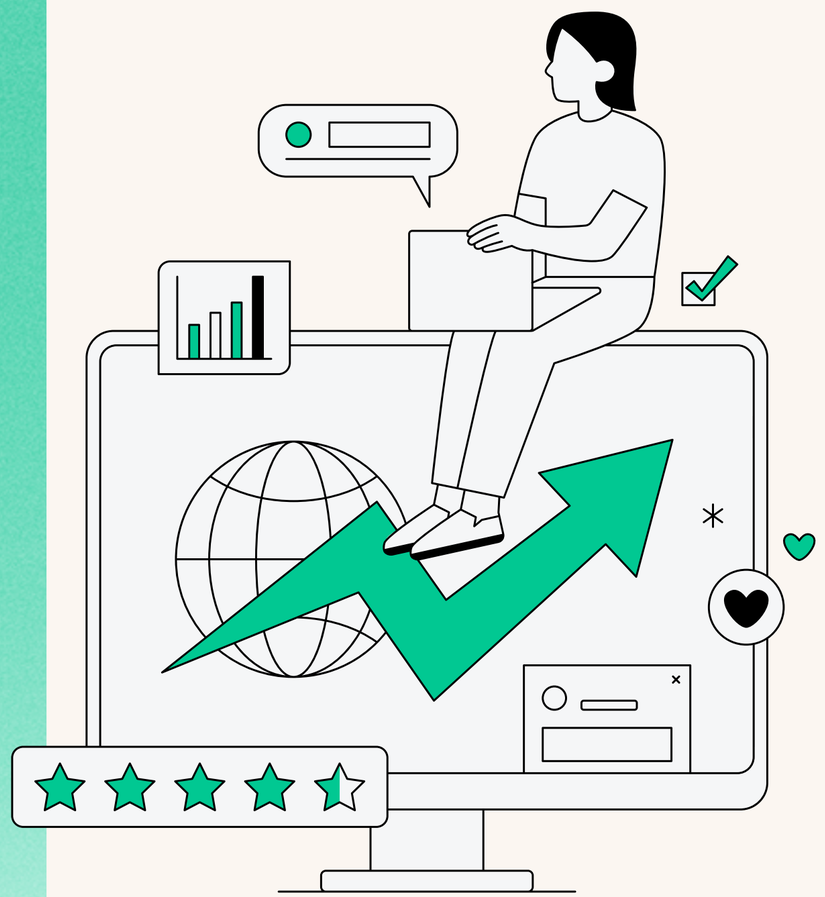
Minor Project-02

Stock Market Price Forecasting



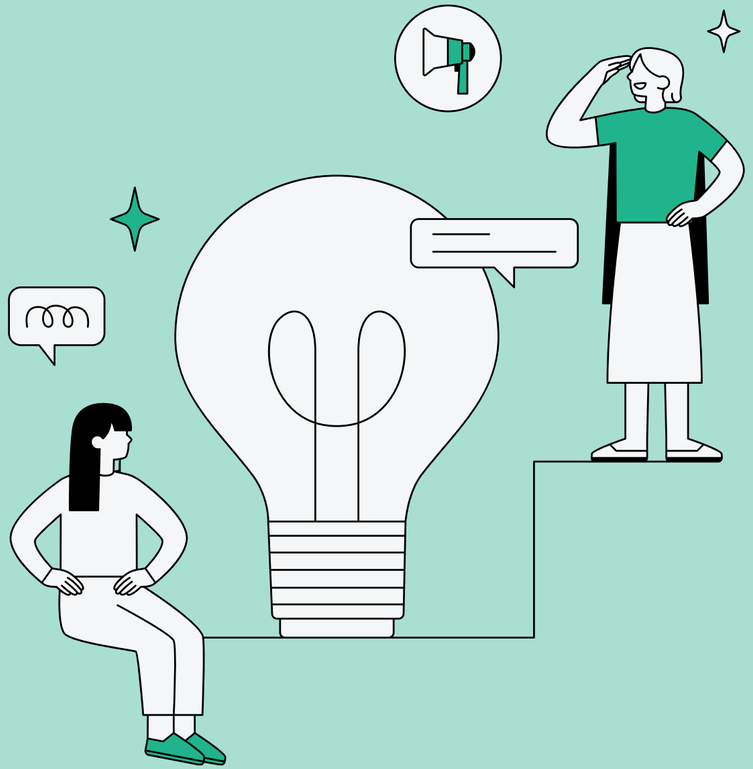
160121771076- Ch.Chandana
160121771088- P.Kathyayini
160121771095- A.Tejas

Contents

Introduction

- **Objective:** Analyze and Forecast stock market prices using data science and deep learning techniques.
- **Scope:** Enhance investment strategies and decision-making processes.
- **Key Components:**
  - **Data Preprocessing:** Clean and prepare financial data.
  - **Exploratory Data Analysis:** Uncover patterns and insights.
  - **Modeling:** Forecast stock performance.
- **Goal:** Develop a data-driven strategy for high-potential stock identification and optimal portfolio management.
- **Outcome:** Provide valuable tools and insights for investors to achieve superior risk-adjusted returns.

