Journal: Texila International Journal of Academic Research
Publisher: Texila International Journal
ISSN: 2520-3088
Volume 13 Issue 1, 2026
DOI: 10.21522/TIJAR.2014.13.01.Art027

A Comparative Analysis of Use of Machine Learning Algorithms for Customer Churn Prediction at Supermarkets in Lagos Nigeria

Temiye Oluwaseun[1*], Isaac John, Emmanuel Ephraim Etukudo[1], Dominic Essien[1], Chioma Esther Osumuo[2]

[1]*Information Technology, Texila American University, Georgetown, Guyana*
[2]*Research and Data Analytics, Renewal Research Institute, Abuja, Nigeria*

Abstract

The retail industry needs customer churn prediction to create successful customer retention strategies. The identification of churn customers in Nigerian supermarkets relies on conventional manual survey methods which take up time and generate limited predictive results. The research investigated customer behavior through transaction data analysis to predict churn while identifying the most effective supervised machine learning models for customer churn risk assessment. The study used transaction data from multiple supermarkets based in Lagos, Nigerian. Retrospective data from Plenty Africa, a digital loyalty platform, was used for this comparative analysis. The research employed predictive modeling methods to study customer characteristics and shopping patterns which helped detect patterns that lead to customer loss. The model performance was evaluated through conventional assessment metrics which led to a comparison for determining which approach performed better. The research showed that machine learning methods succeed in predicting customer departure through supermarket data collection which enables businesses to develop effective customer retention strategies. The research also shows how supermarkets can enhance their operational performance by using data-driven churn prediction systems which generate better results than human-based methods in retail operations.

Keywords: *AdaBoost, Customer Churn, Ensemble Methods, Lagos, Logistic Regression, Machine Learning, Random Forest, Retail Loyalty Data, Supermarket Analytics, XGBoost.*

Introduction

Several establishments, banks, and telecommunication corporations are beginning to implement new cutting-edge machine learning model applications and technology to identify customers who are likely to quit their services for other competitors. This is done in order to provide their customers with the best possible service and keep them, as losing them would cause a huge financial overturn for the company. This process of predicting customers' attrition is referred to as Churn Prediction. Churning customers are classified into two types: voluntary churners and non-voluntary churners. Non-voluntary churners are the simplest to spot because they are customers who have had their service or subscription terminated by the store or supermarkets. A corporation may revoke a customer's service or account for a number of reasons, including misuse of service and non-payment of service. Voluntary churn is more difficult to quantify since it occurs when a customer makes the intentional decision to discontinue service or patronage with the retail store or supermarket. Voluntary churn is classified into two categories: Incidental churn and planned churn

Received: 08.11.2025 ***Accepted: 14.12.2025*** ***Published on: 30.01.2026***
****Corresponding Author: temiye92@gmail.com***

[1]. Incidental churn occurs when changes in circumstances prohibit the customer from further requiring the offered service.

The initial studies about customer churn revealed that customers on purpose left their service providers to become customers of competing businesses which created significant problems for these service companies. [2] established service quality and pricing and customer experience as essential factors which drive customers to leave their service providers according to their research on customer behavior when switching between providers. Scientists can perform current research using machine learning with ensemble-based methods to enhance churn prediction accuracy in telecommunications and banking and e-commerce sectors [3, 16, 27] according to the initial research findings. The development of new methods has improved machine learning model performance for supermarket churn prediction but researchers have not fully investigated these methods in Nigerian developing economic conditions. The research fills this knowledge gap through its assessment of different machine learning approaches which predict customer departure from supermarket operations.

Therefore, with increased competition in the retail industry, particularly among supermarkets and outlets, supermarkets must employ customer retention strategies in order to increase their market share by attracting new customers. More so as it has been established that raising a supermarket's retention rate by up to 5% can increase earnings by up to 85% [7], even as attracting new clients is more expensive than retaining old ones, which is more likely to return profit. As a result, supermarkets should maintain their competitive advantage by utilizing effective methods to predict customer churn, turnover, retention, and loyalty. The conventional method of prediction of customer churn in supermarkets in Nigeria is through the rigorous manual customer survey analysed using excel. Machine learning algorithms/models such as XGboost, AdaBoost, Random Forest and Logistic Regression are alternative methods for prediction of customer churn. Some studies have applied ML optimization and soft computing methods to solve churn-related problems and concluding that use of ML model is effective, less time-consuming and requiring minimal resource as compared to conventional manual customer survey methods. Such studies include [2] who used particle swarm optimization (PSO) algorithm to treat unbalanced data, minimum redundancy and maximum connection for reducing the feature and use of random forest to predict churn; [4] who used a hybrid model based on the learning system to obtain a more accurate result, with model integrating supervised and unsupervised approaches to predict customer behaviour; [5] who used the re-sampling method with the support vector machine (SVM) to solve the unbalanced data problem in predicting customer churn in the telecommunications company; and [6] who used a genetic algorithm (GA) to develop a neural network model to predict customer turnover in a communications firm. While evidence of use and effectiveness of ML models have been studied in different settings like banks, telecommunication and organisation, evidence of use and effectiveness of the ML models in predicting customer churn in supermarkets -to inform action for customer retention- remains poorly studied. This study therefore developed and assessed the effectiveness of a predictive model for customer churn, adopting a supervised machine learning approaches (XGboost, AdaBoost, Random Forest (RF) and Logistic Regression (LR)) to predict customer churn and comparatively analyse performance of the four ML models against each other. This was done using customer level data from 20 supermarkets in Nigeria, which are registered merchants on the Plentiafrica.com, an app and loyalty platform used for data collection in supermarket.

Aim of this Study

Objective

The research establishes an optimal predictive system which helps retail stores and supermarkets identify their customers who are likely to leave their business after their initial two months of shopping.

Specific Objectives

1. The system is designed to develop a churn prediction model which will estimate customer churn risk when customers first join the service.
2. Organizations can determine customer loyalty maintenance through evaluation procedures which operate independently of customer segments and product choices.
3. The system should help supermarkets create specific promotional offers which will keep their at-risk customers from leaving the store.

Novelty of the Work

The research develops an original data-based system which predicts customer departure from retail supermarkets through analysis of customer actions during their initial two months of store visits. The research investigates which first signs of customer departure businesses can identify before their ability to recover lost customers becomes unavailable. The research method uses Figure 1. to show its conceptual framework which starts with customer retail scanner data and customer demographic information before moving to data preprocessing and feature extraction.

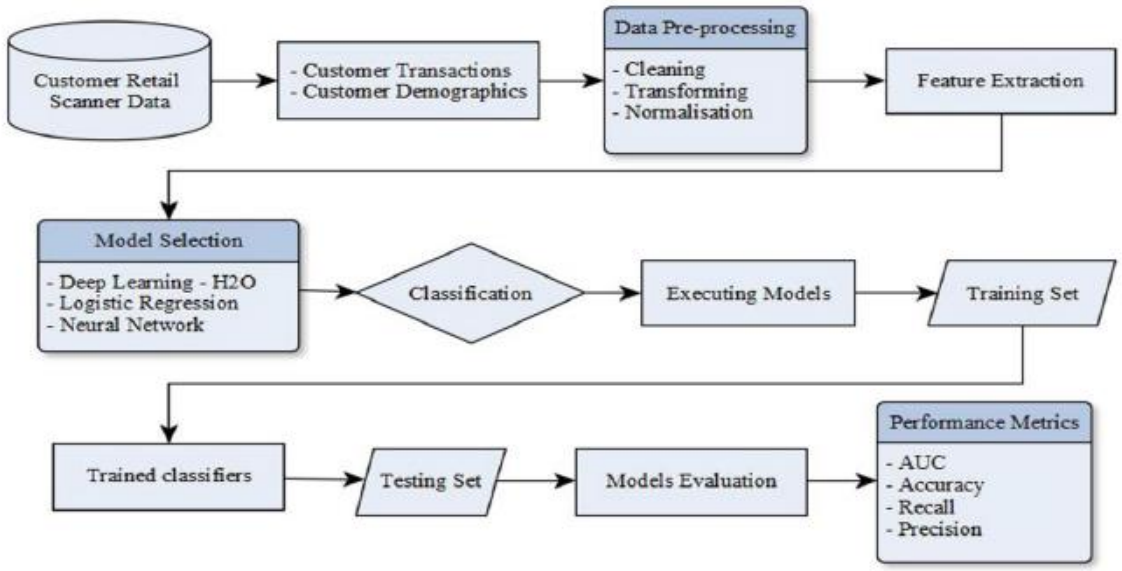

Figure 1. Machine Learning Pipeline for Customer Classification

Figure 1 presents the conceptual machine learning pipeline adopted in this study, illustrating the progression from customer data collection and preprocessing to model training, testing, and evaluation. As shown in *Figure 1*. the framework uses classification-based with supervised machine learning algorithm including logistic regression and ensemble-based supervised machine learning models (Random Forest, AdaBoost, and XGBoost). The model execution process follows a sequence which includes training and testing and evaluation through performance metrics that include AUC and accuracy and precision and recall for complete predictive performance assessment. The analytical pipeline shown in Figure 1 shows the original research approach through its operational system which helps supermarkets identify vulnerable customers at an early stage to create targeted loyalty programs.

Related Work

Customer Churn and Predicting Customer Churn

The process of customer churn prediction involves identifying which customers will stop using their service provider or retail outlet. Research studies have established two main categories of customer churn which include non-voluntary churn that service providers cause through their actions and voluntary churn that customers choose through their own decisions. The prediction of voluntary customer churn becomes difficult because it depends on various behavioral and experiential elements which differ from the direct termination signals used for prediction. The research by [1], divided voluntary customer departure into two categories which included incidental and planned churn because customers switch due to life changes and make deliberate choices to leave.

