Impact of Covid-19 on Economy

Authors:

Sineha Aneel (17k-3897)

Avinash (17k-3918)

Shayan Shahid (17k-3851)

SETTING THE RESEARCH GOAL

The novel COVID-19 virus has left the world reeling with lacs of deaths and millions of cases. The entire human life has come to a standstill with the majority of the countries enforcing shutdown/lockdown and travel bans. Due to the pandemic and lockdown organizations have lost their revenue, people have lost their jobs and the stock market is at its lowest. This has highly affected the economy worldwide which will eventually lead to the decline of GDP in the majority of the countries. Some of the major setbacks were the 2020 stock market crash on 20th Februarys [3], followed by IMF stating on 14th April that all the G7 nations are facing recession[2]. Furthermore, the IMF suggests that the impact of COVID-19 will be far worse than it was in 2009 due to the "Great Recession". The United Nations (UN) predicted in April 2020 that global unemployment COVID wipes out 6.7 percent of working hours globally in the second quarter of 2020—equivalent to 195 million full-time workers [1]. During the crash, global stock markets made unprecedented and volatile swings, mainly due to extreme uncertainty in the markets. This crisis has brought a common question in everyone's mind about to what extent will COVID-19 destroy the economy. Our project will demonstrate the impact of COVID-19 on the economy along with predictions using various data science and machine learning algorithms.

RETRIEVING DATA

The data was collected after extensive research on the internet. It has been collected from various different websites. The dataset mainly contains inflation details, share prices, and details about COVID-19. We shortlisted 5 countries that are the United States of America, the United Kingdom, Spain, Italy, and Germany. The reason to choose these states was that data regarding their economy was easily available on the internet mainly because these countries are a part of OCED. We tried to find data on Pakistan's economy but were unable to find data related to our project. The inflation dataset was collected from OCED's official website [4] since our selected countries are a part of OECD. Inflation is a macroeconomic measure of the change in prices of goods and services, and as such provides a particular measure of growth and macro-economic activity in a country. The dataset is with reference to the Consumer Price Index (CPI) along with the time it was collected. A similar dataset for inflation was also available on datahub.io and on the World Bank’s official website however the most recent data it had was of 2019. The next dataset used was regarding share prices [5], similar to the inflation dataset this one was also collected by OCED with respect to share prices. A share price is the price of a single share of a number of saleable stocks of a company, derivative or another financial asset. The location index of these datasets refers to the countries by their short forms mostly use i.e. Spain is indexed as ESP. Various other datasets were available on OCED's website regarding the economy as housing prices and Producer Price Indices (PPI) however that was not of much use regarding our research problem. Another dataset used is named gdp_growth_rate for the above mentioned 5 countries. This dataset contains the GDP growth rate starting from 1980, in case the growth rate is negative a minus sign has been attached to the value. Furthermore, we have COVID-19's dataset for all the countries which have been cleaned to fit our use. Some of the columns that are used are total cases, total deaths, total recoveries, location, total cases per million, total deaths per million etc. whereas the remaining fields are cleaned thoroughly.

DATA PREPARATION

This is from where the actual implementation of the project started. Initially, the datasets were loaded and the combined data of COVID-19 was used. The dataset had a lot of columns that were not related to our project hence we had to drop them. The rows containing null values were also dropped along with duplicate rows. Furthermore, since our research was mainly based on selected 5 countries hence data of the remaining countries regarding COVID-19 was ignored. No specific indexing was done on the dataset because it was not useful hence basic indexing was used denoted by row numbers. A new column of the day was added to our dataset to track the number of days since the outbreak. Then the dataset which contained COVID-19 data of all countries affected by the dataset only chose 5 specified countries and loaded to another CSV file. The new CSV file with only 5 countries' data was further used in the entire code. The data were then grouped and renamed according to need. This was mainly done on the newly generated CSV file containing shortlisted countries COVID-19 data along with share price, CSV and inflation.csv. Some columns were merged while some were dropped then by using the function of SimpleImputer() in sklearn library the blank values were filled with the mean and the data was then transformed using imputer.fit_trasnform() according to its fit. Countries were also renamed in the fivecountrydata.csv to their short forms the way it was in share price and inflation dataset files i.e. SPAIN became ESP so that it could be properly coherent with the other dataset files in our use for the project.

