Systems and Computing
Volume 1, Issue 1, 27-43, 2025.
https://doi.org/10.64409/sycom.v1.i1.14

Research Article Open Access

Customer Churn Prediction in Neobanking System Using Predictive Analytics and Feature Selection

Abdulrauph Olanrewaju Babatunde[1], Sheriffdeen Ade Yinusa[1,2*], Idowu Dauda Oladipo[1], Ayisat Wuraola Asaju-Gbolagade[1]

1. *Department of Computer Science, University of Ilorin, Ilorin, Nigeria.*
2. *Department of Computer Science, West Mindlands Open University, Lagos, Nigeria.*

Corresponding author: Sheriffdeen Ade YINUSA (yinusa.etutor@westmidlands.university)

Manuscript Review Record:
Submitted:
June 24, 2025

Accepted:
July 6, 2025

Published:
July 19, 2025

Cite This:
A. O. Babatunde, S. A. Yinusa, I. D. Oladipo , A. W. Asaju-Gbolagade, "Customer churn prediction in neobanking system using predicive analytics and feature selection". *Systems and Computing*, Volume 1, Issue 1, 27-43, 2025.
https://doi.org/10.64409/sycom.v1.i1.14

Abstract- Context: Customer behavior, including loyalty and satisfaction, is increasingly volatile due to rapid technological and market changes. Neobank startups, in particular, face significant challenges related to customer churn, which can severely impact profitability and reputation. **Objective:** To develop a predictive model that identifies customers at risk of churning in the neobanking sector, enabling proactive retention strategies. **Method:** This study employed data mining techniques using three classification algorithms, Logistic Regression, Naïve Bayes, and Decision Tree, implemented in the WEKA platform. Feature selection methods based on accuracy, precision, and correlation were applied to identify key churn indicators. **Results:** The Decision Tree algorithm outperformed the others, achieving an accuracy of 80.5%. It demonstrated superior performance, particularly when all features were included in the model. Key predictive features were successfully identified through feature selection techniques. **Conclusions:** The findings confirm that Decision Trees are effective in predicting customer churn in neobanks. Understanding and targeting the key factors influencing churn can help neobanks retain customers and maintain a competitive edge.

Keywords: Classification, Customer Churn, Churn prediction, Data Mining.

Acknowledgement

All praise is due to Almighty Allah for His guidance, wisdom, and the strength to complete this MSc programme. I deeply appreciate my supervisor, Dr. Abdulrauph O. Babatunde, for his mentorship and support. Gratitude goes to the PG Coordinator, Dr. R. M. Abdulraheem, and the HOD, Prof. R.O. Oladele.

1. Introduction

Customer churn, also referred to as attrition, is the loss of customers who discontinue using a service or switch to a competitor. In the context of neobanking, which represents digital-only banks that operate without physical branches, churn presents a significant challenge. While neobanks offer convenience and innovative digital services, they are just as susceptible to customer attrition as traditional banks. In Nigeria, the emergence of digital banks has improved financial inclusion and accessibility; however, customer retention remains a persistent challenge due to intense market competition and constantly shifting customer preferences. To mitigate churn, neobanks rely on predictive modeling to identify customers who are likely to disengage. Machine learning enables the analysis of behavioral and transactional data, offering timely insights for implementing customer retention strategies [1]. Churn prediction is a well-studied problem in the telecommunications sector, where customer switching is driven by service dissatisfaction and enhanced by mobile number portability (MNP) [2]. In Nigeria, MNP has made switching providers easier since 2001, resulting in higher churn rates. These insights are increasingly relevant to the banking sector, where customer attrition can severely impact revenue, brand trust, and customer lifetime value [3].

Retaining existing customers is generally more cost-effective than acquiring new ones, and understanding churn drivers is essential to maintaining profitability. Prior studies show that service dissatisfaction, poor pricing, and superior alternatives are key factors influencing customer exits [4]. The loss of high-value customers can be especially damaging. While telecom research has provided foundational insights, the banking industry, particularly neobanking, has yet to fully capitalize on advanced churn prediction models[5]. Notably, heterogeneous ensemble techniques have demonstrated superior performance in predictive tasks but remain underutilized in the neobanking domain [6]. As Nigeria's neobank ecosystem continues to grow, churn typically manifests as account inactivity or closure. Predicting and managing churn is crucial for maintaining financial performance, operational stability, and competitive positioning [7]. For example, with 3.7 billion real-time payment transactions recorded in 2022, Nigerian neobanks face direct revenue risks when customers disengage [8]. Churn reduction strategies not only help stabilize income but also enhance customer satisfaction and reinforce brand loyalty. The future of digital banking in Nigeria depends on building strong customer relationships through data-driven insight and targeted interventions [9].

This study addresses the pressing need to develop an effective churn prediction model for Nigeria’s neobanking sector. Using machine learning techniques, the research aims to design and implement a predictive system that can accurately identify customers at risk of churning. The proposed model is built and tested using WEKA software and evaluated using standard performance metrics. In addition, the study compares the proposed model with related works to validate its effectiveness. By focusing on user demographics, transactional behavior, and satisfaction indicators, the study contributes meaningful insights into predictive retention strategies for neobanks. The findings of this work are intended to assist Nigerian neobanks in reducing churn by applying interpretable and efficient machine learning techniques. These insights can inform the design of proactive retention strategies, ultimately improving customer satisfaction and overall business performance. Moreover, this research adds to the existing literature on churn prediction and provides a foundation for further exploration of AI-driven solutions in digital financial services.

The remainder of this paper is organized as follows: Section 2 presents the materials and methodologies, including data preparation and model selection; Section 3 outlines the experimental results and performance evaluation; and Section 5 concludes the study and offers directions for future research.

2. Materials and methodologies

2.1 The Proposed Architecture

The proposed architecture in this research illustrated in Figure 1. shows the steps followed in the Prediction and Analysis of Customer status from neobank Customers dataset. In the proposed architecture, the Customer data were taken as input and stored in the dataset and this dataset contains the neobank Customers 12 data list who are currently active and not active status. To validate the proposed model, we use the neobank Customers dataset [10]. The proposed architecture passes through six main stage that come after Business Understanding meaning understanding the problem to select the relevant attribute and Data understanding it construct tasks for relevant attribute used customer prediction,

then Data Preprocessing tasks to clean attribute that exist in datasets, Data Splitting: 80% for Training of the entire dataset used to train the model and 20% for Test dataset used to evaluate the performance of the model, Classification that used three (3) supervised ML such as Logistic regression, Naive Bayes and Decision Tree to obtain the Train model, then Train model and the Test dataset are the same pass to the customer prediction output, finally customer prediction model predict customer churn or Not Churn.

2.2 Data preparation

Data preparation is the primary tasks that highly determine the Machine Learning results. The model built mainly depends on how thoroughly and carefully the necessary data is obtained, analyzed and preprocessed. After description of the original data the next subsequent sections present the selected relevant attribute, and preprocessing tasks done to clean attribute in the dataset were performed.

