Proceedings of the Second International Conference on Automation, Computing and Renewable Systems (ICACRS-2023)
IEEE Xplore Part Number: CFP23CB5-ART; ISBN: 979-8-3503-4023-5

2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3503-4023-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACRS58579.2023.10404319

Comparative Performance Analysis using Machine Learning for Churn Prediction in E-commerce

1st R Alexander
Department of Computing Technologies
SRM Institute of Science and Technology, Kattankulathur
Chennai, India
ar4204@srmist.edu.in

2nd Maria Nancy A*
Department of Computing Technologies
SRM Institute of Science and Technology, Kattankulathur
Chennai, India
marianaa@srmist.edu.in

3rd Aswini. E
Department of Computing Technologies
SRM Institute of Science and Technology, Kattankulathur
Chennai, India-603203
aswiniettiyappan@gmail.com

4th Parwaz Singh Sarao
Department of Computing Technologies
SRM Institute of Science and Technology, Kattankulathur
Chennai, India
ps4989@srmist.edu.in

***Abstract*—New clients cost a business significantly more money in e-commerce than keeping its existing clients. Companies can boost consumer retention, which will result in more revenue and faster growth, by anticipating which customers will quit. There are several products and solutions in today's competitive industry. Because of this, most clients are accustomed to quickly switching from one brand to another and from one supplier to another in their search for the best possible product or item to fulfill their requirements. This problem, known as "client churn," affects e-commerce enterprises. Due to their ability to process large volumes of data and recognize complex patterns, machine learning algorithms have emerged as a powerful tool for predicting client churn in recent years. Using a publicly accessible dataset, the proposed model examines various machine learning methods for predicting customer churn in this study. Also, by using performance metrics, the proposed model compares how well different algorithms perform.**

Keywords—Customer Conduct, Customer churn prediction, Machine Learning, Data Frame, E-Commerce

I. INTRODUCTION

Customers are significant to a business. It's terrible for any firm to invest a lot of money to acquire new clients and lose them after two months. Since it has a sizable impact on the organization's revenue, customer churn or attrition is crucial in assessing business growth and performance. The costs of acquiring new consumers are often six to ten times higher than the costs of keeping the existing ones [1]. In such a situation, marketing professionals work to reduce customer turnover by shifting their attention from obtaining new clients to maintaining existing ones. With the help of e-commerce, consumers have numerous options to compare goods from other companies and switch businesses with little effort. Customer churn is a significant issue for businesses as a result.

An industry must examine the factors that lead customers to sever their ties with a business by ceasing to use their services or purchase their goods. A lower churn rate results from the ability of e-commerce professionals of a specific firm to adapt their present activities and offers. The ability to predict customer churn is an important research topic that helps businesses determine how many subscribers will remain active over a specific period. The industries must choose a practical methodology to estimate churn forecasts to sustain their revenue. This churn forecast may be calculated using data mining and machine learning algorithms. Once an efficient model [2] has been created, it can tell churners from non-churners. This study focuses on predicting client churn in e-commerce.

II. LITERATURE SURVEY

A. Machine Learning

Machine learning uses a variety of different algorithms with a variety of functions to extract hidden patterns and key concepts from datasets. Electronics is employed to collect information [3]. Business decisions can be made better by utilizing this underutilized information resource. Hence, predicting client attrition can be done using machine learning. When combined with machine learning and artificial intelligence, electronics make signal processing, pattern recognition, and precise signal output adjustments possible. Mathematics and machine learning both use algorithms to gather data and generate predictions. Machine learning, however, aids in predictive analysis when applied to statistics.

B. I am predicting online review ideas with improved machine learning techniques.

Many Different algorithms were used in Praphula Kumar Jain's suggested strategy [4]. They also pre-processed the input characteristics using various methods before feeding them into their training and validation models. Finally, they assessed the predictive recommendation system's capacity for generalization on the testing set. With the help of the suggested method, future airline suggestions may be predicted, and prospective purchases can be informed before they are made. Compared to other ways, the unique approach that has been offered performs significantly better.