DATA ANALYSIS

Data Analysis was then done to find hidden insights of the dataset however no undetected errors or outliers were found indicating that data cleaning was properly done. Initially, a basic plot was visualized using matplotlib to see the rate at which the number of countries affected by COVID-19 increased. The visualization started by 31st December 2019 and followed. A comparison between total deaths and total cases was done another comparison of total cases against total recoveries can also be done. Another visualization was done combining the five countries we chose, a basic line plot was drawn which represented the total deaths, total recoveries with respect to the date on the same plot of all five countries to have a clearer insight about the data.

DATA MODELLING

Our next step in the process was to perform data modeling for our project. It started off as training and testing of data. For the x domain, a new case rate and the death rate were considered whereas for our y domain inflation rate was used. The test data was 33% and the train data was 67%, Linear Regression was then used to find the relation between these variables. Initially, the model was fitted using training data and then y prediction was done using testing data of x, the r2 score was then calculated using y's testing and prediction values. Our r2 score was on the negative side which meant linear regression is not working properly with our model hence after searching online and studying in-depth we figured that we could use Lasso Regression for our model to obtain a subset of predictors that minimize prediction error by imposing a constraint in the model which is represented as alpha using the value of 0.3. Training and testing data was used the same way it did for linear regression i.e. the model was fitted using training data and then y prediction was done using testing data of x, the r2 score was then calculated using y's testing and prediction values. The r2 score was a little reduced but was still in the negative domain. Another regression was then tried named as Ridge Regression. It is a technique for analyzing multiple regression data that suffer from multicollinearity. By adding a degree of bias to the regression estimates ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. The ridge regression then uses the alpha value as 0.3 to improve the smoothness moderately since if we increase the alpha constraint majorly then a case similar to overfitting might occur. After consecutive failed attempts we shifted to logistic regression to explain the relationship. The f1 score for inflation was then calculated as 0.75 which was satisfactory.

Once we were done with inflation we shifted our regression techniques to share price data using logistic regression. Since logistic regression has a small set of values present the f1 score was calculated as 0.8 which was a very good score. Initially, the model was fitted using training data and then y prediction was done using testing data of x, the r2 score was then calculated using y's testing and prediction values. Finally, we were done with the regression techniques so we used the confusion matrix to describe the classification of the model for further analysis. The confusion matrix using logistic regression for inflation and share price was then illustrated. Then a precision-recall curve was drawn to summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. The area under the curve was calculated to be 0.85

which was pretty high inferring both high recall and high precision, where high precision relates to a low false-positive rate, and high recall relates to a low false-negative rate. ROC curve was also illustrated summarizing the trade-off between the true positive rate and false-positive rate for a predictive model using different probability thresholds. The process was then followed by using a model called Autoregressive Moving Average (ARMA) for forecasting. Initial research was done for basic AR and MA models but then the combination of their properties impressed us to use the ARMA model. The ARMA model is used to forecast trends by obtaining information from the variables used. It commonly uses the philosophy of *"let the variable speak for itself"*. This model helped us to make an informed decision regarding our analysis and prediction. ARMA model is used for stationary values due to this specific feature we were inclined towards this model even more. By using the test statistic and ARMA model we were able to test our hypothesis along with confidence values. After ARMA we used the upgraded version of AR model that was Variable Autoregressive (VAR) model in which each variable is modeled as a linear combination of past values of itself and the past values of other variables in the system due to having multiple time series so one equation per time series is modeled. The model was fit using 3 lags.

PRESENTATION AND ANNOTATION

The data models of ARMA and VAR modeling were then presented properly. ARMA model was drawn with respect to one significant country followed by its autocorrelation and partial correlation function. The extensive white noise that is the variables are independent and identically distributed with a mean of zero was observed hence seeing almost no correlation the use of ARMA model was not correct. However in the VAR model we were able to forecast the GDP data accordingly. The model was fit by using a lag order of 3 followed by forecasting for 3 periods. Impulse response function (IRF) were also illustrated showing the system's response for various function with different country combinations along with cumulative responses to enhance interpretation was used separately for country combinations too.

REFERENCES

[1]https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_740893/lang--en/index

[2]https://www.bloombergquint.com/business/global-great-lockdown-will-dwarf-the-great-recession

[3]https://www.bbc.com/news/business-51984470

[4]https://data.oecd.org/price/inflation-cpi.htm

[5]https://data.oecd.org/price/share-prices.htm

[6]https://data.oecd.org/gdp/real-gdp-forecast.htm

[7]https://notes.quantecon.org/submission/5db25b54831cf4001af7e506