Forecasting daily COVID-19 spread in regions around the world.

Frank Lawrence Nii Adoquaye Acquaye*
Faculty of Computer Science
National Research University Higher School of Economics
fakvey@edu.hse.ru

Bense Valerie Caroline
Faculty of Computer Science
National Research University Higher School of Economics
vbense@edu.hse.ru

Abstract

1 Problem statement:

The year 2020 will forever be remembered as the year the earth stood still. This is primarily due to the spread of COVID-19. As Data Scientists we seek to provide solutions to problems facing humanity and the world at large. In this regard we seek to develop a forecasting model that will predict the daily spread of COVID-19 in regions around the world. Our model predicts the number of daily new cases in regions around the world in order to help policy makers plan and manage the COVID-19 pandemic.

2 Dataset summary and EDA:

2.1 Background of dataset:

The White House Office of Science and Technology Policy (OSTP) pulled together a coalition of research groups and companies (including Kaggle) to prepare the COVID-19 Open Research Dataset (CORD-19) to attempt to address key open scientific questions on COVID-19. Those questions are drawn from National Academies of Sciences, Engineering, and Medicine's (NASEM) and the World Health Organization (WHO).

2.2 Data sources:

The sources of data used in this project can be obtained from Kaggle Dataset

2.3 Actual data:

Since the accuracy of such a model is dependent on the freshness of the data, the most up to date data can be found here

*http://acquayefrank.github.io

	Id	County	Province_State	Country_Region	Population	Weight	Date	Target	TargetValue
0	1	NaN	NaN	Afghanistan	27657145	0.058359	2020-01-23	ConfirmedCases	0.0
1	2	NaN	NaN	Afghanistan	27657145	0.583587	2020-01-23	Fatalities	0.0
2	3	NaN	NaN	Afghanistan	27657145	0.058359	2020-01-24	ConfirmedCases	0.0
3	4	NaN	NaN	Afghanistan	27657145	0.583587	2020-01-24	Fatalities	0.0
4	5	NaN	NaN	Afghanistan	27657145	0.058359	2020-01-25	ConfirmedCases	0.0

Figure 1: Preview of first five line items in training dataset.

	ForecastId	County	Province_State	Country_Region	Population	Weight	Date	Target
0	1	NaN	NaN	Afghanistan	27657145	0.058359	2020-04-27	ConfirmedCases
1	2	NaN	NaN	Afghanistan	27657145	0.583587	2020-04-27	Fatalities
2	3	NaN	NaN	Afghanistan	27657145	0.058359	2020-04-28	ConfirmedCases
3	4	NaN	NaN	Afghanistan	27657145	0.583587	2020-04-28	Fatalities
4	5	NaN	NaN	Afghanistan	27657145	0.058359	2020-04-29	ConfirmedCases

Figure 2: Preview of first five line items in test dataset.

2.4 Actual data used in project:

In this project we use frozen dataset i.e dataset that has been frozen in time and this dataset can be found here

2.5 Basic exploratory data analysis

The dataset for training consists of 8 primary variables with a total of 914232 line items. 1.6% of the line items contained missing data. *Figure 1* shows the a preview of the first five line items in training dataset. Upon further investigations we realise that most instances had $Country = NaN$, also $Province\_State = Nan$ except in cases where the $Country\_Region = U.S.A$. For this reason we exempted these two variables or attributes. In the exploratory stage, there is no clear description of the weight parameter, but we may experiment with it, to see how it impacts predictions but most likely we may drop it.

Figure 2 shows the first five rows of the data for testing. One can easily realise that the target value was not supplied, for this reason we will drop the test dataset and split our training dataset in a manner that allows us to test our models.

We split our training data into *confirmed cases* and *fatalities*. This can be confirmed from *Figure 3*. A simple summary statistic of confirmed cases and fatalities is shown is shown in *Figure 4* and *Figure 5*.
We tried to have a fair understanding of the growth rate of confirmed cases globally by week and it seems exponential. This can be seen if *Figure 6*

In order to have a better understanding of the data, kindly follow the links below to view a detailed report of our EDA. To preview the profile of [train.csv] follow this link. To preview the profile of [test.csv] follow this link. To view a plot of fatalities vs first infections kindly follow this link

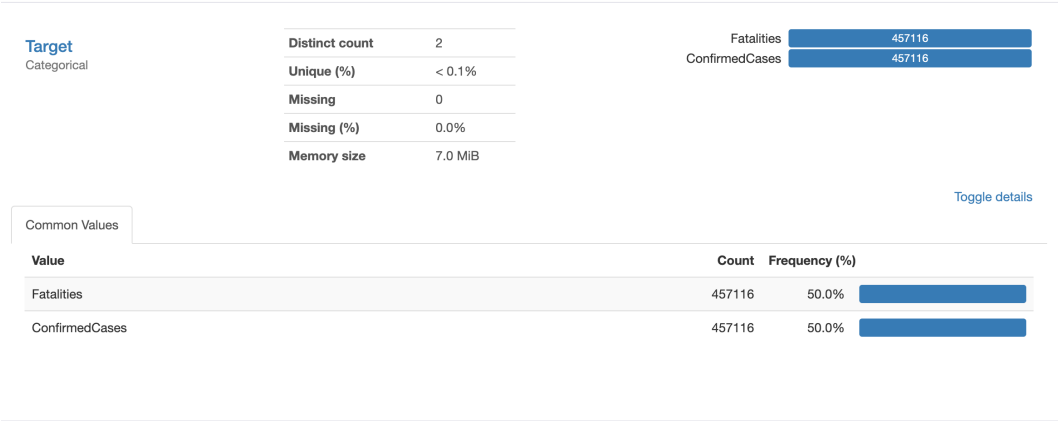

Figure 3: Instances of confirmed cases and instances of fatalities.

