Stock Market Price Prediction Using Machine Learning

By

Abrar Karim

ID: 202-35-3129

Supervised by

Ms. Nusrat Jahan

Assistant Professor

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of B.Sc. in Software Engineering

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

SPRING-2024

APPROVAL

This Thesis titled "**Stock Market Price Prediction Using Machine Learning**", submitted by **Abrar Karim**, ID: **202-35-3129** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

10/7/24
-- **Chairman**

Dr. Engr. AKM Masum
Professor
Department of Software Engineering
Faculty of Science and Information
Technology Daffodil International University

10/7/24
-- **Internal Examiner 1**

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **Internal Examiner 2**

Md. Rajib Mia
Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

-- **External Examiner**

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering
Jagannath University

THESIS DECLARATION

I thus certify that, as a partial achievement of my work, I completed my thesis, "Stock Market Price Prediction using Machine Learning," under the guidance of Ms. Nusrat Jahan, Assistant Professor, Department of Software Engineering, Daffodil International University. Furthermore, I declare that this project has not been presented elsewhere as a partial fulfillment for a degree from this university or any other. Furthermore, I hereby declare that neither this thesis nor any portion of it has been submitted for consideration for a bachelor's degree or else elsewhere.

Abrar

Submitted by

Abrar Karim

ID: 202-35-3129

Department of Software Engineering

Daffodil International University

Supervised by,

Ms. Nusrat Jahan

Assistant Professor

Department of Software Engineering

Daffodil International University

ACKNOWLEDGEMENT

I owe a debt of appreciation to **Ms. Nusrat Jahan**, Assistant Professor in the Department of Software Engineering at Daffodil International University. Her expertise and passion for the subject of "**Machine Learning**" provided invaluable guidance throughout this research journey. Without her unwavering patience, intellectual guidance, continuous encouragement, rigorous supervision, constructive feedback, and meticulous review of multiple drafts, this achievement would not have been possible.

I extend my heartfelt regards to **Dr. Imran Mahmud**, Head of the Software Engineering Department at DIU, as well as to other professors and staff members who supported me in completing my research. Their assistance was instrumental in successfully defending my thesis and graduating with honors.

Lastly, I am deeply grateful to my parents for their enduring patience and unwavering support throughout my life.

Stock Market Price Prediction using Machine Learning

ABSTRACT

Stock market price prediction is a vital component for investors and financial analysts, offering the dual benefits of maximizing profits and managing risks effectively. This study delves into the application of sophisticated machine learning algorithms to forecast stock prices. It leverages a comprehensive dataset that spans several years and encompasses key financial metrics, including opening, closing prices, high, and low prices, as well as trading volumes. In this research, five regression models were employed to forecast future stock prices: DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor. Training and testing samples were taken from the dataset in order to assess the performance of these models. The analysis shows that ensemble methods, specifically GradientBoostingRegressor and RandomForestRegressor, deliver superior forecast accuracy than the other models. The study meticulously details the actions required in data preprocessing, the execution of the models, and a thorough performance evaluation. It sheds light on the strengths and weaknesses of each algorithm throughout the stock market forecasting. The results indicate that machine learning provides a solid foundation for creating predictive models that can accommodate the complex and dynamic nature of financial markets. This study highlights the opportunities for machine learning to enhance predictive analytics in finance, providing valuable insights for developing robust investment strategies.

Keywords: Machine Learning, Stock Market Prediction, Financial Forecasting, Regression Models, Ensemble Methods, Predictive Analytics, Data Science, Algorithm Evaluation, Investment Strategies.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction

The stock market is a vital gauge of the state of the economy, reflecting the collective sentiments of investors and their expectations for the future. As a result, the ability to accurately predict stock prices is of paramount importance for investors seeking to maximize their returns, financial analysts aiming to provide sound investment advice, and policymakers responsible for steering the economy. Successful stock market predictions can lead to substantial financial gains, the development of improved investment strategies, and the formulation of more informed economic policies. However, because the stock market is so complex and dynamic, it is intrinsically difficult to anticipate stock prices. The price of stocks is affected by a wide array a number of elements, including financial metrics such as GDP growth, rates of interest, inflation, in addition to company-specific information, industry trends, and even global events like political elections and natural disasters. Moreover, the market's susceptibility to sudden shifts in investor sentiment and unforeseen events adds another layer of complexity to the prediction process.

The stock market's dynamic nature and its significance in the global economy have spurred significant interest in accurately predicting stock prices. Reliable forecasts can offer substantial benefits to investors, traders, and financial analysts by informing their decisions and strategies. Machine learning (ML) has become a potent instrument for handling complex data and generating predictions in various fields, including finance. The purpose of this study is to investigate how well machine learning methods forecast stock market values, focusing on the analysis and model development outlined in the provided notebook.

1.2 Background

Predicting future stock prices using historical data, present market circumstances, and other factors is known as stock market price prediction also with various economic indicators. The ability to predict stock prices accurately is of immense value in finance, offering potential gains through informed investment strategies and risk management.

Though useful, traditional stock price forecasting techniques frequently come to reduced because they cannot fully capture the intricate and inconsistent correlations found in financial information. This limitation has prompted the exploration of advanced techniques, particularly machine learning algorithms. A branch of artificial intelligence called machine learning is very good at finding trends in big datasets and anticipating them. These algorithms are ideal for the dynamic nature of financial markets because they can handle large volumes of data, identify subtle patterns, and continually learn and adapt. We explore the use of several machine learning algorithms for stock market price prediction in this paper. Our dataset provides a comprehensive perspective of market movements and stock performance, encompassing many years of daily stock data and major financial indicators like starting price, highest price, lowest price, closing price, and trading volume.

1.3 Research Motivation

Comparing the effectiveness of several machine learning algorithms in stock price prediction is the main goal of this study. This paper attempts to give a detailed evaluation of the suitability of these algorithms for financial forecasting by carefully examining their prediction accuracy, advantages, and disadvantages. The results of this study will add to the growing body of information about machine learning's application in finance and offer valuable insights for investors and analysts seeking to leverage these technologies for improved decision-making. The significance of this study extends beyond its potential to enhance stock price prediction, as it also has broader implications for the field of financial analytics. With the continuous evolution of machine learning, its applications in finance will be estimated to grow, resulting in the creation of increasingly complex tools for managing financial risks and capitalizing on opportunities. This research represents a step towards realizing this potential, showcasing how advanced algorithms can be utilized to improve our understanding and forecasting of market behavior.

1.4 Problem Statement

Accurate stock market price prediction remains a complex and critical challenge due to the inherent volatility and multifaceted influences on stock prices, including economic indicators, market sentiment, and unforeseen events. Technical analysis and statistical models, two common traditional forecasting techniques, frequently fall short of capturing the complex patterns alongside the dynamic nature of the market. Although machine learning (ML) offers having the ability to enhance prediction accuracy by leveraging large datasets and advanced algorithms, current ML approaches face significant limitations. These include overfitting to historical data, inadequate integration of diverse data sources, insufficient feature engineering, lack of model interpretability, and poor adaptability to real-world trading environments.

The goal of this study is to cope with these limitations by developing robust models for machine learning that effectively predict stock prices, incorporating advanced feature engineering and diverse primary sources of data including mood analysis and indicators of macroeconomic health. The objective is to enhance model performance while ensuring adaptability to evolving market conditions, and integrating risk management strategies to handle market anomalies. The study serves as a foundation for exploring these challenges, showcasing machine learning's capability in financial prediction and highlighting areas for improvement to create more accurate, practical, and resilient predictive systems for stock market forecasting.

1.5 Research Gap

Despite the advancements in machine learning to forecast stock market prices, several research gaps remain. Current models often struggle with the non-stationarity and volatility inherent in stock prices, as traditional classifiers like neural networks, decision trees, and support vector machines may fail to adapt dynamically to sudden market shifts or rare events. Additionally, many studies focus on short-term predictions without adequately addressing the long-term forecasting challenges or the impact of outside elements like geopolitical developments and economic strategies. The integration of multi-source data, such as news sentiment and macroeconomic indicators, into machine learning classifiers is underexplored, limiting the models' ability to fully capture the complexity of market dynamics. Furthermore, interpretability and transparency of these models are often overlooked, which makes it challenging to comprehend how they make decisions and build trust among financial professionals. Addressing these gaps requires a holistic approach that combines advanced algorithms, comprehensive feature sets, and a focus on model explainability.

