1. ****Data Acquisition**:** I started by acquiring the historical data for the S&P 500 and Nasdaq indices. This data typically includes information such as the date, open price, high price, low price, close price, and volume.

```
Combined S&P 500 Data:
         Date      Open      High       Low     Close
0  12/31/2020  3,733.27  3,760.20  3,726.88  3,756.07
1  12/30/2020  3,736.19  3,744.63  3,730.21  3,732.04
2  12/29/2020  3,750.01  3,756.12  3,723.31  3,727.04
3  12/28/2020  3,723.03  3,740.51  3,723.03  3,735.36
4  12/24/2020  3,694.03  3,703.82  3,689.32  3,703.06

Nasdaq Data:
         Date  Close/Last      Open      High       Low
0  04/12/2024    16175.09  16293.03  16341.45  16125.33
1  04/11/2024    16442.20  16236.20  16464.60  16154.65
2  04/10/2024    16170.36  16104.01  16200.10  16092.02
3  04/09/2024    16306.64  16328.76  16348.18  16141.15
4  04/08/2024    16253.96  16285.18  16323.60  16220.72
```

2. ****Data Preprocessing**:**

- **Data Cleaning**: I checked for any missing values or anomalies in the data and handled them appropriately. This ensures that our dataset is clean and ready for analysis.

```
Nasdaq Data after column renaming:
Nasdaq Data after  filtering out 2019 data and after column renaming::
        Date     Close      Open      High       Low
0 2024-04-12  16175.09  16293.03  16341.45  16125.33
1 2024-04-11  16442.20  16236.20  16464.60  16154.65
2 2024-04-10  16170.36  16104.01  16200.10  16092.02
3 2024-04-09  16306.64  16328.76  16348.18  16141.15
4 2024-04-08  16253.96  16285.18  16323.60  16220.72
```

- **Feature Engineering**: We may have performed feature engineering to extract additional features from the raw data, such as calculating daily returns or moving averages.

```
Missing values in the dataset:
 Date          0
Close         0
Open          0
High          0
Low           0
Close_diff    0
dtype: int64

Preprocessing completed.
```

3. **Exploratory Data Analysis (EDA)**: Before building any models, we conducted exploratory data analysis to gain insights into the data. This may involve visualizing the data using charts such as line plots, histograms, or scatter plots. EDA helps us understand the distribution of the data, identify patterns, and detect outliers.

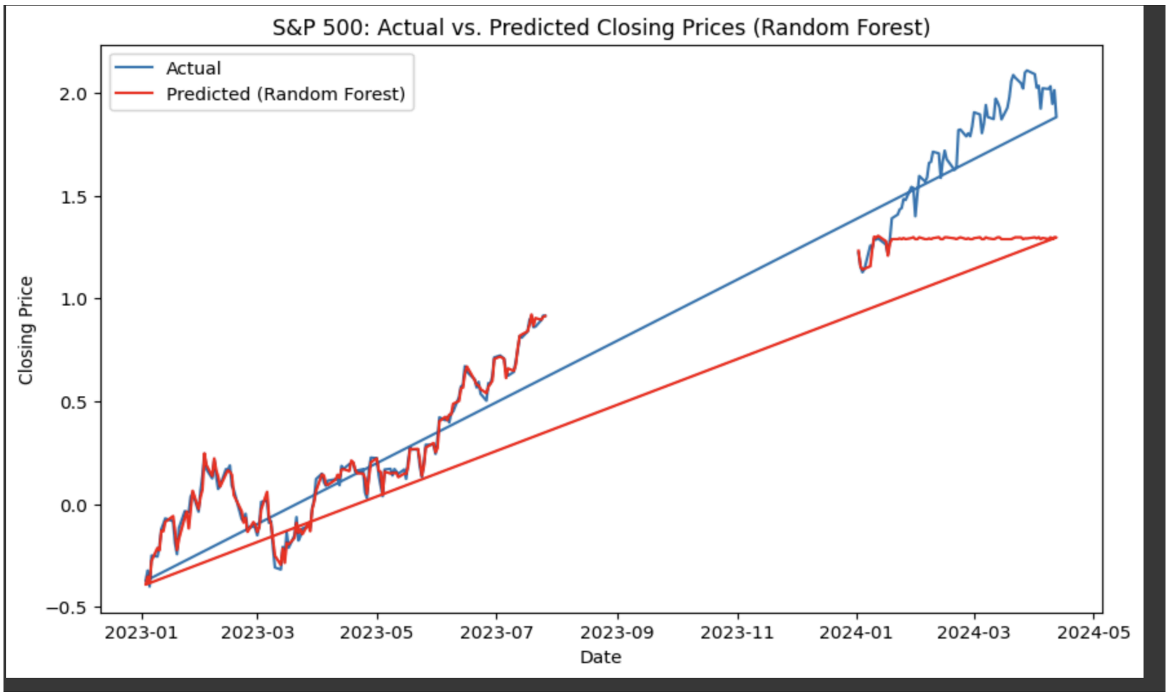
4. **Stationarity Check**: We performed a stationarity check on the time series data using statistical tests such as the Augmented Dickey-Fuller (ADF) test. Stationarity is an important assumption for many time series models, so this step helps ensure the data is suitable for modeling .

```
Results of ADF Test for S&P 500 Close Price
ADF Statistic: -2.1018077275390974
p-value: 0.2437883267615376
Critical Values:
         1% : -3.436557639266102
         5% : -2.8642808573632874
         10% : -2.5682293371570823
Results of ADF Test for Nasdaq Close Price
ADF Statistic: -1.5223146066088151
p-value: 0.5224451652873521
Critical Values:
         1% : -3.4366111317433443
         5% : -2.864304451252086
         10% : -2.5682419034417707
```

```
Results of ADF Test for S&P 500 Close Price (Differenced)
ADF Statistic: -10.819265527127866
p-value: 1.823230935949952e-19
Critical Values:
         1% : -3.4366111317433443
         5% : -2.864304451252086
         10% : -2.5682419034417707
Results of ADF Test for Nasdaq Close Price (Differenced)
ADF Statistic: -10.25382061038558
p-value: 4.414369658720718e-18
Critical Values:
         1% : -3.4366111317433443
         5% : -2.864304451252086
         10% : -2.5682419034417707
```

