Stock price forecasting: A machine learning approach

Sergio Hörtner

Springboard

Introduction

Importance of stock price forecasting:

- Stock price forecasting is crucial in financial analysis. It guides investment decisions, risk management, and portfolio optimization by predicting future market trends.
- Accurate forecasts help investors and analysts in strategizing and maximizing returns while mitigating financial risks.

Challenges: complex financial data, need for sophisticated analytical models, inherent unpredictability and volatility of financial markets.

Goal of our project: to explore the capabilities of machine learning and deep learning models for stock price forecasting.

Machine learning for stock price forecasting

Why machine learning?

- Traditional methods for stock price forecasting often struggle with the dynamic nature of markets, the vast amount of data and complex non-linear patterns in financial data.
- Machine learning (ML) offers advanced computational techniques capable of analyzing vast datasets, learning from market trends, and uncovering non-linear patterns.
- In a data-driven world, as financial markets become more complex and interconnected, the need for advanced, scalable, and efficient analytical methods becomes more pronounced.

Machine learning for stock price forecasting

Expected Advantages:

- Enhanced predictive accuracy by processing high-dimensional data.
- Ability to adapt to new data, making forecasts more responsive to market changes.
- Automated analysis reduces human error and bias.

Challenges:

- Overfitting: Models might learn noise as signal in complex market data.
- ML models are highly dependent on the quality and quantity of the data fed into them.
- Interpretability: Advanced models can act as ”black boxes,” making it hard to understand the decision-making process.

Our approach

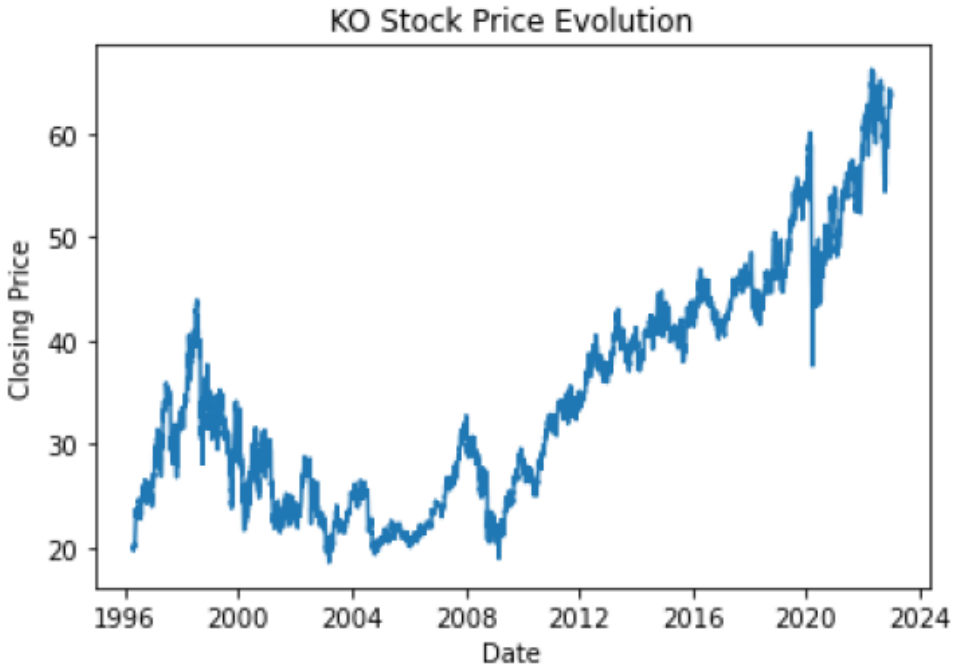
- **Objective:** to develop robust machine learning models for forecasting stock prices.
- We will explore the capabilities of **Random Forests** and **Long Short-Term Memory** (LSTM) networks.
- **Scope:** we will focus on two stocks with very different behavior and volatility, Coca-Cola (KO) and Tesla (TSLA).
- **Methodology:** follow the usual steps of data collection, exploratory data analysis, feature engineering and predictive modeling.

