The Investigation of LSTM-Random Search with Various Standardization and Normalization Technologies

Xinyu Ma *

School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, China

* Corresponding author: 1345227791@stu.xjtu.edu.cn

Abstract. The stock market, being a complex and chaotic system, presents a formidable challenge for prediction. This paper explores the intricacies of stock prediction by focusing on data preprocessing, specifically investigating the impact of different standardization and normalization techniques on LSTM structures. To enhance predictive accuracy, this study employed a stochastic search technique in this study. It is tested using 1520 data points of the closing prices of Bitcoin and Amazon that were obtained from Kaggle between May 1, 2013, and May 14, 2019. It is discovered that the prediction performance under the LSTM framework is significantly impacted by various data preprocessing techniques. The normalization techniques employing MinMaxScaler and MaxAbsScaler performed better for AMZN, a stock with stable prices, and the RobustScaler normalization method yielded the best results for Bitcoin stock prediction, suggesting that it is more useful for datasets with higher volatility. These results provide a useful guide for future research on stock price prediction by highlighting the significance of selecting the right preprocessing technique based on data characteristics and highlighting the benefits of dynamically modifying the lstm structure for stochastic searches.

Keywords: LSTM, Stock Prediction, Normalization, Standardization.

1. Introduction

With the vigorous development of artificial intelligence technology and its gradual application in the financial field, Financial Technology, Fintech came into being [1], and then the wave of using machine learning for stock prediction swept the world. The process of projecting future stock prices based on historical prices is known as stock price prediction, which can help investors make more money from trading stocks [2]. The two conventional methods of prediction are fundamental analysis and technical analysis. While fundamental analysis concentrates on the state of the economy as a whole, the company's finances, and its management, technical analysis looks for patterns in previous data [3].

Among the various methods to obtain features from past data for prediction, Long Short-Term Memory (LSTM) [4] is a very suitable method. The important hyperparameters of the deep learning model LSTM used include batch size, number of epochs, number of hidden layers, etc. Activation functions (sigmoid, tanh, etc.); Optimizers (Adam, RMSprop, etc.) [5]. The model's stock prediction performance varies noticeably depending on the combination of these hyperparameters because they comprise a very vast solution space. To improve a model and change parameters, it is worthwhile to employ better techniques. For instance, Bao et al combined genetic algorithm and LSTM model to adjust parameters, and used genetic algorithm to design evaluation mechanism to get a better index model [6]. Chen found an algorithm model with high accuracy for forecasting the price of Bitcoin the following day by contrasting random forest regression and the LSTM model. And the factors that affect the price of Bitcoin were clarified. A model with only one lag of the explanatory variables had the best predictive accuracy for predicting the price of bitcoin the following day, according to the relationship between the number of periods and the accuracy of the explanatory variables brought by his model [7].

Srivastava et al. investigated a variety of stock market prediction algorithms, comparing the performance and accuracy of each, from straightforward ones like simple mean and linear regression to more complex ones like AutoRegressive Integrated Moving Average (ARIMA) and LSTM. Lastly,

a research method is presented that makes use of an enhanced (LSTM) version of the Recurrent Neural Network (RNN) and maintains each data variable's weight through random gradient descent. helped them achieve more accurate and efficient results than stock price prediction systems that were previously in place [8].

This paper presents a method to adjust the structure of the LSTM model using Stochastic search, and makes a prediction study on two very representative stocks (AMZN and BTC). A variety of typical standardized methods are used to predict the closing price of two stocks, and the results of different standardized methods are analyzed and compared.

The contribution of this work is twofold. First, this work proposes a method adjusting model structure based on Stochastic search. This will surely be a good tool in obtaining better LSTM model structure and improving accuracy. Second, it systematically compares the effects of different standardization methods on the LSTM model prediction. This contributes to effectively choose suitable standardized procedures for stocks with various attributes.

