Stock prices forecasting in USA during Covid-19 using Long Short-Term Memory (LSTM)

Milestone 2 Report

Aparna Krishna Bhat :- 1001255079

Naveena Mullapudi :- 1001821645

Mirat Gajera :- 1001829052

1 Introduction

One year ago, the U.S. stock market hit rock bottom after the coronavirus outbreak resulting in 30+ percent drops for each of the three major stock market indices from their previous peaks. But, a year later the world is still in crisis, but stock prices are near all-time highs.Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. In the midst of the COVID-19 pandemic and the dynamics of the global financial market during the year 2020, majority of global stock index reference indexes have experienced a significant decline. Forecasting stock prices has always been considered a challenging task due to the fact that the stock market tends to be non-stationary, nonlinear, and highly noisy. Artificial intelligence (AI) algorithms have been proven successful in solving problems such as predicting stock prices. Numerous factors influence financial market performance, and even financial experts find it complicated to make accurate predictions. This project seeks to utilize Deep Learning models, Long-Short Term Memory (LSTM) Neural Network algorithm, to predict stock prices.
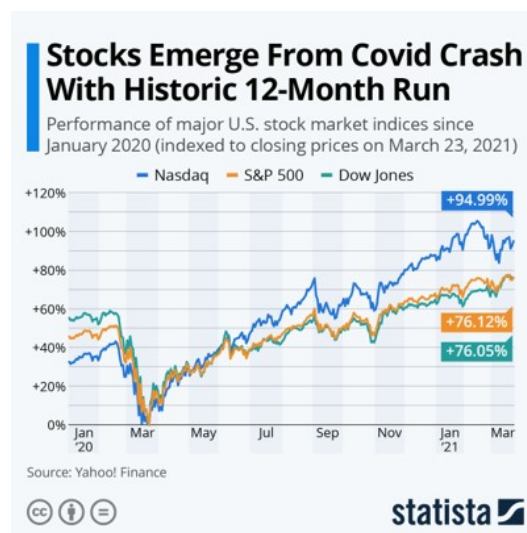

Figure 1: Stock market performance before and during COVID-19

2   Problem Statement and Goal

Problem statement:- Stock market process is full of uncertainty; hence stock prices forecasting is very important in finance and business. COVID-19 have had a significant influence on the global economy, as well as an impact on the financial markets. The challenge of this project is to accurately predict the future stock price of the given stock across a given period of time in the future considering the impact of COVID- 19 on the economy

Goal:- In this project, we are trying to investigate the impact of COVID-19 on the stock prices of companies like Amazon, Netflix, Airbnb, Airliners and Facebook on NASDAQ index. We will be making use of a Long Short Term Memory networks, Regression models to predict the future stock price of these companies on the NASDAQ using a data-set of past prices.

3   Datasets

1) Yahoo Finance:- Historical data of the stock prices. Yahoo Finance is the largest business and financial news site in the world, with unrivaled access to data, insights, and content.

2) NASDAQ Website:- Historical data of the stock prices.

3) Data of COVID19 confirmed cases are publicly available and operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

4   Sample Dataset

Date	Open	High	Low	Close	Adj Close	Volume
4/14/2021	3404.04	3404.13	3326	3333	3333	3145200
4/15/2021	3371	3397	3352	3379.09	3379.09	3233600
4/16/2021	3380	3406.8	3355.59	3399.44	3399.44	3183300

Figure 2: Amazon stock price

Date:- Day on which the stock is traded.

High:- Highest price at which a stock is traded during the course of the day.

Open:- Price of the first trade for any listed stock is its daily opening price.

Low:- Lowest price at which a stock trades over the course of a trading day.

Volume:- The number of shares or contracts traded in a security or an entire market during a given period of time.

5   Algorithms and Techniques

Use different machine learning and deep learning models available and compare them in terms of graphical analysis.

5.1   Linear Regression

Linear Regression algorithm is a supervised machine learning algorithm to model between dependent variable and one or more independent variables to produce a best fit linear line.