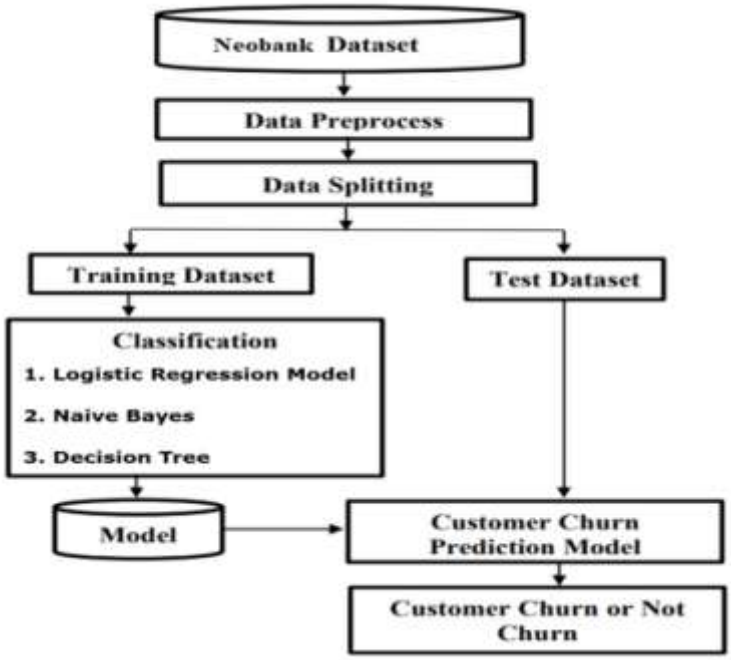

Figure 1. Proposed architecture for neobank customer churn prediction model

To conduct the machine learning (ML) project successfully, this study includes all available information about customer related data which are recommended by the neobank managers, digital agents, Expertise, and literatures such as:

1 Customer Demographics data
2 Customer Behavioral data
3 Customer Transactional data
4 Customer Account data

Customer Demographics (CD): This is the geographic and population data of a given customer or, information about a group living in particular area. As listed customer-related Demographic predictors (e.g., income, age, sex, region, status, marital status, occupation.

Customer Behaviors: This is any behavior related to a customer's bank account. It used to help make key business decisions via market segmentation and predictive analytics. This information is used by businesses for direct marketing, site selection, and customer relationship management. Such predictors are as product, customer support, and revenue from customers, frequency, subscription packages, and customers' type and so on.

Customer Transactional Data: Transactional data is broadly defined as information that records exchanges of number of debits and credits. The transaction details such as date, location, and amount spent, how much amount is debited or credited information and even the category the transaction were left in the research because it has been done by the researcher in [11].

2.3 Business understanding

This phase is also known as Problem understanding and entails the processes used to comprehend the opportunities and business purposes of the company. Understanding of the business domain under investigation is a very important step for clearly stating the business objectives, the ML goal, and for assessing the current situations of the business.
To understand the business domain of Nigeria's Neobank and coin Machine Learning problems - the researcher acquired knowledge through different techniques such as:

1 Observe various neobanks' contact centers, agent-branches.
2 Interviews and discussion also made with senior managers; domain expert consultation had been made to have brief understanding on the problem area.
3 Collected dataset and information from database and online sources of various neobanks' public websites.
4 Written documentations used such as flyers, procedure, report, and magazine.

2.4 Attribute selection

Features Selection is one of the core concepts in data mining which hugely impacts the performance of your model. The data variables that shall be used to train your models have a huge influence on the performance one can achieve. Variables selection and Data cleaning should be the first and most important step of your model designing. Variables Selection is the process where one automatically or manually select those features which contribute most to your prediction variable or output in which researchers are interested in [12]. Feature selection is a technique that removes noisy and redundant features to improve the accuracy and generalizability of a prediction model. Although feature selection is important, it adds yet another step to the process of building a bug prediction model and increases its complexity [13]. As stated and suggested by [11] and through made discussion with senior domain experts in neobanks we identify, there are a few customers characteristic specific to neobank customers and characteristics factor that may be considered as the main factors for customer churn prediction. This section provides the description of the main factors used to predict the customer churn in neo bank. The customer's characteristics such as age, marital status, gender and having other sources of customer class, are the variables that can influence the customer churn's The financial related characteristics include the opening balance, years of customer for neobanks, product type, factor also significant in influencing customer churns [14].

Attribute selection from the dataset was done based on the objective of the study at hand. Hence the account number, customer's names and branch code attributes are removed in order to reduce the data to only most important ones; this would minimize the effort required for further processing.

Table 1. Attributes of dataset used for neobanks customer churn perdition

SN	Attributes	Description
1	ID	Customer Account Number
2	Age	Age of the customers
3	Sex	Gender of the customers
4	Region Status	Region of customer
5	Current Balance	Current balance in Naira on the account
6	Marital status	Married or single
7	Referral	Number of referrals
8	Customer status	Either active or Not active
9	Mobile App User	Customers uses mobile application

10	ATM user	Customers uses ATM money withdrawal
11	Customer	Is the account for individual customer?
12	Business	Is the account for individual business or agent?

After understanding the problem to be addressed, the next step was understanding and analyzing the available data. The outcome of Machine Learning and knowledge discovery heavily depends on the quantity of the available data. To do this the researcher included domain expert, data analysts, and Database Administrator for understanding and preparing data for mining. Domain experts and data analysts specify the data required for solving the problem. The dataset used in this research was customers' related real data; the data is collected from one of the neobanks data warehouse department. The dataset consists randomly sampled 600 customer's information from June 2022 up to Dec 2022 with twelve (12) attributed [15]. The bank as a policy does not expose the privacy of its customers in any way. So, fields such as: Account Number, customer code, Name, Address, Telephone Number and other sensitive information are filtered by the bank's Data Base Administrator themselves. Such data also do not have contribution for the outcomes of the research.

Several fundamental issues related to the data have a significant impact on the quality of the outcome of a knowledge discovery process. In this regard, the initial dataset has been statistically described and visualized using Microsoft Excel to examine the properties of the whole dataset records and to obtain high level information regarding the Machine Learning questions. Simple statistical analysis has been performed to verify the quality of the dataset, addressing questions such as: Does the data cover all cases required? Is the data correct or does it contains errors? Are there missing values in the data? See more about the statistical descriptions of all attributes in the initial dataset.

