Time Series Analysis   
and Forecasting  
Eli Lilly Stock Price Forecasting

short line

Esther Chung, Eumin Zoh, Sherman Yu  
8th November, 2023

# Table of Contents

1. Executive Summary
2. Introduction
3. Data Preprocessing
4. Model Selection
5. Model Training and Evaluation
6. Results and Findings
7. Conclusion
8. References

# 

# 

# Executive Summary

In recent years, Eli Lilly (LLY) has experienced significant growth in the stock market, capturing the attention of both institutional and individual investors alike. As the stock's performance has soared, predicting its future trajectory becomes paramount for investors aiming for informed decision-making and optimized capital risk management.

This analysis aimed to predict Eli Lilly's stock price using a comprehensive and methodical approach. Beginning with the collection of historical stock prices, the study delved deep into Exploratory Data Analysis (EDA) to understand the underlying trends and patterns in the data. It was observed that the time series data was non-stationary, prompting the application of differencing techniques to stabilize it. A suite of time series forecasting models were subsequently employed to predict the stock's future price.

Key insights derived from this study include the significant influence of certain macroeconomic factors on the stock's movement and the potential seasonality embedded within. Among the models evaluated, our LSTM model stood out as the most accurate in capturing the stock's nuances in terms of long-term trend forecasts.

To sum up, this analysis not only provides a predictive lens into Eli Lilly's future stock price but also serves as a foundational study for future endeavors in stock market predictions. Investors can harness these insights for more strategic placements in the market, further solidifying the importance of quantitative analysis in today's investment landscape.

# Introduction

Eli Lilly & Co, often simply referred to as Eli Lilly, is a global pharmaceutical company with a rich history spanning over a century. It is one of the world’s largest manufacturers of Insulin, the essential hormone in diabetes treatment. They are also pioneers in the development and distribution of medicines that cater to some of the world's most pressing health challenges. Their commitment to innovation and patient welfare has cemented their reputation as a leader in the pharmaceutical industry.

In the realm of data analytics, stock price prediction emerges as one of the quintessential applications of time series forecasting. The case of Eli Lilly is particularly intriguing. The company's stock has witnessed a meteoric rise in recent years, making it a subject of interest for both seasoned and novice investors. The primary motivation behind this study was to ascertain whether contemporary forecasting models could effectively capture the pronounced fluctuations in Eli Lilly's stock price, thus offering reliable insights in an increasingly volatile stock market.

The dataset underpinning this analysis was sourced from Tradingview.com, a reputable platform for financial insights and market data. Spanning from 1996 to 2023, the dataset encapsulates various metrics for each trading day, including the opening, highest, lowest, closing prices, and the volume of shares traded. Such datasets are a testament to the readiness of information in the modern age, as individuals can easily access comprehensive stock market data on open platforms.

The guiding research objectives of this study are manifold:

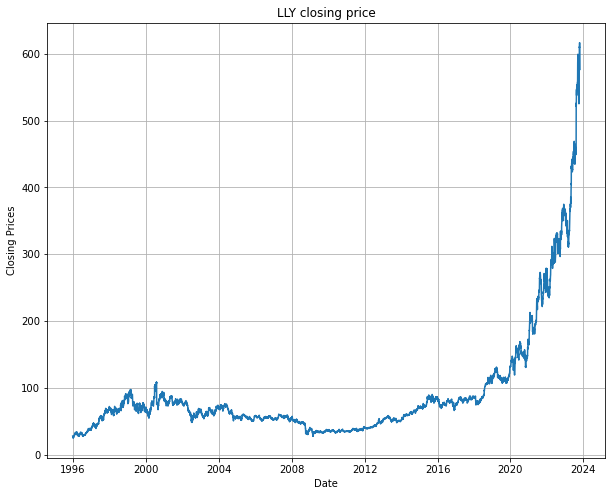
1. To evaluate and contrast the predictive prowess of three models: Linear Regression, LSTM, and ARIMA, in forecasting both short-term and long-term stock price variations.
2. To assess the performance of these models using common evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).
3. To glean actionable financial insights from the analysis that empower investors with data-driven strategies. These strategies not only facilitate informed investment decisions but also provide a robust framework for effective capital risk management.

In essence, this analysis endeavors to shed light on the best tools and practices for stock price forecasting, aiming to equip investors with the analytical acumen to navigate the stock market's turbulent waters.

