KERNEL FUZZY C-MEANS AND RNN-BASED FRAMEWORK FOR CUSTOMER CHURN ANALYSIS IN TELECOM

S. Induja[1], Dr.V.P. Eswaramurthy[2]

[1] Research Scholar, Periyar University, Salem – 636011, Tamilnadu, India.,
[2] Assistant Professor, Department of Computer Science, Government Arts and Science College, Komarapalayam, Tamilnadu, India.

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S.I(1).pp554-561

KEYWORDS
Customer Relationship Management (CRM), Kernel based Fuzzy C-Means algorithm, Optimized XGBoost algorithm, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Support Vector Machine (SVM), Kernel Extreme Learning Machines (KELM)

Received on:

20-07-2024

Accepted on:

10-12-2024

Abstract

A strategic method named customer relationship management (CRM) that builds profitable, offering long-lasting connections with important customers and other stakeholders. These contributions are considered to be the main objective of CRM. In the telecom sector, Customer churn is considered as one of the biggest CRM challenges. Then, maintaining current customers are far less expensive than finding new customers. This made churn as a major financial burden. For telecom sectors, the customer churn's evolving problem has become a major concern. Getting novel customers is significantly more exclusive than holding current ones and it is revealed in this study. To address customer churn analysis and to improve the prediction models efficacy, sophisticated classification algorithms was employed in this study. Predicting individual customer's churn behavior are the primary objective of the study. Effective data pre-processing is the main factor for this approach, as it includes the creation of derived features. Data comprehension is the initial step that supports in commercial significance.

In data pre-processing, limited data accessibility problem can be effectively addressed by implementing kernel-based fuzzy C-Means (KFCM) algorithm, as it converts the incomplete datasets into complete ones. Then, the churn analysis will support in detecting the signs of churn in highly valuable customers. Changes in customer status can be classified as: either partial defection (active use to suspension or non-use) or entire defection (active use to churn), as it will aid in differentiating the two. The mediating effects of partial defection on the connection among the churn determinants and total defection are examined by employing the Optimised XGBoost algorithm. This algorithm acts as a Base Classifier (BC). Here, complex interactions can be effectively handles by this BC, as it will also facilitate in detecting the non-linear patterns (NL) in the data. The Enhanced Gradient Boosting Machine (EGBM) has the ability in selecting the deriving features from customer profiles and call records.Proposed Recurrent Neural Network (PRNN)-based deep learning (DL) approaches like LSTM and GRU form the basis for churn prediction (CP). The standard models like Support Vector Machines (SVM), Logistic Regression (LR), Boosting Algorithms (BA), and Kernel Extreme Learning Machines (KELM) are compared with the proposed RNN model. Robust separation of churn data is offered and it was demonstrated by the experimental outcomes. It also offers high scalability and accuracy. The potential of DL techniques for addressing churn analysis problem in the telecom sector was demonstrated.

INTRODUCTION

For businesses, enduring and profitable relationships with customers was built by the strategy like CRM. Rising customer acquisition costs are facilitated by the CRM, as it has a major part in the telecom sector with highly competitive market [1]. Then, customer churn is considered as one of the major challenge that companies face. Customers discontinuing their services are the main challenge. Significant revenue losses are the outcome of churn, as it also increases the operational costs. Thus, it is significantly less expensive to maintain current clients than to find new ones [2]. In this sector, understanding customer churn patterns and developing effective predictive models are crucial.

The behavioural patterns are identified by the customer churn analysis, as it is inherently complex. This analysis will help in identifying the probability of customers in discontinuing their services [3]. An in-depth data exploration, pre-processing, and feature engineering was demanded in this procedure for the purpose of ensuring the predictive accuracy. Valuable insights regarding churn patterns was offered by the telecom companies, as it helps in collecting vast amounts of data like customer profiles and call detail records [4]. Here, pre-processing is the crucial step in this analysis, because this data has the issues like incompleteness, noise, and high dimensionality (HD) [5]. The churn can be effectively predicted by the companies using advanced methodologies. Thus, the valuable customers can be maintained by executing the procedures.

There are many advantages and it also has many challenges. Differentiating among partial defection and total defection is the main problem. Customers temporarily suspend or reduce service usage are named as partial defection. Complete disengagement represents the total defection [6]. Model is needed for handling complicated and non-linear patterns, as the consistent high-value customers show small signals of churn [7]. Then, the issues in the data including limited data availability, incomplete datasets, and noise complicates the procedure. Thus, generalizability and predictive power becomes less in this model. In the telecom sector, the efficacy of churn analysis is improved by addressing the above mentioned challenges.

A robust methodology was suggested in this study, as it integrates advanced classification and DL techniques for overcoming those issues. To convert incomplete datasets into beneficial ones, and the data pre-processing is executed with the implementation of KFCM algorithm. Then, EGBM is used for selecting the derived key features. The optimal feature representation was also ensured by this EGBM. The LSTM and GRU included in PRNN and this PRNN was employed by the churn analysis model. This PRNN is well-suited for sequential data patterns. With the conventional methods like LR, BA, SVM, and KELM, the suggested model is compared. The suggested DL method overtakes the standard methods and it was demonstrated in the simulation outcomes. High scalability and predictive accuracy was offered by this suggested method. In telecom sector, critical challenges of churn analysis are addressed.

1. Literature Review

In order to predict customer churn in TCI, a novel Customer CP (CCP) architecture called ChurnNet was suggested by Saha et al. [8] . To improve performance, the suggested ChurnNet combines a Spatial Attention Module (SAM), Squeeze-and-Excitation Block (SEB), and Residual Block (RB)with a 1D convolution layer (CL). While the SEB and SAM enable ChurnNet to capture interdependencies among channels and within individual channels. The RB addresses the Vanishing Gradient Problem (VGP). Three publicly accessible datasets were used in trials to evaluate its performance. 3 data balancing methods: SMOTE, SMOTEEN, and SMOTETomek were used because of the datasets' notable Class Imbalance (CI). ChurnNet surpassed State-Of-The-Art (SOTA) models with accuracies of 95.59%, 96.94%, and 97.52% on the three benchmark datasets through thorough experimentation and 10-fold cross-validation (CV).