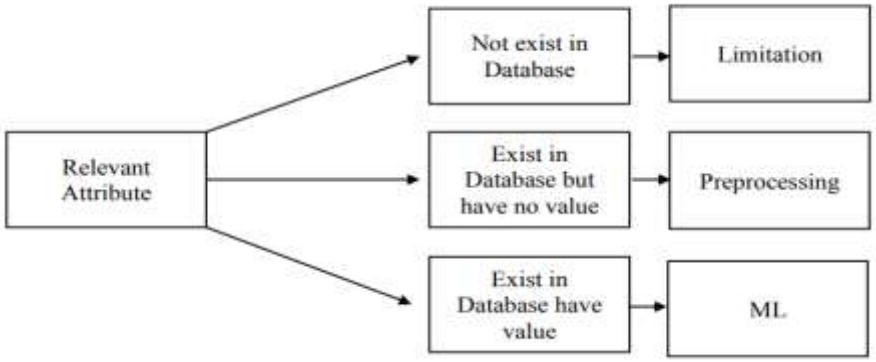

Figure 2. Optimal process flow of the selected attribute through relevant phases

2.5 Data collection

For this research, all the data was collected from a recognized neobank company. The data included new attributes which are not considered in prior studies that include Demographic data of customers, categories of customer's class and total years of customership with neobanks. The dataset also included additional attributes such marital status, ownership, industry, product type, and sector. Records contain a dependent field representing either a "churn customer" or "Not Churn customer's field. The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. The research explored different non Machine Learning studies related to customer churn analysis and prediction using financial data and customer churn assessment strategies in commercial banks. This helped us to ensure whether attributes (column names) are complete and adequate for prediction of customer churn. For this research the data from neobanks contained required customer details for analysis and prediction of customer churn.

Moreover, data representation and aggregations of the data is necessary because it minimize the variations of the attribute values in some of the fields and also to make results more expressive and simply interpretable. In the dataset, age attribute is varying in values and it is transformed in to more aggregated values "young, "adult" and old" for the different age groups determined based on experiences and consultation with the domain experts. Hence young is con-

sidered to be between 18-25, while adult is considered to be between 26-55 and old age was considered to be those with ages 55 and above. From the source of the dataset, the names and values of the attributes have been changed to some generic symbols for the sake of simplicity for experiment and to have a more accurate representation of the variables. Below it shows the short summary of the changed variable names and values which are used throughout the experiment and analysis.

2.5.1. Data preprocessing

The data preprocessing refers preparing the dataset in the form that it is ready to Machine Learning task. In this research the processes applied include data cleaning, parsing, data selection and aggregating on the extracted data in order to make the data more suitable for the experiment to improve the overall Machine Learning task. Data preprocessing which includes Data Preprocessing (Data Cleaning and Data Set Splitting). The Data Preprocessing deployed combining datasets, choosing part of the data as subgroup, combining rows, developing new columns, arranging the data to be used in the modeling, taking care of problematic figures (blank or missing) and dividing into training and test datasets.

2.5.2. Data cleaning

Here tasks are done to clean attribute that exist in Database but, requires cleaning need to fill the missing value therefore among the total dataset extracted, 29 of them have missed values for attributes such as marital status and age. Instead of filling values on these attributes, it was found easier and more logical to remove the records that make up 4.61% of the dataset. As a result, the remaining 95.38% of the original dataset which amounted to 600 records were kept for further processing.

2.5.3. Imbalance data and splitting the dataset

An imbalance dataset is such a case where there is major difference in the number of classification categories as stated by [11]. In our dataset domain, the classification categories consist of “churn customer” and “Not Churn customers” are categorized as active and inactive customers, where the number of “Not Churn customer” cases outnumber the “churn customer” cases. Almost 57 percent of the datasets are non-churn customers (Active). In such a situation a model becomes more inclined to the majority class and cannot properly identify the minority class. To solve this issue, the possibilities are either we can over sample the minority class or under sample the majority class. But under sampling the majority class will act as a difficulty in properly understanding the trends in our independent attributes. Furthermore, only over sampling the minority class will not solve this as the techniques lying behind the over sampling which also matter greatly. Thus, in such a scenario the research used Synthetic Minority Oversampling Technique (SMOTE) and up sampling [16]. This technique is used for both oversampling and under sampling. Synthetic instances of the minority class are created to reduce the margin between the majority and minority class. For the model the research used up sampling to increase the minority class and keep equal number of churned and non-churned. But we have to mention that up sampling was only applied on the training set keeping the test set pure and untouched. And therefore, this helped us to properly classify the borrowers keeping the model aware of both the output classes [17].

For the purpose of checking the performance of any Machine Learning model in an effective manner, splitting the dataset is a fundamental task. It helps to prevent over fitting by evaluating the performance of the model on a portion of the dataset upon which the model has not been trained. In most empirical studies like. It is shown that the dataset is split in to 70:30 ratios which has become most common in many other studies. Therefore, this research used 70:30 train-test split ratio for the supervised model.

2.6 System layout of ML techniques

This section describes the details of the study, methodologies used, and modules. The system involved in the analysis of customer churning uses three (3) different algorithms mentioned below.

1. Logistic regression model

2. Naïve bayes
3. Decision tree

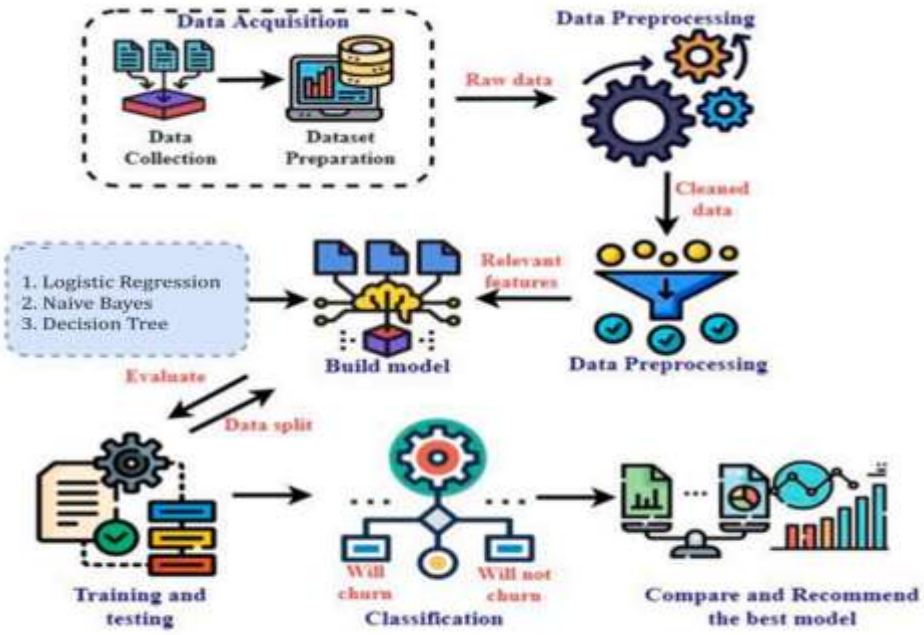

Figure 3. System layout

2.6.1 Logistic regression model

The logistic Regression can classify our observations because the client "will churn" or "won't churn" from the platform. This architecture is depicted in Fig. 4. This model will attempt to figure out the likelihood of happiness in at least one cluster or another.

