رواد مصر الرقمية

Sales Forecasting and Optimization

DEPI Graduation Project

Meet our team

- Mariam Hassan
- Rawan Sotohy
- Sara Abdelrahman
- Toka Khaled

Under supervision: Eng. Mahmoud Khorshid

View The Full Code Notebooks From Here

View The Project Web Application From Here

Table of contents

01

Introduction

Data Description

The dataset contains 9,800 rows and 18 columns describing customer details, order information, product categories, and financial transactions. Key variables include:

- Customer demographics (Segment, City, State, Region)
- Product details (Category, Sub-Category, Product Name)
- Order information (Order Date, Ship Date, Ship Mode)
- Target variable: Sales amount (in USD)

Sample Data

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City
1	CA-2017-15215	8/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
2	CA-2017-15215	8/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson
3	CA-2017-13868	12/6/2017	16/06/2017	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles
4	US-2016-10896	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale
5	US-2016-10896	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale
6	CA-2015-11581	9/6/2015	14/06/2015	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
7	CA-2015-11581	9/6/2015	14/06/2015	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
8	CA-2015-11581	9/6/2015	14/06/2015	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
9	CA-2015-11581	9/6/2015	14/06/2015	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles
10	CA-2015-11581	9/6/2015	14/06/2015	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles

State	Postal Coc	Region	Product ID	Category	Sub-Category	Product Name	Sales
Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96
Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacki	731.94
California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Type	14.62
Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectang	957.5775
Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.368
California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions Wood and Plastic	48.86
California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell 322	7.28
California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5320 IP Phone VoIP phone	907.152
California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View Binders with Locking	18.504
California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C206VTEL 6 Outlet Surge	114.9

Data Dictionary

Column Name	Description	Datatype
Row ID	Unique numeric identifier for each row in the dataset	int64
Order ID	Unique identifier for each order.	object
Order Date	The date when the order was placed.	object
Ship Date	The date when the order was shipped.	object
Ship Mode	Shipping method used (e.g., Second Class, Standard Class).	object
Customer ID	Unique identifier for each customer.	object
Customer Name	Name of the customer who placed the order.	object
Segment	Customer segment (e.g., Consumer, Corporate, Home Office).	object
Country	Country of the customer.	object

Data Dictionary

Column Name	Description	Datatype
City	City of the customer.	object
State	State of the customer.	object
Postal Code	Postal/ZIP code (may contain missing values).	float64
Region	Geographic region of the customer (e.g., West, South).	object
Product ID	Unique identifier for each product.	object
Category	Unique identifier for each customer.	object
Sub-Category	Sub-category of the product (e.g., Chairs, Labels).	object
Product Name	Name of the product.	object
Sales	**Target** — Sale amount (in USD) for the product line.	float64

Data Exploration

We began by examining the dataset structure:

- Checked basic statistics (mean, min, max values)
- Identified data types and missing values
- Analyzed unique values in categorical columns
- Visualized distributions of key variables

Key findings:

- The dataset had minimal missing values (only in Postal Code column)
- Significant outliers were present in the Sales variable
- High cardinality in Product Name and City columns

Data Preprocessing

Handling Missing Values

- Dropped rows with missing Postal Codes (only 11 rows affected)

Outlier Treatment

- Used IQR method to cap outliers in the Sales variable
- Visualized distribution before and after treatment

Data Preprocessing

Encoding Categorical Variables

- Applied Label Encoding to all categorical columns including:
  - Ship Mode (4 categories)
  - Segment (3 categories)
  - Region (4 categories)
  - Product Category/Sub-category
  - Geographic locations (City, State)

Data Preprocessing

Feature Engineering

- Date Features:
  - Extracted Year, Month, Day, and Weekday from Order Date
  - Calculated Shipping Duration (days between order and shipment)
- Text Processing:
  - Standardized Product Names by converting to lowercase and removing punctuation

Data Preprocessing

Feature Selection

We dropped unnecessary columns that wouldn't contribute to modeling:

- Row ID, Customer ID, Product ID (unique identifiers)
- Customer Name (personal information)
- Order ID (transaction identifier)
- Ship Date (redundant since we have shipping duration)

