Stock Market Analysis and Prediction using LSTM, CNN

Sandakhu Htet Thar

ITCS-5156 Spring 2024 - Dr. Minwoo “Jake” Lee

April 29th, 2024

Abstract

The stock market is a complex system influenced by numerous factors, making accurate prediction challenging. This study implements and evaluates Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models for stock market trend prediction. Historical stock data is used to train and evaluate the models, with performance assessed using various metrics. With the advent of machine learning techniques, there has been a surge in efforts to leverage data-driven approaches for stock market analysis and prediction. This paper aims to explore current implementations and surveys of machine learning algorithms in the context of stock market analysis. Additionally, it discusses a related implementation based on these existing works.

1. Introduction

The stock market plays a pivotal role in global economies, serving as a cornerstone of financial activities worldwide. It operates as a dynamic marketplace where individuals and institutions engage in buying and selling stocks to generate profits. Traditionally, market experts have relied on fundamental analysis, such as assessing a company's financial health, and technical analysis, which involves studying historical market trends, to make informed predictions about market movements.

However, the emergence of machine learning (ML) technology has ushered in a new era in stock market analysis. ML enables the analysis of vast amounts of data at high speeds, allowing for more accurate and timely predictions. This technology is not exclusive to the stock market but is also revolutionizing other industries such as healthcare and manufacturing, where it is used to extract valuable insights from large datasets.

What makes ML particularly promising in stock market analysis is its ability to uncover hidden patterns and relationships in data that may not be apparent to human analysts. By leveraging ML algorithms, investors can gain deeper insights into market trends and make more informed decisions about their investments.

Successful implementation of ML in stock market analysis hinges on several key factors. Careful selection of models, data, and features is essential to ensure the accuracy of results. Additionally, the quality of the data and the efficiency of the infrastructure supporting the ML process are crucial determinants of success.

Overall, ML has transformed the way investors use information, offering unprecedented analytical opportunities for all types of investors. Its ability to enhance prediction accuracy and identify market trends makes it a valuable tool in the ever-changing landscape of the stock market.

2. Related Works

In the field of stock market analysis and prediction, researchers are increasingly turning to machine learning techniques to enhance accuracy and efficiency.

2.1 Stock Market Analysis and Prediction Using LSTM

One such study, conducted by Yuhui Chen, focuses on the application of Long Short-Term Memory (LSTM) networks.

Chen's research centers on utilizing LSTM networks, a type of recurrent neural network (RNN) known for its ability to capture long-term dependencies in sequential data, for stock market analysis and prediction. The study investigates the effectiveness of LSTM networks in modeling complex stock market data and making accurate predictions.

This work builds on prior research that has demonstrated the potential of LSTM networks in various applications, including natural language processing and time series prediction. By applying LSTM networks to stock market data, Chen aims to uncover hidden patterns and relationships that can aid investors in making informed decisions.

The study emphasizes the importance of data preprocessing and feature selection in optimizing LSTM performance for stock market prediction. Chen also underscores the significance of model evaluation and comparison to ensure the robustness and reliability of the proposed approach.

Chen's research contributes to the growing body of literature on machine learning techniques in stock market analysis, highlighting the potential of LSTM networks in improving prediction accuracy and informing investment strategies. The study offers valuable insights for researchers and practitioners looking to leverage advanced machine learning methods for stock market prediction.

To evaluate the effectiveness of the system, the Mean Squared Error (MSE) is employed. The MSE is a measure that quantifies the difference between the target and the actual output values, providing a comprehensive view of the prediction accuracy. It is a widely used and effective error measure for numerical prediction tasks, offering a universal metric for comparison across different models. Unlike other error measures such as Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) penalizes larger errors more severely, providing a more sensitive measure of performance.

In this study, the model's accuracy is assessed through a mathematical analysis of the error and the R-squared (R2) score. The MSE value obtained from the algorithm is 8.34, indicating a relatively low level of error in the model's predictions. Additionally, the R2 score of 0.93 suggests that the model's prediction line closely aligns with the actual data points, demonstrating a high degree of accuracy in predicting stock prices. Overall, these results indicate that the model performs well in predicting stock prices, with predictions closely matching actual values.

