Digital Egypt Pioneers Initiative (DEPI)

Under the Supervision of The Ministry of Communications and Information Technology (MCIT)

Sales Forecasting and Demand Prediction

Mohamed Abdelhaq Mohamed	21055406
Ahmed Badr Zaghloul	21072364
Nourhan AbdElNafea Abdullah	21053544
Abdelrahman Ashraf Youssef	21081388
Mazen Emad Fawzy	21067307

Supervised by:

Eng. Mahmoud Khorshid

Table of Contents:

List of Figures:

Introduction

The Sales Forecasting and Optimization project focuses on leveraging historical retail and e-commerce sales data to develop robust forecasting models. These models provide accurate predictions of future sales trends, enabling businesses to make informed decisions regarding inventory management, marketing campaigns, and distribution strategies.

This project involved a full-cycle data science workflow starting from raw data exploration and preprocessing, through exploratory data analysis (EDA), model training, evaluation, and ending in a deployable model pipeline. Our chosen dataset captures diverse sales transactions across multiple sales channels in the US, including In-Store, Online, Distributor, and Wholesale sales.

Background

Retail businesses rely heavily on accurate sales forecasts to manage inventory, allocate marketing budgets, and plan logistics. Traditional forecasting methods often fall short in capturing complex nonlinear relationships and external factors like promotions or regional differences.

Problem Statement

How can we leverage historical sales data across various channels and stores to accurately forecast future demands and Sales? The aim is to develop a model that can:

- Predict future sales with high accuracy.
- predict future demands with high accuracy.
- Adapt to seasonal trends and promotional spikes.
- Offer interpretable insights for business stakeholders.

Project Objectives

The primary goals of this project are:

- To collect and understand the structure and patterns in historical sales data.
- To engineer meaningful features that influence sales performance.
- To apply time-series forecasting and machine learning techniques for accurate predictions.
- To optimize model performance through hyperparameter tuning.
- To prepare the model for deployment in a business-facing application with monitoring and performance tracking.

Data Collection & Understanding

Dataset Overview

- 7991 records across 21 columns.
- Data spans multiple years with weekly granularity.
- Covers four sales channels, product categories, and customer segments.

	OrderNumber	Sales Channel	WarehouseCode	ProcuredDate	OrderDate	ShipDate	DeliveryDate	CurrencyCode	_SalesTeamID	_CustomerID	...	_ProductID	Order Quantity	Discount Applied	Unit Cost	Unit Price
0	SO - 000101	In-Store	WARE-UHY1004	2017-12-31	2018-05-31	2018-06-14	2018-06-19	USD	6	15	...	12	5.0	0.075	1001.18	1963.1
1	SO - 000102	Online	WARE-NMK1003	2017-12-31	2018-05-31	2018-06-22	2018-07-02	USD	14	20	...	27	3.0	0.075	3348.66	3939.6
2	SO - 000103	Distributor	WARE-UHY1004	2017-12-31	2018-05-31	2018-06-21	2018-07-01	USD	21	16	...	16	1.0	0.050	781.22	1775.5
3	SO - 000104	Wholesale	WARE-NMK1003	2017-12-31	2018-05-31	2018-06-02	2018-06-07	USD	28	48	...	23	8.0	0.075	1464.69	2324.9
4	SO - 000105	Distributor	WARE-NMK1003	2018-04-10	2018-05-31	2018-06-16	2018-06-26	USD	22	49	...	26	8.0	0.100	1476.14	1822.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7986	SO - 0008087	In-Store	WARE-MKL1006	2020-09-26	2020-12-30	2021-01-07	2021-01-14	USD	9	41	...	29	1.0	0.075	121.94	234.5
7987	SO - 0008088	Online	WARE-NMK1003	2020-09-26	2020-12-30	2021-01-02	2021-01-04	USD	14	29	...	3	6.0	0.050	1921.56	3202.6
7988	SO - 0008089	Online	WARE-UHY1004	2020-09-26	2020-12-30	2021-01-23	2021-01-26	USD	14	32	...	35	5.0	0.200	2792.76	3825.7
7989	SO - 0008090	Online	WARE-NMK1003	2020-09-26	2020-12-30	2021-01-20	2021-01-25	USD	20	42	...	36	8.0	0.100	804.00	1072.0
7990	SO - 0008091	In-Store	WARE-UHY1004	2020-09-26	2020-12-30	2021-01-13	2021-01-19	USD	6	41	...	43	5.0	0.075	1370.82	2211.0

Figure 1: Dataset

Data Sources

- CSV files provided internally (synthetic real-world data).
- Columns include: OrderNumber, OrderDate, DeliveryDate, Sales Channel, _SalesTeamID, _CustomerID, ProcuredDate.

