Project Report

on

Stock market prediction using Deep learning

Submitted in partial fulfillment of the requirement for the award of the degree of

Btech/CSE

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Dr.Prashant Johri

Submitted By

Paritosh kumar(18scse1010011)

Mohini(18scse1010453)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

October 13, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project , entitled **"** Stock market prediction using Deep learning**"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of July, 2021 to December and 2021, under the supervision of **Dr.Prashant Johri**, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida .

The matter presented in the project has not been submitted by me for the award of any other degree of this or any other places.

Paritosh Kumar (18scse1010011)

Mohini (18scse1010453)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr.Prashant Johri

<u>CERTIFICATE</u>

The Final Project Viva-Voce examination of Paritosh Kumar 18scse1010656 and Mohini 18scse1010453 has been held on ________________ and his work is recommended for the award of Bachelor of Technology , Computer Science and Engineering .

Signature of Examiner(s)

Signature of Supervisor(s)

Signature of Dean

Date: December, 2021

Place: Greater Noida

Table of Contents

ABSTRACT

A stock (also known as equity) is a security that represents the ownership of a fraction of a corporation and are the backbone of any investment portfolio. Advances in trading technology have opened up markets so that nowadays almost everyone has stocks. From in the last few decades, there has been a dramatic increase in the number of descendants of the average person stock market. In a volatile financial market, such as the stock market, it is important that have the most accurate prediction of future practice. Due to the financial of the record, it is imperative that there be a secure prediction of stock market. In this research paper we will focus on Long-Short-Term Memory(LSTM) Recurrent Neural Network belongs to the family of deep learning algorithms.

While predicting the actual price of a stock is an uphill climb, we can build a model that will predict whether the price will go up or down. It’s important to note that there are always other factors that affect the prices of stocks, such as the political atmosphere and the market.

CHAPTER-1

INTRODUCTION

Stock price predictions are very important among many business people and the public. People can make a lot of money or lose their financial income from a stock market job. Algorithm predictions and models can be used to make future predictions applied to historical data. Predicting the future has been a daunting task, one that many have found difficult to understand. This type of prediction is even more appealing when it involves money and risks such as Stock Market speculation. Researchers are conducting research on stock market forecasts from a variety of fields, including computer science and business. Researchers have tried a variety of methods to predict the market, including different strategies and algorithms and the combination of indicators. The attribute that makes a prediction model depends on factors on which market performance can depend. ShortTerm Memory (LSTM) is one of a variety of RNNs structures. LSTM replaces traditional artificial neurons in the hidden network layer into the most useful memory cells. With these memory cells, networks are able to better associate memory with remote input over time, which is why it is worthwhile to understand the formation of strong data over time with great predictive power. A lot of research has been done on stock price forecasts on a daily basis, using different data sources with many built-in models such as news articles, twitter data, google and Wikipedia data. All of these external factors combined with stock prices and stock technology indicators have shown an impact on stock price movements. How to improve the accuracy of stock prices is an open question in today's society. Time series data is a sequence of data from the occasional behaviour of certain fields such as social science, finance, engineering, physics and economics. Finance, engineering, physics, and economics. Those types of complexity make it very difficult to predict price trends. The main purpose of predicting a series of time series is to construct future value simulation models given their past values. In many cases, the relationship between past and future recognition is not clear, this is tantamount to exposing the distribution of conditional opportunities.

1.2 Formulation of Problem

Before we get into the program's implementation to predict the stock market values, let us visualise the data on which we will be working. Here, we will be analysing the stock value of TATA Ltd of India from the Bombay Stock Exchange located in Mumbai , India. The stock value data will be presented in the form of a Comma Separated File (.csv), which can be opened and viewed using Excel or a Spreadsheet.

TATA has its stocks registered in BSE & NSE and has its values updated during every working day of the stock market. Note that the market doesn't allow trading to happen on Saturdays and Sundays , hence there is a gap between the two dates. For each date, the Opening Value of the stock, Highest and Lowest values of that stock on the same days are noted, along with the Closing Value at the end of the day.

The Adjusted Close Value shows the stock's value after dividends are posted . Additionally, the total volume of the stocks in the market are also given, With these data, it is up to the work of a Machine Learning/Data Scientist to study the data and implement several algorithms that can extract patterns from the TATA historical data.

1.2.1 Technology Used

Programming Language : Python

Tools : Google colab studio, keras library, Sklearn library, CSV datatset

Computer with minimum 4GB ram And Good graphic power

METHODOLOGY

The data in this paper consist of the daily opening prices of two stocks in the New York Stock Exchange NYSE TATA extracted from yahoo finance, for TATA our data series cover the period going from 8/19/2004 to 12/19/2019 and for TATA power the data cover the period from 1/4/2010 to 12/19/2019. To build our model we are going to use the LSTM RNN, our model uses 80% of data for training and the other 20% of data for testing. For training we use mean squared error to optimize our model. Also, we used different Epochs for training data (12 epochs, 25 epochs, 50 epochs and 100 epochs) our model will be structured as follow:

Lstm model summary

Layer (type)	Output Shape	Parameters
lstm_1 (LSTM)	(None, 50, 96)	37632
dropout_1 (Dropout)	(None, 50, 96)	0
lstm_2 (LSTM)	(None, 50, 96)	74112
dropout_2 (Dropout)	(None, 50, 96)	0
lstm_3 (LSTM)	(None, 50, 96)	74112
dropout_3 (Dropout)	(None, 50, 96)	0
lstm_4 (LSTM)	(None, 96)	74112
dropout_4 (Dropout)	(None, 96)	0
dense_1 (Dense)	(None, 1)	97
Total number of parameters : 260,065 Trainable parameters : 260,065 Non-trainable parameters :0		

