SALES FORECASTING
and
DEMAND PREDICTION

Supervised by: Eng. Mahmoud Talaat

Table of Content

01 Introduction

02 Problem Statement

03 Proposed solution

04 Methodology

05 Conclusion

Introduction

Sales Forecasting & Demand Prediction is a data science project aimed at building a predictive system that accurately forecasts future product demand using historical sales data and external variables such as promotions, holidays, and weather conditions. The objective is to help businesses reduce stockouts and overstocking, improve operational and financial planning, and enable more effective, data-driven decisions across marketing, supply chain, and inventory management.

Problem Statement

- Businesses often struggle to predict demand accurately in a world driven by dynamic factors like promotions, seasonality, and customer behavior.
- Traditional methods often lead to stockouts, overstocking, and lost opportunities.
- This creates a critical need for a forecasting system that aligns inventory, marketing, and operations with real-world demand patterns.

Proposed Solution

We built a machine learning model that uses historical sales data and time-based features to forecast future demand.

Key steps included:

- ○ Segment analysis through EDA.
- ○ Feature engineering (lags, promotions, seasonality).
- ○ Model training with XGBoost for high accuracy.
- ○ Deployment using Flask to serve forecasts via API.
- ○ Forecast generation to guide inventory and marketing decisions.

This solution enables accurate, accessible, and data-driven sales planning.

Methodology

1.Problem Understanding:

- Defined business objective: Improve sales forecasting accuracy to optimize inventory and planning.

2. Data Collection:

- Collected historical sales data, customer segments, promotional events, holidays, and weather factors.

3. Data Preprocessing:

- Handled missing values, outliers, and duplicates.
- Performed feature engineering

4. Exploratory Data Analysis (EDA):

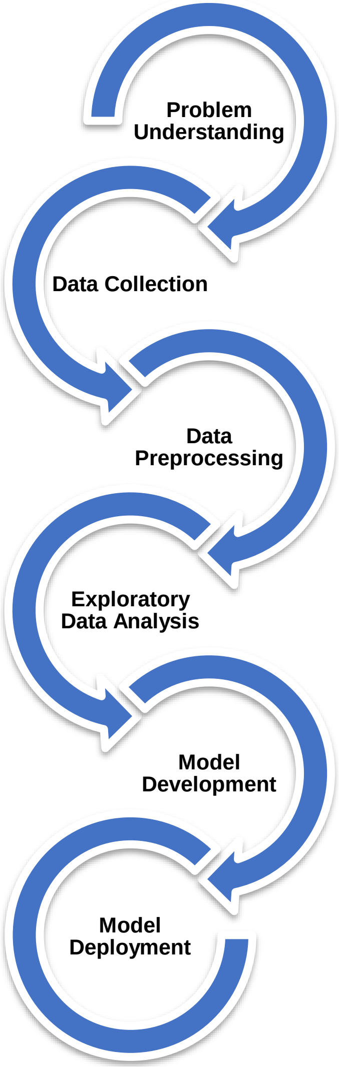
- Identified trends, seasonality, and high-performing segments
- Visualized segment-wise sales and demand behavior.

5. Model Development:

- Tested multiple models.
- Selected best-performing model based on RMSE and accuracy.

6. Model Deployment:

- Deployed model using Flask and IBM Cloud to generate real-time predictions.



Problem Understanding

Data Collection

Data Preprocessing

Exploratory Data Analysis

Model Development

Model Deployment

Methodology

“Problem Understanding”

Business Objective:

To enhance the accuracy of sales forecasting in order to optimize inventory levels, reduce waste, and align business operations with real demand.

Challenge:

Traditional forecasting methods fail to account for time-based patterns and external influences like promotions or customer segment behavior, leading to overstocking or stockouts.

Methodology

“Data Collection”

Data Source:

- Historical sales data from Kaggle (CSV dataset).
- Included features: Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, Profit.

Data Scope:

- Captures transactions across multiple customer segments and product categories.
- Date range spans multiple years, allowing analysis of trends and seasonality.

Methodology

“Data Preprocessing”

Part 1: Data Cleaning

- **Date Conversion:** Converted ***Order Date*** and ***Ship Date*** to datetime format for time-based operations.
- **Removed Unnecessary Columns:** Dropped irrelevant or duplicate columns not useful for modeling (e.g., IDs).
- **Handled Missing Values:** Checked for and confirmed the absence of missing values in relevant columns.
- **Removed Duplicates & Outliers:** Ensured data uniqueness and integrity by removing any duplicated records or outliers.

Methodology

"Data Preprocessing"

Part 2: Feature Engineering

- **Time-Based Features:** Extracted Year, Month, Day, and Weekday from Order Date.
- **Sales Aggregation:** Grouped sales data monthly by Segment to generate a time-series format.
- **Lag Features:** Created lag-based features (e.g., previous month's sales) to capture momentum.

