Retail Store Sales Forecasting
Final Report

Capstone 2 project
by Denis Grekov

Table of Contents

Final report for Capstone 2 “Retail Store Sales Forecasting”

Chapter 0: Introduction

Project Overview

In this project, we aim to build an accurate sales forecasting model for a retail store. Accurate sales forecasts are essential for effective inventory management, financial planning, and strategic decision-making. By predicting future sales, we can optimize stock levels, reduce holding costs, and improve customer satisfaction.

Dataset Description

We have a rich dataset of historical sales data from the retail store, which includes:

	Invoice_id	Cust_id	Txn_date	Product_Sub_category	Product_Category	Qty	Rate	Tax	Total_Amt	Store_type
0	25890929042	266783	2016-09-23	Women	Footwear	4	1321	554.820	5838.820	e-Shop
1	25890929042	266783	2016-09-24	Women	Footwear	-4	-1321	554.820	-5838.820	e-Shop
2	98477711300	266783	2017-10-21	Mens	Clothing	3	93	29.295	308.295	TeleShop
3	8410316370	266783	2018-02-20	Mens	Clothing	1	869	91.245	960.245	e-Shop
4	16999552161	266783	2018-09-02	Non-fiction	Books	2	835	175.350	1845.350	e-Shop

1. Invoice ID: Unique identifier for each transaction.
2. Customer ID: Unique identifier for each customer.
3. Product Subcategory: Specific subcategory of the product sold.
4. Quantity: Number of units sold in each transaction.
5. Rate: Selling price per unit.
6. Tax: Tax applied to the transaction.
7. Total Amount: Total transaction amount including tax.
8. Store Type: Type of store (e.g., e-Shop, TeleShop, Brick and Mortar).
9. Product Category: General category of the product (e.g., Books, Clothing, Electronics).

Potential Challenges and Roadblocks

Data Quality: Missing values, outliers, and errors need to be addressed.
Seasonality and Trends: Identifying and accounting for seasonal patterns and trends.
Return Transactions: Handling returns to avoid skewed sales figures.

Feature Engineering: Creating meaningful features from raw data.
Model Selection: Choosing the best model from various forecasting techniques.

Primary Goal

Develop a reliable sales forecasting model to accurately predict future sales, involving steps like data cleaning, exploratory data analysis, feature engineering, model building, and evaluation.

Methods

Data Cleaning: Handle return transactions, outliers, and missing values.
Exploratory Data Analysis (EDA): Uncover key trends and insights.
Feature Engineering: Normalize and encode variables, scale features, create new features.
Model Building: Train various forecasting models (e.g., Prophet, baseline models).
Model Evaluation: Use MAE, RMSE, and MAPE to select the best model.

Chapter 1: Data Cleaning

Methodology

Handling Return Transactions: Separate return transactions to prevent skewed results.
Outlier Detection: Use the Interquartile Range (IQR) method to identify and manage outliers.

