Machine Learning Major Project: Group 15

Harshvardhan Walia *2017B4A71027G*, Ishan Bansal *2017AAPS0356G*, Riddhesh Sawant *2017AAPS0261G*, Amandeep *2017B4A70576G*, Nikhil Srivastava *2017A7PS0091G*

***Abstract*—We were tasked with predicting the spread of the Covid-19 pandemic based on the data and information available and collected throughout the region of Europe. This has been achieved using Machine Learning models: "Prophet" and "Random Forest" and a comparative analysis is done to gain an insight on factors affecting the spread of the disease.**

I. Introduction

THE current world is facing the great challenge of combating against the Covid-19 pandemic. The consequences were especially significant in the European region with countries like France,Germany,Italy and Spain suffering the most cases. Therefore these countries have been taken into the focus of our study. The objective of this project is to collect and visualise the data about spread of Covid-19 along with using Machine Learning models to forecast the full extent. Various factors including vaccination numbers and lockdown methods have been analysed and taken into consideration for the prediction.

II. Machine Learning Model Overview

A. Prophet: Forecasting Procedure

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Holiday effect is caused by events that provide large, somewhat predictable shocks to the data being forecasted and often do not follow a periodic pattern, so their effects are not well modeled by a smooth cycle. In the case of spread of Covid-19 in a country, these can be paralleled with events such as erratic lockdowns or an annual festival. Prophet is therefore robust to missing data and shifts in the trend, and typically handles outliers well.

B. Random Forest Regression

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set. However, it is important to know your data and keep in mind that a Random Forest can't extrapolate. It can only make a prediction that is an average of previously observed labels. In other words, in a regression problem, the range of predictions a Random Forest can make is bound by the highest and lowest labels in the training data. This behavior becomes problematic in situations where the training and prediction inputs differ in their range or distributions.

III. Data Acquisition

The Covid-19 disease was declared a pandemic on 30th March 2020, and had grabbed the attention of the entire world. Both government bodies and private researchers began collecting and recording data regarding its spread to gain a better insight. This made the data be more readily available and could be accessed by the general public. The core entries in the data we collected were the number of daily cases and the number of daily deaths. These numbers could then be correlated to other factors that occurred in the country that led to a spike or downfall in the number of cases, as well as help us gauge the infection and transmission rate of Covid-19. Some of these factors include lockdown duration and methods, vaccination numbers, and other measure taken by the government to curb the spread. The focus of the study being on Europe, four countries were selected to be studied: France, Germany, Spain and Italy. The reasons for selecting these countries were: they are geographic neighbours, and hence expected to show similar characteristics in data; recorded highest number of infected cases; economically strong and urban, meaning both factors such as international travel and medical aid can be analysed.

IV. Data Visualisation and Analysis

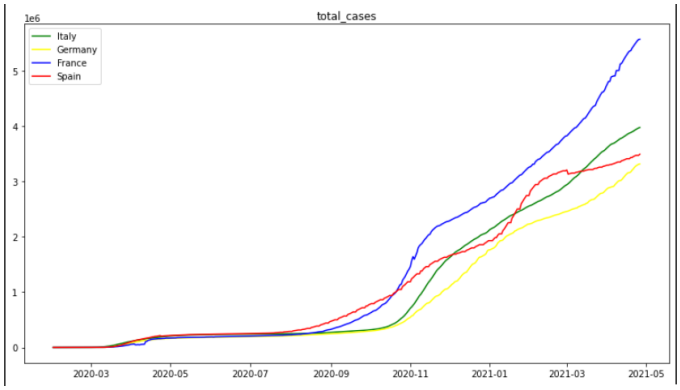

Fig. 1. Cumulative count of cases.

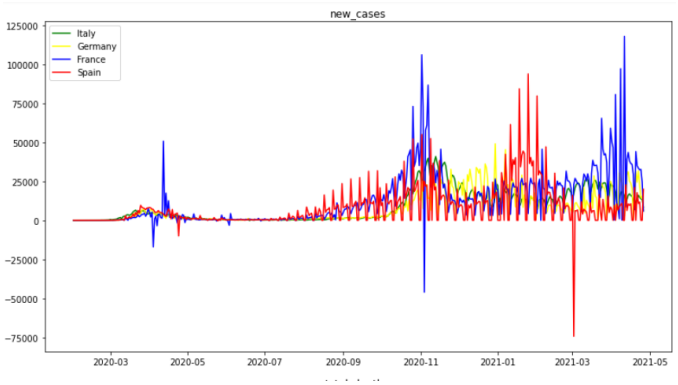

Fig. 2. Daily new cases.

