IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Stock Price Prediction Using Machine Learning

USING LSTM AND LINEAR REGRESSION

Mrs.B. SAILAJA

Assistant professor

Computer Science and Engineering

Vignan's Institute of Engineering for Women, Visakhapatnam

Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

R. Chandini, S. Rupa Sri, N. Krishna Veni, S. Nirmala

B. Tech Students

Computer Science and Engineering

Vignan's Institute of Engineering for Women, Visakhapatnam

Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Abstract: The prediction of stock value is a complex task which needs a robust algorithm background in order to compute the longer-term share prices. Stock prices are correlated within the nature of market; hence it will be difficult to predict the costs.

The proposed algorithm using the market data to predict the share price using machine learning techniques like recurrent neural network named as Long Short-Term Memory, in that process weights are corrected for each data points using stochastic gradient descent. This system will provide accurate outcomes in comparison to currently available stock price predictor algorithms. The network is trained and evaluated with various sizes of input data to urge the graphical outcomes.

The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values. Factors considered are open, close, low, high and volume.

Key words :close, open, high, low, volume,LSTM model and regression .

1. INTRODUCTION

Stock price is the price of a single stock among the number of stocks sold by a company listed in public offering. Having stocks of a public company allows you to own a portion of it. Original owners of the company initially sell the stocks to get additional investment to help the company grow. This initial offering of stocks to the public is called Initial Public Offering (IPO). Stock prices change because of the supply and demand. Suppose, if many people are willing to buy a stock, then the price goes up as there is more demand. If more people are willing to sell the stock, the price goes down as there is more supply than the demand. Though understanding supply and the demand is relatively easy, it is hard to derive what factors exactly contribute to the increase in demand or supply. These factors would generally boil down to socioeconomic factors like market behaviour, inflation, trends and more importantly, what is positive about the company in the news and what's negative. Predicting the accurate stock price has been the aim of investors ever since the beginning of the stock market. Millions of dollars- worth of trading happens every single day, and every trader hopes

to earn profit from his/her investments. Investors who can make right buy and sell decisions will end up in profits. To make right decisions, investors have to judge based on technical analysis, such as company's charts, stock market indices and information from newspapers and microblogs. However, it is difficult for investors to analyse and forecast the market by churning all this information. Therefore, to predict the trends automatically, many Artificial Intelligence (AI) techniques have been investigated. Some of the first research in prediction of stock prices dates back to 1994, in which a comparative study with machine learning regression models was performed. Since then, many researchers were investing resources to devise strategies for forecasting the price of the stock.

2 PROBLEM STATEMENT

Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the quarterly financial ratio using the dataset. Thus, relying on a single dataset may not be sufficient for the prediction and can give a result which is inaccurate. Hence, we are contemplating towards the study of machine learning with various datasets integration to predict the market and the stock trends. The problem with estimating the stock price will remain a problem if a better stock market prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. The movement in the stock market is usually determined by the sentiments of thousands of investors. Stock market prediction, calls for an ability to predict the effect of recent events on the investors. These events can be political events like a statement by a political leader, a piece of news on scam etc. It can also be an international event like sharp movements in currencies and commodity etc. All these events affect the corporate earnings, which in turn affects the sentiment of investors. It is beyond the scope of almost all investors to correctly and consistently predict these hyperparameters. All these factors make stock price prediction very difficult. Once the right data is collected, it then can be used to train a machine and to generate a predictive results.

3. LITERATURE SURVEY

The initial focus of our literature survey was to explore generic online learning algorithms and see if they could be adapted to our use case i.e., working on real-time stock price data. These included Online AUC Maximization, Online Transfer Learning, and Online Feature Selection. However, as we were unable to find any potential adaptation of these for stock price prediction, we then decided to look at the existing system, analyse the major drawbacks of the same, and see if we could improve upon them. We zeroed in on the correlation between stock data (in the form of dynamic, long-term temporal dependencies between stock prices) as the key issue that we wished to solve. A brief search of generic solutions to the above problem led us to RNN's and LSTM. After deciding to use an LSTM neural network to perform stock prediction, we consulted a number of papers to study the concept of gradient descent and its various types. We concluded out literature survey by looking at how gradient descent can be used to tune the weights of an LSTM network and how this process can be optimized.