C. Evaluation of XG-Boost and Enhanced Value Model for Predicting E-commerce Customer Churn.

A different method was introduced, and the conventional paradigm was presented to investigate the behavior of non-contractual clients, such as those who shop online [5]. To identify high-value customers, this model first incorporates social network aspects of e-commerce clients into the RFM. The churn prediction is then carried out using the XG-Boost algorithm. This strategy offers fresh perspectives on predicting customer attrition, and the client division improves

the forecasting rate. Earlier studies tended to focus on the individual qualities of customers rather than the interaction effects between customers.

D. E-Commerce Churn Prediction Using User Modelling.

Decision trees and other data mining models constructed using customer profiles can only categorize customers as churners or non-churners; they cannot provide profitable actionable knowledge. By undertaking the prediction model for likelihood assessment, [6] proposed an effective technique for business operations with cross-clients that require an automated retrieval of business information (PET). Our algorithm then offers the cost-sensitive steps to shift the consumer towards the most achievable, better lucrative position when this exact PET forecasts that they belong to one of the less profitable classes. The byte sequences are enterprises in a novel manner by applying binary arithmetic AND tasks. Using efficient data structures contributes to the suggested method's excellent computing performance. Substantial experiments on cellular line network information, UCI datasets, and different reference repositories show the recommended technique outperforms progressive solutions.

E. Predicting Mobile Customer Attrition with Machine Learning: The Future of E-Commerce

Due to the rapid development of technology, smartphone commerce is often viewed as a driving force in the next generation of online shopping. With the ability to communicate among clients at any time and from any location using digital devices, the Smartphone trade's strength stems from its ability to open up many possibilities for client acquisition and involvement. In the e-commerce era, many individuals believe it takes more work to retain customers, especially in the telecom sector. Putting a lot of effort into maintaining the current clients is crucial because many fiercely competitive service providers exist today.

The feature selection technique was used to identify all the top significant elements in forecasting user attrition [7]. Researchers adopted their method to highlight choice based on closures, in which Particle Swarm Optimization (PSO) is a finding strategy and various classifiers, such as Logistic regression, Decision Tree (DT), k-NN, and Naive Bayes, utilized for evaluation purposes to examine the impact of the legislation on datasets that have been optimally sampled and condensed. In this particular research endeavor, the feature selection approach was utilized to ascertain which characteristics are the most significant in forecasting the rate of customer churn—the use of simulations led to the discovery that this technique has a success rate for predicting churners. As a result, it has the potential to be helpful for the ever-increasing level of competition in the telecommunications industry.

F. System for predicting customer churn using machine learning

One of the most difficult challenges the telecommunications sector must contend with is the projected customer turnover rate or CCP. It has gotten significantly more straightforward to foretell client churn since the emergence of ML and AI. Many parts make up the strategy proposed in [8]. The initial two rounds involve pattern evaluation and processed information. The gravitational search technique was utilized during this subsequent round to examine element choice. Bolstering and aggregation techniques were used with prominent forecast methods, such as traditional machine learning algorithms, to investigate the effects on system reliability.

Moreover, K-fold cross-validation has been done to the train set to tweak the hyperparameters and prevent the models from overfitting. The AUC arc and confusion grid assessed the test set outcomes. With respective values of 81.71% and 80.8%, It was discovered that the most accurate results could be achieved with Adaboost and XGboost Classifier. Compared to other classifiers, Adaboost and XGBoost perform well and achieve the highest AUC score of 84%.

Though several approaches are used in literature for predicting churn, there aren't enough studies done to understand the importance of features and their contribution in predicting churn. The sole identification of these features can greatly contribute to better churn prediction. The detection accuracy of the models proposed in the literature is good.

