

Business Problem

Novel Coronavirus disease 2019 (COVID-19) emerged from Wuhan, China in December 2019. World Health Organization (WHO) officially declared COVID-19 as a worldwide pandemic on 11th March 2020 (Cucinotta & Vanelli, 2020). Since March 2020, the COVID-19 pandemic has led to major lifestyle changes such as constant masking up, regular hand and space sanitization, social distancing and strong restrictive measures such as lockdown, quarantine and closure of almost all sectors (Xu, et al., 2021). From December 2019 to date, the COVID-19 has affected global countries with nearly 314,453,948 confirmed cases identified, more than 5,523,849 deaths occurred and 261,908,298 recovered cases reported (Coronavirus, 2022). The ability to identify the dynamic of the COVID-19 is important in the battle against pandemics. Prediction of the future trajectory of the pandemic will equip national authorities with the tool to plan public health and policymaking to address the pandemic outcomes (Battineni & Chintalapudi, 2020).

Objectives

The proposed study on Russia COVID-19 forecast aims to:

1. Predict the number of confirmed cases and deaths that would occur in Russia using FB Prophet, Linear Regression and Random Forest models
2. Compare and evaluate performance of the predictive models

Methodology

Figure 1 illustrates flow diagram of the proposed methodology for this study.

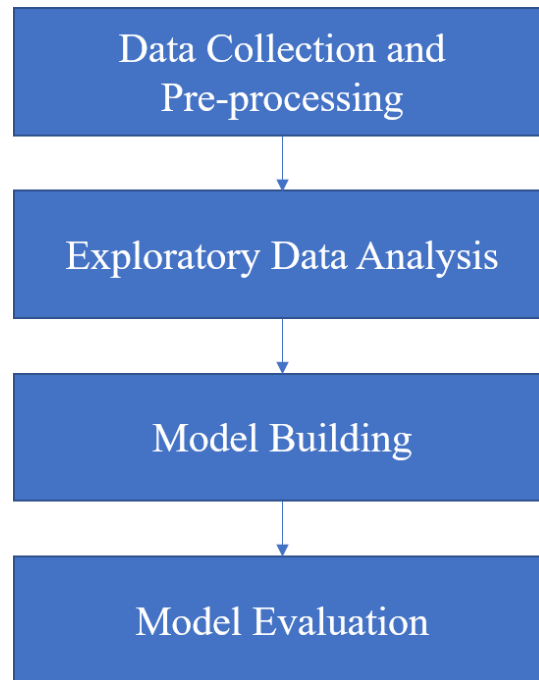


Figure 1. Flow Diagram

Data collection and pre-processing

The datasets used in this study were retrieved from <https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university> (COVID-19 data from John Hopkins University, 2021). It is a daily updating version of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The datasets consist of confirmed cases and deaths on a country level from 23rd January 2020 to 1st December 2021. Figure 2 and Figure 3 display glimpse of the global confirmed cases and death datasets respectively.

Country/Region		Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and Barbuda	Argentina	Armenia	Australia	...	United Kingdom.11	Uruguay	Uzbekistan	Vanuatu
0	Province/State	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Australian Capital Territory	...	NaN	NaN	NaN	NaN
1	1/23/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	1/24/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	1/25/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	1/26/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
...
675	11/27/21	19.0	405.0	163.0	0.0	9.0	0.0	1521.0	517.0	7.0	...	39567.0	291.0	218.0	0.0
676	11/28/21	28.0	418.0	172.0	0.0	5.0	0.0	888.0	409.0	7.0	...	36507.0	167.0	232.0	0.0
677	11/29/21	42.0	195.0	192.0	0.0	11.0	0.0	1968.0	189.0	6.0	...	42144.0	156.0	234.0	0.0
678	11/30/21	29.0	195.0	187.0	403.0	13.0	0.0	2332.0	398.0	4.0	...	39713.0	191.0	143.0	0.0
679	12/1/21	70.0	228.0	192.0	311.0	15.0	0.0	1881.0	502.0	8.0	...	47235.0	271.0	216.0	0.0

680 rows × 281 columns

Figure 2. Global COVID-19 Confirmed Cases Dataset

Country/Region		Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and Barbuda	Argentina	Armenia	Australia	...	United Kingdom.11	Uruguay	Uzbekistan	Vanuatu
0	Province/State	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Australian Capital Territory	...	NaN	NaN	NaN	NaN
1	1/23/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	1/24/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	1/25/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	1/26/20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
...
675	11/27/21	1.0	8.0	6.0	0.0	0.0	0.0	12.0	26.0	0.0	...	131.0	1.0	3.0	0.0
676	11/28/21	0.0	4.0	6.0	0.0	0.0	0.0	12.0	29.0	0.0	...	51.0	4.0	2.0	0.0
677	11/29/21	0.0	3.0	6.0	0.0	0.0	0.0	25.0	21.0	0.0	...	35.0	1.0	2.0	0.0
678	11/30/21	0.0	4.0	7.0	0.0	0.0	0.0	35.0	32.0	0.0	...	159.0	1.0	3.0	0.0
679	12/1/21	1.0	5.0	5.0	0.0	2.0	0.0	8.0	43.0	0.0	...	171.0	0.0	4.0	0.0

680 rows × 281 columns

Figure 3. Global COVID-19 Deaths Dataset

Exploratory Data Analysis

In order to have a clear understanding of the datasets, graphical representations were created to identify the trend of COVID-19 confirmed cases and deaths worldwide. Figure 4 below illustrates the trend of COVID-19 confirmed cases globally from January 2020 to December 2021.

