Explanation of ARIMAX and SARIMAX Models Implementation and Results

Steps for ARIMAX and SARIMAX Implementation

Step-by-Step Process:

1. **Feature Engineering**:
    - Select and create relevant features (such as technical indicators, lag features, etc.) from the raw data.
2. **Scaling the Data**:
    - Apply scaling techniques to standardize the features so that all variables are on the same scale.
3. **Data Splitting**:
    - Split the data into training and test sets (e.g., 80% for training, 20% for testing).
4. **Model Selection**:
    - Use **ARIMAX** (ARIMA + Exogenous Variables) for time series forecasting that incorporates external predictors. SARIMAX adds seasonality to ARIMAX, making it suitable for seasonal data.
5. **Model Tuning**:
    - Perform hyperparameter tuning using methods like grid search to find the best values for parameters such as p, d, q, P, D, and Q.
6. **Model Fitting**:
    - Fit the selected ARIMAX or SARIMAX model on the training data.
7. **Forecasting**:
    - Use the trained model to forecast future values (e.g., next 30 days).
8. **Evaluate Performance**:
    - Calculate forecast error metrics (MAE, MSE, RMSE, MAPE) to evaluate the model's performance.
9. **Residual Analysis**:
    - Check the residuals (errors) to ensure there is no pattern left, indicating a good model fit.

1. Feature Engineering

- **Feature engineering** is the process of selecting and transforming raw data into meaningful features that help improve model performance. For our dataset, we began by analyzing various financial and technical indicators like RSI, MACD, Volume, and VWAP, which are commonly used in time series forecasting. These features help capture trends, seasonality, and volatility, which are essential for predicting future values. By adding lag features (such as Close_Lag1) and rolling statistics (like Rolling_Mean_7), we enhance the model's ability to recognize patterns and make more accurate predictions.

2. Data Scaling

- **Data scaling** is crucial because it standardizes the range of features, preventing features with larger values from dominating the learning process. We scaled the data using techniques like Min-Max scaling or standardization to ensure that all features are on the same scale. This helps the model better capture relationships between features and improves performance. In this case, scaling the data made the models work more effectively, as many machine learning algorithms and statistical models are sensitive to the magnitude of the features.

3. Data Splitting

- Before training the models, we **split the dataset** into training and testing sets. This allows us to assess the model's ability to generalize to unseen data. Typically, the data is divided into:
  - **Training set**: Used to train the model (e.g., 80% of the data).
  - **Test set**: Used to evaluate the model's performance on unseen data (e.g., the remaining 20%).
- By doing this, we ensure that the model is not overfitting (i.e., memorizing the training data) and can predict future values accurately.

4. ARIMAX and SARIMAX Model Results

- **ARIMAX (AutoRegressive Integrated Moving Average with Exogenous variables)** and **SARIMAX (Seasonal ARIMAX)** are both statistical models used for time series forecasting. Here are the results of both models:
  - **ARIMAX Model**:
    - MAE: 0.0259
    - MSE: 0.00125
    - RMSE: 0.0353
    - MAPE: 1.24%
  - **SARIMAX Model**:
    - MAE: 0.0255
    - MSE: 0.00125
    - RMSE: 0.0354

        - MAPE: 1.21%

- These low error metrics indicate that both models are performing well, with **SARIMAX** slightly outperforming **ARIMAX** in terms of MAE and MAPE.
- The **MAPE** values (1.24% for ARIMAX and 1.21% for SARIMAX) demonstrate that the forecasts are highly accurate, with errors of around 1% on average.

Conclusion

- Both **ARIMAX** and **SARIMAX** models performed well, providing forecasts with very low errors (MAPE ~ 1%). **SARIMAX**, with its seasonal component, slightly outperforms **ARIMAX**, making it a better choice for this dataset, which might have seasonal patterns.
- These models are effective tools for forecasting time series data, especially when external factors (exogenous variables) are involved. Proper feature engineering, scaling, and model tuning are essential steps that significantly improve prediction accuracy.