Evaluation of two models for predicting Amazon stock based on machine learning

Zixuan You

Department of Science, University of Sydney, Sydney, NSW
*Corresponding author: zyou2909@uni.sydney.edu.au

Abstract. With the fast growth of artificial intelligence and technology, the use of machine learning techniques in financial markets is gaining popularity. As a result, many opportunities arise, such as predicting future stock movements. Financial markets are complex and constantly evolving environments, so analyzing them can be challenging and interesting. There are no specific rules to predict or estimate the value of a stock in the stock market, so one can do stock price prediction by various methods. In this project, the stock price data of Amazon for the past five years, ‘Date’, ‘Starting price’, ‘Closing price’, ‘Highest price’, ‘Lowest price’, ‘Adjusted close price’, and ‘Volume’ are all included. The data were obtained from Yahoo Finance to predict the future stock price. Two forecasting models, the Linear Regression model, and the Long short-term memory model were analyzed, and based on the comparison of Mean Absolute Error, it was concluded that the LSTM forecasting model was shown to be more effective for forecasting the time-series data category.

Keywords: Machine Learning; Python; Amazon.

1. Introduction

By anticipating market values that are near tangible values, machine learning enhances accuracy. The dataset utilized is critical. Yahoo Finance provided the dataset. Supervised machine learning was applied in this dataset. The dataset contains five variables: open, close, low, high, and volume. They are the various purchase prices of the stock at various times. Separate regression and LSTM models were used. The accuracy of the two prediction models is compared in the study, and the methodologies and practicality of predicting stock price fluctuations are discussed.

The following is how this document is structured. Section 2 presents an overview of Amazon.com, the corporate background, and describes its stock price data for this year; Section 3 discusses data and model selection, performs data preprocessing, and describes forecasting methods and structural analysis to compare two forecasting models; Section 4 discusses and analyzes the conclusions obtained, comparing the advantages and disadvantages of the LR model and the LSTM model; the last section presents the conclusions of the paper [1].

2. Firm Description

2.1 Background Information

In the early years of its existence in the 1990s, Amazon became the number one book site by selling books over the Internet at a very fast pace. Because of the Internet, Amazon, which was very small at the time, was able to ship books to a wide range, fulfilling a large number of orders from the United States and other countries in a month. After two years of improving the site, Amazon officially became a public company in 1997 and continued to expand in 1998, entering the music industry. 1999 Amazon opened an online auction service and began selling toys and electronics, but with fierce competition from many start-ups, Amazon's stock price continued to fall. After firing thousands of employees, closing distribution facilities, and signing third-party deals with several well-known retailers, Amazon's strategy led to a sustained rise in sales beyond the reach of other companies. 2002 saw Amazon open clothing stores and finally make a net profit, and in the years that followed, Amazon acquired several industries, such as pharmaceuticals and smart homes, through the company's long-term strategy [2].
2.2 Stock Price Description

Yahoo.com provides data on Amazon's stock price [3]. The paper analyzed Amazon's stock data from 2017 to 2022. Figure 1 shows Amazon's stock price closing trend and the trading volume for the last five years. There was a considerable increase occurred from 2017 to 2018. However, between 2018 and 2019 Amazon's stock price first experienced a small decline and then rose back to a position comparable to that of 2018. The stock price trend from 2019 to 2020 tends to be flat. 2020 starts with a sharp rise in Amazon's stock price to nearly twice its original value, reaching a small peak, but the stock market then plateaus and trends downward. 2021 sees a near five-year peak in Amazon's stock price and then continues its slow decline, rises briefly in early 2022 before plunging, and then begins a gradual rise. Looking at the five-year closing price and volume charts, Amazon's overall share price is significantly higher in the pre-2020 period when it was lower than in the post-2020 period when it was higher. Higher stock prices mean lower trading volumes (Fig 1).
The MA is a popular technical indicator among investors. The 5-days MA is calculated by dividing the sum of the closing prices from the previous five trading days by five. The 5-day MA is more appropriate for short-term traders (Fig 2). In terms of its function, the 5-day MA has the impact of alerting changes, supporting and repressing the stock price. The MA calculates the average variation over time in a set of data. An uptrend in a MA may imply a rise in a security's price or momentum, whilst a downtrend would indicate a drop [4].

Although Amazon stock had risen sharply during the global covid-19 epidemic, unstable market conditions had caused it to lose most of its gains. After a sharp rally in mid-summer, the stock is still down more than 15% in 2022. But in Amazon’s earnings release on July 28, it reported higher-than-expected revenue and a positive forecast for the third quarter. Most analysts, as well as investors, believe that Amazon will show extremely high potential in the long term.

3. Methodology and Result Analysis

3.1 Stock Price Description

3.1.1 Strategy and Selection Criteria

This prediction model is a prediction study of Amazon stock price based on machine learning, aiming to explore the application of machine learning models on the stock price prediction. Two models are used in the article to forecast Amazon's stock closing price (Fig 3).

