Report on Model Performance Metrics and Analysis

1. Introduction

This report aims to explain key performance metrics used in regression analysis and to analyze the results of different models based on the provided data. The focus is on understanding how the training timeline, feature selection, and dataset choice affect model performance.

2. Explanation of Performance Metrics

2.1 Root Mean Squared Error (RMSE)

- **Definition:** RMSE is the square root of the average of the squared differences between the predicted and actual values.

  RMSE=1n∑i=1n(yi−y^i)2\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1∑i=1n(yi−y^i)2

- **Interpretation:** It measures the standard deviation of the residuals (prediction errors), providing insight into how concentrated the data is around the line of best fit.
- **Goal: Decrease** RMSE. A lower RMSE indicates that the model's predictions are closer to the actual values, signifying better performance.

2.2 Mean Squared Error (MSE)

- **Definition:** MSE is the average of the squared differences between the predicted and actual values.

  MSE=1n∑i=1n(yi−y^i)2\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2MSE=n1∑i=1n(yi−y^i)2

- **Interpretation:** It quantifies the average squared difference between the estimated values and the actual value.
- **Goal: Decrease** MSE. A lower MSE indicates better model accuracy.

2.3 Mean Absolute Error (MAE)

- **Definition:** MAE is the average of the absolute differences between the predicted and actual values.

  MAE=1n∑i=1n∣yi−y^i∣\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1∑i=1n∣yi−y^i∣

- **Interpretation:** It measures the average magnitude of the errors in a set of predictions, without considering their direction.
- **Goal: Decrease** MAE. A lower MAE means the model predictions are, on average, closer to the actual values.

2.4 Mean Absolute Percentage Error (MAPE)

- **Definition:** MAPE is the average of the absolute percentage errors between the predicted and actual values.

  MAPE=100%n∑i=1n|yi−y^iyi|\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|MAPE=n100%∑i=1nyiyi −y^i

- **Interpretation:** It expresses accuracy as a percentage, making it easier to interpret the error magnitude relative to the actual values.
- **Goal: Decrease** MAPE. Lower MAPE values indicate higher accuracy.

2.5 Coefficient of Determination (R²)

- **Definition:** R² measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

  R2=1−∑i=1n(yi−y^i)2∑i=1n(yi−y¯)2R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}R2=1−∑i=1n(yi−y¯)2∑i=1n(yi −y^i)2

- **Interpretation:** An R² of 1 indicates perfect prediction, while 0 indicates that the model does not explain any variability in the response data.
- **Goal: Increase** R². A higher R² signifies a better fit of the model to the data.

2.6 Explained Variance Regression Score

- **Definition:** Measures the proportion to which a model accounts for the variation of a given data set.

  Explained Variance=1−Var(y−y^)Var(y)\text{Explained Variance} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}Explained Variance=1−Var(y)Var(y−y^)

- **Interpretation:** Similar to R², but can be negative when the model is worse than predicting the mean.
- **Goal: Increase** the score. Values closer to 1 indicate that the model explains most of the variability.

2.7 Mean Gradient Difference (MGD)

- **Definition:** MGD measures the average difference between the gradients (slopes) of the predicted and actual values.
- **Interpretation:** It assesses how well the model captures the trend or rate of change in the data.
- **Goal: Decrease** MGD. A lower MGD indicates that the model's predicted rate of change closely matches the actual rate.