Methodology

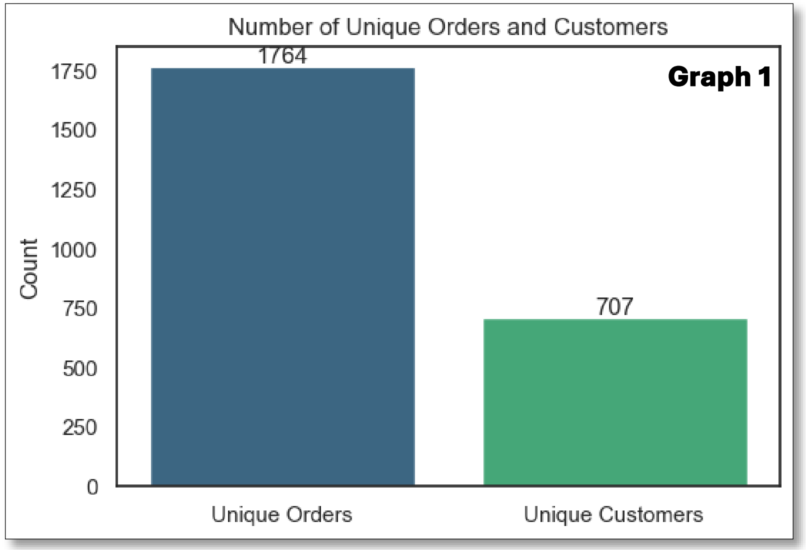
"Data Preprocessing"

Part 3: Encoding & Transformation

- **Categorical Encoding:** Encoded segment and categorical columns where necessary.
- **Data Formatting:** Reformatted the grouped data into a structured time series dataframe for modeling.

Methodology

"Exploratory Data Analysis"



Graph 1: Number of Unique Orders vs. Customers

Description:

This bar chart compares the total number of unique orders (1764) to the number of unique customers (707) in the dataset.

Key Insight:

Customers place multiple orders on average (~2.5 per customer), indicating repeat business or bulk purchasing behavior.

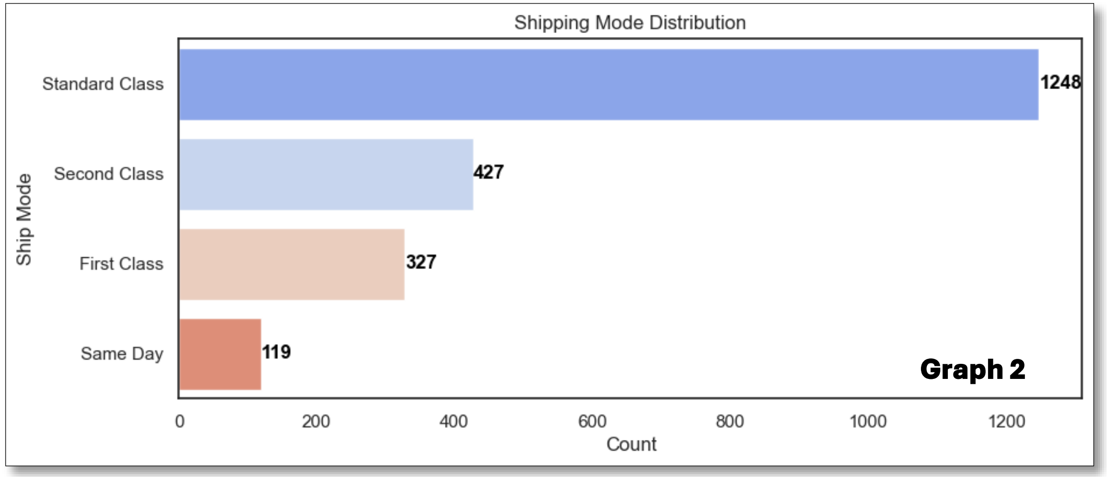
Graph 2: Shipping Mode Distribution

Description:

This horizontal bar chart shows how often each shipping mode was used.

Key Insight:

Most customers prefer cost-effective Standard Class , while faster options like Same Day are rarely chosen.



Methodology

"Exploratory Data Analysis"

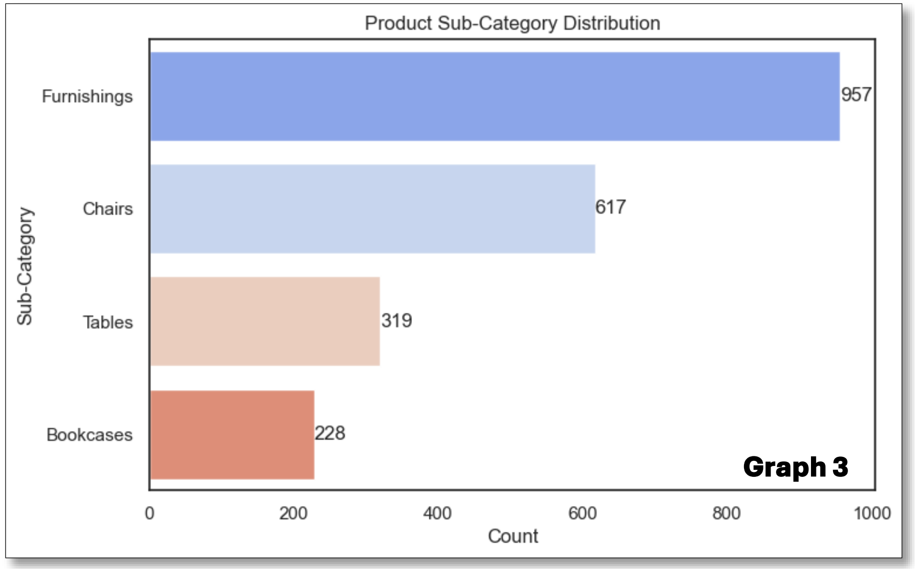
Graph 3: Product Sub-Category Distribution

Description:

This horizontal bar chart shows how frequently different furniture sub-categories appear in the dataset.

Key Insight:

Customers buy more furniture accessories (Furnishings) and seating (Chairs) than larger items like Tables or Bookcases .



