

**Technological Institute of the Philippines****Quezon City - Computer Engineering**

---

Course Code:	CPE 019
Code Title:	Emerging Technologies in CpE 2 - Fundamentals of Computer Vision
2nd Semester	AY 2023-2024

---

**ACTIVITY NO.****PRELIM EXAMINATION**

<b>Name</b>	Dela Cruz, Irish & Gullas, Rainer
<b>Section</b>	CPE32S3
<b>Date Performed:</b>	02/28/2024
<b>Date Submitted:</b>	03/06/2024
<b>Instructor:</b>	Engr. Roman M. Richard

---

## OBJECTIVES

In this exam, you will show the application of the following algorithms:

### Part 1: Linear Regression

- Singular Linear Regression
- Multiple Linear Regression
- Polynomial Linear Regression

### Part 2: Logistic Regression

### Part 3: Decision Tress

### Part 4: Random Forest

## SCENARIO / BACKGROUND

In this exam, we will show each application of algorithms that listed above, including the datasets which contains the 2,516 entries.

The tasked is to analyzed historical stock market data for the past 10 years to identify patterns and build predictive models for stock price movements.

The goal is to develop models that can accurately predict wheter the stock price increase or decrease in the future based on various factors.

## REQUIRED RESOURCES

- PC / Laptop with Internet Access
- **Python libraries:** pandas, sklearn, IPython.display, numpy, and seaborn.
- **Additional application:** Graphviz
- **Datafiles:** Meta Stock Price Technical Indicator.csv

✓ The datasets contain the following variables:

Variable	Description
1. Date	Date for which data is recorded
2. Open	Opening price stock on trading day
3. High	Highest price at which stock traded on during trading day
4. Low	Lowest price at which stock traded on during trading day.
5. Close	Final price at which stock is valued for the day
6. Volume	No. of shares or contrast traded in the security/market during a given period.
7. RSI_7	<b>7-day Relative Strength Index</b> <ul style="list-style-type: none"> <li>• Measure the magnitude of recent prices changes to evaluate overbought/oversold conditions in price of stock</li> </ul>
8. RSI_14	<b>14-day Relative Strength Index</b> <ul style="list-style-type: none"> <li>• Calculated over 14 days</li> </ul>
9. CCI_7	<b>7-day Commodity Channel Index</b> <ul style="list-style-type: none"> <li>• Measure the difference between current price and historical average price</li> </ul>
10. CCI_14	<b>14-day Commodity Channel Index</b> <ul style="list-style-type: none"> <li>• 14 days for medium-term trends</li> </ul>

11. SMA_50	<b>50-day Simple Moving Average</b> <ul style="list-style-type: none"> <li>• Average the closing prices of stock over the past 50 days.</li> </ul>
12. EMA_50	<b>50-day Exponential Moving Average</b> <ul style="list-style-type: none"> <li>• Weight to recent prices</li> </ul>
13. SMA_100	<b>100-day Simple Moving Average</b> <ul style="list-style-type: none"> <li>• Average of closing prices of stock for over the past 100 days</li> </ul>
14. EMA_100	<b>100-day Exponential Moving Average</b> <ul style="list-style-type: none"> <li>• Responsive to recent price changes</li> </ul>
15. MACD	<b>Moving Average Convergence Divergence</b> <ul style="list-style-type: none"> <li>• Show the relationship between 2 moving averages of stock price</li> </ul>
16. Bollinger	Price envelope
17. TrueRange	Measure the volatility that consider the range between high, low, and close of stock
18. ATR_7	<b>7-day Average True Range</b> <ul style="list-style-type: none"> <li>• measures the market volatility by decomposing the entire range of stock for that period</li> </ul>
19. ATR_14	<b>14-day Average True Range</b>

## ✓ Import the libraries and data

```
#importing the important libraries that needed in the study
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

```
#load the Dataset and named as "stock"
stock = pd.read_csv("/content/Meta Stock Price Technical Indicators (10 Years).csv")
```

```
#select the first 20 rows of the data
stock.head(20)
```

	date	open	high	low	close	volume	rsi_7	rsi_14	
0	2014-01-02	54.830002	55.220001	54.189999	54.709999	43195500	51.917475	58.077822	-6
1	2014-01-03	55.020000	55.650002	54.529999	54.560001	38246200	50.604988	57.387622	-4
2	2014-01-06	54.419998	57.259998	54.049999	57.200001	68852600	67.483917	65.221525	4
3	2014-01-07	57.700001	58.549999	57.220001	57.919998	77207400	70.672584	67.003189	15
4	2014-01-08	57.599998	58.410000	57.230000	58.230000	56682400	72.049420	67.768804	10
5	2014-01-09	58.650002	58.959999	56.650002	57.220001	92253300	61.139242	62.667059	6
6	2014-01-10	57.130001	58.299999	57.060001	57.939999	42449500	65.485896	64.706800	5
7	2014-01-13	57.910000	58.250000	55.380001	55.910000	63010900	47.872555	55.499705	-3
8	2014-01-14	56.459999	57.779999	56.099998	57.740002	37503600	59.366740	60.900791	-1
9	2014-01-15	57.980000	58.570000	57.270000	57.599998	33663400	58.220894	60.297824	4
10	2014-01-16	57.259998	58.020000	56.830002	57.189999	34541800	54.619116	58.472055	-2
11	2014-01-17	57.299999	57.820000	56.070000	56.299999	40849200	47.220911	54.606761	-9
12	2014-01-21	56.599998	58.580002	56.500000	58.509998	48669200	62.094919	61.425729	8

```
#Calculating the summary statistics of data
stock.describe()
```

	open	high	low	close	volume	rsi_7	
<b>count</b>	2516.000000	2516.000000	2516.000000	2516.000000	2.516000e+03	2516.000000	2516
<b>mean</b>	178.035517	180.328621	175.826359	178.129122	2.617010e+07	54.339039	54
<b>std</b>	81.123720	82.200952	80.118216	81.166653	1.782460e+07	15.808831	11
<b>min</b>	54.020000	54.939999	51.849998	53.529999	5.467500e+06	14.083263	21
<b>25%</b>	115.787498	117.452497	114.007502	115.565003	1.563175e+07	43.513651	46
<b>50%</b>	170.114998	172.110001	168.224998	170.245002	2.106275e+07	55.357801	54
<b>75%</b>	220.297504	221.827499	216.492501	219.864994	3.022008e+07	66.148583	62
<b>max</b>	381.679993	384.329987	378.809998	382.179993	2.323166e+08	93.579562	86