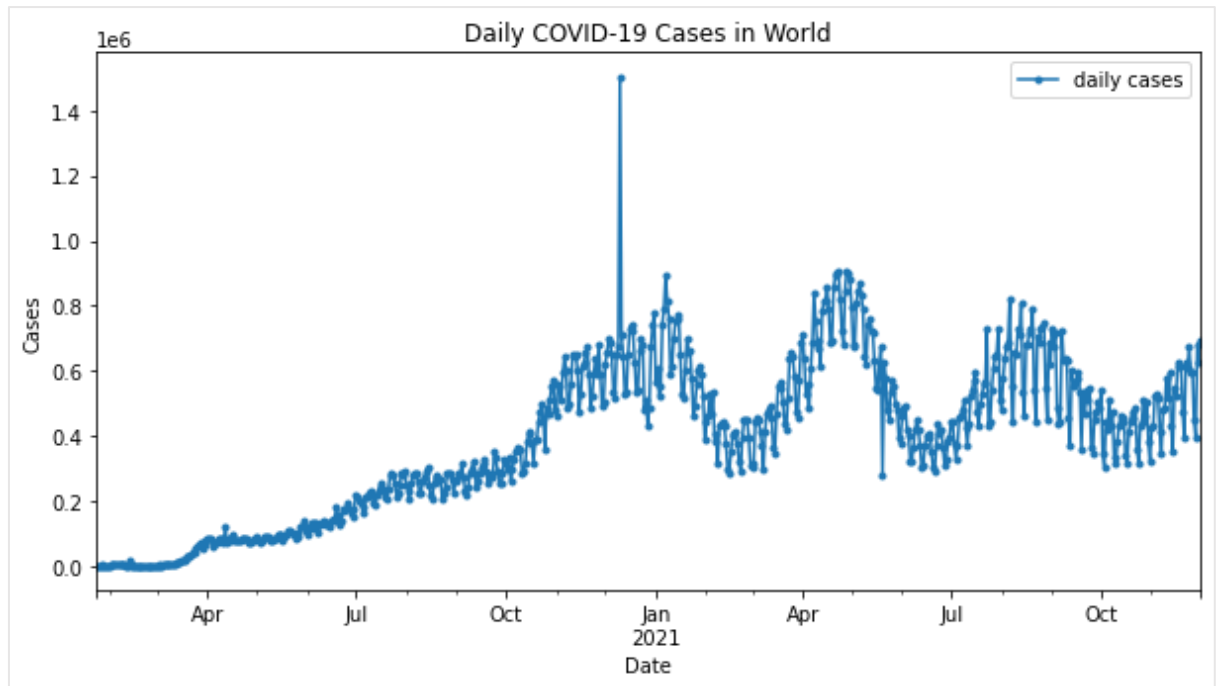


Figure 4. Global COVID-19 Confirmed Cases Trend

Figure 5 displays the worldwide COVID-19 fatality trend from January 2020 to December 2021.

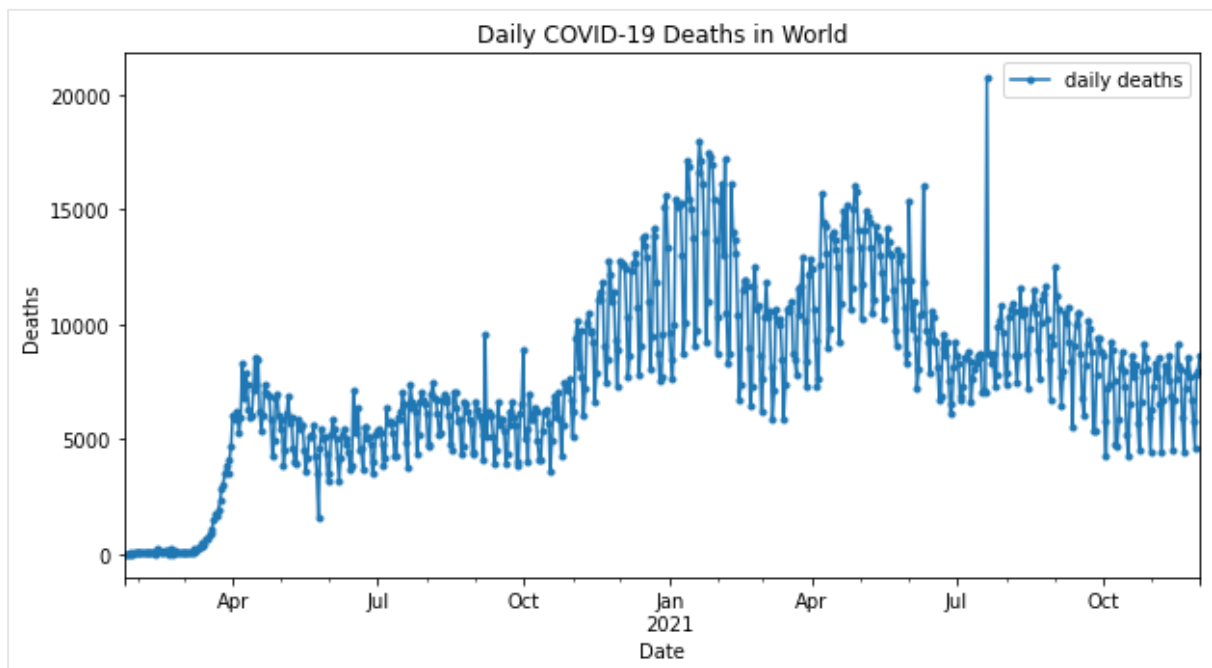


Figure 5. Global COVID-19 Deaths Trend

As can be observed in Figure 6, United States (US) has the highest number of COVID-19 confirmed cases followed by India, Brazil, United Kingdom (UK) and Russia.

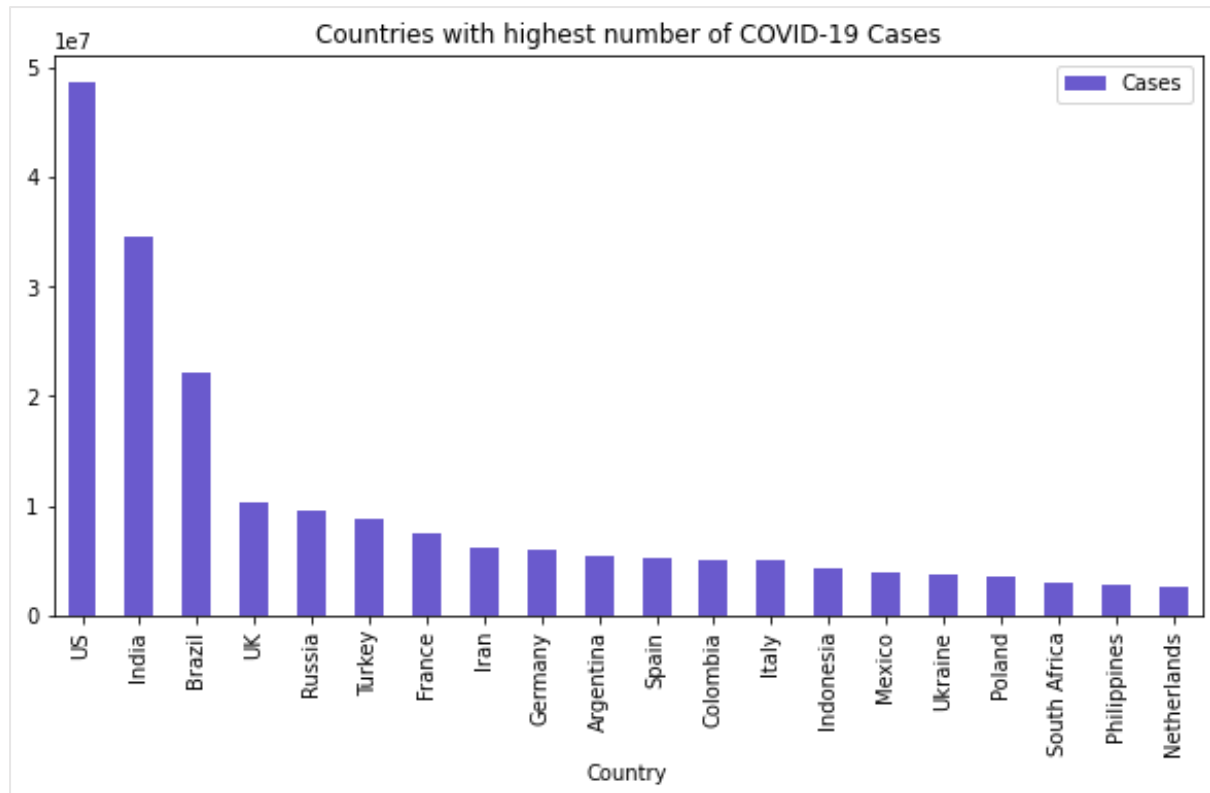


Figure 6. Top 20 Countries with Highest Number of COVID-19 Cases

Based on Figure 7, it can be observed that US also has the highest number of COVID-19 deaths followed by Brazil, India, Mexico and Russia.

