*Journal of Basic Physical Research ISSN: 2141-8403 Prints, 2141-8411 Online*
*Vol.14, Issue 2, October, 2025, pp. 236– 248, Available online at https://journals.unizik.edu.ng/jbpr/*

# A BIG DATA-DRIVEN ANALYTICAL FRAMEWORK FOR PREDICTING CUSTOMER PRODUCT PREFERENCES AND CHURN IN NIGERIA

**[1]Emmanuel Cosmas Patrick and [2]Virginia E. Ejiofor**
*[1]Computer Science Department, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka*
*[1]Department of Computer Science and Dean, Faculty of Physical Sciences*
*Nnamdi Azikiwe University, Awka*
*ve.ejiofor@unizik.edu.ng & cenptech1@gmail.com*

***Abstract***
*The rapid growth of Big Data has transformed organizational decision-making and operational efficiency across industries, particularly in the telecommunications sector, where understanding customer behaviour is crucial for reducing churn and optimizing revenue. This study presents a Big Data-driven analytical framework for predicting customer product preferences and churn in MTN Nigeria, leveraging a dataset of 974 anonymized customers with 158,943 transactional records collected between 2024 and early 2025. The methodology integrates data preprocessing, normalization, class imbalance correction using SMOTE, and supervised machine learning algorithms, including Random Forest (RF), Backpropagation Neural Network (BPNN), K-Nearest Neighbours (KNN), and Naïve Bayes (NB). Models were trained and evaluated using accuracy, precision, recall, and F1-score metrics, with RF achieving the highest overall performance (accuracy = 89%, F1-score = 0.94). The framework was implemented as a Python-based prototype system with modules for customer segmentation, product targeting, churn risk prediction, and personalized marketing communication. Results demonstrate that the proposed system can accurately predict customer preferences, reduce unsolicited promotions, and improve customer retention and revenue. This study highlights the potential of integrating Big Data analytics and machine learning into telecom CRM workflows and provides a scalable model applicable to other sectors requiring behaviour-driven customer engagement strategies.*
***Keywords:*** *Big Data, Machine Learning, Customer Churn Prediction, Telecommunications, Personalized Marketing, Customer Segmentation*

## Introduction

Since the early 1990s, the concept of Big Data has continued to impact organizational activities and the quality of resource management. Moreover, in the early 21st century, the idea of Big Data underwent an evolutionary shift from traditional decision support systems to a form of business intelligence (O'Donovan et al., 2019). Globally, it has been recognized as a key factor in driving industrial management and operational processes, business innovations, and resource optimization to a more productive landscape due to the characteristics, which include volume, velocity, and variety of data inherent in the system (Mathrani & Lai, 2021).

Big Data refers to datasets that exceed the processing capacity of conventional database systems, requiring sophisticated techniques and technologies for data capture, storage, management, and analysis (Mach-Krol, 2022). These datasets may include structured data (such as customer records), semi-structured data (like XML files), and unstructured data (like social media feeds, videos, and emails) (Li et al., 2021). Due to the significant revelations uncovered through big data analysis, companies have increasingly relied on it to solve complex problems associated with business administration, marketing, production, and security, among other areas, especially in the current competitive business landscape (Al-Jumaili et al., 2023). One major area that has continued to adopt Big Data to facilitate decision-making is the telecommunications sector (Zhang et al., 2022). Telecom operators generate and manage an enormous amount of data from call logs, SMS, internet usage, geolocation, customer feedback,

and device types. In Nigeria, a country with over 222 million mobile phone subscribers (Statista, 2024), the telecommunications sector is not only one of the largest but also one of the most competitive. Providers must constantly innovate to retain customers, improve service delivery, and remain profitable. However, challenges such as customer churn, poor customer satisfaction, and revenue leakage persist (Pejic et al., 2021; Kim et al., 2020).

Customer churn, in particular, has become a critical concern. High churn rates result in reduced profitability and increased customer acquisition costs. Research indicates that retaining an existing customer is significantly less expensive than acquiring a new one. To tackle these problems, Big Data was applied to develop analytical frameworks that facilitated processes like digital marketing, innovative ideas, and the advertisement of new products, packages, data bundles, promotions, and bonuses that stirred up customers' interest, thereby increasing their patronage (De Bock and De Caigny, 2021). However, Xu et al. (2021) revealed that the heterogeneous nature of customer behaviour made it challenging to identify the type of customer required for a particular product.