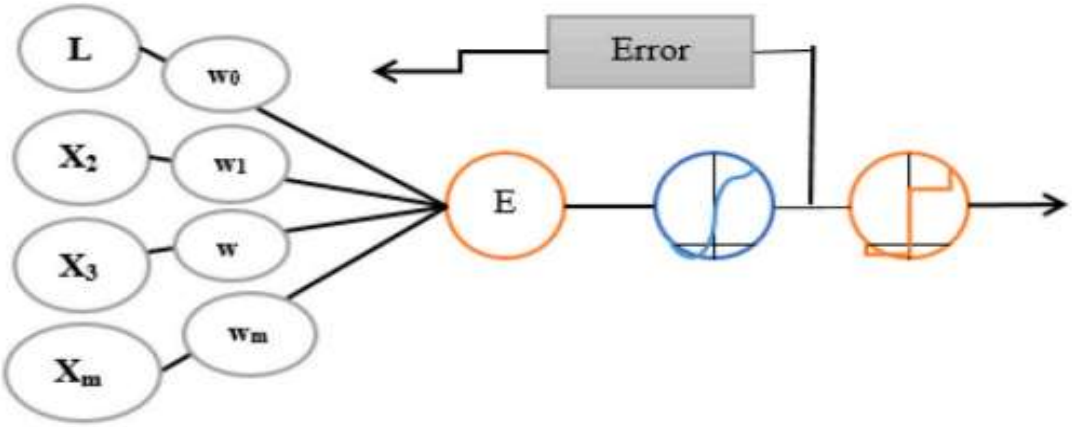

Figure 4. Logistic regression architecture

Steps carried in LR calculation:
Step 1. Import the necessary libraries
Step 2. Peruse and get the information
Step 3. Exploratory Data Analysis
Step 4. Information Preparation

Step 5. Building Logistic Regression Model
Step 6. Making Predictions on Test Set
Step 7. Appointing Scores according to anticipated likelihood values

2.6.2 Naïve bayes

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these useful for very large data sets. Along with simplicity, Naïve Bayes is known to outperform even highly sophisticated classification methods.

2.6.3 Decision tree classification algorithm

The Decision Tree is a supervised learning technique applicable to both classification and regression tasks, although it is most commonly used for classification problems. It is structured as a tree, where internal nodes represent features from the dataset, branches correspond to decision rules, and each leaf node indicates an outcome or class label. A typical decision tree consists of two types of nodes: decision nodes and leaf nodes. Decision nodes are used to split the data based on specific feature values and can have multiple branches. Leaf nodes, on the other hand, represent final outcomes and do not branch any further. Decisions in a tree are made based on the feature values of the input data. This model provides a graphical representation that outlines all possible outcomes for a given decision-making process under defined conditions. The name "decision tree" comes from its resemblance to a natural tree structure, it starts with a root node, which branches out recursively into sub-nodes, ultimately forming a tree-like diagram. To construct a decision tree, the CART (Classification and Regression Tree) algorithm is commonly used. At each step, the algorithm asks a question based on one of the features (e.g., "Is balance > 5000?"). Depending on the answer (yes or no), the tree splits into subtrees, and this process continues until a stopping criterion is met. With many algorithms available in machine learning, selecting the most suitable one for a particular dataset and problem is crucial. The Decision tree is often favored for the following two primary reasons:

1 Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
2 The logic behind the decision tree can be easily understood because it shows a tree-like structure.

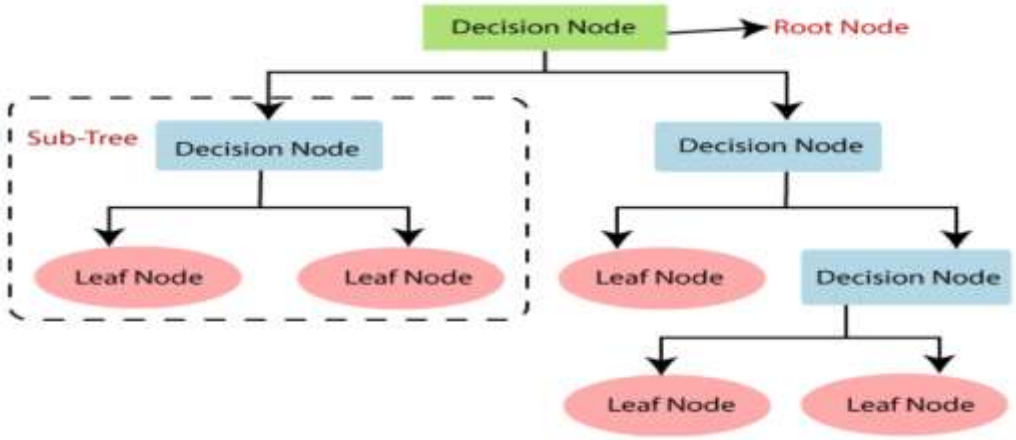

Figure 5. Decision trees architecture

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with

the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

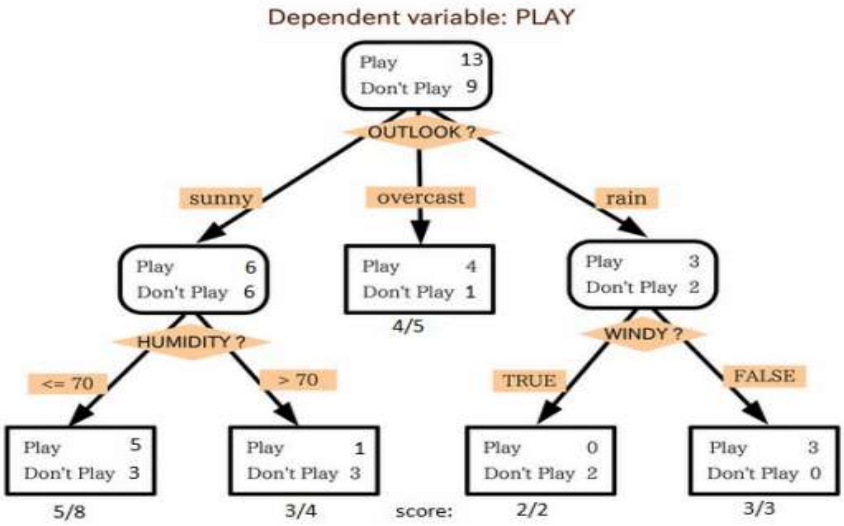

Figure 6. Decision trees classification architecture

3. Results

In this research work, the data mining tool weka was used. The dataset was loaded into the weka explorer. The classify tab enables the user to apply various data mining algorithms like, classification and regression on the dataset. To evaluate the performance of the algorithm various measures like, confusion matrix, evaluation time, correctly classified instances, in correctly classified instances and the error performance measures like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE) error performance measures are all calculated by the tool. The algorithms Logistic regression, Naive Bayes and Decision Tree are executed on the data set to predict the customer churn and to evaluate the performance efficiency of these algorithms, the above said measures has been considered. After applying various algorithms on the dataset, performance of the respective models was analyzed. Some measures were used to predict the customer churn in the dataset. The following are the measures that were used to analyze the performance of the classifiers: Correctly Classified instances, incorrectly classified instances, time taken to build model, MSE, RMSE, RAE, RRSE.

