**Predicting Stock Market Trends Using Machine Learning**

Preethi Bandari, Vamsi Krishna Alaparthi, Raja Thadikamalla

(bandarip4@mail.sacredheart.edu)(alaparthiv@mail.sacredheart.edu)(thadikamallar@mail.sacredheart.edu)

Sacred Heart University, 5151 Park Avenue, Fairfield, CT, USA

# Abstract

Predicting stock market trends using machine learning aims to create a predictive model capable of forecasting stock price movements, such as increases, decreases, or stability. This project utilizes historical stock data, trading volumes, and market sentiment indicators, along with economic factors, to train machine learning algorithms to provide actionable predictions.

**Index Terms** :

Stock Market, Trend Prediction, Classification, Machine Learning,Stock Data,Trading Volumes

# 1. Introduction

Stock prediction has always been a challenging problem for statistics experts and finance. The main reason behind this prediction is buying stocks that are likely to increase in price and then selling stocks that are probably to fall. Generally, there are two ways for stock market prediction. Fundamental analysis is one of them and relies on a company’s technique and fundamental information like market position, expenses, and annual growth rates. The second one is the technical analysis method, which concentrates on previous stock prices and values. This analysis uses the data from the YFinance. Stock markets were normally predicted by financial experts in the past time. However, data scientists have started solving prediction problems with the progress of learning techniques. Also, computer scientists have begun using machine learning methods to improve the performance of prediction models and enhance the accuracy of predictions. Employing deep learning was the next phase in improving prediction models with better performance. Stock market prediction is full of challenges, and data scientists usually confront some problems when they try to develop a predictive model. Complexity and nonlinearity are two main challenges caused by the instability of the stock market and the correlation between investment psychology and market behavior.

There are always unpredictable factors such as the public image of companies or the political situation of countries, which affect stock market trends. Therefore, if the data gained from stock values are efficiently preprocessed and suitable algorithms are employed, the trend of stock values and index can be predicted. In stock market prediction systems, machine learning and deep learning approaches can help investors and traders through their decisions. These methods intend to automatically recognize and learn patterns among large amounts of information. The algorithms can be effectively self-learning and can tackle the predicting task of price fluctuations to improve trading strategies.

In this work supervised learning, models are trained using labeled data, where both the input variables (features) and the output variable (target) are known. Linear regression specifically deals with regression tasks, where the goal is to predict continuous values, such as predicting house prices or stock values based on input data.

Overall, regarding the above literature, prior studies often concentrated on macroeconomic or technical features with recent machine learning methods to detect stock index or valued movement without considering appropriate preprocessing methods.

# 2. Research Data

In this study, the data of last year is being taken from the YFinance where live data is getting fetched directly from this and it’s been uploaded into the database of the backend. YFinance module is a Python Library that allows users to get the historical market data from Yahoo Finance. It Provides easy access to financial data such as stock prices, historical data, dividends, and market statistics making it useful for financial analysis, stock market research, and algorithmic trading. The main features of YFinance include acquiring historical stock prices, such as open, high, low, close, adjusted close, and volume. Fetching real-time data and live market updates. Retrieving company financials, earnings data, balance sheets, and cash flow statements. Getting dividends and stock splits information.

This script fetches historical stock data from Yahoo Finance for a specified stock symbol and inserts the data into a PostgreSQL database. The data includes information such as

• Opening price - The opening price refers to the price at which a stock, commodity, or other financial instrument starts trading when the market opens for the day. It is the first price that is matched between buyers and sellers after the market opens.

• Closing price - The closing price in stocks refers to the last price at which a stock, commodity, or other financial instrument is traded during a regular trading session before the market closes for the day. It represents the final agreed-upon price between buyers and sellers at the close of the market.

• High - In the context of stocks, the high refers to the highest price at which a stock or other financial instrument was traded during a specific period. This could be during a trading day (daily high), week, month, or even longer timeframes.

• Low - In the context of stocks, the low refers to the lowest price at which a stock or other financial instrument was traded during a specific period. Like the high, this can be measured over different timeframes such as the Daily Low – The lowest price of the stock during that trading day

In this study, the data of last year is being taken from the YFinance where live data is getting fetched directly from this and it’s been uploaded into the database of the backend. YFinance module is a Python Library that allows users to get the historical market data from Yahoo Finance. It Provides easy access to financial data such as stock prices, historical data, dividends, and market statistics making it useful for financial analysis, stock market research, and algorithmic trading. The main features of YFinance include acquiring historical stock prices, such as open, high, low, close, adjusted close, and volume. Fetching real-time data and live market updates. Retrieving company financials, earnings data, balance sheets, and cash flow statements. Getting dividends and stock splits information.

This script fetches historical stock data from Yahoo Finance for a specified stock symbol and inserts the data into a PostgreSQL database. The data includes information such as

• Opening price - The opening price refers to the price at which a stock, commodity, or other financial instrument starts trading when the market opens for the day. It is the first price that is matched between buyers and sellers after the market opens.