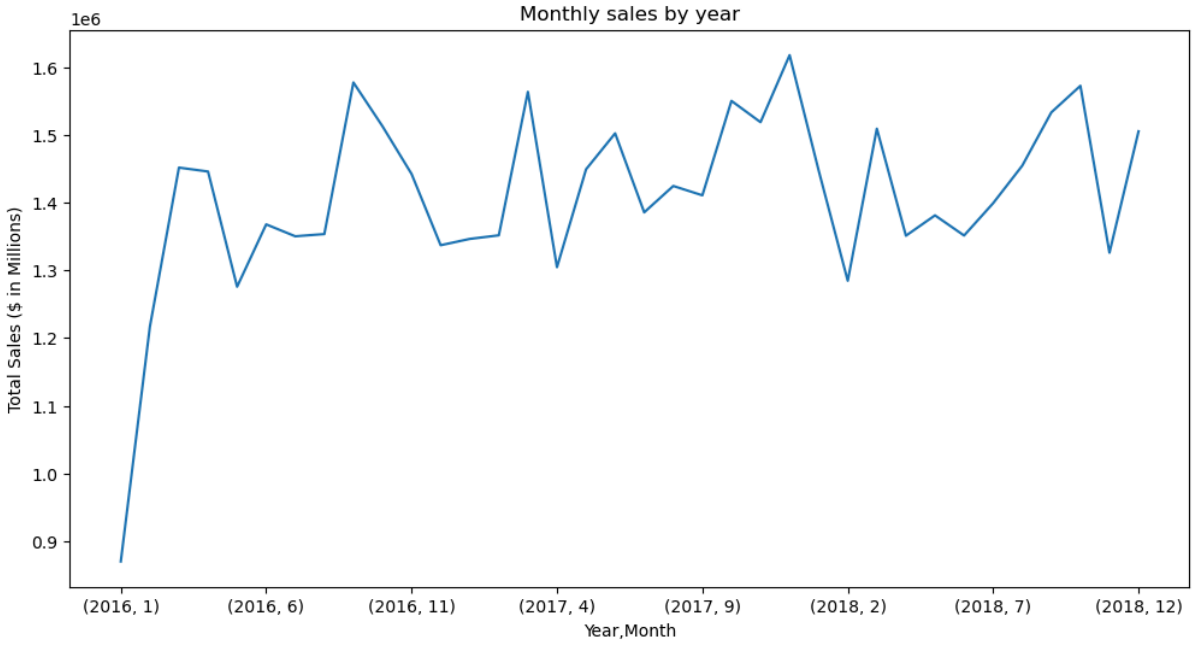
Conclusions

A solid data cleaning process ensures the integrity and accuracy of the data for subsequent analysis and modeling.

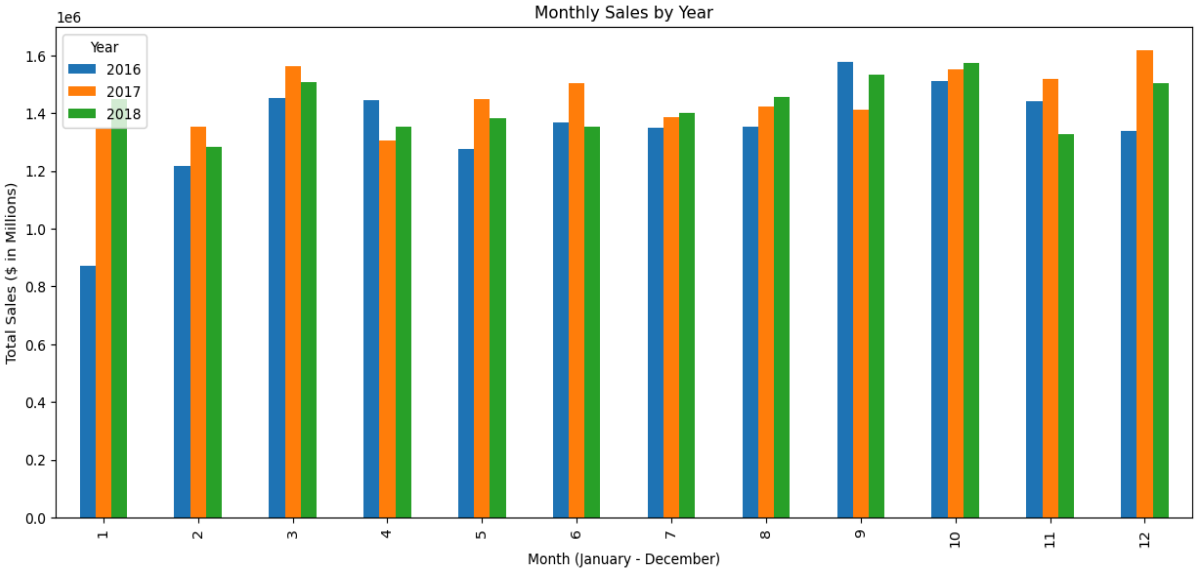
Chapter 2: Data Analysis

Methodology

Monthly Sales Over Time: This analysis helps identify seasonal trends by examining monthly sales data over several years.

