International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Novel Stochastic Gradient Descent Based Logistic Regression (SGD-LR) Framework for Customer Churn Prediction

[1]Dr. Suja Sundram, [2]Dr. D. Poornima, [3]Dr. Praveenkumar G. D., [4]Mr. C. Balakumar, [5]Dr. D. Sasikala, [6]Sardor Omonov

Submitted: 25/12/2023 **Revised**: 27/01/2024 **Accepted**: 05/02/2024

Abstract: Customer churn is a crucial issue in any company or organization, and it describes the loss of customers as a result of them switching to competitors. When there is an opportunity to discover customer churn sooner, the organization may take steps such as providing important knowledge for keeping and boosting the client count. Deep Learning (DL) models have recently gained popularity because to their remarkable performance boost in a variety of fields. In this paper, a DL-based Customer Churn Prediction (CCP) is introduced using Stochastic Gradient Descent Based Logistic Regression (SGD-LR) with an LR classifier model. Effective categorization may be achieved by combining SGD with LR. The provided SGD-LR model is evaluated against a benchmark dataset, with the outcomes examined over a range of epochs. Furthermore, a comparison study with the outcomes of existing approaches is conducted. The implementation results demonstrated that the provided SGD-LR model outperformed the current CCP models on the same dataset.

Keywords: *Customer Churn, Prediction, Logistic Regression, Stochastic Gradient Descent, Deep Learning.*

1. Introduction

Customer churn is a key aspect of business, and it is defined as the loss of consumers as a part of them migrating to a competitor. Being able to predict turnover rates ahead of time gives a business a good reputation in terms of providing and developing its client base. As a result, the telecom company has produced numerous attempts to predict churning subscribers before they actually quit a service. Predicting telecom churners has usually attracted the attentions of researchers, and as a consequence, many have studied in this area to forecast telecom customer churn. For telecom services, forecasting attrition is a tricky problem. Hence, they have a large number of customers.

Churn occurs in a company as a result of customer dissatisfaction. A few parameters are essential to detect consumer disappointment. A client is unlikely to churn as a result of a single disappointment. Before a customer completely stops doing business with a company, there are usually a few occurrences of disappointment. The associations make note of a few properties relevant to the customers and their technique of communication with the association.

Churn relates to the existing and potential phenomenon of consumer turn over. Churn activities pertaining to the loyalty techniques used by stakeholders to control churn. The purpose of churn management is to keep profitable customers engaged with the company Churn is a notion that appears in all industries, although it is particularly prominent in subscription-based organizations such as Internet service providers, digital television, and telecommunication. During the 1990s and 2000s, business managers and researchers tried to find a way to use consumer data to reduce time and improve revenues, and also use data to find useful, hidden business patterns.

Berson.et al., (2000) discovered that a 'client beat' is established as the cycle of supporters (either pre-paid or post-paid) moving from one service provider to another. Stir can be dynamic/conscious, pivot, or aloof/unwilling [1]. We can minimize the weakness to beat and enhance the benefit of the organization by efficiently maintaining clients. A mechanism should be put into effect to investigate productivity credit. Agitate Prediction could also be characterized as a technique for pinpointing churners ahead of time [2].

Predictive analytics is typically engaged with projecting how a client will act one day based on earlier behavior. One example of predictive analytics is predicting

[1]*Assistant Professor, Department of Business Administration, Jubail Industrial College, Kingdom of Saudi Arabia, sundrams@rcjy.edu.sa, ORCID:0000-0001-6668-2233*

[2]*Assistant Professor, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu, India. Orcid::https://orcid.org/0000-0002-7380-8882. Email: poorniramesh2011@gmail.com*

[3]*Assistant Professor, Department of Computer Technology -UG Kongu Engineering College- Erode, Tamilnadu, India. Email: erodegd@gmail.com*

[4]*Assistant Professor, Department of Computer Applications, Faculty of Arts, Science, Commerce and Management, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India. Email: balakumar.cbk@gmail.com*

[5]*Assistant Professor, Department of Computer Science, Sri Vasavi College, Erode, Tamilnadu, India. Email: sasikalad@gmail.com*

[6]*The Department of Corporate Finance and Securities, Tashkent Institute of Finance, Tashkent, Uzbekistan, Email: s_omonov@tfi.uz. Orchid: 0000-0003-2679-6599*

customers who are more likely to churn. With the purpose of provide consumer-degree elements that indicate the possibility that a consumer would execute a unique action, predictive analytics is used to inspect client Relationship management (CRM) data and DM. The arrangements are still mostly connected to revenue, advertising, and client preservation [3].

There is a myriad of units that can be used to determine the difference between churners and non-churners in an organization. Ordinary items or strategy (RA and DT) along with tender calculating methods will be separated into these units (FL and NN). The target of process modeling is to forecast what will likely happen in the future based on previous events [4].

Long-term clients are also more valuable since they are less likely to be recruited by other competitors, are more likely to consider new customers, and are less costly to visit. As a consequence, even a little increase in customer retention can have a major impact on a telecommunications company's growth and long-term sustainability. As a result, a strong churn prediction model should have excellent interpretation capabilities, which are critical for identifying churn clients and identifying the root cause of churn.

Customer churn does have a direct impact on a gross profitability, making it a significant challenge for service-based businesses. Finding a suitable procedure to lower the churn rate becomes essential, but most businesses put a greater priority on customer retention. They are prepared to spend and reduce profits in order to make the customers away. Because acquiring new clients is difficult, any service firm's priority is on retaining existing customers. Reduced turnover rate also contributes to lengthy business growth. The cause of churn should be recognized ahead of time to avoid churn.

Machine learning entails developing algorithms that can learn from current data and make predictions based on historical evidence. A design is produced by using input data and then used to generate predictions on new data. Machine learning is now becoming extremely relevant as the data that is collected grows exponentially each day. During last decade, churn prediction has received a lot of attention.

Data from the past is indeed the cornerstone for predicting future consumer turnover. Customers' data that had already churned (reaction) and their characteristics / behaviour (predictors) before the churn were carefully investigated. By developing a statistical model that links predictors to reaction, and trying to predict response for current customers. Just in case you required another buzzing term, this method belongs to the supervised learning category [5].

