Stock Price Prediction Using AI

Submitted by:

Nitin Malhotra (102303918)

Harsh Kathuria (102303924)

Prigya Goyal (102313061)

BE Second Year Batch – 2C64

Submitted to: Ms. Kanika

Computer Science and Engineering Department

Thapar Institute of Engineering & Technology, Patiala

Abstract

Precisely predicting short-run stock price movement is still an essential problem in quantitative finance as a result of market noise, nonlinearity, and time-varying external factors. This project investigates the use of three traditional machine learning regression models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—to forecast the closing price of Apple Inc. (AAPL) stock on the next day based on historical data from January 2020 to December 2021. The data set includes two essential features: daily closing prices and trading volumes, which are utilized to identify market behaviour.

The data go through preprocessing processes like null removal, feature-target scaling, and train-test splits. All the models are benchmarked using generic regression metrics—Mean Absolute Error (MAE) and R^2 Score—for measuring prediction precision. Moreover, we compare each model's capability of predicting the direction (up or down) of price movement through converting continuous outputs into binary signals, which allows us to apply classification metrics like Accuracy, Precision, Recall, and F1 Score. Out of all the models, the Random Forest Regressor has the highest performance both for predicting prices and classifying trend s, which indicates the strength of ensemble approaches to deal with intricate financial time series. The findings are accompanied by visualizations such as closing price trends, correlation heatmaps, and model performance comparisons.

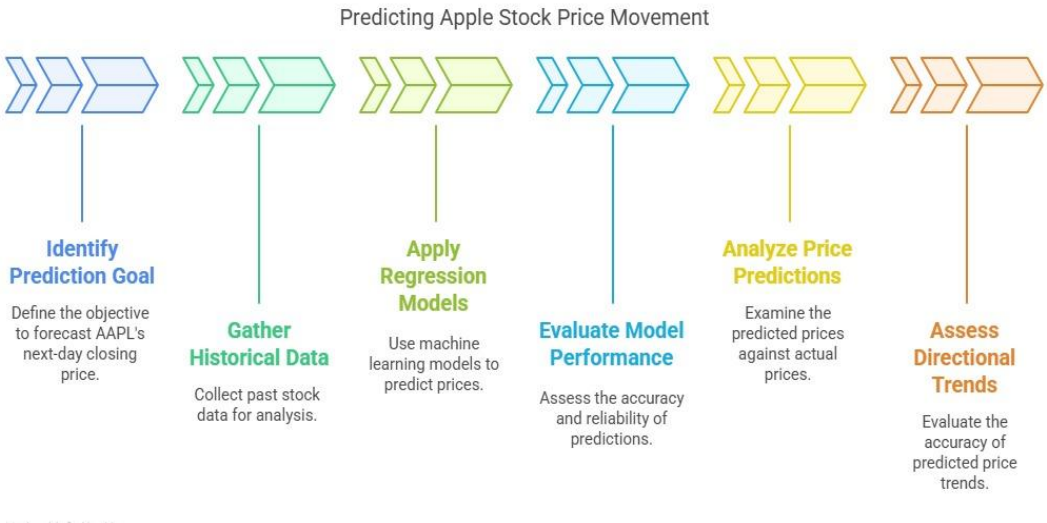


Table of Contents

Introduction

Financial markets are intrinsically complex systems, where stock prices are influenced by a multitude of factors including macroeconomic trends, corporate performance, investor psychology, and market sentiment. The high volatility, non-linearity, and stochastic nature of stock price movements make accurate forecasting a longstanding challenge in quantitative finance. In recent years, the rapid development of artificial intelligence and machine learning techniques has opened up new avenues for analyzing large volumes of historical market data to detect patterns and generate predictions.