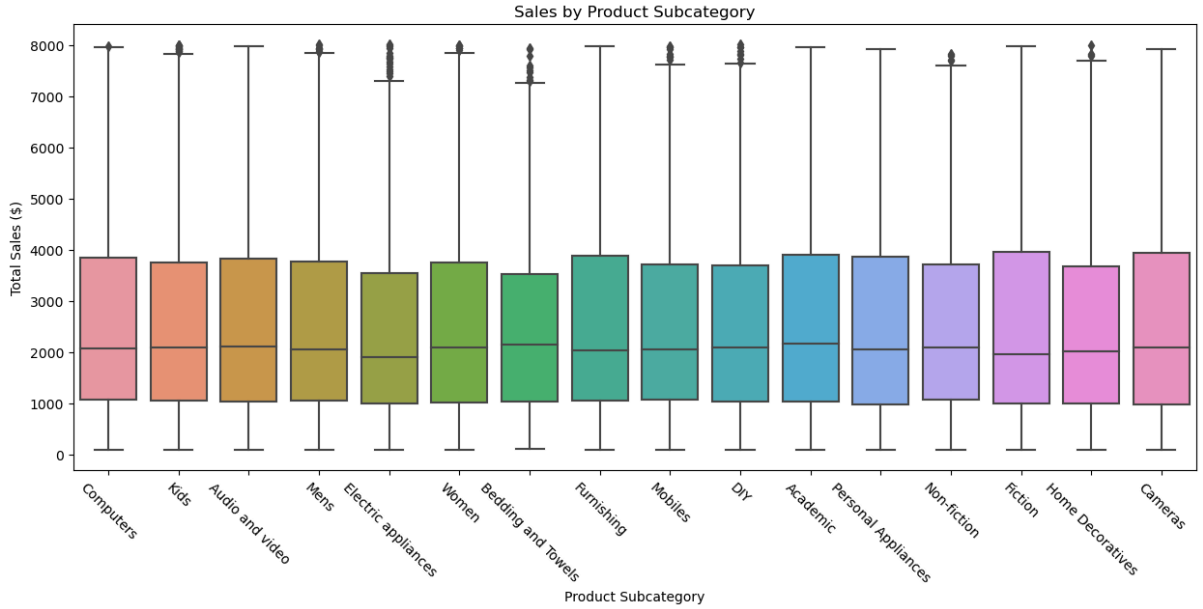


Monthly Sales by Year: Highlights seasonal sales trends, showing how sales vary across different months and years.

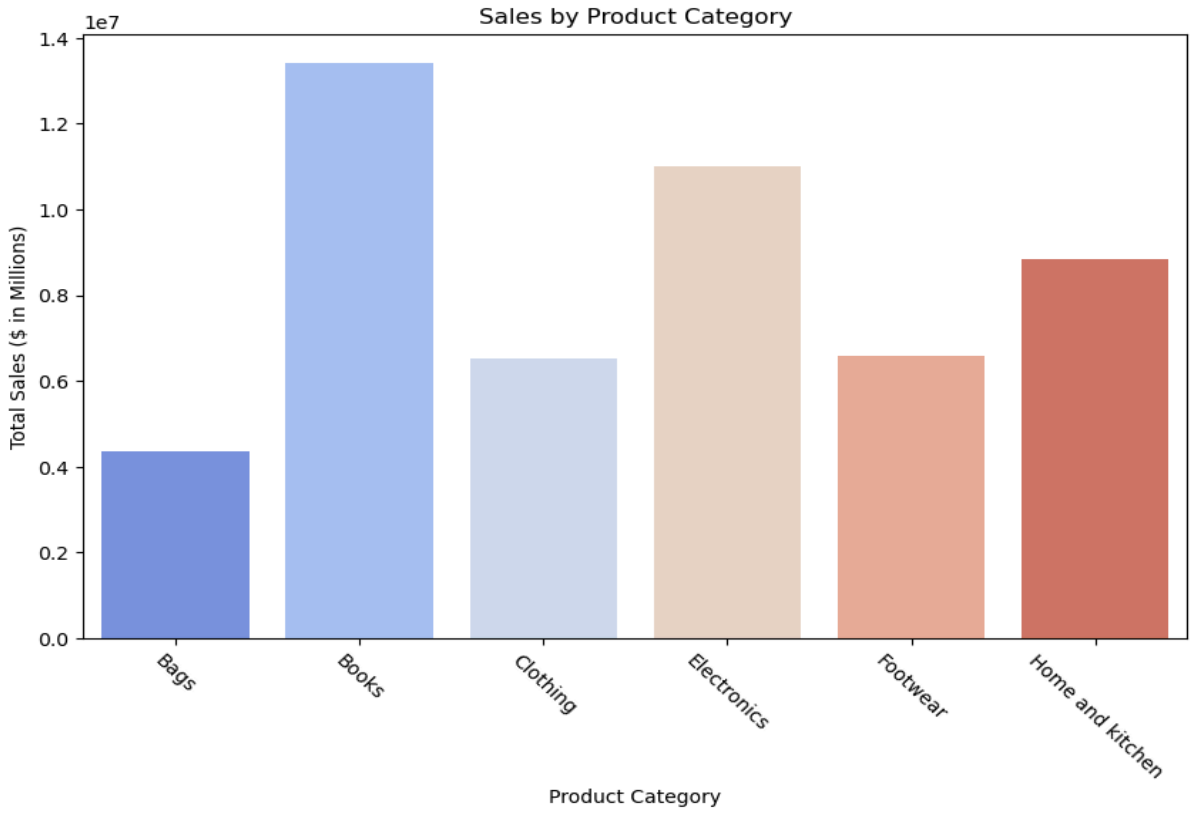


Sales by Product Subcategory: This chart highlights which subcategories have higher median sales, allowing us to identify top-performing

