Duke
UNIVERSITY

FORECASTING STOCK MARKET TRENDS USING LEARNING-BASED TIME-SERIES MODELING TECHNIQUES

Authors:

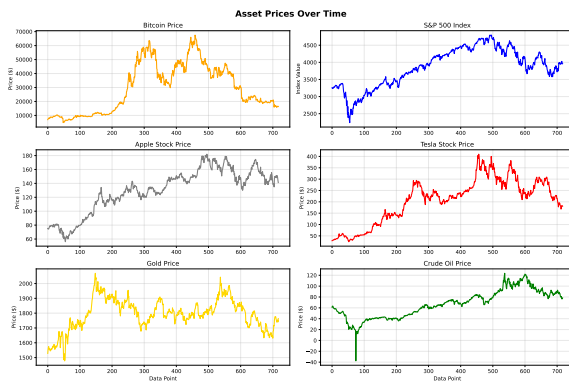Mohammad Afrazi

Patrick Lo

Problem Description

Predicting stock market movements is highly challenging due to the market's volatility, non-linearity, and chaotic dynamics [1]. Traditional time-series models such as ARIMA struggle to capture complex temporal dependencies in financial data, limiting forecasting accuracy and decision-making in finance [2].

We argue and hypothesize that learning-based time-series models - specifically Long Short-Term Memory (LSTM) networks, Temporal Convolutional Networks (TCNs), and Transformer- based models - can better model long-term dependencies in stock price data. This project will develop and compare these models for forecasting the daily closing price of the **S&P 500** index using historical price. Via evaluating LSTM, TCN, and Transformer approaches on a recent dataset of U.S. market data, we aim to advance financial forecasting methods and provide more reliable tools for investment and risk management [3].

Data Description

To address our research question, we utilize the "US Stock Market Dataset" from Kaggle, which contains over five years of daily stock and commodity data for a diverse set of U.S.-listed companies and assets. Each record includes the following features: *No, Date, Natural_Gas, Natural_Gas_Vol., Crude_oil, Crude_oil_Vol., Copper, Copper_Vol., Bitcoin, Bitcoin_Vol., Ethereum, Ethereum_Vol., S&P_500, Nasdaq_100_Price, Nasdaq_100_Vol., Apple, Apple_Vol., Tesla, Tesla_Vol., Microsoft, Microsoft_Vol., Silver, Silver_Vol., Google, Google_Vol., Nvidia, Nvidia_Vol., Berkshire, Berkshire_Vol., Netflix, Netflix_Vol., Amazon, Amazon_Vol., Meta, Meta_Vol., Gold, Gold_Vol.*

To mitigate the effects of major scale differences across time periods, we restrict our analysis to data from January 2, 2020, to November 29, 2022, yielding a total of 718 daily observations. Figures 1a and 1b provide visual summaries of the dataset, illustrating overall asset price dynamics and a sample of the input features used in our analysis.

(a) Asset price trends.

(b) Data preview.

Figure 1: Visualization of the dataset showing asset price dynamics and a sample of the input data.

Preprocessing

For preprocessing, we first compute the log returns of each asset to normalize price movements and stabilize variance across different scales. The log return for time t is calculated as

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t represents the asset price at time t. This transformation converts multiplicative price changes into additive ones, making the data more stationary and suitable for modeling. It also helps mitigate the effects of large price discrepancies between assets, ensuring that model training focuses on relative changes rather than absolute price levels.

Work Breakdown:

- **Mohammad Afrazi:** Data preprocessing, feature engineering, writing, and model training/evaluation.

- **Patrick Lo:** Data preprocessing, feature engineering, writing, and model training/evaluation.

References

[1] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

[2] Mehdi Khashei and Mehdi Bijari. A novel hybridization of artificial neural networks and arima models for time series forecasting. *Applied Soft Computing*, 11(2):2664–2675, 2011.

[3] David MQ Nelson, Adriano CM Pereira, and Renato A De Oliveira. Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, pages 1419–1426. Ieee, 2017.