Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(110)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101581

Customer churn prediction model in enterprises using machine learning

Yamini B[1], K.Venkata Ramana[2], M.Nalini[3], D.Chitra Devi[3], B.Maheswari[4] and Siva Subramanian.R[5]*

Assistant Professor, Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India[1]
Assistant Professor, Department of Computer Science, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad-500090, Telangana, India[2]
Associate Professor, Department of Computer Science, S.A. Engineering College, Poonamalle-600077, Tamil Nadu, India[3]
Assistant Professor, Department of Computer Science, R.M.K Engineering College, Kavaraipettai-601206, Tamil Nadu, India[4]
Associate Professor, Department of Computer Science, R.M.K College of Engineering and Technology, Puduvoyal-601204, Tamil Nadu, India[5]

Received: 12-May-2023; Revised: 16-December-2023; Accepted: 18-December-2023

Abstract

The customer is a vital and integral component of all organizations worldwide. The success of any business hinges on its ability to engage customers effectively in terms of the products or services offered. Research into customer churn is a critical element in the operations of any enterprise. Customer churn prediction (CCP) helps to understand customer interactions with a business and often to identify when customers are likely to cease doing business with the company. In this study, Machine Learning (ML) algorithms are utilized for effective CCP. This study considers various supervised learning models, as the dataset used is labeled. The ML models employed include logistic regression (LR), k-nearest neighbors (k-NN), decision tree (DT), random forest (RF), XGBoost (XGB), light gradient boosting machine (LightGBM), and CatBoost. The ML model is applied to a dataset of customer churn sourced from the Kaggle repository. The results of the experiments are evaluated using several validity metrics, such as accuracy, recall, precision, area under the curve (AUC), and F1-score. The experimental data reveal that LR excels in terms of recall (0.5275) and accuracy (0.581), while the CatBoost model leads in AUC (0.8415), precision (0.6564), and F1-score (0.581). Moreover, LightGBM achieves results close to those of LR and CatBoost. The research findings indicate that the use of ML contributes to the prediction of customer churn. Additionally, the experimental results suggest that LR, CatBoost, and LightGBM outperform other ML models. Utilizing this knowledge, enterprises can develop more effective strategies for customer retention and enhance their business's financial performance.

Keywords

Customer churn, Enterprises, Prediction, ML, CatBoost, LightGBM, Logistic regression.

1.Introduction

The business of any enterprises is highly contingent upon their consumers. Understanding how the customers perform with the enterprises business is crucial one. Since the firm's business profit is based upon the customer satisfaction. To comprehend how the customer relationship with enterprises business can be analysed using the customer churn prediction (CCP). The term customer churn refers to frequency at which the client cease doing the business with the enterprises [1]. The customer churn can be measured using no product purchased during the time of interval by the consumer. Customer churn is significant issue in any enterprises around the globe.

Due to numerous factors contributing to customer churn, enterprises often face adverse effects, including revenue loss and the need to invest significant resources in acquiring new customers. The primary causes of customer churn include:

- Attracting the wrong customers.
- Mismatch between product offerings and customer needs.

*Author for correspondence

- Failure to achieve expected outcomes from customers.
- Lack of customer engagement.
- Absence of proactive support.

The enterprises can surmount the issue by devising new design strategies to enhance the customer attrition, business model and profitability. By overcoming the customer churn the enterprises can increase the sales & revenue and lower the customer lifetime value (CLV) with the enterprises [2]. CLV refers to how long the client remain active with any enterprises business.

Each enterprise should conduct the customer churn owing to:
comprehend the no of discontinued customers-here grasp the reason why the consumer cease doing the business with the enterprise. What are the faults with the enterprises and why the consumers migrate out first we have to comprehend [3].
Expenses in acquiring customers-By Minimizing the customer churn the enterprises can reduce the resource allocation to acquire new customers. By this the enterprises cost can be reduced and greater service can be given.
Customer lifecycle value–Comprehending the customer churn the enterprises can enhance the customer lifecycle value with the firms. The CLV refers to how long the client remain active with any enterprises business.
Better enterprises functioning- By understanding customer attrition, we can get the insights about the customer and further using this information to enhance the customer value with the enterprises. To conduct efficient customer churn the use of machine learning (ML) algorithm is contemplated. ML is classified into three categories, and a ML technique is considered and employed contingent on the dataset and challenge [4, 5]. In this study various supervised learning models are considered, Since the dataset employed are labelled one. Logistic regression (LR), *k*-nearest neighbours (*k*-NN), decision trees (DT), random forest (RF), XGBoost (XGB), light gradient boosting machine (LightGBM), and CatBoost are the ML models used. The ML model is applied to a dataset of customer churn obtained from the Kaggle repository. The results of the investigations are recorded using several validity metrics such as accuracy, recall, precision, area under curve (AUC), and F1-score. Finally, the model that produces the greatest results is shown and predicted. By figuring out which customers are most likely to depart, ML models help businesses take steps to keep customers from fleeing. This could mean offering focused bargains, making customer service better, or giving consumers personalised suggestions to keep them as customers. Overall, ML models are used to predict customer loss because they can study large and complicated datasets, learn from past data, get more accurate over time, and enable effective actions to stop customers from fleeing. By using these models, companies can increase the number of consumers they retain, bring in more money, and make their customers more loyal.

The following compelling causes prompted this research:
Business value: Customer turnover is important for many industries. A company's revenue and profitability depend on understanding and reducing customer attrition. This research is driven by the goal to help organizations overcome this massive impediment and improve their financial health.

Economic impact high customer churn reduces revenue, client acquisition costs, and CLV. Businesses have a financial incentive to understand churn and reduce it. Customer focus today's customer-centric atmosphere makes retaining customers cheaper than attracting new ones. This research aims to increase consumer loyalty, pleasure, and satisfaction.
Data-driven decisions: ML algorithms analyse massive, complex information to reveal consumer behaviour and churn patterns. These models are used to boost customer retention via data-driven decision-making.
Competitive advantage: Companies that retain consumers and learn about their preferences and behaviour have a competitive advantage.
To provide insightful advice and suggestions that will help companies create proactive plans to retain more customers, lower customer churn, and ultimately boost revenue and profitability.

The aim of the work is to get better insights about the customer churn, its effects, and the application of ML models to anticipate and manage it efficiently. It also provides organizations with actionable advice on how to maximize customer retention and financial success.

The remainder of this work is organized into five sections. Section 2, explores into existing work in the field of CCP, providing insights into previous research and methodologies. Section 3, details the overarching approach and techniques used in this research, offering a comprehensive view of the methods applied. In Section 4, the focus shifts to the experimental procedures that were conducted and the results that were derived from these experiments. Finally, Section 5, brings together all the findings, summarizing the key outcomes and insights gained from the research.

