Analyzing and Predicting US Stocks amid Coronavirus to understand its impact on various industries

1st Chirag Sharma
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, United States
csharma3@stevens.edu

2nd Prem Patel
Dept. of Information Systems
Stevens Institute of Technology
Hoboken, United States
ppate101@stevens.edu

3rd Shreya Bhargava
Dept. of Information Systems
Stevens Institute of Technology
Hoboken, United States
sbharga2@stevens.edu

***Abstract*—The project is designed to predict stocks from various industries based on the potential effects that the coronavirus (COVID-19), will have on the economy. We would like to recommend stocks in a way such that an individual could profit off the market amidst the global viral outbreak. It is important to consider past outbreaks and come to a conclusion that the market will react adversely to such situations in the short run however the market does eventually correct itself in the long run.**

***Index Terms*—COVID-19, coronavirus, stock market, economy, industries, machine learning, linear regression, long short term memory, support vector machine, dataset, data mining**

I. Introduction

A. *COVID-19*

Coronavirus is an infectious disease caused by a newly discovered coronavirus. Most people infected with COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. Protect yourself and others from infection by washing your hands or using an alcohol based rub frequently and not touching your face. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. At this time, there are no specific vaccines or treatments for COVID-19. However, there are many ongoing clinical trials evaluating potential treatment. [1]

B. *Financial Crisis*

As countries implement necessary quarantines and social distancing practices to contain the pandemic, the world has been put in a Great Lockdown. In addition, many countries now face multiple crises—a health crisis, a financial crisis, and a collapse in commodity prices, which interact in complex ways. Under the assumption that the pandemic and required containment peaks in the second quarter for most countries in the world, the World Economic Outlook project global growth to fall by 3 percent. This is a downgrade of 6.3 percentage points from January 2020, a major revision over a very short period. [2] Emerging market and developing economies face additional challenges with unprecedented reversals in capital flows as global risk appetite wanes, and currency pressures, while coping with weaker health systems, and more limited fiscal space to provide support. Moreover, several economies entered this crisis in a vulnerable state with sluggish growth and high debt levels.

C. *The Project*

As the coronavirus began to spread across the world, the volatility of the market has reached levels comparable with the Global Financial Crisis of 2008. Falling share prices reflect expectations of future profits, and the virus has been dampening economic activity which could result in worse corporate profits. This pandemic has not only created more churn in the market but has also amplified uncertainty. Two months later, the United States became the first country in the world with more than 100,000 cases, the economy has ground to a near standstill, and the virus has killed more than 1,000 people in New York State alone.This is how the major stock index (Dow Jones) has been reacting to the major outbreak.

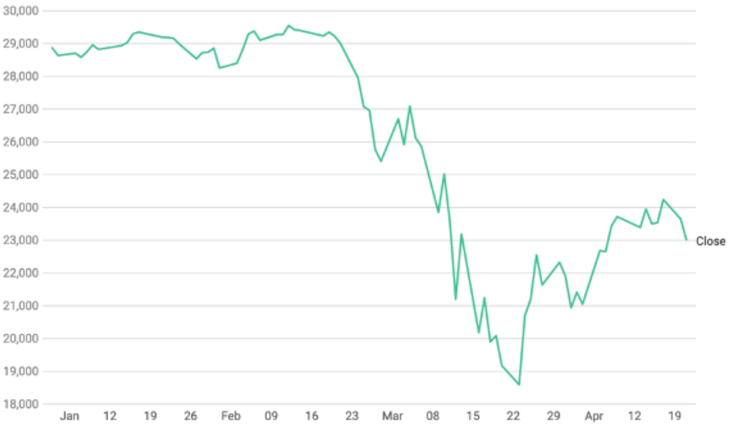


Fig. 1. Dow Jones Index value in 2020

Gradually the cases, threat of COVID-19 increased and stocks tumbled resulting in market volatility. The economic impact of this pandemic has introduced extraordinary volatility in the global financial markets, and it is hard for investors to operate during this crisis. [3] Keeping this pandemic and its threat in mind, through our project we aim at analyzing and predicting stocks across the States alongside COVID-19 to share the best practice and mitigate risk for market-makers to operate and exchange trade.

