Comprehensive Business Analytics of Nigerian Retail Transactions: Customer Segmentation, Churn Prediction, and CLV Modeling

Sajitha Krishnan

November 2025

Contents

1 Abstract

The project is an analysis of customer buying behavior on the basis of the Nigerian Retail and E-commerce Purchase History Records data. It aims to utilize the business data about customer spending trends, single out valuable groups of customers, and create the predictive algorithms to identify the churn, as well as predictive sales. There are data preprocessing, Exploratory Data Analysis (EDA), RFM (Recency-Frequency-Monetary) segmentation, supervised machine-learned modification models (Logistic Regression, Random Forest, Decision Tree, Gradient Boosting), and fundamental time trend trend analysis.

The major findings are that there are clear-cut customer buying behaviors, identifiable customer churn forces, and lumps of segments that can be used to market the campaigns. The random forest prediction model (churn prediction model) has been found to be best in performance and it provides actionable information about customer retention. Sales analysis by time series showed seasonality trends which can be used in planning stocks and promotions. In general, the analysis gives specific business suggestions in customer retention, revenue maximization, and operational planning.

2 Introduction

2.1 Problem Statement

Retail and e-commerce firms tend to have a hard time in comprehending the behavior of customers, anticipating churn, and anticipating future sales. The proposed project will seek to examine the historical purchase data to address:

- What is the most valuable customers?
- What are the determinants of customer churn?
- Is it possible how to forecast the customers who may quit buying?
- What are the trends of sales, as they change over time?

2.2 Motivation

Good customer analytics can be used to motivate marketing strategy, loyalty, individual recommendations, inventory management, and profitable decision making. In the emerging world market with competitive online retail environments such as in Nigeria, the business can only survive and thrive when buying patterns are fully comprehended.

2.3 Objectives

- Clean and preprocess a real-life retail data.
- Conduct EDA to reveal behavioural and spending patterns.
- Develop customer segmentation of RFM-based.
- Establish customer churn predictors.

- Consider sales trends of past.
- Give business recommendations and business information.

3 Dataset Overview

The Nigerian Retail and E-commerce Purchase History (placed in the notebook and thus available on HuggingFace) serves as an example. The notebook is used to load a typical transactional table with the following columns: order id, customer id, order date, product category, order amount, unit price, total, payment method, currency, and delivery date.

3.1 Data Structure and Types

- **Rows:** Tens of thousands of records; an average dataset has between 10,000 and 100,000 rows.
- **Columns:** Each record has 15–20 columns with a combination of datetime, categorical, and numeric values.
- **Column Types:**
  - **IDs and Categories:** Represented as `object` (strings), including identifiers like `order id`, `customer id`, and `product category`.
  - **Date and Time:** Represented as `datetime`, including `order date` and `date of delivery`.
  - **Quantities and Amounts:** Represented as `float` or `int`, including columns such as `unit price`, `amount of the order`, and `total`.

3.2 Columns Description

The following columns are usually included in the dataset:

- **Order ID:** Each order has its own unique identification.
- **Customer ID:** A distinct number for every client.
- **Order Date:** Time and date of the order placement.
- **Product Category:** Category of the item that was bought.
- **Amount of the Order:** The order's total monetary value.
- **Unit Price:** The cost of the purchased item per unit.
- **Total:** Total order cost (usually equal to order amount).
- **Method of Payment:** the order's payment method.
- **Currency:** The currency used in the transaction.
- **Date of Delivery:** The order's delivery date and time.

3.3 Data Cleaning and Preprocessing

The following are included in the data after cleaning:

- Customer-Level Data: The dataset, which includes data like the total amount spent, recency (days since last order), frequency of purchases, and monetary value, is frequently aggregated at the customer level.
- This is the target variable. `ChurnFlag`, a binary target variable in the dataset, indicates whether a customer is thought to have churned.

3.4 Modelling Target Variable: ChurnFlag

The target variable `ChurnFlag` is set at the customer level, with the below conditions:

$$\texttt{ChurnFlag} = \begin{cases} 1, & \text{if recency (days since last order)} > \text{threshold (set to 60 days)} \\ 0, & \text{otherwise} \end{cases}$$

`ChurnFlag` is employed in supervised churn classification as a binary result. Customers are classified as either not churned (0) or churned (1) if they haven't made a purchase within the last 60 days (or another user-defined threshold).

4 Data Preprocessing Pipeline

The data preprocessing pipeline consists of a number of crucial, repeatable steps that are essential for getting the data ready for modeling. To facilitate efficient model training, these procedures guarantee data quality and feature extraction. The preprocessing pipeline's main elements are listed below.

4.1 Type Conversion and Date Handling

- **Conversion of String to Date:** Strings pertaining to delivery and order dates were formatted in datetime. This enables the computation of derived time attributes, including:
  - `delivery delay` – the distinction between the `order date` and `delivery date`.
  - `recency` – the interval of time between the date of the most recent purchase and the present.