[2] conducted their first empirical study in 2004 to prove that service quality along with pricing and customer satisfaction serve as vital elements which drive customers to choose alternative options. Scientists now apply machine learning methods to customer behavior modeling because their first research results showed this approach produces superior churn prediction outcomes. Research studies have proven that supervised learning models together with ensemble-based learning models work well in all three domains which include telecommunications and banking and e-commerce. The available research about supermarket method implementation in developing nations remains scarce. The situation requires researchers to conduct more studies about machine learning algorithm operations which predict customer churn in these particular business settings.

To combat this kind of churn, it is essential to have a system in place that can predict and identify churn occurrences based on existing features and historical data along with information that can motivate customers to be retained. In addition, ensuring that high standards of customer care and the availability of up-to-date products are maintained are needed to proactively prevent churn. Doing the above will achieve [7] recommendations which established that raising a supermarket's retention rate by up to 5% can increase earnings by up to 85%, even as attracting new clients is more expensive than retaining old ones, which is more likely to return profit. As a result, supermarkets should maintain their competitive advantage by utilizing effective methods to predict customer churn, turnover, retention, and loyalty.

Though customer churn prediction using machine models have been successfully implemented and published across various domains and industries, such telecommunication, retail, banking and e-commerce [3-6]. There is however a need to understand use of machine learning (ML) in predicting customer churn in all and other settings like the supermarkets.

Major Components for Developing Loyalty and Customer Retention

1. Attitude, satisfaction, trust, and dedication are essential ingredients for creating a successful Loyalty program.
2. A positive attitude is essential to forming a meaningful bond with customers [8].
3. Satisfaction is the assessment of the gap between previous expectations and the performance of a good or service [9].
4. Trust is a critical aspect for establishing customer loyalty, and suppliers must earn customers' trust and protect the confidentiality of their data.
5. Commitment is also necessary to sustain loyalty as it is the result of a rational and emotional connection [8].
6. Managers must manage loyalty programs carefully to ensure customer loyalty over the long term [10].

7. Incentives inspire customers to make a single purchase, but also to become loyal to a company [11].
8. Rewarding customers with discounts or free items not only allows them to save money but also provides them with a sense of accomplishment and pride [10]. When customers are made to feel special, they develop a greater affinity to the company [11].

Customer Prediction Modelling and Churn Analysis

This section provides an overview of supervised machine learning approaches commonly used for churn prediction. Supervised learning is used to identify patterns from a given data set and predict target outcomes, such as those pertaining to a continuous or categorical variable (i.e. one with a constrained set of values). The outcome is referred to as binomial and the model as a binary classifier if there are only two potential states (i.e. 0 or 1, non-churn and churn). Numerous statistical models have been effective binary churn classifiers, such as evolutionary computing (such as genetic Algorithm and ant colony optimization), support vector machines, logistic regression, decision trees, boosted trees, gradient boosted decision trees, random forests, neural networks, and an ensemble of hybrid techniques [4]. Artificial neural networks demonstrate suitable performance for churn prediction because they can process complex non-linear patterns which exist in customer behavior data that heterogeneous customer attributes [15]. With a large set of observations, patterns can be generalized to accurately predict outcomes for unseen data. These developments reflect broader advances in machine learning and deep learning techniques, which have expanded their application beyond traditional domains to include structured customer and transactional data analysis [14]. By leveraging such models, organizations can use their data to gain insight into customer behaviour and take appropriate actions to reduce churn and retain customers.

Overview of Machine Learning Algorithms/Models

Regression Classification and Logistic Regression

Regression and classification are two of the widely used techniques in Machine learning and data analysis. Regression is a supervised learning technique used to predict a continuous or numerical outcome (such as house prices or stock prices) from a set of independent variables. While classification is a supervised learning approach used to predict a categorical outcome (e.g., a Yes/No decision) from a set of independent variables. Classification models are used to predict a categorical outcome from a set of independent variables. Examples of classification models include decision trees and Logistic regression. Logistic Regression is a type of regression used to predict a binary outcome (Yes/No) from a set of independent variables. In Logistic regression models, the relationship between the dependent variable and the given feature set can be represented using probability functions [12] and it can be used with discrete, continuous, or categorical explanatory variables. The model is favored by many since it’s straightforward to implement and interpret, in addition to being robust. In this case, the dependent variable is a probability between 0 and 1. Decision trees are a kind of machine learning algorithm that uses necessary set of rules to classify a data set. SVMs are a type of machine learning algorithm used to classify data points by creating a hyperplane that divides the data points into two categories.

Model Evaluation and Performance Measures

To evaluate and measure performance of the ML models, the following measures can be adopted.

The Receiver Operating Characteristic (ROC)

ROC curve is an important tool for comparing classification methods. As the discrimination threshold of ROC is changed, it visibly depicts the trade-off between TP and FP.

The Area Under the ROC Curve (AUC)

AUC is the likelihood that a classifier would score a randomly selected positive instance higher than a randomly selected negative example; the greater the AUC value, the better the prediction accuracy [18].

The F-score

The F-score commonly known as the F1-score, is a model's accuracy on a dataset. It is used to assess binary classification algorithms that categorize examples as 'positive' or 'negative.' [18].

Model Accuracy

Model accuracy is defined as the number of correct classifications predicted by a model divided by the total number of predictions produced. It is one method of evaluating a model's performance, but it is far from the only one.

$$Accuracy = \frac{(TP + TN)}{P+N} \qquad (1)$$

Model entropy provides insight into the predictive usefulness of individual variables by measuring how well they contribute to distinguishing between churn and non-churn outcomes. The greatest potential predictor is one that has examples with the same value as the guide variable in each generated bin; thus, the guide variable can be correctly predicted.

Precision

Precision is one measure of a machine learning model's performance since it measures the accuracy of a positive prediction provided by the model. Precision is calculated by dividing the number of true positives by the total number of positive predictions.

$$Accuracy = \frac{(TP + TN)}{P+N} \; ; \; F_1 = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)} \qquad (2)$$

Recall

The recall is determined as the proportion of Positive samples that were correctly identified as Positive to the total number of Positive samples. The recall of the model assesses its ability to recognize positive samples. The more positive samples identified, the larger the recall.

Confusion Matrix

The confusion matrix is a matrix that is used to determine how well classification models perform for a given set of test data. It can only be determined if the genuine test data values are known.

Data Description and Churn Analysis

Churn rate (also referred to as attrition) is defined as the annual turnover of customers over a predefined period [19]. Customers need to be segregated into active and non-active and periodically monitored overtime to determine the churn rate. The churn rate is calculated as the proportion of active customers who discontinue their patronage over a predefined period.

$$Active\ Clients_{\{(at\ end\ year)\}} = \Sigma Active\ Clients_{\{(at\ begin\ year)\}} + \Sigma new\ Clients_{\{(enrolled\ within\ year)\}} - \Sigma Churned\ Clients_{\{(discharged\ during\ year)\}} - \Sigma Non - churn\ Client\ exits_{\{(discharged\ during\ year)\}} \qquad (3)$$

Equation 3: Churn Rate is a percentage of active customer who churned over a predefined period.

$$Churn\ Rate\ (\%) = \sum_{\{s=1\}}^{\{S\}} \left(\frac{Churned\ Clients_s}{Active\ Clients_s}\right) \qquad (4)$$

Equation 4: where S is set of supermarkets s1, s 2, ..., sn**.**

Applied Churn Prediction Model

Prediction models have been widely employed in numerous areas, such as forecasting customer churn, weather forecasting, fraud detection, risk mitigation, etc. In many industries, many statistical models have been utilized as churn classifiers. Research has shown that customer behavior patterns and purchase records serve as strong indicators which help predict customer churn in non-contractual businesses setting such as logistics and retail industries [13].

[20] conducted a study on features extracted from customers' transaction history from the point of sale (POS) system that can be used to predict churn within the retail industry. Data from a local supermarket was used for this study. Convolution Neural Networks (CNN) and Restricted Boltzmann Machine learning algorithm were the deep learning techniques used for prediction. The Restricted Boltzmann Machine attained the best results of 83% in predicting customer churn compared to CNN, which had an accuracy of just 74%. [19].

In [21] study in a novel e-churn model for improving customer churn prediction through increased recall. For churn prediction, this model employed the ensemble technique. Different algorithm combinations were investigated, including C5, QUEST, CHAID, CRT, and logistic regression, with the combination of C5 and QUEST yielding the best results (93.4%). There was no mention of a dataset for the experimentation and evaluation. The accuracy of these algorithms can be improved by including customer data history.[22] also discovered a significant relationship between historic data and churn prediction. The Authors used customer behaviour data of a B2C e-commerce enterprise to test the predictive ability of the SVM and LR models to evaluate the prediction performance of the two models, the k-means algorithm was first used for clustering subdivision to classify into three types of customers, and then predictions were made for these three types of customers. Accuracy, recall, precision, and AUC were calculated. A total of 987,994 customer record was used for this study. The accuracy of the SVM model prediction was higher than that of the LR model prediction with 92% and 90% accuracy respectively.[23] compared CNN and LR for predicting customer churn in telecommunication industry in Nigeria. It showed that CNN, which is mostly used for image classification, can also be used for churn prediction. CNN gave 89% accuracy compared to LR which gave an accuracy of 80%. Recall, F1-score and confusion matrix were used on our datasets for sampling, accuracy and to check how well the classification models behaved.