Two types of Linear Regression

1) Simple Linear Regression

Simple Linear Regression is used for finding the relationship between the independent variable and the dependent variable. The goal of Simple Linear Regression algorithm is to obtain a line that best fits the data which is achieved by minimizing the loss function.

2) Multiple Linear Regression

In real-life scenarios, there will never be a single variable that predicts a target.

Multiple Linear Regression is used for finding the relationship between dependent variables and one or more independent.

Linear Regression Model (Initial Findings)

Linear regression is an approach for predictive modeling to showcase the relationship between a scalar dependent variable 'Y', (in our case, we have 'Close' attribute) and one or more independent variables 'X' ('Date' attribute).
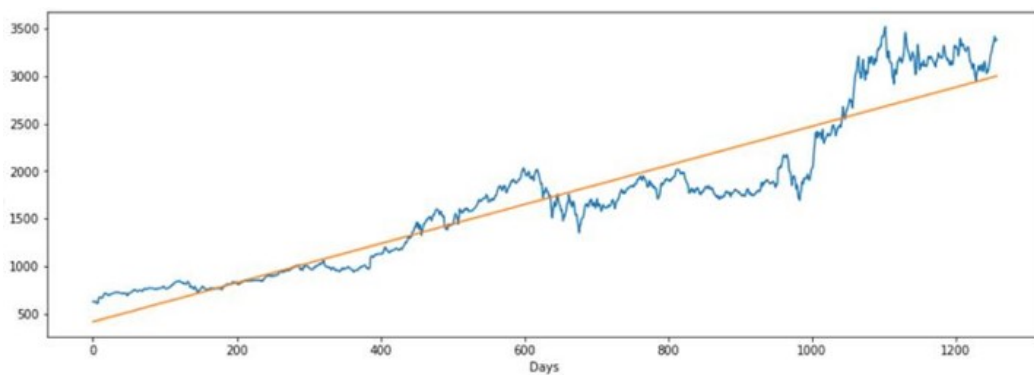

Figure 3: Linear Regression

Dataset Used

	Close	Year	Month	Week	Day	Dayofweek	Dayofyear
0	627.9	2016	4	16	19	1	110
1	632.99	2016	4	16	20	2	111
2	631	2016	4	16	21	3	112
3	620.5	2016	4	16	22	4	113
4	626.2	2016	4	17	25	0	116

Figure 4: Dataset For Linear Regression Model

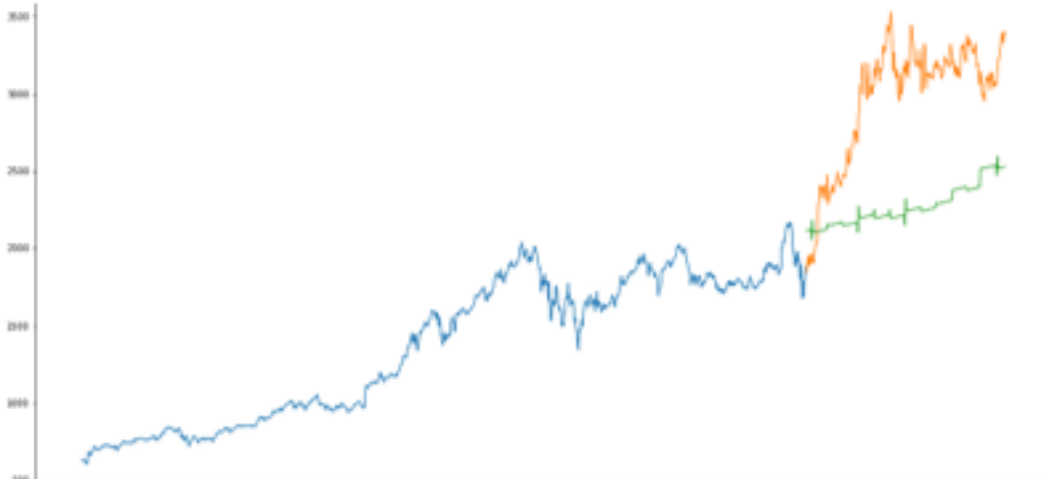**Evaluation Metric Used:- RMSE (Root Mean Square Error)**

Figure 5: Visualization between the actual value and the predicted value.

