Covid- 19 Reproduction Rate Estimator

KSchool

Madrid, May 2021

Lucas Botella Roca

Index

The presented document will serve as a guide on how to understand and explore the presented project that contains main information and code in GitHub.

GitHub repo link: https://github.com/lucasbotellaroca/Covid-19_Reproduction_Rate_Estimator

1. Introduction

2. Problem Statement and Modelling

The intention of the presented project is to explain and understand how restrictions, mobility trends, temperature, demographic and special characteristics of each region affect the spread of Covid-19. In order to achieve this, we will try to predict the reproduction number for each week. We will refer to *effective reproduction number* or *reproduction rate* as *Rt*. *Rt* tells you the average number of people who will contract a contagious disease from one person with that disease. The *basic reproduction number* or *R0* specifically applies to a population of people who were previously free of infection and haven't been vaccinated, however for our specific case of study we will take into account people that have been vaccinated or that have contracted the disease in order to make our predictions more precise, therefore predicting *Rt*.

In general terms, as mentioned, *Rt* estimates in average how many infections may be caused by one infected individual, this of course is related with mobility trends, restrictions applied by governments and many other factors.

- If *Rt < 1* one infected person will cause less than one infection. In this case the disease will eventually die out.
- If *Rt > 1* one infected person will cause more than one infection. In this case the disease will increase and eventually cause an outbreak or pandemic.
- If *Rt = 1* one infected person will cause one infection. In this case the disease will still be transmitted and there is a risk of outbreak or pandemic.

Coronavirus data specially cases and deaths reported by governments are not very trustworthy, especially in the toughest times of the epidemic. Taking this into account our proxy variable to detect possible infections will be the excess mortality recorded. Excess mortality is a measure of the excess number of deaths recorded in 2020 and 2021 in relation with previous years by week, such difference will of course, indicate us, the number of deaths caused by coronavirus disease, making the assumption that there are no other causes that may cause an excess of deaths.

Our approach for the presented problem is based on SIR model which is a standard used in epidemiology for a disease spread in the population.

The standard SIR model in discrete times describes the reproduction rate of a virus based on three components referred as: susceptible (S_t), infected (I_t), and recovered (R_t) in time t. β_t is the transmission rate, and γ_t is the transition rate from infected to recovered in time t
Note the difference between *Rt* which we refer as *effective reproduction number* and R_t which we refer to as *recovered* individuals in time t.

Let´s note that $N = St + It + Rt$. The original SIR problem is stated as shown below:

$$S_t = S_{t-1} - \beta_t I_{t-1} S_{t-1} \frac{S_{t-1}}{N}$$
$$I_t = I_{t-1} - \beta_t I_{t-1} S_{t-1} \frac{S_{t-1}}{N} - \gamma I_{t-1}$$
$$R_t = R_{t-1} - \gamma I_{t-1}$$

To simplify things, *R0* is defined as for whatever defined time period as $R0=\beta/\gamma$. *Rt* is defined as shown in the equation below. *It* is referred to as the number of individuals infected in time *t*.

$$Rt = 1 + \frac{I_t - I_{t-1}}{I_t \gamma}$$

Rt therefore is a value that measures how the virus is increasing or decreasing in time. For our specific problem we will not try to exactly replicate this idea, but our dataset structure will be based on the equation system shown above. Key points taken from this model is that *Rt* is dependent of infections in the time period defined *t,* accumulated infections or recovered *Rt*, and for our specific case the restrictions applied. If all infected individuals were isolated from the rest of the population for γ time, then the disease would disappear.

Excess mortality is recorded weekly on Sundays, and that value is the sum of deaths in the deferred week. In this project we will take excess mortality as an indicator or proxy variable of both accumulated and recovered individuals together with infected individuals. Accumulated will be the sum of excess mortality in time t, such value is calculated by summing deaths for every country until time n. It has been recorded that the average time between a person contracting the virus and dying is 18 days (Verity et al., 2020). It has also been recorded, that people most infectious period is between 5 and 12 days after infection. We will take 7 as average, which is the value that best fits our data structure (weekly). Then the amount of infected of individuals in week n will be reflected as excess deaths in day n+11- Therefore infections will be estimated as the excess mortality recorded in day n+(18-7) = n+11.

As mentioned, excess mortality data is unfortunately retrieved weekly, hence, every entry in our dataset will be a week estimate of value *Rt*, which in the end makes our dataset smaller and more aggregated which may affect the results.

Therefore, our problem is stated as shown below and will be referred as **no lags** methodology:

Restrictions = Restrictions applied by governments in week n.
Mobility Trend = Mobility trends provide by Google.
Others = Demographic and other variables unique for each country that may affect the spread of the disease.
Recovered = Accumulated excess deaths to week n.
Infected = Excess deaths in next 11 days.

$$f_{week\ n}\ (Restrictions,\ Mobility\ Trends,\ Others,\ Recovered,\ Infected) = Rt_{week\ n}$$

According to the structure mentioned above, every row in our dataset will contain restrictions, mobility trends in week n, country characteristics and other metrics detailed in section below, recovered population until week n, and infected individuals in week n. Infected individuals as mentioned are estimated as the excess mortality in n+11 days from the selected week, which will serve as proxy or approximation of infected individuals. The target variables of course, the reproduction rate in the selected week.

The approach will be to try to estimate the value of reproduction rate with regression models in order to evaluate the effect and importance of each of the features included in the dataset. This first data configuration not including lags is to serve and its intention is to try to understand the sign and importance of features included. Therefore, this first approach is more of an explanatory model of how covid spreads in relation with features included. A forecast model will be included using lags, which is detailed in this section.

It is important to note, that as seen there are many imperfections that may affect our predictions, we are working with mean values therefore, a deviation with the target variable is already expected. When grouping values weekly and grouping to closest Sunday we are of course losing information, however data is aggregated weekly, and we will treat it as it is.
As mentioned, the proxy variable used for infections and accumulated total is excess mortality. Excess mortality variable is pretty reliable since accounted deaths worldwide seems to be trustful, however there are some key points that need to be taken into account. The improvements in

treatments given to patients has predictably increased over time, therefore excess of deaths accounted in March 2020 are not the same as the excess of deaths accounted in February 2021 in terms of survival to treatments, however there is no analytical correction in these terms in the presented project.

The detailed grouping refers to data configuration version grouped weekly, which in our GitHub repo refer to as models without lags tag. However, for a second methodology, which will be referred as **lags methodology** and second approach to the problem another model was implemented including lags of reproduction rates as a variable in our dataset. In order to achieve this, we will simply group data over two-week intervals taken the mean of selected weeks. The reason in this case grouping data every 2 weeks is due to the fact that adjacent values between week n and week n-1 of reproduction rate are very close, when grouping to two weeks, such relations decreased since reproduction rate values tend to change more over time.

Also, this model would serve more as a forecast model, since it does not include mobility indexes, and is a 2-week forward forecast, which in the end is how restrictions have been applied over time in all countries in two weeks' time windows.

Therefore, our data configuration for the mentioned methodology is:

Restrictions = Restrictions applied by governments in selected 2-week period.
Others = Demographic and other variables unique for each country that may affect the spread of the disease.
Recovered = Accumulated excess deaths to selected 2-week period.
Reproduction Rate Lag = Reproduction rate of week n-2.

$$f_{2\text{-week } n}\,(Restrictions, Others, Recovered, Infected, Reproduction\ Rate\ Lag) = Rt_{2\text{-week } n}$$

It is important to note that in this case we are not including mobility indexes. In this case our proxy variable infected is not included, as our indicator of virality or of current number of infected individuals is *Reproduction Rate Lag.* With this second data configuration we intend to make better and precise predictions, since we are going to be including a lag variable whose collinearity with the variable of two weeks are generally close.

Therefore, after this statement we have two data configurations one not including lags grouped weekly which its main intention is to understand and interpret the magnitude and significance of each feature, and another one including lags grouped every 2 weeks which its main goal is to accurately estimate the value in relation with the restrictions applied.

3. Data Gathering and Preparation

Once the problem has been stated and what the approach will be for this project will firstly detail our sources of data, features, range of values and description.