Data Collection

- We build two dataframes for KO and TSLA by collecting historical daily stock price data and earnings per share (EPS).
- Collected features: High, Low, Open and Close prices, trading volume, adjusted closing price and EPS.
- Data sources: Yahoo Finance for stock prices, Alpha Vantage for earnings.
- KO stock prices available from 1962, but earnings per share restrict usable data to 1996: 6500 datapoints.
- TSLA stock prices and earnings per share available for the whole company's existence: 3000 datapoints.

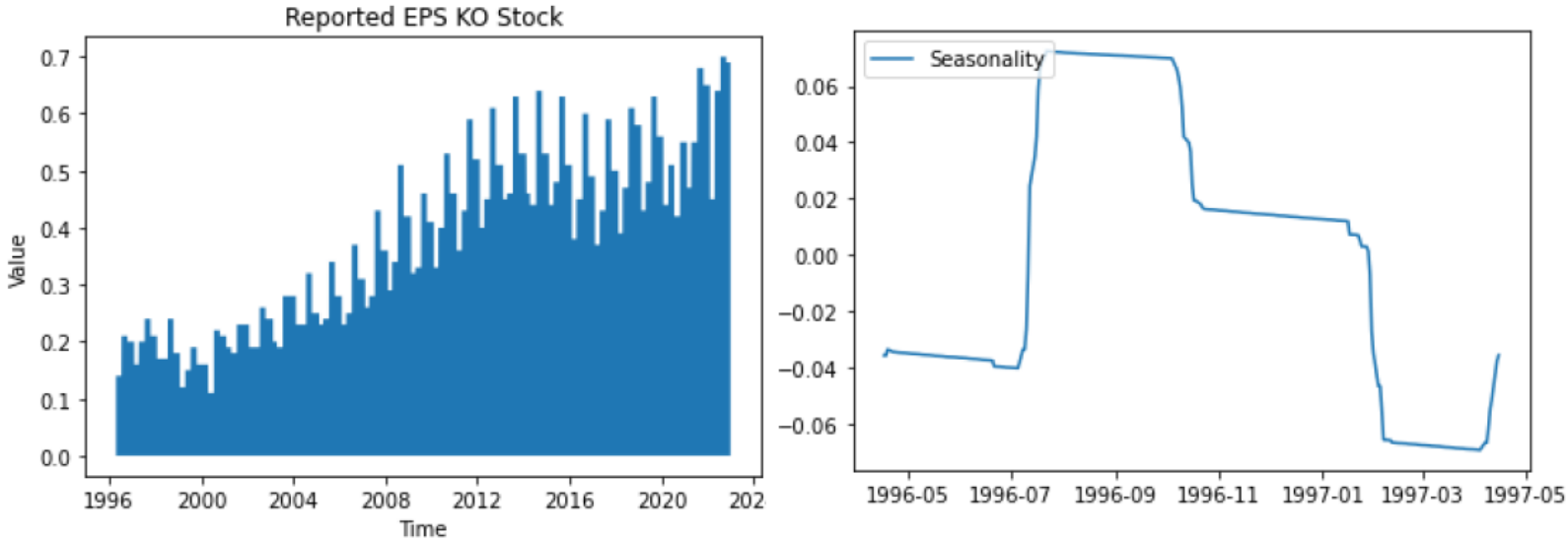
Exploratory Data Analysis

KO stock



- Phase of growth from 1996 to 1999. Contracting trend until 2006
- Short phase of exponential growth is killed by the 2008 crisis
- Growth at different rates from 2009, affected by the COVID pandemic in 2020

KO Earnings Per Share



- General trend of growth over time with some modulation
- Marked seasonality signal indicating higher earnings in the spring and summer.