Another churn prediction model used in CCP is the Random Forest (RF) model, Gradient Boosted Model (GBM), and versions of linear regression with a 9.2% churn rate in the organization, Big Data technologies were used to acquire and prepare a massive dataset. The RF model was chosen with AUC (area under the ROC curve) as the evaluation criterion. Several methodologies, including up-sampling, down-sampling, and weighing strategies, were used to investigate class imbalance. [31] developed a churn prediction model based on the Genetic Programming (GP) algorithm with AdaBoost. AdaBoost was utilized in an iterative way to find distinct reasons for customer churn. To balance the dataset, the Particle Swarm Optimization (PSO) sampling approach was applied. This model solved several complex issues very effectively.

In another study, [32] evaluated how well the BP neural network and AdaBoost performs in predicting customer churn on e-commerce data from the B2C industry. Data segmentation into three categories was employed to facilitate model evaluation. The results indicated that the AdaBoost model was more accurate than the BP neural network model, with Accuracy, Recall, and Precision metrics of 0.9555, 0.9316, and 0.9604 respectively. Additional metrics such as ROC and AUC for the AdaBoost model

were also better than those for the BP neural network model. Consequently, the amount authors suggest that the AdaBoost model is more suitable for predicting customer churn in this context.

Using historical purchasing behaviour of customer in an insurance company [33] developed a new customer profitability metric for the insurance businesses that measures real insurance customer contribution using liability reserve. They applied Random Forecast Regression, a method for Big Data analytics, to forecast insurance customer profitability. The five crucial variables to forecast insurance client profitability were discovered to be region, age, insurance status, sex and customer source. The proposed RF model performs very well in the training and testing datasets with a pseudo-R-square of 0.9903 and 0.9901, respectively. It is also found that the RF model has the smallest Root Mean Squared Error compared to other models, such as SVM and generalized boosted, decision tree and linear regression models. For customer relationship management and premium actuarial in insurance businesses, RF was concluded to have significant reference value.

Hyperparameter Tunning and Parameter Selection

The task of selecting a collection of ideal hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning in machine learning. A hyperparameter is a parameter whose value governs the learning process. Other parameters (usually node weights) are learned in contrast. A parameter sweep is a frequent approach in machine learning model training that involves exhaustively scanning a subset of variables deliberately selected by the researcher. This method is known as grid search, and it is one of the most commonly utilized strategies due to its simplicity and reproducibility. However, executing a grid search over every possible value combination quickly turns it into an approach that is far too computationally expensive to be useful in real-world applications. Performing a random search across the same parameter domain can result in a model that is as good as or better than the one validated by grid search while requiring a fraction of the calculation time. A random search selects parameter values at random from a specified distribution rather than directly from a grid. Samples are collected for each parameter over a predetermined number of rounds and utilized for model validation This enables us to plan a processing budget that is independent of the number of available parameters and the number of possible values for each. As a result, including parameters with little effect on performance will not reduce the efficiency of the cross-validation process for no gain. Some of the benefits of RandomizedSearchCV over grid search include that searching many different parameters at once may be computationally infeasible and RandomizedSearchCV searches a subset of the parameters, and you control the computational "budget".

System Design and Methodology

Methodology

Methodology used is a structured system analysis. This section further describes the methods and materials used to carry out the study and arrive at outcome/result of the study. This includes study location, study population, study design, sample size, sampling technique, intervention approach, data collection methods and tools, data analysis, ethical consideration, and study schedule/duration.

Study Location

The study was conducted in Lagos, Nigeria, which is the biggest city in Nigeria, thus having diverse number of supermarkets.

Study Population

The study was conducted among supermarkets using customer level data.

Study Design

This is a retrospective comparative analytic study that assessed the effectiveness of the use of four machine learning algorithms/models (XGboost, AdaBoost, random forest and logistic regression) for prediction of customer churn using data from ''Plenti Africa'' app in supermarkets and compared the effectiveness of each of the four models against each other.

Sample Size and Sampling Technique

According to finelib.com, a Nigeria directory n of supermarkets, there are 62 large supermarkets in Lagos. Using raosoft sample size calculator [34], at sample size *n* and margin of error E given by $x = Z(^{c}/_{100})^{2}r\ (100\text{-}r)$, $n = {}^{N\ x}/_{((N-1)E^{2}+x),}$ $E = \text{Sqrt}[{}^{(N-n)x}/_{n(N-1)}]$, sample size of 19 was calculated, using Raosoft, (2004). Additional 1 was added as allowance for 5% attrition/eventualities, thus total of 20 supermarkets was randomly selected for the study.

Data Collection and Data Description

Data collection was done using data collection app known as PLENTI AFRICA, deployed in supermarkets in Lagos Nigeria. Plenti Africa is a loyalty collation platform that helps supermarkets manage their loyalty program. The solution has a mobile app that is used to collect customer transaction information and thereafter give the customers points which can be collated over a period and analysed to predict customer churn. Data collectors, who are supermarket operators, were trained to use app for data collection. Retrospectively, data were retrieved from transaction history on the data collection app-Plenti Africa from Jan 2019 to October 2022 from the 20 selected supermarkets.

System and Data Analysis

The Dataset and Data Schema Extract

The data schema extract from the plentiafrica.com database was used for the study. The schema contains 5 major tables that hold customer demographics, customer balance, customer airtime usage, customer credit cards and customer transaction logs. The data extracted for the analysis was sourced from customer transaction logs, customer demography and transaction stats (i.e., this table 1 stores the aggregate balances for each customer) tables respectively. To enable easier analysis, some variables like age categorization were derived. Thus, age band variable was constructed by sorting customers into four distinct age groups: Teenagers (ages <19), Young Adults (ages 20-40), Middle Aged (ages 41-60), and senior citizen (>60 years).

Table 1. Shows the Attributes that were Extract Using SQL Join Queries from the Plenti Africa Database

Variable	Description
Customer ID	The Identification number of the customer
Mobile Number	The mobile number of the customer
Gender	The sex of the customer
DOB	The date of birth of the customer
Age	The age of the customer
State of residence	The State of residence of the customer
Current balance	The current balance of the customer
Date of registration	Date the customer was registered into the program
Last visit date	The last visit date of the customer
Total transaction	Total transaction of the customer

System Implementation and Analysis

Source Data Entity Relationship Diagram (ERD)

The ERD model in *Figure 2.* was used to display the relationships amongst attributes being derived. Figure 2 presents the source data Entity Relationship Diagram used in this study. To obtain total transaction by age groups new attributes were created. Age group was derived from the current age of the customers and was further categorized into teenager (<19 years), young adult (20-40 years), middle age (41-60 years) and senior citizen (>60 years). While total transactions per customers were derived from counting all customer transactions per supermarket The data thus contained variables related to customers' transactions which include the amount they spent, date of transaction, age, sex, state of residence, the date they registered on the platform etc.

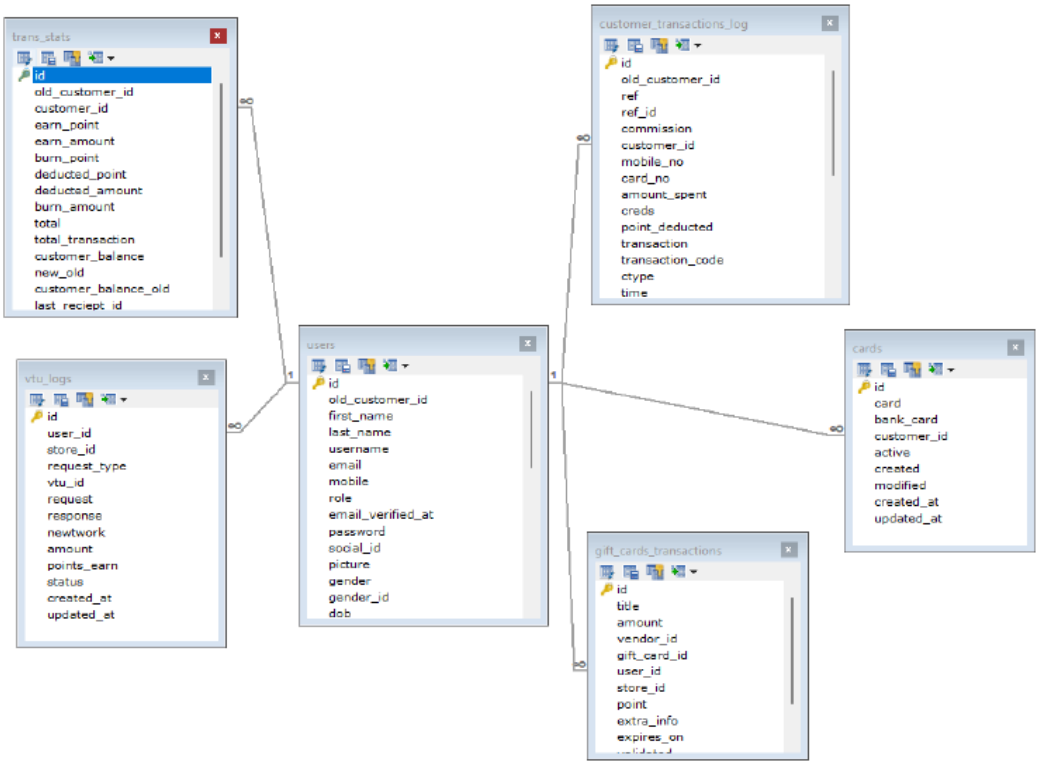

Figure 2. Source Data Entity Relationship Diagram

After converting the multi-dimensional customer data into a flat file (a single file with columns representing the variables and rows representing each customer records that were generated), the file was assessed and pre-processed to identify anomalies and unusable records. The list of final attributes can be seen in *Table 2.* In the data transformation phase, some of the variables were aggregated (e.g Age was categorised to teenage, young adult, middle aged and senior citizen). All variables were aggregated to binary for better interpretation for the predictive modelling. Additionally, some of the variables were used to derive new variables.

