**Project Report**

**Stock Price Analysis of Yahoo Finance S&P500**

**Zainab Hashmi**

**Abstract :**

* In this stock market analysis project, we have embarked on a journey to unravel the mysteries of the S&P 500 index using advanced machine learning techniques. Our primary goal is to assist investors, traders, and financial analysts in making informed decisions by providing accurate forecasts of stock prices.
* Our project commenced with a meticulous exploration of the S&P 500 dataset sourced from Yahoo Finance. We have diligently cleaned and preprocessed this data to ensure its suitability for analysis. Through exploratory data analysis (EDA), we have unearthed valuable insights regarding historical stock prices, trading volumes, and price trends. Furthermore, we have ventured into the realm of time series analysis to comprehend the underlying patterns, seasonality, and stationarity inherent in the data.
* To accomplish our goal of predicting stock prices, we have harnessed the power of various machine learning models, including linear regression, Random Forest Regressor, and XGBoost. We have meticulously optimized these models using advanced techniques such as GridSearchCV. The results of our analysis have illuminated the potential of these models to forecast stock prices with a high degree of accuracy.
* Transitioning to the realm of classification analysis, we have explored the classification of stock price trends, discerning whether they are trending up, down, or remaining stable. A Random Forest Classifier has been rigorously trained and evaluated for this purpose, providing valuable insights for investors seeking to make data-driven trading decisions.
* Additionally, we have introduced advanced analysis techniques such as rolling averages and trend indicators to enhance the classification model's performance. Furthermore, we have implemented an ARIMA model for stock price forecasting, offering a glimpse into potential future price trends.
* As we conclude this project, we present a comprehensive summary of our key findings, highlighting our successes and the challenges we encountered during our journey. We acknowledge that financial markets are dynamic and ever-changing, and we recognize the need for continuous research and improvement.
* Through this project, our mission is to empower investors and financial analysts with the tools and insights required to navigate the complex world of stock market analysis. We aspire to facilitate data-driven decision-making in pursuit of financial success.

# **Introduction:**

* This project is a journey into the world of stock market analysis, with a specific focus on the influential S&P 500 index. The S&P 500 serves as a critical barometer of the U.S. economy, comprising 500 prominent American companies. Our central aim is to leverage advanced machine learning techniques to develop predictive models for stock price movements.
* Our journey begins with the collection of historical stock price data from Yahoo Finance, emphasizing the importance of high-quality and complete data as the foundation for our analysis. Data preprocessing is a pivotal step, involving tasks such as cleaning, handling missing values, and feature transformation to prepare the data for analysis.
* Feature engineering is the next stage, where we identify and craft relevant features from the raw data, enhancing the models' predictive capabilities. We explore regression models like Linear Regression, Random Forest Regressor, and XGBoost to predict stock prices as continuous values.
* Shifting our focus to classification, we employ a Random Forest Classifier to predict stock price movements, categorizing them as up, down, or stable. This provides actionable insights for investors seeking guidance on potential price trends.
* Recognizing the time-dependent nature of stock prices, we delve into time series analysis. Here, the ARIMA model takes center stage, enabling us to forecast stock prices and capture time-dependent trends.
* Model performance evaluation is a rigorous process, with the use of appropriate metrics and techniques such as hyperparameter tuning for optimization.
* In summary, this project embodies the essence of stock market analysis in the modern era. It combines data collection, preprocessing, feature engineering, and a variety of machine learning techniques to empower investors and financial analysts with valuable insights for informed decision-making.

# Literature Review :

* The literature review on stock price analysis and prediction underscores the significant growth and relevance of this field. Machine learning techniques have emerged as powerful tools for enhancing predictive accuracy and understanding the complexities of financial markets.
* Key themes and approaches highlighted in the review include:
* Regression Models: Linear Regression, Random Forest Regressors, and XGBoost are commonly used regression models to predict stock prices. They offer the flexibility to capture both linear and non-linear patterns in data.
* Classification Models: Beyond predicting continuous stock prices, classification models like Random Forest Classifiers have gained popularity for categorizing price movements (up, down, or stable). These models provide actionable insights for investors.
* Time Series Analysis: Given the time-dependent nature of stock prices, time series analysis, especially ARIMA models, has been pivotal in modeling and forecasting stock prices, accounting for temporal dependencies.
* Feature Engineering and Data Preprocessing: Effective feature engineering and data preprocessing are crucial for model performance. These steps involve creating meaningful features from raw data and ensuring data is appropriately scaled and transformed.
* Evaluation Metrics: Evaluation metrics such as MSE, RMSE, R2 for regression models, and precision, recall, and F1-score for classification models are commonly used to assess model performance.
* Challenges in this domain include the dynamic nature of financial markets influenced by various factors, the risk of overfitting, and the fundamental assumption of stationarity in time series analysis.
* In conclusion, the literature review showcases the evolving landscape of stock price prediction, where machine learning techniques offer promise but also present challenges. This project contributes to this field by applying machine learning models to analyze historical data of the S&P 500 index, with the goal of providing valuable insights to stakeholders in finance.

# **Econometric Theory :**

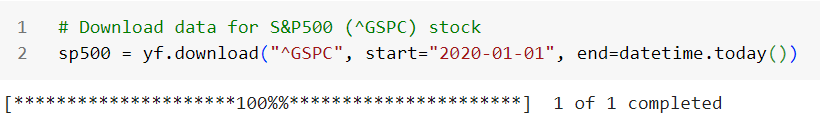
* Efficient Market Hypothesis (EMH): EMH, proposed by Eugene F. Fama, asserts that stock prices reflect all available information, making it challenging to consistently outperform the market. It includes Weak, Semi-Strong, and Strong forms, influencing how we analyze stock predictability.
* ARIMA: ARIMA combines auto-regressive, integrated, and moving average components to model time series data, particularly for stock price forecasting, capturing trends and seasonality.
* Cointegration: This concept highlights long-term relationships between non-stationary time series, crucial for pairs trading strategies involving multiple assets.
* Risk-Return Tradeoff: It emphasizes that higher expected returns often come with higher risk (volatility), a vital consideration for stock selection and portfolio construction.
* Model Validation: Ensuring model robustness and generalization through out-of-sample testing is critical, especially in financial applications.
* Stock Behavior and Its Impact on Investors:
* Volatility and Risk Perception: Volatility measures price variation, influencing investors' risk perceptions. Stocks with turbulent histories are often seen as riskier.
* Market Sentiment and Behavioral Finance: Emotional reactions and cognitive biases like overconfidence and herd behavior can drive investor decisions and market behavior.
* Technical Analysis and Chart Patterns: Technical analysts use historical price patterns, moving averages, and support/resistance levels to predict future price movements.
* Fundamental Analysis and Long-Term Investment: Fundamental factors, like earnings and growth prospects, guide long-term investors. Dividend yield matters for income-focused investors.
* Trading Strategies and Technical Signals: Short-term traders use technical indicators and algorithms for rapid decisions, responding to stock behavior.
* Psychological Factors and Investment Goals: Risk tolerance and investment objectives impact stock selection and behavior alignment with individual goals.
* Impact of News and Events: News-driven events significantly influence stock behavior, with investors reacting to corporate events and economic news.

