Data Mining and Time Series - Group 03

S&P500 Time Series Forecasting

Nicola Cecere | Francesco Mattioli | Luca Petracca

The Goal

Time Series forecasting models are a powerful tool to help the decision-making process

Accurate predictions are crucial for informed investment decisions

We are aiming to predict the next day 'Close' price for a given stock, based on the past 5 days

Data Preparation

1

2

3

Visualization

- Close price over time
- Feature Relationships
- Correlation Matrix

Exploration

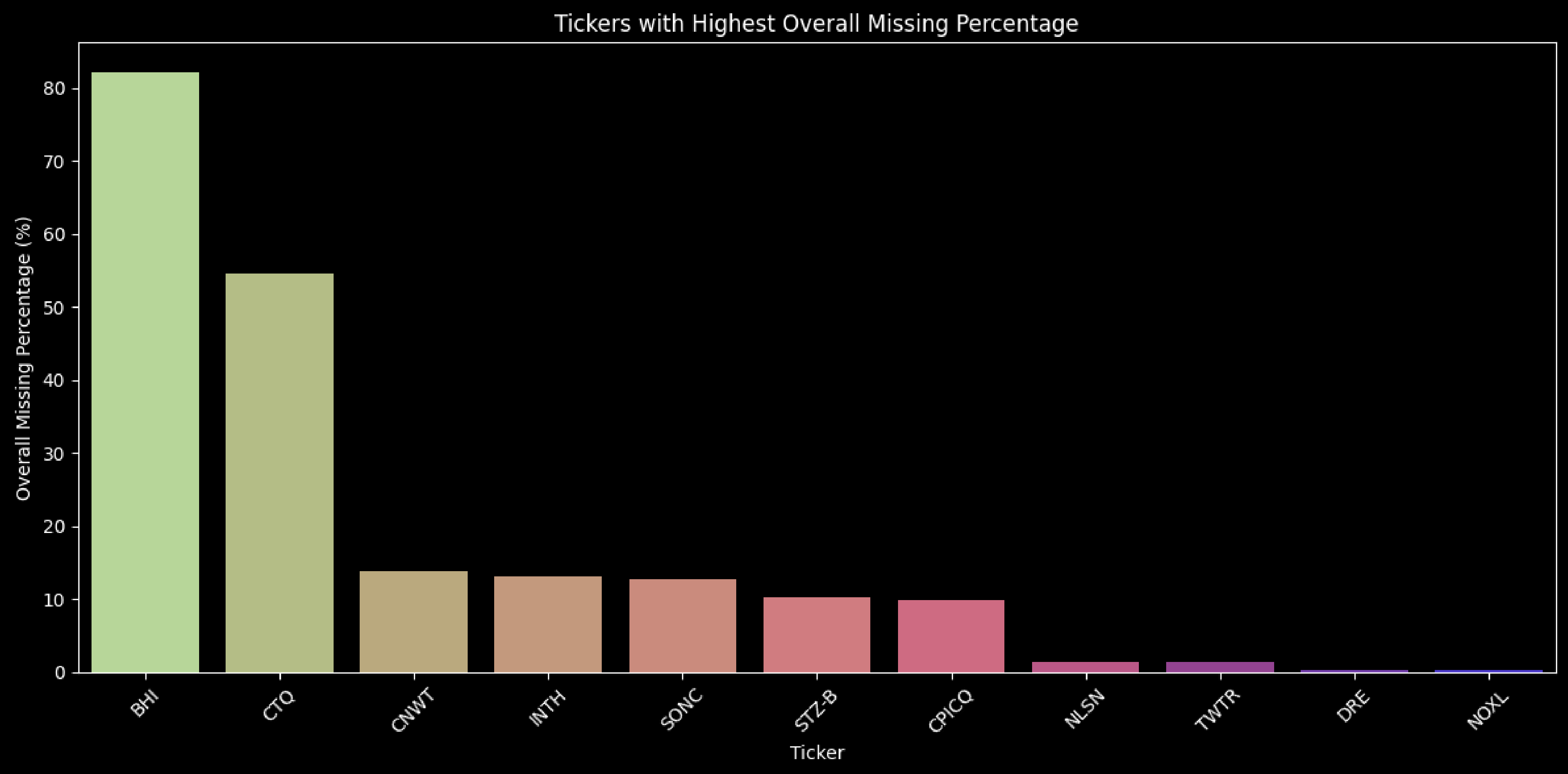
- Rows per year
- Missing values
- Outlier analysis

Preparation

- Data cleaning
- Data splitting (80-10-10)