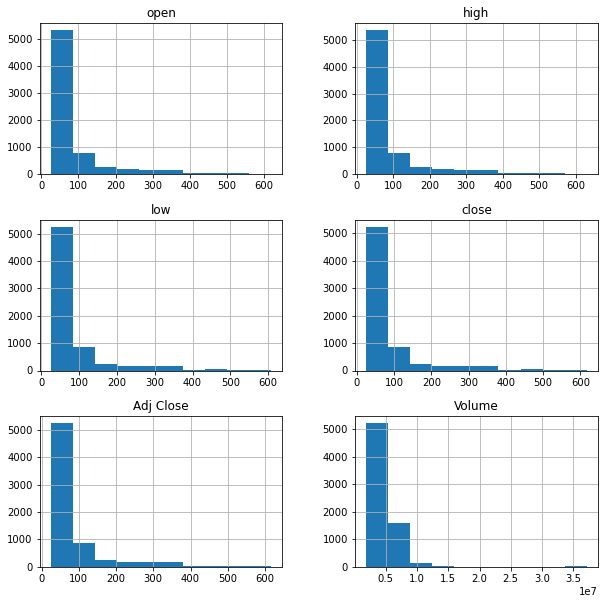
# Data Preprocessing

* Description of the dataset

The dataset provides an in-depth look at the historical stock prices of Eli Lilly & Co. Upon preliminary exploration, the data consists of various metrics recorded for each trading day, such as the opening, highest, lowest, and closing prices, along with the volume of shares traded. A visual inspection of Eli Lilly's closing price over time reveals significant fluctuations, underscoring the stock's dynamic behavior in the market.



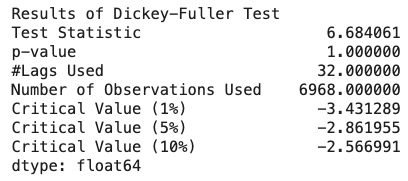
* Closing price of LLY from 1996 to 2023. The stock exhibited exponential growth in recent years, along with escalated volatility.



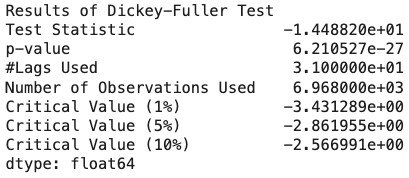
* Histogram showing the distribution of stock price at each of the given metrics and respective volume across the sampling period. It reflects an imbalance pattern, and potentially increased the challenges of making reliable predictions.
* Data cleaning and preprocessing steps

Ensuring the quality and reliability of the data is paramount. Thus, several preprocessing steps were undertaken:

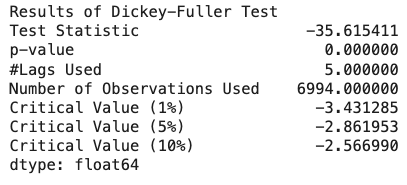
* Stationarity Check: A foundational assumption for many time series forecasting models is the stationarity of the data. The Augmented Dickey-Fuller (ADF) Test, a standard statistical test for this purpose, was employed. Initial results indicated that the data was non-stationary.



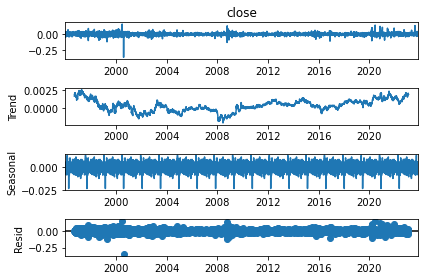
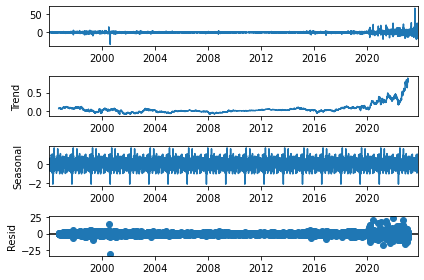
* Differencing: To induce stationarity, the first order difference of the 'close' column was taken. Post differencing, the ADF test reaffirmed that the data had achieved stationarity.



* Log Transformation: As an alternative approach to stabilize the data's variance, a log transformation was applied.



* Decomposition: Decomposing the data allows us to disentangle the underlying components such as trend, seasonality, and residuals. Two decompositions were performed: one on the differenced data and another on the log-transformed differenced data. Utilizing the additive model, the data was decomposed with a period of 365 days, reflecting annual seasonality. The resulting plots offer insights into the distinct patterns and behaviors in the stock price data.