Strategies to Address Research Gaps:

- Online Learning: Implement online learning techniques that revise the model incrementally whenever fresh information becomes accessible, allowing it to adapt to real-time changes in the market.
- Reinforcement Learning: Explore reinforcement learning approaches where the model learns to make trading decisions based on continuous feedback from the market, optimizing for long-term returns.
- Adaptive Algorithms: Use algorithms that adjust their parameters dynamically in response to changing market conditions, ensuring the model remains relevant and accurate. Use early stopping

- regularization strategies to reduce overfitting in machine learning models. For better model generalization performance, take into account methods like hyperparameter adjustment and cross-validation.

  .

1.6 Research Scope

The research scope for machine learning for predicting stock market prices involves a comprehensive approach to developing and refining predictive models through various stages of data processing, feature engineering, model development, evaluation, and practical application. By concentrating on these topics, the study aims to create robust, accurate, and interpretable models that can effectively forecast stock prices and be applied in real-world trading environments. This scope ensures that the research not only addresses current challenges but also opens the door for more developments in the sector of financial prediction via machine learning. The study inquires into traditional models for machine learning and deep learning techniquesto identify intricate patterns as temporal dependencies in stock prices, with evaluation metrics including MAE, MSE, and financial metrics like Sharpe ratio to ensure model efficacy. Emphasizing interpretability through methods like SHAP and LIME can be done, the research also addresses the deployment of models in live trading systems, adaptation to market changes, and integration of risk management strategies to handle market anomalies and extreme events, aiming to develop accurate, practical, and adaptive predictive systems for real-world financial applications.

1.7 Research Question

1. How can machine learning models be effectively developed and optimized to accurately predict stock market prices, incorporating diverse data sources and advanced feature engineering?

2. How can the performance of models such as logistic regression, decision trees, naïve bias, and random forest be improved by lowering the values of R-squared, Mean Absolute Error (MAE) and Mean Squared Error (MSE)?

1.8 Research Objective

Develop and Evaluate Predictive Models: The primary objective is to constructmachine learning algorithms that can forecasting stock prices in a really accurate manner. This involves leveraging historical stock data, which includes a wide range of features include trade volume, closing and opening prices, and technical indications like moving averages and momentum oscillators.The research will explore various machine learning algorithms, including both traditional statistical models and advanced techniques like neural networks and ensemble methods, to ascertain the best strategy for stock pricing prediction. In order to compare and evaluate the efficacy of the constructed models, a thorough analysis will be carried out utilizing suitable performance measures, including mean squared error, mean absolute, R-squared error, and mean squared error.

Enhance Predictive Accuracy: The research aims to increase the precision of the predictive models by employing advanced techniques. Hyperparameter tuning involves optimizing the internal parameters of machine learning algorithms to fine-tune their performance and achieve better results. Feature selection techniques will be utilized to determine the most pertinent characteristics that significantly participate in stock price prediction, reducing noise and improving model efficiency. Model ensembles, which combine multiple models, will be explored to leverage the strengths of different algorithms and potentially achieve higher accuracy than individual models.

1.9 Summary

One important measure of the state of the economy, reflecting investor sentiment and future expectations is the stock market. Accurately predicting stock prices is vital for investors, financial analysts, and policymakers. Successful predictions can lead to financial gains, better investment strategies, and informed economic policies. However, it's difficult to anticipate stock values because of the market's dynamic nature. A number of variables, such as business news, industry trends, economic indicators, and world events, can affect stock prices. The market's susceptibility to sudden shifts in sentiment and unforeseen events further complicates the prediction process.

Chapter 2

Literature Review

2.1 Introduction

Stock market price prediction using machine learning classifiers has attracted a lot of interest in academic and professional circles, reflecting advancements in computational finance and data analytics. The literature reveals a diverse array of approaches, where traditional statistical models are often outperformed by sophisticated machine learning techniques. Classifiers like as the capacity of algorithms like Random Forests (RF), Neural Networks (NN), and Support Vector Machines (SVM) to capture nonlinear correlations while participating in complex patterns in stock data. Studies have demonstrated that these models, when trained on features derived from historical prices, technical indicators, and macroeconomic variables, can provide superior predictive accuracy compared to conventional methods.

2.2 Literature Review

The forecasting of stock market values has historically been a difficult and important field of study because of the market's inherent complexity and volatility. Traditional approaches to forecasting stock prices primarily relied on statistical models and econometric techniques. Many techniques have been employed, including GARCH (Generalized Autoregressive Conditional Heteroskedasticity) and AUTOREMISA (Autoregressive Integrated Moving Average). ARIMA models, developed by Box and Jenkins in 1976, aim to predict future points in a chronological series built off its own historical data. While effective for short-term predictions, ARIMA models often struggle with the non-linearities present in financial data. GARCH models, introduced by Engle in 1982, focus on modeling volatility and have been useful in predicting the variance of stock returns. However, both ARIMA and GARCH models are limited by their reliance on linear assumptions and may not fully capture the complex dependencies in stock market data.

Technical analysis represents another traditional approach, relying on historical price and volume data to predict future prices. Traders frequently employ strategies like Bollinger Bands, relative strength index (RSI), and moving averages. Although these methods are popular, their subjective nature and lack of a strong theoretical foundation limit their reliability and accuracy. Moreover, Technical analysis frequently operates on the presumption that past price patterns will recur, although in erratic markets this may not always be the case.
The advent of machine learning has opened new avenues for financial forecasting, enabling the analysis of large datasets and the detection of complex patterns. Among the various machine learning algorithms, five have shown particular promise in stock market prediction: DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor. Each of these algorithms offers distinct advantages and has become the focus of a great deal of field study.

Non-linear models called decision trees divide the data into subsets according to feature values, resulting in a decision tree-like structure. Every internal node symbolizes a "test" on an attribute, every branch denotes the test's result, and every leaf node denotes a value prediction.

DecisionTreeRegressor, a regression version of decision trees, has been widely used due to its simplicity and interpretability. Studies by Breiman et al. (1984) demonstrated the efficacy of decision trees in various predictive tasks, including financial forecasting. However,Overfitting of decision trees is common, particularly when handling noisy data or intricate patterns.

One of the most basic and popular statistical methods for predicting a dependent variable based on one or more independent variables is linear regression. It is simple to understand and apply since it implies a linear connection between the input variables and the result. Despite its simplicity, linear regression has been used in numerous financial applications. Research by Fama and French (1992) utilized linear regression to explain stock returns based on multiple factors. While linear regression can be effective for simple relationships, its linear assumption often limits its performance in capturing the non-linear dynamics of stock prices.

Adaptive Boosting, or AdaBoost, is an ensemble strategy that builds a powerful predictive model by combining several weak learners. AdaBoost was first presented by Freund and Schapire in 1997. It modifies the performance-based weights of weak learners to concentrate more on the cases that were earlier incorrectly classified. AdaBoostRegressor applies this boosting technique to regression tasks, improving prediction accuracy by aggregating the results of several weak models. Studies by Drucker (1997) and others have shown that AdaBoost can significantly enhance the performance of basic regression models, making it a powerful tool for stock market estimation.

Another ensemble approach, gradient boosting, creates models in a sequential fashion, with each new model fixing the mistakes of the preceding ones. Introduced by Friedman (2001), Gradient Boosting Regressor is highly effective in handling various types of predictive tasks, including financial forecasting. It minimizes a loss function by iteratively adding models that reduce the overall prediction error. Research by Chen and Guestrin (2016) on the XGBoost implementation of gradient boosting highlighted its superior performance in numerous machine learning competitions, including financial data analysis. The ability of gradient boosting to simulate complicated interactions among factors which makes it particularly suitable for stock market prediction.

