Predicting Next-Day Stock Closing Prices Using Multivariate Linear Regression

Introduction

The goal is to predict the next day's closing price of a stock using multiple linear regression. It can be challenging to predict stock prices because there are many variables that affect it. Hence, why multivariate linear regression is used instead of simple linear regression. This allows the model to incorporate multiple features such as the stock's Open, High, Low, Close, and Volume to make more informed predictions. A single stock is used for this prediction but ideally one should train per stock (separately), compare how different stocks behave, group similarly behaving stock together (feature coefficient), and then the model used should be based on their behaviour.

Data Preparation

The dataset consists of historical stock prices fetched using Yahoo Finance. The key features considered for predicting the next day's closing price include:

- Open Price – The stock's price at the start of the trading day.
- High Price – The highest price the stock reached during the day.
- Low Price – The lowest price the stock dropped to during the day.
- Close Price – The stock's price at the end of the trading day.
- Volume – The number of shares traded during the day.

The target variable is the Next day's Close price, which represents the stock's closing price for the following trading day. To ensure data integrity:

- Missing values are handled using forward-fill or dropping NaN values.
- The dataset is split chronologically in a time based manner.

Model Building

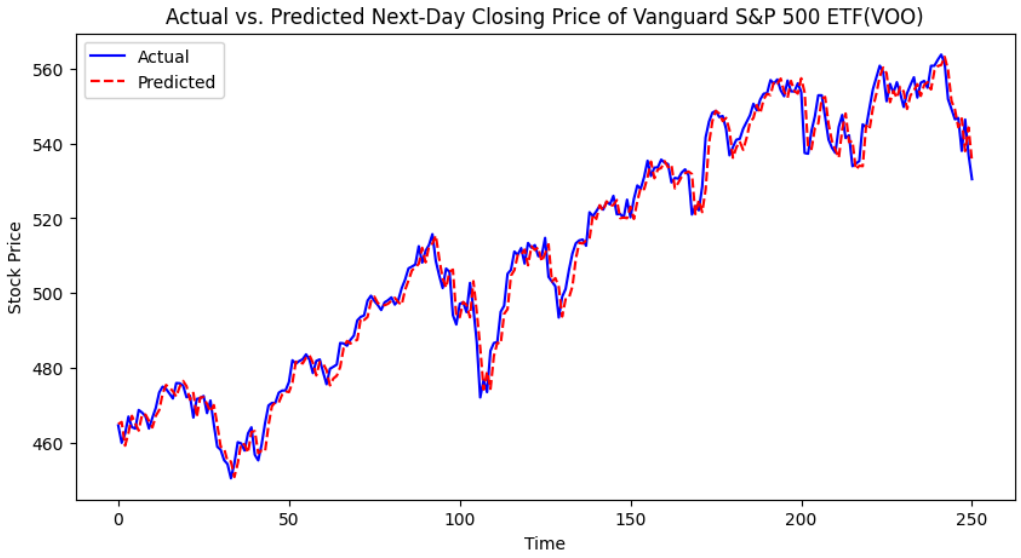
- The dataset is divided into training (80%) and testing (20%) using a time-based split
- A Linear Regression model is trained on the historical data.
- The trained model predicts the next-day closing price using the test data.

Results

The model's performance is evaluated using:

- MSE = 18.0218
  Indicates a well-fitting model
- R^2 Score = 0.9825
  Indicates that the model explains 98.25% of the variance in stock prices

Actual vs. Predicted prices:



Coefficient Analysis

Feature	Coefficient	Analysis
Open	-0.43	A $1 increase in today's open price leads to a $0.43 decrease in tomorrow's close.
High	+0.39	A $1 increase in today's high leads to a $0.39 increase in tomorrow's close.
Low	+0.56	A $1 increase in today's low leads to a $0.56 increase in tomorrow's close.
Close	+0.48	A $1 increase in today's close leads to a $0.48 increase in tomorrow's close.
Volume	+7.76e-8	Minimal effect on price changes.

Limitations & Conclusions

Multivariate linear regression proves to be an effective method for stock price prediction. However, the model assumes linearity, which may not always hold due to market volatility, external factors, and sudden economic changes.