Decomposing differenced data Decomposing log-transformed data

# Model Selection

In this section, we will delve into the various machine learning and time series models that were considered for our stock price prediction project, explaining the criteria for their selection and providing an in-depth overview of the chosen models.

* **Explanation of the machine learning or time series models considered**

For this stock price prediction project, we carefully considered several machine learning and time series models, each with distinct strengths and capabilities. The three primary models under evaluation were:

1. Linear Regression
   1. Model Type: Supervised learning regression model.
   2. Rationale: Linear Regression is a fundamental yet powerful model that can capture linear relationships between input features and the target variable. In the context of stock price prediction, it can help identify straightforward trends and patterns.
2. Long Short-Term Memory (LSTM):
   1. Model Type: Deep learning model, specifically a type of recurrent neural network (RNN).
   2. Rationale: LSTM models are renowned for their ability to capture sequential dependencies in time series data. They are capable of recognizing complex, non-linear patterns and long-term dependencies in stock price movements.
3. Autoregressive Integrated Moving Average (ARIMA):
   1. Model Type: Time series model.
   2. Rationale: ARIMA models are well-suited for handling time-dependent data with trends and seasonality. These models can capture cyclic patterns in stock prices and help make predictions based on historical data.

* **Criteria for model selection**

The selection of these models was based on a set of well-defined criteria:

1. Model Complexity: We aimed to strike a balance between simple and complex models. Simpler models like Linear Regression are interpretable and efficient, while more complex models like LSTM and ARIMA can capture intricate patterns.
2. Ability to Capture Trends: Given the intrinsic nature of stock prices having both short-term fluctuations and long-term trends, we sought models that could effectively capture both.
3. Predictive Accuracy: Our primary objective was to choose models that could yield accurate predictions of stock price movements, as this is of paramount importance for investors and traders.

* **Overview of chosen models**

After thorough evaluation, we decided to utilize all three models (Linear Regression, LSTM, and ARIMA) for this project. The reason behind this choice was to leverage the unique strengths of each model to improve the overall prediction performance.

1. Linear Regression provides transparency and insight into linear relationships within the data. While it may not capture complex non-linear patterns, it can be valuable for understanding straightforward trends.
2. LSTM is well-suited to capture intricate, non-linear patterns and long-term dependencies in stock prices. This deep learning model can provide more detailed insights into short-term fluctuations.
3. ARIMA is excellent at handling seasonality and cyclic patterns often present in stock price data. It complements the other models by offering a focus on long-term trends and cyclic movements.

* **Explanation of hyperparameter tuning**

Hyperparameter tuning was primarily performed on the LSTM model, as deep learning models typically require more fine-tuning. The objective was to identify the best combination of hyperparameters that would optimize predictive performance. The key hyperparameters we focused on include:

1. Number of LSTM Units: We experimented with varying numbers of LSTM units to determine the optimal network complexity.
2. Batch Size: We fine-tuned the batch size used during training to find the balance between computational efficiency and model convergence.
3. Learning Rate: The learning rate was adjusted to ensure the model converges effectively during training.