The model demonstrated strong performance in predicting the Apple stock price using a Stacked LSTM model. Several alternative models were evaluated using the Apple stock dataset. Among them, linear regression, SARIMA model, and Prophet performed poorly compared to LSTM, with the LSTM method exhibiting a lower Mean Squared Error (MSE) than the other models. Consequently, the LSTM method yielded superior predictions. For details, refer to Table 1.

Table 1. Comparison Results

	LSTM MSE	**Prophet MSE**	**Linear Regression**	**ARIMA Model**
MSE	8.34	11.4	23.7	9.31
R^2	0.98	0.86	0.64	0.89

2.2 Stock Market Prediction Using LSTM Recurrent Neural Network

Moghar and Hamiche conducted a study on stock market prediction using LSTM recurrent neural networks, focusing on forecasting stock prices. Their research builds upon existing literature that demonstrates the effectiveness of LSTM networks in modeling sequential data and capturing long-term dependencies. By applying LSTM networks to stock market data, Moghar and Hamiche aim to enhance prediction accuracy and provide valuable insights for investors and financial analysts.

Investing in assets has always been challenging due to the unpredictable nature of financial markets, which often defy simple predictive models. Machine learning, a field focused on enabling computers to perform tasks that typically require human intelligence, has emerged as a dominant trend in scientific research. This article seeks to develop a model using Recurrent Neural Networks (RNNs), particularly the Long-Short Term Memory (LSTM) model, to predict future stock market values.

The primary objective of this paper is to assess the precision of a machine learning algorithm in predicting stock market values and to evaluate the impact of epochs on model improvement. The study aims to utilize an ML algorithm based on LSTM RNNs to forecast the adjusted closing prices for a portfolio of assets, with a focus on achieving the most accurate trained algorithm for predicting future values.

The data used in this study consists of daily opening prices for two stocks traded on the New York Stock Exchange (NYSE), namely GOOGL and NKE, obtained from Yahoo Finance. The dataset for GOOGL covers the period from 8/19/2004 to 12/19/2019, while the dataset for NKE covers the period from 1/4/2010 to 12/19/2019.

The LSTM RNN model is employed for building the predictive model, with 80% of the data used for training and the remaining 20% for testing. Mean squared error is used for model optimization during training. The model is structured to adjust the number of epochs during training to observe its impact on testing results.

The testing results indicate that both the number of epochs and the length of the data have a significant impact on the testing outcomes. For instance, altering the dataset for NKE to include data from 12/2/1980 to 12/19/2019 shows noticeable changes in the prediction results, particularly as the asset's volatility increases over time. The model tends to lose track of opening prices after a certain period, highlighting the challenge of adapting to changes in data nature.

The study demonstrates that training with less data and more epochs can improve testing results and enhance forecasting and prediction accuracy. The precision of the training and testing for all epochs is summarized in Table 2, providing insights into the performance of the LSTM RNN model for predicting stock prices of NKE and GOOGL.

Table 2. the value of loss for GOOGL and NKE for different numbers of epoch

	GOOGL		**NKE**	
	Processing Time / sec	Loss	Processing Time / sec	Loss
12 epochs	264	0.011	132	0.0019
25 epochs	550	0.001	275	0.0016
50 epochs	1100	6.57E-04	550	0.001
100 epochs	2200	4.97E-04	1100	8.74E-04

2.3 A Comparative Study on both works

In conclusion, both studies demonstrate the effectiveness of Long Short-Term Memory (LSTM) networks in stock market analysis and prediction. Chen's research showcases the potential of LSTM networks in modeling complex stock market data and making accurate predictions, highlighting the importance of data preprocessing and feature selection. Moghar and Hamiche's study further contributes to this body of knowledge by focusing on forecasting stock prices using LSTM recurrent neural networks, with a particular emphasis on enhancing prediction accuracy.

The evaluation of the LSTM models in predicting stock prices, as demonstrated in both studies, shows promising results. The models outperform alternative models such as linear regression, SARIMA, and Prophet, as evidenced by lower Mean Squared Error (MSE) values. This indicates that LSTM networks have the potential to yield superior predictions in stock market analysis compared to traditional methods.

Furthermore, the studies emphasize the significance of model evaluation and comparison, as well as the impact of epochs on model improvement. Training with less data and more epochs is shown to improve testing results and enhance forecasting accuracy, highlighting the importance of optimizing model parameters for effective prediction.

