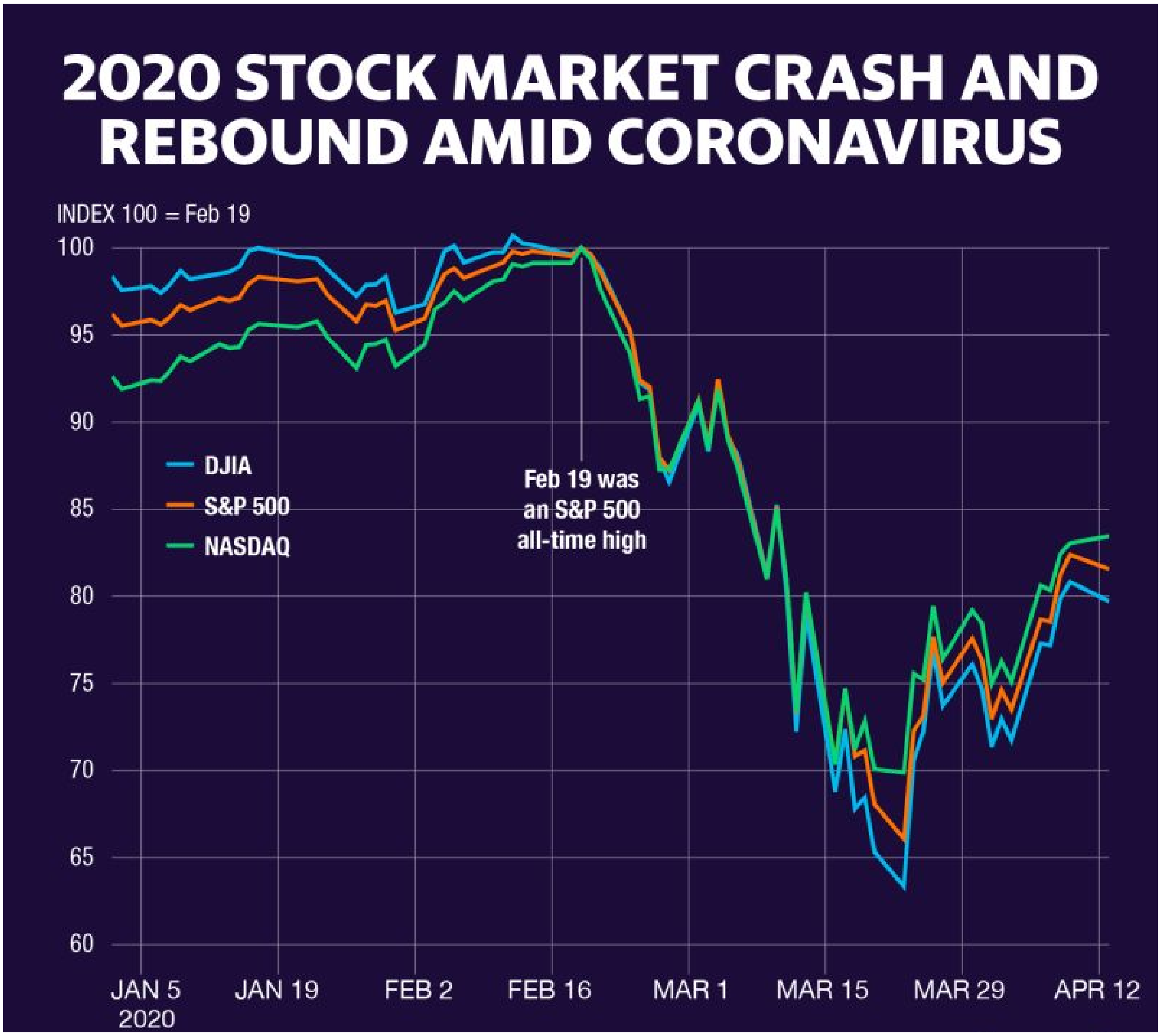
Machine Learning Project Report

May 07, 2020



TEAM MEMBERS:

Anubhav

Former Software Engineer - Accenture with 2 years of work experience. Currently pursuing MS in Computer Science.

Bibek

Graduate from the Indian Statistical Institute and the University of Warwick. Currently pursuing Ph.D. in Mathematics.

Chiranjeevi

Former Software Engineer - Tata Consultancy Service with 3 years of work experience. Currently pursuing MS in Computer Science.

Ishrak

1 year of teaching experience as a full time Lecturer in the department of Computer Science and Engineering at BRAC University. Currently pursuing MS in Computer Science.

Madhu

MS student in computer science.

INDEX

STORYLINE

ANALYSIS OF MACHINE LEARNING MODEL’S PREDICTION OF 2020 STOCK MARKET CRASH AND REBOUND AMID CORONAVIRUS PANDEMIC: Stock market prices are highly unpredictable and volatile. This means that there are no consistent patterns in the data that allow you to model stock prices over time near-perfectly. Princeton University economist Burton Malkiel, who argues in his 1973 book, "A Random Walk Down Wall Street," that if the market is truly efficient and a share price reflects all factors immediately as soon as they're made public, a blindfolded monkey throwing darts at a newspaper stock listing should do as well as any investment professional.