KO Volatility

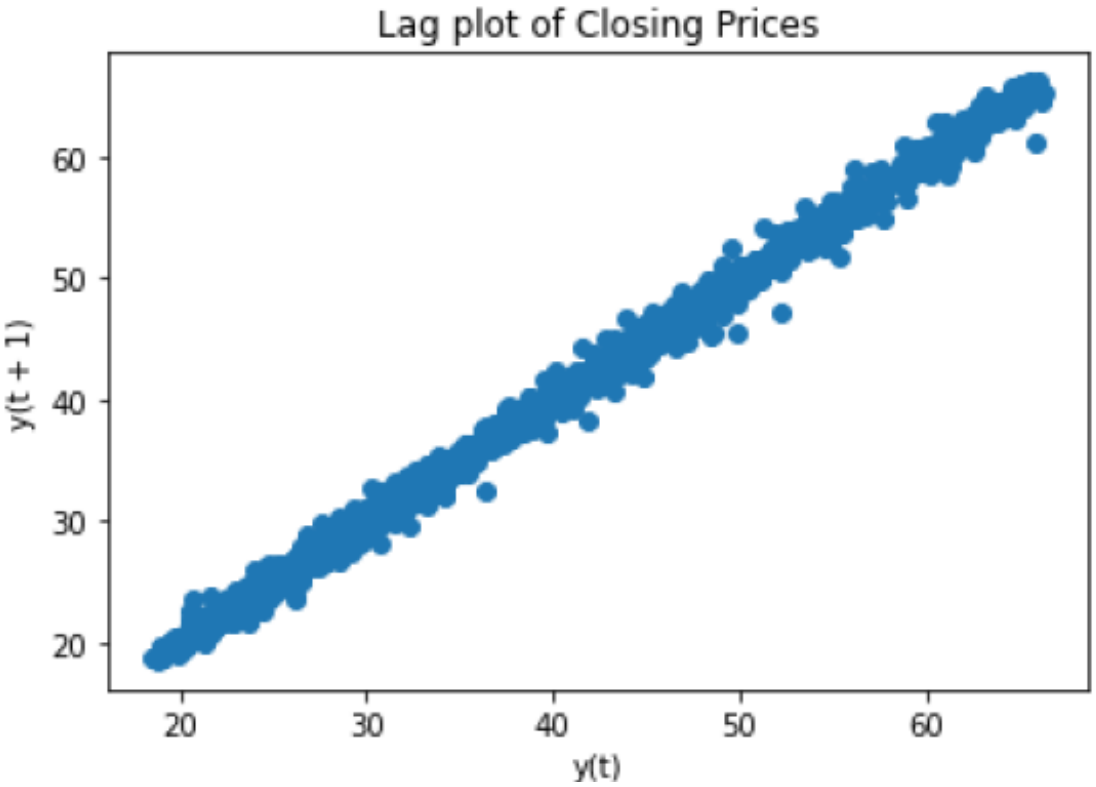
Volatility:

- Statistical measure of the degree of variation in a time series. Used to estimate the risk related to the changes in the values of a stock price.
- In finance, volatility is typically calculated as the standard deviation of logarithmic returns.
- Volatility above 30% is high. Below 20% is low.

KO stock: volatility 22%$\rightarrow$ Low!

This is in agreement with the analysis of outliers: in 26 years, only the years 1996, 1999, 2011, 2006 and 2022 show outlier values.