2. Method

The purpose of this paper is to forecast future stock prices using historical stock price data. It is crucial to modify the LSTM's structure and the associated parameters. In order to quickly find better parameter combinations, this article employs random search. The author of this article used 1520 sets of data from May 1, 2013, to May 14, 2019, the closing prices of Bitcoin and Amazon stocks, respectively, that were sourced from Kaggle [9].

2.1. Dataset Preparation

2.1.1. Data Analysis

In the realm of investing, Bitcoin (BTC) and Amazon (AMZN) are regarded as significant assets. The first and most well-known cryptocurrency, Bitcoin, has a high degree of price volatility and is heavily impacted by investor sentiment, regulatory frameworks, and general world economic conditions. Because its market is open around-the-clock, unlike traditional stocks, it is a high-risk, high-yield investment option with extremely volatile stock prices. Fig. 1 illustrates how the price of bitcoin has increased quickly from $1,000 since 2017. It peaked at 18,972.32 on December 18, 2017, and then fell sharply until 2019, when it started to gradually rise again. Meanwhile, one of the biggest online retailers in the world, Amazon is also a leader in cloud computing, AI, and digital entertainment. Over the past few years, the company's stock value has increased steadily, which is indicative of its strong performance and profitability across several categories. Numerous factors, such as conventional macroeconomic factors, corporate earnings reports, industry competition, and strategic mergers and acquisitions, have affected Amazon stock, which has generally shown a flat trend. The overall Amazon trend is depicted in Fig. 1 as slowing down and peaking at 2039.51.

Following standardization, Fig. 2 unequivocally demonstrates that the stock price of BTC is significantly more volatile than that of AMZN. This implies that, within the same LSTM model framework, various standardization and normalization techniques will yield distinct outcomes, and that this particular method can preserve this volatility attribute. Results from a standardized approach with Maxminscaler can be more favorable.

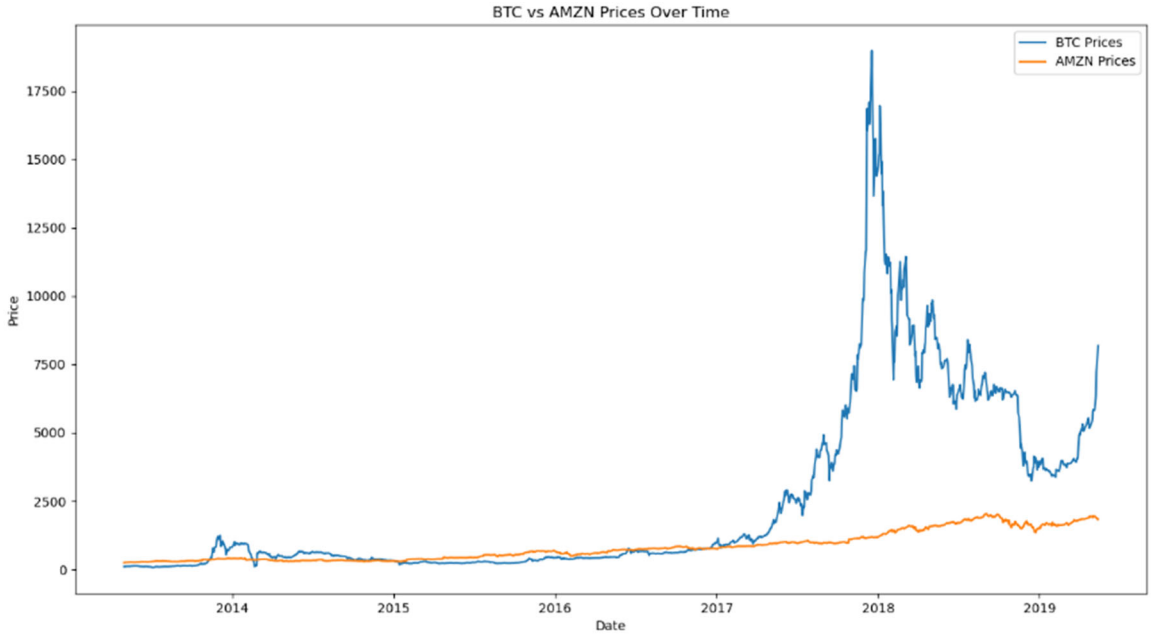

Figure 1. BTC vs AMZN Prices Over Time (Photo credit: Original)