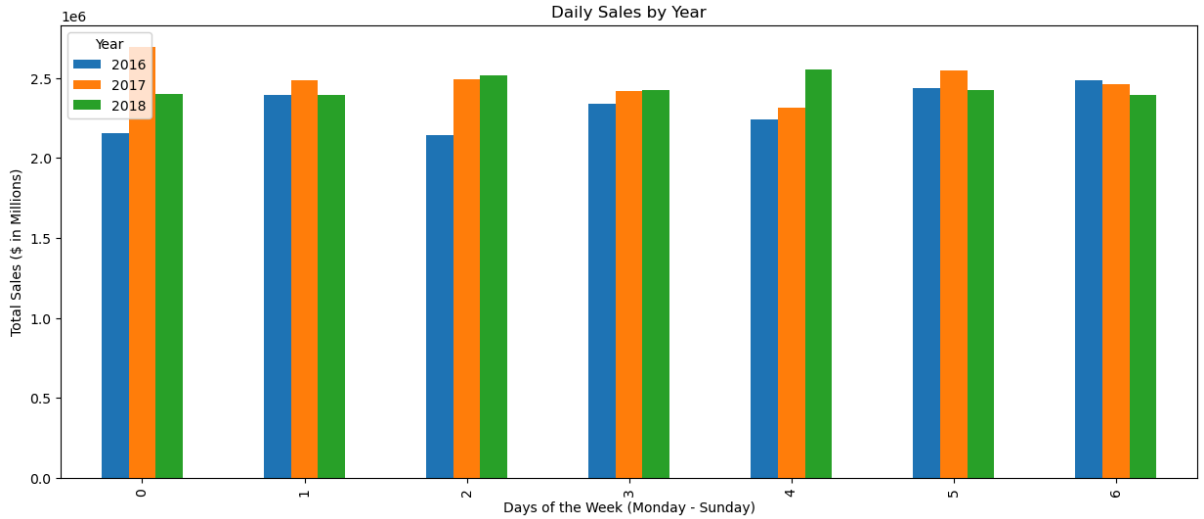
products.

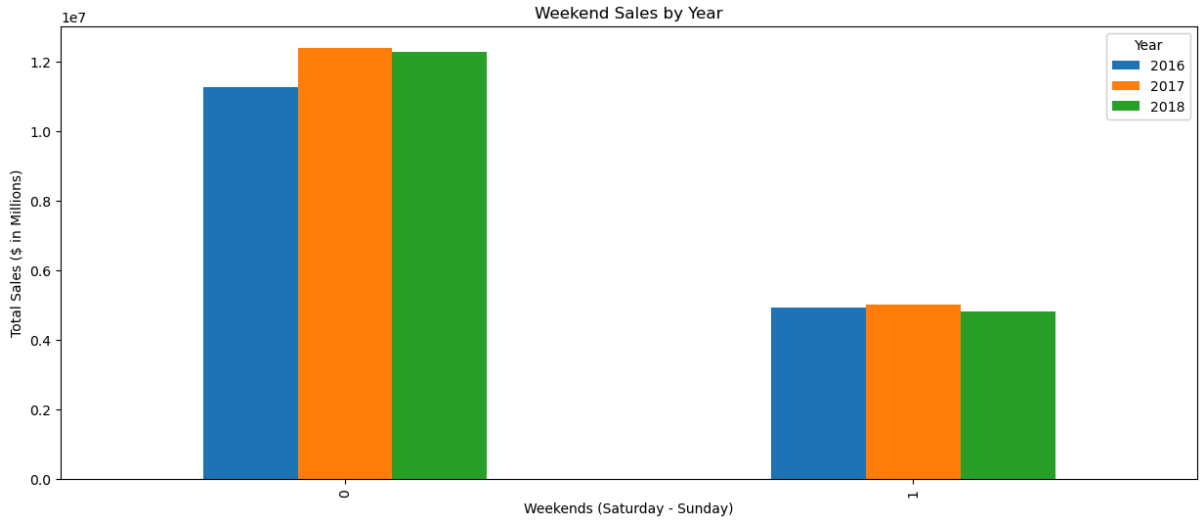


Sales by Product Category: Here we can determine which categories contribute most significantly to overall sales.



Weekend vs. Weekday Sales: This analysis compares sales patterns across different days of the week, typically showing higher sales during weekends.





Analysis

Seasonal Trends: Sales peak during certain months, likely due to holidays and promotional events.
Day of Week Patterns: Sales show variation across different days of the week, with weekends typically higher.
Year-over-Year Trends: Sales growth is observed across the years, indicating an overall positive trend.