Stock Market prediction using Machine Learning is not really meant for knowing the share price, rather it is used to get an insight about the trends from the data. These insights are based on how well a machine learning model predicts the data based on the previously seen data and trends upon which it has been trained. However, the prediction’s accuracy depends on the factors which have been already witnessed by the model. Coronavirus outbreak called for sudden lockdowns and travel bans across the major economies of the world (by GDP). These were factors which were never seen before and no set of training data could have ever contained these sudden fluctuations. Therefore, we anticipated that unlike the experiments conducted before 10th of March 2020, the experiments conducted post declaration of emergency, travel ban, layoffs and mass deaths would vary in accuracy of prediction of stock prices.

How long can the “Stay-at-Home” or “Shelter-in-Place” or “Lockdown” continue? One day, slowly, the economy must reopen. But how should they reopen? Where should they reopen and most importantly based on what factors should the government reopen the economy?

It is a fact there is no vaccine currently for coronavirus and no one knows about the pricing and logistics of the vaccines which may come in the market in near future.

Therefore, we have presented our work on the SIR statistical model where we have studied the spread of a contagious disease per 100,000 population. This gives a relationship between the three factors: “*Susceptible population*” - “*Infected population*” - “*rate of Recovery with immunity*” and two parameters: “*effective contract rate of the disease*” and “*average time period in which an infected person can infect others*.”

ECLECTIC PROJECT PATH

Multi-Dimensional Work to Zoom In

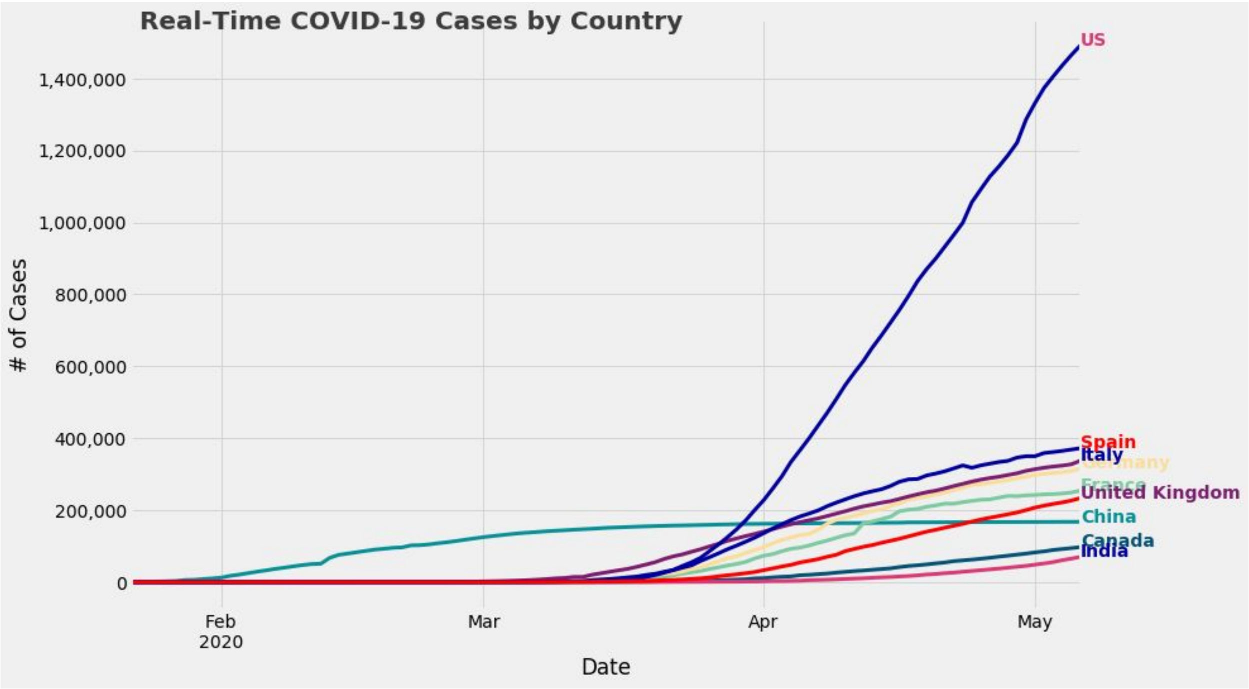
The approach towards the completion of this project has been eclectic. To give you some information, here are a few pointers:

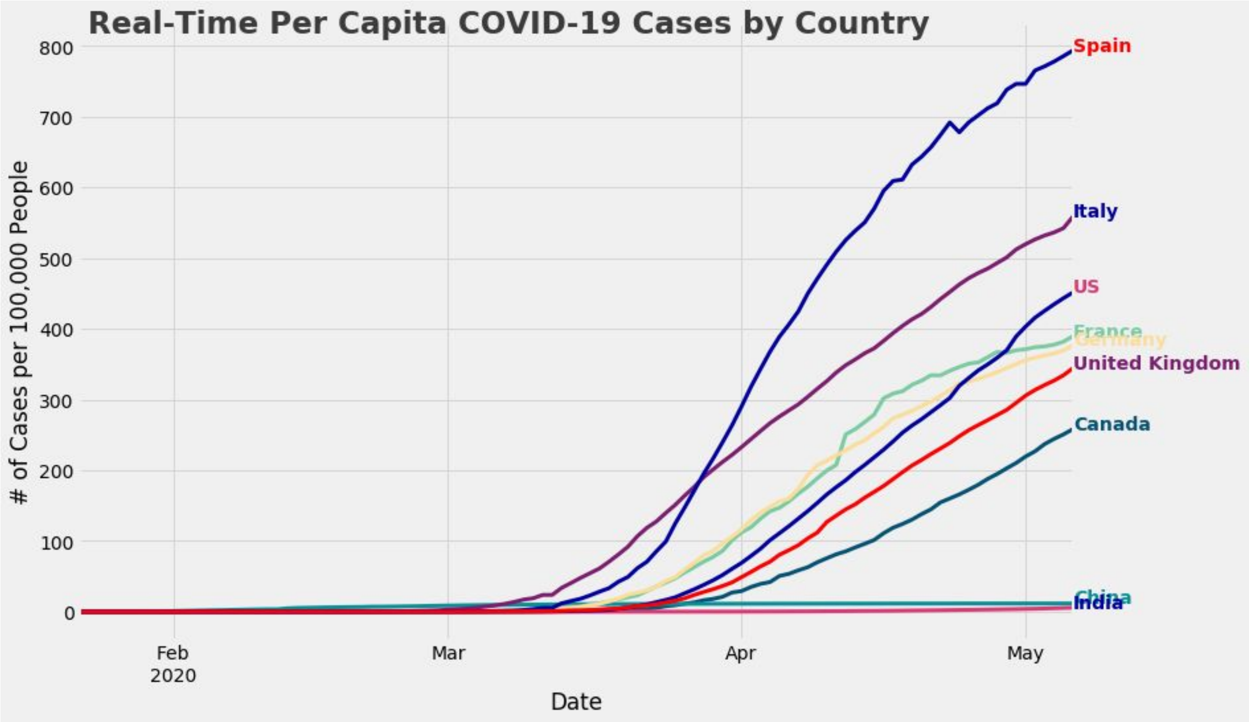
- Right from using various approaches towards getting real time data like Pandas Datareader for remote access of data to using Quandl API for financial data.
- Analysing the pandemic and its level of devastation based on the number of cases per 100,000 of the population.
- Applying Machine Learning Models across multiple stock markets like Nasdaq (USA), London Stock Exchange (UK), Bombay Stock Exchange (India)
- Applying machine learning models across various industries like the Airlines industry which has been badly hit during the pandemic
- Applying Machine Learning models on the stock price of one of the largest conglomerates in the world
- Analysing work using statistical knowledge.
- Extensive work on SIR statistical models which predicts the risk, spread and recovery (by immunity) per 100,000 population taking into consideration relationship between the three factors: "*Susceptible population*" - "*Infected population*" - "*rate of Recovery with immunity*" and two parameters: "*effective contract rate of the disease*" and "*average time period in which an infected person can infect others*."

CHANGES IN PROJECT PLAN: Earlier we proposed working on the stock market and the real estate market. However, instead of working the real estate, we have focused on the Airline industry which has been one of the worst hit industries during the coronavirus pandemic and also focused on the SIR mathematical model which shows the rate of recovery by immunity and rate of spread per 100,000 population.

Experiment 1

Coronavirus pandemic has left more than 100 countries devastated. The effect on the economy has been severe too. Although the numbers in the United States are baffling, it is not the United States which ranks top when it comes to the intensity of the effect. We analysed data from several countries based on the total number of cases as well as total number of cases per 100,000.





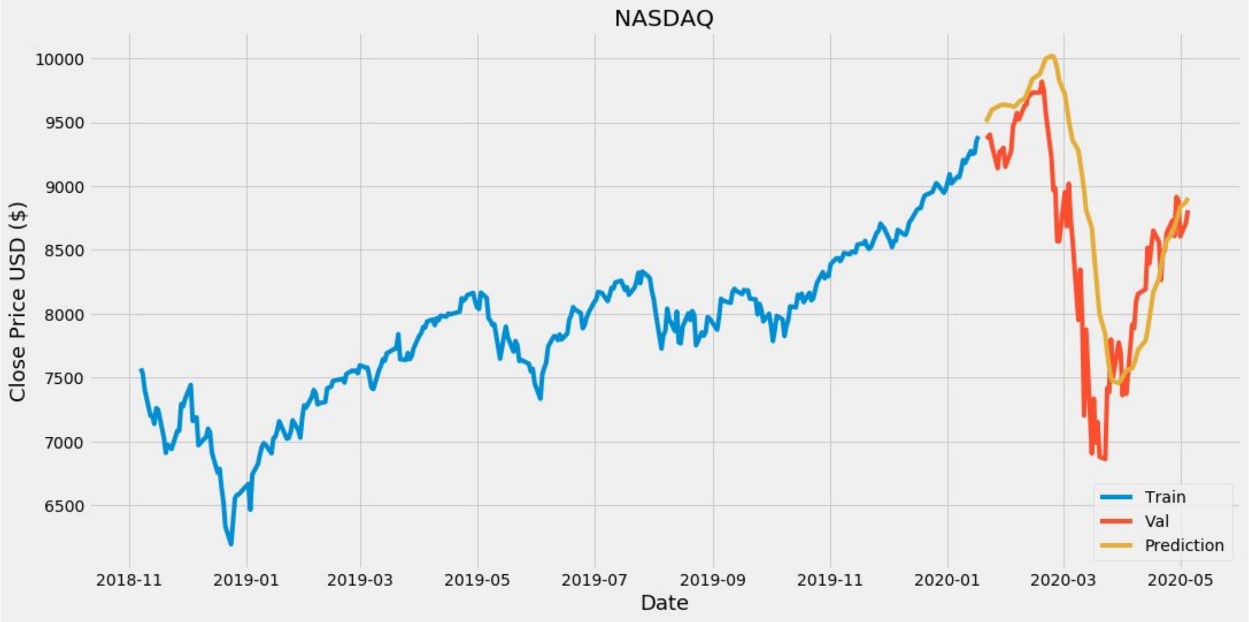
(Above results are from Experiment 1- Part 1. Link to the code is here: https://drive.google.com/file/d/1V3w1637zeX5ndWgrz9Z2oU-bgoifTHDz/view?usp=sharing)