Fig. 1 and Fig. 2 are the initial data visualisation plots of daily and cumulative count of positive cases in all four countries. Our initial hypothesis that these neighbouring countries which have similar geopolitical environment and culture would showcase similar trends in spread came out to be true. In all countries the total case count started to rise in late October 2020, with daily new case counts showing peaks in November, February and April. Few reasons that support the high correlation between countries is that inter country travel between them is rich with an average 1 million to 3 million annual tourists. Also similar geopolitical environment encourages similar level of safety policies during similar time periods.

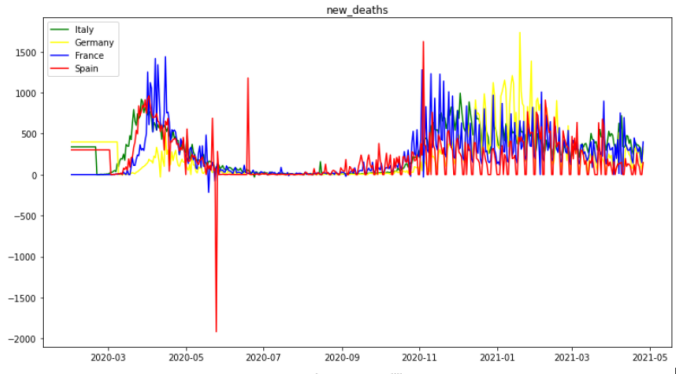

Fig. 3. Daily new deaths due to Covid-19.

The daily new deaths follows the same trends as the daily new positive case count which could be indicative of the fact that the lethality of the virus has not increased drastically even with the news of different variants and mutations.

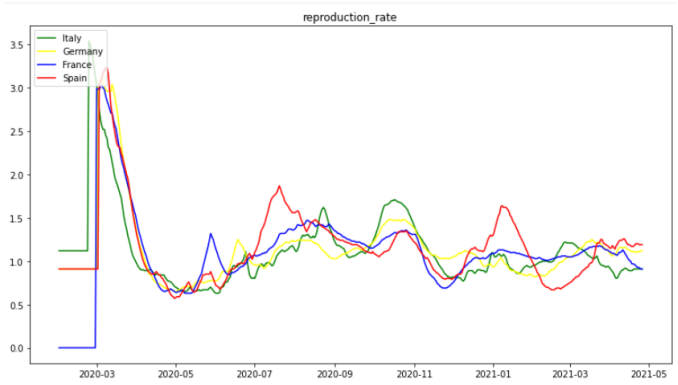

Fig. 4. Infectivity and Transmission reat of the Virus.

From Fig. 4 which shows the transmission rate of the virus from one person into the society we might initially draw a conclusion that the virus has diminished its potency, however this is actually due to the impact of public awareness and safety measures. From government policy news and data most of these countries implemented lockdowns and either recommended or enforced wearing of masks. These factors led to the decline of the inital transmission rate of the disease. Thereafter each country easied up its policy according to their current situation.

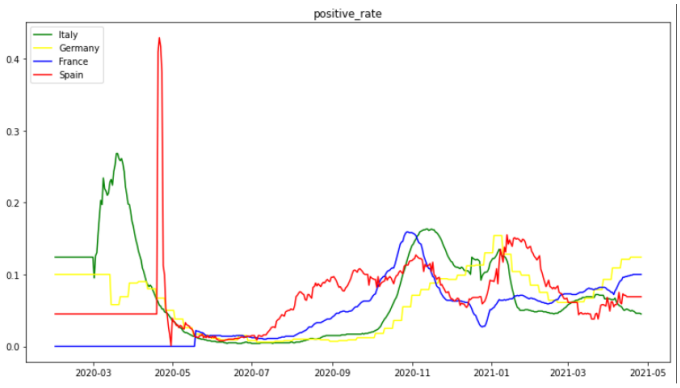

Fig. 5. Percentage of people sampled that were positive.

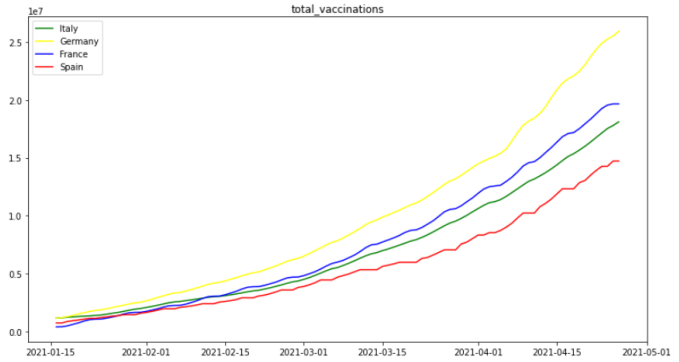

Fig. 6. Cumulative count of Vaccines succesfully administered.