The minority class has fewer instances in the data set, culminating in skewed classifier training due to the overwhelming presence of majority class occurrences. Telecom businesses also collect a lot of information on their clients, such as invoices, payments, call history, demographics, and so on, resulting in a large number of features in the dataset to account for. As a result, the predictors for identifying churners suffer from the feature space and an imbalanced distribution of the telecom datasets.

As a consequence, a churn prediction technique that can successfully mitigate the telecom dataset's imbalanced nature and high - dimensional concerns is very desirable. Customer dissatisfaction is the primary cause of churn. Several factors must be considered while determining consumer dissatisfaction. A consumer doesn't really usually churn as a result of a single unhappiness event [6-7]. Before a customer totally stops doing business with a company; there are frequently multiple instances of discontent. The businesses keep track of several properties related with the customer and their style of activity with the company.

Over the last decade, churn prediction has attracted a lot of attention. Several strategies, strategies, and methods have been suggested, most of which rely on both static and dynamic analysis. Due to advancements in machine learning technology, competitive companies' marketing strategies have shifted from being product-oriented to becoming consumer [8].

The ultimate goal of this research work is to create a comprehensive solution that includes a database and a model for predicting churn. The accuracy and complexity of a churn prediction model are desirable qualities. Churn prediction is a continuous process that occurs in a changing corporate environment, rather than a one-time effort.

As a result, a churn prediction solution must have the following properties:

❖ It has a high degree of accuracy in predicting churn, allowing it to collect as many churners as feasible.

❖ It can assign a churn likelihood score to each client, indicating the likelihood that the client will leave in the near future.

❖ It has a reasonable implementation time, allowing for regular model updates and predictions.

❖ It can be seamlessly integrated into the business's regular operations.

❖ It is simple to improve and update to reflect any changes in the market environment.

DL models have recently gained popularity because to their remarkable performance boost in a variety of fields.

In this paper, a DL-based CCP is introduced using SGD with an LR classifier model. Effective categorization may be achieved by combining SGD with LR. The provided Stochastic Gradient Descent Based Logistic Regression (SGD-LR) (SGD-LR) model is evaluated against a benchmark dataset, with the outcomes examined over a range of epochs. Furthermore, a comparison study with the outcomes of existing approaches is conducted. The implementation results demonstrated that the provided SGD-LR model outperformed the current CCP models on the same dataset.

2. Related Works

Customer Relationship Management (CRM) is a relationship that aids in the collection of consumer data, which the organization then uses to meet client needs [8].CRM apparatuses have been developed to improve and examine client acquisition and upkeep, as well as to aid in the vision display and organization of various coherent tasks [9]. Information mining plays an important role in media transmission companies' showcasing efforts, detecting deception, and effectively managing their media transmission networks [10].

Due to the rapid production of a massive amount of data, significant market competition, and an increase in the agitate rate, information mining procedures are used in broadcast communications for CRM [11]. CRM devices can be used to increase client acquisition and maintenance by enhancing benefit and supporting logical errands [11]. As a result of the hidden information in telecom ventures, a great deal of degree has advanced its approach for analysts to investigate the information and introduce the entire data for advancing their firm.

The authors' epic model [12] shows a half breed model that combines a modified K-means bunching computation with an example principle inductive technique (FOIL) for anticipating client beat behaviour. An examination was never truly based on six different tactics. These included hybrid strategies such as k-NN-LR and SePI, as well as unique k means, choice tree, computed relapse, PART, SVM, KNN, and One R.

The suggested methodology outperformed all six classifiers, half breed models, and benchmark datasets by a factor of ten. The normal AUC esteems (estimation of expected precision) for each arrangement approach were then calculated, with the half breed model having the most extreme normal worth.

The experts discovered that decision tree computations outperform neural organisations when using the Gain Measure as their assessment model. Furthermore, they stated that a combination of Decision Tree calculations was required to enhance the model exhibition. The authors [13] explained how to use grouping choice tree approaches for beat analysis in the media transmission business.

The author of [14] predicted a churn catastrophe in telecommunication using a model based on generic programming and Ada Boost. Two standard datasets were used to test this model. Orange and cell 2 cell datasets had 89 percent and 63 percent accuracy, respectively, for cell 2 cell and other datasets. The author predicted a client attrition issue in big data platforms in [15].

The authors of [16] employed a fuzzy classifiers algorithm to forecast customer turnover in the telecommunications industry. They employed a real-world dataset with 722 variables and 600,000 subscribers from a South Asian telecommunications business. That was the first research they knew of that used a fuzzy classification method to forecast turnover in the telecommunications industry. The authors identified the most essential 84 variables based on domain knowledge and feature selection.

The model was tested on engineering and selecting criteria. The model's performance was demonstrated using the AUC curve. By extracting Social Network Analysis (SNA) attributes, a consumer social network prediction model was also created [17]. To forecast customer turnover, a new hybrid model was constructed combining logistic regression and the Nave Bayes approach, and it was proven that the hybrid method outperforms individual application of classification methods [18].

Many service companies now run retention campaigns to motivate consumers and advise cost-cutting strategies. These ads reduced churn by increasing the number of potential consumers, leading in service firm recommendations [19]. To forecast churning from call pattern changes, a churn prediction technique has been developed for mobile communications providers. The proposed technique uses a multi classifier to develop a model for churn prediction using data from a one-month interval.

The model's predictive power was proved in the results [20]. The majority of churn prediction models are shown to suggest customer retention behaviour but do not explain why churn occurs.

By taking organisational competitiveness strategy into account, a new model has been proposed. The churn predictor was modelled using factor analysis [21]. In service sectors, customer churn is more valuable. Gradient boosting and weighted random forest techniques were used to create a unified framework to deal with the imbalance in churn prediction, and the results were impressive. On publicly available data sets, a survey of six sample approaches and four rule-generation algorithms was conducted.

In comparison to other sampling strategies and rule-generation algorithms, the mega- trend diffusion function on genetic algorithm performed well. To measure the turnover rate of Telecom customers, a Rough Set Theory-based categorization model was presented. When compared to previous models, the results reveal a considerable improvement in accuracy [22-24]. Deep learning techniques can handle enormous data sets, uncover hidden patterns, and predict patterns relevant to the telecom sector's underlying hazards.

