Supermarket Sales Prediction Capstone Project Report

BY

Anas Yunusa Adamu

(FE/24/154693239)

3MTT Fellow,

Cohort 1 DeepTech Program

January 2026

1. Executive Summary

This project aims to optimize inventory management and business strategy for a supermarket chain by developing a predictive machine learning model for sales forecasting. By analyzing historical transaction data, we identified key drivers of sales and built a robust Random Forest regression model. The solution has been deployed as an interactive web application using Streamlit, allowing stakeholders to simulate sales scenarios based on product details, location, and timing. To ensure relevance, the dataset was localized to reflect major Nigerian commercial hubs (Lagos, Abuja, Port Harcourt).

2. Project Objectives

The primary goal of this capstone project is to leverage data science techniques to solve retail challenges. The specific objectives are as follows:

1. **Analyze the Dataset:** Conduct a thorough Exploratory Data Analysis (EDA) to identify patterns, anomalies, and correlations that affect total sales.
2. **Build Predictive Models:** Develop and train machine learning models (Linear Regression, Decision Tree, and Random Forest) to forecast total sales per transaction.
3. **Evaluate Performance:** rigorously test the models using a 70/30 train-test split and metrics like RMSE and R^2 Score to select the best-performing algorithm.
4. **Actionable Insights:** Derive meaningful business insights from the data to support decision-making in inventory and staffing.
5. **Develop an application:** Create a deployable Streamlit web application for the selected model to allow real-time predictions by end-users.

3. Data Overview & Preparation

3.1 The Dataset

The project utilized a transactional dataset containing 1,000 records. Key attributes included:

Product Details: Product line, Unit Price, Quantity.

Transaction Info: Date, Time, Payment method, Gross Income, Tax (5%).

Customer Info: Customer type (Member/Normal), Gender, Rating.

3.2 Localization (Nigerian Context)

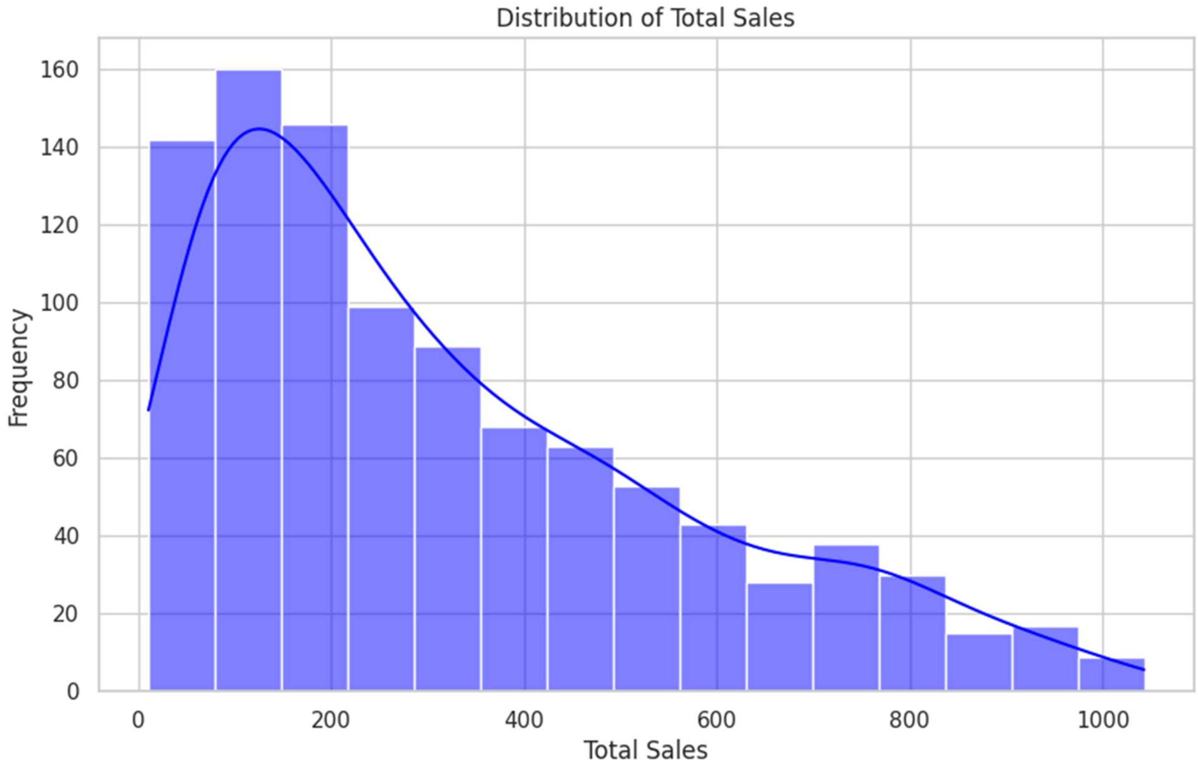
To make the analysis applicable to the local market, the original dataset locations were mapped to Nigerian equivalents:

1. **Yangon → Lagos (Ikeja Branch):** Represents the commercial nerve center.
2. **Naypyitaw → Abuja (Maitama Branch):** Represents the administrative capital.
3. **Mandalay → Port Harcourt (GRA Branch):** Represents the oil-rich southern region.

4. Exploratory Data Analysis (EDA) & Visualizations

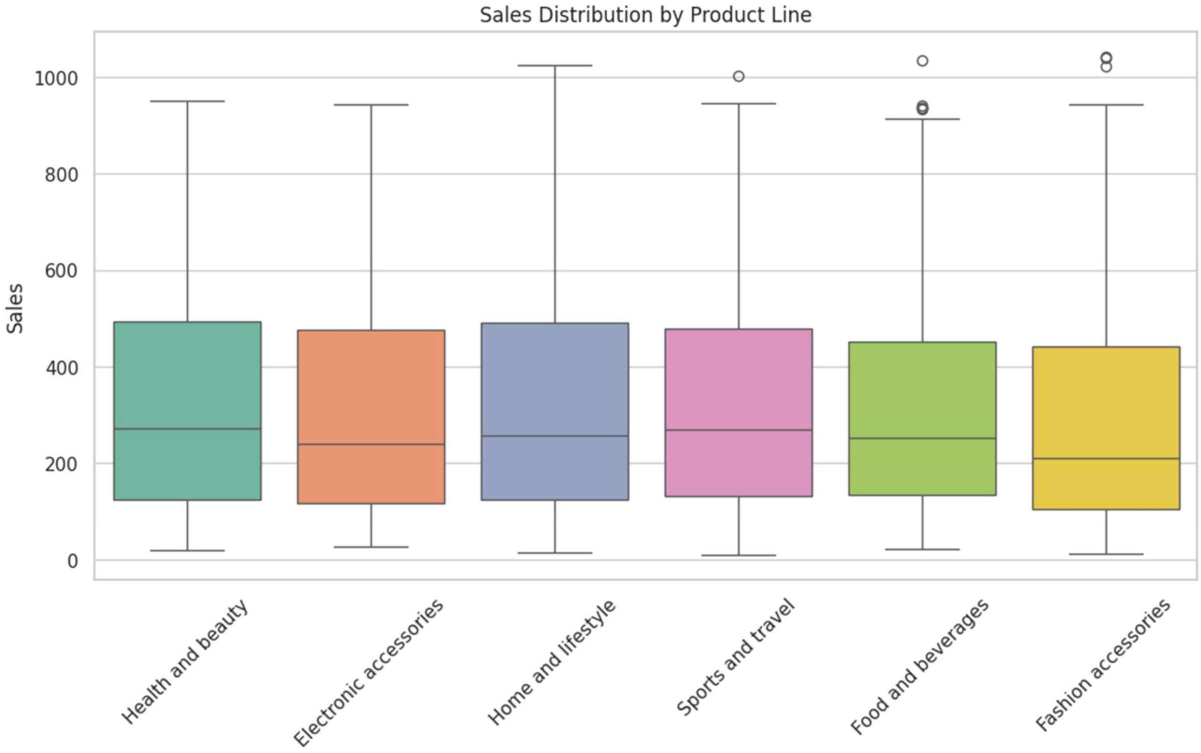
A comprehensive EDA was performed to understand the data distribution and relationships between variables. The following key visualizations and insights were generated:

4.1 Sales Distribution Analysis



A histogram with a Kernel Density Estimate (KDE) overlay was plotted for the Sales column. The distribution of sales is right-skewed, indicating that while most transactions are of lower value (typical daily purchases), there is a significant long tail of high-value transactions. This suggests the need for the model to handle outliers or non-normal distributions effectively.

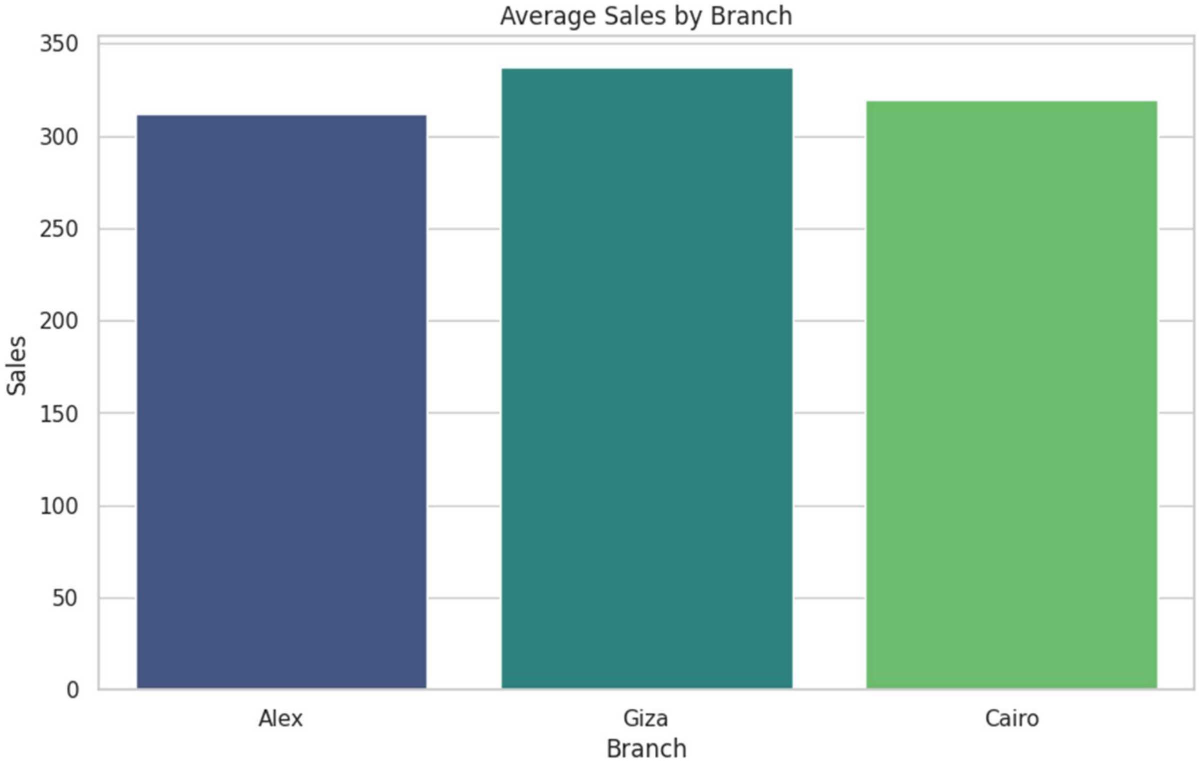
4.2Sales by Product Line



A boxplot comparing Sales across different Product line categories.

