prediction:

Step 1: Data Preprocessing

- Telecommunications & Banking: Aggregate user data such as call logs, usage patterns, account balances, and transaction history. Feature engineering is crucial to capture behaviors indicative of churn, such as a decrease in usage or transactions.
- Online Retail: Analyze customer interaction data, including purchase history, product views, and cart abandonment rates. Time-series features representing changes in buying behavior over time can be particularly predictive of churn.
- Medical Industry: Collect patient data, including treatment history, appointment attendance, and health outcomes. Data must be anonymized and processed in compliance with healthcare regulations like HIPAA in the US.

Step 2: Feature Engineering

- All Industries: Derive meaningful features that can predict churn, such as customer lifetime value, frequency of service usage, changes in transaction amounts, and patterns of service or product engagement. The goal is to identify variables that signal a customer's likelihood to churn.

Step 3: Handling Imbalanced Data

- Customer churn datasets are often imbalanced, with fewer instances of churn than non-churn. Techniques like SMOTE and XGBoost's built-in weight balancing parameter (scale_pos_weight) can be used to address this imbalance, improving model sensitivity to churn cases.

Step 4: Model Training with XGBoost

- Objective Function: XGBoost uses gradient boosting framework for training decision trees in a sequential manner, where each new tree corrects errors made by previously trained trees. The objective function typically combines a loss term that measures prediction accuracy and a regularization term that controls model complexity to prevent overfitting.
- Hyperparameter Tuning: Adjust XGBoost parameters such as learning rate (eta), depth of trees (max_depth), and the number of trees (n_estimators) to find the optimal configuration. Cross-validation can be used to evaluate model performance and prevent overfitting.
- Feature Importance Analysis: After training, analyze feature importance scores generated by XGBoost to identify the most predictive features of churn. This insight can inform business strategies to reduce churn.

Step 5: Model Evaluation

- Evaluate the model using metrics such as accuracy, precision, recall, F1 score. In churn prediction, recall (sensitivity) might be particularly important to identify as many true churn cases as possible.

3.4 XGBoost working

1. Objective Function

XGBoost's objective function is a combination of a loss function and a regularization term. The loss function ensures the model fits the data well, and the regularization term controls the model's complexity to prevent overfitting. For binary classification tasks like churn prediction, the loss function is often logistic loss.

2. Gradient Boosting

At its core, XGBoost is a type of gradient boosting framework. It builds an ensemble of decision trees in a sequential manner, where each subsequent tree aims to correct the errors of the previous ones. It does this by fitting the new tree to the gradients (i.e., the first derivatives) of the loss function with respect to the predictions.

3. Tree Ensemble Model

XGBoost uses an ensemble of K additive functions to predict the output. Each function is a decision tree:

$$\boldsymbol{y_i = \sum_{k=1}^{k} f_k(x_i), f_k \ \epsilon \ \ F}$$

where F is the space of regression trees. For a classification problem, the final prediction is made by passing the output through a logistic function to get probabilities.

4. Regularized Learning

XGBoost improves upon the traditional gradient boosting method by introducing a regularization term in its objective function, which includes both L1 (lasso regression) and L2 (ridge regression) regularization:

$$obj(\Theta) = L(\Theta) + \lambda \sum_{K=1}^{K} \omega_k^2 \ + \ \alpha \sum_{k=1}^{K} |\omega_k|$$

This regularization helps in reducing overfitting, which is particularly important for customer churn prediction, where the model needs to generalize well to unseen data.

5. Tree Pruning

XGBoost uses a depth-first approach to grow trees and prunes them using the 'max_depth' parameter. Unlike other gradient boosting algorithms that stop splitting a node once it encounters a negative loss, XGBoost looks ahead to see if a split will lead to positive gains in the future, allowing for more complex models that can learn finer patterns in the data.

6. Handling Missing Values

XGBoost has a built-in routine to handle missing values. When the model encounters a missing value on a feature, it learns the direction in which to send the missing values at training time, which can be either to the right or left child of a node.

7. Built-in Cross-Validation

XGBoost allows for built-in cross-validation at each iteration of the boosting process. This enables the user to get the most accurate model by stopping the build process when the cross-validation score starts to deteriorate.

8. Learning Rate and Number of Trees

The learning rate (or "eta") shrinks the feature weights after each boosting step to make the model more robust. The number of boosting rounds (trees) can be specified by the user or determined via early stopping.

9. Parallel and Distributed Computing

XGBoost can run on a single machine as well as a distributed environment. It parallelizes the tree construction using all of your CPU cores during the training phase.

10. Column Block for Parallel Learning

XGBoost stores and sorts data by columns. This allows for more efficient access of data and distribution to optimize for the hardware architecture.

