Stock Price Prediction using LSTM

Trushita Maurya
Student
Dalhousie University
Halifax
tr711348@dal.ca

Nishit Mistry
Student
Dalhousie University
Halifax
ns209086@dal.ca

1 Introduction

The stock market represents ownership claims through the purchase of shares. The investors buy shares at a cheap price and later on sell them at a higher price. A prominent factor for receiving profit from shares is through analysis of the market. To ensure a possible prediction in the share market can be done, statistical analysis of the past prices of the closing shares during a particular period can help in providing good insights while understanding the prices of the share.

In recent years, machine learning has become popular in various fields that have provided promising results. Technical advancements have led to the introduction of predictive systems in the financial sector. In this paper, we are using Long Short-Term Memory (LSTM) to predict the stock market and the programming language used is Python.

2 Business Problem

As per the resource provided by Adil Moghar et al.[5], stock price analysis is highly volatile and dynamic, which makes it difficult to predict. The key reason to select this problem was to have a platform to predict the stock value of the company that could help yield significant insights into profit.

The stock price fluctuates every day which could lead to loss of finances as its uncertainty is very high. People who are new to the stock market or existing holders can benefit from this analysis to understand the future value of the company and make educated decisions while purchasing or selling stocks on the exchange.

2.1 Data Understanding and Preprocessing

Stock Price Prediction requires a large amount of data for the model to accurately predict the future stock price outcomes. For this purpose, we are using two different stock price datasets that will help us predict the price of stocks more efficiently and effectively. We will integrate the from different sources and train the model from different data sources.

Two Datasets have been used for five different companies. The first data is for the company Tata Global Beverages Limited which has all the information regarding the stock price for a particular period. The dataset consists of features: Date, Open, High, Low, Last, Close, Total Trade Quantity, and Turnover (Lacs) [6]. The other dataset is the Stock Price Dataset which provides the historical prices of stocks like Apple, Microsoft, Tesla, and Facebook. The dataset consists of features: Date, Open, High, Low, Close, Volume, OpenInt, and Stock [7].

A feature description of the Dataset is provided in **Table 1**. Dataset has been taken from two different sources which have almost the same features.

Table 1: Dataset Features

Name	Description
Date	Date of the stock price.
Open	Open price of a stock.
Close	Closing price of a stock.
Low	Lowest price of a stock in a day.
High	Highest price of a stock in a day.
Stock	Company name of a stock.
Last	Last price of a stock in a day.
Turnover (Lacs)	Total turnover by stock in a day.
Volume/Total Trade Quantity	Number of stocks traded in a day.
OpenInt	Total number of future contracts held at the end of the trading day.

The merging between the two datasets was done after some preprocessing that included some cleaning and transformation of the dataset. The first dataset that included Tata stocks was cleaned by dropping the "Last" and "Turnover (Lacs)" features so that it better fits the other dataset and is easy to merge. The name of the "Total Trade Quantity" was also changed to "Volume" and another feature titled "Stock" was added that includes all five brands. For the other stock dataset, the feature "OpenInt" was dropped as it only contained 0's. The merging of the dataset was done using concatenating the two data frames on the rows. Also, the dataset was sorted according to Date.

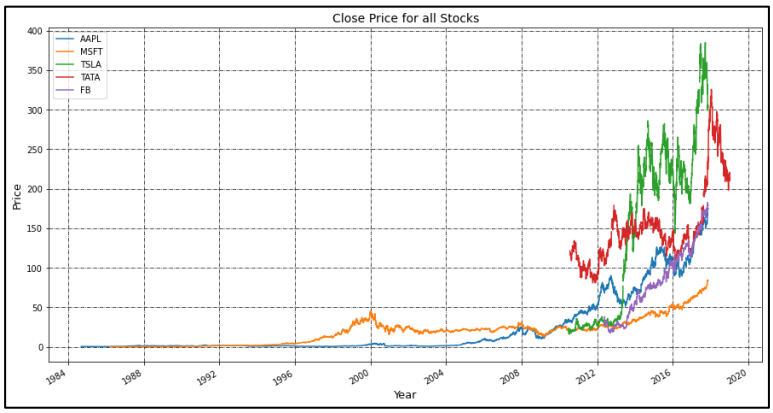

Figure 1: Closing Prices for all Stocks

From **Figure 1,** we can see the closing prices for all the different stocks throughout the various years. The stock prices for Apple are growing steadily with fewer dips but the stock prices for Tesla have grown very high but also have high dips. Microsoft stocks are almost stable and increasing in a very slow manner, while Facebook's and Tata's stocks are also growing slowly. This shows how volatile the stock market is for different companies.