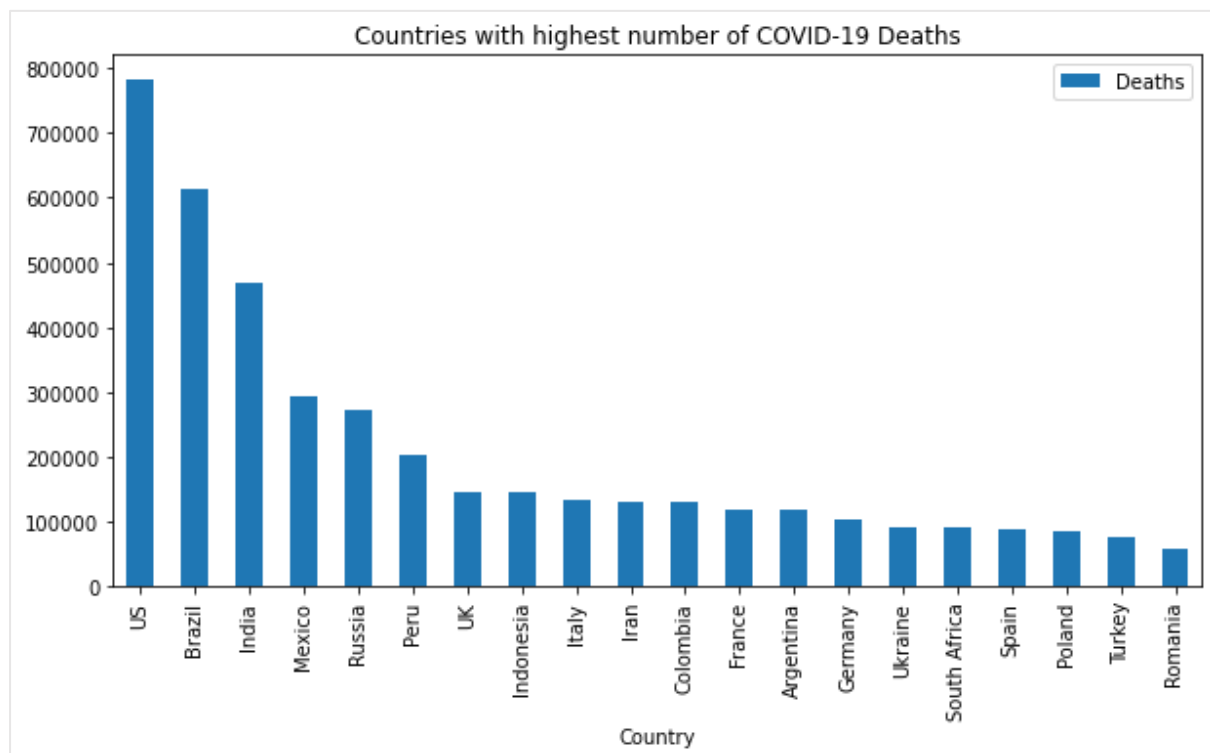


Figure 7. Top 20 Countries with Highest Number of COVID-19 Deaths

It is notable that US, Brazil, India and Russia are the 4 countries most affected by COVID-19 pandemic. Based on Figure 8 and 9 illustrated below, it can be observed that towards final quarter of 2021, the trend of COVID-19 confirmed cases and deaths in both US and Russia are on rise. Given the fact that numerous current relevant works are related to US COVID-19 forecast, this study is interested in Russia COVID-19 situation as it is not explored as much. Hence, the data of Confirmed Cases and Deaths in Russia were used to train the predictive models in this research.

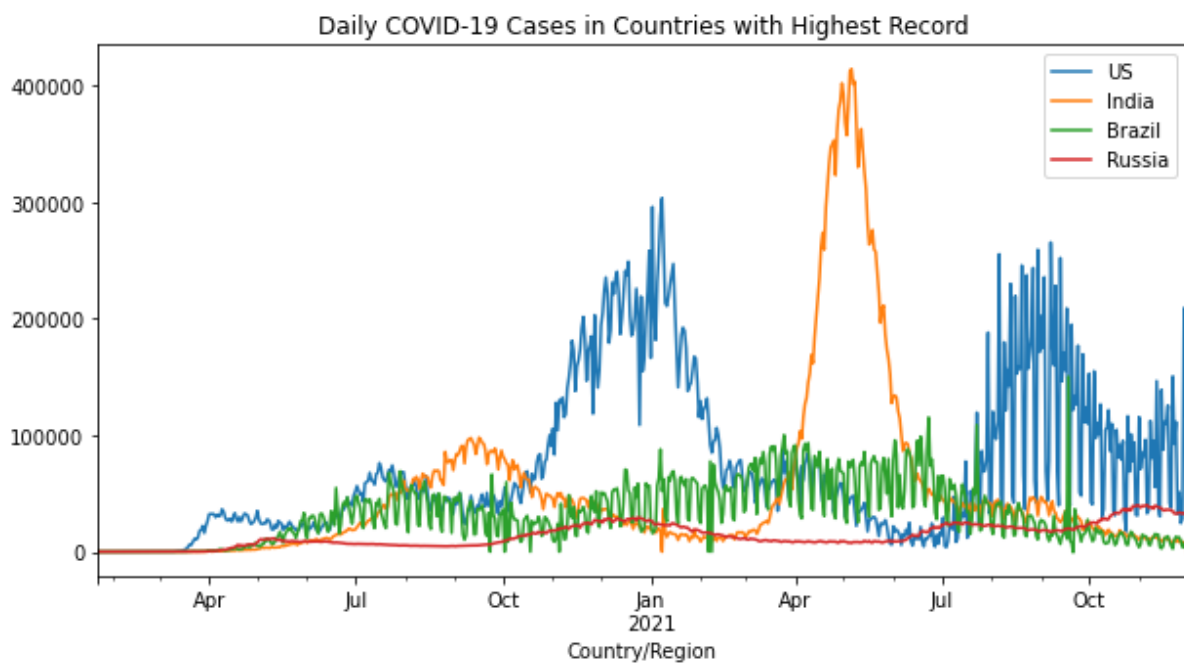


Figure 8. Dynamic in Top 4 Countries with Highest Number of COVID-19 Confirmed Cases