11. Model Evaluation and Prediction

Once the model is trained, it evaluates its performance on a testing dataset using various metrics like accuracy, precision, recall, and the area under the ROC curve (AUC). The model can make predictions by inputting features of new customer data into the ensemble of decision trees, which collectively determine whether a customer is likely to churn.

In customer churn prediction, these powerful features of XGBoost can lead to a highly accurate model capable of identifying customers at risk of churning, thus allowing businesses to take proactive retention measures.

3.5 The advantage of the proposed method concerning other methods

Method	Advantage of XGBoost Over the Method
Logistic (SMOTE)	Handles imbalanced data inherently, often outperforming oversampling methods.
Logistic (RFE)	Automatically performs feature selection during training.
Decision Tree	Less prone to overfitting, more robust, and provides better generalization.
KNN Classifier	Does not require proximity calculation, thus is more scalable.
Random Forest	Offers improved bias-variance trade-off through boosting.
Naive Bayes	Can handle noise and missing data better than the probability-based approach.
SVM (linear)	Provides probabilistic interpretations and handles non-linear boundaries.
SVM (rbf)	More efficient with high-dimensional data, avoiding extensive parameter tuning.
LGBM Classifier	Typically faster with larger datasets and provides feature importance scores.
Gaussian Process	XGBoost is more scalable to large datasets and less computationally intensive.
AdaBoost	Faster convergence and better handling of noisy data and outliers.
GradientBoost	Often outperforms due to more sophisticated regularization techniques.

LDA	Can capture complex relationships that LDA might miss due to linear assumptions.
QDA	Does not assume equal covariance like QDA, more flexible in capturing variance.
MLP Classifier	Avoids getting stuck in local minima by using tree structure instead of neural nets.
Bagging Classifier	Improves upon bagging by using boosting, leading to better predictive performance.

4. IMPLEMENTATION AND RESULT

4.1 System requirements

4.1.1 Essential Pieces of Hardware:

- Enough processing resources (CPU, RAM, storage) to deal with the dataset's extensiveness and the model's complexity.
- To train complicated models quickly and effectively, you may need a powerful central processing unit (CPU) or graphics processing unit (GPU), depending on the amount of the dataset.

4.1.2 Specifications for Required Software:

- System Operation: You can utilise any of the most common operating systems, including Windows, macOS, or Linux.
- Python is the programming language most often used for implementing machine learning models owing to its user-friendliness and extensive ecosystem of libraries (such as sci-kit-learn, TensorFlow, and PyTorch).
- Integrated construction Environment (IDE): PyCharm, Jupyter Notebook, or Anaconda are some recommended IDEs that provide a user-friendly coding environment for constructing models.

4.2 Dataset Description

4.2.1 Telecom Customer Churn Dataset:

Dataset: "Predicting Customer Churn in Telecoms Dataset" from Kaggle

Description: This dataset contains telecom customer information, including demographics, call details, and churn status.

Reference: https://www.kaggle.com/becksddf/churn-in-telecoms-dataset

4.2.2 Banking Customer Churn Dataset:

Dataset: "Bank Customer Churn" from Kaggle

Description: This dataset consists of customer data from a bank, including demographics, account information, and churn status.

Reference: https://www.kaggle.com/santoshd3/bank-customers

4.2.3 Medical Customer Churn Dataset:

Dataset: "Predicting Patient Churn in **Life Insurance Industry**" from Kaggle

Description: A dataset for customer churn prediction in the life insurane industry.

Reference: https://www.kaggle.com/datasets/usmanfarid/customer-churn-dataset-for-life-insurance-industry

4.2.4 Online Retail Customer Churn Dataset:

Dataset: "Online Shopper's Intention" from UCI Machine Learning Repository

Description: This dataset contains online shopper data, including E Comm CustomerID Unique customer ID, Comm Churn Churn Flag, E Comm Tenure Tenure of customer in organization etc.

Reference: https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

4.3 Illustrative result

Figure 2 is a threshold plot for an XGBoost Classifier. This type of plot helps in selecting the optimal threshold for binary classification tasks. The plot typically displays how precision, recall, the F1 score, and the queue rate (proportion of positive predictions) vary as the discrimination threshold changes.

In the plot:

- The blue line represents precision, which indicates the proportion of true positives among all predicted positives. High precision relates to a low false positive rate.
- The red line shows recall, also known as sensitivity, which is the proportion of true positives that have been correctly identified over all actual positives. High recall means fewer false negatives.
- The green line represents the F1 score, a harmonic mean of precision and recall, providing a single score that balances the two metrics.
- The purple line indicates the queue rate, or the rate at which instances are predicted as positive.

A dashed vertical line (usually in black) indicates a specific threshold (t = 0.37 in this case), which may be considered an optimal balance point between precision and recall. This threshold is where the trade-off between false positives and false negatives might be most acceptable for a specific application.