# **Methodology :**

The methodology section outlines the approach, tools, and techniques used to conduct stock price analysis and build predictive models. In this project, we employ a comprehensive methodology that encompasses data collection, preprocessing, exploratory data analysis (EDA), and the development of both regression and classification models. Below is an overview of the methodology:

# **Data Source :**

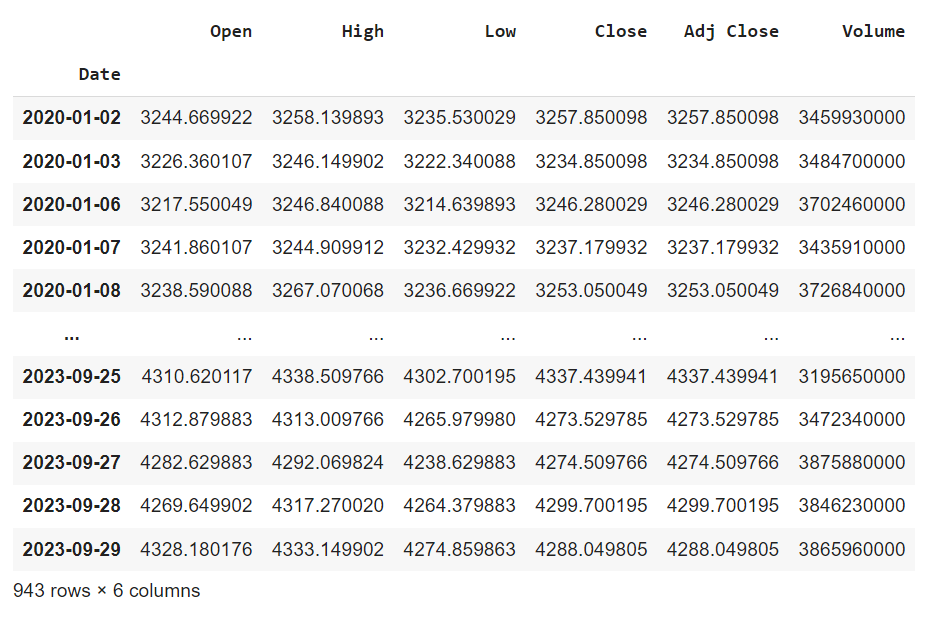
The primary data source for this project is Yahoo Finance, a widely respected platform known for its extensive coverage of global financial markets. Yahoo Finance offers historical and real-time data on various financial instruments, including stocks, indices, currencies, and commodities. Its user-friendly interface makes it a preferred choice for investors, analysts, and researchers worldwide.

We utilized Python's Yahoo Finance API to access historical stock price data for our selected time frame. Our focus was on the S&P 500 index (^GSPC), which represents 500 major publicly traded U.S. companies. This index serves as a benchmark for assessing the overall performance of the U.S. stock market. The dataset obtained from Yahoo Finance covers daily historical data for the S&P 500 index, spanning from January 1, 2020, to the present date.

# **Data Set :**

The dataset used for this project is a treasure trove of information about the S&P 500 index, which is like a yardstick for the U.S. stock market. It's like a giant Excel sheet that records what happened to this index every day, starting from January 1, 2020, up to today

Now, what kind of information does this dataset hold? Well, it tells us things like the opening and closing prices of the S&P 500, which are like the first and last prices of the day. It also tracks the highest and lowest prices during each trading day, which gives us a sense of how much the market swings. And, it counts how many shares were bought and sold each day, which we call trading volume.Imagine it's like a diary for the S&P 500. Every day, it writes down what the index did, and we can use this information to learn some really cool stuff. For example, we can see if the stock market is going up or down over time, or if there are patterns in how it moves. This dataset is like a window into the world of finance, and it's our job to analyze it and uncover its secrets.



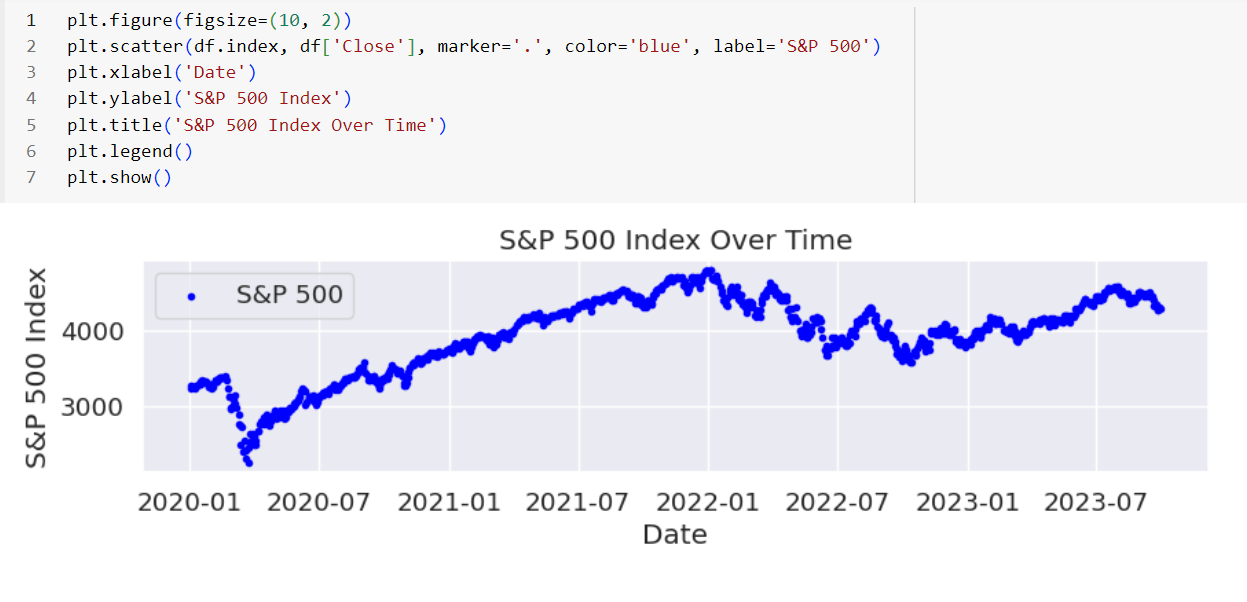
**Data Cleaning and Preprocessing :**

The dataset used in this project was exceptionally clean, with no presence of null values or duplicate entries. It provided a comprehensive daily record of stock prices, making it a reliable and valuable resource for our analysis.