4 EXISTING SYSTEMS AND DRAWBACKS

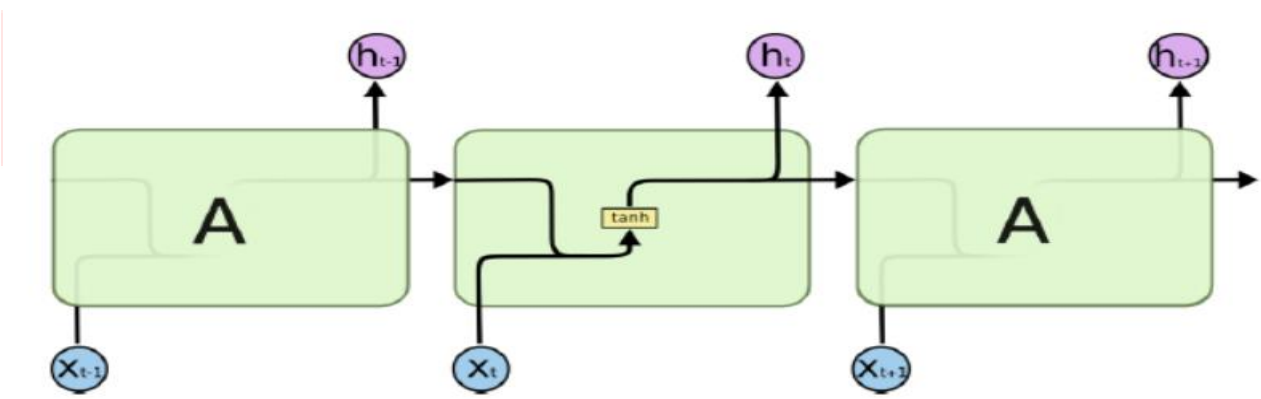
Traditional approaches to stock market analysis and stock price prediction include fundamental analysis, which looks at a stock's past performance and the general credibility of the company itself, and statistical analysis, which is solely concerned with number crunching and identifying patterns in stock price variation. The latter is commonly achieved with the help of Genetic Algorithms (GA) or Artificial Neural Networks (ANN's), but these fail to capture correlation between stock prices in the form of long-term temporal dependencies. Another major issue with using simple ANNs for stock prediction is the phenomenon of exploding / vanishing gradient, where the weights of a large network either become too large or too small (respectively), drastically slowing their convergence to the optimal value. This is typically caused by two factors: weights are initialized randomly, and the weights closer to the end of the network also tend to change a lot more than those at the beginning. An alternative approach to stock market analysis is to reduce the dimensionality of the input data and apply feature selection algorithms to shortlist a core set of features (such as GDP, oil price, inflation rate, etc.) that have the greatest impact on stock prices or currency exchange rates across markets. However, this method does not consider long-term trading strategies as it fails to take the entire history of trends into account; furthermore, there is no provision for outlier detection.

5 PROPOSED SYSTEM

In this proposed system, we focus on predicting the stock values using machine learning algorithms like Linear Regression and Long Short-Term Memory (LSTM). We proposed the system "Stock price prediction" we have predicted the stock market price using the LSTM algorithm. In this proposed system, we were able to train the machine from the various data points from the past to make a future prediction. We took data from the previous year stocks to train the model. We majorly used two machine-learning libraries to solve the problem. The first one was numpy, which was used to clean and manipulate the data, and getting it into a form ready for analysis. The other was scikit, which was used for real analysis and prediction. The data set we used was from the previous-years stock markets collected from the public database available online, 80 % of data was used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data. We used the python pandas library for data processing which combined different datasets into a data frame. The tuned-up data-frame allowed us to prepare the data for feature extraction. The data-frame features were date and the closing price for a particular day. We used all these features to train the machine on Linear Regression and LSTM and predicted the object variable, which is the price for a given day. We also quantified the accuracy by using the predictions for the test set and the actual values. The proposed system touches different areas of research including data pre-processing, Linear Regression and LSTM.