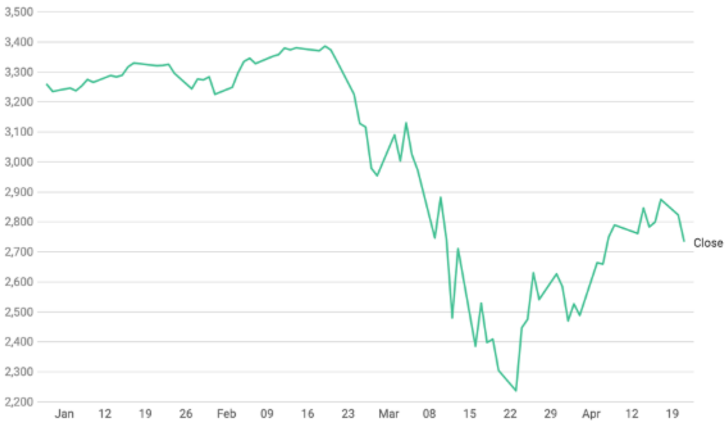


Fig. 2. S&P values in 2020

On the broader-based S&P 500 shed 226 points, respectively, which blew their previous largest single-day point declines out of the water. On a percentage basis, the Nasdaq gave back 7.29%, its 10th-biggest percentage loss in history, whereas the S&P 500 gave up 7.6% for its 17th-largest percentage loss of all time.

II. Literature Survey

A. *Solutions to the Pandemic*

For COVID-19 there is a lot of ongoing research and the daily numbers across the world have been used to make charts and figures to examine its trend. The data is available across different states which can be ploted to derive relationships between death, confirmed cases and recovered patients with time using Tableau. There are dashboards prepared to map the confirmed deaths across various countries and how rapidly they this grows. Using the COVID-19 data, many projects aim to analyze the effect of this outbreak on different sectors such as Information Tcehnology, consumer utilization, telecommunications and Helath care.

B. *Solutions for Stock Market Prediction*

Stock market prediction has been existing since long however it has either predicted the future value of one specific company or the value of indexes.

a) Particle Swarm optimization: PSO has been used mainly because of its intuitiveness, ease of implementation and ability to solve nonlinear optimization problem. The basic principle is that it moves from a set of points to another set of points in a single iteration improving and combining deterministic and probabilistic rules. However the PSO algorithm has a very low convergence rate in case of complex problems such as stock prediction and easily tends to fall into local optimum in high dimensional space. PSO cannot avoid the problem of scattering and furthermore it is difficult to implement it with initial design parameters.

b) Multi-Score Multiple nstance Learning: The modern web has proved to be a very useful tool in making task easier. The interconnection of data it is easier to establish relationships between variou svaribales and rougly scope out a pattern of regular investments. There have bene similarities iin investment patterns and the key is to successfully predict stocks, expliting same consistencies between the data. The way stock market information can be predicted successfully is by using more than just technical historical data, and using sentiment analyzer to derive emotions to a certain stock.

We look forward to predict stocks based on industry and how adversely it has been affected by the outbreak. Since, COVID-19 is fairly new, there is no existing solution available which integrates Covid-19 and its impact on stock market and loss of economy.

III. Implementation

A. *Description of the Dataset*

To implement and get results for our desired project we have gathered three different datasets, each of them serving different purposes at different stages of our analysis. The first dataset of COVID-19 includes columns like date, confirmed, recovered and number of deaths across the world. Since we are only interested in analyzing data for the US, all other country data is not used for analysis. After filtering our data, we calculated the number of new cases from last week for each day. This will be plotted against the total number of cases. We have not used the traditional method of plotting the data against time as COVID-19 spread is not dependent on time but on the actual spread of the disease – new cases from last week. We have achieved the following graph.