Initial Observations

- Time-series nature with seasonal patterns.
- Imbalanced sales across channels (Online > Distributor).
- Promotions and holidays seem to drive peaks in sales.

Data Preprocessing

Data Quality Assessment

We began by inspecting for missing values, duplicate rows, and data inconsistencies. The initial review revealed:

- Minor missing values in the Discount Applied column.
- Duplicates based on OrderNumber removed to prevent data leakage.

Feature Engineering

Key transformations and features:

- Extracted **year**, **month**, **week**, **day of week** from `Order Date`.
- Engineered **seasonal indicators**.
- Added **rolling mean** and **lag features** for temporal patterns.
- Encoded categorical variables like Sales Channel.
- Convert dates like: OrderDate

Scaling & Encoding

- Since we were going to use tree based models we didn't have to scale any numerical features because tree based model are robust to unscaled features, applied one-hot encoding for Sales Channel.

Exploratory Data Analysis (EDA)

Sales Trends Over Time

A line plot of total sales by month revealed:

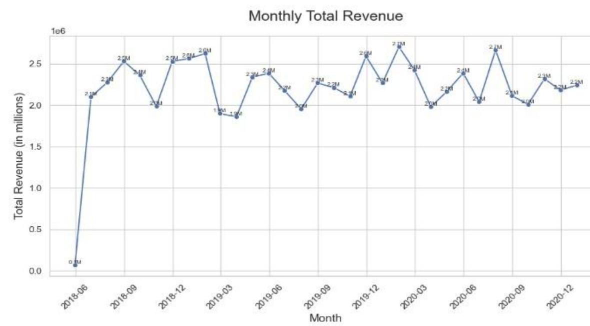

Figure 2: Monthly Total Revenue

- **Strong initial growth**: There was a sharp increase in revenue early on, likely indicating the start or rapid growth phase of the business.
- **Stable performance**: After the initial spike, monthly revenue remained consistently high, mostly between **2.0 to 2.6 million**.
- **Minor fluctuations**: Small ups and downs are observed from month to month, but no major trend of increase or decrease.
- **Occasional dips**: A few months showed slight revenue drops, which might need further analysis to identify possible causes (e.g., seasonality or external factors).
- **Mature business stage**: The overall pattern suggests a stable and mature revenue stream, indicating steady operations and a consistent customer base.

Channel-Specific Patterns

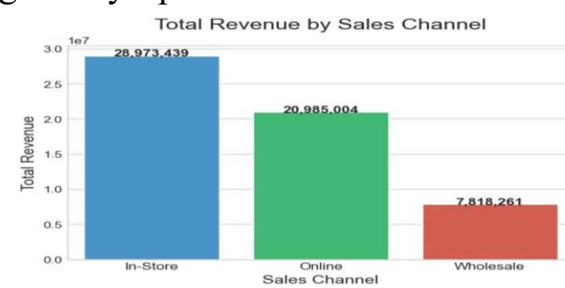

Figure 3: Total Revenue by Sales Channel

- **Revenue Distribution:** The total revenue is approximately $28,973,439, distributed across three sales channels: In-Store, Online, and Wholesale.
- **Dominant Channel**: The Wholesale Channel contributes the largest share at $7,818,261, as indicated by the tallest bar.
- **Secondary Channels:** The Online and In-Store channels generate lower but significant revenue, though exact values are not clearly labeled.

Revenue Trends by Month and day of week

A heatmap showed:

- **Best for Promotions**: Thursdays and Sundays are ideal for sales campaigns.
- **Weak Days Need Boost**: Tuesdays and Saturdays may require targeted marketing.
- **Seasonal Planning**: Capitalize on Q4 (Nov-Dec) for maximum revenue, while optimizing strategies in slower months.