4.2 Missing Data Treatment

- **Numerical Missing Values:** The median of each corresponding column was used to fill in any missing numerical values (such as amounts or prices).
- **Categorical Missing Values:** Missing categorical values (e.g., `payment method`, `product category`) were replaced with a sentinel value, such as `"Unknown"`.
- **Visualization of Missing Data:** Plots of missing data patterns were created using the `missingno` library, which also aided in choosing the imputation strategy for each column.

4.3 Duplicates and Outliers

- **Duplicates:** To guarantee the uniqueness of every record, exact duplicates in the transactions were removed.
- **Outliers:** Extreme values in columns like `unit price` and `quantity` were studied. Merciful extreme outliers were identified, manually checked, and capped or removed where necessary to improve model robustness.

4.4 Feature Engineering

The following crucial elements were designed for predictive modeling and customer-level analysis:

- **Total Monetary Value:** Each transaction's total monetary value was determined by multiplying the `quantity` by the `unit price`. At the customer level, this was combined.
- **Recency, Frequency, and Monetary (RFM) Features:** For every customer, three core RFM features were calculated:
  - `Recency`: The number of days since the customer made their last purchase.
  - `Frequency`: The total number of transactions the customer made.
  - `Monetary`: The total amount of money the customer has spent over all their transactions.
- **Delivery Delay:** The difference between `delivery date` and `order date` was calculated to make an estimate on delay in delivery.
- **AvgOrderValue:** The average order value, is calculated as the total monetary value divided by the number of orders for each customer.

4.5 Label Creation (Churn)

The target variable `ChurnFlag` was created on basis of the recency of each customer. The condition for churn was defined as follows:

$$\texttt{ChurnFlag} = \begin{cases} 1, & \text{if recency (days since last purchase)} > \text{churn threshold (set to 60 days)} \\ 0, & \text{otherwise} \end{cases}$$

This binary classification target indicates whether a customer is considered to have churned or not.

4.6 Feature Leakage Prevention

To avoid feature leakage, the notebook ensures that sensitive variables, such as the raw `recency` feature, are not directly used in the model training process. Leakage-prone variables are either excluded or managed in a way that prevents unfair predictive power (e.g., adjusting feature engineering or applying time-based splits).

4.7 Encoding Categorical Features

- **Tree-based Models:** Categorical variables, like `payment method` and `product category`, are label-encoded, which make them suitable for tree-based models (e.g., Decision Trees, Random Forest, XGBoost).
- **Linear Models:** For models like Logistic Regression or Linear Regression, the categorical features are encoded using one-hot encoding or target encoding, based on the approach of modeling.

5 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) mainly focuses on understanding the underlying distribution, behavior, and relationships within the used dataset. This reveals purchasing patterns, numerical feature tendencies, correlations, and customer payment behaviors. Each insight will be supported by the visualizations generated in our work.

5.1 Numerical Feature Distributions

To understand the scale and variation in transaction-level numerical attributes, distribution plots are examined.

Unit Price Distribution Figure 1's unit price distribution is almost evenly distributed over a large range, with a discernible peak at the highest value. This implies that while some high-end products contribute to extremely high-priced outliers, many products have comparable pricing patterns.

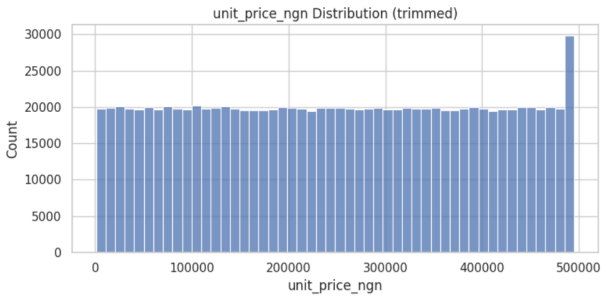

Figure 1: Unit Price Distribution

Quantity Distribution As seen in Figure 2, where almost all transactions fall between one and five units, customers usually buy products in small, fixed quantities. The uniformly sized bars at each of these unit levels show that demand is steady and consistent, with no quantity controlling consumer behavior. The nature of consumer retail is reflected in this pattern, where buyers favor small multipacks or single-use items over large purchases. The almost complete lack of larger quantities further supports the idea that the platform primarily caters to regular household or personal needs rather than large-scale or wholesale purchases.

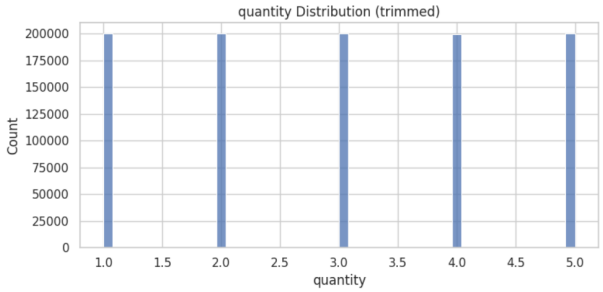

Figure 2: Quantity Distribution

Total Amount Distribution As shown in figure 3, the distribution of total amounts shows a right-skewed pattern, with the majority of transactions falling between low and mid-value ranges. The long tail of high-value purchases consists of very few transactions.