III. PROPOSED WORK

The present study is crucial because it will help e-commerce businesses predict and identify which customers are likely to migrate. Retaining customers is a difficult problem faced by businesses worldwide. Thus, analyzing the different transactions that took place over time may aid in recommending to customers the right items, improving their experience, and completing customization. Our work utilized an e-commerce dataset with a transaction set that was conducted utilizing 5630 clients and 20 features out of the various transaction datasets for our study. The information on the feature set and feature distribution is shown in Table 1. The dataset preprocessing is carried out in stages such as cleaning, filtering, and scaling. The process of cleaning concentrates on eliminating the unusable data and this non-usability comes because it holds a null value thus these blank spaces are filled with the mean values. The outliers are identified using the box plot and then capping is performed for the removal of the outliers. Exploratory data analysis is carried out using the univariate analysis. Then the SMOTE analysis is carried out to balance out the class distribution. This is followed by the employment of six distinct machine learning algorithms and one deep learning method. After creating several bar graphs and confusion matrices, various performance measures of various algorithms were compared. The overall system diagram including the training phase and prediction phase is showcased in Fig. 1

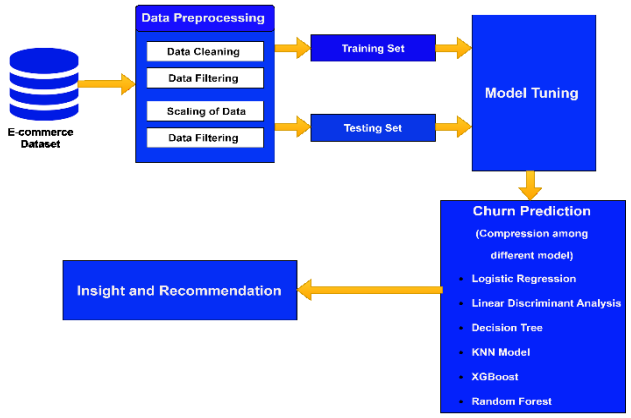

Fig. 1. System Architecture of the proposed model

IV. METHODOLOGY

A. Data Acquisition

The E-commerce dataset is taken from Kaggle. The consumption information of customers who purchased from the website was used for research and forecasting. The collection includes historical behavioral interactions when utilizing the platform, as well as consumption records from customers.

Table 1 consists of the dataset information in tabular style, the variable, and their description.

TABLE I. DATASET DISTRIBUTION

Variable	Description
Customer ID	Unique customer ID
Attrition	Attrition Alert
Duration	The client stays within that organization
Registration Device of Choice	User's favoured access gadget
City Tier	Level of the city
Warehouse to Home	The distance from the distribution center to the customer's place of residence
Preferred Payment Mode	Preferred payment method of customer
Gender	Customer identifier based on their gender
Time Spent using Application	Time consumed on a smartphone app or blog
Number of Devices Registered	The total amount of scams that have been perpetrated on a particular customer
Preferred Order Cat	Client order type that was given preference in the prior period
Satisfaction Score	Whether the user is happy with the product
Relationship History	Relationship History of the user
Number of Address	where the user lives
oppose	user liked service or not
Order Amount Hike From preceding Year	In a sequence of annual growth from the previous Year
Coupon Used	The overall amount of tickets utilized during the preceding period
Order Count	The entire amount of transactions made in the prior period
Recent Ordered Date	Day since the user's last purchase
Cashback Amount	Previous season's payout median

B. Data Cleaning

Handling null values and outliers was the data-cleaning process's main topic. Several techniques imputed missing values, including mean, median, and mode. Box plots and other schemes were used to identify outliers, eliminated or assigned using the Interquartile run (IQR) or the Tukey method [9].

C. Data Analysis

Each variable's distribution and summary statistics in the dataset were examined using univariate analysis in Fig 2. Bivariate analysis was used to investigate the connections between the dataset's variable pairs in Fig 3. This made it easier to detect the relationships or dependencies between variables, which assist in feature selection and model construction.

D. Clustering

Using a bottom-up method called hierarchical clustering, each client is first treated as a separate cluster, and the closest groups are then iteratively combined to form one large set that contains all of the customers. The average linkage approach was also utilized to calculate the separation between clusters. The generated dendrogram assisted us in identifying the ideal number of groups and visualizing the data's hierarchical structure in Fig 4.

K-means clustering is a partitioning technique that divides clients into k clusters according to how similarly they behave when making purchases. Each customer is assigned to the closest centroid depending on their distance after k centroids are randomly chosen. Utilizing an evaluation tool known as the silhouette coefficient, the ideal value of k was established.