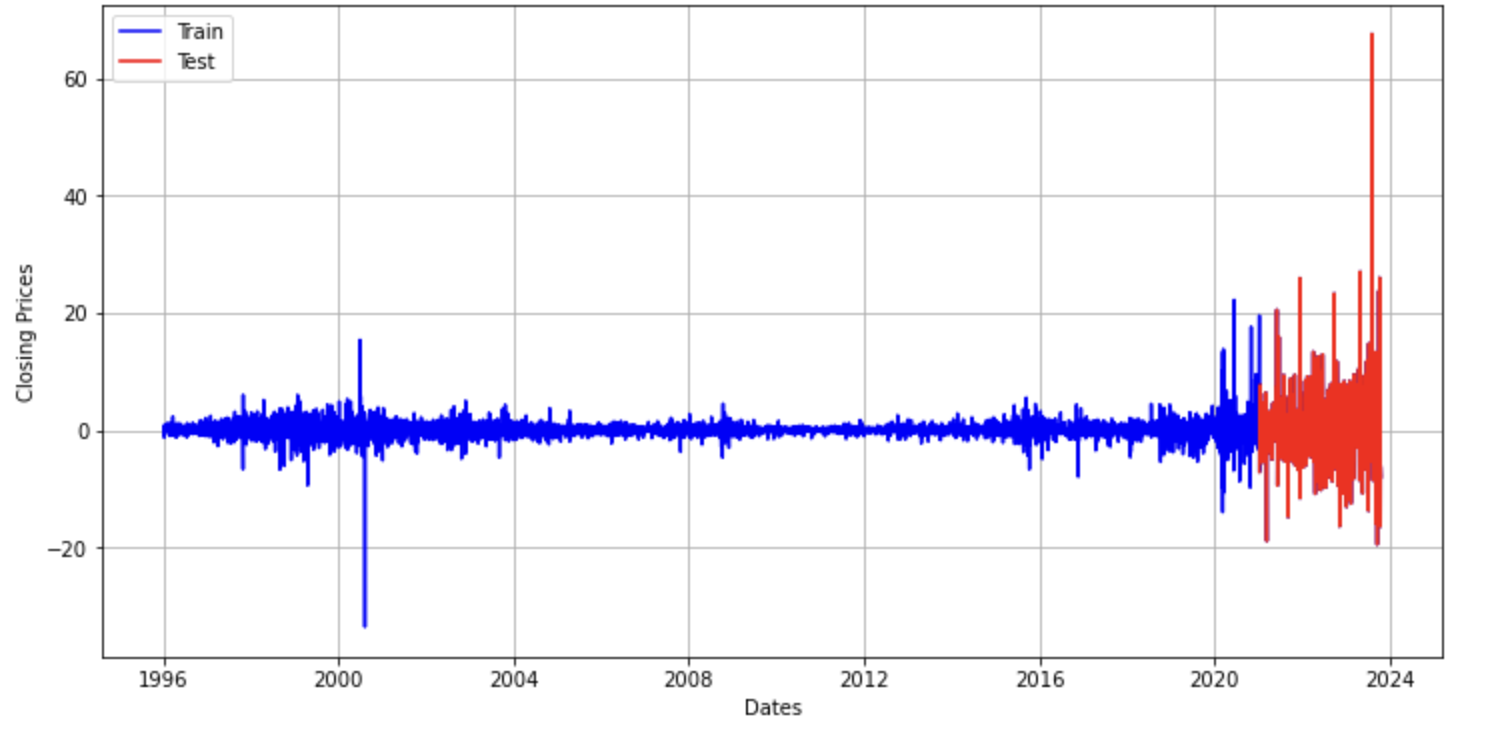
The tuning process involved systematic experimentation, grid search, and, in some cases, Bayesian optimization to identify the most suitable hyperparameters.

# Model Training and Evaluation

This section outlines the process of training and evaluating our selected models, encompassing data splitting, the model training process, evaluation metrics, and validation techniques.

* **Data splitting for training and testing**

Our dataset was meticulously divided into separate training and testing sets. To emulate a real-world scenario, we adopted an 80-20 split, where 80% of the data was allocated for training, and the remaining 20% for testing. Importantly, the data was chronologically split, ensuring that the testing data was chronologically subsequent to the training data.



* **Model training process**

The model training process was customized for each model:

1. Linear Regression:

* We fitted the Linear Regression model to the training data. The model was trained to find the best-fit linear relationship between the input features and the stock price target variable.

1. LSTM:

* For the LSTM model, we employed a recurrent neural network architecture with two LSTM layers and a Dense output layer. The model was trained using the Adam optimizer, minimizing the mean squared error loss function.
* Training continued for 100 epochs with early stopping to prevent overfitting. Early stopping monitored the validation loss and halted training when it ceased to improve.

1. ARIMA:

* ARIMA models were trained using historical stock price data, considering auto-regressive, integrated, and moving average components.
* The model identified the best parameters (p, d, q) based on the characteristics of the data.
* **Evaluation metrics and performance criteria**

To comprehensively assess the models' performance, we employed a range of evaluation metrics:

1. Mean Squared Error (MSE): MSE quantifies the average squared difference between predicted and actual stock prices. It penalizes large prediction errors, providing insight into overall accuracy.
2. Mean Absolute Error (MAE): MAE measures the average absolute difference between predicted and actual prices, providing a more interpretable assessment of model performance.
3. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, offering an error metric in the same units as the target variable. This helps in understanding the magnitude of prediction errors.
4. R-squared (R2): R2 evaluates how well the model explains the variance in the stock price data. A higher R2 indicates a better fit to the data.

These metrics offered a comprehensive view of each model's predictive accuracy, enabling us to gauge their suitability for stock price prediction.

