

A Comparison of Linear Regression, LSTM model and ARIMA model in Predicting Stock Price A Case Study: HSBC's Stock Price

Shiyue Kuang*

School of Economics and Management, South China Normal University, Guangzhou, China

*Corresponding author: 20200741041@m.scnu.edu.cn

Abstract. It is widely assumed that the stock market is a key aspect of the financial market and stock price forecasting has become a popular topic of an in-depth study by financial technologists. With the advent of financial markets, stock prices face several analytical and forecasting challenges. In this paper, the stock price trend forecast analysis is carried out and the survey object was the transaction data of HSBC from 2010 to 2019. Linear regression, LSTM model and ARIMA model are used to forecast stock price trends. The models' prediction accuracy is demonstrated after the cross-validation process by combining error indicators and trading performance. As a consequence of the analysis, we discover that LSTM has the lowest error value, which implies that it is ideally suited to stock market forecasting. On the contrary, linear regression holds the largest mean square error, implying that this method has the weakest fitting for the stock market.

Keywords: Linear regression; LSTM model; ARIMA model; stock price prediction; HSBC.

1. Introduction

The research on stock price prediction has never stopped since the birth of the stock market. A number of forecasting systems emerged during this period. Machine learning algorithms have been introduced to the field of stock forecasting to achieve superior performance because current forecasting methods can only meet the requirements for accuracy to a certain extent. Machine learning algorithms are able to mimic the specific properties of items to the greatest extent feasible and it also has huge advantages in terms of volume and complexity of data processing and predictions. Stock market forecasting has become an important focus for investors and academics because a better understanding of the stock market can yield superior returns.

Experience shows that machine learning techniques can successfully predict daily stock prices and trading volumes [1].

It is the case that three types of models are used in this paper to do price forecasting. First of all, linear regression is a common mathematical research tool that allows you to assess and quantify projected effects given many input variables [2]. It's a data analysis and modeling technique that develops linear relationships between dependent and independent variables. From the analysis and learning to the current training results, this strategy would model relationships between dependent and independent variables. Applying linear regression to stock forecasting is fitting and forecasting based on the date and the daily closing price of the stock market.

What's more, the LSTM (Long-Short-Term Memory) network is a type of recurrent network that has been shown to be very efficacious on a variety of problems owing to its capacity to distinguish between recent and early examples by assigning specific weights to each while neglecting memory that it considers irrelevant to predict the next output. [3]. In this respect, it is more capable of handling long input sequences than other recurrent neural networks, which can merely memorize short sequences. Historic price data (candlesticks) from stock will be used as a source of information for the network, evaluated and will attempt to predict whether the price of a particular stock will go up or down.

Additionally, the Autoregressive Integrated Moving Average (ARIMA) is based on the ARMA model, which is widely used to predict linear time series data. As an essential model to study time series, the Autoregressive and Moving Average (ARMA) model combines the concept of

autoregressive (AR) and moving average (MA) models, which was formulated by the works of Yule, Slutsky, Walker and Yaglom[4]. In univariate time series model identification, parameter estimation, and forecasting, the ARIMA approach provides a great deal of flexibility [5]. Stock prices are not randomly generated numbers; rather, they can be thought of as a discrete-time series model, with the trend evaluated and predicted as such. Stock forecasting has a number of goals, one of which is monetary gain. Because it's critical to build a model that examines stock price movements with relevant data for decision-making, it's advised that converting time series using ARIMA is a better technique than forecasting directly, as it may yield more accurate findings.

After applying these three models, the accuracy of these predictions will then be assessed using cross-validation. Each model has its own set of disadvantages and advantages. The aim of this research is to compare the forecasting data and the actual historical one in order to determine which model fits with the most accurate trend or even specific data. Eventually, find the best one for the current stock market prediction.

2. Data Collection

2.1 The stock HSBC

For the selected stocks, the dataset in this paper comes from the stock price of HSBC (Hongkong and Shanghai Banking Corporation Limited). As it is one of the world's largest banking and financial institutions, the Hongkong and Shanghai Banking Corporation Limited is a huge financial group institution whose business scope covers more than 100 countries and regions around the world. As a result, HSBC's share price tends to fluctuate regularly and cyclically and is hard to be influenced by extreme situations or market crises. It provides data that is relatively consistent with common stocks, which is an advantage for generalizing the results of these models.