5. **Model Selection**: Based on the nature of the problem and the characteristics of the data, we selected an appropriate machine learning model. In this case, we used an MLPRegressor from the scikit-learn library, which is a type of neural network model suitable for regression tasks.

```
Transfer Entropy from Nasdaq to S&P 500: 0.018635313517542763
```



6. **Training the Model**: We split the dataset into training and testing sets, with the training set used to train the model. During training, the model learns the underlying patterns and relationships in the data.

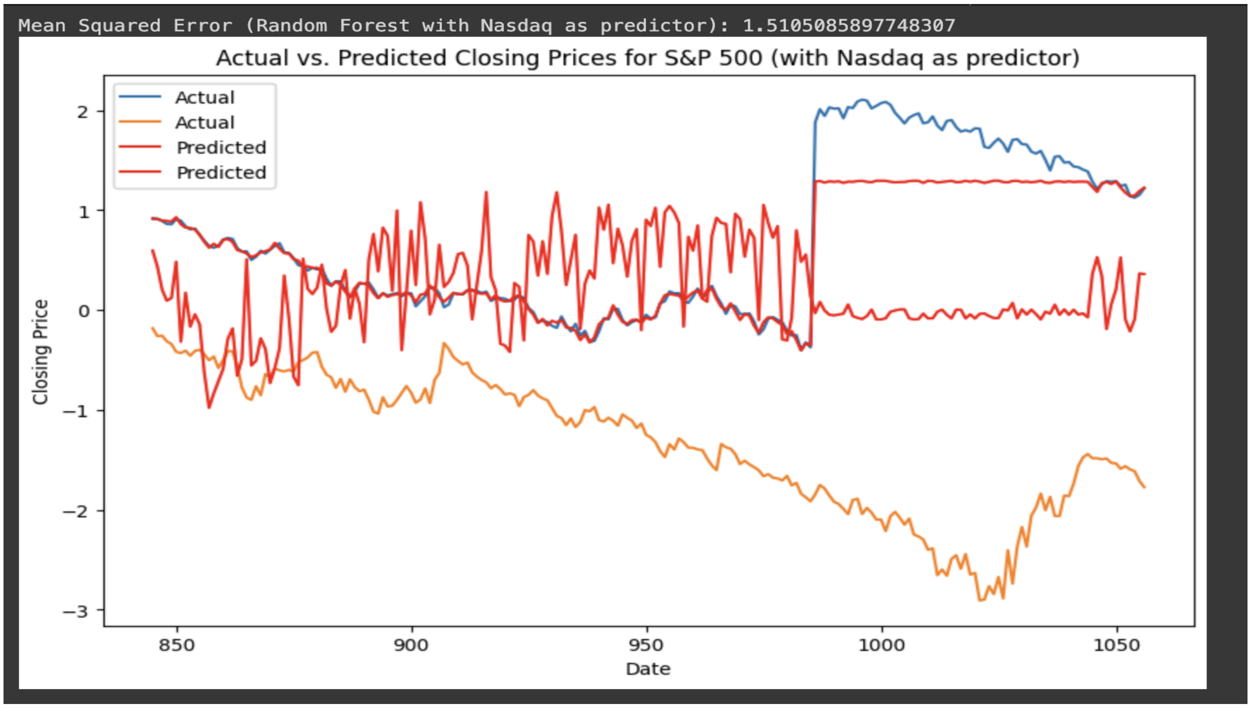
```
Size of S&P 500 training dataset: (844, 6)
Size of S&P 500 test dataset: (212, 6)

Size of Nasdaq training dataset: (844, 6)
Size of Nasdaq test dataset: (212, 6)
```