3.1 Data Gathering and Preparation Explanation

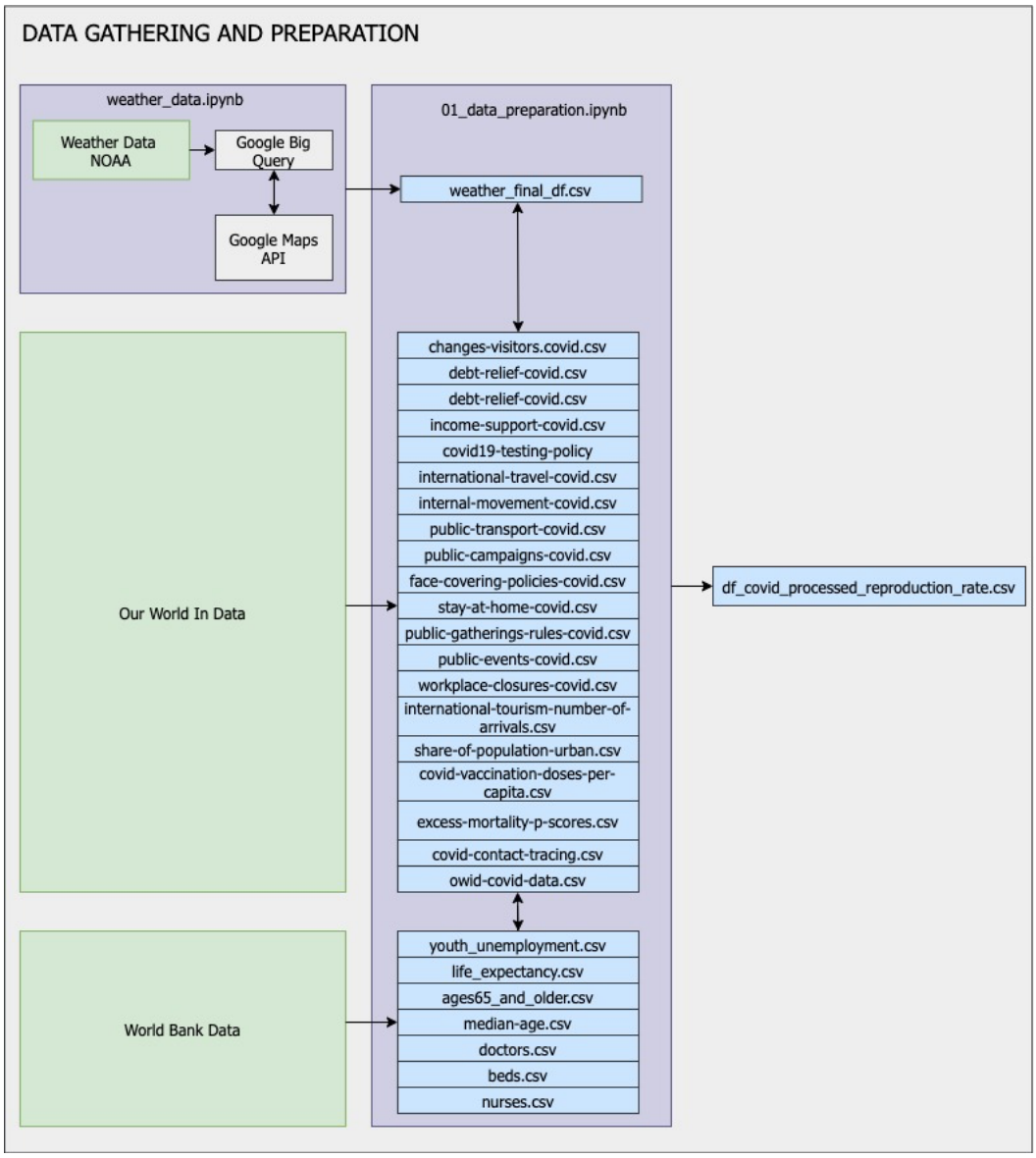
Data has been gathered form different sources:

- Our World in Data
- World Bank of Data
- National Oceanic and Atmospheric Administration (NOAA) Us department of commerce. (Accessed using Google Big Query)

All data has been retrieved from the sites mentioned. Afterwards data has been uploaded to a drive directory containing subdirectories. We will access data from the code directly from Google Drive. Our data preparation section may be summarized in two main phases

(1) Obtain weather data

(2) Obtain and process restrictions, mobility indexes, and characteristics per country including weather information, generating our final dataset for later processing. Below there is a scheme illustrating the data preparation phase.



Let's now detail each section separately:

1) Weather Data

The aim of this is to retrieve weather data including temperature and precipitation for all countries, since it may have effect on the spread of the virus.
As seen our first notebook "weather_data.ipynb" is able to invoke NOAA Weather Data database using Big Query. Information from all stations recorded in the NOAA database for all countries has been retrieved. This data is mostly retrieved daily depending on the station. For our specific case we will only retrieve precipitation and temperature information. However, while processing data, it was encountered, that countries abbreviation associated with the stations did not follow any of the country standard, and they followed a convention given by NOAA which was not feasible to access.

Therefore, since we had for every station, the date, latitude, longitude, temperature and precipitation it was decided to obtain the country of each station with conventional ISO Country Code, taking usage of Google Maps API.

Once all data has been accessed for all stations, we group them by countries taken the average, since our information for restrictions, mobility etc., comes grouped by countries.

Therefore, our data extraction for our weather data can be resumed as follows:

1) Retrieve date, latitude, longitude, temperature and precipitation for all stations worldwide.
2) Invoke Google Maps API with latitude and longitude parameter in order to extract country ISO Code associated with each station.
3) Group information on every date obtained by country mean.

Here is a look at our weather dataset:

	Code	Date	temp	prcp
0	AFG	2020-01-12	3.178571	0.037143
1	AFG	2020-01-19	2.465465	0.309009
2	AFG	2020-01-26	-0.958463	0.100000
3	AFG	2020-02-02	-0.615304	0.082075
4	AFG	2020-02-09	1.666667	0.031604
...	...	...	...	...
11962	ZWE	2021-03-14	23.790598	0.082308
11963	ZWE	2021-03-21	24.935185	0.000333
11964	ZWE	2021-03-28	24.713992	0.002963
11965	ZWE	2021-04-04	23.766667	0.000400
11966	ZWE	2021-04-11	23.986111	0.000000

11967 rows × 4 columns

Once all this data has been processed, it is exported as a .csv "weather_data.csv" which will be used in the data preparation notebook. Unfortunately, this code is not possible to be replicable. Queries executed by big query and requests to Google Maps API are very costing and were able to be executed in the past due to free trial access, and in case of activating another account it would just be possible to execute it once. However as said the complete dataset "weather_data.csv" can be accessed.

2) Generate Complete Dataset

As it has been stated in point 2 of the presented project (Problem Statement and Modeling) every row in our dataset will be defined as shown below for **no lag methodology**:

$$f_{week\ n}\ (Restrictions,\ Mobility\ Trends,\ Others,\ Recovered,\ Infected) = Rt_{week\ n}$$

As it has also been stated, infected data comes out every week, since it is a proxy variable of excess deaths per week, and such information comes out daily. In order to group variables accordingly the approach has been to add 11 days to all restrictions, mobility tends, others and recovered together with our target variable *Rt.* Therefore, all data retrieved follows the same approach add 11 days to date associated and group to closest Sunday, therefore linking such information with excess deaths in 11+ days, since deaths occurred in 11+ days, are associated with the amount of population infecting the referenced week. This is due to the fact, as mentioned, that one person in average dies 18 days after contracting the virus and starts being contagious 7 day after contracting the virus, therefore 11 days is our average number of days for our proxy variable. However of course, this approach may occur in imprecisions, but is the best we can do we the data is available.

For our **lag methodology version**, we will simply group data over 2-week intervals, exclude infections variable and replace it with reproduction rate of two weeks prior, however, this is not done in the data preparation file, it is done directly in the notebook for analysis. Therefore, for our lagged version we are left with the definition shown below:

$$f_{2\text{-}week\ n}\ (Restrictions,\ Others,\ Recovered,\ Infected,\ Reproduction\ Rate\ Lag) = Rt_{2\text{-}week\ n}$$

Once it has been detailed how our data has been gathered, generated and processed for both methodologies **(a)** weekly estimation including mobility with no lags and **(b)** 2-week estimation including lag variable of reproduction rate; in the next section will detail the description of each field and explore and make an explanatory data analysis on our dataset, getting some first insights around it.

4. Data Exploration

In this section we will explore our dataset in order to extract some insights around it. In this paper we will detail each field and make a description of the dataset, however in order to access the explanatory data analysis it is required to follow the notebook. This notebook contains various graphs that are commented and detailed there. Also, conclusions on exploratory data analysis can be found there. Therefore, we will now detail the structure and fields of the dataset.