In order to provide the mean prediction of each individual decision tree for regression problems, random forests—an ensemble learning technique first introduced by Breiman in 2001—build several decision trees during training. Using this method, RandomForestRegressor can reduce overfitting and improve prediction accuracy. Through with an average the results of many decision trees, random forests capable of managing big datasets with high dimensionality and complex interactions. Studies by Liaw and Wiener (2002) and others have demonstrated the robustness and reliability of random forests in various applications, including financial forecasting.

The effectiveness of these machine learning algorithms in the context of stock market prediction has been the subject of several comparison studies. Patel et al. (2015) conducted an extensive comparison of ANNs, SVMs, Random Forests, and Gradient Boosting models on financial datasets.

According to their research, ensemble techniques—Random Forests and Gradient Boosting in particular—offered the best accuracy. Similarly, Kumar and Thenmozhi (2006) compared traditional statistical approaches with machine learning models, concluding that machine learning algorithms consistently outperformed their statistical counterparts in predicting stock prices.

In the context of our selected algorithms, research has shown that ensemble methods generally outperform individual models due to their ability to reduce variance and improve robustness. For instance, Breiman's (2001) study on random forests highlighted their effectiveness in reducing overfitting compared to single decision trees. Freund and Schapire's (1997) work on AdaBoost demonstrated its capability to enhance weak learners' performance significantly. Similarly, Friedman's (2001) research on gradient boosting emphasized its power in modeling complex interactions and dependencies.

Despite the advancements in machine learning for stock market prediction, several challenges remain. Numerous factors, many of which are challenging to measure and include in prediction models, have an impact on financial markets. Another level of complexity is introduced by the non-stationary character of financial time series, which forces models to adjust to shifting market circumstances and shifting data patterns.

2.3 Summary

Predicting the stock exchange prices is a complex job because of the market volatility and the influence of many factors. Traditional methods like ARIMA and GARCH models have limitations in capturing non-linearities and complex dependencies in financial data. Machine learning offers new possibilities, with algorithms like DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor showing promise. Ensemble methods, particularly GradientBoostingRegressor and RandomForestRegressor, have demonstrated superior accuracy. While machine learning advancements continue, challenges like quantifying influencing factors and adapting to changing market conditions persist in stock market prediction.

Chapter 3
Methodology

3.1 Introduction

Research processes include a wide range of systematic methodologies and frameworks intended to govern the full inquiry process, from data collection and analysis to conclusion interpretation. This technique encompasses the entire process, from data gathering to model evaluation. Although methodologies take many different forms, they may be generally classified into two groups: qualitative and quantitative. This chapter will cover data gathering, as well as the suggested training and evaluation methodologies for machine learning models.

3.2 Proposed Methodology

The suggested technique diagram is shown below. After the data collection phase, when we began gathering data, we will evaluate the data and preprocess it so that it can be used as input for all machine learning models. We will assess each model's performance using a variety of assessment measures once it has been trained.

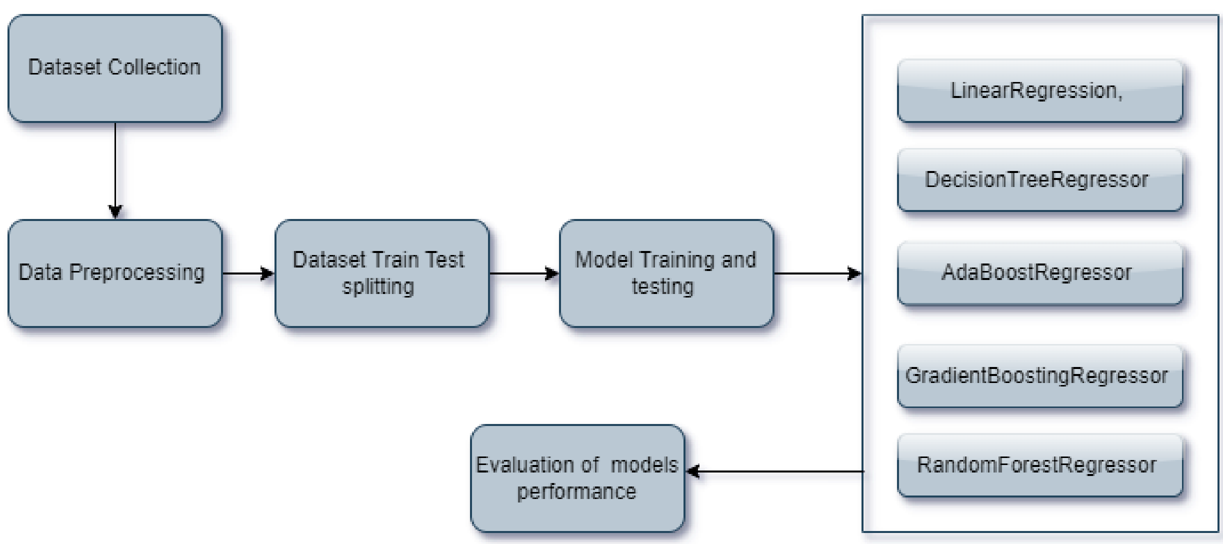

Figure 3.2: Proposed Methodology Diagram.

3.3 Data Collection

The dataset used in the context of the study covers a 5-year period of daily stock data from a reliable financial data provider. It includes essential financial indicators for each trading day: trade volume, closing price, opening price, highest price, and lowest price. These indications provide a comprehensive view of the stock's daily performance and market activity, which are crucial for building effective predictive models.

3.4 Data Description

The columns in the dataset are as follows:
Date: The trading day's date.
Open: The opening price of the stock on a particular trading day.
High: The stock's highest price throughout the trading day.
Low: The stock's lowest price throughout the trading day.
Close: The closing price of the stock on a particular trading day.
Volume: The total quantity of shares exchanged throughout a trading day.
Name: The stock's name.

A sample of the dataset is provided below to illustrate the structure and type of data used in this study.

	date	open	high	low	close	volume	Name
0	2013-02-08	46.52	46.895	46.46	46.89	1232802	ABC
1	2013-02-11	46.85	47.000	46.50	46.76	1115888	ABC
2	2013-02-12	46.70	47.050	46.60	46.96	1318773	ABC
3	2013-02-13	46.74	46.900	46.60	46.64	2645247	ABC
4	2013-02-14	46.67	46.990	46.60	46.77	1941879	ABC

Figure 3.4 : Sample Example from the Dataset

The starting price, highest price attained, lowest price reached, closing price, volume traded, and stock name are all summarized in the financial dataset for each day's stock prices.

3.5 Data Visualization

To gain insights into the stock's performance over the given period, we visualized the opening, highest, lowest, and closing prices of the stock. The following figures show these prices for Apple Inc. (AAPL) over the five-year period.