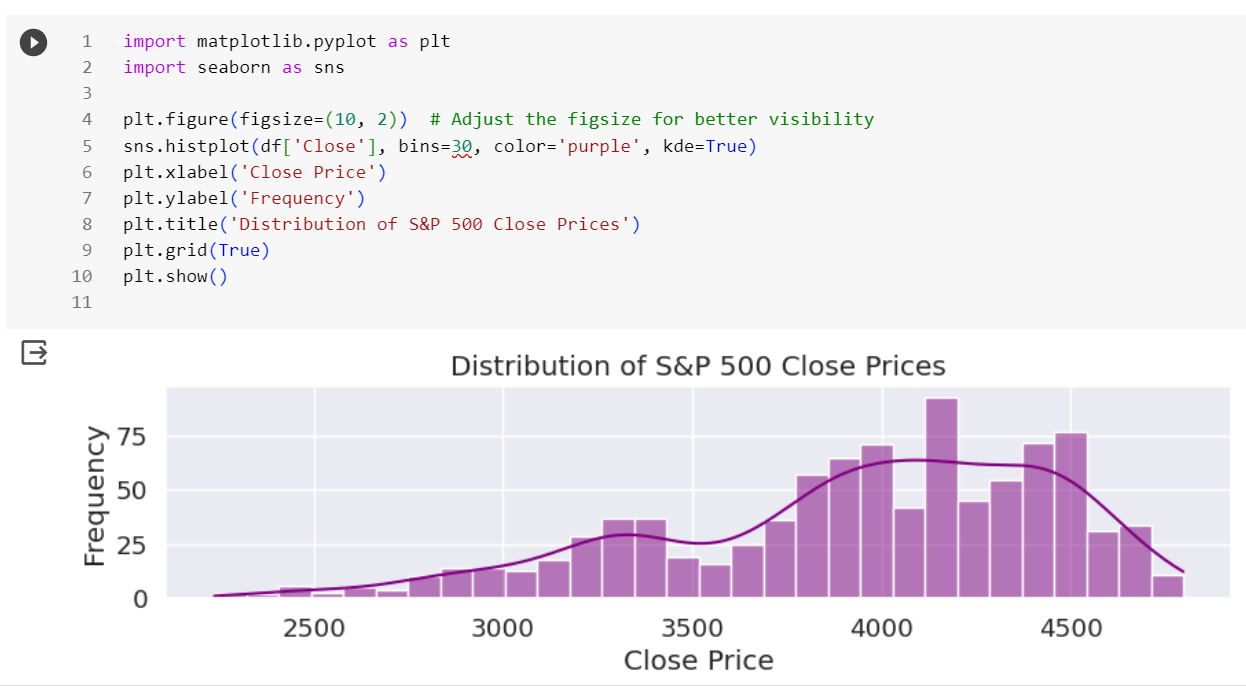
**Exploratory Data Analysis (EDA)**

During our exploratory data analysis (EDA), we initially checked the data description and found that there were no missing values, indicating that the dataset was complete. We then proceeded to visualize the historical trend of the S&P 500 Index using a line chart.

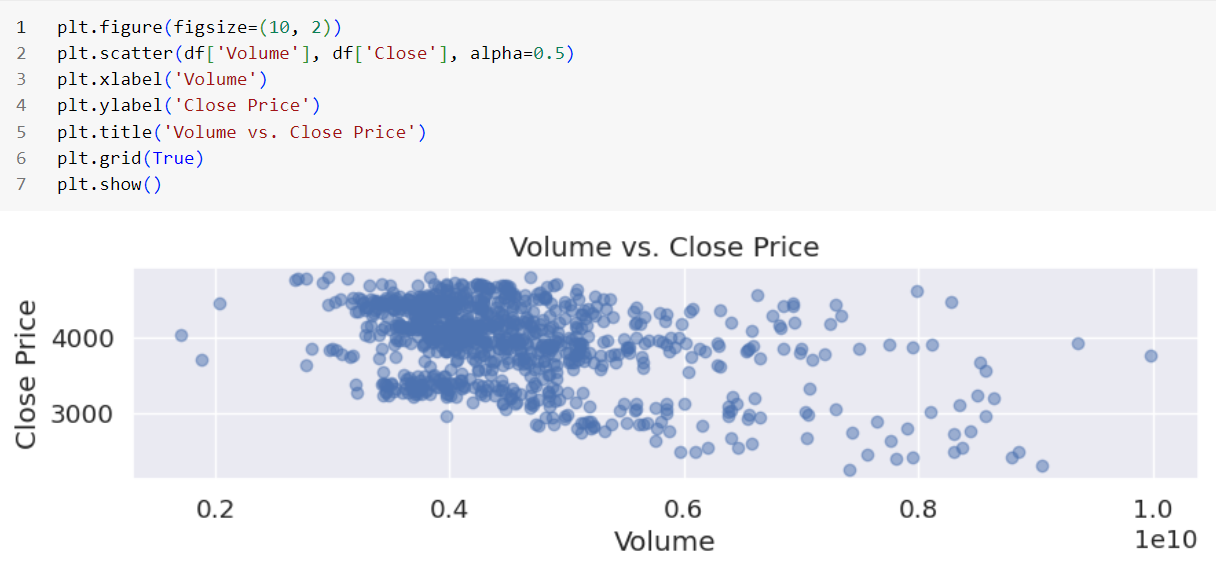
The line chart clearly illustrates an upward trend in the S&P 500 Index from January 2020 to July 2023. The positive slope of the line graph indicates a consistent increase in the index value during this period, with the exception of a notable downturn in 2020.



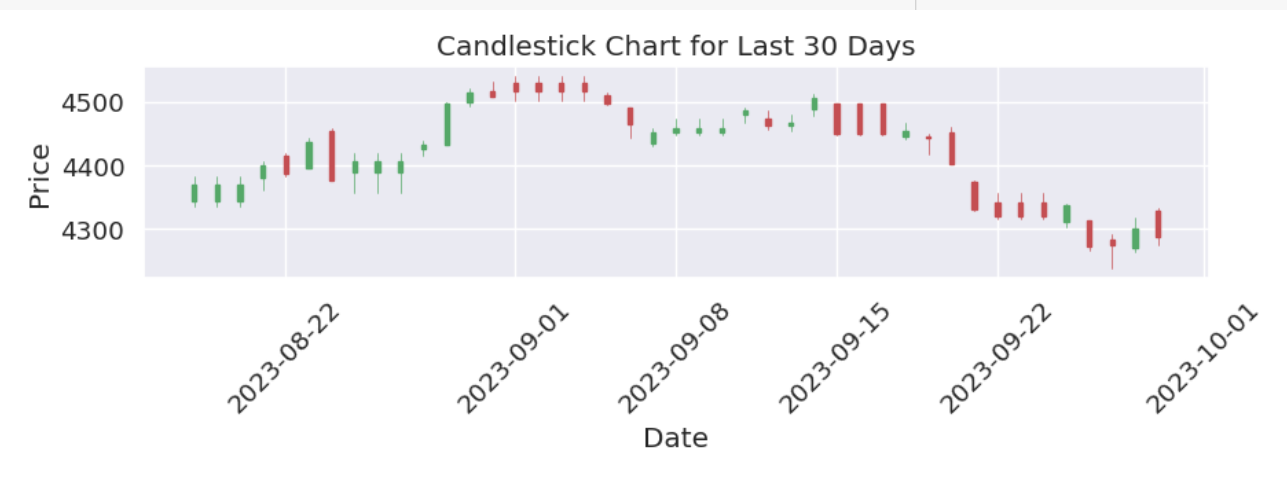
Moving on to the distribution of the closing prices, we observed that it displayed a left-skewed pattern. This left-skewness implies that a majority of the closing price data points are concentrated to the right of the mean value, while there are some extreme values on the left side of the distribution.