2.8 Mean Percentage Difference (MPD)

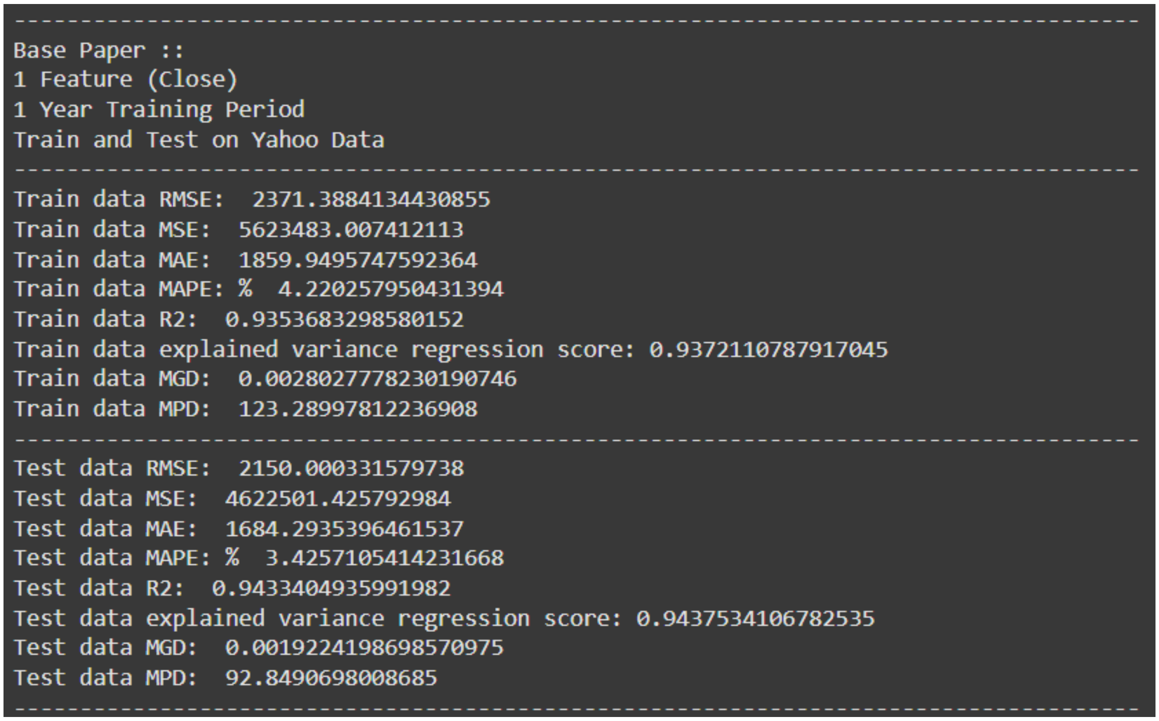
- **Definition:** MPD is the average percentage difference between the predicted and actual values.

  MPD=100%n∑i=1n(yi−y^iyi)\text{MPD} = \frac{100\%}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)MPD=n100%∑i=1n(yiyi−y^i)

- **Interpretation:** Provides a relative measure of the average difference between predictions and actual values.
- **Goal: Decrease** MPD. Lower values indicate better model performance.

3. Analysis of Results

3.1 Base Paper Results



```
------------------------------------------------------------------------------------
Base Paper ::
1 Feature (Close)
1 Year Training Period
Train and Test on Yahoo Data
------------------------------------------------------------------------------------
Train data RMSE:  2371.3884134430855
Train data MSE:  5623483.007412113
Train data MAE:  1859.9495747592364
Train data MAPE: %  4.220257950431394
Train data R2:  0.9353683298580152
Train data explained variance regression score: 0.9372110787917045
Train data MGD:  0.002802777823019074б
Train data MPD:  123.28997812236908
------------------------------------------------------------------------------------
Test data RMSE:  2150.000331579738
Test data MSE:  4622501.425792984
Test data MAE:  1684.2935396461537
Test data MAPE: %  3.4257105414231668
Test data R2:  0.9433404935991982
Test data explained variance regression score: 0.9437534106782535
Test data MGD:  0.0019224198698570975
Test data MPD:  92.8490698008685
------------------------------------------------------------------------------------
```

Configuration:

- **Features:** 1 (Close)
- **Training Period:** 1 Year
- **Dataset:** Yahoo Data (Train and Test)