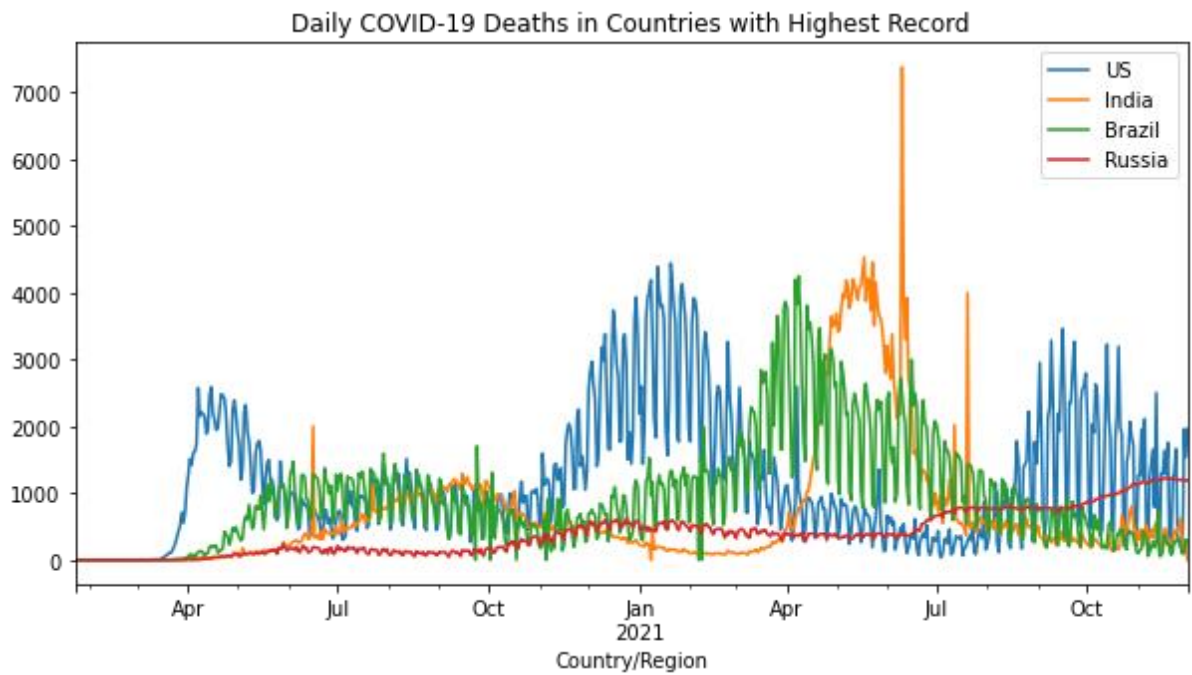


Figure 9. Dynamic in Top 4 Countries with Highest Number of COVID-19 Deaths

Figures 10 and 11 provide bigger and closer picture of the dynamic of COVID-19 confirmed cases and deaths in Russia respectively.

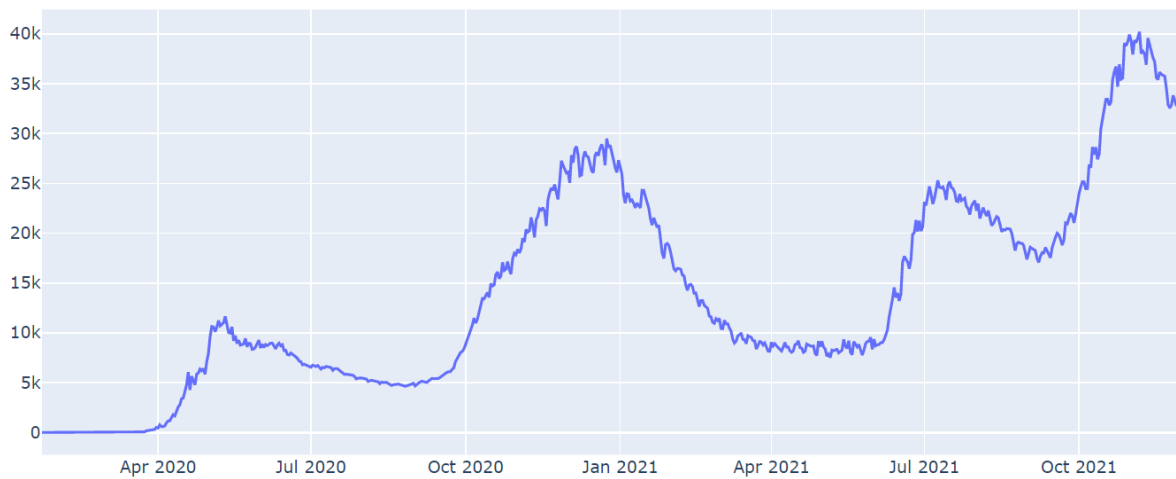


Figure 10. Trend of COVID-19 Confirmed Cases in Russia

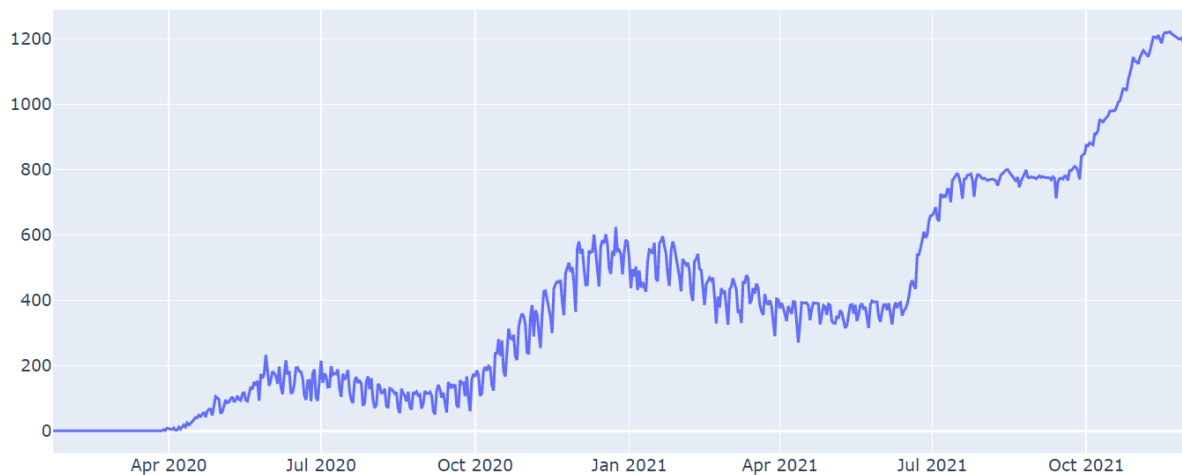


Figure 11. Trend of COVID-19 Deaths in Russia


Build Forecast Model

This study has used 3 prediction models to forecast COVID-19 confirmed cases and deaths in Russia.

1) Facebook Prophet

It is an open-source time series forecast framework created by Facebook. Its non-linear trends are based on daily, weekly, yearly and holiday effects seasonality (Yenidoğan, Çayır, Kozan, Dağ, & Arslan, 2018). Russia confirmed cases and deaths were extracted from global confirmed cases and deaths datasets. 'ds' and 'y' are given as the input for the FB Prophet model. 'ds' stands for datestamp while 'y' is the number of confirmed cases or deaths depending on the prediction. Figure 12 shows the data transformation of COVID-19 Confirmed Cases in Russia.