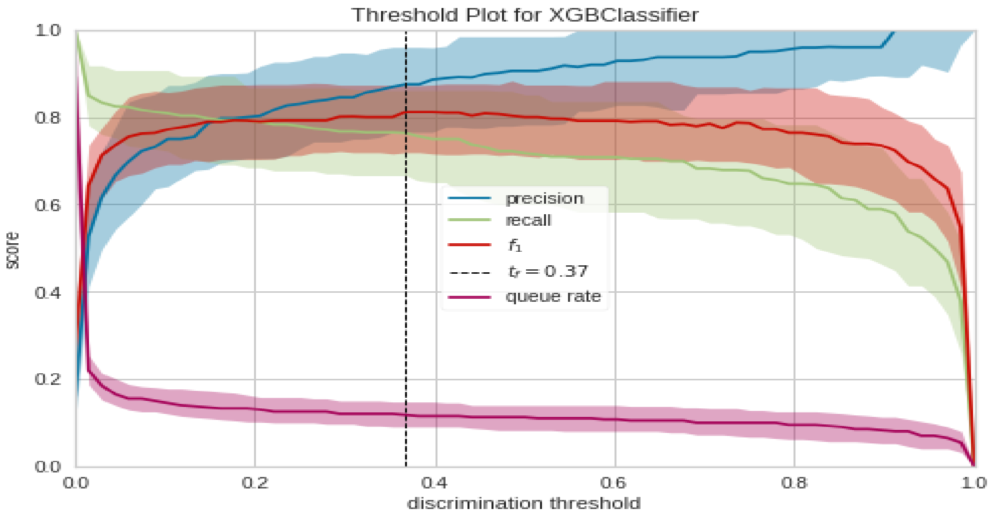

Figure 2. Threshold plot for an XGBoost Classifier

4.4 Confusion matrices for models

The figure 3 provided confusion matrix graphics for various machine learning models applied to a dataset reveal insights into each model's performance in predicting churn. Models like Logistic Regression with SMOTE and RFE modifications, SVM with RBF kernel, and Naive Bayes show a balanced trade-off between true positives and false negatives, suggesting a moderate sensitivity to churn prediction. The KNN Classifier, on the other hand, indicates a higher propensity for false negatives, potentially leading to underestimation of churn risk. Notably, the XGBoost Classifier and LGBM Classifier exhibit a high number of true positives with fewer false negatives, highlighting their superior performance in accurately identifying churn. The Gradient Boost and Bagging Classifier also perform well, albeit with a slightly higher number of false negatives. These visualized results underscore the importance of selecting and tuning the right machine learning model to effectively manage and anticipate customer churn, with ensemble methods like XGBoost and LGBM emerging as particularly powerful tools for this purpose.