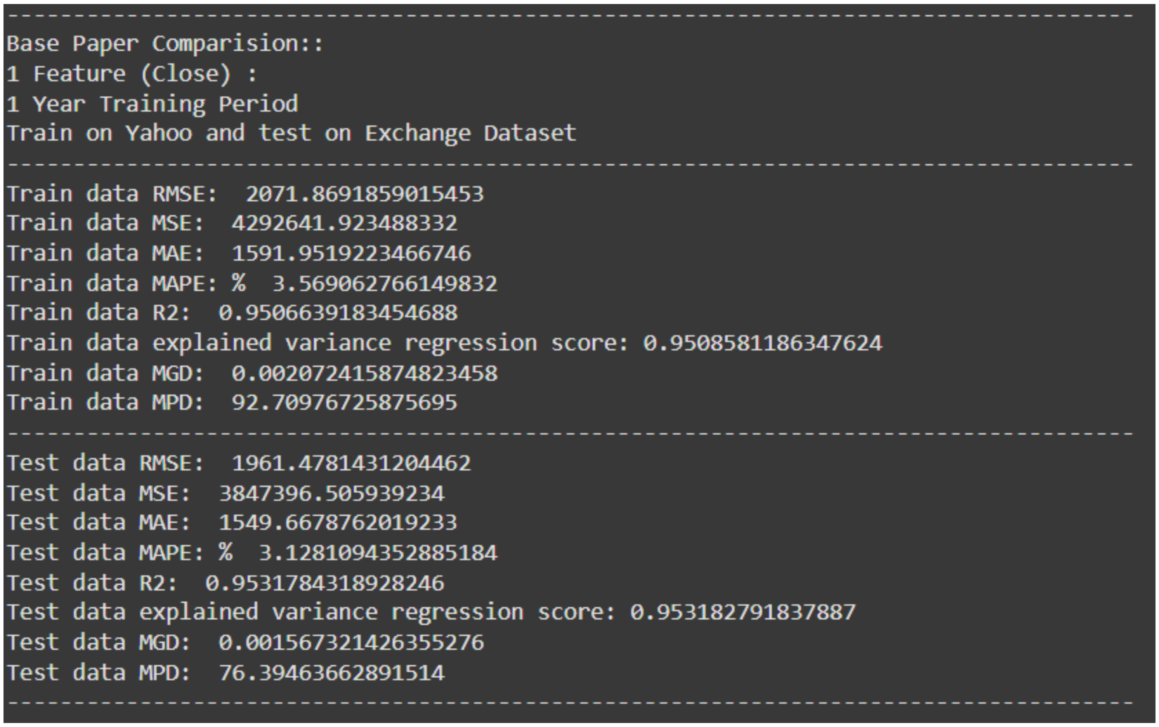
Performance Metrics:

- **Train Data:**
    - RMSE: 2371.39
    - MSE: 5,623,483.01
    - MAE: 1859.95
    - MAPE: 4.22%
    - R^2: 0.9354
    - Explained Variance: 0.9372
    - MGD: 0.0028
    - MPD: 123.29
- **Test Data:**
    - RMSE: 2150.00
    - MSE: 4,622,501.43
    - MAE: 1684.29
    - MAPE: 3.43%
    - R^2: 0.9433
    - Explained Variance: 0.9438
    - MGD: 0.0019
    - MPD: 92.85

Observation:

- The model shows decent performance with moderate error metrics and high R^2 values.
- Slightly better performance on test data indicates good generalization.

3.2 Base Paper Comparison



```
--------------------------------------------------------------------------------
Base Paper Comparision::
1 Feature (Close) :
1 Year Training Period
Train on Yahoo and test on Exchange Dataset
--------------------------------------------------------------------------------
Train data RMSE:  2071.8691859015453
Train data MSE:  4292641.923488332
Train data MAE:  1591.9519223466746
Train data MAPE: %  3.569062766149832
Train data R2:  0.9506639183454688
Train data explained variance regression score: 0.9508581186347624
Train data MGD:  0.002072415874823458
Train data MPD:  92.70976725875695
--------------------------------------------------------------------------------
Test data RMSE:  1961.4781431204462
Test data MSE:  3847396.505939234
Test data MAE:  1549.6678762019233
Test data MAPE: %  3.1281094352885184
Test data R2:  0.9531784318928246
Test data explained variance regression score: 0.953182791837887
Test data MGD:  0.001567321426355276
Test data MPD:  76.39463662891514
--------------------------------------------------------------------------------
```