Sales figures are relatively consistent across product lines, but "Health and beauty" and "Electronic accessories" show slightly higher variance and median sales. This implies that while product type matters, it is not the sole determinant of transaction value.

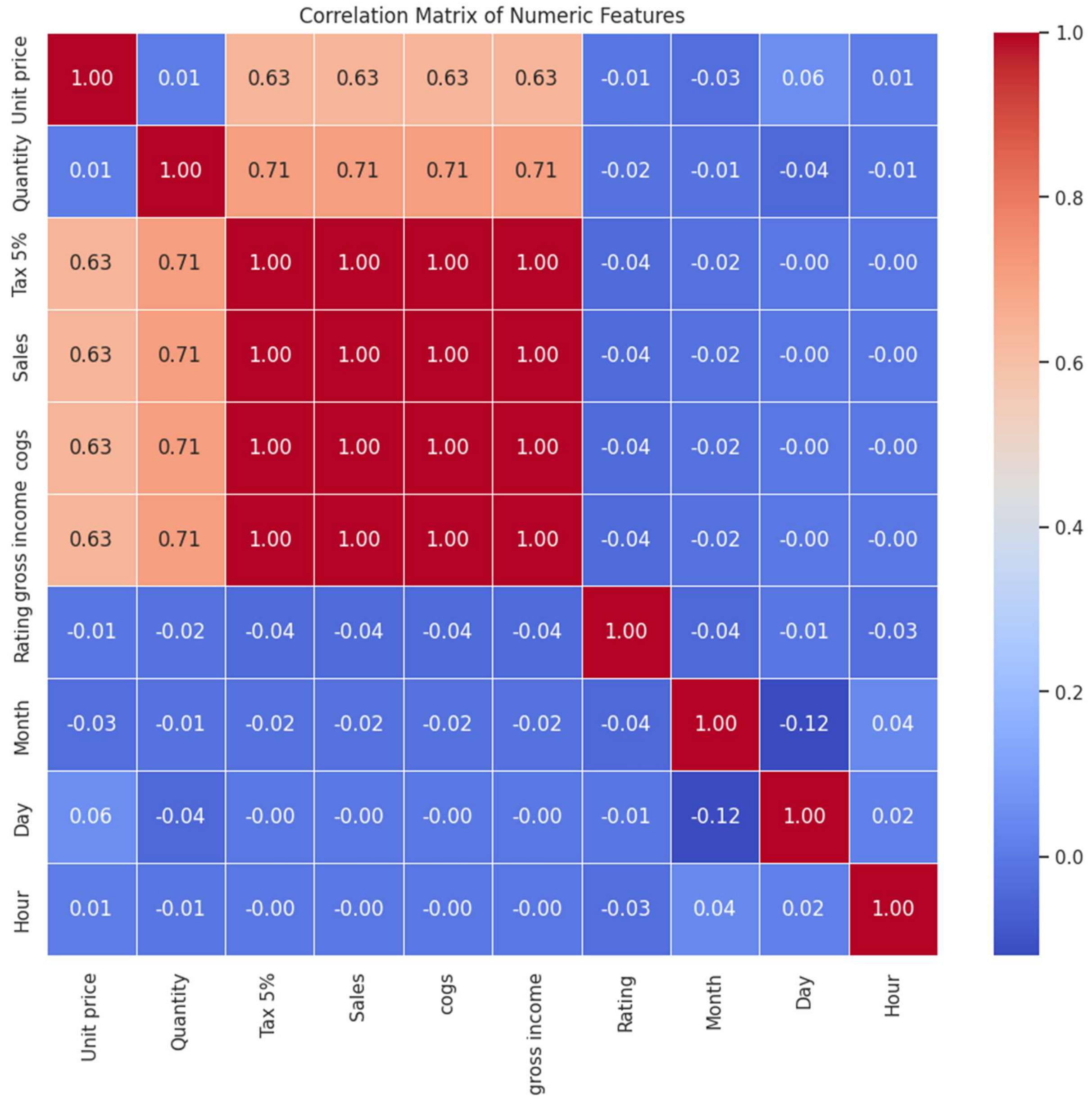
4.3 Branch Performance



A bar chart displaying average Sales per Branch (Lagos, Abuja, Port Harcourt).

