

# **COVID-19 Reproduction Rate Estimator**

**Master in Data Science**

**23<sup>rd</sup> Edition - KSchool**

Madrid, May 2021

Lucas Botella Roca

# Table of Contents

<b>1. Introduction .....</b>	<b>3</b>
<b>2. Problem Statement and Modelling .....</b>	<b>3</b>
<b>3. Data Gathering and Preparation .....</b>	<b>6</b>
<b>3.1 Data Gathering and Preparation Explanation .....</b>	<b>6</b>
<b>4. Data Exploration.....</b>	<b>8</b>
<b>4.1 Dataset Description Detail.....</b>	<b>8</b>
<b>4.2 Data Categorization.....</b>	<b>12</b>
<b>4.3 Data Grouping and Multi Collinearity .....</b>	<b>13</b>
<b>5. Data Modelling and Evaluation .....</b>	<b>15</b>
<b>5.1 Modelling and Evaluation Process .....</b>	<b>15</b>
5.1.1 Train Test Split Approach.....	15
5.1.2 Assessing the R <sup>2</sup> mismatch between train and test .....	17
<b>5.2 Modelling 1: No Lag Methodology .....</b>	<b>18</b>
5.2.1 Modelling and Evaluation for Defined Phases.....	18
5.2.1 Conclusions Methodology 1 .....	22
<b>5.3 Modelling 2: Lag Methodology .....</b>	<b>23</b>
5.3.1 Modelling and Evaluation for Defined Phases.....	23
5.2.1 Conclusions Methodology 2 .....	27
<b>6. Model Features Evaluation with SHAP .....</b>	<b>28</b>
<b>6.1 SHAP Evaluation Modelling 1: No Lag Methodology .....</b>	<b>28</b>
Tree Explainer Plot.....	28
<b>6.2 SHAP Evaluation Modelling 2: Lag Methodology.....</b>	<b>31</b>
Tree Explainer Plot.....	32
<b>Conclusion .....</b>	<b>32</b>
<b>7. Frontend Applications with Streamlit.....</b>	<b>33</b>
<b>8. Final Results and Conclusion .....</b>	<b>34</b>
<b>Annex: Data Sources .....</b>	<b>37</b>

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

GitHub repo link: [https://github.com/lucasbotellaroca/Covid-19\\_Reproduction\\_Rate\\_Estimator](https://github.com/lucasbotellaroca/Covid-19_Reproduction_Rate_Estimator)

## 1. Introduction

The intention of the presented project is to estimate and understand how restrictions, mobility trends, temperature, demographic and special characteristics of each country affect the spread of Covid-19.

In order to achieve this, we will try to predict the evolution and changes of the reproduction rate index for a set of 61 countries with data relative from March 2020 till April 2021. The value of reproduction rate has been retrieved from [Our World In Data](#) and the detail of how such value was calculated is detailed in [Arroyo Marioli et al. \(2020\)](#), Central Bank of Chile and Humboldt University of Berlin. In general terms, reproduction rate estimates in average how many infections may be caused by one infected individual.

Some approaches and state of art best projects which have been taken into account are: [Interpretable Sequence Learning for COVID-19 Forecasting Google](#), [COVID-19 Forecasts: Deaths, Centers for Disease Control and Prevention](#), [Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter](#), [Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries](#).

Our approach for the presented problem is based on [SIR Model](#) which is a standard used in epidemiology for disease spread in the population, but we will include the effect and impact of mentioned features in order to understand how the spread of the virus behaves on different situations.

The following project will follow a complete CRISP-DM cycle. Business Understanding (Problem Statement and Modelling), Data Understanding (Problem Statement and Modelling, Data Exploration), Data preparation (Data Gathering and Preparation), Modeling and Evaluation (Data Modelling and Evaluation) and Deployment (Frontend Applications with Streamlit)

## 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, two models will be created. We will refer to effective reproduction number or reproduction rate as  $R_t$ .  $R_t$  tells you the average number of people who will contract a contagious disease from one person with that disease in our case Covid-19. The basic reproduction number or  $R_0$  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  $R_t$ .

In general terms, as mentioned,  $R_t$  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  $R_t < 1$  one infected person will cause less than one infection. In this case the disease will eventually die out.
- If  $R_t > 1$  one infected person will cause more than one infection. In this case the disease will increase and eventually cause an outbreak or pandemic.
- If  $R_t = 1$  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  $t$ .  $\beta_t$  is the transmission rate, and  $\gamma_t$  is the transition rate from infected to recovered in time  $t$ . Note the difference between  $R_t$  which we refer as effective reproduction number and  $R_t$  which we refer to as recovered individuals in time  $t$ .

Let's note that  $N = S_t + I_t + R_t$ . The original SIR problem is stated as shown below:

$$\begin{aligned} S_t &= S_{t-1} - \beta_t I_{t-1} S_{t-1} \frac{S_{t-1}}{N} \\ I_t &= I_{t-1} - \beta_t I_{t-1} S_{t-1} \frac{S_{t-1}}{N} - \gamma I_{t-1} \\ R_t &= R_{t-1} + \gamma I_{t-1} \end{aligned}$$

To simplify things,  $R_0$  is defined as for whatever defined time period as  $R_0 = \beta/\gamma$ .  $R_t$  is defined as shown in the equation below. It is referred to as the number of individuals infected in time  $t$ .

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

$R_t$  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  $R_t$  is dependent of infections in the time period defined  $t$ , accumulated infections or recovered  $R_t$ , and for our specific case the restrictions applied which affects the  $\beta_t$  coefficient. If all infected individuals were isolated from the rest of the population for  $\gamma$  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  $R_t$  for model one, and 2-week period for our model two,.

Therefore, our problem is stated as shown below for Model 1 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.

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

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 Model 2, 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. Values in correlation matrix decreased from 0.9 to 0.8.

Also, this model would serve more of 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.

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

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 either, 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

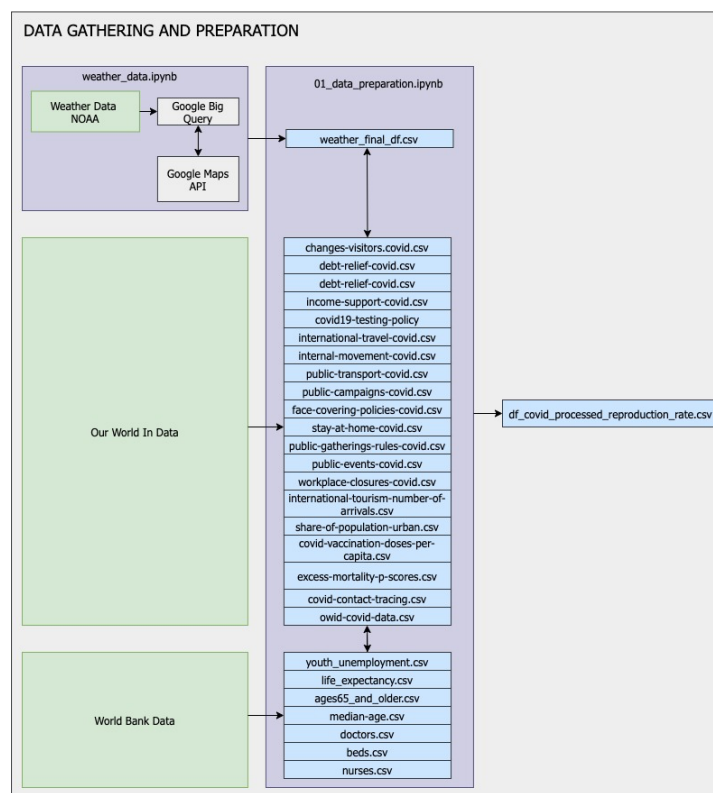
Data has been gathered form different sources (links in annex):

- 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 stage 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 x 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 and placed in one of our folders in our google drive repository. Unfortunately, this code is not possible to be replicated. 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

All files retrieved and detailed above are the merged together in notebook [01\\_data\\_preparation.ipynb](#) the logic behind all that merging of information is detailed below:

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:

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



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 trends, others and recovered together with our target variable  $R_t$ . 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 capture the dataset previously generated for no lag methodology and 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:

$$f_{2\text{-week } n}(\text{Restrictions, Others, Recovered, Infected, Reproduction Rate Lag}) = R_{t2\text{-week } n}$$

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 below.

### 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.



		<p>This index is smoothed to the rolling 7-day average.</p> <p>Range: [-100,100]</p>
residential	Float	<p>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).</p> <p>This index is smoothed to the rolling 7-day average.</p> <p>Range: [-100,100]</p>
transit_stations	Float	<p>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).</p> <p>This includes public transport hubs such as subway, bus, and train stations.</p> <p>This index is smoothed to the rolling 7-day average.</p> <p>Range: [-100,100]</p>
parks	Float	<p>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).</p> <p>This includes places like local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens.</p> <p>This index is smoothed to the rolling 7-day average.</p> <p>Range: [-100,100]</p>
workplaces	Float	<p>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).</p> <p>This index is smoothed to the rolling 7-day average.</p> <p>Range: [-100,100]</p>
contact_tracing	Integer	<p>Government policies on contract tracing for COVID-19.</p> <ul style="list-style-type: none"> <li>- No tracing - 0</li> <li>- Limited tracing (Only some cases) - 1</li> <li>- Comprehensive tracing (All cases) - 2</li> <li>-</li> </ul> <p>Range: [0,2]</p>
testing_policy	Integer	<p>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.</p> <ul style="list-style-type: none"> <li>- No testing policy- 0</li> <li>- 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</li> <li>- Testing of anyone showing COVID-19 symptoms - 2</li> <li>- Open public testing (e.g “drive through” testing available to asymptomatic people) - 3</li> </ul> <p>Range: [0,3]</p>
international_travel_controls	Integer	<p>Government policies on restrictions on international travel controls.</p> <ul style="list-style-type: none"> <li>- No measures - 0</li> <li>- Screening - 1</li> <li>- Quarantine from high-risk regions - 2</li> <li>- Ban on high-risk regions - 3</li> <li>- Total border closure - 4</li> <li>-</li> </ul> <p>Range: [0,4]</p>
restrictions_internal_movements	Integer	<p>Government policies on restrictions on internal movement/travel between regions and cities.</p> <ul style="list-style-type: none"> <li>- No measures - 0</li> <li>- Recommend movement restriction - 1</li> <li>- Restrict movement - 2</li> </ul> <p>Range: [0,2]</p>
close_public_transport	Integer	<p>Government policies on public transport closures</p> <ul style="list-style-type: none"> <li>- No measures - 0</li> <li>- Recommended closing (or reduce volume) - 1</li> <li>- Required closing (or prohibit most using it) - 2</li> </ul>

