MBA737A

TRAINING INDEX MODEL FOR SECURITY PRICE PREDICTION

PRESENTED TO

DR. ABHINAVA TRIPATHI
ASSISTANT PROFESSOR,
DEPARTMENT OF MANAGEMENT SCIENCES
IIT KANPUR

PRESENTED BY
GROUP-8
AKANSHA PATEL-210081
AMISHA PATEL-210119
ENNA GUPTA-210371
RIYA SILOTIYA-210867
SAI VEDANT-210901

INTRODUCTION

This analysis focuses on predicting security 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 subsequent analysis compares their predictions, residuals, coefficients, and various statistical tests to gauge the models' efficacy.

DATA ANALYSIS

- **Skewness:** measures the asymmetry of a distribution.
    - Security: -11.9967 (Highly Negative) indicating the distribution is highly skewed to the left, with a long tail on the negative side
    - Nifty: -0.1776 (Moderately Negative) indicating a less extreme skewness compared to Security and a more balanced distribution around the mean.
- **Kurtosis:** measures tail heaviness.
    - Security: 415.1945 (Very High) indicating heavier tails and excess peaked ness in the distribution.
    - Nifty: 7.3196 (Moderate) same but to a lesser extent compared to Security.

Statistical Testing

- **D'Agostino Skewness Test:**
    - Highly significant skewness in both security and Nifty.
    - Rejecting null hypotheses, indicating non-normal skewness.
- **Anscombe-Glynn Kurtosis Test:**
    - Both Security and Nifty kurtosis significantly differ from normal distribution kurtosis (3).
    - Rejecting null hypotheses, since p-value<0.05

DATA ANALYSIS

Normality Testing

- **Jarque-Bera Normality Test**: check whether sample data have the skewness and kurtosis matching a normal distribution
    - Significantly low p-values (< 2.2e-16) reject normal distribution assumption for security and Nifty.
- Interpretation:
    - Large negative skewness and kurtosis in Security and Nifty lead to rejection of normality.

Stationarity Testing

It indicates whether the mean and variance of the process are constant or changing with time.

- **Augmented Dickey-Fuller Test:**
    - Both Security and Nifty returns display stationary characteristics.
    - Highly significant statistics suggest rejection of the presence of a unit root.
- **Phillips-Perron Test:**
    - Further confirms the stationarity of Secuirty and Nifty returns.
- **KPSS Test:**
    - Accepts the null hypothesis, indicating stationary characteristics for Security and Nifty returns.
    - Test-statistics below critical values support the stationary nature of the data.

MACHINE LEARNING

- **Preparation:** The dataset is split into training (2007-2016) and testing (2016 onwards) sets
- **Four different models are trained as follows:-**