2.Literature survey

The researcher performs an in-depth work on why and for the customer churn is faced by enterprises. To understand and monitor the customer churn an efficient churn prediction model is needed. Based upon the above-mentioned problem, the author applies a swish recurrent neural network (S-RNN) model which efficiently performs churn prediction [6]. The customer churn dataset is used to conduct the experiment, and the outcomes of the suggested model are compared to those of the recurrent neural network (RNN), deep neural networks (DNN), convolutional neural network (CNN), and artificial neural network (ANN) models. The study of the outcome suggests that the suggested method is superior than the others.

The author discusses the problem of telecom business client attrition. Further, the researcher implies the use of a ML model aids to forecast customer churn efficiently and reduces the customer churn within an enterprise [7]. Based upon the study different ML models are considered: LR, Naive Bayes (NB), Support vector machine (SVM), RF, boosting, ensemble, and decision tree (DT). A customer churn dataset is used to conduct the experiment, and validity ratings are applied to project & context the results of several models. The study of the results suggests that the AdaBoost and XGB models are better than the others.

The importance of customer churn in the telecom sector & its impact on market performance are addressed by the author. The author uses social network analysis (SNA) features in conjunction with many ML models, including DT, RF, gradient boosting (GBM) and XGB, to solve the aforementioned issue [8]. The SyriaTel dataset is used to conduct the experimental process, and validity ratings are used to project & context the results of several models. The study of the results suggests that the XGB is better than the others.

The importance of customer churn research in the financial sector has been explored, with recommendations to utilize ML for effective analysis of customer churn. Traditional approaches often rely on structured datasets for conducting customer churn analysis. However, in this research, unstructured datasets were utilized to achieve a more comprehensive understanding of customer churn. The experimental process involves the application of XGB, RF, LR, and NB models. This approach allows for a more nuanced analysis compared to conventional methods.

The author explores the importance and role of the customers in enterprises and the problem of customer churn. In this research the author proposes a loss prediction that is based upon *k*-means customer segmentation and SVM to perform customer churn analysis [10]. Further, SVM and LR are applied to perform customer churn. The experimental results show after customer segmentation the results are improved significantly and with compare to the two ML models, SVM archives superior results.

The author explores the loss of a consumer due to heavy competition in the telecom industry. To understand the customer churn in this research the author develop a churn prediction model. LR & Fisher discriminant equations are considered to build a churn prediction model. The empirical process is conducted out with the Chinese telecom companies' dataset and outcome generated indicate that LR has superior results [11].

The author performs an in-depth study about customer churn in the banking sector. Based upon above mentioned study the researcher use of filesystem (FS) and the ML helps to conduct effective analysis about the customer churn analysis. In ML, chi-square automatic interaction detection (CHAID), classification & regression (CR), neural network (NN), bayesian network (BN), and C5 are applied. The empirical process is conducted with the bank credit dataset [12].

The author suggests that without a thorough understanding and analysis of customer behavior, enterprises cannot effectively mitigate customer churn. Based on the study, the author proposes the use of a deep backpropagation artificial neural network (Deep-BP-ANN) combined with two feature selection (FS) approaches: Variance thresholding and LR methods. The experimental process utilizes datasets from IBM Telco and Cell2Cell, applying various ML approaches such as XGB, LR, NB, *k*-NN, and Deep-BP-ANN. The results of the analysis indicate that the Deep-BP-ANN model outperforms the others [13].

The researchers imply the need of customers in enterprises and further address the need for customer churn analysis. Based upon above mentioned study the researcher applies four different ML models to perform CCP. ML models like DT, *k*-NN, SVM, and RF applied. The empirical procedure is conducted out with a customer churn dataset and results imply RF performs superior compared to others [14].

The researchers perform an important work about customer churn in enterprises. To overcome customer churn and to retain the customer for a long time, the study of customer patterns is considered as important. Based on the study the author applies FS with the ML model to perform customer churn analysis. In ML, SVM, multi-layer perceptron (MLP), RF, and NB are applied. The FS approach combines IG & ranker methods. The empirical process is done using the customer churn dataset & the outcome imply using FS with ML performs superior compared to others [15].

Telecommunications companies struggle to estimate client attrition owing to rivals' departing offerings and network difficulties. Churn rate greatly affects client lifetime value and income. Churn models affect industry revenue; thus, companies use them. ML methods like DT, RF, and XGB forecast which consumers will terminate their subscriptions, enhancing services and lowering churn. This strategy makes telecom services lucrative, improving client retention and revenue [16].

CCP is important for enterprises businesses in industries like telecommunications, banking, insurance, and e-commerce. ML approaches are used to develop effective models, using past customer data to identify churn predictors. A study applied popular supervised ML algorithms to predict churn in the telecommunications industry. The study found that ML is a viable method for predicting customer churn, with ensemble learners outperforming single-base learners. A balanced learning dataset is expected to enhance classifier performance [17].

In Businesses retaining consumers, and churn reduction is a big issue. This article describes using a Stacking Classifier to predict ecommerce turnover using gender and tenure. The ensemble learning classifier incorporates metaclassifiers from *k*-NN, SVM, RF Classifiers, and DT. The Stacking Classifier surpasses other ML methods with 98.2% accuracy, 98.1% Area under receiver operating characteristic curve (AUROC), and 95.0% F-Measure score [18].

A BiLSTM-CNN model is integrated with RNNs and CNNs in parallel to increase CCP accuracy. Bank data was used to compare the AttnBLSTM-CNN model. Compared to the bidirectional long short-term memory networks (BLSTM) model, the AttnBiLSTM-CNN model increased accuracy by 0.2%, churn rate by 1.3%, F1 value by 0.0102, and AUC value by 0.0103[19].

E-commerce companies prioritize customer churn reduction owing to competitiveness and high expenditures. Customer churn forecasting frameworks use AI for insight and advice to increase accuracy and discover non-churn users. The framework includes exploratory data analysis (EDA), data pretreatment, model tuning, model comparison, and insight and recommendation. The suggested strategy accurately predicts customer attrition, with CatBoost performing best in the dataset. The work addresses a research gap and adds to CCP literature. The system improves e-commerce efficiency by using the best classifier for insight and suggestion [20].

Service providers compete fiercely in the fast-growing telecom business. This study created Improved_ XGB, a churn prediction model based on XGB. The model performed above 99% in accuracy, precision, recall, and F1-measure utilizing South Asia GSM and churn-big datasets [21].

Large industries struggle with customer attrition, which lowers income. Telecom firms may use a ML model to anticipate churn and develop efficient marketing retention tactics. The model analyses datasets using pre-processing, data visualization, and existing datasets. Supervisory computer learning methods like LR, SVM, DT, and RF generate classification models. Performance parameters like accuracy and F1 recall determine the optimal algorithm. This product aids telecom corporations in decision-making and customer attrition management [22].

To improve business perks and client retention, the study predicts consumer behaviour using Cat Boost Classifier. The model outperforms conventional approaches with 95% prediction accuracy, improving QoS and customer retention [23].