3. Proposed Methodology

Figure 1 shows the processes involved in the SGD-LR model. At the earlier stage, a setof three preprocessing steps were carried out namely format conversion, data transformation and data normalization. Then, sample selection and 10-fold cross validation process takes place. Subsequently, data classification is performed using SGD-LR. Effective categorization may be achieved by combining SGD with LR. The provided SGD-LR model is evaluated against a benchmark dataset, with the outcomes examined over a range of epochs.

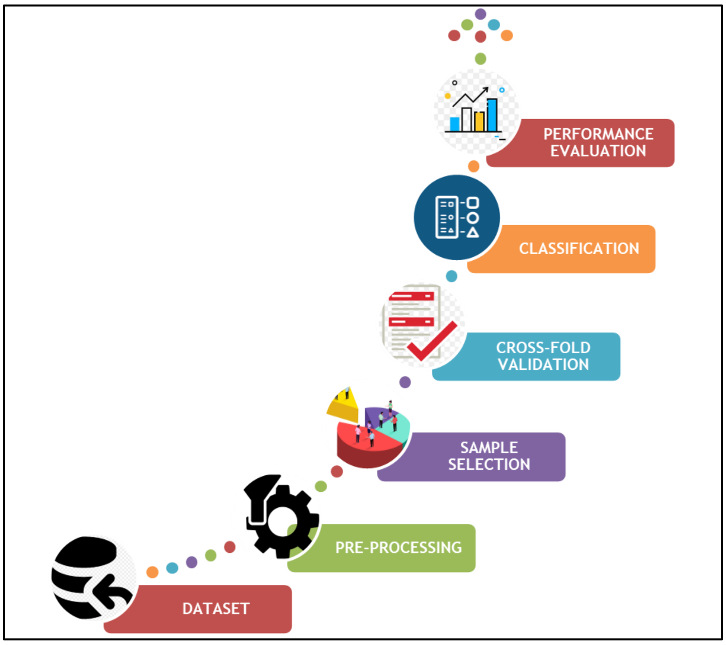


Fig 1. Overall Process of SGD-LR Method

Stochastic Gradient Descent (SGD)

SGD is a popular ML as well as DL oriented optimizing model. SGD is a productive method which requires an individual monitoring that is stored in memory storage. Forexample, assume a dataset that comprises of numerous instances. Usually, SGD counts more number of observations for all iterations. However, in SGD, variable calculationis improved with the help of single observation simultaneously and the application to evaluate web learning, that has new observations in an effective manner.

Every z sample is a pair (x, y) consisting of a random input x and a scalar result y. In this case, the loss function P(y, y) is used to assess the cost of detecting y if the right answer isy, and it picks a family F of functions ffww(x) characterized by a weight vector w. The function ff F can reduce the loss function QQ(z, w) = P(ffww(x), y) determined on instances. Regardless of the unknown distribution dP(z), the Laws of Nature are always defined on a sample z1... zn.

$$E(f) = \int l(f(x),y)dP(z)E_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(x_i),y_i\right) \text{ ---------- (1)}$$

The empirical risk En(ff) demonstrates the training set function. As a result, the targeted risk E(ff) computes the anticipated generalizing operation on each succeeding occurrence. According to statistical learning theory, decreasing the empirical risk rather than the projected danger if a selected family F is constrained.

Typically, it is shown to use Gradient Descent (GD) to lower the empirical risk En(ffww). Each iteration increases the weights w based on a gradient.

$$w_{t+1} = w_t - \gamma\frac{1}{n}\sum_{i=1}^{n}\nabla_w Q(z_i,w_t) \text{---------- (2)}$$

where $\gamma\gamma$ denotes a suitably chosen learning rate. In the case of sufficient regularity functions, if the primary estimate w_0 is close to the optimal value, and in the case of a low learning rate $\gamma\gamma$, this approach achieves linear convergence, which is, log t, where denotes the remaining error.

A better optimizing strategy might be built by replacing the scalar learning value $\gamma\gamma$ with an optimistic finite matrix Γt that finds the disparity of the Hessian function with a minimal cost:

$$w_{t+1} = w_t - \tau_t\frac{1}{n}\sum_{i=1}^{n}\nabla_w Q(z_i,w_t) \text{---------- (3)}$$

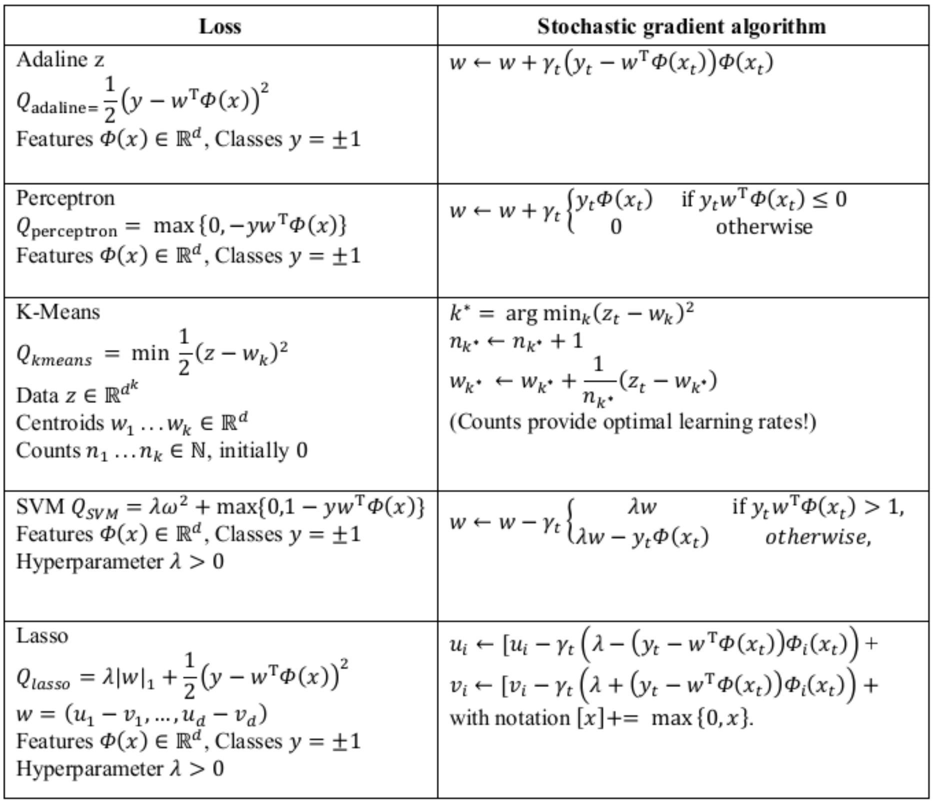
The widely used Newton model is an application of the second order GD (2GD). With properly optimizing regularity considerations, which is supplied as w0 is close to the optimum, 2GD achieves quadratic convergence. If the cost is quadratic and the transformation matrix is correct, the approach achieves a maximum value in only one iteration. Otherwise, with adequate smoothness, these yields $-loglog\ \rho \sim t$.