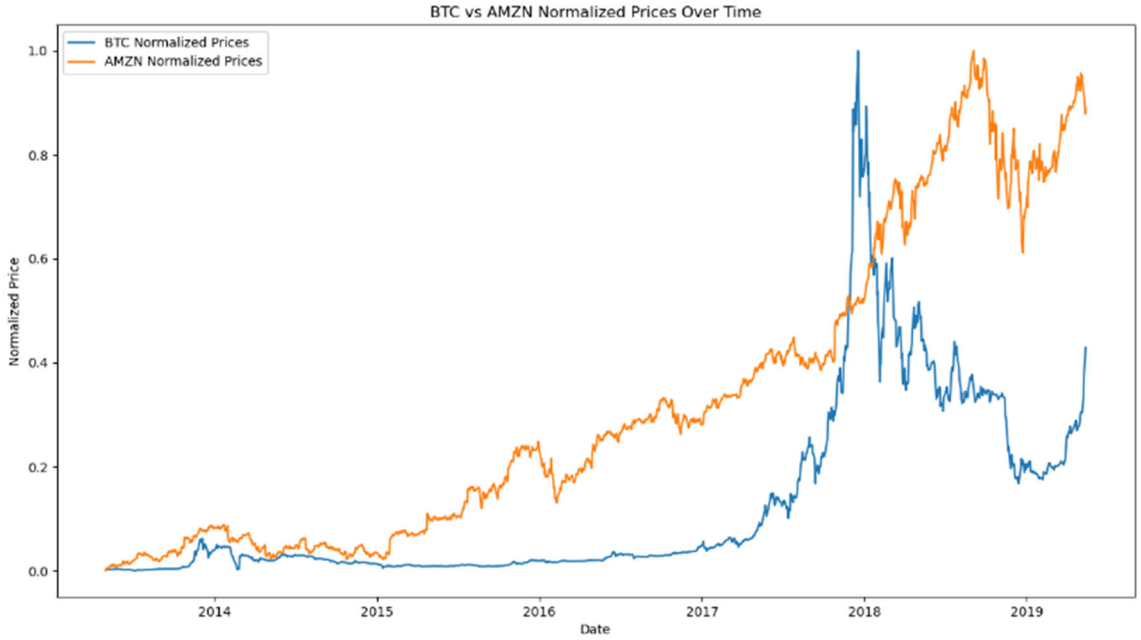

Figure 2. BTC vs AMZN Prices with Maxminscaler Over Time (Photo credit: Original)

2.1.2. Data Preprocessing

The fundamental principle of min-max normalization is to rescale the original data to fall within a predefined range and to scale it linearly to a specified minimum and maximum value, typically [0,1]. This ability to scale the data to a specific range aid in improving and accelerating the gradient descent optimization algorithm's convergence process, but it is vulnerable to the influence of outliers, which reduces the scaling range of other data. Nonetheless, it is appropriate for data that is roughly distributed within a small within the parameters of the situation, shown in Equation (1) below.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \times (max' - min') + min' \quad (1)$$

The original data is scaled to the range [-1,1] by absolute maximum normalization. The data is scaled using this method in accordance with the data's highest absolute value rather than being

centralized. It can handle both dense and sparse data without distorting the data's scale, shown in Equation (2) below.

$$x' = \frac{x}{\max(|x_{max}|,|x_{min}|)} \tag{2}$$

When using Z-Score standardization, the original data can be transformed into a standard normal distribution, with a mean of 0 and a standard deviation of 1. Z-Score normalized data is insensitive to outliers in the source data since outliers have a negligible effect on the mean and standard deviation. especially helpful for data that exhibit normal distributional properties, shown in Equation (3) below.