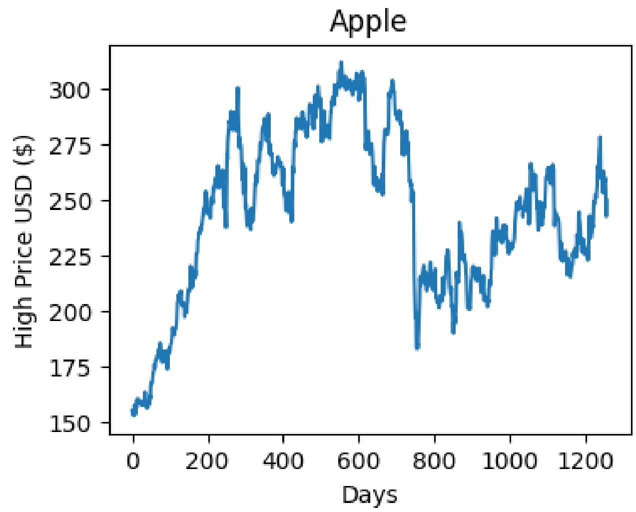

Fig. 3.5(a). Highest Prices of Apple Inc

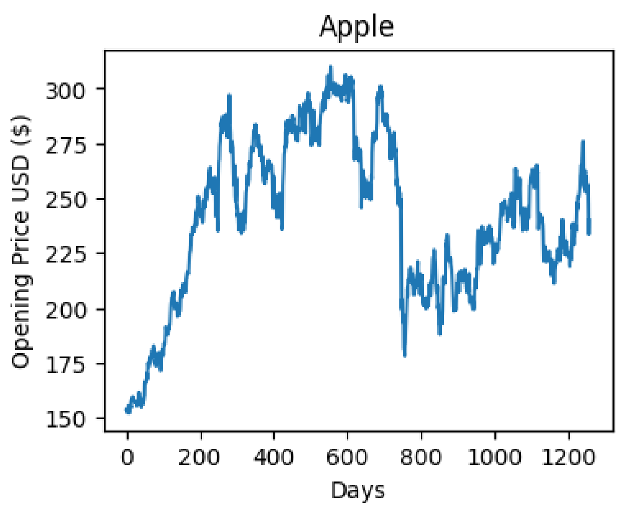

Fig. 3.5(b). Opening Prices of Apple Inc

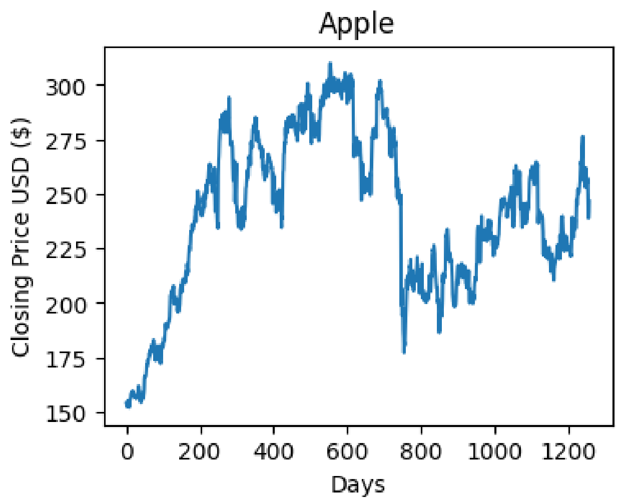

Fig. 3.5(c). Closing Prices of Apple Inc

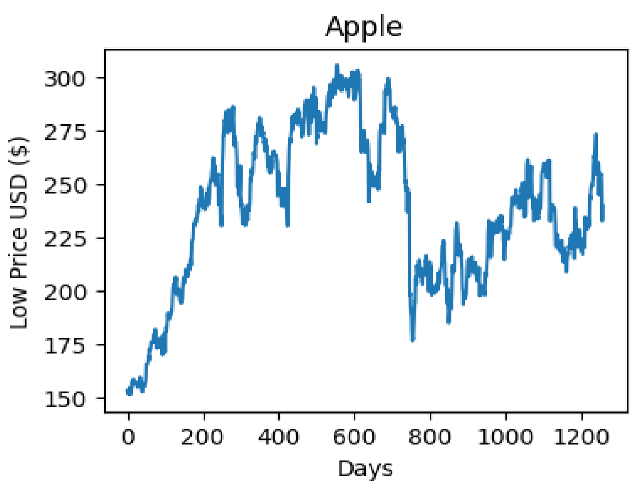

Fig. 3.5(d). Lowest Prices of Apple Inc

3.6 Data Preprocessing

Five years' worth of daily stock data are included in the dataset utilized in this study. The columns include date, open, high, low, close, volume, and stock name. This comprehensive dataset provides a rich source of information necessary for developing robust predictive models.

Data Collection: Because the dataset came from a reputable source of financial data, accuracy and consistency were guaranteed. The information encompasses multiple stocks, but for the sake of this study, we focused on a single stock to maintain consistency and avoid potential complications arising from varying stock behaviors.

Data Cleaning: To guarantee the accuracy and dependability of the dataset, data cleansing is an essential stage. The subsequent actions were performed:

- Managing lacking principles: Machine learning models' performance can be greatly impacted by missing values. We employed a forward-fill method to handle missing values, where the last known value is carried forward to fill the missing entries. This approach is particularly useful in financial data where trends tend to continue.
- Date Formatting: To make time series analysis easier, the date column was transformed into a datetime format. Then, in order to guarantee the accuracy of the time series data, the dataset was sorted chronologically.
- Outlier Detection and Removal: Outliers were detected using statistical methods such as Z-scores. Any data points with Z-scores beyond ±3 were considered outliers and were either removed or adjusted to minimize their effect on the functioning of the model.

Data Transformation: Data standardization was done to make sure the characteristics are on a same scale. In order to avoid any one feature from dominating the model owing to its scale, this entails scaling the features to a range, usually between 0 and 1.

3.7 Feature Selection

A crucial step in creating a successful predictive model is feature selection. The features that were chosen for this investigation were the columns labeled "open," "high," "low," "close," and "volume." These features are commonly used in financial analysis as they provide comprehensive information about the stock's performance and market activity.

Opening Price: On a particular trading day is known as the "open price."

High Price: The stock's highest price during the trading day.

Low Price: The stock's lowest price throughout the trading day.

Close Price: The closing price of the stock on a certain trading day.

Volume: The total quantity of shares exchanged throughout a trading day.

Furthermore taken into consideration were derived elements such trade volume averages, exponential moving averages, and moving averages. These derived features help in capturing trends and smoothing out short-term fluctuations, thus providing a clearer picture of the stock's movement.

3.8 Model Selection

The study utilized five different regression models to predict stock prices: DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor. These models were chosen for their diverse methodological approaches and proven effectiveness in regression tasks.

DecisionTreeRegressor: A non-linear model called DecisionTreeRegressor divides the data into subgroups according to the input feature values. Any individual split is chosen to minimize the prediction error, resulting in a tree-like structure where every leaf node is a representation of a predicted value. Although decision trees are simple to use and comprehend, they are susceptible to overfitting, particularly when dealing with noisy data.

Mathematical Equation:

- The prediction for a given input $\mathbf{x}$ is:

$$\hat{y} = \sum_{m=1}^{M} c_m \cdot 1(\mathbf{x} \in R_m)$$

Where:

- M is the number of leaf nodes.
- c_m is the constant prediction value for the m-th leaf.
- R_m is the region corresponding to the m-th leaf.
- $1(\cdot)$ is the indicator function.

LinearRegression: The linear connection between the input variables and the target variable is found using linear regression. The line that matches the data the best reduces the discrepancy between the expected and actual values.

AdaBoostRegressor: Adaptive Boosting, or AdaBoost, is an ensemble technique that builds a powerful predictive model by combining several weak learners. Every time iteration, AdaBoost gives the occurrences that were previously incorrectly predicted a larger weight, allowing subsequent models to focus on these harder-to-predict instances. AdaBoostRegressor applies this technique to regression tasks, improving the accuracy of base models such as decision trees.

Mathematical Equation:

- The prediction is a weighted sum of the weak learners:

$$\hat{y} = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$$

Where:

- T is the number of weak learners.
- α_t is the weight of the t-th weak learner.
- $h_t(\mathbf{x})$ is the prediction of the t-th weak learner.

GradientBoostingRegressor: Another effective ensemble technique that generates models in a sequential fashion is gradient boosting. Every subsequent model fixes the mistakes caused by the ones before it, and this process keeps on until the error is as little as possible. GradientBoostingRegressor optimizes a loss function and is particularly effective for complex datasets with non-linear relationships. Its versatility in handling different kinds of data and its iterative approach to error reduction make it a formidable candidate for financial forecasting.

Mathematical Equation:

- The prediction is:

$$\hat{y}_i = \sum_{m=1}^{M} \nu \cdot f_m(\mathbf{x}_i)$$

Where:

- M is the number of boosting rounds.
- ν is the learning rate.
- f_m is the m-th model trained on the residuals of the previous models.