Finally, we examined the relationship between the volume and the closing price by creating a scatter plot. In this scatter plot, we did not observe a clear linear relationship between the two variables. The data points were scattered rather than forming a distinct straight line. As a result, it's likely that the correlation coefficient between volume and closing price is close to 0, indicating the absence of a strong linear relationship between these variables.



**The Candlestick Chart** displays price movements over a specified time period, using individual "candlesticks" to represent each day's trading data, including open, close, high, and low prices. The color and shape of these candlesticks convey information about price changes. In this 30-day candlestick chart, red candlesticks indicate closing prices lower than opening prices, while green ones indicate the opposite. The chart reveals a general 30-day downtrend with more red than green candlesticks and lower closing prices, along with occasional price increases. Traders and investors use candlestick patterns to analyze trends and identify potential trading opportunities, such as bullish or bearish reversals, based on price patterns and trends.



# **Regression Model :**

**Linear regression:**

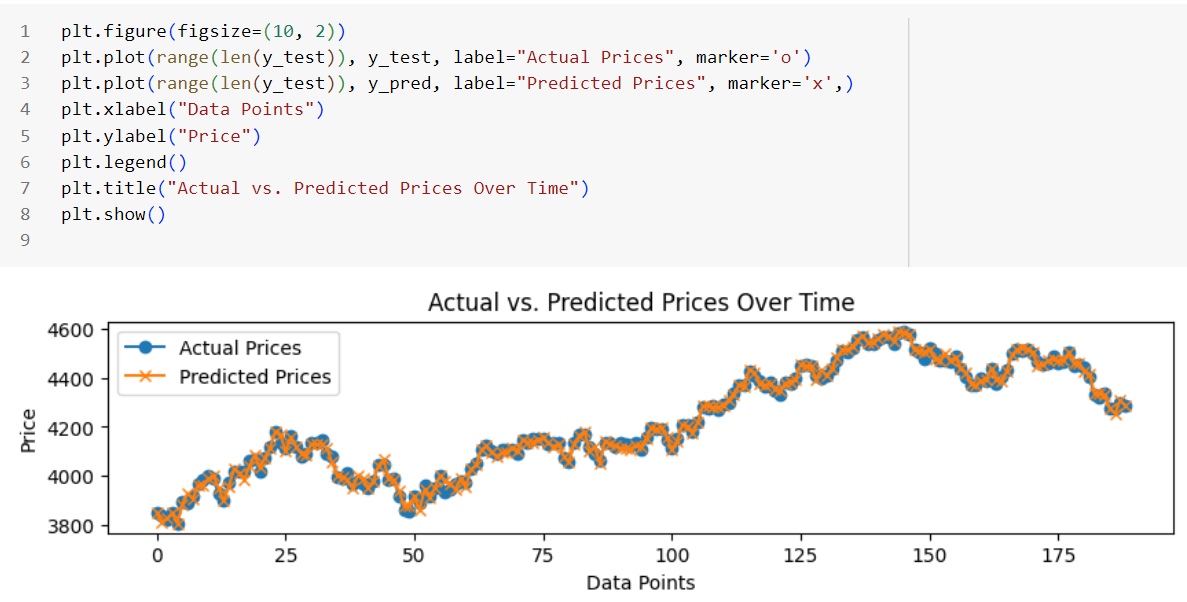
Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the features and the target and aims to find the best-fitting linear equation to make predictions. So, we have used Regression Models to predict the next day’s closing price.

Root Mean Squared Error: 12.531365245190138 Accuracy : 99.65% Predicted Price for the next day : 4287.8875050512215

* Mean Squared Error (MSE): This metric measures the closeness of predicted values to observed values. Lower MSE indicates a better fit, implying less error in the model's predictions.
* Root Mean Squared Error (RMSE): It is the square root of MSE, providing an easier-to-interpret measure in the original units of the data.
* R-squared (R2): This metric assesses how well the model explains data variation. Higher R2 suggests a better fit, indicating that the model accounts for more of the data's variability.

The RMSE of our model is 12.53136524590138 and the accuracy is 99.65%. The accuracy is the percentage of correct predictions made by the model.

The increasing trend in the line graph shows that the actual and predicted prices of the product are rising over time. This means that the product is becoming more expensive and more valuable.



**Evaluation and Optimization for Regression Model ( Linear Regression ):**

GridSearchCV is employed to optimize hyperparameters such as regularization L1, Ridge regression, a type of regularized linear regression that adds a penalty term to prevent overfitting. GridSearchCV is employed to find the optimal value of the regularization parameter (alpha).

Comparing this to a linear regression model, the Ridge regression model outperforms it significantly with a much lower RMSE. This improvement is likely due to Ridge regression's ability to reduce multicollinearity and the effect of outliers through regularization. However, Ridge regression slightly sacrifices variance for bias, explaining slightly less variation in the data.

Furthermore, the predicted prices for the next day differ between the two models. The Ridge regression predicts a price of 4288.89, while the linear regression predicts 4287.89. This difference is likely attributed to distinct coefficient estimates in the two models.