Client attrition in digital banking requires intelligence models to improve customer satisfaction. A classifier-based model for CCP utilizing *k*-NN, LR, AdaBoost, gigabytes (GB), and RF is presented in this work. Hyperparameter optimization and tweaking improve model performance, with an experimental case study obtaining 87% accuracy [24].

Business and finance utilize customer churn to represent the progressive loss of customers. Companies may identify prospective customers who will quit to reduce churn. Age, location, gender, credit card information, and balance may be utilized to anticipate client departures using ML methods like LR and NB. These models can make probabilistic predictions, with NB outperforming LR. Understanding these consumers helps banks avoid client loss [25].

The following points are summarized from the above survey:

1. Due to its detrimental effects on revenue and client retention, customer churn is a major problem for several businesses, including e-commerce, banking, and telecom.
2. A number of algorithms, including XGB, AdaBoost, RF, and Deep-BP-ANN, have shown promise in various scenarios when it comes to ML models, which are extensively used to forecast and minimize customer turnover.
3. Feature selection strategies and ensemble learning approaches are often used to improve churn prediction models' accuracy.
4. There is no one-size-fits-all ML model; instead, the choice of model often relies on the particular dataset and enterprises.
5. To enhance classifier performance in churn prediction, a well-balanced training dataset is essential.
6. Churn prediction algorithms have a big effect on revenue, CLV, and general company success.
7. Thorough data pretreatment, effective model tweaking, and visualization are essential for enhancing churn prediction models.
8. In summary, organizations may utilize data analysis and ML approaches to help them make wise choices and put in place successful retention strategies that lower customer attrition and raise customer satisfaction.

3.Methodology

Customer churn is the occurrence in which consumers discontinue utilizing a company's goods or services. The purpose of the research is to identify drifting customer in enterprises and to surmount the problem and to increase the customer lifecycle value with the enterprises. Customer churn leads to cessation of using the business service due to various reason by the customers [26]. This makes to cause loss to enterprises business and burden in acquiring new consumers. To overcome the customer churn problem, we have to mark the consumers are who possible to be expelled from the business and devise new strategies to enhance the customer value. Based upon the above problem, in this investigation CCP is conducted using ML algorithm. The objective is to conduct improved CCP using ML models. LR, *k*-NN, DT, RF, XGB, lightgbm, and catboost are the ML models used in CCP. The ML model is applied to a dataset of customer churn obtained from the Kaggle repository. The results of the investigations are recorded using several validity metrics such as accuracy, recall, precision, AUC, and F1-score. Finally, the method that produces the greatest results is shown and predicted. Overall methodology is projected in the *Figure 1*.

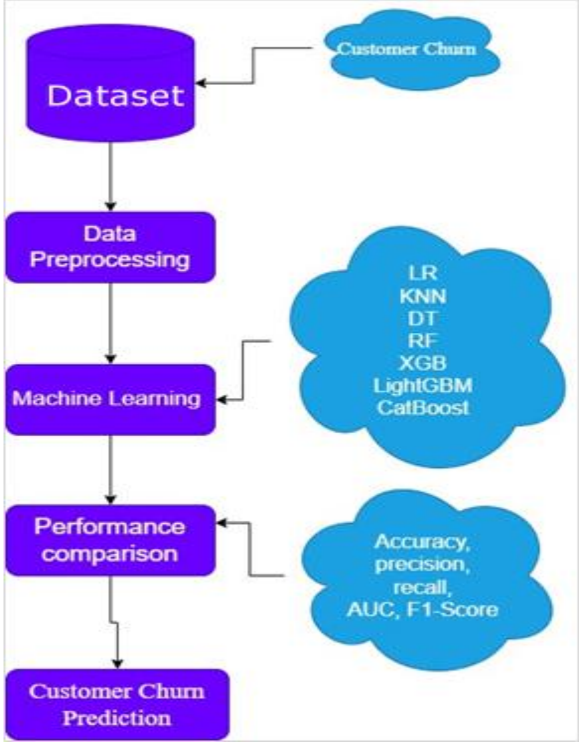

Figure 1 Overview of methodology

3.1Data collection

Data acquisition is a crucial stage in predicting customer churn. In this study, a dataset on customer churn is utilized. The dataset is captured from Kaggle repository which consists of 14 attributes with 10000 entries.

3.2Data pre-processing

Data preparation is a crucial stage in both data analysis and ML. It means putting raw data in a form that can be used for research. By doing data preparation jobs well, data scientists and ML experts can build models that are more accurate, reliable, and effective at fixing business problems. Here mean & median are calculated for the non-missing values and used in computing of the missing values.

3.3 Prediction models

ML is a branch of AI that teaches computers to anticipate and act without being taught. ML algorithms can analyze massive, complicated information, find patterns, and make predictions or judgments [27]. This model uses LR, *k*-NN, DT, RF, XGB, LightGBM, and CatBoost ML models.

The models used are determined by taking into account the kind of data, the intricacy of the problem, and the balance between interpretability and forecast accuracy. With its unique benefits and potential for superior performance in certain situations, each model is a useful instrument for mitigating loss of customers across a range of sectors. The particulars and limitations of the current churn prediction issue determine which model is best. For the purpose of training and testing the dataset is split into 80 and 20 ratios.

Logistic regression: LR solves binary classification issues. LR is supervised learning. Using feature values, it predicts a new observation's class probabilistically. The logistic function (sigmoid function) transfers every input value to a probability value between 0 and 1 in the LR model. Labelled data trains the LR model. The approach optimizes model parameters to minimize the discrepancy between expected probability and actual labels. Gradient descent optimizes this. LR can classify binary and multi-class issues. Multi-class LR may employ a softmax function to forecast class probabilities [28]. LR is straightforward and interpretable, making it easy to understand and execute. It works effectively for datasets with several characteristics. The hyperparameter applied in the LR are penalty (Regularization term (l2 norm), Inverse regularization strength (C =1.0), Maximum number of iterations(max_iter), and Optimization solver(solver='lbfgs').

K-nearest neighbors: The *k*-NN method finds the *k* closest training dataset neighbors to an input location. Euclidean distance is used to compute the distance between the input location and each of the *k* neighbors. The technique outputs the class label most prevalent among *k* neighbors for classification issues. Regression outputs the average of *k* neighbors' values. *k*-NN predicts using labelled training data since it is supervised. *k*, distance metric, and data scaling greatly affect algorithm performance [29]. A bigger value of *k* smoothes the decision boundary but may over fit, whereas a lower value may over fit. *k*-NN is simple, flexible, and handles non-linear decision bounds. However, huge datasets and high-dimensional data make it computationally costly. Image recognition, recommendation systems, and anomaly detection employ *k*-NN. It is beneficial for datasets with complicated decision boundaries that other algorithms cannot simulate. The hyperparameter applied in the *k*-NN are Number of neighbors to consider(n_neighbors=5), Weighting scheme (uniform or distance) and Algorithm to compute nearest neighbors(algorithm=auto).

