# Stock Price Prediction using LSTM

**Group Members**

**Moiz Ullah Khan CS-172061**

**Muhammad Qasim CS-172029**

**Haider Naseer CS-172041**

**Abstract:**

Machine Learning is growing day by day because the problems never end. In the previous years we are using traditional programming concept for solving the problems. Now with the passage of time problems are becoming advance so the technologies to the solve the problem are also become advance. Machine Leaning are growing in health sector, business, climate change and classification-based problems.

Now we are moving towards our topic which is related to business. In this paper we are discussing the hot favorite top of industry which is stock price prediction. Researches and developers are exploring this topic from year 2010 till now. We are discussing different aspect of machine learning and the methodology we are using.

**Introduction:**

As I said earlier in this paper that the stock price is the hot favorite top of industry. Why is it? Because when any person put money in business, they want return. Similarly in stock market when you take share of 10,000 rupees (PKR. per share 20rs) you are expecting the return of 10,000 are more. Machine Learning are now going to easier this by predicting the stock price. Predicting the stock market and take the decision whether you buy more share or sell it.

In this paper we are discussing the methodology. We take the data from Kaggle which is the best platform for Artificial Intelligence, Machine Learning and Deep Learning practitioners. We have 13 years of Pakistan stock market from 2008 till 2021(February). This data has 7 columns which are defining different characteristics of data. The six columns are:

1. Date – Defining the stock price at a particular day, we set date as index.
2. Open – Defining the Opening of the stock market at particular date.
3. Close – Defining the closing of market at particular date.
4. High – Defining the highest-level market reaches.
5. Low – Defining the lowest level of market at particular date.
6. Change – Defining the change in market with respect to previous day.
7. Volume – Total market volume or total share buying and selling.

These are the columns which are defining and describe different properties of stock market. In our work we are going to predict the stock closing price so anyone can see that the market is going bearish or bullish trend.

**Methodology:**

**Workflow of Our Methodology**

**Step 01 Step 02 Step 03 Step 04 Step 05 Step 06**

Test

LSTM

Data

Predictions

EDA

Train

**Input Data Analysis and Preparation Results & Output**

Here is the complete workflow of how we are doing the things. There are six steps we have done in this project and they are divided into further parts also. Discussing the first Step which is Data.

**Data:**

13 years of statistics with date, opening, closing index, High and close price, change and total volume were mentioned in the dataset. We have the data from 2008 till 2021(February). As we have already stated above that our aim is to develop the framework to predict the stock closing prices of Pakistan stock Exchange (PSx-100). Every element which has been using in the dataset define its own characteristics. The elements are defined in the following:

(I) Date

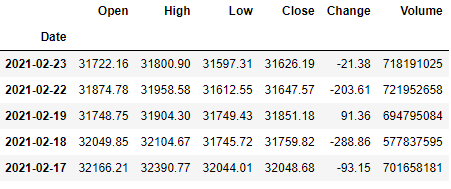
(ii) Open Price

(iii) Closing Price

(iv) High

(v) Low

(vi) Volume of market

****

**Figure 01**

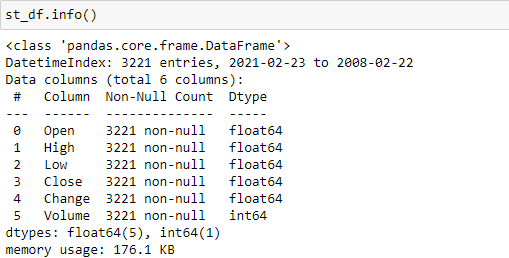
By using these six elements within the dataset. After the whole process of cleaning, we use these seven factors to predict the stock closing price. We use the regression approach in order to predict the closing price. We are moving towards our next step which is Exploratory Data Analysis – EDA.

**Step 02:**

**Exploratory Data Analysis – EDA:**

EDA which is the abbreviation of is Exploratory Data Analysis – EDA. After reading the data we apply some exploratory data analysis EDA in order to get the hidden pattern of data. We basically use the matplotlib and seaborn for data visualization. We take the correlation between the variables to check the dependencies over one another.