Table 2. Variables Derived from Database for Customer Churn

Variable	Description
Customer ID	The Identification number of the customer
Mobile Number	The mobile number of the customer
Gender	The sex of the customer
DOB	The date of birth of the customer

Age	The age of the customer
Age category	The Age category of the customer
Tenure	The tenure of the customer
State of residence	The State of residence of the customer
Current balance	The current balance of the customer
Date of registration	Date the customer was registered into the program
Visited 3months Age	Check if a customer has visited any store in the last 3 months
Last visit date	The last visit date of the customer
Total transaction	Total transaction of the customer
Churn	Check if the customer churns or not

Handling Missing Data

In dataset, the issue of missing values was addressed. After a thorough analysis of the data, it was concluded that the number of null or missing values was negligible in comparison to the total sample size. As a result, all missing data from the State of residence, Gender and Age Band fields (comprising 45 Gender, 48 Age Category and 144 State of residence fields) were deleted in accordance with stating that most of the models are unable to fit and predict values when supplied with missing values.

Data Cleansing

Prior to describing the data and calculating correlation metrics, data collected needed to be cleansed of duplicates. In the absence of a pre-existing definition and monitoring of customers churn, each distinct customer was thus labelled as “churned” or “not-churned”.

A de-duplication procedure was carried out to find and eliminate any duplicate customer entries to ensure the uniqueness of the customer on the database. The process involved using a customer's full name and phone of find perfect duplicates and deleting them. As a result, the final dataset was 6,188 as extracted from the database. Subsequently, all unique customers were categorized as ‘churned =1’ if customer had not been to the supermarket in the last 2 months and as ‘non-churned=0’, if customer had been to the supermarket in the last 2 months.

Imbalanced Data

Issue of imbalance data did not need to be addressed as the number of churned and non-churned customers were not significantly different (3015 churned and 3173 non-churned) [24–26]. When predicting customer attrition with machine learning, the imbalance between the classes of customers who have stayed versus customers who have left is a common problem [24-26] as more data are usually available on customers who have stayed than those who have left. However, data was available for all customers in this study where churned or non-churned.

Feature Selection

Machine learning model development requires feature selection as an essential process because it helps researchers decrease data dimensions while removing unneeded information which results in better model performance. Two consecutive stages of feature selection were used in this study. Firstly, correlation analysis was used to find linear relationships between input data points which showed which variables had the strongest connection to each other. The purpose of this step was to check for duplicate variables and prevent multicollinearity by identifying strongly related variables which could affect the models.

Secondly, model-based feature importance techniques were employed using ensemble

learning methods. Specifically, AdaBoost and XGBoost algorithms were used to rank features based on their relative contribution to churn prediction. The algorithms determine feature importance through training-based assessment which produces performance-related scores for each feature. The models revealed essential characteristics which future predictive modeling required so these characteristics were retained for additional research. The method used statistical analysis together with predictive modeling to choose features which showed both statistical importance and predictive ability which produced models that were more stable and simpler to understand.

Correlation Method

Pearson correlation method was used for the study as displayed in *Figure 3*. The correlation heatmap in Figure 3 shows that the predictor variables have weak relationships with each other which justifies their inclusion for future model construction. The Pearson correlation method was used to identify the variables that should be kept for the current analysis. The variables tenure and total transaction had a weak correlation of 0.14483. Since correlation was considered relatively small compared to the target variable 'churn', the variables were retained.

This process was repeated for all variables, and all features were kept since they showed weak correlations with one another. This ensured that all relevant variables were accounted for.

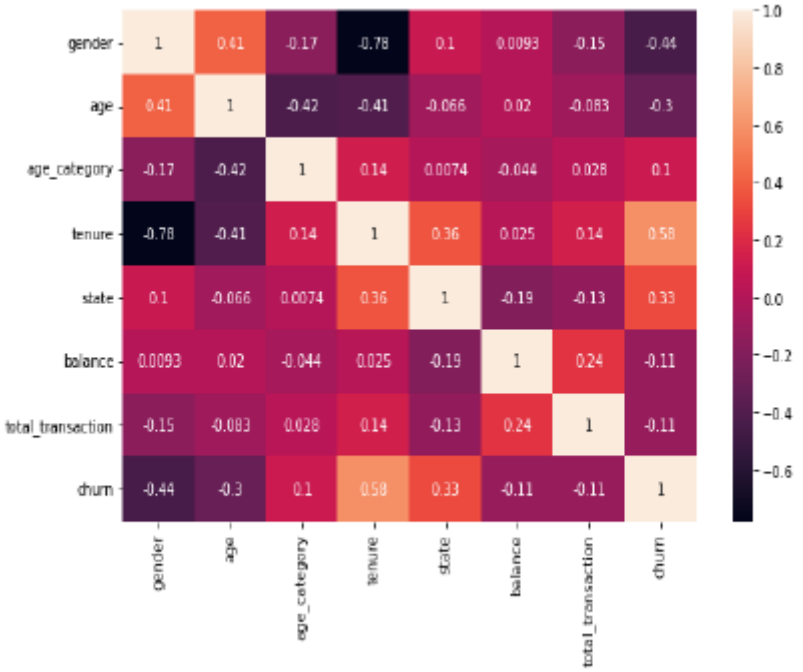

Figure 3. Correlation Heatmap showing Pairwise Relationships between Variables

Prediction Modelling of Customer Churn using Machine Learning

Prediction modelling discusses why specific classifiers were chosen. It demonstrated the usage of grid search and 10-fold cross-validation to discover the optimal hyperparameter combinations for the classifiers. The accuracy, F-score, and AUC scores of each model are shown when the default parameters and grid-searched parameter tuning approach are employed. The confusion matrices and ROC curves for the two models that performed the best were then displayed. Furthermore, the churn prediction model's

creation and test results are discussed, beginning with an overview of the approach and common feature sets used by all models.

Following the preliminary data preprocessing and correlation-based feature screening described in Section 4.5, Adaboost and XGboost feature selection algorithm were used to identify features with the highest predictive relevance. Highly correlated features were dropped resulting in a common set of features used across candidate prediction models ensuring consistency in feature representation across all algorithms and improving the robustness of the comparative analysis. The data was then analysed from each of the models (AdaBoost, XGboost, Random Forest and Logistic regression) and the most successful models were ranked based on comparative effectiveness to answer the study questions. Study questions include.

1. Question 1. "Is the kind of customer data collected on daily basis suitable for use by machine learning algorithm?
2. Question 2: Which machine learning algorithms are more suitable for predicting customer retention from the increasing volumes of data that supermarkets have about their customers?
3. Question 3: How can the data collection process be improved to achieve customer data suitable for use by machine learning algorithms?

Model Development

It is assumed that customer past purchase behaviour determines the future outcomes [17]. Therefore, customer historical behaviour was analysed for future outcomes using the four ML algorithms/models. The data used for this study was derived by setting cut-off transaction date to November 2022, as this was the date when the supermarkets under study discovered a huge drop in daily transaction as compared to previous months. This attribute is a key variable in labeling instances as churn or non-churns.

All models generated churn prediction probabilities. The probability of each instance is rounded to nearest class membership of either churn (1) or non-churn (0) and threshold used was 50%. To visualize prediction accuracy with changing threshold values, Receiver Operating Characteristic (ROC) curve was used and the area under the AUC was computed to compare prediction accuracies, the overall prediction accuracy, which measures are correct classifications over the population, and if the main criteria was used for the final best model selection.

Table 3. below displays summary of the Models. All steps in the model development were included in a single program written in Python programming Language and published at

https://colab.research.google.com/drive/1XxdlXrf70rPNtMFYbkg-UWGWhdw3nPHL

Table 3. Model Development Summary

Dataset	Customer transaction date from 20 supermarket in Lagos Nigeria Binary outcome variable (Churn or No Churn)
Feature Selection	XGBoost and AdaBoost
Candidate prediction models	Logistic regression, Random Forest, AdaBoost and XGBoost
Prediction probability threshold	50% for all models
Model Comparison	F1 score, ROC, Accuracy, Confusion Matrixes
Validation	10-fold cross validation (90%-10% training/test split)

Data Preparation and Handling the Class Imbalance

The dataset, which was previously cleansed for descriptive analysis as described in previous chapter, needed to be further prepared for prediction modelling. In data preparation, the following were undertaken:

1. Variables that have values missing for up to 50% were removed.
2. Negative values set to 0 or Null.
3. While missing numerical values were deleted, the missing categorical values were set to most used value.
4. Also, defaults were assigned as required (e.g., customers with null age values were set to young adults since this age range was more common in the data set).
5. Quartile-based boundaries replaced outliers (i.e. 3 times inter-quartile range).
6. Normalization was applied to numeric variables like the amount spent by customers. Z-score normalisation was also performed only for the logistic regression model.
7. Furthermore, the experiment in this study was randomly under sampled to a class ratio of 1:1. Testing different ratios for the class distribution was thus one of the experiments that were performed in this study.

Result and Discussion

Customers' Demographic Characteristics

Figure 4. and Table 4. below displays demographic characteristics of customers.

Sex: Out of the 6,188 customers analysed, the majority, 3713 (60%) were female while 2475 (40%) were male.

