SALES FORECASTING FOR INVENTORY OPTIMIZATION

Y.Sri laxmi
A.Samanvitha
K.Jahnavi

Problem Statement

To enhance inventory management, this project builds a time series model to forecast daily sales for the next 90 days by analyzing historical data, including order dates and product details, to optimize stock levels and improve operational efficiency.

Objectives

01 Data Preparation

02 Exploratory Data Analysis (EDA)

03 Feature Engineering

04 Model Selection

05 Model Building and Training

06 Model Evaluation

07 Forecasting

08 Visualization of results

Dataset

The dataset has multiple rows, where each row corresponds to a unique sales transaction over time

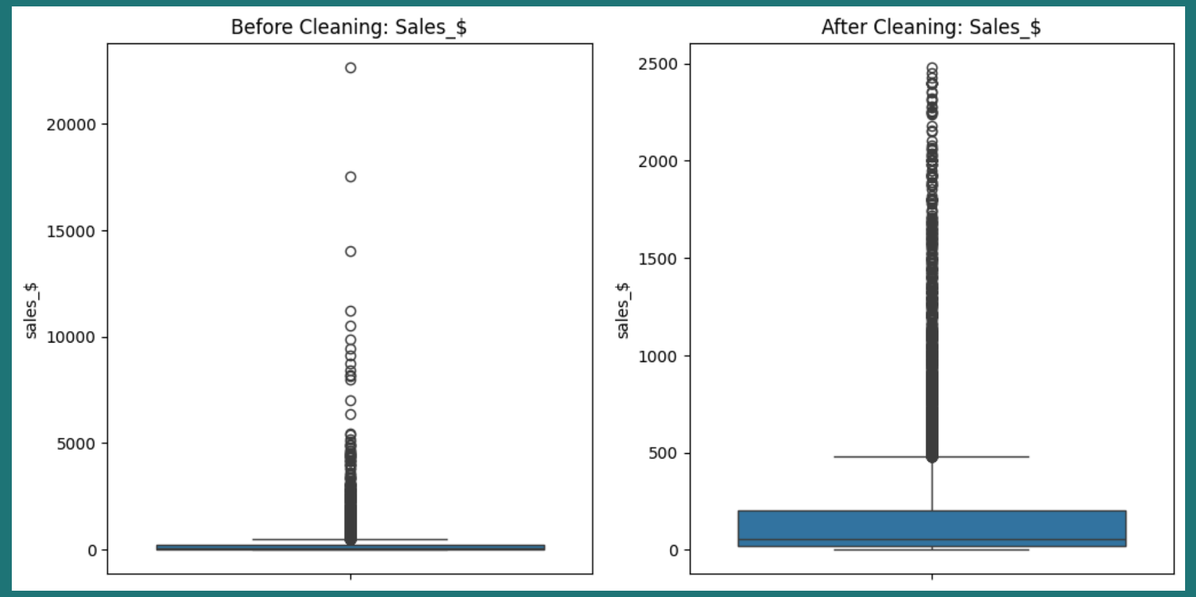
- Number of Rows: 9,800
- Number of Columns: 15

The columns represent key attributes like:

- Order ID, Order Date, Ship Date
- Ship Mode, Customer ID
- Location: Country, City, State, Postal Code, Region
- Product ID, Category, Sub-Category, Product Name
- Sales ($)