![Figure 3. Observation of Amazon Stock Closing Price](image)

3.1.2 Data Selection

As the analysis data, this study gathers Amazon.com trade data from Yahoo.com for a total of 1258 trading days from 2007 to 2022. The daily indicators of the opening price, highest price, lowest price, closing price, and volume primarily reflect the stock trading situation, so the article selects six indicators that can reflect the trading situation and uses the data situation of the six indicators for the first 1058 days to estimate the closing price for the next 200 days. The six indicators for each day from the first to the tenth trading day are chosen as the model's independent variables, and the closing price for each day from the tenth to the tenth trading day is chosen as the model's dependent variable. The six indicators for each day from the 1st to the 1058th trading day are chosen as the model's independent factors, and the closing price for each day from the 1059th to the 1258th trading day is
chosen as the model’s dependent factors, as shown in the chart below for the daily closing prices of Amazon stock over the last five years [4].

3.1.2 Data Preprocessing

When training a model, a common thing to do is to preprocess the data in order to make the model converge as soon as possible. The article normalizes the data by using the skl preprocess module when processing the data. MinMaxScaler has an important parameter, feature_range, to control the range to which the data is compressed, [0,1] [5].

3.1.3 Model Selection

Linear Regression The price and time period of a stock determines the systematic parameters of linear regression. Accurately predicting the price of a stock is something that many people seek to do. The accurate prediction may not be practical, but even simple linear models can sometimes be very close. This article discusses how to use Python's linear regression model to train a regression model to predict future stock prices using previous stock price data and technical indicators. First, load the historical stock price data into a Data Frame in Pandas and add technical indicators to be used as features in our linear regression model. Next, extracted the data intended to use from the Data Frame. Here the "date" column is converted to a DateTime index and plotted with a visual representation of the open, high, low, and close data (Fig 4).

![Figure 4. Visualization of Amazon Stock Price](image)

The article then analyzes the accuracy of the model, plots the results, and considers the magnitude of the error. A new column is created, containing the forecast values, to predict the closing price for the next 200 days based on the previous data. Since 'label' and 'Adj_close' are not needed for forecasting, they are directly dropped and the data is divided into training and test data, after which the model is evaluated using the absolute coefficient R Squared. The prediction starts from February 12, 2022, with a total of 200 days of closing prices.

Long Short-Term Memory Machine learning algorithms are commonly used to forecast stock market direction, prices, or returns based on past data [6]. LSTM networks are widely used to deal with time series problems because of their special type of recurrent neural networks. They can identify long-term trends in serial data. LSTM networks can remember significant prior data while discarding irrelevant data. The article makes use of the expected daily closing price of each stock based on historical data. To ensure that the results are as reproducible as possible, this study sets the relevant parameters timestamp to 40, epochs to 20, and chooses a single LSTM model for prediction. The closing prices of the first (1258-200) days are then utilized as a training dataset, while the closing
prices of the previous 200 days are used as the testing sample, with data from the first timestamp days being X and data from timestamp plus one day being Y. Different layers of the LSTM models are built and the model structures used are output [7].

Model: "sequential_1"

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>lstm_1 (LSTM)</td>
<td>(None, 50)</td>
<td>10400</td>
</tr>
<tr>
<td>dense_1 (Dense)</td>
<td>(None, 1)</td>
<td>51</td>
</tr>
</tbody>
</table>

Total params: 10451
Trainable params: 10451
Non-trainable params: 0

**Figure 5.** Model structures

The loss function is used to observe the loss values with MAE, the test dataset is input into the model for prediction, the predicted and real data are reduced and the comparison curves between the real and predicted data are plotted (Fig 5).

### 3.2 Result Analysis

Using Linear regression, a large gap between the true and predicted values was obtained.

![Figure 6. Comparison of Real Value and Predicted Value](image)

The results show an R Squared value of 0.46157 and a Mean Absolute Error value of 22.04023. The mean of the prediction results combined with the mean absolute error can be inferred as the range of the rise and fall of the stock data predicted by the model for a single prediction (Table 1), which is plus or minus 1.68% [8] (Fig 6).
Table 1. Evaluation of Linear Regression Model

<table>
<thead>
<tr>
<th></th>
<th>Linear Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>22.04023</td>
</tr>
<tr>
<td>R Squared</td>
<td>0.46157</td>
</tr>
</tbody>
</table>

The feedback mechanism of LSTM makes it one of the greatest tools for forecasting time series data. It has an input layer, a hidden layer, a dense layer, and an output layer as its fundamental structure. When using the LSTM research method, the loss values are observed and no accuracy is observed, and only the loss values are displayed at each epoch iteration of the display. This step of the study yields the most ideal case, where the training loss decreases and validation loss decreases, which means that the training network is normal.