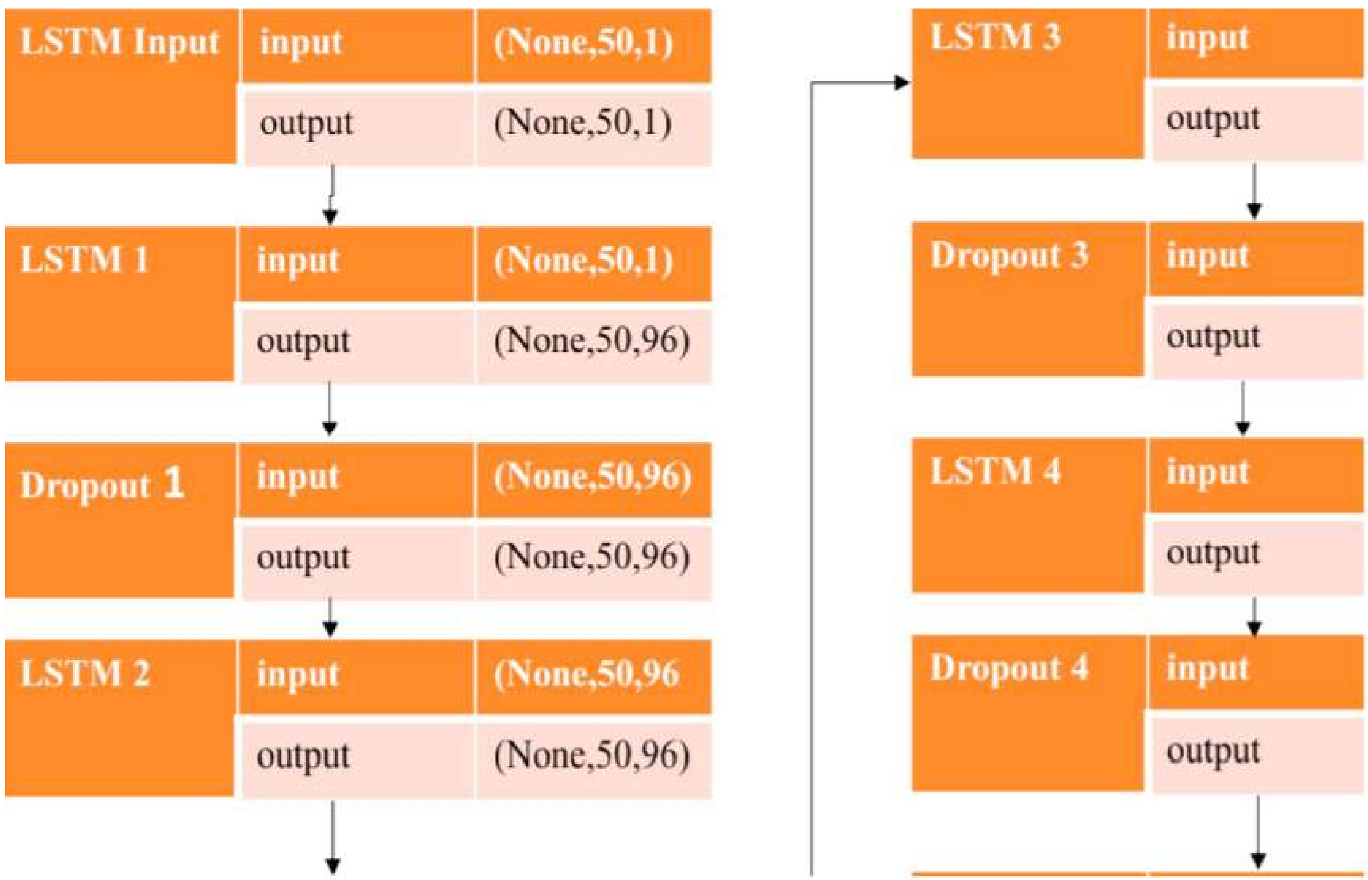


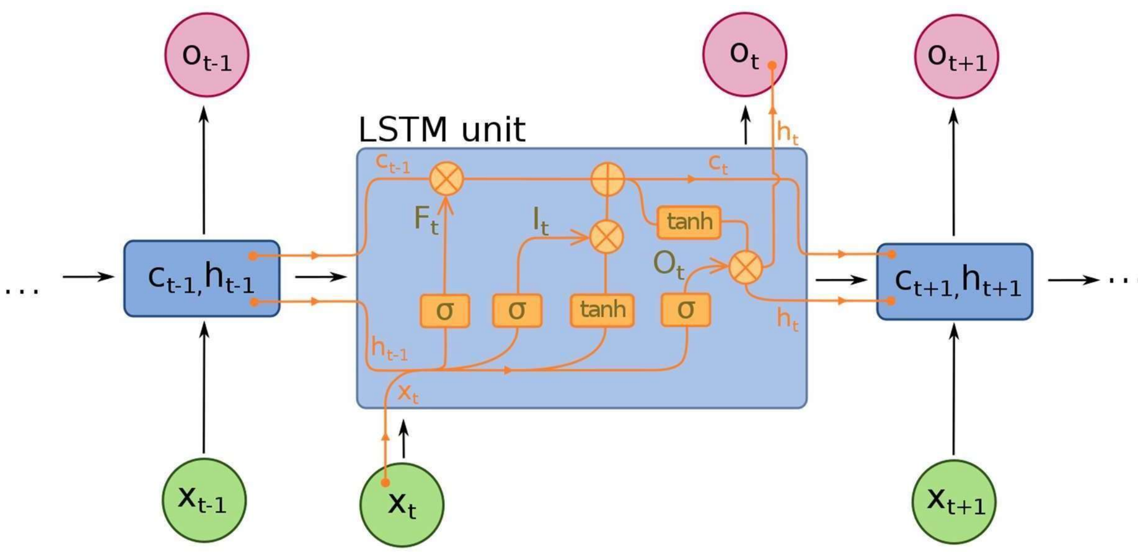
Figure 2: the LSTM model structure

CHAPTER-2
LITERATURE SURVEY

The first focus of our literature review was to evaluate different algorithms and models to determine whether stock price predictions could be made on real stock prices. However, as we have not been able to detect a possible change in this stock price forecast, we decided to look at existing plans, analyse major issues, and improve ourselves. A brief search of common solutions to the above problem led us to LSTM. After deciding to use the LSTM neural network to make stock forecasts, time series data is collected from stock firm prices of the stock and related macroeconomic variables over a period of 10 years .