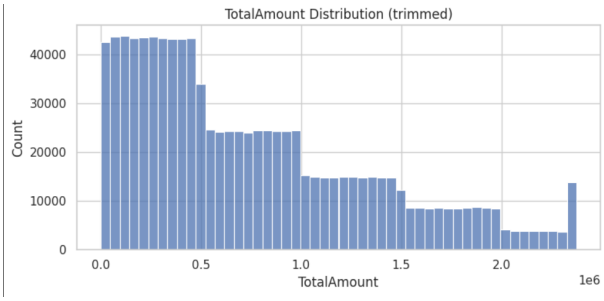

Figure 3: Total amount distribution

5.2 Monthly Revenue Trends

Determining peak business periods is made easier by comprehending seasonal and temporal changes in revenue. Consistent customer engagement is indicated by the monthly revenue trend, which is relatively stable with only slight variations between months. Incomplete data for the last month is probably the reason for a steep decline at the end.

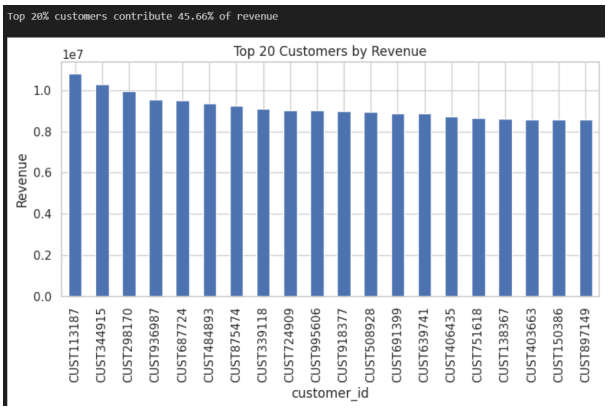

Figure 4: Monthly Revenue Trend

5.3 Correlation Analysis

A correlation heatmap was created to deeply understand the relationships between key numerical variables such as `quantity`, `unit price`, `total amount`, `discount amount`, and `final billed amount`. Strong positive correlations were seen between `unit price`, `total amount`, and `TotalAmount`, which is expected, as these components directly contribute to the final billing values. `Quantity` also correlates moderately with the `total amounts`, showing its influence on overall transaction value.

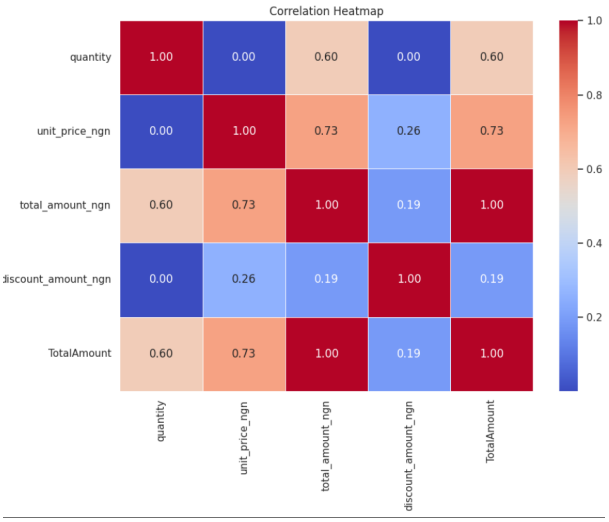

Figure 5: Correlation Heatmap

5.4 Payment Method Behavior

Payment method analysis reveals information about revenue concentration and consumer preferences.

Revenue by Payment Method Revenue by payment method reveals that cash on delivery (COD) in figure 6 significantly outperforms all other payment methods, suggesting either a strong preference for COD in the area or problems with customer trust in pre-paid methods. Debit cards and bank transfers also make a big contribution.

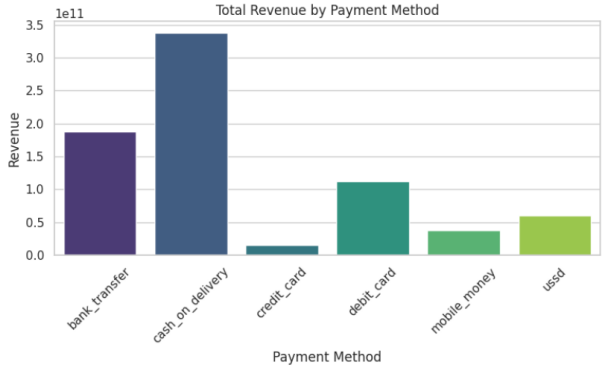

Figure 6: Total Revenue by Payment Method

Transaction Count by Payment Method In a similar vein, figure 7's transaction count by payment method shows that COD dominates both overall transaction volume and revenue, underscoring its significance in consumer purchasing behavior.