Taking the case study of MTN Nigeria, which currently has different packages for data bundles, social media, and calls. These packages are all integrated as one, with diverse prices allocated to diverse subscription times. The need to identify the appropriate type of customer that can fit in with each of these packages in order to avoid non-cost effectiveness and nuisance has presented a complex problem (MTN, Nigeria, 2024). In a bid to solve this issue of identifying a customer's fitness for a particular package, Mach-Król (2022), Mathrani et al. (2021), and O'Donovan et al. (2019) developed Big Data analytical frameworks that were capable of improving an organization's decision-making. Though with this advancement in research, there is a gap in developing a Big Data analytics framework that considers the diversity of customers' behaviour and attributes, particularly demography, age, and other historical information, to appropriately inform them based on the type of products that fit their requirements.

Hence, this research focused on developing a comprehensive model utilizing big data, which characterized customer attributes, and then applying the necessary data analytical processes to solve issues of data imbalance, noise, feature importance, and transformation, before training machine learning algorithms to generate models for big data analysis and prediction of products that fit customer needs. This, when achieved, will improve customer patronage, address customer churn, solve issues of power service quality, and most importantly, increase revenue generated through sustainable customer patronage. In addition, it will stop the series of annoying messages sent by telecommunication companies to customers and help identify the class of customers that needs a particular message.

The expected outcome is a robust and intelligent system that is capable of dynamically aligning telecom products with customer needs and will not only improve customer satisfaction and engagement but also drive revenue growth, operational efficiency, and competitive advantage. Additionally, the proposed approach will significantly reduce the volume of unsolicited promotional messages sent to customers, replacing them with personalized offers based on actual user profiles and preferences. Hence, in the long term, such a model could be extended beyond the Nigerian context to other developing countries with similar telecommunications challenges, and it could equally serve as a foundational framework for deploying recommender systems and Customer Relationship Management (CRM) solutions in sectors such as e-commerce, banking, construction, and healthcare where understanding customer behaviour is essential for growth and sustainability.

## Methodology

The methodology adopted in this study is the Behavioral Driven Development (BDD) approach. In a digital marketing context, it is a suitable methodology for implementing an efficient Big Data analytics framework for Nigeria's telecommunication sector due to its focus on analyzing user behaviors, trends, and patterns. This approach leverages dynamic customer data on call durations, internet usage, and subscription preferences, with MTN Nigeria as the case study. Primary data comprising 974 anonymized customer records with 17 key attributes (including age, gender, state, tenure, device type, satisfaction level, subscription plans, data usage, revenue, and churn status) was collected directly from MTN Nigeria's internal transaction records in compliance with NDPR guidelines. Following collection, the dataset underwent rigorous preprocessing involving missing value imputation, categorical label encoding, numerical feature normalization, exploratory analysis (class distribution, product preferences, correlation matrices), and class imbalance correction using SMOTE on the training set only. Four supervised classification algorithms, such as K-Nearest Neighbors, Random Forest, Naïve Bayes, and Backpropagation Neural Network (MLP Classifier), were then trained on an 80:20 stratified train-test split. Model performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score, confusion matrices, ROC-AUC, and learning curves, with comparative analysis identifying Random Forest as the superior performer. The selected model was subsequently integrated into a Python-based prototype analytical system (built with pandas, scikit-learn, imbalanced-learn, and Streamlit) featuring modules for data upload, customer segmentation, product targeting, churn risk prediction, and simulated personalized marketing communication, enabling data-driven, behavior-responsive recommendations tailored to the Nigerian telecom market.

## Data Collection

The dataset used in this research was sourced from MTN Nigeria, one of the leading telecommunications service providers in the country. The data reflects customer subscription behavior and churn patterns, captured in 2024 and the first quarter of 2025. A total of 974 unique customers were considered, and a sample size of 158,943 records of these customers was used, offering a representative sample of customer interactions within the time frame under review. Each record in the dataset comprises 15 different attributes, encompassing demographic information, service usage metrics, customer account details, and churn status. These attributes provide critical insights into customer behavior and allow for an in-depth analysis of factors influencing churn within the Nigerian telecommunications industry. Table 1 presents the data description.

