Duke Datathon 2020

The Economic Impact of COVID-19

Abstract

With the continuous outbreak of COVID-19 around the world, many health experts believe that the pandemic will not end in 2021, and the cease of the pandemic will likely involve a mix of efforts. The main objective of this report is to provide a comprehensive analysis and concrete data-driven recommendations on the devastating long-term economic implications of COVID-19.

To define economic impact, we found three important economic metrics: GDP, unemployment rate, and stock market index. We used principal component analysis to aggregate these three metrics into a single score which we call Economic Develop Index, EDI. We then build a regression model to see which attributes are significant. Finally, we used time-series forecasting to predict how a country would perform in the foreseeable future.

Our model shows increases in *Population Density*, *Death Rate*, *New Deaths Per Million*, *Stringency Index*, and *Hospital Beds Per Thousand* will decrease the overall EDI. The findings match our expectation because of the increase of death rate, stringency index, new death counts, and hospital beds indicate the increase of severity for COVID-19. Moreover, we used the United States to demonstrate how our findings could be used to predict the EDI economic indicator. The same approach could be applied to other countries/cities with appropriately processed data. Finally, we made recommendations to countries facing the devastating long-term economic implications of this pandemic.

Introduction

As of today, there are a total of 46 million global cases of COVID-19 with a total of 1.2 million deaths in the world. Many health experts believe COVID-19 outbreak will not end in 2021, and the cease of pandemic will likely involve a mix of efforts that stopped historic outbreaks: social-control measures, medications and a vaccine [1]. The main objective of this report is to provide a comprehensive analysis of the devastating long-term economic implications of the pandemic and quantifying the association of other factors. In addition, through assessing the results of our Exploratory Data Analysis (EDA), predictive model and time-series forecasting model, we will provide concrete data-driven recommendations helping countries to effectively mitigate the impact of COVID-19 on the economy.

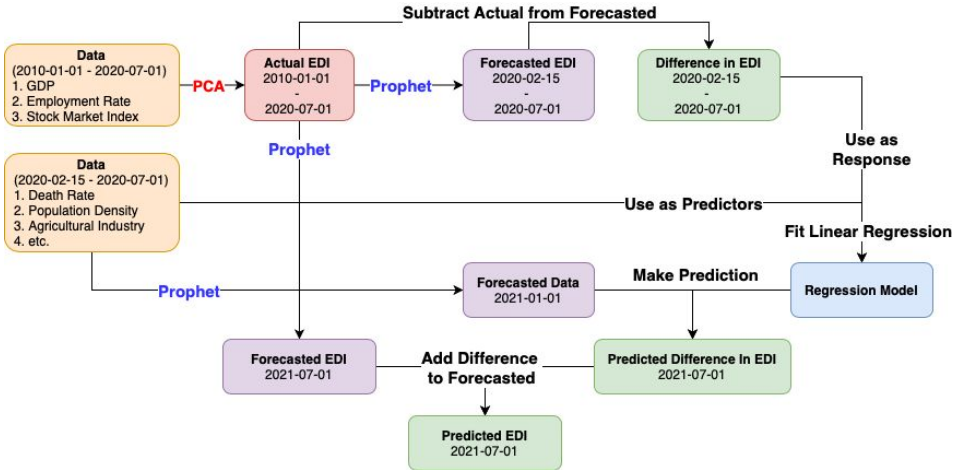
Problem Statements

1) What attributes about a city makes the most significant contribution to impacting its economy?
2) Given these attributes, provide a case study of what the economic indicator will be like 1,2, or 5 years from now?
3) Using the cases available, suggest an approach to scale this problem to the global level.
4) What kind of data-driven solutions will you offer to countries facing devastating long-term economic implications of this pandemic?

Assumptions

Out of all the potential predictors we picked, only COVID-19 predictors like case counts and death counts can be represented at the city-level (please refer to Table 1 in the reference section). This is partly due to the scarcity of city-level data online.The rest of potential predictions related to national economy and national policies can only be represented at the national-level. However, we believe our predictive and forecasting models are capable of accurately quantifying the impact of city-level predictors on the national economy. Our final dataset included data extracted from community mobile reports provided by *Google Movement Data* [2], GDP data extracted from *Kaggle* [3], and COVID-19 related data from *Kaggle* [4,5]. Our pandemic-related potential predictors are selected based on the results of various medical studies conducted by accredited organizations [6].