Data Preparation

Missing Values Handling



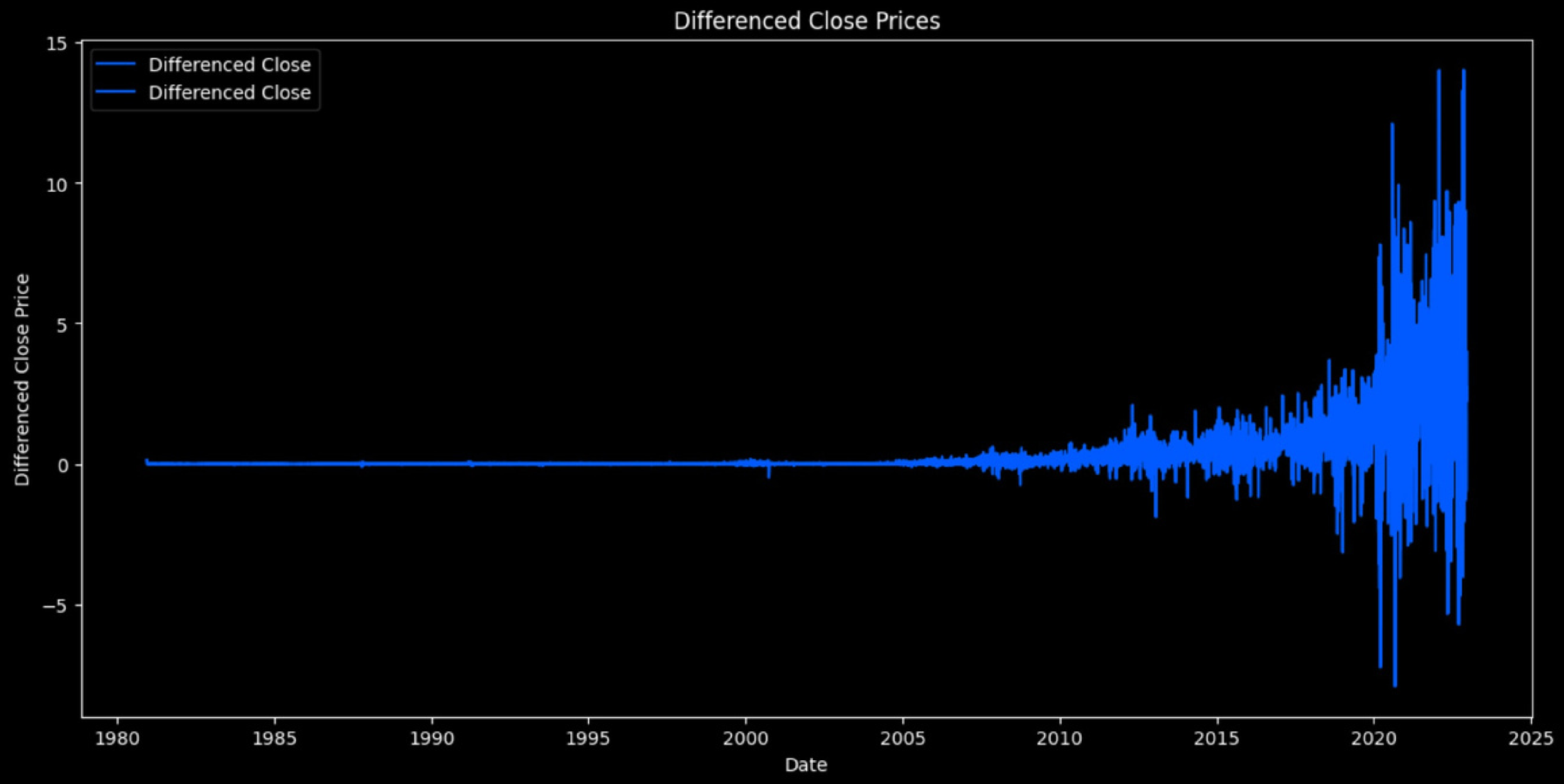
We identified the ticker with the highest number of missing values.

We handled the missing values by eliminating tickers with > 80% of missing values.

We handled the other missing values by simply eliminating the records.

