Illustration of Activities in the Forecasting Process Using Multiple Inputs

The forecasting process involves several systematic steps that integrate data from multiple sources to predict future trends. Let's illustrate this process with an example from retail sales forecasting:

1. Problem Definition

The first step defines the objective of the forecast.

- **Example:** Predicting monthly sales for a product to optimize inventory and reduce stockouts.

2. Data Collection

Relevant historical data is gathered, including sales, promotions, holidays, weather conditions, and competitor activity.

- **Example:** Sales data from a store (primary input) combined with weather data and holiday schedules (secondary inputs).

3. Data Analysis

Analysing patterns and relationships in the data is crucial.

- **Example:**
    - A **time series plot** might reveal trends (e.g., increasing sales over months).
    - Seasonal variations may show higher sales during summer.
    - Scatter plots might reveal correlations, such as higher sales during holidays.

4. Model Selection and Fitting

Select an appropriate model that suits the data characteristics.

- **Example:** Using a **multivariate time series model** to include both sales data (primary variable) and weather as predictors.

5. Model Validation

Evaluate the model's accuracy using part of the historical data (validation set).

- **Example:** Train the model on data from 2020-2022 and validate it on 2023 data to test predictive accuracy.

6. Forecasting Model Deployment

Deploy the validated model for operational use.

- **Example:** Using the model to forecast sales for the next quarter, with forecasts automatically updated weekly based on the latest inputs.

1. Analyse the Missing Data

- **Nature of Missing Data:**
  Use exploratory data analysis (EDA) to identify the type and pattern of missingness:
    - **MCAR (Missing Completely at Random):** If the missing data is random, simpler imputation methods may suffice.
    - **MAR (Missing at Random):** Identify relationships with other variables to inform imputation.
    - **MNAR (Missing Not at Random):** Collaborate with domain experts to understand the root cause and address potential biases.

---

2. Correct Missing Data Using Imputation Methods

(a) **Mean Value Imputation**

- Replace missing values with the mean of the non-missing values in the column.
- **Example:** For patient age, replace missing values with the mean age of all patients.
- **Suitability:** When the data lacks trends or seasonal patterns.

(b) **Stochastic Mean Value Imputation**

- Add random noise to the mean value to better reflect data variability.
- **Example:** For missing lab results, replace them with the mean result plus a small random variation.

(c) **Regression Imputation**

- Predict missing values using a regression model based on other related variables.
- **Example:** Predict a missing blood pressure value using patient weight, age, and diagnosis.

(d) **Hot Deck Imputation**

- Replace missing values with observed values from similar patients (based on clustering or matching criteria).
- **Example:** For missing medication adherence, use values from patients with similar demographics and diagnoses.

(e) Cold Deck Imputation

- Use external reference datasets, such as historical healthcare records, to impute missing values.
- **Example:** Replace missing socioeconomic data with values from similar patient records in a previous year's dataset.

---

3. Validate the Imputation

- **Cross-validation:**
  Split the dataset into training and testing sets, imputing only the training set to evaluate the effectiveness of the imputation.
- **Error Analysis:**
  Assess how well imputed values match actual values where data is available.

---

4. Ensure Data Quality

- **Completeness:**
  Perform thorough checks to confirm all imputations align with the data context.
- **Accuracy:**
  Verify imputed values make sense clinically, e.g., no negative lab results.

Example in Context

1. Centralized Repository for Historical Data

- A data warehouse stores data related to **sales, inventory, supplier performance**, and other business metrics from all stores.
- This consolidated database enables access to consistent and structured historical data, which is essential for identifying trends and seasonality in demand forecasting.

Example from Forecasting:
Sales data from multiple stores over the past five years is stored and used to create a time series model, capturing patterns like peak sales during holidays.

---

2. Data Extraction and Cleaning

- Data is extracted from internal systems (e.g., POS systems) and external sources (e.g., economic indicators, weather conditions) and cleaned before loading into the warehouse.
- **Cleaning Processes:**
    - Missing data is imputed using techniques like **mean imputation** or **regression imputation**.
    - Outliers or inconsistent entries are corrected during the transformation stage.

Example from PDF:
Data cleaning ensures completeness, accuracy, and consistency, addressing challenges like missing inventory records or errors in sales data.

---

3. Data Transformation for Forecasting

- **Smoothing Techniques:**
  The data is pre-processed to identify patterns like trends and seasonal effects.
    - **Example:** Smoothing reduces noise, making it easier to detect the seasonal demand for winter clothing or summer products.
- **Feature Engineering:**
  Variables like day-of-week effects, regional holidays, and promotions are added to enhance the forecasting model.

---

4. Integration of Diverse Data Sources

- A **data warehouse** integrates multiple sources, such as transactional data, supplier data, and customer demographics.
- External data, such as weather reports, is added to enrich the model and improve accuracy.

Example from PDF:
Combining internal sales records with weather data helps forecast demand for raincoats or winter clothing during specific months.

---

5. Ensuring Data Quality

- **Key Dimensions of Data Quality:**
    - **Accuracy:** Ensures records reflect true values.
    - **Completeness:** All relevant fields, like product demand and inventory, are populated.
    - **Timeliness:** Data is updated regularly to allow for rolling forecasts.

Example from PDF:
For forecasting demand for perishable goods, timely updates ensure accurate inventory decisions.

---

6. Supporting Model Development

- **Time Series Models:**
  Models like ARIMA, Holt-Winters, and regression-based approaches rely on historical data stored in the warehouse.
    - The data warehouse enables fitting and validating these models over extensive datasets.

Example from PDF:
Historical trends and seasonal cycles are extracted to fit models like ARIMA for predicting monthly sales in all stores.

---

7. Real-Time Updates for Rolling Forecasts

- **Rolling Horizon Forecasting:**
  Data is regularly updated in the warehouse to produce revised forecasts based on the latest trends.
    - **Example from PDF:** New sales data from the past week automatically updates forecasts for the next month.