The SGD approach is a significant simplification approach. Each successive iteration calculates the gradient by substituting the GD ofEn(ffww) with a single valuezt:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t) \text{---------- (4)}$$

Because the SGD model must remember the instance to be approached from previous iterations, it handles fly instances in a designed fashion. As the samples are obtained from a ground truth distribution, an SGD is optimized with the desired issue. Table 1 depicts an SGD strategy for conventional ML techniques. It is mostly used for Perceptron, Adaline, and k-Means mapping. Traditional optimizing models were used to define the SVM and Lasso. A hyper-parameter controls the regularization term in both QQ_{svm} and QQ_{lasso}. Because QQ_{means} is a non-convex function, the K-means technique converges for local minimum. Furthermore, the projected update rule incorporates 2GD learning values, ensuring speedy convergence. Using the SGD rule for these parameters and ensuring the positivity tends to sparser solutions.

Table 1. Stochastic Gradient Algorithms



Loss	Stochastic gradient algorithm
Adaline z $Q_{\text{adaline=}}\frac{1}{2}\left(y - w^{\mathrm{T}}\Phi(x)\right)^2$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$	$w \leftarrow w + \gamma_t\left(y_t - w^{\mathrm{T}}\Phi(x_t)\right)\Phi(x_t)$
Perceptron $Q_{\text{perceptron}} = \max\{0, -yw^{\mathrm{T}}\Phi(x)\}$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$	$w \leftarrow w + \gamma_t \begin{cases} y_t\Phi(x_t) & \text{if } y_t w^{\mathrm{T}}\Phi(x_t) \leq 0 \\ 0 & \text{otherwise} \end{cases}$
K-Means $Q_{kmeans} = \min \frac{1}{2}(z - w_k)^2$ Data $z \in \mathbb{R}^{d^k}$ Centroids $w_1 \ldots w_k \in \mathbb{R}^d$ Counts $n_1 \ldots n_k \in \mathbb{N}$, initially 0	$k^* = \arg\min_k (z_t - w_k)^2$ $n_{k^*} \leftarrow n_{k^*} + 1$ $w_{k^*} \leftarrow w_{k^*} + \frac{1}{n_{k^*}}(z_t - w_{k^*})$ (Counts provide optimal learning rates!)
SVM $Q_{SVM} = \lambda\omega^2 + \max\{0,1 - yw^{\mathrm{T}}\Phi(x)\}$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$ Hyperparameter $\lambda > 0$	$w \leftarrow w - \gamma_t \begin{cases} \lambda w & \text{if } y_t w^{\mathrm{T}}\Phi(x_t) > 1, \\ \lambda w - y_t\Phi(x_t) & otherwise, \end{cases}$
Lasso $Q_{lasso} = \lambda\lvert w\rvert_1 + \frac{1}{2}\left(y - w^{\mathrm{T}}\Phi(x)\right)^2$ $w = (u_1 - v_1, \ldots, u_d - v_d)$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$ Hyperparameter $\lambda > 0$	$u_i \leftarrow [u_i - \gamma_t\left(\lambda - \left(y_t - w^{\mathrm{T}}\Phi(x_t)\right)\Phi_i(x_t)\right) +$ $v_i \leftarrow [v_i - \gamma_t\left(\lambda + \left(y_t - w^{\mathrm{T}}\Phi(x_t)\right)\Phi_i(x_t)\right) +$ with notation $[x]+= \max\{0, x\}$.

The convergence of SGD is widely studied in a stochastic approximation work. Converging outcome often acquires minimized learning values to satisfy the constraints of $\sum_t \gamma\gamma^2 < \infty$ and$\sum_t \gamma\gamma_t < \infty$. The Robbins-Siegmund theorem to implement almost assuring convergence with unexpected conditions, such as loss function is$t$non-smooth.

The converging speed of SGD is reduced by a noisy approximation of positive gradient. If the learning value gets reduced, then the variance of a parameter estimate w_t is gradually minimized. In case of decreased learning rates, the expectation of a variable estimate w_t consumes longer duration to attain the optimal solution.

- If the Hessian matrix of a cost function is optimums which are conditionally positive, the best converging speed can be attained with the application of learning values $\gamma\gamma_t \sim t^{-1}$. The desire of remaining error is further reduced with same speed, where, D(Б) $\sim t^{-1}$. Such theoretical converging values are observed frequently.
- Generally, the functions of D(ρ) $\sim t^{-1/2}$ are converged. Experimentally, the convergence can be demonstrated at the time of final stage while processing optimization task.

It is not considered as main factor as the optimization process is terminated before attaining the required option.

Second order stochastic gradient descent (2SGD) combines the gradients through positive finite matrix Γ_t which seeks for inverse of the Hessian:

$$w_{t+1} = w_t - \gamma_t \tau_t \nabla_w Q(z_t, w_t) \text{---------- (5)}$$

Unexpectedly, the variation cannot minimize the stochastic noise and no enhancement ofw_t. Though constants are increased, expectation of remaining error again reduces oft^{-1}, (i.e.), D(ρ) $\sim t^{-1}$ at best. As a result, the optimizing model of DGS is gradually slower when compared with general batch method.

In LR is a classification approach employed in several applications, like biomedical studies, commercial and economy, criminology, ecology, engineering, healthcare, andso on. LR belongs to the class of methods named as generalized linear techniques. The main goal of generalized linear methods in case of binary oriented parameters as well as LR techniques for continuous variable helps to evaluate the regression functionwhich compares the required value of dependent parameter y more than one predictorvalues.