Data Preparation

Stationarity Analysis



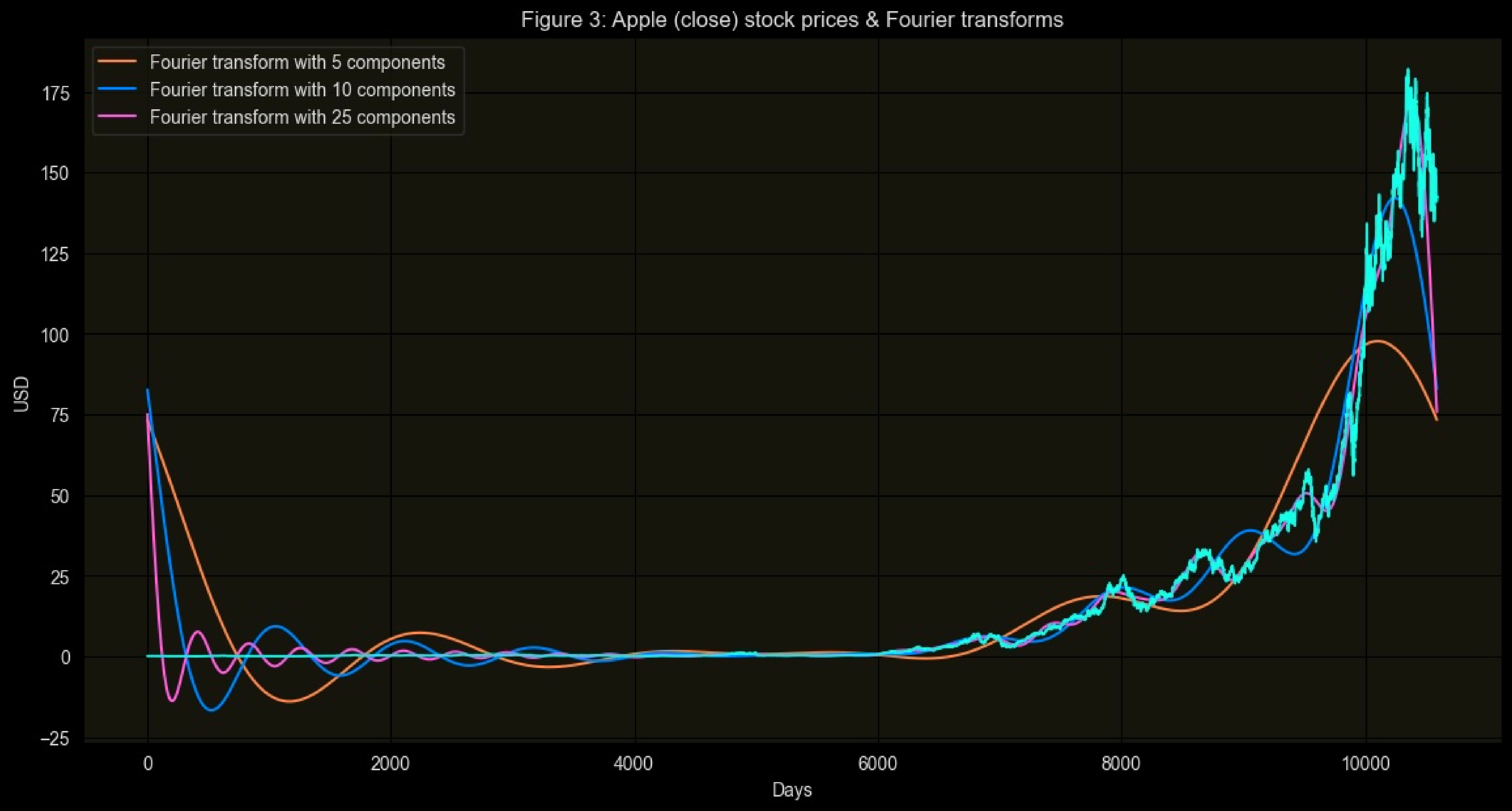
We applied ADF and KPSS testing to verify the stationarity.

Our time series resulted not stationary.