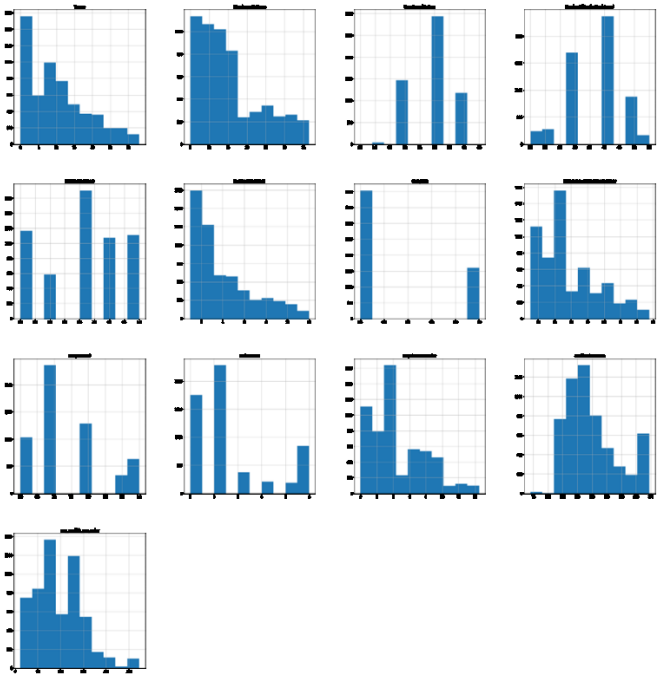

Fig. 2. Univariate Analysis

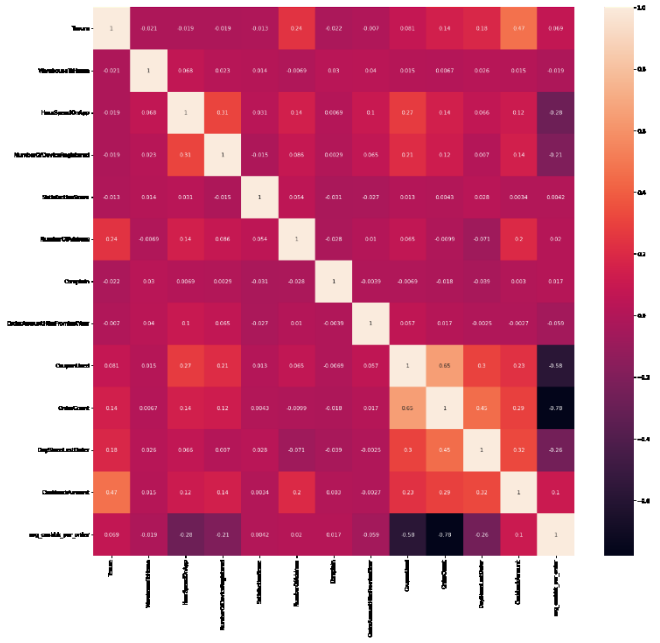

Fig. 3. Bivariate Analysis

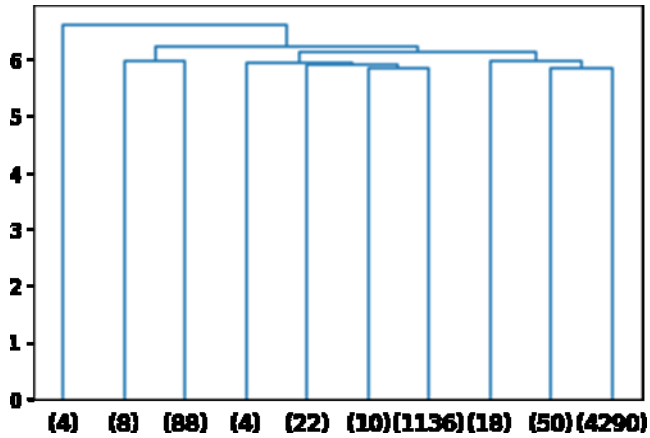

Fig. 4. Dendrogram

E. Synthetic minority oversampling technique (SMOTE)

To mitigate the category imbalance that is characteristic of churn datasets, the proposed model made use of SMOTE to generate additional minority samples artificially. The core idea behind this strategy was to interleave closely spaced minority courses with virtually distributed minority samples. To generate a new minority sample, each existing minority sample is sought among its k nearest neighbors, and these k neighbors are picked randomly from the entire set of points. Therefore, if the results of the dataset were to be unbalanced, SMOTE could increase the performance of any machine learning technique shown in Table 2.

