PREDICTION OF COVID-19 CONFIRMED CASES & FATALITIES

ECE 884 – DEEP LEARNING

FINAL PROJECT

Syed Kashif Kamoonpuri

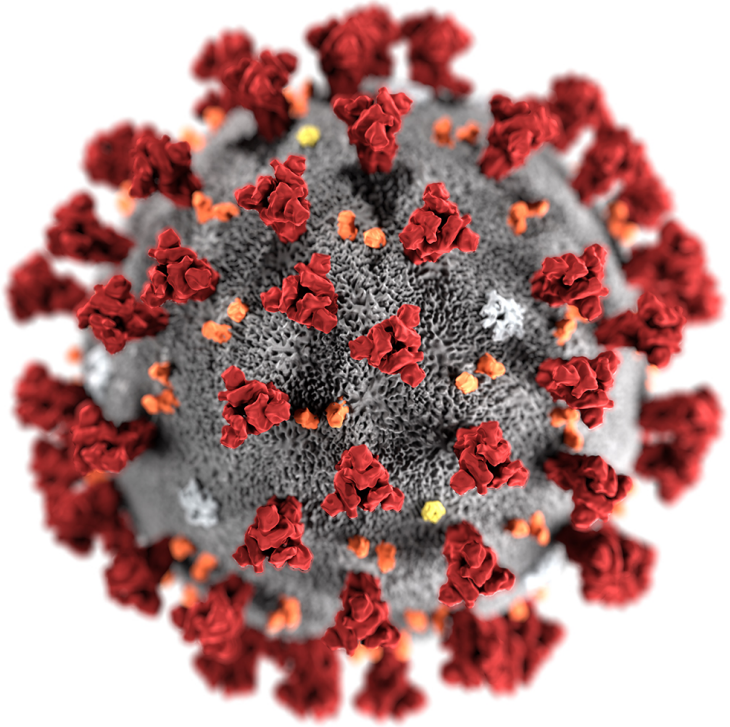
kamoonpu@msu.edu

INTRODUCTION

The objective of this project is to build a machine learning model that predicts the expected number of confirmed cases and fatalities due to COVID-19 in a region, by analyzing past trends.

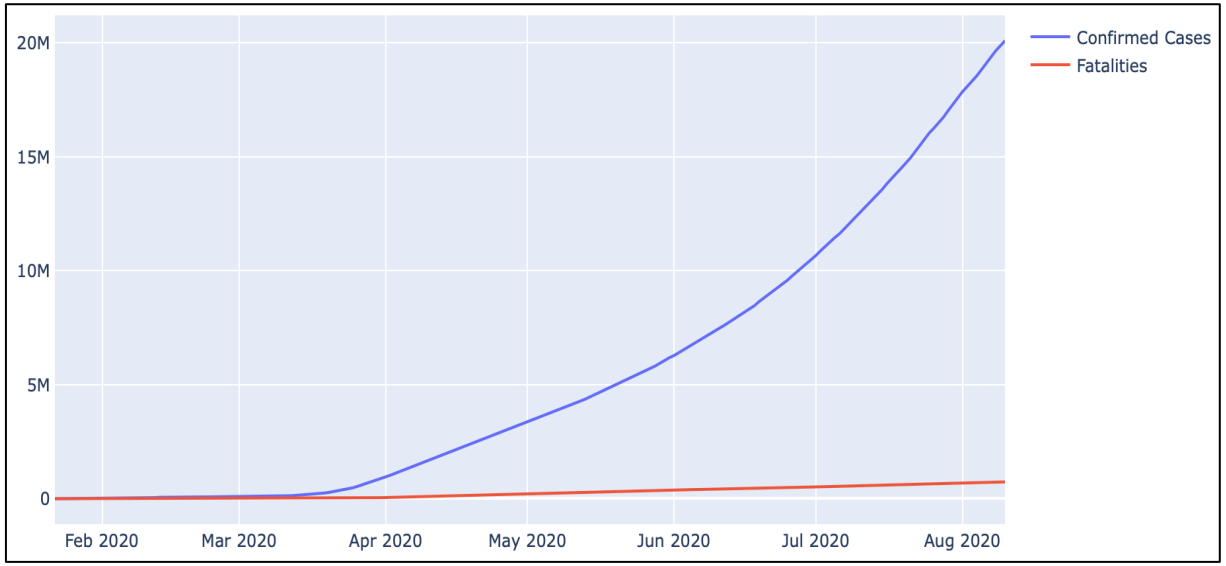
For this project, I have used historical data of COVID-19 cases and fatalities of 188 countries, from January 22 till August 10.