3.1 The experimental setup

Weka, a free software written under the JAVA General Public License that provides a collection of machine learning algorithms, was used in the study experiments. Initially, the dataset was processed to prepare for the experiment. Three attributes were removed from the data using Excel, and then outliers were detected and removed using two filters (Interquartile Range and RemoveWith-Values). The issue of data imbalance was resolved in the quest to optimize classification, considering the cost of errors and overfitting issues, which can also lead to suboptimal results due to the high costs associated with misclassification of the minority class. The transformation step was critical to improving the model's performance and making the features easier to understand. The class value was transformed from a numeric value to a nominal value by WEKA using an unsupervised attribute filter (NumericToNominal). It was then combined

with two attributes, and the labels for one attribute were encoded. Additionally, by binning the data, two numeric attributes were discretized; this step was completed using an unsupervised attribute filter (Discretize). The classifiers were evaluated on the selected attributes using the the10- fold cross-validation and optimum parameters for each classifier. Table 2 to Table 18 illustrate the results. With these results, different partition ratios were used to implement the classifiers. The highest accuracy for logistics regression, Decision Tree and naïve bayes were found to be achieved with a (70:30) ratio, 70 percent for training data and 30 percent for testing data. Tables shown below summarize the findings. Finally, the final studies have been performed using the optimum options for the best subset features to achieve the best results in terms of cross-validation or partition ratio outcome.

3.2 Predictive modelling

Predictive modeling is the general concept of building a model that is capable of making predictions [19]. Typically, such a model includes a Machine Learning algorithm that learns certain properties from a training dataset in order to make those predictions. Predictive modeling is a name given to a collection of mathematical techniques or models that helps in finding a mathematical relationship between a target or dependent variables and the predictor or independent variables [11]. It helps in predicting the probability of an outcome when a set of independent variables passes through the model. Logistic Regression and Naïve Bayes Models can be used for prediction purpose.

The principal goal of this research is to analyze the existing neobank customer behavior and predict customer churns using computational algorithms. With this in mind, the research aimed at identifying Machine Learning technique which is better in predicting customer churn. Throughout this work different Machine Learning algorithms were explored and effective ones were used to find model in the data. The performance of each Machine Learning models was explored and analyzed in previous related works. Then based on successfulness in making predictions and stability in their best performing algorithms were selected. The Machine Learning techniques selected for this thesis were Logistic regression, Naive Bayes, and Decision Tree. Each one is selected based on their advantages and past performance seen in other research. In different literatures, it has been reported that they were widely used classifier algorithms for prediction and classification are Logistic regression, Naive Bayes, and Decision Tree. The research was use the above three (3) different algorithms to build different models for this customer churn prediction and classification model [12].

Naïve Bayes is selected due to the following reasons; it is easy to implement, Naïve Bayes classifiers can be trained quickly, classification process is quick compared to other models, it can handle a large and discrete amount of data, it is not sensitive to irrelevant features. Logistic Regression is a predictive analysis. It takes independent features and returns output as categorical output. The probability of occurrence of a categorical output can also be found by Logistic Regression model by fitting the features in the logistic curve. Logistic Regression is included in this work because; it is easy to implement and no linear relationship between independent and dependent variable, multiple explanatory variables can be used, no confounding effects because Logistic Regression allows quantified values for strength of association between explanatory variables and less prone to over-fitting due to simplicity and low variance.

3.3. Evaluation methods and proposed metric

In this study, to produce more accurate results, four performance measures were considered to evaluate each classifier: accuracy, precision, recall, and f-measure. All of these measures are based on the following possibilities:

- TP: True Positive is the total of instances that a churn customer was correctly classified.
- FP: False positive is the total of instances that a churn customer was incorrectly classified.
- TN: True Negative is the total of instances that a non_churn customer was correctly classified.
- FN: False negative is the total of instances that a non_churn customer was incorrectly classified.

$$\text{Accuracy: } \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$\text{Precision: } \frac{TP}{TP+FP} \tag{2}$$

$$\text{Sensitivity: } \frac{TP}{TP+FN} \tag{3}$$

$$\text{Specificity: } \frac{TN}{TN+FP} \tag{4}$$

For training and testing of data set, holdout method is used with 70 – 30% (training – test) rates. To evaluate model performance, accuracy, sensitivity, specificity, precision, and F-score values are considered, churn customers' class labels are used as reference values. In prediction studies such as churn analysis, it is important to identify churn customers rather than identifying the classes of all samples. When calculating the accuracy, the number of classes predicted accurately is checked for all instances, regardless of the class. For example, 100 of the records in a dataset are churn customers, of which 900 of them belong to non-churn customers. Consider that the class predictions of 900 records in the dataset are correct and only 10 of these accurate predictions are churned. In general, this model has an accuracy rate of 90%; however, the true prediction rate for churn customers is only 10%. Therefore, a modified accuracy calculation approach is presented in this study. The equation for measuring accuracy is given below:

$$\text{Modified Accuracy} = \text{MA} = \frac{TP}{TP+FN+FP} \tag{5}$$

According to this equation, customers with no churn status and those correctly classified by the model are excluded from the accuracy rate formulation.

3.4. Modeling with WEKA

3.4.1 Modeling using logistic regression

In this research work, a Prediction model building is done using Logistic Regression algorithm. Logistic Regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name Logistic Regression is used when the dependent variable has only two values [20], such as 0 and 1 or Yes and No. The main thing here is to determine a mathematical equation that can be used to predict the probability of event 1(Not Churn). After the equation is established, it can be used to predict the Y (customer status) attributes when only the X's (other 15 attributes) are known using selected dataset.