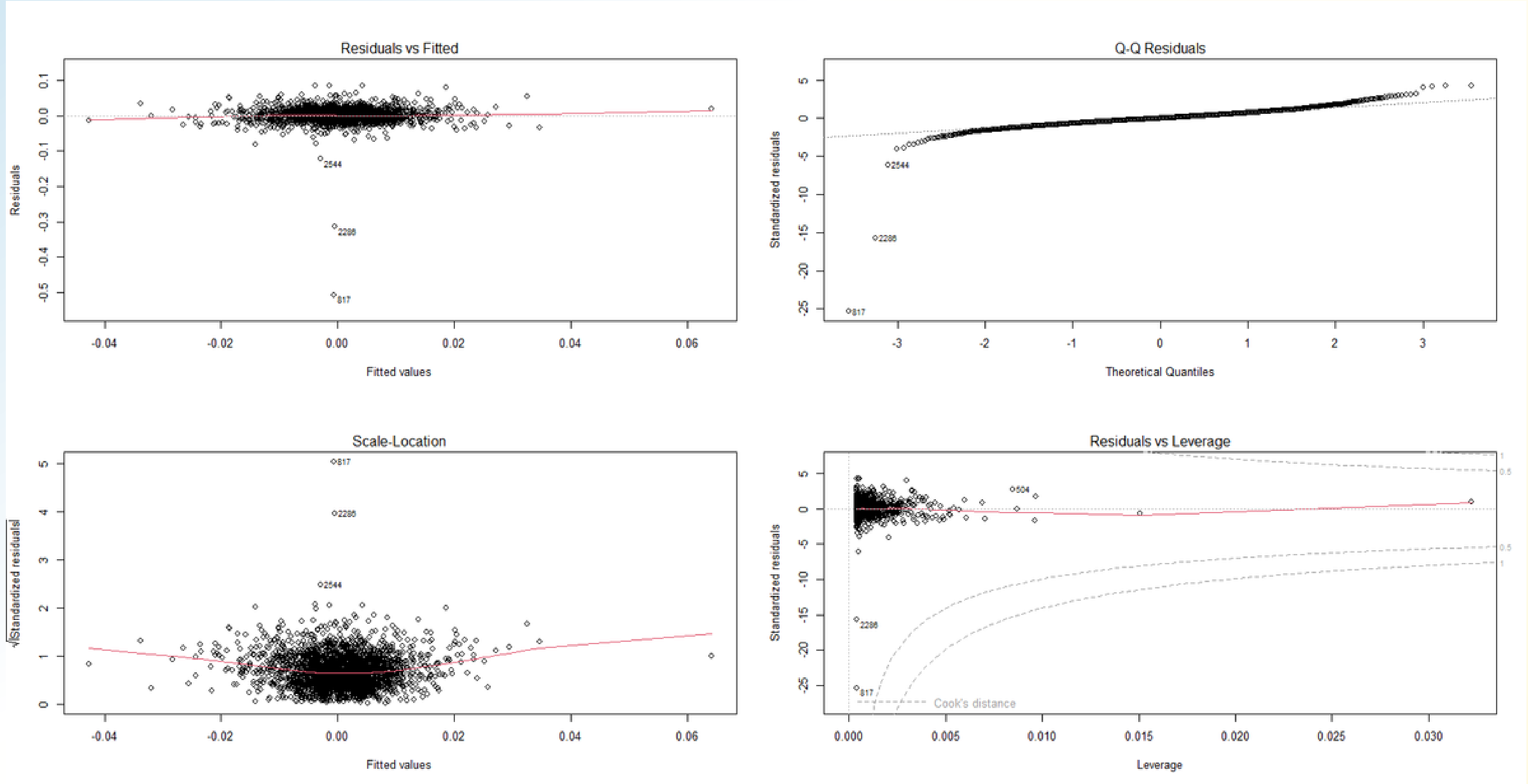
1. Simple Linear Regression (SLR)- returns on the security is taken as the dependent variable and NIFTY returns are taken as the independent variable.
2. Multiple Linear Regression (MLR)- The returns on the security is taken as the dependent variable. Independent variables: NIFTY returns ,BSE-SENSEX 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
3. Naïve 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.
4. Non Linear SLR-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 and the predictor variable (Nifty).

MACHINE LEARNING

Statistical Significance of SLR Model Results

- **Residuals:**

1. Low (Rsq) residual standard error (0.02) indicates a relatively tight fit of the model to the data.
2. The 'Multiple R-squared' (0.1085) the model explains about 10.85% of the variability.
3. 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.
4. The 'F-statistic' (317.6) 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.