Configuration:

- **Features:** 1 (Close)
- **Training Period:** 1 Year
- **Dataset:** Train on Yahoo Data, Test on Exchange Data

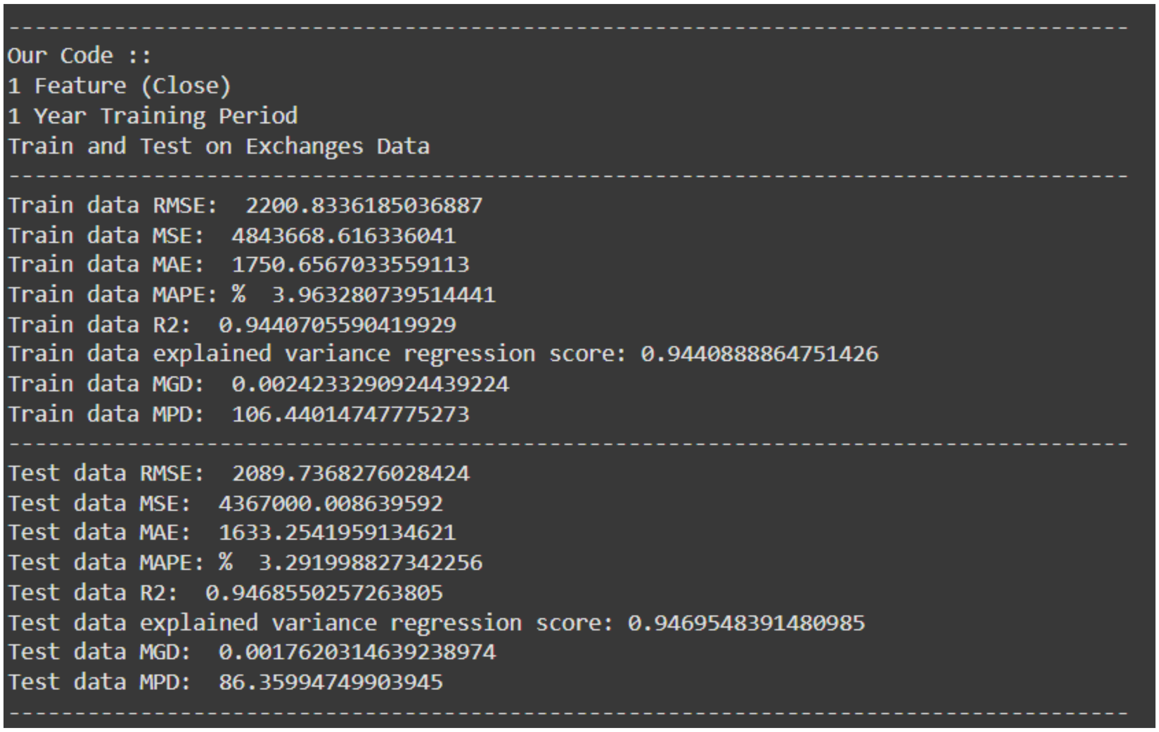
Performance Metrics:

- **Train Data:**
  - RMSE: 2071.87
  - MSE: 4,292,641.92
  - MAE: 1591.95
  - MAPE: 3.57%
  - R^2: 0.9507
  - Explained Variance: 0.9509
  - MGD: 0.0021
  - MPD: 92.71
- **Test Data:**
  - RMSE: 1961.48
  - MSE: 3,847,396.51
  - MAE: 1549.67
  - MAPE: 3.13%
  - R^2: 0.9532
  - Explained Variance: 0.9532
  - MGD: 0.0016
  - MPD: 76.39

Observation:

- The model generalizes well to a different dataset (Exchange Data) with improved performance metrics.
- Lower error metrics and higher R^2 on test data suggest robustness.