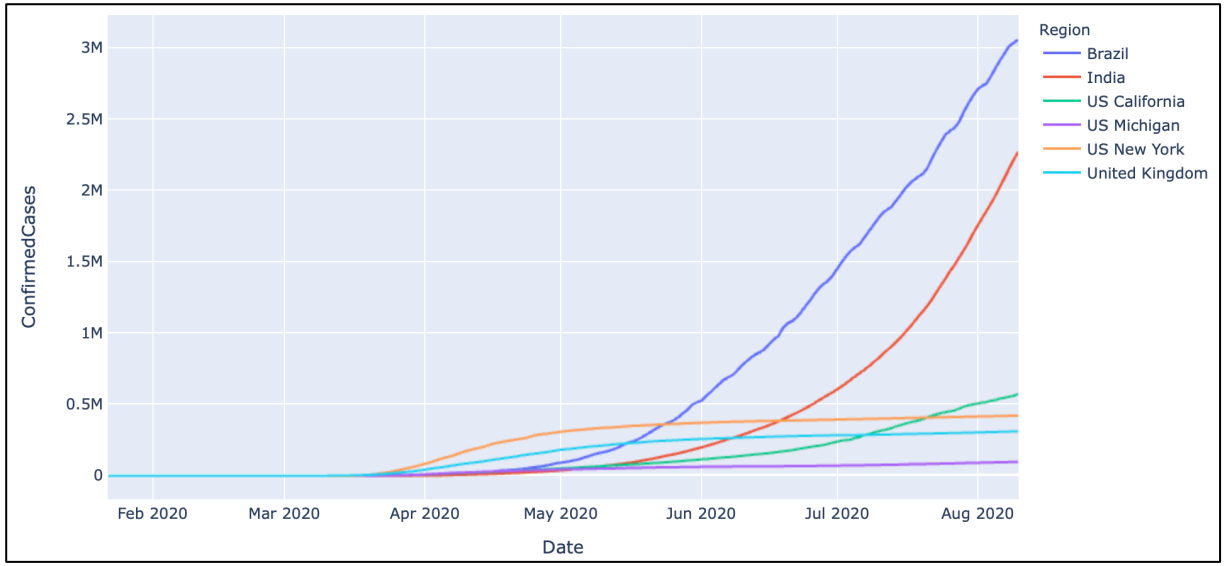
The dataset was downloaded from the GitHub repository of Center for Systems Science and Engineering, Johns Hopkins Whiting School of Engineering.[1]



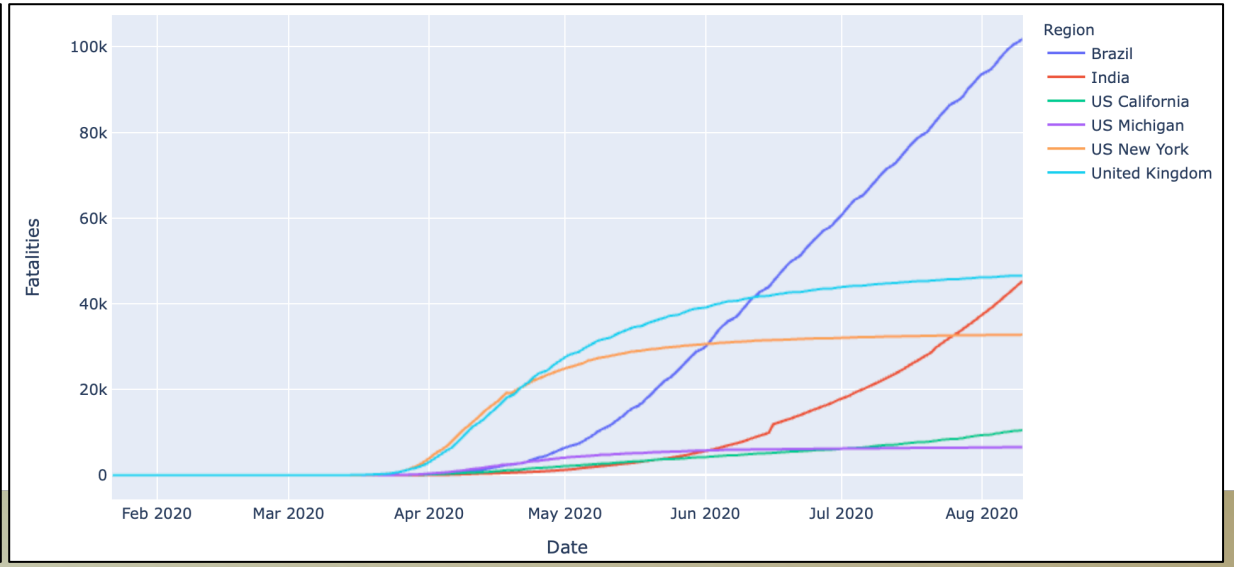
1. Available at: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