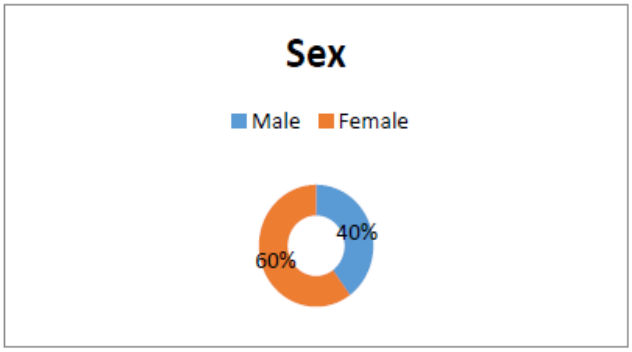

Figure 4. Sex Distribution of Customers

Table 4. Age Categories of Customers

Age category	Total	% Total
Teenagers (<19)	4409	71.25
Young adult (19 to 40)	1134	18.33
Middle adult (41 to 60)	309	4.99
Seniors (>60)	336	5.43
Total	6188	100

Outcome of Customer Churn Analysis

Table 6 below shows rate of customer churn, displayed by age category. Almost half (49% (2654)) of the entire 6188 customers were churn and 51.2% remained not churned (active). By age category, teenagers had highest churn rate at 54.0%, followed by seniors at

44.0%. Young adults and middle-aged adults have the lowest customer churn rate at 32.8% and 35.2% respectively.

Feature Selection by AdaBoost and XGBoost

*Figure **5.** and Figure 6.* demonstrate that tenure and total transaction are the most significant features selected by Adaboost and XGboost (the two algorithms/models used for feature selection). Figure 5 and Figure 6. Feature selection with XGboost Algorithm/model illustrates the feature importance ranking generated by the AdaBoost and XGboost algorithm, identifying tenure and total transaction as key predictors of customer churn. Tenure and total transaction are two of the most important metrics for gauging customer loyalty in the supermarket industry.

Tenure indicates how long a customer has been shopping in the supermarket, giving an indication of their loyalty and how likely they are to continue shopping there in the future. Total transaction indicates the number of transactions the customer has done in the supermarket, which can be used to measure their loyalty and the impact they have had on the supermarket's profits.

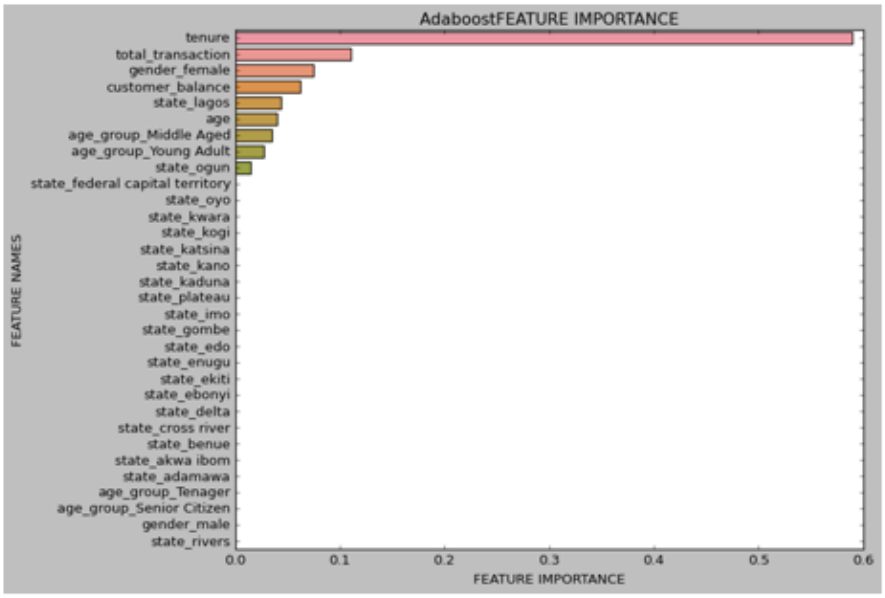

Figure 5. Feature Selection by AdaBoost Algorithm/Model

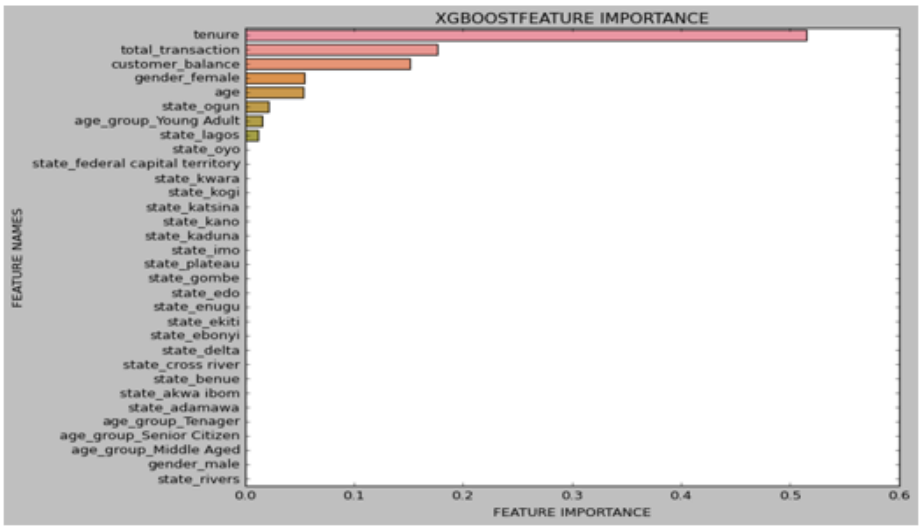

Figure 6. Feature Selection with XGboost Algorithm/Model

Outcome of Algorithm/Model Prediction

Logistic Regression

The most important hyperparameters for logistic regression are regularization or C, solver, penalty, and random state.

1. Penalty intends to reduce model generalization error and is meant to disincentivize and regulate overfitting. Technique discourages learning a more complex model, to avoid the risk of overfitting.
2. Solver is the algorithm to use in the optimization problem. The choices are (*newton-cg', lbfgs', liblinear', sag', saga')*, default='lbfgs'.
3. Regularization or C is the inverse regularization parameter to the lambda parameter in the LR. The default setting for logistic regression in (Scikit-learn is solver = "liblinear", penalty = 'l2', regularization/C =1.0).

LR Report Before Parameter Tuning

The results for datasets before the RandomizedSearchCV is displayed in Table 7 below. Where default=1.0, before parameter tuning, precision, recall and F1-score for Not Churn were relatively less effective with scores 0.74, 0.86 and 0.80 respectively while Churn were 0.88, 0.76 and 0.81 respectively. Whereas accuracy for F1-score was 0.80. With these scores, Logistic regression was considered effective.

LR Report After Parameter Tuning

After running RandomizedSearchCV with the dataset, the following parameters were found to be the best parameter.

1. Solver = newton-cg,
2. Penalty =none
3. Maxiter = 1000
4. C = 0.0006951927961775605.

The result after parameter tuning is displayed in Table 7 below.

Random Forests (RF)

Many parameters of RFs can be changed to find the optimal match. The following parameters were tested:

1. Minimum samples per leaf (default=1) is the number of samples necessary at the leaf node.
2. Minimum samples per split (default=2) sets the number of samples necessary in a node before it can be split.
3. Bootstrapping (default=true), a Boolean value that specifies whether bootstrapped samples or the entire dataset is utilized to create each tree.
4. Maximum tree depth (default=None), which limits the tree's maximum depth to 45.
5. Max features (default 2) which is the number of features to consider when looking for the best split.
6. The number of estimators (default=10) defines how many trees there are in the forest.

RF Report Before Parameter Tuning

Results with default values are shown in Table 7 below. Where default=1.0, before parameter tuning, precision, recall and F1-score for Not Churn were 0.89, 0.87 and 0.88 respectively while Churn were 0.90, 0.90, and 0.90 respectively. Whereas accuracy for F1-score was 0.89. With these scores, Random Forest was considered effective.

RF Report After Parameter Tuning

After running RandomizedSearchCV with the dataset, the following parameters were found to be the best parameter as shown in Table 7 below.

1. Bootstrapping = false
2. Maximum tree depth = 4
3. Maximum features considered for a split = "sqrt"
4. Minimum samples per leaf = 1
5. Minimum samples per split = 2
6. Number of estimators = 41

7. Criterion ='entropy'

XGBoost

XGboost has many parameter which are classified into 3 types: General parameters, booster parameter and task parameters. XGboost has 35 different parameters, but the most important ones are the N estimators, Subsample, Maxdepth, Learning rate, Gamma, Regalpha, Reglambda. Other special purpose parameters include Scale Pos Weight, Monotone constraints, booster, missing, Val metric. For this study the following parameter were tried before parameter tuning.

base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bytree=0.5, gamma=0.4, learning_rate=0.1,max_delta_step=0, max_depth=6, min_child_weight=7, missing=None, n_estimators=100, n_jobs=1, nthread=None,objective='binary:logistic',random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=True, subsample=1.

XGBoost Report Before Parameter Tuning

The results for datasets before the RandomizedSearchCV is displayed in Table 7 below.

Where default=1.0, before parameter tuning, precision, recall and F1-score for Not Churn were 0.90, 0.86 and 0.88 respectively while churn were 0.89, 0.92, 0.91c respectively. Whereas accuracy for F1-score was 0.90. With these scores, XGboost was considered effective.

XGBoost Report After Parameter Tuning

After the random search CV tuning procedure, the best parameters were found and were the best parameters for the model. The final result after parameter tuning is shown in Table 7 below.

1. Solver='liblinear'
2. Penalty = 'none',
3. Max iter: 5000,
4. C: 0.0018329807108324356

AdaBoost

AdaBoost is a boosting algorithm that combines multiple weak learners into a strong learner. It is a sequential technique that works by fitting a classifier on the original dataset and then fitting additional copies of the classifier on the same dataset, but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

The default parameters used were.

1. n_estimators default=50 which is the maximum number of estimators at which boosting is terminated. In case of perfect fit, the learning procedure is stopped early.
2. learning_rate**,** default=1.0 a weight applied to each classifier at each boosting iteration.
3. A higher learning rate increases the contribution of each classifier. The results generated by AdaBoost for datasets before and after the RandomizedSearchCV are displayed in Table 7.