As the expert says that Exploratory Data Analysis is the most crucial part before implementing the model. Because the model is work on data and if the data is not accurate then our model is not accurate to predict the price. So there are steps in order to analysis, cleaning and process the data for getting the accurate results. The first steps we have to do is getting the information about the columns. So we have we use info function:



**Figure 02**

This function tells us about the complete information about data. The first the things we have found the total number of columns, associated the non-Null column. This defines the total non-null values in our columns. This function also defines the data type in front of each column.

Now with this function we have found that we have to make the date column as index. So, we know the everyday price of stock.

**Data Preparation:**

As you can see the above image the date column type is object so we need to change the type by using pandas library function to\_datetime () and then set date as index. Every column has object type so change the type of columns before moving onwards.

Datatype code figure

We have use to\_numeric so they can automatically get the suitable type for every column. The open, high, low, close and change column are float type because they contain floating point values and volume column data type is integer.

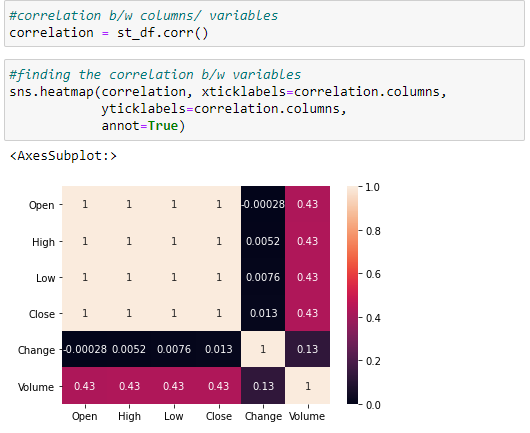
**Removing comma in prices:**

After changing the data type, we use head function to visualize the data. There are commas in the prices. So, we use replace function with regex parameter. Now our data have enough capable to do things.

We are finding that our data has null values or not. We use isnull function to get the total null values in each column of dataset. We remove the duplicates in our dataset because we have to away from overfitting the model. By dropping duplicates rows we use drop\_duplicate function.

**Visualization, Correlation and Boxplot:**

Now our data have enough to get the hidden patterns and visualize some results. we basically uses the matplotlib and seaborn for data visualization.

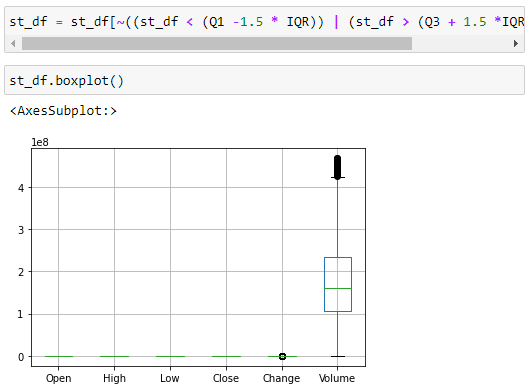


**Figure 03**

We take the correlation between the variables to check the dependencies over one another.

In the above diagram the dark color defines the negatively correlated and the light color shows the high correlation among variables. Open column is correlated with close variable. Open, high, close, low have same correlation which is 0.43 with volume. Soo, its mean that volume is depending on these factors. Change variable also correlated with volume variable by 0.13.

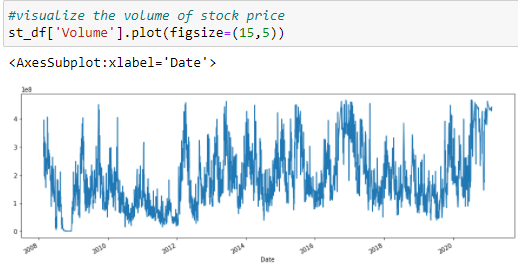
**Boxplot:**



**Figure 04**

We have used the describe function to get the description of data. When we compare the 75 percentiles with the max column there is a lot of difference between them. So, its mean there are outlier exits in our dataset. By using the boxplot, we found that the volume column of data contain the outliers so we take the quantile q1 and q3 to find out the inter quantile range. After inter quantile range we discard the data less than 25% and above 75% by using the or-bitwise operator. We again plot the boxplot by using seaborn, we somehow decreased the number of outliers in our dataset.