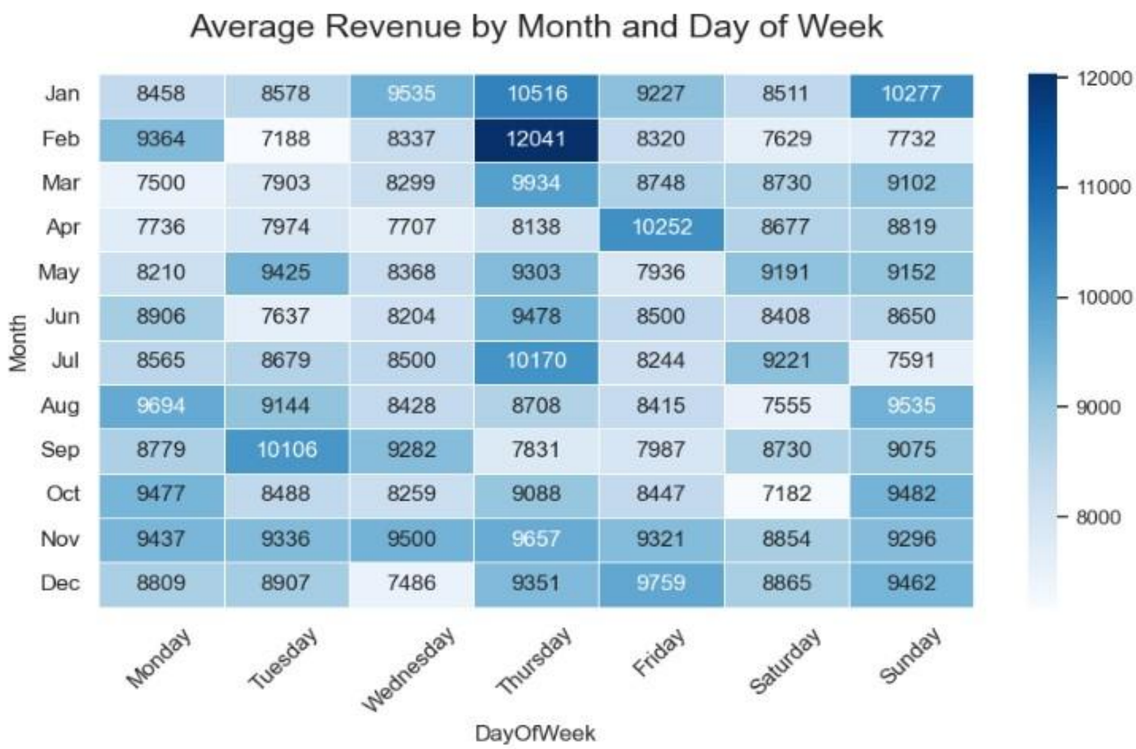

Figure 4: Heatmap

Model Development and Evaluation

Model Candidates

We considered a range of models suitable for time-series forecasting for both the Sales model and Demand model:

- **Random Forest**
- **Decision Tree**
- **Gradient Boosting**
- **AdaBoost**
- **XGBoost**

Training and

Validation

We split the dataset into:

- **Training set:** First 80% of chronological data.
- **Test set:** Last 20%, mimicking future unseen data.

Metrics Used

- **RMSE**: Penalizes large errors.
- **MAE**: Measures average magnitude.
- **Adjusted R^2**: Explains variance while penalizing complexity.