Dataset Description



Stock Index Considered for Analysis:

- BANKNIFTY : an index comprised of the most liquid and large capitalised Indian banking stocks.

Number of rows: Updating daily, around 4090 as of now.

Source: https://finance.yahoo.com/

Columns:

- Date: Given in "yyyy-mm-dd" format (eg: 2024-09-17)
- Open: The opening price of the stock for the corresponding date, given upto 2 decimal points
- Close: The closing price of the stock for the corresponding date, given upto 2 decimal points
- High: The highest price reached by the stock on the corresponding date, given upto 2 decimal points
- Low: The lowest price reached by the stock on the corresponding date, given upto 2 decimal points
- Adj. Close: The closing price after adjustments for all applicable splits and dividend distributions.
- Volumes: Number of options, contracts bought or sold on a given trading day

Tasks Overview

- **Preprocessing:**

1. Converting all columns to numeric datatype
2. Printing the summary of data
3. Counting missing values in data
4. Imputing missing values with moving averages
5. Replacing rows with value “0” in Volume column
6. Feature Engineering
7. Downloading the modified dataset

- **Data Analysis:**

1. Closing Price vs Volume Line Chart
2. Opening Price vs Closing Price Scatter Plot
3. Date vs Closing, Opening Price

Tasks Overview (Contd..)

- **Data Analysis:**

4. Date vs High, Low Price
5. Avg. Closing-Opening Difference per month
6. Avg. Same-day Point Difference per month
7. Avg. “Today point Difference“, “Prev. and today’s point Difference” Over Time
8. Candlestick graph of the year 202
9. Candlestick Feb-Apr 2020, Apr-Jun 2020
10. Candlestick graphs of Feb-Apr 2021,2022,2023

- **Modeling:**

1. ARIMA Model Forecasting
2. LSTM Model Forecasting

Preprocessing Tasks

1. Converting columns to numeric datatype

- Given attributes are “character” type, hence, to carry out tasks, all attributes need to be converted to numeric
- Function used: as.numeric and lapply

2. Summary of Data

- Computing key statistics and understanding the distribution of data in a dataset.

```
     Date                Open            High            Low
 Length:4090        Min.   : 3385   Min.   : 3447   Min.   : 3315
 Class :character   1st Qu.:10308   1st Qu.:10414   1st Qu.:10180
 Mode  :character   Median :18386   Median :18539   Median :18227
                    Mean   :20869   Mean   :21033   Mean   :20676
                    3rd Qu.:30228   3rd Qu.:30480   3rd Qu.:29960
                    Max.   :48880   Max.   :49057   Max.   :48669
                    NA's   :303     NA's   :303     NA's   :303
     Close          Adj.Close         Volume
 Min.   : 3340   Min.   : 3340   Min.   :0.000e+00
 1st Qu.:10289   1st Qu.:10289   1st Qu.:0.000e+00
 Median :18373   Median :18372   Median :0.000e+00
 Mean   :20856   Mean   :20856   Mean   :6.583e+05
 3rd Qu.:30216   3rd Qu.:30215   3rd Qu.:4.165e+04
 Max.   :48987   Max.   :48987   Max.   :1.798e+09
 NA's   :303     NA's   :303     NA's   :303
```

Preprocessing Tasks

3. Counting missing values

- An important step in data cleaning and preprocessing.
- Missing values can arise from various sources

1. data entry errors,
2. data corruption
3. unavailability of data.

- Identifying and handling these missing values is crucial for accurate data analysis and modeling.

4. Imputing missing values with moving averages

- Imputing missing values with moving averages

1. used in time series analysis
2. replace missing data points with the average of neighboring values.

- Maintains the temporal structure and trends within the data.

Preprocessing Tasks

5. Replacing “0” values in Volume Column

- As we are dealing with financial and time series data, zero values may indicate missing or incorrect data.
- Need to be corrected by understanding trends in data.

6. Feature Engineering

- To extract meaningful features helps to analyze the data better and improves model performance as well
- Two extra features are added:

1. Difference between Today’s Closing and Opening Prices
2. Difference between Yesterday’s Closing and Today’s Opening Price

- Provides significant insights in stock market prediction, forecasting and analyzing trading strategies.

