Sales Forecasting for a Retail Store

Data Science Capstone Project

by Denis Grekov

May 2024

1. Project Overview

This project aims to develop machine learning models to predict future sales using historical sales data. By accurately forecasting sales, we can help with:

- **Inventory Management**: Maintain optimal stock levels to meet customer demand without overstocking
- **Financial Planning**: Improve budgeting, resource allocation by managing cash flows more efficiently.
- **Strategic Decision-Making:** Inform marketing campaigns, promotions, and pricing strategies, ensuring alignment with market trends and customer needs.

2. Dataset Description

Objective: Develop an accurate sales forecasting model.

- Invoice_ID
- Cust_ID
- Txn_date *
- Product_Sub_category *
- Product_Category *
- Qty *
- Rate
- Tax
- Total_Amt *
- Store_type *

* - for forecasting store sales we will be focusing on these columns

3. Challenges and Roadblocks

Data Quality: Address missing values, outliers, and errors.

Seasonality and Trends: Identify and account for patterns.

Return Transactions: Handle returns to avoid skewed figures.

Feature Engineering: Create meaningful features.

Model Selection: Choose the best forecasting techniques.

4. Project Goals and Methods

Goal: Develop a reliable sales forecasting model.

Approach:

1. Data Cleaning: Ensure accurate and reliable data by addressing return transactions and outliers.
2. Exploratory Data Analysis (EDA): Analyze data to uncover trends and patterns.
3. Feature Engineering: Transform raw data into useful features.
4. Model Building: Train and test forecasting models.
5. Model Evaluation: Use MAE, RMSE, and MAPE to assess model performance.

5. Data Cleaning

Standard Python Methods: standard Python libraries and techniques were used for data cleaning tasks such as handling missing values and correcting errors.

Outlier Detection: Use the Interquartile Range (IQR) method to identify and manage outliers.

6. Data Analysis

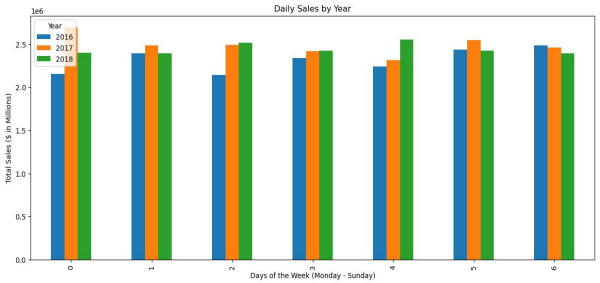
Seasonal Trends: Peaks during certain months due to holidays and promotions.

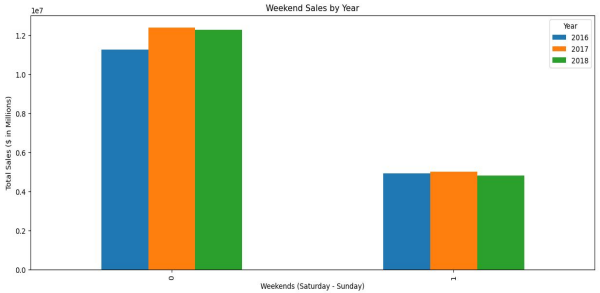
Day of Week Patterns: Higher sales on weekends.

Year-over-Year Trends: Overall positive sales growth.

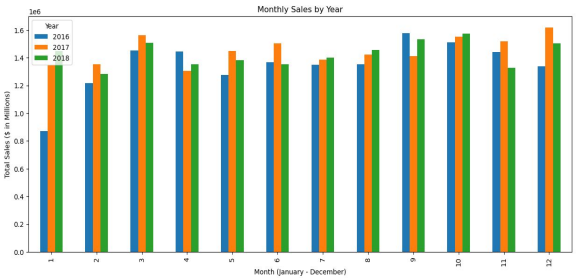
7. Charts

Monthly, daily sales:





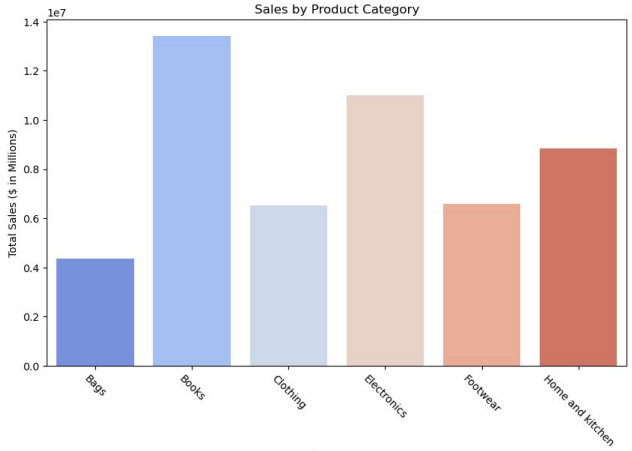
Year-over-Year Trend, Monthly Sales:

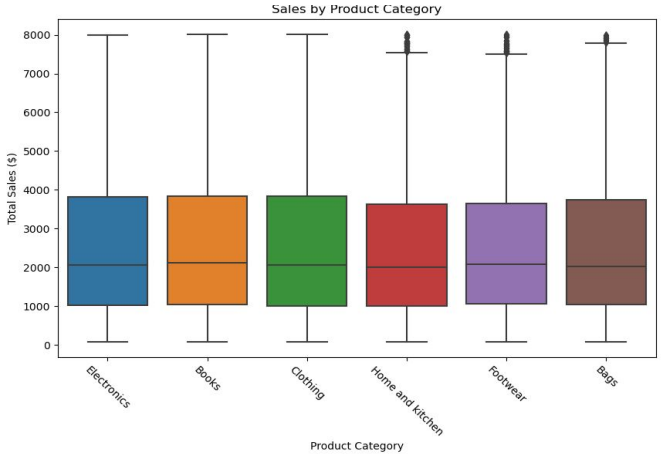


The charts indicate that sales are consistently high on weekends and beginning of the week. Additionally, sales exhibit steady patterns throughout the years, with noticeable peaks in the middle months, reflecting seasonal trends.

8. Category sales

The sales analysis reveals that Books and Electronics lead in total sales, with Books significantly outperforming other categories. Other categories like Clothing, Home and Kitchen, Footwear, and Bags also demonstrate significant sales but with more variability in their performance.





9. Preprocessing and Training Data

Encoding and Standardizing data

Encode Categorical Variables: Convert categorical variables into numerical format to make them usable for machine learning algorithms.

Standardize Features: Scale features to ensure uniform contribution and improve model performance. This step is important for algorithms that measure distances between data points.

10. Modeling and Evaluation

Model Selection: Choose the best forecasting models, focusing on models like Prophet.

Training: Train the selected models using historical sales data to learn patterns and trends.

Validation: Validate the models on a separate dataset to ensure they generalize well to new, unseen data.

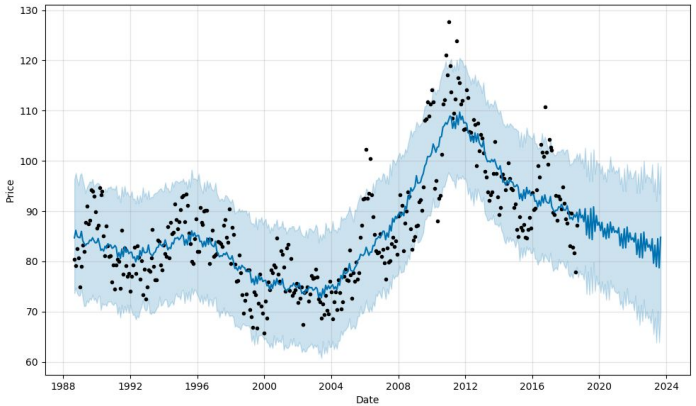
Evaluation Metrics: Assess model performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics help quantify the accuracy and reliability of the forecasts.

11. Why FB Prophet

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.

An Example of FB Prophet forecasting → → →



12. Comparison of Sales Forecasts and Actual Sales

Objective: Compare model forecasts against actual sales data.