2.1 PROPOSED SYSTEM

We propose to use LSTM (Long Short Term Memory) algorithm to provide efficient stock price prediction



A special type of RNN, which can learn long-term dependence, is called Long-Short Term Memory (LSTM). LSTM enables RNN to remember long-term inputs. Contains information in memory, similar to computer memory. It is able to read, write and delete information in its memory. This memory can be seen as a closed cell, with a closed description, the cell decides to store or delete information. In LSTM, there are three gates: input, forget and exit gate. These gates determine whether new input (input gate) should be allowed, data deleted because it is not important (forget gate), or allow it to affect output at current timeline (output gate).

1. **Forget gate**: The forget gateway determines when certain parts of the cell will be inserted with information that is more recent. It subtracts almost 1 in parts of the cell state to be kept, and zero in values to be ignored.
2. **Input gate** : Based on the input (e.g., previous output o (t-1), input x (t), and the previous state of cell c (t-1)), this network category reads the conditions under which any information should be stored (or updated) in the state cell.
3. **Output gate**: Depending on the input mode and the cell, this component determines which information is forwarded in the next location in the network.

2.2 Advantages of LSTM

The main advantage of LSTM is its ability to read intermediate context. Each unit remembers details for a long or short period without explicitly utilizing the activation function within the recurring components. An important fact is that any cell state is repeated only with the release of the forget gate, which varies between 0 and 1. That is to say, the gateway for forgetting in the LSTM cell is responsible for both the hardware and the function of the cell state activation. Thus, the data from the previous cell can pass through the unchanged cell instead of explicitly increasing or decreasing in each step or layer, and the instruments can convert to their appropriate values over a limited time. This allows LSTM to solve a perishable gradient problem - because the amount stored in the memory cell is not converted in a recurring manner, the gradient does not end when trained to distribute backwards.

2.3 PROJECT DESIGN

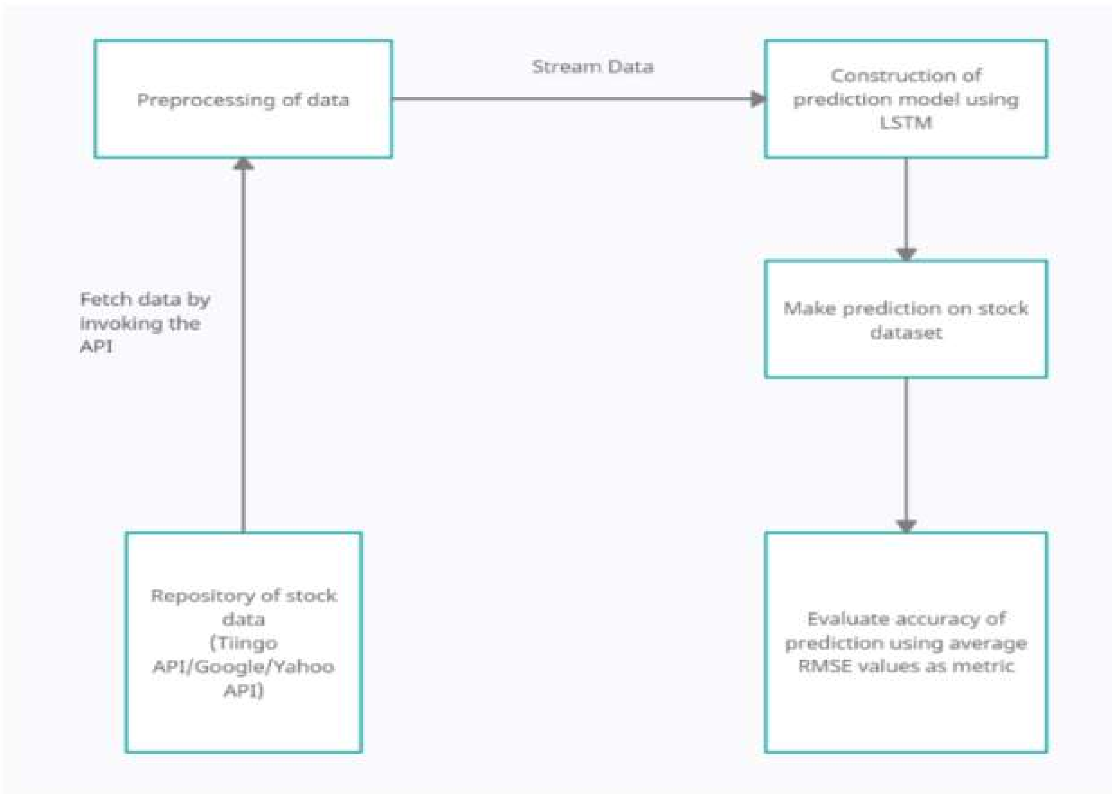


Figure 2 System Model

Obtaining dataset and pre-processing The obtained data contained five features:

1. Date: Date of stock price.
2. Opening price: When trading begins each day this is opening price of stock.
3. High: The highest price at which the stock was traded during a period(day).
4. Low: The Lowest price at which the stock was traded during a period(day).
5. Volume: How much of a given financial asset has traded in a period of time.
6. Close Interest: The last price at which a particular stock traded for the trading session.