EXPLORATORY DATA ANALYSIS

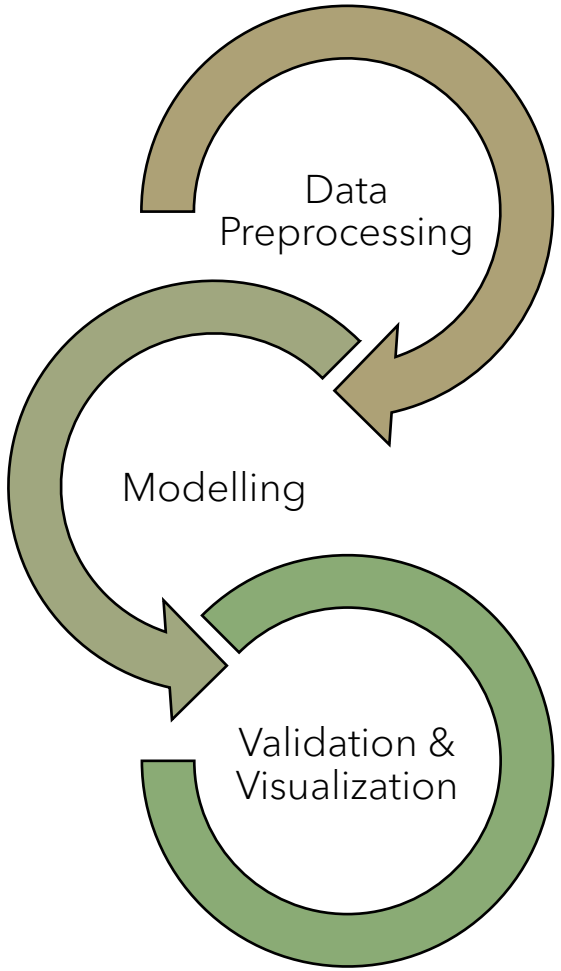




Region	Date	ConfirmedCases	Fatalities
Brazil	2020-08-10	3057470	101752
India	2020-08-10	2268675	45257
Russia	2020-08-10	890799	14973
US California	2020-08-10	574231	10476
South Africa	2020-08-10	563598	10621



PROJECT PHASES



DATA OVERVIEW

Column Name	Description
Date	Ranges from '01-22-2020' to '08-10-2020'
Country_Region	Country Name
Province_State	Province Name
ConfirmedCases	Total number of confirmed cases due to COVID-19 as of respective date
Fatalities	Total number of fatalities due to COVID-19 as of respective date

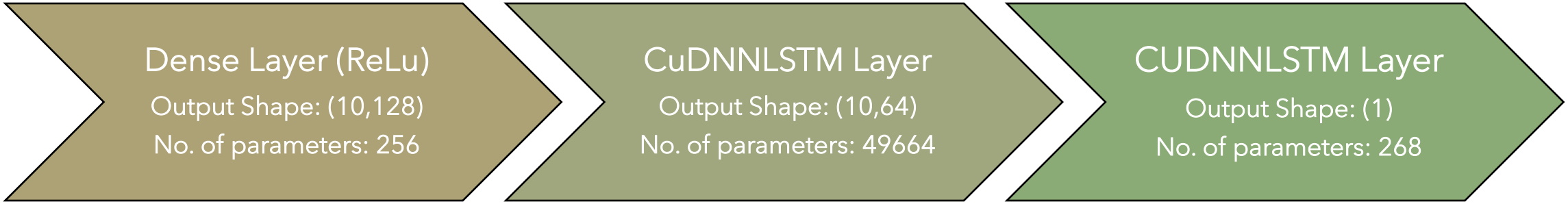
PHASE 1: Data Preprocessing

Before the data was ready for modelling, I had to take following steps to preprocess the data:

- Converting data from wide to long format
- Concatenating global COVID-19 dataset with United State's COVID-19 dataset
- Dealing with missing values in Province_State column
- Normalizing values for ConfirmedCases & Fatalities using MinMaxScaler()
- Adding confirmed cases and fatalities for the last 10 days in each row
- Splitting the data into train and test set (150 days for training and 41 days for testing)

PHASE 2: Modelling

The architecture for the model is as follows:



Please note that the input shape for the model is (10,1)

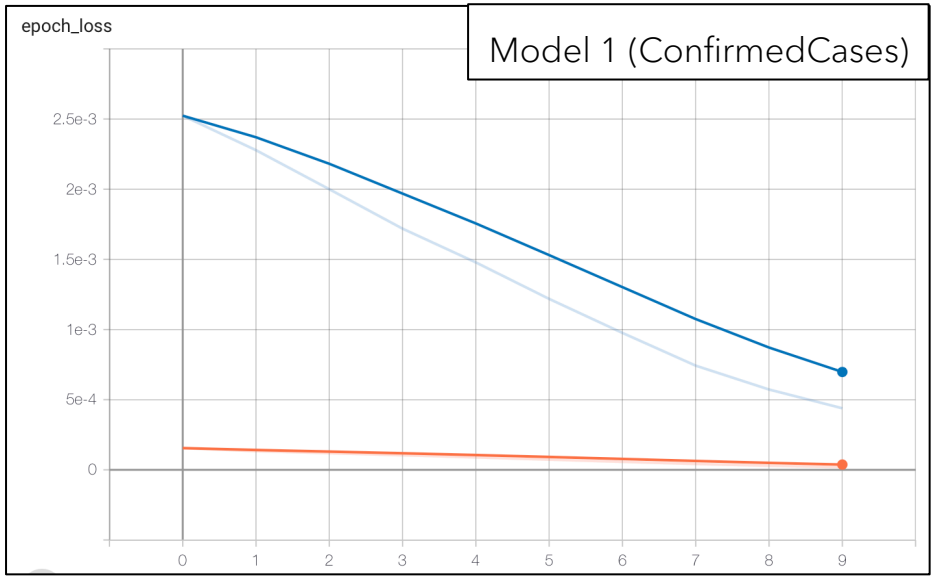
PHASE 3: Validation & Visualization

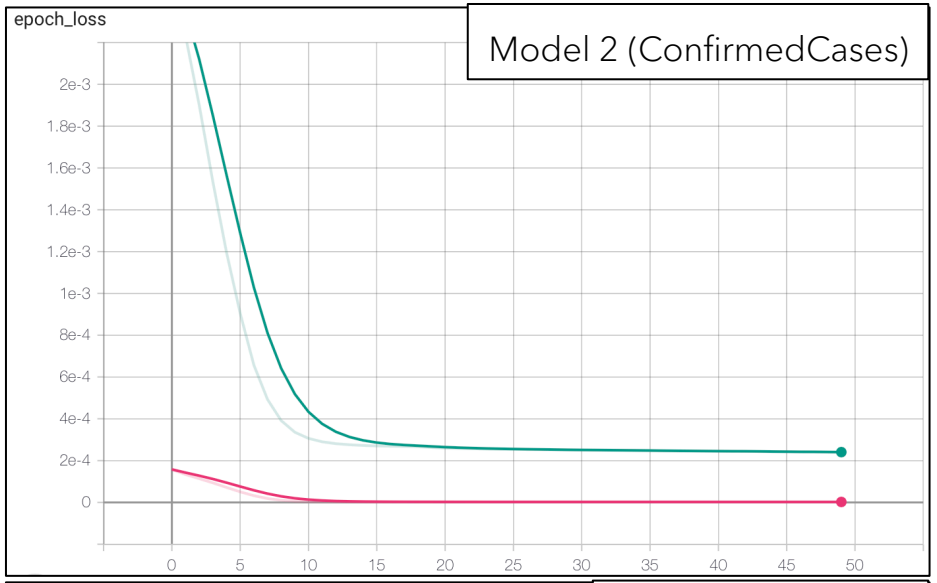
The model was trained and tested on two sets of hyper-parameters:

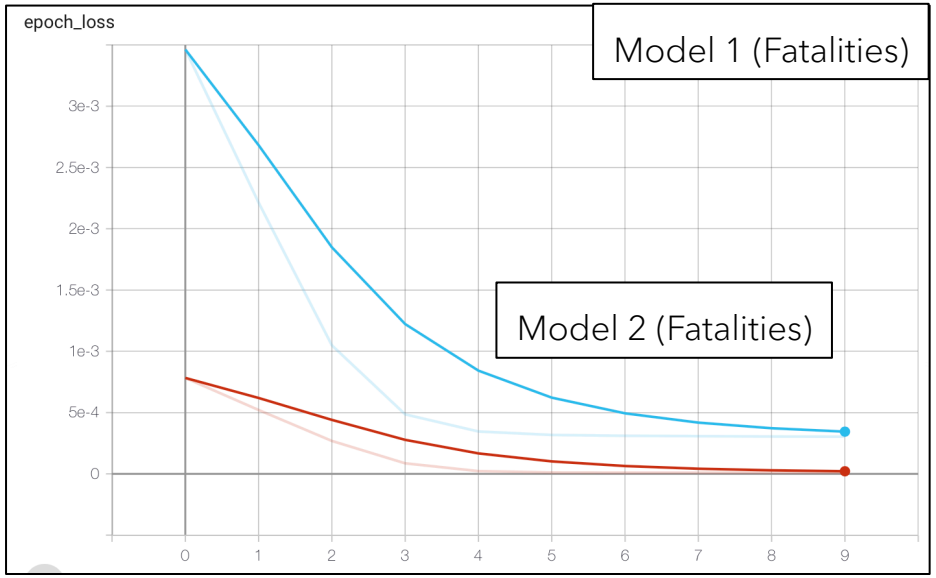
- Loss: SGD and epochs: 10
- Loss: adam and epochs:50

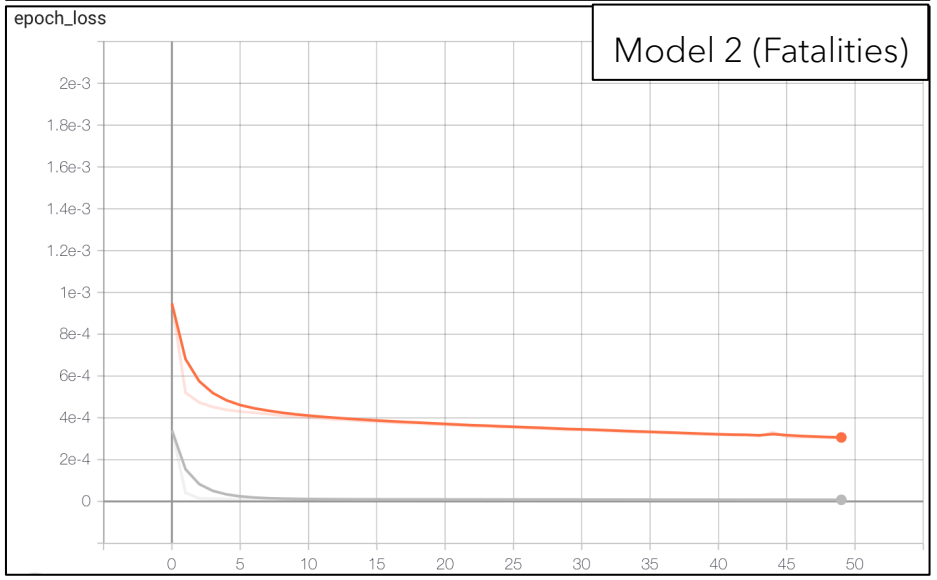
Model #	Parameters	Target Variable	Loss	RMSE
1	Optimizer: SGD Epochs: 10	ConfirmedCases	0.0007	0.027
		Fatalities	0.0003	0.019
2	Optimizer: SGD Epochs: 50	ConfirmedCases	0.0002	0.015
		Fatalities	0.0003	0.017

Epoch vs loss graph for all models:

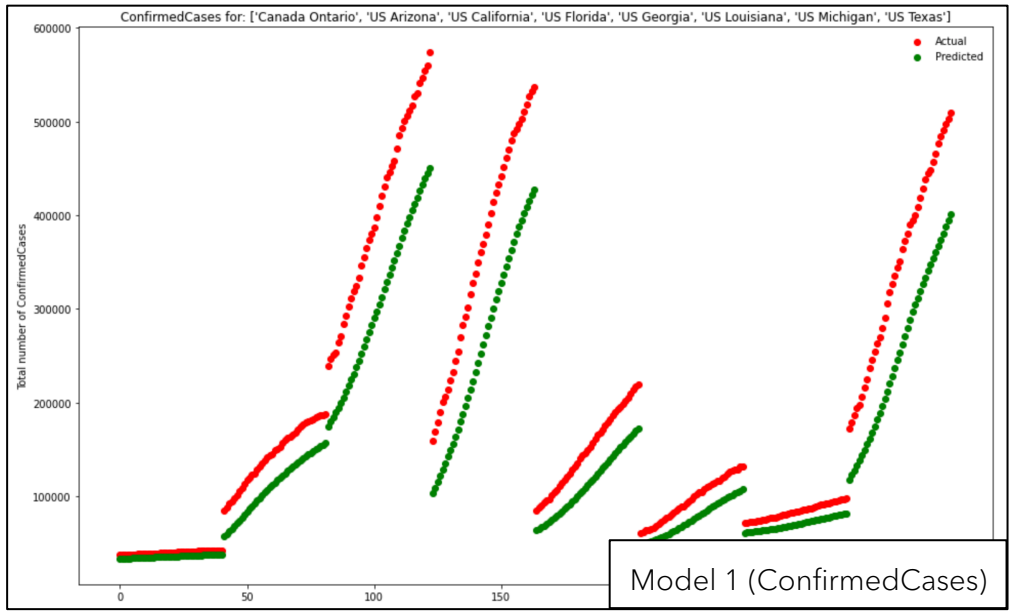




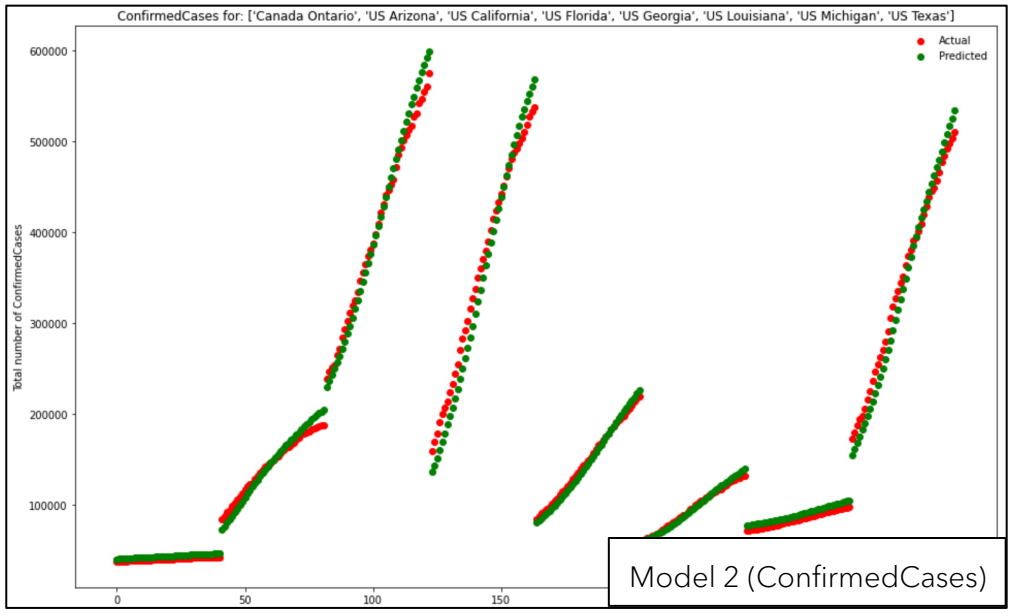




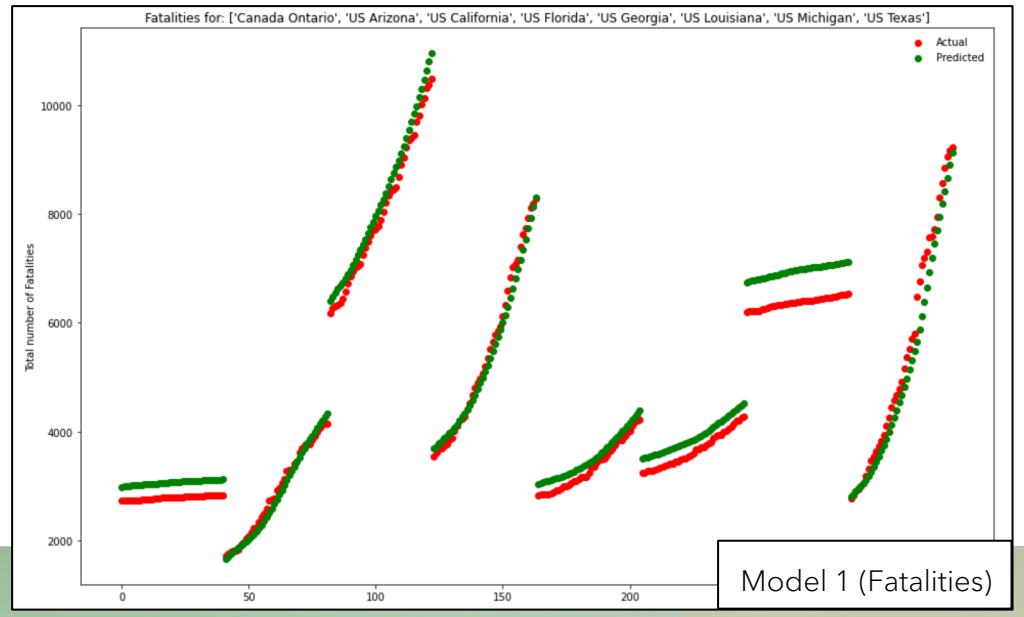
Visualizing model performance:



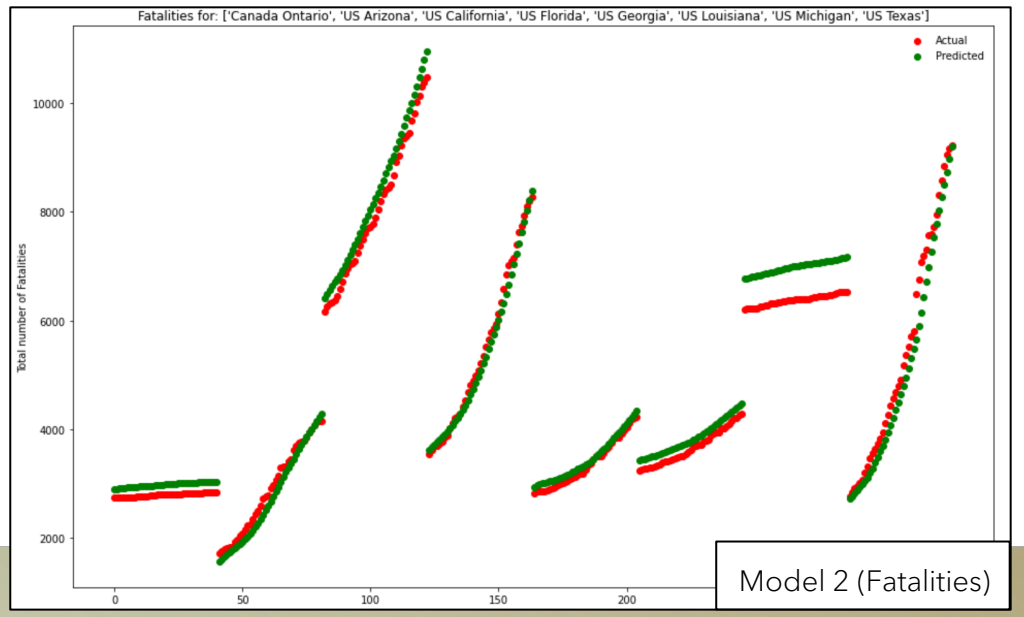
Model 1 (ConfirmedCases)



Model 2 (ConfirmedCases)



Model 1 (Fatalities)



Model 2 (Fatalities)

EXPERIMENTATION

In order to improve the model, I downloaded mobility_report.csv from Google Cloud Platform.[2] The dataset contains following values for each country/sub-region:

- retail_and_recreation_percent_change_from_baseline
- grocery_and_pharmacy_percent_change_from_baseline
- parks_percent_change_from_baseline
- transit_stations_percent_change_from_baseline
- workplaces_percent_change_from_baseline
- residential_percent_change_from_baseline

However, I noticed that the performance of the neural network model decreased after adding these parameters. Hence, I did not include these parameters in the final model.