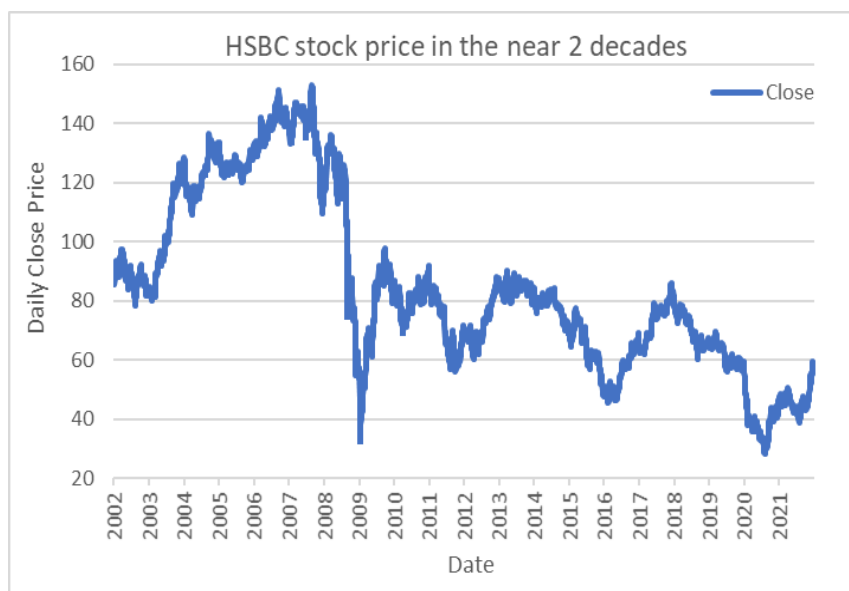


Figure 1. HSBC stock price in the near 2 decades

Taking the 2008-2009 financial crisis and the COVID-19 pandemic since March 2020 as examples, the price movement of this stock is indeed a good reflection of the major changes and dramatic fluctuations in the world financial and economic markets. Conversely, during the 2010s, this stock's price can fluctuate relatively steadily and cyclically, indicating that its operations have some stability.

2.2 Data

The time range from 2010 to 2019 was chosen because no major events hit financial and economic markets during that time period. The global financial crisis (2008-2009) almost ended in 2010 and the

rush of COVID-19 (widely spread since March 2020) had not occurred then. As shown in Figure 1, the HSBC stock price trend has obvious periodicity and regularity from 2010 to 2019, which makes the model easier to build.

Before applying models, it is necessary to do some data processing. In this stage, the historical stock price data of HSBC from 2010 to 2019 is collected from Yahoo Finance (<https://www.yahoo.com/finance>). The data set includes feathers 'Data', 'Open', 'High', 'Low', 'Close', 'Adj Close' and 'Volume', which is ready to be used for the prediction of future stock price.

3. Methodology

3.1 Data Preprocessing

- Data Cleaning

There is some missing value in the dataset that needs to be filled in to assure the forecasting techniques in subsequent steps can be performed properly. In order to ensure data continuity, the stock price and other relevant feathers of the former day are chosen to put in the blank.

- Setting Training and Testing Data Set

After converting the dataset to a clean dataset, the dataset is divided into training and testing parts. From 2010 to October 2019, a hundred and twenty months of data have been taken. The former 118 months' stock price is set to be training data to predict the last two months' data in 2019, which assure there is enough data to train.

- Feature Extraction

In this stage, the feature from the data set 'Date' and 'close' will be chosen. It is widely acknowledged that the closing price of the stock market is a both useful and vital piece of information for every short-term trader [6]. To some degree, close prices are even extremely significant for swing and position traders alike. In many day trading systems, it also has ramifications for practical day trading. The closing price level of the stock market offers critical information regarding investor sentiment. It reveals a lot about the mindset of huge investors who invest significant sums of money in the stock market for asset management objectives. Therefore, we mainly use the close price of the stock on that day to do the following prediction.

3.2 Start Modeling

Then the processed data set is used to define models with three methodologies and the test set is used to validate whether the models are fit for prediction. After that, the defined models are used to predict stock prices for the next 2 months.

3.3 Models

- Linear Regression

John Wiley & Sons have already study about the regression model whose research states the opinion that regression analysis is a statistical technique for investigating and modeling the relationship between variables [7]. It is the foundation of data science and analytics, which is broadly applicable to a wide range of problems including stock price forecasting.

$$y = \beta_0 + \beta_1 x + \epsilon$$

The equation above is a linear regression model. Take this equation as an example, x is called the independent variable and y is called the dependent variable. β_1 is the slope parameter and β_0 is the intercept parameter. ϵ refers to the error term or disturbance, which is a random variable that indicates that the model cannot accurately fit the data.