```
#Provides the summary of Dataframe.
#Including the index, columnss, data type, and non-null values.
stock.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2516 entries, 0 to 2515
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   date                  2516 non-null  object
1   open                  2516 non-null  float64
2   high                  2516 non-null  float64
3   low                   2516 non-null  float64
4   close                 2516 non-null  float64
5   volume                2516 non-null  int64
6   rsi_7                 2516 non-null  float64
7   rsi_14                2516 non-null  float64
8   cci_7                 2516 non-null  float64
9   cci_14                2516 non-null  float64
10  sma_50                2516 non-null  float64
11  ema_50                2516 non-null  float64
12  sma_100               2516 non-null  float64
13  ema_100               2516 non-null  float64
14  macd                  2516 non-null  float64
15  bollinger             2516 non-null  float64
16  TrueRange             2516 non-null  float64
17  atr_7                 2516 non-null  float64
18  atr_14                2516 non-null  float64
19  next_day_close        2516 non-null  float64
dtypes: float64(18), int64(1), object(1)
memory usage: 393.2+ KB
```

```
#Checking if there's a missing or null values in Dataframe.  
#isnull() and sum() calculates the total number of missing values for each column.  
stock.isnull().sum()
```

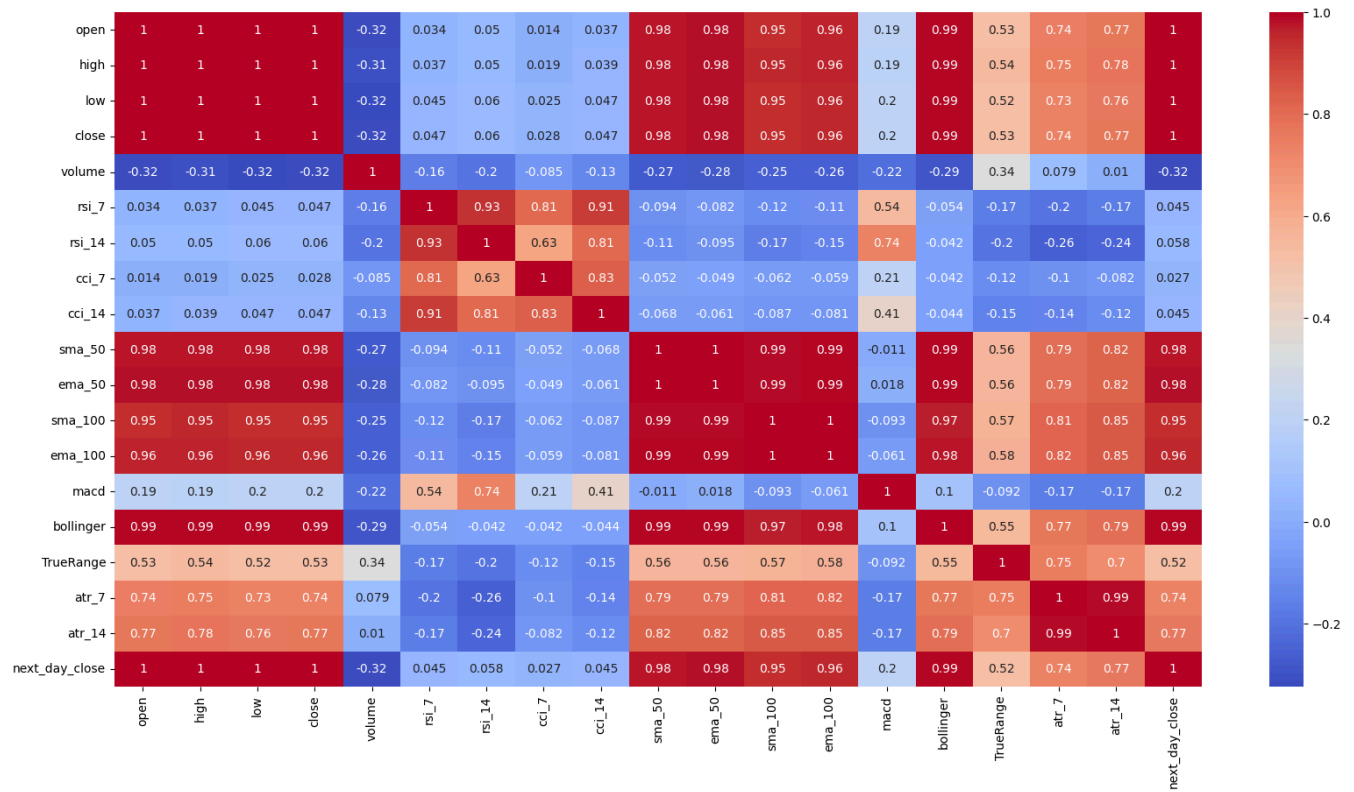
```
date          0  
open          0  
high          0  
low           0  
close         0  
volume        0  
rsi_7         0  
rsi_14        0  
cci_7         0  
cci_14        0  
sma_50        0  
ema_50        0  
sma_100       0  
ema_100       0  
macd          0  
bollinger     0  
TrueRange     0  
atr_7         0  
atr_14        0  
next_day_close 0  
dtype: int64
```

## ✓ Analysis

Once there's missing values the statistic analysis can lead to inaccurate results and predictions. That's why we check if there's a missing value above to fill the missing values to ensure that the analysis will be based on reliable information. Also, there's algorithms cannot handle missing values and produces errors.

```
#importing seaborn for better visualization with color palettes that looks professional  
import seaborn as sns  
import matplotlib.pyplot as plt  
  
#annotation - used to display actual values by each cell  
#color map - 'coolwarm' for better visualization  
plt.figure(figsize=(20,10))  
sns.heatmap(stock.corr(), annot= True, cmap='coolwarm')  
plt.show()
```

```
<ipython-input-25-d23ea36618d7>:6: FutureWarning: The default value of numeric_only in [
sns.heatmap(stock.corr(),  annot= True, cmap='coolwarm')
```



# Analysis

We used heatmap with coolwarm tone to easily visualized the correlation coefficient of the data that indicates the relationship between 2 or more variables if they are closely related to each other.

Here in the visual aids, when the diagonal indicate as 1 it signifies a perfect positive correlation. The independent variables (open, high, low, close) have a perfect positive correlation with dependent variable (next\_day\_close) that can used to predict or explain the values with high certainty.

## Part 1: Linear Regression

In this part of exam, we will build and evaluate 3 types of Linear Regression: **Singular Regression, Multiple Regression, Polynomial Regression**. These models will be trained on labeled dataset containing technical indicator for stock prices over the past 10 years.

Our goal is to predict the classification of Meta Stock Price Movements from 2014 - 2023 based on indicators.

### ✓ Singular Linear Regression

Statistical method used to model relationship between dependent variable (target or response) and independent variables (predictors or features). It assumes if there's a linear relationship between the independent and dependent variables.