order_id	order_dat	ship_date	ship_mod	customer_	country	city	state	postal_co	region	product_ic	category	sub_categ	product_n	sales_$
CA-2017-1	########	########	Second Cl	CG-12520	United Sta	Hendersor	Kentucky	42420	South	FUR-BO-1(	Furniture	Bookcases	Bush Some	261.96
CA-2017-1	########	########	Second Cl	CG-12520	United Sta	Hendersor	Kentucky	42420	South	FUR-CH-1(	Furniture	Chairs	Hon Delux	731.94
CA-2017-1	########	16/06/17	Second Cl	DV-13045	United Sta	Los Angele	California	90036	West	OFF-LA-10	Office Sup	Labels	Self-Adhes	14.62
US-2016-1	########	18/10/16	Standard C	SO-20335	United Sta	Fort Laude	Florida	33311	South	FUR-TA-1(	Furniture	Tables	Bretford C	957.5775
US-2016-1	########	18/10/16	Standard C	SO-20335	United Sta	Fort Laude	Florida	33311	South	OFF-ST-10	Office Sup	Storage	Eldon Fold	22.368
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	FUR-FU-1(	Furniture	Furnishing	Eldon Expr	48.86
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	OFF-AR-1(	Office Sup	Art	Newell 32	7.28
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	TEC-PH-1(	Technolog	Phones	Mitel 5320	907.152
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	OFF-BI-10(	Office Sup	Binders	DXL Angle-	18.504
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	OFF-AP-10	Office Sup	Appliances	Belkin F5C	114.9
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	FUR-TA-1(	Furniture	Tables	Chromcraf	1706.184
CA-2015-1	9/6/2015	14/06/15	Standard C	BH-11710	United Sta	Los Angele	California	90032	West	TEC-PH-1(	Technolog	Phones	Konftel 25	911.424
CA-2018-1	15/04/18	20/04/18	Standard C	AA-10480	United Sta	Concord	North Carc	28027	South	OFF-PA-10	Office Sup	Paper	Xerox 196	15.552
CA-2017-1	########	########	Standard C	IM-15070	United Sta	Seattle	Washingtc	98103	West	OFF-BI-10(	Office Sup	Binders	Fellowes P	407.976
US-2016-1	22/11/16	26/11/16	Standard C	HP-14815	United Sta	Fort Worth	Texas	76106	Central	OFF-AP-10	Office Sup	Appliances	Holmes Re	68.81
US-2016-1	22/11/16	26/11/16	Standard C	HP-14815	United Sta	Fort Worth	Texas	76106	Central	OFF-BI-10(	Office Sup	Binders	Storex Dur	2.544
CA-2015-1	########	18/11/15	Standard C	PK-19075	United Sta	Madison	Wisconsin	53711	Central	OFF-ST-10	Office Sup	Storage	Stur-D-Sto	665.88
CA-2015-1	13/05/15	15/05/15	Second Cl	AG-10270	United Sta	West Jord	Utah	84084	West	OFF-ST-10	Office Sup	Storage	Fellowes S	55.5

Data Preparation



- **Handling Null Values**

The null values in the postal_code column are replaced with the most frequent value (mode) from the same column.

- **Date Format Conversion and Validation**

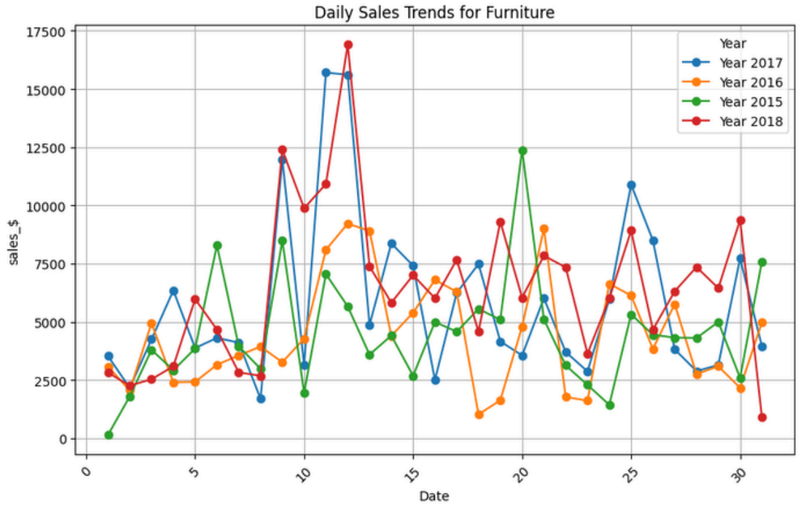
This process ensures that the date columns (order_date and ship_date) are in a consistent format for accurate time series analysis.