Visualization plays the important role for getting the better representation of data. We plot multiple visualization but we will highlight some visualization which are important for us.



**Figure 05**

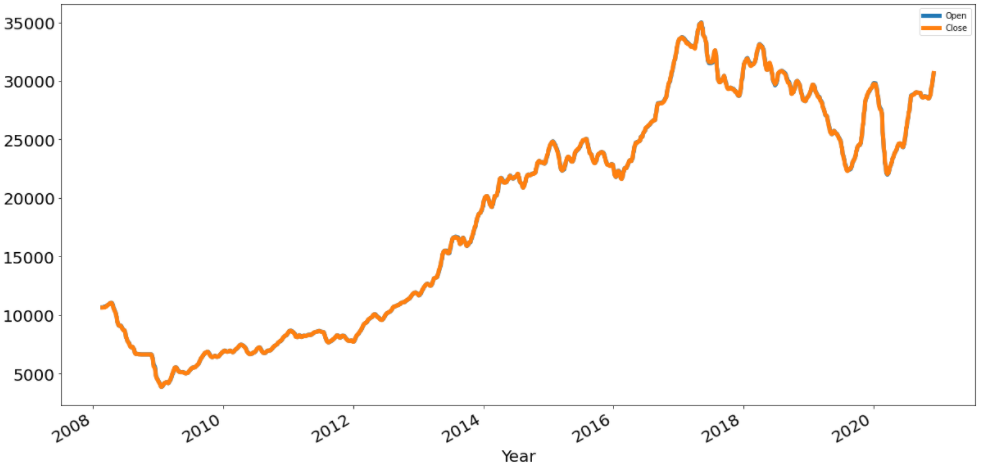
The first we visualize the volume column. We see that the volume column gives us the total volume of our market at a particular day. There is a lot of fluctuation in the volume.

**Open close:**

The last thing which we have want to share you is the open and close variable representation.

We have seen that the open and close values are almost same but there is a minor difference between these columns.

As you have seen that the stock market is crash in the 2009, we have shared the news so anyone knows that what is the reason behind that.



**Figure 06**

Now we are moving to make the train and test set. We import some libraries like Keras, sklearn and TensorFlow for multiple purposes. After all this now we have much understanding about data. We split the data into two measure parts. The first one is the training set and the other is test set. We scale the data by using the minmax scalar.

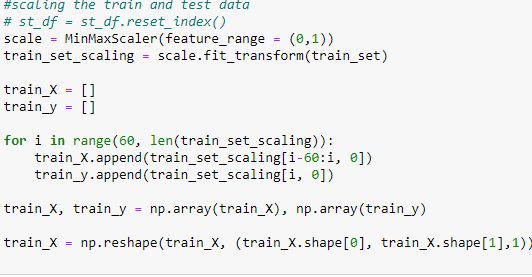
**Step 03:**

**Making Train Set:**

For training data set, we take the data and scale the data value by minmax scalar because data have some high and very low values.

Preprocessing the data is essential part so we scale the data in order to normalize it because to highlight the each and every feature of dataset. After that we give the training data

to model for training and the model will train on this dataset. After completing the training of our model, we move towards testing.



**Figure 07**

**Step 04:**

**LSTM Model:**

There are several models for predicting the stock prices but the thing is that few peoples know what model to apply on it.

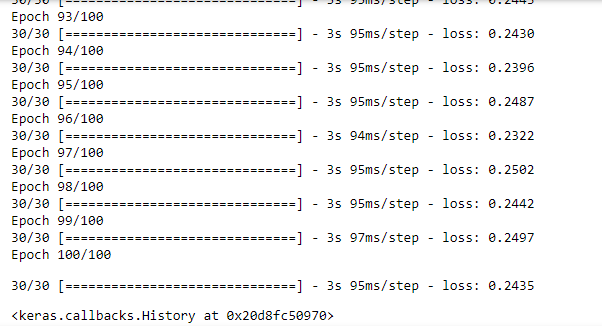
In the era of deep learning, people use convolutional neural networks for predicting the stock market price but the **CNN** almost takes 3 to 4 weeks of data to perform good training while the Long Short Term Model LSTM which we have uses takes the one week of data to give you the best output with good accuracy.

