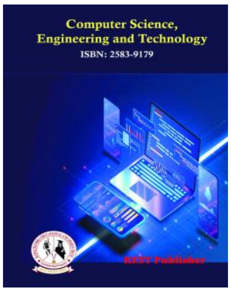
Computer Science, Engineering and Technology
Vol: 1(1), March 2023
REST Publisher; ISSN: 2583-9179 (Online)
Website: https://restpublisher.com/journals/cset/
DOI: https://doi.org/10.46632/cset/1/1/12



Identifying Customer Churn for E-Commerce Using Ensemble Learning Method

* S. G. Balakrishnan, R. Janaki, S. Geetha, C. Vaikunda Perumal
Mahendra Engineering College, Namakkal, Tamil Nadu, India.
**Corresponding Author Email: balakrishnansg@mahendra.info*

Abstract: *Organizations in the E-Commerce (E-Comm) sector, particularly those belonging to the business-to-consumer category are in strong fight for their survival, attempting to acquire accessibility with their competitors' client groups while preventing existing customers from leaving. The expense of obtaining novel customers is increasing as additional competitors enter the market with large initial investments as well as cutting-edge penetration methods, making client retention critical for these businesses. The most appropriate course for proceeding in this situation is to recognize potential churning clients and avert churn using temporary maintenance tactics. Understanding wherefore the consumer opted for the route is also necessary in order to implement specific win-back methods. The e-commerce corporation keeps every client's records on file comprising searches done, purchase fascinated, purchase frequency, reviews left, feedback provided, and additional data. The combination of Machine Learning (ML) with clickstream data analysis has transformed E-Comm platforms. In the E-Comm context, this study provides a comprehensive examination of the ML usage to estimate client purchase decisions, identify engagement variables, and improve user interfaces. The data used in this research employs ML to assess client behavior in E-Comm was obtained from a variety of sources among 2004 and September 2023. These models provide as the foundational learners in an ensemble structure. A weighted-average ensemble approach integrates them, yielding an impressive 99.84% accuracy on a reserved test dataset. This strategy is more promising than previous approaches for detecting client attrition. In marketing, company relationships are critical for transactions over the internet.*

Keywords: *E-Commerce, customer churn, customer satisfaction, weighted average ensemble method, machine learning*

1. INTRODUCTION

E-comm has completely changed the retail scene, ushering in a new era of accessibility and ease [1]. Customers may explore a broad range of items, comparing prices, as well as making purchases with just a few clicks or taps without ever having to leave the comforts of their homes [2]. But as the digital market gets more crowded, companies have to contend with the difficult challenge of drawing in and keeping discriminating customers in the face of intense competition. Understanding and forecasting customer behaviors, preferences, and motives is crucial to meeting this challenge, and traditional market study techniques frequently fall short in this regard. The rise of the internet retail company has heightened industry contest. Because the E-Comm industry has such a high client churn rate, business managers must look into measures to reduce churn in online purchases. Because the customer's behavior is predictable, it is possible to forecast the customer's future trading preferences using the appropriate data and essential analysis. To reduce the amount of lost customers, business owners can identify consumers with a losing tendency and do the necessary pre-control activity. In the last decade, E-Comm industry research has focused heavily on predicting customer attrition for online purchases [3]. Customer churn is commonly referred as attrition that happens when a current customer discontinues their relationship with the business by not using the services it offers or buying their goods. Customers in an E-Comm setting might be classified as active, new, inactive, or churned. A company's sustainability is totally dependent on its ability to maintain customers involved for a prolonged term. Due to the industry's severe competition, acquiring new customers can be relatively costly. In the case of a new customer, a company's economic breakeven is often attained once the few transactions is done [4]. Active customers are the foundation of any E-Comm business, and companies ought to offer close attention to individuals whose might grow inactive as well as churn subsequently. Methodologies for predicting probable loss of customers include statistical methods and ML techniques [5].