Decision trees: DT it is a tree-like supervised system that represents choices and their probable outcomes. The decision tree algorithm recursively partitions incoming data by feature values. A feature's value determines the tree node's choice. The tree branches reflect the two consequences of the choice. This method continues until all data points are classified or a stopping requirement is reached [30]. The decision tree algorithm helps visualize and understand decision-making. Non-experts can understand and visualize the tree to explain predictions. The method handles category and numerical data, making it adaptable. The decision tree method can handle non-linear input-output interactions. It can manage missing numbers and outliers, making it noise-resistant. If the tree is complicated or the training dataset is short, the method may over fit. Finance, healthcare, and marketing employ DT. They can anticipate consumer purchase probability, illness prediction, and credit application risk. The hyperparameter applied in the DT are Maximum depth of the tree, Minimum number of samples required to split an internal node(min_samples_split=2), and Minimum number of samples required to be at a leaf node(min_samples_leaf=1)

Random forest: RF is a common classification and regression technique. The ensemble learning approach uses many DT to increase prediction accuracy and resilience. The RF approach trains several DT using a random selection of the training data and input attributes. The DT are integrated by averaging (regression) or majority voting (classification) the individual tree forecasts. RF outperform single DT. It eliminates overfitting and improves prediction accuracy and generalizability. It handles high-dimensional data and missing values, making it useful for many applications. The RF method measures feature significance, a major benefit. Averaging feature relevance scores across all trees in the forest lets you discover which input characteristics are most relevant for predictions. This information may increase data interpretation and uncover feature engineering opportunities. Image and voice recognition, fraud detection, and medical diagnosis employ RF. They work well for huge datasets with complicated input-output interactions [31]. The hyperparameter applied in the RF are Number of trees in the forest(n_estimators=100), Split criterion(criterion=gini), Maximum depth of the trees(max_depth=None), Minimum number of samples required to split an internal node(min_samples_split=2), and Minimum number of samples required to be at a leaf node(min_samples_leaf=1)

XGBoost: XGB is a popular regression & classification technique. Gradient boosting combines numerous weak models (typically DT) into a powerful model. The XGB technique builds DT repeatedly to repair the faults of the preceding tree. Regularised goal functions avoid overfitting and improve model performance. XGB can handle huge, high-dimensional, and missing data. It accepts numerical, category, and text input data. XGB is fast and scalable. It efficiently processes big datasets with millions of rows and thousands of columns due to its parallel processing. Distributed computing is built-in, making it suited for large data settings. XGB handles skewed datasets well. It automatically adjusts input data weights to equalize class distribution, improving prediction performance on unbalanced datasets. Advertising, finance, and healthcare utilize XGB. It predicts customer turnover, detects fraud, and diagnoses ailments from medical photos [32]. The hyperparameter applied in the XGB are Step size shrinkage to prevent overfitting(learning_rate=0.1), Number of boosting rounds(n_estimators=100), Maximum depth of a tree, Subsample ratio of the training instances(subsample=1.0), Subsample ratio of columns when constructing each tree(colsample_bytree=1.0)

LightGBM: Lightgbm is a fast, open-source ML method used for classification, regression, & ranking. It handles big datasets with millions or billions of rows and thousands of columns using gradient boosting. Each decision tree in LightGBM is trained to rectify the faults of the preceding tree. LightGBM has numerous differences from previous gradient boosting methods. Gradient-based one-side sampling (GOSS) reduces computational complexity and speeds up training.

Based on the loss function gradients, GOSS discards data items that are unlikely to improve the model's performance. Second, LightGBM utilizes a leaf-wise growth approach instead of a level-wise method, reducing the tree's decision nodes and improving generalization. Finally, LightGBM supports category, numerical, and missing values. LightGBM outperforms gradient boosting methods. For big datasets, it is quicker and more memory-efficient than other techniques. It's flexible and accurate, handling numerical and categorical data. Recommendations, fraud detection, and images categorization employ LightGBM. It recommends items, detects fraud, and classifies photographs of things and animals [32]. The hyperparameter applied in the LightGBM are Step size shrinkage to prevent overfitting(learning_rate=0.1), Number of boosting rounds(n_estimators=100), Maximum depth of a tree (-1 means no limit), Subsample ratio of the training instances(subsample=1.0), Subsample ratio of columns when constructing each tree(colsample_bytree=1.0)

CatBoost: CatBoost is an open-source gradient boosting toolkit for classification, regression, and ranking. Yandex, a Russian technology firm, created it to handle categorical characteristics and high-dimensional data. CatBoost trains an ensemble of DT to repair each other's mistakes. CatBoost varies from gradient boosting methods in key ways. First, ordered boosting optimizes gradient boosting by ordering feature pairings. This method reduces noisy input and improves model generalization. CatBoost handles category characteristics with the Taylor series expansion. This approach translates categorical characteristics into numerical values and retains category ordering. Thirdly, CatBoost handles numerical and missing values without preprocessing. CatBoost outperforms gradient boosting methods. For big datasets, it is quicker and more memory-efficient than other techniques. It handles category and numerical data. It outputs feature significance ratings and handles unbalanced datasets. Recommender systems, fraud detection, and consumer segmentation employ CatBoost. It recommends items, detects fraudulent financial activities, and groups clients by behaviour and preferences [32]. The hyperparameter applied in the

CatBoost are Step size shrinkage to prevent overfitting(learning_rate=0.1), Number of boosting rounds(iterations=100), Depth of the trees, L2 regularization term(l2_leaf_reg=3), and Random seed for reproducibility(random_seed=42).

To understand the efficiency of the model the empirical results captured from the methodology is presented & projected using different metrics like: accuracy, AUC, recall and precision [33, 34].

4.Experimental results

This section describes in detail about the experimental results carried out in the methodology.

4.1Experimental Results of CCP using various ML models

The results of the task to predict customer churn using seven different ML models are shown in *Figure 2*. The efficiency of each model is judged based on a metric called accuracy, which determines how accurate the model's predictions are by calculating the proportion of times they are true. Findings reveals, the LR model had the maximum accuracy, which is represented by the value 0.8045. This value shows that the model accurately predicted 80.45% of the instances. CatBoost came up at number two when it came to accuracy, scoring 0.7998 out of 1.00, followed closely by LightGBM, which scored 0.7961 out of 1.00. The other models also accomplished acceptable levels of accuracy, with RF reaching an accuracy of 0.791, XGB earning an accuracy of 0.7868, and *k*-NN achieving an accuracy of 0.7614 respectively. The accuracy of the DT model was the lowest of all the models, coming in at 0.7291, the lowest of any model.