We first initiate the model and add up the first layer of 30 neurons and with 0.3 dropout rate. Similarly, we add the multiple layers of 50,100,50 and 30 neurons.

The output layers contain only one neuron because the predicted price is only one which is closing price. We use the two-activation function just before the output layer and on the output layer. On the second last layer we use rectified linear unit **relu** and on the output layer we use **SoftMax** because **SoftMax**

have better understand of value and then backpropagate it by checking the loss.





**Figure 08**

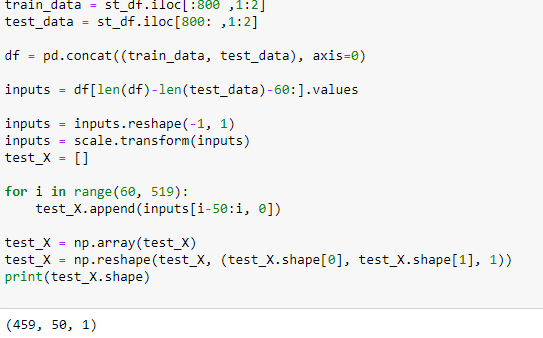
**Step 05:**

**Making Test Set:**

For testing data set, we take the data and scale the data value by minmax scalar because data have some high and very low values.

Preprocessing the data is essential part so we scale the data in order to normalize it because to highlight the each and every feature of dataset. After that we give the test data

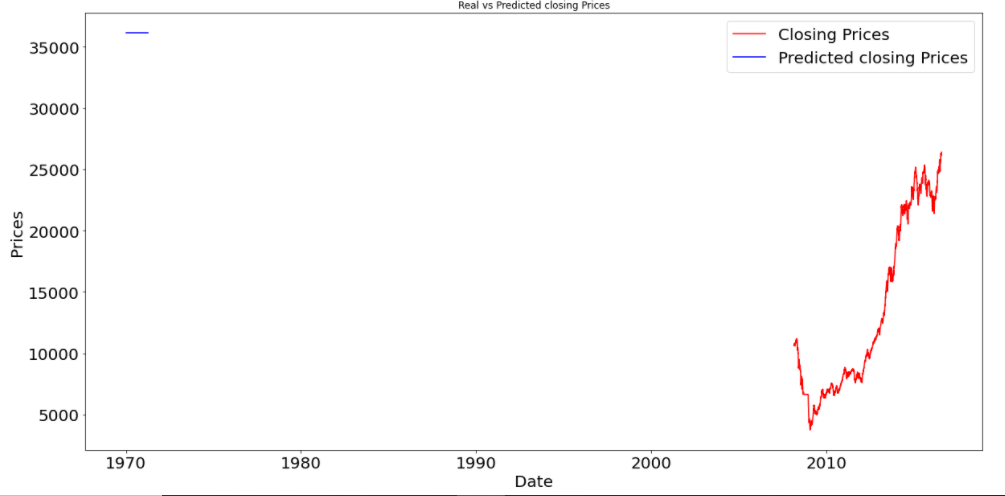
to model for evaluation and the model will give us good performance. By testing data model predict the closing price of stock market.



**Figure 09**

**Step 06 Final Step:**

**Predictions:**

****

**Figure 10**

Here is the prediction of our model. The x-axis shows the year and y-axis shows the closing price which our model is predicted. After all that analyzes, we are going to predict the markets closing price. We were using LSTM which are a part of Deep learning (branch of Machine Learning). Because LSTM takes one week data for prediction. LSTM gives us good result. The last closing price of stock market hit the 30,000 and our model predicted the closing price of stock market is 35,000 with the loss of 24% percent