We applied Fractional Differencing to make the time series stationary

Data Preparation

Fourier Transformation

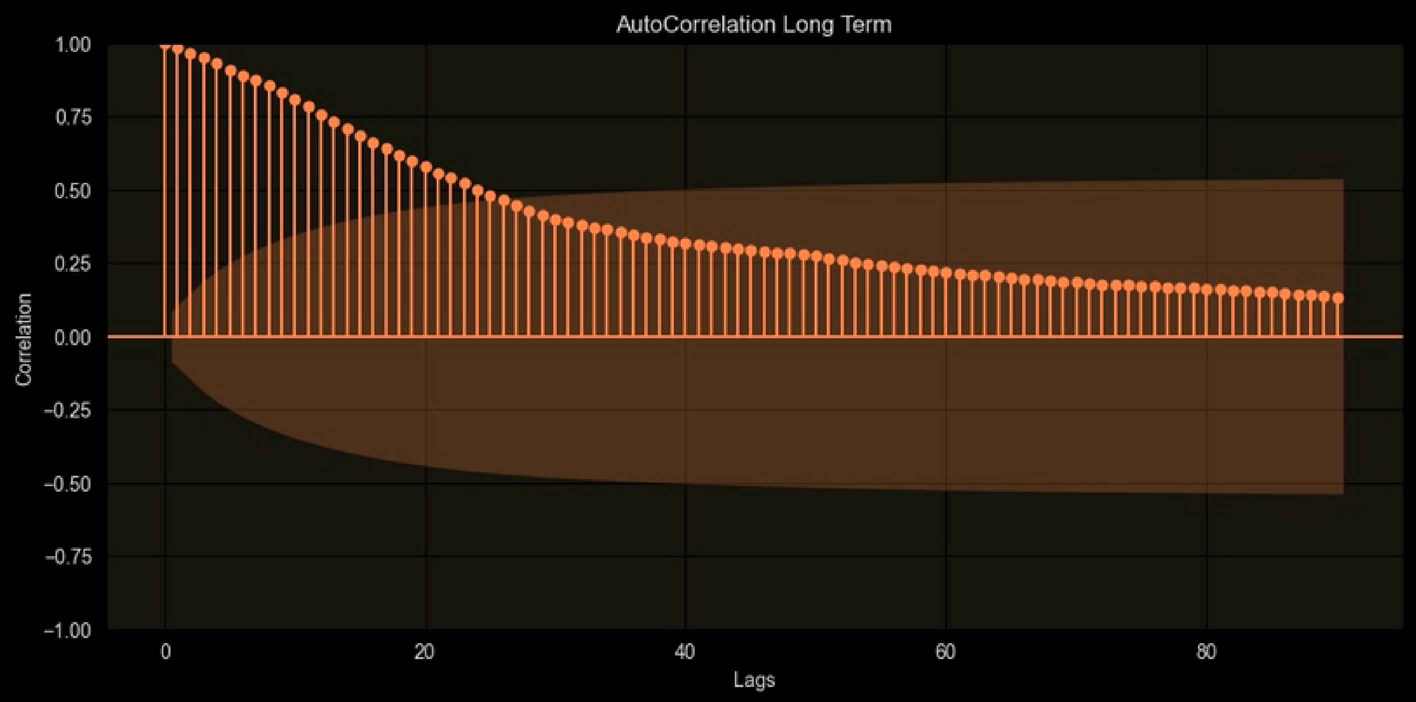


Transforming the closing price data in the frequency domain using the Fast Fourier Transform (FFT) method

We compared 5-10-25 components to approximate the stock's price and reveal underlying patterns and trends in the data

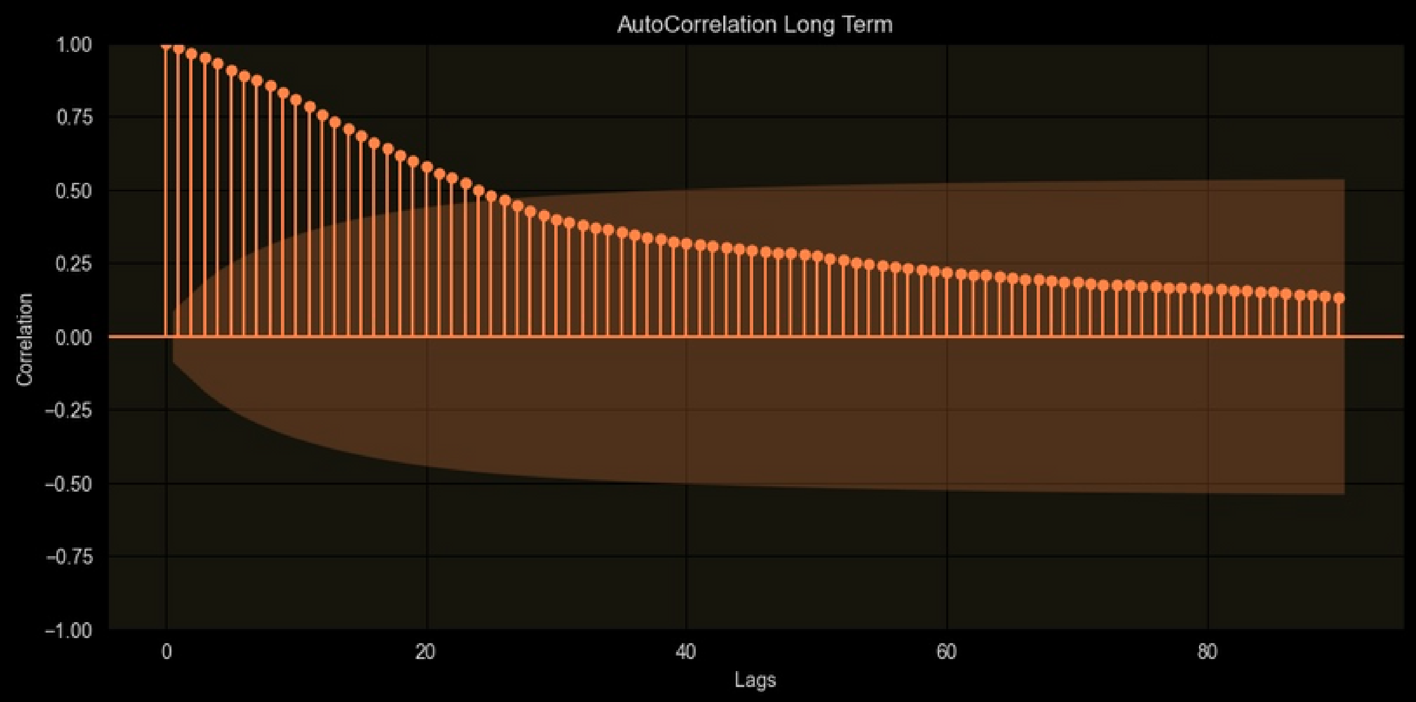
Data Preparation

Autocorrelation



- The autocorrelation plots demonstrate a gradual decay in correlation as the number of lags increases, indicating a slow decline in the relationship over time.