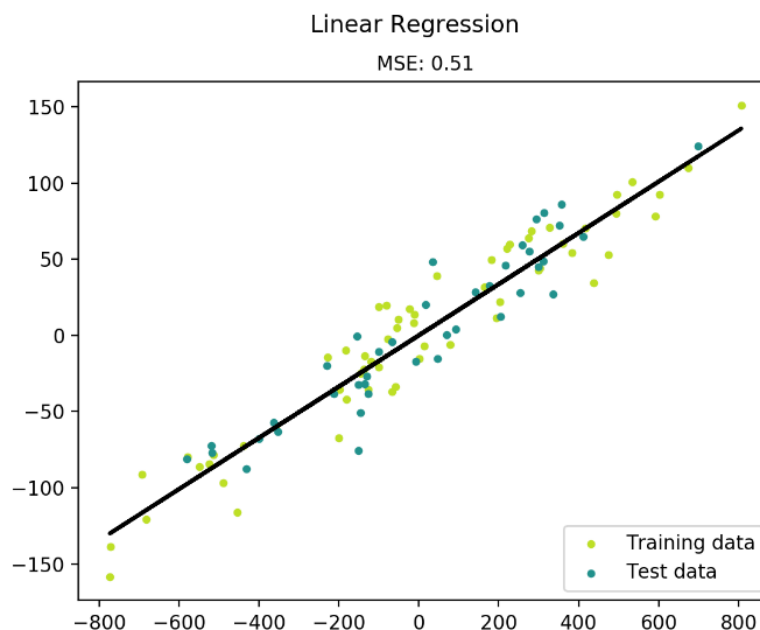


Figure 2. A sample of linear regression

Applying in the study, x will be the date (range from 2010 to 2019) and y refers to stock prices of different dates. With the help of linear regression models, it is easy to use years of historical data to roughly characterize stock price movements and predict the next day's value. (Figure 2 source: <https://zhuanlan.zhihu.com/p/32600338>)

- LSTM

In a previous study, it is justified that Recurrent Neural Networks (RNNs) is one of the most powerful models for processing sequence data. In addition, the LSTM model is even one of the most successful RNN architectures [8]. It is essentially an RNN for processing time-series data and it even improves the accuracy of previous models while overcoming some of the shortcomings of the RNN model.

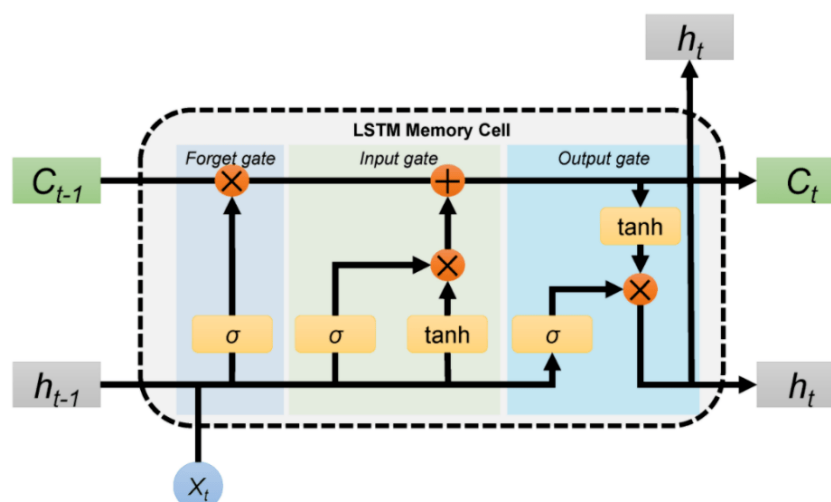


Figure 3. Procedure of the LSTM model

In the buried layer of the network, LSTM introduces the memory cell which is a computational unit that substitutes typical artificial neurons. Then, networks can efficiently correlate memories and input remote in time with these memory cells, allowing them to understand the structure of data dynamically through time with great prediction capacity. After that, the cells resemble a transport line (the top line in each cell) that connects one module to the next, transporting data from the past and gathering it for

the present. Data in each cell can be disposed of or filtered due to the use of some gates in each cell. (Figure 3 source: <https://www.sohu.com/picture/443110431>).

- ARIMA

The Autoregressive Integrated Moving Average model (ARIMA) is a generalized autoregressive moving average (ARMA) model that creates composite time series models by combining autoregressive (AR) and moving average (MA) methods. Although the word "different" does not appear in ARIMA's English name, it is a vital component.

The ARIMA (p, d, q) model is an extension of the ARMA (p, q) model. the ARIMA (p, d, q) model can be expressed as:

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

where L is the lag operator.