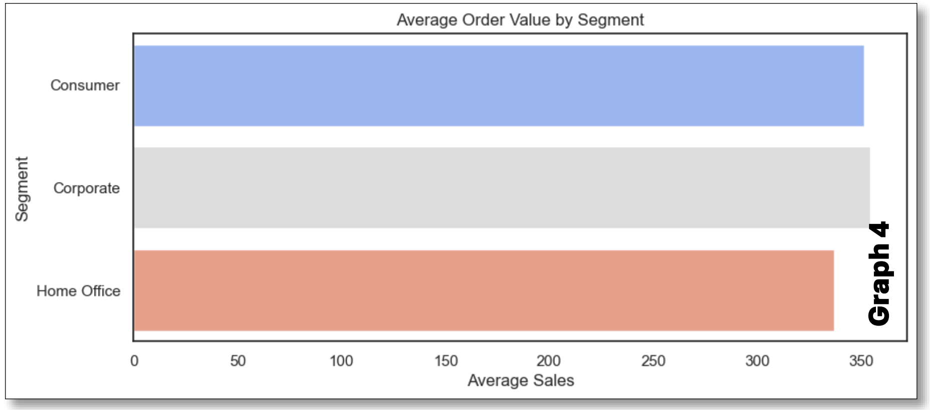
Graph 4: Average Order Value by Segment

Description:

This horizontal bar chart shows the average order value for each customer segment.

Key Insight:

Consumer and Corporate customers spend about the same per order, while Home Office customers spend slightly less..



Methodology

"Exploratory Data Analysis"

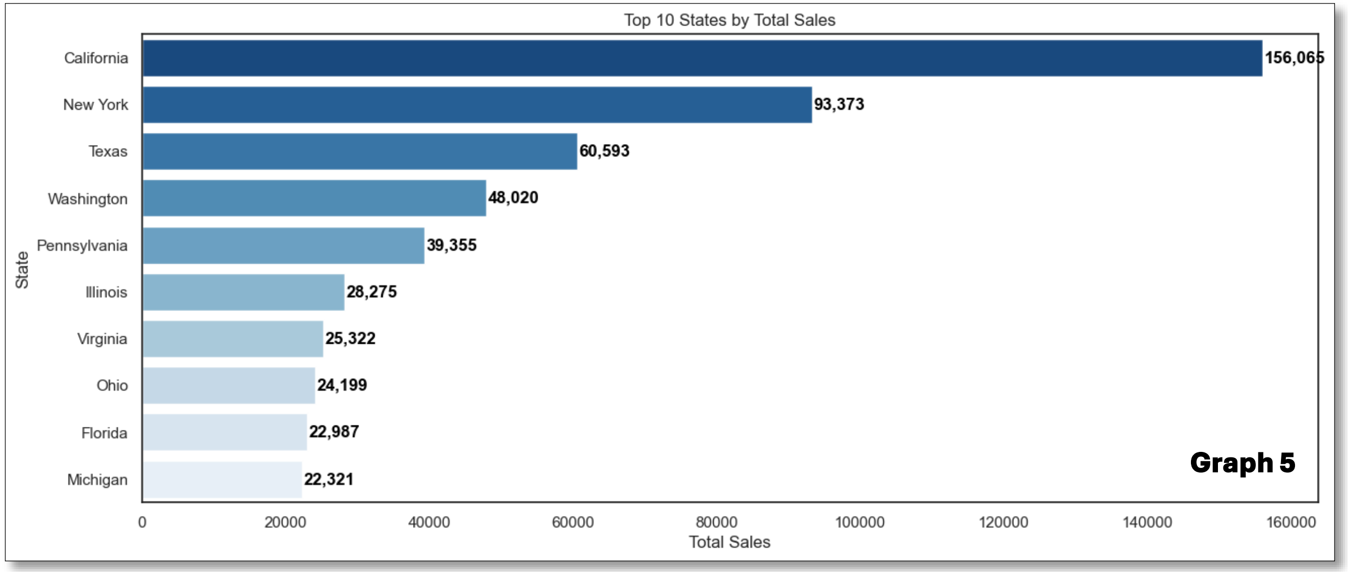
Graph 5: Top 10 States by Total Sales

Description:

This horizontal bar chart ranks the top 10 states based on their total sales. The chart shows the following.

Key Insight:

California Dominates Sales : California has significantly higher sales compared to other states, indicating it is a major market for the business .



Methodology

“Exploratory Data Analysis”