TABLE II. SMOTE TABLE

	Customers Retained (0)	Churned Customers (1)
SMOTE Before	4682	948
SMOTE After	4682	4682

V. MACHINE LEARNING MODELS

A. Logistic Regression

One of the main techniques for analysis is logistic regression, which typically utilizes conditional results. The likelihood of a unique occurrence is modeled using the technique of logistic regression [10]. For categorization problems, logistic regression is commonly employed, especially when assessing whether a specimen fits within a subclass appropriately. The results showed that 85% accuracy was reached with the logistic regression model. The metrics for its evaluation were also used, as shown in Table 3.

TABLE III. LOGISTIC REGRESSION

	Precision	Recall	F1- score	Support
False	0.86	0.85	0.86	1171
True	0.85	0.87	0.86	1172
Accuracy			0.86	2342
Macro average	0.86	0.86	0.86	2342
Weighted average	0.86	0.86	0.86	2342

B. Linear Discriminant Analysis

A method for reducing dimensionality is linear discriminant analysis or LDA. The LDA seeks to cast those features onto a smaller space to evade the dimensionality constraint while saving resources and temporal expenditures in a larger area. A method for reducing dimensionality is linear discriminant analysis, or LDA [11]. The LDA aims to escape the dimensionality curse and minimize energy and dimensional costs by projecting the features in a larger environment onto a smaller area. With an accuracy rate of 85%, the LDA algorithm accurately predicted client churn in Table 4. The table shows that its evaluation measures were also removed.

TABLE IV. LINEAR DISCRIMINANT ANALYSIS

	Precision	Recall	F1- score	Support
False	0.87	0.84	0.85	1171
True	0.84	0.87	0.86	1172
Accuracy			0.86	2342
Macro average	0.86	0.84	0.86	2342
Weighted average	0.86	0.84	0.86	2342

C. Decision Tree

The decision tree is an autonomous machine-learning algorithm for classifying data or making predictions. By learning simple decision rules based on information properties, decision trees are used to develop systems that anticipate the worth of events [12]. The Decision Tree model's accuracy of 96% demonstrated that it was effective 4at foretelling client churn in Table 5. The decision tree's evaluation metrics were also removed.

TABLE V. DECISION TREE

	Precision	Recall	F1-score	Support
False	0.97	0.96	0.98	1171
True	0.97	0.97	0.98	1172
Accuracy			0.98	2342
Macro average	0.97	0.97	0.98	2342
Weighted average	0.97	0.97	0.98	2342

D. KNN Model

The K Nearest Neighbor Classifier classifies data by utilizing the query's closest neighbors as examples and establishing its type depending on its peers [13]. This classification strategy is particularly intriguing since traditional run-time efficiency considerations do not yet apply to computational resources. KNN model's 91% accuracy rate demonstrated its reliability in predicting client attrition in Table 6.

TABLE VI. KNN MODEL

	Precision	Recall	F1- score	Support
False	1.00	0.85	0.92	1171
True	0.87	1.00	0.93	1172
Accuracy			0.92	2342
Macro average	0.94	0.92	0.92	2342
Weighted average	0.94	0.92	0.92	2342

E. XGBoost

With the help of numerous weak prediction models, the ensemble learning technique XGBoost builds a robust predictive model [14]. The method incorporates a regularization method to help avoid overfitting this training batch of information, increasing its dependability and

accuracy while projecting fresh data. According to the analysis, the accuracy rate of the XGBoost model is 90.5%. The Random Forest algorithm outperformed XGBoost because the input data better matched Random Forest's assumptions, which explains why the accuracy of Random Forest is greater than that of XGBoost. This algorithm's evaluation metrics are listed in Table 7.