Customer data acquired by B2C E-Comm enterprises is a valuable data source because it enables the analysis for various customer buying patterns. Churn prediction estimates the possibility of customer retention. It lowers the cost of gaining prospective customers while assisting with customer retention. It requires more advertising time as well as money for gaining a new customer than to retain an existing one. Customers that have been apprehensive about making a purchase or are prepared to transfer online stores owing to budgetary constraints might be persuaded as well as grasped. Customers are entitled to consistent and diverse product offers. Customers have the right to leave for important and unavoidable causes. Despite our investment in involuntary churners, the results are solid. Targeted advertising helps connect with as well as interact with customers. In a contractual arrangement, a relationship with customers implies both the business and the customer collaborate. The contract explicitly outlines both parties' respective rights and responsibilities. To take advantage of the right liberty once establishing a contract with the business, the client has to perform the essential responsibilities. Extending reductions, tailoring items to customer preferences, as well as sending out triggered emails are the strategies to dissuade volunteer churners. Concentrating primarily on volunteer churners will lower the cost of offering benefits to new as well as previously churned customers [6]. The prediction of client attrition with unorganized data gives inadequate outcomes as well as generates a class imbalance. To avoid these concerns, there needs to be significant variation between the percentages of churn and non-churn in the prior information. In these circumstances, an intriguing potential in improving e-commerce systems has emerged: the combination of clickstream data analysis with advanced ML methods, particularly with Large Language Models (LLMs) [7] [8]. Through the utilization of natural language processing and predictive analytics, language learning models (LLMs) provide previously unattainable approaching towards customer behavior which permits businesses to precisely customize their goods and user experiences [9]. The way customers interact with e-commerce platforms revolutionized by LLMs, as they forecast purchase choices, identify important engagement drivers, facilitate real-time interface adjustments, and nurture tailored and intuitive shopping experiences [10]. This study intends to fill in important knowledge gaps and identify promising directions for future research and innovation in addition to clarifying the existing status of the field. We hope to foster discussion and cooperation in the academic and industry communities by highlighting issues like interpretability, ethics, and multimodal integration [11]. This will help to advance e-commerce analytics and the field of machine learning in consumer behavior analysis as a whole. Customer data ought to be pre-processed, as well as feature selection ought to be applied for predicting important attributes. It spares both money and time on forecasts. Various classification and prediction methods may be employed to foresee client attrition. Efficient methods based on voting softly on accuracy have been demonstrated and the ensemble model is chosen as the best option for applying in future study.

2. LITERATURE REVIEW

Successfully forecasting for potential revenue is crucial for internet businesses' strategic decision-making. Companies are able to supervise their employees and improve their supply chain management process when they can predict sales for the E-Comm platform as well as comprehend their finances. Based on predicting sales, E-Comm platform using a sales prediction may make better educated judgments about pricing, promotional timing and inventory levels [12] [13]. Based on forecasting sales can reveal information on the establishment, decline of an E-Comm platform, stability as well as the influence of short-term product goals namely promotion, season, internet ranking and pricing on sales over the long term [14]. A variety of models are being investigated in the framework of ML applications toward customer churn prediction. For example, deep learning algorithms have demonstrated potential in increasing prediction accuracy in complicated e-commerce situations, especially when client behavior is extremely variable [15]. Studies emphasize the importance of correcting class imbalances, which might distort outcomes as well as impair the model's capacity to accurately recognize churners. Approaches such as SMOTE are widely used to address this problem, resulting in improved accuracy as well as recall in a range of uses [16]. This research majorly focuses on developing a conceptual framework to explore the influence of customer retention and relationship level on the customer maintenance procedure [17]. The current state of programmed management approaches, as illustrated by researchers is a type of suggestion administration. There have been several studies on recommender frameworks like the one which examined the efficacy of three ML methods are neural networks, relapse trees and relapse for predicting customer churn [18]. A mix of SMC method and naïve Bayesian estimates of consumer behavior on E-Comm websites. The novel Naive Bayesian model has outperformed the SMC method. They developed a three-stage method for analyzing customer turnover in telecoms, which included clustering analysis and DT. Bugajev et al. used six different categorization approaches to get the high accuracy over predicting customer attrition in the telecom sector. According to Alkhayrat et al., using ML in consumer strategy support, improving customer experience and profiling can assist businesses flourish in the sales and services of customer-centric [19]. However, ensemble learning technique assists in integrating multiple ML approaches, each of which learns from the same dataset to solve a single issue [20]. Thus, ensemble learning method is a familiar strategy for improving performance by combining a set of classifiers, which may involve two or more ML algorithms [21].

3. RESEARCH METHODOLOGY

The dataset for the present research was obtained from Kaggle and includes 5,597 records once cleanup of data. It provides a variety of features including client tenure, favorite login device, satisfaction ratings, and city tier. The goal variable is binary representing if a customer has churned '1' or not churned '0'. Table 1 illustrates the feature description of key features involved in this research whereas the data preprocessing involves with crucial stages for ensuring dataset quality and consistency. Numerical variable with missing values have been feed as an input through median but in the case of categorical variable has imputed using mode measurement. Thus, the outlier features such as order amount and hike from last year are eliminated by the quartile method that eliminates 33 extreme records.