```
#Column name
feature_names = ['open','high', 'low', 'close', 'volume', 'rsi_7', 'rsi_14', 'cci_7', 'cci_1

#Display the Feature name and important variables
print("Number of features:", len(feature_names))
print("Number of feature importances:", len(feature_importances))

# Create horizontal bar plot
plt.figure(figsize=(10, 6))
ax = model_ranks.plot(kind='barh', color='pink')

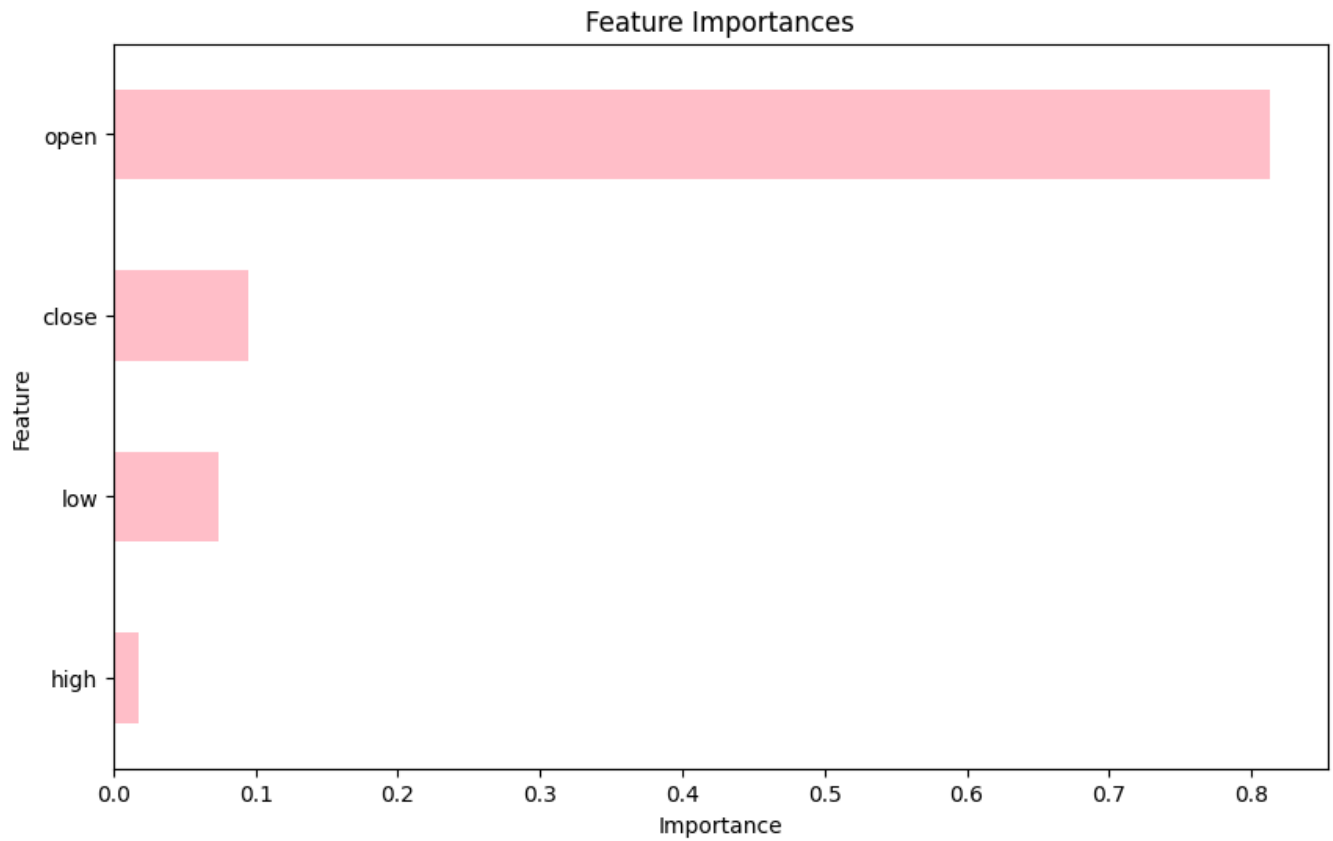
# Set plot title and labels
plt.title('Feature Importances')
plt.xlabel('Importance')
plt.ylabel('Feature')
```



Number of features: 19

Number of feature importances: 4

Text(0, 0.5, 'Feature')



```
import seaborn as sns

# Define predictors and target variable
X = stock[['open']] # Independent variables
y = stock['next_day_close'] # Dependent variable

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=23)

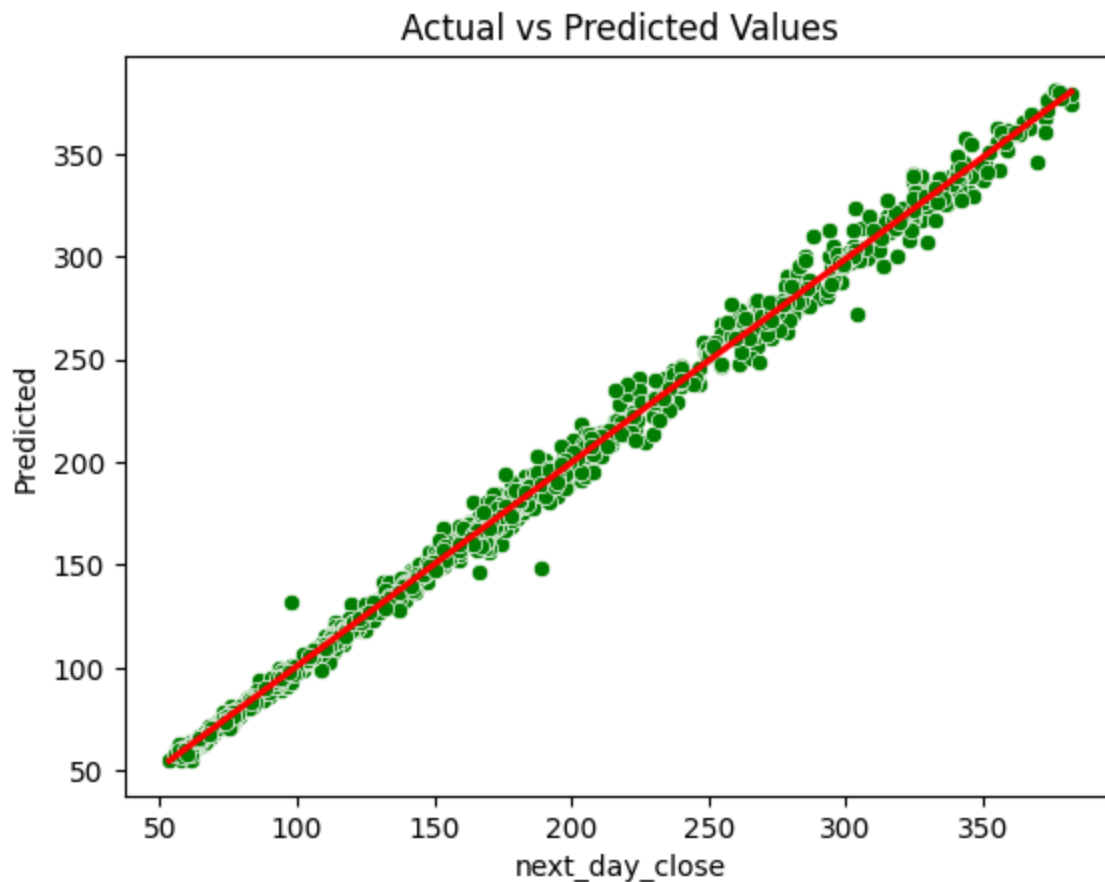
# Create a linear regression model
model = LinearRegression()