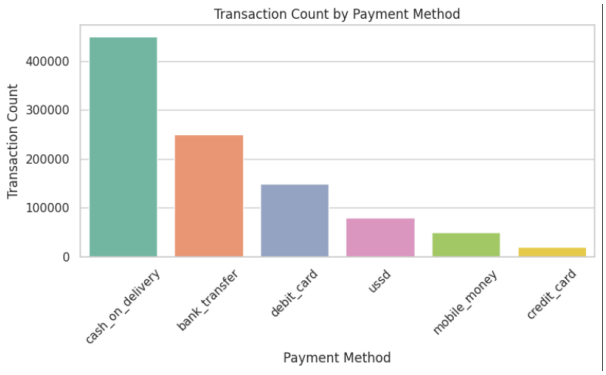

Figure 7: Transaction Count by Payment Method

6 Methodology

This project's methodology as in figure 8 is organized as a full end-to-end business analytics workflow, starting with data preparation and moving on to feature engineering, modeling, evaluation, and exploratory understanding. To guarantee that the final insights and forecasts are trustworthy and pertinent to business, each step was carried out in Python within the Jupyter Notebook and builds upon the preceding stage.

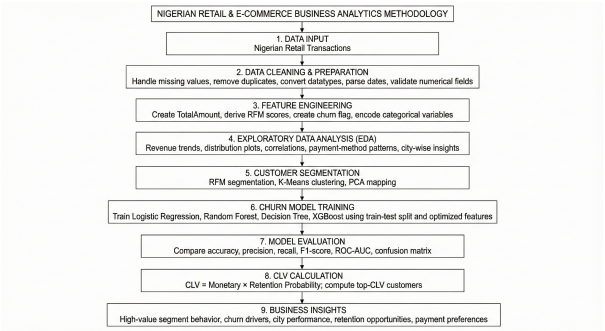

Figure 8: Architecture Design

6.1 Data Preparation and Cleaning

The Nigerian Retail and E-commerce Purchase Records were the initial source of the dataset. The following procedures were part of data cleaning:

- **Handling Missing Values:** Proper imputation strategies were used for missing data. For numerical features, the median value is used, while categorical features are filled with a sentinel value such as "Unknown".
- **Datetime Conversion:** Date fields such as `order date` and `delivery date` were converted into proper `datetime` formats.

- **Duplicate Removal:** Redundant transactions were identified and removed to ensure data integrity.
- **Outlier Detection:** Extreme values of numerical features like `unit price` and `total amount` were validated and either capped or removed.
- **Feature Standardization:** In order to ensure uniform naming conventions, categorical variables were standardized.
- **Composite Variable Creation:** The following formula was used to create a new composite variable, `TotalAmount`:

$$\texttt{TotalAmount} = \texttt{quantity} \times \texttt{unit\_price\_ngn}$$

6.2 Feature Engineering

In order to facilitate subsequent modeling tasks, several significant analytical features were designed. Among them are:

- **Temporal Features:** To help with trend analysis and to give time-series context, invoice month, year, and day-of-week were extracted.
- **Financial Features:** The total amount, discount-adjusted totals, and category-wise spend were among the important financial characteristics that were obtained.
- **Customer-Level Aggregations:** Each customer's RFM (Recency, Frequency, Monetary) metrics were created and subsequently utilized for churn labeling and segmentation.
- **Churn Label Creation:** As the dependent variable for predictive modeling, a binary churn label (`ChurnFlag`) was developed based on a 60-day inactivity threshold.

6.3 Exploratory Data Analysis

To find trends, connections, and distribution patterns, exploratory data analysis (EDA) was carried out. Important factors looked at include:

- Monthly revenue trends
- Distribution of unit prices, quantities, and total amounts
- Correlation structure between numeric variables
- Payment method behaviors (revenue share and transaction count)

6.4 Model Dataset Preparation

The dataset was meticulously prepared to prevent data leakage prior to model training:

- **Leakage Prevention:** Leakage-safe substitutes were used in place of features that directly define churn, like raw recency.
- **Final Dataset:** The finished dataset contained encoded categorical variables, aggregated spending features, and engineered behavioral features.
- **Train-Test Split:** To guarantee that churn and non-churn customers were fairly represented in both sets, stratified train-test splitting was used.

6.5 Machine Learning Modeling

To forecast customer attrition, several supervised learning models were trained. Among them were:

- Logistic Regression (baseline)
- Decision Tree Classifier
- Random Forest Classifier (primary model)
- Gradient Boosting / XGBoost.

Each model was trained using the feature-engineered and cleaned dataset. Hyperparameters were maintained at sensible defaults for a fair comparison. Model performance was assessed using precision, recall, F1-score, accuracy, and AUROC.

