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

Chirag Sharma
Dept. of Electrical & Computer Engineering
Stevens Institute of Technology
csharma3@stevens.edu

Prem Patel
Dept. of Information Systems
Stevens Institute of Technology
ppate101@stevens.edu

Shreya Bhargava
Dept. of Information Systems
Stevens Institute of Technology
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, support vector machine, decision tree**

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 treatments.[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[2] project global growth to fall by 3%. This is a downgrade of 6.3 percentage points from January 2020, a major revision over a very short period. 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

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. 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. 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. This is how the major stock indexes have been reacting to the major outbreak.[3]

Fig 1. Dow Jones values in 2020

Fig 2: S&P values in 2020

II. EXISTING SOLUTIONS

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 thereby plotting and deriving the relationship 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 grow as compared to other countries.

Stock market prediction has been existing since long however it has either predicted the future value of one specific company or predicted the value of indexes. As per our literature survey is concerned particle swarm optimization (PSO) has been widely used to predict the stocks of one specific company. 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.

Another basic technique used to predict stocks are k-nearest neighbours and linear regression. The disadvantage of using them is that it gives a high RMSE value which usually determines that the model is poor and overfits the data. It is safe to say that linear regression models do not perform well for stock market prediction.[4]

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. SOLUTION

A. *DESCRIPTION OF THE DATASET*

To implement and get results for our desired project we have gathered three different datasets. The COVID-19 dataset has data which includes date, confirmed, recovered and number of deaths all across the world. Since we are only interested in analyzing data for the US, all other country data isn't used for analysis.

In order to get the stock market stats, we willing be retrieving data from yahoo-finance. We use this to get historical data for fortune 500 companies. This is linked to another data module S&P 500 companies with financial information. The S&P is a free float, capitalization-weighted index of the top 500 publicly listed stocks in the US. It is necessary to link the two so as to obtain the sector of each stock.

B. *STEPS FOR PROCESSING AND CLEANING*

1. COVID-19 data has been collected through the URL Request method to download the csv and dump it into a *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.

2. These are three different datasets for *Confirmed*, *Recovered*, and *Deaths* with *Province/State*, *Country/Region*, *Last Update* as common columns in all three.

3. We remove unnecessary columns and only keep *Date*, *Confirmed*, *Recovered*, and *Death* merging all three columns into one *dataframe* using *pandas.merge*.

4. 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 a list.

5. For stock market data, we are using *yfinance* 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. Some companies' data is not found as they might be delisted.

C. LIBRARIES

We are importing Pandas and NumPy libraries for data management. Pandas provides easy to use structures and data analysis tools, in-memory 2D table object called dataframe which resembles a spreadsheet with columns and row labels. NumPy library provides objects for multi-dimensional arrays and computing functionalities that are designed for high-level mathematical functions and scientific computation. We have also imported the Scikit Learn Metrics which implements functions assessing prediction errors for specific purposes such as regression metrics. In future we will use this to calculate MSE and regression plots between various countries and the training dataset which will be a part of our prediction analysis. We have used polynomial regression fit to avoid underfitting or overfitting of the data, because it would increase the complexity of the model and make it fit better with lesser noise.

D. MACHINE LEARNING ALGORITHMS

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.

One of the most common techniques for regression analysis is Linear Regression, where we try to find a line (or a more complex linear combination) 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 commonly 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.

In conjunction to Linear Regression we will also use Support Vector Machine (SVM) and more 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.

E. PRELIMINARY RESULTS

Most of the reports that we currently receive only focuses on the exponential growth and only few focuses on when is it going to end. We focused on trying to know where we are headed with the coronavirus COVID – 19 disease trends and are we making any detectable progress, because, we can not have exponential growth forever, which is the current case. At some point the virus will run out of people to infect either because most people have already been infected or we get the situation under control following stringent steps. However, when we are in the middle of the in the process of exponential growth it is very difficult to determine where we are going to end. Exactly knowing where the exponential growth end is hugely important because it lets us know how many people are going to get infected.

We plot a graph (Total Confirmed Cases vs. New Confirmed Cases in the Past Week) visualizing the epidemic. We compare the results for two countries in particular – China, where the situation is almost under control and not many new cases are being reported, and the second being United States of America, where huge numbers of new cases are being reported daily. This graph uses the real data and shows the countries travel along the axes of exponential growth and it makes it clear if the exponential spread of disease has stopped because it plummets downwards from the axes. The key ideas behind this graph being plotting against a Logarithmic Scale (which is the natural scale for exponential growth) where the scale the scale ticks in the multiples of 10 rather than ticking in increments of 10 in a Linear Scale, doing this scales up small numbers and scales down large numbers making the growth equally apparent on all scales. The other key idea is not plotting against time because the spread of the disease is not dependent upon the time but on the number of people already infected and the number of people that will be infected currently i.e.

the total number of infections and the growth rate of infections. Therefore, the number of new cases is proportional to the number of existing cases, which helps us plot the exponential growth in a straight line. In most of the countries the number cases increase at almost the same rate until at a point it stops and falls from its trajectory.

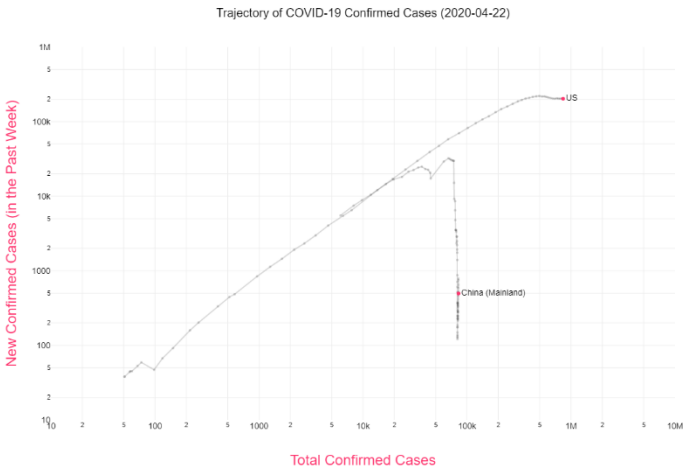

Fig 3: Total Confirmed Cases vs New Confirmed Cases on Logarithmic Scale

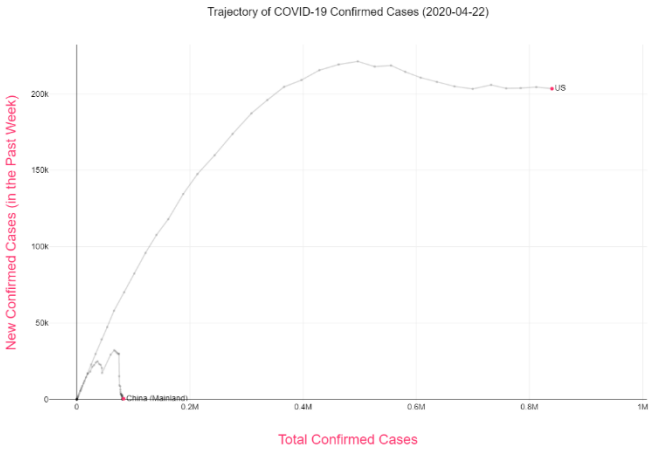

Fig 4: Total Confirmed Cases vs New Confirmed Cases on Linear Scale

IV. FUTURE DIRECTIONS

We would be implementing Long Short-Term Memory, Random Forest algorithm along with support vector machine and decision tress. Comparing the results from various algorithms using a confusion matrix we should be able to predict stocks and how it has been impacted by COVID-19.

V. REFERENCES

[1] https://www.who.int/health-topics/coronavirus#tab=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://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/