Correlation

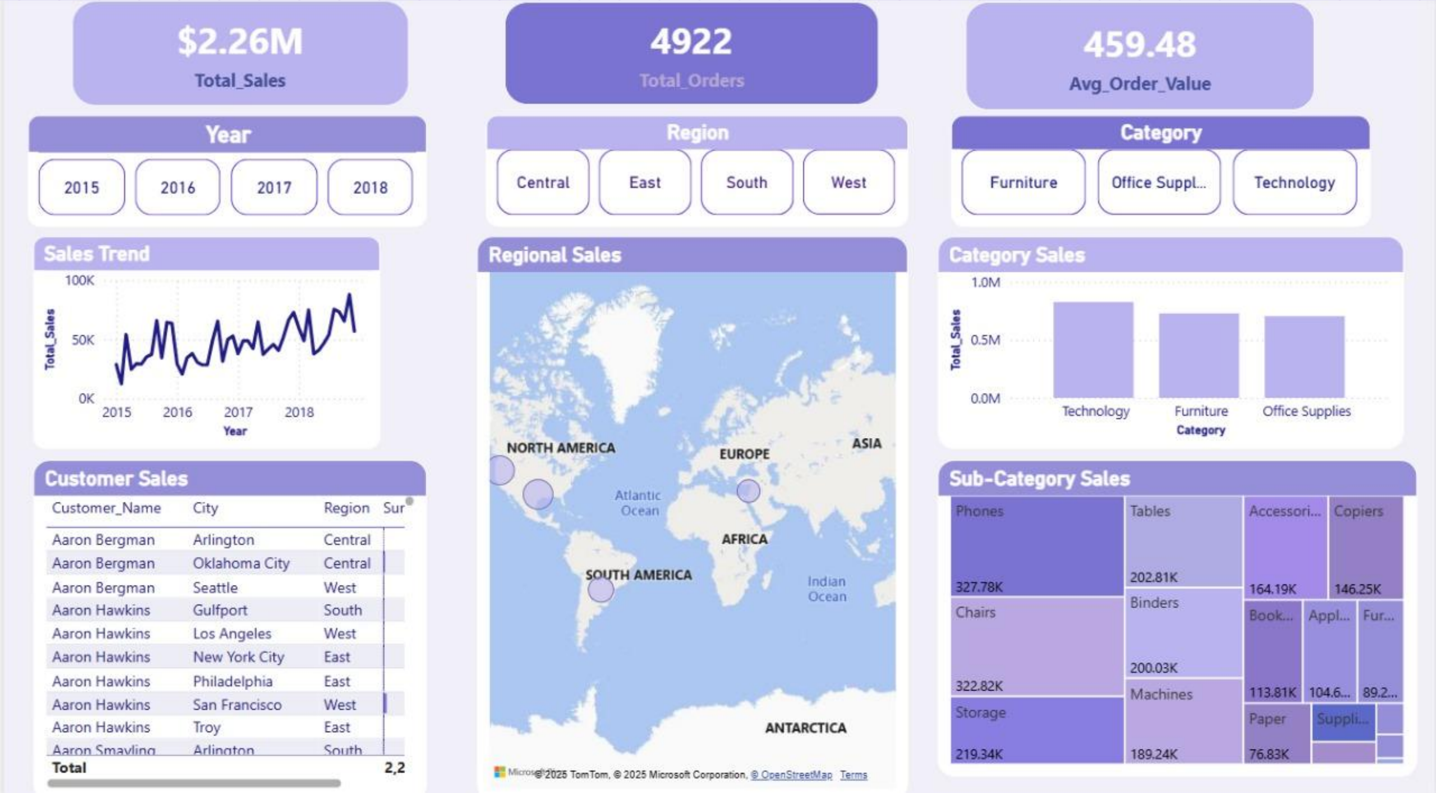
- We calculated the correlation matrix to understand the linear relationships between numerical features and the target variable (**Sales**).