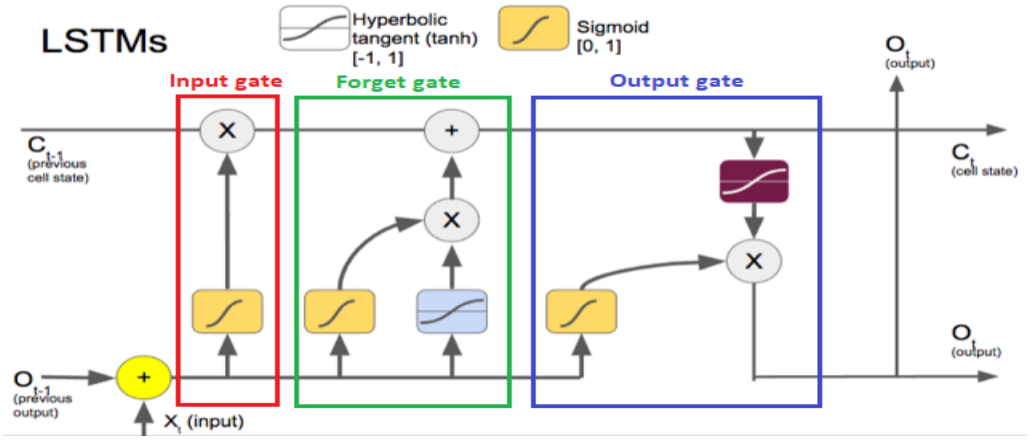
5.1 LSTM (LONG SHORT-TERM MEMORY)

Long short-term memory (LSTM) is a type of recurrent neural-network architecture in which the vanishing gradient problem is solved. LSTMs are capable of learning very long-term dependencies and they work tremendously well on a large variety of problems. LSTMs are first introduced by Hochreiter et al. in 1997 [10]. In addition to the original authors, many researchers contributed to the architecture of modern LSTM cells. Recurrent neural networks are generally designed in a chain like structure, looping back to the previous layers. In standard RNNs, this looping module will have a very simple structure, as shown in the Figure 4. This structure can be a simple tanh layer controlling the flow.



5.1 simple structure of RNN's Looping module

Whereas in LSTMs, instead of this simple structure, they have four network layers interacting in a special way. General architecture of LSTMs is shown in Figure 5. Normally LSTM's are augmented by gates called "forget" gates. By controlling these gates, errors can be backpropagated through any number of virtual layers. This mechanism enables the network to learn tasks that depend on events that occurred millions of time steps ago.



5.2 LSTM Memory Unit Cell

1. **Forget gate:** The forget gate decides when specific portions of the cell state are to be replaced with more recent information. It outputs values close to 1 for parts of the cell state that should be retained, and zero for values that should be neglected.

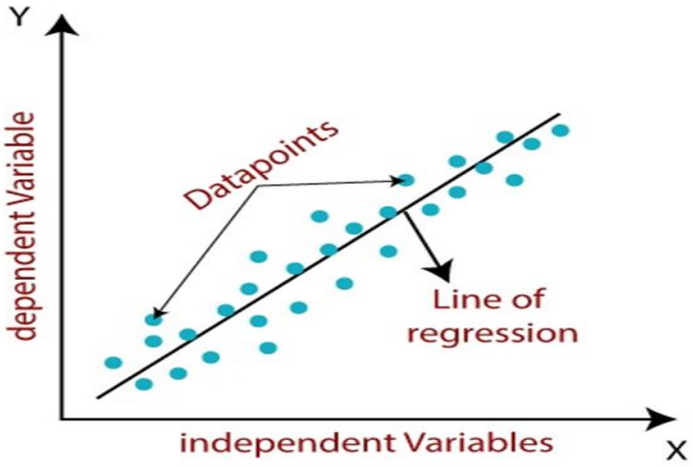
2. **Input gate:** Based on the input (i.e., previous output o(t-1), input x(t), and previous cell state c(t-1)), this section of the network learns the conditions under which any information should be stored (or updated) in the cell state

3. **Output gate:** Depending on the input and cell state, this portion decides what information is propagated forward (i.e., output o(t) and cell state c(t)) to the next node in the network.

Thus, LSTM networks are ideal for exploring how variation in one stock's price can affect the prices of several other stocks over a long period of time. They can also decide (in a dynamic fashion) for how long information about specific past trends in stock price movement needs to be retained in order to more accurately predict future trends in the variation of stock prices.

5.2 LINEAR REGRESSION

Linear regression can be used to find a relationship between two or more variables of interest and allows us to make predictions once these relationships are found. In simple linear regression, there are only two variables: one dependent variable and one independent variables.



Linear Regression in Machine Learning Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Linear regression is the simplest and most widely used statistical technique for predictive modelling. It basically gives us an equation, where we have our features as independent variables, on which our target variable [sales in our case] is dependent upon.

So what does the equation look like? Linear regression equation looks like this:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \dots \theta_n X_n$$

Here, x1, x2,….xn represent the freelance variables whereas the coefficients θ1, θ2, …. θn represent the weights.. Here, we have Y as our dependent variable, X's are the independent variables and all that as are the coefficients. Coefficients are basically the weights assigned to the features, based on their importance. For example, if we believe that sales of an item would have higher dependency upon the type of location as compared to size of store, it means that sales in a tier 1 city would be more than that of store size.