# Fit the model to the training data
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Plot actual vs predicted values
sns.scatterplot(x=y_test, y=y_pred, color='green')
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted Values')

# Plot regression line
sns.regplot(x=y_test, y=y_pred, scatter=False, color='red')
plt.show()
```



## Analysis

We can visually assess the performance of linear regression model in predicting the 'next\_day\_closing' prices based on the opening prices. The points are closely align along diagonal line which indicates that the model's prediction closely match to actual values.

### ✓ Evaluate the model (Singular Linear Regression)

```
#Computation for MSE and R-squared score
mse = mean_squared_error(y_test, y_pred) #ave sqr diff. between actual and predicted values
r_squared = r2_score(y_test, y_pred) #measure the proportion of variance in x and y

print("Mean Squared Error:", mse)
print("R-squared Score:", r_squared)
```

```
Mean Squared Error: 27.762230502330063
R-squared Score: 0.995662789441985
```

# Analysis

MSE is approximately 27.76 which is low value indicating that the model's prediction are close to actual values on average, which suggest a good accuracy.

The ranges of r2-score from 0 to 1, the r-squared score of the data was 0.996 closes to 1. Which suggest that 99.6% of independent variables in model are effective in explaining the variation in dependent variable.

**Note:** Closes to 1 means a perfect fit for the variability of response data around it mean

## ✓ Multi-linear Regression

Statistical technique used to analyzed relationship between two or more independent variables and dependent variables by fitting a linear equation to observed data.

```
import seaborn as sns

# Define predictors and target variable
X = stock[['open', 'high', 'low', 'close']] # Independent variables
y = stock['next_day_close'] # Dependent variable

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=23)

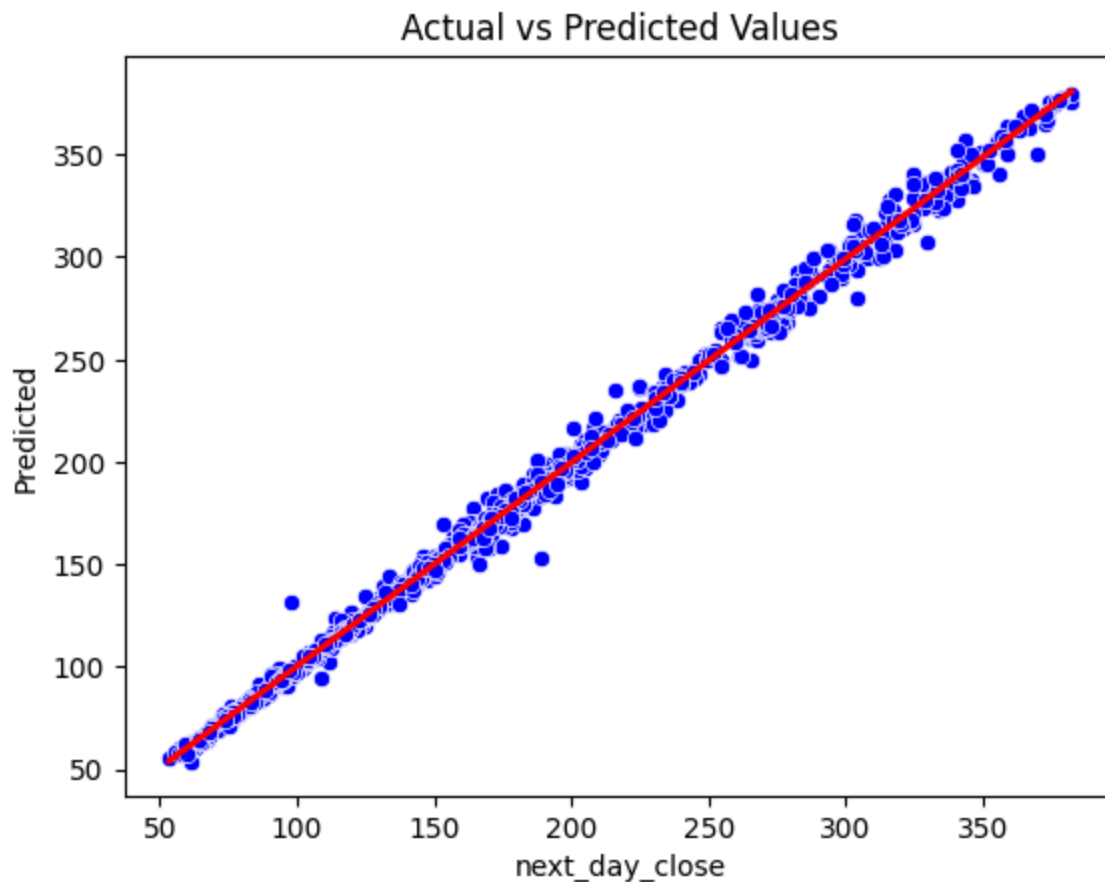
# Create a linear regression model
model = LinearRegression()

# Fit the model to the training data
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Plot actual vs predicted values
sns.scatterplot(x=y_test, y=y_pred, color='blue')
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted Values')

# Plot regression line
sns.regplot(x=y_test, y=y_pred, scatter=False, color='red')
plt.show()
```



## Analysis

The points closely follow the red line, which suggests that the model's prediction are accurate and have a strong linear relationship between the independent variable (open, high, low, close) and dependent variable (next\_day\_close)

Note: If the points from tigh cluster around the regression line, it indicates that the model's prediction closely match the actual values and model capturing the patterns in data effectively.

## ✓ Evaluate the model (Multi-linear Regression)

```
# Calculation of MSE and R2-score
mse = mean_squared_error(y_test, y_pred)
r_squared = r2_score(y_test, y_pred)
print("Mean Squared Error:", mse)
print("R-squared Score:", r_squared)
```

Mean Squared Error: 19.765984763393266  
R-squared Score: 0.9969120190901752

# Analysis

The MSE is approximately 19.77 - low value which means the predicted values are close to the actual values. While the  $r^2$ -score is equivalent to 0.997 or 99.7% which explains the model effectively captures the relationship of independent and dependent variable as evidenced by the high R-squared and low MSE.

## ✓ Polynomial Regression

Regression analysis used to model the relationship between a dependent variable and one or more independent variable by fitting the polynomial function to observed data.

Note: Polynomial Regression fits a curve.