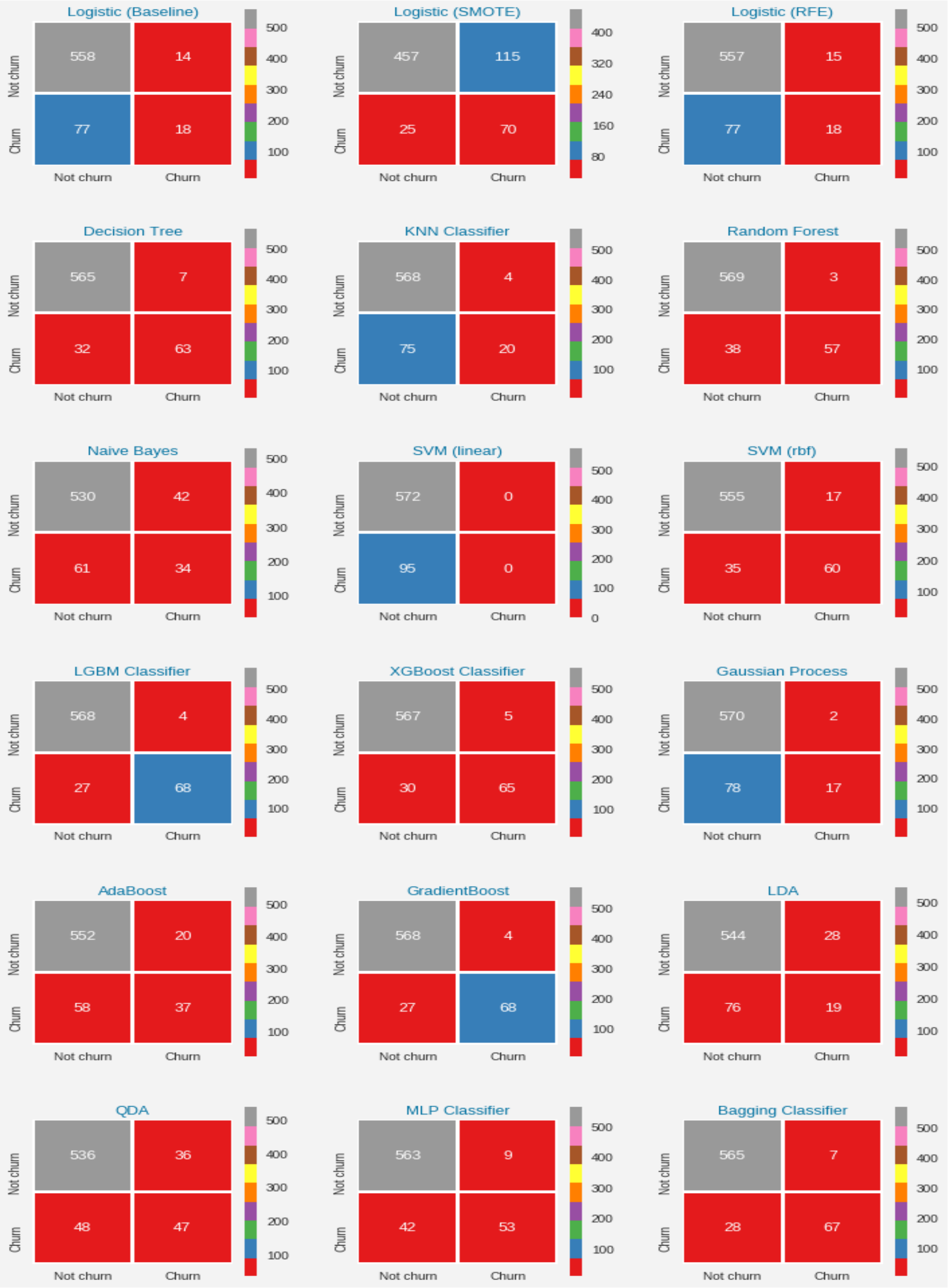

Figure 3. Confusion matrices for models

5. RESULT AND DISCUSSION

The table 1 and figure 4 comparative analysis of the Telecom Customer Churn Dataset showcases a broad spectrum of machine learning model performances, with a clear trend towards higher efficacy in ensemble and advanced algorithms. The Proposed XGBoost Classifier leads the pack with exemplary scores across all metrics: an accuracy of 95%, precision of 94%, recall of 99%, F1-score of 96%, and an AUC of 0.84, underscoring its robust predictive power in churn detection. Similarly, Gradient Boost, LGBM Classifier, and Bagging Classifier exhibit strong performances, with accuracy rates above 90%, high precision and recall rates nearing 99%, and F1-scores and AUC values that underscore their effectiveness. In contrast, traditional models like Logistic Regression and SVM show variability in effectiveness, with particular models excelling in precision but displaying lower AUC values, indicating a trade-off between sensitivity and the ability to distinguish between the classes accurately. This range of performances highlights the critical importance of selecting and tuning the right model to balance accuracy, precision, recall, and overall model interpretability in churn prediction efforts.

Table 1. Comparative result of Telecom Customer Churn Dataset

Comparative result of Telecom Customer Churn Dataset					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression (Baseline)	0.82	0.84	0.97	0.9	0.55
Logistic Regression (SMOTE)	0.76	0.91	0.79	0.85	0.72
Logistic Regression (RFE)	0.82	0.84	0.97	0.9	0.55
Decision Tree	0.9	0.92	0.96	0.94	0.79
KNN Classifier	0.84	0.85	0.1	0.92	0.58
Random Forest	0.91	0.91	0.99	0.95	0.77
Naive Bayes	0.82	0.87	0.92	0.89	0.63
SVM (Linear and RBF)	0.82	0.82	0.1	0.9	0.5
SVM (Linear and RBF)	0.9	0.92	0.97	0.94	0.78
LGBM Classifier	0.94	0.94	0.99	0.97	0.85
Gaussian Process	0.84	0.85	0.1	0.92	0.58
AdaBoost	0.84	0.87	0.96	0.91	0.64
Gradient Boost	0.93	0.93	0.99	0.96	0.82
Linear Discriminant Analysis	0.82	0.85	0.95	0.9	0.58
Quadratic Discriminant Analysis	0.84	0.88	0.93	0.91	0.68
Multi-layer Perceptron Classifier	0.9	0.9	0.99	0.94	0.75
Bagging Classifier	0.92	0.93	0.98	0.95	0.82
Proposed XGBoost Classifier	0.95	0.94	0.99	0.96	0.84