Residuals vs Fitted

Q-Q Residuals

Scale-Location

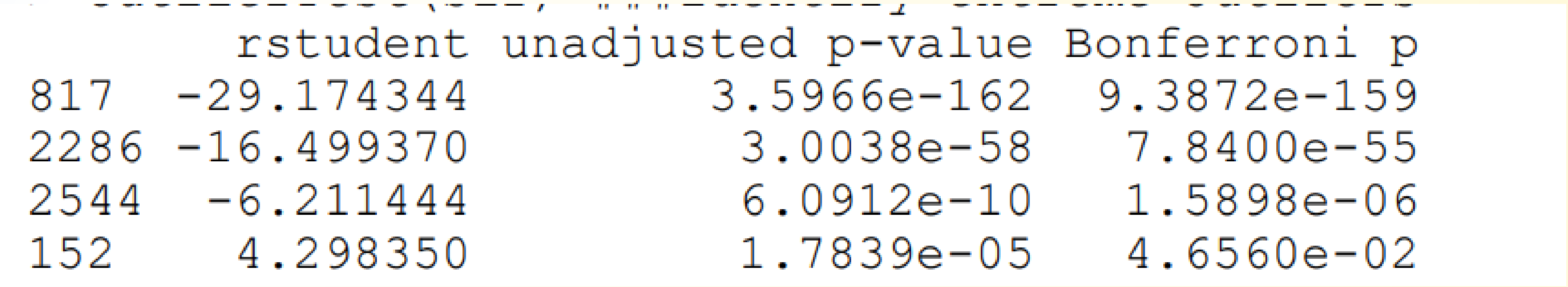
Residuals vs Leverage

MACHINE LEARNING

- **t and p values of Coefficients:**

1. For the NIFTY variable, the 'Estimate' of 0.3985891- Slope of Line
2. A t-value far from zero (1.006) indicates that the intercept is significantly different from zero. p-value of 0.315, suggesting that the intercept is not statistically significant at conventional significance levels (e.g., 0.05).
3. t-value of 17.820 is associated with a very low p-value (< 2e-16), indicating that the NIFTY coefficient is highly statistically significant.

- Outliers are identified and removed:



```
       rstudent unadjusted p-value Bonferroni p
817  -29.174344        3.5966e-162  9.3872e-159
2286 -16.499370         3.0038e-58   7.8400e-55
2544  -6.211444         6.0912e-10   1.5898e-06
152    4.298350         1.7839e-05   4.6560e-02
```

- **Heteroscedasticity Tests**

1. Non-constant Variance Score Test: Low p-value (0.0073695) rejects the null hypothesis. The data is heteroscedastic.
2. Breusch-Pagan test: Low p-value (0.007369) rejects the null hypothesis. The data is heteroscedastic.
3. Studentized Breusch-Pagantest: The higher p-value (0.7795) cannot reject the null hypothesis. The data is homoscedastic.

Heteroscedasticity goes away with using studentized errors.

MACHINE LEARNING

- **Autocorrelation Tests**

1. Durbin-Watson test : The DW statistic is 2.0229, which is close to 2 suggests no autocorrelation. The high p-value (0.7204) cannot reject the null hypothesis. No positive autocorrelation in the residuals.
2. Breusch-Godfrey test for serial correlation of order up to 1 : LM test statistic 0.35414 is small and the p-value of 0.5518 is relatively high cannot reject the null hypothesis.

No positive autocorrelation in the residuals.

- **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 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 all of the following error methods.

1. Heteroscedasticity-Corrected Covariance Matrices (HCCM): Robust to Heteroscedasticity.
2. Heteroscedasticity and Autocorrelation Consistent (HAC) Covariance Matrix: Robust to both heteroscedasticity and autocorrelation.
3. Heteroscedasticity-Consistent Covariance Matrix Estimation (HC): Robust to Heteroscedasticity.
4. Newey-West HAC Covariance Matrix: Robust to autocorrelation.