All four countries in focus have a high standard and capability in the field of medical facilities. As soon as the pandemic

status of Covid-19 was announced on 11 March 2020, every country started vigorous public testing showing high inital peaks in percentage of positive cases(Fig. 5). In accordance they were quick to dispatch and administer vaccines to the general public as seen in Fig. 6

V. Methodology

Based on the above initial analysis and assumptions we proceeded towards estimating the total number of cases. As a data pre-processing step multiple provinces data within a country is aggregated based on dates and final table is replaced with mean values. Also NULL values have been changed to estimated values. Multiple mobility features are used as feature to train the model and are assigned binary or index values relative to base mobility before pandemic.

A. Prophet

For all Plots from Fig. 7 to Fig. 11: * represents actual data (smoothed); light blue is the prediction range with eighty percent confidence and dark blue line is the predicted data line. (Higher degree of fluctuations can be attributed to fluctuations of mobility and testing data based on various days of the week, and imposition of partial lockdowns)

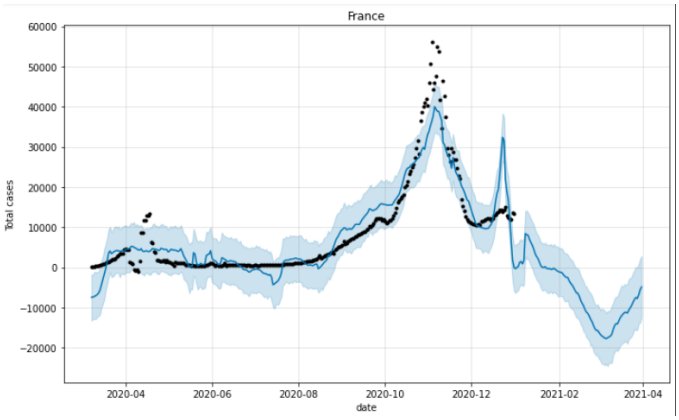

Fig. 7. Forecast of Daily Cases in France.

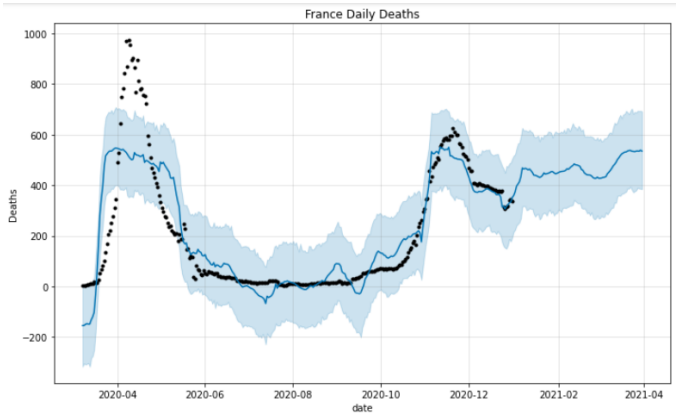

Fig. 8. Forecast of Daily Deaths in France.

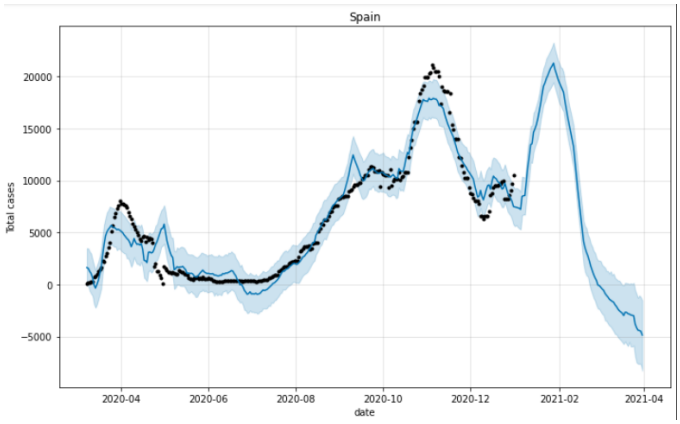

Fig. 9. Forecast of Daily Cases in Spain.