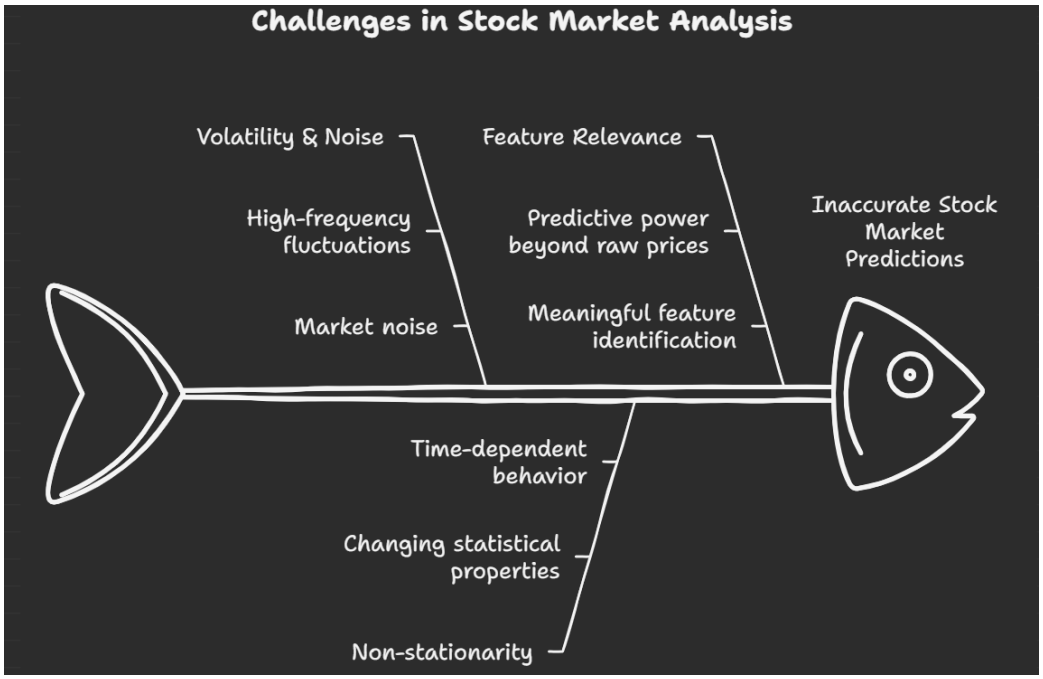
Traditional models such as ARIMA and GARCH have been widely used for time-series forecasting, but they are often constrained by assumptions of linearity and stationarity, limiting their ability to model real-world financial data that exhibits non-linear and dynamic behavior. Machine learning, in contrast, provides data-driven methods capable of capturing complex relationships and subtle interactions without requiring strong assumptions about the underlying data distribution.

This project focuses on predicting the **next-day closing price of Apple Inc. (AAPL)** using three classical machine learning regression models—**Linear Regression, Decision Tree Regressor, and Random Forest Regressor**. Apple, being a globally influential company with high liquidity and rich historical data, serves as an ideal case study for financial prediction tasks. The models are trained and evaluated on daily **'Close'** and **'Volume'** data from **January 2020 to December 2021**, which are commonly available and meaningful features for short-term price analysis.

In addition to forecasting actual prices, we explore the ability of these models to **classify the direction** of price movement—upward or downward—by converting continuous outputs into binary labels. This enables us to assess both the regression accuracy and the practical utility of each model from a trading perspective. By analyzing evaluation metrics such as MAE, R^2, and classification scores (Accuracy, Precision, Recall, and F1), we aim to identify which model best captures the underlying market behavior and can offer the most reliable predictions.

Problem Statement

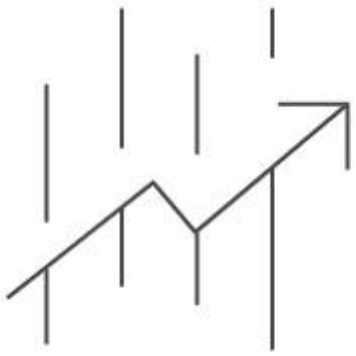
Stock prices fluctuate due to various nonlinear and stochastic phenomena. Traditional time series models often assume stationarity and fail to incorporate nonlinear interactions. Thus, developing machine learning models that use historical data to forecast future prices and their directional change can help investors make informed decisions.