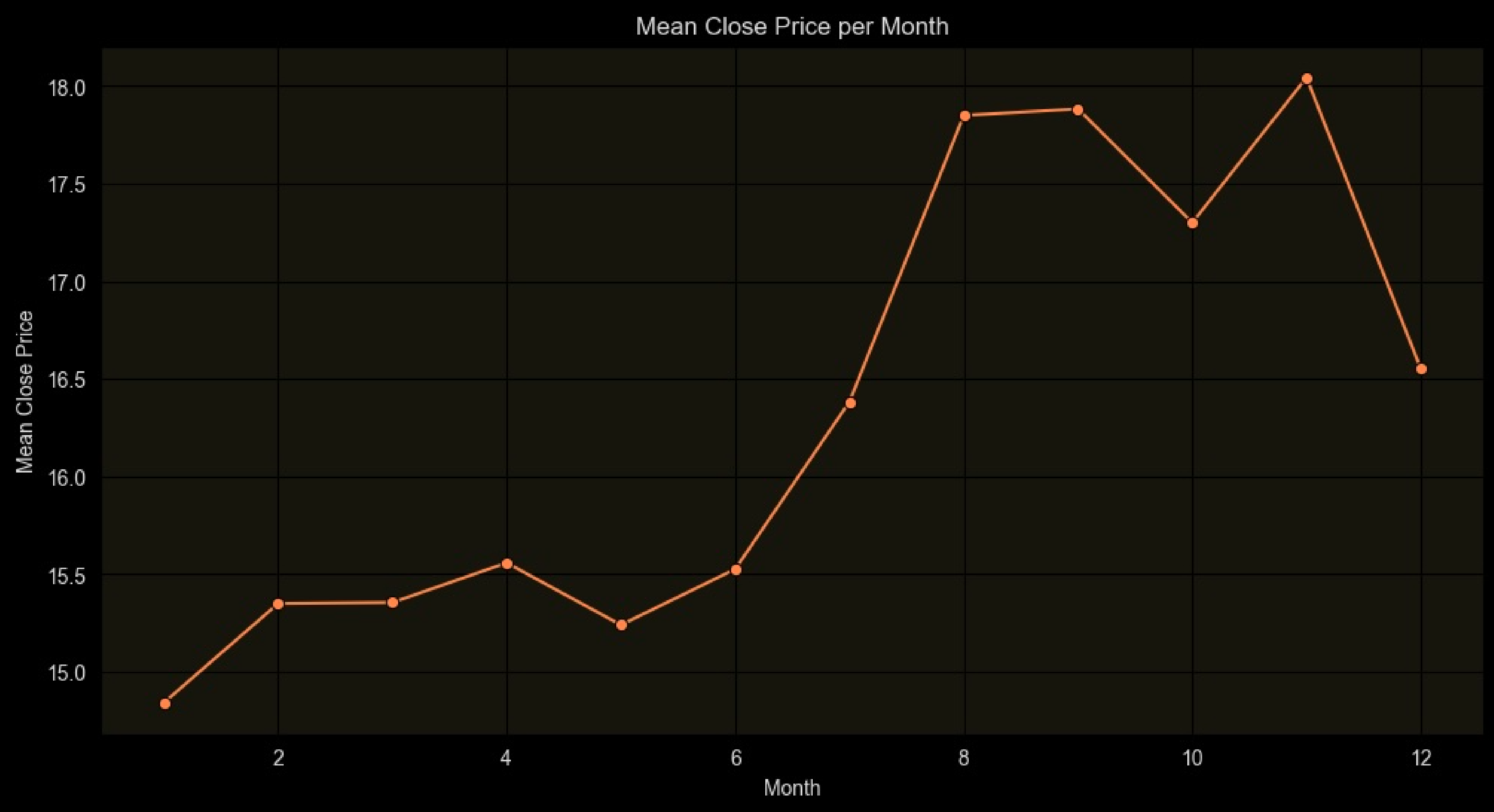
Partial Autocorrelation



- The partial autocorrelation plots suggest a potential AR(1) process where only the immediate past value has a direct impact on the current value.