Chapter 3: Preprocessing and Training Data Cleaning

Methodology

Encoding Categorical Variables: Convert categorical variables to numerical format.
Feature Scaling: Standardize data to ensure uniform contribution of features.

Chapter 4: Modeling

Introduction

We're aiming to build and evaluate sales forecasting models to pick the best one for predicting future sales. This involves selecting the right models, training and validating them, and checking their performance with different metrics.

Methodology

Model Selection:
We tried out several forecasting models, focusing mainly on the Facebook Prophet model because it's great at handling time series data, including seasonality and holidays.

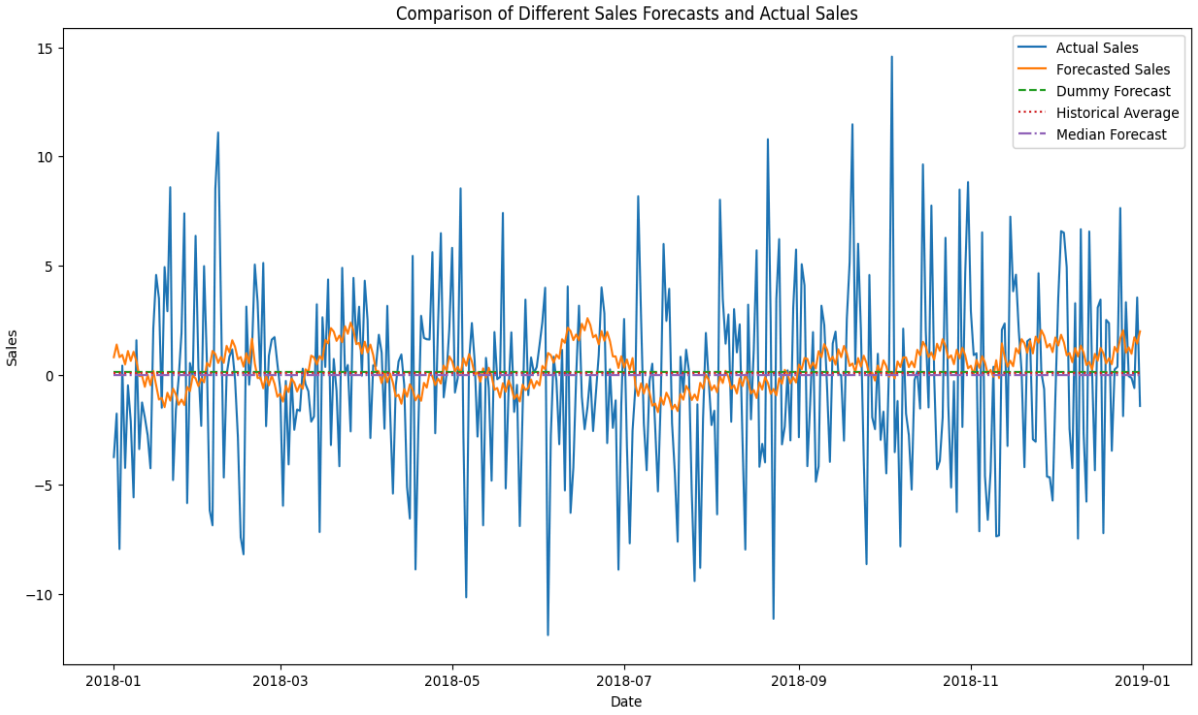
Why Prophet Model?

Advantages: Handles seasonality and holidays effectively, provides interpretable components (trend, seasonality), and is robust to missing data.
Potential Downsides: May not perform well with very noisy data, can be computationally intensive, and requires careful parameter tuning.

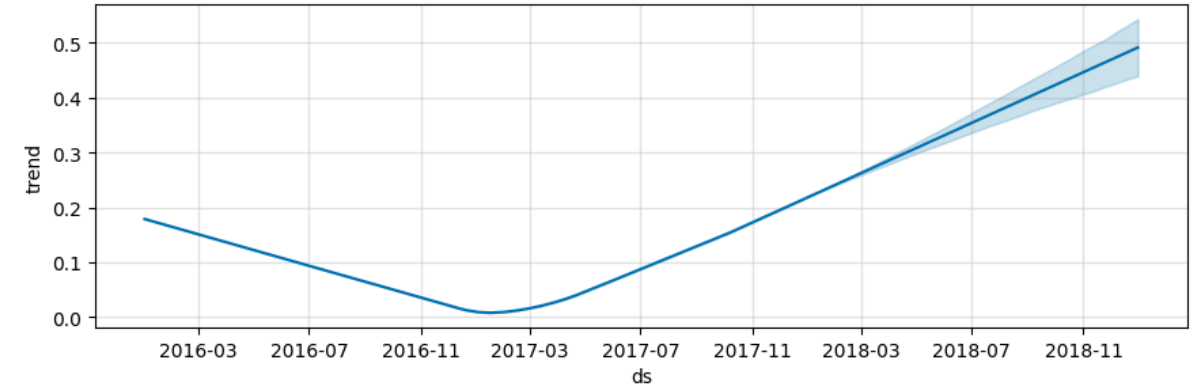
Training and Validation:
We trained the models on past sales data and validated them to ensure they work well with new data.

