SALES FORECASTING MODELS FOR DIRECT SELLING BUSINESS: A DATA-DRIVEN APPROACH TO PREDICTIVE ANALYTICS

SIVARAJAN A/L S.ESVARAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The research methodology employed is discussed in this chapter to develop an accurate sales forecasting model for the direct selling businesses. The methodology outlines the process of collecting data from the primary direct selling website, followed by data preprocessing, feature engineering, model development, and evaluation using machine learning techniques to forecast future sales performance. This study aims to generate meaningful insights from historical sales data and external factors that influence sales in direct selling environments. Ultimately it will provide valuable forecasting capabilities for business planning and decision making.

3.2 Research Framework

The framework for this research includes the following stages:

1. Identifying the Research Problem and Reviewing Existing Literature.
2. Data Collection from Amway Business Operations.
3. Preprocessing the Data: Preparing and cleaning data for detailed analytical tasks
4. Exploratory Data Analysis (EDA): Time Series Analysis and Pattern Recognition
5. Sales Forecasting Models: Implementing multiple forecasting algorithms (Linear Regression, Random Forest, LSTM and ARIMA)

6. Model Evaluation: Comparing model performance using forecasting evaluation metrics.

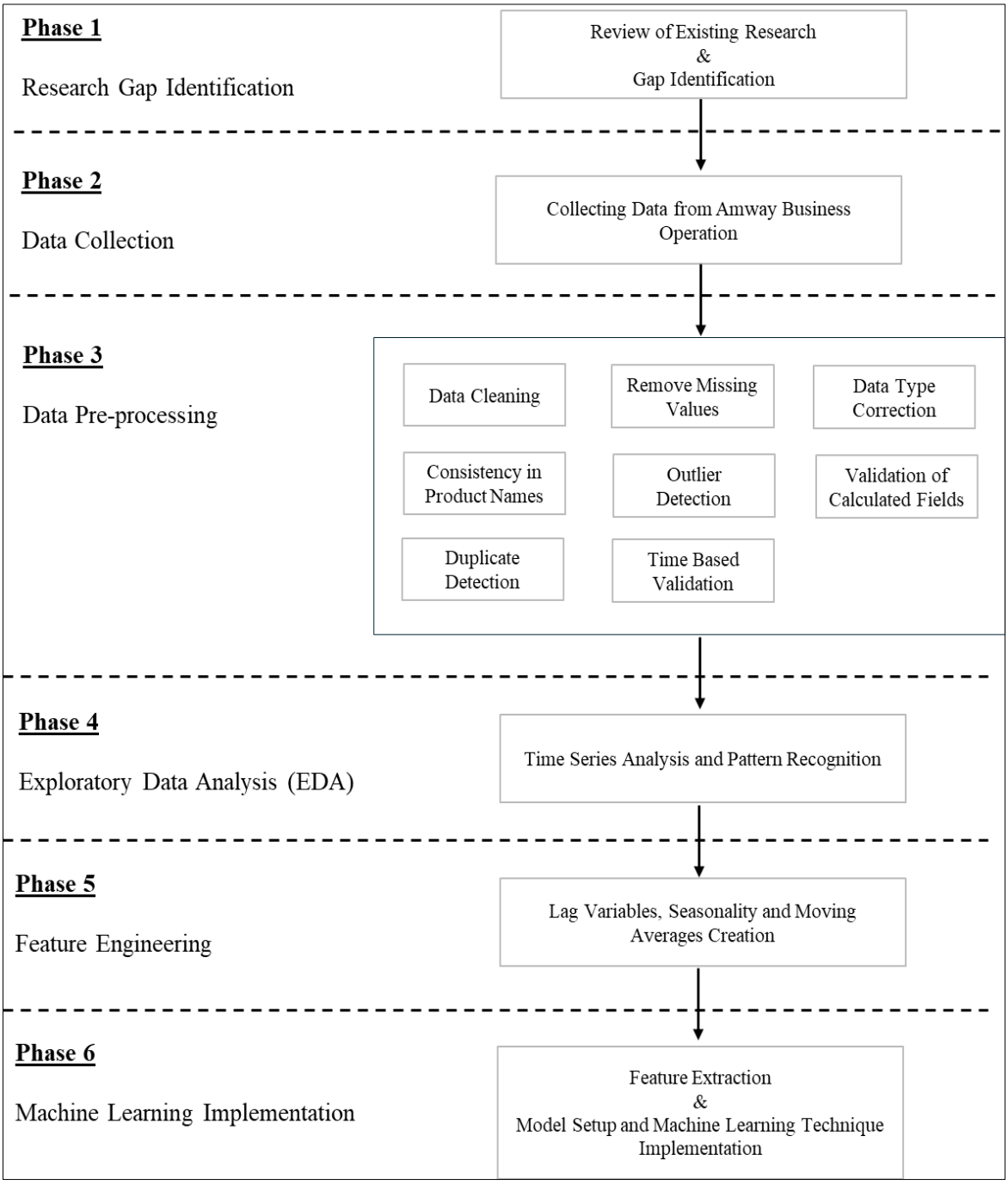

Figure 3.1 Research Framework for a Sales Forecasting Model in Direct Selling Business

3.3 Problem Formulation

This study aims to build a precise sales forecasting system for direct selling distributors using machine learning techniques. Thus providing valuable insights for strategic planning. However. Ensuring precise and dependable forecasting requires tackling a number of key issues:

a) Identifying key factors that influence sales performance in direct selling business.
b) Handling seasonality, trends and cyclical patterns in sales data.
c) Comparing the performance of Random Forest Linear Regression, LSTM and ARIMA algorithms in sales forecasting.
d) Developing robust model that can adapt to changing market conditions.

3.4 Data Collection