4.1 Dataset Description Detail

Our final dataset has the following structure shown in the table, together with type and description for each field:

Features used for the Analysis		
Name	**Type**	**Description**
Code	String	Country in ISO 3166-1 alpha-2 Code
Date	Date	Date in yyyy-mm-dd format. Date contains only Sundays since it is grouped by week, all the rest of features are aggregated under this constraint.
retail_and_recreation	Float	Shows how the number of visitors to places of retail and recreation has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This includes places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters. This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
grocery_and_pharmacy	Float	Shows how the number of visitors to grocery and pharmacy stores has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This includes places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies. This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
residential	Float	Shows how the number of visitors to residential areas has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
transit_stations	Float	Shows how the number of visitors to transit stations has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This includes public transport hubs such as subway, bus, and train stations. This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
parks	Float	Shows how the number of visitors to parks and outdoor spaces has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This includes places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens.

		This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
workplaces	Float	Shows how the number of visitors to workplaces has changed compared to baseline days (the median value for the 5-week period from January 3 to February 6, 2020). This index is smoothed to the rolling 7-day average. **Range: [-100,100]**
contact_tracing	Integer	Government policies on contract tracing for COVID-19. - No tracing – 0 - Limited tracing (Only some cases) - 1 - Comprehensive tracing (All cases) – 2 - **Range: [0,2]**
testing_policy	Integer	Government policies on testing for COVID-19. Note that this relates to PCR testing for the virus only; it does not include non-PCR, antibody testing. - No testing policy- 0 - Testing only for those who both (a) have symptoms AND (b) meet specific criteria (e.g. key workers, admitted to hospital, came into contact with a known case, returned from overseas) - 1 - Testing of anyone showing COVID-19 symptoms - 2 - Open public testing (e.g "drive through" testing available to asymptomatic people) – 3 **Range: [0,3]**
international_travel_controls	Integer	Government policies on restrictions on international travel controls. - No measures - 0 - Screening - 1 - Quarantine from high-risk regions - 2 - Ban on high-risk regions - 3 - Total border closure – 4 - **Range: [0,4]**
restrictions_internal_movements	Integer	Government policies on restrictions on internal movement/travel between regions and cities. - No measures - 0 - Recommend movement restriction - 1 - Restrict movement – 2 **Range: [0,2]**
close_public_transport	Integer	Government policies on public transport closures - No measures - 0 - Recommended closing (or reduce volume) - 1 - Required closing (or prohibit most using it) – 2 **Range: [0,2]**
public_information_campaigns	Integer	Public information campaigns on COVID-19. - None – 0 - Public officials urging caution – 1 - Coordinated information campaign – 2 **Range: [0,2]**
facial_coverings	Integer	Government policies on the use of face coverings outside-of-the-home. Countries are grouped into five categories: - No policy - 0 - Recommended - 1 - Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible - 2 - Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible - 3 - Required outside the home at all times regardless of location or presence of other people – 4

		Note that there may be sub-national or regional differences in policies on face coverings. The policy categories shown may not apply at all sub-national levels. A country is coded based on its most stringent policy at the sub-national level. **Range: [0,4]**
stay_home_requirements	Integer	Government policies on stay-at-home requirements or household lockdowns. - No measures - 0 - Recommended not to leave the house - 1 - Required to not leave the house with exceptions for daily exercise, grocery shopping, and 'essential' trips - 2 - Required to not leave the house with minimal exceptions (e.g. allowed to leave only once every few days, or only one person can leave at a time, etc.) – 3 **Range: [0,3]**
restriction_gatherings	Integer	Government policies on restrictions on public gatherings. Countries are grouped into five categories: - No restrictions - 0 - Restrictions on very large gatherings (the limit is above 1000 people) - 1 - Restrictions on gatherings between 100 to 1000 people - 2 - Restrictions on gatherings between 10 to 100 people - 3 - Restrictions on gatherings of less than 10 people – 4 **Range: [0,4]**
cancel_public_events	Integer	Cancellation of public events. No measures - 0 Recommended cancellations - 1 Required cancellations – 2 **Range: [0,2]**
workplace_closures	Integer	Government policies on workplaces closures. No measures – 0 Recommended - 1 Required for some - 2 Required for all but key workers – 3 **Range: [0,3]**
school_closures	Integer	Government policies on school closures. No measures - 0 Recommended - 1 Required (only at some levels) - 2 Required (all levels) - 3 Note that there may be sub-national or regional differences in policies on school closures. The policy categories shown may not apply at all sub-national levels. A country is coded as 'required closures' if at least some sub-national regions have required closures. **Range: [0,3]**
debt_relief	Integer	Governments provide debt or contract relief to citizens during the COVID-19 pandemic. No relief – 0 Narrow relief – 1 Broad relief – 2 **Range: [0,2]**
income_support	Integer	Governments provide income support to workers during the COVID-19 pandemic. No income support – 0 Covers <50% of lost salary – 1 Covers >50% of lost salary – 2 **Range: [0,2]**
holiday	Integer	Number of holidays in the selected time period **Range: [0,7]**

temp	Float	Average temperature in celsius of all stations in the selected time period. **Range: [-20,40]**
prcp	Float	Average precipitation in mmph of all stations in the selected time period **Range: [0,3]**
doctors_per_1000	Float	Number of doctors per 1000 habitants last year recorded **Not used**
nurses_per_1000	Float	Number of nurses per 1000 habitants last year recorded **Not used**
beds_per_1000	Float	Number of hospital beds per 1000 habitants last year recorded **Not used**
number_of_arrivals	Float	Number of tourism arrivals last year recorded
urban_population	Float	Percentage of urban population last year recorded **Range: [0,100]**
total_vaccinations_per_100	Float	Share of the total population that received at least one vaccine dose. This may not equal the shares that are fully vaccinated if the vaccine requires two doses. **Range: [0,200]**
youth_unemployment	Float	Percentage of youthment unemployment last year recorded **Range: [0,100]**
life_expectancy	Float	Average life expectancy at birth last year recorded **Not used**
%df_population_gr_65	Float	Percentage of population with age 65 or higher last year recorded **Not used**
UN Population Division (Median Age) (2017)	Float	Median age last year recorded **Not used**
accumulated	Float	Accumulated percentage of deaths. **Range: [0,100]**
infections_value	Float	This value is the infection value related to week n. Hence this value is the excess mortality recorded for n+11 days from week n. Calculated as mentioned in section 1. **Range: [0,100]**
reproduction_rate	Float	This value is calculated as the increment/ decrement of deaths from in week n

After various try and error and changes in the approach of the problem there are some variables that were initially conceived to be used but were later on deprecated due to non-relation with our target variable reproduction rate.

The mentioned variables are:

- **Doctors, nurses and beds per 1000**, such variables should affect our proxy variable excess deaths, however it does not fit well in our model since doctors, nurses and beds have no relation in the spread of the virus, it may have a relation with the number of deaths but not with the reproduction rate, therefore they have not been used.
- **Life expectancy, population greater than 65 and median age** they all might have a relation with the number oof excess deaths however they don't fit in this model as well,

since its effect is related with one of the features (our proxy variable for infections which is excess deaths) but has no relation with our target variable reproduction rate.

4.2 Data Categorization

Since there are lots of variables and it may be confusing trying to understand all, below there is a scheme on the variables shown and their categorical classification.
There are four categories defined.
Mobility Factors: These refer to data retrieved by Google Mobility, this data shows how the number of visitors (or time spent) to different types of places has increased in pandemic times in relation with previous years.
Population Virus Infections and Immunity Factors: these factors represent the state of the virus in a certain population, in our case it is the country to which the data is referring to. These variables define the current situation of the country in terms of infected individuals, and people who are "immune" to the virus because they have already contracted the virus or been vaccinated.
Country Characteristics Factors: These factors are unique to each country and serve as a measure on related to mobility, and the type of mobility associated which in the end has effect on the virus spread.
Political Measures Factors: These refer to measures taken by governments.

Mobility Factors	Population Virus Infections and Immunity Factors	Country Characteristics Factors	Political Measures Factors
• retail_and_recreation • grocery_and_pharmacy • residential • transit_stations • parks • workplaces	• infections_value • accumulated • total_vaccinations_per_100	• temp • prcp • number_of_arrivals • urban_population • youth_unemployment • holiday	• debt_relief • income_support • testing_policy • international_travel_controls • restrictions_internal_movements • close_public_transport • public_information_campaigns • facial_coverings • contact_tracing • stay_home_requirements • restriction_gatherings • cancel_public_events • workplace_closures • school_closures

4.3 Data Grouping to Remove Multi Collinearity

Based on the data exploration section we have conducted that lots of variables have a high correlation, which is something expected.
Mobility is of course affected by the measures taken, and the other way around. Also, restrictions and measures do have a high correlation between each other, restrictions are usually applied together and with similar strictness.

Based on the article shown (https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea) there are three main approaches to assess the refered problem.

1. Drop variables, select meaningful features.
2. Transforming variables.
3. PCA: Principal Component Analysis

For our specific case, we will take three approaches.