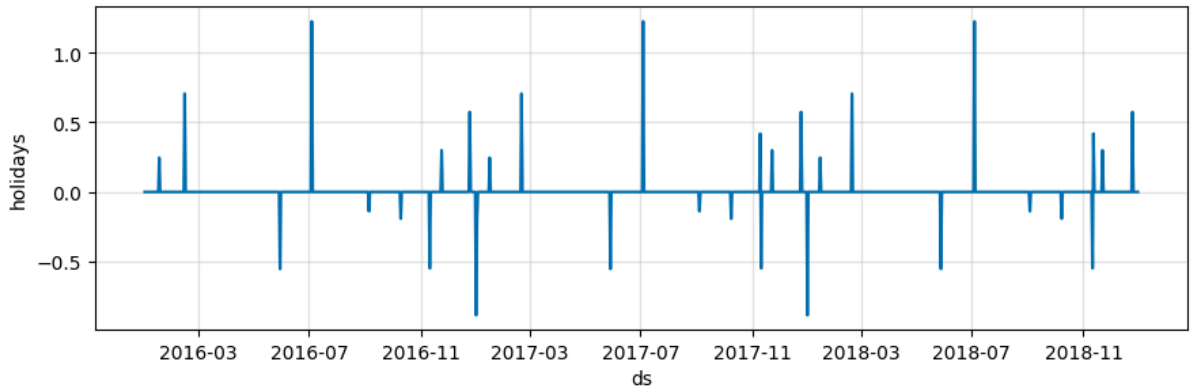
Visualizations and Key Findings

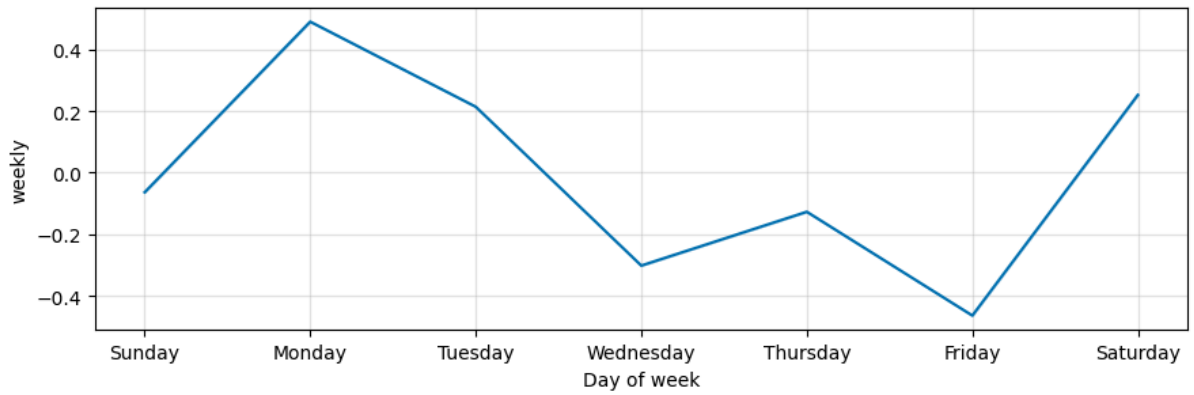
Comparison of Sales Forecasts and Actual Sales:
This chart shows how the different model forecasts stack up against the actual sales data.