$$z = \frac{x-\mu}{\sigma} \tag{3}$$

Based on the median and interquartile range (IQR), robust standardization is a data scaling technique. It processes the data without converting it to a fixed range and is insensitive to data outliers. The initial data can be kept. Features of larger and smaller values, shown in Equation (3) below.

$$r = \frac{x-Q_2}{Q_3-Q_1} \tag{4}$$

2.2. LSTM Model

The Long Short-Term Memory Network, or LSTM model, is a unique kind of recurrent neural network (RNN) created especially to address the issue of standard RNNs' long-term dependency on lengthy sequences. By introducing gating mechanisms (forgetting gates, input gates, and output gates) that allow the network to learn long-distance dependencies over sequential data, it regulates the retention and forgetting of information. Because of this structure, LSTM performs exceptionally well in tasks like time series prediction and natural language processing that call for an understanding of long-term relationships between data points. The structure is shown in Fig.3.

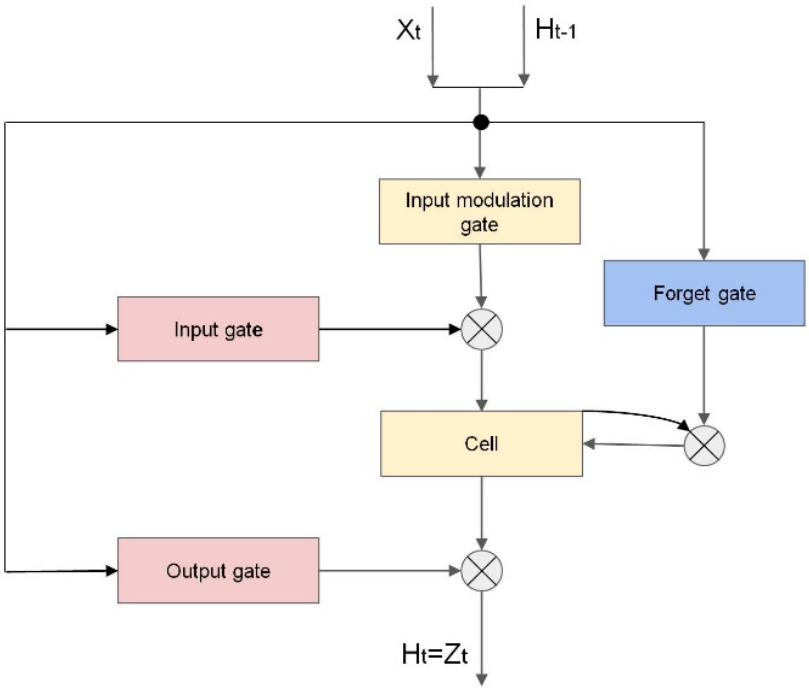

Figure 3. LSTM Structure [10].

Creating multiple LSTM models by randomly choosing different combinations of parameters within a variable hyperparameter space is the foundation of the stochastic search for dynamically

tuned LSTM structures. Predefined performance metrics (such as loss on the validation set) are used to assess each model. The random search finds the optimal structure and parameter configurations by comparing the performance of various models. This allows the LSTM network to be optimized automatically, without the need for manual tuning, to better fit a given time series prediction task. This method raises the likelihood of finding high-performance models while simultaneously increasing the efficiency of the model search process.

In this work, up to three dropout layers and three LSTM layers can be configured. Effectively discovering parameter combinations with greater accuracy in fewer trials can be achieved by conducting random searches for different batch size and LSTM structure combinations.

2.3. Implementation Details

This work used Early Stopping, a technique to avoid potential overfitting when employing stochastic search to dynamically modify the structure of LSTM models. A regularization strategy called "early stopping" is used to stop deep learning models from overfitting while they are being trained. It does this by keeping an eye on how well the model performs on the validation dataset. When the model no longer performs better on the validation set, Early Stopping will end training. In particular, Early Stop tracks a metric (such as loss on the validation set) and stops training if, after a predetermined number of training rounds (called "patience" in this paper; it is set to 10), the metric does not improve.