Data was collected from an individual Amway distributor sales records to develop a sales forecasting system that helps distributors to enhance their business growth and sales performance. The dataset represents actual transaction level sales data from an Amway distributor's business operations. This sales data provides detailed insights into customer purchasing patterns, product performance and sales trends.

Dataset Information:

1. **Collection Period**: April 2023 to April 2025. (24 consecutive months)
2. **Data Format**: Monthly PDF sales reports downloaded from the distributor portal.
3. **Conversion Process**: Python-based PDF to CSV conversion.
4. **Variables:** 12 comprehensive features per sales record
5. **Final Dataset**: 553,542 sales records in structured CSV format.

3.4.1 PDF to CSV Conversion Process

Python-based Conversion Methodology

The conversion from PDF to CSV format was accomplished using Python libraries specifically designed for PDF data extraction and processing.

Step 1: Environment Setup

```
python# Required libraries for PDF to CSV conversion
import pandas as pd
import PyPDF2
import tabula
import pdfplumber
import os
import glob
from datetime import datetime
import re

# Install required packages
# pip install pandas PyPDF2 tabula-py pdfplumber openpyxl
```

Figure 3.2 Importing Modules for PDF Handling

Step 2: PDF Data Extraction Function

```
python
def extract_amway_sales_from_pdf(pdf_file_path):
    """
    Extract sales data from Amway monthly PDF reports
    """
    try:
        # Method 1: Using tabula-py for table extraction
        tables = tabula.read_pdf(pdf_file_path,
                                 pages='all',
                                 multiple_tables=True,
                                 pandas_options={'header': 0})

        # Combine all tables from the PDF
        if tables:
            combined_df = pd.concat(tables, ignore_index=True)
            return combined df
```

Figure 3.3 Function to Extract tables from Sales PDF Reports

Step 3: Batch Processing All Monthly PDFs files

```
python
def process_all_monthly_pdfs(pdf_folder_path):
    """
    Process all monthly PDF files and combine into single dataset
    """
    all_monthly_data = []
    pdf_files = glob.glob(os.path.join(pdf_folder_path, "amway_sales_*.pdf"))

    print(f"Found {len(pdf_files)} PDF files to process")

    for pdf_file in sorted(pdf_files):
        print(f"Processing: {os.path.basename(pdf_file)}")

        # Extract data from current PDF
        monthly_data = extract_amway_sales_from_pdf(pdf_file)
```