Based on above results and our assumption that the Coronavirus outbreak which called for sudden lockdowns and travel bans across the major economies of the world (by GDP) introduced factors which were never seen before and no set of training data could have ever contained these sudden fluctuations. Therefore, we anticipated that unlike the experiments conducted before 10th of March 2020, the experiments conducted post declaration of emergency, travel ban, layoffs and mass deaths would vary in accuracy of prediction of stock prices. We chose a few countries (USA, UK and UK) and tried to test the accuracy of our Machine Learning model’s prediction of stock price.

Failure in prediction = More than 10% Error in prediction.

NASDAQ (FIGURES IN USD)

DATE	03/01/2020	03/14/2020	03/30/2020	04/10/2020
PREDICTED STOCK PRICE	9717	8157	7742	7789
ACTUAL STOCK PRICE	9628	6989	7774	8153
ERROR%	0.92%	16.71%	0.41%	4.68%



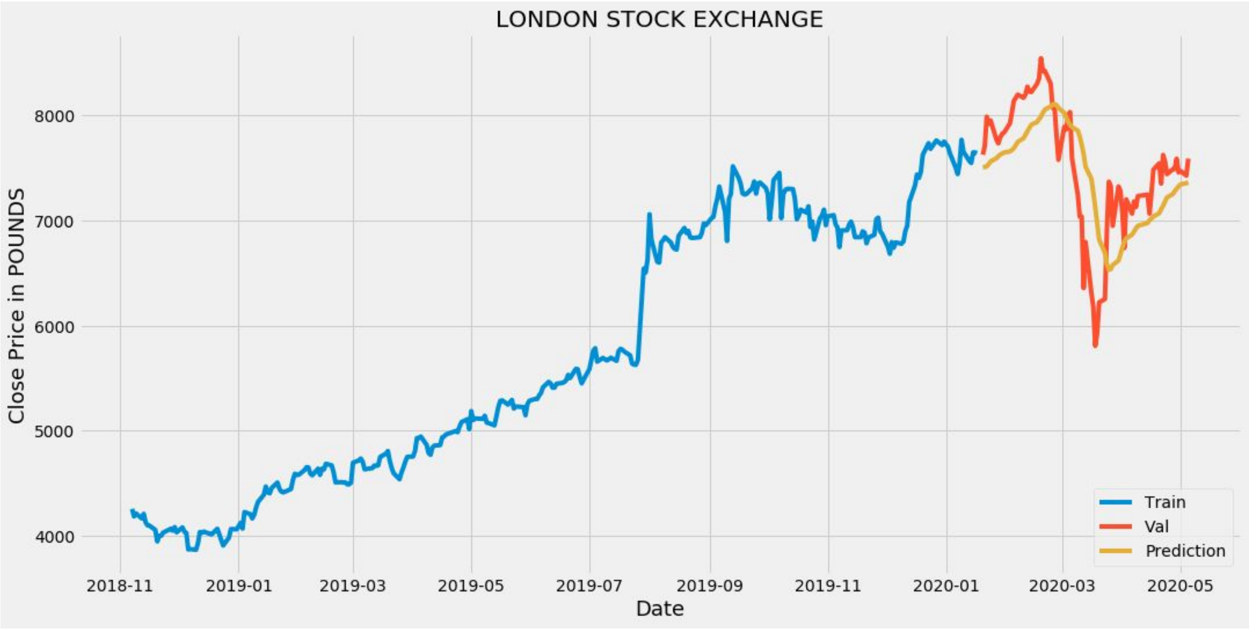
(LINK TO EXPERIMENT FILE:
https://drive.google.com/file/d/1jtt1YjXVPPwaopV-tMEjE77Kf9Lj_08Q/view?usp=sharing)

← What is significant about the date 03/14/2020 when the error rate is 16.71%?

On this date, the president of the United States declared an emergency and stay-at-home order. The machine learning models cannot anticipate such sudden changes and hence the prediction fails.