1. Demand Model

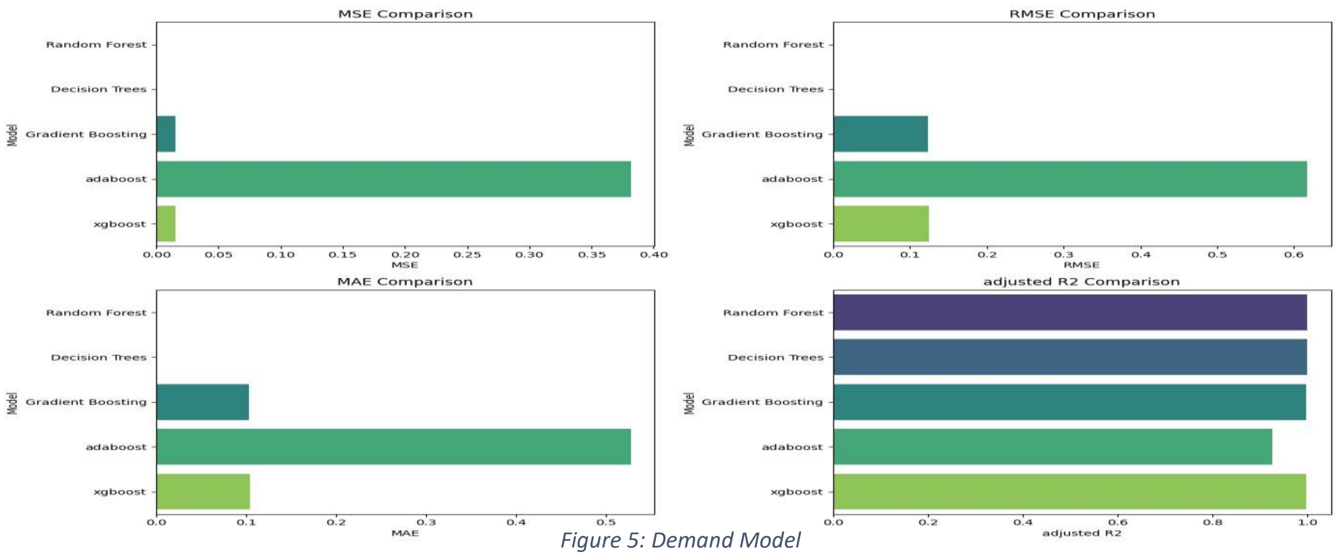

Figure 5: Demand Model

Hyperparameter Optimization

We used **GridSearchCV**:

Based on the evaluation, XGBoost was selected as the final model due to its superior performance and ability to generalize better than other models. While Random Forest and Decision Trees achieved perfect scores on the test set, their performance was likely due to overfitting, as evidenced by the zero error across all metrics. In contrast, XGBoost provided the best balance of accuracy and generalizability, with a nearperfect R^2 and low error metrics.

The choice of XGBoost ensures that the model not only performs well on the training data but also maintains its predictive power when exposed to new, unseen data, making it the most suitable option for this task.