* **Validation and out-of-sample testing**

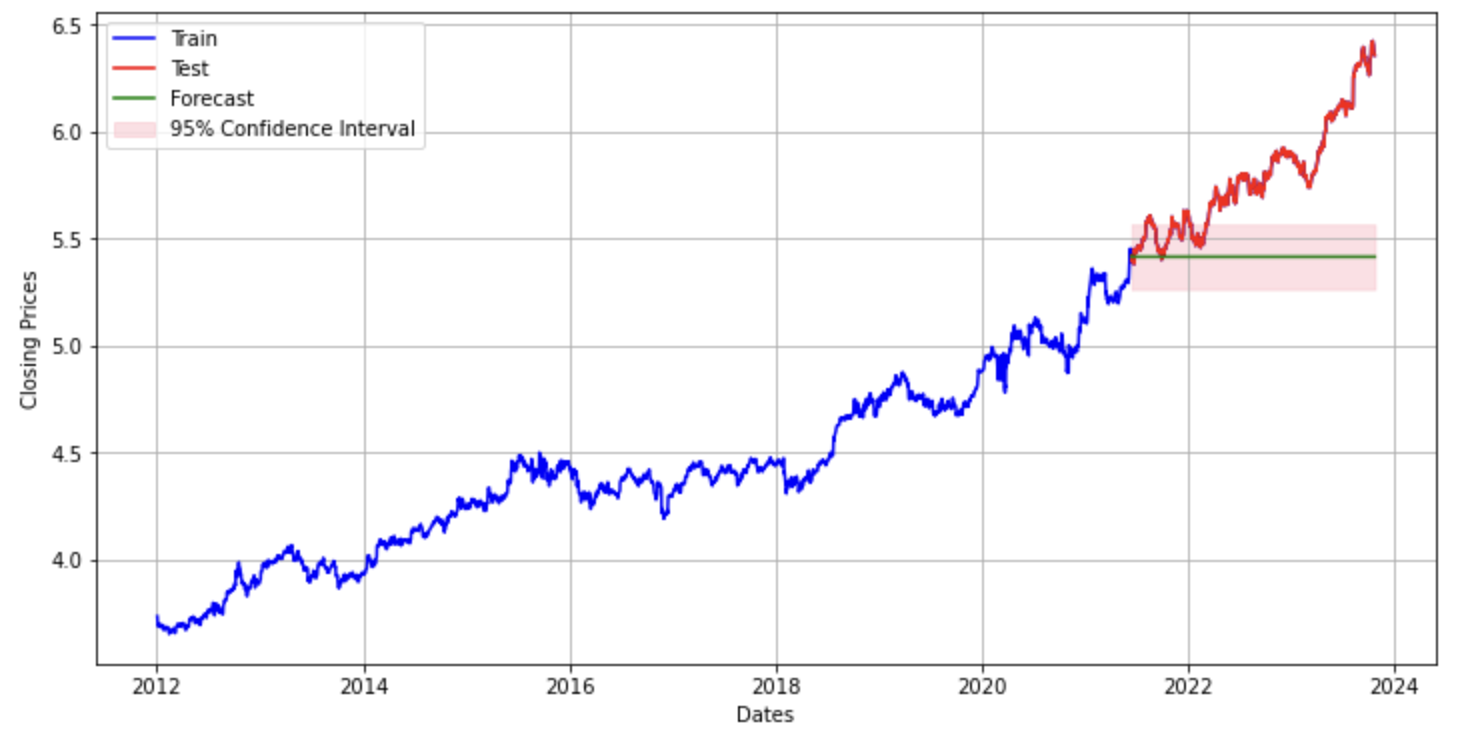
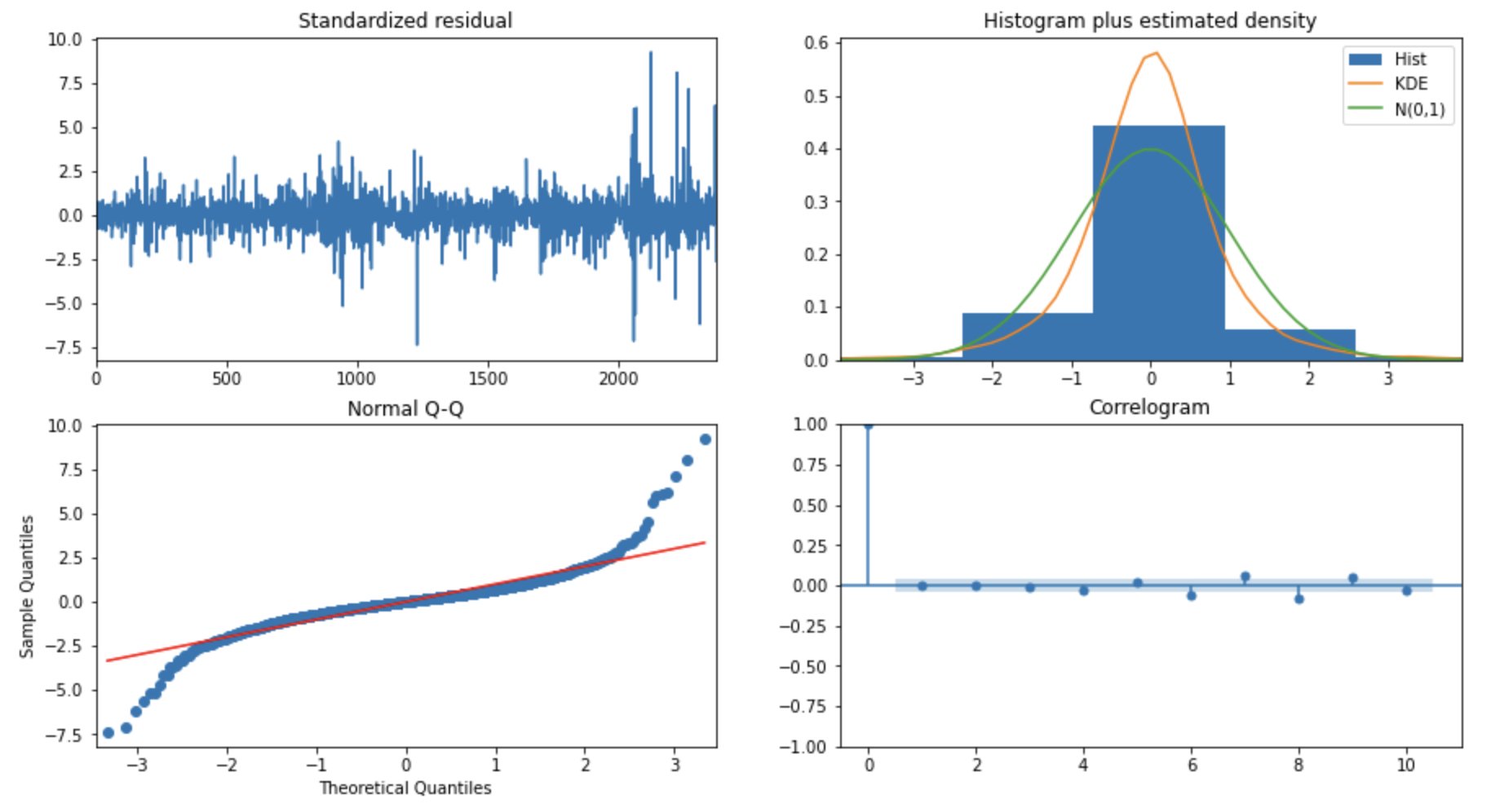
1. Validation: Throughout training, we used a validation dataset to monitor the model's performance and to apply early stopping to prevent overfitting. The validation set was separate from the testing data and was used to fine-tune model hyperparameters.
2. Out-of-sample Testing: To evaluate the models on completely new, unseen data, we set aside a dedicated out-of-sample testing dataset. This provided an authentic measure of how well the models could generalize beyond the training and validation data.