3 Literature Review

V. Kranthi Sai. Reddy et al.[1] suggested that the Support Vector Machine (SVM) does not give any problem of overfitting and easily works on large dataset values gathered from different global financial markets. This is usually the case with SVM that they predict correctly with higher accuracy but the disadvantage of SVM is it is slow to train and takes a lot of time and resources.

Pramod. B. S et al.[2] indicated that an LSTM-based stock price predictor that works on adam optimization and mean squared error loss with dropout regularization produces good results with very low loss and error rate. LSTM is a special kind of Recurrent Neural Network (RNN) that remembers the previous data and predicts future outputs. The solution ahead is also going to use LSTM for predicting the stock prices.

Shreya Pawaskar et al.[3] shows that Random Forest Regressor has a high R Squared (R^2) score and lower Root Mean Squared Error (RMSE) and is a highly accurate model that can be used for Stock Price Prediction. It is an ensemble model that uses the concept of bagging underneath and is highly parallelizable as well.

Mehar Vijh et al.[4] indicates that Artificial Neural Networks (ANN) is a better technique in comparison with Random Forest as they provide a better RMSE and Mean Absolute Percentage Error (MAPE). ANNs are the versatile choice between the two and prove to continue the same in this finding as well.

Adil Moghar et al.[5] show that using LSTM and training with fewer data and more epochs resulted in an improved testing result with better prediction values. It also verifies that increasing the number of epochs leads to more time in training but drastically reduces the loss for predictions. LSTM units are better and more sophisticated as the gates present in them flow the information more efficiently. They are also widely used in time-series data as they store the previous timestamp information.

4 Solution

The proposed system that predicts stocks uses LSTM. The LSTM is an ANN that is mostly used in time-series and deep learning fields and hence a special type of RNN. The special thing that LSTM has is that it remembers the previous timestamp data and keeps it in memory. As per the resource provided by Adil Moghar et al.[5], an LSTM has three gates – Input Gate chooses which data should be stored by first applying a sigmoid function and then a tanh layer, Forget Gate tells whether to keep the data or completely forget it, and Output Gate decides what the output of the cell will be after filtering the fresh data. This leads to an advantage that vanilla RNNs do not have that LSTMs partially solve the vanishing gradient problems as it passes the gradients during backpropagation unchanged[1]. However, LSTM does suffer from long training times, and exploding gradient problems.

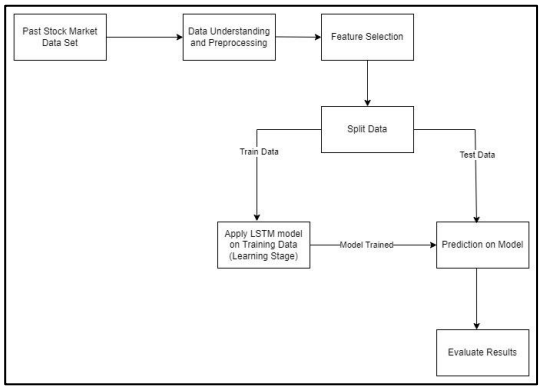

Figure 2: System Architecture

Figure 2 represents the overall system architecture for predicting stock prices using LSTM. The merged dataset is pre-processed and sorted according to the Date feature. The features are manually selected and the data is split into training and test sets. This is a supervised (regression) problem where the target variable is Close.

The LSTM-based model is trained on the training split and the model is tested using the test dataset. The predicted results are then plotted and evaluated using the evaluation metrics that include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R Squared (R^2), and Mean Absolute Percentage Error (MAPE). Also, the training loss is plotted to check the progress of loss vs epochs.

5 Data Analysis, Results, and Evaluation

Dataset has been taken from two different sources having stocks from different companies which have almost the same features. The elected features are Open, Close, Stock, Date, Low, High, and Volume. These attributes were selected based on the correlation matrix as they provided relevant information to predict the closing price. The dataset is imbalanced based on the company shares. For example, the apple share quantity is more than Facebook shares. In such a scenario, the model might make wrong predictions for a different company. Therefore, we trained different models for each company to ensure a good accuracy for prediction. Evaluation metrics used were Mean Absolute Error, Root Mean Squared Error, Mean Absolute Percentage Error and R Squared as this is a regression problem.

[1] "Long short-term memory - Wikipedia", *En.wikipedia.org* [Online]. Available: https://en.wikipedia.org/wiki/Long_short-term_memory. [Accessed: August 01, 2022].