KO Volatility

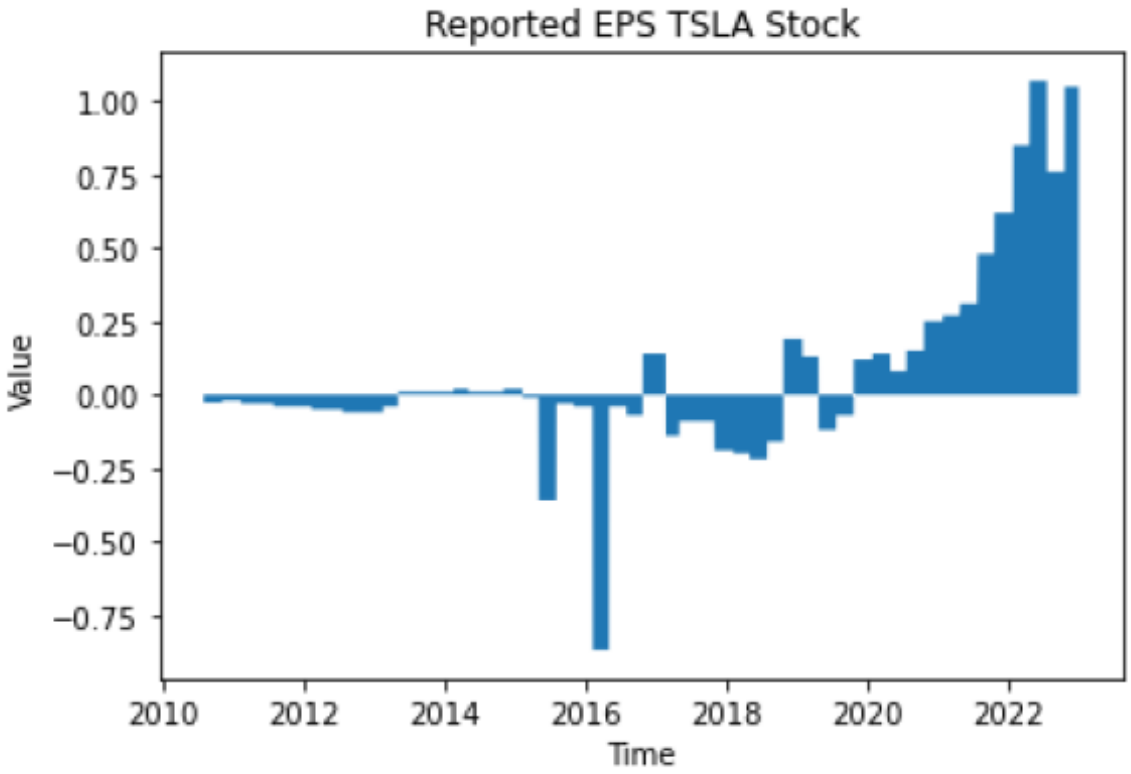


The lag plot shows homoscedasticity: KO stock's volatility can be considered approximately constant over time

TSLA stock

- First phase of almost constant closing price that lasts from 2010 to 2020
- Two phases of exponential growth between 2020 and 2022, separated by a slight decline in the stock price.
- In 2022, the stock price has declined throughout the year.