Figure 3.4 Function to Process and Combine Monthly Sales Reports

Step 4: Export to Final CSV

```python
def export_to_csv(data, output_filename='Amway_Sales_Records_100k.csv'):
    """
    Export processed data to CSV format
    """
    # Final data validation
    print("Final dataset summary:")
    print(f"Total records: {len(data)}")
    print(f"Columns: {list(data.columns)}")
    print(f"Date range: {data['Date'].min()} to {data['Date'].max()}")

    # Export to CSV
    data.to_csv(output_filename, index=False)
    print(f"✓ Data exported to {output_filename}")
```

Figure 3.5 Function to Processed Sales Data to CSV

Number of rows: 553542
Number of columns: 12

	Order_ID	Date	Time	Customer_ID	Product_ID	Product_Name	Quantity	Unit_Price	Total_Amount	Return_Status	Customer_Age	Customer_Name
0	1000536015	2023-04-01	00:01:00	7010445678	125895A	Hand Sanitizer 400ml	5	59.0	295.0	No	36	Murugaiah A/L Ahyanari
1	1000526177	2023-04-01	00:04:00	7010045678	121697	DOUBLE X Refill Pack 186tab	3	198.0	594.0	No	45	Manirajah A/L Velu
2	1000045590	2023-04-01	00:06:00	7009583801	126457	Anti-Hair Fall Shampoo 280ml	2	72.0	144.0	No	39	Santhi A/P Sinnkoladai
3	1000119832	2023-04-01	00:07:00	7013409072	230727	Sanita Ultra Thin Wings 20/pk	5	13.0	65.0	No	43	Sumathi A/P Supaya
4	1000246702	2023-04-01	00:07:00	7024498970	102735	ClearGuardâ„¢ 180tab	4	139.0	556.0	No	44	Sheela A/P Berabakaran
...	...	...	...	...	...	...	...	...	...	...	...	...
553537	1000926116	2025-04-30	23:46:00	7010389012	123785	Renewing Reactivation Cream 50ml	4	250.0	1000.0	No	54	Perabakaran A/L Ramasamy
553538	1000234084	2025-04-30	23:48:00	7013658426	309177	Vergold Drip Coffee 10 sachets x 11g (Medium R...	5	58.0	290.0	No	39	Yugneswari A/P Rajendran
553539	1000076280	2025-04-30	23:49:00	7010045678	387800	Pursueâ„¢ Disinfectant Cleaner One Step 1l	3	30.8	92.4	No	45	Manirajah A/L Velu
553540	1000829916	2025-04-30	23:56:00	7686255	319372M	White Tea Toothpaste 200g	4	29.0	116.0	No	51	Tamil Selvi A/P Velayutham
553541	1000915819	2025-04-30	23:57:00	7010156789	592300	Garlic with Licorice 150tab	4	97.0	388.0	No	54	Arjunan A/L Pachappan

Figure 3.6 The Dataset Preview

3.5 Data Pre-Processing

Conducting an initial analysis is essential before proceeding with further preprocessing to ensure a thorough understanding of the dataset's feature availability. Several data processing and transformation procedures will be applied to prepare the data for time series forecasting.

Data Pre-Processing	Purpose
Preliminary Analysis	To analyse the given dataset and extract meaningful insights that will support the subsequent modelling phase.
Data Cleaning	Remove Missing Values, inconsistent records and outliers.

Table 3.1 Data Pre- Processing Method

3.5.1 Preliminary Analysis

Preliminary analysis plays a vital role in any data analysis project, which helps for a good understanding of the dataset, including its variables, format and structure. This preliminary identification helps to uncover potential issues. For example, outliers, inconsistencies and missing values need to be identified to ensure reliable and accurate results.

The Preliminary Analysis on this case involves two key stages:

a) Identifying common patterns within the raw data.
b) Analyzing data distribution based on time and relevant keywords.

3.5.2 Data Cleaning

In sales forecasting, data cleaning plays a vital role in guaranteeing that the data is precise, pertinent, and ready for processing by predictive models. The following steps outline the data cleaning procedures applied on the Amway Distributor Sales Dataset.

1. **Remove the Missing Values**

   Missing values are important to identify any rows and columns where data is missing such as blank entries. These missing values can lead to incorrect analysis or errors during modelling. Therefore, it's essential to remove rows or

columns with critical missing information or fill non-critical ones using reasonable estimates.

2. **Data Type Correction**

   Convert data into correct formats for example, date strings to datetime or text numbers to integers/floats. A proper data types allow for accurate calculations and visualizations.