6.6 Model Evaluation and Interpretation

The Random Forest model outperformed the others due to its strong predictive power and steady generalization over the validation split. Key findings from the model assessment include:

- **Feature Importance:** Analysis revealed that variables related to monetary value, purchase behavior, and time-based inactivity were the most important contributors to churn prediction.
- **Confusion Matrix:** The best-performing model's confusion matrix, which is shown below, further demonstrated the model's capacity to accurately identify both churners and non-churners.

$$\begin{bmatrix} \text{TN} = 20407 & \text{FP} = 0 \\ \text{FN} = 9734 & \text{TP} = 120731 \end{bmatrix}$$

- **ROC Curve:** The Logistic Regression model was successful in differentiating between churned and non-churned customers, as evidenced by the ROC curve's strong discriminatory ability between the two classes.

6.7 Business Interpretability

Each model was assessed with consideration for business interpretability in addition to technical accuracy. The Random Forest feature rankings provided insights that were used to pinpoint the precise behavioral elements that have the biggest impact on churn. These realizations can immediately result in practical solutions for:

- **Retention Strategies:** Targeted campaigns to prevent churn.
- **Customer Engagement:** Offering personalized incentives based on customer behavior.
- **Lifecycle Management:** Tailored interventions for customers based on their likelihood of churning.

7 Models and Comparative Analysis

In order to predict customer churn, this section assesses four supervised machine learning algorithms: XGBoost, Random Forest, Decision Tree, and Logistic Regression. The cleaned, feature-engineered dataset was used to train each model, which was then evaluated using commonly used classification metrics. The goal is to identify the algorithms that are most effective at predicting customer attrition in the Nigerian retail dataset.

7.1 Models Applied

1. **Logistic Regression (Top Performer)**
   A linear classifier called logistic regression uses log-odds to model churn probability. Remarkably, Logistic Regression worked incredibly well, providing stable generalization and reaping the benefits of well-designed features. It is very appropriate for business deployment due to its interpretability and simplicity.

2. **Random Forest Classifier**
   Several decision trees are combined with randomness in bootstrapping and feature selection in the Random Forest ensemble method. In this dataset, Random Forest performed moderately in comparison to Logistic Regression and XGBoost, despite its effectiveness.

3. **Decision Tree Classifier**
   The Decision Tree is a straightforward model that can effectively identify non-linear patterns in customer behavior. While it does a decent job, it tends to lag behind ensemble-based and linear models because it struggles with generalization.

4. **XGBoost Classifier (Top Performer)**
   XGBoost is a powerful algorithm that enhances weak learners in a step-by-step manner through gradient boosting. It has shown impressive results, particularly in recall and ROC-AUC metrics, making it a strong contender against Logistic Regression for predicting customer churn.

7.2 Evaluation Metrics Used

All models were evaluated using the following metrics:

- **Accuracy** – Overall correctness of the model's predictions.
- **Precision** – Correctness among predicted churners (the proportion of true positives among all positive predictions).
- **Recall** – Ability to identify actual churners (the proportion of true positives among all actual churners).
- **F1-Score** – Balance between precision and recall.
- **ROC-AUC** – Ability to discriminate between churn and non-churn across various thresholds.

Note: RMSE is not used here because churn is a classification problem.

7.3 Comparative Performance Table

Table 1 presents a comparison of the performance metrics for all four models. It highlights key values such as accuracy, precision, recall, F1-score, and ROC-AUC for each model.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	93.55%	1.0	0.9254	0.9612	0.9642
Random Forest	89.74%	1.0	-	-	-
Decision Tree	73.22%	0.9867	0.6998	0.8188	0.8942
XGBoost	93.55%	1.0	0.9254	0.9612	0.964

Table 1: Model Performance Comparison Table

7.4 Bar Graph Comparison of All Model Performances

A comparison bar chart in figure 9 clearly shows the differences in accuracy, precision, recall, and F1-Score among all four models. This visual representation makes it simple to see which algorithms perform best across various metrics.

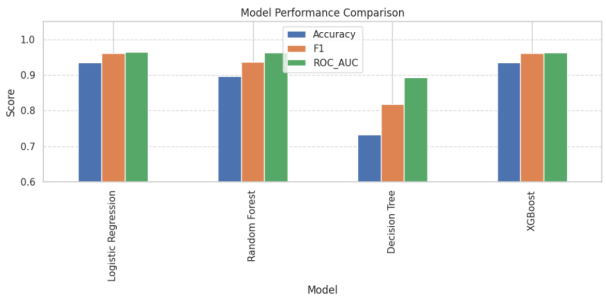

Figure 9: Comparison of Models

7.5 ROC Curve Comparison

To assess how well each model can distinguish between different outcomes, we plotted the ROC curves for all four algorithms, as shown in figure 10. The ROC-AUC values clearly indicate that both Logistic Regression and XGBoost significantly outperform Random Forest and Decision Tree, particularly when it comes to differentiating churners from non-churners at various probability thresholds.