TABLE 1. E-Commerce dataset key feature and its description

Key Feature	Feature description
Customer ID	Each customer identified with unique identifier
City Tier	Tier's of customer in the city
Preferred login device	Login device preferred by customer (Desktop, mobile, etc.)
Tenure	Number of year with the customer within the company
Warehouse to Home	Distance from warehouse to customer's house
Satisfaction Score	Customer satisfaction score (1 to 5)
Churn Status	Customer churn (1) and customer not churn (0)

To boost the accuracy of models, a variety of methods for feature engineering were used, include one-hot encoding for categorical characteristics including the introduction of additional interactions. Standard Scaling was used to standardize all numerical characteristics and reduce unit inconsistencies. Weighted-average ensembles are a form of ensemble ML methods which incorporates the forecasts of multiple independent models by taking a weighted average of the results they generate. The essential premise is that the ensemble approach incorporates the best features of multiple approaches while minimizing their faults.

Working of Weighted Average Ensemble Method (WAEM)

Based on the proposed ensemble learning model, this research focuses on stacking pattern with 'n' number of classifier are selected for ensemble classification. In this experimental research, three traditional ML classifier is implemented for modeling to predict the e-learning concept through IoT. The three classifier are CF1 as Random Forest (RF), CF2 as Naïve Bayes (NB) and the CF3 as Support Vector Classifier (SVC). The bagging based ensemble as staking through weighted approach of the above three classifier predictions have been accomplished through Mean (Average). However, the ensemble using staking (weight based) are concentrated in minimizing the correlation between 'n' estimators through ensemble model named WAEM and it classifies the features by random sampling instead of entire feature set. Hence, the mechanism of Weighted Average (WA)is measured through weights to specific features in accordance with significance of target variable as well as accomplishing the final result. Figure 1 illustrates the mechanism of WA based stacking as an ensemble model.

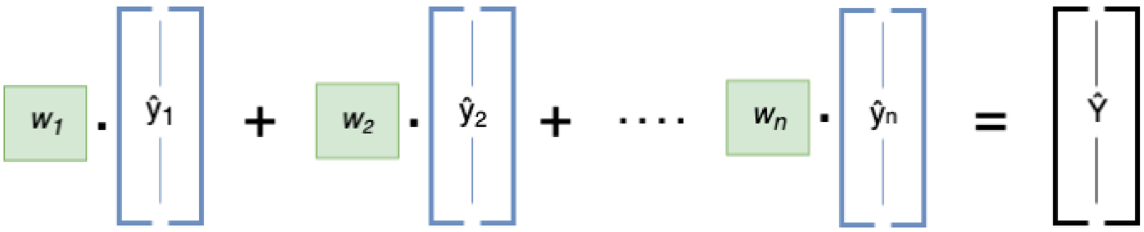


FIGURE 1. WAbased stacking as ensemble model

The technique of ensemble learning through stacking mechanism of various classifier ML model. The traditional algorithms are capable by an entire train data set in which the Meta model has learned by eventual findings of features using the complete base level model. Moreover, this experimental research has includes bagging as well as boosting technique in maintaining bias as well as variance which assisting in understanding the stacks that have increasing the predicting accuracy of ensemble model. Hence, the alternate of voting based classifier is stacking based WA by a blending level which helps to learn the obtained best aggregation of individual outcomes. Thus, the stacking mechanism of proposed WAEM is used to identify the specific weighted features with each traditional ML algorithm has resulted with complexness for regression. The WAEM has been illustrated in the equation 1.

$$WA_{CF} = w_1 * P_1 + w_2 * P_2 + w_3 * P_3 + \cdots + w_n * P_n \qquad (1)$$

Where,
WA_{CF} = Weighted average of Meta classifier model
P_n = Prediction of nth traditional ML classifier model
w_n = Weight of definite feature of nth traditional ML classifier model

As a result, the WAEM-assisted LR accurately predicted customer satisfaction which determined through customer churn status in e-commerce review from the respective product. The WAEM-based model is illustrated briefly for predicting the customer churn status in e-commerce business dataset.