Using the Naïve Bayes (NB) classifier along with a Genetic Algorithm (GA) (a subset of an Evolutionary Algorithm)(EA)) for feature weighting, Amin et al. [9] introduced an adaptive learning technique for addressing the difficult problem of CCP. The performance of the suggested approach is assessed on publicly accessible datasets, such as BigML Telco Churn, IBM Telco, and Cell2Cell, and shows notable gains over baseline classifiers like CNN, DBP-ANN, LR, XGBoost, KNN, Logit Boost, SVM, PCALB, and NB with default settings. The method attains 0.97, 0.97, and 0.98 average Precision (P), Recall (R)rates of 0.84, 0.94, and 0.97, F1-scores of 0.89, 0.96, and 0.97, MCC values of 0.89, 0.96, and 0.97, and Accuracies (Acc) of 0.95, 0.97, and 0.98 on the respective datasets. These results highlight the efficiency of the suggested method in enhancing CP performance.

A hybrid strategy that integrates clustering and classification algorithms within an ensemble framework was put out by Bilal et al. [10]. First, every clustering technique: K-means, K-medoids, X-means, and random clustering was evaluated independently using two CP datasets. These clusters were then paired with seven distinct classification algorithms to generate hybrid models, and ensembles were used for evaluations.

The analysis was tested with 2 benchmark telecom datasets from the BigML and GitHub platforms. In both BigML and GitHub datasets, the suggested model attains the best prediction accuracy of 92.43% and 94.7%, and it was revealed in the outcomes. The suggested model executes better than the current CP models and it is validating by a sophisticated comparison.

A CP model was developed for the purpose of predicting telecom customer churn via customer segmentation, this CP model was suggested by Zhang et al. [11]. The data was supplied by 3 major Chinese telecom operators. By using LR analysis and Fisher discriminant equations, CP model was developed. Higher prediction accuracy of 93.94% was attained by the regression-based CP and it was revealed in the outcomes. To predict customer churn, telecom companies with an effective tool was utilized in this study. To retain customers and boost profitability, this study allows in implementing targeted measures.

For the purpose of CCP, Pustokhina et al. [12] introduced the ISMOTE-OWELM model, which combines the optimal weighted extreme learning machine (OWELM) with the enhanced synthetic minority over-sampling approach (SMOTE). Preprocessing, categorisation, and balancing the imbalanced dataset are all included in the suggested model. Two important tasks are optimised using a Multi-Objective (MO) rain optimisation algorithm (ROA) (MOROA): fine-tuning (FT) the WELM parameters and optimising the sample rate for SMOTE. First, class labelling and normalisation are applied to the customer data. The dataset imbalance is then addressed using ISMOTE, and the optimal sampling rate is found using the ROA. The data is then classified using the WELM model. The ISMOTE-OWELM model was validated on the CCP Telecommunication dataset after extensive tests. Using datasets, I, II, and III, the ISMOTE-OWELM model achieved accuracies of 0.94, 0.92, and 0.909, respectively, proving its superiority over other models.

The probability of Data Mining (DM) in identifying prospective churn in a telecom sector was discussed by Khalid et al. [13]. To evaluate its effectiveness and the ability of predicting potential churns, several tests were performed, and several classification algorithms were evaluated. To retain customer satisfaction, this data is crucial. Continued subscription to their services was also ensured.

In addition to data transformation and Feature Selection (FS) strategies, Sana et al. [14] provided a variety of Machine Learning (ML) models. A univariate FS process was used to improve the prediction models, and the Grid Search (GS) method was used to find the best Hyperparameters (HP). The models' performance was then assessed using well accepted measures, including as AUC, P, R, and F-measure, in experiments on a number of publicly accessible TCI datasets. The study demonstrated the benefits of using feature selection and data transformation strategies in training optimised CCP models through extensive experimentation. The suggested method significantly improved prediction performance, with gains of up to 26.2% in AUC and 17% in F-measure.

Using a hard tuner to efficiently optimise HP, an enhanced architecture for the Deep Neural Network (DNN) method was suggested by Nalatissifa and Pardede [15]. To find the number of nodes in each Hidden Layer (HL), dropout rate, and Learning Rate (LR), the HP tuning procedure uses Random Search (RS). Stochastic gradient descent (SGD), adaptive moment estimation (Adam), adaptive gradient algorithm (Adagrad), Adadelta, and root mean square propagation (RMSprop) are the five optimisers that are examined in this study, along with three different variations in the number of HL and two Activation Functions (AF) (rectified linear unit (ReLU) and Sigmoid). Using 3 HL with nodes configured as RMSprop optimiser, a dropout rate of 0.1, and a LR of 0.01,the DNN method with RS HP tuning yields an accuracy of 83.09%, according to experimental data. The fastest tuning time was 21 seconds. Comparison techniques like k-nearest neighbour (K-NN), random forest (RF), and decision tree (DT) are outperformed by this method.

Through the use of diverse ML architectures, hybrid models, data transformation methods, and optimisation methodologies, the evaluated research show improvements in CCP. However, a notable research gap lies in the integration of real-time (RT) data streams and adaptive learning mechanisms to handle dynamic customer behavior in ever-evolving telecom environments. Additionally, while several studies address class imbalance and hyperparameter optimization, limited attention has been given to explainability and interpretability of CPmodels, which are critical for practical business applications. Future research could focus on combining real-time analytics with interpretable AI models to enhance prediction accuracy and usability in decision-making (DM) processes.