Feature Engineering

1. Encoding Seasonality



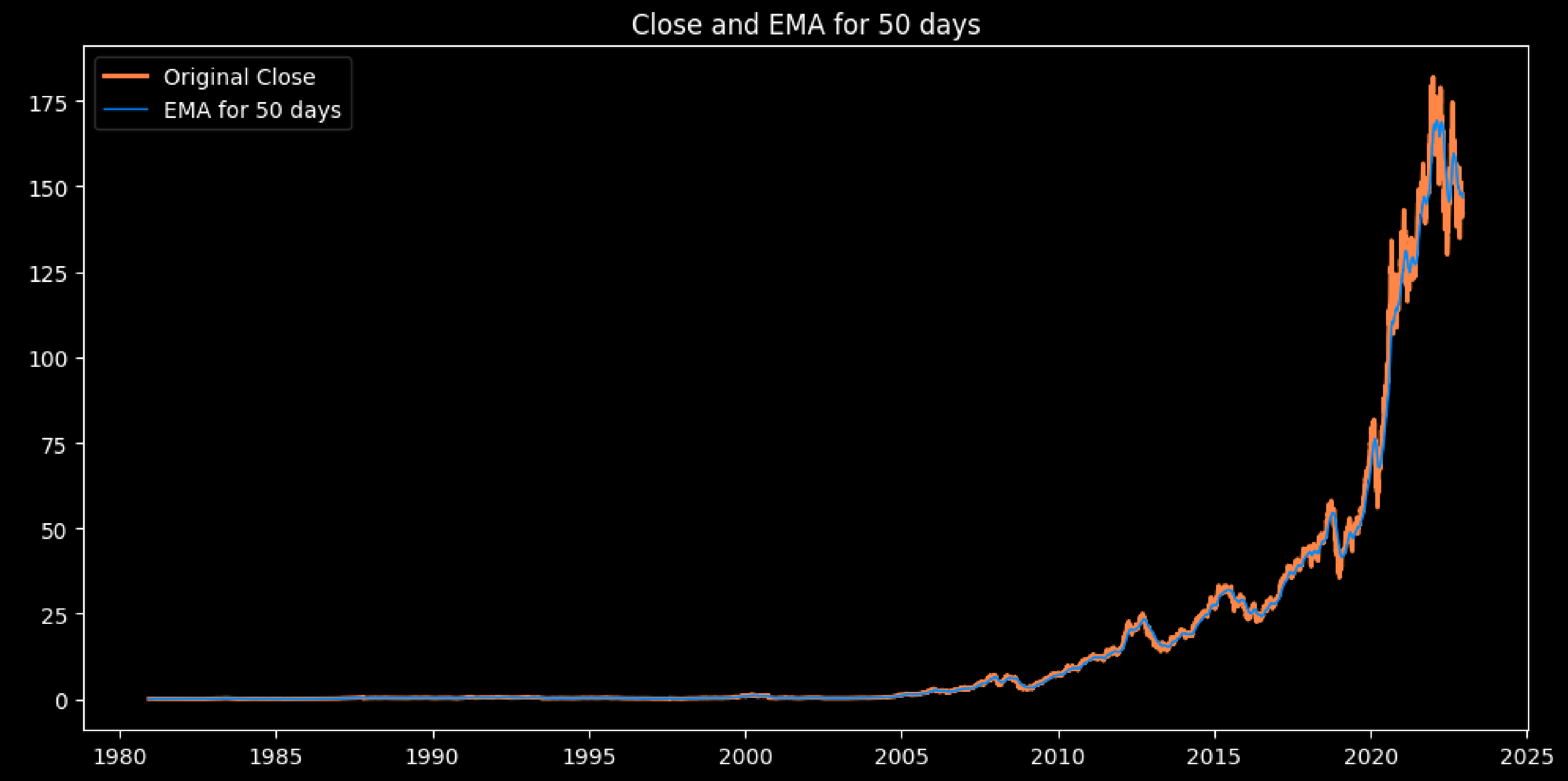
Applying trigonometric transformations to the day of the week to capture cyclical behavior

One-hot encoding the month categories in 'Bullish' , 'Bearish' and 'Normal' based on mean close price per month

Three Lag features represent the 'Close' prices of the stock from the previous three days

Feature Engineering

2. Moving Average



We explored the use of both Simple Moving Average (SMA) and Exponentially Weighted Moving Average (EWMA) in financial forecasting

We utilized the ROC, a momentum indicator, to analyze trends based on a 50-day moving average

The categorical encoded EWMA-based ROC features were used in our forecasting model

Feature Engineering

3. Clustering

DBSCAN

K-MEANS

We explored clustering using DBSCAN and K-Means. DBSCAN is known for its ability to form clusters of arbitrary shapes, while K-Means is effective in partitioning data into K distinct, non-overlapping subgroups.

For both DBSCAN and K-Means, we conducted extensive hyperparameter tuning using the Optuna framework, with over 200 iterations

K-Means emerged as the superior algorithm, achieving a higher Silhouette Score of 0.49

