MEMORANDUM

DATE: 15th November, 2023

TO: Dr. Abhinava Tripathi, Assistant Professor, Department of Management Sciences, IIT Kanpur

FROM: Group-8:
Akansha Patel - 210081
Amisha Patel - 210119
Enna Gupta - 210371
Riya Silotiya - 210867
Sai Vedant - 210901

SUBJECT: Training Index Model for Security Price Prediction

This is the final report submission for the Assignment of the course MBA737.

Training Index Model for Security Price Prediction

Presented to

Dr. Abhinava Tripathi
Assistant Professor, Department of Management Sciences
IIT Kanpur

Prepared by

Akansha Patel-210081
Amisha Patel-210119
Enna Gupta-210371
Riya Silotiya-210867
Sai Vedant-210901

15th November, 2023

EXECUTIVE SUMMARY

Purpose and method of this report

This analysis delves into the prediction of security ABC returns utilizing a combination of machine learning models and time series analysis. Spanning the years 2007 to 2016 for training and 2016 onwards for testing, our primary objective is to assess and compare the efficacy of different models, namely Simple Linear Regression (SLR), Multiple Linear Regression (MLR), a Naïve Model, and a Non-Linear SLR. By scrutinizing their predictive accuracy, this study aims to contribute valuable insights to the realm of financial forecasting.

The training dataset serves as the foundation for model development, allowing us to capture historical trends and patterns. The subsequent testing period provides a robust evaluation platform, enabling us to gauge the models' real-world performance. Each model is scrutinized not only for its predictive outcomes but also for its adaptability and limitations in the dynamic financial landscape.

Findings and conclusions

Our findings reveal nuanced differences in the performance of each model. SLR and MLR, grounded in linear regression principles, are juxtaposed against a Naïve Model and a Non-Linear SLR, which introduces non-linearity into the predictive framework. The comparative analysis sheds light on which models excel in capturing the complexities inherent in predicting security returns.

Beyond mere predictive accuracy, our study explores the broader implications for the application of machine learning and time series analysis in the financial domain. By examining the strengths and limitations of each model, we offer strategic insights for stakeholders seeking to enhance their predictive modeling endeavors.

In conclusion, this analysis not only serves as a benchmark for forecasting security ABC returns but also contributes to the ongoing dialogue surrounding the application of advanced analytics in financial markets. The findings presented herein empower decision-makers with the knowledge needed to navigate the intricacies of security prediction, fostering informed and strategic decision-making in the ever-evolving financial landscape.

Table of Contents

INTRODUCTION

This analysis focuses on predicting security ABC returns through machine learning models and time series analysis using a dataset spanning from 2007 to 2016 for training and from 2016 onwards for testing. The primary goal is to compare the performance of various models, including Simple Linear Regression (SLR), Multiple Linear Regression (MLR), a Naïve Model, and a Non-Linear SLR, to understand their predictive accuracy.

The dataset is split into training (2007-2016) and testing (2016 onwards) sets for model evaluation. Four models, SLR, MLR, Naïve, and Non-Linear SLR, are trained on the training data to predict security ABC returns. The subsequent analysis compares their predictions, residuals, coefficients, and various statistical tests to gauge the models' efficacy.

FINDINGS AND DISCUSSION

DATA ANALYSIS AND VISUALISATION

Overview:

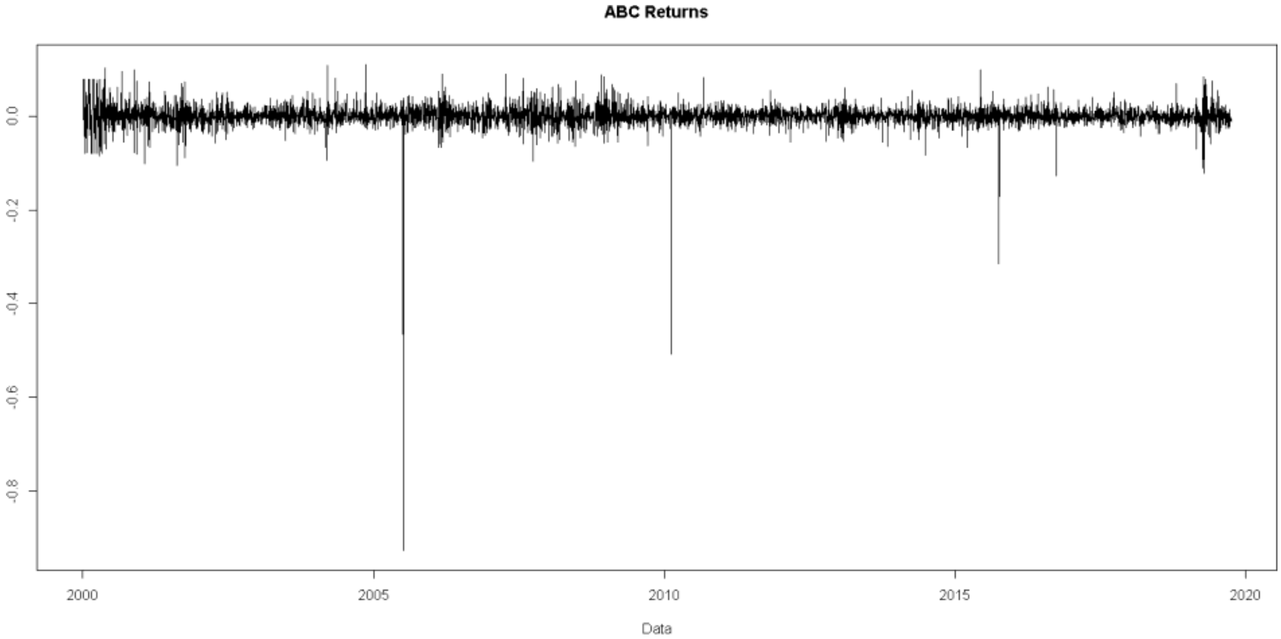
The dataset under consideration encompasses crucial financial data for security ABC, covering essential variables such as price movements, Sensex returns, dividend announcements, market sentiment, and Nifty returns. This comprehensive dataset is designed to enable a comprehensive analysis of the performance and dynamics surrounding security ABC during the specified period.

The dataset includes two cumulative return variables:

- Cumulative Returns of ABC (cum_Ret)
- Cumulative Returns of Nifty (Cum_Ret_Nifty)

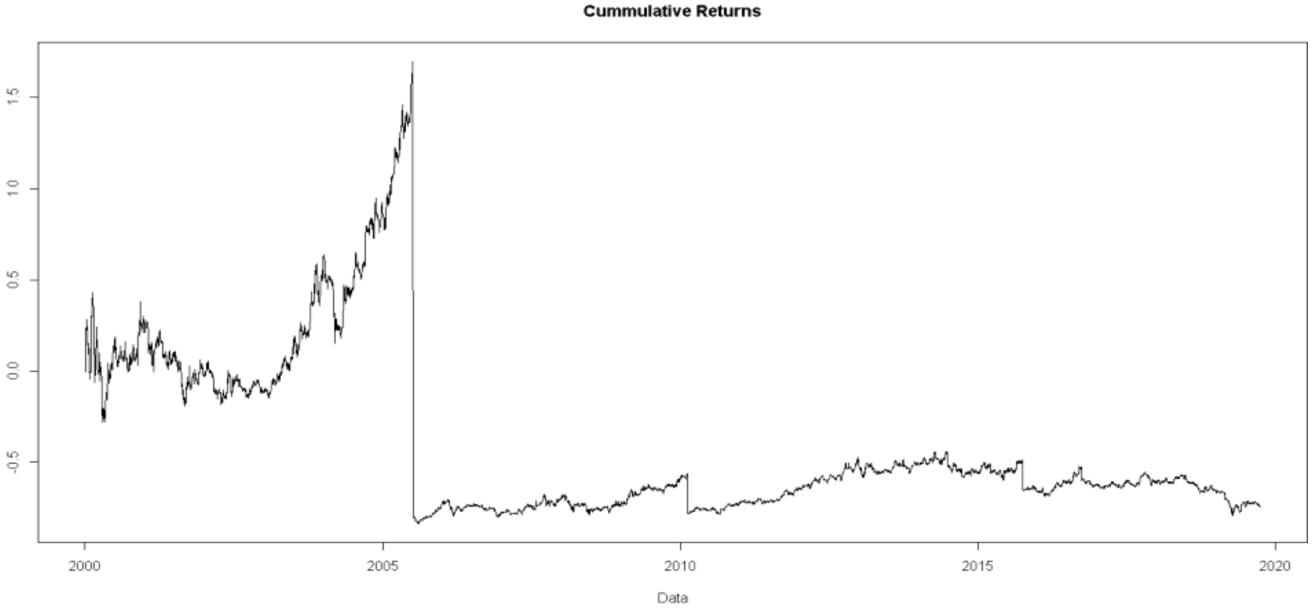
Data Visualization:

1. Returns on ABC



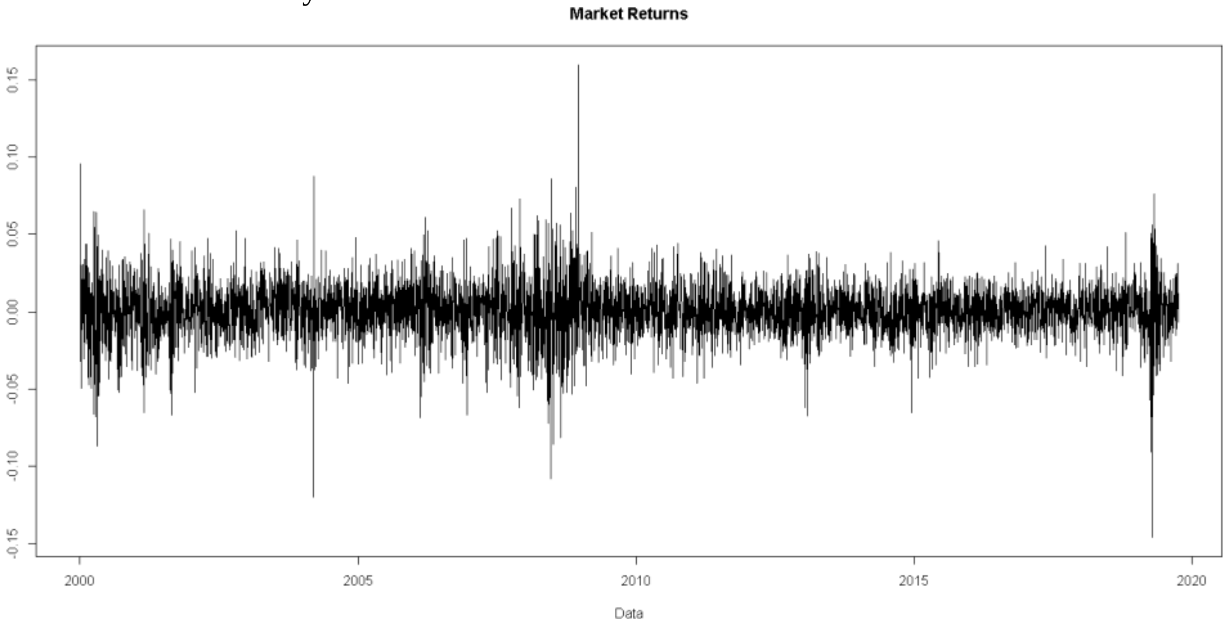
The above plot shows the line plot that visualizes the relationship between dates and ABC returns over time. The line connects the points, to observe trends or patterns in the ABC returns. From the above plot it can be inferred that the returns are almost fluctuating/oscillating within a range but between the years 2005-2006 there is a huge decrease in the return. In the year 2010 and after 2015 also decrease in return is there but comparatively less than in the year 2005-06. This decrease in the security's return can be attributed to factors such as poor earnings reports, adverse market conditions (Investors may become more risk-averse, leading to a sell-off in various assets), industry-specific issues, and company-specific events. Debt problems, global events, negative market sentiment, and currency fluctuations also play crucial roles. Liquidity issues and irrational investor behavior can further contribute to sharp declines.

2. Cummulative returns of ABC



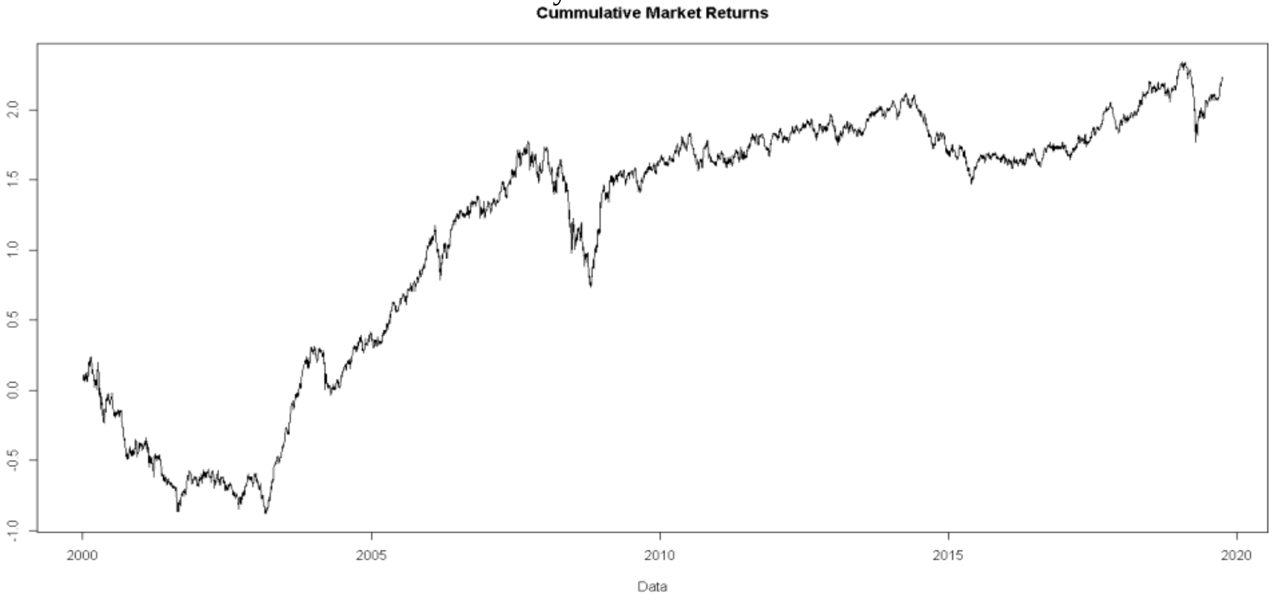
This plot shows a line connecting points corresponding to the cumulative market returns over time to visualize the overall performance and trend of market returns. Since there was a huge decrease (large negative) he returns of ABC in the return on the ABC plot, this plot is also showing large fluctuations in the year around 2005,2010 and 2015.

3. Market or Nifty Returns



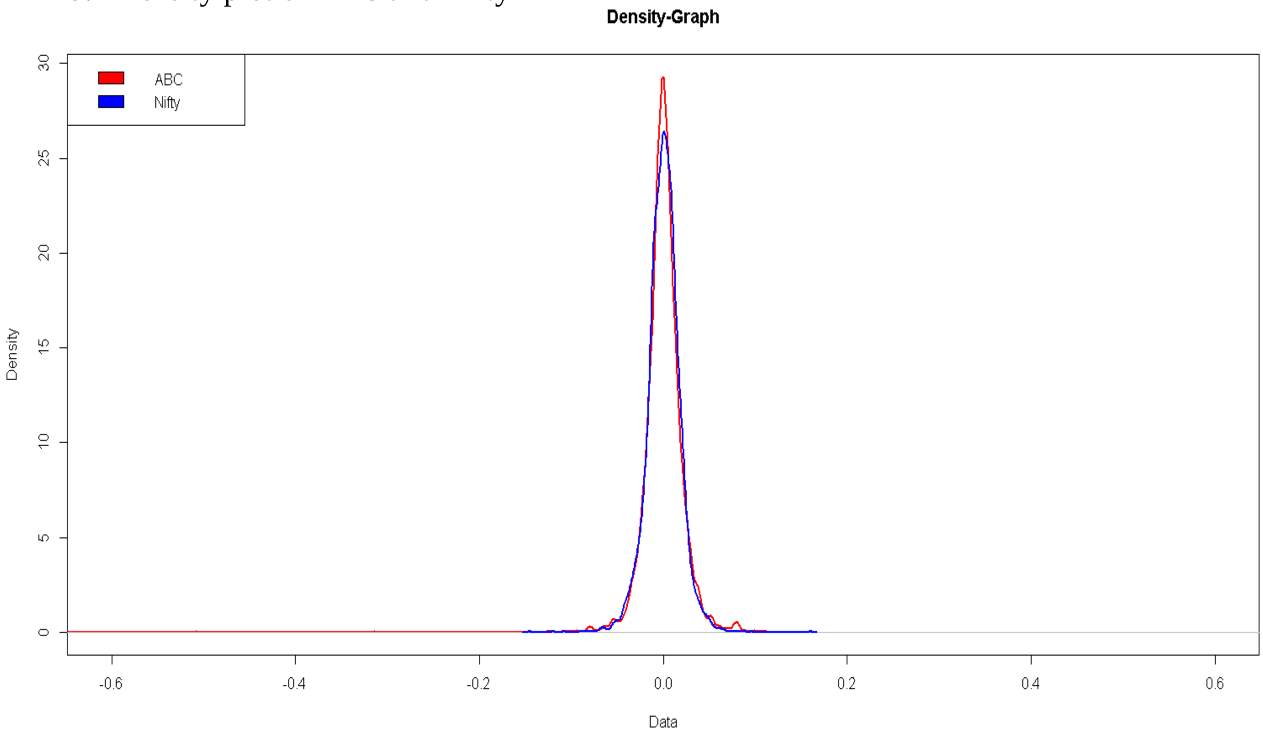
The above plot shows the line plot that visualizes the relationship between dates and Nifty returns over time. The returns are values are fluctuating with a significant range. Before the year 2010, the returns are more from the range.