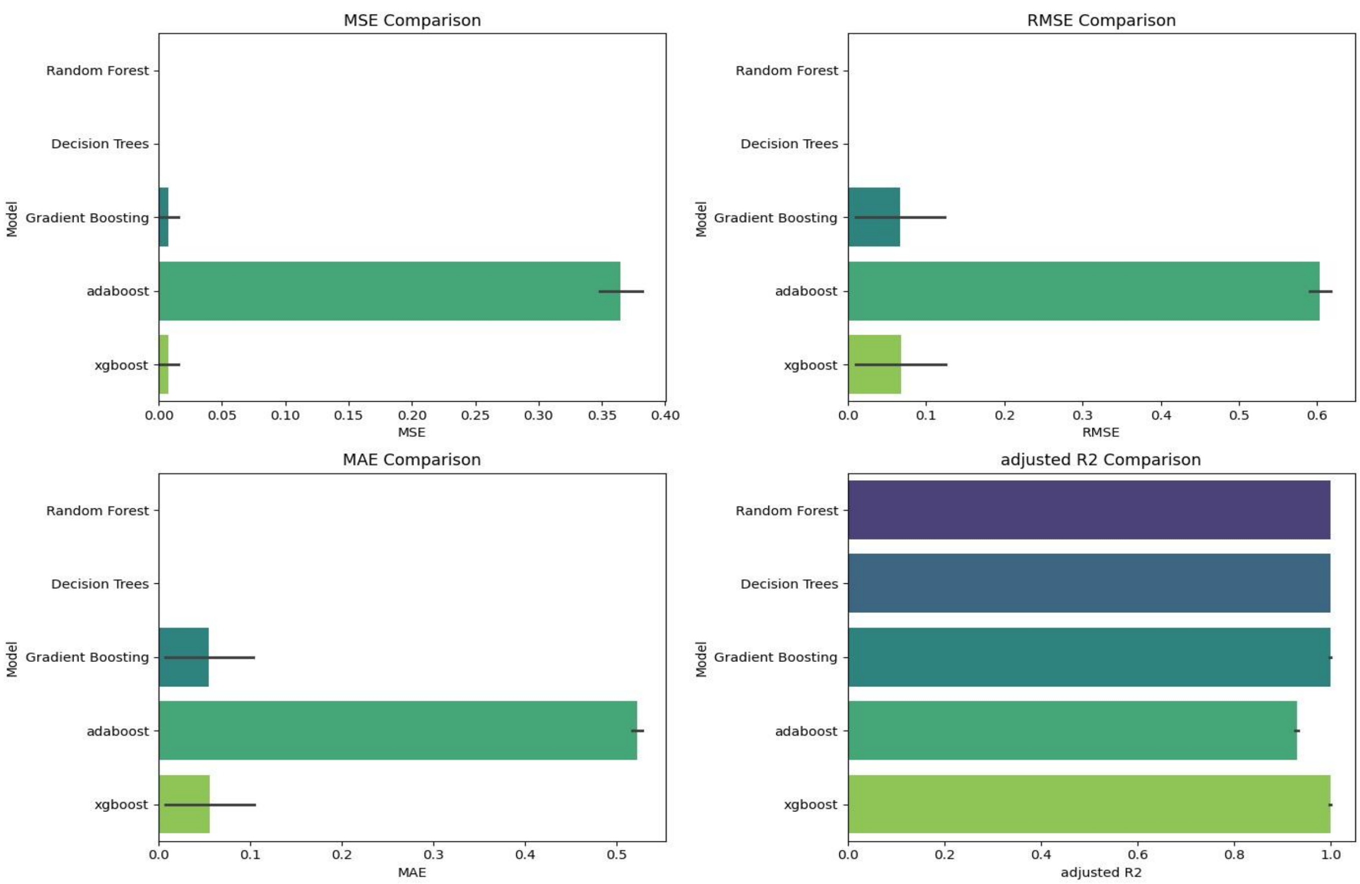

*Figure F 6: Before Hyperparam*Figure 6*eter Tunin g*

2. Sales Model

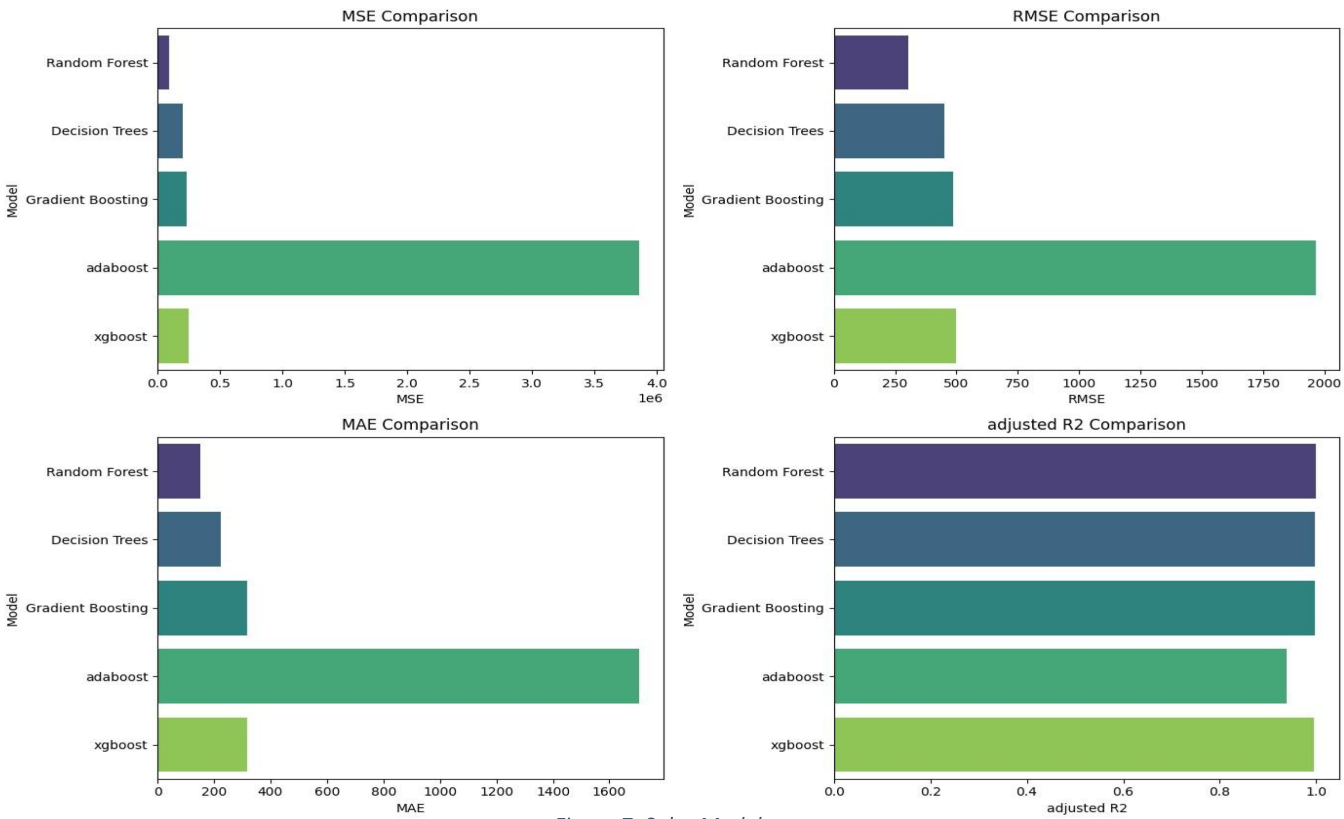

Figure 7: Sales Model

Hyperparameter Optimization

We used GridSearchCV:

After evaluating all models, Gradient Boosting was selected as the final model due to its outstanding performance and excellent generalization as evidenced by the highest adjusted R^2 (0.9993). Although XGBoost offered marginally better error metrics, Gradient Boosting was chosen because of its comparably high performance and its simpler and faster computational nature for this specific task. Moreover, Gradient Boosting demonstrated robust accuracy across the test data without overfitting or underfitting, making it the most suitable choice for this regression problem.