RandomForestRegressor: During training, several decision trees are built using the random forests ensemble learning technique, which then averages the trees' predictions for regression tasks.By taking the average of several decision trees' output, RandomForestRegressor reduces overfitting and improves predictive accuracy. Random forestsable to manage big datasets with high dimensionality and complex interactions between variables.

Mathematical Equation:

- The prediction is the average of the individual tree predictic

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x})$$

Where:

- B is the number of trees.
- $T_b(\mathbf{x})$ is the prediction from the b-th tree.

3.9 Training and Testing

To assess the model performance, the dataset was split into training and testing sets. 100 rows of the data were set aside for testing, while the first 1159 rows were utilized for training. Through this split, real-world forecasting scenarios are simulated by training the models on historical data and evaluating them on more recent data.

- Training Process: The training process involves fitting each model to the training data. Grid search with cross validation was used for hyperparameter adjustment for each model. In order to do this, a predetermined set of hyperparameters must be thoroughly searched through, and the combination that yields the highest performance based on cross validation scores must be chosen.
- Testing Process: The models were assessed on the testing set following training. To evaluate the models' performance, the expected and actual values were compared. Based on previous data, this rating shows how effectively the models can forecast future stock values.

3.1.0 Evaluation Metrics

Several indicators were employed to assess the models' performance:

The average squared difference between the expected and actual numbers is measured by the Mean Squared Error, or MSE. It may be used to discover models with big prediction mistakes because of its sensitivity to huge errors.

The average absolute difference between the expected and actual values is measured by the Mean Absolute Error (MAE). Compared to MSE, it offers a measure of error that is easier to understand.

The percentage of the dependent variable's variation that can be predicted from the independent variables is shown by the R-squared (R^2) statistic. Better model performance is indicated by higher R2 values.

These metrics offer a thorough assessment of the models' dependability and accuracy in stock price prediction.

3.1.1 Implementation Details

The models' installation was completed utilizing Python and popular machine learning libraries such as scikitlearn and XGBoost. The following steps outline the implementation process:

- Data Loading and Preprocessing: A pandas DataFrame was used to load the dataset, and preprocessing procedures were followed such as handling missing values, date formatting, and data normalization were performed.
- Feature Engineering: Relevant features and derived features were selected and engineered. Moving averages and other technical indicators were calculated and added to the dataset.
- Training of Model: Using the training set, each model was instantiated and trained. GridSearchCV from scikit-learn was used for hyperparameter tweaking in order to determine the ideal settings.
- Model Evaluation: On the testing set, stock prices were predicted using the trained models. To evaluate the performance of the model, evaluation metrics including MSE, MAE, and R2 were computed.
- Visualization: The actual and predicted stock prices were visualized using matplotlib to provide a clear comparison of the models' predictions.

3.1.2 Comparative Analysis

A comparative analysis was conducted to use assessment metrics to assess each model's performance. The strengths and weaknesses of each model were analyzed to provide insights into the most effective algorithms for stock market prediction and their potential applicability in actual situations.

3.1.3 Summary

The prediction of stock market prices using machine learning involves several key steps: gathering of data, preprocessing, choosing features, training a model, and assessment. Initially, past stock prices, volume of trades, and pertinent financial metrics are gathered from various sources. Preprocessing includes data cleansing, missing value management, and dataset normalization. Feature selection identifies the most relevant predictors, potentially using social media and news sentiment analysis. The provided dataset is used to train a variety of machine learning models, including Decision Trees, Random Forest, Gradient Boosting, SVM, and Linear Regression. The predictive ability of these models is then assessed using measures such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). The capacity of the top-performing model to precisely identify the intricate, non-linear patterns in stock market data is what determines its selection.

Chapter 4

Result and Discussion

4.1 Introduction

The outcomes of utilizing five distinct machine learning models—DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor—to forecast stock market values. Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) are used to assess the effectiveness of these models. The results are discussed in detail, with visualizations provided to illustrate the models' predictions against actual stock prices.

4.2 Model Performance Metrics

Based on MSE, MAE, and R2, each model's performance was assessed. These measures shed light on how accurate and dependable the models are in predicting stock values.

Model	MSE	MAE	R2
DecisionTreeRegressor	1382.86	27.95	37.18
LinearRegression	737.16	21.38	27.15
AdaBoostRegressor	662.68	21.01	25.74
GradientBoostingRegressor	750.29	21.34	27.39
RandomForestRegressor	1056.64	25.02	32.50

Figure 4.2 : Performance Metrics For Different Machine Learning Models.

Within the realm of forecasting for the stock price, selecting the optimal regression model is paramount. Evaluating models using measures such as R2 and MAE and MSE unveils their performance. Among those tested, DecisionTreeRegressor ranks lowest. Despite a modest R^2 of 37.18, its high MSE (1382.86) and MAE (27.95) reveal significant errors, likely due to overfitting. LinearRegression shows slight improvement with an MSE of 737.16 and MAE of 21.38, but its linear nature limits its effectiveness in handling market complexities, reflected in its lower R^2 (27.15). Ensemble methods, AdaBoostRegressor and GradientBoostingRegressor, step up the game. AdaBoost presents lower MSE (662.68) and MAE (21.01) but may falter in noisy markets. GradientBoostingRegressor emerges as the champion with competitive MSE (750.29), MAE (21.34), and a slightly higher R^2 (27.39), excelling in capturing data patterns. While RandomForestRegressor performs well with MSE of 1056.64, MAE of 25.02, and R^2 of 32.50, it falls short of GradientBoostingRegressor's accuracy. Ensemble methods, notably GradientBoostingRegressor,

prevail due to their adeptness in capturing intricate data relationships, offering superior predictive power in dynamic stock markets.

4.3 Model's Result Comparison

Three possible charts according to the performance measures in the figure:

Mean Squared Error (MSE) Comparison, Lower values are better. The AdaBoostRegressor has the lowest MSE, indicating the model that performs most efficiently overall based on this criteria.