**Table 1: Data description**

| Attribute | Data Format | Description | Validation Rules |
|---|---|---|---|
| **Customer ID** | VARCHAR (20), PRIMARY KEY | Unique alphanumeric customer identifier | Must be unique, non-null |
| **Full Name** | VARCHAR (100), NOT NULL | Customer's legal full name | Maximum 100 characters |
| **Date of Purchase** | DATE, NOT NULL | Initial service activation date | Valid date format (YYYY-MM-DD) |
| **Age** | INTEGER, Range: 18-100 | Customer age at time of data collection | Must be between 18-100 |
| **State** | VARCHAR (50) | Geographical state of primary usage | Valid Nigerian state names |
| **MTN Device** | VARCHAR (100) | Mobile handset model and type | Free text, device names |

| | | | |
|---|---|---|---|
| **Gender** | ENUM ('Male','Female','Other') | Self-reported gender identity | Predefined categories |
| **Satisfaction Rate** | INTEGER, Range: 1-10 | CSAT score from recent surveys | 1(Very Poor)-10(Excellent) |
| **Customer Review** | TEXT, Optional | Open-ended customer comments | Maximum 500 characters |
| **Customer Tenure** | INTEGER, Calculatud | Months since first purchase | Auto-calculated from purchase date |
| **Subscription Plan** | VARCHAR (50) | Current active service plan | Must match available plans |
| **Unit Price** | DECIMAL (8,2) | Monthly recurring charge | Positive decimal value |
| **Number of Purchases** | INTEGER | Count of plan renewals/changes | Non-negative integer |
| **Total Revenue** | DECIMAL (10,2) | Lifetime customer value | Sum of all payments |
| **Data Usage** | DECIMAL (8,2) | Average monthly data usage | In gigabytes (GB) |
| **Churn Status** | ENUM ('Active',' Churned',' At Risk') | Current customer status | Based on activity rules |
| **Reasons for Churn** | TEXT, Conditional | Documented churn reasons | Only for churned customers |

**Data Processing**

Effective data processing is a crucial step in the development of accurate and reliable predictive models. In this study, the collected dataset underwent a series of preprocessing techniques to ensure that the data was complete, properly scaled, and suitably enhanced for machine learning tasks. The key procedures implemented include imputation and normalization. Data imputation was performed to address missing or incomplete values within the dataset. Incomplete data can lead to biased results or cause algorithms to malfunction. Normalization was applied to rescale numerical attributes to a standard range, typically between 0 and 1. Min-Max Scaling was selected, given its effectiveness in compressing the range of data without altering the distribution shape. To address the potential issue of class imbalance often observed in churn datasets where non-churners dominate the data, data augmentation techniques were introduced. Synthetic Minority Oversampling Technique (SMOTE) was applied to generate artificial samples for the minority class (churners). This method synthesizes new samples by interpolating between existing minority class instances, preserving meaningful variance within the class.

**The Proposed Machine Learning Algorithms**

This section outlines the set of algorithms employed in the development of the intelligent telecom customer management system. The primary algorithms implemented include supervised learning techniques such as Gradient Descent Back Propagation Neural Networks, Random Forest, K-Nearest Neighbors, and Naïve Bayes.

**1. The Gradient Descent Back-Propagation Neural Network Algorithm**

1. Load data set
2. Normalize feature values
3. Split the dataset into training and testing sets.

4. Network Initialization
5. Define architecture: input layer, one or more hidden layers, output layer.
6. Randomly initialize weights and biases for all neurons.
7. Forward Propagation
8. For each input, compute weighted sum for each neuron: $z = w_i x_i + b$
9. Apply activation function to each (z).
10. Pass signals through layers to calculate the output.
11. Loss Calculation
12. Compute error using a loss function
13. Compute gradients of loss considering weights and biases
14. Update weights and biases using gradient descent:
15. $w_{new} = w_{old} - \eta \frac{\delta L}{\delta w}$
16. Where (ŋ) is the learning rate.
17. Repeat Training
18. Repeat forward propagation and backpropagation for all epochs
19. Model generation and testing
20. End

**2. Random Forest Algorithm**

1. Load and Data Preparation
2. Split data into training and testing sets.
3. Bootstrap Sampling
4. Generate multiple random samples from the dataset with replacement (bagging).
5. Train Multiple Decision Trees
6. For each bootstrap sample, train a decision tree:
7. At each node, randomly select a subset of features.
8. Split based on the best feature using Gini index or entropy.
9. Aggregate Predictions
10. For classification: use majority voting among all trees.
11. Evaluation
12. End