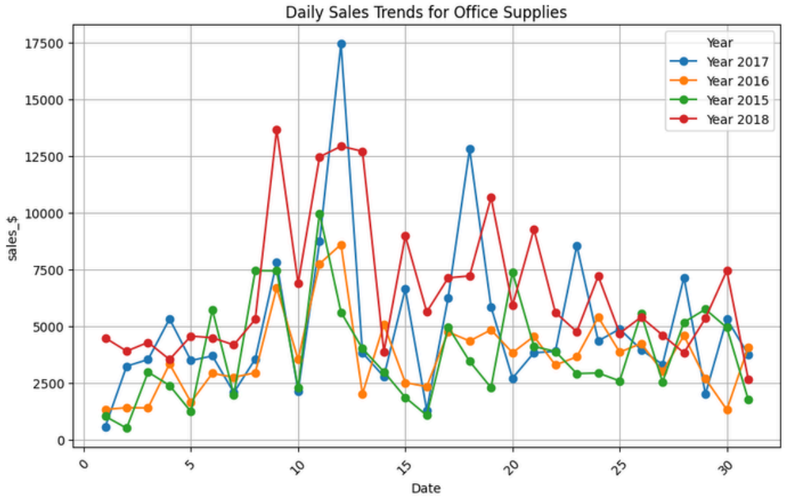
- **Handling Outliers**

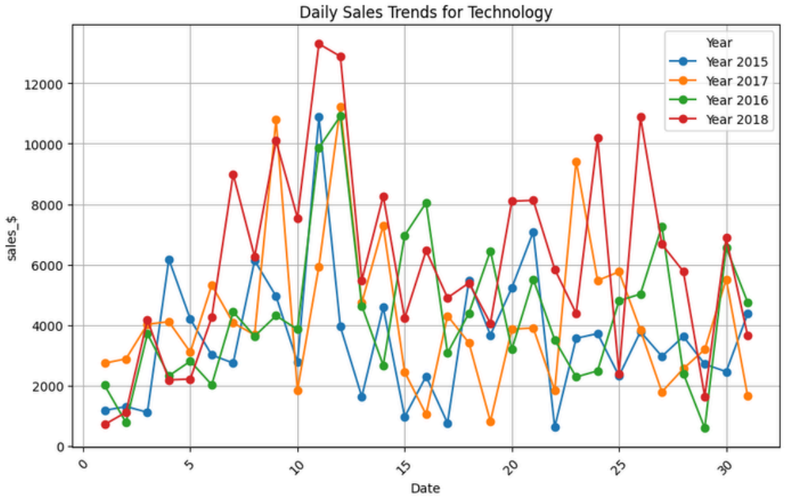
The 99th percentile of the sales_$ column was calculated to identify extreme outliers, and the dataset was filtered to remove values above this threshold.