AdaBoost Report Before Parameter Tuning

Where default=1.0, before parameter tuning, precision, recall and F1-score for Not Churn were 0.84, 0.91 and 0.87 respectively while churn was 0.92, 0.92, 0.86 and 0.89 respectively. Whereas accuracy for F1-score was 0.88. With these scores, AdaBoost was considered effective.

AdaBoost Report After Parameter Tuning

Where default=1.0, after parameter tuning, precision, recall and F1-score for Not Churn were 0.84, 0.91 and 0.87 respectively while churn was 0.92, 0.92, 0.86 and 0.89 respectively. Whereas accuracy for F1-score was 0.88. With these scores, AdaBoost was considered effective. The final result after parameter tuning is displayed in Table 5 below.

Table 5. Outcome of Algorithm/Model Prediction

		Precision	**Recall**	**F1-score**	**Support**
Classification Report before LR parameter tuning	Not Churn	0.74	0.86	0.80	265
	Churn	0.88	0.76	0.81	335
	Accuracy			0.80	600
	Macro avg	0.81	0.81	0.80	600
	Weighted Avg	0.81	0.80	0.80	600
Classification Report after LR parameter tuning.	Not Churn	0.78	0.83	0.80	272
	Will Churn	0.85	0.80	0.83	328
	Accuracy			0.82	600
	Macro avg	0.82	0.82	0.82	600
	Weighted Avg	0.82	0.82	0.82	600
RF Report before parameter tuning	Not Churn	0.89	0.87	0.88	272
	Will Churn	0.90	0.90	0.90	328
	Accuracy			0.89	600
	Macro Avg	0.89	0.89	0.89	600
	Weighted Avg	0.89	0.89	0.89	600
RF Report after parameter tuning	Not Churn	0.88	0.88	0.88	264
	Will Churn	0.91	0.91	0.91	336
	Accuracy			0.90	600
	Macro Avg	0.89	0.89	0.89	600
	Weighted Avg	0.90	0.90	0.89	600
XGBoost Report before parameter tuning	Not Churn	0.90	0.86	0.88	264
	Will Churn	0.89	0.92	0.91	336
	Accuracy			0.90	600
	Macro Avg	0.90	0.89	0.89	600
	Weighted Avg	0.90	0.90	0.89	600
XGBoost Report after parameter tuning	Not Churn	0.91	0.89	0.90	265
	Will Churn	0.91	0.93	0.92	335
	Accuracy			0.91	600
	Macro Avg	0.91	0.91	0.91	600
	Weighted Avg	0.91	0.91	0.91	600
AdaBoost Report before parameter tuning	Not Churn	0.84	0.91	0.87	272
	Churn	0.92	0.86	0.89	328
	Accuracy			0.88	600
	Macro Avg	0.88	0.88	0.88	600
	Weighted Avg	0.88	0.88	0.88	600

AdaBoost Report after parameter tuning	Not Churn	0.87	0.92	0.89	272
	Will Churn	0.93	0.88	0.90	328
	Accuracy			0.90	600
	Macro Avg	0.90	0.90	0.90	600
	Weighted Avg	0.90	0.90	0.90	600

Effectiveness of ML models using Evaluation Metrics

Confusion Metrics

The classification results of each model became visible through confusion matrix analysis which measured their ability to predict customer churn correctly based on actual customer behavior. Figure 7. presents the confusion matrices for XGBoost, AdaBoost, Random Forest, and Logistic Regression models, highlighting their respective classification performance. The confusion matrix contains four elements which include true positives (TP) and true negatives (TN) and false positives (FP) and false negatives (FN). In the context of this study, a true positive represents a customer correctly predicted as churned, while a true negative represents a customer correctly predicted as not churned. A false positive happens when the system predicts customer churn, but the customer stays active. The system produces a false negative when it shows a customer as active, but the customer actually left the company. The classification results from XGBoost and AdaBoost and Random Forest and Logistic Regression appear in Figure 8. Confusion metrics of XGBoost, AdaBoost, Random Forest and Logistic regression which shows their distinct performance levels. The ensemble-based models (XGBoost, AdaBoost, and Random Forest) produced better true positive and true negative rates than Logistic Regression which showed they performed better for prediction tasks.

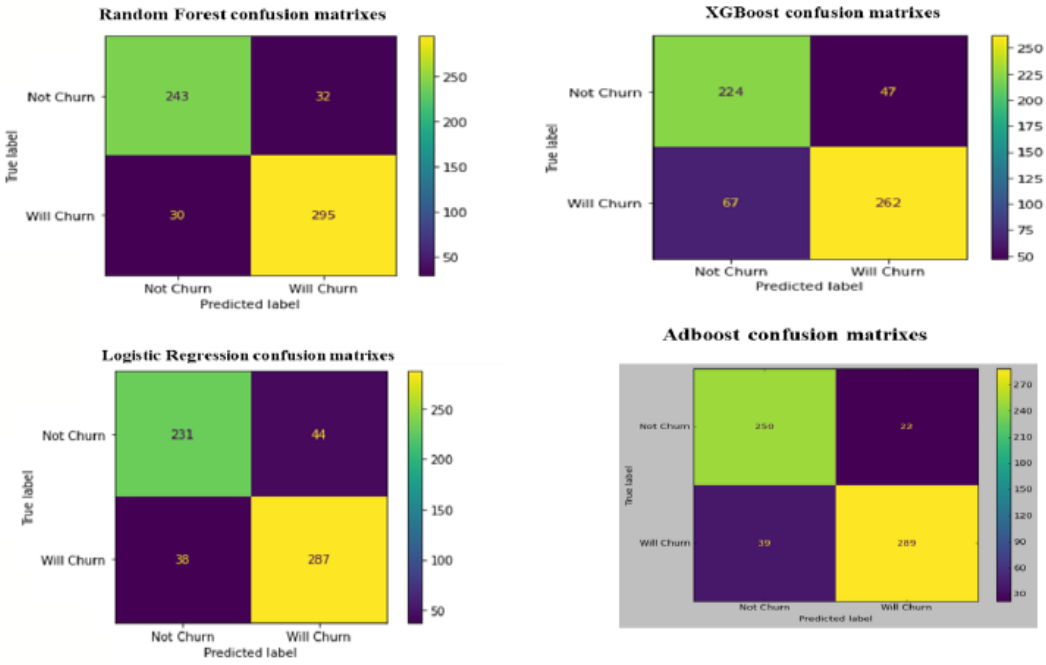

Figure 7. Confusion Metrics of XGBoost, AdaBoost, Random Forest and Logistic Regression

Receiver Operating Characteristic (ROC) Curves

The ROC-curve demonstrates that the models had effective performance when compared to the random dotted line. There is significant difference in the ROC curve of the 4 different models especially with better performance seen in XGboost and AdaBoost. Figure 8 shows the Receiver Operating Characteristic (ROC) curves comparing the classification performance of the four machine learning models. In order of performance, XGboost, AdaBoost and Random Forest are more effective as compared to Logistic Regression.

Table 6. Summary of Evaluation Metrics

	RF	**LR**	**XGBoost**	**AdaBoost**
AUC	0.869093	0.797466	0.896919	0.879677
F-Score	0.885278	0.798712	0.911243	0.896755
Accuracy	0.871667	0.791667	0.900000	0.883333
Recall	0.911234	0,801234	0.931123	0.681222

Metrics evaluation outcome of the four ML models using AUC, F-score, Accuracy and Recall

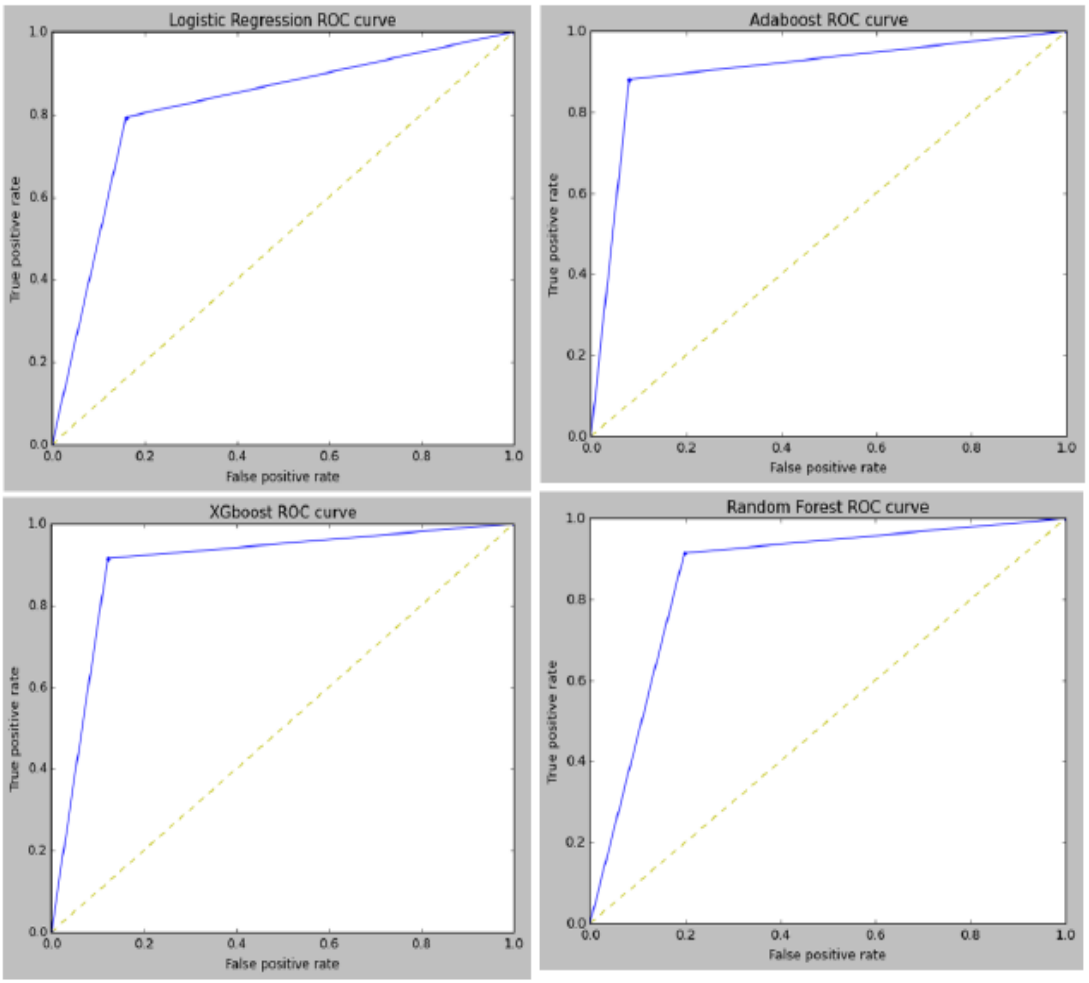

Figure 8. The ROC Curves Further Illustrate the Comparative Classification Performance of the Four Machine Learning Models