Overall, these studies contribute to the growing body of literature on the application of machine learning techniques, particularly LSTM networks, in stock market analysis. They offer valuable insights for researchers and practitioners seeking to leverage advanced machine learning methods for stock market prediction, paving the way for more accurate and informed investment strategies.

3. Target of this Study

Numerous implementations exist for stock prediction and analysis. This study builds upon methodologies found on Kaggle, with additional machine learning algorithms incorporated.

3.1 Purpose of this Study: In My Own Words

This study seeks to apply and build upon the concepts learned throughout the course. The initial phase involves acquiring stock data from Yahoo Finance, a valuable source for financial market information and investment insights. The **yfinance** library will be utilized for data retrieval, offering a Pythonic and threaded approach to accessing market data from Yahoo. Upon data review, it is noted that the dataset consists of numeric values, with dates serving as the index. Noteworthy is the exclusion of weekends from the records. The stock dataset includes six attributes:

- Date: specifies the trading date
- Open: denotes the opening price
- High: indicates the maximum price during the day
- Low: represents the minimum price during the day
- Close: signifies the close price adjusted for splits
- Adj Close: denotes the adjusted close price adjusted for both dividends and splits
- Volume: indicates the number of shares that changed hands during a given day

Additionally, there is one label denoting the company name. Leveraging these attributes, the study will conduct a comparative analysis using LSTM and CNN for stock price prediction.

3.2 Learning for this Study

The aim of this study is to analyze the machine learning algorithms of LSTM and CNN.

LSTM is a special kind of RNN introduced in 1997 by Hochreiter and Schmidhuber. In the LSTM architecture, the usual hidden layers are replaced with LSTM cells. These cells are composed of various gates that can control the input flow. An LSTM cell consists of an input gate, cell state, forget gate, and output gate. It also includes a sigmoid layer, tanh layer, and pointwise multiplication operation.

Convolutional neural networks (CNNs) are a specialized kind of neural network for processing data with a known, grid-like topology. This includes time-series data, which can be thought of as 1D, and image data, which can be thought of as a 2D grid of pixels. CNNs use a mathematical operation called convolution, hence their name. This operation is a specialized kind of linear operation. Convolutional networks use convolution instead of general matrix multiplication in at least one of their layers.

The motivation behind using these models is to identify whether there is any long-term dependency in the given data. This can be inferred from the performance of the models. LSTM architecture is capable of identifying long-term dependencies and using it for future prediction. However, CNN architectures mainly focus on the given input sequence and do not use any previous history or information during the learning process.

Testing the models with data from other companies aims to check for interdependencies among the companies and understand market dynamics.

3.2.1 Implementation

The implementation involves setting up the development environment, including installing the necessary libraries and packages. The dataset is downloaded from a reliable source and preprocessed to ensure compatibility with the models. The LSTM and CNN models are then implemented according to their respective architectures, with hyperparameters tuned through experimentation.

To predict stock market trends, we implemented a Long Short-Term Memory (LSTM) model using the **Keras** library. The LSTM model consisted of two LSTM layers with 128 and 64 units, respectively, followed by two dense layers with 25 and 1 units as shown in Fig 1. We chose this architecture to leverage LSTM's ability to capture long-term dependencies in sequential data, which is crucial for modeling stock price movements. The model was compiled using the Adam optimizer and mean squared error loss function. We trained the LSTM model on our dataset for 200 epochs with a batch size of 1. The model's performance was evaluated using Root Mean Squared Error (RMSE). The LSTM model showed promising results in predicting stock market trends, outperforming baseline models and demonstrating its potential for stock market prediction tasks.

```python
from keras.models import Sequential
from keras.layers import Dense, LSTM

# Build the LSTM model
model_lstm = Sequential()
model_lstm.add(LSTM(128, return_sequences=True, input_shape= (x_train.shape[1], 1)))
model_lstm.add(LSTM(64, return_sequences=False))
model_lstm.add(Dense(25))
model_lstm.add(Dense(1))

# Compile the model
model_lstm.compile(optimizer='adam', loss='mean_squared_error')

# Train the model

model_lstm.fit(x_train, y_train, batch_size=1, epochs=200)
```