LONDON STOCK EXCHANGE (FIGURES IN POUNDS)

DATE	03/01/2020	03/14/2020	03/30/2020	04/10/2020
PREDICTED STOCK PRICE	7908	7394	6677	6975
ACTUAL STOCK PRICE	7864	6356	7322	7232
ERROR%	0.55%	16.33%	8.80%	3.55%



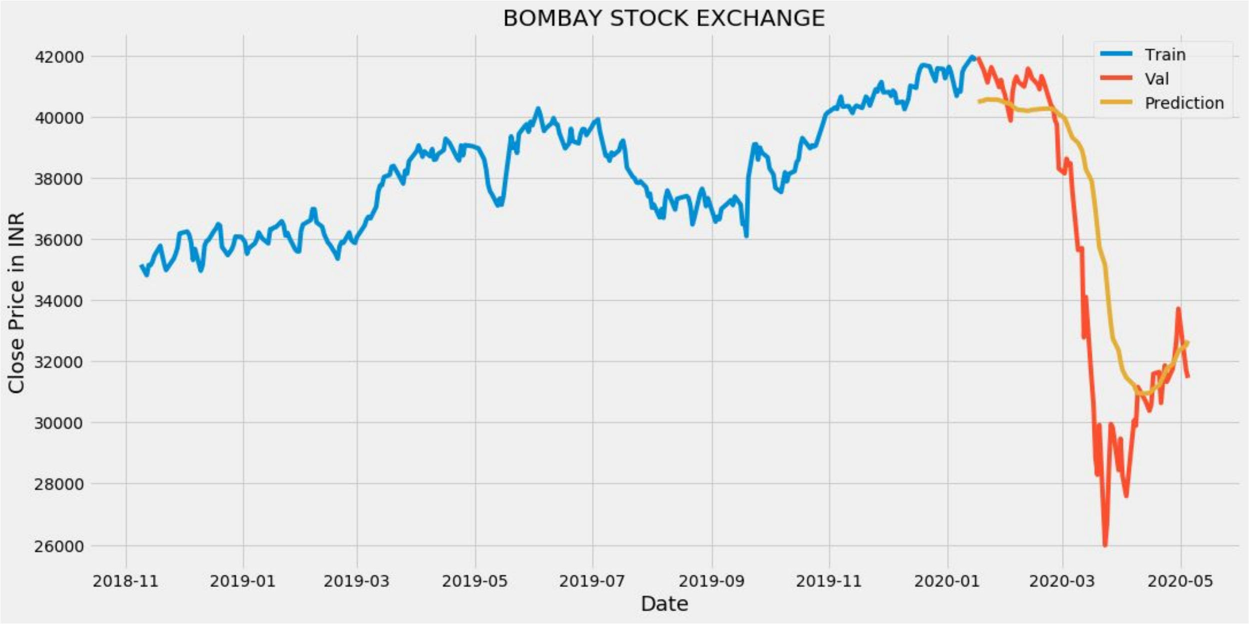
← What is significant about the date 03/14/2020 when the error rate is 16.33%?

On this date, the United States of America declared a travel ban from the UK which houses the European headquarters of several banking and financial institutions.

(LINK TO FILE: https://drive.google.com/file/d/1xXJ2dRz_myI8r35aGqX93l-1gvI9YvAM/view?usp=sharing)

BOMBAY STOCK EXCHANGE (FIGURES IN INR)

DATE	03/01/2020	03/14/2020	03/23/2020	04/10/2020
PREDICTED STOCK PRICE	39472	37904	34479	30939
ACTUAL STOCK PRICE	38409	34103	25981	31160
ERROR%	2.76%	11.14%	32.70%	0.07%



<-- Two Significant Dates: What went wrong 14th of March and on 23rd of March? Note that 23rd of March shows an error of 32.70%.

14th March showed the impact of global travel bans and lockdowns. However, on 22nd of March 2020, India declared the world's largest lockdown. The state and central governments together called for a blanket ban on flights, trains, buses, in fact on any kind of movement outside the home. This caused a massive crash in the stock market and as usual, no machine learning model could have predicted such a fall. LINK FOR BSE: https://drive.google.com/file/d/1dng24GWku-e-4xvLVgtiZ6CxMltfMM1m/view?usp=sharing

Technical Details of Experiment 1

LSTM

LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not.

LSTM is a special kind of 'recurrent neural network(RNN).' Note that in RNN, previous information is being connected with the present task compared to a traditional neural network in which every time the task is started from scratch. In LSTM, this is taken to another level in which the algorithm is capable of learning long-term dependencies. LSTMs have a chain-like structure in which each repeating module will have four layers (compared to one simple layer in standard RNNs). We will describe below in a bit more detail how this works.

Common Architecture of LSTM:

- Cell state: This is like a conveyor belt which runs entire information in almost linear fashion. The information does not change in this state.
- Forget Gate: This is a sigmoid layer and this decides which information from the cell state should be thrown. A value of 1 means all information passes through and value of 0 means nothing goes through.
- Input Gate: This decides which new information we need to store in the cell-state. This also has a tanh layer which creates new candidate values that could be added to the state. These two informations are finally combined to create an update to the state.
- Output Gate: This decides which state will finally come out. This depends upon a filtered version of the cell state in which the cell state will be put through a tanh gate so that the values of output will lie between -1 and 1.

Each of the above steps mentioned in file is performed through a mathematical function in which the inputs and outputs are vectors depending upon the problem in hand.