![Figure 7. Loss](image)

Trains of x and y, as well as Batch size 64 and epochs 20, are supplied to fit the LSTM model (Fig 7-8). The amount of training samples collected by the model in one iteration is referred to as a batch size. Following the completion of the model's training phase, the predict method is employed to make predictions based on the x test data, and the predicted data is then reduced to the real data by inverting the normalization from zero to one to the original range and plotting the comparison curve between the real data and the predicted data [9].

![Figure 8. Comparison of Real Value and Predicted Value](image)
Now check the MAE as well as R2 of the linear regression, which is 7.92757 and 0.74147, respectively. In addition, check the RMSE and MSE, which are 9.80034 and 96.04668, respectively [6] (Table 2).

<table>
<thead>
<tr>
<th>Table 2. Evaluation of LSTM Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
</tr>
<tr>
<td>R Squared</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MSE</td>
</tr>
</tbody>
</table>

4. Discussion

The article finds that LSTM models predict stock prices better than linear regression models in this investigation of stock forecasting approaches. The assumptions applied in econometric models do not always apply in practice. Artificial neural networks, on the other hand, are not constrained by these assumptions and may find nonlinear correlations in data features. As a result, numerous academics have begun researching stock value prediction based on such models.

Econometrics portrays economic principles as a quantifiable mathematical model, i.e., an econometric model, and then uses statistical inference methods to numerically represent this mathematical model. This analytical approach has three features: it mixes theory with observation data, providing theory empirical meaning; it includes the effect of random elements on economic interactions into the analysis, and the results obtained are probabilistic in nature [5].

However, the straightforward use of mathematical equations to define economic rules makes it hard to represent and deal with difficult aspects of socioeconomic situations. Furthermore, the theoretical foundation of empirical analysis via econometric models is overlooked, and the models are simply used for economic analysis. Furthermore, data for developing econometric models is few or of poor quality (Fig 9).

The repeating module in an LSTM contains four interacting layers.

**Figure 9. LSTM Architecture [10]**

LSTM is a form of neural network known as a recurrent neural network. One of the model’s shortcomings is that there is no memory linked with it. This is an issue when dealing with sequential data, such as text or time series. Sequence-to-sequence LSTM models are the state of the art in translation. They also have a wide range of applications, such as time series prediction.

LSTM model works well for some problems, but some of the drawbacks are that the LSTM model takes longer to train, the LSTM model requires more memory to train, the LSTM model is prone to overfitting, Dropout is more difficult to implement in the LSTM model, and LSTM model is sensitive to different random weight initializations.
The prediction of stock prices is affected by many factors, and it is difficult to accurately and effectively predict the future stock price direction based on historical prices alone. However, the results of the study suggest that using LSTM prediction models may be more accurate compared to Linear Regression prediction models, and the findings can be used to inform the application of machine learning in the stock market [11].

<table>
<thead>
<tr>
<th>Methods</th>
<th>Linear Regression</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>22.04023</td>
<td>7.92757</td>
</tr>
<tr>
<td>R Squared</td>
<td>0.46157</td>
<td>0.74147</td>
</tr>
</tbody>
</table>

5. Conclusion

The article obtained historical Amazon stock prices from yahoo finance and predict stock prices by two prediction methods. A mean absolute error of 22.04023 was obtained using the Linear Regression forecasting model and an MAE of 7.92757 was obtained using the LSTM forecasting model. The paper presents a comparative analysis of the gap between the true and predicted values and discusses the advantages and disadvantages of the two methods. Finally, some conclusions are drawn: a robust system or indicator must be less affected by outliers. MSE and RMSE may not be as robust as MAE because the squared error imposes higher importance on the outliers. Therefore, the article compared the MAE of the two models. It is clear that the MAE of LSTM is smaller than that of Linear Regression, which means that the LSTM model predicts better than Linear regression. The LSTM may be more suited to forecasting time series data, whereas linear regression is better suited to non-time series data. LSTM may be more suited to predicting time series data, but linear regression is better suited to forecasting non-time series data.

Thus, holding Amazon stock may be a loss in the short term. However, as inflation data eases, the long-term outlook for Amazon stock remains bright and in the value range. Those who have precise investment objectives and can afford the risk can choose to buy Amazon stock (Table 3).

The research in this paper has several limitations. Deep learning relies on artificial neural networks, a type of machine learning that necessitates a vast quantity of training data. Deep learning performance, whether it can be improved or not, is determined by the size of the data set, therefore deep learning often requires a significant quantity of data as support, which frequently leads to the overfitting issue if a sufficient amount of effective training cannot be completed. Moreover, this study cannot judge the correctness of the data. Deep learning can duplicate what is in the data without comprehending it; it does not reject any data and does not detect hidden biases in the data, which may lead to unobjective outcomes.

A future research direction may be to improve the accuracy of stock market forecasts. By considering macroeconomic data and other factors that affect stock prices, prediction accuracy can be improved. Furthermore, the LSTM structure and parameters should be refined to avoid adopting a single-layer LSTM model. Similar forecasting approaches can be employed in various markets and investment kinds.

References