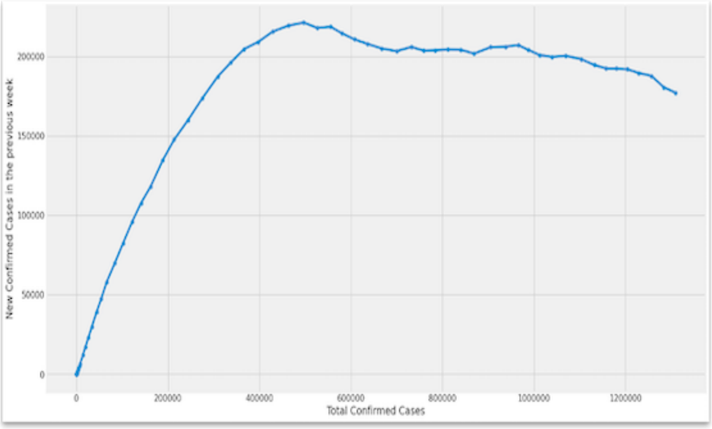


Fig. 3. New Confirmed Cases vs Total Confirmed Cases

The purpose behind this plot is to understand the confidence level that we could have on our further analysis. As the graph suggests, the number of new cases have started to decrease and as a result we can assume that our predictions would be more accurate as the ambiguity if not reduced is under control. We extracted a static data of S&P 500 companies to give us a list of companies, its ticker (stock market symbol) and their industry. The S&P is a free float, capitalization weighted index of the top 500 publicly listed stocks in the US. We extracted the stock market stats for all these 500 companies from a Python module called yahoo-finance. This data consisted daily company's stock details. The details included data, open, close, high, adj-close, volume columns. Combining these columns with COVID-19 confirmed cases data, we have tried to predict adj-close column for selected companies. We have also considered company's industry while drawing insights from our analysis.

Although the data gathered in the previous section is certainly a good start, with missing intradya prices that is the prices minute by minute. It is certainly possible to guide and recommend on interday level. However, intraday prices are considered as a commodity in itself and not readily available.One more missing attribute could be order book. It is the live trading of a particular stock and consist of amount of stocks each trader will buy or sell. For instance, the predicted price of a stock can be the weighted average of orders.

B. Machine Learning Algorithms Considered

In time – series analysis and statistical modelling, regression analysis is a process used for estimating relationships between one or more independent variables (also termed as 'predictors', 'covariates', 'features') and the dependent variables (also termed as 'outcome variable', 'target'). One of the primary purposes of regression analysis is prediction and forecasting, which is the process of making predictions based on the past historical and present data and mostly analysis of trends.

- Linear Regression

Here we try to find a best fit line that fits the data very closely according to a specific mathematical criterion. We can optimize the fit of this line even further by using the Gradient Descent Method. This helps us to estimate the conditional expectation of the dependent variable when the independent variables take on a given set of values. It is a type of regression analysis which is studied rigorously and is most used in practical applications. If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.

- Support Vector Machine

Particularly one of its extension termed as Support Vector Regression (SVR) in order to compare the results and accuracy of the two algorithms and help us predict better. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

- Long Short Term Memory

This gives an edge over conventional feed-forward neural networks and Recurrent Neural Networks (RNN) because of their property of selectively remembering patterns for long durations of time, making them idea for stock market prediction where there is a long duration history involved. RNNs have a hard time carrying information if the sequence is too long as they suffer from a short-term memory. During back propagation, RNNs suffer from the vanishing gradient problem [4]. A small gradient with a small value does not contribute much to the learning, making the model redundant. All these issues are resolved by using LSTM models for stock market prediction.

C. Data Cleaning and Preprocessing

- COVID-19 data has been collected through the URL Request method to download the csv and dump it into a Python dataframe. We are saving the files on the local machines to keep a copy of the latest file used as these datasets are updated daily.
- There are three different dataframes created for Confirmed, Recovered, and Deaths with Province/State, Country/Region, Last Update as common columns in all three. We remove unnecessary columns and only keep Date, Confirmed, Recovered, and Death; merging all three columns into one dataframe using pandas.merge.
- Since, the S&P 500 companies' data is static, we load this data from our local machine through pandas.read_csv method and dump the company symbols in an array while creating another Python dataframe that includes name of the company, it's ticker (symbol for stock market), and the industry that it belongs to.
- For stock market data, we are using yahoo-finance module to download data for all 500 companies by looping in the symbol list created earlier and create new columns for each company's Open, Low, High, Date, Close, and Adj_Close columns from 1st Jan, 2018 to present. Some companies' data is not found as they might be delisted.
- All the above data is dumped into one single Python dataframe with Open, Low, High, Close, and Adj_Close columns for each company with distinct names and a common Date column.
- To short list companies from about 500 companies, we will be taking companies with highest and lowest average Adj_Close for each Sector. This is to extract the best and the worst fairing company since 2018 for each sector to get a fair idea of an industry.
- Using the shortlisted companies, we download their stock market data again and append confirmed COVID-19 cases data. As the data is from 2018, there is no COVID-19 data present in the major first portion of the data. We will fill those rows with zeros for Confirmed COVID-19 cases.