So, firstly let us try to understand linear regression with only one feature i.e., only one independent variable, therefore our equation becomes,

$$Y = \theta_1 * X + \theta_0$$

6 METHODOLOGY

Step1: Raw Stock Price Dataset Day-wise past stock prices of selected companies are collected from the BSE (Bombay Stock Exchange) official website. The obtained data contain five features they are

- **Date** of the observation
- **Opening** price of the stock
- **High** highest intra-day price reached by the stock
- **Low** lowest intra-day price reached by the stock
- **Volume** number if shares or contracts bought and sold in the market during the data
- **Open** interest how many futures contracts are currently outstanding in the market

The above data was than transformed into format suitable for use with our prediction model by performing the following steps:

TRANSFORMATION OF TIME-SERIES DATA INTO INPUT-OUTPUT COMPONENTS FOR SUPERVISED LEARNING

1. Scaling the data to the [-1,+1] range

Step2: Pre-processing

Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

This step incorporates the following:

a) Data discretization: Part of data reduction but with particular importance, especially for numerical data
b) Data transformation: Normalization Stock price prediction using LSTM on Indian Share Market
c) Data cleaning: Fill in missing values
d) Data integration: Integration of data files. After the dataset is transformed into a clean dataset, the dataset is divided into training and testing sets so as to evaluate.

Step3: Feature Selection: In this step, data attributes are chosen that are going to be fed to the neural network. In this study Date and Close Price are chosen as selected features.

Step 4: Train the NN model: The NN model is trained by feeding the training dataset. The model is initiated using random weights and biases. Proposed LSTM model consists of a sequential input layer followed by 3 LSTM layers and then a dense layer with activation. The output layer again consists of a dense layer with a linear activation function.

Step5: Output Generation: The RNN generated output is compared with the target values and error difference is calculated. The Backpropagation algorithm is used to minimize the error difference by adjusting the biases and weights of the neural network.

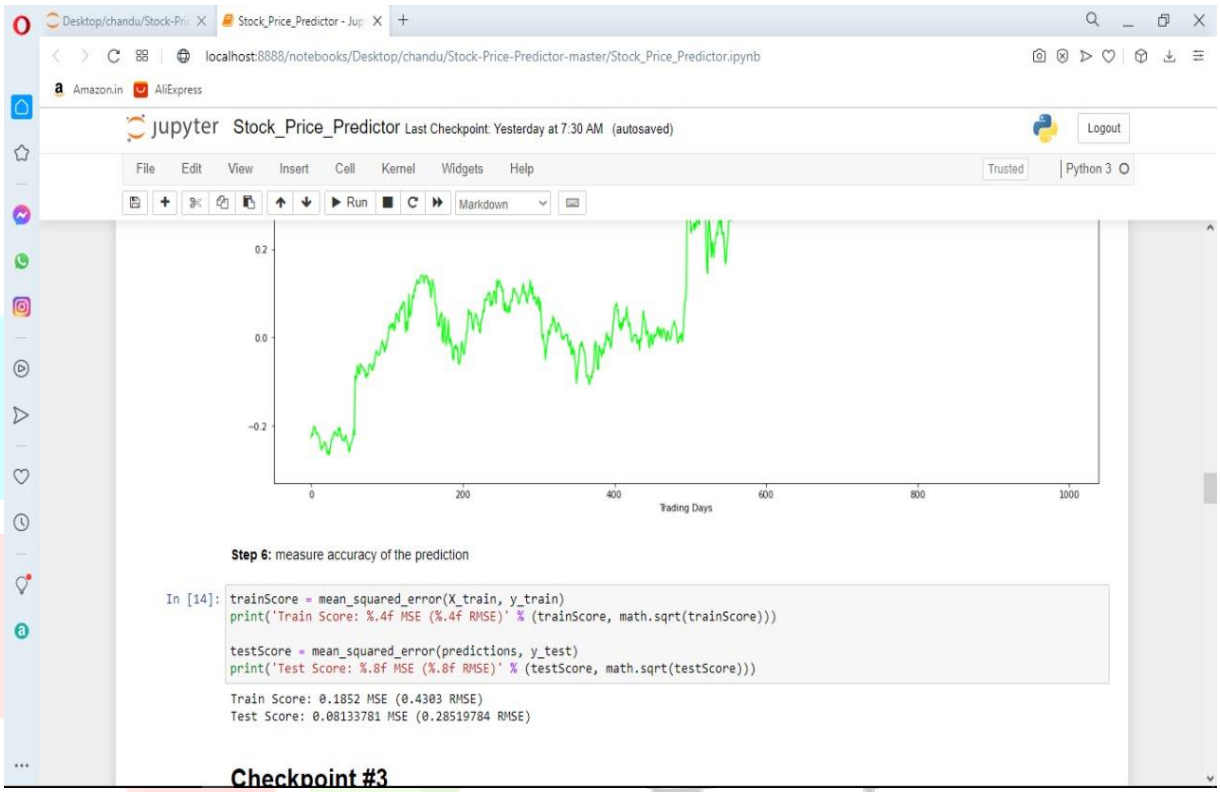
Step 6: Test Dataset Update: Step 2 is repeated for the test data set.