		Range: [0,2]
public_information_campaigns	Integer	<p>Public information campaigns on COVID-19.</p> <ul style="list-style-type: none"> <li>- None - 0</li> <li>- Public officials urging caution - 1</li> <li>- Coordinated information campaign - 2</li> </ul> <p>Range: [0,2]</p>
facial_coverings	Integer	<p>Government policies on the use of face coverings outside-of-the-home.</p> <p>Countries are grouped into five categories:</p> <ul style="list-style-type: none"> <li>- No policy - 0</li> <li>- Recommended - 1</li> <li>- Required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible - 2</li> <li>- Required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible - 3</li> <li>- Required outside the home at all times regardless of location or presence of other people - 4</li> </ul> <p>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.</p> <p>Range: [0,4]</p>
stay_home_requirements	Integer	<p>Government policies on stay-at-home requirements or household lockdowns.</p> <ul style="list-style-type: none"> <li>- No measures - 0</li> <li>- Recommended not to leave the house - 1</li> <li>- Required to not leave the house with exceptions for daily exercise, grocery shopping, and 'essential' trips - 2</li> <li>- 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</li> </ul> <p>Range: [0,3]</p>
restriction_gatherings	Integer	<p>Government policies on restrictions on public gatherings.</p> <p>Countries are grouped into five categories:</p> <ul style="list-style-type: none"> <li>- No restrictions - 0</li> <li>- Restrictions on very large gatherings (the limit is above 1000 people) - 1</li> <li>- Restrictions on gatherings between 100 to 1000 people - 2</li> <li>- Restrictions on gatherings between 10 to 100 people - 3</li> <li>- Restrictions on gatherings of less than 10 people - 4</li> </ul> <p>Range: [0,4]</p>
cancel_public_events	Integer	<p>Cancellation of public events.</p> <p>No measures - 0 Recommended cancellations - 1 Required cancellations - 2</p> <p>Range: [0,2]</p>
workplace_closures	Integer	<p>Government policies on workplaces closures.</p> <p>No measures - 0 Recommended - 1 Required for some - 2 Required for all but key workers - 3</p> <p>Range: [0,3]</p>
school_closures	Integer	<p>Government policies on school closures.</p> <p>No measures - 0 Recommended - 1 Required (only at some levels) - 2 Required (all levels) - 3</p>

		<p>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.</p> <p>Range: [0,3]</p>
debt_relief	Integer	<p>Governments provide debt or contract relief to citizens during the COVID-19 pandemic.</p> <p>No relief - 0 Narrow relief - 1 Broad relief - 2</p> <p>Range: [0,2]</p>
income_support	Integer	<p>Governments provide income support to workers during the COVID-19 pandemic.</p> <p>No income support - 0 Covers &lt;50% of lost salary - 1 Covers &gt;50% of lost salary - 2</p> <p>Range: [0,2]</p>
holiday	Integer	<p>Number of holidays in the selected time period</p> <p>Range: [0,7]</p>
temp	Float	<p>Average temperature in celsius of all stations in the selected time period.</p> <p>Range: [-20,40]</p>
precip	Float	<p>Average precipitation in mmph of all stations in the selected time period</p> <p>Range: [0,3]</p>
doctors_per_1000	Float	<p>Number of doctors per 1000 habitants last year recorded</p> <p>Not used</p>
nurses_per_1000	Float	<p>Number of nurses per 1000 habitants last year recorded</p> <p>Not used</p>
beds_per_1000	Float	<p>Number of hospital beds per 1000 habitants last year recorded</p> <p>Not used</p>
number_of_arrivals	Float	<p>Number of tourism arrivals last year recorded</p>
urban_population	Float	<p>Percentage of urban population last year recorded</p> <p>Range: [0,100]</p>
total_vaccinations_per_100	Float	<p>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.</p> <p>Range: [0,200]</p>
youth_unemployment	Float	<p>Percentage of youthment unemployment last year recorded</p> <p>Range: [0,100]</p>
life_expectancy	Float	<p>Average life expectancy at birth last year recorded</p> <p>Not used</p>
%df_population_gr_65	Float	<p>Percentage of population with age 65 or higher last year recorded</p> <p>Not used</p>
UN Population Division (Median Age) (2017)	Float	<p>Median age last year recorded</p> <p>Not used</p>
accumulated	Float	<p>Accumulated percentage of deaths.</p> <p>Range: [0,100]</p>

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. Extracted from <a href="#">Arroyo Marioli et al. (2020)</a>

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 of 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
<ul style="list-style-type: none"> <li>• retail_and_recreation</li> <li>• grocery_and_pharmacy</li> <li>• residential</li> <li>• transit_stations</li> <li>• parks</li> <li>• workplaces</li> </ul>	<ul style="list-style-type: none"> <li>• infections_value</li> <li>• accumulated</li> <li>• total_vaccinations_per_100</li> </ul>	<ul style="list-style-type: none"> <li>• temp</li> <li>• prep</li> <li>• number_of_arrivals</li> <li>• urban_population</li> <li>• youth_unemployment</li> <li>• holiday</li> </ul>	<ul style="list-style-type: none"> <li>• debt_relief</li> <li>• income_support</li> <li>• testing_policy</li> <li>• international_travel_controls</li> <li>• restrictions_internal_movements</li> <li>• close_public_transport</li> <li>• public_information_campaigns</li> <li>• facial_coverings</li> <li>• contact_tracing</li> <li>• stay_home_requirements</li> <li>• restriction_gatherings</li> <li>• cancel_public_events</li> <li>• workplace_closures</li> <li>• school_closures</li> </ul>

## 4.3 Data Grouping and Multi Collinearity

Based on the data exploration section we have conducted that lots of variables have a high correlation with each other, 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 referred problem.

- Drop variables, select meaningful features.
- Transforming variables.
- 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."*

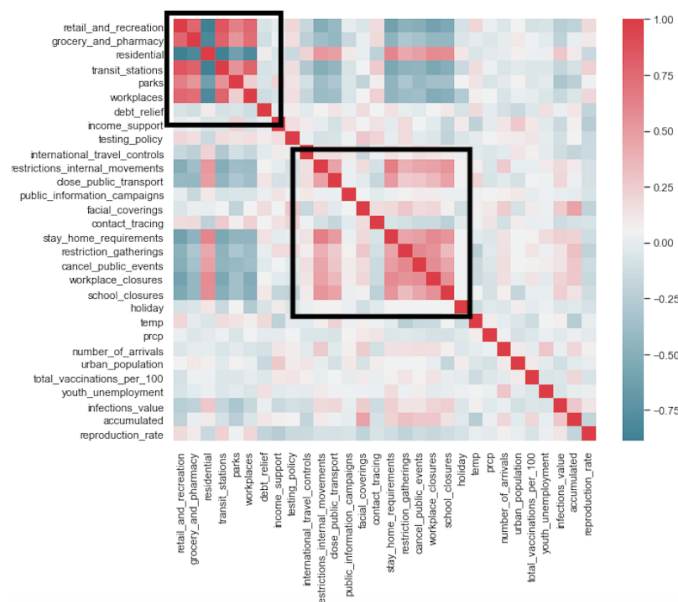
Even though when grouping variables, we might not have better predictions this might be useful in order to interpret the model afterwards, since one feature may absorb the effect of another. Hence, we will explore three approaches and observe which one best fits our purpose which as stated is not only 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 appropriate 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.

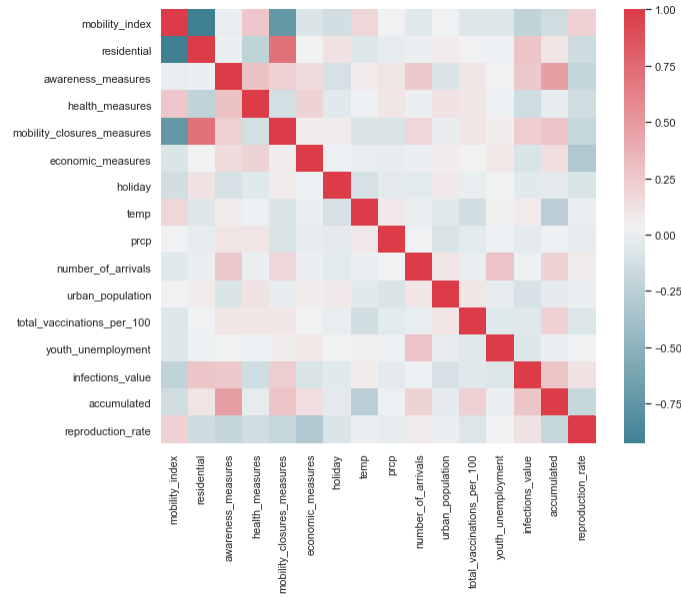


Those set of variables will be grouped under one single group since they have high correlation:

- 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 concieny 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  $r^2$  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.

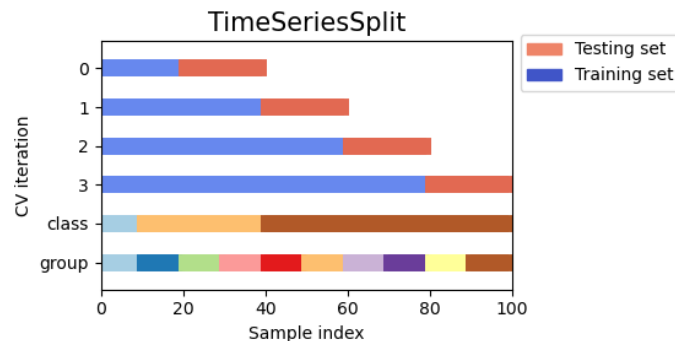
#### 5.1.1 Train Test Split Approach

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 iteration 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 (approach 2) 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 the three phases have been stated, let´s look at our way of evaluating using time series split in phase 3.