- Now that our data is ready to be dumped into a model, we will use the three models – Linear Regression, Support Vector Regression, and Long Short-Term Memory to perform our analysis.
- We will compare results of every company and each industry for all models and draw our insights based on the analysis.

IV. Algorithm Implementation

A. Linear Regression

Linear regression can be the most basic machine learning algorithm used for stock market prediction. It helps in giving a continuous output between variables. The dependent variable here was 'adjusted close' value and the independent variables were 'open', 'close', 'volume', 'high', 'low' and 'confirmed corona cases'. The data was divided into train and test explicitly to check how our model performs. After teh model was developed using the train data, we tested the model through our test data and the predictions were accurate but were proned to overfit. We then calculated its accuracy through mean square error, root mean square error and variance. The closer the variance to value '1', the better the model is said to me. The root mean square error should be as minimal as possible.

B. Support Vector Regression

Support Vector Machines makes uses of labelled training data and is classified as a discriminative classifier. It constitutes of a hyperplane which is a boundary for the new dataset. They are considered to be associated learning algorithms making use of supervised learning models. The tuning parameters for Support vector regression are:

- Kernel: they are categorized as linear, polynomial and radial basis function kernels calculates the prediction line. Linear kernels work as dot product of the input and the support vector. We have used radial basis function kernel as linear and poly model didn't provide realistic results as compared to rbf.
- C Parameter: Regularization parameter determines how many misclassifications are we ready to accommodate. It evaluates the accuracy of our model. We have chosen C=1e3, the default value. Lower the regularization value, less the misclassification.
- Gamma Parameter: Measures the influence with every single training data is added to the model. High gamma value describes the closeness form the boundary whereas low gamma value tell us how far they are from the plausible margin.

The goal of using support vector regression is to identify n- dimensional space into differentiable categories. There are multiple hyperplanes out of which we choose the best one that distinguishes our data. We use this in order to maximise our margins (distance between data points of two classes). The only benefit in maximising the margin is that it provides reinforcement which enables quicker classification of future data [5]. Based on their closeness, the data points are attributed to different classes. The dimension of hyperplanes depend on the number of features, in our case (Confirmed cases, open, high, low, volume).

C. Long Short Term Memory

From the results achieved with the help of the Regression Models – Linear Regression and Support Vector Regression, we can see how they are highly prone to outliers and noise, as a result, they are highly overfitted. To overcome these problems, we agreed upon the third algorithm to be a Neural Network Algorithm.

We cannot use the conventional Feed – Forward Neural Networks or the Recurrent Neural Networks (RNNs) as they are fundamentally developed to work on smaller sets of data and have a Short – Term Memory. Also, RNNs are highly prone to the Exploding Gradient Problem and the Vanishing Gradient Problem. The dataset that we have includes historical data for the stock market price details of the companies. Using the Long Short – Term Memory Neural Network serves our purpose of training this large dataset as it is not vulnerable to the problems discussed above.

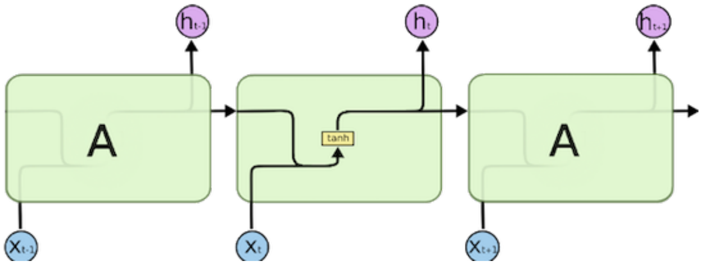


Fig. 4. Architecture of RNN

An LSTM has a similar control flow as a recurrent neural network. It processes data passing on information as it propagates forward. The differences are the operations within the LSTM's cells. The core concept of LSTM's is the cell state, and its various gates. The cell state act as a transport highway that transfers relative information all the way down the sequence chain. Think of it as the "memory" of the network. The cell state can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make its way to later time steps, reducing the effects of short-term memory. As the cell state goes on information gets added or removed to the cell state via gates. The gates are different neural networks that decide which information is allowed on the cell state. The gates can learn what information is relevant to keep or forget during training.