Step 7: Error and companies' net growth calculation: By calculating deviation we check the percentage of error of our prediction with respect to actual price. To calculate the error we use RMSE, Where RMSE is the Root Mean Square Error the error or the difference between the target and obtained output value is minimized by using RMSE value. RMSE is the square root of the mean/average of the square of all of the error. The use of RMSE is highly common and it makes an excellent general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

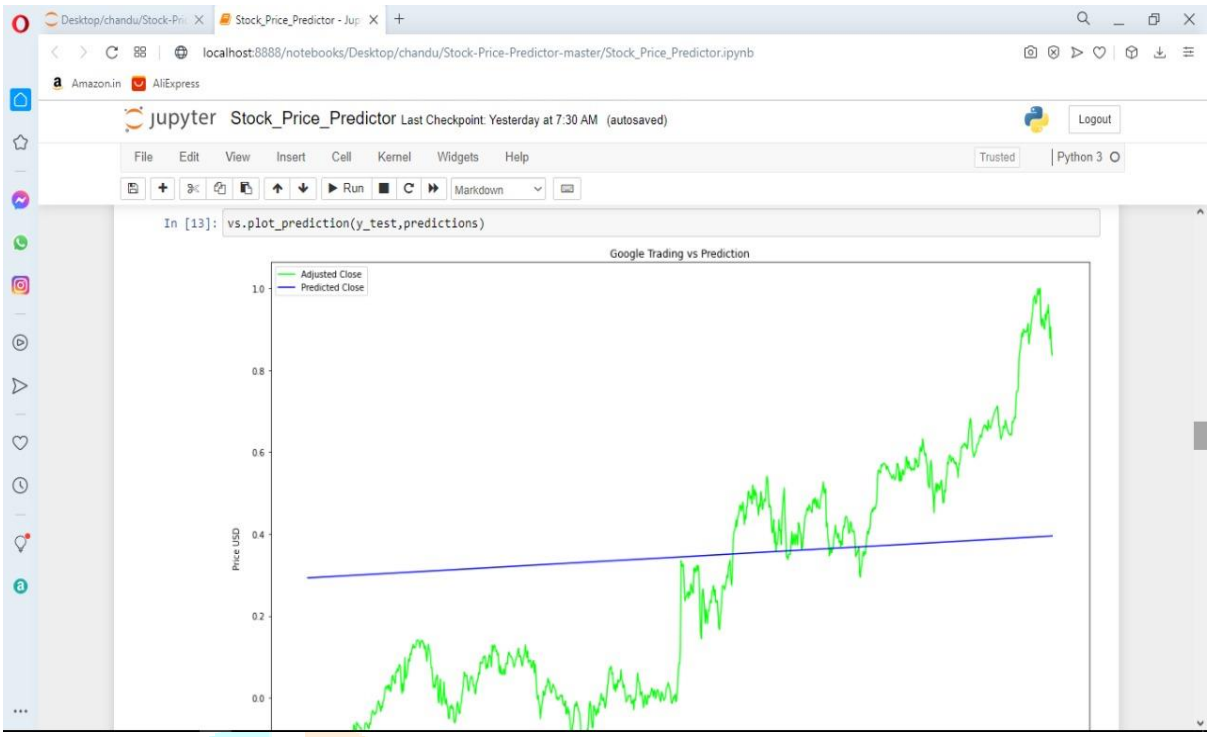
$$RMSE=\sqrt{1/N\sum(Y_i-Y_i)^2}$$

Step 8: Visualization: Using Keras[21] and their function APIs the prediction is visualized.