Blue curve:- Training Data, Orange Curve:- Test Data, Green Curve:- Predicted Data by the model

5.2 Long Short-Term Memory

According to my research, best possible model to use for solution is LSTM Neural Net Model.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.

Recurrent Neural Network + LSTM (How it works?)

The key to LSTM is the Memory cell state which stores the information. It runs straight down the entire chain.LSTM has the ability to remove or add information to these cell state, regulated by structures called gates. Gates are composed of Sigmoid neural net layer and a multiplication operation.

There are three gates to protect and control the cell states.

1) Input Gate:- The input gate adds information to the cell state.

2) Forget Gate:- The forget gate removes the information that is no longer required by the model

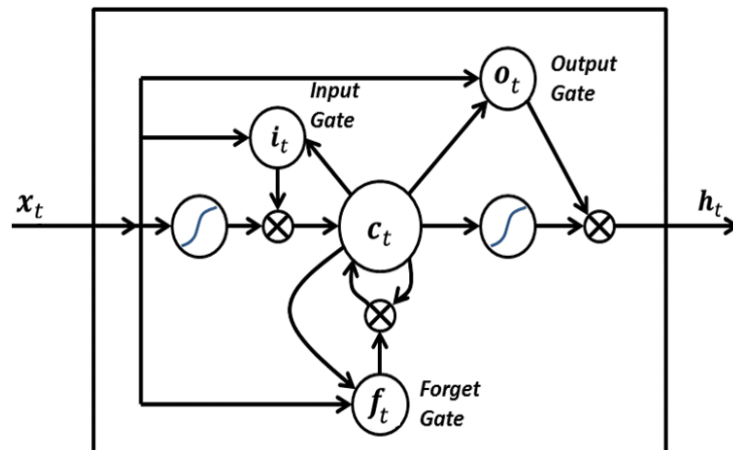**3) Output Gate:-** The output gate at LSTM selects the information to be shown as output.