02

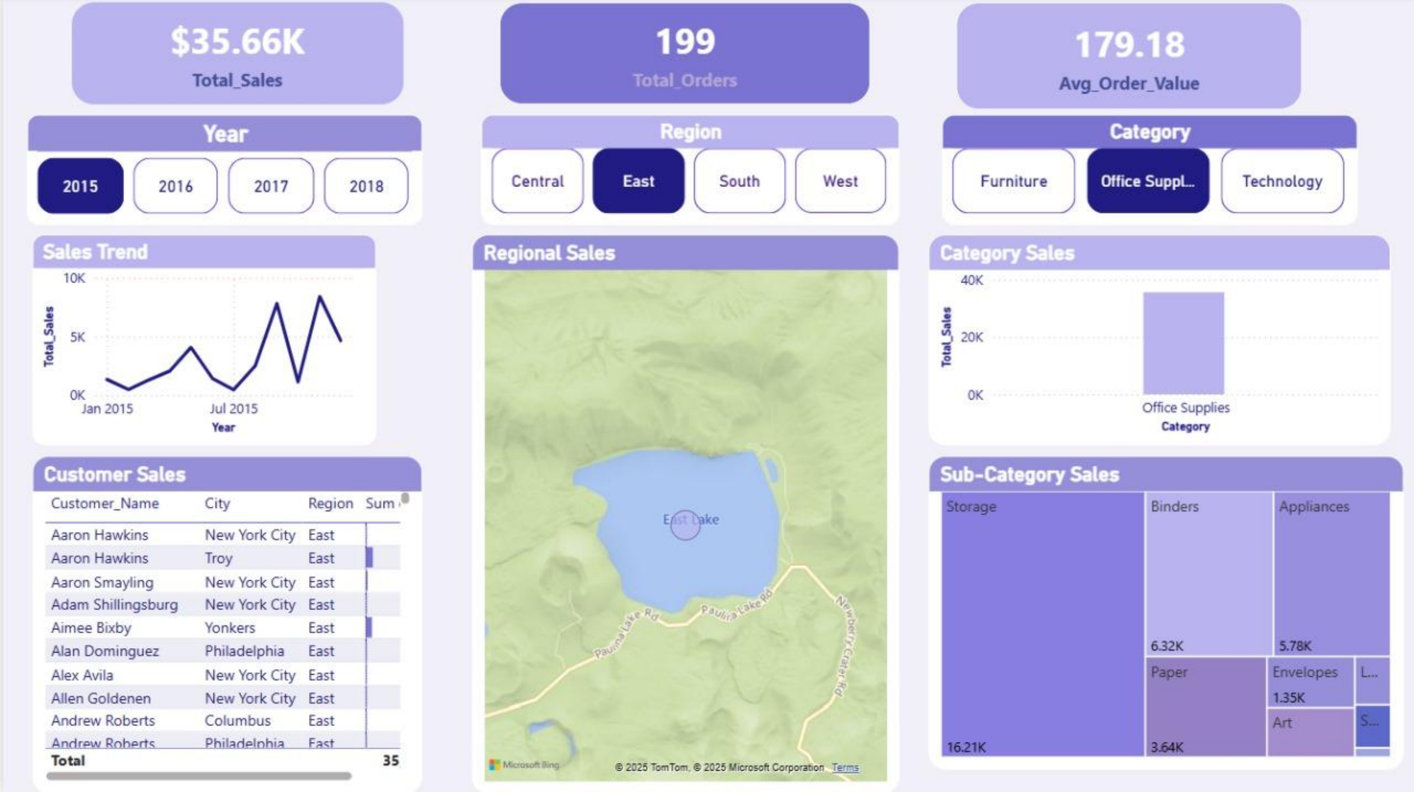
Dashboard

View the Dashboard from here

Dashboard

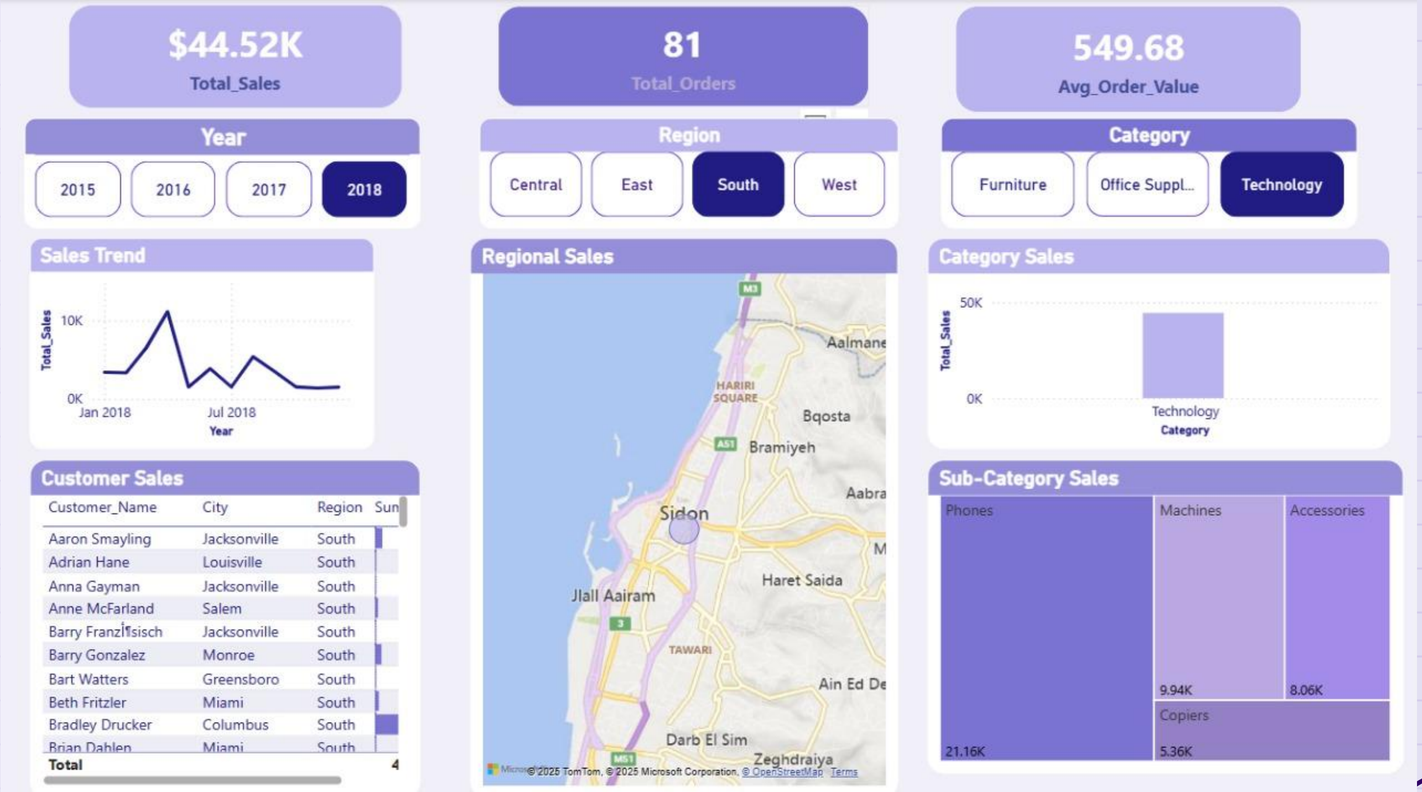

$2.26M
Total_Sales
4922
Total_Orders
459.48
Avg_Order_Value
Year
2015
2016
2017
2018
Region
Central
East
South
West
Category
Furniture
Office Suppl...
Technology
Sales Trend
100K
50K
0K
Total_Sales
2015
2016
2017
2018
Year
Regional Sales
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
ANTARCTICA
© 2025 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms
Category Sales
1.0M
0.5M
0.0M
Total_Sales
Technology
Furniture
Office Supplies
Category
Customer Sales
Customer_Name
City
Region
Sur
Aaron Bergman Arlington Central
Aaron Bergman Oklahoma City Central
Aaron Bergman Seattle West
Aaron Hawkins Gulfport South
Aaron Hawkins Los Angeles West
Aaron Hawkins New York City East
Aaron Hawkins Philadelphia East
Aaron Hawkins San Francisco West
Aaron Hawkins Troy East
Aaron Smayling Arlington South
Total 2,2
Sub-Category Sales
Phones 327.78K
Tables 202.81K
Accessori... 164.19K
Copiers 146.25K