Based on observed testing and iteration 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. Such function is used for both methodologies in phase 3.

The models selected for modelling are XGB Regressor, LGBM Regressor, Gradient boosting Regressor, KNN Regressor, Histogram Gradient Boosting Regressor and NuSVR Regressor. We will focus our evaluation not only in the metrics gathered which will mainly be MAE, MSE, RMSE and MAPE but also in the sign and significance of features. In order to achieve this and retrieve good interpretability of models we will plot the SHAP values representation for every model. It is important to note that some models are not eligible for SHAP values, however they are still included since it is of interest to see our data perform in non-tree models.

Hence we not only want good metrics, but we also want a model that is able to interpret how restrictions, mobility trends, etc are able to alter and modify our response variable reproduction rate.

Evidency and intuition on whether SHAP Values seem to have good values are based on our expert knowledge in the field, which can be backed up in the following researches:

- Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter. <https://doi.org/10.2139/ssrn.3581633>
- Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries <https://spiral.imperial.ac.uk:8443/bitstream/10044/1/77731/10/2020-03-30-COVID19-Report-13.pdf>

Once we have defined the three phases followed, and what metrics and evaluation logic will be followed we can now assess the issue found with  $R^2$  score.

### 5.1.2 Assessing the $R^2$ 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  $R^2$  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  $R^2$  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  $R^2$  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  $R^2$  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.  $R^2$  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 let's 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 governments 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.

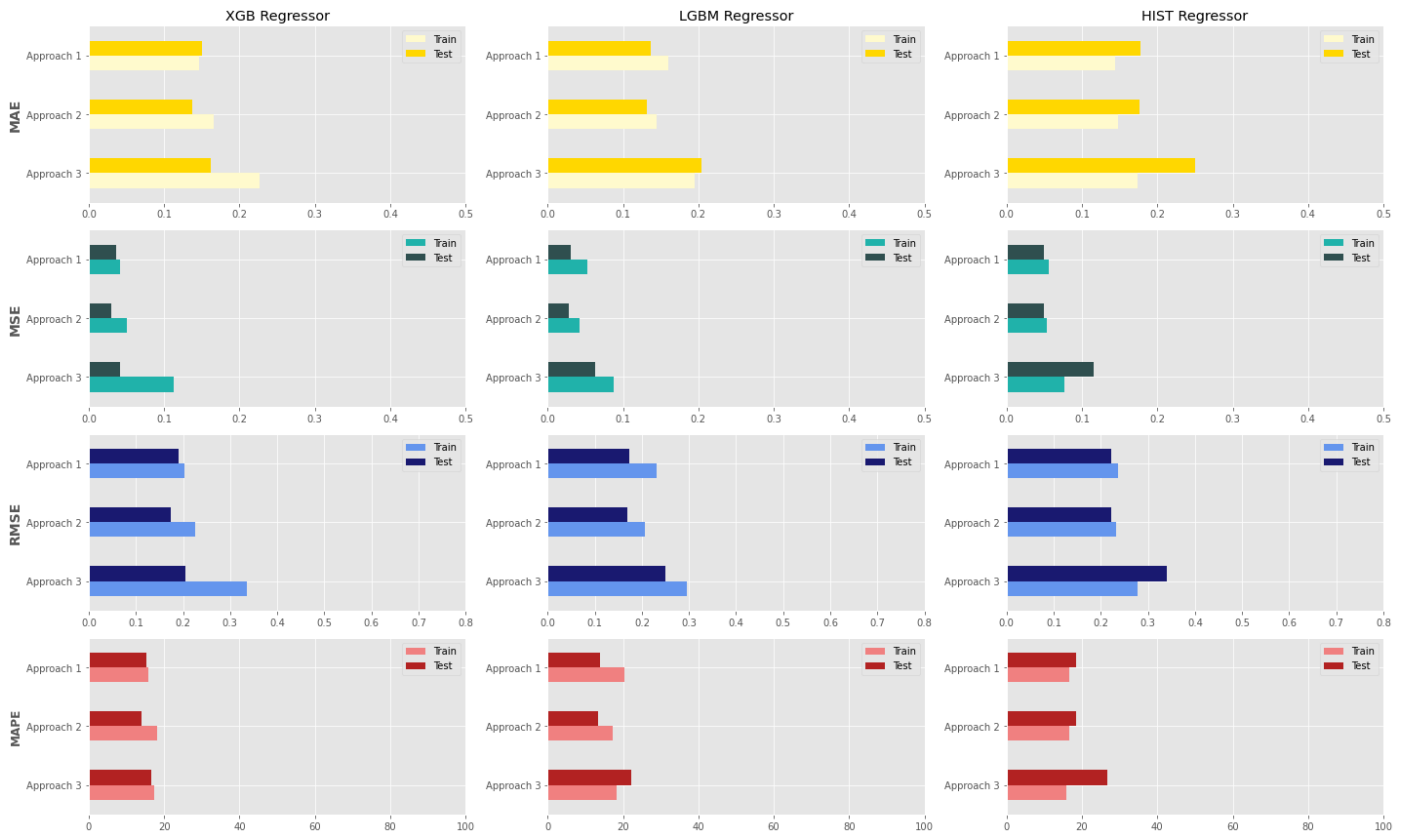
## 5.2 Modelling 1: No Lag Methodology

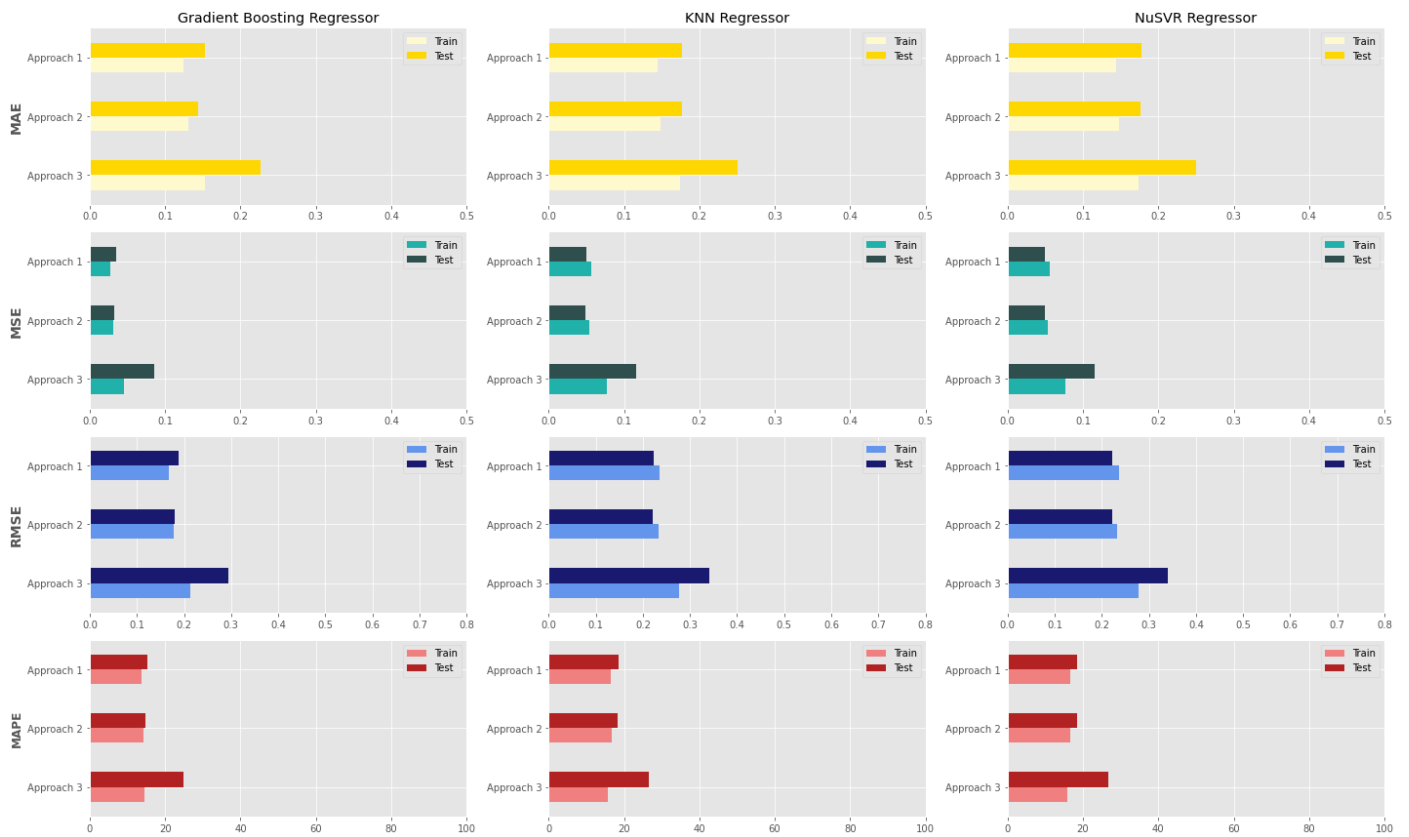
In this Modelling methodology our goal is to include all variables excluding reproduction rate of week n-2. The reason for this is that variable has of course a lot of weight in the prediction, since it carries lots of information on how the virus is spreading in a certain situation. Therefore, when 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

#### 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. Below there is a summary of all metrics for three approaches and all models.