2. Proposed Methodology

Data preparation and customer behaviour analysis (CBA) are covered in this part. After that, it explores CP and attribute selection (AS) strategies. The system overview, which comes last, provides specifics on the overall functioning of the suggested system.

The functionality of the suggested system is depicted in Figure 1. It begins with the initial stage of data preparation, where data is collected, integrated, and meticulously cleaned. During this stage, data pre-processing is conducted using the kernel based Fuzzy C-Mean algorithm. Following this, customer behavior is analyzed and classified into four distinct categories using the Optimized XGradient Boost algorithm. Subsequently, customer churn behavior is evaluated based on factors (H1, H2, H3, H4), which correlate with the likelihood of customer churn. These probabilities are calculated using boosting techniques. Additionally, attribute selection is performed through the Enhanced Gradient Boosting Machine algorithm to identify relevant features and eliminate redundant or unnecessary ones. At last, classifiers built with RNN utilising DL techniques are fed the refined dataset, which now only contains pertinent attributes.

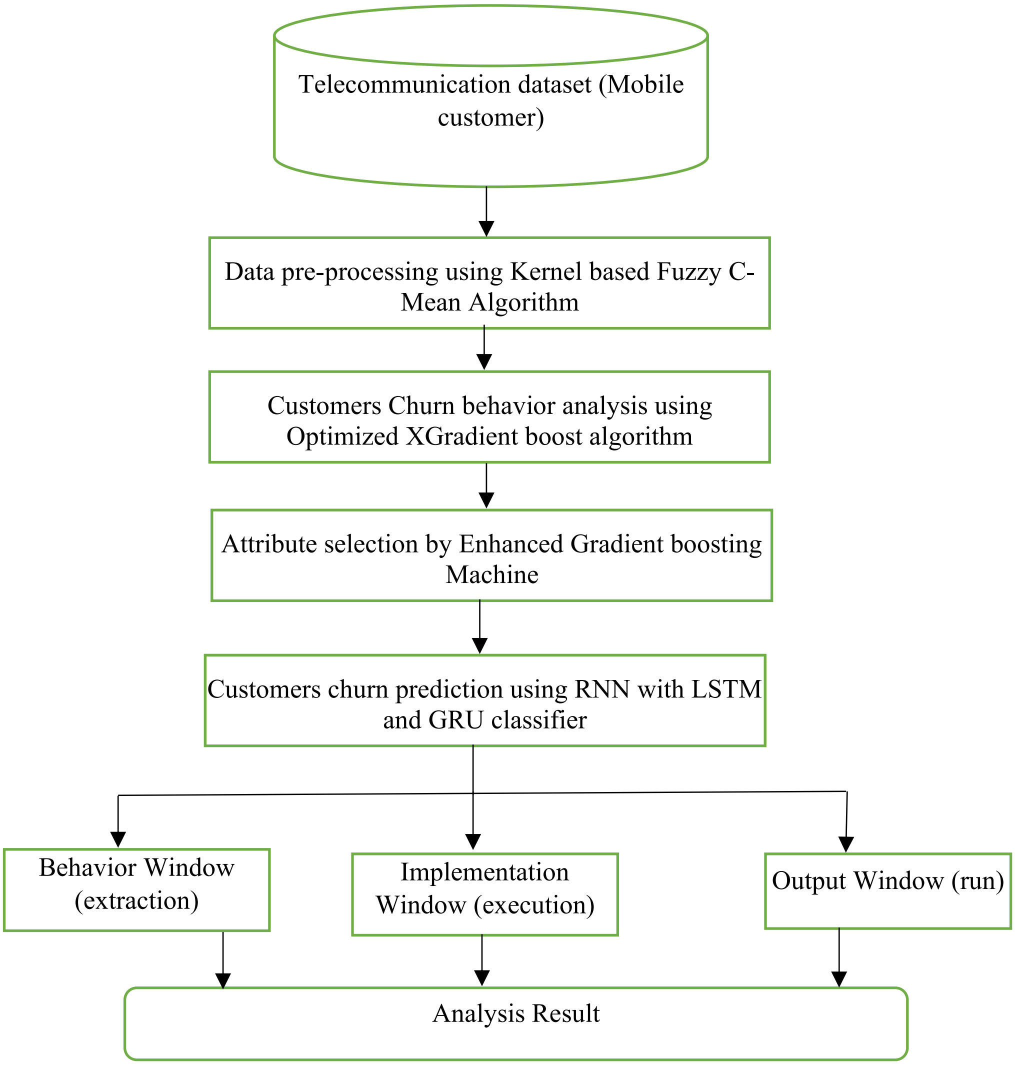

Figure 1. Architecture diagram of the suggested method

Researchers have decided to examine a data set from a telecom provider that comprises a subset of mobile users (millions) who were active at a certain moment in 2010. Base research on the annual churn records of 2010 because the churn prediction model could be skewed by a specific observed point in time. The first step is to extract dataset variables from the mobile customer database, such as billing, usage, product holding, mobile plan and contract information, and inbound and outbound customer care data. A data set representative of a sector of mobile consumers was chosen from a telecommunications company for this study.
The information is taken from the IBM Watson Analytics Telco Customer Churn dataset, that can be found at https://community.watsonanalytics.com/predictive-insights-in-the-telco-customer-churn-data-set/.
Data preparation entails pre-processing of the raw data which converts incomplete dataset into complete dataset by using KFCM Algorithm which containing limited information.