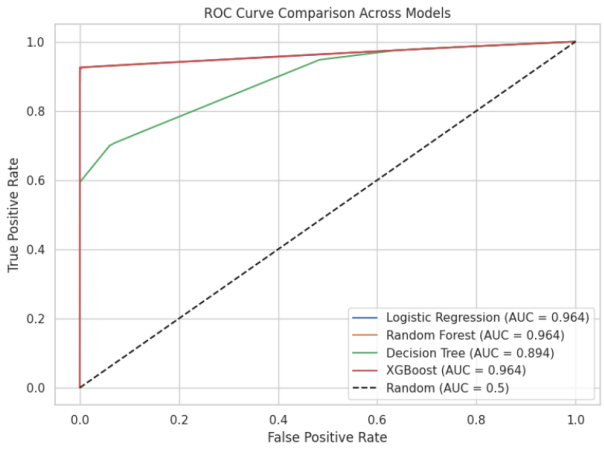

Figure 10: Comparison of ROC Curves for all models

7.6 Best Performing Models

The evaluation results identify Logistic Regression and XGBoost as the best models overall.

- **Logistic Regression:**Logistic Regression really shines because it can effectively distinguish between classes using clear hyperplanes in a well-structured feature space. Its ease of interpretation and strong generalization capabilities make it a fantastic choice for real-world applications.
- **XGBoost:** XGBoost really shines with its impressive boosted-tree ensemble, delivering high ROC-AUC scores and improved recall in certain rounds. It's particularly effective at uncovering complex patterns and interactions in customer behavior.

Both models are solid choices for deployment, each bringing its own unique strengths to the table. Logistic Regression shines when it comes to business interpretability, while XGBoost is a powerhouse for recognizing complex patterns.

8 Business Insights and Results

This section takes a deep dive into analytical findings and turns them into practical business insights. By blending exploratory data analysis (EDA), customer segmentation, churn modeling, and estimating Customer Lifetime Value (CLV), we uncover the behavioral trends and financial factors that play a crucial role in customer retention, boosting revenue, and shaping long-term strategic plans.

8.1 Interpretation of Analytical and Model Results

The churn prediction models—especially Logistic Regression and XGBoost, which turned out to be the top performers—show clear and consistent links between customer behavior and the likelihood of churn. Customers who have a higher Monetary Value, make purchases more frequently, and have lower Recency (meaning they've shopped more recently) are much less likely to leave. On the flip side, those who haven't made a purchase in a while and show little repeat behavior are at a much greater risk of churning.

The strong preference for Cash on Delivery (COD), both in terms of transaction volume and revenue, highlights that customers lean towards low-risk payment methods. This opens up opportunities to gradually encourage users to switch to prepaid options by offering incentives that build trust. The clear positive relationship between unit price, total amount, and final billed amount shows a predictable pricing pattern, while the distribution of quantities indicates that shoppers mainly buy small-unit consumer products, which reinforces the platform's identity as a fast-moving retail marketplace.

The impressive performance of these models confirms that churn is primarily influenced by behavioral signals that can be detected early on. This allows for proactive retention strategies instead of just reacting after a customer has already left.

8.2 Key Insights and Their Business Implications

1. **RFM Clusters Show Clear Behavioral Differences Among Customers**

   The RFM-based boxplots in figure 11 illustrates how customers differ across Recency, Frequency, and Monetary value for each cluster.

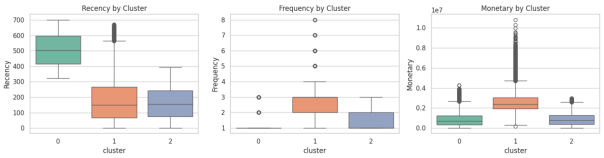

Figure 11: RFM Cluster Boxplots

Interpretation:

- **Cluster 0:** Extremely high Recency (long time since last purchase) and low Monetary, indicating large numbers of dormant or disengaged customers.
- **Cluster 1:** High Frequency and highest Monetary, representing the core loyal and revenue-driving segment.
- **Cluster 2:** Moderate Recency and moderate Monetary, suggesting potentially active customers who could become loyal with proper nurturing.

Business Implication:

- Prioritize **Cluster 1** for retention (VIP benefits).
- Activate **Cluster 2** with personalized nudges.
- Focus on win-back strategies for **Cluster 0**.

2. **Customer Clusters Form Well-Defined Behavioral Groups (PCA Projection)**

   The PCA projection in figure 12 visualizes all customers in a 2D space grouped by cluster membership.