Table 2: Evaluation Metrics

Stock Name	MAE	RMSE	R^2	MAPE
Apple	2.559	3.020	0.984	0.022
Microsoft	1.986	2.546	0.613	0.012
Tesla	2.002	2.310	0.934	0.030
Tata	3.360	4.408	0.820	0.015
Facebook	7.547	9.487	0.755	0.022

Error metric is used for evaluating the performance of the model. As noted, the highest RMSE is for Facebook's share as shown in Table 2. This shows that the difference between the model's predicted values and the expected value is higher than the others denoting the model is performing badly for Facebook shares and therefore more data would be required to train the model.

LSTM model is evaluated on test data and the output of the model shows that it correctly predicts the closing price as shown in Figure 3. The prediction highlighted in yellow is for test data and green shows the actual closing price for test data. As we increased the epochs, the training loss of the model is drastically reduced after one epoch, thereafter it is observed that loss becomes very less. So, to avoid overfitting, early stopping was done by training up to 5 epochs.

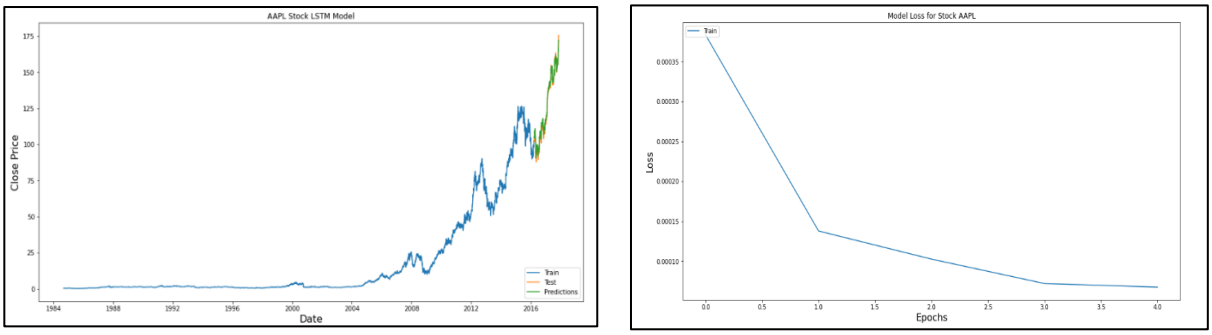

Figure 3: Stock Prediction and Loss for Apple shares

Similarly, the stock prediction was done for the other companies and it was noted that the model was predicting the closing prices correctly.

6 Conclusion

In this project, we proposed the use of the LSTM model as it works well with time series data. LSTM is used to predict the closing prices for the stocks and various evaluation metrics such as MAE, RMSE, R2, and MAPE have been used to judge the model's performance. It is seen that the model is capable of learning and remembering the historic data and predicting the evolution of the stock prices correctly.

Short-term predictions are made to get a probabilistic estimate of share price in the next few days but it is difficult to get long-term predictions due to sudden news that impacts the market making it unpredictable such as political influence or pandemic. This dataset solely captures the statistical features of the stock market. Therefore, to solve some of these limitations, there is a need to capture non-financial analysis during the same period to efficiently map the stock prices with the external factors.

For future work, better deep learning models with more layers can be developed. Grid Search or Random Search can be implemented to find the best hyperparameters for training the model and add more data for the model to learn as training with fewer data can lead to biased results. Other ANNs like CNN (Convolutional Neural Networks) can also be used to map the input data to output variables with weight optimization to improve predictions. Lastly, a machine learning web application can be developed that will take the input from the user with a company or brand name, and show the prediction graph for the next few days, so, it becomes easier for a user to invest in shares correctly with low-risk factor involved.

References

[1] V. Sai Reddy, "Stock Market Prediction Using Machine Learning", *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, pages 1032-1035, 2018.

[2] B. Pramod and M. Shastry P. M., "Stock Price Prediction Using LSTM", *Test Engineering and Management*, vol. 83, pages 5246-5251, 2020.

[3] S. Pawaskar, "Stock Price Prediction using Machine Learning Algorithms", *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. 1, pages 667-673, 2022.

[4] M. Vijh, D. Chandola, V. Tikkiwal and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", *International Conference on Computational Intelligence and Data Science (ICCIDS 2019)*, pages 599-606, 2020.

[5] A. Moghar and M. Hamiche, "Stock Market Prediction Using LSTM Recurrent Neural Network", *International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI 2020)*, pages 1168-1173, 2020.

[6] "Nasdaq Data Link", *Data.nasdaq.com* [Online]. Available: https://data.nasdaq.com/data/NSE/TATAGLOBAL-tata-global-beverages-limited. [Accessed: August 01, 2022].

[7] "Data-Visualization/stock_data.csv at master · pierpaolo28/Data-Visualization", *GitHub* [Online]. Available: https://github.com/pierpaolo28/Data-Visualization/blob/master/Dash/stock_data.csv. [Accessed: August 01, 2022].