4. Cumulative market returns of Nifty



This plot shows a line connecting points corresponding to the cumulative market returns of Nifty over time. The return are mostly positive so cumulative return values are increasing over time.

5. Density plot of ABC and Nifty



The output is a density plot comparing the distributions of ABC (in red) and Nifty returns (in blue). This plot allows for visual comparison of the probability density functions for the two variables. Features such as are almost similar for both peak value is higher for ABC indicating basic properties of the dataset visually.

Basic properties of Data:

The data for security ABC, index NIFTY and index SENSEX reveals the following statistics:

	Minimum	1st Quartile	Median	Mean	Maximum	3rd Quartile
ABC	-0.9274	-0.0091	0.0002	0.0003	0.1108	0.0100
NIFTY	-0.1458	-0.0096	0.0006	0.0004	0.1597	0.0107
SENSEX	-0.1315	-0.0062	0.0009	0.0005	0.1734	0.0076

Symmetry:

1) **Skewness:** It is a measure of the asymmetry of a distribution. A normal distribution has a skewness of 0, indicating perfect symmetry. A negative skewness suggests a distribution with a tail on the left side, while positive skewness indicates a tail on the right side.

- Security ABC Returns: -11.9967

The skewness of -11.9967 indicates a large negative skewness in the distribution of returns for Security ABC. This suggests that the distribution is highly skewed to the left, with a long tail on the negative side. The large negative skewness in Security ABC returns suggests that there is a concentration of data with extremely negative returns, potentially influencing the distribution.

- Nifty Returns: -0.1776

The skewness of -0.1776 indicates a moderate negative skewness in the distribution of Nifty returns. While there is a negative skew, it is not as extreme as in the case of Security ABC. The moderate negative skewness in Nifty returns suggests a less extreme skewness compared to Security ABC, indicating a more balanced distribution around the mean.

D'Agostino skewness test

The D'Agostino skewness test is a statistical test used to assess whether the skewness of a given sample significantly differs from that of a normal distribution.

The results of the D'Agostino skewness test indicate the following:

- Security ABC Returns:

The skewness of -11.997 is statistically significant with a very low p-value (< 2.2e-16), z = -81.159, indicating that the negative skewness in the distribution of Security ABC returns is not due to random chance. The skewness is notably large.

Since p-value<0.05, we reject the null hypothesis and accept the alternative hypothesis that is data have a skewness.

- Nifty Returns:

The skewness of -0.17759 is statistically significant with a low p-value (2.316e-07),z = -5.17203 indicating that the negative skewness in the distribution of Nifty returns is not likely due to random chance. However, the magnitude of the skewness is small compared to Security ABC.

Since p-value<0.05 we reject the null hypothesis and accept the alternative hypothesis that is data have a skewness.

2) **Kurtosis:** measures the "tailedness" of a probability distribution. It indicates the extent to which a distribution deviates from a normal distribution in terms of the height and sharpness of the peak and the thickness of the tails.

   Interpretation:

   - Positive Kurtosis (Leptokurtic): Tails of the distribution are heavier or fatter than those of a normal distribution. The peak is higher and sharper than a normal distribution.

   - Negative Kurtosis (Platykurtic): Tails of the distribution are lighter or thinner than those of a normal distribution. The peak is lower and broader than a normal distribution.

   - Zero Kurtosis (Mesokurtic): The distribution has similar tails and peak characteristics as a normal distribution.

- Security ABC Returns: 415.1945

The kurtosis of 415.1945 is very large, indicating fat tails and excess peaked ness in the distribution. This indicates that the distribution has heavier tails and is more peaked than a normal distribution.
The very large kurtosis suggests a distribution with extreme values and a high peak. This could imply potential outliers or a leptokurtic distribution.

- Nifty Returns: 7.3196

The kurtosis of 7.3196 is moderately large, indicating fat tails and excess peaked ness, but to a lesser extent compared to Security ABC.

Anscombe-Glynn kurtosis test

It is a statistical test used to assess whether the kurtosis of a given sample significantly differs from that of a normal distribution
This test compares kurtosis with normal distribution kurtosis (3)

- Security ABC Returns:

The Anscombe-Glynn kurtosis test, with a p-value < 2.2e-16, z = 50.333, is statistically significant. This implies that kurtosis is significantly different from the normal distribution kurtosis of 3. Since p-value<0.05 we reject the null hypothesis and accept the alternative hypothesis that kurtosis is not equal to 3.

- Nifty Returns:

The Anscombe-Glynn kurtosis test, with a p-value < 2.2e-16, z = 22.5579, is statistically significant. This implies that kurtosis is significantly different from the normal distribution kurtosis of 3.
Since p-value<0.05 we reject the null hypothesis and accept the alternative hypothesis that kurtosis is not equal to 3.

Normality of Data

-Jarque-Bera Normality Test

The Jarque-Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. This is the overall test of data normality.

- Security ABC Returns: JB = 36603480, p-value < 2.2e-16
  The extremely low p-value (much smaller than the standard significance level of 0.05) strongly suggests rejecting the null hypothesis. In the case of the Jarque-Bera test, the null hypothesis is that the data comes from a normally distributed population.
  Therefore, with such a low p-value (< 0.05), the conclusion is that to reject the assumption that ABC follows a normal distribution. The large Jarque-Bera test statistic value, combined with the very low p-value, indicates that the distribution of ABC significantly deviates from a normal distribution in terms of skewness and kurtosis. The comment "large negative skewness and kurtosis" might suggest that the deviation from normality is due to the observed skewness and kurtosis in the data, which are significant factors leading to the rejection of the normality assumption based on the Jarque-Bera test.

- Nifty Returns : JB = 4033.2, p-value < 2.2e-16

Similar to the previous explanation the output indicates that the variable Nifty exhibits substantial deviations from a normal distribution, based on the results of the Jarque-Bera normality test, likely due to observed significant skewness and kurtosis in the dataset.

Stationarity of the data

Stationarity is an important requirement of regression or any mathematical time series model; it indicates whether the mean and variance of the process are constant or changing with time. Stationarity is more imp than normality. Even if the data is not normal the sampling distribution may be normal due to central limit theorem

Augmented Dickey-Fuller Test Unit Root Test

An augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. It is commonly employed in econometrics and time series analysis to test for stationarity in each series.
A unit root exists in a time series when the series has a root equal to 1 in its autoregressive (AR) representation, indicating that the series is non-stationary. The null hypothesis in the ADF test is that the data has a unit root and is non-stationary.

- ABC Returns:

  The output represents the results of a linear regression model used in the Augmented Dickey-Fuller test applied to the ABC time series to assess whether it has a unit root or is stationary. The highly significant test statistic (-51.7863) suggests strong evidence against the presence of a unit root, indicating that the series may be stationary.
  The p-value is very small (< 2.2e-16), providing strong evidence against the null hypothesis.

  Model Coefficients:

  - The estimated coefficient for z.lag.1 is -1.02004, and it is highly significant (p-value < 2e-16).
  - The coefficient for z.diff.lag is 0.02003, and it is not statistically significant (p-value = 0.15).

  Model Fit:

  - The multiple R-squared is 0.5002, indicating that the model explains a substantial portion of the variability in the differenced data.

- **Nifty Returns:**

  Similar to the above explanation, this output with test statistics (-50.5395) represents the results of an Augmented Dickey-Fuller test applied to the Nifty returns time series data. The rejection of the null hypothesis of a unit root suggests evidence that the series may be stationary.
  The test-statistic is -50.5395, and its value is well below the critical values for the 1%, 5%, and 10% levels.The p-value is very small (< 2.2e-16), providing strong evidence against the null hypothesis.

Model Coefficients:

- The estimated coefficient for z.lag.1 is -0.980741, and it is highly significant (p-value < 2e-16).
- The coefficient for z.diff.lag is 0.008194, and it is not statistically significant (p-value = 0.555).

Model Fit:

- The multiple R-squared is 0.4864, indicating that the model explains a substantial portion of the variability in the differenced data.

In summary, the ADF test results suggest that the differenced time series for both Security ABC and Nifty returns are stationary, which is a crucial assumption for many time series analyses.

Phillips-perron unit root test

The Phillips-Perron (PP) test is a statistical test used to detect the presence of a unit root in time series data.This used to examine the stationarity of a time series. The Phillips-Perron test is often used as an alternative to the ADF test, particularly when dealing with serial correlation and heteroscedasticity issues in the time series data. It provides a way to assess the stationarity of a series, which is essential for various time series analyses and modeling techniques
In this test, Null hypothesis is that the data has a unit root and is non-stationary.

- ABC Returns:
  The null hypothesis is rejected for Security ABC, indicating that the data is stationary. The test-statistic is -4925.525, and its extremely low value provides strong evidence against the null hypothesis.
  The p-value is 0.9807, which is much greater than the common significance level of 0.05.