3. **Consistency in Product Names and ID's**

   This to ensure product names in IDs are standardized across all the sales records. Inconsistent naming such as "G&H Lotion" vs "GH Lotion" leads to incorrect grouping. To address this, replace typos manually or use string normalization techniques.

4. **Outlier Detection**

   This steps involve to finds extreme or unrealistic values such as negative sales and absurd quantities. Outliers can skew analysis and mislead insights. To manage it, statistical methods (IQR, Z-scores) can be applied or simple filtering such as ensuring quantities are greater than zero can also be used.

5. **Validation of Calculated Fields**

   This step ensures calculated fields like Total Sale Amount reflect expected values (Quantity x Unit Price). Miscalculations can lead to inaccuracies in revenue analysis. To prevent this, recalculate and compare with any mismatches flagged for review or corrections.

6. **Duplicate Detection**

   This step focuses on identifying and removing duplicate rows for example transactions have been recoded twice. Duplicates can distort sales counts and totals, which leads to inaccurate analysis. Use drop duplicates () based on key identifying columns like Product ID to ensure data accuracy.

7. **Time Based Validation**

This validation method is to check whether dates fall within valid ranges. For example, ensuring they are not in the future unless they represent pre-orders. The invalid dates can affect time series analysis. This will be done by each date against the current date or a defined date range.

8. **Apply Pre-Processing**

This step integrates all the previously mentioned procedures into a single preprocessing pipeline, ensuring that the entire dataset is cleaned and processed consistently and uniformly.

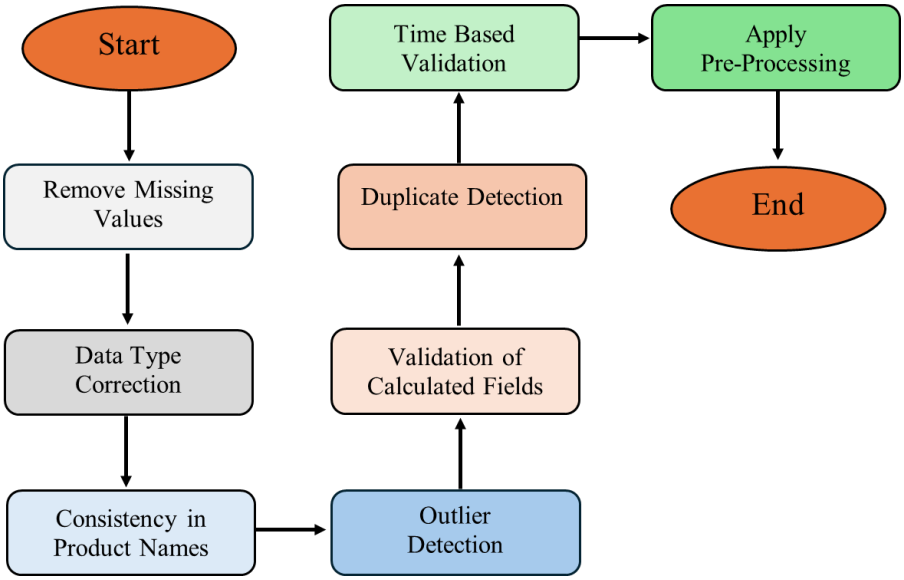

Figure 3.7 Flow Data Cleaning and Preparation

In this section, data cleaning is performed to identify and eliminate rows and columns with missing values. As illustrated in Figure 3.7, the data pre-processing steps include data type correction, standardizing the product names, outlier detection, validating calculation fields, normalizing, duplicate detection and time-based validation.

```python
import pandas as pd
import numpy as np

# Load Amway sales dataset
df = pd.read_csv('Amway Sales Records_100k.csv')

# Execute complete cleaning pipeline
print("Starting Amway data cleaning pipeline...")

# Apply all cleaning steps
df = remove_missing_values(df)
df = correct_data_types(df)
df = standardize_product_names(df)
df = detect_outliers(df)
df = validate_calculations(df)
df = remove_duplicates(df)
df = validate_time_data(df)
df = apply_preprocessing(df)

# Save cleaned data
df.to_csv('amway_sales_cleaned.csv', index=False)
```

Figure 3.8 Flow Data Cleaning Process

3.6 Exploratory Data Analysis

EDA plays a key role in data science by systematically exploring the relationships, underlying patterns, and characteristics of a dataset before any modelling is performed. EDA serves as the foundation for identifying temporal patterns, key business insights and customer behaviours that directly influence sales performance and inform subsequent forecasting model development in the context of sales forecasting for Amway distributors.