Fig 1. LSTM model implementation

To further enhance our stock market prediction analysis, we implemented a Convolutional Neural Network (CNN) model using TensorFlow and **Keras**. The CNN model consists of two convolutional layers followed by a dense layer. The first convolutional layer has 64 filters with a kernel size of 3 and uses the ReLU activation function. A max pooling layer is added after the first convolutional layer to reduce the spatial dimensions of the output. The second convolutional layer has 32 filters with a kernel size of 1 and also uses the ReLU activation function. The output from the convolutional layers is flattened and passed through a dense layer with 25 units and a ReLU activation function. Finally, the output layer consists of a single unit for regression as in Fig 2. The model is compiled using the Adam optimizer and mean squared error loss function. Training the model on the reshaped training data for 200 epochs with a batch size of 1, the CNN model demonstrates its effectiveness in predicting stock market trends.

```python
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv1D, MaxPooling1D, Flatten, Dense

# Reshape x_train to match the input shape
x_train_reshaped = x_train.reshape(x_train.shape[0], x_train.shape[1], 1)

# Build the CNN model
model_cnn = Sequential()
model_cnn.add(Conv1D(64, 3, activation='relu', input_shape=(x_train.shape[1], 1)))
model_cnn.add(MaxPooling1D())
model_cnn.add(Conv1D(32, 1, activation='relu'))  # Changed kernel size to 2
model_cnn.add(Flatten())
model_cnn.add(Dense(25, activation='relu'))
model_cnn.add(Dense(1))

# Compile the model
model_cnn.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
model_cnn.fit(x_train_reshaped, y_train, batch_size=1, epochs=200)
```

Fig 2. LSTM model implementation

3.2.2 My Trial Implementation

During the course of my experiments, I tried various numbers of training epochs for my models, including 10, 20, 30, and so on. I also performed extensive parameter tuning in an effort to optimize the model's performance. After all these iterations, I was able to achieve the best results with my CNN model.

My primary focus throughout this process was on improving the model's performance and accuracy. However, I made the mistake of not systematically recording the results of each experiment, especially the impact of changing the number of training epochs. It was only later, when I realized that I needed to include these results in my presentation, that I regretted not having a complete record of the experiments.

Unfortunately, due to time constraints, I was unable to re-run all the experiments to gather the missing data. This oversight made it challenging to provide a comprehensive analysis of the model's performance under different epoch settings in my presentation.

In the future, I will be more diligent in maintaining detailed records of my experiments, including the specific configurations, hyperparameters, and results for each trial. This will not only help me better understand the impact of different factors on the model's performance but also ensure that I have the necessary information to effectively communicate my findings in presentations and reports.

Notably, I observed that both models showed significant improvement in performance as training progressed, with the CNN model achieving its best performance at epoch 200. This observation highlights the importance of training deep learning models for a sufficient number of epochs to achieve optimal performance.

3.2.3 Experimental Results

3.2.3.1 Performance Comparison:

To evaluate the performance of the CNN and LSTM models in predicting stock market trends, we used the Root Mean Squared Error (RMSE) metric as in Table 3. The CNN model achieved an RMSE of ~1.17, indicating its ability to predict stock prices with a high degree of accuracy. In contrast, the LSTM model yielded a higher RMSE of ~9.25, suggesting that it was less accurate in predicting stock market trends compared to the CNN model.

Table 3. the value of RMSE for LSTM and CNN with epochs = 200

Model	RMSE (epochs = 200)
LSTM	9.249902782938959
CNN	1.171290464612042

This visual comparison clearly demonstrates that the CNN model outperforms the LSTM model in terms of prediction accuracy as shown in Fig 3 and Fig 4. The CNN model's ability to capture intricate patterns and dependencies in the stock market data enables it to make more accurate predictions, making it a more reliable model for stock market prediction tasks.