LR is a representation of discriminative classifier which learns the direct mapping frominput x to result y by designing a posterior probability $P(y \mid x)$ directly. The parametric method presented by LR is described in the following. Among other models, the simple function can denote the LR as sigmoid function provided by Eqn. (6).

$$\sigma(x) = \frac{1}{1+e^{-z}} \text{---------- (6)}$$

Then, it is described as loss function which reveals the 0-1 losses for a technique.

$$Loss_{\frac{0}{1}}(z) = \begin{cases} 1, & if\ z < 0 \\ 0, & otherwise \end{cases} \text{---------- (7)}$$

Assume $y\epsilon\{-1, 1\}$ as well as$z = y.\, w^T x$. Where z implies positive if$y$ and $w^T x$ consistof similar sign, otherwise, negative symbol.

$$p(y = -1|x) = \frac{1}{1+\exp\,(w_o + \sum_{i=1}^{d} w_i x_i)} \text{---------- (8)}$$

$$p(y = 1|x) = 1 - P(y = -1|x) \text{---------- (9)}$$

The major operation of LR is to reduce$w$, thus the maximum value of $0 - 1$ loss isreduced than training points.

$$min \sum_{i=1}^{n} l_{\frac{0}{1}} \,(y^{(i)} \cdot w^T \cdot x^i) \text{---------- (10)}$$

$$w = [w_0, w_1, w_2, \dots., w_d] \leftarrow$$
$$\arg \underset{w}{max} \prod_k P\,(y^{(k)}|x^{(k)}, w) \text{---------- (11)}$$

According to the plotting of 0/1 loss function, the regression method is converted as logistic function where the measures differ from 0 to 1 as z is from −∞ to +∞.

$$l_{log}(z) = \log\,(1 + e^{-z}) \text{---------- (12)}$$

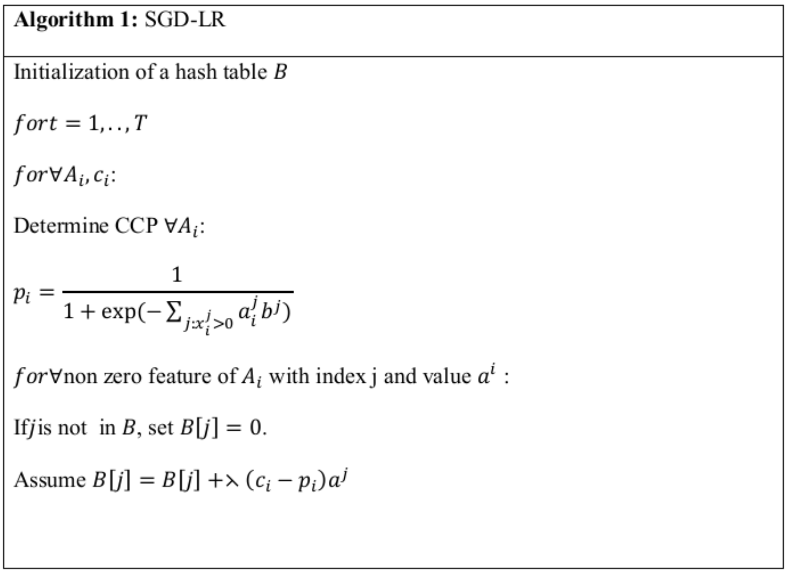
Furthermore, weight w with the application of GD rule. The main purpose of developing LR is to handle continuous features; it is capable of dealing with nominal feature as well as missing values in an effective manner. Fig. 5.2 shows the allocationof logistic losses which occurs frequently.

By including regularization with learning method eliminates over-fitting by rejecting the irregular features from data set. Mostly, regularization can be attained on the basis of $L1$ and $L2$ which results in sparseness of minimizing the difficulty.

Regularization oriented LR learns mapping (w) which decreases the logistic loss on training data including regularization norm. To perform regularization in LR, a greater number of likelihood functions is used as formulated in Eq. 13.

$$\min_w \sum_{i=1}^{n} l_{log}\left(y^{(i)} \cdot w^T \cdot x^i\right) + \lambda||w||_2^2 \text{---------- (13)}$$

Eq. 13 has 2 units namely, training log-loss function as well as model difficulty. The λ obtained from model complexity is a regularizing attribute. It computes the w variables to be inflated. With the application of Eq. 13 as a cost function, the resultof a hypothesis can decrease the over-fitting. When w is selected as large value, it smoothens and leads to under-fitting. By frequently observing the $L1$ regularization, several methods result in minimizing the variables to 0, and provides the simulation outcome of a parameter vector has to be sparse.



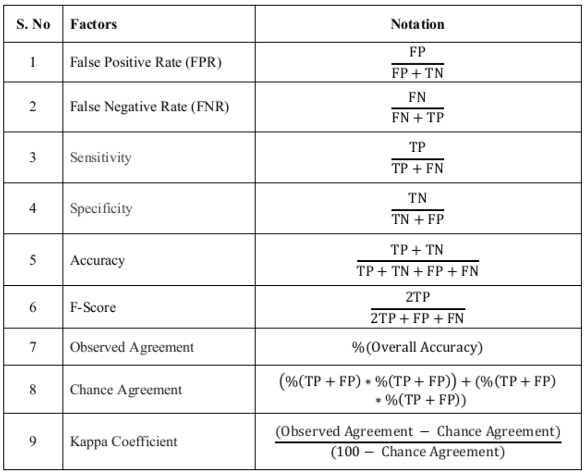
Algorithm 1: SGD-LR

```
Initialization of a hash table $B$
$fort = 1, . . , T$
$for \forall A_i, c_i$:
Determine CCP $\forall A_i$:
$p_i = \frac{1}{1 + \exp(-\sum_{j:x_i^j > 0} a_i^j b^j)}$
$for \forall$non zero feature of $A_i$ with index j and value $a^i$ :
If$j$is not  in $B$, set $B[j] = 0$.
Assume $B[j] = B[j] + ⋋ (c_i - p_i)a^j$
```

Experimental Results and Performance Evaluation

A comprehensive study was performed on datasets 1 and 2 to validate the proposed method's performance. FPR, FNR, sensitivity, specificity, accuracy, F-score, and Kappa are the metrics used to analyze the findings. The set of performance indicators used to evaluate the results of the suggested models is shown in Table 2.