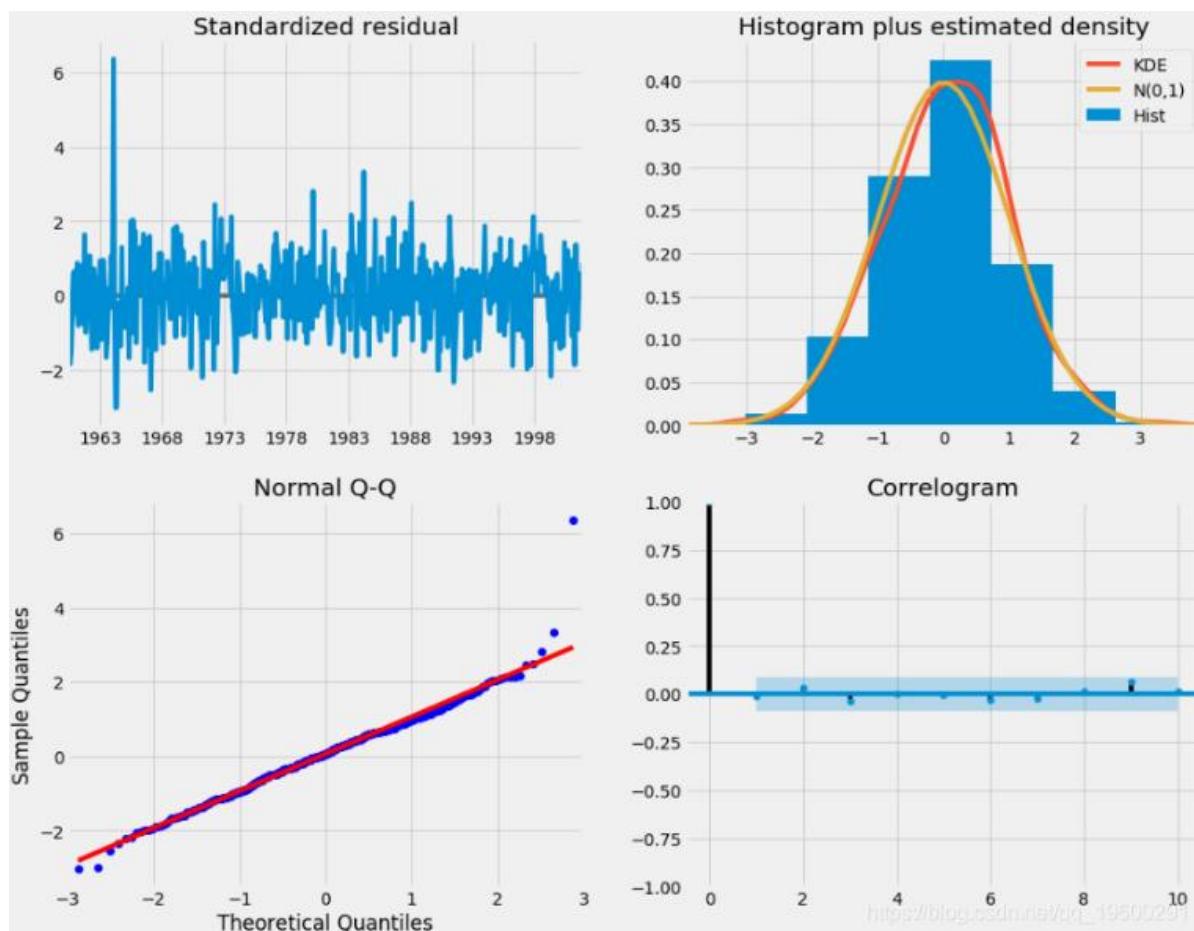


Figure 4. A sample of processing and graphing data with the ARIMA model

ARIMA predicts that a non-stationary problem is transformed into a stationary problem through multiple differencing, so there are specific requirements for the data owing to the difference. Due to the presence of the difference process, the temporal data dispersion increases, and the ARIMA model's fit is usually poor. The ARIMA model does not always suit well [9]. (Figure 4 source: https://blog.csdn.net/qq_19600291/article/details/113939485).

- Cross-validation

As a measure to evaluate and compare learning algorithms, cross-validation is a statistical method that divides data into two segments and detects the goodness-of-fit of the model through the sum and average value of errors. The data using cross-validation will generally be split into two parts: one for learning or training and the other for validating the model. In classical cross-validation, the training

and validation sets must cross over in successive rounds so that each data point can be validated against each other [10].