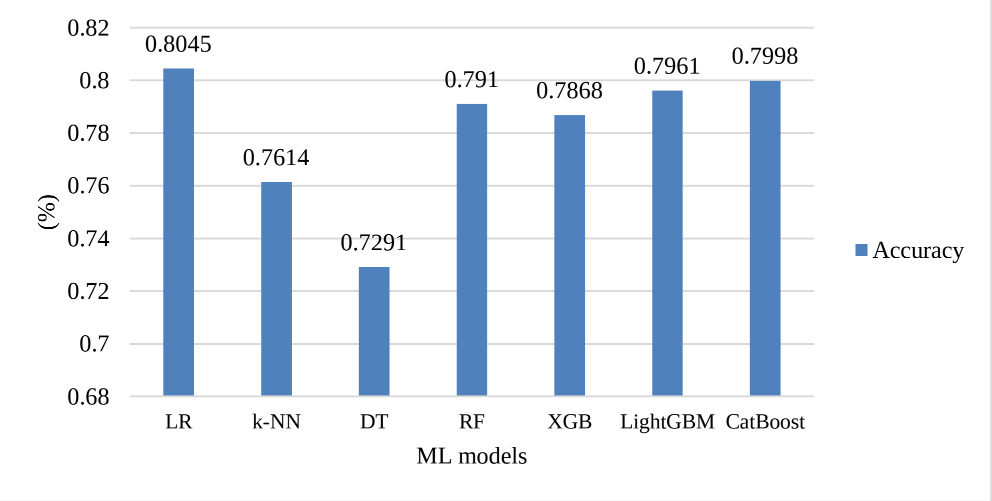

Figure 2 CCP using 7 different ML models using accuracy parameter

The AUC parameter was utilised as the evaluation metric for predicting customer churn using 7 distinct ML models, as presented in *Figure 3*. A greater AUC value signifies an enhanced ability of the model to discriminate between positive and negative instances. The experiment employs a range of models, including LR, *k*-NN, DT, RF, XGB, LightGBM, & CatBoost. The AUC metric is a commonly employed tool for assessing model performance, quantifying the extent to which the positive and negative classes can be distinguished from one another. Based on the findings, it is evident that LR exhibits the greatest AUC score of 0.8443, while CatBoost and LightGBM follow closely with AUC values of 0.8415 and 0.8355, respectively. In contrast, the Decision Tree model exhibits a relatively low AUC value of 0.6581, thereby suggesting its inadequacy for the given task. In general, the findings indicate that ensemble techniques, namely RF, XGB, LightGBM, and CatBoost, exhibit superior performance in forecasting customer churn compared to conventional models, including LR, *k*-NN, and DT.

Figure 4 shows the results of seven different ML models that were used to predict customer churn. The Recall parameter was used as a measure of how well each model did. Recall is the number of true positives as a percentage of the total number of true positives plus fake negatives. LR, *k*-NN, DT, RF, XGB, LightGBM, & CatBoost. are the models that are used in this experiment. From the results it is evident that the LR has the highest recall value with 0.5393, followed by LightGBM with 0.5275 and CatBoost with 0.5217. On the other hand, *k*-NN's Recall number of 0.4414 is the lowest. Overall, the results show that none of the models in this experiment are very good at predicting customer churn based on the Recall measure, since they all have low numbers. But it looks like LR is the best model out of the seven in this Figure. The findings of seven different ML models that were used in the process of forecasting customer turnover are shown in *Figure 5*, with the Precision parameter serving as the assessment measure. In this experiment, the models used include LR, *k*-NN, DT, RF, XGB, LightGBM, and CatBoost. The results indicate that the LR model achieves the highest precision score of 0.6626, followed by the CatBoost model with a score of 0.6564, and the LightGBM model with 0.6432. Conversely, the DT model shows the lowest Precision at 0.4906. Overall, based on the Precision metric, LR, CatBoost, and LightGBM appear to be the most effective models for predicting customer attrition in this study.