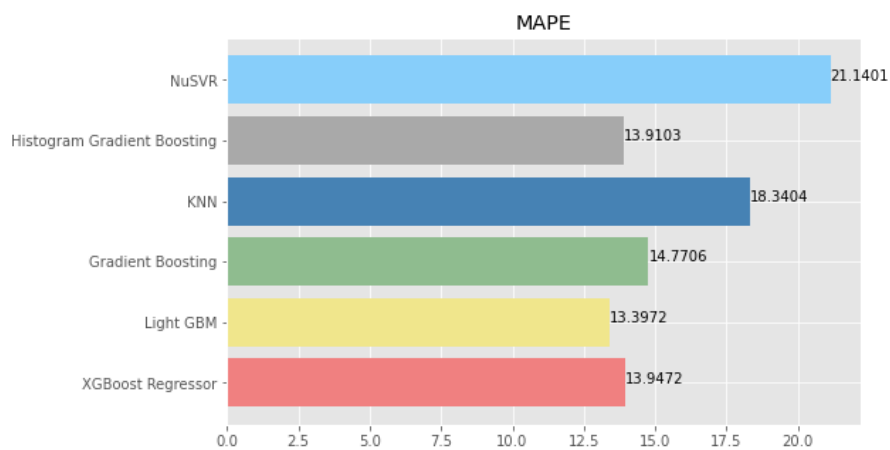


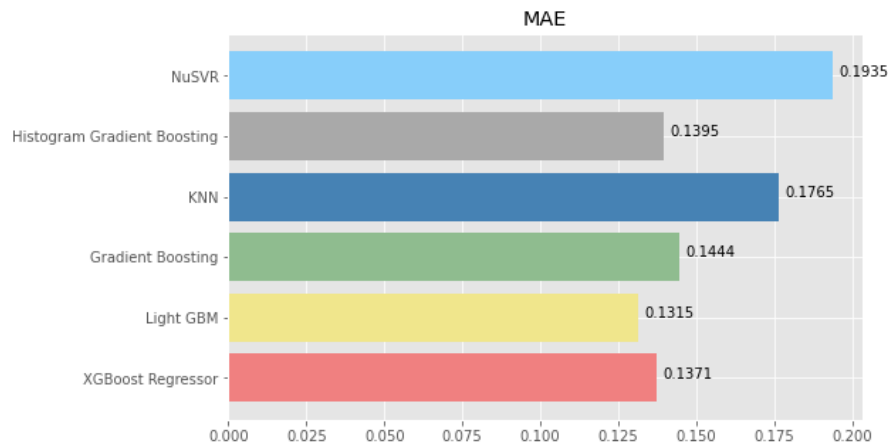


- In conclusion, approaches 1 and 2 seem to be getting similar results.
- Approach 2 seems to be getting best results in terms of metrics and explainability of shap values.
- There are no signs of overfitting based on the metrics and residual plots visualized.
- Boosting models (XGB, LGBM, Gradient boosting and Histogram Gradient Boosting) seem to be getting the best results based on the fits and metrics observed.

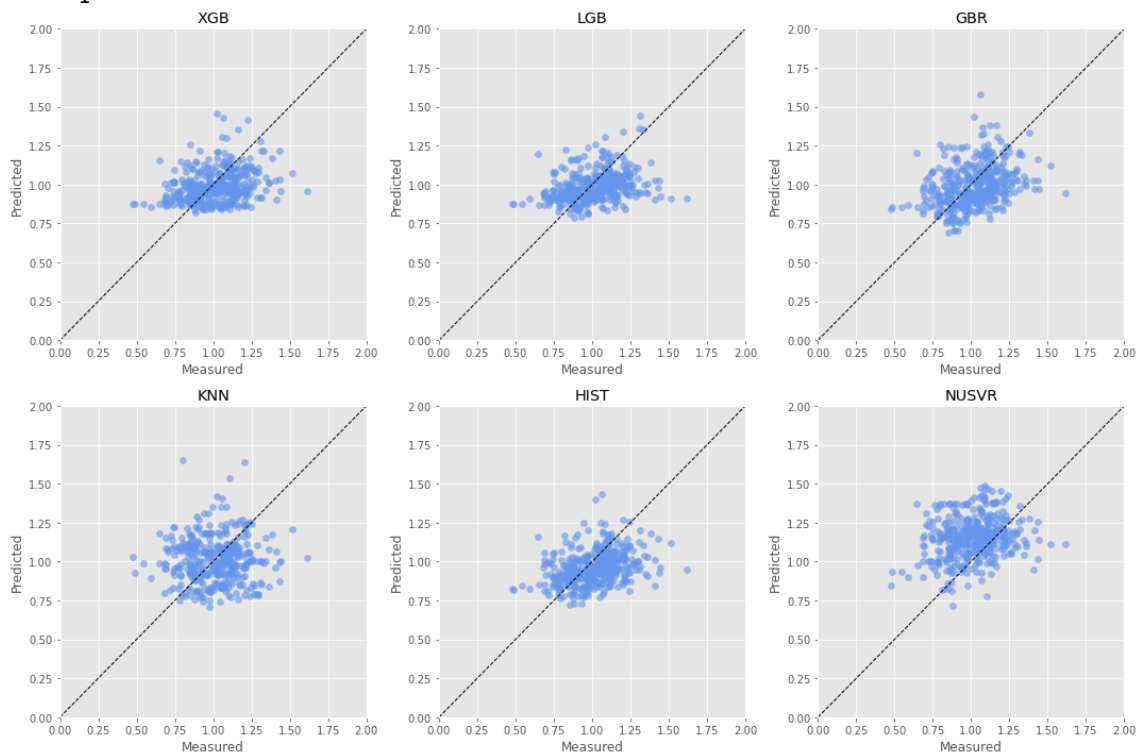
## Phase 2

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 will be used in phase 3. Below we show MAPE and MAE metrics for all models:





- Based on the metrics shown it can be seen that XGB, LGBM, Gradient boosting and Histogram Gradient Boosting are getting the best results, we will perform in detail analysis to them.
- Also based on the SHAP values observed for the different iterations it can be seen that models are mostly well adjusting to parameters based on evidence, and therefore all models mentioned are worth of time series split analysis. Now we will inspect and analyze scatter plots of predictions versus measured for all six models.

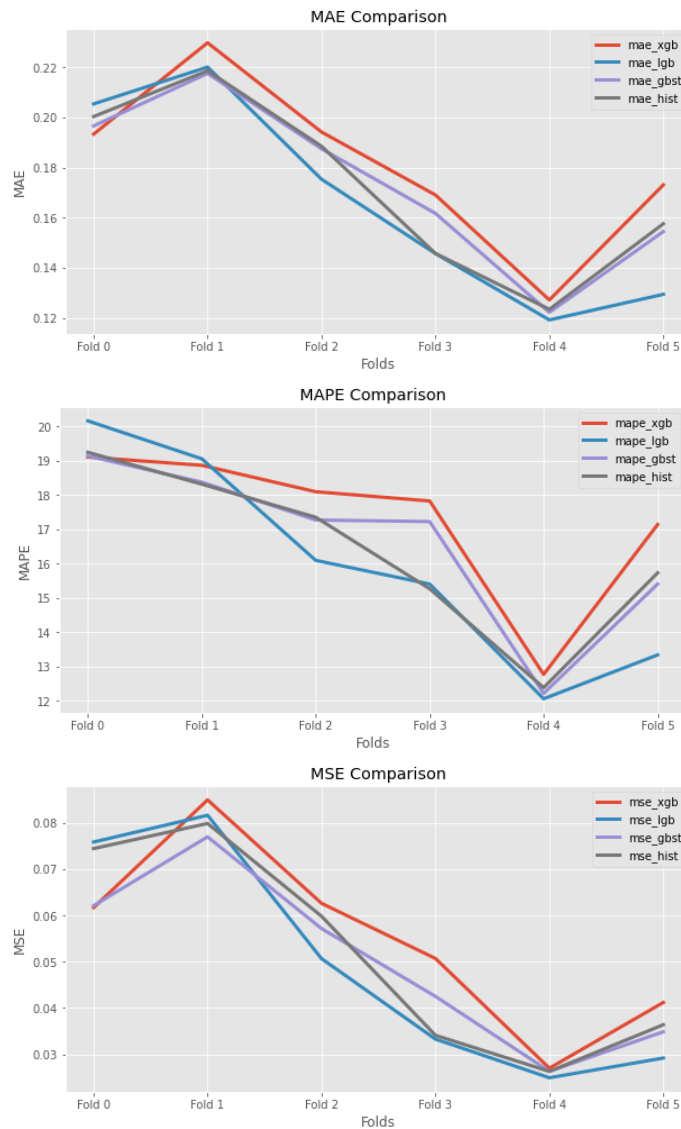


It can be appreciated that the best approaches seem to be the ones obtained by the metrics. We will perform time series evaluation on the selected models.

We will perform 6 splits, with the selected top models, and compare them.

### 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).



- All models seem to be performing better as number of folds increase, except last fold where 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.



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.

### 5.2.1 Conclusions Methodology 1

- In the presented methodology we have covered three different approaches working without lagged values and including mobility indexes. This is due to the fact that we wanted to have more visibility on the sign of features, and leaving estimation to this features would let us have more in detail impressions on the effect of features. Analysis has been performed with 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 offers best explainability for SHAP values and is able to reduce multicollinearity. 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. In terms of metrics, approach 3 using PCA was getting the best results in phase 1.
- 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 LGBM regressor 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. Predicticably 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 behaviour 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.
- In detail evaluation of SHAP values will be performed for the selected model in order to asses anomalies in the sign and magnitude of features that mainly have to deal with multicollinearity.
- Metrics for our LGBM Regressor model are:



- Average Test Mape: 15.95328653434539
- Average Test Mse: 0.049478868842694194
- Average Test Mae: 0.16651967323832476

- A more precise evaluation of the model will be done studying shap values in file