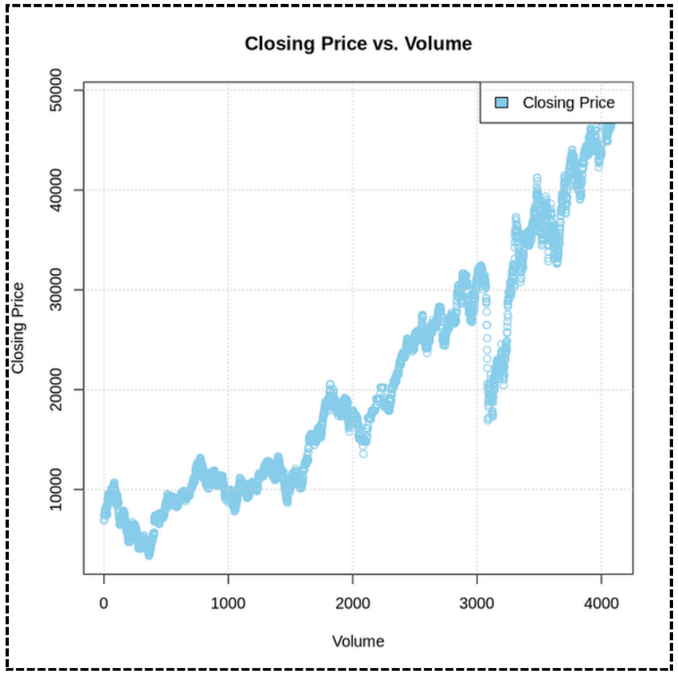
Preprocessing Tasks

7. Final modified dataset

Date	Open	High	Low	Close	Adj.Close	Volume	Today_point_difference	yesterday_	today_open	closing_opening_difference
17-09-2007	6898	6977.2	6843	6897.1	6897.02	388941	-0.899902	6897.1	6898	-0.899902
18-09-2007	6921.15	7078.95	6883.6	7059.65	7059.568	334057	138.5	6897.1	6921.1499	-24.049804
19-09-2007	7111	7419.35	7111	7401.85	7401.764	324021	290.850098	7059.65	7111	-51.350098
20-09-2007	7404.95	7462.9	7343.6	7390.15	7390.064	360996	-14.800293	7401.85	7404.9502	-3.100097
21-09-2007	7378.3	7506.35	7367.15	7464.5	7464.413	426317	86.200195	7390.15	7378.2998	11.850097
24-09-2007	7514.4	7661.05	7514.4	7650.9	7650.811	324506	136.5	7464.5	7514.3999	-49.899902
25-09-2007	7658.5	7694.25	7490.2	7629.15	7629.061	393626	-29.350098	7650.9	7658.5	-7.600098
26-09-2007	7647.1	7829.85	7591.8	7755.9	7755.81	245403	108.799804	7629.15	7647.1001	-17.950196
27-09-2007	7804.55	7866.5	7747.1	7833.65	7833.559	265160	29.100097	7755.9	7804.5498	-48.649903
28-09-2007	7838.25	8082.85	7836.05	8042.2	8042.107	259133	203.950195	7833.65	7838.25	-4.600098
01-10-2007	8008.55	8085.15	7913.3	7987.5	7987.407	383203	-21.049805	8042.2	8008.5498	33.65039
03-10-2007	8029.8	8235.8	7820.25	8097.9	8097.806	345254	68.100097	7987.5	8029.7998	-42.299805