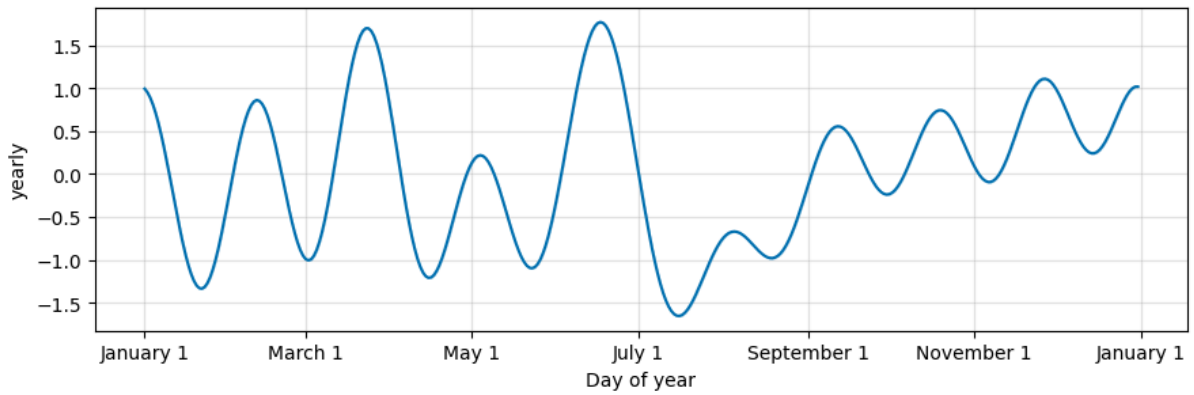


Trend, Holidays, Weekly, and Yearly Components: Prophet breaks down the time series data to show how trends, holidays, and weekly/yearly patterns affect sales.

