Sales Forecasting Project using ARIMA, ETS, and Random Forest

Author: Eyesly Meribha Johnson Paulraj

1. Introduction

This project focuses on building an end-to-end solution for forecasting sales using both classical time series techniques and modern machine learning models. It demonstrates how businesses can leverage historical sales data, external factors like oil prices and holidays, and machine learning to make better inventory and marketing decisions.

2. Business Objective

The objective is to forecast daily product sales across multiple stores and product families. Accurate sales predictions can help optimize stock levels, manage supply chains, plan promotions, and ultimately reduce overstock and stockouts.

3. Dataset Overview

The dataset consists of multiple files:

- train.csv: Historical sales data
- test.csv: Data used for generating predictions
- oil.csv: Daily oil prices
- holidays_events.csv: Holiday and event calendar
- stores.csv: Store metadata
- transactions.csv: Store-level daily transaction data
- sample_submission.csv: Submission format

4. Data Cleaning and Preprocessing

Each dataset was carefully cleaned and merged based on the date and store identifiers. Missing values were filled, oil prices were forward-filled, and categorical values were encoded. The training set was enriched with time features like year, month, day, and whether the day was a weekend.

5. Exploratory Data Analysis

EDA included line plots to identify seasonal trends, bar charts for daily and family-level sales distributions, and transaction volume studies. This helped understand the variability in sales across different product families and stores.

6. Time Series Forecasting (ARIMA & ETS)

ARIMA and Exponential Smoothing (ETS) models were trained on aggregated daily sales. Forecasts were made for the next 30 days and evaluated using RMSE. RMSE Scores:

- ARIMA: 107,251
- ETS: 161,663

7. Random Forest Model

A Random Forest Regressor was trained using aggregated store-family-day-level features. The model was trained on a sampled subset of the data to improve speed. Feature importance was calculated to interpret which variables had the most impact.

RMSE: 371 (Best among all models)

8. Power BI Dashboard

A 5-page interactive dashboard was created to visually present the insights and predictions:

1. Overview & Business Objective
2. Data Exploration
3. Time Series Forecasting (ARIMA & ETS)
4. Machine Learning Forecast (Random Forest)
5. Final Submission & Recommendations

The dashboard includes KPIs, line charts, bar plots, tables, and slicers for interactivity.

9. Final Insights & Recommendations

The Random Forest model significantly outperformed the classical methods, showing the advantage of combining engineered features and external factors. Key business takeaways include:

- Sales spike before holidays and weekends
- Promotions impact varies across product families
- Store clusters can be targeted based on past sales patterns

10. Tools & Technologies Used

- Python (Pandas, Scikit-learn, Statsmodels, Matplotlib)
- Power BI (for dashboard visualization)
- Git & GitHub (for version control and code sharing)
- Jupyter Notebook (for model development and documentation)

11. Power BI Dashboard Link

Power BI dashboard:
https://drive.google.com/drive/u/0/folders/1SA_Tl1PJRE7xYKICO3M07leajcOfvjiT