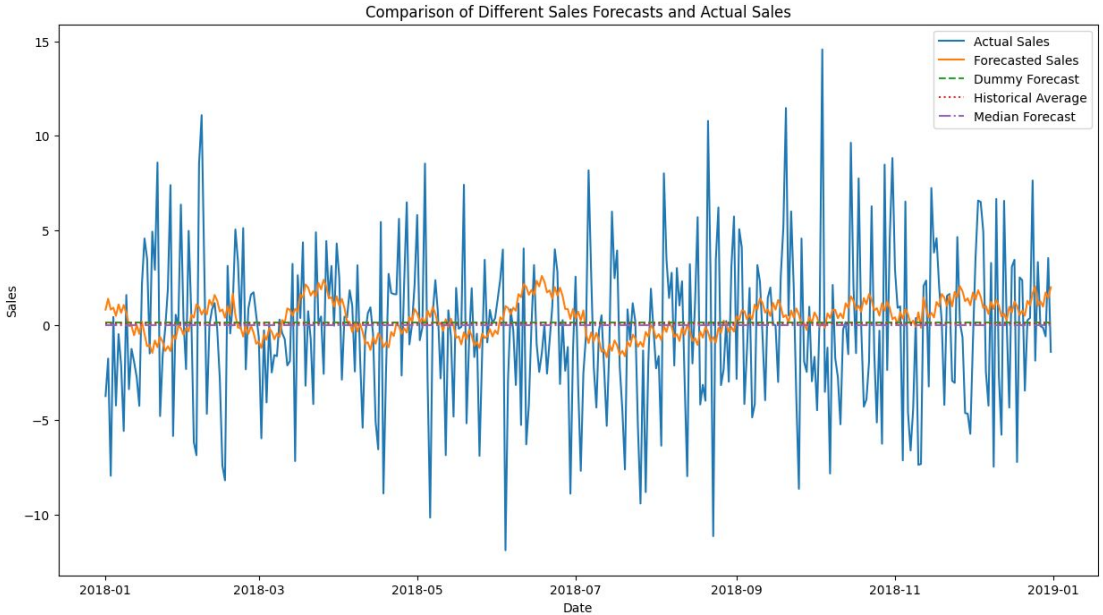
Prophet Model: Evaluate how well the Prophet model forecasts align with actual sales.

Dummy Forecast: Use a basic model to set a baseline for comparison.

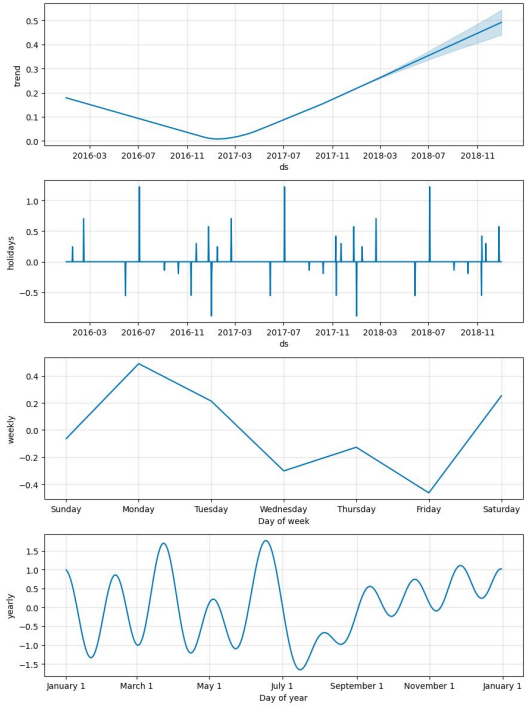
Historical Average Forecast: Compare with forecasts based on historical sales averages.

Actual vs. Forecasted: Visualize differences to identify the best-performing model.