This work evaluated the prediction accuracy using Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE).

3. Results and Discussion

3.1. Results

Using the LSTM model, the following two stocks—BTC and AMZN—are projected to rise. The actual and predicted values are plotted in a single line chart, with the actual value shown in gray and the predicted value in blue, to make the prediction results easier to understand overall. The training set's data is represented by color, the verification set's data by green, and the test set's data by red shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7. The first 75% of the 1520 data points are the training set, the middle 10% are the verification set, and the remaining data points are the test set. There's a window of time between the two sets to stop data leaks.

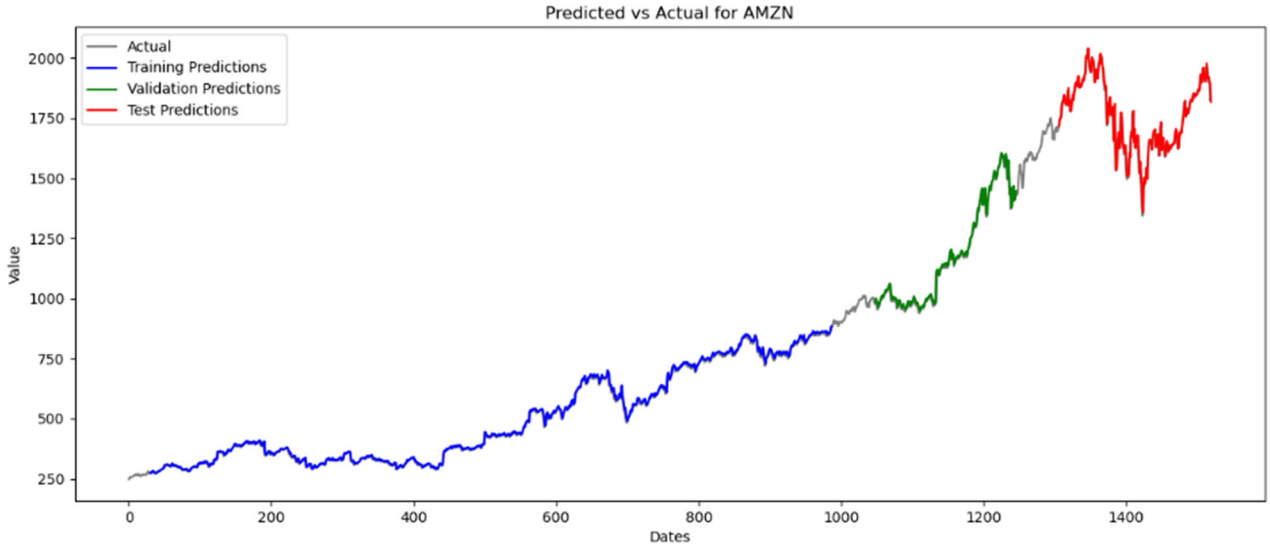

Figure 4. AMZN-Maxmin (Photo credit: Original)