Figure 6: The inner working of LSTM

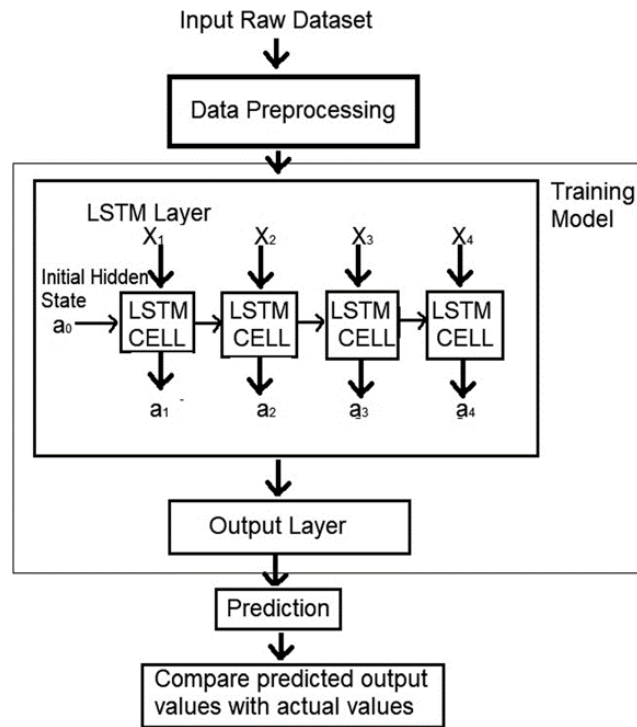4

LSTM Architecture and Initial Findings

Figure 7: Current LSTM architecture

Model Parameters

Stacked LSTMs

1) Loss function:- Mean Squared Error

2) Optimizer:- Adam
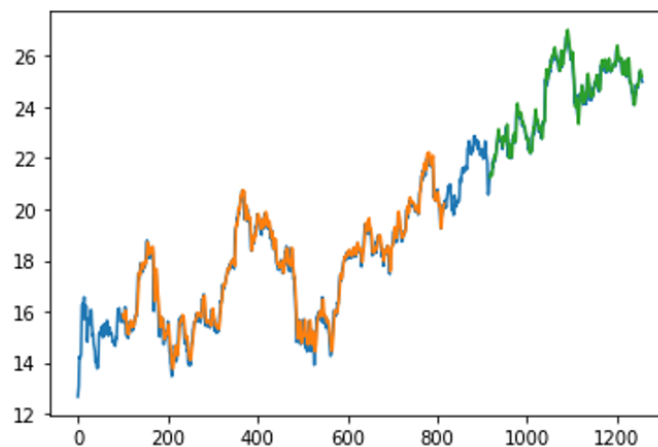
3) Epochs:- 100

4) Batch Size:- 64

Figure 8: Averaging (High + Low) / 2

Blue curve - Original data, Orange curve - Training model prediction, Green curve - Testing model prediction
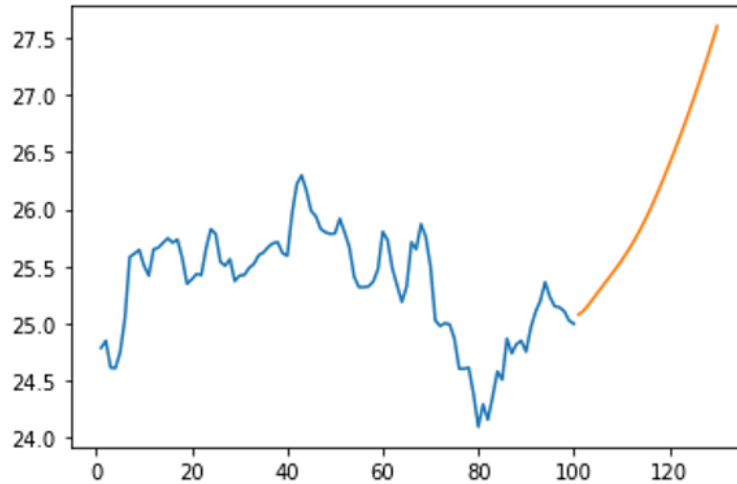
Figure 9: **Blue curve - last 100 days data, Orange curve - next 30 days prediction**

6 Challenges

* Feature selection for the stock price prediction

* Understanding the historical CSV data obtained form Yahoo Finance.

* Analyzing the obtained results to draw preliminary conclusions.

* Modifying stacked LSTM to function better.

* Choosing the correct hyper parameters for building better models.

7 Next Steps for Next Milestone

Proposed architecture for LSTM
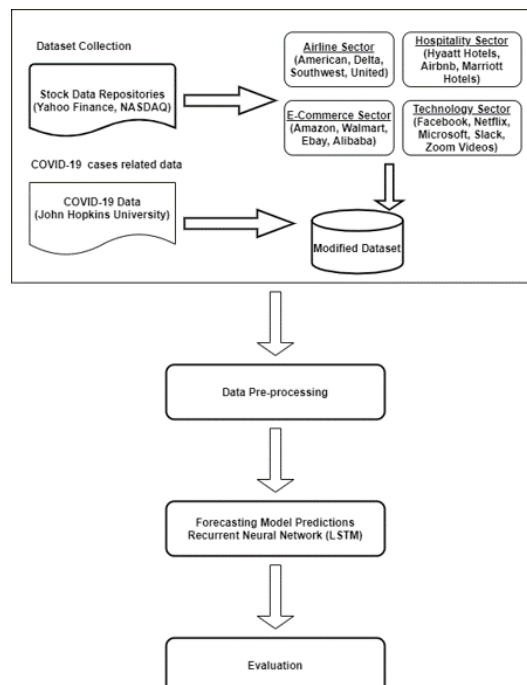
* Integration of COVID-19 data.

* Comparing the results for LSTM and Linear Regression Models.

* Drawing the insights and conclusions from the comparison.

8   Research Papers (References)

1) Impact of COVID-19 on Forecasting Stock Prices: An Integration of Stationary Wavelet Transform and Bidirectional Long Short-Term Memory

https://downloads.hindawi.com/journals/complexity/2020/1846926.pdf

2) Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM).

https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-021-00430-0.pdf