TABLE VII. XGBOOST

	Precision	Recall	F1- score	Support
False	0.91	0.92	0.92	1171
True	0.92	0.91	0.92	1172
Accuracy			0.92	2342
Macro average	0.92	0.92	0.92	2342
Weighted average	0.92	0.92	0.92	2342

F. Random Forest

It utilizes decision trees to tackle specific slotted masses and skew issues and gives accurate forecasts, mainly while multiple trees are unconnected [14]. Its ensemble learning method combines many decision trees to boost the model's accuracy and robustness. Estimates of the likelihood of a client leaving are precise and well-adapted to capture the dynamic nature of customer behavior. With a high accuracy of 98%, the Random Forest model accurately predicted customer attrition in Table 8.

TABLE VIII. RANDOM FOREST

	Precision	Recall	F1-score	Support
False	0.99	0.98	0.99	1171
True	0.98	0.99	0.99	1172
Accuracy			0.99	2342
Macro average	0.99	0.99	0.99	2342
Weighted average	0.99	0.99	0.99	2342

VI. RESULTS DISCUSSION

Random Forest demonstrated excellent performance, outperforming all the algorithms tested in this study regarding accuracy. Also, the accuracy of the Random Forest method, which is 98%, is higher than that of [16], which shows an accuracy of 80.03% [17], which displays an accuracy of 89%.

So, it is clear from Table 8 that random forest has the highest evaluation metrics.

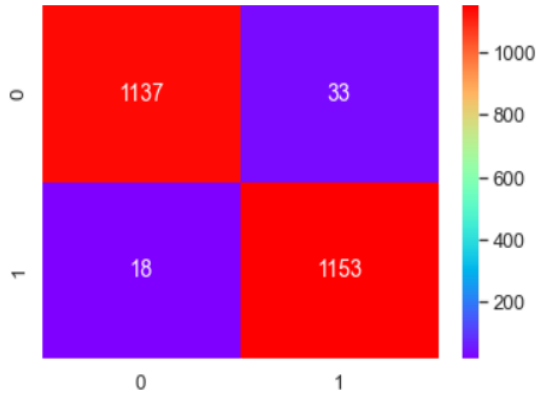

Fig. 5. Confusion matrix

So from this confusion matrix, Fig 5. Compared to the 1137 customers who did not churn, we can observe that Random Forest made an accurate prediction that 1153 consumers would churn.

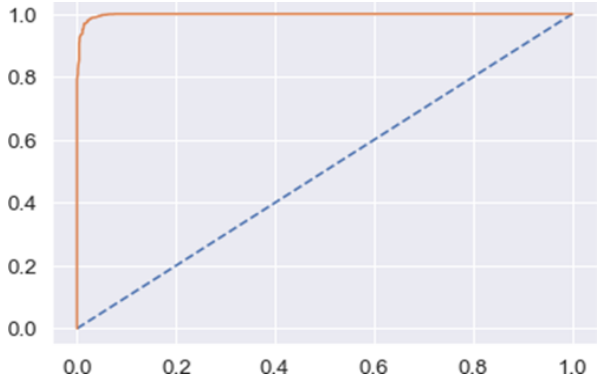

Fig. 6. ROC plot for the proposed system

The ROC curve's ideal point in Fig 6, which symbolizes perfect classification, is located closest to the top left corner of the plot. This point has a TPR of 1 (all real positives are accurately classified) and an FPR of 0(no actual negatives are incorrectly classified as positives). Thus, An AUC of 1 represents a perfect classifier.

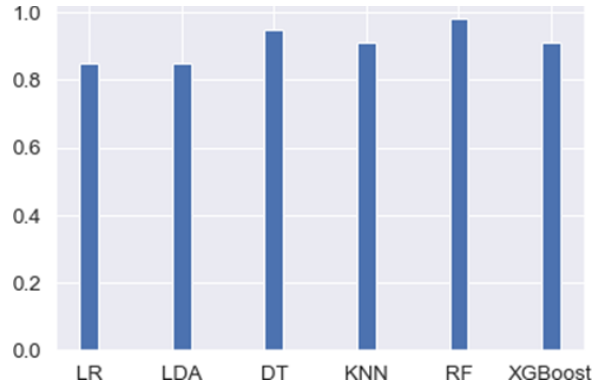

Fig. 7. Comparative analysis of machine learning model

The varied performance analysis of each of the algorithms is shown in Figure 7. As shown in the graph, the random forest is the best-performing model followed by the decision tree and logistic regression.