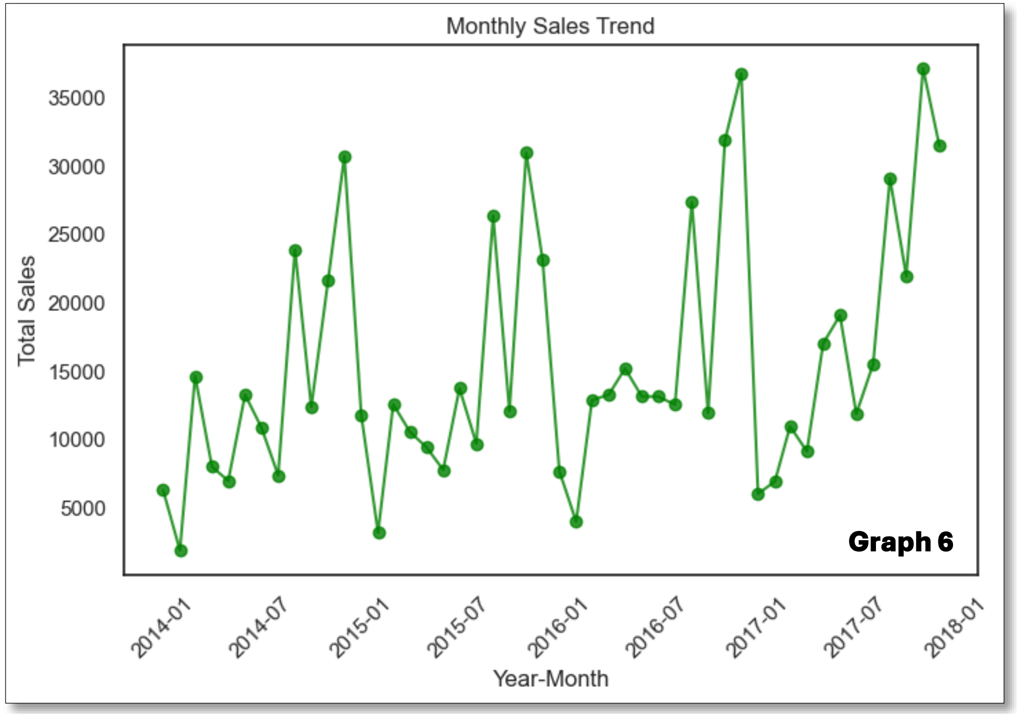
Graph 6: Monthly Sales Trend

Description:

This line chart illustrates the total sales over time, measured monthly from January 2014 to January 2018 . The x-axis represents the year-month, and the y-axis shows the total sales in dollars..

Key Insight:

- Steady sales growth from 2014 to 2018.
- Seasonal peaks suggest strong holiday and mid-year demand.
- Mid-2016 dip indicates a possible market challenge or supply issue.
- Post-2016 recovery shows resilience and growth potential.



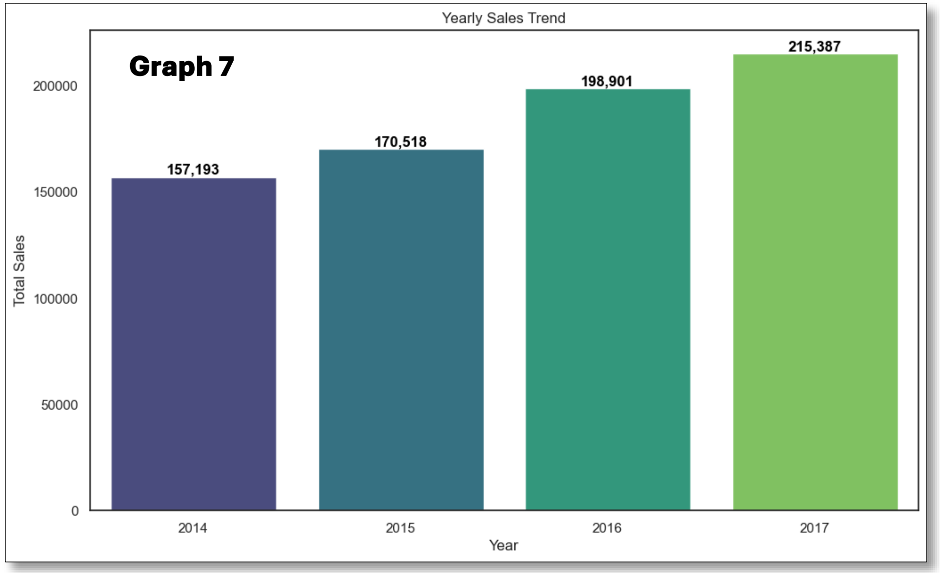
Methodology

“Exploratory Data Analysis”

Graph 7: Yearly Sales Trend

Description:
This bar chart shows the total sales for each year from 2014 to 2017.

Key Insight:
Sales have consistently increased year-over-year , with a steady growth trend over the four years.



Methodology

"Exploratory Data Analysis"