Chairs 322.82K
Binders 200.03K
Book... 113.81K
Appl... 104.6...
Fur... 89.2...
Storage 219.34K
Machines 189.24K
Paper 76.83K
Suppli...

Dashboard



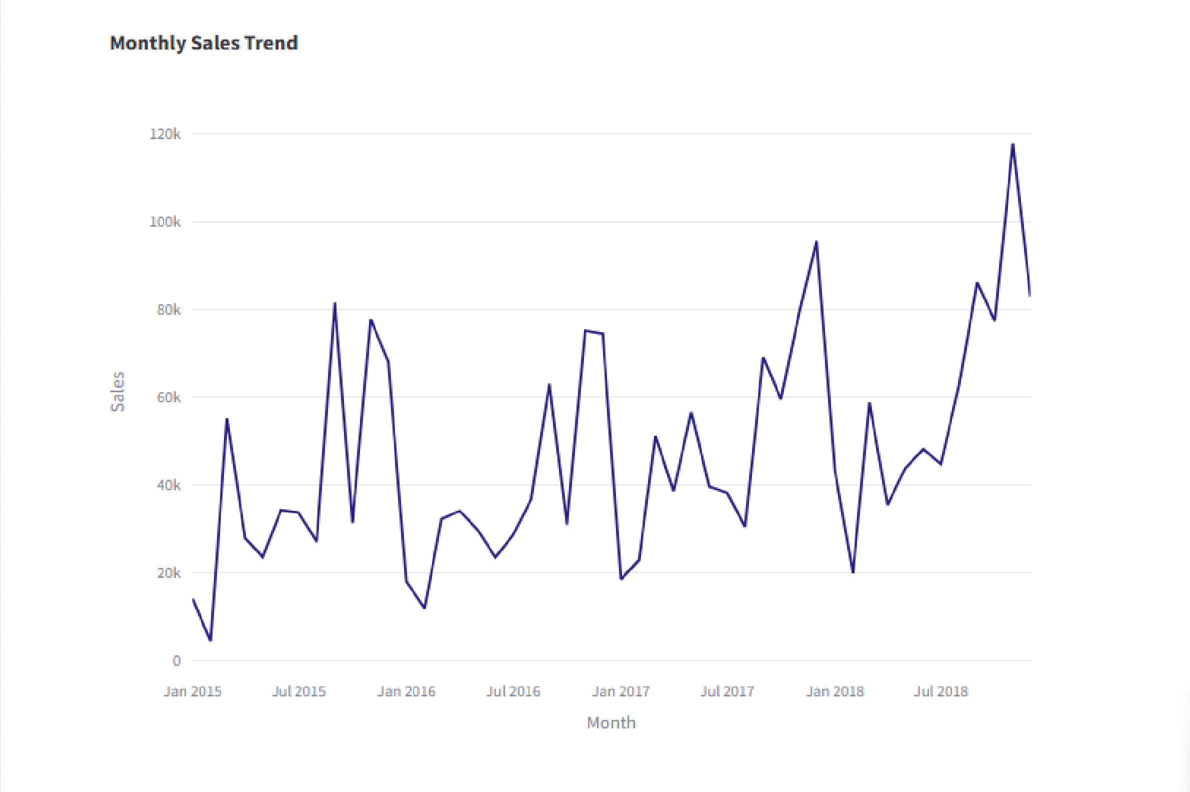
$35.66K
Total_Sales
199
Total_Orders
179.18
Avg_Order_Value
Year
2015
2016
2017
2018
Region
Central
East
South
West
Category
Furniture
Office Suppl...
Technology
Sales Trend
10K
5K
0K
Total_Sales
Jan 2015
Jul 2015
Year
Regional Sales
East Lake
Paulina Lake Rd
Newberry Crater Rd
© 2025 TomTom, © 2025 Microsoft Corporation Terms
Category Sales
40K
20K
0K
Total_Sales
Office Supplies
Category
Customer Sales
Customer_Name
City
Region
Sum
Aaron Hawkins
New York City
East
Aaron Hawkins
Troy
East
Aaron Smayling
New York City
East
Adam Shillingsburg
New York City
East
Aimee Bixby
Yonkers
East
Alan Dominguez
Philadelphia
East
Alex Avila
New York City
East
Allen Goldenen
New York City
East
Andrew Roberts
Columbus
East
Andrew Roberts
Philadelphia
East
Total
35
Sub-Category Sales
Storage
16.21K
Binders
6.32K
Appliances
5.78K
Paper
3.64K
Envelopes
1.35K
Art
L...
S...