Data Analysis Tasks

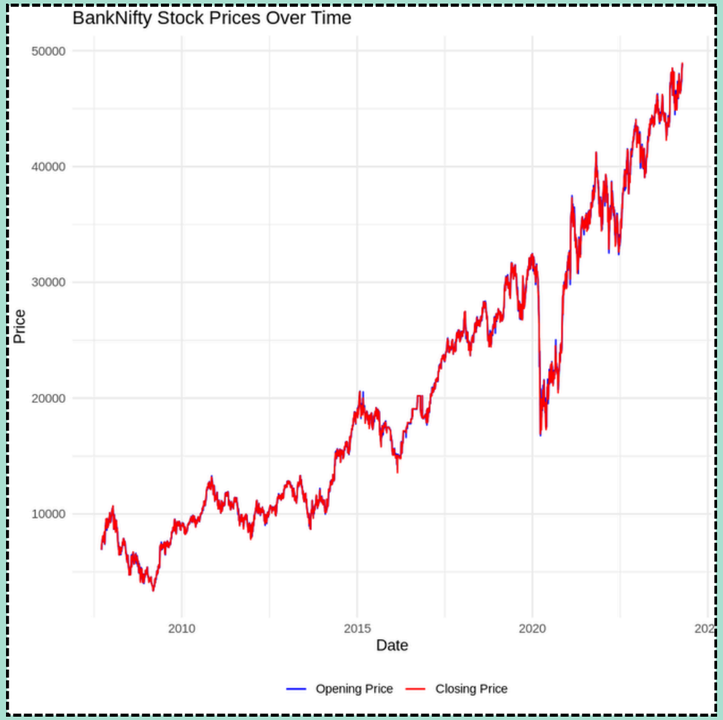
1. Closing Price vs Volume



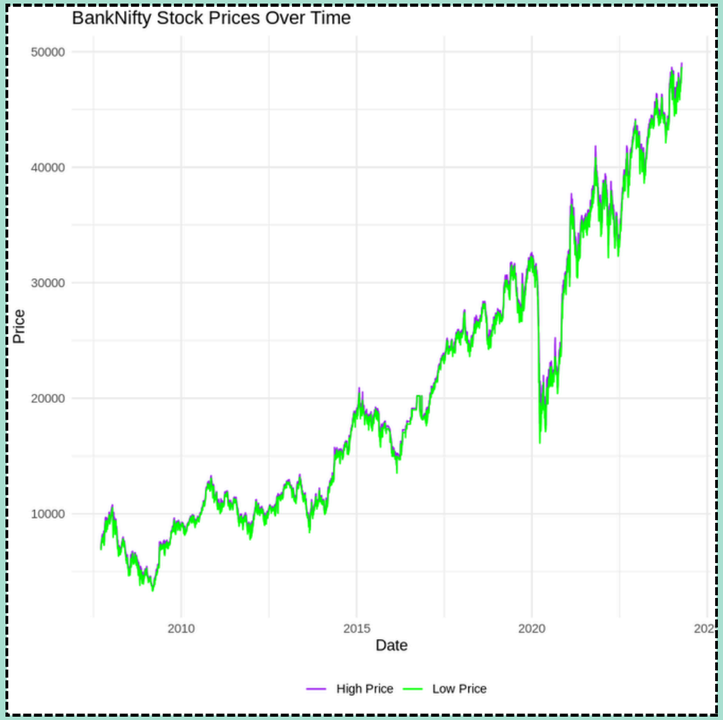
2. Opening Price vs Closing Price

Data Analysis Tasks

3. Date vs Closing, Opening Price

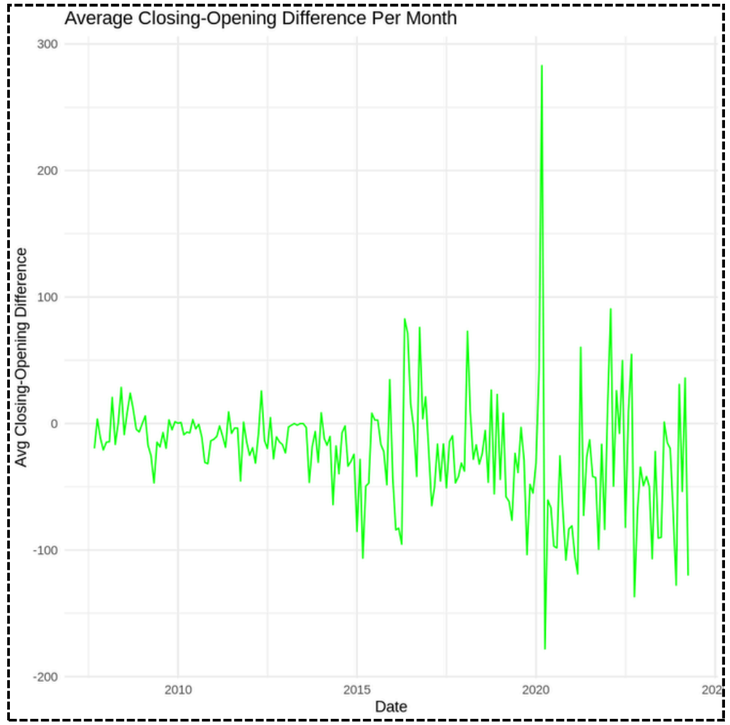


4. Date vs High, Low Price

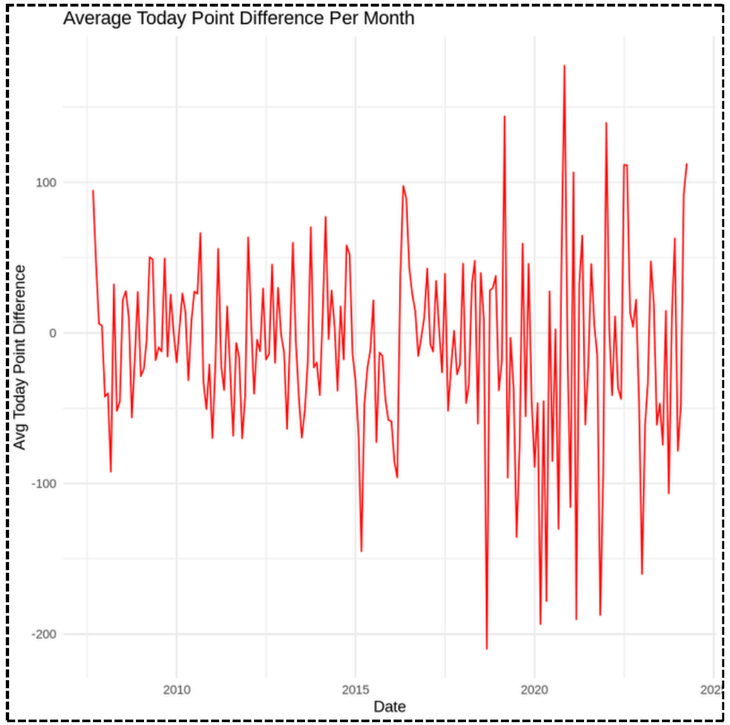


Data Analysis Tasks