- **Approach 1**: Raw data no grouping of variables or PCA Analysis - We first want to know how the model performs withOut any changes in our dataset.
- **Approach 2**: Grouping variables, decreasing dimensionality and correlation - We will perform feature engineering grouping variables with similar correlations on our target variable base on the data exploration section.
- **Approach 3: PCA**: Since variables have a high correlation, we will group variables taking use of Principal Component Analysis.

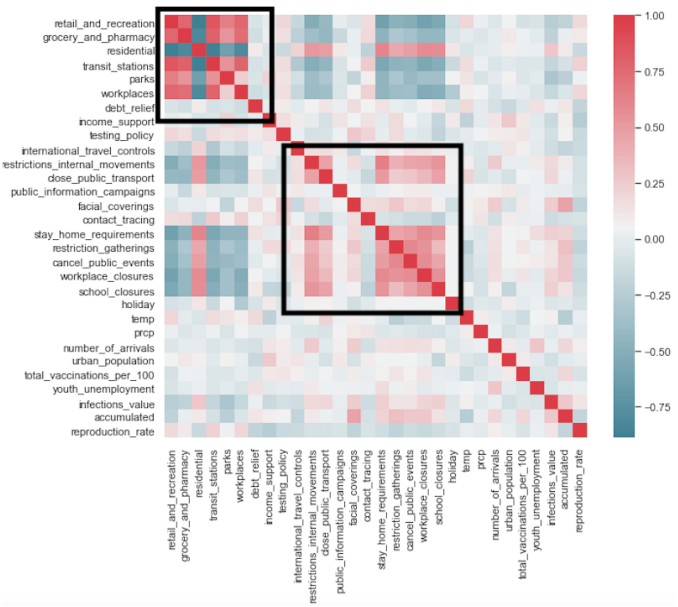
However, it is stated by some staticians that grouping variables when correlated does not necessarily mean an improvement in the predictions.
"The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations. —Applied Linear Statistical Models, p289, 4th Edition."

Even though when grouping variables, we might not have better predictions this might be useful in order to interpret the model afterwards, since one feature may absorb the effect of another. Hence, we will explore three approaches and observe which one best fits our purpose which as stated is not only getting precise predictions but also, a good interpretably and explanation of the model itself.
In the presented problem we will work with some models that work with null values and others that do not. In order to have appropiate datasets for all approaches with all the models evaluated we will create 2 sets of datasets for each approach**,** one with raw data df_appch_x**,** no null treatment "df_appch_x_clean"**,** and another with null values treatment and standard scaling applied. Approach 3 only contains one dataset since in order to perform PCA we need the dataset to be "clean".

Let's detail what data preprocessing we are going to use for every approach of the ones mentioned.

Approach 1: Raw data no grouping of variables or PCA Analysis
No data transformation done; we will train models with raw data.

Approach 2: Grouping variables, decreasing dimensionality and correlation
Since variables seem to have high multicollinearity will group variables with similar correlation. Grouping of variables based on previous knowledge acquired in the data exploration section. Below can be shown correlations between variables, they all refer to restrictions and mobility indexes, since mobility indexes tend to increase and decrease at the same time. That can also be observed whit restrictions, governments tend to apply restrictions at the same time with the same severity.

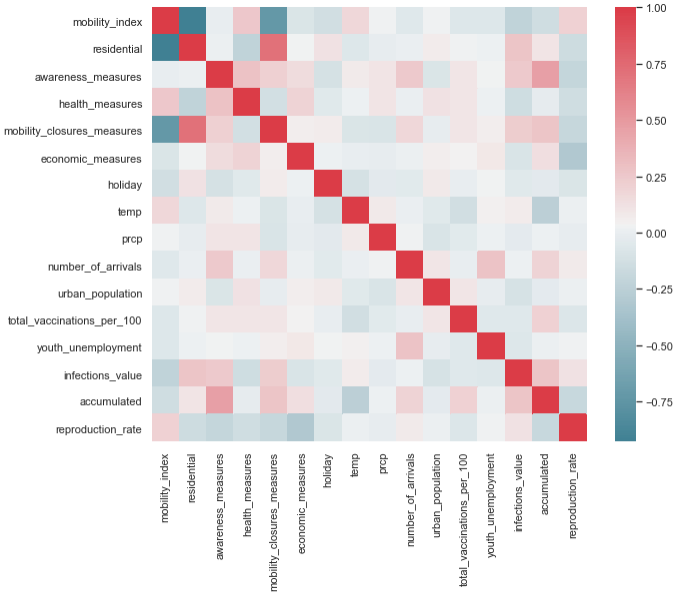


Those set of variables will be grouped under one single group since they have high correlation:

- Variables grouped as **mobility_index** are variables that have a positive relation in the spread of the virus. Those variables are "retail_and_recreation", "transit_stations", "grocery_and_pharmacy", "workplaces". They have been ponderated to an approximation of more or less effect on the response variable.
- Variables grouped as **awareness_measures** are variables that represent the conciency and awareness given from governments to the population, most meaningful one is "facial_coverings".
- Variables grouped as **economic_measures** are measures that support debt and income, which prevents workers from going to work and thus decreasing mobility and activity.
- Variables grouped as **health_measures** are measures that support tracing and testing of cases.
- Variables grouped as **mobility_measures_and_closure_measures** are measures that limit mobility and interactions between individuals of different households and cities. They also include measures that limit usual work and school life together with events.

There are some other variables such as: 'residential', 'holiday', 'temp', 'prcp', 'number_of_arrivals', 'urban_population','total_vaccinations_per_100', 'youth_unemployment', 'infections_value', 'accumulated' are not modified nor grouped since they appear to be independent from each other and it effect on the response variable. Once this transformation has been done, we get the following correlation matrix.

As it can be visualized in the image, we have removed most of correlation between features, however we are left with some correlations.



- Awareness measures seems to be highly correlated with closure measures, however it is of interest to evaluate the effect of them separately, since ones represent closures of schools, workplaces and public events, while the others represent restrictions in mobility and stay at home requirements.
- Mobility index is negatively correlated with residential.
- Infectious value is positively correlated with accumulated.
- Awareness measures are positively correlated with accumulated.

All approaches mentioned will be evaluated for both methodologies mentioned no lags and lags methodologies.

In this section we have detailed every field in our dataset, and we have also detailed three approaches taken for our dataset **(1)** raw data **(2)** grouping variables **(3)** PCA for both methodologies used **(a)** weekly estimation including mobility with no lags and **(b)** 2-week estimation including lag variable of reproduction rate. In the next section we will detail the modelling and evaluation phase for both methodologies over the three different approaches.

5. Data Modelling and Evaluation

Once Data Preparation, and Data Exploration phases have been covered, we are now able to understand and explain, the Data Modelling and Evaluation phase. In this section we will detail modelling and results over both methodologies covered and compare them. In order to do this we will first state our approach on evaluation of models and also asses the r2 score issue encountered for this specific type of evaluation found, which is shared by both methodologies.

5.1 Modelling and Evaluation Process

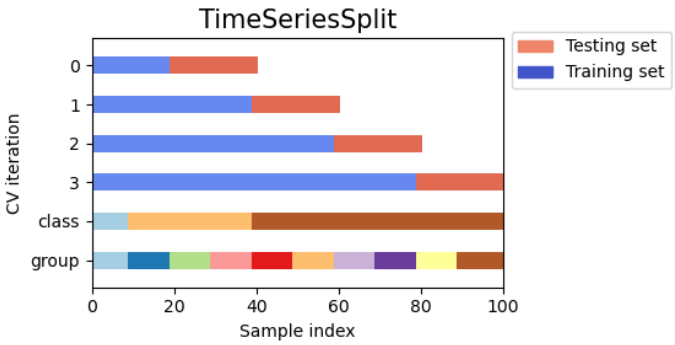
As said, we will now explain and detail the structure followed to modellize and evaluate our models for both methodologies.

5.1.1 Train Test Split Approach

Once the three approaches mentioned have been stated let's explain how we will advance in the following phase for both methodologies. We have to firstly analyze and test all approaches with all models selected in order to select an approach. Once an approach has been selected a more in-depth analysis will be performed in order to select the best fit model with the selected approach.

After various iterations and processing of the models included, it has been stated that Approach 2 gets good results overall in relation with the other two approaches and offers a better explain ability in terms of feature sign and impact evaluation. This is due to the already assessed issue experienced with multicollinearity. Grouping features and reducing number of features allows us to better interpret the models. However, we will still evaluate and explore three model approaches to make our analysis and evaluation more precise overall.

Before we start testing and applying models to our dataset of three approaches, we need to define our test split definition. It has been conducted that the best approach to test and evaluate our models due to the particularity of our dataset is to use time series split image extracted from (scikit-learn.org).