Dashboard



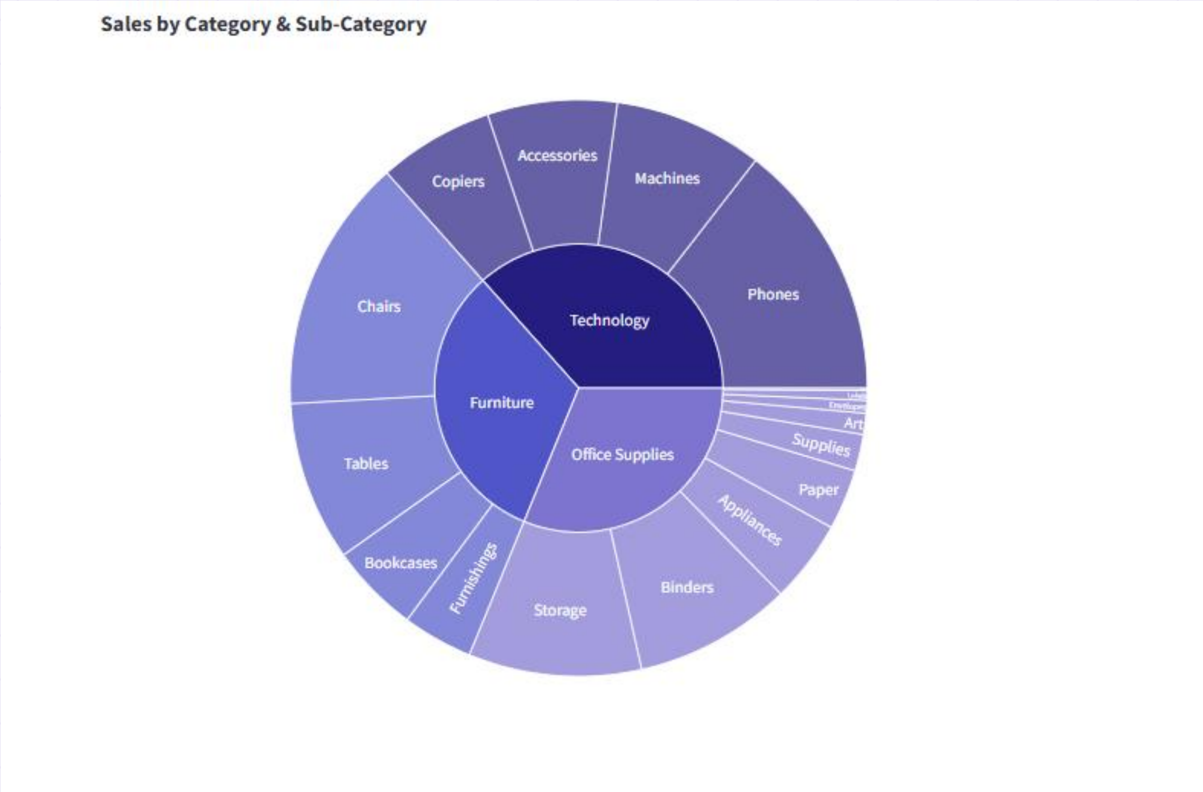
$44.52K
Total_Sales
81
Total_Orders
549.68
Avg_Order_Value
Year
2015
2016
2017
2018
Region
Central
East
South
West
Category
Furniture
Office Suppl...
Technology
Sales Trend
10K
0K
Total_Sales
Jan 2018
Jul 2018
Year
Regional Sales
Aalmane
HARIRI SQUARE
Bqosta
Bramiyeh
Aabra
Sidon
Haret Saida
Jlall Aairam
TAWARI
Ain Ed De
Darb El Sim
Zeghdraiya
© 2025 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms
Category Sales
50K
0K
Total_Sales
Technology
Category
Customer Sales
Customer_Name City Region Sum
Aaron Smayling Jacksonville South
Adrian Hane Louisville South
Anna Gayman Jacksonville South
Anne McFarland Salem South
Barry FranzÌ¶sisch Jacksonville South
Barry Gonzalez Monroe South
Bart Watters Greensboro South
Beth Fritzler Miami South
Bradley Drucker Columbus South
Brian Dahlen Miami South
Total 4
Sub-Category Sales
Phones
21.16K
Machines
9.94K
Accessories
8.06K
Copiers
5.36K