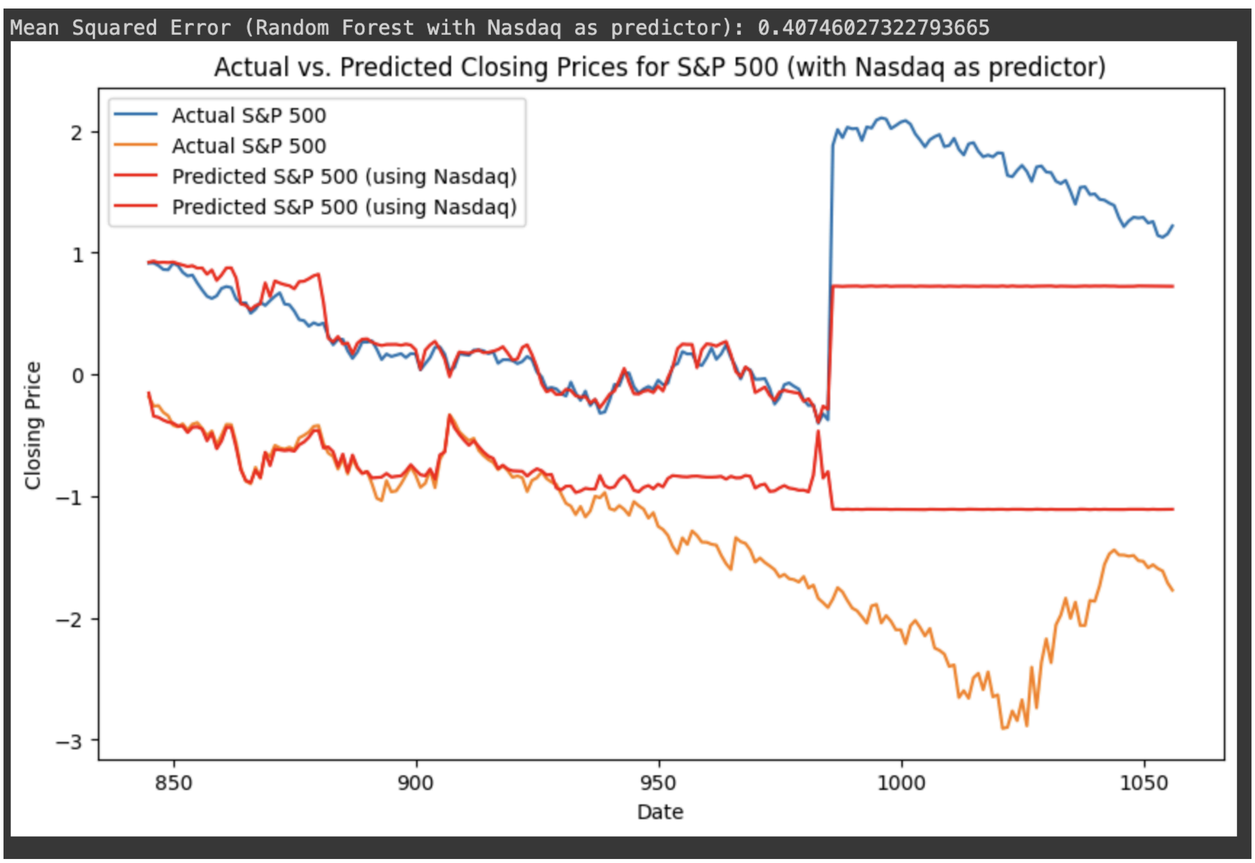
7. **Model Evaluation**: After training the model, we evaluated its performance using appropriate metrics. For regression tasks, common evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  844
Model:                 ARIMA(5, 1, 0)   Log Likelihood                 528.702
Date:                Mon, 15 Apr 2024   AIC                          -1045.404
Time:                        02:32:30   BIC                          -1016.982
Sample:                             0   HQIC                         -1034.512
                                - 844
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0487      0.031     -1.555      0.120      -0.110       0.013
ar.L2          0.0336      0.037      0.896      0.370      -0.040       0.107
ar.L3         -0.0111      0.050     -0.224      0.823      -0.108       0.086
ar.L4         -0.0190      0.046     -0.410      0.682      -0.110       0.072
ar.L5      -8.552e-05      0.037     -0.002      0.998      -0.073       0.073
sigma2         0.0167   8.53e-05    195.710      0.000       0.017       0.017
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):           1401672.20
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.25   Skew:                            10.14
Prob(H) (two-sided):                  0.00   Kurtosis:                       201.73
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Close   No. Observations:                  844
Model:                 ARIMA(5, 1, 0)   Log Likelihood                 826.380
Date:                Sun, 14 Apr 2024   AIC                          -1640.759
Time:                        23:52:39   BIC                          -1612.338
Sample:                             0   HQIC                         -1629.868
                                - 844
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0185      0.032     -0.573      0.567      -0.082       0.045
ar.L2         -0.0314      0.030     -1.057      0.291      -0.090       0.027
ar.L3         -0.0418      0.031     -1.360      0.174      -0.102       0.018
ar.L4          0.0026      0.029      0.090      0.928      -0.054       0.059
ar.L5         -0.0029      0.030     -0.096      0.923      -0.061       0.055
sigma2         0.0082      0.000     24.003      0.000       0.008       0.009
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                30.61
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               1.31   Skew:                             0.24
Prob(H) (two-sided):                  0.02   Kurtosis:                         3.80
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