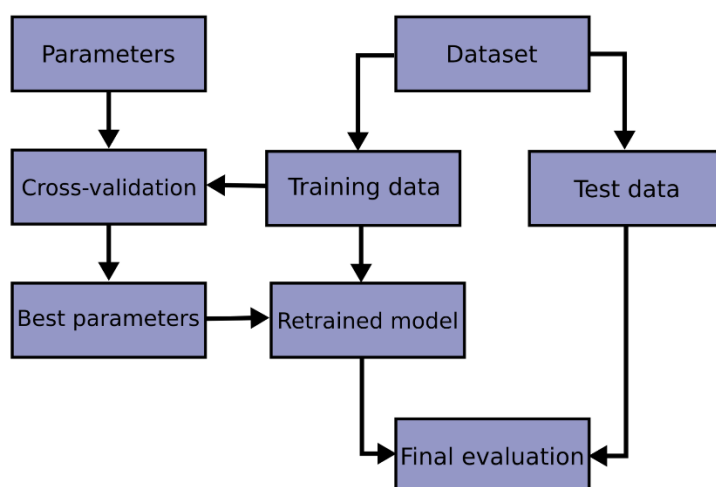


Figure 5. The procedure of cross-validation

In the data set, the errors of the training set and the test set are found one by one. Eventually, these errors are averaged to obtain the error of the model.

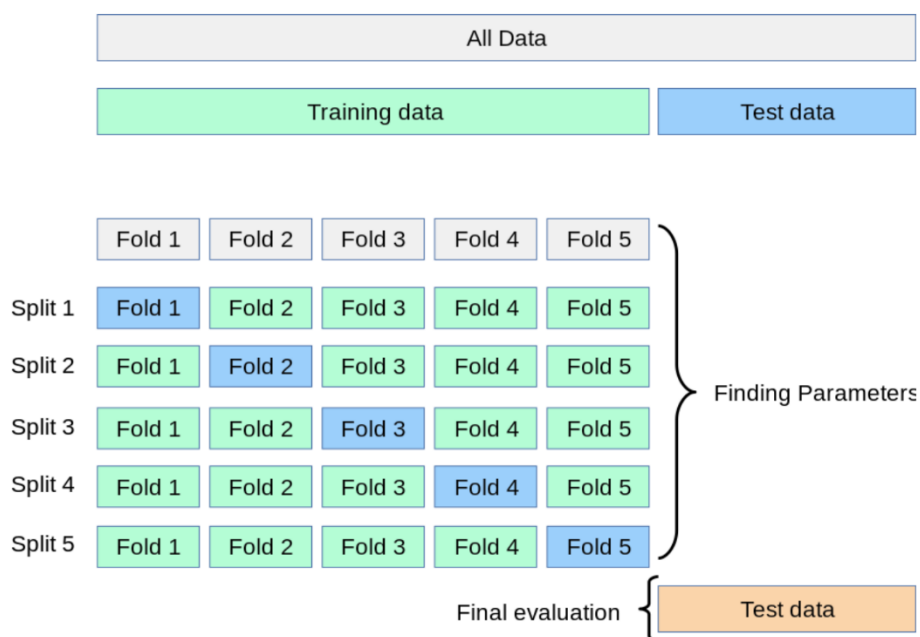


Figure 6. Dividing data, getting the errors and calculating its mean value

Cross-validation will be performed in this paper to measure model accuracy and identify which model is best for predicting the stock price with the smallest error value.

(Figure 5 and Figure 6 source: <https://blog.csdn.net/iterate7/article/details/102139750>)

4. Results and Discussion

4.1 General Tendency

The ultimate outcome is to plot graphs after training, modeling, testing, and estimating. As shown in Figure 7, it is the predicted result of test data, forecasting the future stock price of HSBC during November and December 2019.