Country/Region		Russia
0	Province/State	NaN
1	1/23/20	0.0
2	1/24/20	0.0
3	1/25/20	0.0
4	1/26/20	0.0
...
675	11/27/21	33119.0
676	11/28/21	32786.0
677	11/29/21	33170.0
678	11/30/21	31990.0
679	12/1/21	32196.0
680 rows × 2 columns		



ds		y
1	1/23/20	0.0
2	1/24/20	0.0
3	1/25/20	0.0
4	1/26/20	0.0
5	1/27/20	0.0
...
675	11/27/21	33119.0
676	11/28/21	32786.0
677	11/29/21	33170.0
678	11/30/21	31990.0
679	12/1/21	32196.0
679 rows × 2 columns		

Figure 12. Data Transformation of Confirmed Cases in Russia

Figure 13 shows below the modelling of Facebook Prophet for COVID-19 Confirmed Cases in Russia. The similar steps were performed for forecast of COVID-19 deaths in Russia.

```
#Create FBProphet Model and visualize the model output for Russian confirmed cases data

class Fbprophet(object):
    def fit(self,data):

        self.data = data
        self.model = Prophet(weekly_seasonality=True,daily_seasonality=False,yearly_seasonality=False)
        self.model.fit(self.data)

    def forecast(self,periods,freq):

        self.future = self.model.make_future_dataframe(periods=periods,freq=freq)
        self.df_forecast = self.model.predict(self.future)

    def plot(self,xlabel="Months",ylabel="Values"):

        self.model.plot(self.df_forecast,xlabel=xlabel,ylabel=ylabel,figsize=(9,4))
        self.model.plot_components(self.df_forecast,figsize=(9,6))

    def R2(self):
        return r2_score(self.data.y, self.df_forecast.yhat[:len(RUS_confirmed_cases)])

model1 = Fbprophet()
model1.fit(RUS_confirmed_cases)
model1.forecast(30,"D")
model1.plot()
```

Figure 13. FB Prophet Model

The monthly forecast (the dark blue line) of COVID-19 confirmed cases and deaths in Russia are illustrated in Figures 14 and 15 respectively.

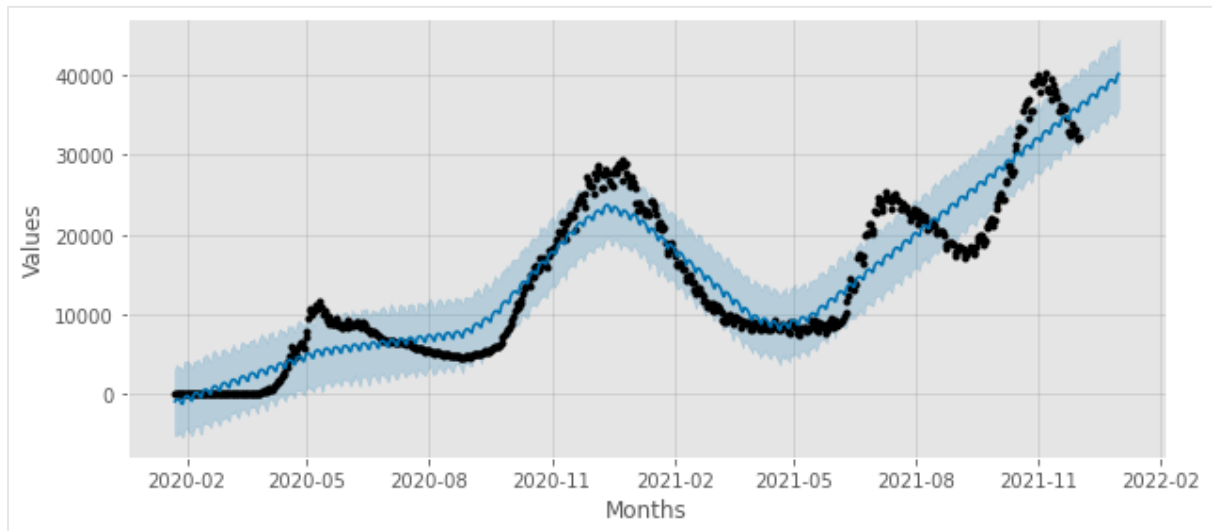


Figure 14. Forecast of COVID-19 Confirmed Cases in Russia

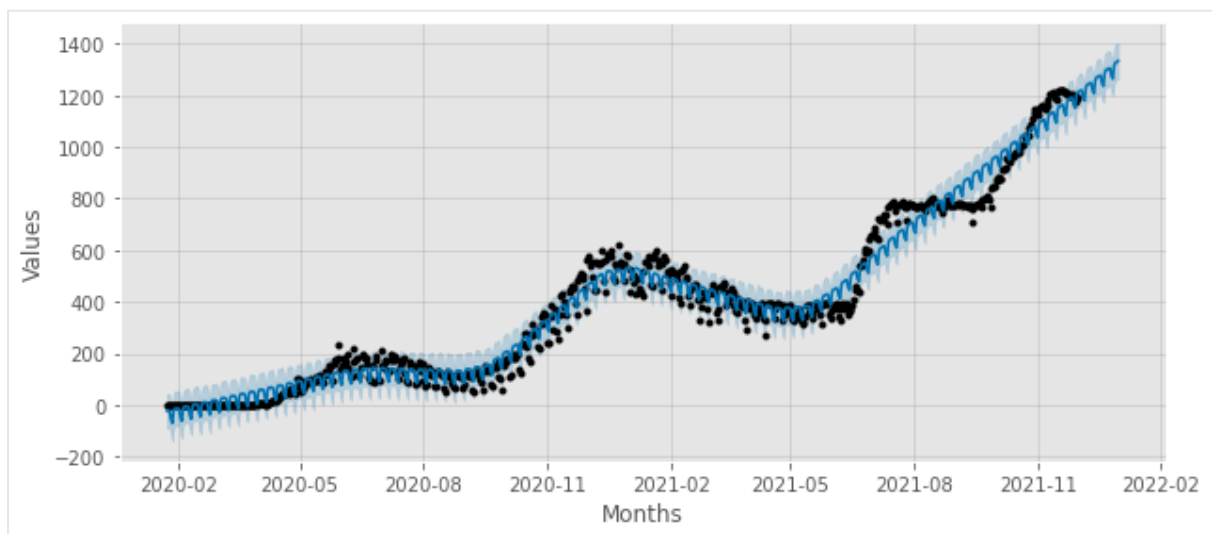


Figure 15. Forecast of COVID-19 Deaths in Russia

The weekly forecast of COVID-19 confirmed cases and deaths in Russia are illustrated in Figures 16 and 17 respectively.