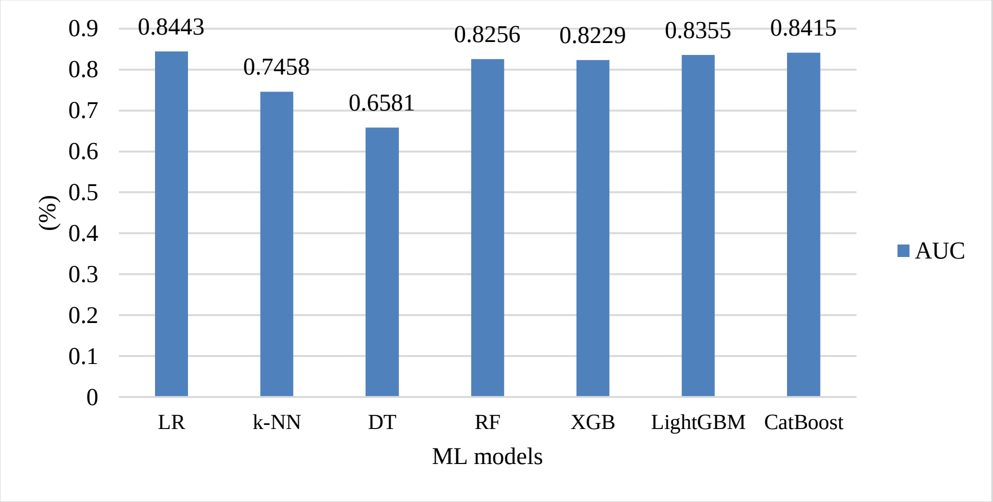

Figure 3 CCP using 7 different ML models using AUC parameter

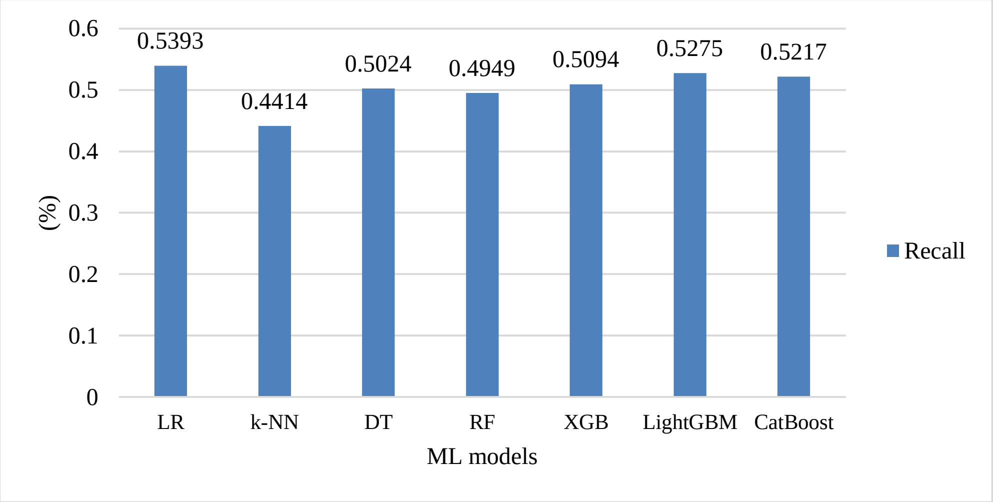

Figure 4 CCP using 7 different ML models using recall parameter

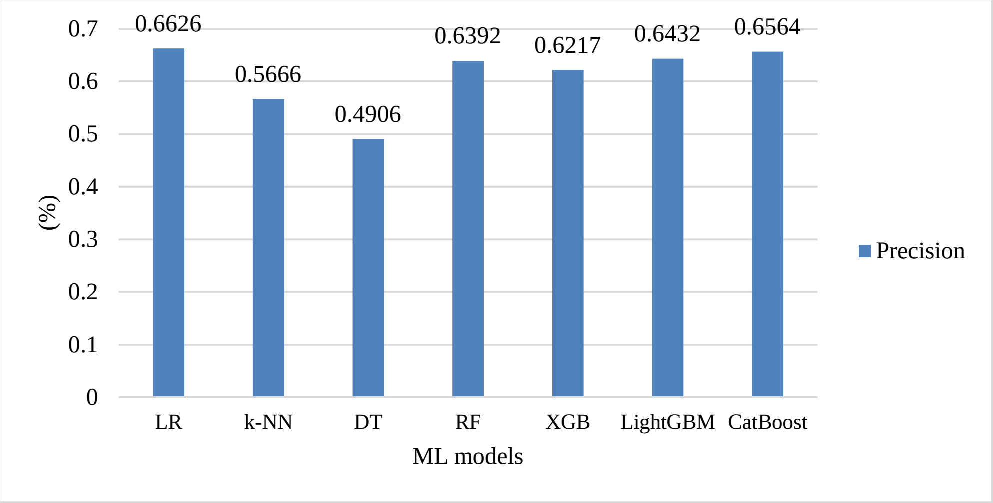

Figure 5 CCP using 7 different ML models using precision parameter

The findings of seven distinct ML models that were used in the process of forecasting customer turnover are shown in *Figure 6*, with the F1-score parameter serving as the assessment score. It is the harmonic mean of accuracy and recall, and it combines both metrics to offer an overall performance score. The F1-score is also known as the F1-score. LR, *k*-NN, DT, RF, XGB, LightGBM, & CatBoost are the models that have been used in this experiment. LR has the highest F1-score value of 0.5944, as can be seen from the data; this is followed by LightGBM with 0.5789 and CatBoost with 0.581. On the other side, *k*-NN's F1-score value of 0.4956 is the lowest of all the scores. According to the findings of the research, the usage of ML aids in the prediction of customer churn. Furthermore, the experimental findings suggest that LR, CatBoost, and LightGBM outperform other ML models. Using this information, the enterprise can develop better methodology to retain the customer and improve the enterprises business financial turnover.

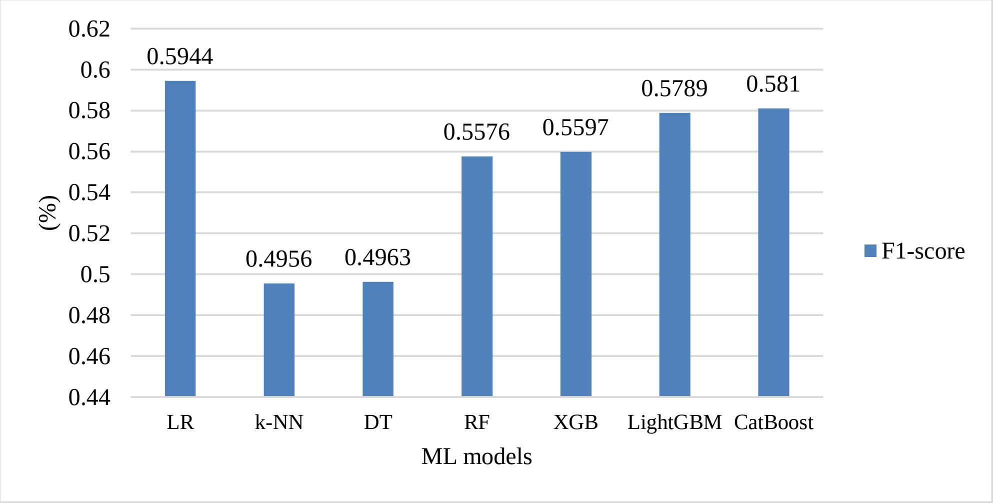

Figure 6 CCP using 7 different ML models using F1-score parameter

4.2Result discussion

Figure 2 to *6* represent five performance metrics were used in the analysis: accuracy, recall, AUC, precision, and F1-score.

1. Top Models for Accuracy: LightGBM (0.7961), CatBoost (0.7998), and LR (0.8045) got the closest accuracy.
2. Best Model for Recall: LightGBM (0.5275) and XGB (0.5094) came in second and third,

respectively, behind LR in recall (the capacity to detect churners).

3. AUC Rankings: With a score of 0.8415, CatBoost outperformed LR (0.8443) and LightGBM (0.8355) in terms of AUC (the ability to differentiate between churners & non-churners).
4. CatBoost had the greatest accuracy rating of 0.6564, followed by LightGBM (0.6432) and LR (0.6626).
5. F1-score: LightGBM (0.5789), LR (0.5944), and CatBoost (0.581) had the greatest F1-score (harmonic mean of accuracy and recall).

In summary, LightGBM, CatBoost, and LR are the most promising models for predicting customer turnover; their efficacy varies based on the performance metric that is given priority.

4.3Result findings

1. Objective: To perform efficient CCP using the ML Models.
2. Models: In this study, 7 various ML models are considered and applied.
3. Validity Scores: In this research, 5 different validity scores are considered and applied
4. Performing Models: Compare to 7 ML models, LR, CatBoost and LightGBM performs superior compare to other ML models. LR perform superior compare to others in terms of recall (0.5275) and accuracy (0.581) and CatBoost model perform superior compare to others in terms of AUC (0.8415), precision (0.6564), and F1-score (0.581). Also, LightGBM achieves closer results to LR and CatBoost
5. Application: The proposed methodology is useful to perform CCP and help business to develop better strategies to retain customers and improve their values.

4.4Interpretations

1. From the experimental results its clear choosing the best model depends upon the customer churn challenges and the dataset.
2. LR is best suited when accuracy and recall is considered.
3. CatBoost is best suited when precision is considered.
4. LightGBM makes the balanced choice between accuracy and recall.

4.5Implications

The findings have significant ramifications on businesses trying to lower client attrition. They may choose the best model for their particular use case based on their objectives. This may result in client retention tactics that are more successful.