First, we have the forget gate. This gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep. To update the cell state, we have the input gate. First, we pass the previous hidden state and current input into a sigmoid function. That decides which values will

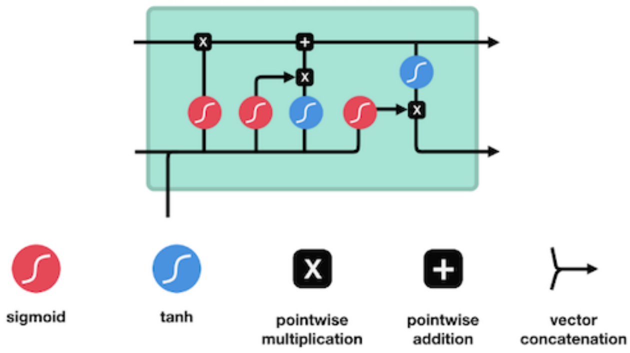
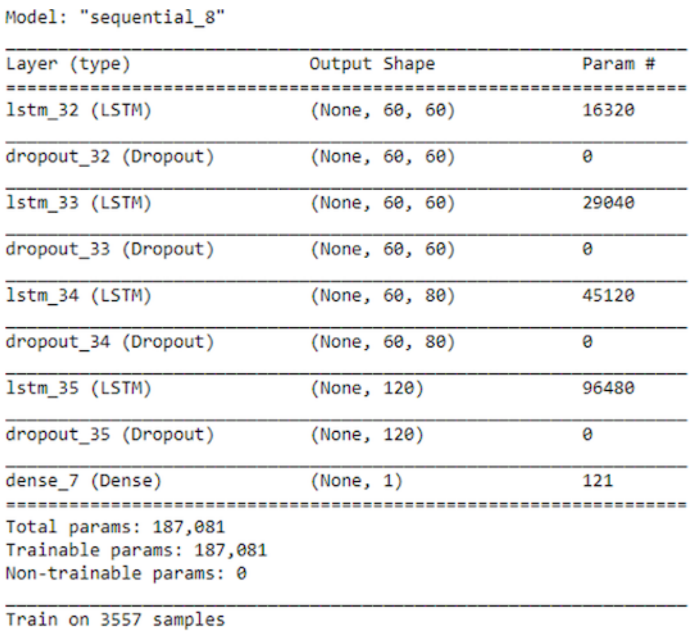


Fig. 5. LSTM Cell and its operation

be updated by transforming the values to be between 0 and 1. 0 means not important, and 1 means important. Now we should have enough information to calculate the cell state. First, the cell state gets pointwise multiplied by the forget vector. This has a possibility of dropping values in the cell state if it gets multiplied by values near 0. Then we take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant. That gives us our new cell state. Last, we have the output gate. The output gate decides what the next hidden state should be. The hidden state contains information on previous inputs. The hidden state is also used for predictions. First, we pass the previous hidden state and the current input into a sigmoid function. Then we pass the newly modified cell state to the Rectified Linear Unit ‘ReLU’ function. We multiply the relu output with the sigmoid output to decide what information the hidden state should carry. The output is the hidden state. The new cell state and the new hidden is then carried over to the next time step. For our project, we have built the LSTM model as follows:

- It is a sequential model with a stack of 4 LSTM Layers and a single Dense Layer. There is a Dropout Layer included after each LSTM layer for better generalization and avoid overfitting
- The activation function used for all the LSTM layers is Rectified Linear Unit ‘ReLU’
- The first LSTM layer consists of 60 memory units and it returns sequences to ensure that the next LSTM layer receives sequences and not just randomly scattered data
- The second, third and fourth LSTM layer consists of 60, 80 and 120 memory units respectively
- All the Dropout Layers are set to drop 20
- The Dense Layer consists of a single memory unit and the activation function used is ‘Linear’
- To compile the model ‘Adam’ optimizer is used along with the loss function being ‘Mean Squared Error



```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
lstm_32 (LSTM)               (None, 60, 60)            16320     
_________________________________________________________________
dropout_32 (Dropout)         (None, 60, 60)            0         
_________________________________________________________________
lstm_33 (LSTM)               (None, 60, 60)            29040     
_________________________________________________________________
dropout_33 (Dropout)         (None, 60, 60)            0         
_________________________________________________________________
lstm_34 (LSTM)               (None, 60, 80)            45120     
_________________________________________________________________
dropout_34 (Dropout)         (None, 60, 80)            0         
_________________________________________________________________
lstm_35 (LSTM)               (None, 120)               96480     
_________________________________________________________________
dropout_35 (Dropout)         (None, 120)               0         
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 121       
=================================================================
Total params: 187,081
Trainable params: 187,081
Non-trainable params: 0
_________________________________________________________________
Train on 3557 samples
```