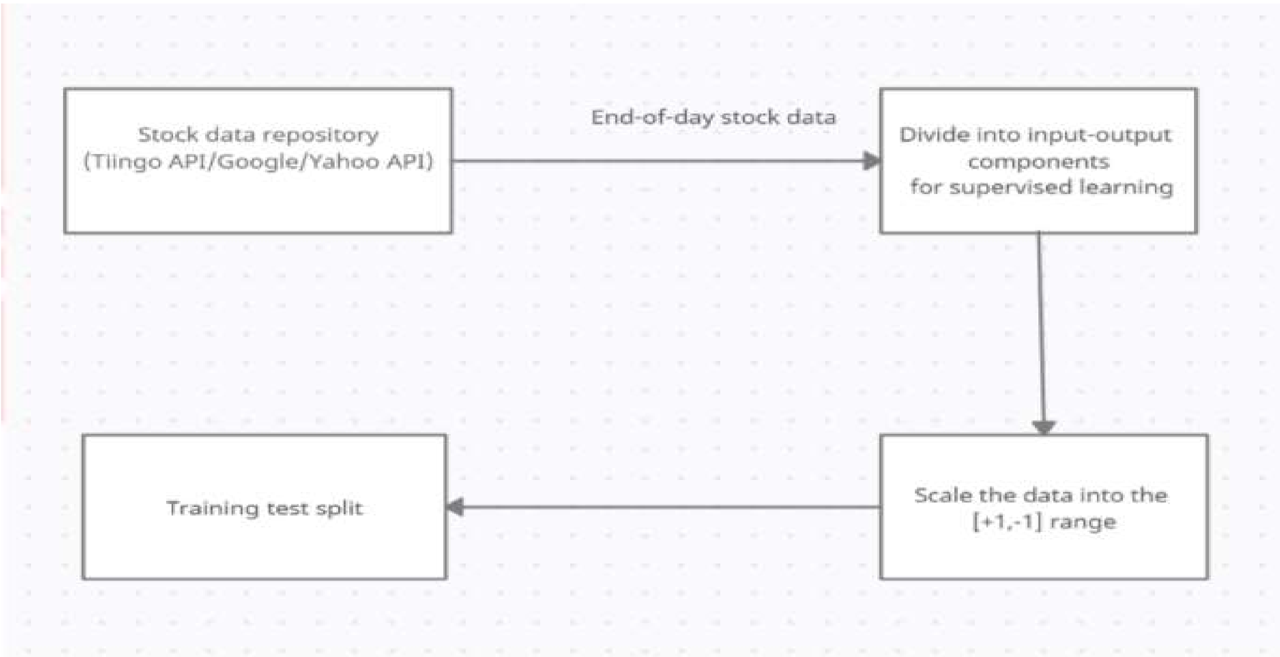


Figure 3 Data Preprocessing

Stock market information is available from key sources: Tiingo API, Yahoo and Google Finance. These websites give APIs from which stock dataset can be obtained from various companies by simply specifying parameters. The data is processed into a format suitable to use with prediction model by performing the following steps: 1. Transformation oftime-seriesdata intoinput-output components for supervised learning. 2. Scaling the data to the [-1, +1]range.

2.4 RELATED WORK

There has been several research work on implementing machine learning algorithm for predicting stock market. A study is done by implementing machine learning algorithms on Karachi Stock Exchange (KSE) in [10]. It compared Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF) and Support Vector

Machine (SVM). MLP performs best as compared to others. A comparison of four techniques Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and Naive-Bayes is done in [11]. A study used unsupervised learning as a precursor for

supervised tasks [12]. A study compared various machine learning techniques like Random Forest, AdaBoost, Kernel Factory, Neural Networks, Logistic Regression, Support Vector Machine, KNN, on dataset of European companies [13]. An application of various machine learning algorithms (SVM, NaÃ¯ve Bayes, Random Forest) was done and it was found that random forest gives the highest F- score [14]. A research applied RNN, LSTM, Gated Recurrent Unit (GRU) on google stock dataset and found that LSTM outperforms other algorithms [15]. An application of LSTM to predict Nifty prices is done in [16]. A proposal of an algorithm of multi layer feedforward networks on Chinese Stock dataset is done in [17]. A study applied random forest on Shenzhen Growth Enterprise Market (China) in [18]. It aims to predict stock price and also interval of growth rate. They found that this method is better than some existing methods in terms of accuracy. A proposal of a model that consists of LSTM and GRU is done in [19]. They applied it on S&P dataset. The result was better than some existing neural network approach. A research compared SVM (supervised) and K-means clustering (unsupervised) on S&P 500 dataset [20]. They

perform Principal Component Analysis (PCA) for dimensionality reduction. They found that both algorithms give similar performance. The accuracy of SVM is 89.1% and accuracy of K-means is 85.6%. An algorithm is proposed in [21] which combined the Hierarchical Agglomerative Clustering (HAC) and reverse K-means clustering to predict the stock market. It compared HRK model with HAC, K-means, reverse K-means, SVM. The study found that the proposed system is better than SVM in terms of accuracy. An analysis of a model by AprioriALL algorithm (association rule learning) and K-means clustering is done in [22]. It converted data into charts and clustered using K-means to analyze patterns. A paper proposed a clustering method on the Stock Exchange of

Thailand (SET) and found that the proposed method is better than other methods of stock market prediction [23].

2.5 Dataset

The dataset is downloaded from kaggle. The dataset represent data of National Stock Exchange of India for the years 2016 and 2017. The description of dataset is given in Table1.

Table 1. Description of dataset

Feature	Description
Symbol	Symbol of the listed company
Series	Series of the equity(EQ, BE, BL, BT, GC, IL)
Open	Starting price at which a stock is traded in a day
High	Highest price of equity symbol in a day
Low	Lowest price of share in a day
Close	Final price at which a stock is traded in a day
Last	Last traded price of the equity symbol in a day
Prevclose	The previous day closing price of equity symbol in a day
TOTTRDQTY	Total traded quantity of equity symbol on the date
TOTTRDVAL	Total traded volume of equity symbol on the date

2.6 DATA PRE-PROCESSING

The dataset is in raw format. The dataset needs to be converted into a format that can be analysed. Therefore there are some steps that are performed before building the model:

1. Handling missing data

2. One Hot Encoding: It converts categorical data to quantitative variable as any data in the form of string or object does not help in analysing data. First step is to convert the columns to category data type. Second step is to apply label encoding in order to convert it into numerical values which will be valuable for analysis. Third step is to convert the column into binary value (either 0 or 1).

3. Data Normalization: It is often possible that if data is not normalized, the column with high values will be given more importance in prediction. In order to tackle that, we scale the data.