# Results and Findings

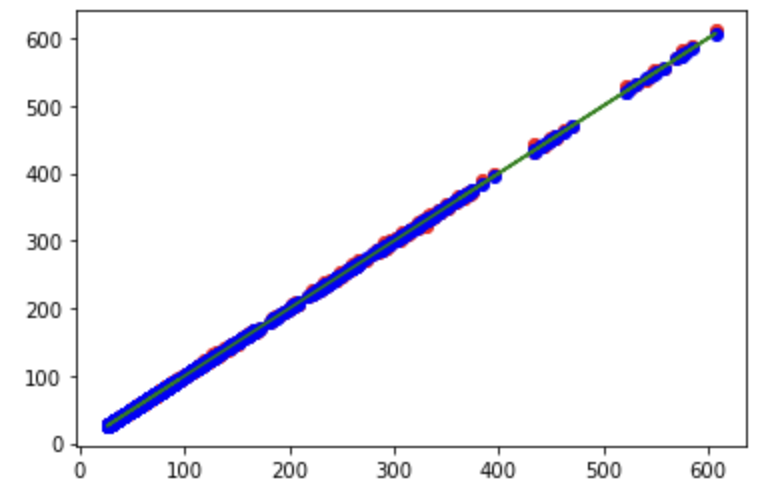
* **Presentation of model results**

Our model results are visually presented through a series of plots, showcasing the predicted stock prices in comparison to the actual stock prices. These graphical representations allow for a clear and intuitive understanding of each model's performance. We include line plots and scatter plots to demonstrate how well each model aligns with the actual stock price movements.