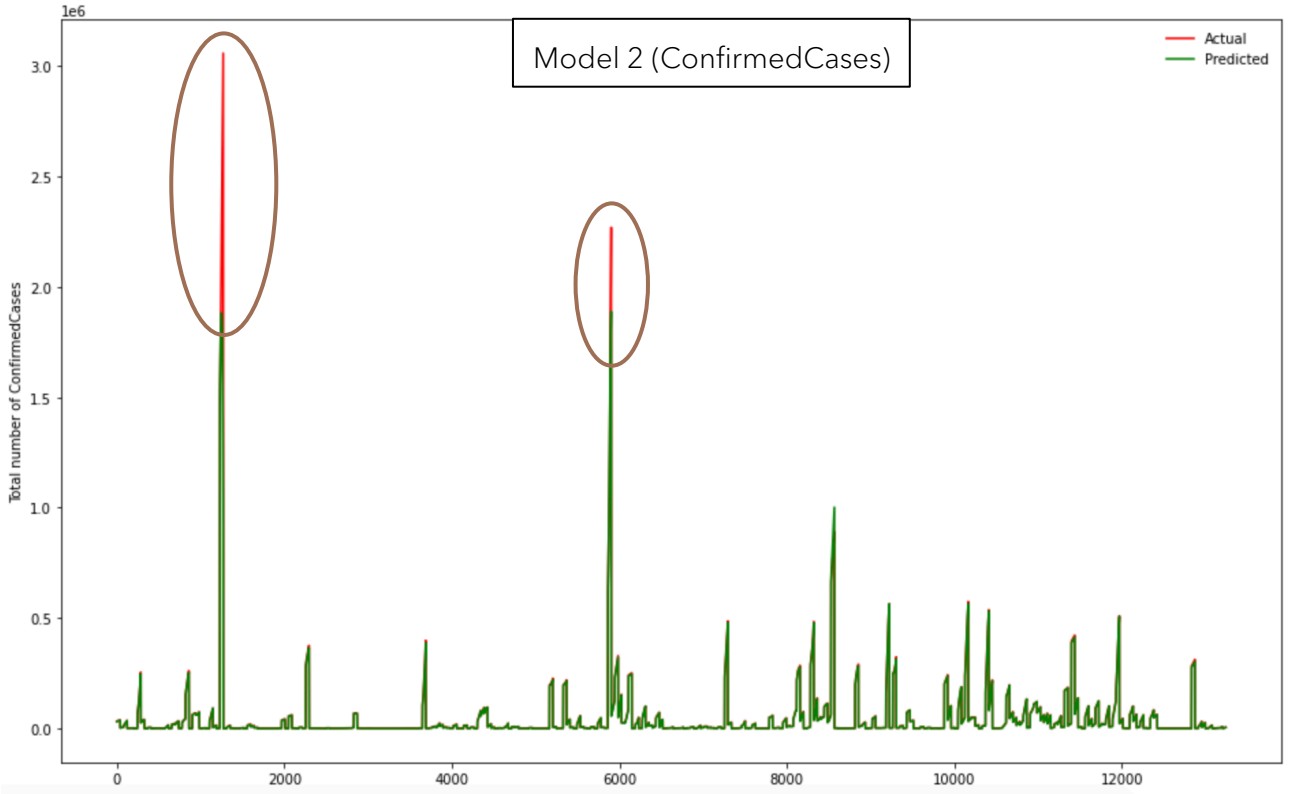
country_region	sub_region_1	date	retail_and_recreation_	grocery_and_pharmacy_	parks_percent_	transit_stations_	workplaces_	residential_
United Arab Emirates	Abu Dhabi	2/15/20	1	6	-2	-1	2	1
United Arab Emirates	Abu Dhabi	2/16/20	-2	5	2	-2	2	1
United Arab Emirates	Abu Dhabi	2/17/20	-3	2	4	-3	2	1
United Arab Emirates	Abu Dhabi	2/18/20	-3	2	1	-2	2	1
United Arab Emirates	Abu Dhabi	2/19/20	-3	1	0	-1	2	1
United Arab Emirates	Abu Dhabi	2/20/20	-2	2	3	-3	1	1

2. Available at: https://console.cloud.google.com/bigquery?project=upbeat-aspect-285714&folder=&organizationId=&p=bigquery-public-data&d=covid19_google_mobility

FUTURE SCOPE

The model performs quite well for most of the data except the circled region, where actual numbers are very different from predicted numbers. These two spikes contain data for Brazil and India.

Brazil and India have 2nd and 3rd highest number of COVID-19 cases, respectively. US has the highest number of COVID-19 cases, but the data for US has been recorded for each state. In order to improve the model, the state-wise data for Brazil and India should be collected.



Thank You!

Questions?