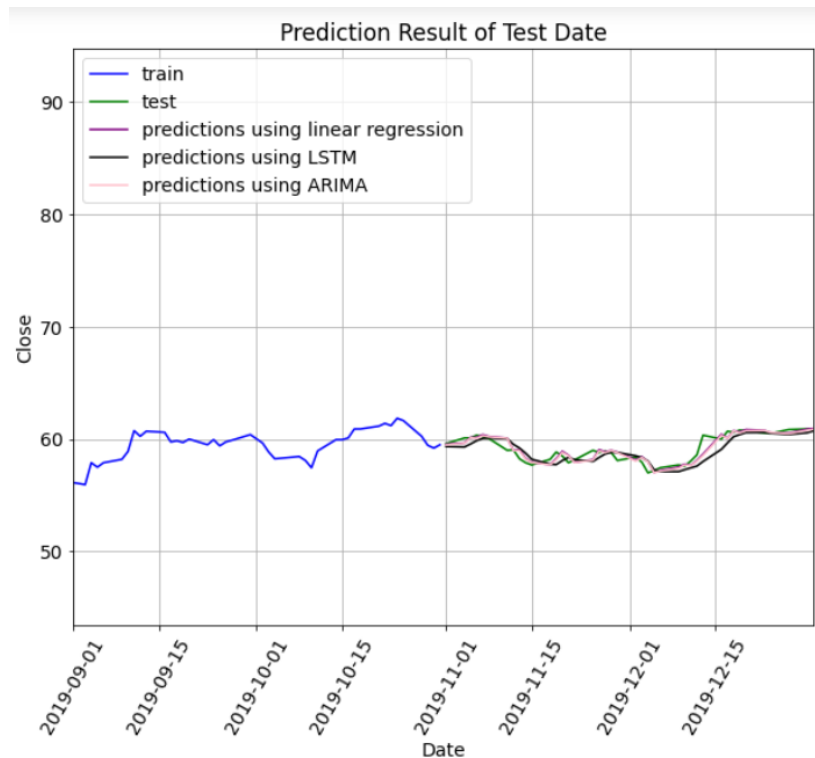


Figure 7. Picture of the prediction using 3 models

In general, the three estimated curves are slightly different from the real one. They fluctuate up and down around the curve that depicts real data. It is worth noticing that during the first week of November and the data after December 20th, the estimated curves almost perfectly match the actual one. That means when the curve flattens out, these three models are able to forecast more accurate future data. Then comes the conclusion that the time series models are likely to perform better and the result will be more precise when the data is relatively smooth. This could be because the models can capture the pattern of changes when there isn't much diversity in the train set.

Additionally, the stock price increased rapidly from December 10th to December 15th. Here comes the problem that the prediction models are seemingly unable to keep up with the sudden change. The LSTM model is thought to have the weakest fit since the black curve is the furthest away from the green curve that represents the real stock price. And the pink and purple curves nearly overlap, demonstrating that the linear regression and ARIMA models are nearly identical in terms of fit. To conclude, in the presence of rapidly changing data, the fitting accuracy of all three models declines, with the LSTM being the weakest fit. Besides, the linear regression model fits best in this case, because the purple curve is slightly closer to the green line than the pink line.

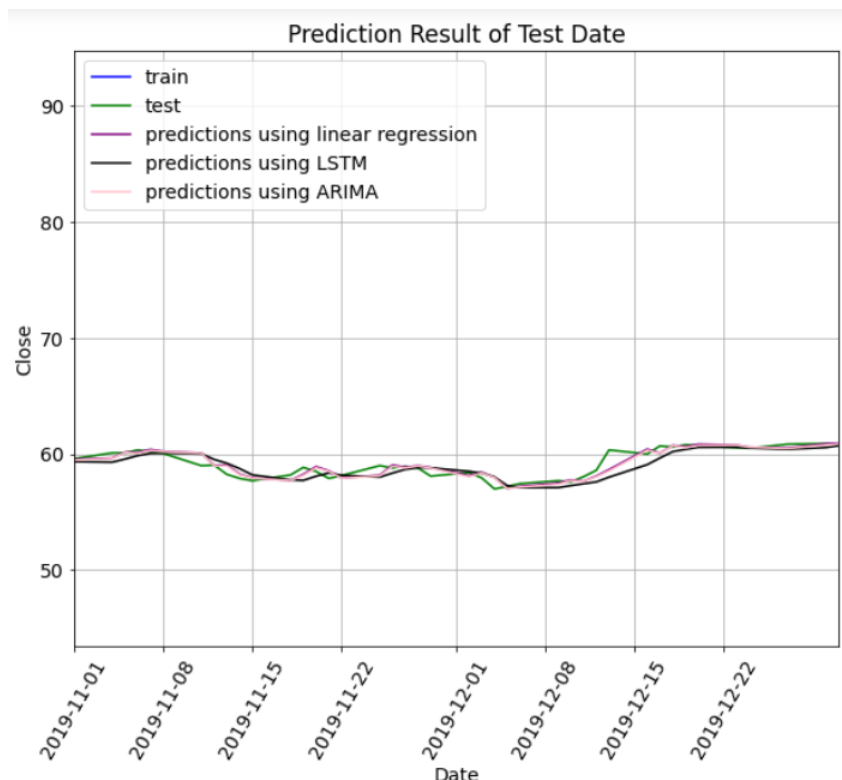


Figure 8. Specific Estimated Data

Furthermore, to gain a better grasp of the forecasting data, it is beneficial to obtain more analysis in Figure 8. The graphic is very similar to Figure 7, but it contains only the predicted data and the timeline is stretched longer.

As Figure 8 depicts, the estimated curves always trail behind the actual curve in terms of ups and downs. This situation is particularly obvious in the data during late November. There is no denying that forecasting lag is unfavorable to investors who make decisions. Especially when there are frequent and tiny fluctuations in stock prices, such a lag is likely to cause investors to make incorrect judgments and even make investment decisions that are completely contrary to the trend of stock market developments.

4.2 Judging Accuracy with RMSE

Generally, with the naked eye, it can be tricky to determine which model is the best forecasting as these lines seem to be heading in almost the same direction. In addition, every model makes accurate or incorrect predictions at times. Therefore, it is necessary to use other techniques such as cross-validation to conduct a more thorough investigation.

