**Forecasting Sales Data: A Comprehensive Analysis**

**1. Introduction**

**1.1 Objective**

The primary objective of this analysis is to develop and evaluate various forecasting models to predict the sales quantity of a different products in an electronics store. The focus is on comparing different algorithms, including traditional statistical methods and advanced machine learning techniques, to determine the most accurate forecasting approach.

**1.2 Data Description**

The dataset used includes sales information from an electronics store, spanning from January 2018 to June 2024. Key columns in the dataset include 'Product Name', 'Purchase Date', and 'Quantity'. The analysis specifically targets the sales trend of any selected product.

The dataset generation script simulates sales data for an electronics store in Mumbai from 2018 to June 2024. The script considers various categories, brands, and factors that impact sales, such as seasonality, promotions, COVID-19 impact, and festive periods.

**Key Steps in Dataset Generation:**

1. **Define Categories and Brands:**
   * Categories include Audio, Video, Mobile Devices, Appliances, and Computers.
   * Each category has associated brands, e.g., Sony, Samsung, LG for Audio.
2. **Sales Distribution:**
   * Specifies the proportion of sales for each category.
3. **Functions to Generate Product Details:**
   * generate\_product\_id(i): Generates a unique product ID.
   * generate\_product\_name(category): Generates a random product name based on the category.
   * generate\_unit\_price(category, product\_name): Determines the unit price based on the category and product name.
4. **Quantity and Purchase Date Generation:**
   * generate\_quantity(year, category, seasonality, promotion, covid\_impact, festive\_period, month): Calculates quantity based on year, category, seasonality, promotions, COVID-19 impact, and festive periods.
   * generate\_purchase\_date(year, month): Generates a random purchase date within the specified month and year.
5. **COVID-19 and Festive Periods:**
   * Simulates COVID-19 lockdowns and their impact on sales.
   * Defines festive seasons like Diwali, Dussehra, and Ganesh Chaturthi to boost sales during these periods.
6. **Product ID to Name Mapping:**
   * Ensures consistent product names for the same product IDs across the dataset.
7. **Data Generation for 2018 to June 2024:**
   * Increases the number of products each year.
   * Generates a base number of records per month, ensuring diversity in product details, purchase dates, and sales impacts.
8. **Additional Data Generation:**
   * Ensures the dataset has at least 50,000 rows by generating additional records as needed.
9. **Final DataFrame and Export:**
   * Combines all generated data into a pandas DataFrame.
   * Exports the DataFrame to an Excel file named electronics\_store\_sales\_data.xlsx.

**Example Data Fields:**

* **Product ID:** Unique identifier for each product.
* **Product Name:** Randomly generated product name.
* **Category:** Product category (e.g., Audio, Video).
* **Brand:** Randomly selected brand from the defined list.
* **Unit Price:** Price of a single unit of the product.
* **Quantity:** Number of units sold.
* **Total Price:** Total revenue from the sale (Unit Price × Quantity).
* **Purchase Date:** Date of purchase.
* **Seasonality:** Impact of seasonal trends (binary: High or Low).
* **Promotion:** Whether the product was on promotion (binary: Yes or No).
* **COVID Impact:** Impact of COVID-19 on sales (e.g., Lockdown, Online Surge, Normal).
* **Festive Period:** Whether the purchase occurred during a festive period (binary).

**Summary:**

The generated dataset simulates realistic sales trends considering historical events, seasonality, promotions, and festive impacts, providing a comprehensive overview of the store's sales from 2018 to mid-2024. The approach ensures a rich and varied dataset that can be used for various analyses, including trend analysis, forecasting, and inventory management.

**2. Methodology**

**2.1 Data Extraction and Cleaning**

The extracting\_cleaning function performs the following steps:

1. **Filtering Data:** Extracts relevant data for the specified product.
2. **Aggregation:** Aggregates the quantity data on a monthly and quarterly basis.
3. **Handling Missing Values:** Reindexes the data to ensure a complete date range and fills missing values with zero.
4. **Visualization:** Plots the monthly and quarterly quantity trends to visualize the sales patterns.

**2.2 Forecasting Models**

**2.2.1 Machine Learning Models**

1. **Linear Regression:** Utilizes a linear model to predict future quantities based on historical data.
2. **Random Forest:** Employs an ensemble of decision trees to capture complex patterns in the data.
3. **Support Vector Regression (SVR):** Uses a support vector machine to handle non-linear relationships.
4. **Decision Tree Regression:** Constructs a tree-based model to predict sales quantities.

**2.2.2 Deep Learning Models**

1. **LSTM (Long Short-Term Memory):** A type of recurrent neural network designed for sequence prediction. The model is trained on historical sales data to forecast future values.

**2.2.3 Statistical Models**

1. **Exponential Smoothing (ETS):** Applies smoothing techniques with additive trend and seasonal components to forecast future quantities.
2. **ARIMA (AutoRegressive Integrated Moving Average):** Utilizes historical data to model and forecast future values, focusing on trend and seasonality.
3. **Prophet:** A forecasting tool developed by Facebook, which handles seasonal effects and holidays in the prediction.

**2.3 Model Evaluation**

The performance of each model is evaluated based on the Mean Absolute Error (MAE) metric. The models' predictions are compared against a test dataset (Quantity sold from Jan 2024 to Jul 2024) to determine their forecasting accuracy.

**2.4 Hybrid Modeling**

A hybrid model is created by combining forecasts from different models. Weights are assigned based on the inverse of each model's MAE, aiming to enhance prediction accuracy by leveraging the strengths of various models.

**3. Results**

**3.1 Model Performance**

The performance results for each product, MAE is calculated, and the forecasts are assessed for accuracy.

**3.2 Best Model Selection**

Complete prediction for all the unique products has to be done. Random products from different categories were tested against our models and they are providing accuracy on average >70%.

**3.3 Forecast Visualization**

Forecasts for the product are visualized using time series plots. These plots include:

* Historical data
* Forecasted data

**4. Conclusion**

**4.1 Summary**

The analysis demonstrated that hybrid modeling provides the most accurate forecasts for sales data compared to individual models. The combination of multiple forecasting techniques allowed for better handling of the data's complexities and trends.

**4.2 Future Work**

Future work could involve:

* Incorporating additional features or external data to enhance model accuracy. Creating a multi-variate forecasting model.
* Testing other deep learning architectures like 1-D CNN and LLMs.
* Implementing real-time forecasting and updating models with new data.
* Implementing Big Data Handling
* And creating a better interface using streamlit to find demand forecast.

**4.3 Practical Implications**

The forecasting models and methods developed in this analysis can be applied to improve inventory management, optimize stock levels, and make informed business decisions based on accurate sales predictions.