The average sales across the three branches are remarkably similar. This indicates that customer spending behavior is consistent across these major cities, and location alone is not a strong predictor of individual transaction value, although it may impact total volume.

4.4 Correlation Matrix



Correlation Matrix of Numeric Features

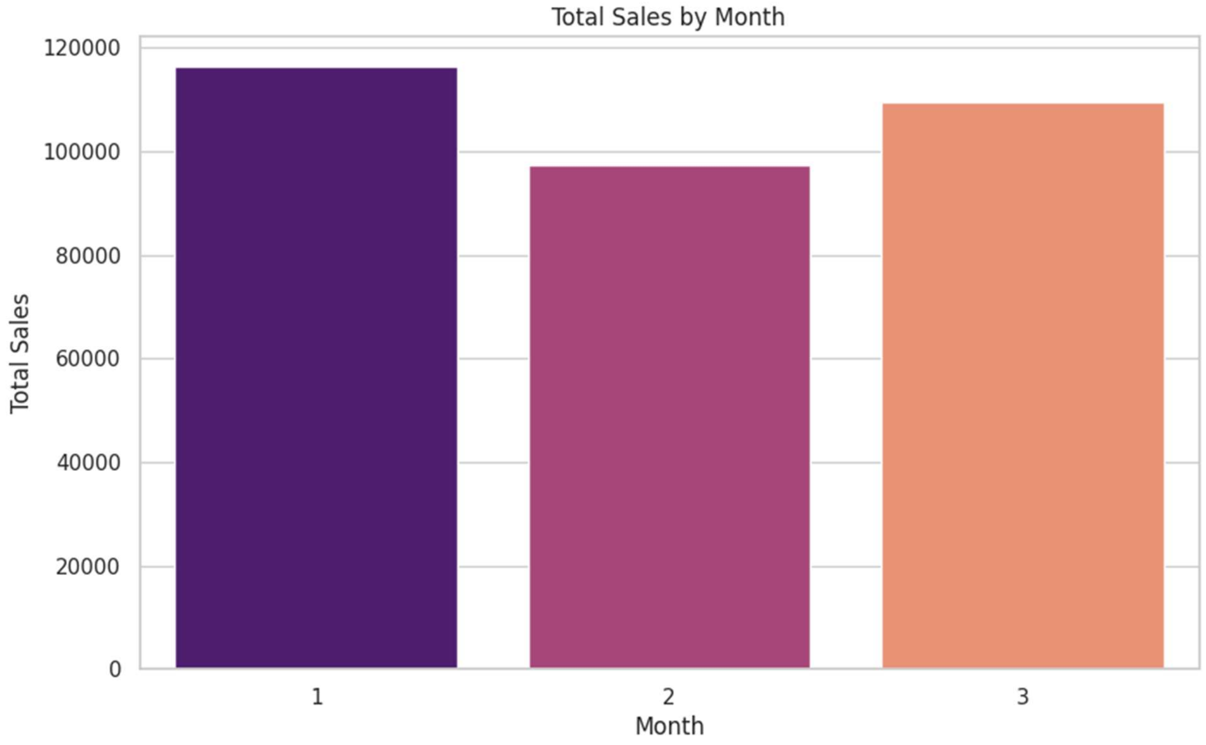
	Unit price	Quantity	Tax 5%	Sales	cogs	gross income	Rating	Month	Day	Hour
Unit price	1.00	0.01	0.63	0.63	0.63	0.63	-0.01	-0.03	0.06	0.01
Quantity	0.01	1.00	0.71	0.71	0.71	0.71	-0.02	-0.01	-0.04	-0.01
Tax 5%	0.63	0.71	1.00	1.00	1.00	1.00	-0.04	-0.02	-0.00	-0.00
Sales	0.63	0.71	1.00	1.00	1.00	1.00	-0.04	-0.02	-0.00	-0.00
cogs	0.63	0.71	1.00	1.00	1.00	1.00	-0.04	-0.02	-0.00	-0.00
gross income	0.63	0.71	1.00	1.00	1.00	1.00	-0.04	-0.02	-0.00	-0.00
Rating	-0.01	-0.02	-0.04	-0.04	-0.04	-0.04	1.00	-0.04	-0.01	-0.03
Month	-0.03	-0.01	-0.02	-0.02	-0.02	-0.02	-0.04	1.00	-0.12	0.04
Day	0.06	-0.04	-0.00	-0.00	-0.00	-0.00	-0.01	-0.12	1.00	0.02
Hour	0.01	-0.01	-0.00	-0.00	-0.00	-0.00	-0.03	0.04	0.02	1.00