**3. K-Nearest Neighbors (KNN)**

1. Load Data and Preprocessing
2. Standardize features to normalize scales (Euclidean distance is scale-sensitive).
3. Split dataset into training and testing sets.
4. Choose Hyperparameter (k)
5. Set the number of neighbors to consider.
6. Optimize (k) using cross-validation if needed.
7. Distance Measurement
8. For a given test point, compute the distance to every training point:
9. $d(x, x') = \sqrt{\underset{j=1}{\overset{n}{}}(x_j - x'_j)^2}$
10. Find Nearest Neighbors
11. Select the (k) closest training samples to the test point.
12. Predict Class
13. Use majority voting among the (k) neighbors to assign a label.
14. Evaluation
15. Model generation
16. End

**4. Naïve Bayes Classifier**

1. Load Data
2. Encode target classes and features if not already numerical.
3. Split into training and testing sets.
4. Estimate Prior Probabilities
5. $P(C_k) = \frac{Number\ of\ samples\ in\ class\ C_k}{Total\ samples}$
6. Estimate Likelihoods
7. For continuous data, assume feature values follow a Gaussian distribution:
8. $P(x_j|C_k) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp - \frac{(x_j - u_{kj})^2}{2\pi\sigma_{kj}^2}$
9. Apply Bayes' Theorem
10. $P(C_k|X) = \frac{P(X|C_k).P(C_k)}{P(X)}$
11. For categorical data, compute frequency ratios per class.
12. Classification
13. Assign test data to the class $C_k$ with the highest posterior probability.
14. Model Evaluation
15. Evaluate performance

**Training of the Models**

The objective of this phase was to train, validate, and compare the performance of four supervised machine learning models: gradient descent BPNN, RF, NB, and K-NN. Each model was trained using the pre-processed dataset to predict customer churn in the MTN Nigeria customer base. For the neural network, training was conducted using the Adam optimizer, with the loss function defined as binary cross-entropy. An 80/20 train-test split validation strategy was employed, and early stopping was implemented to prevent overfitting. The backpropagation algorithm iteratively updated the network weights through gradient descent until the model converged. The RF classifier was trained on the entire feature set, leveraging its inherent ability to model feature interactions and capture non-linear behaviour. By using random feature selection at each node split and aggregating multiple decision trees, the model effectively reduced overfitting while providing interpretable feature importance insights. The Naïve Bayes classifier was trained by calculating the conditional probabilities of each feature given the class label. Due to its assumption of feature independence, NB offered an efficient and computationally lightweight baseline model for churn prediction. The K-NN model utilized a lazy learning approach, where no explicit model was trained. Instead, all pre-processed data instances were stored, and predictions were made by locating the K closest samples based on a distance metric, such as Euclidean distance. A grid search was performed to determine the optimal value of K, and feature normalization ensured that all attributes contributed equally to the distance computation.