In essence, model evaluation is crucial to compare their performance, select the best model for Our data and problem, and identify potential issues or areas for improvement, such as overfitting or underfitting. It helps quantify how well the models fit the data and make accurate predictions, guiding further model tuning and refinement.Optimized Root Mean Squared Error (RMSE): 12.531365227625248 Optimized R-squared (R2): 99.64880886247659

**XGBoost Regressor:**

XGBoost is a powerful ensemble machine learning algorithm that can be used for regression tasks. It uses decision trees as base learners and combines their predictions to make accurate predictions. It's known for its high performance and ability to handle complex datasets. Root Mean Squared Error: 16.72237519972091 Accuracy: 99.37% Predicted Price for the next day: 4288.001953125

**Random Forest Regressor:**

Random Forest is another ensemble algorithm used for regression. It builds multiple decision trees and averages their predictions, which helps reduce overfitting and improve accuracy.Root Mean Squared Error: 16.40006722131821 Accuracy : 99.4% Predicted Price for the next day : 4287.624970703125

**Conclusion:**

* Both Ridge Regression and Linear Regression have the lowest RMSE (12.53), indicating that they have the smallest prediction errors. This suggests that they provide the best fit to the data in terms of minimizing prediction errors.
* Linear Regression and Ridge Regression have the highest accuracy (99.65%), which means they make the most correct predictions compared to Random Forest and XGBoost.
* While Random Forest and XGBoost have very high accuracy and relatively low RMSE, they slightly underperform compared to Ridge and Linear Regression in terms of minimizing prediction errors.
* The predicted prices for the next day are quite similar across all models, with only minor variations.
* In conclusion, both Ridge Regression and Linear Regression perform exceptionally well on Our dataset. They provide the lowest prediction errors and the highest accuracy. The choice between these two models may come down to interpretability and simplicity, as Linear Regression is more straightforward to explain and understand. However, Ridge Regression offers regularization to handle multicollinearity and potential overfitting. Therefore, the best model for Our dataset could be either Ridge Regression or Linear Regression.

# **Classification Model :**

**Creating the Target Variable for Classification:**

In our classification model, we began by creating the target variable. This variable, known as "Target," is essential for predicting whether the stock price will go up or down on the next day. We achieved this by comparing the closing price of one day with the closing price of the following day. If the closing price increased, we labeled it as 1 (indicating "up"), and if it decreased, we labeled it as 0 (indicating "down").

**Random Forest Classifier:**

We employed the Random Forest Classifier, a powerful machine learning technique. This ensemble learning method utilizes multiple decision trees to make predictions. Each decision tree in the forest contributes to the final prediction, leading to more accurate and robust results.

Optimization of the Classification Model after Using Data Mining Techniques:

To enhance the performance of our classification model, we implemented various data mining techniques. This involved:

**Columns Added in the Model to Improve Precision Score:**

We introduced additional columns in our model to enhance its precision score. These columns were derived from rolling averages and trends for different time horizons. Specifically, we calculated the closing price ratio by dividing the closing price by the rolling average of closing prices for a specified time window. We also computed trends by summing up the actual targets for the same time horizons. These new features aimed to capture meaningful patterns and movements in the stock price data, contributing to improved prediction accuracy.

**Predictions Whether the Stock Will Go Up or Down:**

Our model's primary objective was to predict whether the stock price would exhibit an upward or downward movement on the next trading day. We achieved this through the Random Forest Classifier, which assigned binary class labels: 0 for "down" and 1 for "up."

**The Class Labels:**

In our classification, the class labels were defined as follows:

Class 0: Represents a prediction that the stock price will go "down" on the next day.

Class 1: Indicates a prediction that the stock price will go "up" on the next day.

**Conclusion:**

In conclusion, we built and evaluated a classification model using the Random Forest Classifier to predict stock price trends based on the S&P 500 index data. We introduced new features derived from rolling averages and trends to enhance the model's precision score. Despite improvements, our model exhibited a bias towards predicting up movements and had room for further enhancement. The Random Forest algorithm played a pivotal role in achieving accurate predictions in this financial context.

# **Time Series Analysis :**

**Introduction:**

In this time series analysis, we delved into the intricacies of stock price data with a focus on forecasting for the first week of October. Our approach involved various techniques and tools to uncover patterns, make predictions, and assess the model's performance.

**Key Insights and Actions Taken:**

**Data Preprocessing:**

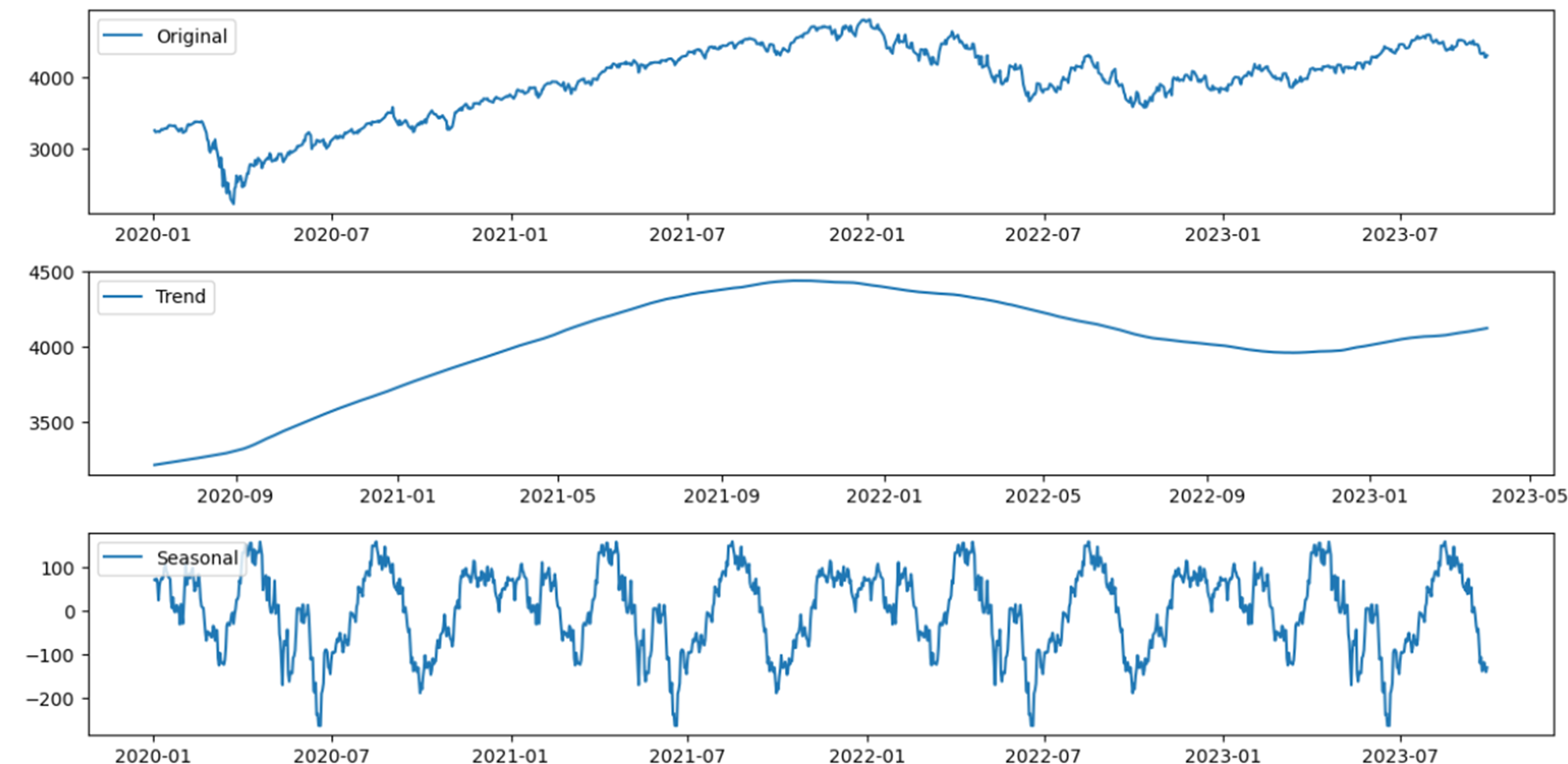
We resampled the daily stock price data to ensure consistent time intervals.