```
# Plotpolly - visualization of polynomial regression models.
# model - callable variable that can pass to independent and get vack the predicted values c
# independent - used in regression / dependent - observed values
# name - label for x-axis (independent variable)
def PlotPolly(model, independent_variable, dependent_variabble, Name):
    # Genearte 100 space between 15 & 55, used to create smooth curve
    x_new = np.linspace(15, 55, 100)
    # predict the dependent variable based on model and x variable
    y_new = model(x_new)

    # Plots the original data points as dots and polynomial curve
    plt.plot(independent_variable, dependent_variabble, '.', x_new, y_new, '-')
    #Title Plot
    plt.title('Polynomial Fit with Matplotlib for Stock Meta ~ Length')
    # Current axes instance
    ax = plt.gca()
    # Set the background color
    ax.set_facecolor((0.898, 0.898, 0.898))
    # Current figure instance
    fig = plt.gcf()
    # label for x-axis as name
    plt.xlabel(Name)
    # plot y-axis as next_day_close
    plt.ylabel('next_day_close')
    # display plot
    plt.show()
```

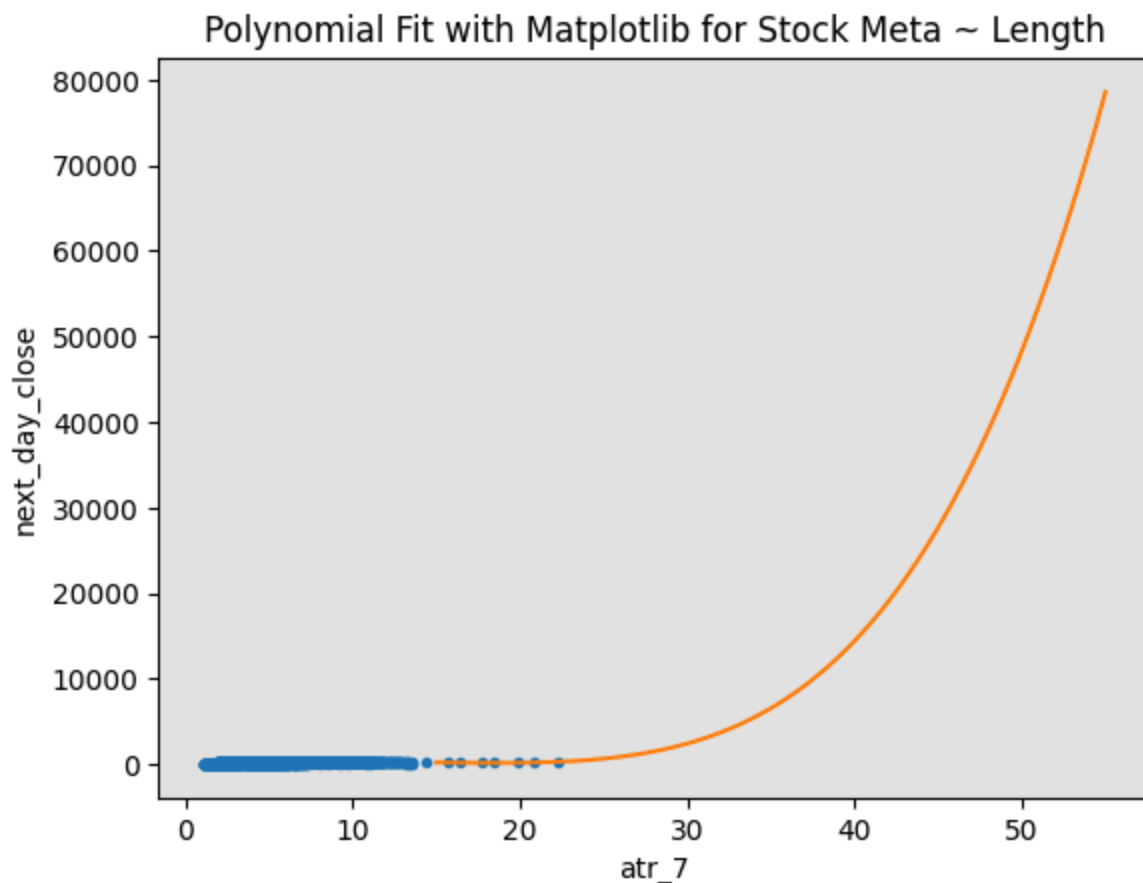
## ✓ Analysis

Designed that provide a convenient way to visualized the polynomial fit for regression model along side the original data points. It helps to capture the relationship between independent and dependent variable

```
# Two important variables that set as example for this model
x = stock['atr_7']
y = stock['next_day_close']

# Inside the polynomial model 'p' used to predict values of dependent based on independent v
f = np.polyfit(x, y, 4)
p = np.poly1d(f)

# Plot the polynomial fit
PlotPolly(p, x, y, 'atr_7')
```



## ✓ Analysis

The graph show the relationship between 2 variables: 'atr\_7' and 'next\_day\_close'. The dots represent as actual stock prices. The blue lines represents the polynomial fit that capture the general trend of stock as 'atr\_7' changes.

This suggest that the volatility of stock increases ('atr\_7' value), the prices tend to decrease or show movement.

y\_pred

```
array([100.49389125,  98.99541718, 103.73844079, ..., 237.93001415,  
       229.14277395, 234.43512425])
```

## ✓ Evaluation of Polynomial Linear Regression

```
# Calculate predicted values using the polynomial regression model  
y_pred = p(x)
```

```
# Calculate R-squared score  
r_squared = r2_score(y, y_pred)  
print('R-squared score:', r_squared)
```

```
# Calculate Mean Squared Error (MSE)  
mse = mean_squared_error(y, y_pred)  
print('Mean Squared Error:', mse)
```

```
R-squared score: 0.6170762448390501  
Mean Squared Error: 2523.689159316617
```

## Analysis

The MSE was equivalent to 2523.689, indicate the squared difference between the actual and predicted stock prices. The higher MSE suggest a larger predictions error, so there's another factor aside from 'atr\_7' that affecting the stock prices.

r2-score of 0.617 or 61.7% of the variation in dependent variable (next\_day\_close) is explained by indepdent variable ('atr\_7'), which suggest that the model doesn't not explain much of variability in the data.



## Part 2: Logistic Regression

Statistical method used for binary classification task, where target variable has 2 possible outcomes. It used when dependent variable is categorical.

Note: Linear Regression - predicts continuous values, while Logistic Regression - is a binary classification algorithm, it requires the target variable to be categorical with two classes.

### ✓ Ridge Regression instead of Logistic Regression

(continuous values are not suitable for logistic regression)

variant of linear regression that used when the data suffers from multicollinearity (high correlation among predictor variables) or when overfitting in the model. It adds penalty term to ordinary least objective function, which helps to reduce the complexity of model and prevent overfitting.