• Closing price - The closing price in stocks refers to the last price at which a stock, commodity, or other financial instrument is traded during a regular trading session before the market closes for the day. It represents the final agreed-upon price between buyers and sellers at the close of the market.

• High - In the context of stocks, the high refers to the highest price at which a stock or other financial instrument was traded during a specific period. This could be during a trading day (daily high), week, month, or even longer timeframes.

• Low - In the context of stocks, the low refers to the lowest price at which a stock or other financial instrument was traded during a specific period. Like the high, this can be measured over different timeframes such as the Daily Low – The lowest price of the stock during that trading day

**2.1 Volume Requirement**

- In the stock market, volume refers to the total number of shares of a stock or contracts of security traded during a specific period, typically within a trading day.

Volume requirement refers to the minimum level of trading activity (or volume) that must be met in certain scenarios, such as:

1. Liquidity Requirements: Some stocks or securities must maintain a certain trading volume to ensure sufficient liquidity. Higher volume means more active trading, which makes it easier for investors to buy or sell without causing significant price changes.

2. Margin or Trading Strategies: Certain trading strategies or margin accounts may have volume requirements. For example, day traders often look for stocks with high volume, as it allows them to enter and exit positions quickly.

3. Exchange Listings: Some stock exchanges have minimum volume requirements for listed companies to ensure there is enough investor interest and activity in the stock. If a stock falls below this requirement for a prolonged period, it could be at risk of being delisted from the exchange.

4. Options and Derivatives: For options and other derivatives, there may be volume requirements to open certain trades, as low volume can lead to illiquid markets, making it difficult to execute orders at favorable prices.

**2.2 Preliminary Investigation Phase**

**2.2.1 Summary of Problems, Opportunities, and/or Directives**

Predicting stock market trends using machine learning involves developing a sophisticated model that can accurately forecast whether the price of a stock will rise, fall, or remain stable in the future. This complex task is fraught with potential challenges, including several data-related issues such as incomplete, inaccurate, or non-representative data. Identifying relevant features is another critical hurdle, as the model must discern which variables most significantly impact stock prices. The risk of overfitting the model to historical data, where it performs well on past data but poorly on new data, is also a significant concern. Furthermore, the presence of noise in stock market data, characterized by random fluctuations and irrelevant information, can obscure true patterns and complicate the prediction process. The stock market's dynamic and ever-changing nature adds a layer of difficulty, as the model must continuously adapt to new information and shifting trends.

**2.2.2 Statement of Preliminary Scope**

The Scope of Predicting Stock Market trends using machine learning is to create a reliable tool that can assist investors in making informed decisions by providing insights into potential stock movements. By doing this the model aims to reduce uncertainty and risk associated with stock market investments thereby enhancing the overall investment strategy and potentially increasing results.

**2.2.3 What type of data describes the system being studied?**

•Predicting stock market trends using machine learning examines the following data:

•Data is being taken from the API Y Finance ( a Python library that provides a simple interface to retrieve stock market data from Yahoo Finance.

•The Stock for the various companies such as Apple, Google, Microsoft, Tesla, Amazon

•PostgreSQL is configured to store the live and historical stock data

# 3. Problem Analysis Phase

**3.1Study the problem domain**

The potential problems that can arise while predicting the stock market trends using machine learning can be data-related issues, Identifying Relevant Features, Overfitting, and Noise – as the stock market data can contain a lot of noise makes it difficult to identify true patterns that can predict future trends.

**3.2 Data-Related Issues:**

•Incomplete or Inconsistent Data: Financial datasets are often vast but can contain missing or inconsistent entries due to various reasons such as market holidays, trading suspensions, or data errors. This can disrupt training and lead to inaccurate predictions if not handled properly.

•High Dimensionality: Stock market data typically involves multiple variables (e.g., opening price, closing price, volume, economic indicators). Managing and processing high-dimensional data can strain computational resources and lead to slower, less efficient models.

**3.3 Data Collection by the System**

Our application leverages the Yfinance library to gather comprehensive stock market data from Yahoo Finance, which is then systematically stored in a structured database. This setup includes tables for storing historical stock prices, key performance indicators (KPIs), and user information. Each stock’s daily data, such as opening and closing prices, high and low values, and trading volume, is captured and organized in a stock\_prices table, indexed by the stock symbol and date. In addition, the stock\_kpi table maintains essential metrics, including opening balance, annual income, total balance, and year-end summaries. By storing this data in a database, we enable efficient retrieval, manipulation, and analysis, allowing users to query and visualize historical trends, analyze stock performance, and make informed financial decisions. This setup is scalable and facilitates further processing, such as calculating technical indicators, generating automated reports, and running predictive models, enhancing our ability to provide real-time insights and in-depth financial analysis.