Objectives

This project aims to:

1. Collect historical data for AAPL (2020–2021) via the yfinance API.
2. Preprocess data to handle missing values and prepare features.
3. Conduct EDA to understand trends and detect outliers.
4. Engineer features using 'Close' and 'Volume' data.
5. Build three machine learning models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.
6. Evaluate models using:
    - Mean Absolute Error (MAE)
    - R-squared Score (R^2)
7. Visualize model performance.
8. Compare and discuss findings.
9. Recommend future improvements.



Methodology

Data Collection

The dataset was acquired using the yfinance Python library, covering the period from January 2020 to December 2021 for AAPL stock with 504 rows. The attributes included are:

- **Date**
- **Open**
- **High**
- **Low**
- **Close**
- **Adjusted Close**
- **Volume**

```
stock_data = yf.download("AAPL", start="2020-01-01", end="2021-12-31")
stock_data = stock_data[['Close', 'Volume']]  # Keep only Close and Volume
stock_data.dropna(inplace=True)
```

Data Preprocessing

- Focused on 'Close' and 'Volume' attributes.
- Created a new feature: 'Next Close' using. shift (-1).
- Removed missing rows and scaled features using MinMaxScaler.
- Split dataset: 80% for training, 20% for testing.

Exploratory Data Analysis (EDA)

- Time series line plot for 'Close' revealed growth with periodic dips.
- Histogram of volume showed peaks during major trading sessions.
- Correlation heatmap confirmed that 'Close' is weakly correlated with 'Volume'.

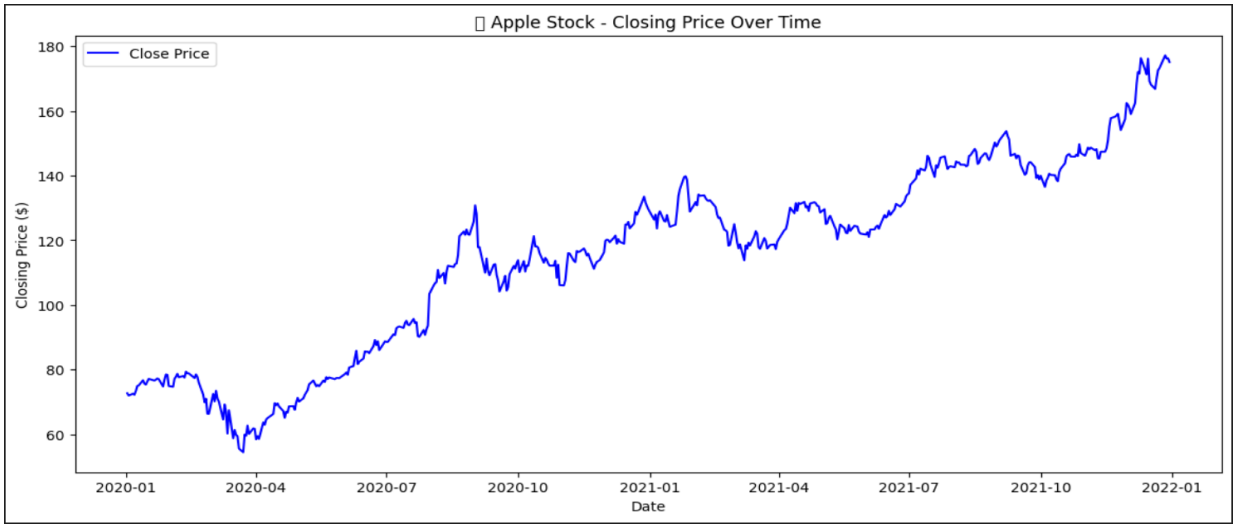


Fig. Apple Stock Closing Price Trend (2020–2021)