5. Avg. Closing-Opening Diff. P.M.

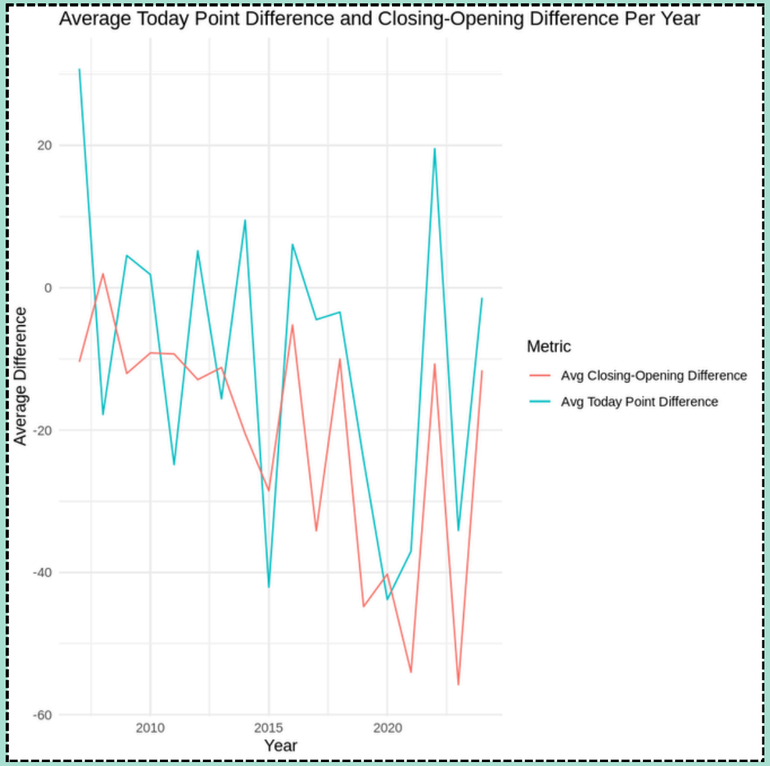


6. Avg. Today Point Diff. P.M.



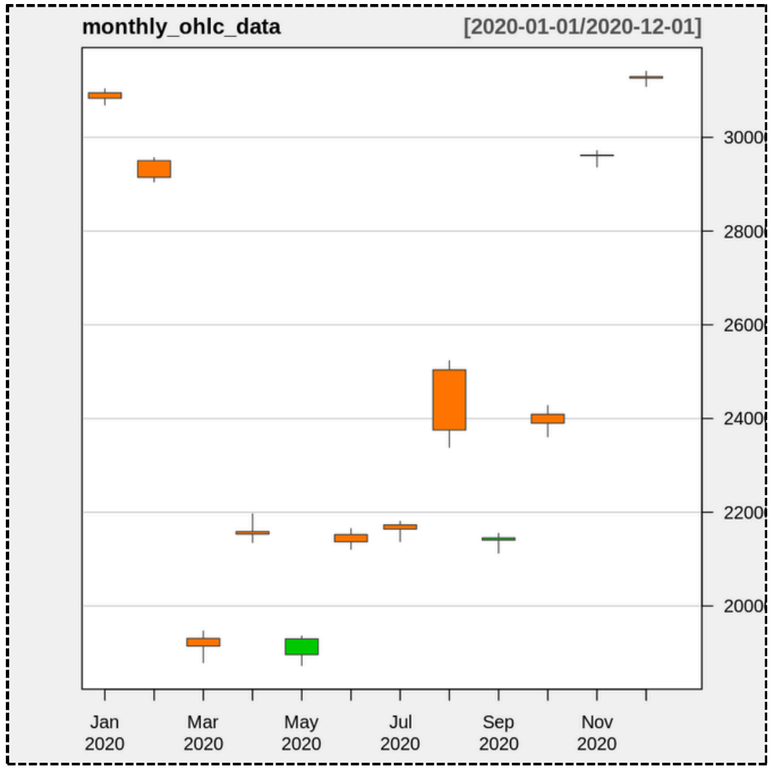
Data Analysis Tasks

7. Avg. Today Point Difference and Avg. Closing-Opening Difference Per Year



Data Analysis Tasks

8. Candlestick graph of the year 2020



Data Analysis Tasks