TSLA Earnings Per Share

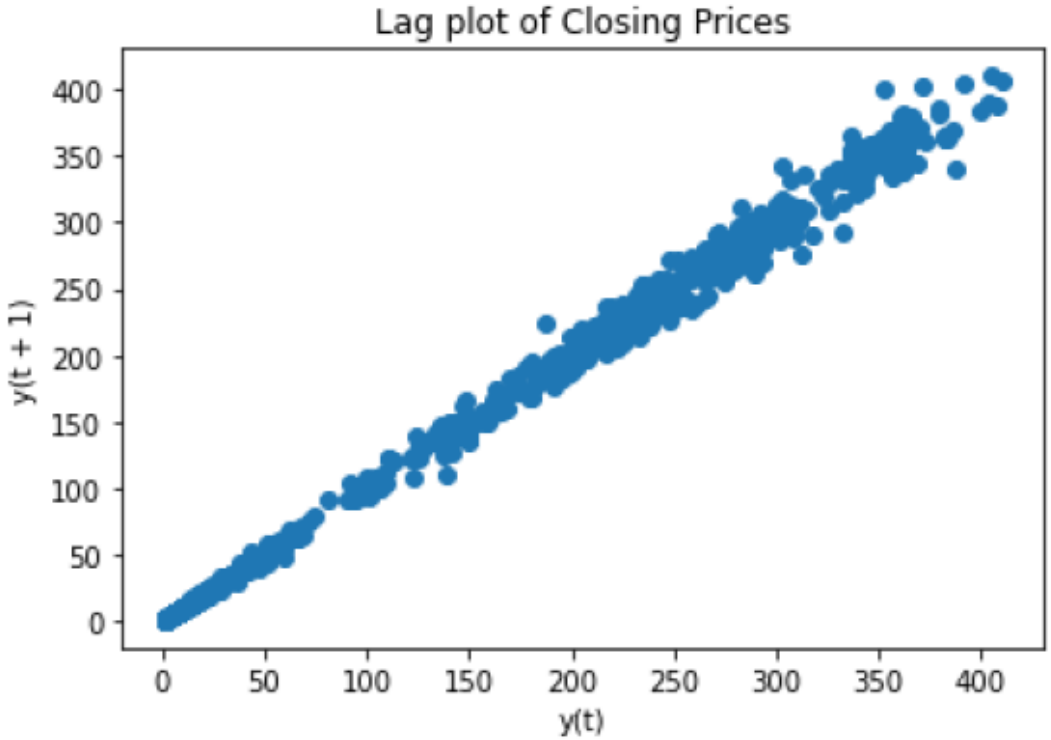


- From 2010 to 2020: EPS are negative most of the times. Not an uncommon scenario for tech startups.
- Profitable phase of exponential growth followed from 2020 to 2022, with oscillations ever since.

TSLA Volatility

TSLA stock: volatility 56%$\rightarrow$ Very high!

The lag plot shows heteroscedasticity: volatility is increasing over time.



Feature Engineering: adding financial indicators

- Prices and EPS alone are often not sufficient for stock price forecasting in machine learning models due to the complex nature of financial markets.
- Effective forecasting models generally benefit from a more comprehensive set of features that encompass various dimensions of market data: market sentiment, financial indicators, etc.
- Financial indicators provide insights into stock momentum and potential reversals, and aid in the identification of overbought or oversold conditions.
- We add to our dataframe six indicators as features: RSI, K value, MACD, ROC, OBV and P/E.

Modeling: Random Forest

Why Random Forests?

- Capable of handling non-linear relationships in financial data and complex interactions among features.
- Reduced overfitting and high accuracy.
- Can handle large datasets with numerous features.

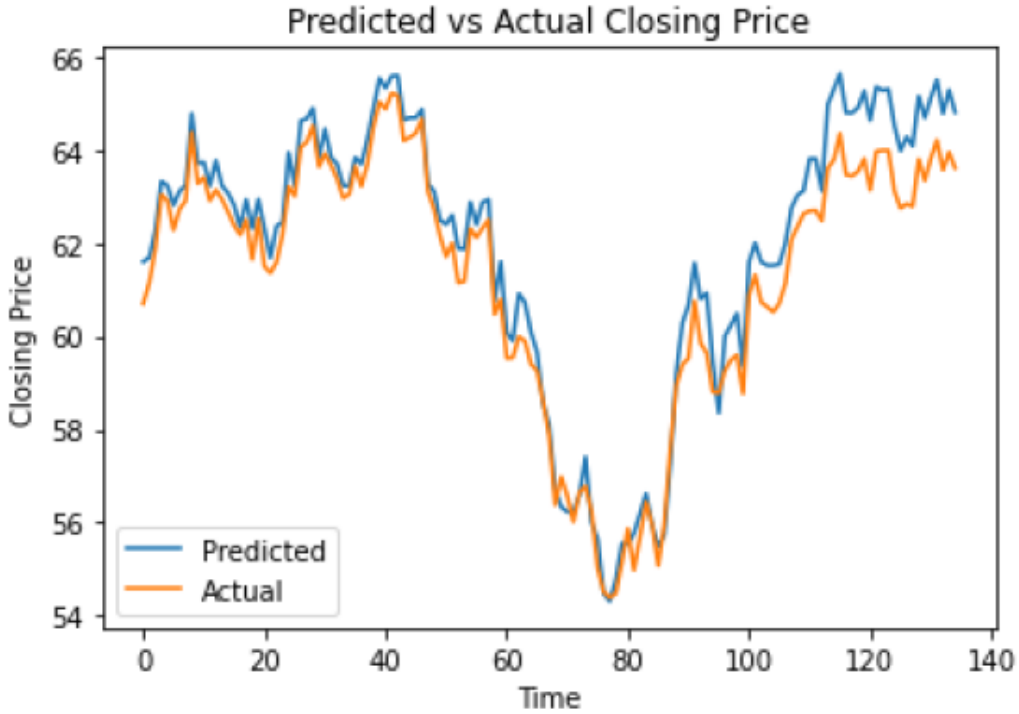
Challenges:

- They do not take into account the temporal structure of time series data. Can be an issue for stocks with a high temporal dependence.
- Lack of direct interpretability due to averaging process.