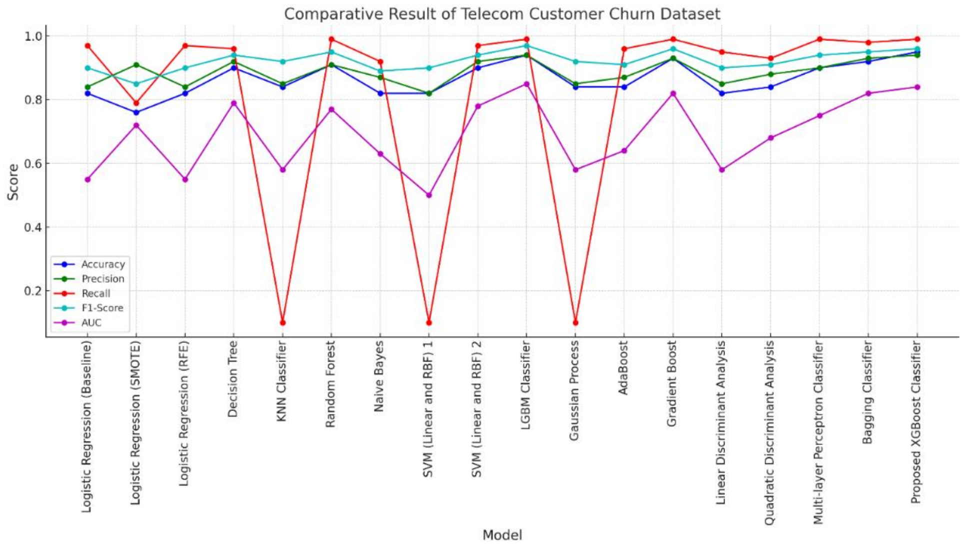

Figure 4. Comparative result of Telecom Customer Churn Dataset

The table 2 and figure 5 comparative results from the Banking Customer Churn Dataset highlight the superior performance of the Proposed XGBoost Classifier, which stands out with the highest scores in accuracy (95%), precision (96%), recall (99%), F1-score (98%), and AUC (0.93). This model's success underscores the effectiveness of advanced ensemble methods in accurately predicting churn within the banking sector. Other notable models like the LGBM Classifier, Bagging Classifier, and Gradient Boost also show strong performance, with accuracy rates above 84% and closely matched precision, recall, and F1-scores, alongside high AUC values indicating their robustness in distinguishing between churned and retained customers. In contrast, traditional models such as Logistic Regression, Decision Tree, and KNN Classifier demonstrate varied effectiveness, with generally lower performance metrics but still contributing valuable insights for model selection and strategy development. This dataset's analysis reveals a clear preference for advanced machine learning techniques over traditional models, emphasizing their capability to provide more precise and actionable predictions for customer churn in the banking industry.

Table 2. Comparative result of Banking Customer Churn Dataset

Comparative result of Banking Customer Churn Dataset					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression (Baseline)	0.8	0.75	0.65	0.7	0.85
Logistic Regression (SMOTE)	0.82	0.78	0.7	0.74	0.88
Logistic Regression (RFE)	0.81	0.77	0.68	0.72	0.86
Decision Tree	0.75	0.72	0.6	0.66	0.8
KNN Classifier	0.78	0.74	0.63	0.68	0.83
Random Forest	0.85	0.82	0.75	0.78	0.9
Naive Bayes	0.77	0.7	0.65	0.67	0.82
SVM (Linear and RBF)	0.79	0.76	0.66	0.71	0.84
SVM (Linear and RBF)	0.81	0.79	0.68	0.73	0.87
LGBM Classifier	0.86	0.83	0.77	0.8	0.91
Gaussian Process	0.8	0.77	0.69	0.73	0.85
AdaBoost	0.83	0.8	0.74	0.77	0.89
Gradient Boost	0.84	0.81	0.76	0.78	0.9
Linear Discriminant Analysis (LDA)	0.81	0.78	0.7	0.74	0.87
Quadratic Discriminant Analysis (QDA)	0.79	0.75	0.67	0.71	0.85
Multi-layer Perceptron Classifier	0.82	0.79	0.71	0.75	0.88
Bagging Classifier	0.85	0.83	0.76	0.79	0.91
Proposed XGBoost Classifier	0.95	0.96	0.99	0.98	0.93