WAEM algorithm

Step 1 – Initiate data preprocessing and normalizing the data of e-commerce dataset of e-learning and split into train and test.
Step 2 - Considering the data split to model the training data with various as well as distinct classifiers namely CF$_1$, CF$_2$ and CF$_3$.
Step 3 - Each traditional classifier has trained independently using similar hypotheses as well as represent as sub individual model and predicted the target in term of $\mathrm{P}_1, \mathrm{P}_2$ and P_3.
Step 4 - The prediction of classifier techniques are weighted namelyw_1, w_2 and w_3 from the individual classifier.
Step 5 - Stacking based WA of specific features correlated and results the classifier with the WAEM as WA_{CF}based on equation 1.
Step 6 - Ensemble model has resulted predictions with Meta classifier using LR as high accuracy.

In order to achieve an improved outcome when predicting a customer churn status as ‘Churn’ and ‘Not Churn’. By analyzing records, the WAEM has generated high accuracy through choosing the exact correlated featured identified from the various classifiers.

4. EXPERIMENTAL RESULT

Based on the experimental research, the e-commerce dataset involved 80 instances after wrangling data in which 80% dataset has measured the sample of training dataset and the enduring20% as random sample is considered to be testing dataset. However, the better prediction is determined by three other CF models as an individual classifier. The prediction of three individual classifiers has recognized and the parameter value of confusion matrix for SVC, NB, RF and the proposed WAEM with LR is illustrated in Table 2. Therefore, the confusion matrix of WAEM with LR is shown in figure 4and Jupiter notebook is used for carrying the implementation performance. The obtained result for evaluating the performance in WAEM with LR model has compared with confusion matrix parameter values of existing traditional classifier prediction.

TABLE 2. Confusion Matric value in WAEM with other existing classifier

ML classifier Method	**True Positive (TP)**	**True Negative (TN)**	**False Negative (FN)**	**False Positive (FP)**
RF	11	7	0	2
SVC	11	7	1	1
NB	11	6	1	2
WAEM with LR	13	6	0	1

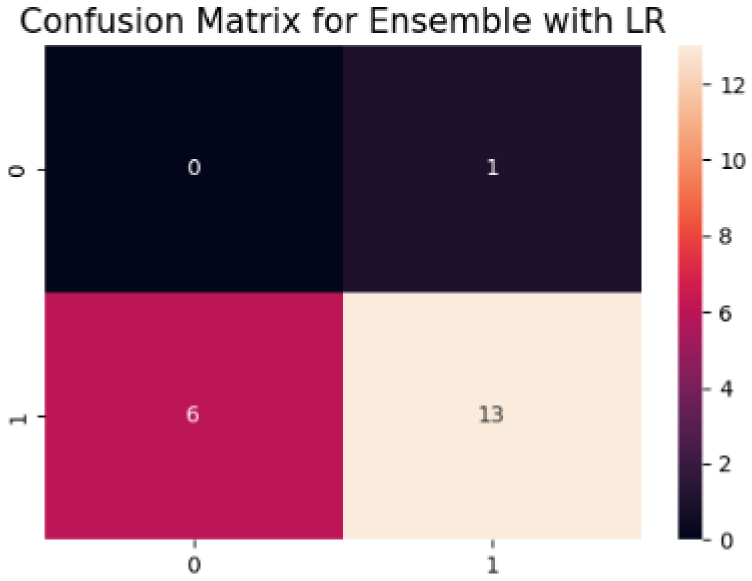


FIGURE 4. Ensemble with LR method confusion matrix

The performance of confusion matrix is assisted in evaluating the model effectiveness in performance of predicting e-commerce customer churn status effectively. The accuracy metric has estimated by correct prediction count from the overall customer responded for e-commerce review involved in online website of the respective e-commerce.

Accuracy (Acc)

The Acc is defined as the amount in exact prediction of customer churn from the overall review response from the e-commerce business dataset.

$$\text{Acc} = \frac{TN + TP}{TN + TP + FN + FP}$$

Precision

Precision is a measure about correctness presented for positive prediction of customer churn that perform better which illustrates positive prediction customer churn as resulted as an actually ‘churn’ as positive.

$$Precision = \frac{TP}{FP + TP}$$

Recall

It is defined as TP is predicted from all positive based customer churn present in the e-commerce business dataset which is also named as true positive rate.

$$Recall = \frac{TP}{TP + FN}$$

Moreover, the sensitivity and specificity represent TP rates and FP rates respectively and the suggested WAEM performs with high accuracy.