Dashboard



Monthly Sales Trend
120k
100k
80k
60k
40k
20k
0
Sales
Jan 2015
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Jul 2018
Month

Dashboard


Sales by Category & Sub-Category
Copiers
Accessories
Machines
Phones
Chairs
Technology
Furniture
Office Supplies
Tables
Art
Supplies
Paper
Appliances
Bookcases
Furnishings
Storage
Binders

03

Prediction

View the Prediction form from here

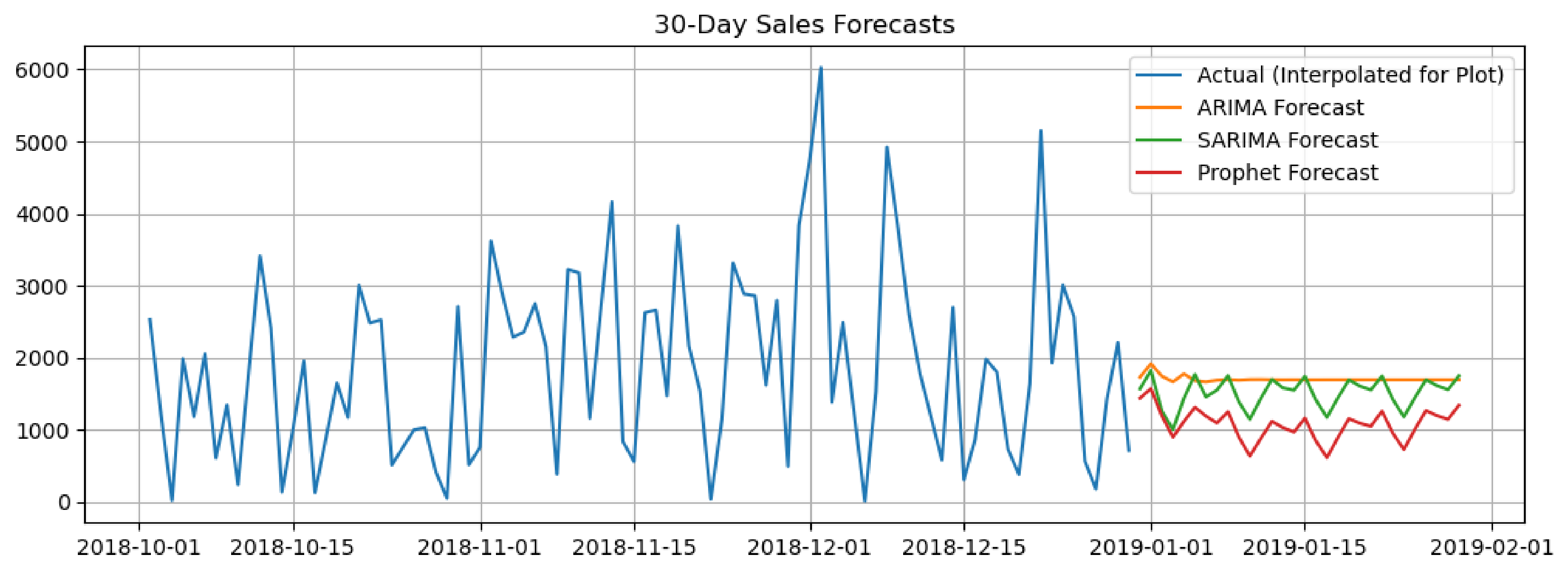
Forecasting Models

Objective:

- Predict future sales using three different models: ARIMA, SARIMA and Prophet.

Steps:

- Data Preparation:
  - Group sales by date (Order Date).
  - Handle missing values with interpolation.
- Models Used:
  - ARIMA (5,1,2): Captures trends and basic patterns.
  - SARIMA (1,1,1)(1,1,1,7): Adds weekly seasonality.
  - Prophet: Automatically detects trends and seasonality.
- Forecast Horizon: 30 days ahead.



30-Day Sales Forecasts
Actual (Interpolated for Plot)
ARIMA Forecast
SARIMA Forecast
Prophet Forecast
6000
5000
4000
3000
2000
1000
0
2018-10-01
2018-10-15
2018-11-01
2018-11-15
2018-12-01
2018-12-15
2019-01-01
2019-01-15
2019-02-01

Ensemble Boosting Models

Objective:

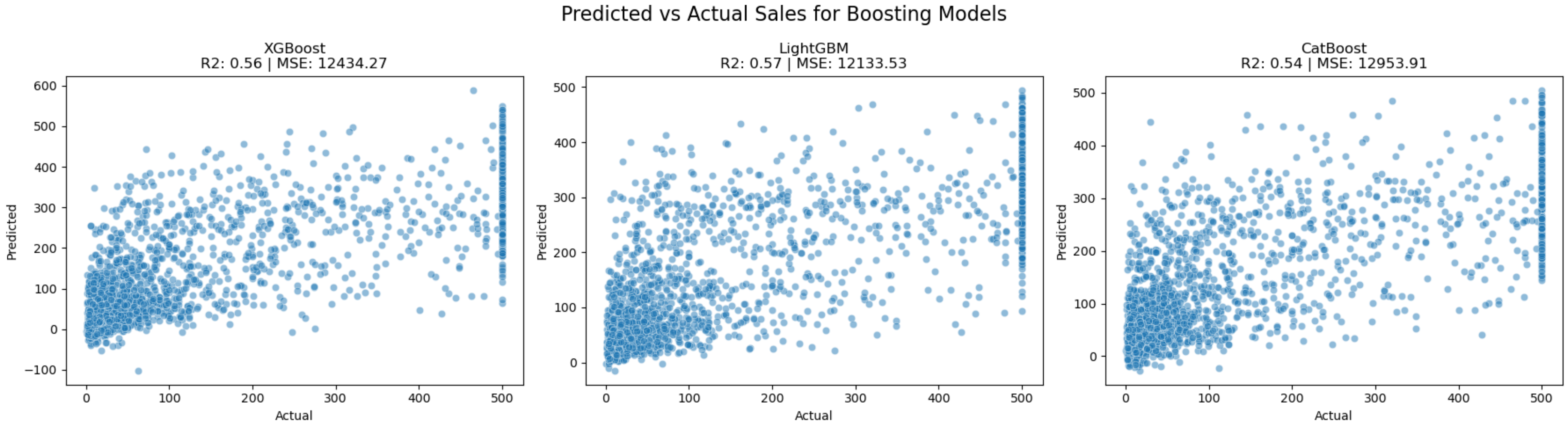
- Predict sales using ensemble boosting models (XGBoost, LightGBM, CatBoost) and optimize performance via hyperparameter tuning.

Steps:

- Data Preparation:
  - Split into features (X) and target (y = Sales).
  - Train-test split (80-20) with scaling (StandardScaler).
- Models Used:
  - XGBoost: High flexibility, handles complex patterns.
  - LightGBM: Faster training, good for large datasets.
  - CatBoost: Robust to categorical features, minimal preprocessing.
- Evaluation Metrics:
  - MSE (Mean Squared Error): Lower = Better.
  - R^2 Score: Closer to 1 = Better fit.

Model Performance & Tuning

- **Initial Results (R^2 Scores):**
  - XGBoost = 0.56.
  - LightGBM = 0.57.
  - CatBoost = 0.54.
- **Hyperparameter Tuning:**
  - Used RandomizedSearchCV to optimize:
    - n_estimators, learning_rate, max_depth (XGBoost).
    - num_leaves (LightGBM).
    - depth (CatBoost).
- **Tuning Impact (R^2 Scores):**
  - XGBoost = 0.571.
  - LightGBM = 0.570.
  - CatBoost = 0.528.


Predicted vs Actual Sales for Boosting Models
XGBoost
R2: 0.56 | MSE: 12434.27
LightGBM
R2: 0.57 | MSE: 12133.53
CatBoost
R2: 0.54 | MSE: 12953.91
Predicted
Actual
−100
0
100
200
300
400
500
600

Bagging Models

Objective:

- Predict future sales using ensemble (Random Forest) and single-tree (Decision Tree) approaches.

Models Used:

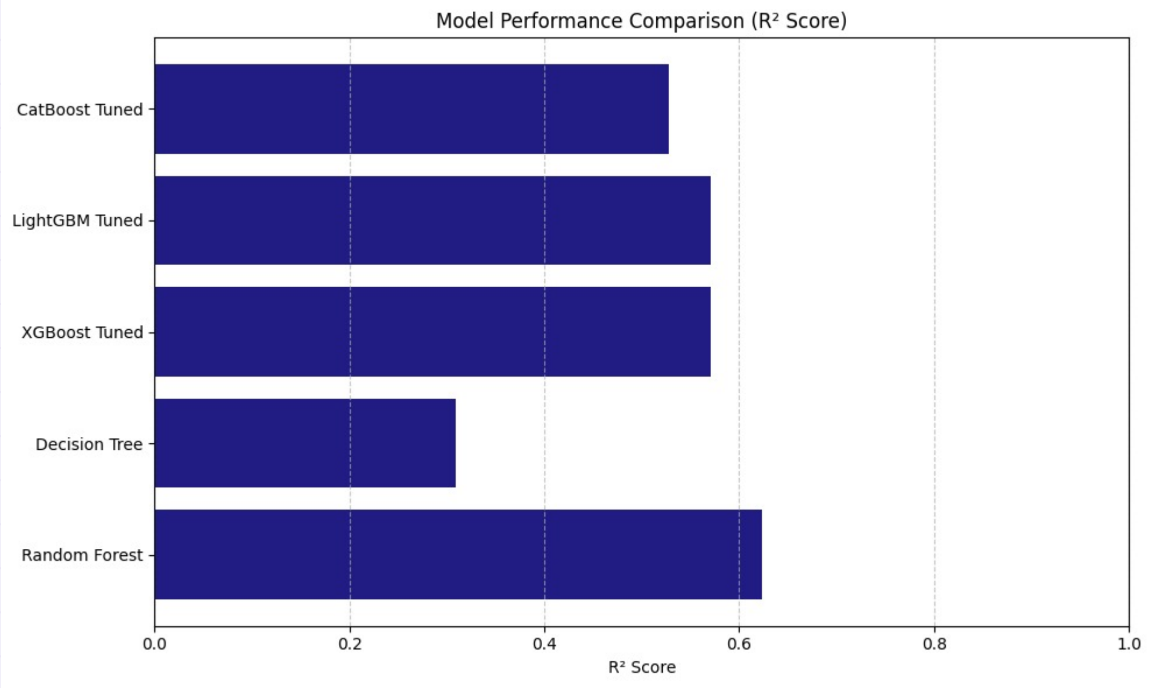
- Decision Tree:
  - Simple splits (max_depth=5, random_state=42).
  - Prone to overfitting but interpretable.
- Random Forest:
  - Ensemble of 100 trees (n_estimators=100).
  - Robust to overfitting, handles non-linearity.