9. Candlestick Feb-Apr 2020

10. Candlestick Apr-Jun 2020

Data Analysis Tasks

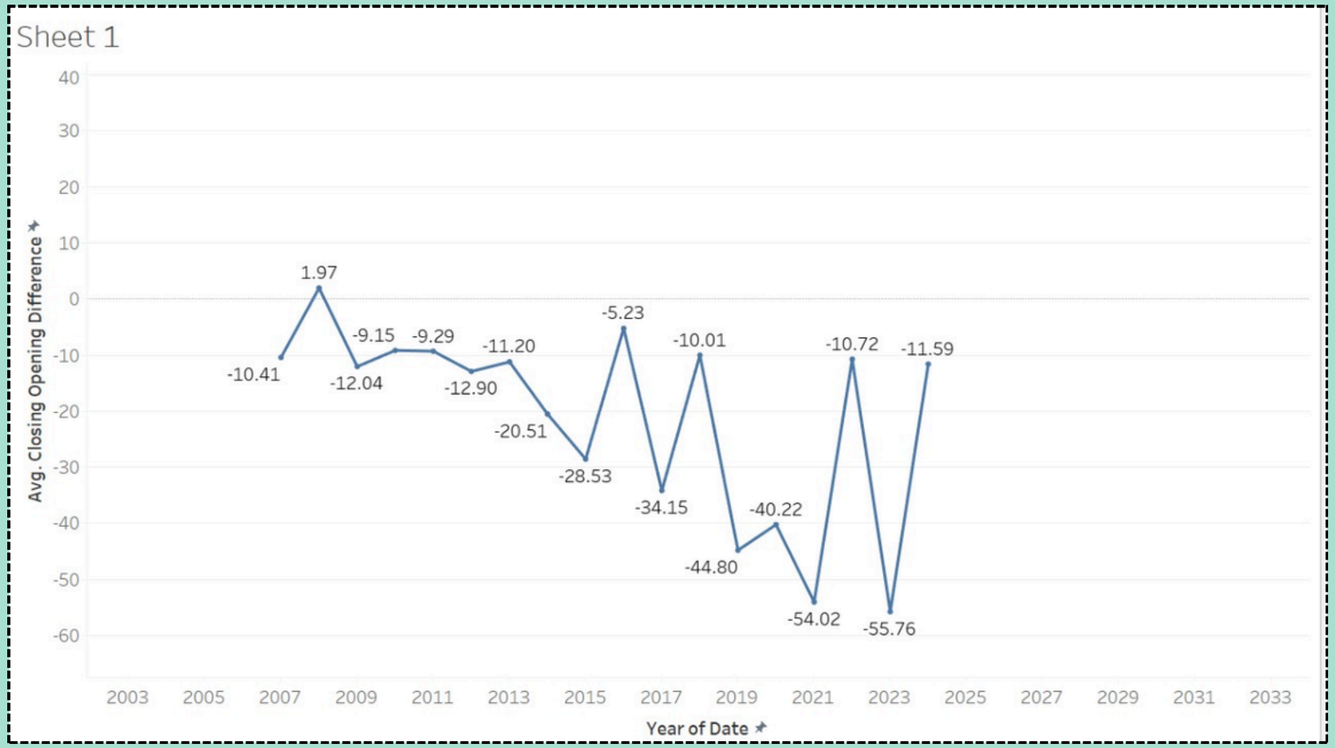
11.Feb-Apr 2021

12. Feb-Apr 2022

13. Feb-Apr 2023

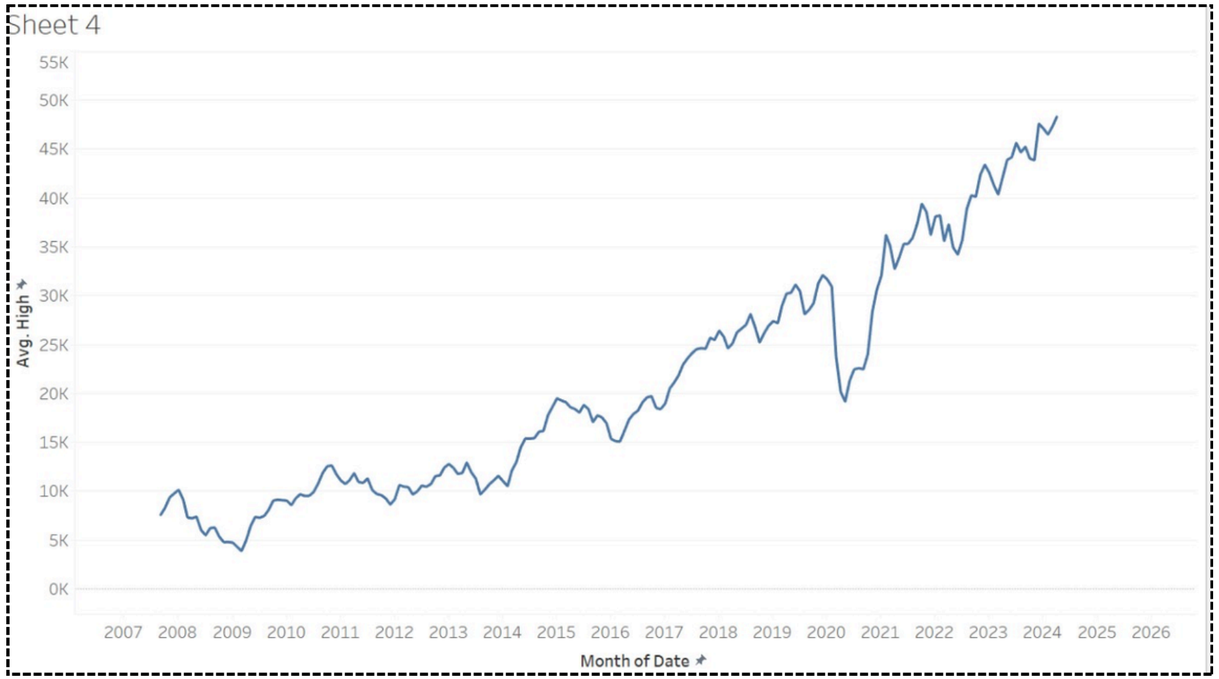
Data Analysis Tasks

14. Avg. Closing Opening Differences vs years



Data Analysis Tasks

15. Avg. High Price for each month over time



Modeling Tasks