TABLE 3. Confusion matric metrics value for distinct ML methods

ML classification methods	Accuracy	Recall	Precision	Sensitivity	Specificity
NB	0.85	0.9167	0.8462	0.9167	0.750
RF	0.90	1.00	0.8462	1.00	0.778
SVM	0.90	0.9167	0.9167	0.9167	0.875
Ensemble LR	0.95	1.00	0.9286	1.00	0.857

Table 3 illustrates the WAEM algorithm with LR is performing the high prediction of customer churn from the e-commerce datasets. Generally, the WAEM with LR accuracy is relatively higher than particular conventional RF, NB and SVC methods. Moreover, the WAEM has generated high accuracy for predicting customer churn status precisely and it will assist the other customer to buy the product effectively. Hence, each model with WAEM algorithm with LR model is analyzed are shown in figure 5.

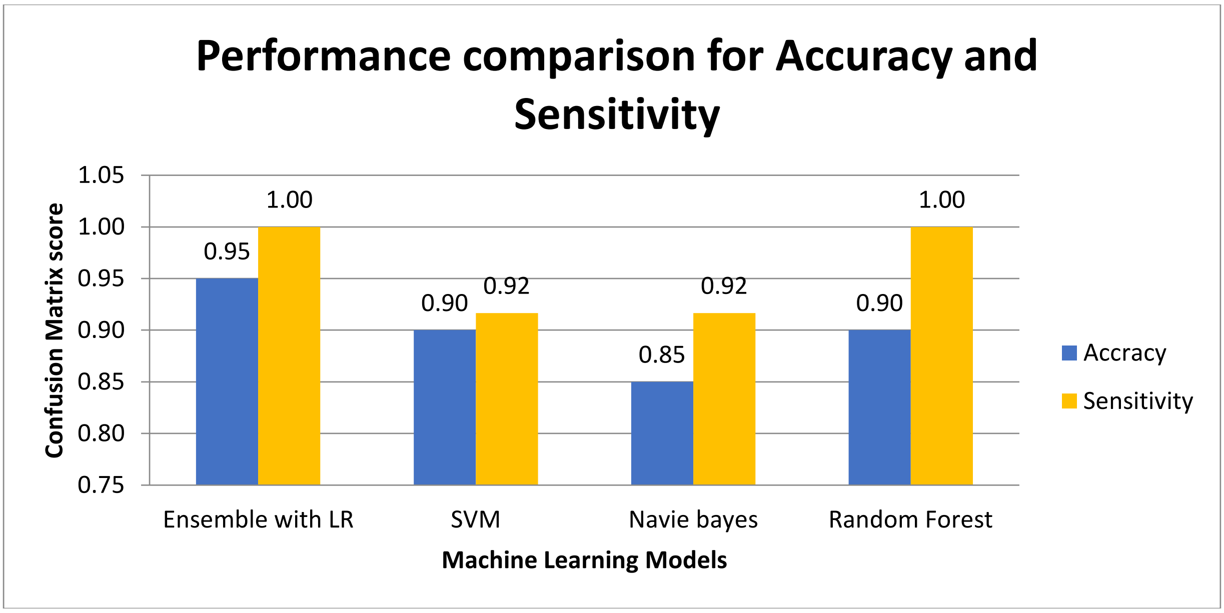


FIGURE 5. Performance comparison of customer churn status by accuracy and sensitivity

5. CONCLUSION

The study concludes that an ensemble ML model and several combinations of preprocessing and data splitting strategies can outperform the most recent work on customer churn prediction using ensemble LR in a dataset. Unlike previous research, our study focuses solely on developing a ML approach for customer attrition, has a time constraint with big data sets, and does not involve customer segmentation to generate strategies based on prediction as well as segmentation outcomes. The result finding of ensemble learning with LR has better accuracy as 95% which is comparatively better than traditional ML methods which determined that feature weight of other method has been effectively accumulated for generating better precision. Future study should focus on improving the efficiency and effectiveness of ML methods in huge dataset instances. In this study, the investigation merely employ structural data reveals the customer churn, unstructured data, such as consumer comments on social media or another platform, may influence the results. The integration of structural and non-structural data will provide a more dependable model of customer satisfaction with the company's product.

REFERENCES

[1]. L. T. Khrais, “Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce,” *Future Internet*, vol. 12, no. 12, p. 226, Dec. 2020, doi: https://doi.org/10.3390/fi12120226.

[2]. R. Ketipov, V. Angelova, Lyubka Doukovska, and R. Schnalle, “Predicting User Behavior in e-Commerce Using Machine Learning,” Cybernetics and Information Technologies, vol. 23, no. 3, pp. 89–101, Sep. 2023, doi: https://doi.org/10.2478/cait-2023-0026.