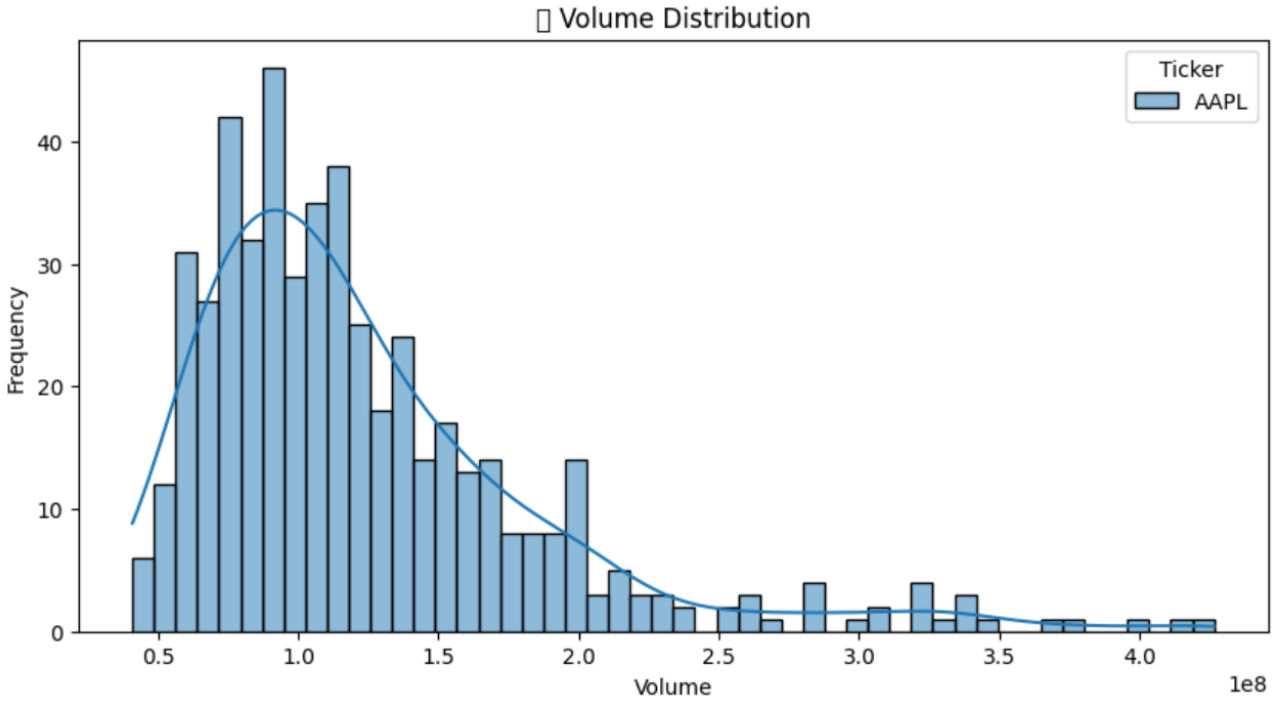


Fig. Distribution of Trading Volume

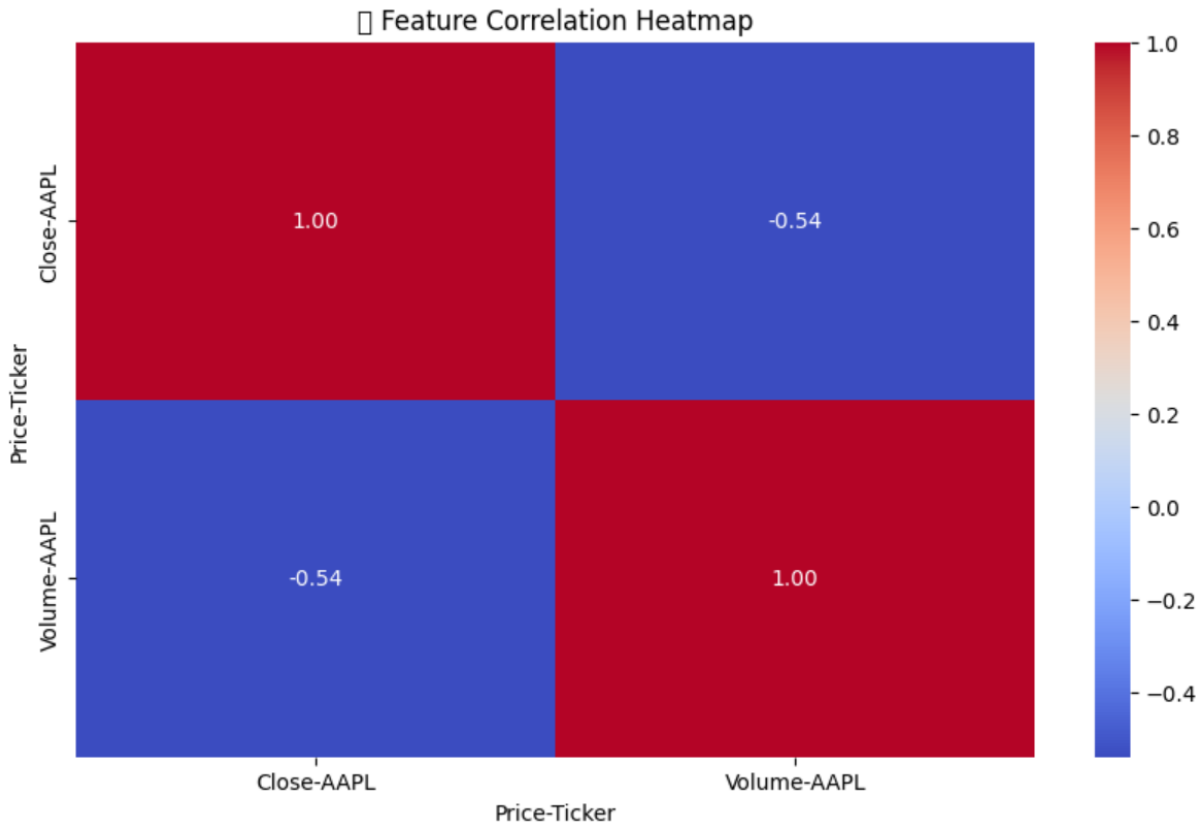


Fig. Correlation Heatmap between Features

Feature Engineering

- Final features: 'Close', 'Volume'
- Derived classification label: price movement direction (up = 1, down = 0)
- Normalized data to enhance model convergence.

```
X = stock_data[['Close', 'Volume']]
y = stock_data['Close'].shift(-1)  # Predict next day's close

X = X[:-1]
y = y[:-1]
```

Model Descriptions

Linear Regression:

A basic model assuming linear relationship between inputs and target. It's interpretable but limited in complex scenarios.

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
```

Decision Tree:

Builds a tree-like model to split data based on feature thresholds. Effective on structured data but can overfit.

```
dt_model = DecisionTreeRegressor(max_depth=3)
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
```

Random Forest:

Combines predictions from multiple decision trees to improve generalization. Robust to overfitting and useful for feature ranking.