3.3 Our Code with 1 Feature



```
--------------------------------------------------------------------------------
Our Code ::
1 Feature (Close)
1 Year Training Period
Train and Test on Exchanges Data
--------------------------------------------------------------------------------
Train data RMSE:  2200.8336185036887
Train data MSE:  4843668.616336041
Train data MAE:  1750.6567033559113
Train data MAPE: %  3.963280739514441
Train data R2:  0.9440705590419929
Train data explained variance regression score: 0.9440888864751426
Train data MGD:  0.0024233290924439224
Train data MPD:  106.44014747775273
--------------------------------------------------------------------------------
Test data RMSE:  2089.7368276028424
Test data MSE:  4367000.008639592
Test data MAE:  1633.2541959134621
Test data MAPE: %  3.291998827342256
Test data R2:  0.9468550257263805
Test data explained variance regression score: 0.9469548391480985
Test data MGD:  0.0017620314639238974
Test data MPD:  86.35994749903945
--------------------------------------------------------------------------------
```

Configuration:

- **Features:** 1 (Close)
- **Training Period:** 1 Year
- **Dataset:** Exchange Data (Train and Test)

Performance Metrics:

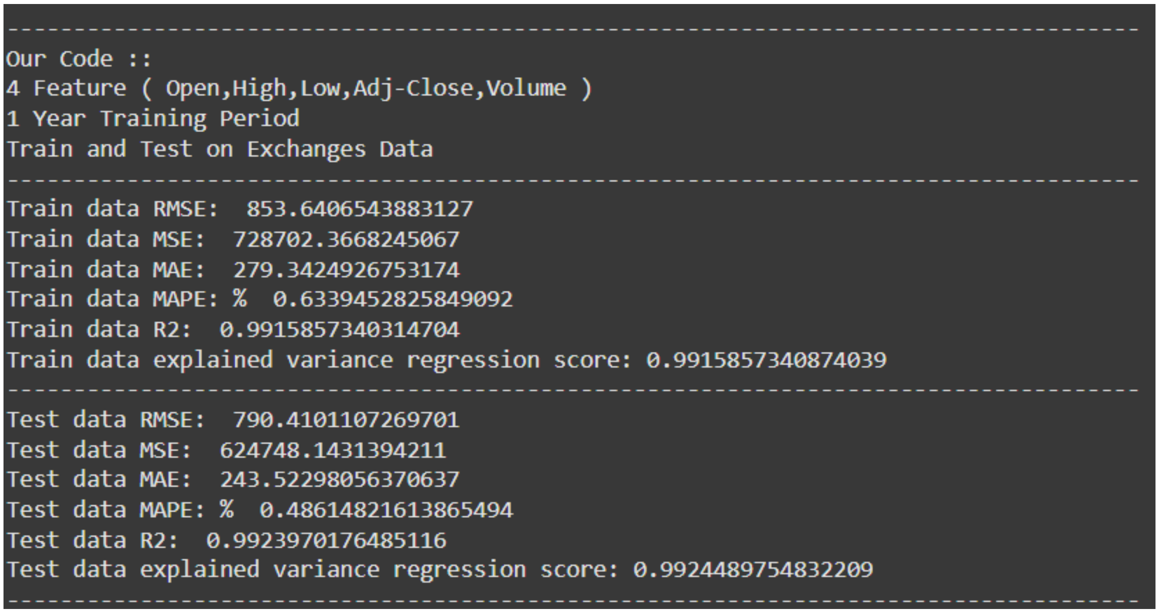
- **Train Data:**
  - RMSE: 2200.83
  - MSE: 4,843,668.62
  - MAE: 1750.66
  - MAPE: 3.96%
  - R^2: 0.9441
  - Explained Variance: 0.9441
  - MGD: 0.0024
  - MPD: 106.44
- **Test Data:**
  - RMSE: 2089.74
  - MSE: 4,367,000.01

  - MAE: 1633.25
  - MAPE: 3.29%
  - R^2: 0.9469
  - Explained Variance: 0.9470
  - MGD: 0.0018
  - MPD: 86.36

Observation:

- Similar performance to the base paper, indicating consistent results when using the same feature set.
- Slight improvements in error metrics and R^2 values on test data.