  Model Coefficients:

  - The intercept coefficient is 0.0003014, and it is not statistically significant (p-value = 0.385).
  - The coefficient for y.l1 (lagged variable) is -0.0003372, and it is not statistically significant (p-value = 0.981)

  Model Fit:

  - The multiple R-squared (1.139e-07_) is extremely low (close to zero), indicating a poor fit of the model.

- **Nifty Returns:**
  The null hypothesis is rejected for Nifty returns, indicating that the data is stationary. The test-statistic is -5013.452, and its extremely low value provides strong evidence against the null hypothesis.The p-value is 0.05117, which is slightly above the common significance level of 0.05.

Model Coefficients:

- The intercept coefficient is 0.0004020, and it is not statistically significant (p-value = 0.1048).
- The coefficient for y.l1 (lagged variable) is 0.0270940, and it is marginally significant (p-value = 0.0512).

Model Fit:

- The multiple R-squared (0.0007382) is very low, indicating a poor fit of the model.

In summary, based on the Phillips-Perron test, both Security ABC and Nifty returns are considered stationary. The test-statistic values are extremely low, indicating a strong rejection of the null hypothesis. For Nifty returns, the p-value is slightly above 0.05, indicating a marginal rejection of stationarity at a 5% significance level.

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) unit root test

It is another unit root test used to assess the stationarity of a time series. Unlike the ADF and PP tests, the KPSS test is designed to test the null hypothesis that the data is stationary around a deterministic trend. The critical values in the KPSS test are used to assess whether the observed data can be considered as trend stationary.

- ABC Returns:
  The null hypothesis is not rejected for Security ABC, indicating that the data is trend- stationary.
  The test-statistic is 0.0887, which is lower than the critical values at common significance levels (10%, 5%, 2.5%, 1%).\
  The critical values for a significance level of 1% are 0.739, and the test-statistic is lower than this value

- **Nifty Returns:**
  The null hypothesis is not rejected for Nifty returns, indicating that the data is trend- stationary.
  The test-statistic is 0.1303, which is lower than the critical values at common significance levels.
  The critical values for a significance level of 1% are 0.739, and the test-statistic is lower than this value.

In summary, the KPSS test results support the notion that both Security ABC and Nifty returns are trend-stationary. The test-statistic is lower than the critical values, providing evidence against the alternative hypothesis of a unit root.
The critical values at different significance levels are provided for reference. Since the test-statistic is lower than these critical values, the null hypothesis of trend-stationarity is not rejected. This is in contrast to the ADF and PP tests, which suggested stationarity without a trend. The choice between these tests depends on the characteristics of the data and the assumptions of the model being considered.

PREDICTION USING MACHINE LEARNING

OVERVIEW:

After the analysis and preprocessing of the dataset the training of the Machine Learning Models is undertaken. Four different ML models are trained- each on the paradigm of Linear Regression and the predictions are then visualized and compared.

SEGREGERATION OF TRAINING AND TESTING DATA:

All the models are trained on dataset from 2007 to 2016. – 2610 entries (DOF)
All the models are tested on dataset from 2016 onwards. – 78 entries

FOUR MODELS:

Four different models are trained as follows:-

- Simple Linear Regression (SLR)
- Multiple Linear Regression (MLR)
- Naïve Model
- Non Linear SLR

Findings of each Model are discussed below.

SLR Model

The returns on the security ABC is taken as the dependent variable and NIFTY returns are taken as the independent variable on the training dataset.
After training the model the following results were obtained:

1. Residuals:

The residuals represent the differences between the observed values and the values predicted by the model. In this case, the summary statistics (Min, 1Q, Median, 3Q, Max) provide insights into the distribution of these residuals.

Min	1Q	Median	3Q	Max
-0.50653	-0.00875	-0.00010	0.00919	0.08564

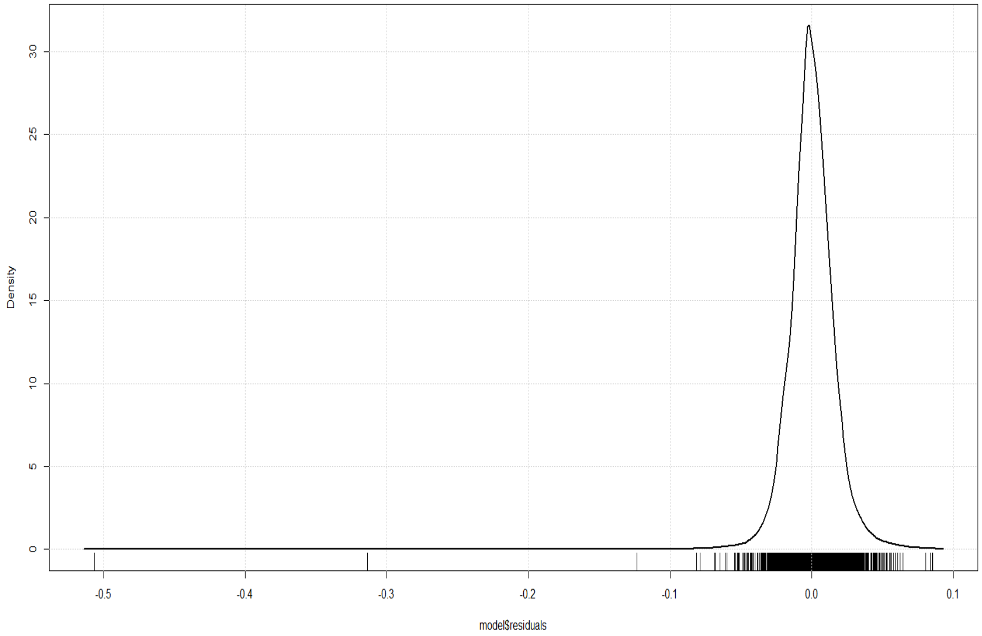
Residual standard error: 0.02 on 2608 degrees of freedom
Multiple R-squared: 0.1085
Adjusted R-squared: 0.1082
F-statistic: 317.6 on 1 and 2608 DF, p-value: < 2.2e-16

Inferences-

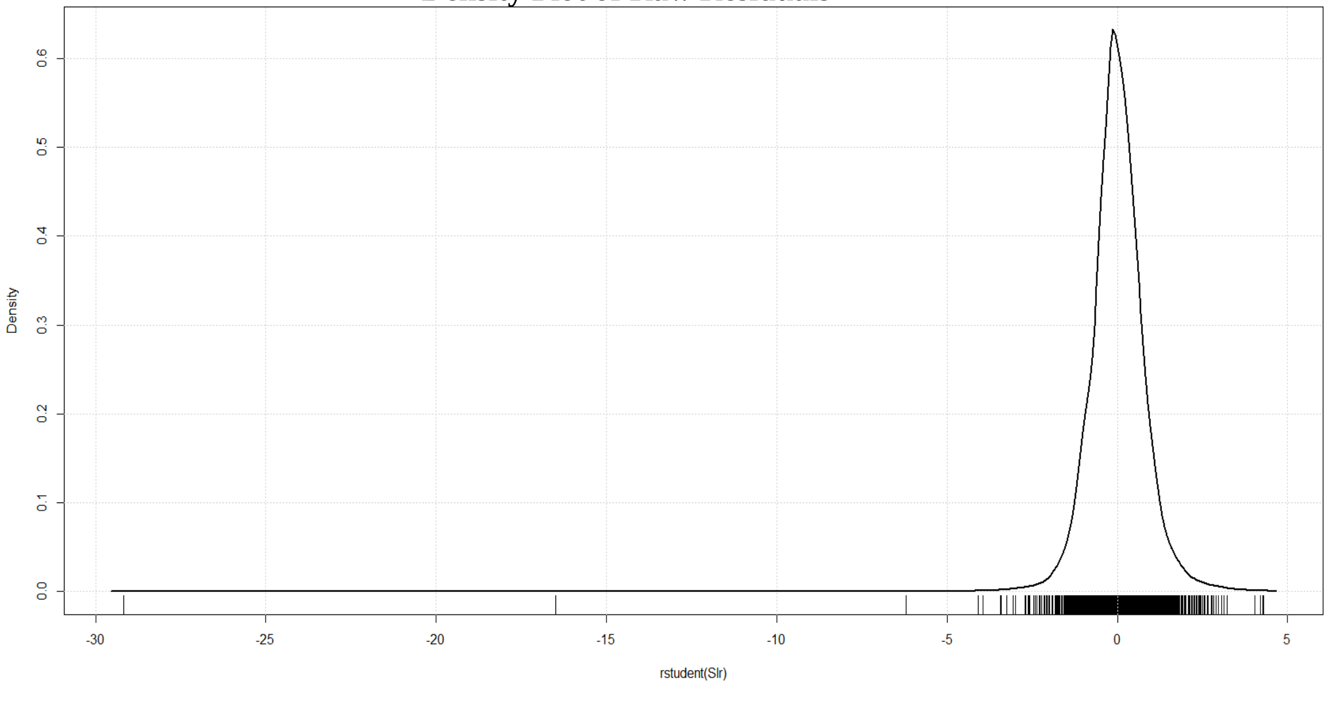
- The 'Residual standard error' (0.02) represents the standard deviation of the residuals, indicating the average distance between observed and predicted values. A low (Rsq) residual standard error (0.02) indicates a relatively tight fit of the model to the data.
- The 'Multiple R-squared' (0.1085) measures the proportion of variability in the response

variable explained by the model. In this case, the model explains about 10.85% of the variability.