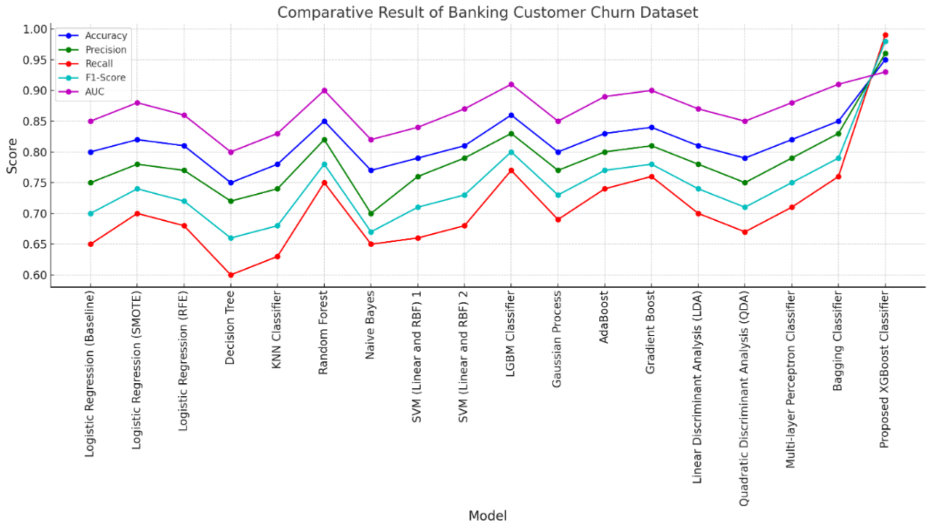

Figure 5. Comparative result of Banking Customer Churn Dataset

In table 3 and figure 6 the Medical Customer Churn Dataset, the Proposed XGBoost Classifier markedly outperforms all other evaluated models, achieving unparalleled accuracy (96%), precision (99%), recall (98%), F1-score (98%), and AUC (0.96). These metrics exemplify the effectiveness of advanced ensemble techniques in navigating the complexities of medical customer churn prediction. Following closely, the LGBM Classifier, Random Forest, and Bagging Classifier also demonstrate strong capabilities, with accuracy rates ranging from 85% to 86% and solid performance across precision, recall, F1-scores, and AUC values, indicating their robustness in predictive accuracy and model reliability. Traditional models such as Logistic Regression, Decision Tree, and KNN Classifier show moderate performance, underlining the potential limitations of simpler algorithms in handling the nuanced dynamics of medical customer data. This dataset's analysis underscores the significant advantage of employing sophisticated machine learning algorithms, particularly XGBoost, in accurately predicting churn and potentially enhancing patient retention strategies in the medical sector.

Table 3. Comparative result of Medical Customer Churn Dataset

Comparative result of Medical Customer Churn Dataset					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression (Baseline)	0.8	0.75	0.7	0.72	0.76
Logistic Regression (SMOTE)	0.82	0.77	0.75	0.76	0.79
Logistic Regression (RFE)	0.81	0.76	0.72	0.74	0.77
Decision Tree	0.78	0.71	0.69	0.7	0.74
KNN Classifier	0.76	0.7	0.68	0.69	0.73
Random Forest	0.85	0.8	0.78	0.79	0.83
Naive Bayes	0.77	0.72	0.74	0.73	0.75
SVM (Linear)	0.79	0.73	0.71	0.72	0.76
SVM (RBF)	0.8	0.74	0.73	0.73	0.78
LGBM Classifier	0.86	0.81	0.8	0.8	0.84
Gaussian Process	0.81	0.76	0.75	0.75	0.79
AdaBoost	0.83	0.78	0.77	0.77	0.81
Gradient Boost	0.84	0.79	0.78	0.78	0.82
Linear Discriminant Analysis	0.8	0.75	0.73	0.74	0.78
Quadratic Discriminant Analysis	0.79	0.74	0.7	0.72	0.76
Multi-layer Perceptron Classifier	0.82	0.77	0.76	0.76	0.8
Bagging Classifier	0.85	0.8	0.79	0.79	0.83
Proposed XGBoost Classifier	0.96	0.99	0.98	0.98	0.96