**System Flow Diagram**

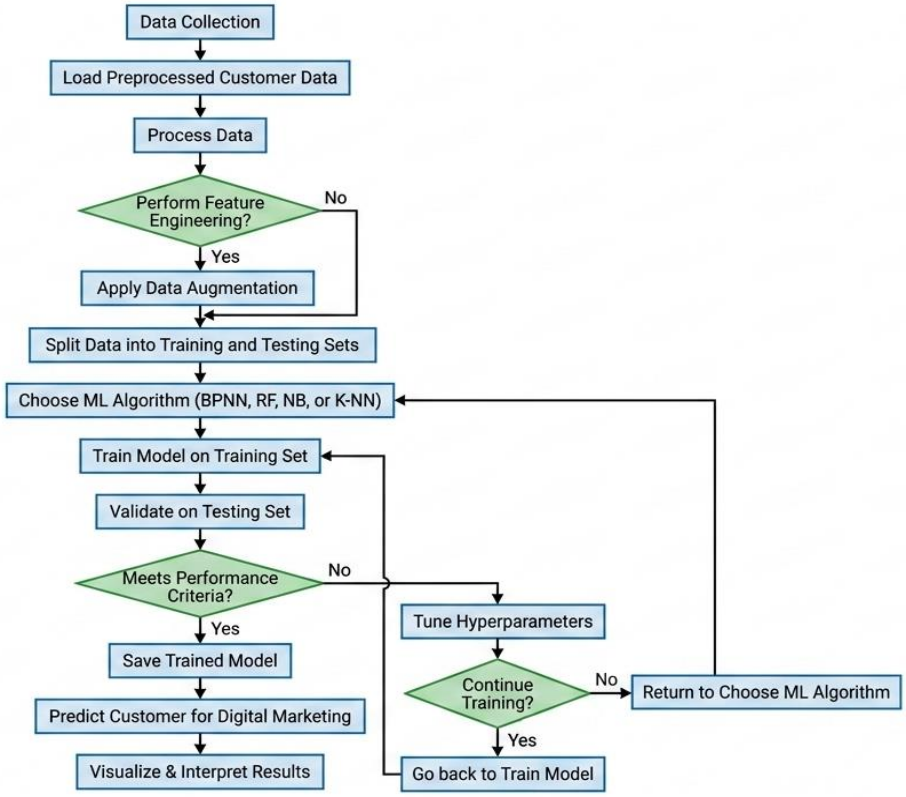


*Figure 1: Flow diagram of the Big Data Analytical Model*

The customer behaviour analysis flow diagram in Figure 1 begins by loading the pre-processed customer data and selecting relevant features for churn prediction. If necessary, feature engineering is applied to enhance data quality, followed by splitting the data into training and testing sets. A suitable machine learning algorithm, such as a BPNN, RF, Naïve Bayes, or KNN, is then chosen for training. The model undergoes iterative training, validation, and hyperparameter tuning until desired performance metrics are achieved. Once a satisfactory model is obtained, it is saved for further use in predicting customer churn probabilities. The final stage involves visualizing and interpreting the results to support data-driven decision-making for customer retention strategies.

**System Implementation**

The implementation phase translated the designed machine learning models and methodologies into a functional system capable of predicting customer churn and analyzing behavioral patterns in the MTN Nigeria dataset. Python was selected as the primary programming language due to its extensive machine learning libraries, such as Scikit-learn, TensorFlow, and Pandas, which facilitated data processing, model training, and evaluation. The system was structured into modular components, including data preprocessing, feature selection, model training, evaluation, and prediction. Each component was executed within a Jupyter Notebook environment to enable seamless experimentation and visualization. The model was integrated as software for big data analysis and customer product preference prediction.

**System Results**

This section presents the training results of the four machine learning algorithms. The results were reported considering accuracy, precision, F1-score, and recall. The results were generated during the training and evaluation process. This performance evaluation was carried out when test data was fed to the respective trained model, and then the evaluation metrics were applied to assess their ability to correctly classify customer product preference to prevent churn. The result of the neural network training was reported in Table 2.

**Table 2: Training result of the back-propagation neural network**

| Test | Precision | recall | F1-score |
|---|---|---|---|
| 0 | 0.58 | 0.80 | 0.67 |
| 1 | 0.88 | 0.92 | 0.90 |
| 2 | 0.48 | 0.63 | 0.55 |
| 3 | 0.00 | 0.00 | 0.00 |
| 4 | 0.22 | 0.42 | 0.29 |
| 5 | 0.43 | 0.57 | 0.49 |
| 6 | 0.78 | 0.88 | 0.83 |
| 7 | 0.98 | 0.90 | 0.94 |
| Macro average | 0.89 | 0.97 | 0.93 |

**Table 3: Training result of the Random Forest**

| Test | Precision | recall | F1-score |
|---|---|---|---|
| 0 | 0.64 | 0.83 | 0.72 |
| 1 | 0.89 | 0.92 | 0.91 |
| 2 | 0.58 | 0.67 | 0.62 |
| 3 | 0.00 | 0.00 | 0.00 |
| 4 | 0.08 | 0.08 | 0.08 |
| 5 | 0.48 | 0.62 | 0.54 |
| 6 | 0.84 | 0.91 | 0.87 |
| 7 | 0.98 | 0.92 | 0.95 |
| Macro average | 0.89 | 1.00 | 0.94 |

**Table 4: Training result of the K-Nearest Neighbor**

| Test | Precision | recall | F1-score |
|---|---|---|---|
| 0 | 0.40 | 0.57 | 0.47 |
| 1 | 0.76 | 0.81 | 0.78 |
| 2 | 0.31 | 0.57 | 0.40 |
| 3 | 0.00 | 0.00 | 0.00 |
| 4 | 0.16 | 0.42 | 0.23 |
| 5 | 0.29 | 0.51 | 0.37 |
| 6 | 0.20 | 0.49 | 0.28 |
| 7 | 0.97 | 0.77 | 0.86 |
| Macro average | 0.89 | 0.99 | 0.93 |