The plotted closing prices over time revealed clear trends and non-linearity in the data.

**Decomposition:**

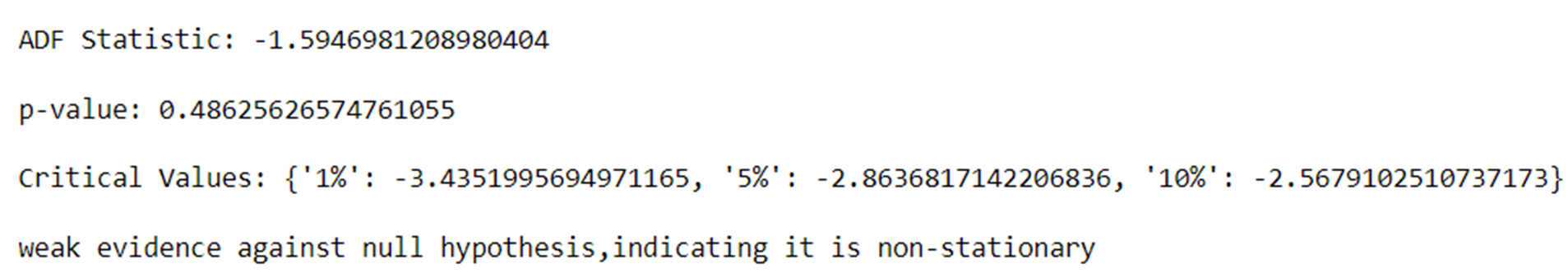
Through decomposition, we broke down the time series into its constituent elements: trend, seasonality, and residuals.

This decomposition helped us identify the presence of underlying patterns and variations, notably showcasing trends and seasonality.



**Stationarity Test:**

We conducted the Augmented Dickey-Fuller (ADF) test to gauge the stationarity of the data.

The results indicated non-stationarity, implying the presence of trends or seasonality.****

**Lag Plot:**

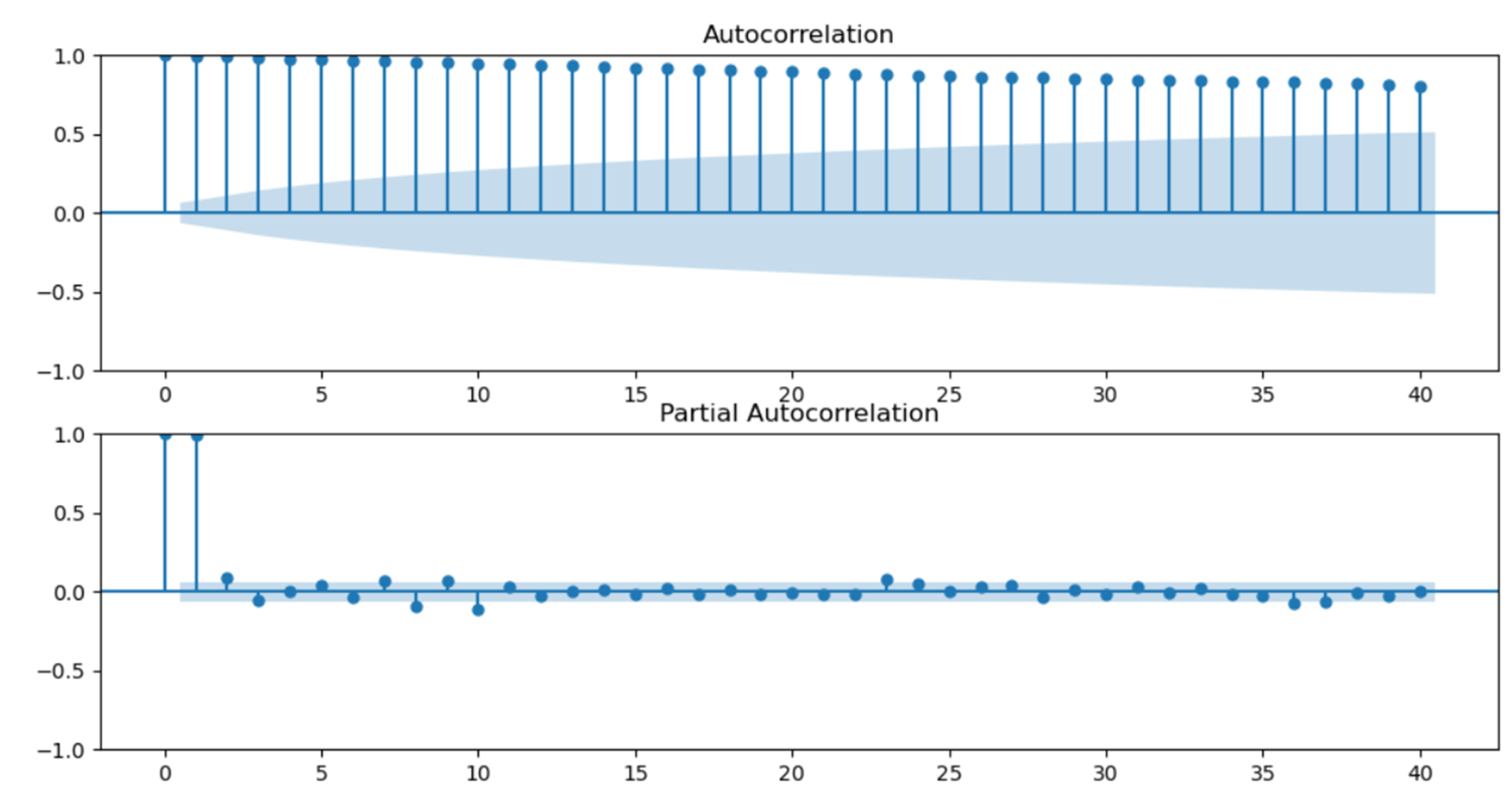
A lag plot provided a visual representation of autocorrelation within the time series.

Autocorrelation and Partial Autocorrelation Analysis:

Autocorrelation and partial autocorrelation plots were generated to assess correlations between data points at different time lags.