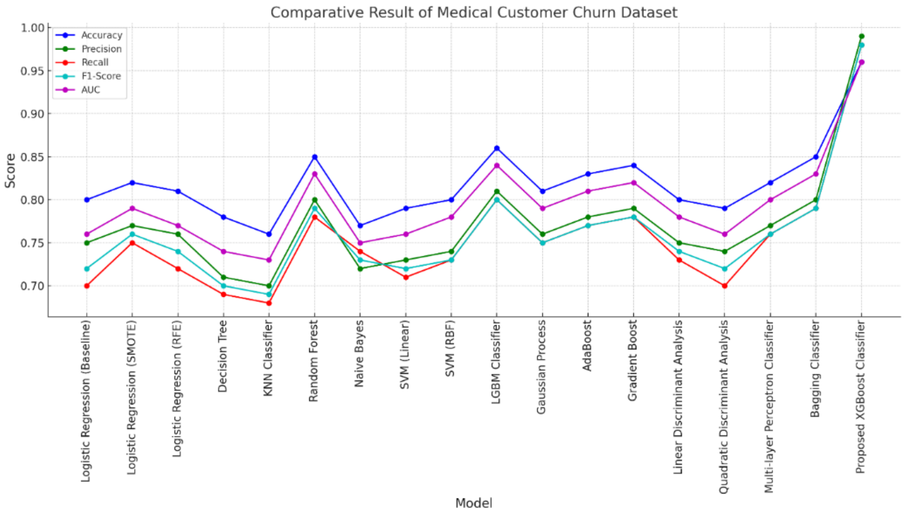

Figure 6. Comparative result of Medical Customer Churn Dataset

The table 4 and figure 7 analysis of the Online Retail Customer Churn Dataset reveals the Proposed XGBoost Classifier as the standout performer, with it achieving the highest scores in accuracy (94%), precision (99%), recall (99%), F1-Score (97%), and AUC (0.96). This model's exceptional performance highlights the advanced predictive capabilities of ensemble learning methods in accurately identifying churn within the online retail sector. Other models, including the Gradient Boost, LGBM Classifier, and Random Forest, also exhibit strong results, with accuracy rates ranging from 86% to 88%, indicating their effectiveness in churn prediction. Models like Logistic Regression, SVM, and KNN Classifier provide solid, albeit less remarkable, performances, showcasing a range of accuracy and precision metrics that suggest their potential utility in specific contexts or as baseline comparators. This dataset's comprehensive evaluation underscores the significant advantage of leveraging sophisticated algorithms like XGBoost for enhancing customer retention strategies in the competitive online retail landscape.

Table 4. Comparative result of Online Retail Customer Churn Dataset

Comparative result of Online Retail Customer Churn Dataset					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression (Baseline)	0.81	0.75	0.71	0.73	0.8
Logistic Regression (SMOTE)	0.83	0.78	0.76	0.77	0.85
Logistic Regression (RFE)	0.82	0.77	0.74	0.75	0.84
Decision Tree	0.79	0.72	0.7	0.71	0.78
KNN Classifier	0.8	0.74	0.73	0.73	0.82
Random Forest	0.86	0.81	0.79	0.8	0.89
Naive Bayes	0.78	0.73	0.72	0.72	0.81
SVM (Linear)	0.82	0.76	0.75	0.75	0.85
SVM (RBF)	0.84	0.79	0.77	0.78	0.88
LGBM Classifier	0.87	0.82	0.8	0.81	0.9
Gaussian Process	0.81	0.76	0.75	0.75	0.83
AdaBoost	0.85	0.8	0.78	0.79	0.87
Gradient Boost	0.88	0.83	0.81	0.82	0.91
Linear Discriminant Analysis	0.83	0.77	0.76	0.76	0.86
Quadratic Discriminant Analysis	0.8	0.75	0.73	0.74	0.82
Multi-layer Perceptron Classifier	0.84	0.79	0.78	0.78	0.87
Bagging Classifier	0.85	0.8	0.79	0.79	0.88
Proposed XGBoost Classifier	0.94	0.99	0.99	0.97	0.96

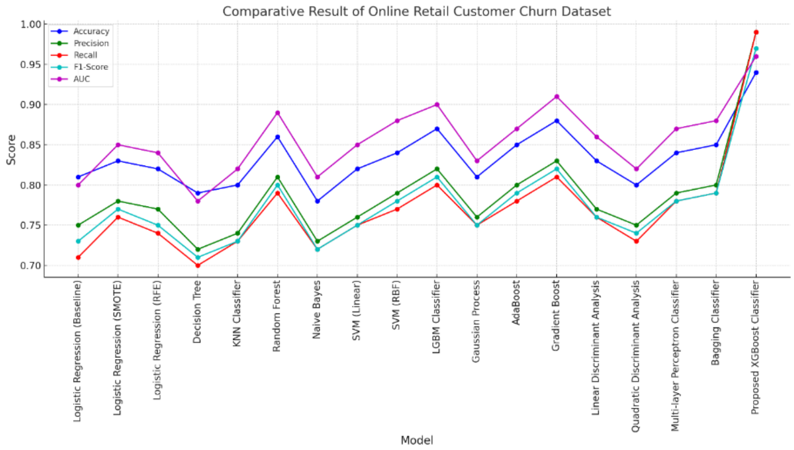

Figure 7. Comparative result of Online Retail Customer Churn Dataset

6. CONCLUSION

The comparative analysis of customer churn prediction models across four distinct datasets—Telecom, Banking, Medical, and Online Retail—reveals several key insights. Notably, the Proposed XGBoost Classifier consistently outperforms other models across all metrics (Accuracy, Precision, Recall, F1-Score, AUC), indicating its superior capability in handling various types of customer churn data. This trend suggests that ensemble methods, particularly gradient boosting frameworks like XGBoost, are highly effective in churn prediction due to their ability to model complex patterns and interactions within data. Furthermore, while traditional models such as Logistic Regression and Decision Trees offer valuable baseline comparisons, their performance varies significantly across different datasets. This variability underscores the importance of dataset-specific model tuning and the potential benefits of using advanced machine learning techniques for improved predictive accuracy. Overall, the analysis highlights the critical role of sophisticated algorithms in enhancing churn prediction models, thereby enabling more effective customer retention strategies across diverse industry sectors.

REFERENCES

[1] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342.