[3]. Ponnan Suresh, J Robert Theivadas, V.S. HemaKumar, Daniel Einarson, Driver monitoring and passenger interaction system using wearable device in intelligent vehicle, Comput. Electr. Eng. 103 (2022), 108323, https://doi.org/10.1016/j. compeleceng.2022.108323. ISSN 0045-7906.

[4]. S. Markkandan, R. Logeshwaran, N. Venkateswaran, Analysis of precoder decomposition algorithms for MIMO system design, IETE J. Res. (2021), https:// doi.org/10.1080/03772063.2021.1920848.

[5]. G. Klimantaviciute, Customer churn prediction in E-commerce industry, J. Mach. Learn. Res. 1 (2021) 1–14.

[6]. Ponnan Suresh, J Robert Theivadas, V.S. HemaKumar, Daniel Einarson, Driver monitoring and passenger interaction system using wearable device in intelligent vehicle, Comput. Electr. Eng. 103 (2022), 108323, https://doi.org/10.1016/j. compeleceng.2022.108323. ISSN 0045-7906.

[7]. S. Sharma and A. A. Waoo, "Customer Behavior Analysis in E-Commerce using Machine Learning Approach: A Survey. ," IJSRCSEIT, vol. 9, pp. 163–170., 2023.

[8]. S.-C. Necula, "Exploring the Impact of Time Spent Reading Product Information on E-Commerce Websites: A Machine Learning Approach to Analyze Consumer Behavior," Behavioral Sciences, vol. 13, no. 6, p. 439, Jun. 2023, doi: https://doi.org/10.3390/bs13060439.

[9]. L. Li, "Analysis of e-commerce customers' shopping behavior based on data mining and machine learning," Soft Computing, Jul. 2023, doi: https://doi.org/10.1007/s00500-023-08903-5.

[10]. Z. J. Wen, W. Lin, and H. Liu, "Machine-Learning-Based Approach for Anonymous Online Customer Purchase Intentions Using Clickstream Data," Systems, vol. 11, no. 5, pp. 255–255, May 2023, doi: https://doi.org/10.3390/systems11050255.

[11]. A. A. Tudoran, "A machine learning approach to identifying decision-making styles for managing customer relationships," Electronic Markets, Jan. 2022, doi: https://doi.org/10.1007/s12525-021-00515-x

[12]. Nie Chen, "Research on E-Commerce Database Marketing Based on Machine Learning Algorithm", Computational Intelligence and Neuroscience, vol. 2022, Article ID 7973446, 13 pages, 2022. https://doi.org/10.1155/2022/7973446

[13]. Liu C-J, Huang T-S, Ho P-T, Huang J-C, Hsieh C-T (2020) Machine learning-based e-commerce platform repurchase customer prediction model. PLoS ONE 15(12): e0243105. https://doi.org/10.1371/journal.pone.0243105

[14]. Sarvjeet Kaur Chatrath, G.S. Batra, Yogesh Chaba, Handling consumer vulnerability in e-commerce product images using machine learning, Heliyon, Volume 8, Issue 9, 2022, e10743, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2022.e10743.

[15]. Pondel, M., Wuczynski, M., Gryncewicz, W., Lysik, L., Hernes, M., Rot, A., & Kozina, A. (2021). Deep Learning for Customer Churn Prediction in E-Commerce Decision Support. , 3-12. https://doi.org/10.52825/bis.v1i.42.

[16]. Zimal, S., Shah, C., Borhude, S., Birajdar, A., & Patil, P. (2023). Customer Churn Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2023.49142.

[17]. V. Urbancokova, M. Kompan, Z. Trebulova, M. Bielikova, BEHAVIOR-BASED customer demography prediction in e-commerce, J. Electron. Commer. Res. 21 (2) (2020) 96–112.

[18]. T.T. Leonid, R. Jayaparvathy, Classification of elephant sounds using parallel convolutional neural network, INTELLIGENT AUTOMATION AND SOFT COMPUTING 32 (3) (2022) 1415–1426.

[19]. Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data, 7(1), 9.

[20]. M. A. Shehab and N. Kahraman, "A weighted voting ensemble of efficient regularized extreme learning machine," Computers & Electrical Engineering, vol. 85, 2020

[21]. K. HASAN, A. ALAM, D. DAS, E. H. 3 and M. HASAN, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," IEEE Access, vol. 8, 2020.