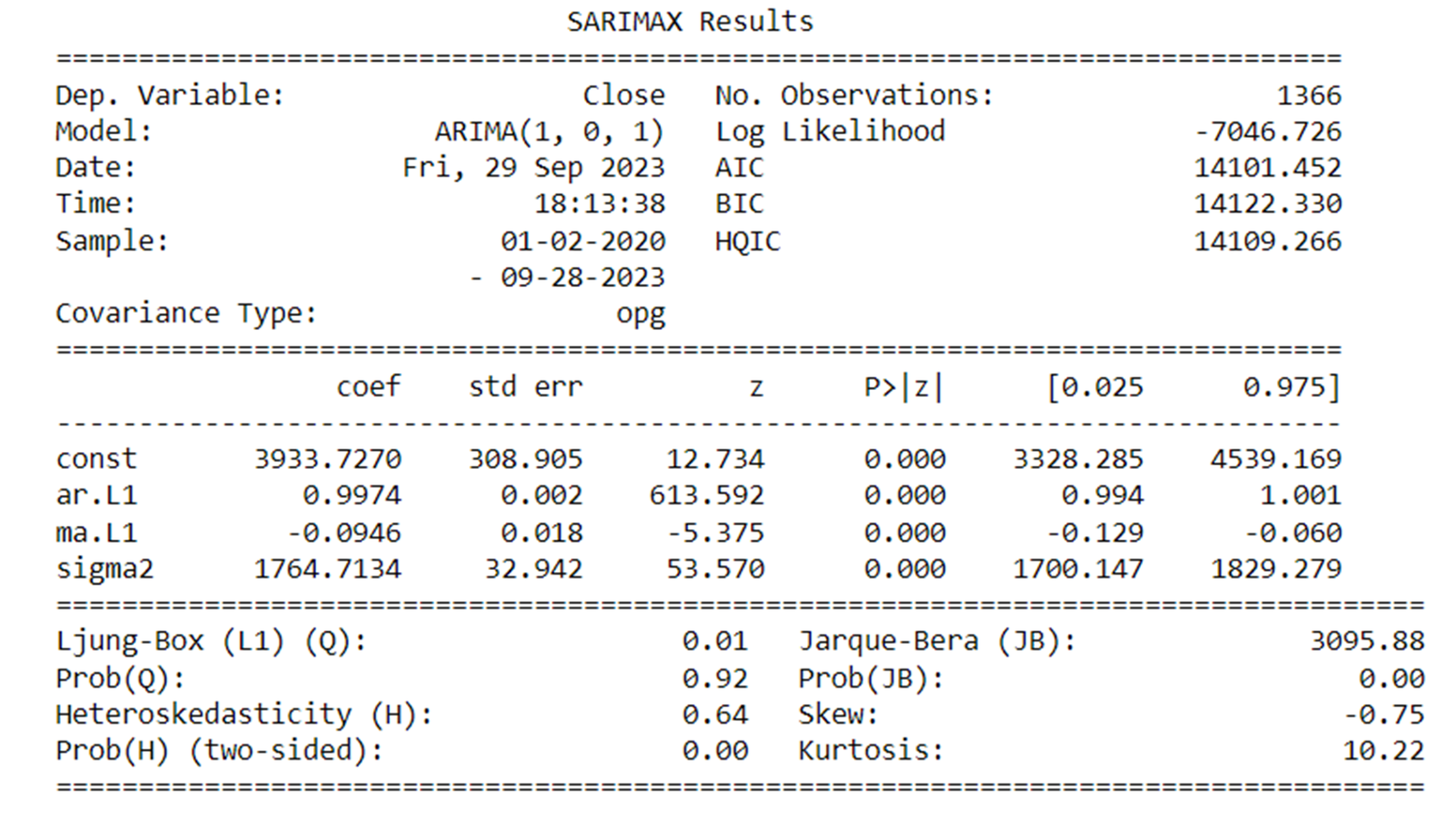
Differencing:

To achieve stationarity, we applied differencing by computing the first-order difference of the time series data, denoted as y.diff().