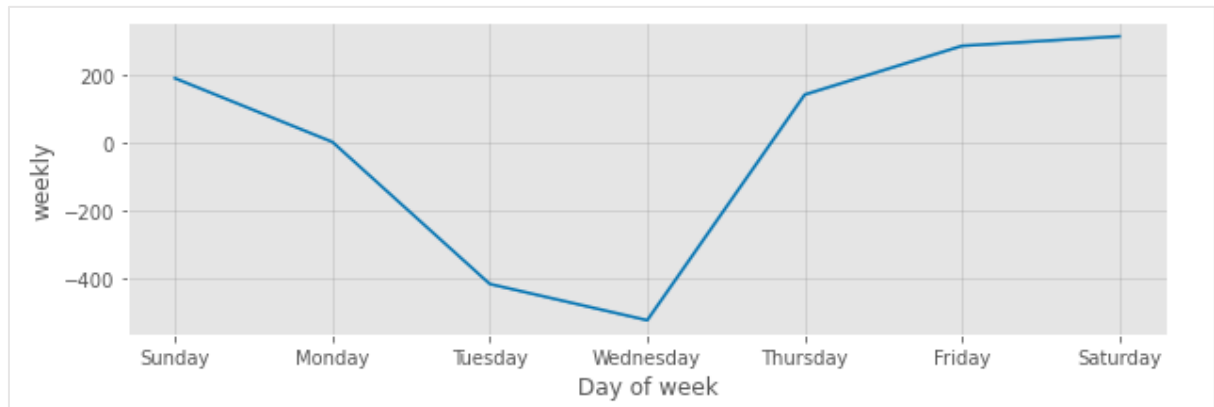


Figure 16. Weekly COVID-19 Confirmed Cases in Russia

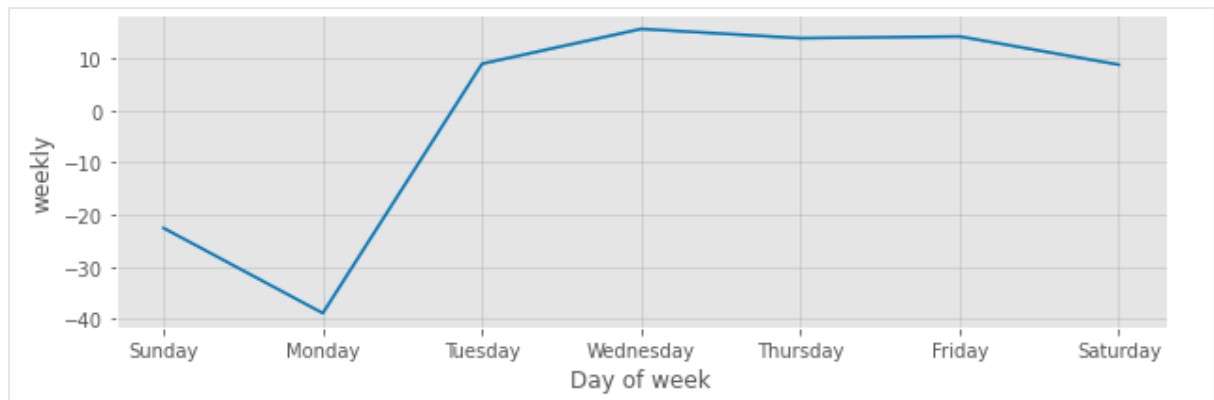


Figure 17. Weekly COVID-19 Deaths in Russia

2) Linear Regression

It is a simple model which can be used to discover the relation between dependent variable and an independent variable (Pandeya, Chaudhary, Gupta, & Pal, 2020). Figures 18 and 19 show the implementation of Linear Regression to forecast COVID-19 confirmed cases and deaths in Russia respectively.

```
X = RUS_confirmed_cases[['month', 'year', 'dayofmonth']]
Y = RUS_confirmed_cases[['Confirmed cases']]

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3)

from sklearn.linear_model import LinearRegression
LR=LinearRegression()
LR.fit(X_train,Y_train)

LinearRegression()

LR.score(X_test,Y_test)

0.6691953600139898
```

Figure 18. Linear Regression (Confirmed Cases)

```
X = RUS_deaths[['month', 'year', 'dayofmonth']]
Y = RUS_deaths[['Deaths']]

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3)

from sklearn.linear_model import LinearRegression
LR=LinearRegression()
LR.fit(X_train,Y_train)

LinearRegression()

LR.score(X_test,Y_test)

0.8523768051424707
```

Figure 19. Linear Regression (Deaths)

Figures 20 and 21 display glimpse of data used for the COVID-19 confirmed cases and deaths in Russia predictive models.

	Confirmed cases	month	year	dayofmonth
Date				
2020-01-23	0.0	1	2020	23
2020-01-24	0.0	1	2020	24
2020-01-25	0.0	1	2020	25
2020-01-26	0.0	1	2020	26
2020-01-27	0.0	1	2020	27
...
2021-11-27	33119.0	11	2021	27
2021-11-28	32786.0	11	2021	28
2021-11-29	33170.0	11	2021	29
2021-11-30	31990.0	11	2021	30
2021-12-01	32196.0	12	2021	1

679 rows × 4 columns

Figure 20. Dataset for COVID-19 Confirmed Cases Prediction

	Deaths	month	year	dayofmonth
Date				
2020-01-23	0.0	1	2020	23
2020-01-24	0.0	1	2020	24
2020-01-25	0.0	1	2020	25
2020-01-26	0.0	1	2020	26
2020-01-27	0.0	1	2020	27
...
2021-11-27	1203.0	11	2021	27
2021-11-28	1190.0	11	2021	28
2021-11-29	1178.0	11	2021	29
2021-11-30	1195.0	11	2021	30
2021-12-01	1191.0	12	2021	1

679 rows × 4 columns

Figure 21. Dataset for COVID-19 Deaths Prediction

The forecast of COVID-19 confirmed cases and deaths in Russia using Linear Regression are illustrated in Figures 22 and 23 respectively.