2.1. PREPROCESSING WITH KFCM CLUSTERING

FUZZY CLUSTERING ALGORITHM (FCM)

Using a similarity/dissimilarity measure, clustering algorithms divide the n data objects in a given data set $X = \{x1, x2, , xn\}$ into c groups $C = \{C1, C2, , Cc\}$. The following equation represents the Euclidean distance (ED), which is used as an Objective Function (OF) to be minimised in the typical FCM:

$$J_{FCM}(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m}\|x_j - v_i\|^2 \qquad (1)$$

Here, m is the weighting exponent or the degree of fuzzifier in FCM, and v_i is the cluster centre of cluster c_i.Then, $\sum_{i=1}^{c}\mu_{ij} = 1,\ \forall j = 1,2,3,\dots,n.$ $U = [\mu_{ij}]_{c\times n}$ is a fuzzy partition matrix and the membership degree(MD) of data object x_j to cluster v_i is represented as $\mu_{ij}\epsilon[0,1]$.
Then, v_i and the MD μ_{ij} are modified along the iterations as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \qquad (2)$$

$$c_i = \frac{\sum_{j=1}^{n}\mu_{ij}^{m}x_j}{\sum_{j=1}^{n}\mu_{ij}^{m}} \qquad (3)$$

Until the fuzzy partition matrix changes very little or another stopping criterion is satisfied, it repeats steps (2) and (3).

KERNEL-BASED FCM

In order to overcome the limitations of the FCM algorithm, specifically the inadequacy of the data distribution features to clustering results, the kernel-based fuzzy clustering added the kernel method into the FCM algorithm. $\phi: x \to \phi\,(x)\epsilon\, F,$ is the definition of a NL map. $Here, x\epsilon X,$ the data space is represented by X. The transformed feature space of larger or even infinite dimension is denoted by F [16].
Equation (4-5) defines the KFCM OF.

$$J_{KFCM} = \sum_{i-1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m} \|\phi(x_j) - \phi(v_i)\|^2 \quad (4)$$

Where

$$\|\phi(x_j) - \phi(v_i)\|^2 = K(x_j, x_j) + K(v_i, v_i) - 2K(x_j, v_i) \quad (5)$$

Equation (6) uses the Gaussian Function as its kernel function,

$$K(x, v) = exp\,[-(x-v)^2/\sigma^2],\ \mathrm{K\ (x,\ x)} = 1, \quad (6)$$

where σ is Gaussian Kernel parameter, Equation (7) is expressed below:

$$J_{KFCM}(U, V) = 2\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{m}\left[1 - K(x_j, v_i)\right] \quad (7)$$

Under the constraint of U, Equation (8-9) are simplified, (8)

$$\mu_{ij} = \frac{\left(1 - K(x_j, v_i)\right)^{-1/(m-1)}}{\sum_{1=1}^{c}\left(1 - K(x_j, v_i)\right)^{-1/(m-1)}}$$

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m}(x_j, v_i)x_j}{\sum_{j=1}^{n} \mu_{ij}^{m} K(x_j, v_i)} \quad (9)$$

2.2. Customer Behaviour Analysis Using Optimized XG BA

For tasks involving supervised learning (SL), particularly regression and classification problems, a widely used and effective ML technique known as "XG-Boost" is employed. XGBoost, which stands for "Extreme Gradient Boosting," is an ensemble learning (EL) technique that combines the predictions of many DT to get a more accurate and reliable outcome [17]. In the kth prediction equation (10), the XG-Boost technique combines the best tree model with the existing classification framework using a hybrid learning approach.

$$y_i^k = y_i^{k-1} + f_k(x_i) \quad (10)$$

Where $f_k(x_i)$= optimized tree model

y_i^{k-1}= previous classification model

y_i^k= next prediction classification model

XG-boost OF

$$O^k = \sum_{i=1}^{n} loss_{XG}\left(y_i, y_i^k\right) + \sum_{i}^{k} \Omega\,(f_i) \quad (11)$$

Where

$\Omega\,(f_i)$= Reguralization Function use for overfitting, O^k = Objective Function

The new O^k obtained by applying equations (11) and (12) is

$$O^k = \sum_{i=1}^{n} loss_{XG}\left(y_i, y_i^{k-1} + f_k(x_i)\right) + \sum_{i}^{k} \Omega\,(fi) + \ Constant \quad (12)$$

Taylor Expression

$$f(x + \Delta x) \cong f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \quad (13)$$

Utilize Taylor expression in $O^k$with the equation (13-15)

$$O^k = \sum_{i=1}^{n}\left[loss_{XG}\left(y_i, y_i^{k-1} + g_i F_k(x_i)\right) + \frac{1}{2}h_i f_k^2(x_i)\right] + \sum_{i}^{k} \Omega\,(f_i) + \ Constant \quad (14)$$

Where $g_i = \partial_{yi}\ (k-1)Loss_{XG}(y_i, y_i^{k-1})$

$$h_i = \partial_{yi}^2(k-1)Loss_{XG}\left(y_i, y_i^{k-1}\right) \quad (15)$$

Equation (16) represents the new O^kafter the constant function has been eliminated.

$$O^k = \sum_{i=1}^{n}\left[g_i F_k(x_i)_{kj}) + \frac{1}{2}h_i f_k^2(x_i)\right] + \sum_{i}^{k}\Omega(f_i) \quad (16)$$

Refinement of tree mentioned in the equation (17);

$$f_k(m) = w_{q(x)}(w\epsilon R^k, q{:}\,R^d \rightarrow 1,2,2.\dots\dots K) \quad (17)$$

$q_{(x)} = \ sample\ at\ leaf\ node\,.\ w_{q(x)}$ = score value at leaf node

$\Omega(f_k)$ is redefined as the equation (18)

$$\Omega(f_k) = \ \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{k} w_j^2 \quad (18)$$

Value of $\Omega(f_k)$ put in equation (19 & 20)

$$O^k = \sum_{i=1}^{n}\left[g_i F_k(x_i)) \ + \frac{1}{2}h_i f_k^2\ (x_i)\right] + \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{k} w_j^2 \quad (19)$$

$$\sum_{j=1}^{k}\left[\left(\sum_{i\epsilon I_l} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\epsilon I_j} h_i + \lambda\right)w_j^2\right] + \gamma M \quad (20)$$