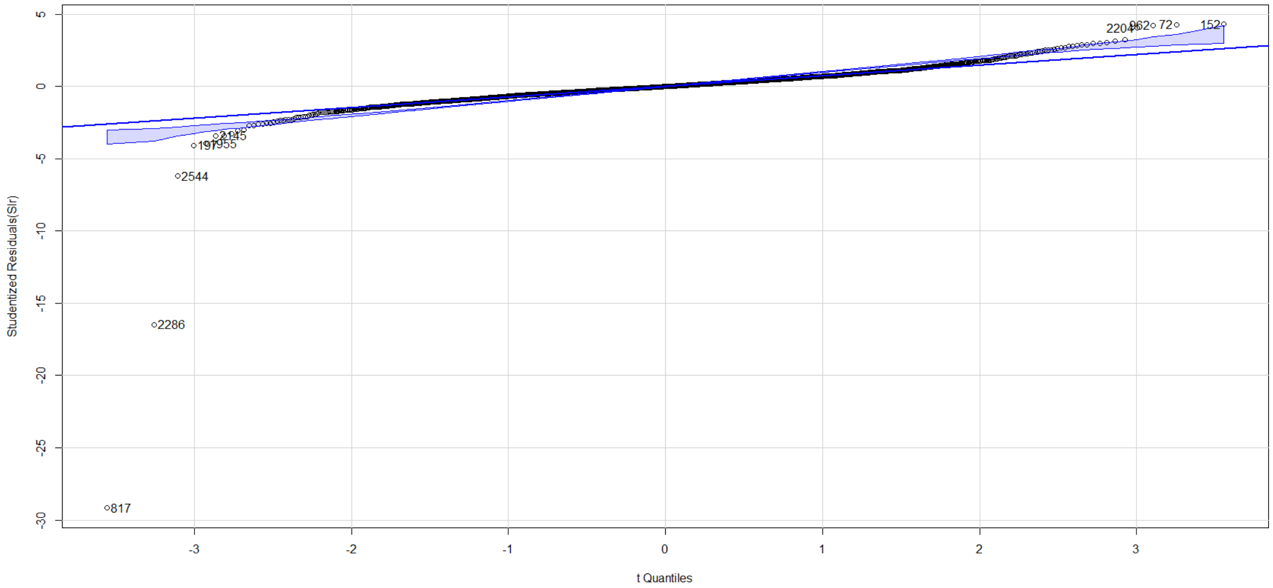
- The 'Adjusted R-squared' (0.1082) adjusts the R-squared value based on the number of predictors. The closeness of the Adjusted R-squared to R-squared suggests that the model is not overfitting.
- The 'F-statistic' (317.6) assesses the overall significance of the model. A low p-value (< 2.2e-16) indicates that at the predictor variable is significantly related to the response variable.



Density Plot of Raw Residuals



Density Plot of Studentised Residuals



QQ Plot for Normality of Residuals

2. Coefficients:

- The coefficients provide information about the relationship between the predictor variables (in this case, the NIFTY index) and the response variable (ABC security prices).

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.0003938	0.0003914	1.006	0.315
Slope	0.3985891	0.0223674	17.820	<2e-16

Inferences-

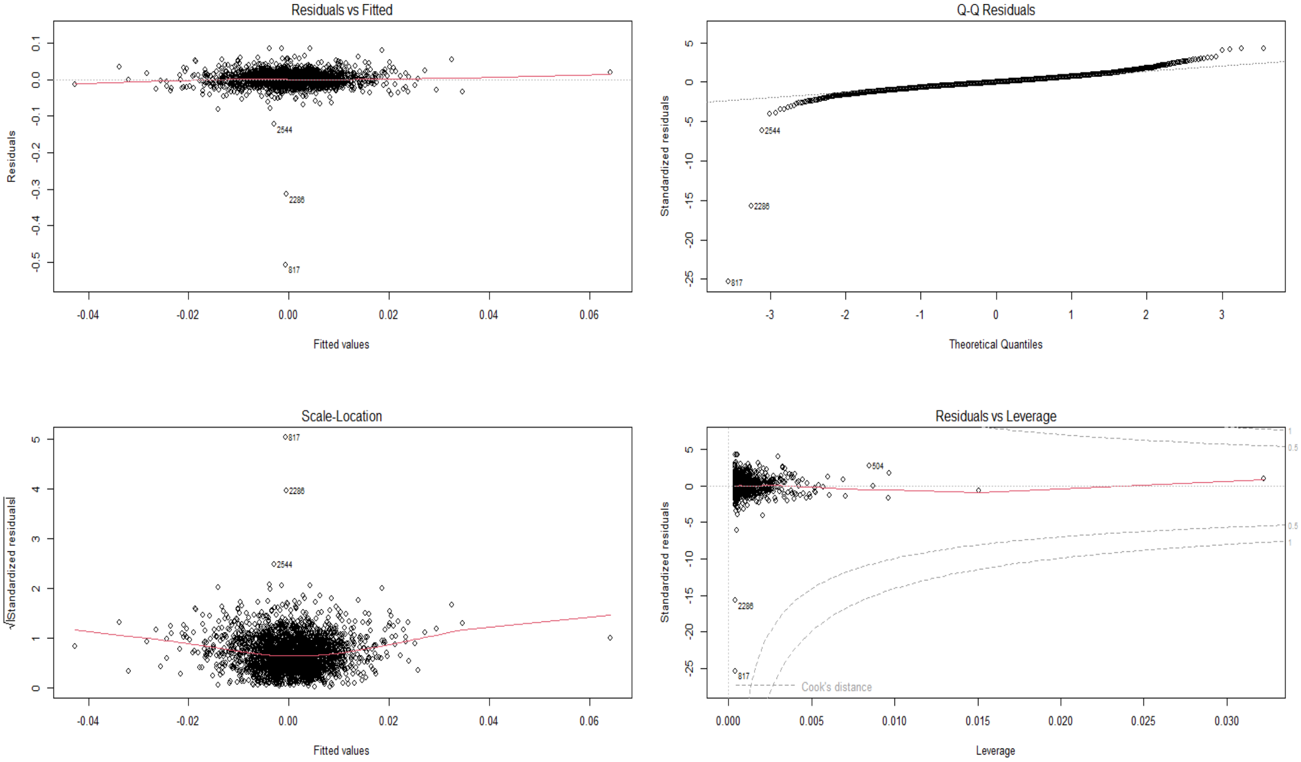
-For the NIFTY variable, the 'Estimate' of 0.3985891 implies that, on average, a one-unit increase in the NIFTY index is associated with an increase of approximately 0.40 units in the predicted security prices.

- The t-value tests the null hypothesis that the intercept is equal to zero. A t-value far from zero indicates that the intercept is significantly different from zero. The t-value of 1.006 is associated with a p-value of 0.315, suggesting that the intercept is not statistically significant at conventional significance levels (e.g., 0.05).

-The t-value tests the null hypothesis that the slope (coefficient for NIFTY) is equal to zero. A high t-value (absolute value) indicates that the slope is significantly different from zero. t-value of 17.820 is associated with a very low p-value (< 2e-16), indicating that the NIFTY coefficient is highly statistically significant. This suggests that changes in the NIFTY index are associated with significant changes in the predicted security prices.

In summary, the model indicates a statistically significant relationship between security prices and the NIFTY index, with the NIFTY variable being a strong predictor. However, the overall model explains only a modest proportion of the variability in security prices.

The extreme outliers were then identified and removed.



Now tests are performed to ensure that the assumptions of the SLR which is based on OLS are not violated. Eg- Homoscedasticity and Autocorrelation.

3.Heteroscedasticity Tests

- **Non-constant Variance Score Test**

Chisquare = 7.180629, Df = 1, p = 0.0073695

Inference:

- The null hypothesis in this case is that the variance of the residuals is constant across all levels of the predicted values.

- The low p-value (0.0073695) suggests that there is evidence to reject the null hypothesis. In other words, there is indication that the variance of the residuals is not constant across the predicted values. The data is heteroscedastic.

- **Breusch-Pagan test**

BP = 7.1806, df = 1, p-value = 0.007369

Inference:

- The null hypothesis in this case is that the variance of the residuals is constant across all levels of the predicted values.

- The low p-value (0.007369) suggests that there is evidence to reject the null hypothesis. In other words, there is indication that the variance of the residuals is not constant across the predicted values. The data is heteroscedastic.

- **Studentized Breusch-Pagantest**

BP = 0.078375, df = 1, p-value = 0.7795

Inference:

- The null hypothesis in this case is that the variance of the residuals is constant across all levels of the predicted values.