**Table 5: Training result of the Naïve Bayes**

| Test | Precision | recall | F1-score |
|---|---|---|---|
| 0 | 0.39 | 0.83 | 0.53 |
| 1 | 0.57 | 0.82 | 0.68 |
| 2 | 0.24 | 0.38 | 0.29 |
| 3 | 0.01 | 0.75 | 0.02 |

| | | | |
|---|---|---|---|
| 4 | 0.12 | 0.75 | 0.21 |
| 5 | 0.28 | 0.68 | 0.40 |
| 6 | 0.17 | 0.52 | 0.25 |
| 7 | 1.00 | 0.55 | 0.71 |
| Macro average | 0.89 | 0.96 | 0.92 |

Tables 2 to 5 present the individual training results of the four ML algorithms for digital marketing. In Table 6, a summary of the results was presented for discussion.

**Table 6: Result of experimental training performance**

| Model | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Back-propagation neural network | 0.89 | 0.97 | 0.93 | 0.86 |
| Random forest | 0.89 | 1.00 | 0.94 | 0.89 |
| K-NN | 0.89 | 0.99 | 0.93 | 0.88 |
| NB | 0.89 | 0.96 | 0.92 | 0.85 |

Table 6 presents the experimental performance of four models: BPNN, RF, KNN, and Naïve Bayes used for customer product prediction. The BPNN demonstrated strong learning capability with a precision of 0.89, recall of 0.97, F1-score of 0.93, and an accuracy of 0.86. These results show that the neural network effectively captures nonlinear customer behaviour patterns and correctly identifies a large proportion of customer classes. However, despite its strong recall, the overall accuracy is slightly lower than that of the RF and K-NN models, indicating that it still misclassifies a notable number of customers.

**Result of the big data analytical framework for digital marketing**

This section presents the big data analytical framework designed for customer product preference prediction and digital marketing. Figure 2 presents the login page. This page ensures that only registered users are able to gain access to the software to carry out big data analysis. The users are of different categories, such as management, analyst, and marketing. The reason was to ensure access control to the software and improve security. Figure 3 presents the dashboard, while Figure 4 presents the product target module.