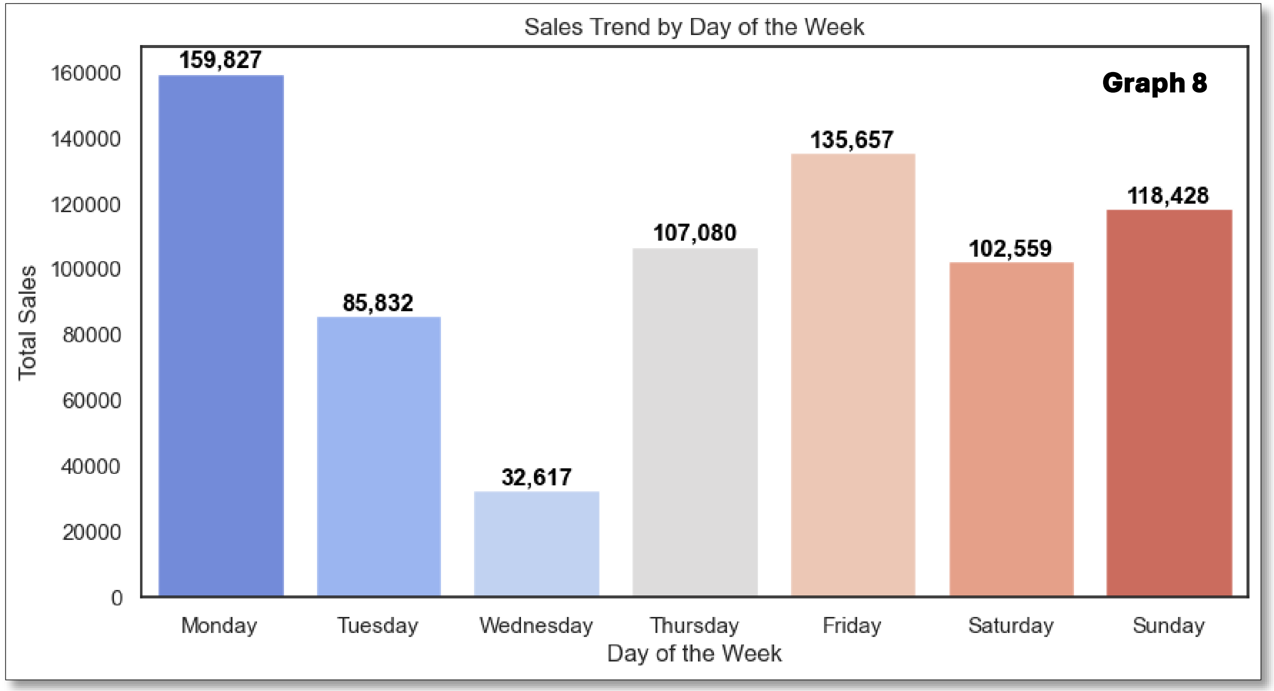
Graph 8: Sales Trend by Day of the Week

Description:

This bar chart shows the total sales for each day of the week.

Key Insight:

- Highest Sales on Monday.
- Lowest Sales on Wednesday.
- **Midweek Recovery :** Thursday and Friday show a recovery in sales, suggesting increased customer engagement toward the end of the week.
- **Weekend Performance :** Saturday and Sunday maintain relatively high sales, showing consistent demand throughout the weekend.



Methodology

"Exploratory Data Analysis"

Graph 9: Sales Seasonality - Monthly Sales Trend

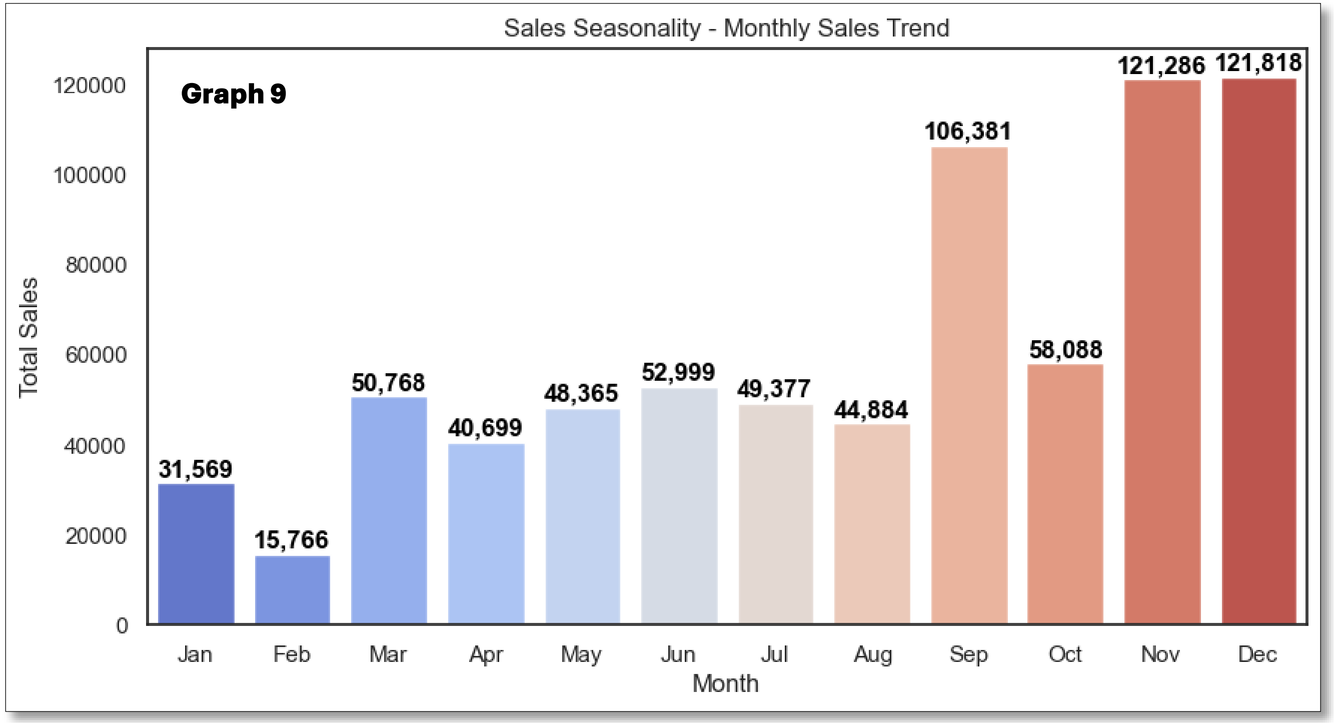
Description:

This bar chart shows the total sales for each month of the year, highlighting seasonal trends in sales performance.

Key Insight:

- **Seasonal Patterns :** Sales see a low point in January and February after the holidays, then steadily increase from March to September, reaching their highest point in November and December due to holiday shopping.

- **Holiday Impact :** The significant spike in November and December reflects strong consumer spending during the holiday season.