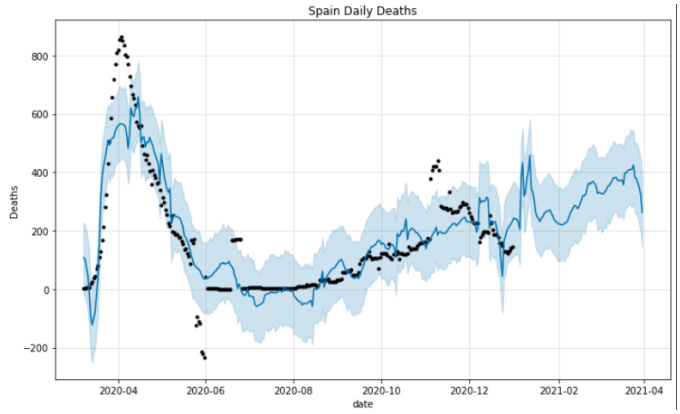

Fig. 10. Forecast of Daily Deaths in Spain.

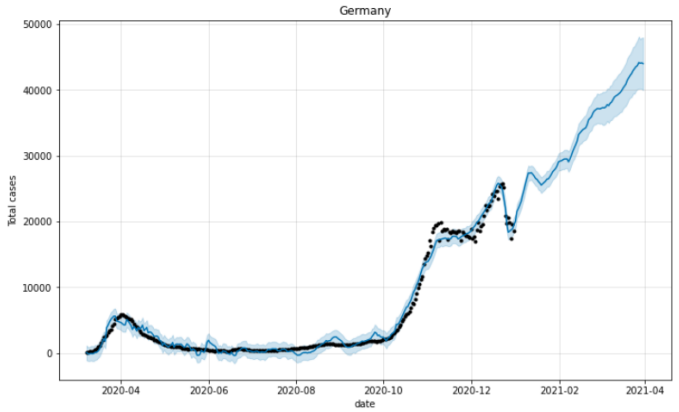

Fig. 11. Forecast of Daily Cases in Germany.

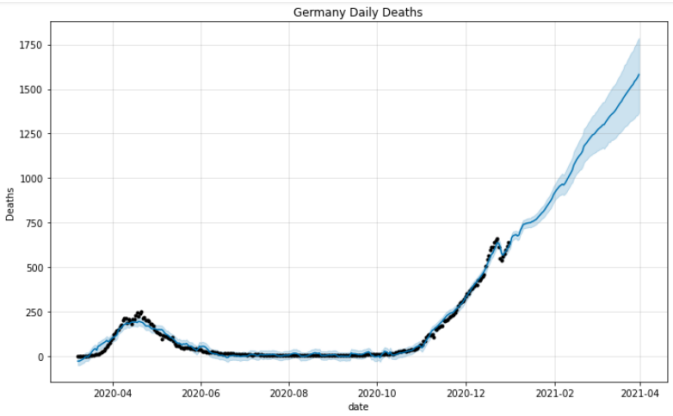

Fig. 12. Forecast of Daily Deaths in Germany.

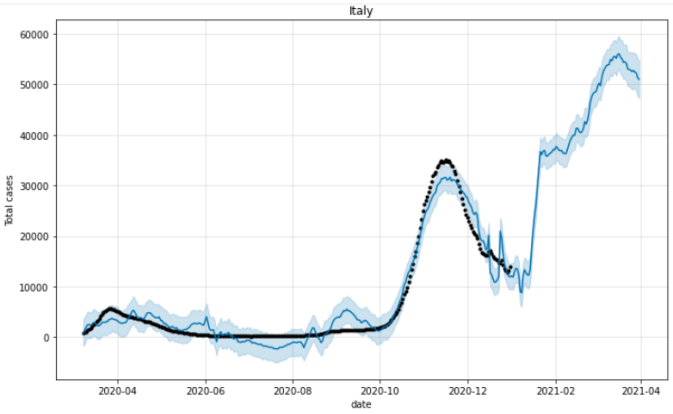

Fig. 13. Forecast of Daily Cases in Italy.

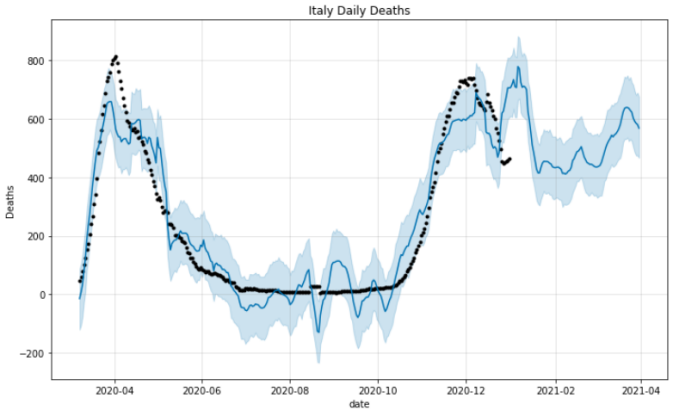

Fig. 14. Forecast of Daily Deaths in Italy.