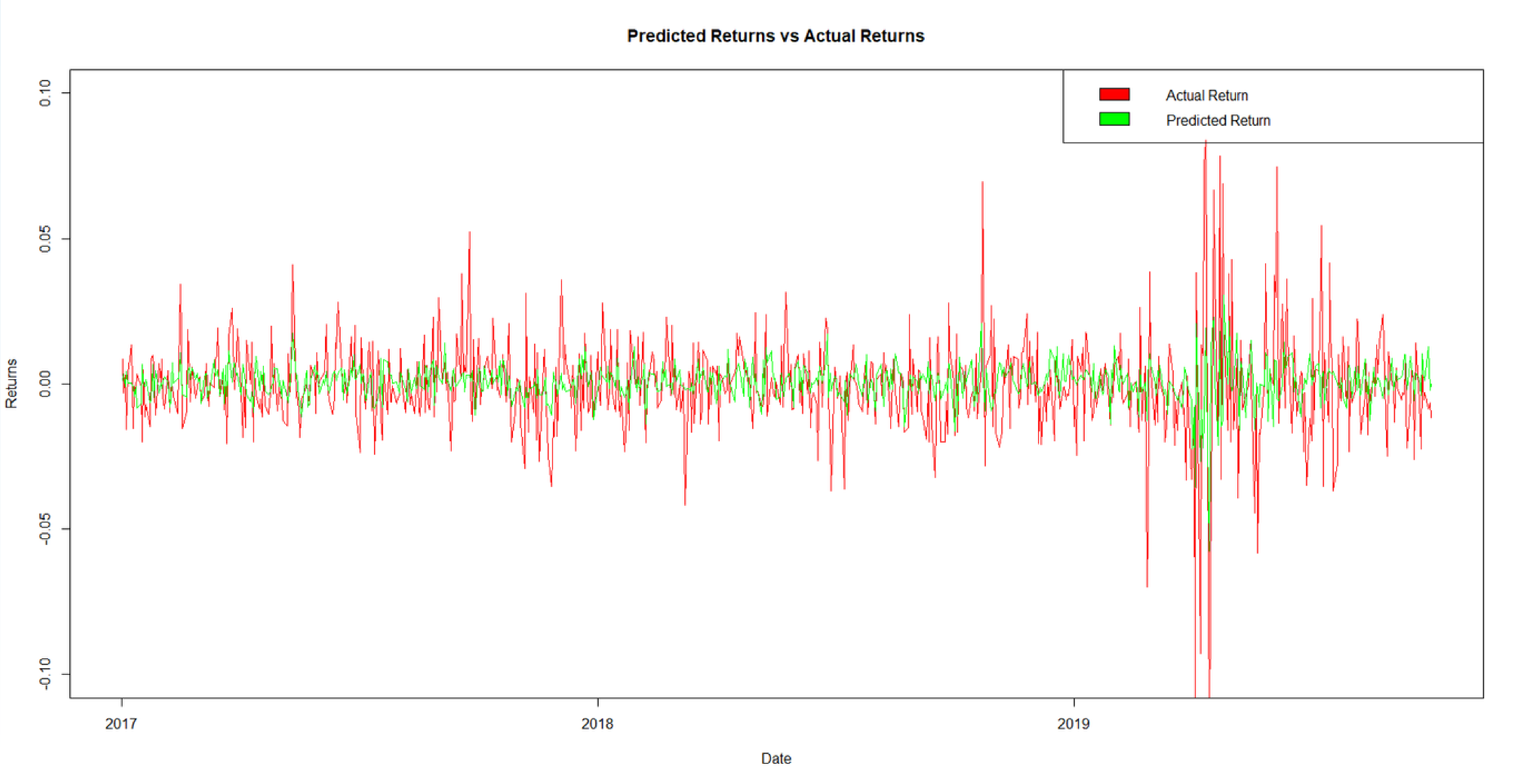
MACHINE LEARNING

- **Prediction:**

Pearson's product-moment correlation Pearson's Correlation Coefficient: 0.4235387 indicates a moderate linear relationship.

t-value : 12.511. High. Degrees of Freedom: 716. p-value : < 2.2e-16 Low.

The 95 percent confidence interval for the true correlation coefficient is provided as (0.3615762, 0.4817763). Narrow. The results are considered statistically reliable.



MACHINE LEARNING

Statistical Significance of MLR Model Results

- **Residuals:**

1. 'Residual standard error' (0.0179) low (Rsq) residual standard error (0.0179) indicates a relatively tight fit of the model to the data.
2. The 'Multiple R-squared' (0.2867) the model explains about 28.67% of the variability.
3. 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.
4. 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.