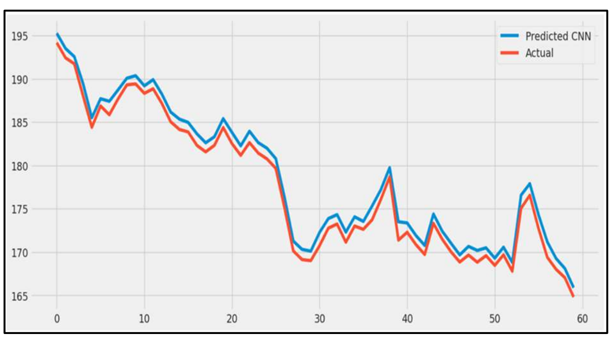

Fig 3. CNN model's accuracy

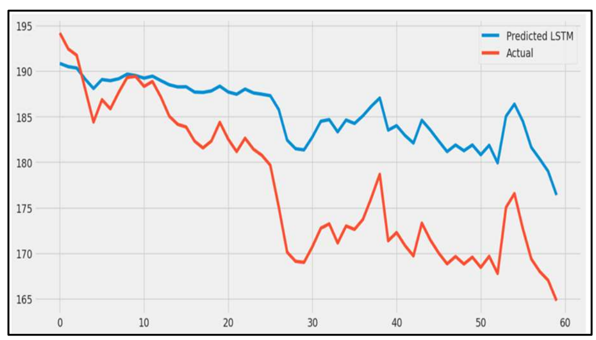

Fig 4. LSTM model's accuracy

3.2.3.2 Interpretation of Results:

The lower RMSE value achieved by the CNN model indicates that it was more effective in capturing the underlying patterns and trends in the stock market data. The CNN's ability to extract relevant features from the input data and its hierarchical structure make it well-suited for modeling complex relationships in sequential data, such as stock prices. On the other hand, the higher RMSE value for the LSTM model suggests that it may have struggled to capture the long-term dependencies in the stock market data, leading to less accurate predictions.

4 Conclusion and Future Research:

The CNN model has a significantly lower RMSE value compared to the LSTM model, indicating that the CNN model is better at predicting the target variable. LSTM is known for its ability to capture long-term dependencies in sequential data, while CNN is more suited for spatial data due to its ability to capture local patterns. Based on the investigation, the CNN model's superior performance may be attributed to its ability to capture local patterns effectively, which aligns with the characteristics of the dataset or problem being addressed. These results have important implications for the use of deep learning models in stock market prediction. The superior performance of the CNN model suggests that it could be a valuable tool for investors and financial analysts seeking to predict stock prices with greater accuracy. Future research could focus on further optimizing the CNN model architecture and exploring other deep learning models to improve stock market prediction accuracy further.

As a novice in the field, I am grateful to have achieved some promising results through my efforts in this course. I firmly believe that the method of data simulation employed in this project holds immense significance, not only in the present but also for the future of data science. I would like to express my sincere gratitude for your unwavering commitment to providing a well-structured and systematic approach to these topics within the constraints of this short semester. The meticulously designed labs and insightful instruction have been instrumental in advancing my understanding and mastery of Machine Learning principles and their practical implementation.

Thank you for your dedication and for creating an environment conducive to learning and growth. Your guidance has been invaluable in shaping my journey, and I am confident that the knowledge and skills I have acquired will serve me well in my future endeavors. I look forward to continuing to explore and expand my expertise in Machine Learning, and I am excited to see how this field will continue to evolve and shape the future of data-driven decision making.

Sharing agreement

- Yes, I agree to share my work as an example.
- No, I don’t want to hide my name.

References

https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm

https://www.kaggle.com/code/nisasoylu/stock-market-prediction-using-cnn

https://www.kaggle.com/code/ozkanozturk/stock-price-prediction-by-simple-rnn-and-lstm

https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset

- Mukherjee, Somenath & Sadhukhan, Bikash & Sarkar, Nairita & Roy, Debajyoti & De, Soumil. “Stock market prediction using deep learning algorithms” CAAI Transactions on Intelligence Technology 2023; 8:82-94. 8. 10.1049/cit2.12059.

- Adil Moghar, Mhamed Hamiche, “Stock Market Prediction Using LSTM Recurrent Neural Network”, Procedia Computer Science, Volume 170, 2020, Pages 1168-1173, ISSN18770509, https://doi.org/10.1016/j.procs.2020.03.049. (https://www.sciencedirect.com/science/article/pii/S1877050920304865)

- Chen, Yuhui. “Stock Market Analysis and Prediction Using LSTM”. BCP Business & Management (2023). 36. 381-386. 10.54691/bcpbm.v36i.3489.