```
from sklearn.metrics import r2_score

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Creating and fitting the Ridge regression model
ridge_model = Ridge(alpha=1.0) # You can adjust the regularization strength by changing the
ridge_model.fit(X_train, y_train)

# Making predictions on the test set
predictions = ridge_model.predict(X_test)

# Calculating Mean Squared Error
mse = mean_squared_error(y_test, predictions)

# Calculating R-squared score
r_squared = r2_score(y_test, predictions)

print("Mean Squared Error:", mse)
print('R-squared score:', r_squared)
```

```
Mean Squared Error: 1.7115637211977852
R-squared score: 0.9997318045506529
```

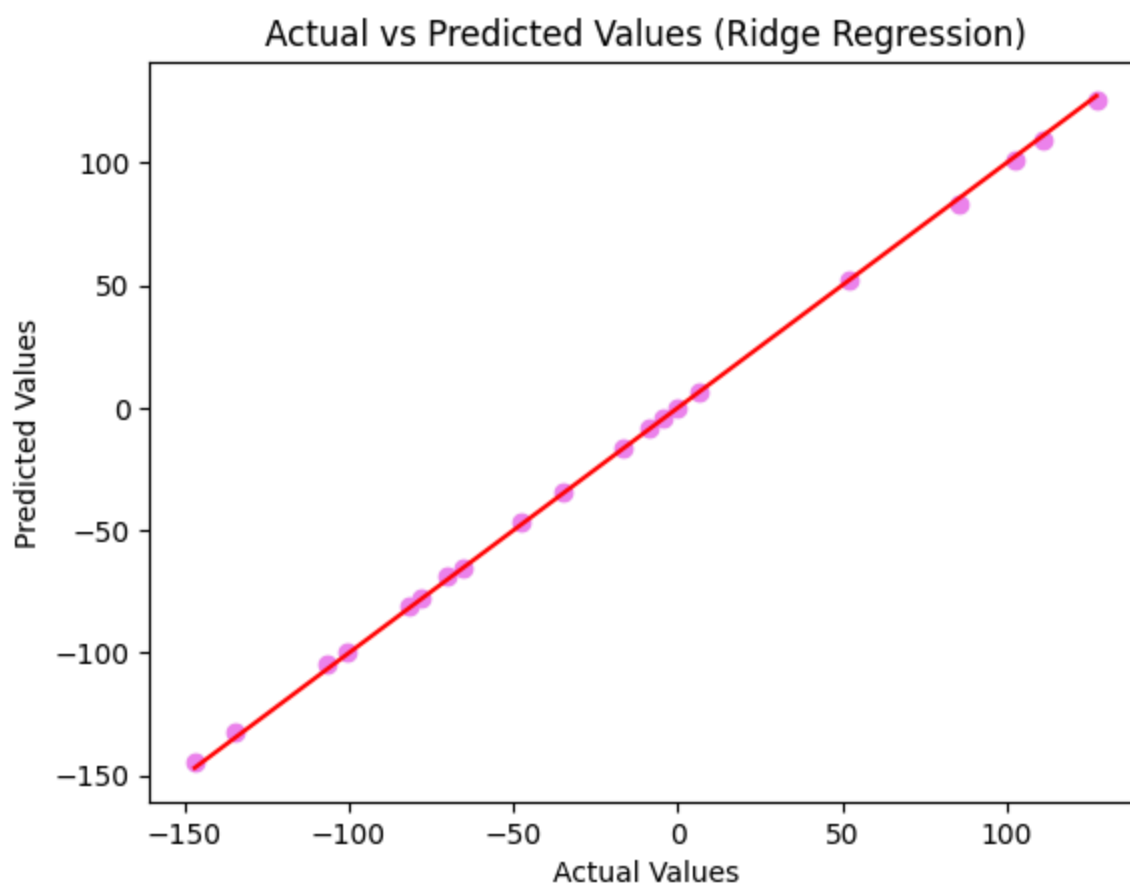
### ✓ Analysis

Our target variables are continuous that's why we used ridge regression as a substitute for the logistic.

The MSE value of approximately 1.71 indicates that, on average, the squared difference between the actual and predicted target values is 1.71 closer to actual value

R-squared score of approximately 0.9997 suggests that the Ridge regression model explains about 99.97% perfectly explain the variance in the target variable.

```
# Plotting the actual vs predicted values
plt.scatter(y_test, predictions, color='violet')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='-')
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs Predicted Values (Ridge Regression)')
plt.show()
```



## Analysis

The visual aod seems that the Actual Value was 50 and 100. Then the Predicted Values are: 100, 50, -50, -100, -150, -150. The model's predictions seem to be lower that actual in some instances and higher to others, However the Actual value of 100 was accurate. Since that all feautres are on the

same scale then it capturing the true relationship between the features and target variable accurately.

## Part 3: Decision Trees

A flowchart structure used for classification or regression, where each internal node represents a test on an attribute (feature), branches represent the outcome of the test, and leaf nodes represent the class label (classification) or numerical value (regression).

Decision tree or if-else statements used to determine the best feature and decision rule for splitting data, minimizing squared error at each node.

### ✓ Installing graphviz for displaying the image

```
!pip install graphviz
```

```
Requirement already satisfied: graphviz in /usr/local/lib/python3.10/dist-packages (0.20.1)
```



### ✓ Importing Libraries

```
from sklearn.tree import DecisionTreeRegressor, export_graphviz
from IPython.display import display, Image
import graphviz
```

```
#data["condition"].fillna(data["condition"].mean(), inplace = True)
```

```
# Dependent variable
y_target = stock["next_day_close"].values
```

```
# 4 important variables
columns = ["open", "close", "low", "high"]
X_input = stock[list(columns)].values
```

```
# Creating Decision Tree Regression
regressor = DecisionTreeRegressor(max_depth=3) # depth 3 meand that tree will make split up

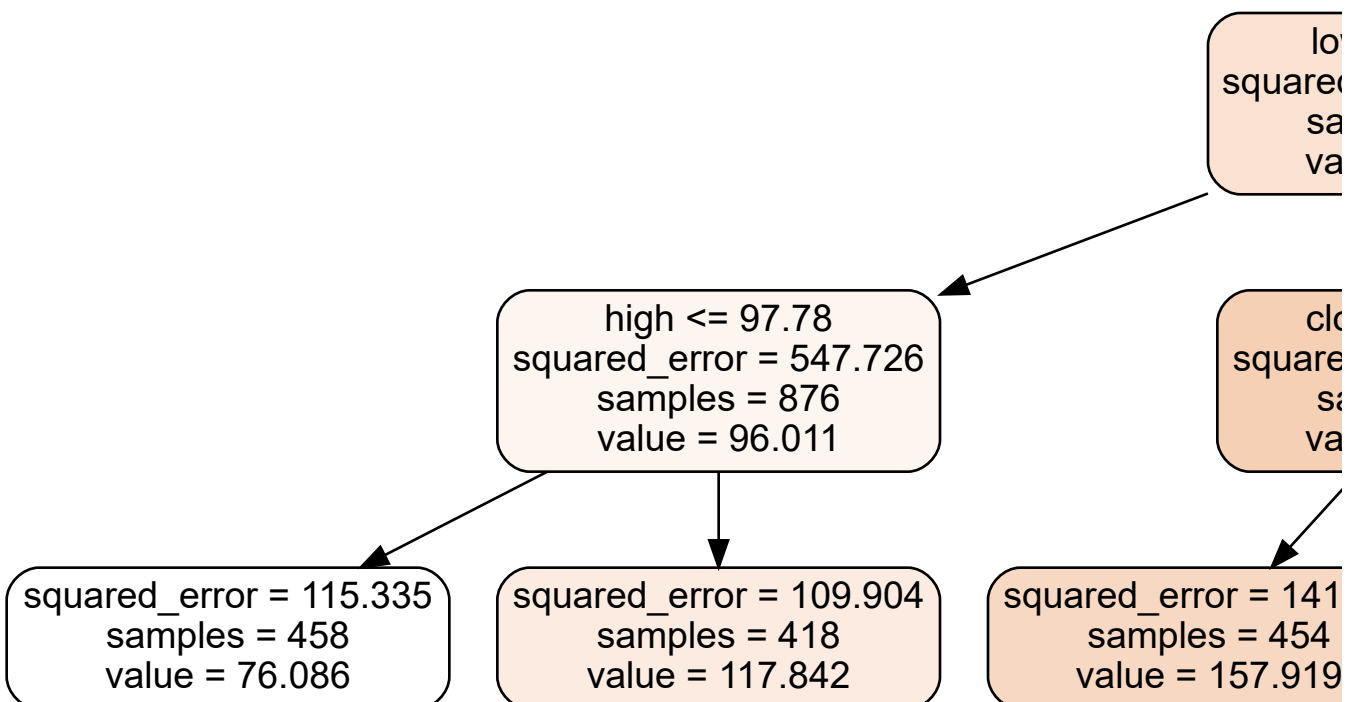
#used to make prediction on new data
regressor.fit(X_input, y_target)
```