- **t and p values of Coefficients:**

1. SENSEX- t-value far from 0 (13.714) and p-value very low (<2e-16). Highly statistically significant.
2. Sentiment- t-value far from 0 (19.649) and p-value very low (<2e-16). Highly statistically significant.

3. Dividend Announced- t-value sufficiently far from 0 (2.553) and p-value is less than 0.05 (0.0107). Statistically significant.

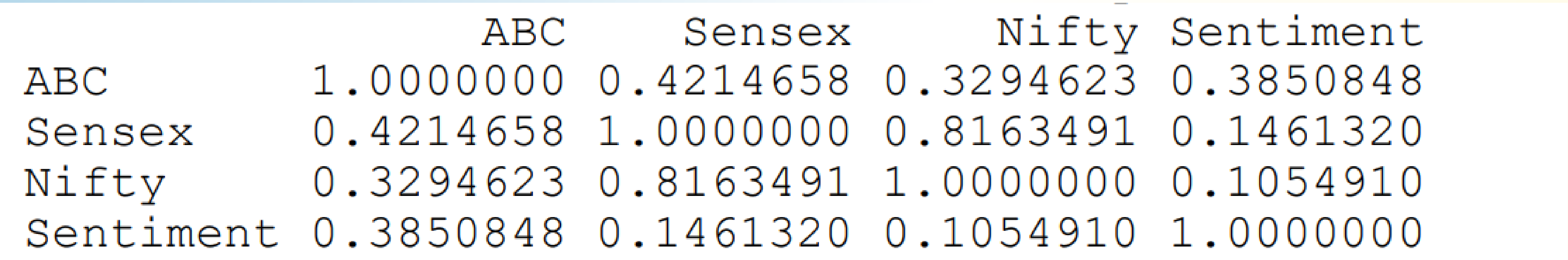
4.NIFTY- The NIFTY coefficient is insignificantly small and negative. Ideally it should be positive and significant; this appears to be happening on account of multicollinearity between Nifty and Sensex