Resources and credits

- https://in.finance.yahoo.com/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAL7E6_r6OazDyMaYH9n6e32-AUdgNR5VSdIr1fw1AJ4iNy_IhLx27u8Ztxzn8ltiShuAfghxFlKsHJDOFssDZeKue921oqEbS-z-dBghsNVDHShUgBe86rzzmpTgrm_92bP3HuT2LxVqXVspK_lxWhTf1kmVs2KNPGo7e-nxoxbJ
- https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/
- https://stackabuse.com/time-series-analysis-with-lstm-using-pythons-keras-library/
- https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/
- https://medium.com/@dclengacher/keras-lstm-recurrent-neural-networks-c1f5febde03d

Pre-Processing Transformations

In practice it is nearly always advantageous to apply the following to the input data before it is presented to a network.

- Pre-processing transformations
- Scaling
- Normalization

3D Expectations of LSTM

A LSTM network expects the input to be 3-Dimensional in the form [samples, time steps, features]:

- Samples is the number of data points (or rows/ records) we have,
- Time steps is the number of time-dependent steps that are there in a single data point (60),
- Features/indicators refers to the number of variables we have for the corresponding true value in Y, since we are only using one feature 'Close',
- The number of features/indicators will be one -- Reshape the data into the shape accepted by the LSTM

The root mean squared error (RMSE):RMSE is a commonly used scale to measure the accuracy of predictions from various models.

- In simple words, it is the average of Euclidean distance between the original vectors (or data points in n-dimensional space) and the prediction vectors.
- RMSE is a good measure of how accurately the model predicts the response.
- The less the value the better the model is in terms of prediction.
- RMSE is the standard deviation of the residuals (prediction errors).

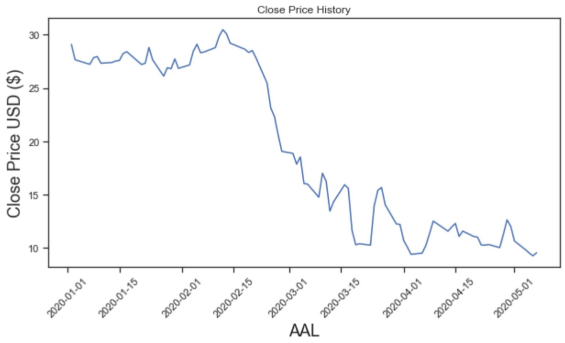
Some of the other commonly used norms are maximum norm, absolute value norm and L_p norms for various values of p each having its own advantage and utility. However, for this project for simplicity we will go with RMSE.

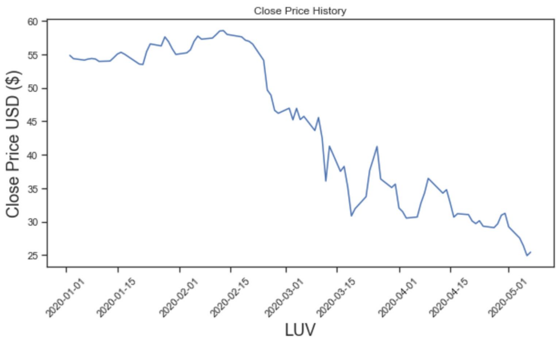
EXPERIMENT 2

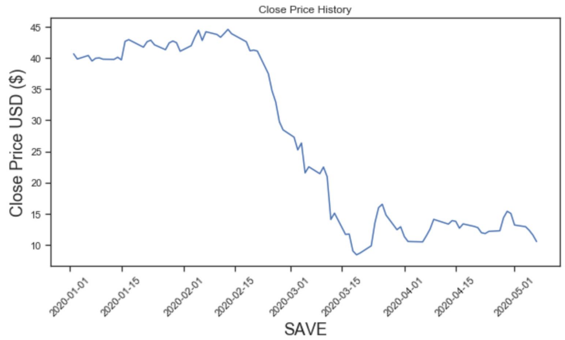
Experiment 2 has two facets, first which has focused on exploratory analysis of the devastating impact of Covid-19 on the stock prices of various US based airlines. We have chosen US based airlines as these have seen the impact which is based on reduced consumer demand.

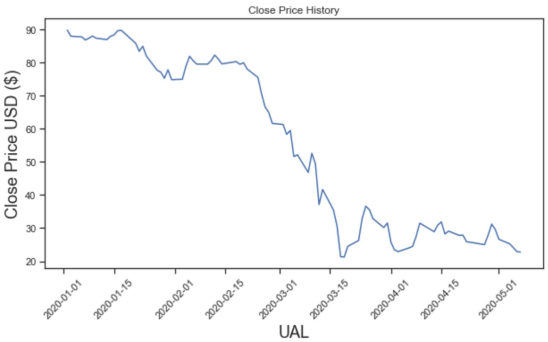
The motivation for the initial signal similarity analysis was to find a group of stocks having similar impact due to the COVID-19 crisis. Eventually, we wish to find out if the stocks belonging to a similar group demonstrate a similar recovery trajectory after the COVID-19 crisis. For projecting stock prices, we have experimented with a Generative Adversarial Network. We can then cross refer to the outcome of the initial similarity analysis and the eventual GAN projection to find out the degree of similarity among the forecast of those similar stocks.

Phase 1: Similarity Analysis of US Airlines stocks using DTW









As the DTW scores suggest and from visual inspection, we can see that the airlines industry was homogeneously impacted by the COVID-19 crisis.