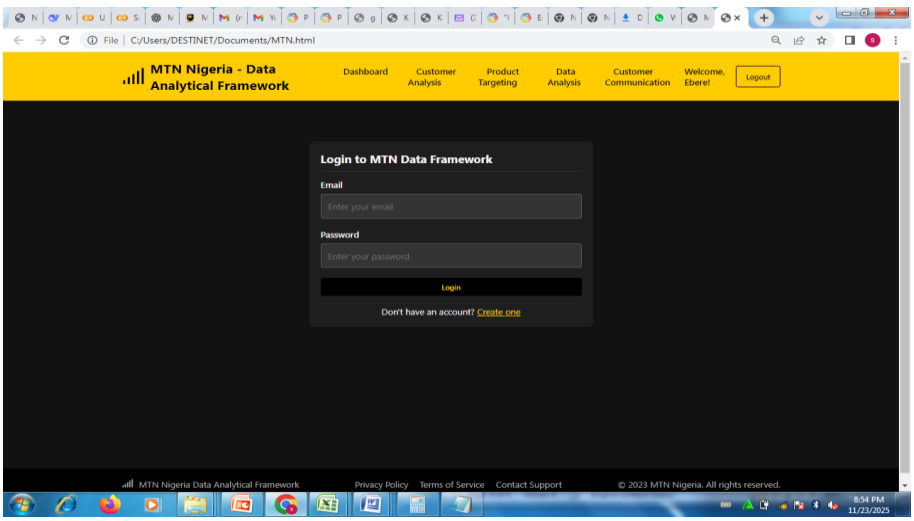


*Figure 2: Login page of the big data framework*

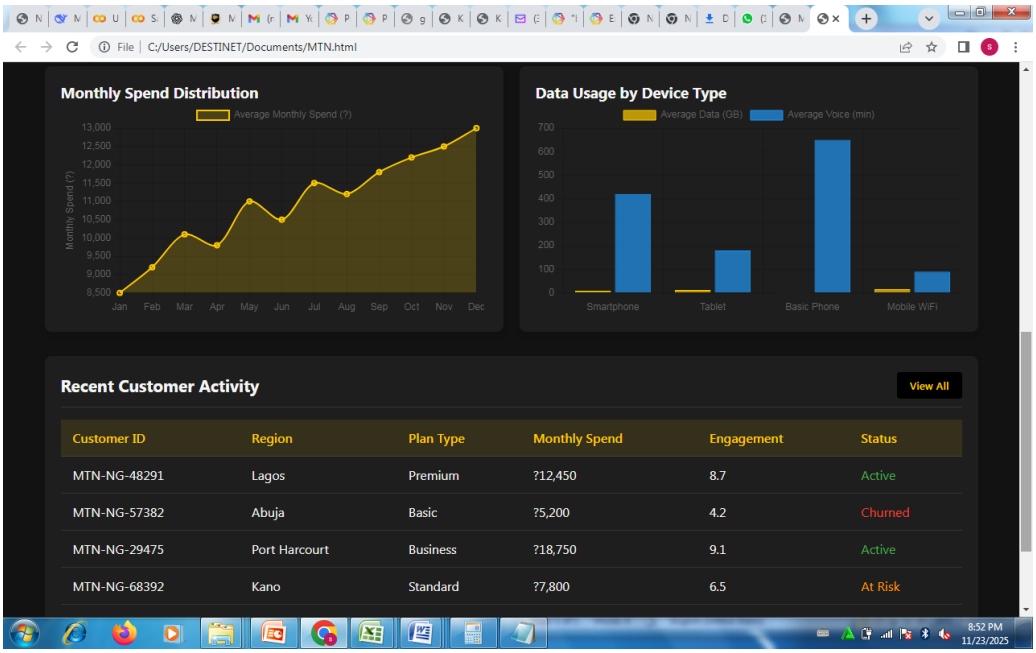


*Figure 3: the system dashboard*

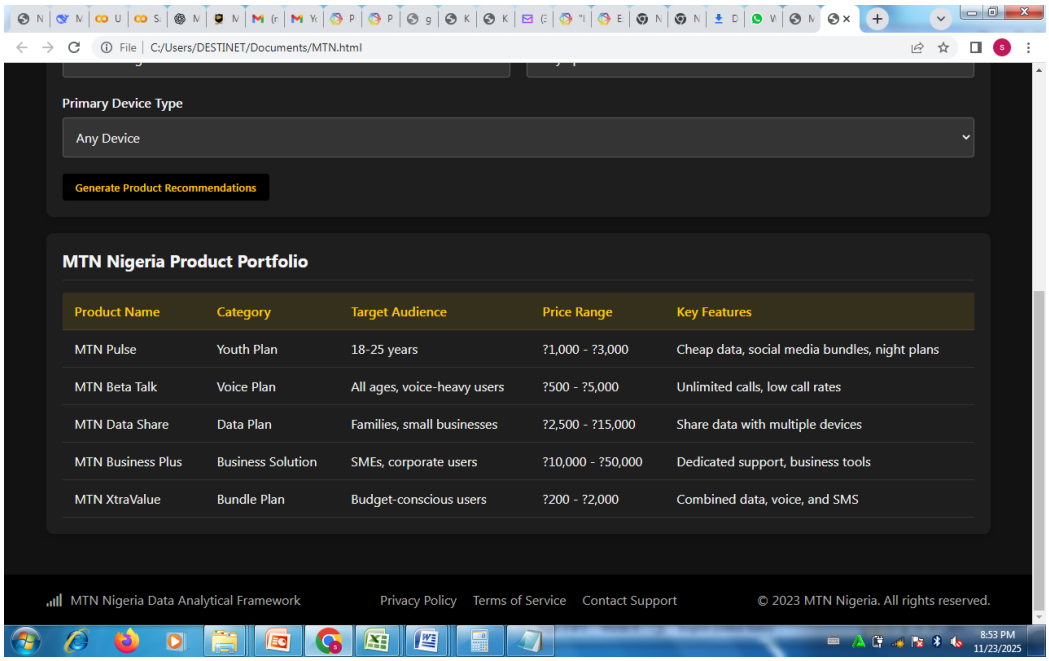


*Figure 4: Product target modules*

Dashboard Module in Figure 3 provides a comprehensive overview with real-time KPIs, interactive charts, and customer activity tracking to monitor key business metrics and performance indicators. Product Targeting Module in Figure 4 uses the trained and selected RF model to recommend optimal MTN products based on customer profiles, usage patterns, and spending behaviors to maximize uptake and revenue potential. Figure 5 presents the customer analysis module, while Figure 6 presents the big data analysis module.