Methodology

“Model Development”

After completing data preprocessing and exploratory analysis, we developed and evaluated several machine learning models to predict weekly sales, our objective was to identify the most accurate and generalizable model to help optimize inventory, staffing, and marketing strategies.

Models Explored:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor
- XGBoost with Log-Transformed Target

Methodology

"Model Development"

1.Linear Regression:

- **Target:** Weekly Sales Data
- **Processing:**
  - Train-Test split: 80% – 20%
  - Standardized features

Interpretation:

○ Model explains ~42.7% of the variance in test data.

○ Good start but shows underfitting – fails to capture complex, non-linear trends.

Performance Metrics	
RMSLE	1.00
RMSE	179.2
MSE	32,111.44
MAE	129.65

R² Score	
Training	0.0478
Testing	0.427

Methodology

"Model Development"

2. Decision Tree Regressor:

- **Target:** Weekly Sales Data
- **Processing:**
  - Random state set to ensure reproducibility.

Interpretation:

○ Better performance than Linear Regression.
○ Captures more complex relationships.
○ High train R^2 indicates possible overfitting.

Performance Metrics	
RMSLE	0.62
RMSE	165.25
MSE	27,307.89
MAE	101.26

R^2 Score	
Training	**1.000**
Testing	0.513

Methodology

“Model Development”

3. Random Forest:

- **Hyperparameters:**
  - n_estimators = 100
  - min_samples_leaf = 5
  - max_depth = None

Interpretation:

○ Strong generalization (less overfitting than Decision Tree).

○ RMSE is only ~52.7% of average weekly sales (241.83).

Performance Metrics	
RMSLE	0.54
RMSE	127.42
MSE	16,235.72
MAE	79.87

R^2 Score	
Training	0.88
Testing	0.71

Methodology

“Model Development”

4. XGBoost:

- **Hyperparameters:**
  - n_estimators = 100
  - learning_rate = 0.1
  - max_depth = 6

Performance Metrics	
RMSLE	0.55
RMSE	128.84
MSE	16,599.59
MAE	80.31

Interpretation:

- ○ Excellent training fit and strong generalization.
- ○ RMSE is ~53.3% of average weekly sales.
- ○ Fast training with great accuracy

R² Score	
Training	0.97
Testing	0.70

Methodology

"Model Development"

5. XGBoost with Log-Transformed Target :

Why Log Transformation?
- Sales data is skewed with extreme high values.
- Log transformation (log1p) stabilizes variance and reduces impact of outliers.

Interpretation:

○ Highest test R^2 across all models.
○ RMSLE is the lowest overall.
○ Indicates improved generalization and smoother predictions.

Performance Metrics	
RMSLE	0.48
RMSE	132.2
MSE	17,477.38
MAE	77.32

R^2 Score	
Training	0.97
Testing	0.87

Methodology

"Cross Validation"

Interpretation:

- XGBoost (Log Target) achieved the lowest RMSLE (0.464) and MAE (69.5), making it the most reliable for minimizing relative and absolute errors.

- XGBoost with Log Target is the best choice for deployment as it balances accuracy and error robustness effectively.

Model	RMSLE	RMSE	MSE	MAE
Linear Regression	0.999	178.697	31932.487	130.277
Decision Tree	0.619	158.045	24978.293	94.348
Random Forest	0.515	119.429	14263.239	75.771
XGBoost	0.552	116.891	13663.398	73.344
XGBoost (Log Target)	0.464	118.420	14023.196	69.469

Methodology

“Model Deployment”

Deployment Strategy:

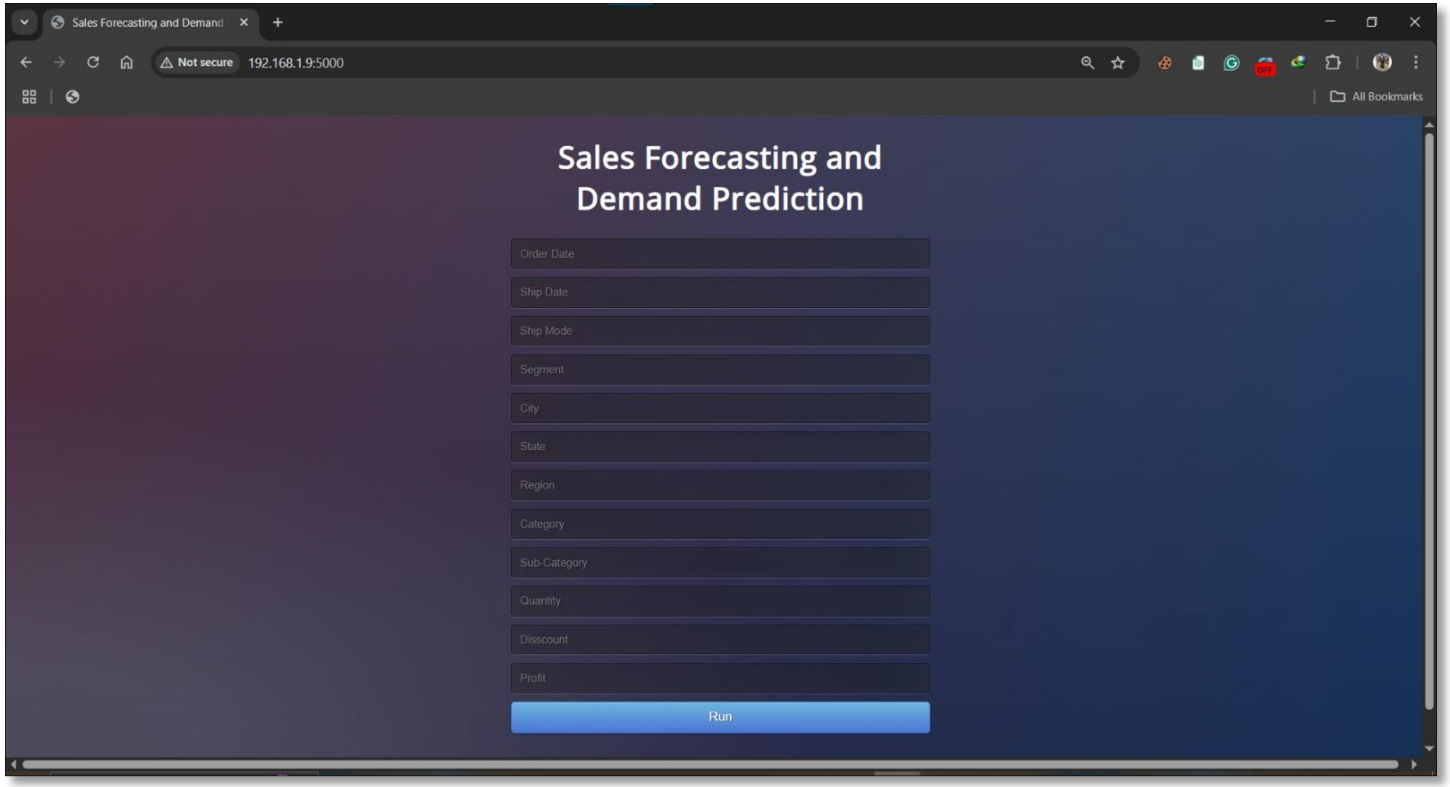
- ○ Developed a Flask-based web application to host the forecasting model.
- ○ Designed with HTML and CSS for a responsive and user-friendly interface.
- ○ The user interface allows input of 12 key features like order date, ship mode, category, and profit for real-time sales predictions.
- ○ Output displays the expected weekly sales based on the provided inputs.