Time Series Analysis

Time series analysis forms a crucial component of EDA for sales forecasting, focusing on identifying temporal patterns such as trends, seasonality, and cyclical behaviours that are essential for accurate prediction models. Through time series decomposition, researchers can separate the underlying trend from seasonal variations and random noise, providing clear insights into long-term business growth patterns and recurring seasonal effects that impact Amway sales performance.

```
# Time Series Analysis

print("\n--- Time Series Analysis ---")

# Set Date as index for time series analysis
df_ts = df.set_index('Date').sort_index()

# Resample data to a specific frequency (e.g., daily, weekly, monthly)
and aggregate sales
# Example: Monthly Sales
if 'Total_Amount' in df_ts.columns:
    monthly_sales = df_ts['Total_Amount'].resample('M').sum()

    # Plot Time Series of Monthly Sales
    plt.figure(figsize=(14, 7))
    monthly_sales.plot()
    plt.title('Monthly Total Sales')
    plt.xlabel('Date')
    plt.ylabel('Total Sales Amount')
    plt.grid(True)
    plt.show()
```

Figure 3.9 Time Series Data Preparation for EDA

Pattern recognition analysis

Pattern recognition analysis within EDA encompasses customer behaviour analysis, product performance evaluation, and geographic market assessment to identify actionable business intelligence. This includes RFM analysis (Recency, Frequency, Monetary) for customer segmentation, cross-selling pattern identification, and regional market performance analysis that reveals opportunities for business expansion and optimization.

```
# Pattern Recognition Analysis

print("\n--- Pattern Recognition Analysis ---")

# Seasonal Patterns (using the extracted time features)
# Example: Sales by Month
if 'Sales Month' in df.columns and 'Total_Amount' in df.columns:
    monthly_avg_sales = df.groupby('Sales Month')['Total_Amount'].mean()
    plt.figure(figsize=(10, 6))
    monthly_avg_sales.plot(kind='bar')
    plt.title('Average Sales Amount by Month')
    plt.xlabel('Month')
    plt.ylabel('Average Total Sales Amount')
    plt.xticks(rotation=0)
    plt.show()
```

Figure 3.10 Pattern Recognition Analysis

3.7 Feature Engineering

Feature Engineering is an essential data preprocessing phase that is comprised of creating, transforming, and choosing the most significant variables (features) from raw data to optimize the accuracy and effectiveness of machine learning forecasting models. This phase is geared towards transforming raw Amway sales data into significant prediction features that reflect the underlying relations and trends necessary in sales forecasting.

Time-Based Feature Extraction was done by grabbing features like Sales Year, Sales Month, and Sales Day, and Sales Day of Week from the Date column. Such features enable the models to capture temporal nuances, such as vertical seasonality for example monthly pattern, horizontal sales trend inside a day, and weekday effect, that are important in the direct sale cycle.

The Return Status and Customer Demographics were represented through integrating Return_Status as transaction outcome indicator and Customer_Age representing stages of customer lifecycle and its influence on purchase behaviour.

Price-Based Features were constructed by Price_Per_Quantity, which normalizes the value of a sale by the number of units sold; and Avg_Item_Price_Order, which is the average price per item in each order. These help the model to comprehend the price sensitivity, and product's perceived value.

Customer Behaviour Metrics had been introduced for summarising customer activity and buying behaviour. These include:

Customer_Order_Count: total orders for the customer.

Customer_Total_Quantity: total quantity of item purchased per customer.

Customer_Total_Spend: total spend per customer.

Customer_Avg_Order_Value: average order value against a customer.

These metrics will help in predicting sales based on customer loyalty, spending habit, and order rate.

Recency Features were obtained by formulating Days_Since_Last_Purchase, which is the days since a customer’s last purchase. It is a significant feature for customer-purchase-cycle modelling, churn prediction, and repeat purchasing behaviour modelling.