Phase 2: Generative Adversarial Network(GAN) for forecasting

We have designed our GAN model with two sub-networks that are a generator and a discriminator. The discriminator network comprises a Convolutional layer with Batch Normalization. LeakyReLU activation has been used since it handles dead neurons better. The output layer is a single neuron that will output 1 for true samples and 0 for fake samples. The model is compiled with ‘binary_crossentropy’ loss function and ‘adam’ optimizer. The discriminator will be pre-trained to discriminate between authentic and synthetic samples.

The generator will generate fake samples a few timesteps into the future. These future timesteps will define the forecast period. Gated Recurrent Unit (GRU) is used as the generator. LSTM (Long term short term memory) is another option. However, since both GRU and LSTM give comparable performance and LSTM takes longer time to train, we have explored the usage of GRU as a sequence generator. The number of cells of the output GRU unit will be equal to the forecast period.

After pretraining the discriminator model, its training parameters are frozen. We create a third logical GAN model by vertically stacking up the generator and the discriminator. The idea is for the generator to produce forecast sequences and if the discriminator identifies the produced sequence as fake, only the generator updates its weights to produce better and more credible sequences the next time. Since we have frozen the discriminator models’ training, weight updates are propagated directly to the generator.

So, with enough epochs the generator should be able to generate credible future sequences. For example, from the existing training data, if we have a sequence of 0 to 60 days, we can use that sequence as input to the GRU generator which will generate a sequence of next 60 days. We can use the forecasted sequence with the original sequence from 1 to 61 days in the discriminator to train the generator to produce more credible sequences. We can continue the GAN iterations for multiple 60 days sequences and reach an iteration of the generator that will be robust to generate a credible forecast for the next 60 days from the current day.

Links to experiments:

Phase1:https://colab.research.google.com/drive/137sKRcUN_kfJE7wODXiZHZBAi0oVEOhB?usp=sharing
Phase2:https://colab.research.google.com/drive/1KVDCHHakvs96TfJ9ra1w2k8VegZvDoK4?usp=sharing

Resources and Credits:

a. https://machinelearningmastery.com/how-to-develop-a-generative-adversarial-network-for-a-1-dimensional-function-from-scratch-in-keras/
b. https://towardsdatascience.com/aifortrading-2edd6fac689d

Experiment 3

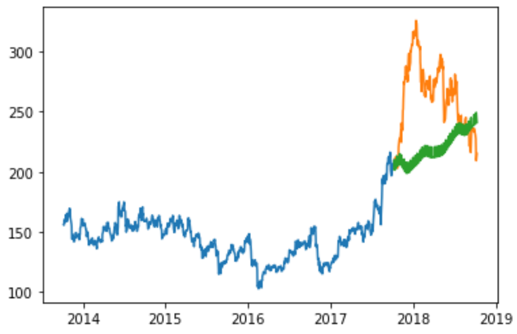
Stocks of one of the largest conglomerates in the world, TATA, and one of the largest companies, Apple Inc. have been analysed in these experiments.

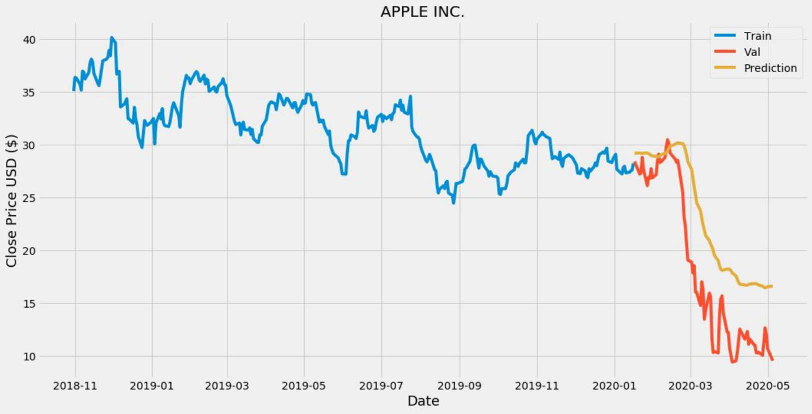
Quandl Data Acquisition Code:
https://drive.google.com/file/d/1YhR_sZdxF_7-I7ZED06xzF291rfKfzKa/view?usp=sharing

Experiment Code:
https://drive.google.com/file/d/1DKTf5T2KolQqZccZGF1oC7DAcRq62KdM/view?usp=sharing

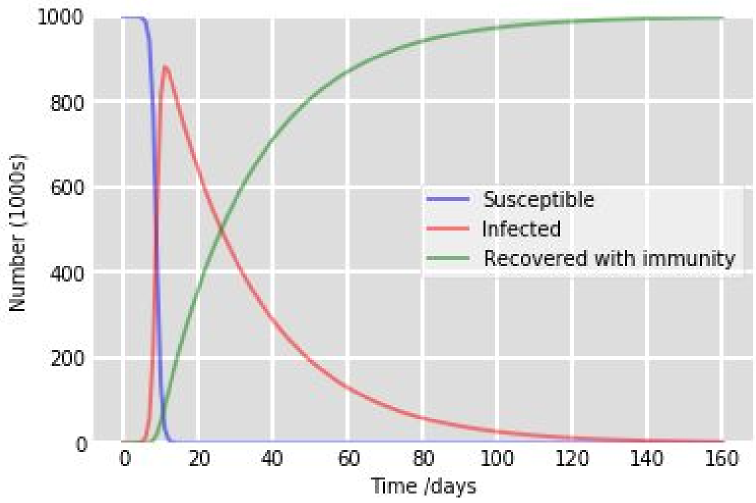
There are multiple variables in the dataset – date, open, high, low, last, close, total_trade_quantity, and turnover.The columns Open and Close represent the starting and final price at which the stock is traded on a particular day. High, Low and Last represent the maximum, minimum, and last price of the share for the day. Total Trade Quantity is the number of shares bought or sold in the day and Turnover (Lacs) is the turnover of the particular company on a given date.

