Duke
UNIVERSITY

FORECASTING STOCK MARKET TRENDS USING ADVANCED TIME-SERIES MODELING TECHNIQUES

Authors:

Mohammad Afrazi

Patrick Lo

Problem Description

Predicting stock market movements is a notoriously difficult task due to the market's inherent volatility, non-linearity, and chaotic nature [?]. Traditional time-series analysis models, such as ARIMA, often fail to capture the complex temporal dependencies and non-linear patterns present in financial data. This limitation presents a significant challenge for investors and financial institutions who rely on accurate forecasts for risk management and strategic decision-making [?]. The development of more robust predictive models could lead to substantial improvements in financial forecasting, enabling more informed investment strategies.

Our central hypothesis is that deep learning approaches, specifically advanced time-series architectures such as Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs), Temporal Convolutional Networks (TCNs), and Transformer-based models, can more effectively model the long-term sequential dependencies in stock price data than traditional methods. We propose to develop and evaluate models based on these architectures to forecast the daily closing price of a stock using its historical price and volume data. The project's impact lies in its potential to provide a more accurate and reliable tool for financial forecasting, a problem of enduring interest to both the finance industry and academic researchers. Previous work has shown promise in using LSTMs for this purpose, but performance varies significantly based on model architecture and the specific dataset used [?]. Our work aims to build on this foundation by implementing and comparing carefully structured models across LSTM, TCN, and Transformer frameworks, evaluating their performance on a comprehensive, recent dataset of US stocks.

Data Description

To address our research question, we plan to use the ”US Stock Market Dataset,” a publicly available dataset on Kaggle. This dataset contains daily stock data for a wide range of U.S.-listed companies spanning at least ten years. The features for each stock include daily Open, High, Low, Close, and Volume values—standard metrics used in financial analysis.

We have already downloaded this data, so there are no obstacles to its acquisition. The dataset is provided in a clean CSV format, which will simplify the preprocessing stage. The historical, sequential nature of this data is essential for our proposed approach. The daily closing prices will serve as our target variable for prediction, while the sequence of past prices and trading volumes will be used as input features to train our time-series models, including LSTM, TCN, and Transformer architectures. This rich historical context is precisely what such models are designed to leverage, allowing them to learn temporal and structural patterns over time that may be predictive of future movements.

Tentative Approach

Our planned methodology consists of a multi-stage process, beginning with data preparation and concluding with model evaluation.

First, we will perform data preprocessing. This involves selecting a specific stock from the dataset (e.g., Apple Inc. - AAPL) for our initial univariate analysis. We will then normalize the data, likely using a MinMaxScaler, to scale all feature values between 0 and 1. This step is critical for ensuring stable training of the neural network.

Second, we will implement multiple time-series model architectures, including Long Short-Term Memory (LSTM) networks, Temporal Convolutional Networks (TCNs), and Transformer-based models, using TensorFlow and Keras. Each model will be designed to take a sequence of historical data (e.g., data from the previous 60 days) as input to predict the closing price of the next day. For the LSTM architecture, several stacked LSTM layers will be used with Dropout layers in between to prevent overfitting. The TCN and Transformer architectures will be adapted for financial time-series forecasting to capture both local temporal patterns and long-range dependencies in the data.

Third, we will train and evaluate each model. The dataset will be divided into training and testing sets, using approximately 80% for training and 20% for testing. Model performance will be evaluated using the Root Mean Squared Error (RMSE) metric, which quantifies the average deviation between predicted and actual stock prices. Additionally, we will visualize the predicted prices against the actual prices to qualitatively assess each model's forecasting accuracy and temporal alignment.

This approach is highly feasible. LSTM, TCN, and Transformer models are well-established and powerful tools for time-series forecasting, and the required software libraries (TensorFlow, Scikit-learn, Pandas) are open-source and well-documented. The dataset is clean and of sufficient size to train these deep learning models effectively, enabling robust comparison and analysis of their relative performance.

Work Breakdown:

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