INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

Department of Computer Science & Engineering

Stock Market Prediction System Using Machine Learning (Linear Regression)

Course: CS 306 – Machine Learning

Name: Vemali Deepika

Roll No: 2301243

Stock Market Prediction System Using Machine Learning (Linear Regression)

ABSTRACT:

This project focuses on developing a machine-learning-based stock market prediction system using historical price data of the S&P 500 index.
The system uses numerical features derived from stock OHLCV data to predict the next-day closing price using a Multiple Linear Regression Model.

To enhance predictive accuracy, several technical indicators are engineered, including SMA (Simple Moving Average), EMA (Exponential Moving Average), MACD, lag-based closing prices, and daily returns. The system utilizes time-series aware train–test splitting, evaluation metrics (MAE, RMSE, R^2 Score), and complete visualization to observe residuals, actual vs predicted trends, and error distributions.

This report documents the dataset, preprocessing steps, feature engineering, algorithm details, experimental results, performance metrics, and interpretation of findings.

1. DATASET DESCRIPTION:

The dataset used in this project is a structured CSV file containing daily stock market information for the S&P 500 index.

Columns in the Dataset

Column	Description
Date	Trading date (converted to datetime format)
Open	Opening price of the index
High	Highest price on the day
Low	Lowest price on the day
Close	Closing price (target for prediction)
Volume	Number of shares traded
Adj Close	Adjusted close (removed to avoid redundancy)

These features represent the standard financial OHLCV structure required for time series modeling.

Why These Columns Are Used

- OHLC prices reflect intraday price movement.
- Volume indicates liquidity and market participation.
- Close is the most commonly analyzed price and is used as the prediction target.
- The dataset spans many years, enabling long-term market behavior analysis.

Dataset Preprocessing:

Date Conversion

The Date column is converted to datetime and sorted chronologically.

Removing Redundant Columns

Adj Close is removed as it duplicates "Close" after adjustment.

Missing Values

Missing entries are removed to maintain accuracy.

Volume Cleaning

Volume is standardized as float values.

Index Reset

The entire dataset is cleaned and saved as:
cleaned_dataset.csv

2. FEATURE ENGINEERING

To enhance prediction quality, the following features were created:

Moving Averages

Feature	Description
SMA_10	Average closing price over last 10 days
SMA_20	Average closing price over last 20 days
EMA_12	Exponential moving average (12 days)
EMA_26	Exponential moving average (26 days)

These indicators help capture trend behaviour and smooth fluctuations.

MACD (Moving Average Convergence Divergence)

MACD = EMA_12 – EMA_26
This captures momentum shifts and trend reversals.

Daily Return

Return = percentage change in the closing price.

Lag Features

To capture temporal dependencies:

Feature	Meaning
Close_lag_1	Yesterday's closing price
Close_lag_2	Close from 2 days ago
Close_lag_3	Close from 3 days ago
Close_lag_4	Close from 4 days ago
Close_lag_5	Close from 5 days ago

Target Variable

Target = Close shifted by -1 day
→ Predict tomorrow's closing price.

3. MACHINE LEARNING MODEL

The system uses ***Multiple Linear Regression***, which fits a linear relationship between engineered features and next-day price.

Why Linear Regression?

- Highly interpretable
- No complex hyperparameters
- Fast to train on large datasets
- Suitable for baseline financial prediction
- Works well with numerical and time-based features

Train-Test Split

- **Train:** 80% earliest data
- **Test:** 20% latest data
- ***No shuffling***, because time-series order must be preserved.

4. PERFORMANCE EVALUATION

The following evaluation metrics were used:

Metric	**Meaning**
MAE	Average absolute prediction error
RMSE	Root mean squared error (punishes large errors)
R^2 Score	Ability of the model to explain variance

These metrics are saved in: **metrics.csv**

Evaluation Results:

The results indicate:

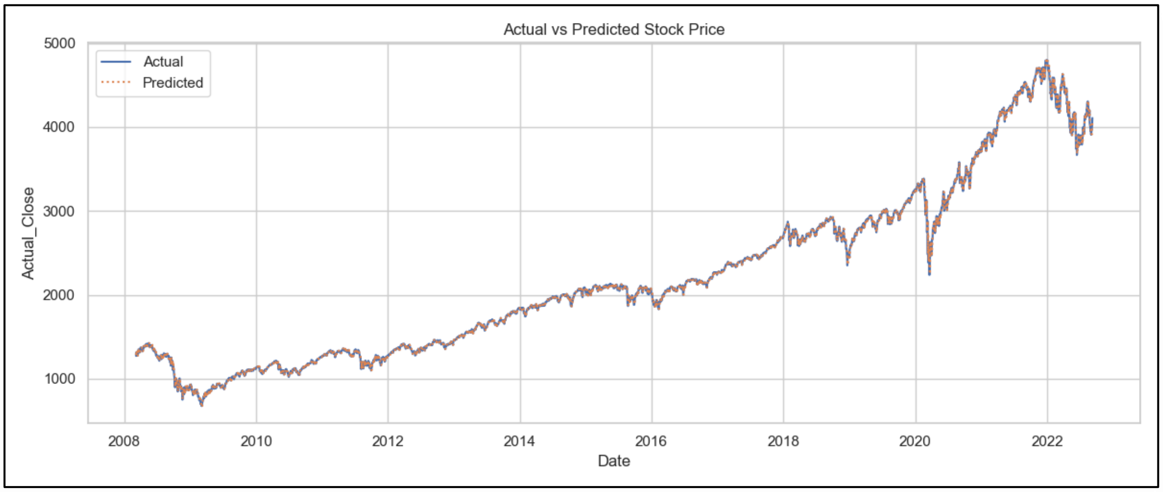
- Errors are relatively small considering large price values.

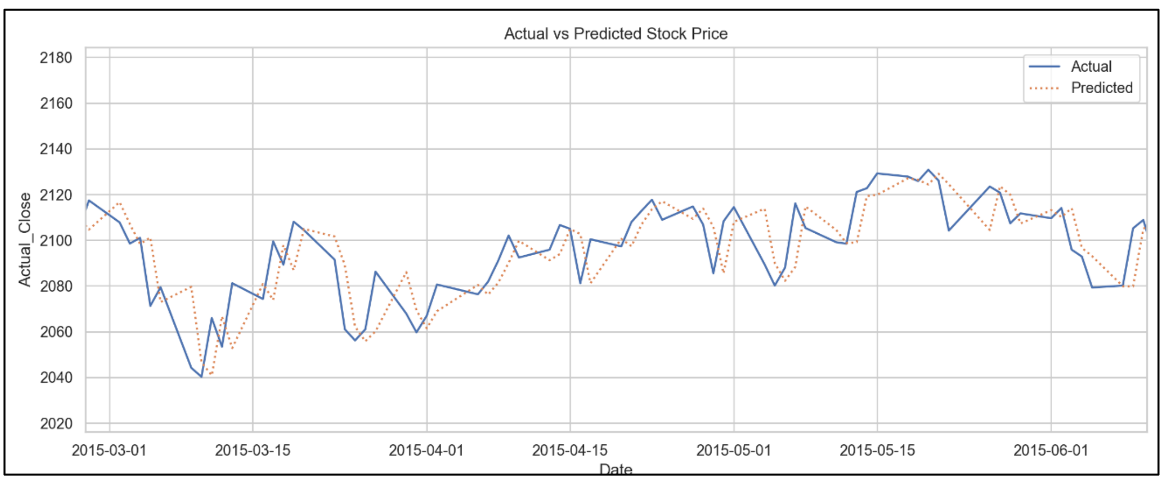
- R^2 Score shows the model fits extremely well on trend-driven long-term data.

5. VISUALIZATIONS AND ANALYSIS

All graphs provide insight into the model behaviour and prediction quality.

Actual vs Predicted Stock Price (Line Plot)





The predicted line follows the actual stock price very closely.

Only small differences appear during sharp market movements.

Actual Closing Prices Plot

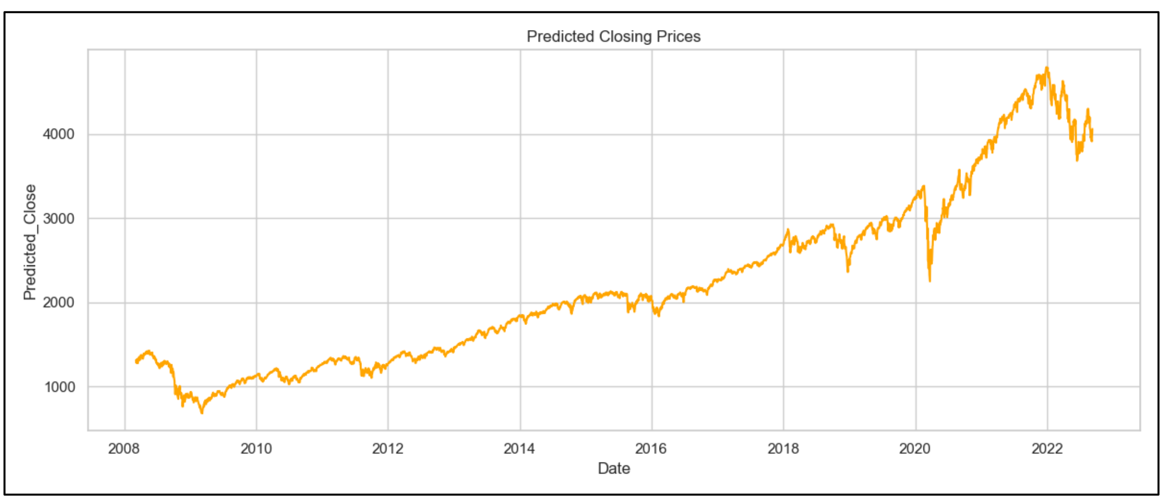
Shows complete long-term index movement.

