Digital Egypt Pioneers

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

DEPI Graduation Project IBM Data Scientist Track

Sales Forecasting and Optimization

Team 7

Team Members:

Toka Khaled

Mariam Hassan

Rawan Sotohy

Sara Abdelrahman

Supervisor: Eng. Mahmoud Khorshid

Table of Contents

1. Introduction

This project aims to predict sales and optimize business operations using historical sales data. Accurate forecasts help improve strategic decisions in inventory management, marketing, and logistics.

---

2. Data Collection and Description

- **Source:** Sales transaction data (9,800 records, 18 features) from Kaggle
- **Key Features:**
  - Customer Demographics: Segment, City, State, Region
  - Order Information: Order Date, Ship Date, Ship Mode
  - Product Details: Category, Sub-Category, Product Name
  - Target: Sales (USD)

---

3. Data Preprocessing and Feature Engineering

- **Column Renaming:** To maintain a consistent and programming-friendly format, several columns were renamed. For example, "Ship Mode" became "Ship_Mode", and "Product Name" was renamed to "Product_Name".
- **Duplicate Check:** We verified the dataset for duplicate rows and confirmed that there were no duplicates, ensuring data integrity.
- **Missing Values:** Removed 11 rows with missing postal codes
- **Outlier Treatment:** Capped sales outliers using the IQR method
- **Encoding:** Label encoding on categorical features (e.g., Ship Mode, Segment, Region)
- **Standardized Product Names:** Converted product names to lowercase and removed punctuation for consistency and to reduce redundancy in categorical values.
- **Feature Engineering:**

    - Extracted Year, Month, Day, Weekday from order dates
    - Calculated shipping duration

---

4. Feature Selection

- **Dropping Unnecessary Columns:**
    - Columns that were either identifiers, redundant, or not useful for predictive modeling were removed: 'Row ID', 'Customer ID', 'Product ID', 'Customer Name', 'Order ID', and 'Ship_Date'.
- **Correlation Analysis:**
    - We computed the absolute correlation values between each numerical feature and the target variable **Sales**.
    - Features with low or negligible correlation were reviewed for potential removal.
- **Additional Feature Dropping Based on Correlation and Relevance:**
    - The following features were dropped due to low correlation with Sales or redundancy: 'Segment', 'Weekday', 'Day', and 'Year'.
- **Final Selected Features for Modeling:**
  Ship_Mode, City, State, Region, Category, Sub_Category, Product_Name, Month, Shipping_Duration, Postal Code

---

5. Exploratory Data Analysis

- Checked for nulls, data types, and unique categories
- Visualized distributions and relationships
- Noted high cardinality in cities and product names
- Created a correlation matrix to identify relationships with the target variable

---

6. Modeling and Evaluation

Forecasting Models:

- **ARIMA, SARIMA, Prophet**: Used for time-series predictions for the next 30 days

Machine Learning Models:

- **Ensemble Boosting Models (with Hyperparameter Tuning):**
  - **XGBoost:** $R^2 = 0.571$
  - **LightGBM:** $R^2 = 0.570$
  - **CatBoost:** $R^2 = 0.528$
- **Bagging Models:**
  - **Random Forest:** $R^2 =$ **0.624** (Best performing model)
  - **Decision Tree:** $R^2 = 0.310$

Final Selected Model: Random Forest due to its high performance, low error, and generalization ability.

---

7. Deployment Strategy

The final machine learning model was deployed using **Streamlit Cloud**, offering an interactive and user-friendly interface for sales prediction, data exploration, and dashboard analytics. The deployment includes multiple functional pages, each tailored for different use cases:

1) **Home Page – Dataset Overview**

- Provides an introduction to the dataset, including:
  - Data description, objectives, and structure
  - A **Data Dictionary** with column names, data types, and explanations
- Links to:
  - The presentation (hosted on Google Drive)
  - The project notebook (hosted on GitHub)

2) Prediction Page

- Allows users to input order details including:
  - Sub-Category, Region, State, City, Order Date, Ship Date, Postal Code, Ship Mode, Product Name, and Category
- Displays the predicted **sales value** in real-time

3) Plotly Dashboard Page

- Offers rich visual insights based on:
  - RFM analysis (Recency, Frequency, Monetary)
  - Monthly sales trends, delivery times, and churn insights
  - Top customers by revenue
  - Segmented visualizations using sunbursts, treemaps, pie charts, and histograms
- Enhances decision-making with **interactive charts**

4) Power BI Dashboard Page

- Embeds an online **Power BI report** directly into the app
- Displays high-level analytics including:
  - Profitability, segment performance, churn behavior
  - Filterable views by region, category, and time period
- Complements Plotly visuals with a business-optimized dashboard interface

---

8. Model Performance Summary

Model	**R² Score**
Random Forest	**0.624**
Decision Tree	0.310

Model	R² Score
XGBoost (Tuned)	0.571
LightGBM (Tuned)	0.570
CatBoost (Tuned)	0.528

9. Business Impact

- **Informed Decisions:** Enables teams to forecast sales and plan inventory, marketing, and resource allocation
- **Revenue Optimization:** Aligns product availability with demand trends
- **Cost Efficiency:** Avoids overstocking and understocking
- **Customer Satisfaction:** Ensures better product availability and quicker service

10. Conclusion

The project successfully combined advanced forecasting models with a user-friendly deployment, driving real business value. By integrating dashboards, predictive insights, and seamless deployment, the system supports smarter, data-driven decision-making.