## 5.3 Modelling 2: Lag Methodology

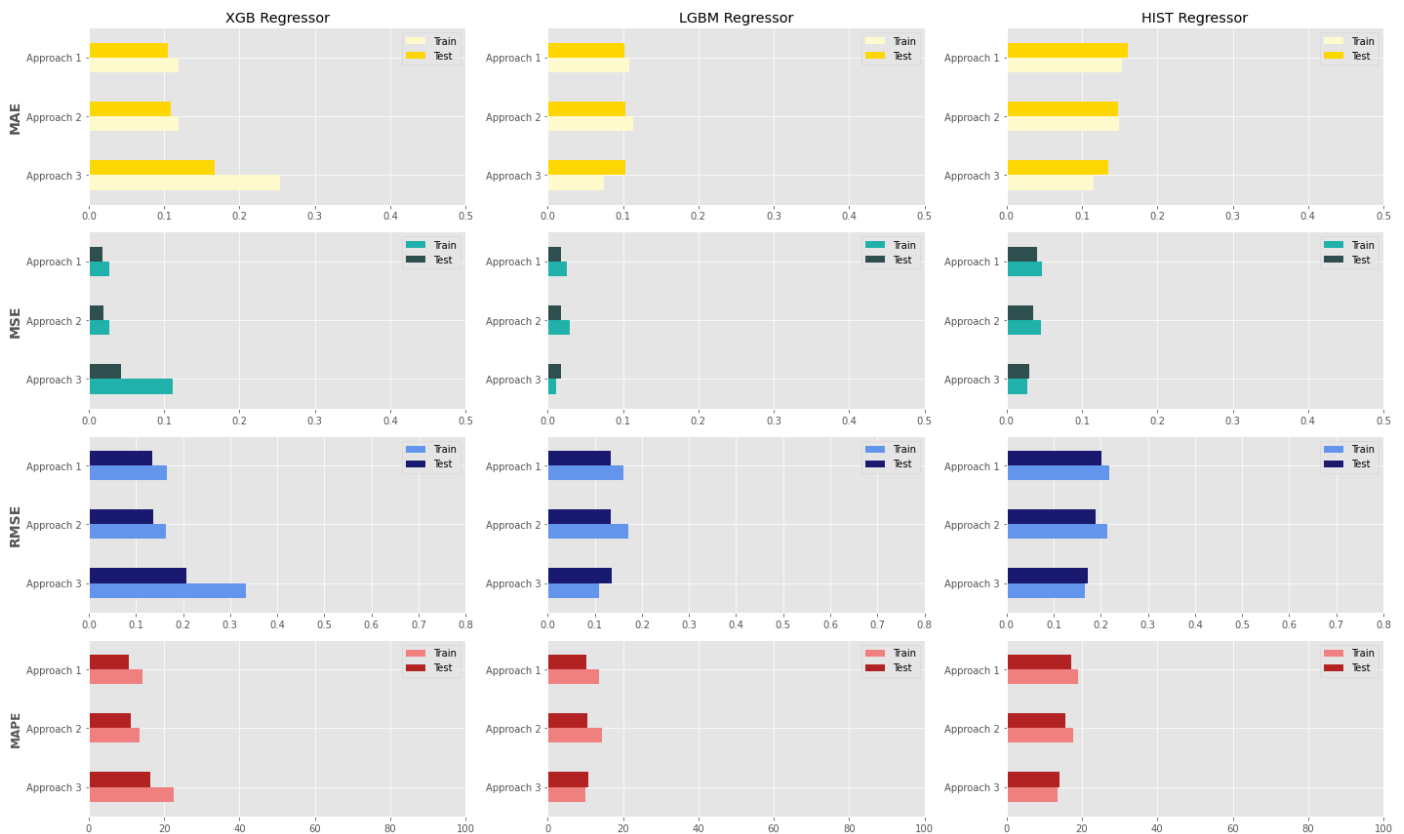
In this modelling methodology our goal is to accurately estimate or forecast reproduction rate based only on restrictions applied. This model is to serve as a forecast tool on what value of reproduction rate is expected in a 2-week time window when applying certain measures together with other features detailed.

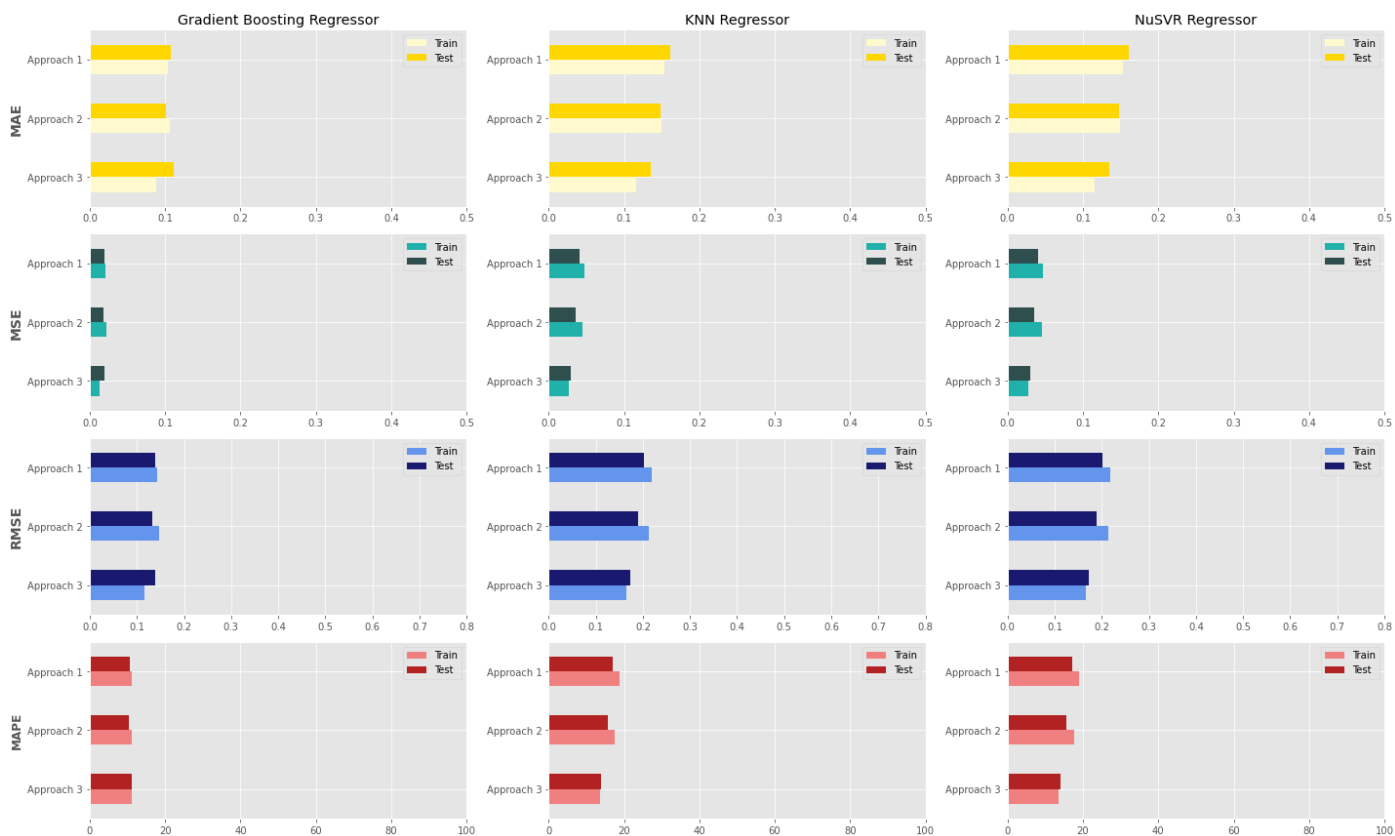
In order to achieve this, we will group our data on 2-week periods, including the value of reproduction rate of week  $n-2$ . Since this model is to serve as a tool, the information of mobility indexes is not included, and infections is not included either since reproduction rate will serve as our infections index. It will be detailed below the resume of evaluation for methodology 2 with three different approaches, however to find full explanation and detail, please refer to the [notebook](#) found on GitHub.

### 5.3.1 Modelling and Evaluation for Defined Phases

#### 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. Below there is a summary of all metrics for three approaches and all models.

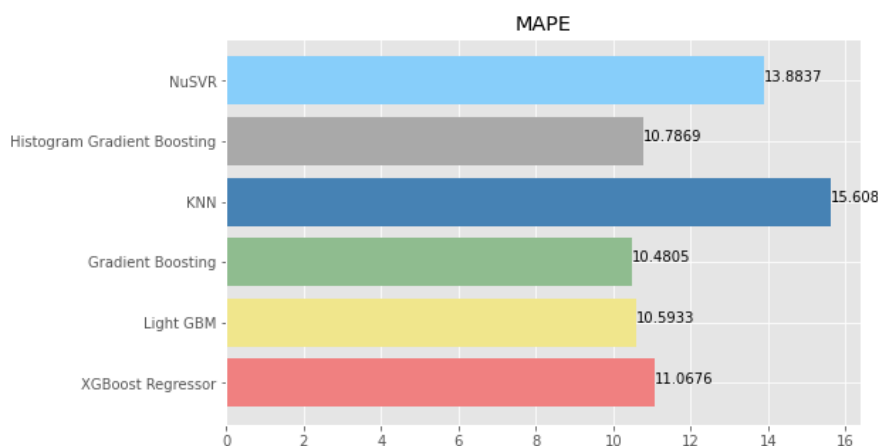


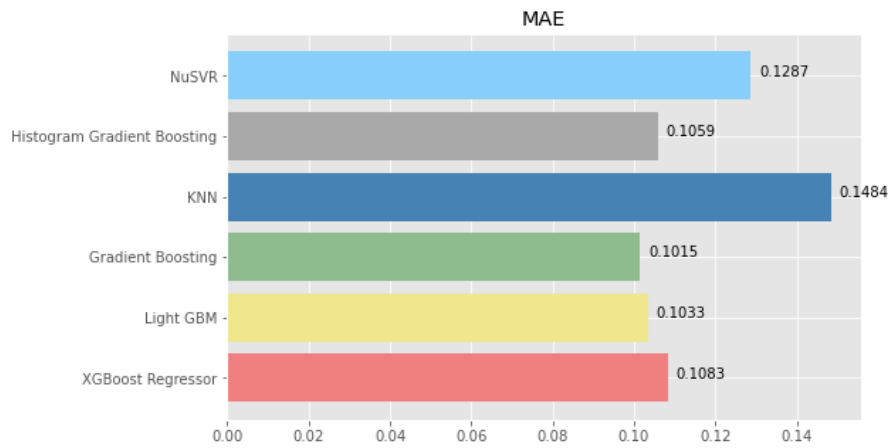


- In conclusion, all approaches seem to be getting similar metric results.
- There are no signs of overfitting based on the metrics and residual plots visualized.
- Boosting models (XGB, LGBM, Gradient boosting and Histogram Gradient Boosting) seem to be getting the best results based on the fits and metrics observed.
- Results seem to have improved from methodology without lagged values