2.7 Classifier

Classifiers are given training data, it constructs a model. Then it is supplied testing data and the accuracy of model is calculated. The classifiers used in this paper are:

1) Random Forest Classifier:

It is a supervised algorithm and a type of ensemble learning program. It is a very versatile algorithm capable of performing regression as well as classification. It is built on decision trees. It basically builds multiple decision tree and merges them for producing result. In this algorithm, only a subset of features is taken into consideration. It has same hyperparameters as a decision tree. Advantages of Random Forest are that it works very effectively on large dataset. It can work for both regression and classification problems. It adds more randomness to the model which makes it a better model. The disadvantage of this model is that it makes use of large number of trees that makes it slow.

Algorithm:

1. Randomly select m features.
2. For a node, find the best split.
3. Split the node using best split.
4. Repeat the first 3 steps.
5. Build the forest by repeating these 4 steps.

2) SVM (Support Vector Machine):

It is a supervised learning algorithm which classifies cases by a separator. It works by mapping data to a high dimensional feature space and then finds a separator. It finds n-dimensional space that categorizes data points. This algorithm finds the best plane. This plane must have a maximum margin. The boundary that classifies data points is called hyperplanes. The data points are classifies on the basis of position with respect to hyperplanes. Kernel parameter, gamma parameter and regularization parameter are tuning parameters of SVM. Linear kernel predicts new input by dot product between input and support vector. Mapping data to a higher dimensional space is called kernelling. Kernel function can be linear, polynomial, RBF and Sigmoid. Regularization parameter is the C parameter with default value of 10. Less regularization means wrong classification. Small value of gamma means not able to find the region of data. One can improve the model by increasing the importance of classification of each data. Advantages of SVM are that it is a good algorithm for estimation in high dimensional space and it is very memory efficient. Disadvantages of SVM are that it can suffer from over-fitting and that it works very well on small datasets.

3) KNN (K-nearest neighbour):

It is an algorithm for classifying similar cases. It produces results only when they are requested. Therefore, it is called lazy learner because there is no learning phase. Advantages of KNN are that it is one of the simplest algorithms as it has to compute the value of k and the Euclidean distance only. It is sometimes faster than other algorithms because of its lazy learning feature. It works well for multiclass problem. Disadvantages of KNN are that the algorithm may not generalize well as it does not go through the learning phase. It is slower for a large dataset as it will have to calculate sort all the distances from the unknown item. Data normalization is necessary for KNN algorithm in order to get best result.

Algorithm for KNN-

a. Choose k.

b. Calculate the Euclidean distance of all cases from unknown case.

The Euclidean distance (also called the least distance) between sample x and y is

:

Where,

xi is the ith element of the instance x, yi is the ith element of the instance y, n is the total number of features in the data set.

c. k number of data points are chosen near unknown data.

d. The unknown data will belong to the majority cases in chosen k neighbours.

CHAPTER-3
WORK ON PROJECT

Loading the Dataset

The next step is to load in our training dataset and select the Open and Highcolumns that we'll use in our modeling.

```
dataset_train = pd.read_csv('NSE-TATAGLOBAL.csv')
training_set = dataset_train.iloc[:, 1:2].values
```

Feature Scaling

```
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler(feature_range = (0, 1))
training_set_scaled = sc.fit_transform(training_set)
```

Creating Data with Timesteps

LSTMs expect our data to be in a specific format, usually a 3D array. We start by creating data in 60 timesteps and converting it into an array using NumPy. Next, we convert the data into a 3D dimension array with X_train samples, 60 timestamps, and one feature at each step.

```
X_train = []
y_train = []
for i in range(60, 2035):
    X_train.append(training_set_scaled[i-60:i, 0])
    y_train.append(training_set_scaled[i, 0])
X_train, y_train = np.array(X_train), np.array(y_train)
```

X_train = np.reshape(X_train, (X_train.shape[0], X_train.shape[1], 1))

Building the LSTM

from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import Dropout

regressor = Sequential()

regressor.add(LSTM(units = 50, return_sequences = True, input_shape = (X_train.shape[1], 1)))
regressor.add(Dropout(0.2))

regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(0.2))

regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(0.2))

regressor.add(LSTM(units = 50))
regressor.add(Dropout(0.2))

regressor.add(Dense(units = 1))

regressor.compile(optimizer = 'adam', loss = 'mean_squared_error')

```
regressor.fit(X_train, y_train, epochs = 100, batch_size = 32)
```