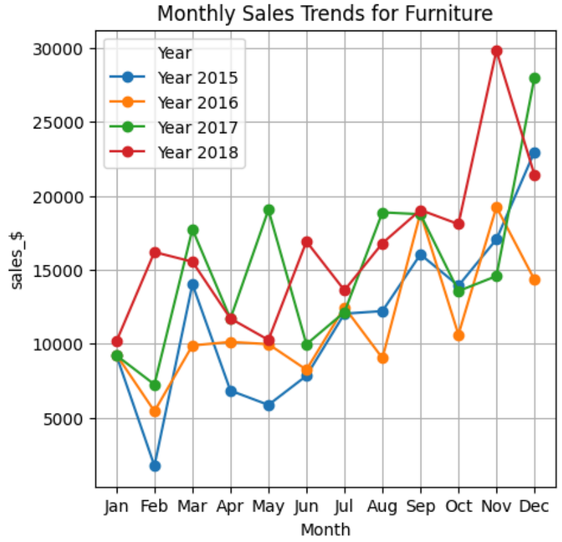
Exploratory Data Analysis

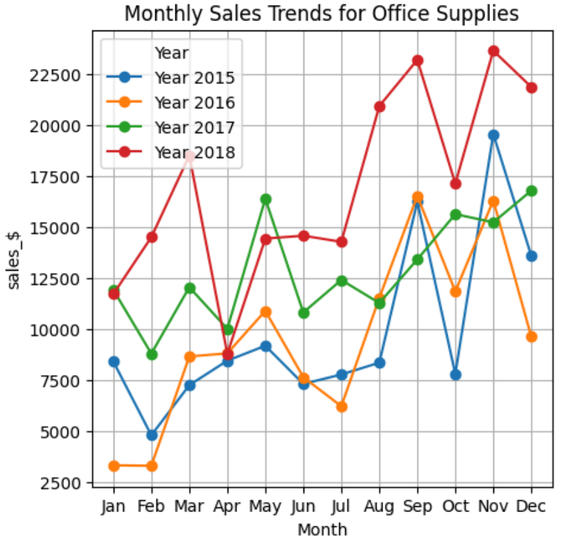
Time Series Analysis and Visualization

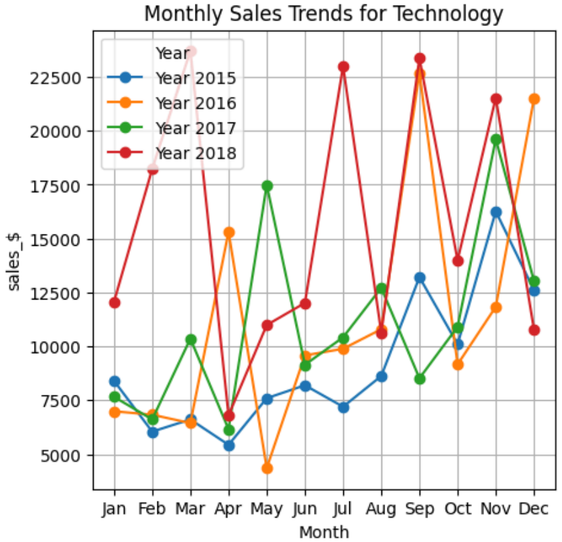




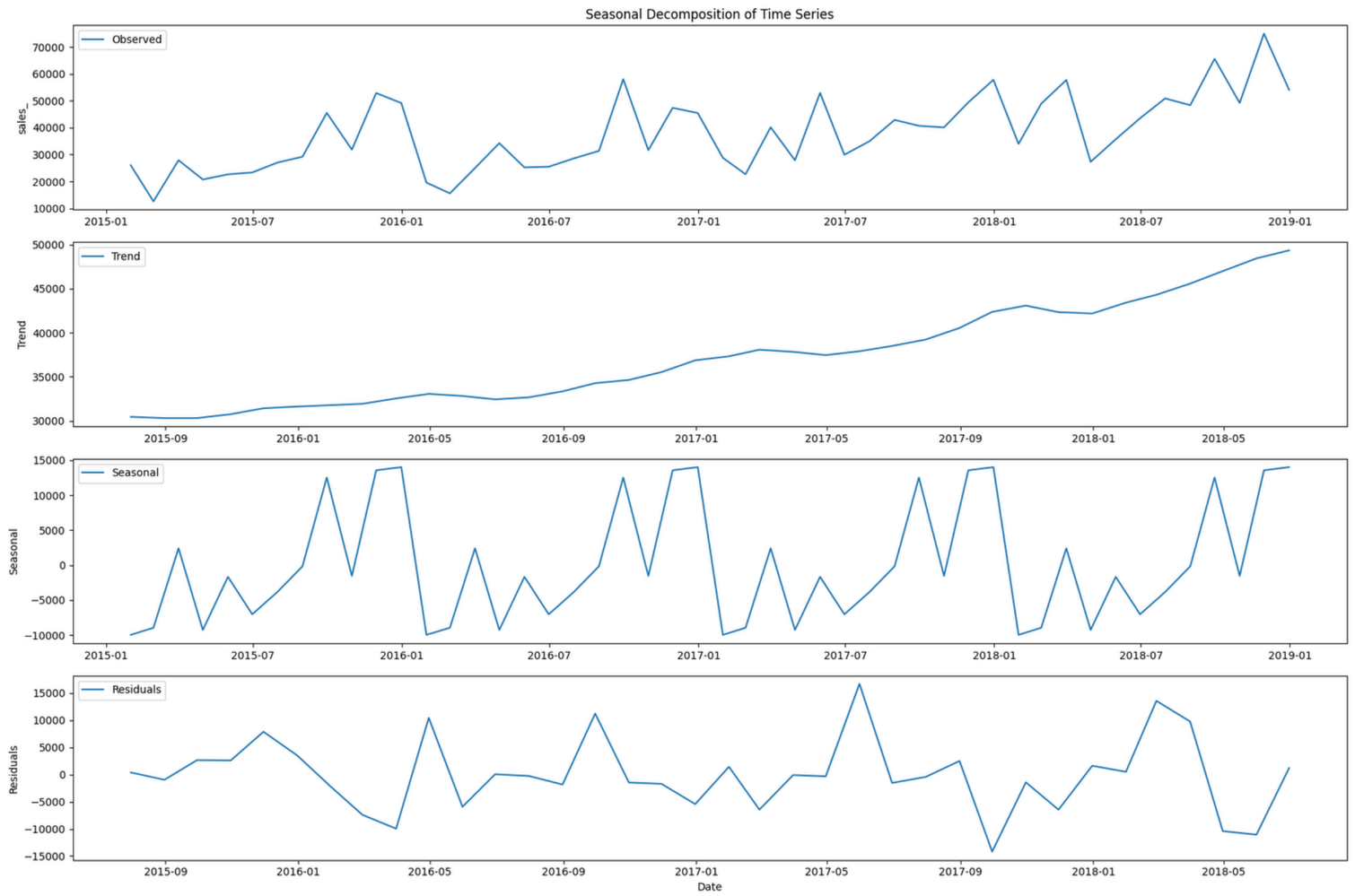








Seasonal Decomposition



Feature Engineering

Aggregating Sales Data

Grouping and aggregating sales data by order_date, category, and sub_category to calculate the total sales amount for performance analysis.

Adding Additional Features

Additional features were introduced :

- Lagged Features: Captures trends
- Moving Averages: Smooths fluctuations
- Exponential Moving Averages (EMA): Highlights recent data
- Day/Month: Captures seasonality
- Day of the Week: Identifies weekly patterns
- Holidays: Identifies Holidays

Feature Imporatance

Analyzed correlations with sales_$, and applied PCA to identify top 10 and top 5 features based on explained variance.

Model Selection

- XGBoost: Selected for capturing complex patterns in sales data across various product categories and enhancing prediction accuracy for inventory optimization.
- SARIMAX: Chosen for modeling seasonality, trends, and external factors in sales time series.
- ARIMA: Applied for capturing autocorrelation and patterns in univariate time series forecasting.