However, since we have three different approaches, making such evaluation with all approaches and model could be an inconvenient in terms of understating and complexity of the validation. Taking this into account the approach has been the following.

- **Phase 1:** Evaluate best model approach with one single time series split year 2020 for train and year 2021 for test. This will be an evaluation upon all approaches, in order to select best approach. After various iterance on modelling, as mentioned, it has been detected that approach number 2 is the one that provides better specially in terms of explicability of the models due to multicollinearity between variables. Therefore, after all Approach 2 has been selected in order to perform in detail analysis to select best model based on metrics and SHAP values. However, we will still view how models respond to different dataset approaches.

- **Phase 2:** Evaluate every model in depth with the selected approach and select best models for our approach. Overview of results of approach 2 with selected models in order to perform in detail analysis in phase 3, therefore, best models obtained in this phase will be used in phase 3.

- **Phase 3:** Perform in detail time series split (multiple splits) with selected approach and narrowed filtered models from phase 2. Select best model for selected approach and evaluate results.

Once we have defined the three phases followed, we can now assess the issue found with r2 score.

5.1.2 Assessing the R2 mismatch between train and test

As it can be appreciated in the predictions for the initial split of 2020 (train) 2021 (test) in the notebooks for both methodologies it can be seen that R2 score highly differs for all three approaches in all models. However, there is no sign of overfitting. So, in order to understand this, we will try to understand the R2 formula first and secondly evaluate what is happening without train vs our dataset.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

As seen on the right term of R2 there is a coefficient which in the denominator contains the equivalent to the sum of mae squared and on the denominator, it contains the sum of the variance squared, the variance is divided by N, but this is a fixed term therefore it does not imply changes in the following analysis.
The main idea is that if the variance in both test and train are the same, then any changes in the value of R2 will mean that there is a difference in the predictions in train and test. In case of the mae been equivalent, therefore the predictions staying the same, it will be due to a difference in the variance of train and test sets, which is our specific case. **R2 score is so different in train and test because the variation of values in train and test are widely different**.

In order to explain this lets just look at variance in test and train for the previous evaluation:
Train set: 0.17225297677258844 variation
Test set 0.029528047387486104 variation

As seen, variance in the trainset is 0.172 and variance in the test set is 0.029. This is a significant difference actually, it is 0.17/0.029=**5.86 times greater**

And this is the cause why our R2 score is so different in train and test, and this will be more appreciated in the later analysis when performing folds.
The reason behind this is that **governments are able to apply measures and restrictions in order to fit the desired reproduction rate of 1**, which by looking at the data seems to be the intention of governemnts worldwide. That has been appreciated in the plots as well, values **ranging from (0.0, 4)** for 2021 and values **ranging from (0.5, 1.5) in 2021**

We will assess the mentioned result and evaluation in phase 3 for both methodologies.

5.2 Modeling 1: No Lag Methodology

In this Modelling phase 1 our goal is to include all variables but reproduction rate of week n-2. The reason for this is that variable has of course a lot of weight in the prediction, since it carries lots of information on how the virus is spreading in a certain situation. Therefore, when making predictions and when later evaluating the results of the SHAP values, it has been appreciated how features selected affect our target variable, and we are able to better understand the effect of our features, due to the fact that their importance is higher when reproduction rate of week n-2 is not present. It will be detailed below the resume of evaluation for methodology 1 with three different approaches, however to find full explanation and detail, please refer to the notebook found on GitHub.

5.2.1 Modelling and Evaluation for Defined Phases

The models selected for modelling are XGB Regressor, LGBM Regressor, Gradient boosting Regressor, KNN Regressor, Histogram Gradient Boosting Regressor and NuSVR Regressor.

5.2.1.1 Phase 1

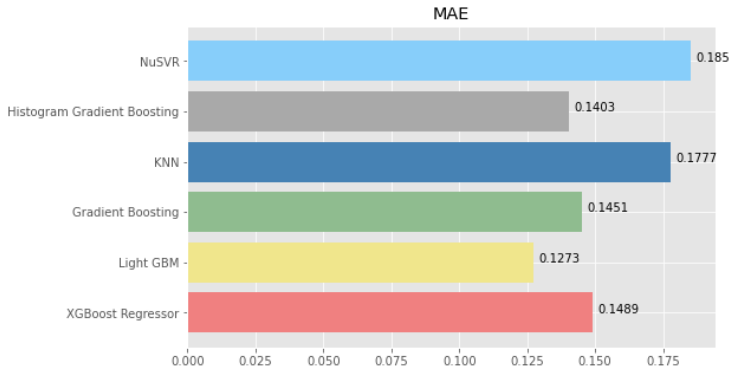
We apply models to three approaches and compare results between train and test set in order to prevent overfitting. For more in detail results of every model, showing graphs and metrics, please refer to the notebook. Below can be seen results for all metrics included for all models selected.

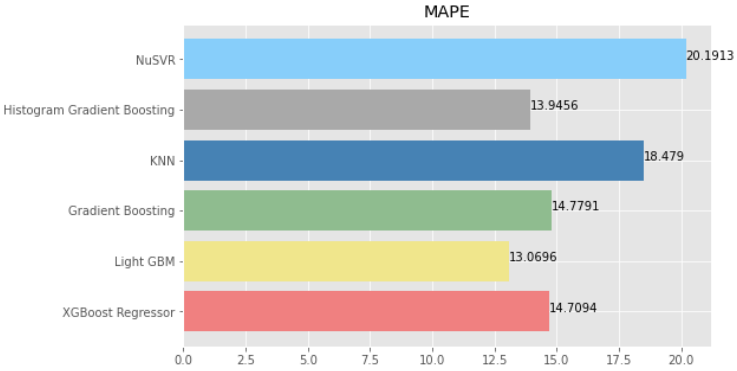
Conclusions Phase 1:

- In conclusion, approaches 1 and 2 seem to be getting simillar results.
- Approach 2 seems to be getting best results in terms of metrics and plot, however it offers no explanability at all, therefore it will not be taken into account.
- There are no signs of overfitting based on the metrics and residual plots visualized.
- Boosting models (XGB, LGBM, Gradient boosting and Histogram Gradient Boosting) seem to be getting the best results based on the fits and metrics observed.

5.3.2 Phase 2

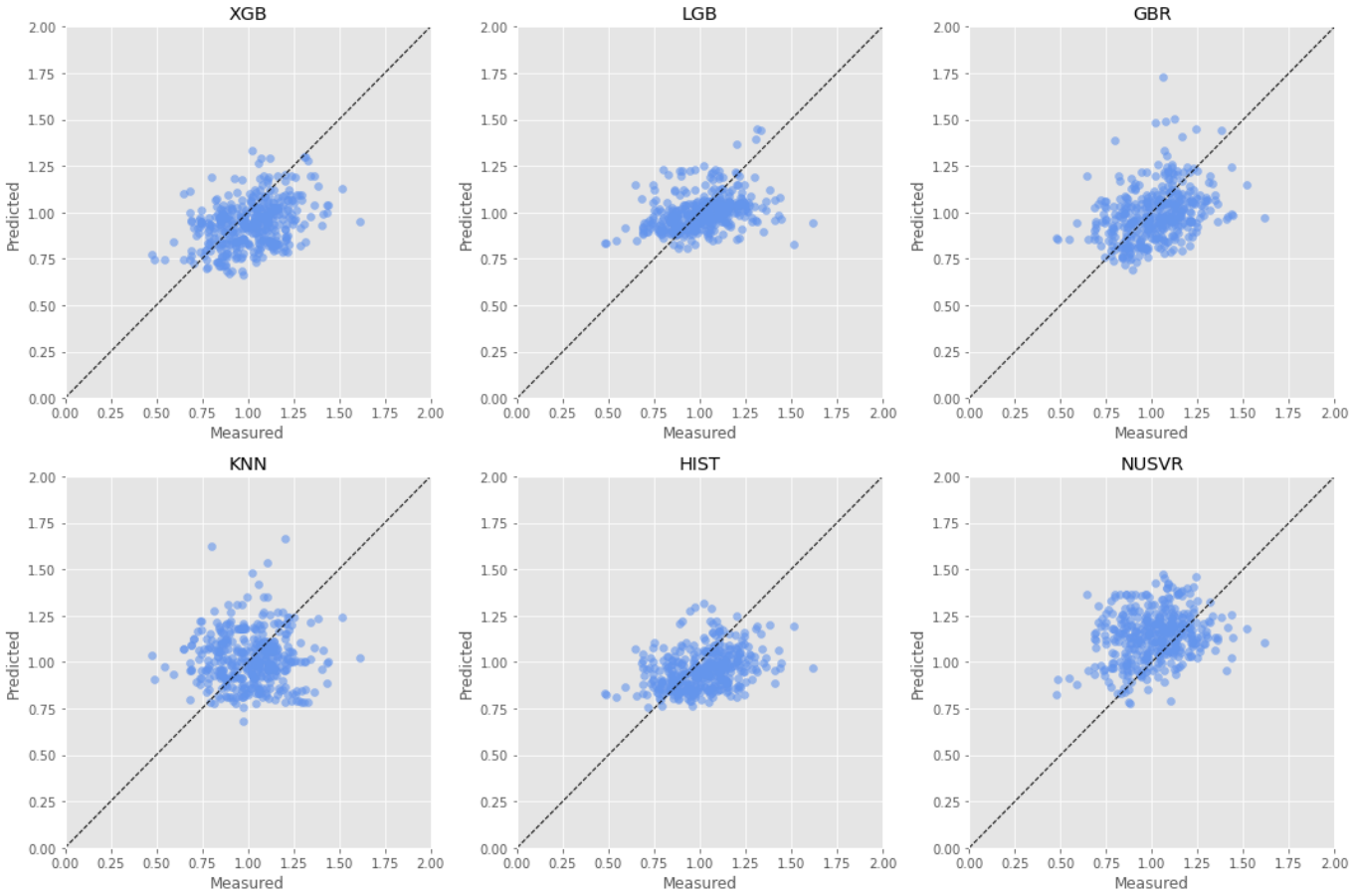
Once it has been stated in phase 1 that approach 2 is going to be chosen, we will now select best models for selected approach. In order to select best models, we will firstly look at values o selected metrics and secondly view plots fit for every model. For the presented resume we will show the main two metrics used in this project mae and mape for all six models for Approach 2.