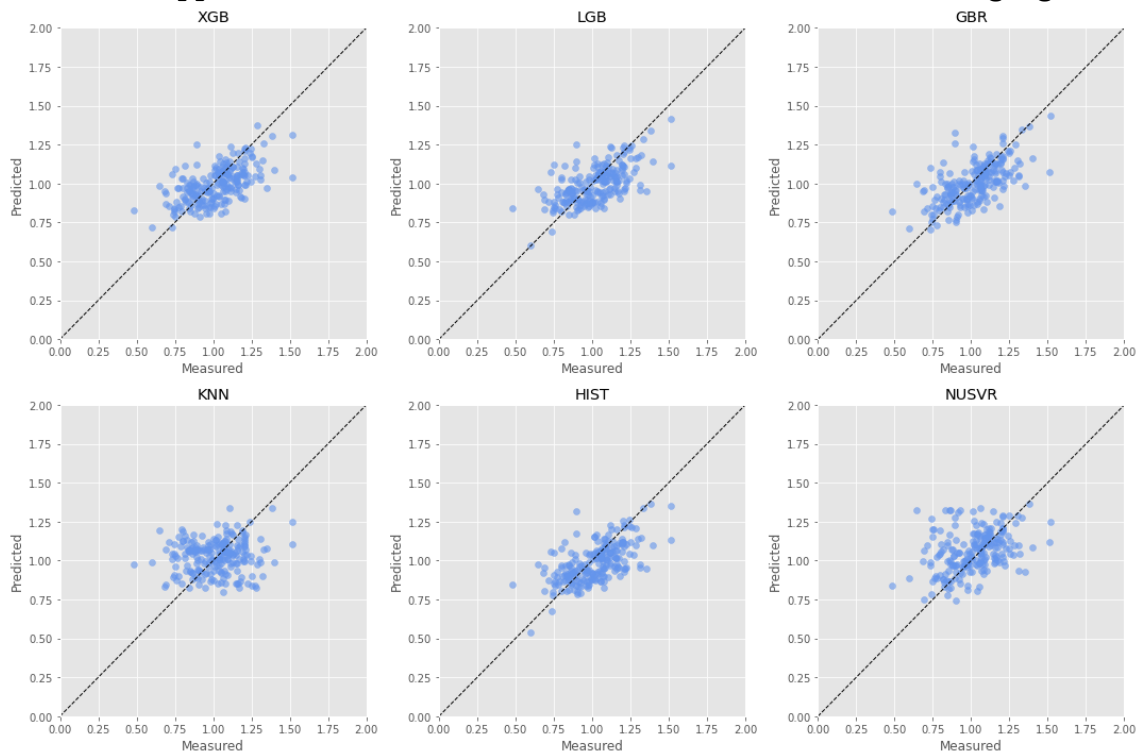
## Phase 2

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 will be used in phase 3. Below we show MAPE and MAE metrics for all models:





- Based on the metrics shown it can be seen that XGB, LGBM, Gradient boosting and Histogram Gradient Boosting are getting the best results, we will perform in detail analysis to them.
- Also based on the SHAP values observed for the different iterations it can be seen that models are mostly well adjusting to parameters based on evidence, and therefore all models mentioned are worth of time series split analysis. Now we will inspect and analyze scatter plots of predictions versus measured for all six models.
- However, some features seem to have very low importance, remarkable low value of awareness measures and health measures.
- Temperature has different sign from other SHAP values evaluation in methodology with no lag values.
- It can also be appreciated how  $r^2$  scores are now much better for boosting algorithms.

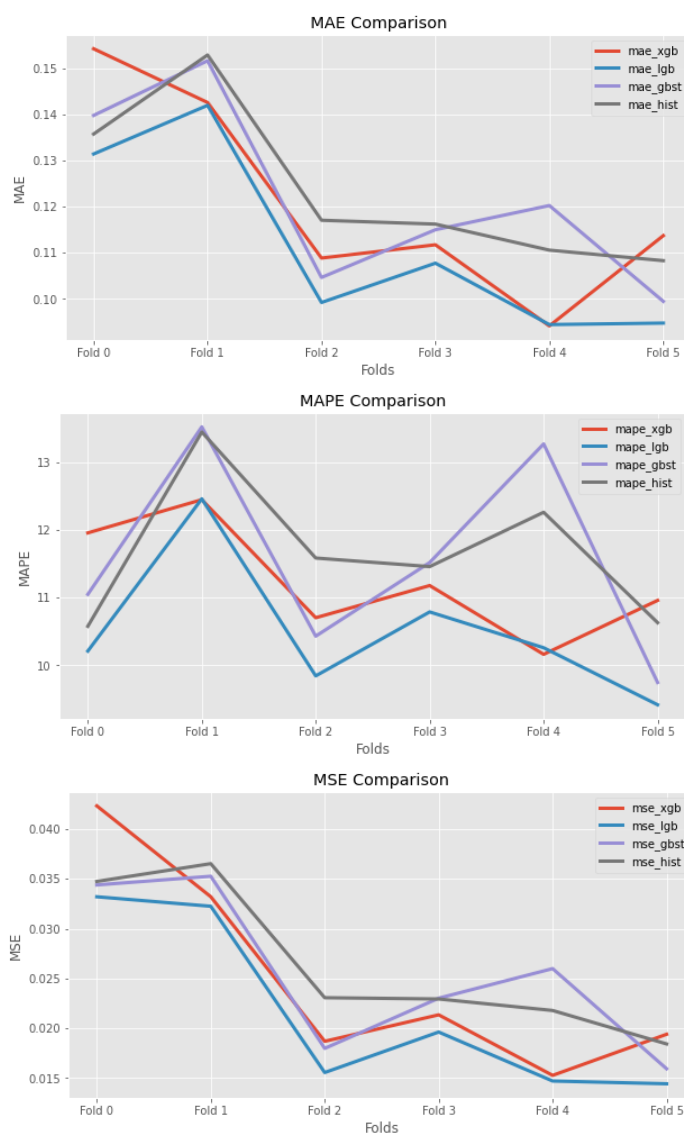


It can be seen that the best approaches seem the ones obtained by the metrics. We will perform time series evaluation on the selected models.

We will perform 6 splits, with the selected top models, and compare them.

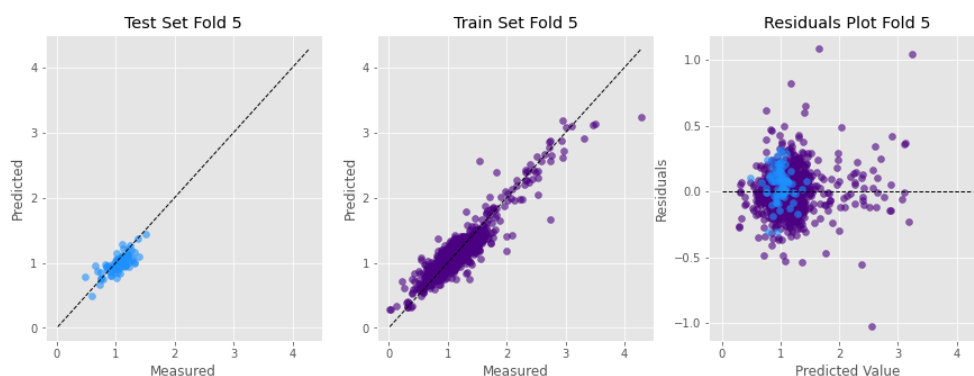
Phase 3

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).



- As represented in the plot, all models seem to be getting similar results, however LGBM regressor, seems to be getting best results among all in average of all folds, XGB is the best fitted model according to the metrics included.
- Results seem to have improved from methodology without lagged values.

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



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.

### 5.2.1 Conclusions Methodology 2

- In the presented methodology including lag values we have covered three different approaches working with lagged values and excluding mobility indexes from the analysis. This is due to the fact that we wanted to create a model that would be able to predict, and in order to get more precise predictions we would need to include reproduction rate as seen improving results, since the significance of that variable is big. Also, if a prediction is intended to be made, mobility indexes are not available when applying restrictions, and also, they are found to have high multicollinearity with closure and mobility restrictions, therefore the effect of those restrictions is annulled by mobility indexes.
- Overall the following three approaches were explored based on the presented methodology using lagged values of reproduction rate.
  - 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 as it has been mentioned, offers the most explanatory view due to the fact that grouping variables reduces multicollinearity. 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. However this is less visible in this methodology including lagged values since predictions are a lot more precise.
- In conclusion this model offers good forecast of reproduction rate based on the variables included and the metrics studied including reproduction rate lags.
- A more precise evaluation of the model will be done studying shap values.
- Our last LGBM Regressor model mean value over 6 folds has the following metrics:
  - Average Test Mape: 10.497015775567002
  - Average Test Mse: 0.021632609470518
  - Average Test Mae: 0.11160438536628081
- Overall metrics have significantly improved from model without lagged values, r2 score is still low when performing folds, explain ability is still the issue detailed with variance.

## 6. Model Features Evaluation with SHAP

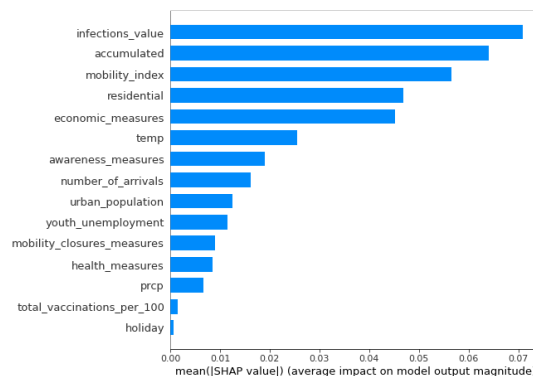
Once an approach and model has been selected, we will now try to understand what the model is doing and what features included in the analysis are the most important, and even more if the value of the feature has positive or negative impact on our target variable reproduction rate. In order to do that we will take advantage of SHAP library, which reflects feature importance and sign 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 referring to last fold of both model methodologies separately.

### 6.1 SHAP Evaluation Modelling 1: No Lag Methodology

In this section we will evaluate the sign and importance of SHAP values associated with our model for methodology 1 without lags and including mobility indexes on training set referring to last fold obtained in phase 3 of modelling section.