4.6Limitations

1. Model and Dataset: The ML model and the dataset applied is not optimal around the universe. The results may change according to the problem and dataset
2. Ensemble Model: In this research ensemble models are applied.
3. Assumption of Stationarity: Customer patterns do not remain constant. They change over time, leading to potential inaccuracies in the model

4.7Recommendations

The following suggestions might be made considering these findings:

1. Organizations should carefully consider their goals before selecting a churn prediction model. LR, CatBoost, and LightGBM are competitive options with distinct functions.
2. Since customer behaviours and habits may vary over time, it is essential that churn prediction models be continuously monitored and updated.
3. Investigating more characteristics or outside data sources may improve churn forecast accuracy.

4.8Comparative analysis

When comparing the models, LR stands out with the highest overall accuracy and recall. CatBoost excels at distinguishing churners from non-churners, as evidenced by its high AUC score. Meanwhile, LightGBM achieves the best F1-score, striking a balance between precision and recall. In summary, this study highlights the importance of careful model selection when predicting customer attrition. Through such strategic selection, companies can reduce churn and make more informed decisions regarding customer retention strategies, ultimately enhancing both customer satisfaction and financial outcomes.

A complete list of abbreviations is summarised in *Appendix I.*

5.Conclusion

For organizations, customer churn is a serious issue since losing clients may result in a drop in income, reputation, and market share. CCP is the practice of identifying customers who are likely to leave a firm and taking steps to keep them. This is done using ML algorithms. Based on past customer information, ML models may be taught to spot trends and other causes of customer turnover as well as identify the clients most likely to depart in the future. These models may then be used to create focused retention measures to keep clients who are at danger of leaving, such as personalized offers or proactive customer care. Customer retention may be improved and corporate profitability can be raised by using CCP in a variety of sectors, including telecom, banking, retail, and healthcare. The findings of the research indicate that ML models may be useful for forecasting customer turnover, and the findings also indicate that different models have varying degrees of success with respect to various assessment measures. LR had the greatest values for accuracy and recall, while CatBoost had the highest values for precision and F1-score. LR, CatBoost, and LightGBM all had the top values for AUC. Based on these results, it seems that a mix of models may be required to get the highest overall performance in forecasting customer turnover. The precise requirements and priorities of the organization will determine which models should be used in conjunction with one another. Overall, the research indicates the potential of ML in detecting and maintaining consumers who are at danger of leaving a firm, which may be beneficial for organizations in terms of boosting customer retention and enhancing profitability.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data Availability

The dataset used in this study is publicly available at https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling/data.

Author's contribution statement

Yamini B, D.Chitra Devi: Responsible in software setup and analysis, Writing – original draft, Writing – review and editing. **K.Venkata Ramana, B.Maheswari:** Responsible as in writing contributors including grammar checking, Writing – original draft, Writing – review and editing. **M.Nalini, Siva Subramanian.R:** Responsible for entire study and analysis of experiment, Writing – original draft, Writing – review and editing.

References

[1] Mishra A, Reddy US. A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. In international conference on inventive computing and informatics 2017 (pp. 721-5). IEEE.

[2] Qi J, Zhang Y, Zhang Y, Shi S. TreeLogit model for customer churn prediction. In Asia-Pacific conference on services computing 2006 (pp. 70-5). IEEE.

[3] Prabha D, Subramanian RS. A survey on customer relationship management. In 4th international conference on advanced computing and communication systems 2017 (pp. 1-5). IEEE.

[4] Hopkins E. Machine learning tools, algorithms, and techniques. Journal of Self-Governance and Management Economics. 2022; 10(1):43-55.

[5] Muneer A, Ali RF, Alghamdi A, Taib SM, Almaghthawi A, Ghaleb EA. Predicting customers churning in banking industry: a machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science. 2022; 26(1):539-49.

[6] Sudharsan R, Ganesh EN. A swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science. 2022; 34(1):1855-76.

[7] Lalwani P, Mishra MK, Chadha JS, Sethi P. Customer churn prediction system: a machine learning approach. Computing. 2022:1-24.

[8] Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data. 2019; 6(1):1-24.

[9] Vo NN, Liu S, Li X, Xu G. Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems. 2021; 212:106586.

[10] Xiahou X, Harada Y. B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research. 2022; 17(2):458-75.

[11] Zhang T, Moro S, Ramos RF. A data-driven approach to improve customer churn prediction based on telecom customer segmentation. Future Internet. 2022; 14(3):1-19.

[12] AL-najjar D, Al-rousan N, AL-najjar H. Machine learning to develop credit card customer churn prediction. Journal of Theoretical and Applied Electronic Commerce Research. 2022; 17(4):1529-42.

[13] Fujo SW, Subramanian S, Khder MA. Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters. 2022; 11(1):185-98.

[14] Rahman M, Kumar V. Machine learning based customer churn prediction in banking. In 4th international conference on electronics, communication and aerospace technology 2020 (pp. 1196-201). IEEE.

[15] Saheed YK, Hambali MA. Customer churn prediction in telecom sector with machine learning and information gain filter feature selection algorithms. In international conference on data analytics for business and industry 2021 (pp. 208-13). IEEE.

[16] Thorat MA, Sonawane VR. Customer churn prediction in the telecom industry using machine learning algorithms. Computer Integrated Manufacturing Systems. 2023; 29(4):1-11.

[17] Kumar S, Logofatu D. Comparative study on customer churn prediction by using machine learning techniques. In Asian conference on intelligent

information and database systems 2023 (pp. 339-51). Cham: Springer Nature Switzerland.

[18] Awasthi S. Customer churn prediction on E-commerce data using stacking classifier. Authorea Preprints. 2023:1-10.

[19] Liu Y, Shengdong M, Jijian G, Nedjah N. Intelligent prediction of customer churn with a fused attentional deep learning model. Mathematics. 2022; 10(24):1-16.

[20] Jahan I, Sanam TF. An improved machine learning based customer churn prediction for insight and recommendation in E-commerce. In 25th international conference on computer and information technology 2022 (pp. 1-6). IEEE.

[21] Swetha P, Dayananda RB. A customer churn prediction model in telecom industry using improved_XGBoost. International Journal of Cloud Computing. 2023; 12(2-4):277-94.

[22] Pandithurai O, Ahmed HH, Sriman B, Seetha R. Telecom customer churn prediction using supervised machine learning techniques. In international conference on advances in computing, communication and applied informatics 2023 (pp. 1-7). IEEE.

[23] Angelina JJ, Subhashini SJ, Baba SH, Reddy PD, Reddy PS, Khan KS. A machine learning model for customer churn prediction using CatBoost classifier. In 7th international conference on intelligent computing and control systems 2023 (pp. 166-72). IEEE.

[24] Galal M, Rady S, Aref M. Enhancing customer churn prediction in digital banking using ensemble modeling. In 4th novel intelligent and leading emerging sciences conference 2022 (pp. 21-5). IEEE.

[25] Agarwal V, Taware S, Yadav SA, Gangodkar D, Rao AL, Srivastav VK. Customer-churn prediction using machine learning. In 2nd international conference on technological advancements in computational sciences 2022 (pp. 893-9). IEEE.