	Population	Weight	TargetValue
count	4.571160e+05	457116.000000	457116.000000
mean	2.720127e+06	0.965218	1.319908
std	3.477773e+07	0.175551	28.767670
min	8.600000e+01	0.474908	-1918.000000
25%	1.213300e+04	0.864488	0.000000
50%	3.053100e+04	0.968379	0.000000
75%	1.056120e+05	1.063404	0.000000
max	1.395773e+09	2.239186	4591.000000

Figure 4: Summary statistics of fatalities.

	Population	Weight	TargetValue
count	4.571160e+05	457116.000000	457116.000000
mean	2.720127e+06	0.096522	22.328864
std	3.477773e+07	0.017555	407.011027
min	8.600000e+01	0.047491	-10034.000000
25%	1.213300e+04	0.086449	0.000000
50%	3.053100e+04	0.096838	0.000000
75%	1.056120e+05	0.106340	0.000000
max	1.395773e+09	0.223919	36163.000000

Figure 5: Summary statistics of confirmed cases.

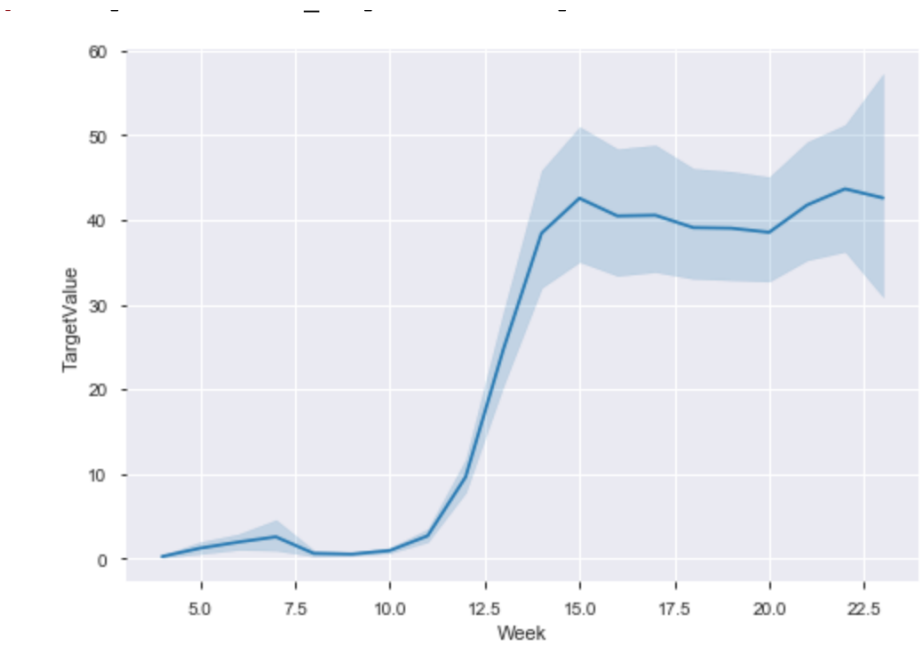

Figure 6: Number of confirmed cases by week.

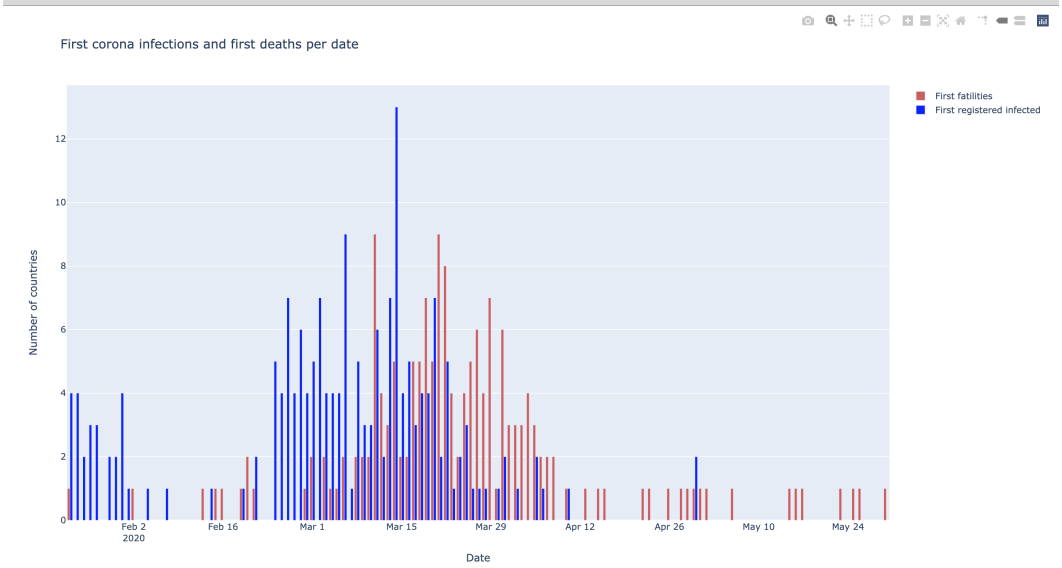

Figure 7: A distribution of first infections vs first deaths.