Methodology

“Model Deployment”

Step 1:

- Prepare all the data (mentioned in the input cells) needed by the model to predict your sales properly.
- Make sure that all the data is accurate.

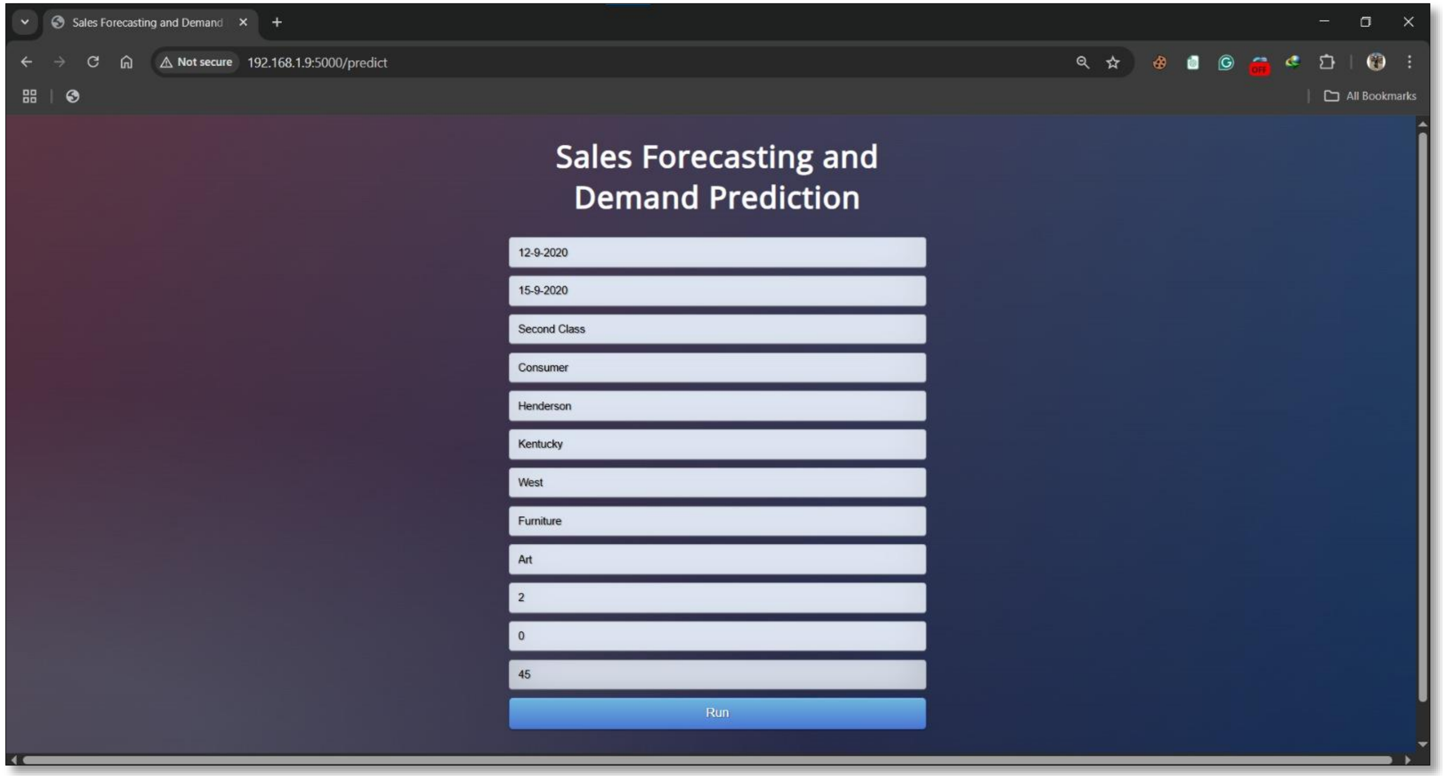


Methodology

“Model Deployment”

Step 2:

- Fill in the cells.
- Review your inputs.
- Press the run button.