Methodology



The figure on the left depicts the approach we have taken to address our problem statement. We collected two groups of data for constructing our response and predictors, respectively. However, response data spams from 2010 so that its underlying trend and seasonality could be captured with Prophet to estimate economy behaviour without the disturbance of COVID-19. Our response variables then attempt to isolate the effects of pandemic induced impacts on the economy. The impacts (derived using data from February to July 2020) are then treated as observations and regressed on our predictor variables to provide a model that we could use to quantify these effects as well as to perform forecasting.

Economic metric

Since we want to answer the economic impact of COVID-19, we have to define a metric for tracking economic status. Many indexes could reflect a country's economic condition such as GDP, interest rate, credit rating etc [7]. To capture the most information about a country's economic state, we defined a new economic index that aggregates multiple metrics and we named it: Economic Development Index (EDI). More specifically, we picked GDP (gross domestic income), which is defined as the monetary measure of the market value of all the final goods and services produced in a specific time; Employment rate, which is the percentage of jobless labor force; and stock market index which is a aggregated value of market performance. We used these three metrics because together, they capture the consumer, industrial and service conditions of a country. Moreover, they are coincidental and leading metrics meaning they reflect the current and future state of economics. We found these metrics on the World Bank's global economic monitor website. Finally, to compute EDI, we aggregate the three metrics using principal component analysis. This way our metrics conveys most of the information that the three metrics capture. Below is a plot of EDI over the past 10 years. And it captures the pandemic crisis very well with a drop.

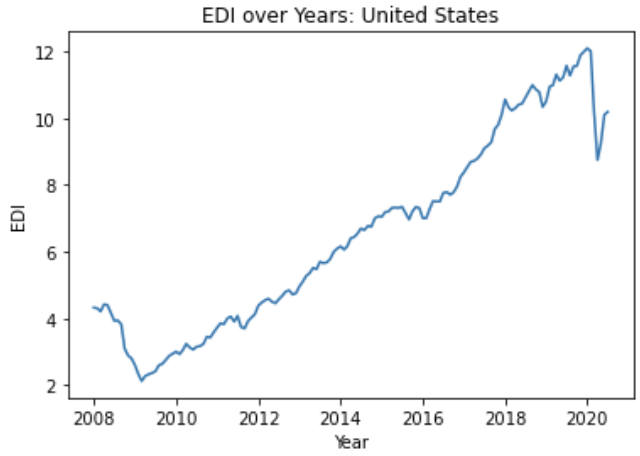


Figure 2. USA's EDI from 2008-2020

(Note: The economic crash in 2008 and the pandemic crisis in 2020 causes significant drop in EDI)

Regression

We built a multiple linear regression model to predict the effect of selected predictors in Table 1 on EDI difference, ΔEDI. The model is trained on the timeframe window from February 15th 2020 to July 1st 2020. We defined EDI difference to be the forecasted EDI assuming the pandemic did not happen $EDI_{no\ pandemic}$ and the actual EDI where the Coronavirus happened $EDI_{pandemic}$.

$$\Delta EDI = EDI_{no\ pandemic} - EDI_{pandemic}$$

In the way we defined ΔEDI, positive ΔEDI means $EDI_{pandemic}$ is lower than $EDI_{no\ pandemic}$, which highlights the negative impact of pandemic on the national economy. To calculate $EDI_{pandemic}$, we simply use PCA to compute EDI from the 3 metrics we chose from the timeframe between February 15th 2020 to July 1st 2020. $EDI_{no\ pandemic}$ is forecasted using Facebook's Prophet library. We used past data from 2010 to 2019 to predict EDIs for the period between February 15th 2020 to July 1st 2020. Since past data is not influenced by the pandemic the forecasting will also not take consideration of the pandemic. We could then take the difference between the two EDIs to simulate the impact of COVID-19 on economics. After building the model, we looked at the significance of our predictors and interpret our findings in the Analysis section

Time series forecasting

To answer which countries will be affected by COVID-19 the most over the next 1,2, and 5 years. We first forecasted the predictors from our regression model for the next 5 years using Prophet library. We then used the forecasted predictors and our regression model to estimate the EDI difference. We reason, the larger the EDI, the

more acute the economic impact is to the country. Therefore, with the projected EDI differences in the next 5 years, we could infer which country is mostly affected.