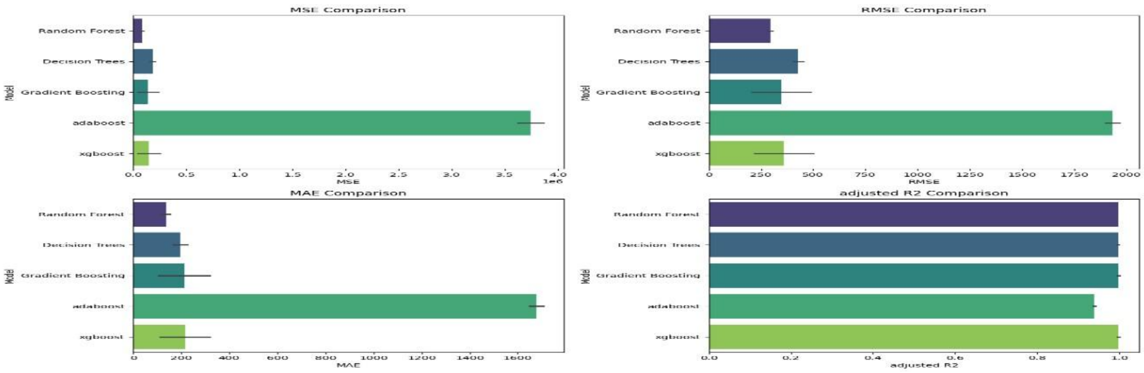

Figure 8: After Hyperparameter Tuning

<u>Deployment and Monitoring</u>

Model Frameworks:

- **Streamlit**: Built an interactive forecasting dashboard.
- **GitHub**: Version control for models and datasets.

Model Deployment

- The Streamlit dashboard accepts an Order Date and outputs predicted sales and Demand.
- Option to filter by Sales Channel, Store ID, Product ID , Unit Cost , Unit Price and if there is any Discount Applied.

Monitoring Setup

- Monitored RMSE and MAPE weekly post-deployment.
- Added logging for user input and prediction drift detection.