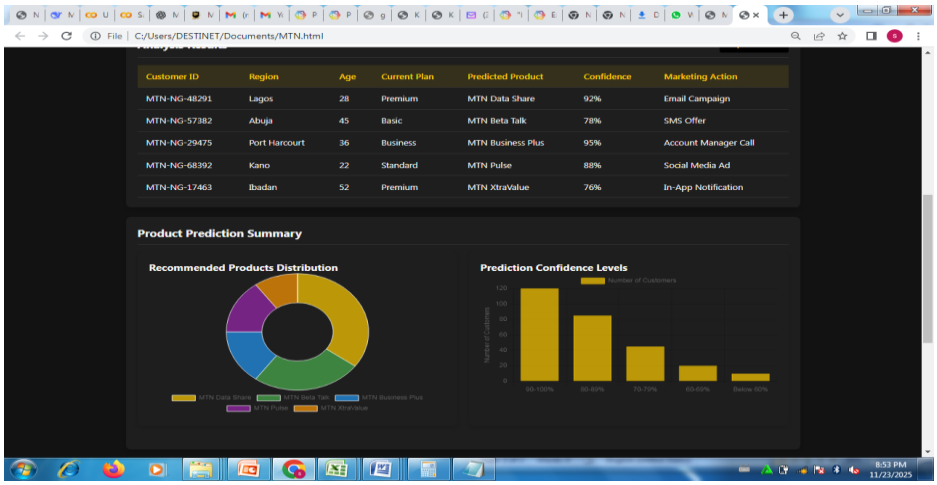


*Figure 5: customer analysis module*

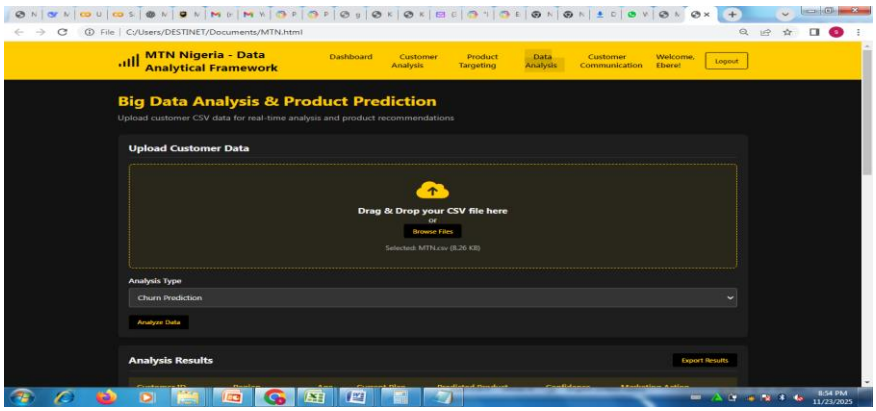


*Figure 6: Big data analysis section*

The customer analysis module in Figure 4.24 segments users by behavior patterns, usage trends, and demographic data while predicting churn risks and identifying high-value customer segments for targeted interventions. Data Analysis Module in Figure 4.25 processes uploaded CSV files for predictive analytics, generating product recommendations and digital marketing strategies through simulated machine learning algorithms. Customer Communication Module in Figure 7 enables targeted email campaigns with filtering capabilities, template management, and bulk messaging for personalized customer engagement and retention efforts.

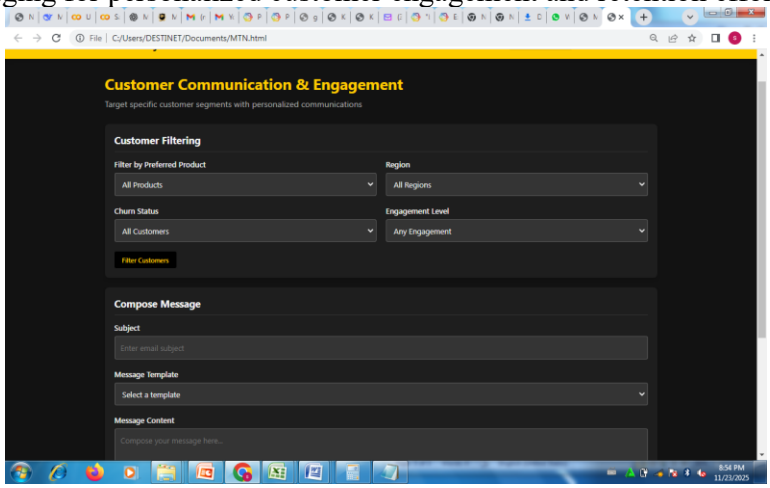


*Figure 7: Customer communication module*