Table 2. Performance Evaluation Metrics



S. No	Factors	Notation
1	False Positive Rate (FPR)	$\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
2	False Negative Rate (FNR)	$\frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$
3	Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
4	Specificity	$\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$
5	Accuracy	$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
6	F-Score	$\frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$
7	Observed Agreement	$\%(\text{Overall Accuracy})$
8	Chance Agreement	$(\%(\mathrm{TP} + \mathrm{FP}) * \%(\mathrm{TP} + \mathrm{FP})) + (\%(\mathrm{TP} + \mathrm{FP}) * \%(\mathrm{TP} + \mathrm{FP}))$
9	Kappa Coefficient	$\frac{(\text{Observed Agreement} - \text{Chance Agreement})}{(100 - \text{Chance Agreement})}$

Results Analysis on Dataset 1

An experiment is conducted on the applied dataset by varying the number of epochs interms of 100, 200, 300, 400 and 500 as provided in Table 3. Under the presence of 100, 200, 300, 400 and 500 epochs are a similar outcome, the total of 2850 instances are correctly classified as churn and a number of 0 instances are properly classified as non-churn.

Table 3. Confusion Matrix of Different Runs by SGD based DNN for Dataset 1

	Epochs-100		Epochs-200		Epochs-300		Epochs-400		Epochs-500	
Experts	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn
Churn	2850	0	2850	0	2850	0	2850	0	2850	0
Non-Churn	483	0	483	0	483	0	483	0	483	0

Next, the classification results are derived from the confusion matrix and the results are tabulated in terms of several measures namely FPR, FNR, sensitivity, specificity, accuracy, F-score and kappa value. Here, the values of the evaluation parameters FPRand FNR should be low. At the same time, the values of sensitivity, specificity, accuracy, F-score and kappa values should be high. The classification outcome attained under the existence of different number of epochsis presented in Table 4 and Figure 3. From the table and figure, it is noticed that the FNR of 14.49, sensitivity if 85.50%, accuracy of 85.51% and F-score value of 92.19% is attained under the presence of 100 epochs. Similarly, under the presence of 200, 300, 400 and 500 epochsare presence that the same classifiers results.

Table 4. Performance Evaluation of Different Runs on Dataset 1

No. of Runs	FPR	FNR	Sensitivity (%)	Specificity (%)	Accuracy (%)	F-Score (%)	Kappa (%)
Epochs-100	35.01	14.49	85.5	65.99	86.51	92.19	38.07
Epochs-200	34.04	15.5	84.6	65.95	82.6	92.19	43.72
Epochs-300	33.62	14.49	85.5	66.38	87.51	92.19	43.54
Epochs-400	33.02	14.49	80.6	66.97	89.68	92.19	45.36
Epochs-500	32.44	16.5	85.5	67.56	85.51	92.19	45.84

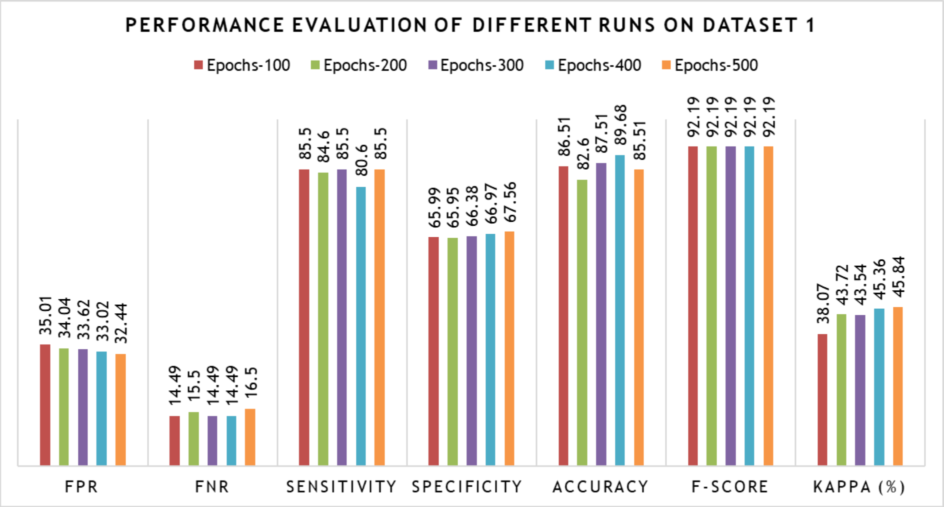


Fig 2. Performance Evaluation of Different Runs on Dataset 1

Results Analysis on Dataset 2

An experiment is conducted on the applied dataset by varying the number of epochs interms of 100, 200, 300, 400 and 500 as provided in Table 5. Under the presence of 100 epochs, a total of 4722 instances are correctly classified as churn and a number of877 instances are properly classified as non-churn. Likewise, under the presence of 200 epochs, a total of 4697 instances are correctly classified as churns and a number of 924 instances are properly classified as non-churn. Similarly, under the presence of300 epochs, a total of 4698 instances are correctly classified as churns and a number of 940 instances are properly classified as non-churn. In the same way, under the presence of 400 epochs, a total of 4704 instances are correctly classified as churns anda number of 953 instances are properly classified as non-churn. Finally, under the presence of 500 epochs, a total of 4715 instances are correctly classified as churns anda number of 956 instances are properly classified as non-churn. From these values, it can be observed that maximum classifier outcome is exhibited under the increasing number of epochs.

Table 5. Confusion Matrix of Different Runs by SGD-LR for Dataset 2

Experts	**Epochs-100**		**Epochs-200**		**Epochs-300**		**Epochs-400**		**Epochs-500**	
	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn	Churn	Non-Churn
Churn	4722	452	4697	477	4698	476	4704	470	4715	459
Non-Churn	992	877	945	924	929	940	916	953	913	956

The classification outcome attained under the existence of different number of epochsis presented in Table 6 and Figure 3. From the table, it is noticed that the FPR of 34.01, FNR of 17.36, sensitivity if 82.63%, specificity of 65.99%, accuracy of 79.49%, F-score of 86.74%, Kappa value of 42.07% is attained under the presence of 100 epochs. Similarly, under the presence of 200 epochs, it is noted that the FPR of 34.04,FNR of 16.74, sensitivity if 83.25%, specificity of 65.95%, accuracy of 79.81%, F-score of 86.85%, Kappa value of 43.72% respectively. In the same way, a FPR of 33.62, FNR of 16.51, sensitivity if 83.49%, specificity of 66.38%, accuracy of 80.05%, F-score of 86.99%, Kappa value of 44.54% is attained under the presence of 300 epochs. Similarly, under the presence of 400 epochs, it is noted that the FPR of 33.02,FNR of 16.29, sensitivity if 83.70%, specificity of 66.97%, accuracy of 80.32, F-score of 87.15, Kappa value of 45.36% respectively. Finally, under the presence of 500 epochs, it is interesting that the FPR of 32.44, FNR of 16.22, sensitivity if 83.78%, specificity of 67.56, accuracy of 80.52%, F-score of 87.29%, Kappa value of 45.84% respectively. From these values, it is noticeable that a maximum accuracy of 80.52% is attained under the presence of 500 epochs implying that the classifier results are increased with an increase in number of epochs.