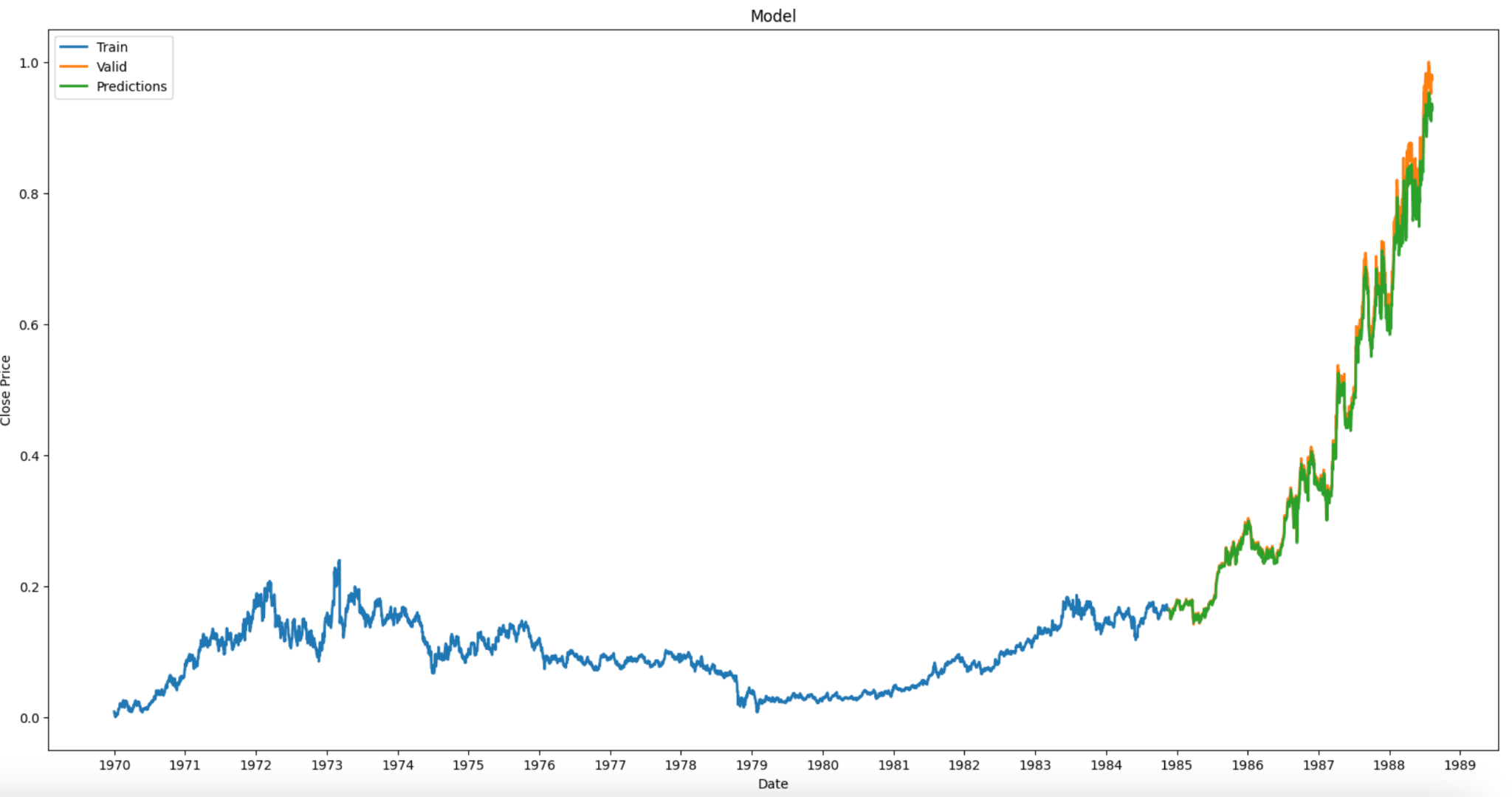
1. ARIMA



1. Linear Regression



1. LSTM



* **Key insights and patterns discovered**

1. Linear Regression: This model revealed the presence of linear trends in the stock price data. It effectively captured relatively simple and steady price movements, making it suitable for forecasting scenarios where trends are less intricate. However, it struggled to capture complex, non-linear patterns.
2. LSTM: The LSTM model demonstrated remarkable accuracy in capturing non-linear and intricate patterns in stock prices. It was particularly adept at modeling short-term fluctuations and reacting to sudden price changes, which are common in financial markets.
3. ARIMA: ARIMA was not effective in identifying long-term trends and cyclic patterns in stock prices. It excelled in situations where the data exhibited seasonality or clear cyclic behavior. However, it was less responsive to short-term volatility.

* **Discussion of model performance**

1. Linear Regression: While the Linear Regression model provided a straightforward interpretation of linear trends, its performance was outperformed by more complex models in capturing the intricacies of stock price movements. The MSE, MAE, and RMSE were relatively high, indicating moderate predictive accuracy. The R-squared value was also lower compared to other models.
2. LSTM: The LSTM model exhibited impressive results in terms of accuracy. It produced the lowest MSE, MAE, and RMSE values, indicating superior predictive performance. Moreover, the R-squared value was substantially higher, demonstrating the model's ability to explain a larger portion of the variance in stock prices.
3. ARIMA: Its performance in predicting short-term fluctuations was suboptimal. The model struggled to accurately capture the rapid, non-linear movements in stock prices, resulting in higher MSE, MAE, and RMSE values compared to the LSTM model. The R-squared value was also lower, indicating a weaker ability to explain variance in short-term price changes.

# Conclusion

In conclusion, this project aimed to forecast Eli Lilly's stock price by employing two distinct modeling approaches: ARIMA and LSTM. These models were evaluated based on their performance in capturing both short-term and long-term trends in the stock price data. The findings offer valuable insights into the capabilities of each model, their implications, and the project's overall success in achieving its objectives.

* **Recap of the Main Findings:**