# 3.4 Project Setup

# • Framework: The project is built using Flask, a lightweight WSGI web application framework in Python.

# • Database: PostgreSQL is used for storing stock data and user authentication details. SQL Alchemy is utilized as the ORM (Object Relational Mapper) for handling database operations.

# 4. Model

Model Selection:

Linear Regression: The Linear Regression model is used to predict the closing price of a stock based on features like opening price, high, low, and volume.

Data Handling and Preprocessing:

Data is fetched from a PostgreSQL database (using SQL Alchemy to create a connection engine) and is processed using pandas for further preparation.

The data is split into training and testing sets using train\_test\_split from sklearn.model\_selection.

Performance Metrics:

The model's performance is evaluated with Mean Squared Error (MSE) and R-squared (R²), using mean\_squared\_error and r2\_score.

Saving the Model:

The model is saved as a .joblib file in the "models" directory using joblib, allowing it to be reused later without retraining.

The Equation :

Y = w1x1+w2x2+w3x3+w4x4+b

Terms :

Y – is the target variable which is the closing price of the stock

X1,x2,x3,x4 – the input features ( independent variables ) :

X1- opening price , x2 – high price , x3- low price , x4- volume

W1,w2,w3,w4 – The weights or coefficients assigned to each input feature. learned by the model during training

b- the intercept term, which is also learned during the training and adjusts the output value of y

# 5. Model Training



**Figure1.** X: Scatter plot of actual vs. predicted stock close prices using Linear Regression. The red line y=x indicates the ideal line where predicted values would equal actual values. The model achieved a high R2R score of 0.96, suggesting strong predictive accuracy.

**How This Graph Supports Model Reliability:**

This plot visually confirms that the Linear Regression model performs well in capturing the relationships within the stock data. This graph provides evidence of the model's reliability and helps to validate that Linear Regression is a suitable choice for predicting stock prices based on the given features.

**6. Model’s Comparison Table :**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Data Used** | **Accuracy Metric (e.g., MSE/R²)** | **Computation Time** | **Strengths** | **Limitations** | **Suitability for Stock Prediction** |
| Linear Regression | stock prices, volume | MSE: 20.56, R²: 0.96 | Low | Interpretable, efficient, handles linear relationships | Struggles with non-linear data | High |
| Decision Tree | Stock prices, market sentiment, trading volumes | Moderate accuracy | Moderate | Captures non-linear relationships | Prone to overfitting without pruning | Moderate |
| Random Forest | Stock prices, economic indicators | Higher accuracy than Linear Regression | Moderate | Reduces overfitting, handles complex relationships | Requires more computation, less interpretable | High |

# 7. Why Choose Linear Regression

Simplicity and Interpretability - Its output includes coefficients that directly show the relationship between each feature (e.g., opening price, high, low, volume) and the target variable (closing price). See Figure 1.1 and Table 1.1(Model Comparison Table )

This makes it easy for researchers and investors to understand which factors are influencing stock price predictions and by how much., Efficiency in Computation - Linear Regression is computationally efficient and fast to train, especially with relatively straightforward features (open, high, low, volume). This makes it well-suited for initial analysis or if computational resources are limited. Complex models like Random Forest, SVM, or LSTM can be resource-intensive and require more tuning. Adequate Performance for Trend Prediction - If the goal is to capture general trends or patterns in stock prices without needing complex patterns, Linear Regression often provides satisfactory performance, especially in cases where there are strong linear relationships within the dataset. Your model’s high R² score of 0.96 suggests that Linear Regression is capturing the main patterns well in this case, Feature Simplicity and Linearity - Since the model uses features such as open, high, low, and volume to predict close prices, Linear Regression works well if these relationships are close to linear. For example, if closing prices correlate linearly with opening prices and other inputs, Linear Regression can perform quite effectively without the need for more complex models. Focus on Reliability and Predictive Stability

# 8. Overall Summary

The research project on predicting stock market trends using machine learning (ML) aims to develop a predictive model that forecasts stock price movements based on various data sources. By leveraging historical stock data, trading volumes, market sentiment, and economic indicators, the study employs various ML algorithms, including linear regression and decision trees, to identify patterns and enhance predictive accuracy. The project emphasizes the importance of data preprocessing and feature engineering, enabling the models to recognize complex relationships within the data. Ultimately, this research seeks to assist investors by providing insights that may help reduce risks and improve returns in stock investments. The approach signifies a paradigm shift in investment strategies, utilizing artificial intelligence and big data to analyze market behavior and enhance decision-making processes. The study also highlights the strengths and limitations of different models, concluding that linear regression offers simplicity, interpretability, and adequate performance for initial analyses.

**References :**

1. Machine Learning for Asset Managers" by Marcos López de Prado

2. Hands-On Machine Learning for Algorithmic Trading" by Sourav Ghosh

3. Financial Machine Learning: A Guide to Machine Learning in Finance -Tony Guida

4. Python for Finance: Mastering Data-Driven Finance" by Yves Hilpisch

5.M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques", *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, pp. 322-327, 2016.