#### Importance Bar Plot

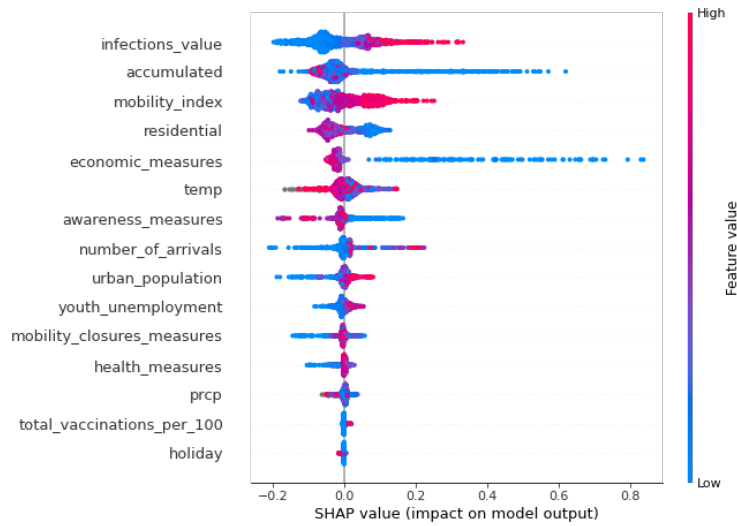
We will plot our feature importance bar plot; this will represent the importance of each feature in terms of predictions of our target variable reproduction\_rate.



- 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
  - mobility\_index which represents the mobility increase or decrease in retail and recreation, transit stations, groceries, pharmacy and workplace.
  - 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
- Other variables that have smaller coefficients are temperature, number of arrivals, closure measures, awareness measures and urban population

#### Tree Explainer Plot

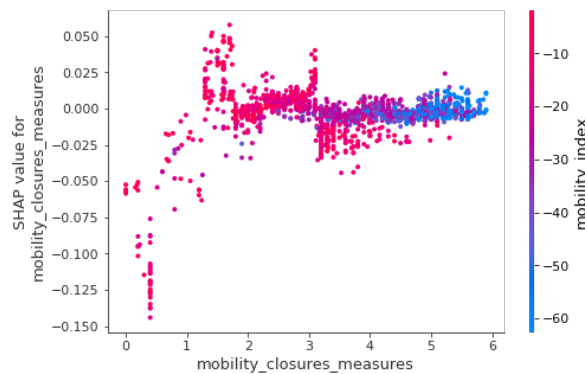
The figure shown below reveals the impact of each variable in our target variable reproduction rate.



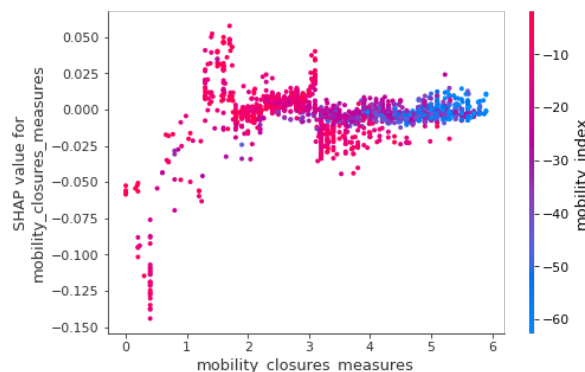
As seen red colored values are related with high values of the feature elected and SHAP value mean its impact on the target variable. In this case all variables seem to have good signs and importance based on expert knowledge in the field. However, there are some features that have signs which don't seem to be correct, will analyze those in detail.

Mobility closure measures seem to have negative impact with low values of restrictions and positive impact on high values of restrictions. We are experiencing the same issue with health measures.

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.

Therefore, the sign and impact of this variable is corrupted due to multicollinearity, therefore it can't be assumed that mobility and closure measures have positive impact on our target variable. The effect of this variable is being captured by a variable that has high collinearity with this variable which is mobility index. It can also be viewed in the correlation matrix how this variable has negative impact on the target variable.

#### 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 in relation with number of positive cases among a lot other variables and 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 predictably 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 neither 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.

#### Conclusion

- Overall the presented model as seen in the modelling section as a forecast and predictive model does not seem to be very good, although mae and mape scores are relatively good in terms of deviation against the target variable it has been appreciated and detailed the issue of  $r^2$  score.
- Multicollinearity between restrictions and mobility indexes are corrupting our interpretation of the features, however most signs and importance seem to make sense. Based on this premise of collinearity it can be assumed that number of infections, accumulated and mobility index seem to be the most important variables in terms of prediction.
- From an explanatory point of view as seen in this notebook most features have signs and effects that match with expertise knowledge and signs in the correlation matrix. However there are two features regarding restriction, mobility measures and health measures that do not have the expected behaviour, however as it has been assessed this is due to multicollinearity issues detailed. Vaccinations do not have an effect as it has also been detailed due to lack of data of vaccination period.

Based on the output of SHAP values it can be estimated that top features have the following ranges effect on our target variable reproduction rate, which ranges (0, 4.8). This table represents in descending order the impact of each feature included in our target variable and in what range it can modify our target variable as recorded by our model.

Feature	Range
infections_value	[-0.19,0.33]
accumulated	[-0.17,0.62]
mobility_index	[-0.12,0.25]
economic_measures	[-0.05,0.83]
residential	[-0.09,0.12]
temp	[-0.16,0.14]
awareness_measures	[-0.18,0.16]
urban_population	[-0.18,0.08]

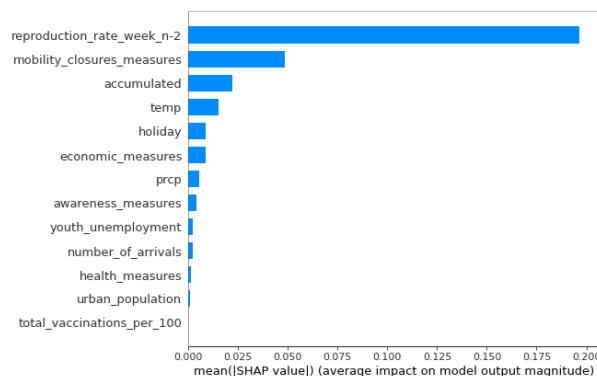
Economic measures do seem to be independent from mobility index since they are included in the presented table. Overall, the presented model is able to give some insights and acknowledge the effect of restrictions and mobility indexes over our target variable.

## 6.2 SHAP Evaluation Modelling 2: Lag Methodology

In this section we will evaluate the sign and importance of SHAP values associated with our model for methodology 2 including lag values, on training set referring to last fold obtained in phase 3 of modelling section.

### Importance Bar Plot

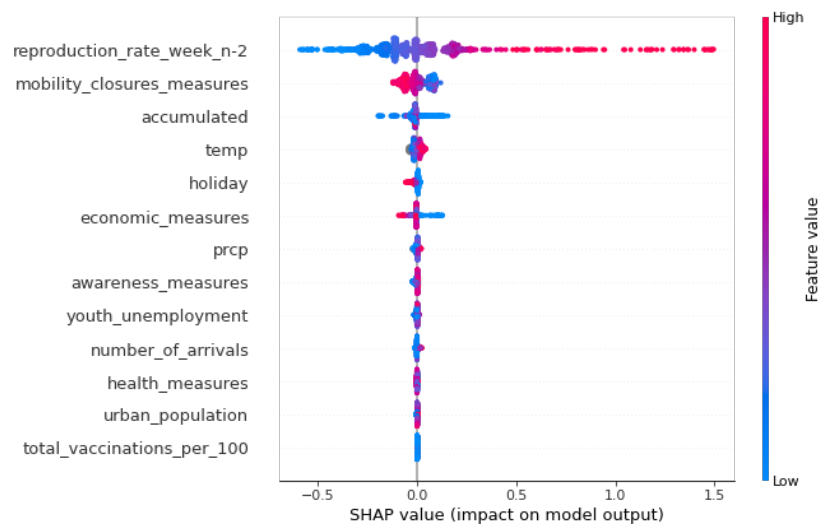
We will plot our feature importance bar plot; this will represent the importance of each feature in terms of predictions of our target variable reproduction\_rate.



- As seen variables that seem to have a higher impact on our target variable reproduction rate are:
  - reproduction\_rate\_week\_n-2 which represents the reproduction rate value of 2 weeks previous to the predicted value.
  - mobility\_closures\_measures which represent the restrictions in mobility and closings applied.
  - accumulated which represents the amount of people that have already contracted the virus
  - mobility\_index which represents the mobility increase or decrease in retail and recreation, transit stations, groceries, pharmacy and workplaces
  - temp which represents the mean temperature presented in 2 week period
- Other variables that have smaller coefficients are economic\_measures, infectious value, prcp, holiday, youth unemployment, awareness\_measures, health\_measures, number\_of\_arrivals, urban\_population and total\_vaccinations\_per\_100 don't seem to have a big impact on the target variable based on the mean shap values shown.

## Tree Explainer Plot

The figure shown below reveals the impact of each variable in our target variable reproduction rate.



As seen red colored values are related with high values of the feature elected and SHAP value mean its impact on the target variable. In this case all variables seem to have good signs and importance based on expert knowledge in the field. However, there are some features that have signs which don't seem to be correct, will analyze those in detail.

The only variables that do have significant impact on our target variable are reproduction\_rate\_week\_n-2, mobility\_closure\_measures, accumulated, temp, holiday and economic measures. The rest of features seem to have little to no impact on the output variable, the coefficients and importance of those variable may be absorbed by some other variables previously mentioned.

It is also remarkable that the variable temp has different sign on this methodology model than in first methodology model, therefore our conclusions and estimations on the first model may be wrong. In the end and based on these results we can't precisely affirm if temperature is a good or bad factor with our target variable reproduction rate. Based on evidence and studies found it is estimated that high values of temperature have negative impact on the spread of the virus. *"Several studies indicate that the transmission of COVID-19 is affected by temperature. An inverse correlation was found between temperature and the daily number of infections."* [Does Temperature Affect COVID-19 Transmission?](#)

Based on this evidence the sign retrieved by this model for our target variable temp may be incorrect.

## Conclusion

- Overall the presented model as seen in modelling section offers good and precise predictions according to metrics mae and mape studied. It even records a positive  $r^2$  score in contrast with model obtained in methodology 1. Therefore it could be used to estimate predictions, however it is a model that highly relies on the values of reproduction rate prior two weeks, therefore it might not capture big changes in restrictions or tendencies, and other features included.