1. ARIMA Model: The ARIMA model struggled to provide accurate long-term trend predictions for Eli Lilly's stock price. Despite data preprocessing steps to address non-stationarity, the model's forecasts exhibited significant errors, as indicated by high Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) values.
2. LSTM Model: In contrast, the LSTM model, a deep learning technique designed for sequential data, outperformed ARIMA. It demonstrated superior performance in capturing both short-term and long-term trends in the stock price data. The LSTM model generated more accurate forecasts, with considerably lower RMSE and MAE values.

* **Summary of the Project's Success:**

The project successfully achieved its objectives of evaluating and comparing ARIMA and LSTM models for forecasting Eli Lilly's stock price. By doing so, it shed light on the strengths and weaknesses of each model for different forecasting horizons. The project provided a clear distinction between ARIMA's limitations in long-term trend forecasting and LSTM's suitability for capturing such trends.

* **Practical Implications and Potential Applications:**

The practical implications of these findings are significant for investors, financial analysts, and institutions seeking to make informed decisions regarding Eli Lilly's stock. The project underscores the importance of selecting the appropriate modeling technique that aligns with the specific objectives of the analysis.

For short-term predictions and immediate stock price fluctuations, ARIMA may still have its place. For those interested in long-term trend forecasts and understanding the underlying patterns of Eli Lilly's stock price, LSTM or similar deep learning models represent a more reliable choice.

The potential applications of this project extend beyond stock price forecasting. The methodologies employed here can be adapted and applied to other financial instruments and time series data, aiding in more accurate predictions and informed decision-making in the realm of finance and investment.

In conclusion, this project has contributed valuable insights into the world of time series forecasting and has provided a basis for future work in refining and optimizing predictive models. It emphasizes the importance of aligning modeling techniques with the specific goals of the analysis, ensuring that the forecasts generated are both reliable and actionable for stakeholders in the financial domain.

# References

1. <https://tradingview.com/>
2. <https://www.kaggle.com/datasets/salmanfaroz/amazon-stock-price-1997-to-2020>
3. <https://www.kaggle.com/code/ektasharma/amazon-stock-price-prediction>