Table 2. Detailed accuracy by class for logistic regression

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.526	0.285	0.608	0.526	0.564	0.245	0.661	0.628	YES
	0.715	0.474	0.642	0.715	0.676	0.245	0.661	0.672	NO
Weighted Avg.	0.628	0.388	0.626	0.628	0.625	0.245	0.661	0.652	

Table 3. Table of confusion matrix for logistic regression

Actual Class	**Predicted Class**	
	Yes (1)	No (0)
Yes (1)	240 (TP)	217 (FN)
No (0)	155 (FP)	388 (TN)

Table 4. Summary of stratified cross-validation (1)

Logistic Regression		
Correctly Classified Instances	628	62.8333 %
Incorrectly Classified Instances	372	37.1667 %

Table 5. Summary of stratified cross-validation (2)

Mean Absolute Error (MAE)	**Mean Absolute Error (MAE) Percentage**	**Root Mean Squared Error (RMSE)**	**Root Mean Squared Error (RMSE)**
0.4483	90.3342 %	0.4787	96.0951 %

Table 6. Cogent summary of stratified cross-validation (3)

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Accuracy	Time Taken (second)
Logistic Regression	628	372	0.2429	62.8%	0.07

3.4.2 Modeling using naive bayes

In this research work, a Prediction model building is done using Naive Bayes Classification algorithm and WEKA used to predict the customer.

Table 7. Detailed accuracy by class for naïve bayes

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.544	0.279	0.621	0.544	0.580	0.269	0.703	0.650	YES
	0.721	0.456	0.653	0.721	0.685	0.269	0.703	0.734	NO
Weighted Avg.	0.640	0.375	0.638	0.640	0.637	0.269	0.703	0.696	

Table 8. Table of confusion matrix for naïve bayes

Actual Class	Predicted Class	
	Yes (1)	No (0)
Yes (1)	248 (TP)	208 (FN)
No (0)	152 (FP)	392 (TN)

Table 9. Summary of stratified cross-validation (1)

Naïve Bayes		
Correctly Classified Instances	640	64 %
Incorrectly Classified Instances	360	36 %

Table 10. Summary of stratified cross-validation (2)

Mean Absolute Error (MAE)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Root Mean Squared Error (RMSE)
0.4248	85.5903 %	0.4696	94.2684 %

Table 11. Cogent summary of Stratified cross-validation (3)

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Accuracy	Time Taken (second)
Naïve Bayes	640	360	0.2673	64 %	0.02

3.4.3 Modeling using decision tree

In this research work, a Prediction model building is done using Decision Tree Classification algorithm and WEKA used to predict the customer.

Table 12. Detailed accuracy by class for decision tree

	TP Rate	FP	Precision	Recall	F-Measure	MCC	ROC	PRC	Class

		Rate					Area	Area	
	0.763	0.160	0.801	0.763	0.781	0.606	0.865	0.815	YES
	0.840	0.237	0.808	0.840	0.824	0.606	0.865	0.862	NO
Weighted Avg.	0.805	0.202	0.805	0.805	0.805	0.606	0.865	0.841	

Table 13. Table of confusion matrix for decision tree

Actual Class	Predicted Class	
	Yes (1)	No (0)
Yes (1)	348 (TP)	108 (FN)
No (0)	87 (FP)	457 (TN)

Table 14. Summary of stratified cross-validation (1)

Decision Tree		
Correctly Classified Instances	805	80.5 %
Incorrectly Classified Instances	195	19.5 %

Table 15. Summary of stratified cross-validation (2)

Mean Absolute Error (MAE)	Relative Absolute Error (RAE) Percentage	Relative Squared Error (RRSE)	Root Relative Squared Error (RMSE)
0.3032	61.095 %	0.379	76.093 %

Table 16. Cogent summary of stratified cross-validation (3)

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Accuracy	Time Taken (second)
Decision Tree	805	195	0.6056	80.5%	0.16

3.5. Comparison of Machine Learning Models

The below (Table 17) shows a summary of the result obtained from all the three (3) models when all models are trained on around 1000 instances. It has been found that Decision Tree (with accuracy of 80.5%) resulted into highest accuracy. But at the prediction level Logistic regression, Naive Bayes, and Decision Tree model performed well in terms of accuracy and precision. These machine learning models evaluations (comparison) are done through the model accuracy and precision. The three (3) Machine learning models (Logistic regression, Naive Bayes and Decision Tree) were used to predict the Neobank dataset on the 12 selected attributes.

Table 17. Comparison of ML models

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Accuracy	Precision	Time Taken (second)
Logistic Regression	628	372	0.2429	62.8%	0.626	0.07
Naïve Bayes	640	360	0.2673	64 %	0.638	0.02
Decision Tree	805	195	0.6056	80.5%	0.805	0.16

Table 18. Comparison of performance error value of various classifiers

Classifier	Mean Absolute Error (MAE)	Relative Absolute Error (RAE) Percentage	Relative Squared Error (RRSE)	Root Relative Squared Error (RMSE)

Logistic Regression	0.4483	90.3342 %	0.4787	96.0951 %
Naïve Bayes	0.4248	85.5903 %	0.4696	94.2684 %
Decision Tree	0.3032	61.095 %	0.379	76.093 %

As stated in above table, it was recorded highest accuracy with Decision Tree (80.5%) than the rest two models with Logistic Regression (62.8%) and Naïve Bayes (64%) show a comparable performance. The result of the study is presents to the domain experts those are Neobank managers and experts and they proved technical evaluation to assurance of the results of the study and quality of the data.