```
▼ DecisionTreeRegressor
DecisionTreeRegressor(max_depth=3)
```

```
# exporting the train decision tree model into textual representation in DOT format which is
dot_data = export_graphviz(regressor, out_file=None, feature_names=columns, filled=True, rou
# represent the decision tree graph from DOT data
graph = graphviz.Source(dot_data)

# set the length and width of the graph
graph.format = 'png' # or 'pdf', 'svg', etc.
graph.render("decision_tree_graph", format='png', cleanup=True, view=True, directory=None, c

# Displaying the decision tree
display(graph)
```



# Analysis

The squared errors for specific condition and no. of samples indicates as data points that meet that condition.

For the first line the condition is met ( $\text{high} \leq 97.78$ ), the squared error is 115.335 calculated based on 458 samples.

For the second line the condition is met ( $\text{high} \leq 97.78$ ) is true, used in calculation 76.086

The squared errors and no of sample are calculated in different consditions for ex. " $\text{high} \leq 97.78$ ", " $\text{low} \leq 133.705$ ", and " $\text{close} \leq 173.83$ " represent different rules or threshold for splitting data based on values of features.

The true condition lead to one branch of decision tree as true, false lead to another branch as false.

## ✓ Evaluation for Decision Tree

```
# Predict the target variable using the trained model
y_pred = regressor.predict(X_input)

# Compute Mean Squared Error (MSE)
mse = mean_squared_error(y_target, y_pred)
print("Mean Squared Error:", mse)

# Compute R-squared (R2) score
r_squared = r2_score(y_target, y_pred)
print("R-squared:", r_squared)
```

```
Mean Squared Error: 139.11747590258972
R-squared: 0.9788914628869912
```

# Analysis

Since we used Decision Tree Regressor, the evaluation metrics will be MSE and R2 score.

Results suggest that the regression model has performed well, with a relatively low MSE as 139.12 and a high R2 score as 0.979 Or 97.9%. It indicates that the model's predictions are close to the actual values, and the model explains a large portion of the variability in the target variable.

Note:

- `y_test` or `y_pred` contained continuous values instead of discrete class we can't use confusion matrix for that.
- Decision Tree Classifier used for classification problems where target variable is categorical "class labels".

## ✓ Part 4: Random Forest

An ensemble learning method that combines multiple decision trees to improve prediction accuracy and prevent overfitting.

```
# Contain the values of all columns in data except 'date'  
X = stock.drop(["date"],axis=1).values
```

```
# Removing rows with missing values  
x = stock.dropna(inplace = True)
```

```
# Train the random forest regressor model
regressor = RandomForestRegressor()
regressor.fit(X_train, y_train)

# Get feature importances from the trained random forest regressor model
feature_importances = regressor.feature_importances_

# Get the names of the features
feature_names = ['open', 'high', 'low', 'close']

# Create a Series with feature importances and their corresponding column names
model_ranks = pd.Series(feature_importances, index=feature_names, name='Importance')

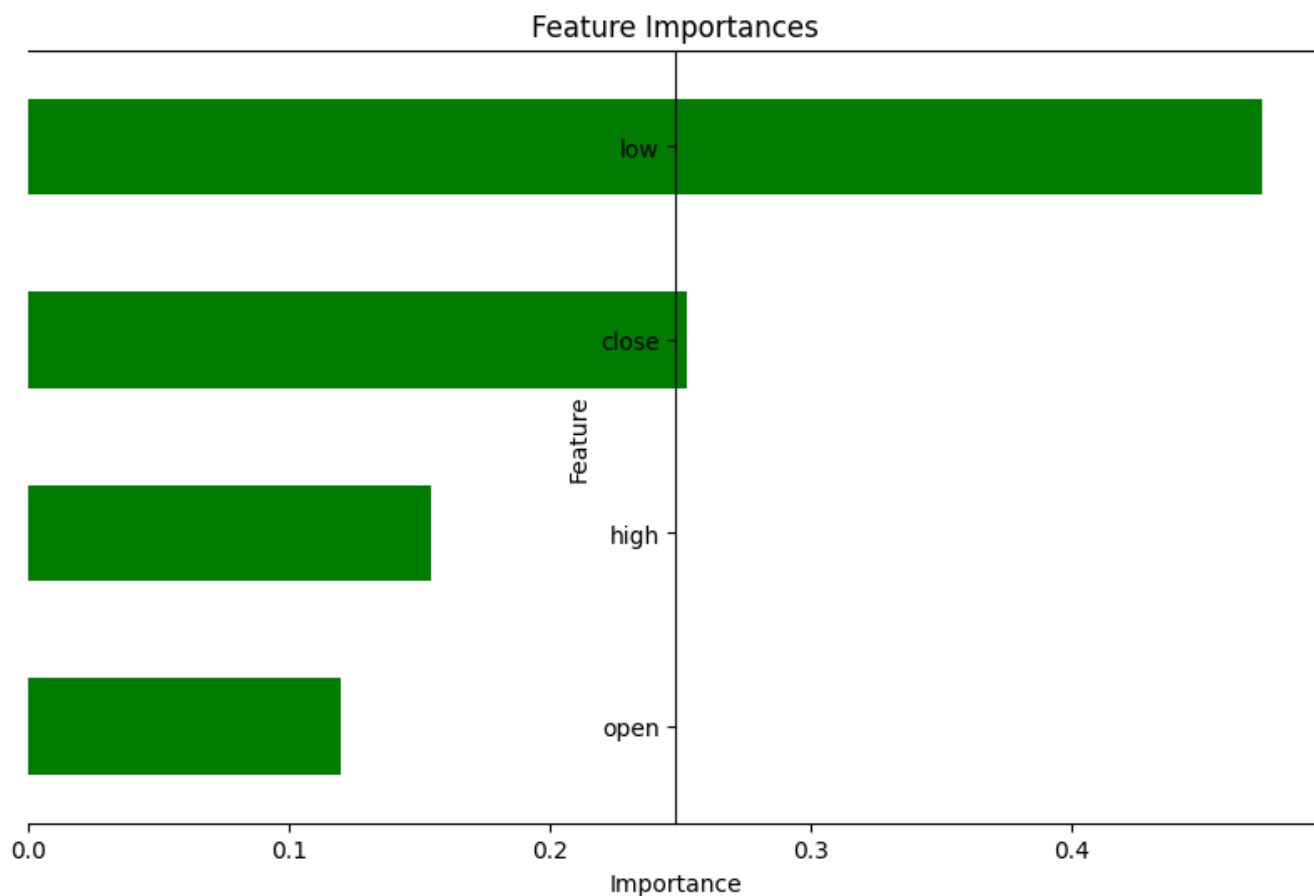
# Sort feature importances
model_ranks = model_ranks.sort_values(ascending=True)