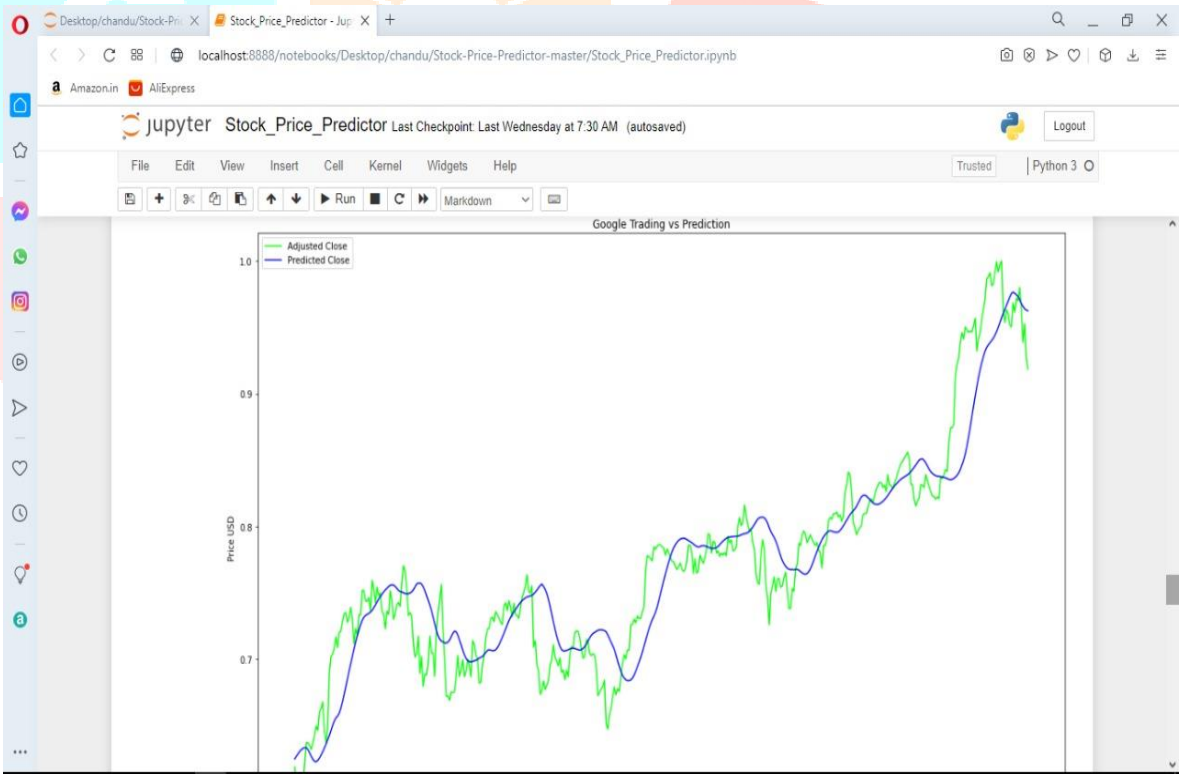
7 VISUALIZATION OF RESULTS



7.1 Visualization of the error rate using RMSE AND MSE



7.1 Linear Regression



7.2 Future Stock price Prediction using LSTM

CONCLUSION

Two techniques have been utilized in this paper: LSTM and Regression, on the Yahoo finance dataset. Both the techniques have shown an improvement in the accuracy of predictions, thereby yielding positive results. Use of recently introduced machine learning techniques in the prediction of stocks have yielded promising results and thereby marked the use of them in profitable exchange schemes. It has led to the conclusion that it is possible to predict stock market with more accuracy and efficiency using machine learning techniques. In the future, the stock market prediction system can be further improved by utilizing a much bigger dataset than the one being utilized currently. This would help to increase the accuracy of our prediction models. Furthermore, other models of Machine Learning could also be studied to check for the accuracy rate resulted by them.

REFERRENCES

[1] M.Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322-327.

[2] K. Raza, "Prediction of Stock Market performance by using machine learning techniques," 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), Karachi, 2017, pp. 1-1.

[3] Nazar, Nasrin Banu, and Radha Senthilkumar. “An online approach for feature selection for classification in big data.” Turkish Journal of Electrical Engineering & Computer Sciences 25.1 (2017): 163-171.

[4] Soulas, Eleftherios, and Dennis Shasha. “Online machine learning algorithms for currency exchange prediction.” Computer Science Department in New York University, Tech. Rep 31 (2013).

[5] Suresh, Harini, et al. “Clinical Intervention Prediction and Understanding using Deep Networks.” arXiv preprint arXiv:1705.08498 (2017)