Analysis

Our study began with EDA with the goal of checking the association of predictor variables and the response variable, and highlight the preliminary concerns we have based on the results of EDA. According to the results of our EDA, all predictors except *Birth Rate* appeared to have association with response variables. In addition, we decided to categorize *Population Density* into three different levels due to the insufficiency of data for some data range. Stepwise selection method was performed to find the optimal model with the lowest BIC score. Before dropping the predictor variables, F-test was implemented to assess the impact of variables on the predictive model. Lastly, we checked to ensure our final model fulfilled the linear regression assumptions (linearity, normality, independence and equal variance). Our final model contained **11 predictors** with all **p-values lower than 0.05**. Our adjusted R-squared is approximately **0.73**, meaning **73%** of the variation can be explained through our model.

Findings

Table 1. Interpretation of Regression model

Variable Name	Variable Type	Interpretation (With all other variables constant)
Population Density	General Predictors	As the population density changes from level 1 (<96) to level 2 (>296) , change of EDI will **increase a multiplicative effect of 55**.
Birth Rate	General Predictors	With 1 unit increase, change of EDI will **decrease by 64%..**
Death Rate	General Predictors	With 1 unit increase, change of EDI will **increase by 50%.**
Retail & Recreation Percent Change From Baseline	Economic Predictors	With 1 unit increase, change of EDI will **decrease by 0.3%.**
Agriculture	Economic Predictors	With 1 unit increase , change of EDI will **decrease by 0.002%.**
Industry	Economic Predictors	With 1 unit increase , change of EDI will be close to 0%. The change is negligible despite its statistical significance.
Service	Economic Predictors	With 1 unit increase , change of EDI will be close to 0%. The change is negligible despite its statistical significance.
New Cases Per Million	COVID19 Predictors	With 1 unit increase, change of EDI will **decrease by 0.003%.**
New Deaths Per Million	COVID19 Predictors	With 1 unit increase , change of EDI will **increase by 1.5%.**
Stringency Index	COVID19 Predictors	With 1 unit increase , change of EDI will **increase by 0.1%.**
Hospital Beds Per Thousand	COVID19 Predictors	With 1 unit increase , change of EDI will **increase by 1.5%.**

(Note: Grey text refers to insignificant predictor, green refers to positive predictor, red refers to negative predictor.)

Table 2. Economic impact of Covid -19 in the United States on July 1st, 2021.

Forecasted EDI	Predicted Difference in EDI	Predicted EDI
12.53	-2.18	10.35

Note that difference in EDI is a vector that we have derived in an attempt to quantify the economic impact while controlling for factors associated with COVID-19. Therefore, In the case of United states, we expect that there will be a negative impact in US economy, assuming that the trends of all predictors we described above will stay consistent until July 1st, 2021

Conclusion

Our results show an increase of *Population Density*, *Death Rate*, *New Deaths Per Million*, *Stringency Index*, and *Hospital Beds Per Thousand* will decrease the overall EDI. The findings match our expectation because the increase of death rate, stringency index, new death counts, and hospital beds indicate the increase of severity for COVID-19. Scientific studies have shown the impact of social distancing on the infection rate of COVID-19; therefore, increase in population density will likely increase the infection rate of COVID-19. In contrast, increases of *Birth Rate*, *Retail & Recreation Percent Change From Baseline*, *Agriculture*, and *New Cases Per Million* will increase the overall EDI. The less severe the pandemic, the more likely an individual will pick public indoor activities like shopping during leisure time. As the results, increase the national EDI due to the increase of customer consumption. Furthermore, since high birth rate is often an indicator of higher distribution of people in the young adult groups, an increase of birth rate will increase the national EDI (more people in the workforce). However, it is not logical to conclude that an increase of new case counts will result in higher EDI. Therefore, there might be potential confounders.

We used the United States to demonstrate how our findings could be used to predict the EDI economic indicator. The same approach could be applied to other countries/cities with appropriately processed data. Concretely, we used the Prophet to forecast what the selected predictors will be like in the U.S on July 1st, 2021 (approximately 1 year from now). The output was then fed into our trained multivariate regression model to predict the difference in EDI and to produce the Predicted EDI, as illustrated in Table 2 above. Our result indicated that the U.S economy will still be impacted negatively due to factors associated to COVID-19 approximately 1 year from now. However, further analysis suggested that our approach is not applicable to make long term projections. This could due to insufficient sample size, which could lead to increasing uncertainty in long term forecasting.