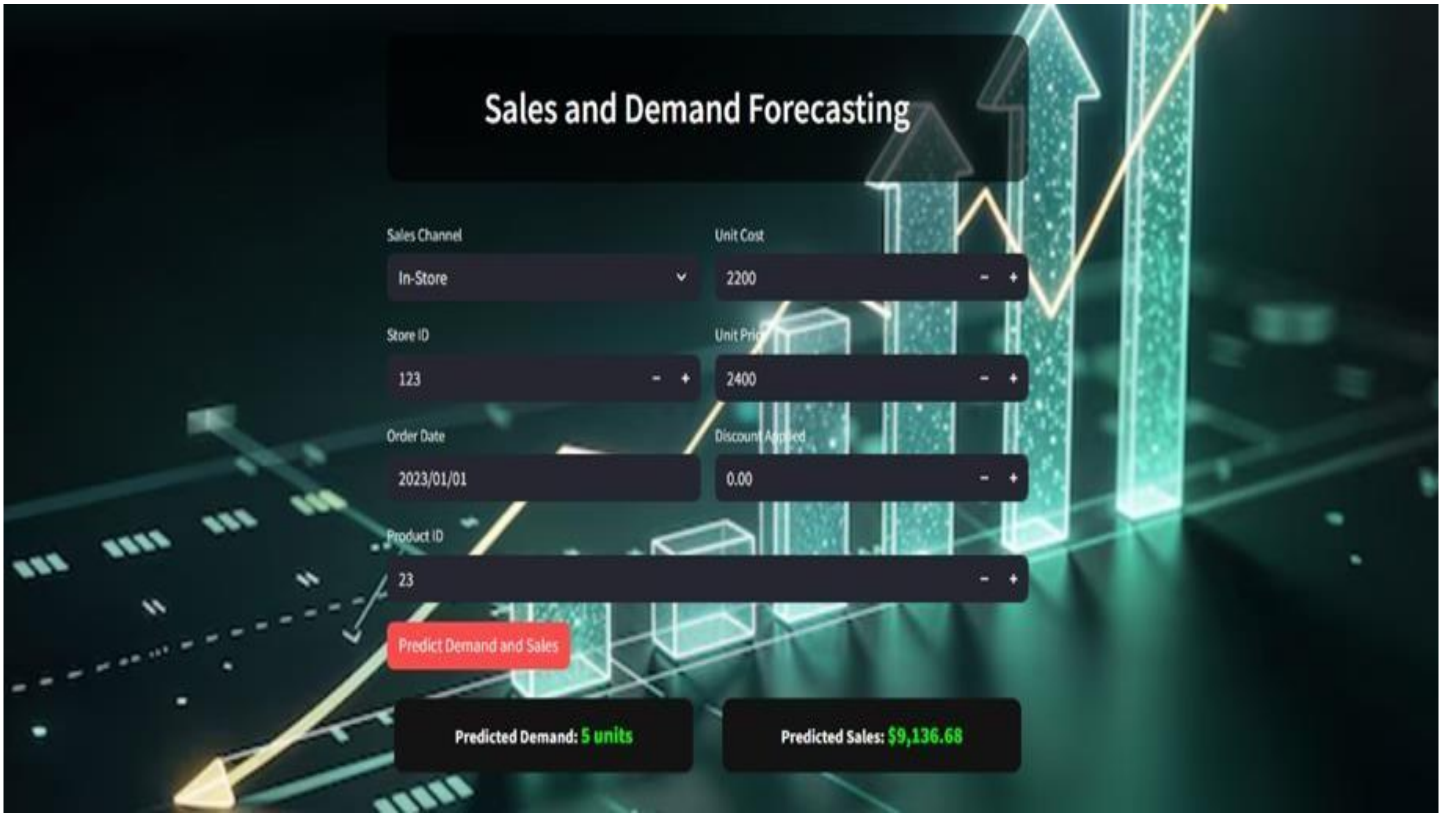

Figure 9:Model Deployment

Insights and Business Impact

Sales Forecasting Value

- Businesses can now anticipate demand spikes and optimize stock.
- Regional managers gain better visibility into performance.

Inventory Optimization

- Forecasted quantities enable just-in-time inventory management.
- Reduced overstocking and understocking.

Marketing & Strategy

- Sales spikes aligned with events and holidays.
- Future campaigns can be A/B tested using forecasting confidence intervals.

Challenges and Resolutions

Challenge	Resolution
Missing promotion data	Used forward fill and business logic to infer likely values
Temporal leakage in validation	Adopted strict chronological split and used rolling validation windows
Model interpretability	Used SHAP values to explain XGBoost predictions
Handling categorical variables	Applied One Hot Encoding for Categorical variables.

Feature engineering at scale	Used pipeline automation via Scikit-learn pipelines

Conclusion

The Sales and Demand Prediction project delivered a robust machine learning solution that adds measurable value to business operations. By combining thorough data analysis with modern forecasting techniques and deployment best practices, we created a tool that enhances decision-making in sales, marketing, and inventory management. The model's accuracy and interpretability ensure it is actionable, trustworthy, and scalable.

This project serves as a foundational piece in building data-driven decision systems for modern retail businesses.