Table 6. Performance Evaluation of Different Runs on Applied Dataset 2

No. of Runs	**FPR**	**FNR**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**F-Score (%)**	**Kappa (%)**
Epochs-100	34.01	17.36	82.63	65.99	79.49	86.74	42.07
Epochs-200	34.04	16.74	83.25	65.95	79.81	86.85	43.72
Epochs-300	33.62	16.51	83.49	66.38	80.05	86.99	44.54
Epochs-400	33.02	16.29	83.70	66.97	80.32	87.15	45.36
Epochs-500	32.44	16.22	83.78	67.56	80.52	87.29	45.84

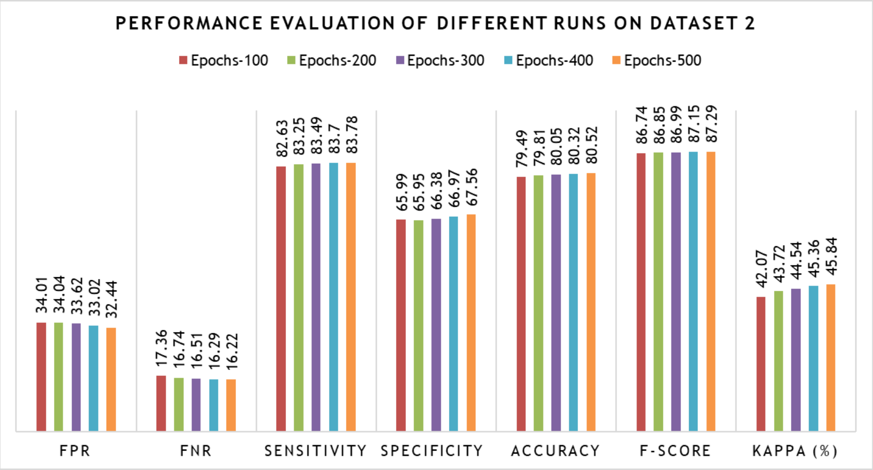


Fig 3. Performance Evaluation of Different Runs on Dataset 2

Comparison with Existing Methods for Applied Datasets

Table 7 showcases the comparison of previous models for using dataset 1, and dataset 2by means of Accuracy and F-Measure. Followed by, a comparative analysis with traditional methods on employed dataset 1 in terms of accuracy and F-Measure is demonstrated in the Table 7 and in Fig. 4. The rates provided in table points that the existing technique attains high accuracy of 84.00% and F-Measure of 54.29%. But, the SGD-LR approach offers additional classification task by generatingthe accuracy of 85.51% and F-measure of 92.19% correspondingly. Hence, the abovedetailed experimental analysis validates that the SGD-LR approach is a proper classification tool for CCP.

Table 7. Comparison of Proposed Method with Existing Methods for Applied Datasets

Methods	Dataset-1		Dataset-2	
	Accuracy(%)	F-Measure (%)	Accuracy(%)	F-Measure (%)
SGD-LR	90.50	92.19	80.52	87.29
LR	84.00	57.89	74.00	61.76
SVM	84.00	54.29	78.00	66.15
Naïve Bayes	85.33	54.17	77.00	66.99
Decision Tree	84.75	55.47	75.00	64.29
KNN	85.40	56.29	74.20	63.04
XGBoost	84.67	54.90	74.83	62.53
RF	84.86	57.60	74.43	62.00
GBM	84.63	58.02	74.13	61.31

On the applied dataset 2, the measures given in the table pointed out that the previous models attain a higher accuracy of 78.00% and F-Measure of 66.15%. Therefore, the SGD-LR technique exhibits outstanding classification performance with the accuracy of 90.52% and F-measure of 87.29% correspondingly.As a result, the provided experimental analysis observed that the SGD-LR frameworkis an optimal classification tool for CCP.