Model Building and Training

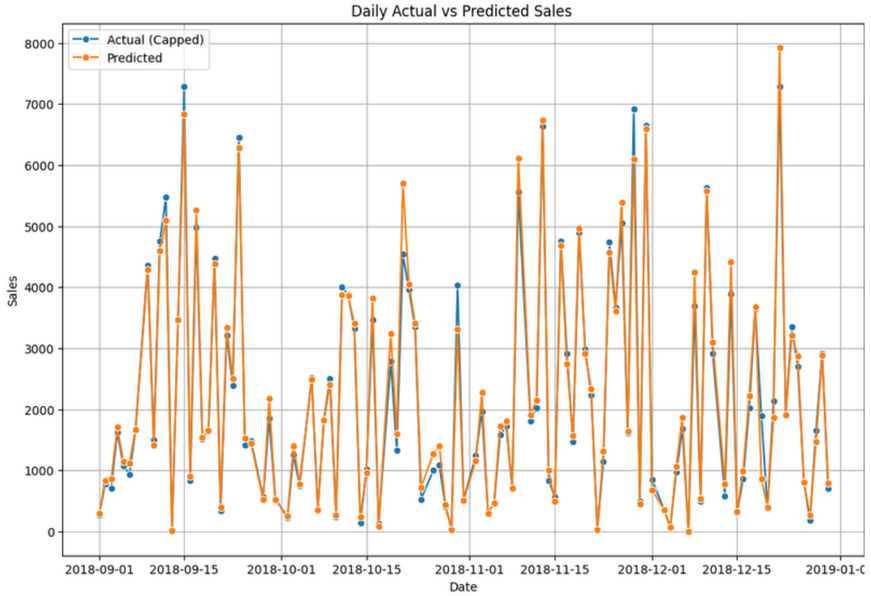
- **XGBoost chosen:** Outperformed other models in accuracy and performance.
- **Combined Dataset:** Built and trained XGBoost using all product categories together.
- **Individual Categories:** Trained separate XGBoost models for Furniture, Technology, and Office Supplies datasets.