Calculate for optimal w_j^* in XG Boost algorithm by using the equation (21)

$$w_i^* = -\frac{\sum_{i\epsilon I_j} g_i}{\sum_{i\epsilon I_j} h_i + \lambda} \quad (21)$$

New O^k is obtained after applying Eqn (22)

$$O^k = -\frac{1}{2}\sum_{j=1}^{k}\frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_j} h_i + \lambda} + \gamma M \quad (22)$$

$G_j = \sum_{i\epsilon I_j} g_i$ and $H_j = \sum_{i\epsilon I_j} h_i$ O^kas in the equation (23)

$$O^k = \sum_{j=1}^{k}\left[G_j w_j + \frac{1}{2}(H_j + \lambda)w^2\right] + \gamma M \quad (23)$$

Set $w_j^* = -\frac{G_j^2}{H_{j+\lambda}}$ then modified objective function in the equation (24)

$$O^k = -\frac{1}{2}\sum_{j=1}^{k}\frac{G_j^2}{H_j + \lambda} + \gamma M \quad (24)$$

XGBoost remains a go-to algorithm for competitive and production-level machine learning tasks, thanks to its speed, flexibility, and predictive power.

2.3. Selection Attribute Using Enhanced Gradient Boosting (GB) Machine (EGBM)

This section provides a description of the developed CP-EGBM and an overview of the GBM.

EGBM:

In order to create a strong learner, a group of weak learners (WL) is compiled by the GBM by concentrating on the error that arises at every repetition [18].

To represent training examples, let $D = \{x_n, y_n\}_{n=1}^{N}$.

In this situation, the purpose of GB is to get the best compute $F(x)$ of an approximation function $F^*(x)$.

For each training example distribution in equation (25) to minimise the predicted value of a given (LF) loss function $L(y, F(x))$.

$F^*(x)$ translates the instances x_n to y_n.

$$F^*(x) = \ argmin_{F(x)} L_{x,y}(y, F(x)) \quad (25)$$

For classification tasks, GBM estimates the approximation function $Ly, F(x) = y - F(x)$ using a logistic loss function. By minimising the LF over each boosting step, EGBM strengthens prediction performance by first fitting each WL to rectify the errors caused by the preceding WL, $F(x)$, which is typically a constant value. The local minimum proportional finds the local minimum at each stage by taking steps to the negative gradient of the LF. The gradient direction of the LF at the i[th] boosting step can be found using equation (26).

$$r_{i,n} = -\left|\frac{\partial L(y_n, F(x_n))}{\partial F(x_n)}\right|_{F(x)=F_{i-1(x)}} \tag{26}$$

As weak-learners, regression trees with parameter an are typically parameterised functions of the input variables x, where ∂ denotes the partial derivative. EGBM expands the range of calculations for the gradient in these cases. The following equation (27) can be solved to obtain the tree:

$$a_i = argmin_{a,\beta} \sum_{n=1}^{N} \left[r_{i,n} - \beta h(x_n, a)\right]^2 \tag{27}$$

Here, the weight value, or the WL's expansion coefficient, is denoted by β.

Then, $a_i$is a parameter that is acquired at iteration I.
At every iteration i with t = 1 to the number of iterations T, the model $F_i(x)$ is updated after the optimal length pi has been established. It is presented in the following steps 5 and 6, Algorithm 1 describes EGBM in detail.

Algorithm 1: Training Process of State- of- the-art EGBM Model

Input: Training dataset $D = \{x_n, y_n\}_{n=1}^{N}$, the B- highest boosting stages

Output: GBM $F_i(x)$

1. $F_O(x) = argmin_p \sum_{n=1}^{N} L(y_n, p)$
2. For m=1to B do
3. $r_{i,n} = -\left|\frac{\partial L(y_n, F(x_n))}{\partial F(x_n)}\right|_{f(x)=F_{i-1(x)}}$
4. $a_i = argmin_{a,\beta} \sum_{n=1}^{N} \left[r_{i,n} - \beta h(x_n, a)\right]^2$
5. $p_i = argmin_p \sum_{n=1}^{N} L\left(y_n, F_{i-1}(x_n) + ph(x_n, a_i)\right)$
6. $F_i(x) = F_{i-1}(x) + p_i h(x, a_i)$
7. End for

2.4. Customer Churn Analysis Using LSTM AND GRU Neural Network (NN)for Analysis

Because of the ability to retain long-term dependencies, Recurrent NN (RNN), were first employed for language models. Nevertheless, RNN gradients may vanished when they unfold into extremely deep feed forward neural networks, as time lags increase [19]. Certain RNN structures, such LSTM and GRU, were suggested with forget units to address the GVP. It is intended to enable the memory cells to choose the best time lags by enabling them to decide when to forget certain information. The LSTM and GRU NN structures are shown in the remaining portion of this section.