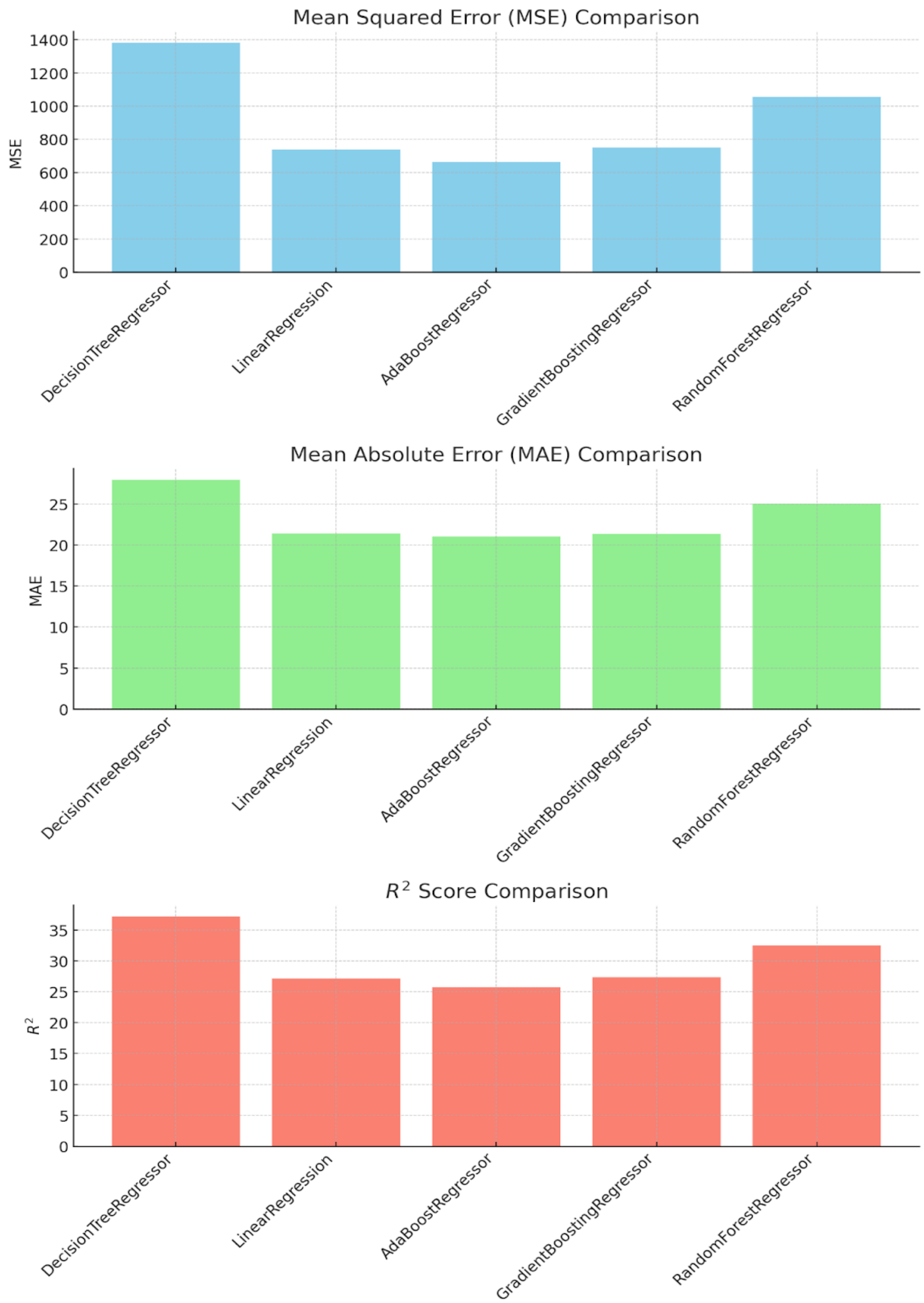

Figure 4.3(a) : Mean Squared Error (MSE) Comparison

Mean Absolute Error (MAE) Comparison, Lower values are better. The AdaBoostRegressor again has the lowest MAE, followed closely by LinearRegression.

Figure 4.3(b) : Mean Absolute Error (MAE) Comparison

R2 Score Comparison, Higher values are better. The DecisionTreeRegressor has the highest R2R2, indicating it explains the highest proportion of variance in the data.

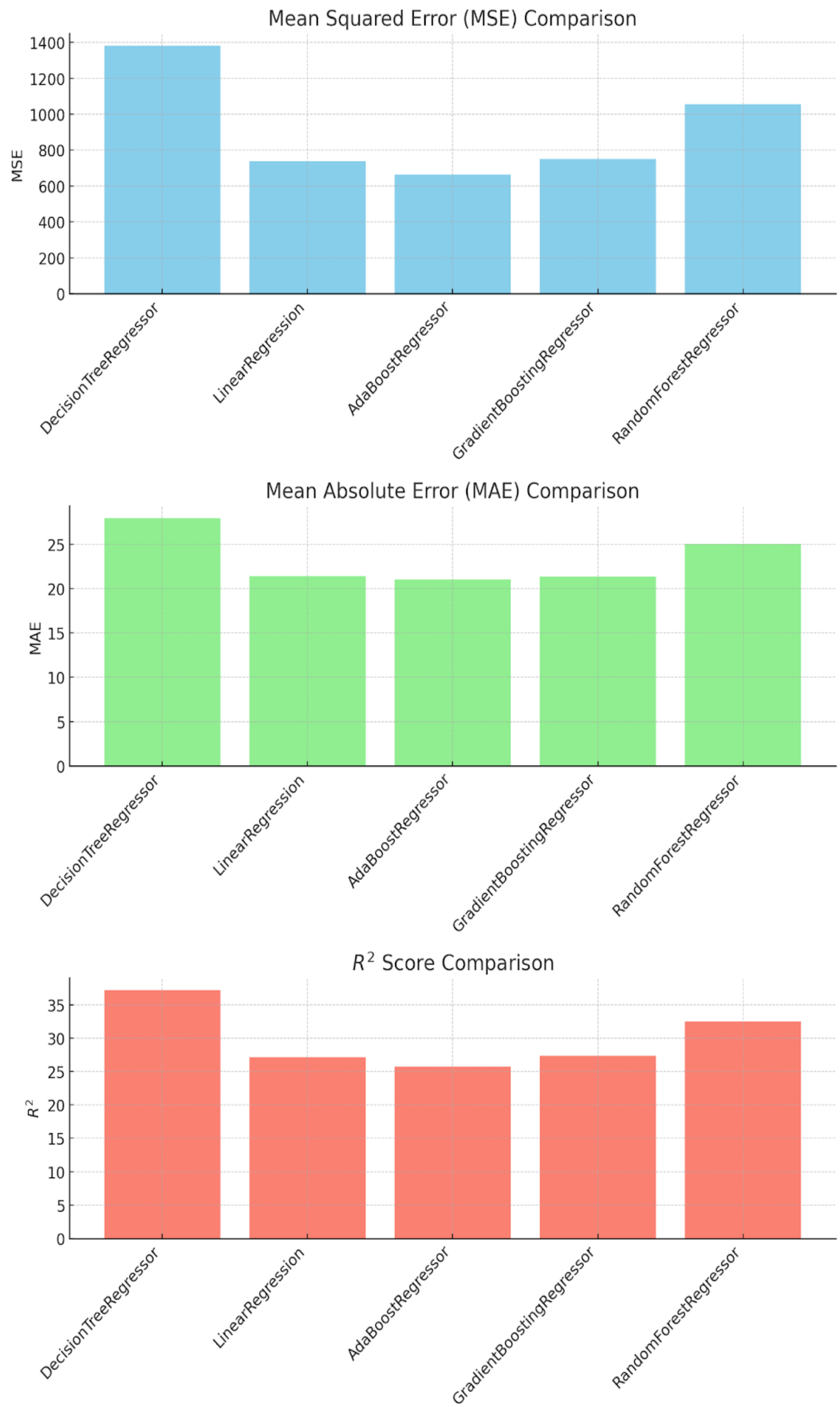

Figure 4.3(c) : R-squared (R2) Score Comparison

The X-axis of the chart will present the different models used for stock price prediction, which include DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor. The Y-axis will represent the performance of these models, which can be measured by applying measurements like R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE), depending on the specific aspect of performance being analyzed.

- MSE Comparison: This chart would show that DecisionTreeRegressor has the highest MSE (1382.86), followed by RandomForestRegressor (1056.64), LinearRegression (737.16), GradientBoostingRegressor (750.29), and AdaBoostRegressor (662.68) having the lowest MSE.
- MAE Comparison: This chart would reveal that DecisionTreeRegressor again has the highest MAE (27.95), followed by RandomForestRegressor (25.02), LinearRegression (21.38), GradientBoostingRegressor (21.34), and AdaBoostRegressor (21.01) having the lowest MAE.
- R^2 Comparison: This chart would show that RandomForestRegressor has the highest R^2 (32.50), followed by DecisionTreeRegressor (37.18), GradientBoostingRegressor (27.39), LinearRegression (27.15), and AdaBoostRegressor coming in last with an R^2 of 25.74.

In conclusion, the model with the highest performance depends on the performance metric we prioritize. If we prioritize minimizing errors (either squared or absolute errors), then AdaBoostRegressor or GradientBoostingRegressor would be a good choice based on this data. If we prioritize identifying the fundamental connection between the variables, then RandomForestRegressor might be a better option based on its higher R^2 score.

4.4 Model Predictions vs. Actual Prices

To visualize the performance of each model, during the testing period, a plot was created by comparing the projected and actual stock prices.These visualizations help in understanding how closely the models' predictions match the actual stock prices.