Insights:

- Market recovered from 2008 crash.
- Strong bullish trend from 2012–2021.
- COVID-19 crash is clearly visible.
- Regression successfully fits this long-term structure.

Predicted Closing Prices Plot

Isolates the predicted values for clarity.
Predictions follow real market direction extremely closely.

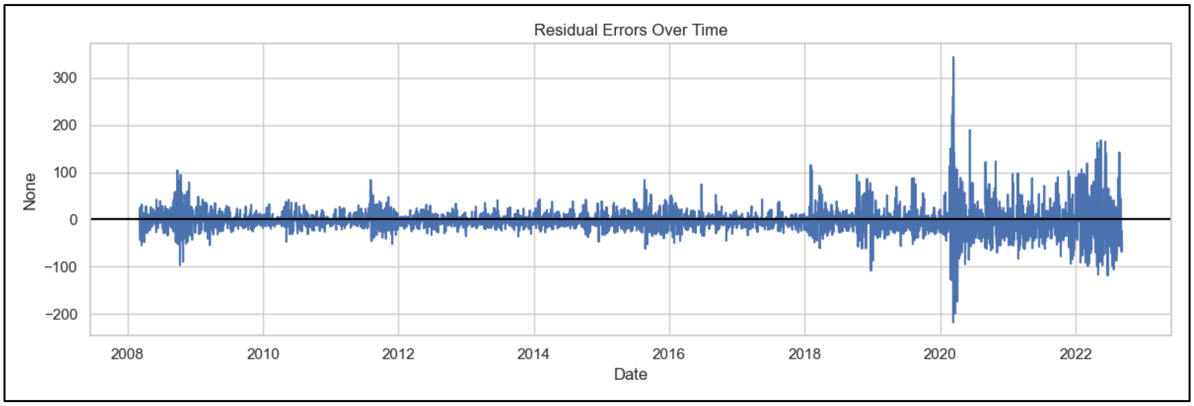


Residual Error Plot

Residual = Predicted − Actual

Interpretation:

- Errors mostly revolve around zero.
- Few spikes occur during highly volatile periods (e.g., 2020 crash).
- No major long-term drift → good model stability.

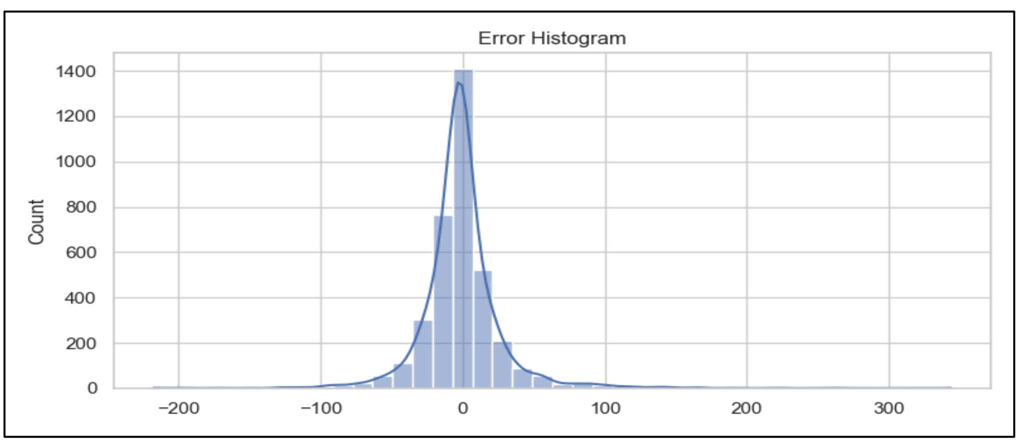


Error Histogram

Residual distribution is:

- Centered around zero
- Slightly spread during volatile years
- Nearly normal distribution

This indicates the model generalizes well.