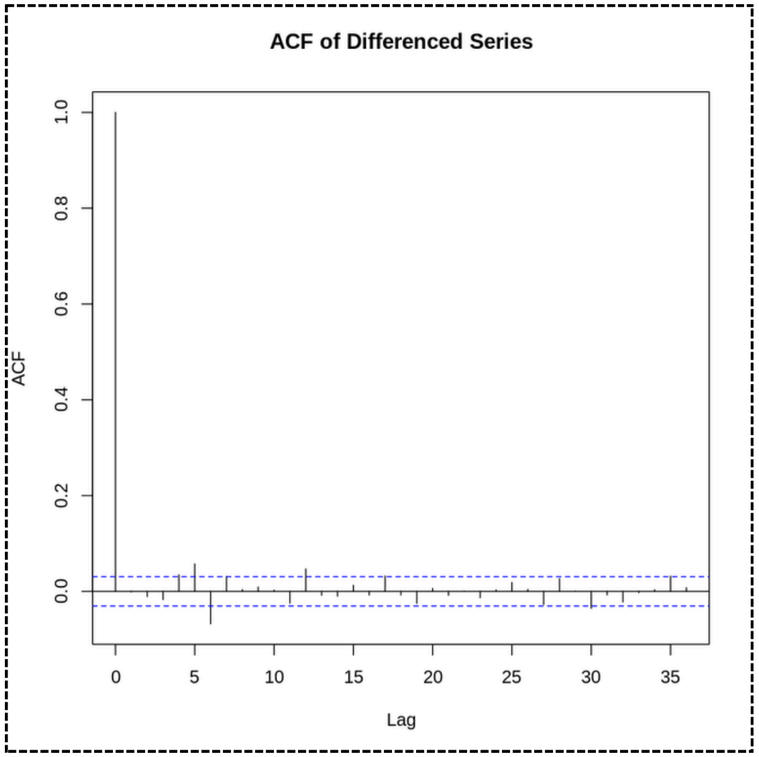
1. ARIMA (Auto-regressive Integrated Moving Averages)

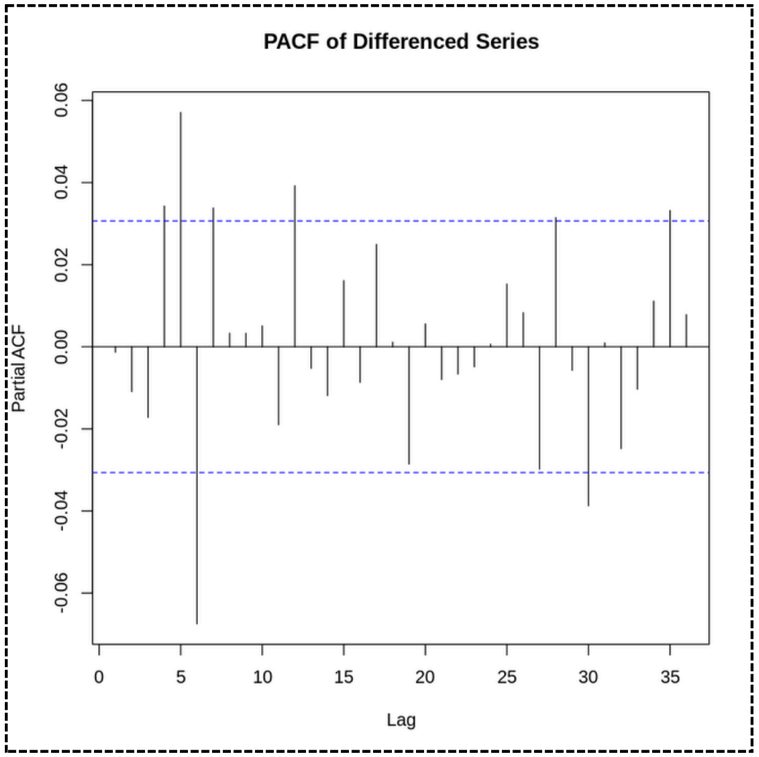
- a statistical analysis model that uses time series data
- helps better understand the data set or to predict future trends.
- Autoregressive - if it predicts future values based on past values.
- **Components:**