Predicting Future Stock using the Test Set

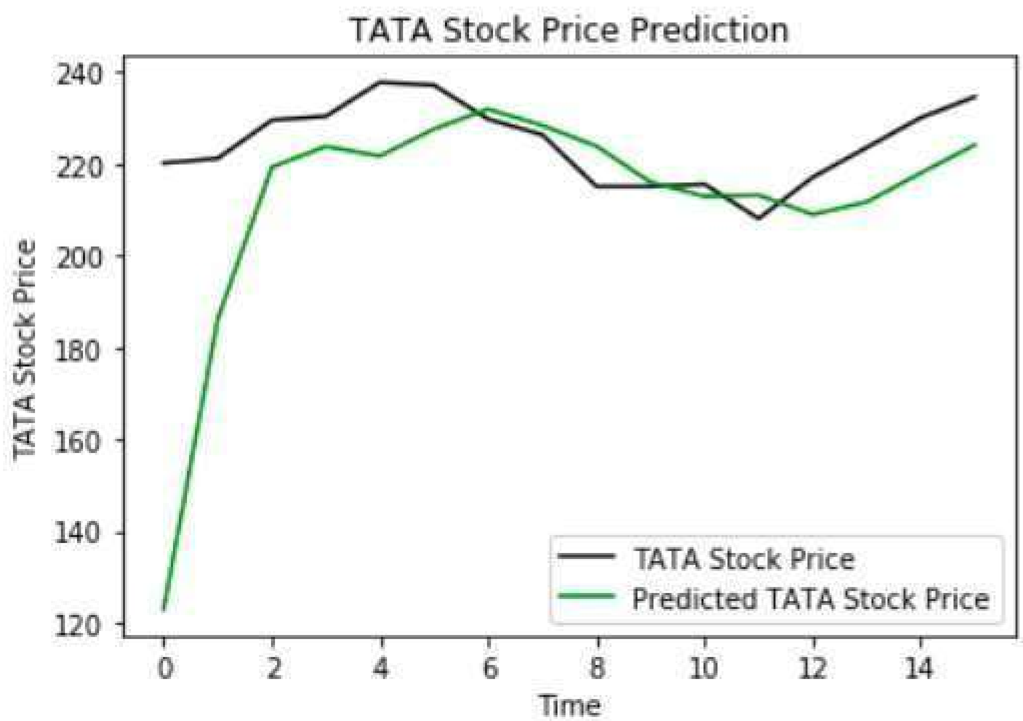
```
dataset_test = pd.read_csv('tatatest.csv')
real_stock_price = dataset_test.iloc[:, 1:2].values

dataset_total = pd.concat((dataset_train['Open'], dataset_test['Open']), axis = 0)
inputs = dataset_total[len(dataset_total) - len(dataset_test) - 60:].values
inputs = inputs.reshape(-1,1)
inputs = sc.transform(inputs)
X_test = []
for i in range(60, 76):
    X_test.append(inputs[i-60:i, 0])
X_test = np.array(X_test)
X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))
predicted_stock_price = regressor.predict(X_test)
predicted_stock_price = sc.inverse_transform(predicted_stock_price)
```

Plotting the Results

```
plt.plot(real_stock_price, color = 'black', label = 'TATA Stock Price')
plt.plot(predicted_stock_price, color = 'green', label = 'Predicted TATA Stock Price')
```

```
plt.title('TATA Stock Price Prediction')
plt.xlabel('Time')
plt.ylabel('TATA Stock Price')
plt.legend()
plt.show()
```



CHAPTER-4

RESULT AND DISCUSSION

After training our NN the result of our testing has shown different results, the number of epochs as well as the length of the data have both significant impact on the result of testing. For example, if we changed the dataset for NKE giving it a time set going from 12/2/1980 to 12/19/2019 the results will appear as follow:

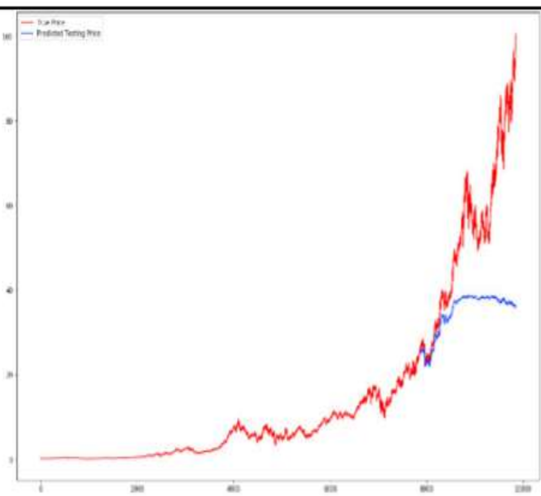
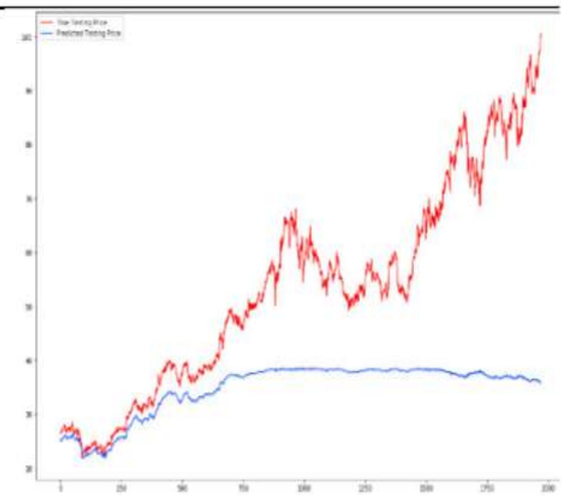
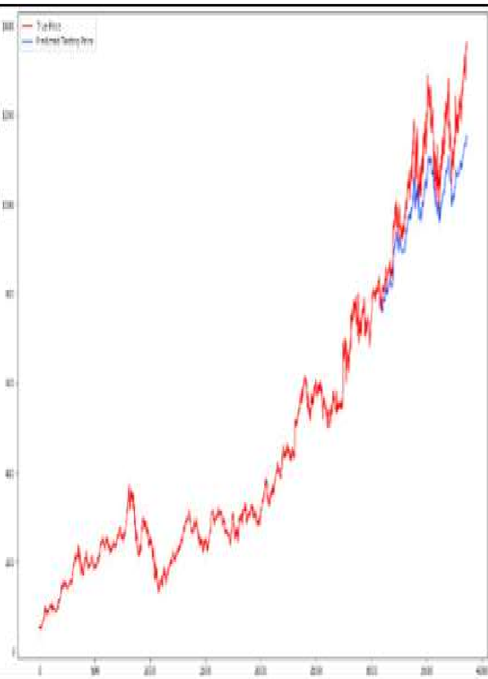
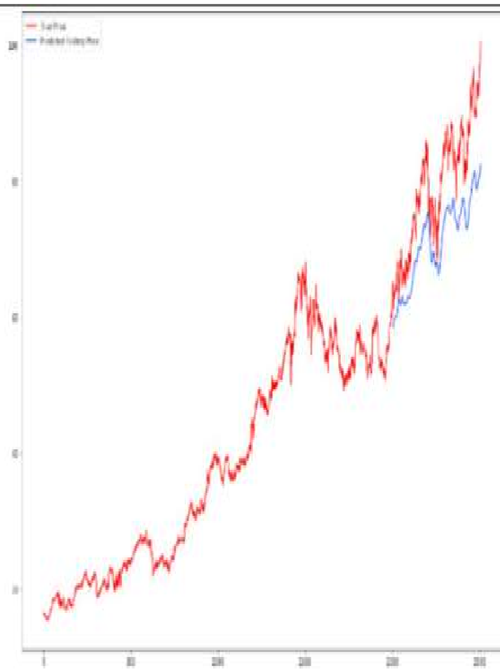
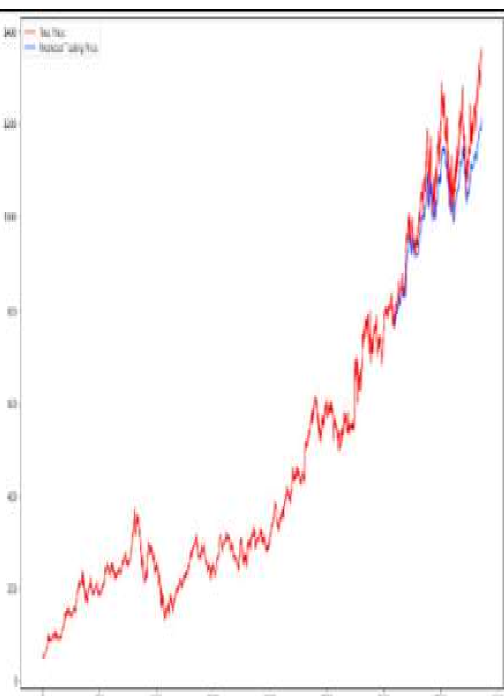
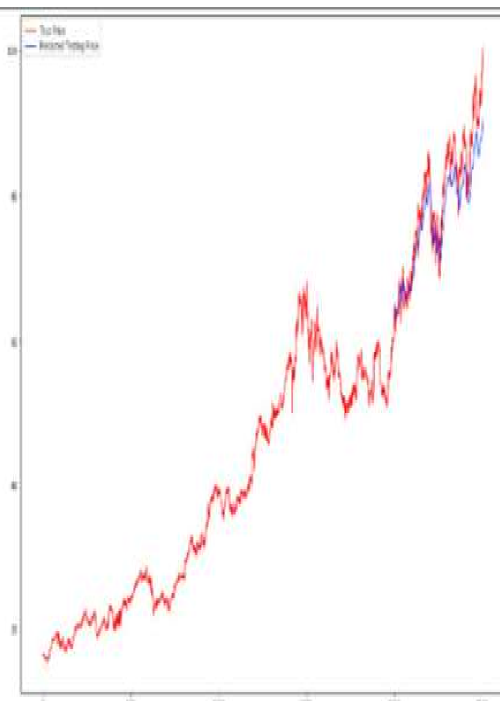
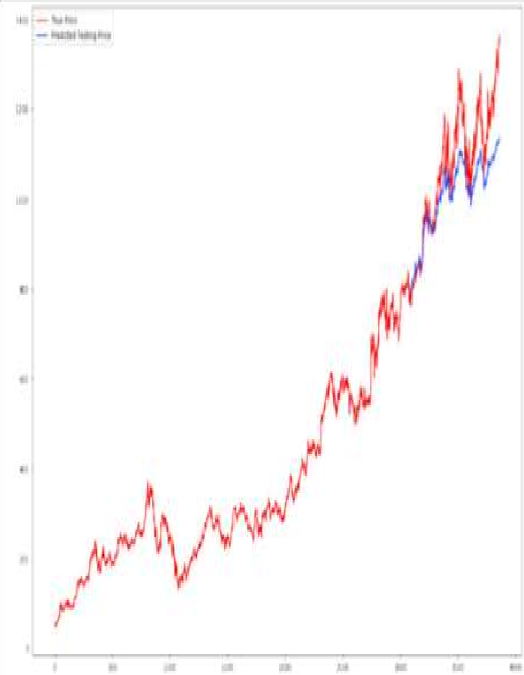
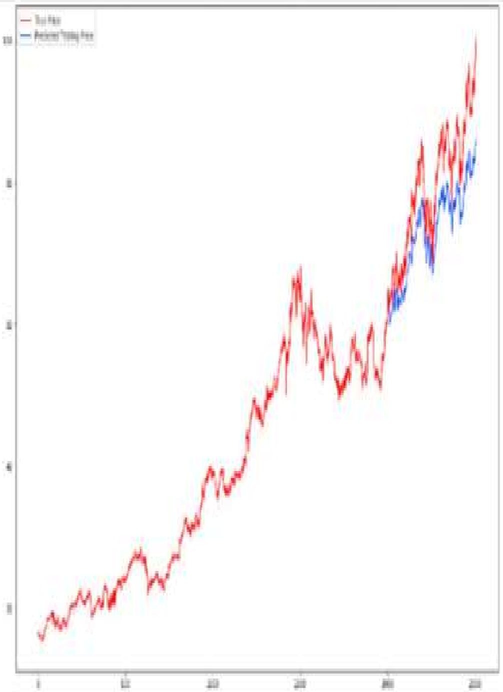
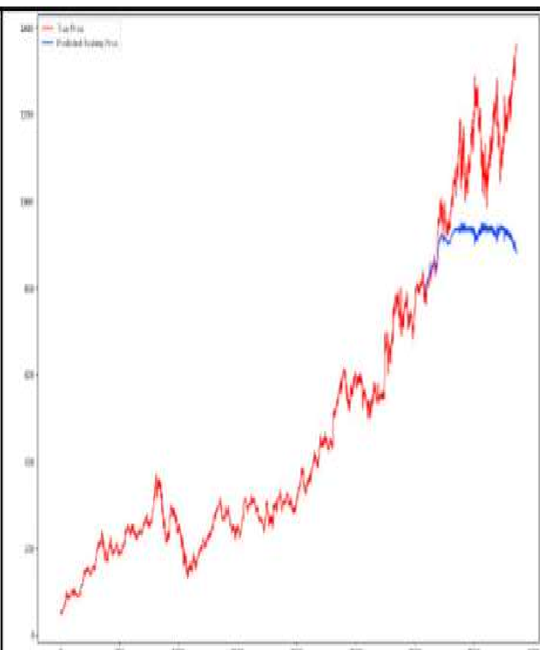
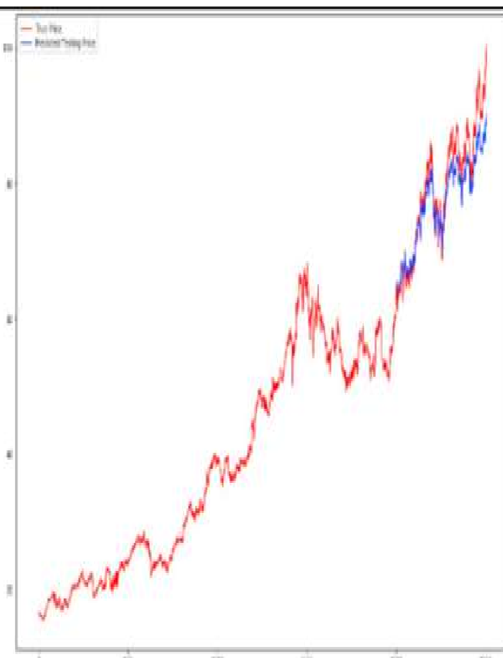
Epochs	NKE	NKE zomed view
25 Epochs		