# Create horizontal bar plot
plt.figure(figsize=(10, 6))
ax = model_ranks.plot(kind='barh', color='green')

# Set plot title and labels
plt.title('Feature Importances')
plt.xlabel('Importance')
plt.ylabel('Feature')

# Move x-axis to the center
ax.spines['left'].set_position('center')

# Show plot
plt.show()
```



## ✓ Analysis:

It's very essential to know what's the important variable in the model. So it's easy to identify what features contribute the most model's predictions. Which variables have little to no impact on prediction.

We used to display the importance variable in order for us to provide insight in the relationship between features and target variables that can useful for feature selection.



```

from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

X = stock[['open', 'high', 'low', 'close']] # Independent variables
y = stock['next_day_close'] # Dependent variable

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

# Calling the module
# 100 = parameters/ reproducibility
rf = RandomForestRegressor(n_estimators = 100, random_state=42)
rf.fit(X_train, y_train)

# Generates the predictions in given data
y_train_pred = rf.predict(X_train)
y_test_pred = rf.predict(X_test)

import pandas as pd

# Contains predicted values of training set in y_train_pred
y_train_pred = pd.DataFrame(y_train_pred, columns =['yPredict'])
y_train_pred

```

	yPredict
<b>0</b>	337.834301
<b>1</b>	180.930000
<b>2</b>	111.489999
<b>3</b>	231.566498
<b>4</b>	167.470501
...	...
<b>1756</b>	175.202799
<b>1757</b>	275.835602
<b>1758</b>	127.633797
<b>1759</b>	148.666199
<b>1760</b>	233.130495

1761 rows × 1 columns

## ✓ Analysis

it's important to know the `y_train_pred` for evaluating the performance model on training data. By comparing the predicted values with actual values.

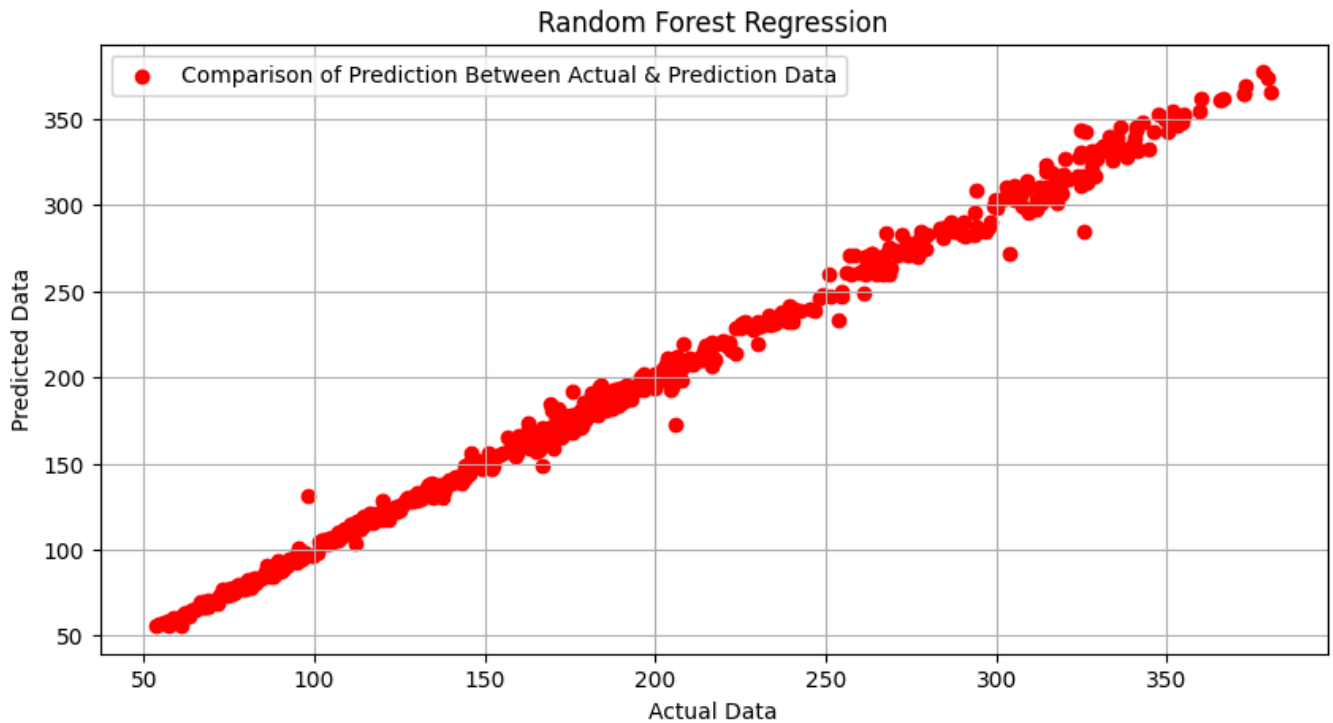
This information can be used to fine-tune the model to prevent overfitting.

```
# Generates prediction in x_test that stored to y_test_pred
y_test_pred

# Size of data and numbers of features.
stock.shape

(2516, 21)

plt.figure(figsize=(10, 5))
plt.scatter(y_test, y_test_pred, color='red', label='Comparison of Prediction Between Actual
plt.legend()
plt.grid()
plt.title('Random Forest Regression')
plt.xlabel('Actual Data')
plt.ylabel('Predicted Data')
plt.show()
```



## Analysis

The visual aid compares the actual and predicted data for random forest regression. It suggests that the model was able to predict values for the test set with accuracy.

### ✓ Evaluation for Random Forest (Actual vs. Predicted)

```
from sklearn.metrics import mean_squared_error, r2_score
```

```
# Make predictions on the training and testing data
```

```
y_train_pred = model.predict(X_train)
```

```
y_test_pred = model.predict(X_test)
```

```
# Calculate MSE for training and testing data
```