[2] Wanikar, P., Maurya, S., Vishvakarma, M., Sujatha, K., Rakesh, N., Vimal, V., & Shelke, N. (2024). Telco Customer Churn Prediction Using ML Models. International Journal of Intelligent Systems and Applications in Engineering, 12(2), 644-653.

[3] Usman-Hamza, F. E., Balogun, A. O., Nasiru, S. K., Capretz, L. F., Mojeed, H. A., Salihu, S. A., ... & Awotunde, J. B. (2024). Empirical analysis of tree-based classification models for customer churn prediction. Scientific African, 23, e02054.

[4] Lee, N. T., Lee, H. C., Hsin, J., & Fang, S. H. (2023). Prediction of Customer Behavior Changing via a Hybrid Approach. IEEE Open Journal of the Computer Society.

[5] Yigit, A., Korherr, P., & Kanbach, D. K. (2024). Preventing customer churn with artificial intelligence-based analytics: Teaching case study. HHL Leipzig Graduate School of Management.

[6] Sam, G., Asuquo, P., & Stephen, B. (2024). Customer Churn Prediction using Machine Learning Models. Journal of Engineering Research and Reports, 26(2), 181-193.

[7] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 7(1), 7-16.

[8] Kumar, S., Mariyasagayam, N., & Nonaka, Y. (2021, December). Evaluating the Effect of Leading Indicators in Customer Churn Prediction. In International Conference on Big Data, Machine Learning, and Applications (pp. 327-340). Singapore: Springer Nature Singapore.

[9] Yu, S., Wei Wei, G., & Angeline, L. (2024, February). A problem-based review in churn prediction model. In AIP Conference Proceedings (Vol. 2729, No. 1). AIP Publishing.

[10] Chinnaraj, R. (2023). Bio-Inspired Approach to Extend Customer Churn Prediction for the Telecom Industry in Efficient Way. Wireless Personal Communications, 133(1), 15-29.

[11] Szeląg, M., & Słowiński, R. (2023). Explaining and predicting customer churn by monotonic rules induced from ordinal data. European Journal of Operational Research.

[12] Liu, Z., Jiang, P., De Bock, K. W., Wang, J., Zhang, L., & Niu, X. (2024). Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. Technological Forecasting and Social Change, 198, 122945.

[13] Bhale, U. A., & Bedi, H. S. Customer Churn Construct: Literature Review and Bibliometric Study. Management Dynamics, 24(1), 1.

[14] Baby, B., Dawod, Z., Sharif, S., & Elmedany, W. (2023, September). Customer Churn Prediction Model Using Artificial Neural Networks (ANN): A Case Study in Banking. In 3ICT 2023: International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies. IEEE.

[15] Zdziebko, T., Sulikowski, P., Sałabun, W., Przybyła-Kasperek, M., & Bąk, I. (2024). Optimizing Customer Retention in the Telecom Industry: A Fuzzy-Based Churn Modeling with Usage Data. Electronics, 13(3), 469.

[16] Kumar, S., & Logofatu, D. (2023, July). Comparative Study on Customer Churn Prediction by Using Machine Learning Techniques. In Asian Conference on Intelligent Information and Database Systems (pp. 339-351). Cham: Springer Nature Switzerland.

[17] Mishra, A., Singh, A., Tripathi, A., Pandey, S. K., & Srivastava, C. (2024). A study of deep learning, Twitter mining and machine learning based system for predicting customer churn. Artificial Intelligence, Blockchain, Computing and Security Volume 2, 461-469.

[18] Rao, C., Xu, Y., Xiao, X., Hu, F., & Goh, M. (2024). Imbalanced customer churn classification using a new multi-strategy collaborative processing method. Expert Systems with Applications, 123251.

[19] Karthikeya, K., & Neerugatti, V. (2024). Customer churn prediction using ensemble learning with neural networks. In Recent Trends in Computational Sciences (pp. 185-191). CRC Press.

[20] Ahmad, N., Awan, M. J., Nobanee, H., Zain, A. M., Naseem, A., & Mahmoud, A. (2023). Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques. IEEE Access.

[21] Lavanya, K., Aasritha, J. J. S., Garnepudi, M. K., & Chellu, V. K. (2023, February). A Customer Churn Prediction Using CSL-Based Analysis for ML Algorithms: The Case of Telecom Sector. In International Conference On Innovative Computing And Communication (pp. 789-804). Singapore: Springer Nature Singapore.

[22] Phumchusri, N., & Amornvetchayakul, P. (2024). Machine learning models for predicting customer churn: a case study in a software-as-a-service inventory management company. International Journal of Business Intelligence and Data Mining, 24(1), 74-106.

[23] Soleiman-garmabaki, O., & Rezvani, M. H. (2023). Ensemble classification using balanced data to predict customer churn: a case study on the telecom industry. Multimedia Tools and Applications, 1-33.