Modeling: Random Forest

KO stock results

- $n = 5$ trees, test size 2% (140 days)

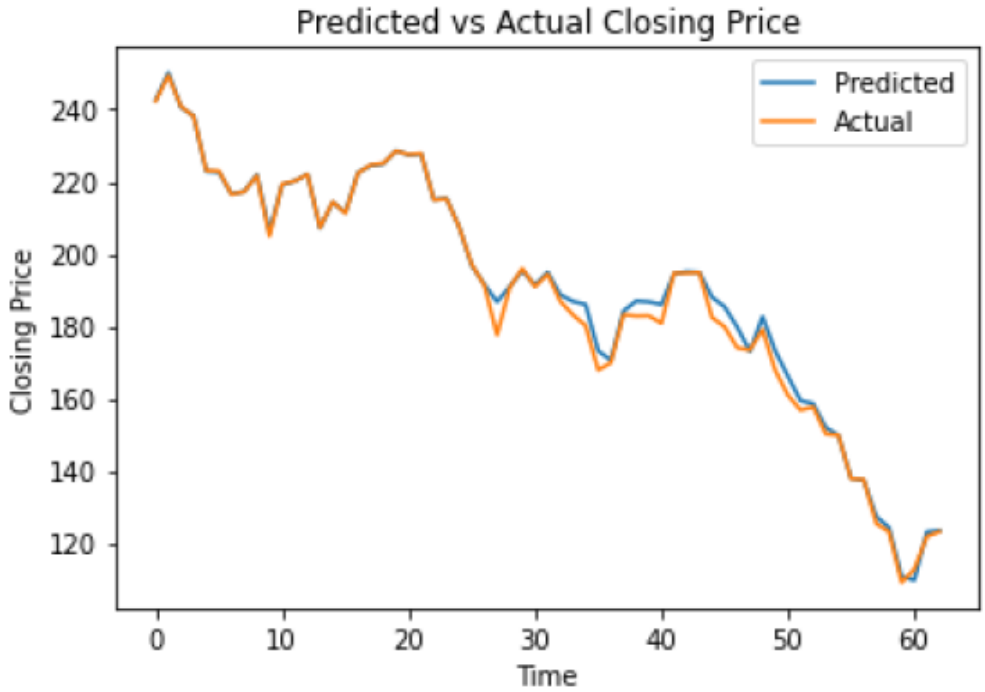


- MSE$\approx$ 0.17
- Cross validation: mean MSE$\approx$ 0.13, $\sigma \approx 0.003$

Modeling: Random Forest

TSLA stock results

- Feature importance analysis: only keep Open, High, Low and Adj Close.
- $n = 17$ trees, test size 2% (60 days)



- MSE$\approx$ 6.7
- Cross validation: mean MSE$\approx$ 6.56, $\sigma \approx$ 0.25

Modeling: LSTM Neural Networks

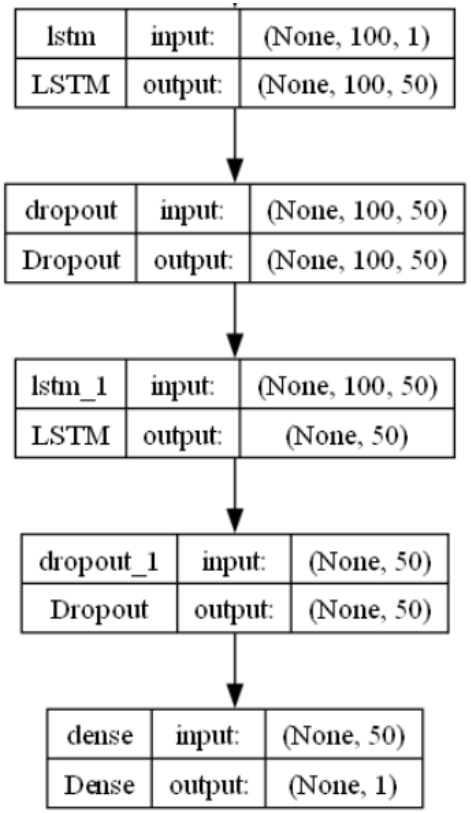
Long Short-Term memory networks (LSTM):

- Designed to remember patterns in sequential data over long-term time.
- Can model complex non-linear patterns and dependencies in historical stock price data.
- Improved ability to forecast future stock prices with higher accuracy.