A. LSTM

Figure 1 shows the typical cell configuration of the LSTM NN. Input gate (i_t), input modulation gate, forget gate (f_t), and output gates (O_t)are the 4 main gates that make up a typical LSTM NN cell. An i_t processes newly arriving data by accepting a new input point from the outside. In the final cycle, the memory cell input gate receives input from the LSTM NN cell's output [20]. The f_t chooses the ideal time lag for the input sequence by determining when to forget the output results. The O_t generates output for the LSTM NN cell after calculating all the results. In language models, the end outcome of the NN is typically found by adding a soft-max (SM) layer.
Consider $X = (x_1, x_2, \dots. X_n)$ for the input time series and $H = (h_1, h_2, \dots. h_n)$ for the hidden state of the memory cells. $Y = (y_1, y_2, \dots. y_n)$ represents the output time series. LSTM NN use the following formula to perform the computation (28-29).

$$h_t = H(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \tag{28}$$

$$p_t = W_{hy}y_{t-1} + b_y \tag{29}$$

Here, bias vectors are represented by b and weight matrices by W. The following formula (30–34) calculates the memory cells' hidden state:

$$i_t = \sigma(W_{ix}x_t + W_{hh}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{30}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{hh}h_{t-1} + W_{fc}c_{t-1} + b_f\right) \tag{31}$$

$$c_t = f_t * c_{t-1} + i_t * g(W_{cx}x_t + W_{hh}h_{t-1} + W_{cc}c_{t-1} + b_c) \tag{32}$$

$$O_t = \sigma(W_{ox}x_t + W_{hh}h_{t-1} + W_{oc}c_{t-1} + b_o) \tag{33}$$

$$h_t = o_t * h(c_t) \tag{34}$$

Here, the scalar product of two vectors or matrices can be denoted as *. Then, g and h are the extends of the regular sigmoid function (SF) with the range shifting to [-2, 2] and [-1, 1], and σ represents the conventional SF defined in Equation (35).

$$\sigma(x) = \frac{1}{1 + e^x} \tag{35}$$

The following Eqn (36) uses square LF for OF.

$$e = \sum_{t=1}^{n} (y_t - p_t)^2 \tag{36}$$

Here, p is the predicted traffic flow and y is the real output. Back propagation through time (BPTT) uses the Adam optimiser (AO), a variant of the stochastic gradient descent (SGD) optimiser with adaptive learning rates, to reduce training error while avoiding local minimal points. NN are particularly prone to overfitting and are renowned for their extraordinary expressive capabilities. To lessen overfitting, numerous regularisation techniques have been suggested, yet NN training has always been a challenging problem. It has been challenging to apply dropout to language models of RNNs, because of the recurring nature of these networks. Dropout techniques have reportedly been effectively used with RNNs up to this point [13].

b. GRU

Though easier to calculate and apply, GRU is comparable to LSTM [21]. Fig. 2 depicts the usual structure of GRU cells.
A typical GRU cell consists of two gates: reset gate r and update gate z. Equation (37), which presents the input time series value at time t and the hidden state of time t-1, are used to calculate the hidden state output at time t, just like in an LSTM cell.

$$h_t = f(h_{t-1}, x_t) \tag{37}$$

Reset gates and LSTM forget gates serve a similar purpose. Researchers won't get into the specific formula because GRU NNs and LSTM NNs share a number of similarities. Researchers employ the same optimisation strategy and regression section for GRU NNs as they do for LSTM NN in this study.

3. Result and Discussion

To evaluate the performance of the PRNN model alongside existing churn prediction models, these models are trained using a dataset containing customer information collected over a six-month period in the telecommunications industry. Each customer is classified into one of two predefined categories, with their churn propensity continuously monitored and updated based on their most recent three-month data. This approach effectively mirrors the real-world scenario of CP. A CP system's effectiveness is assessed by its ability to identify churners for targeted marketing strategies. The PRNN prediction framework is thoroughly assessed in this study using the Receiver Operating Characteristic (ROC) curve.

3.1. DATA COLLECTION

This research utilizes a dataset from a telecommunications organization, representing a segment of mobile customers. The dataset, obtained from the IBM Watson Analytics Telco Customer Churn data at https://community.watsonanalytics.com/predictive-insights-in-the-telco-customer-churn-data-set/, offers valuable insights into customer behavior to aid retention strategies. The telecommunications company is facing challenges with customers switching from their landline services to competitors offering cable services. The goal is to identify which customers are leaving and the reasons behind their departure. Details of the customer's account, including duration, kind of contract, mode of payment, electronic bill payment, cost per month, and the entire cost, the services each customer subscribes to (e.g., phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies), and customers who churned within the last month (shown in the Churn column) are all included in the dataset.

3.2. EVALUATION CRITERIAS

Table 1. Confusion matrix

No of samples (N)	Predicted: Non-churner	Predicted: churner
Actual: Non-churner	TN	FP
Actual: churner	FN	TP

Another name for the confusion matrix (CM) is the error matrix. CM are a certain table arrangement that makes it possible to visualise how well an algorithm performs, typically in supervised learning (SL) (it is typically referred to as a matching matrix in unsupervised learning). The instances in a predicted class are represented by each column of the matrix, while the occurrences in an actual class are represented by each row. In this work, the accuracy of the predictive models is measured using the area under the receiver curve (AUC), sensitivity, and specificity [39]. The ease with which it is possible to ascertain whether the system is confusing two classes has become the source of the term.

With two dimensions ("actual" and "predicted") and identical sets of "classes" in both, it is a unique type of contingency table. Every possible combination of dimension and class is a variable in the contingency. TN stands for True Negatives, FP for False Positives, FN for False Negatives, and TP for True Positives in the CM table 1.

Specificity: The proportion of negative cases which are predicted to be negative [38].

$$\text{Specificity} = \frac{TN}{(TN + FP)} \tag{38}$$

Sensitivity: The proportion of positive cases which are predicted to be positive [39].