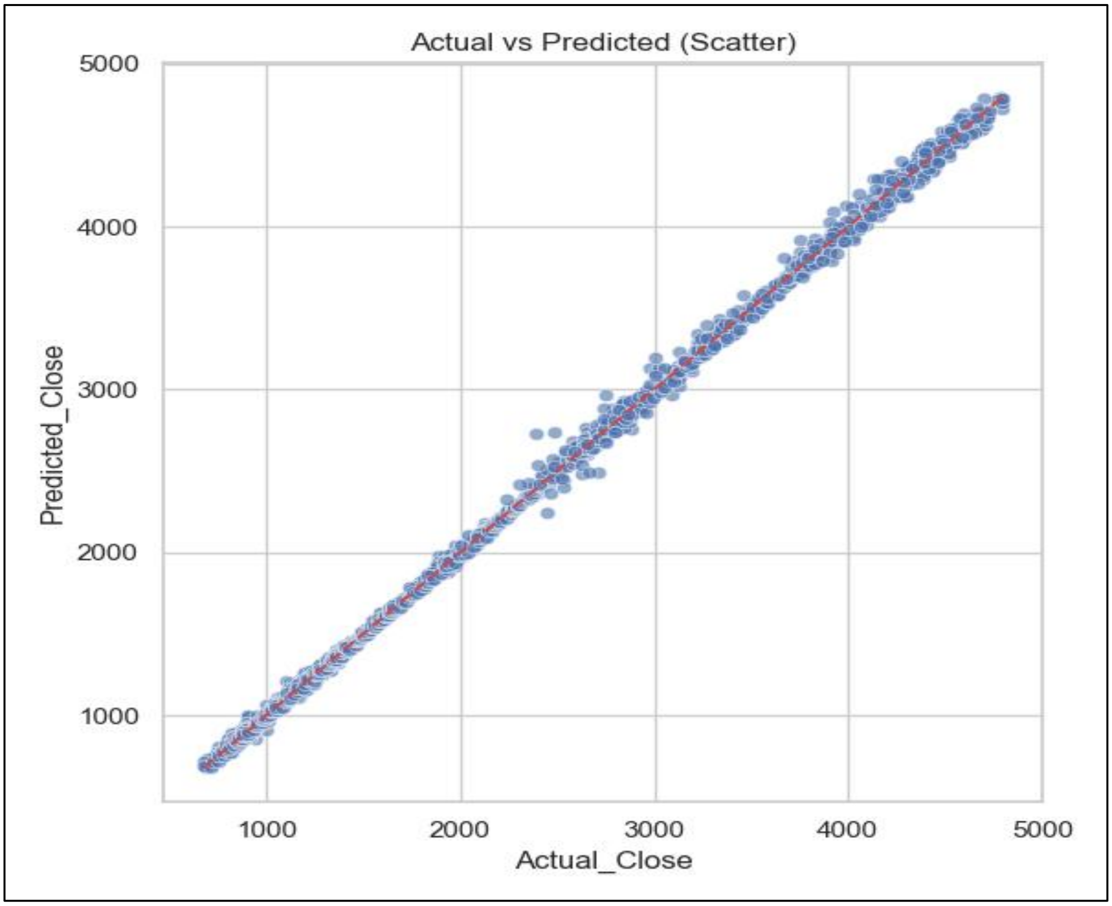


Actual vs Predicted Scatter Plot

Displays a nearly perfect straight diagonal line.

Interpretation:

- Predictions ≈ Actual values
- Very high accuracy
- Minimal outliers



Metrics:

Metric	Value
MAE	17.26148
RMSE	28.23818
R2 Score	0.999221

6. TECHNICAL INTERPRETATION

Model Strengths

- Captures long-term trend very well
- Strong predictive capability due to engineered features
- Uses simple and explainable ML
- Computationally fast
- Easily extendable to other stocks

Model Limitations

- Linear models cannot capture sudden market shocks
- Momentum-based features may lag during reversals
- Not designed for minute or hour-level trading
- Only predicts next-day closing price (short horizon)

Business Interpretation

The system can support:

- Financial analysis
- Trend forecasting
- Investor education
- Baseline modeling for algorithmic trading
- Market anomaly detection via residual spikes

7. CONCLUSION

This project successfully builds a complete Stock Market Prediction System using engineered technical indicators and a multiple linear regression model. The system:

- Performs robust cleaning and feature engineering
- Uses financial indicators widely used in technical analysis

- Predicts next-day closing prices with high accuracy
- Produces meaningful visual analysis
- Generates performance reports and predicted outputs