3.4 Our Code with 4 Features (1-Year Training)



```
---------------------------------------------------------------------------------
Our Code ::
4 Feature ( Open,High,Low,Adj-Close,Volume )
1 Year Training Period
Train and Test on Exchanges Data
---------------------------------------------------------------------------------
Train data RMSE:  853.6406543883127
Train data MSE:  728702.3668245067
Train data MAE:  279.3424926753174
Train data MAPE: %  0.6339452825849092
Train data R2:  0.9915857340314704
Train data explained variance regression score: 0.9915857340874039
---------------------------------------------------------------------------------
Test data RMSE:  790.4101107269701
Test data MSE:  624748.1431394211
Test data MAE:  243.52298056370637
Test data MAPE: %  0.48614821613865494
Test data R2:  0.9923970176485116
Test data explained variance regression score: 0.9924489754832209
---------------------------------------------------------------------------------
```

Configuration:

- **Features:** 4 (Open, High, Low, Adj-Close, Volume)
- **Training Period:** 1 Year
- **Dataset:** Exchange Data (Train and Test)

Performance Metrics:

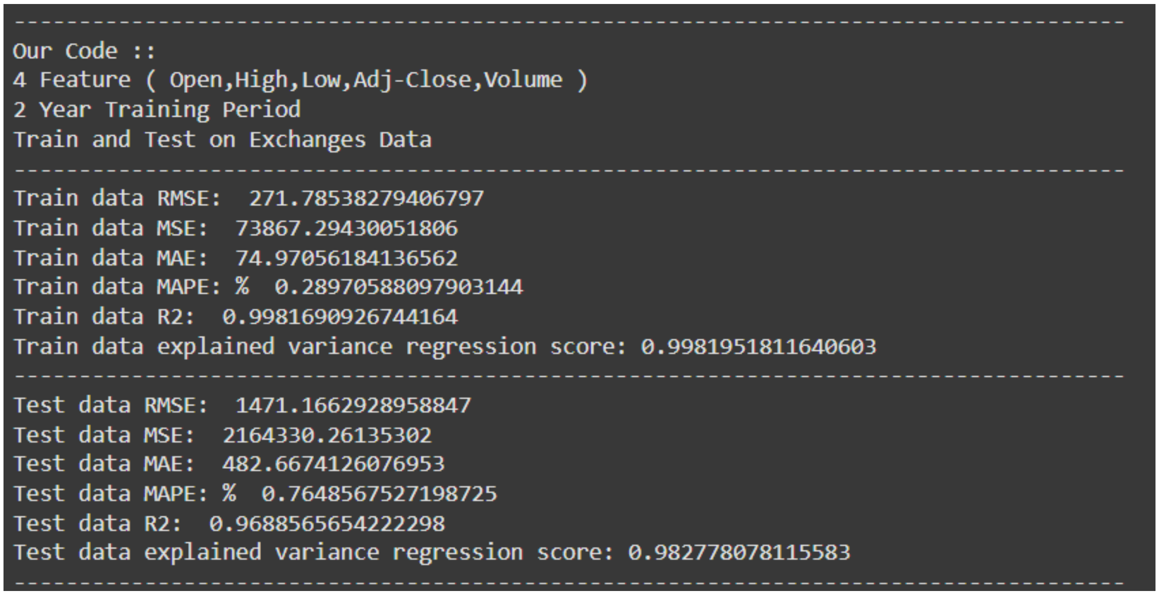
- **Train Data:**
  - RMSE: 853.64
  - MSE: 728,702.37
  - MAE: 279.34
  - MAPE: 0.63%
  - R^2: 0.9916
  - Explained Variance: 0.9916
- **Test Data:**

  - RMSE: 790.41
  - MSE: 624,748.14
  - MAE: 243.52
  - MAPE: 0.49%
  - R^2: 0.9924
  - Explained Variance: 0.9924

Observation:

- Significant reduction in RMSE and other error metrics compared to models using only the 'Close' feature.
- High R^2 values indicate a strong fit.
- Including additional features greatly improves model accuracy.