Table 1. RMSE Estimated by Cross-validation

	LR_RMSE	LSTM_RMSE	ARIMA_RMSE
fold1	1.534607	0.036764	1.237239
fold2	1.032934	0.03238	0.833236
fold3	0.94388	0.022922	0.743819
fold4	0.927977	0.033758	0.750949
fold5	0.850113	0.019453	0.683137

As shown in Table 1, after cross-validation, the fitting errors of the three models for the real data were reported as root mean square error (RMSE). The number of validations is given in the first column of indicators. The more you train the data set, the more likely you are to make smaller errors.

When it comes to the result shown in the table, the LSTM has the smallest RMSE, which is far lower than the others. ARIMA's RMSE is almost 40 times that of LSTM. The Linear Regression ranked last, with an estimated error that was nearly 25% greater than the ARIMA's.

Similar to the results obtained in the previous section 'General Tendency', the outcomes of RMSE values about ARIMA and linear regression are highly comparable, which means these two models have the prediction of the stock price with similar accuracy. As a consequence, the LSTM model is the best model for predicting future stock price, whereas, with a slightly larger RMSE value than the ARIMA model, a linear regression model is the worst when forecasting.

4.3 Comparison of curve shapes

To get a more detailed observation of the predicted stock price curves, the range of vertical coordinates in Figure 8 is restricted and shown in Figure 9.

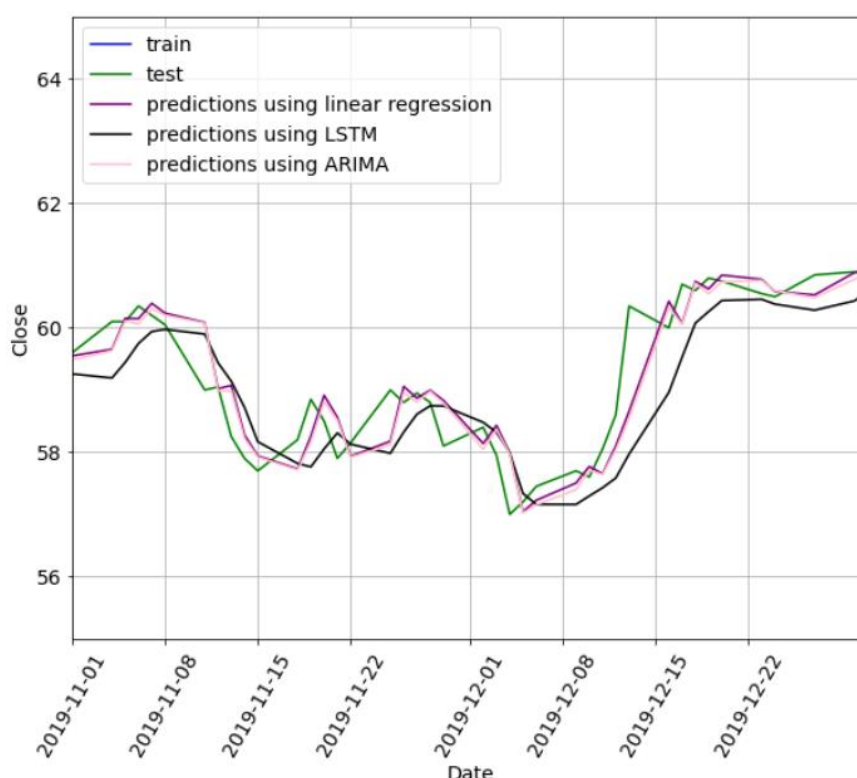


Figure 9. Forecast stock price chart after narrowing the range of vertical coordinate

As it is displayed in Figure 9, the curves overlap less and are easier to spot in detail after the vertical coordinate range is narrowed. As can be observed, not only the highs and lows of the linear regression model and ARIMA model are more pronounced, but also the extremes of them are more conspicuous, resulting in the two predicting curves being more similar in length to historical data.

However, the LSTM model predicts a somewhat more consistent and smoother stock price curve in which the maximum and minimum values do not stand out. It seems to portray merely an approximate price trend rather than a real day-to-day stock market movement. Therefore, although the LSTM model predicts data with minimal error, it is not a good choice for real-time or time-scale precision stock market forecasting and related investments as there is no obvious representation of extreme values. Overall, despite the LSTM model can forecast stock prices with the smallest error, the best scientific approach to investment decisions in the short term might be to employ ARIMA models to produce projections and then apply the forecast data, which depicts a certain future time range, to current stock market investments.