LSTM Model

Model Structure

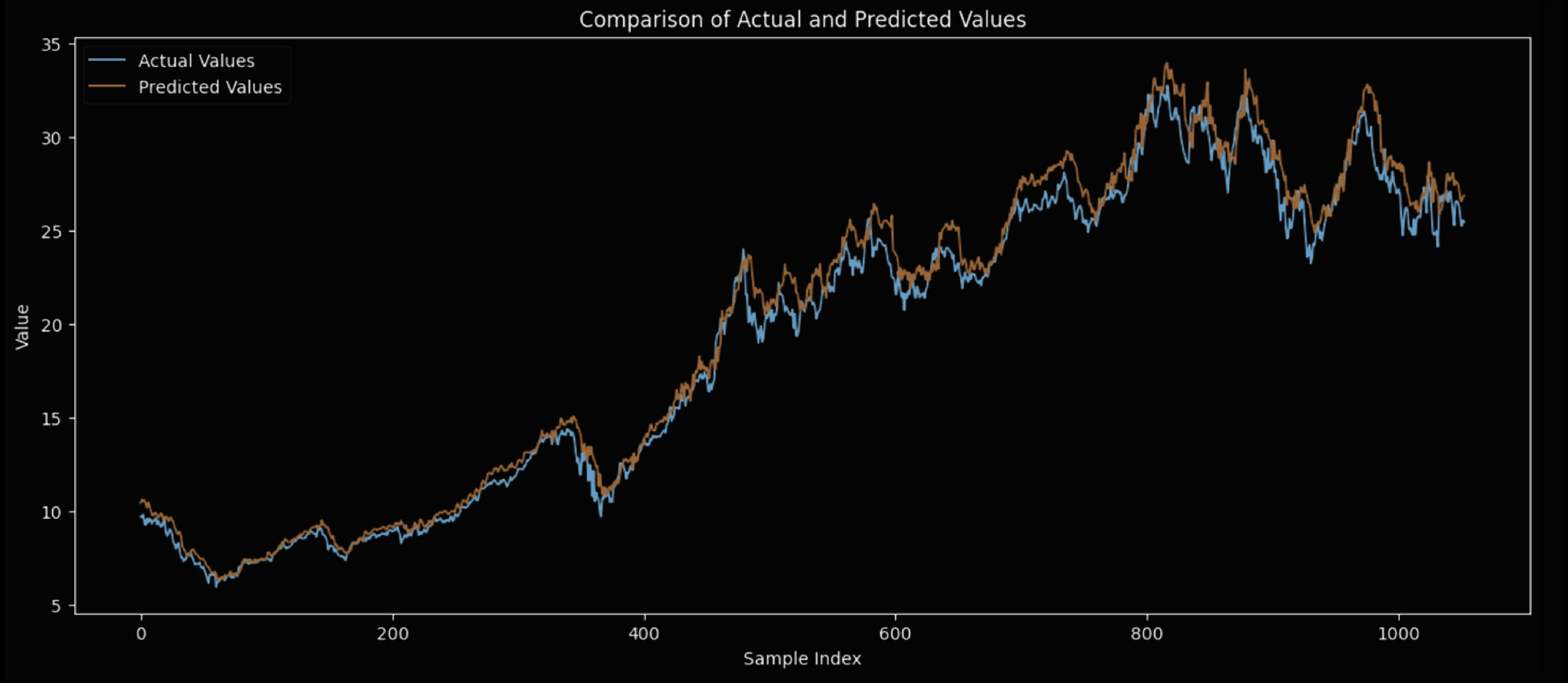
Layer Type	Output Shape	Total Params (4433)
Input (InputLayer)	[(None, 5, 19)]	0
lstm_24 (LSTM)	(None, 5, 16)	2304
lstm_25 (LSTM)	(None, 16)	2112
dense_12 (Dense)	(None, 1)	17