One approach we came up with is to use clustering algorithms such as K-means to find out countries or cities that are within the same cluster as cases we have studied. This way we could scale our findings to cities and counties from the entire world and would be alot easier than finding datasets to do forecasting. We could simply cluster cities and countries using demographics and COVID-19 policies. We suspect countries with similar demographics and policies will experience a similar pandemic economic trend.

For countries that are facing devastating long-term economic effects of the pandemic, we will recommend to enforce more stringent policies on school closures, workplace closures, and travel bans to reduce new case counts and death counts. In addition, we will recommend countries with a high GDP percentage in the service sector to formulate a strategy to boost GDP in other sectors like agriculture, allowing employees to work in a outdoor setting to reduce infection rate.

Limitations

There exist some key limitations in our analysis and modeling. Firstly, we only looked at 1 year in the future. We tried to project longer than 1 year; However, due accumulated uncertainty the further we move in the future, the confidence interval of our predicted features approaches infinitely large. We believe it is impractical to forecast the long-term future since we only have approximately 5 month of pandemic data. Another limitation is that we used interpolation to generate daily data observations. Since the most granular time frame of our response variable is a month, we do not have enough data points given we can only find pandemic data for 5 month. The solution we took is to interpolate daily observations using data points of consecutive months.

References

1. Denworth, L. (2020, June 01). How the COVID-19 Pandemic Could End. Retrieved November 01, 2020, from https://www.scientificamerican.com/article/how-the-covid-19-pandemic-could-end1/
2. G. (n.d.). Community Mobility Reports. Retrieved November 1, 2020, from https://www.google.com/covid19/mobility/
3. R. (2019, June 15). GDP per capita (1990-2017). Retrieved November 01, 2020, from https://www.kaggle.com/robstepanyan/gdp-per-capita-19902017
4. Coalition, R. (2020, August 10). UNCOVER COVID-19 Challenge. Retrieved November 01, 2020, from https://www.kaggle.com/roche-data-science-coalition/uncover
5. AI, A. (2020, October 30). COVID-19 Open Research Dataset Challenge (CORD-19). Retrieved November 01, 2020, from https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
6. T. (n.d.). Global Economic Monitor (GEM). Retrieved November 01, 2020, from https://databank.worldbank.org/source/global-economic-monitor-(gem)
7. T. (n.d.). Economic Indicators: List By Category. Retrieved November 01, 2020, from https://tradingeconomics.com/indicators

Table 1. Datasets Variables Description

Variable Name	**Variable Type**	**Explanation**
Response	Response Label	Difference in Economic development index (EDI) for no-corona vs corona situation. See economic metric section in Methodology for more detail.
Date (MM-DD-YYYY)	General Predictors	Range: 2020-02-15~2020-07-01
Population Density	General Predictors	Measurement of population per unit area (per square miles)
Birth Rate	General Predictors	Total number of live births per 1,000 population divided by the length of the period in years.
Death Rate	General Predictors	Total number of deaths per 1,000 population divided by the length of the period in years.
Retail & Recreation Percent Change From Baseline	Economic Predictors	Mobility trends for places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.
Grocery & Pharmacy Percent Change From Baseline	Economic Predictors	Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.
Parks Percent Change From Baseline	Economic Predictors	Mobility trends for places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens.
Transit Stations Percent Change From Baseline	Economic Predictors	Mobility trends for places like public transport hubs such as subway, bus, and train stations.
Residential Percent Change From Baseline	Economic Predictors	Mobility trends for places of residence.
Agriculture	Economic Predictors	% of GDP Contributed by Agriculture Sector
Industry	Economic Predictors	% of GDP Contributed by Industry Sector
Service	Economic Predictors	% of GDP Contributed by Service Sector
New Cases Per Million	COVID19 Predictors	Country Level
New Deaths Per Million	COVID19 Predictors	Country Level

Stringency Index	COVID19 Predictors	Composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest)
Hospital Beds Per Thousand	COVID19 Predictors	Hospital beds per 1000 people.