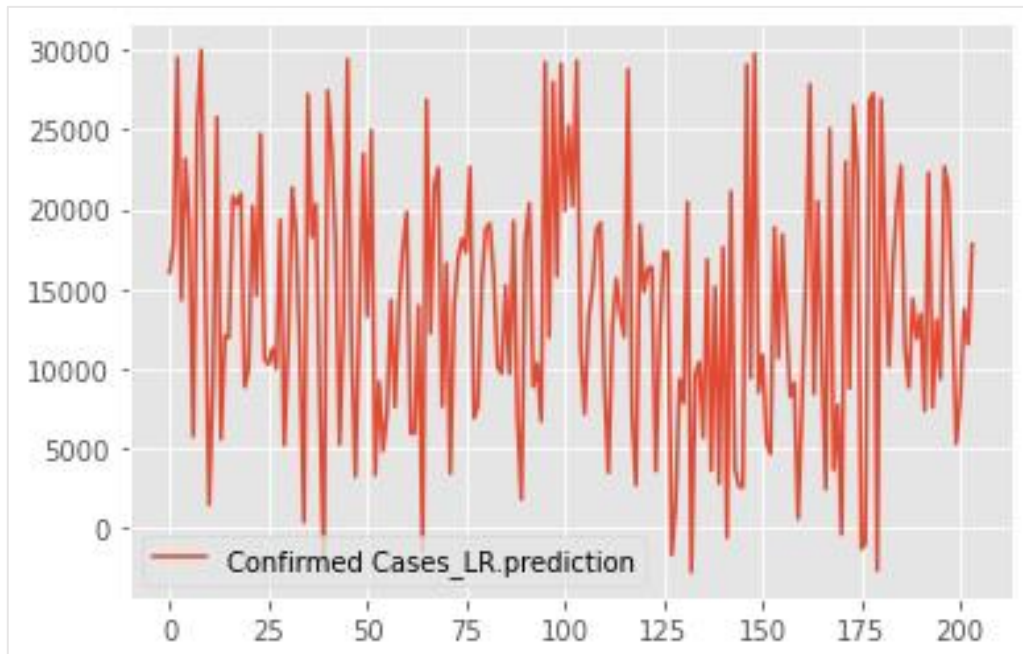


Figure 22. Prediction of COVID-19 confirmed cases in Russia

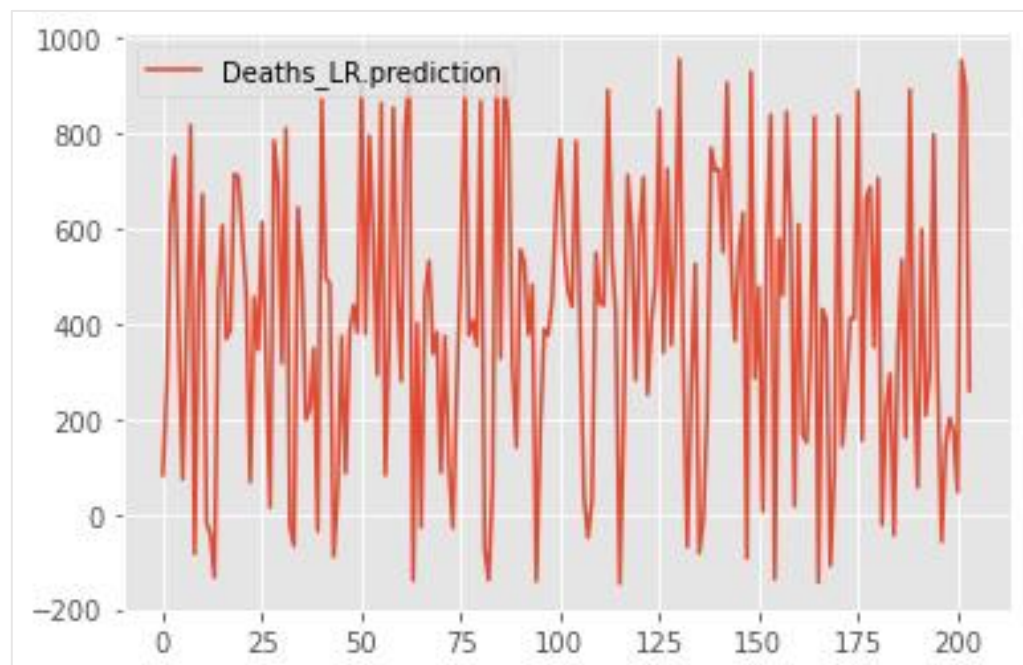


Figure 23. Prediction of COVID-19 deaths in Russia

3) Random Forest

It is a bagging ensemble-based model which is fast, robust and able to handle the randomness of the time series (Ribeiro, Silva, Mariani, & Coelho, 2020). This model was implemented with the same train and test dataset used in Linear Regression model. Figure 24 presents the implementation of Random Forest to forecast COVID-19 confirmed cases and deaths in Russia.

```
from sklearn.ensemble import RandomForestClassifier
tree_model= RandomForestClassifier(n_estimators=100, max_depth=200,
                                  random_state=1)
tree_model.fit(X_train,Y_train)
RF_prediction = tree_model.predict(X_test)
```

Figure 24. Random Forest

The forecast of COVID-19 confirmed cases and deaths in Russia using Random Forest model are illustrated in Figures 25 and 26 respectively.

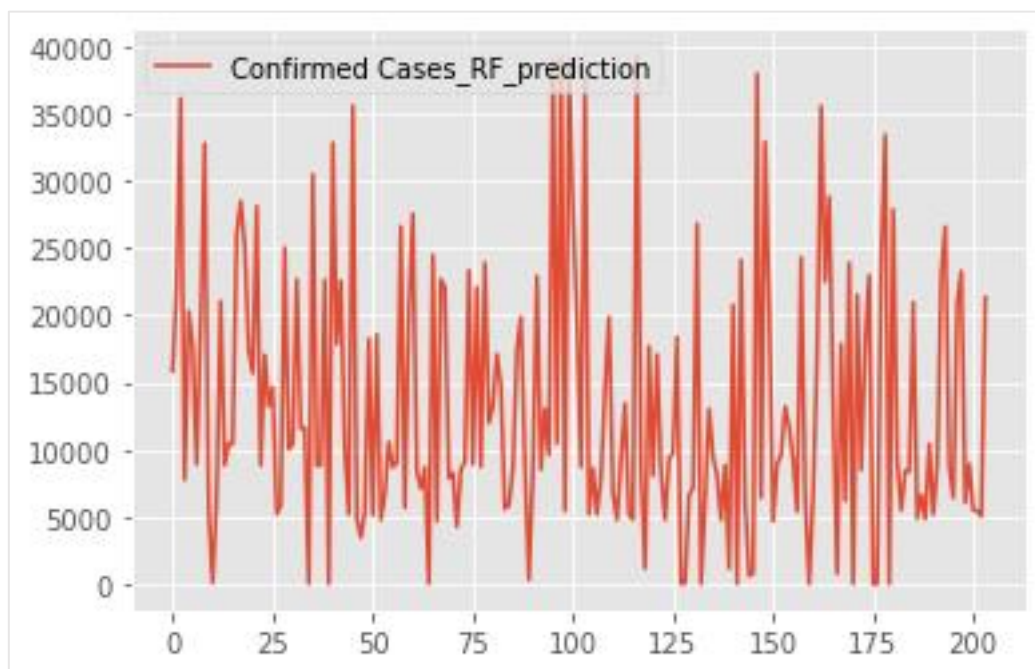


Figure 25. Prediction of COVID-19 Confirmed Cases in Russia

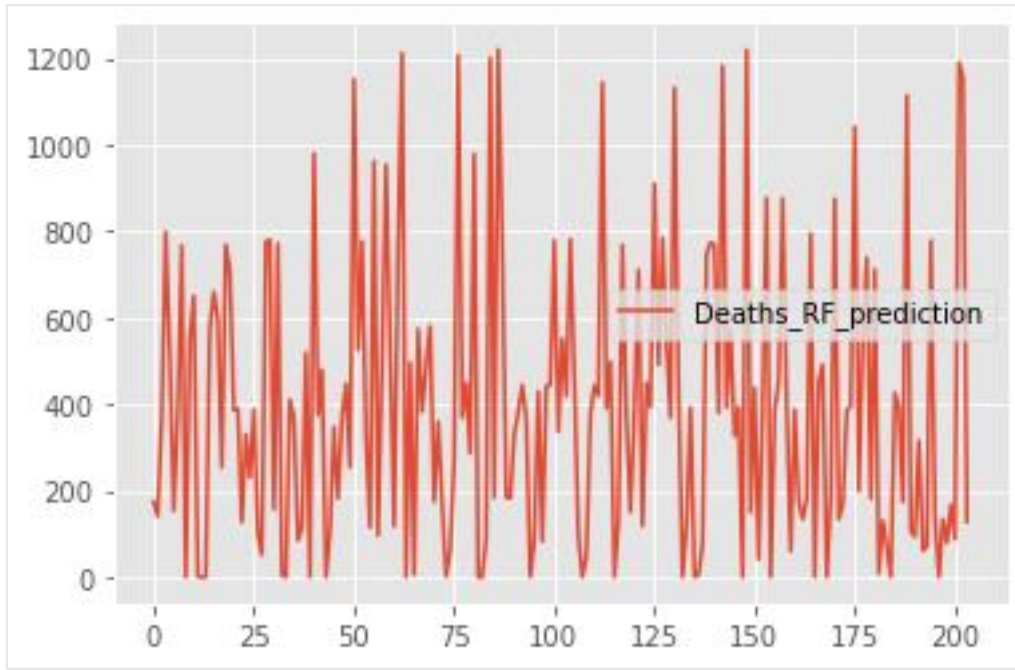


Figure 26. Prediction of COVID-19 Deaths in Russia

Model Evaluation

There are various evaluation metrics that are used to evaluate the predictive models such as Confidence Interval, Coefficient of Determination (R^2), Accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Relative Squared Error (RRSE) and Mean Absolute Percentage Error (MAPE). This study was evaluated in terms of Coefficient of Determination (R^2).