MACHINE LEARNING

- **Multicollinearity:**

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

1. 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
```

The correlation across variables is pretty high. Further Strengthening our hypothesis.

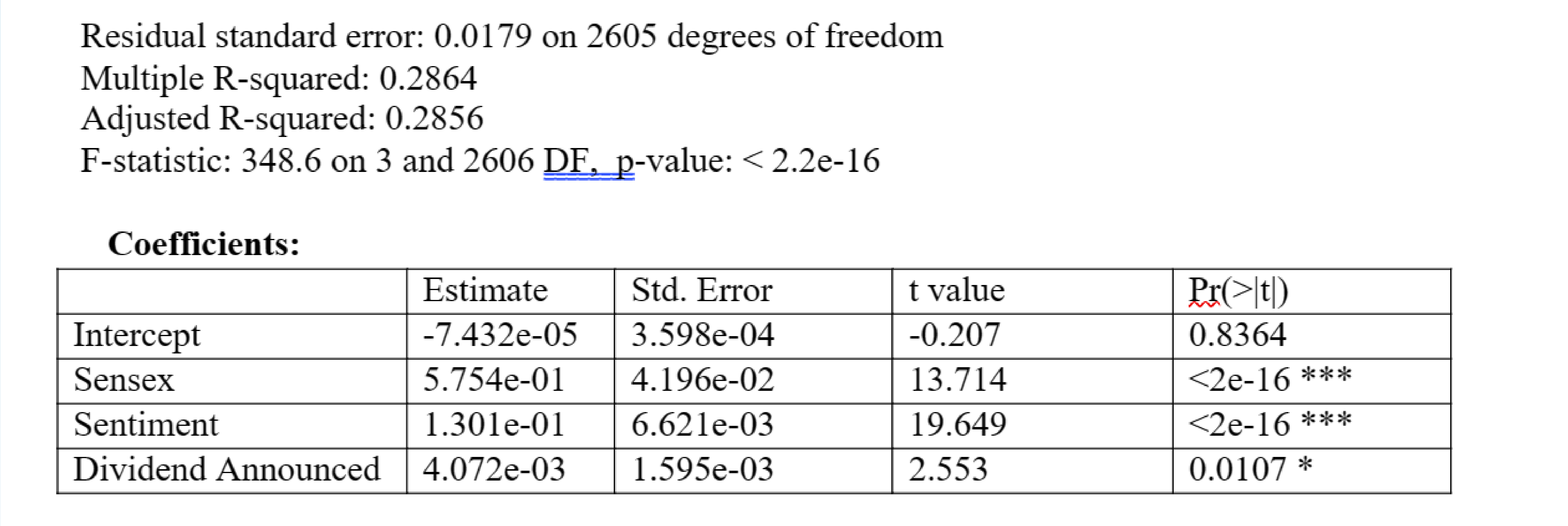
2. Variance Inflation Factor(VIF): Obtained by regressing each variable with the other independent variables and using the R2 values to compute VIF.

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

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.

MACHINE LEARNING

- **MLR Results after removing NIFTY:**



