PYTHON PROJECT

TOPIC: Predictive Modelling of stock Prices using Linear Regression

---

Group 5:
441: Vansh Gupta
459: Chirag Sharma

DATA ANALYSIS AND INTERPRETATION

```
stock.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 639 entries, 0 to 638
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Date       639 non-null    object 
 1   Open       639 non-null    float64
 2   High       639 non-null    float64
 3   Low        639 non-null    float64
 4   Close      639 non-null    float64
 5   Adj Close  639 non-null    float64
 6   Volume     639 non-null    int64  
dtypes: float64(5), int64(1), object(1)
memory usage: 35.1+ KB
```

1. The dataset consists of 639 rows and 7 columns.

2. The data types include:
object: Date
float64: Open, High, Low, Close, Adj Close
int64: Volume

3. There are no null or NA values in the dataset.

EXPLANATORY DATA ANALYSIS

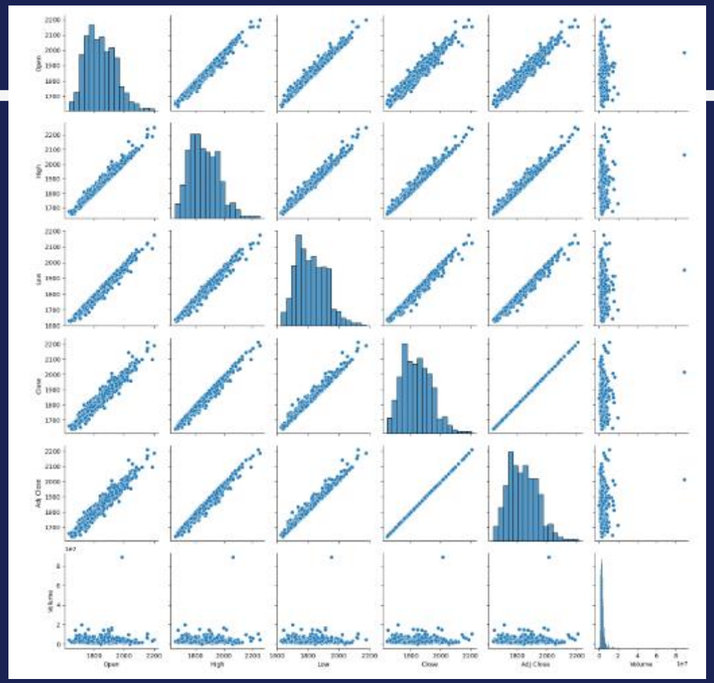
>>> Most variables show normal distributions, but Volume is peaked at a specific time.

>>> Correlation Analysis:

- The Open, High, Low, Close, and Adjusted Close Prices show substantial positive correlations, suggesting strong similarity in their trends.
- Conversely, Volume exhibits weaker correlations with the other variables.
- Volume seems to display no or negative correlation with the other variables, hinting that its changes might not directly align with price fluctuations.

>>> The Date variable is not considered significant in linear regression modeling due to its lack of numerical relevance, making it unsuitable as an explanatory variable.

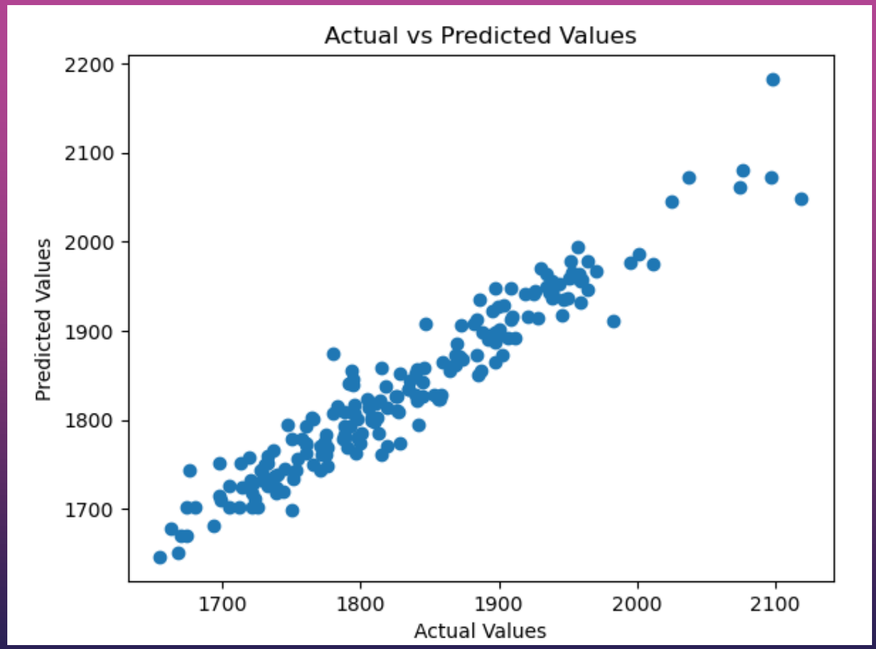
>>> The response variable chosen for the linear regression model is the closing price, with the explanatory variables being the open, high, low, adjusted close prices, and volume.