4.4 Figures and Tables

In this paper, the stock price of HSBC is applied to the three methods, linear regression, LSTM model and ARIMA model. As a highly integrated and broad-based financial group, HSBC's stock market price changes largely reflect the movements of the world's finance and economy. The prediction of HSBC's future movements is of great reference value to institutions in the related financial, banking or other relevant industries. Additionally, it is also of great guidance to individuals or institutional investors in their stock market investment decisions.



Figure 10. Close Price of HSBC stock during 2010-2019

As it is shown in Figure 10, in the time range from 2010 to 2019, HSBC was experiencing a relatively smooth cyclical fluctuation development, which theoretically made it easier for the models to capture the pattern of stock price changes.

However, for the linear regression model, trying to capture the pattern of change through simple regression and making predictions still seems to be a challenging task. Because there are so many factors that influence HSBC's stock prices, it's tricky for the linear regression method to account for almost all of the elements in a more comprehensive way. When it comes to the ARIMA model, it works best with data sets that have fairly smooth forecasts. Data that is extremely volatile, such as the stock prices movements of HSBC, is not ideal for this method. As for the LSTM model, there is no denying that it fits the best with the smallest RMSE. This is certainly the best choice among these three models for stock valuers in a more stable market, such as the data set of HSBC's stock price from 2010 to 2019.

5. Conclusion

5.1 General Models vs. Deep Learning Models: Deep Learning Model is Better

It is acknowledged that the general model has the ability to be both very easy to code and flexible enough to cover a large range of different variables through a simple structure. Besides, it has a wide range of applicability and a large number of off-the-shelf models and algorithms than deep learning models [11], which enables people to apply the ready-to-use models to a variety of investment circumstances. However, the general model does not account for a number of aspects that can influence the forecast findings. Some variables are difficult to capture by the model and some are directly ignored by the model. This can lead to significant inaccuracies and poor fitting.

5.2 LSTM vs. ARIMA: LSTM Wins the Prediction

The Namin's have already published a paper on the accuracy of LSTM and ARIMA when evaluating both financial and economic data. They found the RMSE data reveal that LSTM-based models beat ARIMA-based models substantially [12]. With ARIMA's RMSE almost 40 times that of LSTM, the results in this paper also confirm this conclusion.

When comparing ARIMA and LSTM models, both were applied to time series data to predict future stock prices, and their accuracy was evaluated as a predictive time representation algorithm for time series data. However, LSTM outperforms ARIMA for various reasons:

LSTM is a terrific RNN variant that inherits most of RNN's characteristics and solves the gradient disappearance and gradient explosion problems in the gradient backpropagation process, making it ideal for dealing with highly correlated time series problems like the stock prediction problem in this paper. In addition, it is believed that LSTM models are well suited for problems involving time series [13]. When used to deal with multiple regression, the most powerful property of LSTM is that it is resistant to correlations between variables, covariance and nonlinearity of variables.

However, on contrary, while the ARIMA model has the benefit of being simple and requiring just endogenous variables rather than exogenous variables, the data needs to be stable when using the ARIMA model to predict time series data. Otherwise, capturing the pattern is impractical. In this case, stock prices are not an acceptable dataset for ARIMA in this case because stock market conditions are constantly changing, causing stock prices to fluctuate.

References

- [1] Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.
- [2] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1 (4), 140 - 147.
- [3] Nelson, D. M., Pereira, A. C., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In 2017 International joint conference on neural networks (IJCNN) (pp. 1419-1426). IEEE.
- [4] Chen, S., Lan, X., Hu, Y., Liu, Q., & Deng, Y. (2014). The time series forecasting: from the aspect of network. *arXiv preprint arXiv: 1403.1713*.
- [5] Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4 (2), 13.
- [6] Seethalakshmi, R. (2018). Analysis of stock market predictor variables using linear regression. *International Journal of Pure and Applied Mathematics*, 119 (15), 369 - 378.
- [7] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.
- [8] Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*, 6 (4), 1754 - 1756.
- [9] Jarrett, J. E., & Kyper, E. (2011). ARIMA modeling with intervention to forecast and analyze Chinese stock prices. *International Journal of Engineering Business Management*, 3 (3), 53 - 58.
- [10] Refaeilzadeh, P., Tang, L., Liu, H., Liu, L., & Özsu, M. T. (2009). Encyclopedia of database systems. In Cross-validation (pp. 532-538). Springer.
- [11] Gharehchopogh, F. S., Bonab, T. H., & Khaze, S. R. (2013). A linear regression approach to prediction of stock market trading volume: a case study. *International Journal of Managing Value and Supply Chains*, 4 (3), 25.
- [12] Siامي-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv preprint arXiv:1803. 06386*.
- [13] Wenjuan Ding. (2021). Comparison of ARIMA Model and LSTM Model Based on Stock Forecast. *Industrial control computers*, 34 (7), 109 - 112.