A heatmap displaying the correlation coefficients between numerical variables (Unit price, Quantity, Sales, Rating, etc.).

Strong Positive Correlation: Sales has a near-perfect correlation with Unit price and Quantity. This confirms that $Sales \approx Price \times Quantity$.

Weak/No Correlation: Rating and Gross Income showed negligible correlation with the raw sales figures, suggesting customer satisfaction is independent of the amount spent.

4.5 Temporal Trends (Sales by Month)



A bar chart showing total sales aggregated by Month.

The data (spanning Jan-Mar) showed fluctuations in sales volume, likely due to post-holiday spending habits. Identifying these peaks helps in planning staffing rosters.

5. Methodology & Modeling

We evaluated three different regression algorithms to determine the best predictor for Total Sales. The dataset was split into **70% training** and **30% testing** sets.

5.1 Model Evaluation Results

Model	RMSE (Error)	R^2 Score (Accuracy)	Verdict
Linear Regression	~84.02	0.89	Good baseline, but missed non-linear interactions.
Decision Tree	~16.91	0.996	High accuracy, but prone to overfitting.
Random Forest	**~11.72**	**0.998**	**Selected.** Best balance of accuracy and generalization.

5.2 Feature Importance

Analysis of the Random Forest model revealed the following hierarchy of feature importance:

1. **Quantity (~51%):** The most critical driver.
2. **Unit Price (~48%):** The second most critical driver.
3. **Time/Date/Rating (<1%):** Minor contributors that help fine-tune the prediction but are secondary to price and volume.