Fig. 6. LSTM Model Summary

V. Results and Comparison

A. Results through Linear Regression

From the plots we can see that the predicted values fit the actual values a bit to closely, which can be due to overfitting. Also, it was very difficult to establish a stable relationship between the dependent and independent variables. The metrics we used here are Mean Squared Error (MSE) and Variance Score. By taking in these metrics we can infer that the company Google belonging to the Information Technology Sector has been moderately impacted by COVID – 19 with high MSE and high Variance Score. COVID – 19 had a low impact on Mettler Toledo from the Healthcare Sector with high MSE and low Variance score. Whereas, Verizon from Telecommunications sector was highly impacted by COVID – 19 with low MSE and low Variance Score.

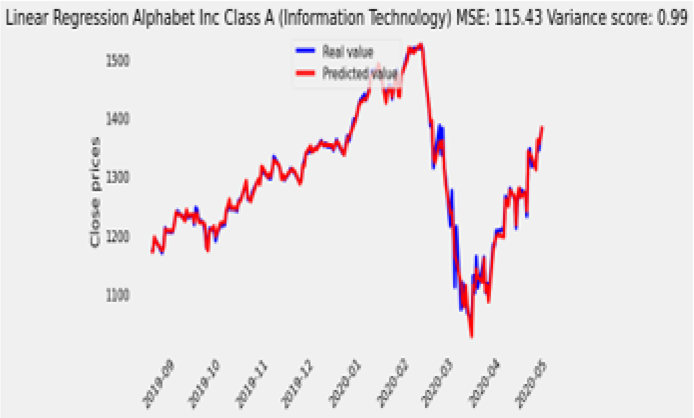


Fig. 7. Linear Regression: Information Technology

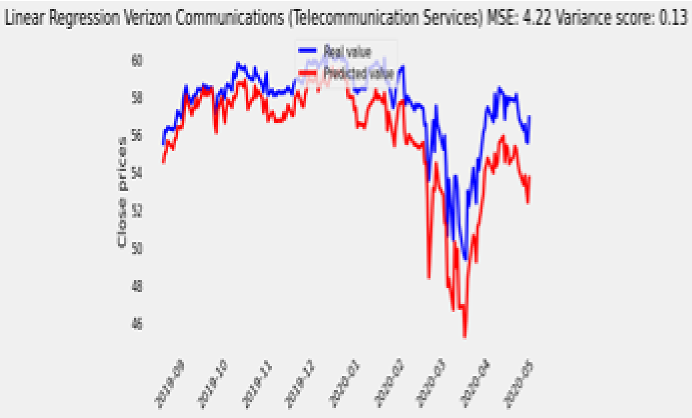