Candlestick Graph for given data

BUILDING AND FITTING THE MODEL

- We added additional columns such as Open-close, High-Low, High/Low, Next_Open, and Next_Close to analyze the trend effectively.
- Linear regression analysis is used to forecast the closing price by utilizing explanatory variables. The dataset is divided into a 70% training set and a 30% test set to both train the model and assess its accuracy.
- To identify the most relevant explanatory variables from a pool of 10 variables, we utilized the Sequential Feature Selector (SFS) with the R2 (R squared) scoring setting.
- We have two highly explanatory variables, namely High and Next_Open, with an R-squared value of 0.92643, equivalent to 92.64%.
- The analysis concludes that the predicted closing prices closely align with the actual values, indicating an accurate model performance, suggesting that the selected explanatory variables, collectively contribute to the model's ability to predict the closing price.



25.9842

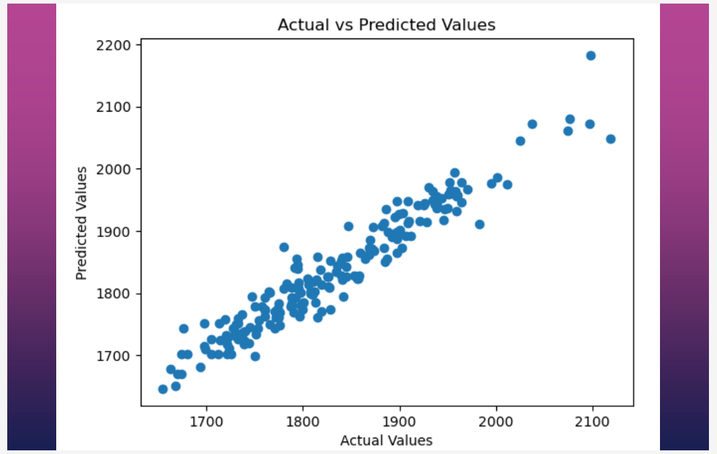
MODEL EVALUATION METRICS

- A lower MAE, MSE, or RMSE suggests improved accuracy, indicating that the model's performance appears satisfactory with relatively low error metrics.
- The R^2 value in regression measures how much of the target variable's variance is explained by the features. An R^2 close to 1 indicates a good fit, with a value of around 0.9264 showing the model explains 92.64% of the target variable's variance, indicating strong predictive ability.
- The residuals are normally distributed, showing that the model meets its assumptions and effectively captures data patterns, explaining target variable variability.

MAE (Mean Absolute Error)
19.5385

MSE (Mean Squared Error)
672.8018

RMSE (Root Mean Squared Error)
25.9384



ASSUMPTION TESTING

1. AUTOCORRELATION TEST:

The Durbin-Watson test statistic, around 1.9522, suggests little autocorrelation in residuals, approaching the optimal value of 2, which is essential for independence in linear regression.

2. MULTICOLLINEARITY TEST:

"Next_Open" and "High/Low" have low VIF values, suggesting minimal multicollinearity. In contrast, "Adj Close" has a higher VIF value (6.340360e+06), indicating substantial multicollinearity. Removing highly correlated variables could affect the model's effectiveness, so retaining them may be beneficial.

3. NORMALITY OF ERROR TEST:

Shapiro Wilks test statistic is 0.9867 which is close to 1, thereby residuals appear relatively normally distributed. The p-value of 0.0689 is greater than 0.05, indicating no significant deviation from normality.

4. HOMOSCEDASCITY TEST (Errors have constant variance):

The Breusch-Pagan test statistic is 0.00043, indicating smaller-than-expected residual variance. The p-value of 7.1397e-08 suggests evidence to reject the null hypothesis of homoscedasticity. Thus, the residuals appear heteroscedastic, which can impact the reliability and validity of the regression analysis.

TESTING THE MODEL WITH VALIDATION DATASET

The validation dataset was imported, and a new column named 'Close' was created with 'NaN' values. Subsequently, 'Close' was utilized as the target variable, while other variables served as predictors to predict the closing prices.

Summing up, the model demonstrates strong predictive skills in forecasting closing prices through selected explanatory factors. The assessment encompassed a range of tests, such as assumption validation and multicollinearity verifications, along with metrics like MAE, MSE, RMSE, and R-squared. In general, the model exhibits effective performance with minimal errors and notable explanatory power, as evidenced by a high R-squared value.

R-SQUARED EVALUATION FOR RIDGE, LASSO AND ELASTIC NET

Linear Regression

The R-squared value of "0.926534" suggests that around 92.6534% of the variability in the dependent variable is accounted for by the independent variables in the linear regression model.

Ridge Regression

The R-squared value of 0.926568 suggests that around 92.6568% of the variability in the dependent variable is accounted for by the independent variables in the linear regression model.

Rasso Regression

The R-squared value of "0.9267909 " suggests that around 92.67909% of the variability in the dependent variable is accounted for by the independent variables in the linear regression model.

Linear Regression

The R-squared value of 0.926728 suggests that around 92.6728% of the variability in the dependent variable is accounted for by the independent variables in the linear regression model.

Thank You