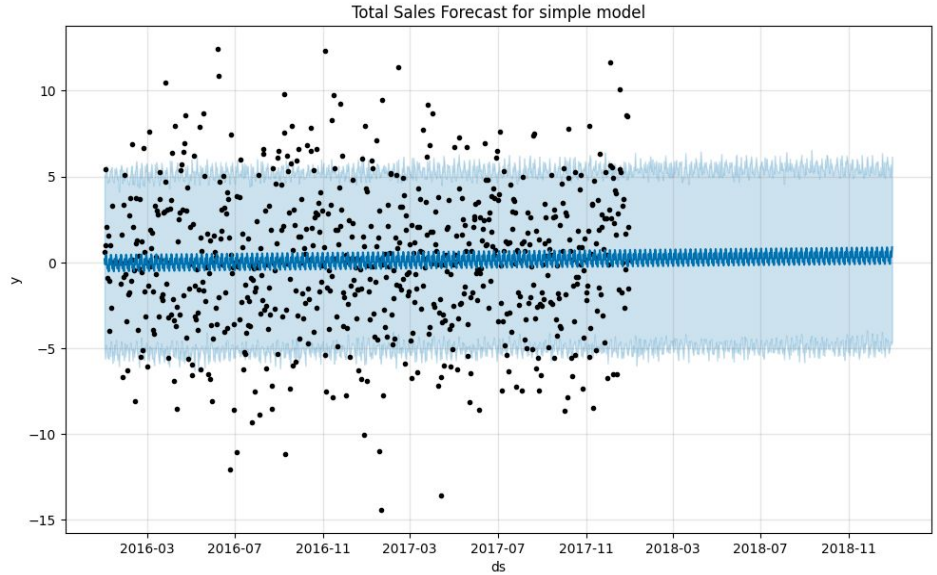
13. Chart Sales Forecasts vs Actual Sales

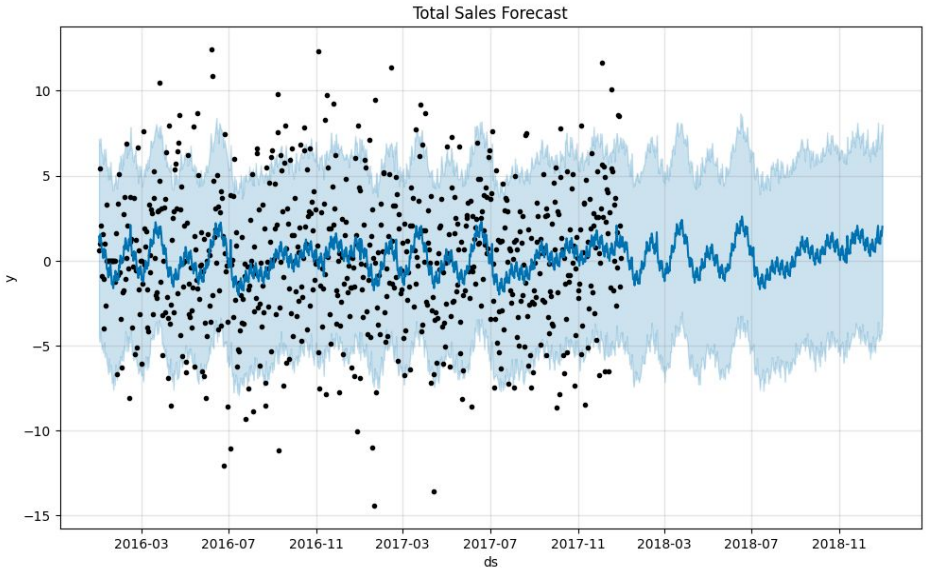


14. Trend, Holidays, weekly and Yearly Components



15. Total sales forecast Simple model vs Tuned model





16. Evaluation Metrics and Analysis

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 model.

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.

The model's limitations may be due to the lack of additional predictive features like economic indicators, promotional events, and weather data.

17. Future Improvements

Additional Features: Incorporate external economic indicators, promotional events, and weather data. The additional features can provide more context and improve the model's accuracy by accounting for factors that influence sales.

Advanced Models: Experiment with sophisticated models like neural networks and ensemble methods. These models can capture complex patterns and interactions in the data that simpler models might miss.

Hyperparameter Tuning: Optimize model parameters using grid search and cross-validation. Ensuring that the model is performing at its best by finding the optimal set of parameters.

Data Augmentation: Increase training data by combining data from similar stores or using synthetic data. Additional data can help the model learn better and generalize well to new data.

18. Summary and Conclusions

Summary:

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.

Conclusions:

The complexity of the data requires 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 systems.

Thank you!