But:

- Prone to overfitting, especially with smaller or noisier datasets.
- Computational intensity: they require significant computational resources and time for training.

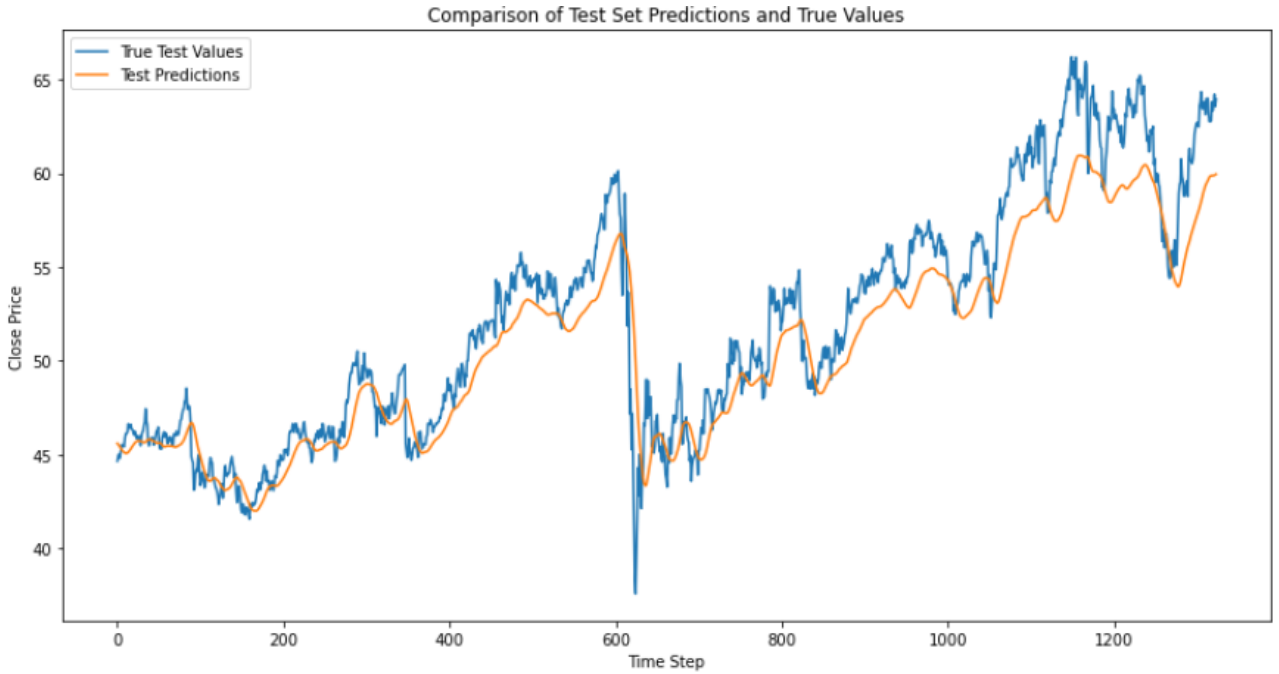
Our LSTM architecture



Regularization of the LSTM layers weights by L2 norm + dropout

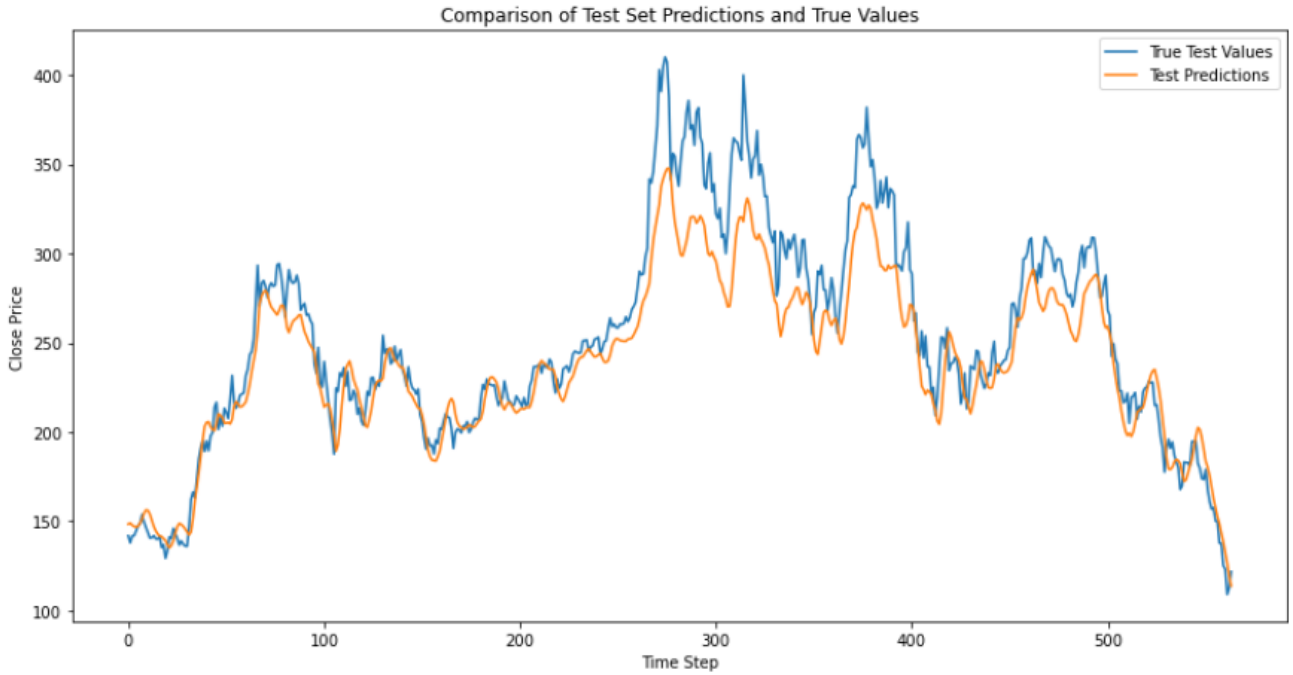
Total: 50451 trainable parameters

KO stock results



- MSE$\approx$ 5.60
- Cross validation: mean MSE$\approx$ 2.27, $\sigma \approx$ 1.80
- Overfitting reduced. Variability most likely due to COVID onset noise.

TSLA stock results



- Model captures overall behavior but suffers from a high MSE$\approx$ 533!
- Increasing the number of units in the first two layers reduces MSE but the model overfits.
- Address overfitting by regularization and hyperparameter tuning.

Conclusions

Random Forests:

- Limited in handling the temporal dependencies of financial data.
- Best suited for stable markets with less temporal dependency or forecasting over short periods of time.
- Accurate predictions on small-sized test sets for KO and TSLA.
- They struggle for larger test sets or during periods of high volatility.

LSTM networks:

- Excelled in modeling time-dependent data. Ideal for markets with strong temporal dependencies and longer-term trends.
- Excellent performance over long term for stable stocks such as KO. Overfitting can be reduced in these conditions.
- However, they may produce high errors in highly volatile markets, such as TSLA stock.

Recommendations

- Data scope: incorporating a broader range of data, such as market sentiment or macroeconomic factors, could potentially enhance model accuracy. This could be particularly relevant for the TSLA stock.
- Overfitting: future work could explore more sophisticated regularization techniques or model architectures to address this overfitting, especially for highly volatile stocks such as TSLA.
- Real-time forecasting capability: adapting these models for real-time forecasting remains a challenge.
- More complex models: CNNs for pattern recognition in time series, or more complex RNN variants like GRU might capture complex patterns in stock price data more effectively.