Based on the metrics shown it can be seen taht XGB, LGBM, Gradient boosting and Histogram Gradient Boosting are getting the best results, we will perform in detail analysis to them. KNN and NuSVR are getting the worst mae and mape scores among all models.

Now we will inspect and analyze scatter plots of predictions versus measured for all six models



As seen in the image below, models KNN and NuSVR also seem to be having bad fits, while in the other hand models XGB, LGBM, Gradient boosting and Histogram Gradient Boosting appear to be making best fits of the measured versus predicted line.

Conclusions Phase 2:

- It can be seen that the best approaches seem the ones obtained by the metrics (XGB, LGBM, Gradient boosting and Histogram Gradient Boosting). We will perform time series evaluation on the selected models. We will perform 6 splits, with the selected top models, and compare them.

Over the evaluation of the presented models with all three approaches there was an issue encountered. All metrics were the same in both train and test, however a huge difference between R2 scores was found in train and test. Before going into phase 3, we will explain and explore the presented issue.

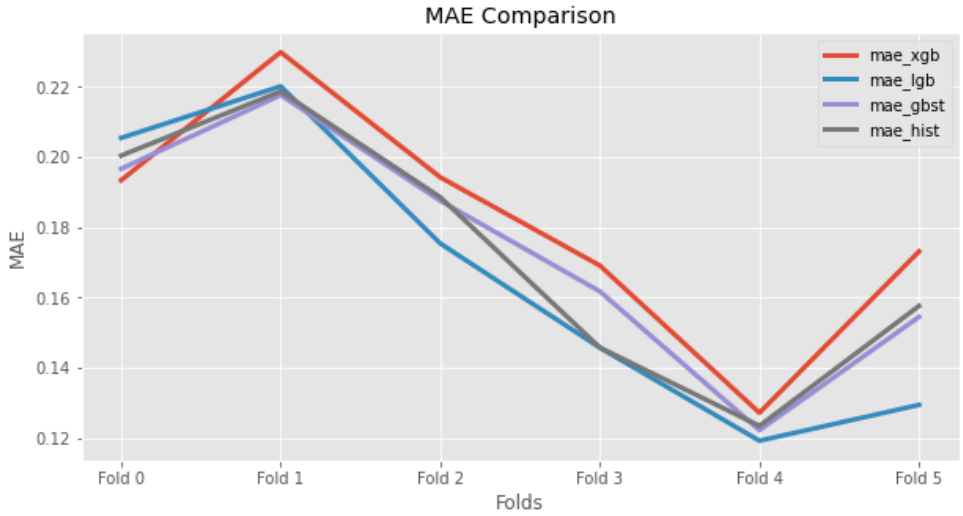
5.3.3 Phase 3

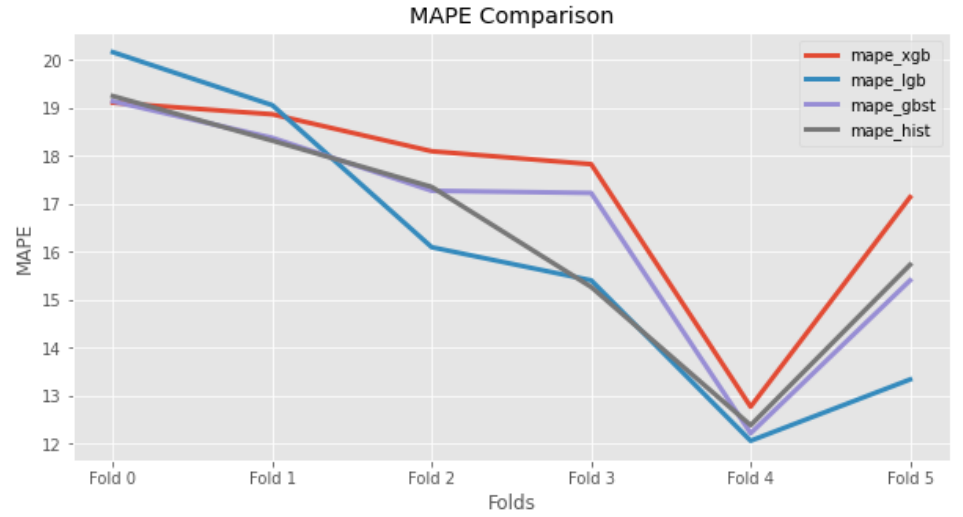
From phases 1 and 2 we have concluded that we will perform in depth analysis using time series split for approach 2 and models (XGB, LGBM, Gradient boosting and Histogram Gradient Boosting). Once the mismatch in R2 has been stated, let´s look at our evaluation of each model based on the time series train test split.

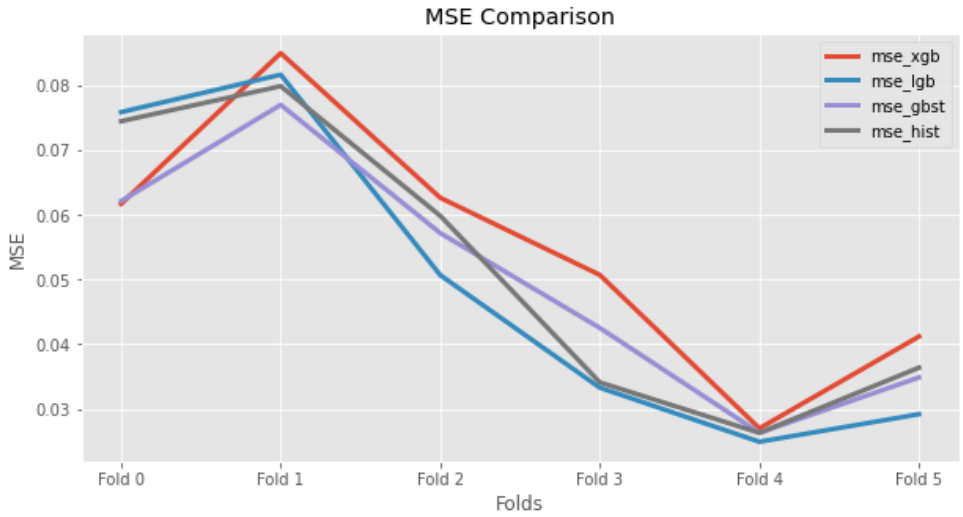
Based on observed testing and iterance it has been decided to evaluate it as time series split wise but with a little modification, based on the referred article (https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8)

Traditional time series split provided by sklearn starts with a small portion of the dataset**,** therefore, the first iterations in our folding have some bad results, since the model needs a minimum amount of data in order to extract meaningful estimated values. Therefore, we have created a custom auxiliary function *time_test_split_custom(model)* that given a model perform time series evaluation from a start date determined. In our case it is the first 1000 rows which is equivalent to September 2020, therefore minimum training dataset contains data from March 2020 till September 2020, and such dataset start being larger in every fold of the 6 folds performed.

Once all folds have been performed over models XGB, LGBM, Gradient boosting and Histogram Gradient Boosting, Mse, Mae and Mape results over 6 folds are shown below. For more in detail analysis of each of the models, please refer to the notebook.