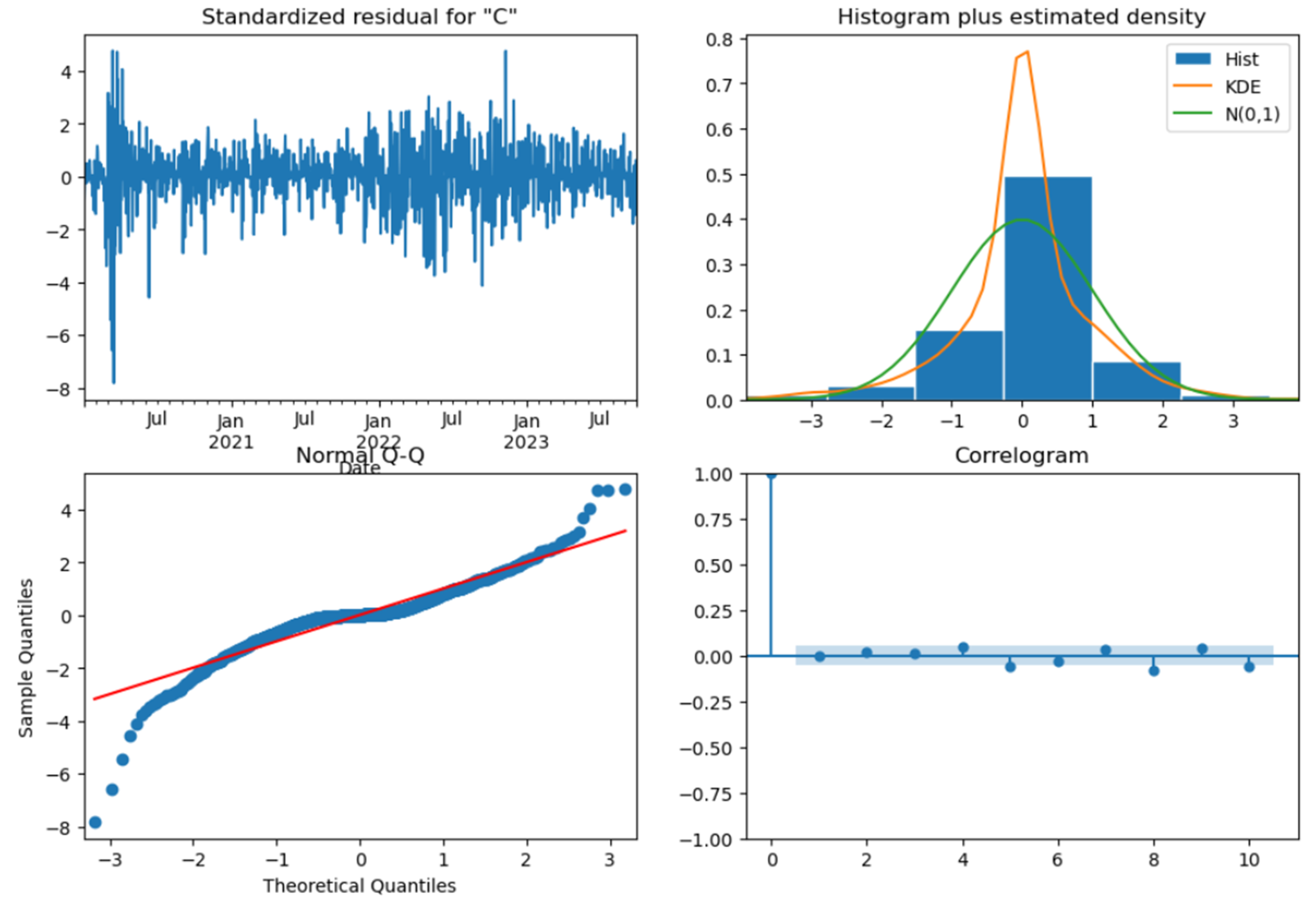


**ARIMA Modeling:**

We employed an ARIMA (AutoRegressive Integrated Moving Average) model with specific parameters (p=1, d=0, q=1) based on the differenced data.



Model fitting and summary statistics were carried out, revealing essential model information and diagnostics.



**One-Step Ahead Forecast:**

Utilizing the ARIMA model, we performed one-step ahead forecasts for the year 2023.

The forecasted values were plotted alongside observed data, complete with confidence intervals.

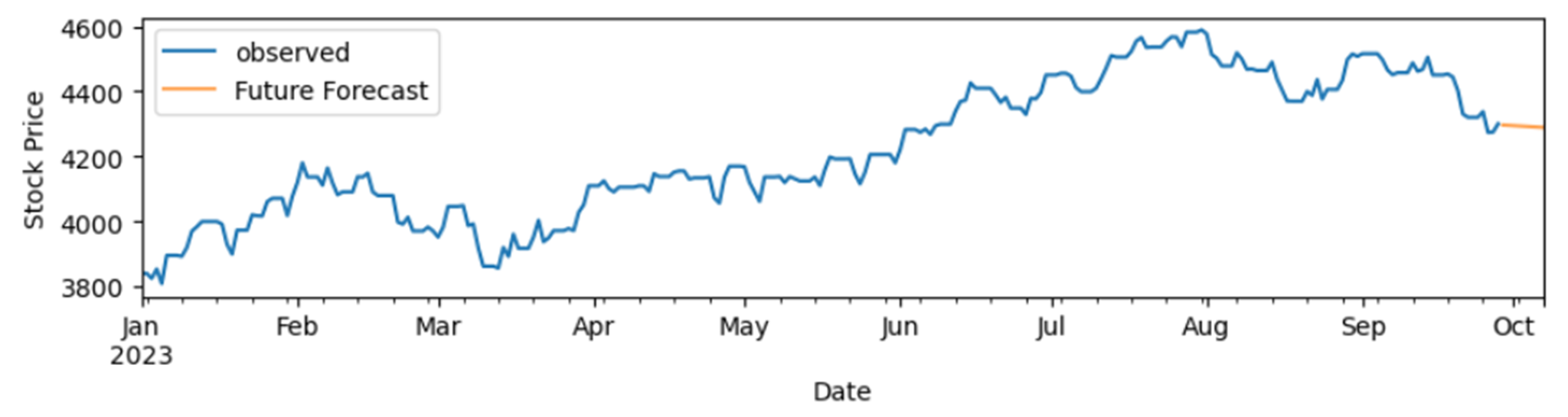
**Forecast Evaluation:**

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated to assess the accuracy of our forecasts.

The MSE was found to be 845.29, with an RMSE of 29.07.

Weekly Forecast: We generated a forecast specifically for the first week of October.

Extracted forecasted values for each day of this week were presented.



**Time Series Conclusion:**

In conclusion, this comprehensive time series analysis illuminated the stock price data's nuances, uncovering trends, seasonality, and autocorrelation patterns. The ARIMA model was instrumental in making forecasts for the upcoming year, and we evaluated its performance. We noted that the data initially lacked stationarity but became stationary after differencing. The forecasted values for the first week of October were provided, offering insights into potential future stock price trends. To enhance predictive accuracy, further model refinement and evaluation may be required.

# **Result:**

You have conducted a thorough analysis of stock price data, focusing on the S&P 500 index.

Data preprocessing and exploratory data analysis (EDA) were carried out effectively, revealing important insights into the data's characteristics.

Regression models, including Linear Regression and Ridge Regression, were developed and evaluated for predicting the next day's closing prices.

Classification models, particularly the Random Forest Classifier, were employed to predict whether the stock price would go up or down.

Time series analysis was conducted to forecast stock prices for the first week of October, utilizing an ARIMA model.

# **Conclusion:**

Your project showcases a comprehensive and systematic approach to analyzing stock price data, making it suitable for both investors and researchers.

The regression models achieved high accuracy and relatively low error rates, suggesting their effectiveness in predicting stock prices.

The Random Forest Classifier demonstrated its potential for predicting stock price trends, though there is room for further improvement.

The time series analysis successfully uncovered trends and seasonality in the data, and the ARIMA model provided forecasts for future stock prices.

Stationarity was achieved through differencing, which is crucial for time series analysis.

The project highlights the importance of data preprocessing, feature engineering, and model evaluation in financial analysis.

# **Summary:**

* In summary, your project presents a comprehensive exploration of stock price analysis through various methodologies:
* Data Analysis and Preprocessing: You meticulously collected and cleaned data from Yahoo Finance, ensuring its reliability for analysis. The EDA process provided a clear understanding of the S&P 500 index's historical performance.
* Regression Models: You employed regression models, including Linear Regression and Ridge Regression, to predict stock prices. These models exhibited high accuracy and are valuable for short-term price forecasting.
* Classification Model: The Random Forest Classifier was used to predict whether stock prices would go up or down. While it performed well, there is room for further enhancement to address potential biases.
* Time Series Analysis: Through time series analysis, you delved into the intricate patterns within the data. The ARIMA model provided forecasts for future stock prices, and stationarity was achieved through differencing.
* Conclusion and Future Directions: The project's conclusion summarizes the key findings and insights gained from your analysis. It also highlights the need for further refinement and evaluation of the models to enhance predictive accuracy.

# **References:**

* [Predict The Stock Market With Machine Learning And Python](https://www.youtube.com/watch?v=1O_BenficgE)
* <https://towardsdatascience.com/time-series-forecasting-predicting-stock-prices-using-an-arima-model-2e3b3080bd70>
* [Forecasting Future Sales Using ARIMA and SARIMAX](https://www.youtube.com/watch?v=2XGSIlgUBDI)
* https://www.geeksforgeeks.org/stock-price-prediction-using-machine-learning-in-python/