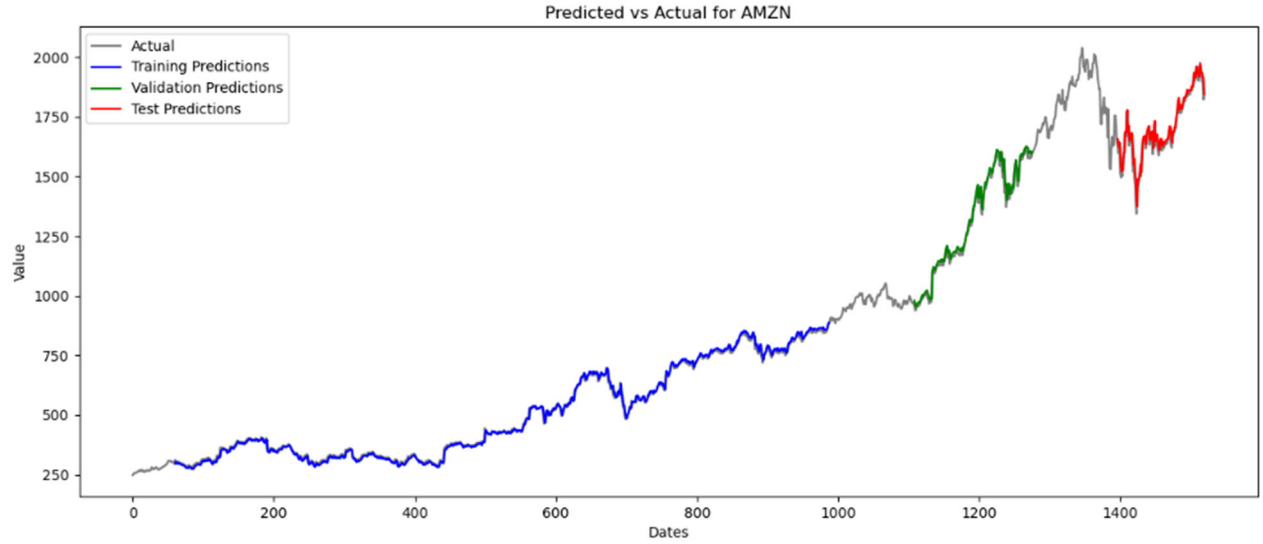

Figure 5. AMZN-Maxabs (Photo credit: Original)

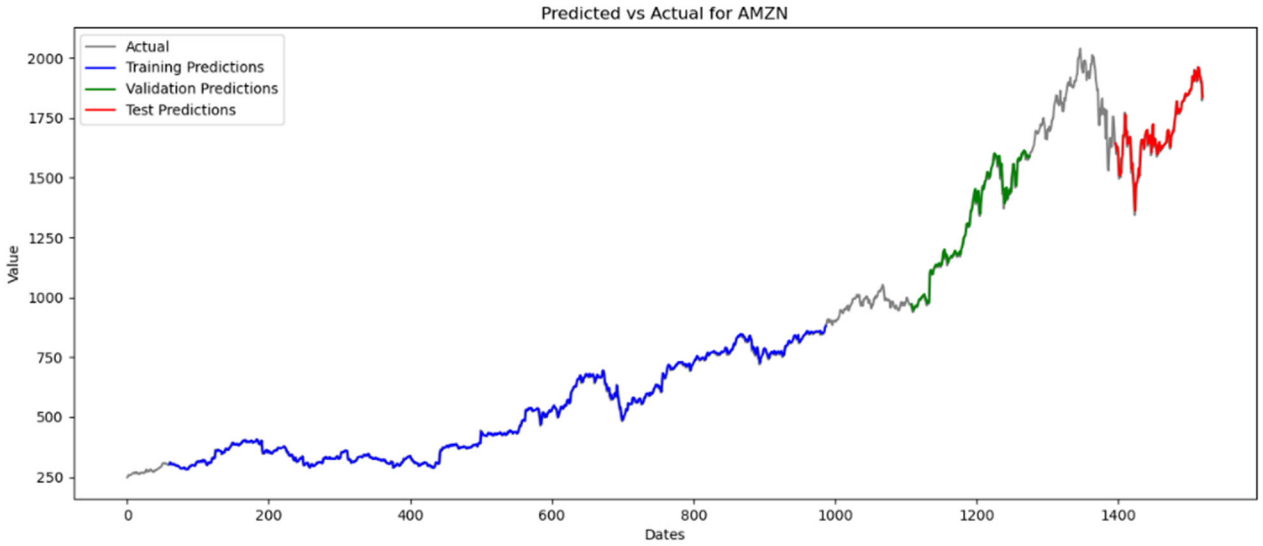

Figure 6. AMZN-Robust (Photo credit: Original)