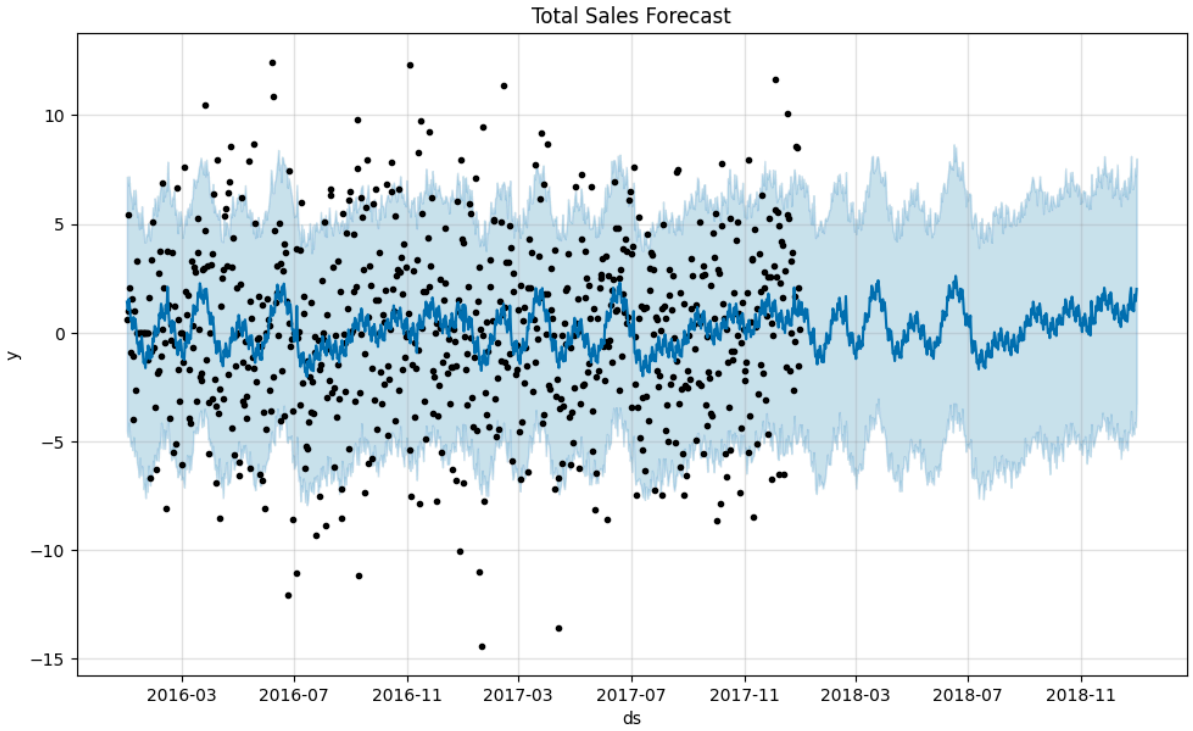




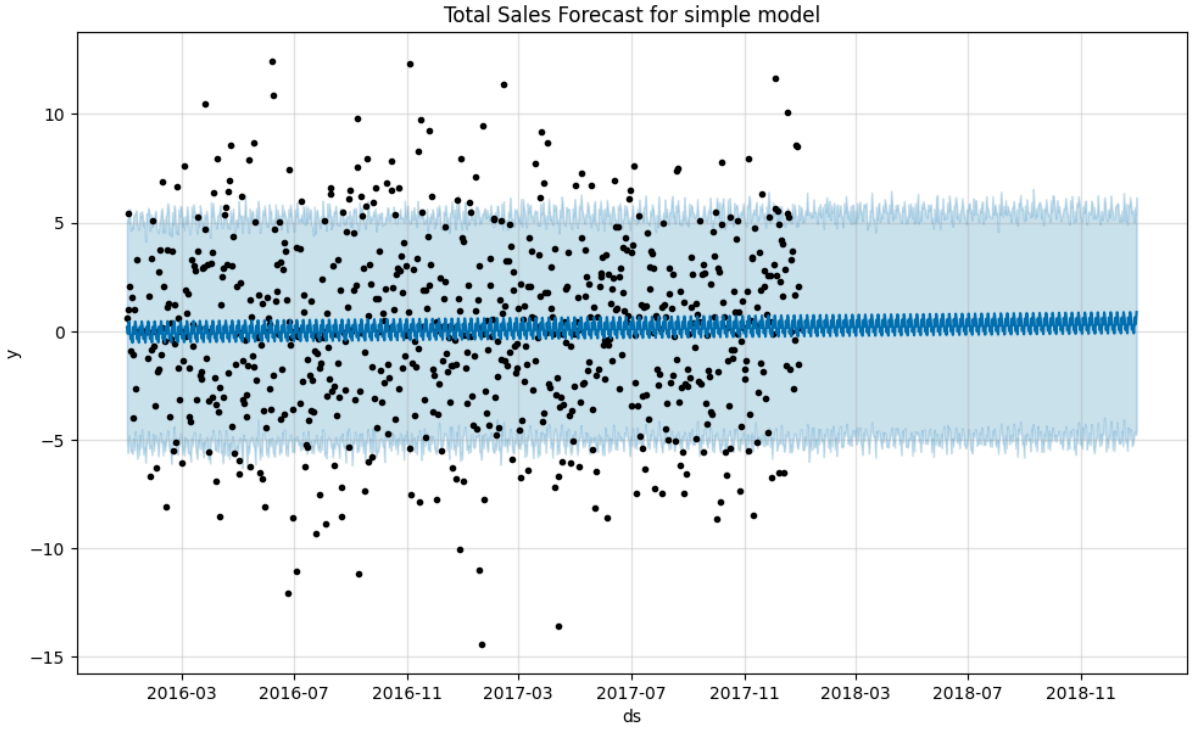




Total Sales Forecast: This chart shows our sales forecasts with confidence intervals, giving a range of expected future sales.



Total Sales Forecast for Simple Model (with no fine tuning): Compares simple model forecasts with actual sales to see how the basic model performs.



Chapter 5: Model Evaluation

Evaluation Metrics

We looked at the performance of our models using these metrics:

Prophet Model: MAE, RMSE, and MAPE values show it captures overall trends well but has trouble with fluctuations.
Dummy Forecast: A basic model to compare against more complex ones.
Historical Average Forecast: Uses past average sales for predictions, serving as a simple benchmark.
Analysis
The Prophet model captures the overall trend well, but its performance is not better than simpler baseline models.
High MAPE suggests the model struggles with percentage errors.
Deviations between forecasted and actual sales indicate potential issues with handling fluctuations.

Using Dummy Model and Other Models

Dummy Model: Serves as a baseline to compare the performance of more sophisticated models.
Historical Average Forecast: Provides a simple benchmark by using the average of historical data.
Prophet Model: Selected for its ability to handle complex seasonality and provide interpretable forecasts.

Challenges with the Dataset

Our dataset is quite hectic, with many outliers and fluctuating trends, making it hard to build a reliable forecasting model.

Practical Applications

Had the models performed exceptionally well, we could have achieved:

Optimized Inventory Management: Precise inventory levels to meet demand without overstocking.
Reduced Operational Costs: Lower holding costs and reduced wastage.
Improved Customer Satisfaction: Consistent availability of products leading to better customer experiences.
Strategic Decision-Making: Data-driven decisions on promotions, pricing, and stocking strategies.

Future Improvements

To improve the model's performance, further tuning is necessary. This may include incorporating additional features, using more advanced models, or applying more sophisticated preprocessing techniques. Exploring ensemble methods and leveraging external data sources for more context could also enhance model robustness.

Conclusions

The current performance indicates the need for refinement before relying on the forecasts for strategic decision-making. Continuous improvements will be essential to develop a more accurate and reliable sales forecasting model.

Summary

In summary, while we have made significant progress, the complexity of the data necessitates ongoing efforts to refine our models and approaches to achieve better forecasting accuracy. The hectic nature of the dataset presents a significant challenge, but with continuous improvements, we can aim to develop a more reliable forecasting system in the future.