- All models seem to be performing better as number of folds increase, except last fold were all models get worse scores than on previous fold.
- Mae, mse and mape seem to be getting the best results in last three iterations.
- As represented in the plot and table shown, all models seem to be getting similar results, however **LGBM regressor**, seems to be getting similar or better results of models in almost all folds, specially in fold 5, for both metrics mae and mape.

Let's take a look at plot fold results for LGBM Regressor on last Fold which is the latest reproduction rate estimation in time.

Conclusions Phase 3:

The best model for Approach 2 is LGBM Regressor, we then decide to select LGBM model with data configuration Approach 2 among all approaches and models, for our SHAP values evaluation and our front-end interaction with the model.

5.4 Modelling and Evaluation Conclusions

In the Modelling and Evaluation phase we have covered three different approaches:

- **Approach 1:** Raw data no grouping of variables or PCA Analysis
- **Approach 2:** Grouping variables, decreasing dimensionality and correlation
- **Approach 3:** PCA

- All approaches have been evaluated with 6 different models: **XGB Regressor, LGBM Regressor, KNN, NuSVR, Histogram Gradient Regressor and Gradient Boosting Regressor**.

- Over all approaches, Approach 2 was selected since it was getting the best results. Over all 6 models, Boosting models (XGB Regressor, LGBM Regressor, Histogram Gradient Regressor and Gradient Boosting Regressor) were the ones fitting and estimating best our target variable.

- A more in detail analysis was performed using time series split with Boosting Models and approach 2. Once this last analysis was performed it was evaluated that LGBM Regressor was the model getting the most accurate results.

- Overall, this model seems to be getting good metrics for mae, mse, rmse, and mape. However, we have the opened issue detailed for R2 score. As mentioned, this is due to the fact that test sets tend to have less variance and therefore resulting in worst scores of R2. Predictably the root cause behind this is that governments tend to apply restrictions to reach the desired reproduction rate of 1, and all values passed the first wave (March-April 2020) seem to be pointing in that direction.

- It can be affirmed that our model does not interpret or is able to estimate correctly the variance when the test set has low variance, getting even negative results or close to zero. Therefore, in terms of forecast and if only if the test set has low variance, a model that always predict the mean value would return better results than the presented model. However, the presented model does perform better when a test dataset with high variance is predicted. Such model of course offers a better estimation than the mean, since features offer information and are able to interpret changes in mobility, restrictions as it will be

detailed in the shap model evaluation notebook. Therefore, our model does recognize and is able to correctly estimate values with the metrics shown.

- In conclusion this model offers a good interpretation and estimation of the target variable when there is high variance in the test set giving parametric relations that explain the behavior of the target variable. When low variance is encountered our models does offer a good functional relation, meaning it offers information on the positive or negative centered in the sign of the features included than in the parameter value itself.

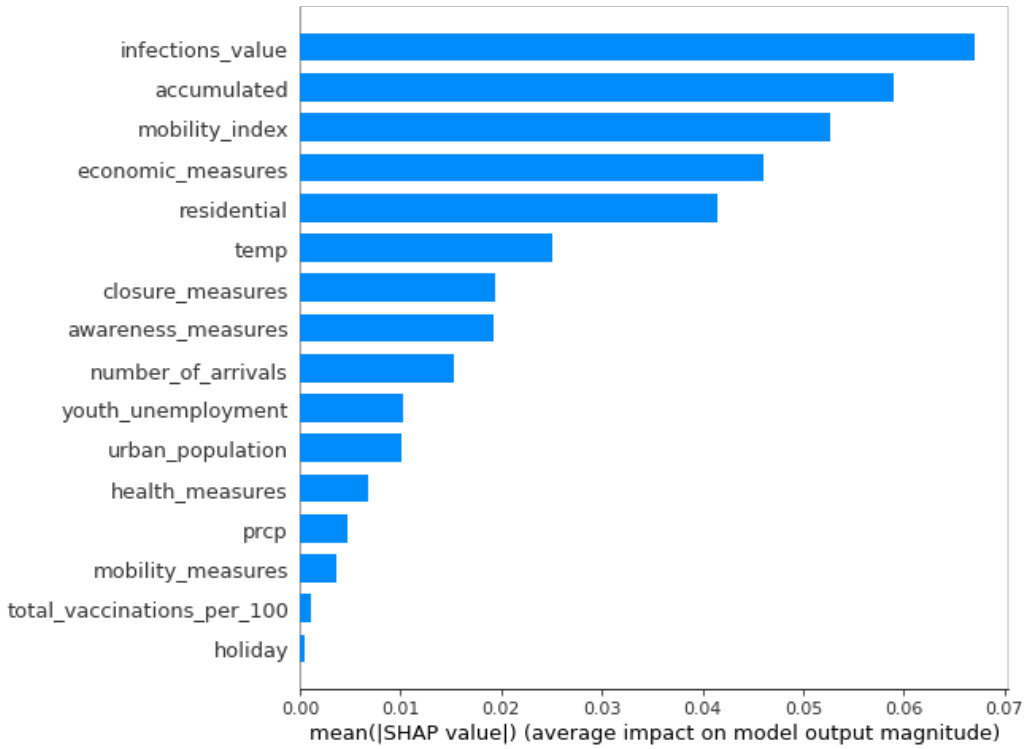
5.4 Modelling 2: Lags Methodology

6. Model Features Evaluation with SHAP

Once an approach and model has been selected, we will now try to understand what the model is doing and what features included in the analysis are the most important, and even more what if the value of the feature has positive or negative impact on our target variable reproduction rate. In order to do that we will take advantage of SHAP library, which reflects feature importance and sing of each feature in relation with our target variable. In order to do this, we will perform evaluation and analysis of the impact of variables on the training set predictions (year 2020).

6.1 Tree Explainer Bar Plot

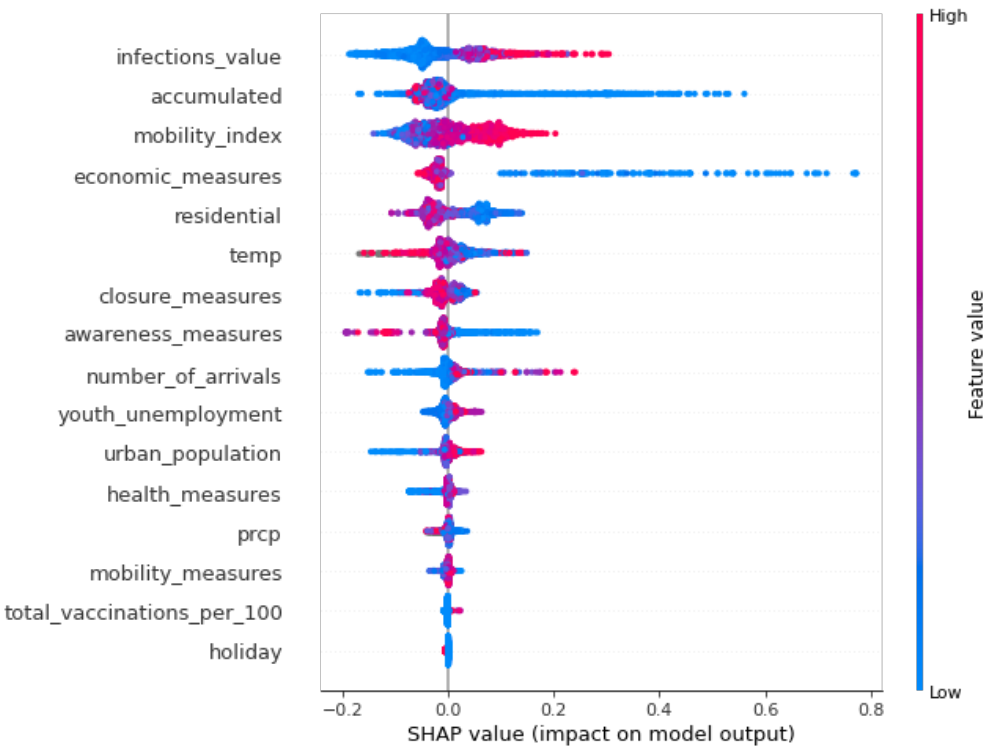
We will plot our feature importance bar plot; this will represent the importance of each feature in terms of predictions of our target variable reproduction_rate.



- As seen variables that seem to have a higher impact on our target variable reproduction rate are:
  - **infections_value** which represents the amount of current people infecting
  - **accumulated** which represents the amount of people that have already contracted the virus
  - **residential** which represents the amount of time people spend time at home

    - **economic_measures** which represents the economic support given by governments in terms of debt relief and income support
    - **mobility_index** which represents the mobility increase or decrease in reatil and recreation, transit stations, groceries, pharmacy and workplaces
- Other variables that have smaller coefficients are **temperature, number of arrivals, closure measures, awareness measures and urban population**

6.2 Tree Explainer Plot



We will perform a resume of the variables and its impact on our target variable reproduction rate, sorted by importance in our model.