To prepare this model and related RMSE, we have used Quandl API BeatifulSoap parsing, Linear Regression, KNN, ARIMA, LSTM and FB Prophet.





STATISTICAL MODEL FOR RECOVERY AND REOPENING OF THE ECONOMY



How long can the “Stay-at-Home” or “Shelter-in-Place” or “Lockdown” continue? One day, slowly, the economy must reopen. But how should they reopen? Where should they reopen and most importantly based on what factors should the government reopen the economy?

It is a fact there is no vaccine currently for coronavirus and no one knows about the pricing and logistics of the vaccines which may come in the market in near future.

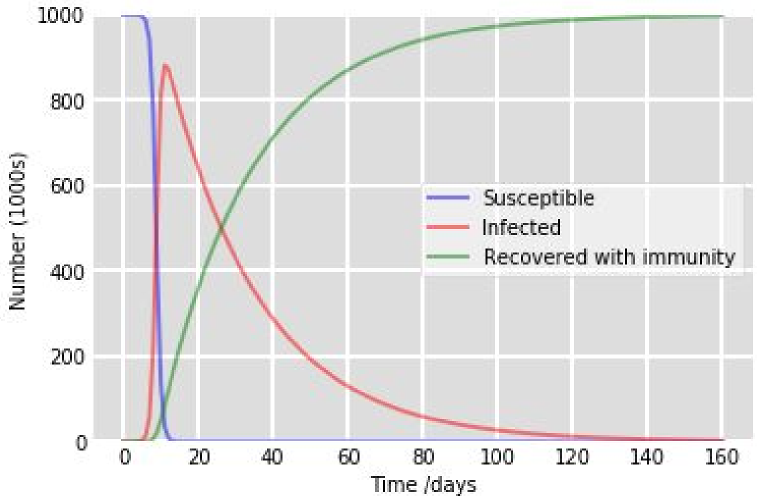
Therefore, we have presented our work on the SIR statistical model where we have studied the spread of a contagious disease per 100,000 population. This gives a relationship between the three factors: "*Susceptible population*" - "*Infected population*" - "*rate of Recovery with immunity*" and two parameters: "*effective contract rate of the disease*" and "*average time period in which an infected person can infect others*."

SIR model for epidemics:

The SIR model is a simple mathematical model to study the progress of an epidemic in a population. This is one of the standard compartmental models used for studying infectious disease i.e. the whole population is divided into a number of compartments. For example, in the SIR model, the whole population is divided into three parts: S stands for the number of 'Susceptible' persons, I stands for 'Infected' and R stands for 'Recovered'(including the deceased persons).

The model also involves two constants which are usually computed from various other experiments and analysis. The first one is denoted as beta which stands for effective contact rate of the disease. That means if we denote the total population in a city (or country) as N, then beta.N is the number of people in that population who will come in contact with a single infected person. The second constant is denoted as gaama which stands for the recovery rate. In simple words 1/gamma gives the average time period in which an infected person can infect others.

We are applying this model below to a hypothetical city having population (N= 1,000,000), beta= 1.6 and gamma value= 0.04. (Values are based on alpha and beta values of generic contagious diseases)



This mathematical model will tell us where the risk of spread is low and where the recovery rate is high. Those are the places where economies should be reopened, cautiously.

LINK FOR CODE:
https://drive.google.com/file/d/1UNMocF6NXj-IzbJZ2qBbQk3DZvMVbJVh/view?usp=sharing

LINKS TO CODES:

EXPERIMENT 1:

- https://drive.google.com/file/d/1V3w1637zeX5ndWgrz9Z2oU-bgoifTHDz/view?usp=sharing
- https://drive.google.com/file/d/1jtt1YjXVPPwaopV-tMEjE77Kf9Lj_08Q/view?usp=sharing
- https://drive.google.com/file/d/1xXJ2dRz_myI8r35aGqX93l-1gvI9YvAM/view?usp=sharing

- https://drive.google.com/file/d/1dng24GWku-e-4xvLVgtiZ6CxMltfMM1m/view?usp=sharing

EXPERIMENT 2:

- https://colab.research.google.com/drive/137sKRcUN_kfJE7wODXiZHZBAi0oVEOhB?usp=sharing
- https://colab.research.google.com/drive/1KVDCHHakvs96TfJ9ra1w2k8VegZvDoK4?usp=sharing

EXPERIMENT 3:

- https://drive.google.com/file/d/1DKTf5T2KolQqZccZGF1oC7DAcRq62KdM/view?usp=sharing
- https://drive.google.com/file/d/1YhR_sZdxF_7-I7ZED06xzF291rfKfzKa/view?usp=sharing

SIR MATHEMATICAL MODEL FOR RECOVERY

- https://drive.google.com/file/d/1UNMocF6NXj-IzbJZ2qBbQk3DZvMVbJVh/view?usp=sharing