3.5 Our Code with 4 Features (2-Year Training)



```
------------------------------------------------------------------------------------
Our Code ::
4 Feature ( Open,High,Low,Adj-Close,Volume )
2 Year Training Period
Train and Test on Exchanges Data
------------------------------------------------------------------------------------
Train data RMSE:  271.78538279406797
Train data MSE:  73867.29430051806
Train data MAE:  74.97056184136562
Train data MAPE: %  0.28970588097903144
Train data R2:  0.9981690926744164
Train data explained variance regression score: 0.9981951811640603
------------------------------------------------------------------------------------
Test data RMSE:  1471.1662928958847
Test data MSE:  2164330.26135302
Test data MAE:  482.6674126076953
Test data MAPE: %  0.7648567527198725
Test data R2:  0.9688565654222298
Test data explained variance regression score: 0.982778078115583
------------------------------------------------------------------------------------
```

Configuration:

- **Features:** 4 (Open, High, Low, Adj-Close, Volume)
- **Training Period:** 2 Years
- **Dataset:** Exchange Data (Train and Test)

Performance Metrics:

- **Train Data:**
  - RMSE: 271.79
  - MSE: 73,867.29
  - MAE: 74.97
  - MAPE: 0.29%
  - R^2: 0.9982
  - Explained Variance: 0.9982

- **Test Data:**
  - RMSE: 1471.17
  - MSE: 2,164,330.26
  - MAE: 482.67
  - MAPE: 0.76%
  - R^2: 0.9689
  - Explained Variance: 0.9828

Observation:

- Training error metrics improved further, indicating a better fit on training data.
- However, test error metrics worsened compared to the 1-year training model, suggesting potential overfitting.
- The model performs exceptionally well on training data but not as well on unseen data.

4. Conclusions

4.1 Effect of Training Timeline

- **Observation:** Increasing the training period from 1 year to 2 years reduced training errors but increased test errors.
- **Conclusion:** A longer training period led to overfitting, where the model learns the training data too well, including its noise, and fails to generalize to new data.
- **Implication:** There is a need to balance the amount of training data to avoid overfitting. Techniques like cross-validation, regularization, or pruning the model can help mitigate this issue.

4.2 Effect of Feature Selection

- **Observation:** Including additional features (Open, High, Low, Adj-Close, Volume) significantly improved performance metrics.
- **Conclusion:** More relevant features provide the model with better information, leading to more accurate predictions.
- **Implication:** Feature engineering and selection are critical steps in model development. Careful selection of informative features can enhance model performance.

4.3 Effect of Dataset Choice

- **Observation:** Training on Yahoo data and testing on Exchange data yielded better performance compared to training and testing on the same dataset.
- **Conclusion:** The model's ability to generalize across different datasets indicates robustness. However, variations in data distributions can affect performance.
- **Implication:** It's essential to consider dataset characteristics and ensure that training data is representative of the data the model will encounter in production.

5. Recommendations

- **Optimize Training Period:** Avoid overfitting by selecting an appropriate training timeline. Consider using validation sets to monitor model performance on unseen data.
- **Enhance Feature Set:** Continue exploring additional relevant features and perform feature importance analysis to understand their impact.
- **Dataset Analysis:** Evaluate the consistency between training and testing datasets. If discrepancies exist, data normalization or augmentation techniques might be necessary.
- **Model Regularization:** Implement regularization methods to prevent overfitting when using larger training datasets.

6. Summary

The analysis demonstrates that:

- **Training Timeline:** Longer training periods can lead to overfitting, negatively impacting model generalization.
- **Feature Selection:** Incorporating multiple relevant features significantly improves model accuracy.
- **Dataset Choice:** Models trained on one dataset can perform well on another if the data distributions are similar, but care must be taken to ensure data compatibility.

By carefully considering these factors, we can develop more robust and accurate predictive models.