1. **infections_value** which represents the amount of current people infecting.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate.

2. **accumulated** which represents the amount of current people infecting.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate.

3. **mobility_index** which represents the mobility increase or decrease in retail and recreation, transit stations, groceries, pharmacy and workplaces.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate.

4. **economic_measures** which represents the economic support given by governments in terms of debt relief and income support.

    - High values have negative impact on reproduction rate.
    - Low values have positive impact on reproduction rate.

5. **residential** which represents the amount of time people spend time at home.

    - High values have negative impact on reproduction rate.
    - Low values have positive impact on reproduction rate.

6. **temp** which represents the average temperature in the deferred week.

    - High values have negative impact on reproduction rate.
    - Low values have positive impact on reproduction rate.

7. **closure_measures** which are the restrictions in closings applied by governments.

    - High values have negative impact on reproduction rate.
    - Low values have positive impact on reproduction rate. However, there are some very low values related with negative impact. Will analyze it more in detail later on.

8. **number_of_arrivals** which is the usual number of tourists received but that represents countries with high or small mobility worldwide.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate.

9. **youth_unemployment** which is the percentage of youth unemployed in the referenced country.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate.

10. **awareness_measures** which are variables that represent the conciency and awareness given from governments to the population, most meaningful one is facial_coverings.

    - High values have negative impact on reproduction rate.
    - Low values have positive impact on reproduction rate. However, there are some very low values related with negative impact. Will analyze it more in detail later on.

11. **urban_population** which is an indicator of the percentage of people that live in urban areas and that tries to simulate population density.

    - High values have positive impact on reproduction rate.
    - Low values have negative impact on reproduction rate. However, there are some very low values related with negative impact. Will analyze it more in detail later on.

- The rest of the coefficients seem to have little influence in the target variable and the coefficients and importance of those variable may be absorbed by some other variables previously mentioned.

12. **health_measures** and **mobility_measures** seem to be getting confusing shap values, will analyze those in detail taking usage of single variables dependence plots analysis.

However, based on the graph it can be estimated that:

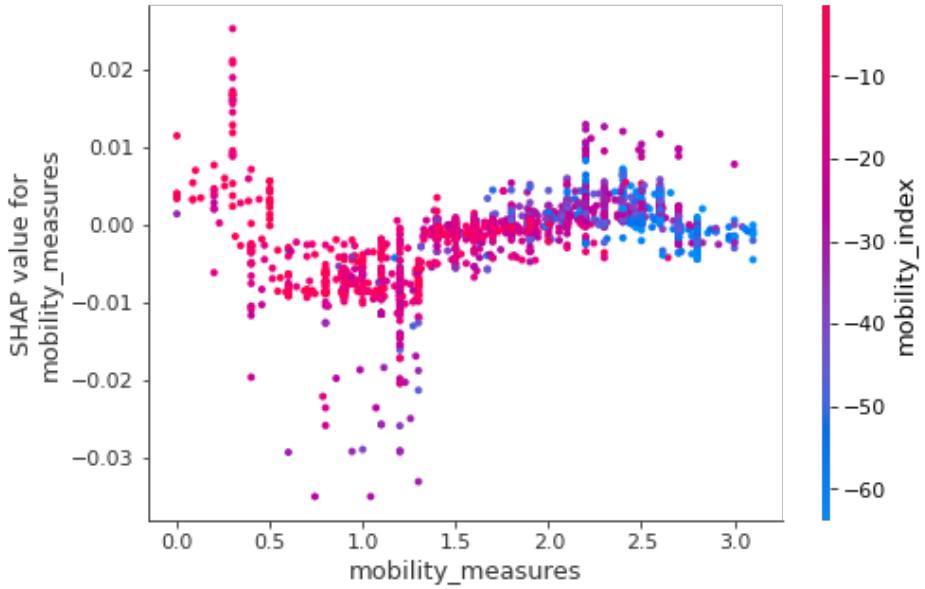
- **precipitations** are related with reproduction rate negatively.
- **holiday** is related with reproduction rate positively on holidays with high values, however low values of holiday don't seem to have impact on the response variable.

Some other variables may have such coefficients due to multicollinearity with other variables, based on the correlation matrix shown below. In order to detect this multicollinearity issues will

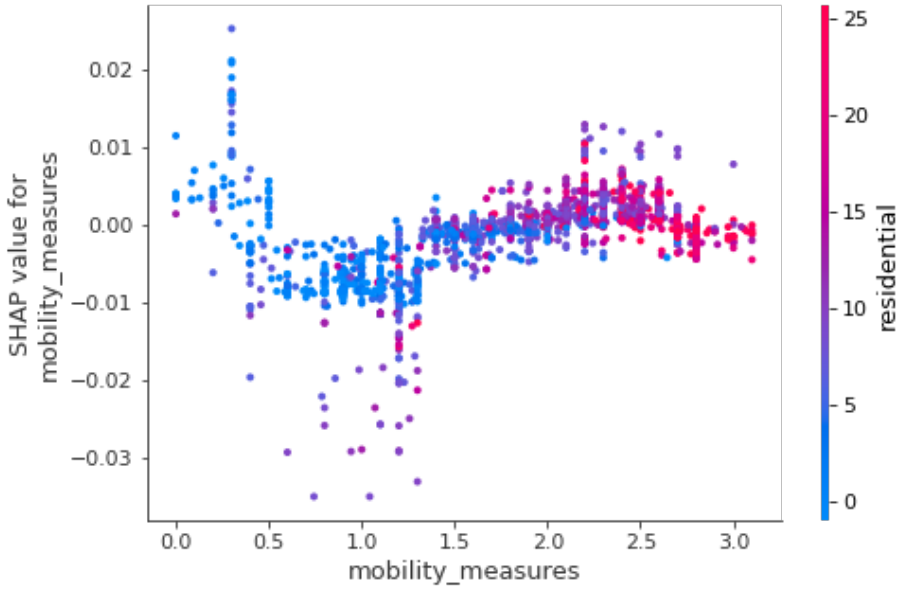
analyze each of these features separately and its possible relation with some other more dominant features.

6.3 Single Variables Dependence Plots Analysis

mobility_measures



As seen mobility_measures have a relation with mobility index, therefore on the previous plots analyzed the impact of this variable is not really positive, it is just that variable mobility_index has captured its effect. This can be seen in the graph, low values of mobility_measures are related with high values in mobility_index which in fact affect the positive increase of reproduction rate. High values of mobility measures are related with low values of mobility_index and therefore negative impact on the response variable.



It can also be seen how high values of mobility measures are related with high values in residential index, therefore, even though this variable is not being reflected as important in the SHAP analysis, it does have relation with two of the main features, mobility index and residential.

health_measures

This variable doesn't seem to have relation with any other feature in the dataset. This variable reflects government measures on testing and tracing cases. The referenced variable does not appear to have a big impact on the target variable, which is something confusing, however, no possible explanation has been found.

However our target variable in the end is a very precise estimation taken from a reliable source as detailed in memory.pdf. And such variable is calculated i relation with number of positive cases among a lot other variablesand complex calculations, maybe an increasing number of testing and contact tracing is related in this case with the sign of the referenced variable.

total_vaccinations_per_100

This variable is of course predicticably the variable that should predict the most together with accumulated and infections, in fact it should have the same effect as accumulated, however since there is very few data collected for train of the model since vaccinations started in 2021 approximately, the model is not including it as an important feature, however in future analysis with more amount of data this variable will predictably have the same weight as accumulated.

prcp

Precipitations don't seem to have no relation with the mobility index netiher residential. However it does seem to have negative impact on the target variable base on tree explainer plot meaning, high values are related with low values of reproduction rate.

6.4 SHAP Values Evaluation Conclusion

The model mostly predicts reproduction rate based on top variables mentioned

- Our World in Data
    - https://ourworldindata.org/excess-mortality-covid
    - https://ourworldindata.org/policy-responses-covid
    - https://ourworldindata.org/grapher/international-tourism-number-of-arrivals
    - https://ourworldindata.org/urbanization
    -
- World Bank of Data
    - https://data.worldbank.org/indicator/SH.MED.NUMW.P3
    - https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS
    - https://data.worldbank.org/indicator/SH.MED.BEDS.ZS

- National Oceanic and Atmospheric Administration (NOAA) Us department of commerce. (Accesed using Google Big Query)
    - https://www.noaa.gov/