[26] Tsai TY, Lin CT, Prasad M. An intelligent customer churn prediction and response framework. In 14th international conference on intelligent systems and knowledge engineering 2019 (pp. 928-35). IEEE.

[27] Günesen SN, Şen N, Yıldırım N, Kaya T. Customer churn prediction in FMCG sector using machine learning applications. In IFIP international workshop on artificial intelligence for knowledge management 2021 (pp. 82-103). Cham: Springer International Publishing.

[28] Siva SR, Prabha D. Optimizıng naive bayes probability estimation in customer analysis using hybrid variable selection. In computer networks and inventive communication technologies: proceedings of third ICCNCT 2021 (pp. 595-612). Springer Singapore.

[29] Subramanian RS, Prabha D, Aswini J, Maheswari B, Anita M. Alleviating NB conditional independence using multi-stage variable selection (MSVS): banking customer dataset application. In journal of physics: conference series 2021 (pp. 1-10). IOP Publishing.

[30] Subramanian RS, Prabha D, Maheswari B, Aswini J. Customer analysis using machine learning algorithms: a case study using banking consumer dataset. Recent Trends in Intensive Computing. 2021; 689-94.

[31] Subramanian RS, Prabha D. Ensemble variable selection for naive bayes to improve customer behaviour analysis. Computer Systems Science & Engineering. 2022; 41(1):339-55.

[32] Al DE. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering. 2019; 13(1):6-10.

[33] Maheswari B, Bushra SN, Nirmala G, Anita M, Smys S, Kamel KA, et al. Enhancing customer prediction using machine learning with feature selection approaches. Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems. 2023; 563.

[34] Prabha D, Subramanian RS, Balakrishnan S, Karpagam M. Performance evaluation of naive bayes classifier with and without filter based feature selection. International Journal of Innovative Technology and Exploring Engineering. 2019; 8(10):2154-8.

Dr. Yamini B obtained her Bachelor of Engineering in Computer Science and Engineering from Mailam Engineering College in 2003. She then completed her Master of Technology in Information Technology at Sathyabama University, Chennai, in 2007. In 2020, she earned her Doctor of Philosophy in Computer Science and Engineering from Sathyabama Institute of Science and Technology, Chennai. Throughout her career, Dr. Yamini has published numerous papers in various international and national conferences and journals. Currently, she serves as an Assistant Professor in the Department of Networking and Communications at the SRM Institute of Science and Technology, College of Engineering and Technology, School of Computing. Her research interests encompass Network Security, Cyber Forensics, Image Processing, Information Retrieval Systems, Machine Learning, Deep Learning, and Cloud Computing.

Email: yamini.subagani@gmail.com

Dr. K.Venkata Ramana earned his Ph.D. from Jawaharlal Nehru Technological University Hyderabad, India, in 2021, following his Master's in Computer Science and Engineering from the same university in 2010. He is currently serving as an Assistant Professor in the Department of Computer Science & Engineering at the VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India. Previously, he spent 10 years as the Head of the Master of Computer Applications Department at Bhoj Reddy Engineering College for Women in Hyderabad. With over 20 years of teaching experience, his research interests include Data Mining for Software

Engineering, Machine Learning, Deep Learning, and Cloud Computing. Dr. Ramana has published 15 papers in various esteemed international journals and conferences, focusing on Source Code Mining.
Email: venkataramana_k@vnrvjiet.in

Dr. M.Nalini is an Associate Professor in the Department of Computer Science and Engineering at S.A Engineering College, Chennai, India. She earned her B.E. in Computer Science and Engineering from Anna University, Chennai, in 2010, and her M.Tech. in the same field from B.S. Abdur Rahman Crescent Institute of Science & Technology, Chennai, in 2012. She received her Ph.D. in Computer Science and Engineering from St. Peter's Institute of Higher Education and Research, Chennai, in 2018. With a decade of teaching experience, her research focuses on Data Mining, Machine Learning, Big Data Analytics, and Networking. She has published numerous articles in various esteemed journals.
Email: nalini.tptwin@gmail.com

Dr. D.Chitra Devi is an Associate Professor in the Department of Computer Science and Engineering at S.A. Engineering College, Chennai, India. She completed her M.Sc. in Mathematics from Madras University, Chennai, in 2000 and her M.E. in Systems Engineering and Operations Research from the College of Engineering, Chennai, in 2009. In 2020, she earned her Ph.D. in Computer Science and Engineering from the College of Engineering, Chennai. With 19 years of teaching experience, her research interests include Cloud Computing, Green Computing, Data Mining, Machine Learning, and Big Data Analytics. She has published numerous research articles in esteemed journals.
Email: drchitradevi@saec.ac.in

B.Maheswari is currently working as an Assistant Professor in the Department of Computer Science and Engineering at R.M.K Engineering College, Chennai, Tamil Nadu, India. She holds a B.E. degree in Computer Science and Engineering and an M.E. degree in Software Engineering, both from Anna University, India. She is also pursuing her Ph.D. in Computer Science and Engineering at Anna University, Chennai. Her research interests are primarily in the fields of the Internet of Things and Big Data.
Email: mahesasi23@gmail.com

Siva Subramanian.R is an Associate Professor in the Department of Computer Science and Engineering at RMK College of Engineering and Technology, Chennai, India. He earned his B.E. in Computer Science and Engineering from Anna University, Chennai, in 2009, and his M.Tech. in the same field from Bharath University, Chennai, in 2013. With a decade of teaching experience, his research interests focus on Data Mining, Machine Learning, Big Data Analytics, and Networking. He has published numerous articles in well-regarded journals.
Email: sivasubramanian12@yahoo.com

Appendix I

S. No.	Abbreviation	Description
1	ANN	Artificial Neural Network
2	AUC	Area Under Curve
3	AUROC	Area Under Receiver Operating Characteristic Curve
4	BLSTM	Bidirectional Long Short-Term Memory Networks
5	BN	Bayesian Network
6	CCP	Customer Churn Prediction
7	CLV	Customer Lifetime Value
8	CHAID	Chi-Square Automatic Interaction Detection
9	CNN	Convolutional Neural Network
10	CR	Classification & Regression
11	Deep-BP-ANN	Deep Backpropagation Artificial Neural Network
12	DNN	Deep Neural Networks
13	DT	Decision Trees
14	EDA	Exploratory Data Analysis
15	FS	Filesystem
16	GB	Gigabytes
17	GBM	Gradient Boosting
18	*k*-NN	K-Nearest Neighbours
19	LightGBM	Light Gradient Boosting Machine
20	LR	Logistic Regression
21	MLP	Multi-layer perceptron
22	NN	Neural Network
23	ML	Machine Learning
24	RF	Random Forest
25	RNN	Recurrent Neural Network
26	S-RNN	Swish Recurrent Neural Network
27	SNA	Social Network Analysis
28	SVM	Support Vector Machine
29	XGB	XGBoost