The Coefficient of Determination defines the degree of variance in the dependent variable which are number of COVID-19 confirmed cases and deaths that can be explained by the independent variable which are Day of Month, Month and Year. Coefficient of Determination describes how well the data provided fits the model using the following formula (Car, Šegota, Anđelić, Lorencin, & Mrzljak, 2020):

$$R^2 = 1 - \frac{S_{RESIDUAL}}{S_{TOTAL}} = 1 - \frac{\sum_{i=0}^m (y_i - \hat{y}_i)^2}{\sum_{i=0}^m (y_i - 1/m \sum_{i=0}^m y_i)^2},$$

Coefficient of Determination is defined in the range [0,1], whereby if the value is 0.0 the data provided does not fit the model at all while the value 1.0 fits the model perfectly.

Results and Analysis

As can be observed in Figure 23, Random Forest and Facebook Prophet models has R^2 score of 0.89 and 0.894 respectively.

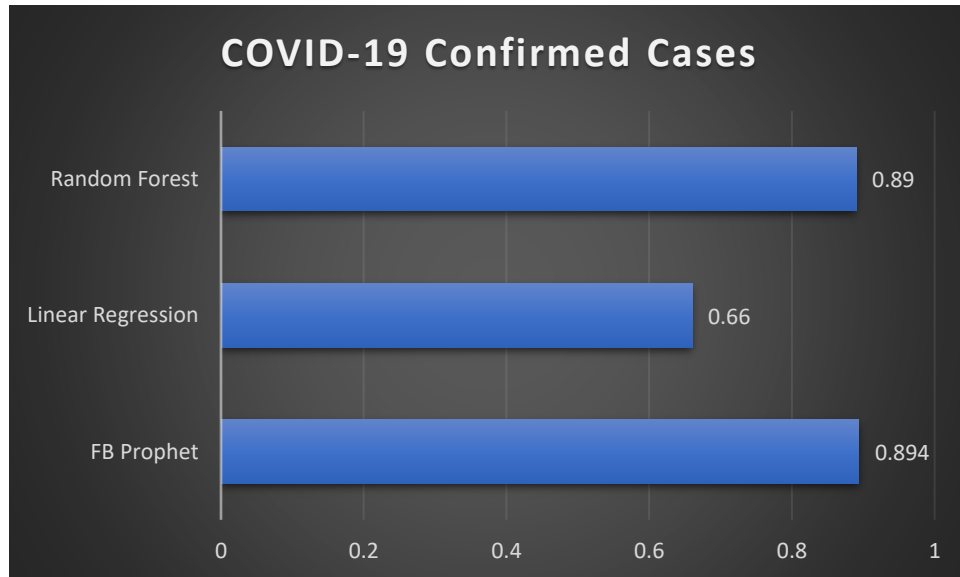


Figure 27. R^2 score for Russia COVID-19 Confirmed Cases Prediction

Figure 24 below presents the R^2 score comparison for Russia COVID-19 deaths prediction. Random Forest and Facebook Prophet models has R^2 score of 0.98 and 0.975 respectively. This indicates that these predictive models fit both the Russia COVID-19 confirmed cases and deaths datasets better compared to Linear Regression model and are suitable for the studied prediction.

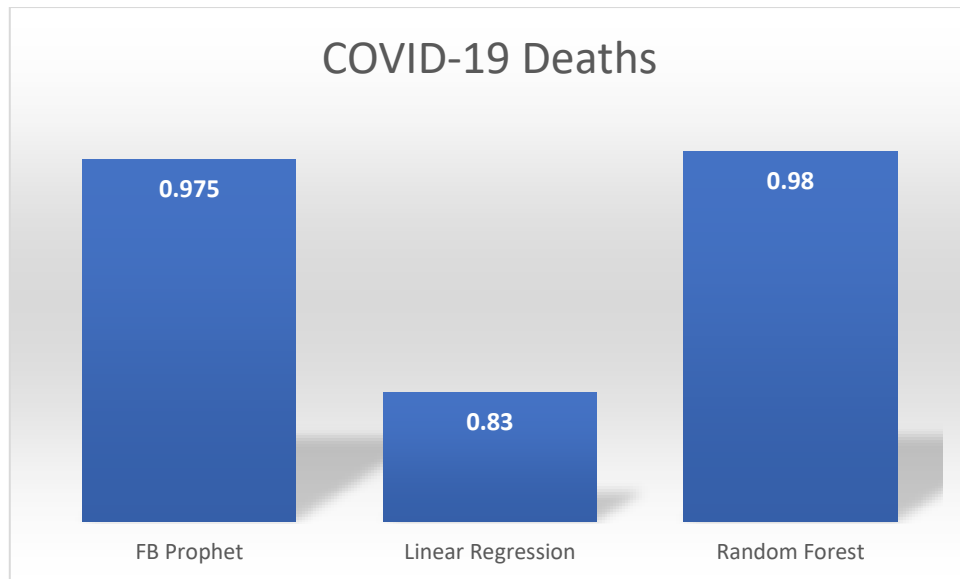


Figure 28. R2 score for Russia COVID-19 Deaths Prediction

Code Repository

The source codes for this study are available here:

<https://github.com/TarashiniSuthesan/Forecast-of-Covid-19-Cases-in-Russia>

References

- Battineni, G., & Chintalapudi, N. (2020). Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model. *Applied Computing and Informatics*.
- Car, Z., Šegota, S. B., Anđelić, N., Lorencin, I., & Mrzljak, V. (2020). Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. *Computational and Mathematical Methods in Medicine*.
- Coronavirus. (2022, January 12). Retrieved from Worldometer: <https://www.worldometers.info/coronavirus/>
- COVID-19 data from John Hopkins University. (2021, December 1). Retrieved from Kaggle: <https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university>
- Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed*, 157-160.
- Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model. *Big Data Mining and Analytics*.
- Kumar, N., & Susan, S. (2020). COVID-19 Pandemic Prediction using Time Series Forecasting Model. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.
- Majhi, R., Thangeda, R., Sugasi, R. P., & Kumar, N. (2021). Analysis and prediction of COVID-19 trajectory: A machine learning approach. *Journal of Public Affairs*.
- Pandeya, G., Chaudhary, P., Gupta, R., & Pal, S. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India.
- Ribeiro, M. H., Silva, R. G., Mariani, V. C., & Coelho, L. S. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals*.
- Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons and Fractals*.
- Xu, Y., Su, S., Jiang, Z., Guo, S., Lu, Q., Liu, L., . . . Lu, L. (2021). Prevalence and Risk Factors of Mental Health Symptoms and Suicidal Behavior Among University Students in Wuhan, China During the COVID-19 Pandemic. *Front. Psychiatry*.

Yenidođan, I., ayir, A., Kozan, O., Dađ, T., & Arslan, . (2018). Bitcoin Forecasting Using ARIMA and PROPHET. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, (pp. 621-624).