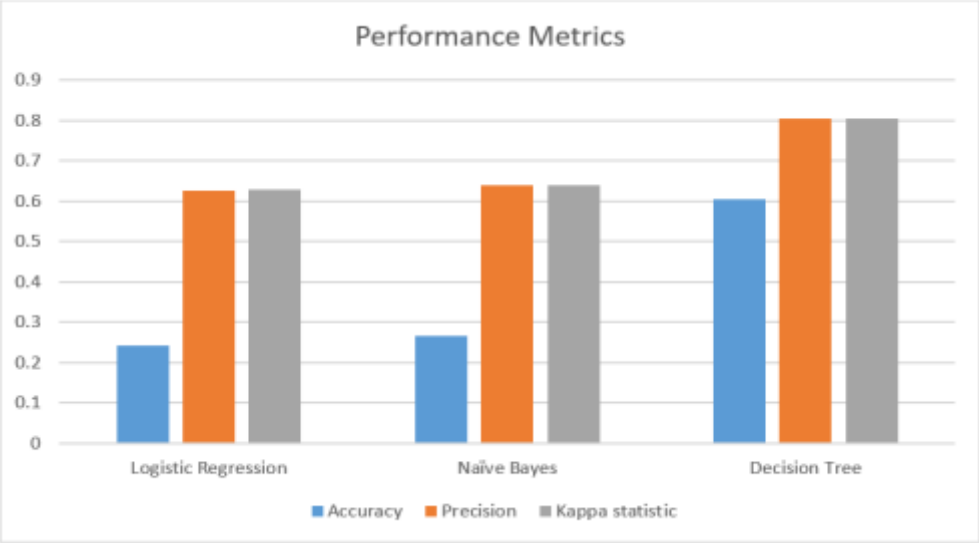

Figure 7. Comparison of performance metrics of the three models

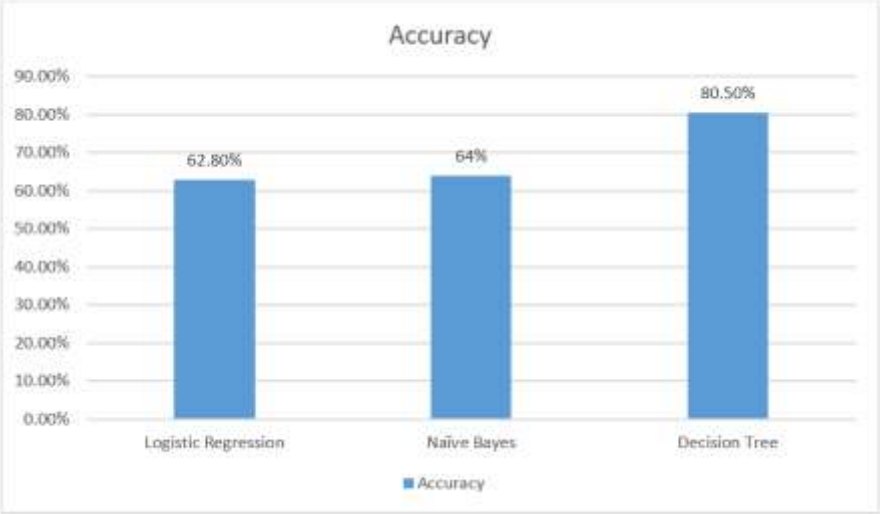

Figure 8. Comparison of performance accuracy of the models

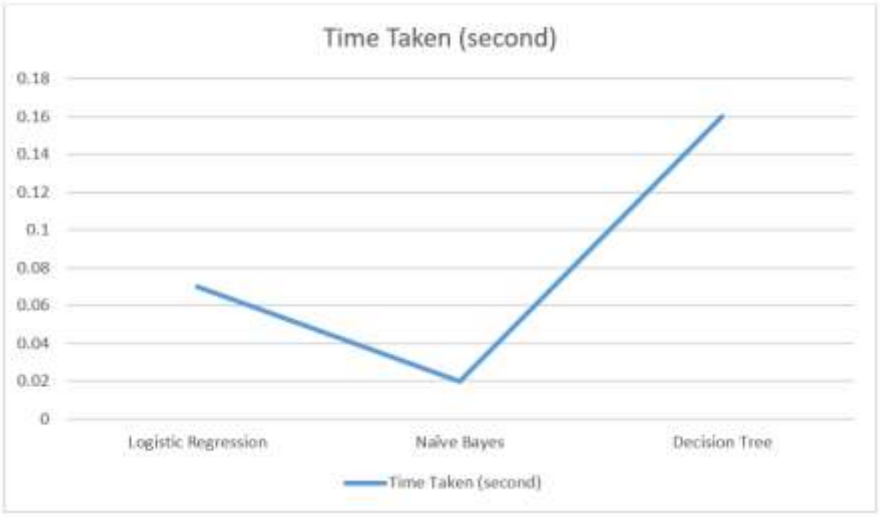

Figure 9. Comparison of time taken to build model

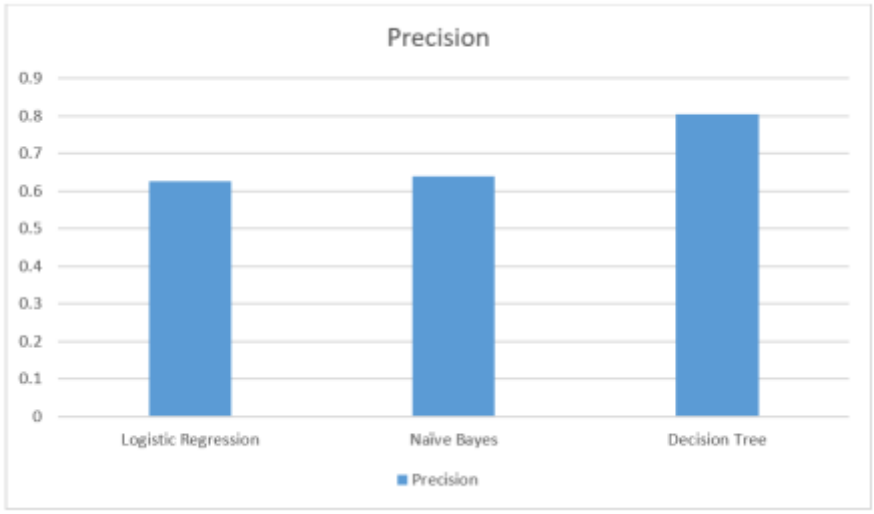

Figure 10. Comparison of Performance Precision of various classifiers

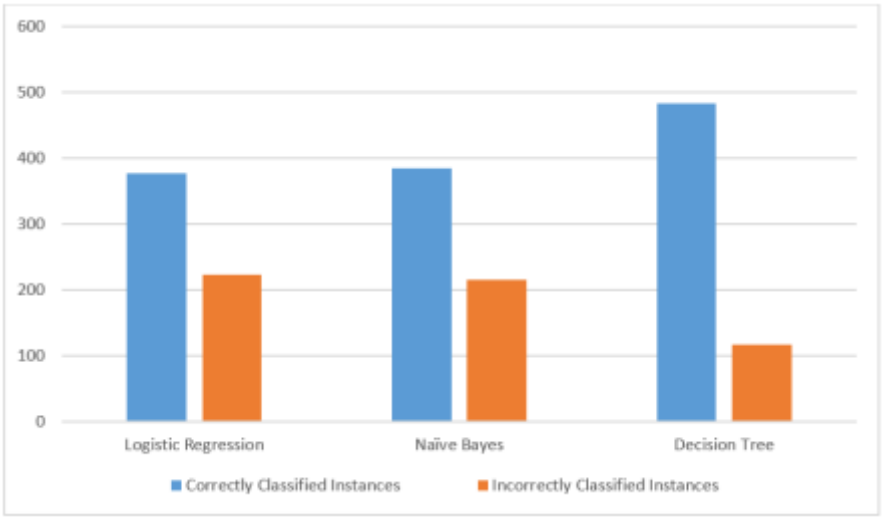

Figure 11. Comparison of Performance Accuracy of various classifiers

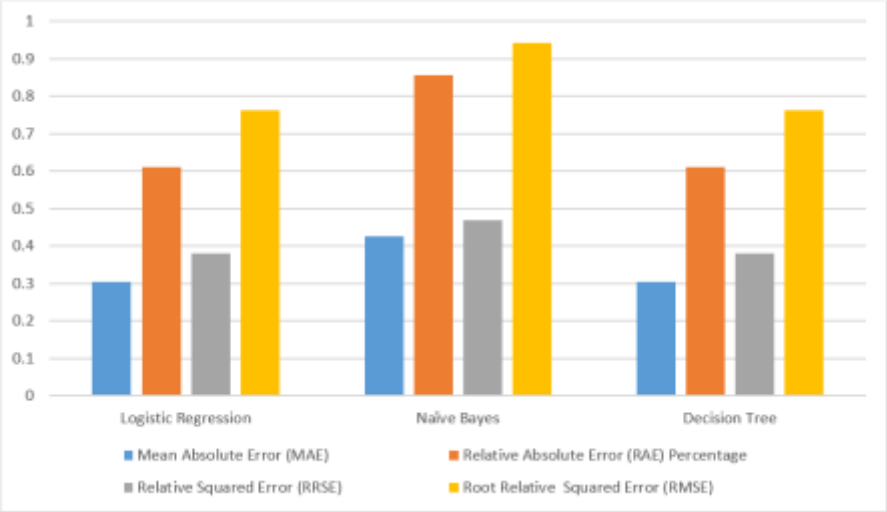

Figure 12. Comparison of performance error value of various classifiers.