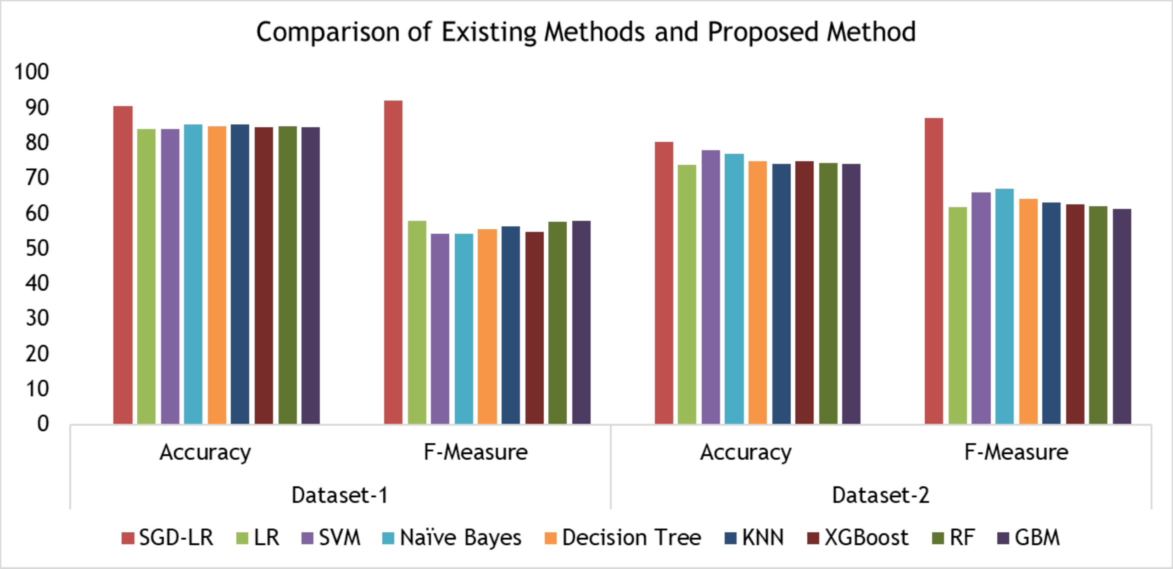


Fig 4. Comparison of Proposed Method with Existing Methods for Applied Datasets

In this objective, the SGD with LR classifier model is applied for the CCP. By the integration of SGD and LR, effective classification can be accomplished. The presented SGD-LR model is tested against a benchmark dataset and the results are investigated under varying number of epochs. The presented model shows extraordinary classification performance with the accuracy of 85.51%, 80.52%, 94.53% and F-measure of 92.19%, 87.29%, and 95.09% for applied three datasets respectively. The above detailed experimental analysis verified that the presented model is an appropriate classification tool for CCP. In the next objective, a set of different MRI models have been applied for CCP.

4. Conclusion

This article has addressed the research methods used to estimate customer churns in a four-fold manner. The general system structure was created and detailed with several levels. For the CCP, the SGD with LR classifier model is used in this study. Effective categorization may be achieved by combining SGD with LR. The provided SGD-LR model is evaluated against a benchmark dataset, with the outcomes examined over a range of epochs. The provided model performs exceptionally well in classification, with accuracy of 85.51%, 80.52%, and 94.53% and F-measures of 92.19%, 87.29%, and 95.09% for the three datasets used. The extensive experimental study described above confirmed that the presented model is an adequate CCP classification tool.

References

[1] Jagadeesan, A.P. Bank customer retention prediction and customer ranking based on deep neural networks. Int. J. Sci. Dev. Res.2020, 5, 444–449.

[2] Amuda, K.A.; Adeyemo, A.B. Customers churn prediction in financial institution using artificial neural network. arXiv 2019,arXiv:1912.11346.

[3] Kim, S.; Shin, K.-S.; Park, K. An application of support vector machines for customer churn analysis: Credit card case. In Proceedings of the International Conference on Natural Computation, Changsha, China, 27–29 August 2005; Springer: Berlin/Heidelberg, Germany; pp. 636–647.

[4] Kumar, D.A.; Ravi, V. Predicting credit card customer churn in banks using data mining. Int. J. Data Anal. Tech. Strateg. 2008,1, 4–28.

[5] Keramati, A.; Ghaneei, H.; Mirmohammadi, S.M. Developing a prediction model for customer churn from electronic banking services using data mining. Financ. Innov. 2016, 2, 10.

[6] Bastan, M.; Akbarpour, S.; Ahmadvand, A. Business dynamics of iranian commercial banks. In Proceedings of the 34th International Conference of the System Dynamics Society, Delft, The Netherlands, 17–21 July 2016.

[7] Bastan, M.; Bagheri Mazrae, M.; Ahmadvand, A. Dynamics of banking soundness based on CAMELS Rating system. In Proceedings of the 34th International Conference of the System Dynamics Society, Delft, The Netherlands, 17–21 July 2016.

[8] Iranmanesh, S.H.; Hamid, M.; Bastan, M.; Hamed Shakouri, G.; Nasiri, M.M. Customer churn prediction using artificial neural network: An analytical CRM application. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Bangkok, Thailand, 5–7 March 2019; pp. 23–26.

[9] Domingos, E.; Ojeme, B.; Daramola, O. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. Computation 2021, 9, 34.

[10] Chen, S.C.; Huang, M.Y. Constructing credit auditing and control & management model with data mining technique. Expert Syst. Appl. 2011, 38, 5359–5365.

[11] Pankajavalli, P. B., & Karthick, G. S. (2022). An Independent Constructive Multi-class Classification Algorithm for Predicting the Risk Level of Stress Using Multi-modal Data. Arabian Journal for Science and Engineering, 47(8), 10547-10562.

[12] Risselada, H.; Verhoef, P.C.; Bijmolt, T.H. Staying power of churn prediction models. J. Interact. Mark. 2010, 24, 198–208.

[13] Kim, H.S.; Yoon, C.H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommun. Policy 2004, 28, 751–765.

[14] Xia, G.; He, Q. The research of online shopping customer churn prediction based on integrated learning. In Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018), Qingdao, China, 30–31 March 2018; pp. 30–31.

[15] Olaniyi, A.S.; Olaolu, A.M.; Jimada-Ojuolape, B.; Kayode, S.Y. Customer churn prediction in banking industry using K-means and support vector machine algorithms. Int. J. Multidiscip. Sci. Adv. Technol. 2020, 1, 48–54.

[16] Pankajavalli, P. B., & Karthick, G. S. (Eds.). (2019). Incorporating the Internet of Things in Healthcare Applications and Wearable Devices. IGI Global.

[17] Seng, J.L.; Chen, T.C. An analytic approach to select data mining for business decision. Expert Syst. Appl. 2010, 37, 8042–8057.

[18] Karthick, G. S. (2023). Energy-Aware Reliable Medium Access Control Protocol for Energy-Efficient and Reliable Data Communication in Wireless Sensor Networks. SN Computer Science, 4(5), 449.

[19] Rahman, M.; Kumar, V. Machine learning based customer churn prediction in banking. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; IEEE:

[20] Piscataway, NJ, USA; pp. 1196–1201.

[21] 20. Karthick, G. S., & Pankajavalli, P. B. (2020). A review on human healthcare internet of things: a technical perspective. SN Computer Science, 1(4), 198.

[22] Miguéis, V.L.; Van den Poel, D.; Camanho, A.S.; e Cunha, J.F. Modeling partial customer churn: On the value of first productcategory purchase sequences. Expert Syst. Appl. 2012, 39, 11250–11256.

[23] Kolajo, T.; Adeyemo, A.B. Data Mining technique for predicting telecommunications industry customer churn using both descriptive and predictive algorithms. Comput. Inf. Syst. Dev. Inform. J. 2012, 3, 27–34.

[24] Kaya, E.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B. Behavioral attributes and financial churn prediction. EPJ Data Sci. 2018, 7, 41.

[25] Miao, X.; Wang, H. Customer churn prediction on credit card services using random forest method. In Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022), Online, 14–16 January 2022; Atlantis Press: Paris, France; pp. 649–656.