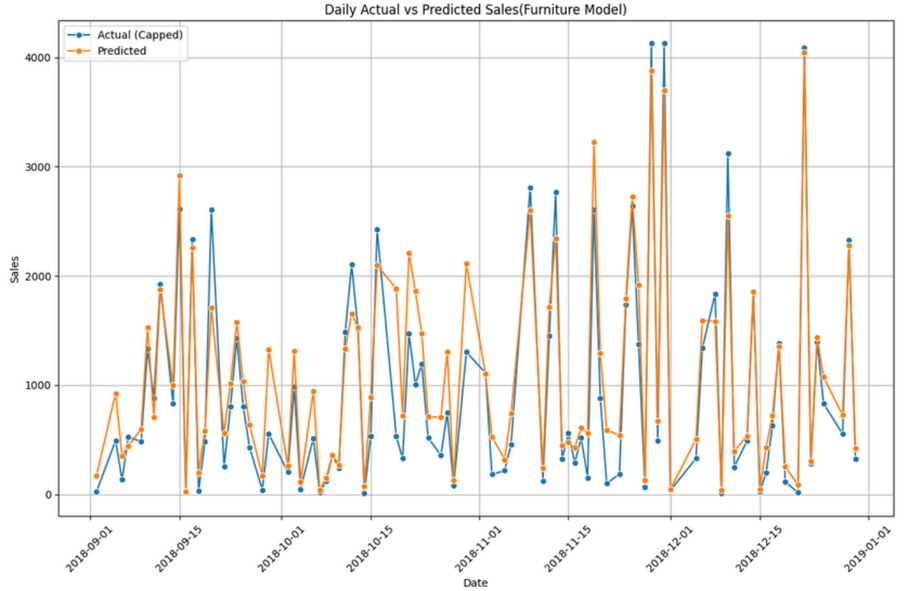
Model Performance

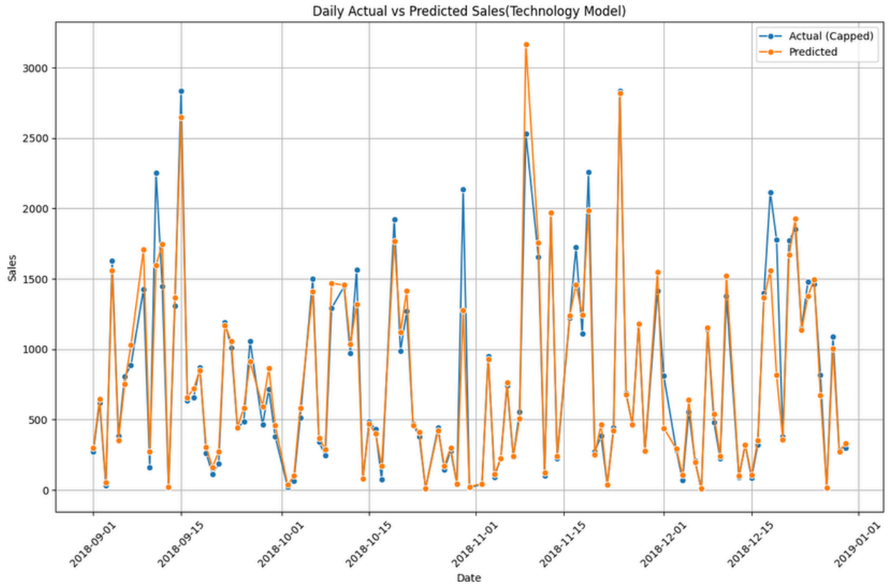
Model	MAE	RMSE
XGBoost	**28.366850992566146**	**76.89681749570681**
Sarimax	190.45344374038325	317.3626057549045
Arima	190.76281641626966	315.29964326715964

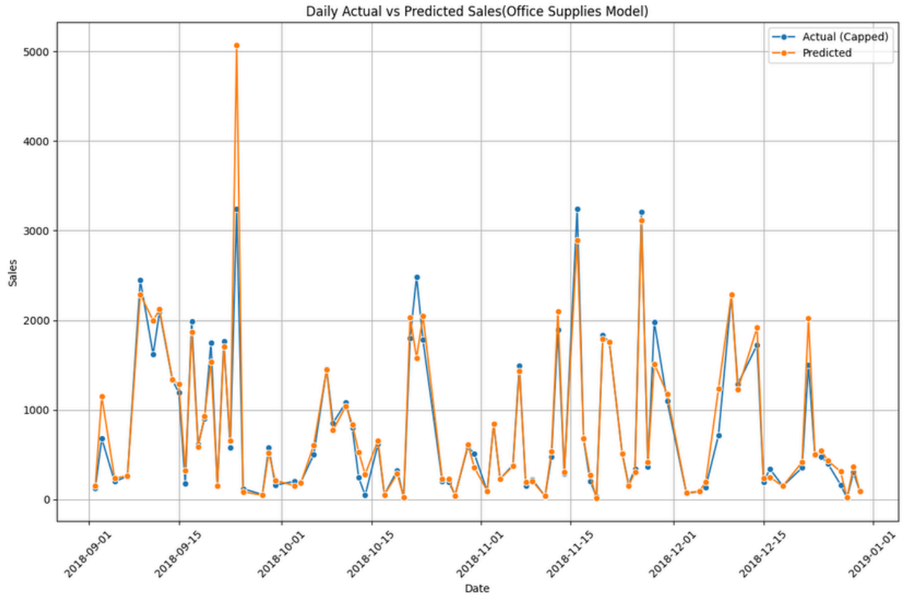
Category	MAE	RMSE
Furniture DataSet	111.13370338230732	198.96108909268915
Technology DataSet	20.457743122823832	70.91133306538153
Office Supplies DataSet	63.14648277231663	128.686106375402

Visualizations for Actual Vs Predicted Sales









Forecasting future sales by category for the next 90 days

- Iterative Forecasting: Begins with the most recent test data.
- Daily Predictions: Model forecasts sales using lag variables, moving averages, and day-of-week indicators.
- Update Data: After each prediction, the forecasted value is added to the dataset.
- Feature Recalculation: Sales lags and moving averages are updated for the next day's prediction.
- 90-Day Forecast: Process continues for 90 days, generating a sequential sales forecast.

Future Scope

- Making Predictions Iteratively.
- Experiment with advanced models or ensemble techniques.
- Enhancing outlier management
- Expand predictions to sub-categories for greater precision.

Thank You