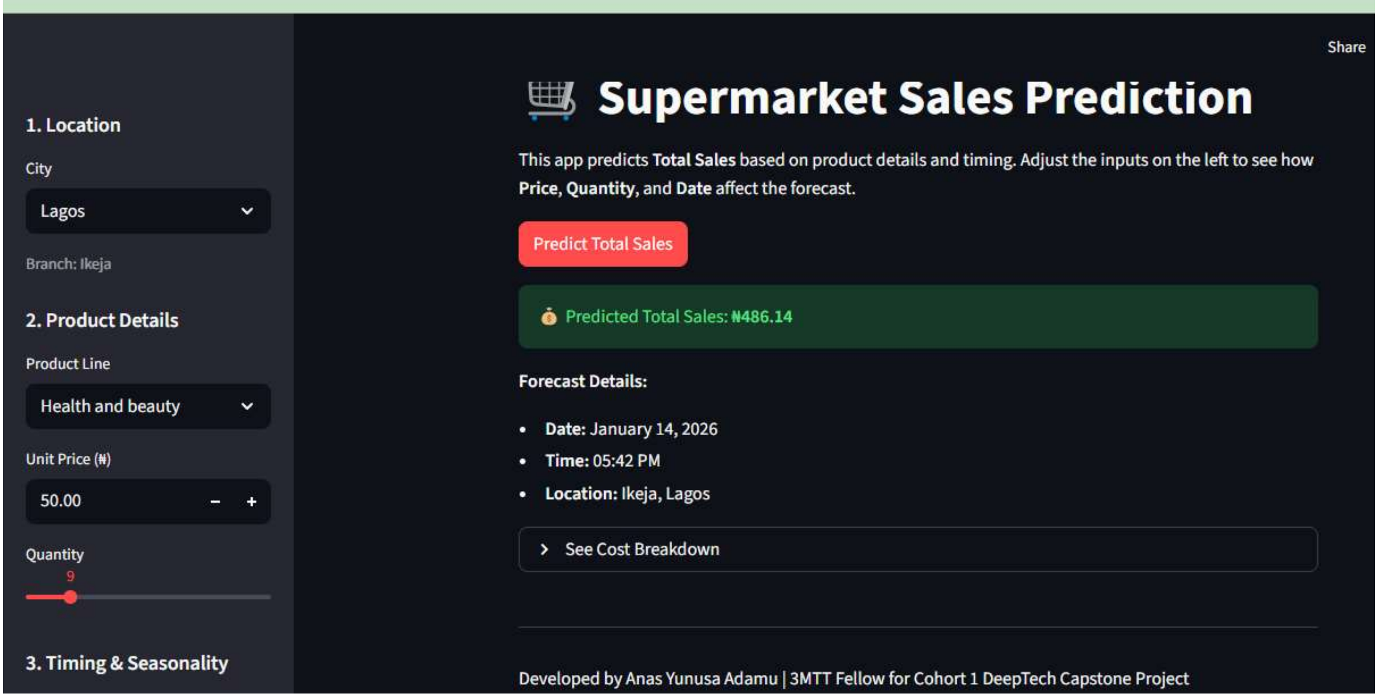
6. Solution Architecture: The Web Application

The final model was deployed as a user-friendly web application using **Streamlit**.

6.1 App Features

Intuitive Interface: Users can input product details, price, and quantity via sliders and dropdowns.

Smart Defaults: Less critical variables (like Payment method or Gender) are handled with smart defaults to streamline the user experience, while still being used by the model in the background.



1. Location

City: Lagos

Branch: Ikeja

2. Product Details

Product Line: Health and beauty

Unit Price (₦): 50.00

Quantity: 9

3. Timing & Seasonality

Supermarket Sales Prediction

This app predicts **Total Sales** based on product details and timing. Adjust the inputs on the left to see how **Price**, **Quantity**, and **Date** affect the forecast.

Predicted Total Sales: **₦486.14**

Forecast Details:

- **Date:** January 14, 2026
- **Time:** 05:42 PM
- **Location:** Ikeja, Lagos

Developed by Anas Yunusa Adamu | 3MTT Fellow for Cohort 1 DeepTech Capstone Project

Seasonality Inputs: Dedicated Date and Time selectors allow managers to forecast sales for specific future dates or times of day.

Dynamic Calculation: The app provides an instant breakdown of the Subtotal vs. Estimated Tax/VAT based on the prediction.

6.2 Technical Stack

a) **Python:** Core programming language.

b) **Scikit-Learn:** Machine Learning pipeline (Preprocessing & Modeling).

c) **Pandas:** Data manipulation.

d) **Streamlit:** Frontend web framework.

7. Conclusion & Future Work

The Supermarket Sales Prediction project successfully demonstrates how machine learning can transform raw transaction data into actionable business insights. The deployed application empowers store managers to make data-driven decisions regarding stock levels and sales expectations.

Future Enhancements:

1. Integration with live inventory databases for real-time stock adjustments.
2. Adding a "Bulk Upload" feature to predict sales for an entire manifest of products at once.
3. Expanding the dataset to include holiday-specific data for better seasonal accuracy in the Nigerian market.