    1. Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
    2. Integrated (I): represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
    3. Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Modeling Tasks

ACF & PACF Graphs of the Time Series:





Modeling Tasks

Forecasted Output

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
4091	48996.89	48576.92	49416.87	48354.59	49639.20
4092	49007.19	48413.25	49601.13	48098.84	49915.54
4093	49017.48	48290.06	49744.90	47904.98	50129.98
4094	49027.77	48187.82	49867.73	47743.17	50312.38
4095	49038.07	48098.97	49977.17	47601.84	50474.30

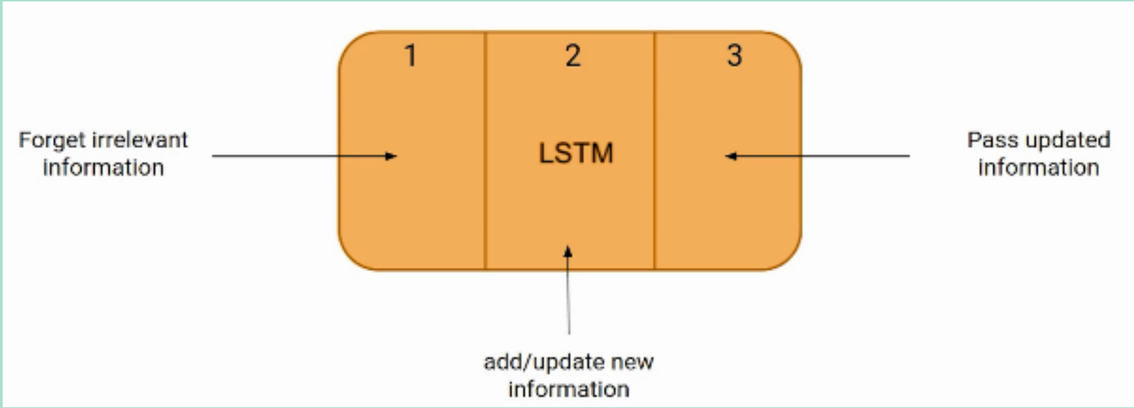
```
Forecasted closing prices for the next 5 days:
Time Series:
Start = 4091
End = 4095
Frequency = 1
[1] 48996.89 49007.19 49017.48 49027.77 49038.07
```

Modeling Tasks

2. LSTM (Long Short Term Memory)

- A deep learning, sequential neural network which is capable of handling the vanishing gradient problem faced by RNN.
- designed to avoid long-term dependency problems.

Architecture:



Modeling Tasks

Procedure Followed:

- Normalize the data
- Create sequences of data for LSTM
- Split data into training and testing sets
- Re-shape input data
- Prepare input and output variables
- Build the model
- Train the model
- Evaluate the model
- Make predictions
- Denormalize predictions
- Compare actual and predicted values
- Forecast future values

Forecasted Output

```
Forecasted closing prices for the next 5 days:
 48915.79 49088.34 49257.15 49426.4 49594.81
```

Modeling Tasks

Performance of ARIMA model:

```
Mean Absolute Error (MAE):  1452.682
Mean Squared Error (MSE):  2193842
Root Mean Squared Error (RMSE):  1481.162
Mean Absolute Percentage Error (MAPE):  3.057958
```

Performance of LSTM model:

```
Mean Absolute Error (MAE):  1691.7
Mean Squared Error (MSE):  2978181
Root Mean Squared Error (RMSE):  1725.741
Mean Absolute Percentage Error (MAPE):  3.559951
```

Modeling Tasks

Inferences:

1. The ARIMA model has a lower MAE compared to the LSTM model, indicating that, on average, the ARIMA model's forecasts are closer to the actual values.
2. The ARIMA model has a significantly lower MSE than the LSTM model. This suggests that the ARIMA model has smaller squared errors, meaning fewer large deviations from the actual values compared to the LSTM model.
3. The ARIMA model has a lower RMSE, indicating better overall accuracy since RMSE gives higher weight to larger errors.
4. The ARIMA model has a lower MAPE, indicating that its relative forecasting errors are smaller compared to the LSTM model.

Conclusion

- Engineered features improved model accuracy.
- Analyzed and forecasted Banknifty prices with machine learning.
- ARIMA and LSTM provided valuable price forecasts.
- Real-time predictions via deployed models and dashboards.
- Key insights: market sentiment and volatility management.
- Future work: integrate more data, refine models, explore advanced methods.

THANK YOU!