1. Introduction

This solution aims to predict stock prices based on historical date values using a Long Short-Term Memory (LSTM) model. The problem at hand involves predicting the missing stock prices (NaN values) in a given test dataset, where the price values for specific dates are not available.

The steps in the solution include:

- Preprocessing the date data.
- Training the model on historical stock price data.
- Applying the trained model to predict the missing values in the test dataset.
- Storing the predicted values in a new CSV file.

---

2. Problem Statement

The given test.csv file contains two columns:

- date: Dates corresponding to stock prices.
- price: The stock prices corresponding to each date, with some values marked as NaN.

The objective is to fill in these missing (NaN) price values by predicting them using an LSTM-based model trained on past stock price data.

---

3. Data Overview

Test Data Format: The dataset contains two columns:

1. date - Date in the format DD-MM-YYYY.
2. price - Stock price (with missing values represented by NaN).

Example of Test Data:

date,price

01-10-2017,NaN

02-10-2017,NaN

03-10-2017,NaN

04-10-2017,NaN

05-10-2017,NaN

---

4. Methodology

4.1 Data Preprocessing

- **Date Handling**: The date column is first parsed and checked for correctness. In case of any formatting issues, extra spaces are removed, and the date values are converted to the appropriate format.
- **Missing Data**: Missing price values are handled by using the LSTM model to predict and replace these values.

- **Days Calculation**: A new column called days is added to the dataset, representing the number of days since the earliest date in the dataset. This is the main feature used by the model to predict the stock prices.

4.2 Scaling the Data

- The days column is scaled using a MinMaxScaler to ensure that the input data is normalized for the LSTM model. The scaler used for this transformation (scaler.pkl) was trained on the historical data and applied consistently to both training and test data.

4.3 LSTM Model

- The model used for stock price prediction is an LSTM model, which is effective for time-series data like stock prices.
- The LSTM model was trained on historical stock prices, where the input feature was the number of days (since the first date), and the output was the stock price.
- The trained model is loaded from a file (stock_price_model.h5) and used for generating predictions on the test dataset.

---

5. Steps for Prediction

1. **Load Test Data**:

- The test.csv file is loaded into a DataFrame and parsed with date handling.
- The date column is converted into a string format and stripped of any spaces.

2. **Add Days Column**:
    - The days column is calculated by finding the difference in days from the earliest date in the dataset.
3. **Scale the Data**:
    - The days column is scaled using the pre-trained MinMaxScaler.
4. **Prepare Data for LSTM**:
    - The scaled days values are reshaped to fit the input requirements of the LSTM model.
5. **Model Prediction**:
    - The LSTM model is used to predict the stock prices. The model generates predictions that are in the scaled form.
6. **Inverse Scaling**:
    - The predicted values are inverse-transformed to return them to the original price scale.
7. **Replace NaN Values**:

    - The missing price values (NaN) in the test dataset are replaced by the predicted stock prices.

8. **Save the Results**:
    - The updated dataset with the predicted values is saved to a new CSV file (PREDICTION_RESULT.CSV).

---

7. Results

The predictions for the stock prices are generated by the LSTM model and saved in the PREDICTION_RESULT.CSV file. The file contains the date and the predicted price values, replacing the original NaN values.

Example Output:

csv

Copy code

```
date,price
01-10-2017,150.25
02-10-2017,152.35
03-10-2017,153.50
04-10-2017,155.10
05-10-2017,157.45
...
```

---

8. Conclusion

This solution successfully predicts the missing stock prices based on historical date information using an LSTM-based model. The steps involved preprocessing the date data, scaling the features, training the model, and using it for prediction on test data. The final output is saved in a CSV file, providing predicted prices for the missing values.

---

9. Future Work

- **Model Improvement**: Experimenting with different architectures or hyperparameters for the LSTM model could improve prediction accuracy.
- **Additional Features**: Including other features like historical prices or external factors (e.g., market data, news) could enhance model performance.