Figure 3: result of training for the NKE stocks with different dataset time

After observing our data, we can see that at first the data was less volatile and have lower values, in the figure, the red lines represent the real market value and the blue lines represent the predicted price value, after the NKE start peeking bigger values, the asset become more volatile, then the nature of this asset changed. In our case is better to avoid this type of change. Our model has lost trace of opening prices around 600 to 700 day of testing which conform the change in data nature. The result for our dataset for different number of epochs is giving by the following Figure:

	TATA	TATA POWER
12 Epochs		
25 Epochs		

50 Epochs		
100 Epochs		

For different data set we can observe that training with less data and more epochs can improve our testing result and at the same time allow us to have beater forecasting and prediction values. The following table shows the precision of our training and testing for all the epochs for both TATA and TATA POWER asset price.

	processing Time / sec	Loss	processing Time / sec	Loss
12 epochs	264	0.0011	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

above confirms that the precision of our forecasting increases if we add more epochs of training for our model.

CHAPTER-4
Conclusion and Future Scope

We proposed a model that uses LSTM to predict the trend in stock prices that would be more accurate. LSTM introduces the memory cell, a unit of computation that replaces traditional artificial neurons in the hidden layer of the network. In this work by increasing the Epochs and batch size, the accuracy of prediction is more. In proposed method, we are using a test data that is used to predict which gives results that are more accurate with the test data. The proposed method of tracing and prediction of stock market and the prediction will produce higher and accurate results. In our above model we are getting accurate results which will be more useful to stock analysts, Business analysts, Stock Market Investors.

Stock investing has attracted the interest of many investors around the world. However, making a decision is a difficult task as many things are involved. By investing successfully, investors are eager to predict the future of the stock market. Even the slightest improvement in performance can be enormous. A good forecasting system will help investors make investments more accurate and more profitable by providing supporting information such as future stock price guidance. In addition to historical prices, other related information could affect prices such as politics, economic growth, financial matters and the atmosphere on social media. Numerous studies have proven that emotional analysis has a significant impact on future prices. Therefore, the combination of technical and basic analysis can produce very good predictions

CHAPTER-6
REFERENCES

[1] Kim, S., Ku, S., Chang, W., & Song, J. W. (2020). Predicting the Direction of US Stock Prices Using Effective Transfer Entropy and Machine Learning Techniques. IEEE Access, 8, 111660-111682.

[2] Nandakumar, R., Uttamraj, K. R., Vishal, R., & Lokeswari, Y. V. (2018). Stock price prediction using long short term memory. International Research Journal of Engineering and Technology, 5(03).

[3] Roondiwala, Murtaza, Harshal Patel, and Shraddha Varma. "Predicting stock prices using LSTM." International Journal of Science and Research (IJSR) 6.4 (2017): 1754-1756.

[4] Pahwa, Kunal, and Neha Agarwal. "Stock market analysis using supervised machine learning." 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, 2019.

[5] Rajput, Vivek, and Sarika Bobde. "Stock market forecasting techniques: literature survey." International Journal of Computer Science and Mobile Computing 5.6 (2016): 500-506.

[6] Singh, S., Madan, T. K., Kumar, J., & Singh, A. K. (2019, July). Stock market forecasting using machine learning: Today and tomorrow. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 738-745). IEEE.

[7] Yu, Pengfei, and Xuesong Yan. "Stock price prediction based on deep neural networks." Neural Computing and Applications 32.6 (2020): 1609-1628.

[8] Nayak, Aparna, MM Manohara Pai, and Radhika M. Pai. "Prediction models for Indian stock market." Procedia Computer Science 89 (2016): 441-449.

[9] Ghosh, Achyut, Soumik Bose, Giridhar Maji, Narayan Debnath, and Soumya Sen. "Stock price prediction using LSTM on Indian share market." In Proceedings of 32nd international conference on, vol. 63, pp. 101-110. 2019.

[10]Wei, Dou. "Prediction of stock price based on LSTM neural network." 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM). IEEE, 2019.

[11]Sadia, K. H., Sharma, A., Paul, A., Padhi, S., & Sanyal, S. (2019). Stock market prediction using machine-learning algorithms. Int. J. Eng. Adv. Technol, 8(4), 25-31.

[12]Usmani, Mehak, Syed Hasan Adil, Kamran Raza, and Syed Saad Azhar Ali. "Stock market prediction using machine learning techniques." In 2016 3rd international conference on computer and information sciences (ICCOINS), pp. 322-327. IEEE, 2016.

[13]Cont., R. Empirical properties of asset returns: stylized facts and statistical issues. Quant. Finance 2001, 1, 223–236. [14]Cheng, J.; Huang, K.; Zheng, Z. towards Better Forecasting by Fusing Near and Distant Future Visions. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020