Summary of Finding

Suitability of Customer Level Data Collected on Daily Basis for Use by Machine Learning Algorithms/Models

To answer the research question ‘‘Is the kind of customer data collected on daily basis suitable for use by machine learning algorithm?’’, the finding demonstrated that customer data collected on daily basis were suitable for use by ML algorithms/models. >60% of data collected were complete (6188/10,000) and used for analysis of customer churn using the four-machine algorithm. Thus, the alternate hypothesis was accepted that **‘’**Customer data collected on daily basis are suitable for use by machine learning algorithm’’.

Effectiveness of use of Machine Learning Algorithms/Models to Predict Customer Churn Using Data Collected in Supermarkets

To answer the research question ‘‘Is Machine learning algorithms/models effective in predicting customer churn in supermarkets?’’, the findings showed that though Logistic Regression has a low performance, all four ML algorithms are effective in predicting customer churn in supermarkets. Evaluation matrixes also demonstrated their reliability in use performance. Therefore, the alternate hypothesis ‘’ Machine learning algorithm/models are effective in predicting customer churn in supermarkets’’ is accepted.

Comparative Analysis of the Four Machine Learning Model in Prediction of Customer Churn

To answer the research question ‘’ Which machine learning algorithms are more effective in predicting customer churn? comparative analysis of the four machine learning models evaluated using evaluation matrixes showed that XGboost has the highest performance and is more effective, followed by AdaBoost, and then Random Forest with average AUC of 0.90,0.88,0.87 respectively. Logistic Regression has the least performance with average AUC of 0.82.

The final XGboost model fitted hyperparameters were Solver = 'liblinear', Penalty = 'none', and Maxiter: 5000. The XGboost and Adabost after hyperparameters tuning yielded comparable results (AUC of 0.91 and 0.90, respectively) Whereas Random Forest and Logistic Regression realized the lowest AUC values of 0.89 and 0.82 respectively.

However, XGBoost and AdaBoost required longer time than Random Forest and Logistic Regression because their boosting algorithms and parameter tuning procedures needed additional time to complete. The XGBoost model achieved superior training results through hyperparameter optimization which resulted in better performance than its initial state. Boosting-based models achieve better prediction results for churn prediction tasks because they need additional computational power than other models.

F1-score demonstrated that the XGboost model is the best technique in comparison for balancing the precision-to-recall trade-off with value of 0.911243.

Confusion matrix applied on XGboost model analysis classified 67 customers falsely as not churn (23.02% of all not Churn) and 47 as will churn (15.30% of all will churn). This is further reflected by the model’s precision (0.91) and recall (0.93). Thus, there is a high proportion of both correctly predicted ‘’will churn or churn’’ among all correctly and falsely predicted as ‘’will churn/churn’’ (84.69%) and correctly predicted ‘’will not churn’’ among all actual ‘’will not churn’’ (76.98%).

The results seen in previous studies were similar to the findings from this study. Effective performance of XGboost, AdaBoost and Random Forest seen in this study concurs with studies of [31].

On the other hand, the low performance of the Logistic Regression demonstrated in this

study is in accordance to finding by [28], which concluded that Logistic Regression model performed poorly as it does not fit well into more massive datasets. However, another study concluded that LR performance in simple form has been seen to show average results [17]. In the same study which compared Random Forest and Logistic Regression models performance in prediction, it was concluded that Random Forest outperformed Logistic Regression [17], Comparative studies have also shown that boosting-based algorithms such as XGBoost better than traditional machine learning classifiers in structured prediction tasks [27], thus agreeing with findings in this study. However, on the contrary, a finding by [28] disagrees with finding from this study as it concluded that Logistic regression performs the best for most of the datasets, and it outperforms the boosting methods (AdaBoost and XGBoost) for the datasets with a 5% minority class.

Findings from this study also show that XGboost model has the highest performance (performance with accuracy of 0.91 and F-score of 0.90- 0.92) which agrees with finding from [16] where it was demonstrated that XGboost approaches typically outperformed other machine learning (ML) models/algorithms by emphasizing functional space when decreasing model cost.

XGboost thus performed well on the datasets collected daily at the supermarkets, by having the highest average score in all metrics (including the F-score where AB is lightly lesser to it by 0.01).

Random Forest however has the best outcome in terms of effectiveness to identify false positive as seen in Confusion matrix which shows that Random Forest has less false positives across all datasets.

Furthermore, in comparative analysis of the ROC-curves of XGboost and AdaBoost, it was discovered that XGboost appears to perform better. However, XGboost was shown to be a computationally heavy model that consumed much more memory than other models. Complex and large models may imply very low entropy and thus very complex models, which may also result in poor generalizability and overfitting.

AdaBoost is the second-best performing model on the datasets. The differences between AUC and accuracy metrics on average are not that considerable, even though the differences in ROC curves are noticeable. By looking at CMs, XGboost introduced more false positives regarding churning customers when compared to RF.

Random Forest was not computationally demanding, which was a plus for RF. Memory consumption was manageable, and predictions were accurate. RF is considered the third-best model. It is practically tied with LR in terms of average scores, but it may outperform in terms of F-score.

XGboost performs quite similarly than AdaBoost in this dataset. Due to the stochastic nature of most ML algorithm all models where ran 30 times each and The average Accuracy, F-score, and AUC values across repeated runs were used for the final comparative evaluation of the models.

Summary of Literature

Table 4 presents a summary of the literature review's findings. The table shows that support vector machines (SVM), logistic regression (LR), artificial neural networks (ANN), decision trees (DT), and random forests are the most used models. These models come in different iterations, but they all fall under the same general category. Additionally, the area under the curve (AUC), receiver operating characteristics curve (ROC), percentage properly classified/accuracy (PCC), and top decile lift appear to be the most popular validation methods. However, there is also widespread use of the confusion matrix (CM) and training and evaluation (T/E).

Table 7. Summary of Client Churn Prediction Models Reviewed

Tittle of paper	Industry	Data Set	Models and Techniques	Accuracy
Customer profitability forecasting using Big Data analytics: A case study of the insurance industry	Telecom	18,000	RF, SVM, decision tree and linear regression	RF had highest accuracy with 99.67%
Optimizing Coverage of Churn Prediction in Telecommunication Industry	Telecommunication Industry	Telco BSS and OSS data	C5, QUEST, CHAID, CRT, and logistic regression	QUEST performed better with 93.4% accuracy
Customer churn prediction by hybrid neural networks. Expert Systems with Applications	Multimedia Industry		ANN, C5.0 decision trees	C5.0 decision trees and ANN) improved prediction accuracy
Adaptive Particle Swarm Optimization Algorithm Ensemble Model Applied to Classification of Unbalanced Data	Telecom		Random Forest and Naive Bayes	Random Forest outperformed Naive Bayes, with an accuracy of 71.99%
Machine-learning techniques for customer retention: A comparative study of churn in telecommunication	CRM systems	3,333	Random Forest and AdaBoost	AdaBoost achieved the highest accuracy of 96%
B2C E-Commerce Customer Churn Prediction Based on K-Means and B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM	E-Commerce	987,994	SVM, LR	SVM, LR achieved a higher accuracy of92%
Predicting Customer Churn in Telecommunication Industry Using Convolutional Neural Network Model	Telecommunication industry in Nigeria	Customers transaction	CNN and LR	SVM, LR achieved a higher accuracy of 89%

Research Gap

Research studies have thoroughly examined customer churn prediction through multiple industries which include telecommunications and banking and insurance and e-commerce while using supervised learning and ensemble machine learning approaches. Research studies achieve high prediction results through their extensive well-organized datasets which focus on customers who have subscription-based or contractual agreements.

The field of churn prediction has received limited research regarding its application to supermarket and retail settings because their customers do not have contractual agreements and their buying activities follow unpredictable patterns. There is limited research available about churn prediction which uses supermarket loyalty and transaction data from developing economies such as Nigeria. Existing studies

also tend to evaluate individual models in isolation rather than conducting a comprehensive comparative analysis across multiple machine learning algorithms using the same dataset.

The available research lacks sufficient evidence about how daily customer information from digital loyalty platforms works for machine learning-based supermarket customer churn prediction. The research lacks proper methodological and contextual analysis which creates an obvious knowledge deficit. The research fills this knowledge gap through two main objectives which evaluate supermarket transaction data for churn prediction and compare different supervised machine learning models to find the best method for supermarket churn prediction in Lagos Nigeria.