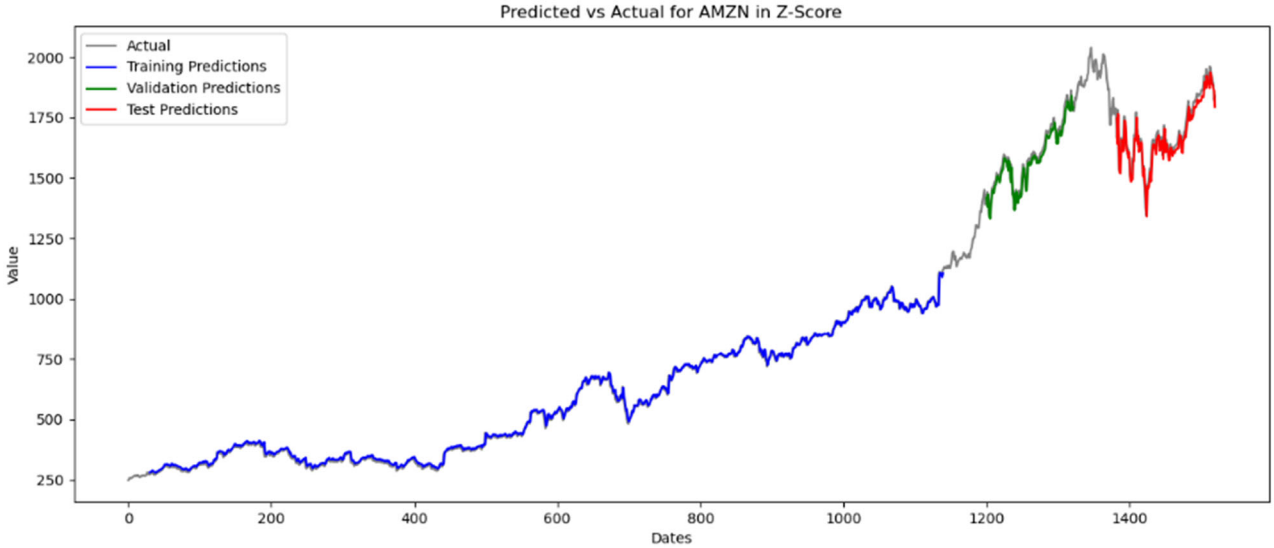

Figure 7. AMZN-Z-score (Photo credit: Original)

3.2. Results Analysis

When processing AMZN data in Table 1, the outcomes of MaxMinScaler and MaxAbsScaler are highly similar, suggesting that these two approaches function similarly. They have very close and relatively high accuracy, and their RMSE and MAPE values are both low. This demonstrates that when normalizing the data, maintaining the proportionate relationship of the original data is advantageous to the model's prediction effect. Out of the three methods, RobustScaler performs the worst, with lower accuracy, significantly increased MAPE, and higher RMSE. The reason for this could be that AMZN's stock change trend is relatively stable, which increases the accuracy of MaxMinScaler and MaxAbsScaler's prediction effects, or it could be that the outliers in AMZN's data set are not essential to the prediction task.

Table 1. Summary of equation of LSTM in AMZN test set.

AMZN-TEST	MAXMIN	MAXABS	ROBUST	Z-SCORE
RMSE	37.201	37.109	70.518	46.699
MAPE	1.631%	1.623%	3.663%	2.205%
ACCURACY	98.369%	98.377%	96.337%	97.795%

Table 2. Summary of equation of LSTM in BTC test set

BTC-TEST	MAXMIN	MAXABS	ROBUST	Z-SCORE
RMSE	232.78	228.201	215.966	222.963
MAPE	3.270%	3.252%	3.006%	3.173%
ACCURACY	96.730%	96.748%	96.994%	96.827%

RobustScaler offers the best overall predictive performance for the Bitcoin dataset. This is probably due to the fact that RobustScaler is more resilient to outliers and that extreme value changes are typically associated with Bitcoin price fluctuations, which is in line with the initial analysis of the overall data trends. Z-Score is a legitimate normalization technique even though it is marginally less accurate than RobustScaler; this is because its RMSE is quite similar to RobustScaler's. On this specific test set, MaxMinScaler and MaxAbsScaler perform poorly, particularly on the MAPE metric. This could indicate that the RobustScaler and Z-Score methods handle fluctuations and extremes in the data better than these two methods.