Fig. 8. Linear Regression: Telecommunications

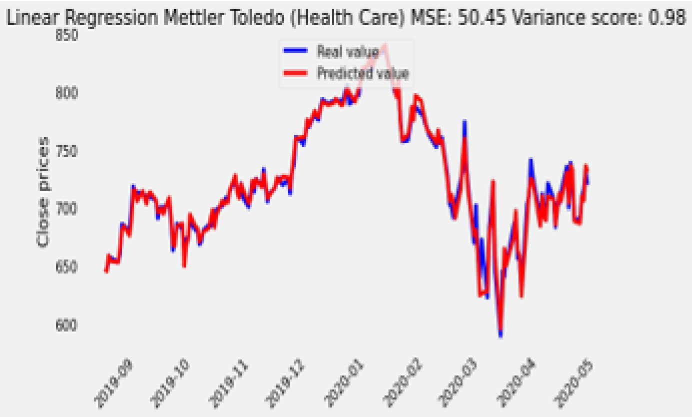


Fig. 9. Linear Regression: Healthcare

B. Results through Support Vector Machine

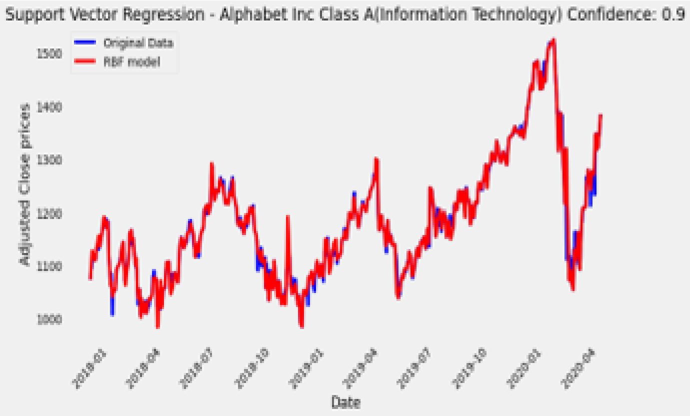


Fig. 10. SVR: Information Technology

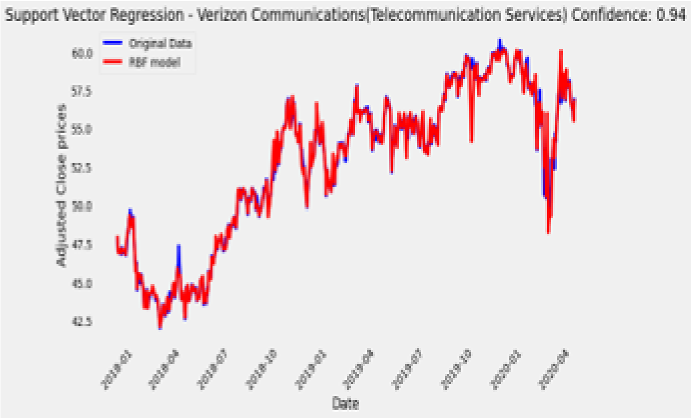


Fig. 11. SVR: Telecommunications

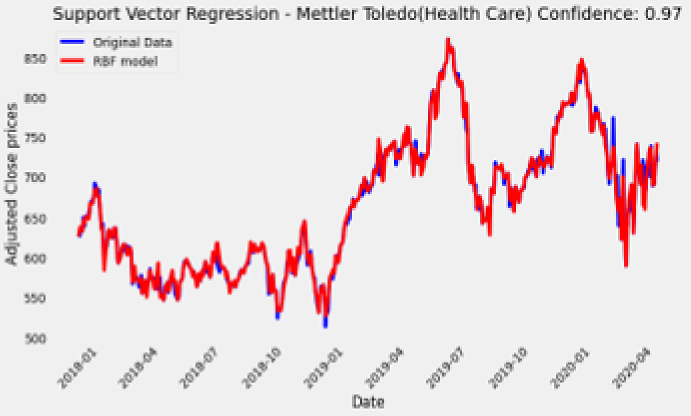


Fig. 12. SVR: Healthcare

From the plots of Support vector Regression we can see that even here there is overfitting of the predicted values. This can be due to outliers. The metric used here is Confidence score (this is the 'r2̂' score). We can infer that the Information Technology Sector is highly impacted by COVID – 19, Telecommunications Services is moderately impacted and there is low impact of COVID – 19 on the Healthcare sector.

C. Results through LSTM

After compilation, we train the model for 50 epochs to achieve the following results:

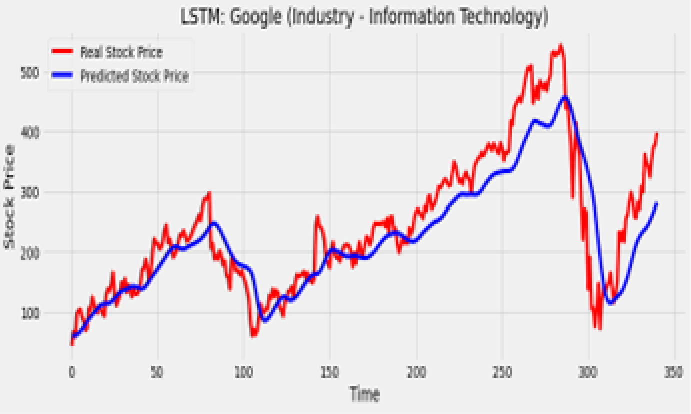
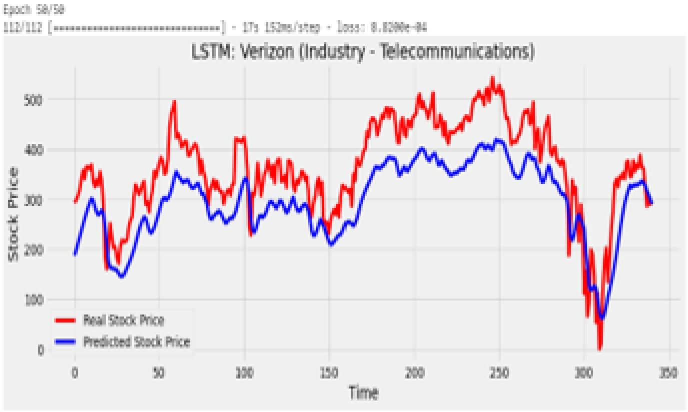
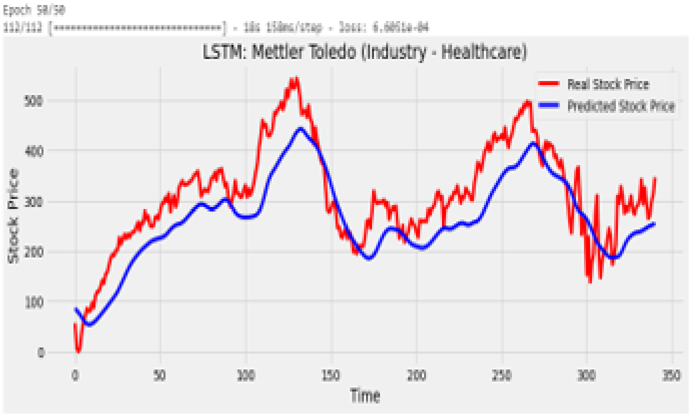


Fig. 13. LSTM: Information Technology



```
Epoch 50/50
112/112 [==============================] - 17s 152ms/step - loss: 8.8200e-04
```

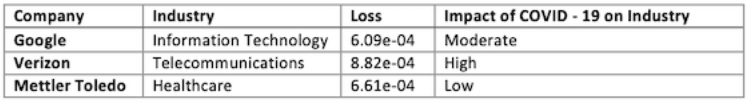
Fig. 14. LSTM: Telecommunications



```
Epoch 50/50
112/112 [==============================] - 18s 158ms/step - loss: 6.6051e-04
```

Fig. 15. LSTM: Healthcare

From the above plots for three companies belonging to different industrial sectors, we can infer that, there is no overfitting as seen in the case of Linear Regression and Support Vector Regression (SVR). Also, the value for the metric 'loss' is very low after training for the last epoch. The loss for the three companies and the impact of COVID – 19 on their respective industries are summarised below:



Company	Industry	Loss	Impact of COVID - 19 on Industry
Google	Information Technology	6.09e-04	Moderate
Verizon	Telecommunications	8.82e-04	High
Mettler Toledo	Healthcare	6.61e-04	Low

Fig. 16. LSTM: Model Loss

VI. Future Enhancement

The project can be improved with adding more financial statistical parameters like company's financial ratios and dividends while also including sentimental analysis of the company but monitoring company's overall sentiment on Twitter and other social medias. These parameters are considered to increase the accuracy. The whole project can be moved to cloud to automize data collection, processing, running the algorithms, creating visual reports and timely exporting them to a dashboard. The dashboard can have custom filters like companies, industries, and impact. Push notifications about alerts regarding companies to perform well and poor taking into the account of our analysis.

VII. Conclusion

COVID-19 crisis is stable than before, and results are more reliable because of curve dipping/flattening. Our findings tell us that Linear Regression and Support Vector Regressors give accurate results at the cost of overfitting the data heavily. Linear Regression does not give reliable results with fewer features. Support Vector Regression does not support perform against large datasets and noise. Industries like Healthcare and Telecommunication are relatively less affected by the COVID-19 crisis while Information Technology and Consumer Discretionary are relatively highly affected. LSTM proves to be the best algorithm for stock market prediction. However, the loss drastically increases with decrease in quantity of data and iterations. The computational power required is also high. The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data with improved accuracy and has been chosen after being tested on a sample data. The project demonstrates the machine learning model to predict the stock value with more accuracy as compared to previously implemented machine learning models.

References

[1] https://www.who.int/health-topics/coronavirustab=tab_1

[2] https://blogs.imf.org/2020/04/14/the-great-lockdown-worst-economic-downturn-since-the-great-depression/

[3] https://www.marketplace.org/2020/04/17/how-the-markets-are-reacting-to-covid-19/

[4] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[5] https://www.ijeat.org/wp-content/uploads/papers/v8i4/D6321048419.pdf