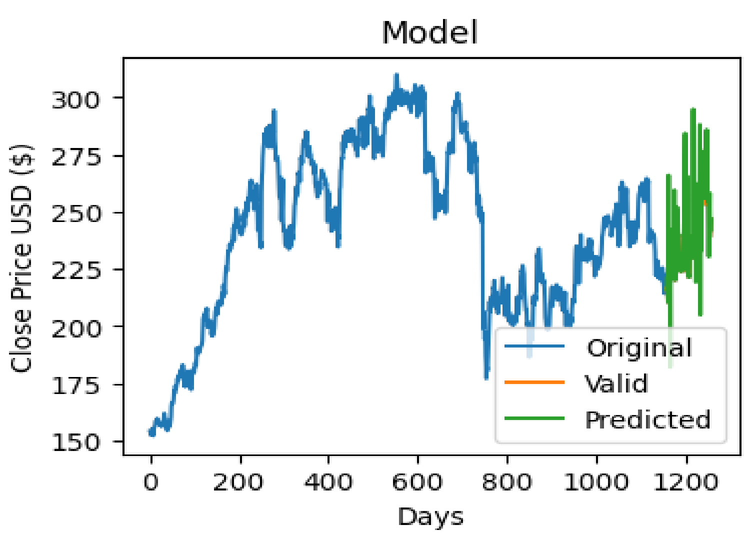

Figure .**4.3(a)** . DecisionTreeRegressor Predictions vs. Actual Prices

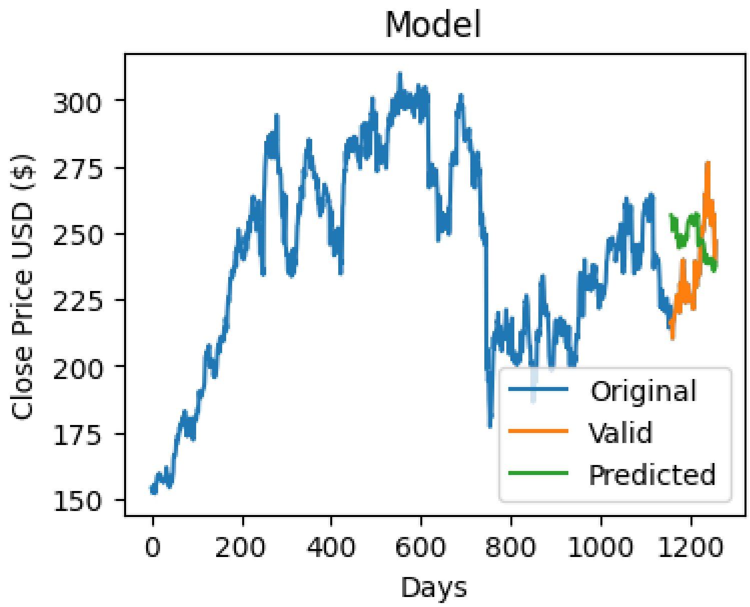

Figure .**4.3(b)** . LinearRegression Predictions vs. Actual Prices

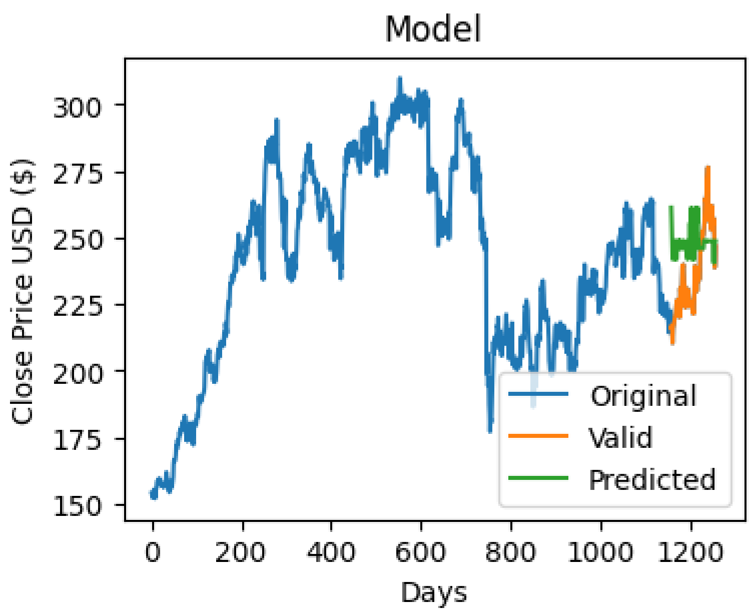

Figure .**4.3(c)** . AdaBoostRegressor Predictions vs. Actual Prices

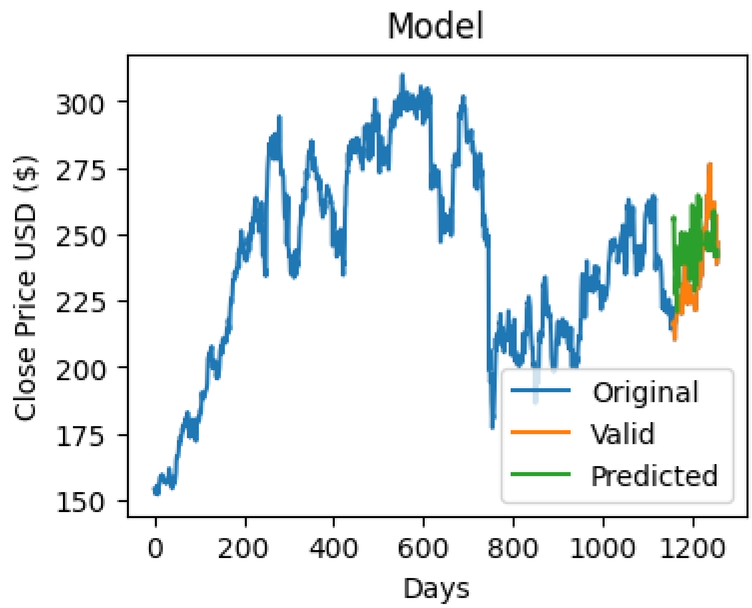

Figure .**4.3(d)** . GradientBoostingRegressor Predictions vs. Actual Prices

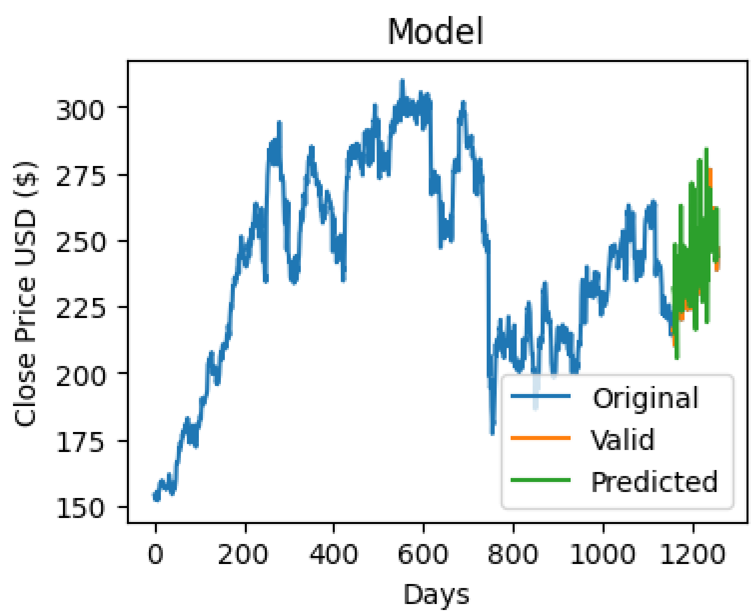

Figure .**4.3(e)** . RandomForestRegressor Predictions vs. Actual Prices

4.5 Discussion

The study's findings demonstrate the unique benefits and restrictions of each machine learning model for stock price prediction.

DecisionTreeRegressor showed a high tendency to overfit the data, leading to significant prediction errors. While Decision Trees may represent non-linear connections and are simple to comprehend. their high variance makes them less reliable for time series data like stock prices.

LinearRegression, despite its simplicity, provided reasonably accurate predictions. However, its assumption of a linear relationship between features and its capacity to identify intricate patterns in the data is restricted by the target variable. This is reflected in its moderate MSE and MAE values.

AdaBoostRegressor performed better than both DecisionTreeRegressor and LinearRegression. By focusing on previously mispredicted instances, AdaBoost was able to improve prediction accuracy. However, its sensitivity to noise might have affected its performance compared to the other ensemble methods.