Using prophet, we try to understand the effects of vaccination drives in these countries. For this we use the before vaccination i.e up to December 2020 data of smoothed everyday new cases with the features: retail and recreation percent change from baseline, grocery and pharmacy percent change from baseline, parks percent change from baseline, transit stations percent change from baseline, workplaces percent change from baseline, residential percent change from baseline, new tests performed (smoothed). This is used to train the prophet model after which data of features is used to predict new cases everyday from January 2021. This model accurately predicts dates for peaks of new cases within few days of error but predicts a much higher number of cases (comparison of prophet predictions with Fig. 2) at these peaks implying a strong positive effect of these vaccination drives.

B. Random Forest

We used RandomForestRegressor to try and predict the ratio of new cases 7 days from today. At first, a model was trained, and then used to try and predict future cases from the database. After this was found to be too simplistic and failed to predict long term cases, a dynamic approach was needed. For each instance we decided to train a fresh model using prior data and predict the ratio of new cases 7 days later. This model performed fairly well on Italy and Germany, whose graphs have been included.

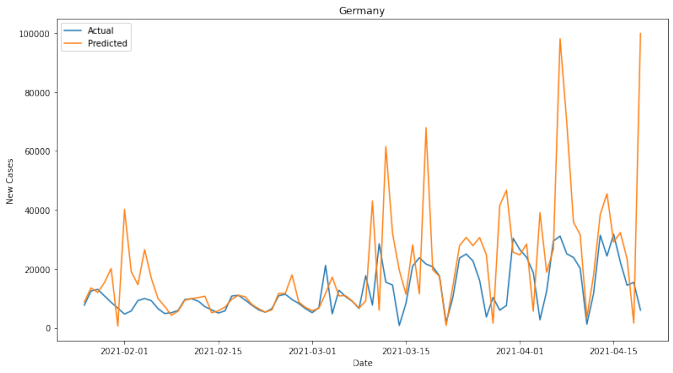

Fig. 15. Actual vs Predicted New Cases in Germany on testing data

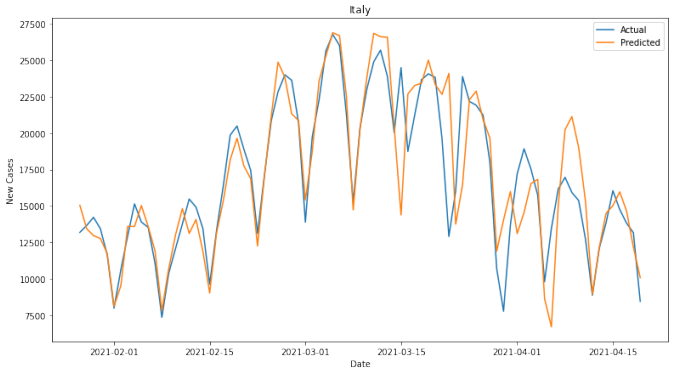

Fig. 16. Actual vs Predicted New Cases in Italy on testing data

VI. Results

The prophet model has been successful in capturing the trends followed in new cases very well. It has high accuracy and the peaks have been captured well on the basis of mobility data. The mobility is also affected by the lockdown restrictions and thus is indirectly affecting the final output. As the vaccination data was not used as features, we can see that

post 1st Jan (initial roll out of vaccines in most of the Europe) there has been a dip in actual cases, while our model predicts the cases to go higher based on the mobility, which is justified as the restrictions were up lifted. The predicted values show that there should have been a further rise in cases but in reality they have decreased which shows the effect of vaccinations has been positive and similarly we can observe for deaths as well, that the number of deaths have decreased which shows that vaccinations decrease the mortality rate. Vaccination seems to be the only way ahead to move out of this pandemic.

The Random Forest dynamic model results in Italy are promising. They show the model can accurately predict numbers. In Germany the peaks predicted by the model are higher. Similarly in France and Spain the predictions are a little varying but showing similar trends. The high peaks could've been avoided because of the effect of vaccines.

VII. References

Data References:

1]https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

2]https://www.google.com/covid19/mobility/

3]https://covid19.apple.com/mobility

4]https://www.bts.gov/covid-19

5]https://www.ecdc.europa.eu/en/covid-19/data

6]https://ourworldindata.org/policy-responses-covid

Others:

7]https://github.com/facebook/prophet

8]https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f