A Seasonality Feature Engineering step was then applied, which also included features such as Sales Week of Year and Sales Quarter, "' The model could then learn these cyclic business patterns that we noticed such as peak sales quarters or weekly campaign effects. The Is Weekend feature was also designed as a tool to distinguish between weekend and weekday purchase behaviour that can be affected operationally and from a marketing perspective by sales patterns.

In total, 28 features were engineered after including this novel, newly established ones, making the original raw sales data rich with predictors, representing temporal trends, customer activities, pricing tactics, and transaction details as potential predictors. This extensive feature engineering results in a firm foundation for more powerful and precise machine learning models for Amway sales prediction.

3.8 Classification Models and Techniques

The concluding phase for producing sales forecasts is the application and evaluation of the data model using various statistical and machine learning approaches, including Linear Regression, Random Forest, LSTM, and ARIMA. Four machine learning and statistical procedures are applied in sales forecasting:

Linear Regression: A model that creates a straightforward relationship between features and sales, effective for analyzing the impact of variables such as price, seasonal fluctuations, and past sales data

Random Forest: A form of ensemble learning where multiple decision trees are integrated to forecast sales values, capable of efficiently managing non-linear relationships and feature interactions while offering resilient predictions against overfitting.

Long Short-Term Memory (LSTM): A deep learning intelligent retrieval architecture especially designed for consecutive data that captures periodic patterns and long-term reliable in sales time series data.

Autoregressive Integrated Moving Average (ARIMA): A conventional time series forecasting model that forecasts future sales based on trends, past values and seasonal behaviour through autoregressive and moving average components.

All the four models will be compared extensively based on measures like R-squared, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) in order to pick the best performance. Model performances are assessed based on the following measures:

a. R-squared (R^2): Coefficient of determination that shows the proportion of variance in sales data explained by the model
b. Root Mean Square Error (RMSE): Square root of the average squared differences, penalizing larger prediction errors.
c. Mean Absolute Error (MAE): Average absolute difference between predicted and actual sales values.
d. Mean Absolute Percentage Error (MAPE): Percentage-based error metric that measures prediction accuracy relative to actual sales values.

```python
# Calculate metrics
def calculate_metrics(y_true, y_pred):
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    mape = np.mean(np.abs((y_true - y_pred) / y_true)) * 100
    r2 = r2_score(y_true, y_pred)
    return mae, rmse, mape, r2

# Displays accuracy results and model performance
lr_mae, lr_rmse, lr_mape, lr_r2 = calculate_metrics(y_test, lr_pred)
print(f"Linear Regression - MAE: {lr_mae:.2f}, RMSE: {lr_rmse:.2f}, MAPE: {lr_mape:.2f}%, R2: {lr_r2:.3f}")

rf_mae, rf_rmse, rf_mape, rf_r2 = calculate_metrics(y_test, rf_pred)
print(f"Random Forest - MAE: {rf_mae:.2f}, RMSE: {rf_rmse:.2f}, MAPE: {rf_mape:.2f}%, R2: {rf_r2:.3f}")

lstm_mae, lstm_rmse, lstm_mape, lstm_r2 = calculate_metrics(lstm_y_test, lstm_pred)
print(f"LSTM - MAE: {lstm_mae:.2f}, RMSE: {lstm_rmse:.2f}, MAPE: {lstm_mape:.2f}%, R2: {lstm_r2:.3f}")

arima_mae, arima_rmse, arima_mape, arima_r2 = calculate_metrics(y_test, arima_pred)
print(f"ARIMA - MAE: {arima_mae:.2f}, RMSE: {arima_rmse:.2f}, MAPE: {arima_mape:.2f}%, R2: {arima_r2:.3f}")
```

Figure 3.11 Function to Calculate Forecasting Evaluation Metrics

3.9 Summary

This chapter outlines an integrated methodology to build precise sales forecasting models in direct selling industry. It incorporates conventional statistical methods with recent machine learning approaches to provide sound and dependable forecasting power. Data quality, feature construction, and careful assessment are stressed in the methodology to provide meaningful insights for business decisions. A systematic methodology ensures that the forecasting system remains flexible in response to evolving market conditions but continues to be precise and business oriented.