Performance (R^2 Scores):

- Decision Tree = 0.31.
- Random Forest = 0.62.

Best Model:

Random Forest was chosen as the final model based on performance.



04

Deployment

View The Project Web Application From Here

Chosen Model:

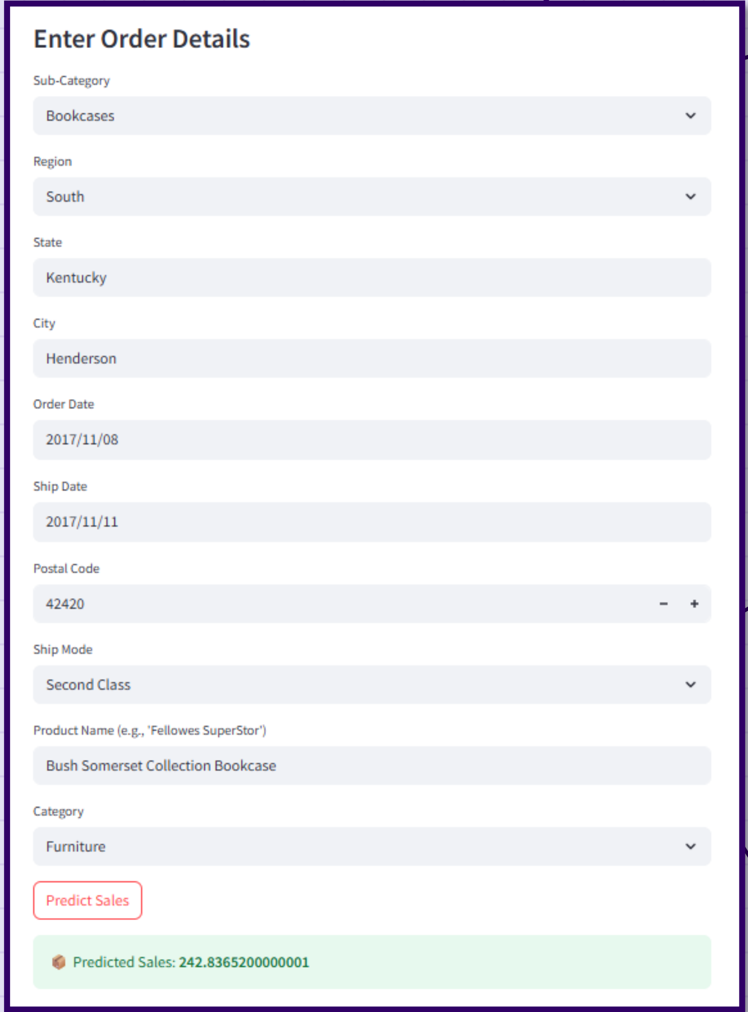
- After evaluation, Random Forest was selected for deployment due to its:
  - ✓ Lowest MSE (best R^2 score)
  - ✓ Fast prediction speed
  - ✓ Handling of complex patterns

Deployment Steps:

1. Save Model: Used joblib to serialize and save the trained model.
2. Build Web App: Created an interactive Streamlit application.

User-Friendly Interface:

- **Input Order Details**: Users can enter features (e.g., category, order date, region).
- **Real-Time Prediction**: Instantly displays predicted **Sales** value.



Business Impact

- **Better Decision-Making:** Improves planning across inventory, marketing, and operations using accurate sales forecasts.
- **Revenue Growth:** Increases sales by aligning stock and promotions with customer demand.
- **Reduced Operational Costs:** Lowers costs by avoiding overstocking, stockouts, and inefficient logistics.
- **Improved Customer Satisfaction:** Ensures product availability, leading to a better customer experience and higher loyalty.

Thanks!

Do you have any questions?

View The Project Web Application From Here

Scan Me