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \tag{39}$$

ROC curve

The performance of a binary classifier system as its discrimination threshold is changed is depicted graphically by ROC curve [40], which is used in statistics. The TP Rate (TPR) (y) is plotted against the FP Rate (FPR) (x) at different threshold values to form the curve. Sensitivity is another name for the TPR in ML. One way to compute the FPR, sometimes called the likelihood of false alarm, is to calculate 1− specificity. The sensitivity as a function of fall-out is therefore represented by the ROC curve.

x=1− specificity, y=sensitivity

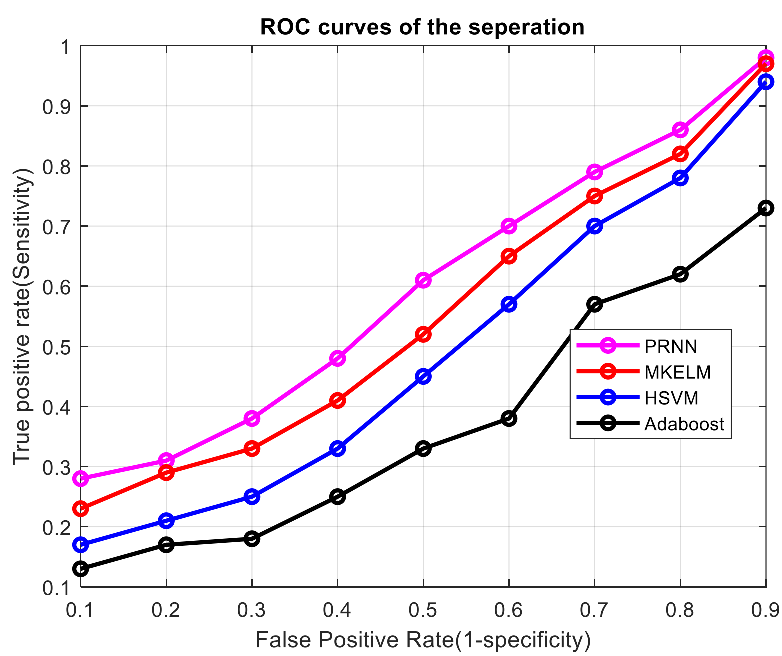

Figure 2. ROC curves of the separation

Figure 2 represents that the graph illustrates the Receiver Operating Characteristic (ROC) curves for four models: Adaboost, HSVM, MKLEM, and PRNN, in a churn prediction task. The analysis of the framework's efficiency is provided by the ROC curve, which displays the TPR (sensitivity) versus the FPR (1-specificity).Among the models, PRNN demonstrates the highest sensitivity across all false positive rates, indicating superior predictive accuracy. MKLEM and HSVM also perform well, with Adaboost showing the lowest sensitivity, suggesting it is the least effective model. The closer a curve approaches the top-left corner, the better the model's performance, making PRNN the best performer in this comparison.

Processing Time Comparison

Compared to the current HSVM and AdaBoost prediction models, which are displayed in Fig. 4, the PRNN prediction model finds frequent CP and significantly higher efficiency. In comparison to the current system, the suggested PRNN prediction model requires less computation time for CCP.

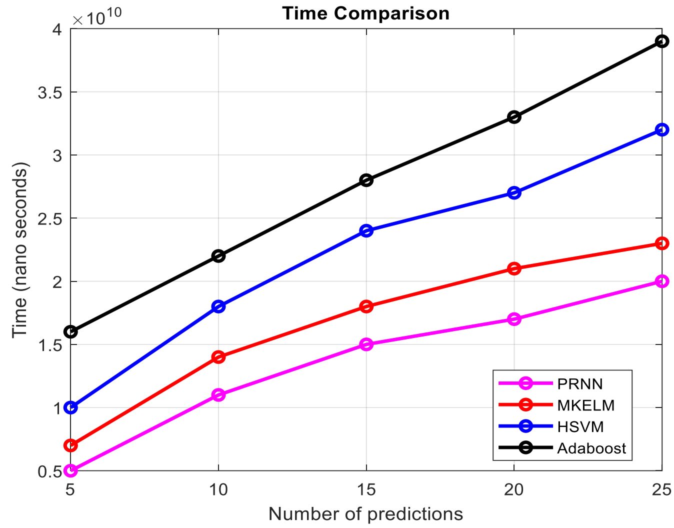

Figure 3. Processing time comparison

Figure 3 shows that the graph compares the processing times of four models—Adaboost, HSVM, MKLEM, and PRNN—across different numbers of analyses. Adaboost consistently has the highest processing time, indicating it is computationally the most expensive model. PRNN, on the other hand, exhibits the lowest processing time across all analysis scales, making it the most efficient model. HSVM and MKLEM fall in between, with MKLEM being slightly faster than HSVM in most cases. The trends suggest that PRNN is well-suited for applications requiring lower computational overhead.

Accuracy Comparison

As illustrated in Figure 5, the PRNN prediction model outperforms the current AdaBoost and LR prediction models in terms of accuracy and frequent CP. The accuracy of the outcome rises as the number of predictions rises. By increasing the Area Under the Receiver Curve, the suggested PRNN outperforms the current system in terms of accuracy rate. Reports indicate that PRNN is more effective than HSVM and AdaBoost, and it frequently yields higher accuracy rates.