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:

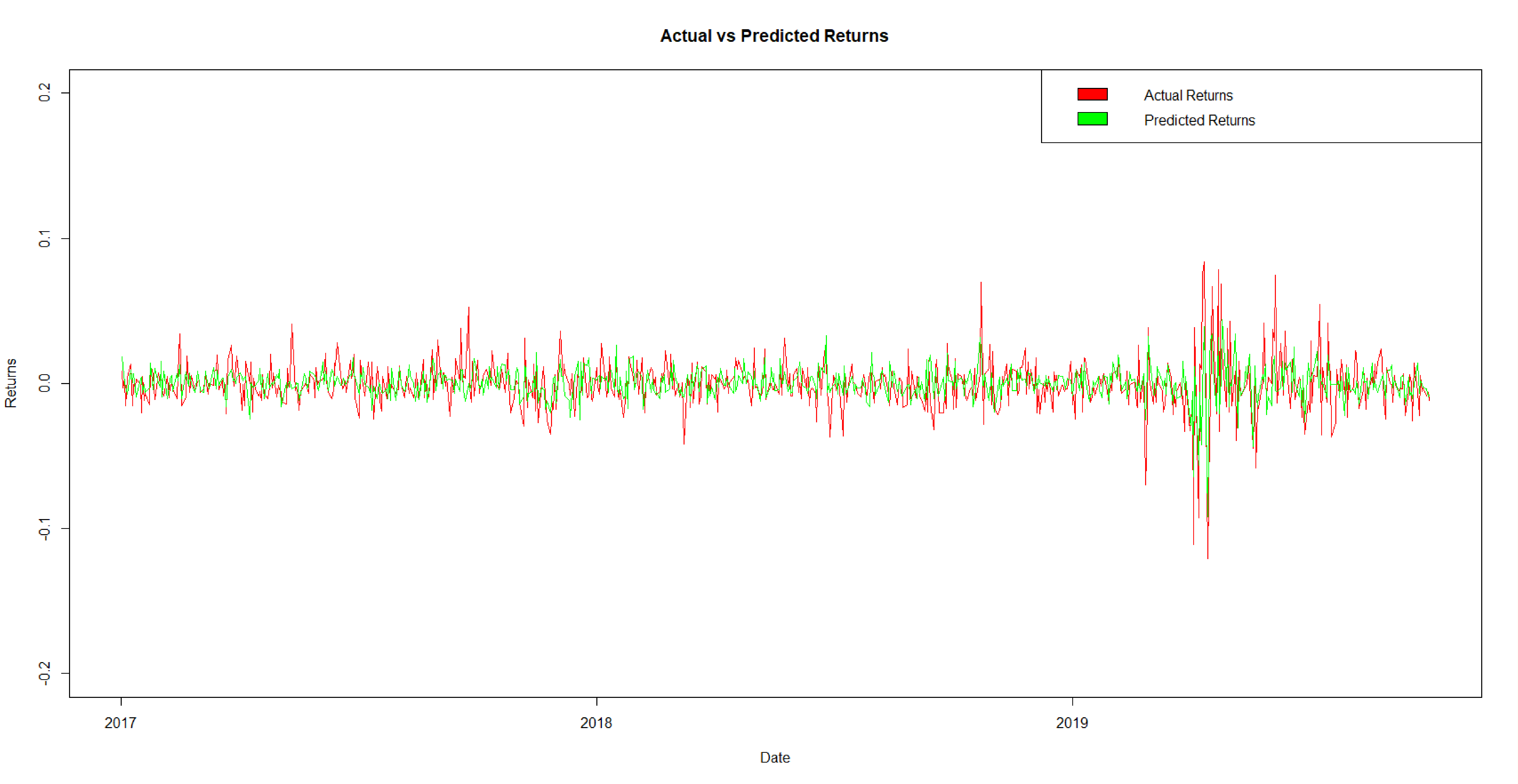
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-7.432e-05	3.598e-04	-0.207	0.8364
Sensex	5.754e-01	4.196e-02	13.714	<2e-16 ***
Sentiment	1.301e-01	6.621e-03	19.649	<2e-16 ***
Dividend Announced	4.072e-03	1.595e-03	2.553	0.0107 *

The values seem to be well fitted and statically significant.

MACHINE LEARNING

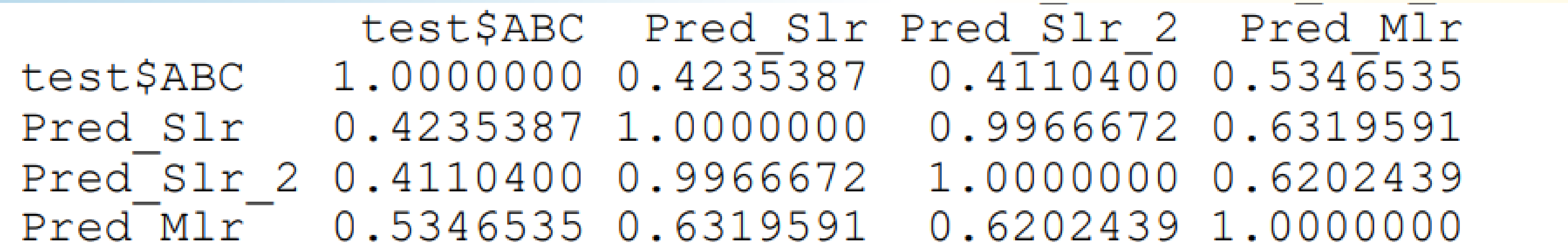
- **Prediction:**

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



EVALUATING MODEL PERFORMANCE

1. Correlation of Predicted Data with Actual Data:



```
             test$ABC  Pred_Slr Pred_Slr_2  Pred_Mlr
test$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.
- - 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 (more closely compared to the other models.

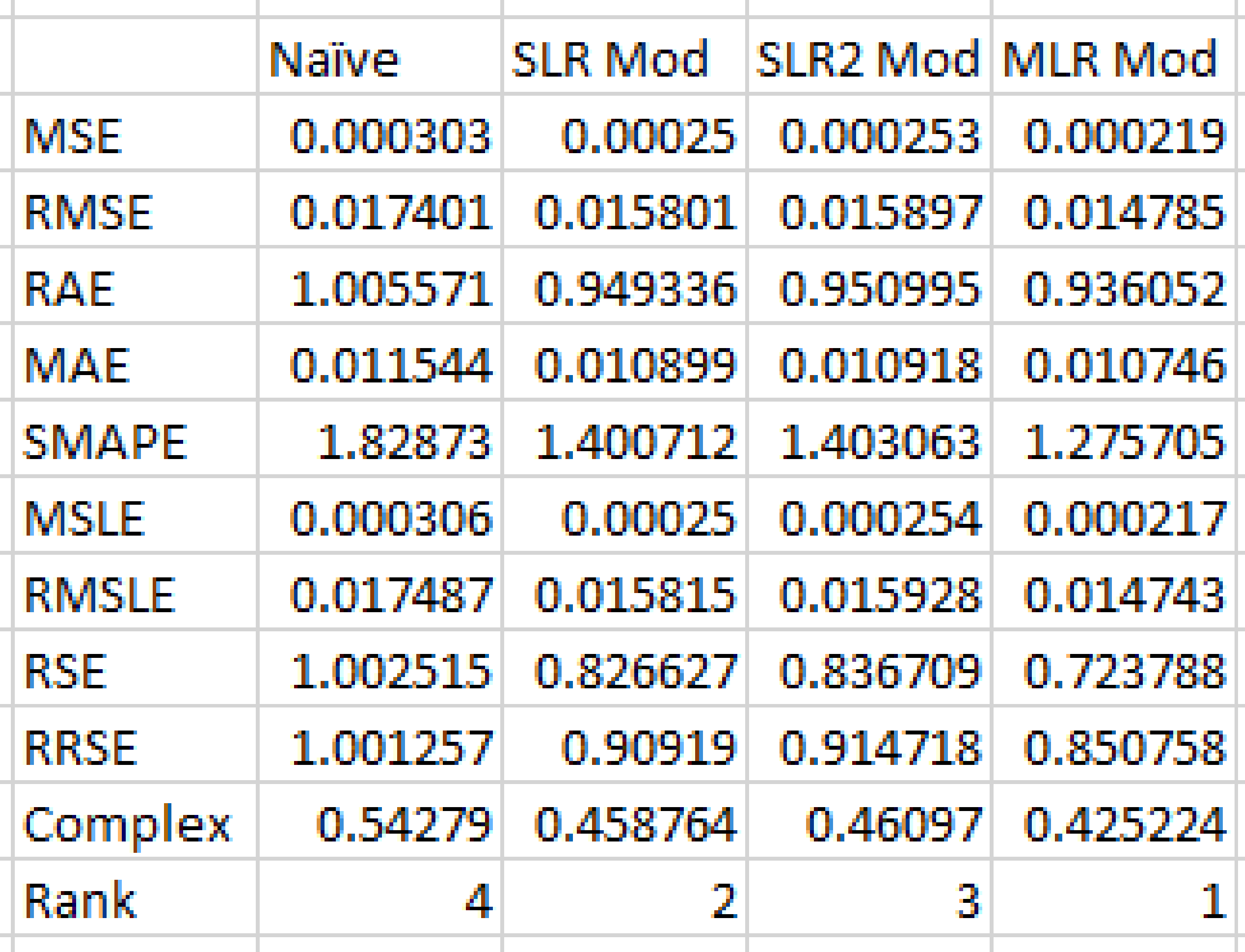
EVALUATING MODEL PERFORMANCE

2. Assessment of Error Metrics:

- Various error metrics (MSE, RMSE, RAE, MAE, SMAPE, MSLE, RMSLE, RSE, RRSE) provide unique insights into model accuracy and fit.
- Utilizing the mean of errors for model comparison aids in comparative analysis.
- Choosing the most appropriate model demands considering multiple error metrics.

3. Model Rankings:

EVALUATING MODEL PERFORMANCE



	Naïve	SLR Mod	SLR2 Mod	MLR Mod
MSE	0.000303	0.00025	0.000253	0.000219
RMSE	0.017401	0.015801	0.015897	0.014785
RAE	1.005571	0.949336	0.950995	0.936052
MAE	0.011544	0.010899	0.010918	0.010746
SMAPE	1.82873	1.400712	1.403063	1.275705
MSLE	0.000306	0.00025	0.000254	0.000217
RMSLE	0.017487	0.015815	0.015928	0.014743
RSE	1.002515	0.826627	0.836709	0.723788
RRSE	1.001257	0.90919	0.914718	0.850758
Complex	0.54279	0.458764	0.46097	0.425224
Rank	4	2	3	1

The overall as well as the individual ranking of the models are same, that is

1. **MLR**
2. **SLR**
3. **SLR Non Linear**
4. **Naive**

MLR model seems to work the best.

Thank You