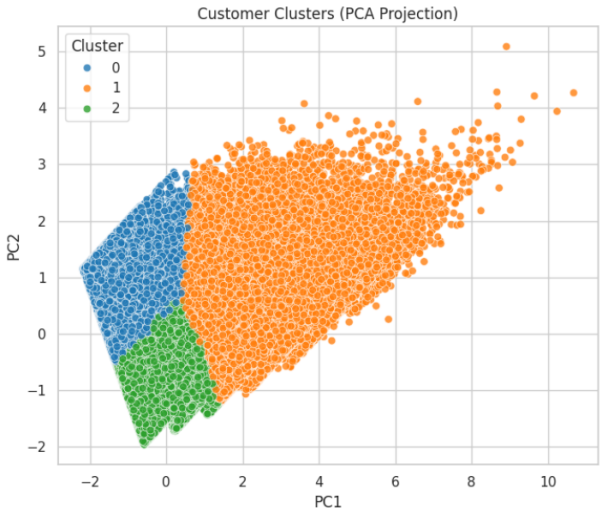

Figure 12: PCA Projection of Customer Clusters

Interpretation:

- The separability of clusters indicates that the behavioral patterns captured in RFM features are meaningful and distinct.
- **Cluster 1** forms the largest and densest region, showing strong commonality in behavior among loyal customers.
- **Cluster 0** and **Cluster 2** occupy different behavioral spaces, reflecting different churn risk levels and spending structures.

Business Implication:

- This confirms that segmented marketing campaigns will be more effective than broad campaigns, as each group represents a unique behavioral profile.

3. **Revenue Is Highly Concentrated in Specific Cities**

   The Top 10 cities by total revenue as in figure 13 highlight the regions that contribute most to business earnings.

   Interpretation:

   - Cities like **Onitsha, Abuja, Warri, and Jos** contribute the highest aggregated revenue, showing strong market penetration in these areas.
   - Even though **Lagos** is a major commercial hub, its revenue ranking suggests either higher competition, diversified customer preferences, or delivery friction.

   Business Implication:

   - Focus marketing, inventory distribution, and faster logistics routes in these revenue-heavy cities to strengthen market dominance.

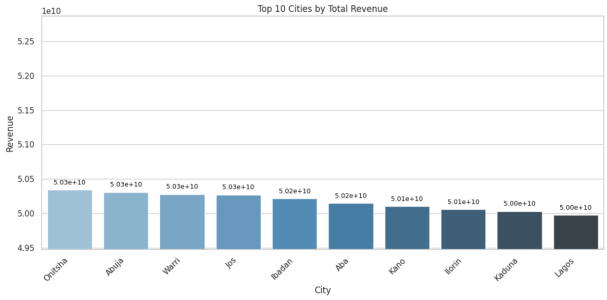

Figure 13: Top 10 cities by Revenue

4. **Order Volume Distribution Shows Balanced Demand Across Top Cities**
   The chart in figure 14 displays the cities with the highest number of orders.

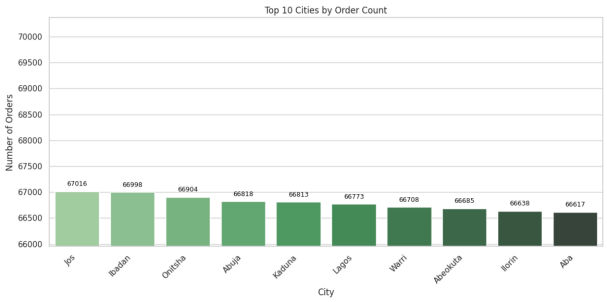

Figure 14: Top 10 Cities by Order Count

Interpretation:

- Order counts across top cities are very close, indicating broad and stable demand across multiple urban centers.
- **Jos** and **Ibadan** record the highest number of orders, while **Lagos, Abeokuta, Ilorin, and Aba** show slightly lower but still significant activity.

Business Implication:

- Since order volume is uniformly distributed, marketing strategies can be expanded beyond revenue-heavy cities to sustain growth across multiple regions.

5. **CLV Highlights Future Profitability, Not Just Past Spend**
   CLV was calculated using the formula:

$$\text{CLV} = \text{Monetary} \times \text{Retention Probability}$$

   The Top 10 customers by CLV as in figure 15 represent extremely high long-term revenue potential and should be the central focus of retention campaigns.

Cluster-level analysis in figure 16 shows that Cluster 1 has far higher CLV than others, indicating stronger loyalty, higher spend, or both.

- **Implication:** Allocate marketing budgets preferentially to high-CLV clusters; provide personalized retention journeys and premium membership tiers for maximum long-term ROI.