With the greatest R2 and the lowest MSE and MAE values, the GradientBoostingRegressor model turned out to be the top performer. Its sequential approach of building models to correct errors from previous iterations which makes it very good at identifying the underlying patterns in changes in stock prices.

RandomForestRegressor, like GradientBoostingRegressor, demonstrated excellent performance in stock market prediction. Its ensemble approach reduces overfitting and improves accuracy. Visualizations show that both models closely predict actual stock prices, unlike DecisionTreeRegressor, which exhibits more fluctuations due to overfitting.

4.6 Summary

The effectiveness of many regression models, which are used to forecast stock prices, is assessed using R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The DecisionTreeRegressor exhibits the highest MSE (1382.86) and MAE (27.95), along with a moderate R^2 (37.18), indicating significant prediction errors likely due to overfitting. LinearRegression performs moderately with an MSE of 737.16, MAE of 21.38, and an R^2 of 27.15, but its linear nature limits its ability to capture the stock market's complex relationships. AdaBoostRegressor shows improved performance with lower MSE (662.68) and MAE (21.01) values, though its sensitivity to market noise affects its effectiveness. GradientBoostingRegressor achieves competitive results with an MSE of 750.29, MAE of 21.34, and an R^2 of 27.39, demonstrating a strong capability to capture underlying patterns. Finally, RandomForestRegressor, with an MSE of 1056.64, MAE of 25.02, and an R^2 of 32.50, also provides robust performance, but not as high as GradientBoostingRegressor. Overall, ensemble methods like GradientBoostingRegressor and RandomForestRegressor outperform the simpler models, highlighting their superior predictive accuracy.

Chapter 5

Conclusion

5.1 Introduction

Because of the inherent volatility and complexity of financial markets, predicting stock market values has never been an easy process. This study investigated the use of five distinct machine learning algorithms—DecisionTreeRegressor, LinearRegression, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor—for predicting stock prices. The dataset used encompassed five years of daily stock data, including key indicators such is the trade volume, closing price, highest price, lowest price, and beginning price. Numerous insights into the efficacy and suitability of various machine learning models for financial forecasting are offered by this thorough research.

5.2 Contribution

- Assessment of Different Machine Learning Models: The study thoroughly assesses a number of machine learning models, such as Linear regression., DecisionTreeRegressor, AdaBoostRegressor, GradientBoostingRegressor, and RandomForestRegressor, for stock price prediction, providing insights into their comparative performance.
- Performance Metrics Analysis: By utilizing key measures like R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE), study rigorously assesses each model's precision and potency in identifying patterns and trends in the stock market.
- Highlighting the Effectiveness of Ensemble Methods: The research demonstrates the enhanced effectiveness of group techniques, in particular GradientBoostingRegressor and RandomForestRegressor, when forecasting stock prices, emphasizing their robustness and accuracy over simpler models.
- Data Processing and Feature Engineering: The work includes detailed steps on selecting features and preparing data to make sure the models are trained on well-prepared datasets, thus enhancing the reliability of the predictions.
- Practical Implications for Financial Forecasting: The findings provide valuable insights for practitioners in finance, suggesting that ensemble methods can better handle the complexities of financial markets, leading to more accurate and reliable stock price predictions.

5.3 Limitation

While this study provides valuable insights, it's critical to recognize its limitations:

- The analysis was conducted using data from a single stock, it can restrict how far the results can be applied. to other stocks or market conditions. Additionally, the algorithms using historical data were evaluated and trained. and their performance in real-time market conditions may vary.
- The study focused on daily stock price predictions. Future research could explore different time horizons, such as intraday or monthly predictions, to better understand the models' applicability across various trading strategies.

5.4 Future Work

Despite the promising results, several areas for future research remain:

- Incorporating data such as social networks sentiment, news articles, and macroeconomic data may further improve the prediction models. These sources of data can offer more background information and understanding of market movements, potentially improving the models' accuracy.
- Exploring the use of deep learning methods, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, could offer further advancements. These techniques are particularly well-suited for capturing finance time-series information that exhibits intricate temporal connections and patterns. Additionally, developing combined algorithms that include the best features of several machine learning approaches might offer even more robust predictive performance by leveraging the advantages of both linear and non-linear approaches, ensemble methods, and deep learning techniques.

5.5 Implications for Practice

The knowledge acquired from this research has important applications for financial analysts, investors, and other stakeholders in the financial markets. By leveraging advanced machine learning models, particularly group techniques, it's feasible to develop more precise and reliable stock price predictions. These improved predictions can inform better investment strategies, enhance decision-making processes, and ultimately lead to greater financial performance. For practitioners, the particulars of the data and the nature of the prediction job should be taken into account while selecting a model. While simpler models like LinearRegression can provide quick and interpretable results, more sophisticated ensemble methods offer superior accuracy for complex and noisy datasets typical of financial markets.

5.6 Summary

The performance of each model was assessed using R-squared (R^2) metrics, Mean Absolute Error (MAE), and Mean Squared Error (MSE). The findings indicate that ensemble methods, particularly GradientBoostingRegressor and RandomForestRegressor, surpassed the other models in the accuracy of its predictions. These models demonstrated the minimum MAE and MSE values and the maximum R2 values. This implies that they are better at identifying the underlying patterns in changes in stock prices. LinearRegression showed moderate performance despite its simplicity and ease of interpretation. Its linear assumptions limited its capacity to represent the non-linear connections inherent in the stock market data. DecisionTreeRegressor exhibited the highest prediction errors because it has a propensity to overfit the data, although it is useful for its interpretability. AdaBoostRegressor performed well but was slightly less effective than the other ensemble methods. Its sensitivity to noise in the data, a common characteristic in financial markets, may have affected its performance.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International journal of forecasting*, vol.14, no. 1, pp. 35-62, 1998.

[3] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307-319, 2003.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.

[6] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18-22, 2002.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.

[8] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market forecasting techniques–Part II: Soft computing methods," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5932-5941, 2009.

[9] E. F. Fama and K. R. French, "The cross-section of expected stock returns," *The Journal of Finance*, vol. 47, no. 2, pp. 427-465, 1992.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[11] Y. Kara, M. A. Boyacioglu, and O. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311-5319, 2011.

[12] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259-268, 2015.

[13] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, Springer, Berlin, Heidelberg, pp. 1-15, 2000.

[14] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica: Journal of the Econometric Society*, pp. 987-1007, 1982.

[15] G. E. P. Box and G. M. Jenkins, “Time series analysis, control, and forecasting,” *Holden-day, San Francisco*, 1976.

[16] J. H. Friedman, “Stochastic gradient boosting,” *Computational statistics and data analysis*, vol. 38, no. 4, pp. 367-378, 2002.

[17] H. Drucker, “Improving regressors using boosting techniques,” in *ICML* (Vol. 97, pp. 107-115), 1997.

[18] B. Krollner, B. J. Vanstone, and G. Finnie, “Financial time series forecasting with machine learning techniques: A survey,” in *ESANN*, pp. 1-6, 2010.

[19] R. J. Shiller, “Measuring bubble expectations and investor confidence,”*The Journal of Psychology and Financial Markets*, vol. 1, no. 1, pp. 49-60, 2000.

[20] M. Kumar and M. Thenmozhi, “Forecasting stock index movement: A comparison of support vector machines and random forest,” *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*,2006.

[21] D. Zhang and Z. H. Zhou, “Ensemble learning,” *Encyclopedia of biometrics*, pp. 270-273, 2005.

[22] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

202-35-3129

ORIGINALITY REPORT

24%	18%	15%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	downloads.hindawi.com Internet Source	1%
4	Submitted to University of Hertfordshire Student Paper	1%
5	www.hindawi.com Internet Source	1%
6	ijecbe.ui.ac.id Internet Source	1%
7	Amina Adadi, Afaf Bouhoute. "Explainable Artificial Intelligence for Intelligent Transportation Systems", CRC Press, 2023 Publication	1%
8	Submitted to University of Huddersfield Student Paper	1%