The findings highlight the significance of selecting appropriate normalization strategies for maximizing the predictive performance of LSTM models. They also suggest that RobustScaler could be a useful tool for handling financial time series data with extreme volatility and outliers. Furthermore, this study finds it can be found that dynamically modifying the lstm's structure through a stochastic search greatly aids in the prediction of stock prices and improves overall prediction accuracy.

3.3. Limitations

This experiment uses an LSTM framework to test various normalization techniques and predicts the price performance of AMZN and BTC. Although the results are insightful, the framework has certain limitations. First, it's possible that the experiment only used a smaller dataset, which would have limited how broadly the findings could be applied. Second, a broader range of data processing techniques and sophisticated predictive models were not addressed, and only a small number of preprocessing methods and a single model architecture were taken into consideration. Furthermore, no analysis of other performance metrics was included in the evaluation metrics, which instead concentrated on RMSE, MAPE, and accuracy. Lastly, because the stock market is an extremely intricate chaotic system, the model ignores outside variables that influence the financial market's volatility, such as economic indicators and market sentiment, which may affect the accuracy of the forecasts.

4. Conclusion

By contrasting the performance of two stocks, AMZN and BTC, this study observes significant variations in data preprocessing outcomes based on the utilization of different normalization or standardization techniques. Robust techniques demonstrate enhanced accuracy for highly volatile stocks, whereas standardization techniques prove more effective for stable stock prices. Furthermore, it is noted that a single stock can achieve commendable prediction results after a certain number of search rounds when employing stochastic search to dynamically adjust the LSTM structure. However, the stock market is a highly intricate and volatile system. The stock prediction only takes into account a portion of the rules found in the time series of a single stock. The impact of a single stock on the overall stock market index must be taken into account in the next stage of this work. The impact of other factors, like news, can be taken into account more thoroughly and methodically.

References

[1] Zhou, F., Chen, X. D., Zhong, T., et al. Overview of deep learning technologies for financial technology. Journal of Computer Science, 49 (S2), 20 - 36, 2022.

[2] Yadav, A., Jha, C. K., Sharan, A. Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, 2091 - 2100, 2020.

[3] Lam, M. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decision Support Systems, 37 (4), 567 - 581, 2004.

[4] Hochreiter, S., Schmidhuber, J. Long short-term memory. Neural Computation, 9 (8), 1735 - 1780, 1997.

[5] Persio, L. D., & Honchar, O. Artificial Neural Networks architectures for stock price prediction: comparisons and applications. International Journal of Circuits, Systems and Signal Processing, 10, 403 - 413, 2016.

[6] Bao, Z. S., Guo, J. N., Xie, Y., et al. Stock price prediction model based on LSTM-GA. Journal of Computer Science, 2020.

[7] Chen, J. Analysis of bitcoin price prediction using machine learning. Journal of Risk and Financial Management, 16 (1), 51, 2023.

[8] Srivastava, P., Mishra, P. K. Stock Market Prediction Using RNN LSTM. In 2021 2nd Global Conference for Advancement in Technology (GCAT) (pp. 1-5). IEEE, 2021.

[9] AMZN, DPZ, BTC, NTFX adjusted May 2013-May2019, kaggle.com.

[10] Thethi, J. K., Pandit, A., Patel, H., et al. Stock Market Prediction and Portfolio Management using ML techniques. International Journal of Engineering Research Technology (Ijert) Ntasu-2020, 9 (03), 2021.