- However, it is able to detect changes in trend as seen above, were reproduction\_rate\_week\_n-2 does not have all the explanation of the prediction but it does have most of it.
- From an explanatory point of view as seen, most features have zero to very small effect on the target variable, and the signs of each feature seem to be confusing as well for some cases. However just as with the non lagged version, this is due to multicollinearity, variable mobility\_closures\_measures captures all the effect of the rest of restrictions, since restrictions tend to be applied at the same time with the same severity.

It is also important to note that for both methodologies' vaccinations seem to have no impact. This variable is of course predictably the variable that should predict the most together with accumulated and infections or reproduction rate of week n-2, 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.

Based on what has been covered over bot methodologies and models captured from each one it can be concluded that model 1 working without lag values offers good explanatory analysis based on the features captures. Model 2 working with lag values offers good predictions in terms of estimation and metrics, but its explain ability and usage is rather low.

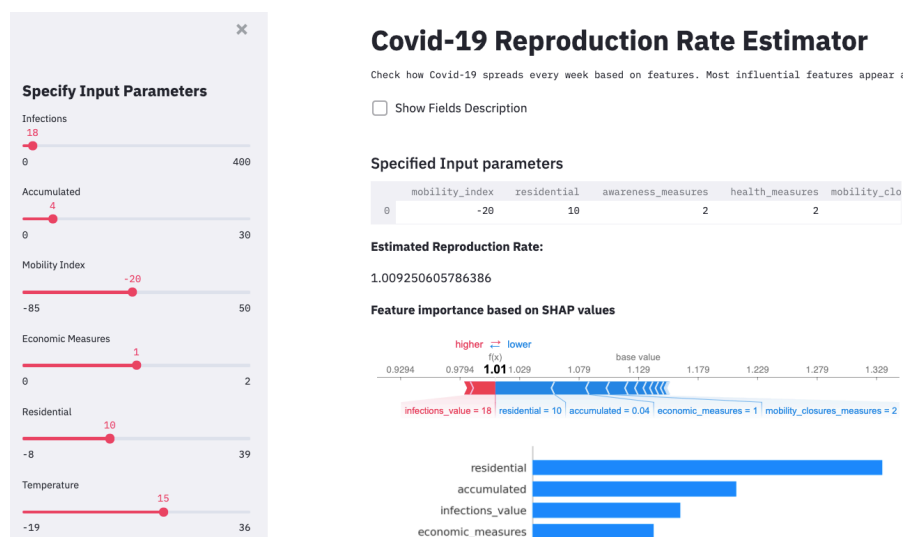
## 7. Frontend Applications with Streamlit

In order for users to be able to interact with both generated models, two applications have been developed taking usage of Streamlit. A tool that helps us create frontend applications using Python.

Both applications follow same structure, however each serves a different purpose and the inputs that both models retrieve is of course different. Its code is also available on GitHub. Instructions of used can be found on README.md of GitHub repository.

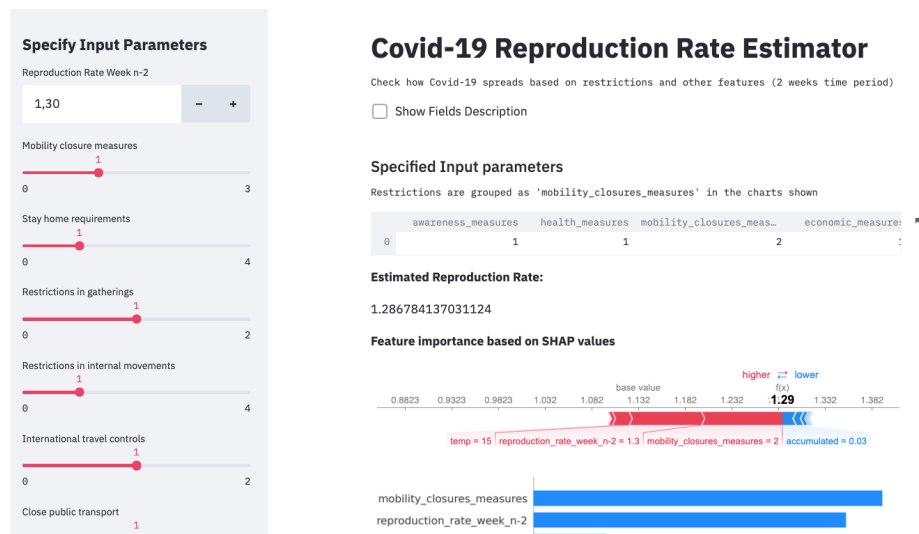
It is important to note that as it has been appreciated on SHAP analysis, there are some variables that do not seem to have impact on our target variable based on the models, therefore when applying modifications to these features they may not be revealed or appreciated as a response in the target variable in the app.

covid-19-app.py (No lags methodology)



As seen this application allows user to enter values of selected features, once they enter the information, estimated reproduction rate value is shown with its interpretation with SHAP values, and a bar importance plot of each feature in the selected prediction. It also includes a help text if “Show Fields Description” is pressed, showing detail of each feature and its value meaning.

## covid-19-app-lags.py (Lags methodology)



In this case the functionality is the same as the app without lags, however this one of course does not have the ability to include mobility indexes as parameters. On the other hand, this app includes the ability to include the value of reproduction rate on week n-1. Since measures variables of mobility and closures that are the most important have been grouped under variable “mobility\_closure\_measures”, in this case we allow the user to enter information for each of these variables, we then make the calculation and output the result.

Once the information has been entered it shows the SHAP values interpretation as well. It also includes a description text with insights and details on the value of each variable.

Both applications have been developed in order to offer visibility and usability of both models.

## 8. Final Results and Conclusion

The goal of the presented project was trying to estimate and understand how restrictions, mobility trends, and demographic features of each country affect the spread of Covid-19, in order to measure this, spread we used reproduction rate index, which is an standard used in epidemiology for virus spread, which is a measure used by countries in order to manage the pandemic. In order to do this, we gathered information from 61 countries from March 2020 till April 2021. The project can be divided in five phases:

- 1) Problem Statement and Modelling
- 2) Data Gathering and Preparation
- 3) Data Exploration
- 4) Modelling and Evaluation
- 5) Frontend app Creation

On phase one we had to make a statement of the problem approach and what modelling was intended to be used. In order to achieve this, we based our data structure and modelling on the SIR model, however there were many approaches that were treated before selecting SIR. Such approaches include solutions such as clustering per countries, raw death forecast and multiclassification of excess deaths. However due to bad results, and unfeasibility they were deprecated. It is important to note that there were some variables included that based on the state

of art in terms of covid forecast models and current evaluated projects have not been included, such as temperature, precipitation and youth unemployment.

On phase two there were two challenges, firstly gather weather data from NOAA using Google cloud, and later on merge that information with a set of files each containing different restrictions and demographic characteristics for each country. Overall, this process was very time consuming and costly overall. Custom functions to iteratively retrieve files from a single folder in drive were also developed in order to avoid multiple calls to different files.

On phase three our goal was to get some insights and first idea on how the data was distributed and its relationship with our target variable reproduction rate. On this phase we concluded that the dataset had an issue with multicollinearity that was inherent to data. Restrictions are applied at the same time in every country with same severity, and this does not only affect multicollinearity between restrictions, but also restrictions and mobility trends, which in the end would highly difficult our analysis. VIF analysis was done in order to evaluate the magnitude and significance of this issue. In order to fix this, we defined three different approaches of analyzing our dataset.

It was also concluded that values of reproduction rate as times passed tended to be around one with very low variance, this is due, as mentioned to the fact that government's goal is to find this reproduction rate value of 1, and all values since June 2020 range from (0.5, 1.5).

On phase four based on results there were two methodologies taken and two models created. A first approach was to estimate the value of reproduction without including lag of reproduction rate of previous week since it was highly correlated and would leave to no explain ability of other variables. Therefore model 1 was created, however since metrics and results were not good overall it was decided to create a second model, in this case grouped in two-week time periods in order to forecast accurately the target variable. In order to achieve this regressor models such as XGB Regressor, LGBM Regressor, Gradient boosting Regressor, KNN Regressor, Histogram Gradient Boosting Regressor and NuSVR Regressor, and evaluating all this models with mae, mse, rmse and mape taking use of a custom function created based on time series split metrics, together with deep analysis on SHAP values.

Therefore, in the end two models were created,

- 1)** Model 1 no lags: One first model that offered a good explain ability and interpretability of the features included based on SHAP values and expertise knowledge gathered, with metrics, average mae: 0.16 and average mape:15.95, bad fit in measured vs predicted line.
- 2)** Model 2 lags: Model with good and precise forecast estimation average mape: 10.49 and average mae: 0.11, very good fit in measured vs predicted line, but low explain ability based on the analysis performed with SHAP values.

Overall, in this phase we have to again highlight multicorrelation between variables. As seen when studying SHAP value signs and importance it was concluded that models that work with correlated variables for our case tend to charge the impact of those variables over one single one. It was appreciated how in model 1, mobility index captured all the information that other variables were including. We have also encountered as mentioned some features with unexpected signs and magnitude and this is due to multicollinearity.

However, these effects gathered from restrictions even when it is attracted by one restriction, are apporporting information, since all restrictions are applied at the same time, in the end, it means how much all restrictions applied are able to modify the evolution of the epidemic. Therefore, even though multicollinearity is an issue for model interpretability in machine learning problems it was stated that the sign and effect of variables are the conglomerate sum of them all, and its effect separately may differ in case of them not been applied together.

On phase five two apps were created in order to deploy and create an interactive tool for users to work with the models presented, such applications were developed with Streamlit, and include model response interpretability with SHAP values.

Overall a complete cycle of a Data Science project following CRISP-DM methodology was covered from collecting data and manipulating it from scratch, exploring and analyzing features for later modelling and evaluation using multiple algorithms, metrics, approaches and model types, ending with a deployed solution of the two models selected.



## Annex: Data Sources

- Our World in Data. Oxford University Covid-19 Data and Google Mobility trends
  - <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>
  - <https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv>
- 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/>