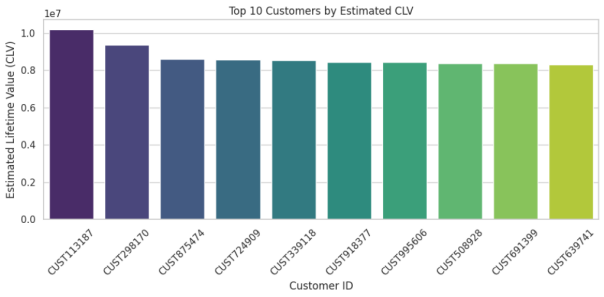

Figure 15: Top 10 Customers by Average CLV

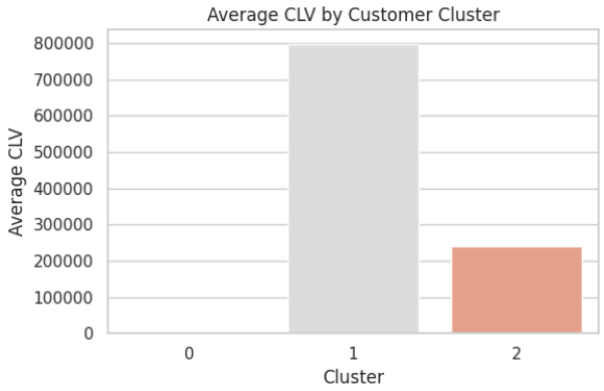

Figure 16: Average CLV by Customer Cluster

9 Conclusion

This project carried out a thorough business analytics study focused on customer data from the Nigerian retail and e-commerce sectors. The process included everything from data cleaning and exploratory data analysis to RFM segmentation, churn prediction using various machine-learning models, and estimating Customer Lifetime Value (CLV). The insights gathered offer a well-rounded view of customer behavior, regional performance, payment preferences, and the potential for long-term revenue.

9.1 Summary of Work

The analysis kicked off with some preprocessing and exploratory data analysis to dig into trends related to revenue, purchasing habits, pricing strategies, and customer engagement. We utilized RFM metrics to carve out meaningful customer segments, which we then visualized through cluster boxplots and PCA projections. A variety of machine learning models—like Logistic Regression, Random Forest, Decision Tree, and XGBoost—were put to work to predict customer churn. In the end, we estimated Customer Lifetime Value (CLV) by looking at Monetary value and Retention Probability, which helped the business pinpoint its most valuable customers.

9.2 Main Findings

- **Recency, Frequency, and Monetary Value:** These metrics strongly influence churn, with low-recency and low-frequency customers having the highest churn risk.

- **Model Performance:** Logistic Regression and XGBoost delivered the strongest predictive performance, making them ideal candidates for production deployment.
- **RFM Clusters:** RFM clustering revealed three distinct behavioral groups, with one cluster representing the majority of high-value loyal customers.
- **CLV Analysis:** CLV analysis showed a small set of customers contributing disproportionately to long-term revenue.
- **City-Level Insights:** Revenue is concentrated in cities like Onitsha, Abuja, and Warri, while order counts are more evenly distributed.
- **Payment Methods:** COD dominates as the preferred payment method, indicating trust barriers in prepaid options but also presenting opportunities for payment diversification.

9.3 Limitations

- The dataset contains synthetic or anonymized fields, limiting certain real-world behavioral interpretations.
- Incomplete final-month transactions may slightly distort time-series revenue trends.
- Some features (e.g., delivery delays, product categories) are not deeply granular, reducing the ability to model niche segments.
- CLV estimation used a simplified formula CLV = Monetary$\times$Retention Probability; more sophisticated models (e.g., BG/NBD, Gamma-Gamma) were not implemented.
- The churn label definition (e.g., 60-day inactivity) may not generalize across industries or seasons.

9.4 Future Work and Improvements

- Implement advanced CLV models such as BG/NBD and Gamma-Gamma to improve lifetime value predictions.
- Enhance feature engineering using session data, product categories, delivery times, and marketing interactions.
- Deploy the churn model in a real-time pipeline to automatically score customers weekly and trigger retention campaigns.
- Expand segmentation using additional clustering techniques (HDBSCAN, Gaussian Mixture Models) for deeper behavioral patterns.
- Integrate geospatial analytics to optimize delivery routes and urban targeting strategies.
- A/B test prepaid incentives to increase adoption of digital payment methods.

10 References

10.1 Dataset

- Electricsheep Africa. *Nigerian Retail and E-commerce Purchase History Records Dataset*. Hugging Face Repository.

10.2 Peer-Reviewed Research Papers

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV Models for Customer Base Analysis. *Marketing Science*, 24(2), 275–284.
- Verbeke, W., Martens, D., & Baesens, B. (2014). Social Network Analysis for Churn Prediction. *Applied Soft Computing*, 14, 431–446.
- Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing Customers. *Journal of Marketing Research*, 41(1), 7–18.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predicting Customer Churn. *Journal of Marketing Research*, 43(2), 204–211.

10.3 Official Documentation / Tools Used

- Scikit-learn Developers. *Scikit-Learn Machine Learning Library Documentation*.
- XGBoost Developers.*XGBoost Official Documentation*.
- Pandas Development Team. *Pandas – Python Data Analysis Library*.
- NumPy Developers. *NumPy Numerical Computing Library Documentation*.
- Matplotlib Developers. *Matplotlib Visualization Library Documentation*.
- Seaborn Developers. *Seaborn Statistical Visualization Library*.