LSTM Model

Decay of Validation Loss

LSTM Model

Performance on Test Set



Evaluation

Epochs: 50

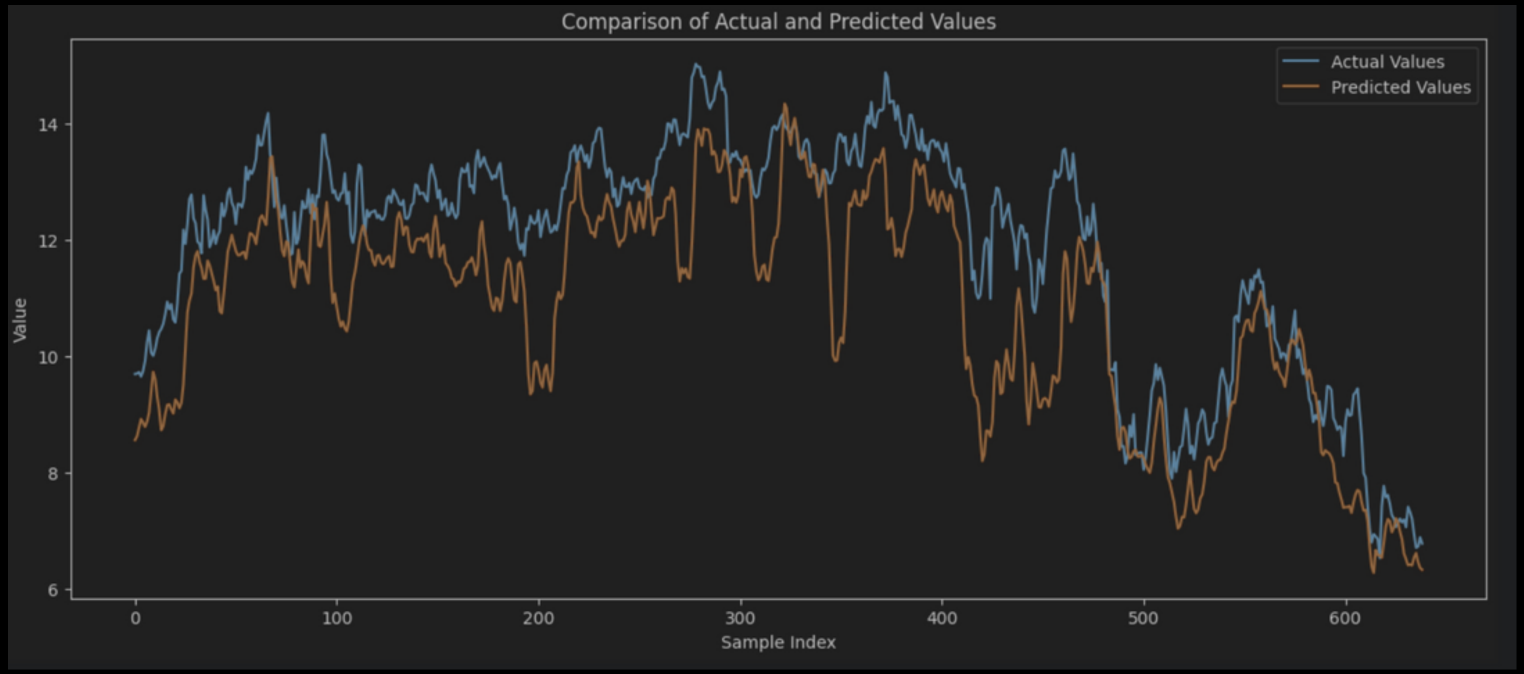
Loss: 1.2495

MAE: 0.8377

Model Comparison

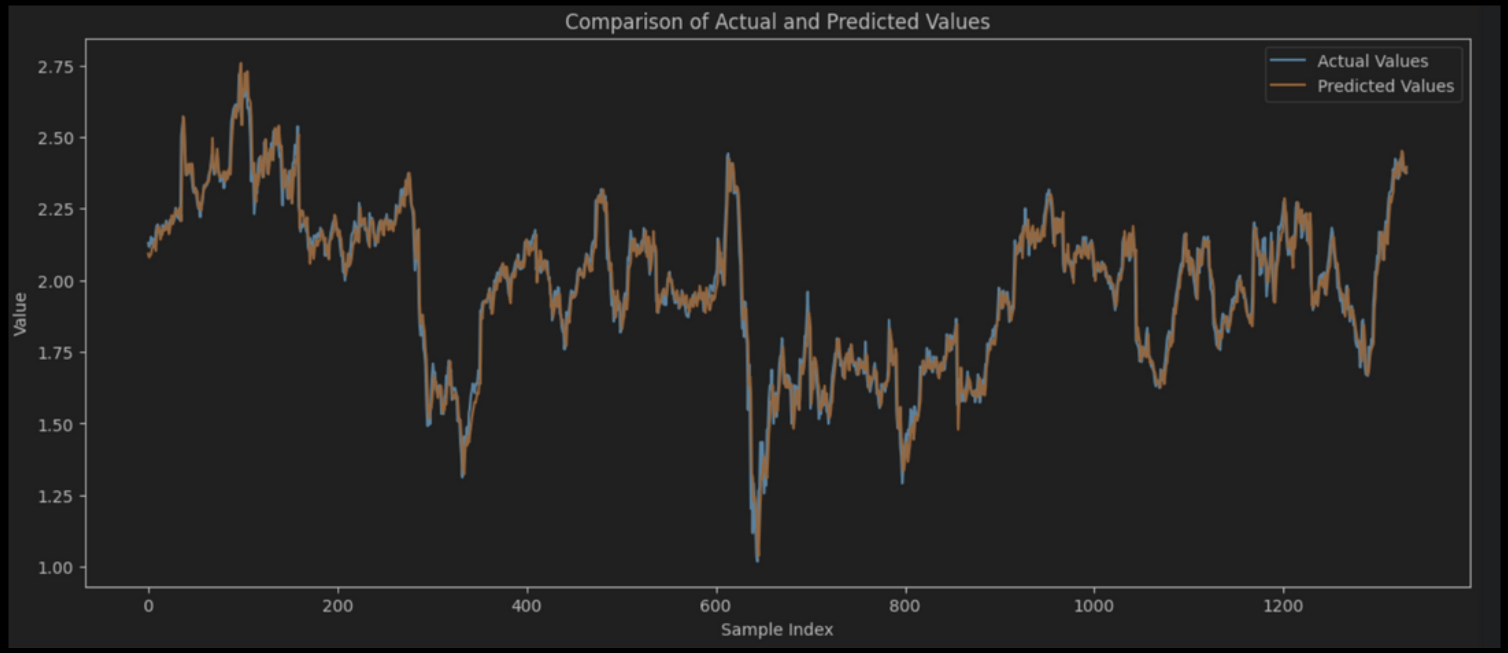
These results on both Amazon and IBM stocks reinforce the adaptability of the LSTM model. Despite being built analyzing data from Apple, it shows promise in its ability to understand and predict stock market dynamics for other companies in the IT sector.

Performance on AMZN Stock



- Loss: 1.9337
- Mae: 1.1223

Performance on IBM Stock



- Loss: 0.0033
- Mae: 0.403

Group 03

Thank you!
Q&A?

Nicola Cecere | Francesco Mattioli | Luca Petracca