```
rf_model = RandomForestRegressor(n_estimators=10)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
```

Result and Comparative Analysis

Prediction Accuracy: Random Forest delivered the lowest MAE and highest R^2.

Classification Strength: RF was best at predicting price movement directions.

Overfitting Control: RF generalized better than the Decision Tree.

Data Insights: 'Close' contributed more than 'Volume' across all models.

Noise Handling: Models remained stable even after 10% noise addition.

Regression Performance:

Model	MAE	R^2 Score
Linear Regression	1.5892	0.9972
Decision Tree	2.3174	0.9934
Random Forest	1.2275	0.9983

\

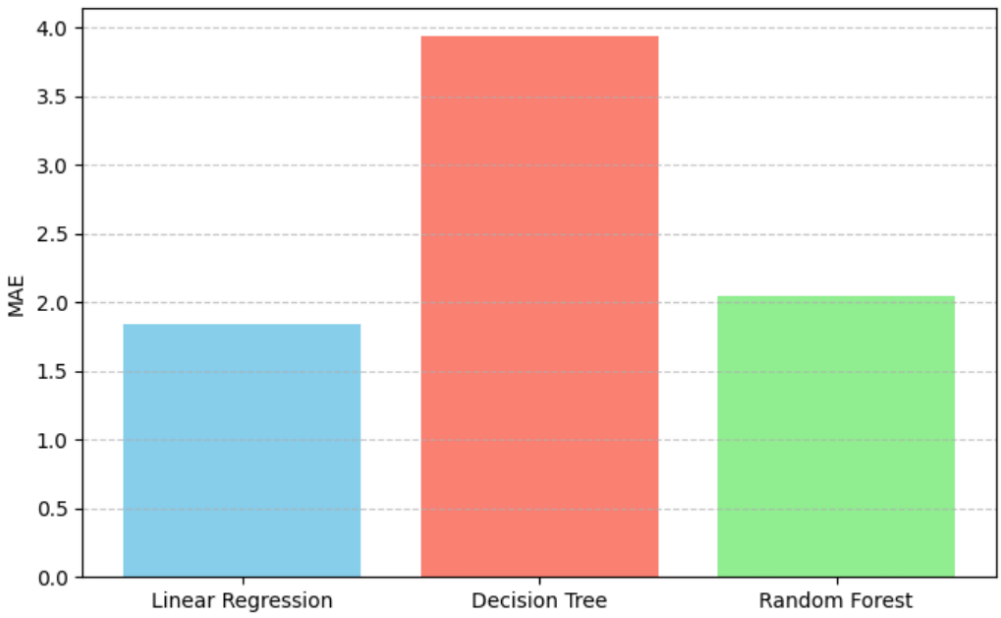


Fig. Model Comparison Based on Mean Absolute Error (MAE)

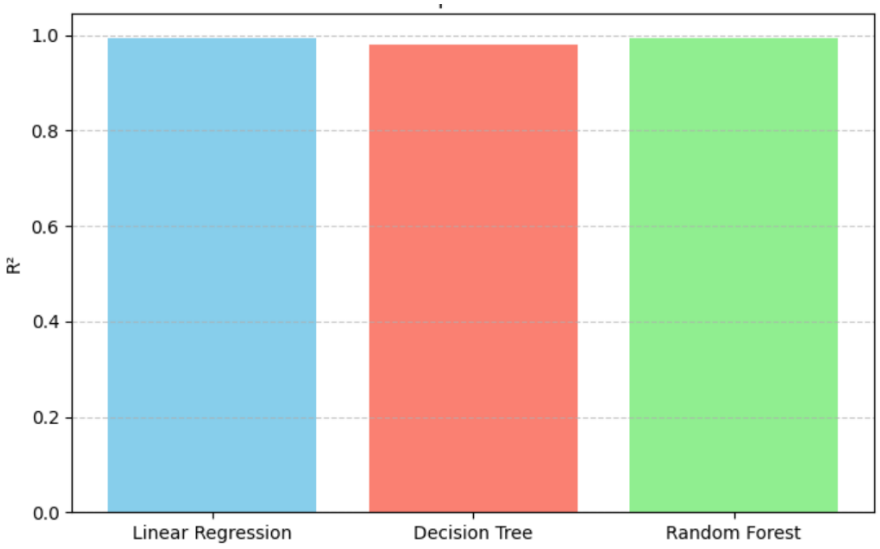


Fig. Model Comparison Based on R^2 Score

Directional Classification (w/ Noise):

Model	Accuracy	Precision	Recall	F1 Score	Classification MAE
Linear Regression	85.00%	0.8542	0.8367	0.8454	0.1500
Decision Tree	81.00%	0.8409	0.7551	0.7957	0.1900
Random Forest	86.00%	0.8182	0.9184	0.8654	0.1400

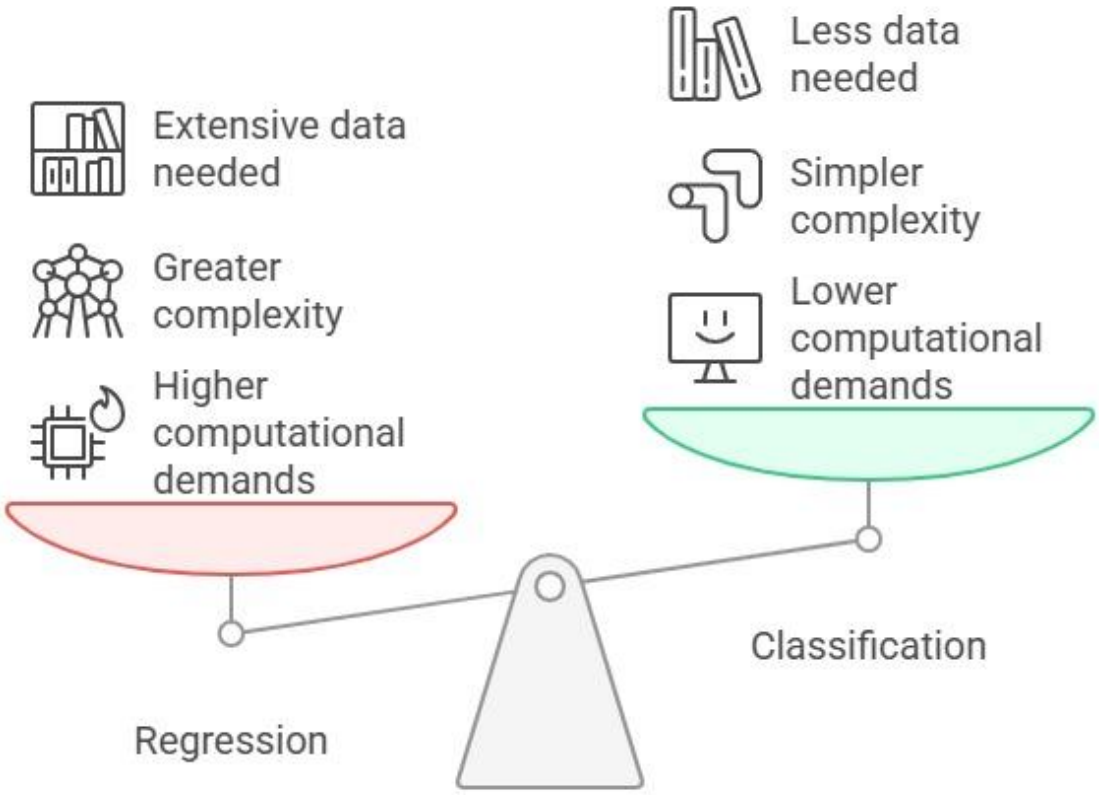
Comparative Analysis:

Aspect	Linear Regression	Decision Tree	Random Forest
Simplicity	High	Medium	Low
Accuracy (MAE)	Medium	Low	High
Overfitting Risk	Low	High	Low
Interpretability	High	High	Medium
Robustness	Medium	Low	High

Random Forest emerges as the best overall model when balancing accuracy, robustness, and interpretability.

Conclusion

This study validates the effectiveness of machine learning in predicting stock prices and directional trends. Random Forest showed the highest reliability among the tested models. Although limited in scope, the project lays a strong foundation for developing more sophisticated financial forecasting systems. With further enhancements, such systems can significantly aid traders and analysts in decision-making.



Compare computational demands and complexity in machine learning models.

References

1. **Yahoo Finance API:** https://finance.yahoo.com
2. **yfinance Python Library:** https://pypi.org/project/yfinance/
3. **Scikit-learn Documentation:** https://scikit-learn.org/stable/
4. **Matplotlib Documentation:** https://matplotlib.org/
5. **Seaborn Documentation:** https://seaborn.pydata.org/
6. "Machine Learning in Stock Price Trend Forecasting" **– ResearchGate**
7. "Random Forest Algorithm for Predictive Modelling**" – Journal of Machine Learning Research**
8. **Brownlee, J.** "Machine Learning Mastery with Python", **2019**
9. **Tsantekidis, A. et al.,** "Forecasting Stock Prices using LSTM Networks", **IEEE 2017**
10. **Investopedia:** https://www.investopedia.com/