VII. CONCLUSION

Using a real-world dataset and a variety of machine-learning techniques, our proposed attempt to forecast customer churn. Our study's goal was to examine the overall performance metrics of several algorithms. From the results, it is evident that the Random Forest (RF) strategy leverages the highest accuracy and F1 score when compared to all other algorithms that were utilized. The proposed model can improve the algorithms' performance metrics by fine-tuning these algorithms, making the models more trustworthy and effective in forecasting client churn. In the area of e-commerce, our study has shown that the random forest algorithm, in particular, is a very accurate and efficient method for tackling this problem. These data-driven insights can assist online retailers in making decisions that will increase consumer retention and loyalty, leading to increased revenue and growth.

REFERENCES

[1] Amin, A., Shah, B., Khattak, A. M., Lopes Moreira, F. J., Ali, G., Rocha, A., and Anwar, S. 2019. "Cross-Company Customer Churn Prediction in Telecommunication: A Comparison of Data Transformation Methods," International Journal of Information Management 46: 304-319.

[2] Jain, H., Khunteta, A., & Srivastava, S. 2021. "Telecom Churn Prediction and Used Techniques, Datasets and Performance Measures: A Review." Telecommunication Systems, (76)4: 613-630. https://doi.org/ 10.1007 /s11235-020-00727-0

[3] Gaikwad, T. S., Jadhav S. A., Vaidya R. R., Kulkarni S. H. 2020 "Machine Learning Amalgamation of Mathematics, Statistics and Electronics" International Research Journal on Advanced Science Hub 2:100-108.

[4] Jain, P. K., Yekun E. A., Pamula R., Srivastava G. 2021 "Consumer Recommendation Prediction in Online Reviews Using Optimized Machine Learning Models." Computers & Electrical Engineering 95: 107397.

[5] Zhuang Y. 2018 "Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. Management Science and Engineering" (12) 3: 51-56

[6] Muneiah, J. N. and Rao D. V. 2019 "An Efficient Probability Estimation Decision Tree Post Processing Method for Mining Optimal Profitable Knowledge for Enterprises with Multi-Class Customers." Intelligencia Artificial 64 (22)

[7] Yaseen A. 2021 "Next-Wave of E-commerce: Mobile Customers Churn Prediction using Machine Learning" Research Journal of Computer Science and Information Technology 5 (2): 62-72

[8] Lalwani, P., Mishra M. K., Chadha J. S., and Sethi P. 2022 Customer Churn Prediction System: A Machine Learning Approach. Computing 104: 2-7.

[9] Graham, U. And Ian C. 1996 "Understanding Statistics". Oxford University Press Pp 55 ISBN 0-19-914391-9.

[10] Edgar, T.W. and David O. M. 2017 " Research Methods for Cyber Security" Cambridge, MA: Syngress, an imprint of Elsevier

[11] Tharwat A, Gaber T, Ibrahim A 2017 "Linear Discriminant Analysis: A detailed tutorial." AI Commun. 30(2):169–190.

[12] Kübler, R. 2022 "Understanding by Implementing: Decision Tree" Towards Data Science

[13] Cunningham, P., Delany, S. J. 2020 "K- Nearest Neighbour Classifiers. 2nd ed. MachineLearning. 2:1–15.

[14] Kiangala, S.K. and Wang, Z., 2021. A practical adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. Machine Learning with Applications, 4, p.100024.

[15] Iwendi C. 2020 "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm" Frontiers in public health https://doi.org/10.3389/fpubh.2020.00357

[16] Hiya N. and Pardede H. F. 2021 "Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data." Journal Elektronika dan Telekomunikasi 21.2: 122-127

[17] Irfan U. 2019 "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in the telecom sector." IEEE Access 7: 60134-60149.