Recommendation

1. Currently, the supermarkets referred to in this study use manual/analogue method of customer frequency as model for predicting customer churn. However, no classification-based predictions about churning are currently in used. This study demonstrated the effectiveness of Machine Learning algorithms in accurately predicting customer churn and also classifying the predictions.
2. Results show that XGBoost and AdaBoost are more effective in predicting customer churn using data from supermarkets as compared to Random Forest and Logistic Regression. Thus, XGboost and AdaBoost are recommended for use in supermarket settings.
3. Furthermore, customer data collection using Plenti Africa app at the supermarkets is suitable in customer churn prediction by the ML models. Thus, data collected using the Plenti Africa app is recommended for use in ML model to predict customer churn.
4. When comparing the performance of features selected by ML algorithms/model to those provided manually by the supermarket, the ML models have significantly better performance. It was demonstrated that by adding two additional characteristics chosen via a feature selection method, it was possible to improve the performance of the model on different datasets. These results were the same regardless of customer balance or tenure. These further buttresses recommendation of use of ML model in supermarkets to predict customer churn often and effectively.

Conclusion

From the findings of this study, it is an effective method in performance and accuracy to predict customer churn in supermarkets using ML models (XGBoost: 0.91, AdaBoost: 0.90, RF:0.89 and LG: 0.82), by using the customer level data provided at the supermarkets. In addition, XGBoost and AdaBoost are better performing models as compared to Random Forest and Logistic Regression.

Limitations and Future Research

The following limitations were encountered on this study and future research are thus recommended.

1. Larger dataset is required for maximal use of the Boost models (XGBoost and AdaBoost). However, for this study, barely 6000 datasets were analysed from only 20 supermarkets who use the data collection tool (Plenti Africa). Future research is therefore recommended to be done using larger dataset in more supermarkets across Nigeria and beyond.
2. Though XGBoost and AdaBoost were concluded to be the best performing models as compared to Random Forest and Logistic Regression. However, the computer facilities needed to use XGBoost and AdaBoost would not be affordable for supermarkets as compared to simpler computer facilities used for RF and LR. Thus, further studies are needed to assess

suitability and availability of computer facilities used for machine learning.

3. Customer behavior was not added to features as it wasn't part of the data collection tool. This information would be needed to determine reasons for customer churn. Further studies on risk factors to customer churns should be undertaken.

Conflict of Interest

No conflicts of interest were reported by the author on this project. From start to finish, fairness and accuracy guided every step - writing, thinking, sharing.

Acknowledgements

The author gratefully acknowledges the support from the University of Edinburgh Napier and his supervisor, Dr. Powers, Simon, for providing him with the necessary knowledge and guidance to complete this work, and also to CEO of Plenti Africa, Mr. Dominic Essien, for granting me access to Plenti Africa's database and the invaluable data it provided. My family has been a pillar of strength throughout this journey, and I cannot thank them enough for their unwavering support.

Ethical Approval

The research used Plenty Africa provided de-identified customer transaction data which underwent secondary analysis. The research did not require direct human subject contact and used protected data, so it did not need official ethical approval according to NHREC of Nigeria guidelines. The data usage followed all regulations which protect personal data and maintain confidentiality standards.

Data Availability

The datasets used for this study were collected from app known as PLENTI AFRICA, deployed in supermarkets in Lagos Nigeria. Plenti Africa is a loyalty collation platform that helps supermarkets manage their loyalty program. Therefore, can only be retrieve by writing to the Plenti Africa team.

Author Contributions

Oluseun Temiye led the study through every phase, from shaping the research design to collecting and analysing the data and ultimately drafting the manuscript. Chioma Osumuo and Isaac John contributed technical expertise, offering guidance on the study's design and supporting the interpretation of the results. All authors read, reviewed, and approved the final version of the manuscript.

Funding

The research project did not obtain funding from any public or commercial or not-for-profit sector organization. The study was self-supported with institutional guidance from Edinburgh Napier university.

References

[1]. Hadden, J., Tiwari, A., Roy, R., & Ruta, D., 2007, Computer-assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10), 2902–2917.

[2]. Kim, H. S., & Yoon, C. H., 2004, Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, 28(9–10), 751–765.

[3]. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B., 2012, New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

[4]. Huang, Y., & Kechadi, T., 2013, An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*, 40(14), 5635–5647.

[5]. Jian, C., Gao, J., & Ao, Y., 2016, A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, 193, 115–122.

[6]. Pendharkar, P., 2009, Misclassification cost-minimizing fitness functions for genetic algorithm-based artificial neural network classifiers. *Journal of the Operational Research Society*, 60(8), 1123–1134.

[7]. Chawla, V., Lyngdoh, T., Guda, S., & Purani, K., 2020, Systematic review of determinants of sales performance: Verbeke et al.'s 2011 classification extended. *Journal of Business & Industrial Marketing*, 35(8), 1359–1383.

[8]. Dölarslan, E. S., 2014, Assessing the effects of satisfaction and value on customer loyalty behaviors in service environments: High-speed railway in Turkey as a case study. *Management Research Review*, 37(8), 706–727.

[9]. Gómez, B., Arranz, A., & Cillán, J., 2006, The role of loyalty programs in behavioral and affective loyalty. *Journal of Consumer Marketing*, 23(7), 387–396.

[10]. Hofman-Kohlmeyer, M., 2016, Customer loyalty programs as a tool of customer retention: A literature review. *In CBU International Conference Proceedings* (Vol. 4, p. 199).

[11]. Meyer-Waarden, L., 2008, The influence of loyalty programme membership on customer purchase behavior. *European Journal of Marketing*, 42(1/2), 87–114.

[12]. Sucki, O., 2019, Predicting customer churn with machine learning methods: Case study of private insurance customer data (Master's thesis). *Lappeenranta–Lahti University of Technology*.

[13]. Chen, K., Hu, Y. H., & Hsieh, Y. C., 2015, Predicting customer churn from valuable B2B customers in the logistic industry: A case study. *Information Systems and E-Business Management*, 13(3), 475–494.

[14]. Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E., 2018, Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, Article 7068349.

[15]. Grossi, E., & Buscema, M., 2007, Introduction to artificial neural networks. *European Journal of Gastroenterology & Hepatology*, 19(12), 1046–1054.

[16]. Jhaveri, S., Khedkar, I., Kantharia, Y., & Jaswal, S., 2019, Success prediction using random forest, CatBoost, XGBoost, and AdaBoost for Kickstarter campaigns. *In Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC 2019)* (pp. 1170–1173).

[17]. Coussement, K., & Van den Poel, D., 2008, Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3), 6127–6134.

[18]. Freed, M., & Lee, J., 2015, Krylov iterative methods for support vector machines to classify galaxy morphologies. *Journal of Data Analysis and Information Processing*, 3(3), 72–86.

[19]. Singh, H., & Samalia, H. V., 2014, A business intelligence perspective for churn management. *Procedia – Social and Behavioral Sciences*, 109, 51–56.

[20]. Dingli, A., Marmara, V., & Fournier, N. S., 2017, Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International Journal of Machine Learning and Computing*, 7(5), 128–132.

[21]. Anjum, A., Usman, S., Zeb, A., Afridi, I., Shah, P. M., Anwar, Z., Raza, B., Malik, A., & Malik, S., 2017, Optimizing coverage of churn prediction in the telecommunication industry. *International Journal of Advanced Computer Science and Applications*, 8(6), 145–153.

[22]. Xiahou, X., & Harada, Y., 2022, B2C e-commerce customer churn prediction based on K-means and support vector machine. *Theoretical and Applied Electronic Commerce Research*, 17(2), 458–475.

[23]. Amatare, S. A., & Ojo, A. K., 2020, Predicting customer churn in the telecommunication industry using convolutional neural network model. *IOSR Journal of Computer Engineering*, 22(3), 1–7.

[24]. Seymen, O. F., Dogan, O., & Hiziroglu, A., 2021, Customer churn prediction using deep learning. *In Advances in Intelligent Systems and Computing* (Vol. 1383, pp. 520–529).

[25]. Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M., 2020, A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588–596.

[26]. Sabbeh, S. F., 2018, Machine-learning techniques for customer retention: A comparative study. International *Journal of Advanced Computer Science and Applications*, 9(2), 273–281.

[27]. Bentejac, C., Csörgő, A., & Martínez-Muñoz, G., 2019, A comparative analysis of XGBoost. *Artificial Intelligence Review*, 52(2), 1937–1967.

[28]. Lai, S. B. S., Shahri, N. H. N. B. M., Mohamad, M. B., Rahman, H. A. B. A., & Rambli, A. B., 2021, Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data. Mathematics and Statistics, 9(3), 379–385. https://doi.org/10.13189/ms.2021.090320

[29]. Tsai, C. F., & Lu, Y. H., 2009, Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10), 12547–12553.

[30]. Cenggoro, T. W., Wirastari, R. A., Rudianto, E., Mohadi, M. I., Ratj, D., & Pardamean, B., 2021, Deep learning as a vector embedding model for customer churn. *Procedia Computer Science*, 179, 624–631.

[31]. Kiangala, S. K., & Wang, Z., 2021, An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications*, 4, 100024.

[32]. Xiahou, X., & Harada, Y., 2022, Customer churn prediction using AdaBoost classifier and BP neural network techniques in the e-commerce industry. *American Journal of Industrial and Business Management*, 12(3), 277–293.

[33]. Fang, K., Jiang, Y., & Song, M., 2016, Customer profitability forecasting using big data analytics: A case study of the insurance industry. *Computers & Industrial Engineering*, 101, 554–564.

[34]. Raosoft, Inc. 2004, Sample size calculator. http://www.raosoft.com/samplesize.html