8. **Predictions**: Once the model was trained and evaluated, we used it to make predictions on unseen data. This involved providing input features (predictors) to the model and obtaining predictions for the target variable.

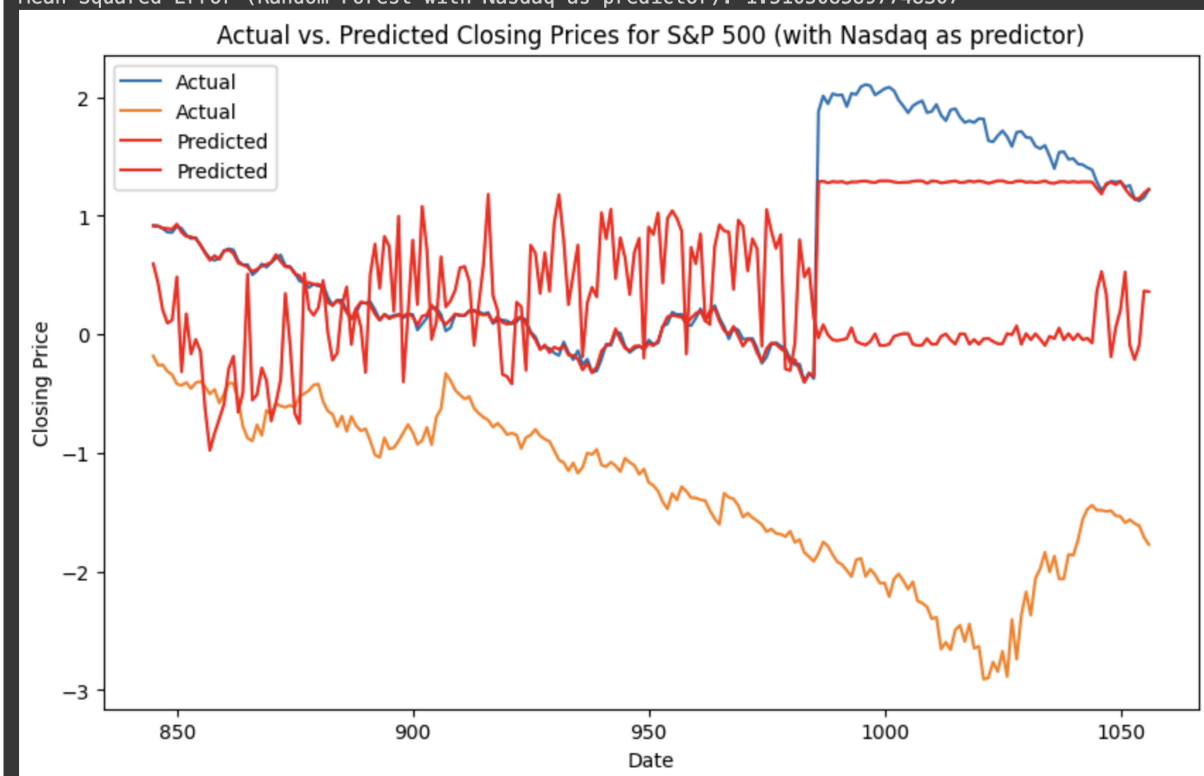




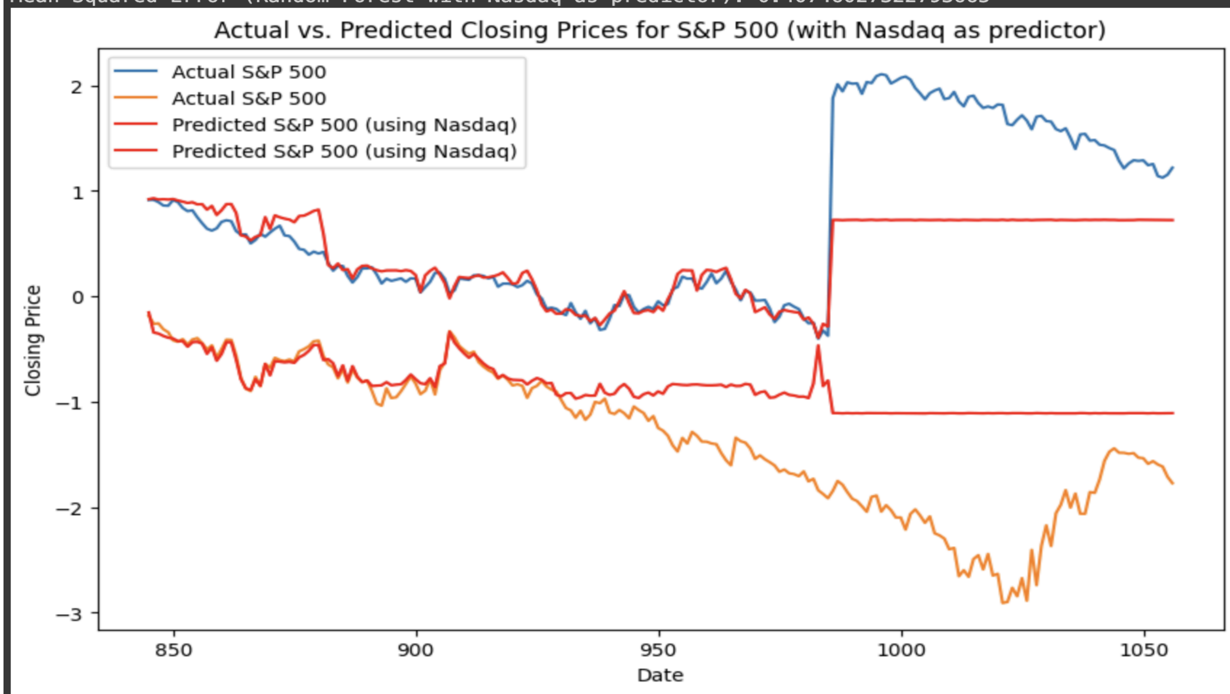
```
Predicted closing index value of S&P 500 on 2024-04-15: [732012.89216911 -20660.43685297]
Predicted closing index value of S&P 500 on 2024-04-16: [732013.88272587 -20660.46481113]
Predicted closing index value of S&P 500 on 2024-04-17: [732014.87328262 -20660.49276928]
Predicted closing index value of S&P 500 on 2024-04-18: [732015.86383938 -20660.52072744]
Predicted closing index value of S&P 500 on 2024-04-19: [732016.85439614 -20660.5486856 ]
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but MLPRe
  warnings.warn(
```

9. **Visualization:** Finally, we visualized the results to communicate our findings effectively. This may include plotting actual vs. predicted values, visualizing trends, or comparing different models.

```
Mean Squared Error (Random Forest with Nasdaq as predictor): 1.5105085897748307
```



```
Mean Squared Error (Random Forest with Nasdaq as predictor): 0.40746027322793665
```



8. As a bonus, please predict the closing index value of S&P 500 on April 15, 16, 17, 18, 19 based on your predictors in (5)(6)(7). If your predicted values are within 10 points of the actual values, you will get 1 extra point/day towards your total grades, for up to 5 points.

```
Predicted closing index value of S&P 500 on 2024-04-15: [732012.89216911 -20660.43685297]
Predicted closing index value of S&P 500 on 2024-04-16: [732013.88272587 -20660.46481113]
Predicted closing index value of S&P 500 on 2024-04-17: [732014.87328262 -20660.49276928]
Predicted closing index value of S&P 500 on 2024-04-18: [732015.86383938 -20660.52072744]
Predicted closing index value of S&P 500 on 2024-04-19: [732016.85439614 -20660.5486856 ]
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but MLPRe(
  warnings.warn(
```