The study used a dataset of 1,000 records from a neobank, cleaned and preprocessed for modeling with Logistic Regression, Naïve Bayes, and Decision Tree algorithms using WEKA. Confusion matrix, accuracy, and sensitivity were evaluated. Decision Tree performed best with 80.5% accuracy, followed by Naïve Bayes (64%) and Logistic Regression (62.8%). Key predictive features included balance, customer activity, referral, and ATM usage. The Decision Tree model was identified as the most suitable for churn prediction. Overall, all research questions were answered, confirming Machine Learning's effectiveness in predicting and managing customer churn in Nigeria's neobanking sector.

4. Conclusions and future research

This study concludes that Machine Learning (ML) algorithms can effectively predict customer churn in Nigerian neobanks with acceptable accuracy [21]. These models facilitate early detection of potential churners, thereby supporting more informed and timely investment and retention decisions. However, the performance of the models is highly dependent on the quality, representativeness, and volume of the dataset used. In this study, the use of only 1,000 data points constituted a significant limitation, potentially restricting the generalizability of the findings. Future research would benefit from incorporating larger and more diverse datasets, encompassing various neobank startups and broader customer demographics. This expansion would not only improve the robustness of the models but also enable the capture of nuanced behavioral patterns across different customer segments. Moreover, exploring additional or more advanced ML techniques—such as ensemble learning, deep learning architectures, or hybrid approaches—could lead to improved prediction accuracy and model interpretability. In conclusion, broader data coverage, methodological enhancements, and continuous model evaluation are essential steps for advancing churn prediction in Nigeria's evolving neobanking landscape.

References

[1] Y. Huang, L. Li, "Naïve Bayes Classification Algorithm Based on Small Sample Set". In *Proc. IEEE CCIS 2011 – International Conference on Cloud Computing and Intelligent Systems*,, Beijing, China, Sept. 15–17, 2011

[2] M. Bogaert, L. Delaere, "Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art". *Mathematics*, 11, 1137, 2023. https://doi.org/10.3390/math11051137

[3] G. Buchi, M. Cugno, A. Zerbetto, R. Castagnoli, "New Banks in the 4th Industrial Revolution: A Review and Typology". *Proceedings of 22nd Excellence in Services International Conference*. Thessaloniki, Greece, Aug. 29–30, 2019, pp. 47–63. Available: https://sites.les.univr.it/eisic/wp-content/uploads/2019/11/6-Buchi-Fasolo-Cugno-Zerbetto-Castagnoli-2.pdf

[4] Chih-Fong Tsai & Yu-Hsin Lu, "Customers churn prediction by hybrid neural networks". *Expert Systems with Applications*, 36, 12547–12553, 2009. http://dx.doi.org/10.1016/j.eswa.2009.05.032

[5] J. Fu and M. Mishra, "Combatting fraudulent and predatory fintech apps with machine learning," Proceedings of the *2023 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, Dec. 2023, pp. 1234–1242.

[6] B. Huang, M. T. Kechadi, B. Buckley, "Customer churn prediction in telecommunications". *Expert Systems with Applications*, 39(1), 1414–1425, 2012.

[7] IBM Corporation, "Working with telecommunications", Minimizing churn in the telecommunications industry, 2010, United States of America.

[8] L. Jie, Xu, Xu "A novel model for global customer retention using data mining technology, data mining and knowledge discovery in real life applications", Julio Ponce and Adem Karahoca (eds.), 438, I-Tech, 2009, Vienna, Austria.

[9] T. Kolajo, A. B. Adeyemo, "Data mining technique for predicting telecommunications industry customer churn using both descriptive and predictive algorithms". *Computing Information Systems & Development Informatics Journal*, 3(2), 27–34, 2012.

[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector". *IEEE Access*, 7, 60134–60149, 2019.

[11] A.V. Komarov, V.M. Martyukova, "Neobanking As A Direction Of Development Of Modern Financial Technologies". *Vestnik Universiteta*, (3):134–142, 2020. https://doi.org/10.26425/1816-4277-2020-3-134-142

[12] I. Martinčević, S. Črnjević, I. Klopotan, "Novelties and Benefits of Fintech in the Financial Industry". *International Journal of E-services and Mobile Applications*, 14(1), 1–25, 2022. https://doi.org/10.4018/ijesma.2022010107

[13] E. O. Oyatoye, S. O. Adebiyi, B. B. Amole, "Modeling switching behaviour of Nigeria global system for mobile communication multiple SIMs Subscribers` using Markov chain analysis". *Journal of Operations Management*, 14(1), 7–31, 2020.

[14] N. B. Syam, J. D. Hess, "Acquisition versus retention: competitive customer relationship management", Working Paper, University of Houston, Houston, Texas, 2006, USA.

[15] T. Xu, "Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping". *Applied Sciences*, 2021, 11, 4742. https://doi.org/10.3390/app11114742

[16] R. Saxena, "Introduction to k Nearest Neighbor Classification and Condensed Nearest Neighbour Data Reduction". Data Science, Machine Learning Journal, blog post, December 23, 2016.

[17] S. B. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview". *Artificial Intelligence Review*, 42, 157 (2014). https://doi.org/10.1007/s10462-011-9230-1

[18] S. Portela, R. Menezes, "Detecting customer defections: an application of continuous duration models". *Journal of Global Strategic Management*, 9, 22–30, 2011. http://dx.doi.org/10.20460/jgsm.2011515809

[19] Z. Lei and G. Junfeng, "Customer Classification of Discrete Data Concerning Customer Assets Based on Data Mining". Int. Conf. Intell. Transp. *Big Data Smart City*, pp. 352–355, 2019.

[20] OECD, Digital Disruption in Banking and its Impact on Competition, 2020. http://www.oecd.org/daf/competition/digital-disruption-in-financial-markets.htm

[21] P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, E. Hossain, "Investigating customer churn in banking: A machine learning approach and visualization app for data science and management". *Data Science and Management*, 7 (2024) 7–16. https://doi.org/10.1016/j.dsm.2023.09.002