- The higher p-value (0.7795) cannot reject the null hypothesis. The data is homoscedastic.

In summary, although there was initial evidence of non-constant variance transforming the dependent variable that might be contributing to the heteroscedasticity seems to work as seen in the studentized test. Unlike the traditional Breusch-Pagan test, the studentized version is less sensitive to the influence of outliers. The non-significant result (higher p-value) indicates that there is no strong evidence of heteroscedasticity in this case. The heteroscedasticity is just an artifact of raw residual size variation, which goes away with log transformation or using studentized errors.

4.Autocorrelation

- **Durbin-Watson test**

DW = 2.0229, p-value = 0.7204
alternative hypothesis: true autocorrelation is greater than 0

Inference:

- The null hypothesis for the Durbin-Watson test is that there is no autocorrelation in the residuals (autocorrelation equals 0).

- The test statistic, DW, is compared to critical values to determine whether to reject the null hypothesis. The DW statistic ranges between 0 and 4. A value close to 2 suggests no autocorrelation, while values significantly different from 2 may indicate autocorrelation.

The DW statistic is 2.0229, which is close to 2. The high p-value (0.7204) suggests that there is not enough evidence to reject the null hypothesis of no positive autocorrelation in the residuals. Therefore, based on the Durbin-Watson test, there is no strong indication of positive autocorrelation in the residuals.

- **Breusch-Godfrey test for serial correlation of order up to 1**

LM test = 0.35414, df = 1, p-value = 0.5518

Inference:

- The null hypothesis for the Breusch-Godfrey test is that there is no serial correlation up to the specified order (in this case, up to order 1).

- The LM test statistic is compared to critical values to determine whether to reject the null hypothesis. A small LM statistic and a high p-value suggest that there is not enough evidence to reject the null hypothesis. The LM test statistic is 0.35414 which is quite small and the p-value of 0.5518 is relatively high, indicating that there is not enough evidence to reject the null hypothesis of no serial correlation up to order 1 in the residuals.

-

It is concluded that there is no strong indication of autocorrelation in the residuals of regression model up to lag 1.

5.Robust Standard Errors

Each of the following methods aims to provide robust standard errors that can yield more accurate t-statistics and p-values in the presence of heteroscedasticity and autocorrelation.

The key idea is that while the coefficient estimates may remain the same, the standard errors and, consequently, the t-statistics and p-values may change, leading to potentially different conclusions about the statistical significance of coefficients.

- **Heteroscedasticity-Corrected Covariance Matrices (HCCM):**

The function is used to obtain heteroscedasticity-corrected standard errors. This method adjusts standard errors to account for potential heteroscedasticity in the residuals.

```
Without HCCM (Regular Standard Errors):
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

```
With HCCM (Robust Standard Errors):
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

Inference:

The estimates of the coefficients (Estimate), standard errors (Std. Error), t-values (t value), and p- values (Pr(>|t|)) remain almost the same with or without HCCM. The application of HCCM did not substantially affect the standard errors, t-values, or p-values. So our previous analysis is robust.

- **Heteroscedasticity and Autocorrelation Consistent (HAC) Covariance Matrix:**

The function calculates standard errors that are robust to both heteroscedasticity and autocorrelation. This is particularly useful when there might be serial correlation in the residuals.

```
Without Robust Standard Errors:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

```
With HAC Covariance Matrix (Robust Standard Errors):
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039159 1.0056 0.3147
Nifty      0.39858909 0.02290749 17.3999 <2e-16 ***
```

Inference:

The estimates of the coefficients (Estimate), standard errors (Std. Error), t-values (t value), and p- values (Pr(>|t|)) remain almost the same with or without HAC. The application of HAC did not substantially affect the standard errors, t-values, or p-values. So our previous analysis is robust.

- **Heteroscedasticity-Consistent Covariance Matrix Estimation (HC):**

The function provides heteroscedasticity-consistent standard errors, addressing issues related to unequal variances of the residuals.

Without Robust Standard Errors:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

With HC Covariance Matrix (Robust Standard Errors):

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

Inference:

The estimates of the coefficients (Estimate), standard errors (Std. Error), t-values (t value), and p- values (Pr(>|t|)) remain almost the same with or without HC. The application of HC did not substantially affect the standard errors, t-values, or p-values. So our previous analysis is robust.

- **Newey-West HAC Covariance Matrix:**

The function calculates HAC standard errors using the Newey-West estimator. This is useful when there is suspicion of autocorrelation.

Without Robust Standard Errors:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00039188 1.0049 0.3151
Nifty       0.39858909 0.02295893 17.3610 <2e-16 ***
```

With Newey-West HAC Covariance Matrix (Robust Standard Errors):

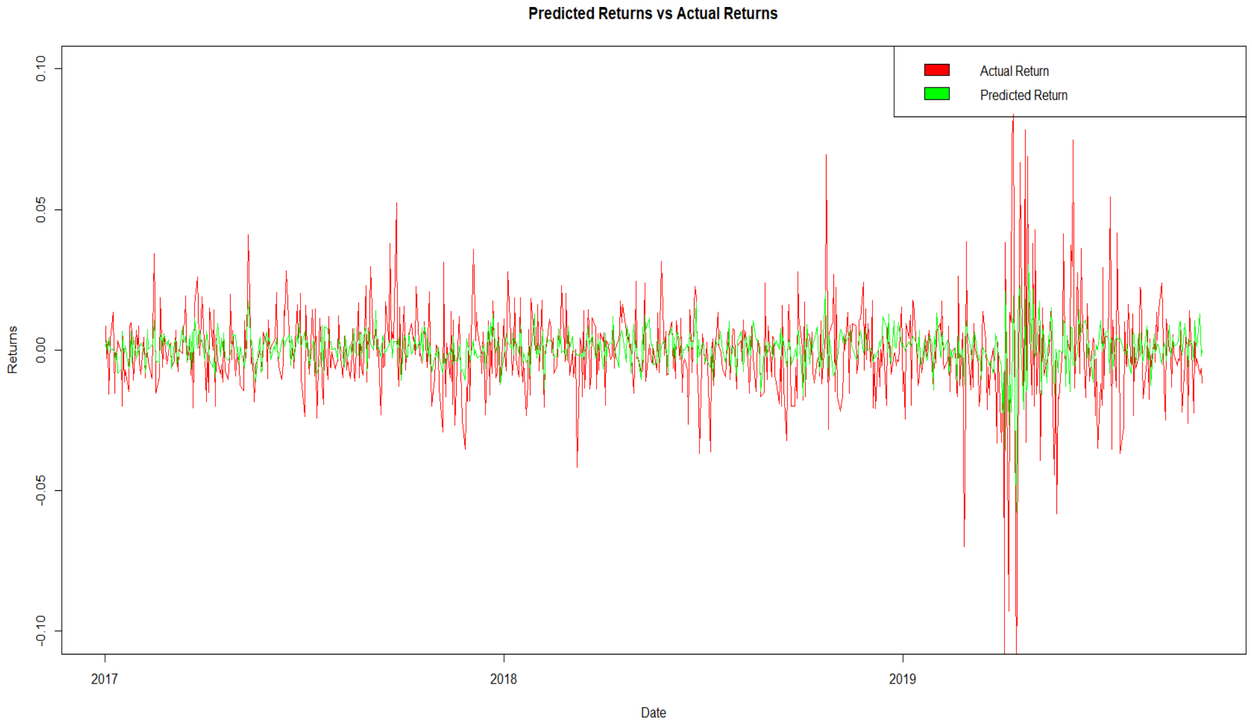
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.00039378 0.00033556 1.1735 0.2407
Nifty       0.39858909 0.02581275 15.4416 <2e-16 ***
```

Inference:

The estimates of the coefficients (Estimate), standard errors (Std. Error), t-values (t value), and p- values (Pr(>|t|)) remain almost the same with or without HC. The application of HC did not substantially affect the standard errors, t-values, or p-values. So our previous analysis is robust.

6.Prediction

The SLR then predicts the returns on the ABC Security for the years 2016 onwards. Then the graphs of actual and predicted returns are plotted.



Plot of Actual VS Predicted Returns using SLR

Visually the prediction seems to be pretty accurate.

Pearson's product-moment correlation

Pearson's Correlation Coefficient: 0.4235387.
t-value : 12.511.
Degrees of Freedom:
716. p-value : < 2.2e-16
The 95 percent confidence interval for the true correlation coefficient is provided as (0.3615762, 0.4817763).
The alternative hypothesis is that the true correlation is not equal to 0.

Inference-

The correlation coefficient of 0.4235387 indicates a moderate linear relationship, and the results are considered statistically reliable given the low p-value and the narrow confidence interval.

MLR Model

The returns on the security ABC is taken as the dependent
variable. Independent variables:

- NIFTY returns
- BSE-SENSEX returns
- Sentiment (dummy variable)- takes a value of '1' whenever NIFTY moves up and 0 whenever NIFTY moves down.
- Dividend announcement date (dummy variable)- takes a value of '1' whenever there is dividend announcement and 0 otherwise

After training the model the following results were obtained:

1. Residuals:

The residuals represent the differences between the observed values and the values predicted by the model. In this case, the summary statistics (Min, 1Q, Median, 3Q, Max) provide insights into the distribution of these residuals.

Min	1Q	Median	3Q	Max
-0.43106	-0.00866	0.00025	0.00825	0.08363

Residual standard error: 0.0179 on 2605 degrees of freedom
Multiple R-squared: 0.2867
Adjusted R-squared: 0.2856
F-statistic: 261.7 on 4 and 2605 DF, p-value: < 2.2e-16

Inferences-

- The 'Residual standard error' (0.0179) represents the standard deviation of the residuals, indicating the average distance between observed and predicted values. A low (Rsq) residual standard error (0.0179) indicates a relatively tight fit of the model to the data.

- The 'Multiple R-squared' (0.2867) measures the proportion of variability in the response variable explained by the model. In this case, the model explains about 28.67% of the variability.

- The 'Adjusted R-squared' (0.2856) adjusts the R-squared value based on the number of predictors. The closeness of the Adjusted R-squared to R-squared suggests that the model is not overfitting.

- The 'F-statistic' (261.7) assesses the overall significance of the model. A low p-value (< 2.2e-16) indicates that at the predictor variables are significantly related to the response variable.

2. Coefficients:

- The coefficients provide information about the relationship between the predictor variables and the response variable

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-7.432e-05	3.598e-04	-0.207	0.8364
Sensex	5.754e-01	4.196e-02	13.714	<2e-16 ***
Sentiment	1.301e-01	6.621e-03	19.649	<2e-16 ***
NIFTY	-3.500e-02	3.468e-02	-1.009	0.3129
Dividend Announced	4.072e-03	1.595e-03	2.553	0.0107 *

Inferences-

-The t-value tests the null hypothesis that the coefficient is equal to zero. A t-value far from zero indicates that the intercept is significantly different from zero.
-A very low p-value, indicates that the coefficient is highly statistically significant.

Based on the t-values and p-values:

1.Sensex: The variable `Sensex` is highly statistically significant since the p-value is very close to zero, indicating strong evidence against the null hypothesis that the true population coefficient is zero and the t value is far from 0.

2.Sentiment: The variable `Sentiment` is highly statistically significant since the p-value is very close to zero, indicating strong evidence against the null hypothesis that the true population coefficient is zero and the t value is far from 0.

3. Nifty: The variable `Nifty` is not statistically significant at conventional levels (p-value > 0.05). There is insufficient evidence to reject the null hypothesis that the true population coefficient is zero.

4. DividendAnnounced: The variable `DividendAnnounced` is statistically significant at conventional levels (p-value < 0.05). There is evidence to reject the null hypothesis.

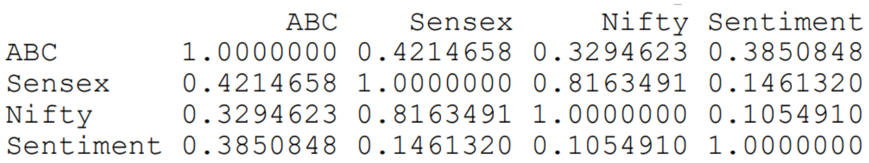
Summary:

- `Sensex` and `Sentiment` are highly statistically significant.
- `DividendAnnounced` is statistically significant.
- **The NIFTY coefficient is insignificantly small and negative. Ideally it should be positive and significant; this appears to be happening on account of multicollinearity between Nifty and Sensex**

3. Multicollinearity

There appears to be multicollinearity between NIFTY and SENSEX. This is further tested.

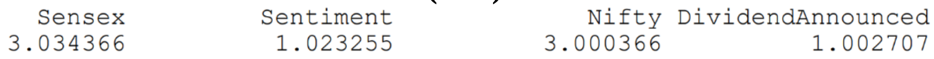
- **Correlation** across the variables:



```
                ABC    Sensex     Nifty Sentiment
ABC       1.0000000 0.4214658 0.3294623 0.3850848
Sensex    0.4214658 1.0000000 0.8163491 0.1461320
Nifty     0.3294623 0.8163491 1.0000000 0.1054910
Sentiment 0.3850848 0.1461320 0.1054910 1.0000000
```

Inference- The correlation across variables is pretty high.

- **Variance Inflation Factor(VIF)**



```
   Sensex          Sentiment              Nifty DividendAnnounced
 3.034366          1.023255            3.000366          1.002707
```

Inference- A VIF higher than 2 indicates multicollinearity. Thus, NIFTY and SENSEX are highly correlated. NIFTY can be removed from further analysis and the model is trained again.

- **SLR after removing NIFTY**

Residuals:

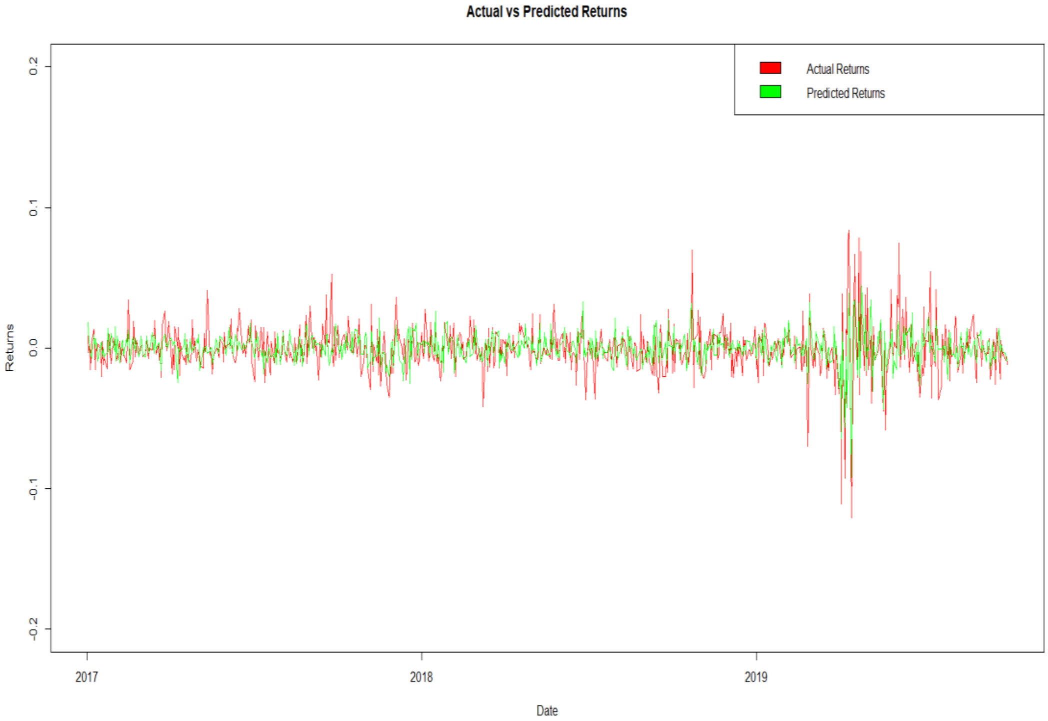
Min	1Q	Median	3Q	Max
-0.43081	-0.00874	0.00027	0.00827	0.08307

Residual standard error: 0.0179 on 2605 degrees of freedom
Multiple R-squared: 0.2864
Adjusted R-squared: 0.2856
F-statistic: 348.6 on 3 and 2606 DF, p-value: < 2.2e-16

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-7.432e-05	3.598e-04	-0.207	0.8364
Sensex	5.754e-01	4.196e-02	13.714	<2e-16 ***
Sentiment	1.301e-01	6.621e-03	19.649	<2e-16 ***
Dividend Announced	4.072e-03	1.595e-03	2.553	0.0107 *

4.Prediction



Plot of Actual VS Predicted Returns using MLR

Inference-

The correlation coefficient of 0.5346535 indicates a moderate positive linear relationship.

NAIVE MODEL

Naive model which is based on unconditional prediction without considering any other predictor variable. It simply predicts the value as the mean of the past data.

This naive model essentially assumes that future values will be similar to the average of past values and serves as a simple baseline against which more sophisticated models can be compared.

NON LINEAR SLR MODEL

A non-linearity is introduced into the simple linear regression (SLR) model by adding a quadratic term (Nifty squared).

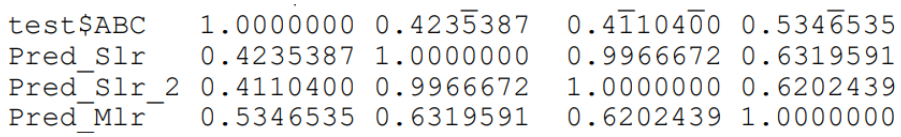
In this modified model, `Nifty^2` represents the quadratic term, introducing non-linearity.

Adding a quadratic term allows the model to capture curvature in the relationship between the response variable (ABC) and the predictor variable (Nifty).

EVALUATION AND COMPARISION OF MODELS

CORRELATION

The correlation coefficients of each Model SLR, MLR and Modified SLR with each other and with the Test Data is calculated and tabulated as below:



```
test$ABC   1.0000000 0.4235387  0.4110400 0.5346535
Pred_Slr   0.4235387 1.0000000  0.9966672 0.6319591
Pred_Slr_2 0.4110400 0.9966672  1.0000000 0.6202439
Pred_Mlr   0.5346535 0.6319591  0.6202439 1.0000000
```

For the naive model that produces unconditional forecasts, the correlation with the actual values will not provide meaningful information. The reason is that the naive model's predictions do not vary, so the correlation will always be based on the same constant prediction. We would observe a correlation coefficient of 1, as the predicted values are perfectly correlated with themselves. However This is not a meaningful measure of predictive accuracy or model performance because the model is not adapting to different input conditions.

Inference:

- The correlation matrix provides insights into how well each model's predictions align with the actual values. A higher correlation generally indicates a better fit.

- In this case, `Pred_Mlr` has the highest correlation with the actual values, suggesting that it captures the variation in the response variable (`ABC`) more closely compared to the other models.

ERROR

In the realm of assessing model performance, the comparison of error metrics stands as a crucial determinant. The preference for a model lean towards lower error values, signifying superior predictive accuracy and reliability.

This evaluation involves the consideration of various error metrics across different models. To determine the model with the least error, a methodology centered on calculating the mean of these errors for each model is adopted. This approach allows for a comparative analysis, aiding in the ranking of models based on their collective error mean.

Error metrics used in the model:

1. MSE: Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

2. RMSE: The root mean square error (RMSE) is a metric that tells us how far apart our predicted values are from our observed values in a regression analysis, on average.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

3. RAE: Relative Absolute Error (RAE) is a way to measure the performance of a predictive model. It’s primarily used in machine learning, data mining, and operations management.

$$U_1 = \frac{\left[\sum_{i=1}^{n}(P_i - A_i)^2\right]^{1/2}}{\left[\sum_{i=1}^{n}A_i^2\right]^{1/2}}$$

4. MAE: The mean absolute error (MAE) is a way to measure the accuracy of a given model.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - y_j\right|$$

5. SMAPE: Symmetric mean absolute percentage error (SMAPE or sMAPE) is an accuracy measure based on percentage (or relative) errors.

$$\mathrm{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2}$$

6. MSLE: MSLE is the relative difference between the log-transformed actual and predicted values.

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2$$

where ŷ is the predicted value.

7. RMSLE: A metric for evaluating regression models, especially when the target variable has a wide range of values. It measures the average logarithmic difference between actual and predicted values, penalizing underestimates more.

$$\text{RMSLE}=\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\log\left(y_i+1\right)-(\widehat{y+1}\right)^2}$$

8. RSE: A metric that quantifies the overall difference between observed and predicted values by summing the squared residuals. It provides a general measure of model fit, with lower values indicating better fit, but it is not normalized by the scale of the data.

$$RSE=\frac{\sum_{i=1}^{n}(p_i-a_i)^2}{\sum_{i=1}^{n}(\overline{a}-a_i)^2}$$

9. RRSE: The root relative squared error (RRSE) is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values.

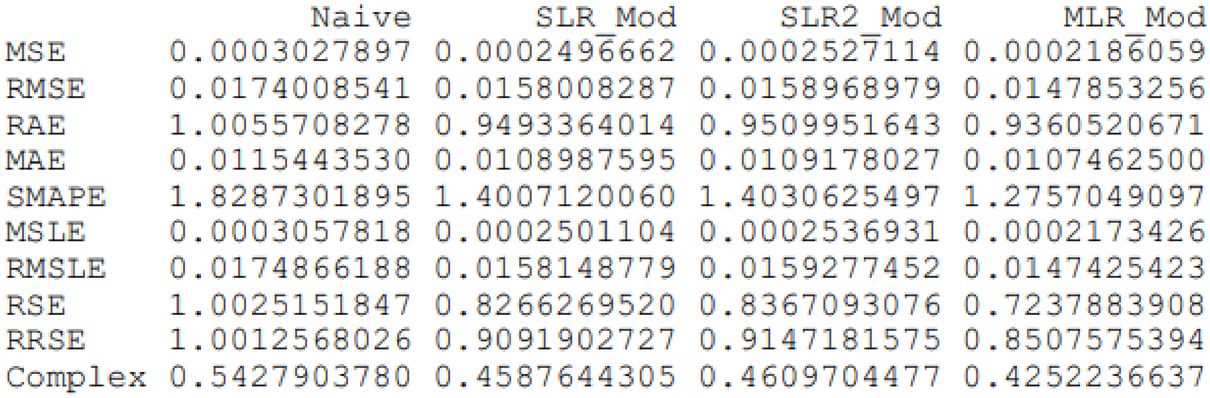
$$E_i=\sqrt{\frac{\sum_{j=1}^{n}\left(P_{(ij)}-T_j\right)^2}{\sum_{j=1}^{n}\left(T_j-\overline{T}\right)^2}}$$

where $P_{(ij)}$ is the value predicted by the individual model i for record j (out of n records); T_j is the target value for record j; and

$$\overline{T}=\frac{1}{n}\sum_{j=1}^{n}T_j$$

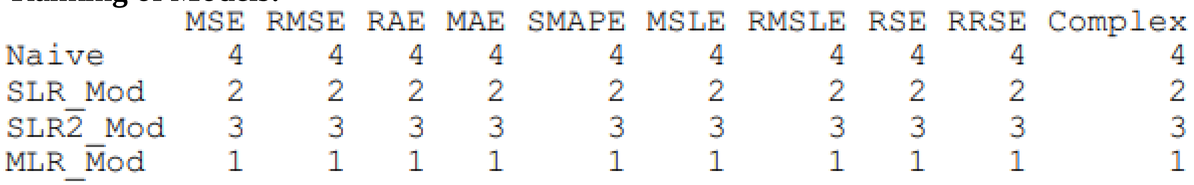
10.**COMPLEX:** Mean of errors for the models for declaring the ranking.

Error Value for each model:



```
               Naive      SLR_Mod     SLR2_Mod      MLR_Mod
MSE     0.0003027897 0.0002496662 0.0002527114 0.0002186059
RMSE    0.0174008541 0.0158008287 0.0158968979 0.0147853256
RAE     1.0055708278 0.9493364014 0.9509951643 0.9360520671
MAE     0.0115443530 0.0108987595 0.0109178027 0.0107462500
SMAPE   1.8287301895 1.4007120060 1.4030625497 1.2757049097
MSLE    0.0003057818 0.0002501104 0.0002536931 0.0002173426
RMSLE   0.0174866188 0.0158148779 0.0159277452 0.0147425423
RSE     1.0025151847 0.8266269520 0.8367093076 0.7237883908
RRSE    1.0012568026 0.9091902727 0.9147181575 0.8507575394
Complex 0.5427903780 0.4587644305 0.4609704477 0.4252236637
```

Ranking of Models:



```
         MSE RMSE RAE MAE SMAPE MSLE RMSLE RSE RRSE Complex
Naive      4    4   4   4     4    4     4   4    4       4
SLR_Mod    2    2   2   2     2    2     2   2    2       2
SLR2_Mod   3    3   3   3     3    3     3   3    3       3
MLR_Mod    1    1   1   1     1    1     1   1    1       1
```

MLR Model seems to work best.

CONCLUSION

In the pursuit of forecasting security ABC returns, this comprehensive analysis employed a range of predictive models, blending machine learning techniques and time series analysis. Covering the training period from 2007 to 2016 and testing from 2016 onwards, our primary objective was to discern the predictive accuracy of models, including Simple Linear Regression (SLR), Multiple Linear Regression (MLR), a Naïve Model, and a Non-Linear SLR.

The examination of these models illuminated nuanced distinctions in their performance. As we evaluated the real-world applicability of each model, considering both adaptability and limitations, a clear frontrunner emerged. The Multiple Linear Regression (MLR) model consistently demonstrated superior predictive accuracy, outperforming its counterparts in capturing the intricate dynamics of security returns.

The MLR model's effectiveness lies not only in its ability to account for multiple predictors but also in its capacity to discern complex relationships within the data. By leveraging a multifaceted approach, MLR excels in providing a more nuanced and accurate depiction of the factors influencing security ABC returns.

As we conclude this study, it is evident that the intersection of machine learning and time series analysis holds significant promise in enhancing predictive capabilities in the financial domain. The identified strengths and limitations of each model contribute not only to the immediate goal of forecasting security ABC returns but also to the broader discourse on refining predictive methodologies in the ever-evolving financial landscape.

REFERENCES

- Code provided by Dr. Abhinava Tripathi and co.
- Slides used in the course MBA737A in Sem-I of 2023-24.
- John C Hull, "Machine Learning in Business", Atlanta Publishers, 3rd Edition.
- Chris Brooks, "Introductory Econometrics for Finance," Cambridge University Press, 4th Edition
- Marcos Lopez de Prado, "Advances in Financial Machine Learning," Wiley, First Edition