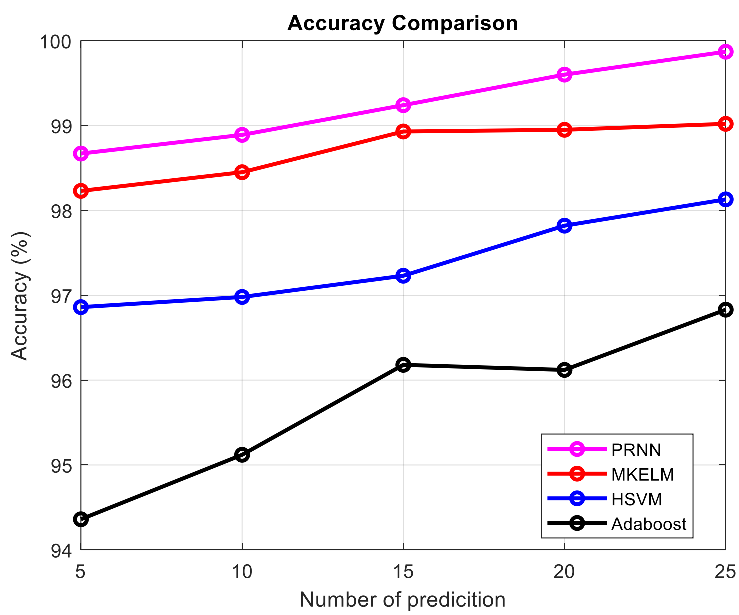

Figure 4. Accuracy comparison

Figure 4 illustrates that the graph compares the accuracy of four models—Adaboost, HSVM, MKLEM, and PRNN—across varying numbers of analyses. PRNN consistently demonstrates the highest accuracy, nearing 100%, making it the most reliable model for churn prediction. MKLEM and HSVM show competitive performance, with MKLEM slightly outperforming HSVM as the number of analyses increases. Adaboost exhibits the lowest accuracy, although it improves gradually with more analyses. Overall, PRNN stands out as the most accurate model, followed closely by MKLEM and HSVM.

CONCLUSION

In this study, proposed a hybrid framework integrating Kernel Fuzzy C-Means (KFCM) clustering and proposed Recurrent Neural Networks (PRNN) for customer churn analysis in the telecom industry. The KFCM algorithm effectively segmented customers into distinct clusters, capturing underlying patterns in customer behavior, while the PRNN model leveraged temporal data to predict churn with high accuracy. The combined approach demonstrated superior performance compared to traditional methods, achieving enhanced precision in identifying high-risk churners and facilitating targeted marketing interventions. This framework not only improves CP accuracy but also provides valuable insights for customer retention strategies, ensuring better resource allocation and higher profitability for telecom operators. Future work can extend this approach by incorporating advanced optimization techniques and additional data sources for further enhancement.

REFERENCES

- Ferrer-Estévez, M. and Chalmeta, R., 2023, Sustainable customer relationship management. *Marketing Intelligence & Planning*, vol.*41, no.*2, pp.244-262.
- Fernando, E., Sutomo, R., Prabowo, Y.D., Gatc, J. and Winanti, W., 2023, Exploring customer relationship management: Trends, challenges, and innovations. *Journal of Information Systems and Informatics*, vol.*5*, no.3, pp.984-1001.
- Srinivasan, R., Rajeswari, D. and Elangovan, G., 2023, Customer churn prediction using machine learning approaches. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pp. 1-6.
- Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., 2022, Customer churn prediction system: a machine learning approach. *Computing*, vol.*104, no.*2, pp.271-294.
- Manzoor, A., Qureshi, M.A., Kidney, E. and Longo, L., 2024, A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners, vo.12, pp.70434-70463.
- Al-Najjar, D., Al-Rousan, N. and Al-Najjar, H., 2022, Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and applied electronic commerce research*, vol.*17*, no.4, pp.1529-1542.
- Jiang, P., Liu, Z., Abedin, M.Z., Wang, J., Yang, W. and Dong, Q., 2024, Profit-driven weighted classifier with interpretable ability for customer churn prediction. *Omega*, vol.*125*, p.103034.
- Saha, S., Saha, C., Haque, M.M., Alam, M.G.R. and Talukder, A., 2024, ChurnNet: Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry. *IEEE Access*.
- Amin, A., Adnan, A. and Anwar, S., 2023, An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. *Applied Soft Computing*, vol.*137*, p.110103.
- Bilal, S.F., Almazroi, A.A., Bashir, S., Khan, F.H. and Almazroi, A.A., 2022, An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Computer Science*, vol.*8*, p. e854.
- Zhang, T., Moro, S. and Ramos, R.F., 2022, A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet*, vol.*14*, no.3, p.94.
- Pustokhina, I.V., Pustokhin, D.A., Nguyen, P.T., Elhoseny, M. and Shankar, K., 2021, Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. *Complex & Intelligent Systems*, pp.1-13.
- Khalid, L.F., Abdulazeez, A.M., Zeebaree, D.Q., Ahmed, F.Y. and Zebari, D.A., 2021, Customer churn prediction in telecommunications industry based on data mining. In *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, pp. 1-6.
- Sana, J.K., Abedin, M.Z., Rahman, M.S. and Rahman, M.S., 2022, A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *Plos one*, vol.*17*, no.12, p. e0278095.
- Nalatissifa, H. and Pardede, H.F., 2021, Customer decision prediction using deep neural network on telco customer churn data. *Jurnal Elektronika Dan Telekomunikasi*, vol.*21*, no.2, pp.122-127.
- Munusamy, S. and Murugesan, P., 2020, Modified dynamic fuzzy c-means clustering algorithm–Application in dynamic customer segmentation. *Applied Intelligence*, vol.*50*, no.6, pp.1922-1942.
- Marcus, L., 2023, Predicting user churn using temporal information: Early detection of churning users with machine learning using log-level data from a MedTech application.
- AlShourbaji, I., Helian, N., Sun, Y., Hussien, A.G., Abualigah, L. and Elnaim, B., 2023, An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. *Scientific Reports*, vol.*13*, no.1, p.14441.
- Fu, R., Zhang, Z. and Li, L., 2016, Using LSTM and GRU neural network methods for traffic flow prediction. In *2016 31st Youth academic annual conference of Chinese association of automation (YAC)*, pp. 324-328.
- Kumar, M.R., Priyanga, S., Anusha, J.S., Chatiyode, V., Santiago, J. and Chaudhary, D., 2024, Enhancing Telecommunications Customer Retention: A Deep Learning Approach Using LSTM for Predictive Churn Analysis. In *2024 International Conference on Data Science and Network Security (ICDSNS)*, pp. 01-07.
- Abualkishik, A.Z., Almajed, R. and Thompson, W., 2023, Intelligent Model for Customer Churn Prediction using Deep Learning Optimization Algorithms. *Journal of Intelligent Systems & Internet of Things*, vol.*8*, no.1.