Sales Forecasting and Demand Prediction

12-9-2020

15-9-2020

Second Class

Consumer

Henderson

Kentucky

West

Furniture

Art

2

0

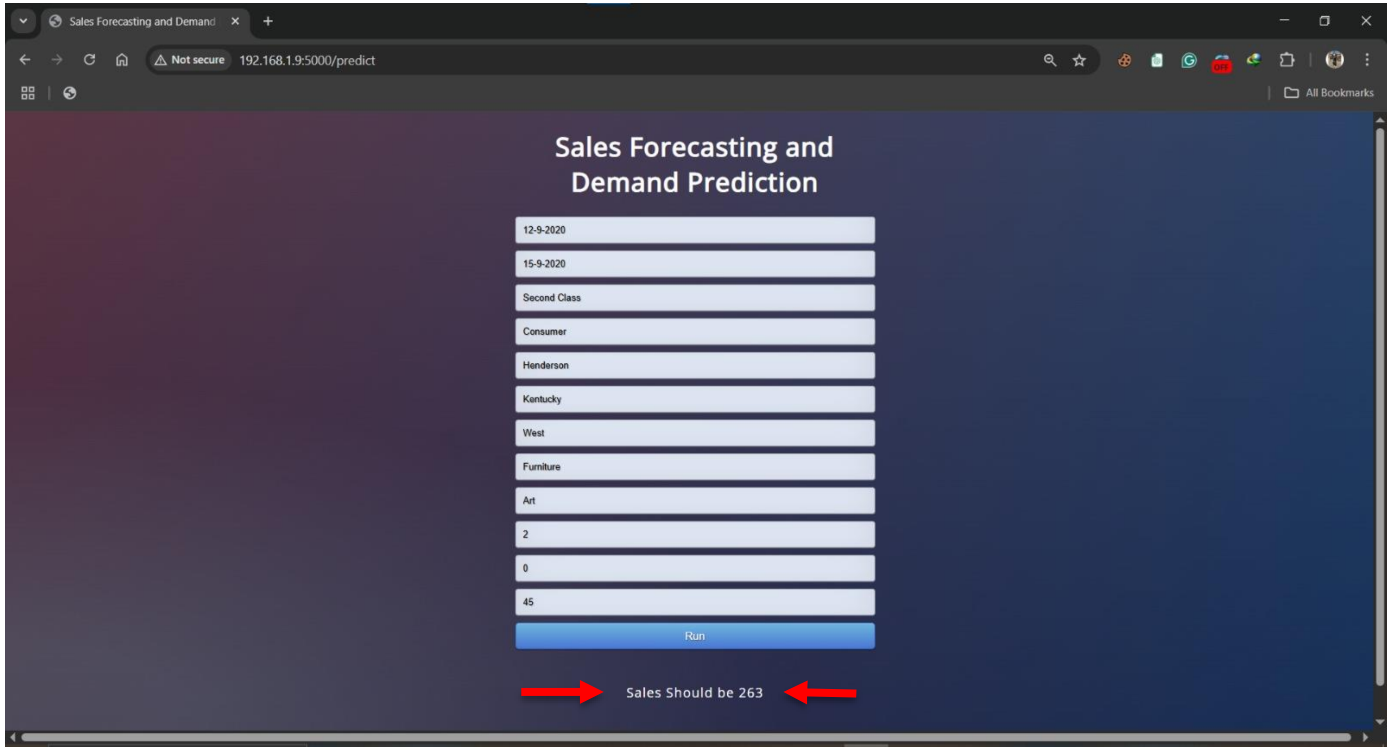
45

Methodology

"Model Deployment"

Step 3:

- ○ The predicted weekly sales value will be displayed.
- ○ Use it to enhance your future business strategy.



Sales Forecasting and Demand Prediction

12-9-2020

15-9-2020

Second Class

Consumer

Henderson

Kentucky

West

Furniture

Art

2

0

45

Run

Sales Should be 263

Conclusion

Key Achievements:

- ○ Developed a robust machine learning model (XGBoost with log transformation) for accurate sales forecasting, achieving an R^2 score of 0.87 on test data.
- ○ Identified critical seasonal trends, customer behaviors, and high-performing segments through comprehensive EDA.
- ○ Engineered features like time-based variables and lag features to enhance model performance.
- ○ Deployed a user-friendly Flask API to enable real-time demand predictions for business decision-making.

Business Impact:

- ○ Reduces stockouts and overstocking by aligning inventory with predicted demand.
- ○ Improves operational efficiency through data-driven planning for marketing, staffing, and supply chain.

Tools & Technologies Used

kaggle

pandas

NumPy

CSV

jupyter

scikit learn

Flask

matplotlib

seaborn

Thank you

Presented by:

Abdallah Gasem

Youssef Ali

Mazen Karam

Micheal Emad

Abdallah Sayed

Mostafa Ahmed

Group:

GIZ2_AIS4_S9

Contact:

☎ (+20) 102 534 7679 ✉ ag.ellsayed@gmail.com